Preparatory Course in Statistics for the M.Phil. Programme

Donald Robertson*
Faculty of Economics, University of Cambridge

September 2020

1 Probability theory and statistics

1.1 Elements of probability theory

1.1.1 Definition of probability

To define probability it is useful to think in terms of experiments. We know that each experiment has an **outcome** (denoted by ω); however, this outcome cannot be known *a priori*, so that it is uncertain: therefore the experiment is called a **random experiment**. The set of all possible outcomes Ω will be referred to as the **sample space**.

Example 1 *Suppose we toss a coin twice; we have four possible mutually exclusive outcomes*

$$\omega_1 = (H, H) \quad \omega_2 = (H, T) \quad \omega_3 = (T, H) \quad \omega_4 = (T, T)$$

while the set Ω *is*

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

Now, given the experiment, we can invent a measure of how rare or surprising its various possible outcomes are. We will refer to this measure as

*Please report any typos or mistakes to dr10011@cam.ac.uk.

probability. An **event** $A \subseteq \Omega$ is a set of distinct outcomes. Intuitively probability of the event A can be thought of as a frequency of outcomes leading to A.

Example 2 *Continuing the coin toss example, denote*

$$A_1 = \{\emptyset\} \quad A_2 = \{(H,H),(H,T)\} \quad A_3 = \{(T,H),(T,T)\}$$

so that $A_i \cap A_j = \emptyset, i, j = 1, 2, 3, i \neq j$. *Then*

$$\begin{aligned} \Pr[A_1 \cup A_2 \cup A_3] &= \Pr[\Omega] = 1 \\ &= \Pr[A_1] + \Pr[A_2] + \Pr[A_3]. \end{aligned}$$

We can now introduce the following result:

Proposition 1 *(**The additive law**) Let* A_1 *and* A_2 *be any two events. Then*[1]

$$\Pr[A_1 \cup A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 \cap A_2].$$

We say that two events are **mutually exclusive** if they cannot both occur at the same time, i.e., mutually exclusive events A_1 and A_2 have property that $\Pr[A_1 \cap A_2] = 0$.

1.1.2 Conditional probability

Having defined what we mean by probability we can now introduce a further concept. In particular we will consider situations where events have an effect on each other. To understand what we mean, consider the following example.

Example 3 *For a single toss of a fair dice* $\Pr[6] = \frac{1}{6}$, *while* $\Pr[\textit{even number}] = \frac{1}{2}$. *However, the probability of the outcome being* 6 ***knowing*** *that the outcome is an even number is* $\Pr[6 \,|\textit{even number}] = \frac{1}{3}$. *Therefore the probability of a six has changed once we have the information that an even number has occurred. The effect of the information is to reduce the set of possible outcomes from* $\{1,2,3,4,5,6\}$ *to* $\{2,4,6\}$.

[1] Note that drawing a Venn diagram helps understanding this law.

Definition 1 *(**Conditional probability**) Consider two events A and B. The probability of event A given event B is equal to the probability of A and B divided by the probability of B:*

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Consider now a situation where knowing that an event B has happened does not provide any information about the occurrence of event A; formally

$$\Pr[A|B] = \Pr[A] \Rightarrow \Pr[A \cap B] = \Pr[A|B]\Pr[B] = \Pr[A]\Pr[B].$$

Definition 2 *(**Independence**) Two events A and B are said to be **independent** if*

$$\Pr[A \cap B] = \Pr[A]\Pr[B].$$

One of the most useful results of conditional probability is Bayes' Rule, which is extensively used in Game Theory.

Proposition 2 *(**Bayes' Rule**) Suppose that $\Omega = A_1 \cup \ldots \cup A_n$, with A_i disjoint, $\Pr[A_i] > 0$, $\forall i = 1, \ldots, n$. Then for any event B such that $\Pr[B] > 0$*

$$\Pr[A_i|B] = \frac{\Pr[A_i \cap B]}{\Pr[B]} = \frac{\Pr[A_i]\Pr[B|A_i]}{\sum_{j=1}^{n}\Pr[A_j]\Pr[B|A_j]}.$$

1.2 Random Variables

1.2.1 A random variable

So far we have defined the outcome of a random experiment ω and the set of possible outcomes Ω. In order to make the outcomes of random experiments somehow tractable we need to assign a numerical value to each of them. This is done by introducing the concept of random variable.

Definition 3 *(**Random variable**) A **random variable** X is a function that assigns a real number to each element of the sample space Ω:*

$$X : \Omega \to \mathbb{R}.$$

By convention we denote by X the random variable itself, and by x the *actual realisation* of the random variable X, that is

$$x = X(\omega).$$

Example 4 *In the coin toss example, a toss can be classified as either a success (denoted by* 1*) or a failure (denoted by* 0*). For each toss* $\Omega = \{H, T\}$ *we can define the following random variable*

$$\begin{aligned} X(\{H\}) &= 1 \\ X(\{T\}) &= 0. \end{aligned}$$

Random variables are classified as either discrete or continuous, depending on the set of values they can take.

1.2.2 Discrete Random Variables

Definition 4 *(**Discrete random variable**) A random variable* X *is said to be discrete if it takes values in a countable subset of* $\mathbb{R}$ *(typically* $\mathbb{N}$ *or a subset of it).*

In order to compute probabilities of events, a discrete random variable can be directly characterised in terms of a function called probability distribution.

Definition 5 *Consider a discrete random variable* X*. A **probability mass function** is a function that links each value the random variable* X *can take to the corresponding probability. Formally, denoting by* $p_X(\cdot)$ *the probability mass function of the discrete random variable* X *taking values in the set* $\{x_1, x_2, \ldots\}$*, then*

$$p_X(x) = \Pr[X = x].$$

Example 5 *In the coin toss example, if we define* $\Pr[X = 1] = \pi$ *then*

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad x = 0, 1$$

A probability mass function satisfies the following properties:

1. $p_X(x) \geq 0$ for all x.
2. $\sum_{i=1}^{\infty} p_X(x_i) = 1$.

Another useful concept is that of cumulative distribution function (cdf), usually denoted by $F_X()$. It is equivalent to that of probability mass function so it can be used as an alternative way to define a random variable.

