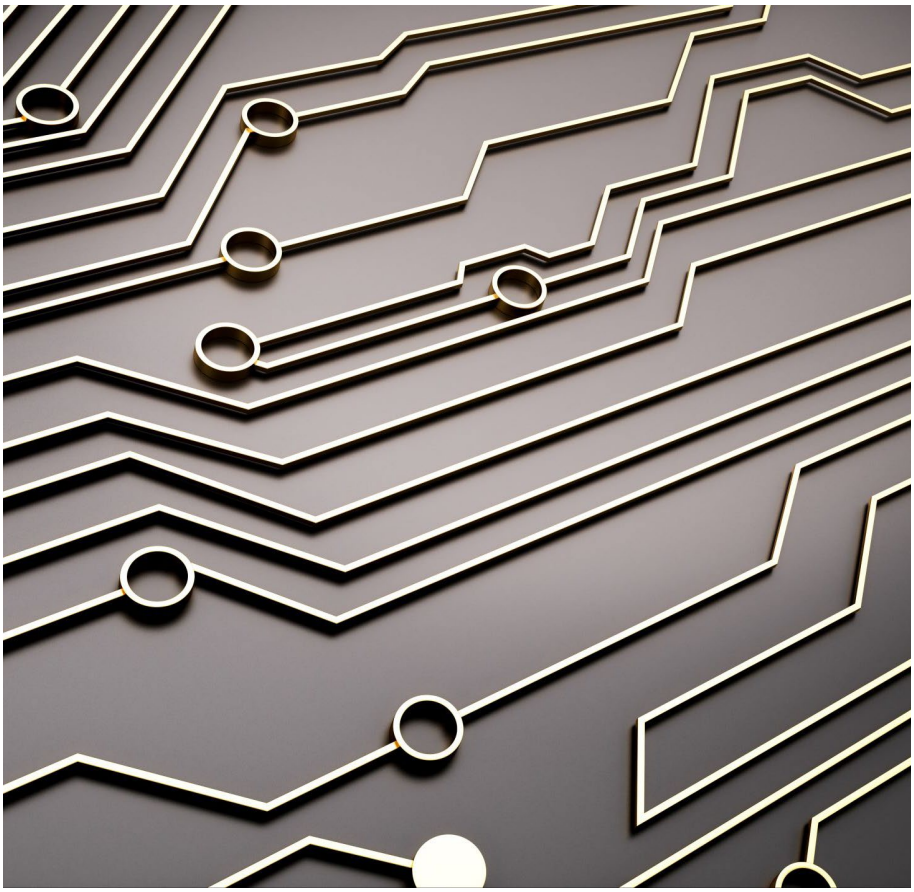

Anthony Almudevar, PhD

INTRODUCTION TO STATISTICAL METHODOLOGY

A First Course in Statistics

(Exercises with Solutions)

Preface. This document contains practice exercises intended to supplement the volume *Introduction to Statistical Methodology* by Anthony Almudevar.

Solutions are given in blue font for all problems. Plots which are part of a solution are distinguished by blue captions.

The organization of the problems does not match exactly the chapter organization of *Introduction to Statistical Methodology*, as many questions rely on concepts covered in multiple chapters. However, following the main text, problems dealing primarily with probability or statistical methods are listed in separate chapters.

In addition, a third chapter titled "Data Projects" contains more extended problems which make use of datasets from several data repositories accessible from the `R` platform. These cover multiple topics, and are intended to give the reader practice and training in actual data analysis.

Contents

Chapter 1

Probability

1.1 Basic Rules of Probability

Problem 1.1 Three 6-sided dice are tossed independently. Label the dice red, green and blue. Suppose we define the following events:

$$\begin{aligned} A_1 &= \{\text{red dice} = \text{green dice}\} \\ A_2 &= \{\text{red dice} = \text{blue dice}\} \\ A_3 &= \{\text{green dice} = \text{blue dice}\}. \end{aligned}$$

Calculate the probabilities:

$$\begin{aligned} P(A_i), &\quad i = 1, 2, 3; \\ P(A_i \cap A_j), &\quad i \neq j; \\ P(A_1 \cap A_2 \cap A_3). & \end{aligned}$$

Are the events A_1, A_2, A_3 independent? Are they pairwise independent?

SOLUTION: There are $6 \times 6 = 36$ outcomes involving two dice, and the two dice are equal for 6 of them:

$$P(A_i) = 6/36 = 1/6$$

for $i = 1, 2, 3$. Next, note that for any $i \neq j$

$$A_i \cap A_j = \{\text{all three dice are equal}\} = A_1 \cap A_2 \cap A_3.$$

There are $6 \times 6 \times 6 = 216$ outcomes involving three dice, and the three dice are equal for 6 of them:

$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = P(A_1 \cap A_2 \cap A_3) = 6/216 = 1/36.$$

This means

$$(1/6)^3 = P(A_1)P(A_2)P(A_3) \neq P(A_1 \cap A_2 \cap A_3) = 1/36,$$

so that A_1, A_2, A_3 *are not independent.* However, for any $i \neq j$,

$$(1/6)^2 = P(A_i)P(A_j) = P(A_i \cap A_j) = 1/36,$$

so that A_1, A_2, A_3 *are pairwise independent.*

Problem 1.2 A coin is tossed twice, independently. Define the three events:

$$\begin{aligned} A_1 &= \{ \text{ first toss is Heads } \} \\ A_2 &= \{ \text{ second toss is Heads } \} \\ A_3 &= \{ \text{ two outcomes are the same } \}. \end{aligned}$$

Prove that these events are pairwise independent but not independent.

SOLUTION: The sample space is $S = \{HH, HT, TH, TT\}$. We have

$$\begin{aligned} A_1 &= \{HH, HT\} \\ A_2 &= \{HH, TH\} \\ A_3 &= \{HH, TT\}. \end{aligned}$$

We have $P(A_i) = 1/2$ for $i = 1, 2, 3$.

To prove **pairwise independence** it suffices to show that $P(A_iA_j) = P(A_i)P(A_j) = 1/4$ for each pair $i \neq j$. We have

$$\begin{aligned} P(A_1A_2) &= P(\{HH\}) = 1/4, \\ P(A_1A_3) &= P(\{HH\}) = 1/4, \\ P(A_2A_3) &= P(\{HH\}) = 1/4. \end{aligned}$$

establishing pairwise independence.

To disprove **independence** we note

$$P(A_1 \cap A_2 \cap A_3) = P(\{HH\}) = 1/4 \neq P(A_1)P(A_2)P(A_3) = 1/8.$$

Problem 1.3 The hour hand on a 12-point clock is positioned at 12. The hand moves backwards or forwards one position with equal probability N times. All moves are independent. Determine the probability that the hand rests at 3 if:

(a) $N = 9$,
(b) $N = 10$,
(c) $N = 19$.

SOLUTION: An outcome consists of a sequence of N (F)orward or (B)ackwards directions, say, $BFF \ldots BFB$. The problem is best approached by defining a random outcome X, defined as

$$X = \text{The number of } F\text{'s in the sequence.}$$

This is because the problem can be resolved by knowing X. By the rule of product, each outcome has the same probability $1/2^N$. The number of sequences of length N that have exactly k F's is $\binom{N}{k}$, since we are making an unordered selection of k positions for the F's from the N available. So,

$$P(X = k) = \binom{N}{k}(1/2)^N.$$

To determine if the final position of the hour is 3 we use the rule

$$\begin{aligned}(\#F - \#B) \mod 12 &= 3, \quad \text{or equivalently} \\ (2X - N) \mod 12 &= 3.\end{aligned}$$

Note that $x = y \mod n$ if $y - x$ is divisible by n. So $-9 \mod 12 = 3$. The order of the moves does not matter.

(a) There are two ways for the hour hand to rest on position 3 if $N = 9$. Either it moves forward 6 and backwards 3 positions, or it moves backwards 9 positions. The values of X for which which this occurs are 0 and 6, so

$$P(\text{ hand rests on 3 }) = P(X = 0) + P(X = 6) = \frac{\binom{9}{0} + \binom{9}{6}}{2^9} = \frac{85}{512} \approx 0.166.$$

(b) The hour hand cannot rest on 3 if $N = 10$, so

$$P(\text{ hand rests on 3 }) = 0.$$

(c) For $N = 19$ the hour hand rests on 3 if $X = 5, 11$ or 17. This means

$$\begin{aligned}P(\text{ hand rests on 3 }) &= P(X = 5) + P(X = 11) + P(X = 17) \\ &= \frac{\binom{19}{5} + \binom{19}{11} + \binom{19}{17}}{2^{19}} \\ &= \frac{87381}{524288} \approx 0.167.\end{aligned}$$

Note that the answers to (a) and (c) are both close to, but not exactly, 1/6.

Problem 1.4 (The Inclusion-Exclusion Principle). If events $A_1, \ldots, A_n$ are mutually exclusive, the probability of the union is

$$P(\cup_{i=1}^n A_i) = P(A_1) + \ldots + P(A_n).$$

If they are not mutually exclusive, then calculation of the probability of their union can become quite complex. For two events, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2).$$

This can be extended to n events using the *inclusion-exclusion identity*

$$\begin{aligned}P(\cup_{i=1}^n A_i) &= \sum_i P(A_i) \\ &\quad - \sum_{i<j} P(A_i A_j) \\ &\quad + \sum_{i<j<k} P(A_i A_j A_k) \\ &\quad \vdots \\ &\quad -1^{n+1} P(A_1 A_2 \ldots A_n).\end{aligned} \tag{1.1}$$

(a) Write explicitly the *inclusion-exclusion identity* for $n = 3$.
(b) Suppose any integer from 1 to 100 inclusive is chosen at random with equal probability, which will be denoted N. What is the probability that N is divisible by at least one of 3, 4 or 7?

SOLUTION:

(a) We have directly

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

(b) Let E_i be the event that N is divisible by i. So,

$$E_7 = \{N \in \{7, 14, 21, 28, 35, 42, 49, 56, 63, 70, 77, 84, 91, 98\}\},$$

giving

$$P(E_7) = \frac{|E_7|}{100} = \frac{14}{100}.$$

Also, we must have

$$E_i \cap E_j = E_{i\times j}, \;\; E_i \cap E_j \cap E_k = E_{i\times j\times k}.$$

So, by the inclusion-exclusion identity we have

$$\begin{aligned} P(E_3 \cup E_4 \cup E_7) &= P(E_3) + P(E_4) + P(E_7) - P(E_3E_4) - P(E_3E_7) - P(E_4E_7) + P(E_3E_4E_7) \\ &= P(E_3) + P(E_4) + P(E_7) - P(E_{12}) - P(E_{21}) - P(E_{28}) + P(E_{84}). \\ &= \frac{1}{100} \times (33 + 25 + 14 - 8 - 4 - 3 + 1) = \frac{58}{100} \end{aligned}$$

Problem 1.5 Make use of the inclusion-exclusion principle (Problem 1.4) to answer the following question. Suppose any integer from 1 to 75 inclusive is chosen at random with equal probability, which will be denoted N. What is the probability that N is divisible by at least one of 5, 7 or 11?

SOLUTION: Let E_i be the event that N is divisible by i. So,

$$E_5 = \{N \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75\}\},$$

giving

$$P(E_5) = \frac{|E_5|}{73} = \frac{15}{75}.$$

Also, we must have (since 5, 7 and 11 are prime numbers)

$$E_i \cap E_j = E_{i\times j}, \;\; E_i \cap E_j \cap E_k = E_{i\times j\times k}.$$

So, by the inclusion-exclusion identity we have

$$\begin{aligned} P(E_5 \cup E_7 \cup E_{11}) &= P(E_5) + P(E_7) + P(E_{11}) - P(E_5E_7) - P(E_5E_{11}) - P(E_7E_{11}) + P(E_5E_7E_{11}) \\ &= P(E_5) + P(E_7) + P(E_{11}) - P(E_{35}) - P(E_{55}) - P(E_{77}) + P(E_{385}). \\ &= \frac{1}{75} \times (15 + 10 + 6 - 2 - 1 - 0 + 0) = \frac{28}{75}. \end{aligned}$$

Problem 1.6 Make use of the inclusion-exclusion principle (Problem 1.4) to answer the following question. Suppose any integer from 1 to 105 inclusive is chosen at random with equal probability, which will be denoted N. What is the probability that N is divisible by at least one of 2, 9 or 13?

SOLUTION: Let E_i be the event that N is divisible by i. So,

$$E_2 = \{N \in \{2, 4, \ldots, 102, 104\}\},$$

giving

$$P(E_2) = \frac{|E_2|}{105} = \frac{52}{105}.$$

Similarly,

$$\begin{aligned} P(E_9) &= \frac{|\{9, 18, \ldots, 99\}|}{105} = \frac{11}{105} \\ P(E_13) &= \frac{|\{13, 26, \ldots, 104\}|}{105} = \frac{8}{105}. \end{aligned}$$

In general,

$$P(E_i) = \frac{\text{floor}(i/105)}{105}.$$

Also, note that 2, 9 and 13 have no common factors, **and if this condition holds**, we have

$$E_i \cap E_j = E_{i\times j}, \quad E_i \cap E_j \cap E_k = E_{i\times j\times k}.$$

So, by the inclusion-exclusion identity we have

$$\begin{aligned} P(E_2 \cup E_9 \cup E_{13}) &= P(E_2) + P(E_9) + P(E_{13}) - P(E_2E_9) - P(E_2E_{13}) - P(E_9E_{13}) + P(E_2E_9E_{13}) \\ &= P(E_2) + P(E_9) + P(E_{13}) - P(E_{18}) - P(E_{26}) - P(E_{117}) + P(E_{234}). \\ &= \frac{1}{105} \times (52 + 11 + 8 - 5 - 4 - 0 + 0) = \frac{62}{105}. \end{aligned}$$

Problem 1.7 Make use of the inclusion-exclusion principle (Problem 1.4) to answer the following question. A bin contains n balls labeled $1, \ldots, n$. The balls are selected in order, at random. We say the ball labeled k was *selected correctly* if it is in position k of the selection order. For example, if $n = 5$, and the selection order was $5, 2, 1, 4, 3$ then balls 2 and 4 were selected correctly.

(a) Give an expression in terms of n for the probability that a specific ball was selected correctly.
(b) Suppose B is a specific subset of $m \leq n$ balls Give an expression in terms of n and m for the probability that all balls in B were selected correctly.
(c) Use the inclusion-exclusion principle to derive a formula for the probability that *no* ball is selected correctly (such a permutation is known as a *derangement*).
(d) Write an `R` program to calculate the probability of a derangement for $n = 1, 2, \ldots, 25$. Comment briefly on the resulting sequence.

SOLUTION:

(a) There are $D = n!$ possible selections. Let A_j be the set of ordered selections in which ball j is selected correctly. We can select from A_j using 2 tasks. First, put ball j in its correct place ($n_1 = 1$). Second, select positions for the remaining balls ($n_2 = (n-1)!$). Note that more than one ball can be selected correctly in A_j, as long as ball j is selected correctly. So $N = 1 \times (n-1)!$, giving

$$P(A_j) = \frac{N}{D} = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

(b) Again, we do the selection in 2 tasks. First, place the balls in B in their correct place ($n_1 = 1$). Then permute the remaining balls ($n_2 = (n-m)!$). Let A_B be the set of ordered selections in which all balls $j \in B$ are selected correctly. This gives

$$P(A_B) = \frac{N}{D} = \frac{(n-m)!}{n!}. \tag{1.2}$$

(c) Let A_j be the event that ball j is selected correctly. Then

$$P(derangement) = 1 - P(\cup_{j=1}^{n} A_j).$$

Write equation (1.1)

$$P(\cup_{j=1}^{n} A_j) = \sum_{j=1}^{n} (-1)^{j-1} S_j,$$

where S_j are the appropriate sums. Each term in S_j is equal to $(n-j)!/n!$ from (1.2). Furthermore, each sum S_j contains $\binom{n}{j}$ terms, giving

$$P(\cup_{i=1}^{n} A_i) = \sum_{j=1}^{n} (-1)^{j-1} \binom{n}{j} \frac{(n-j)!}{n!} = \sum_{j=1}^{n} (-1)^{j-1} \frac{1}{j!}.$$

If we subtract from 1, we get

$$P(derangement) = \sum_{j=0}^{n} (-1)^{j} \frac{1}{j!}.$$

(d) The following code performs the required calculations:

```
> nmax = 25
> prob = rep(NA,nmax)
>
> for (n in 1:nmax) {
+   sm = 0
+   for (i in 1:n) {sm = sm + ((-1)^(i-1))/prod(1:i)}
+   prob[n] = 1 - sm
+ }
> prob
 [1] 0.0000000 0.5000000 0.3333333 0.3750000 0.3666667 0.3680556 0.3678571 0.3678819
 [9] 0.3678792 0.3678795 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794
[17] 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794 0.3678794
[25] 0.3678794
```

The derangement probability converges to a specific number as n increases (it can be shown that this number is exactly e^{-1}). More informally, for large enough n, the probablity of a derangement depends very little on n.

Problem 1.8 The formal axioms of probability can be stated as follows. Let the set S be the sample space. Let $\mathcal{F}$ be some collection of subsets of S. Not all subsets of S need be in $\mathcal{F}$, but S and $\emptyset$ must be. We assign a probability $P(E)$ to each $E \in \mathcal{F}$. The three axioms are as follows:

Axiom 1. For any $E \in \mathcal{F}$, $P(E) \geq 0$.

Axiom 2. $P(S) = 1$.

Axiom 3. Suppose $E_1, E_2, \ldots, E_i, \ldots$ is a countable collection of mutually exclusive sets from $\mathcal{F}$. Suppose that the union $\cup_{i=1}^{\infty} E_i$ is also in $\mathcal{F}$. Then

$$P\left\{\cup_{i=1}^{\infty} E_i\right\} = \sum_{i=1}^{\infty} P(E_i).$$

Axiom 3 is referred to as *countable additivity*. For most probability models, we assume that if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ as well. Furthermore if the collection of sets E_i are in $\mathcal{F}$, then so is any union or intersection of any combination of these sets. We will make these assumptions below.

(a) Suppose in the following statements A,B are in $\mathcal{F}$. Prove that Axioms 1-3 imply each of the following statements.

(i) $P(\emptyset) = 0$.

(ii) $A \cap B = \emptyset$ implies $P(A \cup B) = P(A) + P(B)$.

(iii) $P(A^c) = 1 - P(A)$.

(iv) $A \subset B$ implies $P(A) \leq P(B)$.

HINT: The set $\emptyset$ is disjoint to all other sets, including $\emptyset$ itself. This means $S = S \cup \{\cup_{i=1}^{\infty} \emptyset\}$. Also note that countable additivity and finite additivity (*ie* statement (ii) above) are distinct statements.

(b) Let $E_1, E_2, \ldots$ be a countable collection of sets in $\mathcal{F}$. It is sometimes useful to construct an associated collection of sets, denoted

$$\begin{aligned} \bar{E}_1 &= E_1, \\ \bar{E}_i &= E_i \cap E_{i-1}^c \cap \ldots \cap E_1^c, \quad i \geq 2. \end{aligned}$$

Verify the following properties of $\bar{E}_i$:

(i) $\bar{E}_i \subset E_i$ for all $i \geq 1$.

(ii) The sets $\bar{E}_1, \bar{E}_2, \ldots$ are mutually exclusive.

(iii) $\cup_{i=1}^{\infty} E_i = \cup_{i=1}^{\infty} \bar{E}_i$.

(c) Prove **Boole's Inequality**: Let $E_1, E_2, \ldots$ be a countable collection of sets. Then $P\left\{\cup_{i=1}^{\infty} E_i\right\} \leq \sum_{i=1}^{\infty} P(E_i)$.

(d) Suppose world records for a given sport are compiled annually. For convenience, label the years $i = 1, 2, \ldots$ with $i = 1$ being the first year records are kept. Define the events

$$E_i = \{ \text{ World record broken in year } i \ \}, \ \ i \geq 1.$$

Then, let Q_i be the probability that from year i onwards, the current world record is never broken. Prove that if $\sum_{i=1}^{\infty} P(E_i) < \infty$, then $\lim_{i\to\infty} Q_i = 1$. Verify that, in particular, if $P(E_i) \leq c/i^k$ for some finite constants $c > 0$ and $k > 1$ then $\lim_{i\to\infty} Q_i = 1$.

SOLUTION:

(a) (i) We have $S = S \cup \{\cup_{i=1}^{\infty} \emptyset\}$. By Axiom 3 we must have

$$P(S) = P(S) + \sum_{i=1}^{\infty} P(\emptyset).$$

This implies $P(\emptyset) = 0$.

(ii) We can write $A \cup B = A \cup B \cup \{\cup_{i=1}^{\infty} \emptyset\}$. Then by Axiom 3 and Part (a) we have

$$\begin{aligned} P(A \cup B) &= P\left(A \cup B \cup \{\cup_{i=1}^{\infty} \emptyset\}\right) \\ &= P(A) + P(B) + \sum_{i=1}^{\infty} P(\emptyset) \\ &= P(A) + P(B). \end{aligned}$$

(iii) We have $A \cup A^c = S$. In addition A and A^c are always disjoint. By Part (b) (finite additivity),

$$1 = P(S) = P(A) + P(A^c),$$

therefore $P(A^c) = 1 - P(A)$.

(iv) We have $B = A \cup (BA^c)$. By Part (b) (finite additivity), we have

$$P(B) = P(A) + P(BA^c).$$

By Axiom 1, $P(BA^c) \geq 0$, therefore $P(A) \leq P(B)$.

(b) (i) First note that for any $i \geq 1$ we may write $\bar{E}_i = E_i \cap A$ for some set A. It follows that $\bar{E}_i \subset E_i$.

(ii) Let $j > i$. By construction, we can see that $\bar{E}_j \subset E_i^c$ and $\bar{E}_i \subset E_i$. We must therefore have $\bar{E}_i \bar{E}_j \subset E_i E_i^c = \emptyset$.

(iii) First, note that if $A \subset B$ and $A \supset B$, then $A = B$. Clearly, $\cup_{i=1}^{\infty} \bar{E}_i \subset \cup_{i=1}^{\infty} E_i$ by Part (i). Conversely, suppose we have element $\omega \in \cup_{i=1}^{\infty} E_i$. There must be at least one finite index n' such that $\omega \in E_{n'}$. Then suppose n' is the smallest such index. Then $\omega \notin E_i$ for any $i < n'$, equivalently $\omega \in \bar{E}_{n'} \subset \cup_{i=1}^{\infty} \bar{E}_i$. This implies $\cup_{i=1}^{\infty} \bar{E}_i \supset \cup_{i=1}^{\infty} E_i$, giving $\cup_{i=1}^{\infty} \bar{E}_i = \cup_{i=1}^{\infty} E_i$.

[Alternative Method] This result is often proven using induction. For convenience set $\hat{E}_n = \cup_{i=1}^{n} E_i$. Then by De Morgan's law $\bar{E}_{n+1} = E_{n+1} \cap \hat{E}_n^c$. We proceed by induction. Suppose for some specific index n we provisionaly assume that

$$\cup_{i=1}^{n} \bar{E}_i = \hat{E}_n.$$

Then (assuming the induction hypothesis holds),

$$\begin{aligned} \cup_{i=1}^{n+1} \bar{E}_i &= \bar{E}_{n+1} \cup \left\{ \cup_{i=1}^{n} \bar{E}_i \right\} \\ &= \left\{ E_{n+1} \cap \hat{E}_n^c \right\} \cup \hat{E}_n \\ &= E_{n+1} \cup \hat{E}_n \\ &= \hat{E}_{n+1}. \end{aligned}$$

Then for $n = 1$, we have $\cup_{i=1}^{n} \bar{E}_i = \hat{E}_n$. By induction, it also follows for all finite n. Then

$$\cup_{n=1}^{\infty} \bar{E}_n = \cup_{n=1}^{\infty} \cup_{i=1}^{n} \bar{E}_i = \cup_{n=1}^{\infty} \hat{E}_n = \cup_{n=1}^{\infty} E_n.$$

(c) Use the sets $\bar{E}_i$ of Part (b). Then note that

$$\begin{aligned} P\left\{ \cup_{i=1}^{\infty} E_i \right\} &= P\left\{ \cup_{i=1}^{\infty} \bar{E}_i \right\} \\ &= \sum_{i=1}^{\infty} P(\bar{E}_i) \\ &\leq \sum_{i=1}^{\infty} P(E_i). \end{aligned}$$

(d) Let A_n be the event that the world record is broken sometime during or after year n.

$$A_n = \cup_{i=n}^{\infty} E_i,$$

which means, using Boole's Inequality

$$P(A_n) \leq \sum_{i=n}^{\infty} P(E_i).$$

However, if $\sum_{i=1}^{\infty} P(E_i) < \infty$, then we must have

$$\lim_{n \to \infty} P(A_n) \leq \lim_{n \to \infty} \sum_{i=n}^{\infty} P(E_i) = 0.$$

The conclusion follows by noting $Q_n = 1 - P(A_n)$. In particular, if $P(E_i) \leq c/i^k$, then

$$\sum_{i=1}^{\infty} P(E_i) \leq \sum_{i=1}^{\infty} c/i^k \leq c + \int_{x=1}^{\infty} c/x^k dx = c + c/(k-1) < \infty,$$

provided $k > 1$.

1.2 Combinatorics and Equiprobable Distributions

Problem 1.9 An urn contains 20 balls. Exactly two are labeled by number i, for $i = 1, \ldots, 10$. A random sample (without replacement) of 4 balls is taken. What is the probability that no number is represented twice?

SOLUTION: It will be useful to temporarily label each numbered pair red and green, so that we have 20 distinct balls. We have

$$D = \binom{20}{4} = \frac{20 \times 19 \times 18 \times 17}{4!} = 4845$$

possible selections. Use the rule of product:

Task 1: Select 4 unique labels from 10, $n_1 = \binom{10}{4} = \frac{10 \times 9 \times 8 \times 7}{4!} = 210$.
Task 2: Select a color for each label, $n_2 = 2^4 = 16$.

There are

$$N = n_1 \times n_2 = 210 \times 16 = 3360$$

such selections, so

$$P(\text{ No number represetned twice }) = \frac{3360}{4845} \approx 0.693.$$

Problem 1.10 The letters ALABAMA are permuted at random. What is the probability that no two 'A's are next to each other?

SOLUTION: The total number of permutations is

$$D = \binom{7}{1,1,1,4} = \frac{7!}{1 \times 1 \times 1 \times 4!} = 7 \times 6 \times 5 = 210.$$

Let N be the number of permutations for which no two 'A's are next to each other. To calculate N use the *rule of product*. First note that for such a permutation the 'A's must occupy position 1,3,5,7, and the letters 'L','B','M' must occupy positions 2,4,6.

Task 1: There is exactly one way to arrange the 'A's, so $n_1 = 1$.
Task 2: Then 'L','B','M' can be freely permuted among the remaining positions, so $n_2 = 3! = 6$.

There are

$$N = n_1 \times n_2 = 1 \times 6 = 6,$$

so

$$P(\text{No consecutive A's}) = \frac{N}{D} = \frac{6}{210} = \frac{1}{35}.$$

Problem 1.11 Ten birds are arranged in a row on a powerline. Four are red, six are blue. Suppose the ordering is random, that is, all permutations of the ordering of the ten birds are equally likely.

(a) What is the probability that the four red birds are all adjacent?
(b) What is the probability that no two red birds are adjacent?

HINT: In a problem like this, it can be useful to assign labels to the objects to make them completely distinct. So, the blue birds will be labeled $B_1, \ldots, B_6$, the red birds will be labeled $R_1 \ldots, R_4$. The events in question do not depend on this labeling, so this won't affect the answer.

Next, suppose the blue birds are ordered $B_1, B_2, B_3, B_4, B_5, B_6$. We will make 'slots' on each side of the blue birds:

$$\_\_\_ B_1 \_\_\_ B_2 \_\_\_ B_3 \_\_\_ B_4 \_\_\_ B_5 \_\_\_ B_6 \_\_\_$$

Thus, to construct an ordering, we first order the blue birds, then place each red bird in one of the 7 slots.

SOLUTION: We use the $P(E) = N/D$ rule. The number of arrangements of the 9 labelled birds is $D = 10! = 3,628,800$.

(a) If all 4 red birds are adjacent, they occupy a single slot. So by the rule of product, we have the $k = 3$ tasks:

Task 1: Arrange 6 blue birds, $n_1 = 6! = 720$.
Task 2: Select 1 slot for the red birds, $n_2 = 7$.
Task 3: Arrange the 4 red birds within the slot, $n_3 = 4! = 24$.

There are

$$N = n_1 \times n_2 \times n_3 = 720 \times 7 \times 24 = 120,960$$

such selections, so

$$P(\text{red birds are adjacent}) = \frac{120,960}{3,628,800} = \frac{1}{30} \approx 0.0333.$$

(b) If no red birds are adjacent, they each occupy distinct slots. So by the rule of product, we have the $k = 3$ tasks:

Task 1: Arrange 6 blue birds, $n_1 = 6! = 720$.
Task 2: Select 4 of 7 slots for the red birds, $n_2 = \binom{7}{4} = 35$.
Task 3: Arrange the 4 red birds within the slot, $n_3 = 4! = 24$.

There are

$$N = n_1 \times n_2 \times n_3 = 720 \times 35 \times 24 = 604,800$$

such selections, so

$$P(\text{red birds are adjacent}) = \frac{604,800}{3,628,800} = \frac{1}{6} \approx 0.167.$$

Note that only Task 2 changes.

Problem 1.12 A lunch counter has 16 stools in a single row. Currently, 5 stools are occupied by customers. You notice that no two customers sit next to each other. Assume that customers select stools at random (that is, all selections of 5 stools are equally likely to be occupied).

(a) What is the probability that no two adjacent stools are occupied?
(b) What is the probability that all 5 customers sit in adjacent stools (that is, there are no empty stools between any two customers)?

HINT: A selection of 5 from 16 stools can be represented as an arrangement of 5 O's and 11 E's, such as EOEEOEEOEEEOEOEEE or EEEOOEOEEOEEOEEE. In the second example, but not the first, there are two customers sitting next to each other. Then, an arrangement can be constructed in the following way. Make 'slots' on each side of the 11 E's:

_E_E_E_E_E_E_E_E_E_E_E_

There are 12 slots (they should also be placed at the ends of the arrangement). Then, the O's can be placed in the slots according to some relevant rule.

SOLUTION: The number of arrangements of 5 O's and 11 E's is $D = \binom{16}{5}$.

(a) In order to construct an arrangement with no consecutive O's, we may assign each O to a distinct slot. There are $N = \binom{12}{5}$ ways to do this. Therefore

$$P\{\text{ No two adjacent stools are occupied }\} = \frac{N}{D} = \frac{\binom{12}{5}}{\binom{16}{5}} = \frac{12 \times 11 \times 10 \times 9 \times 8}{16 \times 15 \times 14 \times 13 \times 12} = 33/182 \approx 0.1813.$$

(b) In order to construct an arrangement in which all O's are adjacent, we may assign all O's to a single slot. There are $N = 12$ ways to do this. Therefore

$$P\{\text{ All 5 occupied stools are adjacent }\} = \frac{N}{D} = \frac{12}{\binom{16}{5}} = \frac{12 \times 5 \times 4 \times 3 \times 2 \times 1}{16 \times 15 \times 14 \times 13 \times 12} = 1/364 \approx 0.002747.$$

Problem 1.13 The letters in MISSISSIPPI are randomly permuted.

(a) What is the probability that there are no consecutive S's (for example ISMPISPSIIS)?
(b) What is the probability that the S's are consecutive (for example, IPSSSSIIMPI)?

SOLUTION: The number of permutations of MISSISSIPPI is

$$D = \binom{11}{1,2,4,4} = \frac{11!}{2! \times 4! \times 4!} = 34650.$$

(a) Use the *rule of product*.

Task 1: Permute the letters other than S, $n_1 = \binom{7}{1,2,4} = 105$.
Task 2: Once the remaining 7 letters have been permuted, we need to select for each S, uniquely, a position before, after or in between the letters. There are 8 such positions, so $n_2 = \binom{8}{4} = 70$.

There are

$$N = n_1 \times n_2 = 105 \times 70$$

such permutations, so

$$P(\text{No consecutive S's}) = \frac{N}{D} = \frac{105 \times 70}{34650} \approx 0.2121.$$

(b) Permuting the letters of MISSISSIPPI such that the S's are consecutive is equivalent to replacing the 4 S's with 1 S, then counting the permutations. This gives

$$N = \binom{8}{1,1,2,4} = \frac{8!}{2! \times 4!} = 840,$$

so

$$P(\text{All S's consecutive}) = \frac{N}{D} = \frac{840}{34650} \approx 0.024.$$

Problem 1.14 A bin contains 5 white and 5 black balls. A random selection of 2 balls is made. Let X be the number of white balls among the 2 selected. Determine $P(X = k)$ for $k = 0, 1, 2$.

SOLUTION: We may temporarily label the balls within each color $1, \ldots, 5$, so that they are all distinct. Then the total number of selections is

$$D = \binom{10}{2} = 45.$$

To enumerate the selections for which $X = k$ use the *rule of product*.

Task 1: Selection combination of k from 5 white balls, $n_1 = \binom{5}{k}$.
Task 2: Selection combination of $2 - k$ from 5 black balls, $n_2 = \binom{5}{2-k}$.

There are

$$N = n_1 \times n_2 = \binom{5}{k} \times \binom{5}{2-k}$$

such combinations. We then have the general expression:

$$P(X = k) = \frac{N}{D} = \frac{\binom{5}{k}\binom{5}{2-k}}{\binom{10}{2}}.$$

This gives

$$\begin{aligned} P(X = 0) &= \frac{\binom{5}{0}\binom{5}{2}}{\binom{10}{2}} = \frac{1 \times 10}{45} = \frac{2}{9} \\ P(X = 1) &= \frac{\binom{5}{1}\binom{5}{1}}{\binom{10}{2}} = \frac{5 \times 5}{45} = \frac{5}{9} \\ P(X = 2) &= \frac{\binom{5}{2}\binom{5}{0}}{\binom{10}{2}} = \frac{10 \times 1}{45} = \frac{2}{9}. \end{aligned}$$

Problem 1.15 A container contains 2 balls each of n colors (a total of $2n$ balls). The two balls of the same color are considered identical. Derive an expression for

$$\alpha_n = P(\text{All colors are adjacent in a random permutation of all balls}).$$

SOLUTION: We can use the multinomial coefficient. There are n types of balls, with $n_i = 2$ of each type, $i = 1, \ldots, n$. The number of permutations is therefore

$$D = \binom{2n}{2, \ldots, 2} = \frac{(2n)!}{\prod_{i=1}^{n} 2!} = \frac{(2n)!}{2^n}.$$

The number of permutations for which all colors are adjacent is equal to the number of permutations of the n colors:

$$N = n!$$

So,

$$\alpha_n = \frac{N}{D} = \frac{2^n n!}{(2n)!}.$$

ALTERNATIVE SOLUTION We can use the *rule of product*. Temporarily label the balls in each color pair 1 and 2. Then, to construct a permutation with adjacent colors use the following *tasks*:

Task 1: Select permutation of colors, $n_1 = n!$.
Task 2: Select ordering of temporary labels within each color pair, $n_2 = 2^n$.

There are

$$N = n_1 \times n_2 = n! \times 2^n$$

such permutations. There are a total of

$$D = (2n)!$$

(temporarily labelled) permutations, so

$$\alpha_n = \frac{N}{D} = \frac{2^n n!}{(2n)!}.$$

1.3 Games of Chance

Problem 1.16 A dice game is played in the following way. A player continues to toss a dice as long as the current outcome is strictly higher than the previous outcome. The score is the number of such outcomes. For example, for the sequence 1,3,5,2 the player stops at the fourth toss, and scores $X = 3$. What is the probability that the player scores at least $X = 3$?

SOLUTION: Noting that the player will always toss a dice at least twice, we define events

$$\begin{aligned} E &= \{ \text{ Player score at least } X = 3 \ \}, \\ A_{i,j} &= \{ \text{ First two tosses are } i, j \ \}. \end{aligned}$$

Consider event $A_{i,j}$. If $i \geq j$ then E cannot occur. If $i < j$, then E occurs with probability $(6 - j)/6$. Each event $A_{i,j}$ has probability $P(A_{i,j}) = 1/36$. This is expressed as conditional probabilities:

$$P(E \mid A_{i,j}) = \begin{cases} (6 - j)/6 & ; \quad i < j \\ 0 & ; \quad i \geq j \end{cases}.$$

By conditioning on the events $A_{i,j}$ we have (including only those events for which $P(E \mid A_{i,j}) > 0$)

$$\begin{aligned} P(E) &= P(E \mid A_{1,5})P(A_{1,5}) + P(E \mid A_{2,5})P(A_{2,5}) + P(E \mid A_{3,5})P(A_{3,5}) + P(E \mid A_{4,5})P(A_{4,5}) \\ &\quad + P(E \mid A_{1,4})P(A_{1,4}) + P(E \mid A_{2,4})P(A_{2,4}) + P(E \mid A_{3,4})P(A_{3,4}) \\ &\quad + P(E \mid A_{1,3})P(A_{1,3}) + P(E \mid A_{2,3})P(A_{2,3}) \\ &\quad + P(E \mid A_{1,2})P(A_{1,2}) \\ &= \left(4 \times \frac{1}{6} \times \frac{1}{36}\right) + \left(3 \times \frac{2}{6} \times \frac{1}{36}\right) + \left(2 \times \frac{3}{6} \times \frac{1}{36}\right) + \left(1 \times \frac{4}{6} \times \frac{1}{36}\right) \\ &= \frac{4+6+6+4}{6 \times 36} = \frac{5}{54}. \end{aligned}$$

Problem 1.17 Make use of the inclusion-exclusion principle (Problem 1.4) to answer the following question. A dice game has the following rules. Three dice are tossed, and the three values from 1 to 6 are noted. Points are awarded for the following combinations:

(1) The sum of the values is at least 16.
(2) The three values are the same.
(3) The three values are each even numbers.

A dice toss may match more than one of these combinations, and would receive additional points accordingly. If the dice toss does not match any of these combinations, no points are awarded. Use the inclusion-exclusion principle to determine the probability that no points are awarded for a given toss.

SOLUTION: To solve the problem, it will be convenient to regard the dice as distinct, say, colored red, blue and green. Then there are $D = 6^3 = 216$ outcomes, by the rule of product. Define the events

$$\begin{aligned} A_1 &= \{ \text{ The sum of the values is at least 16 } \}, \\ A_2 &= \{ \text{ The three values are the same } \}, \\ A_3 &= \{ \text{ The three values are each even numbers } \}. \end{aligned}$$

We then need to find the probabilities of A_1, A_2, A_3, and then the probabilities of all intersections A_1A_2, A_1A_3, A_2A_3 and $A_1A_2A_3$. We can approach this by finding the cardinality N of each required event E, then applying the formula $P(E) = N/D$. We take the following steps:

(1) The only outcomes in A_1 are those with unordered values (4,6,6), (5,6,6), (6,5,5) and (6,6,6). There are 3 ways to 'color' the values (4,6,6), (5,6,6) and (6,5,5); and 1 way to color the values (6,6,6). Applying the rule of product to each set of unordered values, we have

$$|A_1| = 3 + 3 + 3 + 1 = 10.$$

(2) There are clearly only 6 outcomes in A_2. So,

$$|A_2| = 6.$$

(3) Each color dice value can be 2, 4 or 6, in any combination. Then there are $3^3 = 27$ outcomes, by the rule of product. So,

$$|A_3| = 27.$$

(4) Of all outcomes in A_1 only 1 (the outcome with all 6's) is also in A_2. So,

$$|A_1A_2| = 1.$$

(5) Of all outcomes in A_1 there are 4 which are also in A_3, in particular $\{466, 646, 664, 666\}$, taking the dice in the order, say, red/blue/green. So,

$$|A_1A_3| = 4.$$

(6) There are 3 outcomes in both A_2 and A_3, in particular $\{222, 444, 666\}$. So,

$$|A_2A_3| = 3.$$

(7) There is 1 outcome in all three events, A_1, A_2 and A_3, in particular, $\{666\}$. So,

$$|A_1A_2A_3| = 1.$$

So, by the inclusion-exclusion identity we have

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3) \\ &= \frac{1}{216}\left[|A_1| + |A_2| + |A_3| - |A_1A_2| - |A_1A_3| - |A_2A_3| + |A_1A_2A_3|\right] \\ &= \frac{1}{216}\left[10 + 6 + 27 - 1 - 4 - 3 + 1\right] \\ &= \frac{36}{216} \\ &= \frac{1}{6}. \end{aligned}$$

Then

$$P(\text{ No points awarded }) = 1 - P(A_1 \cup A_2 \cup A_3) = \frac{5}{6}.$$

Problem 1.18 *Yahtzee* is a game of chance played with five standard 6-sided dice. A player may roll the dice up to three times, and is awarded points based on various combinations. For example, a *yahtzee* (all dice the same number) earns 50 points. A *full house* occurs when two dice are of one number, and three dice are of another number. An example would be the five dice showing 2 fours and 3 sixes. This combination earns 25 points.

What is the probability of throwing a *full house* on a single toss? **HINT:** Carefully list the *tasks* used in an application of the *rule of product*. It may also be advantageous to assume that the dice are ordered, or otherwise distinctly colored or labelled.

SOLUTION: We can have

$$D = 6^5 = 7,776$$

possible tosses. Then the tasks are:

Task 1: Select 1 from 6 sides for the three of a kind, $n_1 = 6$.

Task 2: Select 1 from 5 sides for the two of a kind, $n_2 = 5$.
Task 3: Select 3 from 5 dice for the three of a kind, $n_3 = \binom{5}{3} = 10$.

There are

$$N = n_1 \times n_2 \times n_3 = 6 \times 5 \times 10 = 300$$

such selections, so

$$P(\text{ Full House }) = \frac{N}{D} = \frac{300}{7,776} \approx 0.0386.$$

Problem 1.19 A standard 52 card playing deck assigns a unique combination of 13 ranks

$$(2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A)$$

and 4 suits

$$(Clubs, Diamonds, Hearts, Spades)$$

to each card ($13 \times 4 = 52$). Suppose 5 cards are selected at random. Derive the probability that each of the following events occurs.

(a) All cards are face cards (J,Q,K) of a single color (all black [spades or clubs] or all red [hearts or diamonds]).
(b) All suits are represented at least once.
(c) All ranks are distinct, and of a single color.

Carefully list the *tasks* used in the application of the *rule of product*.

SOLUTION: Recall that we can have

$$D = \binom{52}{5} = 2,598,960.$$

possible hands. Use the rule of product for each problem:

(a) **All cards are face cards of a single color.** There are 6 red, or black, face cards.

Task 1: Choose color, $n_1 = 2$.
Task 2: Select combination of 5 from 6 cards, $n_2 = \binom{6}{5} = 6$.

There are

$$N = n_1 \times n_2 = 2 \times 6 = 12$$

such selections, so

$$P(\text{ All cards are face cards of a single color }) = \frac{12}{2,598,960}.$$

(b) **All suits are represented at least once.** One suit must be represented twice, the rest each once.

Task 1: Choose the suit represented twice, $n_1 = 4$.

Task 2: Select combination of 2 from 13 cards from the suit represented twice, $n_2 = \binom{13}{2} = 78$.
Task 3: Select 1 rank from each of the remaining suits, $n_3 = 13^3 = 2,197$.

There are

$$N = n_1 \times n_2 \times n_3 = 4 \times 78 \times 2,197 = 685,464$$

such selections, so

$$P(\text{ All suits are represented at least once }) = \frac{685,464}{2,598,960}.$$

(c) **All ranks are distinct, and of a single color.**

Task 1: Choose color, $n_1 = 2$.
Task 2: Select 5 from 13 ranks, $n_2 = ch135 = 1,287$.
Task 3: For each rank, select 1 of 2 suits of the selected color, $n_3 = 2^5 = 32$.

There are

$$N = n_1 \times n_2 \times n_3 = 2 \times 1,287 \times 32 = 82,368$$

such selections, so

$$P(\text{ All ranks are distinct, and of a single color }) = \frac{82,368}{2,598,960}.$$

Problem 1.20 A standard 52 card playing deck assigns a unique combination of 13 ranks

$$(2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A)$$

and 4 suits

$$(Clubs, Diamonds, Hearts, Spades)$$

to each card ($13 \times 4 = 52$). Suppose 10 cards are selected at random. Derive the probability that each of the following events occurs.

(a) Only 1 suit is represented.
(b) Exactly 2 suits are represented in equal number.
(c) No rank is represented more than once.

Carefully list the *tasks* used in the application of the *rule of product*.

SOLUTION: SOLUTION Recall that we can have

$$D = \binom{52}{10} = 15,820,024,220$$

possible hands. Use the rule of product for each problem:

(a) **Only 1 suit is represented.**

Task 1: Choose suit, $n_1 = 4$.
Task 2: Select combination of 10 from 13 cards, $n_2 = \binom{13}{10} = 286$.

There are

$$N = n_1 \times n_2 = 4 \times 286 = 1,144$$

such selections, so

$$P(\text{ Only 1 suit is represented }) = \frac{1,144}{15,820,024,220} \approx 7.231 \times 10^{-8}.$$

(b) **Exactly 2 suits are represented in equal number.**

Task 1: Choose 2 suits, $n_1 = \binom{4}{2} = 6$.
Task 2: Select combination of 5 from 13 cards for first suit, $n_2 = \binom{13}{5} = 1,287$.
Task 3: Select combination of 5 from 13 cards for second suit, $n_3 = \binom{13}{5} = 1,287$.

There are

$$N = n_1 \times n_2 \times n_3 = 6 \times 1,287 \times 1,287 = 9,938,214$$

such selections, so

$$P(\text{ Exactly 2 suits are represented in equal number }) = \frac{9,938,214}{15,820,024,220} \approx 0.0006282.$$

(c) **No rank is represented more than once.**

Task 1: Choose 10 ranks, $n_1 = \binom{13}{10} = 286$.
Task 2: Select suit for each rank, $n_2 = 4^{10} = 1,048,576$.

There are

$$N = n_1 \times n_2 = 286 \times 1,048,576 = 299,892,736$$

such selections, so

$$P(\text{ No rank is represented more than once }) = \frac{299,892,736}{15,820,024,220} \approx 0.01896.$$

Problem 1.21 A game is played in the following way. First a 6-sided dice is tossed. Suppose the dice shows N. Then a coin is tossed N times. The player wins if the coin shows the same face for each of the N tosses. What is the probability that the player wins? Use the law of total probability.

SOLUTION: Let W be the event that the player wins. If $N = n$, then by independence

$$P(\text{All Heads}) = P(\text{All Tails}) = (1/2)^n,$$

so that

$$P(W \mid N = n) = (1/2)^n + (1/2)^n = (1/2)^{n-1}.$$

The sample space is partitioned by the 6 events $A_1 = \{N = 1\}, \ldots, A_6 = \{N = 6\}$. By the law of total probability

$$\begin{aligned} P(W) &= \sum_{i=1}^{6} P(W \mid N = i) P(N = i) \\ &= \sum_{i=1}^{6} (1/2)^{i-1} \times \frac{1}{6} \\ &= \frac{1}{6} \times (1 + 1/2 + 1/4 + 1/8 + 1/16 + 1/32) \\ &= \frac{1}{6} \times 2 \times \frac{63}{64} \\ &= \frac{21}{64}. \end{aligned}$$

Problem 1.22 A standard 52 card playing deck assigns a unique combination of 13 ranks

$$(2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A)$$

and 4 suits

$$(Clubs, Diamonds, Hearts, Spades)$$

to each card ($13 \times 4 = 52$). Suppose a hand of 5 cards is selected at random. Using the *rule of product* calculate the probability that the cards form the following hands:

(a) **One Pair.** Exactly two cards on one rank, the remaining cards of distinct rank.
(b) **Two Pairs.** Two distinct ranks represented by exactly two cards, the remaining card of distinct rank.
(c) **Three of a Kind.** Exactly three cards of one rank, the remaining cards of distinct rank.
(d) **Straight.** Five distinct ranks in sequence, with at least two suits represented. Note that $A, 2, 3, 4, 5$ is a sequence.
(e) **Flush.** All cards of the same suit, but not in rank sequence. Note that $A, 2, 3, 4, 5$ is a sequence.
(f) **Full House.** Two cards of one rank, three cards of a different rank.
(g) **Four of a Kind.** Exactly three cards of one rank, the remaining cards of distinct rank.
(h) **Straight Flush.** All cards of the same suit, and in rank sequence. Note that $A, 2, 3, 4, 5$ is a sequence. Excludes Royal Flush.
(i) **Royal Flush.** All cards of the same suit, in rank sequence $10, J, Q, K, A$.

Carefully list the *tasks* used in the application of the *rule of product*.

SOLUTION: Recall that we can have

$$D = \binom{52}{5} = 2,598,960$$

possible hands. Use the rule of product for each problem:

(a) **One Pair**

Task 1: Select rank for pair, $n_1 = 13$.
Task 2: Select combination of 2 from 4 cards for pair rank, $n_2 = \binom{4}{2} = 6$.
Task 3: Select 3 distinct ranks for remaining cards, $n_3 = \binom{12}{3} = 220$.
Task 4: Select 1 of 4 suits for each of the remaining cards, $n_4 = 4^3 = 64$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 6 \times 220 \times 64 = 1,098,240$$

such selections, so

$$P(\text{ One Pair }) = \frac{1,098,240}{2,598,960} \approx 0.4226.$$

(b) **Two Pairs**

Task 1: Select combination of 2 from 13 ranks for the pairs, $n_1 = \binom{13}{2} = 78$.
Task 2: Select 2 from 4 cards for first pair rank, $n_2 = \binom{4}{2} = 6$.
Task 3: Select 2 from 4 cards for second pair rank, $n_3 = \binom{4}{2} = 6$.
Task 4: Select 1 of 44 = 52 - 8 remaining cards, $n_4 = 44$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 78 \times 6 \times 6 \times 44 = 123,552$$

such selections, so

$$P(\text{ Two Pairs }) = \frac{123,552}{2,598,960} \approx 0.04754.$$

(c) **Three of a Kind**

Task 1: Select rank for three cards of common rank, $n_1 = 13$.
Task 2: Select combination of 3 from 4 cards for common rank, $n_2 = \binom{4}{3} = 4$.
Task 3: Select 2 distinct ranks for remaining cards, $n_3 = \binom{12}{2} = 66$.
Task 4: Select 1 of 4 suits for each of the remaining cards, $n_3 = 4^2 = 16$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 66 \times 16 = 54,912$$

such selections, so

$$P(\text{ Thee of a Kind }) = \frac{54,912}{2,598,960} \approx 0.021.$$

(d) **Straight**

Task 1: Select from 10 possible sequences (low card A to 10), $n_1 = 10$.
Task 2: Select a suit for each card, excluding selections of common suits, $n_2 = 4^5 - 4 = 1020$.

There are

$$N = n_1 \times n_2 = 10 \times 1020 = 10,200$$

such selections, so

$$P(\text{ Straight }) = \frac{10,200}{2,598,960} \approx 0.0039.$$

(e) **Flush**

Task 1: Select 1 from 4 suits, $n_4 = 4$.
Task 2: Select 5 from 13 ranks, excluding sequential ranks, $n_2 = \binom{13}{5} - 10 = 1277$.

There are

$$N = n_1 \times n_2 = 4 \times 1277 = 5,108.$$

such selections, so

$$P(\text{ Flush }) = \frac{5,108}{2,598,960} \approx 0.0020.$$

(f) **Full House**

Task 1: Select 1 from 13 ranks for the *three of a kind*, $n_1 = 13$.
Task 2: Select 3 from 4 cards for the *three of a kind*, $n_2 = \binom{4}{3} = 4$.
Task 3: Select 1 from 12 remaining ranks for the *two of a a kind*, $n_3 = 12$.
Task 4: Select 2 from 4 cards for the *two of a kind*, $n_4 = \binom{4}{2} = 6$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 12 \times 6 = 3,744$$

such selections, so

$$P(\text{ Full House }) = \frac{3,744}{2,598,960} \approx 0.001441.$$

(g) **Four of a Kind**

Task 1: Select rank for four cards of common rank, $n_1 = 13$.
Task 2: Select remaining card, $n_2 = 48$.

There are

$$N = n_1 \times n_2 = 13 \times 48 = 624$$

such selections, so

$$P(\text{ Four of a Kind }) = \frac{624}{2,598,960} \approx 2.401 \times 10^{-4}.$$

(h) **Straight Flush**

Task 1: Select from 9 possible sequences (low card A to 9), $n_1 = 9$.
Task 2: Select a common suit, $n_2 = 4$.

There are

$$N = n_1 \times n_2 = 9 \times 4 = 36$$

such selections, so

$$P(\text{ Straight Flush }) = \frac{36}{2,598,960} \approx 1.385 \times 10^{-5}.$$

(h) **Royal Flush**

Task 1: Select a common suit, $n_1 = 4$.

There are

$$N = n_1 = 4$$

such selections, so

$$P(\text{ Royal Flush }) = \frac{4}{2,598,960} \approx 1.539 \times 10^{-6}.$$

1.4 Games and Strategies

Problem 1.23 This question is adapted from *Introduction to Probability Models* (10th Edition), S.M. Ross. Three prisoners, labeled A, B and C, are informed by a guard that one of them has been chosen at random to be executed the following day. Prisoner A asks the guard, privately, to name one of the other prisoners who will be released. We then have the competing claims:

(a) The guard argues that by eliminating one prisoner from the execution pool the probability that A is executed changes from 1/3 to 1/2.
(b) Prisoner A argues that since it is already known that at least one of prisoners B or C will be released, the probability that A is executed remains 1/3.

Assume that if the guard names a prisoner to be released, and both B and C are to be released, the guard will name either one with equal probability. Otherwise, the guard names the only prisoner other than A being released. Define the following events.

$$\begin{aligned} E_A &= \{\text{Prisoner } A \text{ chosen for execution}\} \\ E_B &= \{\text{Prisoner } B \text{ chosen for execution}\} \\ E_C &= \{\text{Prisoner } C \text{ chosen for execution}\} \\ F_B &= \{\text{Guard informs prisoner } A \text{ that prisoner } B \text{ is being released}\} \\ F_C &= \{\text{Guard informs prisoner } A \text{ that prisoner } C \text{ is being released}\}. \end{aligned}$$

So, the event that B is to be released is equivalent to E_B^c, and the event that A is informed that B is to be released is F_B. Calculate the following probabilities:

1. $P(E_B^c)$,
2. $P(F_B)$,
3. $P(E_A \mid E_B^c)$,
4. $P(E_A \mid F_B)$.

Who is correct, the guard or prisoner A?

SOLUTION:

1. We have $P(E_B) = 1/3$, so $P(E_B^c) = 1 - P(E_B) = 1 - 1/3 = 2/3$.
2. We must have

$$\begin{aligned} P(F_B \mid E_A) &= 1/2 \\ P(F_B \mid E_B) &= 0 \\ P(F_B \mid E_C) &= 1 \end{aligned}$$

so by the law of total probability:

$$\begin{aligned} P(F_B) &= P(F_B \mid E_A)P(E_A) + P(F_B \mid E_B)P(E_B) + P(F_B \mid E_C)P(E_C) \\ &= (1/2)(1/3) + 0(1/3) + 1(1/3) \\ &= 1/2. \end{aligned}$$

3. Noting that $E_A \subset E_B^c$, and so $E_A E_B^c = E_A$ we have

$$P(E_A \mid E_B^c) = \frac{P(E_A E_B^c)}{P(E_B^c)} = \frac{P(E_A)}{P(E_B^c)} = \frac{1/3}{2/3} = 1/2.$$

4.

$$P(E_A \mid F_B) = \frac{P(E_A F_B)}{P(F_B)} = \frac{P(F_B \mid E_A)P(E_A)}{P(F_B)} = \frac{(1/2)(1/3)}{1/2} = 1/3.$$

Prisoner A is correct.

Problem 1.24 The *Monty Hall problem* is a good example of the often counterintuitive nature of probability. It is based on the television game show *Let's Make a Deal* (starring Monty Hall). There are three doors. Behind one is a car, and behind the other two are goats. The contestant picks one door. Then, one the other doors is opened, revealing a goat (this can always be done, since there are two goats). The contestant is offered the choice of staying with the original choice, or switching to the one remaining door. The contestant wins whatever is behind the selected door. Assume the contestant makes the first choice at random, and has decided in advance whether or not to switch. Determine the probability of winning the car if the contest doesn't switch, and if the contestant does switch.

SOLUTION:

We can, without loss of generality, assume that the car is behind door 1 (the argument is identical wherever the car is). Define events

$$\begin{aligned} D_i &= \{ \text{ Contestant initially picks door } i \}, \quad i = 1, 2, 3. \\ W &= \{ \text{ Contestant wins car } \}. \end{aligned}$$

Then $P(D_i) = 1/3$.

If the **contestant doesn't switch**, we clearly have $P(W \mid D_1) = 1$ and $P(W \mid D_2) = P(W \mid D_3) = 0$. By conditioning on the initial selection (the Law of Total Probability) we have

$$P(W) = P(W \mid D_1)P(D_1) + P(W \mid D_2)P(D_2) + P(W \mid D_3)P(D_3) = 1 \times 1/3 + 0 \times 1/3 + 0 \times 1/3 = 1/3.$$

So the probability of winning the car is 1/3 if the contestant doesn't switch.

If the **contestant does switch**, we have $P(W \mid D_1) = 0$, since the contestant necessarily switches to a goat. However, if the contestant initially picks door 2 (which has a goat) then it has to be door 3 which is opened. The contest would have to switch to door 1, thus winning the car. This means $P(W \mid D_2) = 1$, and also $P(W \mid D_3) = 1$ by an identical argement. By conditioning on the initial selection (the Law of Total Probability) we have

$$P(W) = P(W \mid D_1)P(D_1) + P(W \mid D_2)P(D_2) + P(W \mid D_3)P(D_3) = 0 \times 1/3 + 1 \times 1/3 + 1 \times 1/3 = 2/3.$$

So the probability of winning the car is 2/3 if the contestant does switch.

Problem 1.25 (The Shrewd Prisoner's Dilemma) In some little known kingdom convicted prisoners are offered the possibility of a pardon according to a game of chance. The prisoner is given n red balls and n green balls. He/she then places the balls into two bins in any manner he/she chooses. The king then (i) selects a bin at random; (ii) selects one ball at random from that bin. If that ball is green the prisoner is pardoned. Note that the prisoner can leave one bin empty, and if that bin is selected by the king, then no green ball can be chosen, so no pardon is granted.

How should the balls be placed in order to maximize the probability $P(G)$ of selecting green (and therefore winning a pardon)?

HINT: There are several ways to solve this. One way is to solve the following sub-problems, from which the optimal choice can be deduced:

P1 In a *simple allocation* all balls are in one bin. What is $P(G)$ for this case?

P2 In an *even allocation* each bin has the same number of balls, independent of their color. Show that $P(G)$ is the same for all even allocations, and derive this number.

P3 In an *uneven allocation* one bin (the *light bin*) has strictly fewer balls than the other (the *heavy bin*), but both have at least one. Show that for any uneven allocation the heavy bin has at least one green ball. Next, show that if the light bin has at least one red ball, and if a red ball from the light bin is exchanged with a green ball from the heavy bin, then $P(G)$ will strictly increase.

P4 Show that for an uneven allocation in which the light bin has no red balls and at least two green balls, if a green ball is moved from the light bin to the heavy bin, then $P(G)$ will strictly increase.

It also helps to assume $n > 1$. The case $n = 1$ can then be consider separately.

SOLUTION: SOLUTION 1: Let $P_k(G)$ be the probability of selecting a green ball from bin k. Then

$$P(G) = [P_1(G) + P_2(G)]/2.$$

Suppose we put r and g red and green balls into bin 1. It will be convenient to set $m = r + g$. If bin k is empty we have $P_k(G) = 0$. Otherwise,

$$\begin{aligned} P_1(G) &= \frac{g}{m}, \quad m \geq 1 \\ P_2(G) &= \frac{n-g}{2n-m}, \quad m \leq 2n-1 \\ P(G) &= \frac{1}{2}\left[\frac{g}{m} + \frac{n-g}{2n-m}\right], \quad 1 \leq m \leq 2n-1 \end{aligned} \tag{1.3}$$

We can, without loss of generality, assume that bin 1 is the light bin (or empty bin) for an uneven (or simple) allocation.

P1 We have $r = g = m = 0$, so $P(G) = 1/4$.

P2 In an *even allocation* we have $m = n$, but g may vary from 0 to n. This gives

$$P(B) = \frac{1}{2}\left[\frac{g}{n} + \frac{n-g}{n}\right] = \frac{1}{2}\left[\frac{n-g+g}{n}\right] = 1/2,$$

which does not depend on g.

P3 There are n red balls. Since the heavy bin contains more than n balls, at least one must be green. First, note that m is unchanged by the exchange. Let $P(G)$, $P'(G)$ be the probabilities of selecting green before and after the exchange. Then

$$\begin{aligned} P'(G) - P(G) &= \frac{1}{2}\left[\frac{g+1}{m} + \frac{n-g-1}{2n-m}\right] - \frac{1}{2}\left[\frac{g}{m} + \frac{n-g}{2n-m}\right] \\ &= \frac{1}{2}\left[\frac{1}{m} - \frac{1}{2n-m}\right] \\ &= > 0, \end{aligned}$$

where the inequality follows from the fact that $m < 2n - m$.

P4 First, note that $P_1(G) = 1$ before and after the move. Following the move, the number of green balls in the heavy bin increases and the number of red balls is unchanged, so $P_2(G)$, and therefore $P(G)$, must strictly increase.

The number of allocations is finite, so at least one allocation gives a maximum probability. We then proceed by elimination. From P1 and P2 we can eliminate the simple allocation, since any even allocation is strictly better. From P3 we can eliminate any uneven allocation with at least one red ball in the light bin. From P4 we can eliminate any uneven allocation in which the light bin contains no red balls and at least two green balls. This means that if an uneven allocation is optimal it must be $g = 1, r = 0, m = 1$ (or $g = n-1, r = n, m = 2n-1$). For this allocation we have

$$P(G) = \frac{1}{2}\left[\frac{1}{1} + \frac{n-1}{2n-1}\right] = \frac{3n-2}{4n-2}. \tag{1.4}$$

If we assume $n > 1$ then this number is greater than $1/2$, which eliminates the even allocation, exhausting all remaining possibilities. Finally, if $n = 1$, no uneven allocation is possible, so the even allocation is optimal, and equation (1.4) holds for this case also. Therefore, placing exactly one green ball in one of the bins yields the maximum value of $P(G)$, given in (1.4), for any $n \geq 1$. This value approaches $3/4$ from below as n increases (so larger n is more favorable to the prisoner).

SOLUTION 2: We have an optimization problem in two dimensions. Sometimes, such problems can be solved by holding one dimension fixed, solving the simpler one-dimensional problem under that constraint, then relaxing the constraint.

Let $q_m(g) = P(G)$ for m and g. First assume that $m \leq n$. In that case, this function is defined for $g = 0, 1, \ldots, m$ (see notaton for SOLUTION 1). Then let

$$q_m^* = \max_{g=0,1,\ldots,m} q_m(g).$$

We know (from SOLUTION 1) that $q_0(g) = 1/4$ and $q_n(g) = 1/2$ for all admissible g, so $q_0^* = 1/4$, $q_n^* = 1/2$.

Next, if $1 \leq m < n$ then it is easy to show that $q_m(g) = P(G) = ag + b$ for some constants a, b, for which $a > 0$ (other than that, the exact values of a, b don't matter). Therefore

$$q_m^* = q_m(m) = \begin{cases} \frac{1}{2}\left[1 + \frac{n-m}{2n-m}\right] & ; \quad m = 1, \ldots, n \\ 1/4 & ; \quad m = 0 \end{cases}.$$

This covers the cases $m = 0, 1, \ldots, n$. By symmetry (all we need to do is exchange the bin labels) we have $q_{2n-m}^* = q_m^*$, which exhausts all cases. The problem is then solved by maximizing q_m^*, achieved by setting

$m = 1, g = 1$ or $m = 2n - 1, g = n - 1$, again yielding maximum probability (1.4).

A final note. If the two solutions are compared, it can be seen that they are very similar.

Problem 1.26 Someone proposes playing a dice game, and kindly offers to provide the dice. You suspect that the dice may be *loaded*, that is, at least one outcome has a probably other than 1/6. Suppose E is an event involving a die with the following special property. Let $P_f(E)$ be the probability of the event for a *fair* die (each outcome has probability 1/6). Let $P_{uf}(E)$ be the probability of the event for some other die. If this special property holds, then if that die is loaded in any way, we have $P_{uf}(E) > P_f(E)$. Note that E can involve more than one toss *of the same* die. Can you think of an event with this property? If so, you can propose a bet that favors you if the dice is loaded, and is fair otherwise, without having to know how the dice is loaded.

SOLUTION: Toss a single die twice. Set $E = \{$ outcome is the same for both tosses $\}$. Then

$$P(E) = \sum_{i=1}^{6} p_i^2,$$

where p_i is the probability of tossing i. No matter what the loading, we must have $p_1 + \ldots + p_6 = 1$. It can then be shown that the sum is uniquely minimized by setting $p_i = 1/6$ for each i (using, for example, the Lagrange multiplier method).

1.5 Probability and Genetics

Problem 1.27 In genetics, a genotype consists of two genes, each of which is one of (possibly) several types of alleles. When two organisms mate, each passes one allele, selected at random, to the offspring, forming that offspring's genotype. Suppose a gene of a type of flower has two alleles, `r` and `R`. A plant possessing genotype `rr`, `rR` or `RR` has white, pink or red petals, respectively. A trait like this, in which both alleles determine the trait, is called *codominant*.

(a) Suppose a white and pink flower produce offspring A. Give the color distribution of A (that is, the probability that A is a given color, for each color).
(b) Suppose that A mates with a pink flower to produce offspring B. Give the color distribution of B.
(c) Suppose that C is the offspring of two pink flowers, and that A mates with C to produce offspring D. Give the color distribution of D.

SOLUTION: The following table gives the offspring genotype probability distribution for each combination of parents:

		Offspring Genotype		
Parent 1	Parent 2	`rr`	`rR`	`RR`
`rr`	`rr`	1	0	0
`rr`	`rR`	1/2	1/2	0
`rr`	`RR`	0	1	0
`rR`	`rR`	1/4	1/2	1/4
`rR`	`RR`	0	1/2	1/2
`RR`	`RR`	0	0	1

(a) The probabilities are given in row 2: $P(\text{white}) = 1/2$, $P(\text{pink}) = 1/2$.

(b) Let G_A, G_B be the genotypes of A and B. Using the law of total probability we have

$$\begin{aligned}
P(G_B = \texttt{rr}) &= P(G_B = \texttt{rr} \mid G_A = \texttt{rr})P(G_A = \texttt{rr}) + P(G_B = \texttt{rr} \mid G_A = \texttt{rR})P(G_A = \texttt{rR}) \\
&\quad + P(G_B = \texttt{rr} \mid G_A = \texttt{RR})P(G_A = \texttt{RR}) \\
&= \left[\frac{1}{2} \cdot \frac{1}{2}\right] + \left[\frac{1}{4} \cdot \frac{1}{2}\right] + 0 = \frac{3}{8} \\
P(G_B = \texttt{RR}) &= P(G_B = \texttt{RR} \mid G_A = \texttt{rr})P(G_A = \texttt{rr}) + P(G_B = \texttt{RR} \mid G_A = \texttt{rR})P(G_A = \texttt{rR}) \\
&\quad + P(G_B = \texttt{RR} \mid G_A = \texttt{RR})P(G_A = \texttt{RR}) \\
&= 0 + \left[\frac{1}{4} \cdot \frac{1}{2}\right] + 0 = \frac{1}{8}.
\end{aligned}$$

Then $P(G_B = \texttt{rR})$ is obtainable from the other probabilities, giving $P(\text{white}) = 3/8$, $P(\text{pink}) = 1/2$, $P(\text{red}) = 1/8$.

(c) It can be argued that the answers to (b) and (c) must be the same. In (b) the genetic contribution from A's mate comes from rR. In (c) it comes from one of C's parents, both of which are rR. So there is no difference, statistically. This is an acceptable answer.

The direct solution is as follows. Let G_C, G_D be the genotypes of C and D. Using the law of total probability we have (noting that $P(G_A = \texttt{RR}) = 0$):

$$\begin{aligned}
P(G_D = \texttt{rr}) &= P(G_D = \texttt{rr} \mid G_A = \texttt{rr}, G_C = \texttt{rr})P(G_A = \texttt{rr}, G_C = \texttt{rr}) \\
&\quad + P(G_D = \texttt{rr} \mid G_A = \texttt{rr}, G_C = \texttt{rR})P(G_A = \texttt{rr}, G_C = \texttt{rR}) \\
&\quad + P(G_D = \texttt{rr} \mid G_A = \texttt{rr}, G_C = \texttt{RR})P(G_A = \texttt{rr}, G_C = \texttt{RR}) \\
&\quad + P(G_D = \texttt{rr} \mid G_A = \texttt{rR}, G_C = \texttt{rr})P(G_A = \texttt{rR}, G_C = \texttt{rr}) \\
&\quad + P(G_D = \texttt{rr} \mid G_A = \texttt{rR}, G_C = \texttt{rR})P(G_A = \texttt{rR}, G_C = \texttt{rR}) \\
&\quad + P(G_D = \texttt{rr} \mid G_A = \texttt{rR}, G_C = \texttt{RR})P(G_A = \texttt{rR}, G_C = \texttt{RR}) \\
&= \left[1 \cdot \frac{1}{2} \cdot \frac{1}{4}\right] + \left[\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right] + 0 + \left[\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4}\right] + \left[\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}\right] + 0 \\
&= \frac{3}{8}
\end{aligned}$$

and

$$\begin{aligned}
P(G_D = \texttt{RR}) &= P(G_D = \texttt{RR} \mid G_A = \texttt{rr}, G_C = \texttt{rr})P(G_A = \texttt{rr}, G_C = \texttt{rr}) \\
&\quad + P(G_D = \texttt{RR} \mid G_A = \texttt{rr}, G_C = \texttt{rR})P(G_A = \texttt{rr}, G_C = \texttt{rR}) \\
&\quad + P(G_D = \texttt{RR} \mid G_A = \texttt{rr}, G_C = \texttt{RR})P(G_A = \texttt{rr}, G_C = \texttt{RR}) \\
&\quad + P(G_D = \texttt{RR} \mid G_A = \texttt{rR}, G_C = \texttt{rr})P(G_A = \texttt{rR}, G_C = \texttt{rr}) \\
&\quad + P(G_D = \texttt{RR} \mid G_A = \texttt{rR}, G_C = \texttt{rR})P(G_A = \texttt{rR}, G_C = \texttt{rR}) \\
&\quad + P(G_D = \texttt{RR} \mid G_A = \texttt{rR}, G_C = \texttt{RR})P(G_A = \texttt{rR}, G_C = \texttt{RR}) \\
&= 0 + 0 + 0 + 0 + \left[\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}\right] + \left[\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4}\right] \\
&= \frac{1}{8}
\end{aligned}$$

Then $P(G_D = \texttt{rR})$ is obtainable from the other probabilities, giving $P(\text{white}) = 3/8$, $P(\text{pink}) = 1/2$, $P(\text{red}) = 1/8$.

Problem 1.28 In genetics, a genotype consists of two genes, each of which is one of (possibly) several types of alleles. Suppose we consider only two alleles, `r` and `R`. Furthermore, suppose the allele `r` exists in a population with frequency q. Under Hardy-Weinberg equilibrium, genotypes are essentially random samples of 2 alleles, one sampled from each parent. Since the genes are usually not ordered (because we don't know which is maternal and which is paternal), the probability of each possible genotype is

$$\begin{aligned} P(\texttt{rr}) &= q^2, \\ P(\texttt{rR}) &= 2(1-q)q, \\ P(\texttt{RR}) &= (1-q)^2. \end{aligned}$$

We then note that genotypes of unrelated individuals are independent, but genotypes of related individuals are not independent. For example, when two organisms mate, each passes one allele, selected at random, to the offspring, forming that offspring's genotype (this predicts Mendels' Law of Inheritance). If G_o is an offspring genotype then

$$\begin{aligned} P(G_o = \texttt{rr}) &= q^2, \text{ but} \\ P(G_o = \texttt{rr} \mid \text{ both parents have genotype } \texttt{rR}) &= 1/4. \end{aligned}$$

Now, suppose `r` is a *recessive* allele, meaning that it only determines a trait when the genotype is `rr`. Such a trait is a *recessive trait*. Typically, a genetic disease is a recessive trait, and `r` is a *rare allele*, meaning q is very small.

Next, suppose we may determine without error whether or not an individual possesses a recessive trait (and therefore has genotype `rr`). Let G_m, G_f, G_1, G_2 be the genotypes of a mother, father and two offspring (ie. siblings). Define events

$$\begin{aligned} R_m &= \{ \text{ Mother has recessive trait } \}, \\ R_f &= \{ \text{ Father has recessive trait } \}, \\ R_1 &= \{ \text{ Offspring 1 has recessive trait } \}, \\ R_2 &= \{ \text{ Offspring 2 has recessive trait } \}. \end{aligned}$$

(a) Determine the conditional probability $P(R_1 \mid R_f^c \cap R_m^c)$, the probability that an offspring possesses the recessive trait given that neither parent does.
(b) Determine the conditional probability $P(R_1 \mid R_2 \cap R_f^c \cap R_m^c)$, the probability that an offspring possesses the recessive trait given that a sibling does, and that neither parent does.
(c) Does the probability defined in Part (b) depend on q? Give a brief explanation for this.

SOLUTION:

(a) We have

$$\begin{aligned} R_1 &= \{G_1 = \texttt{rr}\} \\ R_f^c &= \{G_f \neq \texttt{rr}\} = \{G_f = \texttt{rR} \text{ OR } G_f = \texttt{RR}\} \\ R_m^c &= \{G_m \neq \texttt{rr}\} = \{G_m = \texttt{rR} \text{ OR } G_m = \texttt{RR}\}. \end{aligned}$$

Thus the condition $R_f^c \cap R_m^c$ consists of 4 maternal/paternal genotype combinations. So we have

$$\begin{aligned} P(R_1 \cap R_f^c \cap R_m^c) &= P\{G_1 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{RR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_f = \mathtt{RR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_f = \mathtt{RR}, G_m = \mathtt{RR}\} \\ &= P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{RR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{RR}\} \\ &\quad + P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{RR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{RR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{RR}, G_m = \mathtt{RR}\} P\{G_f = \mathtt{RR}, G_m = \mathtt{RR}\} \end{aligned}$$

Note that of the 4 conditional probabilities above, only 1 is not zero, since if $G_1 = \mathtt{rr}$ then each parent must have at least one allele $\mathtt{r}$. This means

$$\begin{aligned} P(R_1 \cap R_f^c \cap R_m^c) &= P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &= \frac{1}{4} \times [2q(1-q)]^2 \end{aligned}$$

using Mendel's Law, and Hardy-Weinberg Equilibrium. Then, using Bayes Theorem,

$$\begin{aligned} P(R_1 \mid R_f^c \cap R_m^c) &= \frac{P(R_1 \cap R_f^c \cap R_m^c)}{P(R_f^c \cap R_m^c)} \\ &= \frac{P\{G_1 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\}}{P\{G_f \neq \mathtt{rr}, G_m \neq \mathtt{rr}\}} \\ &= \frac{\frac{1}{4} \times [2q(1-q)]^2}{[1-q^2]^2} \\ &= \left[\frac{q}{1+q}\right]^2 \end{aligned}$$

(b) The procedure for Part (b) is very similar. We have

$$\begin{aligned} P(R_1 \cap R_2 \cap R_f^c \cap R_m^c) &= P\{G_1 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{RR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr}, G_f = \mathtt{RR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr}, G_f = \mathtt{RR}, G_m = \mathtt{RR}\} \\ &= P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{RR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{RR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{RR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{RR}, G_m = \mathtt{rR}\} \\ &\quad + P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{RR}, G_m = \mathtt{RR}\} P\{G_f = \mathtt{RR}, G_m = \mathtt{RR}\}. \end{aligned}$$

As for Part (a), only one of the above terms is nonzero, so that

$$\begin{aligned} P(R_1 \cap R_2 \cap R_f^c \cap R_m^c) &= P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\} P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\} \\ &= \left(\frac{1}{4}\right)^2 \times [2q(1-q)]^2, \end{aligned}$$

since once the parental genotypes are fixed, the genotypes of siblings are independent (but are not independent unconditionally). Then, using Bayes Theorem

$$\begin{aligned}
P(R_1 \mid R_2 \cap R_f^c \cap R_m^c) &= \frac{P(R_1 \cap R_2 \cap R_f^c \cap R_m^c)}{P(R_2 \cap R_f^c \cap R_m^c)} \\
&= \frac{P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\}}{P\{G_2 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\}} \\
&= \frac{P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\}P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\}}{P\{G_2 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\}P\{G_f = \mathtt{rR}, G_m = \mathtt{rR}\}} \\
&= \frac{P\{G_1 = \mathtt{rr}, G_2 = \mathtt{rr} \mid G_f = \mathtt{rR}, G_m = \mathtt{rR}\}}{P\{G_2 = \mathtt{rr}, G_f = \mathtt{rR}, G_m = \mathtt{rR}\}} \\
&= \frac{\left(\frac{1}{4}\right)^2}{\frac{1}{4}} \\
&= \frac{1}{4}.
\end{aligned}$$

(c) If R_2 occurs the sibling must have genotype `rr`. This means each parent must have at least one allele `r`. But if R_m^c and R_f^c also occur then neither parent can have genotype `rr`. The only possibility left is that each parent have geonotype `rR`. Therefore, the conditional probability in Part (b) must be 1/4.

Problem 1.29 In genetics, a genotype consists of two genes, each of which is one of (possibly) several types of alleles. Suppose we consider only two alleles, `r` and `R`. Furthermore, suppose the allele `r` exists in a population with frequency q. Under Hardy-Weinberg equilibrium, genotypes are essentially random samples of 2 alleles, one sampled from each parent. Since the genes are usually not ordered (because we don't know which is maternal and which is paternal), the probability of each possible genotype in an individual is

$$\begin{aligned}
P(\mathtt{rr}) &= q^2, \\
P(\mathtt{rR}) &= 2(1-q)q, \\
P(\mathtt{RR}) &= (1-q)^2.
\end{aligned}$$

We then note that genotypes of unrelated individuals are independent, but genotypes of related individuals are not independent. For example, when two organisms mate, each passes one allele, selected at random, to the offspring, forming that offspring's genotype (this predicts Mendels' Law of Inheritance). If G_o is an offspring genotype then

$$\begin{aligned}
P(G_o = \mathtt{rr}) &= q^2, \text{ but} \\
P(G_o = \mathtt{rr} \mid \text{ both parents have genotype } \mathtt{rR}) &= 1/4.
\end{aligned}$$

Now, suppose `r` is a *recessive* allele, meaning that it only determines a trait when the genotype is `rr`. Such a trait is a *recessive trait*. Typically, a genetic disease is a recessive trait, and `r` is a *rare allele*, meaning q is very small.

Next, suppose we may determine without error whether or not an individual possesses a recessive trait (and therefore has genotype `rr`). Let G_m, G_f, G_1, G_2 be the genotypes of a mother, father and two

offspring (ie. siblings). Assume the parental genotypes are independent (ie. the parents are unrelated). Define events

$$\begin{aligned} R_m &= \{ \text{ Mother has recessive trait } \}, \\ R_f &= \{ \text{ Father has recessive trait } \}, \\ R_1 &= \{ \text{ Offspring 1 has recessive trait } \}, \\ R_2 &= \{ \text{ Offspring 2 has recessive trait } \}. \end{aligned}$$

(a) Determine the conditional probability $P(R_1 \mid R_m)$, the probability that an offspring possesses the recessive trait given that the mother does.
(b) Determine the conditional probability $P(R_1 \mid R_2)$, the probability that an offspring possesses the recessive trait given that a sibling does.
(c) Give a lower bound for the conditional probabilities of Parts (a) and (b). In other words, to what value does each conditional probability approach as q approaches zero?

