MSc Mas6002 Introductory Material

Block A
Introduction to Probability and Statistics

1 Probability

1.1 Multiple approaches

The concept of probability may be defined and interpreted in several different ways, the chief ones arising from the following four approaches.

1.1.1 The classical approach

A game of chance has a finite number of different possible outcomes, which by symmetry are assumed 'equally likely'. The probability of any *event* (i.e. particular outcome of interest) is then defined as the *proportion* of the total number of possible outcomes for which that event *does* occur.

Evaluating probabilities in this framework involves *counting* methods (e.g. permutations and combinations).

1.1.2 The frequency approach

An experiment can be repeated indefinitely under essentially identical conditions, but the observed outcome is random (not the same every time). Empirical evidence suggests that the proportion of times any particular *event* has occurred, i.e. its *relative frequency*, converges to a limit as the number of repetitions increases. This limit is called the *probability* of the event.

1.1.3 The subjective approach

In this approach, an *event* is a statement which may or may not be true, and the *(subjective) probability* of the event is a measure of the degree of belief which the subject has in the truth of the statement. If we imagine that a 'prize' is available if and only if the statement does turn out to be true, the subjective probability can be thought of as the proportion of the prize money which the subject is prepared to gamble in the hope of winning the prize.

1.1.4 The logical approach

Formal logic depends on relationships of the kind $A \longrightarrow B$ ('A implies B') between propositions. The logical approach to probability generalizes the concept of implication

to *partial* implication; the conditional probability of B given A measures the extent to which A implies B. In this approach, all probabilities are 'conditional'; there are no 'absolute' probabilities.

Some pros and cons of the four approaches:

	Classical	Frequency	Subjective	Logical
PRO	Calculation straight-forward	Objective and empirical	Applicable to wide range of 'events'	Extends formal logic in consistent way
CON	Limited in scope	Depends on infinite repeatability	Depends on individual	Yields no way of assigning probabilities

1.2 Axiomatic probability theory

Because there is not a uniquely best way of defining probabilities, it is customary to lay down a set of rules (axioms) which we expect probabilities to obey (whichever interpretation is put on them). A mathematical theory can then be developed from these rules. The framework is as follows. For any probability model we require a *probability space* (Ω, F, P) with

- the set Ω of all possible outcomes, known as the *sample space* (for the game, experiment, etc.)
- a collection F of subsets of Ω, each subset being called an *event* (i.e. the event that the observed outcome lies in that subset).
- a *probability measure* defined as a real valued function P of the elements of F (the events) satisfying the following *axioms of probability:*

  A1 $P(A) \geq 0$ for all $A \in F$

  A2 $P(\Omega) = 1$

  A3 If A_1, A_2,... are *mutually exclusive* events (i.e. have no elements in common), then $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$ [The sum may be finite or infinite].

Example 1. In a 'classical' probability model, Ω consists of the n equally likely outcomes $\{a_1, a_2, ..., a_n\}$ say, F consists of *all* subsets of Ω, and P is defined by

$$P(A) = \frac{\text{no. of elements in } A}{n} \quad \text{for} \quad A \in F.$$

It is easy to show that the P satisfies the axioms of probability.

Theorems about probabilities may be proved using the axioms. The following is a simple example.

Theorem 1. If $A^c = \Omega \backslash A$, (the 'complement' or 'negation' of A), then $P(A^c) = 1 - P(A)$.

Proof. A and A^c are mutually exclusive with union Ω. Therefore

$$\begin{aligned} P(A) + P(A^c) &= P(\Omega) && \text{(by axiom A3)} \\ &= 1 && \text{(by axiom A2).} \end{aligned}$$ □

1.3 Conditional probability

For any event A with $P(A) > 0$, we can 'condition' on the occurrence of A by defining a new probability measure, P_A say, which is obtained from P by reducing all probability outside A to zero and rescaling the rest so that the axioms are still satisfied. Thus for any event B in the original sample space Ω, we define

$$P_A(B) = \frac{P(A \cap B)}{P(A)} = \text{probability of } \ B \ \textit{conditional on} \ A \ \text{or} \ \textit{given} \ A$$

$P_A(B)$ is normally written $P(B \mid A)$.

Conditional probabilities are often easier to specify than unconditional ones, and the above definition may be rearranged to give

$$P(A \cap B) = P(A)P(B \mid A)$$

which is sometimes known as the *multiplication rule*. This may be extended (by an easy induction argument) to a sequence $A_1, A_2, ..., A_n$ of events

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

which is useful when it is easy to write down the probability of each event in the sequence conditional on all the previous ones having occurred.

Another important relationship involving conditional probabilities is the *law of total probability* (sometimes known as the *elimination rule*). This involves the notion of a *partition of the sample space*, which is a sequence $A_1, A_2, \ldots$ (finite or infinite) such that

$$A_1 \cup A_2 \cup \ldots = \Omega$$

and $A_i \cap A_j = \phi$ whenever $i \neq j$. In other words, '$A_1, A_2, \ldots$ are mutually exclusive and exhaustive' or alternatively 'one and only one of $A_1, A_2, \ldots$ must occur'. If $A_1, A_2, \ldots$ is a partition and B is an arbitrary event, then the law of total probability states that

$$P(B) = \sum_i P(A_i)P(B \mid A_i)$$

This follows from the axioms and the definition of conditional probability, since $A_1 \cap B, A_2 \cap B, \ldots$ are mutually exclusive with union B, and $P(A_i \cap B) = P(A_i)P(B \mid A_i)$ for each i.

Example 2. In a multiple choice test, each question has m possible answers. If a candidate knows the right answer (which happens with probability p) he gives it; if he thinks he knows it but is mistaken (which has probability q) he gives the answer he thinks is

correct; and if he does not think he knows it (with probability $r = 1 - p - q$) then he chooses an answer at random. What is the probability that he answers correctly?

Let A_1 be the event that the candidate knows the right answer, A_2 the event that he thinks he knows it but is mistaken, and A_3 the event that he does not think he knows it. Let B be the event that he answers correctly.

$$\begin{aligned} P(A_1) &= p & P(B \mid A_1) &= 1 \\ P(A_2) &= q & P(B \mid A_2) &= 0 \\ P(A_3) &= r & P(B \mid A_3) &= 1/m \end{aligned}$$

therefore

$$P(B) = p \times 1 + q \times 0 + r \times \frac{1}{m} = p + \frac{r}{m}.$$

Sometimes we wish to relate two conditional probabilities. This is easily achieved

$$P(A \mid B) = \frac{P(B \cap A)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

This result is known as *Bayes' Theorem* (or Bayes' rule). It is often used in conjunction with a partition where we wish to know the probability of one (or more) of the events of the partition having occurred conditional on the 'secondary event' B having been observed to occur:

$$P(A_j \mid B) = \frac{P(A_j)P(B \mid A_j)}{P(B)} = \frac{P(A_j)P(B \mid A_j)}{\Sigma P(A_i)P(B \mid A_i)}$$

Example 3. (Model as above). What is the probability that the candidate knew the right answer given that he answered correctly?

$$P(A_1 \mid B) = \frac{p \times 1}{p \times 1 + q \times 0 + r \times \frac{1}{m}} = \frac{p}{p + \frac{r}{m}}$$

Bayes' Theorem lies at the foundation of a whole branch of statistics – Bayesian statistics. See MAS6004.

1.4 Independence

Event B is said to be *independent* of event A if

$$P(B \mid A) = P(B)$$

i.e. conditioning on A does not affect the probability of B. The relationship is more usually written equivalently as

$$P(A \cap B) = P(A)P(B)$$

which indicates that the relationship is symmetrical: A and B are mutually independent. This form my be generalized to longer sequences: events $A_1, A_2, \ldots$ are said to be (mutually) independent if

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k})$$

for *any* finite collection of distinct subscripts $i_1, i_2, \ldots i_k$.

Example 4. For three events A, B, C to be mutually independent, all the following relationships must hold:

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(A \cap C) &= P(A)P(C) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned}$$

Fortunately, independence is usually used as an *assumption* in constructing probability models, and so we do not need to check a whole collection of relationships such as the above.

1.5 Worked examples

Example 5. Consider the roll of an ordinary six-sided die. Then the sample space Ω is the set of outcomes, i.e. $\{1, 2, 3, 4, 5, 6\}$.

Let A be the event that we roll an even number, and let B be the event that we roll at least 4. As subsets of Ω, $A = \{2, 4, 6\}$ and $B = \{4, 5, 6\}$.

Now look at the various set operations applied to these two events:

- $A \cup B = \{2, 4, 5, 6\}$: we roll an even number *or* we roll at least 4. (NB inclusive "or": both 4 and 6 are included.)
- $A \cap B = \{4, 6\}$: we roll an even number *and* we roll at least 4, i.e. we roll 4 or 6.
- $A^c = \{1, 3, 5\}$: we do not roll an even number, i.e. we roll an odd one.

If the die is fair we can assume that the elements of Ω are equally likely. We can then work out probabilities of events by simply counting the number of elements and dividing by the total number of elements in Ω (here 6): $P(A) = 3/6 = 1/2$, $P(B) = 3/6 = 1/2$, $P(A \cup B) = 4/6 = 2/3$, etc.

Example 6. A bag contains 8 balls labelled $1, 2, \ldots, 8$. Three balls are drawn randomly without replacement. Calculate (a) the probability that the ball labelled 4 is drawn, (b) the probability that the three balls drawn have consecutive numbers (e.g. 3,4,5). Are the events in (a) and (b) independent?

The number of outcomes here is the number of ways of selecting 3 from 8, i.e. $\binom{8}{3} = 56$. (The order that the balls are drawn in is not important; otherwise the number of outcomes would be $8 \times 7 \times 6 = 336$.) (a) If ball 4 is drawn, there are $\binom{7}{2} = 21$ possibilities for the other two, so, given equally likely outcomes, the probability is $21/56 = 3/8$. (b) There are six outcomes here, so the probability is $6/56 = 3/28$.

Call the events A and B respectively. Then $A \cap B$ requires the balls to be $\{2, 3, 4\}$, $\{3, 4, 5\}$ or $\{4, 5, 6\}$, and hence has probability $3/56$. But $P(A)P(B) = 9/224 \neq 3/56$, so the two events are not independent. (The conditional probability $P(A|B) = P(A \cap B)/P(B) = (3/56)/(3/28) = 1/2 > 3/8$, so B occurring increases the chance of A occurring.)

