MA1501: Statistical Inference

Dr Jonathan Gillard
GillardJW@Cardiff.ac.uk
M1.26

Topics Covered (informal description)

1. Sampling distributions
2. Point estimation and confidence intervals
3. Hypothesis testing
4. Linear regression (fitting a straight line)

How will this module be lectured?

Main mode of lecturing will be 'fill in the blanks' on provided lecture notes. Additional handouts, and material presented via the board/computer may also be given. Electronic lecture notes are not available for this module.

Example lectures will just be used as normal lectures.

Assessment

The summative assessment (100%) is through an unseen written examination (two hours).

Formative assessments will be given out. Feedback on these formative assessments will be given via (i) written comments, (ii) tutorial sessions (which will go over the assessments) and (iii) videos available on Learning Central which will go through the solutions of some of the problems on the formative assessments (which I will call Example Sheets).

Reading List

Probability and Statistical Inference, by Hogg and Tanis.

Many other suitable books - too many to mention (just look in library for any introductory statistics text).

1 Sampling distributions

1.1 Formal definitions

For the moment, consider a population as a collection of objects, such as people, families, cars etc. A sample is a sub-collection or part of the population. Populations are studied because they have some property or characteristic that varies among different members of the population. Such a characteristic is called a variable e.g. the monthly income of families, the fuel consumption of a car etc.

A variable identifies a property of interest, and is the basis upon which values are associated with members of the population. Formal definitions are:

Definition 1.1 *A* population *is the collection of all values of the variable under study.*

Definition 1.2 *A* sample *is any sub-collection of the population.*

Definition 1.3 *A* population parameter *is some numerical measure associated with the population as a whole.*

Definition 1.4 *Let X denote a random variable. A* random sample from the distribution *of X is a set of independent and identically distributed (i.i.d) random variables $X_1, X_2, \ldots, X_n$.*

Definition 1.5 *The values taken by $X_1, X_2, \ldots, X_n$ in an actual sample are denoted by $x_1, x_2, \ldots, x_n$ and are called the* sample values*.*

Definition 1.6 *A* statistic *is a function of $X_1, X_2, \ldots, X_n$ and is thus itself a random variable. It does not contain any unknown parameters.*

1.2 Common sample statistics

Definition: The sample mean $\bar{x}$ is defined to be

Definition: The sample variance s^2 is defined to be

We use $(n-1)$ in the denominator compared with the definition of the variance of the population (n). We use $(n-1)$ instead of n for a practical reason, and shall prove why it is beneficial to use $(n-1)$ later.

To emphasize the occasions when we are treating the sample mean and sample variance as random variables, we shall use the notation $\bar{X}$ and S^2 respectively.

Lemma 1.7 *s^2 may also be written as* $s^2 = \dfrac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\} = \dfrac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}$.

Proof:

1.3 Sampling distributions

1.3.1 Motivating example

A purse contains six coins: one 5p, two 10p, one 20p, two 50p coins. Consider the selection of two coins from the purse, obtained with replacement. Let X be the value of a randomly chosen coin, and let X_1 and X_2 be the values of the two selected coins.

We can construct the probability distribution of X:

x	5p	10p	20p	50p
$P[X = x]$				

We can also compute $E[X] =$

and we can compute $Var[X] =$

Now consider the distribution of (X_1, X_2):

(x_1, x_2)	No. of ways	Prob.	$\bar{x}$	s^2
(5p,5p)	1	$\frac{1}{36}$	5	0
(5p,10p)	4	$\frac{4}{36}$	7.5	12.5

We can now work out the sampling distribution for $\bar{X}$:

$\bar{x}$	5	7.5	10	12.5	15	20	27.5	30	35	50
$P[\bar{X} = \bar{x}]$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{8}{36}$	$\frac{4}{36}$	$\frac{4}{36}$

and $E[\bar{X}] = 24.1667 = E[X]$, $Var[\bar{X}] = 176.736 = \dfrac{Var[X]}{2}$.

The sampling distribution for S^2 is:

s^2	0	12.5	50	112.5	450	800	1012.5
$P[S^2 = s^2]$	$\frac{10}{36}$	$\frac{4}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{8}{36}$	$\frac{4}{36}$

and $E[S^2] = 353.4722 = Var[X]$. We can also compute $Var[S^2]$ if we wish. We can also do the same for other sample statistics, such as the sample minimum.

1.3.2 Sampling from a distribution: commonly used results

Theorem 1.8 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \dfrac{\sigma^2}{n}$.*

Proof:

Theorem 1.9 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then $E[S^2] = \sigma^2$.*

Proof:

Theorem 1.10 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with cumulative distribution function $F(x)$ and probability density function $f(x)$. The probability density function of the maximum and minimum of $X_1, X_2, \ldots, X_n$ are given by $g(z) = nf(z)[F(z)]^{n-1}$ and $h(w) = nf(w)[1 - F(w)]^{n-1}$, where $z = \max(x_1, x_2, \ldots, x_n)$ and $w = \min(x_1, x_2, \ldots, x_n)$.*

Proof:

It is also possible to prove a similar result for discrete distributions, but we will not do that in this module.

Example 1.11 *The random variable X has probability density function*

$$f(x) = 12x^2(1-x) \;\; 0 \leq x \leq 1.$$

Obtain the probability density function of the sample maximum, when a random sample of size n is taken from X. Hence, or otherwise, find the probability that the largest maximum is $\frac{1}{2}$.

Solution:

1.4 Special distribution results and commonly used distributions for inference

1.4.1 Normal distribution

Let X be a normally distributed random variable. Its probability density function is given by

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

We have that $E[X] =$

and $Var[X] =$

For brevity we write $X \sim N[\mu, \sigma^2]$.

Graphs of this probability density function for different μ and σ^2 are given below:

If X is normally distributed, the so-called standardized random variable $Z = \dfrac{X-\mu}{\sigma}$ is also normally distributed, with mean 0 and variance 1. Its probabilities and percentiles are tabulated.

Example 1.12 *Let $X \sim N[4, 3^2]$. Find $P[X < 5]$, $P[X > 7]$ and the value of x such that $P[X < x] = 0.95$.*

Here we use statistical tables (available on Learning Central):

Theorem 1.13 *If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables, such that each X_i has mean μ_i and variance σ_i^2, then for any constants a_i, $i = 1, 2, \ldots, n$, the random variable $\sum_{i=1}^{n} a_i X_i$:*

a. is normally distributed,

b. has mean $\sum_{i=1}^{n} a_i \mu_i$,

c. has variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$.

This tells us that 'a linear combination of normally distributed variables is also normally distributed'.

Lemma 1.14 *If $X_1, X_2, \ldots, X_n$ are independent and identically distributed normal random variables, such that each X_i has mean μ and variance σ^2, then $\sum_{i=1}^{n} X_i \sim N\left[n\mu, n\sigma^2\right]$.*

Proof:

Lemma 1.15 *If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables, such that each X_i has mean μ and variance σ^2, then $\bar{X} \sim N\left[\mu, \frac{\sigma^2}{n}\right]$.*

Proof:

Its standardized version which is used in practise is:

Example 1.16 *The weights of sacks of potatoes are normally distributed with mean $25kg$ and standard deviation $1kg$. Find:*
(i) the probability that the mean weight of a random sample of four sacks is greater than $26kg$;
(ii) the probability that the total weight of a random sample of 16 sacks lies between $396kg$ and $405kg$;
(iii) the sample size necessary for the sample mean to be within $0.25kg$ of the true mean $25kg$ at least 95% of the time.

