Lecture Notes on MS237

Mathematical statistics

Lecture notes by Janet Godolphin

2010

Contents

MS237 Mathematical Statistics

Level 2 **Spring Semester** **Credits 15**

Course Lecturer in 2010

D. Terhesiu **email:** d.terhesiu@surrey.ac.uk

Class Test

The Class Test will be held on **Thursady 11th March** (week 5), starting at 12.00.

Class tests will include questions of the following types:

- examples and proofs previously worked in lectures,
- questions from the self-study exercises,
- previously unseen questions in a similar style.

The Class Test will comprise 15% of the overall assessment for the course.

Coursework

Distribution: Coursework will be distributed at 14.00 on **Friday 26th March**.
Collection: Coursework will be collected on **Thursday 29th April** in Room LTB.

The Coursework will comprise 10% of the overall assessment for the course.

Chapter 1

Chapter 1 contains and reviews prerequisite material from MS132. Due to time constraints, students are expected to work through at least part of this material independently at the start of the course.

Objectives and learning outcomes

This module provides theoretical background for many of the topics introduced in MS132 and for some of the topics which will appear in subsequent statistics modules.

At the end of the module, you should

(1) be familiar with the main results of statistical distribution theory;

(2) be able to apply this knowledge to suitable problems in statistics

Examples, exercises, and problems

Blank spaces have been left in the notes at various positions. These are for additional material and worked examples presented in the lectures. Most chapters end with a set of self-study exercises, which you should attempt in your own study time *in parallel* with the lectures.

In addition, six exercise sheets will be distributed during the course. You will be given a week to complete each sheet, which will then be marked by the lecturer and returned with model solutions. It should be stressed that completion of these exercise sheets is not compulsory but those students who complete the sheets do give themselves a considerable advantage!

Selected texts

Freund, J. E. *Mathematical Statistics with Applications*, Pearson (2004)

Hogg, R. V. and Tanis, E. A. *Probability and Statistical Inference*, Prentice Hall (1997)

Lindgren, B. W. *Statistical Theory*, Macmillan (1976)

Mood, A. M., Graybill, F. G. and Boes, D. C. *Introduction to the Theory of Statistics*, McGraw-Hill (1974)

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. *Mathematical Statistics with Applications*, Duxbury (2002)

Useful series

These series will be useful during the course:

$$
\begin{aligned}
(1-x)^{-1} &= \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots \\
(1-x)^{-2} &= \sum_{k=0}^{\infty} (k+1)x^k = 1 + 2x + 3x^2 + 4x^3 + \cdots \\
e^x &= \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots
\end{aligned}
$$

Chapter 1

Introductory revision material

This chapter contains and reviews prerequisite material from MS132. If necessary you should review your notes for that module for additional details. Several examples, together with numerical answers, are included in this chapter. It is strongly recommended that you work independently through these examples in order to consolidate your understanding of the material.

1.1 Basic probability

Probability or **chance** can be measured on a scale which runs from **zero**, which represents **impossibility**, to **one**, which represents **certainty**.

1.1.1 Terminology

A **sample space**, Ω, is the set of all possible **outcomes** of an experiment. An **event** $E \in \Omega$ is a subset of Ω.

Example 1 *Experiment:* roll a die twice. Possible *events* are $E_1 = \{\text{1st face is a 6}\}$, $E_2 = \{\text{sum of faces} = 3\}$, $E_3 = \{\text{sum of faces is odd}\}$, $E_4 = \{\text{1st face - 2nd face} = 3\}$. Identify the sample space and the above events. Obtain their probabilities when the die is **fair**.

Answer:

		second roll					
	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
first	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
roll	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

$p(E_1) = \frac{1}{6}$; $p(E_2) = \frac{1}{18}$; $p(E_3) = \frac{1}{2}$; $p(E_4) = \frac{1}{12}$.

Combinations of events

Given events A and B, further events can be identified as follows.

- The **complement** of any event A, written $\bar{A}$ or A^c, means that A does **not** occur.
- The **union** of any two events A and B, written $A \cup B$, means that A **or** B **or** both occur.
- The **intersection** of A and B, written as $A \cap B$, means that both A **and** B occur.

Venn diagrams are useful in this context.

1.1.2 Probability axioms

Let $\mathcal{F}$ be the class of all events in Ω. A **probability (measure)** P on $(\Omega, \mathcal{F})$ is a real-valued function satisfying the following three axioms:

1. $P(E) \geq 0$ for every $E \in \mathcal{F}$
2. $P(\Omega) = 1$
3. Suppose the events E_1 and E_2 are mutually exclusive (that is, $E_1 \cap E_2 = \emptyset$).

Then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Some consequences:

(i) $P(\bar{E}) = 1 - P(E)$ (so in particular $P(\emptyset) = 0$)

(ii) For *any* two events E_1 and E_2 we have the **addition rule**

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Example 1: (continued)
Obtain $P(E_1 \cap E_2)$, $P(E_1 \cup E_2)$, $P(E_1 \cap E_3)$ and $P(E_1 \cup E_3)$

Answer: $P(E_1 \cap E_2) = P(\emptyset) = 0$
$P(E_1 \cup E_2) = P(E_1) + P(E_2) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}$
$P(E_1 \cap E_3) = P(6,1),(6,3),(6,5) = \frac{3}{36} = \frac{1}{12}$
$P(E_1 \cup E_3) = P(E_1) + P(E_3) - P(E_1 \cap E_3) = \frac{1}{6} + \frac{1}{2} - \frac{1}{12} = \frac{7}{12}$

[*Notes on axioms:*
(1) In order to cope with infinite sequences of events, it is necessary to strengthen axiom 3 to
3'. $P(\cup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} P(E_i)$ for any sequence $(E_1, E_2, \cdots)$ of mutually exclusive events.
(2) When Ω is noncountably infinite, in order to make the theory rigorous it is usually necessary to restrict the class of events $\mathcal{F}$ to which probabilities are assigned.]

1.1.3 Conditional probability

Supose $P(E_2) \neq 0$. The **conditional probability** of the event E_1 given E_2 is defined as

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}\,.$$

The conditional probability is undefined if $P(E_2) = 0$. The conditional probability formula above yields the **multiplication rule**:

$$\begin{aligned} P(E_1 \cap E_2) &= P(E_1)P(E_2|E_1) \\ &= P(E_2)P(E_1|E_2) \end{aligned}$$

Independence

Events E_1 and E_2 are said to be **independent** if

$$P(E_1 \cap E_2) = P(E_1)P(E_2)\,.$$

Note that this implies that $P(E_1|E_2) = P(E_1)$ and $P(E_2|E_1) = P(E_2)$. Thus knowledge of the occurrence of one of the events does not affect the likelihood of occurrence of the other.
Events $E_1, \ldots, E_k$ are **pairwise independent** if $P(E_i \cap E_j) = P(E_i)P(E_j)$ for all $i \neq j$. They are **mutually independent** if for all subsets $P(\cap_j E_j) = \prod_j P(E_j)$. Clearly, mutual independence $\Rightarrow$ pairwise independence, but the converse is false (see question 4 of the self study exercises).

Example 1 (continued): Find $P(E_1|E_2)$ and $P(E_1|E_3)$. Are E_1, E_2 independent?

Answer: $P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = 0$, $P(E_1|E_3) = \frac{P(E_1 \cap E_3)}{P(E_3)} = \frac{1/12}{1/2} = \frac{1}{6}$

$P(E_1)P(E_2) \neq 0$ so $P(E_1 \cap E_2) \neq P(E_1)P(E_2)$ and thus E_1 and E_2 are not independent.

Law of total probability (partition law)

Suppose that $B_1, \ldots, B_k$ are **mutually exclusive** and **exhaustive** events (*i.e.* $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\cup_i B_i = \Omega$).
Let A be any event. Then

$$P(A) = \sum_{j=1}^{k} P(A|B_j)P(B_j)$$

Bayes' Rule

Suppose that events $B_1, \ldots, B_k$ are mutually exclusive and exhaustive and let A be any event. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Example 2: (Cancer diagnosis) A screening programme for a certain type of cancer has reliabilities $P(A|D) = 0.98$, $P(A|\bar{D}) = 0.05$, where D is the event "disease is present" and A is the event "test gives a positive result". It is known that 1 in 10, 000 of the population has the disease. Suppose that an individual's test result is positive. What is the probability that that person has the disease?

Answer: We require $P(D|A)$. First find $P(A)$.

$P(A) = P(A|D)P(D) + P(A|\bar{D})P(\bar{D}) = 0.98 \times 0.0001 + 0.05 \times 0.9999 = 0.050093$.

By Bayes' rule; $P(D|A) = \frac{P(A|D)P(D)}{P(A)} = \frac{0.0001 \times 0.98}{0.050093} = 0.002$.

The person is still very unlikely to have the disease even though the test is positive.

Example 3: (Bertrand's Box Paradox) Three indistinguishable boxes contain black and white beads as shown: [ww], [wb], [bb]. A box is chosen at random and a bead chosen at random from the selected box. What is the probability of that the [wb] box was chosen given that selected bead was white?

Answer: $E \equiv$ 'chose the [wb] box', $W \equiv$ 'selected bead is white'. By the partition law: $P(W) = 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2}$. Now using Bayes' rule $P(E|W) = \frac{P(E)P(W|E)}{P(W)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}$ (i.e. even though a bead from the selected box has been seen, the probability that the box is [wb] is still $\frac{1}{3}$).

1.1.4 Self-study exercises

1. Consider families of three children, a typical outcome being bbg (boy-boy-girl in birth order) with probability $\frac{1}{8}$. Find the probabilities of

   (i) 2 boys and 1 girl (any order),

   (ii) at least one boy,

   (iii) consecutive children of different sexes.

   Answer: (i) $\frac{3}{8}$; (ii) $\frac{7}{8}$; (iii) $\frac{1}{4}$.

2. Use $p_A = P(A)$, $p_B = P(B)$ and $p_{AB} = P(A \cap B)$ to obtain expressions for:

   (a) $P(\bar{A} \cup \bar{B})$,

   (b) $P(\bar{A} \cap B)$,

   (c) $P(\bar{A} \cup B)$,

   (d) $P(\bar{A} \cap \bar{B})$,

   (e) $P((A \cap \bar{B}) \cup (B \cap \bar{A}))$.

   Describe each event in words. (Use a Venn diagram.)

   Answer: (a) $1-p_{AB}$; (b) p_B-p_{AB}; (c) $1-p_A+p_{AB}$; (d) $1-p_A-p_B+p_{AB}$; (e) $p_A+p_B-2p_{AB}$.

3. (i) Express $P(E_1 \cup E_2 \cup E_3)$ in terms of the probabilities of E_1, E_2, E_3 and their intersections only. Illustrate with a sketch.

   (ii)Three types of fault can occur which lead to the rejection of a certain manufactured item. The probabilities of each of these faults (A, B and C)

occurring are 0.1, 0.05 and 0.04 respectively. The three faults are known to be interrelated; the probability that A & B both occur is 0.04, A & C 0.02, and B & C 0.02. The probability that all three faults occur together is 0.01.

What percentage of items are rejected?