Definition 6 *Consider a discrete random variable X taking values in the set $\{x_1, x_2, \ldots\}$; the **cumulative distribution function** is defined as*

$$F_X(x) = \Pr[X \leq x] = \sum_{a \leq x} p_X(a).$$

The cdf of a discrete random variable satisfies the following properties:

1. $\lim_{x \to -\infty} F_X(x) = 0$.
2. $\lim_{x \to \infty} F_X(x) = 1$.
3. If $x_1 > x_2$ then $F_X(x_1) \geq F_X(x_2)$.
4. $p(x_h) = F_X(x_h) - F_X(x_{h-1})$. This is because

$$\begin{aligned} F_X(x_h) - F_X(x_{h-1}) &= \sum_{i=1}^{h} p_X(x_i) - \sum_{i=1}^{h-1} p_X(x_i) \\ &= p_X(x_h) + \sum_{i=1}^{h-1} p_X(x_i) - \sum_{i=1}^{h-1} p_X(x_i) \\ &= p_X(x_h). \end{aligned}$$

1.2.3 Continuous Random Variables

Definition 7 *(**Continuous random variable**) A random variable X is said to be continuous if there exists a function $f_X(x)$ such that for any two real numbers a and b with $b > a$,*

$$\Pr[a < X < b] = \int_a^b f_X(x)\, dx,$$

*$f_X(x)$ being called the **probability density function**[2] (pdf) of X.*

[2]Note that taking close or open intervals does not change the probabilities:

$$\begin{aligned} \Pr[a < X < b] &= \Pr[a \leq X < b] \\ &= \Pr[a < X \leq b] \\ &= \Pr[a \leq X \leq b]. \end{aligned}$$

The probability density function is analogous to the probability mass function in the discrete case and it satisfies the following properties

1. $f_X(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Example 6 *Consider the random variable X with pdf given by*

$$f_X(x) = \begin{cases} c\left(4x - 2x^2\right) & : \quad 0 < x < 2 \\ 0 & : \quad otherwise \end{cases}$$

and suppose we want determine the value of c. In order to do that we need to impose the condition

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\,dx &= 1 \\ \int_{-\infty}^{0} 0dx + \int_{0}^{2} c\left(4x - 2x^2\right)dx + \int_{2}^{+\infty} 0dx &= 1 \\ \int_{0}^{2} c\left(4x - 2x^2\right)dx &= 1 \\ c\left[2x^2 - \frac{2}{3}x^3\right]_0^2 &= 1 \end{aligned}$$

so that

$$c = \frac{3}{8}.$$

Analogously to the discrete case, the cdf can be defined also for continuous random variables

Definition 8 *Consider a continuous random variable X; the **cumulative distribution function** is defined as*

$$F_X(x) = \Pr[X \leq x] = \int_{-\infty}^{x} f_X(a)\,da$$

Therefore, whenever $F_X(x)$ has a derivative, we have that

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

The cdf of a continuous random variable satisfies the following properties, which are analogous to the case of a discrete random variable:

1. $\lim_{x\to-\infty} F_X(x) = 0$
2. $\lim_{x\to\infty} F_X(x) = 1$
3. If $x_1 > x_2$ then $F_X(x_1) \geq F_X(x_2)$
4. $\Pr[a < X < b] = F_X(b) - F_X(a)$. This implies that

$$\Pr[X = a] = F_X(a) - F_X(a) = 0$$

1.2.4 Random Vectors

So far we have dealt with univariate random variables. We can take a more general approach by working with multivariate random variables. This is done by introducing the concept of **random vector**, that is of an ordered collection of univariate random variables. The concepts of probability distribution and probability density function can be generalised in a straightforward way in order to be applied to random vectors.

Definition 9 *Consider the discrete random vector (X, Y) with X taking values in $\{x_1, x_2, \ldots\}$ and Y taking values in $\{y_1, y_2, \ldots\}$. The **joint probability mass function** $p_{XY}(x, y)$ is defined as*

$$p_{XY}(x, y) = \Pr(X = x, Y = y).$$

The joint probability mass function satisfies the following properties:

1. $0 \leq p_{XY}(x, y) \leq 1, \forall (x, y)$.
2. $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{XY}(x_i, y_j) = 1$.

Definition 10 *Consider a continuous random vector (X, Y). The **joint density function** $f_{XY}(x, y)$ is such that*

$$\Pr\left[\underline{x} < X < \overline{x}, \underline{y} < Y < \overline{y}\right] = \int_{\underline{y}}^{\overline{y}} \int_{\underline{x}}^{\overline{x}} f_{XY}(x, y)\, dxdy.$$

The joint density function satisfies the following properties:

1. $f_{XY}(x, y) \geq 0, \forall (x, y)$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dxdy = 1$.

When working with random vectors we sometimes need to consider the probability distribution or the density of a single component, e.g. $p_X(\cdot)$ or $f_X(\cdot)$ defined in sections 1.2.2 and 1.2.3 respectively. In this context we refer to $p_X()$ as the **marginal mass function of** X and to $f_X(\cdot)$ as the **marginal density of** X. Note that the joint distribution contains more information than the marginal distributions of all the components of the vector. In particular, the marginal mass function can be obtained from the joint mass function as

$$p_X(x) = \sum_{j=1}^{\infty} p_{XY}(x, y_j)$$

while the marginal densities are given by

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

Obviously the concepts introduced in this section can be extended to collections of n random variables $(X_1, X_2, \ldots, X_n)$.

Analogously to the univariate case, we can also define the **joint cumulative distribution function**:

$$F_{XY}(x, y) = \Pr[X \leq x, Y \leq y]$$

and are evaluated as

$$F_{XY}(x, y) = \sum_{u \leq x} \sum_{v \leq y} p_{XY}(u, v)$$

if the variables are discrete, and as

$$F_{XY}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(u, v)\, du dv$$

if the variables are continuous. Further, they both satisfy the following properties:

1. $\lim_{x \to -\infty} F_{XY}(x, y) = \lim_{y \to -\infty} F_{XY}(x, y) = 0$.
2. $\lim_{x \to \infty} F_{XY}(x, y) = F_Y(y)$.
3. $\lim_{y \to \infty} F_{XY}(x, y) = F_X(x)$.
4. $\lim_{x \to \infty} [\lim_{y \to \infty} F_{XY}(x, y)] = 1$.
5. $f_{XY}(x, y) = \dfrac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$ for continuous random variables.

1.2.5 Conditional distribution, conditional density and independence

In section 1.1.2 we saw that given two events A and B, the conditional probability of A given B (denoted by $\Pr[A|B]$) is given by

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

provided that $\Pr[B] > 0$. An analogous concept can be used in the context of random variables.

