Random Vectors

So far we have considered scalar random variables

If X_1 and X_2 are 2 scalar random variables, the 2×1 vector

$$X = \begin{pmatrix} X_1 \\ \\ X_2 \end{pmatrix}$$

is a random vector, i.e. a vector whose elements are scalar random variables

Our definitions of distribution functions and density functions extend to random vectors

Initially we focus on the bivariate case, i.e. random vectors with only 2 elements

Distribution Function

$$F(x_1, x_2) = \mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant x_2)$$

where $\mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant x_2)$ now denotes the probability that $X_1 \leqslant x_1$ and that $X_2 \leqslant x_2$

$F(x_1, x_2)$ describes the joint distribution of the 2 random variables X_1 and X_2

$F(x_1, x_2)$ satisfies

i) $\lim_{x_1, x_2 \to -\infty} F(x_1, x_2) = 0$ and $\lim_{x_1, x_2 \to \infty} F(x_1, x_2) = 1$

ii) $F(x_1, x_2)$ is monotonically non-decreasing in x_1, x_2

iii) $F(x_1, x_2)$ is right-continuous

Density Function

For continuous random variables

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

For discrete random variables, we define the left-limits $F(x_1^-, x_2)$, $F(x_1, x_2^-)$ and $F(x_1^-, x_2^-)$ similarly to $F(x^-)$ for the scalar case previously [so, for example, we have $F(x_1^-, x_2) = \lim_{h \to 0} F(x_1 - h, x_2)$ for $h > 0$]; then we have

$$f(x_1, x_2) = F(x_1, x_2) - F(x_1^-, x_2) - F(x_1, x_2^-) + F(x_1^-, x_2^-)$$

As in the scalar case, for discrete random variables it may be simpler to consider the probability mass function $f(x_1, x_2) = \mathrm{P}(X_1 = x_1, X_2 = x_2)$ directly

Joint density functions integrate or sum to one, so that we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2) dx_2 dx_1 = 1$$

for continuous random variables, and

$$\sum_{x1}\sum_{x2} f(x_1, x_2) = 1$$

for discrete random variables

NB. We use the notation $F(x_1, x_2)$ and $f(x_1, x_2)$ to describe the joint distribution function and the joint density function of the 2 random variables X_1 and X_2, even when we think of them jointly as the column vector X

That is, if we define

$$X = \begin{pmatrix} X_1 \\ \\ X_2 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_1 \\ \\ x_2 \end{pmatrix}$$

we still write the joint distribution function as $F(x)$ [or perhaps as $F_X(x)$] with the interpretation

$$F(x) = \mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant x_2)$$

and we still write the joint density function as $f(x)$ [or perhaps as $f_X(x)$]

Independence

Recall that 2 random variables X_1 and X_2 are said to be independent if and only if, for all x_1, x_2

$$\mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant x_2) = \mathrm{P}(X_1 \leqslant x_1)\mathrm{P}(X_2 \leqslant x_2)$$

For 2 random variables that are independent, knowing the value taken by one tells us nothing at all about the distribution of the other

From this definition, for 2 independent random variables we have

$$F_X(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$$

$$f_X(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

The joint distribution function factorises into the product of the 2 marginal distribution functions, and the joint density function factorises into the product of the 2 marginal density functions

For n independent scalar random variables $X_1, X_2, ..., X_n$, we have

$$F_X(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

$$f_X(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

For n independent and identically distributed scalar random variables $X_1, X_2, ..., X_n$, with the same distribution function and density function, these expressions simplify further

$$F(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} F(x_i)$$

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

From the definitions of $F(x_1)$ and $F(x_1, x_2)$ we can *always* obtain the marginal distribution from the joint distribution by considering

$$F(x_1, \infty) = \mathrm{P}(X_1 \leqslant x_1, X_2 \leqslant \infty) = \mathrm{P}(X_1 \leqslant x_1) = F(x_1)$$

since the outcome $X_2 \leqslant \infty$ is certain to occur

This is called **marginalisation**, or marginalising the joint distribution

Provided the joint density function exists, we have

$$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(a, b) da db$$

and

$$F(x_1) = F(x_1, \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_1} f(a, b) da db$$

Then

$$f(x_1) = \frac{\partial F(x_1)}{\partial x_1} = \int_{-\infty}^{\infty} f(x_1, b) db$$

We integrate the joint density $f(x_1, x_2)$ over x_2 to obtain the marginal density $f(x_1)$

For discrete random variables, the corresponding expression is

$$f(x_1) = \mathrm{P}(X_1 = x_1, X_2 \leqslant \infty) = \sum_b f(x_1, b)$$

where the sum is taken over all possible values of X_2

Example - suppose that X_1 and X_2 denote the outcomes from 2 tosses of a fair coin, each coded 0/1 (heads/tails)

There are 4 possible outcomes, each of which is equally likely

X_1 and X_2 are independent so, for example

$$\mathrm{P}(X_1 = 0, X_2 = 0) = \mathrm{P}(X_1 = 0)\mathrm{P}(X_2 = 0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The joint density $f(x_1, x_2)$ describes the probabilities of each of these 4 possible outcomes

$f(x_1, x_2)$	$X_2 = 0$	$X_2 = 1$	$f(x_1)$
$X_1 = 0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$X_1 = 1$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$f(x_2)$	$\frac{1}{2}$	$\frac{1}{2}$	

To obtain $f(x_1)$, we sum $f(x_1, x_2)$ over the 2 possible values of X_2, giving

$$f(0) = \mathrm{P}(X_1 = 0) = \mathrm{P}(X_1 = 0, X_2 = 0) + \mathrm{P}(X_1 = 0, X_2 = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$f(1) = \mathrm{P}(X_1 = 1) = \mathrm{P}(X_1 = 1, X_2 = 0) + \mathrm{P}(X_1 = 1, X_2 = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

And similarly for $f(x_2)$, since we have $X_1 \overset{\mathrm{D}}{=} X_2$ in this example

NB. Two (or more) independent random variables which have the same distribution are said to be **independent and identically distributed**, or iid

Aside - if instead we consider the total number of tails ($X_i = 1$) in the outcome of the 2 coin tosses, we have a scalar random variable which may take 3 possible values: $0, 1$ or 2

Note that there are 2 ways to obtain an outcome with 1 tail ($X_1 = 1, X_2 = 0$ or $X_1 = 0, X_2 = 1$), but only 1 way of obtaining the outcome with 0 tails, and only 1 way of obtaining the outcome with 2 tails

For this random variable, we have the density function

$$f(0) = \frac{1}{4}, \qquad f(1) = \frac{1}{2}, \qquad f(2) = \frac{1}{4}$$

This is an example of a discrete random variable with a **binomial** distribution