Answer: (i) $P(E_1)+P(E_2)+P(E_3)-P(E_1\cap E_2)-P(E_1\cap E_3)-P(E_2\cap E_3)+P(E_1\cap E_2\cap E_3)$

(ii) $P(A\cup B\cup C) = .01 + .05 + .04 - (.04 + .02 + .02) + .01 = 0.12$

4. Two fair dice rolled: 36 possible outcomes each with probability $\frac{1}{36}$. Let $E_1 = \{\text{odd face 1st}\}, E_2 = \{\text{odd face 2nd}\}, E_3 = \{\text{one odd, one even}\}$, so $P(E_1) = \frac{1}{2}, P(E_2) = \frac{1}{2}, P(E_3) = \frac{1}{2}$. Show that E_1, E_2, E_3 are pairwise independent, but not mutually independent.

   Answer: $P(E_2|E_1) = \frac{1}{2} = P(E_2)$, $P(E_3|E_1) = \frac{1}{2} = P(E_3)$, $P(E_3|E_2) = \frac{1}{2} = P(E_3)$, so E_1, E_2, E_3 are pairwise independent. But $P(E_1\cap E_2\cap E_3) = 0 \neq P(E_1)P(E_2)P(E_3)$, so E_1, E_2, E_3 are not mutually independent.

5. An engineering company uses a 'selling aptitude test' to aid it in the selection of its sales force. Past experience has shown that only 65% of all persons applying for a sales position achieved a classification of 'satisfactory' in actual selling and of these 80% had passed the aptitude test. Only 30% of the 'unsatisfactory' persons had passed the test.

   What is the probability that a candidate would be a 'satisfactory' salesperson given that they had passed the aptitude test?

   Answer: A = pass aptitude test, S = satisfactory. $P(S) = 0.65$, $P(A|S) = 0.8$, $P(A|\bar{S}) = 0.3$. Therefore $P(A) = (0.65\times 0.8) + (0.35\times 0.3) = 0.625$ so $P(S|A) = P(S)P(A|S)/P(A) = (0.65\times 0.8)/0.625 = 0.832$.

1.2 Random variables and probability distributions

1.2.1 Random variables

A random variable X is a real-valued function on the sample space Ω; that is, $X : \Omega \to \mathcal{R}$. If P is a probability measure on $(\Omega, \mathcal{F})$ then the induced probability measure on $\mathcal{R}$ is called the **probability distribution** of X.

A **discrete** random variable X takes values $x_1, x_2, \ldots$ with probabilities $p(x_1), p(x_2), \ldots$, where $p(x) = \text{pr}(X = x) = P(\{\omega : X(\omega) = x\})$ is the **probability mass function** (pmf) of X. (E.g. X = place of horse in race, grade of egg.)

Example 4: (i) Toss a coin twice: outcomes HH, HT, TH, TT. The random variable X = number of heads, takes values 0, 1, 2.
(ii) Roll two dice: X = total score. Probabilities for X are $\text{P}(X = 2) = \frac{1}{36}$, $\text{P}(X = 3) = \frac{2}{36}$, $\text{P}(X = 4) = \frac{3}{36}$ *etc.*

Example 5: X takes values $1, 2, 3, 4, 5$ with probabilities $k, 2k, 3k, 4k, 5k$. Calculate k and P($2 \leq X \leq 4$).

Answer: $1 = \sum_{x=1}^{5} P(x) = k(1 + 2 + 3 + 4 + 5) = 15k$ so $k = \frac{1}{15}$.
$P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = \frac{2}{15} + \frac{3}{15} + \frac{4}{15} = \frac{3}{5}$.

A **continuous** random variable X takes values over an interval. *E.g.* X = time over racecourse, weight of egg. Its **probability density function** (pdf) $f(x)$ is defined by

$$\text{pr}(a < X < b) = \int_a^b f(x)dx\,.$$

Note that $f(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example 6: Let $f(x) = k(1 - x^2)$ on $(-1, 1)$. Calculate k and $\text{pr}(|X| > 1/2)$.

Answer: $1 = \int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{1} k(1 - x^2)dx = k[x - \frac{1}{3}x^3]_{-1}^{1} = \frac{4k}{3} \Rightarrow k = \frac{3}{4}$
$P(|X| > 1/2) = 1 - P(-\frac{1}{2} \leq X \leq \frac{1}{2}) = 1 - \int_{-\frac{1}{2}}^{\frac{1}{2}} k(1 - x^2)dx = 1 - \frac{11k}{12} = \frac{5}{16}$

A **mixed** discrete/continuous random variable is such that the probability is shared

between discrete and continuous components with $\sum p(x) + \int f(x)dx = 1$, *e.g.* rainfall on given day, waiting time in queue, flow in pipe, contents of reservoir.
The **distribution function** F of the random variable X is defined as

$$F(x) = \text{pr}(X \leq x) = P(\{\omega : X(\omega) \leq x\}).$$

Thus $F(-\infty) = 0, F(\infty) = 1, F$ is monotone increasing, and pr$(a < X \leq b) = F(b) - F(a)$.

Discrete case: $F(x) = \sum_{u \leq x} p(u)$

Continuous case: $F(x) = \int_{-\infty}^{x} f(u)du$ and $F'(x) = f(x)$.

1.2.2 Expectation

The **expectation** (or **expected value** or **mean**) of the random variable X is defined as

$$\mu = E(X) = \begin{cases} \sum xp(x) & X \text{ discrete} \\ \\ \int xf(x)dx & X \text{ continuous} \end{cases}$$

The **Variance** of X is $\sigma^2 = \text{Var}(X) = E\{(X - \mu)^2\}$. Equivalently $\sigma^2 = E(X^2) - \{E(X)\}^2$ (exercise: prove).

σ is called the **standard deviation**.

Functions of X:

(i) $E\{h(X)\} = \begin{cases} \sum h(x)p(x) & X \text{ discrete} \\ \\ \int h(x)f(x)dx & X \text{ continuous} \end{cases}$

(ii) $E(aX + b) = aE(X) + b, \ \text{Var}(aX + b) = a^2\text{Var}(X)$.

Proof (for discrete X)

(i) $h(X)$ takes values $h(x_1), h(x_2), \ldots$ with probabilities $p(x_1), p(x_2), \ldots$, so, by definition, $E\{h(X)\} = h(x_1)p(x_1) + h(x_2)p(x_2) + \cdots = \sum h(x)p(x)$.

(ii) $E[aX + b] = \sum(aX + b)P(x) = a\sum xP(x) + b\sum P(x) = aE[X] + b$

$\text{Var}[aX+b] = E[(aX+b) - E[aX+b]^2] = E[aX+b-aE[X]-b^2] = E[a^2(X-E[X])^2] = a^2\text{Var}[X]$

Example 7: $X = 0, 1, 2$ with probabilities $1/4, 1/2, 1/4$.
Find $E(X),\ E(X-1),\ E(X^2)$ and Var(X).

Answer: $E[X] = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$

$E[X-1] = E[X] - 1$, $E[X^2] = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = \frac{3}{2}$

$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{1}{2}$.

Example 8: $f(x) = k(1+x)^{-4}$ on $(0, \infty)$. Find k and hence obtain $E(X),\ E\{(1+X)^{-1}\},\ E(X^2)$ and Var(X).

Answer: $1 = k\int_0^\infty (1+x)^{-4}dx = k[-\frac{1}{3}(1+x)^{-3}]_0^\infty = \frac{k}{3} \Rightarrow k = 3$

$E[X] = 3\int_0^\infty x(1+x)^{-4}dx = 3\int_1^\infty (u-1)u^{-4}du = 3[-\frac{1}{2}u^{-2} + \frac{1}{3}u^{-3}]_1^\infty = 3(\frac{1}{2} - \frac{1}{3}) = \frac{1}{2}$

$E[(1+X)^{-1}] = 3\int_0^\infty (1+x)^{-5}dx = 3[-\frac{1}{4}(1+x)^{-4}]_0^\infty = \frac{3}{4}$

$E[X^2] = 3\int x^2(1+x)^{-4}dx = 3\int_1^\infty (u-1)^2u^{-4}du = 3[-u^{-1}+u^{-2}-\frac{1}{3}u^{-3}]_1^\infty = 1$

$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{3}{4}$.

1.2.3 Self-study exercises

1. X takes values 0, 1, 2, 3 with probabilities $\frac{1}{4}$, $\frac{1}{5}$, $\frac{3}{10}$, $\frac{1}{4}$. Compute (as fractions) $E(X)$, $E(2X+3)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(2X+3)$.

   Answer: $E(X) = \frac{31}{20}$, $E(2X+3) = 2E(X)+3 = \frac{61}{10}$, $E(X^2) = \frac{73}{20}$, so $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{499}{400}$, $\mathrm{Var}(2X+3) = 4\mathrm{Var}(X) = \frac{499}{100}$.

2. The random variable X has density function $f(x) = kx(1-x)$ on (0,1), $f(x) = 0$ elsewhere. Calculate k and sketch $f(x)$. Compute the mean and variance of X, and pr $(0.3 \leq X \leq 0.6)$.

   Answer: $k = 6$, $E(X) = \frac{1}{2}$, $\mathrm{Var}(X) = \frac{1}{20}$, pr $(0.3 \leq X \leq 0.6) = 0.432$.

Chapter 2

Random variables and distributions

2.1 Transformations

Suppose that X has distribution function $F_X(x)$ and that the distribution function $F_Y(y)$ of $Y = h(X)$ is required, where h is a strictly increasing function. Then

$$F_Y(y) = \text{pr}(Y \leq y) = \text{pr}(h(X) \leq y) = \text{pr}(X \leq x) = F_X(x)$$

where $x \equiv x(y) = h^{-1}(y)$. If X is continuous and h is differentiable, then it follows that Y has density

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dy} = f_X(x)\frac{dx}{dy}\,.$$

On the other hand, if h is strictly decreasing then

$$F_Y(y) = \text{pr}(Y \leq y) = \text{pr}(h(X) \leq y) = \text{pr}(X \geq x) = 1 - F_X(x)$$

which yields $f_Y(y) = -f_X(x)(dx/dy)$. Both formulae are covered by

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|\,.$$

Example 9: Suppose that X has pdf $f_X(x) = 2e^{-2x}$ on $(0, \infty)$. Obtain the pdf of $Y = \log X$.

Probability integral transform. Let X be a continuous random variable with distribution function $F(x)$. Then $Y = F(X)$ is uniformly distributed on $(0, 1)$.
Proof. First note that $0 \leq Y \leq 1$. Let $0 \leq y \leq 1$; then

$$\text{pr}(Y \leq y) = \text{pr}(F(X) \leq y) = \text{pr}(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y,$$

so Y has pdf $f(y) = 1$ on $(0, 1)$ (by differentiation), which is the density of the uniform distribution on $(0, 1)$.
This result has an important application to the simulation of random variables:

2.1.1 Self-study exercises

1. X takes values 1, 2, 3, 4 with probabilities $\frac{1}{10}$, $\frac{1}{5}$, $\frac{3}{10}$, $\frac{2}{5}$ and $Y = (X-2)^2$.

   (i) Find $E(Y)$ and $\text{Var}(Y)$ using the formula for $E\{h(X)\}$.

   (ii) Calculate the pmf of Y and use it to calculate $E(Y)$ and $\text{Var}(Y)$ directly.

2. The random variable X has pdf $f(x) = \frac{1}{3}$, $x = 1, 2, 3$, zero elsewhere. Find the pdf of $Y = 2X + 1$.

3. The random variable X has pdf $f(x) = e^{-x}$ on $(0, \infty)$. Obtain the pdf of $Y = e^X$.

4. Let X have the pdf $f(x) = \left(\frac{1}{2}\right)^x$, $x = 1, 2, 3, \ldots$, zero elsewhere. Find the pdf of $Y = X^3$.