Solution:

1.4.2 Students' t-distribution

In probability and statistics, Students' t-distribution (or simply the t-distribution) is a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean. Let the random variable X have the Students' t-distribution. Its probability density function is given by:

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

It depends on one parameter ν, called the *degrees of freedom*. For brevity we write $X \sim t(\nu)$. Note that as $\nu \to \infty$ the Students' t-distribution behaves like the normal distribution.

Plots of the probability density function for the Students' t-distribution for different ν are given below:

Lemma 1.17 *If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then the random variable $T = \dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ has the Students' t-distribution with $(n-1)$ degrees of freedom.*

Most statistical tables give only the percentiles of T, for example, with $n = 11$, $P[T < 1.812] = 0.95$.

1.4.3 Chi-squared distribution

For n a positive integer, the chi-squared distribution is the distribution of

$$X_1^2 + X_2^2 + \ldots + X_n^2$$

where $X_1, X_2, \ldots, X_n$ are all independently and identically normally distributed with zero mean, and unit variance.

The chi-squared distribution depends on one parameter, which again is called the degrees of freedom.

Shortly, if X is a random variable with chi-squared distribution with ν we can write $X \sim \chi^2(\nu)$. Here $E[X] = \nu$ and $Var[X] = 2\nu$.

Note that this distribution is not symmetric. A plot of the chi-squared distribution for different degrees of freedom ν is given below:

Lemma 1.18 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then $\dfrac{(n-1)S^2}{\sigma^2}$ has a chi-squared distribution with $(n-1)$ degrees of freedom.*

1.4.4 F-distribution

The F-distribution is another skewed distribution which depends on two parameters, again known as two degrees of freedom m and n. If $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$ are two independent random variables, then

$$Z = \frac{X/m}{Y/n} \sim F(m, n)$$

is said to have the F distribution with m and n degrees of freedom. A plot of the F distribution with different degrees of freedom is given below:

Note that if $Z \sim F(m, n)$ then $1/Z \sim F(n, m)$.

Lemma 1.19 *Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be random samples from $X \sim N[\mu_X, \sigma_X^2]$ and $Y \sim N[\mu_Y, \sigma_Y^2]$ respectively. Let S_X^2 and S_Y^2 be the corresponding sample variances. It follows that $\dfrac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1)$ and $\dfrac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1)$. Hence*

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1).$$

1.4.5 The Central Limit Theorem

Theorem 1.20 *Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. Then:*

a. $\displaystyle\sum_{i=1}^{n} X_i \sim N\left[n\mu, n\sigma^2\right]$, *approximately*

b. $\bar{X} \sim N\left[\mu, \dfrac{\sigma^2}{n}\right]$, *approximately.*

The approximation improves as $n \to \infty$.

Example 1.21 *The number of typing errors made on a page follows a Poisson distribution with mean 2. Use the central limit theorem to calculate (approximately) the probability that there are more than 950 typing errors in a 450 page book.*

Solution:

2 Point estimation and confidence intervals

2.1 Point estimation

In statistics, we take a sample which we use to infer things regarding a population. Thats why this module is called 'Statistical Inference'.

Statistical inference is needed to answer questions such as:

- What are the voting intentions before an election? *Market research, opinion polls, surveys*
- What is the effect of obesity on life expectancy? *Epidemiology*
- What is the average benefit of a new cancer therapy? *Clinical trials*
- What proportion of temperature change is due to man? *Environmental statistics*
- What is the benefit of speed cameras? *Traffic studies*
- What portfolio maximises expected return? *Financial and actuarial applications*
- How confident are we the Higgs Boson exists? *Science*
- What are possible benefits and harms of genetically-modified plants? *Agricultural experiments*
- What proportion of the UK economy involves prostitution and illegal drugs? *Official statistics*
- What is the chance Liverpool will best Arsenal next week? *Sport*

In the population we have population parameters, properties of the population that we may or may not know. Upon taking a sample, we can compute sample statistics. For example, to estimate the population mean, we may take a sample, then compute a sample mean, and this seems a reasonable approach.

However, sometimes, there may be several alternative estimators that can be used to estimate a parameter. We therefore need criteria to compare estimators and to decide which is the 'best' in a particular situation.

Example 2.1 *Let X be the random variable denoting the duration between calls to 999. X is known to follow an exponential distribution, with probability density function*

$$f(x;\lambda) = \frac{1}{\lambda}\exp(-x/\lambda), \;\; x \geq 0\,.$$

You are asked to compute the probability that the duration between calls is less than 1 minute. You correctly note that to do this, you need to estimate λ. Suggest a strategy to estimate λ.

Methods of finding estimators are covered in later modules. Here we concentrate on

1. 'Obvious' estimators, or ones that can be deduced from sensible statistical reasoning

2. Properties of estimators

3. Potential ways to compare estimators

For notation, and to be general, let θ denote a general population parameter that we wish to estimate. Estimators are often denoted by a circumflex above the parameter e.g. $\hat{\theta}$ is an estimator of θ.

Definition:

Definition:

Definition 2.2 *The mean square error (MSE) of an estimator $\hat{\theta}$ of θ is defined as*

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

Theorem 2.3 *The mean square error (MSE) of an estimator $\hat{\theta}$ of θ may be written as*

$$MSE[\hat{\theta}] = Var[\hat{\theta}] + \{bias[\hat{\theta}]\}^2$$

Proof:

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of a parameter θ, $\hat{\theta}_1$ is said to be a better estimator, in mean square error, than $\hat{\theta}_2$ if $MSE[\hat{\theta}_1] < MSE[\hat{\theta}_2]$.

Example 2.4 *A scientist wants to estimate the volume v of a cuboid whose unequal edges are of lengths x, y and z. The scientist can estimate the volume using either of these two methods.*

Method (i): Obtain a direct measurement V of the volume. V can be regarded as a random variable having mean v and variance σ_V^2.
Method (ii): Obtain measurements X, Y and Z of the unequal edges, and estimate the volume using $\hat{v} = XYZ$. It may be assumed that X, Y and Z are independent random variables with respective means x, y and z and common variance σ^2.

Show that method (ii) gives an unbiased estimate of the volume, but that method (i) is preferable to method (ii) if

$$\sigma_V^2 < \sigma^6 + \sigma^4(x^2 + y^2 + z^2) + \sigma^2(x^2y^2 + x^2z^2 + y^2z^2).$$

Solution:

2.2 Confidence intervals

The discussed properties of estimators provide valuable information on comparing and choosing an estimator, but say rather little about the quality of a particular estimate.

