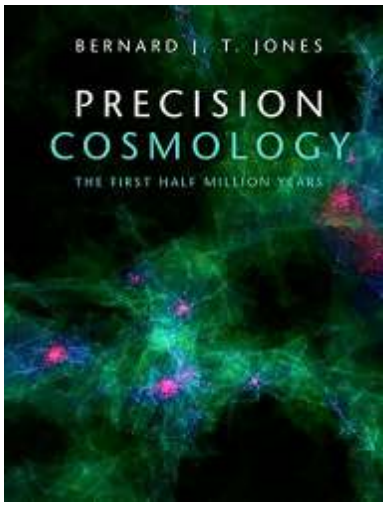
Probability and Statistics Primer

A Supplement to “Precision Cosmology”

Bernard Jones



A short introductory course on Probability and Statistics. This is mainly built from my older lectures and lecture notes. It shares some material with my *Strasbourg Lectures* and there is also a small overlap with my ”*Precision Cosmology*”. My own sources for this material, aside from the lecture notes, are diverse texts, most notably the advanced texts of Cox and Hinkley (1979) and Papoulis and Pillai (2002).

This is one of a set of Supplementary Notes and Chapters to “Precision Cosmology”.
Some of these Supplements might have been a chapter in the book itself, but were regarded either as being somewhat more specialised than the material elsewhere in the book, or somewhat tangential to the main subject matter.
The are mostly early drafts and have not been fully proof-read.
Please send comments on errors or ambiguities to “PrecisionCosmology(at)gmail.com”.

Contents

1 Probability and Statistics - the basics

> The unsatisfactory state of probability theory is bound to tell on modern physics in which the former plays such a prominent part. ... I do admit that the readiness with which we physicists and many others who use the theory in practice adopt Frequency as the basis does mean taking things a little too easy. There are grave objections to this, mainly that we thereby cut ourselves off from ever applying rational probability considerations to a single event. (Schrödinger, 1945-1948)

1.1 Probability - the rules

Probability is the mathematics of chance. The mathematical study of chance goes back to Girolamo Cardano's *Liber de ludo aleae* ("On Casting the Die")[1] which was written in 1526, but not published until 1663. The theory was developed by Pierre de Fermat, Blaise Pascal (1654) and by Christian Huygens. It was Huyghens who in 1657 developed the first scientific treatment of the subject in his *De ratiociniis in aleae ludo* (Calculations in Games of Chance) (Huygens, 1889). The early 18^{th} century saw the important publication in 1711 of de Moivre's *Doctrine of Chances* (de Moivre, 1718, 1738) and Jacob Bernoulli's *Ars Conjectandi* (Bernoulli, 1713).

Perhaps surprisingly, the notion of what "probability" is and what it measures is highly controversial. Jaynes in his *Probability Theory: The Logic of Science* (Jaynes, 2003) says that "People who believe in physical probabilities are like those who believe in astrology: they never ask what would constitute a controlled experiment capable of proving or disproving their belief". This may be a good point, but it is not very helpful in coming to terms with our naive intuition that probability measures something that describes the frequencies with which particular outcomes of some observation or experiment occur.

This naïve notion is the basis of the "frequentist" interpretation of probability. The frequentist dogma was set by Fisher (1925 (1^{st} edition)) is his *Statistical Methods for Research Workers*, a book which went to 14 editions and dominated the 20^{th} century statistical literature [2].

The notion that physical probabilities have some kind of reality independently of human

[1] Cardano (1501-1576?) was a friend of Leonardo da Vinci, and is a noted mathematician and physician. In 1554 he had computed the horoscope of Jesus, and in 1570 was accused of heresy and imprisoned.

[2] Jaynes (2003, p.492 of 7^{th} printing, 2010) comments that "... it acquired such an authority over scientific practice in some fields such as medical testing found it impossible to get their work published if they failed to follow Fisher's recipes to the letter".

information is a fundamental tenet of the frequentist view. Alternatives to the frequentist view are that

1 Probability is a measure of our state of knowledge
2 Probability is a measure of our degree of belief that an event will occur.

Neither of these comes close to our naïve notion that probability has something to do with throwing a dice, in fact they don't seem to have much to do with Physics. This vagueness arises because they both involve rather abstract notions: "our state of knowledge" and "our degree of belief". There were early discussions of this issue was by the noted economist Keynes (1921) and , somewhat later, by Kendall (1949). Schrödinger (1945-1948) and Feynman (1951) addressed the issue with regard to the interpretation of the wave-function that characterised their versions of quantum theory. See also Feynman and Hibbs (1965, pp. 5-8) [3].

What are these "measures"? The view put forward by Salmon (1967) is that any theory of probability must, at least, satisfy the following three criteria. The theory must be

Admissible It must satisfy the precepts of some probability calculus
Ascertainable It must produce numbers that we can determine
Applicable The values so-determined must relate to experience or some theory

These are perhaps self-evident: we want our measure to be based on a rigorous and consistent mathematical foundation (otherwise we are asking for trouble), we want to be able to use the measure (otherwise why bother?), and the measure must bear some relationship to our experiences (otherwise nothing is clarified or explained).

There may be additional requirements. For example we may require the theory to be *causal* in the sense that some events may be influenced by other events. This would impose a sense of time ordering on the theory, which poses difficulties since time ordering and causation are merely empirical, as opposed to logical.

Mathematicians, and notably Kolmogorov (1933); Doob (1934, 1941); von Mises (1941), have tried to formalise the concept in an abstract way. But like so many formal mathematical systems it is often difficult to discern in this approach the elements of our naive (frequentist) perception of what a probability is.

Bayesian probability interprets the concept of probability as "a measure of a state of knowledge". Precisely what this measure is is a subject of much discussion. So in the

[3] Feynman (1951) discussed the interpretation of probability in quantum mechanics where probability is determined by the square modulus of the value of a complex function ϕ: $P = |\phi|^2$. He illustrated this by considering an electron passing through one of two holes in a screen when the other was closed. The probability associated with hole '1' when hole '2' was closed is $P_1 = |\phi_1|^2$, and, likewise, when hole '1' was closed, $P_2 = |\phi_2|^2$. He asserted that when both holes were open

$$P = |\phi|^2, \quad \phi = \phi_1 + \phi_2$$

and so $P \neq P_1 + P_2$: probabilities in quantum mechanics are not additive. This result is verified by observing what happens on the back side of the screen. We can put this non-classical result to a test and do the experiment simply by watching the electron. In doing that experiment we deduce that the electron passes through only one of the holes and we can infer that $P = P_1 + P_2$. This is one of the fundamental paradoxes posed by quantum mechanics, yet nobody would doubt the efficacy of quantum mechanics as a theory of our experience on sub-atomic scales.

present context it is best to simply regard probability as a measure on a space of events and to interpret that measure formally in your favourite way.

1.1.1 Axioms of Probability

One way to proceed is to lay down the rules, or axioms, of a model for probability. Probability is about *events*, and so we have to say what we mean by an *event*, and about a measure on the space of events that we shall refer to as the *probability*. We have to define rules for that measure. At the highest level of mathematical formality we can follow this route through the work of Kolmogorov (1933).

However, for the present purposes it is useful to go to a more intuitive level. Let us start with a couple of definitions describing what we are talking about: an experiment that has a number of outcomes. The outcome might be a state ("black", "red"), a value, X, or a value within some range or interval $(X, X + dX)$.

Definition 1.1
Sample Space: The *sample space* is the collection of all possible different outcomes of an experiment. We can denote the set of all possible outcomes by the symbol $\mathbf{\Omega}$.

Definition 1.2
Point: A *point* in the sample space is a possible outcome of an experiment.

Definition 1.3
Event: An *event* is any set of points in the sample space. So we can think of a family $\mathcal{E}$ of subsets of $\mathbf{\Omega}$ that represents the events. The empty set, $\varnothing$ should be a member of $\mathcal{E}$: $\varnothing \in \mathcal{E}$.

Definition 1.4
Probability: Each point is assigned a *probability*: a real number $\mathbb{P}$ between 0 and 1. Formally, $\mathbb{P}$ would be a function defined on the subsets $\mathcal{E}$. We insist that $\mathbb{P}[\varnothing] = 0$ and $\mathbb{P}[\mathbf{\Omega}] = 1$.

We also need to stipulate that each outcome is represented by precisely one point in the sample space. Thus we speak in terms of the *probability of a given outcome* for the exper-

iment. The three components defined above, $\{\boldsymbol{\Omega}, \boldsymbol{\mathcal{E}}, \mathbb{P}\}$, form what is called a *probability space*: a model for understanding what happens in the real world[4].

Quite what this probability measure actually measures has been the subject of debate for many years (von Mises and Doob, 1941, for example). Naïvely we might think it reasonable to expect that, in some limit, it reflects the frequency with which a particular outcome would occur if the experiment were to be repeated an infinite number of times [5]. See Jaynes (2003, Appendix A) for an overview of some of the various approaches to interpreting probability.

This allows us to conjure up another definition:

Definition 1.5
Probability function: A *probability function* is a function that assigns probabilities to the events of a sample space.

Traditionally, we denote events by the symbols $A, B, ...$ According to Definition 1.3 we are to think of these events as sets in the sense of set theory: e.g: $E \in \boldsymbol{\mathcal{E}}$.

We are going to need the following definition:

Definition 1.6
Mutually exclusive: Two events having no points in common are said to be *mutually exclusive*

Following the precepts of set theory, two special sets are defined: the empty set $\varnothing$ and the set of all possible outcomes for the experiment, denoted by $\boldsymbol{\Omega}$. We then have the following axioms:

Axiom 1 For any event E, $0 \leq \mathbb{P}[E] \leq 1$

Axiom 2 $\mathbb{P}[\varnothing] = 0, \quad \mathbb{P}[\boldsymbol{\Omega}] = 1$
This asserts that the probability of an event that cannot occur is zero and that the probability of entire sample space, $\boldsymbol{\Omega}$, is 1. The latter is sometimes referred to as the axiom of certainty.

Axiom 3 If A and B are mutually exclusive events, then
$\mathbb{P}[A + B] = \mathbb{P}[A] + \mathbb{P}[B]$
where $A + B$ denotes the event that one of A and B occurs [6].

By *mutually exclusive* we mean that either A or B can occur, but not both.

[4] From a mathematical point of view there are further conditions needed to fully define a probability space. See, for example Nelson (1992)

[5] The notion that probability merely reflects a fraction of outcomes in a series of repeated measurements is dubbed *the frequentist view*. Recognising that this is what naïve intuition tells us, we could interpret this as some kind of "frequentist limit" of our probability theory. It has long been recognised that this line of argument leads to a somewhat limited view of what *probability* might be and what it might measure. However, it serves as a useful, if naïve starting point for this discussion.

[6] In set theory parlance this is $A \cup B$ when A and B are mutually exclusive, *i.e.* $A \cap B = \varnothing$.

We can use the notation

AB	both A and B occur
$A + B$	at least one of A and B occurs

These correspond to the intersection and union of the events A and B: the event $A + B$ can be denoted by $A \cup B$ and the event AB by $A \cap B$, which exploit the obvious analogy with set theory. The $A \cup B$, $A \cap B$ notation is generally used in mathematical set-theoretic approaches to probability.

We can denote probabilities $\mathbb{P}[(\,]X)$ for a variety of events X as follows

$\mathbb{P}[A\,]$	A occurs
$\mathbb{P}[B\,]$	B occurs
$\mathbb{P}[\overline{A}\,]$	A does not occur
$\mathbb{P}[A \mid B\,]$	A occurs, given B
$\mathbb{P}[AB\,]$	A and B both occur
$\mathbb{P}[A + B\,]$	at least one of A and B occurs

and so on.

It is axiomatic (or obvious - depends how you look at it) that

$$\mathbb{P}[A\,] + \mathbb{P}[\overline{A}\,] = 1 \tag{1.1}$$

$$\mathbb{P}[A + B\,] = \mathbb{P}[A\,] + \mathbb{P}[B\,] - \mathbb{P}[AB\,] \tag{1.2}$$

Equation (1.2) is the logical extension of Axiom 3 and can be understood on the basis of the set theoretic view of events [7]. We shall discuss equation (1.2) at length below (see section (1.2.1) *et seq.*).

Events A and B are said to be *mutually exclusive* when they cannot happen together [8]

Definition 1.7

Mutually exclusive events: If events A and B are mutually exclusive

$$\mathbb{P}[AB\,] = 0 \tag{1.3}$$

in which case we have, as per equation (1.2) and Axiom 3

$$\mathbb{P}[A + B\,] = \mathbb{P}[A\,] + \mathbb{P}[B\,], \quad A \text{ and } B \text{ mutually exclusive.} \tag{1.4}$$

Using the alternate notation $A \cap B$ for AB, we appreciate that mutually exclusive events have zero intersection: $A \cap B = \varnothing$. Clearly the events A and $\overline{A}$ are mutually exclusive.

[7] Conversely, we could regard (1.2) as an axiom and then what we have chosen to be Axiom 3 would follow with $\mathbb{P}[AB\,] = 0$ as a definition of independence.

[8] As when throwing a dice: it cannot come up both '3' and '6' in the same throw. Since the probability of A: throwing '3' is $\frac{1}{6}$ and the probability of B: throwing a '6' is also $\frac{1}{6}$, the probability of throwing either a '3' or a '6' is $\mathbb{P}[A + B\,] = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

1.2 Derived concepts

1.2.1 Conditional probabilities

The basic framework is established and the next step is to introduce a way of working with events that are not independent, *i.e.* $\mathbb{P}[AB] \neq 0$, by building on equation (1.2). We start with an assertion of what we mean by *independent events* that we will take to be a definition [9]

Definition 1.8
Independent events: Events A and B are said to be *independent* if

$$\mathbb{P}[AB] = \mathbb{P}[A]\,\mathbb{P}[B], \quad A \text{ and } B \text{ independent} \tag{1.5}$$

This is the first time in our discussion of probability that we have given a meaning to the product of probabilities $\mathbb{P}[A]\,\mathbb{P}[B]$, and additionally insisted that this product $\mathbb{P}[A]\,\mathbb{P}[B]$ can itself meaningfully be interpreted as a probability.

