STAT2001: Probability and Inference (0.5 unit)

STAT3101: Probability and Statistics II (0.5 unit)

Lecturer: Yvo Pokern

Aims of course

To continue the study of probability and statistics beyond the basic concepts introduced in previous courses (see prerequisites below). To provide further study of probability theory, in particular as it relates to multivariate distributions, and to introduce some formal concepts and methods in statistical estimation.

Objectives of course

On successful completion of the course, a student should have an understanding of the properties of joint distributions of random variables and be able to derive these properties and manipulate them in straightforward situations; recognise the χ^2, t and F distributions of statistics defined in terms of normal variables; be able to apply the ideas of statistical theory to determine estimators and their properties in relation to a range of estimation criteria.

Application areas

As with other core modules in probability and statistics, the material in this course has applications in almost every field of quantitative investigation; the course expands on earlier modules by introducing general-purpose techniques that are applicable in principle to a wide range of real-life situations.

Prerequisites and Workload

STAT1004 and STAT1005 or MATH7501 or their equivalent.
Lectures: 3 hours per week during term 1.

Tutorials: 1 hour per week during term 1.

Office hour

I will be available for consultation in Room 144 (Department of Statistical Science), at a time to be announced at the start of the course. If you cannot make it to the office hour, you may also contact me by email at `y.pokern@ucl.ac.uk` to arrange an appointment but please note that I do not generally discuss mathematics by email as this is often inefficient. Please ask questions in the Moodle online discussion forum instead.

Exercise Sheets

There will be ten weekly exercise sheets in total. Section A on these exercise sheets is for you to do at home; it serves as a warm-up for section B. Your answers to section B should be handed in using the lockers in the undergraduate common room in the Department of Statistical Science (Room number 117) by the deadline stated on the exercise sheet. One randomly selected section B question will be marked and the best seven out of eight marks make up the 10% in-course assessed component. Your exercises will be handed back and solutions as well as common mistakes will be discussed at the weekly 1h tutorial. Finally, exercise sheets contain a section C with questions which in part are from past exams or in-course assessments or which are of a similar style. Full solutions to section A as well as succinct solutions to section B will be available on Moodle as soon as they have been discussed in tutorials. Very succinct answers to section C will be made available on Moodle in time for exam preparation.

Summer Exam

$2\frac{1}{2}$ hour (STAT2001) or 2 hour (STAT3101) written examination paper in term 3. All questions need to be answered, past papers are available on Moodle. The final mark will be a 9 to 1 weighted average of the written examination and the in-course assessment marks. Only standard UCL calculators may be used for these assessments.

Attending Tutorials

If you do not attend tutorials (which are compulsory), then you will be asked to discuss your progress with the Departmental Tutor. In an extreme case of non-participation in tutorials, you may be banned from taking the summer exam for the course, which means that you will be classified as ‘not complete’ for the course (in practice this means that you will fail the course). Finally, exercise sheets contain a section C with questions which in part are from past exams or in-course assessments or which are of a similar style.

Full solutions to section A as well as succinct solutions to section B will be available on Moodle as soon as they have been discussed in tutorials. Very succinct answers to section C will be made available on Moodle in time for exam preparation.

Feedback

Feedback in this course will be given mainly through two channels: written feedback on your weekly exercise sheet and discussion of the exercise sheet, in particular of common mistakes, in your tutorial. Additionally, there will be regular questions during the lecture where you can contribute your answers and you can also come to the office hours to discuss any questions you may have on the course material in greater detail. There will also be occasional polls and quizzes on Moodle which provide instant feedback in addition to the online forum.

Texts

The following texts are a small selection of the many good books available on this material. They are recommended as being especially useful and relevant to this course. The first book listed is particularly recommended. It includes large numbers of sensible worked examples and exercises (with answers to selected exercises) and also covers material on data analysis that will be useful for other statistics courses. Books marked ‘*’ are slightly more theoretical and cover more details than given in the lectures.

- J. A. Rice: *Mathematical Statistics and Data Analysis.* (Third edition; 2006) Duxbury.

- D. D. Wackerly, W. Mendenhall & R. L. Scheaffer: *Mathematical Statistics with Applications.* (Sixth edition; 2002) Duxbury.

- L. J. Bain & M. Engelhardt: *Introduction to Probability and Mathematical Statistics.* (Second edition; 1992) Duxbury.

- R. V. Hogg & E. A. Tanis: *Probability and Statistical Inference.* (Sixth edition; 2001) Prentice Hall.

- H. J. Larson: *Introduction to Probability Theory and Statistical Inference.* (Third edition; 1982) Wiley.

* G. Casella & R. L. Berger: *Statistical Inference.* (Second edition;2001) Duxbury.

* V. K. Rohatgi & E. Saleh: *An Introduction to Probability and Statistics.* (Second edition; 2001) Wiley.

Contents

1 Joint Probability Distributions 13

1.1 Revision of basic probability 13

1.2 Revision of random variables (univariate case) 16

1.2.1 What are Random Variables? 16

1.2.2 Expectation of a Random Variable 17

1.2.3 Functions of a random variable 18

1.3 Joint distributions 20

1.3.1 The joint CDF 20

1.3.2 Joint distribution: the discrete case 21

1.3.3 Joint Distribution: the continuous case 26

1.4 Further results on expectations 29

1.4.1 Expectation of a sum 29

1.4.2 Expectation of a product 30

1.4.3 Covariance 30

1.4.4 Correlation 31

1.4.5 Conditional variance 34

1.5 Standard multivariate distributions 36

1.5.1 From bivariate to multivariate 36

1.5.2 The multinomial distribution 37

1.5.3 The multivariate normal distribution 40

1.5.4 Reminder: Matrix notation 43

1.5.5 Matrix Notation for Multivariate Normal Random Variables . . . 45

2 Transformation of Variables **49**

2.1 Univariate case 49

2.1.1 Discrete case 49

2.1.2 Continuous case 50

2.2 Bivariate case 51

2.2.1 General Transformations 51

2.2.2 Sums of random variables 53

2.3 Multivariate case 53

2.4 Approximation of moments 54

2.5 Order Statistics 55

3 Generating Functions **59**

3.1 Overview 59

3.2 The probability generating function (pgf) 60

3.2.1 Definition of the pgf 60

3.2.2 Moments and the pgf 60

3.3 The moment generating function (mgf) 61

3.3.1 Definition 61

3.3.2 Moments and the mgf 62

3.3.3 Linear Transformations and the mgf 63

3.4 Joint generating functions 63

3.5 Linear combinations of random variables 66

3.5.1 The Central Limit Theorem 69

4 Distributions of Functions of Normally Distributed Variables 75

4.1 Motivation 75

4.2 Reminder: Random Sample from a Normal Population 75

4.3 The chi-squared (χ^2) distribution 76

4.3.1 The distribution of $\sum(X_i - \overline{X})^2$ 79

4.3.2 Student's t distribution 81

4.4 The F distribution 82

5 Statistical Estimation 85

5.1 Overview 85

5.2 Criteria for good estimators 85

5.2.1 Terminology: 86

5.3 Cramér-Rao lower bound: 87

5.3.1 Overview 87

5.3.2 What is the Cramér-Rao lower bound? 87

5.3.3 Proof of the Cramér-Rao bound 88

5.4 Methods for finding estimators 89

5.4.1 Overview 89

5.4.2 The method of moments 90

5.4.3 Least squares 90

5.4.4 Maximum likelihood 91

Foreword

This course continues the study of probability and statistics beyond the basic concepts introduced in previous courses, such as STAT1004 and STAT1005, MATH7501, or STATD001. In particular we will consider the following topics:

Joint (or multivariate) distributions
Models that describe the joint behaviour of more than one random variable. It is important to study the properties of and to be able to manipulate joint distributions since many applications deal with the dependence structure among *several* variables.

- Duration of unemployment is typically associated with education, age and gender of a person. Other important factors may be identified by a careful multivariate analysis.
- The joint distribution of various physiological variables in a population of patients is often of interest in medical studies.
- Yields on shares from different companies may show a complex interrelationship reflecting economic revivals and downturns in industrial sectors.

Transformation of variables
This will be considered for both the univariate and multivariate case. Transformations are useful in practice for finding simpler distributions of random variables, *e.g.* the log–transformation is often applied to skewed distributions in order to get a symmetric distribution. But transformations are also helpful for deriving the distribution of commonly used statistics. For example, the sample mean, sample median, sample variance and sample minimum are all transformations of the sample variables. The t–test statistic is a transformation of the sample variables and has the appealing feature that its distribution does not depend on the unknown parameters.

Generating functions
Like the previous topic, generating functions are mainly used as a means to the end of simplifying calculations for probability distributions. In particular, we will use the moment generating function for identifying the distribution of sums of independent random variables.

Distributions of functions of normally distributed variables
The above tools are applied to prove some crucial results on transformations of normal

variables. Most of these results are known from earlier courses, such as the relation between the normal distribution and the χ^2–distribution or t–distribution. These results are useful for deriving statistical tests and confidence intervals.

Statistical estimation

An important aspect of statistical analysis is the estimation of unknown parameters of a population from a sample and the need to *quantify the uncertainty* in this estimation. Unknown parameters could be the population mean, the strength of the association between two variables, or the intensity for the occurrence of a specific disease given a patient's history. We therefore have to address criteria for good estimators and the question of how to find good estimators.

How to use these lecture notes

The present lecture notes contain all relevant material for the course, *i.e.* definitions and results as well as the methods required to derive these results. However, we will work through the examples in detail during the lecture and additional explanations not contained in this booklet will also be given in the lectures. It is therefore essential to attend the lectures and supplement the lecture notes with your own notes. The lecture notes would be woefully incomplete without the weekly exercise sheets that will be handed out separately and discussed in tutorials.

The intranet: Moodle

All the exercise sheets handed out in lectures will be available on the course information pages accessible from the UCL Moodle login page at `https://moodle.ucl.ac.uk/`. In addition, the course information pages will include (i) answers to section A and B questions on the exercise sheets after these have been discussed in the tutorials, (ii) very succinct answers to section C questions (eventually), (iii) past in course assessments and exams as well as very succinct answers, and (iv) news and discussion fora to debate your questions online.

Learning outcomes

At the end of each chapter and of important sections you will find a list of *Learning Outcomes.* These summarize key aspects, and point out what you are expected to be able to do once you have 'learned' the material. You can use them to monitor your own progress and to check whether you are well prepared for in–course assessments or the exam. The learning outcomes will be reflected in the examples and exercises given throughout the course.

Chapter 1

Joint Probability Distributions

Joint probability distributions (or **multivariate distributions**) describe the joint behaviour of two or more random variables. Before introducing this new concept we will revise the basic notions related to the distribution of only one random variable.

1.1 Revision of basic probability

The fundamental idea of probability is that chance can be measured on a scale which runs from zero, which represents *impossibility*, to one, which represents *certainty*.

Sample space, Ω: the set of all outcomes of an experiment (real or hypothetical).

Event, A: a subset of Ω, written $A \subseteq \Omega$. The elements $\omega \in \Omega$ are called **elementary events** or **outcomes**.

Event Space, $\mathcal{A}$: The family of all events A whose probability we may be interested in. $\mathcal{A}$ is a family of sets, so e.g. the events $A_1 \subseteq \Omega$ and $A_2 \subseteq \Omega$ may be contained in it: $A_1 \in \mathcal{A}$, $A_2 \in \mathcal{A}$. The event space always contains Ω, i.e $\Omega \in \mathcal{A}$.

Probability measure, P: a *mapping* from the event space to $[0, 1]$. To qualify as a probability measure P must satisfy the following *axioms* of probability:

1. $\mathrm{P}(A) \geq 0$ for any event $A \in \mathcal{A}$;

2. $\mathrm{P}(\Omega) = 1$;

3. *Countable additivity:* If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint sets (*i.e.* $A_i \cap A_j = \emptyset$, for all $i \neq j$) then
$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathrm{P}(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

If Ω is **countable** (*i.e.* $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$) then the event space $\mathcal{A}$ can be chosen to include *all* subsets $A \subseteq \Omega$. We will always make this choice in this course.

If Ω is **uncountable**, like the real numbers, we have to define a 'suitable' family of subsets, i.e. the event space $\mathcal{A}$ does not contain *all* subsets of Ω. However, in practice the event space can always be constructed to include all events of interest.

From the axioms of probability one can mathematically prove the **addition rule**:

For *any* two events A and B we have:
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

Events A and B are said to be **independent** if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.

Events $A_1, A_2, \ldots, A_n$ are **independent** if
$$\mathrm{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathrm{P}(A_{i_1}) \ldots \mathrm{P}(A_{i_k})$$
for all possible choices of k and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. That is, the product rule must hold for every subclass of the events $A_1, \ldots, A_n$.

Note: In some contexts this would be called mutual independence. Whenever we speak of independence of more than two events or random variables, in this course, we mean mutual independence.

Example 1.1. *Consider two independent tosses of a fair coin and the events $A =$ 'first toss is head', $B =$ 'second toss is head', $C =$ 'different results on two tosses'.*
Find the sample space, the probability of an elementary event and the individual probabilities of $A, B,$ and C.
Show that $A, B,$ and C are not independent.

Suppose that $\mathrm{P}(B) > 0$. Then the **conditional probability** of A given B, $\mathrm{P}(A|B)$ is defined as
$$\boxed{\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}}$$

i.e. the relative weight attached to event A within the restricted sample space B. The conditional probability is undefined if $\mathrm{P}(B) = 0$. Note that $\mathrm{P}(\cdot|B)$ is a probability measure on B. Further note that if A and B are independent events then $\mathrm{P}(A|B) = \mathrm{P}(A)$ and $\mathrm{P}(B|A) = \mathrm{P}(B)$.

The above conditional probability formula yields the **multiplication rule**

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A|B)\mathrm{P}(B) \\ &= \mathrm{P}(B|A)\mathrm{P}(A) \end{aligned}$$

Note that if $\mathrm{P}(B|A) = \mathrm{P}(B)$ then we recover the multiplication rule for independent events.

Two events A and B are **conditionally independent** given a third event C if

$$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C).$$

Conditional independence means that once we know that C is true A carries no information on B. Note that conditional independence does not imply marginal independence, nor vice versa.

Example 1.2 (1.1 ctd.)**.** *Show that A and B are not conditionally independent given $\overline{C}$.*

The **law of total probability**, or **partition law** follows from the additivity axiom and the definition of conditional probability: 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 $\cup_i B_i = \Omega$) and let A be any event. Then

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

Bayes theorem follows from the law of total probability and the multiplication rule. Again, let $B_1, \ldots, B_k$ be mutually exclusive and exhaustive events and let A be any event with $\mathrm{P}(A) > 0$. Then Bayes theorem states that

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

1.2 Revision of random variables (univariate case)

1.2.1 What are Random Variables?

A random variable, X, assigns a real number $x \in \mathbb{R}$ to each element $\omega \in \Omega$ of the sample space Ω. The probability measure P on Ω then gives rise to a probability distribution for X.

More formally, any (measurable) function $X : \Omega \to \mathbb{R}$ is called a **random variable**. The random variable X may be **discrete** or **continuous**. The probability measure P on Ω induces a **probability distribution** for X. In particular, X has **(cumulative) distribution function (cdf)** $F_X(x) = \mathrm{P}(\{\omega : X(\omega) \leq x\})$, which is usually abbreviated to $\mathrm{P}(X \leq x)$. It follows that $F_X(-\infty) = 0, F_X(\infty) = 1$. Also, F_X is non-decreasing and right-continuous (though not necessarily continuous), and $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.

Example 1.3. *Give an example of a random variable whose cdf is right-continuous (it has to be) but not continuous.*

Discrete random variables

X takes only a finite or countably infinite set of values $\{x_1, x_2, \ldots\}$. F_X is a step-function, with steps at the x_i of sizes $p_X(x_i) = \mathrm{P}(X = x_i)$, and $p_X(\cdot)$ is the **probability mass function (pmf)** of X. (*E.g.* X = place of horse in race, grade of egg.) CDFs of discrete random variables are only right-continuous but not continuous.

Example 1.4 (1.1 ctd. II)**.** *Consider the random variable X = number of heads obtained on the two tosses. Obtain the pmf and cdf of X. Sketch the cdf – is it continuous?*

Example 1.5. *Consider the random variable $X \sim \mathrm{Geo}(p)$ with $P(X = k) = (1-p)^{k-1}p$ where $k \in \mathbb{N}$. Compute the cdf and sketch it. Is X a discrete or a continuous random variable?*

Continuous random variables

When F_X can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u.$$

for a non-negative function $f_X \geq 0$ which integrates to one, i.e. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$, then F_X is the cdf of a continuous random variable. f_X is called the **probability density function (pdf)** of X. Continuous random variables X take values in a non-countable set and

$$\mathrm{P}(x < X \leq x + \mathrm{d}x) \simeq f_X(x)\,\mathrm{d}x.$$

Thus $f_X(x)\,\mathrm{d}x$ is the probability that X lies in the infinitesimal interval $(x, x+\mathrm{d}x)$. Note that the probability that X is *exactly* equal to x is zero for all x (*i.e.* $\mathrm{P}(X = x) = 0$).

If F_X is a valid cdf which is continuous with piecewise derivative g, then F_X is the cdf of a continuous random variable and the pdf is given by g.

Example 1.6. *Suppose $f_X(x) = k(2 - x^2)$ on $(-1, 1)$. Calculate k and sketch the pdf. Calculate and sketch the cdf. Is the cdf differentiable? Calculate $P(|X| > 1/2)$.*

A **mixed** discrete/continuous random variable is such that the probability is shared between discrete and continuous components with $\sum_{x_i} p_X(x_i) + \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$, *e.g.* rainfall on given day, waiting time in queue, flow in pipe, contents of reservoir. The probability of an event A is specified through a combination of the probability mass function p_X and the probability density function f_X:

$$P(X \in A) = \int_A f_X(x)dx + \sum_{x_i \in A} p_X(x_i)$$

In a mixed random variable, we have $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x < 1$ as well as $\sum_{x_i} p_X(x_i) < 1$, although the continuous and discrete parts sum to one, of course. The cdf of a mixed random variable is non-decreasing and has jumps at those points x_i at which $p_X(x_i) > 0$.

1.2.2 Expectation of a Random Variable

A distribution has several characteristics that could be of interest, such as its shape or skewness. Another one is its **expectation**, which can be regarded as a summary of the 'average' value of a random variable.

Discrete case:

$$\mathrm{E}[X] = \sum_i x_i\, p_X(x_i) = \sum_\omega X(\omega)\mathrm{P}(\{\omega\})\,.$$

That is, the averaging can be taken over the (distinct) values of X with weights given by the probability distribution p_X, *or* over the sample space Ω with weights $\mathrm{P}(\{\omega\})$.

Continuous case:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x.$$

Mixed case:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x + \sum_{x_i} x_i p_X(x_i)$$

Note: Integration is applied for continuous random variables, summation is applied for discrete random variables. Make sure not to confuse the two.

Example 1.7. *The amount of water flowing in a pipe is zero litres per hour with probability* 0.9*. With probability* 0.1*, the flow is uniformly distributed between* $0\frac{\ell}{\mathrm{h}}$ *and* $360\frac{\ell}{\mathrm{h}}$*. Sketch the cdf for the flow of water and compute the expectation.*

1.2.3 Functions of a random variable

Let ϕ be a real-valued function on $\mathbb{R}$; that is, $\phi : \mathbb{R} \to \mathbb{R}$. Then the random variable $Y = \phi(X)$ is defined by

$$\boxed{Y(\omega) \equiv \phi(X)(\omega) = \phi(X(\omega))}$$

Since $X : \Omega \to \mathbb{R}$, it follows that $\phi(X) : \Omega \to \mathbb{R}$. Thus $Y = \phi(X)$ is also a random variable and the above definitions apply. In particular, we have

$$\begin{aligned}\mathrm{E}\{\phi(X)\} &= \sum_i \phi(x_i)p_X(x_i). \\ &= \sum_\omega \phi(X(\omega))\mathrm{P}(\{\omega\})\end{aligned}$$

The first expression on the right-hand side averages the values of $\phi(x)$ over the distribution of X, whereas the second expression averages the values of $\phi(X(\omega))$ over the probabilities of $\omega \in \Omega$. A third method would be to compute the distribution of Y and average the values of y over the distribution of Y.

Example 1.8 (1.1 ctd. III)**.** *Let X be the random variable indicating the number of heads on two tosses. Consider the transformation ϕ with $\phi(0) = \phi(2) = 0$ and $\phi(1) = 1$. Find $E[X]$ and $E[\phi(X)]$.*

The **variance** of X is

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}[X - \mathrm{E}[X]]^2.$$

Equivalently $\sigma^2 = \mathrm{E}[X^2] - \{\mathrm{E}[X]\}^2$ (*exercise: prove*). The square root, σ, of σ^2 is called the **standard deviation**.

Example 1.9 (1.1 ctd. IV)**.** *Find $Var(X)$ and $Var\{\phi(X)\}$.*

Linear functions of X. The following properties of expectation and variance are easily proved (*exercise/previous notes*):

$$\boxed{\mathrm{E}[a + bX] = a + b\mathrm{E}[X],\ \mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)}$$

Example 1.10 (1.1 ctd. V)**.** *Let Y be the excess of heads over tails obtained on the two tosses of the coin. Write down $E[Y]$ and $Var(Y)$.*

Standard distributions. For ease of reference, Appendices 1 and 2 provide definitions of standard discrete and continuous distributions given in earlier courses.

