Session 1

Probability: Introduction, definitions and axioms

1.1 Course overview and organisation

In sessions 1 to 5 we introduce the concept of probability, some of the most commonly encountered probability distributions in medical statistics, and important concepts. We will cover the following topics:

- Session 1
  - Probability as a concept
  - Axioms of probability
  - Conditional probabilities and independence
  - Theorem of total probability
- Session 2
  - Bayes theorem
  - Discrete random variables
  - Expectation and variance of random variables
- Session 3
  - Combinatorics
  - The binomial distribution
  - The Poisson distribution
- Session 4
  - Continuous probability distributions and density functions
  - Examples of continuous distributions, including the normal
- Session 5
  - Joint distributions
  - Covariance and correlation
  - The Central Limit Theorem

1.1.1 Overall Objectives

By the end of the 5 sessions you should be able to:

- explain basic concepts of probability theory
- draw a probability tree and obtain probabilities from it
- apply Bayes' Theorem to clinical examples and appreciate its important role in the area of screening
- state the probability distributions for the Normal, Poisson and Binomial distributions
- calculate the expectation and variance for these (and other) distributions

1.1.2 Recommended Textbooks

The following two texts contain the basics of probability theory required for medical statistics:

- Essential Medical Statistics, Kirkwood B. R. and Sterne J. A. C. Wiley-Blackwell, 2nd Edition, 2003.
- An Introduction to Medical Statistics, Bland J. M. OUP, 3rd Edition, 2000.

For a more in depth exposition, see: A First Course in Probability. Ross S. Pearson, 8th Edition, 2008.

1.2 Introduction

In session 1 we introduce the basic concepts of probability theory. By the end of this session you should appreciate why probability theory is used in medical statistics, understand the concepts of independence and conditional probability, be aware of the three fundamental axioms of probability and be able to draw a probability tree and derive probabilities from it.

1.2.1 What is probability?

Think about what we mean when we say that the probability of an outcome is 1%, 10%, 100% etc

One way to define probability is through the notion of *relative frequency*. If we carry out an experiment (e.g. rolling of a die) a large number of times, then the proportion of times a particular result (e.g. a six) occurs is known as the relative frequency. If we repeat the experiment a very large number of times then the limiting value of the relative frequency (i.e. the value that the relative frequency approaches as the number of times approaches infinity) is called a *probability*.

$$P(\text{roll a six}) = \lim_{n\to\infty} \frac{\text{number of 6s in } n \text{ rolls}}{n}$$

Figure 1.1 shows the relative frequency of 6s when a die is rolled repeatedly. The first roll was not a 6, and so the relative frequency is zero. The second roll was a 6, so the relative frequency is 0.5. As we then roll the die further, the relative frequency fluctuates. But from the figure we see that as the number of rolls increases, the relative frequency converges to 1/6.

1.3 Why is probability useful in medical research?

Probability is useful in the context of medical research due to the concept of uncertainty. When we carry out an experiment the results that we obtain have a degree of uncertainty.

Consider the hypothetical situation in the bottom half of Figure 1.2, which depicts the situation where we know the prevalence of disease in a population. If we then draw a random sample of individuals from

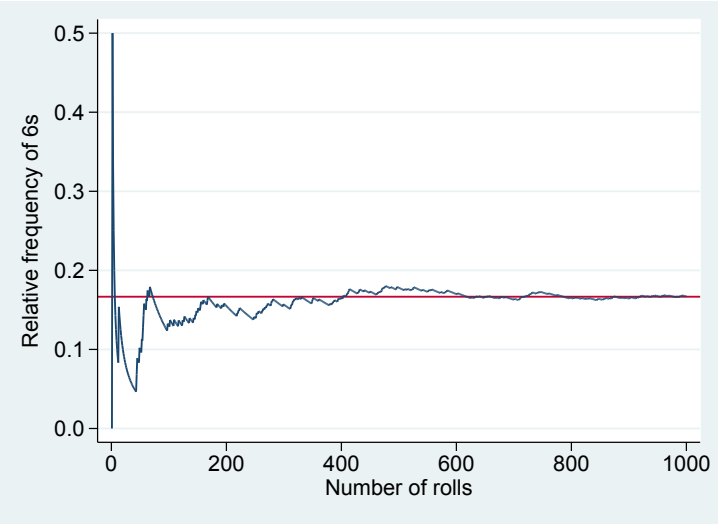

Figure 1.1: Relative frequency of 6s against trial number

this population, we can use *probability theory* to calculate the probability that a particular number x of these individuals have the disease. So probability theory is used to describe the uncertainty about how many diseased individuals we will obtain in our sample. Here we are assuming that characteristics (called parameters) of the population, such as the population prevalence, are known.

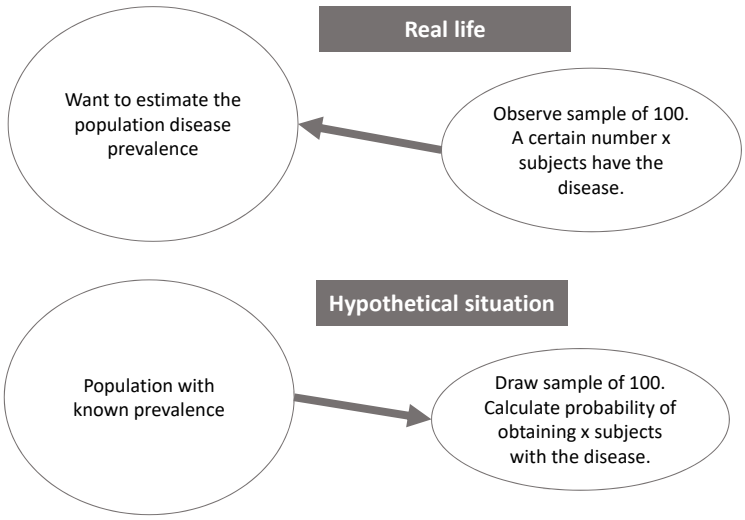

Figure 1.2: The role of probability theory in medical statistics

Real life is not like this though! Throughout medical studies, we observe a sample of individuals from a population whose parameters are unknown and we use statistical inference to make statements about the larger population from which this sample was drawn (the upper half of Figure 1.2). It turns out that probability theory is also needed in this situation, which is referred to as *inference*, i.e. how to make statements about the unknown population based on an observed sample. Here probability theory is being used to quantify our uncertainty about population parameters, given the data we have observed.

1.4 Definitions and notation

Experiment: A process that produces one outcome from some set of alternatives.

Sample space: The set of points representing all the possible outcomes of an experiment.

Suppose the experiment involved selecting an individual at random from the above population and denoting their smoking and asthma status. If we let A denote having asthma and S being a smoker, we can write the sample space as $\{AS, A\bar{S}, \bar{A}S, \bar{A}\bar{S}\}$, where $\bar{A}$ denotes not having asthma.

Event: A subset of the sample space, e.g. the event that a randomly selected individual is a smoker is $\{AS, \bar{A}S\}$.

Venn diagram: Venn Diagrams are sometimes used to represent probabilities in the whole sample space graphically. The whole diagram (bordered by the rectangle) represents the sample space and events within it are drawn with areas proportional to their probabilities.

Figure 1.3 shows a Venn diagram for the sample space considering smoking and asthma.

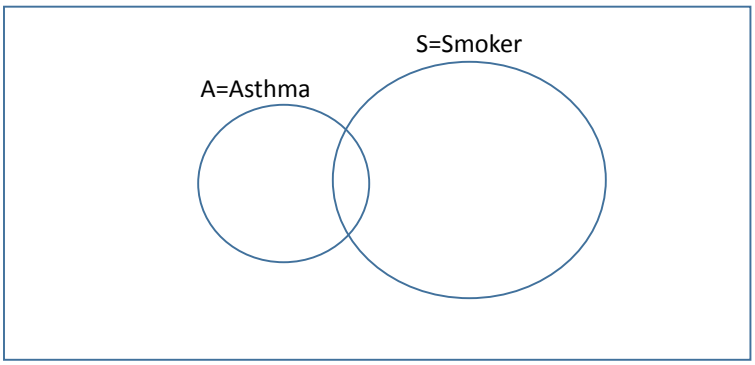

Figure 1.3: Venn diagram for smoking and asthma

1.4.1 Set notation for events

The **union** of two events X and Y, denote $X \cup Y$, is the event that either X occurs, or Y occurs, or both occur. For example, $A \cup S$ is the event that the randomly selected individual has asthma, is a smoker, or has asthma and is a smoker. This is represented by the total area included in the two ellipses in Figure 1.3.

The **intersection** of two events X and Y, denoted $X \cap Y$, is the event that both X and Y occur. It is sometimes referred to as the joint probability of X and Y. For example $A \cap S$ is the event that the randomly selected individual has asthma and is a smoker. This area is the intersection of the two ellipses in Figure 1.3.

The **complement** of an event X, denoted $\bar{X}$, is the event that X does not occur. The complement of X is also sometimes denoted X'.

1.5 Axioms of probability

The probabilities of events must follow the *axioms* of probability theory:

1. $0 \leq P(A) \leq 1$ for every event A.
2. $P(\Omega) = 1$ where Ω is the total sample space.
3. For disjoint (mutually exclusive) events $A_1, .., A_n$:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

   The events $A_1, .., A_n$ are disjoint if at most one of them can occur, i.e. there are no intersections between any of the events.

Axiom 3 is sometimes referred to as the *additive rule* of probability.

1.6 Consequences of the axioms

Given these axioms we can prove a number of useful results. Drawing a Venn diagram often helps us in constructing a proof, by appealing to a our visual intuition.

The first result we shall prove is that $P(\bar{A}) = 1 - P(A)$, for any event A (the 'complement rule'). Figure 1.4 shows a Venn diagram with the event A in the ellipse. Everything outside of the ellipse is the event $\bar{A}$.

To prove the result, note that A and $\bar{A}$ are disjoint, and so from Axiom 3 we have that $P(A \cup \bar{A}) = P(A) + P(\bar{A})$. Then since $A \cup \bar{A} = \Omega$, and $P(\Omega) = 1$ (Axiom 2), we have $1 = P(A) + P(\bar{A})$, which implies $P(\bar{A}) = 1 - P(A)$.

A and its complement $\bar{A}$ are **exhaustive**, meaning that it is certain that at least one of them will occur — you either suffer from asthma or you do not.

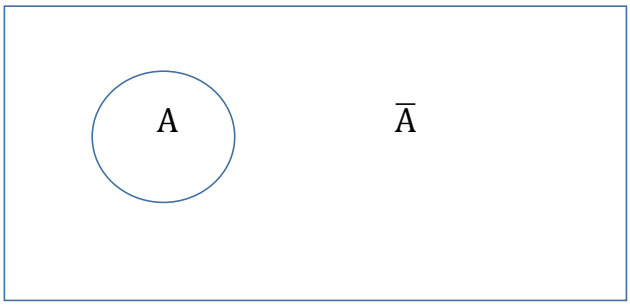

Figure 1.4: Venn diagram for proving the 'complement rule'

An extremely useful result is that $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. Figure 1.5 shows a Venn diagram which is useful in proving this result. The proof will be set as a practical question.

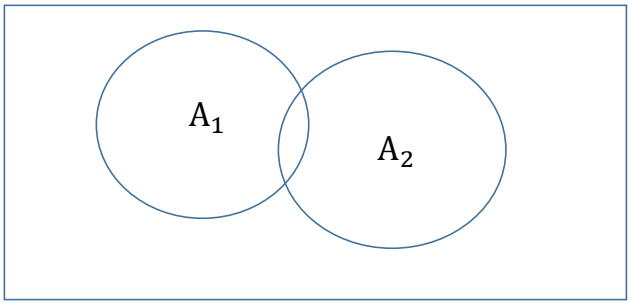

Figure 1.5: Venn diagram to aid the proof of $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

1.7 Conditional probability

Suppose that we know that an individual is a smoker. Given that event, what is the probability that the individual suffers from asthma?

We can write this as $P(A|S)$. This is called the conditional probability of A given S, i.e. the sample space is now reduced to S as we know the individual is a smoker. Once S has occurred, the event of interest will occur only if the observed outcome is in $A \cap S$.

In the original sample space $A \cap S$ has probability $P(A \cap S)$ with denominator $P(\Omega) = 1$, but in reduced sample space it has the re-weighted probability $P(A \cap S)/P(S)$. The reduced sample space is represented by the white area in Figure 1.6 (with the relevant event in blue stripes).

