

Preliminary Statistics

I will be presenting some basic ideas on probability and statistics. The statistics material supports Quantitative Methods and the Econometrics courses. The probability material is used there but also in decision making under uncertainty in various Economics and Finance courses.

Entire books are devoted to my topics.

For a mathematical treatment e.g.

D. Wackerly, W. Mendenhall & R. Scheaffer *Mathematical Statistics with Applications*.

For a less mathematical treatment e.g.

E. Mansfield *Statistics for Business and Economics*.

The topics are covered (in less detail) in chapters 2 and 3 of Stock and Watson's *Introduction to Econometrics*, the

textbook for Quantitative Methods.

Probability SW ch 2

The probability of an event can be understood in different ways, there are alternative interpretations of probability.

- An objective interpretation—relative frequency—is more common in Econometrics/Statistics.
- A subjective interpretation—degree of belief—is more common in Economics/Finance.

The mathematical formalism is the same.

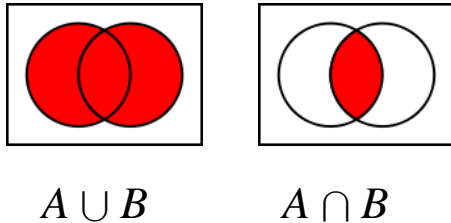
Events

Suppose an *experiment* or chance process has associated with it a *sample space* S , covering all the possible outcomes; S is also called the *certain event*.

In a Venn diagram we represent S by the inside of a box and events A and B by the

insides of circles—such as the circles in the diagrams below.

The event A or B or both is also called the union of A and B and is represented by $A \cup B$. The intersection $A \cap B$ is the event according to which both A and B occur.



If there is no overlap, i.e., the events cannot occur together then the events are called mutually exclusive or disjoint events.

Probability axioms

To every event A in S (A is a subset of S) we assign a number $P(A)$, so that the following axioms hold

Axiom 1

$$P(A) \geq 0$$

Axiom 2

$$P(S) = 1$$

Axiom 3 (Addition Rule for Mutually Exclusive Events)

If A_1, A_2, A_3, \dots form a sequence of pairwise mutually exclusive events in S , then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = \sum P(A_i)$$

Interpretations

- In the relative frequency interpretation $P(A)$ is the limiting proportion of times in which A occurs as the experiment is repeated indefinitely.
- In the subjective interpretation $P(A)$ is a measure of your degree of belief in the occurrence of A . It is related to the odds at which you are prepared to bet on the

occurrence of A .

Formalising these interpretations and showing that the axioms hold for them is a big task and I will not attempt it.

Proofs

I prove one theorem to give the flavour of axiomatic reasoning in probability.

Theorem *If event A has probability $P(A)$ then not- A the complementary event, \bar{A} , has probability $P(\bar{A}) = 1 - P(A)$.*

Proof

A and \bar{A} are mutually exclusive so from Axiom 3

$$P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

A and \bar{A} contain all possibilities, i.e.

$$A \cup \bar{A} = S.$$

So using Axiom 2 we have

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

rearranging gives the theorem.

Conditional Probability

The probability that an event A will occur, given that another event B has occurred or is certain to occur, is a *conditional probability*. The new space of possibilities—the new sample space—are subsets of the event B .

The symbol for the conditional probability of A given B is $P(A|B)$ and it is defined by

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

The definition of conditional probability rewritten as

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

called the *multiplication rule*. Reversing the roles of A and B gives the equally valid

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

Combining the two equations gives a relationship between $P(A|B)$ and $P(B|A)$:

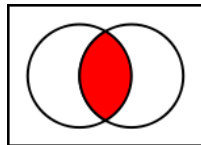
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Total probability

The *law of total probability* relates the probability of an event to any underlying conditional probabilities

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

The first equation holds because $(B \cap A)$ and $B \cap \bar{A}$ are disjoint events whose union is B :



B the circle on the right divides into $B \cap A$ (the red area) and $B \cap \bar{A}$, the white area.

The second equation uses the definition of conditional probability.

Example

At a university in northern California for a male the chance of being accepted is 0.19 and for a female 0.115.

The university's lawyer argues that this is NOT discrimination. The chance of admission to Arts is the same for males and females and so is the chance of admission to Science (there are only 2 faculties). HOWEVER females are more likely to apply to Arts where it is more difficult to get in.

For a female the probability of being accepted by the university is

$$\begin{aligned} P(A) &= P(A|Arts)P(Arts) + P(A|\overline{Arts})P(\overline{Arts}) \\ &= 0.1 \cdot 0.9 + 0.25 \cdot 0.1 \\ &= 0.115. \end{aligned}$$

where $P(A|Arts)$ is the probability of being accepted into Arts and $P(Arts)$ is the probability of applying to Arts. Because there are only 2 faculties \overline{Arts} is Science.

For a male the probability of being accepted

$$\begin{aligned}
P(A) &= P(A|Arts)P(Arts) + P(A|\overline{Arts})P(\overline{Arts}) \\
&= 0.1 \cdot 0.6 + 0.25 \cdot 0.4 \\
&= 0.16.
\end{aligned}$$

So the admission probabilities $P(A|Arts)$ and $P(A|\overline{Arts})$ are the same for men and women BUT the popularity of the subjects, $P(Arts)$ and $P(\overline{Arts})$ are quite different.

Bayes

Take

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

and apply the law of total probability to obtain

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

known as *Bayes' theorem* or *Bayes' rule*.

The theorem has numerous uses but a very important one is in association with subjective probability—a description of how the subjective probability of A is

updated in the light of the information that B has occurred—how $P(A)$ is transformed to $P(A|B)$. Hence the expression *Bayesian updating*.

Example

Testing for AIDS.

Event A is that a person has AIDS; event B is the bad outcome, that the test is positive.

Suppose .006 of the population has AIDS so for a randomly chosen person $P(A) = .006$ and $P(\overline{A}) = .994$.

The testing procedure is good but not perfect: $P(B|A) = .999$ and $P(B|\overline{A}) = .01$.

$$\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \\
&= \frac{.999 \cdot .006}{.999 \cdot .006 + .01 \cdot .994} \\
&= .38
\end{aligned}$$

The Bayesian updating is impressive: a prior probability of .006 has been transformed to a posterior probability of

.38.

Note that $P(\bar{A}|B) = 1 - P(A|B) = .62$ so that 62% of those testing positive do NOT have AIDS.

2 factors are involved

- The condition is rare.
- The testing technology is not perfect.

Independence

Underlying many applications of probability is the notion of *independence*. There are also many situations in which *dependence* is modelled by building a process out of independent components.

The events A and B are **independent** if

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

The occurrence of B does not alter the probability of occurrence of A and vice versa.

If A and B are independent then by the multiplication rule

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

This is sometimes called the product rule.

Exercise 1

1. The impossible or empty event, \emptyset , is the complementary event to the certain event S

$$\emptyset = \bar{S}.$$

Show from the axioms that $P(\emptyset) = 0$.

2. Prove that if A_1 and A_2 are not disjoint

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

3. Show that if A and B are independent then A and \bar{B} are independent.

4. A university in southern California discriminates against men yet the probability of a male applicant being admitted is still greater than the probability for a female. Invent some numbers to demonstrate this possibility. (This phenomenon is sometimes called *Simpson's paradox*.)

5. Consider Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}.$$

a) Show that the right hand side reduces to $P(A)$ when A and B are independent.

b) Suppose A and B are mutually exclusive (disjoint) events. (Using the symbol introduced in question 1, $A \cap B = \emptyset$.) What is the value of $P(A|B)$ in this case?

6. Consider testing for AIDS. Redo the calculation of the posterior probabilities for the cases:

a) For a subgroup where AIDS is more common, $P(A) = .06$.

b) When the testing technology has become perfect: $P(B|A) = 1$ and $P(B|\bar{A}) = 0$.

Random Variables

A **random variable** is a numerical quantity the value of which is determined by the outcome of an experiment.