Now we introduce the concept of *conditional probability* which relates the probability $\mathbb{P}[AB]$ of both of A and B occurring to the probability of either $\mathbb{P}[A]$ or $\mathbb{P}[B]$:

Definition 1.9
Conditional Probability:

$$\mathbb{P}[A\,|\,B] = \frac{\mathbb{P}[AB]}{\mathbb{P}[B]}, \quad \mathbb{P}[B] > 0 \tag{1.6}$$

Clearly A and B are independent if $\mathbb{P}[A\,|\,B] = \mathbb{P}[A]$: the event B has no influence on A. In other words, when A and B are independent, knowing anything about B does not provide any information about A, and vice versa, in which case we recover equation (1.5) [10].

Using equation (1.6) we deduce that

$$\mathbb{P}[AB] = \mathbb{P}[A\,|\,B]\,\mathbb{P}[B] = \mathbb{P}[B\,|\,A]\,\mathbb{P}[A] \tag{1.7}$$

The first equality says that the probability of both A and B occurring is the probability that A occurs, given that B has occurred, multiplied by the probability the B occurs, and likewise for the second equality, reversing the roles of A and B. A key consequence of equation (1.7)is *Bayes Theorem*:

[9] Independent events are not to be confused with exclusive events! For the event of throwing two dice, one red and one blue, the probability of A: throwing a '3' on the red dice and B: throwing a '6' on the blue one is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. The result A does not influence B, and so the events A and B are independent but not exclusive. This is a frequent cause of confusion.

[10] The motivation is simple: consider events A and B and count the number of times, n_{AB}, both A and B occur and the number of times, n_B, B occurs in n observations. Intuitively, we can see that as n gets very large, $\mathbb{P}[A\,|\,B] \simeq n_{AB}/n_B = (n_{AB}/n)/(n_B/n) \simeq \mathbb{P}[AB]/\mathbb{P}[B]$.

Bayes Theorem

$$\mathbb{P}[A\,|\,B\,] = \mathbb{P}[B\,|\,A\,]\frac{\mathbb{P}[A\,]}{\mathbb{P}[B\,]} \tag{1.8}$$

As a simple example, consider drawing two cards from a standard deck of 52 cards. The probability that either of the cards is a king (K) is $\frac{4}{52}\frac{48}{51} + \frac{48}{52}\frac{4}{51} + \frac{4}{52}\frac{3}{51} = \frac{396}{2652} \simeq 0.149$, this is a simple counting exercise [11]. Likewise the probability of getting a king, K, and a queen, Q in the two cards is $2\times\frac{4}{52}\frac{4}{51} = \frac{16}{2652} \simeq 0.012$. So what is the probability that one of the two cards is a K, given that the other is a Q, i.e. $\mathbb{P}[K\,|\,Q\,]$? Equation (1.6) allows us to express $\mathbb{P}[K\,|\,Q\,]$ as $\mathbb{P}[KQ\,]/\,\mathbb{P}[Q\,] \simeq 0.012/0.149 = 0.081$. This is frequently a source of surprise to some people who might think that knowing there is a Q among the two cards has no effect on what the other might be. The fallacy there is that $\mathbb{P}[KQ\,] \neq \mathbb{P}[K\,]\,\mathbb{P}[Q\,]$: the events K and Q are not independent.

Ex 1.2.1 Use equation (1.6) to show that

$$\mathbb{P}[A\,|\,B\,] + \mathbb{P}[\overline{A}\,|\,B\,] = 1 \tag{1.9}$$

Since the events $A\,|\,B$ and $\overline{A}\,|\,B$ are mutually exclusive, we recognise from equation (1.22) the ratio of the two terms $\mathbb{P}[A\,|\,B\,]/\,\mathbb{P}[\overline{A}\,|\,B\,]$ on the left of (1.9) as being the odds for $A\,|\,B$.

Following directly from equation (1.6) we have the *chain-rule relationships*:

Chain rules:

$$\mathbb{P}[AB\,] = \mathbb{P}[A\,]\,\mathbb{P}[B\,|\,A\,] \tag{1.10}$$

$$\mathbb{P}[ABC\,] = \mathbb{P}[A\,]\,\mathbb{P}[B\,|\,A\,]\,\mathbb{P}[C\,|\,AB\,] \tag{1.11}$$

$$\dots$$

$$\mathbb{P}[A_1A_2\dots A_n\,] = \mathbb{P}[A_1\,]\,\mathbb{P}[A_2\,|\,A_1\,]\,\mathbb{P}[A_3\,|\,A_1A_2\,]\dots\,\mathbb{P}[A_n\,|\,A_1A_2\dots A_{n-1}\,] \tag{1.12}$$

and so on. Equation (1.12) is proved by induction starting with (1.10) [12].

[11] The first term is the probability of getting K on the first card and not on the second, the second term is the probability of not getting a K on the first card and getting K on the second, and the third term is the probability of getting two K's

[12] This is also seen written out as

$$\mathbb{P}[A_1\cap A_2\cap\dots\cap A_n\,] = \mathbb{P}[A_1\,]\,\mathbb{P}[A_2\,|\,A_1\,]\,\mathbb{P}[A_3\,|\,A_2\cap A_1\,]\,\mathbb{P}[A_4\,|\,A_3\cap A_3\,\cap A_1\,]\dots \mathbb{P}[A_n\,|\,A_{n-1}\cap A_{n-2}\cap\dots\cap A_2\cap A_1\,] \tag{1.13}$$

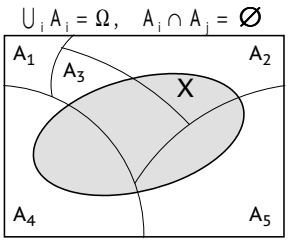


Fig. 1.1 Total Probability. Here the subsets $\{A_i\}$ that cover $\mathbf{\Omega}$ are independent, they form a partition of $\mathbf{\Omega}$. Hence $\bigcup_i A_i = \mathbf{\Omega}$ and $A_i \cap A_j = \varnothing$ for all $i \neq j$. The event X (shown shaded) can then be expressed as a sum over all its intersections $X \cap A_i$

Ex 1.2.2 Use equations (1.10) and (1.11) to show that

$$\mathbb{P}[AB|C] = \mathbb{P}[A|BC]\,\mathbb{P}[B|C] = \mathbb{P}[B|AC]\,\mathbb{P}[A|C] \tag{1.14}$$

Another important relationship can be derived when the events $A_1, A_2, \ldots, A_n$ are pairwise exclusive and cover the space of events $\mathbf{\Omega}$ (see figure 1.1). The probability of any random event X occurring is then

$$\begin{aligned}\mathbb{P}[X] = {} & \mathbb{P}[A_1]\,\mathbb{P}[X|A_1] + \\ & \mathbb{P}[A_2]\,\mathbb{P}[X|A_2] + \cdots + \mathbb{P}[A_n]\,\mathbb{P}[X|A_n] \\ & \mathbb{P}[A_i] > 0 \quad \text{for } i = 1, \ldots, n\end{aligned} \tag{1.15}$$

which can be written more succinctly as

Total probability theorem

$$\mathbb{P}[X] = \sum_i \mathbb{P}[X|A_i]\,\mathbb{P}[A_i], \quad A_i \cap A_j = \varnothing, \quad \forall i \neq j. \tag{1.16}$$

This result is known as the *total probability theorem*.[13]

Conditional probability has been a serious point of debate over the past century. One of the issues is that equation (1.8) is symmetric under interchange of A and B. In a causal system B might precede A and so there is no sense in being able to talk about the probability of B occurring given A. The reason that this arises is that, in our probability space model $\{\mathbf{\Omega}, \mathcal{E}, \mathbb{P}\}$, the space of events $\mathcal{E}$ is not an ordered set. Put another way, our notation is inadequate to handle the situation. The way around this is to model causal systems using Markov models which describe time evolution in terms of transitions that can be assigned probabilities.

[13] This can be seen as follows. Because the A_i are a complete and mutually exclusive set of events, X is the union of the intersections of X with all the A_i: $X = XA_1 + \cdots + XA_n$. The events XA_i are mutually exclusive and so we have $\mathbb{P}[X] = \mathbb{P}[XA_1] + \cdots + \mathbb{P}[XA_n]$. Our result follows because $\mathbb{P}[XA_i] = \mathbb{P}[X|A_i]\,\mathbb{P}[A_i]$.

1.2.2 Expectation

While we may have no innate concept of probability (de Finetti, 1937, 1974; Jaynes, 2003) we certainly have a clear intuitive notion of "expectation" as in "what is the expected outcome of this procedure", to which the response might be "There is about a 12% chance that ...". The level at which you might anticipate a particular outcome is formulated as a probability relative to a set of possible outcomes. This is, in a sense, a counting exercise: list how often each of those possible outcomes has happened in the past [14].

If we assign a numerical value $X(a)$ to each of the events $a \in \mathbf{\Omega}$, we can define the *expectation* of the function X over the set of events $\mathbf{\Omega}$ as

Definition 1.10
Expectation: The *expectation* of the function $X(a)$ over the set of events $a \in \mathbf{\Omega}$ is

$$\mathbb{E}[X] = \sum_{a \in \mathbf{\Omega}} X(a)\ \mathbb{P}[a] \tag{1.17}$$

This is perhaps one of the most important concepts in probability: it tells us how to assign probabilities to quantities that depend on the events under consideration.

1.2.3 Indicator random variables

We can illustrate this definition of expectation with the following important example, the concept of an *Indicator random variables*. An indicator random variable, I_A, has the defining property that it takes the value 1 or 0 according as to whether a specified event, A, occurs or not. If the probability of A occurring is $p = \mathbb{P}[A]$, then the expectation, $\mathbb{E}[I_A]$, and variance $\mathrm{Var}(I_A) = \mathbb{E}[I_A^2] - \mathbb{E}[I_A]^2$ of the random variable I_A are:

$$\mathbb{E}[I_A] = p \tag{1.18}$$

$$\mathrm{Var}(I_A) = p(1-p). \tag{1.19}$$

The proof of the first of these follows directly from the definition (**??**) of expectation:

$$\mathbb{E}[I_A] = 1.\,\mathbb{P}[A] + 0.\,\mathbb{P}[\bar{A}] = p \tag{1.20}$$

The second follows similarly and is left as an exercise.

1.2.4 Odds

We can introduce a useful definition [15]

[14] At this point we hear screams of "Frequentism!". There is no accoutning for any other prior information we may have at our disposal, at which point the Frequentists scream "Bayesians!". See Jaynes (2003, §3.4)

[15] The notation follows Good (1975) who attributes this to **?** and to a private communication with Turing in 1940. See also Kass and Raferty (1995).
I.J. Good went to Bletchley in May 1941, after finishing his undergraduate mathematics degree in Cambridge, to work with Turing. At Bletchley, he developed what is now referred to as "Good-Turing frequency estima-

Definition 1.11
Odds: Odds associated with probability p:

$$O(p) = \frac{p}{1-p} \tag{1.21}$$

Then the odds of the event A is ratio of the two terms $\mathbb{P}[A]$ and $\mathbb{P}[\overline{A}]$:

$$O(\mathbb{P}[A]) = \frac{\mathbb{P}[A]}{\mathbb{P}[\overline{A}]} = \frac{\mathbb{P}[A]}{1-\mathbb{P}[A]} \tag{1.22}$$

In other words this is the odds of A compared with the the alternative $\overline{A}$, this conforms to the well known gambling terminology. Note that the odds is not a measure of probability. The logarithm of the odds $\ln O(p)$ is sometimes used [16].

1.3 Univariate discrete and continuous distributions

We now move on to practical descriptors of probabilities describing the possible outcomes of measurements. To this end it is conventional to talk of one or more random variables $X, Y, \ldots$ taking on values $x, y, \ldots$ respectively. Thus x is the value of a possible outcome of a measurement of the random variable X. The $X, Y, \ldots$ can be regarded as the set of all possible outcomes of one or more experiments and so the values of the outcomes, $x, y, \ldots$, can be assigned probabilities. Apart from the fact that as yet we have no prescription for assigning those probabilities, this is all fairly self-evident [17].

The space of $x, y, \ldots$ values can be all or part of the real line, it could be a set of complex numbers or even a set of matrices. It could be a qualitative, non-numeric, attribute like colour, taste or smell or even an opinion (such as: good or bad, true or false). In practical terms we need to distinguish these various possibilities. In the past statisticians would speak of different *measurement scales*: Nominal scales, Ordinal scales, Interval scales and Ratio scales. A nominal scale is a set of attributes that do not have numerical values. The attributes may be identified by a number, but there will be no sense or order in the assigned numbers. Nominal data is also referred to as *categoric data*. In an ordinal scale, the values assigned have a sense of order even if the attributes are non-numerical (eg: "bad" = -1, "indifferent" = 0, "good" = 1). An interval scale is an ordinal scale in which the size of the interval between successive elements has relevance. A ratio scale is an interval scale in which the ratios of the assigned values has relevance.

tion", an essential step in the deciphering of the Naval Enigma cypher (see, for example, Good (1979, 2000)). Good wrote many lucid books and erudite articles, notably on Bayesian statistics. In Good's obituary, *The Independent* newspaper commented that he had "almost a thousand papers to his credit" ranging over "statistics, physics, mathematics and philosophy"). Jaynes (2003, p.709), however, criticises Good as one who "persisted in believing in the existence of physical probabilities that have some kind of reality independent of human information".

[16] Generally speaking all logarithms used in relation to probability and statistics are to the base e: $\ln x = \log_e x$.

[17] This can be formally expressed in terms of the probability space $\{\boldsymbol{\Omega}, \mathcal{E}, \mathbb{P}\}$

The importance of this classification of data types was to make it clear what could or could not be done with a particular data set. It makes sense to talk about the mean and standard deviation of interval and ratio data, but not of nominal or ordinal data. However, with ordinal data it is possible to use powerful "non-parametric" statistical methods based on rank (order) to analyse and compare data samples.