This leads to the definition of conditional probability

$$P(A|S) = \frac{P(A \cap S)}{P(S)}.$$

Multiplying through by $P(S)$, we see this implies that

$$P(A \cap S) = P(A|S)P(S).$$

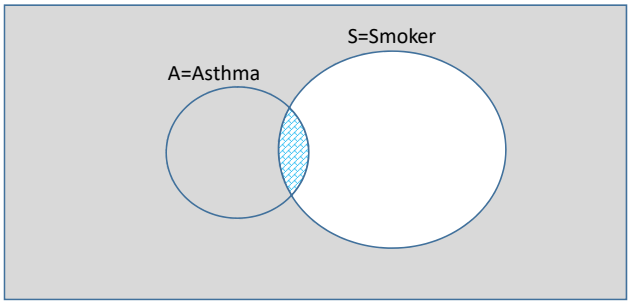

Figure 1.6: Venn diagram motivating the definition of $P(A|S)$

1.8 Probability trees

A good way of displaying and calculating simple conditional probabilities is through the use of a *probability tree*. The different outcomes are represented by the different branches of the tree. The events in a probability tree are exhaustive (every possible event is accounted for) and mutually exclusive, so you must go along one branch and only one branch at each stage.

The simplest form of a probability tree is shown in Figure 1.7. The second branches are all conditional probabilities, conditional on the events in each of the previous branches.

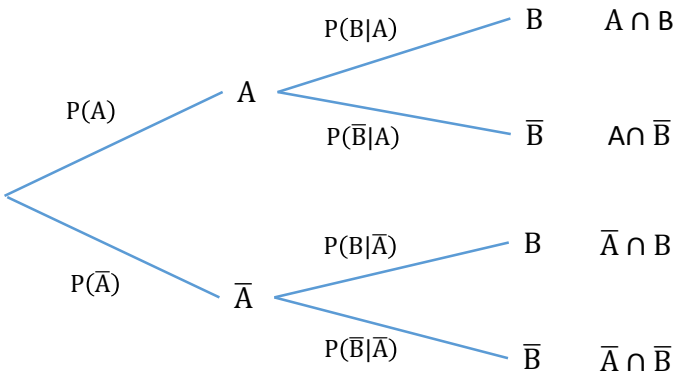

Figure 1.7: Basic probability tree

To calculate the probability of taking a particular path through the tree, we multiply the probabilities of the corresponding branches. That this is correct is a consequence of our earlier result that

$$P(A \cap S) = P(A|S)P(S).$$

For example, suppose in the population we know that the prevalence of smoking is 20% among adults in general and that 9% of smokers suffer asthma whereas 7% of non-smokers have asthma. This information can be represented in a probability tree. Note that we are given $P(A)$, $P(A|S)$ and $P(A|\bar{S})$, so the first branch of the tree is for smoking status and the second branch is asthma status conditional on smoking status (Figure 1.8).

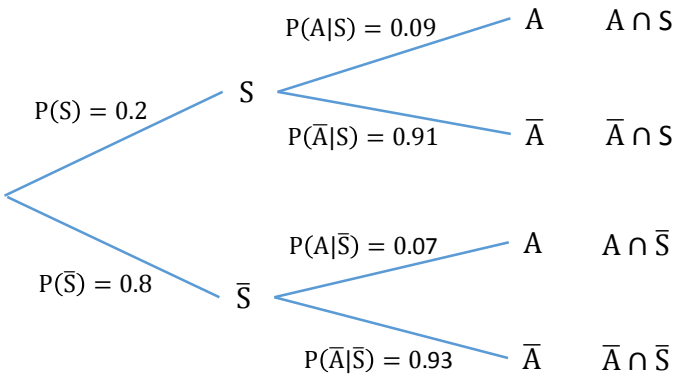

Figure 1.8: Probability tree for asthma and smoking

The probability of any intersection of asthma and smoking events can then be obtained by multiplying the probabilities along the branches of the tree leading to the tip which corresponds to the desired event. For example

$$\begin{aligned} P(A \cap S) &= P(S) \times P(A|S) \\ &= 0.2 \times 0.09 = 0.018. \end{aligned}$$

i.e. less than 2% of the population are both asthma sufferers and smokers.

Each tip of the tree represents an outcome of the 'experiment', and together the corresponding probabilities sum to 1.

1.9 Independence

A concept of fundamental importance in probability theory and medical statistics in particular is that of independence between events. Suppose that the occurrence of an event A_1 provides no information at all about the probability of a second event A_2 occurring (e.g. if knowing someone was a smoker gave no information about whether they had asthma). In that case, $P(A_2|A_1) = P(A_2)$. i.e. knowing that A_1 has occurred does not affect the probability that A_2 will occur. In this case

$$P(A_1 \cap A_2) = P(A_2|A_1)P(A_1) = P(A_2)P(A_1).$$

The latter is usually how independence between two events A_1 and A_2 is formally defined, i.e. A_1 and A_2 are said to be *independent* if

$$P(A_1 \cap A_2) = P(A_1) \times P(A_2). \tag{1.1}$$

This is known as the *multiplicative rule* of probability.

The concept of *independence* is often the central issue when we examine the association between two variables (e.g. using the χ^2 test — see Foundations: Analytical Techniques 4.4).

Conditional probability and independence play important roles in both the theory and application of medical statistics.

1.9.1 Example

Does taking vitamin supplements (as opposed to nothing) reduce the incidence of cold?

Define the probabilities of cold (C) and taking vitamin supplements (V) using the probabilities in Table 1.1.

	Cold (C)	No cold ($\bar{C}$)	Total
Vitamin supp. (V)	π_{CV}	$\pi_{\bar{C}V}$	π_V
No vitamin supp. ($\bar{V}$)	$\pi_{C\bar{V}}$	$\pi_{\bar{C}\bar{V}}$	$\pi_{\bar{V}}$
Total	π_C	$\pi_{\bar{C}}$	1

Table 1.1: Probabilities of having a cold (C) according to whether vitamin C was taken (V) or not ($\bar{V}$)

If vitamin C does not reduce incidence of cold, then the event C will be independent of the event V. We can check whether this holds by checking whether equation (1.1) holds, i.e. if

$$P(C \cap V) = P(C) \times P(V),$$

i.e. whether

$$\pi_{CV} = \pi_C \times \pi_V.$$

1.10 Theorem of total probability

The theorem of total probability allows us to express the probability of an event A occurring in terms of other events which *partition* the sample space.

1.10.1 Partitions

The events $B_1, B_2, ...B_n$ partition the sample space Ω if

1. $P(B_i) > 0$ for all i (all events in the partition have non-zero probability of occurring)
2. $\bigcup_{i=1}^{n} B_i = \Omega$ (i.e. the union of the events = the sample space)
3. $B_i \cap B_j = \emptyset$ (empty) for all $i \neq j$. i.e. B_i and B_j are disjoint

That is, the events $B_1, ..., B_n$ partition the sample space if

- all events are possible
- at least one event must occur, but
- no two events can occur simultaneously

Figure 1.9 shows an example of a partitioned sample space.

Some examples of partitions of sample spaces are:

- Blood groups A, B, AB and O partition the sample space of blood groups.
- Age groups 0-4, 5-9, 10-14 and 15-19 years partition the sample space of the ages of people under age 20 years.
- Men and women partition the sample space of sex outcomes.

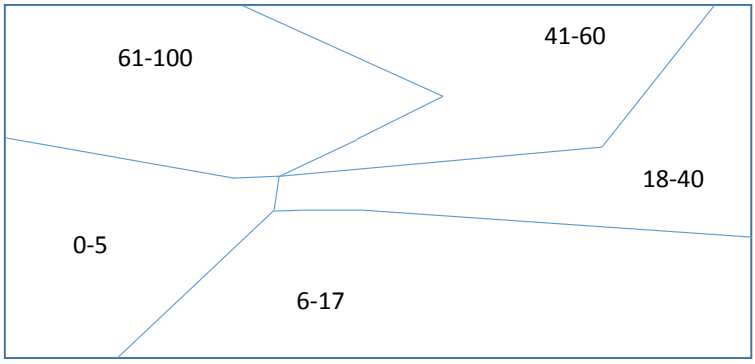

Figure 1.9: Partition of the sample space of ages (in a population where age ≤ 100)

1.10.2 Theorem of total probability

Let A be some event, and let $B_1, .., B_n$ be a partition of the sample space Ω. Then the theorem of total probability says that

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n).$$

That is, the total probability of A can be calculated if the conditional probability of A given B_i and $P(B_i)$ are known, for all i.

For example, given our partition $B_1, .., B_5$ of the sample size (Figure 1.10), we can express $P(A)$ using the probabilities $P(A|B_1), .., P(A|B_5)$ and $P(B_1), .., P(B_5)$.

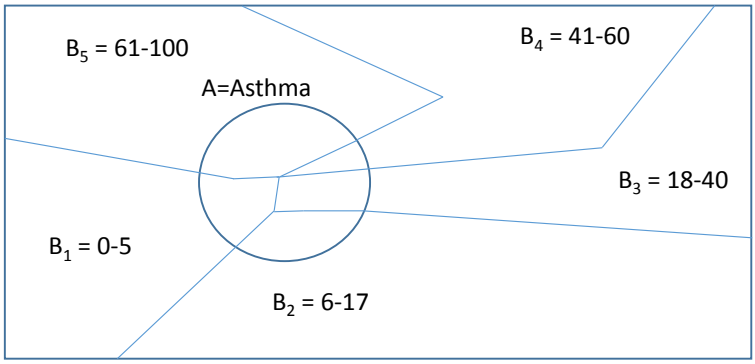

Figure 1.10: Expressing $P(A)$ based on a partition

Proof

First we note that we can express the set A as

$$(A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n).$$

Then, because the events $B_1, ...B_n$ are mutually exclusive (disjoint), the sets $(A \cap B_1), (A \cap B_2), ..., (A \cap B_n)$ are also mutually exclusive. Because they are mutually exclusive we can use Axiom 3 to give

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n).$$

Finally, we use the fact that $P(A \cap B_i) = P(A|B_i)P(B_i)$ to express $P(A)$ as

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n).$$

The theorem of total probability has important application in Bayes' theorem.

1.10.3 Example: smoking and asthma

From Section 1.8 we know the probability of being a smoker and the probability of having asthma conditional on smoking status. To calculate the total probability of having asthma, we can use the theorem of total probability, since the events of 'being a smoker' or 'not' form a partition of the sample space.

$$\begin{aligned} P(A) &= P(A|S)P(S) + P(A|\bar{S})P(\bar{S}) \\ &= 0.09 \times 0.2 + 0.07 \times 0.8 \\ &= 0.074. \end{aligned}$$

Session 2

Bayes' theorem, random variables, expectation and variance

2.1 Objectives

This session is concerned with some important further theory and applications of probability: Bayes' Theorem, screening, the concepts of random variables, expectation and variance.

By the end of the session you should:

- understand the concepts of conditional probability and Bayes' Theorem
- understand the concepts of a discrete random variable and its cumulative distribution function
- know how the expectation and variance of a discrete random variable are defined
- understand the notion of a joint distribution and of independence

2.2 Bayes theorem

In the theorem of total probability, $P(A)$ is calculated using the knowledge of the conditional probabilities of A given B, $P(A|B)$. In many situations we would like to make statements about the probability of B knowing or conditional on A, i.e. we would like to reverse the conditioning. Bayes' Theorem provides a useful and very powerful theorem to do this.

From the definition of conditional probability we can express $P(A \cap B_i)$ in two different ways

$$\begin{aligned} P(A \cap B_i) &= P(A|B_i)P(B_i) \\ &\text{or} \\ P(A \cap B_i) &= P(B_i|A)P(A). \end{aligned}$$

Equating the two, we therefore have

$$\begin{aligned} P(A|B_i)P(B_i) &= P(B_i|A)P(A) \\ \Rightarrow P(B_i|A) &= \frac{P(A|B_i)P(B_i)}{P(A)}. \end{aligned}$$

From the theorem of total probability we know that

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n) = \sum_{j=1}^{n} P(A|B_j)P(B_j),$$

(notice that we must now use a different index to i for the summation) so we can write

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

This is *Bayes' Theorem*. It has important and powerful applications in medical statistics, particularly in the interpretation of odds ratios estimated in case-control studies. Another common use is in the area of screening.