Learning Outcomes: *Most of the material in STAT1004 and STAT1005 (or STAT7501) is relevant and important for STAT2001/3101. Students are* **strongly** *advised to revise this material if they don't feel confident about basic probability.*

In particular, regarding subsections 1.1 and 1.2, you should be able to

1. *Explain the concept of (mutual) independence of events and apply it to new situations and examples;*
2. *Define conditional independence and verify it in a concrete situation;*
3. *name and check properties of pdfs, cdfs and pmfs in concrete examples and decide whether a given random variable is discrete, continuous or mixed*
4. *Compute the expectation (of a transformation) of a discrete or continuous random variable.*
5. *Be familiar with standard discrete and continuous distributions.*

1.3 Joint distributions

1.3.1 The joint CDF

Let us first consider the bivariate case. Suppose that the two random variables X and Y share the same sample space Ω (*e.g.* the height and the weight of an individual). Then we can consider the event

$$\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}$$

and define its probability, regarded as a function of the two variables x and y, to be the **joint (cumulative) distribution function** of X and Y, denoted by

$$\begin{aligned} F_{X,Y}(x,y) &= \mathrm{P}(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}) \\ &= \mathrm{P}(X \leq x, Y \leq y). \end{aligned}$$

It is often helpful to think geometrically about X and Y: In fact, (X,Y) is a random point on the two-dimensional Euclidean plane, $\mathbb{R}^2$, i.e. each outcome of the pair of random variables X and Y, or equivalently each outcome of the bivariate random variable (X,Y) corresponds to the point in $\mathbb{R}^2$ whose horizontal coordinate is X and whose vertical coordinate is Y. For this reason, (X,Y) is also called a *random vector.* $F_{X,Y}(x,y)$ is then simply the probability that the point lands in the semi-infinite rectangle $(-\infty, x] \times (-\infty, y] = \{(a,b) \in \mathbb{R}^2 \,:\, a \leq x \,\text{and}\, b \leq y\}$.

The joint cumulative distribution function (cdf) has similar *properties* to the univariate cdf. If the function $F_{X,Y}(x,y)$ is the joint distribution function of random variables X and Y then

1. $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$ and $F_{X,Y}(\infty,\infty) = 1$ and
2. $F_{X,Y}$ is a non decreasing function of each of its arguments
3. $F_{X,Y}$ must also be *right-continuous.* That is, $F_{X,Y}(x+h, y+k) \to F_{X,Y}(x,y)$ as $h, k \downarrow 0$ for all x, y.

The **marginal** cdfs of X and Y can be found from

$$F_X(x) = \mathrm{P}(X \leq x, Y < \infty) = F_{X,Y}(x, \infty)$$

and

$$F_Y(y) = \mathrm{P}(X < \infty, Y \leq y) = F_{X,Y}(\infty, y)$$

respectively.

We already know in the univariate case that $\mathrm{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Similarly, we find in the bivariate case that

$$\mathrm{P}(x_1 < X \leq x_2, \;\; y_1 < Y \leq y_2) =$$
$$F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1)$$

Understanding this expression is straightforward using the geometric interpretation: To calculate the probability of (X, Y) lying in the rectangle $(x_1, x_2] \times (y_1, y_2]$ one takes the probability of lying in the rectangle $(-\infty, x_2] \times (-\infty, y_2]$ and subtracts two probabilities: firstly, the probability of landing in the rectangle $(-\infty, x_1] \times (-\infty, y_2]$ and secondly the probability of landing in the rectangle $(-\infty, x_2] \times (-\infty, y_1]$. Unfortunately, we have now subtracted the probability that we land in the rectangle $(-\infty, x_1] \times (\infty, y_1]$ twice, so we need to add it again to compensate for this mistake. It may help for you to draw a sketch of all those rectangles here:

Example 1.11. *Consider the function*

$$F_{X,Y}(x, y) = x^2y + y^2x - x^2y^2\,, \quad 0 \leq x \leq 1,\, 0 \leq y \leq 1\,.$$

extended by suitable constants outside $(0, 1)^2$ *such as to make it a cdf. Show that* $F_{X,Y}$ *has the properties of a cdf mentioned above. Find the marginal cdfs of* X *and* Y*. Also find* $P(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2})$.

1.3.2 Joint distribution: the discrete case

Cumulative distribution functions fully specify the distribution of a random variable - they encode everything there is to know about that distribution. However, as Example 1.11 showed, they can be somewhat difficult to handle. Making additional assumptions about the random variables makes their distribution easier to handle, so let's assume in this part that X and Y take only values in a countable set, i.e. that (X,Y) is a discrete bivariate random variable. Then $F_{X,Y}$ is a step function in each variable separately and we consider the **joint probability mass function**

$$p_{X,Y}(x_i, y_j) = \mathrm{P}(X = x_i, Y = y_j).$$

It is often convenient to represent a discrete **bivariate** distribution — a joint distribution of two variables — by a **two-way table**. In general, the entries in the table are the *joint probabilities* $p_{X,Y}(x,y)$, while the row and column totals give the *marginal* probabilities $p_X(x)$ and $p_Y(y)$. As always, the total probability is 1.

Example 1.12. *Consider three independent tosses of a fair coin. Let X = 'number of heads in first and second toss' and Y = 'number of heads in second and third toss'. Give the probabilities for any combination of possible outcomes of X and Y in a two-way table and obtain the marginal pmfs of X and Y.*

In general, from the joint distribution we can use the law of total probability to obtain the **marginal pmf** of Y as

$$\begin{aligned} p_Y(y_j) = \mathrm{P}(Y = y_j) &= \sum_{x_i} \mathrm{P}(X = x_i,\ Y = y_j) \\ &= \sum_{x_i} p_{X,Y}(x_i, y_j)\,. \end{aligned}$$

Similarly, the **marginal pmf** of X is given by

$$p_X(x_i) = \sum_{y_j} p_{X,Y}(x_i, y_j)\,.$$

The marginal distribution is thus the distribution of just one of the variables.

The joint cdf can be written as

$$F_{X,Y}(x,y) = \sum_{x_i \le x} \sum_{y_j \le y} p_{X,Y}(x_i, y_j)\,.$$

Note that there will be jumps in $F_{X,Y}$ at each of the x_i and y_j values.

Independence

The random variables X and Y, defined on the sample space Ω with probability measure P, are **independent** if the events

$$\{X = x_i\} \text{ and } \{Y = y_j\}$$

are *independent events*, for all possible values x_i and y_j. Thus X and Y are independent if

$$p_{X,Y}(x_i, y_j) = \mathrm{P}(X = x_i, Y = y_j) = p_X(x_i)p_Y(y_j) \tag{1.1}$$

for all x_i, y_j. This implies that $\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B)$ for all sets A and B, so that the two events $\{\omega : X(\omega) \in A\}, \{\omega : Y(\omega) \in B\}$ are independent. (*Exercise:* prove this.)

NB: If x is such that $p_X(x) = 0$, then $p_{X,Y}(x, y_j) = 0$ for all y_j and (1.1) holds automatically. Thus it does not matter whether we require (1.1) for all *possible* x_i, y_j *i.e.* those with positive probability, or all *real* x, y. (That is, $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all x, y would be an equivalent definition of independence.)

If X, Y are independent then the entries in the two-way table are the products of the marginal probabilities. In Example 1.12 we see that X and Y are *not* independent.

Conditional probability distributions

These are defined for random variables by analogy with conditional probabilities of events. Consider the conditional probability

$$\begin{aligned}\mathrm{P}(X = x_i \mid Y = y_j) &= \frac{\mathrm{P}(X = x_i,\, Y = y_j)}{\mathrm{P}(Y = y_j)} \\ &= \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}\end{aligned}$$

as a function of x_i, for fixed y_j. Then this is a probability mass function — it is non-negative and

$$\sum_{x_i} \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)} = \frac{1}{p_Y(y_j)} \underbrace{\sum_{x_i} p_{X,Y}(x_i, y_j)}_{p_Y(y_j)} = 1,$$

and it gives the probabilities for observing $X = x_i$ given that we already know $Y = y_j$. We therefore *define* the **conditional probability distribution** of X given $Y = y_j$ as

$$\boxed{p_{X|Y}(x_i|y_j) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}}$$

Conditioning on $Y = y_j$ can be compared to selecting a subset of the population, *i.e.* only those individuals where $Y = y_j$. The conditional distribution $p_{X|Y}$ of X given $Y = y_j$ then describes the distribution of X within this subgroup.

From the above definition we immediately obtain the multiplication rule for pmfs:

$$\boxed{p_{X,Y}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)}$$

which can be used to find a bivariate pmf when we know one marginal distribution and one conditional distribution.

Note that if X and Y are *independent* then $p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_i)$ so that $p_{X|Y}(x_i|y_j) = p_X(x_i)$ *i.e.* the *conditional* distribution is the same as the *marginal* distribution.

In general, X and Y are independent *if and only if* the conditional distribution of X given $Y = y_j$ is the same as the marginal distribution of X *for all* y_j. (This condition is equivalent to $p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j)$ for all x_i, y_j, above).

The conditional distribution of Y given $X = x_i$ is defined similarly.

Example 1.13 (1.12 ctd.)**.** *Obtain the conditional pmf of X given $Y = y$. Use this conditional distribution to verify that X and Y are not independent.*

Example 1.14. *Suppose that R and N have a joint distribution in which $R|N$ is Bin(N, π) and N is Poi(λ). Show that R is Poi$(\pi\lambda)$.*

Conditional expectation

Since $p_{X|Y}(x_i|y_j)$ is a probability distribution, it has a mean or expected value:

$$\mathrm{E}[X|Y = y_j] = \sum_{x_i} x_i \, p_{X|Y}(x_i|y_j)$$

which represents the average value of X among outcomes ω for which $Y(\omega) = y_j$. This may also be written $\mathrm{E}_{X|Y}[X|Y = y_j]$. We can also regard the conditional expectation $\mathrm{E}[X|Y = y_j]$ as the mean value of X in the subgroup characterised by $Y = y_j$.

Example 1.15 (1.12 ctd. II)**.** *Find the conditional expectations $E[X|Y = y]$ for $y = 0, 1, 2$. Plot the graph of the function $\phi(y) = E[X|Y = y]$. What do these values tell us about the relationship between X and Y?*

In general, what is the relationship between the *unconditional* expectation $\mathrm{E}[X]$ and the *conditional* expectation $\mathrm{E}[X|Y = y_j]$?

Activity 1.1. *Collect the joint distribution of X: gender ($x_1 = M, x_2 = F$) and Y: number of cups of tea drunk today ($y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 3\,\mathrm{or\,more}$). Are X and Y independent? What is the expectation of Y? What is the conditional expectation $Y|X = M$ and $Y|X = F$?*

We see from the above activity that the overall mean is just the average of the conditional means. We now prove this fact in general. Consider the conditional expectation $\phi(y) = \mathrm{E}_{X|Y}[X|Y = y]$ as a function of y. This function ϕ may be used to transform the random variable Y, i.e. we can consider the new random variable $\phi(Y)$. This random variable is usually written simply $\mathrm{E}_{X|Y}[X|Y]$ because the possibly more correct notation $\mathrm{E}_{X|Y}[X|Y = Y]$ would be even more confusing! We may then compute the expectation of our new random variable $\phi(Y)$, i.e. $\mathrm{E}[\phi(Y)] = \mathrm{E}_Y[\mathrm{E}_{X|Y}[X|Y]]$. But, from the definition of the expectation of a function of Y, we have $\mathrm{E}[\phi(Y)] = \sum_{y_j} \phi(y_j)p_{_Y}(y_j)$, so that

$$\mathrm{E}_Y[\mathrm{E}_{X|Y}[X|Y]] = \sum_{y_j} \underbrace{\mathrm{E}[X|Y = y_j]}_{\text{function of } y_j}\; p_{_Y}(y_j)$$

This gives the marginal expectation of $\mathrm{E}[X]$ as will be shown in the lectures. That is,

$$\boxed{\mathrm{E}[X] = \mathrm{E}_Y[\mathrm{E}_{X|Y}[X|Y]]}$$

which is known as the **iterated conditional expectation** formula. It is most useful when the conditional distribution of X given $Y = y$ is known and easier to handle than the joint distribution (requiring integration/summation to find the marginal of X if it is not known).

Example 1.16 (1.12 ctd. II)**.** *Verify that $E[X] = E_Y[E_{X|Y}[X|Y]]$ in this example.*

Example 1.17 (1.14 ctd.)**.** *Find the mean of R using the iterated conditional expectation formula.*

Note that the definition of expectation generalises immediately to functions of two variables, *i.e.*

$$\begin{aligned}
\mathrm{E}\left[\phi(X,Y)\right] &= \sum_{\omega} \phi(X(\omega),Y(\omega))\mathrm{P}\left(\{\omega\}\right) \\
&= \sum_{x_i}\sum_{y_j} \phi(x_i,y_j)\mathrm{P}\left(\{\omega : X(\omega)=x_i, Y(\omega)=y_j\}\right) \\
&= \sum_{x_i}\sum_{y_j} \phi(x_i,y_j)p_{X,Y}(x_i,y_j)
\end{aligned}$$

and that the above result on conditional expectations generalises too, since

$$\begin{aligned}
\mathrm{E}\left[\phi(X,Y)\right] &= \sum_{x_i}\sum_{y_j} \phi(x_i,y_j)p_{X|Y}(x_i|y_j)p_Y(y_j) \\
&= \sum_{y_j} p_Y(y_j)\underbrace{\sum_{x_i}\phi(x_i,y_j)p_{X|Y}(x_i|y_j)}_{\mathrm{E}_{X|Y}[\phi(X,y_j)|y_j]} \\
&= \mathrm{E}_Y[\mathrm{E}_{X|Y}[\phi(X,Y)|Y]]\,.
\end{aligned}$$

Taking out what is known (TOK)

$$\mathrm{E}_{X|Y}[\phi(Y)\psi(X,Y)|Y] = \phi(Y)\mathrm{E}_{X|Y}[\psi(X,Y)|Y]$$

This will be shown in lectures for discrete random variables only. It also holds for continuous and mixed random variables, however.

Example 1.18. *Consider two discrete random variables X and Y, where the marginal probabilities of Y are $P(Y=0)=3/4$, $P(Y=1)=1/4$ and the conditional probabilities of X are $P(X=1|Y=0)=P(X=2|Y=0)=1/2$ and $P(X=0|Y=1)=P(X=1|Y=1)=P(X=2|Y=1)=1/3$. Use the iterated conditional expectation formula to find $E(XY)$.*

1.3.3 Joint Distribution: the continuous case

We consider now the case where both X and Y take values in a continuous range (i.e. their set of possible values is uncountable) and their joint distribution function $F_{X,Y}(x,y)$ can be expressed as

$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u}$$

where $f_{X,Y}(x,y)$ is the **joint probability density function** of X and Y. In short, we consider a bivariate continuous random variable (X,Y).

Letting $y \to \infty$ we get

$$F_X(x) = F_{X,Y}(x,\infty) \; = \; \int_{-\infty}^{x} \Big(\int_{-\infty}^{\infty} f_{X,Y}(u,v)\mathrm{d}v \Big) \mathrm{d}u \, .$$

But from §1.2 we also know that $F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$. It follows that the **marginal** density function of X is

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,v)\mathrm{d}v}$$

Similarly, Y has **marginal** density

$$\boxed{f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u,y)\mathrm{d}u}$$

As for the univariate case, we have

$$\begin{aligned} \mathrm{P}(x < X \le x + \mathrm{d}x, y < Y \le y + \mathrm{d}y) \;&=\; \int_{x}^{x+\mathrm{d}x} \int_{y}^{y+\mathrm{d}y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u \\ &\simeq\; f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \, . \end{aligned}$$

That is, $f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$ is the probability that (X,Y) lies in the infinitesimal rectangle $(x, x+\mathrm{d}x) \times (y, y+\mathrm{d}y)$. As in the univariate case, $\mathrm{P}(X = x, \; Y = y) = 0$ for all x, y.

Example 1.19. *Consider two continuous random variables X and Y with joint density*

$$f_{X,Y}(x,y) = \begin{cases} 8xy & 0 \le x \le y \le 1 \\ 0 & \textit{otherwise} \end{cases}$$

Sketch the area where $f_{X,Y}$ is positive. Derive the marginal pdfs of X and Y.

Independence

By analogy with the discrete case, two random variables X and Y are said to be **independent** if their joint density factorises, *i.e.* if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x, y.$$

An equivalent characterisation of independence reads as follows:

Two continuous random variables are independent if and only if there exist functions $g(\cdot)$ and $h(\cdot)$ such for all (x, y) the joint density factorises as $f_{X,Y}(x, y) = g(x)h(y)$, where g is a function of x only and h is a function of y only.

Proof. If X and Y are independent then simply take $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. For the converse, suppose that $f_{X,Y}(x, y) = g(x)h(y)$ and define

$$G = \int_{-\infty}^{\infty} g(x)dx, \qquad H = \int_{-\infty}^{\infty} h(y)dy.$$

Note that both G and H are finite (why?). Then the marginal densities are $f_X(x) = g(x)H$, $f_Y(y) = Gh(y)$ and either of these equations implies that $GH = 1$ (integrate wrt. x in the first equation or wrt. y in the second equation to see this). It follows that

$$f_{X,Y}(x, y) = g(x)h(y) = \frac{f_X(x)}{H}\frac{f_Y(y)}{G} = f_X(x)f_Y(y)$$

and so X and Y are independent. □

The advantage of knowing that under independence $f_{X,Y}(x, y) = g(x)h(y)$ is that we don't need to find the marginal densities $f_X(x)$ and $f_Y(y)$ (which would typically involve some integration) to verify independence. It suffices to know $f_X(x)$ and $f_Y(y)$ up to some unknown constant.

Example 1.20 (1.19 ctd.). *Are X and Y independent?*

Conditional distributions

For the conditional distribution of X given Y, we cannot condition on $Y = y$ in the usual way, as for any arbitrary set A, $\mathrm{P}(X \in A \text{ and } Y = y) = \mathrm{P}(Y = y) = 0$ when Y is continuous, so that

$$\mathrm{P}(X \in A \mid Y = y) = \frac{\mathrm{P}(X \in A, Y = y)}{\mathrm{P}(Y = y)}$$

is not defined (0/0). However, we can consider

$$\frac{\mathrm{P}(x < X \leq x + \mathrm{d}x, y < Y \leq y + \mathrm{d}y)}{\mathrm{P}(y < Y \leq y + \mathrm{d}y)} \quad \simeq \quad \frac{f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y}{f_Y(y)\mathrm{d}y}$$

and interpret $f_{X,Y}(x, y)/f_Y(y)$ as the **conditional density** of X given $Y = y$ written as $f_{X|Y}(x \mid y)$.

Note that this *is* a probability density function — it is non-negative and

$$\int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)} \mathrm{d}x = \frac{1}{f_Y(y)} \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x}_{f_Y(y)} = 1.$$

If X and Y are independent then, as before, the conditional density of X given $Y = y$ is just the marginal density of X.

Example 1.21 (1.19 ctd. II)**.** *Give the conditional densities of X given $Y = y$ and of Y given $X = x$ indicating clearly the area where they are positive. Also, find $E[X|Y = y]$ and $E[X]$, using the law of iterated conditional expectation for the latter. Compare this with the direct calculation of $E[X]$.*

1.4 Further results on expectations

1.4.1 Expectation of a sum

Consider the sum $\phi(X)+\psi(Y)$ when X, Y have joint probability mass function $p_{X,Y}(x, y)$. (The continuous case follows similarly, replacing probability mass functions by probability densities and summations by integrals.) Then

$$\begin{aligned}
\mathrm{E}_{X,Y}[\phi(X) + \psi(Y)] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i) + \psi(y_j)\}\, p_{X,Y}(x_i, y_j) \\
&= \sum_{x_i} \phi(x_i) \underbrace{\sum_{y_j} p_{X,Y}(x_i, y_j)}_{p_X(x_i)} + \sum_{y_j} \psi(y_j) \underbrace{\sum_{x_i} p_{X,Y}(x_i, y_j)}_{p_Y(y_j)} \\
&= \mathrm{E}_X[\phi(X)] + \mathrm{E}_Y[\psi(Y)]\,.
\end{aligned}$$

Note that the subscripts on the E's are unnecessary as there is no possible ambiguity in this equation, and also that this holds regardless of whether or not X and Y are independent.

In particular we have $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$. Note the power of this result: there is no need to calculate the probability distribution of $X + Y$ (which may be hard!) if all we need is the mean of $X + Y$.

1.4.2 Expectation of a product

Now consider $\phi(X)\psi(Y)$. Then

$$\begin{aligned} \mathrm{E}_{X,Y}[\phi(X)\psi(Y)] &= \sum_{x_i}\sum_{y_j}\{\phi(x_i)\psi(y_j)\}p_{X,Y}(x_i,y_j) \qquad (1.2) \\ &= \ ? \end{aligned}$$

If X and Y are *independent*, then $p_{X,Y}(x_i,y_j) = p_X(x_i)\ p_Y(y_j)$ and the double sum in (1.2) factorises, that is

$$\mathrm{E}_{X,Y}[\phi(X)\psi(Y)] = \underbrace{\sum_{x_i}\phi(x_i)p_X(x_i)}_{\mathrm{E}_X[\phi(X)]}\ \underbrace{\sum_{y_j}\psi(y_j)p_Y(y_j)}_{\mathrm{E}_Y[\psi(Y)]}.$$

Thus, *except* for the case where X and Y are *independent*, we typically have that

E(product) $\neq$ product of expectations

even though, from above, it is *always* true that

E(sum) $=$ sum of expectations

Slogan:

Independence means Factorising

1.4.3 Covariance

A particular function of interest is the **covariance** between X and Y. As we will see, this is a measure for the strength of the linear relationship between X and Y. The covariance is defined as

$$\boxed{\mathrm{Cov}(X,Y) = \mathrm{E}\ [(X - \ \mathrm{E}[X])(Y - \mathrm{E}[Y])]}$$

An alternative formula for the covariance follows on expanding the bracket, giving

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}\ [XY - X\mathrm{E}[Y] - Y\mathrm{E}[X] + \mathrm{E}[X]\ \mathrm{E}[Y]] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\ \mathrm{E}[Y] - \mathrm{E}[X]\ \mathrm{E}[Y] + \mathrm{E}[X]\ \mathrm{E}[Y] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\ \mathrm{E}[Y] \end{aligned}$$

Note that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$, giving the familiar formula $\mathrm{Var}(X) = \mathrm{E}[X^2] - \{\mathrm{E}[X]\}^2$.