2.2 Some standard discrete distributions

2.2.1 Binomial distribution

Consider a sequence of independent trials in each of which there are only two possible results, 'success', with probability π, or 'failure', with probability $1 - \pi$ (**independent Bernoulli trials**).
Outcomes can be represented as binary sequences, with 1 for success and 0 for failure, *e.g.* 110001 has probability $\pi\pi(1-\pi)(1-\pi)(1-\pi)\pi$, since the trials are independent.
Let the random variable X be the number of successes in n trials, with n fixed. The probability of a particular sequence of r 1's and $n - r$ 0's is $\pi^r(1-\pi)^{n-r}$, and the event $\{X = r\}$ contains $\binom{n}{r}$ such sequences. Hence

$$p(r) = \text{pr}(X = r) = \binom{n}{r} \pi^r (1-\pi)^{n-r},\ r = 0, 1, \ldots, n\,.$$

This is the pmf of the **binomial** (n, π) **distribution**. The name comes from the binomial theorem

$$\{\pi + (1-\pi)\}^n = \sum_{r=0}^{n} \binom{n}{r} \pi^r (1-\pi)^{n-r}\,,$$

from which $\sum_r p(r) = 1$ follows.
The mean is $\mu = n\pi$:

The variance is $\sigma^2 = n\pi(1-\pi)$ (see exercise 3).

Example 10: A biased coin with pr(head) $= 2/3$ is tossed five times. Calculate $p(r)$.

2.2.2 Geometric distribution

Suppose now that, instead of a fixed number of Bernoulli trials, one continues until a success is achieved, so that the number of trials, N, is now a random variable. Then N takes the value n if and only if the previous $(n-1)$ trials result in failures and the nth trial results in a success. Thus

$$p(n) = \text{pr}(N = n) = (1-\pi)^{n-1}\pi,\ n = 1, 2, \ldots .$$

This is the pmf of the **geometric** (π) **distribution**: the probabilities are in geometric progression. Note that the sum of the probabilities over $n = 1, 2, \ldots$ is 1.

The mean is $\mu = 1/\pi$:

The variance is $\sigma^2 = (1-\pi)/\pi^2$ (see exercise 4).
Eg. Toss a biased coin with pr(head) $= 2/3$. Then, on average, it takes three tosses to get a tail.

2.2.3 Poisson distribution

The pmf of the **Poisson** (λ) **distribution** is defined as

$$p(r) = \frac{e^{-\lambda}\lambda^r}{r!},\ r = 0, 1, 2, \ldots,$$

where $\lambda > 0$. Note that the sum of the probabilities over $r = 0, 1, 2, \ldots$ is 1 (exponential series).
The mean is $\mu = \lambda$:

The variance is $\sigma^2 = \lambda$ (see exercise 6).
The Poisson distribution arises in various contexts, one being the limit of a binomial(n, π) as $n \to \infty$ and $\pi \to 0$ with $n\pi = \lambda$ fixed.

Example 11: (Random events in time.) Cars are recorded as they pass a checkpoint. The probability π that a car is level with the checkpoint at any given instant is very small, but the number n of such instants in a given time period is large. Hence X_t, the number of cars passing the checkpoint during a time interval t minutes, can be modelled as Poisson with mean proportional to t. For example, if the average rate is two cars per minute, find the probability of exactly 3 cars in 5 minutes.

2.2.4 Self-study exercises

1. In a large consignment of widgets 5% are defective. What is the probability of getting one or two defectives in a four-pack?

2. X is binomial with mean 2 and variance 1. Compute pr$(X \leq 1)$.

3. Derive the variance of the binomial (n, π) distribution.

   [Hint: find $E\{X(X-1)\}$.]

4. Derive the variance of the geometric (π) distribution.

   [Hint: find $E\{X(X-1)\}$.]

5. A leaflet contains one thousand words and the probability that any one word contains a misprint is 0.005. Use the Poisson distribution to estimate the probability of 2 or fewer misprints.

6. Derive the variance of the Poisson (λ) distribution.

   [Hint: find $E\{X(X-1)\}$.]

2.3 Some standard continuous distributions

2.3.1 Uniform distribution

The pdf of the **uniform** (α, β) **distribution** is

$$f(x) = (\beta - \alpha)^{-1},\ \alpha < x < \beta\,.$$

The mean is $\mu = (\beta + \alpha)/2$:

The variance is $\sigma^2 = (\beta - \alpha)^2/12$ (see exercise 1).

Application. Simulation of continuous random variables via the probability integral transform: see Section 2.1.

2.3.2 Exponential distribution

The pdf of the **exponential** (λ) **distribution** is

$$f(x) = \lambda e^{-\lambda x},\ x > 0,$$

where $\lambda > 0$. The distribution function is $F(x) = \int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}$ (verify).

The mean is $\mu = 1/\lambda$:

The variance is $\sigma^2 = 1/\lambda^2$ (see exercise 4).

Lack of memory property.

$$\text{pr}(X > a + b | X > a) = \text{pr}(X > b)$$

Proof:

For example, if the lifetime of a component is exponentially distributed, then the fact that it has lasted for 100 hours does not affect its chances of failing during the next 100 hours. That is, the component is not subject to *ageing*.

Application to random events in time.
Example: cars passing a checkpoint. The distribution of the **waiting time**, T, for the first event can be obtained as follows:

$$\text{pr}(T > t) = \text{pr}(N_t = 0) = e^{-\lambda t},$$

since N_t, the number of events occurring during the time interval $(0, t)$, has a Poisson distribution with mean λt. Hence T has distribution function $F(t) = 1 - e^{-\lambda t}$, that of the exponential (λ) distribution.

2.3.3 Pareto distribution

The **Pareto** (α, β) **distribution** has pdf

$$f(x) = \frac{\alpha}{\beta(1 + \frac{x}{\beta})^{\alpha+1}} \ , \ x > 0 \, ,$$

where $\alpha > 0$ and $\beta > 0$. The distribution function is $F(x) = 1 - (1 + \frac{x}{\beta})^{-\alpha}$ (verify).

The mean is $\mu = \beta/(\alpha - 1)$ for $\alpha > 1$:

The variance is $\sigma^2 = \alpha\beta^2/\{(\alpha - 1)^2(\alpha - 2)\}$ for $\alpha > 2$.

2.3.4 Self-study exercises

1. Obtain the variance of the uniform (α, β) distribution.

2. The lifetime of a valve has an exponential distribution with mean 350 hours. What proportion of valves will last 400 hours or longer? For how many hours should the valves be guaranteed so that only 1% are returned under guarantee?

3. A machine suffers random breakdowns at a rate of three per day. Given that it is functioning at 10am what is the probability that

   (i) no breakdown occurs before noon?

   (ii) the first breakdown occurs between 12pm and 1pm?

4. Obtain the variance of the exponential (λ) distribution.

5. The random variable X has the Pareto distribution with $\alpha = 3\,, \beta = 1$. Find the probability that X exceeds $\mu + 2\sigma$, where μ, σ are respectively the mean and standard deviation of X.

2.4 The normal (Gaussian) distribution

2.4.1 Normal distribution

The **normal distribution** is the most important distribution in Statistics, for both theoretical and practical reasons. Its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty .$$

The parameters μ and σ^2 are the mean and variance respectively. The distribution is denoted by $N(\mu, \sigma^2)$.
Mean:

The importance of the normal distribution follows from its use as an approximation in various statistical methods (consequence of Central Limit Theorem: see Section 3.4.2), its convenience for theoretical manipulation, and its application to describe observed data.

Standard normal distribution
The **standard normal distribution** is $N(0, 1)$, for which the distribution function has the special notation $\Phi(x)$. Thus

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du .$$

The function Φ is tabulated widely (*e.g.* New Cambridge Statistical Tables). Useful values are $\Phi(1.64) = 0.95, \Phi(1.96) = 0.975$.

Example 12: Suppose that X is $N(0, 1)$ and Y is $N(2, 4)$. Use tables to calculate pr$(X < 1)$, pr$(X < -1)$, pr$(-1.5 < X < -0.5)$, pr$(Y < 1)$ and pr$(Y^2 > 5Y - 6)$.

2.4.2 Properties

(i) If X is $N(\mu, \sigma^2)$ then $aX + b$ is $N(a\mu + b, a^2\sigma^2)$.
In particular, the **standardized variate** $(X - \mu)/\sigma$ is $N(0, 1)$.

(ii) if X_1 is $N(\mu_1, \sigma_1^2), X_2$ is $N(\mu_2, \sigma_2^2)$ and X_1 and X_2 are independent, then $X_1 + X_2$ is

$$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

[Hence, from property (i), the distribution of $X_1 - X_2$ is $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.]

(iii) If $X_i,\ i = 1, \ldots, n$, are independent $N(\mu_i, \sigma_i^2)$, then $\sum_i X_i$ is $N(\sum_i \mu_i, \sum_i \sigma_i^2)$.

(iv) The **moment generating function** (see Section 2.6.3) of $N(\mu, \sigma^2)$ is $M(z) = E(e^{zX}) = e^{\mu z + \frac{1}{2}\sigma^2 z^2}$.
(Properties (i) - (iii) are easily proved via mgfs - see Section 2.6.3.)

(v) Central moments of $N(\mu, \sigma^2)$. Let $\mu_r = E\{(X - \mu)^r\}$, the rth **central moment** of X. Then

$$\mu_r = 0 \text{ for } r \text{ odd}, \quad \mu_r = (\sigma/\sqrt{2})^r r!/(r/2)! \text{ for } r \text{ even}.$$

Note that $\mu_2 = \sigma^2$, the variance of X.

Sampling distribution of the sample mean
Let $X_1, \ldots, X_n$ be independently and identically distributed (iid) as $N(\mu, \sigma^2)$. Then the distribution of $\bar{X} = n^{-1} \sum X_i$ is $N(\mu, n^{-1}\sigma^2)$. This is the **sampling distribution** of the sample mean, a result of fundamental importance in Statistics.

Proof:

2.4.3 Self-study exercises

1. The distribution of lengths of rivets is normal with mean 2.5cm and sd 0.02cm. In a batch of 500 rivets how many would you expect on average to have length

   (i)less than 2.46cm,

   (ii) between 2.46cm and 2.53cm,

   (iii) greater than 2.53cm?

   (iv) What length is exceeded by only 1 in 1000 rivets?

2. Suppose that X is $N(0, 1)$ and Y is $N(2, 4)$. Use tables to calculate pr$(Y - X < 1)$ and pr$(X + \frac{1}{2}Y > 1.5)$.

3. Two resistors in series have resistances X_1 and X_2 ohms, where X_1 is $N(200, 4)$ and X_2 is $N(150, 3)$. What is the distribution of the combined resistance $X = X_1 + X_2$? Find the probability that X exceeds 355.5 ohms.

4. The fuel consumption of a fleet of 150 lorries is approximately normally distributed with mean 15 mpg and sd 1.5 mpg.

   (i) Compute the expected number of lorries that average between 13 and 14 mpg.

   (ii) What is the probability that the average of a random sample of four lorries exceeds 16 mpg?

2.5 Bivariate distributions

2.5.1 Definitions and notation

Suppose that X_1, X_2 are two random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$. Then P induces a **joint distribution** for X_1, X_2. The **joint distribution function** is defined as

$$\begin{aligned} F(x_1, x_2) &= P(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2\}) \\ &= \text{pr}(X_1 \leq x_1, X_2 \leq x_2)\,. \end{aligned}$$

In the discrete case the **joint pmf** is $p(x_1, x_2) = \text{pr}(X_1 = x_1, X_2 = x_2)$. In the continuous case, the **joint pdf** is $f(x_1, x_2) = \frac{\partial F(x_1, x_2)}{\partial x_1 \partial x_2}$.