2.2.1 Example: asthma and smoking

Returning to the example of asthma and smoking, we know the probability being a smoker and the probability of having asthma conditional on smoking status. In order to make statements about how asthma relates to the probability of being a smoker, an individual's asthma status is the known exposure and smoking would be the unknown outcome, i.e. we would like to make a statement about the probability of smoking given your asthma status, $P(S|A)$ and $P(S|\bar{A})$.

For someone with asthma, applying Bayes' Theorem, the probability of being a smoker is

$$P(S|A) = \frac{P(A|S)P(S)}{P(A|S)P(S) + P(A|\bar{S})P(\bar{S})} = \frac{0.09 \times 0.2}{0.09 \times 0.2 + 0.07 \times 0.8} = 0.243.$$

2.2.2 Example: Genetic marker in childhood cancer

In a population, 10% of people develop a particular childhood cancer. Of those who develop the cancer, 1 in 4 carry a genetic marker, M, whereas of those who don't develop the cancer, 1 in 10 carry M. A newly born infant is tested for the genetic marker and is found to carry it, what is the probability that the individual will develop cancer?

We have reversed the conditioning and are now interested in $P(C|M)$. We can obtain this by applying Bayes' Theorem:

$$\begin{aligned} P(C|M) &= \frac{P(M|C)P(C)}{P(M|C)P(C) + P(M|\bar{C})P(\bar{C})} \\ &= \frac{0.25 \times 0.1}{0.25 \times 0.1 + 0.1 \times 0.9} = 0.22 \end{aligned}$$

$P(C|M)$ is called the *positive predictive value* (PPV) of the test. It is the probability, given a positive test result, that the individual actually will develop the disease. i.e. in this case there is a 22% chance that the infant will develop the disease if they tested positive.

This is an example of a screening test. Screening tests are tests that indicate whether disease or high risk subjects may be identified. Several useful measures can be obtained:

- $P(C)$ = prevalence of disease
- $P(C|M)$ = predictive value of a positive test (PPV)
- $P(M|C)$ = sensitivity of test
- $P(\bar{M}|\bar{C})$ = specificity
- $P(M|\bar{C})$ = 1-specificity

The sensitivity of a test is the probability of testing positive, given that the person is actually diseased and the specificity is the probability of testing negative given that the person is not diseased.

2.3 Random variables

A random variable X is a variable which takes a numerical value which depends on the outcome of the experiment under consideration. Random variables which take a value either from a finite or countably infinite set (e.g. the positive integers) are known as *discrete random variables*. In contrast, continuous random variables take values in an uncountable set (e.g. positive real numbers). For now we will consider discrete random variables.

A random variable is characterised by its *probability distribution*. For discrete random variables the *probability mass function* gives the probability that the variable takes each of the values which it might take:

$$P(X = x_i), \quad \text{for each value } x_i \text{ that } X \text{ can take}$$

Suppose X represents the number of boys in a four child family. It is important to understand the distinction between X and x. Capital X denotes the random variable whose value is the number of boys in a four child family to be selected randomly. Little x is used to denote possible values which the random variable X might take. So the expression $P(X = x)$ can be read as 'the probability that the random variable X takes value little x'.

We can tabulate (Table 2.1) or graphically show the probability distribution for the example with the number of boys in a four child family (Figure 2.1).

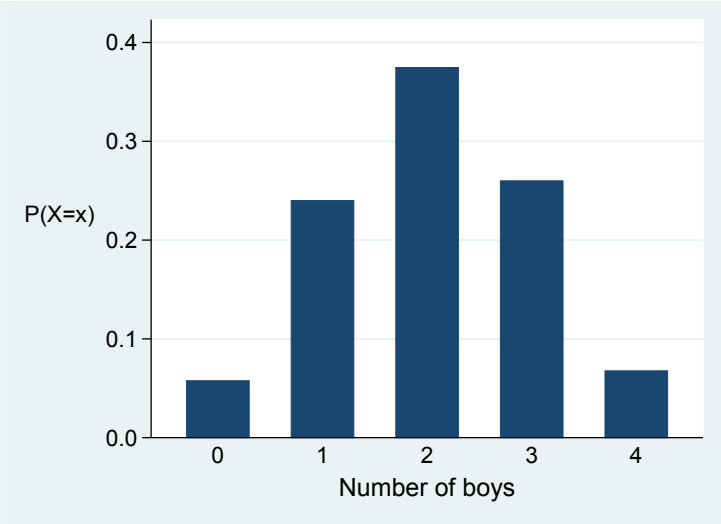

Figure 2.1: The probability distribution for X = number of boys in a four child family

2.3.1 Cumulative distribution function

For a random variable X, its cumulative distribution function (CDF) is given by:

$$F(x) = P(X \leq x)$$

Given the probability mass function for X we can derive the CDF, and vice-versa — they are two different ways of encoding the same information.

Returning to the example of the number of boys in a four child family, we can calculate the CDF based on the probability mass function, as shown in Table 2.1.

x	$P(X = x)$	$F(x) = P(X \leq x)$	$x \times P(X = x)$
0	0.06	0.06	0×0.06
1	0.24	0.3	1×0.24
2	0.37	0.67	2×0.37
3	0.26	0.93	3×0.26
4	0.07	1	4×0.07

Table 2.1: Various quantities concerning X = number of boys in a four child family

2.4 Expectation and variance

2.4.1 Expectation of a random variable

The expectation (or mean) of a random variable X is one measure of the centre of its distribution (another is the median). For discrete random variables X, it is defined as:

$$E(X) = \sum_{x_i} x_i P(X = x_i),$$

where the summation is over all possible x_i values that X can take. One way to think of $E(X)$ is the average value of X over a large number of repetitions of the experiment or random process that produces X. The Greek letter μ is often used for $E(X)$. Note that here we are defining the population mean, which is not the same as the sample mean.

Returning to the example of number of boys in a four child family, for each value x_i we can calculate $x_i \times P(X = x_i)$. This is shown in the last column of Table 2.1. To find $E(X)$ we then simply take the sum of these values across all values of x_i:

$$\begin{aligned} E(X) &= 0 \times 0.06 + 1 \times 0.24 + 2 \times 0.37 + 3 \times 0.26 + 4 \times 0.07. \\ &= 2.04 \end{aligned}$$

Note that we do not actually expect to find 2.04 boys in a 4 child family! Rather, if we repeatedly sample X from a population, and then take the average of the values of X across the samples, the value we expect to get is 2.04, with the value getting closer and closer the more samples we take (see Session 5).

Expectations of functions of random variables satisfy certain rules. For now, we will consider the effects (on the expectation) of multiplying a random variable by a constant a or adding a constant b to the random variable. If we add a constant b to X, the expectation of the newly obtained random variable is simply $E(X) + b$. This is because adding b just shifts the distribution of X by b. Similarly, if we multiply X by a constant a, the new random variable aX has expectation $aE(X)$, since for each value x which X takes, aX takes the value ax. Combining these two results, we have that $E(aX + b) = aE(X) + b$. To prove the result we use the definition of expectation for a random variable X, note that in general $E(g(X)) = \sum_{x_i} g(x_i)P(X = x_i)$:

$$\begin{aligned} E(aX + b) &= \sum_{x_i} (ax_i + b)P(X = x_i) \\ &= \sum_{x_i} (ax_i P(X = x_i) + bP(X = x_i)) \\ &= a \sum_{x_i} x_i P(X = x_i) + b \sum_{x_i} P(X = x_i) \\ &= a \sum_{x_i} x_i P(X = x_i) + b \\ &= aE(X) + b. \end{aligned}$$

Note that $\sum_{x_i} P(X = x_i) = 1$ by the axioms of probability.

2.4.2 Variance of a random variable

Expectation is a measure of the centre of a distribution. In contrast, the variance of a random variable measures the magnitude of the dispersion in the distribution around its expectation. The variance of a discrete random variable X is defined as

$$Var(X) = E((X - \mu)^2),$$

where $\mu = E(X)$. The variance uses the square of the distance from observations to the mean because this is always positive. If we were to define variance instead as $E(X - \mu)$, this would always be equal to zero!

An often more useful expression for $Var(X)$ is $E(X^2) - \mu^2$, whose validity is easily proved:

$$\begin{aligned} Var(X) &= E((X - \mu)^2) \\ &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

In contrast to expectation, adding a constant b to a random variable does not affect its variance. This makes sense intuitively — shifting a distribution does not affect how dispersed the distribution is around its (newly shifted) mean. Multiplication of X by a constant a does affect the variability. Generally, we have that $Var(aX + b) = a^2 Var(X)$, which again is easily proved using the result that $E(aX + b) = aE(X) + b$:

$$\begin{aligned} Var(aX + b) &= E[((aX + b) - E(aX + b))^2] \\ &= E[(aX + b - a\mu - b)^2] \\ &= \sum_{x_i} (ax_i + b - a\mu - b)^2 P(X = x_i) \\ &= \sum_{x_i} (ax_i - a\mu)^2 P(X = x_i) \\ &= a^2 \sum_{x_i} (x_i - \mu)^2 P(X = x_i) \\ &= a^2 Var(X). \end{aligned}$$

2.5 Joint distributions

So far we have considered a single random variable X. Often in medical statistics we are concerned with the associations between two or more random variables, and so we need to be able to characterise how one variable depends on another. The starting point for this is to define the *joint* distribution of two random variables X and Y.

Let X and Y be two discrete random variables. We define their *joint* mass function $P(X = x, Y = y)$ for values x, y which X and Y can take by the probability $P(X = x \cap Y = y)$. We sometimes abbreviate $P(X = x, Y = y)$ by $P(x, y)$. The joint mass function must satisfy:

$$\begin{aligned} P(x, y) &\geq 0 \text{ for all } x, y \\ \sum_x \sum_y P(x, y) &= 1. \end{aligned}$$

The marginal distribution of either X and Y can be found from the joint distribution, e.g.

$$P(X = x) = \sum_y P(x, y).$$

The conditional distribution of one variable given the other can then be found as

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Lastly, the joint cumulative mass function is defined as

$$F(x, y) = P(X \leq x, Y \leq y).$$

2.5.1 Independence

Two random variables X and Y are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all possible values x and y that X and Y take.

2.5.2 Expectation of linear combinations

Now that we have defined the joint distribution of two random variables we can find the expectation of their sum:

$$\begin{aligned} E(X+Y) &= \sum_{x_j}\sum_{y_k}(x_j + y_k)P(X = x_j, Y = y_k) \\ &= \sum_{x_j}\sum_{y_k} x_j P(X = x_j, Y = y_k) + \sum_{x_j}\sum_{y_k} y_k P(X = x_j, Y = y_k) \\ &= \sum_{x_j} x_j P(X = x_j) + \sum_{y_k} y_k P(Y = y_k) \\ &= E(X) + E(Y), \end{aligned}$$

where we are using the fact that $\sum_{y_k} x_j P(X = x_j, Y = y_k) = x_j P(X = x_j)$. More generally we can prove that $E(aX + bY + c) = aE(X) + bE(Y) + c$.

2.5.3 Expectation of a product

If random variables X and Y are independent we can find the expectation of their product as:

$$\begin{aligned} E(XY) &= \sum_{x_j}\sum_{y_k} x_j y_k P(X = x_j, Y = y_k) \\ &= \sum_{x_j}\sum_{y_k} x_j y_k P(X = x_j)P(Y = y_k) \\ &= \sum_{x_j} x_j P(X = x_j) \sum_{y_k} y_k P(Y = y_k) \\ &= E(X)E(Y). \end{aligned}$$

2.5.4 Variance of a linear combination

Again suppose that X and Y are independent. Then we can find the variance of their sum as:

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2 + Y^2 + 2XY) - (E(X) + E(Y))^2 \\ &= E(X^2) + E(Y^2) + 2E(X)E(Y) - E(X)^2 - E(Y)^2 - 2E(X)E(Y) \\ &= Var(X) + Var(Y). \end{aligned}$$

We will return in Session 5 to the variance of the sum when X and Y are not independent.

Session 3

Combinatorics, the binomial distribution and the Poisson distribution

3.1 Objectives

This session considers some important discrete distributions. In order to do this, we first consider combinatorics (methods of counting). Then we use this to derive the binomial probability mass function, and then the Poisson probability mass function.