Definition 2.5 *A* $100(1-\alpha)\%$ *confidence interval for an unknown parameter* θ*, with estimator* $\hat{\theta}$ *is an interval*

$$C = [\hat{\theta} - l, \hat{\theta} + u]$$

for a lower value l *and upper value* u *such that*

$$P[\theta \in C] = 1 - \alpha\,.$$

This interval C is random, because the value of $\hat{\theta}$ depends on the data. In the long run, $100(1-\alpha)\%$ of confidence intervals will contain θ.

2.2.1 Confidence intervals for means, with normally distributed population and variance known

Example 2.6 *Construct a* $100(1-\alpha)\%$ *confidence interval for the unknown mean* μ *of a normally distributed population, with known variance* σ^2*, having observed a random sample* $X_1, X_2, \ldots, X_n$ *from this population.*

Solution:

Example 2.7 *A sample of size $n = 4$ gave values 12.4, 13.6, 12.9 and 13.4 when randomly sampled from a normally distributed population with unknown mean μ, and variance $\sigma^2 = 0.25^2$. Compute a 95% confidence interval for μ.*

Solution:

Example 2.8 *Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two random samples, each from a normal distribution with unknown means μ_1 and μ_2, but known variances σ_1^2 and σ_2^2 respectively.*

A $100(1-\alpha)\%$ *confidence interval for the difference of the unknown means* $\mu_1 - \mu_2$*, is given by*

$$(\bar{x} - \bar{y}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\,.$$

Solution: The idea of the method is entirely the same as the confidence interval derived previously. Make sure you can derive this confidence interval yourself. The main steps are as follows:

Let $\bar{X}$ and $\bar{Y}$ denote the sample means of $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$.

It follows that $\bar{X} \sim N\left[\mu_1, \sigma_1^2/m\right]$ and $\bar{Y} \sim N\left[\mu_2, \sigma_2^2/n\right]$.

Hence $\bar{X} - \bar{Y} \sim N\left[\mu_1 - \mu_2, \sigma_1^2/m + \sigma_2^2/n\right]$ and $Z = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim N\left[0, 1\right]$.

So we can find a value $z_{1-\alpha/2}$ such that $P[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}] = 1 - \alpha$. This equation can be rearranged to yield a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Compare and contrast the steps outlined above with the confidence interval we have derived previously.

Example 2.9 *The mean life of two types of light bulbs, A and B, were compared by testing 20 bulbs of type A and 25 bulbs of type B. The sample means, in obvious notation, were given by* $\bar{x}_A = 1021.3$ *hours and* $\bar{x}_B = 1005.7$ *hours. Assuming that the life of both types of bulbs is normally distributed with standard deviation 30 hours (for both cases), find a* 95% *confidence interval for the difference in the population means for both types.*

Solution:

2.2.2 Confidence intervals for means, with normally distributed population, but variance unknown

In the previous section we considered confidence intervals for means when the population variance was known if we have an independent and identically distributed sample from a normally distributed population. Hence, we use the fact that

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N[0, 1]$$

to construct a probability statement concerning Z, and then algebraically manipulate this to achieve a probability statement for $\bar{X}$.

If we do not know what σ^2 is, we can make use of the following fact considered in the previous chapter:

> If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then the random variable
>
> $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
>
> has the Students' t-distribution with $(n - 1)$ degrees of freedom.

Example 2.10 *Construct a* $100(1 - \alpha)\%$ *confidence interval for the unknown mean* μ *of a normally distributed population, with unknown variance, having observed a random sample* $X_1, X_2, \ldots, X_n$.

Solution:

Example 2.11 *The heights of 16 randomly selected children aged 11 were measured, with sample mean 121cm. The sample variance was calculated to be 25cm. Assuming that these heights are normally distributed, find the* 90% *confidence interval for the population mean height.*

Solution:

We now wish to find confidence intervals for the difference of two population means. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two random samples, each from a normal distribution with unknown means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2 respectively.

If we knew the variances, then we can use the fact that

$$\bar{X} - \bar{Y} \sim N\left[\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right],$$

and derive a confidence interval in the usual way.

If we do not know σ_1^2 and σ_2^2, we again can use Students' t-distribution, but we also can consider two other options:

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$. In this case we assume both samples come from populations which may have equal variances. This means we can use both samples combined to estimate σ^2, rather than using each sample separately, to consequently estimate σ_1^2 and σ_2^2 separately.
2. $\sigma_1^2 \neq \sigma_2^2$. In this case we assume both samples come from populations which do not have equal variances, and so we have to consequently estimate σ_1^2 and σ_2^2 separately.

Case 1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ Let S_X^2 be the sample variance computed from $X_1, X_2, \ldots, X_m$, and let S_Y^2 be the sample variance computed from $Y_1, Y_2, \ldots, Y_n$. Both of these are estimators of σ^2.

The optimal combination of S_X^2 and S_Y^2 to estimate σ^2(in the sense that it is the estimator with minimum variance is):

See the exercise sheets for a proof of this result.

One can show that

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

has the Students' t distribution with $m + n - 2$ degrees of freedom. Hence one may construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm t_{1-\alpha/2}(m+n-2)s\sqrt{\frac{1}{m} + \frac{1}{n}}\,.$$

Case 2. $\sigma_1^2 \neq \sigma_2^2$ Let S_X^2 be the sample variance computed from $X_1, X_2, \ldots, X_m$, and let S_Y^2 be the sample variance computed from $Y_1, Y_2, \ldots, Y_n$. If $\sigma_1^2 \neq \sigma_2^2$ then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$$

can be shown to have a Students' t-distribution with ν degrees of freedom.

But ν can only be estimated from

$$\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{\left(\frac{s_X^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_Y^2}{n}\right)^2}{n-1}}\,.$$

This is known as Satterthwaite's approximation.

Hence one may construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm t_{1-\alpha/2}(\nu)\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}\,.$$

2.2.3 Confidence intervals for means, with normally distributed population, with paired observations

Sometimes the assumption that two random variables X and Y are independent is inappropriate because there is a natural pairing of results. For example, lets imagine that I measure your systolic blood pressure today, and then measure it again tomorrow. You would provide me with two measurements, and it is not appropriate to assume that both measurements are independent.

Suppose however that I still want to obtain a confidence interval for the difference of the population means X and Y.

It is possible to construct such a confidence interval:

Example 2.12 *An experiment was conducted to measure mileages of cars under two types of petrol: Fast Oil, and Slicker. Ten cars are initially used with Fast Oil petrol, and these cars are used again under the same conditions but with Slicker petrol. The results were as follows:*

Car:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Fast Oil	*38.4*	*39.6*	*37.6*	*40.2*	*36.9*	*39.4*	*38.3*	*39.6*	*39.1*	*38.2*
Slicker	*39.8*	*40.3*	*39.7*	*41.2*	*38.6*	*40.6*	*39.9*	*41.1*	*40.8*	*39.8*

Find a 90% *confidence interval for the population mean difference.*

Solution:

2.2.4 Approximate confidence intervals using the Central Limit Theorem

The possibility of generating approximate confidence intervals using the Central Limit Theorem is demonstrated by an example.