Example 13: (discrete) Two biased coins are tossed. Score heads = 1 (with probability π), tails = 0 (with probability $1 - \pi$). Let X_1 = sum of scores, X_2 = difference of scores (1st - 2nd). The tables below show

(i) the possible values of X_1, X_2 and their probabilities,

(ii) the joint probability table for X_1, X_2.

(i)

Outcome	00	01	10	11
X_1				
X_2				
Prob				

(ii)

		X_2 -1	0	1
	0			
X_1	1			
	2			

Example 14: (continuous) Suppose X_1 and X_2 have joint pdf $f(x_1, x_2) = k(1 - x_1 x_2^2)$ on $(0, 1)^2$. Obtain the value of k.

2.5.2 Marginal distributions

These follow from the law of total probability.
Discrete case. Marginal probability mass functions

$$p_1(x_1) = \text{pr}(X_1 = x_1) = \sum_{x_2} p(x_1, x_2) \text{ and } p_2(x_2) = \text{pr}(X_2 = x_2) = \sum_{x_1} p(x_1, x_2)$$

Continuous case. Marginal probability density functions

$$f_1(x_1) = \int f(x_1, x_2) dx_2 \text{ and } f_2(x_2) = \int f(x_1, x_2) dx_1$$

Marginal means and variances. $\mu_1 = E(X_1) = \sum x_1 p_1(x_1)$ (discrete) or $\int x_1 f_1(x_1) dx_1$ (continuous)

$\sigma_1^2 = \text{var}(X_1) = E\{(X_1 - \mu_1)^2\} = E(X_1^2) - \mu_1^2$

Likewise μ_2 and σ_2^2.

2.5.3 Conditional distributions

These follow from the definition of conditional probability.
Discrete case. Conditional probability mass function of X_1 given X_2 is

$$\begin{aligned} p_1(x_1 | X_2 = x_2) &= \text{pr}(X_1 = x_1 | X_2 = x_2) \\ &= \frac{\text{pr}(X_1 = x_1, X_2 = x_2)}{\text{pr}(X_2 = x_2)} = \frac{p(x_1, x_2)}{p_2(x_2)}. \end{aligned}$$

Similarly

$$p_2(x_2 | X_1 = x_1) = \frac{p(x_1, x_2)}{p_1(x_1)}.$$

Continuous case. Conditional probability density function of X_1 given X_2 is

$$f_1(x_1 | X_2 = x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}.$$

Similarly

$$f_2(x_2 | X_1 = x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}.$$

Independence. X_1 and X_2 are said to be **independent** if $F(x_1, x_2) = F_1(x_1) F_2(x_2)$. Equivalently, $p(x_1, x_2) = p_1(x_1) p_2(x_2)$ (discrete), or $f(x_1, x_2) = f_1(x_1) f_2(x_2)$ (continuous).

Example 15: Suppose that R and N have a joint distribution in which $R|N$ is binomial (N, π) and N is Poisson (λ). Show that R is Poisson $(\lambda\pi)$.

2.5.4 Covariance and correlation

The **covariance** between X_1 and X_2 is defined as

$$\sigma_{12} = \text{Cov}(X_1, X_2) = E\{(X_1 - \mu_1)(X_2 - \mu_2)\} = E(X_1 X_2) - \mu_1 \mu_2 \,,$$

where $E(X_1 X_2) = \sum x_1 x_2 p(x_1, x_2)$ (discrete) or $\int x_1 x_2 f(x_1, x_2) dx_1 dx_2$ (continuous).

The **correlation** between X_1 and X_2 is

$$\rho = \text{Corr}(X_1, X_2) = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \,.$$

Example 13: (continued)

Marginal distributions:

$x_1 = 0, 1, 2$ with $p_1(x_1) =$

$x_2 = -1, 0, 1$ with $p_2(x_2) =$

Marginal means:

$\mu_1 = \sum x_1 p_1(x_1) =$

$\mu_2 = \sum x_2 p_2(x_2) =$

Variances:

$\sigma_1^2 = \sum x_1^2 p_1(x_1) - \mu_1^2 =$

$\sigma_2^2 = \sum x_2^2 p_2(x_2) - \mu_2^2 =$

Conditional distributions: *e.g.*

$$p(x_1|X_2 = 0) = \begin{cases} & x_1 = 0 \\ & \\ & x_1 = 2 \end{cases}$$

Independence: *e.g.* $p(1,0) = 0$ but $p_1(0)p_2(1) \neq 0$, so X_1, X_2 are not independent.

Covariance: $\sigma_{12} = \sum x_1 x_2 p(x_1, x_2) - \mu_1\mu_2 =$

Example 14: (continued)

Marginal distributions:

$f_1(x_1) = \int_0^1 k(1 - x_1 x_2^2) dx_2 =$

$f_2(x_2) = \int_0^1 k(1 - x_1 x_2^2) dx_1 =$

Marginal means:

$\mu_1 = \int_0^1 x_1 f_1(x_1) dx_1 =$

$\mu_2 = \int_0^1 x_2 f_2(x_2) dx_2 =$

Variances:

$\sigma_1^2 = \int_0^1 x_1^2 f_1(x_1) dx_1 - \mu_1^2 =$

$\sigma_2^2 = \int_0^1 x_2^2 f_2(x_2) dx_2 - \mu_2^2 =$

Conditional distributions: *e.g.*

$f(x_2|X_1 = \frac{1}{3}) =$

Independence:

$f(x_1, x_2) = k(1 - x_1 x_2^2)$, which does not factorise into $f_1(x_1)f_2(x_2)$ so X_1, X_2 are not independent.

Covariance:

$\sigma_{12} = \int x_1 x_2 f(x_1, x_2) dx_1 dx_2 - \mu_1\mu_2$

Properties

(i) $E(aX_1 + bX_2) = a\mu_1 + b\mu_2,\ \text{Var}(aX_1 + bX_2) = a^2\sigma_1^2 + 2ab\sigma_{12} + b^2\sigma_2^2$

$\text{Cov}(aX_1 + b, cX_2 + d) = ac\sigma_{12}$, $\text{Corr}(aX_1 + b, cX_2 + d) =$ $\text{Corr}(X_1, X_2)$

(note: invariance under linear transformation)

Proof:

(ii) X_1, X_2 independent $\Rightarrow$ $\text{Cov}(X_1, X_2) = 0$. The converse is **false**.

Proof:

(iii) $-1 \leq$ $\text{Corr}(X_1, X_2) \leq +1$, with equality if and only if X_1, X_2 are linearly dependent.

Proof:

(iv) $E(Y) = E\{E(Y|X)\}$ and $\text{Var}(Y) = E\{\text{Var}(Y|X)\} + \text{Var}\{E(Y|X)\}$

Proof:

2.5.5 Self-study exercises

1. Roll a fair die twice. Let X_1 be the number of times that face 1 shows, and let $X_2 = $ [sum of faces/4], where $[x]$ denotes the integer part of x.

   (a) Construct the joint probability table.

   (b) Calculate the two marginal pmfs $p_1(x_1)$ and $p_2(x_2)$ and the conditional pmfs $p_1(x_1|x_2 = 1)$ and $p_2(x_2|x_1 = 1)$. Are X_1 and X_2 independent?

   (c) Compute the means, μ_1 and μ_2, variances, σ_1^2 and σ_2^2, and covariance σ_{12}. Are X_1 and X_2 uncorrelated?

2. X_1 and X_2 have joint density $f(x_1, x_2) = 4x_1x_2$ for $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 1$. Calculate the marginal and conditional densities of X_1 and X_2, their means and variances, and their correlation.

3. Calculate, in terms of the means, variances and covariances of X_1, X_2 and X_3, $E(2X_1 + 3X_2)$, $\text{Cov}(2X_1, 3X_2)$, $\text{Var}(2X_1 + 3X_2)$ and $\text{Cov}(2X_1 + 3X_2, 4X_2 + 5X_3)$.

2.6 Generating functions

2.6.1 General

The generating function for a sequence $(a_n : n \geq 0)$ is $A(z) = a_0 + a_1 z + a_2 z^2 + \cdots = \sum_{n=0}^{\infty} a_n z^n$. Here z is a dummy variable. The definition is useful only if the series converges. The idea is to replace the sequence (a_n) by the function $A(z)$, which may be easier to analyse than the original sequence.

Examples:
(i) If $a_n = 1$ for $n = 0, 1, 2, \ldots$, then $A(z) = (1-z)^{-1}$ for $|z| < 1$ (geometric series).
(ii) If $a_n = \begin{pmatrix} m \\ n \end{pmatrix}$ for $n = 0, 1, \ldots, m$, and $a_n = 0$ for $n > m$, then $A(z) = (1+z)^m$ (binomial series).

2.6.2 Probability generating function

Let (p_n) be the pmf of some discrete random variable X, so $p_n = \text{pr}(X = n) \geq 0$ and $\sum_n p_n = 1$. Define the **probability generating function** (pgf) of X by

$$P(z) = E(z^X) = \sum_n p_n z^n .$$

Properties
(i) $|P(z)| \leq 1$ for $|z| \leq 1$.
Proof:

(ii) $\mu = E(X) = P'(1)$.
Proof:

(iii) $\sigma^2 = \text{ Var}(X) = P''(1) + P'(1) - \{P'(1)\}^2$.
Proof:

(iv) Let X and Y be independent random variables with pgfs P_X and P_Y respectively. Then the pgf of $X + Y$ is given by $P_{X+Y}(z) = P_X(z)P_Y(z)$.
Proof:

Example 16: (i) Find the pgf of the Poisson (λ) distribution.
(ii) Let X_1, X_2 be independent Poisson random variables with parameters λ_1, λ_2 respectively. Obtain the distribution of $X_1 + X_2$.

2.6.3 Moment generating function

The **moment generating function** (mgf) is defined as

$$M(z) = E(e^{zX}) .$$

The pgf tends to be used more for discrete distributions, and the mgf for continuous ones, although note that the two are related by $M(z) = P(e^z)$.

Properties

(i) $\mu = E(X) = M'(0)$, $\sigma^2 =$ Var$(X) = M''(0) - \mu^2$.

Proof:

(ii) Let X and Y be independent random variables with mgfs $M_X(z)$, $M_Y(z)$ respectively. Then the mgf of $X + Y$ is given by $M_{X+Y}(z) = M_X(z)M_Y(z)$.

Proof:

Normal distribution. We prove properties (i) - (iv) of Section 2.4.2.

2.6.4 Self-study exercises

1. Show that the pgf of the binomial (n, π) distribution is $\{\pi z + (1 - \pi)\}^n$.

2. (Zero-truncated Poisson distribution) Find the pgf of the discrete distribution with pmf $p(r) = e^{-\lambda}\lambda^r/\{r!(1 - e^{-\lambda})\}$ for $r = 1, 2, \ldots$. Deduce the mean and variance.

3. The random variable X has density $f(x) = k(1 + x)e^{-\lambda x}$ on $(0, \infty)$ with $\lambda > 0$. Find the value of k. Show that the moment generating function $M(z) = k\{(z - \lambda)^{-2} - (z - \lambda)^{-1}\}$. Use it to calculate the mean and standard deviation of X.