We now think mainly in terms of *categorical data*, i.e. nominal data on the older classification, versus *quantitative* data. In the physical sciences most of the data that comes from a measurement process is quantitative data [18].

Database architectures are much more specific about the types of the data that is entered into the various database fields. Among numeric-valued fields, database technology distinguishes binary, integer (of different size), decimal, date-time, text strings, phone numbers and so on. The reason for this is that the type of field permits only certain operations on their values. You can sort phone numbers, but it makes no sense to take an average and you cannot multiply dates or times.

It will be useful to talk in terms of *data types* for quantitative data where the type reflects the nature of the measurement being made and often the units of measurement. The data describing measurements of a distant galaxy may consist, among other things, of the galaxy redshift, z, its $U-B$ and $B-V$ colours, its apparent magnitude m and its apparent diameter θ. These are the attributes describing the object. The apparent magnitude, m and apparent diameter θ are attributes having different data types, while the $U-B$ and $H-K$ colours are of the same type.

1.3.1 A comment on notation

We shall use the capital letters, as in $X, Y, \ldots$, to denote the *names* of random variables and $x, y, \ldots$ to denote the *values* they take. We thus speak of the probability $\mathbb{P}[X \leq x]$ that the random variable X takes on a value less than or equal to x.

Categoric (nominal) data can take on values like { "Tokyo, "London", "Paris", "Sydney"} which have no immediate sense of order. Boolean data values are values that can be logically combined and generally take on values like "true", "false", or "black", "white". Quantitative data can take on values from sets such as

Type	**Symbol**	**Values**
natural numbers	$\mathbb{N}$	$1, 2, 3, 4, \ldots$
integers	$\mathbb{Z}$	$\ldots, -2, -1, 0, 1, 2, \ldots$
Real numbers 1-dimension	$\mathbb{R}$	$(-\infty, +\infty)$, $[0, \infty)$
Real numbers d-dimensions	$\mathbb{R}^d$	$[0, 1]^d$
Complex real numbers	$\mathbb{C}$	$a + ib$ with $a, b \in \mathbb{R}$

and so on.

Quantitative random variables may take on values that are scalars, $x, y, \ldots$, vectors, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ or tensors of various dimensions, $\{a_{ij}\}$. By *vector* and *tensor* it is meant

[18] Examples of important exceptions to this are the O,A,B,F, ... classification of stellar spectra, the classification of supernova types, and the elements of the Hubble sequence for galaxy classification. The spectral types, however, do map onto a temperature scale.

that transformations among the components of the vectors and tensors are allowed and meaningful: the components of vectors and tensor-valued random variables are in the same space. We use the notation $X, Y, \ldots$ to denote a random variable taking on scalar values, $\mathbf{X}, \mathbf{Y}, \ldots$ for vectors and $\mathbf{X}, \mathbf{Y}, \ldots$ for tensors.

The fact that two random variables X and Y take on scalar values does not mean that they can be regarded as a vector simply by joining them up as a pair $Z = \{X, Y\}$. To do that requires that they are part of the same space and transformations between them, such a $U = X + Y, V = X - Y$ are meaningful. Consider, for example, the maximum rotation speed of a galaxy V and its luminosity L. It certainly makes sense to plot V as a function of L, but we would not consider the random variable $Z = (V, L)$ as a vector in the normal sense. There is a guiding principle here: components of a vector valued random variable will have the same physical dimensions. Ideally, one should consider random variables that are non-dimensional quantities derived from the measured quantities [19].

1.3.2 The distribution function: single variable

From here on we shall, unless otherwise stated, focus on the situation where random variable X takes on ordinal numerical values. The set of values can be continuous, discrete or both. Whether the set of values is continuous or discrete we can define the cumulative distribution function of the random variable as follows. A univariate *cumulative distribution function* (c.d.f.), or more simply a *distribution function*, is a real valued function $F(x)$ if

((a)) $F(x)$ is non-decreasing:

$$F(x_1) \leq F(x_2) \quad \text{for} \quad x_1 \leq x_2 \tag{1.23}$$

((b)) $F(x)$ is everywhere right-continuous

$$F(x) = \lim_{\epsilon \to 0+} F(x + \epsilon) \tag{1.24}$$

((c))

$$F(-\infty) = 0, \quad F(\infty) = 1. \tag{1.25}$$

When these conditions are satisfied, we can interpret $F(x)$ as the probability that the value taken by a random variable X is smaller than x:

$$\mathbb{P}[X \leq x] = F(x) \tag{1.26}$$

In this case we say that $F(x)$ is the cumulative distribution function (c.d.f.) of X, or more simply, the *distribution function*.

The set of values taken by the variable X is generally either continuous or discrete, but it can also be a mixture of both, in which case its distribution function is a combination of continuous monotonically increasing pieces and jump discontinuities. We shall be dealing

[19] There is another possibly more important point of guidance to make here. When fitting a theory to data, map the theory on to the data - do not map the data on to the theory. It is generally bad practice to transform data since that also transforms the errors.

mainly with continuous distributions, but the translation of equations between the two cases is relatively straightforward.

We need one final definition

Definition 1.12
Realisation and Sample:

- A realisation of a random variable X is an instance, x of the variable having distribution function $F(x)$.
- A sample of a random variable X is a set of realisations $\{x\}$, all having the same distribution $F(X)$.

So, if you make one observation of a random process X, that is a sample of size one. Statisticians talk of analysing a data sample in order to assess the statistical properties of that data sample. Such samples are presumed to be repeated realisations of the same underlying physical process.

One of the special aspects of cosmological data is that we have only one Universe, and so we have to deduce everything from a single realisation. Moreover, unlike the situation in most branches of the physical sciences, we cannot experiment with the object of our study except by trying to simulate it. Hence simulations of the Universe are a central part of cosmology.

1.3.3 Continuous distributions

Continuous distributions refer to distribution functions that are defined on the real line. They are characterised by continuity as specified in equation (1.24). In general distributions functions are everywhere continuous in this sense, but there can be points where it is not differentiable. However, almost all practical continuous distributions are differentiable everywhere.

Wherever $F(x)$ is differentiable there exists a function $f(x)$ such that

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(t)\, dt \tag{1.27}$$

The function $f(x)$ is called the *probability density function* (*p.d.f.*) or *frequency function*. The domain of the random variable X is defined by the set of points where $f(x) > 0$ [20].

1.3.4 Discrete distributions

Discrete distributions are characterised by the random variable X being defined on an enumerable set of points $\{\ldots, x_{-1}, x_0, x_1, \ldots\}$. At each point x_s a non-zero point-wise proba-

[20] When we talk of a "statistical distribution" we commonly think in terms of the probability density function $f(x)$ rather than the distribution function $F(x)$. So the commonly used phrase "the distribution is bell-shaped" refers to $f(x)$, not the distribution function.

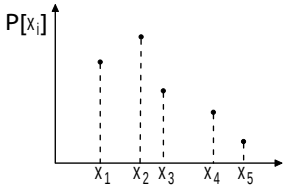


Fig. 1.2 Discrete probability mass function.

bility $p(s)$ is defined such that

$$p(s) = \mathbb{P}[X = x_s] \geq 0, \quad \sum_s p(s) = 1. \tag{1.28}$$

With this, the distribution function $F(x)$ can be written as

$$F(x) = \mathbb{P}[X \leq x] = \sum_{x_s \leq x} p(s) \tag{1.29}$$

In this discrete case the function $p(s)$ is often referred to as the *probability mass function* of the distribution and abbreviated as "*p.m.f.*". This corresponds to the "*p.d.f.*" in the case of a continuous random variable.

1.3.5 Moving between discrete and continuous

There is a strong correspondence between statements about discrete random variables and statements about continuous random variables. To avoid the need to cast results in both forms, we can note a method of translating from the continuous to the discrete. For example, the *distribution function* of a random variable X, denoted by $F(x)$, is the function that gives the probability of X being less than or equal to some value x:

$$F(x) = P(X \leq x) = \sum_{s \leq x} f(s) \quad \text{or} \quad \int_{-\infty}^{x} f(s)ds \tag{1.30}$$

the choice of the sum or integral depending whether the values, x, taken on by the random variable X are discrete or continuous. It would be cumbersome to write all results in this way.

We can unify the discrete and continuous cases by associating a probability density function $f(x)$ with a discrete probability distribution p_i that is specified at points $\{x_i\}$ by writing

$$f(x) = \sum_{i=1}^{n} p_i \, \delta(x - x_i) \tag{1.31}$$

where $\delta(x)$ is the *Dirac delta function* [21] which is loosely defined by

$$\delta(x) = 0 \text{ for } x \neq 0 \text{ and } \int_0^{\infty} \delta(x)dx = 1 \tag{1.32}$$

[21] The function $\delta(x)$ is not an ordinary mathematical function, technically it is a singular generalised function (Lighthill, 1958). Mathematically it is only meaningful when integrated against sufficiently smooth functions.

The rule for using the δ-function is

$$\int_{-\infty}^{\infty} f(x)\delta(x-a)dx = f(a) \tag{1.33}$$

So the *p.d.f.* (1.31) is seen to consist of a weighted sum of δ-function spikes of "strength" p_i. Hence

$$F(x) = \int_{-\infty}^{x} f(s)ds = \sum_{i=1}^{\infty} \int_{-\infty}^{x} p_i\delta(x-x_i)dx = \sum_{i=1}^{i:x\leq x_i} p_i \tag{1.34}$$

since the δ-function picks out the p_i at points $x < x_i$.

Ex 1.3.1 Show that for the *p.d.f.* (1.31)

$$\int_{-\infty}^{\infty} x^k f(x)dx = \sum_{1}^{n} p_i x_i^k \tag{1.35}$$

The first derivative, $\delta'(x)$, of the delta function can be manipulated in the standard way using integration by parts:

$$\int_{-\infty}^{\infty} \delta'(x)f(x)dx = \delta(x)f(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \delta(x)f'(x)dx = -f'(0) \tag{1.36}$$

Thus integrating a function $f(x)$ with the first derivative of a delta function picks out the first derivative of $f(x)$ evaluated at the zero of the delta function argument. If we permit ourselves to think about the n^{th} derivative of the delta function, $\delta^{(n)}(x)$, we have the following important result:

$$\int_{-\infty}^{\infty} \delta^{(n)}(x-x_0)f(x)dx = (-1)^n f^{(n)}(x_0), \quad n \geq 0. \tag{1.37}$$

This is "proved" using repeated integration by parts.

Another specific example of (1.33)which will be of considerable use later is the integral

$$\int_{-\infty}^{\infty} \delta(x)\, e^{i\omega x}dx = 1 \tag{1.38}$$

where again the δ-function is picking out the value of the integrand at $x = 0$.

Related to the Delta function is the Heaviside step function:

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \tag{1.39}$$

which can informally be thought of as

$$H(x) = \int_{-\infty}^{x} \delta(s)\, ds \tag{1.40}$$

1.4 Bivariate distributions

By analogy with the preceding section we can define distributions functions for two or more random variables taken together. In the case of two random variables X and Y that take on values x and y respectively we can define their *joint* distribution function $F_{XY}(x, y)$ as

$$F_{XY}(u, v) = \mathbb{P}[x < u, y < v] \tag{1.41}$$

Note that, in general, the random variables X and Y could describe elements in the data set that are of different data types. X and Y might even have different measurement scales. For example, X might represent measurements of the diameter of a galaxy, while Y might represent its luminosity. A plot of the measured values of Y against he measured values of Y would yield a *scatter plot*. On the other hand Y could represent the spectral type of a star according to the (O, B, A, F, ...) classification scheme (this is an ordinal scale), and X might be the star's age or mass (a ratio scale). We might happily fit a curve to the first scatter plot of luminosity versus diameter, but to do that for the second plot would not make much sense unless we calibrated O, A, B, ... as a temperature (an interval scale).

If X and Y are of the same data type we may sometimes allow transformations of the xy-plane on which the sample values of X and Y are plotted. Thus if X is the $B - V$ colour of a galaxy and Y is the $H - K$ colour, it might make sense to rotate and rescale the axes on which the data is plotted to define a new colour system.

As with a distribution function of a single random variable, we can define the probability density function corresponding to $F_{XY}(x, y)$ by

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \, \partial y} \tag{1.42}$$

whenever this derivative exists. This is called the *joint density* of X and Y.

1.4.1 Marginal density

The *marginal densities* of X and Y are defined as

Definition 1.13
Marginal density:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx \tag{1.43}$$

The notation asserts that these integrals are the probability densities of the random variable X and Y.

The distribution functions corresponding to the densities $f_X(x)$ and $f_Y(y)$ are

$$F_X(x) = F_{XY}(x, \infty), \quad F_Y(y) = F_{XY}(\infty, y) \tag{1.44}$$

We can reverse the process and generate a two point distribution function $f_{XY}(x,y)$ from the distributions $f_X(x)$ and $f_Y(y)$ of random variables X and Y. This is achieved by the function

$$f_{XY}(x,y) = f_X(x)f_Y(y)\left(1+\rho\left[2F_X(x)-1\right]\left[2F_Y(y)-1\right]\right), \quad |\rho|<1 \tag{1.45}$$

where ρ is a constant which we will show later is the *covariance* of the random variables X and Y. Here $F_X(x)$ is the distribution function of X that gives $f_X(x)$ and likewise $F_Y(y)$ for Y. Of course this reconstruction is by no means the only function $f_{XY}(x,y)$ which has $f_X(x)$ and $f_Y(y)$ as its marginal distributions[22].

Example: Two exponentials?

