1 An introduction to probability and distributions

1.1 Probability mass functions, probability density functions, the mean and the variance

Suppose that a random variable X takes a value from the set of possible values $S = (u_1, u_2, \ldots, u_k)$ and the probability that the value u_i is taken is $P(X = u_i) = f(u_i) > 0$. Then the function f is called the *probability mass function (p.m.f.)* of X. Naturally

$$\sum_{x \in S} f(x) = 1$$

The *mean*, or expected value, of X is defined as

$$\mu = \sum_{x \in S} x f(x) = u_1 f(u_1) + \ldots + u_n f(u_n)$$

The *variance* of X is a measure of how spread out the distribution is about this mean, and is given by

$$\sigma^2 = \sum_{x \in S} (x - \mu)^2 f(x)$$

This can be rearranged to the alternative form

$$\sigma^2 = \sum (x - \mu)^2 f(x) = \sum (x^2 - 2x\mu + \mu^2) f(x) =$$

$$\sum x^2 f(x) - 2\mu \sum x f(x) + \mu^2 = \sum x^2 f(x) - \mu^2$$

If we have a continuous distribution (a distribution on a continuous set S), then the probability mass function is replaced by the *probability density function (p.d.f.)*. The probability of any given point is 0, and the probability of a value lying in an interval is given by

$$P[a \leq X \leq b] = \int_a^b f(x) dx$$

Thus

$$\int_{x \in S} f(x) dx = 1$$

and the mean and variance are defined in an analogous way to before, namely the mean of X is defined as

$$\mu = \int_{x \in S} x f(x) dx$$

and the *variance* of X is defined as

$$\sigma^2 = \int_{x \in S} (x - \mu)^2 f(x) dx = \int_{x \in S} x^2 f(x) dx - \mu^2$$

1.2 Some common distributions

We shall give the p.m.f. or p.d.f., the mean and the variance of some common distributions. Each distribution is defined in terms of some key parameters; the mean and variance depend upon the values that these parameters take.

1) The Poisson distribution, parameter λ, has p.m.f.

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots$$

The mean and variance of the Poisson distribution both take value λ.

2) The Binomial distribution $Bin(n, p)$ has p.m.f.

$$f(x) = (^n_x) p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

The mean of the Binomial distribution is np and its variance is $np(1-p)$.

3) The Negative Binomial distribution $NB(m, p)$ has p.m.f.

$$f(x) = (^{x-1}_{m-1}) p^m (1-p)^{x-m} \qquad x = m, m+1, m+2, \ldots,$$

The mean of the Negative Binomial distribution is m/p and its variance is $m(1-p)/p^2$.

4) The Exponential distribution, parameter λ, has p.d.f.

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0$$

The mean of the Exponential distribution is $1/\lambda$ and its variance is $1/\lambda^2$.

5) The Normal distribution $N(\mu, \sigma^2)$ has p.d.f.

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

The mean of the Normal distribution is μ and its variance is σ^2. The standard normal distribution, written as Z, has $\mu = 0, \sigma^2 = 1$ and is written $N(0, 1)$.

In general if X is $N(\mu, \sigma^2)$ then

$$\frac{X - \mu}{\sigma}$$

is Normal (0,1).
If $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$ which is taken from a $N(\mu, \sigma^2)$ distribution, then $\bar{X}$ is $N(\mu, \sigma^2/n)$.

6) The Gamma distribution, parameters α, ν (written $Ga(\alpha, \nu)$), has p.d.f.

$$f(x) = \frac{\nu^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-\nu x} \qquad x > 0$$

where $\Gamma(\alpha)$ is the appropriate constant term so that the integral of the p.d.f. is 1. The mean of the Gamma distribution is α/ν and its variance is α/ν^2. Note that the Gamma distribution is sometimes also written as $Gamma(\alpha, \beta)$ where $\beta = 1/\nu$.

1.3 Sampling distributions

We often take samples from distributions that we assume are normal, but with unknown mean and/or variance. We shall consider three key distributions, the $\chi^2(r)$ distribution, the $T(r)$, distribution and the $F(r_1, r_2)$ distribution. All three of these can be expressed as a function of the standard normal distribution Z. You will see later in the course how important each of these distributions are, and the circumstances in which they are used. You should know that the normal distribution is an awkward one, in the sense that there is no simple closed form for the integral of its p.d.f., and that we make great use of normal tables. As each of these distributions depends on the normal, each also has its own set of tables.

If $Z_1, Z_2, \ldots, Z_r$ are independent standard normal random variables, then

$$U = \sum_{i=1}^{r} Z_i^2$$

has the $\chi^2(r)$ distribution.

If Z is a further standard normal random variable, independent of the others, then

$$T = \frac{Z}{\sqrt{U/r}}$$

has the $T(r)$ distribution.

If U is $\chi^2(r_1)$ and V is $\chi^2(r_2)$ then

$$F = \frac{U/r_1}{V/r_2}$$

has the $F(r_1, r_2)$ distribution.

1.4 The Central Limit Theorem

The importance of the normal distribution comes from the following result, which shows that if we average a number of observations of a particular distribution, even if this distribution is far from normal, it can be well approximated by the normal distribution.

Central Limit Theorem
If $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$, i.e. of size n, from a distribution with a finite mean μ and a finite positive (i.e. non-zero) variance σ^2, then the distribution of

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

converges (in distribution) to the standard normal distribution as $n \to \infty$.

2 Point estimates and confidence intervals

2.1 Point estimates

Suppose that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with mean μ and variance σ^2. Then the natural estimator of the mean of the population is the mean of the data; so if we have actual observations $x_1, \ldots, x_n$ with mean $\bar{x} = \sum x_i/n$ then we estimate μ by $\bar{x}$.

Now consider data that take values of 1 (probability p) or 0 only, so that each data point is a Bernoulli (p) random variable so that the sum of a sample of n such data points, X, is binomial (n, p). The expected value of X is np. Thus if we want to estimate the proportion p, we can do this with x/n.

We will see more theoretically why the above is the case later in the course.

In the rest of this section we shall consider not just our "best guess" for the location of a parameter, but rather find a set of plausible values. Our point estimate will never be precisely correct, and we would like some information on how wrong it is likely to be. Such a set of points will generally take the form of an interval, a *confidence interval.*

2.2 Confidence intervals for a mean

Suppose that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with mean μ and known variance σ^2. The point estimate of μ is $\bar{X}$. We know that $\bar{X}$ itself has a distribution with mean μ and variance σ^2/n.

Thus

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution. Writing z_α as the value from Normal tables for which

$$P[Z > z_\alpha] = \alpha$$

we have

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging the term within the brackets, we can come up with an equivalent expression, which of course has the same probability.

$$
\begin{aligned}
-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \Rightarrow \\
-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow \\
-\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow \\
\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}
\end{aligned}
$$

and so the probability that a random interval

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains μ is $1-\alpha$.

When we obtain a sample the random variable $\bar{X}$ is replaced by the observed mean $\bar{x}$ and we have a $100(1-\alpha)\%$ confidence interval for μ.

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

The centre of this interval is our sample mean, as we would expect, and the size of the interval decreases, so the point estimate becomes more precise, as the number of data points n increases.

Now suppose that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with mean μ and unknown variance σ^2. The point estimator of μ is again $\bar{X}$. This time we shall follow the same procedure, but use the fact that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

has a t distribution with n-1 degrees of freedom, where S^2 is the common unbiased estimator of σ^2 given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Using the term $t_\alpha(m)$ as the value from t-tables which gives

$$P[T(m) > t_\alpha(m)] = \alpha$$

and similarly to before, the probability that a random interval

$$\left[\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

contains μ is $1-\alpha$.

When we obtain a sample the random variables $\bar{X}$ and S^2 are replaced by the observed mean $\bar{x}$ and variance s^2 and we have the following $100(1-\alpha)\%$ confidence interval for μ.

$$\left[\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right]$$

The centre of this interval is again the sample mean, and the size of the interval decreases, as the number of data points n increases.

2.3 Confidence intervals for the difference of two means

Suppose that we are interested in making a comparison between the means of two independent normal distributions. Let our distributions be $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively, where the variances are known but the means are not. We take independent samples $X_1, X_2 \ldots, X_n$ and $Y_1, Y_2 \ldots, Y_m$, with sample means $\bar{X}$ and $\bar{Y}$ having distributions

$$N\left(\mu_X, \frac{\sigma_X^2}{n}\right), N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$$

The distribution of the difference in the means $W = \bar{X} - \bar{Y}$ is thus

$$N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right),$$

Thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

has the standard normal distribution, and so

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \leq z_{\alpha/2}\right) = 1 - \alpha \Rightarrow$$

$$P\left(-z_{\alpha/2}\sigma_w \leq (\bar{X} - \bar{Y}) - (\mu_X - \mu_Y) \leq z_{\alpha/2}\sigma_w\right) = 1 - \alpha \Rightarrow$$

$$P\left((\bar{X}-\bar{Y})-z_{\alpha/2}\sigma_w \le (\mu_X-\mu_Y) \le (\bar{X}-\bar{Y})+z_{\alpha/2}\sigma_w\right)=1-\alpha$$

where $\sigma_w = \sqrt{\sigma_X^2/n+\sigma_Y^2/m}$

After we obtain a sample the random variable $\bar{X}$ is again replaced by the observed mean $\bar{x}$ and we have the following $100(1-\alpha)\%$ confidence interval for the difference in the means $\mu_X - \mu_Y$.

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}\sigma_w, \bar{x}-\bar{y}+z_{\alpha/2}\sigma_w\right]$$

If the variances are unknown, but the sample sizes are large, then we can estimate the variance terms, getting

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}s_w, \bar{x}-\bar{y}+z_{\alpha/2}s_w\right]$$

where s_w is the observed value of σ_w given by $s_w = \sqrt{s_X^2/n+s_Y^2/m}$

If the samples are small we have to take a different approach. We shall assume that the variances are unknown, but equal, so that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. It can be shown that an appropriate confidence interval is given by

$$\left[\bar{x}-\bar{y}-\sqrt{\frac{1}{n}+\frac{1}{m}}s_p t_{\alpha/2}(n+m-2), \bar{x}-\bar{y}+\sqrt{\frac{1}{n}+\frac{1}{m}}s_p t_{\alpha/2}(n+m-2)\right]$$

where s_p^2 is the pooled estimate of the variance, found from

$$s_p^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$$

2.4 Confidence intervals for proportions

Let us now again consider data that take values of 1 or 0 only, so that the sum of a sample of n such data points, X, is binomial (n, p). When finding the confidence interval for such data we use the normal approximation to the binomial distribution. X is approximately normal with mean np and variance $np(1-p)$ and X/n is approximately normal with mean p and variance $p(1-p)/n$. Thus

$$\frac{X-np}{\sqrt{np(1-p)}} = \frac{X/n-p}{\sqrt{p(1-p)/n}}$$

has an approximate standard normal distribution, for sufficiently large n. Thus

$$P\left[-z_{\alpha/2} \le \frac{X/n-p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right] \approx 1-\alpha \Rightarrow$$

$$P\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1 - \alpha$$

We can find the confidence interval by approximating again, and letting $\sqrt{p(1-p)}$ be approximated by

$$\sqrt{\frac{X}{n}\left(1 - \frac{X}{n}\right)}$$

since this is much less variable than X/n itself. This gives

$$P\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}} \leq p \leq \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}}\right] \approx 1 - \alpha$$

giving the confidence interval

$$\left[\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}}, \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}}\right]$$

2.5 Confidence intervals for paired data

Suppose that we have two normal distributions which can have observations taken in pairs (e.g. two measurements on the same patient). If these are X and Y, and we are interested in comparing the means of these distributions, if the values are likely to be dependent, it can make sense to consider the distribution of $X - Y$. Suppose that we have n pairs of dependent measurments $(X_1, Y_1), \ldots, (X_n, Y_n)$ We consider the observations only through $D_i = X_i - Y_i$, and the set $D_1, D_2, \ldots, D_n$ which we consider as a random sample from a normal distribution with mean $\mu_D = \mu_X - \mu_Y$ and variance σ_D^2. If the data are highly correlated, then σ_D^2 can be a lot smaller than σ_X^2 or σ_Y^2. As in the single sample case (assuming the variance is unknown)

$$\frac{\bar{D} - \mu}{S_D/\sqrt{n}}$$

has a t distribution with n-1 degrees of freedom, where S_D^2 is the usual unbiased estimator of σ_D^2. This gives the confidence interval for $\mu_D = \mu_X - \mu_Y$ in the same way as before

$$\left[\bar{d} - t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}, \bar{d} + t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}\right]$$

where $\bar{d}$ and s_D^2 are the observed mean and variance of the sample.

3 An introduction to statistical tests

3.1 Introduction

In this section we shall introduce a very important area of statistical inference, the testing of statistical hypotheses. Suppose that there is good reason to think that a set of observations come from some general class of distributions (e.g. the class of all normal distributions). We are interested in whether they come from a particular sub-class (for example the normal distributions with mean 0). We make a hypothesis that the mean of the underlying distribution of our data is zero, and depending upon information we obtain from the data we decide whether to accept or reject our hypothesis. For instance a process may have in the past had mean 0, so it is possible that the current mean is zero also, but we wish to verify if this is plausible in light of the data.

We state our hypothesis, the *null hypothesis*, and test it against some alternative set of possibilities, the *alternative hypothesis*. Based on the evidence, we either accept or reject the null hypothesis.

There are two possible mistakes that we can make:
A *type I error* is when we reject the null hypothesis when it is in fact true.
A *type II error* is when we accept the null hypothesis when it is in fact false.

We wish to minimise the probability of us making either of these mistakes (though depending on the situation, one mistake may be far more costly than the other).

If the null hypothesis is just represented by a single value, e.g. that the mean is zero, then it is called a *simple hypothesis*. For a simple null hypothesis, the probability of making a type I error is also called the significance level of the test, and is labelled α.

The alternative hypothesis generally consists of a range of possible values; the probability of a Type II error at any given point is called the *power* of the test at that point.

3.2 Testing for the mean with known variance

Suppose that we are sampling from a normal distribution with unknown mean μ and known variance σ^2. We are interested in testing whether the mean μ takes a particular value μ_0. The null hypothesis is thus

$$H_0 : \mu = \mu_0$$

There are three main alternatives for the alternative hypothesis

$$1) H_1 : \mu > \mu_0$$

μ has increased

$$2)H_1 : \mu < \mu_0$$

μ has decreased

$$3)H_1 : \mu \neq \mu_0$$

μ has changed, but it is not known whether it has increased or decreased.

To test the null hypothesis against one of these three alternative hypotheses, a random sample of n data points is taken, and the observed sample mean $\bar{x}$ is found. If $\bar{x}$ is close to μ_0, this lends support to the null hypothesis. What is meant by "close" depends upon the variability of the data, and how much of it we have. The standard error of the mean is $\sigma/\sqrt{n}$ where σ is the known standard deviation of the distribution. The test statistic Z is defined by

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

and will be an observation from a Normal (0,1) distribution if the null hypothesis is true. If it is false, then Z will tend to be further from 0. The critical region at a significance level α (the set of values of z where we reject H_0) for the three alternative hypotheses are

$$1)H_1 : \mu > \mu_0 \quad z \geq z_\alpha$$

$$2)H_1 : \mu < \mu_0 \quad z \leq -z_\alpha$$

$$3)H_1 : \mu \neq \mu_0 \quad |z| \geq z_{\alpha/2}$$

For example, if $\alpha = 0.05$ then from normal tables we obtain $z_{0.05} = 1.645$ and $z_{0.05/2} = z_{0.025} = 1.960$.

3.3 Testing for the mean with unknown variance

Usually the variance σ^2 is unknown. How do we test the above null hypothesis in this case?

For a random sample of n data points from a normal distribution our test statistic becomes

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and will be an observation from a t distribution with $n-1$ degrees of freedom, and will again tend to be further from 0 if the null hypothesis is not true. We have replaced the true variance of the sample mean σ^2/n by its (unbiased) estimate s^2/n. This increases the variability of our statistic, so that the region where H_0 is accepted become slightly larger than before (the critical region becomes smaller). The critical region at a significance level α (the set of values of z where we reject H_0) for the three alternative hypotheses are

$$1)H_1 : \mu > \mu_0 \quad t \geq t_\alpha(n-1)$$

$$2)H_1 : \mu < \mu_0 \quad t \leq -t_\alpha(n-1)$$

$$3)H_1 : \mu \neq \mu_0 \quad |t| \geq t_{\alpha/2}(n-1)$$

3.4 the p-value of a test

The p-value is the probability of getting a result at least as extreme as the observed result, under the assumption that the null hypothesis is true. Thus it is a measure of how unlikely it would be to obtain the type of result that we did under the null hypothesis.

