Math 494 - Mathematical Statistics

Why statistics?

The world is swimming with data.

We need to be able to analyse data, make conclusions and hence make decisions.

Phrases such as: *margin of error*, *statistically significant*, *beyond reasonable doubt* are things you'll see in the media, but as mathematicians we are going to define exactly what these things mean.

Why mathematical statistics?

Our aim is to clearly define a mathematical framework for statistics.

It is one thing to take our average weight of ten chocolate pieces, but then what? What can we say about the true mean of the population?

Do we need to assume anything for our conclusions to be valid?

How do we check out assumptions?

Random Sampling

A random sample on a random variable X is a sequence of independent random variables $X_1, X_2, \ldots, X_n$ each having the same distribution X. If for example our distribution of X is continuous, then the pdf of the joint distribution would be

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_X(x_1)f(x_2)\ldots f_X(x_n)$$

Notation

We denote the underlying random variables of our sample as $X_1, X_2, \ldots, X_n$.

Realisations, that is, the actual observed values are denoted with lower case letters $x_1, x_2, \ldots x_n$.

The goal in statistics is normally to gain an understanding about the distribution X based upon a random sample. This is often termed as making *inference* on X.

We may wish to estimate measures of location, or spread, or a variety of characteristics of the distribution X.

Statistic

To achieve this, the only thing you could do is to take a function of your sample. A *Statistic* is a function of your sample:

$$T = \psi(X_1, X_2, \ldots X_n),$$

and hence it's realisation is

$$t = \psi(x_1, x_2, \ldots x_n).$$

Statistics and estimators

All statistics can be considered as an estimator of a particular characteristic of the population, usually referred to as a parameter.

For example, what did you sample mean of 10 chocolate blocks estimate?

If our statistic T is an estimator of some parameter θ, then in order to make inference on θ we will need to understand the distribution of T.

The sample mean

The sample mean is often denoted as $\hat{\mu}$ or $\bar{x}$.

Note that for any parameter θ, $\hat{\theta}$ typically denotes an estimate of θ.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample standard deviation

The sample standard deviation is usually denoted by $s = \hat{\sigma}$.

The most convenient form to compute s is

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

Distributions of statistics

Typically the mean is the parameter of most interest.

Typically we estimate it using the sample mean.

Question is then, now what? How do we interpret this number?

Distribution of the sample mean

The first natural question is:

What is the mean and standard deviation of the random variable $\bar{X}$?

$$\mathbb{E}(\bar{X}) = \ldots = \mu$$
$$\mathrm{Var}(\bar{X}) = \ldots = \frac{\sigma^2}{n}.$$

So we 'know' the mean and the variance of $\bar{X}$. That doesn't actually tell us what the distribution is though...

Perhaps when n is large we can say something about the distribution of $\bar{X}$?

Asymptotic distributions

In general, finding the exact distribution of $\bar{X}$ when n is large is an incredibly hard problem.

Suppose the distribution of $\bar{X}$ converges to a nice limit as $n \to \infty$. Perhaps we can use that then as an approximation.

Question is, what do we mean by convergence in this scenario?

Convergence in distribution

A sequence of random variables $\{Y_n\}$ converges to Y in distribution, denoted by $Y_n \xrightarrow{d} Y$ if

$$F_{Y_n}(x) \to F_Y(x) \quad \text{as } n \to \infty,$$

for all values x which are continuity points of Y.

It is sufficient to show that the moment generating function of Y_n converges to the moment generating function of Y. (We will not prove why this is sufficient.)

The central limit theorem

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $V(X_i) = \sigma^2$, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty.$$

In other words for large n, $S_n \overset{d}{\approx} N(n\mu, n\sigma^2)$ or $\overline{X} \overset{d}{\approx} N(\mu, \frac{\sigma^2}{n})$.

Example

If $X \overset{d}{=} \exp(1)$, then $\mu = 1$ and $\sigma^2 = 1$. Given a random sample of $n = 100$ observations of X, then

- What is an approximate distribution of $\bar{X}$?
- What is an approximate probability for $\mathbb{P}(0.9 < \bar{X} < 1.1)$?

The sample variance

Recall:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right).$$

So, what is the expectation of this? Does $\mathbb{E}(S^2) = \sigma^2$?

Why the $n-1$ in the denominator?

Variance of S^2

It can also be shown (with a great deal of effort) that

$$\mathrm{Var}(S^2) = \frac{\nu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)},$$

where $\nu_4 = \mathbb{E}((X-\mu)^4)$.

Note that there isn't an asymptotic result for the sample variance similar to the central limit theorem for the sample mean. The only case where we can say something about the distribution of S^2 is when sampling from a normal distribution.

Distributions

The mean and the variance often tell us a great deal about the distribution, but sometimes not a lot about the shape of the distribution.

Recall that the cdf function F, completely characterises a distribution.

Order statistics

Given a sample $x_1, x_2, \ldots x_n$, if we re-order them in increasing value: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ so that $x_{(1)}$ denotes the smallest observation, then $x_{(k)}$ is called the kth *order statistic*.

Sample cdf

The sample cdf (also known as the empirical cdf) can be defined in terms of the order statistics as follows:

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ \frac{k}{n} & \text{if } x_{(k)} \leq x < x_{(k+1)}, \\ 1 & \text{if } x \geq x_{(n)} \end{cases}$$

Quantiles vs. Order statistics

The above expression for the sample cdf allows us to relate sample quantiles to order statistics.

$$\hat{c}_q = x_{(k)} \qquad \text{if} \;\; \frac{k-1}{n} < q < \frac{k}{n},$$

as long as nq is not an integer. If nq is an integer, then

$$\hat{c}_q = \frac{1}{2}(x_{(qn)} + x_{(qn+1)}).$$

Note that there are alternative definitions for sample quantiles.

Box plots

A box plot is a useful way for representing a distribution. A box plot represents five features of the sample: the extremes, the quartiles and the median.

Art!

Distribution of order statistics

Given the pdf f_X, The joint pdf of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f_X(x_1) f_X(x_2) \ldots f_X(x_n)$$

The distribution of $X_{(k)}$ is

$$f_{X_{(k)}}(u) = k \binom{n}{k} F_X(u)^{k-1} (1 - F_X(u))^{n-k} f_X(u)$$

Approximate distribution for quantiles

We can derive rough/suspicious approximations for the mean and variance of our sample quantiles for a continuous random variable.

How?

When in doubt, Taylor expand.

We have

$$\begin{aligned} F(\hat{C}_q) &\approx F(c_q) + (\hat{C}_q - c_q) f(c_q) \\ \Rightarrow \hat{F}(\hat{C}_q) &\approx \hat{F}(c_q) + (\hat{C}_q - c_q) f(c_q) \end{aligned}$$

Since $\hat{F}(\hat{C}_q) \approx q$, some rearrangement gives

$$\hat{C}_q \approx c_q - \frac{\hat{F}(c_q) - q}{f(c_q)}$$

A little more work gives

$$\begin{aligned}\mathbb{E}(\hat{C}_q) &\approx c_q \\ \mathrm{Var}(\hat{C}_q) &\approx \frac{q(1-q)}{nf(c_q)^2}.\end{aligned}$$

Furthermore, it can be shown that as $n \to \infty$,

$$\hat{C}_q \overset{d}{\approx} \mathrm{N}\left(c_q, \frac{q(1-q)}{nf(c_q)^2}\right).$$

Sampling from a normal distribution

We start by considering the special case where $x_1, x_2, \ldots x_n$ are i.i.d. random variables with distribution $\mathrm{N}(\mu, \sigma^2)$.

In this case we know the distribution of the sample mean is exactly normal:

$$\bar{X} \overset{d}{=} \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

χ^2_n distribution

If $Z_1, Z_2, \ldots, Z_n$ are all iid standard normal random variables, and $U = \sum_{i=1}^n Z_i^2$, then we say U has a χ^2 distribution with n degrees of freedom.

We often write $U \overset{d}{=} \chi^2_n$.

It can be shown that the mgf of U is given by

$$M_U(t) = (1 - 2t)^{-\frac{n}{2}}$$

Distribution of S^2

Noting that it can be shown that S^2 and $\bar{X}$ are independent, we can show

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2,$$

and hence using moment generating functions

$$\frac{(n-1)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-1}$$

QQ-plots

QQ-plots, which is short for quantile-quantile plots, are used to check the assumption that the data follow a certain distribution.

The fundamental idea of a QQ-plot is that if our data follows a particular distribution, then the sample/empirical cdf should be similar to the theoretical cdf. Or alternatively, the sample quantiles, should be similar to the quantiles of the assumed distribution.

A normal QQ-plot would plot the points:

$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right).$$

We also know that $x_{(i)} \approx \hat{c}_q$, for $q = \frac{i}{n+1}$ and hence we expect that $x_{(i)} \approx \mu + \sigma \Phi^{-1}\left(\frac{i}{n+1}\right)$.

Hence, the intercept and the slope can be seen as estimates for the mean and standard deviation.

Note that a QQ-plot can be used for any distribution F by plotting the points

$$\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right).$$

Estimation

In statistics, we often assume that the form of the distribution of X is known, but that the parameters are unknown.

For example, we can fit a QQ-plot, decide that the data does follow some normal distribution, but it still remains to estimate the mean and variance.

Suppose that the distribution of X depends on the parameter θ, then given a random sample on X, we wish to estimate θ. An *estimator* of θ is a statistic $T_n = \psi(X_1, X_2, \dots X_n)$.

There are clearly infinitely many possible estimators one could choose, how do we decide if an estimator is 'good'?

Unbiasedness

An estimator T_n is an *unbiased* estimator for θ if $\mathbb{E}(T_n) = \theta$.

For example, we saw earlier that our estimator for sample mean and variance were unbiased.

Consistency

An estimator T_n is a *consistent* estimator for θ if $T_n \xrightarrow{p} \theta$ as $n \to \infty$.

A useful result regarding consistency:

If $\mathbb{E}(T_n) \to \theta$ and $\text{Var}(T_n) \to 0$, then $T_n \xrightarrow{p} \theta$.

This can be easily proven using Chebyshev/Bienayme's inequality. Note that the converse is not true.

Efficiency

As estimator T_n is more *efficient* than another estimator T_n' if $\mathrm{Var}(T_n) < \mathrm{Var}(T_n')$.

It can be shown that for any unbiased estimator of θ, there exists a lower bound for the variance of the estimator. This lower bound is known as the minimum variance bound (MVB), which we will consider later. Using MVB, we define the efficiency of T_n as

$$eff(T_n) = \frac{MVB}{\mathrm{Var}(T_n)}$$

If $eff(T_n) = 1$ then T_n is said to be an efficient estimator.

Examples

- If $\theta = \mathbb{E}(X)$ then $\bar{X}$ is consistent and unbiased for θ.
- If $X \overset{d}{=} R(0, \theta)$, consider the two estimators of θ: $U_n = 2\bar{X}_n$ and $V_n = \frac{n+1}{n} X_{(n)}$.

It should be noted that just because an estimator is biased, does not necessarily mean it is a bad estimator.

Art!

Point estimation

An estimate of θ is the observed value of the estimator. This then gives us a 'point' estimate for the unknown value of θ.

For example, given a sample on $\mathrm{N}(\theta, 5)$,

```
10.26 9.17 13.34 10.36 2.83 6.16 13.86
```

an estimate for θ is $\bar{x} = 9.43$.

But we know so much more! For example, we in theory know the distribution of sample means.

Given a sample, and our point estimate, what else can we say then?

Interval estimation

If we assume that we have an estimator T_n for θ for which we know the distribution, then we could make statements of the form

$$\mathbb{P}(c_{0.025}(T_n) < T_n < c_{0.975}(T_n)) = 0.95.$$

T_n depends on θ, so if we write $a(\theta) = c_{0.025}(T_n)$ and $b(\theta) = c_{0.975}(T_n)$, then

$$\mathbb{P}(a(\theta) < T_n < b(\theta)) = 0.95$$

and if we invert this (usually a and b are increasing functions), then

$$\mathbb{P}(b^{-1}(T_n) < \theta < a^{-1}(T_n)) = 0.95.$$

Confidence interval

This says that the random interval $(b^{-1}(T_n), a^{-1}(T_n))$, has a probability of 0.95 of containing θ.