2 Random variables

2.1 Definition

Often the outcome of an experiment will have *numerical* values, e.g. throwing a die we can take the sample space to be $\Omega = \{1, 2, 3, 4, 5, 6\}$. In other more complicated experiments even though the outcomes may not be numerical we may be interested in a numerical value which is a function of the observed outcome, e.g. throwing three dice we may be interested in 'the total score obtained'. In this latter example, the sample space may be described as

$$\Omega = \{(1,1,1), (1,1,2), (1,1,3), \ldots\ldots (6,6,6)\}$$

but the possible values which 'the total score obtained' can take are given by

$$\Omega_X = \{3, 4, 5, \ldots 18\}$$

which is called the *induced sample space*. The function which takes Ω into Ω_X according to the rule described is called a *random variable*, e.g. 'the total score obtained' is a random variable which we may call X say. Associated with the sample space induced by X are:

(i) F_X = events 'generated' by X (e.g. 'X=4' and 'X is an even number' are events);

(ii) P_X = induced probability measure, known as the *distribution* of X, e.g. $P_X(3,4) = P(X = 3 \text{ or } 4) = P\{(1,1,1), (1,1,2), (1,2,1)(2,1,1)\} = \frac{4}{216}$ assuming dice are fair.

It is often convenient to work with (Ω_X, F_X, P_X), rather than with (Ω, F, P), but we have to be careful if considering more than one random variable defined on the *same* underlying sample space.

2.2 Types of distribution

A random variable X, or its distribution, is called *discrete* if it only takes values in the integers or (possibly) some other countable set of real numbers. (This will automatically be true if the underlying sample space is countable.) In this case the distribution is entirely specified by giving its value on every singleton in the induced sample space:

$$P_X(\{i\}) = p_X(i) \text{ for } i \in \Omega_X.$$

$p_X(i)$ is called the *probability function* of X (or of its distribution). The probability of any event A_X in F_X is then found by summing the values of p_X over the singletons contained in A_X:

$$P(A_X) = \sum_{i \in A_X} p_X(i) \tag{1}$$

In order to satisfy the axioms of probability, it is sufficient that the function p_X satisfies

(i) $p_X(i) > 0$ for all $i \in \Omega_X$ (to satisfy A1)

(ii) $\sum_{i \in \Omega_X} p_X(i) = 1$ (to satisfy A2)

A3 is then automatically satisfied because of (1).

Example 7.

$$p_X(i) = \frac{1}{i(i+1)} \text{ with } \Omega_X = 1, 2, 3, \ldots$$

is a probability function since $p_X(i) > 0$ obviously and also

$$\sum_{i=1}^{\infty} p_X(i) = \sum_{i=1}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1} \right) = 1.$$

A random variable X, or its distribution, is called *(absolutely) continuous* if it takes values on the whole real line or some sub-intervals and the probability that it lies in any interval $(a, b]$ say is given by the *integral* over the interval of a function, known as the *probability density function* (p.d.f.)f_X:

$$P(a < X \leq b) = \int_a^b f_X(x)\ dx\ = P_X((a, b]).$$

Note that it does not matter whether we include the endpoints of the interval or not; the probability of any singleton is zero.

In order to satisfy the axioms of probability it is sufficient that the function f_X satisfies

1. $f_X(x) \geq 0$ for all x;
2. $\int_{\Omega_X} f_X(x)dx = 1$.

Example 8. For what value of c is the function

$$f_X(x) = \frac{c}{1+x^2}$$

a p.d.f.?

Obviously (i) is satisfied provided $c \geq 0$; to satisfy (ii) we must have

$$1 = \int_{-\infty}^{\infty} \frac{c}{1+x^2} dx = c \left[\tan^{-1} x \right]_{-\infty}^{\infty} = c \left[\frac{\pi}{2} - \left(\frac{-\pi}{2} \right) \right] = c\pi, \text{therefore c} = \frac{1}{\pi}.$$

Note. Not all distributions are either discrete or absolutely continuous; for example, a distribution may be *partly* discrete and *partly* absolutely continuous.

The *distribution function* F_X of any distribution is defined as

$$F_X(x) = P(X \leq x) = P_X((-\infty, x]) \text{ for all real } x.$$

If the distribution is discrete, it will be a step function:

$$F_X(x) = \sum_{i \leq x} p_X(i) \text{ for } i \in \Omega_X$$

whereas if the distribution is absolutely continuous, it will be a smooth function with derivative given by the p.d.f.:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\ du$$

therefore $F'_X(x) = f_X(x)$. Because of the axioms of probability, a distribution function is always *non-decreasing*, *continuous from the right*, and satisfies

$$\begin{array}{lll} F_X(x) \downarrow 0 & \text{as} & x \to -\infty \\ F_X(x) \uparrow 1 & \text{as} & x \to +\infty \end{array}$$

Example 9. For Example 7 (here $[x]$ denotes 'whole number part of x', so $[3.2] = 3$)

$$F_X(x) \;=\; \begin{cases} \displaystyle\sum_{i=1}^{[x]} \left(\frac{1}{i} \;-\; \frac{1}{i+1}\right) \;=\; 1 \;-\; \frac{1}{[x]+1} & \text{for} \quad x \geq 0 \\ 0 & \text{for} \quad x < 0 \end{cases}$$

and for Example 8

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+u^2)} du = \frac{1}{2} + \frac{1}{\pi} \;\tan^{-1} x \quad (-\infty < x < \infty).$$

Note. In the discrete case where $\Omega_X = \{0, 1, 2, \ldots\}$ say, we can 'recover' p_X from F_X using the fact that $p_X(i) = F_X(i) - F_X(i-1)$ (difference of successive values).

Note. We usually drop subscripts on p, f and F if it is clear to which variable they refer.

2.3 Expectation and more general moments

If X is a random variable, then its *expectation, expected value* or *mean* $E(X)$ is the number defined by

$$E(X) = \begin{cases} \displaystyle\sum_{i \in \Omega_X} i p_X(i) & \text{(discrete case)} \\ \displaystyle\int_{\Omega_X} x f_X(x) dx & \text{(abs. cont. case)} \end{cases}$$

provided that the sum or integral converges (otherwise the expectation does not exist). It is a weighted *average* of the possible values which X can take, the weights being determined by the distribution. It measures where the *centre* of the distribution lies.

Properties

- If $X > 0$ always, then $E(X) > 0$.
- If $X \equiv c$ (constant) then $E(X) = c$.
- If a and b are constants then $E(a + bX) = a + bE(X)$ (linearity).

If $g(X)$ is a function of a random variable, then, to evaluate $E(g(X))$, it is unnecessary to know the distribution of $g(X)$ because it may be shown that

$$E(g(X)) = \begin{cases} \displaystyle\sum_{i \in \Omega_X} g(i) p_X(i) & \text{(discrete case)} \\ \displaystyle\int_{\Omega_X} g(x) f_X(x) dx & \text{(abs. cont. case)} \end{cases}$$

Of particular importance are *moments* of a random variable, such as the following:

$$\begin{aligned} E(X^r) &= r\text{th moment of } X \\ E((X-\mu_X)^r) &= r\text{th moment of } X \text{ about its mean} \\ &\quad (\text{where } \mu_X = E(X)) \\ E(X(X-1)\dots(X-r+1)) &= r\text{th factorial moment} \end{aligned}$$

The second moment of X about its mean is called the *variance* of X

$$\text{Var}(X) = E(X-\mu_X)^2 = \sigma_X^2$$

and measures the extent to which the distribution is *dispersed* about μ_X. Its positive square root σ_X is called the *standard deviation (s.d.)*.

Properties of variance

- $\text{Var}(X) \geq 0$, with equality if and only if X is a constant random variable
- $\text{Var}(a+bX) = b^2 \,\text{Var}\, X$.

The mean and variance of a distribution are commonly used measures of 'location' and 'dispersion' but there are other possibilities, e.g. the *median* is defined as any value η such that $F_X(\eta) \geq \frac{1}{2}$ and $F_X(\eta^-) \leq \frac{1}{2}$ (η may not be unique) and is a measure of location (the half-way point of the distribution), and the *interquartile range* is defined as

$$\eta_{\frac{3}{4}} - \eta_{\frac{1}{4}}$$

where the upper and lower quartiles used here are defined as the median but with 1/2 replaced by 3/4 and 1/4 respectively; the interquartile range is a measure of dispersion.

Examples of moments will follow in the next section.

2.4 Some standard distributions

2.4.1 The binomial distribution $Bi(n,\theta)$

This is the discrete distribution with probability function given by

$$p(i) = \begin{cases} \binom{n}{i}\theta^i(1-\theta)^{n-i} & \text{for } 0 \leq i \leq n \\ 0 & \text{otherwise} \end{cases}$$

n and θ are the **parameters** of the distribution. Here n is a positive integer and θ a real number between 0 and 1. It arises as the distribution of the number of 'successes' in n independent 'Bernoulli trials' at each of which there is probability of 'success' θ. A combinatorial argument leads to the above formula.

If X has $Bi(n,\theta)$ distribution then we write $X \sim Bi(n,\theta)$ and find

$$E(X) = n\theta \text{ and } \text{Var}(X) = n\theta(1-\theta).$$

2.4.2 The Poisson distribution $Po(\mu)$.

This is the discrete distribution with probability function given by

$$p(i) = \begin{cases} e^{-\mu}\dfrac{\mu^i}{i!} & \text{for } i = 0, 1, 2, \ldots \\ \\ 0 & \text{otherwise} \end{cases}.$$

Here μ is the **parameter** of the distribution, and is a positive real number. It arises as the distribution of the number of 'occurrences' in a time period during which in a certain sense these 'occurrences' are completely random.

The Poisson distribution with parameter μ has mean μ and variance μ. (See Example 10 below.)

One way of deriving the Poisson probability function is to take the limit of $Bi(n, \theta)$ as $n \to \infty$ and $\theta \to 0$ but $n\theta$ remains fixed at μ. If we substitute $\theta = \mu/n$ then the binomial probability function is

$$\begin{aligned} \frac{n!}{i!(n-i)!}\left(\frac{\mu}{n}\right)^i\left(1-\frac{\mu}{n}\right)^{n-i} &= \frac{n.(n-1)\ldots(n-i+1)}{n.n\ldots n}\frac{\mu^i}{i!}\left(1-\frac{\mu}{n}\right)^{-i}\left(1-\frac{\mu}{n}\right)^{n} \\ &\longrightarrow 1 \times \frac{\mu^i}{i!} \times 1 \times e^{-\mu} \quad \text{as} \quad n \to \infty. \end{aligned}$$

For this reason, the binomial distribution $Bi(n, \theta)$ is well approximated by $Po(n\theta)$ when *n is large* — the approximation is most useful when *θ is small*, so that $n\theta$ is not large.