The smaller the p-value is, the greater the evidence against the null hypothesis. What is considered as "more extreme" depends upon which alternative hypothesis is being used. The concept of the p-value is relevant to any hypothesis testing situation. In the case of known variance, the p-value for each of the three possible alternative hypotheses is as follows

$$1)H_1 : \mu > \mu_0 \quad p = P[Z \geq z]$$

$$2)H_1 : \mu < \mu_0 \quad p = P[Z \leq z]$$

$$3)H_1 : \mu \neq \mu_0 \quad p = P[|Z| \geq |z|]$$

where Z is a Normal(0,1) random variable and z is the particular value of our statistic.

Note that in each case if z lies on the boundary of the critical region, the p value is α. The null hypothesis is rejected if and only if the p-value of the test is less than α

3.5 Testing for proportions

Suppose that we have a situation where each data value is not on some continuous scale, but rather is represented by either a 1 or a 0 (success or failure of an engineering component, yes or no in a social survey). On the assumption that each observation is independent and has probability p of being a 1, the total number of 1s in a sample of size n, and so the sum of all the data values, can be assumed to come from a binomial distribution with parameters n and p.

We are typically interested in the value of p, the proportion of 1s in the population at large. As before we consider the null hypothesis

$$H_0 : p = p_0$$

There are again three main alternatives for the alternative hypothesis, p has increased, p has decreased or p has changed, but it is not known whether it has increased or decreased.

In this situation it is generally not possible to find a test with a significance level taking a given precise value of α, since we are considering a discrete distribution, although we can do this approximately if n is sufficiently large.

If n (and np) is sufficiently large, then the binomial distribution can be approximated by a normal dsitribution with the same mean and variance, so that under the null hypothesis Y is approximately normal with mean np_0 and variance $np_0(1 - p_0)$. Substituting these for μ and σ^2 in our normal tests we obtain the critical region for each of the alternative hypotheses as

$$1) H_1 : p > p_0 \quad z = \frac{y/n - p_0}{\sqrt{p_0(1 - p_0)/n}} \geq z_\alpha$$

$$2) H_1 : p < p_0 \quad z = \frac{y/n - p_0}{\sqrt{p_0(1 - p_0)/n}} \leq -z_\alpha$$

$$3) H_1 : p \neq p_0 \quad |z| = |\frac{y/n - p_0}{\sqrt{p_0(1 - p_0)/n}}| \geq z_{\alpha/2}$$

3.6 Testing for the equality of means

Firstly we consider testing for equality of the means. We shall treat this in a similar way as we did when finding a confidence interval for the difference in the means in an earlier section; in particular we consider the same t-statistic as before, on the assumption that the two variances are equal. We test the hyptohesis

$$H_0 : \mu_X = \mu_Y$$

against one of the three obvious alternatives. If the null hypothesis is true then

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) S_p^2}}$$

has a t distribution with $n+m-2$ degrees of freedom, where S_p^2 is again the pooled estimate of the variance

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Then we reject the null hypothesis if this observed value is sufficiently far from zero (in the relevant direction, if appropriate) and this yields the critical regions for the three possible alternative hypotheses as

$$1) H_1 : \mu_X > \mu_Y \qquad t \geq t_\alpha(n+m-2)$$

$$2) H_1 : \mu_X < \mu_Y \qquad t \leq -t_\alpha(n+m-2)$$

$$3) H_1 : \mu_X \neq \mu_Y \qquad |t| \geq t_{\alpha/2}(n+m-2)$$

Note that in the two sided test, the complement of the critical region $|t| \geq t_{\alpha/2}(n+m-2)$ is equivalent to

$$-\sqrt{\frac{1}{n}+\frac{1}{m}} s_p t_{\alpha/2}(n+m-2) \leq \bar{x}-\bar{y} \leq \sqrt{\frac{1}{n}+\frac{1}{m}} s_p t_{\alpha/2}(n+m-2)$$

so that we accept $H_0 : \mu_X - \mu_Y = 0$ if and only if 0 lies in the confidence interval

$$\left[\bar{x}-\bar{y}-\sqrt{\frac{1}{n}+\frac{1}{m}} s_p t_{\alpha/2}(n+m-2), \bar{x}-\bar{y}+\sqrt{\frac{1}{n}+\frac{1}{m}} s_p t_{\alpha/2}(n+m-2)\right]$$

This relationship between two-sided tests and confidence intervals is in fact quite general.

Note that if we know the variances of our two distributions (whether they are equal or not) we use the statistic

$$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

which has the standard normal distribution when the null hypothesis is true. Similarly if we do not know the variances, but our samples are large we can similarly use

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

Finally, what if the samples are not large, but we do not believe that the variances are equal (in particular we think they are far from equal). In such circumstances it is best not to use the t-statistic as described above. In this case a modified t-statistic is used.

There is a wide range of different t- distributions (dependening on the degrees of freedom) so we need to find the degrees of freedom which mostly correspond with the underlying distribution that we will use. There is more than one way of estimating this. Welch's formula gives the estimate of this (expressed as r) as

$$\frac{1}{r} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$$

where

$$c = \frac{s_x^2}{s_x^2 + s_y^2}$$

4 Descriptive statistics, graphical methods and non-parametric tests

Before analysing any data, it is always advisable to represent it pictorially in some way. This is because there are many ways to anaylse data, and the best way depends upon the type of data involved. You may think you know what type of data you have, in which case a plot may suggest possible errors or omissions in the data. We will describe several simple ways of plotting data which do not involve a lot of work, but may save you a lot! We follow this by considering two examples of statistical tests which do not assume the data follow some particular type of distribution (although this does not mean that we make no assumptions) which can be applied in a wide range of circumstances.

4.1 Bar charts, stem and leaf plots and histograms

Quantitative data can be represented in many ways. Three related pictorial representations are bar charts, stem and leaf plots and histograms. Consider the following example, which describes the amount of money spent in the fiscal year 2000 by the US Department of Defence in various categories

Category	Amount (billions of dollars)
Military personnel	76.0
Operation and maintenance	105.9
Procurement	51.6
Research and development	37.6
Military construction	5.1
Total	276.2

This data can be represented both as a bar chart. In the bar chart, each category simply gets a bar whose height corresponds to size of its numerical entry.

If we have a set of numerical data, expressed in decimal form, $x_1, \ldots, x_n$ we can categorise this data by allocating the data in a particular interval to a certain category. The simplest way to do this is to divide each entry into two parts the *stem* and the *leaf*. The stem is the more significant part, and each type of stem will represent a category. For example, our stem could be the data cut off at the first decimal place, and the leaf is the remainder. Thus 2.59 would have stem 2.5 and leaf (9). Every data point with stem 2.5 will be grouped together as a single category.

Suppose we have the following set of data
2.41,2.57,2.35,2.54,2.49,2.37,2.50,2.51,2.32,2.46,2.48,2.44,2.52,2.38,2.47

Choosing the value of the data cut off at the first decimal place, we obtain the following stem and leaf plot

2.3	5728
2.4	196847
2.5	74012

which we rearrange in numerical order as

2.3	2578
2.4	146789
2.5	01247

The aim of this idea is to get an impression of the shape of the distribution. Sometimes the numbers are such that dividing the stems just by the "cutting off" method above is not very informative, and we have to divide the stems differently. For instance it might be convenient to have two categories for 2.5?, with 2.50-2.54 in category 2.5a, and 2.55-2.59 in category 2.5b. The above stem and leaf plot now becomes

2.3	2
2.3	578
2.4	14
2.4	6789
2.5	0124
2.5	7

Thus the stem and leaf plot is a version of a bar chart. Note that, unlike the earlier version with the military data, there is a natural order of the categories, which it would not make sense to rearrange.

A histogram is a pictoral representation of the data, similar in form to a bar chart but with important differences. Suppose again that we have a set of numerical data $x_1, \ldots, x_n$ that we can divide into categories. In a histogram these categories do not have to be equal width. For each interval the **area** of the block representing it is the same as the proportion of data points within that category. Thus if 6 out of 15 data points lie in a category covered by an interval of width 0.5, the proportion of data points is 6/16=0.4, and the height of the block is therefore 0.4/0.5=0.8. The sum of the areas of all the blocks is equal to 1, and the histogram gives an approximation to the shape of the p.d.f. (or p.m.f) of the underlying distribution.

4.2 Five number summary and boxplot

We introduce a way to represent a set of data by five numbers only, whilst retaining most of the important information. Suppose that we have a set of data points $x_1, \ldots, x_n$. We shall now reorder the data points in order of size $y_1, \ldots, y_n$ so that y_1 is the smallest of the xs, y_2 the next smallest up to y_n the largest. The *median, m* of the data is given by the data value in position $(n+1)/2$. Thus it is

$$m = y_{(n+1)/2}(oddn), m = \frac{y_{n/2} + y_{n/2+1}}{2}(evenn)$$

The *lower quartile, Q1* is the data value in position $(n+1)/4$ and similarly the *upper quartile, Q3* is the data value in position $3(n+1)/4$. As in the case of the median, this may not be an integer value, in which case interpolation is used between the two adjacent values. Finally the *minimum* of the data is just y_1 and the *maximum* is y_n. A useful measure of the spread of the data is the *interquartile range*, given by Q3-Q1.

These five numbers (minimum, lower quartile, median, upper quartile and maximum) together make up the five number summary. They provide all of the information required to produce a boxlot of the data.

A good way to make a preliminary comparison between two data sets and their important features such as location, spread and skewness is to produce back to back boxplots.

4.3 Cumulative distribution plots and Normal Q-Q plots

The empirical cumulative distribution function is defined as the fraction of the data smaller than the value y, over all values of y. Thus if x is the kth smallest of our n data values, the value of the function will be $k - 1/n$ for y just below x, and jump to k/n at $y = x$. The idea of this plot, is that the empirical function will look like the distribution function of the underlying distribution, and by looking at it you can get some idea of what that function might be.

Often we are interested in whether the data are normally distributed. For a better method, we can plot the kth smallest observation against the expected value of this out of n observations for a standard normal distribution. if the data come from some normal distribution, this plot should look like a straight line. This is not especially easy to do ‘by hand' but most statistical software packages can do it. It is done in R using qqnorm(x)

4.4 The sign test

Suppose that we have a random sample $X_1, \ldots, X_m$ from some unknown distribution with median m, which we cannot assume to be normal. We wish to test the hypothesis that the median takes some value m_0, against the alternative that it does not i.e.

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

Under the null hypothesis each data point is equally likely to be above or below m_0. If we let Y be the number of observations of X_i which are less than m_0 then under H_0 Y has a Binomial $(n, 0.5)$ distribution (which is symmetric about its mean/median $n/2$). If H_0 is true then values of Y near $n/2$ are more likely than if it is false. Therefore we reject H_0 if our observed value of Y is too far from $n/2$ in either direction. It is again possible to consider either of the one-sided test, so we will have critical regions

$$1)H_1 : m > m_0 \quad Y \leq k_1$$

$$2)H_1 : m < m_0 \quad Y \geq k_2 = n - k_1$$

$$3)H_1 : m \neq m_0 \quad Y \leq \frac{n}{2} - k_3 or Y \geq \frac{n}{2} + k_3$$

A common use of this test is when comparing paired data Y_i and Z_i, and wanting to test whether the two distributions are the same. In this case the mean and the median of $X = Y - Z$ would be zero.

4.5 The Wilcoxon Test

Suppose that we now have two random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from some unknown distributions with means μ_X and μ_Y respectively, which we cannot assume to be normal. We wish to test the hypothesis that the distributions are the same, against the alternative that they are not. There are $m + n$ values; suppose that we order them (irrespective of which sample they come from) from smallest to largest.

If the two distributions were identical, then every possible ordering would be equally likely. There are $(m+n)!$ of them, so each has probability $1/(m+n)!$. Let r_j be the ordering, or *rank*, of Y_j in the pooled group. We shall consider the statistic

$$R = \sum_{j=1}^{n} r_j$$

If the mean μ_Y of the underlying distribution of the Y_js is larger than that of the X_is (μ_X) then R will tend to be large, if it is smaller it will be small, and if they are the same it will be intermediate. There are the three usual alternative hypotheses, with critical regions

$$1) H_1 : \mu_X > \mu_Y \quad R \leq k_1$$

$$2) H_1 : \mu_X < \mu_Y \quad R \geq k_2$$

$$3) H_1 : \mu_X \neq \mu_Y \quad R \leq k_{31} or R \geq k_{32}$$

The values of the ks can be found from tables.

It is also possible to use a version of the Wilcoxon test to consider the median of a distribution as in the sign test above. We are able to utilise more information than in the sign test, and hence the test will be more powerful. We have a random sample $X_1, \ldots, X_m$ from some unknown distribution with median m, and we wish to test the null hypothesis that the median takes some value m_0. We transform our data by allocating $X_i - m_0$ to the series Y if $X_i - m_0 > 0$ and $-(X_i - m_0)$ to the series Z if $X_i - m_0 < 0$.

Thus we have two series Y which are the sizes of the positive differences (above the supposed median m_0), and Z are the sizes of the negative differences. We rank the Y and Zs together as above, and then consider the statistic W which is the sum of the ranks of the Ys minus the sum of the ranks of the Zs. If H_0 is true then about half of the differences will be positive and of about the same size as the negative differences, so W will be near 0. If H_0 is not true then W will be tend to be positive if the median is larger than m_0 and negative if it is less. Thus we will have

$$1) H_1 : m > m_0 \quad W \geq k_1$$

$$2) H_1 : m < m_0 \quad W \leq k_2$$

$$3) H_1 : m \neq m_0 \quad W \leq k_{31} or W \geq k_{32}$$

In particular when n is sufficiently large, under H_0, W is approximately Normal with mean 0 and variance

$$\frac{n(n+1)(2n+1)}{6}$$

so that

$$\frac{W}{n(n+1)(2n+1)/6}$$

is approximately N(0,1).

5 Regression and correlation

5.1 Simple linear regression

Let us consider a situation where we have a set of paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we are interested in finding the relationship between the two values; in particular we are interested in predicting what the value of Y may be if we know the value of X. The series X may be something that is in our control to vary, or simply something we can measure accurately.

We shall assume a straightforward relationship between Y and X, namely that

$$E[Y] = \alpha + \beta X$$

so that on average the relationship between the two is linear. In addition there is a random component, so that we have a probabilistic model

$$Y = \alpha + \beta X + \epsilon$$

where ϵ is an error term, assumed to come from a normal distribution with mean 0 and variance σ^2 for some unknown σ^2. In addition it is assumed that every value of ϵ is independent of every other.

5.2 The method of least squares

Our aim is to find the line which fits the data best. We want to minimise the distance of the data points from the line in some sensible way. Given that it is assumed that we control the x_is but not the y_is, the logical method is to consider the vertical distances from the data points to the fitted line. The *method of least squares* seeks to minimise the sum of the squares of this distance over all of the data points. There are other possible ways to fit the line, but this is the most common (by far).

We thus wish to minimise the sum

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

by choosing the values of α and β appropriately. We differentiate with respect to α and β in turn and set equal to zero.

$$\frac{\partial}{\partial \alpha} S(\alpha, \beta) = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \Rightarrow$$

$$\frac{1}{n}\left(\sum y_i - n\alpha - \beta\sum x_i\right) = \bar{y} - \alpha - \beta\bar{x} = 0$$

$$\frac{\partial}{\partial\beta}S(\alpha,\beta) = \sum_{i=1}^{n} -x_i(y_i - \alpha - \beta x_i) \Rightarrow$$

$$\frac{1}{n}\left(\sum x_i y_i - \alpha\sum x_i - \beta\sum x_i^2\right) = \frac{1}{n}\sum x_i y_i - \alpha\bar{x} - \beta\frac{1}{n}\sum x_i^2 = 0$$

Setting a as the value of α and b as the value of β which solve these equations simultaneously, and so become our estimates of α and β, we obtain

$$a = \bar{y} - b\bar{x}$$

and

$$\frac{1}{n}\sum x_i y_i - (\bar{y} - b\bar{x})\bar{x} - b\frac{1}{n}\sum x_i^2 = 0 \Rightarrow$$

$$\frac{1}{n}\sum x_i y_i - \bar{y}\bar{x} = b\left(\frac{1}{n}\sum x_i^2 - (\bar{x})^2\right) \Rightarrow$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$S_{xx} = \sum x_i^2 - n(\bar{x})^2, S_{xy} = \sum x_i y_i - n\bar{y}\bar{x}$$

We can estimate the variance of the error term, σ^2, by substituting the estimates of α and β into the sum of squares, and weighting by the number of observations, i.e.