We call this a 95% confidence interval for θ.

Example

Suppose we our sample mean of the chocolate pieces was 35.

For now assume that we know the true standard deviation is 5. We will relax this assumption soon.

Example

If $X \overset{d}{=} R(0, \theta)$, then $X_{(n)}$ has cdf $F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n$ for $0 < x < \theta$.

Calculate a 95% confidence interval for θ based upon $X_{(n)}$.

Proportions

If a value of Z where $Z \overset{d}{=} \mathrm{Bi}(n, p)$, is observed for large n, we wish to obtain a 95% confidence interval for p.

For large n, an approximate 95% confidence interval is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Prediction Intervals

A confidence interval tells you information about a parameter of a distribution. However, it does not tell you anything about future observations.

An interval that is likely to contain a future observation is called a *prediction interval*.

Example

Suppose we have a random sample on $X \overset{d}{=} \mathrm{N}(\mu, \sigma^2)$. Let X^* denote a future observation that is independent of our sample $X_1, X_2, \ldots X_n$.

Now $X^* \overset{d}{=} \mathrm{N}(\mu, \sigma^2)$ and $\bar{X} \overset{d}{=} \mathrm{N}(\mu, \sigma^2/\mathrm{n})$. Therefore

$$X^* - \bar{X} \overset{d}{=} \mathrm{N}(0, \sigma^2(1 + 1/\mathrm{n})).$$

And hence,

$$\mathbb{P}\left(\bar{X} - 1.96\sigma\sqrt{1 + \frac{1}{n}} < X^* < \bar{X} + 1.96\sigma\sqrt{1 + \frac{1}{n}}\right) = 0.95.$$

We should take note of the difference between the limits of confidence intervals vs. prediction intervals as $n \to \infty$.

The prediction interval tends towards $\mu \pm 1.96\sigma$, while the confidence interval tends towards the single point μ.

Inference on a normal population

If we have a random sample on $\mathrm{N}(\mu, \sigma^2)$, then the ‘best’ unbiased estimators for μ and σ^2 are the sample mean and the sample variance. We will show why these estimators are the ‘best’ in the following weeks.

If we assume that σ^2 is known, then it is a relatively simple exercise to make inference about μ.

Example

The volume of beer that our ‘buddy’ Jeff (the alcoholic) consumes on a given day has a standard deviation of 400ml. Over a four week period, Jeff drank 70 litres of beer. If we assume that the amount that Jeff drinks on a given day is normally distributed, then find a 95% confidence interval for the average amount of beer Jeff drinks in a day.

Note that the normality assumption is not unreasonable since for large enough n, the central limit theorem kicks in.

Not assuming σ^2 is known

The thing is, in practice does one ever actually known σ^2?

σ^2 is almost always also a parameter that needs to be estimated.

When we knew σ^2 we would use the fact that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{=} \mathrm{N}(0, 1).$$

The obvious solution to not knowing σ^2 is to substitute the unknown σ^2 with it's estimator S^2.

t_ν distribution

If $U \overset{d}{=} \chi^2_\nu$, and $Z \overset{d}{=} \mathrm{N}(0,1)$ and U, Z are independent then $T = \frac{Z}{\sqrt{U/\nu}}$ is said to follow the t-distribution with ν degrees of freedom.

It ‘can be shown’ that T has pdf

$$f_T(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

The most useful fact about the t_ν distribution is that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{d}{=} t_{n-1},$$

and hence we can use the t-distribution to construct confidence intervals.

Example

Chocolate revisited!

Comparison of populations

So we now have (some) tools to make inference about one population.

A more common occurence in practice is that we have two populations that we wish to compare.

For example, are American's more fat than Australians?

Comparison of normal populations

Suppose we have independent samples of two normal random variables

$$\begin{aligned} &n_1 \text{ observations on } X_1 \overset{d}{=} \mathrm{N}(\mu_1, \sigma_1^2): \quad X_{11}, X_{12}, \ldots, X_{1n_1} \\ &n_2 \text{ observations on } X_2 \overset{d}{=} \mathrm{N}(\mu_2, \sigma_2^2): \quad X_{21}, X_{22}, \ldots, X_{2n_2}, \end{aligned}$$

based upon these observations, we wish to make comparisons between the variances σ_1^2 and σ_2^2, and the means μ_1 and μ_2.

The F-distribution

If $U \overset{d}{=} \chi^2_m$ and $V \overset{d}{=} \chi^2_n$ and U, V are independent, then $Z = \frac{U/m}{V/n}$ is said to have F-distribution with m and n degrees of freedom.

We write $Z \overset{d}{=} F_{m,n}$.

It 'can be shown' that for $z > 0$

$$f_Z(z) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B(\frac{m}{2}, \frac{n}{2})} \frac{z^{\frac{m}{2}-1}}{(n + mz)^{\frac{m+n}{2}}},$$

where B denotes the Beta function.

Now recalling that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \overset{d}{=} \chi^2_{n_1-1}, \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \overset{d}{=} \chi^2_{n_2-1}$$

Hence, from the definition of the F-distribution,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \overset{d}{=} F_{n_1-1,n_2-1},$$

and we can use this result to find confidence intervals for σ_1^2/σ_2^2.

Example

Given $n_1 = 25, s_1^2 = 1.27, n_2 = 7, s_2^2 = 2.92$, find a two sided 95% confidence interval for σ_1^2/σ_2^2.

Well, we have $\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \overset{d}{=} F_{24,6}$. Hence we can find

$$\mathbb{P}\left(0.33 < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < 5.11\right) = 0.95$$

And therefore,

$$\mathbb{P}\left(\frac{1}{5.11}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{0.33}\frac{S_1^2}{S_2^2}\right) = 0.95$$

Comparison of means

The first thing to note is

$$\bar{X}_1 - \bar{X}_2 \overset{d}{=} \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{\mathrm{n}_1} + \frac{\sigma_2^2}{\mathrm{n}_2}\right).$$

So the distribution of this depends upon σ_1^2 and σ_2^2 and the assumptions that we make about the variances will affect this distribution, and hence the inference we make on $\mu_1 - \mu_2$.

Known variances

If we know the variances σ_1^2, σ_2^2, then constructing a confidence interval is essentially the same as constructing a one sample confidence interval.

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{d}{=} \mathrm{N}(0, 1)$$

Variances unknown but equal

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then we have

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{d}{=} \mathrm{N}(0,1)$$

Similarly to the one sample case, we would expect that if we replace σ^2 with its estimator S^2, then we would result in a t-distribution.

If we chose

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

we can show

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{d}{=} \mathrm{t}_{n_1+n_2-2}.$$

Note that this estimator S_p^2 is often called the *pooled* estimator for σ^2.

Example

Suppose samples of X_1 and X_2 are as follows:

$n_1 = 10, \bar{x}_1 = 4.3, s_1^2 = 2.2$ and $n_2 = 20, \bar{x}_2 = 3.7, s_2^2 = 2.5$.

Construct a confidence interval for $\mu_1 - \mu_2$ assuming equal variances.

Variances unknown and unequal

In this case we would use

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

The exact distribution of this, is intractable, but can be reasonably approximated using a t-distribution with k degrees of freedom where

$$\frac{1}{k} = \frac{\beta^2}{n_1 - 1} + \frac{(1-\beta)^2}{n_2 - 1},$$

where $\beta = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

This approximation is derived from choosing k such that the variance of $\frac{S_1^2/n_1+S_2^2/n_2}{\sigma_1^2/n_1+\sigma_2^2/n_2}$ is the same as the variance of χ_k^2/k.

Example

Suppose samples of X_1 and X_2 are as follows:

$n_1 = 10, \bar{x}_1 = 4.3, s_1^2 = 2.2$ and $n_2 = 20, \bar{x}_2 = 3.7, s_2^2 = 2.5$.

Construct a confidence interval for $\mu_1 - \mu_2$ without assuming equal variances.

Methods of estimation

So far we have seen numerous estimators, and properties of estimators. However, where exactly do these estimators come from?

What we need is a framework for deriving estimators for the problem at hand.

Method of moments (MM)

The method of moments essentially uses the moments of your sample to estimate the parameters of interest.

If $\mathbb{E}(X) = \phi(\theta)$, then the MM estimate of θ, which we will denote with $\bar{\theta}$ is chosen such that

$$\phi(\bar{\theta}) = \bar{x}.$$

Examples

If $X \overset{d}{=} \exp(\theta)$, then what is the MM estimator for θ?

If $X \overset{d}{=} \mathrm{geo}(\theta)$, then what is the MM estimator for θ?

Comments on MM estimators

Due to the law of large numbers, as long as ϕ has an inverse, MM estimators are consistent. However, in general they are not unbiased, and often not very efficient.

The main reason to use a MM estimator is that they are usually very easily derived when other methods may be too difficult to use.

Multiple parameters

In the case where there are multiple parameters, we use an appropriate number of moments of the distribution in question.

For example, if there were two parameters to be estimated, we would equate the sample variance with the population variance to give us a second equation.

Example

If $X \overset{d}{=} \mathrm{Nb}(\mathrm{r}, \mathrm{p})$, then what is a MM estimator for r and p?

Method of Maximum Likelihood (ML) in words

- Get some data.
- Propose model for said data.
- Hence, calculate the ‘probability’ of achieving actual data.
- Choose parameters such that this probability is maximised.

Method of Maximum Likelihood (ML) in maths

- We begin with sample of iid random variables $X_1, X_2, \dots X_n$, with common density function $f_X(x|\theta)$ where θ is a parameter of the distribution.
- Define the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

- The maximum likelihood estimator of θ, denoted by $\hat{\theta}$ is defined such that

$$L(\hat{\theta}) \geq L(\theta)$$

for all θ.

Generalisations and remarks

- If the distribution of X is discrete, then simply replace the density function with a mass function.
- It is usually easier to compute the logarithm of the likelihood function and maximise that.
- The sample does not necessarily need to be iid. In full generality the likelihood function is the joint distribution of the sample.

Example

Suppose I tossed a two sided coin 84 times, and heads appeared 36 times.

Find the maximum likelihood estimate (MLE) for the probability of a coin toss showing heads.

Example

Given a random sample $X_1, X_2, \ldots X_n$ of X, where X has distribution $\mathrm{N}(\mu, \sigma^2)$, then derive the maximum likelihood estimators for both μ and σ^2.

Example

Suppose a random sample of *n* observations is obtained on $X = \mathrm{R}(0, \theta)$.

What is the maximum likelihood estimator for θ?

Bias?

From previous classes, we know that the sample variance for a normally distributed sample is unbiased, and for the rectangular distribution example, an unbiased estimate is actually $\frac{n}{n+1}x_{(n)}$.

... So it seems that MLEs are often biased.

This isn't good...

So why are we still interested in MLEs then?

MLEs have many nice properties, and asymptotically, they are actually the best.

Regularity conditions

The regularity conditions are given by

1. (R1) The pdfs are distinct, that is $\theta \neq \theta' \Rightarrow f(x|\theta) \neq f(x|\theta')$.
2. (R2) The pdfs have common support (domain) for all θ.
3. (R3) The true value of θ which we denote by θ_0 is an interior point in the space of all possible values of θ.

Remarks

The first assumption is reasonable, it merely states that the parameter is what defines the pdf.

The second assumption eliminates cases where the support depends on θ. For example $\mathrm{R}(0, \theta)$.

The essential idea for having the second assumption is that it allows us to take things like derivatives of the likelihood function, and the third assumption means that we can use the derivative to find the maximum.

Consistency

As we saw in our examples, the MLE can be biased, but at least in our examples, it appears to be asymptotically unbiased, or in other words, consistent.

Is this always true?

Theorem

Let θ_0 denote the true value of a parameter. Under assumptions (R1), (R2) and (R3), and assuming f is differentiable with respect to θ:

As $n \to \infty$, the MLE estimate $\hat{\theta} \xrightarrow{p} \theta_0$.