The distribution functions of the binomial and Poisson distribution are tabulated for various values of the parameters (Neave tables 1.1 and 1.3(a) respectively).

Example 10. Calculate the mean and variance of a $Po(\lambda)$ random variable X.

We have

$$\begin{aligned} E(X) &= \sum_{n=0}^{\infty} n p_X(n) \\ &= \sum_{n=0}^{\infty} n e^{-\lambda}\frac{\lambda^n}{n!} \\ &= \sum_{n=1}^{\infty} e^{-\lambda}\frac{\lambda^n}{(n-1)!} && \text{(the } n = 0 \text{ term is zero)} \\ &= \lambda e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} && \text{(changing variables to } m = n-1) \\ &= \lambda e^{-\lambda}e^{\lambda} = \lambda \end{aligned}$$

For the variance, start off by calculating

$$E(X(X-1)) = \sum_{n=2}^{\infty} n(n-1)e^{-\lambda}\frac{\lambda^n}{n!}.$$

(The sum can start from $n = 2$ as the first two terms would be zero.) Change variables to $m = n - 2$, using the fact that $n! = n(n-1)(n-2)!$, giving

$$E(X(X-1)) = \lambda^2 \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} = \lambda^2.$$

Then $E(X^2) = E(X(X-1)) + E(X) = \lambda^2 + \lambda$, and

$$\mathrm{Var}(X) = E(X^2) = E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

2.4.3 The normal distribution $N(\mu, \sigma^2)$

The normal distribution with parameters μ and σ^2 is the continuous distribution with p.d.f. given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } -\infty < x < \infty$$

It is a symmetrical bell-shaped distribution of great importance. It has mean μ and variance σ^2.

If X has $N(\mu, \sigma^2)$ distribution then $(X - \mu)/\sigma$ has **standard** normal distribution N(0,1), whose density is given the special symbol ϕ and likewise its distribution function Φ. Both Φ and its inverse are tabulated (Neave tables 2.1, 2.3). Any probability involving a $N(\mu, \sigma^2)$ random variable may be obtained from these tables, e.g.

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

2.4.4 The exponential distribution $Ex(\lambda)$, and the Gamma distribution $Ga(\alpha, \lambda)$

Before introducing these distributions, we define the Gamma function

$$\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u}\, du.$$

Integration by parts shows that

$$\begin{aligned} \Gamma(\alpha) &= \Big[-u^{\alpha-1} e^{-u} \Big]_0^{\infty} + \int_0^{\infty} (\alpha - 1) u^{\alpha-2} e^{-u}\, du \\ &= 0 + (\alpha - 1)\Gamma(\alpha - 1) \\ &= (\alpha - 1)\Gamma(\alpha - 1). \end{aligned}$$

Also

$$\Gamma(1) = \int_0^{\infty} e^{-u}\, du = \Big[-e^{-u} \Big]_0^{\infty} = 1.$$

Using the above two results, we can see that for a positive integer α,

$$\Gamma(\alpha) = \prod_{i=1}^{\alpha-1} i = (\alpha-1)!.$$

For example, $\Gamma(3) = 2! = 2$, $\Gamma(7) = 6! = 720$. Another value which is relevant in the theory of probability distributions is $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$).)

The exponential distribution with parameter λ is the continuous distribution with p.d.f. given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \quad (\lambda > 0) \\ 0 & \text{for } x < 0 \end{cases}$$

It occurs commonly as the distribution of a "waiting time" in various processes. It has mean $1/\lambda$ and variance $1/\lambda^2$ (see Example 11 below). Its distribution function is given by

$$F(x) = \begin{cases} \displaystyle\int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x} & \text{for } x \geq 0; \\ 0 & \text{for } x < 0 \end{cases}.$$

Example 11. Calculate the mean and variance of a $Ex(\lambda)$ random variable Y.

We have

$$E(Y) = \int_0^\infty \lambda x e^{-\lambda x}\, dx = \frac{1}{\lambda}\int_0^\infty u e^{-u}\, dx = \frac{1}{\lambda}\Gamma(2),$$

by changing variables to $u = \lambda x$ and the definition of the Gamma function. As $\Gamma(2) = 1$, the mean is $1/\lambda$.

For the variance, calculate

$$E(Y^2) = \int_0^\infty \lambda x^2 e^{-\lambda x}\, dx = \frac{1}{\lambda^2}\Gamma(3)$$

by the same change of variables. As $\Gamma(3) = 2$, we get

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

The Gamma distribution with parameters α and λ is a generalization of the exponential distribution with p.d.f. given by

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{for } x > 0 \quad (\lambda, \alpha > 0). \\ 0 & \text{for } x \leq 0 \end{cases}$$

The appearance $\Gamma(\alpha)$ here ensures that the p.d.f. integrates to 1.

It has mean α/λ and variance α/λ^2. The exponential is the special case $\alpha = 1$; another special case of interest in statistics is the case $\alpha = \nu/2, \lambda = 1/2$, where ν is a positive integer, which is known as the χ^2 **distribution with ν degrees of freedom,** χ^2_ν.

2.5 Transformations of random variables

2.5.1 New densities from old

If X is a random variable then so is $Y = g(X)$ for any function g, and its distribution is related to that of X by

$$P_Y(B) = P(g(X) \in B) = P_X(g^{-1}(B))$$

for $B \subseteq \Omega_Y$, where $g^{-1}(B)$ denotes the inverse image of B under g. In many cases such inverse images are easy to identify, e.g. if g is a *continuous increasing* function and $B = (-\infty, y]$ say, then here the distribution function of Y is given by

$$F_Y(y) = P_Y((-\infty, y]) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

and if X has p.d.f. f_X and g is also differentiable, then Y is also absolutely continuous with p.d.f.

$$f_Y(y) = F_Y'(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)].$$

Example 12. If X has $Ex(\lambda)$ distribution and $Y = \sqrt{X}$ then the above conditions are satisfied (since $X \geq 0$ always) and $g(x) = \sqrt{x}, g^{-1}(y) = y^2$ for $y \geq 0$. Therefore

$$f_Y(y) = f_X(y^2) \frac{d}{dy} y^2 = \lambda e^{-\lambda y^2} 2y = 2\lambda y e^{-\lambda y^2}$$

for $y \geq 0$.

2.5.2 Approximating moments

It is sometimes useful to approximate moments of functions of random variables in terms of moments of the original random variables as follows: suppose the distribution of X is such that the first term Taylor expansion

$$g(X) = g(\mu) + g'(\mu)(X - \mu),$$

where $\mu = E(X)$, is a good approximation over the bulk of the range of X. It then should follow that $E(g(X)) \approx g(\mu)$ since $E(X - \mu) = 0$, and $\operatorname{Var} g(X) \approx [g'(\mu)]^2 \operatorname{Var} X$.

This approximation will be useful if g is smooth at μ.

Example 13. If X has $Ga(\alpha, \lambda)$ distribution, then

$$E(\log X) \approx \log(\alpha/\lambda) \quad \text{and} \quad \operatorname{Var}(\log X) \approx \left(\frac{\lambda}{\alpha}\right)^2 \frac{\alpha}{\lambda^2} = \frac{1}{\alpha}.$$

Example 14. Suppose X has $Bi(n, \theta)$ distribution. Then θ might be estimated[1] by X/n and the odds $\theta/(1-\theta)$ could be estimated by $(X/n)/(1-(X/n)) = X/(n-X)$. The log odds, $\log(\theta/(1-\theta)) = \log\theta - \log(1-\theta)$, can be estimated by $\log(X/n/(1-X/n)) =$

[1] Estimation is developed more fully later, in Section 4

$\log(X) - \log(n - X)$. Now a natural question is what is the (approximate) variance of this estimator.

Here $EX = n\theta$, $\mathrm{Var}(X) = n\theta(1-\theta)$ and $g(x) = \log(x) - \log(n-x)$. Thus

$$g'(x) = \frac{1}{x} + \frac{1}{n-x}.$$

Hence

$$E[\log(X) - \log(n - X)] \approx \log(n\theta) - \log(n - n\theta) = \log(\theta/(1-\theta))$$

and

$$\begin{aligned}\mathrm{Var}[\log(X) - \log(n - X)] &\approx \left(\frac{1}{n\theta} + \frac{1}{n - n\theta}\right)^2 n\theta(1-\theta) \\ &= \frac{1}{n\theta(1-\theta)} = \frac{1}{n\theta} + \frac{1}{n(1-\theta)}.\end{aligned}$$

Now, if you wanted to estimate this variance, you would have to replace θ by its estimator X/n, giving an estimated approximate variance of

$$\frac{1}{n(X/n)} + \frac{1}{n(1-(X/n))} = \frac{1}{X} + \frac{1}{n-X}.$$

This approximation should work reasonably well provided neither X nor $n - X$ is too small. Obviously it breaks down if either of these is zero.

2.6 Independent random variables

Two random variables X and Y are said to be independent of each other if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any choice of events A and B. This clearly extends the notion of independence of events. The generalization to more than two random variables follows: $X_1, X_2, \ldots$ (a finite or infinite sequence) are independent. If the events $X_1 \in A_1$, $X_2 \in A_2$, $\ldots X_n \in A_n$ are independent for all n and for all $A_1 \in F_{X_1}, \ldots A_n \in F_{X_n}$.

Often we use a sequence of independent random variables as the starting point in constructing a model. It can be shown that it is always possible to construct a probability space which can carry a sequence of independent random variables with given distributions.

Example 15. In a sequence of n Bernoulli trials with probability of success θ , let

$$U_k = \begin{cases} 1 & \text{if success occurs on } k\text{th trial } (1 < k < n) \\ 0 & \text{otherwise.} \end{cases}$$

Then $U_1, U_2, \ldots U_n$ are independent (because the trials are independent) and each has the same distribution, given by the probability function

$$\begin{aligned} P_U(0) &= 1 - \theta \\ P_U(1) &= \theta \\ P_U(i) &= 0, \ i \neq 0, \ i \neq 1 \end{aligned}$$

Note also that $E(U_k) = \theta$ and $\mathrm{Var}(U_k) = \theta(1-\theta)$ for each k.

If we have several random variables defined on the same probability space, then we can form functions of them and create new random variables, e.g. we can talk about the sum of them. The following results are important.