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y} + b\bar{x} - bx_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{1}{n}2b\sum_{i=1}^{n}(y_i - \bar{y})(\bar{x} - x_i) + \frac{1}{n}b^2\sum_{i=1}^{n}(\bar{x} - x_i)^2$$

$$= \frac{1}{n}\left(S_{yy} - 2bS_{xy} + b^2S_{xx}\right) = \frac{1}{n}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

where

$$S_{yy} = \sum y_i^2 - n(\bar{y})^2$$

5.3 Assessing the fit of a regression model

A common measure of the strength of the linear relationship between X and Y is related to the correlation coefficient, r, which we shall meet later.

$$r^2 = \frac{(S_{XY})^2}{S_{XX}S_{YY}}$$

is called the *coefficient of determination* and can be thought of as the proportion of the variation within the data explained by the linear regression of Y on X. Thus if r^2 is large (close to 1), then knowledge of X gives a great deal of information about what Y will be, whereas if it is small (near 0) then very little information is provided.

It is possible that there is no relationship between X and Y, of course, but it is also possible that there is a relationship which is not linear. How can we tell if this is the case?

The most obvious thing to do is plot the data and see if there appears to be a linear relationship between X and Y. Amore detailed investigation would involve fitting the regression line and plotting the residuals

$$y_i - a - bx_i$$

against x_i. If the regression model is a good one, then there should be no pattern to this data, since

$$Y_i - \alpha - \beta X_i$$

is a normal random variable with mean 0 and variance σ^2, independent of all others. If there is a clear pattern, then the linear model is unlikely to be true. Particular things to look out for include bunches of positive values, then negative values, then positive again (suggesting a curved relationship) or increasing variance with the size of X (possibly a logarithmic relationship). There are other things to look for, for instance a normal probability plot could show whether the residuals were likely to be from a normal distribution (as we have assumed).

Note that a more formal way to test the validity of the model is to use *Analysis of Variance* which we shall meet later in the course.

5.4 Transforming the data

If it appears that there is not a linear relationship between X and Y it may be possible to transform the data to new values where there is a linear relationship. It may be that the relationship relating Y and X (neglecting the error terms) is

$$Y = \alpha exp(\beta X)$$

in which case, taking logarithms gives

$$log(Y) = log(\alpha) + \beta X$$

so we can regress $log(Y)$ against X. Similarly

$$Y = \alpha X^{\beta} \Rightarrow log(Y) = log(\alpha) + \beta log(X)$$

so we regress $log(Y)$ on $log(X)$. If

$$Y = \alpha + \frac{\beta}{X}$$

we can regress Y on $1/X$ etc.

It should be noted that the underlying error structure of the data is important, and should be regarded as an integral part of the model. For instance

$$Y_i = \epsilon_i \alpha exp(\beta X_i) \Rightarrow log(Y_i) = log(\alpha) + \beta X_i + log(\epsilon_i)$$

so that $log(\epsilon)$ should be $N(0, \sigma^2)$, not ϵ itself.

5.5 Correlation

The related topic of correlation deals with the strength of linear relatedness between two random variables, X and Y. If X is large, does this mean that Y is more likely to be large, to be small or does this tell us little about Y?

Suppose that X has mean μ_X and variance σ_X^2, and Y has mean μ_Y and variance σ_Y^2. We further define the *covariance* of X and Y, $Cov(X, Y)(= \sigma_{XY})$, as

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

The *correlation coefficient* of X and Y is defined as

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Thus we have;
1) if $Y = X$, then $\sigma_{XY} = \sigma_X^2 = \sigma_Y^2$ and $\rho = 1$, perfect positive correlation.
2) if $Y = -X$, then $\sigma_{XY} = -\sigma_X^2 = -\sigma_Y^2$ and $\rho = -1$, perfect negative correlation.
3) if Y and X are independent, then

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y = E[X]E[Y] - \mu_X \mu_Y = 0$$

and so $\rho = 0$, zero correlation. Note that independence implies zero correlation, but zero correlation does not necessarily mean independence.

We have considered the true, theoretical, correlation of the distributions. As in the case of linear regression, if we are presented with a set of paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we are interested in the correlation of X and Y, we will have to estimate it. The sample correlation coefficient is found using the estimates

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

$$\sigma_{\hat{X}Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

to give

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

$$\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

6 The analysis of variance

Previously we have looked at how to compare the means of two normal distributions. Now we consider how to consider a large number of distributions.

6.1 Testing the equality of several means

Suppose that we have m normal distributions, with unknown means $\mu_1, \mu_2, \ldots, \mu_m$ respectively, and an unknown but common variance σ^2. We consider the null hypothesis

$$H_0 : \mu_i = \mu \quad i = 1, \ldots, m$$

for some unspecified μ, against the alternative hypothesis that this is not true, so that there is at least some pair of means which are different.

Let $X_{i1}, X_{i2}, \ldots, X_{ini}$ represent a random sample of size n_i from the distribution $N(\mu_i, \sigma^2)$, for each i.

Further denote the mean of sample i as

$$\bar{X_{i.}} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

and the overall mean of all the samples as

$$\bar{X_{..}} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}$$

where $n = n_1 + \ldots + n_m$.

We wish to form a test of our null hypothesis. We shall initially consider the sum of squares associated with the overall variance of the data, and partition this into two components. We label this the *total sum of squares, SS(T).*

$$SS(T) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X_{..}})^2 =$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X_{i.}} + \bar{X_{i.}} - \bar{X_{..}})^2 =$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X_{i.}})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{X_{i.}} - \bar{X_{..}})^2 +$$

$$2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X_{i.}})(\bar{X_{i.}}-\bar{X_{..}})$$

This third term is just zero, since we can rearrange it as

$$2\sum_{i=1}^{m}(\bar{X_{i.}}-\bar{X_{..}})\sum_{j=1}^{n_i}(X_{ij}-\bar{X_{i.}}) = 2\sum_{i=1}^{m}(\bar{X_{i.}}-\bar{X_{..}})(n_i\bar{X_{i.}}-n_i\bar{X_{i.}}) = 0$$

In a similar way we can rewrite the second term

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X_{i.}}-\bar{X_{..}})^2 = \sum_{i=1}^{m}n_i(\bar{X_{i.}}-\bar{X_{..}})^2$$

Thus we have

$$SS(T) = SS(B) + SS(W)$$

where the *sum of squares within groups, SS(W)* is given by

$$SS(W) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X_{i.}})^2$$

and the *sum of squares between groups, SS(B)*, is given by

$$SS(B) = \sum_{i=1}^{m}n_i(\bar{X_{i.}}-\bar{X_{..}})^2$$

If the null hypothesis is true, then every observation just comes from a single normal distribution with mean μ and variance σ^2. In this case $SS(T)/\sigma^2$ has a $\chi^2(n-1)$ distribution, and $SS(T)/(n-1)$ is unbiased for σ^2.

Considering a single sample i, an unbiased estimator for σ^2 is given by

$$W_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}-\bar{X_{i.}})^2$$

as $(n_i-1)W_i/\sigma^2$ is $\chi^2(n_i-1)$. We can add these terms for all m samples to give

$$\sum_{i=1}^{m}\frac{(n_i-1)W_i}{\sigma^2} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\sigma^2}(X_{ij}-\bar{X_{i.}})^2 = \frac{SS(W)}{\sigma^2}$$

which thus has a χ^2 distribution with $\sum_{i=1}^{m}(n_i-1) = n-m$ degrees of freedom. Thus $SS(W)/(n-m)$ is an unbiased estimator of σ^2.

It can be shown that $SS(W)$ and $SS(B)$ are independent, and that since $SS(W)/\sigma^2$ is $\chi^2(n-m)$ and $SS(T)/\sigma^2$ is $\chi^2(n-1)$ then $SS(B)sigma^2$ is $\chi^2(m-1)$.

6.2 Analysis of variance

Some of the above is based upon the assumption that H_0 is true, and some is not. In particular $SS(W)$ is $\chi^2(n-m)$ whether or not H_0 is true, but the same is not true for the χ^2 distribution of SS(T). An important piece of information that we can use, is that if all the means are not equal $SS(T)$, and hence $SS(B)$, will tend to be larger than if they are. We show this below by considering the expectation of SS(B).

$$E[SS(B)] = E\left[\sum_{i=1}^{m} n_i(\bar{X_{i.}} - \bar{X_{..}})^2\right] = E\left[\sum_{i=1}^{m} n_i\bar{X_{i.}}^2 - n\bar{X_{..}}^2\right] =$$

$$\sum_{i=1}^{m} n_i(Var(\bar{X_{i.}}) + E(\bar{X_{i.}})^2) - n(Var(\bar{X_{..}}) + E(\bar{X_{..}})^2) =$$

$$\sum_{i=1}^{m} n_i\left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$(m-1)\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i^2 - \mu^2) = (m-1)\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i^2 - 2\mu_i\mu + \mu^2) =$$

$$(m-1)\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i - \mu)^2$$

where

$$\mu = \frac{1}{n}\sum_{i=1}^{m} n_i\mu_i$$

Thus if the null hypothesis is true, the expectation of SS(B)/(m-1) is σ^2, and if it is false it is some larger value. Moreover, in some sense, the larger the departure from the null hypothesis that the truth is, the larger this expectation is.

Both of the terms SS(B) and SS(W) contain the unknown variance term σ^2. We can remove this from our calculations by considering the ratio of these two terms.

$$\frac{SS(B)/(m-1)}{SS(W)/(n-m)} = \frac{SS(B)/\sigma^2(m-1)}{SS(W)/\sigma^2(n-m)} = F$$

where F is our test statistic. Under H_0 F has an F distribution with $m-1$ and $n-m$ degrees of freedom, as $SS(B)/\sigma^2$ and $SS(W)/\sigma^2$ are χ^2 and independent. If H_0 is false, the observed F will tend to be too large. Thus the critical region is

$$F \geq F_\alpha(m-1, n-m)$$

The relevant information is usually summarised in an *analysis of variance* table, see below. The mean square term is just the sum of squares divided by the degrees of freedom, so that under H_0 mean squares should be of comparable size. Note that "analysis of variance" is concerned with comparing the means, and not the variances.

Source	Sum of squares	df	Mean square	F-ratio
Between groups	SS(B)	$m-1$	MS(B)=SS(B)/(m-1)	MS(B)/MS(W)
Within groups	SS(W)	$n-m$	MS(W)=SS(W)/(n-m)	
Total	SS(T)			

Suppose that we reject H_0, and so decide that there is some difference between the groups. We may well be interested in which groups differ from which others. We can perform a pairwise test on each pair of groups in term, following a procedure similar to that from Section 3.

Whether H_0 is true for false, we can estimate σ^2 by $SS(W)/(n-m)$, so that the means of groups i and j can be judged as different if

$$\frac{|\bar{X_{i.}} - \bar{X_{j.}}|}{\sqrt{(1/n_i + 1/n_j)SS(W)/(n-m)}} > t_{\alpha/2}(n-m)$$

However, as there are m such groups, there are $m(m-1)/2$ tests being carried out if we wish to compare every pair of means. Given that for each test we 'find' a difference even when there is not one $100\alpha\%$ of the time (typically 5%), if we carry out enough tests it is almost certain that we will find numerous false positives. Thus, following the method of Bonferroni (there are other more complex methods which can be better) we correct for this by 'raising the bar' and making it more difficult to reject the null hypothesis in each case, by dividing our significance level α by the number of tests $m(m-1)/2$.

$$\frac{|\bar{X_{i.}} - \bar{X_{j.}}|}{\sqrt{(1/n_i + 1/n_j)SS(W)/(n-m)}} > t_{\alpha/(m(m-1))}(n-m)$$

This multiple testing problem is particularly common in genetic analysis, where there can be thousands of such comparisons taking place.

7 Analysing tabular data

7.1 Goodness of fit tests

Consider an experiment where there are k possible outcomes $A_1, A_2, \ldots, A_k$ and set p_j to be the probability that the outcome of a particular trial is A_j, so that $\sum p_j = 1$.

The experiment is repeated n times ,and we let Y_j be the number of times that the outcome A_i occurs. The distribution of $Y_1, \ldots, Y_k$ follows a *multinomial distribution*, namely the probability mass function is given by the formula

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = f(y_1, y_2, \ldots, y_{k-1}, y_k) = \frac{n!}{y_1! y_2! \ldots y_k!} p_1^{y1} \ldots p_k^{yk}$$

Note that there are are only $k-1$ "free" terms in this expression, as $\sum Y_j = n$, so that once the first $k-1$ Y_js are known, Y_k is determined.

We claim that

$$Q = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$$

has, approximately, a $\chi^2(k-1)$ distribution.

This approximation works well provided that all of the $np_j \geq 5$ (and sometimes for smaller values).

We shall see why this works in the case when $k = 2$. In this case Y_1 is binomial (n, p_1) and so the central limit theorem tells us that

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

has an approximate Normal (0,1) distribution for large n (in particular when np_1 and $n(1-p_1)$ are at least 5). Thus the square of Z

$$Z^2 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)}$$

has a $\chi^2(1)$ distribution.

$$Z^2 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)} =$$

$$\frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} = Q$$

since $p_2 = 1 - p_1$ and $(Y_1 - np_1) = -(Y_2 - np_2)$. Thus Q is $\chi^2(1)$.

The fact that in general Q is $\chi^2(k-1)$ can be used to test hypotheses about the parameters p_j in a similar way to before.

The standard null hypothesis is of the form

$$H_0 : p_j = p_{j0} \quad j = 1, \ldots, k$$

i.e. that the p_js take some particular set of known values p_{j0}. The alternative hypothesis will usually simply be that this is not true, and that the p_js take some other unspecified set of values, i.e.

$$H_1 : p_j \neq p_{j0}$$

for at least one (in practice this means at least 2) values of j.

In its simplest case this could reduce to

$$H_0 : p_j = 1/k \quad j = 1, \ldots, k$$

so that all probabilities are equal.

7.2 Contingency Tables

One use of the test discribed above is that concerning contingency tables. Consider several multinomial distributions simultaneously. We thus have a number of categories (h) each of which are split into a number of classes (k). Assume that data from category i have probability p_{ij} of being in class j, $j = 1, \ldots, k$ and $i = 1, \ldots, h$. Thus

$$\sum_{j=1}^{k} p_{ij} = 1 \quad i = 1, \ldots, h$$

We test the null hypothesis that all of these distributions are the same

$$H_0 : p_{1j} = p_{2j} = \ldots = p_{hj} = p_j \quad j = 1, \ldots, k$$

against the alternative hypothesis that this is not true.

Supposing that we take a sample of size n_i from the ith distribution, the number Y_{ij} from sample i being of type j, we can obtain the following table

Y_{11}	Y_{12}	$\dots$	Y_{1k}	n_1
Y_{21}	Y_{22}	$\dots$	Y_{2k}	n_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
Y_{h1}	Y_{h2}	$\dots$	Y_{hk}	n_h

We can adapt the result from before that

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i p_j)^2}{n_i p_j}$$

has, approximately, a $\chi^2(h(k-1))$ distribution. We do not know the value of p_j, and so estimate them by pooling the data to give

$$\hat{p_j} = \frac{\sum_{i=1}^{h} Y_{ij}}{\sum_{i=1}^{h} n_i} = \frac{m_j}{n} \qquad j = 1, \dots, k-1$$

where m_j is the total number of occurrences occurrences of event j over all distributions, and n is the total number of all occurrences. Note that the estimate of $p_k, \hat{p_k}$ is just $1 - \hat{p_1} - \dots - \hat{p_{k-1}}$. Our complete table is now

Y_{11}	Y_{12}	$\dots$	Y_{1k}	n_1
Y_{21}	Y_{22}	$\dots$	Y_{2k}	n_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
Y_{h1}	Y_{h2}	$\dots$	Y_{hk}	n_h
m_1	m_2	$\dots$	m_k	n

Under H_0 the revised statistic is

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i \hat{p_j})^2}{n_i \hat{p_j}} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i m_j/n)^2}{n_i m_j/n}$$

which may be more familiar in the form

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{k} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed number of elements in row i and column j of the table and E_{ij} is the expected number of observations in this cell, based upon knowledge of the row sums and column sums alone, and this expected value is just the product of the row sum and the column sum divided by n.

Q has approximate distribution χ^2 with $h(k-1)-(k-1)=(h-1)(k-1)$ degrees of freedom. The number of degrees of freedom are reduced by $k-1$ as we have had to estimate $k-1$ $\hat{p_j}$s (or the number of "free" entries in the contingency tables is (h-1)(k-1) as, if we know the row and column sums, the last row and the last column are determined by the first $h-1$ rows and $k-1$ columns).

We can thus carry out a test of H_0 in the same way as before.

in this we we can find if there is any relationship between the rows and the columns of our tables. Supposing that there is a relationship, how do we determine the nature of the relationship? One obvious way is to look at the contribution of each cell to the overall value of Q; where this number is large, the divergence of observation from expectation can be considered large. it may be that just a single cell stands out, indicating that the probability of occurence of class j from category i is out of line with the other classes. More informative would be if a single row had a number of larger entries, indicating that the probabilities for this category were generally different from the others, so perhaps all other categories are similar with the exception of this one.