Definition 11 *Consider two discrete random variables X and Y. The **conditional probability distribution** of X given Y is defined as*

$$p_{X|Y}(x|y) = \begin{cases} \frac{p_{XY}(x,y)}{p_Y(y)} & \text{if } p_Y(y) > 0 \\ 0 & \text{else} \end{cases}.$$

Definition 12 *Consider two continuous random variables X and Y. The **conditional probability density** of X given Y is defined as*

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{XY}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ 0 & \text{else} \end{cases}.$$

Definition 13 *(**Independence**) Two discrete random variables X and Y are said to be independent if their joint distribution is the product of their marginals:*

$$p_{XY}(x,y) = p_X(x)\, p_Y(y).$$

Similarly, two continuous random variables X and Y are said to be independent if their joint density is the product of their marginals:

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

1.3 Moments

Definition 14 *Consider a random variable X; its **expected value** (when it exists), is a number, denoted by $E[X]$, defined as*

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i p_X(x_i)$$

if X is discrete or as

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

if X is continuous.

The intuition is straightforward: the expected value represents the numerical outcome we should expect from the experiment itself.

If a and b are two real numbers, then the expected value has the following properties:

1. $\mathrm{E}[a] = a$. Formally, this is because[3]

$$\begin{aligned} \mathrm{E}[a] &= \int_{-\infty}^{\infty} a f_X(x)\, dx \\ &= a \int_{-\infty}^{\infty} f_X(x)\, dx \\ &= a. \end{aligned}$$

2. $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$. This is because

$$\begin{aligned} \mathrm{E}[aX + b] &= \int_{-\infty}^{\infty} (ax + b) f_X(x)\, dx \\ &= a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} f_X(x)\, dx \\ &= a\mathrm{E}[X] + b. \end{aligned}$$

Definition 15 *Let $g : \mathbb{R} \to \mathbb{R}$ a continuous function. We define*

$$\mathrm{E}[g(X)] = \sum_{i=1}^{\infty} g(x_i)\, p_X(x_i)$$

if X is discrete and

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

if X is continuous.

[3]From now on, whenever we have to prove a result involving the expected value operator we will only work with continuous random variables, as this is the case you are most likely to encounter during the Econometrics course. The results, however, apply also to the case of discrete random variables.

Similarly to the univariate case, we can define the expected value for random vectors. Formally, this is defined as the vector whose components are the single expected values:

$$\mathrm{E}\left[(X,Y)\right] = \left(\mathrm{E}\left[X\right], \mathrm{E}\left[Y\right]\right).$$

The expected value operator applied to random vectors has the following properties:

1. $\mathrm{E}\left[aX + bY\right] = a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right]$ where a and b two real numbers.

2. Let $g : \mathbb{R}^2 \to \mathbb{R}$ a continuous function. Then

$$\mathrm{E}\left[g\left(X,Y\right)\right] = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} g\left(x_i, y_j\right) p_{XY}\left(x_i, y_j\right)$$

   if X and Y are discrete and

$$\mathrm{E}\left[g\left(X,Y\right)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g\left(x,y\right) f_{XY}\left(x,y\right) dxdy$$

   if X and Y are continuous.

We can also calculate conditional expected values. Recall from section 1.2.5 that given two random variables X and Y, we have

$$p_{X|Y}\left(x\left|y\right.\right) = \begin{cases} \frac{p_{XY}(x,y)}{p_Y(y)} & \text{if } p_Y\left(y\right) > 0 \\ 0 & \text{else} \end{cases},$$

if X and Y are discrete and

$$f_{X|Y}\left(x\left|y\right.\right) = \begin{cases} \frac{f_{XY}(x,y)}{f_Y(y)} & \text{if } f_Y\left(y\right) > 0 \\ 0 & \text{else} \end{cases}, \tag{1}$$

if X and Y are continuous. The conditional expected value is defined as follows:

Definition 16 *(**Conditional expected value**) Consider two random variables X and Y. Then the conditional expected value of X with respect to Y is defined as*

$$\mathrm{E}\left[X\left|Y = y\right.\right] = \sum_{i=1}^{\infty} x_i p_{X|Y}\left(x_i\left|y\right.\right)$$

for the discrete case and

$$\mathrm{E}\left[X\left|Y=y\right.\right]=\int_{-\infty}^{\infty}xf_{X|Y}\left(x\left|y\right.\right)dx$$

for the continuous case.

The conditional and unconditional expected values of a random variable X are linked as follows:

Proposition 3 *(**Law of iterated expectations**) Consider two random variables X and Y; then*

$$\mathrm{E}\left[X\right]=\mathrm{E}_{Y}\left[\mathrm{E}\left[X\left|Y\right.\right]\right]$$

where $\mathrm{E}_{Y}[\cdot]$ indicates the expected value obtained by treating Y as a random variable.

The law of iterated expectations can be proved as follows:

$$\begin{aligned}
\mathrm{E}_{Y}\left[\mathrm{E}\left[X\left|Y\right.\right]\right] &= \int_{y}\left[\int_{x}xf_{X|Y}\left(x\left|y\right.\right)dx\right]f_{Y}\left(y\right)dy \\
&= \int_{y}\left[\int_{x}x\frac{f_{XY}\left(x,y\right)}{f_{Y}\left(y\right)}dx\right]f_{Y}\left(y\right)dy \\
&= \int_{x}x\left[\int_{y}f_{XY}\left(x,y\right)dy\right]dx \\
&= \int_{x}xf_{X}\left(x\right)dx \\
&= \mathrm{E}\left[X\right]
\end{aligned}$$

The concept of expected value can be related to that of independence of two random variables, discussed in section 1.2.5. In particular, if X and Y are independent then

$$\mathrm{E}\left[XY\right]=\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] \tag{2}$$

This is because

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)\,dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)\,f_Y(y)\,dxdy \\
&= \left[\int_{-\infty}^{\infty} xf_X(x)\,dx\right]\left[\int_{-\infty}^{\infty} yf_Y(y)\,dy\right] \\
&= \mathrm{E}[X]\,\mathrm{E}[Y]
\end{aligned}$$

Note that the independence assumption is sufficient but not necessary for (2) to hold, that is (2) may hold even if X and Y are not independent.

The concept of expected value can be generalised to that of $k-th$ moment:

Definition 17 *The $k-th$ **moment** of a random variable X (when it exists), denoted by $\mu_k^{'}$, is the expected value of the $k-th$ power of X. Formally*

$$\mu_k^{'} = \sum_{i=1}^{\infty} x_i^k p_X(x_i)$$

if X is discrete and

$$\mu_k^{'} = \int_{-\infty}^{\infty} x^k f_X(x)\,dx$$

if X is continuous.

