I will be presenting some basic ideas on statistics to support Quantitative Methods and the various Econometrics courses. Ideas about probability are also used in decision making under uncertainty, a topic in several Economics and Finance courses.

I will follow chapters 1-3 of Stock & Watson’s *Introduction to Econometrics*, (2nd edition) the textbook for Quantitative Methods.

Books are devoted to these 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*.
If there is time I will say something about linear algebra.

**Economic questions & data**

“At a broad level, econometrics is the science and art of using economic theory and statistical techniques to analyze economic data.” SW3.

One of the questions SW consider (#2) is whether there is racial discrimination in the market for home loans.

Researchers at the Federal Reserve Bank of Boston found that 28% of black applicants were denied mortgages while only 9% of white were denied. (The details are given in chapter 11 and Appendix 11.1.)

The researchers used **cross-sectional**
data, i.e. data on different entities for a single time period. The entities were applications for mortgages and the observations comprised the fate of the application, the race of the applicant and other characteristics of the applicant that “economic theory” might suggest are relevant to the success of the application—such as ability to repay.

An analogous question #2* takes off from the observation that on average blacks earn less than whites—is this the result of racial discrimination in the labour market? (SW do not discuss this question but on 36-7 they consider differences in earnings between men and women and those with and without a college degree.)

**Probability models**

The statistical methods in SW involve specifying a probability mechanism that has generated the observations.

A very simple mechanism for modelling the binary outcome, denial or
acceptance, is the Bernoulli distribution (SW 20-21).

The Bernoulli or indicator random variable, \( X \), is based on a coding for the occurrence or non-occurrence of an event \( E \): 1 for occurrence (or success or denial here), 0 for non-occurrence (failure, i.e. acceptance).

A probability \( \pi \) (the Greek letter corresponding to \( p \)) is assigned to occurrence so that the random variable \( X \) takes two values with the following probabilities:

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

For brevity we write \( X \sim B(\pi) \).

Standard examples include tossing a coin where success is Heads and rolling a dice where success is an ace (1) and failure is one of the numbers 2,3,4,5,6. With a regular (‘fair’) coin, \( \pi = \frac{1}{2} \) and with a fair dice \( \pi = \frac{1}{6} \).
The function assigning probabilities to the values taken by the random variable is called the **probability distribution**. Write \( p(x) \) for the probability distribution and thus \( p(x) = \Pr(X = x) \) for all values of \( x \).

The probability distribution for the Bernoulli rv is

\[
p(0) = 1 - \pi \quad \text{and} \quad p(1) = \pi
\]

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

The Bernoulli is a discrete rv, i.e. one that takes a finite or countable number of values.

**The denial/acceptance process**

a) The simplest specification: assume that everybody has the same \( \pi \).

b) Complicate things by assuming that there is common value for the blacks, \( \pi_B \), and a (different) common value for the whites, \( \pi_W \).

c) Complicate things still more by assuming that each person has a \( \pi \)
reflecting that person’s race, ability to repay and other personal characteristics. I will discuss a) and b) while c) is treated in SW ch 11 and QM.

**Discrimination over wages**
We can treat the wages issue in the same way.

We model \( W \) as a continuous random variable defined for positive values. A continuous rv takes on a continuum of values: see p. 37 for some possible shapes.

As analogues of a), b) and c), consider

a*) All persons have the same \( W \) distribution.

b*) There is a \( W \)-distribution for the blacks and a different one for the whites.

c*) Each person has a \( W \)-distribution reflecting that person’s race, education, experience and other potentially relevant factors.

I will consider versions of a* and b*. QM
considers versions of c*.

A possible distribution for $W$, a non-negative rv, is the log-normal distribution, i.e. the distribution of $W$ which implies that $\ln W$ is normally distributed.

If $Y = \ln W$ then the probability density of $Y$ is given by

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

with parameters $\mu$ and $\sigma^2$. It is common to write $Y \sim N(\mu, \sigma^2)$.

The case $Z \sim N(0, 1)$ is called the \textbf{standard normal} with density

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

which looks like this:
A useful property of the $N(0,1)$ distribution is that

$$\Pr(-1.96 < Z < +1.96) = 0.95.$$  

For a continuous rv the probability that the rv lies between $a$ and $b$ is given by the integral of the density between $a$ and $b$. The area under the $N(0,1)$ density between $-1.96$ and $+1.96$ is 0.95. The probability that the variable takes any single value is 0.

**Interpreting the parameters $\pi$, $\mu$ and $\sigma^2$**

The Bernoulli distribution has one parameter, $\pi$, which gives the probability of 1, a ‘success.’

The value of $\pi$ controls the average or
The expected value or expectation or mean of a discrete random variable $X$, denoted by $E(X)$ is defined as

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

$E(X)$ may be abbreviated to $EX$.

- 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.

The Bernoulli/indicator random variable $X$ has expected value

$$EX = 0 \cdot (1 - \pi) + 1 \cdot \pi$$

$$= \pi.$$ 

So the parameter $\pi$ is the mean of the distribution.

The expected value, expectation or mean 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 \( X \).

If \( Y \sim N(\mu, \sigma^2) \)

\[ EY = \int_{-\infty}^{\infty} y \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \mu. \]

The parameter \( \mu \) is the mean of the distribution.

**Other expectations**

The **variance** of a random variable 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 $\iff$ high risk.

**Example:** $X$ is a Bernoulli with probability $\frac{1}{2}$. We already know that $EX = \frac{1}{2}$. From the definition of the variance

$$
\sigma_X^2 = E([X - EX]^2) = \sum_{\text{all values of } x} (x - \frac{1}{2})^2 p(x)
$$

$$
= \left(0 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2}
$$

$$
= \frac{1}{4}.
$$

For the Bernoulli with parameter $\pi$, the variance is $\pi(1 - \pi)$.

For the normal random variable the variance is
\[ \sigma_Y^2 = \int_{-\infty}^{\infty} (y - \mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy = \sigma^2. \]

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.

SW 26-29 describes some moment based measures.

The normal distribution is symmetric about its mean but the Bernoulli is not symmetric about its mean unless \( \pi = \frac{1}{2} \).

One measure of symmetry is **skewness** defined by

\[
Skewness = \frac{E(Y - EY)^3}{\sigma_Y^3}
\]

The \( \sigma_Y^3 \) (the cube of the standard deviation) in the denominator ensures
that the measure is free of the units in which $Y$ is measured. The normal distribution has skewness equal to 0 because the third (like all odd moments about the mean) is 0.

**Kurtosis** is a measure of the heaviness of the tails

$$Kurtosis = \frac{E(Y - EY)^4}{\sigma_Y^4}$$

For a normal distribution the kurtosis is 3. A distribution with kurtosis greater than 3 has ‘heavier tails’—more mass given to extreme values—than the normal.

**Expectations in Finance**

Consider a decision maker or gambler choosing between two opportunities one with gain, $G$, with distribution

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

and the other with gain $G^*$,
\[ G^* = \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ 3 \text{ with probability } \frac{1}{2} \end{cases} \]

Check that \( E(G) = E(G^*) = 2 \) and \( \text{var}(G) = 4 \) and \( \text{var}(G^*) = 1 \).

The opportunities have the same expected gain but as \( G^* \) is less risky we might prefer it.

Another approach to decision-making is based on **expected utility**. Assume the utility function \( u(g) = \sqrt{g} \) and consider the expected utility of \( G \) and \( G^* \).

\[
Eu(G) = \sqrt{0} \cdot \frac{1}{2} + \sqrt{4} \cdot \frac{1}{2} = 1
\]

\[
Eu(G^*) = \sqrt{1} \cdot \frac{1}{2} + \sqrt{3} \cdot \frac{1}{2} > 1.
\]

An individual who maximises expected utility will prefer \( G^* \).

Consider a third gamble, a sure thing of 2 which has expected value 2, like \( G \) and \( G^* \). Both \( G \) and \( G^* \) generate less expected utility (than \( \sqrt{2} \)) and are less attractive than the sure thing. The square root function is concave and people with
concave utility functions (risk averters) prefer a sure thing to a gamble with the same expected value.

Exercise 1

0. Do the calculations I have left out and check that my results are correct.

1. Basic facts about expected values include

\[ E(a) = a; \quad E(aX) = aEX; \quad E(X + a) = a + EX. \]

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) \[ \text{var}(aX) = E[(aX - E(aX))^2] = a^2 E(X - EX)^2 = \]

iv) \[ \text{var}(X + a) = E[(X + a - E(X + a))^2] = \text{var}X. \]

2. Use the propositions in question 1 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?

3. Calculate the variance, skewness and kurtosis of the $B(\pi)$ random variable.

4. Repeat my discussion of the choice between $G, G^*$ and the sure thing assuming the convex utility function $u(g) = g^2$.

If you want more practice, do the starred exercises in SW.

**Independence & random samples**

Later I show how to extend these concepts to the case of 2 or more random variables. (Also see SW 29-35.)

For the moment I need one multivariate
concept.

Two random variables $X_1$ and $X_2$ are \textbf{independent} if

$$\Pr(X_1 = x_1 \text{ and } X_2 = x_2) = \Pr(X_1 = x_1) \cdot \Pr(X_2 = x_2)$$

for all $x_1$ and $x_2$.

The definition extends to an arbitrary number of variables. Thus suppose $x_1, \ldots, x_n$ are $n$ independent draws from $B(p)$ then the joint probability function is

$$p(x_1, \ldots, x_n; \pi) = p(x_1; \pi) \cdot \ldots \cdot p(x_n; \pi)$$

$$= \pi^{x_1} (1 - \pi)^{1 - x_1} \cdot \ldots \cdot \pi^{x_n} (1 - \pi)^{1 - x_n}$$

$$= \pi^{\sum x_i} (1 - \pi)^{n - \sum x_i}.$$

Note that $\sum x_i$ is the number of ones (successes) and $n - \sum x_i$ the number of zeros (failures) in $n$ trials.

A collection of $n$ independent random variables $X_1, \ldots, X_n$ each with the same distribution is called a \textbf{random sample} of size $n$ from that distribution.

A standard starting assumption in statistical analysis is that the observations are a random sample from
some distribution (like the Bernoulli). **Experiments** are often designed so that the observations form a random sample. (SW 10)

But for the **observational data** used by economists the assumption of a random sample is an assumption and it may be unrealistic.

### Sums of Random Variables

Questions #2 and #2* began with facts about averages—a proportion is a kind of average. An average is a sum divided by the number of items. So we have an interest in the distribution of the **sum** of a collection of random variables.

The relationship between the distribution of the sum and the distribution of the components is often complicated but a couple of cases should be noted.

Let $X_1, \ldots, X_n$ be a random sample from $B(\pi)$ then $\sum X_i = Y$ has the binomial distribution with parameters $\pi$ and $n$. 
If $Y_1, \ldots, Y_n$ are normal variables then a linear combination of them is also normal—the so-called **reproductive property** of the normal distribution.

Consider the case with two independent normals: if $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ then

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

**Expectation of a sum**

The relationship between the expectation of the sum and the expectation of the components is very simple.

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

$$E\left(\sum X_i\right) = \sum EX_i = \sum \mu_i$$

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

**Example** If the $n$ independent random variables
variables $X_1, \ldots, X_n$ constitute a random sample from a distribution with mean $\mu$ then $\sum EX_i = n\mu$. So that $E\left(\frac{\sum X_i}{n}\right) = \mu$.

**Theorem** The variance of a sum of independent variables is the sum of the variances

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

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

**Exercise 2**

The following questions use material from Ex 1 and the expectation of a sum theorems.

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

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

Averages of many items

SW48-55 discuss 2 important theorems describing the behaviour of $\overline{X} = \frac{\sum X_i}{n}$ when $n$ is large

- the law of large numbers
- the central limit theorem.

A 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

The primitive form of the law of large numbers applies to a fair coin: when we toss the coin a large number of times the proportion of Heads will probably be close to $\frac{1}{2}$. 
More generally let $X_1, \ldots, X_n$ 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 $n \to \infty$

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

In words, ‘the probability limit of $\bar{X}$ is $\mu$’ or $\bar{X}$ converges in probability to $\mu$’ in symbols $p \lim \bar{X} = \mu$.

A proof can be based on the proposition that

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

**Central Limit Theorem**

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with common mean $\mu$ and variance $\sigma^2$.

The quantity $Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}}$ 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 original form of the CLT was the normal approximation to the binomial. See SW 53.

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

**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 to illustrate the main ideas.

1. Observations $X_1, \ldots, X_n$ comprise a random sample from a Bernoulli population with parameter $\pi$. We are making inferences about $\pi$.

2. Observations $X_1, \ldots, X_n$ comprise a random sample from a normal population with mean, $\mu$, and variance $\sigma^2$. We are making inferences about $\mu$.

SW’s question #2 about racial discrimination in the loans market was a hypothesis testing question based on (1) and involving 2 samples: does $\pi_B = \pi_W$?
#2* can be cast as a 2-sample normal question.

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

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

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

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

2. If \( X_1, \ldots, X_n \) are i.i.d. \( N(\mu, \sigma^2) \) then \( \overline{X} \) is \( N(\mu, \frac{\sigma^2}{n}) \) whatever the sample size. (Reproducibility of normals)

The distribution of a statistic is sometimes called the sampling distribution of the statistic; thus the sampling distribution of \( \overline{X} \) is normal with mean \( \mu \) and variance \( \frac{\sigma^2}{n} \).

**Point estimation**

We have already noted that
Given a random sample $X_1, \ldots, X_n$ from a normal population with mean $\mu$, then $E\bar{X} = \mu$ and $p\lim \bar{X} = \mu$.

Given a random sample $X_1, \ldots, X_n$ from a $B(\pi)$ population, $\bar{X}$ is the sample proportion and $E\bar{X} = \pi$ and $p\lim \bar{X} = \pi$.

**What makes a good estimator?**

- When $E\bar{X} = \mu$ we say that $\bar{X}$ is an **unbiassed** estimator of $\mu$. Unbiassedness means that there is no systematic tendency to under- or over-estimate the value of the parameter.

- When $p\lim \bar{X} = \mu$ we say that $\bar{X}$ is a **consistent** estimator of $\mu$. Consistency means that as the sample size tends to infinity the probability of making an error exceeding any given size tends to 0.

- **Efficiency** (of an unbiassed estimator) 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 \( \mu \) (or of \( \pi \)) 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. (It is a special case of the famous Gauss-Markov theorem.)

- Given a random sample \( X_1, \ldots, X_n \) from a population with mean \( \mu \) and variance \( \sigma^2 \), then \( \bar{X} \) is an unbiased estimator of \( \mu \) and more efficient than any other **linear unbiased** estimator. (SW Key Concept 3.3. The sample mean is **BLUE**.)

Here is a sketch of the proof: a linear estimator is of the form \( \sum_{i=1}^{n} a_i X_i \) where the \( a_i \) are constants. The variance of this estimator is \( \sigma^2 \sum_{i=1}^{n} a_i^2 \). An unbiased linear estimator must have \( \sum_{i=1}^{n} a_i = 1 \). 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 3.)

**Finding estimators**

There are several methods for finding good estimators and I will mention three. All three—or variants of them—are used in econometrics.

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

- *least squares* based on the intuitively appealing principle of choosing the value of the parameters to make the agreement between expected and observed as close as possible.

- *maximum likelihood* based on the intuitively appealing principle of choosing the value of the parameters that
makes the observed sample most probable. (SW 398-9)

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

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

Least squares involves forming the square of the difference between each observation and its expected value, summing and finding the value that minimises the sum
\[ S = \sum (x_i - \pi)^2 : \]

\[ \frac{dS}{d\pi} = -2 \sum (x_i - \pi) = 0 \]

\[ \Rightarrow \hat{\pi} = \frac{\sum x_i}{n}. \]

This method also leads to the sample proportion.

Finding the maximum likelihood estimator takes more work. 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 \( X \) is

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

and the joint probability function for a random sample \( x_1, \ldots, x_n \) is
\[ p(x_1, \ldots, x_n; \pi) = p(x_1; \pi) \cdot \ldots \cdot p(x_n; \pi) \]
\[ = \pi^{x_1} (1 - \pi)^{1-x_1} \cdot \ldots \cdot \pi^{x_n} (1 - \pi)^{1-x_n} \]
\[ = \pi^{\sum x_i} (1 - \pi)^{n-\sum x_i}. \]

Interpreted as a function of \(\pi\) for given observations this is 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 

\[
\frac{d}{d\pi} \ln L(\pi; x) = \frac{\Sigma x_i}{\pi} - \frac{(n - \Sigma x_i)}{(1 - \pi)} = 0
\]

\[\Rightarrow \hat{\pi} = \frac{\Sigma x_i}{n}\]

The maximum likelihood estimator \(\hat{\pi}\) is the sample proportion which is also the method of moments estimator and the least squares estimator.

Recall SW’s question #2 about racial discrimination in the market for home loans. The simplest probability specification was random sampling from a \(B(\pi_B)\) for the blacks and random sampling from a \(B(\pi_W)\) for the whites. The proportions of denials are estimates of the \(\pi’s\): \(\hat{\pi}_B = .28\) and \(\hat{\pi}_W = .09\).

As a second example, consider estimating the two parameters \(\mu\) and \(\sigma^2\) 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

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n)$$

$$= f_X(x_1) \cdot \ldots \cdot f_X(x_n)$$

$$= \prod_{i=1}^{n} f_X(x_i)$$

analogous to the case for discrete variables.

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

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2}.$$  

If we now consider this as a function of the parameters $\mu$ and $\sigma^2$ for given $x$ we have the likelihood function
\[ L(\mu, \sigma^2; x) = \frac{1}{(2\pi \sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \]

The method of maximum likelihood has us choose the value of \( \beta \) and \( \sigma^2 \) which maximises \( L \).

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

\[ \ln L(\mu, \sigma^2; x) = -\frac{n}{2} \ln 2\pi \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \]

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 partial 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 \( \mu \) and the sample variance (with divisor \( n \)) the MLE of \( \sigma^2 \).

For each estimation method there is a body of theory describing the performance of the estimator in various circumstances. The theory of maximum likelihood estimation is based on various *large sample* properties:

- 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 asymptotically efficient.
Exercise 3

1. Given a random sample $X_1,\ldots,X_n$ from a population with mean $\mu$ and variance $\sigma^2$ and a linear estimator of $\mu$ 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 $\sum_{1}^{n} a_i = 1$. Check that all the estimators in b) are unbiased.

d) Use the Lagrange multiplier method to show that putting all $a_i = \frac{1}{n}$ will minimise
\[ \sigma^2 \sum_{1}^{n} a_i^2 \text{ subject to the constraint that} \]
\[ \sum_{1}^{n} a_i = 1. \]

2. 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

![Exponential Distribution](image)

Exponential \( \lambda = 1. \)

Show that the maximum likelihood estimator of \( \lambda \) given a random sample of size \( n \) is the reciprocal of the sample mean, \( \frac{\sum_{n} X_i}{n} \).
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 $X_1, \ldots, X_n$ are a random sample from a normal population with mean, $\mu$, and variance $\sigma^2$. We are making inferences about $\mu$ and, to begin with, we assume $\sigma^2$ is known.

From the reproductive property of the normal distribution we know that

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

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

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

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

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 \(\mu\).

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 \(\mu\).

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 \( \mu \).

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

In practice the value of \( \sigma^2 \) 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 \( \sigma^2 \).

Plugging in the estimate \( s \) changes the distribution from the standard normal to Student’s \( t \)-distribution:
\[
\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t^{(n-1)}.
\]

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 \textit{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 \textit{fatter} tails–i.e. the density does not fall away as quickly as the normal.

\[ N(0, 1) \text{ and } t^{(1)} \]

Incidentally fatter–than normal–tails are common in \textbf{financial} data and the analysis of such data requires special

\[ N(0, 1) \text{ and } t^{(1)} \]

Incidentally fatter–than normal–tails are common in \textbf{financial} data and the analysis of such data requires special
handling.
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 \( \hat{\pi} \)

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

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

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 \( \pi \) is unknown.

However it is also true that in large samples

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

and so there is a large sample 95% interval

\[
\left(\hat{\pi} - 1.96\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}\right).
\]
There is a very nice exact theory extending Student’s work to regression (SW ch. 5) but in most 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 $X_1, \ldots, X_n$ are a random sample from a normal population with mean, $\mu$, and variance $\sigma^2$. We assume $\sigma^2$ is known.

We know that

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

and we will be using this fact to test hypotheses about $\mu$.

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.
<table>
<thead>
<tr>
<th></th>
<th>$H_0$ is true</th>
<th>$H_A$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept $H_0$</td>
<td>✓</td>
<td>Type II error</td>
</tr>
<tr>
<td>Accept $H_A$</td>
<td>Type I error</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 $\mu_0$ and $\mu_A$ are specified numbers.

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

c) $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ where $\mu_0$ 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 $\overline{X}$ we have
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.

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

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

  $$P_{H_0}(\bar{X} < 11.5) = P\left( Z < \frac{11.5 - 12}{\sqrt{4}} \right) = 0.4111$$

  $$P_{H_A}(\bar{X} > 11.5) = P\left( Z > \frac{11.5 - 11}{\sqrt{4}} \right) = 0.4111$$

  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.
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 $1 - \Pr(\text{Type II error})$ is called the power of the test. The power of this particular test is about 59%.

In testing there are two common ways to proceed:

- 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

\[ P_{H_A}(\bar{X} > 7.86) = P \left( Z > \frac{7.86 - 11}{\sqrt{4}} \right) \]

\[ = P(Z > -1.57) = 0.94. \]

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 usually 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 of getting it right when an alternative value is true. Here are some values for the situation $\mu = 12$ versus $\mu < 12$ with $\frac{\sigma^2}{n} = 4$ and cut-off 11.5 as before.

We reject $H_0$ (accept $H_A$) when $\bar{X} < 11.5$.

Some values

$Power(\mu = 11)$ (from above):

$$P_{\mu=11}(\bar{X} < 11.5) = 0.5887 = 1 - 0.4113$$

$Power(\mu = 10)$:

$$P_{\mu=10}(\bar{X} < 11.5) = P_{\mu=10}\left(\frac{\bar{X} - 10}{2} < \frac{11.5 - 10}{2}\right)$$

$$= P(Z < 0.75) = 0.7734$$

$Power(\mu = 7)$:

$$P_{\mu=7}(\bar{X} < 11.5) = P_{\mu=7}\left(\frac{\bar{X} - 7}{2} < \frac{11.5 - 7}{2}\right)$$

$$= P(Z < 2.25) = 0.9877$$

For true 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).
For values close to the null hypothesis value (12) the power of the test will be approximately equal to the significance level.

\[ \text{Power}(\mu = 12) \text{ (significance level from above):} \]
\[ P_{\mu=12}(X < 11.5) = 0.4113. \]

The power function looks like:

Power: null 12 and cut-off 11.5.
This is the typical shape for a one-sided test using the left tail of the null hypothesis distribution.

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.

2 sample tests

SW’s question #2 about racial discrimination in the market for loans involved two samples.

The probability of denial for blacks is $\pi_B$, and for whites, $\pi_W$.

No discrimination is the null-hypothesis and discrimination is the alternative.

$H_0 : \pi_B - \pi_W = 0$ versus $H_A : \pi_B - \pi_W \neq 0$.

We assume that the sample sizes $n_B$ and $n_W$ are large enough to use the normal approximation.
The test statistic is $\frac{\hat{\pi}_B - \hat{\pi}_W}{\text{standard deviation}(\hat{\pi}_B - \hat{\pi}_W)}$ which is $N(0, 1)$ when $H_0$ is true.

The standard deviation of the difference is

$$\left(\frac{\pi_B(1 - \pi_B)}{n_B} + \frac{\pi_W(1 - \pi_W)}{n_W}\right)^{\frac{1}{2}}$$

based on the proposition that the variance of a difference is the sum of the variances.

Of course we do not know the probabilities but on the null hypothesis that $\pi_B = \pi_W$ we can estimate this common value by combining the two samples and taking the proportion of denials in that.

**Exercise 4**

1. $X_1, \ldots, X_n$ are a random sample from a normal population with mean, $\mu$, and variance $\sigma^2$.

We found that $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ provides a 95% confidence interval for $\mu$. 
The 99% interval $\overline{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$ 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 $\pi$ 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 $\overline{X}$ 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. A sample of 300 consists of 200 whites
and 100 blacks. 28% of black applicants were denied mortgages while only 9% of white. Test the hypothesis

\[ H_0 : \pi_B - \pi_W = 0 \] versus

\[ H_A : \pi_B - \pi_W \neq 0. \]

Report the prob value.

**Sampling distributions**

**SW39-45**

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 such as the Bernoulli for denial/acceptance;

- 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.

We have already met the most important sampling distribution—the normal
distribution.

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

- As the **exact** distribution of $\overline{X}$ and of $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ when we have a random sample from a normal population with mean $\mu$ and variance $\sigma^2$.

- As an **approximate large sample distribution** of $\overline{X}$ and of $\frac{\overline{X} - \mu}{s / \sqrt{n}}$ when we have a random sample from any population with mean $\mu$ and variance $\sigma^2$.

- 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 $\chi^2$ (*chi-squared*).
Consider $Z_1, \ldots, Z_k$ independent $N(0, 1)$ distributed random variables then

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

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

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

Not so clear but still true is
\[
\left( \frac{X_1 - \bar{X}}{\sigma} \right)^2 + \ldots + \left( \frac{X_n - \bar{X}}{\sigma} \right)^2 \sim \chi^2_{n-1}.
\]

From which it follows that
\[
\frac{(n - 1)s^2}{\sigma^2} \sim \chi^2_{n-1}.
\]

On degree of freedom has been lost because one parameter \( \mu \) 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 \( \chi^2 \).

3. We met Student’s \( t \) when constructing confidence intervals or testing hypotheses on \( \mu \) 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
we notice that the numerator is \( N(0, 1) \) and the denominator is the square root of a \( \chi^2 \) divided by its number of degrees of freedom. The numerator and denominator are independent.

This suggest a characterisation of the \( t \)-distribution. Suppose \( Z \) and \( U \) are independent random variables with \( Z \) standard normal and \( U \) a \( \chi^2 \) 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 \( \chi^2 \) and \( t \).
Suppose $U$ and $V$ are independent $\chi^2$ random variables with $d_1$ and $d_2$ degrees of freedom respectively, then the random variable

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

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 $\chi^2$ in the same way as the $t$ is a modification of the normal. When $d_2 \to \infty$ the denominator of the ratio converges in probability to 1 and the ratio, $X$, converges to $\frac{1}{d_1}$ times a $\chi^2_{d_1}$. 
Joint Distributions (SW 29-35)

We have been considering univariate random variables—the Bernoulli and the normal in particular—but we often need to consider several variables simultaneously.

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 betting on the toss of a fair coin. I win 1 if H and win −1 if T. Suppose we toss the coin twice and
the trials are independent. $X_i$ is what I win 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 $p_{X_1,X_2}$.

<table>
<thead>
<tr>
<th>$X_1 \setminus X_2$</th>
<th>$-1$</th>
<th>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>$1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

where the entries are formed on the pattern

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

Consider now cumulated winnings: $W_1$ denotes my winnings after one toss and $W_2$ my winnings after two tosses. Here are the values of $p_{W_1,W_2}$.

<table>
<thead>
<tr>
<th>$W_1 \setminus W_2$</th>
<th>$-2$</th>
<th>$-1$</th>
<th>$0$</th>
<th>$1$</th>
<th>$2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1$</td>
<td>$\frac{1}{4}$</td>
<td>$0$</td>
<td>$\frac{1}{4}$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1$</td>
<td>$0$</td>
<td>$0$</td>
<td>$\frac{1}{4}$</td>
<td>$0$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Note the relation between $W$ and $X$: 
$$W_1 = X_1$$
$$W_2 = W_1 + X_2.$$  

**Marginal distributions**

A joint distribution for a pair of rvs implies two univariate distributions, only for each variable. Thus

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

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

**Example** Continuing with coin tossing: the marginals $p_{X_i}$ are presented in the last row or column

<table>
<thead>
<tr>
<th>$X_1 \backslash X_2$</th>
<th>$-1$</th>
<th>$+1$</th>
<th>$X_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$+1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td></td>
</tr>
</tbody>
</table>

Thus
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 random variables the conditional probability function (or one of them) is defined

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

This is parallel to the definition of conditional probability.

**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

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

A useful theorem is the so-called *law of*
iterated expectations for random variables (SW 32-3)

\[ E(X_1) = \sum_{\text{all } x_2} E(X_1|X_2 = x_2)p_{X_2}(x_2) \]

\[ = E_{X_2}(E(X_1|X_2)) \]

The proof is as follows

\[ E(X_1) = \sum_{\text{all } x_1, x_2} x_1 p_{X_1X_2}(x_1, x_2) \]

\[ = \sum_{\text{all } x_1, x_2} x_1 p_{X_1|X_2}(x_1|x_2)p_{X_2}(x_2) \]

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

### Covariance and correlation

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.

The correlation between $X_1$ and $X_2$

$$corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{[varX_1 \cdot varX_2]^{\frac{1}{2}}}. $$

is a scaled covariance, scaled so that it lies between -1 and +1.

**Example**: The covariance of $W_1$ and $W_2$.

The values of $p_{W_1, W_2}$ were given above.

<table>
<thead>
<tr>
<th>$W_1 \backslash W_2$</th>
<th>−2</th>
<th>−1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>$p_{W_1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>−1</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$p_{W_2}$</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td></td>
</tr>
</tbody>
</table>

We have $EW_1 = EW_2 = 0$.

The covariance is
\[
E[(W_1 - EW_1)(W_2 - EW_2)] \\
= -1 \cdot -2 \cdot \frac{1}{4} + -1 \cdot 0 \cdot \frac{1}{4} \\
+ 1 \cdot 0 \cdot \frac{1}{4} + 1 \cdot 2 \cdot \frac{1}{4} = 1.
\]

We could have anticipated that the covariance would be positive because when I make a good start, \( W_1 \) is above its mean (0), I expect \( W_2 \) to be above its mean too. As an exercise calculate the correlation.

**Correlation and dependence**

The covariance (and correlation) can be thought of as measures of the linear dependence of random variables. Independent random variables are uncorrelated. This follows from

\[
Cov(X_1, X_2) = E[(X_1 - EX_1)(X_2 - EX_2)] \\
= E(X_1 - EX_1) \cdot E(X_2 - EX_2) \\
= 0.
\]

where I have used the fact that for independent rvs \( Y \) and \( Z \), \( EYZ = EY \cdot EZ \).
However uncorrelated variables are not necessarily independent. Consider $X$ and $X^2$ where

$$X = \begin{cases} 
-1 & \text{with probability } \frac{1}{4} \\
0 & \text{with probability } \frac{1}{2} \\
1 & \text{with probability } \frac{1}{4}
\end{cases}$$

These are variables are uncorrelated although $X^2$ is perfectly predictable from $X$.

**Revisiting the variance of a sum.**

We have often used the fact that the variance of a sum of independent variables is the sum of the variances. The **general** theorem on the variance of a sum is

**Theorem** The variance of the sum of random variables $X_1$ and $X_2$ is given by

$$var(X_1 + X_2) = var(X_1) + var(X_2) + 2Cov(X_1, X_2)$$

This can be shown as follows...
\[ \text{var}(X_1 + X_2) = E[(X_1 + X_2 - E(X_1 + X_2))^2] \]
\[ = E(X_1 - EX_1)^2 + E(X_2 - EX_2)^2 \]
\[ + 2E[(X_1 - EX_1)(X_2 - EX_2)] \]
\[ = \text{var}(X_1) + \text{var}(X_2) + 2\text{Cov}(X_1, X_2) \]

As an immediate consequence we have

**Theorem** The variance of a sum of **uncorrelated** variables is the sum of the variances

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

**Time series**

“Time series data are data from a single entity (person, firm, country) collected at multiple time periods.” SW11.

In the simplest situation there is a single observation for each consecutive time period. Numerous probability schemes are available for modelling the joint distribution of \(Y_1, Y_2, \ldots, Y_t, \ldots, Y_T\).

Successive \(Y\)'s are usually correlated—there is **serial correlation** or
autocorrelation—and the assumption of a random sample is not justified.

One widely-used model for time series is the random walk (SW556)

\[ Y_t = Y_{t-1} + u_t, \quad t = 1, 2, \ldots \]

where \( u_t \) is i.i.d. with 0 mean and variance \( \sigma^2 \). The \( u_t \) are sometimes called shocks or disturbances. The model is widely used in Finance and in Macroeconomics but the cumulated winnings \( W \) process (above) is a version of the random walk.

It is useful to express \( Y_t \) as

\[ Y_t = u_t + u_{t-1} + \ldots + u_1 + Y_0, \quad t = 1, 2, \ldots \]

where \( Y_0 \) is the starting value assumed fixed.

We have

\[ EY_t = Y_0, \quad t = 1, 2, \ldots \]

because \( E u_t = 0 \) for all \( t \).

\( Y_t \) and \( Y_{t-1} \) are correlated because they depend on the same quantities \( u_{t-1}, \ldots, u_1 \).
Specifically
\[ Y_t - EY_t = u_t + u_{t-1} + \ldots + u_1 \]
\[ Y_{t-1} - EY_{t-1} = u_{t-1} + \ldots + u_1. \]

Forming the covariance
\[
\text{covar}(Y_t, Y_{t-1}) = E(Y_t - EY_t)(Y_{t-1} - EY_{t-1})
\]
\[ = E(u_t + u_{t-1} + \ldots + u_1)(u_{t-1} + \ldots + u_1) \]
\[ = E(u_{t-1}^2 + \ldots + u_1^2) = (t - 1)\sigma^2. \]

Terms like \(E(u_t u_{t-1})\) (the covariance of \(u_t\) and \(u_{t-1}\)) vanish because the \(u's\) are independent.

As expected \(Y_t\) and \(Y_{t-1}\) are positively correlated.

One of the features of the random walk is that the variable becomes more and more dispersed over time
\[
\text{var}Y_t = \text{var}Y_{t-1} + \text{var}u_t
\]
\[ = \text{var}Y_{t-1} + \sigma^2. \]

The variance increases each period by a constant amount.
The random walk is a **non-stationary** process for its properties (such as variance and autocorrelation) are not constant over time.

Closely associated with time series is **forecasting**. For a random walk the ‘obvious’ forecast for $Y_{T+1}$ is $y_T$ the actual value at time $t$. There is a theory for such things.

Forecasting is usually combined with estimation for the model usually contains parameters that have to be estimated to create the forecast.

For example when

$$Y_t = \alpha Y_{t-1} + u_t$$

the ‘obvious’ forecast for $Y_{T+1}$ is $\alpha y_T$ but usually we do **not** know $\alpha$. So we estimate it and use $\hat{\alpha} y_T$ as the forecast.

**Exercise 5**

1. Consider the pair $W_1$ and $W_2$ defined earlier.
   
a) Work out the mean and variance of $W_1$
and of $W_2$. Deduce the correlation between $W_1$ and $W_2$.

b) Calculate the values of $p_{W_1|W_2}$. Are $W_1$ and $W_2$ independent?

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

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

and anything else you need 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. Consider the random walk

$$Y_t = Y_{t-1} + u_t, \quad t = 1, 2, \ldots$$

$Y_0$ fixed (non-stochastic).

where $u_t$ is i.i.d $N(0, \sigma^2)$.

a) What is the distribution of $Y_1$?

b) What is the distribution of $Y_t$?

You may find it useful to express $Y_t$ as

$$Y_t = u_t + u_{t-1} + \ldots + u_1 + Y_0, \quad t = 1, 2, \ldots$$
c) Use the formulae for $\text{covar}(Y_t, Y_{t-1})$, $\text{var}(Y_t)$ and $\text{var}(Y_{t-1})$ to find $\text{corr}(Y_t, Y_{t-1})$ called the first-order autocorrelation or serial correlation of $Y_t$.

**Linear equations, vectors & matrices**

This topic is important wherever linear equations are used. Econometrics I & II and Macroeconomics use matrices heavily.

There is a very brief treatment in Appendix 18.1 of SW and a more detailed 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. The ideas carry over to more equations in more unknowns.

**Ex. 1**

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

\[
\begin{align*}
  x_1 + x_2 &= 0 \\
  x_1 + 2x_2 &= 1
\end{align*}
\]

It is easy to see that the equations have a unique solution $x_1 = -1$ and $x_2 = 1$. Plotting the 2 straight lines we see that they intersect at this point.

**Ex. 2**

Consider now

\[
\begin{align*}
  x_1 + x_2 &= 0 \\
  2x_1 + 2x_2 &= 2
\end{align*}
\]

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

\[
\begin{align*}
x_1 + x_2 &= 1 \\
2x_1 + 2x_2 &= 2.
\end{align*}
\]

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:

Ex 1: \[
\begin{pmatrix}
1 & 1 \\
1 & 2
\end{pmatrix}
\]; 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

$$au = b$$

have a unique solution

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

(compare Ex 1) EXCEPT when $a = 0$.

There are then 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).

The reciprocal of a number satisfies

$$a^{-1}a = 1.$$
Ex 1 illustrates the matrix analogue involving the inverse of the matrix.

\[ A^{-1}A = AA^{-1} = I. \]

where the identity matrix \( I \) is a diagonal matrix with 1’s on the diagonal and 0’s elsewhere. In the \( 2 \times 2 \) case

\[ I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \]

Using the inverse we can write

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

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 (\( 1 \times 1 \) matrices) there is only singular number, viz. 0. In square matrices of order 2 or more there are infinitely many.

The solution to a pair of equations \( Ax = b \)
The distinctive terms that appear in the numerator and denominator are called determinants. Note that the same determinant \(a_{11}a_{22} - a_{12}a_{21}\) appears on the bottom of both expressions.

There is a problem when this determinant is zero for the ratios are not defined.

This is the singular case: the matrix of Ex 2 and 3

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

(if there is one) is given by

\[
x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}} : x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{12}a_{21}}
\]

The ‘problem’ with the matrix
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 $3 \times 3$ 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 possibility of no solution is illustrated by

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

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.
- The number (and $1 \times 1$ matrix) $a \neq 0$ has rank 1.

- The zero matrices $0$, \[
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix},
\]
\[
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

... all have rank 0.

To be invertible a matrix must have “full” rank—the rank must equal the number of rows.
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 examples with cleaner geometry.

**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 $45^\circ$ anti-clockwise. Thus the $45^\circ$ vector $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ is changed into $(0, 1)$.

This matrix is invertible because there is a matrix representing a rotation $45^\circ$ clockwise that undoes the action of $A$. The inverse is
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 5**

The singular matrix

$$\left( \begin{array}{cc} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array} \right)^{-1} = \left( \begin{array}{cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array} \right)$$

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

**Ex 6**

The singular matrix

\[
\begin{pmatrix}
1 & 0 \\
0 & 0 \\
\end{pmatrix}
\]

projects points onto the first axis, viz. \((x_1, x_2)\) becomes \((x_1, 0)\).
Exercise 6

1. Calculate the two products:
\[
\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\]
and
\[
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{pmatrix}
\]

2. Suppose the matrices \( C \) and \( D \) are non-singular and of the same dimension so that \( CD \) is defined. Show that the inverse of \( CD \) is given by
\[(CD)^{-1} = D^{-1}C^{-1}.\]

Examples of matrix manipulations

The matrix and vector formalism allows us to work in many dimensions as easily as in 1 dimension.

Provided the matrices are conformable most of the operations performed on numbers can performed on matrices.

The grand exception is that matrix multiplication is not commutative, i.e.

\[AB = BA\]

does NOT hold except in special cases.

Consider averaging \(A\) followed by projection \(P\) onto the axis and vice versa:

\[PAx\] is always on the axis

\[APx\] is always on the 45\(^{\circ}\) line
Example: a dynamical process

Let the vector $x_t$ represent the state of an economy at time $t$ and suppose the evolution of $x_t$ is given by

$$x_t = Ax_{t-1} + b, \quad t = 1, 2, \ldots$$

(an example of a difference equation.)

We can express $x_t$ in terms of $A$, $b$ and the starting value $x_0$ (solve the difference equation) as follows

$$x_t = Ax_{t-1} + b$$
$$= A(Ax_{t-2} + b) + b$$
$$= A^t x_0 + b + Ab + \ldots + A^{t-1} b.$$

(Look back at the similar algebra for the random walk.)

A steady state or equilibrium $x^*$ is defined by

$$x^* = Ax^* + b.$$

(The state $x^*$ once entered persists.)

It can be found (if it exists) as follows
\[ x^* - Ax^* = b \]
\[(I - A)x^* = b \]
\[ x^* = (I - A)^{-1}b. \]

This last step assumes that \((I - A)\) is non-singular.

We may consider whether the system tends to equilibrium as follows. Take
\[ x_t = Ax_{t-1} + b \]
\[ x^* = Ax^* + b \]
and subtract to obtain
\[ (x_t - x^*) = A(x_{t-1} - x^*) \]
\[ = A^2(x_{t-2} - x^*) \]
\[ = A^t(x_0 - x^*). \]

If \(A^t\) tends to the zero matrix then \((x_t - x^*)\) will tend to the zero vector as \(t \to \infty\).