Ex 1.4.1 Show that $f_{XY}(x,y)$ in equation (1.45) has $f_{XY}(x,y) \geq 0$ for all x,y.
Ex 1.4.2 Show that $\int\int f_{XY}(x,y)dxdy = 1$
Ex 1.4.3 Show that $f_X(x)$ is a marginal density of the probability density (1.45).
Hint: using the definition (1.43) of the marginal density, do the integral of $f_{XY}(x,y)$ with respect to y and show that the result is $f_X(x)$. You may need to use the substitution $u = 2F_Y(y) - 1$.

1.4.2 Statistical independence

Two random variables are said to be *independent* if their joint probability density function is simply the product of the individual probability density functions:

$$f_{XY}(x,y) = f_X(x)f_Y(y) \tag{1.46}$$

and it can be shown that

$$f_{XY}(x,y) = f_X(x)f_Y(y) \quad \text{if and only if} \quad F_{XY}(x,y) = F_X(x)F_Y(y) \tag{1.47}$$

Note that this "independent" here means "statistically independent". As we shall see later there is a distinction to be drawn bewteen variables that are functionally dependent and variables that are statistically dependent.

1.4.3 Conditional densities

Consider two random variables X and Y having probability functions $f_X(x)$ and $f_Y(y)$. We put subscripts on the functions to emphasize that the two distributions may have different

[22] There are several ways of creating a joint distribution of two random variables whose marginal distributions are known. This is one example, but the most general way is via the mechanism of, what in statistics is somewhat curiously, named a *copula*. Copulas are distribution functions on the unit interval that provide a way of modelling correlated distributions when only their marginal distributions are known. See section **??**.

functional forms. Let their joint probability function be $f_{XY}(x,y)$. The *conditional probability function* of X given Y, denoted by $f_{X|Y}(x|y)$ is defined as

$$f_{X|Y}(x|y) = \mathbb{P}[X = x, Y = y] \tag{1.48}$$

From equation (1.6) we have

$$f_{X|Y}(x|y) = \mathbb{P}[(X = x\ Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} \tag{1.49}$$

which gives the important equation

Conditional probability density function

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{1.50}$$

expressing the probability function for the individual random variable X, given the probability function of Y.

1.5 Multi-variate distributions

This discussion generalises to functions of n random variables $X_1, X_2, \ldots, X_n$ that take on values $x_1, x_2, \ldots, x_n$. We define

Multi-variate distribution function

$$F_n(u_1, u_2, \ldots, u_n) = \mathbb{P}[x_1 < u_1, x_2 < u_2, \ldots, x_n < u_n] \tag{1.51}$$

The notation here avoids writing $x_1, \ldots, x_n$ or $u_1, \ldots, u_n$ as a vector since the corresponding random variables $X_1, \ldots, X_n$ may have different data types. Because of the definition of probability we have

$$0 \leq F_n(u_1, u_2, \ldots, u_n) \leq 1 \tag{1.52}$$

The left inequality is achieved whenever any of the $u_i \to -\infty$. The right inequality is achieved only when all the $u_i \to +\infty$.

The function $F_n(u_1, u_2, \ldots, u_n)$ is also right continuous in each of the variables:

$$F_n(u_1, \ldots u_i, \ldots, u_n) = \lim_{\epsilon \to 0+} F_n(u_1, \ldots u_i + \epsilon, \ldots, u_n), \quad \text{for each } i \tag{1.53}$$

Moreover, if there exists a function $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ such that

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_2} f_{X_1,\ldots,X_n}(u_1, \ldots, u_n)\, du_1 \ldots du_n \tag{1.54}$$

then $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ is the probability density of the distribution $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

Alternatively we can write

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{\partial x_1 \ldots \partial x_n} \tag{1.55}$$

provided $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ has an appropriate degree of continuity.

n random variables $X_1,\ldots,X_n$ are said to be independent when their joint probability density function factorises into the product of the individual probability densities:

Joint distribution of independent random variables:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1)\ldots f_{X_n}(x_n) \tag{1.56}$$

1.6 Transformations of random variables

Functional transformations of random variables, in which we functionally map a set of n random variables $X_1, X_2, \ldots, X_n$ into another set of m random variables $Z_1, Z_2, \ldots, Z_m$ is of central importance. Not only is it a way of generating new distributions from old, it is also arises in physics experiments where we might measure some function of a set of variables whose distribution is known. As an example of the latter we might study the energy of particles that are moving randomly with velocity components selected from a a normal distribution[23].

The issue that arises when more than on variable is involved in the transformation is the domain of the variable. So whereas the domain of each velocity component is the real line, the energy is always positive. The discussion of Papoulis and Pillai (2002, Ch.6) is highly recommended reading on this subject.

1.6.1 Single variable

First let us deal with a single random variable which has a distribution function

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(s)ds \tag{1.57}$$

This is just equation (1.30). Consider a new random variable Y such that the values y realised by Y are related to x via a monotonic increasing function $y = g(x)$ that has an inverse $y = h(x)$:

$$y = g(x), \quad x = g^{-1}(y) = h(y) \tag{1.58}$$

[23] If the velocity components of a unit mass particle are denoted by (u, v, w) the particle kinetic energy is $E = (u^2 + v^2 + w^2)/2$ and the particle speed is $s = \sqrt{(2E)}$. If the velocity components are normally distributed with zero mean, the distribution of E is the Boltzmann distribution, and the distribution of the s is the Maxwell-Boltzmann distribution.

Then what is the distribution function $\Phi(y)$ of the random variable Y? If $g(x)$ is a monotonic increasing function

$$\Phi(y) = \mathbb{P}[Y \le y] = \mathbb{P}[X \le h(y)] = \int_{-\infty}^{h(y)} f(x)dx = F[h(y)]. \tag{1.59}$$

which gives the relationship between the distribution functions of X and Y. The probability density function $\phi(y)$ of Y is simply the derivative of $\Phi(y)$:

$$\phi(y) = \frac{d\Phi}{dy} = F'(x)\frac{dx}{dy} = f(x)\frac{dh(y)}{dy} \tag{1.60}$$

This is correct under the assumption that $y = g(x)$ is a monotonic increasing function. If it were monotonic decreasing than there would be a sign change in this expression. In the general case, for any monotonic function $g(x)$ which has an inverse $h(y)$ we can write

$$\phi(y) = f[h(y)]\,|h'(y)|, \quad h(y) \text{ monotonic} \tag{1.61}$$

Ex 1.6.1 Show that under the linear transformation $y = ax + b$ the probability density function $\phi(y)$ of y is related to the probability density $f(x)$ of x by

$$\phi(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right) \tag{1.62}$$

1.6.2 One function of two random variables

When we make a monotonic transformation $y = g(x)$ of a random variable as in equation (1.58), there is no difficulty in relating the domain of the new variable y to the domain of the original variable x. That is why the condition of monotonicity is imposed. When we consider a transformation of the form

$$z = g(x, y) \tag{1.63}$$

mapping two random variables X and Y into a single random variable Z, the situation about the domains of X, Y, Z needs careful consideration. We need the 2-variable analogue of equation (1.59), so if $F_Z(z)$ is the distribution function of Z:

$$F_Z(z) = \mathbb{P}[Z \le z] = \mathbb{P}[g(x,y) < z] = \mathbb{P}[(X,Y) \in \mathcal{D}_Z] \tag{1.64}$$

where $\mathcal{D}_Z$ is the domain of the X, Y-plane where $g(x,y) < z$. If $f_{XY}(x,y)$ is the joint distribution of the random variables X and Y, this is

$$F_Z(z) = \int\int_{\mathcal{D}_Z} f_{XY}(x,y)\,dx\,dy \tag{1.65}$$

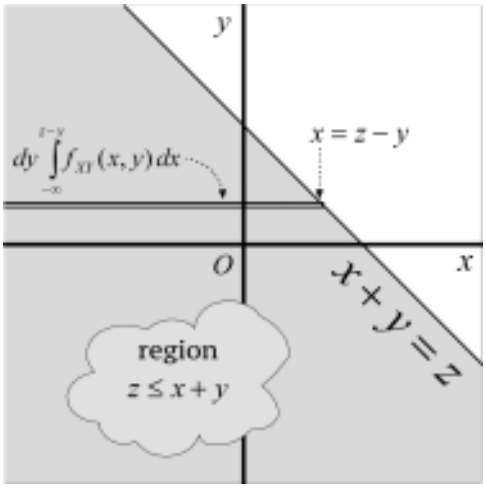


Fig. 1.3 The domain of the random variable $Z = X + Y$: the shaded area is the area where $\mathbb{P}[Z \le z] = \mathbb{P}[X + Y \le Z]$ which defines the domain of the integral in equation (1.66)

Consider a specific example. Let X and Y be two random variables having joint density $f_{XY}(x, y)$, derive the *p.d.f.* of $Z = X + Y$. Since $\mathbb{P}[Z \le z] = \mathbb{P}[X + Y \le Z]$, the distribution function of $F_Z(z)$ or Z is

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_{XY}(x, y)\, dxdy \tag{1.66}$$

The integration limits correspond the the shaded area in figure 1.3.

We can now write an expression for the probability density of Z:

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} \frac{\partial}{\partial z} \left(\int_{-\infty}^{z-y} f_{XY}(x, y)\, dx \right) dy = \int_{-\infty}^{\infty} f_{XY}(z - y, y) dy \tag{1.67}$$

This is as far as we can go without an explicit example of the *p.d.f.* $F_{XY}(x, y)$.

However, if X and Y are independent we have the factorisation $f_{XY}(x, y) = f_X(x)\, f_Y(y)$ and so

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y)\, f_Y(y) dy \tag{1.68}$$

This is just a convolution of the two *p.d.f.* s of X and Y.

1.6.3 Multiple variables

Transformations of two or more variables work in the same way. We can consider the transformation between a pair of random variables X_1, X_2 and Y_1, Y_2 under the monotonic bivariate transformation

$$y_1 = g_1(x_1, x_2) \quad y_2 = g_2(x_1, x_2) \tag{1.69}$$

and its inverse

$$x_1 = h_1(y_1, y_2) \quad x_2 = h_2(y_1, y_2) \tag{1.70}$$

The distribution functions $\Phi(y_1, y_2)$ and $F(x_1, x_2)$ are then related by the transformation

$$\Phi(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} F[h_1(u, v), h_2(u, v)] \left| \frac{\partial(h_1, h_2)}{\partial(u, v)} \right| du\, dv \tag{1.71}$$

where

$$J = \left| \frac{\partial(h_1, h_2)}{\partial(u, v)} \right| = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix} \tag{1.72}$$

is the *Jacobian* of the transformation. The probability density transforms as

$$\phi(y_1, y_2) = f[h_1(y_1, y_2), h_2(y_1, y_2)] \left| \frac{\partial(h_1, h_2)}{\partial(y_1, y_2)} \right| \tag{1.73}$$

This exercise is a bit premature

Ex 1.6.2 Let U and V be independent $N(0, 1)$ distributed random variables Let $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y \in [0, \infty)$ and $\rho \in [0, 1]$ and define new variables X and Y by

$$X = \mu_X + \sigma_X U \tag{1.74}$$
$$Y = \mu_Y + \sigma_Y \rho U + \sigma_Y \sqrt{1 - \rho^2}\, V$$

Show that the joint distribution of X and Y is the bivariate normal distribution

$$f_{X,Y}(x, y\,; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right) \tag{1.75}$$

(see section **??**)

The random variables X and Y created in this example are not independent since Y depends on both U and V. The transformation (1.74) is a simple way of generating correlated random variables.

1.6.4 Products of two random variables

Consider two random variables X and Y and their product

$$Z = XY \tag{1.76}$$

What can we say about the random variable Z?

There are some simple cases where we can make statements about Z. For example,

$$\mathbb{E}[Z] = \mathbb{E}[X]\mathbb{E}[Y], \quad \text{if } X \text{ and } Y \text{ are independent.} \tag{1.77}$$

If X and Y have special mathematical forms, such as exponentials, we can easily make progress. Thus the product of two lognormally distributed random variables is again lognormal. But the product of two normally distributed variables is not normally distributed. We can see this by considering $Z = XY$ with $X = Y$, in which case $Z = X^2$, which is clearly not normally distributed (it is always positive).[24]

When X and Y are independent random variables, then their joint distribution is given by (1.64) with the domain of integration $\mathcal{D}$ being the set of points where $Z < XY$. We can show this as follows using the coordinate transformations of variables $x \to x$ and $y \to z/x$ (see §1.6.3). Given the probability densities of X and Y are $f_X(x)$ and $f_Y(y)$ we can write down the probability $\mathbb{P}[Z < z]$, which is the distribution function of z. We have $\mathbb{P}[Z < z] = \mathbb{P}[XY < z] = \mathbb{P}[Y < z/x]$ for all possible values of $x \in \mathbb{R}$.

To achieve this we use the transformation of variables (x, y) to $(x, z/x)$ for which the Jacobian is

$$J = \left|\frac{\partial(x,y)}{\partial(x,z)}\right| = \begin{vmatrix} \dfrac{\partial x}{\partial x} & \dfrac{\partial x}{\partial z} \\ \dfrac{\partial y}{\partial x} & \dfrac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -z/x^2 & 1/x \end{vmatrix} = \frac{1}{|x|} \tag{1.79}$$

The probability $\mathbb{P}[Y < z/x]$ is then

$$\mathbb{P}[Y > z/x] = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_X(x) f_Y\left(\frac{z}{x}\right)\frac{1}{|x|}dxdz = \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_X(x) f_Y\left(\frac{z}{x}\right)\frac{dx}{|x|}\right]dz \tag{1.80}$$

The last term is just a rearrangement of the order of integration. This expression is just $\mathbb{P}[Z < z]$, *i.e.* the cumulative distribution of the random variable Z. The probability density of Z, $f_Z(z)$ is the derivative of this last expression with respect to z:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y\left(\frac{z}{x}\right)\frac{dx}{|x|} \tag{1.81}$$

1.6.5 Product of two Gaussian random variables

So what is the distribution of a product of two Gaussian random variables X and Y? We note that

$$XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2 \tag{1.82}$$

[24] If $X$and Y are zero mean normal random variables $N(0,\sigma_x)$ and $N(0,\sigma_y)$ respectively, then

$$\mathbb{P}_{XY}[Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_x}e^{-x^2/2\sigma^2}\delta(xu - z)dxdy = \frac{1}{\pi\sigma_x\sigma_y}K_0\left(\frac{|z|}{\sigma_x\sigma_y}\right). \tag{1.78}$$

The function $K_0(x) = \int_0^\infty (x^2+t^2)^{-1/2}\cos t dt$ is a modified Bessel function of the second kind(Gradshteyn and Ryzhik, 2007, 3.754.2). The function $\delta(xy - z)$ is the Dirac delta function and the argument forces the integral to be evaluate on the curve $z = xy$ (see Weinstein, "NormalProductDistribution").