Clearly, the first moment of a random variable X is just its expected value.

The concept of $k-th$ moment can be generalised to that centered moment, which is defined as follows:

Definition 18 *The **centered moment** of a random variable X (when it exists), denoted by μ_k, is the expected value of the $k-th$ power of $X-\mathrm{E}[X]$. Formally*

$$\begin{aligned}
\mu_k &= \mathrm{E}\left[(X-\mathrm{E}[X])^k\right] \\
&= \begin{cases} \sum_{i=1}^{\infty}(x_i-\mathrm{E}[X])^k p_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty}(x-\mathrm{E}[X])^k f_X(x)\,dx & \text{if } X \text{ is continuous} \end{cases}
\end{aligned}$$

Therefore, the centered moment is obtained from the expression for the $k-th$ moment $\mu_k^{'}$ by replacing X with $X - \mathrm{E}[X]$. When $k = 2$ we obtain the variance of the random variable X, formally defined as follows

Definition 19 *Given a random variable X, its **variance** is defined as*

$$\mu_2 = \mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right].$$

The variance can be interpreted as an indicator of variability: the smaller it is, the more concentrated a random variable is around its mean. Rather than from the definition, computation of the variance is often simplified as follows:

Lemma 4 *The variance of a random variable X may be computed as*

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

Proof

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2X\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2(\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.
\end{aligned}$$

The variance of a random variable X has the following properties:

1. If a is a constant then $\mathrm{Var}[a] = 0$.
2. If a and b are real-valued constants then $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$. This is because

$$\begin{aligned}
\mathrm{Var}[aX + b] &= \int_{-\infty}^{\infty} (ax + b - \mathrm{E}[aX + b])^2 f_X(x)\, dx \\
&= a^2 \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 f_X(x)\, dx \\
&= a^2\mathrm{Var}[X]
\end{aligned}$$

Since $\mathrm{Var}[X] \geq 0$ for any random variable X, we can define the **standard deviation** as $\sqrt{\mathrm{Var}[X]}$. The latter is an alternative indicator of variability and has the same scale as X (i.e. if X is a profit measured in pounds, the standard deviation is also an amount expressed in pounds).

In some applications we may be interested in assessing the degree of dependence between two univariate random variables. This can be done by deploying the concept of covariance, which is defined as follows:

Definition 20 *The **covariance** between two random variables X and Y is defined as*

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{3}$$

Obviously, if $X = Y$ then $\mathrm{Cov}[X, Y] = \mathrm{Var}[X]$. Rather than from the definition, computation of the variance is often simplified by using the formula

$$\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]. \tag{4}$$

which is directly obtained from (3).

From (4) two results follow:

1. whenever $\mathrm{E}[X]$ or $\mathrm{E}[Y]$ (or both) are equal to zero then $\mathrm{Cov}[X, Y] = \mathrm{E}[XY]$.

2. if two random variables are independent then $\mathrm{Cov}[X, Y] = 0$ (although the converse is not necessarily true). This results requires taking (2) into account.

From the definition of covariance in (3) the following properties follow:

1. $\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X]$.

2. The covariance between linear combinations of random variables can be expressed as

   $$\mathrm{Cov}[aX, bY] = ab\mathrm{Cov}[X, Y],$$

   where a and b are real numbers.

3. The variance of the sum of two random variables is given by

   $$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y];$$

this is because

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

and implies that when $\mathrm{Cov}\left[X,Y\right]=0$, the variance of $X+Y$ is just the sum of the variances.

A disadvantage of using the covariance as a measure of the degree of dependence between two random variables is that it is unbounded, in the sense that it can take any value in $\mathbb{R}$. In order to overcome this problem, we can define the **(linear) correlation coefficient** as

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

It can be shown that for any random vector (X,Y),

$$-1\leq\rho\left[X,Y\right]\leq 1;$$

when $\rho\left[X,Y\right]=0$, then X and Y are said to be **uncorrelated**. The following properties hold:

1. $\rho\left[X,Y\right]=-1$ if and only if $Y=aX+b, a<0, a, b$ constant.
2. $\rho\left[X,Y\right]=1$ if and only if $Y=aX+b, a>0, a, b$ constant.
3. If X and Y are independent then $\rho\left[X,Y\right]=0$ (although the converse is not true); this means that if two random variable are independent they are also uncorrelated.

The concept of covariance can be generalised so to be applied to random vectors of any dimensions. Consider the $n\times 1$ random vector $\mathbf{x}=(X_1,\ldots,X_n)'$; the **variance matrix** is an $n\times n$ matrix defined as

$$\underset{n\times n}{Var\left[\mathbf{x}\right]} = \mathrm{E}[\underset{n\times 1}{\left(\mathbf{x}-\mathrm{E}\left[\mathbf{x}\right]\right)}\ \underset{1\times n}{\left(\mathbf{x}-\mathrm{E}\left[\mathbf{x}\right]\right)'}].$$

The entries on the diagonal are just the variances of $X_1,\ldots,X_n$. The off-diagonal entries are $\mathrm{Cov}\left[X_i,X_j\right], i\neq j$. The covariance matrix has the useful

property that if the vector $\mathbf{x}$ is premultiplied by an $m \times n$ non-stochastic matrix $\mathbf{A}$ then

$$\underset{m \times 1}{\text{Var}(\quad \mathbf{Ax} \quad)} = \underset{m \times n}{\mathbf{A}} \quad \underset{n \times n}{\text{Var}[\mathbf{x}]} \quad \underset{n \times m}{\mathbf{A}'}.$$

where $\text{Var}[\mathbf{Ax}]$ is an $m \times m$ matrix. In order to prove this result, define

$$\mathbf{y} = \mathbf{Ax}$$

so that

$$\begin{aligned} \text{Var}[\mathbf{Ax}] &= \text{Var}[\mathbf{y}] \\ &= \text{E}\left[(\mathbf{y} - \text{E}[\mathbf{y}])(\mathbf{y} - \text{E}[\mathbf{y}])'\right] \\ &= \text{E}\left[\mathbf{A}(\mathbf{x} - \text{E}[\mathbf{x}])(\mathbf{x} - \text{E}[\mathbf{x}])'\mathbf{A}'\right] \\ &= \mathbf{A}\text{E}\left[(\mathbf{x} - \text{E}[\mathbf{x}])(\mathbf{x} - \text{E}[\mathbf{x}])'\right]\mathbf{A}' \\ &= \mathbf{A}\text{Var}[\mathbf{x}]\mathbf{A}'. \end{aligned}$$

Note that the variance matrix is symmetric and positive semidefinite.