Chapter 3

Further distribution theory

3.1 Multivariate distributions

Let $X_1, \ldots, X_p$ be p real-valued random variables on $(\Omega, \mathcal{F})$ and consider the joint distribution of $X_1, \ldots, X_p$. Equivalently, consider the distribution of the **random vector**

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_p \end{pmatrix}$$

3.1.1 Definitions

The **joint distribution function**

$$F(\underline{x}) = \text{pr}(\underline{X} \leq \underline{x}) = \text{pr}(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

The **joint probability mass function** (pmf)

$$p(\underline{x}) = \text{pr}(\underline{X} = \underline{x}) = \text{pr}(X_1 = x_1, \ldots, X_p = x_p)$$

(discrete case)
The **joint probability density function** (pdf) $f(\underline{x})$ is such that

$$\text{pr}(\underline{X} \in A) = \int_A f(\underline{x}) d\underline{x}$$

(continuous case)

The **marginal distributions** are those of the individual components:

$F_j(x_j) = \text{pr}(X_j \le x_j) = F(\infty, \ldots, x_j, \ldots, \infty)$

The **conditional distributions** are those of one component given another:

$F(x_j|x_k) = \text{pr}(X_j \le x_j | X_k = x_k)$

The X_js are **independent** if $F(\underline{x}) = \prod_j F_j(x_j)$. Equivalently, $p(\underline{x}) = \prod_j p_j(x_j)$ (discrete case), or $f(\underline{x}) = \prod_j f_j(x_j)$ (continuous case).

Means: $\mu_j = E(X_j)$
Variances: $\sigma_j^2 = \text{Var}(X_j) = E\{(X_j - \mu_j)^2\} = E(X_j^2) - \mu_j^2$
Covariances: $\sigma_{jk} = \text{Cov}(X_j, X_k) = E\{(X_j - \mu_j)(X_k - \mu_k)\} = E(X_j X_k) - \mu_j \mu_k$
Correlations: $\rho_{jk} = \text{Corr}(X_j, X_k) = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$

3.1.2 Mean and covariance matrix

The **mean vector** of $\underline{X}$ is $\underline{\mu} = E(\underline{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_p \end{pmatrix}$

The **covariance matrix** (**variance-covariance matrix, dispersion matrix**) of $\underline{X}$ is

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

Since the (i,j)th element of $(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T$ is $(X_i - \mu_i)(X_j - \mu_j)$, we see that $\Sigma = E\{(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T\} = E(\underline{X}\underline{X}^T) - \underline{\mu}\underline{\mu}^T$.

3.1.3 Properties

Let $\underline{X}$ have mean $\underline{\mu}$ and covariance matrix Σ. Let $\underline{a}$, $\underline{b}$ be p-vectors and A be a $q \times p$ matrix. Then

(i) $E(\underline{a}^T \underline{X}) = \underline{a}^T \underline{\mu}$

(ii) $\text{Var}(\underline{a}^T \underline{X}) = \underline{a}^T \Sigma \underline{a}$. It follows that Σ is **positive semi-definite**.

(iii) $\text{Cov}(\underline{a}^T \underline{X}, \underline{b}^T \underline{X}) = \underline{a}^T \Sigma \underline{b}$

(iv) $\text{Cov}(A\underline{X}) = A\Sigma A^T$

(v) $E(\underline{X}^T A\underline{X}) = \text{trace}(A\Sigma) + \underline{\mu}^T A\underline{\mu}$

Proof:

3.1.4 Self-study exercises

1. Let $X_1 = I_1 Y$, $X_2 = I_2 Y$, where I_1, I_2 and Y are independent and I_1 and I_2 take values ± 1 each with probability $\frac{1}{2}$.

   Show that $E(X_j) = 0$, Var$(X_j) = E(Y^2)$, Cov$(X_1, X_2) = 0$.

2. Verify that $E(X_1 + \cdots + X_p) = \mu_1 + \cdots + \mu_p$ and Var$(X_1 + \cdots + X_p) = \sum_{ij} \sigma_{ij}$, where $\mu_i = E(X_i)$ and $\sigma_{ij} =$ Cov(X_i, X_j).

   Suppose now that the X_i's are iid. Verify that $\bar{X}$ has mean μ and variance σ^2/p, where $\mu = E(X_i)$ and $\sigma^2 =$ Var(X_i).

3.2 Transformations

3.2.1 The univariate case

Problem: to find the distribution of $Y = h(X)$ from the known distribution of X. The case where h is a one-to-one function was treated in Section 1.2.3. When h is many-to-one we use the following generalised formulae:
Discrete case: $p_Y(y) = \sum p_X(x)$
Continuous case: $f_Y(y) = \sum f_X(x) \left| \frac{dx}{dy} \right|$
where in both cases the summations are over the set $\{x : h(x) = y\}$. That is, we add up the contributions to the mass or density at y from all x values which map to y.

Example 17: (discrete) Suppose $p_X(x) = p_x$ for $x = 0, 1, 2, 3, 4, 5$ and let $Y = (X-2)^2$. Obtain the pmf of Y.

Example 18: (continuous) Suppose $f_X(x) = 2x$ on $(0, 1)$ and let $Y = (X - \frac{1}{2})^2$. Obtain the pdf of Y.

3.2.2 The multivariate case

Problem: to find the distribution of $\underline{Y} = h(\underline{X})$, where $\underline{Y}$ is $s \times 1$ and $\underline{X}$ is $r \times 1$, from the known distribution of $\underline{X}$.

Discrete case: $p_{\underline{Y}}(\underline{y}) = \sum p_{\underline{X}}(\underline{x})$ with the summation over the set $\{\underline{x} : h(\underline{x}) = \underline{y}\}$.

Continuous case:

Case (i): h is a one-to-one transformation (so that $s = r$). Then the rule is

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{x}(\underline{y})) \left| \frac{d\underline{x}}{d\underline{y}} \right|_+$$

where $\left| \frac{d\underline{x}}{d\underline{y}} \right|$ is the Jacobian of transformation, with $\left(\frac{d\underline{x}}{d\underline{y}} \right)_{ij} = \frac{\partial x_i}{\partial y_j}$.

Case (ii): $s < r$. First transform the s-vector $\underline{Y}$ to the r-vector $\underline{Y}'$, where $Y_i' = Y_i$, $i = 1, \ldots, s$, and Y_i', $i = s+1, \ldots, r$, are chosen for convenience. Now find the density of $\underline{Y}'$ as above and then integrate out $Y_{s+1}', \ldots, Y_r'$ to obtain the marginal density of $\underline{Y}$, as required.

Case (iii): $s = r$ but $h(\cdot)$ is not monotonic. Then there will generally be more than one value of x corresponding to a given y and we need to add the probability contributions from all relevant $\underline{x}$s.

Example 19: (linear transformation) Suppose that $\underline{Y} = A\underline{X}$, where A is an $r \times r$ nonsingular matrix. Then $f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(A^{-1}\underline{y})|A|_{+}^{-1}$.

Example 20: Suppose $f_{\underline{X}}(\underline{x}) = e^{-x_1-x_2}$ on $(0,\infty)^2$. Obtain the density of $Y_1 = \frac{1}{2}(X_1 + X_2)$.

Sums and products If X_1 and X_2 are independent random variables with densities f_1 and f_2, then
(i) $X_1 + X_2$ has density $g(u) = \int f_1(u-v)f_2(v)dv$ (convolution integral)
(ii) X_1X_2 has density $g(u) = \int f_1(u/v)f_2(v)|v|^{-1}dv$.
Proof:

3.2.3 Self-study exercises

1. If $f_X(x) = \frac{2}{9}(x+1)$ on $(-1, 2)$ and $Y = X^2$, find $f_Y(y)$.

2. If X has density $f(x)$ calculate the density $g(y)$ of $Y = X^2$ when

   (i) $f(x) = 2xe^{-x^2}$ on $(0, \infty)$;

   (ii) $f(x) = \frac{1}{2}(1+x)$ on $|x| \le 1$;

   (iii) $f(x) = \frac{1}{2}$ on $-\frac{1}{2} \le x \le \frac{3}{2}$.

3. Let X_1 and X_2 be independent exponential (λ), and let $Y_1 = X_1 + X_2$ and $Y_2 = X_1/X_2$. Show that Y_1 and Y_2 are independent and find their densities.

3.3 Moments, generating functions and inequalities

3.3.1 Moment generating function

The moment generating function of the **random vector** $\underline{X}$ is defined as

$$M(\underline{z}) = E(e^{\underline{z}^T \underline{X}}) .$$

Here $\underline{z}^T \underline{X} = \sum_j z_j X_j$.

Properties

Suppose $\underline{X}$ has mgf $M(\underline{z})$. Then

(i) $\underline{X} + \underline{a}$ has mgf $e^{\underline{a}^T \underline{z}} M(\underline{z})$ and $a\underline{X}$ has mgf $M(a\underline{z})$.

(ii) The mgf of $\sum_{j=1}^k X_j$ is $M(z, \ldots, z)$.

(iii) If $X_1, \ldots, X_k$ are independent random variables with mgfs $M_j(z_j)$, j=1,…,k, then the mgf of $\underline{X} = (X_1, \ldots, X_k)^T$ is $M(\underline{z}) = \prod_{j=1}^k M_j(z_j)$, the product of the individual mgfs.

Proof:

3.3.2 Cumulant generating function

The **cumulant generating function** (cgf) of X is defined as $K(z) = \log M(z)$. The **cumulants** of X are defined as the coefficients κ_j in the power series expansion $K(z) = \sum_{j=1}^{\infty} \kappa_j z^j / j!$.
The first two cumulants are

$$\kappa_1 = \mu = E(X), \quad \kappa_2 = \sigma^2 = \text{Var}(X)$$

Similarly, the third and fourth cumulants are found to be $\kappa_3 = E(X-\mu)^3$, $\kappa_4 = E(X-\mu)^4 - 3\sigma^4$. These are used to define the **skewness**, $\gamma_1 = \kappa_3/\kappa_2^{3/2}$, and the **kurtosis**, $\gamma_2 = \kappa_4/\kappa_2^2$.

Cumulants of the sample mean. Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution with cgf $K(z)$ and cumulants κ_j. Then the mgf of $\bar{X} = n^{-1}\sum_{j=1}^{n} X_j$ is $\{M(n^{-1}z)\}^n$, so the cgf is

$$\log\{M(n^{-1}z)\}^n = nK(n^{-1}z) = n\sum_{j=1}^{\infty} \kappa_j (n^{-1}z)^j / j! \ .$$

Hence the jth cumulant of $\bar{X}$ is κ_j / n^{j-1} and it follows that $\bar{X}$ has mean $\kappa_1 = \mu$, variance $\kappa_2/n = \sigma^2/n$, skewness $(\kappa_3/n^2)/(\kappa_2/n)^{3/2} = \gamma_1/n^{1/2}$ and kurtosis $(\kappa_4/n^3)/(\kappa_2/n)^2 = \gamma_2/n$.

3.3.3 Some useful inequalities

Markov's inequality
Let X be any random variable with finite mean. Then for all $a > 0$

$$\text{pr}(|X| \geq a) \leq \frac{E|X|}{a} \ .$$

Proof:

Cauchy-Schwartz inequality

Let X, Y be any two random variables with finite variances. Then

$$\{E(XY)\}^2 \leq E(X^2)E(Y^2)\,.$$

Proof:

Jensen's inequality

If $u(x)$ is a **convex** function then

$$E\{u(X)\} \geq u(E(X))\,.$$

Note that $u(\cdot)$ is convex if the curve $y = u(x)$ has a supporting line underneath at each point, *e.g.* bowl-shaped.