By the end of this session you should:

- be able to use permutations and combinations for calculating probabilities
- be familiar with the binomial distribution.
- be familiar with the Poisson distribution and the assumptions underlying its derivation

3.2 Permutations and combinations

Recall from Section 1.2.1 that the probability of an event can be defined as the relative frequency of that event in repeated experiments, e.g.

$$P(\text{roll a six}) = \lim_{n\to\infty} \frac{\text{number of 6s in } n \text{ rolls}}{n}$$

In situations like rolling a die, where each of the possible outcomes of the experiment are equally probable, and there are a finite number of possible outcomes, we can calculate the probability of events by dividing the number of outcomes in which the event of interest occurs by the total number of outcomes:

$$P(\text{event } A \text{ occurs}) = \frac{\text{number of outcomes in which event } A \text{ occurs}}{\text{total number of outcomes}}$$

For example, consider the event that when we roll the die we obtain an even number. There are 3 possible outcomes in which this event occurs (2, 4 and 6), and there are 6 possible outcomes. We can then calculate:

$$P(\text{rolling an even number}) = \frac{3}{6} = 0.5.$$

Combinatorics or enumeration is the branch of mathematics that deals with counting and thus can be used to calculate probabilities of discrete events. We now consider two useful formulae for counting, permutations and combinations, which can be used in more complex situations.

3.2.1 Example

In a group of 10 people, 2 are left-handed. Both left-handed people are dyslexic, whilst none of the right-handed people are. These data suggest that there may be an association between handedness and dyslexia, but it is also possible that they have arisen by chance. We can calculate the probability of obtaining such data if there was no association in truth. To find this, consider 'allocation' of the 2 people with dyslexia amongst the 10 people. Assuming these are all equal probable, we can then count the number of allocations in which the 2 dyslexic people are in the left handed group, and take the ratio of this to the total number of possible allocations. Permutations and combinations can help us to calculate the number of allocations.

A *permutation* is an arrangement **with** regard to order, whereas in a *combination* the order does not matter.

3.2.2 Permutations of n objects

For example, the letters AB and BA are two different permutations, but the same combination of letters.

If we have 3 distinct objects, labelled A, B and C, how many permutations are there? There are 6 permutations: ABC, ACB, BAC, BCA, CAB and CBA.

More generally, suppose we have n different objects. In how many ways can n different objects be arranged? Each arrangement (i.e. respecting different orders) is a permutation.

Consider a box with n compartments that we are going to fill with the n different objects.

1	2	.	.	.	n

There are n ways of filling the first slot, leaving $(n-1)$ objects and $(n-1)$ slots.

There are $(n-1)$ ways of filling the second slot, leaving $(n-2)$ objects to choose from.

There are $(n-2)$ ways of filling the third slot, leaving $(n-3)$ objects.

.

.

There is just one way of filling the (last) nth slot.

For each of the n ways of filling the first slot, there are $n-1$ ways of filling the second slot, i.e. there are $n(n-1)$ ways of filling the first two slots. For each of these $n(n-1)$ ways, there are $n-2$ ways to fill the third slot, etc. So there are $n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1$ ways of arranging the n objects, i.e.

$$\begin{aligned} \text{Number of permutations of } n \text{ objects} &= n \times (n-1) \times \ldots \times 3 \times 2 \times 1 \\ &= n! \end{aligned}$$

We denote this expression by $n!$ (pronounced n factorial). As a matter of convention we set $0! = 1$.

3.2.3 Permutations of a subset of x objects chosen from n

If we only select x objects from the total n possible, how many possible permutations are there of these x objects? Once again, we can think of the number of ways that we can fill x slots, as in the diagram below.

1	2	.	.	.	x

There are n ways of filling the first slot, leaving $(n-1)$ objects and $(n-1)$ slots.

There are $(n-1)$ ways of filling the second slot, leaving $(n-2)$ objects to choose from.

There are $(n-2)$ ways of filling the third slot, leaving $(n-3)$ objects.

.

.

There are $(n-x+1)$ ways of filling the xth slot.

So there are $n \times (n-1) \ldots \times (n-x+1)$ ways of arranging (or permuting) x objects from n.

Including $(n-x)(n-x-1)... \times 3 \times 2 \times 1$ in the numerator and denominator, we can express the number permutations of x objects from n as:

$$\begin{aligned} &= \frac{n(n-1)(n-2)...(n-x+1)(n-x)(n-x-1)... \times 3 \times 2 \times 1}{(n-x)(n-x-1)... \times 3 \times 2 \times 1} \\ &= \frac{n!}{(n-x)!}. \end{aligned}$$

We denote this by ${}_nP_x$ or ${}^nP_x$ (the number of permutations of x objects from n different objects).

We can think of the number of permutations of x objects chosen from n as the number of permutations of n objects, divided by the number of of permutations of the $(n-x)$ objects that we don't chose, i.e. $n!/(n-x)!$.

Example

In a 4 digit pin number where each digit cannot be repeated, the number of possible pin-numbers can be thought of as the number of permutations of 4 objects chosen from 10, i.e. the number of possible pin numbers is:

$$\begin{aligned} {}^{10}P_4 &= (10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)/(6 \times 5 \times 4 \times 3 \times 2 \times 1) \\ &= 10 \times 9 \times 8 \times 7 = 5040. \end{aligned}$$

3.2.4 Combinations of x objects chosen from n

Suppose now we select x objects chosen from n, but we are not concerned with the order of the x objects?

We already know that there are $n!/(n-x)!$ ways of permuting x objects from n different objects and that there are $x!$ ways of permuting x objects.

Let k = number of ways of choosing x from n objects without regard to order, i.e. the number of combinations of x objects from n objects.

For each of these combinations (of x objects) there are $x!$ permutations.

Hence the number of combinations of x objects chosen from n is:

$$k = \frac{n!}{x!(n-x)!}$$

Intuitively, there are more permutations than combinations, since for every combination there are several permutations, hence:

$$\frac{n!}{(n-x)!} > \frac{n!}{x!(n-x)!}$$

Thus, the number of combinations of x objects from n, denoted either by ${}^nC_x$ or as the binomial coefficient $\binom{n}{x}$, is given by

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

These quantities will be extremely useful in calculating the probability mass function for the binomial distribution. First, we look at a simpler distribution — the Bernoulli.

3.3 The Bernoulli distribution

The Bernoulli distribution is a discrete distribution that can take only two values: it takes value 1 with probability π and value 0 with probability $1 - \pi$.

$$\begin{aligned} P(X = 1) &= \pi \\ P(X = 0) &= 1 - \pi \end{aligned}$$

The expectation of a Bernoulli X can be found from the definition of expectation as

$$\begin{aligned} E(X) &= \sum_{x_i} x_i P(X = x_i) \\ &= 0 \times (1 - \pi) + 1 \times \pi \\ &= \pi. \end{aligned}$$

Similarly we can find the variance of X using the fact that $Var(X) = E(X^2) - E(X)^2$. First we find $E(X^2)$:

$$\begin{aligned} E(X^2) &= 0^2 \times (1 - \pi) + 1^2 \times \pi \\ &= \pi, \end{aligned}$$

and so the variance is given by

$$Var(X) = \pi - \pi^2 = \pi(1 - \pi).$$

3.4 Binomial distribution

3.4.1 Example

Consider the number of boys among families of four children (i.e. between 0 and 4). If we assume that the proportion of males at birth is 0.51 and that the gender of each birth is an independent event, then we may calculate the probabilities that a family of four children contains no boys, one boy, two boys, three boys and four boys.

Consider first the probability of no boys i.e. the probability of four girls. By applying the multiplication rule (using the assumption of independence between genders of the children) we obtain:

$$\begin{aligned} P(\text{four girls}) &= P(GGGG) \\ &= 0.49^4, \end{aligned}$$

where *GGGG* is shorthand for the event that the first child is a girl, *and* the second is a girl, *and* the third is a girl, *and* the fourth is a girl,

Consider now the probability of one boy and three girls. This may occur in one of four ways: BGGG, GBGG, GGBG and GGGB, each of which has probability $0.49^3 \times 0.51$. Thus

$$P(\text{one boy}) = 4 \times 0.49^3 \times 0.51.$$

A family of 2 boys and 2 girls will arise in one of the following 6 ways: BBGG, BGBG, BGGB, GBBG, GBGB, GGBB each with a probability $0.49^2 \times 0.51^2$ and a total probability of

$$P(\text{two boys}) = 6 \times 0.49^2 \times 0.51^2.$$

With similar reasoning we have that

$$\begin{aligned} P(\text{three boys}) &= 4 \times 0.49 \times 0.51^3 \\ &\text{and} \\ P(\text{four boys}) &= 0.51^4. \end{aligned}$$

We now let X be the random variables which records the number of boys in a randomly selected family of size four. This random variable takes four possible values: 0, 1, 2, 3 or 4. Its probability distribution is given by the following table:

Number of boys x	$P(X = x)$
0	$0.49^4 = 0.0576$
1	$4 \times 0.49^3 \times 0.51 = 0.2400$
2	$6 \times 0.49^2 \times 0.51^2 = 0.3747$
3	$4 \times 0.49 \times 0.51^3 = 0.2600$
4	$0.51^4 = 0.0677$

3.4.2 Definition

In the general situation, consider a sequence of n independent observations/trials (in the example above it was four). Each observation results in a binary outcome, e.g. each trial is a success or a failure. In fact, a Binomial sequence is the sum of n independent Bernoulli trials (i.e. n independent Bernoulli variables).

Let π denote the probability of an individual success (or the defined binary feature, e.g. boy vs. girl). To write that X follows a binomial distribution with these features, we write $X \sim \textit{binomial}(n, \pi)$, (where $\sim$ means 'follows').

How do we obtain the probability distribution for the random variable X which records the number of successes in a sequence of n trials? The possible values for the random variable are $0, 1, .., n-1, n$. We saw from the previous example that the probability of x successes and $n - x$ failures is

$$P(X = x) = \pi^x (1 - \pi)^{n-x} \times \text{number of ways of obtaining } x \text{ successes.}$$

The multiplying factor on the right above is the binomial coefficient, i.e. the number of combinations of x objects chosen from n. The number of ways x successes can be obtained from n observations is equal to ${}^nC_x$ as we are not interested in the order of the successes, only the number of combinations in which such a number of successes could have occurred, and a 'success' can be considered the same as 'choosing' an object: we are 'choosing' x successes and $n - x$ failures out of a 'bag' of n successes and failures.

So the binomial probability distribution can be defined by

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, x = 0, 1, 2, .., n.$$

3.4.3 Application in medical statistics

Suppose X is a random variable for the number of successes out of n subjects selected from a large population of N subjects. If X follows a binomial distribution, the probability of success for each individual is a constant π and their individual failure/success outcomes are independent. π is the overall proportion of successes in the whole population, so $\pi = M/N$ where M is the number of successes in the population and N is the population size. In order for the probability of success for each individual to remain constant, when each of the n subjects is selected, they should be replaced. This process of sampling with replacement would lead to the possibility of selecting the same individual twice. In practice individuals are not replaced, but we can still assume that the probability remains constant provided that M and N are both large, since then the probability of success after the first sample is essentially unaffected.

3.4.4 Expectation and variance of the binomial distribution

We can now use the results we have derived to find the mean and variance of the binomial distribution. We do this by exploiting the fact that a binomial random variable is the sum of independent Bernoulli trials. Let X denote a binomial random variable with n trials and probability of success π. Then if we let X_i denote an individual Bernoulli trial with success probability π, we can express X as

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

Then using the fact that each Bernoulli trial has expectation π and variance $\pi(1-\pi)$, we have that

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} \pi \\ &= n\pi, \end{aligned}$$

and given that each Bernoulli trial is independent of the others, that

$$\begin{aligned} Var(X) &= Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} Var(X_i) \\ &= n\pi(1-\pi). \end{aligned}$$

3.5 Hypergeometric distribution

Recall that when the population size N and number of successes in the population M are large, the variable X recording the number of successes in a sample of size n has a binomial distribution. This approximate result fails when the population size N is small.

Suppose then that we have a population of size N, with M individuals having the characteristic corresponding to 'success', e.g. M out of N individuals in the population have a particular disease. We do not assume that N and M are large. Again, consider sampling n individuals from the population, without replacement. The probability of the first individual having the disease is M/N. However, the probability that the second individual has the disease is $M/(N-1)$ if the first individual did not have the disease but is $(M-1)/(N-1)$ if the first individual did have the disease. The probability

of 'success' is not constant as we sample individuals, and its value depends on the outcomes of the previous observations.