If X and Y are *independent* then, from above,

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$$

and it follows that

$$\mathrm{Cov}(X,Y) = 0\,.$$

However in general $\mathrm{Cov}(X,Y) = 0 \not\Rightarrow X$ and Y are independent! An example for this will be given below in Example 1.23.

Also, if $Z = aX + b$ then $\mathrm{E}[Z] = a\mathrm{E}[X] + b$ and $Z - \mathrm{E}[Z] = a\{X - \mathrm{E}[X]\}$, so that

$$\begin{aligned}\mathrm{Cov}(Z,Y) &= \mathrm{E}[a(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]\\ &= a\mathrm{Cov}(X,Y).\end{aligned}$$

Using a similar argument we get

$$\mathrm{Cov}(X+Y,W) = \mathrm{Cov}(X,W) + \mathrm{Cov}(Y,W).$$

Exercise: Using the fact that $\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y, X+Y)$, derive the general formula $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$.

1.4.4 Correlation

From above, we see that the covariance varies with the scale of measurement of the variables (lbs/kilos etc), making it difficult to interpret its numerical value. The correlation is a standardised form of the covariance, which is *scale-invariant* and therefore its values are easier to interpret.

The **correlation** between X and Y is defined by

$$\boxed{\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}}$$

Suppose that $a > 0$. Then $\mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$ and $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$ and it follows that $\mathrm{Corr}(aX,Y) = \mathrm{Corr}(X,Y)$. Thus the correlation is **scale-invariant**.

A key result is that

$$\boxed{-1 \leq \operatorname{Corr}(X, Y) \leq +1}$$

for all random variables X and Y.

Example 1.22 (1.18 ctd.)**.** *Find the covariance and correlation of X and Y.*

Example 1.23. *Compute the correlation of $X \sim U(-1, 1)$ and $Y = X^2$. Sketch a typical scatter plot of X and Y, e.g. for a sample of size 20. Are X and Y independent?*

STAT3101 only:
To prove this, we use the following trick. For any constant $z \in \mathbb{R}$,

$$\begin{aligned} \text{Var}(zX+Y) &= \text{E}\left[(zX+Y)-(z\text{E}[X]+\text{E}[Y])\right]^2 \\ &= \text{E}\left[z(X-\text{E}[X])+(Y-\text{E}[Y])\right]^2 \\ &= z^2\text{E}\left[X-\text{E}[X]\right]^2 + 2z\text{E}\ \left[(X-\text{E}[X])(Y-\text{E}[Y])\right] + \text{E}\left[Y-\text{E}[Y]\right]^2 \\ &= z^2\,\text{Var}(X) + 2z\,\text{Cov}(X,Y) + \text{Var}(Y) \end{aligned}$$

as a quadratic function of z. But $\text{Var}(zX+Y) \geq 0$, so the quadratic on the right-hand side must have either no real roots or a single repeated root, *i.e.* we must have ("$b^2 \leq 4ac$"). Therefore

$$\{2\text{Cov}(X,Y)\}^2 \leq 4\text{var}(X)\text{var}(Y)$$

which implies that $\text{Corr}^2(X,Y) \leq 1$, as claimed. □

We get the extreme values, $\text{Corr}(X,Y) = \pm 1$, when the quadratic touches the z-axis; that is, when $\text{Var}\,(zX+Y) = 0$. But if the variance of a random variable is zero then the random variable must be a constant (we say that its distribution is *degenerate*). Therefore, letting z be the particular value for which the quadratic touches the z-axis, we obtain

$$zX + Y = \text{constant}. \tag{1.3}$$

Taking expectations of this we find that the constant is given by

$$\text{constant} = z\text{E}[X] + \text{E}[Y]. \tag{1.4}$$

Additionally, we can translate equation (1.3) to say $zX = \text{constant} - Y$ so that taking variances on both sides yields

$$z^2\text{Var}(X) = \text{Var}(Y). \tag{1.5}$$

Thus, the quadratic equation $z^2\text{Var}(X) + 2z\text{Cov}(X,Y) + \text{Var}(Y) = 0$ implies that $z^2\text{Var}(X)+2z\text{Cov}(X,Y)+z^2\text{Var}(X) = 0$ and thus $z = -\text{Cov}(X,Y)/\text{Var}(X)$ follows (the case $z = 0$ corresponds to $\text{Var}(Y) = 0$ and thus $\text{Var}(X) = 0$ in which case both random variables are degenerate).
Now take equation (1.3), subtract equation (1.4) and substitute for z to obtain finally:

$$\boxed{Y - \text{E}[Y] = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - \text{E}[X])}.$$

We therefore see that correlation measures the *degree of linearity* of the relationship between X and Y, and takes its maximum and minimum values (± 1) when there is an *exact* linear relationship between them. As there may be other forms of dependence between X and Y (*i.e.* non–linear dependence), it is now clear that $\text{corr}(X,Y) = 0 \not\Rightarrow$ independence.

1.4.5 Conditional variance

Consider random variables X and Y and the conditional probability distribution of X given $Y = y$. This conditional distribution has a mean, denoted $\mathrm{E}(X|Y = y)$, and a variance, $\mathrm{var}(X|Y = y)$. We have already shown that the marginal (unconditional) mean $\mathrm{E}(X)$ is related to the conditional mean via the formula

$$\mathrm{E}[X] = \mathrm{E}_Y[\mathrm{E}_{X|Y}[X|Y]]\,.$$

In the lectures we will obtain a similar result for the relation between the marginal and conditional variances. The result is that

$$\boxed{\mathrm{Var}(X) = \mathrm{E}_Y[\mathrm{Var}(X|Y)] + \mathrm{Var}_Y\{\mathrm{E}[X|Y]\}}$$

Example 1.24 (1.18 ctd.)**.** *Find the conditional variances of X given $Y = 0, 1$. Compute the marginal variance of X by using the above result.*

Example 1.25 (1.14 ctd. II)**.** *Find the variance of R using the iterated conditional variance formula.*

Learning Outcomes: *Sections 1.3 and 1.4 represent the base of STAT2001/3101. A thorough understanding of the material is essential in order to follow the remaining sections as well as many courses in the second and third year.*

Joint Distributions *You should be able to*

1. *Name and verify the properties of joint cdfs;*
2. *Compute probabilities of rectangles using the cdf;*
3. *Define the marginal and conditional pmf / pdf in terms of the joint distribution;*
4. *Represent the joint distribution of discrete variables in a two-way table and identify the marginal distributions;*
5. *Compute the marginal and conditional distributions (pmf / pdf) from the joint distribution and vice versa;*
6. *Compute probabilities for joint and conditional events using the joint or conditional pmf / pdf or cdf as appropriate.*

Expectation / Variance *You should be able to*

1. *Calculate the expectation of functions of more than two variables; in particular, find expectations of sums and products;*
2. *Find / compute the conditional expectations given a pair of discrete or continuous random variables and their joint or conditional distribution;*
3. *Use the law of iterated conditional expectation to find marginal expectations given only the conditional distribution;*
4. *Apply iterated conditional expectation to find the expectation of a product, and use iterated conditional expectation sensibly to get expectations of more complex transformations;*
5. *Use the "Taking out what is known" rule to simplify conditional expectations*
6. *Compute and interpret conditional variances in simple cases;*
7. *Compute marginal variances when only conditional distributions are given using the result on iterated conditional variance.*

Independence *You should be able to*

1. *Infer from joint or conditional distributions whether variables are independent;*

2. *Apply the main criteria to check independence of two random variables, and identify the one that is easiest to check in a given situation;*
3. *Explain the relation between independence and uncorrelatedness.*

Covariance / Correlation *You should be able to*

1. *Compute the covariance of two variables using the simplest possible way for doing so in standard situations;*
2. *Compute the correlation of two random variables, and interpret the result in terms of linear dependence;*
3. *Derive the covariance / correlation for simple linear transformations of the variables;*
4. *State the main properties of the correlation coefficient;*
5. *Sketch the proof of* $-1 \leq \mathit{Corr} \leq 1$.

1.5 Standard multivariate distributions

1.5.1 From bivariate to multivariate

The idea of joint probability distributions extends immediately to more than two variables, giving general **multivariate** distributions, *i.e.* the variables $X_1, \ldots, X_n$ have a joint cumulative distribution function

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

and may have a joint probability mass function

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{P}(X_i = x_i;\ i = 1, \ldots, n)$$

or joint probability density function

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n),$$

so that a function $\phi(X_1, \ldots, X_n)$ has an expectation with respect to this joint distribution etc.

Conditional distributions of a subset of variables given the rest then follow as before; for example, for discrete random variables X_1, X_2, X_3,

$$p_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) = \frac{p_{X_1,X_2,X_3}(x_1, x_2, x_3)}{p_{X3}(x_3)}$$

is the conditional pmf of (X_1, X_2) given $X_3 = x_3$. Similarly, the discrete random variables $X_1, \ldots, X_n$ are **(mutually) independent** if and only if

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

for all $x_1, \ldots, x_n$. Independence of $X_1, \ldots, X_n$ implies independence of the events $\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$ (*exercise: prove*). Finally, we say that X_1 and X_2 are **conditionally independent** given X_3 if

$$p_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) = p_{X_1|X_3}(x_1 \mid x_3)\, p_{X_2|X_3}(x_2 \mid x_3)$$

for all x_1, x_2, x_3. These definitions hold for continuous distributions by replacing the pmf by the pdf.

1.5.2 The multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution. Suppose that a sample of size n is drawn (*with replacement*) from a population whose members fall into one of $m + 1$ categories. Assume that, for each individual sampled, independently of the rest

$$\mathrm{P}(\text{individual is of type } i) = p_i, \qquad i = 1, \ldots, m + 1$$

where $\sum_{i=1}^{m+1} p_i = 1$. Let N_i be the number of type i individuals in the sample. Note that, since $N_{m+1} = n - \sum_{i=1}^{m} N_i$, N_{m+1} is determined by $N_1, \ldots, N_m$. We therefore only need to consider the joint distribution of the m random variables $N_1, \ldots, N_m$.

The joint pmf of $N_1, \ldots, N_m$ is given by

$$\boxed{\mathrm{P}(N_i = n_i, i = 1, \ldots, m) = \begin{cases} \frac{n!}{n_1! \ldots n_{m+1}!} p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}, & n_1, \ldots, n_m \in \{0, 1, 2, \ldots, n\}, \\ & n_1 + \ldots + n_m \leq n \\ 0 & \text{otherwise.} \end{cases}}$$

where $n_{m+1} = n - \sum_{i=1}^{m} n_i$. This is the **multinomial distribution** with index n and parameters $p_1, \ldots, p_m$, where $p_{m+1} = 1 - \sum_{i=1}^{m} p_i$ (so p_{m+1} is not a ‘free’ parameter).

To justify the above joint pmf note that we want the probability that the n trials result in exactly n_1 outcomes of the first category, n_2 of the second, ... , n_{m+1} in the last category. Any specific ordering of these n outcomes has probability $p_1^{n_1} \dots p_{m+1}^{n_{m+1}}$ by the assumption of independent trials, and there are $\frac{n!}{n_1! \dots n_{m+1}!}$ such orderings.

If $m = 1$ the multinomial distribution is just the **binomial** distribution, *i.e.*
$N_1 \sim \mathrm{Bin}(n, p_1)$, which has mean np_1 and variance $np_1(1 - p_1)$.

Example 1.26. *Suppose that a bag contains five red, five black and five yellow balls and that three balls are drawn at random with replacement. What is the probability that there is one of each colour?*

Marginal distribution of N_i

Clearly N_i can be regarded as the number of successes in n independent Bernoulli trials if we define *success* to be *individual is of type i*. Thus N_i has a binomial distribution, $N_i \sim \mathrm{Bin}(n, p_i)$, with mean np_i and variance $np_i(1 - p_i)$.

STAT3101 only: It is instructive to derive the marginal distribution of N_1 directly from the joint pmf of $N_1, \dots, N_m$ by using the multinomial expansion as follows. To do this, you will need the result that

$$\sum_{n_1} \cdots \sum_{n_{m+1}} \frac{n!}{n_1! \dots n_{m+1}!} p_1^{n_1} \dots p_{m+1}^{n_{m+1}} = (p_1 + \dots + p_{m+1})^n = 1$$

where the sum is taken over all $n_1, \dots, n_{m+1}$ for which $n_1 + \dots + n_{m+1} = n$. Thus, the probabilities of the multinomial distribution are the terms of the *multinomial expansion.*

Example 1.27. *Let N_A, N_B and N_F be the numbers of A grades, B grades and fails respectively amongst a class of 100 students. Suppose that generally 5% of students achieve grade A, 30% grade B and that 5% fail. Write down the joint distribution of N_A, N_B and N_F and find the marginal distribution of N_A.*

Joint distribution of N_i and N_j

Again we can regard individuals as being one of three types, i, j and k={*not i or j*}. This is the **trinomial** distribution with probabilities

$$\boxed{\mathrm{P}(N_i = n_i, N_j = n_j) = \begin{cases} \frac{n!}{n_i!n_j!n_k!} p_i^{n_i} p_j^{n_j} p_k^{n_k}, & n_i + n_j \le n \\ 0 & \text{otherwise} \end{cases}}$$

where $n_k = n - n_i - n_j$ and $p_k = 1 - p_i - p_j$. It is intuitively clear that N_i and N_j are dependent and negatively correlated, since a relatively large value of N_i implies a relatively small value of N_j and conversely. We show this as follows. First, we have

$$\begin{aligned} \mathrm{E}(N_i N_j) &= \sum_{n_i} \sum_{n_j} n_i n_j \mathrm{P}(N_i = n_i, N_j = n_j) \\ &= \sum_{\{n_i, n_j \ge 0, n_i + n_j \le n\}} n_i n_j \frac{n!}{n_i!n_j!n_k!} p_i^{n_i} p_j^{n_j} p_k^{n_k} \\ &= n(n-1)p_i p_j \sum_{\{n_i - 1, n_j - 1 \ge 0, n_i + n_j - 2 \le n - 2\}} \frac{(n-2)!}{(n_i - 1)!(n_j - 1)!n_k!} p_i^{n_i - 1} p_j^{n_j - 1} p_k^{n_k} \\ &= n(n-1)p_i p_j (p_i + p_j + p_k)^{n-2} \\ &= n(n-1)p_i p_j \end{aligned}$$

The manipulations in the third line are designed to create a multinomial expansion that we can sum. Note that we may take $n_i, n_j \ge 1$ in the sum, since if either n_i or n_j is zero then the corresponding term in the sum is zero.

Finally

$$\mathrm{Cov}(N_i, N_j) = \mathrm{E}[N_i N_j] - \mathrm{E}[N_i]\mathrm{E}[N_j] = n(n-1)p_i p_j - (np_i)(np_j) = -np_i p_j$$

and so

$$\mathrm{Corr}(N_i, N_j) = \frac{-np_i p_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}.$$

Note that $\mathrm{Corr}(N_i, N_j)$ is negative, as anticipated, and also that it does not depend on n.

Conditional distribution of N_i given $N_j = n_j$

Given $N_j = n_j$, there are $n - n_j$ remaining independent Bernoulli trials, each with probability of being type i given by

$$\mathrm{P}(\text{type } i | \text{not type } j) = \frac{\mathrm{P}(\text{type } i)}{\mathrm{P}(\text{not type } j)} = \frac{p_i}{1 - p_j}.$$

Thus, given $N_j = n_j, N_i$ has a binomial distribution with index $n - n_j$ and probability $\frac{p_i}{1-p_j}$.

Exercise: Verify this result by using the definition of conditional probability together with the joint distribution of N_i and N_j and the marginal distribution of N_j.

Example 1.28 (1.27 ctd.). *Find the conditional distribution of N_A given $N_F = 10$ and calculate Corr(N_A, N_F).*

Remark: The multinomial distribution can also be used as a model for *contingency tables*. Let X and Y be discrete random variables with a number of I and J different outcomes, respectively. Then, in a trial of size n, N_{ij} will count the number of outcomes where we observe $X = i$ and $Y = j$. The counts N_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, are typically arranged in a contingency table, and from the above considerations we know that their joint distribution is multinomial with parameters n and $p_{ij} = \mathrm{P}(X = i, Y = j)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. This leads to the analysis of *categorical data*, for which a question of interest is often 'are the categories independent?', *i.e.* is $p_{ij} = p_i p_j$ for all i, j? Exact significance tests of this hypothesis can be constructed from the multinomial distribution of the entries in the contingency table.

1.5.3 The multivariate normal distribution

The continuous random variables X and Y are said to have a **bivariate normal distribution** if they have joint probability density function

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right]$$

for $-\infty < x, y < \infty$, where $-\infty < \mu_X, \mu_Y < \infty; \sigma_X, \sigma_Y > 0; \rho^2 < 1$. The parameters of this distribution are μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. As we will see below, these turn out to be the marginal means, variances, and the correlation of X and Y.

The bivariate normal is widely used as a model for many observed phenomena where dependence is expected, *e.g.* height and weight of an individual, length and width of a petal, income and investment returns. Sometimes the data need to be transformed (*e.g.* by taking logs) before using the bivariate normal.

Marginal distributions

In order to simplify the integrations required to find the marginal densities of X and Y, we set

$$\frac{x-\mu_X}{\sigma_X}=u, \qquad \frac{y-\mu_Y}{\sigma_Y}=v.$$

Then, integrating with respect to y, the marginal density of X can be found as

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u^2-2\rho uv+v^2\right\}\right]\sigma_Y\mathrm{d}v \\ &= \frac{1}{\sigma_X\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2+u^2(1-\rho^2)\right\}\right]\mathrm{d}v \end{aligned}$$

where we have *completed the square* in v in the exponent. Taking the term not involving v outside the integral we then get

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma_X\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{(v-\rho u)^2\right\}\right]\mathrm{d}v \\ &= \frac{1}{\sigma_X\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)=\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\} \end{aligned}$$

The final step here follows by noting that the integrand is the density of a $N(\rho u, 1-\rho^2)$ random variable and hence integrates to one. Thus the marginal distribution of X is normal, with mean μ_X and variance σ_X^2.

By symmetry in X and Y we get that $Y \sim N(\mu_Y, \sigma_Y^2)$ is the marginal distribution of Y.

It will be shown later in chapter 3 that the fifth parameter, ρ, also has a simple interpretation, namely $\rho = \mathrm{Corr}(X,Y)$.

Conditional distributions

The conditional distribution of X given $Y = y$ is found as follows. We have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{\sqrt{2\pi}\sigma_Y}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \right.\right. \\ &\qquad \left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - (1-\rho^2)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right] \end{aligned}$$

Now the expression in $[\cdot]$ can be written as

$$\begin{aligned} & -\frac{1}{2\sigma_X^2(1-\rho^2)}\left\{(x-\mu_X)^2 - 2\rho\frac{\sigma_X}{\sigma_Y}(x-\mu_X)(y-\mu_Y) + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}(y-\mu_Y)^2\right\} \\ = & -\frac{1}{2\sigma_X^2(1-\rho^2)}\left\{(x-\mu_X) - \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right\}^2 \end{aligned} \tag{1.6}$$

and so, finally, we get

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\sigma_X^2(1-\rho^2)}} \exp\left[-\frac{1}{2\sigma_X^2(1-\rho^2)}\left\{x - \mu_X - \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right\}^2\right],$$

which is the density of the $N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X^2(1-\rho^2))$ distribution.

The role of ρ. Note that knowledge of $Y = y$ *reduces the variability* of X by a factor $(1-\rho^2)$. The closer the correlation between X and Y, the smaller the conditional variance becomes. Note also that the conditional mean of X is a *linear function* of y. If y is relatively large then the conditional mean of X is also relatively large if $\rho = \text{Corr}(X,Y) > 0$, or is relatively small if $\rho < 0$.

Suppose that $\rho = 0$. Then

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left[-\frac{1}{2}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right] \\ &= f_X(x)f_Y(y) \end{aligned} \tag{1.7}$$

showing that uncorrelated **normal** variables are independent (remember that this is not true in the general case).

