Lecture 10. Essential statistical concepts (Part 1)

10.1. Introduction

This lecture is intended to cover all the basic statistical concepts you will need for the remainder of the QMII course. There are hundreds of basic statistics books you could use to supplement these notes if you wanted to, but you will not need to go into any greater detail for the purposes of this course. The following topics are covered:

Part 1. Section 10.2. Random variables and probabilities.
Section 10.3. The expected value of a random variable.
Section 10.4. The variance of a random variable.
Part 2. Section 10.5. The normal distribution.
Section 10.6. The covariance between two random variables.
Section 10.7. The correlation between two random variables.
Section 10.8. Estimation.
Section 10.9. Unbiasedness.
Section 10.10. Efficiency.
Section 10.11. Consistency.

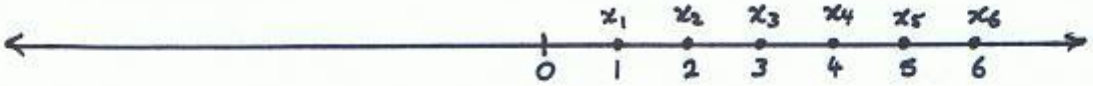
We will do Part 1 in the seminar, and Part 2 in the main lecture on Tuesday.

10.2. Random variables and probabilities

Recall that a 'variable' is something whose magnitude can change i.e. something that can take on different values. For example, the *calendar year* is a variable quantity, because it changes every time the earth completes an orbit around the sun. It is called a *deterministic variable* because the values it can take on are predetermined e.g. after the earth completes its current orbit it will be 2000, after the next one it will be 2001, etc. By contrast, ***a random variable is a variable whose values cannot be predicted with certainty before they actually occur.*** For example, the *highest temperature in Britain tomorrow* is a random variable, because we cannot know for certain what it is going to be before it happens! ***All we know about a random variable are the types of values that can occur, and the probabilities which particular values or sets of values have of occurring.*** Over time, particular values or sets of values of a random variable will be seen to occur more or less often depending on how probable they are compared to others.The more probable they are, the more often they will occur; the less probable they are, the less often they will occur. For example, it is not *impossible* that tomorrow the highest temperature in Britain will lie between -60 degrees Celsius and -50 degrees Celsius, but at this time of year it is more probable that the highest temperature will be somewhere between 0 degrees Celsius and 15 degrees Celsius. Since this is the most probable range of temperatures for this time of year, this is the set of values that will be seen to have occurred most often at this time over the last 500 years, say.

For our course, we must distinguish between two types of random variable: discrete and continuous.

A random variable is discrete if all its possible values can be identified as distinct points on the real line E^1. This means that there are some numbers on the real line that cannot be values of the random variable (e.g. the numbers between any two neighbouring, but distinct, points on the real line). The classic example of a discrete random variable is the score when a dice is thrown. We can use X to denote the random variable i.e. X = the score when a dice is thrown, and we can use $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$, $x_5=5$, $x_6=6$ to denote the possible values of X. Each of the six values can be identified as a distinct point on the real line:



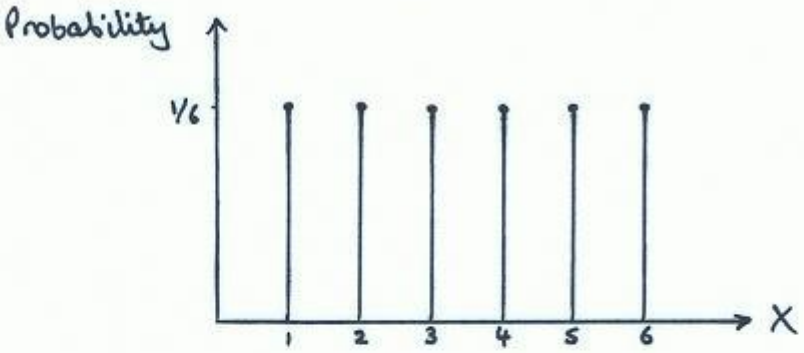
In this case, X is clearly a random variable, because we cannot know for certain what its value is going to be before we throw the dice. All we know is that the value will be an element of the set {1, 2, 3, 4, 5, 6}, and that each of these values has a 'one-in-six chance' of occurring. This is another way of saying that each of the values of X has a probability equal to 1/6 of occurring, which we write as $Prob(X=x_1)=1/6$, $Prob(X=x_2)=1/6$, $Prob(X=x_3)=1/6$, $Prob(X=x_4)=1/6$, $Prob(X=x_5)=1/6$, $Prob(X=x_6) = 1/6$. As you should already know from your previous studies, the sum of the probabilities of all the values of a random variable must equal 1.

The probability distribution of a random variable is a representation of the probabilities for all the possible values. This representation can be *algebraic* (i.e. by means of a formula), or *graphic* (i.e. by means of a diagram), or *tabular* (i.e. in a table). For example, here is the probability distribution of the random variable X in the dice example, in tabular form:

Probability distribution of the random variable X=the score when a dice is thrown

Values of the random variable X	Probabilities
$x_1 = 1$	$Prob(X=x_1) = 1/6$
$x_2 = 2$	$Prob(X=x_2) = 1/6$
$x_3 = 3$	$Prob(X=x_3) = 1/6$
$x_4 = 4$	$Prob(X=x_4) = 1/6$
$x_5 = 5$	$Prob(X=x_5) = 1/6$
$x_6 = 6$	$Prob(X=x_6) = 1/6$

Here is the same distribution in graphical form:



Finally, the distribution can be represented by means of a formula known as the *probability function*, which in this case is written as $p_X(x) = 1/6$, for x=1, 2, 3, 4, 5, 6. ***The probability function is simply a rule that takes you from each possible value of a random variable, to the corresponding probability of that value.***

In the dice example, the random variable X is not only discrete, but also finite, because we can list out all of its values i.e. $\{x_1=1, x_2=2, x_3=3, x_4=4, x_5=5, x_6=6\}$. An example of an infinite discrete random variable is the number of throws of a coin needed before a head first appears. The possible outcomes are 1, 2, 3, ... These values can be identified as distinct points on the real line, and a

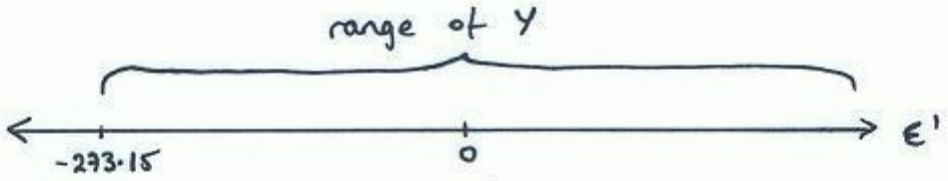
probability can be assigned to each, but you could never list them *all* out, because the list goes on for ever. Thus, the random variable is discrete, but also infinite. Here are some other examples of discrete random variables:

1. The number of fatal car accidents in a city in a given month.
2. The number of customers arriving at a check-out counter in an hour.
3. The number of errors detected in a company's accounts.
4. The number of claims on a medical insurance policy in a particular year.

A random variable is <u>continuous</u> if its possible values cannot be identified as <u>distinct</u> points on the real line E^1. In other words, a random variable is <u>continuous</u> if it can take <u>any</u> value in some subset of the real line. In other words again, there is no number in that subset of the real line that cannot be a value of the random variable. For example, the *highest temperature in Britain tomorrow* is a continuous random variable, because it can take any value along a continuum. Another example is the *height of the tallest person in the room.* ***A crucial difference between continuous and discrete random variables is that, for continuous random variables, we cannot attach probabilities to specific values.*** We cannot do this because of the way we have defined continuous random variables. In order to be able to assign probabilities to specific values of a continuous random variable, we would have to be able to identify the values as distinct points on the real line, which is impossible because of the way we have defined continuous random variables above! ***All we can do with continuous random variables is assign probabilities to <u>ranges</u> of values.*** For example, let us denote the *highest temperature in Britain tomorrow* (a random variable) by the letter Y. Thus

Y = the highest temperature in Britain tomorrow

Now, Y can take on any value along the real line between -273.15 degrees Celsius (absoulte zero) and $+\infty$:



We say that Prob(Y = 10 degrees Celsius) = 0, for example, because 10 is a specific point on the real line. The same applies to any *specific* value of Y in the range of possible values along the real line. All we can do is make statements like Prob(0 degrees Celsius < Y < 15 degrees Celsius) = 0.8, which says *'the probability that Y will be somewhere between 0 degrees Celsius and 15 degrees Celsius is 0.8'*. Here are some other examples of continuous random variables:

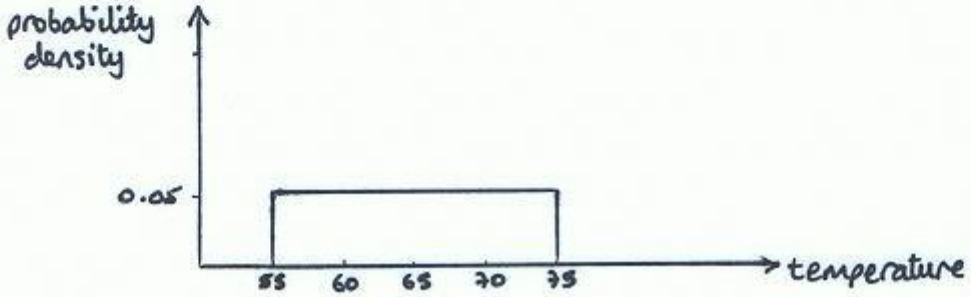
1. The percentage of impurity in a batch of chemicals.
2. The time that elapses between a person's birth and that person's death.
3. Annual rainfall in a particular city.
4. Your weight at 12.01am on 1st January each year.
5. The change in the price of an ounce of gold in a month.