Proof:

Examples

1. Chebyshev's inequality.
Let Y be any random variable with finite variance. Then for all $a > 0$

$$\text{pr}(|Y - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \,.$$

2. Correlation inequality.
$\{\text{Cov}(X, Y)\}^2 \leq \sigma_X^2 \sigma_Y^2$ (which implies that $|\text{Corr}(X, Y)| \leq 1$).

3. $|E(X)| \leq E(|X|)$.
[It follows that $|E\{h(Y)\}| \leq E\{|h(Y)|\}$ for any function $h(\cdot)$.]

4. $E\{(|X|^s)^{r/s}\} \geq \{E(|X|^s)\}^{r/s}$.
[Thus $\{E(|X|^r)\}^{1/r} \geq \{E(|X|^s)\}^{1/s}$ and it follows that $\{E(|X|^r)\}^{1/r}$ is an increasing function of r.]

5. A cumulant generating function is a convex function; *i.e.* $K''(z) \geq 0$.
Proof. $K(z) = \log M(z)$, so $K' = M'/M$ and $K'' = \{MM'' - (M')^2\}/M^2$.
Hence $M(z)^2 K''(z) = E(e^{zX})E(X^2 e^{zX}) - \{E(Xe^{zX})\}^2 \geq 0$, by the Cauchy-Schwartz inequality.
(on writing $Xe^{zX} = (e^{zX/2})(Xe^{zX/2})$)

3.3.4 Self-study exercises

1. Find the joint mgf $M(\underline{z})$ of (X, Y) when the pdf is $f(x, y) = \frac{1}{2}\lambda^3(x + y)e^{-\lambda(x+y)}$ on $(0, \infty)^2$. Deduce the mgf of $U = X + Y$.

2. Find all the cumulants of the $N(\mu, \sigma^2)$ distribution.

   [You may assume the mgf $e^{\mu z + \frac{1}{2}\sigma^2 z^2}$.]

3. Suppose that X is such that $E(X) = 3$ and $E(X^2) = 13$. Use Chebyshev's inequality to determine a lower bound for pr$(-2 < X < 8)$.

4. Show that $\{E(|X|)\}^{-1} \leq E(|X|^{-1})$.

3.4 Some limit theorems

3.4.1 Modes of convergence of random variables

Let $X_1, X_2, \ldots$ be a sequence of random variables. There are a number of alternative modes of convergence of (X_n) to a limit random variable X. Suppose first that $X_1, X_2, \ldots$ and X are all defined on the same sample space Ω.

Convergence in probability

$X_n \xrightarrow{p} X$ if pr$(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. Equivalently, pr$(|X_n - X| \leq \epsilon) \to 1$. Often $X = c$, a constant.

Almost sure convergence

$X_n \xrightarrow{a.s.} X$ if pr$(X_n \to X) = 1$. Again, often $X = c$. Also referred to as convergence 'with probability one'.

Almost sure convergence is a stronger property than convergence in probability. *i.e.* a.s. $\Rightarrow p$, but $p \not\Rightarrow$ a.s.

Example 21: Consider independent Bernoulli trials with constant probability of success $\frac{1}{2}$.

A typical sequence would be $01001001110101100010\ldots$.

Here the first 20 trials resulted in 9 successes, giving an observed proportion of $\bar{X}_{20} = 0.45$ successes.

Intuitively, as we increase n we would expect this proportion to get closer to 1. However, this will not be the case for all sequences: for example, the sequence 11111111111111111111 has exactly the same probability as the earlier sequence, but $\bar{X}_{20} = 1$.

It can be shown that the total probability of all infinite sequences for which the proportion of successes does not converge to $\frac{1}{2}$ is zero; *i.e.* pr$(\bar{X}_n \to \frac{1}{2}) = 1$ so $\bar{X}_n \xrightarrow{a.s.} \frac{1}{2}$ (and hence also $\bar{X}_n \xrightarrow{p} \frac{1}{2}$).

Convergence in *r*th mean

$X_n \xrightarrow{r} X$ if $E|X_n - X|^r \to 0$ as $n \to \infty$.

[rth mean $\Rightarrow p$, but rth mean $\not\Leftarrow$ a.s.]

Suppose now that the distribution functions are $F_1, F_2, \ldots$ and F. The random variables need not be defined on the same sample spaces for the following definition.

Convergence in distribution

$X_n \xrightarrow{d} X$ if $F_n(x) \to F(x)$ as $n \to \infty$ at each continuity point of F. We say that the **asymptotic distribution** of X_n is F.

[$p \Rightarrow d$, but $d \not\Rightarrow p$]

A useful result.

Let (X_n), (Y_n) be two sequences of random variables such that

$X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, a constant. Then

$$X_n + Y_n \xrightarrow{d} X + c\,,\ \ X_nY_n \xrightarrow{d} cX\,,\ \ X_n/Y_n \xrightarrow{d} X/c\ \ (c \neq 0).$$

3.4.2 Limit theorems for sums of independent random variables

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with (common) mean μ. Let $S_n = \sum_{i=1}^n X_i\,,\ \bar{X}_n = n^{-1}S_n$.

Weak Law of Large Numbers (WLLN). If $E|X_i| < \infty$ then $\bar{X}_n \xrightarrow{p} \mu$.
Proof (case $\sigma^2 = \text{Var}(X_i) < \infty$). Use Chebyshev's inequality: since $E(\bar{X}_n) = \mu$ we have, for every $\epsilon > 0$,

$$\text{pr}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$.

Example 21: (continued). Here $\sigma^2 = \text{Var}(X_i) = \frac{1}{4}$ (Bernoulli r.v.) so the WLLN applies to $\bar{X}_n$, the proportion of successes.

Strong Law of Large Numbers (SLLN). If $E|X_i| < \infty$ then $\bar{X}_n \xrightarrow{a.s.} \mu$.
[The proof is more tricky and is omitted.]

Central Limit Theorem (CLT). If $\sigma^2 = \text{Var}(X_i) < \infty$ then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)\,.$$

Equivalently,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)\,.$$

Proof. Suppose that X_i has mgf $M(z)$. Write $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. The mgf of Z_n is given by

$$M_{Z_n}(z) = E(e^{zZ_n}) = \exp\left(-\frac{z\mu\sqrt{n}}{\sigma}\right)\left\{M\left(\frac{z}{\sigma\sqrt{n}}\right)\right\}^n\,.$$

Therefore the cgf of Z_n is

$$\begin{aligned} K_{Z_n}(z) = \log M_{Z_n}(z) &= -\frac{z\mu\sqrt{n}}{\sigma} + nK\left(\frac{z}{\sigma\sqrt{n}}\right) \\ &= -\frac{z\mu\sqrt{n}}{\sigma} + n\left\{\mu\left(\frac{z}{\sigma\sqrt{n}}\right) + \frac{\sigma^2}{2}\left(\frac{z}{\sigma\sqrt{n}}\right)^2\right\} + O\left(\frac{1}{\sqrt{n}}\right) \\ &= -\frac{z\mu\sqrt{n}}{\sigma} + \frac{z\mu\sqrt{n}}{\sigma} + \frac{z^2}{2} + O\left(\frac{1}{\sqrt{n}}\right) \to \frac{z^2}{2} \end{aligned}$$

as $n \to \infty$, which is the cgf of the $N(0,1)$ distribution, as required.

[*Note on the proof of the CLT.* In cases where the mgf does not exist, a similar proof can be given in terms of the function $\phi(z) = E(e^{izX_j})$ where $i = \sqrt{-1}$. $\phi(\cdot)$ is called the **characteristic function** and always exists.]

Example 21: (continued). Normal approximation to the binomial
Suppose now that the success probability is π, so that pr$(X_i = 1) = \pi$. Then $\mu = \pi$ and $\sigma^2 = \pi(1-\pi)$, so the CLT gives $\sqrt{n}(\bar{X}_n - \pi)/\sqrt{\{\pi(1-\pi)\}}$ is approximately $N(0,1)$.
Furthermore, $\bar{X}_n \xrightarrow{p} \pi$ by the WLLN, and it follows from the 'useful result' that $\sqrt{n}(\bar{X}_n - \pi)/\sqrt{\{\bar{X}_n(1-\bar{X}_n)\}}$ is also approximately $N(0,1)$.
Poisson limit of binomial. Suppose that X_n is binomial (n, π) where π is such that $n\pi \to \lambda$ as $n \to \infty$. Then $X_n \xrightarrow{d}$ Poisson(λ).
Proof. X_n is expressible as $\sum_{i=1}^n Y_i$, where the Y_i are independent Bernoulli random variables with pr$(Y_i = 1) = \pi$. Thus X_n has pgf

$$(1 - \pi + \pi z)^n = \{1 - n^{-1}\lambda(1-z) + o(n^{-1})\}^n \to \exp\{-\lambda(1-z)\}$$

as $n \to \infty$, which is the pgf of the Poisson (λ) distribution.

3.4.3 Self-study exercises

1. In a large consignment of manufactured items 25% are defective. A random sample of 50 is drawn. Use the binomial distribution to compute the exact probability that the number of defectives in the sample is five or fewer. Use the CLT to approximate this answer.

2. The random variable Y has the Poisson (50) distribution. Use the CLT to find pr$(Y = 50)$, pr$(Y \leq 45)$ and pr$(Y > 60)$.

3. A machine in continuous use contains a certain critical component which has an exponential lifetime distribution with mean 100 hours. When a component fails it is immediately replaced by one from the stock, originally of 90 such components. Use the CLT to find the probability that the machine can be kept running for a year without the stock running out.

3.5 Further discrete distributions

3.5.1 Negative binomial distribution

Let X be the number of Bernoulli trials until the kth success. Then

$$\begin{aligned}
\mathrm{pr}(X = x) &= \mathrm{pr}(k-1 \text{ successes in first } x-1 \text{ trials, followed by success on } k\text{th trial}) \\
&= \binom{x-1}{k-1} \pi^{k-1}(1-\pi)^{x-k} \times \pi
\end{aligned}$$

(where the first factor comes from the binomial distribution). Hence define the pmf of the **negative binomial** (k, π) **distribution** as

$$p(x) = \binom{x-1}{k-1} \pi^k (1-\pi)^{x-k}, \; x = k, k+1, \ldots$$

The mean is k/π:
The variance is $k(1-\pi)/\pi^2$ (see exercise 1).
The pgf is $\{\pi/(z^{-1} - 1 + \pi)\}^k$:

The name "negative binomial" comes from the binomial expansion

$$1 = \pi^k \{1 - (1-\pi)\}^{-k} = \sum_{x=k}^{\infty} p(x)$$

where $p(x)$ are the negative binomial probabilities. (Exercise: verify)

3.5.2 Hypergeometric distribution

An urn contains n_1 red beads and n_2 black beads. Suppose that m beads are drawn *without replacement* and let X be the number of red beads in the sample. Note that, since $X \leq n_1$ and $X \leq m$, the possible values of X are $0, 1, ..., \min(n_1, m)$. Then

$$
\begin{aligned}
p(x) = \text{pr}(X = x) &= \frac{\text{no. of selections of } x \text{ reds and } m - x \text{ blacks}}{\text{total no. of selections of } m \text{ beads}} \\
&= \frac{\begin{pmatrix} n_1 \\ x \end{pmatrix}\begin{pmatrix} n_2 \\ m - x \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ m \end{pmatrix}}, \quad x = 0, 1, ..., \min(n_1, m).
\end{aligned}
$$

This is the pmf of the **hypergeometric** (n_1, n_2, m) **distribution**.
The mean is $n_1 m/(n_1 + n_2)$ and the variance is $n_1 n_2 m(n_1 + n_2 - m)/\{(n_1 + n_2)^2(n_1 + n_2 - 1)\}$.