Example *The Bernoulli or indicator*

random variable, X , codes for the occurrence or non-occurrence of an event E : 1 for occurrence, 0 for non-occurrence. If $P(E) = \pi$ then:

$$X = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1 - \pi \end{cases}$$

Example Dice throwing with a fair die

$$Y = \begin{cases} 1 & \text{with probability } \frac{1}{6} \\ 2 & \text{with probability } \frac{1}{6} \\ \dots & \\ 6 & \text{with probability } \frac{1}{6} \end{cases}$$

This is an special case of the **discrete uniform** variable for which

$$P(X = i) = \frac{1}{N} \text{ for } i = 1, \dots, N$$

The function which assigns probabilities to the values taken by the random variable is called the **probability function**, or probability distribution or

probability mass function.

Write $p(x)$ for the probability distribution i.e. for $P(X = x)$.

If there is a danger of confusion, it is best to be explicit and include the name of the random variable and write $p_X(x)$; mostly we won't need to.

Example For the Bernoulli r.v. X :

$$p_X(1) = \pi \text{ and } p_X(0) = 1 - \pi.$$

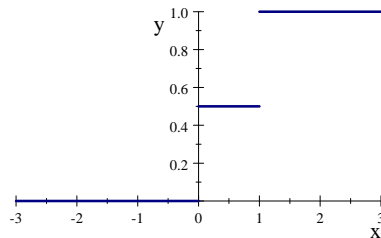
The Bernoulli and the discrete uniform are *discrete* random variables: X and Y take a finite number of values 2 and N respectively. Other discrete variables take a countably infinite number of values (as many values as there are integers). Continuous random variables take uncountably many values.

Random variables: distribution functions

For any random variable we can define the **distribution function**

$$F(x) = P\{X \leq x\}.$$

(SW call it the cumulative distribution.)
 For discrete random variables the distribution function is a step function.
 E.g.

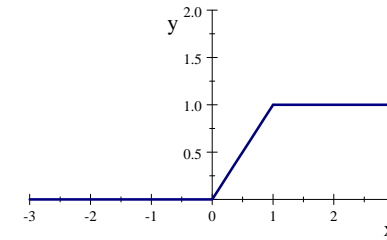


for indicator with $\pi = \frac{1}{2}$

But there also random variables for which the distribution function has no jumps.

Example *The continuous uniform has d.f.*

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$



F for uniform on $(0,1)$

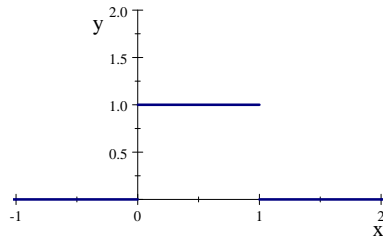
If $F(x)$ be the distribution function for a continuous random variable X then $f(x)$ given by

$$f(x) = \frac{dF(x)}{dx}$$

is called the probability density function of X .

For the uniform

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{dx}{dx} = 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } x \geq 1 \end{cases}$$



for uniform on (0,1)

The p.d.f. plays a similar role for continuous random variables as the mass function for discrete.

Expected Values

The *expected value* or *expectation* or *mean* of a discrete random variable X , denoted by $E(X)$ is defined as

$$E(X) = \sum_{\text{all values of } X} xp(x)$$

$E(X)$ may be written as EX .

- The expectation of X is a *weighted average* of the values of X ; each value is weighted by the probability that value is realised.
- The expectation is also called the **mean** of the random variable and often

represented by μ . If there is danger of ambiguity we might write μ_X to emphasise that it's the mean of X .

Example *The Bernoulli/indicator random variable X has expected value*

$$\begin{aligned} EX &= 0 \cdot (1 - \pi) + 1 \cdot \pi \\ &= \pi. \end{aligned}$$

The **expected value** or **expectation** of a continuous random variable is defined similarly except that summation is replaced by integration

$$EX = \int xf(x)dx.$$

where the integral is over the range of values of X .

Example *For the uniform,*

$$EX = \int_0^1 x \cdot 1dx = \frac{1}{2}.$$

which is not surprising given that the density is symmetric around $\frac{1}{2}$.

Applications

1. In gambling—a form of decision-making under uncertainty—we might be interested in expected gain. Suppose the pay-off is W , defined by

$$W = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ +2 & \text{with probability } \frac{1}{2} \end{cases}$$

The mean $E(W) = 1$.

2. The decision maker may be concerned with expected utility. Consider expected utility of the gamble W when the utility function $u(w) = \sqrt{w}$.

$$Eu(W) = \sqrt{0} \cdot \frac{1}{2} + \sqrt{2} \cdot \frac{1}{2} = \frac{1}{\sqrt{2}}$$

which is less than 1, the expected winnings $E(W)$. A person with such a concave utility function is a risk averter who would prefer a sure thing of 1 to this gamble with expected

value 1. There is a result called *Jensen's Inequality* which says that for any concave function g

$$Eg(X) \leq g(EX).$$

3. In Statistics the random variable may be an estimator $\hat{\theta}$ of some quantity θ . If $E\hat{\theta} = \theta$ then the estimator is said to be *unbiased*; on average it is equal to the true value.

Variance

The **variance** of a random variable X , denoted by $var(X)$ or $\sigma^2(X)$ or σ_X^2 or just σ^2 is the expected value of the squared deviation of the random variable from its mean—in the discrete and continuous cases respectively

$$\sigma_X^2 = E([X - EX]^2) = \sum_{\text{range of } X} (x - EX)^2 p(x)$$

$$\sigma_X^2 = E([X - EX]^2) = \int_{\text{range of } X} (x - EX)^2 f(x) dx.$$

The square root of the variance is the

standard deviation.

In **financial** applications where the return on an asset is treated as a random variable the variance of the return is often used as a measure of the **riskiness** or the uncertainty of the return. High variance \Leftrightarrow high risk.

Example X is the indicator with probability $\frac{1}{2}$. We already know that $EX = \frac{1}{2}$.

Using the definition of the variance

$$\begin{aligned}\sigma_X^2 &= E([X - EX]^2) = \sum_{\text{all values of } x} \left(x - \frac{1}{2}\right)^2 p(x) \\ &= \left(0 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} \\ &= \frac{1}{4}.\end{aligned}$$

Example For the uniform random variable the variance is

$$\sigma_X^2 = \int_0^1 \left(x - \frac{1}{2}\right)^2 \cdot 1 dx = \frac{1}{12}.$$

Moments

The mean and variance of a random variable X , are among the **moments** of X

EX^k is the k -th moment (about the origin)

$E(X - EX)^k$ is the k -th moment about the mean

The mean is the first moment about the origin and the variance is the second moment about the mean.

These are sometimes called *population moments* to distinguish them from *sample moments*

$$\frac{\sum X_i^k}{n}, \text{ } k\text{-th sample moment}$$

$$\frac{\sum (X_i - \bar{X})^k}{n}, \text{ } k\text{-th sample moment about mean.}$$

We come to sample moments later.

Exercise 2

1. Repeat Application 2 about the decision-maker contemplating W but change the utility function to $u(w) = x^2$.
2. The basic facts about expected values are

$$Ea = a; \quad EX = aEX; \quad E(X + a) = a + EX.$$

where X is a random variable and a is a constant.

Use these facts to prove the following propositions:

i) $E(X - EX) = 0.$

ii)
 $E[(X - EX)^2] = E[(X - EX)X] = EX^2 - (EX)^2$

iii)
 $var(aX) = E[(aX - EaX)^2] = a^2E(X - EX)^2 =$

iv)
 $var(X + a) = E[(X + a - E(X + a))]^2 = varX.$

3. Use the propositions in question 2 to

answer the following

- a) Suppose income is measured in £s. How is the mean and variance of income changed if income is measured in millions of £s?
 - b) Suppose £1000 is added to income, how is the mean and variance of income changed?
4. Calculate the variance of the Bernoulli/indicator random variable when the probability is π .

If you want more practice, do the starred exercises in Stock and Watson.

Joint Distributions

When discussing conditional probability we contemplated 2 or more events simultaneously. We can contemplate 2 or more random variables together.

Through the device of indicators we can identify events with 2-valued random variables and generalise the concepts we had for events to random variables taking more values than 0 and 1.

I will give the definitions for the case of 2 *discrete* random variables—they extend to the case of several variables and to the continuous case.

Let X_1 and X_2 be random variables taking values x_1, x_2 . Write

$$P(X_1 = x_1 \text{ and } X_2 = x_2) = p_{X_1, X_2}(x_1, x_2)$$

for all possible values of x_1 and x_2 . The function p_{X_1, X_2} is called the *joint probability function* of X_1 and X_2 .

In the case of continuous variables there is a *joint density* usually denoted by f_{X_1, X_2} .

Example Suppose we are tossing a fair coin twice and the trials are independent. X_i indicates H on the i -th trial. The probability information about the behaviour of X_1 and X_2 can be put into a joint probability table where the entries are the 4 values of p_{X_1, X_2} .

$X_1 \backslash X_2$	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

where the entries are formed on the pattern

$$P(X_1 = 0 \text{ and } X_2 = 1) = P(TH) = \frac{1}{4}$$

Example Suppose we are betting on the toss of a fair coin. I win 1 if H and win -1 if T. We bet twice: W_1 denotes my winnings after one toss and W_2 my winnings after two tosses. Here are the values of p_{W_1, W_2} .

$W_1 \setminus W_2$	-2	-1	0	1	2
-1	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0
1	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$

Marginal distributions

For a pair of events A and B we have

$$P(B) = P(B \cap A) + P(B \cap \bar{A})$$

The corresponding relationship for a pair of random variables is

$$p_{X_1}(x_1) = \sum_{\text{all } x_2} p_{X_1, X_2}(x_1, x_2)$$

The function on the left p_{X_1} is called the *marginal probability function* of X_1 .

Example *Continuing with coin tossing: the marginals p_{X_i} appear in the last row/column*

$X_1 \setminus X_2$	0	1	X_1
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
X_2	$\frac{1}{2}$	$\frac{1}{2}$	

Thus

$$p_{X_1}(0) = \frac{1}{2}; p_{X_1}(1) = \frac{1}{2}$$

The adjective *marginal* distinguishes such distributions from *joint* distributions. Strictly 'marginal' is superfluous; the marginal probability function for X_1 is just the probability function for X_1 .

Conditional distributions

For a pair of events A and B we had the conditional probability defined as follows

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

For a pair of random variables the conditional mass function (or one of them) is defined

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_{X_1X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

As with events there is a concept of independence for random variables. If

$$p_{X_1X_2}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2)$$

for all x_1 and x_2 then X_1 and X_2 are *independent*.

Example *In Statistics a collection of n independent random variables X_1, \dots, X_n each with the same distribution is called a random sample.*

Thus

$$\begin{aligned} p_{X_1 \dots X_n}(x_1, \dots, x_n) &= p_{X_1}(x_1) \cdot \dots \cdot p_{X_n}(x_n) \\ &= p_X(x_1) \cdot \dots \cdot p_X(x_n) \end{aligned}$$

because the marginal distributions are all the same and equal to p_X , say.

Conditional expectations

Given the conditional distribution $p_{X_1|X_2}(x_1|x_2)$, the *conditional expectation* of X_1 given X_2 is defined in a natural way

$$E(X_1|X_2 = x_2) = \sum_{\text{all } x_1} x_1 p_{X_1|X_2}(x_1|x_2).$$

Corresponding to the *law of total probability* for events

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

there is a law of total expectation or *law of iterated expectations* for random variables

$$\begin{aligned} E(X_1) &= \sum_{\text{all } x_2} E(X_1|X_2 = x_2) p_{X_2}(x_2) \\ &= E_{X_2}(E(X_1|X_2)) \end{aligned}$$

The proof is as follows

$$\begin{aligned} E(X_1) &= \sum_{\text{all } x_1, x_2} x_1 p_{X_1X_2}(x_1, x_2) \\ &= \sum_{\text{all } x_1, x_2} x_1 p_{X_1|X_2}(x_1|x_2) p_{X_2}(x_2) \\ &= \sum_{\text{all } x_2} E(X_1|X_2 = x_2) p_{X_2}(x_2). \end{aligned}$$

The Covariance

The *covariance* of X_1 and X_2 is the expected value of the product of the X_1 deviation from its mean and the X_2 deviation from its mean

$$\sigma_{12} = Cov(X_1, X_2) = E[(X_1 - EX_1)(X_2 - EX_2)].$$

For $X_1 = X_2$ this becomes the variance.

If 'on average' when X_1 is above/below its mean and X_2 is above/below its mean, then the covariance will be positive.

Example *The values of p_{W_1, W_2} given earlier*

$W_1 \backslash W_2$	-2	-1	0	1	2	p_{W_1}
-1	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$
1	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
p_{W_2}	$\frac{1}{4}$	0	$\frac{1}{2}$	0	$\frac{1}{4}$	

We have $EW_1 = EW_2 = 0$.

$$\begin{aligned} & E[(W_1 - EW_1)(W_2 - EW_2)] \\ &= -1 \cdot -2 \cdot \frac{1}{4} + -1 \cdot 0 \cdot \frac{1}{4} \\ &\quad + 1 \cdot 0 \cdot \frac{1}{4} + 1 \cdot 2 \cdot \frac{1}{4} \\ &= 1 \end{aligned}$$

We could have anticipated that the covariance would be positive because when I make a good start, W_1 is above its mean (0) then I expect W_2 to be above its mean.

Sums of Random Variables

Often we want to know the distribution of the **sum** of a collection of random variables.

- *Total* earnings from a portfolio of several shares
- *Total* winnings after playing a game several times

The relationship between the distribution of the sum and the distribution of the

components is complicated but the relationship between the expectation of the sum and the expectation of the components is simple.

Theorem Let X_1, \dots, X_n be a collection of random variables with $EX_i = \mu_i$ then the expected value of $U = \sum X_i$ is given by

$$EU = \sum EX_i = \sum \mu_i$$

In words the expected value of a sum is the sum of the expected values.

Example If the n independent random variables X_1, \dots, X_n constitute a random sample from a distribution with mean μ then $\sum EX_i = n\mu$. So that

$$E\left(\frac{\sum X_i}{n}\right) = \mu. \text{ The sample mean } \bar{X} \text{ is}$$

an unbiased estimator of the population mean μ .

Theorem The variance of a sum is the sum of the variances plus twice the covariance.

$$\begin{aligned} \text{var}(X_1 + X_2) &= \text{var}(X_1) + \text{var}(X_2) \\ &\quad + 2\text{covar}(X_1, X_2). \end{aligned}$$

Remark When the random variables have zero covariance the variance of the sum is the sum of the variances

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2).$$

Remark When the variables are *independent* their covariance is 0 and so the variance of the sum of independent random variables is the sum of the variances.

Example If the n independent random variables X_1, \dots, X_n constitute a random sample from a distribution with mean μ and variance σ^2 then $\text{var}(\sum X_i) = n\sigma^2$ and $\text{var}\bar{X} = \frac{\sigma^2}{n}$.

Averages of many items

There are 2 important theorems

describing the behaviour of $\bar{X} = \frac{\sum X_i}{n}$ when n is large

- the law of large numbers

- the central limit theorem.
- I discuss them for the case of i.i.d. variables (random samples) but they hold more widely.
- An important special case is when the X_i are Bernoulli random variables and the sample mean is the proportion of successes in n trials.

Law of large numbers

Let X_1, \dots, X_n be a collection of independent random variables with $EX_i = \mu$ and $varX_i = \sigma^2$.

The (weak) law of large numbers states that for any given $a > 0$, as $n \rightarrow \infty$

$$P(|\bar{X} - \mu| < a) \rightarrow 1.$$

A proof can be based on the proposition that

$$var\bar{X} = \frac{\sigma^2}{n} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

- An alternative statement of the law is ‘the *probability limit* of \bar{X} is μ .’

- If the context is estimating μ then the law states that the sample mean \bar{X} is a *consistent* estimator of μ . The probability of being in error by a or more tends to 0 as n grows however small a is chosen.
- In the special case of Bernoulli random variables, the sample proportion is a *consistent* estimator of the probability of a success.
- More generally if $\hat{\theta}_n$ is an estimator of θ based on n observations and the probability limit of $\hat{\theta}_n$ is θ then $\hat{\theta}_n$ is a consistent estimator of θ .
- The weak law applies much more generally than to means. Typically some form of the law will apply to any estimator you meet in QM.

Central Limit Theorem.

Let X_1, \dots, X_n be independent and identically distributed random variables with common mean μ and variance σ^2 .

The quantity $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ has expectation 0

and variance 1.

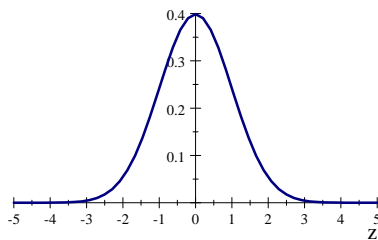
The *central limit theorem* states that when n is large the distribution of Z is approximately normal with mean 0 and variance 1.

- The CLT applies much more generally than to means. Typically some form of the theorem will apply to any estimator you meet in QM.

The probability density function of a normal random variable with mean 0 and variance 1 (called the *standard normal*) is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ for } -\infty < z < +\infty$$

which looks like this:



$N(0, 1)$ density

More generally a *normally distributed*

random variable X with mean μ and variance σ^2 has density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < +\infty$$

It is common to write $X \sim N(\mu, \sigma^2)$.

Normals reproducing

Suppose X_1, \dots, X_n are normal variables then a linear combination of them is also normal—the so-called **reproductive property** of the Normal Distribution.

Some examples show the diverse application of this property.

- In **Statistics**: if X_1, \dots, X_n are independent $N(\mu, \sigma^2)$ —written $X_i \sim IN(\mu, \sigma^2)$ —then

$$\sum X_i \sim N(n\mu, n\sigma^2)$$

$$\frac{1}{n} \sum X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- In **Portfolio Analysis**: an investor owns shares in two companies. Suppose the return on a share of the first company is R_1 and the return on a share of the

second is R_2 and the investor's portfolio has a_1 and a_2 of each kind of share then the return R on the portfolio is given by

$$R = a_1R_1 + a_2R_2$$

It is quite common in portfolio theory to assume that R_1 and R_2 are normally distributed say

$$R_1 \sim N(\mu_1, \sigma_1^2) \text{ and } R_2 \sim N(\mu_2, \sigma_2^2)$$

The reproductive property implies that R the total return is also normal:

$$R \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\sigma_{12})$$

Here σ_{12} is the covariance of R_1 and R_2 .

Exercise 3

1. Consider the random variables W_1 and W_2 introduced above.

a) Supply the missing calculations.

b) Work out the conditional distribution of W_2 given W_1 .

c) Are these variables independent?

2. Use the expected value of a sum theorem, i.e.

$$E(X + Y) = EX + EY,$$

to prove the following:

i) $E[(X - EX)(Y - EY)] = EXY - EXEY.$

ii) $cov(aX, bY) = abcov(X, Y).$

iii) $cov(X, (Y + a)) = cov(X, Y).$

3. Define the correlation, $corr(X, Y)$, between X and Y

$$corr(X, Y) = \frac{cov(X, Y)}{[\text{var}X \cdot \text{var}Y]^{\frac{1}{2}}}.$$

Use the results in question 2 to show that

$$\text{corr}(X, (Y + a)) = \text{corr}(X, Y)$$

$$\text{corr}(aX, bY) = \text{corr}(X, Y).$$

4. Consider the sequence of random variables $W_1, W_2, \dots, W_t, \dots$ defined by

$$W_t = \alpha + W_{t-1} + \varepsilon_t,$$

with $W_0 = 100$ and $E\varepsilon_t = 0$ for $t = 1, 2, \dots$

Show that $EW_t = \alpha + EW_{t-1}$ and

$$EW_t = 100 + \alpha t.$$

(This is an example of a random walk with drift. Such processes are important in Finance.)

Statistical inference SW ch. 3

In statistical inference we have data and wish to make inferences about the process generating the data.

In the simplest situation the data consists of a random sample from a probability distribution.

There are several kinds of inference.

- **Point Estimation:** given the data find a single number that is as good an approximation as possible to the unknown parameter value.
- **Interval Estimation:** given the data, find a range of values likely to contain the unknown parameter value.
- **Hypothesis testing:** given the data, assess the evidence in favour of a hypothesis about the unknown parameter value.
- **Forecasting:** given the data what single number or range of values is the best approximation to a future value from the

same distribution?

I will discuss estimation and testing in 2 situations; these will illustrate the main ideas.

1. Observations X_1, \dots, X_n comprise a random sample from a Bernoulli population with parameter π . We are making inferences about π .
2. Observations X_1, \dots, X_n comprise a random sample from a normal population with mean, μ , and variance σ^2 . We are making inferences about μ .

Inferences are based on functions of the observations, called *statistics*.

Our inferences will be based on the statistic, \bar{X} the sample mean.

We know a lot about its behaviour in these two situations.

1. If X_1, \dots, X_n are i.i.d. Bernoulli then \bar{X} (the sample proportion) is $N(\pi, \frac{\pi(1-\pi)}{n})$ in large samples (central limit theorem).

2. If X_1, \dots, X_n are i.i.d. $N(\mu, \sigma^2)$ then \bar{X} is $N(\mu, \frac{\sigma^2}{n})$ whatever the sample size.

The distribution of a statistic is sometimes called the *sampling distribution* of the statistic; thus the sampling distribution of \bar{X} is normal with mean μ and variance $\frac{\sigma^2}{n}$.

Point estimation

We have already noted that

1. Given a random sample X_1, \dots, X_n from a normal population with mean μ and variance σ^2 , then \bar{X} is an unbiased estimator and a consistent estimator of μ .
2. Given a random sample X_1, \dots, X_n from a Bernoulli population with parameter π , then \bar{X} (the sample proportion) is an unbiased estimator and a consistent estimator of π .

What is goodness for estimators?

- Unbiasedness means that there is no

systematic tendency to under- or over-estimate the value of the parameter.

- Consistency means that as the sample size tends to infinity the probability of making an error exceeding any given size tends to 0.
- Efficiency means that the variance of the estimator is smaller than that of the variance of any other unbiased estimator.

It can be shown that no other unbiased estimator of μ (or of π) has smaller variance than \bar{X} . So the estimator is called *efficient*.

There is a related result which does *not* depend on an assumed form for the population.

- Given a random sample X_1, \dots, X_n from a population with mean μ and variance σ^2 , then \bar{X} is an unbiased estimator of μ and more efficient than any other linear unbiased estimator.

Here is a sketch of the proof: a linear

estimator is of the form $\sum_1^n a_i X_i$ where the a_i are constants. The variance of this estimator is $\sigma^2 \sum_1^n a_i^2$. An unbiased

linear estimator must have $\sum_1^n a_i = 1$. To

minimise the variance of the linear estimator, the a_i must all equal $\frac{1}{n}$ and these are the weights that give the sample mean. (The details are in Exercise 2.)

How to find good estimators

There are several ways but I will describe two.

- *method of moments* based on the intuitively appealing principle of using sample moments to estimate quantities (parameters) that are functions of population moments.
- *maximum likelihood* based on the

intuitively appealing principle of choosing the value of the parameters that is most compatible with the observed sample.

Consider the problem of estimating the probability of a success in Bernoulli trials from a random sample of size n .

We know that for a Bernoulli random variable the first moment satisfies

$$EX = \pi$$

and so the method of moments estimator $\hat{\pi}$ based on the first sample moment is simply

$$\frac{\sum X_i}{n} = \hat{\pi}.$$

Finding the maximum likelihood estimator is more demanding. It involves expressing the probability of obtaining the sample obtained in terms of the value of the parameter and maximising that expression with respect to the parameter value.

Recall that the probability function of the

Bernoulli variable X takes only two values

$$p(0) = (1 - \pi); p(1) = \pi.$$

A neat way of writing p is

$$p(x) = \pi^x(1 - \pi)^{1-x}, x = 0, 1.$$

If we have a random sample x_1, \dots, x_n then the joint probability function is

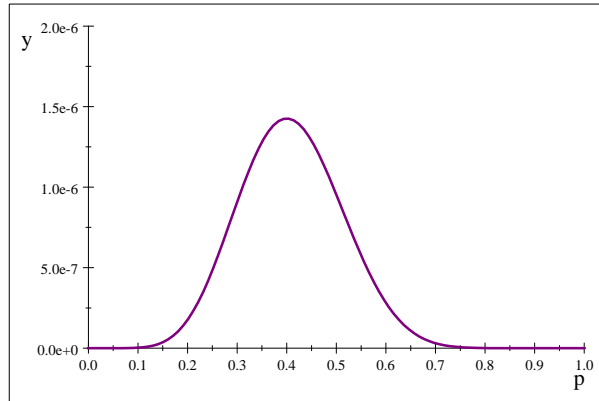
$$\begin{aligned} p(x_1, \dots, x_n; \pi) &= p(x_1; \pi) \cdot \dots \cdot p(x_n; \pi) \\ &= \pi^{x_1}(1 - \pi)^{1-x_1} \cdot \dots \cdot \pi^{x_n}(1 - \pi)^{1-x_n} \\ &= \pi^{\sum x_i}(1 - \pi)^{n - \sum x_i}. \end{aligned}$$

Interpreted as a function of π for given observations this called the *likelihood function* L .

If we have 8 successes in 20 trials then the likelihood function is

$$L(\pi; x) = \pi^8(1 - \pi)^{12}$$

and its plot looks like this



$$\pi^8(1 - \pi)^{12}$$

For finding the maximum it is convenient to take logs

$$\ln L(\pi; x) = \sum x_i \ln \pi + (n - \sum x_i) \ln(1 - \pi)$$

and find the maximum of this function which is in the same place as the maximum of the original function.

Differentiating and finding the maximum

$$\begin{aligned} \frac{d}{d\pi} \ln L(\pi; x) &= \frac{\sum x_i}{\pi} - \frac{(n - \sum x_i)}{(1 - \pi)} = 0 \\ \Rightarrow \hat{\pi} &= \frac{\sum x_i}{n} \end{aligned}$$

The maximum likelihood estimator $\hat{\pi}$ is the sample mean and is the same as the method of moments estimator.

As a second example, consider estimating the *two* parameters μ and σ^2 of a Normal distribution from a random sample of size n .

The derivation of the maximum likelihood estimators begins with the joint density of the observations

$$\begin{aligned} f_{X_1 \dots X_n}(x_1, \dots, x_n) &= f_{X_1}(x_1) \cdot \dots \cdot f_{X_n}(x_n) \\ &= f_X(x_1) \cdot \dots \cdot f_X(x_n) \\ &= \prod_{i=1}^n f_X(x_i) \end{aligned}$$

analogous to the case for discrete variables.

In the case of a random sample from $N(\mu, \sigma^2)$ the joint density is

$$\begin{aligned} f_{X_1 \dots X_n}(x_1, \dots, x_n; \mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum (x_i - \mu)^2}{2\sigma^2}}. \end{aligned}$$

If we now consider this as a function of

the parameters μ and σ^2 for given x we have the *likelihood* function

$$L(\mu, \sigma^2; x) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum (x_i - \mu)^2}{2\sigma^2}}$$

The method of maximum likelihood has us choose the value of β and σ^2 which maximises L .

Again it is convenient to take logs and maximise $\ln L$

$$\ln L(\mu, \sigma^2; x) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{\sum (x_i - \mu)^2}{2\sigma^2}$$

This function has a maximum at the same point as L but is easier to differentiate.

$$\frac{\partial \ln L}{\partial \mu} = \frac{\sum (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{n}{2\sigma^2} - \frac{\sum (x_i - \mu)^2}{2\sigma^4}$$

Setting the partials derivatives equal to 0 and using hats to denote the solutions

$$\frac{\partial \ln L}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum (x_i - \hat{\mu})^2}{n}.$$

So the sample mean is the maximum likelihood estimator of μ and the sample variance (with divisor n) the MLE of σ^2 .

There is a general theory of maximum likelihood estimation in which various *large sample* properties are developed:

- The maximum likelihood estimator is consistent (involving the law of large numbers)
- The maximum likelihood estimator is asymptotically normal (involving the central limit theorem)
- The maximum likelihood estimator is efficient.

Exercise 4

1. Suppose X_1, \dots, X_n are a random sample from a Bernoulli population with parameter π . Given that the variance of a Bernoulli variable with parameter π is $\pi(1 - \pi)$ show that the variance of $\sum X_i$ is $n\pi(1 - \pi)$ and the variance of \bar{X} is $\frac{\pi(1-\pi)}{n}$.

2. If \bar{X} has mean μ and variance σ^2 , then $\frac{\bar{X}-\mu}{\sigma}$ has mean 0 and variance 1. Prove this.

3. Given a random sample X_1, \dots, X_n from a population with mean μ and variance σ^2 and a linear estimator of μ denoted by $\tilde{\mu}$,

$$\tilde{\mu} = \sum_1^n a_i X_i$$

where the a_i are constants.

a) Show that the variance of $\tilde{\mu}$ is

$$\sigma^2 \sum_1^n a_i^2.$$

b) Calculate the variance for the special

cases: $a_1 = 1$ with all other $a_i = 0$;
 $a_1 = \frac{1}{3}$ and $a_n = \frac{2}{3}$ with all other $a_i = 0$;
all $a_i = \frac{1}{n}$.

c) Show that $E\tilde{\mu} = \mu$ if and only if

$$\sum_1^n a_i = 1.$$

d) Use the Lagrange multiplier method to show that putting all $a_i = \frac{1}{n}$ will minimise

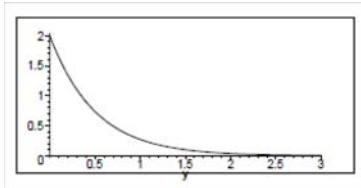
$\sigma^2 \sum_1^n a_i^2$ subject to the constraint that

$$\sum_1^n a_i = 1.$$

4. The exponential distribution is often used for modelling durations (e.g. the length of a spell of unemployment). It is a continuous distribution with density

$$f_X(x; \lambda) = \lambda e^{-\lambda x} \text{ for } 0 < x < +\infty.$$

The expected value of X is $1/\lambda$ and the density looks like this ($\lambda = 1$)



Show that the maximum likelihood estimator of λ given a random sample of size n is the reciprocal of the sample mean, $\frac{\sum X_i}{n}$.

Interval estimation & testing

The techniques of interval estimation and of hypothesis-testing are similar. So it is natural to treat them together.

Interval estimation

Mean of the Normal

Observations X_1, \dots, X_n are a random sample from a normal population with mean, μ , and variance σ^2 . We are making inferences about μ and, to begin with, we assume σ^2 is known.

From the reproductive property of the normal distribution we know that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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

One of the properties of the standard normal distribution is that for $Z \sim N(0, 1)$:

$$P(-1.96 < Z < +1.96) = 0.95.$$

We will adapt this form of statement to the problem of interval estimation.

The statement

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

can be reorganised as follows

$$P\left(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

The final line presents a 95% **confidence interval** for μ .

Each sample brings its own value of \bar{X} and the random interval

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) \text{ or } \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

has the property that with probability 0.95 it covers the mean μ .

In this example the **confidence**

coefficient or level of confidence is 95% but intervals with different confidence coefficients are easily constructed.

The 90% and 99% confidence intervals are respectively

$$\left(\bar{X} - 1.65 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.65 \frac{\sigma}{\sqrt{n}}\right)$$

$$\left(\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}\right)$$

The narrower the interval the bigger the chance that the interval will miss μ .

There is a trade-off between a narrow interval and a high level of confidence.

In practice the value of σ^2 is rarely known and an interval based on an estimate is used. The usual estimator is

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

which is an unbiased estimator of σ^2 .