8 Multiple and polynomial regression

8.1 Confidence intervals and tests for simple linear regression

We shall first return to linear regression and consider the varaince of our two parameter estimates a and b in order to use them to find confidence intervals. Since b is a linear function of independent normal random variables where

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

we have

$$E[b] = \frac{\sum_{i=1}^n (x_i - \bar{x})E[Y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta$$

and

$$Var[b] = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 Var[Y_i] =$$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Similarly a is also a linear function of the Y_is and we have

$$E[a] = E\left[\frac{1}{n}\sum_{i=1}^n Y_i - b\bar{x} \right] = \frac{1}{n}\sum_{i=1}^n E[Y_i] - b\bar{x} = \frac{1}{n}\sum_{i=1}^n (\alpha + \beta x_i) - b\bar{x} = \alpha$$

and it can be shown that

$$Var[a] = \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2$$

We can thus use the standard type of statistical methods that we have seen earlier in the course to obtain *confidence intervals* for the parameters and test the *hypothesis* that there is really no relationship between X and Y.

A confidence interval forβ, the true slope of the regression line, is given by its point estimate b plus or minus the standard error of b multiplied by the appropriate value from t-tables.

$$b - t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}, b + t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}$$

A test of whether the slope is zero (so X and Y would be independent) is directly related to this confidence interval; we reject the null hypothesis of zero slope if and only if 0 does not lie in the confidence interval for β.

We can a similar confidence interval for the mean of Y at a given value of x, using the fact that

$$Var(a+bx) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$$

which gives the confidence interval

$$a + bx - t_{\alpha/2}(n-2)\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}s, a + bx + t_{\alpha/2}(n-2)\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}s$$

Other confidence intervals and tests of appropriate measures can also be made in a similar manner.

8.2 Multiple regression

Suppose that we have a random variable Y which depends on several factors $X_1, X_2, \ldots, X_k$. We shall assume that there is a linear relationship, so that

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

where ϵ is $N(0, \sigma^2)$. If we have n data values of Y, together with the corresponding values of the Xs,
$y_j, x_{1j}, x_{2j}, \ldots, x_{kj} \quad j = 1, \ldots, n$
then we can use the method of least squares to minimise

$$S(\beta_1, \ldots, \beta_k) = \sum_{j=1}^{n}(y_j - \beta_1 x_{1j} - \ldots - \beta_k x_{kj})^2$$

We again differentiate with respect to each β_i in turn and set to zero, to give a set of k simultaneous equations.

$$\frac{\partial}{\partial \beta_i} S(\beta_1, \ldots, \beta_k) = \sum_{j=1}^{n} -2x_{ij}(y_j - \beta_1 x_{1j} - \ldots - \beta_k x_{kj}) \Rightarrow$$

$$\sum x_{ij} y_j - \beta_1 \sum x_{1j} x_{ij} - \ldots - \beta_k \sum x_{kj} x_{ij} = 0 \quad i = 1, 2, \ldots, k$$

These can be solved easily by computer for a large number of factors. Just as for grouped data we can assess the significance of the grouping using an analysis of variance table. We obtain the same kind of F statistic, for each variable in turn, based on a specific ordering which tells us whether the variable should be included given the inclusion of all variables higher up the chain. Thus changing the ordering and comparing results is sensible. The aim will be to find the variables that are important in predicting the value of Y, and those that make no difference once the others are included. This can be problematic if there is a strong linear relationship between some of the variables.

8.3 Polynomial regression

Suppose now that we have a relationship between Y and a single variable X.

$$Y = \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \epsilon$$

At first sight the use of linear regression seems inappropriate, as the relationship between Y and X is clearly non-linear. However the parameters β_i appear in the same linear form as is the case with multiple regression. in fact it is just a simple extension from the methods of multiple regression to obtain estimates of the parameters in the above model. We treat the different powers of X as if they were different variables X_i and perform a multiple regression, which is fine since these powers are not linearly related to each others (although there can be some stability problems because of the obvious high degree of dependence between the different powers of X).

Following the method of the previous section exactly for a series of data pairs $x_i, y_i \;\; i = 1, \ldots, n$, we obtain

$$S(\beta_1, \ldots, \beta_k) = \sum_{j=1}^{n} (y_j - \beta_1 x_j - \ldots - \beta_k x_j^k)^2$$

We differentiate with respect to each β_i in turn and set to zero yet again, to give a set of k simultaneous equations.

$$\frac{\partial}{\partial \beta_i} S(\beta_1, \ldots, \beta_k) = \sum_{j=1}^{n} -2x_j^i (y_j - \beta_1 x_j - \ldots - \beta_k x_j^k) \Rightarrow$$

$$\sum x_j^i y_j - \beta_1 \sum x_j^{i+1} - \ldots - \beta_k \sum x_j^{i+k} = 0 \qquad i = 1, 2, \ldots, k$$

We can again think about whether we include all the terms, although it is generally assumed that all terms up to a maximum power are included, so the ordering problem of multiple regression does not apply.

9 Principal component analysis

9.1 Introduction

Suppose that we have a large number of measurements
$x_{ij}; j = 1, \ldots, n$
taken on a range of subjects $i = 1, \ldots, m$. It may be very complicated to try to estimate how each of these affect certain other variables, and it may be unnecessary to consider them all because some are closely correlated with others. So if we had measured individuals height in centimetres and also in inches we could throw either of these away, as one is simply a constant multiplied by the other. We may thus be interested in reducing the data that we have to a number of important factors. We can do this in particular cases by removing those which do not have a significant effect, as in the case of multiple regression.

What if we want to summarise these data into as small a number of factors as possible for ease of use and interpretation? For example, how do we best combine two factors into a single one? Suppose that we have data values

$x_{11}, x_{12}, \ldots, x_{1n}$
$x_{21}, x_{22}, \ldots, x_{2n}$

It will not generally be best to just throw one of them away; rather we will choose

$C = \alpha_1 X_1 + \alpha_2 X_2$

for some α_1 and α_2 which captures as much of the information as possible. In fact it is only the ratio α_1/α_2 that determines where the line lies, so there is only one 'free' parameter. There are two parameters because we have not yet specified the scaling of our components, which will be made to satisify particular restrictions.

We find the correct ratio by fitting a line to the data in a similar (but not exactly the same) way as fitting a regression line of X_2 on X_1.

9.2 Correlation Matrices

To find our principal components we do not need to consider plots of the data (and these would be multi-dimensional for the cases we would really be interested in doing this for) but we can summarise our data into a single matrix of crucial information, which is the correlation matrix. Simply, we find the correlation of every pair of variables in turn, and the entry in row i and column j (r_{ij}) of the matrix is the correlation of variable i and variable j.

recall that the correlation coefficient is given by

$$r_{ij} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}}$$

Thus $r_{ij} = r_{ji}$ and $r_{ii} = 1$ for every i.

Our correlation matrix is actually a variance-covariance matrix, which has been appropriately scaled. In effect, X_i has been scaled by taking into consideration its variability, so that X_i has been replaced by $U_i = X_i/\sigma_{Xi}$ (divided by its standard deviation).

Our aim to is explain as much of the variability within the data as possible, using the fewest number of components. If we start with m components, then the total amount of variability from this scaled matrix is m. The variance associated with a component given by

$$C = \alpha_1 U_1 + \ldots + \alpha_m U_m$$

is

$$\sum_{i\neq j}\alpha_i\alpha_j Cov(U_i,U_j) + \sum_i \alpha_i^2 Var(U_i) = \sum_{i\neq j}\alpha_i\alpha_j c_{ij} + 1$$

since the scaling restriction that we set on the components, designed to make them all of 'unit' length, is

$$\sum_{i=1}^{m}\alpha_i^2 = 1$$

For instance if every variable is perfectly correlated $c_{ij} = 1$ for all i, this can be shown to be m, and we would only need to select a single component. We need to see how we select the principal components, and how many we need to take.

9.3 Eigenvalues and Eigenvectors

If C is an $m \times m$ matrix, as in the case of our correlation matrix above, then if we can find a vector v (which contains a single column and m rows) such that when we multiply the matrix C by the vector v we obtain some constant λ multiplied by v, then v is an *eigenvector* of the matrix, with associated *eigenvalue* λ.

For those not familiar with matrix multiplication, this is equivalent to

$$\sum_j r_{ij}v_j = \lambda v_i$$

for all values of i.

There is a general method to solve these equations, involving the determinants of matrices, which can be easily done by hand for small matrices and by computer for large. We will see the general solution for two factors only.

The solution involves setting the determinant of the following 2 × 2 matrix equal to zero.

$$\begin{vmatrix} r_{11}-\lambda & r_{12} \\ r_{21} & r_{22}-\lambda \end{vmatrix} \tag{1}$$

is identical to

$$\begin{vmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{vmatrix} \tag{2}$$

whose determinant is $(1-\lambda)^2 - r_{12}^2$. Setting this equal to zero gives

$$1-\lambda = \pm r_{12}$$

so $\lambda = 1 + r_{12}$ or $\lambda = 1 - r_{12}$. It turns out that it is the larger of these two possible values that we select. We then use one of the following equations to find the vector (either will give the same solution). If $r_{12} > 0$ then $\lambda = 1 + r_{12}$ and

$$r_{11}v_1 + r_{12}v_2 = \lambda v_1, r_{21}v_1 + r_{22}v_2 = \lambda v_2$$

$$\Rightarrow v_2 = v_1 \frac{\lambda - 1}{r_{12}} = v_1$$

9.4 Principal components

The principal components of a set of factors are a collection of independent combinations of the factors, each explaining less of the variability in the data than the previous factor. They are in fact identical to the collection of eigenvectors for our correlation matrix C, in order of the size of the eigenvalue. In total there will be m such eigenvectors and their eigenvalues will add up to m.

Thus in our example of two variables, the eigenvalues were $1 + r_{12}$ and $1 - r_{12}$ which add to 2 (if $r_{11} = 1$ then one component explains all of the variability) and if $r_{12} > 0$, then $v_1 = v_2$ so that the principal component is

$$\frac{1}{\sqrt{2}}U_1 + \frac{1}{\sqrt{2}}U_2$$

9.5 How many components to take?

So how many components should we take? if we take all m, we can explain as much as the original data set, but we have as many variable to work with, so we have not acheived anything. typically we wish to take a number noticably smaller than m. One common criterion is the Kaiser criterion, which includes components with eigenvalues greater than 1. Thus unless a component is as explanatory as one of the original variables then we do not include it. There are other methods, such as the scree test, which is a graphical method.

10 Maximum likelihood estimation

10.1 Introduction

We consider random variables for which the functional form of the p.d.f. is known, but that the precise distribution depends upon a single unknown parameter θ, say. We shall take a random sample
$X_1, X_2, \ldots, X_n$
from this distribution to provide information on what value our parameter may take.

The function of these values $u(X_1, \ldots, X_n)$ used to estimate θ is called the estimator, or point estimator, of θ. Ideally we want the actual value of this estimator, $u(x_1, \ldots, x_n)$ to be close to the true value of θ.

10.2 Maximum likelihood estimators

Suppose that X follows the Bernoulli distribution with parameter p, so that it takes value 1 with probability p, and otherwise it is 0. Thus the probability mass function of X is

$$f(x,p) = p^x(1-p)^{1-x} \qquad x = 0, 1$$

Now take a random sample of size n from this distribution. The probability that the values are $x_1, x_2, \ldots, x_n$ in that order is given by

$$P[X_1 = x_1, \ldots, X_n = x_n] = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

which is the joint probability mass function of $X_1, \ldots, X_n$. This is also referred to as the likelihood function, and considered as a function of the variable p.

$$L(p) = L(p; x_1, \ldots, x_n) = \prod f(x_i, p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

This gives the probability of any ordering of the $x_i s$ for a given value of p. If we want to estimate p from this data, a plausible way to proceed is to find the p that makes this likelihood the biggest.

To find the maximum we differentiate with respect to p and set the derivative equal to zero.

$$\frac{dL(p)}{dp} = \sum x_i p^{\sum x_i - 1}(1-p)^{n-\sum x_i} - (n - \sum x_i)p^{\sum x_i}(1-p)^{n-\sum x_i - 1} =$$

$$\left(\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}\right) p^{\sum x_i}(1-p)^{n-\sum x_i} = 0 \Rightarrow$$

$$\sum x_i - np = 0 \Rightarrow p = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

The corresponding statistic $u(X_1, \ldots, X_n) = \bar{X}$ is the maximum likelihood estimator of p.

Since maximising a positive function is the same as maximising its logarithm, this will yield the same estimator. It turns out that this is often a more convenient approach. In the above example

$$ln(L(p)) = \sum x_i ln(p) + (n - \sum x_i) ln(1-p) \Rightarrow$$

$$\frac{dln(L(p))}{dp} = \sum x_i \frac{1}{p} - (n - \sum x_i)\frac{1}{1-p} = 0$$

yielding the same result.

Note that the expectation of $u(X_1, \ldots, X_n) = \bar{X}$ is

$$E[\bar{X}] = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{np}{n} = p$$

so that the expectation of the estimator is the true value of the parameter being estimated. If an estimator has this property it is said to be *unbiased* (and if it does not, it is *biased*).

In general we can find the likelihood for distributions with many parameters, and maximum likelihood estimators for all of these parameters.

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.m.f. or p.d.f. $f(x; \theta_1, \ldots, \theta_m)$ then the likelihood of the vector $(\theta_1, \ldots, \theta_m)$ is given by

$$L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^n f(x_i, \theta_1, \ldots, \theta_m)$$

To find the maximum likelihood estimators, we differentiate with respect to each parameter in turn, set equal to zero, and then solve the m simultaneous equations in m unknowns.

Suppose that $X_1, \ldots, X_n$ is a random sample from a Normal (θ_1, θ_2) distribution, where $\Omega = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$

Thus in the usual notation, $\theta_1 = \mu, \theta_2 = \sigma^2$. The likelihood function is

$$L(\theta_1, \theta_2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta_2}} exp\left(-\frac{(x_i - \theta_1)^2}{2\theta_2}\right) =$$

$$\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n exp\left(-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}\right)$$

Taking the logarithm gives

$$ln(L(\theta_1,\theta_2)) = -\frac{n}{2}ln(\pi\theta_2) - \sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}$$

Differentiating with respect to θ_1 and θ_2 in turn gives

$$\frac{\partial}{\partial\theta_1}(ln(L(\theta_1,\theta_2))) = \sum_{i=1}^{n}\frac{(x_i-\theta_1)}{\theta_2}$$

and

$$\frac{\partial}{\partial\theta_2}(ln(L(\theta_1,\theta_2))) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2$$

Setting these equations equal to zero gives $\theta_1 = \bar{x}$ from the first, and substituting for θ_1 in the second gives

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

It is possible to show that this constitutes a maximum of the likelihood function. Thus the maximum likelihood estimators are $\hat{\theta_1}(=\hat{\mu}) = \bar{X}$, and

$$\hat{\theta_2}(=\hat{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = V$$

10.3 Method of moments estimators

An alternative way to estimate the parameters of a distribution is to equate the sample moments with the theoretical moments (which will be functions of the relevant parameters) and rearrange the equations obtained. For example, the sample mean is $\bar{X}$ whatever the underlying distribution of the process. If the distribution is exponential with unknown parameter λ, the theoretical mean of the distribution is $1/\lambda$, so we equate these to get

$$\bar{X} = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{\bar{X}}$$

giving the *method of moment estimator* $\hat{\lambda} = 1/\bar{x}$. We only needed to consider the first moment, since there was only one parameter to estimate.

If the distribution was gamma (α, β) then the theroetical mean and variance are

$$\alpha\beta, \alpha\beta^2$$

Equating these with the sample mean and variance gives

$$\alpha\beta = \bar{X}, \alpha\beta^2 = S^2$$

which can be rearranged to give

$$\alpha = \frac{\bar{X}^2}{S^2}, \beta = \frac{S^2}{\bar{X}}$$

giving estimates

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2}, \hat{\beta} = \frac{s^2}{\bar{x}}$$

10.4 The Likelihood ratio test

In this section we will examine a test which is based upon the likelihood function. It is assumed that the functional form of the p.d.f. is known, but depends upon one or more unknown parameters, and we set the p.d.f. of X is $f(x; \theta)$. Letting Ω be the entire parameter space (the possible values that the parameter can take), we test the null hypothesis

$$H_0 : \theta \in \omega$$

against the alternative hypothesis

$$H_1 : \theta \notin \omega$$

where ω is a subset of Ω. The likelihood ratio is

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$

where $L(\hat{\omega})$ is the maximum of the likelihood function with respect to θ in the set ω and $L(\hat{\Omega})$ is that maximum in the larger set Ω.

It is easy to see that $0 < \lambda \leq 1$ since the numerator of λ is maximised over a subset of the set that the denominator is, and both are positive. The null hypothesis is more plausible, the larger the value of λ, and the critical region is given by

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \leq k$$

for a suitably chosen k (probably to achieve some required significance level α).