Invariance

Sometimes we may be interested in a function of our parameter.

For example, if we estimated what the average return for a stock portfolio is, the real question of interest is how much money you are expected to make, which is a function of the average rate of return.

And we know probability can be funny, certainly $\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$. So the question is, can we use our MLE estimate for functions of the parameter we estimated?

Theorem

If $\phi = \psi(\theta)$, then the maximum likelihood estimate of ϕ is such that $\hat{\phi} = \psi(\hat{\theta})$.

So far

So we have seen that MLEs can be biased, but they are consistent (under certain conditions).

The last thing to test is the efficiency of a MLE.

We will show that there exists a lower bound for the variance of any unbiased estimator, and furthermore show that the MLE attains this lower bound asymptotically.

Assumptions

Before we start, I'm going to use the simplifying assumption that our likelihood function is 'nice'.

So I mean things like the regularity conditions mentioned earlier, and that we can take derivatives of likelihood functions without any problems.

For a careful use of the assumptions, see the textbook.

D_1 and D_2

Two critical random variables that we will need to study the efficiency of MLEs are the following. Define

$$d_1 = \frac{\partial \log L}{\partial \theta}, \qquad d_2 = \frac{\partial^2 \log L}{\partial \theta^2}.$$

These two objects are functions of $x_1, x_2, \ldots x_n$. The corresponding random variables D_1 and D_2 are defined by replacing the $x_1, x_2, \ldots x_n$ with $X_1, X_2, \ldots X_n$.

Example

Consider a sample of n observations on $X \overset{d}{=} \text{Geo}(\theta)$ for which the pmf is $p_X(x) = \theta(1-\theta)^x$ for $x = 0, 1, 2, \ldots$

Find $\mathbb{E}(D_1)$, $\mathbb{E}(D_1^2)$ and $\mathbb{E}(D_2)$.

We notice that $\mathbb{E}(D_1) = 0$, and $\mathbb{E}(D_2) = -\mathbb{E}(D_1^2)$. Curious.

Theorem

$\mathbb{E}(D_1) = 0$ *and* $\mathbb{E}(D_1^2) = Var(D_1) = -\mathbb{E}(D_2) := I(\theta)$.

$I(\theta)$ is called the *information function*. We call it this because it in some sense encapsulates the total amount of information about θ that the data encodes. This will become more intuitive as we see how we use $I(\theta)$.

Rao-Cramér lower bound

Theorem

If $X_1, X_2, \ldots X_n$ are iid with common distribution, and everything is 'nice', then if $\mathbb{E}(T) = k(\theta)$, that is T is an unbiased estimate of some function of θ, then

$$Var(T) \geq \frac{[k'(\theta)]^2}{I(\theta)}.$$

Example

If we have a sample of n observations on $X \overset{d}{=} \mathrm{N}(\theta, 1)$, then show that the sample mean is an efficient estimator.

Asymptotic normality

So we have derived a minimum variance bound for an unbiased estimator, it remains to be seen how this is related to maximum likelihood estimators.

Theorem

If $\hat{T}$ denotes the maximum likelihood estimator of θ based on a sample of n observations, and everything is still 'nice', then as $n \to \infty$

$$\hat{T} \xrightarrow{d} \mathrm{N}\left(\theta, \frac{1}{\mathrm{I}(\theta)}\right).$$

This means that at worst the maximum likelihood estimator for θ is asymptotically unbiased and asymptotically efficient.

Example

For a sample of n observations on $X \overset{d}{=} \text{Geo}(\theta)$, show that the maximum likelihood estimate of θ is $\hat{T} = \frac{1}{\bar{X}+1}$. Furthermore, specify the asymptotic distribution of $\hat{T}$.

Decision making

The ultimate goal of using statistics is to make decisions.

To be able to answer, yes, Americans are on average more overweight than Australians.

The framework we use for this, is hypothesis testing.

Hypotheses

A statistical hypothesis is usually a statement about the distribution of a random variable X.

We tend to assume that the distribution of X is specified except for a particular parameter.

A hypothesis then takes the form $\theta = 1, \theta = \theta_0, \theta > 3, \theta \neq 6$ etc.

Null hypothesis

The hypothesis under test is called the *null hypothesis*, and we denote this with H_0. It usually represents some kind of known standard, what we expect the results should be.

The onus typically lies with the experimenter to show that an actual ‘effect’ exists, hence we tend to not reject the null hypothesis unless there is strong evidence.

We will usually assume that the null hypothesis always takes the form, $H_0 : \theta = \theta_0$.

Alternative hypothesis

We test the alternative hypothesis against the null hypothesis. The alternative hypothesis is often the opposite of the null hypothesis, that is, the null hypothesis is not true, but not necessarily. For example it is sometimes the case that $H_1 : \theta = \theta_1$, and we will soon see examples of when this is the case.

Example

If we were testing the effect of alcohol on how much nonsense Jeff says, and we assumed that alcohol won't actually make him more sensical, then if θ denotes the percentage of words that come out of his mouth that are nonsense when sober, write down the null and alternative hypotheses.

Conclusions and errors

Given a random sample on X, we then decide whether to reject H_0 or not. Typically there will be some sort of rule for making this decision, for example, reject H_0 if $\bar{X} > c$. With such a decision process, there are two types of errors we need to be aware of.

	accept H_0	reject H_0
H_0 true	correct decision probability $1 - \alpha$	error of type I with probability α aka *size*
H_1 true	error of type II probability β	correct decision probability $1 - \beta$ aka power

Errors

We would typically like the error probabilities α and β to be small, though however typically we focus on making sure α is small, β being small is of lesser importance. This is due to the special importance of H_0.

Example

Given a sample of 10 observations on $X \overset{d}{=} \mathrm{N}(\theta, 1)$, we wish to test the null-hypothesis $H_0 : \theta = 0$ against the alternative hypothesis $H_1 : \theta \neq 0$.

If I wanted the size to be 0.05, then how should I construct this test?

Furthermore what would be the power of this test?

Test statistics and critical regions

Typically, the test can be expressed in the form:

Reject H_0 if $T = \psi(X_1, X_2, \dots X_n) \in A$.

The set A is known as the *critical region* and T is known as the test-statistic. T is typically an estimator for θ.

Therefore, the size of the test is $\mathbb{P}(T \in A | H_0)$.

P-values

An alternate and equivalent approach is to use the P-value which is defined as

$$P = \mathbb{P}(\text{test statistic is as extreme as the value obtained} | H_0),$$

where extreme is defined by the alternative hypothesis.

The test procedure is then to reject H_0 if $P < \alpha$.

Intuition

The P-value tells you how ‘likely’ the observed data is given the null hypothesis.

It is common to quote just the value of P in a statistical analysis, as it gives slightly more information than just accepting or rejecting the null hypothesis.

Example

25 independent observations on $X \overset{d}{=} \mathrm{N}(\theta, 9)$ give $\bar{x} = 11.5$. To test $H_0 : \theta = 10$ against $H_1 : \theta \neq 10$, using a test of size 0.05, what is the P-value and your conclusion?

Why are P-values awful

People, particularly people who are bad at statistics seem to think of P-values as universal cure alls. If you just quote the P-value, then that's it, question answered.

The main reason why P-values are awful is that confidence intervals are better in every single way.

Furthermore, just because we do not reject H_0 does not necessarily mean that H_0 is the right thing to base our decisions on.

Deriving tests

So far, the test statistics have been somewhat arbitrarily chosen.

Similar to when we had estimators, we would like to have some sort of systematic way of deriving the ‘best’ test, or at the very least a ‘good’ test.

The method used to derive the best test is called the *likelihood ratio criterion*.

Simple H_0 vs. simple H_1

Suppose $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.

Define

$$L_1 = \text{likelihood function under } H_1 = L(\theta_1),$$
$$L_0 = \text{likelihood function under } H_0 = L(\theta_0).$$

The likelihood ratio test is then given by,

$$\text{Test: reject } H_0 \text{ if } \frac{L_1}{L_0} > k.$$

Example

Consider a random sample of n observations on $X \overset{d}{=} \text{Pn}(\theta)$. Find the likelihood ratio test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where $\theta_1 > \theta$.

The best test

Theorem

The likelihood ratio test is the best test, that is given all tests with the same size, the likelihood ratio test has the greatest power.

Simple H_0 vs. composite H_1

Typically, our alternative hypothesis isn't that the parameter takes a single value, but rather exists in a range of other values.

Normally we have something like $H_0 : \theta = 3$, and $H_1 : \theta \neq 3$.

Well sometimes this doesn't matter.

Often it will though. If it does, then we use a ‘best of the rest’ option.

We choose to test null hypothesis against the most plausible alternative, that is we take θ_1 to be value that maximises $L(\theta)$ in the set of possible values (A_1) defined by the alternate hypothesis. Thus we define:

$$L_1 = \max_{\theta \in A_1} L(\theta),$$

and the likelihood ratio test is

$$\text{Reject } H_0 \text{ if } \frac{L_1}{L_0} > k.$$

Example

Consider a random sample of n observations on $X \overset{d}{=} \mathrm{N}(\theta, 1)$. Find the likelihood ratio test of $H_0 : \theta = \theta_0$ against the alternative hypothesis of $H_1 : \theta \neq \theta_0$.

Note

Note that typically such a test simply becomes comparing against the maximum likelihood estimate.

Composite H_0 vs composite H_1

Suppose $H_0 : \theta \in A_0$ and $H_1 : \theta \in A_1$. Unsurprisingly this likelihood test ratio takes the form:

$$\text{reject } H_0 \text{ if } \frac{L_1}{L_0} > k$$

where $L_1 = \max_{\theta \in A_1} L(\theta)$ and $L_0 = \max_{\theta \in A_0} L(\theta)$. These sorts of hypotheses typically occur in the case where θ is a vector. There is one case of particular interest.

Example

Consider a random sample of n observations on $X \overset{d}{=} \mathrm{N}(\mu, \sigma^2)$. Find the likelihood ratio test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Show that this likelihood ratio test corresponds exactly to the standard t-test.

Asymptotic LR test statistic

Theorem

Under the null hypothesis $H_0 : \theta = \theta_0$,

$$\log L(\hat{\theta}) - \log L(\theta_0) \xrightarrow{d} \frac{1}{2}\chi_1^2.$$

This means if we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, for large n if H_0 is true, then

$$\log L_1 - \log L_0 \overset{d}{\approx} \frac{1}{2}\chi_1^2.$$

We can therefore specify a critical region with $\frac{1}{2}c_{0.95}(\chi_1^2) = 1.92$ and hence for a size of 0.05, our test is to

$$\text{Reject } H_0 \text{ if } \log \frac{L_1}{L_0} > 1.92.$$

Example

Consider a random sample of n observations from $\mathrm{N}(\theta, 1)$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Find the exact distribution of the log-likelihood ratio.

In this case,

$$\frac{L_1}{L_0} = \exp\left(\frac{1}{2}n(\bar{x} - \theta_0)^2\right).$$

And hence

$$\log\frac{L_1}{L_2} = \frac{1}{2}n(\bar{X} - \theta_0)^2 = \frac{1}{2}\left(\frac{\bar{X} - \theta_0}{1/\sqrt{n}}\right)^2 \overset{d}{=} \frac{1}{2}\chi_1^2.$$

Distribution-free methods

So far for everything we have done, we have always needed to assume something about the underlying distribution.

For example before we can even write down a likelihood function, we need to assume that we know the form of the distribution of X.

Is this always realistic? Are there options where we do not have to assume anything about the distribution of X?

Example

Jeff the alcoholic Australian, claims that X, the number of schooners of beers he drinks in a day, follows the following distribution.

x	0	1	2	3	4	≥ 5
$\mathbb{P}(X = x)$	0.1	0.3	0.2	0.2	0.1	0.1

How can we test such a claim?

Suppose we then observed the following data over 100 days.

x	0	1	2	3	4	≥ 5
$\mathbb{P}(X = x) = p_x$	0.1	0.3	0.2	0.2	0.1	0.1
obs. freq. f_x	0	11	25	22	28	14

How does Jeff's original hypothesis hold up?