More generally, if we have n independent values of a binary random variable X_i with a Bernoulli(θ) distribution, the total number of values with $X_i = 1$, say $Y = \sum_{i=1}^{n} X_i$, has a binomial distribution, with density

$$f(y) = \left(\frac{n!}{y!(n-y)!} \right) \theta^y (1-\theta)^{n-y} \quad \text{for } y = 0, 1, ..., n$$

where $n!$ denotes n factorial $[n! = n \times (n-1) \times (n-2) \times ... \times 1$, with $0! = 1]$

Conditioning

In empirical work in economics, we are often interested in what knowing the value of one variable may tell us about the distribution of another variable, in settings where we observe outcomes for 2 (or more) variables which are not independent

Example: Suppose we observe data on weekly wages and educational attainment for a sample of male, full-time employees, aged 30-35

Consider the binary outcomes $X_{1i} = I(\text{individual } i\text{'s weekly wage} > £500)$ and $X_{2i} = I(\text{individual } i \text{ has a college degree})$

Suppose we observe the sample frequencies described below

$\widehat{f}(x_1, x_2)$	$X_{2i} = 0$	$X_{2i} = 1$	$\widehat{f}(x_1)$
$X_{1i} = 0$	0.4	0.1	0.5
$X_{1i} = 1$	0.2	0.3	0.5
$\widehat{f}(x_2)$	0.6	0.4	

Half the sample earn above £500 per week ($X_{i1} = 1$); 40% of the sample have a college degree ($X_{2i} = 1$); and 30% of the sample earn above £500 per week and have a college degree ($X_{i1} = 1$ and $X_{2i} = 1$)

Among the 40% of the sample with a college degree, three-quarters of these individuals earn above £500 per week, and only one-quarter earn less than or equal to £500 per week

We define the sample frequencies of the outcome X_{1i} **conditional** on the outcome X_{2i} taking the value 1 to be

$$\begin{aligned} \widehat{f}(X_{1i} &= 0|X_{2i} = 1) = \widehat{f}(0|X_{2i} = 1) = \frac{0.1}{0.4} = \frac{1}{4} \\ \widehat{f}(X_{1i} &= 1|X_{2i} = 1) = \widehat{f}(1|X_{2i} = 1) = \frac{0.3}{0.4} = \frac{3}{4} \end{aligned}$$

Similarly we have the sample frequencies of the outcome X_{1i} conditional on the outcome X_{2i} taking the value 0

$$\begin{aligned} \widehat{f}(X_{1i} &= 0|X_{2i} = 0) = \widehat{f}(0|X_{2i} = 0) = \frac{0.4}{0.6} = \frac{2}{3} \\ \widehat{f}(X_{1i} &= 1|X_{2i} = 0) = \widehat{f}(1|X_{2i} = 0) = \frac{0.2}{0.6} = \frac{1}{3} \end{aligned}$$

In this example, observing that an individual has a college degree tells us that this individual is more likely to earn more than £500 per week than a randomly selected individual from the sample; i.e. the conditional sample frequency (0.75) is higher than the unconditional sample frequency (0.5)

Many econometric models can be viewed as generalisations of this very simple idea

Note than observing this association between wages and educational attainment does not tell us *why* the individuals with a college degree are more likely to have a high wage; this could be partly because getting a college degree has tended to make these individuals more productive, and partly because the individuals who get college degrees tend to be intrinsically more able and more productive

The **conditional sample frequencies** in our example can be found using

$$\widehat{f}(X_{1i} = x_1 | X_{2i} = x_2) = \widehat{f}(x_1 | X_{2i} = x_2) = \frac{\widehat{f}(X_{1i} = x_1, X_{2i} = x_2)}{\widehat{f}(X_{2i} = x_2)}$$

or

$$\widehat{f}(x_1 | x_2) = \frac{\widehat{f}(x_1, x_2)}{\widehat{f}(x_2)}$$

for x_1 and x_2 taking the values 0 or 1

Conditional probabilities are the population counterpart of such conditional frequencies

For 2 discrete random variables X_1 and X_2, the probability of X_1 taking the value x_1 (in the support of X_1) conditional on X_2 taking the value x_2 (in the support of X_2) is

$$\mathrm{P}(X_1 = x_1 | X_2 = x_2) = \mathrm{P}(x_1 | X_2 = x_2) = \frac{\mathrm{P}(X_1 = x_1, X_2 = x_2)}{\mathrm{P}(X_2 = x_2)}$$

or

$$\mathrm{P}(x_1 | x_2) = \frac{\mathrm{P}(x_1, x_2)}{\mathrm{P}(x_2)}$$

For both discrete and continuous random variables X_1 and X_2, the **conditional density function** is defined as

$$f(X_1 = x_1|X_2 = x_2) = f(x_1|X_2 = x_2) = \frac{f(X_1 = x_1, X_2 = x_2)}{f(X_2 = x_2)}$$

or

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

Note that for a continuous random variable X_2, the density $f(x_2)$ is well-defined and we can condition on the outcome $X_2 = x_2$, even though the probability of a continuous random variable taking any particular value is vanishingly small

The **conditional distribution function** can also be defined as

$$F(x_1|X_2 = x_2) = \mathrm{P}(X_1 \leqslant x_1|X_2 = x_2)$$

Conditional moments and **conditional quantiles** refer to the moments and quantiles of the conditional distribution function

For 2 random variables X_1 and X_2, the expected value of X_1 conditional on X_2 taking the value x_2 is the **conditional expectation**

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \int x_1 f(x_1|x_2) dx_1 \quad \text{for } X_1 \text{ continuous} \\ &= \sum x_1 f(x_1|x_2) \qquad\quad \text{for } X_1 \text{ discrete} \end{aligned}$$

where the integral or sum is taken over the support of X_1

For 2 random variables Y and X, the expression $\mathrm{E}(Y|X = x) = h(x)$ is a deterministic function of x; this **conditional expectation function** plays a central role in *regression analysis*

For 2 random variables X_1 and X_2, the variance of X_1 conditional on X_2 taking the value x_2 is the **conditional variance**

$$\begin{aligned}\mathrm{Var}(X_1|X_2 = x_2) &= \mathrm{E}\left\{[X_1 - \mathrm{E}(X_1|X_2 = x_2)]^2\,|X_2 = x_2\right\}\\ &= \int [x_1 - \mathrm{E}(X_1|X_2 = x_2)]^2\,f(x_1|x_2)dx_1 \text{for } X_1 \text{ continuous}\\ &= \sum [x_1 - \mathrm{E}(X_1|X_2 = x_2)]^2\,f(x_1|x_2) \quad \text{for } X_1 \text{ discrete}\end{aligned}$$

where the integral or sum is taken over the support of X_1

Law of Iterated Expectations

For 2 random variables Y and X, the unconditional expectation is related to the conditional expectation

$$\mathrm{E}_Y(Y) = \mathrm{E}_X\left[\mathrm{E}_{Y|X}(Y|X = x)\right]$$

or

$$\mathrm{E}(Y) = \mathrm{E}\left[\mathrm{E}(Y|X = x)\right]$$

Law of Total Variance

$$\mathrm{Var}(Y) = \mathrm{Var}\left[\mathrm{E}(Y|X = x)\right] + \mathrm{E}\left[\mathrm{Var}(Y|X = x)\right]$$