(i) $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$.

If $X_1, X_2, \ldots, X_n$ are *independent* then in addition:

(ii) $E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2)\ldots E(X_n)$

and

(iii) $\text{Var}(X_1 + X_2 + \ldots + X_n) = \text{Var}\, X_1 + \text{Var}\, X_2 + \ldots + \text{Var}\, X_n$.

Example 16. (Continuing Example 15) $U_1 + U_2 + \ldots + U_n$ = total no. of successes which has $Bi(n, \theta)$ distribution. Using (i),

$$E(U_1 + U_2 + \ldots + U_n) = n\theta;$$

using (iii)

$$\text{Var}(U_1 + U_2 + \ldots + U_n) = n\theta(1 - \theta)$$

— confirming these facts for $Bi(n, \theta)$.

Often if each of a sequence of independent random variables has a standard form, then their sum has a related standard form. The following summary gives examples, it being understood in each case that the two random variables on the left are independent, and the notation being self-explanatory.

$$\begin{array}{lrcl}
\text{(i)} & Bi_1(n_1, \theta) + Bi_2(n_2, \theta) & = & Bi_3(n_1 + n_2, \theta) \\
\text{(ii)} & Po_1(\mu_1) + Po_2(\mu_2) & = & Po_3(\mu_1 + \mu_2) \\
\text{(iii)} & N_1(\mu_1, \sigma_1^2) + N_2(\mu_2, \sigma_2^2) & = & N_3(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\
\text{(iv)} & Ga_1(\alpha_1, \lambda) + Ga_2(\alpha_2, \lambda) & = & Ga_3(\alpha_1 + \alpha_2, \lambda)
\end{array}$$

3 Multivariate random variables and distributions

3.1 General Concepts

If $X_1, X_2, \ldots, X_k$ are random variables, not necessarily independent, defined on the same sample space, then the induced sample space is (a subset of) k-dimensional space, $\mathbb{R}^k$, and the induced probability measure is called the *joint distribution* of $X_1, X_2, \ldots, X_k$. Alternatively we can think of $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ as a random vector.

The joint distribution may be *discrete*, i.e. given by

$$P\{(X_1, X_2, \ldots, X_k) \in B\} = \sum_{(i_1, i_2, \ldots, i_k) \in B} \sum \cdots \sum p_{X_1, X_2, \ldots, X_k}(i_1, i_2, \ldots, i_k) \qquad (B \subseteq \mathbb{R}^k)$$

where $p_{X_1, X_2, \ldots, X_k}$ is called the *joint probability function* of $X_1, X_2, \ldots, X_k$.

It may also be *absolutely continuous*, i.e. given by

$$P\{(X_1, X_2, \ldots, X_k) \in B\} = \int_B \int \cdots \int f_{X_1, X_2, \ldots, X_k}(x_1, x_2, \ldots, x_k) dx_1 dx_2 \ldots dx_k$$

where $f_{X_1,X_2,\ldots,X_k}$ is called the *joint probability density function* of $X_1, X_2, \ldots, X_k$.

There are many other possibilities, e.g. it might be discrete in the first variable and absolutely continuous in the second variable. For simplicity of notation we shall talk mainly about two random variables (X, Y) say. Most of the concepts generalize naturally.

3.1.1 Discrete case

The joint probability function

$$p_{X,Y}(i,j) = P(X = i, Y = j)$$

must satisfy $p_{X,Y}(i,j) \geq 0$ and

$$\sum_{i \in \Omega_X} \sum_{j \in \Omega_Y} p_{X,Y}(i,j) = 1.$$

The *marginal* probability function of X (i.e. its probability function in the usual sense) is found by summing over $j \in \Omega_Y$ while keeping $i \in \Omega_X$ fixed:

$$p_X(i) = \sum_{j \in \Omega_Y} p_{X,Y}(i,j).$$

Similarly $p_Y(j) = \sum_{i \in \Omega_X} p_{X,Y}(i,j)$.

If (and only if) X and Y are independent then $p_{X,Y}$ *factorizes* into the marginals:

$$p_{X,Y}(i,j) = p_X(i) p_Y(j).$$

More generally we can define the conditional probability functions

$$p_{X|Y}(i \mid j) = P(X = i \mid Y = j) = \frac{p_{X,Y}(i,j)}{p_Y(j)}$$

and similarly the other way round. We then have, e.g.

$$p_Y(j) = \sum_{i \in \Omega_X} p_X(i) p_{Y|X}(j \mid i)$$

by the Law of Total Probability.

3.1.2 Absolutely continuous case

All the analogous results go through, where probability functions are replaced by p.d.f.s and sums by integrals. The definition of conditional p.d.f. as

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

is less obvious but nevertheless works.

3.2 Covariance and Correlation

3.2.1 Two variables

The *covariance* of two random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = E(X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y).$$

If X and Y are independent then $\mathrm{Cov}(X,Y) = 0$, but the converse is not true. The covariance measures how strongly X and Y are (linearly) related to each other, but if we require a *dimensionless* quantity to do this we use the *correlation coefficient*:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{[(\mathrm{Var}\,X)(\mathrm{Var}\,Y)]}}.$$

It may be shown that $\mid \rho(X,Y) \mid \leq 1$, with equality if and only if there is an exact linear relationship between X and Y ($Y = a + bX$ say).

To evaluate the covariance we use

$$E(XY) = \sum_{i \in \Omega_X} \sum_{j \in \Omega_Y} ij p_{X,Y}(i,j)$$

in the discrete case and an analogous integral in the absolutely continuous case.

Covariance is a symmetric function ($\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$) and is linear in each of its arguments, e.g.

$$\mathrm{Cov}(X, aY + bZ) = a\,\mathrm{Cov}(X,Y) + b\,\mathrm{Cov}(X,Z).$$

It is needed when we wish to evaluate the variance of the sum of two (or more) not necessarily independent random variables:

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

This is most easily seen by writing $\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y, X+Y)$ and using the linearity in both arguments.

3.2.2 General case

Let $\boldsymbol{X}$ be a random (column) vector with components X_1, X_2, ..., X_k. Then the mean vector $\boldsymbol{\mu}$ has its ith component given by EX_i, so we can write $E\boldsymbol{X} = \boldsymbol{\mu}$ and then (to practise notation) $\boldsymbol{\mu}_i = (E(\boldsymbol{X}))_i = E(\boldsymbol{X}_i) = E(X_i)$.

Now let

$$\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j),$$

which makes $\boldsymbol{\Sigma}$ a $k \times k$ matrix, and write $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. You can check (by considering the ijth entry on both sides, that

$$E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T = \boldsymbol{\Sigma}.$$

[Notational note: it is also common to use $\mathrm{Var}(\boldsymbol{X})$ for $\mathrm{Cov}(\boldsymbol{X})$.]

If $\boldsymbol{a}$ is a q-vector ($q \leq k$) and $\boldsymbol{B}$ is a $q \times k$ matrix of rank q then

$$E[\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}] = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}$$

and

$$\mathrm{Cov}[\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}] = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T. \tag{2}$$

If $\boldsymbol{y}$ is any vector (or length k) then, $\boldsymbol{y}^T\boldsymbol{X}$ is actually a random variable. Now using the result just given with $\boldsymbol{B} = \boldsymbol{y}^T$,

$$\mathrm{Cov}[\boldsymbol{y}^T\boldsymbol{X}] = \boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y};$$

but $\mathrm{Var}(\boldsymbol{y}^T\boldsymbol{X}) = \mathrm{Cov}[\boldsymbol{y}^T\boldsymbol{X}]$ and variances are always non-negative. Hence the matrix $\boldsymbol{\Sigma}$ must have the property that, for every $\boldsymbol{y}$, $\boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y} \geq 0$. Matrices with this property are called *non-negative definite*.

Multiplying out the definition $\mathrm{Cov}(X) = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T$ and rearrange to give that

$$\begin{aligned} E(\boldsymbol{X}\boldsymbol{X}^T) &= \mathrm{Cov}(\boldsymbol{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^T \\ &= \mathrm{Var}(\boldsymbol{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^T \quad \text{(see notational note on previous page)} \end{aligned}$$

(c.f. $E(X^2) = \mathrm{Cov}(X) + (EX)^2$). Here note that $\boldsymbol{X}\boldsymbol{X}^T$ is a matrix with ijth entry equal to X_iX_j. On the other hand

$$E(\boldsymbol{X}^T\boldsymbol{X}) = E\left(\sum X_i^2\right) = \sum Var(X_i) + (EX_i)^2 = \sum_i \mathrm{Cov}(\boldsymbol{X}))_{ii} + \sum_i \boldsymbol{\mu}_i\boldsymbol{\mu}_i$$

The trace of a square matrix $\boldsymbol{Z}$, $\mathrm{tr}(\boldsymbol{Z})$ is the sum of its diagonal entries, and so the previous equation can be rewritten as

$$E(\boldsymbol{X}^T\boldsymbol{X}) = \mathrm{tr}\left(\mathrm{Cov}(\boldsymbol{X})\right) + \boldsymbol{\mu}^T\boldsymbol{\mu}. \tag{3}$$

Also, if $\boldsymbol{B}$ is a $k \times k$ matrix then $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}$ is just a number and so is equal to its trace. Thus (using the properties of the trace — see Basic Mathematics material)

$$\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}^T),$$

and so

$$\begin{aligned} E(\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}) &= E\left(\mathrm{tr}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}^T)\right) \\ &= \mathrm{tr}(\boldsymbol{B}E(\boldsymbol{X}\boldsymbol{X}^T)) \\ &= \mathrm{tr}(\boldsymbol{B}(\mathrm{Cov}(\boldsymbol{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^T)) \\ &= \mathrm{tr}(\boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{X})) + \mathrm{tr}(\boldsymbol{B}\boldsymbol{\mu}\boldsymbol{\mu}^T) \\ &= \mathrm{tr}(\boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{X})) + \mathrm{tr}(\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu}) \\ &= \mathrm{tr}(\boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{X})) + \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu}. \end{aligned}$$

3.3 Worked examples

Example 17. Take the following joint probability function $p_{X,Y}(x, y)$ for two random variables X and Y:

		y		
$p_{X,Y}(x,y)$		1	2	3
	1	0	0.2	0.1
x	2	0.1	0.1	0.2
	3	0.2	0	0.1

The marginal distribution of X then has probability function given by

$$p_X(1) = 0 + 0.2 + 0.1 = 0.3$$
$$p_X(2) = 0.1 + 0.1 + 0.2 = 0.4$$
$$p_X(3) = 0.2 + 0 + 0.1 = 0.3$$

and that for Y has probability function given by

$$p_Y(1) = 0 + 0.1 + 0.2 = 0.3$$
$$p_Y(2) = 0.2 + 0.1 + 0 = 0.3$$
$$p_Y(3) = 0.1 + 0.2 + 0.1 = 0.4.$$

We can see that X and Y are not independent, as $p_{X,Y}(x,y) \neq p_X(x)p_Y(y)$.