1.3.1 Two useful rules

1. If X is a random variable and $g(.)$ a function then in general

$$E(g(X)) \neq g(E(X))$$

2. If $f(X, \theta)$ is a function differentiable with respect to θ and X a random variable then

$$\frac{\partial}{\partial \theta} E_X(f(X, \theta)) = E_X\left(\frac{\partial}{\partial \theta} f(X, \theta)\right)$$

that is (under mild regularity conditions) one may interchange the expectation and differentiation operators.

1.4 Distribution Theory

1.4.1 The Bernoulli distribution

A random variable X has a Bernoulli distribution if it can only take two values, $X = 0$ or $X = 1$. Its probability density function is given by

$$p_X(x) = \pi^x \cdot (1 - \pi)^{1-x}$$

where π denotes the probability of observing the event $X = 1$. Its expected value and variance are respectively given by

$$\begin{aligned} \mathrm{E}[X] &= \pi \\ \mathrm{Var}[X] &= \pi(1-\pi). \end{aligned}$$

The Bernoulli distribution will be extensively used in the context of binary choice models.

1.4.2 The Poisson distribution

A random variable X has a Poisson distribution if it can take values on $\mathbb{N}$. Its probability density function is given by

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Its expected value and variance are given by

$$\mathrm{E}[X] = \mathrm{Var}[X] = \lambda.$$

1.4.3 The Normal (Gaussian) distribution

If a random variable X is normally distributed with parameters μ and σ^2 (in short $X \sim N(\mu, \sigma^2)$) it has density

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

characterised by the two parameters μ and σ^2. The function is bell-shaped and symmetric around μ. The first two moments of the Gaussian distribution are its parameters:

$$\begin{aligned} \mathrm{E}[X] &= \mu \\ \mathrm{Var}[X] &= \sigma^2. \end{aligned}$$

There is no closed form expression for the cdf $F_X(x)$, but its values are tabulated for the case $\mu = 0$ and $\sigma^2 = 1$ (the *standard* normal distribution). Conventionally, the density function of the standard normal is denoted by

$\phi(z)$, while its cdf by $\Phi(z)$. Any given normally distributed random variable X with parameters $\mu \neq 0$ and $\sigma^2 \neq 1$ can be thought of as a linear transformation of the standard normal with

$$X = \mu + \sigma Z$$

and $Z \sim N(0,1)$. As a result, the probability that X falls in any interval (a,b) may be computed as a function of probabilities involving Z:

$$\begin{aligned}
\Pr[a < X < b] &= \Pr[a < \mu + \sigma Z < b] \\
&= \Pr[a - \mu < \sigma Z < b - \mu] \\
&= \Pr\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right] \\
&= \Pr\left[Z < \frac{b-\mu}{\sigma}\right] - \Pr\left[Z < \frac{a-\mu}{\sigma}\right] \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).
\end{aligned}$$

It is particularly convenient to express the probability in this form because values of $\Phi(z)$ for some useful points z are available from the statistical tables and are known good numerical ways to approximate $\Phi(z)$ to any desired degree of accuracy for any x.

The Gaussian distribution has a multivariate form that can be used to work with random vectors. A random vector $\mathbf{x} = (X_1, \ldots, X_n)'$ has a multivariate normal distribution with expected value $\mathrm{E}[\mathbf{x}] = \boldsymbol{\mu}$ and variance matrix $\mathrm{Var}[\mathbf{x}] = \boldsymbol{\Sigma}$ if its joint distribution is given by

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} \det[\boldsymbol{\Sigma}]^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}.$$

A very useful property is that any linear combination of jointly Gaussian random variables is also Gaussian. Consider an $m \times n$ non-stochastic matrix $\mathbf{A}$. It can be shown that $\mathbf{Ax}$ (a linear combination of the components of $\mathbf{x}$) is an $m \times 1$ Gaussian random vector and in particular

$$\underset{m \times 1}{\mathbf{Ax}} \quad \sim N(\quad \underset{m \times 1}{\mathbf{A}\boldsymbol{\mu}}, \quad \underset{m \times m}{\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'} \quad)$$

1.4.4 The χ^2 distribution

Consider v *independent* random variables $Z_i,\ i = 1, ..., v$ with standard normal distribution, that is $Z_i \sim N(0, 1)$; then the random variable X defined as

$$X = \sum_{i=1}^{v} (Z_i)^2$$

has a chi-squared (χ^2) distribution with v *degrees of freedom*. Formally, the density function of X is given by

$$f_X(x) = \begin{cases} \frac{(1/2)^{v/2}}{\Gamma(v/2)} x^{\frac{v}{2}-1} e^{-x/2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$ is the Gamma function. Its first two moments are

$$\begin{aligned} \mathrm{E}[X] &= v \\ \mathrm{Var}[X] &= 2v. \end{aligned}$$

1.4.5 The Student's t distribution

The Student's t distribution with v degrees of freedom is obtained as the ratio between a standard normal and the square root of a χ^2_v divided by the number of degrees of freedom v. Formally if $Z \sim N(0, 1)$, $X \sim \chi^2_v$ and Z and X are *independent* then

$$\frac{Z}{\sqrt{X/v}} \sim t_v.$$

The density function is bell-shaped and symmetric around zero.

1.4.6 The F distribution

The F distribution with v_1 and v_2 degrees of freedom (where v_1 denotes the degrees of freedom of the numerator and v_2 those of the denominator) is obtained as the distribution of the ratio of two independent random variables with χ^2 distribution, each divided by its degrees of freedom. Formally,

$$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1, v_2}$$

if $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{v_2}$, X_1 and X_2 being *independent*.

The F distributed random variable can only take positive values (since it's the ratio of two random variables that can only take positive values).

1.5 Statistical inference

1.5.1 Definitions

In statistical inference we start with a probability model to describe the behaviour of a population of interest. Some aspects of the model are however unknown, possibly because we know the functional form of a parametric distribution function but not the values of its parameters. The aim of inference is to assign a value to those unknown parameters. We now introduce the concepts of sample, observation and statistic.