Independence and Conditional Independence

If 2 discrete random variables X_1 and X_2 are independent, then we have

$$\begin{aligned} P(X_1 = x_1 | X_2 = x_2) &= \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} \\ &= \frac{P(X_1 = x_1)P(X_2 = x_2)}{P(X_2 = x_2)} = P(X_1 = x_1) \end{aligned}$$

This formalises the idea that conditioning on the realised value of X_2 tells us nothing about the distribution of X_1

If 2 discrete or continuous random variables X_1 and X_2 are independent, we have

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)} = \frac{f(x_1)f(x_2)}{f(x_2)} = f(x_1)$$

It follows that $\mathrm{E}(X_1|X_2 = x_2) = \mathrm{E}(X_1)$

2 discrete random variables X_1 and X_2 are said to be independent conditional on the value taken by a third random variable X_3 if we have

$$\mathrm{P}(X_1 = x_1, X_2 = x_2|X_3 = x_3) = \mathrm{P}(X_1 = x_1|X_3 = x_3)\mathrm{P}(X_2 = x_2|X_3 = x_3)$$

For conditionally independent random variables we have

$$f(x_1, x_2|x_3) = f(x_1|x_3)f(x_2|x_3)$$

Bayes' Theorem

For 2 discrete random variables X_1 and X_2, we have

$$\begin{aligned} \mathrm{P}(X_1 = x_1, X_2 = x_2) &= \mathrm{P}(X_1 = x_1 | X_2 = x_2)\mathrm{P}(X_2 = x_2) \\ &= \mathrm{P}(X_2 = x_2 | X_1 = x_1)\mathrm{P}(X_1 = x_1) \end{aligned}$$

Rearranging gives

$$\mathrm{P}(X_1 = x_1 | X_2 = x_2) = \frac{\mathrm{P}(X_2 = x_2 | X_1 = x_1)\mathrm{P}(X_1 = x_1)}{\mathrm{P}(X_2 = x_2)}$$

This is a version of Bayes' theorem

$$\mathrm{P}(X_1 = x_1 | X_2 = x_2) = \frac{\mathrm{P}(X_2 = x_2 | X_1 = x_1)\mathrm{P}(X_1 = x_1)}{\mathrm{P}(X_2 = x_2)}$$

Suppose we start with a belief about $\mathrm{P}(X_1 = x_1)$

We then observe some data which we use to estimate $\mathrm{P}(X_2 = x_2)$ and $\mathrm{P}(X_2 = x_2 | X_1 = x_1)$

We can then use Bayes' theorem to obtain $\mathrm{P}(X_1 = x_1 | X_2 = x_2)$

This idea is used in models of learning in economic theory

Now suppose that the event $X_1 = x_1$ represents something we are interested in, while $X_2 = x_2$ represents evidence based on our sample

We then have

$$\mathrm{P}(\text{something}|\text{evidence}) = \frac{\mathrm{P}(\text{evidence}|\text{something})\mathrm{P}(\text{something})}{\mathrm{P}(\text{evidence})}$$

In this context, P(something) is the *prior probability* of the thing we are interested in (before we see the evidence), and P(something|evidence) is the *posterior probability* of the thing we are interested in (after we have seen the evidence)

Bayesian statistics/econometrics uses the sample data in this way to update some prior beliefs about things we may be interested in

In contrast, **frequentist** statistics/econometrics uses only the information contained in sample frequencies to learn about population probabilities; prior beliefs play no role

The approach to econometrics that we cover this year follows the frequentist approach

You may encounter Bayesian methods if you continue to study econometrics beyond this year's course

Covariance and Correlation

Covariance and correlation are measures of the strength of the linear association between 2 random variables X and Y

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

where the expected value E[.] is taken over the joint distribution of (X,Y)

More formally

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \int [x - \mathrm{E}(X)][y - \mathrm{E}(Y)] dF(x,y) \\ &= \int\int [x - \mathrm{E}(X)][y - \mathrm{E}(Y)] f(x,y) dx dy \text{ for } X,Y \text{ continuous} \\ &= \sum_y \sum_x [x - \mathrm{E}(X)][y - \mathrm{E}(Y)] f(x,y) \;\text{ for } X,Y \text{ discrete} \end{aligned}$$

The covariance can also be expressed as

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

From which we can see that

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

For linear transformations, we have

$$\mathrm{Cov}(a+bX, c+dY) = bd\mathrm{Cov}(X,Y)$$

for known constants a, b, c, d

If we have either $\mathrm{E}(X) = 0$ or $\mathrm{E}(Y) = 0$ or both

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

The sample covariance in a sample of n observations on (X_i, Y_i) is

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

Correlation is a scaled measure of covariance

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Note that

$$\text{Corr}(X, X) = \frac{\text{Var}(X)}{\sqrt{\text{Var}(X)\text{Var}(X)}} = \frac{\text{Var}(X)}{\text{Var}(X)} = 1$$

Since $\text{Cov}(X, -X) = -\text{Var}(X)$ and $\text{Var}(-X) = \text{Var}(X)$, we also have

$$\text{Corr}(X, -X) = \frac{-\text{Var}(X)}{\sqrt{\text{Var}(X)\text{Var}(X)}} = \frac{-\text{Var}(X)}{\text{Var}(X)} = -1$$

$\text{Corr}(X, Y)$ is bounded by -1 and 1

For linear transformations, we have

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{Corr}(X, Y)$$

for known constants a, b, c, d

If $\mathrm{Corr}(X, Y) = 0$, we say that the 2 random variables are **uncorrelated** or **orthogonal**

The notation $X \perp Y$ is used to indicate that X and Y are uncorrelated

Independence ($X \perp\!\!\!\perp Y$) implies zero correlation ($X \perp Y$)

If $X \perp\!\!\!\perp Y$, we have $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{Cov}(X, Y) = \mathrm{Corr}(X, Y) = 0$

In general the converse is not true: zero correlation does not imply independence

Counter-example: suppose that the discrete random variable X takes one of four possible values $\{-2, -1, 1, 2\}$, each with probability $\frac{1}{4}$

Then the discrete random variable $Y = X^2$ takes one of two possible values $\{1, 4\}$, each with probability $\frac{1}{2}$

And the discrete random variable $XY = X^3$ takes one of four possible values $\{-8, -1, 1, 8\}$, each with probability $\frac{1}{4}$

Since $\mathrm{E}(X) = 0$, we have $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) = 0$ and $X \perp Y$, even though the value taken by Y is entirely determined by the value taken by X

For example

$$\mathrm{P}(Y = 4|X = -2) = 1 \neq \mathrm{P}(Y = 4) = \frac{1}{2}$$

Y and X are not independent, but there is no *linear* association between the values taken by Y and the values taken by X

An *exception* to this general rule applies for normally distributed random variables

If we know that (X, Y) are jointly normally distributed and $\mathrm{Corr}(X, Y) = 0$, then it can be shown that X and Y are independent

The sample correlation is

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

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

The sample correlation is a measure of how close the observed values of (X_i, Y_i) lie to a straight line relationship of the form $Y_i = a + bX_i$, with positive correlation indicating an upward sloping relation $(b > 0)$ and negative correlation indicating a downward sloping relation $(b < 0)$

Expectation and Variance of a Random Vector

For the 2×1 random vector $X = (X_1, X_2)'$, we have

$$\mathrm{E}(X) = \mathrm{E}_X(X) = \mathrm{E}_X \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mathrm{E}_X(X_1) \\ \mathrm{E}_X(X_2) \end{pmatrix} = \begin{pmatrix} \mathrm{E}_{X_1}(X_1) \\ \mathrm{E}_{X_2}(X_2) \end{pmatrix} = \begin{pmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \end{pmatrix}$$

Note that

$$\mathrm{E}_X(X_1) = \int \int x_1 f_X(x_1, x_2) dx_1 dx_2 = \int x_1 f_{X_1}(x_1) dx_1 = E_{X_1}(X_1)$$

which follows from the relation $f_{X_1}(x_1) = \int f_X(x_1, x_2) dx_2$, i.e. marginalising the joint density wrt X_2

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left\{[X-\mathrm{E}(X)]\,[X-\mathrm{E}(X)]'\right\} \\
&= \mathrm{E}\left\{\begin{bmatrix} X_1-\mathrm{E}(X_1) \\ X_2-\mathrm{E}(X_2)\end{bmatrix}[X_1-\mathrm{E}(X_1), X_2-\mathrm{E}(X_2)]\right\} \\
&= \mathrm{E}\begin{pmatrix} [X_1-\mathrm{E}(X_1)]^2 & [X_1-\mathrm{E}(X_1)][X_2-\mathrm{E}(X_2)] \\ [X_1-\mathrm{E}(X_1)][X_2-\mathrm{E}(X_2)] & [X_2-\mathrm{E}(X_2)]^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{E}\left\{[X_1-\mathrm{E}(X_1)]^2\right\} & \mathrm{E}\left\{[X_1-\mathrm{E}(X_1)][X_2-\mathrm{E}(X_2)]\right\} \\ \mathrm{E}\left\{[X_1-\mathrm{E}(X_1)][X_2-\mathrm{E}(X_2)]\right\} & \mathrm{E}\left\{[X_2-\mathrm{E}(X_2)]^2\right\} \end{pmatrix}
\end{aligned}
$$

Thus we have

$$\operatorname{Var}(X) = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) \\ \operatorname{Cov}(X_1, X_2) & \operatorname{Var}(X_2) \end{pmatrix}$$

The variance matrix - also called the covariance matrix - is symmetric

These definitions extend straightforwardly to column vectors with more than 2 rows

Sums of Random Variables

Let X and Y be 2 random variables, and let a and b denote 2 known constants

Expected value

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

NB. This is the same as for the case where $X \perp\!\!\!\perp Y$

Variance

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

NB. This generalises the earlier result for $X \perp\!\!\!\perp Y \Rightarrow \mathrm{Cov}(X, Y) = 0$

Linear transformations of a random vector

Let X be a $p \times 1$ random vector, let A be a known $q \times 1$ vector, and let B be a known $q \times p$ matrix

Then $Y = A + BX$ is a $q \times 1$ random vector

Expected value

$$\mathrm{E}(Y) = \mathrm{E}(A + BX) = A + B\mathrm{E}(X)$$

Variance

$$\mathrm{Var}(Y) = \mathrm{Var}(A + BX) = B\mathrm{Var}(X)B'$$

NB. The preceding results for sums of random variables follow as special cases, for the 2×1 random vector $(X, Y)'$, setting B to be the known 1×2 row vector (a, b) and A to be the known scalar zero, so that

$$A + BX = 0 + \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = aX + bY$$

Now let $X = (X_1, X_2)'$ be a 2×1 random vector (i.e. X_1 and X_2 are random variables), and let $b = (b_1, b_2)'$ be a known 2×1 vector (i.e. b_1 and b_2 are known constants)

Then $b'X = b_1X_1 + b_2X_2$ is a scalar random variable, with $\text{Var}(b'X) \geqslant 0$

We also have $\text{Var}(b'X) = b'\text{Var}(X)b \geqslant 0$

These properties apply for any value of b

The $p \times p$ square matrix Σ is said to be *positive semi-definite* if the scalar $a'\Sigma a \geqslant 0$ for any $p \times 1$ vector a

The notation $\Sigma \geqslant 0$ denotes that Σ is a positive semi-definite matrix

Our example shows that the 2×2 matrix $\text{Var}(X)$ is positive semi-definite

This property holds more generally for the $p \times p$ variance matrix of a $p \times 1$ random vector, and generalises the property that $\text{Var}(X) \geqslant 0$ for a scalar random variable X (which is just the special case with $p = 1$)

Some properties of normally distributed random variables

Normality is preserved by linear transformation

If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y = a + bX$ for known constants a and b, then $Y \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$

Thus if $Z \sim \mathrm{N}(0, 1)$ and $Y = \mu + \sigma Z$ for known constants μ and σ, we have $Y \sim \mathrm{N}(\mu, \sigma^2)$

If $Z \sim \mathrm{N}(0, 1)$ and $Y = Z^2$, then $Y \sim \chi_1^2$ (**chi-squared** with one degree of freedom), with density function

$$f_Y(y) = \left(\frac{1}{\sqrt{2\pi}}\right) y^{\frac{-1}{2}} \exp(-y/2) \quad \text{for } y > 0$$

You may also see this written as $Y \sim \chi^2(1)$

If $Z_1, Z_2, ..., Z_n$ are independent and identically distributed standard normal random variables, so that $Z_i \sim \mathrm{N}(0,1)$ for $i = 1, 2, ..., n$, then $X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$ (**chi-squared** with n degrees of freedom), with density function

$$f_X(x) = \left(\frac{1}{c_n}\right) x^{\frac{n}{2}-1} \exp(-x/2) \quad \text{for } x > 0$$

for some normalisation constant c_n which ensures that $\int_0^\infty f_X(x)dx = 1$

You may see this written as $X \sim \chi^2(n)$

The standard normal distribution and the chi-squared family of distributions are used in *asymptotic inference* - hypothesis tests and confidence intervals based on 'large' sample or asymptotic approximations to the distributions of test statistics and estimators

Two related families of distributions that are used in *exact inference* - hypothesis tests and confidence intervals based on exact distributions of test statistics and estimators in finite samples - are the t distributions and the F distributions

t distribution (William Sealy Gosset, aka A.Student)

Suppose $Z \sim \mathrm{N}(0,1)$ and $W \sim \chi^2(n)$ are two independent random variables. Then

$$T = \frac{Z}{\sqrt{W/n}}$$

has a (Student) t distribution with n degrees of freedom ($T \sim t_n$)

F distribution (Ronald Fisher; George Snedecor)

Suppose $W_1 \sim \chi^2(m)$ and $W_2 \sim \chi^2(p)$ are two independent random variables. Then

$$W = \frac{(W_1/m)}{(W_2/p)}$$

has a (Snedecor) F distribution with (m,p) degrees of freedom ($W \sim F(m,p)$)

If $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ and X and Y are independent, then $X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

The class of normal distributions is said to be *closed under convolution*

If $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ and X and Y are independent, then $X+Y \sim \chi_{m+n}^2$

The class of chi-squared distributions is said to be *closed under convolution*

Estimators and Estimates

When we think of the sample mean for n outcomes of an underlying random variable, we refer to the sample mean as an *estimator*

For example, we might think of n individuals having heights $X_1, X_2, ..., X_n$, each being an outcome for an underlying random variable X, whose distribution describes the distribution of height in our population of interest (for example, male graduate students in Oxford)

We then think of the sample mean as an estimator of the expected value of the random variable X

For example, we might think that the distribution of height for this population is well described by a normal distribution with expected value μ and variance σ^2

In this case, we have a statistical model $X \sim \mathrm{N}(\mu, \sigma^2)$ characterised by the distributional assumption (normality) and 2 unknown parameters μ and σ^2

We then think of the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of the expected value μ

Note that no data on heights has been used in making these statements

An estimator of a parameter μ is often denoted by $\widehat{\mu}$

In contrast, if we observe heights for a sample of n male graduate students in Oxford and calculate the sample mean from this data, we refer to the calculated value of the sample mean as an *estimate*

For example, if I have data on 5 male graduate students with heights $\{1.60\text{m}, 1.65\text{m}, 1.70\text{m}, 1.75\text{m}, 1.80\text{m}\}$, the corresponding value of the sample mean $\overline{X} = 1.70\text{m}$ would be an estimate of the expected value of height in my population of interest

Asymptotic Theory

Asymptotic theory considers the behaviour of estimators in the limit, as the sample size becomes infinitely large

This approach can provide useful approximations to the behaviour of estimators in large finite samples, and in many cases of interest is the only known way of characterising the behaviour of estimators analytically

We start by considering the **limit of a sequence**

Example - Suppose that $X_n = c + \frac{1}{n}$ for $n = 1, 2, ...$ and some constant c

Clearly X_n tends towards the value c as we consider increasing values of n, since $\frac{1}{n}$ approaches zero as we consider increasing values of n

Equivalently the distance $|X_n - c|$ approaches zero as we consider increasing values of n

Formal definition of **convergence**:

For all $\delta > 0$, there exists some value n^* such that for all $n > n^*$ we have $|X_n - c| < \delta$

The notation $\lim_{n\to\infty} X_n = c$, or $X_n \to c$ as $n \to \infty$, is used to denote that X_n converges to c

Now suppose we have $X_n = c + \frac{Y}{n}$ for $n = 1, 2, ...$ and some constant c, where Y is a continuous random variable with $\mathrm{E}(Y) = 1$, finite variance, and Y may take values corresponding to any real number, so that the support of Y is not bounded (for example, we may have $Y \sim \mathrm{N}(1, \sigma^2)$ for $\sigma^2 < \infty$)

There is still a sense in which we expect X_n to approach c as we consider increasing values of n

But since Y is a continuous random variable with unbounded support, for any value of n, there is always a chance that we could obtain an outcome for Y so large that the condition $|X_n - c| < \delta$ would be violated

The standard definition of convergence does not apply

However, as n increases, the probability of Y taking a sufficiently large value to violate the condition $|X_n - c| < \delta$ gets smaller

This observation motivates the formal definition of **convergence in probability**:

For all $\varepsilon, \delta > 0$, there exists some value n^* such that for all $n > n^*$ we have $\mathrm{P}(|X_n - c| < \delta) > 1 - \varepsilon$

An equivalent definition is that $\mathrm{P}(|X_n - c| < \delta) \to 1$ as $n \to \infty$ for all $\delta > 0$, or that $\lim_{n\to\infty} \mathrm{P}(|X_n - c| < \delta) = 1$ for all $\delta > 0$

The notation $\operatorname{p\,lim}_{n\to\infty} X_n = c$, or $X_n \xrightarrow{\mathrm{P}} c$ as $n \to \infty$, is used to denote that X_n converges in probability to c

Convergence in probability is the main convergence concept for random variables that we use in this course

There are other concepts of convergence for random variables, including:

Almost sure convergence, denoted $X_n \overset{\text{a.s.}}{\to} c$:

For all $\varepsilon, \delta > 0$, there exists some value n^* such that we have

$\mathrm{P}(|X_n - c| < \delta, \text{ for all } n > n^*) > 1 - \varepsilon$

Convergence in mean square, denoted $X_n \overset{\text{m.s.}}{\to} c$:

$\mathrm{E}[(X_n - c)^2] \to 0$ as $n \to \infty$

These concepts are both stronger than convergence in probability, in the sense that $X_n \overset{\text{a.s.}}{\to} c$ implies that $X_n \overset{\text{P}}{\to} c$, and $X_n \overset{\text{m.s.}}{\to} c$ implies that $X_n \overset{\text{P}}{\to} c$

We now apply these ideas to study the probability limit ($\operatorname{p\,lim}$) of the sample mean for a sample of iid random variables

Suppose that $X_1, X_2, ..., X_n$ are independent random variables with the same expected value $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean for a sample of size n

We have previously seen that $\mathrm{E}(\overline{X}_n) = \mathrm{E}(X_i) = \mu$, and that $\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_i)}{n} = \frac{\sigma^2}{n}$

Then $\mathrm{E}[(\overline{X}_n - \mu)^2] = \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \rightarrow 0$ as $n \rightarrow \infty$

This establishes that $\overline{X}_n \overset{\text{m.s.}}{\rightarrow} \mu$, which implies that $\overline{X}_n \overset{\text{P}}{\rightarrow} \mu$

This result is a **Law of Large Numbers**

Although much harder to prove, a similar result can be shown without the assumption of finite variance

If $X_1, X_2, ..., X_n$ are independent and identically distributed random variables with the same expected value $\mathrm{E}(X_i) = \mu$, it can be shown that $\overline{X}_n \overset{\text{a.s.}}{\to} \mu$, which implies that $\overline{X}_n \overset{\text{P}}{\to} \mu$

Related notation: small o_P

Consider a sequence of random variables X_n and a sequence of constants a_n for $n = 1, 2, ...$

i) If $X_n \xrightarrow{\text{P}} 0$, we say $X_n = o_P(1)$

ii) If $(X_n/a_n) \xrightarrow{\text{P}} 0$, we say $(X_n/a_n) = o_P(1)$ or $X_n = o_P(a_n)$

Hence if $n^\alpha X_n \xrightarrow{\text{P}} 0$ for some α, we have $(X_n/n^{-\alpha}) \xrightarrow{\text{P}} 0$ and $(X_n/n^{-\alpha}) = o_P(1)$ or $X_n = o_P(n^{-\alpha})$

Example - For $X_1, X_2, ..., X_n$ iid(μ, σ^2), we have from LLN that $\overline{X}_n \xrightarrow{\text{P}} \mu$ so that $Y_n = \overline{X}_n - \mu \xrightarrow{\text{P}} 0$ and $Y_n = o_P(1)$

Viewing the sample mean $\overline{X} = \widehat{\mu}$ as an estimator of the expected value $\mathrm{E}(X_i) = \mu$, we say that $\widehat{\mu}$ is a **consistent** estimator of μ if $\widehat{\mu}$ has the property that $\widehat{\mu} \xrightarrow{\mathrm{P}} \mu$, or if $\mathrm{p}\lim \widehat{\mu} = \mu$

For observations which are independent and identically distributed, the Law of Large Numbers tells us that the sample mean is a consistent estimator of the expected value

Probability limits have a **useful property**:

If $X_n \xrightarrow{\text{P}} c$ and the function $g(a)$ is continuous at the point $a = c$, then $g(X_n) \xrightarrow{\text{P}} g(c)$, or $\operatorname{p\,lim} g(X_n) = g(\operatorname{p\,lim} X_n)$

This result may be referred to as **Slutsky's Theorem**, or as a version of the **Continuous Mapping Theorem**

Note that the function $g(\mu)$ may be a non-linear function

NB. For expected values, we only have the property $\text{E}(a+bX) = a+b\text{E}(X)$ for *linear* transformations of the random variable X

This property of probability limits allows us to establish consistency for estimators that can be expressed as continuous functions of sample means

Example - suppose that $X_1, X_2, ..., X_n$ are independent random variables with a Bernoulli(θ) distribution, with $\mathrm{E}(X_i) = \theta$ and $\mathrm{Var}(X_i) = \theta(1-\theta) < \infty$

The Law of Large Numbers shows that $\overline{X} \xrightarrow{\mathrm{P}} \theta$

We may also be interested in the log odds ratio $\ln[\theta/(1-\theta)]$, which is a continuous function of θ

Thus we have $\ln[\overline{X}/(1-\overline{X})] \xrightarrow{\mathrm{P}} \ln[\theta/(1-\theta)]$

Note that **unbiasedness** and **consistency** are two different properties of an estimator

Unbiasedness - $\mathrm{E}(\widehat{\mu}) = \mu$ - is a *finite sample property*, that holds for any sample size

Consistency - $\mathrm{p}\lim \widehat{\mu} = \mu$ - is an *asymptotic property*, that holds in the limit as the sample size $n \to \infty$

For iid observations, the sample mean is both an unbiased estimator and a consistent estimator of the expected value

In general, neither property implies the other

In the Bernoulli(θ) example, a single observation X_i can be shown to be an unbiased estimator of θ, but a single observation X_i is not a consistent estimator of θ

The non-linear function $\ln[\overline{X}/(1-\overline{X})]$ is a consistent estimator of the log odds ratio, but $\ln[\overline{X}/(1-\overline{X})]$ is not an unbiased estimator of the log odds ratio

Asymptotic Distribution Theory

Asymptotic distribution theory considers the distribution of suitably scaled or transformed estimators, in the limit as the sample size becomes infinitely large

These limit distribution results can be used to approximate the distribution of estimators in large finite samples, which allows confidence intervals to be constructed around the estimated values of objects of interest, and allows statistical tests of claims or hypotheses about objects of interest to be developed

We first require a definition of **convergence in distribution**

Let $X_1, X_2, ..., X_n$ denote a sequence of random variables - for example, X_n could be a statistic computed from a sample of size n

Let X denote a random variable

We say that the sequence X_n converges in distribution to X if

$$\mathrm{P}(X_n \leqslant x) \rightarrow \mathrm{P}(X \leqslant x) \quad \text{as } n \rightarrow \infty$$

for all values of x where $\mathrm{P}(X \leqslant x)$ is continuous

Equivalently the distribution function $F_{X_n}(x) \rightarrow F_X(x)$ as $n \rightarrow \infty$, or $\lim_{n\rightarrow\infty} F_{X_n}(x) = F_X(x)$, for all values of x where $F_X(x)$ is continuous

The notation $X_n \xrightarrow{\text{D}} X$ is used to denote that X_n converges in distribution to X

If X has a known distribution, say $X \sim \text{N}(\mu, \sigma^2)$, we may also write $X_n \xrightarrow{\text{D}} \text{N}(\mu, \sigma^2)$

Convergence in probability implies convergence in distribution, in the somewhat trivial sense that $X_n \xrightarrow{\text{P}} c$ implies that $F_{X_n}(x) \to F_c(x)$, although since c is a constant, the distribution $F_c(x)$ is degenerate

Formally we can still define $F_c(x) = \text{P}(c \leqslant x) = 0$ for $x < c$ and $F_c(x) = \text{P}(c \leqslant x) = 1$ for $x \geqslant c$

Convergence in distribution does not imply convergence in probability, since $X_n \xrightarrow{\text{D}} X$ indicates that in the limit as $n \to \infty$, X_n behaves as a random variable with the distribution function $F_X(x)$

Unless the distribution $F_X(x)$ happens to be degenerate, X_n will not converge in probability to some constant

Central limit theorems state that, under certain conditions, a suitably transformed version of the sample mean converges in distribution to a random variable with a normal distribution

The version of the central limit theorem that we use this term applies to the sample mean for independent and identically distributed random variables with a finite variance

This is called the Lindeberg-Levy central limit theorem, or CLT

Suppose that $X_1, X_2, ..., X_n$ are independent and identically distributed random variables, with expected value $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean for a sample of size n

Let $Z_n = (\overline{X}_n - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\overline{X}_n - \mu)/\sigma$

Then $Z_n \xrightarrow{\mathrm{D}} Z$, where $Z \sim \mathrm{N}(0, 1)$

The result may also be stated in terms $Y_n = \sqrt{n}(\overline{X}_n - \mu)$, in which case we have $Y_n \xrightarrow{\mathrm{D}} Y$, where $Y \sim \mathrm{N}(0, \sigma^2)$

In both cases, the scaling factor $\sqrt{n}$ is critical

We know that, under the stated conditions, $\overline{X}_n \xrightarrow{\text{P}} \mu$ so that $(\overline{X}_n - \mu) \xrightarrow{\text{P}} 0$

As the sample size $n \to \infty$, the distributions of both $\overline{X}_n$ and $(\overline{X}_n - \mu)$ become degenerate

This result is not very helpful for approximating the distribution of $\overline{X}_n$ in large finite samples

However, suitably scaled by $\sqrt{n}$, the CLT tells us that $Y_n = \sqrt{n}(\overline{X}_n - \mu)$ and $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ continue to behave as random variables in the theoretical limit as $n \to \infty$

Indeed, they behave as random variables with particularly convenient distributions

Note that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

and we have seen previously that for iid(μ, σ^2) random variables, we have

$\mathrm{E}(\overline{X}_n) = \mu$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$

Then

$$\mathrm{E}(Z_n) = \left(\frac{1}{\sigma/\sqrt{n}}\right) \mathrm{E}\left[\overline{X}_n - \mu\right] = \left(\frac{1}{\sigma/\sqrt{n}}\right)(\mu - \mu) = 0$$

And

$$\mathrm{Var}(Z_n) = \left(\frac{1}{\sigma^2/n}\right) \mathrm{Var}\left[\overline{X}_n - \mu\right] = \left(\frac{1}{\sigma^2/n}\right) \mathrm{Var}(\overline{X}_n) = \left(\frac{\sigma^2/n}{\sigma^2/n}\right) = 1$$

Showing that Z_n converges to a normal distribution is more challenging

The proof uses the **characteristic function** of a random variable

$$\varphi_X(\lambda) = \mathrm{E}(e^{i\lambda X}) = \int_{-\infty}^{\infty} e^{i\lambda x} dF_X(x)$$

which completely characterises its probability distribution (i and $-i$ are the complex square roots of -1; λ is any real number)

Proof of the (Lindeberg-Levy) CLT uses a result that if $\mathrm{E}(e^{i\lambda W_n}) \rightarrow \mathrm{E}(e^{i\lambda W})$ and $\mathrm{E}(e^{i\lambda W})$ is continous at $\lambda = 0$, then $W_n \xrightarrow{\mathrm{D}} W$

For the CLT, $W_n = Z_n$ and $W = Z \sim \mathrm{N}(0, 1)$, so the idea is to show that the characteristic function of Z_n converges to that of a standard normal random variable - see Hoel, Port and Stone (1971) chapter 8 for a sketch

Related notation: big O_P

Consider a sequence of random variables X_n and a sequence of constants a_n for $n = 1, 2, ...,$ and let X be some random variable

i) If $X_n \xrightarrow{\text{D}} X$, we say that $X_n = O_P(1)$

ii) If $(X_n/a_n) \xrightarrow{\text{D}} X$, we say that $(X_n/a_n) = O_P(1)$ or $X_n = O_P(a_n)$

Hence if $n^\alpha X_n \xrightarrow{\text{D}} X$ for some α, we have $(X_n/n^{-\alpha}) = O_P(1)$ or $X_n = O_P(n^{-\alpha})$

Example - For $X_1, X_2, ..., X_n$ iid(μ, σ^2), we have from CLT that $Z_n \xrightarrow{\text{D}} \text{N}(0, 1)$, so that $Z_n = O_P(1)$. If instead we consider $W_n = \overline{X}_n - \mu$, then we have $\sqrt{n}W_n = Y_n \xrightarrow{\text{D}} \text{N}(0, \sigma^2)$, so that $W_n = O_P(n^{-0.5})$

We can use the limit distribution result for $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ to approximate the distribution of $\overline{X}_n$ in large finite samples as follows

In large finite samples, we expect the distribution of $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ to be well approximated by a standard normal distribution

We denote this by $Z_n \overset{\mathrm{a}}{\sim} \mathrm{N}(0, 1)$

Now multiplying Z_n by $\sigma/\sqrt{n}$ only changes the variance, so we expect the distribution of $(\overline{X}_n - \mu)$ to be well approximated by a $\mathrm{N}(0, \sigma^2/n)$ distribution

And adding μ only changes the mean, so we expect the distribution of the sample mean $\overline{X}_n$ to be well approximated by a $\mathrm{N}(\mu, \sigma^2/n)$ distribution

We denote this by $\overline{X}_n \overset{\mathrm{a}}{\sim} \mathrm{N}(\mu, \sigma^2/n)$

NB. The argument on the previous slide uses the property of a normally distributed random variable that if $Z \sim \mathrm{N}(0,1)$ and $Y = \mu + \sigma Z$ for known constants μ and σ, then $Y \sim \mathrm{N}(\mu, \sigma^2)$

If $X_n \xrightarrow{\mathrm{D}} X$ and the function g is continuous, then $g(X_n) \xrightarrow{\mathrm{D}} g(X)$

This result is also referred to as a version of the **Continuous Mapping Theorem**

If $X_n \xrightarrow{\mathrm{D}} X$ and $Y_n \xrightarrow{\mathrm{P}} c$, and the function $g(a,b)$ is continuous at the point (a,c) for all values of a, then $g(X_n, Y_n) \xrightarrow{\mathrm{D}} g(X,c)$

This result is referred to as **(Generalised) Slutsky's Theorem**

Implications:

Letting $g(a, b) = a + b$ gives $X_n + Y_n \xrightarrow{\text{D}} X + c$

Letting $g(a, b) = ab$ gives $X_n Y_n \xrightarrow{\text{D}} Xc$

Letting $g(a, b) = a/b$ for $b \neq 0$ gives $X_n/Y_n \xrightarrow{\text{D}} X/c$ provided $c \neq 0$

These results allow us to approximate the distribution of functions of statistics

Application:

The central limit theorem

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{\mathrm{D}} Z \sim \mathrm{N}(0,1)$$

does not immediately allow us to construct a test statistic that can be used to test hypotheses about the true value of $\mathrm{E}(X_i) = \mu$, since $\mathrm{Var}(X_i) = \sigma^2$ is also unknown

Instead we construct a test statistic

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}_n}$$

using an estimator $\widehat{\sigma}_n$ of σ that we can calculate from a sample of size n

Write

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}_n} = \left(\frac{\sigma}{\widehat{\sigma}_n}\right)\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}\right) = \left(\frac{\sigma}{\widehat{\sigma}_n}\right) Z_n$$

If we use a consistent estimator $\widehat{\sigma}_n$ with the property

$$\widehat{\sigma}_n \xrightarrow{\mathrm{P}} \sigma \quad \text{or} \quad \left(\frac{\sigma}{\widehat{\sigma}_n}\right) \xrightarrow{\mathrm{P}} 1$$

then T_n is the product of $Z_n \xrightarrow{\mathrm{D}} Z \sim \mathrm{N}(0,1)$ and $\left(\frac{\sigma}{\widehat{\sigma}_n}\right) \xrightarrow{\mathrm{P}} 1$

Applying (Generalised) Slutsky's Theorem we also have

$$T_n \xrightarrow{\mathrm{D}} Z \sim \mathrm{N}(0,1)$$

since $Z \times 1 = Z$

As before, this limit distribution result allows us to approximate the distribution of $\overline{X}_n$ in large finite samples, using the asymptotic approximation

$$T_n = \left(\frac{\overline{X}_n - \mu}{(\widehat{\sigma}_n/\sqrt{n})} \right) \overset{\mathrm{a}}{\sim} \mathrm{N}(0,1)$$

or equivalently (by properties of linear transformations)

$$\overline{X}_n \overset{\mathrm{a}}{\sim} \mathrm{N}(\mu, \widehat{\sigma}^2/n)$$

Notice that replacing the unknown variance parameter σ^2 by a consistent estimator $\widehat{\sigma}^2$ does not change the form of the asymptotic approximation to the distribution of $\overline{X}_n$ (i.e. we still have the result that the distribution of $\overline{X}_n$ is approximately normal)

Both the sample variance estimator

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

and the ‘maximum likelihood’ estimator

$$\widehat{\sigma}_{\mathrm{ML},n}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

can be shown to be consistent estimators of $\sigma^2 = \mathrm{Var}(X_i)$ for independent and identically distributed random variables $X_1, X_2, ..., X_n$

Either estimator could thus be used here

To test a hypothesis about the true value of $\mu = \mathrm{E}(X_i)$ in this setting - for example, the claim that $\mu = 3$, or more generally that $\mu = \mu_0$ - we calculate the estimates $\overline{X}_n$ and (say) s_n^2 from our sample of n observations, and use these to construct the value that the test statistic T_n would take if the hypothesis is true

$$T_n(\mu_0) = \frac{\overline{X}_n - \mu_0}{(s_n/\sqrt{n})}$$

If the hypothesis is true, the distribution of the test statistic $T_n(\mu_0)$ in large finite samples should be approximately standard normal

We can ask if the value of test statistic we calculate is a likely or an unlikely value to obtain for a random variable with a standard normal distribution

For example, if we find $T_n(\mu_0) < -1.96$ or $T_n(\mu_0) > 1.96$, or equivalently if we find $|T_n(\mu_0)| > 1.96$, we might consider these values sufficiently unlikely to cast reasonable doubt on the validity of the claim that $\mu = \mu_0$

Related to this, we could find the range of values for μ_0 that would not lead us to reject the validity of the claim that $\mu = \mu_0$ on this basis

In our setting, if we use the critical values ±1.96 at the 5% level of significance to decide on the outcome of our tests, this would give the interval

$$\overline{X}_n - 1.96\left(\frac{s_n}{\sqrt{n}}\right),\ \overline{X}_n + 1.96\left(\frac{s_n}{\sqrt{n}}\right) \quad \text{or} \quad \overline{X}_n \pm 1.96\left(\frac{s_n}{\sqrt{n}}\right)$$

This is called an approximate (or an asymptotic) 95% confidence interval (or confidence band) for μ

Similar ideas will be used to test hypotheses about, or to construct confidence intervals for, the values taken by individual parameters in econometric models

Note that if we have independent and identically distributed random variables $X_1, X_2, ..., X_n$ with each $X_i \sim \mathrm{N}(\mu, \sigma^2)$ - that is, for iid normal random variables - we can derive the exact distribution of the sum

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

and hence the exact distribution of the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu, \sigma^2/n)$$

Obtaining the exact distribution of the test statistic $T_n = (\overline{X}_n - \mu)/(s_n/\sqrt{n})$ is less straightforward, since replacing the unknown variance parameter σ^2 by the estimator s_n^2 results in an exact distribution of the test statistic which is not normal

Here the exact distribution of the test statistic T_n can be shown to be a t-distribution with $n-1$ degrees of freedom

For iid normal random variables, this result allows hypothesis tests (and confidence intervals) to be based on the exact distribution of the test statistic, so that the resulting statistical inference is valid in *any* finite sample size

In many econometric models of interest, this luxury will not be available

Then we rely instead on approximations to the distribution of test statistics in *large* finite samples, based on the limit distribution results found in asymptotic distribtution theory

One approach that can be used to investigate whether these asymptotic approximations are reliable is to *simulate* the distribution of estimators or test statistics on a computer

Simulation:

- generate a sample of data from a data generation process which mimics some features of the problem we are interested in and using a realistic sample size

- calculate the estimator or test statistic of interest using the generated data

- repeat this process many times

- compare the distribution of the estimator or test statistic in these simulations to the distribution suggested by the asymptotic distribution theory

This approach is called Monte Carlo simulation

Another application of results we have now covered in asymptotic theory is to establish the consistency of the empirical distribution function as an estimator of the distribution function for independent and identically distributed random variables $X_1, X_2, ..., X_n$

Recall that the empirical distribution function can be written as a sample mean, using the indicator function

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x)$$

For iid random variables, the LLN tells us that the sample mean converges in probability to the expected value, so

$$\widehat{F}_n(x) \xrightarrow{\mathrm{P}} \mathrm{E}[I(X_i \leqslant x)]$$

The expected value of the indicator function is the probability that the condition is satisfied, so we have

$$\widehat{F}_n(x) \xrightarrow{\mathrm{P}} \mathrm{P}(X_i \leqslant x) = F_X(x)$$

For each value of x, the empirical distribution function for a sample of size n converges in probability to the distribution function of the underlying random variable at that value of x

Similarly we have a CLT result

$$Z_n = \frac{\sqrt{n}\left[\widehat{F}_n(x) - F_X(x)\right]}{\sqrt{F_X(x)(1 - F_X(x))}} \xrightarrow{\mathrm{D}} Z \sim \mathrm{N}(0,1)$$

which uses $\mathrm{Var}[I(X_i \leqslant x)] = F_X(x)(1 - F_X(x))$

And a (Generalised) Slutsky's Theorem result

$$Z_n = \frac{\sqrt{n}\left[\widehat{F}_n(x) - F_X(x)\right]}{\sqrt{\widehat{F}_n(x)(1 - \widehat{F}_n(x))}} \xrightarrow{\mathrm{D}} Z \sim \mathrm{N}(0,1)$$

For large finite samples, this can be used to construct an approximate 95% confidence interval for $F_X(x)$ at each value of x, or to test a hypothesis about $\mathrm{P}(X_i \leqslant x)$

For example, with data on weekly wages, we may be interested in testing a claim about the proportion of our population who earn less than £400 per week

NB. The empirical distribution function is an example of a *non-parametric* estimator; we have not specified the form of the distribution function for the underlying random variable X in making any of these statements

Stronger results are available for the empirical distribtution function than the *pointwise* consistency and distribution results we have discussed here