SOLUTION: The following table gives the offspring genotype probability distribution for each combination of parents:

		Offspring Genotype		
Parent 1	Parent 2	rr	rR	RR
rr	rr	1	0	0
rr	rR	1/2	1/2	0
rr	RR	0	1	0
rR	rR	1/4	1/2	1/4
rR	RR	0	1/2	1/2
RR	RR	0	0	1

(a) To calculate $P(R_1 \mid R_m)$ we need probabilities $P(R_m)$ and $P(R_1 \cap R_m)$. We already have $P(R_m) = q^2$. To calculate $P(R_1 \cap R_m)$, write

$$\begin{aligned} P(R_1 \cap R_m) &= P\{G_1 = \texttt{rr}, G_m = \texttt{rr}\} \\ &= P\{G_1 = \texttt{rr}, G_m = \texttt{rr}, G_f = \texttt{rr}\} \\ &\quad + P\{G_1 = \texttt{rr}, G_m = \texttt{rr}, G_f = \texttt{rR}\} \\ &\quad + P\{G_1 = \texttt{rr}, G_m = \texttt{rr}, G_f = \texttt{RR}\} \\ &= P\{G_1 = \texttt{rr} \mid G_m = \texttt{rr}, G_f = \texttt{rr}\} P\{G_m = \texttt{rr}, G_f = \texttt{rr}\} \\ &\quad + P\{G_1 = \texttt{rr} \mid G_m = \texttt{rr}, G_f = \texttt{rR}\} P\{G_m = \texttt{rr}, G_f = \texttt{rR}\} \\ &\quad + P\{G_1 = \texttt{rr} \mid G_m = \texttt{rr}, G_f = \texttt{RR}\} P\{G_m = \texttt{rr}, G_f = \texttt{RR}\}. \end{aligned}$$

Note that of the 3 conditional probabilities above, the third is zero, since if $G_1 = \texttt{rr}$ then each parent must have at least one allele `r`. This means

$$P(R_1 \cap R_m) = 1 \times q^4 + (1/2) \times 2(1-q)q^3 = q^3.$$

Then

$$P(R_1 \mid R_m) = \frac{P(R_1 \cap R_m)}{P(R_m)} = \frac{q^3}{q^2} = q.$$

(b) To calculate $P(R_1 \mid R_2)$ we need probabilities $P(R_2) = q^2$ and $P(R_1 \cap R_2)$. To calculate $P(R_1 \cap R_2)$ condition on genotype pairs (G_m, G_f). Note that sibling genotypes are independent **conditional on both parent genotypes**. We then have (including only nonzero terms):

$$\begin{aligned} P(R_1 \cap R_2) &= P\{G_1 = \texttt{rr}, G_2 = \texttt{rr} \mid G_m = \texttt{rr}, G_f = \texttt{rr}\} P\{G_m = \texttt{rr}, G_f = \texttt{rr}\} \\ &\quad + P\{G_1 = \texttt{rr}, G_2 = \texttt{rr} \mid G_m = \texttt{rr}, G_f = \texttt{rR}\} P\{G_m = \texttt{rr}, G_f = \texttt{rR}\} \\ &\quad + P\{G_1 = \texttt{rr}, G_2 = \texttt{rr} \mid G_m = \texttt{rR}, G_f = \texttt{rr}\} P\{G_m = \texttt{rR}, G_f = \texttt{rr}\} \\ &\quad + P\{G_1 = \texttt{rr}, G_2 = \texttt{rr} \mid G_m = \texttt{rR}, G_f = \texttt{rR}\} P\{G_m = \texttt{rR}, G_f = \texttt{rR}\}. \\ &= 1 \times q^4 \\ &\quad + (1/2)^2 \times 2(1-q)q^3 \\ &\quad + (1/2)^2 \times 2(1-q)q^3 \\ &\quad + (1/4)^2 \times 4(1-q)^2 q^2 \\ &= (1/4)q^2(1+q)^2. \end{aligned}$$

Then

$$P(R_1 \mid R_2) = \frac{P(R_1 \cap R_2)}{P(R_2)} = \frac{(1/4)q^2(1+q)^2}{q^2} = (1/4)(1+q)^2.$$

(c) The lower bound of $P(R_1 \mid R_m) = q$ is 0, while the lower bound of $P(R_1 \mid R_2) = (1/4)(1+q)^2$ is 1/4.

Problem 1.30 In genetics, a genotype consists of two genes, each of which is one of (possibly) several types of alleles. When two organisms mate, each passes one allele, selected at random, to the offspring, forming that offspring's genotype. Suppose a gene of a type of flower has two alleles, `r` and `R`. A plant possessing genotype `rr`, `rR` or `RR` has white, pink or red petals, respectively. A trait like this, in which both alleles determine the trait, is called *codominant*.

Suppose a pink flower and a flower A of unknown color produce offspring B. Let G_A, G_B be the respective genotypes, and let $P_A = (p_{rr}, p_{rR}, p_{RR})$ be the genotype probabilities for A.

(a) Evaluate:

$$Odds(B \text{ is white} \mid A \text{ is white}),$$
$$Odds(B \text{ is white} \mid A \text{ is pink}),$$
$$Odds(B \text{ is white} \mid A \text{ is red}).$$

Do these quantities depend on P_A in any way?

(b) Determine $\alpha_w, \alpha_p, \alpha_r$ in the following expressions :

$$\begin{aligned} Odds(A \text{ is white} \mid B \text{ is white}) &= \alpha_w \times Odds(A \text{ is white}), \\ Odds(A \text{ is white} \mid B \text{ is pink}) &= \alpha_p \times Odds(A \text{ is white}), \\ Odds(A \text{ is white} \mid B \text{ is red}) &= \alpha_r \times Odds(A \text{ is white}). \end{aligned}$$

(c) If any of the quantities $\alpha_w, \alpha_p, \alpha_r$ depends on P_A, in what sense do they *not* depend on the prevalence $prev = P(A \text{ is white})$?

(d) Verify that $\alpha_w > \alpha_p > \alpha_r$.

(e) Suppose, as would be predicted under Hardy-Weinberg equilibrium, we have

$$\begin{aligned} p_{rr} &= (1-q_R)^2, \\ p_{rR} &= 2q_R(1-q_R), \\ p_{RR} &= q_R^2, \end{aligned}$$

where q_R is the population frequency of allele R. We say R is a *rare allele* if $q_R \ll 1$. What happens to the quantities $\alpha_w, \alpha_p, \alpha_r$ as q_R approaches 0?

SOLUTION: The following table gives the offspring genotype probability distribution for each combination of parents:

		Offspring Genotype		
Parent 1	Parent 2	rr	rR	RR
rr	rr	1	0	0
rr	rR	1/2	1/2	0
rr	RR	0	1	0
rR	rR	1/4	1/2	1/4
rR	RR	0	1/2	1/2
RR	RR	0	0	1

(a) The genotype probabilities for B conditional on parental genotypes can be evaluated exactly from the table:

$$\begin{aligned} Odds(B \text{ is white} \mid A \text{ is white}) &= \frac{P(G_B = \texttt{rr} \mid G_A = \texttt{rr})}{1 - P(G_B = \texttt{rr} \mid G_A = \texttt{rr})} = \frac{1/2}{1-1/2} = 1, \\ Odds(B \text{ is white} \mid A \text{ is pink}) &= \frac{P(G_B = \texttt{rr} \mid G_A = \texttt{rR})}{1 - P(G_B = \texttt{rr} \mid G_A = \texttt{rR})} = \frac{1/4}{1-1/4} = 1/3, \\ Odds(B \text{ is white} \mid A \text{ is red}) &= \frac{P(G_B = \texttt{rr} \mid G_A = \texttt{RR})}{1 - P(G_B = \texttt{rr} \mid G_A = \texttt{RR})} = \frac{0}{1-0} = 0. \end{aligned}$$

No knowledge of P_A is needed.

(b) Using Baye's Theorem for odds:

$$\begin{aligned} Odds(A \text{ is white} \mid B \text{ is white}) &= \frac{P(B \text{ is white} \mid A \text{ is white})}{P(B \text{ is white} \mid A \text{ is not white})} \times Odds(A \text{ is white}) \\ &= \frac{P(G_B = \texttt{rr} \mid G_A = \texttt{rr})}{P(G_B = \texttt{rr} \mid G_A \neq \texttt{rr})} \times Odds(A \text{ is white}). \end{aligned}$$

We have

$$P(G_B = \texttt{rr} \mid G_A = \texttt{rr}) = 1/2,$$

and

$$P(G_B = \texttt{rr} \mid G_A \neq \texttt{rr}) = \frac{P(G_B = \texttt{rr} \cap G_A = \texttt{rR}) + P(G_B = \texttt{rr} \cap G_A = \texttt{RR})}{P(G_A = \texttt{rR}) + P(G_A = \texttt{RR})} = \frac{\frac{1}{4} \cdot p_{rR} + 0 \cdot p_{RR}}{p_{rR} + p_{RR}},$$

so

$$\alpha_w = 2 \times \frac{p_{rR} + p_{RR}}{p_{rR}}.$$

We next have:

$$\begin{aligned} Odds(A \text{ is white} \mid B \text{ is pink}) &= \frac{P(B \text{ is pink} \mid A \text{ is white})}{P(B \text{ is pink} \mid A \text{ is not white})} \times Odds(A \text{ is white}) \\ &= \frac{P(G_B = \texttt{rR} \mid G_A = \texttt{rr})}{P(G_B = \texttt{rR} \mid G_A \neq \texttt{rr})} \times Odds(A \text{ is white}). \end{aligned}$$

We have

$$P(G_B = \texttt{rR} \mid G_A = \texttt{rr}) = 1/2,$$

and

$$P(G_B = \texttt{rR} \mid G_A \neq \texttt{rr}) = \frac{P(G_B = \texttt{rR} \cap G_A = \texttt{rR}) + P(G_B = \texttt{rR} \cap G_A = \texttt{RR})}{P(G_A = \texttt{rR}) + P(G_A = \texttt{RR})} = \frac{\frac{1}{2} \cdot p_{rR} + \frac{1}{2} \cdot p_{RR}}{p_{rR} + p_{RR}} = 1/2.$$

so

$$\alpha_p = 1.$$

Finally, since

$$P(B \text{ is red} \mid A \text{ is white}) = P(G_B = \texttt{RR} \mid G_A = \texttt{rr}) = 0,$$

and $P(B \text{ is red} \mid A \text{ is not white}) > 0$, we must have

$$\alpha_r = 0.$$

(c) Of the three coefficients $\alpha_w, \alpha_p, \alpha_r$, only α_w is dependent on P_A. It can be written

$$\alpha_w = \frac{2}{P(A \text{ is pink} \mid A \text{ is not white})}.$$

This quantity depends on P_A only through the ratio p_{RR}/p_{rR}, which does not depend on $p_{rr} = P(A \text{ is white})$.

(d) Clearly $\alpha_w \geq 2$. The result follows by noting $\alpha_p = 1$ and $\alpha_r = 0$.

(e) Since α_p and α_r do not depend on P_A in any way, they cannot depend on q_R. On the other hand, we have

$$\alpha_w = 2 \times \frac{2q_R(1-q_R) + q_R^2}{2q_R(1-q_R)} = 2 \times \left(1 + \frac{1}{2}\frac{q_R}{(1-q_R)}\right).$$

This quantity approaches 2 from above as q_R approaches 0. Note that for small q_R $Odds(A \text{ is white})$ is already very large. However, the event $\{B \text{ is white}\}$ is still informative, in the sense that it's occurrence increases the odds of the event $\{A \text{ is white}\}$ by a factor of just over two.

1.6 Random Variables (Basic Concepts)

Problem 1.31 Consider the following CDF for a random variable X:

$$F_X(x) = \begin{cases} 0 & ; \quad x < 0 \\ x/2 & ; \quad x \in [0, 1/3) \\ 1/6 & ; \quad x \in [1/3, 2/3) \\ x/8 + 2/3 & ; \quad x \in [2/3, 2) \\ 1 & ; \quad x \geq 2 \end{cases}$$

(a) Sketch the CDF, then determine the following probabilities:
(b) $P(X > 1)$.
(c) $P(X \in (3/7, 4/7))$.
(d) $P(X = 2/3)$.
(e) $P(X = 2)$.
(f) $P(X \leq 2)$.
(g) $P(X < 2)$.

SOLUTION:

(a) Figure is given in Figure 1.1. This was produced by the following R code:

```
ex1 = expression(paste(italic(x)))
ex2 = expression(paste(italic(F)[italic(X)],'(',italic(x),')',sep=''))

par(mfrow=c(1,1),mar=c(4,2,4,2))
plot(c(0,1/3,2/3),c(0,1/6,1/6),axes=F,xlab='',ylab='',type='l',
     cex=1.25,ylim=c(0,1.1),xlim=c(0,4),lty=1,lwd=2)
lines(c(2/3,2),c(3/4,11/12),lty=1,lwd=2)
arrows(2,1,4,1,lwd=2,length=0.15)
points(c(2/3,2),c(3/4,1),pch=19,cex=1.25)
axis(1,line=-1.2,at=c(0,1/3,2/3,2,4),labels=c('0','1/3','2/3','2',''),
     cex.axis=1.5)
axis(2,line=-1.5,at=c(0,1/6,3/4,11/12,1),
     labels=c('0','1/6','3/4','11/12','1'),cex.axis=1.5)

lines(c(2/3,2/3),c(1/6,3/4),col='gray',lty=2)
lines(c(2,2),c(11/12,1),col='gray',lty=2)

text(0.8,1.0,ex2,cex=2.0)
mtext(ex1,side=1,line=1.5,cex=2.0)
```

(b) $P(X > 1) = 1 - P(X \leq 1) = 1 - F_X(1) = 1 - 1/8 - 2/3 = 5/24$.
(c) First note that

$$P(X \in (1/3, 2/3]) = P(X \leq 2/3) - P(X \leq 1/3) = F_X(2/3) - F_X(1/3) = 1/6 - 1/6 = 0.$$

Then $P(X \in (3/7, 4/7)) = 0$, since $(3/7, 4/7) \subset (1/3, 2/3]$.
(d) $P(X = 2/3) = 3/4 - 1/6$, directly from CDF.
(e) $P(X = 2) = 1 - 11/12 = 1/12$, directly from CDF.
(f) $P(X \leq 2) = F_X(2) = 1$.
(g) $P(X < 2) = P(X \leq 2) - P(X = 2) = 1 - 1/12 = 11/12$.

Problem 1.32 Conditional probabilities can be used to construct new conditional distributions of random variables in a natural way. Let X be a random variable and let A be an event. If we can evaluate $P(X \leq x \mid A)$ then we have a completely defined distribution for X *conditional on* A (this new random

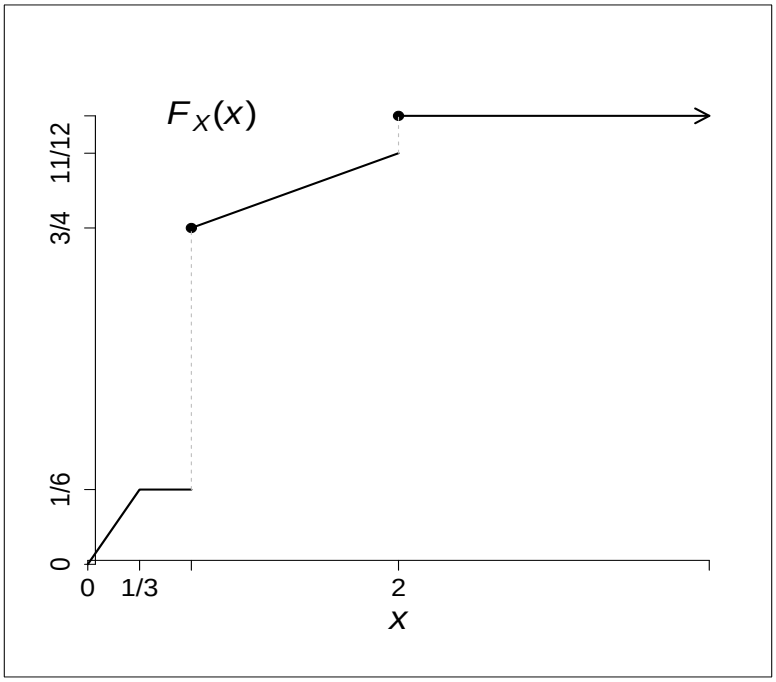

Figure 1.1: CDF for Problem 1.31.

variable may be denoted $X \mid A$). The CDF under this conditional distribution is typically denoted in the literature as $F_{X|A}(x)$ or $F_X(x \mid A)$, and the expected value conditional on A is denoted $E[X \mid A]$.

The law of total probability (Section 2.7) also extends to expected values. Suppose we have a partition of the sample space $A_1, \ldots, A_n$. Then for any random variable X we have

$$E[X] = E[X \mid A_1]P(A_1) + E[X \mid A_2]P(A_2) + \ldots + E[X \mid A_n]P(A_n). \tag{1.5}$$

Calculating expected values by conditioning is often a very useful device. The following problem is a typical example.

A rat attempts to navigate a maze. It is possible for the rat to choose a path which will lead back to the starting point before exiting the maze. If this happens, the rat continues to attempt to exit the maze with no memory of its previous attempts. Let E be the event that the rat returns to the starting point before exiting the maze (it may do this more than once). Suppose $P(E) = p$. Also, let T_0 be the expected time to exit conditional on E^c, and let T_1 be the expected time to return to the starting point conditional on E. If W is the time taken to exit the maze, determine $E[W]$ as a function of p, T_0 and T_1.

SOLUTION: We condition on the event E, leading to

$$\begin{aligned} E[W \mid E^c] &= T_0, \text{ and} \\ E[W \mid E] &= T_1 + E[W]. \end{aligned}$$

The second conditional expectation holds because if the rat returns to the starting point, an expected time of T_1 expires, then the rat "starts again", with an expected time of $E[W]$ remaining until the exit. Using

(1.5) we have

$$\begin{aligned} E[W] &= E[W \mid E^c]P(E^c) + E[W \mid E]P(E) \\ &= T_0(1-p) + (T_1 + E[W])p. \end{aligned}$$

Solving for $E[W]$ gives

$$E[W] = T_0 + T_1 \frac{p}{1-p}.$$

Problem 1.33 A game involving a single dice is played in the following way. The dice is tossed repeatedly, and each outcome is recorded. Let X be the number of tosses needed to see an outcome already observed. For example, if the first four tosses are $3, 1, 6, 3$, then $X = 4$, and the game can stop. The largest value of X decides the winner.

(a) What is the support S_X of X?
(b) Using the rules of combinatorics, derive $P(X > i)$ for each $i \in S_X$.
(c) Use the answer of Part (b) to determine the PMF $p_i = P(X = i)$ for each $i \in S_X$.

SOLUTION:

(a) $S_X = \{2, 3, 4, 5, 6, 7\}$.
(b) If $X > i$, then the first i tosses are distinct. Using the rule of product, there are

$$N = \binom{6}{i} \times i! = \frac{6!}{(6-i)!}$$

such tosses (choose i from 6 distinct outcomes, then choose one of $i!$ orderings. Then there are $D = 6^i$ sequences of tosses of length i of any kind. Then

$$P(X > i) = \frac{N}{D} = \frac{6!}{6^i(6-i)!} = \frac{6}{6} \times \frac{5}{6} \times \cdots \times \frac{6-i+1}{6}.$$

In particular, we have

$$\begin{aligned} P(X > 1) &= \frac{6}{6} = 1 \\ P(X > 2) &= \frac{6}{6} \times \frac{5}{6} \approx 0.833 \\ P(X > 3) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \approx 0.556 \\ P(X > 4) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \approx 0.278 \\ P(X > 5) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \approx 0.0926 \\ P(X > 6) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{1}{6} \approx 0.0154 \\ P(X > 7) &= 0. \end{aligned}$$

(c) For $i = 2, \ldots, 7$ we have

$$\begin{aligned} p_i &= P(X = i) \\ &= P(X > i-1) - P(X > i) \\ &= \frac{6!}{6^{i-1}(6-i+1)!} - \frac{6!}{6^i(6-i)!} \\ &= \frac{6!}{6^{i-1}(6-i+1)!}\left[1 - \frac{6-i+1}{6}\right] \\ &= \frac{6!}{6^{i-1}(6-i+1)!}\left[\frac{i-1}{6}\right] \\ &= \left[\frac{6}{6} \times \frac{5}{6} \times \cdots \times \frac{6-i+2}{6}\right] \times \left[\frac{i-1}{6}\right]. \end{aligned}$$

In particular, we have

$$\begin{aligned} P(X=2) &= \frac{6}{6} \times \frac{1}{6} \approx 0.167 \\ P(X=3) &= \frac{6}{6} \times \frac{5}{6} \times \frac{2}{6} \approx 0.278 \\ P(X=4) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \approx 0.278 \\ P(X=5) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{4}{6} \approx 0.185 \\ P(X=6) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{1}{6} \approx 0.0772 \\ P(X=7) &= \frac{6}{6} \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{1}{6} \times \frac{6}{6} \approx 0.0154. \end{aligned}$$

Problem 1.34 A random variable X possesses a "V"-shaped density on the interval $[0, 4]$. This is constructed by taking the density to be $f_X(x) = cg(x)$ for some constant c, where $g(x)$ is the "V"-shaped function:

$$g(x) = \begin{cases} -(x-2) & ; \quad x \in (0,2) \\ x-2 & ; \quad x \in [2,4) \\ 0 & ; \quad \text{elsewhere} \end{cases}$$

(a) Determine c.
(b) Determine the CDF $F(x) = P(X \leq x)$. Give this as a function of $x \in (-\infty, \infty)$.

SOLUTION:

(a) Directly from the formula for the area of a triangle, we have

$$\int_{-\infty}^{\infty} g(u)du = \int_0^2 -(u-2)du + \int_2^4 (u-2)du = 2 + 2 = 4.$$

Then $c = 1/4$.

(b) First note that $f_X(x) = 0$ for $x \leq 0$ and $x \geq 4$. This means

$$\begin{aligned} F(x) &= 0, \quad x \leq 0, \quad \text{and} \\ F(x) &= 1, \quad x \geq 4. \end{aligned}$$

Then, for $x \in (0, 2)$

$$\begin{aligned} F(x) &= \int_0^x -(1/4)(u-2)du \\ &= (2u - u^2/2)/4\Big|_0^x \\ &= x/2 - x^2/8. \end{aligned}$$

For $x \in [2, 4)$,

$$\begin{aligned} F(x) &= \int_0^2 (1/4)(u-2)du + \int_2^x (1/4)(u-2)du \\ &= 1/2 + (u^2/2 - 2u)/4\Big|_2^x \\ &= 1/2 + (x^2/2 - 2x)/4 - (2^2/2 - 2 \cdot 2)/4 \\ &= 1 + (x^2/2 - 2x)/4. \end{aligned}$$

The complete CDF is then

$$F(x) = \begin{cases} 0 & ; \quad x \leq 0 \\ x/2 - x^2/8 & ; \quad x \in (0, 2] \\ 1 + x^2/8 - x/2 & ; \quad x \in (2, 4) \\ 1 & ; \quad x \geq 4 \end{cases}.$$

Problem 1.35 Recall the "hat" problem. At a party, N men bring identical hats. At the end of the party each man brings home one of the hats chosen at random. Let S_N be the number of men who bring home their own hats. It was shown that $E[S_N] = 1$, which, interestingly, does not depend on N.

For this problem, derive the variance of S_N. **HINT:** Let $X_i = 1$ if the ith man brings home his own hat, and let $X_i = 0$ otherwise. Then $S_N = \sum_{i=1}^n X_i$. What is the covariance of X_i, X_j for any $i \neq j$?

SOLUTION: First, note that X_i is a Bernoulli RV with $E[X_i] = p_i = 1/N$ and variance $\sigma_i^2 = p_i(1-p_i) = (N-1)/N^2$. Then the variance of the sum is

$$\text{var}\,[S_N] = \sum_{i=1}^n \sigma_i^2 + 2\sum_{i<j} \sigma_{ij},$$

where σ_{ij} is the covariance of X_i, X_j, $i \neq j$. First, note that

$$E[X_i X_j] = P(X_i = X_j = 1) = P(i\text{th and } j\text{th men bring home their own hats}).$$

This can be calculated using conditional probabilities:

$$\begin{aligned} P(X_i = X_j = 1) &= P(X_i = 1 \mid X_j = 1)P(X_j = 1) \\ &= \frac{1}{N-1} \times \frac{1}{N}, \end{aligned}$$

since to calculate $P(X_i = 1 \mid X_j = 1)$ we may simply reduce the number of hats by 1. The covariance is then

$$\sigma_{ij} = E[X_i X_j] - E[X_i]E[X_j] = \frac{1}{N-1} \times \frac{1}{N} - \frac{1}{N^2} = \frac{1}{N^2(N-1)}.$$

Since the relevant variances and covariances are all equal, we can write

$$\begin{aligned} \text{var}\,[S_N] &= N \times \sigma_1^2 + 2 \times \binom{N}{2} \sigma_{12} \\ &= N \times \sigma_1^2 + N(N-1)\sigma_{12} \\ &= N(N-1)/N^2 + N(N-1)/[N^2(N-1)] \\ &= (N-1)/N + 1/N \\ &= 1. \end{aligned}$$

Problem 1.36 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} cx^3 & ; \quad x \in [0,4] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine CDF $F_X(x)$.

SOLUTION:

(a) We have

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_0^4 cu^3 du = \frac{cu^4}{4}\Big|_0^4 = c \times 64.$$

This means $c = 1/64$.

(b) for $x \leq 0$ we have $F_X(x) = 0$, and for $x \geq 4$ we have $F_X(x) = 1$. For $x \in [0,4]$ we have

$$F_X(x) = \int_{-\infty}^{x} u^3/64 du = \frac{x^4}{256},$$

so that

$$F_X(x) = \begin{cases} 0 & ; \quad x \leq 0 \\ \frac{x^4}{256} & ; \quad x \in (0,4] \\ 1 & ; \quad x > 4 \end{cases}$$

Problem 1.37 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} cx^3 & ; \quad x \in [0,3] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine the 0.5-quantile for this density.

SOLUTION:

(a) The integral of a density evaluates to 1, so

$$1 = \int_0^3 cx^3 dx = cx^4/4\Big|_0^3 = c \times 81/4,$$

giving $c = 4/81$.

(b) The CDF is

$$F_X(x) = \begin{cases} 0 & ; \quad x < 0 \\ x^4/81 & ; \quad x \in [0,3) \\ 1 & ; \quad x \geq 3 \end{cases}.$$

The 0.5-quantile q is the solution to

$$0.5 = F_X(q) = q^4/81, \quad \text{or} \quad q = (0.5 \times 81)^{1/4} \approx 2.523.$$

Problem 1.38 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} cx^4 & ; \quad x \in [0,2] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine the expected value of X.

SOLUTION:

(a) The integral of a density evaluates to 1, so

$$1 = \int_0^2 cx^4 dx = cx^5/5\Big|_0^2 = c \times 32/5,$$

giving $c = 5/32$.

(b) Then $E[X]$ is

$$\begin{aligned} E[X] &= (5/32)\int_0^2 x \times x^4 dx \\ &= (5/32) \times x^6/6\Big|_0^2 \\ &= (5/32) \times (64/6) = 5/3. \end{aligned}$$

Problem 1.39 Two dice are tossed independently. Let X_1, X_2 be random variables representing the two outcomes, each from sample space $S_X = \{1,2,3,4,5,6\}$. Derive the proability mass function of the following random variables:

(a) $X = X_1 + X_2$,
(b) $X = \max(X_1, X_2)$.

SOLUTION: We have a random experiment with 36 equiprobable outcomes from sample space

$$S = \{(i, j) : i = 1, \ldots, 6, \;\; j = 1, \ldots, 6\}.$$

The random variables X_1, X_2 are determined by representing the outcome as (X_1, X_2).

(a) The PMF for X is given by

$$p_i = P(X = i) = P(X_1 + X_2 = i).$$

The support of X is $\mathcal{S}_X = \{2, 3, \ldots, 11, 12\}$. Then the PMF is given by

$$\begin{aligned}
p_2 = P(X = 2) &= P\left((X_1, X_2) \in \{(1,1)\}\right) = 1/36, \\
p_3 = P(X = 3) &= P\left((X_1, X_2) \in \{(1,2),(2,1)\}\right) = 2/36, \\
p_4 = P(X = 4) &= P\left((X_1, X_2) \in \{(1,3),(2,2),(3,1)\}\right) = 3/36, \\
p_5 = P(X = 5) &= P\left((X_1, X_2) \in \{(1,4),(2,3),(3,2),(4,1)\}\right) = 4/36, \\
p_6 = P(X = 6) &= P\left((X_1, X_2) \in \{(1,5),(2,4),(3,3),(4,2),(5,1)\}\right) = 5/36, \\
p_7 = P(X = 7) &= P\left((X_1, X_2) \in \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}\right) = 6/36, \\
p_8 = P(X = 8) &= P\left((X_1, X_2) \in \{(2,6),(3,5),(4,4),(5,3),(6,2)\}\right) = 5/36, \\
p_9 = P(X = 9) &= P\left((X_1, X_2) \in \{(3,6),(4,5),(5,4),(6,3)\}\right) = 4/36, \\
p_{10} = P(X = 10) &= P\left((X_1, X_2) \in \{(4,6),(5,5),(6,4)\}\right) = 3/36, \\
p_{11} = P(X = 11) &= P\left((X_1, X_2) \in \{(5,6),(6,5)\}\right) = 2/36, \\
p_{12} = P(X = 12) &= P\left((X_1, X_2) \in \{(6,6)\}\right) = 1/36.
\end{aligned}$$

(b) The PMF for X is given by

$$p_i = P(X = i) = P(\max(X_1, X_2) = i).$$

The support of X is $\mathcal{S}_X = \{1, 2, 3, 4, 5, 6\}$. Then the PMF is given by

$$\begin{aligned}
p_1 = P(X = 1) &= P\left((X_1, X_2) \in \{(1,1)\}\right) = 1/36, \\
p_2 = P(X = 2) &= P\left((X_1, X_2) \in \{(1,2),(2,1),(2,2)\}\right) = 3/36, \\
p_3 = P(X = 3) &= P\left((X_1, X_2) \in \{(1,3),(2,3),(3,3),(3,1),(3,2)\}\right) = 5/36, \\
p_4 = P(X = 4) &= P\left((X_1, X_2) \in \{(1,4),(2,4),(3,4),(4,4),(4,3),(4,2),(4,1)\}\right) = 7/36, \\
p_5 = P(X = 5) &= P((X_1, X_2) \in \{(1,5),(2,5),(3,5),(4,5),(5,5),\ldots \\
&\quad \ldots,(5,4),(5,3),(5,2),(5,1)\}) = 9/36, \\
p_6 = P(X = 6) &= P((X_1, X_2) \in \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6),\ldots \\
&\quad \ldots,(6,5),(6,4),(6,3),(6,2),(6,1)\}) = 11/36.
\end{aligned}$$

Problem 1.40 Two dice are tossed independently. Let X_1, X_2 be random variables representing the two outcomes, each from sample space $S_X = \{1, 2, 3, 4, 5, 6\}$. Derive the probability mass function of the following random variables:

(a) $X = X_1 - X_2$,

(b) $X = \begin{cases} X_1 & ; \quad X_1 = X_2 \\ 0 & ; \quad X_1 \neq X_2 \end{cases}$.

SOLUTION: We have a random experiment with 36 equiprobable outcomes from sample space

$$S = \{(i,j) : i = 1,\ldots,6, \;\; j = 1,\ldots,6\}.$$

The random variables X_1, X_2 are determined by representing the outcome as (X_1, X_2).

(a) The PMF for X is given by

$$p_i = P(X = i) = P(X_1 + X_2 = i).$$

The support of X is $\mathcal{S}_X = \{2, 3, \ldots, 11, 12\}$. Then the PMF is given by

$$\begin{aligned}
p_{-5} = P(X = -5) &= P\left((X_1, X_2) \in \{(1,6)\}\right) = 1/36, \\
p_{-4} = P(X = -4) &= P\left((X_1, X_2) \in \{(1,5),(2,6)\}\right) = 2/36, \\
p_{-3} = P(X = -3) &= P\left((X_1, X_2) \in \{(1,4),(2,5),(3,6)\}\right) = 3/36, \\
p_{-2} = P(X = -2) &= P\left((X_1, X_2) \in \{(1,3),(2,4),(3,5),(4,6)\}\right) = 4/36, \\
p_{-1} = P(X = -1) &= P\left((X_1, X_2) \in \{(1,2),(2,3),(3,4),(4,5),(5,6)\}\right) = 5/36, \\
p_0 = P(X = 0) &= P\left((X_1, X_2) \in \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}\right) = 6/36, \\
p_1 = P(X = 1) &= P\left((X_1, X_2) \in \{(2,1),(3,2),(4,3),(5,4),(6,5)\}\right) = 5/36, \\
p_2 = P(X = 2) &= P\left((X_1, X_2) \in \{(3,1),(4,2),(5,3),(6,4)\}\right) = 4/36, \\
p_3 = P(X = 3) &= P\left((X_1, X_2) \in \{(4,1),(5,2),(6,3)\}\right) = 3/36, \\
p_4 = P(X = 4) &= P\left((X_1, X_2) \in \{(5,1),(6,2)\}\right) = 2/36, \\
p_5 = P(X = 5) &= P\left((X_1, X_2) \in \{(6,1)\}\right) = 1/36.
\end{aligned}$$

(b) For any $i = 1, \ldots, 6$ the PMF is given by

$$p_i = P(X = i) = P\left((X_1, X_2) = (i,i)\right) = 1/36.$$

Then

$$p_0 = P(X = 0) = 1 - \sum_{i=1}^{6} p_i = 1 - 6 \times \frac{1}{36} = 5/6.$$

To summarize, $(p_0, p_1, \ldots, p_6) = (5/6, 1/36, \ldots, 1/36)$.

Problem 1.41 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} c(x+1) & ; \quad x \in [1,3] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine $P(X \leq 2)$.

SOLUTION:

(a) We have

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_1^3 c(u+1)du = c(u^2/2+u)\Big|_1^3 = c \times 6.$$

This means $c = 1/6$.

(b) We have

$$P(X \leq 2) = \int_{-\infty}^{2} f_x(u)du = \int_1^2 (u+1)/6\, du = (u^2/2+u)/6\,\Big|_1^2 = 5/12.$$

Problem 1.42 In a certain game, two dice are tossed independently. A player wins if at least one dice shows a 6. The game is played 10 times, and X is the number of times the player wins. Give the mean and variance of X.

SOLUTION: Let $E_i = \{\text{Dice } i \text{ shows } 6\}$, $i = 1, 2$

$$\begin{aligned} P(\text{ At least one dice shows a 6 }) &= P(E_1 \cup E_2) \\ &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= 1/6 + 1/6 - 1/36 \\ &= 11/36 \end{aligned}$$

Then $X \sim bin(n, p)$ where $n = 10$, $p = 11/36$. For any binomial random variable, the mean and variance is given by

$$\begin{aligned} E[X] &= np \\ &= 10 \times 11/36 \\ &= 55/18 \\ &\approx 3.06, \end{aligned}$$

and

$$\begin{aligned} var(X) &= np(1-p) \\ &= 10 \times (11/36) \times (25/36) \\ &= 1375/648 \\ &\approx 2.12. \end{aligned}$$

Problem 1.43 A game of chance is played in the following way. A penny (face value 1 cent), nickel (face value 5 cents), dime (face value 10 cents) and quarter (face value 25 cents) are tossed independently. A player gets to keep any coin that shows *heads*. Let X be the sum of the face values (in cents) of all coins that show *heads*. For example, if only the penny and dime show heads, we would have $X = 11$.

Derive the mean and variance of X.

SOLUTION: We can express X as

$$X = 1 \times I_1 + 5 \times I_2 + 10 \times I_3 + 25 \times I_4,$$

where $I_1, \ldots, I_4$ are independent Bernoulli random variables each of mean $\mu = 1/2$, and therefore variance $\sigma_2 = 1/4$. Then

$$\begin{aligned} E[X] &= 1 \times E[I_1] + 5 \times E[I_2] + 10 \times E[I_3] + 25 \times E[I_4] \\ &= \frac{1}{2}(1 + 5 + 10 + 25) \\ &= \frac{41}{2} = 20.5. \end{aligned}$$

Since the terms in the sum are independent, we have

$$\begin{aligned} var[X] &= var[1 \times I_1] + var[5 \times I_2] + var[10 \times I_3] + var[25 \times I_4] \\ &= 1^2 \times var[I_1] + 5^2 \times var[I_2] + 10^2 \times var[I_3] + 25^2 \times var[I_4] \\ &= \frac{1}{4}\left(1^2 + 5^2 + 10^2 + 25^2\right) \\ &= \frac{751}{4} = 187.75. \end{aligned}$$

Problem 1.44 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} -c(x-2)(x-4) & ; \quad x \in [2,4] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine the CDF $F(x) = P(X \leq x)$. Give this as a function of $x \in (-\infty, \infty)$.

SOLUTION:

(a) We have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(u)du \\ &= \int_2^4 -c(u-2)(u-4)du \\ &= -c(u^3/3 - 3u^2 + 8u)\Big|_2^4 \\ &= -c \times (16/3 - 20/3) \\ &= c \times 4/3. \end{aligned}$$

This means $c = 3/4$.

(b) First note that $f_X(x) = 0$ for $x \leq 2$ and $x > 4$, so that $P(2 < X \leq 4) = 1$. This implies $F(x) = 0$ for $x \leq 2$ and $F(x) = 1$ for $x > 4$. Otherwise

$$\begin{aligned} F(x) &= \int_2^x f_X(u)du \\ &= \int_2^x -3/4 \times (u-2)(u-4)du \\ &= -3/4 \times (u^3/3 - 3u^2 + 8u)\Big|_2^x \\ &= -3/4 \times (x^3/3 - 3x^2 + 8x) + 5, \quad \text{for } x \in (2,4]. \end{aligned}$$

This gives

$$F(x) = \begin{cases} 0 & ; \quad x \leq 2 \\ -3/4 \times (x^3/3 - 3x^2 + 8x) + 5 & ; \quad x \in (2,4] \\ 1 & ; \quad x > 4 \end{cases}.$$

Problem 1.45 A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} c|x|^{1.5} & ; \quad x \in [-1,1] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.
(b) Determine the 0.25-quantile for this density.

SOLUTION:

(a) The integral of a density evaluates to 1, so

$$\begin{aligned} 1 &= \int_{-1}^1 c|x|^{1.5}dx \\ &= 2\int_0^1 cx^{1.5}dx \\ &= 2 \times c \times x^{2.5}/2.5\Big|_0^1 \\ &= 2 \times c/2.5, \end{aligned}$$

giving $c = 5/4$ (the evaluation makes use of the symmetry of $f_X(x)$ about 0).

(b) By symmetry the 0.25-quantile is less than 0. For $x \in [-1,0]$ the CDF is

$$\begin{aligned} F_X(x) &= (5/4)\int_{-1}^x (-x)^{1.5}dx \\ &= -(5/4)(-x)^{2.5}/2.5\Big|_{-1}^x \\ &= (1/2)\left[1 - (-x)^{2.5}\right]. \end{aligned}$$

The 0.25-quantile q is the solution to

$$0.25 = (1/2)\left[1 - (-x)^{2.5}\right],$$

or

$$x = -(1/2)^{1/2.5} \approx -0.7578583.$$

Problem 1.46 Suppose X is a nonnegative random variable, that is, $P(X \geq 0) = 1$. Assume that X is a discrete random variable with sample space $S_X = \{0, 1, 2, \ldots\}$. Show that

$$E[X] = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

Hint: Express X as a sum of Bernoulli random variable.

SOLUTION:

(a) Set $U_n = 1$ if $n \leq X$ and $U_n = 0$ otherwise. Then

$$X = \sum_{i=1}^{\infty} U_i,$$

and so

$$E[X] = \sum_{n=1}^{\infty} E[U_n] = \sum_{n=1}^{\infty} P(X \geq n) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

Problem 1.47 Suppose there are m types of coupons. A collector samples them one at a time. Whenever a collector samples another coupon, assume it is of each type with equal probability, and that the selections are independent. What is the expected number of coupon samples needed to have at least one of each type? [Hint: Suppose T_j is the number of samples at which the jth new coupon is observed by the collector. What is the distribution of $T_{j+1} - T_j$, for any $j = 1, 2, \ldots, m-1$?]

SOLUTION: We wish to find $E[T_m]$. First note that $P(T_1 = 1) = 1$. Then, immediately following sample T_j, there are $m - j$ unseen coupons remaining. The number of samples needed to observe the next new coupon has a geometric distribution with mean $m/(m-j)$. Then,

$$T_m = [T_m - T_{m-1}] + [T_{m-1} - T_{m-2}] + \ldots + [T_2 - T_1] + T_1.$$

By the preceding argument this gives

$$E[T_m] = m + m/2 + \ldots + m/(m-1) + 1.$$

1.7 Random Variables (Commonly Used Distributions)

Problem 1.48 Recall that for a binomial random variable $X \sim bin(n, p)$, the probability mass function is given by

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Then suppose we are given two independent binomial random variables $X_1 \sim bin(n_1, p)$ and $X_2 \sim bin(n_2, p)$. We may have $n_1 \neq n_2$, but the value of p is the same for X_1 and X_2.

Derive an expression for the conditional probability $P(X_1 = s \mid X_1 + X_2 = t)$, where s, t are integers satisfying $0 \leq t \leq n_1 + n_2$ and $0 \leq s \leq \min(n_1, t)$. Does this conditional probability depend on p? **HINT:** What distribution does $X_1 + X_2$ have? Recall that a binomial random variable is a sum of independent identically distributed Bernoulli random variables, which assume values 0 or 1.

SOLUTION: First, note that $X_1 + X_2 \sim bin(n_1 + n_2, p)$. The we have

$$\begin{aligned} P(X_1 = s \mid X_1 + X_2 = t) &= \frac{P(X_1 = s \text{ AND } X_1 + X_2 = t)}{P(X_1 + X_2 = t)} \\ &= \frac{P(X_1 = s \text{ AND } X_2 = t - s)}{P(X_1 + X_2 = t)} \\ &= \frac{P(X_1 = s)P(X_2 = t - s)}{P(X_1 + X_2 = t)} \\ &= \frac{\binom{n_1}{s}p^s(1-p)^{n_1-s} \times \binom{n_2}{t-s}p^{t-s}(1-p)^{n_2-(t-s)}}{\binom{n_1+n_2}{t}p^t(1-p)^{n_1+n_2-t}} \\ &= \frac{\binom{n_1}{s}\binom{n_2}{t-s}p^t(1-p)^{n_1+n_2-t}}{\binom{n_1+n_2}{t}p^t(1-p)^{n_1+n_2-t}} \\ &= \frac{\binom{n_1}{s}\binom{n_2}{t-s}}{\binom{n_1+n_2}{t}}. \end{aligned}$$

No, the conditional probability does not depend on p.

Problem 1.49 For a certain aptitude test, students from Schools A and B have test scores distributed as normal random variables $X_A \sim N(\mu_A, \sigma_A^2)$ and $X_B \sim N(\mu_B, \sigma_B^2)$, respectively. Assume the mean and variances are $\mu_A = 104$, $\sigma_A^2 = 23^2$ and $\mu_B = 123$, $\sigma_B^2 = 14^2$. Suppose one student is randomly chosen from each school.

What is the probability that the student from School A has the higher score? The following table gives the upper tail probability $P(Z > z)$ for selected values of z, where $Z \sim N(0, 1)$ is the standard normal random variable. You may make use of the value which gives the best approximation.

z	$P(Z > z)$	z	$P(Z > z)$
0.00	0.50	0.60	0.27
0.10	0.46	0.70	0.24
0.20	0.42	0.80	0.21
0.30	0.38	0.90	0.18
0.40	0.34	1.00	0.16
0.50	0.31	1.10	0.14

HINT: You may rely on the fact that any linear combination of independent normal random variables is also a normal random variable.

SOLUTION: Let $Y = X_A - X_B$. Then $Y \sim N(\mu_Y, \sigma_Y^2)$, where

$$\begin{aligned} \mu_Y &= \mu_A - \mu_B = 104 - 123 = -19, \\ \sigma_Y^2 &= \sigma_A^2 + \sigma_B^2 = 23^2 + 14^2 = 725. \end{aligned}$$

We then need to calculate

$$\begin{aligned} P(X_A > X_B) &= P(Y > 0) \\ &= P\left(\frac{Y-\mu_Y}{\sigma_Y} > \frac{0-\mu_Y}{\sigma_Y}\right) \\ &= P\left(\frac{Y-\mu_Y}{\sigma_Y} > \frac{19}{\sqrt{725}}\right) \\ &= P\left(Z > 0.7056\right) \approx 0.24 \end{aligned}$$

where $Z \sim N(0,1)$, using the bext approximate value from the table.

Problem 1.50 Let X_1 and X_2 be independent geometric random variables with means $1/p_1$ and $1/p_2$ respectively.

(a) If $Y = \min(X_1, X_2)$, show that Y is a geometric random variable with mean $1/\alpha$, where $\alpha = 1-(1-p_1)(1-p_2)$.
(b) Prove the following equalities:

$$P(X_1 > X_2) = \frac{p_2(1-p_1)}{1-(1-p_1)(1-p_2)}, \tag{1.6}$$

$$P(X_2 > X_1) = \frac{p_1(1-p_2)}{1-(1-p_1)(1-p_2)}, \tag{1.7}$$

$$P(X_1 = X_2) = \frac{p_2 p_1}{1-(1-p_1)(1-p_2)}. \tag{1.8}$$

SOLUTION:

(a) The CDF of $X \sim geom(p)$ is $F_X(k) = 1-(1-p)^k$, and the PMF is $p_X(k) = p(1-p)^{k-1}$, $k \geq 1$. We have

$$\begin{aligned} P(Y > k) &= P(X_1 > k \text{ and } X_2 > k) \\ &= P(X_1 > k)(X_2 > k) \\ &= (1-p_1)^k(1-p_2)^k \\ &= [(1-p_1)(1-p_2)]^k, \end{aligned}$$

that is, $Y \sim geom(1-(1-p_1)(1-p_2))$.
Recall geometric series, for $r < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$$

We have, by independence

$$P(X_1 > X_2 \mid X_2 = k) = P(X_1 > k).$$

By the law of total probability, conditioning on partition $A_k = \{X_2 = k\}$, $k = 1, 2, \ldots$, we have

$$\begin{aligned} P(X_1 > X_2) &= \sum_{k=1} P(X_1 > k) P(X_1 > X_2 \mid X_2 = k) P(X_2 = k) \\ &= \sum_{k=1} (1-p_1)^k p_2 (1-p_2)^{k-1} \\ &= (1-p_1) p_2 \sum_{k=0} [(1-p_1)(1-p_2)]^k \\ &= (1-p_1) p_2 \frac{1}{1-(1-p_1)(1-p_2)}. \end{aligned}$$

Essentially the same argument gives

$$P(X_2 > X_1) = (1-p_2) p_1 \frac{1}{1-(1-p_1)(1-p_2)}.$$

We must then have

$$\begin{aligned} P(X_1 = X_2) &= 1 - P(X_1 > X_2) - P(X_2 > X_1) \\ &= 1 - \frac{(1-p_1)p_2 + (1-p_2)p_1}{1-(1-p_1)(1-p_2)} \\ &= \frac{1-(1-p_1)(1-p_2) - (1-p_1)p_2 - (1-p_2)p_1}{1-(1-p_1)(1-p_2)} \\ &= \frac{p_1 p_2}{1-(1-p_1)(1-p_2)} \end{aligned}$$

Problem 1.51 The distribution of the height of a certain type of plant is normally distributed with mean $\mu = 39.8$ inches and standard deviation $\sigma = 2.05$ inches. What is the probability that of 20 randomly selected plants, exactly 5 have a height of at least 40 inches? Use the fact that $P(Z \leq 0.1) = 0.5398$ for a standard normal random variable $Z \sim N(0, 1)$.

SOLUTION: The height of a single plant is distributed as $X \sim N(39.8, 2.05^2)$. The probability that a single plant has a height of at least 40 inches is

$$\begin{aligned} p &= P(X \geq 40) \\ &= P\left(\frac{X - 39.8}{2.05} \geq \frac{40 - 39.8}{2.05}\right) \\ &\approx P(Z \geq 0.1) \\ &\approx 1 - 0.5398 = 0.4602, \end{aligned}$$

where $Z \sim N(0, 1)$. The number of plants at least 40 inches high has binomial distribution $Y \sim bin(20, p)$, so

$$P(Y = 5) = \binom{20}{5} 0.4602^5 (1 - 0.4602)^{15} \approx 0.03.$$

Problem 1.52 A circle has radius R, circumference C and area A.

(a) If $R \sim exp(1)$ derive the density function for C and A. Which of these has an exponential distribution?
(b) If $R \sim unif(0,1)$ derive the density function for C and A. Which of these has a uniform distribution?

SOLUTION: Either the CDF transformation method or the one-to-one transformation method can be used. In fact, they will be essentially the same method when increasing transformations are used. Below, we use the one-to-one transformation method.

We have $C = 2\pi R$ and $A = \pi R^2$. Then use transformation rule

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y)),$$

for transformation $Y = g(X)$. We have transformations

$$\begin{aligned} C &= g_C(R) = 2\pi R, \\ g_C^{-1}(c) &= c/(2\pi), \\ \frac{dg_C^{-1}(c)}{dc} &= (2\pi)^{-1}, \text{ and if} \\ A &= g_A(R) = \pi R^2, \text{ we have} \\ g_A^{-1}(a) &= \sqrt{a/\pi}, \\ \frac{dg_A^{-1}(a)}{dr} &= 2^{-1}(\pi a)^{-1/2}. \end{aligned}$$

(a) We have support $[0,\infty)$ and density function $f_R(r) = \exp(-r)$ for R. The support of C and A is also $[0,\infty)$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1} \exp\left(-c/(2\pi)\right), \ c \geq 0, \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2} \exp\left(-\sqrt{a/\pi}\right), \ a \geq 0, \end{aligned}$$

with $f_C(c) = 0$ and $f_A(a) = 0$ for $c, a < 0$. Note that $C \sim exp((2\pi)^{-1})$.

(b) We have support $[0,1]$ and density function $f_R(r) = I\{r \in [0,1]\}$ for R. The support of C is now $[0,2\pi]$ and the support of A is now $[0,\pi]$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1} I\{c/(2\pi) \in [0,1]\} = (2\pi)^{-1} I\{c \in [0,2\pi]\} \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2} I\{a \in [0,\pi]\}. \end{aligned}$$

Note that $C \sim unif[0,2\pi]$. These densities are defined on the entire real line, and the support is implicit in the indicator functions. We can also write:

$$f_C(c) = \begin{cases} (2\pi)^{-1} & ; \quad c \in [0,2\pi] \\ 0 & ; \quad ow \end{cases}$$

and

$$f_A(a) = \begin{cases} 2^{-1}(\pi a)^{-1/2} & ; \quad a \in [0,\pi] \\ 0 & ; \quad ow \end{cases}.$$

Problem 1.53 The distribution of the height of a certain type of plant is normally distributed with mean $\mu = 49.2$ inches and standard deviation $\sigma = 1.75$ inches. What is the probability that of 20 randomly selected plants, no more than 3 plants height of no more than 48 inches?

SOLUTION: The height of a single plant is distributed as $X \sim N(49.2, 1.75^2)$. The probability that a single plant has a height of no more than 48 inches is

$$\begin{aligned} p &= P(X \leq 48) \\ &= P\left(\frac{X - 49.2}{1.75} \leq \frac{48 - 49.2}{1.75}\right) \\ &= P\left(Z \leq \frac{48 - 49.2}{1.75}\right) \\ &\approx 0.2464, \end{aligned}$$

where $Z \sim N(0, 1)$. Either use normal probability table entry for -0.69 (giving $P \approx 0.2451$), or R command

```
> pnorm((48 - 49.2)/1.75)
[1] 0.2464466
```

The number of plants of height no more than 48 inches has binomial distribution $Y \sim bin(20, p)$, so

$$P(Y \leq 3) = \sum_{i=0}^{3} \binom{20}{i} p^i (1-p)^{20-i} \approx 0.2361$$

using R commands

```
> pp = pnorm((48 - 49.2)/1.75)
> pbinom(3,20,pp)
[1] 0.2361194
```

Problem 1.54 A certain hospital delivered 10 babies during the last year. Assume that a baby is equally (and independently) likely to be a boy or girl.

(a) What is the probability that 6 of these were boys?
(b) What is the probability that at least 8 of these were boys?
(c) Given that 6 of these were boys, what is the probability that the first six deliveries were all boys?

SOLUTION: Let $X \sim bin(10, 1/2)$ be the number of boys.

(a) Directly from the binomial distribution,

$$P(X = 6) = \binom{10}{6}(1/2)^{10} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210/1024 = 105/512 \approx 0.2051$$

(b) Noting that $p = 1/2$ the expression simplifies to

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= \left[\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right](1/2)^{10} \\ &= [45 + 10 + 1]\,(1/2)^{10} \\ &= 56/1024 = 7/128 \approx 0.05469. \end{aligned}$$

(c) Let $A = \{\text{first 6 deliveries were boys}\}$. Then we need to evaluate

$$\begin{aligned} P(A \mid X = 6) &= \frac{P(A \cap X = 6)}{P(X = 6)} \\ &= \frac{P(\text{first 6 are boys, last 4 are girls})}{\binom{10}{6}(1/2)^{10}} \\ &= \frac{(1/2)^{10}}{\binom{10}{6}(1/2)^{10}} \\ &= \frac{1}{\binom{10}{6}} \\ &= \frac{1 \times 2 \times 3 \times 4}{7 \times 8 \times 9 \times 10} = \frac{1}{210}. \end{aligned}$$

Problem 1.55 Suppose $X_1, \ldots, X_n$ are independent random variables with a common CDF F_X.

(a) Show that the CDF of $Y = \max(X_1, \ldots, X_n)$ is given by $F_Y(t) = F_X^n(t)$.
(b) Suppose $X_1 \sim unif(0, 1)$. Derive the density function and mean of Y.

SOLUTION:

(a) We have, by independence,

$$F_Y(t) = P(Y \leq t) = P\left(\cap_i \{X_i \leq t\}\right) = \prod_i P(\{X_i \leq t\}) = F_X^n(t).$$

(b) For the $unif(0, 1)$ distribution we have

$$F_X(t) = \begin{cases} 0 & ; \quad t < 0 \\ t & ; \quad t \in [0, 1) \\ 1 & ; \quad t \geq 1 \end{cases}$$

so

$$F_Y(t) = \begin{cases} 0 & ; \quad t < 0 \\ t^n & ; \quad t \in [0, 1) \\ 1 & ; \quad t \geq 1 \end{cases}.$$

The density function is the derivative of the CDF, so (apart from $t = 0$ and $t = 1$)

$$f_Y(t) = \frac{dF_Y(t)}{dt} = \begin{cases} nt^{n-1} & ; \quad t \in [0, 1] \\ 0 & ; \quad ow \end{cases}.$$

and we have

$$E[Y] = \int_{t=0}^{1} tnt^{n-1} dt = \int_{t=0}^{1} nt^n dt = \frac{n}{n+1} t^{n+1} \Big|_0^1 = \frac{n}{n+1}.$$

Problem 1.56 A circle has radius R, circumference C and area A.

(a) If $R \sim exp(1)$ derive the density function for C and A. Which of these has an exponential distribution?

(b) If $R \sim unif(0,1)$ derive the density function for C and A. Which of these has a uniform distribution?

Problem 1.57 A bin contains m white and n black balls. A random selection of $k \leq m+n$ balls is made (this is referred to as *sampling without replacement*). Let X be the number of white balls among the k selected. This is known as a *hypergeometric random variable*, which we denote $X \sim hyper(m,n,k)$. Using principles of combinatorics, derive a general expression for the PMF of X. Make sure to state exactly the support $\mathcal{S}_X$ of X. What is $E[X]$?

SOLUTION: We may temporarily label the balls, so that they are all distinct. Then the total number of selections is

$$D = \binom{m+n}{k}.$$

To enumerate the selections for which $X = i$ use the *rule of product*.

(a) Selection combination of i from m white balls, $n_1 = \binom{m}{i}$.
(b) Selection combination of $k-i$ from n black balls, $n_2 = \binom{n}{k-i}$.

There are

$$N = n_1 \times n_2 = \binom{m}{i}\binom{n}{k-i}$$

such combinations. We then have the general expression for the PMF:

$$p_X(i) = P(X=i) = \frac{N}{D} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}.$$

To derive the support $\mathcal{S}_X$ we note that the number of white and black balls selected (i and $k-i$, respectively) must satisfy the following inequalities:

$$\begin{aligned} 0 &\leq i \leq m \\ 0 &\leq k-i \leq n \text{ or } k-n \leq i \leq k, \end{aligned}$$

so that the support is given by

$$\mathcal{S}_X = \{i : \max(0, k-n) \leq i \leq \min(k,m)\}.$$

Next, define Bernoulli random variables $U_1, \ldots, U_k$, where $U_i = 1$ if the ith ball is white, and $U_i = 0$ otherwise. Then

$$X = \sum_{i=1}^{k} U_i.$$

There are m white balls from a total of $m+n$. Therefore, $P(U_i = 1) = p = m/(m+n)$, and $E[U_i] = p$ for $i = 1, \ldots, k$. Therefore

$$E[X] = \sum_{i=1}^{k} E[U_i] = kp = \frac{km}{m+n}.$$

Problem 1.58 The binomial distribution $bin(n,p)$ can be approximated by a Poisson distribution with mean $\lambda = np$. To explore this, suppose we compute the probability $q_\lambda = P(X \leq 8)$, where X is a Poisson random variable with mean $\lambda = 10$. Then we should have

$$q_{n,p} \approx q_\lambda$$

where $q_{n,p} = P(Y \leq 8)$, for $Y \sim bin(n,p)$ with $p = \lambda/n$, provided n is large enough.

To get a sense of how large n should be, using R, construct a plot of $q_{n,p}$ against n, for $n = 10, 11, \ldots, 199, 200$, in each case setting $p = \lambda/n$. Superimpose on the plot a horizontal line at q_λ. Then find the smallest n for which $|q_{n,p} - q_\lambda| \leq 0.01$.

SOLUTION: Fix $\lambda = 10$. The quantity to be approximated is $q_\lambda = P(X \leq 8)$ where $X \sim pois(\lambda = 8)$. This is given by

```
> ppois(8,10)
[1] 0.3328197
```

We then write a function with argument n for sample size. The function calculates $p = \lambda/n = 10/n$, then uses this parameter in the evaluation of $P(X \leq 8)$ where $X \sim bin(n,p)$. This is given by

```
f0 = function(n) {pbinom(8,size=n,prob=10/n)}
```

The plot itself can be constructed using the following code (see Figure 1.2 below):

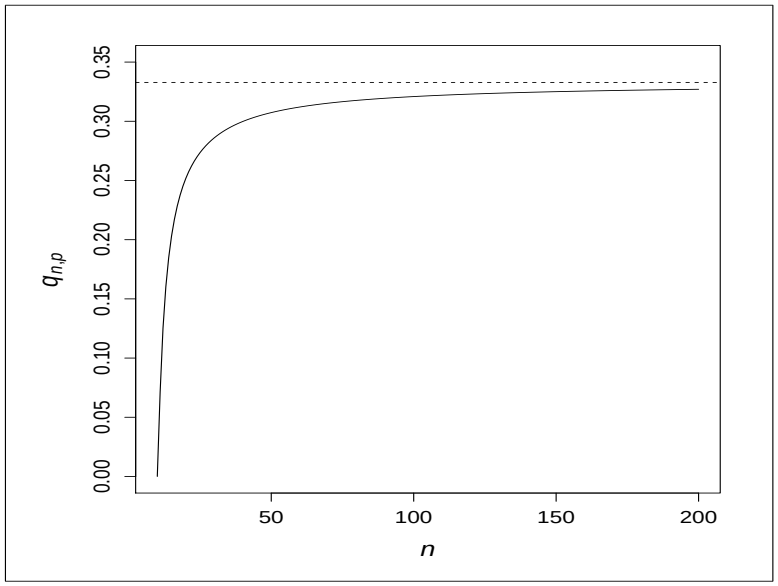

Figure 1.2: Plot for Problem 1.58.

```
ex1 = expression(italic(n))
ex2 = expression(italic(q)[paste(italic(n),',',italic(p),sep='')])
n = seq(10,200,1)
plot(n,f0(n),ylim=c(0,0.35),type='l',xlab=ex1,ylab=ex2)
abline(h = ppois(8,10),lty=2)
```

Note that it is important to specify a large enough y-range to show all required features. This is done with the `ylim` option. There are several ways to find the smallest n for which $|q_{n,p} - q_\lambda| \leq 0.01$ holds. This type of operation can typically be done with a single line of code. In this case, we simply use index subsetting on the original vector of n values used to construct the plot:

```
> min(n[abs(f0(n)-ppois(8,10)) <= 0.01])
[1] 119
```

The correct answer is then $n = 119$.

Problem 1.59 Suppose X_1, X_2 are independent observations from a geometric distribution with mean $1/p$, $p \in (0, 1)$.

(a) Derive the conditional distribution of X_1 conditional on $\{X_1 + X_2 = s\}$ for any $s \geq 2$, in the form of the probability mass function (PMF)

$$p_X(x) = P(X_1 = x \mid X_1 + X_2 = s).$$

(b) How does $p_X(x)$ depend on x and p?

SOLUTION: The PMF for a geometric random variable with parameter p is $p_{X_1}(k) = p(1-p)^{k-1}$, $k \geq 1$. The sum of two *iid* geometric random variables with parameter p is a negative binomial random variable with parameters $r = 2$, p, with PMF

$$p_X(k) = \binom{k-1}{1} p^2(1-p)^{k-2}, \quad k \geq 2.$$

(a) The conditional PMF is then (assuming $s \geq 2$ and $1 \leq k \leq s-1$),

$$\begin{aligned} p_X(k \mid X_1 + X_2 = s) &= \frac{P\left(X_1 = k \text{ and } X_1 + X_2 = s\right)}{P\left(X_1 + X_2 = s\right)} \\ &= \frac{P\left(X_1 = k \text{ and } X_2 = s - k\right)}{P\left(X_1 + X_2 = s\right)} \\ &= \frac{P\left(X_1 = k\right) P\left(X_2 = s - k\right)}{P\left(X_1 + X_2 = s\right)} \\ &= \frac{p(1-p)^{k-1} p(1-p)^{s-k-1}}{\binom{s-1}{1} p^2 (1-p)^{s-2}} \\ &= \frac{p^2(1-p)^{s-2}}{\binom{s-1}{1} p^2 (1-p)^{s-2}} \\ &= \frac{1}{s-1}. \end{aligned}$$

Thus, X_1, conditional on the event $\{X_1 + X_2 = s\}$ is uniformly distributed on the possible outcomes $\{1, \ldots, s-1\}$.

(b) The conditional distribution does not depend either on x or p.

Problem 1.60 Suppose X is a random variable. The *moment generating function* is a function of a real variable t, defined by

$$M_X(t) = E\left[e^{tX}\right]$$

for any fixed t.

(a) Assuming that $M_X(t)$ is finite in some open interval (a, b), where $a < 0$ and $b > 0$, show that the kth moment of X can be calculated by the kth derivative of $M_X(t)$ evaluated at $t = 0$, that is,

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E[X^k], \quad k = 1, 2, \ldots.$$

[HINT: Assume that you may exchange the order of differentiation and integration where needed.]

(b) Show that if X and Y are independent random variables, the moment generating function of $X + Y$ equals the product of the moment generating functions of X and Y, that is,

$$M_{X+Y}(t) = M_X(t)M_Y(t),$$

where they are finite.

(c) Derive the moment generating function for $X \sim bin(n, p)$.

(d) If $X \sim bin(n, p)$ and $Y \sim bin(m, q)$, show that $X + Y$ is a binomial random variable if and only if $p = q$.

SOLUTION:

(a) We have

$$\begin{aligned}\frac{d^k M_X(t)}{dt^k} &= \frac{d^k}{dt^k} E\left[e^{tX}\right] \\ &= E\left[\frac{d^k e^{tX}}{dt^k}\right] \\ &= E\left[X^k e^{tX}\right].\end{aligned}$$

Then substitute $t = 0$ to get $E\left[X^k e^{0 \cdot X}\right] = E\left[X^k\right]$.

(b) Recall from Section 4.8, if random variables X, Y are independent, then for any functions g_1, g_2, we have $E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)]$, assuming the expected values exist. In this case, $g_1(x) = g_2(x) = e^{tx}$. This means

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E\left[e^{tX}e^{tY}\right] = E\left[e^{tX}\right] E\left[e^{tY}\right] = M_X(t)M_Y(t).$$

(c) Since the PMF of $X \sim bin(n, p)$ is $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$, we have

$$\begin{aligned}M_X(t) &= \sum_{x=0}^{n} e^{tx} p(x) \\ &= \sum_{x=0}^{n} \binom{n}{x} e^{tx} p^x (1-p)^{n-x} \\ &= (pe^t + 1 - p)^n,\end{aligned}$$

using the binomial expansion.

(d) From Part (b)-(c), we have

$$M_{X+Y}(t) = (pe^t + 1 - p)^n (qe^t + 1 - q)^m. \tag{1.9}$$

If $X + Y \sim bin(\alpha, s)$ for some α, s, we must have

$$M_{X+Y}(t) = (\alpha e^t + 1 - \alpha)^s. \tag{1.10}$$

Thus, $X + Y$ is binomial if and only if we can find α and s for which expressions (1.9) and (1.10) are equal. First, note that (1.9) is an order $n + m$ polynominal in e^t. Suppose (1.10) is equal to (1.9). Then (1.10) must also be an order $n + m$ polynominal in e^t. This means we must have $s = m + n$. Furthermore, the roots of the polynomials must be equal to each other. Then (1.9) has root $e^t = (p - 1)/p$ of multiplicity n, and root $e^t = (q - 1)/q$ of multiplicity m. But (1.10) has only one root $e^t = (\alpha - 1)/\alpha$ of multiplicity $s = n + m$. Therefore expressions (1.9) and (1.10) are equal if and only if $p = q = \alpha$. Therefore, $X + Y$ is binomial if and only if $p = q$.

Problem 1.61 Suppose X is the sum of n independent Bernoulli random variables $U_1, \ldots, U_n$. Suppose the means are given by $E[U_i] = p_i$. What is the mean and variance of X?

SOLUTION: The variance of U_i is $\sigma_i^2 = p_i(1 - p_i)$. Since the random variables are independent, we have

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} p_i, \text{ and} \\ var(X) &= \sum_{i=1}^{n} p_i(1 - p_i). \end{aligned}$$

Problem 1.62 Assume that over a 25 year period the mean height of adult males increased from 175.5 cm to 179.1 cm, with the standard deviation remaining constant at $\sigma = 5.84$. Suppose the minimum height requirement to join the police force remained unchanged at 172 cm. Assume the heights are normally distributed.

What proportion of adult males would not meet the minimum height requirement at the beginning and end of the 25 year period? [Use the probabilities $P(Z \leq -0.5993) \approx 0.274$, $P(Z \leq -1.2158) \approx 0.112$].

SOLUTION: At the beginning of the period $X \sim N(175.5, 5.84^2)$. The proportion not meeting the height requirement is

$$\begin{aligned} p &= P(X \leq 172) \\ &= P\left(\frac{X - 175.5}{5.84} \leq \frac{172 - 175.5}{5.84}\right) \\ &= P\left(Z \leq -0.5993\right) \\ &\approx 0.274, \end{aligned}$$

where $Z \sim N(0, 1)$.

At the beginning of the period $X \sim N(179.1, 5.84^2)$. The proportion not meeting the height requirement is

$$\begin{aligned} p &= P(X \leq 172) \\ &= P\left(\frac{X - 179.1}{5.84} \leq \frac{172 - 179.1}{5.84}\right) \\ &= P\left(Z \leq -1.2158\right) \\ &\approx 0.112, \end{aligned}$$

where $Z \sim N(0, 1)$.

Problem 1.63 Let U_1 and U_2 be two independent random variables with a $unif[0, 1]$ distribution. If $E \subset S$, where S is a unit square with sides $[0, 1]$, then $P((U_1, U_2) \in E) = area(E)$. This holds whether or not E includes its boundary.

(a) Determine the CDF $F_X(x)$ for $X = U_1 + U_2$.
(b) Determine the density function $f_X(x)$ for X.

SOLUTION:

(a) The support of X is within $[0, 2]$. For $x \in [0, 1]$ the event $E = \{X \leq x\} = \{U_1 + U_2 \leq x\}$ confined to the unit square is a triangle with base x and height x, so

$$F_X(x) = P(X \leq x) = x^2/2, \;\; x \in [0, 1].$$

For $x > 1$ we note that the event $E = \{X > x\} = \{U_1 + U_2 > x\}$ confined to the unit square is a triangle with vertices $(1, x - 1), (x - 1, 1)$ and $(1, 1)$, which has base $2 - x$ and height $2 - x$, so

$$F_X(x) = P(X \leq x) = 1 - P(X > x) = 1 - (2 - x)^2/2, \;\; x \in [1, 2].$$

So the CDF is

$$F_X(x) = \begin{cases} 0 & ; \;\; x < 0 \\ x^2/2 & ; \;\; x \in [0, 1) \\ 1 - (2 - x)^2/2 & ; \;\; x \in [1, 2) \\ 1 & ; \;\; x > 2 \end{cases}.$$

(b) The density is obtained by differentiating F_X:

$$f_X(x) = \begin{cases} x & ; \;\; x \in [0, 1) \\ 2 - x & ; \;\; x \in [1, 2) \\ 0 & ; \;\; otherwise \end{cases}.$$

Problem 1.64 A random variable W has a *Weibull distribution* if there are two parameters $k > 0$ and $\lambda > 0$ such that

$$X = (\lambda W)^k \tag{1.11}$$

has an $exp(1)$ distribution (exponential distrubution with $\lambda = 1$). This will be denoted $W \sim weibull(k, \lambda)$. This distribution is commonly used to model survival times. By convention, k is the *shape parameter*

and θ is the *rate parameter*. Note that in some conventions λ is replaced by, say, $1/\tau$, in which case τ is referred to as a *scale parameter* (be careful, since λ may be used as a scale parameter). Both definitions are equivalent, once the transformation is understood. We use the rate parameter in order to emphasize the relationship with the exponential distribution.

Suppose $W \sim weibull(k, \lambda)$.

(a) Determine the CDF for W. This can be done by using (1.11) and the CDF of $X \sim exp(1)$.
(b) Determine the density of W.
(c) Derive an expression for the ith moment $E[W^i]$ expressed in terms of the gamma function Γ (Appendix B.7). Then determine the mean and variance.

SOLUTION:

(a) The CDF of $X \sim exp(1)$ is $F_X(x) = 1 - \exp(-x)$ for $x \geq 0$, so

$$F_W(w) = P(W \leq w) = P\left(\lambda^{-1} X^{1/k} \leq w\right) = P\left(X \leq (\lambda w)^k\right) = 1 - \exp\left(-(\lambda w)^k\right), \quad w \geq 0,$$

and $F_W(w) = 0$ for $w < 0$.

(b) Take the deriviate of the CDF (or use Theorem 4.3)

$$f_W(w) = \frac{d}{dw}\left\{1 - \exp\left(-(\lambda w)^k\right)\right\} = k\lambda^k w^{k-1} \exp\left(-(\lambda w)^k\right), \quad w \geq 0,$$

and $f_W(w) = 0$ for $w < 0$.

(c) The ith moment is evaluated by

$$E[W^i] = \int_0^\infty k\lambda^k w^{i+k-1} \exp\left(-(\lambda w)^k\right) dw.$$

Use change of variable $u = (\lambda w)^k$, noting that $du/dw = k\lambda^k w^{k-1}$, and the definition of gamma function Γ in Appendix B.7,

$$\begin{aligned} E[W^i] &= \int_0^\infty \frac{k\lambda^k w^{i+k-1}}{k\lambda^k w^{k-1}} \exp(-u)\, du. \\ &= \int_0^\infty w^i \exp(-u)\, du \\ &= \int_0^\infty \lambda^{-i} u^{i/k} \exp(-u)\, du \\ &= \lambda^{-i}\Gamma(1 + i/k). \end{aligned}$$

Note that this expression can also be derived by evaluating the expected value of $W^i = \lambda^{-i} X^{i/k}$ for $X \sim exp(1)$, as discussed in Section 4.6.3. This gives

$$\begin{aligned} E[W] &= \lambda^{-1}\Gamma(1 + 1/k), \\ \text{var}[W] &= E[W^2] - E[W]^2 = \lambda^{-2}\left[\Gamma(1 + 2/k) - \Gamma(1 + 1/k)^2\right]. \end{aligned}$$

Problem 1.65 Let X_1 and X_2 be independent geometric random variables with means $1/p_1$ and $1/p_2$ respectively.

(a) If $Y = \min(X_1, X_2)$, show that Y is a geometric random variable with mean $1/\alpha$, where $\alpha = 1 - (1 - p_1)(1 - p_2)$.
(b) Prove the following equality:

$$P(X_1 > X_2) = \frac{(1 - p_1)p_2}{1 - (1 - p_1)(1 - p_2)}.$$

SOLUTION:

(a) The CDF of $X \sim geom(p)$ is $F_X(k) = 1 - (1 - p)^k$, and the PMF is $p_X(k) = p(1 - p)^{k-1}$, $k \geq 1$. We have

$$\begin{aligned} P(Y > k) &= P(X_1 > k \text{ and } X_2 > k) \\ &= P(X_1 > k)(X_2 > k) \\ &= (1 - p_1)^k(1 - p_2)^k \\ &= [(1 - p_1)(1 - p_2)]^k, \end{aligned}$$

that is, $Y \sim geom(1 - (1 - p_1)(1 - p_2))$.
(b) Recall geometric series, for $r < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1 - r}$$

We have, by independence

$$P(X_1 > X_2 \mid X_2 = k) = P(X_1 > k).$$

By the law of total probability, conditioning on partition $A_k = \{X_2 = k\}$, $k = 1, 2, \ldots$, we have

$$\begin{aligned} P(X_1 > X_2) &= \sum_{k=1}^{\infty} P(X_1 > X_2 \mid X_2 = k)P(X_2 = k) \\ &= \sum_{k=1}^{\infty} P(X_1 > k)P(X_2 = k) \\ &= \sum_{k=1}^{\infty} (1 - p_1)^k p_2 (1 - p_2)^{k-1} \\ &= (1 - p_1)p_2 \sum_{k=0}^{\infty} [(1 - p_1)(1 - p_2)]^k \\ &= \frac{(1 - p_1)p_2}{1 - (1 - p_1)(1 - p_2)}. \end{aligned}$$

1.8 Applications of Random Variables

Problem 1.66 In *Natural Inheritance* by Francis Galton, published in 1889, the paired heights of parents with their adult children were reported. The heights, in inches, of 928 children are summarized in the following table. Essentially, we have a histogram. For example, 165 of the 928 adult children have heights

in the class interval $(64.7, 66.7]$.

We are interested in determining whether or not a normal distribution would be appropriate for modeling these heights. We can extract from the table estimates of mean and standard deviation $\mu \approx 68.1$ and $\sigma \approx 2.60$ (by assuming that each datum is represented by the midpoint of its class interval). In principle, we could use the empirical rule to assess the normality of the data, except for the fact that we could not expect the quantities $\mu \pm K\sigma$, $K = 1, 2, 3$, to land exactly on the endpoints, which we would need in order to obtain the relevant empirical frequencies.

Of course, we can use other quantities to achieve the same goal. For example, we can obtain directly from the table estimates of the CDF $P(X \leq x)$ for each endpoint $x = 60.7, 62.7, \ldots, 74.7$ and compare them directly to the values predicted by the normal distribution, that is, $P(Y \leq x)$ where $Y \sim N(\mu = 68.1, \sigma^2 = 2.60^2)$. Try this, by filling in the table. See Section 4.4.4 of *Biostatistics: A Methodology for the Health Sciences* [L.D. Fisher & G. van Belle] for more on this problem.

Class Interval	Freq.	Cumulative Freq.	Estimated CDF	CDF Predicted by Normal Distribution
(58.7, 60.7]	0	-	-	-
(60.7, 62.7]	12	-	-	-
(62.7, 64.7]	91	-	-	-
(64.7, 66.7]	165	-	-	-
(66.7, 68.7]	258	-	-	-
(68.7, 70.7]	266	-	-	-
(70.7, 72.7]	105	-	-	-
(72.7, 74.7]	31	-	-	-

SOLUTION: The requires numbers can be calculated with the following R script:

```
> x0 = c(58.7, 60.7,62.7,64.7,66.7,68.7,70.7,72.7)
> x = c(60.7,62.7,64.7,66.7,68.7,70.7,72.7,74.7)
> c1 = paste('(',x0,',',x,']', sep='')
> y = c(0,12,91,165,258,266,105,31)
> tab = data.frame(c1,y,cumsum(y),round(cumsum(y)/sum(y),3),
round(pnorm(x, mean=68.1, sd=2.6),3))
> names(tab) = paste('column',1:5)
> tab
column 1 column 2 column 3 column 4 column 5
1 (58.7,60.7]        0        0    0.000    0.002
2 (60.7,62.7]       12       12    0.013    0.019
3 (62.7,64.7]       91      103    0.111    0.095
4 (64.7,66.7]      165      268    0.289    0.295
5 (66.7,68.7]      258      526    0.567    0.591
6 (68.7,70.7]      266      792    0.853    0.841
7 (70.7,72.7]      105      897    0.967    0.962
8 (72.7,74.7]       31      928    1.000    0.994
>
```

This gives:

Class Interval	Freq.	Cumulative Freq.	Estimated CDF	CDF Predicted by Normal Distribution
(58.7,60.7]	0	0	0.00	0.00
(60.7,62.7]	12	12	0.01	0.02
(62.7,64.7]	91	103	0.11	0.10
(64.7,66.7]	165	268	0.29	0.29
(66.7,68.7]	258	526	0.57	0.59
(68.7,70.7]	266	792	0.85	0.84
(70.7,72.7]	105	897	0.97	0.96
(72.7,74.7]	31	928	1.00	0.99

The estimated cumulative frequencies are reasonably close to the frequencies predicted by the normal distribution.

Problem 1.67 Assume that over a 25 year period the mean height of adult males increased from 175.5 cm to 179.1 cm, with the standard deviation remaining constant at $\sigma = 5.84$. Suppose the minimum height requirement to join the police force remained unchanged at 172 cm. Assume the heights are normally distributed.

(a) What proportion of adult males would not meet the minimum height requirement at the beginning and end of the 25 year period?
(b) To what value should the minimum height be changed after 25 years in order to maintain the same proportion which meet the height requirement?
(c) What proportion of adult males would not have met this updated requirement 25 years ago?
(d) Repeat the first three questions using the same values, except that we'll assume that the standard deviation has increased from 5.84 to 10.0 over the 25 year period.

SOLUTION: Set $\mu_1 = 175.5$, $\mu_2 = 179.1$, $\sigma = 5.84$, $\sigma_{new} = 10.0$. Then let $X_1 \sim N(\mu_1, \sigma^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, $X_3 \sim N(\mu_3, \sigma^2_{new})$. Also set $Z \sim N(0, 1)$.

(a) We have

$$p_1 = P(X_1 \leq 172) = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{172 - \mu_1}{\sigma}\right) = P(Z \leq -0.5993) \approx 0.274,$$

$$p_2 = P(X_2 \leq 172) = P\left(\frac{X_2 - \mu_2}{\sigma} \leq \frac{172 - \mu_2}{\sigma}\right) = P(Z \leq -1.2158) \approx 0.112,$$

so that the respective proportions are p_1 and p_2.

(b) Let q_1 be the minimum height required after 25 years. We need the p_1 quantile. For a standard normal distribution this is

$$p_1 = P(Z \leq Z_{p_1})$$

which is solved by $Z_{p_1} = -0.5993$. Then the p_1 quantile for the $N(\mu_2, \sigma^2)$ distribution is

$$q_1 = X_{p_1} = \mu_2 + Z_{p_1}\sigma \approx 179.1 + (-0.5993) \times 5.84 = 175.6.$$

The new minimum height should be $q_1 = 175.6$.