Definition 21 *A **sample** is a collection of random variables (say $X_1, X_2, \ldots, X_n$) from the population. If the random variables constituting the sample are independently and identically distributed, then the sample is said to be **random**.*

Definition 22 *An **observation** is the known value that each collected random variable assumes (say $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$).*

Definition 23 *A **statistic** is a function of the sample (say $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)'$). Therefore, a statistic is random variable.*

Example 7 *Some useful statistics are:*

1. *Sample mean, $\bar{X} = \sum_{i=1}^{n} X_i / n$;*
2. *Sample variance, $\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 / n$;*
3. *Sample covariance (when we jontly sample two variables X and Y)*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(X_i - \bar{X}\right)\left(Y_j - \bar{Y}\right) / n \, .$$

A possible source of confusion arises from the use of the terms *sample moments* and *population moments*. The former are descriptive statistics, and therefore random variables whose value (or realisation) is known. The latter are (usually) unknown constants; see 1.3.

Statistical inference can be classified into point estimation and interval estimation: in the former case we are interested in assigning a value to an unknown parameter; in the latter case we define an interval within which we are confident that the true value of the unknown parameter falls.

1.5.2 Point estimation

Point estimation is an inferential procedure that tries to learn about the value of an unknown parameter, say θ, from the sample information.

Definition 24 *(**Estimator and Estimate**) An **estimator** is a statistic we use to assign a value to an unknown parameter. An **estimate** is the actual value assumed by the estimator. In other words, the estimator is a random variable, and thus has a distribution; an estimate is a particular realised value of the estimator.*

1.5.3 Methods of point estimation

For any estimation problem one can think of different alternative estimators. Possible general strategies to obtain point estimators are the following.

Method of moments The method of moments estimator is based upon substituting unknown population moments by their sample counterparts.

As an example, consider the random sample $X_1, \ldots, X_n$. Suppose we want to estimate the centered population moment

$$\theta = \mathrm{E}\left[\left(X - \mathrm{E}\left[X\right]\right)^k\right]$$

The sample counterpart is given by

$$\hat{\theta}_{MM} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(X_i - \bar{X}\right)^k.$$

Note that in order to estimate θ by method of moments we did not have to fully specify the distribution of the underlying population.

Maximum likelihood A general and powerful method of estimation is that known as maximum likelihood. The idea behind maximum likelihood estimation consists in finding the estimate for the unknown parameter θ that maximises the likelihood of observing the sample.

Suppose we start with a random sample of n observations $X_1, \ldots, X_n$ drawn from a probability density function $f\left(X \mid \theta\right)$ involving an unknown parameter θ. The space Θ is the set of permissible values for θ. If we denote the complete set of observations by the vector $X = (X_1, \ldots, X_n)$, then the joint probability density function of X may be written as

$$f(X \mid \theta) = f(X_1 \mid \theta) f(X_2 \mid \theta) \dots f(X_n \mid \theta)$$

In probability theory the function $f(X \mid \theta)$ gives the probability of observing the sample given the parameter θ (remembering that a random sample gives independent drawings from the population so the probabilities multiply). In particular if

$$f(x_1, \dots, x_n \mid \theta) > f(x_1^{'}, \dots, x_n^{'} \mid \theta)$$

then we may (loosely) say that observing $x_1, \dots, x_n$ is "more likely" than observing $x_1^{'}, \dots, x_n^{'}$. However, in estimation theory we observe the value of X and wish to say something about θ. Now consider $f(X \mid \theta)$ as a function of θ (with X fixed at the observed values). We call this the *likelihood function* of θ. If

$$f(X \mid \theta_1) > f(X \mid \theta_2)$$

we may (loosely) say that θ_1 is a "more plausible" value of θ than θ_2, since θ_1 assigns a larger probability to the observed X than does θ_2. The *principle of maximum likelihood* just says that we should use as our estimator of θ the parameter point for which our observed sample is most likely. That is, we should choose the value of θ which maximises the likelihood function. To emphasise that we have moved from considering the probability of the sample given the parameter θ to considering the likelihood of the parameter θ given the sample we use a different notation for the likelihood function, $L(\theta \mid X)$. So

$$L(\theta \mid X) = f(X \mid \theta)$$

The maximum likelihood estimator is therefore the value of θ that maximises $L(\theta; X_1, \dots, X_n)$, that is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; X_1, \dots, X_n).$$

If the likelihood function is differentiable as a function of θ then possible candidates for the maximum likelihood estimator are the values of θ for which

$$\frac{\partial}{\partial \theta} L(\theta \mid X) = 0$$

Note that solutions are only candidates for the MLE, the derivative being zero is only necessary for a maximum, not sufficient. The second derivative

must be negative for a maximum, though this guarantees only a local maximum, not global. In addition if the parameter θ is restricted to lie in some region Θ then the extrema may occur on the boundary of Θ, and the derivative need not be zero there. Maximum likelihood estimates can also be found by direct maximisation of the likelihood function via numerical search methods.

As an example consider a random sample $X_1, \ldots, X_n$ drawn from a Bernoulli distribution

$$p_X(x) = \pi^x (1-\pi)^{1-x}$$

where π is the unknown parameter. The likelihood function is then given by

$$\begin{aligned} L(\pi; x_1, \ldots, x_n) &= \prod_{i=1}^{n} \left[\pi^{x_i} (1-\pi)^{1-x_i}\right] \\ &= \pi^{\left(\sum_{i=1}^n x_i\right)} (1-\pi)^{\left(n-\sum_{i=1}^n x_i\right)} \end{aligned}$$

Taking logs we obtain the **log likelihood function**[4]

$$l(\pi; x_1, \ldots, x_n) = \sum\nolimits_{i=1}^{n} x_i \log \pi + \left(n - \sum\nolimits_{i=1}^{n} x_i\right) \log(1-\pi);$$

differentiating with respect to π we have

$$\frac{\partial l(\pi; x_1, \ldots, x_n)}{\partial \pi} = \frac{\sum_{i=1}^n x_i}{\pi} - \frac{n - \sum_{i=1}^n x_i}{1-\pi}$$

so that the the maximum likelihood *estimate* of π is obtained by setting this to zero

$$\frac{\sum_{i=1}^n x_i}{\hat{\pi}_{ML}} - \frac{n - \sum_{i=1}^n x_i}{1-\hat{\pi}_{ML}} = 0$$

so

$$\hat{\pi}_{ML} = \frac{\sum_{i=1}^n x_i}{n}$$

while the maximum likelihood *estimator* is

$$\hat{\pi}_{ML} = \frac{\sum_{i=1}^n X_i}{n}.$$

Note that in order to estimate θ by maximum likelihood we have to make assumptions about the distribution of the population.