Using the formula

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)},$$

we can calculate the table of the conditional distribution of X given Y:

		y		
$p_{X\|Y}(x\|y)$		1	2	3
	1	0	2/3	1/4
x	2	1/3	1/3	1/2
	3	2/3	0	1/4

Note that the columns, which each give a conditional distribution function of X, add to 1.

Example 18. Work out the covariance and correlation of the random variables X and Y in Example 17.

We calculate $E(X) = 2$, $E(Y) = 2.1$, $E(XY) = 0.3 \times 2 + 0.3 \times 3 + 0.1 \times 4 + 0.2 \times 6 + 0.1 \times 9 = 4$. So the covariance is -0.2. For the correlation, we need to calculate $\text{Var}(X) = 4.6 - 4 = 0.6$ and $\text{Var}(Y) = 5.1 - 4.41 = 0.69$, giving

$$\rho(X,Y) = \frac{-0.2}{\sqrt{0.414}} = -0.31.$$

3.4 Conditional expectation

If X and Y are two random variables, the expected value of Y given $X = x$ is simply the mean of the conditional distribution. It is denoted by $E(Y \mid X = x)$.

Since this depends upon x, we can write it as

$$g(x) = E(Y \mid X = x).$$

The corresponding function of the random variable X, $g(X)$, is known as the conditional expectation of Y given X, denoted by $E(Y \mid X)$. Note that it is a *random variable* which is a function of X but not of Y (Y has been 'eliminated' by summation or integration). We can similarly talk about $\mathrm{Var}(Y \mid X)$. The following properties of conditional expectation are useful:

- $E(E(Y \mid X)) = E(Y)$
- $E(h(X)Y \mid X) = h(X)E(Y \mid X)$ (Any function of X may be 'factorized out'.)
- $\mathrm{Var}(Y) = \mathrm{Var}(E(Y \mid X)) + E(\mathrm{Var}(Y \mid X))$.

3.5 Transformations of multivariate variables

3.5.1 Densities

The ideas of Section 2.5 generalize naturally. We assume an invertible transformation

$$\begin{aligned} Y_1 &= g_1(\boldsymbol{X}) \\ &\vdots \quad\quad \vdots \\ Y_n &= g_n(\boldsymbol{X}) \end{aligned}$$

with inverse

$$\begin{aligned} X_1 &= G_1(\boldsymbol{Y}) \\ &\vdots \quad\quad \vdots \\ X_n &= G_n(\boldsymbol{Y}) \end{aligned}$$

and Jacobian (generalization of derivative)

$$J(\boldsymbol{y}) = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

where $J(\mathbf{y}) \neq 0$ for invertibility. It can be shown that

$$f_{\boldsymbol{Y}}(y) = f_{\boldsymbol{X}}(G_1(y), \ldots, G_n(y)) \mid J(y) \mid .$$

Frequently, unwanted ys must then be integrated out to obtain the marginal distribution of the Ys of interest (again, take care over the ranges of integration). A particular case in common use is the *convolution integral* for obtaining the distribution of the sum of two variables. Using the transformation $U = X + Y, V = X$,

$$f_U(u) = \int_{R_{V|u}} f_{X,Y}(v, u - v) dv.$$

3.5.2 Approximate means and covariances

Now suppose that the random vector $\boldsymbol{X}$ has length n, mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, see §3.2.2. Consider $\boldsymbol{Y}$, a k-vector, obtained from $\boldsymbol{X}$ by application of the function $\boldsymbol{g}$. Thus $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ or, more explicitly,

$$\begin{aligned} Y_1 &= g_1(\boldsymbol{X}) \\ &\vdots \qquad \vdots \\ Y_k &= g_k(\boldsymbol{X}). \end{aligned}$$

Note that, unlike in §3.5.1, k may be smaller than n. The idea is to get an approximation for the mean vector and covariance matrix of $\boldsymbol{Y}$. The approach show in §2.5.2 works here too, but needs the multivariable version of Taylor's theorem:

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \approx \boldsymbol{g}(\boldsymbol{\mu}) + \boldsymbol{g}'(\boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})$$

where $\boldsymbol{g}'(\boldsymbol{x})$ is the $k \times n$ matrix

$$\begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_k}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_n} \end{pmatrix}.$$

Then, taking expectations, $E\boldsymbol{Y} \approx \boldsymbol{g}(\boldsymbol{\mu})$ and applying (2),

$$\mathrm{Cov}(\boldsymbol{Y}) \approx \boldsymbol{g}'(\boldsymbol{\mu})\boldsymbol{\Sigma}\boldsymbol{g}'(\boldsymbol{\mu})^T,$$

which does indeed give a $k \times k$ matrix, as it should.

This may seem rather abstract, but it is a powerful result when you have an estimated covariance matrix for a set of quantities (coming out of your statistical analysis) and you want to say something about the covariances of some function of them.

3.6 Standard multivariate distributions

3.6.1 The multinomial

If in each of a sequence of n independent trials, one of k different outcomes $A_1, A_2, \ldots, A_k$ may occur, with probabilities $\theta_1, \theta_2, \ldots, \theta_k$ respectively ($\theta_1 + \theta_2 + \ldots + \theta_k = 1$) and we define

$$\begin{aligned} X_1 &= \text{no. of times } A_1 \text{ occurs} \\ X_2 &= \text{no. of times } A_2 \text{ occurs} \\ &\text{etc.} \end{aligned}$$

then the joint distribution of $X_1, X_2, \ldots, X_k$ is called the *multinomial* with parameters $(n; \theta_1, \theta_2, \ldots, \theta_k)$. A combinatorial argument gives its joint probability function as

$$p(i_1, i_2, \ldots, i_k) = \begin{cases} \dfrac{n!}{i_1! i_2! \ldots i_k!} \theta_1^{i_1} \theta_2^{i_2} \ldots \theta_k^{i_k} & \text{if } \; i_1, i_2, \ldots, i_k \geq 0 \\ & \quad i_1 \; + \; i_2 + \ldots + i_k = n \\ 0 & \text{otherwise} \end{cases}$$

The marginal distribution of X_r is $Bi(n, \theta_r)$ and the distribution of X_r+X_s is $Bi(n, \theta_r+\theta s)$, which leads to the following results:

$$E(X)_r = n\theta_r$$
$$\text{Var}(X_r) = n\theta_r(1-\theta_r)$$
$$\text{Cov}(X_r, X_s) = -n\theta_r\theta_s \text{ if } r \neq s.$$

The *likelihood function* (see §4.5) is given by

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = \frac{n!}{X_1!X_2!\ldots X_k!} \ \theta_1^{X_1}\theta_2^{X_2}\ldots\theta_k^{X_k}$$

(assuming $X_1 + X_2 + \ldots + X_k = n$) and the maximum likelihood estimators are $\hat{\theta}_r = \frac{X_r}{n}$ for $r = 1, 2, \ldots, k$ which are unbiased[2].

3.6.2 The multivariate normal distribution

If $\boldsymbol{\mu}$ is a k-vector and $\boldsymbol{\Sigma}$ is a symmetric positive-definite $k \times k$ matrix, then random vector $\boldsymbol{X}$ has multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted by $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the joint p.d.f. is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

As the terminology suggests, $E(\boldsymbol{X})_r = \mu_r$ for each r and $\text{Cov}(X_r, X_s) = \sigma_{rs}$ (the (r, s) element of $\boldsymbol{\Sigma}$).

Despite the forbidding form of the joint p.d.f., the multivariate normal distribution has many nice properties and is the natural analogue of the normal.

Properties

(i) If $\boldsymbol{a}$ is a q-vector ($q \leq k$) and $\boldsymbol{B}$ is a $q \times k$ matrix of rank q then

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

(ii) If $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \ \boldsymbol{x}^{(n)}$ are i.i.d. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)} \sim N\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right).$$

3.7 Example

Here is an extended example that aims to show why being able to do calculations on covariances is important.

Suppose X_1, X_2, X_3, X_4 are from a multinomial on n trials with probabilities $(\theta_1, \theta_2, \theta_3, \theta 4)$. Imagine that these are the counts from a two-way contingency table[3] with the θs being the probabilities of the cells:

[2] See §4.7.1

[3] There is more on contingency tables in Block B.

	A	not A	
B	X_1	X_2	$X_1 + X_2$
not B	X_3	X_4	$X_3 + X_4$
	$X_1 + X_3$	$X_2 + X_4$	n

	A	not A	
B	θ_1	θ_2	$\theta_1 + \theta_2$
not B	θ_3	θ_4	$\theta_3 + \theta_4$
	$\theta_1 + \theta_3$	$\theta_2 + \theta_4$	1

Now the odds of A against not A, when B is true is θ_1/θ_2 and the corresponding odds when B is not true is θ_3/θ_4. The odds ratio is the ratio of these two odds, that is $(\theta_1/\theta_2)/(\theta_3/\theta_4)$ and the log odds ratio is the logarithm of this, given by

$$\psi = \log\theta_1 - \log\theta_2 - \log\theta_3 + \log\theta_4.$$

Notice that the when the two odds are the same it is because the way A is distributed is the same whether B holds of not. Thus the odds ratio being one (and hence the log odds ratio being zero) corresponds to independence of the two characteristics.