Example 1.29. *Let X be the one-year yield of portfolio A and Y be the one-year yield of portfolio B. From past data, the marginal distribution of X is modelled as $N(7, 1)$, whereas the marginal distribution of Y is $N(8, 4)$ (being a more risky portfolio but having a higher average yield). Furthermore, the correlation between X and Y is 0.5. Assuming that X, Y have a bivariate normal distribution, find the conditional distribution of X given that $Y = 9$ and compare this with the marginal distribution of X. Calculate the probability $P(X > 8|Y = 9)$.*

1.5.4 Reminder: Matrix notation

Matrix Basics

First of all, let's remind some general matrix notation, where an m by n matrix, i.e. a matrix containing m rows and n columns of real numbers is denoted by $(a_{i,j})_{i=1,j=1}^{m,n} = \mathbf{A}$ and is thought of as an element of $\mathbb{R}^{m\times n}$. Matrices are added entry-wise and two matrices $\mathbf{A} \in \mathbb{R}^{k\times m}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ can be multiplied to yield a matrix $\mathbf{C} \in \mathbb{R}^{k\times n}$ where the entries are obtained by taking inner products of rows in $\mathbf{A}$ with columns in $\mathbf{B}$, i.e. the entries are obtained as follows:

$$c_{k,j} = \sum_{i=1}^{m} a_{k,i} b_{i,j}$$

Matrices can be multiplied by real numbers and they can act on column vectors (from the right) and row vectors (from the left), so if $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a matrix and $x \in \mathbb{R}^n$ is a vector with entries x_i and $\alpha \in \mathbb{R}$ is a real number we have

$$\begin{aligned} \alpha \mathbf{A} x &= y \in \mathbb{R}^m \\ y_i &= \alpha \sum_{j=1}^{n} a_{i,j} x_j. \end{aligned}$$

Matrices, vectors and scalars satisfy the usual associative and distributive laws, e.g. $\mathbf{A}(x + y) = \mathbf{A}x + \mathbf{A}y$ and $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ etc. However, note that matrix multiplication, contrary to normal multiplication of real numbers, is not commutative, i.e. in general we have $\mathbf{AB} \neq \mathbf{BA}$.

Transpose

The transpose of a column vector $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ is the row vector $x^T = (x_1, x_2, x_3, \dots, x_n)$, although sometimes we will use column and row vectors interchangeably when no confusion can arise. The transpose of a matrix $\mathbf{A} = (a_{i,j})_{i,j=1}^{m,n}$ is just $\mathbf{A}^T = (a_{j,i})_{i=1,j=1}^{n,m}$ (i.e. you mirror the matrix entries across its diagonal) and the following rules apply for transposition:

$$\begin{aligned}(\alpha x + \beta y)^T &= \alpha x^T + \beta y^T \\ (\alpha \mathbf{A} + \beta \mathbf{B})^T &= \alpha \mathbf{A}^T + \beta \mathbf{B}^T \\ (\mathbf{A}x)^T &= x^T \mathbf{A}^T \\ \left(\mathbf{A}^T\right)^T = \mathbf{A} \quad , &\quad \left(x^T\right)^T = x\end{aligned}$$

Here, matrices are denoted by $\mathbf{A}, \mathbf{B}$, and x, y are vectors and α, β are real numbers.

Inner Product

The inner product, also known as scalar product, of two vectors, $x, y, \in \mathbb{R}^n$ is denoted by $x^T y \in \mathbb{R}$. It is symmetric, i.e. $x^T y = y^T x$, and works with the transpose and inverse of invertible matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$ as follows:

$$\begin{aligned}x^T \mathbf{A} y &= (\mathbf{A}^T x)^T y \\ \left(A^{-1}\right)^T &= \left(A^T\right)^{-1},\end{aligned}$$

where the latter equality is the reason for the abbreviated notation $A^{-T} = \left(A^{-1}\right)^T$ – it doesn't matter whether the transpose or the inverse is carried out first.

Determinants

The determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted either $\det(\mathbf{A})$ or simply $|\mathbf{A}|$, satisfies the following rules:

$$\begin{aligned}\det(\alpha \mathbf{A}) &= \alpha^n \det(\mathbf{A}) \\ \det(\mathbf{A}^T) &= \det(\mathbf{A}) \\ \det(\mathbf{A}^{-1}) &= \frac{1}{\det(\mathbf{A})},\end{aligned}$$

where $\mathbf{A}$ is assumed to be invertible for the last line to hold. A determinant can be computed by proceeding in a column first or a row first fashion and proceeding recursively to the determinants of the sub-matrices created, e.g. in dimension $n = 3$ we have

$$\begin{aligned}\det(\mathbf{A}) &= \det\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} \\ &= a_{1,1}\det\begin{pmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{pmatrix} - a_{2,1}\det\begin{pmatrix} a_{1,2} & a_{1,3} \\ a_{3,2} & a_{3,3} \end{pmatrix} + a_{3,1}\det\begin{pmatrix} a_{1,2} & a_{1,3} \\ a_{2,2} & a_{2,3} \end{pmatrix} \\ &= a_{1,1}(a_{2,2}a_{3,3} - a_{3,2}a_{2,3}) - a_{2,1}(a_{1,2}a_{3,3} - a_{3,2}a_{1,3}) + a_{3,1}(a_{1,2}a_{2,3} - a_{2,2}a_{1,3}),\end{aligned}$$

where the simpler 2D rule

$$\det\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$$

has been used.
The matrix is invertible, i.e. $\mathbf{A}^{-1}$ exists, if and only if its determinant is non-zero, i.e. $\det(\mathbf{A}) \neq 0$.

□

1.5.5 Matrix Notation for Multivariate Normal Random Variables

Define

$$\boldsymbol{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Here we call $\boldsymbol{X}$ a **random vector**, $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ is its **mean vector** and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ is the **covariance matrix**, or **dispersion matrix** of $\boldsymbol{X}$. Then

$$\det(\boldsymbol{\Sigma}) = \sigma_X^2\sigma_Y^2(1-\rho^2)\,, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{\det(\boldsymbol{\Sigma})}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}$$

and, writing $\boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}$,

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= (x-\mu_X, y-\mu_Y)\frac{1}{\det(\boldsymbol{\Sigma})}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\ &= \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)}\{(x-\mu_X)^2\sigma_Y^2 - 2(x-\mu_X)(y-\mu_Y)\rho\sigma_X\sigma_Y \\ &\qquad\qquad + (y-\mu_Y)^2\sigma_X^2\} \\ &= \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}.\end{aligned}$$

It follows that the joint density $f_{\boldsymbol{X}}(\boldsymbol{x})$ of X, Y can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.8}$$

on noting that $\det(2\pi\boldsymbol{\Sigma})^{1/2} = 2\pi\det(\boldsymbol{\Sigma})^{1/2}$. The quantity in $\{\cdot\}$ is a **quadratic form** in $\boldsymbol{x}-\boldsymbol{\mu}$. Note that the above way of writing the joint density resembles much more the univariate density than the explicit formula given at the beginning of the section.

The usefulness of this matrix representation is that the bivariate normal distribution now extends immediately to a general **multivariate** form, with joint density given by (1.8), with

$$\boldsymbol{X} = (X_1, \ldots, X_k)^T, \qquad \boldsymbol{x} = (x_1, \ldots, x_k)^T, \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$$

and

$$(\boldsymbol{\Sigma})_{ij} = \text{cov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$$

Further note that, since $\boldsymbol{\Sigma}$ is $k \times k$, we can write $\det(2\pi\boldsymbol{\Sigma})^{1/2} = (2\pi)^{k/2}\det(\boldsymbol{\Sigma})^{1/2}$.

For this k–dimensional joint distribution, denoted by $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\rho_{ij} = \text{Corr}(X_i, X_j)$ and $\text{var}(X_i) = \sigma_i^2$. It can then be shown that X_i has marginal distribution $N(\mu_i, \sigma_i^2)$, that any two of these variables have a bivariate normal distribution as above, and therefore that the conditional distribution of one variable given the other is also normal.

Example 1.30. *Let X_1, X_2, X_3 have a trivariate normal distribution with mean vector (μ_1, μ_2, μ_3) and covariance matrix*

$$\Sigma = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}.$$

Show that $f_{X_1,X_2,X_3} = f_{X_1}f_{X_2}f_{X_3}$ and give the marginal distributions of X_1, X_2, and X_3.

Learning Outcomes:

Multinomial Distribution *You should be able to*

1. *Identify situations where the multinomial distribution is appropriate;*
2. *Deduce the pmf of the multinomial distribution and make use of the link to the multinomial expansion;*
3. *Derive the marginal and conditional distributions;*
4. *Explain why N_i and N_j have a negative correlation.*

Multivariate Normal Distribution *You should be able to*

1. *Recognise the bivariate normal density, describe its shape and interpret the parameters;*
2. *Name the marginal distributions and know how to derive them;*
3. *Give the conditional distributions, know how to derive them and name the characteristic properties of the conditional distributions;*
4. *Relate the correlation parameter to independence between jointly normal random variables;*
5. *With the help of the foregoing points, characterise situations where the multivariate normal distribution is appropriate;*
6. *Compute probabilities of joint and conditional events;*
7. *Explain how the multivariate normal distribution is constructed using matrix notation.*

Chapter 2

Transformation of Variables

In this section we will see how to derive the distribution of transformed random variables. This is useful because many statistics applied to data analysis (*e.g.* test statistics) are transformations of the sample variables.

2.1 Univariate case

Suppose that we have a sample space Ω, a probability measure P on Ω, a random variable $X : \Omega \to \mathbb{R}$, and a function $\phi : \mathbb{R} \to \mathbb{R}$.

Recall from §1.2 : $Y = \phi(X) : \Omega \to \mathbb{R}$ is defined by $Y(\omega) = \phi(X)(\omega) = \phi(X(\omega))$. Since $Y = \phi(X)$ is a random variable it also has a probability distribution, which can be determined either directly from P or via the distribution of X.

2.1.1 Discrete case

$$\begin{aligned}
\mathrm{P}(Y = y) = \mathrm{P}(\{\omega : \phi(X(\omega)) = y\}) &= \sum_{\{\omega:\phi(X(\omega))=y\}} \mathrm{P}(\{\omega\}) \\
&= \sum_{\{x:\phi(x)=y\}} \mathrm{P}(\{\omega : X(\omega) = x\}) \\
&= \sum_{\{x:\phi(x)=y\}} p_X(x).
\end{aligned}$$

So, for example

$$\begin{aligned} \mathrm{E}(Y) &= \sum_{\omega} \phi(X(\omega))\mathrm{P}(\{\omega\}) \quad \text{with respect to P on } \Omega \\ &= \sum_{x} \phi(x)p_X(x) \quad \text{with respect to distribution of } X \\ &= \sum_{y} y p_Y(y) \quad \text{with respect to distribution of } Y = \phi(X) \end{aligned}$$

Example 2.1. *Consider two independent throws of a fair die. Let X be the sum of the numbers that show up. Give the distribution of X. Now consider the transformation $Y = (X-7)^2$. Derive the distribution of Y.*

2.1.2 Continuous case

Suppose that $Y = \phi(X)$ where ϕ is a strictly increasing and differentiable function. Then,

$$F_Y(y) = \mathrm{P}(\phi(X) \leq y) = \mathrm{P}(X \leq \phi^{-1}(y)) = F_X(\phi^{-1}(y)).$$

The first and third equalities arise simply from the definition of the cdf. The middle equality says that the two probabilities on either side are equal because the events are the same, i.e. $\{\omega \in \Omega : \phi(X(\omega)) \leq y\} = \{\omega \in \Omega : X(\omega) \leq \phi^{-1}(y)\}$. In words this simply means that the event that $\phi(X) \leq y$ happens if and only if the event that $X \leq \phi^{-1}(y)$ happens. To see that this is true ...

Therefore, differentiating with respect to y, Y has density

$$f_Y(y) = f_X(\phi^{-1}(y))\,\frac{\mathrm{d}}{\mathrm{d}y}\phi^{-1}(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\bigg|_{x=\phi^{-1}(y)}$$

where the index $x = \phi^{-1}(y)$ means that any x in the formula has to be replaced by the inverse $\phi^{-1}(y)$ because $f_Y(y)$ is a function of y.

Similarly, if ϕ is decreasing then

$$F_Y(y) = \mathrm{P}(\phi(X) \leq y) = \mathrm{P}(X \geq \phi^{-1}(y)) = 1 - F_X(\phi^{-1}(y))$$

so that

$$f_Y(y) = -f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\bigg|_{x=\phi^{-1}(y)}$$

In the first case $\mathrm{d}y/\mathrm{d}x = \mathrm{d}\phi(x)/\mathrm{d}x$ is positive (since ϕ is increasing), in the second it is negative (since ϕ is decreasing) so either way the transformation formula is

$$\boxed{f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{x=\phi^{-1}(y)}}$$

We can check that the right-hand side of the above formula is a valid pdf as follows. Recall that $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$. Changing variable to $y = \phi(x)$ we have, for ϕ increasing,

$$1 = \int \left\{f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\right\}_{x=\phi^{-1}(y)} \mathrm{d}y$$

so that $f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$ is a valid pdf. Similarly for ϕ decreasing.

Example 2.2. *Consider $X \sim$Uniform$[-\frac{\pi}{2}, \frac{\pi}{2}]$, i.e.*

$$f_X(x) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} \le x \le \frac{\pi}{2} \\ 0 & \textit{otherwise.} \end{cases}$$

Derive the density of $Y = \tan(X)$.

When ϕ is a *many-to-one* function we use the generalised formula $f_Y(y) = \sum f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$, where the summation is over the set $\{x : h(x) = y\}$. That is, we add up the contributions to the density at y from all x values which map to y.

Example 2.3. *Suppose that $f_X(x) = 2x$ on $(0, 1)$ and let $Y = (X - \frac{1}{2})^2$. Obtain the pdf of Y.*

2.2 Bivariate case

2.2.1 General Transformations

For the bivariate case we consider two random variables X, Y with joint density $f_{X,Y}(x, y)$. What is the joint density of transformations $U = u(X, Y)$, $V = v(X, Y)$, where $u(\cdot, \cdot)$ and $v(\cdot, \cdot)$ are functions from $\mathbb{R}^2$ to $\mathbb{R}$, such as the ratio X/Y or the sum $X + Y$?

In order to use the following generalisation of the method of §2.1, we need to assume that u, v are such that each pair (x, y) defines a unique (u, v) and conversely, so that $u = u(x, y)$ and $v = v(x, y)$ are differentiable and invertible. The formula that gives the joint density of U, V is similar to the univariate case but the derivative, as we used it above, now has to be replaced by the *Jacobian* $J(x, y)$ of this transformation.

The result is that $U = u(X, Y), V = v(X, Y)$ have joint density

$$\boxed{f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|J(x, y)|_{\substack{x=x(u,v)\\ y=y(u,v)}}}$$

Again, the index $\substack{x=x(u,v)\\ y=y(u,v)}$ means that the x, y have to be replaced by the suitable transformations involving u, v only.

But how do we get the Jacobian $J(x, y)$? It is actually the determinant of the *matrix of partial derivatives* :

$$J(x, y) = \det\frac{\partial(x, y)}{\partial(u, v)} = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

We finally take its absolute value, $|J(x, y)|$. There are two ways of computing this:

(1) Obtain the inverse transformation $x = x(u, v), y = y(u, v)$, compute the matrix of partial derivatives $\partial(x, y)/\partial(u, v)$ and then its determinant and absolute value.

(2) Alternatively find the determinant $J(u, v)$ from the matrix of partial derivatives of (u, v) with respect to (x, y) and then its absolute value and invert this.

The two methods are equivalent since

$$\frac{\partial(x, y)}{\partial(u, v)} = \left\{\frac{\partial(u, v)}{\partial(x, y)}\right\}^{-1}$$

Which way to choose in a specific case will depend on which functions are easier to derive. But note that the inverse transformations $x = x(u, v)$ and $y = y(u, v)$ are required anyway so that the first approach is often preferable.

Example 2.4. *Let X and Y be two independent exponential variables with $X \sim$ Exp(λ) and $Y \sim$ Exp(μ). Find the distribution of $U = X/Y$*

Example 2.5. *Consider two independent and identically distributed random variables X and Y having a uniform distribution on $[0, 2]$. Derive the joint density of $Z = X/Y$*

and $W = Y$, stating the area where this density is positive. Are Z and W independent? Obtain the marginal density of $Z = X/Y$.

2.2.2 Sums of random variables

The distribution of a sum $Z = X + Y$ of two (not necessarily independent) random variables X and Y can be derived directly as follows.

In the discrete case note that the marginal distribution of Z is

$$\mathrm{P}(Z = z) = \sum_x \mathrm{P}(X = x, Z = z) = \sum_x \mathrm{P}(X = x, Y = z - x)$$

That is,

$$\boxed{p_Z(z) = \sum_x p_{X,Y}(x, z - x)}$$

Analogously, in the continuous case we get

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\,\mathrm{d}x.}$$

Example 2.6. *Let X and Y be two positive random variables with joint pdf*

$$f_{X,Y}(x, y) = xye^{-(x+y)}\,,\ \ x, y > 0\,.$$

Derive and name the distribution of their sum $Z = X + Y$.

2.3 Multivariate case

The ideas of §2.2 extend in a straightforward way to the case of more than two continuous random variables. The general problem is to find the distribution of $\boldsymbol{Y} = \phi(\boldsymbol{X})$, where $\boldsymbol{Y}$ is $s \times 1$ and $\boldsymbol{X}$ is $r \times 1$, from the known distribution of $\boldsymbol{X}$. Here $\boldsymbol{X}$ is the random *vector*

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_r \end{pmatrix}.$$

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

$$\boxed{f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}(\boldsymbol{y}))\,|J(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}(\boldsymbol{y})}}$$

where $J(\boldsymbol{x}) = \left|\frac{d\boldsymbol{x}}{d\boldsymbol{y}}\right|$ is the Jacobian of transformation. Here $\frac{d\boldsymbol{x}}{d\boldsymbol{y}}$ is the matrix of partial derivatives $\left(\frac{d\boldsymbol{x}}{d\boldsymbol{y}}\right)_{ij} = \frac{\partial x_i}{\partial y_j}$.

Case (ii): $s < r$. First transform the s-vector $\boldsymbol{Y}$ to the r-vector $\boldsymbol{Y}'$, where $Y_i' = Y_i$, $i = 1, \ldots, s$, and the other $r - s$ random variables Y_i', $i = s+1, \ldots, r$, are chosen for convenience. Now find the density of $\boldsymbol{Y}'$ as in case (i) and then integrate out $Y_{s+1}', \ldots, Y_r'$ to obtain the marginal density of $\boldsymbol{Y}$, as required. (*c.f.* Examples 2.6 & 2.7 in the bivariate case.)

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

Example 2.7 (linear transformation). *Suppose that $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}$ is an $r \times r$ invertible matrix. Then $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}\boldsymbol{y})|\det(\boldsymbol{A})|^{-1}$, where $|\det(\boldsymbol{A})|$ denotes the absolute value of the determinant of $\boldsymbol{A}$.*

2.4 Approximation of moments

Sometimes we may not need the complete probability distribution of $\phi(X)$, but just the first two moments. Recall that $\mathrm{E}[aX+b] = a\mathrm{E}[X]+b$, so the relation $\mathrm{E}[\phi(X)] = \phi(\mathrm{E}[X])$ is true whenever ϕ is a linear function. However, in general if $Y = \phi(X)$ it will not be true that $\mathrm{E}[Y] = \mathrm{E}[\phi(X)] = \int \phi(x) f_X(x)\mathrm{d}x$ is the same as $\phi(\mathrm{E}[X]) = \phi(\int x f_X(x)\mathrm{d}x)$, (or equivalent summations if X is discrete).

To find moments of Y we can use the distribution of X, as above. However, the sums or integrals involved may be analytically intractable. In practice an approximate answer may be sufficient. Intuitively, if X has mean μ_X and X is not very variable, then we would expect $\mathrm{E}[Y]$ to be quite close to $\phi(\mu_X)$.

Suppose that $\phi(x)$ is a continuous function of x for which the following Taylor expansion about μ_X exists (which requires the existence of the derivatives of ϕ):

$$\phi(x) = \phi(\mu_X) + (x - \mu_X)\phi'(\mu_X) + \frac{1}{2}\,(x - \mu_X)^2\phi''(\mu_X) + \ldots$$

Replacing x by X and taking expectations (or, equivalently, multiplying both sides of the above equation by $f_X(x)$ and integrating over x) term by term, we get

$$\mathrm{E}[\phi(X)] = \phi(\mu_X) + \phi'(\mu_X)\ \underbrace{\mathrm{E}[X-\mu_X]}_{=0} + \frac{1}{2}\ \phi''(\mu_X)\,\underbrace{\mathrm{E}\left[(X-\mu_X)^2\right]}_{\sigma_X^2} + \dots$$

so that

$$\boxed{\mathrm{E}[Y] \approx \phi(\mu_X) + \frac{1}{2}\ \phi''(\mu_X)\sigma_X^2}$$

STAT3101 only:
This approximation will be good if the function ϕ is well-approximated by the quadratic (i.e. second order) Taylor expansion in the region to which f_X assigns significant probability. If this region is large, e.g. because σ_X^2 is large, then ϕ will need to be well-approximated by the quadratic Taylor expansion throughout a large region. If σ_X^2 is small, ϕ may only need to be nearly a quadratic function on a much smaller region.
The rougher approximation $\mathrm{E}(Y) \approx \phi(\mu_X)$ will usually only be good if $\phi''(\mu_X)\sigma_X^2$ is small, which will be the case if σ_X^2 is small and/or $\phi''(\mu_X)$ is small; that is, if X is not very variable and/or ϕ is approximately linear at μ_X. Including the next term in the expansion will usually provide a better approximation.