The value of k obviously depends upon the distribution of λ. For large samples we can use an approximation which makes this test a lot easier to work with. This approximation concerns the distribution of $-2log(\lambda)$, which will be a positive term since $\lambda < 1$. If r_1 is the dimension of the space Ω and r_2 is the dimension of the subspace ω, then $-2log(\lambda)$ has an approximate $\chi^2(r_1 - r_2)$ distribution. Using this idea we can find an approximate value for k. We reject H_0 if $-2log\lambda$ is too large, namely larger than $\chi^2_\alpha(r_1 - r_2)$ under our approximation.

$$-2log\lambda \geq \chi^2_\alpha(r_1 - r_2) \Rightarrow$$

$$log\lambda \leq -\frac{1}{2}\chi^2_\alpha(r_1 - r_2) \Rightarrow$$

$$\lambda \leq exp\left(-\frac{1}{2}\chi^2_\alpha(r_1 - r_2)\right) = k$$

11 An introduction to Stochastic Processes

11.1 The Poisson process

We model a situation where events occur spontaneously, at random, in time. Examples are the emissions of α-particles in a radioactive experiment, arrivals of customers at a post office, the passing of cars on a quiet road. The important feature of these events is that they are unpredictable (they occur 'at random').

The average rate λ at which events occur is constant over time (not true for post office customers over a whole day, but good enough for a half hour period).
The occurence of events after time t is independent of what happened up to time t.
We also assume that events can only occur singly (never ≥ 2 simultaneously).

$X(t)$, the number of occurrences of a Poisson process at time t, has a Poisson distribution with parameter λt.

e.g. Fax messages arrive at an office at the mean rate of three per hour according to a Poisson process.
(i) What is the probability that exactly two messages are received between 9.00 and 9.40 ?
(ii) What is the probability that no messages arrive between 10.00 and 10.30 ?
(iii) What is the probability that not more than three messages are received between 10.00 and 12.00 ?

Let an hour be the unit of time that we work with. Thus we have a Poisson process of rate $\lambda = 3$. The number of messages in time t is thus Poisson $(3t)$.
(i) 9.00 to 9.40 - the number of calls is Poisson $(3 \times 2/3 = 2)$, so

$$P(2messages) = e^{-2}\frac{2^2}{2!} = 0.271$$

(ii) 10.00 to 10.30 - the number of calls is Poisson $(3 \times 0.5 = 1.5)$, so
P(0 messages) $= e^{-1.5} = 0.223$

(iii) 10.00 to 12.00 - the number of calls is Poisson $(3 \times 2 = 6)$.
$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) =$

$$e^{-6}\frac{6^0}{0!} + e^{-6}\frac{6^1}{1!} + e^{-6}\frac{6^2}{2!} + e^{-6}\frac{6^3}{3!} = 0.151$$

If T is the time until the next event from a given starting point (since the process is 'memoryless' it does not matter when this starting point is), then T has an exponential distribution with parameter λ.

e.g. In the previous example
(iv) What is the probability that the first message after 10.00 occurs before 11.00 ?

This is a Poisson process with rate 3, $t = 1$, and so $P(T \leq 1) = 1 - e^{-3\times 1} = 0.9502$

11.2 Birth processes

11.3 The simple birth process

Consider a population where each individual alive in the population generates further offspring according to a Poisson process at rate β (new individuals are produced asexually). We assume that the initial population size is x_0 and that there are no deaths, so that the population increases with time. It can be shown that the distribution of $X(t)$, the number of individuals alive at time t, follows a negative binomial distribution with parameters $e^{-\beta t}$ and x_0.

Note that if $x_0 = 1$, this becomes the geometric distribution with parameter $e^{-\beta t}$.

e.g. A population starts at time 0 with a single individual. Let the birth rate be two per week.
(i) What is the probability that after three weeks there are exactly two individuals?
(ii) What is the probability that after one week there are between two and four individuals (inclusive)?

(i) $x_0 = 1$, $\beta = 2$ per week, and $t = 3$ i.e.

$$p_2(3) = \binom{1}{0} e^{-2\times 3}(1 - e^{2\times 3})^1 = e^{-6}(1 - e^{-6}) = 0.00247$$

(ii) $x_0 = 1, \beta = 2, t = 1. P[2 \leq X \leq 4] = P(X = 2) + P(X = 3) + P(X = 4) =$
$e^{-2}(1 - e^{-2}) + e^{-2}(1 - e^{-2})^2 + e^{-2}(1 - e^{-2})^3 = 0.3057$

11.4 The pure death process

We consider a population in which there are no births, just deaths. Observations start with x_0 individuals alive at time 0 - these individuals die independently of each other, and eventually the population dies out completely.

This model is approached best by considering every individual separately. The probability that an individual is alive at time t, $P_a(t) = Ae^{-\nu t} = e^{-\nu t}$. We can use the binomial theorem to deduce that the probability that j individuals are still alive at time t, is given by

$$p_j(t) = (e^{-\nu t})^j (1 - e^{-\nu t})^{x_0 - j} \binom{x_0}{j}$$

In particular the probability that the population is extinct by time t is

$$p_0(t) = (1 - e^{-\nu t})^{x_0}$$

e.g. A population starts at time 0 with 4 individuals. The population follows a pure death process at a rate of 1 every 2 days.
(i) Find the probability that there is exactly one individual alive after a week.
(ii) Find the probability that the population has died out after a week.
(iii) Find the probability that the population has died out after two weeks, given that the total number of survivors after one week was 2.

(i) $t = 7, \nu = 0.5$ i.e.

$$p_1(7) = e^{-3.5}(1 - e^{-3.5})^3 \binom{4}{1} = 0.1102$$

(ii) $t = 7, \nu = 0.5 \Rightarrow$

$$p_0(7) = (1 - e^{-3.5})^4 = 0.8846$$

(iii) The process is memoryless, so that
P(0 after 2 weeks/ 2 after 1 week) = P(0 after 1 week/ 2 after 0 weeks) =

$$(1 - e^{-3.5})^2 = 0.9405$$

11.5 Birth and death processes

We shall now consider a population model, similar to those of the last chapter, but with both births and deaths. In the Poisson process and the simple birth process the only change possible is an increase in the population. In the pure death process the population could only be reduced. Here both may occur.

11.6 The simple birth-death process

In the previous section we considered the simple birth process and the pure death process. Now we shall combine the two.

In the simple birth process, each individual gives birth at rate β, so that when the population is of size x, the birth rate is βx. In the pure death process individuals die at rate ν, so that the death rate is νx. We could look for an expression for $X(t)$, the number of individuals alive at time t. This turns out to be quite complicated to do; we can answer some simpler questions more easily.

For instance, what is the probability that the population eventually becomes extinct? If $\beta \leq \nu$, the population is certain to become extinct. If $\beta > \nu$, the probability of eventual extinction is

$$\left(\frac{\nu}{\beta}\right)^{x_0}$$

If we are not interested in the time that a particular event occurs, but only in its type (is it a birth or a death), we can simplify the model. Relabelling the time of the occurrence of the ith event as i, we obtain a new random process $\{X_i; i = 0, 1, 2 \ldots\}$ in discrete time. This process is said to be *embedded* in the original process, i.e. is an *embedded process.* X_i is the size of the embedded process at time i, and of the population immediately after the ith change.

$$p = \frac{\beta}{\beta + \nu}$$

where p is the probability that any particular event is a birth. We have what is called a simple random walk with

$$P[X_i = x + 1/X_{i-1} = x] = \frac{\beta}{\beta + \nu}$$

and an absorbing barrier at zero. This means that we can use standard results to show;

the expected number of events to extinction is

$$\frac{x_0(\nu + \beta)}{\nu - \beta} \qquad \beta < \nu$$

and is infinite if $\beta \geq \nu$.

The probability that the size of our population reaches m individuals at some point (before possibly becoming extinct), has probability

$$\frac{1 - \left(\frac{\nu}{\beta}\right)^{x_0}}{1 - \left(\frac{\nu}{\beta}\right)^{m}} \qquad \beta \neq \nu$$

$$\frac{x_0}{m} \qquad \nu = \beta$$

e.g. If a simple birth-death process starts with 5 individuals, what is the probability that it reaches 10 given that it becomes extinct if
a) $\beta = 4, \nu = 6$?
b) $\beta = 9, \nu = 6$?

a) P(reaches 10) =

$$\frac{1-\left(\frac{6}{4}\right)^5}{1-\left(\frac{6}{4}\right)^{10}} = 0.1164$$

The process is certain to become extinct, so that P(reaches 10 and becomes extinct) = P(reaches 10) = 0.1164

b) P(reaches 10) =

$$\frac{1-\left(\frac{6}{9}\right)^5}{1-\left(\frac{6}{9}\right)^{10}} = 0.8836$$

P(reaches 10 / becomes extinct)=

P(extinction/reaches 10)P(reaches 10)/P(extinction)

P(extinction)=

$$\left(\frac{6}{9}\right)^5 = 0.1317$$

P(extinction/reaches 10)=

$$\left(\frac{6}{9}\right)^{10} = 0.0173$$

Thus P(reaches 10 /becomes extinct)=

$$\frac{0.0173 \times 0.8836}{0.1317} = 0.116$$

So if we know that the population becomes extinct, it is unlikely to have reached 10 before doing so.

12 Markov Chains

in this section we shall discuss Markov chains with a finite number of states. this will be central to a lot of what follows in later sections. A Markov chain is defined by a number of states
$E_1, \ldots, E_n$
, one of which is occupied at a given time. We follow the process from time
$t = 1, 2, 3, \ldots$
As time moves a step forward, the process moves from one state to another (possibly staying at the same state) following some straightforward probabilistic rules.

The process is memoryless and homogeneous in time. Thus the probability of moving from E_i at time t to E_j at time $t+1$ takes the same value for all t. So only the current state matters in determining future movements; the earlier history of the process is irrelevant.

12.1 The transition matrix

If we are in state E_i then the probability that the process moves to state j in the next time step is labelled p_{ij} for each pair of states i and j. For convenience we group all of these transition probabilities together into a single matrix, labelled P.

$$\begin{array}{|c|ccccc|}
 & E_1 & E_2 & E_3 & \ldots & E_n \\
\hline
E_1 & p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\
E_2 & p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
E_n & p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn}
\end{array} \tag{3}$$

the entry in row i and column j os p_{ij}, the probability of moving form i to j. Since wherever the process stands at time t, it must end up at precisely one state at time $t+1$, the probabilities in every row must sum to 1, i.e.

$$\sum_{j=1}^{n} p_{ij} = 1$$

for every value of i.

It is easy to see that the random walk from the previous section (where at each step except 0 there was a probability p of increasing by 1 and $q = 1 - p$ of decreasing by 1) has transition matrix

$$\begin{array}{|c|cccccccc|} & 0 & 1 & 2 & \dots & k-1 & k & k+1 & \dots \\ \hline 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ 1 & q & 0 & p & \dots & 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ k & 0 & 0 & 0 & \dots & q & 0 & p & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{array} \tag{4}$$

12.2 Absorbing states

The random walk example that we discussed above is a Markov chain with an *absorbing state*, since the state '0' cannot be left when it is reached. Such a Markov chain is easily recognised; if its transition matrix has a '1' on its leading diagonal, then the corresponding state is an absorbing one; if there are no 1s on the leading diagonal, then it does not contain an absorbing state.

For a Markov chain with at least one absorbing state, sooner or later one of its absorbing states will be entered and never left. We may wish to consider what is the probability that we enter a given state, or the expected time to entering a state.

For example, the probability that the size of the population from our birth and death process from Section 11 reaches m individuals at some point (before possibly becoming extinct), was given to be

$$\frac{1-\left(\frac{\nu}{\beta}\right)^{x_0}}{1-\left(\frac{\nu}{\beta}\right)^{m}} \qquad \beta \neq \nu$$

This was found by using the random walk transition matrix from above, with the single change that m becomes an absorbing state, so that row m is all 0s except $p_{mm} = 1$.

If a Markov chain has no absorbing states, the questions that we can ask about it are more varied. To simplify matters, and because these are the key properties of the processes used in Monte Carlo Markov Chains from later sections, we concentrate on Markov chains which are *finite*, *aperiodic* and *irreducible.* A chain is finite if it has a finite number of possible states (as opposed to the birth and death process). A chain is aperiodic if there is no state that can be returned to at regular intervals only (for instance even time points only). For instance

$$\left|\begin{array}{cccc} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{array}\right| \tag{5}$$

is periodic. If a Markov chain has no 0s down its leading diagonal, then it is aperiodic. A chain is irreducible if whatever state we are in, we can reach some other state from that state from some path. For instance

$$\left|\begin{array}{cccc} 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.6 & 0.4 \end{array}\right| \tag{6}$$

is reducible, since once states 3 or 4 have been reached, the process can never return to states 1 or 2.

12.3 Stationary distributions

Suppose that a Markov chain has a transition matrix P which has probability ϕ_j of being in state E_j at time t, $j = 1, \ldots, n$. Suppose that

$$\phi_j = \sum_{i=1}^{n} \phi_i p_{ij} \qquad j = 1, 2, \ldots, n$$

so that the probability of being in state j at time $t+1$ is still ϕ_j for all j. The probability distribution $(\phi_1, \phi_2, \ldots, \phi_n)$ is *stationary.*

To find a stationary distribution, we must solve the set of equations above. For example, for the matrix

$$\left|\begin{array}{cccc} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{array}\right| \tag{7}$$

we obtain
$\phi_1 = 0.5\phi_3 + 0.2\phi_4 \ldots (i)$
$\phi_2 = 0.5\phi_3 + 0.8\phi_4 \ldots (ii)$
$\phi_3 = 0.6\phi_1 + 0.3\phi_2 \ldots (iii)$
$\phi_4 = 0.4\phi_1 + 0.7\phi_2 \ldots (iv)$

The first two equations yield
$0.6\phi_4 = \phi_2 - \phi_1 \ldots (v)((ii) - (i))$
$1.5\phi_3 = 4\phi_1 - \phi_2 \ldots (vi)(4(i) - (ii)$
$1.24\phi_1 - 0.58\phi_2 = 0 \Rightarrow \phi_2 = 2.1379\phi_1(0.6(iv) - (v)$
Thus from (v) $\phi_4 = 1.8965\phi_1$
and from (vi) $\phi_3 = 1.2414\phi_1$
So $\phi_1(1 + 2.1379 + 1.2414 + 1.8965) = 1 \Rightarrow \phi_1 = 0.1593$ giving
$\phi_2 = 0.3407, \phi_3 = 0.1978, \phi_4 = 0.3022$

Notice that we could have solved this more easily by observing that since the process alternates between states 1,2 and 3,4 we must have $\phi_1 + \phi_2 = \phi_3 + \phi_4 = 0.5$

The stationary distribution of

$$\begin{vmatrix} 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.6 & 0.4 \end{vmatrix} \tag{8}$$

can be found easily since it is clear that $\phi_1 = \phi_2 = 0$. Sovling for the other two strategies, gives $\phi_3 = 6/11, \phi_4 = 5/11$

12.4 Markov chains in continuous time

A Markov chain in continuous time can be defined in terms of the transition rates between the states, so that the rate from state E_i to E_j is given by q_{ij}. In fact each of the models from Section 11 are continuous time Markov chains. if we are not worried about the specific times of events, but just the sequence, we can construct an embedded process in precisely the same way as we did in that section, which is then a Markov process in discrete time, with a transition matrix given by

$$p_{jk} = \frac{q_{jk}}{\sum_{i=1, i \neq j}^{n} q_{ji}}$$

Exercise: Show this.

Stationary distributions are found in the same way as before. In fact, the equations become

$$\phi_j = \sum_{i=1}^{n} \phi_i q_{ij} \qquad j = 1, 2, \ldots, n$$

So for the birth and death process of Section 11, for $j > 0$, $q_{j,j+1} = \beta, q_{j,j-1} = \nu$ and $q_{ji} = 0$ otherwise, and we obtain

$$p_{j,j+1} = \frac{q_{j,j+1}}{\sum_{i=1}^{n} q_{ji}} = \frac{\beta}{\beta + \nu}$$

$$p_{j,j-1} = \frac{q_{j,j-1}}{\sum_{i=1}^{n} q_{ji}} = \frac{\nu}{\beta + \nu}$$

and $p_{ji} = 0$ otherwise

13 DNA sequence analysis

13.1 Analysing a single sequence

Before we can analyse a sequence, we need to determine what the sequence is. It turns out that it is not possible to accurately identify long DNA sequences. Rather, many overlapping small sequences are taken (of the order of 500 bases). Then these pieces need to be assembled into a single long sequence by matching overlapping regions on to each other; such an assemblage is called a *contig.* To do this a sufficiently large collection of such short subsequences need to be collected. This is called *shotgun sequencing.* A sequence has nX *coverage* if for the original sequence of length G, the total length of all the fragments is nG. To expect to cover 99% of the sequence, it can be shown that we need to have $4.6X$ coverage. This is the technique that has been used to sequence the whole human genome.

FIGURE 13.1

Suppose that the DNA sequence that we are mapping has length G, and we have N fragments each of length L, then the coverage is given by

$$a = \frac{NL}{G}$$

The mean proportion of the sequence covered by one or more fragments is the probability that a random point has its left hand end in the interval of length L immediately to the left of this point.

FIGURE 13.2

On the (reasonable) assumption that L is a lot smaller than G, the number of fragments whose left-hand edge falls in this region is Binomial with parameters N and L/G. The probability that no fragment lies in this interval is thus

$$\left(1 - \frac{L}{G}\right)^N \approx e^{-a}$$

The probability that the point is covered is thus $1 - e^{-a}$. When $a = 4.6$ this leads to a probability of 0.99, as above.

13.2 Modelling of DNA sequences

The structure of DNA can be thought of as long sequences of nulcletides of four types labelled a, g, c and t. We are interested in the structure of these sequences, based upon the relationship between the type of nucleotide which appears at neighbouring sites. The sequences comprise different regions, where the structure of the components can vary from region to region. The simplest 'structure' conceivable is complete independence, where the nucleotide at every site is independent of all others, and identically distributed with them. A more plausible model is based upon a Markov structure; namely that neghbouring nucleotides are correlated, but that only immediate neighbours influence the distribution at a site. The following table shows the kind of data that we are interested in for the Markov model.

	a	g	c	t	Total
a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
g	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$Y_{2.}$
c	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$Y_{3.}$
t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	$Y_{4.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	Y

Such data can be used to estimate the transition matrix which defines our Markov model. In particular we can test whether the transition probabilities are identical for each nucleotide using a χ^2 test as used for standard categorical data.

$$Q = \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{(Y_{ij} - Y_{i.}Y_{.j}/Y)^2}{Y_{i.}Y_{.j}/Y}$$

will have an approximate χ^2 distribution with $(4-1)(4-1) = 9$ degrees of freedom under the null hypothesis of identical transition probabilities.

We may be interested in the possibility of long repeats of a particular nucleotide, say a. Suppose that for a particular DNA sequence, the probability of a appearing at any given site is p. After any given occurrence of a nucleotide other than a (a 'failure') there will be a run of Y a's ('successes'); this length can be zero if a second failure follows the first. Y has a geometric distribution with parameter p, i.e.

$$P(Y = y) = (1-p)p^y$$

It can be shown that if n such (independent) sequences occur, and Y_{max} is the largest of these, then

$$P(Y_{max} \geq y) = 1 - (1-p^y)^n$$

In a DNA sequence of length N, the number of such sequences of a's is approximately $n = N(1-p)$ (the number of failures) so that

$$P(Y_{max} \geq y) = 1 - (1-p^y)^{N(1-p)}$$

This gives the p-value of the test of the null hypothesis that every successive nucleotide is independent.

e.g. if $N = 100000, p = 0.25$ and the observed $y_{max} = 10$, then P is as follows:

$$P = 1 - (1 - 0.25^{10})^{75000} = 0.0690272$$

13.3 Comparing two DNA sequences

Suppose that we have two DNA sequences. One way of comparing the two is to test whether the proportion of each nucleotide is the same in each sequence. This follows a typical χ^2 test for tabular data as above, where the table is

	a	g	c	t	Total
S1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
S2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$Y_{2.}$
Total	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	Y

In this case there are $(2 - 1) \times (4 - 1) = 3$ degrees of freedom.

We are often interested in finding *alignments* between two sequences. For example, suppose that we have the sequences *cgggtatccaa* and *ccctaggtccca* which may be descended from a single common ancestral sequence (changes occur, by means of substitutions, deletions and insertions). We try to find the best match up between the two sequences, where gaps can be introduced into either sequence (as deletions and insertions alter the position of whole sub-sequences).

The following is an example of an alignment between the two sequences
c g g g t a − − t c c a a
c c c − t a g g t c c c a

How do we decide whether two sequences are well aligned. one way is to give an alignment score to any alignment between two sequences. For example we can use

number of matches-number of mismatches-number of gaps

It makes sense to penalise the use of gaps as this gives us a lot more options in our sequences, and we could get a perfect match if allowed to use an unlimited number of gaps for free. Scoring schemes can be more complicated, of course, but in some cases this one works well.

Once we have a scoring system, we can compare the possible alignments and choose the one with the highest score. If no gaps are allowed, then we can do this for every possible alignment for 2 sequences. The use of gaps makes this not feasible; we need to use an algorithmic method. One such algorithm is the Needleman-Wunsch algorithm. This is an example of a dynamic programming algorithm. We will come back to this and related algorithms at the end of the course.

For example, if we use the above scoring mechanism on the alignment
c t t a g − g − −
c a t − g a g a a
the score is $4 - 1 - 4 = -1$.

One related alignment problem is the *linear gap model.* Given two sequences, we try to find the subsequence of the longer one that can be aligned with the shorter one in the best way (where gaps are allowed).

Let $x = X_1X_2 \ldots X_m$ be the shorter sequence and $y = Y_1Y_2 \ldots Y_n$ be the longer one. Also denote the subsequence of y given by $Y_kY_{k+1} \ldots Y_j$ as $y_{k,j}$. To get a value of $B(x, y_{k,j})$, the score of the best alignment that we can find between x and $y_{k,j}$, the running time of the Needleman-Wunsch algorithm is of the order of $m(j - k)$.

We want to find the best alignment overall, given by

$$max(B(x, y_{k,j}) : 1 \leq k \leq j \leq n)$$

If we did this exhaustively, it would take time of the order of mn^3. It is possible to use another algorithmic approach to improve this considerably, to the order of mn.

13.4 Protein sequences and substitution matrices

When looking at DNA sequences, such simple scoring systems are often effective. However for protein sequences, some substitutions are much more likely than others; an alignment algorithm that takes this into account is a lot more effective. There are two main type of substitution matrix, PAM (Accepted Point Mutation) and BLOSUM (BLOcks SUbstitution Matrices) with rather contrived acronyms, as you can see. We will consider PAM substitution matrices only.

An accepted point mutation is the substitution of one amino acid by another that is ‘accepted’ by evolution, so that effectively this change happens for the whole of a given species. A PAM1 transition matrix is the Markov transition matrix applying to the time period in which we expect 1% of amino acids to undergo accepted point mutations within the given species.

We can find transition matrices for larger distances by raising the matirx to a given power; transitions for n PAM units is given by PAMn,
$M_n = M_1^n$

Suppose that all amino acids are equally frequent (so that $p_j = 0.05$), that all are equally likely to be substituted by some other amino acid in any given time, and all substitutions are equally likely. Then the PAM1 matrix $M_1 = (m_{ij})$ is given by,

$$m_{ii} = 0.99, m_{ij} = \frac{0.01}{19} \quad i \neq j$$

For this matrix it can be shown that these probabilities in the PAMn matrix become

$$m_{ii}^{(n)} = 0.05 + 0.95(94/95)^n$$

$$m_{ij}^{(n)} = 0.05 - 0.05(94/95)^n$$

For $n = 10$ for example, we obtain the probability that any given amino acid is unchanged as

$$m_{ii}^{(10)} = 0.05 + 0.95(94/95)^{10} = 0.9046$$

and the probability that it is replaced by any given alternative amino acid is

$$m_{ij}^{(10)} = 0.05 - 0.05(94/95)^{10} = 0.00502$$

Of course for real populations these probabilities are not so evenly spread, as some amino acids are more common than others, and transition rates vary between sites etc. Data can be used to estimate the values in the PAM1 matrix.

Question: If we find amino acid, A, at a point in the sequence, what is the probability that 2 PAM units ago it was a B?

14 Monte Carlo Markov Chains I

In this section we shall introduce a variant on the Markov chain model and consider some computational statistical techniques which will lead us to the mainstream Monte Carlo Markov chain methods of the next section. The methods of this section, both theoretical and practical, have a wider applicability as well.

14.1 Hidden Markov models

A Hidden Markov model is an extension of the Markov chain idea that we met in the previous sections. Suppose that we have a Markov chain where states are visited according to a transition matrix P, so that a sequence of states are visited which we label $q_1, q_2, q_3 \ldots$. At each state a symbol is emitted from some collection of possibilities, so that we have a sequence of symbols $O_1, O_2, O_3, \ldots$. We can label the sequence of q_is by $\mathbf{Q}$ and the sequence of O_is by $\mathbf{O}$. Often we know the sequence $\mathbf{O}$ but we do not know the underlying sequence of states $\mathbf{Q}$; the sequence $\mathbf{Q}$ is *hidden*.

We may be able to estimate what the underlying sequence $\mathbf{Q}$ is from the information that we have in $\mathbf{O}$, if each state does not emit the symbols with the same probabilities as all of the other states. To do this we thus need two pieces of information; the transition matrix P and a collection of probability distributions

Example 14.1. Consider the Markov chain with transition matrix

$$\begin{vmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{vmatrix} \tag{9}$$

where the process is equally likely to start in either state. There are only two possible symbols that can be emitted 1 and 2; state S_1 emits 1 with probability 0.5 (and so 2 with probability 0.5), state S_2 emit 1 with probability 0.25 (and so 2 with probability 0.75).

Suppose that we observe the sequence 2,2,2. What is the most plausible underlying sequence? There are 8 possibilities, each of which is written below with the probability of obtaining the observed sequence AND underlying sequence calculated (i.e. $P(\mathbf{Q} \bigcap \mathbf{O})$).

$$S_1 \to S_1 \to S_1 \qquad 0.5 \times 0.5 \times 0.9 \times 0.5 \times 0.9 \times 0.5 = 0.0506$$

$$S_1 \to S_1 \to S_2 \qquad 0.5 \times 0.5 \times 0.9 \times 0.5 \times 0.1 \times 0.75 = 0.0084$$

$$S_1 \to S_2 \to S_1 \qquad 0.5 \times 0.5 \times 0.1 \times 0.75 \times 0.8 \times 0.5 = 0.0075$$

$$S_1 \to S_2 \to S_2 \qquad 0.5 \times 0.5 \times 0.1 \times 0.75 \times 0.2 \times 0.75 = 0.0028$$

$$\begin{array}{ll}
S_2 \to S_1 \to S_1 & 0.5 \times 0.75 \times 0.8 \times 0.5 \times 0.9 \times 0.5 = 0.0675 \\
S_2 \to S_1 \to S_2 & 0.5 \times 0.75 \times 0.8 \times 0.5 \times 0.1 \times 0.75 = 0.0113 \\
S_2 \to S_2 \to S_1 & 0.5 \times 0.75 \times 0.2 \times 0.75 \times 0.8 \times 0.5 = 0.0225 \\
S_2 \to S_2 \to S_2 & 0.5 \times 0.75 \times 0.2 \times 0.75 \times 0.2 \times 0.75 = 0.0084
\end{array}$$

The best sequence maximises

$$P(\mathbf{Q}|\mathbf{O}) = \frac{P(\mathbf{Q} \bigcap \mathbf{O})}{P(\mathbf{O})}$$

which is equivalent to maximising the above. Thus the most plausible sequence is S_2, S_1, S_1.

An HMM consists of the following five components

(1) A set of n states $S_1, \ldots, S_n$
(2) An alphabet of distinct observation symbols $A = \{a_1, \ldots, a_M\}$
(3) The transition probability matrix $P = (p_{ij})$ where

$$p_{ij} = P(q_{t+1} = S_j | q_t = S_i)$$

(4) Emission probabilities for each state: if the process is in state S_i then
$b_i(a_j) = P(S_i \text{ emits } a_j)$, so $\sum_{j=1}^{M} b_i(a_j) = 1$
(5) An initial distribution of states $\pi_i = P(q_1 = S_i)$

Components 1,2 and 5 define the underlying Markov chain.

Thus in our example $n = 2$, $M = 2$ with $a_1 = \text{‘}1', a_2 = \text{‘}2'$, P is the 2×2 matrix (e.g. $p_{12} = 0.1$), $b_1(a_1) = 0.5, b_1(a_2) = 0.5, b_2(a_1) = 0.25, b_2(a_2) = 0.75$.
In this simple example it was possible to find our solution by hand,. However real problems often have long sequences with many states, so such a complete calculation cannot be done even by computer.

14.2 Hidden Markov models and multiple sequence alignments

The following figure illustrates how a Hidden Markov Model can be used to model a protein family.

FIGURE 14.1

The example has length 5, although any length is possible. States are labelled m_j, i_j and d_j. m, i and d stand for *match, insert* and *delete.* We start in state m_0 and move from left to right ending on m_5, through some path denoted by the arrows.

Emissions are made from all states on the path through our diagram except the first (m_0) and last (m_5, here). Our alphabet consists of the twenty amino acids, plus a dummy symbol δ representing delete. Each match and insert state has its own distribution over the 20 amino acids and cannot emit δ, a delete state must emit δ.

It could be that all emissions in m and i are uniformly likely, which would mean that random sequences would result, or ms could have one distribution, and is another. It is possible to find the most likely path through the model by using an algorithm, such as the *Viterbi* algorithm.

For example, consider the sequences CAEFDDH and CDAEFPDDH and suppose that their most likely paths are
$m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$ and
$m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$ respectively.

The alignment induced by this model is found by aligning positions generated by the same state, as follows:

FIGURE 14.2

This leads to the induced alignment
C – A E F – D D H
C D A E F P D D H

Consider an extension of this example, with more than two sequences. Suppose that we have sequences CAEFTPAVH, CKETTPADH, CAETPDDH, CAEFDDH, CDAEFPDDH and the corresponding paths are

$m_0m_1m_2m_3m_4m_5m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3m_4m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3d_4m_5m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$
$m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$

The induced alignment is
C – A E F T P A V H
C – K E T T P A D H
C – A E – T P D D H
C – A E F – – D D H
C D A E F – P D D H

14.3 Bootstrapping methods

With the advent of fast powerful computers, it has become possible to use alternative methods to test statistical hypotheses and find confidence intervals.

We use the *empirical probability distribution* of a set of data. If we have a set of data values $x_1, \ldots, x_n$ then the empirical probability distribution of our random variable X is

$$P_e(X = x) = \frac{m_x}{n}$$

where m_x is the number of times x appears in our data set.

Question: What is the expectation of the value of the proportion of the data less than or equal to a given x (this proportion is the empirical distribution function at x)?

Suppose that we want to find an estimate of the mean of the underlying distribution and a confidence interval for this mean. one way to do this is given in Section 2 of the course. This procedure assumed that the sample mean $\bar{X}$ has a normal distribution (at least approximately). If this is not the case, then the methodology used is not valid. The aim of the bootstrap method is to provide a way of finding estimates and confidence intervals which do not need this assumption.

Suppose that we start with the data set $x_1, \ldots, x_n$. The first step of the bootstrap involves sampling from the this collection n times, *with replacement*. Some observations from the original data may not appear at all, others will appear once, others twice etc.

Example 14.2 - We consider the following simple example of $n = 12$ data points (these have actually been randomly generated from an exponential distribution with parameter 0.5, and so mean 2)

0.843, 0.977, 0.003, 3.159, 1.027, 1.009, 3.331, 0.235, 0.476, 0.398, 4.597, 0.434

Each of the 12 observations in each bootstrap sample is selected from the above 12 numbers, each being chosen at random with probability 1/12. The number of times the value 0.843 appears in a sample, for example, follows a binomial distribution with parameters (12, 1/12). On average it appears once (occurring with probability 0.384), but it may not appear at all (probability 0.352) or more than once (the probabilities that it appears 2 and 3 times are 0.192 and 0.058 respectively).

From this sample we replace $x_1, \ldots, x_n$ by the bootstrap sample values, to give $x\bar{_{B1}}$, the mean of the first bootstrap sample. The procedure is then repeated a large number R times, leading to R bootstrap estimates

$$x\bar{_{B1}}, \ldots, x\bar{_{BR}}$$

These can be thought of as giving an empirical distribution of $\bar{X}$. The average of the bootstrap estimates is the *bootstrap estimate* of the mean.

Example 14.2 continued - we shall now take 20 bootstrap samples from the above distribution (recall that we sample with replacement), giving the mean value for each sample. These means are as follows.

Means - 1.510, 1.685, 0.952, 0.961, 1.230, 0.446, 2.211, 1.921, 1.295, 1.669, 0.906, 0.825, 1.495, 1.462, 1.326, 1.261, 1.157, 1.194, 1.262, 1.425

A 95% confidence interval for the mean of the population can be obtained from the above, and is (0.659, 2.218).
Note that this contains the true mean that we know to be 2, due to the method that the data was generated.

Typically bootstrapping is useful when there is good reason to think that the samples that we take may be significantly non-normal. Thus we can find confidence intervals for a wide range of parameters that do not behave as nicely as means. Thus in the above example we may be interested in the variance of the data. With a small sample, the distribution of the sample variance is quite different to the normal distribution. There is an alternative way to carry out tests for the variance that we have seen before; but how about the mean of the reciprocal of the data $1/x_i$ for example. Any function $\theta(X_1, \ldots, X_n)$ can be treated in the same manner.