Example 2.13 *Suppose that X has a Poisson distribution with mean λ. A random sample of size 100 is taken from the distribution of X. Given that the sample mean is 6.1, obtain an approximate 90% confidence interval for λ. Find also an approximate confidence interval for $\exp(-\lambda)$, the probability that X takes the value zero.*

Solution:

2.2.5 Confidence intervals for variances

Recall that

> Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ

and variance σ^2. Then $\dfrac{(n-1)S^2}{\sigma^2}$ has a chi-squared distribution with $(n-1)$ degrees of freedom.

Example 2.14 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Derive a $100(1-\alpha)\%$ confidence interval for σ^2.*

Solution:

Example 2.15 *Suppose a random sample of 25 observations was taken from a normal distribution, and it was found that the sample variance was equal to 10. Find a 95% confidence interval for σ^2.*

Solution:

2.2.6 Confidence intervals for ratio of variances

Recall the following:

> Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be random samples from $X \sim N[\mu_X, \sigma_X^2]$ and $Y \sim N[\mu_Y, \sigma_Y^2]$ respectively. Let S_X^2 and S_Y^2 be the corresponding sample variances. It follows that $\dfrac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1)$ and $\dfrac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1)$.
> Hence $\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1)$.

A $100(1-\alpha)\%$ confidence interval for $\dfrac{\sigma_X^2}{\sigma_Y^2}$ is given by

$$\left[\frac{s_X^2}{s_Y^2}\frac{1}{F_{1-\alpha/2}(m-1,n-1)}, \frac{s_X^2}{s_Y^2}\frac{1}{F_{\alpha/2}(m-1,n-1)}\right]$$

This will be derived in an exercise sheet. Here, $F_{1-\alpha/2}(m-1, n-1)$ is the $100(1-\alpha/2)th$ percentile of the F-distribution with $m-1$ and $n-1$ degrees of freedom.

Example 2.16 *Using the notation introduced in this section, obtain a* 95% *confidence interval for* σ_X^2/σ_Y^2 *if* $m = 10$, $n = 5$, $s_X^2 = 20$ *and* $s_Y^2 = 35.6$.

Solution:

3 Hypothesis testing

3.1 Introduction and key concepts

A statistical hypothesis is a statement concerning the distribution of a random variable e.g. the hypothesis that a coin is fair. This hypothesis concerns the distribution of the random variable generating the process, as tossing coins can be modelled as a binomial random variable, where we toss the coin n times and have a proportion p of heads. The hypothesis that a coin is fair is testing the hypothesis that $p = \frac{1}{2}$ or not. To perform an hypothesis test we need to define lots of terminology, concepts and notation. As far as possible we will remain general and test hypotheses concerning an unknown population parameter θ.

The purpose of the test is to choose between two hypotheses. One is called the null hypothesis, and the other is called the alternative hypothesis. The null hypothesis is denoted H_0 and the alternative hypothesis is denoted H_1.

The null hypothesis is commonly of the form:

Corresponding to this null hypothesis, the alternative hypothesis can be one of the following:

The first option is known as 'two-tailed' whilst the latter two are known as 'one-tailed'.

We thus have two options:

1. We reject H_0, and accept H_1
2. We accept H_0, and reject H_1

Consequently there are two types of errors that we can make:

1. The incorrect rejection of H_0, known as a Type I error, which happens with probability α.
2. The incorrect rejection of H_1, known as a Type II error, which happens with probability β.

The probability of making a Type I error is denoted α, and this is known as the significance level of the hypothesis test. The probability of making a Type II error is denoted β. Another concept used in hypothesis testing is the so-called power of the hypothesis test, and this is given by $1 - \beta$. Naturally for any hypothesis test, we desire α to be as small as possible, and the power $1 - \beta$ to be as large as possible.

We now address the problem of how to create the decision rule of accepting H_0 or rejecting H_0. This is decided on the basis of a test statistic. We will now consider an example to illuminate some of these concepts described so far, and also outline each of the main steps necessary.

Example 3.1 *A beer-dispensing machine is supposed to deliver 20 fluid ounces of beer. The amount dispensed by the machine is thought to be normally distributed. 10 samples are measured from the machine, with the following results: 19.89, 19.90, 19.87, 19.94, 19.92, 19.90, 19.93, 19.91. The sample mean is given by 19.904, and the sample standard deviation is given by 0.0217. Test the hypothesis that the mean amount dispensed by the machine is indeed 20 fluid ounces.*

We will now deconstruct the steps necessary to perform this hypothesis test.

Formally form the hypotheses

Select an appropriate test statistic with a known distribution, and compute it assuming H_0 is true

"Determine how likely the test statistic is to happen"

Clearly state your conclusions

Interpretations of a p-value should be along the following lines:

- $p < 0.01$; there is very strong evidence for rejecting H_0.
- $0.01 \leq p \leq 0.05$; there is strong evidence for rejecting H_0.
- $p > 0.05$; there is insufficient evidence for rejecting H_0.

3.2 Examples

Example 3.2 *It is claimed that a shop makes a profit of £*850*, per week, on average. To test this, the manager records the weekly profit across five randomly selected weeks, and finds the average to be £*905*, with standard deviation £*50*. The manager asks the question "Have profits increased significantly?". Perform a suitable hypothesis test at a* 5% *significance level. Also approximate the p-value.*

Solution:

Example 3.3 *Ten students failed a diagnostic test, and are asked to resit after a period of intensive revision and support. The results are given below:*

Student	*Test 1*	*Test 2*
1	*30*	*40*
2	*32*	*34*
3	*27*	*50*
4	*30*	*38*
5	*35*	*33*
6	*32*	*45*
7	*25*	*40*
8	*20*	*30*
9	*31*	*38*
10	*30*	*42*

Conduct an appropriate hypothesis test to determine if the intensive revision and support was effective.

Solution:

Example 3.4 *A petrol additive is supposed to increase the mpg (miles per gallon) of cars. A company runs a fleet of 30 identical cars. For a period of two months, 20 of the cars were run on petrol with the additive, whilst the other 10 were run without. Cars without the additive gave an average mpg of 38.2 with standard deviation 5.3. Cars with the additive gave an average mpg of 45.6, with standard deviation 4.7. Clearly stating your assumptions, perform a suitable hypothesis test to compare the average mpg between cars that received the additive, and those that didn't.*

Solution:

Example 3.5 *Two samples of size 16 and 26 taken from two normally distributed populations had sample variances of 29.6 and 9.8 respectively. Use an appropriate hypothesis test to test the assertion that the population variances are equal.*

Solution:

3.3 Hypothesis testing using the Central Limit Theorem

The hypothesis tests considered so far are based on scenarios when the data are normally distributed. This small section of notes will show that the central limit theorem can be used to design similar tests when the sample sizes are large enough for this theorem to apply.

Suppose therefore that X is a random variable with mean μ and variance σ^2. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X, then by the central limit theorem the statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

under a null hypothesis $H_0 : \mu = \mu_0$ is approximately $N[0, 1]$ for large n.

This method is widely applicable, and the example that will be demonstrated here involves the binomial distribution. The binomial distribution is characterised by two parameters: p and n. n is the number of trials, and p is the probability of success. Lets imagine that we want to test hypotheses concerning p. The theory is contained here.