Test-statistic

The test statistic we use to assess the goodness of fit is

$$U = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}.$$

To decide whether or not the original hypothesis was reasonable, we are simply testing if U is too large or not.

So the question is, what is the distribution of U?

Intuition

As long as n is large-ish, then what is the approximate distribution of $\frac{(f_i - np_i)}{\sqrt{np_i q_i}}$?

The central limit theorem says this is approximately normal.

If we do a little rescaling, some ‘magic’, it can be shown that as $n \to \infty$,

$$U = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2_{k-1}.$$

Notes

- A χ^2 test is a one sided test. Why?
- Implicit in the χ^2 test is the fact that we are approximating a binomial with a normal, hence we normally stipulate that $np_i > 5$ as a rule of thumb.
- If this is not satisfied, we often combine classes until it is satisfied.
- If U is *too* small, then the first is probably *too good* and may suggest that the experiment was rigged.

Pie

Are the digits of π random?

Well here's a table of the first 100 digits of π.

i	0	1	2	3	4	5	6	7	8	9
f_i	8	8	12	12	10	8	9	8	12	13

Fitting distributions

The null hypothesis could easily take the form of a particular distribution, with unknown parameters.

We first estimate the parameters from the sample, and then use these estimates to determine expected frequencies under H_0.

Example

Jeff likes to think that the number of schooners of beer that he drinks in a day follows the Poisson distribution. (Possibly because Poisson sounds like Poison. Who knows.)

Suppose we then observed the following data over 100 days.

x	0	1	2	3	4	5
obs. freq. f_x	0	11	25	22	28	14

Does he follow a Poisson distribution?

If H_0 is true then $U \overset{d}{=} \chi^2_5$.

We have lost a 'degree of freedom' because we have added a constraint that the sample mean must be equal to the fitted mean.

In general in fitting a distribution in this manner,

$$U \overset{d}{=} \chi^2_{k-p-1},$$

where $k =$ the number of classes and $p =$ the number of parameters estimated.

Independence Tests

Suppose I have two classification random variables that I am interested in, and wish to answer the question of whether there is a relation between the two.

This question essentially boils down to whether the two random variables are independent.

A standard technique to test for this is to use a χ^2 test in a similar manner to what we have just covered.

A 2 x 2 contingency table

Let A denote if a person wears yellow clothing a lot, and A' if not.

Let B denote if a person has been stung by a bee in the past year, and B' if not.

If we had a sample of people, we could summarise the data into a table like follows:

obs freq	B	B'	
A	40	30	70
A'	20	10	30
	60	40	100

χ^2 test

If we set H_0 to be that the classifications are independent, this is equivalent to

$$H_0 : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

$$H_1 : \mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B).$$

If H_0 is true then the expected frequencies are given by

exp freq	B	B'	
A	$np_A p_B$	$np_A p_{B'}$	np_A
A'	$np_{A'} p_B$	$np_{A'} p_{B'}$	$np_{A'}$
	np_B	$np_{B'}$	n

Under H_0 we can then calculate expected frequencies, if we make estimates for p_A and p_B.

Example

	B	B'	
A	40	30	70
A'	20	10	30
	60	40	100

Perform a χ^2 test to see if there is a relationship between A and B.

r x c contingency tables

The χ^2 test generalises to a r x c contingency table in exactly the same way as you would expect.

Given that we only need to estimate $r-1$ of the $p_{A_i}s$ and $c-1$ of the p_{B_j}s, then the degrees of freedom in the χ^2 would be $rc-((r-1)+(c-1))-1=(r-1)(c-1)$.

Inference on the median

We are going to consider a random sample of n observations from a continuous random variable X with median m.

We wish to be able to make inference about m, for example test a hypothesis that $m = m_0$.

To do this we use the binomial distribution...?

Sign test

To test the hypothesis $H_0 : m = m_0$, we can use the test statistic $Z = \text{freq}(X \leq m_0)$.

Under H_0 the distribution of Z should be $\mathrm{Bi(n, 0.5)}$.

This is often called the sign test because we are considering the ‘sign’ of the $x_i - m_0$.

Example

Construct a test based on 20 observations with $H_0 : m = m_0$ against $H_1 : m > m_0$.

So the test will take the form of reject H_0 if $Z \leq c$.

Using R we can then find the size of the test for a given c.

Confidence interval

To obtain a confidence interval for m we only need to find the appropriate order statistics and see how it is related to $Z = \text{freq}(X \leq m)$.

$$\mathbb{P}(r \leq Z \leq s - 1) = \mathbb{P}(X_{r)} < m < X_{(s)}).$$

Hence for a sample of $n = 20$ a two sided confidence interval is given by

$$\mathbb{P}(6 \leq Z \leq 14) = \mathbb{P}(X_{(6)} < m < X_{(15)}) = 0.9586.$$

Asymptotic distribution

It's often troublesome to work with the binomial distribution, so recall that we can always approximate the distribution of Z with a normal distribution if n is large enough,

$$Z \overset{d}{\approx} \mathrm{N}\left(\frac{\mathrm{n}}{2}, \frac{\mathrm{n}}{4}\right).$$

Alternatively, we can also use the asymptotic result:

$$\hat{M} \overset{d}{\approx} \mathrm{N}\left(\mathrm{m}, \frac{1}{4\mathrm{nf}(\mathrm{m})^2}\right),$$

but this is usually problematic because one needs to know f to use this.

Example

Jeff believes that his median number of beers he drinks a day is 3. Test this.

x	0	1	2	3	4	5
obs. freq. f_x	0	11	25	22	28	14

Wilcoxon Rank test

If we are willing to assume that the distribution of X is symmetrical, then there exists a more powerful test for $H_0 : m = m_0$ based upon the 'ranks' of our observations.

We first transform our data X by considering the sample $Y_i = |X_i - m_0|$, then assign ranks to these Y_i by ordering them from smallest to largest.

Our plan is to then compare the ranks of the observations smaller than the hypothesised median m_0 with the ranks of the observations that are larger than m_0.

Example

Suppose we wished to test $H_0 : m = 45$ against $H_1 : m \neq 45$ based on the following 10 observations

```
42 54 47 44 53 43 43 16 51 41
```

Under the null hypothesis, what we would expect the difference of the ranks below and above 45 to be?

Test statistic

If we define T to be the difference of the sum of ranks above and the sum of the ranks below the hypothesised median, then it can be shown that assuming H_0 is true,

$$\mathbb{E}(T) = 0, \quad \text{Var}(T) = \frac{n(n+1)(2n+1)}{6},$$

and hence when n is large we can use a normal approximation.

It should be noted that the exact distribution for the Wilcoxon rank test is known, however we will tend to stick with normal approximations in this course.

Comparative inference of medians

The ideas that we have just covered for a single sample can also be extended to two sample problems.

As usual, the question we are interested in is, given two populations, how can we tell if they have the same distribution?

In the past we have tested for equality of means. We now have the tools to test for equality of medians.

Equality of medians

To test for equality of medians, we can use a contingency table.

Given two random samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$, if $N = n_1 + n_2$ and m^* is the median of the combined samples, a simple test for equality of medians is essentially equivalent to testing an independence hypothesis for a 2 x 2 contingency table.

Equality of medians

Test $H_0 : m_X = m_Y$ against $H_1 : m_X \neq m_Y$ for the following:

```
X: 93 98 103 111 102 112 92 90 106 103
```

```
Y: 89 103 118 96 86 84 99 107 106 101
```

The overall median is $m^* = 101.5$.

Contingency table

	$\leq m^*$	$> m^*$	
X-sample			n_1
Y-sample			n_2
			N

If we can calculate the expected number of observations in each cell, then we can apply a χ^2 test.

Under H_0, we expect each cell to have the same number of observations in it.

Rank tests for comparative inference

We can also extend the rank test for two sample testing.

Given two random samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$, if $N = n_1 + n_2$ we wish to test if there is a difference between the two groups.

We simply order the data, and calculate the sum of the ranks from either group.

Example

Our buddy Jeff the alcoholic wishes to test if his alcoholism affects his ability to play Starcraft II. Jeff (unwisely) decides to measure his ability by the number of actions per minute (APM) he can input. He plays 10 games sober (X), then plays 10 games drunk (Y) and measures his average APM per game.

X: 50 43 48 56 54 40 44 47 45 51

Y: 41 35 42 45 44 27 38 33 37 42

Calculate the ranks and see if there is a difference between his ability when sober and drunk.

Distribution of the rank test

If H_0 is true, then W_X (the sum of the ranks from X) should be the sum of 10 random integers between 1 and 20. So in theory we can calculate an exact P value, but considering how many combinations give us a value for the sum less than W_X.

This is a lot of effort, so we can also use a normal approximation. If n_1 and n_2 are large enough, then

$$W_1 \overset{d}{\approx} \mathrm{N}\left(\frac{1}{2}\mathrm{n}_1(\mathrm{n}_1+\mathrm{n}_2+1), \frac{1}{12}\mathrm{n}_1\mathrm{n}_2(\mathrm{n}_1+\mathrm{n}_2+1)\right).$$

Notes about distribution free methods

- Good compared to previous methods because we don't need to assume anything about underlying distributions.
- These tests tend to have less statistical power though.
- Usually much more robust to outliers.

Monte Carlo

The term *Monte Carlo* is thrown around a lot in statistics, and there isn't really a good definition of what it means.

Where does the term Monte Carlo come from?

In essence, gambling.

Nowadays, when people say Monte Carlo, they typically are referring to generating random variables from a known distribution, and then exploiting this to calculate quantities of interest.

I still think of Monte Carlo as throwing darts at a dart board.

Estimating π

If we can assume we know how to generate a random number between 0 and 1, then how can we estimate π?

Well, we could use a dart board...

Simulation

Usually simulation relies upon the ability to generate observations from $\mathrm{R}[0,1]$.

This is a well studied problem and for this course we will take it for granted that we can do this.

Given a uniform generator, if we know the pmf of a discrete distribution for X, then it is relatively simple to simulate values from X.

Inverse-transform method

Theorem

Let U have distribution $\mathrm{R}[0,1]$, and F be a continuous distribution function. Then the random variable $X = F^{-1}(U)$ has the same distribution function F.

Proof.

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x)\end{aligned}$$

□

The inverse transform essentially gives us a simple technique for simulating continuous random variables.

Using the inverse transform method, show how you can simulate an exponential random variable with rate α.

Problems with inverse transform

The primary problem with the inverse transform method is that, inverting the cdf function for most distributions is not actually possible.

Take for example the normal distribution. There already is no analytic function for the cdf function, let alone its inverse.

So what alternatives do we have?

Basic Accept-Reject algorithm

If X is a random variable with finite support, just draw a big box around it and go from there!

Art!

Proof that Basic AR works

The probability that a simulated value takes the value j has density $f(j)$.

Problems with basic AR

It can be horribly inefficient.

This won't really work for a continuous distribution either.

So can we do better?

(Not terrible) Accept-Reject

The key idea is that instead of drawing a giant rectangle, we use a function that closely resembles the target density.

Suppose we wanted to simulate a random variable X with pdf f, but we can't do it exactly. However, we can simulate a different random variable Y with pdf g, which is hopefully similar(ish) to X. g is often called the instrumental pdf.

Well it is usually true that there should exist some M such that for all x

$$f(x) \leq Mg(x).$$

The AR algorithm

- Generate Y and U where U is $\mathrm{R}[0,1]$.
- If $U < \frac{f(Y)}{Mg(Y)}$, then accept Y as a simulated value of X.
- Otherwise, generate a new Y.

We need to prove that this generates a value with the correct density.

Example

Using the AR algorithm, show how one can simulate the standard normal distribution by using the Cauchy distribution.

Note that if Y follows Cauchy distribution, then it has density

$$g(y) = \frac{1}{\pi(1+y^2)}$$

for all real y.

Sufficient statistics

In the past we introduced the *information function* $I(\theta)$ as a (somewhat) measure of the information that the data contains about a particular parameter θ.