Let X be the random variable recording the number of individuals with the disease of interest in a sample of size n from a population of size N, of which M individuals have the disease. To calculate the probability that X takes a particular value x, we first note that there are ${}^{N}C_{n}$ possible ways of choosing (we do not care here about order) n individuals from a population of N. To find $P(X = x)$, we must now calculate how many of these ways result in x diseased individuals. There are ${}^{M}C_{x}$ ways of choosing x individuals with the disease from the M individuals with the disease in the population, and there are ${}^{N-M}C_{n-x}$ ways of choosing $n - x$ individuals without disease from the $N - M$ without the disease in the population. We therefore have that

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

since each possible way of choosing individuals from the population has equal probability. To succinctly express that X follows this distribution we write $X \sim hypergeometric(M, N, n)$.

Although we do not often (in medical statistics) perform studies in small finite populations, the hypergeometric distribution will be needed later in the course when considering whether two categorical variables are independent of each other ('Fisher's exact test').

3.6 The Poisson distribution

The Poisson distribution is used to model the *number of events* occurring in a fixed time interval T when:

- events occur randomly in time,
- they occur at a constant rate λ per unit time,
- they occur independently of each other.

A random variable X which follows a Poisson distribution can therefore take any non-negative integer value. Examples where the Poisson distribution might be appropriate include:

- Emissions from a radioactive source,
- The number of deaths in a large cohort of people over a year,
- The number of car accidents occurring in a city over a year.

To give a heuristic derivation of the probability mass function of the Poisson, we divide the total time T into a very large number of small intervals (Figure 3.1). As the number of intervals we divide T into increases, at most one event will occur in each interval, and so X will equal the number of intervals in which an event occurs. Since the occurrence of events in each interval are assumed independent of each other, $X \sim Bin(n, \pi)$, where n is the number of intervals and π is the probability of an event occurring in any given interval.

With a rate of λ events per unit of time, we expect $\mu = \lambda T$ events in the whole period, and therefore we expect $\lambda T/n = \mu/n$ events in each interval. Thus $\pi = \mu/n$. Therefore, using the probability mass function for the binomial we have that

$$P(X = x) = \binom{n}{x}\pi^{x}(1-\pi)^{n-x} = \binom{n}{x}\left(\frac{\mu}{n}\right)^{x}\left(1-\frac{\mu}{n}\right)^{n-x}$$

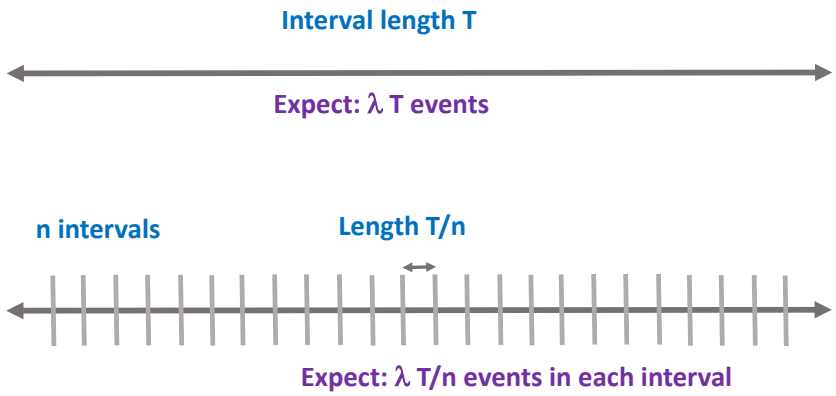

Figure 3.1: Derivation of the Poisson distribution

Then we have that

$$\begin{aligned} P(X=x) &= \binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\ &= \frac{n!}{x!(n-x)!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\ &= \frac{n!}{n^x(n-x)!}\frac{\mu^x}{x!}\left(1-\frac{\mu}{n}\right)^{n-x} \end{aligned}$$

Now:

$$\frac{n!}{n^x(n-x)!} = \frac{n(n-1)...(n-x+1)}{n^x} \to 1 \text{ as } n \to \infty,$$

and

$$\left(1-\frac{\mu}{n}\right)^{n-x} \to \left(1-\frac{\mu}{n}\right)^{n} \to e^{-\mu},$$

so

$$P(X=x) \to \frac{\mu^x}{x!}e^{-\mu} \text{ as } n \to \infty.$$

3.6.1 Definition

We can now define a Poisson distribution for the number of events occurring in a fixed interval T at a constant rate λ with parameter $\mu = \lambda T$, which we write as

$$X \sim \text{Poisson}(\mu = \lambda T)$$

as the distribution which has probability mass function

$$P(X=x) = \frac{\mu^x}{x!}e^{-\mu}, \quad \text{for } x = 0, 1, 2, ...$$

3.6.2 Expectation and variance

The derivation of the expectation and variance of a Poisson random variable X with parameter μ will be set as a practical question.

3.6.3 Example: asthma attacks

A clinical research is interested in modelling the number of asthma attacks that asthma suffers experience in one year. Based on a large sample the researcher has estimated that the average number of attacks in a year is 2.5. Assuming a Poisson distribution for X, the number of attacks a randomly selected asthmatic will suffer in a year, we can calculate $P(X = x)$ for any given value of x. This is shown graphically in Figure 3.2.

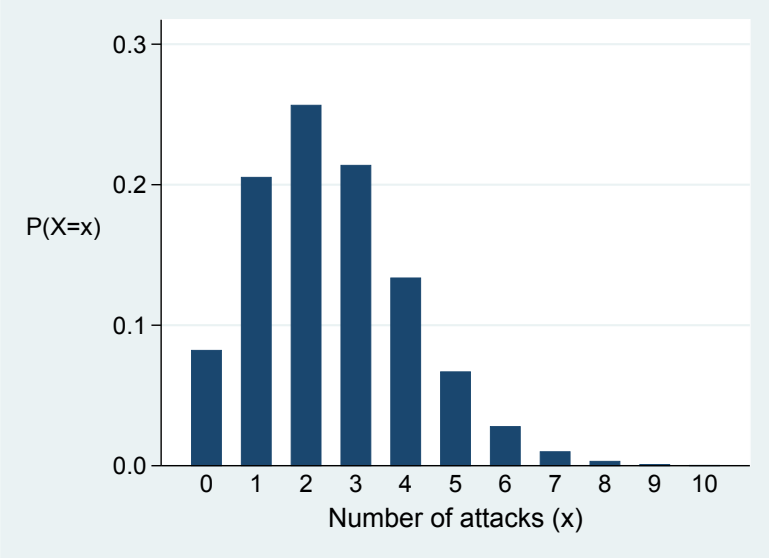

Figure 3.2: Poisson distribution for $\mu = 2.5$

3.6.4 Poisson as an approximation to the binomial

When n is large relative to π, the binomial distribution can be approximated by a Poisson with a mean $n\pi$. That this approximation is reasonable follows directly from our earlier heuristic derivation of how a Poisson distribution arises as an approximation to a binomial distribution when the number of trials tends to infinity.

3.6.5 The negative binomial distribution

In the Poisson distribution, the mean is the same as the variance. Sometimes, we wish to model count data with more (or less) variance than the Poisson allows for. One distribution that lets us do that is called the Negative binomial distribution. There are several different ways to describe the negative binomial distribution mathematically. One way is with parameters μ and θ, called the mean and dispersion parameters respectively. The mean of the distribution is μ and the variance is $\mu + \frac{\mu}{\theta}$. As θ gets bigger, the negative binomial distribution becomes more and more like the Poisson distribution with mean μ. The probability mass function of the negative binomial disruption with these parameters requires mathematics beyond the scope of this course.

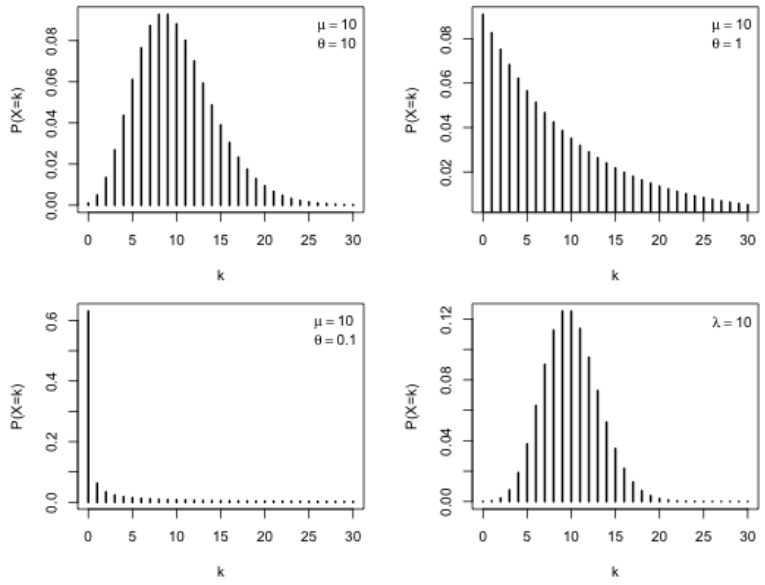

Figure 3.3: Negative binomial distributions with $\mu = 10$ and different values of θ. The bottom right panel shows a Poisson distribution with mean 10.

Session 4

Continuous probability distributions

By the end of this session you should:

- understand the notion of a continuous random variable, and know the definition of probability density functions
- know the definition of expectation and variance for continuous random variables
- know about the key continuous distributions which arise in medical statistics, including the normal

4.1 Continuous random variables and distributions

Recall that a discrete random variable X takes values in a finite or countably infinite set, and is characterised by the probabilities $P(X = x_i)$ for the possible values x_i it might take. For example, e.g. $X =$ weight measured to the nearest kg (for people who weigh between 50kg and 61kg) would be a discrete random variable.

4.1.1 Probability density functions

Continuous random variables arise when a random variable takes values in an *uncountably infinite* set, such as the set of real numbers. For continuous random variables X, the probability that X will take any particular value x is actually zero. This is because although X will take some value, there are *so* many possible values it could take the probability that it will take any particular value is zero. Instead of assigning probabilities to particular values, we define a *probability density function* $f(x)$ such that for any two values a and b,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

So the probability that X takes a value between a and b is equal to the area underneath the density function between $x = a$ and $x = b$, as depicted in Figure 4.1.

Probability density functions must satisfy a number of conditions so that the axioms of probability are not violated:

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)dx = 1$

This second condition is the continuous equivalent of the $\sum_{x_i} P(X = x_i) = 1$ requirement for discrete random variables. Notice that it is possible for $f(x) > 1$, whereas for discrete X, $0 \leq P(X = x_i) \leq 1$.

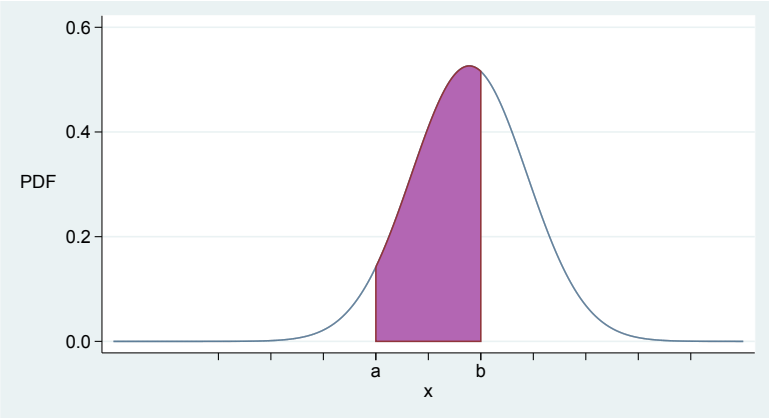

Figure 4.1: Probability density function

4.1.2 Cumulative distribution function

For continuous X, the CDF is defined as:

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

i.e. the area under the curve to the left of x. Probabilities can be expressed using the CDF too:

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx \\ &= F(b) - F(a). \end{aligned}$$

Because the CDF is the integral of the PDF,

$$f(x) = \frac{d}{dx}F(x).$$

4.1.3 Expectation and variance

Expectation is defined for continuous random variables X as for discrete X, but with the summation replaced by an integral:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

The variance is again defined as the expectation of the difference between X and its mean, squared: The variance of continuous X is thus given by:

$$Var(X) = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

4.2 The normal distribution

The normal (Gaussian) distribution is arguably the most important probability distribution in statistics. It lies at the heart of the modelling assumptions of linear regression (see Regression lectures). It also plays a critical role in statistical inference because of the Central Limit Theorem (see Session 5). It is often colloquially known as the bell-shaped curve (Figure 4.1).