Suppose that we observe a sequence of Bernoulli trials with unknown success probability p, and we wish to test the null hypothesis against a suitable one or two tailed alternative. Let Y be the number of successes obtained in n trials and let $\hat{p} = \frac{Y}{n}$ be the observed proportion of successful trials.

It follows that Y has a binomial distribution, $B[n, p]$, which is approximately $N[np, np(1-p)]$ for large n. Thus $\hat{p}$ is approximately $N[p, \frac{p(1-p)}{n}]$ (why?) and the standardised statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

is approximately $N[0, 1]$ under a null hypothesis $H_0 : p = p_0$. We can then reject or accept H_0 by comparing z with the appropriate critical value of the normal distribution.

Example 3.6 *A drug company claims in an advertisement that* 60% *of people suffering from a certain complaint gain instant relief by using a particular product. In a random sample, 106 out of 200 did gain instant relief. Test the validity of this claim, and find the p-value.*

Solution:
The sample is large enough to use the normal approximation. The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.53 - 0.6}{\sqrt{1/200 \times 0.6 \times 0.4}} = -2.021$$

The p-value for a one tailed test (with alternative $p < 0.6$ or $p > 0.6$) is $P[N[0, 1] < -2.021] = 0.022$. This is small and so the null hypothesis $p = 0.6$ is rejected. The p-value for a two tailed test is $2 \times 0.022 = 0.044$. This is still small, and so for a 5% significance level, the null hypothesis is rejected again.

3.4 Relationship between hypothesis testing and confidence intervals

The formal definition of a $100(1-\alpha)\%$ confidence interval is that it is the set of values of a parameter that would be accepted as the null hypothesis in a hypothesis test with significance level α.

So, a 95% confidence interval corresponds to a test with significance level 0.05. However the two should not be interchanged. The role of a confidence interval is to provide an interval estimator for a parameter with an associated level of uncertainty. A hypothesis test is used to gauge the strength of the evidence in the data for or against a specified null hypothesis, and this evidence is most clearly stated by quoting the p-value of the test.

Consider the usual two-tailed test for the difference in means of two independent normally distributed populations, using the significance level approach.

3.5 Principles of hypothesis testing

The principles of statistical hypothesis testing can be summarized in five steps.

1. Set up a null hypothesis about a parameter for testing. Also decide the level of evidence that is needed to reject this null hypothesis (see steps 4 and 5).

2. Take a random sample from the population.

3. Estimate the parameter from the sample and use this to calculate the value of a test statistic, whose distribution is known when the null hypothesis is known.

Steps 4 and 5 depend on whether the significance level approach is adopted, or the p-value approach. The latter is preferred, if it is possible to calculate or approximate the p-value, because it gives more information about the strength of evidence for or against the null hypothesis that is contained in the observations.

4a. In the p-value approach, calculate the probability, using the probability distribution of the test statistic, of a result as extreme or more so than the one observed.

5a. If the probability calculated in Step 4a is judged to be small, reject the null hypothesis, otherwise accept the null hypothesis.

4b. In the significance level approach a pre-determined significance level, a smallish probability (typically 0.05 or 0.01) is predetermined in Step 1 as the level of evidence needed to reject the null hypothesis. Using this significance level a critical value of the test statistic is calculated that determines a critical region and an acceptance region.

5b. If the value of the test statistic is in the critical region, the null hypothesis is rejected (at the pre-determined level of significance). If the value of the test statistic is in the acceptance region, accept the null hypothesis.

3.6 The Power function

In previous sections we have designed tests, using only the significance level α and the probability distribution of the test statistic when H_0 is true. We cannot, in general, compute the Type II error probability (β) as it varies with the parameter values consistent with H_1.

We can however regard it as a function of these values and knowledge of this function helps us to assess the performance of the test. We make the following definition:

Definition: Consider a test of two hypotheses, H_0 and H_1, concerning a parameter θ. The power function of the test, denoted by $\pi(\theta)$ is the probability of rejecting H_0, and hence accepting H_1, expressed as a function of θ. If C is the critical region, we use the notation $P(C;\theta)$ to denote this, i.e.

$$\pi(\theta) = P(C;\theta).$$

The following example shows how the power function can be calculated.

Example 3.7 *A random variable X has the distribution $N[\theta, 1]$. Using a random sample of size 100 and a 5% significance level, compute the power function for the following hypothesis test:*

$$H_0 : \theta = 0; \qquad H_1 : \theta \neq 0\,.$$

Solution:

The graph of the power function for the above example is given below.

Three points are worth noting:

1. $\pi(\theta)$ is a probability and lies between 0 and 1.
2. $\pi(0) = P(C; 0) = \alpha = 0.05$.
3. As $|\theta|$ increases, $\pi(\theta)$ approaches unity. At the same time the type II error probability β approaches zero since $\beta = 1 - \pi(\theta)$.

A good test will reject H_0 with low probability if it is true but will reject it with high probability if it is false. Thus $\pi(\theta)$ should be close to zero for values of θ consistent with H_0 and close to unity otherwise. In the above example the ideal power function is thus:

$$\pi(\theta) = \begin{cases} 0, & \text{if } \theta = 0 \\ 1, & \text{if } \theta \neq 0. \end{cases}$$

The ideal power function is not attainable in practice but the closer the actual power function is to this, the better the test. We can usually improve the power function by increasing the sample size.

Example 3.8 *Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution $N[\mu, 1]$. Construct a test of the hypotheses*

$$H_0 : \mu = 0; \qquad H_1 : \mu \neq 0\,.$$

at the 5% significance level and plot its power function for sample sizes of 25, 100 and 500.

Solution:

Example 3.9 *Let X have a uniform distribution on $[0, \theta]$. Using a random sample of size 12, we wish to test the hypotheses*

$$H_0 : \theta = 6; \qquad H_1 : \theta > 6\,.$$

at the 10% level of significance. Which of the statistics, the sample mean $\bar{X}$ and the sample maximum Z, gives the better test?

Solution:

3.7 Categorical data

This next section of notes is 'self-study'. The reason why this section is 'self-study' is to provide you with an opportunity to practise and enhance your independent mathematical reading. To accompany this section, a video has been placed on Learning Central summarising the main points. This material is fully assessable, and you should not leave it until the last minute to assimilate this material.

3.7.1 Introduction

The vast majority of the hypothesis tests investigated so far assume that the data is sampled from a normally distributed random variable. When the random variable is specified as something else, such as the Poisson distribution, or even if it is unknown, we can use the Central Limit Theorem to construct an approximate hypothesis test.

We will now study hypothesis tests designed to investigate experiments where the outcome can fall into one of a finite number of categories.

3.7.2 The Chi-square test

Suppose that the result of an experiment is classified into one of k categories $c_1, c_2, \ldots, c_k$, and that for a given model the expected numbers of outcomes in each category are $E_1, E_2, \ldots, E_k$ respectively. If $O_1, O_2, \ldots, O_k$ are the observed numbers in each category, the chi-square statistic is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The statistic χ^2 has a distribution which is approximately chi-square with $k - m - 1$ degrees of freedom, where m is the number of parameters that need to be estimated from the data to

fit any theoretical distribution. The approximation is good, provided $E_i > 5$ for all i. If any $E_i < 5$, the outcomes may be grouped to ensure that all $E_i \geq 5$.