It is possible to use the bootstrap procedure with more than one sample. For instance we may be interested in finding a confidence interval for the difference of two means, but not be confident that the underlying variances of the distributions are equal, or that the data is sufficiently close to being normally distributed. The easiest way to do this is to generate R pairs of samples from each data set independently using the method above. If we have n data points from sample 1 and m from sample 2, each bootstrap sample (pair) also contains n and m data points, the mean of each being taken and the difference of the means being the summary statistic from the data. These differences for each sample can then be considered together to provide an estimate and confidence interval for the difference of the means. Note that there are also more elaborate ways to do this, which are preferred by some researchers.

15 Monte Carlo Markov Chains II

In this section we consider two of the most important methods of Monte Carlo Markov Chains, together with an example of their use which relates to earlier problems in sequence analysis. We shall show in each case the mathematical detail of how the method works.

15.1 The Metropolis-Hastings Algorithm

The aim of the Metropolis-Hastings Algorithm is to construct a Markov Chain which is both aperiodic and irreducible, which has a given stationary distribution.

Let us suppose that we have a stationary distribution $\mathbf{v}$, which is of course a vector of probabilities defined on some number of states n. To construct our Markov Chain, firstly we choose some set of constants $q_{ij} \quad i, j = 1, \ldots, n$
such that $q_{ij} > 0$ for all i, j and $\sum_j q_{ij} = 1$ for all i. We shall now define a_{ij} by

$$a_{ij} = min\left(1, \frac{v_j q_{ji}}{v_i q_{ij}}\right)$$

We define p_{ij} by

$$p_{ij} = q_{ij} a_{ij} \qquad i \neq j$$

and

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

Since $q_{ij} > 0$ we can see that $p_{ij} > 0$ for all i, j pairs (including when $i = j$ although this is not as immediately obvious). Thus the Markov Chain defined by the transition probabilities is aperiodic and irreducible. It just remains to show that the stationary distribution of this Markov Chain is $\mathbf{v}$.

This is shown by demonstrating that $v_j p_{ji} = v_i p_{ij}$ for all i, j pairs. This is called the *detailed balance requirement*, and is a sufficient condition for $\mathbf{v}$ to be a stationary distribution of the process. Essentially, the rate of movement from state i to state j is balanced by the rate from state j to state i, and so the process is in (stochastic) equilibrium.

If we suppose that $v_j q_{ji} < v_i q_{ij}$, then from above we have

$$a_{ij} = \frac{v_j q_{ji}}{v_i q_{ij}}, p_{ij} = \frac{v_j q_{ji}}{v_i}, a_{ji} = 1, p_{ji} = q_{ji}$$

and it follows that $v_j p_{ji} = v_i p_{ij}$.
If the reverse inequality holds the reasoning is identical; if $v_j q_{ji} = v_i q_{ij}$ then it also follows easily.

This method can be extended to the case where some q_{ij} are zero (important for the following section) provided that whenever $q_{ij} > 0$, so is q_{ji}.

Why should we wish to do this? We may wish to take a random sample from this distribution, or to find the state with the largest probability from the distribution. But for any reasonably sized number of states, it is easier to compute these from the distribution directly rather than construct the Markov Chain and let it run through many iterations. The answer to this question is that sometimes the number of states is just so large, and the definition of the probabilities of being in the state given in a sufficiently indirect way, that this is impossible. The Markov chain idea can give us a shortcut round this problem, as we shall see in the next section.

15.2 Gibbs Sampling

Suppose that $Y_i, i = 1, \ldots, k$ are discrete finite random variables, and $\mathbf{Y}$ is the random column vector made up of the Y_is, i.e. $(Y_1, \ldots, Y_k)^T$. We can define the distribution of $\mathbf{Y}$, $P_{\mathbf{Y}}(\mathbf{y})$, the probability that $\mathbf{Y}$ takes the vale $\mathbf{y}$ (i.e. $Y_i = y_i$ for all i). We assume that $P_{\mathbf{Y}}(\mathbf{y}) > 0$ for all values of $\mathbf{y}$.

We will construct a Markov Chain whose states are the possible values of $\mathbf{Y}$. We shall order the vectors $\mathbf{y}$ in some manner, to give vectors numbered from 1 to n. We equate vector j with state j in our Markov chain.

We define the Markov chain as follows;
We consider each of the k components of the vector in turn, deciding on a new value for this component only in each step, which may be the same as the old one [alternatively we pick one of the k components at random and only perform a single step -see the next section] . We construct a transition matrix $P^{(1)}$ for the first step as follows. If vectors i and j differ in any component but our chosen one, then we set $p^{(1)}_{ij} = 0$.
If they differ by this first component only (or not at all) we define

$$p^{(1)}_{ij} = P(Y_1 = y_1^* | Y_2 = y_2, \ldots, Y_k = y_k) = \frac{P(Y_1 = y_1^*, Y_2 = y_2, \ldots, Y_k = y_k)}{P(Y_2 = y_2, \ldots, Y_k = y_k)}$$

where vector i is $(y_1, y_2, \ldots, y_k)$ and vector j is $(y_1^*, y_2, \ldots, y_k)$.

After we have changed the first component (or left it the same) we move on to the second component, performing the same operation above to obtain the transition matrix for this component $P^{(2)}$ and so on. A move in the full Gibbs process is a sequence of k moves, one for each component, and thus the full transition matrix for the Gibbs sampler is just the product of all of these transition matrices for each step

$$P = P^{(1)}P^{(2)} \dots P^{(k)}$$

Exercise: Explore the simplest case of this idea, with a two-dimensional vector, each position having two elements only. This is the situation in exercise E2.

This Markov chain is irreducible and aperiodic, since $p_{ii} > 0$ for all i and every state can be reached by a finite number of steps from any other (it is possible to move between two vectors which differ in l places in l steps).

It also has stationary distribution $P_{\mathbf{Y}}(\mathbf{y})$. Choosing $q_{ij} = p_{ij}$, we can show that both $v_j q_{ji}$ and $v_i q_{ij}$ are equal and so $a_{ij} = 1$ and $p_{ij} = a_{ij} q_{ij}$, matching the previous section, confirming the above stationary distribution.

15.3 Gibbs Sampling for multiple sequence alignments

Suppose that we wish to compare more than two sequences to try and find the best alignment between them. If we have N sequences and they are each of approximate length L, then the number of (global) alignments is of the order of

$$(2L)^N$$

which can easily be far too large to handle in the conventional manners described in section 13. We shall use the Gibbs sampling method from the previous section to try to find good alignments (it is usually too ambitious to try to find the "best" alignment out of so many possibilities). We consider the example of protein sequences.

Label the amino acids in some order $1, \dots, 20$. Supposing we now have N protein sequences of lengths $L_1, \dots, L_N$, the aim is to find N segments, one form each sequence, each of some fixed length W which in some way are similar to each other. There are

$$\prod_{i=1}^{N} (L_i - W + 1)$$

possible choices for the locations of these N segments. We assume that N and the L_j are large enough that a simple algorithmic approach is not feasible. We shall consider each of these possible choices as a state in a Markov chain.

The procedure works as follows. We follow a series of simple steps, moving from one state of the Markov chain (i.e. one array) to another. The figures below illustrate this. In Figure 15.1 we choose the third sequence, so the third row is changed in step 1, and the first sequence (so the first row) in step 2.

FIGURE 15.1

In fact each step has two parts; selection of the row to be changed, and the manner of its change. Figure 15.2 shows the first part of the step only. Which row is chosen is a purely random choice [or we could pick each row in turn as described in the previous section]. The aim of the second part, is to improve the overall alignment, by choosing one of the $L_i - W$ alternative segments.

FIGURE 15.2

Suppose that in the first reduced array of Figure 15.2 (i.e. with row 3 removed) amino acid j occurs c_{ij} times in the ith column. From these values, we find a probability estimate

$$q_{ij} = \frac{c_{ij} + b_j}{N - 1 + B}$$

where the b_j are what are called pseudocounts. They can be chosen in a number of ways, but one sensible choice is $b_j = p_j$, the background frequency of amino acid j.

Suppose that the amino acid sequence in a given segment is $\mathbf{x}= x_1, x_2, \ldots, x_W$. The probability of this ordered set under the population amino acid frequencies is

$$P_x = p_{x1} p_{x2} \ldots p_{xW}$$

The estimated probability of this segment under the $N - 1$ segments in our reduced array is

$$Q_x = q_{1x1} q_{2x2} \ldots q_{WxW}$$

The likelihood ratio (which here is an estimate of how much more likely our sequence is due to similarity to the other sequences over pure chance, is $LR(x) = Q_x/P_x$. We select segment $\mathbf{x}$ with probability

$$\frac{LR(x)}{\sum_{m=1}^{L_i - W} LR(m)}$$

where the denominator is the sum of all the likelihood ratios of the $L_i - W$ possible segments.

As time goes on, we should see better and better alignments appearing more frequently.

The relative entropy between $q^*_{ij}(s)$ and p_j is

$$\sum_{i=1}^{W}\sum_{j=1}^{20} q^*_{ij}(s) log \left(\frac{q^*_{ij}(s)}{p_j} \right)$$

States for which the relative entropy is high are those which represent good alignments. Note that this entropy is approximately a linear function of the logarithm of the probability of being in state s

$$C \prod_{i=1}^{W}\prod_{j=1}^{20} \left(\frac{q^*_{ij}(s)}{p_j} \right)^{c_{ij}(s)}$$

This method is an example of the Gibbs sampling method that we described above. Each state of the process is a vector with N elements (one for each of our sequences) and there are a finite number of possible values for each element $L_i - W + 1$ for element i.

FIGURE 15.3

Each step of the process changes a single sequence only (i.e. changes a single element of the vector) so that $p_{ij} = 0$ if elements i and j differ by more than a single element. Since the probability expression

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k)$$

is almost identical to the likelihood ratio $LR(x)$ with the appropriate probabilities defined above, it follows that this is essentially an example of a Gibbs sampling procedure.

16 Genetics

16.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg principle states that, under certain conditions, after a single generation of random mating, the genotype frequencies at a single locus will be fixed at a particular equilibrium value, which is the *Hardy-Weinberg equilibrium.*

Suppose that there are two alleles at a given locus, labelled by A and a, and the proportions of these in the population are p and $q = 1 - p$ respectively. A population is in Hardy-Weinberg equilibrium if the proportion of zygote AA in the population is p^2 and the proportion of the heterozygote Aa is $2pq$. These proportions are those that would be obtained if the genes were selected from the population at random; the number of As at the locus is simply Binomially distributed, with parameters 2 and p.

Exercise: Show that after a generation of random mating a population with two alleles is in Hardy-Weinberg equilibrium, irrespective of the original distribution of zygotes.

The original assumptions for a population to be in Hardy-Weinberg equilibrium were that the population is; diploid, sexually reproducing and randomly mating. in addition the population does not suffer drift, selection, mutation or migration.

The principle can easily be generalised to the case with multiple alleles. Suppose that there are n alleles at a locus

$$A_1, \dots, A_k$$

and these have population frequencies of $A_1, \dots, A_k$ respectively, then under Hardy-Weinberg equilibrium the frequency of homozygote A_iA_i is p_i^2 and the frequency of heterozygote A_iA_j is $2p_ip_j$ for all values of i, j.

We may be interested in whether a particular population is in Hardy-Weinberg equilibrium, or if not, how much (in some sense) that it deviates from it. We shall explore different methods of testing for this.

16.2 The chi-square test

One way of testing for hardy-Weinberg equilibrium is to use the chi-square test that we saw last term in the general context of tabular data. Below we reproduce a revised version of the table from Section 7. In this case we have the entry Y_{ij} $i \leq j$ representing the number of individuals with genotype A_iA_j. The table has the same number of rows and columns (equal to the total number of alleles k). Since A_iA_j and A_jA_i are the same, we only need a single category, and we choose the row i column j entry for $i < j$. The total number of individuals is thus

$$N = \sum_{i \leq j} Y_{ij}$$

$$\begin{array}{cccc}
Y_{11} & Y_{12} & \dots & Y_{1k} \\
Y_{21} & Y_{22} & \dots & Y_{2k} \\
\text{-} & \text{-} & \dots & \dots \\
\text{-} & \text{-} & \text{-} & Y_{kk}
\end{array}$$

The null hypothesis is that the population is in Hardy-Weinberg equilibrium.

Under H_0 the statistic is

$$Q = \sum_{i=1}^{k} \sum_{j=i}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $O_{ij} = Y_{ij}$ is the observed number of individuals with genotype A_iA_j and E_{ij} is the expected number of observations in this cell. E_{ij} is found simply by multiplying the total number of individuals by the probability that a random individual (under Hardy-Weinberg equilibrium) has that genotype. Hence if we knew the values of p_is we would obtain

$$E_{ii} = p_i^2 N \quad E_{ij} = 2p_ip_jN \quad i < j$$

Usually we have to estimate the p_is from the data. We estimate p_i (the proportion of allele A_i in the population) by the proportion of allele A_i in the sample, which is

$$\hat{p_i} = \frac{\sum_{i<j} Y_{ij} + \sum_{i>j} Y_{ji} + 2Y_{ii}}{2N}$$

Q has approximate distribution χ^2 with $k(k-1)/2 + k - k = k(k-1)/2$ degrees of freedom (the number of cells minus the number of terms estimated from the data).

Example 16.1 - suppose that we have a system with two alleles where the zygotes A_1A_1, A_1A_2 and A_2A_2 occur in frequencies 19, 52, 27 respectively. In the above notation, this gives

$$Y_{11} = 19, Y_{12} = 52, Y_{22} = 27, N = 98$$

The proportion of allele A_1 in our data sample is

$$\frac{52 + 2 \times 19}{2 \times 98} = 0.4592$$

and so the proportion of A_2 is 1-0.4592=0.5408. Thus our population proportion estimates are $\hat{p_1} = 0.4592$ and $\hat{p_2} = 0.5408$.

Under the null hypothesis of Hardy-Weinberg equilibrium, the expected frequencies of the three zygotes are
$E_{11} = 98(0.4592)^2 = 20.66$
$E_{12} = 2 \times 98(0.4592)(0.5408) = 48.67$
$E_{22} = 98(0.5408)^2 = 28.66$

$$Q = \sum_{i=1}^{k} \sum_{j=i}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(19 - 20.66)^2}{20.66} + \frac{(27 - 28.66)^2}{28.66} + \frac{(52 - 48.67)^2}{48.67} = 0.457$$

This should be χ_1^2. The 95% point of this distribution is $3.84 > 0.457$. So we do not reject the null hypothesis that the population is in Hardy-Weinberg equilibrium.

16.3 Fisher's exact test

Fisher's exact test is a version of the Likelihood Ratio Test, useful especially when some categories contain small numbers. The likelihood function when Hardy-Weinberg equilibrium is not assumed is

$$\prod_{i=1, j \geq i} q_{ij}^{Y_{ij}}$$

where there are no restrictions on the values of q_{ij}, except that they all add to 1. When there are 2 alleles, this becomes

$$q_{11}^{Y_{11}} q_{12}^{Y_{12}} q_{22}^{Y_{22}}$$

Under the null hypothesis of Hardy-Weinberg equilibrium, $q_{ij}, i \neq j$ is replaced by $2p_ip_j$ and q_{ii} by p_i^2 and for two alleles the likelihood function becomes

$$p^{2Y_{11}}(2p(1-p))^{Y_{12}}(1-p)^{2Y_{22}}$$

where p is the frequency of A_1 in the population.

We proceed to find the maximum likelihood function under the null hypothesis, and without restrictions. The ratio of these two is the likelihood ratio Λ.

It can be shown that with no restriction, the maximum likelihood estimates (the third is implied by the other two, since all probabilities must add to 1) are

$$\hat{q_{11}} = \frac{Y_{11}}{N}, \hat{q_{12}} = \frac{Y_{12}}{N}, \hat{q_{22}} = \frac{Y_{22}}{N}$$

and so the likelihood function is

$$\left(\frac{Y_{11}}{N}\right)^{Y_{11}} \left(\frac{Y_{12}}{N}\right)^{Y_{12}} \left(\frac{Y_{22}}{N}\right)^{Y_{22}}$$

Under the null hypothesis the maximum likelihood estimate of p is

$$\hat{p} = \frac{Y_{11} + Y_{12}/2}{N}$$

and the corresponding likelihood function is

$$\left(\frac{Y_{11} + Y_{12}/2}{N}\right)^{2Y_{11}+Y_{12}} \left(\frac{Y_{22} + Y_{12}/2}{N}\right)^{2Y_{22}+Y_{12}} 2^{Y_{12}}$$

$-2log(\Lambda)$ can be shown to be

$$2(Y_{11}log(Y_{11}) + Y_{12}log(Y_{12}/2) + Y_{22}log(Y_{22}) + Nlog(N)$$

$$-(2Y_{11} + Y_{12})log(Y_{11} + Y_{12}/2) - (2Y_{22} + Y_{12})log(Y_{22} + Y_{12}/2))$$

Under the hypothesis of Hardy-Weinberg equilibrium, this quantity has an approximate χ_1^2 distribution.