Unlike the discrete distributions we have met so far, the normal distribution is characterised by two parameters, usually denoted μ and σ^2. The probability density function for a normal distribution with parameters μ and σ^2, written $X \sim N(\mu, \sigma^2)$, is given by:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The density takes positive values for all real values of x, it is symmetric about $x = \mu$, and it takes its maximum at $x = \mu$.

4.2.1 Expectation and variance

The parameters μ and σ^2 correspond to the mean (expectation) and variance of the distribution. To show this, let $z = (x - \mu)/\sigma$. Then $x = \sigma z + \mu$, and $dx = \sigma dz$. Then the expectation can be derived as

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \int_{-\infty}^{\infty} \frac{\sigma z + \mu}{\sigma\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \sigma dz \\
&= \int_{-\infty}^{\infty} \frac{\sigma z + \mu}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\
&= \int_{-\infty}^{\infty} \frac{\sigma z}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz + \frac{\mu}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\
&= \sigma \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\
&= \sigma E(z) + \mu \\
&= \sigma E\left(\frac{X-\mu}{\sigma}\right) + \mu \\
&= \sigma \frac{(E(X) - \mu)}{\sigma} + \mu \\
&= \mu,
\end{aligned}$$

where we have assumed that $\int_{-\infty}^{\infty} (1/\sqrt{2\pi}) \exp(-z^2/2)$ is equal to one. A slightly longer proof shows that the variance is equal to σ^2.

4.2.2 The standard normal distribution

The 'standard normal distribution' is the normal distribution with $\mu = 0$ and $\sigma^2 = 1$, $Z \sim N(0, 1)$. Its cumulative distribution function (CDF) is often denoted $\Phi(z)$. Statistical tables usually contain $F(z) = \Phi(z)$ for useful values of z (i.e. -4 to 4). This allows us to find $P(a \leq Z \leq b)$ where $Z \sim N(0, 1)$. Some important quantities for the standard normal are:

- $\Phi(0) = 0.5$
- $\Phi(-1.96) = 0.025$
- $\Phi(1.96) = 0.975$

So 95% of the area is contained within 1.96 standard deviations of the mean.

4.2.3 Calculating probabilities for the normal distribution

If we are able to calculate values of the CDF of the standard normal distribution Z, $F(z)$ (Figure 4.2), we can find probabilities for the general normal distribution $X \sim N(\mu, \sigma^2)$ via a transformation. Suppose that we are interested in finding $P(a \leq X \leq b)$. Then since $(X - \mu)/\sigma \sim N(0, 1) \sim Z$, we

can use the CDF of the standard normal to find the desired probability:

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) - P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(Z \leq \frac{a-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

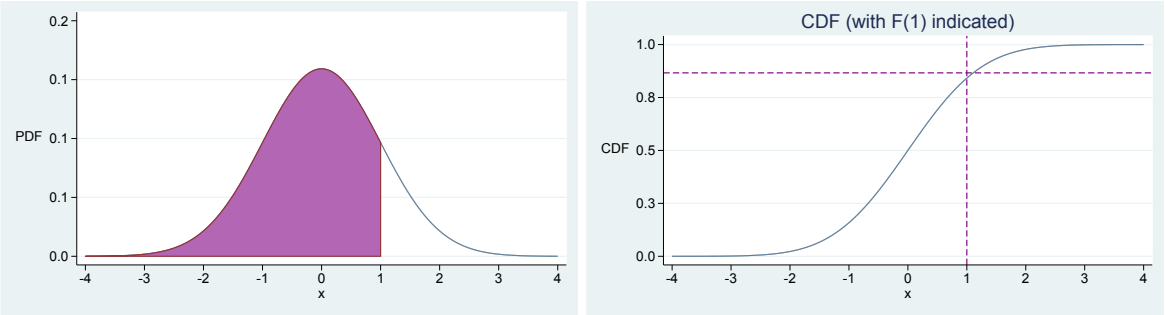

Figure 4.2: Cumulative distribution function for the standard normal Z

4.2.4 Example

Assume that heart rate is normally distributed with mean 74 beats per minute (bpm) and standard deviation 7.5 bpm.

What is the probability of having a heart rate less than or equal to 60 bpm?

Let X be the random variable for heart rate, so $X \sim N(74, 7.5)$. In order to calculate probabilities for a normally distributed variable, we must transform the random variable to a standard normal distribution (i.e. a normal distribution with a mean of zero and a standard deviation if one). To do this, we use the transformation:

$$Z = \frac{X-\mu}{\sigma}$$

In this example, $Z = (X - 74)/7.5$.

$$\begin{aligned} P(X \leq 60) &= P(Z \leq (60-75)/7.5) \\ &= P(Z \leq -1.867) \\ &= \Phi(-1.867) \\ &= 1 - \Phi(1.867) \\ &= 0.031 \text{ from Neave tables.} \end{aligned}$$

4.3 Some other continuous distributions

In the remainder of this session we briefly introduce some of the other continuous distributions which are important in medical statistics.

4.3.1 The (continuous) uniform distribution

The uniform distribution has a constant density function. $X \sim U(0, 1)$ if $f(x) = 1$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise (Figure 4.3). More generally, $X \sim U(a, b)$ if $f(x) = 1/(b - a)$ for $a \leq x \leq b$, and $f(x) = 0$ otherwise. So note that for $X \sim U(0, 1/2)$, $f(x) = 2$ for $0 \leq x \leq 1/2$ and $f(x) = 0$ otherwise.

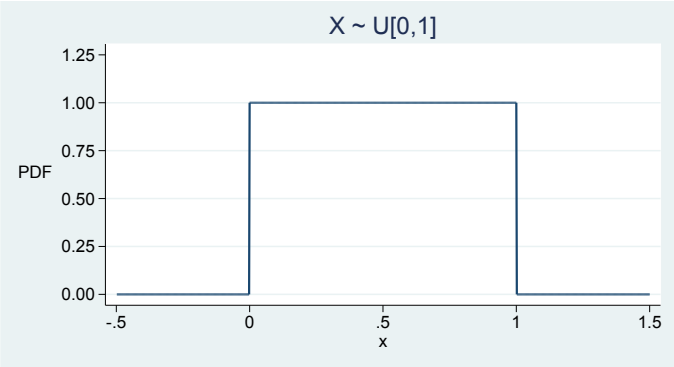

Figure 4.3: The continuous uniform distribution

4.3.2 The exponential distribution

Suppose the number of events in a period follow a Poisson distribution and occur at rate λ. Consider the time between two consecutive events. The mean can be shown to be $1/\lambda$, and the distribution is called the exponential (see Practical). It has PDF (Figure 4.4):

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

and CDF:

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_0^x \\ &= -e^{-\lambda x} + 1 = 1 - e^{-\lambda x} \end{aligned}$$

The expectation is given by:

$$\begin{aligned} E(X) &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \left[-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right] \\ &= 0 - (0 - \frac{1}{\lambda}) \\ &= \frac{1}{\lambda}. \end{aligned}$$

A similar proof shows that the variance is equal to $1/\lambda^2$.

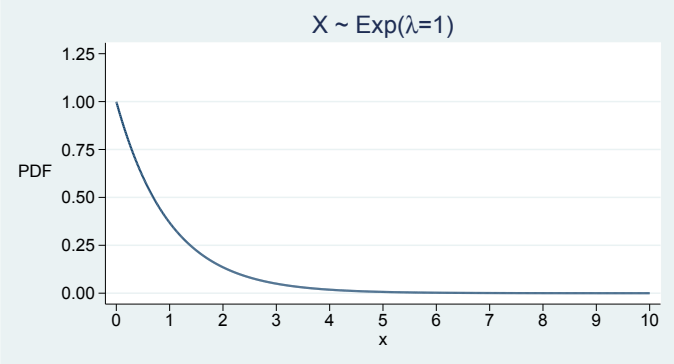

Figure 4.4: Density function of the exponential distribution with $\lambda = 1$

4.3.3 Student's t-distribution

Student's t-distribution arises as the ratio of the sample mean to its standard error, which you will learn more about in the inference course. The t-distribution has a complex density function which

we shall not state here. For now we note that the t-distribution has an additional parameter of sorts, known as the degrees of freedom (d.f.). The density function is similar to the normal, but the t-distribution has heavier tails (Figure 4.5). As the number of degrees of freedom increases the t-distribution gets closer and closer to the normal distribution.

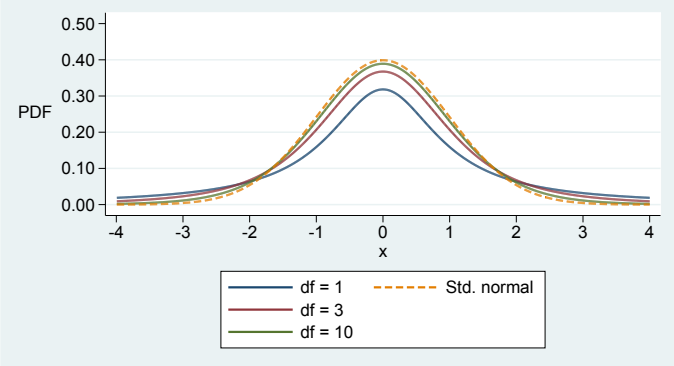

Figure 4.5: t-distribution (with various d.f.) and the standard normal

4.3.4 Chi-squared distribution

The chi-squared distribution with n d.f. arises as the sum of n independent standard normal variables. Thus if $X_1, .., X_n \sim N(0, 1)$ and are independent, then:

$$Q = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

follows a chi-squared distribution on n degrees of freedom. It has mean n and variance $2n$.

The chi-squared distribution is used in a number of areas of statistics, such as when estimating the variance of a normal distribution, and in hypothesis testing to examine if there is evidence of an association between two categorical variables.

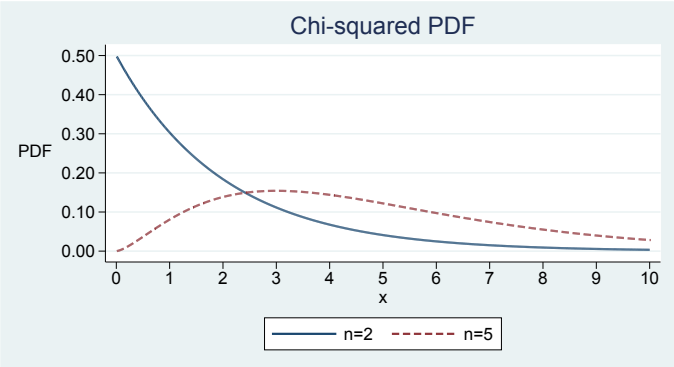

Figure 4.6: The chi-squared distribution with 2 and 5 d.f.

4.3.5 The F-distribution

The F-distribution arises when performing hypothesis tests in an analysis of variance and linear regression. Suppose $U_1 \sim \chi_n^2$, $U_2 \sim \chi_m^2$ and U_1, U_2 are independent, then:

$$F = \frac{U_1/n}{U_2/m}$$

and we write $F \sim F_{n,m}$. Figure 4.9 shows the PDF for the F-distribution with (4,5) degrees of freedom.

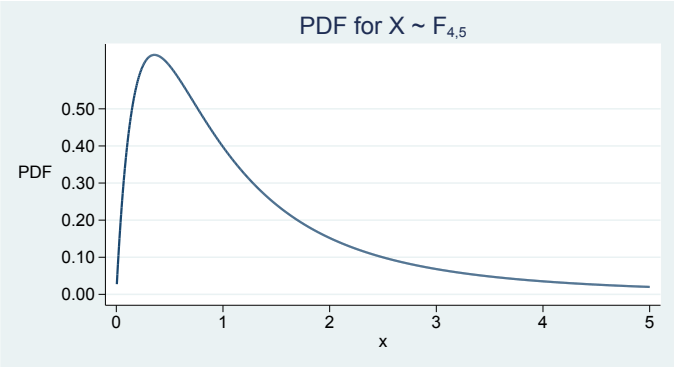

Figure 4.7: The F distribution with (4,5) d.f.