A similar idea exists for a statistic. There exist statistics that contain the maximal information about a parameter θ, we call these *sufficient statistics*.

The idea is that once you know a sufficient statistic, knowing another alternative statistic will not give you any more useful 'information'.

Definition

A statistic of our data $X_1, X_2, \dots X_n = \mathbf{X}$, $T(X_1, X_2, \dots X_n)$ is *sufficient* if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ. That is

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t, \theta) = \mathbb{P}(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t).$$

(Suspiciously) Equivalently,

$$\mathbb{P}(\theta | \mathbf{X} = \mathbf{x}, T = t) = \mathbb{P}(\theta | T = t).$$

Example

Consider $X_1, X_2, \ldots X_n$ as a random sample of Bernoulli random variables with probability of success p.

Show that $Y = X_1 + X_2 + \ldots + X_n$ is a sufficient statistic for p.

Alternate definition

Let $X_1, X_2, \ldots X_n$ denote a random sample that has pdf/pmf $f(x|\theta)$, and $Y = T(X_1, X_2, \ldots X_n)$. Then Y is a *sufficient statistic* for θ if and only if

$$\frac{f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta)}{f_Y(T(x_1, x_2, \ldots x_n)|\theta)} = H(x_1, x_2, \ldots x_n)$$

where H is a function that does not depend on θ.

Example

Let $X_1, X_2, \dots X_n$ denote a random sample from $\mathrm{N}(\mu, 1)$.

Show that $Y = \bar{X}$ is a sufficient statistic for μ.

This definition is still awkward to work with. It requires us to know the distribution of Y, and while this was relatively easy in the Normal case, in general this is a hard problem. Fortunately, we have a solution to this.

Fisher-Neyman factorisation theorem

Let $X_1, X_2, \ldots X_n$ be a random sample from a distribution with pmf/pdf $f(x|\theta)$. A statistic $Y = T(X_1, X_2, \ldots X_n)$ is sufficient if and only if there exist two non-negative functions k_1 and k_2 such that

$$f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta) = k_1(T(x_1, x_2, \ldots x_n)|\theta)k_2(x_1, x_2, \ldots x_n).$$

Examples

Revisit the normal example.

Let $X_1, X_2, \ldots X_n$ be from a $\mathrm{Po}(\lambda)$ distribution. Show that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

Let $X_1, X_2, \ldots X_n$ be from $R[0, \theta]$. Show that $T = \max\{X_1, X_2, \ldots X_n\}$ is a sufficient statistic for θ.

Sufficiency and the MVB

From all our examples, it seems that sufficient statistics always appear to be functions of the maximum likelihood estimate. Is this just chance or is there a reason for this? (Not just chance.)

Furthermore, recall that a MLE asymptotically attains the minimum variance bound (MVB). Can we use sufficiency to find the estimator that attains the minimum variance bound?

MVUE vs MVB

The textbook refers to the minimum variance unbiased estimator.

I refer to it as the MVB, I implicitly assume that it is for an unbiased estimator because it kind of makes no sense to consider this for a biased estimator.

Rao-Blackwell theorem

Given a random sample which depend upon θ, a sufficient statistic Y_1 for θ and Y_2 an unbiased estimator for θ, then

$$\mathbb{E}(Y_2|Y_1) = \phi(Y_1)$$

defines a statistic that is also unbiased, and most importantly the variance of it is less than or equal to $\mathrm{Var}(Y_2)$, that is

$$\mathrm{Var}(Y_2) \geq \mathrm{Var}(\phi(Y_1)).$$

What Rao-Blackwell theorem actually means

If I have a sufficient statistic, then for every unbiased estimator, if we take the conditional expectation with respect to with the sufficient statistic, you get a ‘better’ estimate.

What this means is, that the MVB must be a function of a sufficient statistic.

MLEs and sufficient statistics

If a sufficient statistic Y for θ exists, and if a maximum likelihood estimator $\hat{\theta}$ for θ exists and is unique, then $\hat{\theta}$ is a function of Y.

Example - Exponential distribution

Let $X_1, X_2, \dots X_n$ be a random sample from the exponential distribution with rate parameter λ.

Using a sufficient statistic, can we find the estimator that achieves the MVB for λ?

In the previous example we found what could be a ‘candidate’ for the minimum variance bound, but there could be other ‘candidates’ out there.

However, ‘usually’ there is only one function of the sufficient statistic that is unbiased, and hence that one function gives the MVB.

We should probably define exactly what ‘usually’ means...

Completeness

A family of distributions $f(x|\theta)$ is said to be *complete* if the condition that

$$\mathbb{E}(u(Z)) = 0$$

for all possible values θ implies that $u(z)$ is zero on all points of positive probability, where u must be a function that is independent of θ.

Example

Lets consider the Poisson distribution.

Lehmann-Scheffé Theorem

Given a random sample $X_1, X_2, \ldots X_n$ and Y is a sufficient statistic for θ, then if there is a function of Y, $\phi(Y)$ that is unbiased, and the family of distributions of Y is complete, then $\phi(Y)$ is the unique estimator that achieves the MVB for θ.

Example

Let $X_1, X_2, \ldots X_n$ follow the $R[0, \theta]$ distribution. Find the estimator that achieves the MVB.

Where's Wally?

So the burning question is, how do we find these ‘complete’ sufficient statistics?

It is a bit of a pain to first find a sufficient statistic Y for the sample on X's, and from there check if Y is complete.

Ideally we would like to be able to simply check a condition upon X...?

Exponential class of distributions

Consider a pmf/pdf $f(x|\theta)$ which is non-trivial on $x \in \mathcal{S}$. If there exist functions $p(\theta)$, $K(x)$, $H(x)$ and $q(\theta)$ such that

$$f(x|\theta) = \exp[p(\theta)K(x) + H(x) + q(\theta)] \quad x \in \mathcal{S},$$

where $\mathcal{S}$ does not depend upon θ and $p(\theta)$ is non-trivial, then we say f belongs to the regular exponential family.

Examples

$\mathrm{N}(0, \theta)$.

The binomial distribution with fixed *n*.

Exponential family and complete sufficient statistics

If $X_1, X_2, \ldots X_n$ is a random sample from the regular exponential family and Y a sufficient statistic for X, then

$$Y = \sum_{i=1}^{n} K(X_i)$$

is a sufficient statistic for θ and the family of density functions for Y is complete. Hence Y is a complete sufficient statistic for θ.

From sufficiency to MVB

Now that we have our sufficient statistic, what is the MVB?

Mean and variance of the sufficient statistic

If we have a sample from a the regular exponential family, then a sufficient statistic is $Y = \sum_{i=1}^{n} K(X_i)$ and

$$\begin{aligned} \mathbb{E}(Y) &= -n \cdot \frac{q'(\theta)}{p'(\theta)} \\ \mathrm{Var}(Y) &= \frac{n}{p'(\theta)^3} \left[p''(\theta) q'(\theta) - q''(\theta) p'(\theta) \right] \end{aligned}$$

They key message here is that we *do not need to know the distribution of Y* to find our MVB.

As an example consider $\mathrm{N}(\mu, \sigma^2)$.

Also consider a sample from the Poisson distribution.

Modelling

So far what is the missing step in our process of data analysis?

We've spent a great deal of effort to study properties of estimators etc.

However, we always had to assume a model before applying any of our theory.

So, modelling time.

Example

One day, Jeff begins to wonder, does how much he drinks (water obviously) depend upon the day of the week?

Well we know how to compare two means.

Suppose μ_1 is the mean number he drinks on Monday, μ_2 on Tuesday, μ_3 for Wednesday etc.

We can compare μ_1 and μ_2. In fact, we can compare any pairing. But that's not the question is it?

The question and the hypotheses

So the research question was exactly:

Does the amount of ‘water’ Jeff drink in a day depend upon the day of the week?

So our null hypothesis should be

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7,$$

and the alternate hypothesis is then what?

What we are going to try to do is to jointly test equality of *all* the means. How are we going to derive our test though?

Well we know that the most powerful test is the test that is derived from the likelihood ratio test...

The model

We consider n independent observations on k normally distributed random variances that have equal variances. There are thus $N = nk$ observations in total.

$$\begin{array}{lll}
X_1 \overset{d}{=} \mathrm{N}(\mu_1, \sigma^2) & \text{sample: } X_{11}, X_{12}, \ldots, X_{1n}, & \bar{X}_1, S_1^2 \\
X_2 \overset{d}{=} \mathrm{N}(\mu_2, \sigma^2) & \text{sample: } X_{21}, X_{22}, \ldots, X_{2n}, & \bar{X}_2, S_2^2 \\
\quad\ldots & & \\
X_k \overset{d}{=} \mathrm{N}(\mu_\mathrm{k}, \sigma^2) & \text{sample: } X_{k1}, X_{k2}, \ldots, X_{kn}, & \bar{X}_k, S_k^2
\end{array}$$

Under the hypotheses

Regardless of the truth of H_0, within a sub-sample

$$\frac{(n-1)S_i^2}{\sigma^2} \overset{d}{=} \chi^2_{n-1},$$

and therefore

$$\sum_{i=1}^{k} \frac{(n-1)S_i^2}{\sigma^2} \overset{d}{=} \chi^2_{k(n-1)} = \chi^2_{N-k}.$$

Under H_0

If H_0 is true then all our observations follow the same distribution and form a random sample of N observations from $\mathrm{N}(\mu, \sigma^2)$. So if we denote S_T as the sample variance for the combined sample

$$\frac{(N-1)S_T^2}{\sigma^2} \overset{d}{=} \chi^2_{N-1}.$$

Under H_0 cont.

Under H_0 then the means of each group $\bar{X}_1, \ldots, \bar{X}_n$ are also iid random samples on k observations with $\bar{X} \overset{d}{=} \mathrm{N}\left(\mu, \frac{\sigma^2}{\mathrm{n}}\right)$.

If the sample variance of this sample of sample means is denoted by S_B^2, then we also have

$$\frac{(k-1)S_B^2}{\sigma^2/n} \overset{d}{=} \chi^2_{k-1}.$$

Putting it all together

We thus have

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{\bar{X}})^2 = (N-1)S_T^2 \overset{d}{=} \sigma^2 \chi^2_{N-1} \quad \text{if } H_0 \text{ true}$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k} (n-1)S_i^2 \overset{d}{=} \sigma^2 \chi^2_{N-k} \quad \text{always}$$

$$B = \sum_{i=1}^{k} n(\bar{X}_i - \bar{\bar{X}})^2 = n(k-1)S_B^2 \overset{d}{=} \sigma^2 \chi^2_{k-1} \quad \text{if } H_0 \text{ true}$$

Names

We call T the total sum of squares, or total SS, B the between groups SS, and W the within groups SS.

And those names call it exactly how it is.

Most importantly,

$$T = W + B.$$

The likelihood ratio test

The LR test said we should base our test upon

$$\frac{W}{T}.$$

However instead this is equivalent to testing

$$\frac{\frac{B}{k-1}}{\frac{W}{N-k}}.$$

which we know follows distribution $F_{k-1,N-k}$.

This approach is called Analysis of Variance, or ANOVA for short.

The ANOVA table

	df	SS	MS	F
between				
within				
Total				

Example

Determination of the strength of a fibre after using three methods of treatment were as follows:

1. (control)	49.8	48.5	48.7	47.2
2. (treatment A)	49.3	51.5	50.9	50.1
3. (treatment B)	51.2	52.8	52.3	53.2

Is there a significant difference between methods? Or, in other words, test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$.

ANOVA table

The results are set out as follows:

	df	SS	MS	F
between	2	29.26	14.63	15.66
within	9	8.41	0.93	
Total	11	37.67		

Do we reject the null hypotheses?

The initial hypothesis test of

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

is merely the first step in the analysis of such data.

For starters we are usually interested in the individual parameters, and wish to be able to do inference on those.

Confidence intervals

For any of the parameters estimated, we can also derive a confidence interval, including σ^2.

Question is, how do we first estimate σ^2?