[4]Note that given a function $f(x)$, $\log f(x)$ is a monotonic transformation of $f(x)$. Therefore, $\log f(x)$ achieves the maximum at the same point as $f(x)$.

Least squares In the least squares approach we consider characteristics of the residuals, i.e. the features of the observations x_i that are not explained by the model. The residual $e_i(\theta)$ is defined as the difference between the observation x_i and the predicted value of x_i implied by θ, say $\hat{x}_i(\theta)$, that is

$$e_i(\theta) = x_i - \hat{x}_i(\theta).$$

The sum of squared residuals is

$$S(\theta) = \sum_{i=1}^{n} e_i^2(\theta)$$

The resulting least squares estimator is the value of θ (say $\hat{\theta_{LS}}$) that minimises $S(\theta)$, that is

$$\hat{\theta}_{LS} = \arg\min_{\theta} S(\theta).$$

As an example, suppose we want to estimate the expected value θ of a random variable X. Therefore, given the random sample $X_1, \ldots, X_n$ we have

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (X_i - \theta)^2.$$

Taking derivatives with respect to θ

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{n} (X_i - \theta)^2 = -2 \cdot \sum_{i=1}^{n} (X_i - \theta).$$

Setting it equal to 0 we have

$$\begin{aligned} -2 \cdot \sum_{i=1}^{n} \left(X_i - \hat{\theta}\right) &= 0 \\ \sum_{i=1}^{n} X_i - n\hat{\theta} &= 0 \\ \hat{\theta} &= \frac{1}{n} \sum_{i=1}^{n} X_i \end{aligned}$$

Note that in order to estimate θ by least squares we did not have to make any assumption about the distribution of the population.

1.5.4 Selection criteria for point estimators

Definition 25 *(**Unbiasedness**) An estimator $\hat{\theta}$ of an unknown parameter θ is said to be unbiased if its expected value is equal to θ, that is*

$$\mathrm{E}\left[\hat{\theta}\right] = \theta.$$

Definition 26 *(**Bias**) The **bias** of an estimator is defined as*

$$bias\left(\hat{\theta}\right) = \mathrm{E}\left[\hat{\theta}\right] - \theta$$

Unbiasedness *per se* is not a strong requirement. Many criteria have been developed to select the best estimator among several alternative unbiased estimators. A very important property of any estimator is its variance: estimators with lower variance should be preferred because the probability of getting a point estimate close to the true value is high.

Definition 27 *An estimator is the **best linear unbiased estimator** (**BLUE**) if it is unbiased, it is linear and has the minimum variance among the linear unbiased estimators.*

Theorem 5 *(**Cramer-Rao Lower Bound**) Under some regularity conditions, the variance of an unbiased estimator of an unknown parameter θ will always be at least as large as*

$$\left[I\left(\theta\right)\right]^{-1} = \left[-\mathrm{E}\left(\frac{\partial^2 \log L\left(\theta\right)}{\partial \theta^2}\right)\right]^{-1} = \left\{\mathrm{E}\left[\left(\frac{\partial \log L\left(\theta\right)}{\partial \theta}\right)^2\right]\right\}^{-1}$$

*where $I\left(\theta\right)$ is the **information matrix**.*

Sometimes we are prepared to tolerate a small bias if the estimator has a considerably smaller variance. To choose between alternative estimators when this trade-off is present, we can select the one with the smallest possible mean squared error, defined as

Definition 28 *(**Mean Squared Error**) The **mean squared error** (MSE) of an estimator $\hat{\theta}$ is defined as*

$$\mathrm{MSE}\left[\hat{\theta}\right] = \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right].$$

The MSE takes into account both the variance and the bias of the estimator $\hat{\theta}$. This can more clearly be seen by writing the MSE as

$$\begin{aligned} \text{MSE}\left[\hat{\theta}\right] &= \text{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\ &= \text{Var}\left[\hat{\theta}\right] + \left(bias\left(\hat{\theta}\right)\right)^2. \end{aligned}$$

Obviously, the MSE of an unbiased estimator is equal to its variance.

1.5.5 Interval estimation

In some cases, instead of using a point estimate of an unknown parameter we might be interested in constructing an interval estimate (or confidence interval).

Definition 29 *A **confidence interval** at a given level $\alpha \in (0,1)$ for a parameter θ is an interval in $\mathbb{R}$ such that the true value of the parameter lies within the interval with a preassigned probability equal to $(1-\alpha)$.*

Note that the bounds of the confidence interval depend on the available sample. The key concept behind the construction of a confidence interval is that of pivotal quantity.

Definition 30 ***A pivotal quantity** is a function of both a point estimator and an unknown parameter that has a known distribution.*

Note that a pivotal quantity *is not* a statistic, so its value cannot be computed. We will now see how confidence intervals can be built for the Normal and Student's t distributions.

Case 1 Consider a statistic $\hat{\theta}$, which is known to be normally distributed with mean θ and *known* variance σ^2/n, that is

$$\hat{\theta} \sim N\left(\theta, \frac{\sigma^2}{n}\right);$$

the pivotal quantity can then be obtained by means of the linear transformation

$$q = \frac{\hat{\theta}-\theta}{\sigma/\sqrt{n}} \sim N(0,1).$$

It is easy to see why q is a pivotal quantity: it is a function of $\hat{\theta}$ (a point estimator) and θ (an unknown parameter) and its distribution (the standard normal) is known because it does not depend upon any unknown parameter. Since the distribution of q is known regardless the value of θ, we can find a scalar $c > 0$ from statistical tables so that

$$\Pr\left[-c \leq \frac{\hat{\theta} - \theta}{\sigma / \sqrt{n}} \leq c\right] = (1 - \alpha).$$

Since q is a linear function of $\hat{\theta}$ then

$$\Pr\left[\hat{\theta} - \frac{\sigma c}{\sqrt{n}} \leq \theta \leq \hat{\theta} + \frac{\sigma c}{\sqrt{n}}\right] = (1 - \alpha).$$

Note that $\hat{\theta} - \sigma c / \sqrt{n}$ and $\hat{\theta} + \sigma c / \sqrt{n}$ are known functions of c (which is a known constant) and $\hat{\theta}$, which can be computed from the data. Therefore, the bounds of the confidence interval can be computed and we have the required results. Typical values for α are $\alpha = 0.05, 0.1$.