A usually sufficiently good approximate formula for the variance is based on a first order approximation yielding

$$\begin{aligned}\mathrm{Var}\big\{\phi(X)\big\} &= \mathrm{E}[\phi(X) - \mathrm{E}(\phi(X))]^2 \\ &\approx \mathrm{E}[\phi(X) - \phi(\mu_X)]^2 \approx \mathrm{E}[(X-\mu_X)^2(\phi'(\mu_X))^2] = \{\phi'(\mu_X)\}^2\mathrm{E}[(X-\mu_X)^2]\,,\end{aligned}$$

where we have used the approximation $\mathrm{E}[\phi(X)] \approx \phi(\mu_X)$. Therefore

$$\boxed{\mathrm{Var}(Y) \approx \{\phi'(\mu_X)\}^2\sigma_X^2}$$

Example 2.8. *Consider a Poisson variable X ~Poi(μ). Find approximations to the expectation and variance of $Y = \sqrt{X}$.*

2.5 Order Statistics

Order statistics are a special kind of transformation of the sample variables. Their joint and marginal distributions can be derived by combinatorial considerations.

Suppose that $X_1, \ldots, X_n$ are independent with common density f_X. Denote the ordered values by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. What is the distribution F_r of $X_{(r)}$?

In particular, $X_{(n)} = \max\,(X_1, \ldots, X_n)$ is the **sample maximum** and $X_{(1)} = \min\,(X_1, \ldots, X_n)$ is the **sample minimum**. To find the distribution of $X_{(n)}$, note that $\{X_{(n)} \leq x\}$ and $\{\text{all } X_i \leq x\}$ are the same event – and so have the same probability! Therefore the distribution function of $X_{(n)}$ is

$$F_n(x) = \mathrm{P}(X_{(n)} \leq x) = \mathrm{P}(\text{all } X_i \leq x) = \mathrm{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = \{F_X(x)\}^n$$

since the X_i are independent with the same distribution function F_X. Thus

$$\boxed{F_n(x) = \{F_X(x)\}^n}$$

Furthermore, differentiating this expression we see that the density f_n of $X_{(n)}$ is

$$\boxed{f_n(x) = n\{F_X(x)\}^{n-1} f_X(x)}$$

Using a similar argument for $X_{(1)} = \min\,(X_1, \ldots, X_n)$ we see that

$$F_1(x) = \mathrm{P}(X_{(1)} \leq x) = \mathrm{P}(\text{at least one } X_i \leq x) = 1 - \mathrm{P}(\text{all } X_i > x) = 1 - \{1 - F_X(x)\}^n$$

so the distribution function of $X_{(1)}$ is

$$\boxed{F_1(x) = 1 - \{1 - F_X(x)\}^n}$$

and, differentiating, the pdf f_1 of $X_{(1)}$ is

$$\boxed{f_1(x) = n\{1 - F_X(x)\}^{n-1} f_X(x)}$$

Consider next the situation for general $1 \leq r \leq n$. For $\mathrm{d}x$ sufficiently small we have

$$\begin{aligned} \mathrm{P}(x < X_{(r)} \leq x + \mathrm{d}x) &= \mathrm{P}\left(\begin{array}{l} r-1 \text{ values } X_i \text{ such that } X_i \leq x, \text{ and} \\ \ \text{one value in } (x, x+\mathrm{d}x]\,, \text{ and} \\ n-r \text{ values such that } \ X_i > x + \mathrm{d}x \end{array}\right) \\ &\simeq \underbrace{\frac{n!}{(r-1)!(n-r)!}}_{\text{no. of ways of ordering the } r-1, 1 \text{ and } n-r \text{ values}} \{F_X(x)\}^{r-1} f_X(x)\mathrm{d}x \, \{1 - F_X(x+\mathrm{d}x)\}^{n-r} \end{aligned}$$

Recalling that $f_r(x) = \lim_{\mathrm{d}x \to 0} \mathrm{P}(x < X_{(r)} \leq x + \mathrm{d}x)/\mathrm{d}x$, dividing both sides of the above expression by dx and letting $\mathrm{d}x \to 0$ we obtain the density function of the rth order statistic $X_{(r)}$ as

$$\boxed{f_r(x) = \frac{n!}{(r-1)!(n-r)!} \{F_X(x)\}^{r-1} \{1 - F_X(x)\}^{n-r} f_X(x)}$$

Exercise: Show that this formula gives the previous densities when $r = n$ and $r = 1$.

Activity 2.1. *Collect a random sample of five students from the audience and have their heights* $X_1, \dots, X_5$ *measured. Record the heights and compute the mean height. Reorder the students by height to obtain the first to fifth order statistic,* $X_{(1)}, \dots, X_{(5)}$*. What do the first, third and fifth order statistic,* $X_{(1)}$*,* $X_{(3)}$ *and* $X_{(5)}$ *correspond to?*

Example 2.9. *A village is protected from a river by a dike of height* h*. The maximum water levels* X_i *reached by the river in subsequent years* $i = 1, 2, 3, \dots$ *are modelled as independent following an exponential distribution with mean* $\lambda^{-1} = 10$*. What is the probability that the village will be flooded (in statistical language this would be called a "threshold exceedance") at least once in the next 100 years? How high does the dike need to be to make this probability smaller than* 0.5*?*

The distribution of the sample maximum is an important quantity in the field of extreme value theory as the preceding example showed. Extreme value theory is important for its application in insurance pricing and risk assessment. It turns out that the probability distribution of threshold exceedances follows a universal class of distributions in the limit of high thresholds h, independently of the individual distribution of the X_i – a most remarkable result!

Learning Outcomes: *In Chapter 2, the most important aspects are the following.*

Distributions of Transformations *You should be able to*

1. *Derive the distribution (pmf) of a transformation of a discrete variable for simple cases;*
2. *Apply the general theorem to find the distribution (density) of a transformed continuous variable, specifying the range where the density is positive;*
3. *Apply the general theorem to find the joint distribution of transformations of bivariate random variables, specifying the range where the density is positive;*
4. *Use the Jacobian in the appropriate way for point 3, choosing the simplest of the two possible computations for the specific situation;*
5. *Derive the distribution of the sum of random variables (discrete and continuous case);*
6. *When different methods could be applied, identify the easiest to find the distribution of a transformation.*

Approximation of moments *You should be able to*

1. *Derive an approximation formula for the mean of a general transformation based on the second-order Taylor expansion and compute this for standard cases;*
2. *Derive an approximation formula for the variance of a general transformation based on the first-order Taylor expansion and compute this for standard cases;*
3. *STAT3101 only: Explain in which situations such approximations are good / bad.*

Order Statistics *You should be able to*

1. *Compute the distribution of the rth order statistic (in particular sample maximum and sample minimum) and explain the required probabilistic and combinatorial reasoning;*

Chapter 3

Generating Functions

3.1 Overview

The transformation method presented in the previous chapter may become tedious when a large number of variables is involved, in particular for transformations of the sample variables when the sample size tends to infinity. Generating functions provide an alternative way of determining a distribution (*e.g.* of sums of random variables).

We consider different generating functions for the discrete and continuous case. For the former, these are the probability generating functions (section 3.2) and for the latter both the pgf and the moment generating functions (section 3.3). We will point out the simple connection between pgfs and mgfs and further consider joint generating functions in section 3.4 and apply these to linear combinations of random variables in section 3.5. Finally, in section 3.5.1, we will state and use the Central Limit Theorem with a proof provided for STAT3101 students.

3.2 The probability generating function (pgf)

3.2.1 Definition of the pgf

Suppose that X is a discrete random variable taking values $0, 1, 2\ldots$. Then the **probability generating function** (pgf) $G(z)$ of X is defined as

$$\boxed{G(z) \equiv \mathrm{E}(z^X)}$$

The pgf is a function of particular interest, because it sometimes provides an easy way of determining the distribution of a discrete random variable.

Write $p_i = \mathrm{P}(X = i)$, $i = 0, 1, \ldots$. Then, by the usual expectation formula,

$$\begin{aligned} G(z) &= \mathrm{E}(z^X) = \sum_{i=0}^{\infty} z^i p_i \\ &= p_0 + zp_1 + z^2 p_2 + \cdots \ . \end{aligned}$$

Thus $G(z)$ is a *power series* in z, and p_r is the coefficient of z^r. Note that

$$|G(z)| \leq \sum_i |z|^i p_i \leq \sum_i p_i = 1$$

for all $|z| \leq 1$ and that $G(1) = \sum_i p_i = 1$. The sum is therefore convergent for (at least) $|z| \leq 1$.

We know from the theory of Taylor expansions that the rth derivative $G^{(r)}(0) = p_r r!$, yielding an expression for the probability p_r in terms of the rth derivative of G evaluated at $z = 0$:

$$p_r = \frac{G^{(r)}(0)}{r!}, \qquad r = 0, 1, 2, \ldots \ .$$

In practice it is usually easier to find the power series expansion of G and extract p_r as the coefficient of z^r.

3.2.2 Moments and the pgf

Whereas the probabilities are related to the derivatives of G at $z = 0$, it turns out that the moments of X are related to the derivatives of G at $z = 1$. To see this, note that

$G'(z) = p_1 + 2zp_2 + 3z^2p_3 + \cdots = \sum_{i=1}^{\infty} iz^{i-1}p_i$ so that $G'(1) = \sum_{i=1}^{\infty} ip_i =$ E(X). Thus, we have

$$\boxed{\mathrm{E}(X) = G'(1)}$$

Further, $G''(z) = \sum_{i=2}^{\infty} i(i-1)z^{i-2}p_i$ so that $G''(1) = \sum i(i-1)p_i = \mathrm{E}\{X(X-1)\}$. But we can write $\mathrm{Var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \{\mathrm{E}(X)\}^2$ from which we obtain the formula

$$\boxed{\mathrm{Var}(X) = G''(1) + G'(1) - \{G'(1)\}^2}$$

Example 3.1. *Let $X \sim Poi(\mu)$. Find the pgf $G(z) = E(z^X)$. Also, verify the above formulae for the expectation and variance of X.*

Example 3.2. *Consider the pgf*

$$G(z) = (1 - p + pz)^n \,,$$

where $0 < p < 1$ and $n \geq 1$ is an integer. Find the power expansion of $G(z)$ and hence derive the distribution of the random variable X that has this pgf.

3.3 The moment generating function (mgf)

3.3.1 Definition

Another function of special interest, particularly for continuous variables, is the **moment generating function** (mgf) $M(s)$ of X, defined as

$$\boxed{M(s) \equiv \mathrm{E}(e^{sX}) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\mathrm{d}x}$$

The moment generating function does not necessarily exist for all $s \in \mathbb{R}$, i.e. the integral might be infinite. However, we assume for the following that $M(s)$ is finite for s in some open interval containing zero.

3.3.2 Moments and the mgf

Using the expansion $e^{sx} = 1 + sx + \frac{1}{2!}s^2x^2 + \ldots$ we get

$$M(s) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{s^n x^n}{n!} f_X(x)dx = \sum_{n=0}^{\infty} \frac{s^n}{n!} \int_{-\infty}^{\infty} x^n f_X(x)dx$$

integrating term by term (there are no convergence problems here due to assuming finiteness). It follows that

$$M(s) = \sum_{n=0}^{\infty} \frac{s^n}{n!} \mathrm{E}(X^n)\,.$$

Thus $M(s)$ is a power series in s and the coefficient of s^n is $\mathrm{E}(X^n)/n!$ – hence the name 'moment generating function'.

Again, from the theory of Taylor expansions the rth derivative, $M^{(r)}(0)$, of $M(s)$ at $s = 0$ must therefore equal the rth (raw) moment $\mathrm{E}(X^r)$ of X. In particular we have $M'(0) = \mathrm{E}(X)$ and $M''(0) = \mathrm{E}(X^2)$, so that

$$\boxed{\mathrm{E}(X) = M'(0)}$$

and

$$\boxed{\mathrm{Var}(X) = M''(0) - \{M'(0)\}^2}$$

(Alternatively, and more directly, note that $M'(s) = \mathrm{E}(Xe^{sX})$ and $M''(s) = \mathrm{E}(X^2e^{sX})$ and set $s = 0$.) Note also that $M(0) = \mathrm{E}(e^0) = 1$.

It can be shown that if the moment generating function exists on an open interval including zero then it uniquely determines the distribution.

The pgf tends to be used more for discrete distributions and the mgf for continuous ones, although note that when X takes nonnegative integer values then the two are related by $M(s) = \mathrm{E}(e^{sX}) = \mathrm{E}\{(e^s)^X\} = G(e^s)$.

Example 3.3. *Suppose that X has a gamma distribution with parameters (α, λ). Find the mgf of X. Use this to derive the expectation and variance of X.*

Example 3.4. *Let $X \sim N(\mu, \sigma^2)$ be a normal variable. Find the mgf of X. Use this mgf to obtain the expectation and variance of X.*

3.3.3 Linear Transformations and the mgf

Suppose that $Y = a + bX$ and that we know the mgf $M_X(s)$ of X. What is the mgf of Y? We have

$$M_Y(s) = \mathrm{E}(e^{sY}) = \mathrm{E}\{e^{s(a+bX)}\} = e^{sa}\mathrm{E}(e^{sbX}) = e^{as}M_X(bs)\,.$$

We can therefore easily obtain the mgf of any linear function of X from the mgf of X.

Example 3.5 (3.4 ctd.). *Use the mgf of X to find the distribution of $Y = a + bX$.*

A more general concept is that of the **characteristic function**. This is defined in a similar way to the mgf and has similar properties but involves complex variables. The main advantage over the moment generating function is that the characteristic function of a random variable always exists. However, we will not consider it here.

3.4 Joint generating functions

So far we have considered the pgf or mgf of a single real variable. The joint distribution of a *collection* of random variables $X_1, \dots, X_n$ can be characterised in a similar way by the *joint* generating functions:

The **joint pgf** $G(z_1, \dots, z_n)$ of variables $X_1, \dots, X_n$ is a function of n variables, $z_1, \dots, z_n$, and defined to be

$$\boxed{G(z_1, \dots, z_n) = \mathrm{E}(z_1^{X_1} z_2^{X_2} \cdots z_n^{X_n})}$$

The **joint mgf** $M(s_1, \dots, s_n)$ is a function of n variables, $s_1, \dots, s_n$, and is defined to be

$$\boxed{M(s_1, \dots, s_n) = \mathrm{E}(e^{s_1 X_1 + \cdots + s_n X_n})}$$

These generating functions *uniquely* determine the joint distribution of $X_1, \dots, X_n$. Note that the mgf may also be written in vector notation as

$$M(\boldsymbol{s}) = \mathrm{E}(e^{\boldsymbol{s}^T \boldsymbol{X}})\,,$$

where $\boldsymbol{s} = (s_1, \dots, s_n)^T$ and $\boldsymbol{X} = (X_1, \dots, X_n)^T$.

Generating functions and independence

In both cases (pgf and mgf) we find that if $X_1, \ldots, X_n$ are independent random variables then the pgf / mgf are given as the product of the individual pgfs / mgfs. (Recall: $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ when X, Y are independent.)

$$G(z_1, \ldots, z_n) = \mathrm{E}(z_1^{X_1} \cdots z_n^{X_n}) = \mathrm{E}(z_1^{X_1}) \cdots \mathrm{E}(z_n^{X_n})$$

i.e. joint pgf = product of marginal pgfs

$$M(s_1, \ldots, s_n) = \mathrm{E}(e^{s_1 X_1} \cdots e^{s_n X_n}) = \mathrm{E}(e^{s_1 X_1}) \cdots \mathrm{E}(e^{s_n X_n})$$

i.e. joint mgf = product of marginal mgfs.

The above property can be used to characterise independence because it can be shown that the factorisation of the joint mgf holds *if and only if* the variables are independent.

Marginal mgfs

It is straightforward to see that if $M_{X,Y}(s_1, s_2)$ is the joint mgf of X, Y then the marginal mgf of X is given by $M_X(s_1) = M_{X,Y}(s_1, 0)$.
(*Proof:* $\mathrm{E}(e^{s_1 X}) = \mathrm{E}(e^{s_1 X + 0.Y}) = M_{X,Y}(s_1, 0)$.)

Higher Moments

The joint moment generating function can further be useful to find higher moments of a distribution. More precisely, we can compute $\mathrm{E}(X_i^r X_j^k)$ in the following way.

1. Differentiate $M(s_1, \ldots, s_n)$ r times w.r.t. s_i;

2. further differentiate k times w.r.t. s_j;

3. then evaluate the resulting derivative for $s_1 = \cdots = s_n = 0$.

Following the above steps we get

$$\begin{aligned} \frac{\partial^r M}{\partial s_i^r} &= \mathrm{E}(X_i^r e^{\boldsymbol{s}^T \boldsymbol{X}}) \\ \frac{\partial^{r+k} M}{\partial s_i^r \partial s_j^k} &= \mathrm{E}(X_i^r X_j^k e^{\boldsymbol{s}^T \boldsymbol{X}}), \end{aligned}$$

which gives $\mathrm{E}(X_i^r X_j^k)$ on setting $\boldsymbol{s} = \boldsymbol{0}$.

Linear transformation property

This is the multivariate generalisation of the univariate transformation property. Suppose that $\boldsymbol{Y} = \boldsymbol{a} + b\boldsymbol{X}$. Then the mgf of $\boldsymbol{Y}$ is (*c.f.* §3.3)

$$M_{\boldsymbol{Y}}(\boldsymbol{s}) = \mathrm{E}(e^{\boldsymbol{s}^T\boldsymbol{Y}}) = \mathrm{E}\{e^{\boldsymbol{s}^T(\boldsymbol{a}+b\boldsymbol{X})}\} = e^{\boldsymbol{s}^T\boldsymbol{a}}\mathrm{E}(e^{b\boldsymbol{s}^T\boldsymbol{X}}) = e^{\boldsymbol{a}^T\boldsymbol{s}}M_{\boldsymbol{X}}(b\boldsymbol{s})$$

Example 3.6. *Suppose $X_1, \ldots, X_n$ are jointly multivariate normally distributed. Then, from equation (1.8) in §1.5.3, the density of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is given in matrix notation by*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mid 2\pi\boldsymbol{\Sigma} \mid^{1/2}} \; exp\,\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}\,,$$

where

$$\boldsymbol{x} = (x_1, \ldots, x_n)^T, \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T, \quad \boldsymbol{\Sigma} = \text{ covariance matrix}$$

i.e. *the $\mu_i = E(X_i)$ are the individual expectations and the $\sigma_{ij} = (\boldsymbol{\Sigma})_{ij} = Cov(X_i, X_j)$ are the pairwise covariances (variances if $i = j$).*

We obtain the joint mgf as follows. We have

$$E(e^{s_1X_1+\cdots+s_nX_n}) = \int \frac{1}{\mid 2\pi\boldsymbol{\Sigma} \mid^{1/2}} \; exp\,\{\boldsymbol{s}^T\boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}\mathrm{d}\boldsymbol{x}$$

(Remember that the integral here represents an n-dimensional integral.) To evaluate this integral we need to complete the square in $\{\cdot\}$. The result (derived in lectures) is that

$$\boxed{M(s_1, \ldots, s_n) = exp\,\left\{\boldsymbol{s}^T\boldsymbol{\mu} + \frac{1}{2}\,\boldsymbol{s}^T\boldsymbol{\Sigma}\boldsymbol{s}\right\}}$$

Exercise: from this derive the mgf of the univariate $N(\mu, \sigma^2)$ distribution.

For illustration, we now derive the joint moment $\mathrm{E}(X_iX_j)$. Differentiate first with respect to s_i. Since

$$\boldsymbol{s}^T\boldsymbol{\mu} = \sum_{k=1}^{n} s_k\mu_k, \quad \boldsymbol{s}^T\boldsymbol{\Sigma}\boldsymbol{s} = \sum_{k=1}^{n}\sum_{l=1}^{n} s_ks_l\sigma_{kl}$$

we see that $\partial(\boldsymbol{s}^T\boldsymbol{\mu})/\partial s_i = \mu_i$. Also the terms involving s_i in $\boldsymbol{s}^T\boldsymbol{\Sigma}\boldsymbol{s}$ are

$$s_i^2\sigma_{ii} + 2s_i\sum_{l\neq i} s_l\sigma_{il}$$

giving

$$\partial(\boldsymbol{s}^T\boldsymbol{\Sigma s})/\partial s_i = 2s_i\sigma_{ii} + 2\sum_{l\neq i} s_l\sigma_{il} = 2\sum_l s_l\sigma_{il}$$

Therefore

$$\mathrm{E}(X_i e^{\boldsymbol{s}^T\boldsymbol{X}}) = \partial M(\boldsymbol{s})/\partial s_i = \left(\mu_i + \sum_{l=1}^n \sigma_{il}s_l\right) \exp\{\boldsymbol{s}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Sigma s}\}$$

Now differentiate again with respect to s_j, $j \neq i$, to give

$$\mathrm{E}(X_iX_j e^{\boldsymbol{s}^T\boldsymbol{X}}) = \left\{\sigma_{ij} + \left(\mu_i + \sum_{l=1}^n \sigma_{il}s_l\right)\left(\mu_j + \sum_{l=1}^n \sigma_{jl}s_l\right)\right\} \exp\{\boldsymbol{s}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Sigma s}\}.$$

Setting $\boldsymbol{s} = \boldsymbol{0}$ now gives $\mathrm{E}(X_iX_j) = \sigma_{ij} + \mu_i\mu_j$ and therefore $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_iX_j - \mu_i\mu_j = \sigma_{ij}$.

Example 3.7. *Suppose that*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right).$$

Find $E(X_1X_2)$ (i) directly and (ii) from the joint mgf.

Remark: The multivariate normal density of $X_1, \ldots, X_n$ is only valid when Σ is a non-singular matrix, which can be shown to hold if and only if no exact linear relationships exist between $X_1, \ldots, X_n$. However, the multivariate normal distribution can still be defined when Σ is singular. Note in particular that the joint mgf is valid even when Σ is singular.