Example 3.10 *A die is rolled 60 times, and the outcome of this experiment is given below. Test at both a 5% and 1% significance level the null hypothesis H_0 that the die is fair*

Face	*1*	*2*	*3*	*4*	*5*	*6*
Outcome O_i	*15*	*7*	*4*	*11*	*6*	*17*

Solution:
Here we test the null hypothesis that the die is fair, against the alternative hypothesis that the die is unfair. If the null hypothesis is true, then we would expect to see, after rolling the die 60 times, 10 of each possible outcome. Now due to experimental variation, it is unlikely given 60 rolls of a die, to perfectly observe 10 ones, 10 twos, etc. So given the data we have above, is it reasonable that the die is fair?

In the previous paragraph we have motivated how to calculate the E_i, the expected outcome for each category under the null hypothesis. Thus we can construct the following table

Face	1	2	3	4	5	6
Outcome O_i	15	7	4	11	6	17
Expected E_i	10	10	10	10	10	10

So we compute the test statistic as

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 13.6\,.$$

We now compare the value of the test statistic with percentage points of the chi-square distribution, with 5 degrees of freedom

- The 95th percentage point of the chi-square distribution with 5 degrees of freedom is 11.070. So at 5% significance, we reject the null hypothesis that the die is fair.
- The 99th percentage point of the chi-square distribution with 5 degrees of freedom is 15.086. So at 1% significance, we accept the null hypothesis that the die is fair.

It is possible to write χ^2 in an easier computational form. Prove the following theorem yourself, in the space provided.

Theorem 3.11 *Let N be the total number of observations (*in the previous example $N = 60$*). Then*

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - N\,.$$

Proof:

Example 3.12 *100 squares each 1 meter square were randomly placed in a field where daffodils were growing. The number of clumps of daffodils in each square was counted and the following table of these observed numbers was drawn up.*

	0	*1*	*2*	*3*	*4*	*5*	$\geq$ *6*
O_i	*4*	*17*	*20*	*22*	*15*	*15*	*7*

A Poisson distribution is suggested as a model for the number of clumps of daffodils in a square. Estimate an appropriate value for the mean of this distribution using the data in the table above. Calculate a table of expected numbers to fit the Poisson distribution to these data and perform a chi-square goodness of fit test to determine if the model is a good fit to the observed data.

Solution:
The total is given by $(0 \times 4) + (1 \times 17) + \ldots + (6 \times 7) = 300$. Hence an estimate of the mean is $300/100 = 3$.

The formula for the probability density function for a Poisson distribution is

$$f(x;\mu) = \mu^x \exp(-\mu)/x!\,,$$

where μ is the mean. We need to estimate a value for μ to use this probability density function, hence we estimate μ by the sample mean above, that is $\hat{\mu} = 3$.

We can then use $f(x;3)$ to obtain the following table of expected numbers, by multiplying the Poisson probabilities by 100. Note that the expected number for the last class $x \geq 6$, is found by subtracting the sum of the expected numbers from 0 to 5 inclusive from 100.

	0	1	2	3	4	5	≥ 6
O_i	4	17	20	22	15	15	7
E_i	4.98	14.94	22.40	22.40	16.80	10.08	8.39

From this table the chi-square goodness of fit test statistic is calculated

$$\chi^2 = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} = 3.5658.$$

Since one parameter was estimated to derive the expected numbers, the degrees of freedom are 5. There are 7 classes in total, so we calculate the degrees of freedom by taking one away from 7, but then take another one away as we have estimated μ leaving 5.

From statistical tables the 90th percentile of the chi-square distribution with 5 degrees of freedom is 9.236 thus we conclude that there is no significant discrepancy between the fitted numbers and the observed ones, i.e. the Poisson distribution is a good fit to the data.

3.7.3 Contingency tables

In the previous example, the probabilities used in the null hypothesis were obtained from theoretical considerations. We often want to test the hypothesis that effects are independent without having a theory to predict the relevant probabilities. In this case, the probabilities must be estimated from the contingency table.

Contingency means dependence, so a contingency table is simply a table that displays how two or more characteristics depend on each other. For example, a contingency table, where Effect 1 has three categories I, II, III and Effect 2 has four categories A, B, C, D is:

	A	B	C	D	
I	O_{11}	O_{12}	O_{13}	O_{14}	R_1
II	O_{21}	O_{22}	O_{23}	O_{24}	R_2
III	O_{31}	O_{32}	O_{33}	O_{34}	R_3
	C_1	C_2	C_3	C_4	N

where the O_{ij} is the number of observations observed in the ith row and jth column, R_i are row totals, C_j are column totals, and N is the total number of observations.

Under the null hypothesis that the effects are independent, the probability associated with cell (i, j) is estimated by:

$$P_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} = \frac{R_i C_j}{N^2}$$

and hence the expected number for each cell (i, j) is given by:

$$E_{ij} = N \times P_{ij} = \frac{R_i C_j}{N}\,.$$

The statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$$

where r is the number of rows, and c is the number of columns follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Example 3.13 *The distribution of five plant species A, B, C, D, E is being investigated at three different locations I, II, III. The contingency table of the results is presented below.*

	A	B	C	D	E
I	*10*	*22*	*38*	*8*	*66*
II	*27*	*62*	*120*	*30*	*200*
III	*45*	*100*	*207*	*49*	*342*

Use an appropriate statistical test to address the problem of whether the species of plant is independent of the location.

Solution:
To estimate the probabilities, and thus estimate the expected values, we need to compute row and column totals.

	A	B	C	D	E	
I	10	22	38	8	66	144
II	27	62	120	30	200	439
III	45	100	207	49	342	743
	82	184	365	87	608	1326

Under the null hypothesis of independence (here between plant species and location) we estimate the expected values using the formula given above, giving the following table of expected values:

	A	B	C	D	E	
I	8.9	20.0	39.6	9.4	66	144
II	27.1	60.9	120.8	28.8	201.3	439
III	45.9	103.1	204.5	48.7	340.7	743
	82	184	365	87	608	1326

The test statistic is computed as

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{O_{ij}^2}{E_{ij}} - 1326 = 1.14$$

The $95th$ percentage point of the chi-square distribution with $(r-1)(c-1) = 2\times 4 = 8$ degrees of freedom is 15.507, so the null hypothesis that the distribution of the plants is independent of location, is accepted.

This is the end of the self-study section. To complete your study and to evaluate your understanding of this material, look through some text books to find your own examples. Attempt them, and write up your solutions.

3.8 One-way ANOVA: One-way Analysis of Variance

3.8.1 Introduction and motivation

In this section the analysis is described of an experiment that has several groups of observations. These could either be different levels of the same factor, such as concentrations of a chemical, or several different factors that might affect the observations made. The main objective is to determine if there are significant differences amongst the means of the levels, but the analysis is based on an examination of variation, and is often called the one-way analysis of variance. The one-way factorial experiment can be viewed as an extension of hypothesis tests for two means, to an experiment where there are more than two groups to be compared. The ideas presented in this section can be extended to factorial experiments, where several factors are measured simultaneously at all combinations of levels. These multi-factorial experiments will not be described in this module.