Each of the variables $X+Y$ and $X-Y$ is Gaussian distributed since the sum of two Gaussian random variables is another Gaussian. The square of a Gaussian is Chi-square distributed. So in this last equation $(X+Y)^2$ and $(X-Y)^2$ are both Chi-square distributed with one degree of freedom. The sum of two Chi-squared distributions is also Chi-squared[25], so the distribution of $Z = XY$ is Chi-squared with two degrees of freedom.

1.7 Expectation and moments

Statistical expectation was defined in section 1.2.2 (equation 1.17) in relation to the expressing the expected outcome of an event in terms of the underlying probabilities for the set of possible outcomes. We must now generalise that to apply to random variables and functions of random variables which have a known underlying statistical distribution.

1.7.1 Expectation

The *expected value* of a real valued function $u(X)$ of a random variable X is defined as

Definition 1.14
Expectation:

$$\mathbb{E}[u(X)] = \sum_x u(x) f(x) \quad \text{or} \quad \int_x u(x) f(x) dx \tag{1.83}$$

depending as to whether the random variable X has a discrete *p.m.f.* $p(x)$ or continuous distribution *p.d.f.* $f(x)$ (see section 1.3.5). From here on we shall simply use the case of continuous random variables.

In the physical sciences one often sees the notation

$$\langle u \rangle = \mathbb{E}[u(X)] = \int_x u(x) f(x) dx \tag{1.84}$$

for the statistical expectation of a function $u(x)$ of a random variable X. Here the suffix x on the integral is understood to mean that the integral is taken over the entire domain of the variable x.[26]

The expectation operator is linear in the following sense:

[25] The U and V are Chi-squared with m and n degrees of freedom respectively, then $U+V$ is Chi-squared distributed with $M+n$ degrees of freedom.

[26] In quantum mechanics the expected value of a measurement, $\langle x \rangle$, of a property, X, of a particle is determined by the wave function, $\Psi(x)$, for the property X. $\Psi(x)$ is a complex function whose modulus, $|\Psi^*(x)\Psi(x)|$ is interpreted as a probability density for the property X. In other words, the expected result of the measurement is

$$\langle x \rangle = \frac{\int_{-\infty}^{\infty} \Psi^*(x) x \Psi(x) dx}{\int_{-\infty}^{\infty} \Psi^*(x) \Psi(x) dx} \tag{1.85}$$

The wave function is a solution of the Schrödinger equation for the particle position.

Ex 1.7.1 Show that

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{1.86}$$

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X] \tag{1.87}$$

We can define *conditional expectation* $\mathbb{E}[X \mid Y]$ of random variables X and Y as

Definition 1.15
Conditional expectation:

$$\mathbb{E}[X \mid Y] = \int_x x f(x \mid y) dx \tag{1.88}$$

where $f(x \mid y)$ is defined in equation (1.50).

In other words, this is the conditional *p.d.f.* (equation 1.50) taken at a particular value $Y = y$ of the random variable Y.

Ex 1.7.2 If X and Y are independent and identically distributed random variables, show that

$$\mathbb{E}[X \mid X + Y = k] = \tfrac{1}{2}k \tag{1.89}$$

Hint: exploit the symmetry in X and Y.

1.7.2 Statistical Moments

The *moments* of the distribution of a random variable X are defined as the expectation of X^k:

$$\mu_k = \mathbb{E}[X^k], k = 1, 2, \ldots \tag{1.90}$$

μ_k is referred to as the k^{th}-order moment of X.

We have two important definitions: *orthogonal* random variables and *uncorrelated* random variables:

Definition 1.16

$$X, Y \text{ orthogonal}: \quad \mathbb{E}[XY] = 0 \tag{1.91}$$
$$X, Y \text{ uncorrelated}: \quad \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{1.92}$$

We shall have occasion to exploit orthogonality later.

1.8 Mean and Variance

sketch required

It is of value to be able to parametrise probability densities. Doing so encapsulates the broad features of data that are being fitted with statistical models. Determining these parameters is a primary goal of statistical inference.

The basic parameters of any statistical distribution are the location and spread of the data values, with refinements describing the shape. The fundamental statistical measures of location and spread are the first and second moments: the *mean* and *variance* of the data values.

Different models of the same data, *i.e.* different assumed distribution functions, will inevitably result in different estimates for the mean and variance. Bringing in further parameters, skewness, kurtosis and so on, will better define the distribution.

Several questions now arise. Is the chosen model the best choice? Given the choice, what is the certainty of the determination of the parameter values?

1.8.1 Expectation as a mean value

A particularly important function $u(X)$ is $u(X) = X$. The expectation of X is called the *mean*[27] of the distribution of X:

Definition 1.17
Arithmetic Mean:

$$\mu = \mathbb{E}[X] = \sum_x x\, f(x) \quad \text{or} \quad \int_x x\, f(x)dx \tag{1.93}$$

depending as usual on whether X is a discrete or continuous variable. This is a measure of the location of the distribution, but there are others like the *mode*, the location of the maximum of $f(x)$ which is used in Likelihood estimation (see Chapter **??**).

[27] $\mathbb{E}[X]$ is also referred to as the *population mean* to distinguish it from the *sample mean* which is an estimate derived from measurements, or realisations, of X.

1.8.2 Other sample means

Other measures of location are *median*, the value of x that divides the area under $f(x)$ into two equal halves. For a data sample comprising n items $\{x_i\}$ we have the following:

Arithmetic mean	$A = \frac{1}{n}\sum_{i=1}^{n} x_i$
Mode	$M = \max\{x_i\}$
Median	$Med = x_{[(n+1)/2]}$, where $[a]$ denotes the integer part of a
Geometric mean	$G = (\prod_{i=1}^{n} x_i)^{1/n}$
Harmonic mean	$H = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$, $x_i \neq 0$. Note that $H = G^2/A$.

The relationship $A \geq G \geq H$ holds for any distribution where these quantities are well defined. Which mean to use depends on the nature of the underlying problem [28] [29]. The median is stable against outliers (anomalous data values) among the $\{x_i\}$ and used in many aspects of non-parametric statistical inference. The geometric mean is most relevant when the underlying statistical process is multiplicative, as in population dynamics or when calculating rates of interest. The Harmonic mean finds a use in astrophysics when calculating the potential energy of a self-gravitating distribution of stars.

1.8.3 Variance

Another important function is the *variance* of X. If X has mean μ and probability function $f(x)$, the variance, denoted by σ^2 or $\mathrm{Var}(X)$ is defined as

Definition 1.18
Variance:

$$\sigma^2 = \mathrm{Var}\,(X) = \mathbb{E}\,[(X-\mu)^2] = \int (x-\mu)^2 f(x)\,dx \qquad (1.94)$$
$$= \mathbb{E}\,[X^2] - \mu^2$$

where the last equality in (1.94) follows from using equation (1.83). Since $\mu = \mathbb{E}\,[X]$ this can also be written

$$\mathrm{Var}(X) = \mathbb{E}\,[X^2] - \mathbb{E}\,[X]^2 \qquad (1.95)$$

[28] The word *average* is frequently used in the less scientific articles without specifying which mean this represents. This can give rise to misleading the reader as in "despite the depression, the average salary is rising" when referring to 'mean salaries' rather than 'the most common salary', i.e. the mode of the distribution

[29] As a nice diversion one might note that the Brent-Salamin algorithm for the computation of π involves arithmetic and geometric means of two number sequences, $\{a_i\}, \{b_i\}, i = 0, 1, 2, \ldots$ that start with $a_0 = 1, b_0 = 1/\sqrt{2}$. Successive iterations are $a_{n+1} = (a_n + b_n)/2$, $b_{n+1} = \sqrt{a_n b_n}$. The number

$$c_n = 4a_n b_n \left(1 - \sum_{i=1}^{n} 2^{i+1}(a_i^2 - b_i^2))\right)^{-1} \to \pi, \quad n \to \infty$$

This roughly doubles the number of decimal places at each step, yielding 18 correct places after three iterations. This scheme was used in 1999 to get over 2×10^{11} places of π. However, subsequent computations used other formulae, several coming out of the work of the Indian mathematician S. Ramanujan.

If the mean is zero the variance is just the expectation of the square of the random variable. The square root of the variance, σ, is called the *standard deviation* of the distribution.

The mean describes location of the variables and the variance describes their spread about this mean. Note that not all distributions $f(x)$ have mean or variance, the integrals (1.93) and (1.95) must exist. An example of a distribution where neither the mean nor the variance are defined is the Cauchy distribution (section **??**).

Ex 1.8.1 Show that for random variables X and Y

$$\mathrm{Var}\,(Y\,|\,X\,) = \mathbb{E}\,[Y^2\,|\,X\,] - \mathbb{E}\,[Y\,|\,X\,]^2 \tag{1.96}$$

It should be noted that if a sequence of measurements $\{X_i\}, i = 1\ldots, n$ is drawn from random variable X whose distribution has finite non-zero variance σ^2 then the distribution of random variable S_n that is the sum of the X_i

$$S_n = x_1 + \cdots + X_n, \tag{1.97}$$

becomes the normal distribution in the limit of large n. If the distribution of X has mean μ and variance σ^2 we can put this more formally as

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right] = N(0,1) \tag{1.98}$$

where the symbol $N(0,1)$ denotes the normal distribution having zero mean and unit variance. We shall come back to this later.

1.8.4 Standardised random variables

From a random variable X having mean μ_X and variance σ_X we can create a new "standardised" random variable Z having zero mean $\mu(Z)$ and unit variance $\sigma(Z)$ by means of the linear transformation

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{1.99}$$

Ex 1.8.2 Show that

$$\mu(Z) = \mathbb{E}\,[Z] = 0 \quad \text{and} \quad \sigma(Z) = \mathbb{E}\,[X^2] = 1 \tag{1.100}$$

1.9 Biased estimators

1.9.1 Sampling and sample bias

Data analysis is largely a process of acquiring data and fitting a statistical model, possibly based on some theoretical prejudice, for the data. The model will be parametrised and the data is used to estimate those parameter values. The estimate is only as good as the data and is influenced by the errors that might occur in the data acquisition process.

If we make an estimate $\hat{\theta}$ of the value of a parameter θ from a data sample, the estimate itself is a statistical quantity that will vary from sample to sample. So we can talk about the expectation value of this estimate, $\mathbb{E}[\hat{\theta}]$. The bias of this estimate is formally defined as

Definition 1.19
Bias of a sample estimator:

$$b = \mathbb{E}[\hat{\theta}] - \theta. \tag{1.101}$$

An unbiased estimator has $b = 0$

Ideally we would like a procedure for estimating θ that has $b = 0$.

1.9.2 Sample mean and variance

Given n observations $x_1, \ldots, x_n$ of a random variable X the estimate of the mean $\mathbb{E}[X]$ of the distribution of X is simply

$$\bar{x}_{(n)} = \frac{1}{n}(x_1 + \cdots + x_n) \tag{1.102}$$

This is a forward reference

The central limit theorem (section 1.14.2) tells us that the distribution of this estimate tends to a Gaussian, regardless of the underlying distribution of the random variable X.

Ex 1.9.1 Show that the $\bar{x}_{(n)}$ (equation 1.102) is an unbiased estimator of the population mean $\mu = \mathbb{E}[X]$. i.e.

$$\mathbb{E}[\bar{x}_{(n)}] = \mathbb{E}[X] \tag{1.103}$$

Ex 1.9.2 Show that the population variance of the sample mean $\bar{x}_{(n)}$ is

$$\mathrm{Var}(\bar{x}_{(n)}) = \frac{\sigma^2}{n} \tag{1.104}$$

where $\sigma = \mathrm{Var}(X)$. This result is biased.

The variance of X can be estimated from the observations $x_1, \ldots, x_n$:

$$s^2_{(n)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{(n)})^2 \tag{1.105}$$

s is called the *sample standard deviation* of the observations $x_1, \ldots, x_n$. The reason for dividing the sum of the squares by $n-1$ rather than n is that $s^2_{(n)}$ defined in this way is an *unbiased* estimate (see equation 1.101) of the variance σ^2.

Ex 1.9.3 Show that

$$\mathbb{E}[x_i^2] = \sigma^2 + \mu^2 \tag{1.106}$$

Ex 1.9.4 Show that the population variance (1.9.2) has expectation

$$\mathbb{E}[\bar{x}^2] = \frac{1}{n}\sigma^2 + \mu^2 \tag{1.107}$$

Hint: it might be useful to start with the definition of $\mathrm{Var}(\bar{x})$.

Ex 1.9.5 Show that $s^2_{(n)}$ (equation 1.105) is an unbiased estimator of the population variance. i.e.

$$\mathbb{E}[s^2_{(n)}] = \mathrm{Var}(X) = \sigma^2 \tag{1.108}$$

Caution: this is tricky.

1.10 Skewness and Kurtosis

The mean (1.93) and variance (1.95) are not in general sufficient to characterise the shape of a probability distribution, the notable exception being the Gaussian distribution (see Section **??**). Going to moments of higher order than the mean and variance we come to the *skewness*, which indicates how asymmetric the distribution is about the mean, and the *kurtosis*, which describes how 'peaky' the distribution is.