3.5 Linear combinations of random variables

We will now use the above methods to derive (properties of) the distribution of linear combinations of random variables.

Let $X_1, \ldots, X_n$ be the original variables. A **linear combination** is defined by

$$Y = a_1X_1 + \cdots + a_nX_n$$

for any real-valued constants $a_1, \dots, a_n$. A popular linear combination is for example the sample mean $Y = \overline{X}$, for which $a_i = 1/n$, $i = 1, \dots, n$.

Let us first find the expectation and variance of the linear combination Y in terms of the moments of the X_i. The methods from §1.4 can be used for this purpose. First

$$\mathrm{E}(Y) = a_1\,\mathrm{E}(X_1) + \cdots + a_n\,\mathrm{E}(X_n)$$

regardless of whether or not the X_i are independent. For the variance we have

$$\begin{aligned}\mathrm{Var}\,(Y) &= \mathrm{Cov}\,(\sum_i a_i X_i, \sum_j a_j X_j) \\ &= \sum_i \sum_j a_i a_j\ \mathrm{Cov}(X_i, X_j) \\ &= \sum_i a_i^2\ \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j\ \mathrm{Cov}(X_i, X_j).\end{aligned}$$

In particular, we see that if the X_i are *independent* then

$$\mathrm{Var}(\Sigma a_i X_i) = \Sigma_i\ \mathrm{Var}(a_i X_i) = \Sigma a_i^2\ \mathrm{Var}(X_i)$$

but *not* otherwise. In vector notation, if we write $Y = \boldsymbol{a}^T \boldsymbol{X}$, where $\boldsymbol{X} = (X_1, \dots, X_n)^T$, $\boldsymbol{a} = (a_1, \dots, a_n)^T$ then these relations becomes

$$\mathrm{E}(Y) = \boldsymbol{a}^T \boldsymbol{\mu}, \quad \mathrm{Var}(Y) = \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}\,,$$

where $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}), \Sigma = \mathrm{Cov}(\boldsymbol{X})$.

Now we want to find out about the actual distribution of Y. If we have the joint distribution of $X_1, \dots, X_n$ we could proceed by transformation similar to §2.3, but this could be very tedious if n is large. Instead, let us explore an approach based on the joint pgf / mgf of the X_i. (The result below for the mgf is actually just a special case of the earlier linear transformation property of a joint mgf.)

If Y is discrete we find for its pgf

$$G_Y(z) = \mathrm{E}(z^Y) = \mathrm{E}(z^{a_1 X_1 + \cdots + a_n X_n}) = \mathrm{E}(z^{a_1 X_1} z^{a_2 X_2} \cdots z^{a_n X_n})\,,$$

which is the joint pgf of $X_1, \dots, X_n$ evaluated at $z_i = z^{a_i}$, *i.e.* $G(z^{a_1}, \dots, z^{a_n})$.

Similarly, if Y is continuous its mgf is given as

$$M_Y(s) = \mathrm{E}(e^{sY}) = \mathrm{E}(e^{s(a_1 X_1 + \cdots + a_n X_n)}) = \mathrm{E}(e^{s a_1 X_1 + \cdots + s a_n X_n})\,,$$

which is the joint mgf of $X_1, \dots, X_n$ evaluated at $s_i = sa_i$, *i.e.* $M(sa_1, \dots, sa_n)$.

So, an alternative to the transformation method is to obtain the *joint* pgf or mgf and use this to derive the pgf or mgf of Y. Of course, we still have to get from there to the probability mass function or density if needed — but often the generating function of the univariate Y is known to belong to a specific distribution.

A simplification is available if $X_1, \dots, X_n$ are independent. In this case we have

$$G_Y(z) = \mathrm{E}(z^{a_1 X_1} z^{a_2 X_2} \cdots z^{a_n X_n}) = \prod_{i=1}^{n} \mathrm{E}(z^{a_i X_i}) = \prod_{i=1}^{n} G_{X_i}(z^{a_i}),$$

which is the product of the individual pgfs $G_{X_i}(z^{a_i})$ of the X_i evaluated at $z_i = z^{a_i}$.

Analogously, for the mgf we find

$$M_Y(s) = \mathrm{E}(e^{sa_1 X_1 + \cdots + sa_n X_n}) = \prod_{i=1}^{n} \mathrm{E}(e^{sa_i X_i}) = \prod_{i=1}^{n} M_{X_i}(sa_i),$$

which is the product of the individual mgfs $M_{X_i}(sa_i)$ of the X_i evaluated at $s_i = sa_i$.

Example 3.8 (3.2 ctd.)**.** *Consider independent random variables $X_1, \dots, X_n$ with $X_i \sim$Bin(m_i, p), i.e. they have different numbers of trials m_i but the same success probability p. Find the pgf and the distribution of $Y = \sum X_i$.*

Example 3.9 (3.3 ctd.)**.** *Let $X_1, \dots, X_n$ be independent with $X_i \sim$ Gam(α_i, λ). Find the mgf and the distribution of $Y = \sum X_i$.*

Example 3.10 (3.6 ctd.)**.** *Suppose $X_1, \dots, X_n$ are jointly multivariate normally distributed. Consider again the linear transformation $Y = a_1 X_1 + \cdots + a_n X_n$, or $Y = \boldsymbol{a}^T \boldsymbol{X}$ in vector notation, where $\boldsymbol{a} = (a_1, \dots, a_n)^T$, $\boldsymbol{X} = (X_1, \dots, X_n)^T$. It follows from the general result that the mgf of Y is*

$$M_Y(s) = M_{\boldsymbol{X}}(s\boldsymbol{a}) = E\{\exp(s\boldsymbol{a}^T \boldsymbol{X})\} = exp\left(s\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2} s^2 \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}\right)$$

from §3.4. By comparison with the univariate mgf (see Example 3.4) we see that

$$\boxed{Y = \boldsymbol{a}^T \boldsymbol{X} \sim N(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a})}$$

We have seen earlier that for *any* random vector $\boldsymbol{X}$ we have $\mathrm{E}(\boldsymbol{a}^T \boldsymbol{X}) = \boldsymbol{a}^T \boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{a}^T \boldsymbol{X}) = \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}$, so the importance of the foregoing result is that any linear combination of jointly normal variables is itself *normally* distributed even if the variables are correlated (and thus not independent).

Example 3.11 (3.7 ctd.). *Use the joint mgf of X_1 and X_2 to find the distribution of $Y = X_1 - X_2$.*

3.5.1 The Central Limit Theorem

Possibly the most important theorem in statistics is the Central Limit Theorem which enables us to approximate the distribution of sums of independent random variables. One of the most popular uses is to make statements about the sample mean $\bar{X}$ which, after all, is nothing but a scaled sum of the samples.

Central Limit Theorem:

Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables each with mean μ and variance $\sigma^2 < \infty$. Then

$$\boxed{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z}$$

as $n \to \infty$ where $Z \sim N(0, 1)$.

Here, $\xrightarrow{d}$ denotes convergence in distribution which means that as $n \to \infty$ the cdf of the random variable on the left converges to the cdf of the random variable on the right at each point. Writing this down more carefully, let $Y_n = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)$. Then, the statement simply means $\lim_{n\to\infty} F_{Y_n}(x) = \Phi(x)$ holds for any $x \in \mathbb{R}$, where Φ is the standard normal cdf.

We can use the central limit theorem to compute probabilities about $\overline{X}_n$ from

$$\mathrm{P}(a < \overline{X}_n \leq b) \approx \Phi\left(\frac{\sqrt{n}(b - \mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a - \mu)}{\sigma}\right).$$

NB. There are many generalisations of the theorem where the assumptions of independence, common distribution or finite variance of the X_i are relaxed.

The Central Limit Theorem is Amazing!

From next to no assumptions (independence and identical distribution with finite variance) we arrive at a phenomenally strong conclusion: The sample mean asymptotically follows a Gaussian distribution. Regardless of which particular distribution the X_n follow (exponential, Bernoulli, binomial, uniform, triangle, ...), the resulting distribution is always Gaussian! It is as though the whole of statistics and probability theory reduces to one distribution only!!

Example 3.12. *Let $X_1, \ldots, X_n$ be an i.i.d. sample of exponential variables, i.e. $X_i \sim$ Exp(λ). Find formulae to approximate the probabilities $P(\bar{X}_n \leq x)$ and $P(\sum X_i \leq x)$.*

Example 3.13. *Consider a sequence of independent Bernoulli trials with constant probability of success π, so that $P(X_i = 1) = \pi$, $i = 1, 2, \ldots$. Use the Central Limit Theorem with $\mu = \pi$ and $\sigma^2 = \pi(1 - \pi)$ to derive the normal approximation to the binomial distribution.*

How large does n need to be for the CLT to work?

One of the disadvantages of the Central Limit Theorem is that in practical situations, one can never be quite sure just how large n needs to be for the approximation to work well. As an illustration, consider the sum of independent uniformly distibuted random variables. Approximate probability density functions (obtained through simulation and histograms - if you want to know more about how this works, take STAT7001) are shown in Figures 3.1 and 3.2 below. In the case of $U(0,1)$ variables, it is clear that for $n \geq 5$, the difference in the pdf is nearly invisible. Any rule of thumb as to how large n needs to be suffers from mathematicians' curse: since the actual distribution of the X_i is assumed unknown, one can always find a distribution for which convergence has not yet occurred for the particular value of n considered!

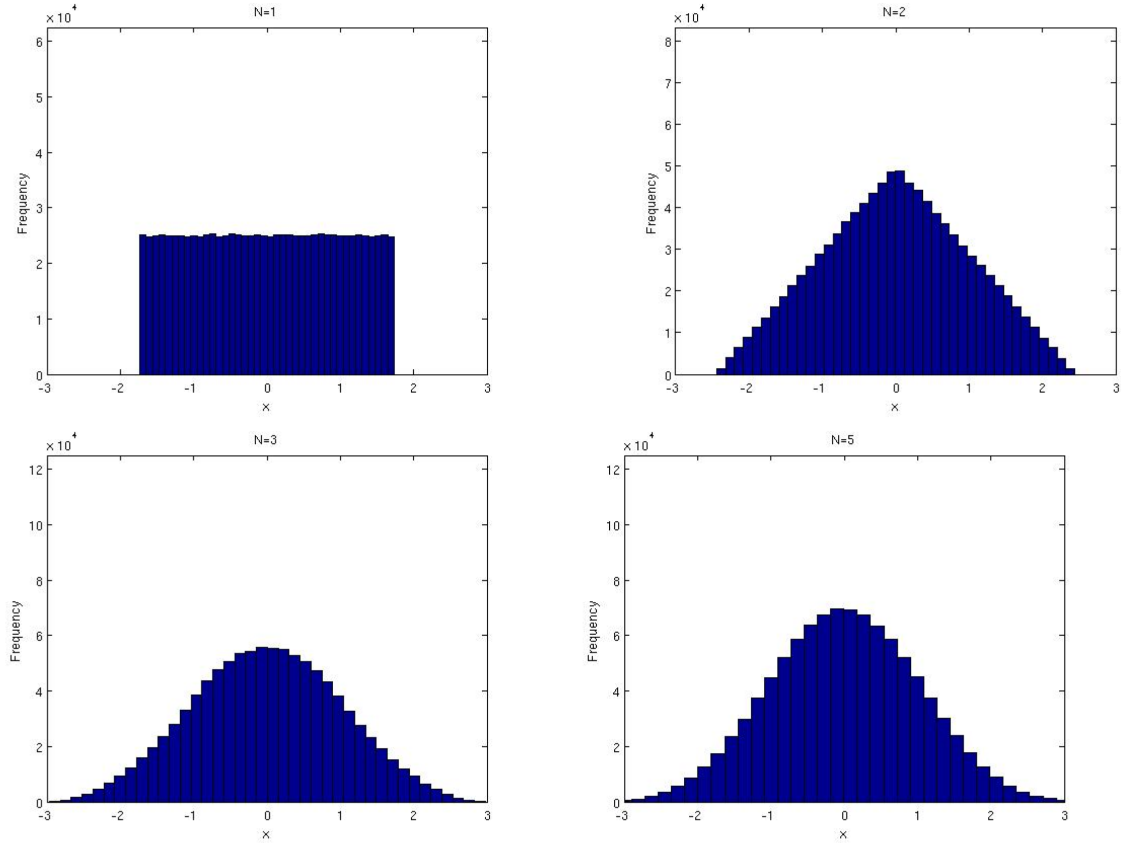

Figure 3.1: Estimated probability density functions standardised to mean zero and variance one. Top left: X_1, top right: $X_1 + X_2$, bottom left: $X_1 + X_2 + X_3$, bottom right: $\sum_{i=1}^{5} X_i$

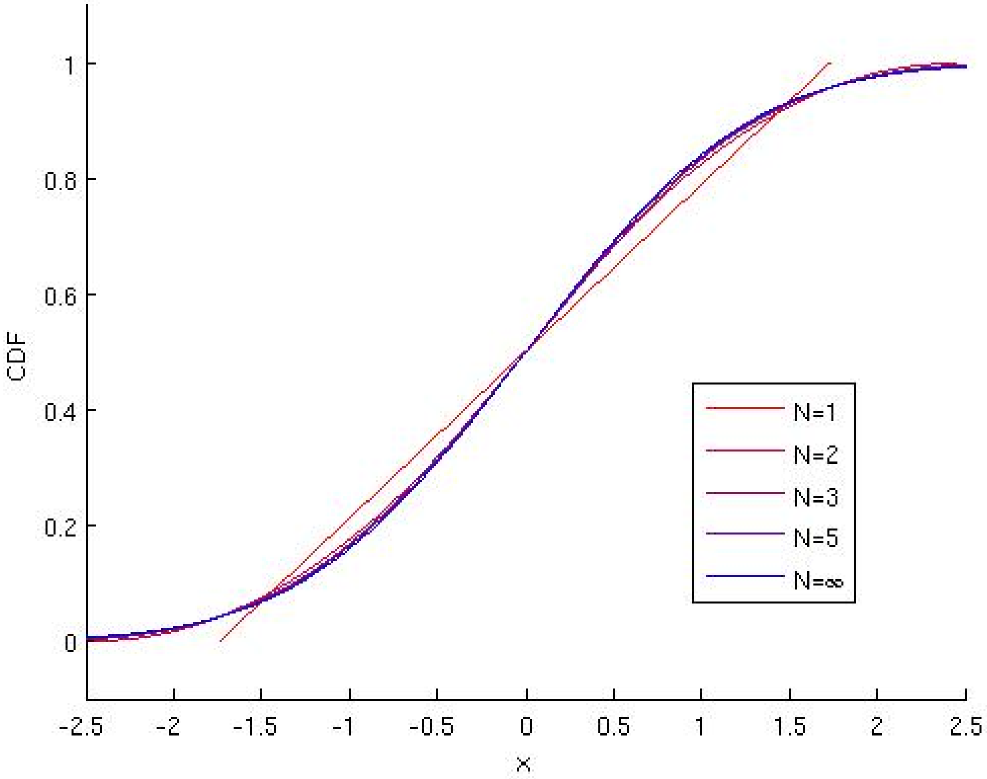

Figure 3.2: Estimated cumulative density functions standardised to mean zero. Note that for $N \geq 5$, the cdfs are essentially indistinguishable from standard normal.

Example 3.14 (Sums of i.i.d. uniformly distributed random variables)**.** *Let* $X_i \sim U(0, 1)$ *be independent random variables. Compute the approximate distribution of* $Y = \sum_{n=1}^{100} X_n$ *and the probability that* $Y < 90$*.*

Finally, as increasingly faster computers become available at steadily decreasing cost, the Central Limit Theorem loses some of its practical importance. It is still regularly being used, however, as an easy first guess.

STAT3101 only:
We prove the theorem only in the case when the X_i have a moment generating function defined on an open interval containing zero. Let us first suppose that $\mathrm{E}(X) = \mu = 0$. We then know that $\mathrm{E}(\overline{X}_n) = 0$ and $\mathrm{var}(\overline{X}_n) = \sigma^2/n$.
Consider the **standardised variable** $Z_n = \sqrt{n} \cdot \overline{X}_n/\sigma$. Then $\mathrm{E}(Z_n) = 0$ and $\mathrm{var}(Z_n) = 1$ for all n. Note that Z_n is a linear combination of the X_i, since

$$Z_n = \frac{\sqrt{n}\sum_{i=1}^n X_i}{\sigma n} = \sum_{i=1}^n \frac{X_i}{\sigma\sqrt{n}} \,.$$

Thus, by the arguments from the previous section, taking $a_i = 1/(\sigma\sqrt{n})$, the mgf of Z_n is given by

$$M_{Z_n}(s) = \prod_{i=1}^n M_{X_i}\left(\frac{s}{\sigma\sqrt{n}}\right) = \left\{M_X\left(\frac{s}{\sigma\sqrt{n}}\right)\right\}^n$$

since the X_i all have the same distribution and thus the same mgf M_X.
As the distribution of the X_i is not given we don't know the mgf M_X. However, M_X has the following Taylor series expansion about 0:

$$M_X(t) = M_X(0) + tM_X'(0) + \frac{t^2}{2}M_X''(0) + \varepsilon_t \,, \tag{3.1}$$

where ε_t is a term for which we know that $\varepsilon_t/t^2 \to 0$ for $t \to 0$. We write this as $\varepsilon_t = o(t^2)$ ('small order' of t^2, meaning that ε_t tends to zero at a rate faster than t^2).
To make use of the above Taylor series expansion, note that (recalling that $\mu = 0$)

$$\begin{aligned} M_X(0) &= \mathrm{E}(e^{0X}) = 1 \\ M_X'(0) &= E(X_i) = 0 \\ M_X''(0) &= E(X_i^2) = \sigma^2. \end{aligned}$$

Inserting these values in (3.1) and replacing t by $s/(\sigma\sqrt{n})$ we find that

$$M_X\left(\frac{s}{\sigma\sqrt{n}}\right) = 1 + 0 + \frac{1}{2}\sigma^2\left(\frac{s}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right) = 1 + \frac{s^2}{2n} + o\left(\frac{1}{n}\right) \,.$$

Now back to Z_n: from earlier, the mgf M_Z of Z_n is now given as the nth power of the above and we find that

$$\begin{aligned} M_Z(s) = \left\{M_X\left(\frac{s}{\sigma\sqrt{n}}\right)\right\}^n &= \left\{1 + \frac{s^2}{2n} + o\left(\frac{1}{n}\right)\right\}^n \\ &= \left(1 + \frac{\frac{1}{2}s^2 + \delta_n}{n}\right)^n \longrightarrow e^{\frac{1}{2}s^2} \end{aligned}$$

as $n \to \infty$, since $\delta_n \to 0$. (Recall: $(1 + \frac{x}{n})^n \to e^x$ as $n \to \infty$.) The limiting mgf is the one that we know as belonging to the standard normal distribution.

It can be shown that convergence of the moment generating functions implies convergence of the corresponding distribution functions at all points of continuity. This proves the claim in the case $\mu = 0$.
In the case that $\mathrm{E}(X_i) = \mu \neq 0$ define $Y_i = X_i - \mu$. Then $\mathrm{E}(Y_i) = 0$ so the result already proved gives $Z_n = \sqrt{n}(\bar{X} - \mu)/\sigma = \sqrt{n}\bar{Y}/\sigma \xrightarrow{d} N(0,1)$ as required. □

Learning Outcomes:

Generating Functions (pgf / mgf) *You should be able to*

1. *Reproduce the definitions of a pgf / mgf and derive their main properties;*
2. *Compute and recognise the pgf / mgf for standard situations;*
3. *Use the pgf to find the pmf of a discrete random variable as well as its expectation and variance for standard cases;*
4. *Use the mgf to find the moments of random variables, in particular the mean and variance, for standard cases.*

Joint Generating Functions *You should be able to*

1. *Reproduce the definition of joint probability / moment generating functions;*
2. *Derive the joint pgf / mgf for a collection of independent random variables;*
3. *Derive the pgf / mgf of linear combinations of a collection of (not necessarily independent) random variables and recognise the resulting distribution (especially for the multivariate normal case);*
4. *Characterise independence between variables with the help of their (joint) generating functions;*
5. *Apply the appropriate formula in order to find higer-order expectations from the joint mgf, especially for the multivariate normal case;*
6. *State the central limit theorem;*
7. *Use the central limit theorem creatively to derive approximate probability statements.*

Chapter 4

Distributions of Functions of Normally Distributed Variables

4.1 Motivation

The Central Limit Theorem (CLT) implies that the normal distribution arises in, or is a good approximation to, many practical situations. Also, the case of a random sample $X_1, \ldots, X_n$ from a normal population $N(\mu, \sigma^2)$ is favourable because many expressions are available in explicit form which is probably at least as important a reason for the widespread use of the normal approximation as the CLT. The random sample from a normal population was treated in STAT1005 and the estimators $\overline{X}$ and S^2 for the mean and variance were exhibited as well as some of their properties shown. In this chapter, we will develop a more careful discussion of the properties and go beyond STAT1005 by establishing results on the joint distribution of these two estimators. Finally, the t- and F-distributions are derived and discussed in order to have them ready for applications in statistics (such as estimation and hypothesis testing).