The distinction we have made between discrete and continuous random variables may appear a bit artificial. After all, we very rarely measure things on a continuum. For example, the highest temperature in Britain tomorrow cannot be reported more precisely than the measurement instrument allows. However, when measurements can be made on such a fine scale that differences between adjacent values are of no significance, it is convenient to act as if they had truly been

made on a continuum. For example, the difference between a highest temperature in Britain tomorrow of 10.3425 and 10.3426 is of very little significance, and the attachment of probabilities to each would be rather pointless.

For practical purposes, then, we treat as <u>discrete</u> all random variables for which probability statements about the individual possible outcomes have worthwhile meaning; all other random variables are treated as <u>continuous</u>.

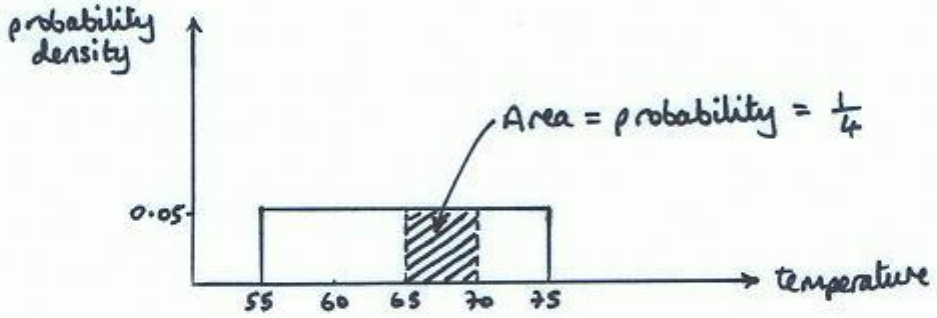
It would obviously be impossible to display the probability distribution of a continuous random variable in tabular form, because there are an infinite number of possible values, and each specific value has zero probability. Instead, what we do is represent the probability distribution graphically and algebraically. Let us illustrate this with the example of the highest temperature in Britain tomorrow, this time measured in degrees Fahrenheit instead of Celsius (just to be awkward!). For the sake of argument, we will assume that this varies within the limits of 55 and 75^0F, and initially we will suppose that it is equally likely to be anywhere within this range. Since all temperatures from 55 to 75^0 are equally likely, the probability is represented as being spread out uniformly, as in the diagram below:



We must always assume that the probability is spread to cover an area equal to 1. This is because the <u>total</u> probability associated with the values of a random variable is always equal to 1. In this case, the probability has been spread to form a rectangle, and since the length of the rectangle is equal to 20, its height is given by the formula for the area of a rectangle

$$20 \times height = 1 \Rightarrow height = 1/20 = 0.05$$

(recall that length×height = area). Having found the height, we can now answer such questions as: 'what is the probability that the highest temperature in Britain tomorrow lies somewhere between 65 and 70^0F ?'. ***The answer is given by the amount of probability lying between 65 and 75^0F, which is the shaded area in the diagram below:***



The base of the shaded area is 5, and its height is 0.05, so the area is 0.25. The probability is 1/4.

The height of the graph at any point represents what is formally called the probability density at that point, and if the height can be written as a function of the random variable, it is known as the probability density function. In this case, the probability density function is given by f(y), where y is the temperature, and

$$f(y) = 0.05 \quad \text{for} \quad 55 \leq y \leq 75$$

We say that $f(y) = 0$ for $y < 55$ or $y > 75$.

You should know from your previous studies that the area under the graph of a function between two points on the horizontal axis is given by the definite integral of the function, where the two limits of integration are the two points on the horizontal axis (actually, this is only true if the value of the function is never negative - which it never is for a probability density function). Thus, the probability above could have been calculated using the probability density function by integrating it as follows:

$$\text{Prob}(65 < y < 70) = \int_{65}^{70} f(y)dy = \int_{65}^{70} (0.05)dy = (0.05)\int_{65}^{70} dy = (0.05)(70 - 65) = 0.25$$

The probability distribution drawn above is the simplest continuous distribution, known for obvious reasons as the uniform distribution. The continuous distribution you will meet most often in econometrics is the normal distribution, to which we will turn after considering the concepts of expectation and variance.

10.3. The expected value of a random variable

The expected value of a random variable is just a weighted sum of all its possible values, where the weights are the probabilities of the values. It is best to consider the case of discrete random variables first, and then extend the discussion to continuous random variables.

Consider a discrete random variable X that can take on n particular values $x_1, x_2, \ldots, x_n$, with probabilities $p_1, p_2, \ldots, p_n$ respectively. Then the expected value of X, denoted by E[X], is given by

$$E[X] = p_1x_1 + p_2x_2 + \cdots + p_nx_n = \sum_{i=1}^{n} x_i p_i$$

In words, the expected value of a discrete random variable is the weighted sum of all its possible values, taking the probability of each value as its weight. For a concrete example, suppose X = the score when a dice is thrown. This random variable has six possible values: $\{x_1=1, x_2=2, x_3=3, x_4=4, x_5=5, x_6=6\}$. Each of these values has a probability of 1/6, so the expected value of X is defined as

$$E[X] = (1)(1/6) + (2)(1/6) + (3)(1/6) + (4)(1/6) + (5)(1/6) + (6)(1/6) = 3.5$$

Notice that the expected value of X in this particular example is a number you could not obtain at all, although this need not be the case with other random variables. The expected value of a random variable is frequently described as its population mean, denoted by the Greek letter μ.

Now consider the case of continuous random variables. Remember that we cannot attach probabilities to specific values of a continuous random variable. In order to define the expected value, therefore, we use the probability density function. Thus, suppose a random variable Y can take on any value between a and b, where a and b are points on the real line such that $a < b$. Further suppose that the probability density function of Y is f(y) for $a < y < b$, and zero otherwise. (Recall that a probability density function evaluated at any particular value y gives the height of

the graph which describes the distribution of probabilities for a continuous random variable). Then the expected value of Y is defined as

$$E[Y] = \int_a^b yf(y)dy$$

This is actually very similar to the definition of the expected value of a discrete random variable. In both cases, the different possible values of the random variable are weighted by some measure of the probability attached to them. In the case of a discrete random variable, the summation is done on a value by value basis, with each value weighted by its probability. In the continuous case, integration replaces summation, and the probability density function f(y) replaces the individual probabilities for each value in the discrete case. However, the principle is exactly the same. ***Note: you will never be asked to find the expected value of a continuous random variable by integration in this course. We will refer to the expected value of any given random variable Z, say, as E[Z], irrespective of whether it is continuous or discrete. Exactly the same rules and principles apply to each, so we do not have to distinguish between them.*** However, just to give you an idea of how you find the expected value of a continuous random variable, consider the simple example considered in the previous section, where Y = the highest temperature in Britain tomorrow, and the density function is $f(y) = 0.05$ for $55 \leq y \leq 75$, and $f(y) = 0$ otherwise. Then, using the elementary rules of integration (which you should know!) we get

$$E[Y] = \int_a^b yf(y)dy = \int_{55}^{75} y(0.05)dy = (0.05)\int_{55}^{75} ydy = (0.05)(75^2 - 55^2)/2 = 65$$

There are three rules for manipulating expected values of random variables which you must learn. They are easy to prove, but I will not do it here. They apply to any random variables, irrespective of whether they are continuous or discrete:

<u>Rule 1</u>. The expected value of a sum of random variables is equal to the sum of their expected values. For example, if you have three random variables X, Y, and Z, then

$$E[X+Y+Z] = E[X] + E[Y] + E[Z]$$

<u>Rule 2</u>. If you multiply a random variable by a constant, you multiply its expected value by the same constant. For example, if X is a random variable and a is a constant, then

$$E[aX] = aE[X]$$

<u>Rule 3</u>. The expected value of a constant is that constant. For example, if a is a constant, then

$$E[a] = a$$

Putting these three rules together, you can simplify more complicated expressions. For example, suppose you wish to work out E[Y], where Y = a+bX, and a and b are constants. Then using the three rules above, E[Y] = a + bE[X].

There are two more things to note. First, suppose that X is a random variable, and that g(X) is some function of X. Then the expected value of g(X) is just

$$E[g(X)] = g(x_1)p_1 + g(x_2)p_2 + \cdots = \Sigma g(x_i)p_i$$

if the random variable X is discrete, or

$$E[g(X)] = \int g(x)f(x)dx$$

if the random variable X is continuous and has a density function f(x). In words, to find the expected value of g(X), you just treat the values of g as if they were the values of the random variable itself.