Similarly, we can apply hypotheses tests.

Example cont.

Calculate confidence intervals for all the sample means, and for σ^2 from the previous example.

Generalising

We have

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = (N-1)S_T^2 \overset{d}{=} \sigma^2 \chi^2_{N-1} \quad \text{if } H_0 \text{ true}$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2 \overset{d}{=} \sigma^2 \chi^2_{N-k} \quad \text{always}$$

$$B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{\bar{X}})^2 \overset{d}{=} \sigma^2 \chi^2_{k-1} \quad \text{if } H_0 \text{ true}$$

Multiple comparisons

In the previous set up where we have categorised our data into k groups, it is a very natural question to ask, which groups are different from each other?

If there are just a few groups, then there's quite a few, but a manageable number of comparisons to be made. But if we introduce more groups, we went up with what is known as the *multiple comparisons* problem.

The problem

Our null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n.$$

We could rewrite this as a joint hypothesis of pairwise hypotheses, and if we were to reject any of the pairwise hypotheses, that would suggest we should reject H_0.

Thing is, each of these pairwise tests would have probability of 0.05 for rejecting its pairwise hypothesis. Then, what is the joint probability of all of the pairwise tests not being rejected?

Family vs. individual size

There are two approaches to this problem, one can in some sense simply ignore it, and maintain confidence intervals that have individually a size of 0.05.

Alternatively, we can come up with a way such that we will have a joint size of 0.05. That is the probability of at least one of the confidence intervals not containing it's true mean is 0.05.

One of these will result in (much) wider confidence intervals.

Bonferroni's correction

A very common and simple (correlated reasons, possibly causal even) way to deal with multiple comparisons is the Bonferroni method. We aim to find the individual confidence levels that maintains a certain overall confidence level.

If we let α_E denote the experimental size, that is the overall size, and α_C denote the size for a single comparison, then in the case that the m comparison tests are independent,

$$\alpha_E = 1 - (1 - \alpha_C)^m.$$

Suppose we wanted a experimental size of α. So to choose α_C we could set $\alpha_C = 1 - (1 - \alpha)^{\frac{1}{m}}$. If α is small, then a reasonable approximation, is $\alpha_C \approx \frac{\alpha}{m}$.

As a conservative but general rule, we can set $\alpha_C = \frac{\alpha}{m}$ which would result in $\alpha_E \leq \alpha$.

Note that this is very conservative, and for larger values of m, the actual value of α_E may be much less than α.

Tukey method

First note: it is Tukey, not Turkey.

Secondly, we introduce the standardised range distribution. (also known as the studentised range distribution)

Standised range distribution

Let $Z_1, \ldots, Z_k$ be independent $\mathrm{N}(\mu, \sigma^2)$ random variables and let R_k denote their range. That is $R_k = Z_{(k)} - Z_{(1)}$.

Suppose we also have an estimator of σ^2, S_ν^2 which is independent of the Z_i's and $\frac{\nu S^2}{\sigma^2} \overset{d}{=} \chi_\nu^2$.

Then $Q_{k,\nu} = \frac{R_k}{S_\nu}$ is called the standardised range distribution with k and ν degrees of freedom.

So given k independent samples each of n observations, let $Z_j = \bar{X}_j - \mu_j$, and S^2 denote the within group MS.

The key idea is that

$$|Z_i - Z_j| < c \quad \text{for all } i, j \quad \Leftrightarrow \quad Z_{(k)} - Z_{(1)} < c,$$

then a statement about *all* the differences in means being less than some value can be written in terms of the standardised range distribution.

Tukey CI

Therefore, our 95% confidence intervals for all the mean differences is:

$$\bar{X}_i - \bar{X}_j \pm c_{0.95}(Q_{k,N-k})\frac{S}{\sqrt{n}}.$$

The model interpretation

We formulated our one way ANOVA using the likelihood ratio test.

To do this we assumed that our observations came from the normal distribution with common variance, but with means depending on which group they came from.

This implicitly suggests a model has been fit.

The one-way ANOVA model

The model interpretation is as follows:

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where *i* indicates which group observation *ij* belongs to, and the ϵ_{ij} are i.i.d. $\mathrm{N}(0, \sigma^2)$.

How do we interpret between and within group SS now?

More than one group

The one-way ANOVA is used when we fit a different mean for each different level of a categorical factor.

For example, we could be interested in the heights of people, and we believe that would depend on one's gender.

However, very rarely does modelling stop at only considering just one variable. Usually there are more to consider. For example we may also be interested to see if ethnicity plays a role in one's height.

Additive models

First we are going to consider what is known as an *additive model*.

The additive model assumes that each level of a factor simply *adds* something to it's expected value.

Example

Suppose being male would contribute 150cm to your height.

Being female contributes 140cm.

Being Australian contributes 30cm.

Being American contributes 10cm.

We can tabulate this to summarise this easier. Question is though, how many parameters are there really? 4?

Contrasts

There are multiple ways of stating the formulation of your model, but all of them are essentially equivalent.

The different formulations are called *contrasts*.

The default R contrast

The default R contrast sets the ‘intercept’ term to the ‘base’ level for your two categorical variables.

Then if considering other levels of your variables, you simply add the appropriate variable.

What is the equation of the model then?

R calls these contrasts, contr.treatment.

Example

The textbook contrast

The textbook uses a different set of constrasts:

Let μ be the overall average of all the data. Then

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$.

R calls these contrasts, contr.sum.

The difference

Absolutely none, except some might argue one is easier to interpret than the other.

Hypothesis tests

What are the primary hypotheses of interest that we wish to test?

- Does the ‘row factor’ have any effect?
- Does the ‘column factor’ have any effect?

Testing of main effects

In an almost identical fashion to the one way ANOVA, we can decompose the total SS into components.

$$T = R + C + W,$$

that is, total SS is equal to the row SS + column SS + within SS.

Similarly to before, we can derive the most powerful test using the likelihood ratio test.

Interaction terms

With the additive model we assumed that the main effects have zero impact on each other. However, this is often not the case. Sometimes the row and column factors may have some sort of multiplicative effect.

So for example, the effect on sweetness by adding sugar to coffee and stirring. Individually, neither of those two factors will have much influence on the outcome. But when both are applied, a great difference occurs.

How do we account for interaction terms? Just add another parameter?

ANOVA

Again, just like before we can decompose the total SS into components such that

$$T = R + C + I + W,$$

that is total SS = row SS + column SS + interaction SS + within SS.

The exact details of this decomposition can be found in the textbook.

Hypothesis testing

We can therefore similarly formulate a hypothesis test for whether the interaction term should be included in our model or not.

As expected it is just a F-test where we compare mean SS for the appropriate columns in the ANOVA table.

Estimation

Nowhere have we actually covered what the estimates for our parameters in the model are.

It turns out that they are exactly as one would except, and they are the maximum likelihood estimates.

Depending on your parametersation, the parameters essentially depend upon the row averages and column averages.

Categorical and Numerical explanatory variables

So we have examined how to model when we take into account differing levels of a categorical variable.

In fact, the model will all interactions included is as general a model as possible.

What about numerical variables though?

One might suspect that the amount of alcohol one can ingest before passing out varies with body weight. This leads to regression.

The problem

The problem is as follows:

We want to estimate what the value of Y would be given different values of X, and we would seek to fit some sort of model based upon observed data (x_i, y_i).

So what we are essentially trying to estimate is

$$\mathbb{E}(Y|x) = \mu(x).$$

The equation

$$\mathbb{E}(Y|x) = \mu(x),$$

is rather general. There certainly exists a true function μ that ideally we are looking for.

If we wanted to find the true function, well there are quite a few functions we could have to consider.

Unfortunately, infinity is a lot. So we put on our applied mathematician hats and assume everything is linear.

Linear regression

The simplest function that we can reasonably consider is a linear function. That is we assume that

$$\mathbb{E}(Y|x) = \alpha + \beta x.$$

We also assume that $\mathrm{Var}(Y_i) = \sigma^2$ and the Y_i are independent. What is the distribution of Y_i then?

Note that this is different from the assumption that Y_i are i.i.d.

Estimation

Given a ‘cloud’ of points, the question is then, with what criteria do we choose α and β?

Art!

The method of least squares

So given our function $\mu(x) = \mathbb{E}(Y|x) = \alpha + \beta x$, we wish to choose the 'best' α and β out of all possible values of α and β to minimise

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

Properties of least squares linear regression

1. If the Y_i are i.i.d. random variables, then the method of least squares is just the method of moments.
2. If the Y_i are normally distributed, then the method of least squares corresponds exactly to the method of maximum likelihood.
3. The method of least squares gives the best linear unbiased estimators. That is out of all the estimators of the form $\sum a_i Y_i$, the least squares estimator is the most efficient.

Least squares vs. maximum likelihood

These two things are exactly equivalent.

The estimates of α and β

If we seek to minimise

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

then the answer is simply to set both partial derivaties to zero and solve for α and β.

. . . but ick?

Reparameterisation

Lets set $u_i = x_i - \bar{x}$.

First thing to note is that $\sum_{i=1}^{n} u_i = 0$. Now

$$E(Y_i) = \alpha + \beta x_i = \alpha + \beta \bar{x} + \beta(x_i - \bar{x}) = \alpha + \beta \bar{x} + \beta u_i = \alpha_0 + \beta u_i,$$

where $\alpha_0 = \alpha + \beta \bar{x}$.

Solutions

$$
\begin{aligned}
\hat{\alpha_0} &= \bar{y}, \\
\hat{\beta} &= \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\
\hat{\alpha} &= \bar{y} - \beta \bar{x}.
\end{aligned}
$$

Properties of the estimators

Let $\hat{A}_0$ be the estimator for α_0 and $\hat{B}$ for β,

$$\begin{aligned}
\mathbb{E}(\hat{A}_0) &= \alpha_0 \\
\mathrm{Var}(\hat{A}_0) &= \frac{\sigma^2}{n} \\
\mathbb{E}(\hat{B}) &= \beta \\
\mathrm{Var}(\hat{B}) &= \frac{\sigma^2}{\sum_{i=1}^n u_i^2} \\
\mathrm{Cov}(\hat{A}_0, \hat{B}) &= 0.
\end{aligned}$$

Ultimately though, we want the mean and variances of our estimators of α and β.

If we set $\hat{A}$ to be the estimator for α, then

$$\begin{aligned}
\mathbb{E}(\hat{A}) &= \alpha \\
\mathrm{Var}(\hat{A}) &= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} u_i^2}\right)\sigma^2 \\
\mathrm{Cov}(\hat{A}, \hat{B}) &= -\frac{\bar{x}}{\sum_{i=1}^{n} u_i^2}\sigma^2
\end{aligned}$$

Estimator variance

Finally what is the variance of our estimator for Y?

Let $\hat{M}(x)$ denote the estimator for $\mu(x)$, that is $\hat{M}(x) = \hat{A}_0 + (x - \bar{x})\hat{B}$, then

$$\mathbb{E}[\hat{M}(x)] = \mu(x)$$

$$\mathrm{Var}(\hat{M}(x)) = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} u_i^2}\right)\sigma^2$$

Estimation of σ^2

We define the residual sum of squares as

$$d^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\beta} u_i)^2,$$

$$D^2 = \sum_{i=1}^{n} (Y_i - \hat{A}_0 - \hat{B} u_i)^2.$$

We note that

$$\sum_{i=1}^{n}(y_i-\alpha_0-\beta u_i)^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}_0-\hat{\beta}u_i)^2+n(\hat{\alpha}_0-\alpha_0)^2+(\hat{\beta}-\beta)^2\sum_{i=1}^{n}u_i^2,$$

which is proved by showing that the cross terms in

$$\sum_{i=1}^{n}\left[(y_i-\hat{\alpha}_0-\hat{\beta}u_i)+(\hat{\alpha}_0-\alpha_0)+(\hat{\beta}-\beta)u_i\right]^2,$$

are zero. Hence

$$\sum_{i=1}^{n}(Y_i-\alpha_0-\beta u_i)^2=\sum_{i=1}^{n}(Y_i-\hat{A}_0-\hat{B}u_i)^2+n(\hat{A}_0-\alpha_0)^2+(\hat{B}-\beta)^2\sum_{i=1}^{n}u_i^2.$$

Hence by taking expectations of both sides we get

$$n\sigma^2 = \mathbb{E}(D^2) + \sigma^2 + \sigma^2.$$

Therefore to estimate σ^2 we would use

$$S^2 = \frac{D^2}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{A}_0 - \hat{B}u_i)^2,$$

and this is an unbiased estimator for σ^2.

Inference on $\hat{A}_0$ and $\hat{B}$

First recall that

$$\begin{aligned} \hat{\alpha_0} &= \bar{y}, \\ \hat{\beta} &= \frac{\sum_{i=1}^{n} u_i y_i}{\sum_{i=1}^{n} u_i^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

If we are assuming the distribution of the Y_i are normal, then what are the distributions of $\hat{A}_0$ and $\hat{B}$?

Distributions of $\hat{A}_0$ and $\hat{B}$

Hence,

$$\hat{A}_0 \overset{d}{=} \mathrm{N}\left(\alpha_0, \frac{\sigma^2}{\mathrm{n}}\right) \quad \text{and} \quad \hat{\mathrm{B}} \overset{\mathrm{d}}{=} \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum \mathrm{u}^2}\right).$$

Furthermore, they are independent, since uncorrellated and normality implies independence.

So how can we apply inference on these two? Usually σ^2 is unknown...

What is the distribution of D^2? We need this to find the distribution of S^2.

Recall that

$$\sum_{i=1}^{n}(Y_i-\alpha_0-\beta u_i)^2=\sum_{i=1}^{n}(Y_i-\hat{A}_0-\hat{B}u_i)^2+n(\hat{A}_0-\alpha_0)^2+(\hat{B}-\beta)^2\sum_{i=1}^{n}u_i^2.$$

Divide everything by σ^2, do a little bit of magic and we see that $D^2 \overset{d}{=} \sigma^2\chi^2_{n-2}$.

Hence $\frac{(n-2)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-2}$.

t-tests

Now since $\hat{A}_0$ and $\hat{B}$ are independent of S^2, we have

$$\frac{\hat{A}_0 - \hat{\alpha}_0}{S/\sqrt{n}} \overset{d}{=} t_{n-2}, \quad \frac{\hat{B} - \beta}{S/\sqrt{\sum u^2}} \overset{d}{=} t_{n-2}.$$

ANOVA approach

Recall that

$$\sum_{i=1}^{n}(y_i-\alpha_0-\beta u_i)^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}_0-\hat{\beta}u_i)^2+n(\hat{\alpha}_0-\alpha_0)^2+(\hat{\beta}-\beta)^2\sum_{i=1}^{n}u_i^2.$$

This can be rewritten as

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2+d^2.$$

ANOVA

We then end up with an ANOVA of the form

	df	SS
regression	1	$\hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
residual	$n - 2$	d^2
total	$n - 1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$

From here we can apply a F-test, which turns out to be equivalent to testing for $\beta = 0$.

Correlation

We have discussed in the past that correlation measures the strength of the linear relationship between two random variables X and Y.

Intuitively, this has to do with β, the slope of a linear regression.

Sample correlation

The *sample correlation coefficient* is defined as

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

This gives us a point estimate for ρ, but if we want to make any inference on ρ as usual we will need to know its distribution.

Relationship between β and ρ

Recall that in our linear regression setting

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

and

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

From these two statements, and noticing that both are unbiased estimators, we get

$$\beta = \rho \cdot \frac{\sigma_Y}{\sigma_X}.$$

Hypotheses

Given the previous assertion, the null hypothesis

$$H_0 : \beta = 0,$$

is exactly the same as testing

$$H_0 : \rho = 0.$$

Therefore, we could test $\beta = 0$ by testing ρ.

Distribution of R if $\rho = 0$

Recalling from the ANOVA setting where we had

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + d^2,$$

and

$$\beta = \rho \cdot \frac{\sigma_Y}{\sigma_X},$$

this implies

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = D^2 + R^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Furthermore we had

$$F = \frac{\text{regression } MS}{\text{residual } MS} = \frac{R^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}{D^2/(n-2)} = \frac{(n-2)R^2}{1-R^2} \overset{d}{=} F_{1,n-2}.$$

Note that $t_{n-2}^2 = F_{1,n-2}$.

So we can base our test statistic upon $\frac{R\sqrt{(n-2)}}{\sqrt{1-R^2}}$.

Note

We can also explicitly derive the pdf of R by change of variables upon the t distribution.

It is not a nice exercise, and is somewhat unnecessary.

The General Linear model

The general linear model is given by

$$\underset{\sim}{Y} = A\underset{\sim}{\theta} + \underset{\sim}{E}$$

where

$\underset{\sim}{Y}$ $= n$ x 1 vector of random observations
A $= n$ x p matrix of known constants
$\underset{\sim}{\theta}$ $= p$ x 1 vector of unknown parameters
$\underset{\sim}{E}$ $= n$ x 1 vector of random errors.

Assumptions

It is assumed that

$$\mathbb{E}(\underset{\sim}{E}) = \underset{\sim}{0}, \quad \mathbb{D}(\underset{\sim}{E}) = \sigma^2 I$$

and

$$\underset{\sim}{Y} \overset{d}{=} \mathsf{N}_n(A\underset{\sim}{\theta}, \sigma^2 I),$$

where N_n means multivariate normal.

Technically the assumption of normality is not necessary for point estimation, it is though for interval estimation and hypothesis testing as we would then require the distribution of estimators.

Model interpretation

So the model can be interpreted as

$$y_i = A_i \utilde{\theta} + \epsilon_i,$$

or

$$\text{response} = \text{deterministic function} + \text{random error}$$

where the deterministic function must be a *linear function* of the parameters $\utilde{\theta}$.

Examples and non-examples of linear models

- $\beta_0 + \beta_1 x$
- $\beta_0 + \beta_1 x + \beta_2 x^2$
- $\begin{cases} \beta_1 & \text{if female} \\ \beta_2 & \text{if male} \end{cases}$
- $\beta_0 + \beta_1 e^{\beta_2 x}$
- $\mu + \alpha_i + \beta_j + \gamma_{ij}$

Estimation

Our aim is to estimate $\underset{\sim}{\theta}$, and to achieve this we use the method of least squares.

That is we choose the vector $\underset{\sim}{t}$ that minimises

$$(\underset{\sim}{y} - A\underset{\sim}{t})^T(y - A\underset{\sim}{t}),$$

and we denote our LS estimate with $\hat{\underset{\sim}{\theta}}$. It can be shown (and will be shown) that $\hat{\underset{\sim}{\theta}}$ satisfies

$$A^T A\hat{\underset{\sim}{\theta}} = A^T \underset{\sim}{y}$$

The solution

Given that $A^T A$ is non-singular, then

$$\underset{\sim}{\hat{\theta}} = (A^T A)^{-1} A^T \underset{\sim}{y}.$$

Futhermore we can show that this estimator is unbiased.

We can also show that the variance/covariance matrix of this estimator is

$$\mathbb{D}(\underset{\sim}{\hat{\Theta}}) = \sigma^2 (A^T A)^{-1}.$$

Examples

- i.i.d. random variables
- linear regression
- one way ANOVA

Proof $\underset{\sim}{\hat{\theta}}$ is the least squares estimate

Note that we will be labelling one of our equations as ‘equation awesome’.

Least squares is BLUE

And by BLUE, I obviously mean the best linear unbiased estimator.

As usual, by best we mean most efficient.

Residuals

The residuals are defined as the difference between the observed and fitted values. Hence the residuals are

$$\underset{\sim}{\hat{e}} = \underset{\sim}{y} - A\underset{\sim}{\hat{\theta}}.$$

As usual we take the sum of the squares of the residuals and this should give us some idea about how well the model fits the data.

$$d^2 = \underset{\sim}{\hat{e}}^T \underset{\sim}{\hat{e}} = (\underset{\sim}{y} - A\underset{\sim}{\hat{\theta}})^T (\underset{\sim}{y} - A\underset{\sim}{\hat{\theta}}).$$

Recalling the 'equation awesome' from our proof that $\underset{\sim}{\hat{\theta}}$ is the least squares estimate, and setting $\underset{\sim}{t_0}^* = \underset{\sim}{\hat{\theta}}$,

$$(\underset{\sim}{y} - A\underset{\sim}{t})^T(\underset{\sim}{y} - A\underset{\sim}{t}) = (\underset{\sim}{y} - A\underset{\sim}{\hat{\theta}})^T(\underset{\sim}{y} - A\underset{\sim}{\hat{\theta}}) + (\underset{\sim}{\hat{\theta}} - \underset{\sim}{t})^T A^T A(\underset{\sim}{\hat{\theta}} - \underset{\sim}{t}).$$

Under H_0, we have that $\alpha_0 = 0, \beta_0 = 0$, or that is $\underset{\sim}{t} = 0$. Which then gives

$$\underset{\sim}{y}^T\underset{\sim}{y} = d^2 + \underset{\sim}{\hat{\theta}}A^T\underset{\sim}{y}.$$

Estimation of σ^2

We can show that

$$\mathbb{E}(D^2) = \mathbb{E}((\underset{\sim}{Y} - A\hat{\underset{\sim}{\Theta}})^T(\underset{\sim}{Y} - A\hat{\underset{\sim}{\Theta}})) = (n-p)\sigma^2,$$

hence our unbiased estimator for σ^2 is

$$S^2 = \frac{D^2}{n-p} = \frac{1}{n-p}(\underset{\sim}{Y} - A\hat{\underset{\sim}{\Theta}})^T(\underset{\sim}{Y} - A\hat{\underset{\sim}{\Theta}}).$$

We set $\underset{\sim}{t} = \underset{\sim}{\theta}$ in 'equation awesome', to show that

$$\mathbb{E}(D^2) = (n-p)\sigma^2.$$

Therefore,

$$s^2 = \frac{1}{n-p}(\underset{\sim}{y}^T \underset{\sim}{y} - \hat{\underset{\sim}{\theta}}^T A^T \underset{\sim}{y}).$$

- i.i.d. random variables
- linear regression
- one-way anova

The correlation of residuals and our estimates

We can show that the residual vector and the estimators are uncorrelated. Recall

$$\underset{\sim}{\hat{\Theta}} = (A^T A)^{-1} A^T \underset{\sim}{Y},$$

and

$$\underset{\sim}{\hat{E}} = \underset{\sim}{Y} - A\underset{\sim}{\hat{\Theta}}.$$

It suffices to show that $\mathbb{E}(\underset{\sim}{\hat{\Theta}}\underset{\sim}{\hat{E}}^T) = 0$.

Geometric approach to the general linear model

We have

$$\underset{\sim}{y} = A\underset{\sim}{\theta} + \underset{\sim}{e}.$$

So $\underset{\sim}{y} \in \mathbb{R}^n$ and $\underset{\sim}{\theta} \in \mathbb{R}^p$. Consider the space

$$V = \{\underset{\sim}{v} : \underset{\sim}{v} = A\underset{\sim}{t}, \underset{\sim}{t} \in \mathbb{R}_p\}.$$

Now set $\underset{\sim}{\mu} = \mathbb{E}(\underset{\sim}{Y}) = A\underset{\sim}{\theta} \in V$.

Now draw a surfboard.

The problem

So the question typically is, given a set of observations $\underset{\sim}{y}$, how should we estimate $\underset{\sim}{\mu}$?

Well we know that $\underset{\sim}{e}$ is a realisation of a vector random variable with mean $\underset{\sim}{0}$. Therefore, we should choose $\hat{\mu}$ such that $\hat{\mu}$ is as close as possible to $\underset{\sim}{y}$.

This is exactly a linear projection problem from linear algebra.

Using projections

We can state that

$$A^T \utilde{y} = A^T (A\hat{\utilde{\theta}} + \hat{\utilde{e}}),$$

but $A^T \hat{\utilde{e}} = 0$ by orthogonality.