4.2 Reminder: Random Sample from a Normal Population

From the mathematical point of view, a random sample of size $n \in \mathbb{N}$ is a collection of independent, identically distributed (iid) random variables $X_1, X_2, \ldots, X_n$. In the case of a normal sample, $X_i \sim N(\mu, \sigma^2)$, the sample mean

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

is known (from STAT1005 or equivalent) to be unbiased, i.e. $\mathbb{E}[\overline{X}] = \mu$ and to have variance $\mathrm{Var}(\overline{X}) = \sigma^2/n$. To estimate the variance, σ^2, in STAT1005 you used the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

which you showed to be unbiased and consistent, i.e. you showed that the variance of S^2 goes to zero as the sample gets larger (i.e. as $n \to \infty$) which implies consistency. To obtain the sampling distribution of this estimator, i.e. the distribution of S^2 thought of as a random variable itself, it is thus necessary to think about the distributions of sums of iid random variables as well as the distribution of $(X_i - \overline{X})^2$ and its sums, i.e. the sums of squares of normal random variables. These distributions will occupy us for the remainder of this chapter whereas the next chapter will consider estimators (and present more detail on unbiasedness, variance, sampling distribution etc.) as well as address the question whether there could be any better estimators than sample mean and sample variance.

4.3 The chi-squared (χ^2) distribution

Preliminaries

Recall the pdf of the $\mathrm{Gam}(\alpha, \lambda)$ distribution (see Appendix 2), its mean and variance $\mathrm{E}(X) = \alpha/\lambda$, $\mathrm{Var}(X) = \alpha/\lambda^2$, its mgf $\{\lambda/(\lambda - s)\}^\alpha$ (see example 3.3) and the additive property: if $X_1, \ldots, X_n$ are independent $\mathrm{Gam}(\alpha, \lambda)$ random variables then $X_1 + \cdots + X_n$ is $\mathrm{Gam}(n\alpha, \lambda)$.

Distribution of Sums of Squares of Normals

We start by showing that if X has the standard normal distribution then X^2 has the gamma distribution with index 1/2 and scale parameter 1/2.

The moment generating function of X^2 is given by

$$\begin{aligned}\mathbb{E}(e^{sX^2}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2 + sx^2\right\} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2(1-2s)\right\} dx \qquad (4.1) \\ &= (1-2s)^{-1/2} \qquad (4.2)\end{aligned}$$

where (4.2) follows by comparison of (4.1) with the integral of the density of a normal variable with mean 0 and variance $(1-2s)^{-1}$. (Alternatively, substitute $z = x\sqrt{1-2s}$ in the integral (4.1).)

By comparison with the gamma mgf, we see that X^2 has a gamma distribution with index and scale parameter each having the value 1/2. This distribution is also known as the **χ^2 distribution with one degree of freedom**. (Thus $\chi^2_1 \equiv$ Gam$(\frac{1}{2}, \frac{1}{2})$.)

Example 4.1. *Verify the result that $X^2 \sim$ Gam$(1/2, 1/2)$ by using the transformation $\phi(x) = x^2$ on a mean zero normal random variable $X \sim N(0, 1)$.*

Now let $X_1, \ldots, X_\nu$ be independent standard normal variables, where ν is a positive integer. Then it follows from the additive property of the gamma distribution that their *sum of squares* $X_1^2 + \ldots + X_\nu^2$ has the gamma distribution Gam$(\frac{\nu}{2}, \frac{1}{2})$ with index $\nu/2$ and scale parameter 1/2. This distribution is also known as the **χ^2 distribution with ν degrees of freedom** and is written as χ^2_ν. Its mgf is therefore $\{\frac{1}{2}/(\frac{1}{2} - s)\}^{\nu/2} = (1-2s)^{-\nu/2}$. (Thus $\chi^2_\nu \equiv$ Gam$(\frac{\nu}{2}, \frac{1}{2})$.)

It further follows from the mean and variance of the gamma distribution that the mean and variance of $U \sim \chi^2_\nu$ are (*verify*)

$$\boxed{\mathrm{E}(U) = \nu\,, \ \ \mathrm{Var}(U) = 2\nu}$$

Example 4.2. *Verify the above expectation and variance of χ^2_ν by considering $U = \sum_{i=1}^{\nu} X_i^2$ with independent $X_i \sim N(0, 1)$.*

The pdf of the χ^2 distribution with ν degrees of freedom ($\nu > 0$) is

$$\boxed{f(u) = \frac{u^{\frac{1}{2}\nu - 1} e^{-\frac{1}{2}u}}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}}$$

for $u > 0$. This can be verified by comparison with the $\mathrm{Gam}(\frac{\nu}{2}, \frac{1}{2})$ density (see Appendix 2).

Question: Could this also be derived using transformation of variables and the formula $f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x)f(z-x)\mathrm{d}x$ for the density of a sum of random variables?

Application to sampling distributions

We know that if X has a normal distribution with mean μ and variance σ^2 then the standardised variable $(X - \mu)/\sigma$ has the standard normal distribution. It follows that if $X_1, \ldots, X_n$ are independent normal variables, all with mean μ and variance σ^2, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$$

because the left-hand side is the sum of squares of n independent standard normal random variables. Since $\sigma^2 = \mathrm{E}(X - \mu)^2$, when μ is known the sample average $S^2_\mu = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an intuitively natural estimator of σ^2. The above result gives the sampling distribution of S^2_μ as $nS^2_\mu \sim \sigma^2 \chi^2_n$. We can now deduce the mean and variance of S^2_μ from those of the χ^2_n distribution:

$$\begin{aligned} \mathrm{E}(S^2_\mu) &= \frac{\sigma^2}{n} \mathrm{E}\left(\frac{nS^2_\mu}{\sigma^2} \right) = \frac{\sigma^2}{n} \times n = \sigma^2 \\ \mathrm{Var}(S^2_\mu) &= \frac{\sigma^4}{n^2} \mathrm{Var}\left(\frac{nS^2_\mu}{\sigma^2} \right) = \frac{\sigma^4}{n^2} \times 2n = \frac{2\sigma^4}{n} \end{aligned}$$

Note that the expectation formula is generally true, even if the X_i are not normal. However, the variance formula is only true for normal distributions.

When μ is unknown, it is natural to estimate it by $\overline{X}$. We already know that $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}$. Since a linear combination of normal variables is again normally distributed, we deduce that

$$\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n} \right)}$$

However, in order to be able to use this result for statistical inference when σ^2 is also unknown, we need to estimate σ^2. An intuitively natural estimator of σ^2 when μ is unknown is the *sample variance* $S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$. The reason for the factor $\frac{1}{n-1}$ rather than $\frac{1}{n}$ is that S^2 is unbiased for σ^2, as we will see in the next chapter. In order to deduce the sampling distribution of S^2 we need to find the distribution of the sum of squares $\sum (X_i - \overline{X})^2$.

4.3.1 The distribution of $\sum(X_i - \overline{X})^2$

Let $X_1, \ldots, X_n$ be independent normally distributed variables with common mean μ and variance σ^2. We will show that $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom. In the following it will become clear why, exactly, we have to reduce the number of degrees of freedom by one.

Result 4.1:
$\overline{X}$ and $S^2 = \sum(X_i - \overline{X})^2/(n-1)$ are independent.

Result 4.2:
The sampling distribution of S^2 is $\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}$.

The results can be established by considering joint moment generating functions.

Derivation of Result 4.1

Let $U_i = X_i - \overline{X}$. If we can show that $\overline{X}$ and the U_i, $i = 1, \ldots, n$, are independent then we are done because S^2 is a function of the U_i only. We show this independence by deriving the joint mgf of the $\overline{X}$ and the U_i, $i = 1, \ldots, n$, and establishing that it factorises.

To find the joint mgf, we relate a linear combination of $\overline{X}$ and the U_i to a linear combination of the original X_i as follows:

$$s_0 \overline{X} + \sum_{i=1}^n s_i U_i = \sum_{i=1}^n a_i X_i \, , \tag{4.3}$$

where $a_i = \frac{s_0}{n} + s_i - \bar{s}$ and $\bar{s} = \sum_{i=1}^n s_i/n$ (*exercise: derive this*). This linear combination is chosen to exploit the fact that, while we wish to find out about the distribution of the left hand side of (4.3) via its mgf, we have a good understanding of the distribution and mgf of the X_i on the right hand side of (4.3) – it is multivariate normal.

The joint mgf of $\overline{X}$ and the U_i is now given by

$$M_{\overline{X}, U_1, \ldots, U_n}(s_0, s_1, \ldots, s_n) = \mathrm{E}(e^{s_0 \overline{X} + \sum s_i U_i}) = \mathrm{E}(e^{\sum a_i X_i}) \, ,$$

which is the joint mgf of $X_1, \ldots, X_n$ evaluated at the above choice of the a_i. As $X_1, \ldots, X_n$ is an independent normal sample we know that this joint mgf (*c.f.* Example

3.6) is given by

$$\mathrm{E}(e^{\sum a_i X_i}) = \Pi_i M_{X_i}(a_i) = \exp\left\{(\sum a_i)\mu + \frac{1}{2}(\sum a_i^2)\sigma^2\right\}$$

But $\sum_{i=1}^n a_i = s_0$ and $\sum_{i=1}^n a_i^2 = \frac{s_0^2}{n} + \sum_{i=1}^n (s_i - \overline{s})^2$, giving

$$M_{\overline{X},U_1,\ldots,U_n}(s_0, s_1, \ldots, s_n) = \exp\left\{s_0\mu + \frac{1}{2n}s_0^2\sigma^2 + \frac{1}{2}\sigma^2 \sum_{i=1}^n (s_i - \overline{s})^2\right\}$$

Therefore the joint mgf $M_{\overline{X},U_1,\ldots,U_n}$ *factorises* into

$$\exp\left\{s_0\mu + \frac{1}{2n}s_0^2\sigma^2\right\} \quad \text{and} \quad \exp\left\{\frac{1}{2}\sigma^2 \sum_{i=1}^n (s_i - \overline{s})^2\right\}.$$

This factorisation implies the following:

1. $\overline{X}$ and $U_1, \ldots, U_n$ are independent.
2. Since S^2 is a function of the U_i only, it follows that $\overline{X}$ and S^2 are independent.
3. The marginal mgf of $\overline{X}$ is $\exp(s_0\mu + \frac{1}{2}s_0^2\sigma^2/n) \Longrightarrow \overline{X} \sim N(\mu, \sigma^2/n)$ (as we already know).

Example 4.3. *Show that $\overline{X}$ and U_i are independent by computing their covariance.*

Derivation of Result 4.2

In order to find the distribution of S^2 note that

$$\sum (X_i - \mu)^2 = \sum (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

so that

$$\frac{\sum (X_i - \mu)^2}{\sigma^2} = \frac{\sum (X_i - \overline{X})^2}{\sigma^2} + \frac{(\overline{X} - \mu)^2}{\sigma^2/n}.$$

But, since $(X_i - \mu)/\sigma$ are independent $N(0,1)$ and $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ is $N(0,1)$, we have

$$\frac{\sum (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2 \quad \text{and} \quad \frac{(\overline{X} - \mu)^2}{\sigma^2/n} \sim \chi_1^2$$

Using mgfs, it follows that

$$\frac{\sum (X_i - \overline{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

since $(\overline{X}-\mu)^2/(\sigma^2/n)$ and $\sum(X_i-\overline{X})^2/\sigma^2$ are independent.

This result gives the sampling distribution of S^2 to be $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$.

From Results 4.1 and 4.2 we can deduce the sampling mean and variance of S^2 to be

$$\begin{aligned} \mathrm{E}(S^2) &= \frac{\sigma^2}{n-1}\mathrm{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^2}{n-1} \times n-1 = \sigma^2 \\ \mathrm{Var}(S^2) &= \frac{\sigma^4}{(n-1)^2}\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^4}{(n-1)^2} \times 2(n-1) = \frac{2\sigma^4}{n-1} \end{aligned}$$

recalling that the mean and variance of χ^2_{n-1} are $n-1$, $2(n-1)$ respectively. As for the case where μ is known, the *unbiasedness* property of S^2 is generally true, even if the X_i are not normal, but the variance formula is only true for normal distributions.

4.3.2 Student's t distribution

If $X_1,\ldots,X_n$ are independent normally distributed variables with common mean μ and variance σ^2 then $\overline{X}$ is normally distributed with mean μ and variance σ^2/n. Recall (STAT1004 & STAT1005, or MATH7502) that if σ^2 is known, then we may test the hypothesis $\mu=\mu_0$ by examining the statistic

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\,.$$

Z is a linear transformation of a normal variable and hence is also normally distributed. When $\mu=\mu_0$ we see that $Z \sim N(0,1)$ so we conduct the test by computing Z and referring to the $N(0,1)$ distribution.

Now if σ^2 is unknown, it is intuitively reasonable to estimate it by S^2 (recalling that $\mathrm{E}(S^2)=\sigma^2$) and use the statistic

$$T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}\,.$$

However, in order to conduct the test we need to know the distribution of this statistic when $\mu=\mu_0$.

Note that we can write T as

$$T = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\ \frac{\sigma}{S} = \frac{Z}{\sqrt{U/(n-1)}}\,,$$

where $U = \sum_i (X_i - \overline{X})^2/\sigma^2$. From the above results we have $Z \sim N(0,1)$, $U \sim \chi^2_{n-1}$ and Z, U are independent random variables. (Note that the distribution of T does not depend on μ_0 and σ^2, but only on the known number n of observations and is therefore suitable as a test statistic.)

We can now find the probability distribution of T by the transformation method that was described in §2.2. Alternatively, it can be derived from the F distribution. The distribution of T, denoted by t_{n-1}, is known as **Student's *t* distribution with *n*-1 degrees of freedom**.

The general description of the t distribution is as follows. Suppose that Z has a standard normal distribution, U has a χ^2 distribution with ν degrees of freedom, and Z, U are independent random variables. Then

$$\boxed{T = \frac{Z}{\sqrt{U/\nu}}}$$

has the Student's t distribution with ν degrees of freedom, denoted by t_ν. T has probability density function

$$\boxed{f_T(t) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}}$$

for $-\infty < t < \infty$. The distribution of T is symmetrical about 0 , so that $\mathrm{E}(T) = 0$ $(\nu > 1)$. It can further be shown that the variance of T is $\nu/(\nu-2)$ for $\nu > 2$.

Example 4.4. *Write down the pdf of a t_1-distribution and identify the distribution by name (look back at Example 2.2). Why is $\nu > 1$ needed for the expected value to be zero?*

STAT3101 only: *What happens as the sample size ν becomes large? Derive the limit of the pdf in this case.*

4.4 The F distribution

Now suppose that we have two *independent* samples of observations: $X_1, \ldots, X_m$ are independent normally distributed variables with common mean μ_X and variance σ^2_X, while $Y_1, \ldots, Y_n$ are independent and normally distributed with common mean μ_Y and variance σ^2_Y. Suppose that we wish to test the hypothesis that the variances σ^2_X and σ^2_Y are equal. If $\sigma^2_X = \sigma^2_Y$ then $\sigma^2_X/\sigma^2_Y = 1$ and it is natural to examine the ratio of the two sample variances, S_X^2/S_Y^2 and compare its value with 1. Here $S_X^2 = \sum_{i=1}^m (X_i - \overline{X})^2/(m-1)$, $S_Y^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2/(n-1)$.

Now, since $(m-1)S_X^2/\sigma_X^2 \sim \chi^2_{m-1}$, and $(n-1)S_Y^2/\sigma_Y^2 \sim \chi^2_{n-1}$, we can write

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{U/(m-1)}{V/(n-1)}\,,$$

where $U = \sum_i (X_i - \overline{X})^2/\sigma_X^2 \sim \chi^2_{m-1}$, $V = \sum_i (Y_i - \overline{Y})^2/\sigma_Y^2 \sim \chi^2_{n-1}$ and U and V are independent variables (since they are based on two independent samples). When $\sigma_X^2 = \sigma_Y^2$, the left hand side is just the ratio S_X^2/S_Y^2.

The above considerations motivate the following general description of the F distribution. Suppose that $U \sim \chi^2_\alpha$, $V \sim \chi^2_\beta$ and U, V are independent. Then

$$\boxed{W = \frac{U/\alpha}{V/\beta}}$$

has the ***F* distribution with $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ degrees of freedom**, denoted $F_{\alpha,\beta}$.

Using the *Beta function* $B(a,b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, the probability density of the F distribution with (α, β) degrees of freedom can be written down as

$$f_W(w) = \frac{\frac{\alpha}{\beta}\left(\frac{\alpha w}{\beta}\right)^{\frac{\alpha}{2}-1}}{B(\frac{\alpha}{2}, \frac{\beta}{2})\left(1 + \frac{\alpha w}{\beta}\right)^{\frac{\alpha+\beta}{2}}}$$

for $w \geq 0$.

Note that, from the definition of the F distribution, we have

$$\frac{1}{W} = \frac{V/\beta}{U/\alpha} \sim F_{\beta,\alpha}$$

i.e. if $W \sim F_{\alpha,\beta}$ then $1/W \sim F_{\beta,\alpha}$.

It can be shown that, for all α and for $\beta > 2$, $\mathrm{E}(W) = \beta/(\beta - 2)$.

We can now apply this result to our test statistic discussed earlier. Under the hypothesis that $\sigma_X^2 = \sigma_Y^2$ we have (taking $\alpha = m-1$, $\beta = n-1$)

$$\boxed{\frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1}}$$

This is the sampling distribution of the variance ratio. Note that this distribution is free from σ_X and σ_Y.

Example 4.5. *Suppose that X, Y, U are independent random variables such that $X \sim N(2, 9)$, $Y \sim t_4$ and $U \sim \chi^2_3$. Give four functions of the above variables that have the following distributions:*
(i) χ^2_1, (ii) χ^2_4, (iii) t_3, (iv) $F_{1,4}$. □

Learning Outcomes: *Chapter 5 derives some fundamental results for statistical theory using the methods presented in Chapters 1 – 3. It should therefore consolidate your usage of and familiarity with those methods. In particular, you should be able to*

1. *Define the chi–squared distribution, t distribution and F distribution in terms of transformations of normal variables and know how they relate to each other,*
2. *State the main properties of the above distributions (mean, variance, shape, meaning of parameters),*
3. *State the relationship between the chi–squared and gamma distributions,*
4. *Remember the relationship between $\overline{X}$ and S^2 and sketch how it is derived,*
5. *Show how the chi–squared distribution can be derived using the properties of mgfs;*

Chapter 5

Statistical Estimation

5.1 Overview

We will now address the problem of estimating an unknown population parameter based on a sample $X_1, \ldots, X_n$. In particular, we focus on how to decide whether a potential estimator is a 'good' one or if we can find a 'better' one. To this end, we need to define criteria for good estimation. Here we present two obvious criteria that state that a good estimator should be 'close' to the true unknown parameter value (accuracy) and should have as little variation as possible (precision). We then go on to describe some methods that typically yield good estimators in the above sense. Results from the previous sections may be used to derive properties of estimators, which can be viewed as transformations of the original sample variables $X_1, \ldots, X_n$.

5.2 Criteria for good estimators

Suppose $X_1, \ldots, X_n$ represent observations on some sample space. Then a single or vector-valued function $T_n(X_1, \ldots, X_n)$ that does not depend on any unknown parameters is called a **statistic**. If its value is to be used as an estimate of an unknown parameter θ in a model for the observations, it is called an **estimator** of θ.

Since T_n is a random variable it has a probability distribution, called the **sampling distribution** of the estimator. The properties of this distribution determine whether or not T_n is a 'good' estimator of θ.

The difference $\mathrm{E}(T_n) - \theta = b_{T_n}(\theta)$ is called the **bias** of the estimator T_n. If $b_{T_n}(\theta) \equiv 0$ (*i.e.* $b_{T_n}(\theta) = 0$ for *all* values of θ) then T_n is an **unbiased estimator** of θ. Although we tend to regard unbiasedness as desirable, there may be a biased estimator T_n' giving values that tend to be closer to θ than those of the unbiased estimator T_n.

If we want an estimator that gives estimates close to θ we might look for one with small **mean square error (mse)**, defined as

$$\boxed{\mathrm{mse}(T_n;\theta) = \mathrm{E}(T_n - \theta)^2.}$$

Note that

$$\begin{aligned}
\mathrm{mse}(T_n,\theta) &= \mathrm{E}\{T_n - \mathrm{E}(T_n) + \mathrm{E}(T_n) - \theta\}^2 \\
&= \mathrm{E}\{T_n - \mathrm{E}(T_n)\}^2 + \{\mathrm{E}(T_n) - \theta\}^2 + 2\{\mathrm{E}(T_n) - \theta\}\mathrm{E}\{T_n - \mathrm{E}(T_n)\} \\
&= \mathrm{Var}(T_n) + \{b_{T_n}(\theta)\}^2
\end{aligned}$$

since $\mathrm{E}\{T_n - \mathrm{E}(T_n)\} = 0$. We therefore have the relation

$$\boxed{\mathrm{mse}(T_n,\theta) = \mathrm{Var}(T_n) + \{b_{T_n}(\theta)\}^2}$$

We see that small mean square error provides a trade-off between small variance and small bias.

Example 5.1. *Let $X_1, \ldots, X_n$ be an iid sample from a normal distribution with mean μ and variance σ^2. Then S^2 is unbiased for σ^2, whereas $\tilde{S}^2 = \frac{1}{n}\sum(X_i - \overline{X})^2$ is biased for σ^2. Compare the mean square errors of S^2 and $\tilde{S}^2$.*

If we restrict ourselves to unbiased estimators then the mse criterion is equivalent to a *variance* criterion and we search for *minimum variance unbiased estimators* (mvue).

5.2.1 Terminology:

The standard deviation of an estimator is also called its **standard error**. When the standard deviation is estimated by replacing the unknown θ by its estimate, it is the *estimated standard error* – but often just referred to as the 'standard error'.

Example 5.2 (Example 5.1 ctd.)**.** *S^2 has standard deviation $\sqrt{2\sigma^4/(n-1)}$. So S^2 is an estimator of σ^2 with standard error $s^2\sqrt{2/(n-1)}$.*

We now give the main result of this section.