Second, suppose you have two random variables X and Y. These are said to be independent if

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$$

for any functions f(X) and g(Y). As an important special case, independence implies that

$$E[XY] = E[X]E[Y]$$

In words, 'the expected value of the product equals the product of the expected values'.

10.4. The variance of a random variable

The variance of a random variable is a measure of the 'dispersion' of its probability distribution around the mean i.e. how 'spread out' the probability distribution is around the expected value of the random variable concerned. ***The variance of a random variable X, denoted by V[X], is defined as the expected value of the square of the difference between X and its mean i.e. the expected value of*** $(X-E[X])^2 = (X-\mu)^2$. ***It is also denoted by the Greek symbol*** σ_X^2.

So suppose that X is a discrete random variable with values $x_1, x_2, \ldots$, and probabilities $p_1, p_2, \ldots$, and with expected value $E[X] = \mu$. Then

$$V[X] \equiv \sigma_X^2 = E[(X-\mu)^2] = (x_1-\mu)^2p_1 + (x_2-\mu)^2p_2 + \cdots = \Sigma(x_i-\mu)^2p_i$$

Similarly, suppose that X is a continuous random variable with density function f(x) and mean μ. Then

$$V[X] \equiv \sigma_X^2 = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)dx$$

As a concrete example, consider the simple random variable X ≡ the score when a dice is thrown. We saw earlier that $E[X] = \mu = 3.5$. Then

$$V[X] \equiv \sigma_X^2 = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + (3-3.5)^2(1/6) + (4-3.5)^2(1/6) + (5-3.5)^2(1/6) + (6-3.5)^2(1/6) = 2.92$$

Another very important indicator of the dispersion of the probability distribution of a random variable is the standard deviation. ***The standard deviation of a random variable X is defined as the square root of its variance, denoted by*** σ_X. ***Thus*** $\sigma_X = \sqrt{V[X]} = \sqrt{\sigma_X^2}$. In the simple case where X ≡ the score when a dice is thrown, the standard deviation of the probability distribution of X is $\sqrt{2.92} = 1.71$.

Using the three rules for manipulating expected values given in the last section, it is easy to show that

$$V[X] \equiv \sigma_X^2 = E[(X-\mu)^2] = E[X^2] - \mu^2$$

I shall prove this here because it is useful. Remembering that $E[X] = \mu$, and that μ is a constant, we have

$$V[X] \equiv \sigma_X^2 = E[(X-\mu)^2] = E[X^2-2\mu X+\mu^2] = E[X^2]-2\mu E[X]+\mu^2 = E[X^2]-2\mu^2+\mu^2 = E[X^2]-\mu^2.$$

(End of Lecture 10 Part 1)

Lecture 10. Essential statistical concepts (Part 2) ①

10.5. The normal distribution

In Part 1 of Lecture 10, we briefly discussed random variables and probabilities, and the concepts of 'expected value' and 'variance'. We saw that there are various ways in which to represent the distribution of probability between the possible values of a random variable. The method that applies to all random variables (whether discrete or continuous) is the algebraic method, in which a formula is used to describe how the probability is 'spread out' over the real line E^1. In the case of discrete random variables, the formula is called the probability function or the probability mass function. In the continuous case, the formula is called the probability density function.

Here are some examples of probability mass functions for various kinds of discrete random variable (you do not have to learn these!):

(1). The discrete uniform distribution

This is the distribution of the simple random variable we looked at in the last lecture: the score when a dice is thrown. More generally, let X be a finite discrete random variable which can take any one of N values in the set {1, 2, . . ., N}. Then the probability mass function of X is

$$p_X(x) = 1/N \qquad \text{for } x = 1, 2, \ldots, N$$
$$p_X(x) = 0 \qquad \text{otherwise}$$

(2). The Bernoulli distribution

This applies to discrete random variables which can only take one of two values: either 0 or 1. For example, suppose X takes the value 1 when you toss a coin and get a head, and X takes the value 0 when you toss a coin and get a tail. Then X is a random variable with a Bernoulli distribution. More generally, let X be a finite discrete random variable which takes the value 1 with probability p, and the value 0 with probability (1-p). Then the probability mass function of X is

$$p_X(x) = p^x(1-p)^{1-x} \qquad \text{for } x = 0, 1$$
$$p_X(x) = 0 \qquad \text{otherwise}$$

(3). The geometric distribution

This is the distribution of the second example of a discrete random variable we looked at in the last lecture i.e. the number of coin tosses before a head is obtained. More generally, suppose X counts the number of 'trials' of an experiment (e.g. coin tosses) before a 'success' occurs (e.g. getting a head), where the probability of a success is p, and the probability of a failure is (1-p). Then the probability mass function of X is

$$p_X(x) = p(1-p)^x \qquad \text{for } x = 0, 1, 2, 3, \ldots$$
$$p_X(x) = 0 \qquad \text{otherwise}$$

(4). The binomial distribution

This is the distribution of the random variable which counts the number of 'successes' (e.g. heads) in a specified number of 'trials' (e.g. 3 coin tosses). So suppose we are considering a sequence of N trials, each of which has two possible outcomes: 'success' (probability = p) or 'failure' (probability = 1-p). Let X be the random variable which counts the number of successes in the N trials. Then the probability mass function of X is

$$p_X(x) = C_x^N\, p^x(1-p)^{N-x} \qquad \text{for } x = 0, 1, 2, \ldots, N$$
$$p_X(x) = 0 \qquad \text{otherwise}$$

where $C_x^N = N!/\{x!(N-x)!\}$, and where ! is the 'factorial' symbol eg. $N! = N(N-1)(N-2)\cdots 1$.

Remember: The above are just 'algebraic' ways of describing how the probability is 'spread out' between all the possible values of the random variable concerned. In a similar way, we can write down many examples of probability density functions for various kinds of continuous random variable. Here are a few:

(1). The continuous uniform distribution

This is the distribution we met in the last lecture (sometimes called the 'rectangular' distribution), when we assumed that the highest temperature in Britain tomorrow was equally likely to be anywhere between 55^0F and 75^0F. More generally, suppose that a random variable Y takes values within the limits of a and b on the real line i.e. $a < Y < b$, and suppose that the probability mass is spread uniformly over this interval. Then the density function of Y is

$$f_Y(y) = 1/(b-a) \qquad \text{for } a < y < b$$
$$f_Y(y) = 0 \qquad \text{otherwise}$$

(2). The exponential distribution with parameter λ

This distribution is typically associated with random variables which measure the 'length of life' or 'time to failure' of various devices and systems (including us!). You will come across it if you do anything related to 'survival analysis' in your dissertation next year. Suppose Y is a nonnegative random variable i.e. $Y \geq 0$ which measures the length of time until a new device 'fails'. Y has the exponential distribution with parameter λ if its density function is

$$f_Y(y) = \lambda e^{-\lambda y} \qquad \text{for } y \geq 0$$
$$f_Y(y) = 0 \qquad \text{otherwise}$$

where e is the physical constant e = 2.71828....

(3). The Weibull distribution with parameters α, λ, and υ

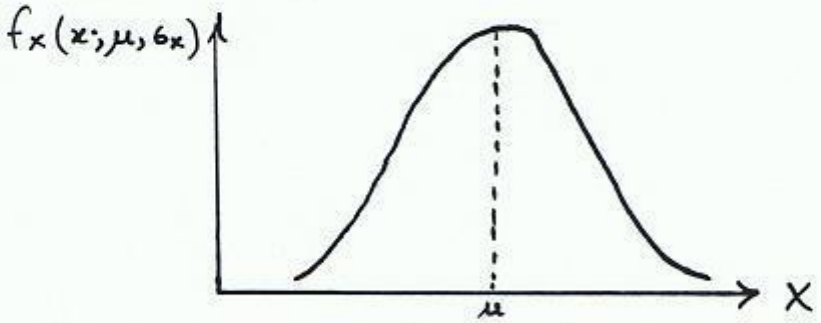
This is a generalisation of the exponential distribution. A nonnegative random variable Y has the Weibull distribution with parameters α, λ, and υ if its density function is of the form

$$f_Y(y) = \alpha\lambda(y-\upsilon)^{\alpha-1}e^{-\lambda(y-\upsilon)^{\alpha}} \qquad \text{for } y - \upsilon \geq 0$$
$$f_Y(y) = 0 \qquad \text{otherwise}$$

In the rest of this section, we will focus on one continuous distribution you will keep coming across again and again in econometrics, namely the normal distribution (so-called because many random variables in 'nature' happen to be distributed approximately like this, so it is the 'normal' thing to find!). Let X be a normally distributed (continuous) random variable, with mean $E[X] \equiv \mu$, and variance $V[X] \equiv \sigma_X^2$, where $-\infty < \mu < \infty$ and $0 < \sigma_X^2 < \infty$. Then the probability density function of X is of the following form:

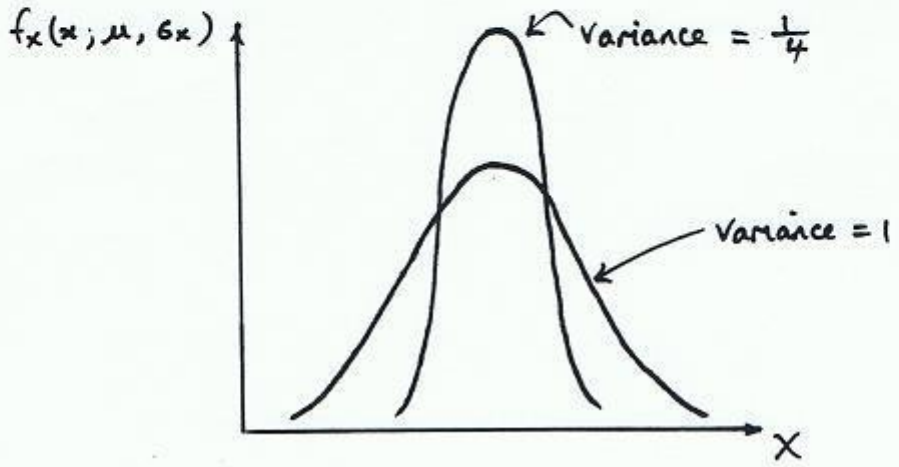
$$f_X(x;\, \mu,\, \sigma_X) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma_X^2}\right\}$$

where x can be any number on the real line, and where π is the physical constant $\pi = 3.14159....$ Remember, all this is doing is telling you how the probability is distributed between the possible values of X. If you plot the graph of $f_X(x;\, \mu,\, \sigma_X)$, it is bell-shaped:



Probability density function for the normal distribution

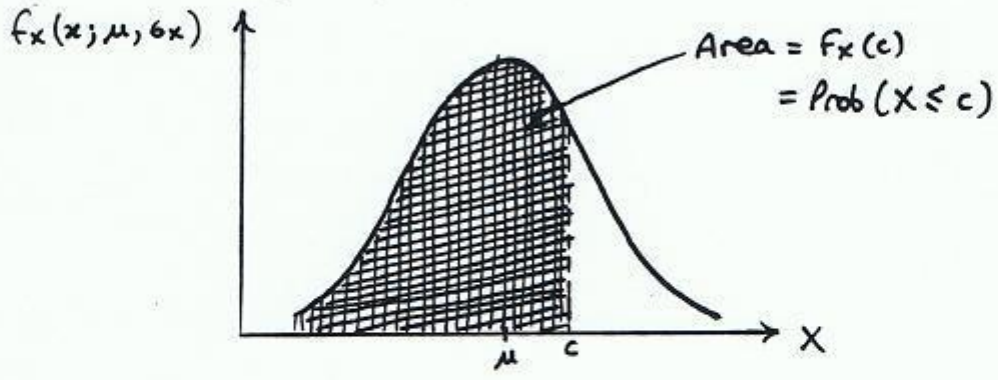
So the probability density function $f_X(x; \mu, \sigma_X)$ is telling you that the probability is distributed in a symmetric bell shape over all the possible values of X. The tallest part of the bell (i.e. the greatest probability density) is directly over the mean μ of the random variable. The variance σ_X^2 tells you how 'spread out' the bell is around this mean. The bigger the variance, the lower is the density at the mean, and the more spread out is the bell shape. ***Remember: as in the simple example of the previous lecture, the area under the bell-shaped graph must always equal 1, because the total probability associated with the values of a random variable must be exactly 1.***



When a random variable is normally distributed with a mean μ and variance σ_X^2, it is customary to write this as $X \sim N(\mu, \sigma_X^2)$. We shall now consider how to find the probability that the value of X lies between two numbers a and b on the real line i.e. $\text{Prob}(a < X < b)$. We begin by introducing the concept of the cumulative distribution function. If $X \sim N(\mu, \sigma_X^2)$, then the cumulative distribution function of X, denoted by $F_X(c)$, is

$$F_X(c) = \text{Prob}(X \leq c) = \int_{-\infty}^{c} f_X(x; \mu, \sigma_X)\, dx$$

This is the area under the probability density function to the left of c, as shown in the diagram:

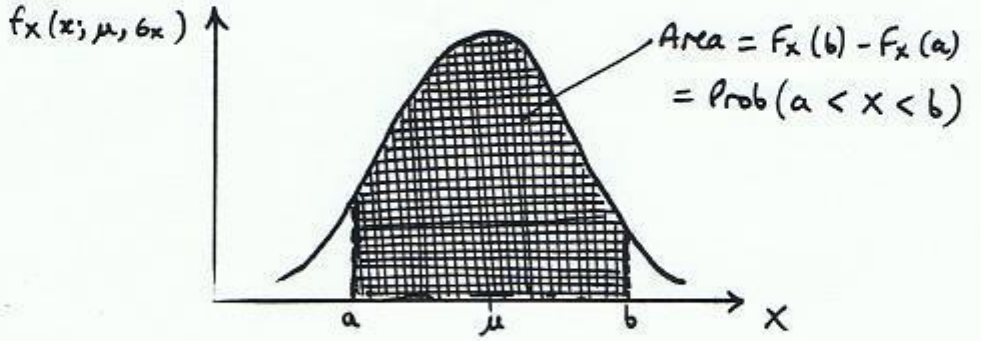


This area is the total probability 'mass' under the curve to the left of point c on the horizontal axis.

Notice that since the total area under the curve is 1, we must have $F_X(\infty) = 1$. After a moment's thought, you should be able to see that the probability $\text{Prob}(a < X < b)$ can be expressed in terms of the cumulative distribution function as follows:

$$\text{Prob}(a < X < b) = F_X(b) - F_X(a)$$

This is the area under the graph of the density function $f_X(x; \mu, \sigma_X)$ between the points a and b on the horizontal axis:



In theory, any required probability can be obtained from the cumulative distribution function in this way. However, problems arise in practice because the integral in the formula for $F_X(c)$ above is tricky to compute. Furthermore, there are many random variables which are normally distributed, but they do not all have the same mean and variance. Since the density function $f_X(x; \mu, \sigma_X)$ depends on these two parameters, each normally distributed variable has a different density function and a different cumulative distribution function! In order to avoid having to do a new integration every time we come across a normally distributed variable (which happens a lot!), mathematicians have tabulated various areas under the normal curve for something called a *standard normal random variable*, usually denoted by Z, which has mean zero and a variance of 1 i.e. $Z \sim N(0, 1)$. Fortunately, probabilities for *any* normal distribution can always be expressed in terms of probabilities for the standard normal variable, because for any X such that $X \sim N(\mu, \sigma_X^2)$, we have

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

If you subtract the mean from a normal random variable, and divide the whole thing by the standard deviation, you end up with a standard normal variable with mean zero and variance 1! It follows immediately that

$$\text{Prob}(a < X < b) = \text{Prob}\left(\frac{a-\mu}{\sigma_X} < \frac{X-\mu}{\sigma_X} < \frac{b-\mu}{\sigma_X}\right) = \text{Prob}\left(\frac{a-\mu}{\sigma_X} < Z < \frac{b-\mu}{\sigma_X}\right)$$

There is some conventional notation associated with the standard normal variable. Its probability density function is denoted by $\phi(z)$, and is of the form

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\}$$

Its cumulative distribution function is denoted by $\Phi(c)$, where

$$\Phi(c) = \text{Prob}(Z \leq c) = \int_{-\infty}^{c} \phi(z)dz$$

For any normal random variable $X \sim N(\mu, \sigma_X^2)$, we can find probabilities such as $\text{Prob}(a < X < b)$ using the cumulative distribution function of the standard normal variable as follows:

$$\text{Prob}(a < X < b) = \text{Prob}\left(\frac{a-\mu}{\sigma_X} < \frac{X-\mu}{\sigma_X} < \frac{b-\mu}{\sigma_X}\right) = \text{Prob}\left(\frac{a-\mu}{\sigma_X} < Z < \frac{b-\mu}{\sigma_X}\right)$$

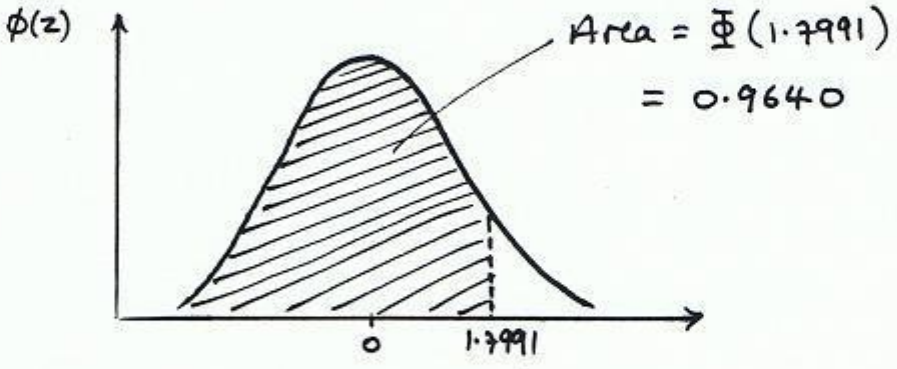
$$= \Phi\left(\frac{b-\mu}{\sigma_X}\right) - \Phi\left(\frac{a-\mu}{\sigma_X}\right)$$