(c) Let p_3 be the proportion not meeting the new minimum requirement q_1 25 years ago. Then

$$p_3 = P(X_1 \leq q_1) = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{175.6 - \mu_1}{\sigma}\right) = P(Z \leq 0.0171) \approx 0.507.$$

(d) The quantities affected by the change from $\sigma = 5.85$ to 10.0 are p_2, q_1 and p_3. The new values are

$$p_2' = P(X_3 \leq 172) = P\left(\frac{X_3 - \mu_2}{\sigma_{new}} \leq \frac{172 - \mu_2}{\sigma_{new}}\right) = P(Z \leq -0.71) \approx 0.239,$$

$$q_1' = X_{p_1} = \mu_2 + Z_{p_1}\sigma_{new} \approx 179.1 + (-0.5993) \times 10.0 = 173.11.$$

$$p_3' = P(X_1 \leq q_1') = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{173.11 - \mu_1}{\sigma}\right) = P(Z \leq -0.4092) \approx 0.341.$$

The following R script can be used to calculate the answers:

```
> ## (a)
>
> p1 = pnorm(172, mean=175.5, sd=5.84)
> p2 = pnorm(172, mean=179.1, sd=5.84)
> p1
[1] 0.2744814
> p2
[1] 0.1120394
>
> ## (b)
>
> q1 = qnorm(p1, mean=179.1, sd=5.84)
> q1
[1] 175.6
>
> ## (c)
>
> p3 = pnorm(q1, mean=175.5, sd=5.84)
> p3
[1] 0.5068309
>
> ####### (d)
>
> ## (a)
>
> pp1 = pnorm(172, mean=175.5, sd=5.84)
> pp2 = pnorm(172, mean=179.1, sd=10.0)
> pp1
[1] 0.2744814
> pp2
[1] 0.2388521
```

```
>
> ## (b)
>
> qq1 = qnorm(pp1, mean=179.1, sd=10.0)
> qq1
[1] 173.1068
>
> ## (c)
>
> pp3 = pnorm(qq1, mean=175.5, sd=5.84)
> pp3
[1] 0.3409814
>
```

Problem 1.68 (The Capture-Recapture Method) A bin contains m white and n black balls. A random selection of $k \leq m+n$ balls is made (this is referred to as *sampling without replacement*). Let X be the number of white balls among the k selected. This is known as a *hypergeometric random variable*, which we denote $X \sim hyper(m, n, k)$.

(a) Using principles of combinatorics, derive a general expression for the PMF of X. Make sure to state exactly the support $\mathcal{S}_X$ of X.

(b) Define a sequence of Bernoulli random variables $U_1, \ldots, U_k$, setting $U_i = 1$ if the ith selected ball is white. Expressing X as their sum, determine the mean and variance of X.

(c) Suppose we make a selection of k balls in the same manner, except that the balls are replaced immediately after being selected, and may be selected again (this is referred to as *sampling with replacement*). Let Y be the total number of white balls selected. What distribution does Y have? Show that $E[X] = E[Y]$ and determine the ratio $var[X]/var[Y]$. Verify that $var[X] \leq var[Y]$ for $k \geq 1$ and $var[X] < var[Y]$ for $k > 1$.

(d) Based on the comparisons of part (c), under what conditions can the distribution of Y be used to approximate the distribution of X?

(e) A lake contains N fish. Suppose J fish are caught, tagged, then released. After a period of time, K fish are caught (assume these K fish are distinct). Suppose X of these have been previously tagged. Assuming both samples are random samples, we have

$$X \sim hyper(J, N-J, K).$$

Derive an expression using X, J, K, denoted $\hat{N}$, which can used as an estimate of total population size N. Use $E[X]$ as a guide. This method is known as *mark and recapture*, and is commonly used to estimate population sizes.

(f) Since X is random, we would like to know how close $\hat{N}$ is to N. Once way to do this is to use a *confidence set* CS of *confidence level* $1-\alpha$. Set x_{obs} to be the observed value of X. Then let N^* be a possible value of N. Then

$$N^* \in CS \quad \text{if and only if} \quad P(Y \leq x_{obs}) > \alpha/2 \ \text{ and } \ P(Y \geq x_{obs}) > \alpha/2, \tag{1.12}$$

where

$$Y \sim hyper(J, N^* - J, K).$$

It may be shown that CS will consist of all integers between some lower and upper bounds, that is,

$$CS = \{N : N_L \leq N \leq N_U\}$$

for some N_L, N_U. Then

$$P(N \in CS) \geq 1 - \alpha,$$

so that the confidence set contains the true value of N with a probability of at least $1 - \alpha$. Write an `R` function which accepts (X, J, K, α) as input, and outputs the bounds N_L, N_U for a confidence set CS for N of confidence level $1 - \alpha$. To do this, set N_U, N_L to be the maximum and minimum values of N^*, respectively, that satisfy condition (1.12). If you use a search algorithm, a good starting point would be $\hat{N}$ (rounded off to the nearest integer), which would be within the bounds N_L, N_U. Make use of `R` function `phyper()`.

(g) Use your function to determine a confidence set for N when $X = 13$, $J = 200$, $K = 100$, $\alpha = 0.05$.

SOLUTION:

(a) We may temporarily label the balls, so that they are all distinct. Then the total number of selections is

$$D = \binom{m+n}{k}.$$

To enumerate the selections for which $X = i$ use the *rule of product*.

(a) Selection combination of i from m white balls, $n_1 = \binom{m}{i}$.
(b) Selection combination of $k - i$ from n black balls, $n_2 = \binom{n}{k-i}$.

There are

$$N = n_1 \times n_2 = \binom{m}{i}\binom{n}{k-i}$$

such combinations. We then have the general expression for the PMF:

$$p_X(i) = P(X = i) = \frac{N}{D} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}.$$

To derive the support $\mathcal{S}_X$ we note that the number of white and black balls selected (i and $k - i$, respectively) must satisfy the following inequalities:

$$\begin{aligned} 0 &\leq i \leq m \\ 0 &\leq k - i \leq n \text{ or } k - n \leq i \leq k, \end{aligned}$$

so that the support is given by

$$\mathcal{S}_X = \{i : \max(0, k - n) \leq i \leq \min(k, m)\}.$$

(b) There are m white balls from a total of $m+n$. Therefore, $P(U_i = 1) = p = m/(m+n)$, and $E[U_i] = p$ for $i = 1, \ldots, k$. Therefore

$$E[X] = \sum_{i=1}^{k} E[U_i] = kp = \frac{km}{m+n}.$$

Since $U_i \sim bern(p)$, we have variance

$$var[U_i] = \sigma_i^2 = p(1-p) = \frac{m}{m+n}\left[1 - \frac{m}{m+n}\right] = \frac{mn}{(m+n)^2}.$$

However, the random variables U_i are not independent. Note that for any $i \neq j$ the product U_iU_j is also a Bernoulli random variable, with $U_iU_j = 1$ if and only if the ith and jth ball are both white. If we condition on the event $U_j = 1$, we effectively remove a white ball before making the next selection. Therefore,

$$P(U_i = 1 \mid U_j = 1) = \frac{m-1}{m+n-1}$$

so that

$$E[U_iU_j] = P(U_i = 1, U_j = 1) = P(U_i = 1 \mid U_j = 1)P(U_j = 1) = \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right)$$

and

$$\begin{aligned} cov[U_iU_j] &= E[U_iU_j] - E[U_i]E[U_j] \\ &= \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right) - \left(\frac{m}{m+n}\right)^2 \\ &= \frac{m}{m+n}\left[\frac{m-1}{m+n-1} - \frac{m}{m+n}\right] \\ &= -\frac{mn}{(m+n)^2(m+n-1)}. \end{aligned}$$

We use the expression

$$\begin{aligned} var[X] &= \sum_i \sigma_i^2 + 2\sum_{i<j} \sigma_{ij} \\ &= \frac{kmn}{(m+n)^2} - \frac{k(k-1)mn}{(m+n)^2(m+n-1)} \\ &= \frac{kmn}{(m+n)^2}\left(\frac{m+n-k}{m+n-1}\right). \end{aligned}$$

(c) Selections are now independent, with constant probability $p = m/(m+n)$ of selecting a white ball for each draw. Therefore $Y \sim bin(k, m/(m+n))$. Then

$$E[Y] = \frac{km}{m+n} = E[X],$$

and

$$var[Y] = kp(1-p) = \frac{kmn}{(m+n)^2}.$$

This means

$$\frac{var[X]}{var[Y]} = \frac{m+n-k}{m+n-1},$$

and the inequalities follow directly.

(d) The means of X and Y are equal for all parameters m, n, k. Otherwise, based on the ratio $var[X]/var[Y]$ the variances are approximately equal if k is small compared to $m + n$. In this case, the binomial distribution may be used to approximate the hypergeometric distribution.

(e) We have J 'white balls', $N - J$ 'black balls', and a selection without replacement of size K. This means

$$X \sim hyper(J, N - J, K).$$

We have

$$E[X] = \frac{KJ}{N},$$

and so

$$N \approx \frac{KJ}{X}.$$

(f) The condition (1.12) can be expressed in terms of the CDF F_Y:

$$P(Y \leq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs}) > \alpha/2$$

and

$$P(Y \geq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs} - 1) < 1 - \alpha/2.$$

Then the following function produces the confidence set:

```
cs.hyper = function(x,j,k,alpha) {

# To find NL:
# start at rounded estimate, decrement n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x,j,n-j,k) > alpha/2) {n = n-1}
NL = n+1

# To find NU:
# start at rounded estimate, increment n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x-1,j,n-j,k) < (1-alpha/2)) {n = n+1}
NU = n-1

cs = c(NL=NL,NU=NU)
return(cs)
}
```

(g) We have output:

```
> j = 200
> k = 100
```

```
> alpha = 0.05
> x = 13
> cs.hyper(x,j,k,alpha)
NL   NU
963 2778
```

Problem 1.69 Candidates for a position are first screened using an aptitude test. The scores are known to have a normal distribution with mean $\mu = 500$ and standard deviation $\sigma = 75$. Candidates who score at least $x = 625$ qualify for an interview.

(a) What is the probability that if 35 candidates take the aptitude test, at least 3 qualify for an interview?
(b) If we want the probability that at least 3 candidates qualify for an interview to be 80%, what should the cutoff score x be.

HINT: This problem can be solved by evaluating the root of an equation numerically using the `uniroot()` function. This calculates numerically the root u of a function f, that is, the solution to $f(u) = 0$. For example, suppose we wanted to find x which satisfies

$$5 = xe^x.$$

A quick sketch shows that the solution is in the interval $x \in [0, 2]$. Then we can use the following code to find that the solution is approximately $x \approx 1.326733$:

```
> f0 = function(x) {x*exp(x)-5}
> uniroot(f0,c(0,2))
$root
[1] 1.326733

$f.root
[1] 6.984409e-05

$iter
[1] 5

$init.it
[1] NA

$estim.prec
[1] 6.103516e-05

> 1.326733*exp(1.326733)
[1] 5.000073
>
```

SOLUTION:

(a) A test score has distribution $X \sim N(500, 75^2)$. Then evaluate

$$\begin{aligned} p &= P(X > 625) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{625-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{625-500}{75}\right) \\ &= P\left(Z \leq 1.667\right) \\ &\approx 1 - 0.9522096 \\ &\approx 0.0478. \end{aligned}$$

where $Z \sim N(0,1)$. Either use the normal probability table entry for 1.67 (giving $p \approx 1 - 0.9525$), or the `R` command `pnorm()` in one of several forms, which will give a more accurate answer:

```
> pnorm(625,mean=500,sd=75)
[1] 0.9522096
> pnorm((625-500)/75)
[1] 0.9522096
>
```

Then let Y be the number of candidates qualifying for an interview. The distribution is $Y \sim bin(35, p)$. Using `R` function `pbinom()` we have

```
> pp = 1-pnorm((625-500)/75)
> 1-pbinom(2,size=35,prob=pp)
[1] 0.2333842
> pbinom(2,size=35,prob=pp,lower.tail=F)
[1] 0.2333842
>
```

so that the required probability is $P(Y \geq 3) \approx 0.2334$. Note the alternative form which uses the option `lower.tail=F`. This evaluates, in general, $P(X > x)$ in place of $P(X \leq x)$.

(b) Here, we can interpret $h(p) = P(Y \geq 3)$ as a function of p, where $Y \sim bin(35, p)$ (it will actually be a quite complicated polynomial). We want to solve $h(p) = 0.8$. We can use the following code:

```
> f0 = function(pp) {pbinom(2,size=35,prob=pp,lower.tail=F) - 0.8}
> uniroot(f0,c(0,1))
$root
[1] 0.1183283

$f.root
[1] 2.792853e-05

$iter
[1] 9

$init.it
```

```
[1] NA

$estim.prec
[1] 6.103516e-05

>
```

So we need to find the 1-0.1183283 = 0.8816717 quantile of the $N(500, 75^2)$ density. We can use the standard R normal quantile function:

```
> qnorm(0.8816717,mean=500,sd=75)
[1] 588.7539
```

so the new cutoff score should be $x \approx 588.75$. If we look up 0.8817 in the standard normal tables, we'd get quantile $z \approx 1.185$, by interpolation. Converting to $N(500, 75^2)$, this would give quantile $\mu + \sigma z = 500 + 75 \times 1.185 = 588.875$. This is close to the previous answer, but would not be as accurate.

Problem 1.70 The table below lists each state's name and population as of June 2014.

Alabama	4849377	Montana	1023579
Alaska	736732	Nebraska	1881503
Arizona	6731484	Nevada	2839099
Arkansas	2966369	New Hampshire	1326813
California	38802500	New Jersey	8938175
Colorado	5355866	New Mexico	2085572
Connecticut	3596677	New York	19746227
Delaware	935614	North Carolina	9943964
Florida	19893297	North Dakota	739482
Georgia	10097343	Ohio	11594163
Hawaii	1419561	Oklahoma	3878051
Idaho	1634464	Oregon	3970239
Indiana	6596855	Pennsylvania	12787209
Iowa	3107126	Rhode Island	1055173
Kansas	2904021	South Carolina	4832482
Kentucky	4413457	South Dakota	853175
llinois	12880580	Tennessee	6549352
Louisiana	4649676	Texas	26956958
Maine	1330089	Utah	2942902
Maryland	5976407	Vermont	626562
Massachusetts	6745408	Virginia	8326289
Michigan	9909877	Washington	7061530
Minnesota	5457173	West Virginia	1850326
Mississippi	2994079	Wisconsin	5757564
Missouri	6063589	Wyoming	584153

(a) Create an R object which stores this data.

(b) For each state, calculate the population proportion.
(c) Rank the proportions in decreasing order, then construct a *log-log* plot to example the powe- law properties of the distribution.
(d) Using the `lines()` function, superimpose on this plot the lines

$$f(k) = \log(p_X(1)) - \alpha \log(k),$$

for $\alpha = 1/3, 2/3, 1$, where k represents the rank of a frequency ($k = 1$ is the largest frequency), and $-\alpha$ is an exponent of a power law. Use the `legend()` function to label the superimposed lines. To do this, select distinct `lty` parameters for the 3 superimposed lines, then use these values in the `legend()` function.
(e) Do state population sizes conform to a power law? Note that the power law may hold for the largest frequencies, but not the smallest.

SOLUTION: The following `R` script produces Figure 1.3 below:

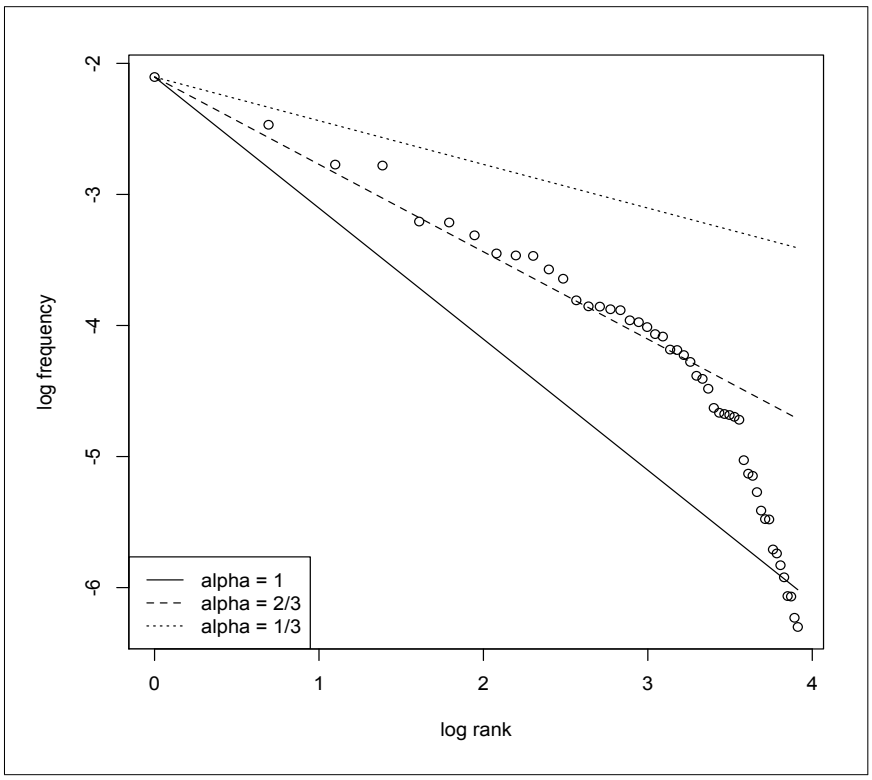

Figure 1.3: Plot for Problem 1.70.

```
> tab = { list containing vector of state names and vector of state populations }
> names(tab) = c("state","pop")
>
> freq = tab$pop/sum(tab$pop)
>
```

```
> y = log(rev(sort(freq)))
>
> par(mfrow=c(1,1),cex=1)
> n = 50
> plot(log(1:n),y,xlab = 'log rank', ylab = 'log frequency')
> alpha = 1
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=1)
> alpha = 2/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=2)
> alpha = 1/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=3)
> legend('bottomleft',legend=paste('alpha = ',c('1','2/3','1/3'),sep=''),lty=1:3)
```

The highest ranking frequencies appear to conform to a power law $p_X(k) \propto 1/k^{2/3}$.

Problem 1.71 The table below lists each state's name and population as of June 2014. In addition, a third column includes categories 1,2,3,4 for regions NE, SE, SW, NW. California is in SW, Maine is in NE, and so on.

Alabama	4849377	2	Montana	1023579	4
Alaska	736732	4	Nebraska	1881503	4
Arizona	6731484	3	Nevada	2839099	3
Arkansas	2966369	2	New Hampshire	1326813	1
California	38802500	3	New Jersey	8938175	1
Colorado	5355866	3	New Mexico	2085572	3
Connecticut	3596677	1	New York	19746227	1
Delaware	935614	1	North Carolina	9943964	2
Florida	19893297	2	North Dakota	739482	4
Georgia	10097343	2	Ohio	11594163	1
Hawaii	1419561	3	Oklahoma	3878051	2
Idaho	1634464	4	Oregon	3970239	4
Indiana	6596855	1	Pennsylvania	12787209	1
Iowa	3107126	4	Rhode Island	1055173	1
Kansas	2904021	2	South Carolina	4832482	2
Kentucky	4413457	2	South Dakota	853175	4
llinois	12880580	1	Tennessee	6549352	2
Louisiana	4649676	2	Texas	26956958	2
Maine	1330089	1	Utah	2942902	4
Maryland	5976407	1	Vermont	626562	1
Massachusetts	6745408	1	Virginia	8326289	2
Michigan	9909877	1	Washington	7061530	4
Minnesota	5457173	4	West Virginia	1850326	2
Mississippi	2994079	2	Wisconsin	5757564	1
Missouri	6063589	2	Wyoming	584153	4

(a) Create a list in `R`, say `tab`, consisting of state name, state population and state region vectors.

(b) Create a vector, say `state.size`, containing the population numbers (column 2 of `tab`). Create a graphics window of 3×2 plots using the `par()` function. Create a histogram of the original data and of the log-transformed (base 10) data. Do the same for boxplots and normal quantile plots. Make sure each plot is identified in the title. Does the log transform appear to 'normalize' the data?

(c) Apply the empirical rule to the original data and to the log-transformed data. Create a 3×3 table. A column should contain proportions of data 1, 2 and 3 standard deviations from the mean. There should be a column for the original data, the log-transformed data and the theoretical probabilities. What does the table suggest regarding the ability of the log-transformation to 'normalize' the data?

(d) Construct side-by-side boxplots of the data categorized by the four regions. Do this for both the original data and the log-transformed data. Create a graphics window of 1×2 plots using the `par()` function. Use the `name` option to label the distributions NE, SE, SW, NW rather than 1,2,3,4. Make sure each plot is identified in the title. Identify the region in each boxplot with the smallest IQR. Why would they differ?

SOLUTION:

(a) We will asssume the required object `tab` has been created.

(b) The following script produces the required plot (Figure 1.4):

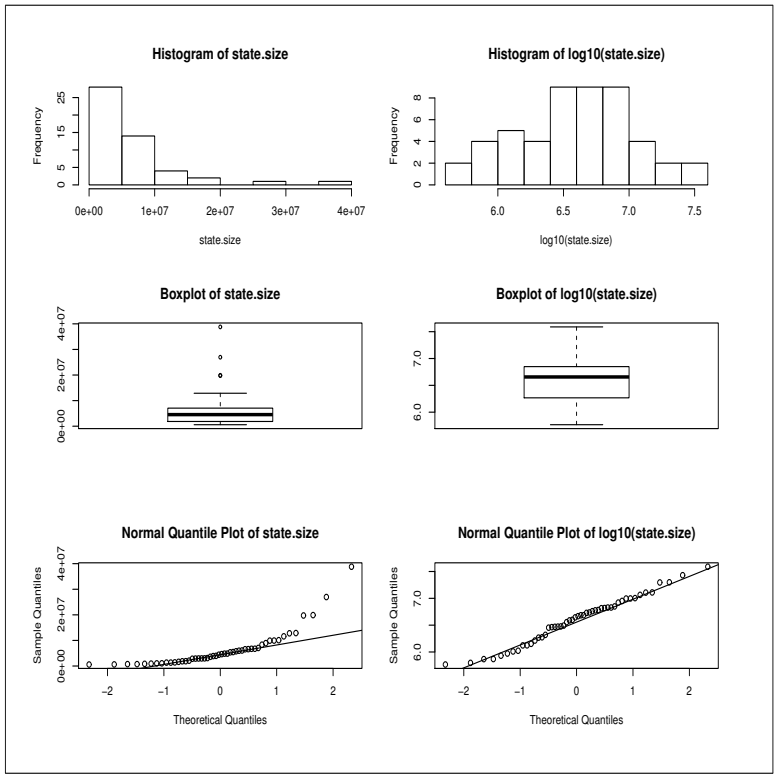

Figure 1.4: Plot for Problem 1.71 (b).

```
par(mfrow=c(3,2))
hist(state.size)
```

```
hist(log10(state.size))
boxplot(state.size, main="Boxplot of state.size")
boxplot(log10(state.size), main="Boxplot of log10(state.size)")
qqnorm(state.size, main="Normal Quantile Plot of state.size")
qqline(state.size)
qqnorm(log10(state.size), main="Normal Quantile Plot of log10(state.size)")
qqline(log10(state.size))
```

Both the histograms and boxplots of the original data show considerable skewness, which is absent after the log-transformation. The quantile plot for the log-transformed data is very close to a straight line, in contrast to that of the original data. Overall, the log-transformation has succeeded in 'normalizing' the data.

(c) The following script produces the required table:

```
er.table = matrix(NA, 3,3)
theoretical.values = c(.68, .95, .997)
sd = sd(state.size)
mn = mean(state.size)
z = (state.size-mn)/sd
er.table[,1] = c(mean(abs(z) <= 1), mean(abs(z) <= 2), mean(abs(z) <= 3))
sd = sd(log10(state.size))
mn = mean(log10(state.size))
z = (log10(state.size)-mn)/sd
er.table[,2] = c(mean(abs(z) <= 1), mean(abs(z) <= 2), mean(abs(z) <= 3))
er.table[,3] = theoretical.values
```

The table itself is:

```
> er.table
[,1] [,2]  [,3]
[1,] 0.92 0.66 0.680
[2,] 0.96 0.98 0.950
[3,] 0.98 1.00 0.997
```

The frequencies of the log-transformed data for 1 and 3 standard deviations are considerably closer to the theoretical values than for the original data. The frequency for 2 standard deviations is closer for the original data, but both are comparable. Overall, the frequencies for the log-transformed data conform to those of a normal distribution much more closely.

(d) The following script produces the required plot (Figure 1.5):

```
par(mfrow=c(1,2))
boxplot(state.size ~ region, names=c('NE', 'SE', 'SW', 'NW'),
main="Boxplots of state.size by Region")
boxplot(log10(state.size) ~ region, names=c('NE', 'SE', 'SW', 'NW'),
main="Boxplots of log10(state.size) by Region")
```

For the original data the NW region has the smallest IQR, while for the log-transformed data the SE region has the smallest IQR. The log function is concave, and therefore reduces the size of increments at higher values by a larger factor relative to smaller values.

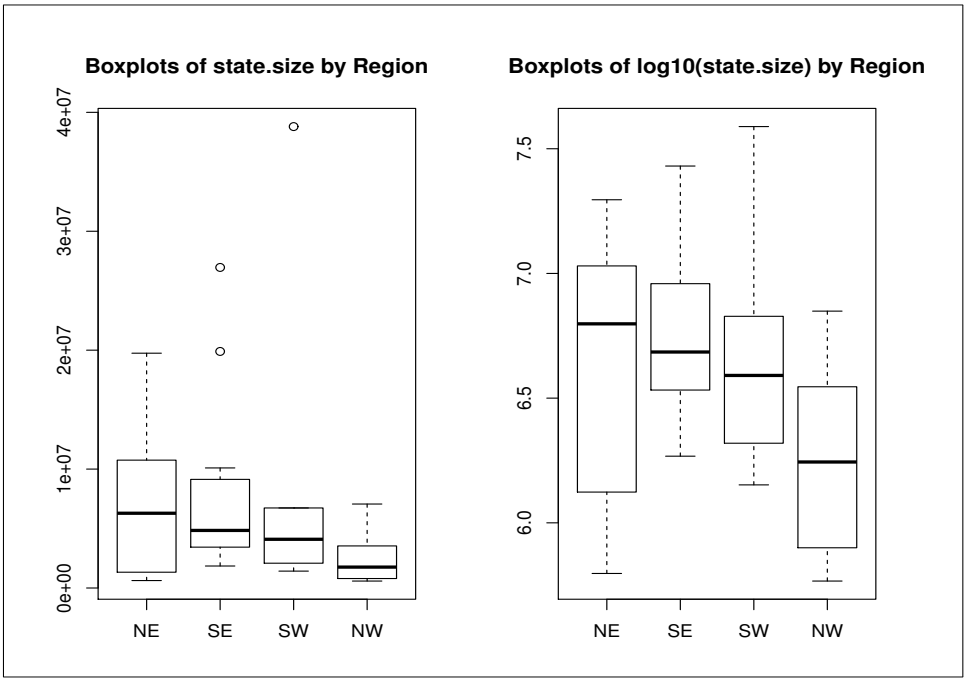

Figure 1.5: Plot for Problem 1.71 (d).

Problem 1.72 Correlation between random variables can be modeled in the following way. Let X and ϵ be random variables with mean 0 and variance 1, and assume X and ϵ are independent. Then set, for constants $\beta_0, \beta_1, \beta_2$, a new random variable Y as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 \epsilon. \tag{1.13}$$

(a) Derive an expression for the correlation between X and Y as a function of β_1 and β_2 (verify that this expression will not depend on β_0).

(b) Suppose we wish to simulate pairs of random variables (X, Y) such that $\mu_X = \mu_Y = 0$, $var(X) = var(Y) = 1$, and the correlation between X and Y is fixed at $\rho \in (-1, 1)$. We can do this by simulating X and ϵ, then using equation (1.13) to generate Y. To achieve this, what values must be used for $\beta_0, \beta_1, \beta_2$? Using `R`, generate four scatter plots from 1000 pairs of normally distributed random variables (X, Y) using this scheme, for $\rho = -0.5, 0, 0.5, 0.9$. Do an independent simulation for each of the four plots. Place all four scatter plots in one graphics window using the `par()` function. Indicate the relevant title for each plot, using the Greek font for ρ (consult `help(plotmath)` and the function `bquote()`).

In addition, for each simulation, summarize in a table the sample variances and correlation of X and Y. Compare these sample values to the theoretical values.

SOLUTION:

(a) The correlation is

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{var(X)var(Y)}} = \frac{E[XY]-\mu_X\mu_Y}{\sqrt{var(X)var(Y)}}.$$

We have, given what we know of ϵ, X and Y,

$$\begin{aligned}
\mu_X &= 0 \\
\mu_Y &= E[\beta_0 + \beta_1 X + \beta_2 \epsilon] = E[\beta_0] + \beta_1 E[X] + \beta_2 E[\epsilon] = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0 \\
var(X) &= 1 \\
var(Y) &= var(\beta_0) + var(\beta_1 X) + var(\beta_2 \epsilon) = 0 + \beta_1^2 + \beta_2^2 = \beta_1^2 + \beta_2^2 \\
E[XY] &= E[\beta_0 X + \beta_1 X^2 + \beta_2 \epsilon X] = \beta_0 E[X] + \beta_1 E[X^2] + \beta_2 E[\epsilon] E[X] = 0 + \beta_1 + 0 = \beta_1.
\end{aligned}$$

So,

$$\rho_{XY} = \frac{\beta_1 - 0 \times \beta_0}{\sqrt{\beta_1^2 + \beta_2^2}} = \frac{\beta_1}{\sqrt{\beta_1^2 + \beta_2^2}}.$$

(b) We have

$$\begin{aligned}
var(Y) &= \beta_1^2 + \beta_2^2, \text{ and} \\
\rho &= \frac{\beta_1}{\sqrt{\beta_1^2 + \beta_2^2}} = \frac{\beta_1}{\sqrt{var(Y)}}.
\end{aligned}$$

If $var(Y) = 1$, this gives (we can assume without loss of generality that $\beta_2 > 0$),

$$\begin{aligned}
\beta_1 &= var(Y)^{1/2} \rho = \rho, \text{ and} \\
\beta_2 &= \sqrt{var(Y) - \beta_1^2} = \sqrt{1 - \rho^2}
\end{aligned}$$

The following R script produces the required plot (Figure 1.6):

```
rho.list = c(-0.5, 0, 0.5, 0.9)
sumtab = matrix(NA,4,3)
par(mfrow=c(2,2))
for (i in 1:4) {
rho = rho.list[i]
eps = rnorm(1000)
x = rnorm(1000)
y = rho*x + sqrt(1 - rho^2)*eps
plot(x,y, main = bquote(rho == .(rho)))
sumtab[i,] = c(var(x), var(y), cor(x,y))
}
```

The table is

```
> sumtab
[,1]      [,2]       [,3]
[1,] 1.0364338 1.0496114 -0.5117990
[2,] 0.9644771 0.9503838 -0.0389420
[3,] 1.0288339 1.0237045  0.5053358
[4,] 0.9148030 0.9224029  0.8937180
```

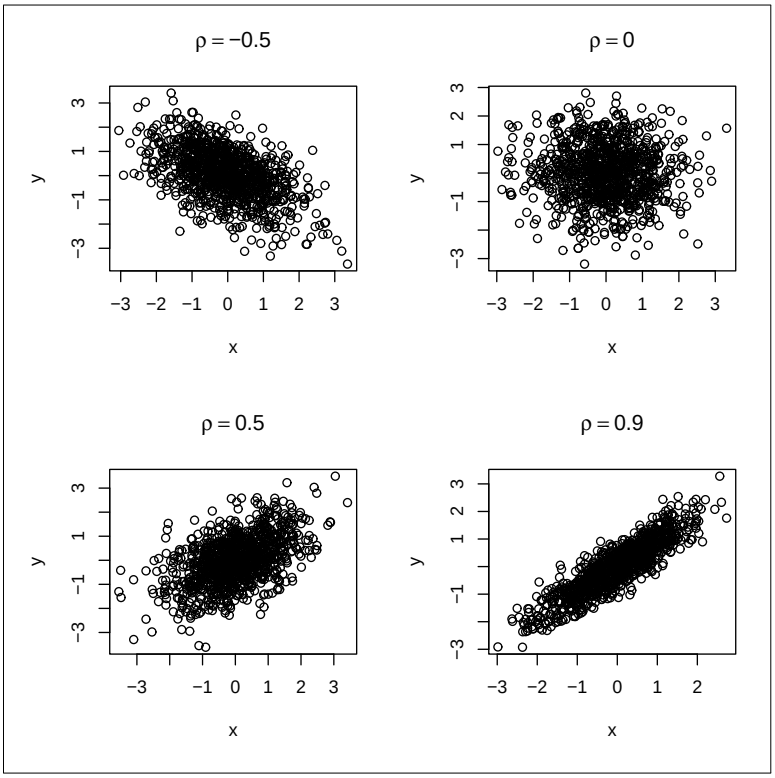

Figure 1.6: Plot for Problem 1.72 (b).

The sample variances and correlations are all close to the theoretical values.

Problem 1.73 If necessary, the number $\pi = 3.141593\ldots$ can be estimated by a simple simulation experiment. The area of a square of unit sides equals 1. A circle of radius $r = 1/2$ can be embedded in this square. The area of a circle is πr^2. Write an `R` function which does the following.

(i) Create an $n \times 2$ matrix, such that each row contains the coordinates of a point uniformly distributed in a unit square centered at $(0, 0)$.

(ii) Count the number of such points contained in the embedded circle described above. Call this number m.

(iii) Use the numbers m and n to estimate π. You can set $n = 10000$.

SOLUTION: The area of the unit square is $A_S = 1$, and the area of the embedded circle is $A_C = \pi/4$. This means

$$\pi = 4 \times \frac{A_C}{A_S}.$$

Suppose (U_1, U_2) are two independent random variables uniformly distributed on interval $(-1/2, 1/2)$. Then (U_1, U_2) is uniformly distributed in the unit square, but is also in the embedded circle only if the event $E = \{U_1^2 + U_2^2 < 1/4\}$ occurs. So, simulate $n = 10000$ pairs (U_1, U_2). Let m be the number of these simulation trials for which E occurs. Then $m/n \approx A_C/A_S$. This gives

$$\pi \approx 4 \times m/n.$$

This is implemented by the following code:

```
> ### Each row of 10000 x 2 matrix contains one (U1,U2) pair
>
> mm = matrix(runif(20000)-0.5,10000,2)
>
> ### Calculate the distance squared U1^2 + U2^2 from the origin for each (U1,U2) pair
>
> r2 = apply(mm,1,function(x) {sum(x^2)})
>
> ### Then calculate 4 * m/n
>
> mean(r2 < 0.25)*4
[1] 3.1628
>
```

Problem 1.74 Suppose a casino has a game in which a player bets x dollars, then with probability p wins back $2x$ dollars (for a net gain of x) and loses the original x dollars with probability $1-p$ (for a net loss of x). Usually, $p < 1/2$. If $p = 1/2$ then the game is *fair*. Probability theory states that in such a fair game, there can be no strategy that results in a positive expected gain.

A commonly claimed counter-example to this is the following strategy. Enter the casino, then play the game, betting $x = 1$ each time, until you have a total gain of 1. For example, the following Win/Loss sequence will accomplish this: *LWLLWWW*, which has gain sequence -1,0,-1,-2,-1,0,1, taking 7 games to reach a gain of 1. The Win/Loss sequence W also achieves a gain of 1 after a single game. Probability theory also states that the probability that a gain of 1 is reached after a finite number of games is 1 (although this *doesn't* hold if $p < 1/2$).

This seems to lead to a contradiction, since if we use this strategy, we can play once a day, and guarantee ourselves a regular income, noting that we can use any value of x we wish. Note that the case of the fair game, $p = 1/2$, is the important one, since if no winning strategy exists for this case, no winning strategy can exist when $p < 1/2$, which settles the matter.

(a) Write an `R` program which simulates this process. Assume $p = 1/2$. For a single simulation, start at *gain* $= 0$, then increase or decrease *gain* after each game by 1. This is the *random walk* introduced in Assignment 1, Question 2. The process stops when *gain* $= 1$. Store the number of games T needed to reach *gain* $= 1$. You may use the `rbinom()` function. Truncate the process at 1000 games. If *gain* $= 1$ has not been reached, indicate this by setting the number of games at, say, $T = 1001$.

(b) Repeat the simulation to get 1000 replicates of T. Estimate the PMF $p_T(k) = P(T = k)$ directly from the data. Construct a *log-log* plot of $\log(p_T(k))$ against $\log(k)$. Note that the frequencies are not sorted in this case. How many times did T exceed 1000? How many times was T within 10 games, inclusive?

(c) Using the `lines()` function, superimpose on this plot the lines

$$f(k) = \log(p_X(1)) - \alpha \log(k),$$

for $\alpha = 1.0, 1.25, 2.0$. Label your plot with the `legend()` function as in Question 4. If you can conclude that T conforms to a power law, what can be said about $E[T]$? The rate at which this strategy earns money is

$$\text{gain rate} = \frac{gain}{\text{number of games played}}.$$

At what rate does this strategy earn money?

SOLUTION: The following `R` script produces the plot in Figure 1.7.

```
> nsim = 1000
> tt = rep(NA, nsim)
> bank = rep(NA, nsim)
>
> for (i in 1:nsim) {
+
+  x = rbinom(1000,size=1,prob=1/2)
+  z = cumsum(2*x-1)
+
+  if (sum(z==1) > 0) {
+  tt[i] = min(which(z==1))
+  bank[i] = min(z[1:tt[i]])
+  }
+  else
+  {
+  tt[i] = 1001
+  bank[i] = min(z)
+  }
+ }
>
> sum(tt <= 10)
[1] 752
> sum(tt == 1001)
[1] 32
>
> xx = as.integer(names(table(tt)))
> par(mfrow=c(1,1),cex=1)
> plot(log(xx), log(table(tt)/1000),xlab = 'log T', ylab = 'log frequency')
> lines(log(xx), max(log(table(tt)/1000)) - 1*log(xx),lty=1)
> lines(log(xx), max(log(table(tt)/1000)) - 1.25*log(xx),lty=2)
> lines(log(xx), max(log(table(tt)/1000)) - 2*log(xx),lty=3)
> legend('bottomleft',legend=paste('alpha = ',c('1.0','1.25','2.0'),sep=''),lty=1:3)
>
```

In this simulation there were 752/1000 simulated values of T within 10, and 32/1000 greater than 1000. Results will vary. From Figure 1.7 the power law $p_X(k) \propto 1/k^\alpha$ holds approximately, with $\alpha \approx 1.25$, and

more generally with $\alpha < 2$. We then have, for some constant c,

$$E[T] = \sum_{k=1}^{\infty} k \frac{c}{k^{\alpha}}$$

noting that the support of T is unbounded. However, $E[T] < \infty$ only if $\alpha > 2$ (compare the summation to the integral $\int_1^{\infty} x^{-\alpha} dx$). If $\alpha < 2$ then in our case $E[T] = \infty$. This means that although the gambler can win a gain of 1 with probability 1 in a finite amount of time, the *gain rate* is 0, since $E[T] = \infty$. If we let G_n be the total gain after the nth game (not day), we would find

$$\lim_{n \to \infty} \frac{G_n}{n} = 0.$$

As a practical matter, using this strategy, we would often find that we cannot play enough games in a single day to achieve the daily gain of 1.

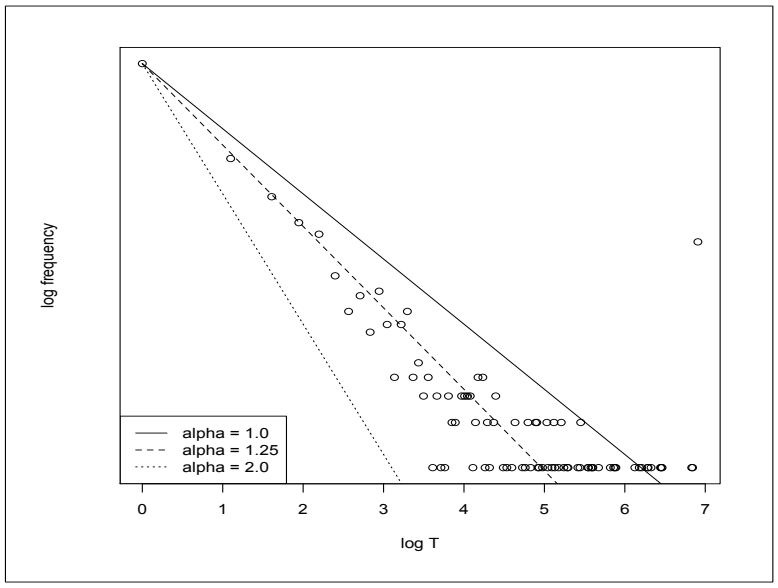

Figure 1.7: Plot for Problem 1.74.

1.9 Stochastic Processes

Problem 1.75 Suppose $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with rates λ_1, λ_2. What type of process is $N(t) = N_1(t) + N_2(t)$?

SOLUTION: At time s, the waiting time until the next arrival for $N(t)$ is $W = \min\{X_1, X_2\}$, where X_1, X_2 are the respective waiting times for $N_1(t)$ and $N_2(t)$. By the memoryless property, X_1, X_2 are exponentially distributed of rates λ_1 and λ_2. This mean W is exponentially distributed with rate $\lambda_1 + \lambda_2$. Therefore, $N(t)$ is a Poisson process of rate $\lambda_1 + \lambda_2$.

Problem 1.76 Suppose an arrival process $N(t)$ is constructed in the following. Starting at time $t = 0$, a coin is flipped every 5 seconds. The coin is biased, so that the probability of tossing a Head is p. Everytime a coin is tossed, an arrival occurs. Give an argument that $N(t)$ is approximately a Poisson process.

SOLUTION: The number of tosses X between arrivals has a geometric distribution with mean $1/p$. However, the geometric distribution shares with exponential distribution the memoryless property (Example 4.20, Theorem 4.14). Thus, an interarrival time W will be approximately exponentially distributed with mean $p^{-1} \times 5$ seconds (the approximation will be more accurate for smaller p, after suitable time rescaling).

Problem 1.77 A *random walk* can be described as follows. We have time points $i = 0, 1, 2, \ldots$. The random walk has value X_i at time point i, according to the following rules:

(1) $X_0 = 0$.
(2) At time point i, +1 or -1 is added to X_i with equal probability, resulting in $X_{i+1} = X_i - 1$ or $X_{i+1} = X_i + 1$. All increments are selected independently.

For example, we could have $X_0 = 0$, $X_1 = 1$, $X_2 = 0$, $X_3 = -1$, $X_4 = -2$, $X_5 = -1$ and so on.

Determine the following probabilities:

(a) $P(X_1 = 1, X_2 = 0, X_3 = -1, X_4 = 0)$,
(b) $P(X_1 = -1, X_2 = -2, X_3 = -3, X_4 = -2)$,
(c) $P(X_4 = 0)$,
(d) $P(X_i > 0 \text{ for } i = 1, 2, 3, 4)$.

SOLUTION: The key here is to recognize that there are 16 possible paths from X_0 to X_4, which can be enumerated with a bit of effort:

Path Index	X_0	X_1	X_2	X_3	X_4
1	0	1	2	3	4
2	0	1	2	3	2
3	0	1	2	1	2
4	0	1	2	1	0
5	0	1	0	1	2
6	0	1	0	1	0
7	0	1	0	-1	0
8	0	1	0	-1	-2
9	0	-1	0	1	2
10	0	-1	0	1	0
11	0	-1	0	-1	0
12	0	-1	0	-1	-2
13	0	-1	-2	-1	0
14	0	-1	-2	-1	-2
15	0	-1	-2	-3	-2
16	0	-1	-2	-3	-4

Then, for parts (a)-(b), we are simply being asked for the probability of a single path, which must be 1/16. For parts (c)-(d) we enumerate the paths which are in the event:

(a) $P(X_1 = 1, X_2 = 0, X_3 = -1, X_4 = 0) = P(\text{Path Index} = 7) = 1/16$,
(b) $P(X_1 = -1, X_2 = -2, X_3 = -3, X_4 = -2) = P(\text{Path Index} = 15) = 1/16$,
(c) $P(X_4 = 0) = P(\text{Path Index} \in \{4, 6, 7, 10, 11, 13\}) = 6/16$,
(d) $P(X_i > 0 \text{ for } i = 1, 2, 3, 4) = P(\text{Path Index} \in \{1, 2, 3\}) = 3/16$.

Problem 1.78 Consider the network shown in Figure 1.8. There are 5 nodes, N_i, $i = 1, \ldots, 5$. Two nodes are considered connected if there is an edge between them. A traveller starts at node N_1 and makes transitions between nodes. The path is denoted $x_1, x_2, \ldots, x_t, \ldots$ with time represented by index $t = 1, 2, \ldots$. The traveller is attempting to reach node N_5, using the following rules:

Rule 1: The initial node $x_1 = N_1$ is fixed. Then x_2 is selected from all nodes connected to N_1 with equal probability.

Rule 2: Following the initial transition, at any $t > 1$, when the traveller is at any node x_t other than N_5 the subsequent node is selected from all nodes connected to x_t *except* for the previously visited node x_{t-1}, with equal probability assigned to each available node. For example, if the traveller transitions from nodes N_3 to N_2, then the next node visited is selected from N_1 and N_5, with probability 1/2 assigned to each.

Rule 3: The traveller remains in node N_5 once it is visited.

We can describe the independence structure this way: when necessary, the traveller can choose the subsequent node by flipping a 2- or a 3- sided coin. The outcome of the toss is independent of all previous coin tosses and on the current choice of coin.

(a) Is $x_1, x_2, \ldots$ a Markov process? Why or why not?
(b) Assuming you answered 'no' to part (a), show how the process can be *Markovianized*. That is, construct a Markov process $z_1, z_2, \ldots$ by defining as a state the transition between pairs of successively visited nodes, rather than the nodes themselves. For example, a transition from node N_2 to node N_1 defines state $N_2 \to N_1$. It may clarify things to introduce a dummy node N_0. The initial state of the Markovianized process would then be the transition $N_0 \to N_1$, which represents the entry of the traveller into the system. The traveller never visits N_0 again, so this transition occurs only once. Enumerate explicitly the state space, and derive the probability transition matrix P. Why is the transition $N_5 \to N_5$ an *absorbing state* in the Markovianized process?
(c) Let T be the time at which the traveller visits node N_5 for the first time. For example, given a path $x_1 = N_1$, $x_2 = N_3$, $x_3 = N_2$, $x_4 = N_5$, we have $T = 4$. Let $P^{(k)}$ be the k-step probability transition matrix of the Markovianized process $z_1, z_2, \ldots$. Explain how the probabilty $P(T < k)$ can be obtained from $P^{(k)}$.
(d) Using `R`, or another suitable computing environment, calculate the probability mass function $p_T(x) = P(T = x)$ for T, and compute its mean. This can be done by successively calculating $P^{(k)}$ for increasing k. Construct a suitable plot of p_T. Note that the support of T is unbounded, so you can confine calculation of $p_T(x)$ to all $x \leq M$, selecting large enough M for which $p_T(M)$ is close to zero.

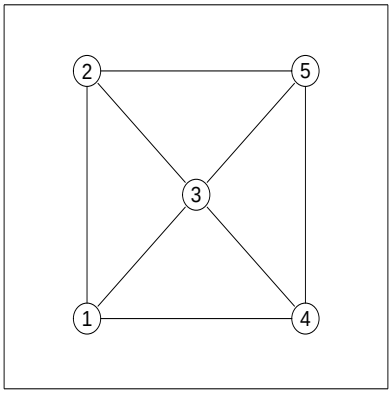

Figure 1.8: Network diagram for Problem 1.78.

SOLUTION:

(a) No, $x_1, x_2, \ldots$ is not a Markov process. The probability $P(x_n = x' \mid x_{n-1}, x_{n-2}, \ldots, x_1)$ depends on both x_{n-1} and x_{n-2}.

(b) The state space consists of transitions $N_j \to N_k$, or $j - k$ for short. For each edge in the network, there are two transitions, except for transitions starting from N_5. In addition, we have the initial transition $0 - 1$ and the final transition $5 - 5$. That makes $2 \times 8 - 3 + 1 + 1 = 15$ transitions, enumerated in the table below. The transition probabilities can be deduced from the rules, yielding the transition matrix given in the table below. The transition $5 - 5$ is and absorbing state because $P(5 - 5 \mid 5 - 5) = 1$.

	0-1	1-2	1-3	1-4	2-1	2-3	2-5	3-1	3-2	3-4	3-5	4-1	4-3	4-5	5-5
0-1	0.00	0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1-2	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1-3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.33	0.00	0.00	0.00
1-4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00
2-1	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2-3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.33	0.00	0.00	0.00	0.00
2-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
3-1	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-2	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00
3-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
4-1	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4-3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.33	0.00	0.33	0.00	0.00	0.00	0.00
4-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
5-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(c) We have

$$P^{(k)}_{0-1,5-5} = P(z_k = 5 - 5 \mid z_1 = 0 - 1) = P(T < k).$$

(d) We calculate $P^{(k)} = P^k$, where P is the (one-step) transition matrix given above. The PMF is then $p_k = P(T < k + 1) - P(T < k)$. The following code completes the problem. See Figure 1.9.

```
#### Construct Transition Matrix

x = c(
```

```
  0,1/3,1/3,1/3,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,1/2,1/2,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,1/3,0,1/3,1/3,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,1/2,1/2,0,
  0,0,1/2,1/2,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,1/3,0,1/3,1/3,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
  0,1/2,0,1/2,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,1/2,0,1/2,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,1/2,0,1/2,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
  0,1/2,1/2,0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,1/3,1/3,0,1/3,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)

pp = matrix(x, 15,15,byrow=T)


#### Iterate matrix multiplication


st = pp[1,15]
ppp = pp
for (i in 1:101) {
  ppp = ppp%*%pp
  st = c(st, ppp[1,15])
}
sum(1 - st) + 1

#### calculate mean and PMF

pa = diff(c(0,st))
mu = sum(pa*(1:length(pa)))

#### output results

ex1 = expression(paste(italic(k)))
ex2 = expression(paste(italic(P),'(',italic(T),'=',italic(k),')',sep=''))
par(mfrow=c(1,1),cex=1,oma=c(2,2,2,2))
barplot(pa[1:50])
axis(1)
mtext(ex1,1,line=3,cex=1.25)
mtext(ex2,2,line=3,cex=1.25)
text(30,.3,paste('mean = ',round(mu,3)))
```

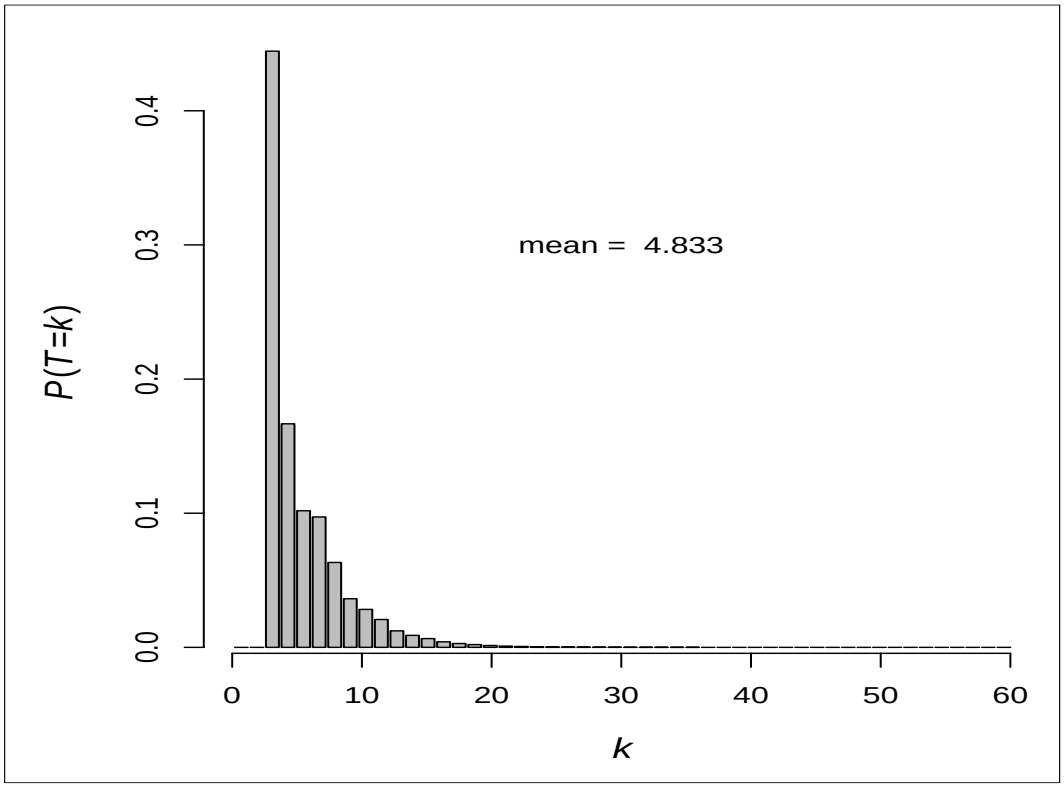

Figure 1.9: Distribution of T for Problem 1.78 (d).

Problem 1.79 Consider the single server queuing system of Example 7.5. This was modeled as a birth and death process with birth rates $\lambda_i = \lambda$, $i \geq 0$ and death rates $\mu_i = \mu$, $i \geq 1$. In Kendall's notation, this is a $M/M/1$ queue. Suppose the queue is modifed to have finite capacity K, that is, there can be no arrivals when there are K customers in the system (including any customer in service). In Kendall's notation, this queueing system is denoted $M/M/1/K$.

(a) Give the birth and death rates for a $M/M/1/K$ queue.
(b) Derive the steady state distribution for the $M/M/1/K$ queue.
(c) Does a steady state distribution always exist? Justify your answer.

SOLUTION:

(a) When the queue is in state $i < K$ customers arrive at rate λ, and for $i \geq K$ customers arrive at rate 0 (that is, arrivals are prohibited). So

$$\lambda_i = \begin{cases} \lambda & ; \quad 0 \leq i \leq K-1 \\ 0 & ; \quad i \geq K \end{cases} .$$

Similarly, the death rate is given by

$$\mu_i = \begin{cases} \mu & ; \quad 1 \leq i \leq K \\ 0 & ; \quad i \geq K+1 \end{cases} .$$

(b) Let $\rho = \lambda/\mu$ be the utilization factor. The steady state probability for state $i = 0$ is, from the balance equations:

$$P_0 = 1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^{i} \mu_k} = 1 + \sum_{i=1}^{K} \rho^i.$$

Then

$$P_i = \frac{\rho^i}{1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} \lambda_k}{\prod_{k=1}^{i} \mu_k} = 1 + \sum_{i=1}^{K} \rho^i}$$

for $0 \leq i \leq K$, and $P_i = 0$, $i > K$.

(c) For any $\rho \geq 0$ we have $P_0 > 0$, so a steady state distribution always exists.

Problem 1.80 Whether or not a Markov chain is an adequate model for a given application is an important question. We'll use a Markov chain model to design a simple tic-tac-toe player. It will play both sides.

Create in `R` the following objects:

(a) The board will consist of a vector of length 9. An unoccupied position is set to 0, otherwise the position is occupied by player 1 or 2.

(b) A tic-tac-toe board has 8 'rows'. The three horizontal rows are (1,2,3), (4,5,6) and (7,8,9), the three vertical rows are (1,4,7), (2,5,8) and (3,6,9). The diagonal rows are (1,5,9) and (3,5,7). Create an 8×3 table which stores these rows.

(c) Each position on the board belongs to certain rows. For example, position 6 belongs to rows (3,6,9) and (4,5,6), and so on. Create a list of length 9, in which the ith element is a vector of indices referencing the rows to which position i belongs.

(d) To choose a move, player 1 examines each position, and assigns each a score. If position i is occupied it is assigned score 0. Otherwise, each row containing i is looked up. The number of positions in that row occupied by 1 and 2 are stored in `n.us` and `n.them` respectively. The row is scored according to the following table:

		`n.them` = 0	1	2
`n.us` =	0	10	100	10,000
	1	1,000	1	0
	2	100,000	0	0

Then, the score for position i is the sum of the scores of the rows containing i. For example, for the following board it is player 1's turn to move. Positions 5 and 7 are scored 0, since they are occupied. To score position 1, note that it is contained in 3 rows, (1,2,3), (1,4,7) and (1,5,9). For row (1,2,3), `n.us` = 0, `n.them` = 0, so this row contributes 10 to the score. For row (1,4,7) `n.us` = 0, `n.them` = 1, and for row (1,5,9) `n.us` = 1, `n.them` = 0, so these rows contribute 100 and 1,000, respectively. The total score for position 1 is then 1,000 + 100 + 10 = 1,110.

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{matrix}$$

(e) After each position's score is calculated the position with the highest score is selected. If more than one position has the maximum score, one of these is chosen at random.
(f) Player 2 uses the same strategy, calculating `n.us` and `n.them` accordingly.
(g) Write an R program to simulate a tic-tac-toe game with alternating players using the same strategy. The game ends after one player completely occupies any row (and therefore wins), or the board is full. Run the simulation 1,000 times, and store the frequency of outcomes (player 1 wins, player 2 wins or the games ends in a draw). What are the frequencies of each outcome?
(h) OPTIONAL BONUS QUESTION: Which of the three outcomes can occur? Justify your answer. Symmetry plays a role here.

SOLUTION: The following `R` program plays the game described in Problem 1.80. All games in the 1,000 simulations end in draws.

```
#### create row.table: 8x3 matrix of row definitions

row.table = matrix( c(1,2,3,4,5,6,7,8,9,1,4,7,2,5,8,3,6,9,1,5,9,3,5,7), ncol=3, byrow=T)

### create row.list. The ith element is a vector of indices to all
### rows in row.table in which position i is located

row.list = vector('list',9)
row.list[[1]] = c(1,4,7)
row.list[[2]] = c(1,5)
row.list[[3]] = c(1,6,8)
row.list[[4]] = c(2,4)
row.list[[5]] = c(2,5,7,8)
row.list[[6]] = c(2,6)
row.list[[7]] = c(3,4,8)
row.list[[8]] = c(3,5)
row.list[[9]] = c(3,6,7)

### score.matrix is a 3 x 3 matrix. Element score.matrix[n.us+1, n.them+1]
### gives the score for (n.us, n.them)

score.matrix = matrix(0, nrow=3, ncol=3)
score.matrix[3,1] = 100000
score.matrix[1,3] = 10000
score.matrix[2,1] = 1000
score.matrix[1,2] = 100
score.matrix[1,1] = 10
```

```
score.matrix[2,2] = 1

### choose.move is a function which accepts the current board,
### the values us.them = c(n.us, n.them), and the objects
### row.table, row.list, score.matrix created above.
### The score is calculated for each position. The highest score is identified.
### If more than one position has the highest score, one of them is chosen at random.
### The output is a list with elements names move (the selected position),
### max.score (the maximum score),
### score.temp (the vector of scores for each position)

choose.move = function(board, us.them, row.table, row.list, score.matrix) {

   score.temp = rep(0,9)

   # calculate score for each position

   for (i in 1:9) {
      if (board[i] == 0) {
         for (j in 1:length(row.list[[i]]) ) {
            row.temp = board[row.table[row.list[[i]][j], ]]
            n.us = sum(row.temp==us.them[1])
            n.them =  sum(row.temp==us.them[2])
            score.temp[i] = score.temp[i] + score.matrix[n.us+1,n.them+1]
         }
      }
   }

   # determine maximum score

   max.score = max(score.temp)

   # identify positions with maximum score

   move.list = which(score.temp==max.score)

   if (length(move.list)==1) {

      # if highest scoring position is unique, copy onto move

      move = move.list[1]
   } else
   {

      # otherwise, select position at random from highest scoring ones
```

```
      move = sample(which(score.temp==max.score),1)
   }

   # return selected position and score

   return(list(move=move, max.score=max.score, score.temp=score.temp))

} ### end choose.move

### simulate nsim games

nsim = 1000

### store result in sv 0=draw, 1=Player 1 wins, 2 = Player 2 wins

sv = rep(0, nsim)

for (iii in 1:nsim) {

   # create playing board as vector of length 9

   board = rep(0,9)

   # flag==1 is used to indicate end of game

   flag = 0
   while (flag == 0) {

      # Player 1 plays

      junk = choose.move(board, c(1,2), row.table, row.list, score.matrix)

      # update board

      board[junk$move] = 1

      # Player 1 wins if score is 100000. Game ends if there are no empty
      # positions on the board

      if ( (junk$max.score >= 100000) | (sum(board==0)==0) ) {
         flag=1
         if (junk$max.score >= 100000) {sv[iii] = 1}
      }
```

```
      # Player 2 plays if flag==0

      if (flag == 0) {

         # Player 2 plays

         junk = choose.move(board, c(2,1), row.table, row.list, score.matrix)

         # update board

         board[junk$move] = 2

         # Player 2 wins if score is 100000. Game ends if there are no empty
         # positions on the board

         if ( (junk$max.score >= 100000) | (sum(board==0)==0) ) {
            flag=1
            if (junk$max.score >= 100000) {sv[iii] = 2}
         }
      }
   }  ### while (flag == 0)
} ### end for (iii in 1:nsim)
```

After the program is run we should see something like this (meaning that every game was a draw).

```
> table(sv)
sv
0
1000
```

OPTIONAL QUESTION: The table below shows the sequence of one game, including scores for each position, and the subsequent move. For the first move, note that when the board is empty, the middle position (5) is scored highest, so Player 1 always starts there.

For Player 2's first move, all four diagonal positions (1,3,7,9) are scored highest. Player 2 selects one of these at random. However, note that by symmetry there is no important difference between these moves.

For Player 1's second move, the two remaining diagonal positions which share a row with Player 2's first position are scored highest. Player 1 chooses one of these at random. Both are essentially identical, after symmetry is accounted for.

If we continue in this way for the remaining moves, we see that after symmetry is accounted for, there is really only one available move at each turn. All games must be essentially identical, and will therefore end in draws.

	Score			Board		
1	30	20	30	0	0	0
	20	40	20	0	1	0
	30	20	30	0	0	0
2	120	110	120	0	0	0
	110	0	110	0	1	0
	120	110	120	0	0	2
3	21	1010	1110	0	0	0
	1010	0	1100	0	1	0
	1110	1100	0	1	0	2
4	111	110	11010	0	0	2
	200	0	1100	0	1	0
	0	101	0	1	0	2
5	1101	1100	0	0	0	2
	2000	0	11000	0	1	1
	0	1001	0	1	0	2
6	1101	1100	0	0	0	2
	10100	0	0	2	1	1
	0	101	0	1	0	2
7	102	1100	0	0	1	2
	0	0	0	2	1	1
	0	1001	0	1	0	2
8	3	0	0	0	1	2
	0	0	0	2	1	1
	0	10001	0	1	2	2
9	3	0	0	1	1	2
	0	0	0	2	1	1
	0	0	0	1	2	2

1.10 Bayes Theorem. Diagnostic Testing and Classification

Problem 1.81 The odds of an event A is denoted $Odds(A)$. Suppose the distribution of a random variable X depends on whether or not event A occurs. In particular, conditional on A, $X \sim bin(4, 0.5)$. Conditional on A^c, $X \sim bin(2, 0.9)$.

Determine the relationship between $Odds(A \mid X = x)$ and $Odds(A)$ for $x = 0, 1, 2, 3, 4$. For which values of x does evidence of the form $\{X = x\}$ increase the odds that A does not occur.

SOLUTION: By Bayes Rule, we have

$$Odds(A \mid X = x) = LR \times Odds(A),$$

where LR is the likelihood ratio

$$LR = \frac{P(X = x \mid A)}{P(X = x \mid A^c)}, \quad x = 0, 1, \ldots, 4.$$

The PMF of $X \sim bin(n,p)$ is $p(x) = \binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \ldots, n$. However, where needed, we can set $p(x) = 0$ for any x not in the set $\{0, 1, \ldots, n\}$. This gives

	$x =$	0	1	2	3	4
A	$p(x) =$	$\binom{4}{0}0.5^4$	$\binom{4}{1}0.5^4$	$\binom{4}{2}0.5^4$	$\binom{4}{3}0.5^4$	$\binom{4}{4}0.5^4$
A^c	$p(x) =$	$\binom{2}{0}0.9^0 0.1^2$	$\binom{2}{1}0.9^1 0.1^1$	$\binom{2}{2}0.9^2 0.1^0$	0	0
	LR	$(0.5^4)/(0.1^2)$	$(4 \times 0.5^4)/(2 \times 0.9 \times 0.1)$	$(6 \times 0.5^4)/(0.9^2)$	∞	∞
		$= 5.25$	≈ 1.39	≈ 0.463	∞	∞

If $LR > 1$, the evidence increases the odds that A occurs, if $LR < 1$ the evidence increases the odds that A does not occur. The only value for which $LR < 1$ is $x = 2$.

Problem 1.82 A test for the presence of an infection is developed. It is administered to subjects whose infection status is known (83 are known be be infected, 420 are known to be not infected). The results are summarized in the Table 1.1. Calculate the *positive predictive value* and the *negative predictive value* of the test for a population with an infection prevalence of $prev = 0.05$.

Table 1.1: Outcomes of Diagnostic Test for Problem 1.82.

		Infection		
		Positive	Negative	Total
Diagnostic Test	Positive	72	5	77
	Negative	11	415	426
	Total	83	420	503

SOLUTION: We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{72}{83} \approx 0.867 \\ spec &= \frac{TN}{TN+FP} = \frac{415}{420} \approx 0.988. \end{aligned}$$

For specific prevalence $prev = 0.03$ we have

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1-spec) \times (1-prev)} \\ &= \frac{0.867 \times 0.05}{0.867 \times 0.05 + (1-0.988) \times (1-0.05)} \\ &\approx 0.793 \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1 - prev)}{spec \times (1 - prev) + (1 - sens) \times prev} \\ &= \frac{0.988 \times (1 - 0.05)}{0.988 \times (1 - 0.05) + (1 - 0.867) \times 0.05} \\ &\approx 0.993. \end{aligned}$$

Problem 1.83 The odds of an event A is denoted $Odds(A)$. Suppose the distribution of a random variable $X \in \{1, 2, 3, 4, 5\}$ depends on whether or not event A occurs. In particular, conditional on A, the PMF of X is given by $(p_1, p_2, p_3, p_4, p_5) = (0, 1/4, 1/4, 1/4, 1/4)$. Conditional on A^c, the PMF of X is given by $(p'_1, p'_2, p'_3, p'_4, p'_5) = (4/10, 3/10, 2/10, 1/10, 0)$.

Determine the relationship between $Odds(A \mid X = x)$ and $Odds(A)$ for $x = 1, 2, 3, 4, 5$. For which values of x does evidence of the form $\{X = x\}$ increase the odds that A occurs?

SOLUTION: By Bayes Rule, we have

$$Odds(A \mid X = x) = LR \times Odds(A),$$

where LR is the likelihood ratio

$$LR = \frac{P(X = x \mid A)}{P(X = x \mid A^c)}, \quad x = 0, 1, \ldots, 4.$$

The particular values of LR are given in the following table:

	$x =$	1	2	3	4	5
A	$p(X = x) =$	0	1/4	1/4	1/4	1/4
A^c	$p(X = x) =$	4/10	3/10	2/10	1/10	0
	$LR =$	0	5/6	5/4	5/2	∞

If $LR > 1$, the evidence increases the odds that A occurs, if $LR < 1$ the evidence decreases the odds that A occurs. The values for which $LR > 1$ are $x = 3, 4, 5$.

Problem 1.84 A test for a certain infection was evaluated experimentally. When administered to a test group of 425 individuals known to have the infection, the test was positive in 401 cases. The test was also administered to a control group of 765 subjects known to be free of the infection. The test was positive in 12 cases.

(a) Estimate the sensitivity and specificity of the test directly from the data.
(b) This test is intended to be used in clinical populations of varying infection prevalence. Use `R` to construct plots of PPV and NPV for values of prevalence ranging from 0 to 20%. Use the `type = 'l'` option of the `plot()` function.
(c) Calculate prevalence, NPV and PPV directly from the data. How do these values compare to those shown in the plots of part (b)?
(d) Give the relationship between the prior and posterior odds of infection for both a positive and negative test result.

SOLUTION: We can summarize the study with the following table:

		Infection		
		Positive	Negative	Total
Diagnostic Test	Positive	401	12	413
	Negative	24	753	777
	Total	425	765	1190

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{401}{425} \approx 0.944 \\ spec &= \frac{TN}{TN+FP} = \frac{753}{765} \approx 0.984. \end{aligned}$$

(b) Given our calculated values of $spec$ and $sens$ we can given PPV and NPV as functions of $prev$:

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1-spec) \times (1-prev)} \\ &= \frac{0.944 \times prev}{0.944 \times prev + (1-0.984) \times (1-prev)} \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1-prev)}{spec \times (1-prev) + (1-sens) \times prev} \\ &= \frac{0.984 \times (1-prev)}{0.984 \times (1-prev) + (1-0.944) \times prev} \end{aligned}$$

The plot itself can be constructed using the following code (see Figure 1.10 below):

```
### Calculate sens, spec as global variables

sens = 401/425
spec = (765-12)/765

### Create functions for PPV and NPV

ppv0 = function(prev,sens,spec) { sens*prev/(sens*prev + (1-spec)*(1-prev)) }
npv0 = function(prev,sens,spec) { spec*(1-prev)/(spec*(1-prev) + (1-sens)*prev) }

### Create range of prev values

prev = seq(0,0.20,by = 0.001)

### Draw plots
```

```
par(mfrow=c(1,2),cex=1.0,cex.lab=1.0,cex.axis=1.0,mar=c(6,6,2,2),pty='s')
plot(prev, ppv0(prev,sens,spec),type='l', xlab='Prevalence',ylab='PPV')
plot(prev, npv0(prev,sens,spec),type='l', xlab='Prevalence',ylab='NPV')
```

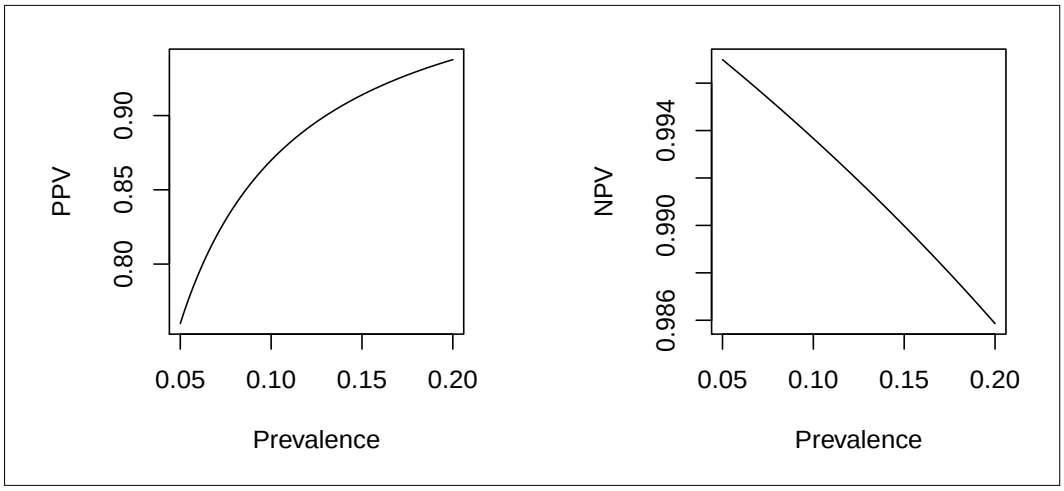

Figure 1.10: Plots for Problem 1.84 (b).

(c) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP+FN}{N} = \frac{425}{1190} \approx 0.357 \\ PPV &= \frac{TP}{TP+FP} = \frac{401}{413} \approx 0.971 \\ NPV &= \frac{TN}{TN+FN} = \frac{753}{777} \approx 0.969. \end{aligned}$$

Directly from the data, the prevalence is $prev = 0.357$, which is much higher than the values used for the plot. Then $PPV = 0.941$ is much higher than the plotted values, and $NPV = 0.969$ is much lower than the plotted values. This is due to the dependence of PPV and NPV on the prevalence.

(d) The positive and negative likelihood ratios are

$$\begin{aligned} LR_+ &= \frac{sens}{1-spec} = \frac{401/425}{12/765} \approx 60.16, \\ LR_- &= \frac{1-sens}{spec} = \frac{24/425}{753/765} \approx 0.0574. \end{aligned}$$

Note that it is better to use the exact values of $sens$ and $spec$ in this calculation to avoid rounding error. Bayes' theorem for odds is then

$$\begin{aligned} Odds(O_+ \mid T_+) &= LR_+ \times Odds(O_+) = 60.16 \times Odds(O_+) \\ Odds(O_+ \mid T_-) &= LR_- \times Odds(O_+) = 0.0574 \times Odds(O_+). \end{aligned}$$

Problem 1.85 A model for a diagnostic test for a certain condition relies on the four events:

$$\begin{aligned} O_- &= \{ \text{ the patient does not have the condition } \} \\ O_+ &= \{ \text{ the patient has the condition } \} \\ T_- &= \{ \text{ the patient tests negative } \} \\ T_+ &= \{ \text{ the patient tests positive } \}. \end{aligned}$$

The quantities *sens* (sensitivity), *spec* (specificity) and *prev* (prevalence) are defined as

$$\begin{aligned} sens &= P(T_+ \mid O_+) \\ spec &= P(T_- \mid O_-) \\ prev &= P(O_+). \end{aligned}$$

Suppose an existing diagnostic test is evaluated by estimating the conditional probabilities $P(O_+ \mid T_+)$ and $P(O_+ \mid T_-)$. For example, these conditional probabilities might be estimated by following up with additional testing patients who have tested positive and negative in a clinical setting.

We would like to verify that $P(O_+ \mid T_+) > P(O_+ \mid T_-)$, and then quantify this difference using some distance function. We will examine three such distances, in each case expressing the difference analytically as a function of *sens*, *spec* and *prev*.

(a) First consider the additive difference of the conditional probabilities:

$$\Delta = P(O_+ \mid T_+) - P(O_+ \mid T_-).$$

Express Δ as a function of *sens*, *spec* and *prev*. Does Δ depend on *prev*? In particular, what is the limit of Δ as *prev* approaches 0, and as *prev* approaches 1?

(b) Next, consider the *relative risk*, which is defined as the ratio of the conditional probabilities:

$$RR = \frac{P(O_+ \mid T_+)}{P(O_+ \mid T_-)}.$$

Express RR as a function of *sens*, *spec* and *prev*. Does RR depend on *prev*? In particular, what is the limit of RR as *prev* approaches 0, and as *prev* approaches 1?

(c) The *odds ratio* is the ratio of the conditional odds:

$$OR = \frac{Odds(O_+ \mid T_+)}{Odds(O_+ \mid T_-)},$$

where, in general, the conditional odds is given by

$$Odds(A \mid B) = \frac{P(A \mid B)}{1 - P(A \mid B)}.$$

Express OR as a function of *sens*, *spec* and *prev*. Verify that OR does not depend on *prev*.

SOLUTION: The relevant quantities are

$$P(O_+ \mid T_+) = PPV = \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)}$$

and

$$P(O_+ \mid T_-) = 1 - NPV = \frac{(1 - sens) \times prev}{spec \times (1 - prev) + (1 - sens) \times prev}.$$

Define the odds $O_{prev} = prev/(1 - prev)$.

(a) We have directly,

$$\begin{aligned}
\Delta &= P(O_+ \mid T_+) - P(O_+ \mid T_-) \\
&= \left(\frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)}\right) - \left(\frac{(1 - sens) \times prev}{spec \times (1 - prev) + (1 - sens) \times prev}\right) \\
&= \left(\frac{prev}{prev + \left(\frac{1-spec}{sens}\right) \times (1 - prev)}\right) - \left(\frac{prev}{prev + \left(\frac{spec}{1-sens}\right) \times (1 - prev)}\right).
\end{aligned}$$

By direct substitution, $\Delta = 0$ for $prev = 0, 1$. For $0 < prev < 1$, $\Delta \neq 0$ unless

$$\frac{1 - spec}{sens} = \frac{spec}{1 - sens},$$

which is equivalent to $sens = 1 - spec$. So Δ depends on $prev$, unless $sens = 1 - spec$.

(b) We can write

$$\begin{aligned}
RR &= \frac{P(O_+ \mid T_+)}{P(O_+ \mid T_-)} \\
&= \frac{\left(\frac{sens \times prev}{sens \times prev + (1-spec) \times (1-prev)}\right)}{\left(\frac{(1-sens) \times prev}{spec \times (1-prev) + (1-sens) \times prev}\right)} \\
&= \frac{prev + \left(\frac{spec}{1-sens}\right) \times (1 - prev)}{prev + \left(\frac{1-spec}{sens}\right) \times (1 - prev)}
\end{aligned}$$

Substituting $prev = 0$ yields

$$RR = \left(\frac{sens}{1 - sens}\right) \times \left(\frac{spec}{1 - spec}\right).$$

Substituting $prev = 1$ yields

$$RR = 1.$$

So, as for Δ, RR depends on $prev$, unless $sens = 1 - spec$.

(c) We may express the odds:

$$\begin{aligned}
Odds(O_+ \mid T_+) &= \frac{Odds(O_+ \mid T_+)}{1 - Odds(O_+ \mid T_+)} \\
&= \frac{sens \times prev}{(1 - spec) \times (1 - prev)},
\end{aligned}$$

and

$$\begin{aligned} Odds(O_+ \mid T_-) &= \frac{Odds(O_+ \mid T_+)}{1 - Odds(O_+ \mid T_+)} \\ &= \frac{(1 - sens) \times prev}{spec \times (1 - prev)}. \end{aligned}$$

The odds ratio is then

$$\begin{aligned} OR &= \frac{Odds(O_+ \mid T_+)}{Odds(O_+ \mid T_-)} \\ &= \frac{\frac{sens \times prev}{(1-spec) \times (1-prev)}}{\frac{(1-sens) \times prev}{spec \times (1-prev)}} \\ &= \frac{sens \times spec}{(1 - sens) \times (1 - spec)}, \end{aligned}$$

which does not depend on $prev$.

Problem 1.86 A test for a certain infection was evaluated experimentally. When administered to a test group of 285 individuals known to have the infection, the test was positive in 256 cases. The test was also administered to a control group of 220 subjects known to be free of the infection. The test was positive in 12 cases.

(a) Estimate the sensitivity and specificity of the test directly from the data.
(b) This test is intended to be used in clinical populations of varying infection prevalence. Use `R` to construct plots of PPV and NPV for values of prevalence ranging from 0 to 10%. Use the `type = 'l'` option of the `plot()` function.
(c) Calculate prevalence, NPV and PPV directly from the data. How do these values compare to those shown in the plots of part (b)?