5.3 Cramér-Rao lower bound:

5.3.1 Overview

Amongst unbiased estimators we prefer those with small variances. Suppose we have an unbiased estimator and have determined its variance. Then it would be useful to know how much better we could do in terms of variance. We will show that there is a lower (positive) bound for the variance of any unbiased estimator. Therefore if the variance of our existing estimator is close to this bound then there is little to be gained by searching for an alternative unbiased estimator. On the other hand, the bound is not necessarily attainable so if the variance of our estimator is far from the bound, the search for a better estimator could still be fruitless. (The third year statistical inference course will explore this further).

5.3.2 What is the Cramér-Rao lower bound?

Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a sample with joint density $f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$, where θ is a real–valued parameter. Let $T_n(\boldsymbol{X})$ be an unbiased estimator of θ. The *Cramér-Rao lower bound* for the variance of $T_n(\boldsymbol{X})$ is

$$\boxed{\mathrm{Var}(T_n(\boldsymbol{X})) \geq \frac{1}{I(\theta)} = \text{ Cramér-Rao bound}}$$

where

$$\boxed{I(\theta) = \mathrm{E}\left\{-\frac{\partial^2}{\partial\theta^2} \log f_{\boldsymbol{X}}(\boldsymbol{X}, \theta)\right\}}$$

is called the **Fisher information** in the sample $\boldsymbol{X}$. In other words, the Cramér-Rao lower bound is equal to the inverse Fisher information. The proof of this inequality is deferred until §5.3.3.

Note that if $X_1, \dots, X_n$ are i.i.d. with density $f(x, \theta)$ then $f_{\boldsymbol{X}}(\boldsymbol{X}; \theta) = \prod_i f(X_i; \theta)$. Therefore

$$\frac{\partial^2}{\partial\theta^2} \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta) = \sum_i \frac{\partial^2}{\partial\theta^2} \log f_X(X_i; \theta)\,,$$

which is a sum of n i.i.d. random variables. It follows that

$$\boxed{I(\theta) = n\,\mathrm{E}\left\{-\frac{\partial^2}{\partial\theta^2} \log f_X(X; \theta)\right\} = ni(\theta)}$$

where $i(\theta)$ is Fisher's information in the observation X. This illustrates the additive property of Fisher information; information increases with increasing sample size.

Example 5.3. *Let $X_1, \ldots, X_n$ be an i.i.d. $N(\mu, \sigma^2)$ sample, where μ is known. Find the Cramér-Rao lower bound for the variance of any unbiased estimator of σ^2.*
Compare this with the variance of the estimator $T = \frac{1}{n}\sum(X_i - \mu)^2$.

Terminology: When the variance of an unbiased estimator attains the Cramér-Rao lower bound, there can be no 'better' estimator. In general, we say that the *best estimator* is the one with the smallest variance. That is, let $T_n^*(\boldsymbol{X})$ be an unbiased estimator for θ. If for all other unbiased estimators T_n we have

$$\mathrm{Var}(T_n^*) \leq \mathrm{Var}(T_n)$$

then T_n^* is the *best unbiased* estimator of θ.

STAT3101 only:
Question: Why don't we just search for minimum mean square error estimators amongst all estimators? That is, find an estimator T_n that gives minimum mse over all possible estimators (*c.f.* Example 5.1). But note that, in general, $\mathrm{mse}(T_n; \theta)$ is a function of the unknown θ. So in comparing T_n and T_n', we would need $\mathrm{mse}(T_n; \theta) \leq \mathrm{mse}(T_n'; \theta)$ *for all* θ. In a lecturecast video it will be shown that an estimator achieving minimum mse *never* exists in non-trivial estimation problems!

5.3.3 Proof of the Cramér-Rao bound

In order to prove the (not very intuitive) Cramér-Rao inequality we need to consider an auxiliary variable. Define the random variable

$$\boxed{V_n(\boldsymbol{X}) = \frac{\partial}{\partial\theta}\, \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}$$

which is called the **score function**. Note that, like Fisher's information, there is an additive property for the score function. Again, if we have n i.i.d. observations with common density $f(x; \theta)$ then $f_{\boldsymbol{X}}(\boldsymbol{X}; \theta) = \prod\limits_i f(X_i; \theta)$ and

$$V_n(\boldsymbol{X}) = \frac{\partial}{\partial\theta}\, \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta) = \sum_i \frac{\partial}{\partial\theta} \log f_X(X_i; \theta) = \sum_i V(X_i)\,,$$

where $V(X)$ is the score function based on a single observation X.

To prove the Cramér-Rao inequality we will show the following facts about $V_n(\boldsymbol{X})$:

1. $\mathrm{E}(V_n) = 0$ (the expected score is zero)
2. $\mathrm{Cov}(V_n, T_n) = 1$
3. $\mathrm{Var}(V_n) = I(\theta)$

It will then follow from $\mathrm{Cov}(V_n, T_n)^2 \leq \mathrm{Var}(V_n)\mathrm{Var}(T_n)$ (*c.f.* §1.4) that

$$\mathrm{Var}(T_n) \geq \frac{\mathrm{Cov}(V_n, T_n)^2}{var(V_n)} = \frac{1}{I(\theta)}$$

as required. The details will be discussed in lectures. □

The conditions required for validity of the Cramér-Rao bound usually break down when the range of X depends on θ (so we cannot pass $\partial/\partial\theta$ under $\int \, d\boldsymbol{x}$.) *E.g.* $X \sim \mathrm{U}(0, \theta)$.

Finally, we note from the proof of the Cramér-Rao bound that an alternative expression for $i(\theta)$ is

$$i(\theta) = \mathrm{E}\left\{\frac{\partial}{\partial\theta} \log f_X(X;\theta)\right\}^2 .$$

The result can be generalised to vector valued θ. $I(\theta)$ then becomes a matrix

$$\left(\mathrm{E}\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f_{\boldsymbol{X}}(\boldsymbol{X}, \theta)\right\}\right)$$

and the covariance matrix of the estimators is bounded below by $I^{-1}(\theta)$.

5.4 Methods for finding estimators

5.4.1 Overview

We have discussed some desirable properties that estimators should have, but so far have only 'inspired guesswork' to find them. We need some *objective* methods that will tend to produce good estimators.

5.4.2 The method of moments

In general, for a sample $X_1, \ldots, X_n$ from a density with k unknown parameters, the **method of moments** (Karl Pearson 1894) is to calculate the first k *sample* moments and equate them to their theoretical counterparts (in terms of the unknown parameters), giving a set of k simultaneous equations in k unknowns for solution. This procedure requires *no* distributional assumptions. However, it might sometimes be helpful to assume a specific distribution in order to express the theoretical moments in terms of the parameters to be estimated. The moments may be derived using the mgf, for example.

Example 5.4. *Let $X_1, \ldots, X_n$ be an i.i.d. sample where X_i has density*

$$f(x) = \begin{cases} \theta x^{\theta - 1}, & 0 < x < 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Find the moment estimator of θ.

Example 5.5 (Example 5.1 ctd.)**.** *What are the moment estimators of μ and σ^2 given an i.i.d. normal sample?*

5.4.3 Least squares

If $X_1, \ldots, X_n$ are observations whose means are functions of an unknown parameter (or vector of parameters), θ, then the **least squares** estimator of θ is that value $\hat{\theta}_{LS}$ of θ which minimises the sum of squares:

$$\boxed{R = \sum_i \{X_i - \mathrm{E}(X_i)\}^2}$$

Example 5.6. *Let $X_1, \ldots, X_n$ be a random sample with common mean $E(X_i) = \mu$. Find the least squares estimator of μ.*

Note that no other assumptions are required — the sample does not need to be independent or identically distributed and *no* distributional assumptions are needed. The *properties* of the estimator, however, will depend on the probability distribution of the sample.

Least squares is commonly used for estimation in (multiple) regression models. For example, the straight-line regression model is

$$\mathrm{E}(X_i) = \alpha + \beta z_i$$

where the z_i are given values. The least squares estimators of α and β are obtained by solving the equations $\partial R/\partial\alpha = \partial R/\partial\beta = 0$, where $R = \sum_i (X_i - \alpha - \beta z_i)^2$.

BLUE property. In linear models least squares estimators have minimum variance amongst all estimators that are linear in $X_1, \ldots, X_n$ and unbiased – best linear unbiased estimator. (The *Gauss-Markov Theorem.*)

5.4.4 Maximum likelihood

This estimation method *does* require the distribution of the data to be *known* (*c.f.* STAT1005/MATH2501).

When $X_1, \ldots, X_n$ are i.i.d. with density (or mass function) $f_X(x_i; \theta)$ — where now we make explicit the dependence of the function on the unknown parameter(s) θ (which can be a vector) — suppose that we observe the sample values $x_1, \ldots, x_n$. Then the joint density or joint probability of the sample is

$$\prod_i f_X(x_i; \theta).$$

If we regard this as a *function of* θ for *fixed* $x_1, \ldots, x_n$ then it is called the **likelihood function** of θ, written $\mathcal{L}(\theta)$. The space of possible values of θ is the **parameter space**, Θ.

The **method of maximum likelihood** estimates θ by that value, $\hat{\theta}_{ML}$, which maximises $\mathcal{L}(\theta)$ over the parameter space Θ, *i.e.*

$$\mathcal{L}(\hat{\theta}_{ML}) = \sup_{\theta \in \Theta} \mathcal{L}(\theta).$$

Since, for a sample of independent observations, the likelihood is a product of the likelihoods for the individual observations, it is often more convenient to maximise the *log-likelihood* $\ell(\theta) = \log\ \mathcal{L}(\theta)$:

$$\ell(\theta) = \log\ \mathcal{L}(\theta) = \log \prod_i f_X(x_i; \theta) = \sum_i \log f_X(x_i; \theta).$$

Example 5.7. *Let $X_1, \dots, X_n$ be an i.i.d. sample of Bernoulli variables with parameter p. Find the maximum likelihood estimator $\hat{p}_{ML}$ of p.*

Example 5.8. *Let $X_1, \dots, X_n$ be an i.i.d. sample of Exponential variables with parameter λ. Find the ML estimator $\hat{\lambda}_{ML}$ of λ.*

Differentiation gives only *local* maxima and minima, so that even if the second derivative $d^2\ell(\theta)/d\theta^2$ is negative (corresponding to a local maximum, rather than minimum) it is still possible that the *global* maximum is elsewhere. The global maximum will be either a local maximum or achieved on the *boundary* of the parameter space. (In Example 5.8 there is a single local maximum and the likelihood vanishes on the boundary of Θ.)

Example 5.9. *Let $X_1, \dots, X_n$ be an i.i.d. sample from a Uniform distribution on $[0, \theta]$. Find the ML estimator $\hat{\theta}_{ML}$ of θ.*
NOTE: The ML estimator cannot be found by differentiating in this case. Why?

Example 5.10. *Let $X_1, \dots, X_n$ be an i.i.d. sample from a Uniform distribution on $[\theta, \theta + 1]$. Then the ML estimator of θ is not unique.*

ML estimators of transformed parameters

Example 5.11. *[Example 5.8 ctd.] Let $X_1, \dots, X_n$ be an i.i.d. sample of Exponential variables with parameter λ. Find the ML estimator of $\mu = E(X_i)$.*

Example 5.11 illustrates an important result, that if we reparameterise the distribution using particular functions of the original parameters, then *the maximum likelihood estimators of the new parameters are the corresponding functions of the maximum likelihood estimators of the original parameters.* This is easy to see, as follows.

Suppose that $\phi = g(\theta)$ where θ is a (single) real parameter and g is an invertible function. Then, letting $\tilde{\mathcal{L}}(\phi)$ be the likelihood function in terms of ϕ, we have

$$\tilde{\mathcal{L}}(\phi) = f(\boldsymbol{x}; \theta = g^{-1}(\phi)) = \mathcal{L}(\theta)$$

and so

$$\frac{d\tilde{\mathcal{L}}(\phi)}{d\phi} = \frac{d\mathcal{L}(\theta)}{d\theta}\frac{d\theta}{d\phi}$$

Assuming the existence of a local maximum, the ML estimator $\hat{\phi}_{ML}$ of ϕ occurs when $d\tilde{\mathcal{L}}(\phi)/d\phi = 0$. Since g is assumed invertible we have $d\phi/d\theta \neq 0$ so it follows that $d\mathcal{L}(\theta)/d\theta = 0$ and hence $\mathcal{L}(\phi)$ is maximum when $\theta = \hat{\theta}_{ML}$, where $\theta = g^{-1}(\phi)$. It follows that $\hat{\phi}_{ML} = g(\hat{\theta}_{ML})$. (This result also generalises to the case when θ is a vector.)

Asymptotic behaviour of ML estimators

Maximum likelihood estimators can be shown to have good properties, and are often used. Under fairly general regularity conditions it can be shown that the maximum likelihood estimator (**mle**) exists and is **unique** for sufficiently large sample size and that it is **consistent**; that is, $\hat{\theta}_{ML} \to \theta$, the true value of θ, 'in probability'. This means that when the sample size is large the ML estimator of θ will (with high probability) be close to θ.

Under some additional conditions we can make an even stronger statement. The result is that $\hat{\theta}_{ML}$ is **asymptotically normally distributed** with mean θ and variance $1/I(\theta)$ where, for an i.i.d. sample, $I(\theta) = n\mathrm{E}\{-\partial^2/\partial\theta^2 \log f_X(X;\theta)\}$ as before. That is,
$\hat{\theta}_{ML} \sim N(\theta, 1/I(\theta))$ approximately for large n. More formally,

$$\boxed{\sqrt{I(\theta)}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} N(0,1)}$$

as $n \to \infty$.

Note that the approximate variance, $1/I(\theta)$ of $\hat{\theta}_{ML}$ is of order $1/n$, which tends to zero as $n \to \infty$. Since the mean of $\hat{\theta}_{ML}$ is approximately θ, the result implies that $\hat{\theta}_{ML}$ is consistent for θ. Thus asymptotic normality is a stronger property than consistency. Furthermore, we see that the Cramér–Rao bound is approximately attained by the ML estimator. Thus the MLE is *asymptotically unbiased* and has, asymptotically, the *minimum possible variance* amongst unbiased estimators.

In addition, since the asymptotic probability distribution is *known* and *normal*, we can use it to construct tests or confidence intervals for θ. However, notice that the variance involves Fisher's information $I(\theta)$, which is unknown since we do not know θ. But it can be shown that the above asympotic result remains true if we estimate $I(\theta)$ by simply plugging in the ML estimator of θ; that is, use $I(\hat{\theta}_{ML})$.

NB: Although we have discussed the method of maximum likelihood in the context of an i.i.d. sample, this is not required. All that is needed is that we know the form of the joint density or probability mass function for the data.

Example 5.12 (Example 5.7 ctd.)**.** *For the i.i.d. Bernoulli sample, find the Cramér–Rao lower bound for the variance of any unbiased estimator of p. Compare this to the actual variance of $\hat{p}_{ML}$. Give the asymptotic distribution of $\hat{p}_{ML}$.*

Example 5.13 (Example 5.8 ctd.)**.** *Obtain the asymptotic distribution of $\hat{\lambda}_{ML}$.*

Learning Outcomes: *Chapter 5 presents the essentials of the theory of statistical estimation. It is important that you are able to*

1. *Define, explain and compare desirable properties of estimators,*
2. *Derive these properties for standard estimators,*
3. *State and derive the relationship between the mean square error, bias and variance,*
4. *Compute the Cramér–Rao bound for standard cases (eg. iid samples),*
5. *Explain the significance of the Cramér–Rao Bound for the statistical theory of estimators,*
6. *Sketch the proof of the Cramér–Rao Bound,*
7. *Explain the ideas of the method of moments, least squares and maximum likelihood in words and name their pros and cons;*
8. *Use the method of moments and of least squares to derive estimators in moderately difficult situations,*
9. *Find the maximum likelihood estimator in standard and slightly non-standard situations,*
10. *State in general, and compute in specific cases, the asymptotic distribution of the ML estimator,*
11. *Compare the finite sample properties of ML estimators with their asymptotic distribution,*
12. *In a given situation, decide which of different possible estimators is most suitable, taking their properties and the benefits / limitations of the above methods into account.*

Appendix A. Standard discrete distributions

Bernoulli distribution

$$X = \begin{cases} 0 & \text{probability } 1-p \\ 1 & \text{probability } p \end{cases}$$

or

$$p_X(x) = p^x(1-p)^{1-x}\,, \quad x = 0, 1\,;\; 0 < p < 1$$

$\mathrm{E}(X) = p\,,\ \mathrm{Var}(X) = p(1-p)$

Denoted by $\mathrm{Ber}(p)$

Binomial distribution

$$p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}\,, \quad x = 0, 1, \ldots, n\,;\; 0 < p < 1$$

$\mathrm{E}(X) = np\,,\ \mathrm{Var}(X) = np(1-p)$

Denoted by $\mathrm{Bin}(n, p)$
X is the number of successes in n independent Bernoulli trials with constant probability of success p.
X can be written $X = Y_1 + \cdots + Y_n$, where Y_i are independent $\mathrm{Ber}(p)$.

Geometric distribution

$$p_X(x) = (1-p)^{x-1}p\,, \quad x = 1, 2, \ldots\,;\; 0 < p < 1$$

$\mathrm{E}(X) = \frac{1}{p}\,,\ \mathrm{Var}(X) = \frac{1-p}{p^2}$

Denoted by $\mathrm{Geo}(p)$
X is the number of trials until the first success in a sequence of independent Bernoulli trials with constant probability of success p.

Negative binomial distribution

$$p_X(x) = \begin{pmatrix} x-1 \\ r-1 \end{pmatrix} p^r(1-p)^{x-r}\,, \quad x = r, r+1, \ldots\,;\ 0 < p < 1,\ r \geq 1$$

$\mathrm{E}(X) = \frac{r}{p}\,,\ \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

Denoted by $\mathrm{NB}(r, p)$
X is the number of trials until the rth success in a sequence of independent Bernoulli trials with constant probability of success p.
The case $r = 1$ is the $\mathrm{Geo}(p)$ distribution.
X can be written $X = Y_1 + \cdots + Y_r$, where Y_i are independent $\mathrm{Geo}(p)$

Hypergeometric distribution

$$p_X(x) = \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}\,, \quad x = 0, 1, \ldots, \min(n, M)\,;\ 0 < n, M \leq N$$

$\mathrm{E}(X) = \frac{nM}{N}\,,\ \mathrm{Var}(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}$

Denoted by $\mathrm{H}(n, M, N)$
X is the number of items of Type I when sampling n items without replacement from a population of size N, where there are M items of Type I in the population.
May be approximated by $\mathrm{Bin}(n, \frac{M}{N})$ as $N \to \infty$

Poisson distribution

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}\,, \quad x = 0, 1, \ldots\,;\ \lambda > 0$$

$\mathrm{E}(X) = \lambda\,,\ \mathrm{Var}(X) = \lambda$

Denoted by $\mathrm{Poi}(\lambda)$
X is the number of random events in time or space, where the rate of occurrence is λ.
Approximation to $\mathrm{Bin}(n, p)$ as $n \to \infty$, $p \to 0$ such that $np \to \lambda$.

Appendix B. Standard continuous distributions

Uniform distribution

$$f_X(x) = \frac{1}{b-a}\,, \quad a < x < b\,;\ a < b$$

$\mathrm{E}(X) = \frac{b+a}{2}\,,\ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$

Denoted by $\mathrm{U}(a,b)$
Every point in the interval (a,b) is 'equally likely'.
Important use for simulation of random numbers.

Exponential distribution

$$f_X(x) = \lambda e^{-\lambda x}\,, \quad x > 0\,;\ \lambda > 0$$

$\mathrm{E}(X) = \frac{1}{\lambda}\,,\ \mathrm{Var}(X) = \frac{1}{\lambda^2}$

Denoted by $\mathrm{Exp}(\lambda)$
X is the waiting time until the first event in a Poisson process, rate λ.

Gamma distribution

$$f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}\,, \quad x > 0\,;\ \lambda > 0,\, \alpha > 0$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

is the gamma function.
$\Gamma(r) = (r-1)!$ for positive integers r, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
$\mathrm{E}(X) = \frac{\alpha}{\lambda}\,,\ \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}$

Denoted by $\mathrm{Gam}(\alpha,\lambda)$
When $\alpha = r$, an integer, X is the waiting time until the rth event in a Poisson process, rate λ. The case $\alpha = 1$ is the $\mathrm{Exp}(\lambda)$ distribution.
When $\alpha = r$, an integer, X can be written $X = Y_1 + \cdots + Y_r$, where Y_i are independent $\mathrm{Exp}(\lambda)$.

More generally we have the following additive property: if $Y_i \sim$ Gam(α_i, λ), $i = 1, \ldots, r$, and are independent then $X = Y_1 + \cdots + Y_r \sim$ Gam$(\sum_{i=1}^{r} \alpha_i, \lambda)$.

Beta distribution

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

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is the beta function.
$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}\,,\ \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Denoted by Beta(α, β)
The case $\alpha = \beta = 1$ is the U$(0, 1)$ distribution.
X sometimes represents an unknown proportion lying in the interval $(0, 1)$.

Normal distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\,, \quad -\infty < x < \infty\,;\ \sigma > 0$$

$\mathrm{E}(X) = \mu\,,\ \mathrm{Var}(X) = \sigma^2$

Denoted by $N(\mu, \sigma^2)$
Widely used as a distribution for continuous variables representing many real-world phenomena, and as an approximation to many other distributions.