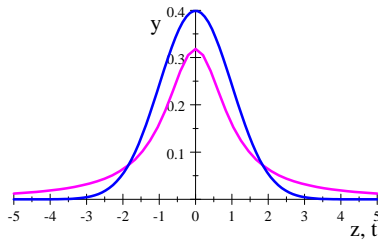
Plugging in the estimate s changes the distribution from the standard normal to *Student's t-distribution*:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t^{(n-1)}.$$

The form was conjectured by W. S. Gosset who used the pen name 'Student'.

The distribution depends on the sample size n through the parameter $n - 1$ called the *number of degrees of freedom*. One degree of freedom is lost because the mean has to be estimated to obtain s^2 .

For large n the distribution looks like the standard normal but for small n it has much *fatter* tails—i.e. the density does not fall away as quickly as the normal.



$N(0, 1)$ and $t^{(1)}$

Incidentally fatter—than normal—tails are common in **financial** data and the analysis of such data requires special

handling. There is a numerical measure of heavy tailedness of a distribution called a **kurtosis measure** based on comparing $E(Y - \mu)^2$ with $E(Y - \mu)^4$.

Some comparisons of 95% intervals for different sample sizes:

$$P\left(\bar{X} - 2.23 \frac{s}{\sqrt{10}} < \mu < \bar{X} + 2.23 \frac{s}{\sqrt{10}}\right)$$

$$P\left(\bar{X} - 2.04 \frac{s}{\sqrt{30}} < \mu < \bar{X} + 2.04 \frac{s}{\sqrt{30}}\right)$$

$$P\left(\bar{X} - 1.98 \frac{s}{\sqrt{120}} < \mu < \bar{X} + 1.98 \frac{s}{\sqrt{120}}\right).$$

There is not much error in using the normal value of 1.96 in realistic size samples.

Bernoulli and large sample methods

Exact confidence intervals for the Bernoulli distribution are available but it is more usual to base them on the large sample normal approximation. For many situations in econometrics there are no tractable exact distributions and so a

large sample argument like the following is resorted to.

The normal approximation to the binomial (CLT) states that the sample proportion $\hat{\pi}$

$$\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1)$$

in large samples.

Reasoning as before we can produce a large sample 95% interval

$$\left(\hat{\pi} - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\pi(1-\pi)}{n}} \right).$$

BUT this is not operational for π is unknown.

However it is also true that in large samples

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \sim N(0, 1)$$

and so there is a large sample 95% interval

$$\left(\hat{\pi} - 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right).$$

There is a very nice exact theory extending Student's work to regression but in practical empirical studies the confidence intervals are either constructed using large sample approximations or using a simulation technique called *bootstrapping*.

Hypothesis testing

Mean of the Normal

As before, observations X_1, \dots, X_n are a random sample from a normal population with mean, μ , and variance σ^2 . We assume σ^2 is known.

We know that

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

and we will be using this fact to test hypotheses about μ .

In the simplest situation there are two hypotheses, a null hypothesis, H_0 , and an alternative hypothesis, H_A , to which there are two responses—accept or reject the null. There are two possible states of the world— H_0 is true and H_A is true.

There are four possible combinations of conclusion and states of the world: two of the combinations represent erroneous conclusions.

	H_0 is true	H_A is true
Accept H_0	✓	Type II error
Accept H_A	Type I error	✓

Just as we cannot expect infallible estimation from probabilistic material we cannot expect the probability of the two kinds of error to be both zero.

Examples of null and alternative

a) $H_0 : \mu = \mu_0$ versus $H_A : \mu = \mu_A$ where μ_0 and μ_A are specified numbers.

b) $H_0 : \mu = \mu_0$ versus $H_A : \mu > \mu_0$ where μ_0 is a specified number.

c) $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ where μ_0 is a specified number.

Examples

We concentrate on the simplest situation, an example of (a).

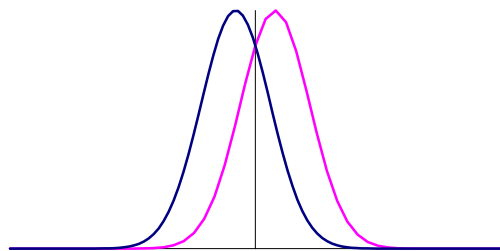
Suppose $\mu_0 = 12$, $\mu_A = 11$, $n = 10$, $\sigma^2 = 40$ then translating this information into hypotheses about the distribution of \bar{X} we have

$$H_0 : \bar{X} \sim N(12,4), H_A : \bar{X} \sim N(11,4).$$

There is a theory of optimal tests but I won't go into it. It justifies focussing on \bar{X} and following the kind of procedure I am about to describe.

The obvious procedure is to interpret a large value of \bar{X} as evidence for H_0 and a low value as evidence for H_A . Suppose we set the *cut-off* at 11.5.

The figure shows the two alternative distributions for \bar{X} . with the cutoff at 11.5.



H_0 (12) versus H_A (11)

Errors occur because sometimes a high value of \bar{X} will come from the distribution on the left and a low value from the distribution on the right. In the case illustrated the test is pretty useless because the distributions overlap so

much. The observations are too noisy and the sample too small for very good discrimination.

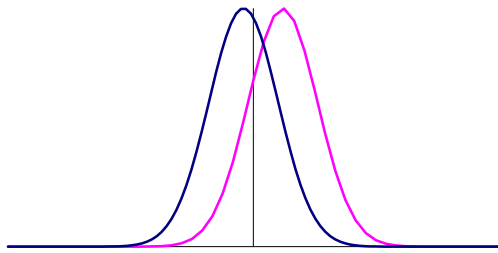
- A **Type I error** occurs when a small value (< 11.5) of \bar{X} is obtained from the H_0 distribution, the $N(12,4)$ distribution.
- A **Type II error** occurs when a large value (> 11.5) of \bar{X} is obtained from the H_A distribution, $N(11,4)$.
- The probabilities are found to be

$$P_{H_0}(\bar{X} < 11.5) = P\left(Z < \frac{11.5 - 12}{\sqrt{4}}\right) = 0.411$$

$$P_{H_A}(\bar{X} > 11.5) = P\left(Z > \frac{11.5 - 11}{\sqrt{4}}\right) = 0.411$$

The test is not a complete waste of time for the probabilities are less than 0.5.

- If the cut-off point is moved to the left the probability of Type I error will be reduced but the probability of Type II error will be increased.



Reduced cutoff: Type I reduced

- We would like both probabilities to be small but there is a trade-off between them.
- The **probability of Type I error** is often called the **significance level** or **size** of the test. Thus the significance level of this particular test is about 41%.
- Sometimes $1 - P(\text{Type II error})$ is called the **power** of the test. The power of this particular test is about 59%.

In testing there are two common ways to proceeding

- Choose a pre-specified significance level such 5% and let that determine the cut-off point defining the rejection region. To find the cut-off value we use the property of the standard normal that $P(Z < -2.57) = 0.05$. The cut-off is

7.86 because

$$0.05 = P\left(Z < \frac{7.86 - 12}{\sqrt{4}}\right) = P_{H_0}(\bar{X} < 7.86).$$

Of course 7.86 is also in the left tail of the H_A distribution and the probability of Type II error will be very large

$$\begin{aligned} P_{H_A}(\bar{X} > 7.86) &= P\left(Z > \frac{7.86 - 11}{\sqrt{4}}\right) \\ &= P(Z > -1.57) = 0.94. \end{aligned}$$

The procedure is that when we have an observed \bar{X} of 9, say, we register that it exceeds the cutoff of 7.86 and say that it is *not* significantly different from 12 at the 5% level.

- Take the observed value of \bar{X} and report the significance level at which it would it would just count against the null hypothesis; this is called the *p-value*. For example if we observed the value is 9 we compute

$$P_{H_0}(\bar{X} < 9) = P\left(Z < \frac{9 - 12}{\sqrt{4}}\right) = 0.0667.$$

With 9 as the value of \bar{X} , H_0 would be rejected at all significance levels greater

than or equal to 6.67%. Statistical packages often report the *p-value* as well as whether the hypothesis is rejected at a prespecified significance level, such as 5% or 1%.