SOLUTION: We can summarize the study with the following table:

		Infection		
		Positive	Negative	Total
Diagnostic Test	Positive	256	12	268
	Negative	29	208	237
	Total	285	220	505

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP + FN} = \frac{256}{285} \approx 0.898 \\ spec &= \frac{TN}{TN + FP} = \frac{208}{220} \approx 0.945. \end{aligned}$$

(b) Given our calculated values of *spec* and *sens* we can given PPV and NPV as functions of *prev*:

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1 - spec) \times (1 - prev)} \\ &= \frac{0.898 \times prev}{0.898 \times prev + (1 - 0.945) \times (1 - prev)} \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1 - prev)}{spec \times (1 - prev) + (1 - sens) \times prev} \\ &= \frac{0.945 \times (1 - prev)}{0.945 \times (1 - prev) + (1 - 0.898) \times prev} \end{aligned}$$

The plot itself can be constructed using the following code (see Figure 1.11 below):

```
### Calculate sens, spec as global variables

sens = 256/285
spec = 208/220

### Create functions for PPV and NPV

ppv0 = function(prev,sens,spec) { sens*prev/(sens*prev + (1-spec)*(1-prev)) }
npv0 = function(prev,sens,spec) { spec*(1-prev)/(spec*(1-prev) + (1-sens)*prev) }

### Create range of prev values

prev = seq(0,0.1,by = 0.001)

### Draw plots

par(mfrow=c(1,2),cex=1.0,cex.lab=1.0,cex.axis=1.0,mar=c(6,6,2,2),pty='s')
plot(prev, ppv0(prev,sens,spec),type='l', xlab='Prevalence',ylab='PPV')
plot(prev, npv0(prev,sens,spec),type='l', xlab='Prevalence',ylab='NPV')
```

(c) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP + FN}{N} = \frac{285}{505} \approx 0.564 \\ PPV &= \frac{TP}{TP + FP} = \frac{256}{268} \approx 0.955 \\ NPV &= \frac{TN}{TN + FN} = \frac{208}{237} \approx 0.878. \end{aligned}$$

Directly from the data, the prevalence is $prev = 0.564$, which is much higher than the values used for the plot. Then $PPV = 0.955$ is much higher than the plotted values, and $NPV = 0.878$ is much lower than the plotted values. This is due to the dependence of PPV and NPV on the prevalence.

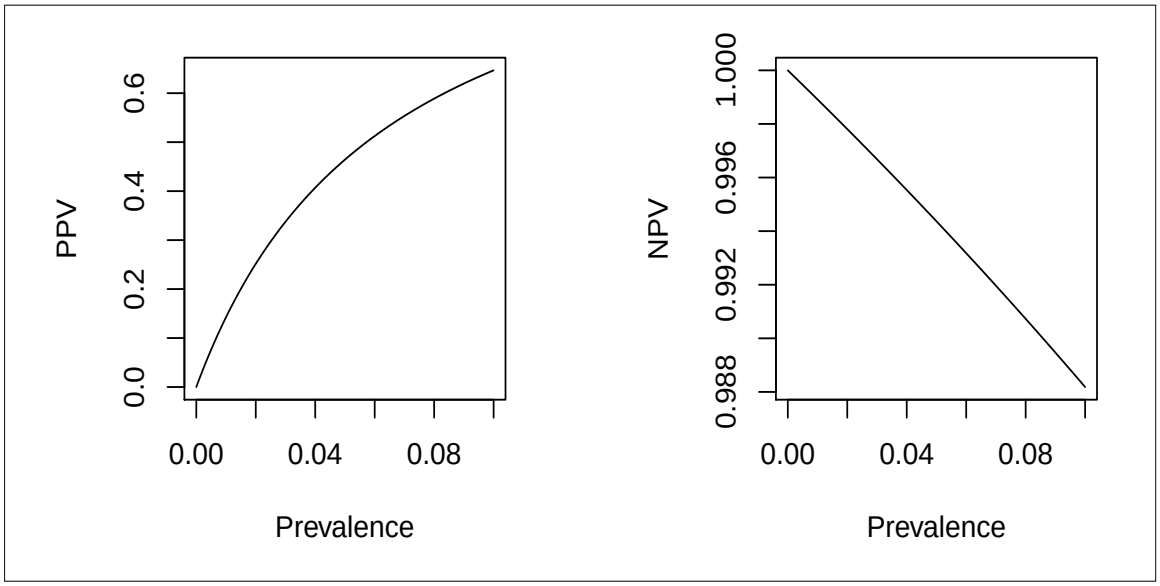

Figure 1.11: Plots for Problem 1.86 (b).

Problem 1.87 A test for the presence of an infection is developed. The test is administered to a test group of 159 individuals known to have the infection. Of this group 154 test positive. The test is also administered to a control group of 325 subjects known to be free of the infection. Of these, 6 test positive.

(a) Estimate the sensitivity and specificity of the test directly from the data.
(b) Calculate prevalence, NPV and PPV directly from the data, then recaclulate assuming a prevalence of 2%.
(c) Give the relationship between the prior and posterior odds of infection for both a positive and negative test result.

SOLUTION: We can summarize the study with the following table:

		Infection		
		Positive	Negative	Total
Diagnostic Test	Positive	154	6	160
	Negative	5	319	324
	Total	159	325	484

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{154}{159} \approx 0.969 \\ spec &= \frac{TN}{TN+FP} = \frac{319}{325} \approx 0.982. \end{aligned}$$

(b) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP+FN}{N} = \frac{159}{484} \approx 0.329 \\ PPV &= \frac{TP}{TP+FP} = \frac{154}{160} \approx 0.963 \\ NPV &= \frac{TN}{TN+FN} = \frac{319}{324} \approx 0.985. \end{aligned}$$

There is a prevalence of 32.9%, which is (hopefully) much higher than any prevalence we would expect to see in any population. If $prev = 0.02$, then

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1-spec) \times (1-prev)} \\ &= \frac{0.969 \times 0.02}{0.969 \times 0.02 + (1-0.982) \times (1-0.02)} \\ &\approx 0.517 \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1-prev)}{spec \times (1-prev) + (1-sens) \times prev} \\ &= \frac{0.982 \times (1-0.02)}{0.982 \times (1-0.02) + (1-0.969) \times 0.02} \\ &\approx 0.999. \end{aligned}$$

(c) We follow Section 5.4.2. The positive and negative likelihood ratios are

$$\begin{aligned} LR_+ &= \frac{sens}{1-spec} = \frac{154/159}{6/325} \approx 52.5, \\ LR_- &= \frac{1-sens}{spec} = \frac{5/159}{319/325} \approx 0.032. \end{aligned}$$

Note that it is better to use the exact values of *sens* and *spec* in this calculation to avoid rounding error. Bayes' theorem for odds is then

$$\begin{aligned} Odds(O_+ \mid T_+) &= LR_+ \times Odds(O_+) = 52.5 \times Odds(O_+) \\ Odds(O_+ \mid T_-) &= LR_- \times Odds(O_+) = 0.032 \times Odds(O_+). \end{aligned}$$

Problem 1.88 A test for Hepatitis-B is developed. The test is administered to a test group of 147 individuals known to have Hepatitis-B. Of this group 123 test positive. The test is also administered to a control group of 220 subjects known to be free of Hepatitis-B. Of these, 15 test positive.

(a) Estimate the sensitivity and specificity of the test directly from the data.

(b) This test is intended to be used in clinical populations of varying infection prevalence. Use `R` to construct plots of PPV and NPV for values of prevalence ranging from 0 to 5%. Use the `type = 'l'` option of the `plot()` function.

(c) Calculate prevalence, NPV and PPV directly from the data. How do these values compare to those shown in the plots of part (b)?

SOLUTION: We can summarize the study with the following table:

		Hepatitis-B		
		Positive	Negative	Total
Diagnostic Test	Positive	123	15	138
	Negative	24	205	229
	Total	147	220	367

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{123}{147} \approx 0.837 \\ spec &= \frac{TN}{TN+FP} = \frac{205}{220} \approx 0.932. \end{aligned}$$

(b) The script shown below produces the plot in Figure 1.12.

```
> sens = 123/147
> spec = (220-15)/220
>
> prev = seq(0,0.05,by = 0.001)
>
> par(mfrow=c(2,1),cex=1.0)
>
> f0 = function(prev,sens,spec) { sens*prev/(sens*prev + (1-spec)*(1-prev)) }
> plot(prev, f0(prev,sens,spec),type='l', xlab='Prevalence',ylab='PPV')
> title('Hepatiti-B Test')
>
>
> f0 = function(prev,sens,spec) { spec*(1-prev)/(spec*(1-prev) + (1-sens)*prev) }
> plot(prev, f0(prev,sens,spec),type='l', xlab='Prevalence',ylab='NPV')
> title('Hepatiti-B Test')
```

(c) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP+FN}{N} = \frac{147}{367} \approx 0.401 \\ PPV &= \frac{TP}{TP+FP} = \frac{123}{138} \approx 0.891 \\ NPV &= \frac{TN}{TN+FN} = \frac{205}{229} \approx 0.895. \end{aligned}$$

There is a prevalence of 40.1%, which is (hopefully) much higher than any prevalence we would expect to see in any population. The PPV estimated directly from the study is much higher than a PPV we would expect to encounter in a population, while the NPV is lower.

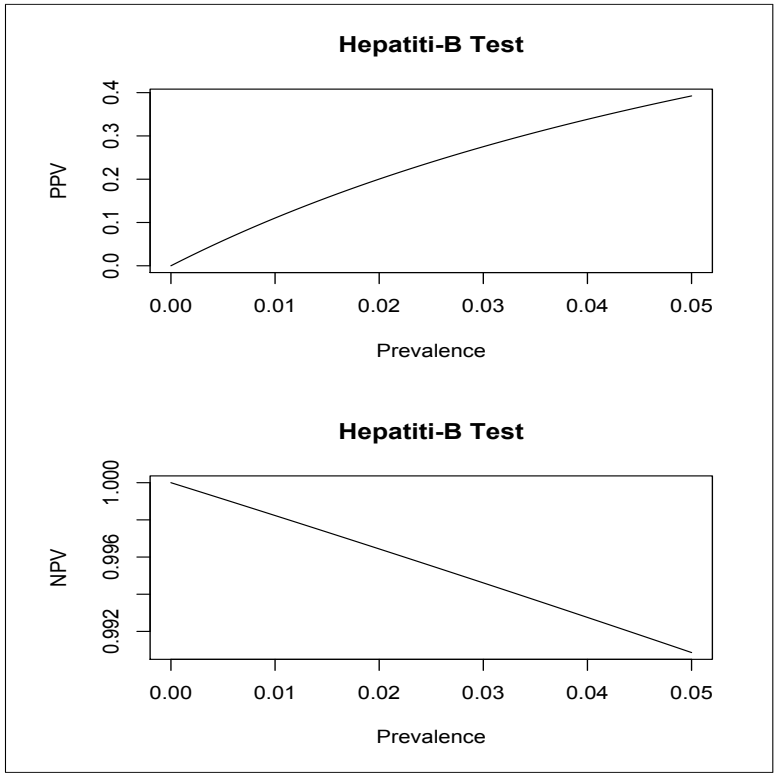

Figure 1.12: Plot for Problem 1.88 (b).

Problem 1.89 A test for the presence of an infection is developed. The test is administered to a test group of 234 individuals known to have the infection. Of this group 229 test positive. The test is also administered to a control group of 705 subjects known to be free of the infection. Of these, 3 test positive.

(a) Estimate the sensitivity and specificity of the test directly from the data.
(b) Calculate prevalence, NPV and PPV directly from the data, then recalculate assuming a prevalence of 10%.
(c) Give the relationship between the prior and posterior odds of infection for both a positive and negative test result.

SOLUTION: We can summarize the study with the following table:

		Infection		
		Positive	Negative	Total
Diagnostic Test	Positive	229	3	232
	Negative	5	702	707
	Total	234	705	939

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{229}{234} \approx 0.979 \\ spec &= \frac{TN}{TN+FP} = \frac{702}{705} \approx 0.996. \end{aligned}$$

(b) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP+FN}{N} = \frac{234}{939} \approx 0.249 \\ PPV &= \frac{TP}{TP+FP} = \frac{229}{232} \approx 0.987 \\ NPV &= \frac{TN}{TN+FN} = \frac{702}{707} \approx 0.993. \end{aligned}$$

If $prev = 0.1$, then

$$\begin{aligned} PPV &= \frac{sens \times prev}{sens \times prev + (1-spec) \times (1-prev)} \\ &= \frac{0.979 \times 0.1}{0.979 \times 0.1 + (1-0.996) \times (1-0.1)} \\ &\approx 0.962 \end{aligned}$$

and

$$\begin{aligned} NPV &= \frac{spec \times (1-prev)}{spec \times (1-prev) + (1-sens) \times prev} \\ &= \frac{0.996 \times (1-0.1)}{0.996 \times (1-0.1) + (1-0.979) \times 0.1} \\ &\approx 0.998. \end{aligned}$$

(c) The positive and negative likelihood ratios are

$$\begin{aligned} LR_+ &= \frac{sens}{1-spec} = \frac{229/234}{3/705} \approx 229.98, \\ LR_- &= \frac{1-sens}{spec} = \frac{5/234}{702/705} \approx 0.0215. \end{aligned}$$

Note that it is better to use the exact values of $sens$ and $spec$ in this calculation to avoid rounding error. Bayes' theorem for odds is then

$$\begin{aligned} Odds(O_+ \mid T_+) &= LR_+ \times Odds(O_+) = 229.98 \times Odds(O_+) \\ Odds(O_+ \mid T_-) &= LR_- \times Odds(O_+) = 0.0215 \times Odds(O_+). \end{aligned}$$

Chapter 2

Statistical Methods

2.1 Inference for Population Means

Problem 2.1 For an *iid* sample from a normal distribution $N(\mu, \sigma^2)$ we are given sample mean $\bar{X} = 11.56$, $n = 26$, sample standard deviation $S = 2.32$.

(a) Is there sufficient evidence to reject the null hypothesis $H_o : \mu = 12$ in favor of the two-sided alternative hypothesis $H_a : \mu \neq 12$ with a significance level of $\alpha = 0.05$?
(b) Calculate a confidence level $1 - \alpha = 0.95$ upper bound for σ.
(c) Using the upper bound for σ calculated in part (b) estimate the sample size needed to obtain a confidence interval for μ with a margin of error of 0.5. Assume the sample size will large. Use confidence level $1 - \alpha = 0.95$.

SOLUTION:

(a)

$$\begin{aligned} T &= \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \\ &= \frac{11.56 - 12}{2.32/\sqrt{26}} \\ &= -0.967. \end{aligned}$$

Note that $|T| < t_{25,0.025} = 2.06$, so do not reject H_o at $\alpha = 0.05$.

(b)

$$\begin{aligned} UB &= \frac{S}{\sqrt{\chi^2_{n-1,1-\alpha}/(n-1)}} \\ &= \frac{2.32}{\sqrt{14.61/25}} \\ &= 3.03. \end{aligned}$$

So,

$$\sigma < 3.03$$

is the 95% upper confidence bound for σ.

(c) Use estimate $\hat{\sigma} = 3.03$ in formula

$$n \approx \left(z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2 = \left(1.96 \times \frac{3.03}{0.5} \right)^2 = 141.51,$$

so round up to $n = 142$.

Problem 2.2 Suppose the true mean and standard deviation of a population of measurements is $\mu = 100.3$ and $\sigma = 7.6$. A random sample of $n = 20$ is collected.

(a) What is the standard deviation of the sample mean?
(b) What is the probability that the sample mean is within 4 units of the true mean?

SOLUTION:

(a) $\sigma_{\bar{x}} = \frac{\sigma}{n^{1/2}} = \frac{7.6}{20^{1/2}} \approx 1.7$.
(b) $P(|\bar{x} - \mu| < 4) = P\left(|Z| < \frac{4}{\sigma_{\bar{x}}}\right) \approx P(|Z| < 2.35) = 1 - 2P(Z < -2.35) = 1 - 2 \times 0.0094 = 0.9812$, using, for example, `R` command `pnorm(-2.35)` or statistical tables.

Problem 2.3

Suppose a random sample of $n = 7$ measurements is collected:

$$25.68, 34.50, 27.06, 22.75, 37.43, 32.66, 32.41.$$

Assume that they are from a normally distributed population.

(a) Construct a confidence level for the population mean using a 90% confidence level.
(b) Do a hypothesis test of $H_o : \mu = 25$ against $H_a : \mu \neq 25$. Report a P-value. Do you reject H_o at significance level $\alpha = 0.05$? What about significance level $\alpha = 0.01$?

SOLUTION:

(a) We have $\alpha = 0.1$, so we need critical value

$$t_{n-1,\alpha/2} = t_{6,0.05} = 1.943$$

from Table A.3, or `R` command `qt(1 - 0.05, df=6)`. We also have

$$\bar{x} = 30.36 \text{ and } S = 5.28.$$

The confidence interval is then

$$\begin{aligned} CI_{.90} &= \bar{x} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \\ &= 30.36 \pm 1.943 \frac{5.28}{\sqrt{7}} \\ &= 30.36 \pm 3.88 = (26.48, 34.23). \end{aligned}$$

(b) We have test statistic

$$\begin{aligned} T_{obs} &= \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \\ &= \frac{30.36 - 25}{5.28/\sqrt{7}} \\ &= 2.68. \end{aligned}$$

The P-value is $P = 0.037$, obtained either exactly by the `R` command `2*pt(-2.68,6)`, or approximated by Table A.3, noting that

$$2.447 = t_{6,0.025} < 2.68 < t_{6,0.01} = 3.143,$$

so that

$$0.02 < P < 0.05.$$

We reject at significance level $\alpha = 0.05$ but not $\alpha = 0.01$.

Problem 2.4 For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 20.292$, $n = 100$, standard deviation $\sigma = 0.34$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.

SOLUTION: We have $\alpha = 0.05$, so we need critical value

$$z_{\alpha/2} = z_{0.025} = 1.96,$$

giving level $1 - \alpha$ confidence interval

$$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 20.292 \pm 1.96 \times 0.034 \\ &= 20.292 \pm 0.0666 = (20.225, 20.359). \end{aligned}$$

Problem 2.5 For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 179.012$, $n = 7$, standard deviation $\sigma = 19.6$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.9$.

SOLUTION: We have $\alpha = 0.1$, so we need critical value

$$z_{\alpha/2} = z_{0.05} = 1.645,$$

giving level $1 - \alpha$ confidence interval

$$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 179.012 \pm 1.645 \times 7.408 \\ &= 179.012 \pm 12.185 = (166.827, 191.197). \end{aligned}$$

Problem 2.6 For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 2.349$, $n = 23$, standard deviation $\sigma = 0.03$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.99$.

SOLUTION: We have $\alpha = 0.01$, so we need critical value

$$z_{\alpha/2} = z_{0.005} = 2.576,$$

giving level $1 - \alpha$ confidence interval

$$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 2.349 \pm 2.576 \times 0.00626 \\ &= 2.349 \pm 0.0161 = (2.333, 2.365). \end{aligned}$$

Problem 2.7 We are given an *iid* sample from a normal distribution $N(\mu, \sigma^2)$.

$$103.58, 97.97, 104.98, 91.40, 98.24,$$

of sample size $n = 5$.

(a) Give an expression for the 75th percentile X_{75} of the normal distribution as a function of μ and σ.
(b) A simple method of estimating X_{75} would be to substitute the sample mean and variance $\bar{X}$ and S^2 for μ and σ^2 in the expression given in Part (a). However, it is decided to construct a conservative estimate $\hat{X}_{75}$ of X_{75}, in the sense that with large probability we expect $X_{75} \leq \hat{X}_{75}$. This is to be done by estimating μ and σ with level 0.95 upper confidence bounds. Carry out this procedure.

SOLUTION:

(a) The 75th percentile of the standard normal distribution is $z_{0.1} \approx 1.282$. So

$$X_{75} = \mu + z_{0.25}\sigma \approx \mu + 0.674 \times \sigma$$

(b) Directly from the data we have $\bar{X} = 99.234$, $S = 5.382$. The 0.90 upper confidence bound for μ is simply

$$\hat{\mu}_{95} = \bar{X} + t_{4,0.05}\frac{S}{\sqrt{5}} = 99.234 + 2.132 \times \frac{5.382}{\sqrt{5}} = 104.37.$$

The level $1 - \alpha$ upper bound for σ is given by ,

$$\sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{4,0.95} = 0.711$. The upper bound is then given by,

$$\hat{\sigma}_{95} = \frac{5.382}{\sqrt{0.711/4}} = 12.77.$$

This leads to

$$\begin{aligned} \hat{X}_{75} &= \hat{\mu}_{95} + z_{0.25} \times \hat{\sigma}_{95} \\ &= 104.37 + 0.674 \times 12.77 \\ &= 112.98. \end{aligned}$$

Problem 2.8 We are given two paired samples from normally distributed populations ($n = 5$). The data is summarized in the table below. Perform a two-sided hypothesis test using hypotheses $H_o : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. Use significance level $\alpha = 0.05$.

	Sample 1	Sample 2	Difference
1	139.6	146.4	-6.8
2	139.3	140.7	-1.4
3	141.4	142.6	-1.2
4	139.1	145.6	-6.5
5	136.1	145.4	-9.3

SOLUTION: We have

$$\begin{aligned} \bar{X}_1 &= 139.1 \\ \bar{X}_2 &= 144.14 \\ \bar{X}_1 - \bar{X}_2 &= -5.04 \\ S_D &= 3.584. \end{aligned}$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{S_D/\sqrt{n}} \\ &= \frac{-5.04}{3.584/\sqrt{5}} \\ &= -3.14. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{4,0.025} = 2.776.$$

Therefore, reject the null hypothesis at a significance level $\alpha = 0.05$.

Problem 2.9 We are given two independent samples from normally distributed populations. The data is summarized in the table below.

	Sample 1	Sample 2
$\bar{X}$	34.8	26.5
S	0.45	1.68
n	5	10

(a) Construct a 90% confidence interval for the difference in means $\mu_2 - \mu_1$. Use two procedures, assuming *i*) equal and *ii*) unequal variances (that is, the pooled procedure and Welch's procedure). How do the confidence intervals differ?

(b) Do a hypothesis test of $H_o : \mu_2 = \mu_1$ against $H_a : \mu_2 < \mu_1$. Again, use two procedures, assuming *i*) equal and *ii*) unequal variances. Report a P-value.

SOLUTION:

(a) *i*) Assuming equal variance, we have pooled sample variance

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(5-1)0.45^2 + (10-1)1.68^2}{5+10-2} = 2.02.$$

Then $\alpha = 0.1$, so the 90% confidence interval for $\mu_2 - \mu_1$ is

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 26.5 - 34.8 \pm 1.77\sqrt{2.02}\sqrt{\frac{1}{5} + \frac{1}{10}} \\ &= -8.30 \pm 1.38 = (-9.68, -6.92), \end{aligned}$$

given critical value $t_{13,0.05} = 1.77$.

ii) Assuming unequal variances, we have the degrees of freedom for Welch's procedure

$$\nu_W = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}} = \frac{\left(\frac{0.45^2}{5} + \frac{1.68^2}{10}\right)^2}{\frac{(0.45^2/5)^2}{5-1} + \frac{(1.68^2/10)^2}{10-1}} = 11.25,$$

then round down to $\nu_W = 11$. The 90% confidence interval for $\mu_2 - \mu_1$ is then

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{\nu_W,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= 26.5 - 34.8 \pm 1.80\sqrt{\frac{0.45^2}{5} + \frac{1.68^2}{10}} \\ &= -8.30 \pm 1.02 = (-9.32, -7.28), \end{aligned}$$

given critical value $t_{11,0.05} = 1.80$. The CI is of smaller width for Welch's procedure.

(b) *i*) Assuming equal variances, the test statistic is

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{26.5 - 34.8}{\sqrt{2.02}\sqrt{\frac{1}{5} + \frac{1}{10}}} \\ &= -10.66. \end{aligned}$$

With $n_1 + n_2 - 2 = 13$ degrees of freedom the P-value is

$$P = P(T < -17.1) = 4.3 \times 10^{-8},$$

so that the P-value is very small, and the evidence against H_o is very strong. If we use Table A.3, we note that $t_{13,0.00025} = 4.597$, so $P < 0.00025$.
ii) Assuming unequal variances, the test statistic is

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\ &= \frac{26.5 - 34.8}{\sqrt{\frac{0.45^2}{5} + \frac{1.68^2}{10}}} \\ &= -14.61. \end{aligned}$$

With $\nu = 11$ degrees of freedom the P-value is

$$P = P(T < -23.4) = 7.5 \times 10^{-9},$$

so that the P-value is very small, and the evidence against H_o is very strong. If we use Table A.3, we note that $t_{13,0.00025} = 4.863$, so $P < 0.00025$.

Problem 2.10 Ten subjects are placed on a diet with the objective of reducing sodium. Sodium levels are measured in mmol/L from a blood sample before and after the diet, and are summarized in the following table:

Table 2.1: Sodium level measurents (mmol/L) for sodium reducing diet study.

Subject	Before Diet	After Diet
1	146.6	148.6
2	150.7	142.7
3	151.9	133.9
4	155.4	146.8
5	159.6	147.6
6	160.1	148.4
7	146.7	137.3
8	160.1	158.5
9	133.8	119.8
10	160.1	149.5

(a) Construct a 95% level confidence interval for the mean change in sodium $\mu_2 - \mu_1$. If you were to test $H_o : \mu_2 = \mu_1$ against $H_a : \mu_2 \neq \mu_1$, would H_o be rejected with significance level $\alpha = 0.05$?
(b) Use the `R t.test()` function to do the lower-tailed test of $H_o : \mu_2 = \mu_1$ against $H_a : \mu_2 < \mu_1$. What is the P-value?

SOLUTION: We first calculate the differences, shown in the following table:

Subject	Before Diet	After Diet	D_i
1	146.6	148.6	2.0
2	150.7	142.7	-8.0
3	151.9	133.9	-18.0
4	155.4	146.8	-8.6
5	159.6	147.6	-12.0
6	160.1	148.4	-11.7
7	146.7	137.3	-9.4
8	160.1	158.5	-1.6
9	133.8	119.8	-14.0
10	160.1	149.5	-10.6

(a) The sample mean and standard deviation of the differences are

$$\bar{D} = -9.19 \text{ and } S_D = 5.79.$$

A 95% CI for $\mu_2 - \mu_1$ is given by

$$\begin{aligned} CI_{1-\alpha} &= \bar{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}} \\ &= -9.19 \pm 2.262\frac{5.79}{\sqrt{10}} \\ &= -9.19 \pm 4.14 = (-13.33, -5.05), \end{aligned}$$

given critical value $t_{9,0.025} = 2.262$. The CI interval does not contain 0, so a two-sided hypothesis test against $H_o : \mu_1 = \mu_2$ would be rejected with significance level $\alpha = 0.05$.

(b) The following `R` script gives the required test

```
> x = c(146.6,150.7,151.9,155.4,159.6,160.1,146.7,160.1,133.8,160.1)
> y = c(148.6, 142.7, 133.9, 146.8, 147.6, 148.4, 137.3, 158.5, 119.8, 149.5)
> t.test(x,y,paired=T,alternative='greater')

Paired t-test

data:  x and y
t = 5.0205, df = 9, p-value = 0.0003593
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
5.834469      Inf
sample estimates:
mean of the differences
9.19
```

Note that the `paired` option is set to `T`, and the `alternative` option is set to `'greater'`, since `t.test()` tests difference $\mu_1 - \mu_2$ and not $\mu_2 - \mu_1$. The P-value is 0.0003593.

Problem 2.11 We are given two paired samples from normally distributed populations ($n = 5$). The data is summarized in the table below. Perform a two-sided hypothesis test for null hypothesis $H_o : \mu_2 - \mu_1 = 0$ against alternative $H_a : \mu_2 - \mu_1 \neq 0$. Use significance level $\alpha = 0.1$.

	Sample 1	Sample 2	Difference $(X_2 - X_1)$
1	13.3	6.6	-6.7
2	14.1	15.4	1.3
3	14.7	6.5	-8.2
4	13.0	5.9	-7.1
5	12.0	3.4	-8.6

SOLUTION: We have the summary statistics:

$$\bar{X}_1 = 13.42, \bar{X}_2 = 7.56, \bar{X}_2 - \bar{X}_1 = -5.86, S_D = 4.077.$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_2 - \bar{X}_1}{S_D/\sqrt{n}} \\ &= \frac{-5.86}{4.077/\sqrt{5}} \\ &= -3.214. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{4,0.05} = 2.132.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.1$.

Problem 2.12 We are given two paired samples from normally distributed populations ($n = 6$). The data is summarized in the table below. Construct level $1 - \alpha = 0.9$ confidence interval for $\mu_1 - \mu_2$.

	Sample 1	Sample 2	Difference
1	46.993	34.846	12.147
2	44.241	45.226	-0.985
3	48.334	45.171	3.163
4	45.816	48.341	-2.525
5	46.837	39.383	7.454
6	48.748	56.129	-7.381

SOLUTION: We have the summary statistics:

$$\bar{X}_1 = 46.828, \quad \bar{X}_2 = 44.849, \quad \bar{X}_D = \bar{X}_1 - \bar{X}_2 = 1.979, \quad S_D = 7.091.$$

Then the CI is given by

$$\begin{aligned} CI &= \bar{X}_D \pm t_{n-1,\alpha/2} \times \frac{S_D}{\sqrt{n}} \\ &= 1.979 \pm 2.015 \times \frac{7.091}{\sqrt{6}} \\ &= 1.979 \pm 5.834 = (-3.85, 7.813). \end{aligned}$$

Problem 2.13 We are given two paired samples from normally distributed populations ($n = 6$). The data is summarized in the table below. Perform a two-sided hypothesis test for a difference in mean, using hypotheses $H_o : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. Use significance level $\alpha = 0.05$.

	Sample 1	Sample 2	Difference
1	11.709	10.699	1.010
2	11.894	10.488	1.406
3	11.427	10.342	1.085
4	11.971	10.558	1.413
5	11.818	10.903	0.915
6	12.348	10.871	1.477

SOLUTION: We have the summary statistics:

$$\bar{X}_1 = 11.861, \bar{X}_2 = 10.643, S_D = 0.242.$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{S_D/\sqrt{n}} \\ &= \frac{1.218}{0.242/\sqrt{6}} \\ &= 12.317. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{5,0.025} = 2.571.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 6.246e-05).

Problem 2.14 We are given two independent samples from normally distributed populations $N(\mu_i, \sigma_i^2)$, $i = 1, 2$. The data is summarized in the following table:

	Sample $i = 1$	Sample $i = 2$
$\bar{X}_i$	43.96	48.98
S_i	7.62	9.42
n_i	23	54

(a) Perform a two-sided hypothesis test for null hypothesis $H_o : \sigma_1^2 = \sigma_2^2$ against alternative $H_a : \sigma_1^2 \neq \sigma_2^2$. Use significance level $\alpha = 0.05$. You can make use of critical values $F_{0.975,22,53} = 0.463$ and $F_{0.025,22,53} = 1.943$.

(b) Construct a level $1 - \alpha = 0.9$ confidence interval for $\mu_2 - \mu_1$. Use the conclusion of part (a) to choose between the pooled procedure for equal variances or Welch's procedure for unequal variances.

SOLUTION:

(a) Use statistic

$$F = \frac{S_1^2}{S_2^2} = \frac{7.62^2}{9.42^2} = 0.654.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F \leq F_{1-\alpha/2,n_1-1,n_2-1} = 0.463 \text{ or } F \geq F_{\alpha/2,n_1-1,n_2-1} = 1.943.$$

Therefore, do not reject the null hypothesis of equal variances at a signficance level $\alpha = 0.05$.

(b) Use the pooled procedure with $\nu = n_1 + n_2 - 2 = 75$ degrees of freedom. Pooled variance is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{22 \times 7.62^2 + 53 \times 9.42^2}{75} = 79.74.$$

The confidence interval is

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 48.98 - 43.96 \pm 1.665 \times 8.93 \sqrt{\frac{1}{23} + \frac{1}{54}} \\ &= 5.02 \pm 3.70 \\ &= (1.32, 8.72). \end{aligned}$$

Problem 2.15 We are given two paired samples from normally distributed populations ($n = 5$). The data is summarized in the table below. Perform a two-sided hypothesis test using hypotheses $H_o : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. Use significance level $\alpha = 0.1$.

	Sample 1	Sample 2	Difference
1	10.94	9.49	1.45
2	13.74	12.86	0.88
3	14.27	11.10	3.17
4	14.30	10.84	3.46
5	13.58	10.09	3.49

SOLUTION: From the table we have:

$$\bar{X}_1 = 13.365,\ \bar{X}_2 = 10.875,\ \bar{X}_2 - \bar{X}_1 = 2.49,\ S_D = 1.233.$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_2 - \bar{X}_1}{S_D/\sqrt{n}} \\ &= \frac{2.49}{1.233/\sqrt{5}} \\ &= 4.517. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{4,0.05} = 2.132.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.1$.

Problem 2.16 Two types of fish attractors, one made from vitrified clay pipes, and the other from cement blocks and brush, were used during 16 different time periods spanning four years at Lake Tohopekaliga, Florida. The following observations are of the average number of fish caught per fishing day.

Table 2.2: Fish caught per day, by time period and fish attractors.

Time Periods	Pipe	Brush
1	6.64	9.73
2	7.89	8.21
3	1.83	2.17
4	0.42	0.75
5	0.85	1.61
6	0.29	0.75
7	0.57	0.83
8	0.63	0.56
9	0.32	0.76
10	0.37	0.32
11	0.00	0.48
12	0.11	0.52
13	4.86	5.38
14	1.80	2.33
15	0.23	0.91
16	0.58	0.79

(a) Use the `R t.test()` function to do a paired t-test to determine whether or not one attractor is more effective. Use level $\alpha = 0.01$

(b) Repeat part (a), but assume the samples are independent. Use procedures for both equal and unequal variances. Does your conclusion change?

SOLUTION:

(a) The test is implemented as follows, specifying the option `paired=T`, noting that a two-sided test against equality of means is the default option.

```
> x = c(6.64,7.89,1.83,0.42,0.85,0.29,0.57,0.63,0.32,0.37,0.00,0.11,4.86,1.80,0.23,0.58)
> y = c(9.73,8.21,2.17,0.75,1.61,0.75,0.83,0.56,0.76,0.32,0.48,0.52,5.38,2.33,0.91,0.79)
>
> t.test(x,y,paired=T)

Paired t-test

data:  x and y
t = -3.0496, df = 15, p-value = 0.00811
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.924855 -0.163895
sample estimates:
mean of the differences
-0.544375

>
```

The t-statistic is $T = -3.0496$ with 15 degrees of freedom, giving $P = 0.00811$. Since $P < \alpha = 0.01$, we reject the null hypothesis of the equality of the mean number of fish caught per day for the two attractors.

(b) We then remove the `paired=T` option, and repeat the test using both option settings `var.equal=T` and `var.equal=F`, for the pooled procedure and the Welch test, respectively.

```
> t.test(x,y,var.equal=T)

Two Sample t-test

data:  x and y
t = -0.56979, df = 30, p-value = 0.5731
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.495559  1.406809
sample estimates:
mean of x mean of y
1.711875  2.256250

> t.test(x,y,var.equal=F)

Welch Two Sample t-test

data:  x and y
t = -0.56979, df = 29.251, p-value = 0.5732
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.497656  1.408906
sample estimates:
mean of x mean of y
1.711875  2.256250

>
```

In both cases $P > 0.5$, so we do not reject the null hypothesis, apparently contradicting the paired procedure. The paired procedure is the correct approach. The larger p-values for the two-sample procedure is due the fact that variation *within* the two samples contributes significantly to the standard error, which measures statistical uncertainty. The pairing procedure removes within-sample variation.

2.2 Inference for Proportions

Problem 2.17 Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 21$, with sample size $n = 80$.

(a) Construct a level $1 - \alpha = 0.95$ confidence interval for p. Use the normal approximation.
(b) Test hypothesis $H_o : p \geq 0.4$ against $H_a : p < 0.4$. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$? Use a normal approximation with a continuity correction

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 21$ and $n = 80$ is

$$\hat{p} = \frac{X}{n} = \frac{21}{80} = 0.262.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.262 \pm 1.96\sqrt{\frac{0.262(1-0.262)}{80}} \\ &= 0.262 \pm 1.96 \times 0.0492 \\ &= 0.262 \pm 0.0964 \end{aligned}$$

or equivalently, $CI = (0.166, 0.359)$.

(b) To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{X - np_0 + 0.5}{\sqrt{np_0(1-p_0)}} = \frac{21 - 32 + 0.5}{\sqrt{80 \times 0.4(1-0.4)}} = -2.396.$$

The critical value is $-z_{.05} = -1.645$. Since $Z'_{obs} < -z_{.05}$ we reject the null hypothesis at an $\alpha = 0.05$ significance level.

Problem 2.18 Suppose we are given the following contingency table:

	Not Vaccinated	Vaccinated	Total
Infection Occurs	34	7	41
No Infection Occurs	933	647	1580
Total	967	654	1621

Construct a level $1-\alpha = 0.95$ confidence interval for the log odds ratio of [Infection Occurs] between groups [Not Vaccinated] and [Vaccinated]. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at a significance level of $\alpha = 0.05$?

SOLUTION: The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{34 \times 647}{7 \times 34} = 3.368.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{34} + \frac{1}{7} + \frac{1}{933} + \frac{1}{647}} \\ &= 0.418. \end{aligned}$$

The level $1-\alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.214 \pm 1.96 \times 0.418 \\ &= 1.214 \pm 0.82 \end{aligned}$$

or equivalently, $CI = (0.395, 2.034)$. Since the CI does not contain 0 we reject the null hypothesis at an α significance level.

Problem 2.19 Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 57$, with sample size $n = 245$.

(a) Construct a level 0.95 confidence interval for p.
(b) Test hypothesis $H_o : p = 0.1$ against $H_a : p \neq 0.1$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.01$? Do the test twice, without and with the continuity correction.
(c) Use `R` function `prop.test()` to verify your answers.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 57$ and $n = 245$ is

$$\hat{p} = \frac{X}{n} = \frac{57}{245} = 0.233.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.233 \pm 1.96\sqrt{\frac{0.233(1-0.233)}{245}} \\ &= 0.233 \pm 1.96 \times 0.027 \\ &= 0.233 \pm 0.0529 \end{aligned}$$

or equivalently, $CI = (0.18, 0.286)$.

(b) We use two methods:
[Without Continuity Correction] The test statistic is

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.233 - 0.1}{\sqrt{\frac{0.1(1-0.1)}{245}}} = 6.921.$$

The P-value is, for $Z \sim N(0, 1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 6.921) = 4.48e - 12.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.

[With Continuity Correction] To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction for the evaluation of a two-sided P-value $\alpha_{obs} = 2P(Z > |Z_{obs}|)$ can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{|X - np_0| - 0.5}{\sqrt{np_0(1-p_0)}} = \frac{|57 - 24.5| - 0.5}{\sqrt{245 \times 0.1(1-0.1)}} = 6.815.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z'_{obs}|) = 2P(Z > 6.815) = 9.45e - 12.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.

(c) The script and output are given below. Note that `prop.test` uses a different method for calculating confidence intervals for single proportions. The equation is the *Wilson score method*, with and without continuity correction, and can be found in (methods 3 and 4 in Section 2):

> Newcombe, R.G. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.

Also note that the χ^2 statistic reported by `prop.test` (labeled `X-squared` in the output) is equal to Z^2 where Z is the Z statistic. So without continuity correction we have $Z = 6.92$, $Z^2 = 47.9$, which is also given by `prop.test` (`X-squared = 47.902`). We can get the exact P-value by refering to the `p.value` element of the list object output by `prop.test`. We get the same value $4.479514e-12$, within rounding error, reported above. Similarly, with continuity correction we get $Z' = 6.815$, $Z'^2 = 46.44$, which is also given by `prop.test` (`X-squared = 46.44`). We also get the P-value $9.447161e - 12$ which is, within rounding error, reported above.

```
>  prop.test(57,245,p=0.1,correct=F)

1-sample proportions test without continuity correction

data:  57 out of 245, null probability 0.1
X-squared = 47.902, df = 1, p-value = 4.48e-12
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
0.1841209 0.2894395
sample estimates:
p
0.2326531
>
>  prop.test(57,245,p=0.1,correct=F)$p.value
[1] 4.479514e-12
>
>  prop.test(57,245,p=0.1,correct=T)

1-sample proportions test with continuity correction

data:  57 out of 245, null probability 0.1
```

```
X-squared = 46.44, df = 1, p-value = 9.447e-12
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
0.1822699 0.2916054
sample estimates:
p
0.2326531
>
>  prop.test(57,245,p=0.1,correct=T)$p.value
[1] 9.447161e-12
```

Problem 2.20 Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 15$, with sample size $n = 34$.

(a) Construct a level 0.95 confidence interval for p.
(b) What sample size would be needed to reduce the margin of error to 0.075? Make sure the sample size would be sufficient for all values of p.
(c) Test hypothesis $H_o : p \leq 0.6$ against $H_a : p > 0.6$. Is the null hypothesis rejected at a significance level of $\alpha = 0.01$? Use a normal approximation with a continuity correction to estimate to evaluate significance.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 15$ and $n = 34$ is

$$\hat{p} = \frac{X}{n} = \frac{15}{34} = 0.441.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.441 \pm 1.96\sqrt{\frac{0.441(1-0.441)}{34}} \\ &= 0.441 \pm 1.96 \times 0.0852 \\ &= 0.441 \pm 0.167 \end{aligned}$$

or equivalently, $CI = (0.274, 0.608)$.

(b) The widest confidence interval is obtained for $p = 0.5$, so we must anticipate $\hat{p} = 0.5$. In this case

$$n = p^*(1-p^*)\left(\frac{z_{0.025}}{ME}\right)^2 = (1/4)\left(\frac{1.96}{0.075}\right)^2 = 170.7378.$$

Round up to $n = 171$.

(c) To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction for the evaluation of an upper-tailed alternative can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{X - np_0 - 0.5}{\sqrt{np_0(1-p_0)}} = \frac{15 - 20.4 - 0.5}{\sqrt{34 \times 0.6(1-0.6)}} = -2.07.$$

Reject H_o if $Z_{obs} > z_\alpha = z_{0.01} = 2.326$. Therefore, we do not reject the null hypothesis at an $\alpha = 0.01$ significance level.

Problem 2.21 In a sample of $n = 80$ randomly selected men, $X = 25$ were observed to be left-handed.

(a) Construct a level 0.95 confidence interval for the proportion of left-handedness p.
(b) Suppose it is conjectured that the proportion of left-handedness among males is 20%. Test hypothesis $H_o : p = 0.2$ against $H_a : p \neq 0.2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.01$? Do the test twice, without and with the continuity correction.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 25$ and $n = 80$ is

$$\hat{p} = \frac{X}{n} = \frac{25}{80} = 0.312.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.312 \pm 1.96\sqrt{\frac{0.312(1-0.312)}{80}} \\ &= 0.312 \pm 1.96 \times 0.0518 \\ &= 0.312 \pm 0.102 \end{aligned}$$

or equivalently, $CI = (0.211, 0.414)$.

(b) We use two methods:
[Without Continuity Correction] The test statistic is

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.312 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{80}}} = 2.516.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 2.516) = 0.0119.$$

Since $\alpha_{obs} > \alpha$ we do no reject the null hypothesis at an α significance level.

[With Continuity Correction] To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then introducing the continuity correction into the evaluation of a two-sided P-value $\alpha_{obs} = 2P(Z > |Z_{obs}|)$ can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{|X - np_0| - 0.5}{\sqrt{np_0(1-p_0)}} = \frac{|25 - 16| - 0.5}{\sqrt{80 \times 0.2(1-0.2)}} = 2.376.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z'_{obs}|) = 2P(Z > 2.376) = 0.0175.$$

Since $\alpha_{obs} > \alpha$ we do no reject the null hypothesis at an α significance level.

Problem 2.22 Suppose a binomial random variable $X \sim bin(n,p)$ is observed to be $X = 32$, with sample size $n = 50$.

(a) Construct a level 0.95 confidence interval for p.
(b) Test hypothesis $H_o : p \geq 0.75$ against $H_a : p < 0.75$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.01$? Do the test twice, without and with the continuity correction.
(c) Use `R` function `prop.test()` to verify your answers.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 32$ and $n = 50$ is

$$\hat{p} = \frac{X}{n} = \frac{32}{50} = 0.64.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.64 \pm 1.96\sqrt{\frac{0.64(1-0.64)}{50}} \\ &= 0.64 \pm 1.96 \times 0.0679 \\ &= 0.64 \pm 0.133 \end{aligned}$$

or equivalently, $CI = (0.507, 0.773)$.

(b) We use two methods:
[Without Continuity Correction] The test statistic is

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.64 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{50}}} = -1.8.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = P(Z < Z_{obs}) = P(Z < -1.8) = 0.0362.$$

Since $\alpha_{obs} > \alpha$ we do no reject the null hypothesis at an α significance level.

[With Continuity Correction] To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction for the evaluation of a lower-tailed P-value $\alpha_{obs} = P(Z < Z_{obs})$ can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{X - np_0 + 0.5}{\sqrt{np_0(1-p_0)}} = \frac{32 - 37.5 + 0.5}{\sqrt{50 \times 0.75(1-0.75)}} = -1.63.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = P(Z < Z'_{obs}) = P(Z < -1.63) = 0.0512.$$

Since $\alpha_{obs} > \alpha$ we do no reject the null hypothesis at an α significance level.

(c) The script and output are given below. Note that `prop.test` uses a different method for calculating confidence intervals for single proportions. The equation is the *Wilson score method*, with and without continuity correction, and can be found in (methods 3 and 4 in Section 2):

> Newcombe, R.G. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.

Also note that the χ^2 statistic reported by `prop.test` (labeled `X-squared` in the output) is equal to Z^2 where Z is the Z statistic. So without continuity correction we have $Z = -1.8$, $Z^2 = 3.24$, which is also given by `prop.test` within rounding error (`X-squared = 3.2267`). The reported P-value = 0.03622 is the same as that obtained above ($P = 0.0362$). Similarly, with continuity correction we get $Z' = -1.63$, $Z'^2 = 2.6569$, which is also given by `prop.test` (`X-squared = 2.6667`). We also get the P-value 0.05124 which is, within rounding error, reported above ($P = 0.0512$).

```
> prop.test(32,50,0.75,alternative='less',correct=F)

1-sample proportions test without continuity correction

data:  32 out of 50, null probability 0.75
X-squared = 3.2267, df = 1, p-value = 0.03622
alternative hypothesis: true p is less than 0.75
95 percent confidence interval:
0.0000000 0.7418033
sample estimates:
p
0.64

> prop.test(32,50,0.75,alternative='less',correct=T)

1-sample proportions test with continuity correction

data:  32 out of 50, null probability 0.75
X-squared = 2.6667, df = 1, p-value = 0.05124
alternative hypothesis: true p is less than 0.75
95 percent confidence interval:
0.0000000 0.7506402
sample estimates:
p
0.64

>
```

Problem 2.23 A drug designed to cure Hepatitis-C is tested on 200 subjects. Half are treated with the drug, and half are given conventional treatment (ie. form a control group). Suppose the cure rates are reported in following contingency table:

	Treatment	Control	
Cured	76	34	110
Not Cured	24	66	90
Total	100	100	200

Construct a level 0.99 confidence interval for the log odds ratio of Cure events between groups Treatment and Control. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.01$?

SOLUTION:

(a) The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{76 \times 66}{34 \times 76} = 6.147.$$

We use critical value

$$z_{\alpha/2} = z_{0.005} = 2.576.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{76} + \frac{1}{34} + \frac{1}{24} + \frac{1}{66}} \\ &= 0.315. \end{aligned}$$

The level $1-\alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.816 \pm 2.576 \times 0.315 \\ &= 1.816 \pm 0.812 \end{aligned}$$

or equivalently, $CI = (1.004, 2.628)$. Since the CI does not contain 0 we reject the null hypothesis at an α significance level.

Problem 2.24 Repeat Prolem 2.23 using a difference of proportions procedures. Let p_1, p_2 be the cure rates for the treatment and control groups respectively. We then have binomial random variables $X_i \sim bin(n_i, p_i)$, $i = 1, 2$, which are observed to be $X_1 = 76$ and $X_2 = 34$, with sample sizes $n_1 = 100$ and $n_2 = 100$

(a) Construct a level 0.99 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.01$?

SOLUTION:

(a) The level $1-\alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.005} = 2.576.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{76}{100}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{34}{100}, \quad \hat{p}_2 - \hat{p}_1 = -0.42.$$

The confidence interval is then given by

$$\begin{aligned} CI &= -0.42 \pm 2.576\sqrt{\frac{0.76(1-0.76)}{100} + \frac{0.34(1-0.34)}{100}} \\ &= -0.42 \pm 2.576 \times 0.0638 \\ &= -0.42 \pm 0.164 \end{aligned}$$

or equivalently, $CI = (-0.584, -0.256)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{76 + 34}{100 + 100} = 0.55.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.34 - 0.76}{\sqrt{0.55(1-0.55)\left(\frac{1}{100} + \frac{1}{100}\right)}} \\ &= \frac{-0.42}{0.0704} \\ &= -5.97 \end{aligned}$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 5.97) = 2.38e-09.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.

Problem 2.25 In a research study, $n_1 = 56$ nonsmokers and $n_2 = 35$ smokers were asked to complete a specific excercise program. It was observed that $X_1 = 20$ nonsmokers and $X_2 = 7$ smokers were able to complete the program within a prescribed time limit of 1/2 hour. Let p_1, p_2 be the respective population proportions able to complete the program.

(a) Construct a level 0.95 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?
(c) Verify your results using the `R` function `prop.test`. Make sure the `correct` option is set to `FALSE`.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{20}{56}, \;\; \hat{p}_2 = \frac{X_2}{n_2} = \frac{7}{35}, \;\; \hat{p}_2 - \hat{p}_2 = -0.157.$$

The confidence interval is then given by

$$\begin{aligned} CI &= -0.157 \pm 1.96\sqrt{\frac{0.357(1-0.357)}{56} + \frac{0.2(1-0.2)}{35}} \\ &= -0.157 \pm 1.96 \times 0.0931 \\ &= -0.157 \pm 0.183 \end{aligned}$$

or equivalently, $CI = (-0.34, 0.0254)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{20 + 7}{56 + 35} = 0.297.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.2 - 0.357}{\sqrt{0.297(1-0.297)\left(\frac{1}{56} + \frac{1}{35}\right)}} \\ &= \frac{-0.157}{0.0984} \\ &= -1.6 \end{aligned}$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 1.597) = 0.11.$$

Since $\alpha_{obs} > \alpha$ we do not reject the null hypothesis at an α significance level.

(c) The following code gives the required confidence interval and hypothesis test, which conforms to the previous calculations.

```
> prop.test(c(20,7),c(56,35),correct=F)

2-sample test for equality of proportions without continuity correction

data:  c(20, 7) out of c(56, 35)
X-squared = 2.5488, df = 1, p-value = 0.1104
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
-0.02536845  0.33965416
sample estimates:
prop 1    prop 2
0.3571429 0.2000000
```

Problem 2.26 Medical records were used to obtain infection rates of children who have been vaccinated, and who have not been vaccinated for a certain strain of flu. The results are given in the table below.

	Not Vaccinated	Vaccinated	
Infection occurs	32	27	59
No infection occurs	988	2993	3981
Total	1020	3020	4040

Construct a level 0.95 confidence interval for the log odds ratio of Infection occurs between Not Vaccinated and Vaccinated groups. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$? Can we conclude that vaccinated children are less likely to be infected?

SOLUTION:

(a) The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{32 \times 2993}{27 \times 32} = 3.59.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{32} + \frac{1}{27} + \frac{1}{988} + \frac{1}{2993}} \\ &= 0.264. \end{aligned}$$

The level $1 - \alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.278 \pm 1.96 \times 0.264 \\ &= 1.278 \pm 0.517 \end{aligned}$$

or equivalently, $CI = (0.761, 1.795)$. Since the CI does not contain 0 we reject the null hypothesis at an α significance level. Since the odds ratio is greater than one, we can conclude that vaccinated children are less likely to be infected.

Problem 2.27 A study examined $n = 969$ male heart attack patients. Each subject was classified as having normal blood pressure (NBP) or high blood pressure (HBP). In each addition, each subject was observed for two years following the initial heart attack. Of 495 subjects with NBP 123 experienced a recurrent heart attack, while out of 474 subjects with HBP, 145 experienced a recurrent heart attack.

	NBP	HBP
Recurrence	123	145
No recurrence	372	329
n_i	495	474
$\hat{p}_i$	0.248	0.306

(a) Construct a confidence interval for the difference in recurrence rates between the NBP and HBP subjects. Use confidence level $1 - \alpha = 0.95$.

(b) Construct a confidence interval for the log odds ratio:

$$\log\left[\frac{Odds(Recurrence \mid NBP)}{Odds(Recurrence \mid HBP)}\right].$$

Use confidence level $1 - \alpha = 0.95$.

(c) What can be concluded from parts (a) and (b) as to whether or not recurrence rates differ between the NBP and HBP groups? Are the conclusions from parts (a) and (b) consistent?

SOLUTION:

(a) We have estimates $\hat{p}_1 = 0.248$, $\hat{p}_2 = 0.306$. The confidence interval is then

$$\begin{aligned} CI_{1-\alpha} &= \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.248 - 0.306 \pm 1.96\sqrt{\frac{0.248(1-0.248)}{495} + \frac{0.306(1-0.306)}{474}} \\ &= -0.0574 \pm 0.0563 = (-0.114, -0.00112). \end{aligned}$$

(b) The estimate of the OR is

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 123 \times 329/(145 \times 372) = 0.750.$$

The standard error is

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{123} + \frac{1}{145} + \frac{1}{372} + \frac{1}{329}} = 0.1440636.$$

An approximate $(1-\alpha)100\%$ confidence interval for $\log(OR)$ is therefore

$$\begin{aligned} CI_{1-\alpha} &= \log(\hat{OR}) \pm z_{\alpha/2}SE(\log(OR)) \\ &= \log(0.750) \pm 1.96 \times 0.1441 \\ &= -0.2874 \pm 0.2824 = (-0.5698, -0.005). \end{aligned}$$

(c) From part (a), the CI for $p_1 - p_2$ is entirely negative. From part (b) the CI for OR is also entirely negative. Both conclusions imply $p_1 < p_2$, and so are consistent.

Problem 2.28 Suppose a study on treatments for migraine headaches compares two preventative treatment regimens. The *standard treatment* relies only on preventative medication. The *enhanced treatment* is a combination of preventative medication and muscle relaxation exercises. Of $N_e = 1045$ subjects given the *enhanced treatment* 100 report going 3 months without a migraine headache. Of $N_s = 687$ subjects given the *standard treatment* 45 report going 3 months without a migraine headache.

Construct a level 0.95 confidence interval for the log odds ratio of the outcome

$$O_+ = \{ \text{ 3 months without a migraine headache } \}$$

between the *enhanced treatment* group and the *standard treatment* group. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$?

SOLUTION: We are given the following contingency table:

	Enhanced Treatment	Standard Treatment	
3 months without a migraine headache	100	45	145
$<$ 3 months without a migraine headache	945	642	1587
Total	1045	687	1732

The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{100 \times 642}{45 \times 100} = 1.51.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{100} + \frac{1}{45} + \frac{1}{945} + \frac{1}{642}} \\ &= 0.187. \end{aligned}$$

The level $1 - \alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 0.412 \pm 1.96 \times 0.187 \\ &= 0.412 \pm 0.366 \end{aligned}$$

or equivalently, $CI = (0.0461, 0.778)$. Since the CI does not contain 0 we reject the null hypothesis at an α significance level.

Problem 2.29 Suppose independent binomial random variables $X_i \sim bin(n_i, p_i)$, $i = 1, 2$ are observed to be $X_1 = 154$ and $X_2 = 96$, with sample sizes $n_1 = 340$ and $n_2 = 120$

(a) Construct a level 0.95 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{154}{340}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{96}{120}, \quad \hat{p}_2 - \hat{p}_1 = 0.347.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.347 \pm 1.96\sqrt{\frac{0.453(1-0.453)}{340} + \frac{0.8(1-0.8)}{120}} \\ &= 0.347 \pm 1.96 \times 0.0454 \\ &= 0.347 \pm 0.089 \end{aligned}$$

or equivalently, $CI = (0.258, 0.436)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{154 + 96}{340 + 120} = 0.543.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.8 - 0.453}{\sqrt{0.543(1-0.543)\left(\frac{1}{340} + \frac{1}{120}\right)}} \\ &= \frac{0.347}{0.0529} \\ &= 6.562. \end{aligned}$$

Reject if $|Z_{obs}| > z_{\alpha/2} = 1.96$. Therefore, reject the null hypothesis at an $\alpha = 0.05$ significance level.

Problem 2.30 Suppose independent binomial random variables $X_i \sim bin(n_i, p_i)$, $i = 1, 2$ are observed to be $X_1 = 24$ and $X_2 = 45$, with sample sizes $n_1 = 85$ and $n_2 = 105$

(a) Construct a level 0.95 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?
(c) Use `R` function `prop.test()` to verify your answers.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{24}{85}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{45}{105}, \quad \hat{p}_2 - \hat{p}_1 = 0.146.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.146 \pm 1.96\sqrt{\frac{0.282(1-0.282)}{85} + \frac{0.429(1-0.429)}{105}} \\ &= 0.146 \pm 1.96 \times 0.0687 \\ &= 0.146 \pm 0.135 \end{aligned}$$

or equivalently, $CI = (0.0116, 0.281)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{24 + 45}{85 + 105} = 0.363.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.429 - 0.282}{\sqrt{0.363(1-0.363)\left(\frac{1}{85} + \frac{1}{105}\right)}} \\ &= \frac{0.146}{0.0702} \\ &= 2.084 \end{aligned}$$

The P-value is, for $Z \sim N(0, 1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 2.084) = 0.0372.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.

(c) The χ^2 statistic reported by `prop.test` (labeled `X-squared` in the output) is equal to Z^2 where Z is the Z statistic. So without continuity correction we have $Z = 2.084$, $Z^2 = 4.34$, which is also given by `prop.test`, within rounding error (`X-squared = 4.3424`). We report a P-value of 0.0372, while `prop.test` reports 0.03717, the same within rounding error. The confidence interval reported by `prop.test` is (-0.28081881, -0.01161817), the same reported above after reversing the labels.

```
> prop.test(c(24,45),c(85,105),correct = F)

2-sample test for equality of proportions without continuity
correction

data:  c(24, 45) out of c(85, 105)
X-squared = 4.3424, df = 1, p-value = 0.03717
alternative hypothesis: two.sided
95 percent confidence interval:
-0.28081881 -0.01161817
sample estimates:
prop 1    prop 2
0.2823529 0.4285714
```

Problem 2.31 Suppose independent binomial random variables $X_i \sim bin(n_i, p_i)$, $i = 1, 2$ are observed to be $X_1 = 15$ and $X_2 = 8$, with sample sizes $n_1 = 37$ and $n_2 = 25$

(a) Construct a level 0.9 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 \leq p_2$ against $H_a : p_1 > p_2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.05} = 1.645.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{15}{37}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{8}{25}, \quad \hat{p}_2 - \hat{p}_1 = -0.0854.$$

The confidence interval is then given by

$$\begin{aligned} CI &= -0.0854 \pm 1.645\sqrt{\frac{0.405(1-0.405)}{37} + \frac{0.32(1-0.32)}{25}} \\ &= -0.0854 \pm 1.645 \times 0.123 \\ &= -0.0854 \pm 0.203 \end{aligned}$$

or equivalently, $CI = (-0.288, 0.118)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{15 + 8}{37 + 25} = 0.371.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.32 - 0.405}{\sqrt{0.371(1-0.371)\left(\frac{1}{37} + \frac{1}{25}\right)}} \\ &= \frac{-0.0854}{0.125} \\ &= -0.683 \end{aligned}$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = P(Z < Z_{obs}) = P(Z < -0.683) = 0.247.$$

Since $\alpha_{obs} > \alpha$ we do not reject the null hypothesis at an α significance level.

Problem 2.32 Suppose we are given the following contingency table, summarizing the infection history of $n = 1621$ subjects.

	Not Vaccinated	Vaccinated	
Infection occurs	34	7	41
No infection occurs	933	647	1580
Total	967	654	1621

Construct a level $1 - \alpha = 0.95$ confidence interval for the log odds ratio for infection occurence between groups Not Vaccinated and Vaccinated. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$?

SOLUTION: The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{34 \times 647}{933 \times 7} = 3.368.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{34} + \frac{1}{7} + \frac{1}{933} + \frac{1}{647}} \\ &= 0.418. \end{aligned}$$

The level $1-\alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.214 \pm 1.96 \times 0.418 \\ &= 1.214 \pm 0.82 \\ &= (0.394, 2.034) \end{aligned}$$

or equivalently, $CI = (0.395, 2.034)$. Since the CI does not contain 0 we reject the null hypothesis at an $\alpha = 0.05$ significance level.

Problem 2.33 Three independent poll samples estimate popular support p for a certain candidate. The estimates and sample sizes are given in the following table.

Poll i	$\hat{p}_i$	n_i
1	0.45	200
2	0.48	1300
3	0.42	750

(a) Suppose we construct a pooled estimate of p by taking the weighted average

$$\hat{p}_{pooled} = \frac{\sum_{i=1}^{3} n_i \hat{p}_i}{\sum_{i=1}^{3} n_i}.$$

Calculate $\hat{p}_{pooled}$ and estimate its standard deviation (use the value of $\hat{p}_{pooled}$ in your estimate).

(b) Suppose we use the following alternative method of constructing a pooled estimate:

$$\hat{p}^*_{pooled} = \frac{\sum_{i=1}^{3} \hat{p}_i}{3}.$$

Calculate $\hat{p}^*_{pooled}$ and estimate its standard deviation (use the value of $\hat{p}^*_{pooled}$ in your estimate).

(c) Which pooled estimator is more accurate?

SOLUTION: Note that in each case $\hat{p}_i = X_i/n_i$ where $X_i \sim bin(n_i, p)$.

(a) First note that if $X = \sum_{i=1}^{3} n_i \hat{p}_i$, then $X \sim bin(p, n_1 + n_2 + n_3)$. Then

$$\hat{p}_{pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2 + n_3\hat{p}_3}{n_1 + n_2 + n_3} = \frac{90 + 624 + 315}{200 + 1300 + 750} = \frac{1029}{2250} = 0.4573.$$

Using $\hat{p}_{pooled}$ to estimate p, we have standard deviation

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_1 + n_2 + n_3}} = \sqrt{\frac{0.4573(1 - 0.4573)}{2250}} = 0.010488.$$

(b) Suppose we use the following alternative method of constructing a pooled estimate:

$$\hat{p}^*_{pooled} = \frac{\sum_{i=1}^{3} \hat{p}_i}{3} = \frac{0.45 + 0.48 + 0.42}{3} = 0.45.$$

Since the estimates are independent,

$$\sigma^2_{\hat{p}} = \frac{1}{3^2}\sigma^2_{\hat{p}_1} + \frac{1}{3^2}\sigma^2_{\hat{p}_2} + \frac{1}{3^2}\sigma^2_{\hat{p}_3}.$$

Using estimate $p \approx \hat{p}^*_{pooled}$ we have

$$\begin{aligned}\sigma^2_{\hat{p}^*} &= \frac{1}{3^2}\hat{p}^*_{pooled}(1-\hat{p}^*_{pooled})\left[\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}\right] \\ &= \frac{1}{3^2}0.45(1-0.45)\left[\frac{1}{200}+\frac{1}{1300}+\frac{1}{750}\right] \\ &= 0.000195.\end{aligned}$$

This gives

$$\sigma_{\hat{p}^*} \approx \sqrt{0.000195} = 0.01396.$$

(c) Since $0.01396 = \sigma_{\hat{p}^*} > \sigma_{\hat{p}} = 0.010488$, the first pooled estimator $\hat{p}_{pooled}$ is more accurate.

Problem 2.34 A certain experimental cancer therapy was evaluated in a clinical trial. A control group of 70 subjects was given standard care, while a treatment group of 50 subjects was given the experimental therapy. The subjects were observed for a 5 year period. Of the control group, 29 subjects experienced recurrence, while of the treatment group, 15 subjects experienced recurrence.

	Control ($i=1$)	Treatment ($i=2$)
X_i = Number of recurrences	29	15
n_i = Group sample size	70	50
$\hat{p}_i$ = Observed recurrence rate	0.414	0.3

(a) Construct a confidence interval for the difference in recurrence rates $p_2 - p_1$ between the Treatment and Control groups. Use confidence level $1-\alpha = 0.95$.

(b) Construct a confidence interval for the log odds ratio:

$$\log\left[\frac{Odds(Recurrence \mid Treatment)}{Odds(Recurrence \mid Control)}\right].$$

Use confidence level $1-\alpha = 0.95$.

(c) What can be concluded from parts (a) and (b) as to whether or not recurrence rates differ between the groups? Are the conclusions from parts (a) and (b) consistent?

SOLUTION:

(a) We have estimates $\hat{p}_1 = 0.248$, $\hat{p}_2 = 0.306$. The confidence interval is then

$$\begin{aligned}CI_{1-\alpha} &= \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.3 - 0.414 \pm 1.96\sqrt{\frac{0.3(1-0,3)}{70}+\frac{0.414(1-0.414)}{50}} \\ &= -0.114 \pm 0.172 \\ &= (-0.286, 0.0573).\end{aligned}$$

(b) We have contingency table:

	Control	Treatment
Recurrence	29	15
No Recurrence	41	35

The estimate of the OR is

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 29 \times 35/(41 \times 15) = 1.65.$$

The standard error is

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{29} + \frac{1}{15} + \frac{1}{41} + \frac{1}{35}} = 0.3926.$$

An approximate $(1-\alpha)100\%$ confidence interval for $\log(OR)$ is therefore

$$\begin{aligned} CI_{1-\alpha} &= \log(\hat{OR}) \pm z_{\alpha/2} SE(\log(OR)) \\ &= \log(1.65) \pm 1.96 \times 0.3926 \\ &= 0.501 \pm 0.769 \\ &= (-0.268, 1.270). \end{aligned}$$

(c) From part (a), the CI for $p_2 - p_1$ contains 0. From part (b) the CI for log OR also contains 0, equivalent to OR = 1. Both conclusions are consistent with the null hypothesis $H_o : p_1 = p_2$ (and so are consistent with each other).

Problem 2.35 A study compares the audit rates of residents of New York state (NY) to residents of Pennsylvania (PA). The audit classifications (ie *Audit* or *No Audit*) of $n = 1621$ randomly selected study subjects, along with their state of residence, are summarized in the contingency table given below. We are interested in estimating the odds ratio $OR = Odds(Audit \mid NY)/Odds(Audit \mid PA)$. Construct a confidence interval for $\log(OR)$ with confidence level $1 - \alpha = 0.95$. Interpret the result as a two-sided test for null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$.

	Audit	*No Audit*
NY	55	12
PA	912	642

SOLUTION: The estimate of the OR is

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{55 \times 642}{12 \times 912} = 3.226.$$

The standard error of $\log(\hat{OR})$ is

$$\begin{aligned} SE(\log(\hat{OR})) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{55} + \frac{1}{12} + \frac{1}{912} + \frac{1}{642}} \\ &= 0.3227527. \end{aligned}$$

The CI is

$$\begin{aligned} CI &= \log(\hat{OR}) \pm z_{\alpha/2} \times SE(\log(\hat{OR})) \\ &= 1.171 \pm 1.96 \times 0.3227527 \\ &= 1.171 \pm 0.6325836 \\ &= (0.539, 1.804). \end{aligned}$$

The CI does not contain 0, equivalent to $\log(OR) = 0$ where $OR = 1$, so reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.36 Categorical counts from two independent samples are summarized in the following table, along with the sample sizes n. Suppose p_1 and p_2 are the respective population proportions of the category. Calculate a confidence interval for $p_1 - p_2$ with confidence level $1 - \alpha = 0.95$.

	Sample 1	Sample 2
X	18	35
n	50	70
$\hat{p}$	0.36	0.50

(a) Construct a confidence interval for the difference in recurrence rates between the NBP and HBP subjects. Use confidence level $1 - \alpha = 0.95$.

(b) Construct a confidence interval for the log odds ratio:

$$\log\left[\frac{Odds(Recurrence \mid NBP)}{Odds(Recurrence \mid HBP)}\right].$$

Use confidence level $1 - \alpha = 0.95$.

(c) What can be concluded from parts (a) and (b) as to whether or not recurrence rates differ between the NBP and HBP groups? Are the conclusions from parts (a) and (b) consistent?

SOLUTION:

(a) We have estimates $\hat{p}_1 = 0.248$, $\hat{p}_2 = 0.306$. The confidence interval is then

$$\begin{aligned} CI_{1-\alpha} &= \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.248 - 0.306 \pm 1.96\sqrt{\frac{0.248(1-0.248)}{495} + \frac{0.306(1-0.306)}{474}} \\ &= -0.0574 \pm 0.0563 = (-0.114, -0.00112). \end{aligned}$$

(b) The estimate of the OR is

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 123 \times 329/(145 \times 372) = 0.750.$$

The standard error is

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} = \sqrt{\frac{1}{123}+\frac{1}{145}+\frac{1}{372}+\frac{1}{329}} = 0.1440636.$$

An approximate $(1-\alpha)100\%$ confidence interval for $\log(OR)$ is therefore

$$\begin{aligned} CI_{1-\alpha} &= \log(\hat{OR}) \pm z_{\alpha/2} SE(\log(OR)) \\ &= \log(0.750) \pm 1.96 \times 0.1441 \\ &= -0.2874 \pm 0.2824 = (-0.5698, -0.005). \end{aligned}$$

(c) From part (a), the CI for $p_1 - p_2$ is entirely negative. From part (b) the CI for OR is also entirely negative. Both conclusions imply $p_1 < p_2$, and so are consistent.

Problem 2.37 Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 21$, with sample size $n = 56$.

(a) Test hypothesis $H_o : p \leq 0.25$ against $H_a : p > 0.25$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$? Use the continuity correction.
(b) Verify your answer using the `R prop.test()` function. What is the P-value without the continuity correction (use the `prop.test()` with the appropriate option to answer this)?

SOLUTION:

(a) To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction for the evaluation of an upper-tailed P-value $\alpha_{obs} = P(Z > Z_{obs})$ can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{X - np_0 - 0.5}{\sqrt{np_0(1-p_0)}} = \frac{21 - 14 - 0.5}{\sqrt{56 \times 0.25(1-0.25)}} = 2.006.$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = P(Z > Z'_{obs}) = P(Z > 2.006) = 0.0224.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.
(b) Use `proptest()` as follows:

```
> prop.test(21,56,p=0.25,alternative='greater',correct=T)

1-sample proportions test with continuity correction

data:  21 out of 56, null probability 0.25
X-squared = 4.0238, df = 1, p-value = 0.02243
```

```
alternative hypothesis: true p is greater than 0.25
95 percent confidence interval:
0.268643 1.000000
sample estimates:
p
0.375

> prop.test(21,56,p=0.25,alternative='greater',correct=F)

1-sample proportions test without continuity correction

data:  21 out of 56, null probability 0.25
X-squared = 4.6667, df = 1, p-value = 0.01538
alternative hypothesis: true p is greater than 0.25
95 percent confidence interval:
0.2766707 1.0000000
sample estimates:
p
0.375
```

With the contiunuity correction, we obtain the same P-value as in Part (a). Without the continuity correction, the P-value decreases from 0.0224 to 0.01538.

Problem 2.38 A chain of grocery stores conducts a customer satisfaction survey at several of its properties. Customers were asked the question: 'Do you shop primarily at this site?'. The following table gives the number surveyed; and the number who answer 'yes' for two different sites.

	Site 1	Site 2
Answered Yes	65	100
Answered No	185	400
Total	250	500

Let p_1, p_2 represent the proportion of each population sampled who would answer 'yes'.

(a) Construct a level 0.95 confidence interval for proportion difference $p_2 - p_1$.
(b) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimates of p_1, p_2 and $p_2 - p_1$ are

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{65}{250}, \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{100}{500}, \quad \hat{p}_2 - \hat{p}_1 = -0.06.$$

The confidence interval is then given by

$$\begin{aligned} CI &= -0.06 \pm 1.96\sqrt{\frac{0.26(1-0.26)}{250} + \frac{0.2(1-0.2)}{500}} \\ &= -0.06 \pm 1.96 \times 0.033 \\ &= -0.06 \pm 0.0647 \end{aligned}$$

or equivalently, $CI = (-0.125, 0.0047)$.

(b) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{65 + 100}{250 + 500} = 0.22.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.2 - 0.26}{\sqrt{0.22(1-0.22)\left(\frac{1}{250} + \frac{1}{500}\right)}} \\ &= \frac{-0.06}{0.0321} \\ &= -1.87 \end{aligned}$$

The P-value is, for $Z \sim N(0,1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 1.87) = 0.0615.$$

Since $\alpha_{obs} > \alpha$ we do not reject the null hypothesis at an α significance level.

Problem 2.39 Suppose in a clinical trial involving 91 liver cancer patients $n_1 = 34$ were treated with an experimental drug, and $n_2 = 57$ were given the conventional treatment. The following contingency table reports the number of each group that experienced a recurrence within 6 months.

	Experimental Treatment	Standard Treatment	
Cancer recurs	3	12	15
Cancer does not recur	31	45	76
Total	34	57	91

Construct a level 0.95 confidence interval for the log odds ratio of recurrence between groups Experimental Treatment and Standard Treatment. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$?

SOLUTION: The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{3 \times 45}{12 \times 3} = 0.363.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{3} + \frac{1}{12} + \frac{1}{31} + \frac{1}{45}} \\ &= 0.686. \end{aligned}$$

The level $1-\alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= -1.01 \pm 1.96 \times 0.686 \\ &= -1.01 \pm 1.345 \end{aligned}$$

or equivalently, $CI = (-2.36, 0.332)$. Since the CI contains 0 we do not reject the null hypothesis at an $\alpha = 0.05$ significance level.

2.3 Sample Size Estimates

Problem 2.40 We are given an *iid* sample from a normal distribution $N(\mu, \sigma^2)$:

$$9.43, 9.85, 10.12, 9.89, 9.81, 10.3,$$

of sample size $n = 6$.

(a) Calculate a level 0.9 upper confidence bound for the standard deviation σ.
(b) Use this upper bound to estimate the sample size needed to construct a 95% confidence interval for the mean μ with a margin of error $E_o = 0.1$ (use a normal approximation).

SOLUTION:

(a) The sample standard deviation is $S = 0.297$. The level $1-\alpha$ upper bound for σ is given by ,

$$\sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{5,0.9} = 1.61$. The upper bound is then given by,

$$\sigma < \frac{0.297}{\sqrt{1.61/5}} = 0.523.$$

(b) To guarantee margin of error $E_o = 0.1$ a sample size of

$$n = \left(\frac{z_{\alpha/2}\hat{\sigma}}{E_o}\right)^2 = \left(\frac{1.96 \times 0.523}{0.1}\right)^2 = 105,$$

should be used.

Problem 2.41 A coin is tossed $n = 2030$ times, and the outcome is heads $X = 1050$ times. Let p be the probability of heads for a single toss.

(a) Construct a confidence interval for p, using a 95% confidence level.
(b) What sample size is needed to guarantee a margin of error $E_o = 0.01$ using a 95% confidence level.

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 1050$ and $n = 2030$ is

$$\hat{p} = \frac{X}{n} = \frac{1050}{2030} = 0.517.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.517 \pm 1.96\sqrt{\frac{0.517(1-0.517)}{2030}} \\ &= 0.517 \pm 1.96 \times 0.0111 \\ &= 0.517 \pm 0.0217 \end{aligned}$$

or equivalently, $CI = (0.496, 0.539)$.

(b) To guarantee margin of error $E_o = 0.01$ calculate n assuming $p^* = 1/2$. Then a sample size of

$$n = \frac{1}{4}\left(\frac{z_{\alpha/2}}{E_o}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

should be used.

Problem 2.42 A coin tossing procedure is tested for fairness. Out of $n = 25$ tosses $X = 11$ result in *heads*.

(a) Let p be the true probability of tossing *heads*. Test the null hypothesis $H_o : p = 0.5$ against two-sided alternative $H_a : p \neq 0.5$. Use significance level $\alpha = 0.01$. Use a normal approximation with continuity correction.
(b) Suppose we wish to construct a confidence interval for p with confidence level $1 - \alpha = 0.99$ and a margin of error no greater than $E = 0.025$. What sample size would be needed?

SOLUTION:

(a) $\hat{p} = 0.44$, Z = -0.6. Reject H_o if $|Z|$ is greater than or equal to $z_{\alpha/2} = 2.576$. Therefore, do not reject the null hypothesis at a significance level $\alpha = 0.01$. P-value = 0.5485. The level $1 - \alpha = 0.99$ CI is (0.184, 0.696). Applying the continuity correction gives P-value = 0.6892 and the exact binomial distribution gives P-value = 0.69.
(b) If $E = 0.025$, use as estimate $p^* = 0.5$. We have $z_{\alpha/2} = 2.576$, so

$$n = p^*(1 - p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.25\left(\frac{2.576}{0.025}\right)^2 = 2653.959.$$

Round up to $n = 2654$.

Problem 2.43 A coin tossing procedure is tested for fairness. Suppose we wish to construct a confidence interval for $p = P(HEADS)$ with confidence level $1 - \alpha = 0.99$ and a margin of error no greater than $E = 0.025$. What sample size would be needed?

SOLUTION: If $E = 0.025$, use as estimate $p^* = 0.5$. We have $z_{\alpha/2} = 2.576$, so

$$n = p^*(1 - p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.25\left(\frac{2.576}{0.025}\right)^2 = 2653.959.$$

Round up to $n = 2654$.

Problem 2.44 Given an *iid* sample of size $n = 253$ we observe a count of $X = 46$ of a certain category. Suppose p is the population proportion of that category.

(a) Calculate a confidence interval for p with confidence level $1 - \alpha = 0.95$.
(b) What sample size is needed to ensure a margin of error $ME \leq 0.02$, assuming $p \leq 0.25$ (again, use $1 - \alpha = 0.95$)?
(c) What sample size is needed to ensure a margin of error $ME \leq 0.02$ if no assumption about p is made (again, use $1 - \alpha = 0.95$)?

SOLUTION:

(a) Estimate is $\hat{p} = 46/253 = 0.1818$.

$$\begin{aligned} CI &= \hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.1818 \pm z_{\alpha/2} \times \sqrt{\frac{0.1818(1-0.1818)}{253}} \\ &= 0.1818 \pm 0.0475 \\ &= (0.134, 0.229). \end{aligned}$$

(b) We have

$$n^* = p^*(1-p^*)\left(\frac{z_{\alpha/2}}{ME}\right)^2 = 0.25 \times (1-0.25)\left(\frac{1.96}{0.2}\right)^2 = 1800.75.$$

Round up to $n^* = 1801$.

(c) We have

$$n^* = p^*(1-p^*)\left(\frac{z_{\alpha/2}}{ME}\right)^2 = 0.5 \times (1-0.5)\left(\frac{1.96}{0.2}\right)^2 = 2401.$$

Use $n^* = 2401$.

Problem 2.45 For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 0.709$, $n = 14$, standard deviation $\sigma = 1.25$.

(a) Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.
(b) What sample size woould be needed to obtain a margin of error $ME = 0.1$ for a confidence level $1-\alpha = 0.99$.

SOLUTION:

(a) We have $\alpha = 0.05$, so we need critical value

$$z_{\alpha/2} = z_{0.025} = 1.96,$$

giving level $1 - \alpha$ confidence interval

$$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 0.709 \pm 1.96 \times 0.334 \\ &= 0.709 \pm 0.655 = (0.0542, 1.364). \end{aligned}$$

(b) We have

$$n^* = \left(\frac{z_{0.01/2}\sigma}{ME}\right)^2 = \left(\frac{2.587 \times 1.25}{0.1}\right)^2 = 1045.7$$

Round up to $n^* = 1046$.

Problem 2.46 For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 138.6$, $n = 15$, sample standard deviation $S = 12.04$.

(a) Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.
(b) Calculate a level $1 - \alpha = 0.95$ upper confidence bound for σ.
(c) Using the upper bound for σ calculated in part (b) estimate the sample size needed to obtain a level $1 - \alpha = 0.95$ confidence interval for μ with a margin of error of 3.0. Use a normal approximation.

SOLUTION:

(a) We have

$$\begin{aligned} CI_{1-\alpha} &= \bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \\ &= 138.6 \pm 2.14 \times \frac{12.04}{\sqrt{15}} \\ &= 138.6 \pm 6.67 = (131.93, 145.27) \end{aligned}$$

(b) We have

$$\begin{aligned} UB &= \frac{S}{\sqrt{\chi^2_{n-1,1-\alpha}/(n-1)}} \\ &= \frac{12.04}{\sqrt{6.57/14}} \\ &= 17.57 \end{aligned}$$

So,

$$\sigma < 17.57$$

is the 95% upper confidence bound for σ.