The *skewness* γ_1 of the distribution of a random variable X is defined as

Definition 1.20
Skewness:

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}[X-\mu)^3]}{(\mathbb{E}[X^2-\mu^2])^{3/2}} \tag{1.109}$$

The *kurtosis* of the distribution is defined as

Definition 1.21
Kurtosis:

$$\gamma_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 = \frac{\mathbb{E}\left[X-\mu\right)^4]}{(\mathbb{E}\left[X^2-\mu^2\right])^2} - 3 \tag{1.110}$$

An alternative definition of the *kurtosis* of a distribution is

$$\mathrm{Kurt}(X) = \mathbb{E}[X^4] - 3(\mathbb{E}[X^2])^2 \tag{1.111}$$

This is also referred to as the *fourth-order cumulant* of the distribution. For a Gaussian random variable this is zero, and it is non-zero for most, but not all, non-Gaussian distributions. Hence this is often used as a measure of how non-Gaussian the distribution of a random variable is.

Ex 1.10.1 Show that if X_1 and X_2 are two independently distributed random variables, then

$$\mathrm{Kurt}(X_1 + X_2) = \mathrm{Kurt}(X_1) + \mathrm{Kurt}(X_2) \tag{1.112}$$

Ex 1.10.2 Show further that

$$\mathrm{Kurt}(\alpha X_1) = \alpha^4\,\mathrm{Kurt}(X_1). \tag{1.113}$$

1.11 Covariance and independence

1.11.1 Formal definitions

Consider two random variables X and Y having means μ_X and μ_Y respectively. Then we define the *covariance* of X and Y to be

Definition 1.22
Covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}\left[XY\right] - \mu_X\mu_Y \tag{1.114}$$

This can be expressed in terms of the probabilities $\mathbb{P}[X]$, $\mathbb{P}[Y]$ that the events X and Y occur[30]:

$$\mathrm{Cov}(X,Y) = \mathbb{P}[XY] - \mathbb{P}[X]\,\mathbb{P}[Y] \tag{1.115}$$

Note that if $X = Y$ then the covariance is simply the variance: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$. It is easy to show the following properties of the covariance:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) \tag{1.116}$$

$$\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y) \tag{1.117}$$

$$\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z) \tag{1.118}$$

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) \tag{1.119}$$

We have the following results for independent random variables:

$$X, Y \text{ independent } \Rightarrow\ \mathrm{Cov}(X,Y) = 0 \tag{1.120}$$

$$X, Y \text{ independent } \Rightarrow\ \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{1.121}$$

$$X, Y \text{ independent and } Y, Z \text{ independent } \not\Rightarrow\ X, Z \text{ independent} \tag{1.122}$$

This last result says that being independent is not a transitive (inherited) relationship. For example, if the three random variables X, Y, Z have joint distribution function that can be written in the form $H(x,y,z) = F(y)G(x,z)$, then Y is manifestly independent of X and Z as per equation (1.122) but X and Z are dependent through their arbitrary joint distribution function $G(x,z)$.

We can derive an important expression for the variance of the sum of a set of random variables. Take $Y(X_1,\ldots,X_n) = X_1 + \cdots + X_n$. Recalling the definition

$$\mathrm{Var}(Y) = \mathbb{E}[(Y-\mu_Y)^2] \tag{1.123}$$

and going through the algebra yields

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_j \sum_{i:\, i\neq j} \mathrm{Cov}(X_i, X_j) \tag{1.124}$$

Note that, because $\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_j,X_i)$, the terms in the double sum of this equation are counted twice. So it can be written as $2\,\boldsymbol{\Sigma}_j\,\boldsymbol{\Sigma}_{i<j}\mathrm{Cov}(X_i,X_j)$.

1.11.2 Mutually independent random variables

If the random variables $X_1,\ldots,X_n$ are mutually independent then

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} \mathrm{Var}(X_i) \quad \text{if} \quad \mathrm{Cov}(X_i,X_j) = 0 \quad \forall\ i \neq j \tag{1.125}$$

If two random variables are independent, their covariance is zero, but the converse is not true.

[30] As before, we use the notation XY to denote the event that both X and Y happen. The alternate notation for XY is $X \cap Y$.

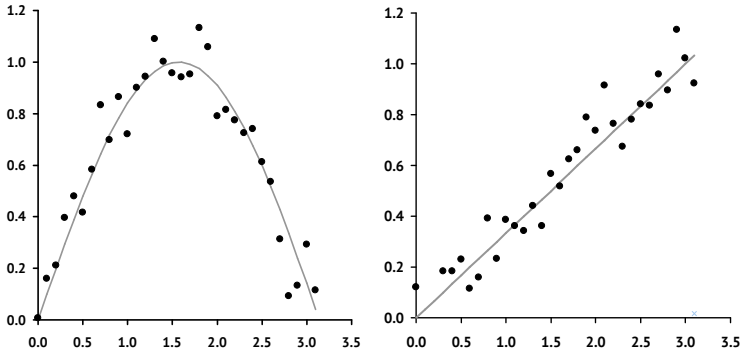


Fig. 1.4 Covariance and correlation are only indicative of a relationship when the data is linear. On the left, the points (x_i, y_i) show data randomly sampled from the function $y = \sin x$ on the interval $0 \leq x \leq \pi$. The correlation coefficient of the set of points $\{(x_i, y_i)\}$ is $\rho_{points} = -0.024$: the data appears only weakly correlated on the basis of ρ. For comparison, the correlation between the *curve* $y = sinx$ and the points is $\rho_{curve} = 0.959$, which is a measure of the goodness of the fit. On the right the data is sampled from a straight line giving a correlation of $\rho = 0.953$.

Ex 1.11.1 Show that the covariance of two independent random variables is zero. Hint: go back to the definition (1.5) of independence.

Ex 1.11.2 Show the result (1.124) and deduce (1.125) for independent random variables.

1.11.3 Zero covariance $\not\Rightarrow$ independence

Zero covariance does not imply independence: this can be seen from the following example. Let X be a Gaussian distributed random variable. Suppose that another random variable, Y is related to X via $Y = X^2$. Then given X you can calculate Y precisely: they are certainly not independent. However, if we compute the covariance of X and Y we find from equation (1.114) that

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X^3] \quad \text{since } \mathbb{E}[X] = 0 \tag{1.126}$$

$$= 0 \quad \text{if the distribution of } X \text{ has zero skewness} \tag{1.127}$$

(We have $\mathbb{E}[X^3] = 0$ since the Gaussian distribution is symmetric and has zero skewness). So here we have two functionally related variables $Y = X^2$ whose covariance is zero. Another example is shown in figure 1.4.

Let X and Y be independent random variables, both uniformly distributed on the interval $[0, 1]$ (i.e. they are both $U(0, 1)$). This exercise shows that the random variables $U = X + Y$ and $V = X - Y$ are functionally related, but that they are uncorrelated:

Ex 1.11.3 Derive the joint probability density distribution $f_{UV}(u, v)$
Ex 1.11.4 Derive the probability density functions of $f_U(u)$ of U and $f_V(v)$ of V.
Ex 1.11.5 Show that $f_{UV}(u, v) \neq f_U(u)\, f_V(v)$ and say what this implies
Ex 1.11.6 Calculate $\mathbb{E}[U, V]$ and $\mathrm{Cov}(U, V)$ and say what this means

Two random variables may be *functionally related* but nonetheless *statistically independent*. See Broffitt (1986) for two further simple examples.

The converse question arises in many areas of research: if two random variables appear to have zero covariance, how could we know whether or not they are independent? The answer lies in part in an important development: the "copula" (Sklar, 1973, the original work was reported in 1959). The copula is a mathematical function that completely describes the joint distribution of a set of n random variables in terms of their 1-dimensional marginal distributions, $F_1, \ldots, F_n$ and a function of these $C(F_1(x_1), \ldots, F_n(x_n)$ that is defined on the unit n-cube. This is covered in more detail in section **??**.

1.11.4 Sample covariance of paired values

The covariance $\mathrm{Cov}(X, Y)$ defined in equation (1.114) is also referred to as the *population covariance* in order to distinguish it from the *sample covariance* determined from n paired observations (x_i, y_i) of X and Y:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - s_X)(y_i - s_Y) \tag{1.128}$$

where s_X and s_Y are the sample means of X and Y:

$$s_X = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s_Y = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1.129}$$

The factor $(n - 1)$ appears in (1.128) so as to make s_{XY} an unbiased estimator of the population covariance.

Ex 1.11.7 Show that

$$\mathbb{E}[s_{XY}] = \frac{n}{n-1} \mathrm{Cov}(X, Y). \tag{1.130}$$

The quantity s_{XY} has played, and continues to play, a key role in statistical analysis of data.

We should recall the relationship $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$ from equations (1.119) and note that the value of s_{XY} depends on the scale which is used to quote the values of the instances (x_i, y_i) of the random variables X and Y. A change in the units of measurement causes the value of s_{XY} to change.

If the variables X and Y are standardised relative to their mean and variance, as in equation (1.99) of section 1.8.4, the corresponding covariance of the standardised variables becomes scale independent[31]:

$$\mathrm{Cor}(X,Y) \equiv \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}, \quad |\rho| \leq 1 \tag{1.131}$$

The covariance function of normalised random variables is called the *Correlation function*, and is denoted by the symbol $\mathrm{Cor}(X,Y)$. It is alternatively denoted by the symbol ρ_{XY}, in which case it is referred to as the *Pearson Correlation Coefficient* (see also section 1.13).

1.12 The Covariance Matrix

In what follows we shall be dealing with vectors and matrices since they provide a compact notation for handling multiple sets of random variables. We will represent an $n-$dimensional *vector* $\mathbf{v}$ having n entries as an $n \times 1$ column matrix having n entries. The transpose of this $n \times 1$ matrix to a $1 \times n$ row matrix is denoted by $\mathbf{v}^T$. The ^{T} superscript in denotes the transpose of the column vector $\mathbf{v}$. Note that $\mathbf{x}\mathbf{y}^T$ is a matrix while $\mathbf{x}^T\mathbf{y}$ only makes sense if the dimensionality of $\mathbf{x}$ and $\mathbf{y}$ are the same, and $\mathbf{x}^T\mathbf{y}$ is then a scalar. The order of multiplication is important.

A set of random variables $\{X_i\}, i = 1, \ldots, n$ will be denoted by the $n-$dimensional random vector $\mathbf{X}$. Realisations, measurements or samples of these random variables will be denoted by $\mathbf{x}$. Matrices will be written in a sans-serif font, as in M, unless the symbol is a Greek letter. The determinant of a matrix M will be denoted by $|\mathsf{M}|$.

1.12.1 Covariance between two random variables

Recall the definition (1.114) of the covariance between two random variables X and Y:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \tag{1.132}$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{1.133}$$

The covariance between two random variables X and Y is a number. It can usefully be normalised relative to the standard deviations of the variables X and Y.

The covariances can between X and Y and themselves be conveniently arranged in the 2×2 *covariance matrix*

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{pmatrix} \tag{1.134}$$

[31] Of course if the scaling factor of one of the variables X and Y is negative, the sign of the covariance changes but the value does not. This is seen from the second of equations (1.119).

Equation (1.133) provides an efficient way of computing covariances.

The normalised covariance is referred to as the *correlation* between X and Y (see section 1.13).

It is worth mentioning some issues that arise when thinking about the covariance matrix. In everyday practise, the variables X and Y may have different dimensional units. We may, for example, investigate the relationship between the the mass of a galaxy, X in $M_\odot$ or gr, and its rotation speed Y in $km\,s^{-1}$. The value of the covariance between these two variables depends on the units, as seen by the second of equations (1.119) [32].

1.12.2 Covariance of the bivariate Gaussian distribution

In the case of the bivariate Gaussian distribution (**??**)

$$f_{X,Y}(x,y\,;\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \quad (1.135)$$

where μ_X and σ_X are the mean and variance of X, with like expressions for Y and ρ is referred to as the *correlation* between the two variables (see section 1.13).

It can be verified that

$$\mathbb{E}[X] = \mu_X \qquad \mathbb{E}[Y] = \mu_Y \quad (1.136)$$

and that the variances and covariances can be written

$$\mathrm{Var}(X) = \sigma_X^2 \quad (1.137)$$

$$\mathrm{Var}(Y) = \sigma_Y^2 \quad (1.138)$$

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = \rho\,\sigma_X\sigma_Y. \quad (1.139)$$

The covariance matrix is then

$$\mathbf{\Sigma}_{Gaussian} = \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \quad (1.140)$$

The determinant of $\mathbf{\Sigma}_{Gaussian}$ is

$$|\mathbf{\Sigma}_{Gaussian}| = \sigma_X^2\sigma_Y^2(1-\rho^2), \quad (1.141)$$

and its inverse is

$$\mathbf{\Sigma}^{-1}_{Gaussian} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)}\begin{pmatrix} \sigma_Y^2 & -\rho\,\sigma_X\sigma_Y \\ -\rho\,\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} \quad (1.142)$$

This provides the link between the concept of covariance and its relationship to the 2- and n-dimensional Gaussian distributions shown in equation (1.148) of section (1.12.4).

[32] Astronomers generally plot logarithmic values of such quantities: that provides an implicit normalisation.

1.12.3 Covariance matrix of a random vector

In order to discuss the covariance between more than two random variables, we simply take them in pairs and construct a covariance matrix from the pairwise covariances.

Consider a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. We will denote their means by $\mu_i = \mathbb{E}[X_i]$ and write this in vector form as $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$. Their covariance is described by the $n \times n$ matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}, \mathbf{X})$ whose elements are

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]. \tag{1.143}$$

We can write out this matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} \tag{1.144}$$

and we should note that the diagonal elements are simply the variances of the corresponding element: $\mathrm{Cov}(X_a, X_a) = \mathrm{Var}(X_a)$.