3.5.3 Multinomial distribution

An urn contains n_j beads of colour j $(j = 1, \ldots k)$. Suppose that m beads are drawn *with replacement* and let X_j be the number of beads of colour j in the sample. Then, for $x_j = 0, 1, \ldots, m$ and $\sum_{j=1}^{k} x_j = m$,

$$
p(\underline{x}) = \text{pr}(\underline{X} = \underline{x}) = \begin{pmatrix} m \\ \underline{x} \end{pmatrix} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k},
$$

where $\pi_j = n_j / \sum_{i=1}^{k} n_i$. This is the pmf of the **multinomial** $(k, m, \underline{\pi})$ **distribution**. Here

$$
\begin{aligned}
\begin{pmatrix} m \\ \underline{x} \end{pmatrix} &= \text{no. of different orderings of } x_1 + \cdots + x_k \text{ beads} \\
&= \begin{pmatrix} \dfrac{m!}{x_1! \cdots x_k!} \end{pmatrix}
\end{aligned}
$$

and the probability of any given order is $\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$. The name "multinomial" comes from the multinomial expansion of $(\pi_1+\cdots+\pi_k)^m$ in which the coefficient of $\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$ is $\begin{pmatrix} m \\ \underline{x} \end{pmatrix}$.

The means are $m\pi_j$:

The covariances are $\sigma_{jk} = m(\delta_{jk}\pi_j - \pi_j\pi_k)$.

The joint pgf is $E(\prod_j z_j^{X_j}) = (\sum_{j=1}^k \pi_j z_j)^m$:

3.5.4 Self-study exercises

1. Derive the variance of the negative binomial (k, π) distribution.

   [You may assume the formula for the pgf.]

2. Suppose that $X_1, \ldots, X_k$ are independent geometric (π) random variables. Using pgfs, show that $\sum_{j=1}^k X_j$ is negative binomial (k, π).

   [Hence, the waiting times X_j between successes in Bernoulli trials are independent geometric, and the overall waiting time to the kth success is negative binomial.]

3. If $\underline{X}$ is multinomial $(k, m, \underline{\pi})$ show that X_j is binomial (m, π_j), $X_j + X_k$ is binomial $(m, \pi_j + \pi_k)$, *etc.*

   [Either by direct calculation or using the pgf.]

3.6 Further continuous distributions

3.6.1 Gamma and beta functions

Gamma function: $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$ for $a > 0$
Integration by parts gives $\Gamma(a) = (a-1)\Gamma(a-1)$.
In particular, for integer a, $\Gamma(a) = (a-1)!$ (since $\Gamma(1) = 1$). Also, $\Gamma(1/2) = \sqrt{\pi}$.

Beta function: $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$ for $a > 0, b > 0$
Relationship with Gamma function: $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

3.6.2 Gamma distribution

The pdf of the **gamma** (α, β) **distribution** is defined as

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)},\ x > 0$$

where $\alpha > 0$ and $\beta > 0$. When $\alpha = 1$, this is the exponential (β) distribution.
The mean is α/β:

The variance is α/β^2 (see exercise 2).
The mgf is $(1 - z/\beta)^{-\alpha}$:

Note that the mode is $(\alpha - 1)/\beta$ if $\alpha \geq 1$, but $f(0) = \infty$ if $\alpha < 1$.

Example 22: The journey time of a bus on a nominal $\frac{1}{2}$-hour route has the gamma $(3, 6)$ distribution. What is the probability that the bus is over half an hour late?

Sums of exponential random variables Suppose that $X_1, \ldots, X_n$ are iid exponential (λ) random variables. Then $\sum_{i=1}^n X_i$ is gamma (n, λ).
Proof:

3.6.3 Beta distribution

The pdf of the **beta** (α, β) **distribution** is

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}\,,\; 0 < x < 1\,,$$

where $\alpha > 0$ and $\beta > 0$.

The mean is $\alpha/(\alpha+\beta)$:

The variance is $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$.

The mode is $(\alpha-1)/(\alpha+\beta-2)$ if $\alpha \geq 1$ and $\alpha+\beta > 2$.

Property If X_1 and X_2 are independent, respectively gamma (ν_1, λ) and gamma (ν_2, λ), then $U_1 = X_1 + X_2$ and $U_2 = X_1/(X_1+X_2)$ are independent, respectively gamma $(\nu_1+\nu_2, \lambda)$ and beta (ν_1, ν_2).

Proof The inverse transformation is

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U_1 U_2 \\ U_1(1-U_2) \end{pmatrix}$$

with Jacobian

$$\left| \frac{d\underline{x}}{d\underline{u}} \right| = \begin{vmatrix} u_2 & u_1 \\ 1-u_2 & -u_1 \end{vmatrix} = -u_1 \, .$$

Therefore

$$
\begin{aligned}
f_U(u) &= \left[\frac{\lambda^{\nu_1}(u_1u_2)^{\nu_1-1}e^{-\lambda u_1u_2}}{\Gamma(\nu_1)}\right]\left[\frac{\lambda^{\nu_2}\{u_1(1-u_2)\}^{\nu_2-1}e^{-\lambda u_1(1-u_2)}}{\Gamma(\nu_2)}\right]|-u_1| \\
&= \left\{\frac{\lambda^{\nu_1+\nu_2}u_1^{\nu_1+\nu_2-1}e^{-\lambda u_1}}{\Gamma(\nu_1+\nu_2)}\right\}\left\{\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)}u_2^{\nu_1-1}(1-u_2)^{\nu_2-1}\right\}
\end{aligned}
$$

on $(0,\infty)\times(0,1)$ and the result follows.

3.6.4 Self-study exercises

1. Suppose X has the gamma $(2,4)$ distribution. Find the probability that X exceeds $\mu+2\sigma$, where μ,σ are respectively the mean and standard deviation of X.

2. Derive the variance of the gamma (α,β) distribution. [Either by direct calculation or using the mgf.]

3. Find the distribution of $-\log X$ when X is uniform (0,1). Hence show that if $X_1,\ldots,X_k$ are iid uniform (0,1) then $-\log(X_1X_2\cdots X_k)$ is gamma $(k,1)$.

4. If X is gamma (ν,λ) show that $\log X$ has mgf $\lambda^{-z}\Gamma(z+\nu)/\Gamma(\nu)$.

5. Suppose X is uniform $(0,1)$ and $\gamma>0$ Show that $Y=X^{1/\gamma}$ is beta $(\gamma,1)$.

Chapter 4

Normal and associated distributions

4.1 The multivariate normal distribution

4.1.1 Multivariate normal

The **multivariate normal** distribution, denoted $N_p(\underline{\mu}, \Sigma)$, has pdf

$$f(\underline{x}) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})\}$$

on $(-\infty, \infty)^p$.
The mean is $\underline{\mu}$ $(p \times 1)$ and the covariance matrix is Σ $(p \times p)$ (see property (v)).

Bivariate case, $p = 2$. Here

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$\begin{aligned} |2\pi\Sigma| &= (2\pi)^2 \sigma_1^2 \sigma_2^2 (1 - \rho^2) \\ \Sigma^{-1} &= (1 - \rho^2)^{-1} \begin{pmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{pmatrix}, \quad \text{giving} \end{aligned}$$

$$f(x_1, x_2) = \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right]}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

4.1.2 Properties

i) Suppose $\underline{X}$ is $N_p(\underline{\mu}, \Sigma)$ and let $\underline{Y} = T^{-1}(\underline{X} - \underline{\mu})$, where $\Sigma = TT^T$. Then $Y_i, i = 1, \ldots, p$, are independent $N(0, 1)$.

(ii) The joint mgf of $N_p(\underline{\mu}, \Sigma)$ is $e^{\underline{\mu}^T \underline{z} + \frac{1}{2}\underline{z}^T \Sigma \underline{z}}$. (*C.f.* property (iv), Section 2.4.2.)

(iii) If $\underline{X}$ is $N_p(\underline{\mu}, \Sigma)$ then $A\underline{X} + \underline{b}$ (where A is $q \times p$ and $\underline{b}$ is $q \times 1$) is $N_q(A\underline{\mu} + \underline{b}, A\Sigma A^T)$.
(*C.f.* property (i), Section 2.4.2.)

(iv) If $\underline{X}_i,\ i = 1, \ldots, n$, are independent $N_p(\underline{\mu}_i, \Sigma_i)$, then $\sum_i \underline{X}_i$ is $N_p(\sum_i \underline{\mu}_i, \sum_i \Sigma_i)$.
(*C.f.* property (iii), Section 2.4.2.)

(v) Moments of $N_p(\underline{\mu}, \Sigma)$. Obtain by differentiation of the mgf. In particular, differentiating w.r.t. z_j and z_k gives $E(X_j) = \mu_j$, Var$(X_j) = \Sigma_{jj}$ and Cov$(X_j, X_k) = \Sigma_{jk}$.
Note that if $X_1, \ldots, X_p$ are all uncorrelated (*i.e.* $\Sigma_{jk} = 0$ for $j \neq k$) then $X_1, \ldots, X_p$ are independent $N(\mu_j, \sigma_j^2)$.

(vi) If $\underline{X}$ is $N_p(\underline{\mu}, \Sigma)$ then $\underline{a}^T\underline{X}$ and $\underline{b}^T\underline{X}$ are independent if and only if $\underline{a}^T\Sigma\underline{b} = 0$. Similarly for $A^T\underline{X}$ and $B^T\underline{X}$.

4.1.3 Marginal and conditional distributions

Suppose that $\underline{X}$ is $N_p(\underline{\mu}, \Sigma)$. Partition $\underline{X}^T$ as $(\underline{X}_1^T, \underline{X}_2^T)$ where $\underline{X}_1$ is $p_1 \times 1$, $\underline{X}_2$ is $p_2 \times 1$ and $p_1 + p_2 = p$. Correspondingly $\underline{\mu}^T = (\underline{\mu}_1^T, \underline{\mu}_2^T)$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.
Note that $\Sigma_{21}^T = \Sigma_{12}$ and $\underline{X}_1$ and $\underline{X}_2$ are independent if and only if $\Sigma_{12} = 0$ (since

the joint density factorises if and only if $\Sigma_{12} = 0$).

The **marginal** distribution of $\underline{X}_1$ is $N_{p_1}(\underline{\mu}_1, \Sigma_{11})$.

Proof:

The **conditional** distribution of $\underline{X}_2|\underline{X}_1$ is $N_{p_2}(\underline{\mu}_{2.1}, \Sigma_{22.1})$, where

$$\begin{aligned} \underline{\mu}_{2.1} &= \underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{X}_1 - \underline{\mu}_1) \\ \Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

(proof omitted). Note that $\underline{\mu}_{2.1}$ is linear in $\underline{X}_1$.

4.1.4 Self-study exercises

1. Write down the joint density of the $N_2\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}\right)$ distribution in component form.

2. Suppose that $\underline{X}_i,\ i = 1, \ldots, n$, are independent $N_p(\underline{\mu}, \Sigma)$. Show that the sample mean vector, $\bar{\underline{X}} = n^{-1}\sum_i \underline{X}_i$ is $N_p(\underline{\mu},\ n^{-1}\Sigma)$.

3. For the distribution in exercise 1, obtain the marginal distributions of X_1 and X_2 and the conditional distributions of X_2 given $X_1 = x_1$ and X_1 given $X_2 = x_2$.