(c) Use estimate $\hat{\sigma} = 5.836$ in formula

$$n \approx \left(z_{\alpha/2}\frac{\hat{\sigma}}{E_o}\right)^2 = \left(1.96 \times \frac{17.57}{3}\right)^2 = 130.43,$$

so round up to $n = 131$.

Problem 2.47 A model predicts that in a given population of trout, on average 13 out of every 40 specimens will possess a certain genetic mutation. Out of a random sample of size $n = 174$ the mutation is observed in $X = 51$ specimens.

(a) Perform an appropriate two-sided hypothesis test to determine whether or not the data is compatible with the model. Use significance level $\alpha = 0.05$. Use a continuity correction.

(b) Assuming that the actual population proportion with the mutation will not exceed $p = 0.4$, what sample size is needed to obtain a confidence interval for p with confidence level $1 - \alpha = 0.95$ and a margin of error no greater than $E = 0.05$?

SOLUTION:

(a) We are testing

$$H_o : p = p_0 \text{ against } H_a : p \neq p_0$$

where $p_0 = 13/40 = 0.325$.

$$\hat{p} = 51/174 = 0.293,$$

With continuity correction, since $\hat{p} < p_0$ use

$$\begin{aligned} Z_{obs} &= \frac{X + 0.5 - np_0}{\sqrt{np_0(1-p_0)}} \\ &= \frac{51 + 0.5 - 174 \times 0.325}{\sqrt{174 \times 0.325 \times (1 - 0.325)}} \\ &= -0.8174. \end{aligned}$$

(Without continuity correction Z = -0.898).
Reject H_o if

$$|Z| \geq z_{\alpha/2} = 1.96.$$

Therefore, so not reject the null hypothesis at a signficance level $\alpha = 0.05$.

(b) If we assume $p \leq 0.4$ then $p(1-p)$ is maximized by substituting $p^* = 0.4$, giving conservative estimate

$$\begin{aligned} n &= p^*(1-p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 \\ &= 0.4(1-0.4)\left(\frac{1.96}{0.05}\right)^2 \\ &= 368.7936. \end{aligned}$$

Rounding up gives sample size $n = 369$.

Problem 2.48 For an *iid* sample from a normal distribution $N(\mu, \sigma^2)$ we are given sample mean $\bar{X} = 9.362$, $n = 56$, sample standard deviation $S = 4.912$.

(a) Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.
(b) Calculate a confidence level $1 - \alpha = 0.95$ upper bound for σ.
(c) Using the upper bound for σ calculated in part (b) estimate the sample size needed to obtain a confidence interval for μ with a margin of error of 0.75. Use confidence level $1 - \alpha = 0.95$.

SOLUTION:

(a) We have

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_n \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \\ &= 9.362 \pm 2.00 \times \frac{4.912}{\sqrt{56}} \\ &= 9.362 \pm 1.315 = (8.047, 10.677) \end{aligned}$$

(b) We have

$$\begin{aligned} UB &= \frac{S}{\sqrt{\chi^2_{n-1,1-\alpha}/(n-1)}} \\ &= \frac{4.912}{\sqrt{38.96/55}} \\ &= 5.836 \end{aligned}$$

So,

$$\sigma < 5.836$$

is the 95% upper confidence bound for σ.

(c) Use estimate $\hat{\sigma} = 5.836$ in formula

$$n \approx \left(z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2 = \left(1.96 \times \frac{5.836}{0.75} \right)^2 = 232.6216,$$

so round up to $n = 233$.

Problem 2.49 For an *iid* sample of size $n = 6$ from a normal distribution $N(\mu, \sigma^2)$ we are given sample standard deviation $S = 1.44$.

(a) Give a level $1 - \alpha = 0.95$ upper confidence bound for σ.
(b) Use the upper confidence bound to estimate the sample size required for a level $1-\alpha = 0.99$ confidence interval with a margin of error of $ME = 0.5$. You may assume that the sample will be large enough to use a critical value from the standard normal distribution.

SOLUTION:

(a) The upper bound is given by

$$\sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha} \right) / (n-1)}} = \frac{1.44}{\sqrt{1.145/5}} = 3.01.$$

(b) Using $\hat{\sigma} = 3.01$, the estimated sample size is

$$n \approx \left(z_{\alpha/2} \frac{\hat{\sigma}}{ME} \right)^2 = \left(2.576 \frac{3.01}{0.5} \right)^2 = 240.48,$$

then round up to $n = 241$.

Problem 2.50 Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 50$, with sample size $n = 80$.

(a) Construct a level $1 - \alpha = 0.95$ confidence interval for p.
(b) What sample size would be needed to construct a level $1 - \alpha = 0.95$ confidence interval for p, such that the margin of error will not exceed 0.05 for any sample?

SOLUTION:

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 50$ and $n = 80$ is

$$\hat{p} = \frac{X}{n} = \frac{50}{80} = 0.625.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.625 \pm 1.96\sqrt{\frac{0.625(1-0.625)}{80}} \\ &= 0.625 \pm 1.96 \times 0.0541 \\ &= 0.625 \pm 0.106 \end{aligned}$$

or equivalently, $CI = (0.519, 0.731)$.

(b) For a margin of error $ME = 0.05$ the sample size is

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{ME}\right)^2,$$

which is maximized by setting $\hat{p} = p^* = 0.5$, so for $ME = 0.05$ and $z_{\alpha/2} = 1.96$:

$$n \approx 0.5 \times 0.5 \times \left(\frac{1.96}{0.05}\right)^2 = 384.16.$$

After rounding up we have $n = 385$.

Problem 2.51 It is important to get repeatable temperature readings for a project. The thermometer to be used is known to have a standard deviation of $\sigma = 0.12$ degrees centigrade. However, a margin of error of 0.1 is needed with a confidence level of 99%. It is decided to use the sample average of n similar thermometers. What should n be?

SOLUTION: We are given

$$\begin{aligned} \alpha &= 0.01, \\ z_{\alpha/2} &= 2.576, \\ \sigma &= 0.12, \\ E &= 0.1, \end{aligned}$$

so that the sample size required is

$$\begin{aligned} n &= \left(z_{\alpha/2}\frac{\sigma}{E}\right)^2 \\ &= \left(2.576 \times \frac{0.12}{0.1}\right)^2 \\ &= 9.56 \end{aligned}$$

which, when rounded up, gives a sample size of $n = 10$.

Problem 2.52 Suppose we wish to test a die for fairness by estimating the proportion of tosses p with an outcome of 1. We want a margin of error no larger than 0.01 for a level 95% confidence interval. Estimate the required sample size:

(a) assuming p is no higher than 0.25;
(b) with no restriction on p.

SOLUTION: We have margin of error $E = 0.01$, and critical value $z_{\alpha/2} = 1.96$, given $0.95 = 1 - \alpha$.

(a) If the maximum proportion is $p^* = 0.25$ we have

$$n = p^*(1-p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.25 \times 0.75 \times \left(\frac{1.96}{0.01}\right)^2 = 7203.$$

We need a sample size of $n = 7203$.

(b) If there is no restriction on p we use $p^* = 0.5$, since this gives the largest sample size. This gives

$$n = p^*(1-p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.5 \times 0.5 \times \left(\frac{1.96}{0.01}\right)^2 = 9603.$$

We need a sample size of $n = 9603$.

2.4 Power Curves

Problem 2.53 Suppose a population of measurements is claimed to be normally distributed with mean no larger that $\mu = 150$ and standard deviation $\sigma = 10$. To test this claim, a random sample of n components is to be collected to do a hypothesis test for $H_o : \mu = 150$ against $H_a : \mu > 150$.

(a) Use `R` to draw a *power curve*, that is plot $Power(\mu) = 1 - \beta(\mu)$ as a function of μ over a suitable range of alternative hypotheses, say $\mu \in (150, 175)$ (use `seq(150,175,by=0.1)` to generate the values of μ for your plot). Do this for a Type I error of $\alpha = 0.05$. Superimpose on the same plot power curves for $n = 5, 10, 15, 20, 25, 30$. Label the appropriate axes μ and $Power(\mu)$. Include a horizontal line at level 0.05, labelled $\alpha = 0.05$. Also, indicate, using the `text()` function, the positions of the $n = 5$ and $n = 30$ curves.
(b) Create a table giving the power for each combination of $n = 5, 10, \ldots, 30$ and

$$\mu = 155, 156, 157, 158, 159, 160.$$

For each of these values of μ, give the minimum sample size (from those considered) required to attain a power of 80%.

SOLUTION: The following `R` code gives the required plot and tables:

```
### Set parameters

mu0 = 150
```

```
sigma = 10
zalpha = -qnorm(0.05)
mugrid = seq(150,175,by=0.01)

### Create text

ex1 = expression(mu)
ex2 = expression(paste('Power(',mu,')',sep=''))
ex3 = expression(paste(alpha,' = 0.05',sep=''))
ex4 = expression(paste(italic(n),' = 5',sep=''))
ex5 = expression(paste(italic(n),' = 30',sep=''))

### Create plot region

plot(range(mugrid),c(0,1),type='n',xlab=ex1,ylab=ex2)

### Draw power curve for each value of n, capturing data for power table

power.table = NULL
for (n in seq(5,30,by=5)) {
pow = 1-pnorm(zalpha - (mugrid-mu0)*sqrt(n)/sigma)
lines(mugrid,pow,type='l')
power.table = cbind(power.table,pow)
}

### Annotate plot

abline(h=0.05,col='gray')
text(165,0.09,ex3)
text(159,0.5,ex4)
text(152,0.75,ex5)

### Create power table

rownames(power.table) = mugrid
colnames(power.table) = seq(5,30,by=5)
final.table = power.table[which(mugrid %in% c(155,156,157,158,159,160)),]
```

(a) The required plot is shown in Figure 2.1.

(b) The required table was produced by the preceding script, and is

```
> final.table
5         10        15        20        25        30
155 0.2991594 0.4745987 0.6147183 0.7228116 0.8037649 0.8629697
156 0.3808638 0.5996777 0.7514109 0.8504646 0.9123145 0.9496513
157 0.4682753 0.7152340 0.8568412 0.9313130 0.9682123 0.9857090
```

```
158 0.5572501 0.8119132 0.9269621 0.9733730 0.9907423 0.9968992
159 0.6434171 0.8851625 0.9671769 0.9913453 0.9978492 0.9994895
160 0.7228116 0.9354202 0.9870641 0.9976528 0.9996034 0.9999365
```

For $\mu = 155$ we attain power 72.3% for $n = 20$, and 80.4% for $n = 25$. Therefore, the minimum sample size from those given needed to attain a power of 80%, is $n = 25$. Using this approach, the minimum sample sizes for each value of μ are given in the following table:

μ	Minimum Sample Size
155	25
156	20
157	15
158	10
159	10
160	10

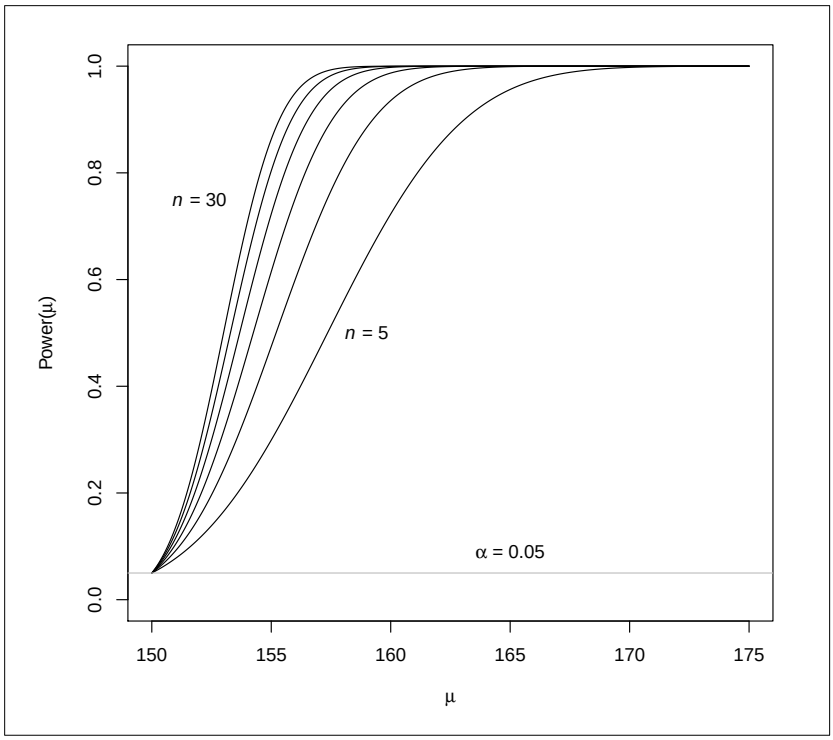

Figure 2.1: Plot for Problem 2.53 (a).

Problem 2.54 You are planning a study that will involve a one sample upper-tailed t-test for hypotheses $H_o : \mu = 100$ against $H_a : \mu > 100$ based on an *iid* sample from a $N(\mu, \sigma^2)$ distribution. The significance level will be 5%.

(a) Construct a power curve in which the vertical axis is $pow = 1 - P($ Type II Error $)$, and the horizontal axis is $\delta = (\mu - 100)/\sigma$. Note that σ doesn't have to be known.

(b) For a given sample size n we can calculate the power for rejecting an alternative for which the effect size is $\delta = (\mu - 100)/\sigma = 2.5$. Construct a plot in which the horizontal axis is sample size $n = 100, 101, \ldots, 150$ and the vertical axis is the power for an effect size $\delta = 2.5$ as a function of sample size n. Determine the smallest sample size needed to attain a power of 0.8.

SOLUTION: To test $H_o : \mu = \mu_0$ against $H_a : \mu > \mu_0$ use test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and reject H_o if $T > t_{n-1,\alpha}$. If $\delta = (\mu - \mu_0)/\sigma$ then

$$P(\text{Type II error}) = \beta(\delta) = P\left(T_{\sqrt{n}\delta} \leq t_{n-1,\alpha}\right),$$

where $T_{\sqrt{n}}$ has a t-distribution with $n-1$ d.f. and $ncp = \sqrt{n}\delta$. Use the `pt()` function as shown below.

```
> par(mfrow=c(1,2),pty='s')
>
> n = 20
> del = seq(0,2,0.02)
>
> plot(del, 1-pt(qt(0.95,df=n-1), ncp=sqrt(n)*del, df=n-1 ), ylim=c(0,1),
type='l', xlab='del', ylab="Power", main = "n = 20")
> lines(range(del), c(0.05, 0.05), lty=2)
> text(0.5, 0.07, "= 0.05")
>
> nlist = c(100:150)
> powl = 1-pt(qt(0.95,df=nlist-1), ncp=sqrt(nlist)*0.25, df=nlist-1)
> plot(nlist, powl, ylim=c(0.75, 0.95), type='l', xlab='Sample size n',
ylab="Power", main="effect size del = 0.25")
>
> cbind(nlist, powl)[1:5,]
nlist      powl
[1,]   100 0.7989855
[2,]   101 0.8024927
[3,]   102 0.8059454
[4,]   103 0.8093443
[5,]   104 0.8126899
```

(a) The required plot is shown in Figure 2.2 (left plot).
(b) The required plot is shown in Figure 2.2 (right plot). The smallest sample size yielding a power of at least 80% is $n = 101$.

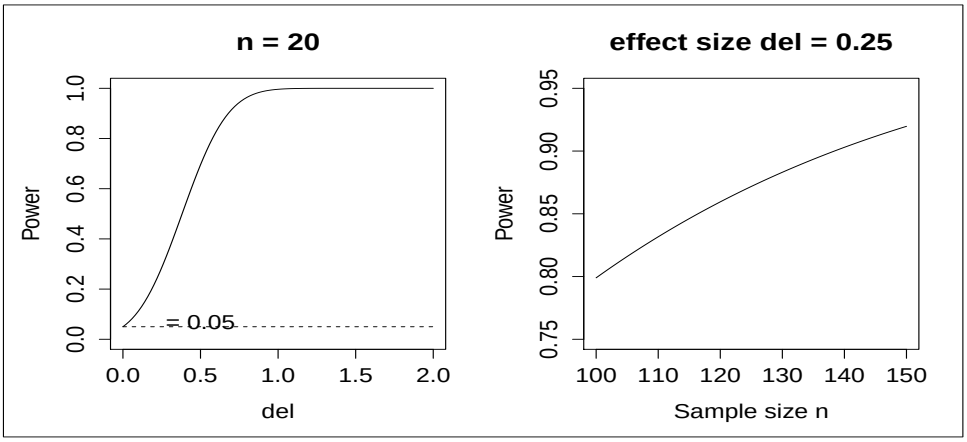

Figure 2.2: Plot for Problem 2.54 (a)-(b).

Problem 2.55 A company has developed a predictive model for the screening of applicants based on a questionnaire. The responses are converted to 3 components:

$$\begin{aligned} X_1 &= \text{Leadership skills} \\ X_2 &= \text{Communication skills} \\ X_3 &= \text{Level of expertise.} \end{aligned}$$

Each component is normally distributed, and has been standardized to have zero mean and standard deviation $\sigma_X = 25$. A composite score believed to be especially predictive of success is given by

$$T = \frac{1}{2}X_1 + \frac{1}{6}X_2 + \frac{1}{3}X_3.$$

The company wishes to use T for screening job applicants. If an applicant's score exceeds a threshold $T \geq t$ they are selected for further interviews. The company wishes to select 10% of applicants for further screening, so it sets the threshold at the value

$$t = \sigma_T \times z_{0.1}$$

where σ_T is the standard deviation of T, and $z_{0.1}$ is the 10% critical value of a standard normal distribution $N(0, 1)$. If σ_T is correctly calculated, and $E[T] = 0$ as expected, we would have

$$P(T > \sigma_T \times z_{0.1}) = 0.1.$$

It is then noted that in order to calculate σ_T, the correlations between X_1, X_2 and X_3 must be known (Sections 4.8 - 4.9 of the lecture notes). Following this, two points of view emerge, which we'll refer to as the null and alternative hypotheses.

H_o Scales from psychometric questionnaires are designed to measure independent constructs. So, although we might expect, say, leaderships skills and communication skills to be positively correlated in everyday life, the scales X_1, X_2 and X_3 are designed to measure these qualities in a manner that is independent of the others. Therefore, we should expect zero correlation between X_1, X_2 and X_3.

H_a A statistical analysis has estimated the following correlations, and these should therefore be used to calculate σ_T.

$$\begin{aligned}\rho_{X_1,X_2} &= 0.56\\ \rho_{X_1,X_3} &= 0.18\\ \rho_{X_2,X_3} &= 0.21.\end{aligned}$$

In order to test these hypotheses, a sample of n test scores T is to be collected. Design a size $\alpha = 0.05$ hypothesis test for null and alternative hypotheses H_o and H_a. Note that only the scores T will be be available, and not the underlying scores X_1, X_2 and X_3. Construct a plot of power against sample size n, for $n = 2, 3, \ldots, 199, 200$. Superimpose a horizontal line at *power* = 90%. What is the minimum sample size needed to attain at least 90% power?

SOLUTION: Write the test statistic

$$T = \frac{1}{2}X_1 + \frac{1}{6}X_2 + \frac{1}{3}X_3 = W_1 + W_2 + W_3,$$

so that

$$\begin{aligned}\sigma^2_{W_1} &= \frac{\sigma^2_X}{2^2} = \frac{25^2}{2^2}\\ \sigma^2_{W_2} &= \frac{\sigma^2_X}{6^2} = \frac{25^2}{6^2}\\ \sigma^2_{W_3} &= \frac{\sigma^2_X}{3^2} = \frac{25^2}{3^2}.\end{aligned}$$

Then

$$\sigma^2_T = \sigma^2_{W_1} + \sigma^2_{W_2} + \sigma^2_{W_3} + 2 \times cov(W_1, W_2) + 2 \times cov(W_1, W_3) + 2 \times cov(W_2, W_3).$$

But we may write, since, $E[W_i] = E[X_i] = 0$ for $i = 1, 2, 3$,

$$cov(W_1, W_2) = E[W_1 W_2] = \frac{1}{2} \times \frac{1}{6} E[X_1 X_2] = \frac{1}{2} \times \frac{1}{6} cov(X_1, X_2) = \frac{1}{2} \times \frac{1}{6} \rho_{X_1,X_2} \sigma_X \sigma_X,$$

carrying out a similar calculation for the other pairs. Under H_o all correlations are zero, so

$$\sigma^2_o = \left(\frac{1}{2^2} + \frac{1}{6^2} + \frac{1}{3^2}\right) 25^2 \ \ [H_o].$$

Under H_a we use the given correlations, so

$$\sigma^2_a = \left(\frac{1}{2^2} + \frac{1}{6^2} + \frac{1}{3^2} + \frac{2 \times 0.56}{2 \times 6} + \frac{2 \times 0.18}{2 \times 3} + \frac{2 \times 0.21}{6 \times 3}\right) 25^2 \ \ [H_a].$$

Next, suppose we have an independent sample of scores $T_1, \ldots, T_n$. Calculate sample variance S^2_T for these scores and test

$$H_o : \sigma^2_T = \sigma^2_o \text{ versus } H_a : \sigma^2_T = \sigma^2_a$$

Use test statistic

$$X^2 = \frac{(n-1)S^2}{\sigma_o^2}.$$

If H_o is true then X^2 is a χ^2_{n-1} random variable. If H_a is true than X^2 is a χ^2_{n-1} random variable muliplied by $\sigma_1^2/\sigma_0^2 > 1$, so we reject H_o for large values of X^2. Therefore, reject for

$$X^2 > \chi^2_{n-1;0.05},$$

where $\chi^2_{n-1;0.05}$ is the appopriate critical value. The power as a function of n is therefore

$$Pow(n) = P\left(X^2 > \frac{\sigma_0^2}{\sigma_1}\chi^2_{n-1;0.05}\right),$$

where X^2 is a χ^2 random variable with $n-1$ degrees of freedom.

The following code can be used to create the plot (Figure 2.3):

```
### Parameters

r12 = 0.56
r13 = 0.18
r23 = 0.21

### Variances

v0 = 225*((1/4) + (1/36) + (1/9))
v1 = 225*((1/4) + (1/36) + (1/9) + 2*r12/(2*6) + 2*r13/(2*3) + 2*r23/(6*3))

### Function gives power for sample size n

pw = function(n) {1 - pchisq( qchisq(0.95,n-1)*(v0/v1), n-1)}

### Plot power vs n

nn = seq(2,200,1)
plot(nn,pw(nn),ylim=c(0,1),type='l',xlab='n',ylab='Power',cex.lab=1.25)
abline(h=c(0.9),col='gray')
```

The minimum sample size needed to attain a power of at least 90% can be given by

```
> ### Minimum sample size giving power = 90%
>
> min(nn[pw(nn) >= 0.9])
[1] 123
```

so that the required sample size is $n = 123$.

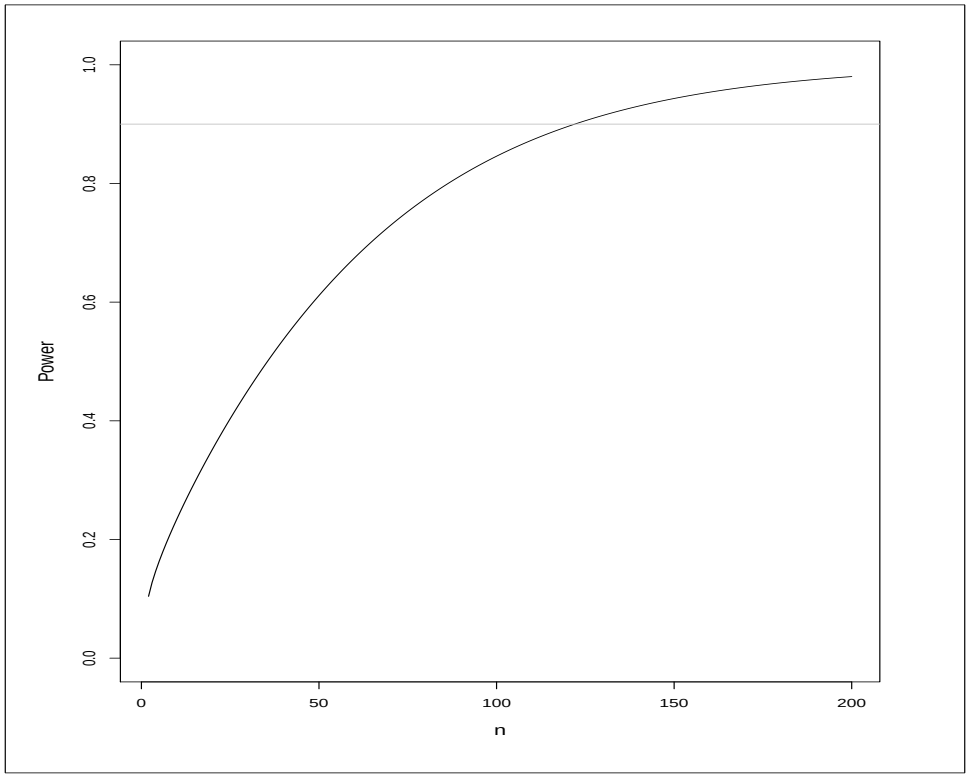

Figure 2.3: Plot for Problem 2.55.

Problem 2.56 Consider the t-test for two independent samples with respective sample sizes n_1, n_2, assuming equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (Section 14.1.2 of lecture notes). We wish to test null hypothesis $H_o : \mu_1 \geq \mu_2$ against alternative hypothesis $H_a : \mu_1 < \mu_2$. Then H_o is rejected for large values of test statistic:

$$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where S_p^2 is the pooled variance.

(a) Suppose $\mu_2 - \mu_1 = \Delta \neq 0$. Show that T_{obs} has a non-central t-distribution with non-centrality parameter:

$$ncp = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

assuming that the populations are normally distributed. One consequence of this is that

$$(n_1 + n_2 - 2)S_p^2/\sigma^2 \sim \chi^2_{n_1+n_2-2},$$

and $\bar{X}_2 - \bar{X}_1 \perp S_p^2$.

(b) Given the form of the non-centrality parameter ncp, show that if the total sample size $N = n_1 + n_2$ is fixed, the most powerful test is obtained by the balanced design $n_1 = n_2 = N/2$, assuming N is even.

(c) Construct power curves for the one-sided pooled variance t-test just described, for $\alpha = 0.05$, with the following features:

(i) Assume a balanced design $n = n_1 = n_2$. The plot will superimpose power curves for $n = 5, 10, 15, 20$ on the same plot.

(ii) The power curve will plot power $1 - \beta$ for one-sided alternatives $\Delta = \mu_2 - \mu_1 > 0$ against Δ/σ, over the range $0 \leq \Delta/\sigma \leq 3$ (use increments of 0.1).

(iii) Label each curve appropriately using the `text()` function.

(iv) The vertical axis should be labeled $1 - \beta$ and the horizontal axis should be labeled Δ/σ, making using the `expression()` function. See `help(plotmath)` for more detail.

(v) A grid should be superimposed with grid size 0.05 for the vertical axis and 0.125 for the horizontal axis. You can use the `abline()` function. Setting option `col = 'gray'` seems to work well.

(d) Suppose we need to determine a per-sample sample size n for a one-sided two-sample pooled variance t-test, testing null hypothesis $H_o : \mu_1 \geq \mu_2$ against alternative hypothesis $H_a : \mu_1 < \mu_2$. A power of 90% is needed for an alternative $\mu_2 - \mu_1 = 5.85$, assuming standard deviation $\sigma = 5.2$ and using $\alpha = 0.05$. Using your power curves, what value of n would you recommend (select from 5,10,15,20)?

SOLUTION:

(a) The non-central t distribution is constructed from

$$T_\delta = \frac{Z + \delta}{\sqrt{W/\nu}}$$

where $Z \sim N(0,1)$, $W \sim \chi^2_\nu$, $Z \perp W$ and δ is the non-centrality parameter. Then we can write

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{\bar{X}_2 - \bar{X}_1 - \Delta + \Delta}{S_p\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \\ &= \frac{\frac{\bar{X}_2 - \bar{X}_1 - \Delta}{\sigma\sqrt{\frac{n_1 n_2}{n_1+n_2}}} + \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}}{S_p/\sigma} \\ &= \frac{Z + \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}}{\sqrt{W/(n_1 + n_2 - 2)}} \end{aligned}$$

where

$$\begin{aligned} Z &= \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{\sigma\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \\ W &= S_p^2/\sigma^2, \end{aligned}$$

with $Z \sim N(0,1)$, $W \sim \chi^2_{n_1+n_2-2}$, $Z \perp W$ and

$$\delta = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

(b) A larger non-centrality parameter implies greater power. If $N = n_1 + n_2$ is fixed, we can write $n_2 = N - n_1$. Then maximize

$$\frac{n_1 n_2}{n_1 + n_2} = \frac{n_1(N - n_1)}{N}$$

with respect to n_1. Taking the derivative verifies that the quantity is maximized by setting $n_1 = N/2$.

(c) The following code produces the power curves in Figure 2.4.

```
### set up labels with mathematical typesetting

ex0 = expression(paste("One-sided two-sample t-test with sample size ",
italic(n)," per sample and ",alpha," = 0.05",sep=''))
ex1 = expression(paste(Delta,'/',sigma))
ex2 = expression(1-beta)

### grid for horizontal axis

del = seq(0,3,by = 0.1)

### set up graphics window, draw empty plot (type='n')

par(mar=c(4,5,2,2), oma=c(4,4,4,4), cex=1,cex.axis=1.2, cex.lab=1.2, cex.main=1.2)
plot(range(del), c(0,1), xlab=ex1, ylab=ex2, type='n')
title(ex0)

### grid

for (x in seq(0,3,by=0.125)) {abline(v=x,col='gray')}
for (y in seq(0,1,by=0.05)) {abline(h=y,col='gray')}

for (n in c(5,10,15,20)) {

# power curve for n

nfactor = sqrt(n/2)
alpha = 0.05
t.crit = qt(1 - alpha,df=2*n-2)
y = 1 - pt(t.crit,ncp=nfactor*del,df=2*n-2)
lines(del, y, type='l')

# label individual plots

ex3 = bquote(italic(n) == .(n))
text(1,y[del==1],ex3)
}
```

(d) We have

$$\frac{\Delta}{\sigma} = \frac{5.85}{5.2} = 1.125.$$

The first curve to equal or exceed $1 - \beta = 0.9$ as n increases is $n = 15$, which is the best choice.

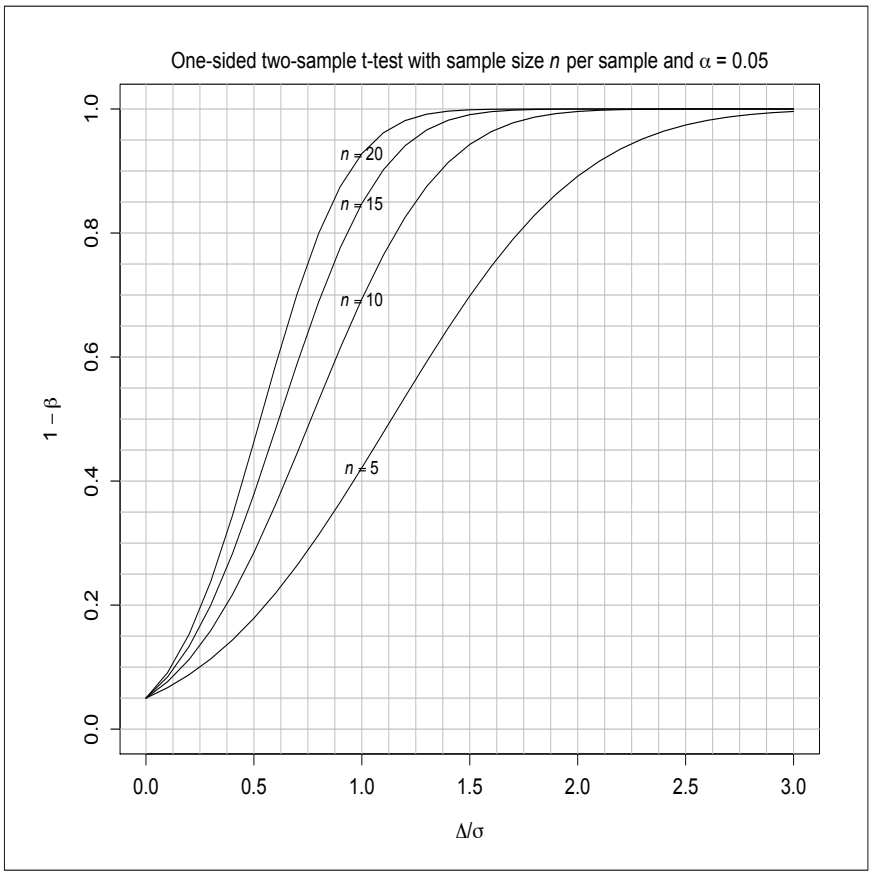

Figure 2.4: Plot for Problem 2.56 (c).

Problem 2.57 Examine the accuracy of the approximations given in Equations (17.8) - (17.9) of Chapter 17 of `INTRODUCTION TO STATISTICAL METHODOLOGY` by constructing two-sided power curves ($\alpha = 0.05$) for alternatives $p \in (1/2, 1)$ against null hypothesis $p_0 = 1/2$ for sample sizes $n = 10, 100, 1000$. Specifically, draw a function of $1 - \beta$ against $p \in [1/2, 1)$, using for β the exact expression in Equation (17.7) and the approximation in Equation (17.8) on the same plot (include the point $p = p_0 = 1/2$ in your plot). Do this separately for $n = 10, 100, 1000$. Use different line types (use the option `lty`) and use the `legend()` function to label the line types. Also, draw horizontal lines at $1 - \beta = 0.025$ and $1 - \beta = 0.05$. Explain the discrepancy when $p \approx p_0$. Given that we are generally interested in evaluating type II errors not greater than $\beta = 0.2$, do these approximations seem sufficiently accurate?

SOLUTION: The following code produces the power curves in Figure 2.5. The required power curves are

based on the following functions:

$$\begin{aligned}
\beta_U(p, p_0, n, \alpha) &= \Phi\left(\frac{z_\alpha\sqrt{p_0(1-p_0)} - \sqrt{n}(p-p_0)}{\sqrt{p(1-p)}}\right), \\
\beta_L(p, p_0, n, \alpha) &= 1 - \Phi\left(\frac{-z_\alpha\sqrt{p_0(1-p_0)} - \sqrt{n}(p-p_0)}{\sqrt{p(1-p)}}\right), \\
\beta_{two}(p, p_0, n, \alpha) &= \beta_U(p, p_0, n, \alpha/2) + \beta_L(p, p_0, n, \alpha/2) - 1, \\
\beta_{two}(p, p_0, n, \alpha) &\approx \beta_U(p \mid p_0, n, \alpha/2) \text{ for } p > p_0.
\end{aligned}$$

Here, Φ is the CDF for the standard normal distribution. The functions β_U, β_L, β_{two} are coded as functions `bu`, `bl`, `bb`.

We expect the approximate power curve to equal $1 - \beta = 0.025$ for p close to p_0, instead of the correct value of 0.05 (recall that α is the probability of rejecting the null hypothesis at $p = p_0$). However, this discrepancy is noticeable only near p_0. For values that would be of practical interest for a power calculation ($1 - \beta \geq 0.8$), the approximation is nearly exact.

```
### Functions bu, bl, bb correspond to Equations (17.4), (17.6) and (17.7)
### of INTRODUCTION TO STATISTICAL METHODOLOGY.

### Type II error for upper-tailed test

bu = function(p,p0,n,alpha) {
zc = qnorm(1-alpha)
sd0 = sqrt(p0*(1-p0))
sd1 = sqrt(p*(1-p))
ans = pnorm( (zc*sd0 - sqrt(n)*(p-p0))/sd1 )
return(ans)
}


### Type II error for lower-tailed test

bl = function(p,p0,n,alpha) {
zc = qnorm(1-alpha)
sd0 = sqrt(p0*(1-p0))
sd1 = sqrt(p*(1-p))
ans = 1 - pnorm( (-zc*sd0 - sqrt(n)*(p-p0))/sd1 )
return(ans)
}

### Type II error for two-sided test

bb = function(p,p0,n,alpha) { bu(p,p0,n,alpha/2) + bl(p,p0,n,alpha/2) - 1}

### construct power curves
```

```
x = seq(0.5,0.99,by=0.01)

ex1 = expression(italic(p))
ex2 = expression(1-beta)

### separate plots for n - 10, 100, 1000

# exact power curve will have lty = 1 (solid line)
# approximate power curve will have lty = 2 (dashed line)

par(mfrow=c(2,2),cex=1,oma=c(1,1,1,1))

n = 10
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)
legend('topleft',legend=c('Exact power','Approximate power'),lty=c(1,2))

n = 100
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)


n = 1000
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)
```

Problem 2.58 An industrial process produces a component with a diameter which is, due to process variability, normally distributed with mean $\mu = 35.5$ mm and standard deviation $\sigma = 0.043$ mm. After some time it is suspected that the process mean has lowered, so a random sample of n components is to be collected to do a hypothesis test for $H_o : \mu = 35.5$ against $H_a : \mu < 35.5$.

(a) Use `R` to draw a *power curve*, that is plot $Power(\mu) = 1 - \beta(\mu)$ as a function of μ over a suitable range of alternative hypotheses, say $\mu \in (35.4, 35.5)$ (use `seq(35.4,35.5,by=0.001)` to generate the values of μ for your plot). Do this for a Type I error of $\alpha = 0.05$ (refer to Section 13.3). Superimpose on the same plot power curves for $n = 5, 10, 15, 20, 25, 30$. Label the appropriate axes μ and $Power(\mu)$.

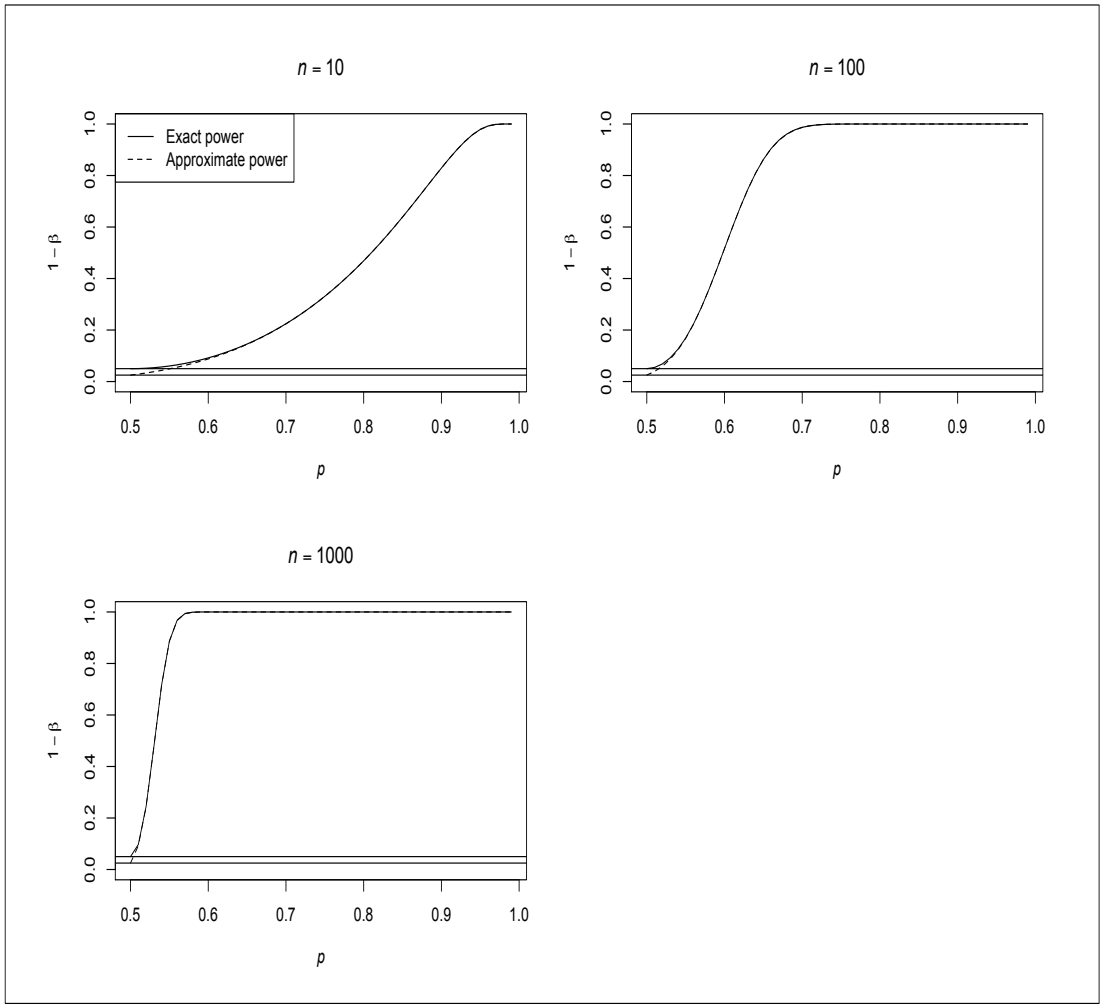

Figure 2.5: Plot for Problem 2.57.

Include a horizontal line at level 0.05, labelled $\alpha = 0.05$. Also, indicate, using the `text()` function, the positions of the $n = 5$ and $n = 30$ curves. What is the value of $Power(35.5)$ for each curve?

(b) Create a table giving the power for each combination of $n = 5, 10, \ldots, 30$ and $\mu = 35.47, 35.48, 35.49$. For each of these values of μ, give either the minimum sample size (from those considered) required to attain a power of 80% or state that the largest sample size considered is not sufficient.

SOLUTION: The following R code gives the required plot and tables:

```
mu0 = 35.5
sigma = 0.043
zalpha = qnorm(0.05)
mugrid = seq(35.4,35.5,by=0.001)
ex1 = expression(mu)
ex2 = expression(paste('Power(',mu,')',sep=''))
```

```
ex3 = expression(paste(alpha,' = 0.05',sep=''))
ex4 = expression(paste(italic(n),' = 5',sep=''))
ex5 = expression(paste(italic(n),' = 30',sep=''))

### use the following object to capture the values used in the plot

power.table = NULL

### start with an empty plot

plot(range(mugrid),c(0,1),type='n',xlab=ex1,ylab=ex2)

### loop through n = 5, 10, ..., 30

for (n in seq(5,30,by=5)) {
pow = pnorm(zalpha - (mugrid-mu0)*sqrt(n)/sigma)
lines(mugrid,pow,type='l')
power.table = cbind(power.table,pow)
}

### annotate plot

abline(h=0.05,col='gray')
text(35.42,0.09,ex3)
text(35.46,0.5,ex4)
text(35.49,0.8,ex5)

### create the table

rownames(power.table) = mugrid
colnames(power.table) = seq(5,30,by=5)
final.table = power.table[which(mugrid %in% c(35.47,35.48,35.49)),]
```

(a) The required plot is shown in Figure 2.6. For all plots, $Power(35.5) = 0.05$. This is because $Power$ is the probability of rejecting H_o, and if μ is the null hypothetical mean, than this is precisely the Type I error, which is set to $\alpha = 0.05$.

(b) The required table was produced by the preceding script, and is

```
> final.table
5         10        15        20        25        30
35.47 0.4662077 0.7127330 0.8547962 0.9299261 0.9673733 0.9852398
35.48 0.2726486 0.4309222 0.5621939 0.6682949 0.7519781 0.8166556
35.49 0.1303289 0.1815589 0.2283899 0.2726486 0.3148806 0.3552889
```

To attain a power of 80%, the minimum sample sizes are $n = 15, n = 30$ for $\mu = 35.47, 35.48$. None of the sample sizes considered attains a power of 80% for $\mu = 35.49$.

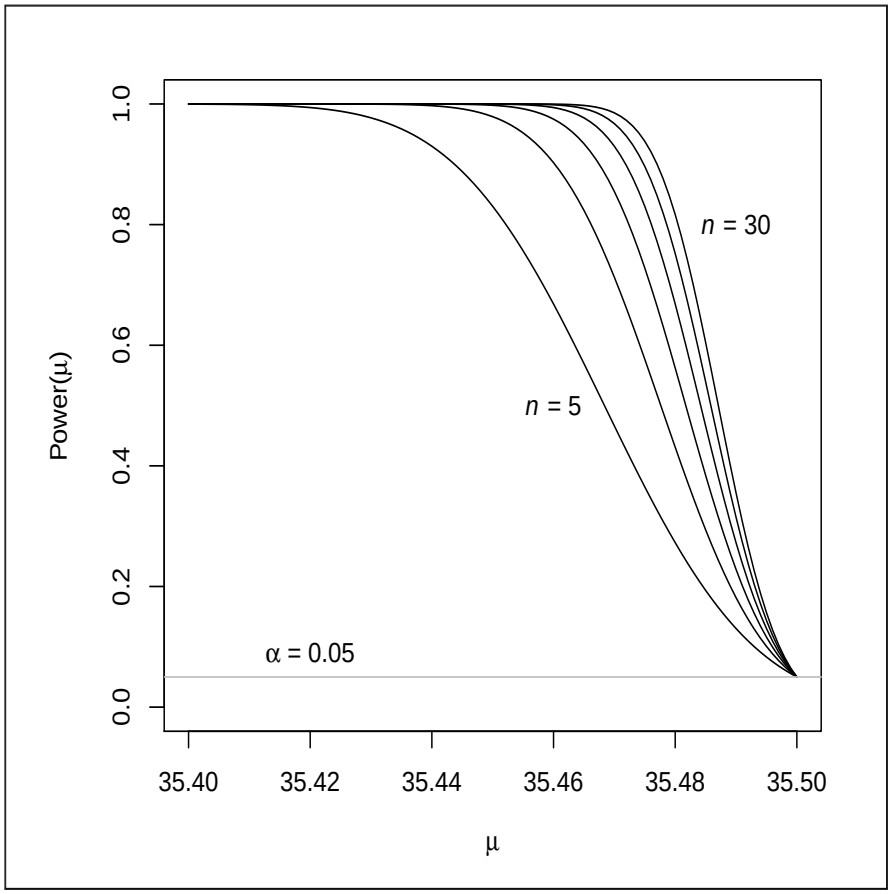

Figure 2.6: Plot for Problem 2.58 (a).

Problem 2.59 We wish to do a power analysis for a hypothesis test of $H_o : p = 1/2$ against $H_a : p > 1/2$ where p is a binomial proportion. It is anticipated that $n = 100$.

(a) Write `R` functions to implement the formula for Type II error, and for sample size, for an upper tailed test for proportions given in Equations (17.4) and (17.5) of Chapter 17 of `INTRODUCTION TO STATISTICAL METHODOLOGY`.

(b) Construct a power curve, using $n = 100$ and $\alpha = 0.05$. Label the axes appropriatetly. The vertical axis can say 'Power', and the horizontal axis should be labeled with 'p' in mathematical font. The plot should be titled

$$\text{Power } H_o : p = 0.5 \text{ vs } H_a : p > 0.5 \ \ (\alpha = 0.05)$$

The plot should be drawn on a grid generated by `seq(0.5,1,by = 0.01)`, and should be a line type (that is, use option `type='l'` in the `plot` function). Include a horizontal dashed line at Power = 0.05, and label this line '$\alpha = 0.05$'.

(c) Does the power curve suggest that there will be a power of 80% for alternative $p = 0.6$? If not, what sample size is required?

SOLUTION:

(a) The required subroutines are given below:

```
bu = function(p,p0,n,alpha) {

### Type II error for upper-tailed test for single proportion
### p0 = null proportion, p = alternative proportion
### alpha = Type I error, n = sample size

zc = qnorm(1-alpha)
sd0 = sqrt(p0*(1-p0))
sd1 = sqrt(p*(1-p))
ans = pnorm( (zc*sd0 - sqrt(n)*(p-p0))/sd1 )
return(ans)
}

nu = function(p,p0,alpha,beta) {

### Sample size for upper-tailed test for single proportion
### p0 = null proportion, p = alternative proportion
### alpha = Type I error, beta = Type II error

zalpha = qnorm(1-alpha)
zbeta = qnorm(1-beta)
sd0 = sqrt(p0*(1-p0))
sd1 = sqrt(p*(1-p))
ans = ( (zalpha*sd0 + zbeta*sd1)/(p-p0))^2
return(ans)
}
```

(b) The following script produces the required graph (Figure 2.7).

```
n = 100
p0 = 0.5
alpha = 0.05
pgrid = seq(0.5,1,by = 0.01)
ex1 = expression(paste(italic(p)))
ex2 = 'Power'
ex3 = expression(paste("Power ",italic(H)[o]," : ",italic(p)," = 0.5 vs ",
italic(H)[a]," : ",italic(p) > 0.5," (",alpha == 0.05,")",sep=""))
ex4 = expression(paste(alpha == 0.05,sep=""))
plot(pgrid, 1 - bu(pgrid,p0,n,alpha),type='l',ylim=c(0,1),
main=ex3,xlab=ex1,ylab=ex2)
abline(h = alpha,lty=2)
text(0.75, 0.1, ex4)
```

(c) From Figure 2.7 the power for $p = 0.6$ is clearly less than 0.8. We use subroutine `nu` to estimate the required sample size. After rounding up, we require a sample size of $n = 153$.

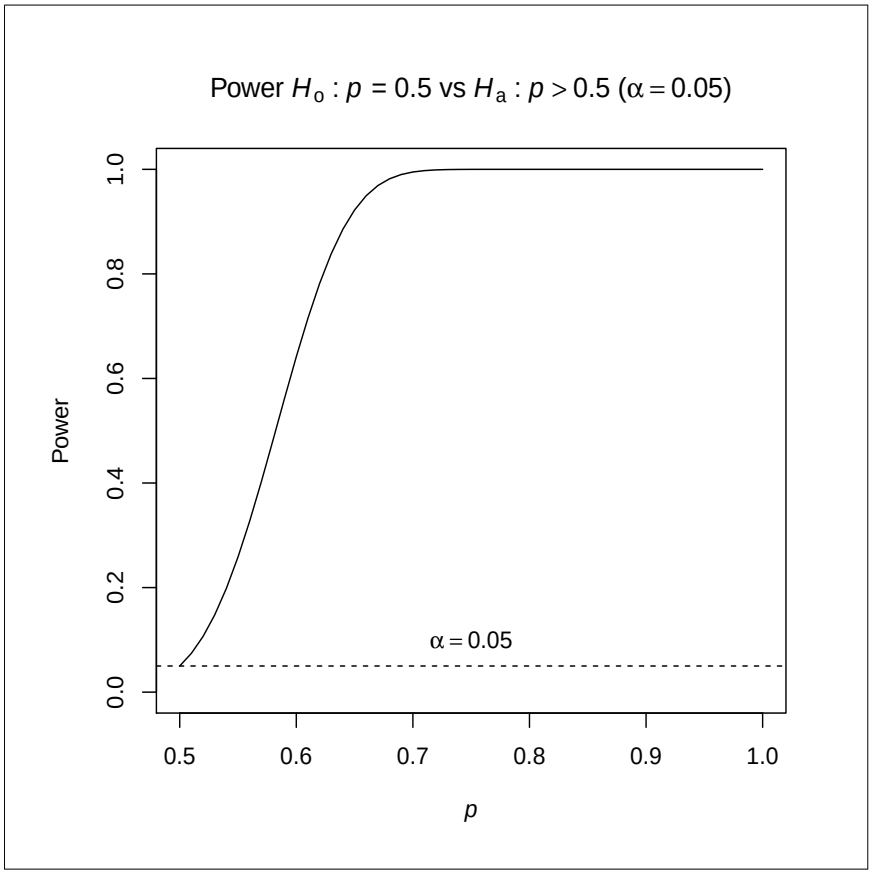

Figure 2.7: Power curve for Problem 2.59 (b).

```
> nu(0.6, 0.5, 0.05, 0.2)
[1] 152.4571
```

2.5 Inference for Variances

Problem 2.60 We are given samples of size $n_1 = 14$ and $n_2 = 18$ from independent normally distributed populations. Suppose we observe sample variances $S_1^2 = 645.16$ and $S_2^2 = 1413.76$. Do a hypothesis test of

$$H_o : \sigma_2^2 = \sigma_1^2$$
$$H_a : \sigma_2^2 \neq \sigma_1^2$$

using an $\alpha = 0.1$ significance level. Give explicitly the rejection regions, and also report a P-value.

SOLUTION: The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{645.16}{1413.76} = 0.456,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{13,17}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.1$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{13,17,0.95} \approx 0.4 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{13,17,0.05} \approx 2.353. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ and $F \leq F_{n_1-1,n_2-1,1-\alpha/2}$ we do not reject the null hypothesis at an $\alpha = 0.1$ significance level. The P-value is 0.157.

Problem 2.61 For an *iid* sample from a normal distribution we are given sample standard deviation $S = 25.3$, with sample size $n = 80$. Calculate a confidence interval for population standard deviation σ, using confidence level $1 - \alpha = 0.95$. Also give the level $1 - \alpha$ lower and upper confidence bounds.

SOLUTION: The level $1 - \alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{79,0.025} = 105.473 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{79,0.975} = 56.309.$$

The confidence interval is then given by

$$\frac{25.3}{\sqrt{105.473/79}} < \sigma < \frac{25.3}{\sqrt{56.309/79}}$$

or equivalently, $CI = (21.896, 29.967)$.

The level $1 - \alpha$ lower bound for σ is given by,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{79,0.05} = 100.749$. The lower bound is then given by,

$$\sigma > \frac{25.3}{\sqrt{100.749/79}} = 22.403.$$

The level $1 - \alpha$ upper bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{79,0.95} = 59.522$. The upper bound is then given by,

$$\sigma < \frac{25.3}{\sqrt{59.522/79}} = 29.147.$$

Problem 2.62 We are given two independent samples from normally distributed populations. The data is summarized in the table below.

	Sample 1	Sample 2
$\bar{X}_i$	12.2780	17.5310
S_i	0.1730	0.6820
n_i	5	10

(a) Use an F-test to test for equality of variances, using significance level $\alpha = 0.05$.
(b) Using the appropriate procedure based on the test for equality of variances (that is, either a pooled variance t-test or Welch's t-test), calculate a confidence interval for $\mu_1 - \mu_2$ with confidence level $1-\alpha = 0.95$.

SOLUTION:

(a) We have

$$F = \frac{0.1730^2}{0.6820^2} = 0.0643.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if F is less than or equal to $F_{1-\alpha/2,n_1-1,n_2-1} = 0.112$ or if F is greater than or equal to $F_{\alpha/2,n_1-1,n_2-1} = 4.718$, where $\alpha = 0.05$. Therefore, reject the null hypothesis of equal variances at a significance level $\alpha = 0.05$.

(b) We conclude that $\sigma_1^2 \neq \sigma_2^2$, so use Welch's procedure for unequal variances. The degrees of freedom is given by

$$\begin{aligned}
\nu_W &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}} \\
&= \frac{\left(\frac{0.1730^2}{5} + \frac{0.6820^2}{10}\right)^2}{\frac{(0.1730^2/5)^2}{5-1} + \frac{(0.6820^2/10)^2}{10-1}} \\
&= 11.05362.
\end{aligned}$$

Round down to $\nu_W = 11$ degrees of freedom. The confidence interval is then

$$\begin{aligned}
CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{\nu_W,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\
&= \bar{X}_1 - \bar{X}_2 \pm t_{11,0.025}\sqrt{\frac{0.1730^2}{5} + \frac{0.6820^2}{10}} \\
&= 12.2780 - 17.5310 \pm 2.201 \times 0.229 \\
&= -5.253 \pm 0.504 \\
&= (-5.757, -4.749).
\end{aligned}$$

The confidence interval is (-5.76,-4.75) or -5.25 ± 0.504.

Problem 2.63 Two measuring instruments are to be compared. We are given from each samples of size $n_1 = 31$ and $n_2 = 12$. We assume they are independent normally distributed samples. Suppose we observe sample variances $S_1^2 = 0.0729$ and $S_2^2 = 0.0121$. Do a hypothesis test of

$$H_o : \sigma_2^2 = \sigma_1^2$$
$$H_a : \sigma_2^2 \neq \sigma_1^2$$

using an $\alpha = 0.05$ significance level. Do the variances of the instruments differ? Give explicitly the rejection regions, and also report a P-value.

SOLUTION: The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{0.0729}{0.0121} = 6.025,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{30,11}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.05$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{30,11,0.975} \approx 0.407 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{30,11,0.025} \approx 3.118. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ we reject the null hypothesis at an α significance level, and we conclude that the instrument's variances differ. The P-value is 0.00321.

Problem 2.64 Consider the following *iid* sample from a normal distribution:

$$100.086, 99.954, 100.242, 99.835, 99.954, 100.083$$

with sample size $n = 6$. Calculate a confidence interval for population standard deviation σ, using confidence level $1 - \alpha = 0.95$. Also give the level $1 - \alpha = 0.95$ lower and upper confidence bounds.

SOLUTION: We have $S = 0.142$. The level $1 - \alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

The appropriate critical values are

$$\chi^2_{n-1,\alpha/2} = \chi^2_{5,0.025} = 12.833 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{5,0.975} = 0.831.$$

The confidence interval is then given by

$$\frac{0.142}{\sqrt{12.833/5}} < \sigma < \frac{0.142}{\sqrt{0.831/5}}$$

or equivalently,

$$CI = (0.0884, 0.347) = \sqrt{(0.0078, 0.1207)}.$$

The level $1-\alpha$ lower bound for σ is given by

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is

$$\chi^2_{n-1,\alpha} = \chi^2_{5,0.05} = 11.07.$$

The lower bound is then given by

$$\sigma > \frac{0.142}{\sqrt{11.07/5}} = 0.0952 = \sqrt{0.00906}.$$

The level $1-\alpha$ upper bound for σ is given by

$$\sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is

$$\chi^2_{n-1,1-\alpha} = \chi^2_{5,0.95} = 1.145.$$

The upper bound is then given by

$$\sigma < \frac{0.142}{\sqrt{1.145/5}} = 0.296 = \sqrt{0.0876}.$$

Problem 2.65 We are given two independent samples from normally distributed populations. The data is summarized in the following table:

	Sample 1	Sample 2
$\bar{X}$	101.379	166.446
S	8.211	8.665
n	25	50

(a) Test for equality of variances, using significance level $\alpha = 0.05$.
(b) Perform a two-sided hypothesis test using hypotheses

$$H_o: \mu_1 - \mu_2 = 0 \text{ against } H_a: \mu_1 - \mu_2 \neq 0.$$

Use significance level $\alpha = 0.05$. Use a two-sample t-test, using the conclusion of part (a) to guide your choice of method.

SOLUTION:

(a) We have

$$F = S_1^2/S_2^2 = 0.898.$$

Reject $H_o: \sigma_1^2 = \sigma_2^2$ if

$$F \leq F_{1-\alpha/2,n_1-1,n_2-1} = 0.474 \text{ or } F \geq F_{\alpha/2,n_1-1,n_2-1} = 1.937$$

where $\alpha = 0.05$. Therefore, do not reject the null hypothesis of equal variances at a signficance level $\alpha = 0.05$. P-value $= 0.7941$.

(b) Use the pooled procedure with $\nu = 73$ degrees of freedom.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{24 \times 8.211^2 + 49 \times 8.665^2}{73} = 72.56$$

or $S_p = 8.51841$. Then

$$T = \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{101.379 - 166.446}{8.51841\sqrt{\frac{1}{25}+\frac{1}{50}}} = -31.18363.$$

Reject H_o if

$$|T| \geq t_{\nu,\alpha/2} = 1.993.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value = 6.184e-44.

Problem 2.66 We are given an *iid* sample from a normal distribution

$$91.84, 103.35, 94.56, 95.42, 105.72,$$

of sample size $n = 5$. Calculate a confidence interval for population standard deviation σ, using confidence level $1-\alpha = 0.95$. Also give the level $1-\alpha = 0.95$ lower confidence bound.

SOLUTION: The sample standard deviation is $S = 6.01$. The level $1-\alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{4,0.025} = 11.143 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{4,0.975} = 0.484.$$

The confidence interval is then given by

$$\frac{6.01}{\sqrt{11.143/4}} < \sigma < \frac{6.01}{\sqrt{0.484/4}}$$

or equivalently, $CI = (3.601, 17.271)$.

The level $1-\alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{4,0.05} = 9.488$. The lower bound is then given by,

$$\sigma > \frac{6.01}{\sqrt{9.488/4}} = 3.903.$$

Problem 2.67 We are given two independent samples from normally distributed populations $N(\mu_i, \sigma_i^2)$, $i = 1, 2$. The data is summarized in the following table:

	Sample 1	Sample 2
$\bar{X}_i$	123.89	153.69
S_i	8.10	9.67
n_i	23	54

(a) Perform a two-sided hypothesis test for null hypothesis $H_o : \sigma_1^2 = \sigma_2^2$ against alternative $H_a : \sigma_1^2 \neq \sigma_2^2$. Use significance level $\alpha = 0.05$.

(b) Construct a level $1 - \alpha = 0.9$ confidence interval for $\mu_2 - \mu_1$. Use the conclusion of part (a) to choose between the pooled procedure for equal variances or Welch's procedure for unequal variances.

SOLUTION:

(a) Use statistic

$$F = \frac{S_1^2}{S_2^2} = \frac{8.10^2}{9.67^2} = 0.702.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F \leq F_{1-\alpha/2, n_1-1, n_2-1} = 0.463 \text{ or } F \geq F_{\alpha/2, n_1-1, n_2-1} = 1.943.$$

Therefore, do not reject the null hypothesis of equal variances at a signficance level $\alpha = 0.05$.

(b) Use the pooled procedure with $\nu = n_1 + n_2 - 2 = 75$ degrees of freedom. Pooled variance is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{22 \times 8.10^2 + 53 \times 9.67^2}{75} = 85.325.$$

The confidence intervals

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2, \alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 153.69 - 123.89 \pm 1.665 \times 9.237 \sqrt{\frac{1}{23} + \frac{1}{54}} \\ &= 29.80 \pm 3.83 \\ &= (25.97, 33.63). \end{aligned}$$

Problem 2.68 We are given two independent samples from normally distributed populations. The sample means, sample standard deviations and sample sizes are summarized in the following table:

	Sample 1	Sample 2
$\bar{X}_i$	44.673	44.299
S_i	0.782	0.923
n_i	35	36

(a) Use an F-test to test for equality of variances, using significance level $\alpha = 0.05$.

(b) Using the appropriate procedure based on the test for equality of variances, calculate a confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha = 0.99$.

SOLUTION:

(a) Use statistic

$$F = \frac{S_1^2}{S_2^2} = \frac{0.782^2}{0.923^2} = 0.718.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F \leq F_{1-\alpha/2,n_1-1,n_2-1} = 0.506 \text{ or } F \geq F_{\alpha/2,n_1-1,n_2-1} = 1.968$$

Therefore, do not reject the null hypothesis of equal variances at a signficance level $\alpha = 0.05$ (P-value $= 0.3361$).

(b) Use the pooled procedure with $\nu = n_1 + n_2 - 2 = 69$ degrees of freedom. Pooled variance is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 0.7335$$

The confidence intervals

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= \bar{X}_2 - \bar{X}_1 \pm t_{69,0.005} S_p \sqrt{\frac{1}{35} + \frac{1}{36}} \\ &= 44.299 - 44.673 \pm 2.649\sqrt{0.7335}\sqrt{\frac{1}{35} + \frac{1}{36}} \\ &= 0.374 \pm 0.539 = (-0.913, 0.165). \end{aligned}$$

Problem 2.69 For an *iid* sample from a normal distribution we are given sample standard deviation $S = 13.65$, with sample size $n = 68$. Calculate a confidence interval for population standard deviation σ, using confidence level $1 - \alpha = 0.95$. Also give the level $1 - \alpha$ lower and upper confidence bounds.

SOLUTION: The level $1 - \alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{67,0.025} = 91.519 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{67,0.975} = 46.261.$$

The confidence interval is then given by

$$\frac{13.65}{\sqrt{91.519/67}} < \sigma < \frac{13.65}{\sqrt{46.261/67}}$$

or equivalently, $CI = (11.679, 16.427)$.

The level $1-\alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{67,0.05} = 87.108$. The lower bound is then given by,

$$\sigma > \frac{13.65}{\sqrt{87.108/67}} = 11.971.$$

The level $1-\alpha$ upper bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{67,0.95} = 49.162$. The upper bound is then given by,

$$\sigma < \frac{13.65}{\sqrt{49.162/67}} = 15.935.$$

Problem 2.70 We are given an *iid* sample from a normal distribution

$$64.2, 26.8, 40.4, 51.2, 30.7,$$

of sample size $n = 5$. Calculate a confidence interval for population standard deviation σ, using confidence level $1-\alpha = 0.9$. Also give the level $1-\alpha$ lower and upper confidence bounds.

SOLUTION: The sample standard deviation is $S = 15.302$. The level $1-\alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{4,0.05} = 9.488 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{4,0.95} = 0.711.$$

The confidence interval is then given by

$$\frac{15.302}{\sqrt{9.488/4}} < \sigma < \frac{15.302}{\sqrt{0.711/4}}$$

or equivalently, $CI = (9.936, 36.302)$.

The level $1-\alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{4,0.1} = 7.779$. The lower bound is then given by,

$$\sigma > \frac{15.302}{\sqrt{9.488/4}} = 10.972.$$

The level $1-\alpha$ upper bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{4,0.9} = 1.064$. The upper bound is then given by,

$$\sigma < \frac{15.302}{\sqrt{0.711/4}} = 29.674.$$

Problem 2.71 A sample of size $n = 6$ from a normal distribution $N(\mu, \sigma^2)$ is collected:

$$X = 100.583, 99.600, 100.045, 98.963, 100.313, 100.097.$$

(a) Construct a level 95% confidence interval for σ. Verify your answer using R.
(b) Construct a level 95% upper confidence bound for σ. Verify your answer using R.
(c) Suppose the data are obtained from a pilot study. The object is to estimate the sample size required to estimate μ from the same population with a margin of error of E_0, with confidence level $1-\alpha$. Using the upper bound of the confidence interval of part (a), it is determined that a sample size of $n = 100$ is required. If the upper confidence bound of part (b) had been used instead, what would the estimated sample size have been?

SOLUTION:

(a) The level $1-\alpha$ confidence interval is given by

$$\sigma \in \left(\frac{S_n}{\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}}}, \frac{S_n}{\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}} \right).$$

The code below calculates the quantities $n = 6$, $S = 0.5758693$, $\chi^2_{1-\alpha/2,n-1} = 0.8312116$, $\chi^2_{\alpha/2,n-1} = 12.8325020$, giving level 95% confidence interval $(0.3594623, 1.4123852)$.

```
> x = c(100.583,99.600,100.045,98.963,100.313,100.097)
> n = length(x)
>
> # critical values from chi.sq distribution with df = n-1
>
> cl = qchisq(0.025,df=n-1)
> cu = qchisq(0.975,df=n-1)
```

```
>
> # construct level 95% confidence  interval
>
> sd0 = sd(x)
> ci = sd0*sqrt((n-1)*c(1/cu,1/cl))
> c(n,sd0,cl,cu,ci)
[1]  6.0000000  0.5758693  0.8312116 12.8325020  0.3594623  1.4123852
>
```

(b) The level $1-\alpha$ upper confidence bound is given by

$$\sigma \leq \frac{S_n}{\sqrt{\frac{\chi^2_{1-\alpha,n-1}}{n-1}}}.$$

The code below calculates $\chi^2_{1-\alpha,n-1} = 1.145476$, giving level 95% upper confidence bound 1.203139.

```
>
> # construct level 95% upper confidence bound
>
> cl = qchisq(0.05,df=n-1)
> ucb = sd0*sqrt((n-1)/cl)
> c(cl,ucb)
[1] 1.145476 1.203139
```

(c) The formula for the sample size required for a level $1-\alpha$ confidence level with margin of error E_o is

$$n = \left(z_{\alpha/2}\frac{\sigma}{E_o}\right)^2$$

To compare the sample sizes n_1, n_2 based on two alternative standard deviations σ_1, σ_2, we take the ratio

$$\frac{n_2}{n_1} = \frac{\left(z_{\alpha/2}\frac{\sigma_2}{E_o}\right)^2}{\left(z_{\alpha/2}\frac{\sigma_1}{E_o}\right)^2} = \frac{\sigma_2^2}{\sigma_1^2}.$$

In this case, we have

$$\frac{n_2}{n_1} = \frac{\sigma_2^2}{\sigma_1^2} \approx \frac{1.203^2}{1.412^2} = 0.726.$$

So, if we originally needed $n = 100$, using the upper confidence bound reduces the sample size estimate to 72.6, rounded up to $n = 73$.

Problem 2.72 We are given samples of size $n_1 = 75$ and $n_2 = 45$ from independent normally distributed populations. Suppose we observe sample variances $S_1^2 = 412.09$ and $S_2^2 = 243.36$. Do a hypothesis test of

$$H_o : \sigma_2^2 = \sigma_1^2$$
$$H_a : \sigma_2^2 \neq \sigma_1^2$$

using an $\alpha = 0.05$ significance level. Give explicitly the rejection regions, and also report a P-value.

SOLUTION: The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{412.09}{243.36} = 1.693,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{74,44}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.05$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{74,44,0.975} \approx 0.597 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{74,44,0.025} \approx 1.736. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ and $F \leq F_{n_1-1,n_2-1,1-\alpha/2}$ we do not reject the null hypothesis at an α significance level. The P-value is 0.0611.

Problem 2.73 We are given samples of size $n_1 = 75$ and $n_2 = 103$ from independent normally distributed populations. Suppose we observe sample variances $S_1^2 = 299.29$ and $S_2^2 = 158.76$. Do a hypothesis test of

$$\begin{aligned} H_o &: \sigma_2^2 = \sigma_1^2 \\ H_a &: \sigma_2^2 \neq \sigma_1^2 \end{aligned}$$

using an $\alpha = 0.05$ significance level. Give explicitly the rejection regions, and also report a P-value.

SOLUTION: The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{299.29}{158.76} = 1.885,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{74,102}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.05$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{74,102,0.975} \approx 0.648 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{74,102,0.025} \approx 1.52. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ we reject the null hypothesis at an α significance level. The P-value is 0.00304.

Problem 2.74 Write an `R` function that accepts two samples as arguments, and returns the respective sample variances and sample sizes, the F statistic appropriate for an equality of variance test, and the p-value for a two-sided test against null hypothesis $H_o : \sigma_1^2 = \sigma_2^2$. Test your function using the following two samples.

```
x = c(6.00, 6.95,8.72,10.83)
y = c(19.38, 19.39, 20.04, 20.57, 19.91, 20.05, 19.82, 20.20)
```

SOLUTION: The following function implements the required F test

```
ftest = function(x,y) {

n1 = length(x)
n2 = length(x)
```

```
var1 = var(x)
var2 = var(y)ffig2


fstat = var1/var2
pval = 2*min(pf(fstat,n1-1,n2-1),1-pf(fstat,n1-1,n2-1))
return(c(var1=var1, var2=var2, n1=n1, n2=n2, fstat=fstat, pval=pval))
}
```

The following code gives the required example:

```
> x = c(6.00, 6.95,8.72,10.83)
> y = c(19.38, 19.39, 20.04, 20.57, 19.91, 20.05, 19.82, 20.20)
> ftest(x,y)
var1        var2          n1          n2       fstat        pval
4.52243333  0.15925714  4.00000000  4.00000000 28.39705179  0.02108341
>
```

Problem 2.75 We are given two independent samples from normally distributed populations, with means and variances $N(\mu_i, \sigma_i^2)$, $i = 1, 2$. The data is summarized in the following table:

	Sample 1	Sample 2
$\bar{X}$	45.93	59.55
S	11.91	12.56
n	11	16

(a) Test the null and alternative hypotheses $H_o : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$. Use significance level $\alpha = 0.05$.

(b) Perform a two-sided hypothesis test for null and alternative hypotheses $H_o : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 \neq 0$. Use an appropriate T-statistic, basing your choice on the conclusion reached in Part (a). Use significance level $\alpha = 0.05$.

SOLUTION:

(a) The F-statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{11.91^2}{12.56^2} = 0.899.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F < F_{n_1-1,n_2-1;\alpha/2} = F_{10,15;0.025} = 3.06,$$

or

$$F > F_{n_1-1,n_2-1;1-\alpha/2} = \frac{1}{F_{n_2-1,n_1-1;\alpha/2}} = \frac{1}{F_{15,10;0.025}} = \frac{1}{3.52} = 0.284.$$

Therefore, do not reject the null hypothesis of equal variances at a significance level $\alpha = 0.05$.

(b) From Part (a) we assume $\sigma_1^2 = \sigma_2^2$ so use the pooled variance procedure with $\nu = n_1 + n_2 - 2 = 25$ degrees of freedom.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{10 \times 11.91^2 + 15 \times 12.56^2}{11+16-2} = 151.39,$$

or $S_p = 12.304$. The T-statistic is then

$$T = \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{59.55 - 45.93}{12.304\sqrt{\frac{1}{11}+\frac{1}{16}}} = 2.826$$

Reject H_o if $|T| > t_{25;0.025}$, where

$$t_{\nu,\alpha/2} = t_{25;0.025} = 2.06.$$

Therefore, reject the null hypothesis at a significance level $\alpha = 0.05$.

2.6 Inference for Correlations

Problem 2.76 A sample correlation coefficient of $r = 0.62$ is observed from $n = 72$ paired observations.

(a) Let ρ be the population correlation coefficient. Test the null hypothesis $H_o : \rho = 0$ against alternative hypothesis $H_a : \rho \neq 0$ using a suitable t-statistic. Assume the samples have a bivariate normal distribution. Can we reject H_o with significance level $\alpha = 0.05$?

(b) Construct a level 95% confidence interval for ρ.

SOLUTION: The sample correlation is $r = 0.62$. We transform to T-statistic

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.62}{\sqrt{\frac{1-0.62^2}{72-2}}} = 6.611.$$

(a) Under H_o, T has a t-distribution with $n-2 = 70$ degrees of freedom, so reject H_o at significance level α if

$$|T| \leq t_{n-2;\alpha/2} = 1.994.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

(b) We use transformation

$$\begin{aligned} V_{obs} &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \\ &= \frac{1}{2}\ln\left(\frac{1+0.62}{1-0.62}\right) = 0.725, \end{aligned}$$

so we first use the confidence interval for $\mu_{V,\rho}$:

$$\begin{aligned} \left(V_{obs} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, V_{obs} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) &= \left(0.725 - \frac{1.96}{\sqrt{69}}, 0.725 + \frac{1.96}{\sqrt{69}}\right) \\ &= (0.489, 0.961), \end{aligned}$$

giving $L_V = 0.489$, $U_V = 0.961$. The confidence interval for ρ is then directly given by inverting the Fisher transformation,then substituting L_V and U_V:

$$\left(\frac{e^{2L_V}-1}{e^{2L_V}+1}, \frac{e^{2U_V}-1}{e^{2U_V}+1}\right) = (0.453, 0.744).$$

Problem 2.77 Using the data from Problem 2.81, we wish to test the null hypothesis $H_o : \rho = 0$ against alternative hypothesis $H_a : \rho \neq 0$, where ρ is the correlation between the two samples. Assume the samples have a bivariate normal distribution.

(a) Calculate a P-value for this test. Can we reject H_o with significance level $\alpha = 0.05$?
(b) In general, for this sample size how large must sample correlation r be in order to reject H_o against a two-sided alternative with $\alpha = 0.05$?
(c) Does the conclusion here in any way contradict the conclusion of Problem 2.81?

SOLUTION: The sample correlation is $r = 0.892$. We transform to T-statistic

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.892}{\sqrt{\frac{1-0.892^2}{9-2}}} = 5.221.$$

(a) Under H_o, T has a t-distribution with $n-2 = 7$ degrees of freedom, so

$$\alpha_{obs} = 2P(T_7 \leq -5.221) = 2 \times 0.0006122 = 0.00122,$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.
(b) The critical value for the t-statistic is $t_{0.025} = 2.3646$. Inverting the transformation gives

$$r_{0.025} = \frac{t_{0.025}}{\sqrt{n-2+t^2_{0.025}}} = \frac{2.3646}{\sqrt{9-2+2.3646^2}} = 0.666.$$

We would reject H_o is $|r| \geq 0.666$.
(c) No. Under the relevant assumption, correlation is independent of differences in mean, or any other measure of location such as the median.

Problem 2.78 Using the data from Problem 2.93, we wish to test the null hypothesis $H_o : \rho = 0$ against alternative hypothesis $H_a : \rho \neq 0$, where ρ is the correlation between the two samples. Assume the samples have a bivariate normal distribution.

(a) Calculate a p-value for this test. Can we reject H_o with significance level $\alpha = 0.05$?
(b) In general, for this sample size how large must sample correlation r be in order to reject H_o against a two-sided alternative with $\alpha = 0.05$?
(c) Construct a level 95% confidence interval for ρ.
(d) Does the conclusion here in any way contradict the conclusion of Problem 2.93?

SOLUTION: The sample correlation is $r = 0.874$. We transform to T-statistic

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.874}{\sqrt{\frac{1-0.874^2}{8-2}}} = 4.412.$$

(a) Under H_o, T has a t-distribution with $n-2=6$ degrees of freedom, so

$$\alpha_{obs} = 2P(T_7 \leq -4.412) = 2 \times 0.00225 = 0.0045,$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

(b) The critical value for the t-statistic is $t_{6,0.025} = 2.447$. Inverting the transformation gives

$$r_{0.025} = \frac{t_{0.025}}{\sqrt{n-2+t^2_{0.025}}} = \frac{2.447}{\sqrt{8-2+2.447^2}} = 0.707.$$

We would reject H_o is $|r| \geq 0.707$.

(c) We use transformation

$$\begin{aligned} V_{obs} &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \\ &= \frac{1}{2}\ln\left(\frac{1+0.874}{1-0.874}\right) = 1.351, \end{aligned}$$

so we first use the confidence interval for $\mu_{V,\rho}$:

$$\begin{aligned} \left(V_{obs} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, V_{obs} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) &= \left(1.351 - \frac{1.96}{\sqrt{5}}, 1.351 + \frac{1.96}{\sqrt{5}}\right) \\ &= (0.474, 2.228), \end{aligned}$$

giving $L_V = 0.474$, $U_V = 2.228$. The confidence interval for ρ is then directly given by inverting the Fisher transformation,then substituting L_V and U_V:

$$\left(\frac{e^{2L_V}-1}{e^{2L_V}+1}, \frac{e^{2U_V}-1}{e^{2U_V}+1}\right) = (0.442, 0.977).$$

(d) No. Under the relevant assumption, correlation is independent of differences in mean, or any other measure of location such as the median.

2.7 Nonparametric Methods

Problem 2.79 We are given two paired samples of sample size $n = 9$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided sign test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$.

SOLUTION: After excluding ties there are X =1 positive differences among $n' = 8$ pairs. The P-value is given by

$$\begin{aligned} \alpha_{obs} &= P(X \leq 1) + P(X \geq 7) \\ &= p_0 + p_1 + p_7 + p_8 \\ &= 0.00390625 + 0.03125 + 0.03125 + 0.00390625 = 0.0703125, \end{aligned}$$

where p_i is the PGF for a $bin(n = 8, p = 0.5)$ distribution. Therefore, we do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	15.9	19.5	-3.6	-
2	17.8	16.9	+0.9	+
3	13.6	16.2	-2.6	-
4	14.8	17.9	-3.1	-
5	15.3	15.3	0.0	0
6	14.4	17.6	-3.2	-
7	13.7	16.3	-2.6	-
8	13.9	16.9	-3.0	-
9	14.8	16.4	-1.6	-

Problem 2.80 We are given two paired samples of sample size $n = 13$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform an upper tailed sign test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D > 0$. Use significance level $\alpha = 0.05$. Calculate a P-value using the exact binomial distribution.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	34.1	25.6	8.5	
2	24.0	20.9	3.1	
3	37.8	36.5	1.3	
4	26.4	15.1	11.3	
5	35.5	34.5	1.0	
6	30.0	15.8	14.2	
7	25.5	19.1	6.4	
8	22.4	8.6	13.8	
9	25.5	16.7	8.8	
10	27.6	16.8	10.8	
11	34.1	21.9	12.2	
12	33.5	17.8	15.7	
13	31.6	36.1	-4.5	

SOLUTION: The sample median of the differences is $\tilde{D} = 8.8$. There are $X = 12$ positive differences among $n' = 13$ pairs (there are no ties). The $P-$value is

$$\begin{aligned} P &= P(X \geq 12) = P(X \leq 1) \\ &= (1-p)^{13} + np(1-p)^{12} \\ &= (1/2)^{13}(1+13) = 14/8192 = 0.0017, \end{aligned}$$

since under H_o we have $X \sim bin(13, 1/2)$, and $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.81 A study is conducted to assess whether or not a cancer treatment increases stress in patients. Nine subjects are enrolled in the study. A psychometric test for stress is given to each subject before and after the treatment. The test scores are given in the table below. Larger scores predict higher stress. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a lower tailed

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	34.1	25.6	8.5	+
2	24.0	20.9	3.1	+
3	37.8	36.5	1.3	+
4	26.4	15.1	11.3	+
5	35.5	34.5	1.0	+
6	30.0	15.8	14.2	+
7	25.5	19.1	6.4	+
8	22.4	8.6	13.8	+
9	25.5	16.7	8.8	+
10	27.6	16.8	10.8	+
11	34.1	21.9	12.2	+
12	33.5	17.8	15.7	+
13	31.6	36.1	-4.5	-

signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D < 0$. Use significance level $\alpha = 0.05$. Do we conclude that the treatment increases stress? Verify your answer with the `R wilcox.test` function (do not use continuity correction).