Given a 'cutoff' point c, you will usually be able to look up the corresponding value $\Phi(c)$ from a table such as the following:

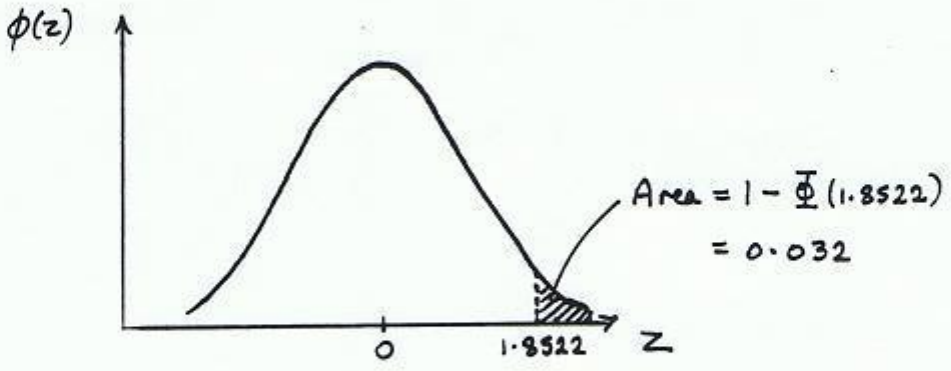
Values of the cumulative distribution function of the standard normal distribution

c	Φ(c)	c	Φ(c)	c	Φ(c)
0.0000	0.5000	1.8808	0.9700	2.3263	0.9900
0.2533	0.6000	1.8957	0.9710	2.3656	0.9910
0.5244	0.7000	1.9110	0.9720	2.4089	0.9920
0.8416	0.8000	1.9268	0.9730	2.4573	0.9930
1.0364	0.8500	1.9431	0.9740	2.5121	0.9940
1.2816	0.9000	1.9600	0.9750	2.5758	0.9950
1.3408	0.9100	1.9774	0.9760	2.6521	0.9960
1.4051	0.9200	1.9954	0.9770	2.7478	0.9970
1.4758	0.9300	2.0141	0.9780	2.8782	0.9980
1.5548	0.9400	2.0335	0.9790	3.0902	0.9990
1.6449	0.9500	2.0537	0.9800	3.2905	0.9995
1.6646	0.9520	2.0749	0.9810	3.7190	0.9999
1.6849	0.9540	2.0969	0.9820	3.8906	0.99995
1.7060	0.9560	2.1201	0.9830	4.2649	0.99999
1.7279	0.9580	2.1444	0.9840	4.4172	0.999995
1.7507	0.9600	2.1701	0.9850	4.7534	0.999999
1.7744	0.9620	2.1973	0.9860	4.8916	0.9999995
1.7991	0.9640	2.2262	0.9870	5.1993	0.9999999
1.8250	0.9660	2.2571	0.9880	5.3267	0.99999995
1.8522	0.9680	2.2904	0.9890	5.6120	0.99999999

For example, according to this table we have $\text{Prob}(Z \leq 1.7991) = \Phi(1.7991) = 0.9640$:



As another example: Prob(Z > 1.8522) = 1 - Prob(Z ≤ 1.8522) = 1 - 0.9680 = 0.032 :



Here is an example involving the transformation of a normal random variable into a standard normal one, for the purposes of calculating probabilities:

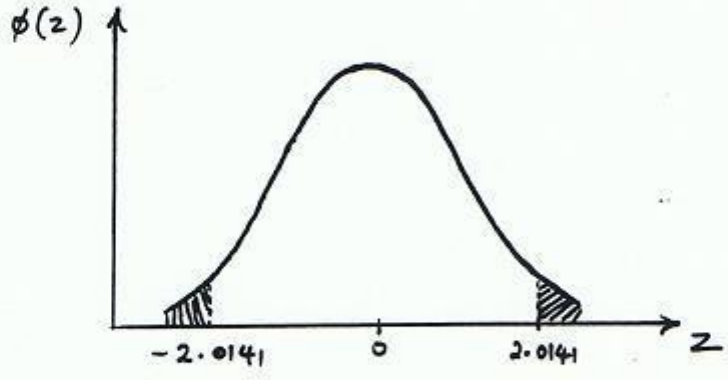
Example: Suppose X~N(5, 9). Find Prob(11.2247 < X < 16.1570).
Solution: In this example, we have $\mu = 5$ and $\sigma_X^2 = 9$, so the standard deviation is $\sigma_X = 3$. So

$$\text{Prob}(11.2247 < X < 16.1570) = \text{Prob}\left(\frac{11.2247 - 5}{3} < \frac{X - 5}{3} < \frac{16.1570 - 5}{3}\right)$$
$$= \text{Prob}(2.0749 < Z < 3.7190)$$
$$= \Phi(3.7190) - \Phi(2.0749)$$
$$= 0.9999 - 0.9810$$
$$= 0.0189$$

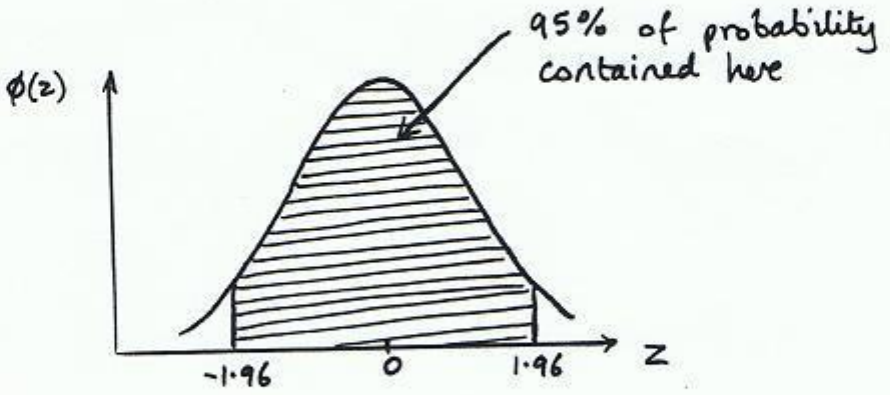
Hopefully, you will have seen lots of these in your previous studies, so we shall leave it at that. There are two more things I want you to note about the standard normal distribution at this stage. Firstly, since the normal curve is symmetric, we have

$$\Phi(-c) = 1 - \Phi(c)$$

For example, Φ(-2.0141) = 1 - Φ(2.0141) = 1 - 0.9780 = 0.022 :



The second thing I want you to note is that
Prob(-1.96 < Z < 1.96) = Φ(1.96) - Φ(-1.96) = Φ(1.96) - {1 - Φ(1.96)} = 0.9750 - 0.025 = 0.95 :



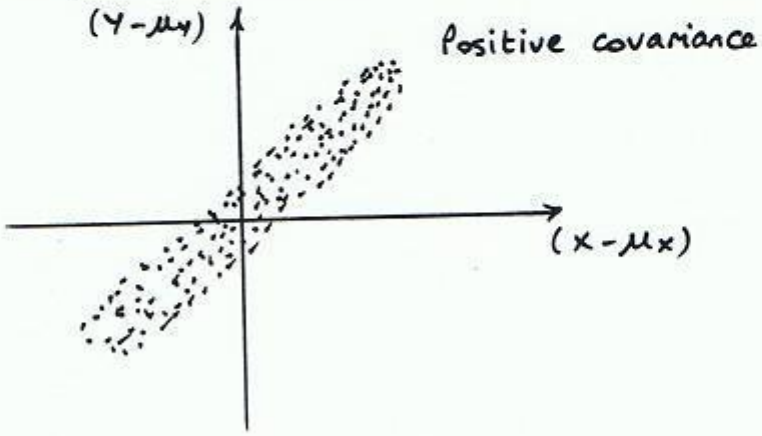
So 95% of the probability mass for a standard normal random variable is contained between the limits -1.96 and 1.96. Since $Z = (X-\mu)/\sigma_X$, we can rewrite this result as

$$\text{Prob}(-1.96 < (X-\mu)/\sigma_X < 1.96) = \text{Prob}(\mu - 1.96\sigma_X < X < \mu + 1.96\sigma_X) = 0.95$$

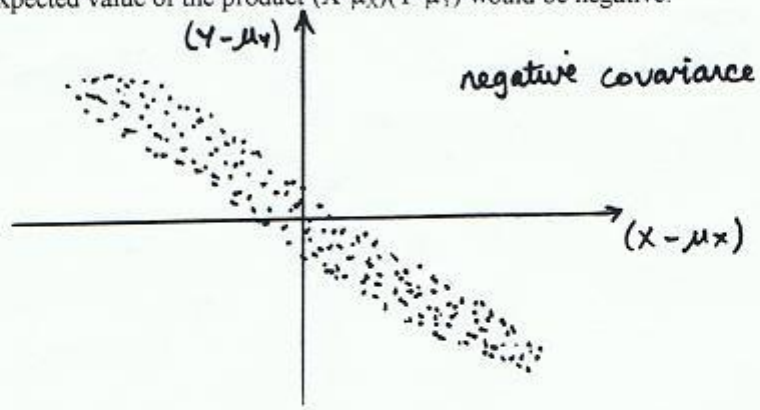
In words, 95% of the probability mass for any normal variable $X \sim N(\mu, \sigma_X^2)$ is contained between the limits $\mu - 1.96\sigma_X$ and $\mu + 1.96\sigma_X$.