4.3.6 The Beta distribution

The beta distribution takes values between 0 and 1. It has two parameters, α and β and can take many shapes. The PDF has form $kx^{\alpha-1}(1-x)^{\beta-1}$, where k is a constant which depends on α and β. The beta distribution is commonly used in Bayesian statistics.

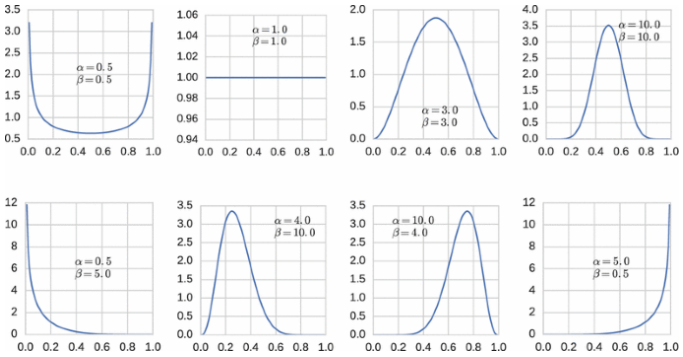

Figure 4.8: The Beta distribution with different values of α and β

4.3.7 The Gamma distribution

The gamma distribution takes values between 0 and ∞. It has two parameters, α and β and can take many shapes. The PDF has form $kx^{\alpha-1}e^{-\beta x}$, where k is a constant which depends on α and β. The gamma distribution is commonly used in Bayesian statistics, in survival analysis, and in models called random effects models which you will meet later in the MSc.

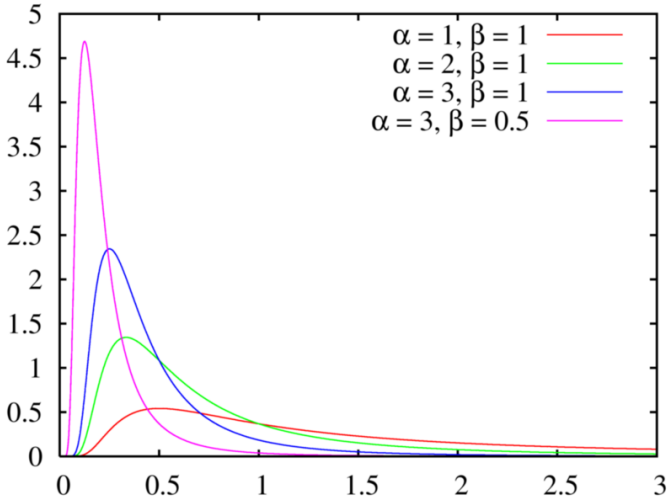

Figure 4.9: The Gamma distribution with different values of α and β

Session 5

Joint continuous distributions, covariance, correlation, the multivariate normal, and the Central Limit Theorem

By the end of this session you should:

- know how joint, marginal and conditional distributions are defined for continuous random variables
- understand the concepts of covariance and correlation
- understand the Central Limit Theorem and its implications
- be familiar with the multivariate normal distribution and some of its properties

5.1 Joint continuous distributions

Earlier, we introduced the concept of the *joint probability distribution* of two discrete random variables X and Y. The idea extends naturally to the case of two continuous random variables X and Y:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dy\ dx.$$

The joint density function must satisfy

$$\begin{aligned} f(x,y) &\geq 0 \text{ for all } x, y \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dy\ dx &= 1. \end{aligned}$$

The marginal distribution of one variable can be obtained from the joint density function:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy.$$

The (joint) cumulative distribution function for (X, Y) is defined by:

$$F(x,y) = P(X \leq x \text{ and } Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dv\ du.$$

Although we will not go into the details further here, we can make corresponding definitions for the joint distribution of mixtures of continuous and discrete random variables.

5.1.1 Independence

Two continuous random variables X and Y are independent if their joint density function can be factorised into the product of their marginal densities, i.e.

$$f(x, y) = f(x)f(y) \text{ for all } x, y$$

5.2 Covariance and correlation

We now introduce two closely related measures of linear dependence, or association, between two random variables.

5.2.1 Covariance

In Session 3 we found an expression for the variance of a sum $X + Y$ of two independent discrete random variables. What happens if we relax the assumption of independence? Let X and Y be two random variables which are not necessarily independent. Then the variance of their sum is found as

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

where we define the covariance between X and Y as:

$$Cov(X, Y) = E\left((X - E(X))(Y - E(Y))\right).$$

Covariance measures the magnitude of *linear* association between X and Y. From its definition we see that if when $X - E(X)$ tends to be positive, $Y - E(Y)$ tends to be positive, then $Cov(X, Y) > 0$. Conversely, if when $X - E(X)$ tends to be positive, $Y - E(Y)$ tends to be negative, then $Cov(X, Y) < 0$ (Figure 5.1).

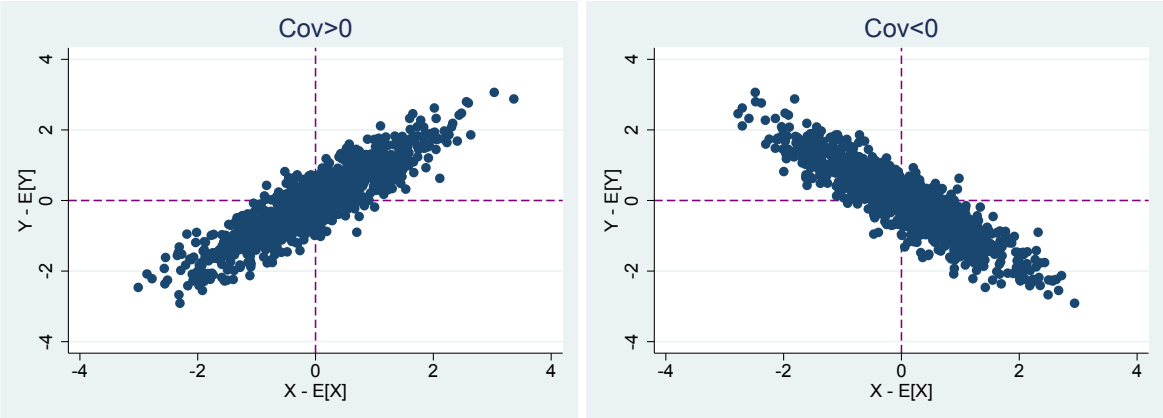

Figure 5.1: X and Y with positive or negative covariance

It is important to remember that covariance is only a measure of linear association. Figure 5.2 shows a plot of a random sample from two variables X and Y which are strongly dependent, yet have zero covariance.

There are various rules which covariance follows, resulting from its definition in terms of expectations.

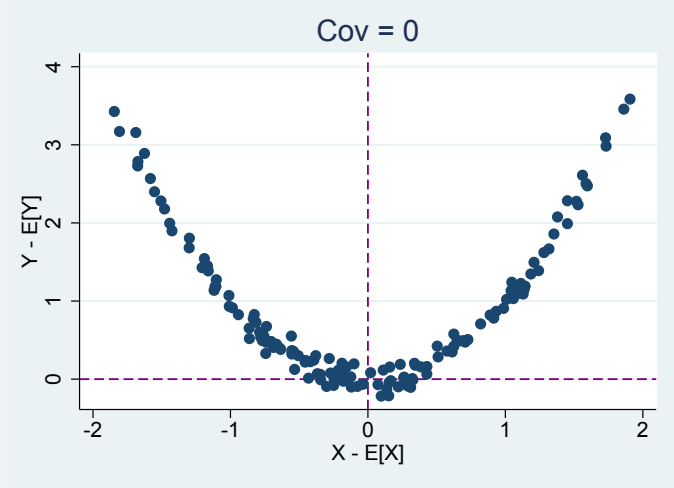

Figure 5.2: X and Y are strongly dependent, but have zero covariance

For example:

$$\begin{aligned} Cov(X,X) &= E((X-E(X))(X-E(X))) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= Var(X). \end{aligned}$$

In fact this result follows from a more general result which is easily proved. Let R, S, X and Y denote random variables, and a, b, c, and d be constants. Then

$$Cov(aR+bS, cX+dY) = acCov(R,X) + adCov(R,Y) + bcCov(S,X) + bdCov(S,Y).$$

By setting $R = X$ and $S = Y$ in this expression we have that:

$$Cov(aX+bY, cX+dY) = acVar(X) + bdVar(Y) + (ad+bc)Cov(Y,X).$$

This result in particular will prove useful in the module 'Analysis of Hierarchical & Other Dependent Data'.

5.2.2 Correlation

The size of $Cov(X,Y)$ depends on the scale/magnitude of variability of X and Y. The correlation between X and Y is equal to their covariance, standardized by their respective standard deviations:

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}.$$

It can be shown that the correlation coefficient lies between -1 and 1. $Corr(X,Y) = 1$ or $Corr(X,Y) = -1$ means X and Y are perfectly correlated. However, it is important to remember that this does **not** necessarily mean X and Y are equal. For example, if $Y = 2X$, they have correlation 1 but are not equal to each other. Conversely, as we saw for covariance (Figure 5.2), it is possible for two variables to be dependent (associated) yet have zero correlation (Figure 5.3).

5.3 Central Limit Theorem

The Central Limit Theorem (CLT) is an important theorem in statistics. It plays a central role in large sample inference theory, as you will find out in the 'Inference' module. In words, the theorem states that if one obtains a random sample of size n from a population and calculates the sample mean

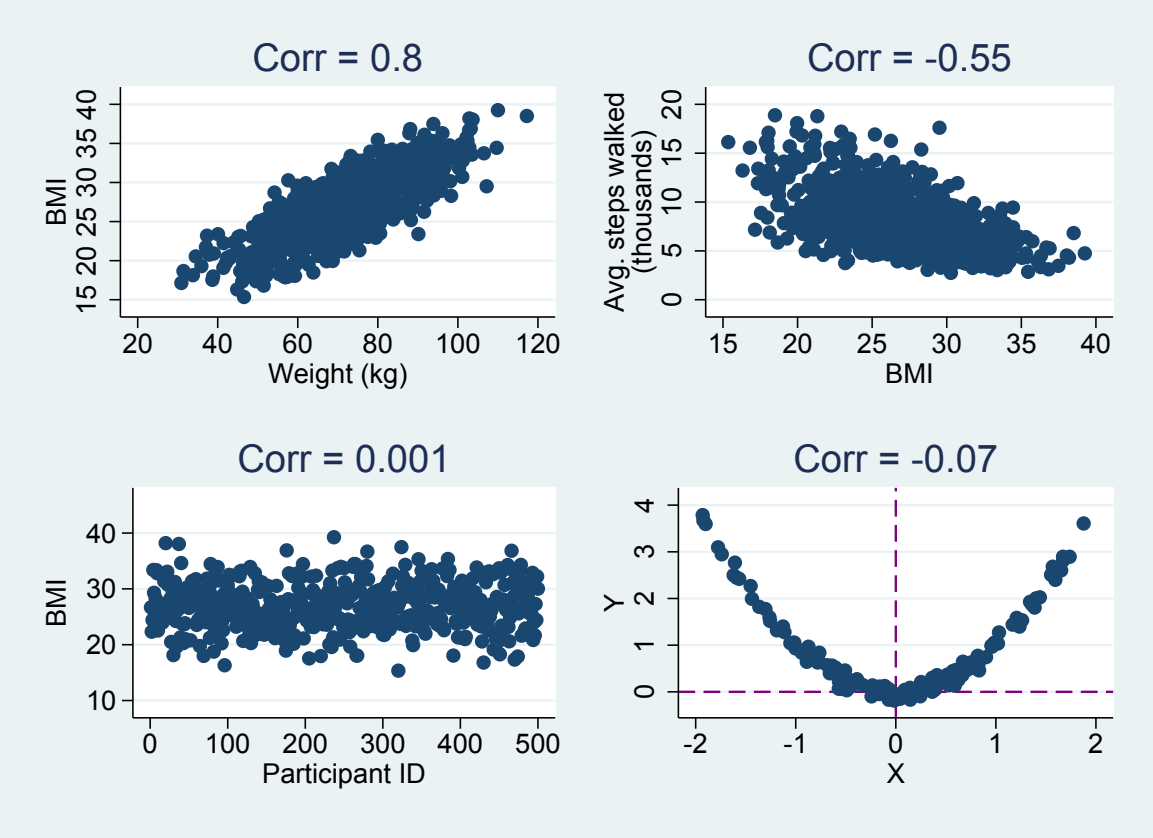

Figure 5.3: Different joint distributions of X and Y and their correlations

$\bar{X}_n$, the *sampling distribution* of $\bar{X}_n$ tends to a normal distribution as $n \to \infty$. More formally, let $X_1, X_2, ..X_n$ be a sequence of iid random variables with mean μ and variance σ^2. Then the distribution of

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

tends to a normal distribution $N(\mu, \sigma^2/n)$ as $n \to \infty$. At the end of Practical 4 you proved that the sample mean has expectation μ and variance σ^2/n. The CLT adds to this by telling us that the distribution of $\bar{X}_n$ is normal, when n is sufficiently large, irrespective of what distribution the individual X_is follow. This holds even if the X_is have a discrete distribution.