	Before Treatment (X)	After Treatment (Y)	Difference ($D = X - Y$)	Sign
1	6.1	8.2	-2.1	-
2	7.5	6.7	0.8	+
3	5.3	3.8	1.5	+
4	6.6	12.9	-6.3	-
5	12.8	18.5	-5.7	-
6	4.5	1.3	3.2	+
7	4.0	6.1	-2.1	-
8	12.6	17.5	-4.9	-
9	12.7	14.6	-1.9	-

SOLUTION: The signed ranks are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D \rvert$	Sign
1	6.1	8.2	-2.1	4.5	-
2	7.5	6.7	0.8	1.0	+
3	5.3	3.8	1.5	2.0	+
4	6.6	12.9	-6.3	9.0	-
5	12.8	18.5	-5.7	8.0	-
6	4.5	1.3	3.2	6.0	+
7	4.0	6.1	-2.1	4.5	-
8	12.6	17.5	-4.9	7.0	-
9	12.7	14.6	-1.9	3.0	-

After excluding ties there are $n' = 9$ pairs remaining. The positive and negative rank sums are,

respectively,

$$T_+ = 1 + 2 + 6 = 9 \text{ and} T_- = (n+1)n/2 - T_+ = 36 - 9 = 27.$$

Giving statistic

$$T_{obs} = \min(T_-, T_+) = \min(9, 27) = 9.$$

The lower tail probability, for $n = 9$ is

$$\alpha_{obs} = P(T_{obs} \leq 9) = 0.06445,$$

noting that $T_+ < T_-$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

Using the normal approximation, we have

$$\mu_T = n(n+1)/4 = 22.5 \text{ and } \sigma_T = \sqrt{n(n+1)(2n+1)/24} = 8.441.$$

This gives Z-score

$$Z = (T_+ - \mu_T)/\sigma_T = -1.6,$$

giving

$$\alpha_{obs} = P(Z < -1.6) = 0.0548.$$

Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

The `wilcox.test` function should be used in the following way, giving the same conclusion.

```
> x = c(6.1,7.5,5.3,6.6,12.8,4.5,4.0,12.6,12.7)
> y = c(8.2,6.7,3.8,12.9,18.5,1.3,6.1,17.5,14.6)
> wilcox.test(x,y,alternative = 'less', paired=T, correct=F)

Wilcoxon signed rank test

data:  x and y
V = 9, p-value = 0.05472
alternative hypothesis: true location shift is less than 0

Warning message:
In wilcox.test.default(x, y, alternative = "less", paired = T, correct = F) :
cannot compute exact p-value with ties
>
```

Problem 2.82 A weight loss program is given to $n = 7$ participants of age 11 years. The weight in pounds for each subject before and after the program are given in the following table. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = Y - X$. Perform a lower tailed signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D < 0$. Use significance level $\alpha = 0.05$. Use both the exact signed rank distribution and the normal approximation. Verify your answer using the `wilcox.test()` function. Do not use any continuity correction.

SOLUTION: The ranks and signs are given in the following table

The sample median of the differences is $\tilde{D}$ = -4.1. The negative and positive rank sums are, respectively, $T_+ = 7$ and $T_- = 21$. Reject H_0 for small values of T_+. Using the exact sign rank distribution we obtain P-value = 0.1484. This is obtained either from tables, or from the `R` function:

	Before Program (X)	After Program (Y)	Difference ($D = Y - X$)
1	91.4	85.6	-5.8
2	101.1	94.6	-6.5
3	97.7	101.8	+4.1
4	95.5	87.8	-7.7
5	100.7	96.6	-4.1
6	102.6	105.5	+2.9
7	84.0	89.1	+5.1

	Before Program (X)	After Program (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	91.4	85.6	-5.8	5.0	-
2	101.1	94.6	-6.5	6.0	-
3	97.7	101.8	+4.1	2.0	+
4	95.5	87.8	-7.7	7.0	-
5	100.7	96.6	-4.1	3.0	-
6	102.6	105.5	+2.9	1.0	+
7	84.0	89.1	+5.1	4.0	+

```
> psignrank(7,7)
[1] 0.1484375
```

The mean and standard deviation of the negative or positive rank sums are $\mu_T = 14$ and $\sigma_T = 5.916$. This gives Z-score $Z = (T_+ - \mu_T)/\sigma_T = (7 - 14)/5.916 = -1.183$. Using the normal approximation gives P-value $= P(Z < -1.183) = 0.118$.

By either method we do not reject H_o with significance level $\alpha = 0.05$.

The following `R` code can be used to perform the test. Note that the `wilcox.test()` function calculates the differences as $X - Y$, so the option `alternative='greater'` should be used.

```
> x = c(91.4,101.1,97.7,95.5,100.7,102.6,84.0)
> y = c(85.6,94.6,101.8,87.8,96.6,105.5,89.1)
> wilcox.test(x,y,paired=T,alternative='greater')

Wilcoxon signed rank test

data:  x and y
V = 21, p-value = 0.1484
alternative hypothesis: true location shift is greater than 0
```

The p-value 0.1484 is the same as that obtained using the exact procedure.

Problem 2.83 We are given two independent samples of sample sizes $n_1 = 7$, $n_2 = 10$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use significance level $\alpha = 0.05$. Use the normal approximation method only. Verify your answer using the `wilcox.test()` function.

Do not use the continuity correction in either case. What method does `wilcox.test()` use when there are ties?

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	31.7	20.9	23.2	30.2	34.4	31.0	26.9				30.2
Sample 2	35.0	40.8	39.8	30.2	40.9	32.7	38.9	35.3	35.2	38.9	37.1

SOLUTION: The pooled ranks are given in the following table:

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	31.7	20.9	23.2	30.2	34.4	31.0	26.9				30.2
Sample 2	35.0	40.8	39.8	30.2	40.9	32.7	38.9	35.3	35.2	38.9	37.1
Ranks 1	7.0	1.0	2.0	4.5	9.0	6.0	3.0				32.5
Ranks 2	10.0	16.0	15.0	4.5	17.0	8.0	13.5	12.0	11.0	13.5	120.5

The sample medians are $\tilde{X}_1 = 30.2$ and $\tilde{X}_2 = 37.1$. The rank sums for samples 1 and 2 are, respectively, $T_1 = 32.5$ and $T_2 = 120.5$. The adjusted rank sum is $W = T_1 - n_1(n_1+1)/2 = 4.5$. The mean and standard deviation of T_1 are $\mu_{T_1} = 63$ and $\sigma_{T_1} = 10.247$. This gives Z-score $Z = (T_1 - \mu_{T_1})/\sigma_{T_1} = -2.976$. Using the normal approximation gives P-value $= 0.0029$. We therefore reject H_o with significance level $\alpha = 0.05$.

The following `R` code can be used to perform the test. Use the `correct=F` option.

```
> x = c(31.7,20.9,23.2,30.2,34.4,31.0,26.9)
> y = c(35.0,40.8,39.8,30.2,40.9,32.7,38.9,35.3,35.2,38.9)
> wilcox.test(x,y,correct=F)

Wilcoxon rank sum test

data:  x and y
W = 4.5, p-value = 0.002881
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x, y, correct = F) :
cannot compute exact p-value with ties
>
```

We obtaim the same statistic $W = 4.5$ and p-value 0.0029.

When there are ties in the ranks, the function `wilcox.test()` uses the normal approximation (see `help(wilcox.test)`).

Problem 2.84 We are given two independent samples of sample sizes $n_1 = 6$, $n_2 = 7$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use significance level $\alpha = 0.01$. Verify your answer with the `R wilcox.test` function (do not use continuity correction).

	1	2	3	4	5	6	7	$\tilde{X}_i$
Sample 1	23.3	21.4	21.3	21.9	20.5	21.5		21.4
Sample 2	21.2	21.0	19.9	21.0	21.4	22.4	21.4	21.2

	1	2	3	4	5	6	7	Total
Sample 1	23.3	21.4	21.3	21.9	20.5	21.5		
Sample 2	21.2	21.0	19.9	21.0	21.4	22.4	21.4	
Ranks 1	13.0	8.0	6.0	11.0	2.0	10.0		50.0
Ranks 2	5.0	3.5	1.0	3.5	8.0	12.0	8.0	41.0

SOLUTION: The pooled ranks are given in the following table:

The sample medians are $\tilde{X}_1 = 21.45$ and $\tilde{X}_2 = 21.2$. The rank sums for samples 1 and 2 are, respectively,

$$T_1 = 50 \text{ and } T_2 = 41.$$

To use tables, for a two-sided test use statistic

$$T_{obs} = \min(T_1, n_1(n_1 + n_2 + 1) - T_1) = \min(50, 6 \times 14 - 50) = \min(50, 34) = 34.$$

From tables, the P-value probability, for $n_1 = 6$, $n_2 = 7$, is

$$\alpha_{obs} = 2P(T_{obs} \leq 34) = 2 \times 0.1474 = 0.295.$$

Using the normal approximation, we have

$$\mu_{T_1} = n_1(n_1 + \mu_2 + 1)/2 = 42 \text{ and } \sigma_{T_1} = \sqrt{n_1 n_2(n_1 + n_2 + 1)/12} = 7.$$

This gives Z-score

$$Z = (T_1 - \mu_{T_1})/\sigma_{T_1} = 1.143.$$

giving

$$\alpha_{obs} = 2P(Z < -1.143) = 2 \times 0.1265 = 0.253.$$

Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.01$.

The `wilcox.test` function should be used in the following way, giving the same conclusion. Note that the adjusted rank sum is $W = T_1 - n_1(n_1 + 1)/2 = 50 - 21 = 29$.

```
> x = c(23.3,21.4,21.3,21.9,20.5,21.5)
> y = c(21.2,21.0,19.9,21.0,21.4,22.4,21.4)
> wilcox.test(x,y,correct=F)

Wilcoxon rank sum test

data:  x and y
W = 29, p-value = 0.2498
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x, y, correct = F) :
cannot compute exact p-value with ties
>
```

Problem 2.85 Q8: We are given two paired samples of sample size $n = 9$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	1136.1	1125.7	10.4	
2	1264.9	1265.1	-0.2	
3	1266.4	1308.5	-42.1	
4	1160.3	1129.4	30.9	
5	1167.3	1175.4	-8.1	
6	1246.5	1184.0	62.5	
7	1125.1	1078.1	47.0	
8	1238.7	1287.5	-48.8	
9	1170.3	1235.2	-64.9	

SOLUTION: The signs and ranks needed for the signed rank test are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	1136.1	1125.7	10.4	3.0	+
2	1264.9	1265.1	-0.2	1.0	-
3	1266.4	1308.5	-42.1	5.0	-
4	1160.3	1129.4	30.9	4.0	+
5	1167.3	1175.4	-8.1	2.0	-
6	1246.5	1184.0	62.5	8.0	+
7	1125.1	1078.1	47.0	6.0	+
8	1238.7	1287.5	-48.8	7.0	-
9	1170.3	1235.2	-64.9	9.0	-

The sample median of the differences is $\tilde{D}$ = -0.2. After excluding (zero) ties there are $n' = 9$ pairs remaining. The negative and positive rank sums are, respectively,

$$T_- = 1 + 2 + 5 + 7 + 9 = 24, \text{ and } T_+ = 3 + 4 + 6 + 8 = 21.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = n(n+1)/4 = 9 \times 10/4 = 22.5,$$

and

$$\sigma_T = \sqrt{n(n+1)(2n+1)/24} = \sqrt{9 \times 10 \times 19/24} = 8.441.$$

Since $\min(T_-, T_+) = T_+$ we have Z-score

$$Z = \frac{T_+ - \mu_T}{\sigma_T} = -0.178.$$

Reject H_o if

$$|Z| \geq z_{\alpha/2} = 1.96.$$

Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.86 We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 6$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a lower tailed rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 < 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	1	2	3	4	5	6	$\tilde{X}_i$
Sample 1	5.6	6.3	5.5	6.6	4.6		5.6
Sample 2	7.5	7.0	6.5	7.2	6.3	6.5	6.8

SOLUTION: The pooled ranks are given in the following table:

	1	2	3	4	5	6	$\tilde{X}_i$
Sample 1	5.6	6.3	5.5	6.6	4.6		5.6
Sample 2	7.5	7.0	6.5	7.2	6.3	6.5	6.8
Ranks 1	3.0	4.5	2.0	8.0	1.0		18.5
Ranks 2	11.0	9.0	6.5	10.0	4.5	6.5	47.5

The sample medians are $\tilde{X}_1 = 5.6$ and $\tilde{X}_2 = 6.75$. The rank sums for samples 1 and 2 are, respectively,

$$\begin{aligned} T_1 &= 3.0 + 4.5 + 2.0 + 8.0 + 1.0 = 18.5 \\ T_2 &= 11.0 + 9.0 + 6.5 + 10.0 + 4.5 + 6.5 = 47.5 \end{aligned}$$

The mean and standard deviation of T_1 are

$$\begin{aligned} \mu_1 &= n_1(n_1 + n_2 + 1)/2 = 5 \times 12/2 = 30, \\ \sigma_W &= \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12} = \sqrt{5 \times 6 \times 12/12} = \sqrt{30} = 5.477. \end{aligned}$$

This gives Z-score

$$Z = \frac{T_1 - \mu_1}{\sigma_W} = \frac{18.5 - 30}{5.477} = -2.1.$$

Reject H_o if

$$Z \leq z_\alpha = -1.645.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.87 We are given two paired samples from normally distributed populations with respective means μ_1, μ_2 $(n = 6)$. The data represents temperature readings in degrees Fahrenheit taken at an elevation of 1000 feet and at sea level at 6 geographical locations. The data is summarized in the table below.

Location	1000 feet	Sea level
1	51.7	54.7
2	60.4	63.3
3	60.8	62.0
4	61.3	61.5
5	65.6	76.1
6	60.6	66.2

(a) Perform a two-sided hypothesis test for null and alternative hypotheses $H_o: \mu_1 - \mu_2 = 0$ and $H_a: \mu_1 - \mu_2 \neq 0$. Your test should be based on an appropriate T-statistic. Use significance level $\alpha = 0.05$.
(b) Use a sign test to test null and alternative hypotheses $H_o: \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ and $H_a: \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$, where $\tilde{\mu}_1$, $\tilde{\mu}_2$ are the respective medians. Use significance level $\alpha = 0.05$.

SOLUTION: The differences are given in the following table:

Location	1000 feet	Sea level	D_i
1	51.7	54.7	-3.0
2	60.4	63.3	-2.9
3	60.8	62.0	-1.2
4	61.3	61.5	-0.2
5	65.6	76.1	-10.5
6	60.6	66.2	-5.6

(a) We can calculate directly $\bar{D} = \bar{X}_1 - \bar{X}_2 = $ -3.9, $S_D = 3.721$. This gives T-statistic

$$T = \frac{\bar{D}}{S_D\sqrt{n}} = \frac{-3.9}{3.721/\sqrt{6}} = -2.567.$$

Reject H_o if $|T| > t_{n-1;\alpha/2}$ where

$$t_{n-1;\alpha/2} = t_{5;0.025} = 2.571.$$

Therefore, do not reject the null hypothesis at a significance level $\alpha = 0.05$.
(b) Let X equal the number of positive differences. We observe $X = 0$. Under H_o we have $X \sim bin(6, 1/2)$. The P-value of the sign test is

$$P = 2P(X = 0) = 2 \times 1/2^6 = 0.03125 < \alpha = 0.05,$$

so reject H_o. Alternatively, X equals the number of negative differences, under H_o we have $X \sim bin(6, 1/2)$, $P = 2P(X = 6) = 2 \times 1/2^6 = 0.03125 < \alpha$.

Problem 2.88 A cholesterol reduction treatment is assessed on 9 subjects. Blood samples are collected before and after treatment. Observed triglyceride levels (a measure of cholesterol) are given in the following table. Use a signed rank test to determine if there is any evidence of a difference in triglyceride levels before and after the treatment. Use a normal distribution approximation to assess significance. Do not use a continuity correction. Use significance level $\alpha = 0.05$.

Subject	Before Treatment (X)	After Treatment (Y)
1	410.9	308.3
2	371.2	309.6
3	364.7	309.4
4	423.6	358.2
5	438.5	394.6
6	338.4	376.0
7	412.6	340.5
8	382.4	406.6
9	443.9	474.0

Subject	Before (X)	After (Y)	$D = X - Y$	Rank $\|D\|$	Sign
1	410.9	308.3	-102.6	9.0	-
2	371.2	309.6	-61.6	6.0	-
3	364.7	309.4	-55.3	5.0	-
4	423.6	358.2	-65.4	7.0	-
5	438.5	394.6	-43.9	4.0	-
6	338.4	376.0	+37.6	3.0	+
7	412.6	340.5	-72.1	8.0	-
8	382.4	406.6	+24.2	1.0	+
9	443.9	474.0	+30.1	2.0	+

SOLUTION: The differences, ranks, and signs are given in the following table:

There are no ties, so $n = 9$. The positive and negative rank sums are

$$T_+ = 1 + 2 + 3 = 6, \;\; T_- = n(n+1)/2 - T_- = 45 - 6 = 39.$$

Then set

$$T_{obs} = \min\{T_-, T_+\} = 6.$$

To use a normal approximation for the distribution of T_{obs} under H_o set mean and standard deviation:

$$\begin{aligned} \mu_T &= n(n+1)/4 = 9 \times 10/4 = 22.5, \\ \sigma_T &= \sqrt{n(n+1)(2n+1)/24} = \sqrt{9 \times 10 \times 19/24} = 8.441. \end{aligned}$$

The appropriate Z-score is therefore:

$$Z = \frac{T_{obs} - \mu_T}{\sigma_T} = \frac{6 - 22.5}{8.441} = -1.955.$$

Reject H_o if $|Z| > z_{\alpha/2} = z_{0.025} = 1.96$. Therefore, do not reject H_o.

Problem 2.89 We are given two independent samples of sample sizes $n_1 = 4$, $n_2 = 12$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use a normal distribution approximation to assess significance. Do not use a continuity correction. Use significance level $\alpha = 0.05$.

	1	2	3	4	5	6	7	8	9	10	11	12	$\tilde{X}_i$
Sample 1	97.6	107.8	87.4	95.4									96.5
Sample 2	86.5	80.0	70.3	79.5	84.7	84.9	75.7	80.1	69.2	79.3	79.1	70.2	79.4

	1	2	3	4	5	6	7	8	9	10	11	12	$\tilde{X}_i$
Sample 1	97.6	107.8	87.4	95.4									96.5
Sample 2	86.5	80.0	70.3	79.5	84.7	84.9	75.7	80.1	69.2	79.3	79.1	70.2	79.4
Ranks 1	15.0	16.0	13.0	14.0									58.0
Ranks 2	12.0	8.0	3.0	7.0	10.0	11.0	4.0	9.0	1.0	6.0	5.0	2.0	78.0

SOLUTION: There are no ties, so $n_1 = 4$, $n_2 = 12$. The pooled ranks are
Since $n_1 < n_2$ we rely only on the rank sum for the first sample:

$$T_1 = 13 + 14 + 15 + 16 = 58.$$

To use a normal approximation for the distribution of T_1 under H_o set mean and standard deviation:

$$\begin{aligned} \mu_{T_1} &= \frac{n_1(n_1+n_2+1)}{2} = \frac{4(4+12+1)}{2} = 34, \\ \sigma_{T_1} &= \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} = \sqrt{\frac{4\times 12\times(4+12+1)}{12}} \approx 8.246. \end{aligned}$$

This gives Z-score

$$Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{58-34}{8.246} = 2.91.$$

Reject H_o if $|Z| > z_{\alpha/2} = z_{0.025} = 1.96$. Therefore, reject H_o.

Problem 2.90 We are given two paired samples of sample size $n = 9$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$, making use of the normal approximation.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	5.4	4.9	0.5		
2	4.7	5.1	-0.4		
3	7.2	7.9	-0.7		
4	5.4	9.0	-3.6		
5	6.6	10.4	-3.8		
6	0.5	-3.0	3.5		
7	6.5	8.9	-2.4		
8	1.9	3.3	-1.4		
9	2.7	-0.3	3.0		

SOLUTION: The signs and ranks required for the signed rank test are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\|D\|$	Sign
1	5.4	4.9	0.5	2.0	+
2	4.7	5.1	-0.4	1.0	-
3	7.2	7.9	-0.7	3.0	-
4	5.4	9.0	-3.6	8.0	-
5	6.6	10.4	-3.8	9.0	-
6	0.5	-3.0	3.5	7.0	+
7	6.5	8.9	-2.4	5.0	-
8	1.9	3.3	-1.4	4.0	-
9	2.7	-0.3	3.0	6.0	+

There are no ties, so $n' = 9$. The negative and positive rank sums are, respectively,

$$T_+ = 2 + 7 + 6 = 15 \text{ and } T_- = (10 \times 9)/2 - 15 = 45 - 15 = 30.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = 22.5 \text{ and } \sigma_T = 8.441.$$

This gives

$$Z = \frac{T_+ - \mu_T}{\sigma_T} = \frac{15 - 22.5}{8.441} = -0.889.$$

Reject H_o if

$$|Z| \geq z_{\alpha/2} = 1.96.$$

Therefore, do no reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 0.374).

Problem 2.91 We are given two paired samples of sample size $n = 7$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a lower tailed signed rank test using null hypothesis $H_o : \tilde{\mu}_D = 0$ against alternative $H_a : \tilde{\mu}_D < 0$. Use significance level $\alpha = 0.05$. Enter the corresponding ranks and signs where indicated in the table. Use a normal approximation to construct a test statistic.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\|D\|$	Sign
1	17.8	21.1	-3.3		
2	23.0	22.8	0.2		
3	26.0	30.1	-4.1		
4	19.4	25.2	-5.8		
5	26.8	31.9	-5.1		
6	31.6	35.3	-3.7		
7	20.1	18.4	1.7		

SOLUTION: The signs and ranks required for the signed rank test are given in the following table:

The sample median of the differences is $\tilde{D}$ = -3.7. After excluding ties there are $n' = 7$ pairs remaining. The negative and positive rank sums are, respectively, $T_- = 25$ and $T_+ = 3$. The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = n(n+1)/4 = 7 \times 8/4 = 14$$

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	17.8	21.1	-3.3	3.0	-
2	23.0	22.8	0.2	1.0	+
3	26.0	30.1	-4.1	5.0	-
4	19.4	25.2	-5.8	7.0	-
5	26.8	31.9	-5.1	6.0	-
6	31.6	35.3	-3.7	4.0	-
7	20.1	18.4	1.7	2.0	+

and

$$\sigma_T = \sqrt{n(n+1)(2n+1)/24} = \sqrt{7 \times 8 \times 15/24} = 5.916.$$

We reject for small values of T_+, which gives Z-score

$$Z = \frac{T_+ - \mu_T}{\sigma_T} = \frac{3 - 14}{5.916} = -1.86.$$

Since $Z < -z_\alpha = -1.645$ we reject H_o for $\alpha = 0.05$.

Problem 2.92 We are given two paired samples of sample size $n = 6$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a lower tailed sign test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D < 0$. Use significance level $\alpha = 0.05$.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	10.5	15.7	-5.2	
2	8.2	11.5	-3.3	
3	12.0	12.4	-0.4	
4	14.5	11.6	2.9	
5	8.3	12.0	-3.7	
6	4.0	4.0	0.0	

SOLUTION: The required signs are given in the followign table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	10.5	15.7	-5.2	-
2	8.2	11.5	-3.3	-
3	12.0	12.4	-0.4	-
4	14.5	11.6	2.9	+
5	8.3	12.0	-3.7	-
6	4.0	4.0	0.0	0

After excluding ties there are $X = 1$ positive differences among $n' = 5$ pairs. Using the **binomial distribution** directly:

$$\alpha_{obs} = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.5^5 + 5 \times 0.5^5 = 0.1875.$$

So, do not reject H_o with $\alpha = 0.05$ significance level.

Using a **normal approximation**, under the null hypothesis, $X \sim bin(5, 1/2)$, with mean and variance $\mu = np = 5 \times 1/2 = 2.5$ and $\sigma^2 = np(1-p) = 5 \times 1/2 \times 1/2 = 1.25$. Then we have z-score

$$\begin{aligned} Z &= \frac{1-\mu}{\sigma} = \frac{1-2.5}{\sqrt{1.25}} = -1.34 \text{ without continuity correction,} \\ Z &= \frac{1.5-\mu}{\sigma} = \frac{1.5-2.5}{\sqrt{1.25}} = -0.89 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \leq -z_\alpha = -1.645$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Problem 2.93 We are given two paired samples of sample size $n = 8$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)
1	97.9	94.7	3.2
2	104.7	107.0	-2.3
3	91.9	90.3	1.6
4	96.7	94.1	2.6
5	106.4	103.3	3.1
6	103.4	100.1	3.3
7	94.2	89.9	4.3
8	97.7	102.1	-4.4

SOLUTION: The signed ranks are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D \rvert$	Sign
1	97.9	94.7	3.2	5.0	+
2	104.7	107.0	-2.3	2.0	-
3	91.9	90.3	1.6	1.0	+
4	96.7	94.1	2.6	3.0	+
5	106.4	103.3	3.1	4.0	+
6	103.4	100.1	3.3	6.0	+
7	94.2	89.9	4.3	7.0	+
8	97.7	102.1	-4.4	8.0	-

There are no ties, so $n = 8$ pairs. The negative and positive rank sums are, respectively,

$$T_- = 10 \text{ and } T_+ = 26.$$

Then

$$T_{obs} = \min(T_-, T_+) = \min(10, 26) = 10.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = \frac{n(n+1)}{4} = 18 \text{ and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 7.141.$$

This gives z-score

$$\begin{aligned} Z &= \frac{T_{obs} - \mu_T}{\sigma_T} = \frac{10 - 18}{7.141} \approx -1.12 \text{ without continuity correction,} \\ Z &= \frac{T_{obs} + 0.5 - \mu_T}{\sigma_T} = \frac{10.5 - 18}{7.141} \approx -1.05 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \leq -z_{\alpha/2} = -1.96$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Problem 2.94 We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 10$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use significance level $\alpha = 0.05$.

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	30.0	23.1	25.8	25.9	23.5						25.8
Sample 2	22.3	23.2	20.3	21.7	25.6	22.9	22.9	20.7	22.3	19.0	22.3

SOLUTION: The signed ranks are given in the following table:

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	30.0	23.1	25.8	25.9	23.5						25.8
Sample 2	22.3	23.2	20.3	21.7	25.6	22.9	22.9	20.7	22.3	19.0	22.3
Ranks 1	15.0	9.0	13.0	14.0	11.0						0.0
Ranks 2	5.5	10.0	2.0	4.0	12.0	7.5	7.5	3.0	5.5	1.0	0.0

The rank sums for samples 1 and 2 are, respectively,

$$T_1 = 62 \text{ and } T_2 = 58.$$

We only need T_1. The mean and standard deviation of T_1 is

$$\mu_{T_1} = \frac{n_1(n_1 + n_2 + 1)}{2} = 40 \text{ and } \sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 8.165.$$

This gives z-score (noting that $T_1 > \mu_{T_1}$),

$$\begin{aligned} Z &= \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{62 - 40}{8.165} \approx 2.694 \text{ without continuity correction,} \\ Z &= \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{62 - 0.5 - 40}{8.165} \approx 2.633 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \geq z_{\alpha/2} = 1.96$. So, reject H_o with $\alpha = 0.05$ significance level.

Problem 2.95 We are given two independent samples of sample sizes $n_1 = 7$, $n_2 = 10$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using null hypothesis $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against alternative $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use significance level $\alpha = 0.05$. Enter the corresponding ranks where indicated in the table. Use a normal approximation to construct a test statistic.

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	208.7	211.6	197.6	209.1	198.1	204.1	192.8	-	-	-	204.1
Sample 2	167.6	162.7	164.4	164.3	172.2	165.9	169.8	157.6	162.9	164.7	164.6
Ranks 1								-	-	-	-
Ranks 2											-

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	208.7	211.6	197.6	209.1	198.1	204.1	192.8				204.1
Sample 2	167.6	162.7	164.4	164.3	172.2	165.9	169.8	157.6	162.9	164.7	164.6
Ranks 1	15	17	12	16	13	14	11				
Ranks 2	8	2	5	4	10	7	9	1	3	6	

SOLUTION: The pooled ranks are given in the following table:

The sample medians are $\tilde{X}_1 = 204.1$ and $\tilde{X}_2 = 164.55$. The rank sums for samples 1 and 2 are, respectively, $T_1 = 98$ and $T_2 = 55$. The mean and standard deviation of T_1 are

$$\mu_1 = n_1(n_1 + n_2 + 1)/2 = 7 \times (7 + 10 + 1)/2 = 63$$

and

$$\sigma_W^2 = n_1 n_2(n_1 + n_2 + 1)/12 = 105, \;\; \sigma_W = \sqrt{105} = 10.247.$$

This gives Z-score

$$Z = \frac{T_1 - \mu_1}{\sigma_W} = \frac{98 - 63}{10.247} = 3.416.$$

Since $Z > z_{\alpha/2} = 1.96$ we reject H_o for $\alpha = 0.05$.

Problem 2.96 We are given two paired samples of sample size $n = 9$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$, making use of the normal approximation.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D \rvert$	Sign
1	5.4	4.9	0.5		
2	4.7	5.1	-0.4		
3	7.2	7.9	-0.7		
4	5.4	9.0	-3.6		
5	6.6	10.4	-3.8		
6	0.5	-3.0	3.5		
7	6.5	8.9	-2.4		
8	1.9	3.3	-1.4		
9	2.7	-0.3	3.0		

SOLUTION: The signed ranks are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\|D\|$	Sign
1	5.4	4.9	0.5	2.0	+
2	4.7	5.1	-0.4	1.0	-
3	7.2	7.9	-0.7	3.0	-
4	5.4	9.0	-3.6	8.0	-
5	6.6	10.4	-3.8	9.0	-
6	0.5	-3.0	3.5	7.0	+
7	6.5	8.9	-2.4	5.0	-
8	1.9	3.3	-1.4	4.0	-
9	2.7	-0.3	3.0	6.0	+

There are no ties, so $n' = 9$. The negative and positive rank sums are, respectively,

$$T_+ = 2 + 7 + 6 = 15 \text{ and } T_- = (10 \times 9)/2 - 15 = 45 - 15 = 30.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = 22.5 \text{ and } \sigma_T = 8.441.$$

This gives

$$Z = \frac{T_+ - \mu_T}{\sigma_T} = \frac{15 - 22.5}{8.441} = -0.889.$$

Reject H_o if

$$|Z| \geq z_{\alpha/2} = 1.96.$$

Therefore, do no reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 0.374).

Problem 2.97 We are given two paired samples of sample size $n = 7$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)
1	23.9	12.3	11.6
2	28.1	31.5	-3.4
3	26.7	24.4	2.3
4	32.4	29.7	2.7
5	20.7	21.9	-1.2
6	23.0	26.1	-3.1
7	27.8	26.5	1.3

SOLUTION: The signed ranks are given in the following table:

There are no ties, so we have $n = 7$ pairs. The negative and positive rank sums are, respectively,

$$T_- = 12 \text{ and } T_+ = 16.$$

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	23.9	12.3	11.6	7.0	+
2	28.1	31.5	-3.4	6.0	-
3	26.7	24.4	2.3	3.0	+
4	32.4	29.7	2.7	4.0	+
5	20.7	21.9	-1.2	1.0	-
6	23.0	26.1	-3.1	5.0	-
7	27.8	26.5	1.3	2.0	+

Then

$$T_{obs} = \min(T_-, T_+) = \min(12, 16) = 12.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = \frac{n(n+1)}{4} = 14 \text{ and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 5.916.$$

This gives z-score

$$Z = \frac{T_{obs} - \mu_T}{\sigma_T} = \frac{12 - 14}{5.916} \approx -0.338$$

We reject H_o if $Z \leq -z_{\alpha/2} = -1.96$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Problem 2.98 We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 11$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu}_1 - \tilde{\mu}_2 = 0$ against $H_a : \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	1	2	3	4	5	6	7	8	9	10	11
Sample 1	86.2	79.7	79.8	92.8	75.9						
Sample 2	125.9	122.9	120.0	120.3	97.5	122.0	127.3	127.0	101.8	90.1	123.5

SOLUTION: The pooled ranks are given in the following table:

	1	2	3	4	5	6	7	8	9	10	11
Sample 1	86.2	79.7	79.8	92.8	75.9						
Sample 2	125.9	122.9	120.0	120.3	97.5	122.0	127.3	127.0	101.8	90.1	123.5
Ranks 1	4.0	2.0	3.0	6.0	1.0						
Ranks 2	14.0	12.0	9.0	10.0	7.0	11.0	16.0	15.0	8.0	5.0	13.0

The sample medians are $\tilde{X}_1 = 79.8$ and $\tilde{X}_2 = 122$. The rank sum for samples 1 and 2 are, respectively, $T_1 = 16$ and $T_2 = 120$. The mean and standard deviation of T_1 are

$$\mu_1 = n_1(n_1 + n_2 + 1)/2 = 5 \times (5 + 11 + 1)/2 = 42.5$$

and

$$\sigma_W^2 = n_1 n_2 (n_1 + n_2 + 1)/12 = 77.92, \quad \sigma_W = \sqrt{105} = 8.827.$$

This gives Z-score

$$Z = \frac{T_1 - \mu_1}{\sigma_W} = \frac{16 - 42.5}{8.827} = -3.00.$$

Since $Z < -z_{\alpha/2} = 1.96$ we reject H_o for $\alpha = 0.05$.

Problem 2.99 A conventional treatment for sleep apnea is reported to improve sleep quality within 1 week for 60% of subjects. A new experimental treatment is studied using 12 subjects. A sleep quality index is measured at the start of treatment and after 1 week of treatment for each subject (higher values signify better sleep quality). The results are given in the following table. Is there evidence that more than 60% of subjects experience improved sleep quality within 1 week? Report an exact P-value, and use significance level $\alpha = 0.05$.

Subject	Start of treatment (X)	1 week of treatment (Y)	Difference ($D = Y - X$)
1	29.1	32.4	3.3
2	30.0	39.2	9.2
3	23.3	34.5	11.2
4	26.2	30.8	4.6
5	25.9	39.5	13.6
6	29.7	24.0	-5.7
7	22.8	25.8	3.0
8	23.6	29.3	5.7
9	27.6	36.0	8.4
10	25.1	30.6	5.5
11	31.4	42.4	11.0
12	27.2	33.0	5.8

SOLUTION: Let p be the proportion experiencing improved sleep quality. Then $X = 11$ be the observed number of subjects experiencing improved sleep quality. Then $X \sim bin(12, p)$. The appropriate hypotheses are $H_o : p \leq 0.6$ against $H_a : p > 0.6$. The p-value is

$$P(X \geq 11) = \binom{12}{11} \times 0.6^{11} \times 0.4^1 + \binom{12}{12} \times 0.6^{12} = 12 \times 0.00145 + 0.00218 = 0.0196$$

where $X \sim bin(12, 0.6)$. Since $P < \alpha$ we reject H_o.

Problem 2.100 We are given two paired samples of sample size $n = 9$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$.

	Sample 1 (X)	Sample 2 (Y)	$D = X - Y$
1	20.1	16.9	3.2
2	16.7	20.3	-3.6
3	18.3	16.5	1.8
4	14.4	14.9	-0.5
5	16.6	16.0	0.6
6	15.2	14.1	1.1
7	15.7	15.0	0.7
8	13.1	12.9	0.2
9	11.7	11.7	0.0

(a) Perform a lower tailed signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D > 0$ with significance level $\alpha = 0.05$. Use both a normal approximation and the exact method.

(b) Perform a sign test for the same hypotheses, and report a P-value. Do you reach a different conclusion than that of Part (a)?

SOLUTION:

(a) In the following table the absolute values of the differences are

	Sample 1 (X)	Sample 2 (Y)	$D = X - Y$	$\lvert D\rvert$	Rank $\lvert D\rvert$	Sign
1	20.1	16.9	3.2	3.2	7.0	+
2	16.7	20.3	-3.6	3.6	8.0	-
3	18.3	16.5	1.8	1.8	6.0	+
4	14.4	14.9	-0.5	0.5	2.0	-
5	16.6	16.0	0.6	0.6	3.0	+
6	15.2	14.1	1.1	1.1	5.0	+
7	15.7	15.0	0.7	0.7	4.0	+
8	13.1	12.9	0.2	0.2	1.0	+
9	11.7	11.7	0.0	0.0	0	0

The sample median of the differences is $\tilde{D} = 0.6$. After excluding ties there are $n' = 8$ pairs remaining. The negative and positive rank sums are, respectively, $T_- = 10$ and $T_+ = 26$. We reject $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D > 0$ for small enough T_-.

Normal Approximation: The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = n(n+1)/4 = 8 \times 9/4 = 18$$

and

$$\sigma_T = \sqrt{n(n+1)(2n+1)/24} = \sqrt{8 \times 9 \times 17/24} = 7.141.$$

This gives Z-score

$$Z = \frac{T_- - \mu_T}{\sigma_T} = \frac{10 - 18}{7.141} \approx -1.12.$$

Then reject H_o if $Z < -z_\alpha = -1.645$. So we fail to reject H_o at significance level $\alpha = 0.05$. The P-value is $P = P(Z < -1.12) \approx 0.131$, using `R` function

```
> pnorm(-1.12)
[1] 0.1313569
```

Exact Method: To use the exact method we evaluate the tail probability $P = P(T \leq T_-)$ for $T_- = 10$, $n = 8$. Using Table A.22 we have $P = P(T \leq 10) = 0.1563$. We can also use the R function

```
> psignrank(10,n=8)
[1] 0.15625
```

Since $P > \alpha = 0.05$, we fail to reject H_o.

(b) Let X be the number of negative differences. We observe $X = 2$. We reject H_o for small values of X. Under the null hypothesis $X \sim bin(n', 1/2)$, with P-value $P = P(X \leq 2) \approx 0.145$. This value can be obtained by the R function

```
> pbinom(2,prob=1/2,size=8)
[1] 0.1445313
```

Since $P > \alpha = 0.05$, we do not reject the null hypothesis, which is the same conclusion reached in Part (a).

Problem 2.101 We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 12$. The data is summarized in the table below. Suppose $\tilde{\mu}_i$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu_1} - \tilde{\mu_2} = 0$ against $H_a : \tilde{\mu_1} - \tilde{\mu_2} \neq 0$. Use significance level $\alpha = 0.01$.

	1	2	3	4	5	6	7	8	9	10	11	12	$\tilde{X}_i$
Sample 1	24.5	19.3	32.5	28.5	23.9								24.5
Sample 2	32.5	31.6	40.4	35.8	34.9	40.7	39.5	41.0	35.1	38.8	39.1	41.0	38.95

SOLUTION: The sample medians are $\tilde{X}_1 = 24.5$ and $\tilde{X}_2 = 38.95$. The pooled ranks are given in the following table:

	1	2	3	4	5	6	7	8	9	10	11	12	$\tilde{X}_i/T_i$
Sample 1	24.5	19.3	32.5	28.5	23.9								24.5
Sample 2	32.5	31.6	40.4	35.8	34.9	40.7	39.5	41.0	35.1	38.8	39.1	41.0	38.95
Ranks 1	3.0	1.0	6.5	4.0	2.0								16.5
Ranks 2	6.5	5.0	14.0	10.0	8.0	15.0	13.0	16.5	9.0	11.0	12.0	16.5	136.5

The rank sums for samples 1 and 2 are, respectively, $T_1 = 16.5$ and $T_2 = 136.5$.
The mean and standard deviation of T_1 are given by $\mu_{T_1} = 45$ and $\sigma_{T_1} = 9.487$. This gives Z-score

$$\mu_{T_1} = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{5(5 + 12 + 1)}{2} = 45$$

and

$$\sigma_{T_1} = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{5 \times 12 \times (5 + 12 + 1)}{12}} \approx 9.487.$$

This gives Z-score

$$Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{16.5 - 45}{9.487} = -3.004111.$$

To calculate a P-value use `R` function

```
> pnorm(-3.004111)
[1] 0.001331791
```

giving $P(Z < -3.004111) \approx 0.0013$. Since the test is two-sided, we have P-value $P = 2 \times 0.0013 = 0.0026$, so we reject H_o at significance level $\alpha = 0.01$.

2.8 Goodness of Fit Tests and Contingency Tables

Problem 2.102 A market survey asked respondents to give their preference of one of four brands of a certain consumer item. The observed counts for $k = 4$ categories based on a random sample of size $n = 224$ are given in the following table, and represent the number of respondents stating their preference for each brand. Use a suitable χ^2 statistic to test the null hypothesis H_o that the preference proportions are the same for each brand, against the alternative hypothesis that they differ. Use significance level $\alpha = 0.05$, and employ Yate's correction procedure.

BRAND	A	B	C	D	Totals
Observed counts O_i	47	53	25	99	224

SOLUTION: The null hypothesis is:

$$H_o : p_i 1/4, \;\; i = 1, 2, 3, 4.$$

The expected counts are given by $E_i = np_i$. For example, $E_1 = 224 \times 1/4 = 56.0$. Then, using Yate's correction the test statistic is

$$X^2 = \sum_i X_i^2,$$

where

$$X_i^2 = \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

For example,

$$X_1^2 = \frac{(|47 - 56.0| - 0.5)^2}{56.0} \approx 1.29.$$

The remaining expected counts and terms are then

	1	2	3	4	Totals
Observed counts O_i	47	53	25	99	224
Expected counts E_i	56.0	56.0	56.0	56.0	224
$(\|O_i - E_i\| - 0.5)^2/E_i$	1.29	0.11	16.61	32.25	50.27

Reject H_o if X^2 is greater than or equal to

$$\chi^2_{k-1;\alpha} = \chi^2_{3;0.05} = 7.815.$$

With Yate's correction: $X^2 = 50.268 > 7.815$. Therefore, reject the null hypothesis at a significance level $\alpha = 0.05$.

It also suffices to note that the 3rd and 4th terms are $X^2_3 = 16.61$, $X^2_4 = 32.25$, each of which exceeds the critical value $\chi^2_{3;0.05}$. So H_o can be rejected on the basis of either term alone.

Problem 2.103 A probability distribution on the positive integers $1, 2, 3, \ldots, N$ conforms to Zipf's Law if the PMF satisfies $p_i \propto 1/i$. It is observed in many types of data. For example, if the frequencies of word occurrences are given in decreasing order, then Zipf's Law is often observed, where p_i is the frequency of the ith most common word. Suppose to test this idea, the frequencies of the four most commonly used words in a certain text are compiled in the following table, and we wish to construct a hypothesis test with the null hypothesis that the observed frequencies conform to Zipf's Law.

Frequency Rank	1	2	3	4	Totals
Observed counts O_i	162	77	63	47	349
Observed frequencies $\hat{p}_i$	0.46	0.22	0.18	0.13	1.00

(a) Formulate precisely the null and alternative hypotheses as a goodness of fit test.
(b) Carry out a χ^2 test and state your conclusion. Use significance level $\alpha = 0.05$. Perform the test without and with Yates' correction. Construct a rejection region, and also give a P-value.
(c) Use the `R` function `chisq.test` to verify your conclusion (without Yate's correction only).
(d) If in this application we rank the words using the observed frequencies (as opposed to using some prior hypothesis), do we violate any assumption used in the χ^2 test?

SOLUTION:

(a) The first four probabilities in Zipf's distribution are proportional to $p_i \propto 1/i$, $i = 1, \ldots, 4$. We therefore normalize by

$$p_i = \frac{1/i}{1/1 + 1/2 + 1/3 + 1/4} = \frac{12}{i \times 25},$$

giving the hypothetical probabilities in Table 2.3. The hypotheses are then

$$H_o : p_i = \frac{12}{i \times 25}, \quad i = 1, 2, 3, 4 \quad \text{against } H_a : p_i \neq \frac{12}{i \times 25}, \quad \text{for some } i.$$

Table 2.3: Calculations for Problem 2.103.

Frequency Rank	1	2	3	4	Totals
Observed counts O_i	162	77	63	47	349
Expected counts E_i	167.52	83.76	55.84	41.88	349.00
Hypothetical p_i	0.48	0.24	0.16	0.12	1.00
$(O_i - E_i)^2/E_i$	0.18	0.55	0.92	0.63	2.27
$(\lvert O_i - E_i\rvert - 0.5)^2/E_i$	0.15	0.47	0.79	0.51	1.92

(b) We have $k = 4$ categories and therefore $k - 1 = 3$ degrees of freedom. Thus, the critical value for a χ^2 test with $\alpha = 0.05$ is $\chi^2_{k-1,\alpha} = 7.815$.
Without Yates' correction, from Table 2.3

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i = 2.27.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value $= P(\chi^2_3 > 2.27) = 0.518$.
With Yate's correction, from Table 2.3

$$X^2 = \sum_{i=1}^{4} (|O_i - E_i| - 0.5)^2 / E_i = 1.92.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value $= P(\chi^2_3 > 1.92) = 0.5887$.

(c) The following code give the appropriate test, which conforms to the conclusion above (without Yate's correction):

```
> x = c(162,77,63,47)
> p0 = 1/1:4
> p0 = p0/sum(p0)
> chisq.test(x,p = p0)

Chi-squared test for given probabilities

data:  x
X-squared = 2.2715, df = 3, p-value = 0.518

>
```

(d) Yes it does. In constructing the test, the theoretical frequencies $p_1, \ldots, p_4$ are assumed to be associated with specific words, p_1 being associated with the most common word, and so on. We also assume that we are sampling from a larger population in which these frequencies hold. We therefore need to assume that the observed frequencies are in the same order as the theoretical ones, which need not be true. The question is whether or not the true ordering is correctly observed. If the probability that this occurs is close to one, than this assumption violation will not be important.

Problem 2.104 A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 187$ is given below (Table 2.4). Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The marginal population frequencies for the row i and column j categories are given by r_i and c_j, respectively.

(a) Use a χ^2 test for the null hypothesis of row and column independence

$$H_o : p_{i,j} = r_i c_j \text{ for all } i, j \text{ against } H_a : p_{i,j} \neq r_i c_j \text{ for some } i, j.$$

Use significance level $\alpha = 0.05$. Perform the test without and with Yates' correction. Construct a rejection region, and also give a P-value.

Table 2.4: Observed counts $O_{i,j}$ for Problem 2.104.

	1	2	3	Totals
1	46	44	42	132
2	8	11	36	55
Totals	54	55	78	187

(b) Use the `R` function `chisq.test` to verify your conclusion (without Yate's correction only).

SOLUTION:

(a) We have $n_r = 2$ rows and $n_c = 3$ columns and therefore $(2-1)(3-1) = 2$ degrees of freedom. Thus, the critical value for a χ^2 test with $\alpha = 0.05$ is $\chi^2_{(n_r-1)(n_c-1),\alpha} = 5.991$. Expected counts $E_{i,j} = R_i C_j / N$ are given in Table 2.5.

Table 2.5: Expected counts $E_{i,j}$ for Problem 2.104.

	1	2	3	Totals
1	38.12	38.82	55.06	132.00
2	15.88	16.18	22.94	55.00
Totals	54.00	55.00	78.00	187.00

Without Yate's correction, the terms of the X^2 statistic are given in Table 2.6, giving

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} (O_{i,j} - E_{i,j})^2 / E_{i,j} = 18.42.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{(n_r-1)(n_c-1),\alpha} = 5.991$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value $= P(\chi^2_2 > 18.42) = 0.0001$. Note that to perform the hypothesis test alone, is suffices to show that $X^2 > 5.991$, which can be done by calculating, for example, only the term $(O_{2,3} - E_{2,3})^2 / E_{2,3} = 7.43 > 5.991$. Of course, to calculate the P-value, the entire statistic must be calculated.

Table 2.6: X^2 statistic terms $(O_{i,j} - E_{i,j})^2 / E_{i,j}$ for Problem 2.104.

	1	2	3	Totals
1	1.63	0.69	3.10	5.42
2	3.91	1.66	7.43	13.00
Totals	5.54	2.35	10.53	18.42

With Yate's correction, the terms of the X^2 statistic are given in Table 2.7, giving

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} (|O_{i,j} - E_{i,j}| - 0.5)^2 / E_{i,j} = 16.52.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 5.991$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value $= P(\chi^2_2 > 16.52) = 0.000259$.

Table 2.7: X^2 statistic terms (with Yates's correction) $(|O_{i,j} - E_{i,j}| - 0.5)^2/E_{i,j}$ for Problem 2.104.

	1	2	3	Totals
1	1.43	0.56	2.86	4.86
2	3.43	1.35	6.88	11.66
Totals	4.86	1.92	9.74	16.52

(b) The following code give the appropriate test, which conforms to the conclusion above (without Yate's correction):

```
> mm = matrix(c(46,44,42,8,11,36), nrow=2, byrow=T)
> chisq.test(mm)

Pearson's Chi-squared test

data:  mm
X-squared = 18.419, df = 2, p-value = 0.0001001

>
```

Problem 2.105 A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 276$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$. Do not employ Yate's correction procedure.

	1	2	3	Totals
1	40	46	88	174
2	44	28	30	102
Totals	84	74	118	276

SOLUTION: Under the null hypothesis, the expected count for cell i, j is

$$E_{ij} = \frac{R_i C_j}{N},$$

where $N = 276$, the total sample size, and R_i, C_j are the row i and column j totals. For example,

$$E_{11} = \frac{R_1 C_1}{N} = \frac{174 \times 84}{276}. = 52.957.$$

The remaining expected counts are then:

E_{ij}	1	2	3	Totals
1	52.96	46.65	74.39	174
2	31.04	27.35	43.61	102
Totals	84	74	118	276

The χ^2 statistic is given by

$$X^2 = \sum_{ij} X_{ij}^2,$$

where

$$X_{ij}^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

For example,

$$X_{11}^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(40 - 52.957)^2}{52.957} \approx 3.17.$$

The remaining terms are then (without Yate's correction):

X_{ij}^2	1	2	3	Totals
1	3.17	0.01	2.49	5.67
2	5.41	0.02	4.25	9.67
Totals	8.58	0.02	6.74	15.34

Reject H_o if X^2 is greater than or equal to

$$\chi^2_{(n_r-1)(n_c-1);\alpha} = \chi^2_{2;0.05} = 5.991.$$

Without Yate's correction: $X^2 = 15.339 > \chi^2_{2;0.05}$, therefore, reject the null hypothesis at a significance level $\alpha = 0.05$.

Problem 2.106 Gregor Mendel hypothesized that a certain pattern of breeding would yield a generation of pea plants consisting of four types: RY (round, yellow); WY (wrinkled, yellow); RG (round, green); WG (wrinkled, green). Furthermore, these types would occur in approximately the ratios:

$$RY{:}WY{:}RG{:}WG = 9 : 3 : 3 : 1. \tag{2.1}$$

Suppose an experiment yields a generation of pea plants with observed type frequencies given in the following table.

	RY	WY	RG	WG	Total
Observed counts O_i	123	101	92	51	367

(a) Formulate precisely the null and alternative hypotheses for a goodness of fit test the conjecture that the type frequencies conform to the ratios given in Equation (2.1).

(b) Carry out a χ^2 test and state your conclusion. Use significance level $\alpha = 0.05$. Perform the test without Yates' correction.

SOLUTION:

(a) The hypothetical frequencies are obtained using normalizing constant $9 + 3 + 3 + 1 = 16$:

$$(p_1^o, p_2^o, p_3^o, p_4^o) = (9/16, 3/16, 3/16, 1/16). \tag{2.2}$$

The null and alternative hypotheses are

$$H_o : p_i = p_i^o \text{ for all } i = 1, 2, 3, 4 \text{ against } H_a : p_i \neq p_i^o \text{ for some } i = 1, 2, 3, 4,$$

where the hypothetical frequencies p_i^o are give in Equation (2.2).

(b) The expected counts are

$$E_i = np_i^o, \quad i = 1, 2, 3, 4,$$

and are given in the table below. There are $k = 4$ categories, so the χ^2 statistic has $k - 1 = 3$ degrees of freedom. The appropriate ciritcal value is $\chi^2_{3,0.05} = 7.815$. The statistic is given by

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i = 90.92552,$$

using the terms from the following table:

	RY	WY	RG	WG	Totals
Observed counts O_i	123	101	92	51	367
Expected counts E_i	206.4375	68.8125	68.8125	22.9375	367
$(O_i - E_i)^2/E_i$	33.723604	15.055915	7.813408	34.332595	90.92552

Note that the terms for $i = 1, 2, 4$ each exceed $\chi^2_{3,0.05}$ so only one of these needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.107 A random sample of categorical data is collected. The sample is of size $n = 124$. The observed counts for $k = 4$ categories are given in the table below. Use a χ^2 test for the null hypothesis

$$H_o : p = (1/15, 2/15, 4/15, 8/15).$$

Use significance level $\alpha = 0.05$. Do not use Yate's correction. Verify your answer using the `R` function `chisq.test()`

	1	2	3	4	Totals
Observed counts O_i	8	19	31	66	124
Hypothetical frequencies p_i	1/15	2/15	4/15	8/15	1.00

SOLUTION: The expected counts and terms of the χ^2-statistic are given in the following table:

	1	2	3	4	Totals
Observed counts O_i	8	19	31	66	124
Expected counts E_i	8.27	16.53	33.07	66.13	124.00
$(O_i - E_i)^2/E_i$	0.0086	0.3680	0.1292	0.00027	0.506

From the table we have $X^2 = 0.506$. Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$, $k - 1 = 3$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

The following `R` code can be used to perform the test.

```
> x = c(8,19,31,66)
> p0 = c(1/15,2/15,4/15,8/15)
> chisq.test(x,p=p0)
```

```
Chi-squared test for given probabilities

data:  x
X-squared = 0.50605, df = 3, p-value = 0.9176

>
```

We obtain the same value for $X^2 = 0.506$.

Problem 2.108 It has been reported that left-handedness is more common in males. Suppose the following table summarizes data from $n = 6500$ individuals categorized as male/female and left/right handedness. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence: $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$. Use Yate's correction. Verify your answer using the `R` function `chisq.test()`. How is your conclusion related to the conjecture that left-handedness is more common in males?

	Right Handed	Left Handed	Totals
Male	2597	425	3022
Female	3128	350	3478
Totals	5725	775	6500

SOLUTION: With Yate's correction: $X^2 = 24.261$. The degrees of freedom are $(n_r - 1)(n_c - 1) = 1$. Reject H_o if X^2 is greater than or equal to $\chi^2_{1,\alpha} = 3.841$. The following tables contain the expected counts and the terms of the χ^2-statistic:

Table 2.8: Expected counts $E_{i,j}$

	Right Handed	Left Handed	Totals
Male	2661.68	360.32	3022.00
Female	3063.32	414.68	3478.00
Totals	5725.00	775.00	6500.00

Table 2.9: X^2 statistic terms (with Yates's correction) $(|O_i - E_i| - 0.5)^2/E_i$

	Right Handed	Left Handed	Totals
Male	1.55	11.43	12.98
Female	1.34	9.93	11.28
Totals	2.89	21.37	24.26

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

The following `R` code can be used to perform the test. Note that Yate's correction is applied by default.

```
> xtab = matrix(c(2597,3128,425,350),2,2)
> xtab
```

```
     [,1] [,2]
[1,] 2597  425
[2,] 3128  350
> chisq.test(xtab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  xtab
X-squared = 24.261, df = 1, p-value = 8.414e-07
```

We obtain the same test statistic $X^2 = 24.261$.

If the frequency of left-handedness did not differ by gender, then the null hypothesis of row/column independence would hold. We therefore conclude that left-handedness frequency does differ by gender.

Problem 2.109 A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 231$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Perform a χ^2 test for the null hypothesis of row and column independence against the alternative hypothesis:

$$H_o : p_{i,j} = r_i c_j \text{ for all } i, j \text{ against } H_a : p_{i,j} \neq r_i c_j \text{ for some } i, j.$$

Use significance level $\alpha = 0.05$. Perform the test without Yates' correction.

Observed counts $O_{i,j}$

	1	2	3	Totals
1	128	40	19	187
2	17	10	17	44
Totals	145	50	36	231

SOLUTION: The expected counts are

$$E_{i,j} = R_i C_j / n, \;\; i = 1, 2; \;\; j = 1, 2, 3,$$

where R_i, C_j are the row and column totals and $n = 231$ is the total counts. The the expected counts are given in the table below.

Table 2.10: Expected counts $E_{i,j}$

	1	2	3	Totals
1	117.38	40.48	29.14	187.00
2	27.62	9.52	6.86	44.00
Totals	145.00	50.00	36.00	231.00

The terms of the χ^2-statistic are given in the following table:

Table 2.11: X^2 statistic terms $(O_i - E_i)^2/E_i$

	1	2	3	Totals
1	0.96	0.01	3.53	4.50
2	4.08	0.02	15.00	19.11
Totals	5.04	0.03	18.53	23.61

The χ^2 statistic has $(n_r - 1)(n_c - 1) = 1 \times 2 = 2$ degrees of freedom. The appropriate ciritcal value is $\chi^2_{2,0.05} = 5.991$. The statistic is given by

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}(O_{i,j} - E_{i,j})^2/E_{i,j} = 23.61.$$

Note that the term for $i, j = 2, 3$ exceeds $\chi^2_{2,0.05}$ so only this one needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.110 Four subspecies of a type of plant are hypothesized to occur in a given habitat according to the hypothetical relative frequencies $(p_1, p_2, p_3, p_4) = (0.45, 0.15, 0.15, 0.25)$. A random sample of $n = 260$ plants results in the observed counts for each subspecies:

	1	2	3	4	Totals
Observed counts O_i	133	57	51	19	260
Hypothetical frequencies p_i	0.45	0.15	0.15	0.25	1.00
Expected counts					260

(a) For each subspecies, calculate the expected count assuming that the hypothetical frequencies are true. Place you answers in the table above. Does one subspecies appear to be overrepresented?

(b) Use a χ^2 test for null hypothesis H_o : p_i are the true population frequencies. Use significance level $\alpha = 0.05$. Yate's correction is not needed.

SOLUTION:

(a) According to the table, species 4 is *under-represented* in the sample.

	1	2	3	4	Totals
Observed counts O_i	133	57	51	19	260
Expected counts E_i	117.00	39.00	39.00	65.00	260
$(O_i - E_i)^2/E_i$	2.19	8.31	3.69	32.55	46.74

(b) Without Yate's correction: $X^2 = 46.742$. Reject H_o if

$$X^2 \geq \chi^2_{k-1,\alpha} = 7.815.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 3.944e-10).

Problem 2.111 A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 555$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$. Do not use Yate's correction.

Table 2.12: Observed counts $O_{i,j}$

	1	2	3	Totals
1	162	72	78	312
2	77	82	84	243
Totals	239	154	162	555

SOLUTION: The following tables contain the expected counts and the terms of the χ^2-statistic:

Expected counts $E_{i,j}$

	1	2	3	Totals
1	134.00	87.00	91.00	312.00
2	105.00	67.00	71.00	243.00
Totals	239.00	154.00	162.00	555.00

X^2 statistic terms $(O_i - E_i)^2/E_i$

	1	2	3	Totals
1	6.00	2.00	2.00	10.00
2	7.00	3.00	2.00	12.00
Totals	13.00	5.00	4.00	22.00

Without Yate's correction: $X^2 = 22.877$. Reject H_o if

$$X^2 \geq \chi^2_{(n_r-1)(n_c-1),\alpha} = 5.991.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 1.077e-05).

Problem 2.112 The observed counts for $k = 3$ categories based on a random sample of size $n = 224$ are given in the following table. Hypothetical population frequencies p_i^o are also given in the table. Use a χ^2 goodness of fit test for null and alternative hypotheses:

$$H_o : p_i = p_i^o \text{ for all } i = 1, 2, 3 \text{ against } H_a : p_i \neq p_i^o \text{ for some } i = 1, 2, 3,$$

where $p_i^o = 1/3$, $i = 1, 2, 3$. Use significance level $\alpha = 0.05$. Do not use Yates's correction.

$i =$	1	2	3	Totals
Observed counts O_i	39	80	105	224
Hypothetical frequencies p_i^o	1/3	1/3	1/3	1.00
Observed frequencies $\hat{p}_i$	0.17	0.36	0.47	1.00

SOLUTION: The expected counts are

$$E_i = np_i^o = 224 \times 1/3 = 74.67, \;\; i = 1, 2, 3,$$

and are given in the table below. There are $k = 3$ categories, so the χ^2 statistic has $k - 1 = 2$ degrees of freedom. The appropriate ciritcal value is $\chi^2_{2,0.05} = 5.991$. The statistic is given by

$$X^2 = \sum_{i=1}^{3} (O_i - E_i)^2/E_i = 17.04 + 0.38 + 12.32 = 29.74,$$

using terms given in the table below. Note that the terms for $i = 1, 3$ each exceed $\chi^2_{2,0.05}$ so only one of these needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

	1	2	3	Totals
Observed counts O_i	39	80	105	224
Expected counts E_i	74.67	74.67	74.67	224.00
$(O_i - E_i)^2/E_i$	17.04	0.38	12.32	29.74

Problem 2.113 A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 231$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$. Do not use Yates's correction.

Table 2.13: Observed counts $O_{i,j}$

	1	2	3	Totals
1	134	23	37	194
2	7	17	13	37
Totals	141	40	50	231

SOLUTION: The expected counts are

$$E_{i,j} = R_i C_j/n, \;\; i = 1, 2; \;\; j = 1, 2, 3,$$

where R_i, C_j are the row and column totals and $n = 231$ is the total count. The values are given in the table below. The χ^2 statistic has $(n_r - 1)(n_c - 1) = 1 \times 2 = 2$ degrees of freedom. The appropriate critcal value is $\chi^2_{2,0.05} = 5.991$. The statistic is given by

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} (O_{i,j} - E_{i,j})^2/E_{i,j} = 37.364,$$

using terms given in the table below. Note that the term for $i, j = 2, 1$ and $i, j = 2, 2$ exceeds $\chi^2_{2,0.05}$ so only this one needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

Table 2.14: Expected counts $E_{i,j}$

	1	2	3	Totals
1	118.42	33.59	41.99	194.00
2	22.58	6.41	8.01	37.00
Totals	141.00	40.00	50.00	231.00

Table 2.15: X^2 statistic terms $(O_i - E_i)^2/E_i$

	1	2	3	Totals
1	2.05	3.34	0.59	5.98
2	10.75	17.51	3.11	31.38
Totals	12.81	20.85	3.70	37.36

Problem 2.114 According to the theory of Mendelian inheritance, the phenotypic ratios of a dihybrid cross (of which there are 4) are given by the ratios 9:3:3:1. Suppose in a sample of $n = 78$ species of plants believed to be a dihybrid cross, the four relevant traits were observed with the frequencies reported in the following table.

Traits	1	2	3	4	Totals
Observed counts O_i	19	13	19	27	78
Hypothetical frequencies p_i	1/16	3/16	3/16	9/16	1.00
Observed frequencies $\hat{p}_i$	0.24	0.17	0.24	0.35	1.00

Hypothetical population frequencies p_i are also given in the table. Use a χ^2 test for null hypothesis

$$H_o : p_i \text{ are the true population frequencies.}$$

Use significance level $\alpha = 0.05$. Use Yate's correction procedure.

SOLUTION: The expected counts are given by $E_i = np_i$. For example, $E_1 = 78 \times 1/16 = 4.875$. Then, using Yate's correction the test statistic is

$$X^2 = \sum_i X_i^2,$$

where

$$X_i^2 = \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

For example,

$$X_1^2 = \frac{(|19 - 4.875| - 0.5)^2}{4.875} \approx 38.08$$

The remaining expected counts and terms are then

	1	2	3	4	Totals
Observed counts O_i	19	13	19	27	78
Expected counts E_i	4.88	14.62	14.62	43.88	78
$(\lvert O_i - E_i\rvert - 0.5)^2/E_i$	38.08	0.09	1.03	6.11	45.30

With Yate's correction: $X^2 = 45.305$. Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. The P-value = 7.97e-10 can be obtained by the R command

```
> 1-pchisq(45.305,3)
[1] 7.969746e-10
```

Problem 2.115 In statistical genetics a *dihybrid cross* produces phenotype pairs in the ratio 9:3:3:1. Suppose for this type of cross in a certain plant, the phenotypes Tall/Yellow, Short/Yellow, Tall/Green, Short/Green are expected to conform to these ratios. Suppose a sample of $n = 120$ of this cross yields the following counts:

	Tall/Yellow	Short/Yellow	Tall/Green	Short/Green	Totals
Observed counts O_i	68	22	20	10	120
Hypothetical frequencies p_i	9/16	3/16	3/16	1/16	1
Expected counts					120

(a) For each phenotype pair, calculate the expected count assuming that the hypothetical frequencies are true. Place you answers in the table above.
(b) Perform a χ^2 test against the null hypothesis H_o : p_i are the true population frequencies. Use significance level $\alpha = 0.05$. Yate's correction is not needed.

SOLUTION:

(a) The expected counts $E_i = np_i$ are given in the following table:

	Tall/Yellow	Short/Yellow	Tall/Green	Short/Green	Totals
Observed counts O_i	68	22	20	10	120
Expected counts E_i	67.5	22.5	22.5	7.5	120.0
$(O_i - E_i)^2/E_i$	0.0037	0.0111	0.2778	0.8333	1.125926

(b) Without Yate's correction

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \approx 0.0037 + 0.0111 + 0.2778 + 0.8333 = 1.126.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

Problem 2.116 A contingency table with $n_r = 3$ rows and $n_c = 3$ columns based on a random sample of size $n = 200$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Perform a χ^2 test against the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$.

Observed counts $O_{i,j}$				
	1	2	3	Totals
1	20	21	51	92
2	12	21	6	39
3	29	37	3	69
Totals	61	79	60	200

SOLUTION: The expected counts $E_{i,j} = R_i C_j / n$ are given in the following table, where R_i, C_j are the row and column totals. For example,

$$E_{1,3} = \frac{R_1 C_3}{n} = \frac{92 \times 60}{200} = 27.6.$$

The expected counts and the terms of the χ^2-statistic are givren in the following tables:

Expected counts $E_{i,j}$				
	1	2	3	Totals
1	28.06	36.34	27.60	92.00
2	11.89	15.40	11.70	39.00
3	21.05	27.25	20.70	69.00
Totals	61.00	79.00	60.00	200.00

X^2 statistic terms $(O_{i,j} - E_{i,j})^2 / E_{i,j}$				
	1	2	3	Totals
1	2.32	6.48	19.84	28.63
2	0.00	2.03	2.78	4.81
3	3.01	3.48	15.13	21.63
Totals	5.32	11.99	37.75	55.07

Reject H_o if X^2 is greater than or equal to $\chi^2_{(n_r-1)(n_c-1),\alpha} = \chi^2_{4,0.05} = 9.488$. Without Yate's correction, we have

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \approx 55.066.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. Note that is suffices to calculate only

$$\frac{(O_{1,3} - E_{1,3})^2}{E_{1,3}} = 19.84 \text{ or } \frac{(O_{3,3} - E_{3,3})^2}{E_{3,3}} = 15.13,$$

since we could then conclude that $X^2 \geq \chi^2_{4,0.05}$, without the need to calculate X^2 (since all other terms are positive).

Problem 2.117 Suppose the ability to reduce asthma symptoms of an experimental treatment is compared to a standard treatment. The experimental treatment is administered to $n_1 = 178$ subjects, and the standard treatment is administrated to $n_2 = 79$ subjects. For each subject, whether or not the treatment succeeded in reducing asthma symptoms was observed. Suppose the results are summarized in the following contingency table:

	Experimental Treatment	Standard Treatment	
Reduced asthma symptoms	154	53	207
Asthma symptoms unaffected	24	26	50
Total	178	79	257

(a) Test hypothesis $H_o : p_1 = p_2$ against $H_a : p_1 \neq p_2$, where p_1, p_2 are the respective proportion of subjects for which symptoms were reduced for each treatment. Use a two-sample difference in proportion test. Report a P-value. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$?

(b) Construct a level 0.95 confidence interval for the log odds ratio of *Reduced asthma symptoms* between groups *Experimental Treatment* and *Standard Treatment*. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$?

(c) Suppose we may interpret the table as a contingency table with $n_r = 2$ rows and $n_c = 2$ columns based on a random sample of size $n = 257$. Assume the probability of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$.

(d) Verify that the null hypotheses of all three tests are the same.

SOLUTION:

(a) The pooled estimate under the null hypothesis of $p_0 = p_1 = p_2$ is

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{154 + 53}{178 + 79} = 0.805.$$

The test statistic is

$$\begin{aligned} Z_{obs} &= \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1 - \hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.671 - 0.865}{\sqrt{0.805(1 - 0.805)\left(\frac{1}{178} + \frac{1}{79}\right)}} \\ &= \frac{-0.194}{0.0535} \\ &= -3.63 \end{aligned}$$

The P-value is, for $Z \sim N(0, 1)$,

$$\alpha_{obs} = 2P(Z > |Z_{obs}|) = 2P(Z > 3.63) = 0.000283.$$

Since $\alpha_{obs} \leq \alpha$ we reject the null hypothesis at an α significance level.

(b) The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{154 \times 26}{53 \times 154} = 3.148.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{154} + \frac{1}{53} + \frac{1}{24} + \frac{1}{26}} \\ &= 0.325. \end{aligned}$$

The level $1-\alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.147 \pm 1.96 \times 0.325 \\ &= 1.147 \pm 0.637 \end{aligned}$$

or equivalently, $CI = (0.51, 1.783)$. Since the CI does not contain 0 we reject the null hypothesis at an α significance level.

(c) The row totals are $R_1 = 207$, $R_2 = 50$, the column totals are $C_1 = 178$, $C_2 = 79$, with $N = 257$. The expected counts are then

$$E_{ij} = \frac{R_i C_j}{N}.$$

For example, $E_{11} = R_1 C_1 / N = 207 \times 178/257 \approx 143.37$. The remaining expected counts are then

E_{ij}	1	2	Totals
1	143.37	63.63	207.00
2	34.63	15.37	50.00
Totals	178.00	79.00	257.00

The χ^2 statistic is then

$$X^2 = \sum_{ij} X_{ij}^2,$$

where

$$X_{ij}^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

For example,

$$X_{11}^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(154 - 143.37)^2}{143.37} \approx 0.79.$$

The remaining terms are then (without Yate's correction)

X_{ij}^2	1	2	Totals
1	0.79	1.78	2.56
2	3.26	7.35	10.62
Totals	4.05	9.13	13.18

Then $X^2 = 13.18$. Reject H_o if X^2 is greater than or equal to $\chi^2_{(n_r-1)(n_c-1),\alpha} = 3.841$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. The P-value $= 0.000283$ can be obtained from the `R` commands

```
> 1-pchisq(13.18,1)
[1] 0.0002829528
```

(d) To show that the null hypothesis of Part (b) is the same as that of Part (a), note that the odds ratio is interpretable as

$$OR = \frac{Odds(\text{Reduced Symptoms} \mid \text{Experimental Treatment})}{Odds(\text{Reduced Symptoms} \mid \text{Standard Treatment})}.$$

Then note that $OR = 1$ if and only if

$$P(\text{Reduced Symptoms} \mid \text{Experimental Treatment}) = P(\text{Reduced Symptoms} \mid \text{Standard Treatment}).$$

But from the definition of p_1, p_2 we have

$$\begin{aligned} p_1 &= P(\text{Reduced Symptoms} \mid \text{Experimental Treatment}) \\ p_2 &= P(\text{Reduced Symptoms} \mid \text{Standard Treatment}). \end{aligned}$$

So $OR = 1$ if and only if $p_1 = p_2$, so that the null hypotheses of Parts (a) and (b) are the same. Next, in Part (c) we are given cell probabilities p_{ij}. We may relate p_1, p_2 of Part (a) to these cell probabilities by

$$\begin{aligned} p_1 &= \frac{p_{11}}{p_{11}+p_{21}} = \frac{p_{11}}{c_1} \\ p_2 &= \frac{p_{12}}{p_{12}+p_{22}} = \frac{p_{12}}{c_2}. \end{aligned}$$

Then under the null hypothesis

$$p_{11} = r_1 c_1$$

which is equivalent to

$$p_1 = \frac{p_{11}}{c_1} = r_1.$$

Similarly, under the null hypothesis,

$$p_{12} = r_1 c_2,$$

and so

$$p_2 = \frac{p_{12}}{c_2} = r_1.$$

Therefore, under the null hypothesis, $p_1 = p_2$, which is the same null hypothesis as for Parts (a) and (b).

2.9 ANOVA

Problem 2.118 Independent samples for $k = 3$ treatments are summarized in the table below (sample means, sample standard deviations and sample sizes are given). Assume sample j is from a normally distributed population with mean μ_j and fixed variance σ^2. An ANOVA table is also given.

(a) Based on the ANOVA table, can the null hypothesis $H_o: \mu_i = \mu_j$ for all i, j be rejected in favor of the alternative hypothesis $H_a: \mu_i \neq \mu_j$ for some i, j? Use significance level $\alpha = 0.05$.

Data Summary

	$\bar{X}_i$	S_i	n_i
Treatment 1	21.83	6.13	13
Treatment 2	23.68	5.57	12
Treatment 3	28.68	3.64	16

ANOVA Table

	SS	DF	MS	F
Treatment	368.72	2	184.36	7.07
Error	990.61	38	26.07	
Total	1359.33	40		

(b) Using the Bonferroni procedure, construct simultaneous confidence intervals for those treatment differences required to decide whether or not Treatment 3 has the largest mean. Use a familywise error rate of $\alpha_{FWE} = 0.05$.

SOLUTION:

(a) $F = 7.072$. Reject H_o if $F \geq F_{k-1,n-k,\alpha} = 3.245$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

(b) We need to construct confidence intervals for $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$, so there are $m = 2$ comparisons. If $\alpha_{FWE} = 0.05$ we set $\alpha = 0.05$ in the following confidence intervals:

$$\bar{X}_i - \bar{X}_j \pm t_{n-k,\alpha/(m2)} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

so we need critical value

$$t_{n-k,\alpha/(m2)} = t_{41-3,0.05/4} = t_{38,0.0125} = 2.334.$$

From the ANOVA table $MSE = 26.07$.
For $\mu_1 - \mu_3$

$$\begin{aligned} CI &= 21.83 - 28.68 \pm 2.334\sqrt{26.07\left(\frac{1}{13} + \frac{1}{16}\right)} \\ &= -6.86 \pm 4.449 \\ &= (-11.299, -2.401). \end{aligned}$$

For $\mu_2 - \mu_3$

$$\begin{aligned} CI &= 23.68 - 28.68 \pm 2.334\sqrt{26.07\left(\frac{1}{12} + \frac{1}{16}\right)} \\ &= -5.00 \pm 4.550 \\ &= (-9.550, -0.450). \end{aligned}$$

From the confidence intervals (which contain only negative values) we conclude $\mu_3 > \mu_1$ and $\mu_3 > \mu_2$ with $\alpha_{FWE} = 0.05$.

Problem 2.119 Independent samples for $k = 3$ treatments are summarized in the table below. Assume sample j is from a normally distributed population with mean μ_j and fixed variance σ^2.

	1	2	3	4	5	$\bar{X}_i$	S_i	n_i
Treatment 1	13.13	15.16	10.60	16.41	17.99	14.66	2.88	5
Treatment 2	9.04	6.94	9.30	9.00	10.69	8.99	1.34	5
Treatment 3	20.55	16.37	20.20	14.39	21.89	18.68	3.16	5

(a) Construct an ANOVA table (fill in the 9 spaces in the ANOVA table below).

	SS	DF	MS	F
Treatment	_____	_____	_____	_____
Error	_____	_____	_____	
Total	_____	_____		

(b) Use an F-test for null hypothesis $H_o: \mu_i = \mu_j$ for all i, j. Use significance level $\alpha = 0.05$.

SOLUTION:

(a) The ANOVA table is

ANOVA Table

	SS	DF	MS	F
Treatment	236.83	2.00	118.42	17.70
Error	80.27	12.00	6.69	
Total	317.11	14.00		

(b) $F = 17.702$. Reject H_o if

$$F \geq F_{k-1,n-k,\alpha} = 3.885.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$ (P-value = 0.0002631).

Problem 2.120 A new type of insecticide was tested against 3 standard alternatives. Each of the four insecticides was tested in 6 separate plots (requiring 24 separate plots). The percentage crop loss was recorded for each of the 24 plots at the end of the experiment. For each insecticide, the sample mean and sample standard deviation of the 6 outcomes is given in the following table (for example, the average percentage crop loss for the six plots using Standard Insecticide B was 9.91). The ANOVA table for the data is also given. Using a Bonferroni multiple comparison procedure, determine whether or not the new insecticide resulted in the lowest average percentage crop loss of all the insecticides tested. Use a familywise error rate of $\alpha_{FWE} = 0.05$.

	$\bar{X}_i$	S_i	n_i
New Insecticide	5.17	3.34	6
Standard Insecticide A	10.65	3.04	6
Standard Insecticide B	9.91	3.20	6
Standard Insecticide C	9.88	2.51	6

	SS	DF	MS	F
Treatment	113.69	3	37.90	4.10
Error	184.65	20	9.23	
Total	298.35	23		

SOLUTION: We need $m = 3$ comparisons, to compare $\mu_1 - \mu_i$, $i = 2, 3, 4$. Since $n_i = 6$ for $i = 1, 2, 3, 4$, the CIs take form

$$\begin{aligned} CI &= \bar{X}_i - \bar{X}_j \pm t_{n-k,\alpha_{FWE}/(m2)} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm t_{20,0.05/6} \sqrt{9.23\left(\frac{1}{6} + \frac{1}{6}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm 2.613 \sqrt{9.23\left(\frac{1}{6} + \frac{1}{6}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm 4.58. \end{aligned}$$

The CIs are given in the following table. We can conclude with confidence $1 - \alpha_{FWE} = 0.95$ that μ_1 is the smallest mean, since $\mu_1 - \mu_i < 0$ for $i = 2, 3, 4$ within each comparison.

Multiple comparisons (Bonferroni procedure)

	Treatment 1	Treatment 2	Difference	Margin of Error	LB	UB
Comp 1	1	2	-5.48	4.58	-10.06	-0.90
Comp 2	1	3	-4.73	4.58	-9.32	-0.15
Comp 3	1	4	-4.71	4.58	-9.30	-0.13

Problem 2.121 The monthly power consumption in kwh of samples of $k = 4$ brands of humidifier were monitored, with results given in Table 2.16. Assume the sample for Brand i is from a normally distributed population with mean μ_i and fixed variance σ^2.

Table 2.16: Humidifier Data for Problem 2.121.

	1	2	3	4	5	$\bar{X}_i$	S_i	n_i
Brand 1	24.85	22.08	25.91	28.74	21.99	24.71	2.83	5
Brand 2	19.00	14.43	23.73	15.74	23.10	19.20	4.20	5
Brand 3	17.14	13.37	18.64	15.31		16.12	2.28	4
Brand 4	14.38	17.28	14.65	9.13		13.86	3.41	4

(a) Construct an ANOVA table.