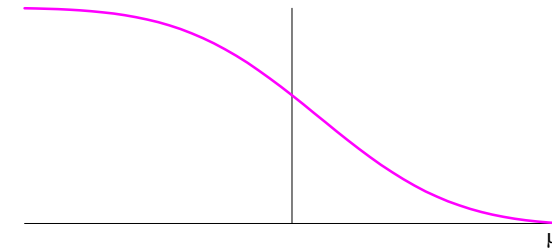
More complicated situations

The researcher seldom has a specific alternative value in mind and it is more common to consider alternatives like $H_A : \mu \neq \mu_0$, $H_A : \mu < \mu_0$ or $H_A : \mu > \mu_0$. Zero is a particularly common choice for the null hypothesis value for it often corresponds to **no** effect and there is rarely a stand-out alternative value.

Let's focus on the case $H_A : \mu > \mu_0$. The probability of Type II error is no longer a single number—as in the simple example. There will be a probability of Type II error for every possible value of the parameter $\mu (> \mu_0)$.

The error performance of the test under the alternative hypothesis can be summed up in the **power function** of the

test. This is the probability of *not* making a Type II error and for the pair $\mu = \mu_0$ versus $\mu < \mu_0$ the power function looks like:



Power: null 12 and cut-off 11.5.

For values *close* to the null hypothesis value (12) the power of the test will be approximately equal to the significance level. For values much less than 12 false acceptance of H_0 will be unlikely and the power will be nearly 1 (i.e. the chance of a Type II error will be almost 0).

We won't go through the various possible situations as we did with confidence intervals

- Small sample situations in which tests on the mean of the normal distribution require the use of the *t*-distribution.

- Large sample tests on the parameter of the Bernoulli distribution using the normal approximation
- Other forms of H_A , including $\mu > \mu_0$ and $\mu \neq \mu_0$. These alternatives require different rejection regions, the other tail or both tails. See SW.

Sampling distributions

In the analysis of any statistical problem the distributions involved fall into two classes

- the distribution(s) appearing as part of the specification of the population/process/model–‘primitive’ distributions;
- the distributions of the various estimators and test statistics calculated for making inferences about the parameters of the model–these *sampling distributions* are derived from the primitive distributions.

Some important sampling distributions

The normal distribution is the most important.

It appears in the following situations, reflecting the reproductivity of the normal distribution and the CLT.

- As the **exact** distribution of \bar{X} and of $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ when we have a random sample from a normal population with mean μ and variance σ^2 .
- As an **approximate large sample distribution** of \bar{X} and of $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ when we have a random sample from any population with mean μ and variance σ^2 .
- As an **approximate large sample distribution** of any average. Thus $\frac{\sum X_i^k}{n}$ from a random sample will be normally distributed in large samples.

Remark *Because of the central limit theorem estimators will often have a limiting normal form even though the primitive distribution is not normal.*

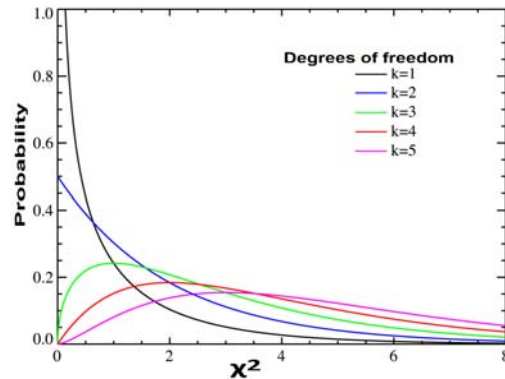
The next most important sampling distribution is χ^2 (*chi-squared*).

Consider Z_1, \dots, Z_k independent $N(0, 1)$

distributed random variables then

$$Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi_k^2$$

a *chi-squared random variable with k degrees of freedom*. The densities look like this.



- This appears as an **exact distribution** in connection with the distribution of the sample variance of a normal random sample.

It is clear that

$$\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \dots + \left(\frac{X_n - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

Not so clear but still true is

$$\left(\frac{X_1 - \bar{X}}{\sigma}\right)^2 + \dots + \left(\frac{X_n - \bar{X}}{\sigma}\right)^2 \sim \chi_{n-1}^2.$$

From which it follows that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

One degree of freedom has been lost because one parameter μ has been replaced by one estimated quantity \bar{X} .

- We often have to deal with the squares of large sample normal variables and that involves us with χ^2 .

3. We met Student's t when constructing confidence intervals or testing hypotheses on μ the mean of the normal distribution. We used the fact that for a random sample from a $N(\mu, \sigma^2)$ population

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t^{(n-1)}.$$

If we rewrite the ratio as

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}}$$

we notice that the numerator is $N(0, 1)$

and the denominator is the square root of a χ^2 divided by its number of degrees of freedom. The numerator and denominator are independent.

This suggests a characterisation of the t -distribution. Suppose Z and U are independent random variables with Z standard normal and U a χ^2 with k degrees of freedom, then the random variable

$$X = \frac{Z}{\sqrt{\frac{U}{k}}}$$

is said to be distributed as t with k degrees of freedom.

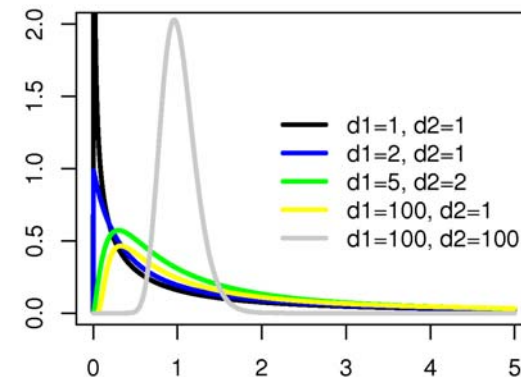
The final common sampling distribution related to the normal distribution is the F distribution which can be thought of as generalising the χ^2 and t .

Suppose U and V are independent χ^2 random variables with d_1 and d_2 degrees of freedom respectively, then the random variable

$$X = \frac{\frac{U}{d_1}}{\frac{V}{d_2}} \sim F(d_1, d_2)$$

is said to be distributed as F with d_1 and d_2 degrees of freedom.

The densities look like this



- The F -distribution can be regarded as a generalisation of the t -distribution for squaring a t with k degrees of freedom gives an $F(1, k)$.
- The F -distribution can also be thought of as a modification of the χ^2 in the same way as the t is a modification of the normal. When $d_2 \rightarrow \infty$ the denominator of the ratio converges in probability to 1

and the ratio, X , converges to $\frac{1}{d_1}$ times a $\chi_{d_1}^2$.

Linear equations, vectors & matrices

I will just make a few comments on this topic. It is very important in Econometrics I & II but not elsewhere: there is a very brief treatment in Appendix 16.1 of Stock and Watson and a more thorough account in W. H. Greene's *Econometric Analysis*. There are books on the subject.

I apologise for the pictures—most are unfinished and I will add finishing touches in the lectures.

Linear equations

Linear equations appear everywhere and you should know some of the basic ideas. I introduce them for the simple case of 2 equations in 2 unknowns

Ex. 1

Consider the following 2 linear equations in the unknowns x_1 and x_2 :

$$x_1 + x_2 = 0$$

$$x_1 + 2x_2 = 1$$

It is easy to see that the equations have a unique solution $x_1 =$ and $x_2 =$.

Plotting the 2 straight lines we see they intersect at this point.

Ex. 2

Consider now

$$x_1 + x_2 = 0$$

$$2x_1 + 2x_2 = 2$$

It is easy to see that this pair of equations has no solution: the equations are contradictory—the first says that x_1 and x_2 sum to 0 and the second says that they sum to 1.

If we plot the equations we have parallel straight lines that do not intersect.

Ex. 3

Change the intercepts in the equations of Ex 2 to obtain

$$x_1 + x_2 = 1$$

$$2x_1 + 2x_2 = 2.$$

We find that there are infinitely many solutions. Both equations say that x_1 and x_2 sum to 1. If we plot the two equations, the two lines coincide.