10.6. The covariance between two random variables

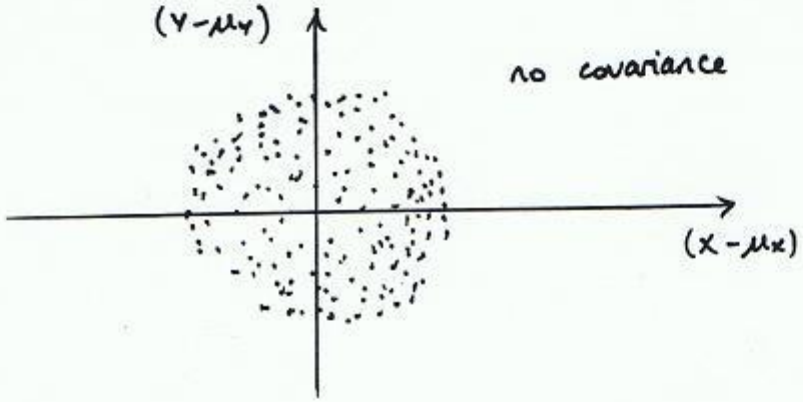
In the last lecture, we reviewed the concept of the expected value of a random variable. We saw how we could also find the expected value of a *function* of a random variable. An important application of the expectation of a function of random variables is to the covariance. Let X and Y be any two random variables (either discrete or continuous), and let $E[X] \equiv \mu_X$ and $E[Y] \equiv \mu_Y$. Often, we would like some measure of the nature and strength of the relationship between two variables such as these. This is difficult to achieve, because the variables could in principle be related in any number of ways. To make things easier, we restrict our attention to the linear association. For example, a high value of X might be associated on the average with a high value of Y, and a low value of X with a low value of Y, in such a way that, to a good approximation, a straight line might be drawn through the associated values when plotted on a graph. Consider the product $(X-\mu_X)(Y-\mu_Y)$. If high values of X tend to be associated with high values of Y, and low values of X with low values of Y, we would expect this product to be positive, and the stronger the association, the larger the expectation of $(X-\mu_X)(Y-\mu_Y)$:



By contrast, if high values of X are associated with low values of Y, and low values of X with high values of Y, the expected value of the product $(X-\mu_X)(Y-\mu_Y)$ would be negative:



An expectation of 0 for $(X-\mu_X)(Y-\mu_Y)$ would imply the absence of a linear association between X and Y:



Thus, as a measure of the linear association between two random variables X and Y, we are led to an examination of the expected value of $(X-\mu_X)(Y-\mu_Y)$. ***The expected value of $(X-\mu_X)(Y-\mu_Y)$ is called the <u>covariance</u> between X and Y, denoted by Cov[X, Y], or by the Greek symbol σ_{XY}.*** For discrete random variables,

$$\text{Cov}[X, Y] = E[(X-\mu_X)(Y-\mu_Y)] = \Sigma\Sigma(x_i - \mu_X)(y_i - \mu_Y)P_{XY}(x_i, y_i)$$

where $P_{XY}(x_i, y_i) \equiv \text{Prob}(X = x_i \text{ and } Y = y_i)$. For continuous random variables, we replace summation by integration, and we replace the (bivariate) <u>probability</u> by the (bivariate) <u>density</u>:

$$\text{Cov}[X, Y] = E[(X-\mu_X)(Y-\mu_Y)] = \iint (x-\mu_X)(y-\mu_Y)f_{XY}(x, y)dxdy$$

To fix ideas, consider the following simple example. Let X and Y be a pair of discrete random variables measuring, respectively, a consumer's satisfaction with food shops in a particular town, and the number of years of residence in that town. Suppose that X can take values 1, 2, 3, or 4,

ranging from low to high satisfaction levels, and Y takes the value 1 if the consumer has lived in the town less than six years, and 2 otherwise. The table below shows the eight joint probabilities for X and Y:

Probabilities for consumer satisfaction (X) and time in residence (Y)

Y	X				
	1	2	3	4	*Totals*
1	0.04	0.14	0.23	0.07	0.48
2	0.07	0.17	0.23	0.05	0.52
Totals	0.11	0.31	0.46	0.12	1.00

For example, according to the table, $P_{XY}(2, 1) \equiv \text{Prob}(X = 2, Y = 1) = 0.14$, and $P_{XY}(3, 1) = 0.23$. The first step in finding the covariance between X and Y is to find their means:

$E[X] = \mu_X = (1)(0.11) + (2)(0.31) + (3)(0.46) + (4)(0.12) = 2.59$

$E[Y] = \mu_Y = (1)(0.48) + (2)(0.52) = 1.52$

Then the covariance is given by

$$\text{Cov}[X, Y] = E[(X-\mu_X)(Y-\mu_Y)] = \Sigma\Sigma(x_i - \mu_X)(y_i - \mu_Y)P_{XY}(x_i, y_i)$$
$$= (1-2.59)(1-1.52)(0.04) + (1-2.59)(2-1.52)(0.07) + (2-2.59)(1-1.52)(0.14)$$
$$+ (2-2.59)(2-1.52)(0.17) + (3-2.59)(1-1.52)(0.23) + (3-2.59)(2-1.52)(0.23)$$
$$+ (4-2.59)(1-1.52)(0.07) + (4-2.59)(2-1.52)(0.05) = -0.05$$

This calculation could have been made easier by using the fact that

$$\text{Cov}[X, Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

To prove that we get the same result, note that

$$E[XY] = \Sigma\Sigma x_i y_i P_{XY}(x_i, y_i) = (1)(1)(0.04) + (1)(2)(0.07) + (2)(1)(0.14) + (2)(2)(0.17)$$
$$+ (3)(1)(0.23) + (3)(2)(0.23) + (4)(1)(0.07) + (4)(2)(0.05) = 3.89$$

Thus $\text{Cov}[X, Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y = 3.89 - (2.59)(1.52) = -0.05$

This negative value for the covariance indicates some tendency for high values of consumer satisfaction to be associated with a low period of time in residence in the town i.e. a *negative* association between this pair of random variables.

There are three rules for manipulating covariances which you must learn:

(1). If Y = U + V, then Cov[X, Y] = Cov[X, U] + Cov[X, V]
(2). If Y = aZ, where a is a constant, then Cov[X, Y] = aCov[X, Z]
(3). If Y = a, where a is a constant, then Cov[X, Y] = 0

10.7. The correlation between two random variables

We saw in the previous section that the covariance between two random variables is one possible measure of the nature and strength of the linear association between them. However, it is not as good an indicator of the *strength* of the linear association as it might be, because it is affected by the scales in which the variables are measured. For example, if X and Y are random variables measuring the returns from two shares, we get a much smaller value for Cov[X, Y] if we measure

the returns in pounds rather than pence. Since the scale of measurement is totally arbitrary, we would ideally like a pure, scale-free measure of the linear association between X and Y. Such a measure can easily be obtained by dividing the covariance by the product of the individual standard deviations. The resulting quantity is called the correlation coefficient.

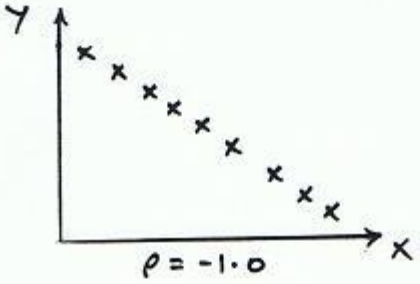
Let X and Y be a pair of random variables, with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2. A measure of the strength of their linear association is provided by the correlation coefficient ρ, which is defined as $\rho = Cov[X, Y]/\sigma_X \sigma_Y$. It is also denoted by the symbol Corr[X, Y].

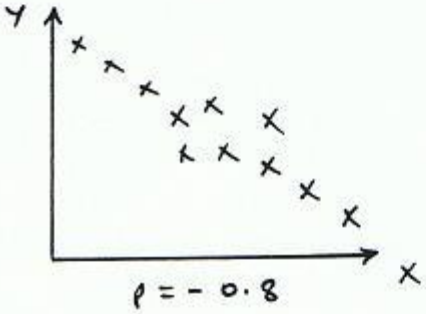
It can be shown that the correlation coefficient must lie between -1 and 1. That is

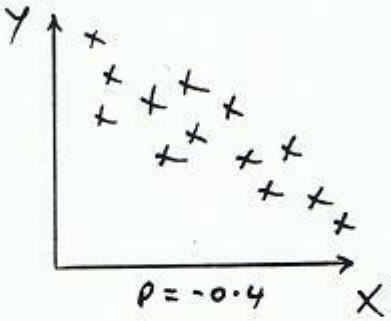
$$-1 \leq \rho \leq 1$$

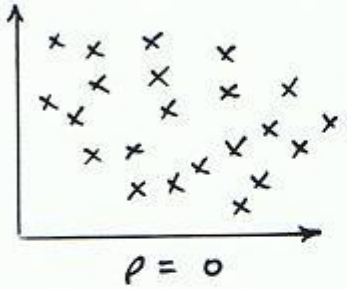
and the values have the following interpretations:
(1). A correlation of -1 implies perfect negative linear association.
(2). A correlation of 1 implies perfect positive linear association.
(3). A correlation of 0 implies no linear association.
(4). The larger the absolute value of the correlation, the stronger the linear association between the random variables.