(b) Use an F-test for null hypothesis

$$H_o : \mu_i = \mu_j \text{ for all } i, j \text{ against } H_a : \mu_i \neq \mu_j \text{ for some } i, j.$$

Use significance level $\alpha = 0.05$. Construct a rejection region, and also give a P-value.

(c) Before the study was carried out, it was conjectured that Brand 1 had the highest mean power consumption. In the study, Brand 1 did have the highest sample mean. However, rejection of the null hypothesis of equal treatment means does not imply by itself that the observed sample mean ordering is the same as the true mean ordering. Construct multiple confidence intervals using an appropriate Bonferroni procedure to determine whether or not Brand 1 has the highest power consumption. Use a familywise error rate of $\alpha_{FWE} = 0.05$.

(d) Construct side by side boxplots of the data for the 4 humidifiers. Does the equality of variance assumption seem reasonable? Confirm using the `bartlett.test` function in `R`.
(e) Verify parts (a) and (b) using the `aov` functions in `R`.
(f) Use the `TukeyHSD` in `R` function to construct simultaneous confidence intervals for all paired differences in mean, using $\alpha_{FWE} = 0.05$. Do you reach the same conclusion as in part (c)?

SOLUTION:

(a) Total sample size is

$$n = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 4 + 4 = 18.$$

The total mean is

$$\hat{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_2\bar{X}_2 + n_4\bar{X}_4}{n_1 + n_2 + n_3 + n_4} = \frac{5 \times 24.71 + 5 \times 19.20 + 4 \times 16.12 + 4 \times 13.86}{18} = 18.86.$$

The SST is

$$SST = \sum_{i=1}^{4} n_i(\bar{X}_i - \hat{X})^2 = 5(24.71-18.85)^2+5(19.20-18.85)^2+4(16.12-18.85)^2+4(13.86-18.85)^2 = 301.7.$$

The SSE is

$$SSE = \sum_{i=1}^{4}(n_i - 1)S_i^2 = (5-1)2.83^2 + (5-1)4.20^2 + (4-1)2.28^2 + (4-1)3.41^2 = 153.1.$$

The Treatment DF is $k - 1 = 4 - 1$, the Error DF is $n - k = 18 - 4 = 14$ and the Total DF is $n - 1 = 18 - 1 = 17$, so we have

$$MST = SST/(k-1) = 301.7/3 = 100.6 \text{ and } MSE = SSE/(n-k) = 153.1/14 = 10.9$$

and F-statistic

$$F = MSE/MST = 100.6/10.9 = 9.2,$$

with P-value $P = P(F_{3,14} > 9.2) = 0.00129$. The ANOVA table is given in Table 2.17.

Table 2.17: ANOVA Table for Problem 2.121 (a).

	SS	DF	MS	F	Pval
Treatment	301.7	3	100.6	9.2	0.00129
Error	153.1	14	10.9		
Total	454.8	17			

(b) From Table 2.17 $F = 9.2$. Reject H_o if $F \geq F_{k-1,n-k,\alpha} = 3.344$. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. P-value $= 0.00129$ (from Table 2.17).
(c) To resolve the question, construct confidence intervals for $\mu_i - \mu_j$ for $i = 1$ and $j = 2, 3, 4$, so $m = 3$ confidence intervals are needed, given by

$$\bar{X}_i - \bar{X}_j \pm t_{\alpha/(m2),n-k}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Since $m = 3$, and $\alpha/(m2) = 0.05/6 = 0.00833$ the required critical value is

$$t_{\alpha/(m2),n-k} = t_{0.00833,14} = 2.718,$$

with $MSE = 10.9$ from Table 2.17, so the CIs are given by

$$\bar{X}_i - \bar{X}_j \pm 2.718 \times \sqrt{10.9\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

The CIs for the appropriate indices are given in Table 2.18. For example, for $i = 1$, $j = 2$ we have

$$24.71 - 19.20 \pm 2.717\sqrt{10.9 \times (1/5 + 1/5)} = (-0.17, 11.20).$$

We can conclude that $\mu_1 > \mu_3$ and $\mu_1 > \mu_4$ but not $\mu_1 > \mu_2$, since the CI for $\mu_1 - \mu_2$ contains 0, but the others do not.

Table 2.18: Multiple comparisons (Bonferroni procedure) for Problem 2.121 (c).

	Brand i	Brand j	Difference	Margin of Error	LB	UB
1	1	2	5.51	5.68	-0.17	11.20
2	1	3	8.60	6.03	2.57	14.62
3	1	4	10.85	6.03	4.83	16.88

(d) The following code draws the required graph (Figure 2.8). Judging by the IQRs, there is little indication that the variances differ, with the possible exception of Brand 2. The P-value for rejecting equality of variances is $P = 0.747$ using Bartlett's test. The assumption of equal variances is therefore reasonable.

```
Brand1 = c(24.85 , 22.08 , 25.91 , 28.74 , 21.99)
Brand2 = c(19.00 , 14.43 , 23.73 , 15.74 , 23.10)
Brand3 = c(17.14 , 13.37 , 18.64 , 15.31)
Brand4 = c(14.38 , 17.28 , 14.65 , 9.13)
y = c(Brand1, Brand2, Brand3, Brand4)
x = factor(c(rep(1,5),rep(2,5),rep(3,4),rep(4,4)))
p = bartlett.test(y~x)$p.value
boxplot(y~x, xlab='Brand',ylab='Monthly Power Consumption (kwh)')
title(paste("P = ",signif(p,3)))
```

(e) The following code calculates the ANOVA table and gives the appropriate P-value. The results conform to those given above, within a small rounding error.

```
> fit = aov(y~x)
> summary(fit)
Df Sum Sq Mean Sq F value  Pr(>F)
x            3  302.1  100.69   9.208 0.00128 **
Residuals   14  153.1   10.93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

(f) The following code implements the Tukey procedure. We reach the same conlcusion as in part (c) (the confidences intervals for $\mu_i - \mu_j$ contain zero for $i, j = 2, 1$, but not $i, j = 3, 1$ and $i, j = 4, 1$.

```
> fit = aov(y~x)
> summary(fit)
Df Sum Sq Mean Sq F value  Pr(>F)
x            3  302.1  100.69   9.208 0.00128 **
Residuals   14  153.1   10.93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(fit)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = y ~ x)

$x
diff        lwr        upr     p adj
2-1  -5.514 -11.592711  0.5647113 0.0813865
3-1  -8.599 -15.046447 -2.1515530 0.0080579
4-1 -10.854 -17.301447 -4.4065530 0.0012011
3-2  -3.085  -9.532447  3.3624470 0.5248661
4-2  -5.340 -11.787447  1.1074470 0.1212242
4-3  -2.255  -9.051206  4.5412058 0.7712737
>
```

2.10 Linear Regression

Problem 2.122 We are given a multiple regression model, with sample size sample $n = 81$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

The following coefficient table is output:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3105	12.8428	0.569	0.571
x1	0.6285	0.9608	0.654	0.515
x2	0.8546	0.7705	1.109	0.271

However, an appropriate F-test shows that the full model significantly reduces the SSE compared to the null model $y = \beta_0 + \epsilon$. Suppose we are given the following error sums of squares SSE:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	88,748.85
M_1	$y = \beta_0 + \beta_1 x_1 + \epsilon$	82,952.39
M_2	$y = \beta_0 + \beta_2 x_2 + \epsilon$	82,112.42
M_{12}	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	81,664.43

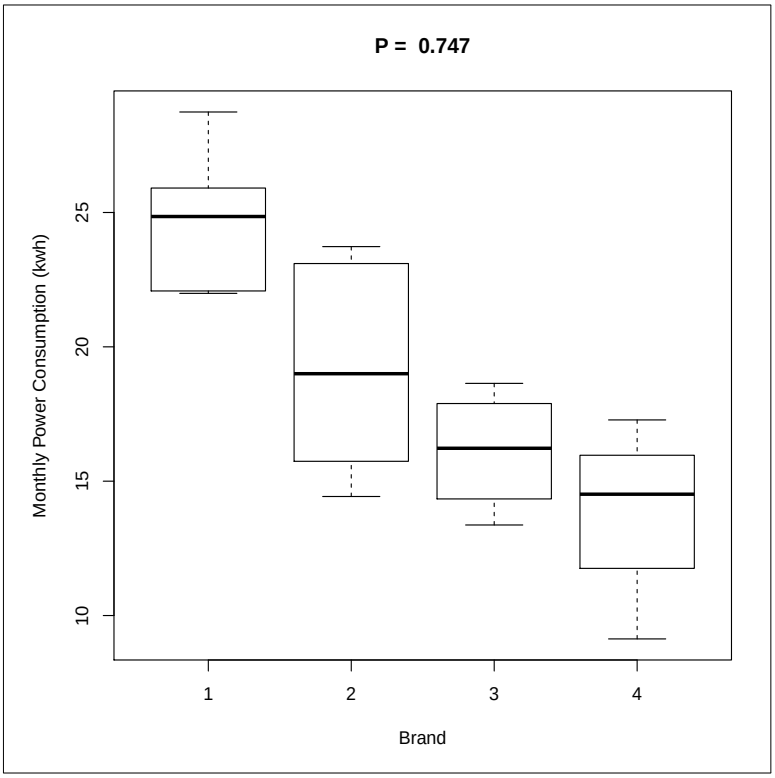

Figure 2.8: Plot for Problem 2.121 (d).

Of the four models, which has the highest value of R^2, and which has the highest value of R^2_{adj}? Justify your answer in each case, and give the actual highest value.

SOLUTION: The largest R^2 must be for the full model (ie it must have the smallest SSE). The total sum of squares $SSTO$ is the SSE for the null model M_0, so $SSTO = 88748.85$. So the largest R^2 is

$$R^2[M_{12}] = 1 - \frac{SSE[M_{12}]}{SSTO} = 1 - \frac{81,664.43}{88,748.85} = 0.0798.$$

As for R^2_{adj}, this is zero for M_0. We can also eliminate M_1, since $SSE[M_1] > SSE[M_2]$, and both models have the same number of parameters. So,

$$\begin{aligned} R^2[M_2] &= 1 - \frac{SSE[M_2]/(n-2)}{SSTO/(n-1)} = 1 - \frac{82,112.42/79}{88,748.85/80} = 0.06307, \\ R^2[M_{12}] &= 1 - \frac{SSE[M_{12}]/(n-3)}{SSTO/(n-1)} = 1 - \frac{81,664.43/78}{88,748.85/80} = 0.05623. \end{aligned}$$

The largest value is $R^2[M_2] = 0.06307$.

Problem 2.123 Two variables Y and X are believed to have the following relationship:

$$Y = aX^b$$

for two constants a, b. According to a certain conjecture, Y is proportional to the square root of X. In order to resolve this question paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled, where $n = 51$. The simple linear regression model

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

is fit, with the following output:

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.3137646 0.02512828 92.078108 1.447613e-56
log(x)      0.4985705 0.05774591  8.633867 2.084971e-11
```

Formulate appropriate null and alternative hypotheses for this question in terms of the regression coefficients β_0 and/or β_1. Is there evidence at an $\alpha = 0.05$ significance level with which to reject the conjecture?

SOLUTION: The hypotheses are

$$H_o : \beta_1 = 1/2 \text{ against } H_a : \beta_1 \neq 1/2.$$

The appropriate t-statistic is

$$T = \frac{\hat{\beta}_1 - 1/2}{SE_{\hat{\beta}_1}} = \frac{0.4985705 - 1/2}{0.05774591} \approx -0.0247.$$

Since $|T| < t_{49,0.025}$ we do not reject the conjecture at significance level $\alpha = 0.05$.

Problem 2.124 The following full linear regression model is considered:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

An all-subsets model selection procedure is to be used to determine which of the predictors and interactions to retain. The relevant SSE values are given in the following table. The sample size is $n = 25$. Which model possesses the largest coefficient of determination R^2? Which model possesses the largest adjusted coefficient of determination R^2_{adj}?

	Model	SSE
1	y=1	10638.06
2	y=x1	7810.40
3	y=x2	2210.16
4	y=x1+x2	29.23
5	y=x1+x2+x1*x2	27.39

SOLUTION: The model with the highest R^2 must be model 5, since all other models are reduced models.

The total sum of squares is given by the null model (model 1):

$$SSTO = 10638.06$$

	Model	SSE	q	R^2_{adj}
1	y=1	10638.06	0	0.00000
2	y=x1	7810.40	1	0.23388
3	y=x2	2210.16	1	0.78321
4	y=x1+x2	29.23	2	0.99700
5	y=x1+x2+x1*x2	27.39	3	0.99706

The formula is

$$R^2_{adj} = 1 - \frac{SSE/(n-(q+1))}{SSTO/(n-1)}.$$

where q is the number of predictors. We can construct table:

Model 5 has the highest R^2_{adj}.

Problem 2.125 Given a single predictor x and response y, a polynomial regression model is considered:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon.$$

Suppose we consider four models (full model and 3 reduced models). Suppose further that the sample size is $n = 41$, and that the four models are fit, yielding the following error sums of squares SSE:

Table 2.19:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	229.26
M_1	$y = \beta_0 + \beta_1 x + \epsilon$	35.388
M_2	$y = \beta_0 + \beta_2 x^2 + \epsilon$	31.923
M_{12}	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$	31.294

Suppose we rank the four models according to the following two rules:

Rule 1: If two models M and M' have the same number of predictors, the one with the smaller SSE is ranked higher.

Rule 2: If model M' can be obtained from model M by adding a single predictor, and the p-value for an F-test comparing full model M' to reduced model M is ≤ 0.05, then M' is ranked higher than M, otherwise it is ranked lower.

The selected model is the one of highest rank. Assuming the models of Table 2.19 can be consistently ranked, what is the highest ranked model?

SOLUTION: Take the following steps:

(1) Directly from Table 2.19 we have $M_2 > M_1$, since M_2 has the smaller SSE.

(2) The F-statistic for comparing M_2 and M_0 is

$$F = \frac{(SSE_0 - SSE_2)}{SSE_2/(n-2)} = \frac{(229.26 - 31.923)}{31.923/39} = 241.0846$$

The appropriate critical value is $F_{1,39;\alpha} = 4.091$, so we conclude $M_2 > M_0$.

(3) The F-statistic for comparing M_{12} and M_2 is

$$F = \frac{(SSE_2 - SSE_{12})}{SSE_{12}/(n-3)} = \frac{(31.923 - 31.294)}{31.294/38} = 0.764$$

The appropriate critical value is $F_{1,38;\alpha} = 4.098$, so we conclude $M_2 > M_{12}$.

This analysis suffices to conclude that M_2 is the highest ranked model.

Problem 2.126 Two variables Y and X are believed to have the following relationship:

$$Y = aX^b$$

for two constants a, b. There is special interest in knowing whether or not this relationship is concave (equivalently, $b < 1$). In order to resolve this question paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled, where $n = 34$. The simple linear regression model

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

is fit, with the following output:

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.7115358 0.09647197  7.37557 2.174990e-08
log(x)      0.8948970 0.03583328 24.97391 1.483693e-22
```

Formulate appropriate null and alternative hypotheses for this question in terms of the regression coefficients β_0 and/or β_1. Is there evidence at an $\alpha = 0.05$ significance level that $b < 1$?

SOLUTION: The appropriate hypotheses are $H_o : b \geq 1$ and $H_a : b < 1$, since we are looking for evidence that $b < 1$. In terms of the regression coefficients this is equivalent to

$$H_o : \beta_1 \geq 1 \text{ and } H_a : \beta_1 < 1.$$

The test statistic is

$$T = \frac{\hat{\beta}_1 - 1}{S_{\hat{\beta}_1}} = \frac{0.8948970 - 1}{0.03583328} = -2.933112,$$

which has a t-distribution with $n-2$ degrees of freedom under H_o. Since $t_{32,0.05} = 1.69$, we reject H_o at a 0.05 significance level, and conclude that $b < 1$.

Problem 2.127 We are given a multiple regression model, with sample size $n = 11$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

This is interpreted as the full model, and we wish to compare it to each reduced submodel. The error sums of squares (SSE) for the full model, and each reduced model, is given in the following table:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	15936.0
M_1	$y = \beta_0 + \beta_1 x_1 + \epsilon$	8790.5
M_2	$y = \beta_0 + \beta_2 x_2 + \epsilon$	5909.5
M_{12}	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	5705.3

Using an appropriate F-test, determine which of the three reduced model M_0, M_1 and M_2 is significantly improved by the full model M_{12}. Use a significance level of $\alpha = 0.05$ for each test.

SOLUTION: The F-statistic for comparing full and reduced models with $q > p$ predictors, respectively, (excluding intercept term) is

$$F = \frac{(SSE_p - SSE_q)/(q-p)}{SSE_q/(n-(q+1))}.$$

Under the null hypothesis of model equivalence the statistic is distributed as $F \sim F_{q-p,n-(q+1)}$.

To compare M_{12} and M_0, $p = 0, q = 2$:

$$F = \frac{(SSE_{M_0} - SSE_{M_{12}})/(2-0)}{SSE_{M_{12}}/(11-3)} = \frac{(15936.0 - 5705.3)/2}{5705.3/8} = 7.172769.$$

The appropriate critical value is $F_{2,8;0.05} = 4.459$. Since $F > F_{2,8;0.05}$ we reject equivalence of M_0 and M_{12} at significance level $\alpha = 0.05$.

To compare M_{12} to M_1, $p = 1, q = 2$:

$$F = \frac{(SSE_{M_1} - SSE_{M_{12}})/(2-1)}{SSE_{M_{12}}/(11-3)} = \frac{(8790.5 - 5705.3)/1}{5705.3/8} = 4.326083.$$

The appropriate critical value is $F_{1,8;0.05} = 5.318$. Since $F < F_{1,8;0.05}$ we do no reject equivalence of M_1 and M_{12} at significance level $\alpha = 0.05$.

To compare M_{12} to M_2, $p = 1, q = 2$:

$$F = \frac{(SSE_{M_2} - SSE_{M_{12}})/(2-1)}{SSE_{M_{12}}/(11-3)} = \frac{(5909.5 - 5705.3)/1}{5705.3/8} = 0.2863303.$$

The appropriate critical value is $F_{1,8;0.05} = 5.318$. Since $F < F_{1,8;0.05}$ we do no reject equivalence of M_2 and M_{12} at significance level $\alpha = 0.05$.

So M_{12} improves M_0, but not M_1 OR M_2. Note that the F-statistic for the M_{12} versus M_2 comparison is necessarily smaller than for the M_{12} versus M_1 comparison. So the non-improvement of M_{12} over M_2 follows from the non-improvement of M_{12} over M_1.

Problem 2.128 A client hires a consulting firm to conduct a study of two types of mutual funds (we'll call them simply Type A and Type B). It uses a simple regression model

$$Y = e^{\beta_0 + \beta_1 X + \epsilon}$$

where $X = 1$ for a Type A mutual fund, and $X = 0$ otherwise; $\epsilon \sim N(0, \sigma^2)$; and Y is the value of an original investment of \$1 after a year (that is, if $Y = 1.05$, the yearly rate of return is 5%). The model is first log-transformed, giving

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; i = 1, \ldots, n. \tag{2.3}$$

A random sample of $n = 62$ paired observations (Y_i, X_i), $i = 1, \ldots, 62$ is collected. A simple least squares regression model is used to fit the model (2.3), producing the following coefficient table:

Coefficient	Estimate	Standard Error	t-value	Pr(>|t|)
$\hat{\beta}_0$	0.0745	0.0040	18.8376	2.43×10^{-34}
$\hat{\beta}_1$	0.0333	0.0056	5.9514	4.13×10^{-8}

The consultant believes Type A mutual funds have a higher average yield, but the client currently purchases mutual funds of Type B, and there would be a significant cost to switching to Type A. Therefore, the consultant will only recommend switching to Type A if there is significant statistical evidence that $\beta_1 > 0.015$ (approximately, that the rate of return of Type A mutual finds exceeds Type B mutual funds by more than 1.5%). Using a significance level of $\alpha = 0.05$, can the consultant recommend switching?

SOLUTION: The hypotheses are

$$H_o : \beta_1 \leq 0.015 \text{ against } H_a : \beta_1 > 0.015.$$

The appropriate t-statistic is

$$T = \frac{\hat{\beta}_1 - 0.015}{SE_{\hat{\beta}_1}} = \frac{0.0333 - 0.015}{0.0056} \approx 3.268.$$

Since$T > t_{60,0.05} = 1.671$ we do not reject the conjecture at significance level $\alpha = 0.05$.

Problem 2.129 There is often interest in determining whether or not two quantities X and Y have a *power-law* relationship:

$$Y = aX^b \tag{2.4}$$

for two constants a, b. Suppose, given $n = 41$ independent paired observations (X_i, Y_i) of these quantities, we fit model

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \beta_2 \log(X_i)^2 + \epsilon_i, \;\; i = 1, \ldots, n,$$

assuming any relevant distributional assumption holds, and get output:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.33180   19.50917   1.965   0.0568 .
log.x         -5.93521    2.67072  -2.222   0.0323 *
log.x.squared  0.02227    0.08868   0.251   0.8031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) How can the output be used to assess the validity of the power-law relationship of Equation (2.4)? What do you conclude?

(b) Give level 95% confidence intervals for parameters a and b, based on this output.

SOLUTION:

(a) If we take a log-transform of the model $\log(Y) = \log(a) + b\log(X)$. We can therefore equate $\beta_0 = \log(a)$ and $\beta_1 = b$ in the linear model. In addition, if the model holds we must have $\beta_2 = 0$. The 2-sided p-value for rejecting null hypothesis $H_o : \beta_2 = 0$ is $p = 0.8031$ from the output. We do not reject the power-law relationship, therefore.

(b) Since $\beta_0 = \log(a)$ and a 95% CI for β_0 is

$$\hat{\beta}_0 \pm 2 \times SE = 38.33180 \pm 2 \times 19.50917 = (-0.68654, 77.35014)$$

a 95% CI for a is

$$e^{\hat{\beta}_0 \pm 2 \times SE} = e^{38.33180 \pm 2 \times 19.50917} = (0.5033, 3.9 \times 10^{33}).$$

Since $\beta_1 = b$ a 95% CI for b is identical to the 95% CI for β_1:

$$\hat{\beta}_0 \pm 2 \times SE = -5.93521 \pm 2 \times 2.67072 = (-11.27665, -0.59377).$$

Problem 2.130 Relationships between the size of two physiological components Y, X of a species of animal often obey a power relationship

$$Y = KX^r,$$

where K and r are two fixed constants. Of course, the value r is not necessarily $r = 1$, but will depend on the size measure, and any number of scaling principles. Suppose we have paired observations (X_i, Y_i), $i = 1, \ldots, n$. Then K and r may be estimated using simple linear regression, after taking the double-log transformation:

$$\log Y_i = \log K + r \log X_i, \tag{2.5}$$

that is, we have intercept and slope $\beta_0 = \log K$, $\beta_1 = r$.

For this problem use data set `Animals` from the `MASS` package, which contains X = average body (kg) and Y = brain (g) weights for 28 species of land animals. We may expect Y to be positively associated with an animal's cognitive abilities. However, we also expect X and Y to be positively associated for reasons having nothing to do with cognitive abilities. Thus, the *encephalization quotient* (EQ) measures the relative brain size after controlling for body size.

(a) Construct a scatter-plot of $\log Brain$ against $\log Body$. Use $\log Body$ as the horizontal axis. Instead of plotting symbols, plot the actual name of the species (here, the `text` command may be used). Does there seem to be a linear trend on the double-log scale? Identify the three most obvious outliers. How do they differ from the remaining species?

(b) Fit the model (2.5), with and without the outliers, and superimpose each fitted line on the scatter-plot. Give the estimates of K and r for each fit.

(c) The *encephalization quotient* (EQ) can be formally defined as the ratio of the actual brain mass to the predicted brain mass based on the species size. If model (2.5) is used to predict brain mass, show that for species i

$$EQ \approx \exp(e_i)$$

where e_i is the residual from the regression fit for that species.

(d) After removing the outliers, rank the species by their EQ. How would the EQ of the outlier species rank?

SOLUTION:

(a) The following code will produce the required plot (Figure 2.9). There seems to be a clear linear trend, with the exception of three outliers, which are all dinosaurs.

```
par(mfrow=c(1,1),pty='m')
with(Animals,plot(log(body),log(brain),type='n',xlim=c(-4,13)))
with(Animals,text(log(body),log(brain),rownames(Animals),cex=0.9))
```

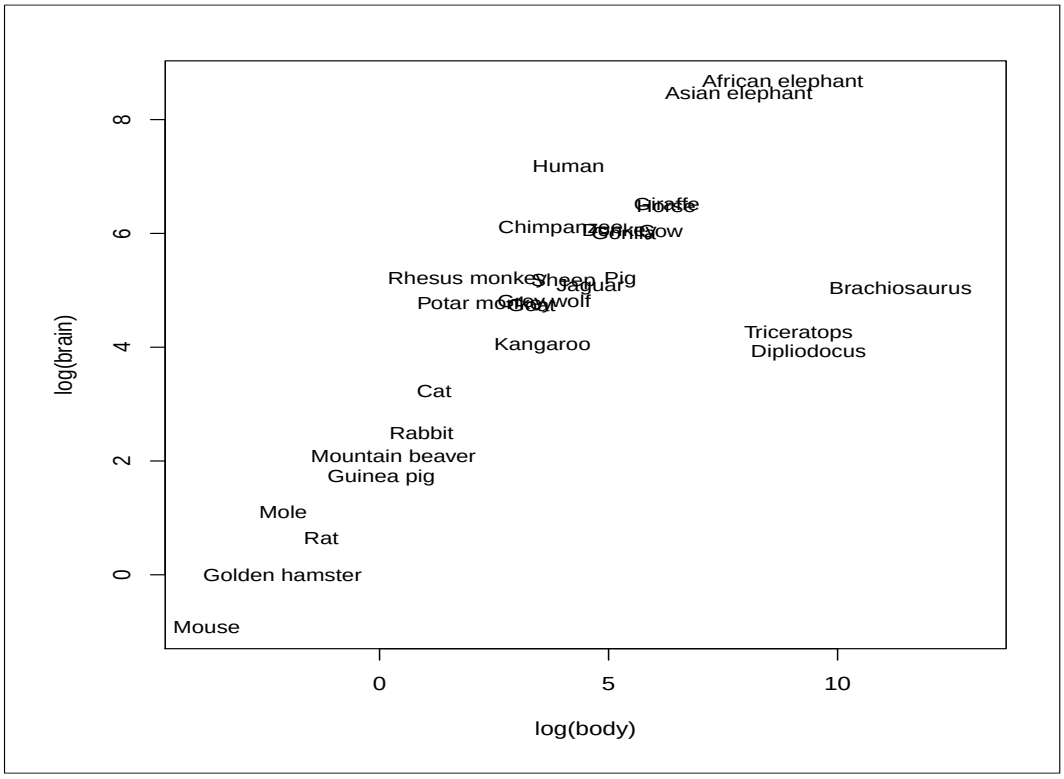

Figure 2.9: Plot for Problem 2.130 (a).

(b) The following code will produce the required plot (Figure 2.10) and calculations. There seems to be a clear linear trend, with the exception of three outliers, which are all dinosaurs. From the fit summaries we have $\log \hat{K} = 2.55490$, $\hat{r} = 0.49599$ (with dinosaurs); $\log \hat{K} = 2.15041$, $\hat{r} = 0.75226$ (without dinosaurs)

```
> # remove dinosaurs
>
> Animals2 = subset(Animals, !(rownames(Animals)
                               %in% c("Triceratops","Dipliodocus","Brachiosaurus")))
>
```

```
> # fit with and without dinosaurs
>
> fit = lm(log(brain) ~ log(body), data=Animals)
> fit2 = lm(log(brain) ~ log(body), data=Animals2)
>
> # redraw plot, with fits
>
> par(mfrow=c(1,1),pty='m')
> with(Animals,plot(log(body),log(brain),type='n',xlim=c(-4,13)))
> with(Animals,text(log(body),log(brain),rownames(Animals),cex=0.9))
> abline(fit$coef,lty=2)
> abline(fit2$coef)
> legend('topleft',legend=c('With dinosaurs','Without dinosaurs'),lty=c(2,1))
>
> # give coefficient summary
>
> summary(fit)

Call:
lm(formula = log(brain) ~ log(body), data = Animals)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2890 -0.6763  0.3316  0.8646  2.5835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.55490    0.41314   6.184 1.53e-06 ***
log(body)    0.49599    0.07817   6.345 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.532 on 26 degrees of freedom
Multiple R-squared:  0.6076,Adjusted R-squared:  0.5925
F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06

> summary(fit2)

Call:
lm(formula = log(brain) ~ log(body), data = Animals2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9125 -0.4752 -0.1557  0.1940  1.9303
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.15041    0.20060   10.72 2.03e-10 ***
log(body)    0.75226    0.04572   16.45 3.24e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7258 on 23 degrees of freedom
Multiple R-squared:  0.9217,Adjusted R-squared:  0.9183
F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14
```

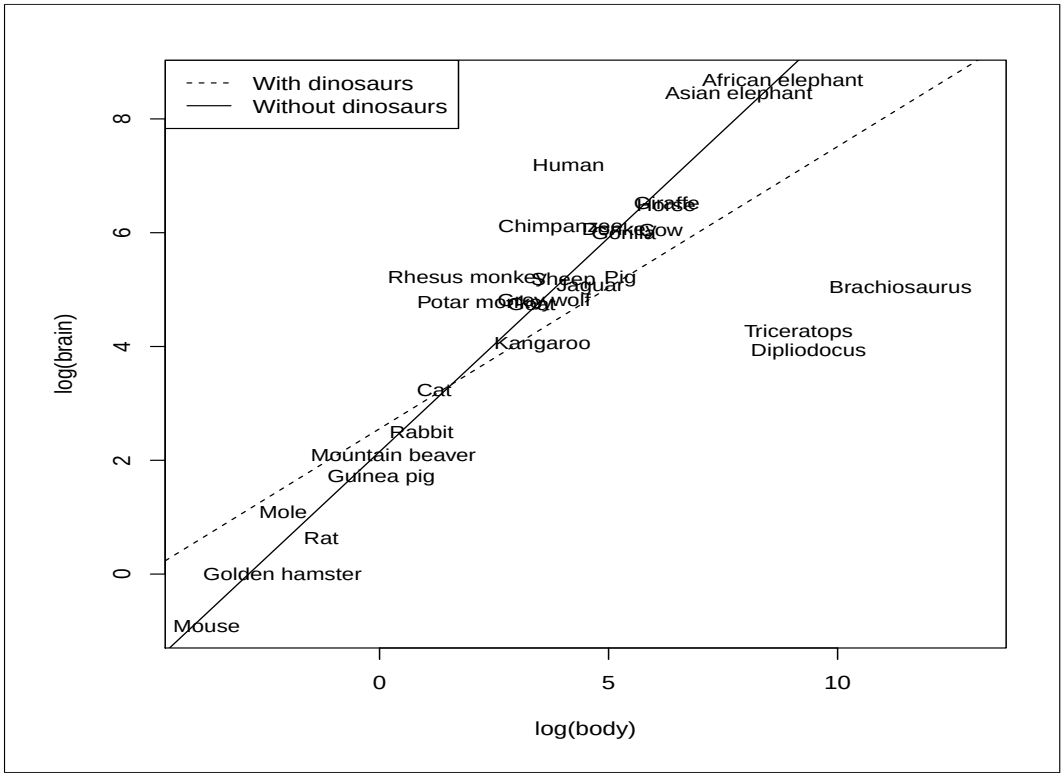

Figure 2.10: Plot for Problem 2.130 (b).

(c) The actual brain mass is Y_i, and the predicted brain mass is KX_i^r. Then

$$EQ_i = \frac{Y_i}{KX_i^r},$$

and

$$\log(EQ_i) = \log(Y_i) - \log(KX_i^r) = \log(Y_i) - [\log(K) + r\log(X_i)].$$

After we substitute estimates $\hat{\beta}_0 = \log(\hat{K})$ and $\hat{\beta}_1 = \hat{r}$ we have

$$\log(EQ_i) \approx \log(Y_i) - \left[\hat{\beta}_0 + \hat{\beta}_1 \log(X_i)\right] = e_i,$$

where e_i is the ith residual of the linear model (1). Then

$$EQ_i \approx \exp(e_i).$$

(d) A sorting of the residuals ranks the EQ_i values. `Pig` has the smallest and `Human` has the highest. From Figure 2.10 it can be seen that the dinosaurs have observed brain mass Y well below predicted brain mass KX^r. They would have the smallest EQ_i.

```
> sort(fit2$residuals)
            Pig        Kangaroo             Cow          Jaguar
-0.912462456     -0.799609067     -0.723453348     -0.558454892
Golden hamster             Rat       Guinea pig           Horse
-0.555421133     -0.550956139     -0.475168201     -0.371731840
       Rabbit          Giraffe  Mountain beaver           Mouse
-0.346496158     -0.345737520     -0.284304924     -0.228979028
      Gorilla African elephant        Grey wolf          Donkey
-0.155653772     -0.122218673     -0.069700540     -0.048100839
         Sheep             Goat              Cat   Asian elephant
-0.006993271      0.097024005      0.194039293      0.384317819
         Mole     Potar monkey       Chimpanzee    Rhesus monkey
 0.530756811      0.862375734      0.961686125      1.594948144
       Human
 1.930293870
```

2.11 Simulation Methods

Problem 2.131 The distributions of ratios of random variables are often difficult to evaluate, making formal inference methods for parameter ratios difficult to develop. The bootstrap procedure can be useful in this case.

Suppose we observe a sample of n paired observations (Y_1, Y_2). The sample is independent, but the components within each pair might not be. The respective means are μ_1, μ_2. Suppose we are interested in the solution to the equation:

$$\mu_1 + \mu_2 x = 165,$$

which is

$$x_0 = \frac{165 - \mu_1}{\mu_2}.$$

We estimate x_0 using the sample means

$$\hat{x}_0 = \frac{165 - \bar{Y}_1}{\bar{Y}_2}.$$

The problem is then to estimate the standard error $S_{\hat{x}_0}$. We will use a simulation study to investigate the accuracy of a bootstrap estimate of $S_{\hat{x}_0}$.

(a) Write a function that returns a simulated independent sample of $Y = (Y_1, Y_2)$. We first assume that Y is bivariate normal with $\mu_1 = 90$, $\mu_2 = 120$, $\sigma_1^2 = \sigma_2^2 = 10$, and $cov(Y_1, Y_2) = 2.5$. The sample size is $n = 250$. Use function `rmvnorm()` from the `mvtnorm` library.

(b) Write a function that returns the estimate $\hat{x}_0$, given an $n \times 2$ matrix of data.

(c) The first step is to estimate the true standard error by simulating the actual model (of course, in an actual inference problem, this step is not available). Simulate the model data $N_s = 10000$ times, in each case computing and storing the estimate $\hat{x}_0$. The standard deviation of these estimates is approximately $S_{\hat{x}_0}$.

(d) Next, write the bootstrap function, which accepts as input the data matrix, and N_b, the number of bootstrap replicates. The function should first compute the sample size n. Then for each replicate take the following steps:

(i) A new data set Y^* is created by sampling n observation pairs with replacement from the input data.

(ii) A new estimate $\hat{x}_0^*$ is calculated using Y^* and stored in an array.

The bootstrap estimate of $S_{\hat{x}_0}$ is the standard deviation of the N_b replicates $\hat{x}_0^*$.

(e) Finally, for a total of N_s replications, do the following steps:

(i) Simulate a sample Y using the function of part (a).

(ii) Use the function of part (d) to estimate $S_{\hat{x}_0}$.

Set $N_s = 1000$, and report the approximate true value of $S_{\hat{x}_0}$ calculated in part (c) and the median, 5th and 95th percentiles of the bootstrap estimates of $S_{\hat{x}_0}$ from part (e). Comment on the accuracy of the bootstrap procedure.

(f) Repeat part (e), this time setting (Y_1, Y_2) to be independent Poisson random variables with the same means $\mu_1 = 90$, $\mu_2 = 120$.

SOLUTION:

(a) We use the`rmvnorm()` from the `mvtnorm` library.

```
give.data = function() {

   # set up mean vector and covariance matrix

   mu = c(90,120)
   Sigma = matrix(c(10,2.5,2.5,10),nrow=2)

   # sample size n = 250

   n = 250
   y = rmvnorm(n,mu,Sigma)
   return(y)

}
```

(b) The following function returns the required estimator:

```
give.est = function(y) {
   yest = apply(y,2,mean)
   est = (165-yest[1])/yest[2]
```

```
   return(est)
}
```

(c) The following function restimates $S_{\hat{x}_0}$:

```
n.true = 10000
est = numeric(n.true)
for (i in 1:n.true) {
   y = give.data()
   est[i] = give.est(y)
}
true.se = sd(est)
```

The estimated value of $S_{\hat{x}_0}$ is:

```
> true.se
[1] 0.002190398
```

(d) The following function gives a bootstrap estimate of $S_{\hat{x}_0}$:

```
bs.fun = function(y,nb) {

   # determine sample size

   n = dim(y)[1]
   estb = numeric(nb)

   # generate nb bootstrap replicates, using function sample() to resample the data
   # with replacement

   for (i in 1:nb) {
      ind = sample(n,n,replace=T)
      yb = y[ind,]
      estb[i] = give.est(yb)
   }
   return(sd(estb))
}
```

(e) The following code implements the simulation study:

```
# use nb = 2000 bootstrap replicates for each simulated data set

nb = 2000

# Do the simulation 1000 times

n.bs = 1000
```

```
# main loop

bs.se = numeric(n.bs)
for (i in 1:n.bs) {
   y = give.data()
   bs.se[i] = bs.fun(y,nb)
}

# Summarize the bootstrap estimates

se.quant = quantile(bs.se,c(0.05,0.95))
sm = c(true.se,se.quant)
```

The (approximately) true estimate of $S_{\hat{x}_0}$ is:

```
> true.se
[1] 0.002190398
```

and the boostrap estimates are summarized by

```
> sm
                      5%         95%
0.002161288 0.002003698 0.002342916
```

The bootstrap estimates are close to $S_{\hat{x}_0} \approx 0.00219$, usually within about 8%.

(f) The function `give.data()` in part (a) can be replace by the following:

```
give.data = function() {

   # means and sample sizes

   mu = c(90,120)
   n =  250

   # generate the Poisson RVs

   y = cbind(rpois(n,lambda=mu[1]), rpois(n,lambda=mu[2]))
   return(y)
}
```

We can run the remaining code unaltered, giving summaries:

```
> true.se
[1] 0.006141413
> sm
                      5%         95%
0.006141413 0.005673434 0.006650371
```

The bootstrap estimates are close to $S_{\hat{x}_0} \approx 0.00614$, again usually within about 8%.

Problem 2.132 We are given the following two samples, each of size $n = 10$.

```
x = c(7.3, 6.3, 5.0, 5.5, 5.2, 5.3, 3.9, 5.0, 4.0, 6.4)
y = c(6.4, 7.6, 7.1, 6.1, 7.0, 7.5, 7.2, 6.1, 7.8, 6.4)
```

(a) Perform a two sided t-test for equality of means. Report the p-value.
(b) Perform a two sided t-test for equality of means.
  (i) For a single replication, randomly reassigning the pooled data to two groups of size $n = 10$, then recalculating the t-test.
  (ii) Do 1000 such replications, for each one storing both the t-statistic and the p-value.
  (iii) Add the original t-statistic T_{obs} to the replicated sample.
  (iv) The empirical p-value is the proportion of replicate t-statistics T^* satisfying $|T^*| \geq |T_{obs}|$.
(c) Compare the empirical p-calue to the theoretical p-value obtained in Part (a).
(d) Why would we add the original t-statistic T_{obs} to the replicated sample? Using this procedure, what is the minimum possible empirical p-value?
(e) The empirical ciritical value can be taken to be the 95th percentile of the replicated absolute values $|T^*|$. Compare the empirical critical value to the theoretical critical value used in the conventional t-test. For this calculation do not include T_{obs} in the replicated sample.
(f) Plot a histogram of the replicated p-values (do not include the original p-value from Part (a)). Is the shape of the histogram what you would expect?

SOLUTION: The following code may be used to complete the problem.

```
>
> ### Load data
>
> x = c(7.3, 6.3, 5.0, 5.5, 5.2, 5.3, 3.9, 5.0, 4.0, 6.4)
> y = c(6.4, 7.6, 7.1, 6.1, 7.0, 7.5, 7.2, 6.1, 7.8, 6.4)
>
>
> ### Store original test result
>
> true.test = t.test(x,y)
>
> ### repeat the test 1000 on randomly permuted data
>
> nsim = 1000
> pvec = rep(NA,nsim)
> tvec = rep(NA,nsim)
> for (i in 1:1000) {
+
+   xy = sample(c(x,y))
+   xperm = xy[1:10]
```

```
+   yperm = xy[11:20]
+
+   junk = t.test(xperm, yperm)
+
+   pvec[i] = junk$p.value
+   tvec[i] = junk$statistic
+
+ }
>
> ### add the original test statistic to the replications
>
> tvec.plus = c(tvec,true.test$statistic)
>
> ### Report theoretical and empirical p-values
>
> true.test$p.value
[1] 0.001362216
> mean(abs(tvec.plus) >= abs(true.test$statistic))
[1] 0.001998002
>
> ### Compare the theoretical and empirical cricital values
>
> t.crit = qt(0.975,df=18)
> t.crit.perm = quantile(abs(tvec),0.95)
> c(t.crit, t.crit.perm)
              95%
2.100922 2.079977
>
> ### Plot a hisogram of the replicated p-values
>
> pdf('figTperm.pdf')
> hist(pvec)
> dev.off()
RStudioGD
        2
>
```

(a) From the above code, the theoretical and empirical p-values are, respectively, $P = 0.001362216$ and $P = 0.001998002$.

(b) In general, statistical procedures should be *conservative*, in the sense that when they cannot be exact they should underestimate rather than overestimare statistical significance. If the observed statistic is not included with the replications, then the empirical p-value could be zero. If it is included, the empirical p-value will be at least $1/(N+1)$, where N is the number of replications.

(c) From the above code, the theoretical and empirical critical values are, respectively, $t = 2.100922$ and $t = 2.079977$. The theoretical critical value is the 0.975 critical value from a t-distribution with 18

degrees of freedom.

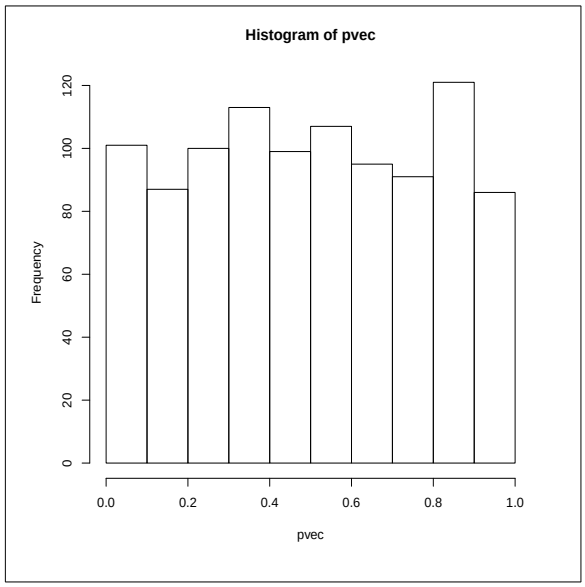

Figure 2.11: Plot for Problem 2.132 (d)

(d) The histogram is given in Figure 2.11, and appears to sampled from a uniform distribution. In general, p-values are approximately uniformly distributed under a null hypothesis. In effect, the permutation procedure creates an empirical null hypothesis, so the histrogram if what we would expect.

Problem 2.133 We wish to investigate the accuracy of the t-distribution approximation for the transformed correlation of Section 21.1 when applied to the Spearman rank correlation coefficient.

(a) The `R` function `sample(n)` outputs a random permutation of the numbers $1, 2, \ldots, n$. How can this be used to simulate the Spearman rank correlation coefficient when the true correlation is $\rho = 0$?
(b) Simulate a random sample (of size 10,000) of Spearman rank correlation coefficients for $n = 25$ under the null hypothesis $H_o : \rho = 0$.
(c) Calculate the t-distribution transformation for each simulated value (Equation (21.1) of Section 21.1).
(d) Plot a histogram of the transformed sample. Superimpose a t-density on the same plot, using the approriate degrees of freedom. Is the t-distribution an accurate approximation? [Make sure you use the `probability = T` option when you plot the histogram].

SOLUTION:

(a) Suppose we have output `x = sample(25)`. Then set `y = 1:25`. Calculate the correlation of `x` and `y` (you can use either the `method = 'spearman'` or the default `method = 'pearson'` method).
(b) The following code creates the sample:

```
f0 = function(i) {
x = sample(25)
y = 1:25
```

```
return(cor(x,y,method='spearman'))
}
cor.sample = sapply(1:10000,f0)
```

(c) The following function calculates the transformation:

```
f.transform = function(x) {x/sqrt((1-x^2)/23)}
```

(d) The following code creates the plot (Figure 2.12):

```
t.sample = f.transform(cor.sample)
tgrid = seq(-3,3,0.01)
hist(f.transform(cor.sample),probability=T,nclass=15)
lines(tgrid,dt(tgrid,df=23))
```

The t-distribution (with 23 degrees of freedom) accurately models the distribution.

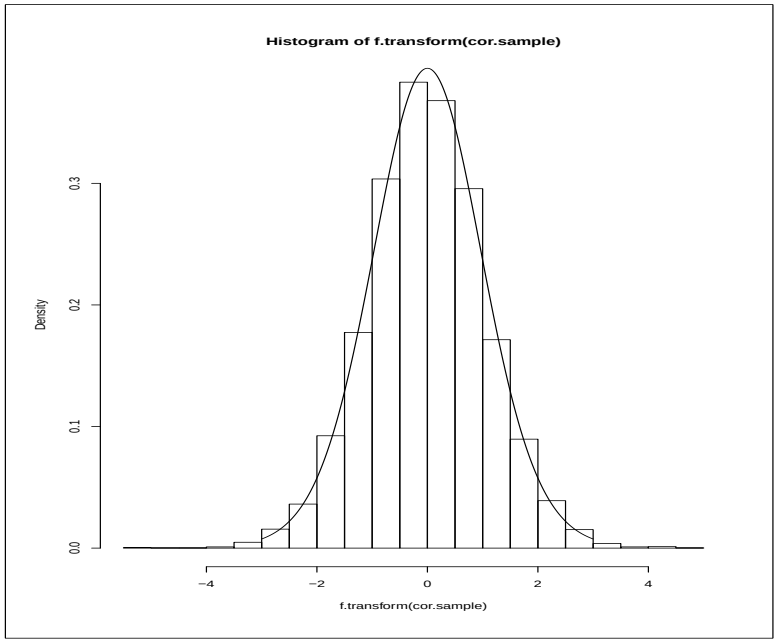

Figure 2.12: Plot for Problem 2.133 (d).

2.12 ROC Curves

Problem 2.134 A biomarker risk score for two-year cancer recurrence is assessed on 10 subjects, with the outcomes given in the following table. The object is to determine a threshold T, so that a more aggressive treatment will be recommended if the risk score exceeds T. It is required that the sensitivity of the test be at least 80%. Based on the available data, what value would you recommend for T?

Risk Score	1 = Recurrence
2.5	0
5.5	1
6.6	0
7.3	0
13.2	0
16.6	1
21.6	1
22.4	0
24.0	1
56.4	1

SOLUTION: We can take each available risk score as a threshold value. For each threshold value we can calculate sensitivity and specificity. Note that at threshold T we have

$$spec = \text{Number of 0's} < T \ / \ \text{Number of 0's}$$
$$sens = \text{Number of 1's} \geq T \ / \ \text{Number of 1's}$$

so the table is easily constructed. If we require $spec \geq 0.8$ we can obtain maximum $sens = 0.8$ at $T = 16.6$ from the resulting table.

Risk Score	1 = Recurrence	N 0's $< T$	$spec$	N 1's $\geq T$	$sens$
2.5	0	0	0.0	5	1.0
5.5	1	1	0.2	5	1.0
6.6	0	1	0.2	4	0.8
7.3	0	2	0.4	4	0.8
13.2	0	3	0.6	4	0.8
16.6	1	4	0.8	4	0.8
21.6	1	4	0.8	3	0.6
22.4	0	4	0.8	3	0.6
24.0	1	5	1.0	2	0.4
56.4	1	5	1.0	1	0.2

Problem 2.135 A biomarker designed to predict 3 year mortality in melanoma patients yielded the ROC curve of Figure 2.13. A new biomarker was assessed, yielding the outcomes (predicted vs. actual mortality) in the table given below. Is it possible to conclude whether or not the new biomarker is more accurate? If so, how do they compare?

	Predicted Mortality	Predicted Survival
Subject did not survive	120	29
Subject survived	26	106

SOLUTION: From the table we have true positive rate $TP = sens = 120/149 \approx 0.805$ and false positive rate $FP = 1 - spec = 12/132 \approx 0.245$. If we locate FP on the horizontal axis of 2.13, we find that this corresponds to $TP \approx 0.62$ for the original biomarker. We may conclude that the new biomarker is more accurate.

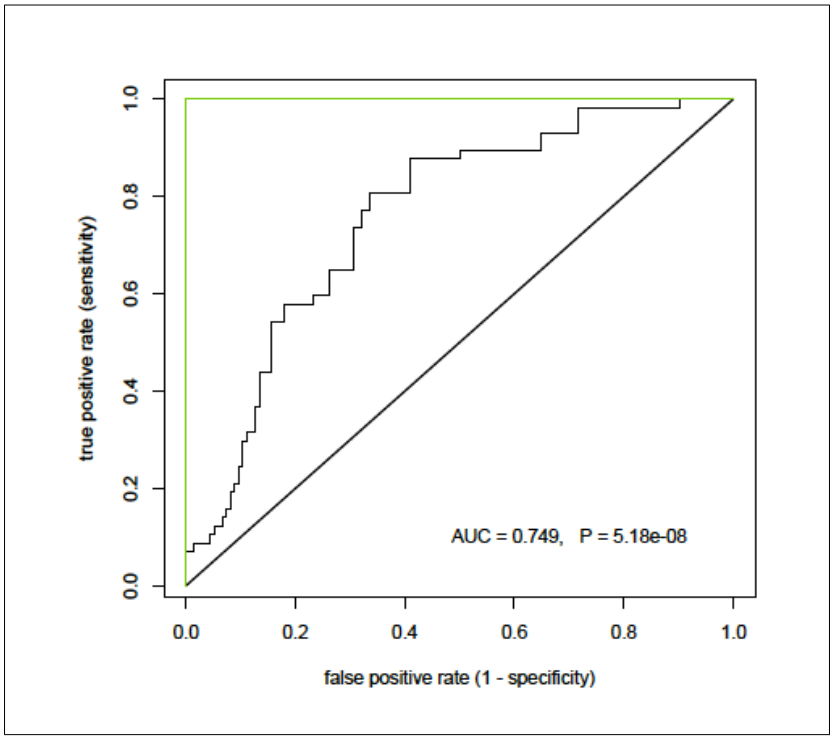

Figure 2.13: ROC curve for Problem 2.135.

Problem 2.136 A biomarker risk score for a certain disease is tested on 5 disease +ve and 5 disease -ve subjects, with the results given in the following table.

Risk Score	Outcome
0.57	-ve
1.69	-ve
2.88	-ve
5.38	-ve
8.75	-ve
7.22	+ve
11.83	+ve
24.91	+ve
28.27	+ve
31.92	+ve

(a) What will be the AUC statistic for the ROC curve?
(b) What is the p-value for testing against the null hypothesis $H_o : AUC = 0.5$?

SOLUTION:

(a) The AUC may be calculated using the following equation:

$$AUC = \frac{\sum_{i\in\text{-ve}} \sum_{j\in+\text{ve}} I\{score_j > score_i\} + 0.5 \times I\{score_j = score_i\}}{n_- \times n_+}.$$

Note that there are no ties. The number of negatives and positives are $n_- = n_+ = 5$. The denominator of the AUC statistic is $n_- \times n_+ = 2$. The numerator is the number of pairs of risk scores, one taken from each outcome, for which the risk score for the positive outcome is larger. There is only one such pair (8.75, 7.22) for which this condition does not hold, therefore $AUC = 24/25$.

(b) This test is equivalent to the Wilcoxon two sample rank sum test, with the two samples defined by the binary outcome. The following code calculates a p-value of $P = 0.01587$.

```
> x = c(0.57,1.69,2.88,5.38,8.75,7.22,11.83,24.91,28.27,31.92)
> y = rep(c(0,1),each=5)
>
> wilcox.test(x ~ y)

Wilcoxon rank sum exact test

data:  x by y
W = 1, p-value = 0.01587
alternative hypothesis: true location shift is not equal to 0
```

Problem 2.137 This problem will make use of the `biopsy` data set from the `MASS` library. From the help file:

> This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known. There are 699 rows and 11 columns.

The data contains features labeled `V1`, ..., `V9` and a response labeled `class` with binary tumor outcomes `benign` and `malignant`. We will use variable `V1` (clump thickness) to predict `class`. Set up binary response `Y` = 1 (malignant) or `Y` = 0 (benign). We will make use of the `ROCR` package to evaluate this predictor.

(a) Verify that the false positive rate is equal to 1 - specificity.
(b) Determine the AUC statistic for this predictor.
(c) Create an ROC curve.
(d) Extract the data used to plot the ROC curve, and superimpose these points on the ROC curve.
(e) What is the p-value for testing against the null hypothesis $H_o : AUC = 0.5$?

SOLUTION: The following code may be used for Parts (b)-(e). Note that `ROCR` produces `S4` objects.

```
library(ROCR)

library(MASS)

data(biopsy)

# Binary response y = 1 (malignant) or y = 0 (benign)
```

```
y = as.numeric(biopsy$class == "malignant")

# Use  clump thickness as predictor

x = biopsy$V1

# Set up ROCR prediction object

pred = prediction( biopsy$V1,  biopsy$class )

### To get AUC we can do this:

perf = performance(pred, measure = "auc")
perf@y.values[[1]]

### To set up ROC curve use the performance() function this way

perf = performance(pred,"tpr","fpr", measure = "auc")
plot(perf.roc.curve)
abline(0,1,lty=2)

### To obtain thetdata used to plot the curve, do this:

x.roc = perf.roc.curve@x.values[[1]]
y.roc = perf.roc.curve@y.values[[1]]
points(x.roc,y.roc,pch=19)

### To get p-value use Wilcoxon rank sum test

wilcox.test(x ~ y)
```

(a) Let FN, FP, TN, TP be the false negative, false positive, true negative and true positive frequencies. Then the false positive rate is

$$fpr = \frac{FP}{FP + TN} = 1 - \frac{TN}{FP + TN} = 1 - spec$$

where *spec* is specificity.

(b) We get $AUC = 0.9098416$.

```
>
> ### To get AUC we can do this:
>
> perf = performance(pred, measure = "auc")
> perf@y.values[[1]]
[1] 0.9098416
>
```

(c) See Figure 2.14.
(d) See Figure 2.14.
(e) We get $P < 2.2 \times 10^{-16}$ (ie. a very small p-value).

```
### To get p-value use Wilcoxon rank sum test

> wilcox.test(x ~ y)

Wilcoxon rank sum exact test

data:  x by y
W = 1, p-value = 0.01587
alternative hypothesis: true location shift is not equal to 0
```

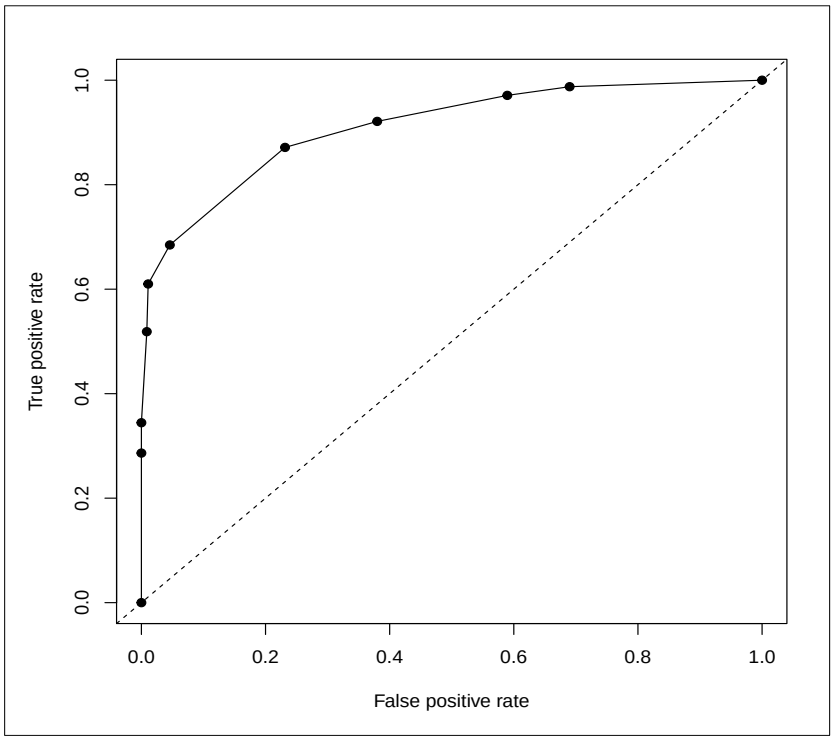

Figure 2.14: ROC curve for Problem 2.137 (c)-(d).

Chapter 3

Data Projects

The problems in this chapter are intended to provide experience in statistical modelling using ANOVA and linear regression. They deal with a number of the more practical aspects of data analysis, and emphasize some of the problems which may arise. For example, Problem 3.5 demonstrates the importance of validating distributional assumptions, and makes use of the bootstrap procedure, Box-Cox transformations, and multiple comparison methods, while Problem 3.6 provides an exercise in model selection.

The data sets can be found in one of two `R` packages which contain libraries of data. The `MASS` library is included in the standard `R` distribution, and contains functions and datasets to support *Modern Applied Statistics with S, 4th ed* by Venables and Ripley (Springer). A list of the datasets can be viewed from the `R` command line with the following commands:

```
> library(MASS)
> help(package='MASS')
```

The `ISLR` package accompanies *An Introduction to Statistical Learning: with Applications in R* by James, Witten, Hastie and Tibshirani (Springer). This package can be installed from the `CRAN` repository directly from `R` with the command

```
> install.packages("ISLR")
```

then the contents can be examined with the commands

```
> library(ISLR)
> help(package='ISLR')
```

Problem 3.1 This question will make use of the `Cars93` data set from the `MASS` library. To access and view this data set, use the commands:

```
> library(MASS)
> class(Cars93)
> head(Cars93)
> help(Cars93)
```

This data set contains various specifications of 93 car models. We will be interested in two of these: `MPG.city` and `Passengers`.

We first consider various transformation methods.

(a) Let $Y = (Y_1, \ldots, Y_n)$ be the observations in `MPG.city`. Create a histogram of Y, and comment on the skewness of the distribution.

(b) Create a function which implements the empirical rule. This function should input a single vector containing a sample, and output a 3×2 matrix summary. Rows 1 to 3 correspond to $k = 1, 2, 3$ standard deviations. The first column should contain the proportion of the data within k standard deviations of the mean, and the second column should contain the theoretical proportion for a normal distribution. The rows and columns should be suitably labelled. Test your function using a simulated random sample from a normal, exponential and uniform distribution. Use $n = 10000$ for each. Note that the choice of parameters will not make a difference, so just use the default values.

(c) Assuming right-skewness was detected in Y, we might consider a log-transformation of Y. However, the properties of the log-transformation (in particular, its ability to induce symmetry) can be affected by any data offset. So we might, more generally, consider the transformation $\log(Y_i + a)$ where a is some constant. The log-transformation can then be standardized by, for example, selecting $a = -\min(Y) + 1$, so that the lower bound of the transformed data $\log(Y_i - \min(Y) + 1)$ will be zero. To investigate this, consider the original data Y, and two transformations:

$$\begin{aligned} Y' &= \log(Y) \\ Y'' &= \log\left(Y - \min(Y) + 1\right) \end{aligned}$$

On a single plot window, plot a histogram and normal quantile plot for Y, Y' and Y'' (use a 3×2 plot grid, using one row for each version of the data). Then apply your empirical rule function to each version of the data. Is at least one version of the transformation able to induce an approximate normal distribution?

SOLUTION:

(a) The following code can be used. See Figure 3.1. The data is notably right-skewed.

```
car2 = subset(Cars93)
y = car2$MPG.city
par(mfrow=c(1,1))
hist(y,xlab = 'MPG (city)')
```

(b) The following code can be used.

```
emp.rule = function(y) {
z = (y-mean(y))/sd(y)
er.tab = NULL
for (i in 1:3) {er.tab = rbind(er.tab,  c(mean(abs(z) <= i), 1-2*pnorm(-i)))}
dimnames(er.tab) = list(paste(1:3,'SD'),c('Emp','Theor'))
return(er.tab)
}
```

Applying the function to the normal, exponential and uniform distributions gives the following tables. The empirical frequencies for the simulated normal sample are very close to the theoretical frequencies, but this does not hold for the remaining distributions.

```
> emp.rule(rnorm(10000))
Emp     Theor
1 SD 0.6818 0.6826895
2 SD 0.9564 0.9544997
3 SD 0.9975 0.9973002
> emp.rule(rexp(10000))
Emp     Theor
1 SD 0.8652 0.6826895
2 SD 0.9502 0.9544997
3 SD 0.9822 0.9973002
> emp.rule(runif(10000))
Emp     Theor
1 SD 0.5764 0.6826895
2 SD 1.0000 0.9544997
3 SD 1.0000 0.9973002
>
```

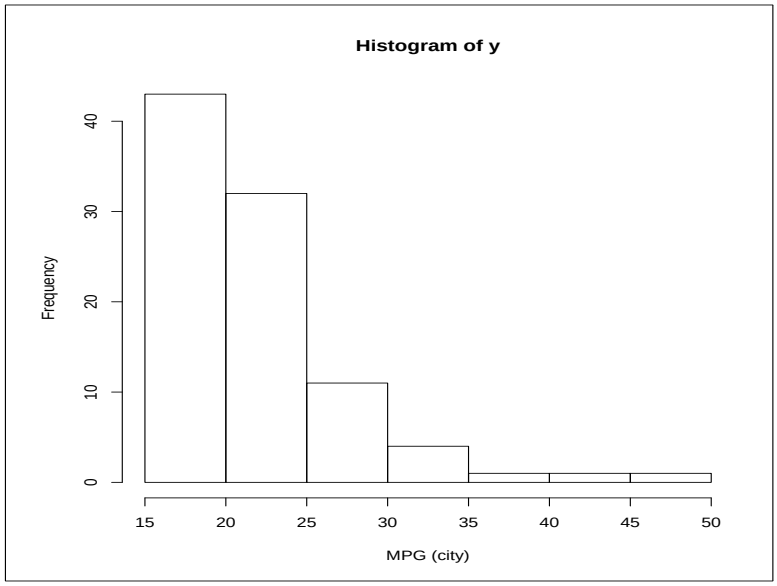

Figure 3.1: Histogram for Problem 3.1 (a).

(c) The following code can be used to construct the plots. See Figure 3.2.

```
# create list of the three data sets

y.list = list(y0 =y, y1  = log(y), y2 = log(y-min(y)+1))
y.labs = c('Untransformed','Transform 1','Transform 2')

# apply plotting routines to each data set
```

```
par(mfrow=c(3,2))
for (i in 1:3) {
ytemp = y.list[[i]]
hist(ytemp,main=y.labs[i])
qqnorm(ytemp)
qqline(ytemp)
}
```

Apply the empirical rule.

```
# Apply emp.rule to each data set, using the lapply() function.

> lapply(y.list, emp.rule)
$y0
Emp     Theor
1 SD 0.7741935 0.6826895
2 SD 0.9677419 0.9544997
3 SD 0.9784946 0.9973002

$y1
Emp     Theor
1 SD 0.6881720 0.6826895
2 SD 0.9677419 0.9544997
3 SD 0.9892473 0.9973002

$y2
Emp     Theor
1 SD 0.6881720 0.6826895
2 SD 0.9569892 0.9544997
3 SD 1.0000000 0.9973002
```

While the $\log(Y)$ transform is able to reduce skewness, the offset log transformation $\log(Y - \min(Y) + 1)$ is better able to linearize the normal quantile plot. See Figure 3.2. In addition, the offset log transformation Y'' conforms more closely to the empirical rule, especially for the $k = 2$ and 3 standard deviation components.

Problem 3.2 This question will make use of the `Cars93` data set from the `MASS` library. To access and view this data set, use the commands:

```
> library(MASS)
> class(Cars93)
> head(Cars93)
> help(Cars93)
```

This data set contains various specifications of 93 car models. We will be interested in two of these: `MPG.city` and `Passengers`.

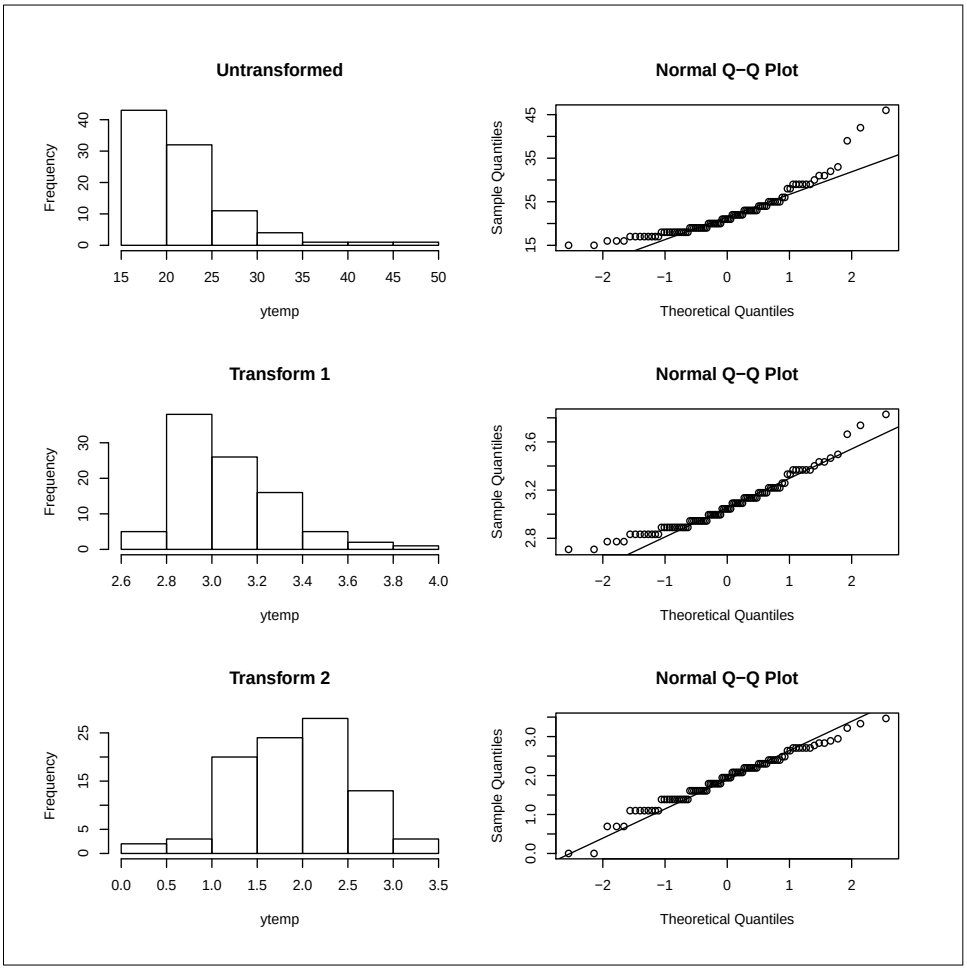

Figure 3.2: Plots for Problem 3.1 (c).

We consider the possibility that lack of normality can be caused by pooling heterogeneous sources of data.

(a) To prepare the data, use the `subset()` function to create a new data frame containing only cars with passenger capacity `Passengers` equal to 4,5,6 or 7. Create side-by-side boxplots of MPG rating Y, grouped by `Passengers`, and include in the same plot a single boxplot for the complete data Y. Comment on what you see.

(b) We might expect `MPG.city` to be negatively correlated with passenger capacity (that is, larger cars tend to have lower MPG ratings). So, we can normalize Y using means and standard deviations which are allowed to depend on the variables `Passengers`. Let $\bar{X}_m$ and S_m^2 be the sample mean and variance of Y restricted to the condition `Passengers` $= m$, $m = 4, 5, 6, 7$. Then for each observation Y_i create an adjusted Z-score

$$Z_i = \frac{Y_i - \bar{X}_m}{S_m},$$

where m is the value of `Passengers` for observation Y_i. Then plot a histogram and normal quantile

plot for the transformed values $Z = (Z_1, \ldots, Z_n)$. Also, apply your empirical rule functions. Are the values Z approximately normal?

SOLUTION:

(a) The following code can be used. See Figure 3.3. The variance within car class is lower than the pooled data, with the exception of 4 passenger cars. The within-class distributions appear symmetric, again, with the possible exception of 4 passenger cars.

```
# create data subset consisting of passenger capacities 4 to 7 inclusive.

car2 = subset(Cars93, Passengers %in% c(4,5,6,7))
y = car2$MPG.city #*car2$Passengers

# split the data by passenger capacity, then append the complete data.
# rename list accordingly.

junk = split(y,car2$Passengers)
junk[[5]] = y
names(junk) = c(names(junk)[1:4],'All')

# then draw boxplot

par(mfrow=c(1,1))
boxplot(junk)
```

(b) The following code can be used. See Figure 3.4. The histogram is approximately symmetric and the normal quantile plot is approximately linear. Both plots conform to the normal distribution.

```
# create data set of z-scores, normalized using
# class specific means and variances.

junk = split(y,car2$Passengers)
z = NULL
for (i in 1:4) {
yyy = junk[[i]]
z = c(z,(yyy-mean(yyy))/sd(yyy))
}

# create histogram and normal quantile plot

par(mfrow=c(1,1))
par(mfrow=c(1,2),pty='s')
hist(z)
qqnorm(z)
qqline(z)
```

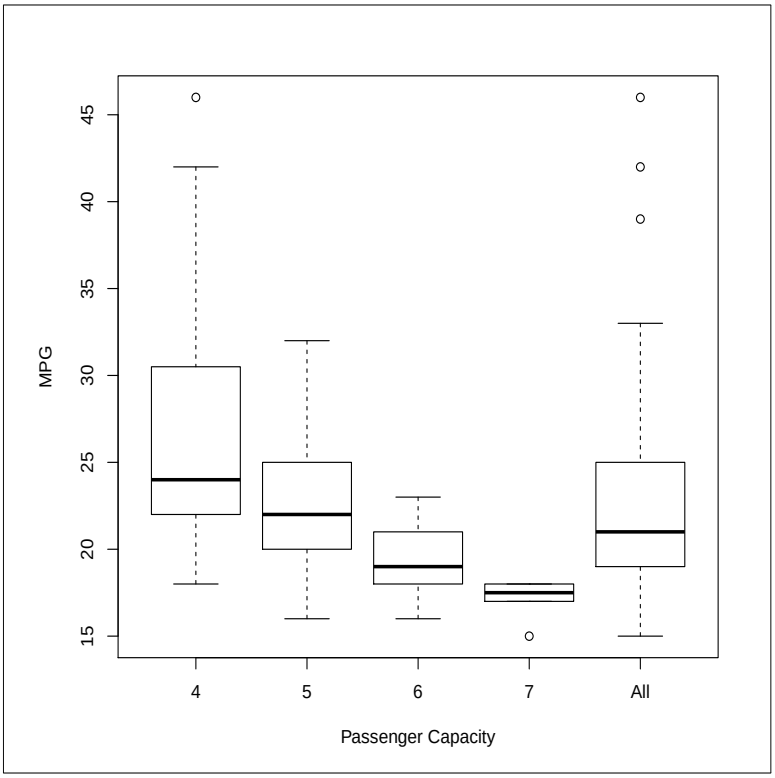

Figure 3.3: Plot for Problem 3.2 (a)

Then, apply the empirical rule. The Z-scores conform to the normal distribution quite closely, without the need for a log-transformation.

```
# apply empirical rule

> emp.rule(z)
Emp     Theor
1 SD 0.6777778 0.6826895
2 SD 0.9555556 0.9544997
3 SD 1.0000000 0.9973002
```

Problem 3.3 This question will make use of the `Cars93` data set from the `MASS` library. To access and view this data set, use the commands:

```
> library(MASS)
> class(Cars93)
> head(Cars93)
> help(Cars93)
```

This data set contains various specifications of 93 car models. We will be interested in two of these: `MPG.city` and `Passengers`.

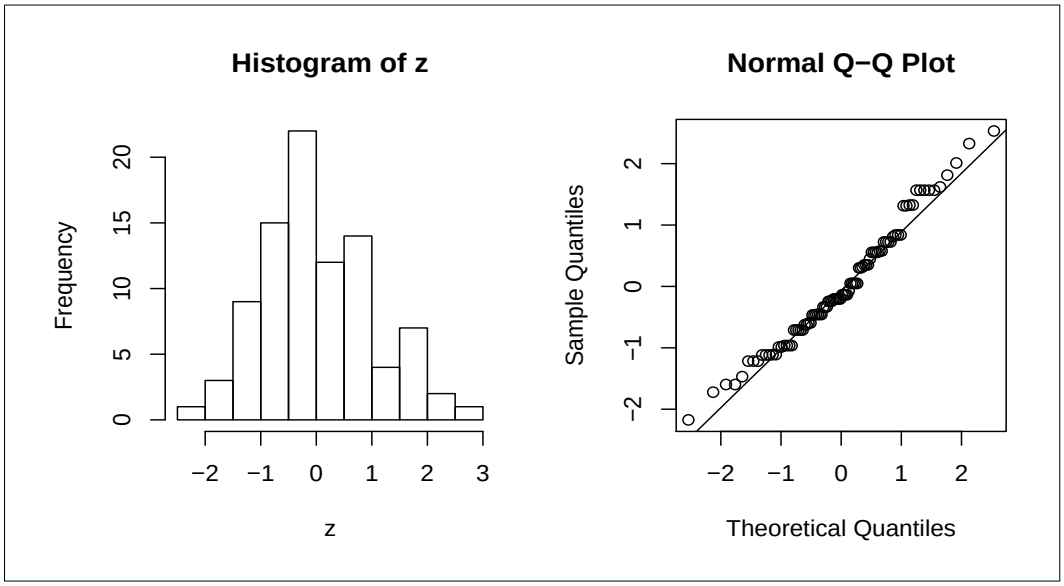

Figure 3.4: Plot for Problem 3.2 (b)

This question considers how to construct and interpret confidence intervals.

(a) Create a function which constructs a level $1-\alpha$ confidence interval for a mean μ given a random sample from a normal $N(\mu, \sigma^2)$ distribution. The function should accept a vector containing the data and α. Note that σ^2 should be assumed unknown.

(b) Apply the function of Part (a) to MPG observations Y separately for each value of `Passengers` $= m$, for $m = 4, 5, 6, 7$. Use confidence level 95%. Do the confidence intervals for `Passengers` $= 4$ and 5 overlap (that is, are there values of μ contained in both)?

(c) It can be shown that if a level $1-\alpha$ confidence interval for a mean doesn't contain 0, then a two-sided hypothesis test for null hypothesis $H_o : \mu = 0$ would reject H_o with an observed significance smaller than α. However, the situation for a difference in means is more complicated. Suppose we are given standard confidence intervals

$$\begin{aligned} \bar{X}_1 &\pm t_{n_1-1,\alpha/2}\frac{S_1}{n_1} \\ \bar{X}_2 &\pm t_{n_2-1,\alpha/2}\frac{S_2}{n_2}, \end{aligned}$$

for means μ_1, μ_2 respectively. Assume the samples used for each confidence interval are independent of each other. Then suppose the lower bound of the first confidence interval is larger than the upper bound of the second, so that the two do not overlap. Is this equivalent to rejecting the null hypothesis $H_o : \mu_1 = \mu_2$ based on the conventional T-statistic:

$$T = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

assuming unequal variances? Why or why not?

(d) Do a formal two-sided T-test against null hypothesis $H_o : \mu_4 = \mu_5$, where μ_m is the mean MPG for cars with passenger capacity m. Assume that variances are not equal. Can we reject H_o with a $\alpha = 0.05$ significance level? Does this contradict the results of Part (b)?

SOLUTION:

(a) The following code can be used to create the required function.

```
tci = function(y,alpha) {

# estimate
est = mean(y)

# sample size
n = length(y)

# calculate margin of error
me = qt(1-alpha/2,df=n-1)*sd(y)/sqrt(n)

# create output vector
ans = c(est,me,est+me*c(-1,1),n)
names(ans) = c('est','m.e.','lb','ub','n')
return(ans)

}
```

(b) Apply the confidence interval for separate car classes using the `by()` function. For 4 and 5 passenger cars, the CIs are $(23.254992, 29.875443)$, and $(21.558712, 24.051044)$, so MPG values between 23.254992 and 24.051044 are contained in both.

```
> ci.list = by(y,car2$Passengers, function(x) {tci(x,0.05)})
> ci.list
car2$Passengers: 4
est      m.e.        lb        ub         n
26.565217  3.310225 23.254992 29.875443 23.000000
------------------------------------------------
car2$Passengers: 5
est      m.e.        lb        ub         n
22.804878  1.246166 21.558712 24.051044 41.000000
------------------------------------------------
car2$Passengers: 6
est      m.e.        lb        ub         n
19.277778  1.016341 18.261437 20.294119 18.000000
------------------------------------------------
car2$Passengers: 7
est       m.e.         lb         ub          n
17.2500000  0.8439111 16.4060889 18.0939111  8.0000000
>
```

(c) Suppose the lower bound for the first CI is larger than the upper bound of the second CI. This can be written

$$\bar{X}_1 - t_{n_1-1,\alpha/2}\frac{S_1}{n_1} > \bar{X}_2 + t_{n_2-1,\alpha/2}\frac{S_2}{n_2},$$

or equivalently.

$$\bar{X}_1 - \bar{X}_2 > t_{n_1-1,\alpha/2}\frac{S_1}{n_1} + t_{n_2-1,\alpha/2}\frac{S_2}{n_2}.$$

In effect, the test statistic is now

$$T^* = \frac{\bar{X}_2 - \bar{X}_1}{t_{n_1-1,\alpha/2}\left[\frac{S_1}{n_1}\right] + t_{n_2-1,\alpha/2}\left[\frac{S_2}{n_2}\right]},$$

which is not the same test statistic as

$$T = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Therefore, rejecting or accepting equality of means based on overlapping confidence intervals is not equivalent to a T-test. But see Sections 23.3-23.4.

(d) The following code implements the required T-test. We can reject $H_o : \mu_4 = \mu_5$ at a significance level $\alpha = 0.05$, since $P = 0.0366 < \alpha$. This does not contradict the results of Part (b), since, as shown in Part (c), the T-test cannot be implemented by comparing confidence intervals.