These simple examples of 2 equations with 2 unknowns illustrate the 3 possibilities that apply to all systems of linear equations:

- unique solution
- no solution
- infinitely many solutions.

For dealing with the general case it is useful to have some terms and symbols.

The 4 coefficients are known collectively as a matrix.

Two matrices were involved in the examples:

$$\text{Ex 1: } \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}; \text{ Ex 2 and 3: } \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}.$$

These matrices each have 2 rows and 2 columns.

We can also consider the variables x_1 and x_2 together, writing them either as a 1 row matrix or as a 1 column matrix

- (x_1, x_2) also called a row vector
- or $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ also called a column vector.

The final component of the equations, the 2 intercepts, can also be written as single entities.

The intercepts in Ex 1 can be combined into the column vector $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Look again at Ex 1

$$x_1 + x_2 = 0$$

$$x_1 + 2x_2 = 1$$

and rewrite it using the matrix/vector symbols as

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

where we think of the matrix of coefficients as multiplying the vector of x 's to get the vector of constants.

Implicit in this representation are certain rules for multiplying matrices and saying when they are equal. I won't spell them out.

All 3 sets of equations can be written in the form

$$Ax = b$$

where Ex 2 and 3 have the same A , coefficient matrix, but different b 's, constant vectors.

There are 2 useful ways of thinking about matrices—algebraically and geometrically.

Algebra of matrices

Consider the analogy between $Ax = b$ and an equation in a single variable u , say

$$4u = 3.$$

We can solve for u to obtain

$$u = \frac{3}{4}.$$

Equations in one variable

$$au = b$$

always have a unique solution

$$u = a^{-1}b.$$

(compare Ex 1) EXCEPT when $a = 0$
when there are 2 possibilities

$$0u = b (\neq 0) \text{ and } 0u = 0$$

where in the first case there is no solution
(cf. Ex 2) and in the second there are
infinitely many (cf. Ex 3).

In Ex 1 we can invert the matrix A and
write

$$Ax = b \Rightarrow x = A^{-1}b.$$

but we cannot do the same for the
coefficient matrix of Ex 2 and 3. That
matrix does not have an inverse: it's
called a singular matrix. Invertible
matrices are called non-singular

matrices.

In numbers (one row/one column
matrices) there is only singular number,
viz. 0. In squares matrices (of order 2 or
more) there are infinitely many.

The solution to a pair of equations $Ax = b$

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

(if there is one) is given by

$$x_1 = \frac{b_1 a_{22} - a_{12} b_2}{a_{11} a_{22} - a_{12} a_{21}} : x_2 = \frac{a_{11} b_2 - a_{21} b_1}{a_{11} a_{22} - a_{12} a_{21}}$$

The distinctive terms that appear in the
numerator and denominator are called
determinants.

Note the same determinant appears on
the bottom of both expressions and there
is a problem when

$$a_{11} a_{22} - a_{12} a_{21} = 0$$

for the ratios are not defined.

This is the singular case: in Ex 2

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$$

$$a_{11}a_{22} - a_{12}a_{21} = 1 \cdot 2 - 1 \cdot 2 = 0.$$

The 'problem' with the matrix

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$$

is that the rows are not independent: the second row is a multiple of the first.

Lack of independence can be less obvious in higher-order matrices: consider the 3×3 matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 3 & 2 & 3 \end{pmatrix}$$

that I have constructed by forming the 3rd row by summing the second row and twice the first row.

The equation system

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 3 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

has infinitely many solutions because the elements of b satisfy the same condition as the rows of A . Although there are 3 equations, one is redundant and can be derived from the other two—so the 2 equations in 3 unknowns do not pin down the 3 unknowns to a unique value.

The equation system

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 3 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

has no solutions.

The number of independent rows is called the rank of the matrix.

- $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ has rank 2.
- $\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$ has rank 1.
- $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 3 & 2 & 3 \end{pmatrix}$ has rank 2.

To be invertible a matrix must have “full” rank.

Geometry of linear transformations

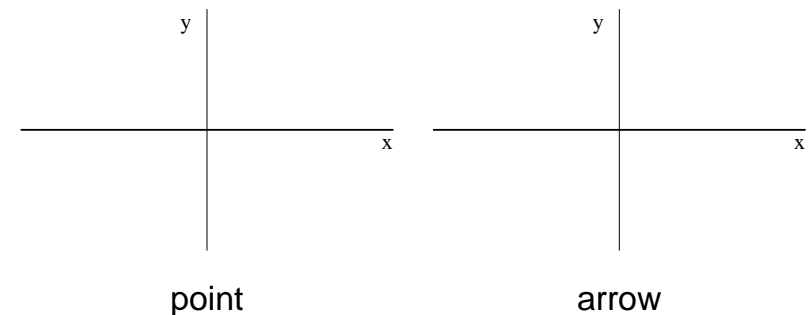
The equation

$$Ax = b$$

can be considered geometrically.

A point x in 2-dimensional space is transformed into another point in 2-dimensional space, b , by the action of A .

The vector, as an ordered collection of numbers, can be interpreted geometrically as a point in space or the tip of an arrow



The geometric interpretation of the matrices in Ex 1 and 2 is messy and so I use cleaner examples.

Ex 4

The nonsingular matrix

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

rotates vectors through 45° anti-clockwise. The 45° vector $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ is changed into $(0, 1)$.

This matrix is invertible because there is matrix representing a rotation 45° clockwise that reverses the action of A .

If we start at x and go to b then a clockwise rotation takes us back to where we started.

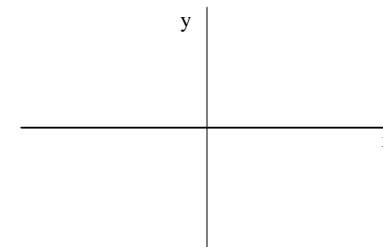
$$Ax = b \Rightarrow x = A^{-1}b.$$

Ex 4

The singular matrix

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

transforms arbitrary points into points on the 45° line, viz. (x_1, x_2) becomes $(\frac{x_1+x_2}{2}, \frac{x_1+x_2}{2})$.



averaging

In this case if b is on the 45° line then x can be any of infinitely vectors. If b is NOT on the 45° line then NO x could have been the starting value.

Exercise 5

1. X_1, \dots, X_n are a random sample from a normal population with mean, μ , and variance σ^2 .

We found that $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ provides a 95% confidence interval for μ .

The 99% interval $\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$ is wider.

How many more observations are needed if we want to make the 99% interval as narrow as the 95% interval based on n observations?

2. Calculate the 95% large sample confidence interval for π when a) $\hat{\pi}$ (sample proportion) is 0.4 and $n = 100$; b) $\hat{\pi}$ is 0.4 and $n = 400$; c) $\hat{\pi}$ is 0.4 and $n = 1600$.

3. Consider the simple testing situation again: $\mu_0 = 12$, $\mu_A = 11$, $n = 10$, $\sigma^2 = 40$. Foolishly I decide to *reject* H_0 when \bar{X}

exceeds 11.5 what is the probability of Type I error and the probability of Type II error?

A bored student decides to ignore the data and toss a coin: if it lands Heads, he rejects H_0 ; if it lands Tails, he accepts H_0 . What is the probability of Type I error and the probability of Type II error of his procedure?

4. Consider the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Describe the geometric action of A . Is A singular? You can reason from the geometric action or calculate the determinant.

5. If we apply the matrix A twice, it is natural to write A^2 for this action. Consider the rotation matrix

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

What does A^4 do; what does A^8 do?

6. Suppose an economy changes from state x_{t-1} at time $t-1$ to state x_t at time t according to

$$x_t = Ax_{t-1}.$$

Show that

$$x_t = A^t x_0$$

where x_0 is the initial state.

7. Some things are **different** about matrix multiplication: contrast the effects of

- first multiplying by A and then by B (resulting in the operation BA)
- first multiplying by B and then by A (resulting in the operation AB).

where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$