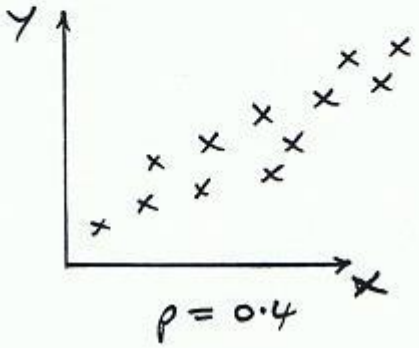
Here are some plots of various values of two hypothetical random variables, with the correlation coefficient they would give rise to:

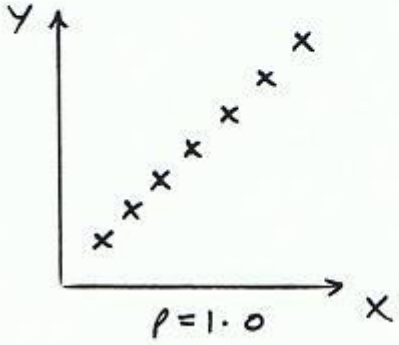












To finish off this section, let us calculate the correlation between the variables X and Y in the example of the last section. We found that the covariance between X and Y is Cov[X, Y] = -0.05.

To find the correlation, we need the two variances σ_X^2 and σ_Y^2. These are

$V[X] = E[(X-\mu_X)^2] = \Sigma(x_i-\mu_X)^2P_X(x_i) = (1-2.59)^2(0.11) + (2-2.59)^2(0.31) + (3-2.59)^2(0.46) + (4-2.59)^2(0.12) = 0.7019$

$V[Y] = E[(Y-\mu_Y)^2] = \Sigma(y_i-\mu_Y)^2P_Y(y_i) = (1-1.52)^2(0.48) + (2-1.52)^2(0.52) = 0.2496$

Therefore $Corr[X, Y] = Cov[X, Y]/\sigma_X\sigma_Y = -0.05/(\sqrt{0.7019})(\sqrt{0.2496}) = -0.1195$

10.8. Estimation

So far, we have assumed that we have exact information about the random variables under discussion, in particular that we know the probability mass function in the case of a discrete random variable, or the probability density function in the case of a continuous random variable. With this information, it is possible to work out the population mean and variance, and any other population characteristics we might be interested in. In real life, however, we usually do not know the exact probability mass function or density function. It follows that we do not know the population mean or variance. However, we would like to obtain an estimate of them.

The basic procedure is always the same. ***You take a sample of N observations, and derive an estimate of the population characteristic using some appropriate formula.*** The formula is called an estimator. The actual number you get when you apply an estimator to data is the estimate. Be sure to carefully distinguish between estimators and estimates: ***the estimator is a general rule or formula, whereas the estimate is a specific number that will vary from sample to sample.*** The following are the usual estimators of the two most important population characteristics of a single random variable:

Estimator of the population mean μ: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Estimator of the population variance σ_X^2: $s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

Note that these are the *usual* estimators of the population mean and variance. They are not the only ones. The following are the usual estimators of the covariance between two random variables X and Y, and the correlation between them (it is assumed that there are N observations on the two variables):

Estimator of the covariance σ_{XY}: $s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

Estimator of the correlation coefficient ρ: $r = s_{XY}/s_Xs_Y$

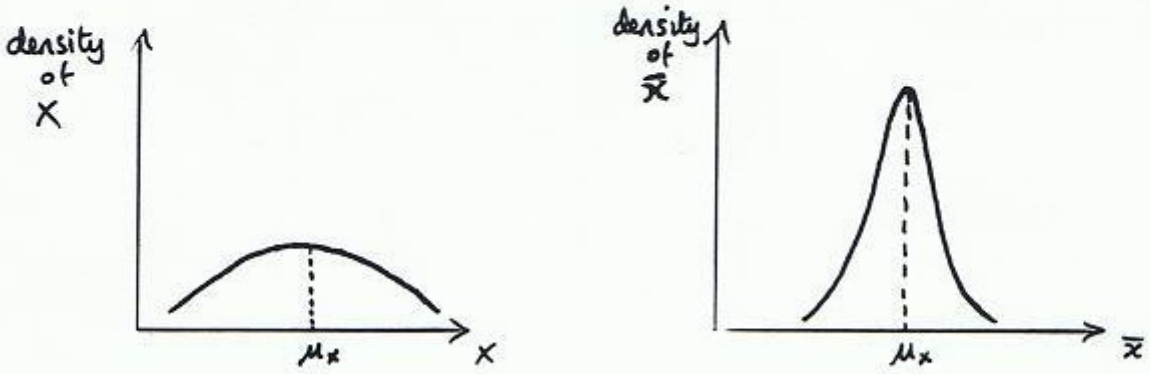
The important thing to realise is that there are many possible alternative estimators for the population parameters above. The reason we use the above estimators is that they are the 'best' according to two very important criteria: unbiasedness and efficiency.

10.9. Unbiasedness

An estimator is a special case of a random variable. This is because it is a combination of the values of X in a sample, and, since X is a random variable, a combination of a set of its values must also be a random variable. For example, take $\bar{x}$, the estimator of the mean:

$$\bar{x} = (1/N)(x_1 + x_2 + \cdots + x_N)$$

Suppose X has a normal distribution, with mean μ_X and variance σ_X^2. The diagram below shows the probability density functions for X and $\bar{x}$:



You will see that the distributions of both X and $\bar{x}$ are centered over μ_X, the population mean. The difference between them is that the distribution of $\bar{x}$ is narrower and taller, so that $\bar{x}$ is likely to be closer to μ_X than a single observation on X. This is because $\bar{x}$ 'averages out' the random errors associated with specific values of X in estimating μ_X. Similarly, s_X^2, the estimator of the population variance of X, is also a random variable.

Since estimators are random variables, it follows that only by coincidence will an estimate be exactly equal to the population characteristic. Usually, there will be some degree of error. Although we must accept this, we would like an estimator that is accurate *on average*. Technically, we want the expected value of the estimator to be equal to the population characteristic. If this is true, the estimator is said to be unbiased. If it is not, the estimator is said to be biased, and the difference between its expected value and its population characteristic is called the bias.

Consider the sample mean. It is easy to show that $E[\bar{x}] = \mu_X$. This is true because

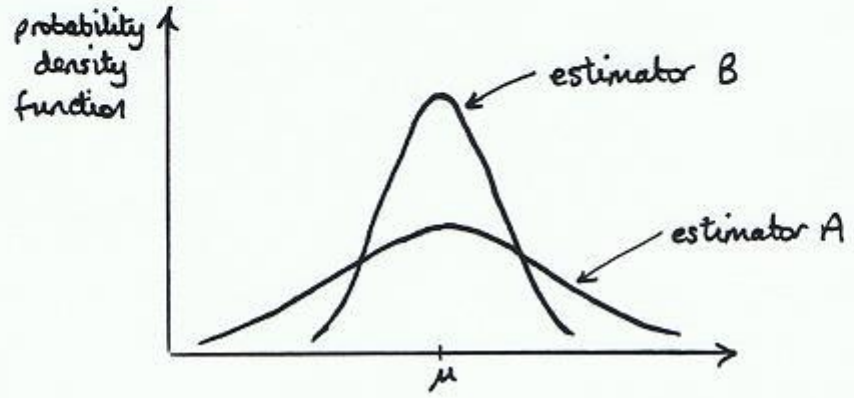
$$E[\bar{x}] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \frac{1}{N}\sum_{i=1}^{N} \mu_X = \frac{N}{N}\mu_X = \mu_X$$

where I have used X_i to denote the ith observation of the random variable X. Thus, $\bar{x}$ is unbiased. It can also be shown straightforwardly that $E[s_X^2] = \sigma_X^2$, so that the sample variance s_X^2 is an unbiased estimator of σ_X^2. Similarly, the estimators of the covariance and correlation between two random variables are also unbiased.