We can see how this is related to the definitions (1.132) and (1.133) by the following simple argument. Using the definition (1.114) and the fact that the expectation of the matrix $\boldsymbol{\Sigma}$, $\mathbb{E}[\boldsymbol{\Sigma}]$, is the matrix of the expectation values of the elements, we have

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left[\begin{pmatrix} (X_1 - \mu_1)^2 & \ldots & (X_1 - \mu_1)(X_n - \mu_n) \\ \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & \ldots & (X_n - \mu_n)^2 \end{pmatrix}\right] \qquad (1.145) \\ &= \mathbb{E}\left[\begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_n - \mu_n \end{pmatrix}\begin{pmatrix} X_1 - \mu_1 & \ldots & X_n - \mu_n \end{pmatrix}\right] \\ &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \\ &= \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \qquad (1.146) \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T \qquad (1.147) \end{aligned}$$

Equations (1.146) and (1.147) are the vector generalisations of equations (1.132) and (1.133).

The covariance matrix $\boldsymbol{\Sigma}$ is a real symmetric matrix. It can be shown that it is also positive semi-definite [33] and so its eigenvalues are all non-negative. The inverse of the covariance matrix $\boldsymbol{\Sigma}$ is called the *Fisher Information matrix*: $\mathbf{F} = \boldsymbol{\Sigma}^{-1}$, or, alternatively and perhaps more commonly in statistics, the *precision matrix*.

Note that if the $\{ij\}^{th}$ component of $\boldsymbol{\Sigma}^{-1}$ is zero then the variables X_i and X_j are conditionally independent, given the other variables. However, in the case of an underlying

[33] A positive semi-definite matrix has all eigenvalues non-negative, whereas a positive definite matrix has all eigenvalues positive. Another way of putting this is that a $p \times p$ matrix $\mathbf{A}$ is non-negative definite if $\mathbf{a^T A a} \geq 0$ for all p-vectors $\mathbf{a}$, and positive definite if $\mathbf{a^T A a} > 0$ for all $p-$vectors $\mathbf{a} \neq 0$.

Gaussian distribution, $(\mathbf{\Sigma}^{-1})_{ij} = 0$ implies that the variables X_i and X_j are in fact independent, regardless of other variables.

1.12.4 Gaussian random vectors

The multi-variate normal distribution provides an important model for the statistical distribution of correlated components of random vectors. Given the covariance matrix $\mathbf{\Sigma}$, as in (1.144), we have the *p.d.f.*

$$f(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{1.148}$$

We saw the two dimensional formulation of this in section 1.12.2 where this expression was written out in full.

1.12.5 Cross-covariances of two vectors

Consider now the relationship between two vectors $\mathbf{X} = \{X_i\}$ and $\mathbf{Y} = \{Y_i\}$, not necessarily of the same dimension. The *cross-covariance* between $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{Y}-\mathbb{E}[\mathbf{Y}])^T\right] \tag{1.149}$$

where, following our convention, $\mathbf{X}$ and $\mathbf{Y}$ are regarded as single-column matrices. So if $\mathbf{X}$ has m components and $\mathbf{Y}$ has n components, $\mathrm{Cov}(\mathbf{X},\mathbf{Y})$ is an $m \times n$ matrix:

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \begin{pmatrix} \mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \dots & \mathrm{Cov}(X_1,Y_n) \\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \dots & \mathrm{Cov}(X_2,Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m,Y_1) & \mathrm{Cov}(X_m,Y_2) & \dots & \mathrm{Cov}(X_m,Y_n) \end{pmatrix} \tag{1.150}$$

and the cross-covariance is no longer symmetric unless $m = n$:

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \mathrm{Cov}(\mathbf{Y},\mathbf{X})^T \tag{1.151}$$

Note that if the vectors $\mathbf{X}$ and $\mathbf{Y}$ have different dimensions it makes no sense to add them, as in $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$. However, if they are of the same dimension, n and are in the same space, $\mathbb{R}^n$, we can write

$$\mathrm{Var}(\mathbf{X}+\mathbf{Y}) = \mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{Y}) + 2\mathrm{Cov}(\mathbf{X},\mathbf{Y}), \quad \mathbf{X},\mathbf{Y} \in \mathbb{R}^n \tag{1.152}$$

1.12.6 Independent samples of a random vector

Now we move on to describing an array of n independent samples of a random vector in d-dimensions. Generally speaking $n \geq d$ and so this is rather like a generalisation of the previous case discussing the cross-covariance of two vectors, except here we have n vectors all of the same dimension. Here we use the cross-covariance between any two of the vectors selected from the n.

First we introduce the concept of the *data matrix* of n independent samples of $d-$dimensional

row vector. The i^{th} sample can be written as $\mathbf{x}_i = (x_{i1}, \ldots, x_{id}),\ i = 1, \ldots, n$. These can be assembled into an $n \times d$ matrix whose rows are the components of the i^{th} sample:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ x_{21} & x_{22} & \ldots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nd} \end{pmatrix} \tag{1.153}$$

If we regard each of the components x_{ij} as a sample of a random variable X_{ij} then we can look at the matrix $\mathbf{x}$ as a realisation the n d-dimensional vectors $\mathbf{X}_i$ arranged in a random matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1d} \\ X_{21} & X_{22} & \ldots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \ldots & X_{id} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nd} \end{pmatrix} \leftarrow i^{th} \text{ data vector } \mathbf{X}_i \tag{1.154}$$

Each of the n row vectors $\mathbf{X}_i$ is a different (and independent) realisation or measurement of a d-dimensional vector quantity. Since the samples are taken as being independent, the covariance matrix elements $\mathrm{Cov}\,(X_{ai}, X_{bj})$ for such data sets must be zero if $a \neq b$. Hence the covariances will be

$$\mathrm{Cov}\,(X_{ai}, X_{bj}) = \delta_{ab}\sigma_{ij} \tag{1.155}$$

where $\delta_{ab} = 1$ if $a = b$ and 1 otherwise. σ_{ij} is the covariance between the i^{th} and j^{th} components of the vector $\mathbf{X}_a$.

We can develop equations analogous to (1.128) and (1.129) by defining some useful constant matrices [34]

$$\mathbf{I} = \mathrm{diag}\,(1, \ldots, 1) \tag{1.156}$$

$$\mathbf{1} = (1, \ldots, 1)^T \tag{1.157}$$

$$\mathbf{W} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \tag{1.158}$$

There is something wrong with the second equation here

With this notation the sample mean and covariance are expressed in matrix notation as

$$\mathbf{s}_\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}\mathbf{x}^T\mathbf{1} \tag{1.159}$$

$$\mathbf{s}_{\mathbf{X}\mathbf{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{s}_\mathbf{X})(\mathbf{x}_i - \mathbf{s}_\mathbf{X})^T = \frac{1}{n-1}\mathbf{x}^T\mathbf{W}\mathbf{x} \tag{1.160}$$

In order to understand these expressions remember that $\mathbf{x}_i$ is a single column matrix, and

[34] Appendix **??** provides a brief overview of some of the more advanced concepts of matrix algebra that might be needed here. Standard advanced texts on matrices are Golub and van Loan (1996, *Matrix Computations*) and Horn and Johnson (2012, *Matrix Analysis*) . See also Press et al. (2007).

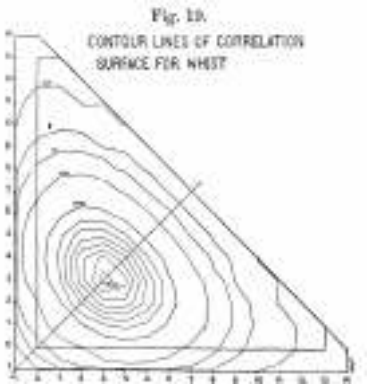


Fig. 1.5 Perhaps the first ever sketch of a correlation function contour map in a scientific journal, drawn by Pearson (1895, Figure 19, Plate 9). There is no figure caption in the article, nor does the article discuss the diagram!

hence so is $\mathbf{s_X}$. Likewise, $(\mathbf{x}_i - \mathbf{s_X})(\mathbf{x}_i - \mathbf{s_X})^T$ is a matrix product of a single column matrix and its transpose row matrix, which leads to an $n \times n$ square covariance matrix.

1.13 Correlation

The *correlation* ρ_{XY} between X and Y is their covariance normalised by their standard deviations:

Definition 1.23
Correlation coefficient:

$$\rho_{XY} = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{\mathrm{Var}\,(X)}\,\sqrt{\mathrm{Var}\,(Y)}}, \quad |\rho| \leq 1 \tag{1.161}$$

In the statistical literature ρ is referred to as the *Pearson correlation coefficient*, it plays an important role in data analysis. The normalisation has the advantage of producing numbers in the range $|\rho| \leq 1$, unlike the covariances where the values are not particularity meaningful. The correlation matrix has 1's down its main diagonal.

Nevertheless, there is no match between the values of ρ and the perceived degree of correlation, especially when the trend in the paired values (x_i, y_i) is not linear (see figure 1.4).

If we look at equation (1.140) we see that the correlation matrix $\mathbf{R}$ for the bivariate normal distribution is simply

$$\mathbf{R}_{Gaussian} = \begin{pmatrix} 1.0 & \rho \\ \rho & 1.0 \end{pmatrix} \tag{1.162}$$

This, by construction, is the covariance matrix for the bivariate Normal distribution where both variates are standardised (see section 1.8.4).

1.13.1 Inverting the correlation matrix

In two and three dimensions, the correlation matrix can be inverted analytically.

First consider the simplest case of the correlation matrix $\mathbf{R}$ of two random variables, it has a simple inverse:

$$\mathbf{R} = \begin{pmatrix} 1.0 & r_{12} \\ r_{12} & 1.0 \end{pmatrix}, \qquad \mathbf{R}^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1.0 & -r_{12} \\ -r_{12} & 1.0 \end{pmatrix} \tag{1.163}$$

The situation for three random variables is a little more complicated but still analytically tractable:

$$\mathbf{R} = \begin{pmatrix} 1.0 & r_{12} & r_{31} \\ r_{12} & 1.0 & r_{23} \\ r_{31} & r_{23} & 1.0 \end{pmatrix}, \qquad \mathbf{R}^{-1} = \frac{1}{|\mathbf{R}|} \begin{pmatrix} 1 - r_{31}^2 & r_{31}r_{23} - r_{12} & r_{12}r_{23} - r_{31} \\ r_{31}r_{23} - r_{12} & 1 - r_{31}^2 & r_{12}r_{31} - r_{23} \\ r_{12}r_{23} - r_{31} & r_{12}r_{31} - r_{23} & 1 - r_{12}^2 \end{pmatrix} \tag{1.164}$$

$$|\mathbf{R}| = 1.0 - r_{12}^2 - r_{23}^2 - r_{31}^2 + 2r_{12}r_{23}r_{31} \tag{1.165}$$

Kopp (2008) presents a thorough discussion on diagonalising these matrices.

1.13.2 Sample correlation

The simplest estimator, r, of the correlation, ρ, between n paired samples (x_i, y_i) taken from random variables X and Y is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2}} \tag{1.166}$$

This can be rewritten in terms of the sample standardised variables as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right) = \frac{s_{XY}}{s_X s_Y} \tag{1.167}$$

Here s_X and s_Y are the unbiased estimators of sample standard deviations of the data sets $\{x_i\}$ and $\{y_i\}$ (see equation (1.105)) and s_{XY} is the sample covariance of equation (1.128) [35]. Writing the correlation coefficient this way shows how r is directly related to the standardised variables. Note that if we know we are sampling from distribution having zero mean, μ, and unit variance σ^2 equation (1.167) reduces simply to

$$r = \frac{1}{n-1}\left(\sum_{ij} x_i y_i\right), \quad \mu = 0,\ \sigma^2 = 1 \tag{1.168}$$

[35] In the statistical literature ρ is often referred to as the *population* correlation coefficient while r is the *sample* correlation coefficient.

Equation (1.166) involves the means $\bar{x}$ and $\bar{y}$. These can be written directly in terms of the given data to give the handy equation

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{1.169}$$

In practise, use of r is not straightforward since it is rather sensitive to outlying data points. Many alternatives have been proposed over the decades since Pearson wrote on this, notably Spearman's Rank Correlation and Kendall's Tau statistic[36]. These are discussed extensively in Press et al. (2007, *Numerical Recipes* sections 14.5 and 14.6). We shall come back to this in the chapter on Data Analysis.

1.13.3 Kendall's Tau-statistic

There are several non-parametric measures of correlation, one of which is Kendall's Tau statistic. Consider n realisations (X_i, Y_i) of a random sample taken from a bivariate distribution of random variables X and Y. There are $n(n-1)/2$ sample pairs to be considered. For each pair (X_i, Y_i), (X_j, Y_j) we can ask about the sign of $m_{ij} = (Y_i - Y_j)/(X_i - X_j)$ [37]. If for a pair (ij), $m_{ij} > 0$ we call the pair (ij) concordant, and if $m_{ij} < 0$ we call them discordant. If $X_i = X_j$ or $Y_i = Y_j$ we say that the pair (ij) is tied.

We count the number N_c of concordant pairs, N_d of discordant pairs and the number of ties, N_t. Then the Kendall Tau statistic is

$$\tau = \frac{N_c - N_d}{n(n-1)/2} \tag{1.170}$$

If the ties were to count as half concordant and half discordant, then the contribution of ties to N_c and N_d would be equal and so the ties would contribute nothing to τ. However, if there were ties then τ could not achieve values ± 1. Goodman and Kruskal (1963) suggested a simple modification to τ and called the resulting statistic γ:

$$\gamma = \frac{N_c - N_d}{N_c + N_d} \tag{1.171}$$

In practice τ (or γ) is computed by sorting the pairs (X_i, Y_i) in ascending order of X_i and then simply comparing the $Y_i - Y_j$ values to get N_c and N_d. The test statistic used in hypothesis testing is $T = N_c - N_d$ (Conover, 1999, §5.4).

[36] The Spearman coefficient is simply the Pearson ρ calculated from the ranks of the data entries. Kendall's τ also uses ranked data, but looks at the difference between the probability that any 2 points will agree on the relative ranks of the variables compared to the probability that they will disagree

[37] We could equally consider the sign of the product $(X_i - X_j)(Y_i - Y_j)$.