5.3.1 The use of the CLT in large sample inference

In the 'Inference' module you will learn about methods for drawing inferences about population parameters based on (usually) random samples from the population. For example, suppose we are interested in estimating the mean of a variable X in a population. Figure 5.4 shows the distribution of X in the population. It is continuous, and quite skewed, and certainly not normally distributed. In order to construct measures of how precise an estimate is we need to know the sampling distribution of the estimate (or estimator) in repeated sampling, which you will learn about in the 'Inference' module. If the distribution of X were normal, the distribution of the mean would also be normal, but here the distribution of X is quite non-normal.

The top left plot in Figure 5.5 shows the sampling distribution of $\bar{X}_n$ with a sample size of $n = 3$. That is, it shows a histogram of the means found by repeatedly sampling from the population of X, with $n = 3$ observations sampled each time. The following plots then show the same for $n = 10$, $n = 30$ and $n = 100$. We see that as the sample size increases, the sampling distribution of $\bar{X}_n$ becomes more and more normal, as predicted by the CLT. Therefore provided our sample size is not too small, for the purposes of conducting inferences about the population mean of X we can assume that the sampling distribution of $\bar{X}_n$ is normal.

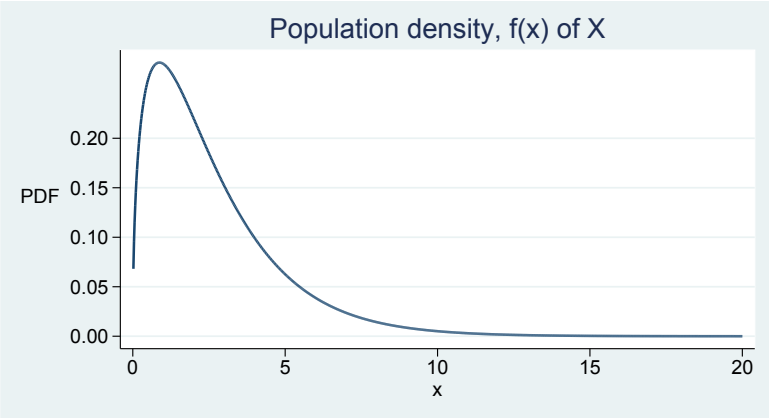

Figure 5.4: Distribution of X, whose mean we wish to estimate, in the population

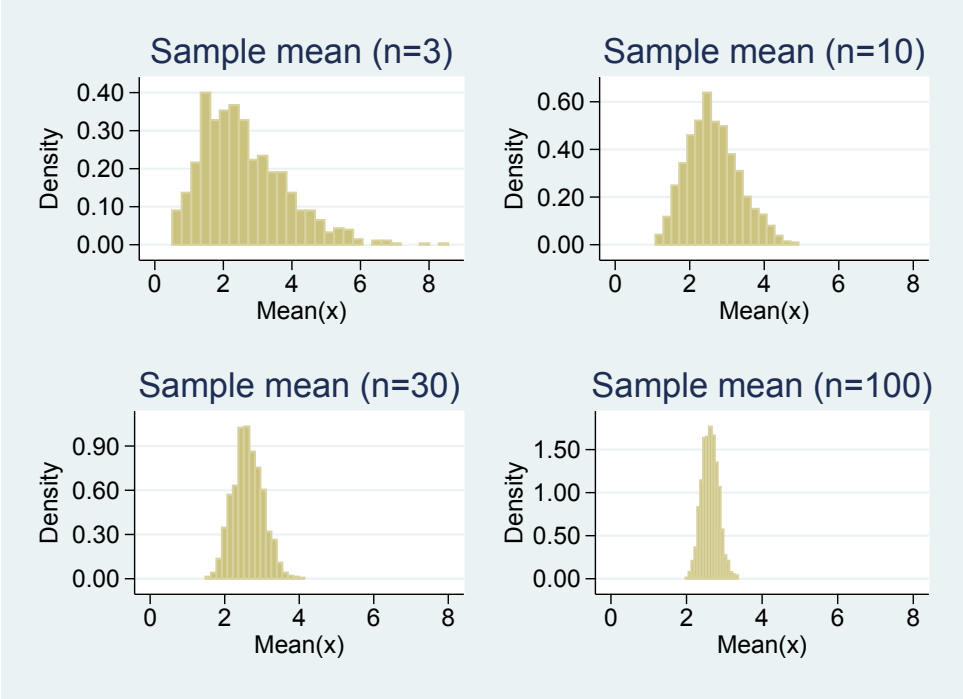

Figure 5.5: Distribution of $\bar{X}_n$, for $n = 3, 10, 30, 100$

5.3.2 Approximating distributions by a normal

Suppose we have a binomial distribution X with large n. Then it will be difficult to calculate the probabilities that X takes a particular value x (unless x is small or close to n) because the binomial coefficient ${}^nC_x$ is then extremely large. One solution is to use the CLT to approximate the distribution of X by a normal. As stated, the CLT tells us how the distribution of the sample mean behaves as $n \to \infty$. An alternative formulation of its result is that

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

as $n \to \infty$.

As in Session 3, we can express $X \sim Bin(n, \pi)$ as

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

where each X_i is an iid Bernoulli trial with success probability π. Then if n is large the CLT says we can approximate the distribution of X as $N(n\mu, n\sigma^2)$, where μ and σ^2 are the mean and variance of the individual X_i Bernoulli trials. Recall that the Bernoulli has expectation π and variance $\pi(1-\pi)$. There, for large n, $X \sim N(n\pi, n\pi(1-\pi))$. In fact the first version of the CLT was derived by DeMoivre in 1733 for the particular case of binomial X with $\pi = 0.5$.

We have approximated a discrete distribution by a continuous one (the normal). What if we are interested in calculating $P(X = x)$ for a given value of x? Recalling that the value of the density function does not give probabilities, we can approximate this probability by calculating $P(x - 0.5 \leq X \leq x + 0.5)$, based on the normal approximation.

We can similarly use the CLT to approximate a Poisson distribution by the normal. Let X be Poisson with expectation/variance μ. Then if we divide the time period under consideration into n equally spaced intervals, we can again express X as $\sum_{i=1}^{n} X_i$ where the X_i are iid Poisson variables with mean μ/n. Then the CLT says we can approximate the distribution of X by a normal $N(\mu, \mu)$.

5.4 The multivariate normal distribution

We have explored the concept of joint distributions, but have not yet explicitly defined any. We conclude the module by introducing the (arguably) most important multivariate continuous distribution, the multivariate normal. You will make use of the multivariate normal distribution extensively in the 'Analysis of Hierarchical & Other Dependent Data' module.

5.4.1 Definition

The vector $\mathbf{X} = (X_1, .., X_n)^T$ (T stands here for matrix transposition) follows a multivariate normal (MVN) distribution if its joint density function is

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

where $\mathbf{x} = (x_1, ., x_n)^T$, $\boldsymbol{\mu} = (\mu_1, .., \mu_n)^T$ denotes an $n \times 1$ parameter vector, and Σ denotes an $n \times n$ matrix parameter. The vector $\boldsymbol{\mu}$ is the vector of means of the components of $\mathbf{X}$, whilst the matrix Σ is the matrix of variances (on the diagonal) of the components and covariances (off the diagonal) between the components. That is,

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} \\ Var(\mathbf{X}) &= \Sigma, \end{aligned}$$

where the variance of the $n \times 1$ vector $\mathbf{X}$ is defined as the $n \times n$ matrix whose diagonals contains the variances and off diagonals contain the covariances.

5.4.2 The bivariate normal distribution

For example, with $n = 2$ we have the bivariate normal distribution. There are two mean parameters contained in $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and three parameters in the covariance matrix Σ:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Here $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$, whilst $Var(X_1) = \sigma_1^2$ and $Var(X_2) = \sigma_2^2$. Lastly $\sigma_{12} = Cov(X_1, X_2)$. Notice that there are only three parameters in the 2×2 covariance matrix Σ. This is because $Cov(X_1, X_2) = Cov(X_2, X_1)$.

Figure 5.6 shows the density function for the bivariate normal, with mean $\boldsymbol{\mu} = (0, 0)^T$ and covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Each component thus has variance 1 and the covariance (and correlation, since the variances are 1) between X_1 and X_2 is 0.5. The z co-ordinate in the plot represents the value of the joint density function, whilst the x and y co-ordinates correspond to the values of x_1 and x_2.

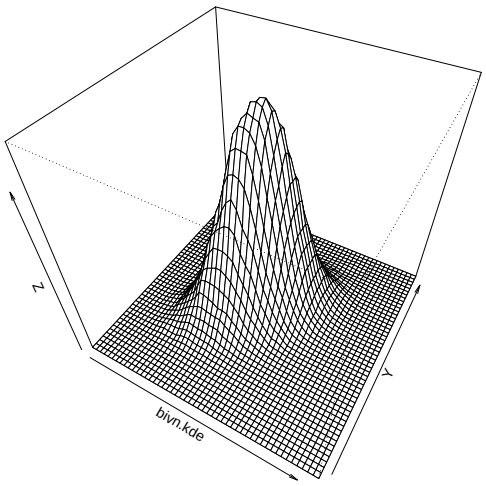

Figure 5.6: The bivariate normal density, with correlation 0.5

5.4.3 Marginal and conditional distributions

One of the nice properties of the MVN is it each of its marginal and conditional distributions is normal. That is, the marginal distribution of (say) X_1 is normal, with mean μ_1 and variance σ_1^2. Similarly, if $\mathbf{X}$ is MVN with n components, the marginal distribution of any subset of the components, e.g. (X_1, X_2, X_3) is also multivariate normal, with means and covariance matrix as given by the corresponding components in $\boldsymbol{\mu}$ and Σ.

Each conditional distribution is also normal. Thus in the case of the bivariate normal, the conditional distribution of X_1 given X_2 is normal, with conditional expectation

$$E(X_1|X_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(X_2 - \mu_2),$$

and conditional variance

$$Var(X_1|X_2) = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}.$$

The conditional variance is defined (in general) as

$$Var(X_1|X_2) = E((X_1 - E(X_1|X_2))^2|X_2).$$

This is the usual definition of variance, except that the expectation is taken conditional on X_2, and we are measuring the squared distance from X_1 to its conditional expectation $E(X_1|X_2)$, as opposed to its marginal expectation $E(X_1)$.

5.4.4 Example: systolic and diastolic blood pressure

Suppose that we model systolic and diastolic blood pressure (BP) using the bivariate normal model. Based on a random sample of data, we estimate that the mean parameter is equal to $\boldsymbol{\mu} = (130, 90)^T$, the standard deviation (SD) of systolic BP is 15 mmHg, the SD of diastolic BP is 10 mmHg, and

the correlation between the two is 0.75. To complete the specification we must find the covariance $\sigma_{12} = Cov(sys, dia)$. This can be found using by re-arranging the definition of correlation as

$$\begin{aligned} Cov(sys, dia) &= Corr(sys, dia) \times SD(sys) \times SD(dia) \\ &= 0.75 \times 15 \times 10 \\ &= 112.5. \end{aligned}$$

Using the properties of the MVN we can now derive both the conditional distribution of *sys* given *dia* and of *dia* given *sys*. For example, the distribution of *sys* given *dia* is normal, with conditional expectation:

$$E(sys|dia) = 130 + \frac{112.5}{10^2}(dia - 90)$$

and the conditional variance is equal to

$$\begin{aligned} Var(sys|dia) &= 15^2 - \frac{112.5^2}{10^2} \\ &= 98.4 \end{aligned}$$

This gives a conditional SD of $\sqrt{98.4} = 9.2$, which, as we should expect, is smaller than the unconditional SD of 15.