All of this is a lead in to wanting to estimate the log odds ratio and approximate its variance. The obvious estimate of θ_i is X_i/n. So the estimate of ψ is

$$\log(X_1/n) - \log(X_2/n) - \log(X_3/n) + \log(X_4/n).$$

Let $g(x_1, x_2, x_3, x_4) = \log(x_1/n) - \log(x_2/n) - \log(x_3/n) + \log(x_4/n)$. The results in Section 3.5.2 show that $g(X_1, X_2, X_3, X_4)$ has approximate mean ψ, but what about getting an approximation to its variance? Here $g'(\boldsymbol{x})$ is the 1×4 matrix

$$\left(\frac{1}{x_1}, -\frac{1}{x_2}, -\frac{1}{x_3}, \frac{1}{x_4}\right)$$

From Section 3.6.1 the covariance matrix of the X's is

$$\begin{pmatrix} n\theta_1(1-\theta_1) & -n\theta_1\theta_2 & -n\theta_1\theta_3 & -n\theta_1\theta_4 \\ -n\theta_1\theta_2 & n\theta_2(1-\theta_2) & -n\theta_2\theta_3 & -n\theta_2\theta_4 \\ -n\theta_1\theta_3 & -n\theta_2\theta_3 & n\theta_3(1-\theta_3) & -n\theta_3\theta_4 \\ -n\theta_1\theta_4 & -n\theta_2\theta_4 & -n\theta_3\theta_4 & n\theta_4(1-\theta_4) \end{pmatrix}.$$

Hence the approximate variance is

$$\left(\tfrac{1}{n\theta_1}, -\tfrac{1}{n\theta_2}, -\tfrac{1}{n\theta_3}, \tfrac{1}{n\theta_4}\right) \begin{pmatrix} n\theta_1(1-\theta_1) & -n\theta_1\theta_2 & -n\theta_1\theta_3 & -n\theta_1\theta_4 \\ -n\theta_1\theta_2 & n\theta_2(1-\theta_2) & -n\theta_2\theta_3 & -n\theta_2\theta_4 \\ -n\theta_1\theta_3 & -n\theta_2\theta_3 & n\theta_3(1-\theta_3) & -n\theta_3\theta_4 \\ -n\theta_1\theta_4 & -n\theta_2\theta_4 & -n\theta_3\theta_4 & n\theta_4(1-\theta_4) \end{pmatrix} \begin{pmatrix} \frac{1}{n\theta_1} \\ -\frac{1}{n\theta_2} \\ -\frac{1}{n\theta_3} \\ \frac{1}{n\theta_4} \end{pmatrix}.$$

Multiplying the first two components out produces extensive simplification to give

$$(1, -1, -1, 1) \begin{pmatrix} \frac{1}{n\theta_1} \\ -\frac{1}{n\theta_2} \\ -\frac{1}{n\theta_3} \\ \frac{1}{n\theta_4} \end{pmatrix} = \frac{1}{n\theta_1} + \frac{1}{n\theta_2} + \frac{1}{n\theta_3} + \frac{1}{n\theta_4}.$$

As in Example 14, it would be usual to want to estimate this variance, by replacing the θ's by their estimates, giving an estimated approximate variance of

$$\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \frac{1}{X_4},$$

which is a nice memorable form. With a little bit more work we could show that the same result is true if our original table had been $r \times s$, instead of 2×2 and we had the intersections of any picked any two rows and columns to identify the θ's we were interested in.

4 Introduction to inference and likelihood

4.1 Introduction

In *probability theory* we use the axioms of probability and their consequences as 'rules of the game' for *deducing* what is likely to happen when an 'experiment' is performed. In *statistical inference* we observe the outcome of an 'experiment' — which we call the *data* — and use this as evidence to *infer* by what mechanism the observation has been generated. Because of the apparent random variation in the data we assume that this mechanism is a probability model; the data help us to reach some *decision* about the nature of this model. Often in order to make the problem of inference tractable, we make some broad assumptions about the form of the underlying probability model; the data then help to narrow our choices within these assumptions.

As a common example, a collection of data of the same type, $x_1, x_2, ..., x_n$ say, may be regarded as an observation on a sequence of independent identically distributed (i.i.d.) random variables, $X_1, X_2, ..., X_n$ say, whose common distribution is of some standard form, e.g. normal or Poisson, but whose parameters are unknown. Such data constitute a *random sample*. The data then give us information about the likely values of the parameters.

These notes concentrate on *classical* (or *frequentist*) inference in which we assume that the values of our underlying parameters are unknown, but fixed. An alternative approach is *Bayesian* inference (see MAS6004) in which we express our initial uncertainty about the true values of parameters by assigning them probability distributions. The data then combine with these prior ideas to offer an updated distributional assessment of the parameter values.

4.2 Types of (classical) inference

Three common types of inference are as follows

Point estimation A single number (which is a function of the observed data) is given as an *estimate* of some parameter. The formula specifying the function is called an *estimator*.

Interval estimation Instead of a single number, we specify a *range* (usually an interval) of values in which the parameter is thought to lie on the basis of the observed data.

Testing hypotheses A hypothesis about a parameter value, or more generally about the underlying probability model, is to be *accepted* or *rejected* on the basis of the observed data.

Many other types of action are possible, e.g. 'collect more data'.

4.3 Sampling distributions

Inference about a population is usually conducted in terms of a *statistic* $t = T(x_1, x_2, ..., x_n)$ which is a function of the observed data. Regarded as a function of the corresponding random variables, $T = T(X_1, X_2, ..., X_n)$, this statistic is itself a random variable. Its distribution is known as the *sampling distribution* of the statistic. In principle this distribution can be derived from the distribution of the data; in practice this may or may not be possible. The sampling distribution is important because it enables us to construct good inferential procedures and to quantify their effectiveness.

Example 19. If $X_1, X_2, ..., X_n$ is a random sample from the normal distribution $N(\mu, \sigma^2)$ then two important statistics are the *sample mean*

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

and the *sample variance*

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

which are used to estimate μ and σ^2 respectively. It may be shown that

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and that $\overline{X}$ and S^2 are independent.

The first result above, for instance, tells us how precise $\overline{X}$ is as an estimator of μ: it has 'mean square error'

$$E\left(\overline{X} - \mu\right)^2 = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

which decreases as the sample size n increases.

4.4 Normal approximations for large samples

In the preceding example, we note that if the probability model is correct (i.e. if X_1, X_2, ..., X_n are i.i.d. random variables each with $N(\mu, \sigma^2)$ distribution) then the sampling distribution of $\overline{X}$ is known *exactly*. But the *central limit theorem* tells us that even if the distribution of the Xs is *not normal* in form, the sampling distribution of $\overline{X}$ is approximately normal in form if the sample size n is large. We say that the sampling distribution of $\overline{X}$ is *asymptotically* normal.

The important consequence of this result is that we can conduct inference using the sampling distribution of $\overline{X}$ without being too concerned what the underlying distribution of the Xs is.

There are many other statistical models in which the sampling distribution of a statistic is *asymptotically normal*; in other cases it may be e.g. *asymptotically* χ^2 (which typically has behind it a result about some multivariate statistic being multivariate normal by the central limit theorem). The beauty of these results is that what might be an intractable distributional problem is made tractable by an approximation using a standard distribution. Such methods are known as *large sample methods*.

4.5 Likelihood

A key tool in many classical and Bayesian approaches to inference is the *likelihood function*.

We often have some reason (for example a probabilistic model for the process producing the data) to assume a particular form for the joint p.d.f. of $\boldsymbol{X}$, usually involving some standard distribution. However, the parameters of the standard distribution will usually be unknown, and our aim in analysing the data will be to obtain information about the values of these unknown parameters.

General parameters are usually denoted by $\boldsymbol{\theta}$; we represent $\boldsymbol{\theta}$ as a vector, although in any particular case the set of parameters can be a scalar (a single number), a matrix or some other structure.

If we have a statistical model with parameter values $\boldsymbol{\theta}$ then $f(\boldsymbol{x}|\boldsymbol{\theta})$ is a p.d.f. (in the continuous case) and as such it defines the distribution of $\boldsymbol{X}$, given values of $\boldsymbol{\theta}$.

(If we have discrete random variables, then we would have a probability function instead of a p.d.f. The theory in this case is very similar, so it is common to use the same notation in both cases.)

Once we have observed data values $\boldsymbol{x}$ (a realisation of $\boldsymbol{X}$) we can consider $f(\boldsymbol{x}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. Note that, considered this way, it is no longer a p.d.f.; in particular there is no requirement for it to integrate to 1.

Definition:

When we regard $f(\boldsymbol{x}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, for the fixed (observed) data $\boldsymbol{x}$, it is called the *likelihood function* (or just the likelihood). We will denote it by $L(\boldsymbol{\theta}; \boldsymbol{x})$ and we say that $L(\boldsymbol{\theta}; \boldsymbol{x})$ is the likelihood of $\boldsymbol{\theta}$ based on data $\boldsymbol{x}$.

The likelihood is a function of the parameter $\boldsymbol{\theta}$. Thus, to describe completely the likelihood, it is important, when we calculate $L(\boldsymbol{\theta}; \boldsymbol{x})$, to identify the set of the possible parameter values $\boldsymbol{\theta}$, that is the domain of L. We will denote the set of possible parameter values by Θ.

In a situation in which $X_1, \ldots, X_n$ are not a random sample (so they are not i.i.d.), the likelihood is still just the joint p.d.f. or probability function of the variables, but this will no longer be a simple product of terms of identical form.

This function provides the basic link between the data and unknown parameters so it is not surprising that it is used throughout classical inference in determining (theoretically or graphically) estimates, test statistics, confidence intervals and their properties and underpins the entire Bayesian approach. In many situations it is more natural to consider

the natural logarithm (but we usually use log for this rather than ln) of the likelihood function

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log(L(\boldsymbol{\theta}; \boldsymbol{x})).$$

This is often denoted simply $l(\boldsymbol{\theta})$.

4.6 Maximum likelihood estimation

If we approach inference from a likelihood-based perspective, then a natural choice of point estimator for $\boldsymbol{\theta}$ is the *maximum likelihood estimator* $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_n)$ where $\hat{\theta}$ is such that

$$L(\hat{\boldsymbol{\theta}}; \boldsymbol{X}) = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{X})$$

This is the value which gives the observed data the highest likelihood (i.e., in the discrete case, the highest probability of having occurred).

Example 20. Let $x_1, x_2, ..., x_n$ be a random sample from $Po(\mu)$.

$$L(\mu; \boldsymbol{x}) = e^{-\mu}\frac{\mu^{x_1}}{x_1!} \ldots e^{-\mu}\frac{\mu^{x_n}}{x_n!}.$$

It is easier to maximize

$$\ell(\mu) = \log L(\mu; \boldsymbol{x}) = -n\mu + \left(\sum_{i=1}^{n} x_i\right) \log \mu - \sum_{i=1}^{n} \log x_i!.$$

Differentiating,

$$\frac{d\ell}{d\mu} = -n + \frac{\sum x_i}{\mu} = 0 \quad \text{when} \quad \mu = \frac{\sum x_i}{n} = \overline{x}.$$

Therefore $\hat{\mu} = \hat{\mu}(X_1, X_2, \ldots, X_n) = \overline{X}$ is the maximum likelihood estimator of μ. You can check that the second derivative is negative to confirm that this is indeed a maximum.