```
> junk = split(y,car2$Passengers)
> t.test(junk$‘4‘,junk$‘5‘)

Welch Two Sample t-test

data:  junk$‘4‘ and junk$‘5‘
t = 2.1926, df = 28.68, p-value = 0.0366
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.2510077 7.2696710
sample estimates:
mean of x mean of y
26.56522  22.80488
```

Problem 3.4 This question will make use of the `Cars93` data set from the `MASS` library. To access and view this data set, use the commands:

```
> library(MASS)
> class(Cars93)
> head(Cars93)
> help(Cars93)
```

This data set contains various specifications of 93 car models. We will be interested in two of these: `MPG.city` and `Passengers`.

While it seems clear that cars with larger passenger capacity have lower MPG, we may also consider fuel efficiency on a per-passenger basis. A 6 passenger car might have a rating of 19.3 MPG. Thus, one mile of travel consumes $1/19.3$ gallons. However, if we assume for the moment that the vehicle is used to full passenger capacity, then we can say that each *passenger* consumes $1/(6 \times 19.3)$ gallons. We can therefore define a *passenger-mile per gallon* rating as

$$PMPG = N_P \times MPG,$$

where N_P is the number of passengers.

(a) Suppose we are given the following parameters and summary statistics for MPG ratings for $m = 4$ and $m = 5$ passenger cars:

	MPG	
	4 Passengers	5 Passengers
Population mean	μ_4	μ_5
Population variance	σ_4^2	σ_5^2
Sample size	n_4	n_5
Sample mean	$\bar{X}_4$	$\bar{X}_5$
Sample variance	S_4^2	S_5^2

Suppose we wish to construct a two-sided hypothesis test with a null hypothesis that the PMPG rating is the same for each class of car. Give the null and alternative hypotheses, as well as the appropriate T-statistic (assuming unequal variances). Use the quantities given in the preceding table.

(b) Carry out the T-test of Part (a) of this question. You can do this by creating the PMPG ratings directly as the product of the MPG ratings and the passenger capacities, then applying the `t.test()` function. Use an $\alpha = 0.05$ significance level, and assume unequal variances. How does your result compare to Part (d) of Question 3?

(c) Create side-by-side boxplots of the PMPG ratings, grouped by `Passengers`. As in Part (b) of Question 3, create level 95% confidence intervals for PMPG separately for each value of `Passengers` $= m$, $m = 4, 5, 6, 7$. Superimpose these directly on the boxplots. Does there seem to be a big difference in PMPG between different car classes? (**HINT:** The location of the ith boxplot on the horizontal axis is i.)

SOLUTION: After the transformation

$$PMPG = N_P \times MPG,$$

the summary statistics are, in terms of the original table,

	PMPG	
	4 Passengers	5 Passengers
Population mean	$4\mu_4$	$5\mu_5$
Population variance	$16\sigma_4^2$	$25\sigma_5^2$
Sample size	n_4	n_5
Sample mean	$4\bar{X}_4$	$5\bar{X}_5$
Sample variance	$16S_4^2$	$25S_5^2$

(a) The hypotheses are

$$H_o : 4\mu_4 = 5\mu_5, \quad \text{versus} , \quad H_a : 4\mu_4 \neq 5\mu_5.$$

The T-statistics is now

$$T = \frac{5\bar{X}_5 - 4\bar{X}_4}{\sqrt{\frac{16S_4^2}{n_4} + \frac{25S_5^2}{n_5}}}.$$

The rejection region is otherwise defined in the same way.

(b) The test can be carried out in the same way as in Part (d) of Question 3. The only difference is that the data is first transformed by the formula `y = car2$MPG.city*car2$Passengers`. In this case, the P-value is now $P = 0.2868$, so we do not reject the null hypothesis of equality of PMPG between 4 and 5 passenger cars, at significance level $\alpha = 0.05$. In Part (d) of Question 3, the equality of PMPG between 4 and 5 passenger cars *was* rejected at significance level $\alpha = 0.05$.

```
> car2 = subset(Cars93, Passengers %in% c(4,5,6,7))
> y = car2$MPG.city*car2$Passengers
> junk = split(y,car2$Passengers)
> t.test(junk$‘4‘,junk$‘5‘)

Welch Two Sample t-test

data:  junk$‘4‘ and junk$‘5‘
t = -1.0926, df = 32.447, p-value = 0.2826
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-22.229129   6.702088
sample estimates:
mean of x mean of y
106.2609  114.0244
```

(c) The following code can be used. See Figure 3.5. In general, there does not appear to be evident differences in PMPG between cars of differing passenger capacity.

```
# create CIs

ci.list = by(y,car2$Passengers, function(x) {tci(x,0.05)})

# add CIs graphically

par(mfrow=c(1,1))
boxplot(y~car2$Passengers)
for (iii in 1:4) {
ci = ci.list[[iii]]
lines(c(iii,iii),ci[1]+ci[2]*c(-1,1),type='l',col='blue')
points(c(iii,iii),ci[1]+ci[2]*c(-1,1),pch=20,col='blue')
}
```

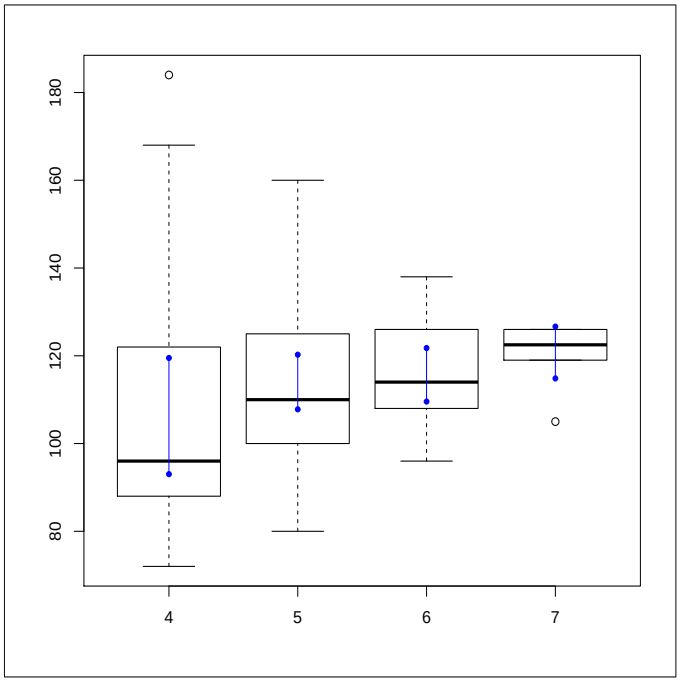

Figure 3.5: Plot for Problem 3.4 (c)

Problem 3.5 For this question, use the `OJ` data set from the `ISLR` package. This represents data from sales of two brands of orange juice. Each of the $n = 1070$ observations represents a single sales transaction. We will make use of the variables:

(1) `Purchase` = Purchased brand was either *Citrus Hill* (`= CH`) or *Minute Maid* (`= MM`).
(2) `StoreID` = ID of store at which purchase was made (`StoreID = 1,2,3,4,7`).
(3) `LoyalCH` = Customer brand loyalty score for `CH` on a scale of 0 to 1.

The objective is to determine whether or not customer loyalty differs significantly between stores.

(a) Construct side-by-side boxplots of `LoyalCH` using `Purchase` as the group variable. Interpret what you see. Use a Wilcoxon rank sum test to determine if there is a significant difference in the median of `LoyalCH` score between purchase groups.
(b) Construct side-by-side boxplots of `LoyalCH` using `StoreID` as the group variable. Fit an ANOVA model using `LoyalCH` as response and `StoreID` as the treatment variable. Is there evidence that mean loyalty score varies by store? (At this point, you need not consider any transformation of the response variable).
(c) Using Tukey's pairwise procedure, what can be said about the rankings of the mean loyalty scores, using a family-wise error rate of $\alpha_{FWE} = 0.05$. You can use function `TukeyHSD`.
(d) Noting that `LoyalCH` is constrained to be between 0 and 1, it is important to assess whether or not the distributional properties of the responses permit the reporting of accurate observed significance levels. Construct a normal quantile plot of the residuals from your ANOVA fit. Then apply the empirical rule to assess the normality of the residuals. What do you conclude? Why should this be done using the model residuals instead the response variable directly?

(e) We can use simulation methods to judge the accuracy of the observed significant levels. Suppose the true mean value μ of `LoyalCH` does not vary by store. Then a response from any store can be modeled as $y = \mu + \epsilon$, where ϵ is a zero mean error term. The distribution of ϵ can then be estimated using the residuals.
We can do this using a *bootstrap* procedure (Section 10.2 of lecture notes, Section 5.2 of ISLR). Suppose `y` is the response vector of length n, and `x` is the factor variable identifying the store. We have already fit the model `y` $\sim$ `x`. Suppose we then let `y.boot` be a random sample of size n (with replacement) of the residuals from some fitted model. This is equivalent to simulating a sample from $y = \mu + \epsilon$, where $\mu = 0$. This suffices for our purpose, since the actual value of μ will play no role in the procedure (and so can be zero with no loss of generality).
If we then fit the ANOVA model `y.boot` $\sim$ `x`, the null hypothesis of equal treatment means will hold, therefore the P-value of the F-test should possess a uniform distribution on $[0, 1]$. Of course, this depends on the correctness of the distributional assumptions (that ϵ is normally distributed), and so provides a means of assessing whether or not those hold (or at least that any deviation from normality does not significantly affect the accuracy of the reported level of significance).
To carry out the procedure, simulate M bootstrap samples, capturing the P-value from the F-test for each one. If the distributional assumptions required for the F-test hold, then the replicated P-value distribution should be approximately uniform.

(i) Suppose X is a continuous random variable with CDF $F(x) = P(X \leq x)$. Verify that $F(X)$ and $1 - F(X)$ have a uniform distribution on $[0, 1]$. How does this verify the claim made above that the P-value is uniformly distributed under the null hypothesis?

(ii) Carry out this bootstrap procedure for the ANOVA model fit in Part (b). Use $M = 100,000$. Draw a histogram of the replicated P-values, using the `nclass = 25` option. Report the proportion of the replicated P-values, say $\hat{\alpha}$, below $\alpha = 0.001, 0.01, 0.05, 0.1$. In addition, for each value of α, report $Z = (\hat{\alpha} - \alpha)/SE$, where $SE = \sqrt{\alpha(1-\alpha)/M}$ is the standard error of $\hat{\alpha}$. Interpret your results.

(iii) What is the standard error of $\hat{\alpha}$ for $\alpha = 0.1$ and M? What does this tell you about the overall accuracy of the bootstrap procedure.

(f) We will next carry out an experiment to assess the sensitivity of the bootstrap method. Consider the transformation $y^* = 1/(1 - y)$, and apply it to the responses of store `StoreID == 7`. Repeat the bootstrap procedure just described, except that the replicated response vector `y.boot` will be constructed by sampling n responses with replacement from the transformed responses y^*. Note that although we only sample responses from store `StoreID == 7`, responses for all original stores are being simulated. Would the observed significance levels using data with this distribution be accurate?

(g) Repeat Part (f), but apply a Box-Cox transformation to the replicated samples. You will need to load the `MASS` package. Under this transformation, would the observed significance levels be accurate?

SOLUTION: The solution makes use of the following code:

```
>
> library(ISLR)
> library(MASS)
>
> set.seed(91885)
>
```

```
> pdf('fig.prob.oj.1.pdf')
> par(mfrow=c(3,2),oma=c(2,2,2,2))
>
>
> ### (a)
>
> boxplot(LoyalCH~Purchase, data=OJ, ylab='CH Loyalty Score')
> wilcox.test(LoyalCH~Purchase, data=OJ)

Wilcoxon rank sum test with continuity correction

data:  LoyalCH by Purchase
W = 238000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> title("(a)")
>
> ### (b)
>
> # extract response and treatment variable (convert to factor)
>
> y = OJ$LoyalCH
> x = as.factor(OJ$StoreID)
>
> # Fit ANOVA model
>
> fitaov = aov(y~x)
> summary(fitaov)
              Df Sum Sq Mean Sq F value Pr(>F)
x              4  18.94   4.734   61.21 <2e-16 ***
Residuals   1065  82.37   0.077
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> boxplot(LoyalCH~StoreID, data=OJ, xlab="Store ID", ylab='CH Loyalty Score')
> title("(b)")
>
>
> ### (c)
>
> # Tukey's pairwise comparisons
>
> TukeyHSD(fitaov)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = y ~ x)

$x
           diff         lwr         upr     p adj
2-1 -0.05302437 -0.13226671  0.02621798 0.3576340
3-1 -0.19716965 -0.27856011 -0.11577918 0.0000000
4-1  0.22021359  0.13171167  0.30871550 0.0000000
7-1  0.10989969  0.03709693  0.18270246 0.0003846
3-2 -0.14414528 -0.21862665 -0.06966391 0.0000015
4-2  0.27323795  0.19104516  0.35543075 0.0000000
7-2  0.16292406  0.09793706  0.22791106 0.0000000
4-3  0.41738323  0.33311750  0.50164896 0.0000000
7-3  0.30706934  0.23947964  0.37465904 0.0000000
7-4 -0.11031389 -0.18631750 -0.03431028 0.0007422

>
> ### (d)
>
> # extract residuals
>
> fit.res = fitaov$residuals
>
> # normal quantile plot
>
> qqnorm(fit.res,main='(d)')
> qqline(fit.res)
>
>
> # vectorized empirical rule calculation for k = 1,2,3 standard deviations
>
> emp.rule = sapply(1:3, function(k) {
+   z = fit.res/sd(fit.res)
+   mean(abs(z) <= k)
+ })
>
> # report theoretical and observed tail probabilities
>
> cbind(1-2*pnorm(1:3,lower.tail=F), emp.rule)
               emp.rule
[1,] 0.6826895 0.6130841
[2,] 0.9544997 0.9747664
[3,] 0.9973002 1.0000000
>
> ### (e)
```

```
>
> # list of alpha values
>
> p.list = c(0.001,0.01,0.05,0.1)
>
> # create simple summary function
>
> f0 = function(px) {
+   pest = mean(pvs <= px)
+   pse = sqrt(pest*(1-pest)/length(pvs))
+   psenull = sqrt(px*(1-px)/length(pvs))
+   ans = c(px,pest, (pest-px)/psenull)
+ }
>
> # set number of bootstrap replications
>
> nboot = 100000
>
> # Create subroutines needed for bootstrap simulations
>
> # Fitting function
>
> aovbc = function(y,x) {
+   return(aov(y~x))
+ }
>
> # This subroutine extracts the p-value for the F-test
>
> aovbc.boot = function(y,x) {
+   yb = sample(y,length(x),replace=T)
+   fit = aovbc(yb,x)
+   return(anova(fit)[1,5])
+ }
>
> # Calculate bootstrap replications
>
> boot1 = sapply(1:nboot, function(i) aovbc.boot(y,x))
>
> # Report summaries of bootstrap replications
>
> pvs = boot1
> hist(pvs,nclass=25,main='(e)')
> sapply(p.list, f0)
          [,1]     [,2]       [,3]      [,4]
[1,]  0.0010000  0.01000  0.0500000 0.1000000
```

```
[2,]  0.0009800  0.00952  0.0499000 0.1007100
[3,] -0.2001001 -1.52554 -0.1450953 0.7484057
>
> ### (f)
>
> # Extract responses for store ID = 7
>
> ytr = 1/(1-y[x==7])
>
> # Repeat bootstrap with new response vectors
>
> boot2 = sapply(1:nboot, function(i) aovbc.boot(ytr,x))
>
> pvs = boot2
> hist(pvs,nclass=25,main='(f)')
> sapply(p.list, f0)
         [,1]      [,2]      [,3]      [,4]
[1,] 0.001000  0.010000   0.05000   0.10000
[2,] 0.001470  0.009010   0.04126   0.08403
[3,] 4.702352 -3.146427 -12.68132 -16.83386
>
> ### (g)
>
> # Use data from Part (f), but create new fitting function that incorporates
> # the Box-Cox transformation. This should be redone for each replication.
>
> aovbc = function(y,x) {
+   junk = boxcox(y~1,plotit=F)
+   lam = junk$x[which.max(junk$y)]
+   ybc = (y^lam-1)/lam
+   return(aov(ybc~x))
+ }
>
> boot3 = sapply(1:nboot, function(i) aovbc.boot(ytr,x))
>
> pvs = boot3
> hist(pvs,nclass=25,main='(g)')
> sapply(p.list, f0)
          [,1]       [,2]      [,3]       [,4]
[1,]  0.001000  0.0100000 0.0500000  0.1000000
[2,]  0.000770  0.0097800 0.0502900  0.0994400
[3,] -2.301151 -0.6992059 0.4207762 -0.5902918
>
> dev.off()
null device
```

```
          1
RStudioGD
        2
>
```

See Figure 3.6 for plots.

(a) Loyalty scores are clearly higher for `CH` purchases (`p-value < 2.2e-16`). See Figure 3.6 (a).

(b) From output $F = 61.21$, with 4/1065 numerator/denominator degrees of freedom, $P <$ `2e-16`. We conclude that loyalty varies by score. See Figure 3.6 (b).

(c) From the output we conclude, with $\alpha_{FWE} = 0.05$

$$\begin{aligned}
\mu_3 &< \mu_1, \\
\mu_4 &> \mu_1, \\
\mu_7 &> \mu_1, \\
\mu_3 &< \mu_2, \\
\mu_4 &> \mu_2, \\
\mu_7 &> \mu_2, \\
\mu_4 &> \mu_3, \\
\mu_7 &> \mu_3, \\
\mu_7 &< \mu_4.
\end{aligned}$$

More concisely,

$$\mu_4 > \mu_7 > \max\{\mu_1, \mu_2\} \geq \min\{\mu_1, \mu_2\} > \mu_3.$$

(d) From Figure 3.6 (d) the residual distribution deviates noticeably from normality. However, it remains symmetric. The observed empirical rule frequencies are 0.613, 0.975, 1.00 for 1,2,3 standard deviations, which differ from the normal theoretical frequencies of 0.68, 0.95, 0.997, at least for the one standard deviation case.
The responses should not be used because the means may vary, so that the responses could not be interpreted as a sample from a single normal distribution. However, if the assumptions hold, the residuals will be close to a homogenous normally distributed sample (although this does not hold exactly).

(e) (i) Suppose $F(x)$ is the CDF of X. Let $U = F(X)$, and set the CDF of U to be $F_U(t) = P(U \leq t)$ Then

$$F_U(t) = P(F(X) \leq t) = P(X \leq F^{-1}(t)) = F(F^{-1}(t)) = t.$$

But this is the CDF of the uniform random variable with support $(0, 1)$. Finally, it is easily verified that if U possesses a uniform distribution on $(0, 1)$, then so does $1 - U$. The P-value can be expressed $P = 1 - F'(F_{obs})$, where F' is the CDF of the observed F-statistic under the null hypothesis of equal treatment means.

(ii) The histogram appears uniformly distributed (Figure 3.6 (e)). Each value of $\hat{\alpha}$ is close to the corresponding α (see output above). The Z-scores are the appropriate statistic for testing the null hypothesis $H_o : E[\hat{\alpha}] = \alpha$. Then $Z \sim N(0, 1)$, approximately, under H_o. In this example, for all values of α considered we observe $|Z| < 1.53$, so we do not reject H_o, which is consistent with the P-values being uniformly distributed. This represents evidence that the reported significance levels are accurate.

(iii) We have $\sqrt{0.1 \times 0.9/100000} \approx 0.00095$. This value is smaller for $\alpha < 0.1$. The estimates $\hat{\alpha}$ of α are accurate to within, say, ± 0.002.

(f) From Figure 3.6 (f) the P-value distribution deviates noticeably from the uniform. The reported Z-scores are all large ($|Z| > 3.14$). The P-values are not uniformly distributed, and the reported significance levels cannot be assumed to be accurate.

(g) From Figure 3.6 (g) the P-value distribution appears uniform. The reported Z-scores are generally small, except possibly for $\alpha = 0.001$ ($Z = -2.3$). The P-values appear uniformly distributed, and the reported significance levels can be assumed to be accurate.

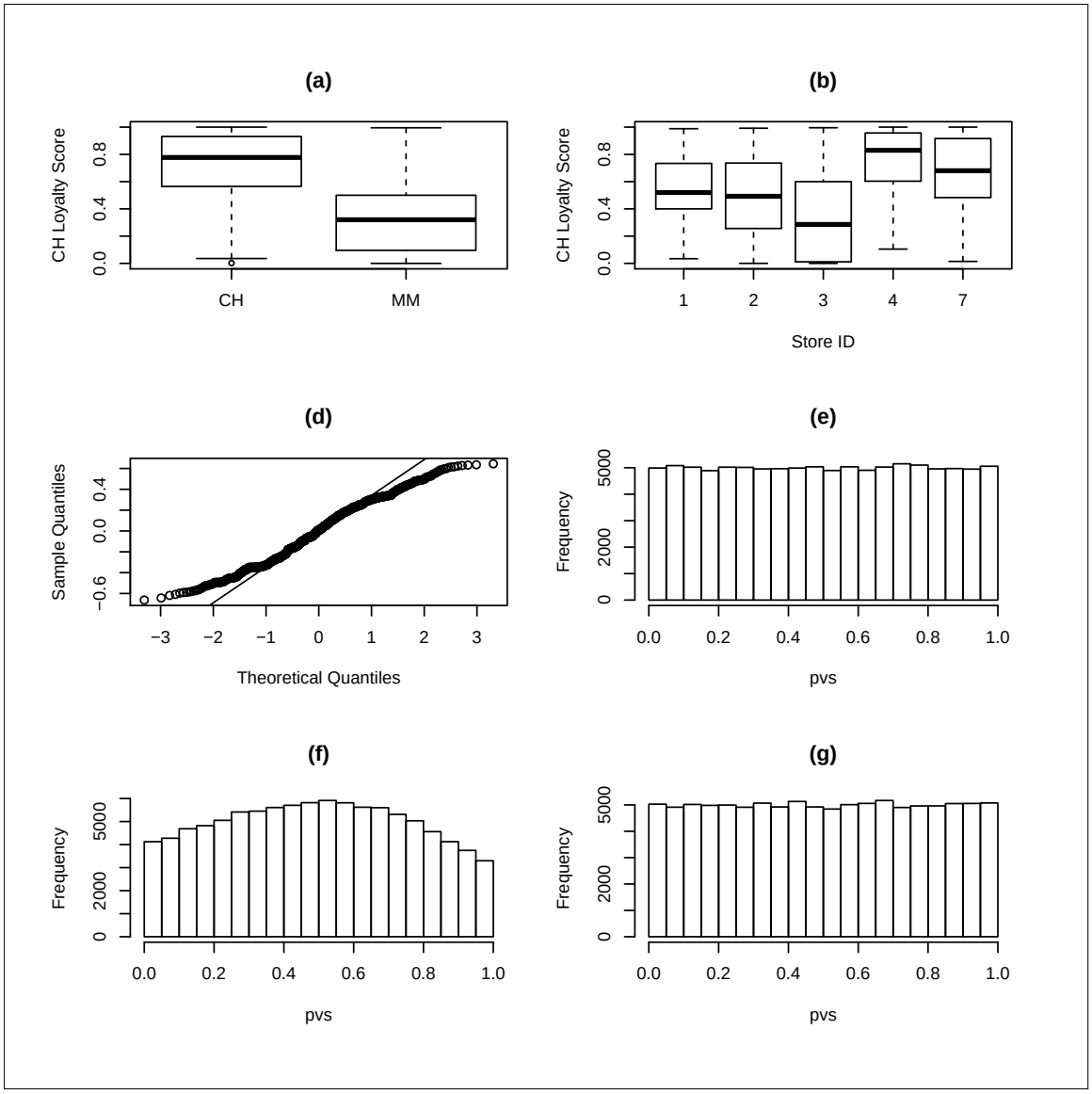

Figure 3.6: Plot for Problem 3.5.

Problem 3.6 We will add a new variable to the analysis of problem Problem 3.5, namely `PriceDiff` (sale price of MM less sale price of CH).

(a) Fit the linear model

$$\texttt{LoyalCH} = \beta_0 + \beta_1 \times \texttt{PriceDiff}. \text{ [Model 1]}$$

Is there significant evidence that `LoyalCH` varies with `PriceDiff`? If so, in what way?

(b) Fit the linear model (or one equivalent to):

$$\texttt{LoyalCH} = \beta_0 + \beta_1 \times I\{\texttt{StoreID == 1}\} + \ldots + \beta_4 \times I\{\texttt{StoreId == 4}\}. \text{ [Model 2]}$$

How does this compare to the ANOVA model of **Q1** Part (b)? **HINT:** Using the `lm` function, it is not necessary to construct individual indicator variables. Using `R` formula objects, this is done automatically if a formula such as `y ~ x.factor` is used, where `x.factor` is a vector of `factor` type.

(c) We next combine the two predictor variables. This can be done additively (using `R` formula notation):

$$\texttt{LoyalCH} \sim \texttt{PriceDiff} + \texttt{StoreID}, \text{ [Model 3]}$$

or by including all interactions:

$$\texttt{LoyalCH} \sim \texttt{PriceDiff} * \texttt{StoreID}. \text{ [Model 4]}$$

For each of the four models, superimpose these fits graphically on a scatterplot of the original data. For Models 2,3,4 present clearly the regression line separately for each store. **HINT:** One way to do this is to use the `newdata` option in the `predict` method. Set `newdata` equal to a data frame with columns `PriceDiff` and `StoreID`. For each of the five levels of `StoreID` include two rows, with `PriceDiff` equal to the endpoints of the range of `PriceDiff`. After some rearranging of the output of `predict`, the `matplot` function can be used to draw 5 separate lines on a single plot. In addition, this same method can be used for all models.

(d) Construct a table giving the SSE and the SSE (residual) degrees of freedom for each model. Which models are nested? Do a goodness of fit test to determine if Model 3 improves Model 1 or 2, and if Model 4 improves model 3. Do this directly using the SSE values. Then verify your result using the `anova` function.

(e) We next consider the possibility that `LoyalCH` varies with `PriceDiff` within some but not other stores. We can do this separately for, say, `StoreID == 1` by adding the term

$$\beta' \times \texttt{I(PriceDiff*(StoreID==1))}$$

to Model 2. Do this for each `StoreID` level, and report a P-value for the F-test comparing the full and reduced model.

(f) If you apply the Bonferroni multiple test procedure to the output of Part (e), with a family-wise error rate of $\alpha_{FWE} = 0.05$, for which stores is there evidence that `LoyalCH` varies with `PriceDiff`?

SOLUTION: The solution makes use of the following code:

```
>
> library(ISLR)
>
```

```
>
> pdf('figa1q2.pdf')
> par(mfrow=c(2,2),oma=c(2,2,2,2))
>
> ### (a)
>
> # Fit model 1
>
> fit.m1 = lm(LoyalCH~PriceDiff,data=OJ)
> summary(fit.m1)

Call:
lm(formula = LoyalCH ~ PriceDiff, data = OJ)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5947 -0.2273  0.0162  0.2729  0.5119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54847    0.01064  51.544  < 2e-16 ***
PriceDiff    0.11819    0.03450   3.426 0.000636 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3063 on 1068 degrees of freedom
Multiple R-squared:  0.01087,Adjusted R-squared:  0.009944
F-statistic: 11.74 on 1 and 1068 DF,  p-value: 0.000636

>
>
> ### (b)
>
> # Fit model 2 (make sure StoreID is converted to a factor)
>
> fit.m2 = lm(LoyalCH~as.factor(StoreID),data=OJ)
> summary(fit.m2)

Call:
lm(formula = LoyalCH ~ as.factor(StoreID), data = OJ)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6631 -0.2339  0.0053  0.2296  0.6448
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.54773    0.02220  24.678  < 2e-16 ***
as.factor(StoreID)2 -0.05302    0.02900  -1.828   0.0678 .
as.factor(StoreID)3 -0.19717    0.02979  -6.619 5.70e-11 ***
as.factor(StoreID)4  0.22021    0.03239   6.799 1.75e-11 ***
as.factor(StoreID)7  0.10990    0.02664   4.125 4.00e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2781 on 1065 degrees of freedom
Multiple R-squared:  0.1869,Adjusted R-squared:  0.1839
F-statistic: 61.21 on 4 and 1065 DF,  p-value: < 2.2e-16

>
> # Identical to ANOVA fit
>
> summary(aov(LoyalCH~as.factor(StoreID),data=OJ))
                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(StoreID)    4  18.94   4.734   61.21 <2e-16 ***
Residuals          1065  82.37   0.077
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ### (c)
>
> # Fit additive and full models
>
> fit.m3 = lm(LoyalCH~PriceDiff+as.factor(StoreID),data=OJ)
> fit.m4 = lm(LoyalCH~PriceDiff*as.factor(StoreID),data=OJ)
>
> # To loop through the four models, first create a list
>
> fit.list = list(fit.m1,fit.m2,fit.m3,fit.m4)
>
> # Create data frame for prediction method (then see what it looks like)
>
> newx = data.frame(PriceDiff = rep(range(OJ$PriceDiff),5),
                    StoreID = as.factor(rep(c(1,2,3,4,7),each=2)))
> newx
   PriceDiff StoreID
1      -0.67       1
2       0.64       1
3      -0.67       2
4       0.64       2
```

```
5      -0.67       3
6       0.64       3
7      -0.67       4
8       0.64       4
9      -0.67       7
10      0.64       7
>
> # For each model plot data, then plot fitted lines for LoyaltyCH against PriceDiff
> # for each store.
>
> for (iii in 1:4) {
+
+   # get fitted values
+
+   newfit = predict(fit.list[[iii]],newdata=newx)
+
+   # plot data (not a bad idea to make the individual points small)
+
+   plot(OJ$PriceDiff,OJ$LoyalCH,pch=20,cex=0.1,main=paste('Model',iii))
+
+   # Prepare fitted values for use in the matplot function.
+   # Note that for Model 1, there is no variation by store, so matplot is not needed.
+
+   newfitm = matrix(newfit,nrow=2)
+   colnames(newfitm) = c(1,2,3,4,7)
+
+   if (iii == 1) {
+     lines(c(-0.6,0.6),newfitm[,1],col=c(1),lty=1,lwd=2)
+   } else {
+     matlines(c(-0.6,0.6),newfitm,col=c(1,2,3,4,5),lty=1,lwd=2)
+     legend('bottomright',legend=c(1,2,3,4,7),col=c(1,2,3,4,5),lty=1,cex=0.5)
+   }
+ }
>
> ### (d)
>
>
> sse.table = sapply(fit.list, function(obj) {as.numeric(anova(obj)['Residuals',1:2])})
> dimnames(sse.table) = list(c('df','SSE'), paste('Model',1:4))
> sse.table
      Model 1    Model 2    Model 3    Model 4
df  1068.0000 1065.00000 1064.00000 1060.00000
SSE  100.2047   82.37044   81.45389   80.83364
>
> my.f.test = function(df.reduced, df.full, sse.reduced, sse.full) {
```

```
+   f.stat = ((sse.reduced-sse.full)/(df.reduced-df.full))/(sse.full/df.full)
+   df.num = df.reduced-df.full
+   df.den = df.full
+   p.value = pf(f.stat,df.num,df.den,lower.tail=F)
+   ans = c(f.stat,df.num,df.den,p.value)
+   names(ans) = c('f.stat','df.num','df.den','p.value')
+   return(ans)
+ }
> my.f.test.wrapper = function(m.red,m.full) {
+   my.f.test(sse.table[1,m.red],sse.table[1,m.full],sse.table[2,m.red],sse.table[2,m.full])
+ }
>
> # Models 1,3
>
> my.f.test.wrapper(1,3)
      f.stat       df.num       df.den      p.value
6.123367e+01 4.000000e+00 1.064000e+03 1.364478e-46
> anova(fit.m1,fit.m3)
Analysis of Variance Table

Model 1: LoyalCH ~ PriceDiff
Model 2: LoyalCH ~ PriceDiff + as.factor(StoreID)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   1068 100.205
2   1064  81.454  4    18.751 61.234 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Models 2,3
>
> my.f.test.wrapper(2,3)
      f.stat       df.num       df.den      p.value
1.197248e+01 1.000000e+00 1.064000e+03 5.613295e-04
> anova(fit.m2,fit.m3)
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ PriceDiff + as.factor(StoreID)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1065 82.370
2   1064 81.454  1   0.91655 11.973 0.0005613 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Models 3,4
```

```
>
> my.f.test.wrapper(3,4)
      f.stat       df.num       df.den      p.value
2.033378e+00 4.000000e+00 1.060000e+03 8.762724e-02
> anova(fit.m3,fit.m4)
Analysis of Variance Table

Model 1: LoyalCH ~ PriceDiff + as.factor(StoreID)
Model 2: LoyalCH ~ PriceDiff * as.factor(StoreID)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1064 81.454
2   1060 80.834  4   0.62025 2.0334 0.08763 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ### (e)
>
> # fit five new models, adding each Store x PriceDiff interaction separately
>
> fit11 = lm(LoyalCH~as.factor(StoreID)+I(PriceDiff*(StoreID==1)),data=OJ)
> fit12 = lm(LoyalCH~as.factor(StoreID)+I(PriceDiff*(StoreID==2)),data=OJ)
> fit13 = lm(LoyalCH~as.factor(StoreID)+I(PriceDiff*(StoreID==3)),data=OJ)
> fit14 = lm(LoyalCH~as.factor(StoreID)+I(PriceDiff*(StoreID==4)),data=OJ)
> fit15 = lm(LoyalCH~as.factor(StoreID)+I(PriceDiff*(StoreID==7)),data=OJ)
>
> # Collect the fits into a single list
>
> fit.list2 = list(fit11,fit12,fit13,fit14,fit15)
>
> # Compare each fit to Model 2 using a goodness-of-fit F-test
>
> ftest.table = lapply(fit.list2, function(obj) {anova(fit.m2,obj)})
> ftest.table
[[1]]
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ as.factor(StoreID) + I(PriceDiff * (StoreID == 1))
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1   1065 82.37
2   1064 81.92  1   0.45005 5.8454 0.01579 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[[2]]
```

```
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ as.factor(StoreID) + I(PriceDiff * (StoreID == 2))
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1065 82.370
2   1064 82.369  1  0.001413 0.0183 0.8926

[[3]]
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ as.factor(StoreID) + I(PriceDiff * (StoreID == 3))
  Res.Df   RSS Df  Sum of Sq  F Pr(>F)
1   1065 82.37
2   1064 82.37  1 3.7098e-06  0 0.9945

[[4]]
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ as.factor(StoreID) + I(PriceDiff * (StoreID == 4))
  Res.Df   RSS Df  Sum of Sq      F Pr(>F)
1   1065 82.37
2   1064 82.37  1 0.00073793 0.0095 0.9222

[[5]]
Analysis of Variance Table

Model 1: LoyalCH ~ as.factor(StoreID)
Model 2: LoyalCH ~ as.factor(StoreID) + I(PriceDiff * (StoreID == 7))
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1065 82.370
2   1064 81.286  1    1.0846 14.197 0.0001737 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
> dev.off()
null device
          1
>
```

See Figure 3.7 for plots.

(a) From the coefficient table we have $\hat{\beta}_1 = 0.11819$, $T = 3.426$, $df = 1068$, $P = 0.000636$. There is a

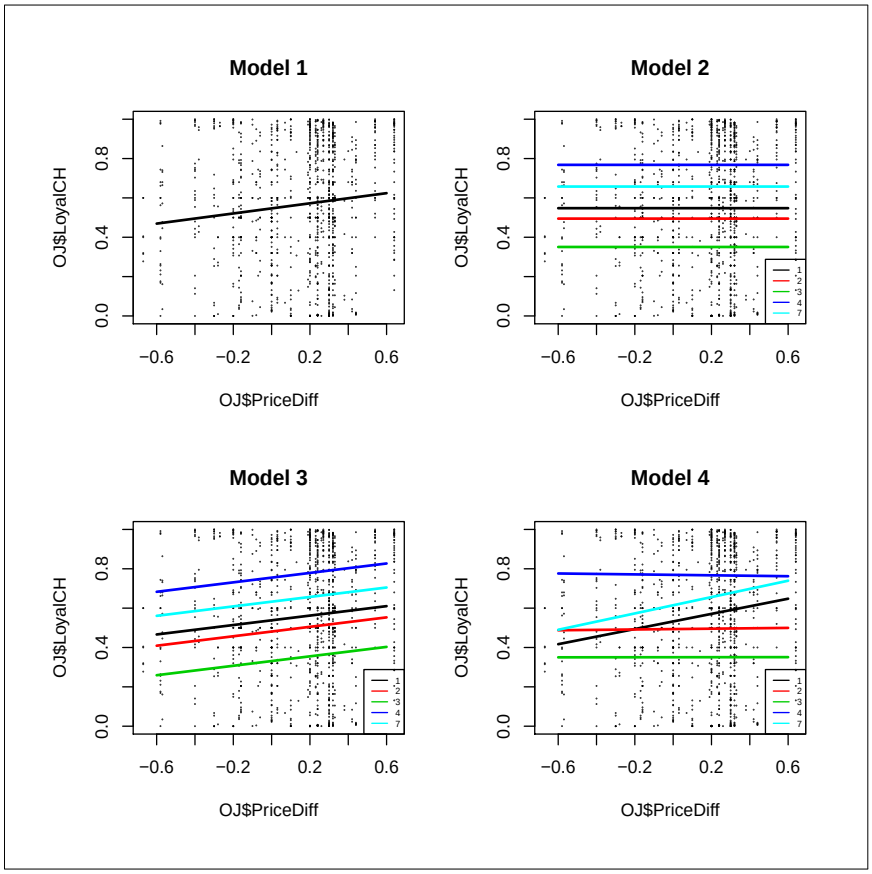

Figure 3.7: Plot for Problem 3.6.

significant positive relationship between `LoyalCH` and `PriceDiff`.

(b) Model 2 is identical to the ANOVA model of **Q1** (b). Both yield $F = 61.21$ with 4/1065 numerator/denominator degrees of freedom.

(c) See Figure 3.7.

(d) The following table (taken from the above output) contains the SSE and d.f. values.

```
> sse.table
      Model 1    Model 2    Model 3    Model 4
df  1068.0000 1065.00000 1064.00000 1060.00000
SSE  100.2047   82.37044   81.45389   80.83364
```

Model 1 is a sub-model of Models 3 and 4 (but not Model 2). Model 2 is a sub-model of Models 3 and 4. Model 3 is a sub-model of Model 4. The function `my.f.test.wrapper` inputs the indices of reduced and full models, and outputs the F-statistic

$$F = \frac{(SSE_r - SSE_f)/(df_r - df_f)}{SSE_f/df_f},$$

the degrees of freedom and the upper-tailed P-value. Here SSE_r, df_r are the reduced SSE and SSE degrees of freedom, and SSE_f, df_f are the full SSE and SSE degrees of freedom.

Model 3 improves Model 1 ($F = 61.234$, $P <$ `2.2e-16`). Model 3 improves Model 2 ($F = 11.973$, $P = 0.0005613$). Model 4 does not significantly improve Model 3 at a $\alpha = 0.05$ significance level ($F = 2.0334$, $P = 0.08763$).

(e) The P-values for the respective `PriceDiff` x `StoreID` interactions are, from the output above

```
StoreID==1, P = 0.01579
StoreID==2, P = 0.8926
StoreID==3, P = 0.9945
StoreID==4, P = 0.9222
StoreID==7, P = 0.0001737
```

We conclude, separately, at an $\alpha = 0.05$ significance level that `LoyalCH` varies linearly with `PriceDiff` for `StoreID == 1` and `StoreID == 7`.

(f) The results of Part (e) are essentially $m = 5$ distinct hypothesis tests. If we control for a family-wise error rate of $\alpha_{FWE} = 0.05$ using the Bonferroni correction procedure, we reject null hypotheses for individual P-values less than $\alpha = \alpha_{FWE}/m = 0.05/5 = 0.01$. Then we may conclude, as a group inference, that `LoyalCH` varies linearly with `PriceDiff` for `StoreID == 7` (note that we have not concluded that this effect exists only for `StoreID == 7`).

Problem 3.7 For this question, use the `Carseats` data set from the `ISLR` package. This data set is simulated, but is intended to represent sales data from 400 different stores. We will make use of the variables:

(1) `Sales` = Unit sales (in thousands) at each location.
(2) `Population` = Population size in region (in thousands).
(3) `ShelveLoc` = A factor with levels `Bad`, `Good` and `Medium` indicating the quality of the shelving location for the car seats at each site.
(4) `Urban` = A factor with levels `No` and `Yes` to indicate whether the store is in an urban or rural location.

The objective is to determine how important shelf location is to sales volume, since securing advantageous shelf position requires considerable effort. Suppose $\mu_{bad}, \mu_{med}, \mu_{good}$ are the respective mean sales volumes for each shelf location category. Two hypotheses might be

$$H_1 : \mu_{good} > \mu_{med} > \mu_{bad}$$

or

$$H_2 : \mu_{good} > \mu_{med} = \mu_{bad}.$$

If H_1 were true, then it would be worth securing good shelf space, and if that were not possible, it would be worth securing medium shelf space over bad shelf space. On the other hand, if H_2 were true, then there would be nothing to gain by attempting to secure medium shelf space if good shelf space were not available.

Fit an ANOVA model then perform a post-hoc analysis using the following steps:

(a) The analysis will be done for rural stores only. Create a subset of the data including only records with factor level `Urban=='No'`. You can use the `subset()` function.

(b) An analysis should examine any association between the response and other variables. For example, it is possible that `Sales` is associated with `Population`, and should therefore be adjusted. Construct a scatterplot of the two variables and test for correlation (you can use the `cor.test()` function). What do you conclude?

(c) Construct side-by-side boxplots of `Sales` for the factor levels of `ShelveLoc`. Do a Bartlett's test for equality of variance of `Sales` across these factor levels (use `bartlett.test()`). Are the standard assumptions of normality and equality of variance reasonable in this case? You can draw your conclusions from the equality of variance test and the boxplots alone.

(d) Fit an ANOVA model, and report a p-value for the rejection of the null hypothesis

$$H_o : \mu_{good} = \mu_{med} = \mu_{bad}.$$

Describe precisely the test statistic, and its distribution under the null hypothesis.

(e) Using Tukey's pairwise procedure (function `TukeyHSD()`) report confidence intervals for each pairwise difference in the means $\mu_{good}, \mu_{med}, \mu_{bad}$. Use a family-wise error rate of $\alpha_{FWE} = 0.01$. You will have to use the `conf.level` option. What can be said about the rankings of the means with confidence level 99%?

(f) Suppose, anticipating that better shelf location will never result in *lower* sales volume, we attempt to resolve the problem by constructing confidence intervals for the differences $\mu_{good} - \mu_{med}$ and $\mu_{med} - \mu_{bad}$ only, using Bonferroni's procedure to attain a family-wise error rate of $\alpha_{FWE} = 0.01$. What can be said about the rankings of the means with confidence level 99%, using this procedure? Does this contradict the conclusion of Part (e)?

SOLUTION: The analysis is given in the following code. Comments follow.

```
> par(mfrow=c(1,2),pty='s',oma=c(2,2,2,2),cex=1,cex.axis=0.85,cex.lab=0.85)
>
> library(ISLR)
>
> ### (a)
>
> Carseats2 = subset(Carseats, Urban=='No')
>
> ### (b)
>
> plot(Carseats2$Population,Carseats2$Sales,xlab='Sales',ylab='Population')
> cor.test(Carseats2$Population,Carseats2$Sales)

Pearson's product-moment correlation

data:  Carseats2$Population and Carseats2$Sales
t = 0.80446, df = 116, p-value = 0.4228
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1077250  0.2518532
sample estimates:
```

```
       cor
0.07448475

>
> ###(c)
>
> boxplot(Sales~ ShelveLoc, data = Carseats2,xlab='Sales',ylab='Population')
> bartlett.test(Sales~ ShelveLoc, data = Carseats2)

Bartlett test of homogeneity of variances

data:  Sales by ShelveLoc
Bartlett's K-squared = 0.53081, df = 2, p-value = 0.7669

>
> ### (d)
>
> fit = aov(Sales~ ShelveLoc, data = Carseats2)
> summary(fit)
             Df Sum Sq Mean Sq F value   Pr(>F)
ShelveLoc     2  253.5  126.75   21.83 9.16e-09 ***
Residuals   115  667.6    5.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> ### (e)
>
> TukeyHSD(fit,conf.level=0.99)
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = Sales ~ ShelveLoc, data = Carseats2)

$ShelveLoc
                 diff         lwr       upr     p adj
Good-Bad     4.384156  2.34383768  6.424474 0.0000000
Medium-Bad   1.693610 -0.06293306  3.450152 0.0136136
Medium-Good -2.690546 -4.29861786 -1.082475 0.0000069

>
> ### (f)
>
> # Get the treatment sample size
>
> ni = table(Carseats2$ShelveLoc)
```

```
>
> # We need MSE, get this from the ANOVA table
>
> mse = summary(fit)[[1]][2,3]
>
> # We can get the estimated difference from the TukeyHSD object
>
> tr.diff = TukeyHSD(fit)$ShelveLoc[,1]
>
> # Assemble the margins of error
>
> t.crit = qt(1-0.01/4,df=sum(ni)-3)
> nh = c(1/ni[1]+1/ni[2],1/ni[1]+1/ni[3],1/ni[2]+1/ni[3])
> me = t.crit*sqrt(mse*nh)
>
> # We only need the last two comparisons
>
> cbind(tr.diff,tr.diff-me,tr.diff+me)[2:3,]
              tr.diff
Medium-Bad    1.693610  0.002130126  3.385089
Medium-Good -2.690546 -4.239054118 -1.142038
>
```

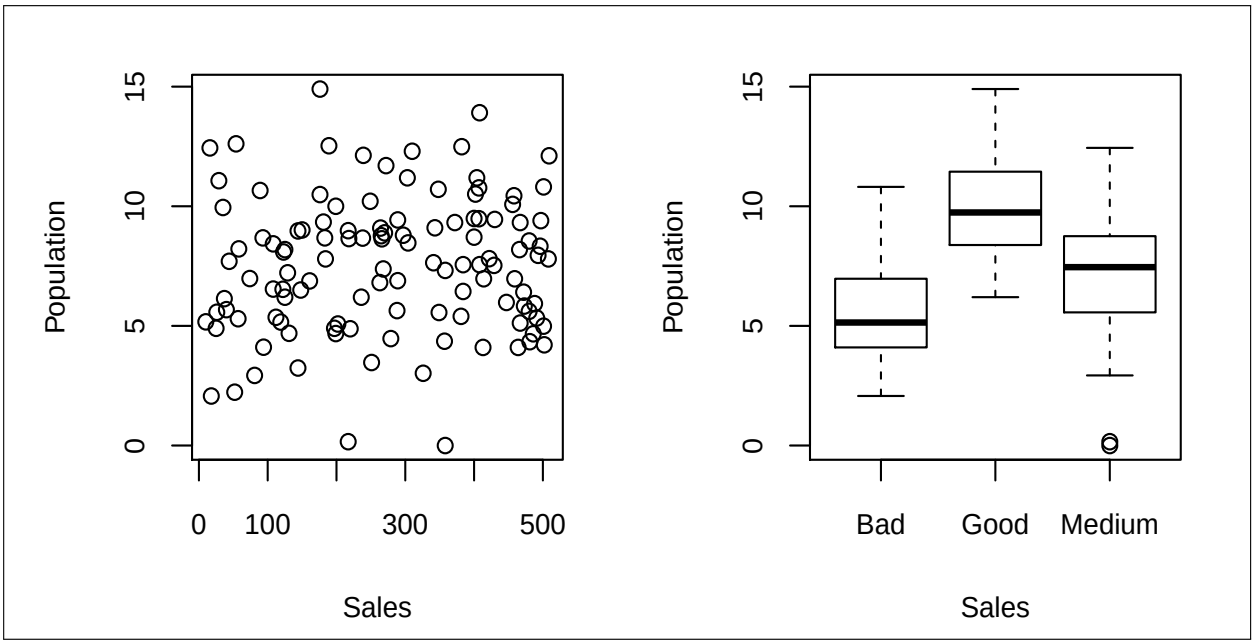

Figure 3.8: Plots for Problem 3.7.

(a) See code.

(b) The correlation is $r = 0.07448475$. The p-value against the null hypothesis $H_o : \rho = 0$ is $P = 0.4228$, so there is no significant correlation. The scatter plot otherwise shows no apparent association between `Sales` and `Population` (Figure 3.8).

(c) Bartlett's test for equality of variance has p-value $P = 0.7669$ against the null hypothesis of equal variances. The boxplots suggest that the distributions for each treatment are symmetric, and of equal variances (Figure 3.8). Based on these diagnostics, the assumptions of normality and equal variance are reasonable.

(d) See code. The test is based on the statistic $F = MST/MSE = 21.83$ (from the ANOVA table). Under H_o, F has an F-distribution with 2 numerator and 115 denominator degrees of freedom. The p-value is $P =$ `9.16e-09`, so we have very strong evidence to reject H_o.

(e) See code. Assuming the confidence intervals are correct (ie contain the true difference in means), we can conclude that $\mu_{good} > \mu_{med}$ and $\mu_{good} > \mu_{bad}$. To summarize, with 99% confidence we can conclude that μ_{good} is the uniquely largest mean. Otherwise, we cannot claim anything regarding the ordering of μ_{med} and μ_{bad}.

(f) See code. To use the Bonferroni procedure, construct confidence intervals for mean differences $\mu_{good} - \mu_{med}$ and $\mu_{med} - \mu_{bad}$ of the form

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha/(m2),n-k}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where $m = 2$, $\alpha = 0.01$, $k = 3$, $n_{good} = 28, n_{med} = 68, n_{bad} = 22$, $n = 28 + 68 + 22 = 118$. Assuming the confidence intervals are correct (ie contain the true difference in means), we can conclude that $\mu_{med} > \mu_{bad}$ and $\mu_{med} < \mu_{good}$. To summarize, with 99% confidence we can conclude that $\mu_{good} > \mu_{med} > \mu_{bad}$. This does not contradict the conclusion of Part (e). If a confidence interval for a difference in means contains zero, it means that the ordering cannot be resolved, and not that the means are equal. So, the conclusion of Part (f) is simply more precise.

Problem 3.8 For this question, use the data set `UScereal` from the `MASS` package. This data contains ingredient quantities taken from the mandatory FDA label printed on 65 brands of cereal. The objective is to determine the ingredient that contributes most to calorie content.

(a) Create side-by-side boxplots of `calories` by manufacturer (given by the categorical variable `mfr`). Identify two outliers (defined as `calories` > 300), and delete from the data all cereals made by the manufacturer responsible for those outliers.

(b) Fit a linear regression model using `calories` as the dependent variable, and all remaining variables as independent variables. This is easily done using the formula `lm(calories ~ ., data = myData)`. Which variables have regression coefficients significantly different from 0 (at a $P < 0.05$ significance level)?

(c) Refit the model, again with `calories` as dependent variable, but including as independent variables only those with significantly nonzero coefficients reported in Part (b). Include the intercept. Do these independent variables remain significant? Do the values of the coefficients change significantly?

(d) Construct side-by-side boxplots for the independent variables of Part (c). In what units are these variables (use `help(UScereal)`)? Which regression coefficient is largest (other than the intercept). Looking at the boxplots, does the independent variable with the largest coefficient necessarily contribute most to calorie content?

(e) Standardize each of the independent variables of the model fit in Part (c) to have zero mean and standard deviation one. Refit the model. Have the t-statistics reported for each independent variable changed? Which predictor contributes most to calorie content?

SOLUTION: The code for the analysis is given below.

(a) See Figure 3.9 for boxplots. The manufacturer responsible for those outliers is labeled `P` (for Post).
(b) The significant independent variables are `protein`, `fat`, `carbo` and `sugars`.
(c) The fit is shown below. The independent variables remain significant. The respective coefficient values in the original fit for these variables are 3.983205, 9.424864, 4.002802, 4.180946. The new values are 3.7978, 8.4661, 4.0402, 4.2139. They are not identical between fits, but are otherwise quite close.
(d) The units used for the four independent variables are grams (in one portion). The largest regression coefficient is associated with `fat`. However, from the boxplots of Figure 3.9 (right side) the variation of `fat` is smaller than for the remaining independent variables. Therefore, the independent variable with the largest coefficient does not necessarily contribute most to calorie content.
(e) The t-statistics are identical. To compare calorie contributions, the coefficients should be calculated with respect to a change in unit standard deviation for each ingredient. Therefore, `sugars` contributes most to calorie content, with $\hat{\beta}_{sugar} = 24.8079$ change in calories per standard deviation change in `carbo`.

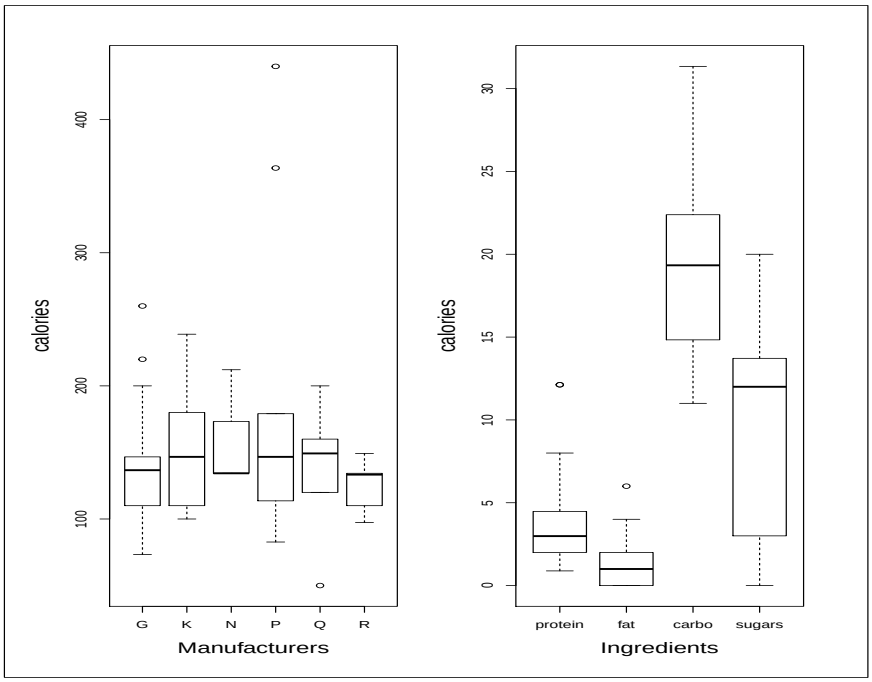

Figure 3.9: Plots for Problem 3.8.

```
> library(MASS)
> par(mfrow=c(1,2))
>
```

```
>
> ### (a)
>
> boxplot(calories~mfr,data=UScereal,ylab='calories',xlab='Manufacturers',cex.lab=1.5)
> data2 = subset(UScereal, mfr!='P')
>
> ### (b)
>
> fit3 = lm(calories ~ ., data=data2)
> summary(fit3)

Call:
lm(formula = calories ~ ., data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8371  -3.1176   0.1532   3.2058  12.8722

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.667136   5.836452  -0.457    0.650
mfrK              3.706381   2.368999   1.565    0.125
mfrN              7.375987   7.187820   1.026    0.311
mfrQ             -2.497673   3.474514  -0.719    0.476
mfrR             -0.452942   3.567393  -0.127    0.900
protein           3.983205   0.862820   4.616 3.82e-05 ***
fat               9.424864   0.910656  10.350 5.30e-13 ***
sodium            0.005558   0.010999   0.505    0.616
fibre             0.668536   0.766720   0.872    0.388
carbo             4.002802   0.222710  17.973  < 2e-16 ***
sugars            4.180946   0.238287  17.546  < 2e-16 ***
shelf             0.427604   1.463232   0.292    0.772
potassium        -0.035009   0.030312  -1.155    0.255
vitaminsenriched -0.274217   3.338727  -0.082    0.935
vitaminsnone     -1.765723   7.271831  -0.243    0.809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.992 on 41 degrees of freedom
Multiple R-squared:  0.9862,Adjusted R-squared:  0.9815
F-statistic: 209.9 on 14 and 41 DF,  p-value: < 2.2e-16

>
> ### (c)
>
```

```
> data3 = data2[,c("calories","protein","fat","carbo","sugars")]
> fit3 = lm(calories ~ ., data=data3)
> summary(fit3)

Call:
lm(formula = calories ~ ., data = data3)

Residuals:
     Min       1Q   Median       3Q      Max
-16.6047  -4.0734   0.1079   3.8670  13.4827

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2281     3.3515  -0.366    0.716
protein       3.7978     0.3753  10.118 8.62e-14 ***
fat           8.4661     0.7449  11.365 1.37e-15 ***
carbo         4.0402     0.1548  26.105  < 2e-16 ***
sugars        4.2139     0.1639  25.713  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.924 on 51 degrees of freedom
Multiple R-squared:  0.9833,Adjusted R-squared:  0.982
F-statistic: 749.3 on 4 and 51 DF,  p-value: < 2.2e-16

>
> ### (d)
>
> boxplot(data3[-1],ylab='calories',xlab='Ingredients',cex.lab=1.5)
>
> ### (e)
>
> data4 = data3
> for (i in 2:5) {data4[[i]] = (data4[[i]] - mean(data4[[i]]))/sd(data4[[i]]) }
> fit4 = lm(calories ~ ., data=data4)
> summary(fit4)

Call:
lm(formula = calories ~ ., data = data4)

Residuals:
     Min       1Q   Median       3Q      Max
-16.6047  -4.0734   0.1079   3.8670  13.4827

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.1199     0.7916  179.53  < 2e-16 ***
protein       9.2881     0.9179   10.12 8.62e-14 ***
fat          11.6437     1.0245   11.37 1.37e-15 ***
carbo        22.4673     0.8606   26.11  < 2e-16 ***
sugars       24.8079     0.9648   25.71  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.924 on 51 degrees of freedom
Multiple R-squared:  0.9833,Adjusted R-squared:  0.982
F-statistic: 749.3 on 4 and 51 DF,  p-value: < 2.2e-16

>
```

Problem 3.9 For this question, use the `cats` data set from the `MASS` package. This data includes the following observations for each of $n = 144$ cats:

```
Sex
sex: Factor with levels "F" and "M".

Bwt
body weight in kg.

Hwt
heart weight in g.
```

(a) Suppose we have linear relationship $y = \beta_0 + \beta_1 x$ between two variables x, y. If $\beta_1 \neq 0$, this can always be written as $x = \beta_0' + \beta_1' y$. Express β_0' and β_1' as functions of β_0 and β_1.

(b) Fit the following linear models using the `lm()` function:

$$\texttt{Hwt} \sim \texttt{Bwt}$$

and

$$\texttt{Bwt} \sim \texttt{Hwt}.$$

Do the least squares coefficients of the two models conform to the equivalence relationship given in Part (a)? Construct a scatter plot of the `Hwt` and `Bwt` paired observations (place `Hwt` on the vertical axis). For both models superimpose on this plot the estimated linear relationship between `Hwt` and `Bwt`. In each case, ensure that `Hwt` is represented on the vertical axis. Provide a brief explanation for your results.

(c) Fit the following three models (expressed using `R`'s model formula notation):

$$\begin{aligned} \texttt{Hwt} &\sim \texttt{Bwt} \text{ [Model 1]} \\ \texttt{Hwt} &\sim \texttt{Bwt} + \texttt{Sex} \text{ [Model 2]} \\ \texttt{Hwt} &\sim \texttt{Bwt} * \texttt{Sex} \text{ [Model 3]} \end{aligned}$$

For each model construct a scatter plot of `Hwt` and `Bwt` (placee `Hwt` on the vertical axis) and superimpose the estimated regression line (plot separate lines for the two `Sex` classes, and use a legend to identify line associated with each class). Is there statistical evidence at an $\alpha = 0.05$ significance level that either Model 2 or 3 improves Model 1?

SOLUTION:

(a) If we have $y = \beta_0 + \beta_1 x$, this may be rewritten

$$x = \frac{y}{\beta_1} - \frac{\beta_0}{\beta_1} = \beta_0' + \beta_1' y,$$

where $\beta_0' = -\beta_0/\beta_1$ and $\beta_1' = 1/\beta_1$.

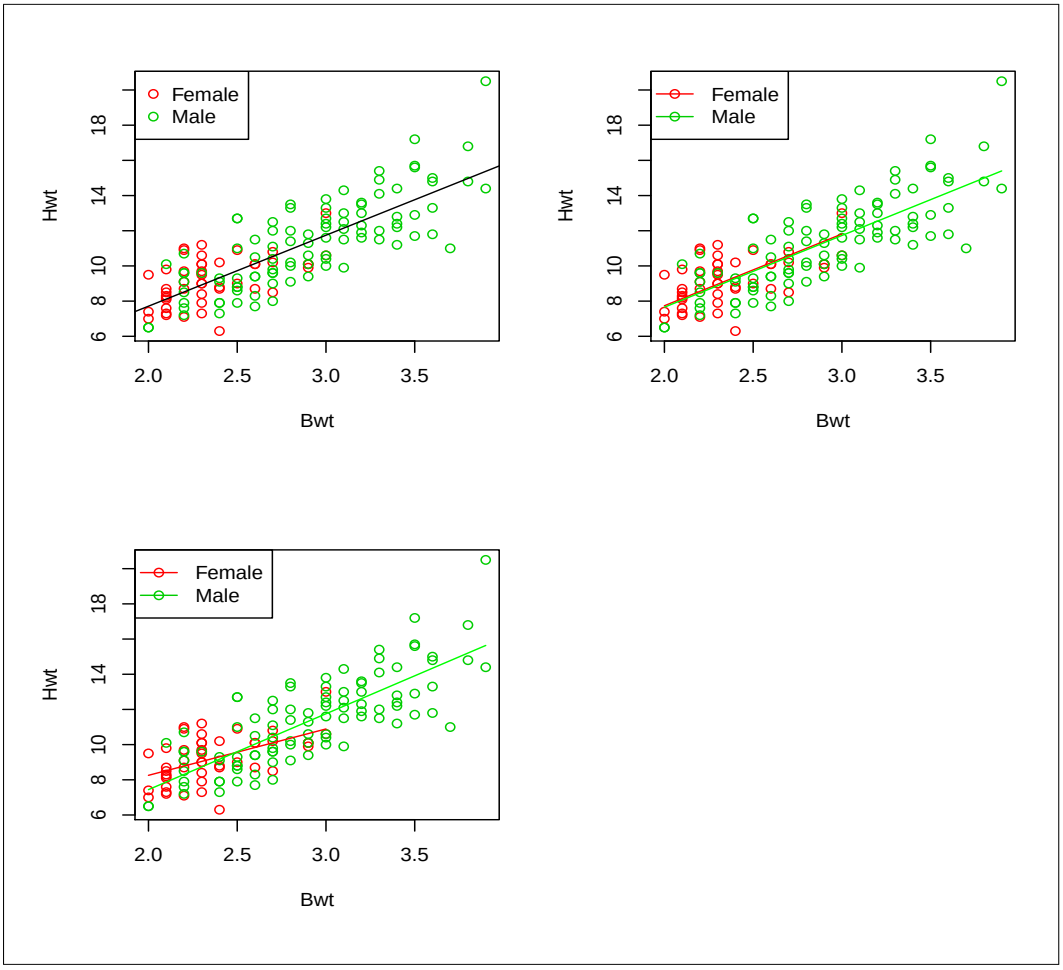

Figure 3.10: Plot for Problem 3.9 (b).

(b) The following `R` code can be used to answer Part (b). See Figure 3.10.

```
> library(MASS)
```

```
>
> ### Plot data Hwt vs Bwt
>
> plot(Hwt~Bwt,data=cats)
>
> ### Calculate both fits
>
> fit1 = lm(Hwt~Bwt,data=cats)
> fit2 = lm(Bwt~Hwt,data=cats)
>
> ### Copy coefficients
>
> cf1 = fit1$coefficients
> cf2 = fit2$coefficients
>
> ### Coefficients for the Hwt~Bwt model
>
> c(cf1)
(Intercept)         Bwt
 -0.3566624   4.0340627
>
> ### Invert Hwt~Bwt model
>
> c(-cf1[1]/cf1[2],1/cf1[2])
(Intercept)         Bwt
 0.08841271  0.24788906
>
> ### Coefficients for the Bwt~Hwt model
>
> c(cf2)
(Intercept)         Hwt
  1.0196367   0.1602902
>
> ### Plot regression line for Hwt~Bwt model
>
> abline(cf1)
>
> ### Plot regression line for Bwt~Wwt model,
> ### after inverting fuction.
> ### Use dashed lines.
>
> abline(-cf2[1]/cf2[2],1/cf2[2],lty=2)
>
> ### Create legend
>
```

```
> legend('topleft',legend=c('Hwt~Bwt model','Bwt~Hwt model'),lty=c(1,2))
>
```

The fitted coefficients for the `Hwt` $\sim$ `Bwt` model are $(\hat{\beta}_0, \hat{\beta}_1) = (-0.3566624, 4.0340627)$. If we use the inversion formula of Part (a) we get

$$\hat{\beta}'_0 = -\hat{\beta}_0/\hat{\beta}_1 = 0.08841271$$
$$\hat{\beta}'_1 = 1/\hat{\beta}_1 = 0.24788906$$

(these values are calculated by the code). However, the fitted coefficients for the `Bwt` $\sim$ `Hwt` fit are $(\hat{\beta}^*_0, \hat{\beta}^*_1) = (1.0196367, 0.1602902)$, which are not equal to the transformed coefficients $(\hat{\beta}'_0, \hat{\beta}'_1)$. In addition, the two fitted models `Hwt` $\sim$ `Bwt` and `Bwt` $\sim$ `Hwt` shown in Figure 3.10, after suitable transformations, are clearly not equal.

Recall that least squares regression minimizes the total squared *vertical* distance over all pairs (x_i, y_i) from the *response* y_i to the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$. If x and y are exchanged, then the new least squares fit is equivalent to minimizing the total squared *horizontal* distance over all pairs (x_i, y_i) from the *predictor variable* x_i to the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$. These are two distinct optimization problems, and we should not expect them to infer the same relationship between x and y.

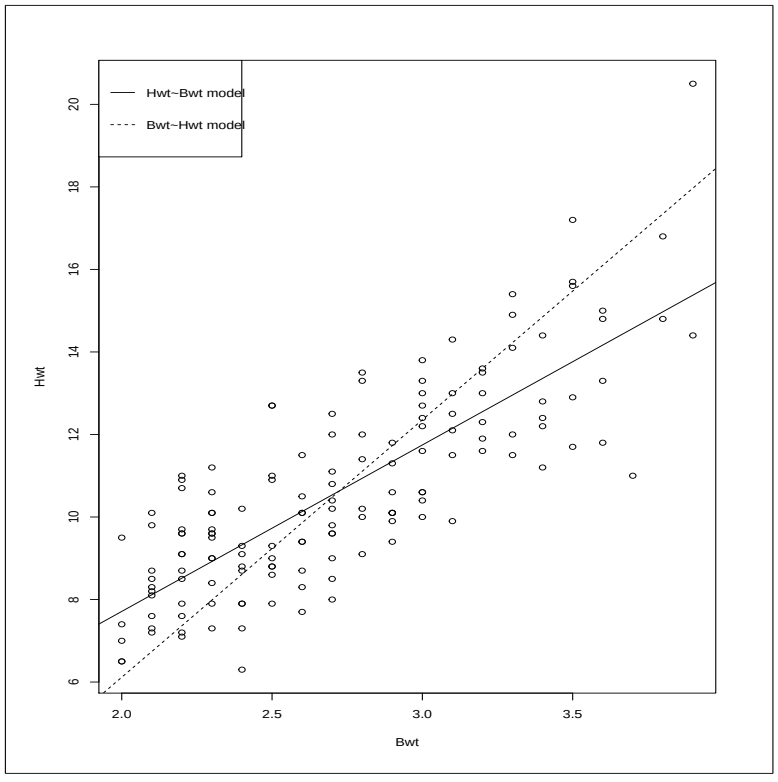

Figure 3.11: Plot for Problem 3.9 (c).

(c) The following `R` code can be used to plot the required graphs for Part (c). See Figure 3.11.

```
par(mfrow=c(2,2))
```

```
### Plot model 1

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
legend('topleft',legend=c('Female','Male'),pch=1,col=2:3)

cf = fit1$coefficients
abline(cf,col='black')

### Plot model 2

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
legend('topleft',legend=c('Female','Male'),lty=1,pch=1,col=2:3)

# Here, we only use the observed range of Bwt for each sex

cf = fit2$coefficients

x = seq(min(cats$Bwt[cats$Sex=="F"]),max(cats$Bwt[cats$Sex=="F"]),0.1)
lines(x, cf[1]+cf[2]*x,col='red')
x = seq(min(cats$Bwt[cats$Sex=="M"]),max(cats$Bwt[cats$Sex=="M"]),0.1)
lines(x, cf[1]+cf[3]+cf[2]*x,col='green')


### Plot model 3

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
legend('topleft',legend=c('Female','Male'),lty=1,pch=1,col=2:3)

# Here, we only use the observed range of Bwt for each sex

cf = fit3$coefficients

x = seq(min(cats$Bwt[cats$Sex=="F"]),max(cats$Bwt[cats$Sex=="F"]),0.1)
lines(x, cf[1]+cf[2]*x,col='red')
x = seq(min(cats$Bwt[cats$Sex=="M"]),max(cats$Bwt[cats$Sex=="M"]),0.1)
lines(x, cf[1]+cf[3]+(cf[2]+cf[4])*x,col='green')
```

The following R code can be used to perform the required tests for Part (c). Directly from the output, the F-test P-value for the Models 1 and 2 comparison is $P = 0.7875$ and the F-test P-value for the Models 1 and 3 comparison is $P = 0.1337$. We cannot conclude that either model 2 or 3 improves model 1 at an $\alpha = 0.05$ sigificance level.

```
>
> ### Use the anova() function to do a goodness of fit F-test
>
```

```
> anova(fit1,fit2)
Analysis of Variance Table


Model 1: Hwt ~ Bwt
Model 2: Hwt ~ Bwt + Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 299.53
2    141 299.38  1    0.1548 0.0729 0.7875
> anova(fit1,fit3)
Analysis of Variance Table

Model 1: Hwt ~ Bwt
Model 2: Hwt ~ Bwt * Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 299.53
2    140 291.05  2    8.4865 2.0411 0.1337
>
```

Problem 3.10 For this question, use the `Insurance` data set from the `MASS` package. This data includes the following observations for each of $n = 64$ insurance companies:

```
District
factor: district of residence of policyholder (1 to 4): 4 is major cities.

Group
an ordered factor: group of car with levels <1 litre, 1-1.5 litre, 1.5-2 litre, >2 litre.

Age
an ordered factor: the age of the insured in 4 groups labelled <25, 25-29, 30-35, >35.

Holders
numbers of policyholders.

Claims
numbers of claims
```

(a) Fit a linear model with response `Claims`, and the remaining variables as predictors. Create a residual plot (residuals against fitted values). Also create a normal quantile plot for the residuals. Do the usual assumptions for linear regression seem reasonable in this case? Comment briefly.

(b) We will try to transform `Claims` using the function $h(x) = \log(x + a)$ (use the natural logarithm). For the standard log-transformation we would set $a = 0$. Why can't we do that here? Repeat Part (a) after replacing response `Claims` with the transformed response $h($`Claims`$)$. Use $a = 1$, then $a = 10$. Which succeeds better in normalizing the residuals?

(c) We can, in principal, consider all models using some subset of the original four predictors, including the original four predictors, and no predictors. We can assume all models include an intercept term. How many such models are there.

(d) Create a list in R of model formulae representing the collection of models defined in Part (c). Note that we can obtain the full model formula, then remove a predictor from the model with the following code:

```
> fit1 = lm(log(Claims+10) ~ .,data=Insurance)
> full.formula = formula(terms(fit1))
> next.formula = update(full.formula, ~ . -District)
> full.formula
log(Claims + 10) ~ District + Group + Age + Holders
> next.formula
log(Claims + 10) ~ Group + Age + Holders
>
```

Use this list to calculate R^2_{adj} for each model. Identify the model with the largest R^2_{adj}.

SOLUTION:

(a) The following R code can be used to plot the required graphs for Parts (a) and (b). See Figure 3.12. From the residual plot (first row of Figure 3.12) it is clear that the variance is larger for larger fitted values. From the normal quantile plot (first row of Figure 3.12) it is clear that the distribution of the residuals is not normal, especially values located in the left and right tails.

```
library(MASS)

### (a)

pdf("fig1A1Q3ab.pdf")
par(mfrow=c(3,2))

# Fit model

fit1 = lm(Claims ~ .,data=Insurance)

# Residual plot

plot(fit1$fitted.values,fit1$residuals,main="Residual plot Q3a")
abline(h=0)

# Normal quantile plot

qqnorm(fit1$residuals,main="Normal quantile plot Q3a")
qqline(fit1$residuals)


### (b)

# a = 1
```

```
fit2 = lm(log(Claims+1) ~ .,data=Insurance)
plot(log(fit2$fitted.values),fit1$residuals,main = "Residual plot Q3b, a = 1")
abline(h=0)
qqnorm(fit2$residuals,main="Normal quantile plot Q3b, a = 1")
qqline(fit2$residuals)


# a = 10

fit2 = lm(log(Claims+10) ~ .,data=Insurance)
plot(log(fit2$fitted.values),fit1$residuals,main = "Residual plot Q3b, a = 10")
abline(h=0)
qqnorm(fit2$residuals,main="Normal quantile plot Q3b, a = 10")
qqline(fit2$residuals)

dev.off()
```

(b) The minimum value of `Claims` is zero, for which the logarithm is not defined. Therefore, the $\log(x)$ transformation cannot be used. Figure 3.12 contains all required plots for this part. Neither transformation succeeds entirely in satisfying the constant variance assumption (see residuals plots). However, the $\log(x+10)$ transformation yields residuals closer to the normal distribution (see normal quantile plots).

(c) If there are 4 predictors, then the number of predictor subsets is $2^4 = 16$, including the empty set. All models include the intercept, so the total number of models is 16.

(d) The following code can be used to answer Part (d). From the resulting table, the full model `log(Claims + 10)  1 + District + Group + Age + Holders` has the highest R^2_{adj} ($= 0.9442990$).

```
> ### (d)
>
> # This is a recursive subroutine that
> # returns all subsets of the elements of
> # a vector x of size k or less
>
> subset.enum = function(x,k) {
+
+   if (k == 0 | length(x) == 0) {
+     return(vector("list",1))
+   } else {
+     list1 = subset.enum(x[-1],k)
+     list2 = subset.enum(x[-1],k-1)
+     list3 = lapply(list2,function(y) {c(x[1],y)} )
+     return(append(list1,list3))
+   }
+ }
>
```

```
>
> # We can get the term labels this way
>
> fit1 = lm(log(Claims+10) ~ .,data=Insurance)
> full.formula = formula(terms(fit1))
> tm = attr(terms(full.formula),"term.labels")
> tm
[1] "District" "Group"    "Age"      "Holders"
>
> # for any subset of terms (say District and Group) we can construct a formula this way
>
> as.formula(paste("log(Claims+10) ~ ", paste(tm[1:2], collapse="+")))
log(Claims + 10) ~ District + Group
>
> # We just need to enumerate all subsets
>
> subset.list = subset.enum(1:4,4)
>
> # This subroutine returns a formula using the predictor subset
> # defined by ondex subset subl.
>
> ff.sub = function(subl) {
+   if (is.null(subl)) {
+     fm = "log(Claims+10) ~ 1"
+   } else {
+     fm = as.formula(paste("log(Claims+10) ~ 1 +", paste(tm[subl], collapse="+")))
+   }
+   return(fm)
+ }
>
> # Create a list of formula for all predictor subsets
> # defined by the index subsets in subset.list.
>
> formula.list = lapply(subset.list, ff.sub)
>
> # For each formula fit the model, then extract the adjusted R-squared.
>
> radj.vector = sapply(formula.list, function(fm) {
+   fit1 = lm(fm,data=Insurance)
+   return(summary(fit1)$adj.r.squared)
+ })
>
> # Display the the adjusted R-squared for each formula.
>
> data.frame(as.character(formula.list),radj.vector)
```

```
                                as.character.formula.list. radj.vector
1                                       log(Claims+10) ~ 1   0.0000000
2                           log(Claims + 10) ~ 1 + Holders   0.6801578
3                              log(Claims + 10) ~ 1 + Age   0.5200682
4                      log(Claims + 10) ~ 1 + Age + Holders   0.7468752
5                             log(Claims + 10) ~ 1 + Group   0.1517469
6                    log(Claims + 10) ~ 1 + Group + Holders   0.7023331
7                       log(Claims + 10) ~ 1 + Group + Age   0.7071737
8              log(Claims + 10) ~ 1 + Group + Age + Holders   0.8100072
9                           log(Claims + 10) ~ 1 + District   0.1668537
10                 log(Claims + 10) ~ 1 + District + Holders   0.7085192
11                    log(Claims + 10) ~ 1 + District + Age   0.7230757
12           log(Claims + 10) ~ 1 + District + Age + Holders   0.8248050
13                  log(Claims + 10) ~ 1 + District + Group   0.3353690
14         log(Claims + 10) ~ 1 + District + Group + Holders   0.7416876
15             log(Claims + 10) ~ 1 + District + Group + Age   0.9318542
16 log(Claims + 10) ~ 1 + District + Group + Age + Holders   0.9442990
>
```

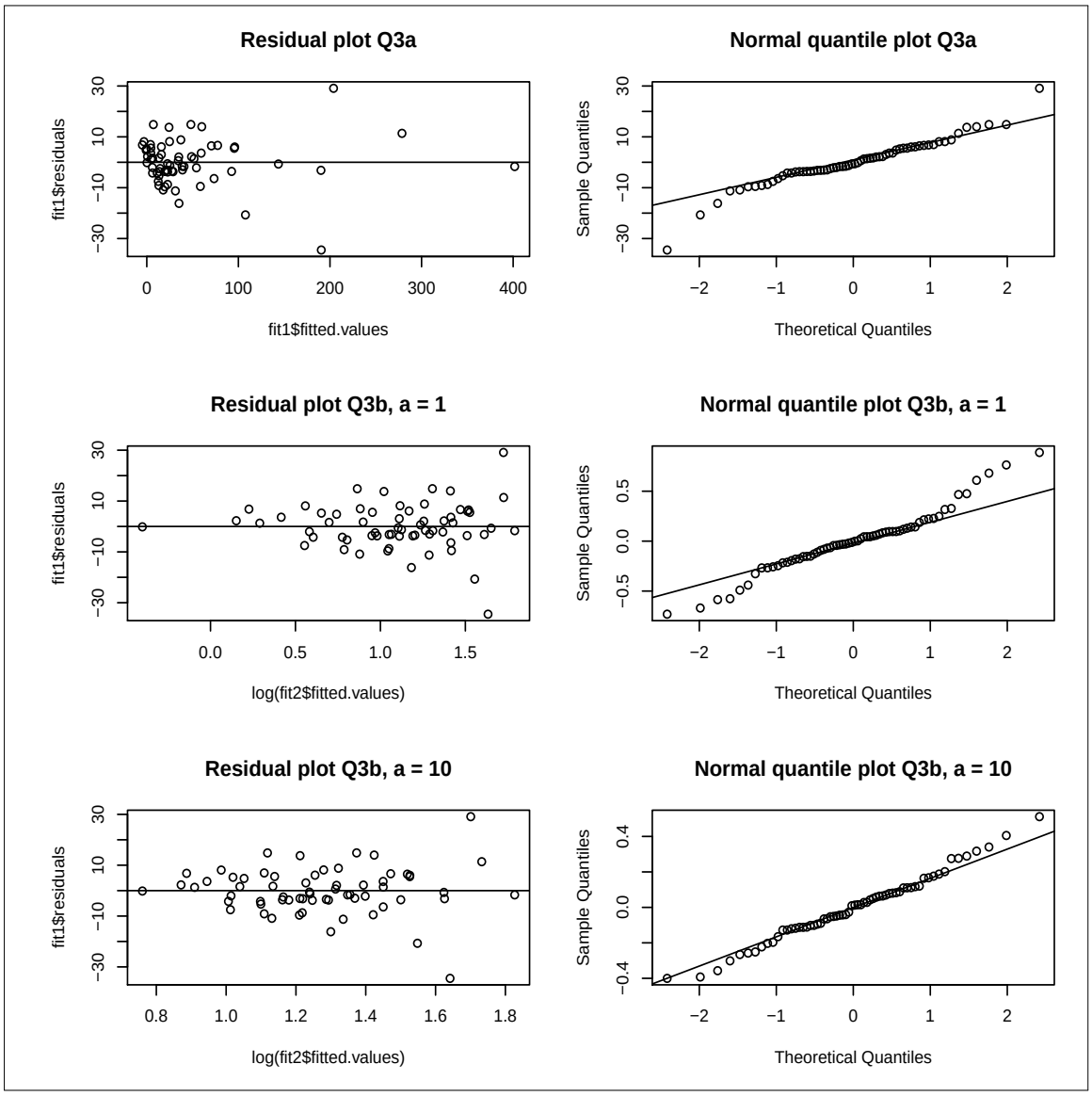

Figure 3.12: Plots for Problem 3.10.