Hence we end up just as before.

Notes

- Projections are unique, indicating that the $A\underset{\sim}{\hat{\theta}}$ is unique.
- Orthogonality also gives the independence of the residuals and the estimates.
- The triangle and Pythagoras give you the sum of squares decomposition.

Summary of the general linear model

Assuming normality of our observations

- $\underset{\sim}{y} = A\underset{\sim}{\theta} + \underset{\sim}{e}$.
- LS estimates = ML estimates
- $A^T A\hat{\underset{\sim}{\theta}} = A^T \underset{\sim}{y}$
- $\hat{\underset{\sim}{\Theta}} \overset{d}{=} \mathrm{N}_p(\underset{\sim}{\theta}, \sigma^2(A^T A)^{-1})$
- $\hat{\underset{\sim}{E}}$ and $\hat{\underset{\sim}{\Theta}}$ are independent
- $\frac{D^2}{\sigma^2} = \frac{(n-p)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-p}$
- $s^2 = \frac{1}{n-p}(\underset{\sim}{y}^T \underset{\sim}{y} - \hat{\underset{\sim}{\theta}}^T A^T \underset{\sim}{y})$

Commonly used applications of the general linear model

In the following we are going to go through in some detail how to apply in the general linear model setting:

- One-way and Two-way ANOVA
- Multiple linear regression

One way classification

Suppose I have three treatments for high blood pressure, and we measured the following data

	reduction in blood pressure
treatment 1	5, 6
treatment 2	4, 6
treatment 3	3, 3

The model

If we assume the observations are independent and

$$\mathbb{E}(Y_{ij}) = \theta_i, \quad \mathsf{Var}(Y_{ij}) = \sigma^2$$

then we can write it in the form $\underset{\sim}{y} = A\underset{\sim}{\theta} + \underset{\sim}{e}$.

We can then solve for the estimates for $\underset{\sim}{\theta}$.

Alternate parameterisation

Suppose we used the following parameterisation

$$\mathbb{E}(Y_{ij}) = \mu + \tau_i,$$

where μ represents the overall mean blood pressure reduction and τ_i represents the effect of treatment i.

Overparameterisation

So we see that the rank of $A^T A$ is 3, and hence there is no unique solution to $\underset{\sim}{\hat{\theta}}$.

This is the idea of *contrasts* that we have discussed in the past.

Constraints

- $\alpha_1 + \alpha_2 + \alpha_3 = 0$, so that μ represents an overall mean, and the α_i represent deviations from this overall mean. (contr.sum)
- $\alpha_1 = 0$, so that group 1 represents the baseline group, and α_2 and α_3 represent the difference between groups 2 and 3 from 1. (contr.treat)
- $\mu = 0$. This is equivalent to the original parameterisation.
- There are many reasonable constraints that can be used, these are just the commonly used ones.

Two-way classification

Suppose Jeff was interested in the effects of alcohol and sleep on his 'mathematical ability'.

'mathematical ability'	drunk	very drunk
slept previous night	15, 18	12
did not sleep previous night	10	5, 3

The additive model

Consider the model

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j, \quad \mathsf{Var}(Y_{ij}) = \sigma^2,$$

where α_i refers to the drunk status, and β_j refers to the sleep status.

The rank of $A^T A$ is 3. Therefore we need two constraints on the parameters. A common choice is

$$\alpha_1 + \alpha_2 = 0, \quad \beta_1 + \beta_2 = 0.$$

As we shall see, we will not get orthogonality of our estimates. This is due to having unequal numbers in the groups.

Calculate 95% confidence intervals for each of the parameters.

How are these related to the difference between treatments?

How would we test for difference in the effects of sleep?

The difference between orthogonal and not orthogonal

Orthogonal estimates allow for simpler subsequent analysis.

It gives us statistical independence for each parameter estimate.

Classification with interaction

Suppose we had the following data

Jeff’s ‘mathematical ability’	drunk	very drunk	beyond very drunk
slept previous night	15, 18	12,10	5, 3
did not sleep previous night	10, 14	5, 3	-5, -2

The general model

The general model for a two way classification is written as

$$\mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

This model is highly overparameterised.

In our example we have 12 parameters, but we should only need at most 6.

Constraints for interactions model

Typically we choose the constraints

$$\sum_{i=1}^{r} \alpha_i = 0,$$

$$\sum_{j=1}^{c} \beta_j = 0,$$

$$\sum_{j=1}^{c} \gamma_{ij} = 0, \ i = 1, 2, ..., r,$$

$$\sum_{i=1}^{r} \gamma_{ij} = 0, \ j = 1, 2, ..., c.$$

This parameterisation gives us orthogonal estimates.

The procedure is then to first test for interaction.

Multiple linear regression

Often we will have more than one continuous variable we wish to do modelling with.

For example suppose we had the following data:

Jeff's maths 'ability'	BAC	Time of day
10	8	10
5	9	15
14	2	7
8	4	22
9	7	11

Orthogonality

There appears to be some, but limited gain in reparameterising our system.

Just feed it into R.

Polynomial regression

Perhaps we may wish to use a quadratic function?

Jeff's maths 'ability'	Time of day
10	10
5	15
14	7
8	22
9	11

Hypotheses testing

In our general linear model setting we have assumed that

$$\underset{\sim}{Y} \overset{d}{=} \mathrm{N}_n(\underset{\sim}{\mu}, \sigma^2 I).$$

Consider hypotheses of the form

$$H_1 : \underset{\sim}{\mu} = A_1 \underset{\sim}{\theta_1},$$
$$H_0 : \underset{\sim}{\mu} = A_0 \underset{\sim}{\theta_0}$$

where $\underset{\sim}{\theta_1}$ and $\underset{\sim}{\theta_0}$ have p_1 and p_0 parameters, and $\mathrm{rank}(A_1) = r_1$ and $\mathrm{rank}(A_0) = r_0$.

Examples

H_0	H_1
$\mathbb{E}(Y_{ij}) = \mu$	$\mathbb{E}(Y_{ij}) = \mu_i$
$\mathbb{E}(Y_{ij}) = \alpha_0$	$\mathbb{E}(Y_{ij}) = \alpha_0 + \beta u_i$
$\mathbb{E}(Y_{ij}) = \mu + \alpha_i$	$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j$
$\mathbb{E}(Y_{ij}) = \mu_{ij}$	$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j$

The procedure

We will be focusing on cases where H_0 can be seen to be a ‘sub-model’ of H_1.

Our approach is to fit each of the two models specified by both hypotheses and then compare the ‘goodness of fit’ of both models.

If H_0 fits ‘as good’ as H_1 fits, then we will accept H_0.

If H_1 gives a ‘much better’ fit, then we will in turn reject H_0 in preference for H_1.

Art!

Let V_0 be the space spanned by $\{A_1 \underset{\sim}{t_1},\ \underset{\sim}{t_1} \in \mathbb{R}_{r_1}\}$ and correspondingly V_0 the space generated by $\{A_0 \underset{\sim}{t_0},\ \underset{\sim}{t_0} \in \mathbb{R}_{r_0}\}$. Note that we set it up such that $V_0 \subset V_1$.

We see that

$$\underset{\sim}{\hat{e}_0}^T \underset{\sim}{\hat{e}_0} = \underset{\sim}{\hat{e}_1}^T \underset{\sim}{\hat{e}_1} + \underset{\sim}{\delta}^T \underset{\sim}{\delta},$$

or

$$d_0^2 = d_1^2 + \delta^2$$

δ^2 gives some idea about how much the fit 'improves' under H_1 compared to H_0. Hence, the test to reject H_0 is to check if δ^2 is too large.

Decomposition

So from our ‘art’ we have

$$D_0^2 = D_1^2 + \Delta^2.$$

We have seen in the past that under H_0: $D_0^2 \overset{d}{=} \sigma^2 \chi^2_{n-r_0}$ and under H_1 : $D_1^2 \overset{d}{=} \sigma^2 \chi^2_{n-r_1}$.

It can be shown that if H_0 is true then

$$\Delta^2 \overset{d}{=} \sigma^2 \chi^2_{r_1 - r_0}.$$

The F-test

A test for H_0 is therefore given by:

Reject H_0 if

$$F = \frac{\Delta^2/(r_1 - r_0)}{D_1^2/(n - r_1)} > c,$$

where the appropriate value of c is chosen using the fact that if H_0 is true then $F \overset{d}{=} F_{r_1 - r_0, n - r_1}$.

As one might expect, this test can be derived using the likelihood ratio test.

ANOVA

In our ANOVAs we typically decompose into model SS and residual SS. Lets do this for both under H_0 and H_1.

H_0 : model SS, $SS_0 = \underset{\sim}{\hat{\mu_0}}^T \underset{\sim}{\hat{\mu_0}} = \underset{\sim}{\hat{\theta}}_0^T A_0^T \underset{\sim}{y}$, and residual SS $d_0^2 = \underset{\sim}{\hat{e}}_0^T \underset{\sim}{\hat{e}}_0 = \underset{\sim}{y}^T \underset{\sim}{y} - \underset{\sim}{\hat{\theta}}_0^T A_0^T \underset{\sim}{y}$.

H_1 : model SS, $SS_1 = \underset{\sim}{\hat{\mu_1}}^T \underset{\sim}{\hat{\mu_1}} = \underset{\sim}{\hat{\theta}}_1^T A_1^T \underset{\sim}{y}$, and residual SS $d_1^2 = \underset{\sim}{\hat{e}}_1^T \underset{\sim}{\hat{e}}_1 = \underset{\sim}{y}^T \underset{\sim}{y} - \underset{\sim}{\hat{\theta}}_1^T A_1^T \underset{\sim}{y}$.

One way classification

Recall the data:

	reduction in blood pressure
treatment 1	5, 6
treatment 2	4, 6
treatment 3	3, 3

Goodness of fit of a straight line

If our data comes on a limited range of x values, we can test the goodness of fit of a linear regression by comparing it to a one way ANOVA.

How is the linear regression model a special case of the one way ANOVA?

Example

x	y
0	3, 6, 6
1	5, 7, 6
2	7, 9, 8
3	12, 11, 7

The procedure is then to calculate the one way classification ANOVA, and the regression ANOVA, then combine them appropriately and apply a F-test.

Two way classification with unequal numbers

Recall our ‘data’

‘mathematical ability’	drunk	very drunk
slept previous night	15, 18	12
did not sleep previous night	10	5, 3

Models to consider

There are four models we should consider:

$$\begin{array}{ll} \text{M:} & \mathbb{E}(Y_{ijk}) = \mu \\ \text{R:} & \mathbb{E}(Y_{ijk}) = \mu + \alpha_i \\ \text{C:} & \mathbb{E}(Y_{ijk}) = \mu + \beta_j \\ \text{A:} & \mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j \end{array}$$

Lack of orthogonality

When we had equal numbers in each of the cells, we could get the sum squares of model A, SS(A) by simply adding the sum squares of both R and C.

WIth unequal numbers it is not so simple. However the additive model is something that is easily fit in R.

Testing hypotheses

To test for row effects, we compare $H_1 = A$ and $H_0 = C$.

Similarly to test for column effects we compare $H_1 = A$ and $H_0 = R$.

Interactions

Suppose we have a r x c classification problem with m observations per cell. The most general model is to have a dfiferent mean for each cell. That is to fit the model with interactions.

Our model is

$$\text{Model G:} \ \ \mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

Note that if $m = 1$ it is impossible to fit the model with interactions as it becomes impossible to differentiate γ_{ij} and e_{ij}.

Using appropriate constraints, we get orthogonality of the sum squares.

$$SS(G) = SS(\mu) + SS(\alpha) + SS(\beta) + SS(\gamma).$$

The procedure

First we apply the appropriate F-test for interaction.

Then if we do not reject the null hypothesis that the interaction terms are 0, we fit the additive model and consider testing the hypotheses about row/col effects.

Example

Jeff's 'mathematical ability'	drunk	very drunk	beyond very drunk
slept previous night	15, 18	12,10	5, 3
did not sleep previous night	10, 14	5, 3	-5, -2