1.14 Fundamental Theorems and inequalities

1.14.1 The Laws of Large Numbers

The usefulness of definition like (1.93) is ensured by what is loosely referred to as "The Law of Large Numbers". If $x_1 \ldots x_n$ are n independent samples taken from a random process X, we can construct the sample mean $\bar{x}_{(n)}$ of these n numbers:

$$\bar{x}_{(n)} = (x_1 + \cdots + x_n)/n \tag{1.172}$$

We are then assured that this converges to $\mathbb{E}[X]$ as the samples size n gets larger. There are basically two forms of the Law of Large Numbers: the weak and the strong, though there are a number of variants of each, differing in their precise mathematical statement of the conditions under which the law holds.

The law was first stated for independent samples of random variables by James Bernoulli in his posthumously published *Ars Conjectandi* (Bernoulli, 1713). This was later generalised by several people: Poisson, Chebyshev and Markov, who generalised it to non-independent random variables. But it was Borel who in 1909 proved what is now known as the *Strong Law of Large Numbers*.

Bernoulli's original statement was about independently and identically distributed samples from the Bernoulli distribution describing a series of experiments which return the value '0' with probability p and '1' with probability $q = 1 - p$. His statement was that the mean of n trials satisfies

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| > \epsilon\right] \leq \frac{pq}{n\epsilon^2} \tag{1.173}$$

We refer to this a "convergence in probability". The weak law of large numbers applies to all distributions, not only the Bernoulli distribution, and can be written as

$$\lim_{n\to\infty} \mathbb{P}[\,|\bar{x}_{(n)} - \mathbb{E}[X]\,| \geq \delta\,] = 0, \quad \forall\ \delta \tag{1.174}$$

where the x_i are n independent and identically distributed realisations taken from the distribution of the random variable X.

The Strong Law of Large Numbers tells us that this convergence is not merely a convergence in probability, but says that the mean of the sample, $(\Sigma_{i=1}^{n} X_i)/n$ tends to $\mathbb{E}[X]$ (*i.e.* p in Bernoulli's statement) with probability 1. This is written as a large n limit:

$$\mathbb{P}[\,\lim_{n\to\infty} \bar{x}_{(n)} \neq \mathbb{E}[X]\,] = 0 \tag{1.175}$$

We describe this as *almost sure convergence*.

The weak law does not assure us that the sample mean will stay near to $\mathbb{E}[X]$ for all arbitrarily large n: there may well be events such that $|\bar{x}_{(n)} - \mathbb{E}[X]| > \epsilon$ for an infinite sequence of n-values. It merely tells us about a probability. On the other hand the strong law states that such violations can occur only a finite number of times, and so there is some value of n beyond which $|\bar{x}_{(n)} - \mathbb{E}[X]| < \epsilon$ for all such n.

The interesting aspect of the statements (1.174) and (1.175) is that, for a sequence of

events $\{A_n\}$, in the former we have $\lim_{n\to\infty} \mathbb{P}[A_n]$, the limit is taken outside the $\mathbb{P}[\,]$, whereas in the latter we have $\mathbb{P}[\lim_{n\to\infty} A_n]$. In order to deal with the probability of a limit, *i.e.* the strong law, we need a mathematical theorem from measure theory known as the *Borel-Cantelli lemma*.

1.14.2 Central Limit Theorem

Formal statement required

This is an important theorem which assures us that, under rather general circumstances, the distribution of the average value measured for samples, taken independently from the same distribution, are Normally distributed, even if the underlying distribution is not itself Gaussian. Moreover, this normal distribution has the same mean as the underlying distribution and variance equal to the variance of the underlying distribution divided by the sample size.

We express this in more mathematical terms as follows. If we make n independent observations $x_1, x_2, \ldots, x_n$ of a random variable X that has mean μ and variance σ^2, the random variable

$$S_n = x_1 + x_2 + \cdots + x_n \tag{1.176}$$

has mean $n\mu$ and variance $n\sigma^2$.

The Central Limit Theorem states that, under certain rather general conditions, as n increases, the probability density distribution $f(x)$ of S_n tends to the Gaussian

$$f(x) \simeq \frac{1}{\sqrt{2\pi\, n}\sigma} e^{-\frac{(x-\mu)^2}{2n\sigma^2}} \tag{1.177}$$

Put slightly more formally, if the distribution of X has mean μ and variance σ^2 then

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = N(0,1) \tag{1.178}$$

where the symbol $N(0,1)$ denotes the normal distribution having zero mean and unit variance.

This is a remarkable result: it is independent of the underlying statistical distribution of the measurements x_i. It requires only that the statistical moments of this underlying distribution exist and, importantly, that the $\{x_i\}$ be independently distributed.

1.14.3 Markov's Inequality

Diagram needed

Markov's inequality imposes an upper bound on a non-negative function of a random variable. If X is a random variable, for any $s > 0$:

$$\mathbb{P}[|X| > a] \leq \frac{\mathbb{E}[|X|]}{a}, \quad a > 0 \tag{1.179}$$

Ex 1.14.1 Let $f(x)$ be the probability density function of the random variable X. Show that the probability that a value x chosen from X exceeds a is

$$\mathbb{P}[x \geq a] = \int_a^\infty x f(x) dx. \tag{1.180}$$

Write $\mathbb{E}[x] = \int_0^\infty x f(x) dx$ and deduce that $\mathbb{E}[x] \geq a\,\mathbb{P}[x \geq a]$

1.14.4 Chebyshev Inequality

Diagram needed

Suppose we wish to make a measurement of a quantity whose actual value is k. Formally we can express the result of making a measurement x that is prone to error by the equation

$$x = k + \delta \tag{1.181}$$

where k denotes the actual value of the quantity and δ, the measurement error, is a random variable having zero mean that reflects the error distribution. Denote the variance of the measurement errors δ by σ^2. Then the probability that a single measurement lies within ϵ of the actual value k satisfies the inequality

$$\mathbb{P}[|x - k| < \epsilon] > 1 - \frac{\sigma^2}{\epsilon^2} \tag{1.182}$$

This is the Chebyshev inequality [38]. It is a simple consequence of the Markov Inequality. It tells us that in order to get close to the real value k with a single measurement it is important
Proof PAP p151 that our measurement error, as quantified by σ, has $\sigma \ll \epsilon$.

We know intuitively that the accuracy can be improved by taking the mean of a number of measurements. If those measurements are denoted by $x_1, x_2, \ldots x_n$ then the sample mean of these is $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$, and $\bar{x}$ is a random variable of zero mean and variance σ^2/n. Hence The σ^2 term on the right hand side of equation (1.182) is decreased by a factor of n and the probability gets that much closer to 1.

1.15 Infinitely divisible distributions

Is this the right place for this?

A random variable X that is expressible as the sum of two independent random variables X_1 and X_2 having the same distribution (but not necessarily the same distribution as X:

$$X = X_1 + X_2 \tag{1.183}$$

are said the be *divisible*. An example of a divisible distribution is the normal distribution where the sum of two independent and identically distributed (*iid* for short) normal variates

[38] Chebyshev's name is also frequently spelled as Tchebycheff. He was Markov's teacher.

is also normal. Infinitely divisible distributions are important in physics since they are closely related to Lévy processes [39] (Sato, 1999).

Clearly this notion generalises to sums of any number n of independent identically distributed variates. The variate X that is the sum of the *iid* variates $X_1, \ldots, X_n$:

$$X = X_1 + X_2 + \cdots + X_n \tag{1.184}$$

can be said to be n-divisible, and if this is the case for all $n \geq 1$ the distribution of X is said to be *infinitely divisible*[40].

What about Saslaw?

Perhaps it is surprising that there are any distributions that are divisible, let alone infinitely divisible. The group of such distributions is indeed small, but it includes the Normal, Poisson, Negative Binomial and Cauchy distributions. A simple example of a distribution that is not indivisible is the Skew-Normal (section **??**).

1.15.1 Gravitational clustering

Introduced by Saslaw and Hamilton (1984) in the context of explaining the clustering properties of points in an N-Body self-gravitating system. Other references: Ahmad et al. (2002, 2006)

[39] A Lévy process is a homogeneous Markov process with spatial homogeneity. Brownian motion and Poisson processes are Lévy processes.

[40] A random variable Z that can be written as the sum of two independent random variables X and Y is said to be *decomposable*. It is *divisible* if the random variables X and Y are independent and identically distributed. Being *indecomposable* is the opposite of being infinitely divisible.

References

Ahmad, F., Saslaw, W. C., and Bhat, N. I. 2002. Statistical Mechanics of the Cosmological Many-Body Problem. ApJ, **571**(June), 576–584.

Ahmad, F., Saslaw, W. C., and Malik, M. A. 2006. Statistical Mechanics of the Cosmological Many-Body Problem. II. Results of Higher Order Contributions. ApJ, **645**(July), 940–949.

Bernoulli, Jakob. 1713. *Ars Conjectandi*. Impensis Thurnisiorum.

Broffitt, J.D. 1986. Zero Correlation, Independence, and Normality. *The American Statistician*, **40**, 276–277.

Conover, W.J. 1999. *Practical Nonparametric Statistics (3rd edition)*. 3rd. edition edn. Series in Probability and Statistics. John Wiley & Son.

Cox, D. R., and Hinkley, D. V. 1979. *Theoretical Statistics*. Chapman & Hall.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré*, **7**, 1–68.

de Finetti, B. 1974. *Theory of Probability (2 vols.)*. New York: John Wiley and Sons.

de Moivre, Abraham. 1718. *The Doctrine of Chances: Or, A Method of Calculating the Probability of Events in Play*. W. Pearson.

de Moivre, Abraham. 1738. *The Doctrine of Chances: Or, A Method of Calculating the Probability of Events in Play*. London: printed for A. Millar.

Doob, J.L. 1934. Probability and Statistics. *Trans. Am. Math. Soc.*, **36**, 759–775.

Doob, J.L. 1941. Probability as Measure. *Ann. Math. Stat.*, **12**, 206–214.

Feynman, R.P. 1951. The Concept of Probability in Quantum Mechanics. Pages 533–541 of: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Univ. Calif. Press.

Feynman, R.P., and Hibbs, A.R. 1965. *Quantum Mechanics and Path Integrals*. McGraw-Hill.

Fisher, R.A. 1925 (1st edition). *Statistical methods for Research Workers*. Oliver and Boyd, Edinburgh.

Golub, G. H., and van Loan, C. F. 1996. *Matrix Computations*. 3rd edn. Johns Hopkins University Press.

Good, I.J. 1975. Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, **30**, 39–73.

Good, I.J. 1979. A. M. Turing's Statistical Work in World War II. *Biometrika*, **66**.

Good, I.J. 2000. Turing's Anticipation of emprical bayes in connection with the cryptanalysis of the naval enigma. *Stat. Computation and Simulation*, **66**, 101–111.

Goodman, L.A., and Kruskal, W.H. 1963. Measures of Association for Cross-classifications II. *J. Amer. Stat. Assoc.*, **58**, 310–364.

Gradshteyn, I.S., and Ryzhik, I.M. 2007. *Table of Integrals, Series, and Products*. Academic Press; 7 edition.

Horn, R.A., and Johnson, C.R. 2012. *Matrix Analysis*. 2nd edition edn. Cambridge Univerity Press.

Huygens, Chrsitiaan. 1889. *Oeuvres complètes de Christiaan Huygens*. La Haye : M. Nijhoff for Societé Hollandaise des Sciences.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science (Vol 1)*. Cambridge University Press.

Kass, R.E., and Raferty, A.E. 1995. Bayes Factors. *J. Amer. Stat. Assoc.*, **90**, 773–795.

Kendall, M.G. 1949. On the Reconciliation of Theories of Probability. *Biometrika*, **36**, 101–116.

Keynes, J.M. 1921. *A Treatise on Probability*. London MacMillan & Co.

Kolmogorov, A. N. 1933. *Grundbergriffe de Wahrscheinlichkeitsrechnung (English translation "Foundations of the Theory of probability", published 1946, 1956, Chelsea Publishing New York)*. J. Springer, Berlin.

Kopp, J. 2008. Efficient numerical diagonalization of hermitian 3x3 matrices. *Int. J. Mod. Phys. C*, **19**, 523–548.

Lighthill, M.J. 1958. *An Introduction to Fourier Analysis and Generalised Functions*. Cambridge Monographs on Mechanics. Cambridge University Press.

Nelson, E. 1992. *Radically Elementary Probability Theory*. Annals of Mathematics Studies. Princeton University Press.

Papoulis, A., and Pillai, S.U. 2002. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Higher Education. 4th. edition.

Pearson, K. 1895. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Phil. Trans. Roy. Soc. A: Math., Phys. and Eng. Sci.*, **186**, 343–414.

Press, W.H., Teukolsky, S., Vetterling, W.T., and Flannery, B.P. 2007. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press; 3rd. edition.

Salmon, W.C. 1967. *The Foundations of Scientific Inference (Pitt Paperback*. University of Pittsburgh Press.

Saslaw, W. C., and Hamilton, A. J. S. 1984. Thermodynamics and galaxy clustering - Nonlinear theory of high order correlations. ApJ, **276**(Jan.), 13–25.

Sato, K. 1999. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Schrödinger, E. 1945-1948. The Foundation of the Theory of Probability -I. *Proc. Roy. Irish Acad. A: Math. and Phys. Sci.*, **51**, 51–66.

Sklar, A. 1973. Random Variables, Joint Distribution Functions, and Copulas. *Kybernetika*, **9**, 449–460.

von Mises, R. 1941. On the Foundations of Probability and Statistics. *Ann. Math. Stat.*, **12**, 191–205.

von Mises, R., and Doob, J.L. 1941. Discussion of Papers on Probability Theory. *Ann. Math. Stat.*, **12**, 215–217.

Weinstein, E. *MathWorld*. Wolfram Web Resource. [Online]. Available: . `http://mathworld.wolfram.com/NormalProductDistribution.html`.