A valid question is "why do we need to construct another hypothesis test to compare the difference of three or more means, when we can do lots of tests comparing pairs of means?". Remember that every single statistical test that you perform runs the risk of making a Type I error. If you do lots of hypothesis tests on the same data, then you accumulate more and more risk of making a Type I error. Hence, if we can do one statistical test that answers all our hypothesis simultaneously, then this is often preferable.

3.8.2 Notation and set-up

Suppose there are m groups of data. The notation for the jth observation in the ith group is x_{ij}.

There can be different numbers of observations from group to group, so the number of observations in the ith group is n_i.

So j runs from 1 to n_i, in the ith group, and i runs from 1 to m.

Thus the data might be arranged as:

$$\begin{array}{rccccc} \text{Group 1:} & x_{11} & x_{12} & \ldots & x_{1n_1} \\ \text{Group 2:} & x_{21} & x_{22} & \ldots & x_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{Group } m\text{:} & x_{m1} & x_{m2} & \ldots & x_{mn_m} \end{array}$$

The null hypothesis here is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$, where μ_i is the population mean corresponding to the ith group. The alternative hypothesis is $H_1 : \mu_a \neq \mu_b$, where $a \neq b$.

Let us introduce the following notation.

3.8.3 Possible sources of variation

The following sketch, for $m = 3$, helps us to identify all the sources of variation possible for our data.

We thus have three sums of squares:

1. Total sum of squares
2. Between group sum of squares
3. Within group sum of squares (also known as error, or residual sum of squares)

We now describe each in turn.

Total sum of squares.

Between group sum of squares.

Within group sum of squares.

3.8.4 ANOVA table

The comparison of sums of squares is done in a table called the analysis of variance table, because the variation in the observations is being partitioned into part that is explained by the data because they are sampled from different groups (the between groups component), and a part that remains unexplained (the residual component).

Source of variation	df	ss	ms	F-ratio
Between groups	$m-1$	$\sum_{i=1}^{m} \frac{X_{i\cdot}^2}{n_i} - \frac{X_{\cdot\cdot}^2}{N}$	s_G^2	$\frac{s_G^2}{s^2}$
Within groups (error or residual)	$N-m$	$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{m} \frac{X_{i\cdot}^2}{n_i}$	s^2	-
Total	$N-1$	$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{X_{\cdot\cdot}^2}{N}$	-	-

The mean squares (ms) are calculated by dividing the sum of squares (ss) by degrees of freedom (df). The F-ratio is the ratio of between groups mean square to error mean square.

Lemma 3.14 *Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$, then the F-ratio $\frac{s_G^2}{s^2}$ follows an F-distribution with $(m-1)$ and $(N-m)$ degrees of freedom.*

The percentiles of the F distribution with $(m-1)$ and $(N-m)$ degrees of freedom are used to judge the significance of the observed F-ratio. If the observed F-ratio exceeds the tabulated 95th percentile, for example, then with a 5% level of significance the null hypothesis (of equal group means) is rejected.

3.8.5 Multiple comparison tests

Although one-way ANOVA can be used to identify that there are statistically significant differences in the means of m groups of data, it does not identify which particular groups differ in this respect from the others.

Various tests have been suggested to do this. The simplest to describe is Fisher's least significant difference (LSD) test.

Fisher's least significant difference test is described as follows.

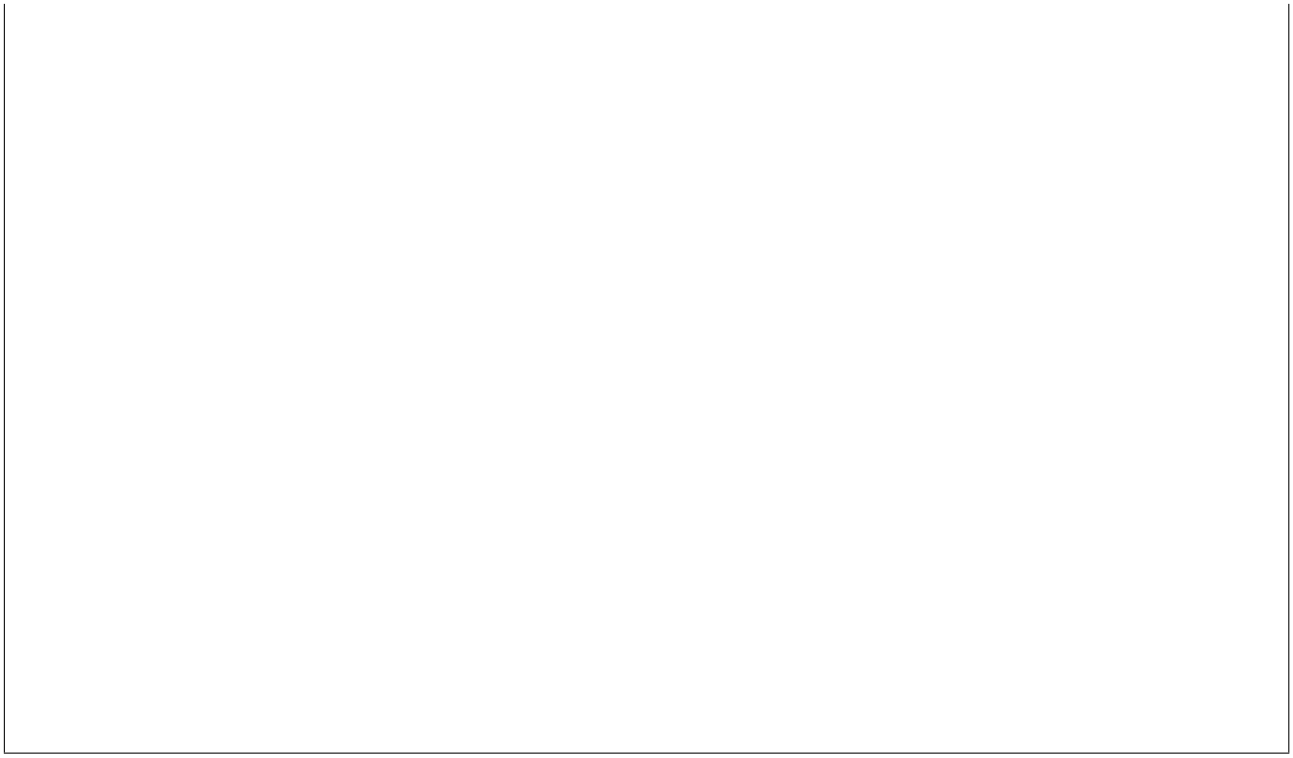
3.8.6 Example

Example 3.15 *The systolic blood pressure was measured for 7 human subjects, with 5 replicate observations being made for each subject. The results obtained (in mmHg) are tabulated below, together with some totals.*

Subject	*Systolic blood pressure*	*Totals*
1	*108 104 108 120 108*	*548*
2	*118 104 118 120 128*	*588*
3	*124 118 120 122 124*	*608*
4	*126 122 120 120 132*	*620*
5	*128 124 118 112 142*	*624*
6	*130 128 138 116 124*	*636*
7	*152 128 132 150 142*	*704*
		4328

Complete an analysis of variance table for these data and test to determine if there are significant differences amongst the means of the subjects. Follow up your analysis by attempting to identify which groups differ from which.