Example 21. Let $x_1, x_2 \ldots, x_n$ be the values of independent observations of exponential random variables with unknown parameter μ (mean $1/\mu$). Then the likelihood function is

$$L(x|\mu) = \prod_{j=1}^{n} \mu e^{-\mu x_j} = \mu^n \prod_{j=1}^{n} e^{-\mu x_j}$$

and the log likelihood $l(x|\mu) = n \log \mu - \mu \sum_{j=1}^{n} x_j$.

To find the maximum likelihood estimate, differentiate $l(x|\mu)$ with respect to μ:

$$\frac{dl(x|\mu)}{d\mu} = \frac{n}{\mu} - \sum_{j=1}^{n} x_j.$$

So at a maximum we will have

$$\frac{n}{\mu} = \sum_{j=1}^{n} x_j,$$

which implies that at a maximum $\mu = n/\sum_{j=1}^{n} x_j = 1/\bar{x}$. To check that this really is a maximum, look at the second derivative:

$$\frac{d^2 l(x|\mu)}{d\mu^2} = -\frac{n}{\mu^2},$$

which is negative, indicating that we have indeed found a maximum.

So the maximum likelihood estimate is $\hat{\mu} = 1/\bar{x}$. (Note that we write $\hat{\mu}$ to distinguish this from the true value of the parameter μ, which was what we were trying to estimate, but the actual value of which remains unknown.)

Notes: 1) It is equivalent and often easier to maximize the natural logarithm of the likelihood rather than L itself. 2) Writing the process in terms of the observed likelihood $L(\boldsymbol{\theta}; \boldsymbol{x})$ leads to the maximum likelihood estimate rather than the maximum likelihood estimator. (The first is a function of the actual values, $\boldsymbol{x}$, the second of the random variables $\boldsymbol{X}$.) This may tempt you into notational inaccuracy when considering properties of m.l.e.s, etc.

4.7 Properties of estimators

4.7.1 Unbiasedness

If $\tau(\theta)$ is some function of a parameter θ and $T(X_1, ..., X_n) = T(\boldsymbol{X})$ is some function of $\boldsymbol{X}$, then T is an *unbiased* estimator of τ if

$$E_\theta[T(\boldsymbol{X})] = \tau(\theta)$$

for all values of θ. [E_θ denotes expected value with respect to the sampling distribution of T when θ is the true value of the parameter]. In other words, 'on average' T gives the right answer.

Example 22. If $X_1, X_2, ..., X_n$ is a random sample from any distribution with mean μ and

$$\overline{\boldsymbol{X}} = \frac{X_1 + X_2 + ... + X_n}{n},$$

then $E\overline{\boldsymbol{X}} = \mu$ as we have seen earlier. Hence $\overline{\boldsymbol{X}}$ is an unbiased estimator of μ.

Unbiasedness is a desirable property *in general*, but not always, as the following example shows.

Example 23. Let X be an observation from $Po(\mu)$, and suppose we wish to estimate $\tau(\mu) = e^{-2\mu}$. If we define

$$T(X) = \begin{cases} 1 \text{ if } X \text{ is even} \\ -1 \text{ if } X \text{ is odd} \end{cases}$$

then

$$E[T(X)] = \sum_{i=0}^{\infty} (-1)^i e^{-\mu} \frac{\mu^i}{i!} = e^{-\mu}.e^{-\mu} = e^{-2\mu}$$

and so T is an unbiased estimator of $e^{-2\mu}$. But is is an absurd estimator, because e.g. $e^{-2\mu}$ cannot take the value -1.

4.7.2 Mean square error and efficiency

If $T(\boldsymbol{X})$ is an estimator of $\tau(\theta)$, then its *mean square error* (m.s.e.) is given by

$$E_\theta[T(\boldsymbol{X}) - \tau(\theta)]^2$$

In particular, if $T(\boldsymbol{X})$ is unbiased for $\tau(\theta)$, then the above is the same as the variance of $T(\boldsymbol{X})$. It is a measure of the likely accuracy of T as an estimator of t: the larger it is, the less accurate the estimator.

If T and T' are two different estimators of t, the *relative efficiency* of T' is given by the ratio

$$\frac{E_\theta(T-\tau)^2}{E_\theta(T'-\tau)^2}$$

This appears to depend on θ, but in many cases it does not.

Example 24. Let $X_1, X_2, ..., X_{2m+1}$ be a random sample (of odd size) from $N(\mu, \sigma^2)$, let $\overline{\boldsymbol{X}}$ be the sample mean, and let M be the sample *median*, i.e. the $(m+1)^{th}$ data value in *ascending order*. $\overline{\boldsymbol{X}}$ and M are both unbiased estimators of μ; we know that

$$\mathrm{Var}(\overline{\boldsymbol{X}}) = \frac{\sigma^2}{2m+1},$$

and it may be shown that

$$\mathrm{Var}(M) = \frac{\pi\sigma^2}{2(2m+1)}$$

for large m. Hence the (asymptotic) relative efficiency of M with respect to $\overline{\boldsymbol{X}}$ as an estimator of μ is

$$\frac{\sigma^2}{2m+1} \cdot \frac{2(2m+1)}{\pi\sigma^2} = \frac{2}{\pi}.$$

M is less efficient than $\overline{\boldsymbol{X}}$ by this factor.

5 Interval estimation and hypothesis testing

5.1 Interval estimation

In point estimation we aim to produce a single 'best guess' at a parameter value θ. Here, in contrast, we provide a region (usually an interval) in which θ 'probably lies'.

Construction is via inversion of a probability statement concerning the data, so if T is a statistic derived from the data we find, by examining the sampling distribution of T, a region $A(\theta)$ such that

$$P(T \in A(\theta)) = 1 - \alpha \tag{4}$$

for some suitable small α, and invert this to identify the region $C(T) = [\theta : T \in A(\theta)]$ which then satisfies

$$P(\theta \in C(T)) = 1 - \alpha. \tag{5}$$

In other words, the 'random region' $C(T)$ 'covers' the true parameter value θ with probability $1-\alpha$. In simple situations $C(T)$ is an interval, called a *confidence interval* for θ. The confidence level is usually expressed as a percentage: 100(1-α)% (e.g. $\alpha = 0.05$ gives a 95% confidence interval).

Example 25. $X_1, X_2, ..., X_n$ a random sample from $N(\mu, \sigma^2)$ with μ unknown, σ^2 known. $T = \overline{X}$ with

$$\overline{X} \sim N(\mu, \sigma^2/n).$$

Thus

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha \tag{6}$$

where $z_{1-\alpha/2}$ solves $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$ (from tables).

Thus

$$P\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{7}$$

so

$$A(\mu) = \left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Inverting (7)

$$P\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

i.e.

$$C(T) = \left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Note: we have chosen the $z_{1-\alpha/2}$ in (6) symmetrically; we could have chosen any z values giving overall probability level $1 - \alpha$.

In general if an estimator $\hat{\theta}$ is asymptotically normally distributed with mean equal to the true value of the parameter θ, then an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\left(\hat{\theta} - z_{1-\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{1-\alpha/2}\sigma_{\hat{\theta}}\right)$$

where $\sigma_{\hat{\theta}}$, the standard deviation of the sampling distribution of $\hat{\theta}$, is called the *standard error*. This may itself have to be estimated from the data.

5.2 Introduction to testing hypotheses

A *hypothesis* is a statement about the underlying probability model which may or may not be true. We do not know which. We can, however, obtain some information by looking at the data. A *test* is a decision rule which attaches a verdict ('do not reject' or 'reject') to each possible set of observed data. Ideally we would like to make the right decision most of the time.

If the model is specified up to an unknown (vector) parameter $\boldsymbol{\theta}$ say, then a hypothesis H is a statement restricting $\boldsymbol{\theta}$ to some subset Θ_0 of the parameter space Θ. If H specifies completely the distribution of the data, it is called a *simple* hypothesis (usually this means Θ_0 is a singleton); otherwise it is *composite*. A *null* hypothesis (typically denoted H_0) is one of particular interest — typically the default/status quo/no effect hypothesis.

5.3 Pure significance tests (for simple H_0)

In addition to our null hypothesis H_0 we need some idea of the type of departures from H_0 which are of interest.

Example 26. Data $X_1, X_2, ..., X_n$; $H_0 : X_i \sim N(0,1)$ i.i.d.

Possible departures of interest:

(a) increase in mean,

(b) increase in variance,

(c) correlation between successive individuals.

To formulate a pure significance test we need to find a statistic $T(\boldsymbol{X})$ such that

(i) the values of T increase with departure from H_0 in the direction of interest;

(ii) the distribution of T is known when H_0 is assumed true.

For the departures from H_0 in Example 26 we might use the following test statistics:

(a) Use $\overline{X} \sim N(0, 1/n)$ on H_0 (note terminology).

(b) Use $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$ on H_0.

(c) Use $\sum_{1}^{n-1} X_i X_{i+1}$ distribution known on H_0

We evaluate t_{obs}, the observed value of T, and find $p_{obs} = P(T \geq t_{obs} \mid H_0 \text{ true})$ — the *p-value* or *observed (or exact) significance level of T*. This is the probability, under H_0, of getting a value of T at least as extreme as the one observed.

In specifying the result of the test we would either quote the p-value directly, or, if $p \leq \alpha$ for some small preassigned value α, say 'the data depart significantly (at level α) from H_0 (in the direction of interest)'.

5.4 Hypothesis tests

Here we specify formally both our null hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ and an *alternative hypothesis* $H_1 : \boldsymbol{\theta} \in \Theta_1$ (often Θ_0^c) representing our 'departures of interest'.

Any test divides the set of possible data sets $\mathcal{X}$ say, into an *non-rejection region* $\mathcal{X}_0$ and a *critical region* $\mathcal{X}_1$ such that if the observed $\boldsymbol{x}$ lies in $\mathcal{X}_1$ we reject H_0 and otherwise we do not reject it.

In principle we can evaluate, for each $\theta \in \Theta$, the probability that $\boldsymbol{X}$ lies in $\mathcal{X}_1$; the function

$$\gamma(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathcal{X}_1)$$

is called the *power function* of the test and summarises its properties. Ideally we wish to choose $\mathcal{X}_1$ so that $\gamma(\boldsymbol{\theta})$ is small (close to 0) for $\boldsymbol{\theta} \in \Theta_0$ and large (close to 1) when $\boldsymbol{\theta} \in \Theta_1$. The number

$$\alpha = \sup_{\theta \in \Theta_0} \gamma(\boldsymbol{\theta})$$

is called the *size* of the test, this is the maximum probability of rejecting H_0 if it is true.