4.2 The chi-square, t and F distributions

4.2.1 Chi-square distribution

The pdf of the **chi-square distribution** with ν **degrees of freedom** ($\nu > 0$) is

$$f(u) = \frac{u^{\frac{1}{2}\nu - 1} e^{-\frac{1}{2}u}}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}\,,\ u > 0\,.$$

Denoted by χ^2_ν. Note that the χ^2_ν distribution is identical to the gamma $(\frac{\nu}{2}, \frac{1}{2})$ distribution (*c.f.* Section 3.6). It follows that the mean is ν, the variance is 2ν and the mgf is $(1 - 2z)^{-\nu/2}$.

Properties

(i) Let ν be a positive integer and suppose that $X_1, \ldots, X_\nu$ are iid $N(0, 1)$. Then $\sum_{i=1}^{\nu} X_i^2$ is χ^2_ν. In particular, if X is $N(0, 1)$ then X^2 is χ^2_1.

(ii) If $U_i,\ i = 1, \ldots, n$, are independent $\chi^2_{\nu_i}$ then $\sum_{i=1}^{n} U_i$ is χ^2_ν with $\nu = \sum_{i=1}^{n} \nu_i$.

(iii) If $\underline{X}$ is $N_p(\underline{\mu}, \Sigma)$ then $(\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu})$ is χ^2_p.

Theorem (Joint distribution of the sample mean and variance)

Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Let $\bar{X} = n^{-1} \sum_i X_i$ be the sample mean and $S^2 = (n-1)^{-1} \sum_i (X_i - \bar{X})^2$ the sample variance.

Then $\bar{X}$ is $N(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2$ is χ^2_{n-1} and $\bar{X}$ and S^2 are independent.

Proof:

4.2.2 Student's t distribution

The pdf of the **Student's** t **distribution** with ν **degrees of freedom** ($\nu > 0$) is

$$f(t) = \frac{1}{B(\frac{1}{2},\frac{\nu}{2})\nu^{\frac{1}{2}}(1+\frac{t^2}{\nu})^{\frac{1}{2}(\nu+1)}}\,,\ -\infty < t < \infty\,.$$

Denoted by t_ν. The mean is 0 (provided $\nu > 1$):

The variance is $\nu/(\nu-2)$ (provided $\nu > 2$).

Theorem If X is $N(0,1)$, U is χ^2_ν and X and U are independent, then

$$T \equiv \frac{X}{\sqrt{U/\nu}} \sim t_\nu.$$

Proof:

4.2.3 Variance ratio (F) distribution

The pdf of the **variance ratio**, or F **distribution** with ν_1, ν_2 **degrees of freedom** ($\nu_1, \nu_2 > 0$) is

$$f(x) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} x^{\frac{1}{2}\nu_1 - 1}}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})(1 + \frac{\nu_1 x}{\nu_2})^{\frac{1}{2}(\nu_1+\nu_2)}}\,, \; x > 0\,.$$

Denoted by F_{ν_1,ν_2}. The mean is $\nu_2/(\nu_2 - 2)$ (provided $\nu_2 > 2$) and the variance is $2\nu_2^2(\nu_1 + \nu_2 - 2)/\{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)\}$ (provided $\nu_2 > 4$).

Theorem. If U_1 and U_2 are independent, respectively $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$, then

$$F \equiv \frac{U_1/\nu_1}{U_2/\nu_2} \sim F_{\nu_1,\nu_2}.$$

Proof:

It follows from the above result that (i) $F_{\nu_1,\nu_2} \equiv 1/F_{\nu_2,\nu_1}$ and (ii) $F_{1,\nu} \equiv t^2_\nu$. (Exercise: check)

4.3 Normal theory tests and confidence intervals

4.3.1 One-sample t-test

Suppose that $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$. Then, from Section 3.2, $\bar{Y} = n^{-1}\sum_i Y_i$ (the sample mean) and $S^2 = (n-1)^{-1}\sum_i(Y_i - \bar{Y})^2$ (the sample variance) are independent, respectively $N(\mu, \sigma^2/n)$ and $\sigma^2\chi^2_{n-1}/(n-1)$. Hence

$$Z = \frac{(\bar{Y} - \mu)}{\sigma/\sqrt{n}}$$

is $N(0,1)$,

$$U = \frac{(n-1)S^2}{\sigma^2}$$

is χ^2_{n-1} and Z, U are independent.

It follows that

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}}$$

is t_{n-1}.

Applications:

Inference about μ: one-sample z-test (σ known) and t-test (σ unknown).

Inference about σ^2: χ^2 test.

4.3.2 Two-samples

Two independent samples. Suppose that $Y_{11}, \ldots, Y_{1n_1}$ are iid $N(\mu_1, \sigma_1^2)$ and $Y_{21}, \ldots, Y_{2n_2}$ are iid $N(\mu_2, \sigma_2^2)$.

Summary statistics: $(n_1, \bar{Y}_1, S_1^2)$ and $(n_2, \bar{Y}_2, S_2^2)$

Pooled sample variance: $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

From Section 4.2, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, then $\bar{Y}_1$ and $(n_1-1)S_1^2$ are independent $N(\mu_1, n_1^{-1}\sigma^2)$, $\sigma^2\chi^2_{n_1-1}$ respectively, and $\bar{Y}_2$ and $(n_1-1)S_2^2$ are independent $N(\mu_2, n_2^{-1}\sigma^2)$, $\sigma^2\chi^2_{n_2-1}$ respectively.

Furthermore, $(\bar{Y}_1, (n_1-1)S_1^2)$ and $(\bar{Y}_2, (n_2-1)S_2^2)$ are independent.

Therefore $(\bar{Y}_1 - \bar{Y}_2)$ is $N(\mu_1 - \mu_2), (n_1^{-1} + n_2^{-1})\sigma^2)$, $(n_1+n_2-2)S^2$ is $\sigma^2\chi^2_{n_1+n_2-2}$ and $(\bar{Y}_1 - \bar{Y}_2)$ and $(n_1+n_2-2)S^2$ are independent.

Therefore

$$T \equiv \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

is $t_{n_1+n_2-2}$.

Also, since S_1^2 , S_2^2 are independent,

$$F \equiv \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}.$$

Applications:

Inference about $\mu_1 - \mu_2$: two-sample z-test (σ known) and t-test (σ unknown).

Inference about σ_1^2/σ_2^2: F (variance ratio) test.

Matched pairs Observations $(Y_{i1}, Y_{i2} : i = 1, \ldots, n)$ where the differences $D_i = Y_{i1} - Y_{i2}$ are independent $N(\mu, \sigma^2)$. Then

$$T = \frac{\bar{D} - \mu}{S/\sqrt{n}}$$

is t_{n-1}, where S^2 is the sample variance of the D_i's.

Application:
Inference about μ from paired observations: paired-sample t-test.

4.3.3 k samples (One-way Anova)

Suppose we have k groups, with group means $\mu_1, \ldots, \mu_k$.
Denote the independent observations by $(Y_{i1}, \ldots, Y_{in_i} : i = 1, \ldots, k)$ with $Y_{ij} \sim N(\mu_i, \sigma^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots k$.
Summary statistics: $((n_i, S_i^2) : i = 1, \ldots, k)$.
Total sum of squares: $ssT = \sum_{ij}(Y_{ij} - \bar{Y})^2$, where $\bar{Y} = n^{-1} \sum_{ij} Y_{ij}$ (the overall mean) and $n = \sum_i n_i$.
Then $ssT = ssW + ssB$ where
$ssW = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2 = \sum_i (n_i - 1)S_i^2$ (the **within-samples ss**)
$ssB = \sum_i n_i(\bar{Y}_i - \bar{Y})^2$ (the **between-samples ss**)

From Sections 4.1 and 4.2, $(n_i - 1)S_i^2/\sigma^2$ is $\chi^2_{n_i-1}$ independent of $\bar{Y}_i$.
Hence ssW/σ^2 is χ^2_{n-k} independent of ssB.
Also, by a similar argument to that of the Theorem in Section 3.2 (proof omitted), ssB is $\sigma^2\chi^2_{k-1}$ when $\mu_i = \mu$, say, for all i.
Hence we obtain the F-test for equality of the group means μ_i:

$$F = \frac{ssB/(k-1)}{ssW/(n-k)}$$

is $F_{k-1,n-k}$ under the null hypothesis $\mu_1 = \cdots = \mu_k$.

4.3.4 Normal linear regression

Observations $Y_1, \dots, Y_n$ are independently $N(\alpha + \beta x_i, \sigma^2)$, where $x_1, \dots, x_n$ are given constants.
The **least-squares estimator** $(\hat{\alpha}, \hat{\beta})$ is found by minimizing the sum of squares $Q(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2$.
By partial differentiation with respect to α and β, we obtain

$$\hat{\beta} = T_{xy}/T_{xx}\,, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

where $T_{xx} = \sum_i (x_i - \bar{x})^2$ and $T_{xy} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$
Note that, since both $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of $\underline{Y} = (Y_1, \dots, Y_n)^T$, they are jointly normally distributed.
Using properties of expectation and covariance matrices, we find that $(\hat{\alpha}, \hat{\beta})^T$ is bivariate normal with mean (α, β) and covariance matrix

$$V = \frac{\sigma^2}{T_{xx}} \begin{pmatrix} n^{-1}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Sums of squares

Total ss: $T_{yy} = \sum_i (Y_i - \bar{Y})^2$;
Residual ss: $Q(\hat{\alpha}, \hat{\beta})$;
Regression ss: $T_{yy} - Q(\hat{\alpha}, \hat{\beta})$

Results:

(a) Residual ss = $T_{yy} - T_{xx}\hat{\beta}^2$, Regression ss = $T_{xx}\hat{\beta}^2 = T_{xy}^2/T_{xx}$
(b) $E(\text{Total ss}) = T_{xx}\beta^2 + (n-1)\sigma^2$, $E(\text{Regression ss}) = T_{xx}\beta^2 + \sigma^2$, $E(\text{Residual ss}) = (n-2)\sigma^2$
(c) By a similar argument to that of the Theorem in Section 3.2 (proof omitted), Residual ss is $\sigma^2\chi^2_{n-2}$ and, if $\beta = 0$, Regression ss is $\sigma^2\chi^2_1$, independently of Residual ss.

Application:

The residual mean square, $S^2 =$ Residual ss$/(n-2)$, is an unbiased estimator of σ^2, $\hat{\beta}$ is an unbiased estimator of β with estimated standard error $S/\sqrt{T}_{xx}$, and $\hat{\alpha}$ is an unbiased estimator of α with estimated standard error $(S/\sqrt{T}_{xx})(\sum_i x_i^2/n)^{1/2}$.
If $\beta = \beta_0$ then

$$T = \frac{\hat{\beta} - \beta_0}{S/\sqrt{T}_{xx}}$$

is t_{n-2}, giving rise to tests and confidence intervals about β.
If $\beta = 0$ then

$$F = \frac{\text{Regression ss}}{S^2}$$

is $F_{1,n-2}$, hence a test for $\beta = 0$.
(Alternatively, and equivalently, use $T = \frac{\hat{\beta}}{S/\sqrt{T_{xx}}}$ as t_{n-2}.)

The **coefficient of determination** is

$$r^2 = \frac{\text{Regression ss}}{\text{Total ss}} = \frac{T_{xy}^2}{T_{xx}T_{yy}}$$

(square of the sample correlation coefficient). The coefficient of determination gives the proportion of Y-variation attributable to regression on x.