Solution:



4 Linear regression (fitting a straight line)

4.1 Introduction

We now consider data consisting of pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where the x and y values may be related. We use the data to investigate the nature of the relationship between the two variables.

Specifically we will consider that the observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ follows a linear relationship of the form

$$Y = \alpha + \beta x + \varepsilon$$

where α and β are constants and ε is a random error term with zero mean. Note that the errors are assumed to be independent. This is equivalent to assuming that $E[Y] = \alpha + \beta x$ so that the mean (or expected value) of Y is a linear function of x.

In general the variable x is called the independent variable and the variable Y is called the dependent variable. Examples include the following:

x	Y
Amount spent advertising a product	Sales of the product
Age	Height
Distance	Time to run distance
Weight of fertilizer	Tomato yield

The primary aim of regression is to estimate the parameters α and β. We consider this in the next section.

4.2 Least squares regression

Suppose we have n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from the model

$$Y = \alpha + \beta x + \varepsilon$$

where α (the intercept) and β (the slope) are constants and ε is a random error term with zero mean.

Here is a sketch of typical data:

The principle of least squares regression, in this scenario is to find α and β such that

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

is as small as possible. The values of α and β which minimize this sum are known as least squares estimates.

Theorem 4.1 *Suppose we have n pairs of observations* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *from the model*

$$Y = \alpha + \beta x + \varepsilon$$

where α *and* β *are constants and* ε *is a random error term with zero mean. The least squares estimators of* α *and* β*, denoted* $\hat{\alpha}$ *and* $\hat{\beta}$ *are given by:*

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,. \end{aligned}$$

Proof:

If we write:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})x_i = \sum_{i=1}^{n}x_i^2 - \frac{(\sum_{i=1}^{n}x_i)^2}{n} = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \\ S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i = \sum_{i=1}^{n}x_i(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - \frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n} \\ &= \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} \end{aligned}$$

where S_{yy} is defined analogously to S_{xx} with the obvious changes, then

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\,.$$

The minimum value of $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$ is given by

$$\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2,$$

and is known as the residual sum of squares. We denote this by SS_{error}.

Lemma 4.2 *Let $\hat{\alpha}$ and $\hat{\beta}$ be defined as in Theorem 4.1. The residual sum of squares, SS_{error} may be written*

$$SS_{error} = \sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Try to prove this yourself. We will use this result later.

4.3 Properties of the least squares estimators: distributions

Recall that we have n pairs of observations from the model $Y_i = \alpha + \beta x_i + \varepsilon_i$.

Let Y_i denote the random variable whose observed value is y_i. It is a random variable due to the addition of the random error term ε_i. To consider properties of the estimator $\hat{\beta}$ (which will enable us to consider properties of the estimator $\hat{\alpha}$ as well as the predicted value of y given x) we write

$$\hat{\beta} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}}.$$

Lemma 4.3 *We may write*

$$\hat{\beta} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})Y_i}{S_{xx}}.$$

We use this form to derive properties of $\hat{\beta}$ (and other parameters).

Proof:

In order to construct confidence intervals and hypothesis tests for estimators of the parameters of the model, and for predictions of y using it, we need to make some distributional assumptions concerning the random error term ε_i of the model:

1. They are all independent.
2. They have zero expectation.
3. They all have the same variance σ^2 (independent of x and i).
4. They are normally distributed.

Straight away, we can derive the following facts using the above assumptions.

$$\begin{aligned} E[Y_i] &= E[\alpha + \beta x_i + \varepsilon_i] = E[\alpha + \beta x_i] + E[\varepsilon_i] = \alpha + \beta x_i, \\ Var[Y_i] &= Var[\alpha + \beta x_i + \varepsilon_i] = Var[\alpha + \beta x_i] + Var[\varepsilon_i] = \sigma^2. \end{aligned}$$

We will use these facts in what follows.

4.3.1 Mean and variance of $\hat{\beta}$

4.3.2 Mean and variance of $\hat{\alpha}$

4.3.3 Mean and variance of $\hat{y_0}$

Suppose we are given a value of x, say x_0. We may be asked to estimate the mean response y_0 at this given value. Since $y_0 = \alpha + \beta x_0$ the obvious estimator of y_0 is $\hat{y_0} = \hat{\alpha} + \hat{\beta} x_0$. We can consider the mean and variance of this estimate.

An examination of the expressions for $\hat{\beta}$, $\hat{\alpha}$ and $\hat{y_0}$ shows that they are all expressible as linear combinations of $Y_1, Y_2, \ldots, Y_n$. It follows that if the ε_i's are normally distributed then so are $\hat{\beta}$, $\hat{\alpha}$ and $\hat{y_0}$. In this case we can summarise the preceding results by stating that

$$\begin{aligned} \hat{\beta} &\sim N\left[\beta, \frac{\sigma^2}{S_{xx}}\right], \\ \hat{\alpha} &\sim N\left[\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right], \\ \hat{y_0} = \hat{\alpha} + \hat{\beta}x_0 &\sim N\left[y_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right]. \end{aligned}$$

These can be used to make confidence intervals and hypothesis tests about these parameters. You will be expected to do these! Derive the confidence intervals for yourself, and think how you could perform hypothesis tests using these results.

4.4 Estimating the error variance σ^2

The value of the error variance σ^2 is usually unknown and must be estimated from the data.

Earlier we stated that the residual sum of squares, SS_{error} may be written

$$SS_{error} = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Lemma 4.4

$$\frac{SS_{error}}{n-2}$$

is an unbiased estimator of σ^2.

It follows from earlier results that if the errors are normally distributed then the random variables

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}, \frac{\hat{\alpha}-\alpha}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}, \frac{\hat{y_0}-y_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}}$$

are have the distribution $N[0,1]$.

It can be shown that when σ^2 is estimated by

$$\hat{\sigma^2} = \frac{SS_{error}}{n-2}$$

then the random variables

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma^2}}{S_{xx}}}}, \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma^2}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}, \frac{\hat{y_0}-y_0}{\sqrt{\hat{\sigma^2}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}}$$

all have the Student's t distribution with $(n-2)$ degrees of freedom. Again these can be used to make confidence intervals, or to perform hypothesis tests in the usual way.

4.5 Example

Example 4.5 *The following table gives measurements of two variables x and y that are known to be linearly related. The variable x is measured without error, but there is a measurement error associated with y that you can assume is normally distributed with zero mean and variance σ^2.*

x	*5.0*	*7.5*	*10.0*	*12.5*	*15.0*
y	*1.23*	*1.39*	*1.52*	*1.66*	*1.81*

Find the least squares estimates of the slope and intercept of the straight line to predict y with x. Also calculate an unbiased estimate of the error variance σ^2. Find a 95% confidence interval for the intercept.

Solution: