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. If you are unable to meet the in-course assessment submission deadline for reasons outside your control, for example illness or bereavement, you ***must*** submit a claim for extenuating circumstances, normally within a week of the deadline. Your home department will advise you of the appropriate procedures. For Statistical Science students, the relevant information is on the DOSSSH Moodle page.

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).

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

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's Moodle site accessible from `https://moodle.ucl.ac.uk/`. In addition, the Moodle site will include (i) information and background videos as well as links to lecturecast recordings to all lectures as far as technically possible (ii) answers to section A and B questions on the exercise sheets after these have been discussed in the tutorials, (iii) very succinct answers to section C questions (eventually), (iv) past in course assessments and exams as well as very succinct answers, and (v) 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

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

Example 1.3. *A child gets to throw a fair die. If the die comes up 5 or 6, she gets to sample a sweet from box A which contains 10 chocolate sweets and 20 caramel sweets. If the die comes up 1,2,3 or 4 then she gets to sample a sweet from box B which contains 5 chocolate sweets and 15 caramel sweets. What is the conditional probability she will get a chocolate sweet if the die comes up 5 or 6? What is the conditional probability she will get a chocolate sweet if the die comes up 1,2,3 or 4? What is her probability of getting a chocolate sweet?*

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 1.4. *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 1.5 (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 1.6. *Consider the random variable $X \sim \text{Geo}(p)$ 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 1.7. *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 $P(|X| > 1/2)$.*

Example 1.8. *Pick a volunteer from the audience and ask his/her height in feet and inches. Then offer the following sequence of bets:*

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:

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

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:

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

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

Example 1.9. *The discrete random variable X has pmf $p_X(k) = \frac{1}{e^\mu - 1} \frac{\mu^k}{k!}$ for $k \in \mathbb{N}$. Compute its 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

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

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 1.10 (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 $\mathbb{E}[X]$ and $\mathbb{E}[\phi(X)]$.*

The **variance** of X is

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

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

Example 1.11 (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*):

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

Example 1.12 (1.1 ctd. V)**.** *Let Y be the excess of heads over tails obtained on the two tosses of the coin. Write down $\mathbb{E}[Y]$ 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 or continuous*
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 1.13. *Consider the function*

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

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 1.13 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 1.14. *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 1.14 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 1.15 (1.14 ctd.)**.** *Obtain the conditional pmf of X given $Y = y$. Use this conditional distribution to verify that X and Y are not independent.*

Example 1.16. *Suppose that R and N have a joint distribution in which $R|N$ is $Bin(N, \pi)$ and N is $Poi(\lambda)$. 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:

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

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

Example 1.17 (1.14 ctd. II)**.** *Find the conditional expectations* $\mathbb{E}[X|Y = y]$ *for* $y = 0, 1, 2$*. Plot the graph of the function* $\phi(y) = \mathbb{E}[X|Y = y]$*. What do these values tell us about the relationship between* X *and* Y*?*

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

Example 1.18. *Collect the joint distribution of* X*: gender (*$x_1 = M, x_2 = F$*) and* Y*: number of cups of tea drunk today (*$y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 3$ 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 example that the overall mean is just the average of the conditional means. We now prove this fact in general. Consider the conditional expectation $\phi(y) = \mathbb{E}_{X|Y}[X|Y = y]$ 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 $\mathbb{E}_{X|Y}[X|Y]$ because the possibly more correct notation $\mathbb{E}_{X|Y}[X|Y = Y]$ would be even more confusing! We may then compute the expectation of our new random variable $\phi(Y)$, i.e. $\mathbb{E}[\phi(Y)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]]$. But, from the definition of the expectation of a function of Y, we have $\mathbb{E}[\phi(Y)] = \sum_{y_j} \phi(y_j)p_Y(y_j)$, so that

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

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

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

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 1.19 (1.14 ctd. II)**.** *Verify that* $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]]$ *in this example.*

Example 1.20 (1.16 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.*

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

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

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

Taking out what is known (TOK)

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

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

Example 1.21. *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 $\mathbb{E}(XY)$.*

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 1.22. *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 1.23 (1.22 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 1.24 (1.22 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 $\mathbb{E}[X|Y = y]$ and $\mathbb{E}[X]$, using the law of iterated conditional expectation for the latter. Compare this with the direct calculation of $\mathbb{E}[X]$.*

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

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

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 $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. 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

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

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

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

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

$$\mathbb{E}(\text{product}) \neq \text{product of expectations}$$

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

$$\mathbb{E}(\text{sum}) = \text{sum of expectations}$$

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

$$\boxed{\text{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]}$$

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

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

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

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

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

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 1.26.

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

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

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

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

for all random variables X and Y.

Example 1.25 (1.21 ctd.)**.** *Find the covariance and correlation of X and Y.*

Example 1.26. *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 $z \in \mathbb{R}$,

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

as a quadratic function of z. But $\mathrm{Var}(zX+Y) \geq 0$, 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

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

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

We get the extreme values, $\mathrm{Corr}(X,Y) = \pm 1$, when the quadratic touches the z-axis; that is, when $\mathrm{Var}\,(zX+Y) = 0$. 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

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

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

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

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

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

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 $\mathrm{corr}(X,Y) = 0 \not\Rightarrow$ 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

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

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

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

Example 1.27 (1.21 ctd.)**.** *Find the conditional variances of X given $Y = 0, 1$. Compute the marginal variance of X by using the above result.*

Example 1.28 (1.16 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* $-1 \leq Corr \leq 1$.

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

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

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, $\ldots$, n_{m+1} in the last category. Any specific ordering of these n outcomes has probability $p_1^{n_1} \ldots p_{m+1}^{n_{m+1}}$ by the assumption of independent trials, and there are $\frac{n!}{n_1! \ldots n_{m+1}!}$ such orderings.

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

Example 1.29. *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, $N_i \sim \text{Bin}(n, p_i)$, 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 $N_1, \ldots, N_m$ by using the multinomial expansion as follows. To do this, you will need the result that

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

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

Example 1.30. *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

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

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

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

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

Finally

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

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 1.31 (1.30 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

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

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 $\rho = \mathrm{Corr}(X,Y) > 0$, or is relatively small if $\rho < 0$.

Suppose that $\rho = 0$. Then

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

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

Example 1.32. *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

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

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 $x^T = (x_1, x_2, x_3, \ldots, x_n)$, 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**, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ 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 $\boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}$,

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

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

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

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 $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\rho_{ij} = \mathrm{Corr}(X_i, X_j)$ and $\mathrm{var}(X_i) = \sigma_i^2$. 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 1.33. *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.*