Often, instead $\boldsymbol{X}$, we work in terms of a statistic $T = T(\boldsymbol{X})$, say, known as the *test statistic*; it is then convenient to define the non-rejection and critical regions in terms of T rather than $\boldsymbol{X}$.

5.5 Constructing tests

5.5.1 Intuition

Often the form of a test (i.e. the form of the critical region) will be suggested by intuition. For example, if T is a good estimator of a parameter θ and the hypotheses take the 'one-sided' form

$$H_0 : \theta \leq \theta_0 \qquad\qquad H_1 : \theta > \theta_0$$

for some fixed given θ_0, then we would expect that the larger the value of T, the greater the evidence in favour of H_1; and so a sensible critical region will be of the form $\{T > c\}$ for some constant c. Then

$$\gamma(\theta) = P_\theta(T > c).$$

To choose c: as c increases, $\gamma(\theta)$ decreases for all values of θ; for $\theta \in \Theta_0$ this is a good thing, whereas for $\theta \in \Theta_1$ it is a bad thing. So we must choose c so as to strike a balance. Often this is done by prescribing the value of α (e.g. 0.05). This is sufficient to determine c. Often *tables* of *critical values* of the test statistic for specified α are produced.

Example 27. $X_1, X_2, ..., X_n$ a sample from $N(\mu, 1)$.

$$H_0 : \mu \leq 0 \qquad\qquad H_1 : \mu > 0.$$

Use $\overline{X}$ as test statistic; critical region will be of the form $\{\overline{\boldsymbol{x}} > c\}$.

$$\gamma(\mu) = P_\mu(\overline{X} > c) = 1 - \Phi\left(\frac{c-\mu}{1/\sqrt{n}}\right) = \Phi\left(\sqrt{n}(\mu - c)\right).$$

This has its maximum value in H_0 when $\mu = 0$; so if α is prescribed we must choose c such that $\Phi(-c\sqrt{n}) = \alpha$. This is easily solved for c (using tables).

If $n = 100, \alpha = 0.05$ $\Phi(-10c) = 0.05$ $c = \frac{-1}{10}\Phi^{-1}(0.05)$ Neave table 2.3(b) gives $c = \frac{-1}{10}(-1.6449) = 0.16449$. So the test is 'reject $H_0 : \mu \leq 0$ at the 5% level if $\overline{x} > 0.16449$'

5.5.2 Neyman-Pearson approach

This form is most easily introduced in the (not very practically useful) case of choosing between two simple hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ (it can be extended to composite hypotheses).

We consider the probabilities of making the two types of error that can occur in any test

$$\alpha = P(\textit{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ true}) = \gamma(\theta_0) \text{ as before;}$$

$$\beta = P(\textit{type II error}) = P(\text{do not reject } H_0 \mid H_1 \text{ true}) = 1 - \gamma(\theta_1).$$

Again, ideally we would like both α and β small, but simultaneous reduction is usually impossible so we compromise by fixing α at an acceptably low level and then minimising β (for this fixed α).

The *Neyman-Pearson lemma* tells us that the test with

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \ : \frac{L(\theta_0; \boldsymbol{x})}{L(\theta_1; \boldsymbol{x})} < c \right\}$$

where c is such that $P(\boldsymbol{X} \in \mathcal{X}_1 \mid \theta_0) = \alpha$ will minimize $\beta = P(\boldsymbol{X} \in \mathcal{X}_0 \mid \theta_1)$, i.e. it is the *most powerful test of level* α.

Typically the form of the test simplifies and we can work with a particular test statistic T and determine probabilities from its sampling distribution.

Example 28. $X_1, X_2, ..., X_n$ a sample from $N(\mu, \sigma^2)$, σ^2 known.

$$H_0 : \mu = \mu_0 \qquad\qquad H_1 : \mu = \mu_1 \quad \text{with } \mu_1 > \mu_0 \text{ say.}$$

Intuition suggests a test with critical region $\{\overline{X} > c\}$ and this is indeed the most powerful test:

$$L(\mu; \boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \right\}.$$

So the N-P lemma says the best test has critical region given by $\boldsymbol{X}$'s for which

$$\exp\left\{ -\frac{1}{2\sigma^2} \left[\sum (X_i - \mu_0)^2 - \sum (X_i - \mu_1)^2 \right] \right\} < c$$

$$\text{i.e.} \quad -\frac{1}{2\sigma^2} \left[\sum X_i^2 - 2\mu_0 \sum X_i + \mu_0^2 - \sum X_i^2 + 2\mu_1 \sum X_i - \mu_1^2 \right] < c'$$

i.e. $\sum X_i(\mu_1 - \mu_0) > c''$ (take care over direction of inequality)

i.e. $\overline{X} > c''' = k$ say recalling $\mu_1 - \mu_0 > 0$.

Now k is determined by the requirement

$$P(\boldsymbol{X} \in \mathcal{X}_1 \mid \mu = \mu_0) = \alpha \quad \text{i.e. } P(\overline{X} > k \mid \mu = \mu_0) = \alpha.$$

The sampling distribution of $\overline{X}$ is $N(\mu_0, \sigma^2/n)$ when H_0 holds. Therefore

$$P(\overline{X} > k \mid \mu = \mu_0) = P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{k - \mu_0}{\sigma/\sqrt{n}} \right) = \alpha.$$

i.e. $k = \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$ evaluated using Neave 2.3(b).

5.5.3 The likelihood ratio procedure

This is a method of suggesting a form of test which in principle always works. It gives the same result as the Neyman-Pearson test in the simple versus simple case (despite the difference in the denominator) and extends easily to more general situations. There is also a useful asymptotic result about it (see later). The test is defined by considering the ratio

$$\frac{\sup_{\theta\in\Theta_0} L(\boldsymbol{\theta};\boldsymbol{x})}{\sup_{\theta\in\Theta} L(\boldsymbol{\theta};\boldsymbol{x})} = \Lambda.$$

This ratio is always < 1 since the supremum in the denominator is over a larger set. The further Λ is from 1, the greater the evidence against H_0. Hence a form of critical region which suggests itself is $\{\Lambda < c\}$.

Example 29. $X_1, X_2, ..., X_n$ a sample from $N(\mu, \sigma^2)$ with both μ and σ^2 unknown.

$$H_0 : \mu = \mu_0 \qquad\qquad H_1 : \mu \neq \mu_0 \quad \text{(a 'two-sided' alternative).}$$

Here $\Theta_0 = \{(\mu, \sigma^2);\ \ \mu = \mu_0, \sigma^2 > 0\}$ and $\Theta = \{(\mu, \sigma^2);\ \ \sigma^2 > 0\}$.
The likelihood is

$$L(\boldsymbol{\theta};\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}\right].$$

This is maximized in Θ_0 by putting $\mu = \mu_0$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$ (check by differentiating!), giving

$$\max_{\theta\in\Theta_0} L(\boldsymbol{\theta};\boldsymbol{X}) = \frac{n^{n/2}\quad e^{-n/2}}{(2\pi)^{n/2}\left[\sum(X_i-\mu_0)^2\right]^{n/2}},$$

It is maximized in Θ by putting $\mu = \overline{X}$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, giving

$$\max_{\theta\in\Theta} L(\boldsymbol{\theta};\boldsymbol{X}) = \frac{n^{n/2}\quad e^{-n/2}}{(2\pi)^{n/2}\left[\sum(X_i-\overline{X})^2\right]^{n/2}}.$$

Hence the critical region takes the form (from the ratio of the previous two expressions)

$$\left(\frac{\sum(X_i-\overline{X})^2}{\sum(X_i-\mu_0)^2}\right)^{n/2} < c.$$

Since $\sum(X_i-\mu_0)^2 = \sum(X_i-\overline{X})^2 + n(\overline{X}-\mu_0)^2$, this is equivalent to

$$\frac{(\overline{X}-\mu_0)^2}{\sum(X_i-\overline{X})^2} > c'$$

or

$$\frac{\overline{X}-\mu_0}{S} > c''$$

where S^2 is the sample variance.

The statistic

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$$

is called the *one-sample t statistic*. Its sampling distribution if $\mu = \mu_0$ does not depend upon σ^2, and is known as the *t distribution with $n-1$ degrees of freedom*. So a test of given size may be constructed using tables of critical values for this distribution (Neave 3.1).

The form of test suggested by the likelihood ratio procedure often simplifies, as above, but the sampling distribution of the simplified test statistic must be derived by 'ad hoc' methods — and this is not always possible. It is, therefore, useful to have a general asymptotic result which works under reasonable regularity conditions as the sample size $n \to \infty$; this is that if H_0 is true $-2\log\Lambda$ *is asymptotically* χ^2_ν where ν, the number of degrees of freedom, is the difference in dimensionality between Θ_0 and Θ.

Example 30. $(X_{11}, ..., X_{1n})$, $(X_{21}, ..., X_{2n})$, ..., $(X_{k1}, ..., X_{kn})$ are samples from $N(\mu_1, \sigma_1^2)$, ..., $N(\mu_k, \sigma_k^2)$ respectively where μ_1, σ_1^2, ..., μ_k, σ_k^2 are all unknown, and we wish to test $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$. Then the unrestricted parameter space Θ is $2k$-dimensional whereas Θ_0 is $(k+1)$-dimensional, since H_0 imposes $(k-1)$ linear constraints. Hence

$$-2\log\Lambda \sim X^2_{k-1} \quad \text{asymptotically.}$$

5.6 Duality between interval estimation and hypothesis testing

There is a duality between interval estimation and hypothesis testing in that possible parameter values which would not be rejected in a size α test are those which constitute the $100(1-\alpha)\%$ confidence interval. This gives us a direct means of establishing $A(\theta)$ in (4).

Let $A(\theta_0)$ denote the acceptance region for a standard test of size α of $H_0 : \theta = \theta_0$ (simple) against $H_1 : \theta \neq \theta_0$ using a test statistic T. Then we have

$$P_{\theta_0}(T \in A(\theta_0)) = 1 - \alpha \quad \text{for all} \quad \theta_0.$$

Defining $C(T)$ as in (5), this may be written

$$P_\theta(\theta \in C(T)) = 1 - \alpha.$$

So $C(T)$ is a confidence interval for θ.