Case 2 Consider a statistic $\hat{\theta}$, which is known to be normally distributed with mean θ and variance σ^2 / n, *both unknown*. In addition suppose that the unknown quantity σ^2 can be estimated by means of a statistic s^2, which is *independent* of $\hat{\theta}$, and is such that

$$\frac{s^2}{\sigma^2} = \frac{Z}{v} \text{ where } Z \sim \chi^2_v.$$

The following ratio

$$\left(\frac{\hat{\theta} - \theta}{\sigma / \sqrt{n}}\right) \Bigg/ \sqrt{\frac{s^2}{\sigma^2}} = \sqrt{n}\frac{\hat{\theta} - \theta}{s} \sim t_v$$

is a pivotal quantity[5]. Therefore, we can construct a confidence interval using Student's t tables:

$$\Pr\left[\hat{\theta} - \frac{st^*}{\sqrt{n}} \leq \theta \leq \hat{\theta} + \frac{st^*}{\sqrt{n}}\right] = (1 - \alpha),$$

where $t^* > 0$ is the value that leaves a probability of $\alpha / 2$ in the right tail of the t_v distribution.

[5]See section 1.4.5 to understand why $\sqrt{n}\frac{\hat{\theta} - \theta}{s} \sim t_v$.

1.5.6 Hypothesis testing

Often in empirical economic studies we need to obtain a rule to assess the validity of a certain assumption about the underlying population. As we have already discussed, the population is described by a model; therefore, imposing an assumption on the population is equivalent to imposing an assumption on the model. For example, we might want to assess whether a certain factor affects the dependent variable of our model. Intuitively, we have to run a test to decide whether our assumption about the model is "right" or should be abandoned in favour of another assumption. Formalisation of this intuition leads us to the following three definitions:

Definition 31 *A statistical* ***test*** *about a certain assumption is a rule that tells us, given the available data, whether we should reject the assumption or not.*

Definition 32 *The* ***null hypothesis*** *(denoted by H_0) makes a specific assertion about the population parameters.*

Definition 33 *The* ***alternative hypothesis*** *(denoted by H_1) is the hypothesis against which the null hypothesis is tested.*

In order to run the test we need a test statistic defined as follows

Definition 34 *A* ***test statistic*** *t is a function of the sample.*

The value taken by a test statistic t can fall into an **acceptance region** (denoted by A) where we *do not reject* the null hypothesis $H_0$and a **rejection region** (denoted by R) where we reject H_0. That is we divide up the sample space Ω (the totality of values that the random sample X can take) into two regions A and R such that they partition the sample space Ω; that is the two regions are mutually exclusive ($A \cap R = \phi$) and exhaustive (together they give the sample space). This partitioning of the sample space ensures that there are no possible outcomes in Ω for which we would

(a) both reject and not reject H_0

or

(b) neither reject nor not reject H_0

Since the sample is a collection of random variables, the *test statistic is a random variable*. Further, the test statistic is such that when H_0 is true its

distribution is known and does not depend upon unknown parameters. Of course, since the true population is unknown, a statistical test can always give rise to errors. Given the form of the statistical test for H_0 we have only two choices, reject or not reject H_0. The possible outcomes are

		Truth	
		H_0true	H_0false
Decision	Not Reject H_0	✓	Type II error
	Reject H_0	Type I error	✓

That is we commit a Type I Error if we reject the hypothesis H_0 when it is in fact true, and a Type II Error if we do not reject H_0 when it is false. Note that this exhausts the possibilities.

Ideally we would like the probabilities of Type I and Type II errors to both be zero. But we can't do this. We can always control for the probability of incurring in a Type I error. This is done by choosing the significance level (or **size**) of the test defined as follows:

Definition 35 *The **significance level** (or **size**) of the test is defined as*

$$\alpha = \Pr\left[Type\ I\ error\right] = \Pr\left[reject\ H_0 \,|H_0\ is\ true\right]$$

Normally the size of the test is set equal to $10\%, 5\%$ or 1%. Clearly, if we decrease the size of the test we decrease the probability of wrongly rejecting a true null. However, this comes to the expense of increasing the probability of committing a Type II error: therefore, there is a trade off between Type I and Type II errors. The probability of incurring in a Type II error leads us to the definition of the power of the test:

Definition 36 *The **power** of the test is defined as the probability of rejecting a false null hypothesis, that is*

$$power = 1 - \beta = 1 - \Pr\left[Type\ II\ error\right].$$

Clearly, for a given size α we want β to be as small as possible (or equivalently the power to be as high as possible). Therefore when choosing amongst several available tests we should always go for the one that has the highest power given the chosen level of significance. This may not be such an easy

task because the power of the test depends upon the particular alternative we are considering.

The actual definition of the rejection region depends on whether the test we wish to run is two-sided or one-sided. If the test is **two-sided** then H_0 and H_1 are specified as

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

In this case, if the distribution of the test statistic has two tails (like the Gaussian or the Student's t) then R is defined as $R = \{\{t\,|t < t^*\} \cup \{t\,|t > t^{**}\}\}$ where t^* and t^{**} (the **critical values**) are chosen so that $\Pr[t\,|t^* < t < t^{**}] = \Pr[t \in A\,|H_0] = 1-\alpha$. Conversely, if the distribution of the test statistic has one tail (like the F or the χ^2) then R is defined as $R = \{t\,|t > t^*\}$, where t^* is such that $\Pr[t\,|t < t^*] = \Pr[t \in A\,|H_0] = 1-\alpha$.

If the test is **one-sided** then H_0 and H_1 are specified as

$$\begin{array}{ll} H_0 : \theta \leq \theta_0 & \\ H_1 : \theta > \theta_0 & \end{array} \quad \text{or} \quad \begin{array}{ll} H_0 : \theta \geq \theta_0 & \\ H_1 : \theta < \theta_0 & \end{array} .$$

In this case, we use a test statistic with a one-tailed distribution. Therefore, if the null hypothesis is $H_0 : \theta \leq \theta_0$, then $R = \{t\,|t > t^*\}$ where t^* is chosen so that $\Pr[t\,|t > t^*] = \alpha$; conversely, if the null hypothesis is $H_0 : \theta \geq \theta_0$, then $R = \{t\,|t < t^*\}$ where t^* is chosen so that $\Pr[t\,|t < t^*] = \alpha$.

To summarise, a hypothesis test consists of the following steps:

1. Define the null hypothesis H_0 and the alternative hypothesis H_1.
2. Decide on the appropriate test statistic.
3. Select the significance level α.
4. Define the rejection region R.
5. Formulate the test criterion: do not reject H_0 when $t \in A$, reject H_0 when $t \in R$.
6. Take your sample.
7. Compute the statistics.
8. Apply the test criterion.