10.10. Efficiency

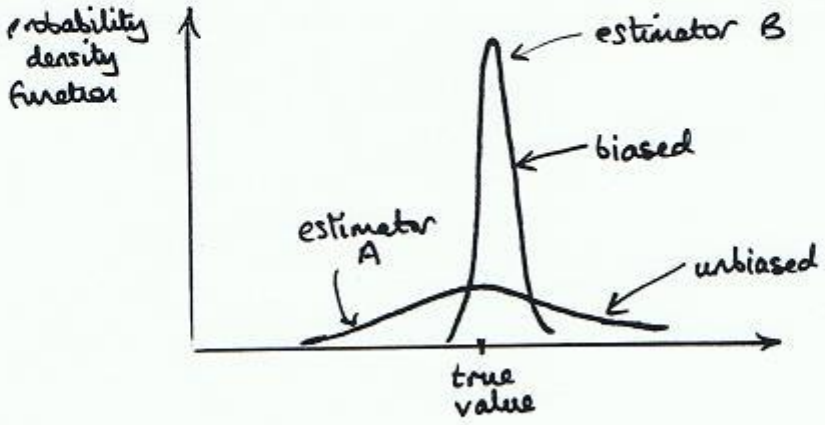
Unbiasedness is one desirable feature of an estimator, but it is not the only one. Unbiasedness means that 'on average', the estimator will tend to give a correct estimate of the population characteristic. What we would also like is for the estimator to have as high a probability as possible of giving a close estimate of the population characteristic, which means that ***we want its probability density function to be as concentrated as possible around the true value***. One way of summarising this is to say that we want its variance to be as small as possible.

Suppose that we have two estimators of the population mean, calculated using the same information, that they are both unbiased, and that their probability density functions are as shown in the following diagram:



Since the probability density function for the estimator B is more highly concentrated than that for estimator A, it is more likely to give an accurate estimate. Technically, it is said to be more efficient. Notice that the definition says 'more likely'. Even though estimator B is more efficient, that does not mean that it will always give the more accurate estimate. It just means that it has a higher probability of producing an accurate estimate of the true mean.

We have said that we want the variance of an estimator to be as small as possible, and that the most efficient estimator is the one with the smallest variance. We also said that we want the estimator to be unbiased. These are two quite different criteria, and they sometimes conflict with each other. ***It sometimes happens that we can construct two estimators of a population characteristic, one of which is unbiased (A in the diagram below), the other being biased but having a smaller variance (B below):***



A is better than B in the sense that it is unbiased, but B is better than A in the sense that its estimates are always close to the true value. How do you choose between them ? It will depend on the circumstances. If you are not bothered by errors, provided that on average they cancel out, then you should probably choose A. On the other hand, if you can tolerate small errors but not large ones, you should choose B.

10.11. Consistency

We shall continue to assume that we are investigating a random variable X with unknown mean μ_X and unknown variance σ_X^2, and that we are using $\bar{x}$ to estimate μ_X. ***How does the accuracy of*** $\bar{x}$ ***depend upon the number of observations N ?***

Not surprisingly, the answer is that as you increase N, you increase the probable accuracy of $\bar{x}$. This is because it can be shown that $V[\bar{x}] = \sigma_X^2/N$. Thus, as N rises, the variance of the probability distribution of $\bar{x}$ decreases. ***The bigger the sample, the smaller is the variance of the probability distribution of*** $\bar{x}$ ***and so the more 'tightly compressed' it is around the true mean.***

This is demonstrated in diagram (A) below for a hypothetical random variable X. As N increases, the probability distribution becomes narrower and taller. If N becomes really large, the probability distribution will be like a vertical line located at $\bar{x} = \mu_X$. This is because in the limit, as N goes to infinity, σ_X^2/N tends to zero and $\bar{x}$ tends to μ_X exactly. This is written mathematically as

$$\lim_{N\to\infty} \bar{x} = \mu_X$$

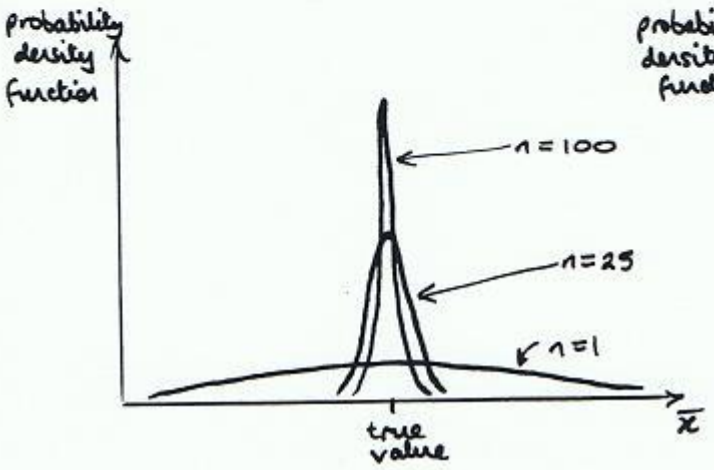
An equivalent and more common way of expressing it is to use the term 'plim', where plim stands for 'probability limit', and emphasises that the limit is being reached in a probabilistic sense:

$$\text{plim}\, \bar{x} = \mu_X$$

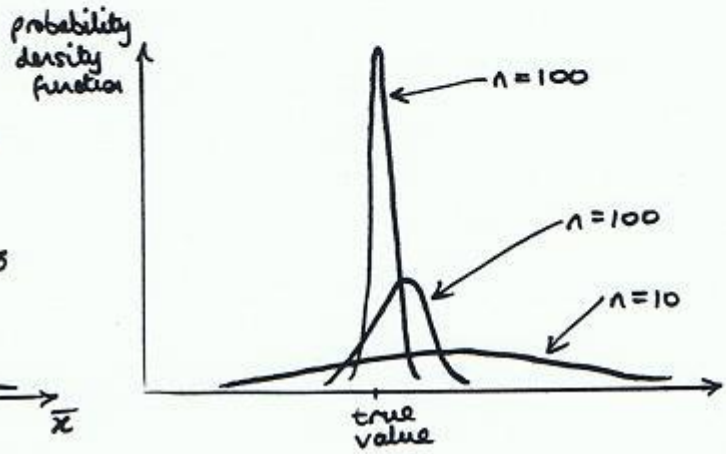
In general, if the plim of an estimator is equal to the true value of the population characteristic, it is said to be consistent. In words, ***a consistent estimator is one that is bound to give an accurate estimate of the population characteristic if the sample is large enough, regardless of the actual observations in the sample.***

In most of the contexts considered in this course, an unbiased estimator will also be a consistent one. It is possible to invent examples for which this is not true, but they are usually artificial. The reason we are interested in consistency is that ***it sometimes happens that an estimator that is biased for small samples may be consistent.*** This can be used to justify its use, despite the fact that it is biased in small samples. Diagram (B) below shows how the probability distribution might look for different sample sizes. The fact that the distribution becomes centred on the true value as the sample size increases, and eventually collapses to a line at the true value, indicates that it is consistent.

(A) Unbiased and consistent



(B) Biased but consistent



As we shall see later in the course, estimators of the type shown in the diagram above are quite important in econometrics. Sometimes it is impossible to find an estimator that is unbiased in small samples. If you can find one that is at least consistent, that may be better than having no estimate at all. However, you should remember that a consistent estimator might perform worse than an inconsistent one in small samples.

10.12. <u>What you must do before the lecture next week (Tuesday, 29th January, 1999, 'Introduction to econometric models')</u>

Read these lecture notes carefully and make sure you understand all the concepts and terms. If you need help, see me or someone else about it *before* next week. Do all the questions on the attached assignment sheet for this lecture. As usual, solutions are attached.

Assignment for Lecture 10. Essential statistical concepts

Question 1 Suppose $X \sim N(10, 100)$. Use the table on page (5) of the lecture notes to find the following probabilities:

(a) $P(27.911 < X < 28.808)$

(b) $P(X > 29.6)$

(c) $P(X < -4.051)$

Question 2 Two discrete random variables X and Y have the following joint and marginal probability mass functions:

X \ Y	0	1	2	$P_X(x)$
0	1/5	1/5	1/5	3/5
1	1/5	0	0	1/5
2	1/5	0	0	1/5
$P_Y(y)$	3/5	1/5	1/5	1.0

(a) Find $E[X]$ and $E[Y]$

(b) Find $V[X]$ and $V[Y]$

(c) Find $Cov[X, Y]$ and $Corr[X, Y]$

(d) Are X and Y independent? Explain

(Hint: see page (9) of lecture notes).

Assignment for Lecture 10. Essential statistical concepts

SOLUTIONS

Question 1

(a) $P(27.911 < X < 28.808)$

$= P\left(\frac{27.911-10}{10} < Z < \frac{28.808-10}{10}\right)$

$= P(1.7911 < Z < 1.8808)$

$= \Phi(1.8808) - \Phi(1.7911) = 0.97 - 0.964$

$= 0.006$

(b) $P(X > 29.6) = P\left(Z > \frac{29.6-10}{10}\right)$

$= P(Z > 1.96) = 1 - \Phi(1.96) = 0.025$

(c) $P(X < -4.051) = P(Z < -1.40511)$

$= P(Z > 1.40511)$ (by symmetry)

$= 1 - P(Z < 1.40511) = 1 - \Phi(1.40511)$

$= 0.08$

Question 2

(a) $E[X] = 3/5 \qquad E[Y] = 3/5$

(b) $V[X] = 16/25 \qquad V[Y] = 16/25$

(c) $Cov[X,Y] = -\frac{9}{25}$

$Corr[X,Y] = \frac{-9}{16}$

(d) No: $E[XY] = 0 \neq E[X] \cdot E[Y] = \frac{9}{25}$