Considering the above example with $Y_{11} = 19, Y_{12} = 52, Y_{22} = 27, N = 98$

$-2log(\Lambda) = (19log(19)+52log(26)+27log(27)+98log(98)-90log(45)-106log(53)) = 0.458$

$0.458 < 3.84$ similarly to above, so we do not reject the null hypothesis of Hardy-Weinberg equilibrium.

16.4 Estimating the Hardy-Weinberg disequilibrium

We have seen above how to test to see whether Hardy-Weinberg equilibrium is plausible for a given set of data. What if it is not (or you think that it may well not be true) so that you wish to estimate the degree of disequilibrium? An estimate can be found by comparing the true proportion of the heterozygote AB, with the predicted level if it was under Hardy-Weinberg. The disequilibrium is defined as half of this difference

$$D = \frac{1}{2}\left(\frac{Y_{12}}{N} - 2\hat{p}(1-\hat{p})\right)$$

We can find an approximate confidence interval by using the Hardy-Weinberg proportions to estimate the variance of $Y_{12}/2N$. This estimate is

$$\frac{\hat{p}(1-\hat{p})}{4N}$$

So that the 95% confidence interval is

$$D - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{4N}}, D + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{4N}}$$

Thus in our example above,

$$D = \frac{1}{2}\left(\frac{52}{98} - 2 \times 0.4592(1 - 0.4592)\right) = 0.0170$$

and the 95% confidence interval is

$$D - 1.96\sqrt{\frac{0.4592(1-0.4592)}{4 \times 98}}, D + 1.96\sqrt{\frac{0.4592(1-0.4592)}{4 \times 98}} = (-0.0323, 0.0663)$$

17 Phylogenetic trees and evolutionary models I

17.1 Introduction to phylogenetic trees

The evolutionary relationships between a set of species with a common ancestor can be represented by a binary tree, as we see in the following figure. Species are represented by points (nodes) connected by lines (edges).

FIGURE 17.1

The lengths of the edges represent evolutionary time, the longer an edge, the more time separates the nodes at the ends of the edge. The trees in Figure 17.1 are examples of rooted trees, where there is a single root node representing the common ancestor. An unrooted tree indicates the relationship between species without showing the direction of evolutionary time, as in Figure 17.2.

FIGURE 17.2

For a set of species there will be some real, unknown, phylogenetic tree connecting the members. Our aim is in infer what this tree is as accurately as possible, given the data. We shall look at several methods of doing this.

17.2 Distances

Several methods and algorithms are based on the concept of a *distance* between species. The most natural measure of distance between two species is the number of years since their most recent common ancestor. This is usually unknown, so in practice we have to use a surrogate distance in its place.

let us assume that exact distances are known. For a set of points S, any measure of distance d must satisfy the following for all elements x, y, z in S.

$$(i) d(x,y) \geq 0$$

$$(ii) d(x,y) = d(y,x)$$

$$(iii) d(x,y) \leq d(x,z) + d(z,y)$$

A distance is a *tree-derived* distance if there is a tree with these species at the leaves such that the distance between x and y is the sum of the lengths of the edges joining them. This automatically satisfies $(i) - (iii)$ above. In fact it turns out that (iii) becomes

$$(iii) d(x,y) < d(x,z) + d(z,y)$$

for such a tree.

For rooted trees joining extant species (non-extinct, so being measured at identical times) there are further restrictions. For the tree in Figure 17.3 the conditions
$d(x, y) = d(x, z)$
and $d(y, z) < d(x, y), d(y, z) < d(x, z)$
must hold.

FIGURE 17.3

Any distance measure which satisfies the condition that for any three members x, y, z of S, two of the three distances between them are equal and the third is smaller than these two, as above, is called *ultrametric*. So a rooted tree of extant species with a tree-derived distance is always ultrametric.

17.3 Tree reconstruction: the ulttrametric case

We shall give an outline of how to show that if an ultrametric distance measure is given for all species then there is a unique rooted tree joining the species (except for trivial changes).

Suppose that we have a set of species $s_1, s_2, \ldots, s_n$ with an ultrametric distance measure $d(x, y)$ between each pair of species x and y.

We shall use an induction argument to show the existence of a unique tree. For only two species, then both must be the same distance from the root and we obtain the following tree.

FIGURE 17.4

Suppose that we can find a unique tree (except for trivial changes like swapping s_1 and s_2 in the above) for m species. If we can show that under this assumption we can do the same for $m + 1$ then we can do it for any number (we can find a tree for 2, and so for 2+1=3, and so for 3+1=4 etc.). For our m species, there is a single root, with a collection of species (labelled S_L) down the left hand branch, and another collection (S_R) down the right-hand one. Let x be some element of S_L and y some element of S_R.

FIGURE 17.5

Now consider the extra species s_{m+1} and its distance from x and y. The ultrametric property indicates that one of x, y and s_{m+1} is equidistant from the other 2. Thus there are three possibilities:

$$(1) d(s_{m+1}, x) = d(s_{m+1}, y) > d(x, y)$$

leads to the following tree (where $a = d(s_{m+1}, x)/2$ and $b = (d(s_{m+1}, x) - d(x, y))/2$)

FIGURE 17.6

We thus have the correct distances between x, y and s_{m+1}. It is not hard to show that this must also work for every other species as well.

$$(2) d(s_{m+1}, y) = d(x, y) > d(s_{m+1}, x)$$

This puts the species s_{m+1} on the same side of the root as S_L. If s_{m+1} is equidistant from every member of S_L we get the following tree

FIGURE 17.7

If not, s_{m+1} is closer to some members of S_L than others, and we can follow similar arguments within S_L to find its correct place.

The third case is

$$(3) d(s_{m+1}, x) = d(x, y) > d(s_{m+1}, y)$$

Case (3) is the same as case (2), just with S_R and S_L interchanged.

17.4 Tree reconstruction: neighbour joining

Two species are *neighbours* if the path between them contains only one node. Thus in the following tree, x and y are neighbours, but x and z are not.

FIGURE 17.8

The following describes the UPMGA (unweighted pair group method using arithmetic averages) algorithm. The algorithm describes distances between groups of species, starting with 'groups' of a single species.

We define the distance between groups u and v as

$$d^*(G_u, G_v) = \frac{1}{n_1 n_2} \sum_{x \in G_u, y \in G_y} d(x, y)$$

where n_1 and n_2 are the sizes of groups G_u and G_v. We find the smallest distance between two groups, and join them with a two-leaf rooted tree, with root "species" r_1 as follows

FIGURE 17.9

The distance from any node z to this root is given by;

$$d(r_1, z) = \frac{1}{2}(d(x, z) + d(y, z) - d(x, y))$$

We now replace the two groups G_r and G_s by the group G_{rs} containing the elements of both groups. We repeat the above procedure, now with one group less, until eventually we have a complete tree.

The *neighbour-joining* algorithm is a more complex procedure, based upon the function

$$\delta(x, y) = (N - 4)d(x, y) - \sum_{z \neq x,y} (d(x, z) + d(y, z))$$

It can be shown that this reaches its minimum value if and only if x and y are neighbours, so that the algorithm joins such pairs immediately. Distances from the node (r_1) joining any such pairs are now calculated, again using;

$$d(r_1, z) = \frac{1}{2}(d(x, z) + d(y, z) - d(x, y))$$

Exercise: Repeat E3 using the neighbour-joining algorithm, instead of the UPGMA.

17.5 Surrogate distances

Unfortunately it is rarely the case that exact distances between species are known, and we must use surrogate distances instead. These are usually derived using DNA information from the species being considered. Evolutionary changes can happen faster along one branch than another for a number of reasons, for example different generation lengths, so we may obtain a tree like Figure 17.10.

FIGURE 17.10

In this example the molecular clock ran slower for the elephant than the others. A surrogate distance is often found by using aligned DNA sequences taken from two species, where the distance is proportional to

$$-log\left(1 - \frac{4}{3}p\right)$$

where p is the proportion of nucleotides where the two sequences differ. This estimate derives from the Jukes-Cantor model, which we shall meet in the next section. We can then use the methods that we have already described to estimate the tree with the surrogate distances instead of the distances (note that the basic properties of a distance (i)-(iii) may not be satisfied, so that negative distances can appear in the inferred trees). Quite often more complex versions of this type of idea need to be used.

17.6 Tree reconstruction: parsimony

Using the method of parsimony, a cost is assigned to each tree, and the optimal tree is the one which minimises the cost. The aim is to find the optimal tree *topology.*

Consider the cost function where unit cost is made for each nucleotide substitution. We find the optimal tree in two steps. Firstly all possible tree topologies must be listed, with species allocated to the leaves. Secondly, for each such choice, the labelling of all internal nodes with a suitable DNA sequence, to ensure minimisation of the cost must be found (this is done using *Fitch's algorithm*).

For a large number of species, the first step is the hardest. For three species there is only one possible topology, with three labelling choices. The following figure shows the five optimal allocations of internal nodes for the simplistic sequences AA, AB and BB.

FIGURE 17.11

The number of topologies increases rapidly with species number. For example, for 20 species there are 8×10^{21} topologies. It is possible to use versions of the above methodology on very large numbers of species. The distinctive feature of this method is that it does not construct distances between species, only the topology.

18 Phylogenetic trees and evolutionary models II

18.1 Introduction to evolutionary models

We shall introduce some models of the evolution of biological data which will be useful for the construction of phlyogenetic trees. We shall assume that there is one dominant nucleotide at any given site for a particular species, so that each species has its own "genome". Over time the nucletides at any site may change; such changes are assumed to occur over negligible times (in evolutionary time this is reasonable). Such a change is called a *substitution*.

18.2 The Jukes-Cantor model

The simplest model of nucleotide substitution is the jukes-Cantor model. The discrete time version considers a Markov chain with four states a, g, c, t with a transition matrix (probability of moving from one state to another in unit time) given by

$$P = \begin{vmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{vmatrix} \tag{10}$$

The value of α will obviously depend upon the chosen timescale. Thus whatever nulceotide exists at whichever site, it is equally likely to be substituted by any of the other three.

The stationary distribution of this Markov chain is just
$\phi = (0.25, 0.25, 0.25, 0.25)$.
It is possible to show that after n steps the transition matrix P^n is

$$P^n = \begin{vmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{vmatrix} \tag{11}$$

$$+(1-4\alpha)^n \begin{vmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{vmatrix} \tag{12}$$

Thus, whatever the dominant nucleotide at time 0, the probability that it is still the dominant one at time n is

$$0.25 + 0.75(1 - 4\alpha)^n$$

and the probability that it is any specific other one is

$$0.25 - 0.25(1 - 4\alpha)^n$$

18.3 The Kimura models

The assumption that all transitions between nucleotides are equally likely is unrealistic, and some are more likely than others. In particular, transitions between a and g, and between c and t are more common than the others. Kimura's first model, allows for two different trnasition probabilities to occur; α is the probability of moving between a and g, or c and t and other changes have rate β. Thus the transition matrix becomes

$$P = \begin{vmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{vmatrix} \tag{13}$$

The stationary distribution of this Markov chain can again be shown to be the simple equal probability one,
$\phi = (0.25, 0.25, 0.25, 0.25)$.
As in the Jukes-Cantor model, we can find the transition matrix P^n.

$$P^n = \begin{vmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{vmatrix} \tag{14}$$

$$+(1-4\beta)^n \begin{vmatrix} 0.25 & 0.25 & -0.25 & -0.25 \\ 0.25 & 0.25 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.25 & 0.25 \\ -0.25 & -0.25 & 0.25 & 0.25 \end{vmatrix} \tag{15}$$

$$+(1-2(\alpha+\beta))^n \begin{vmatrix} 0.5 & -0.5 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & -0.5 \\ 0 & 0 & -0.5 & 0.5 \end{vmatrix} \tag{16}$$

Thus, whatever the dominant nucleotide at time 0, the probability that it is still the dominant one at time n is

$$0.25 + 0.25(1-4\beta)^n + 0.5(1-2(\alpha+\beta))^n$$

The probability that it has been replaced by its "partner" (e.g. a by g) is

$$0.25 + 0.25(1-4\beta)^n - 0.5(1-2(\alpha+\beta))^n$$

and the probability that it is a specific one of the other two is

$$0.25 - 0.25(1-4\beta)^n$$

There are other more complex Kimura models where each of the substitution probabilities differ, but we shall not look at those here. The Felsenstein Model and the HKY Model are two further generalisations on the same theme using Markov transition matrices for nucleotide substitutions.

The Felsenstein model generalises the Jukes-Cantor model in a different way. It recognises that the equilibrium distribution of nulceotides is often not uniform i.e. (0.25,0.25,0.25,0.25) and so gives a transition matrix which gives the equilibrium distribution $(\phi_a, \phi_g, \phi_c, \phi_t)$ for any such values. This matrix is

$$P = \begin{vmatrix} 1-u+u\phi_a & u\phi_g & u\phi_c & u\phi_t \\ u\phi_a & 1-u+u\phi_g & u\phi_c & u\phi_t \\ u\phi_a & u\phi_g & 1-u+u\phi_c & u\phi_t \\ u\phi_a & u\phi_g & u\phi_c & 1-u+u\phi_t \end{vmatrix} \tag{17}$$

It is not too hard to verify that the stationary distribution of this matrix is indeed $(\phi_a, \phi_g, \phi_c, \phi_t)$

The HKY Model is a further generalisation, with features from both the Kimura and Felsenstein models (these are both special cases of it). The transition matrix is given by

$$P = \begin{vmatrix} 1-u\phi_g-v\phi_c-v\phi_t & u\phi_g & v\phi_c & v\phi_t \\ u\phi_a & 1-u\phi_a-v\phi_c-v\phi_t & v\phi_c & v\phi_t \\ v\phi_a & v\phi_g & 1-u\phi_t-v\phi_a-v\phi_g & u\phi_t \\ v\phi_a & v\phi_g & u\phi_c & 1-u\phi_c-v\phi_a-v\phi_g \end{vmatrix} \tag{18}$$

Exercise: Show that the HKY model has the required stationary distribution.

In addition, each of these models can be converted to continuous time models quite easily, with transition probabilities being replaced by transition rates.

18.4 Tree reconstruction: maximum likelihood

Suppose that we already have the topology of the tree, and we are interested in estimating the various lengths of the edges. Finding a likelihood function is relatively straightforward, given an evolutionary model. Thus we choose our model (for instance the Jukes-Cantor model). As an example, suppose that we have data from 5 species, which are arranged in the following tree.

FIGURE 18.1

At a particular collection of sites the nucleotide present for each of the species is known (and we assume that this is all that is known). We shall follow one such site, and label the nucleotides of the five species as A_1, A_2, A_3, A_4 and A_5 (some of these can of course be the same). In the tree there are nine nodes, and our likelihood function will be a product of 9 terms. Firstly the probability that W is the nucleotide at node n_0 is written as ϕ_W (taken from the its overall frequency in the population). There are eight arms leading from one node to another; each has an associated probability of a change from the nucleotide at the start node to the one at the end in the time given by the length of the node. For example, the probability that a change from nucleotide X at node n_1 to A_3 at node s_3 is labelled $P_{XA3}(l_4)$. If the nucleotides at the internal nodes n_1, n_2 and n_3 are X, Y and Z respectively, then the likelihood for our arm lengths AND these nucleotides at the root and internal nodes is given by

$$\phi_W P_{WX}(l_1) P_{WZ}(l_2) P_{XY}(l_3) P_{XA3}(l_4) P_{YA2}(l_5) P_{YA1}(l_6) P_{ZA4}(l_7) P_{ZA5}(l_8)$$

We now compute the above formula for the $4^4 = 256$ possible combinations of nucleotides at the root and the internal nodes, and the sum of all of these terms is the likelihood of the arm lengths $l_1, l_2, \ldots, l_8$.

This procedure is now repeated over all the sites that we possess information on, and a full likelihood function is found, which is just the product of all the likelihood functions at the single sites.

We can, in principle, repeat this for all tree topologies, and then find the largest of these likelihood functions, which will be from the tree which becomes our maximum likelihood estimator. Of course, as with the other methods, when the number of species is large, this can be very complex.

Felsenstein suggested a plan of building a tree (unrooted) by starting with two species and adding other species successively. If at any stage there are $k-1$ species, it turns out that there are $2k-5$ ways of adding a further species. Each of these should be tried in turn and that with the maximum likelihood accepted. This does not necessarily lead to the tree at the end of the process with maximum likelihood, and different trees can be obtained with different starting species. Different orders can be tried, and the final result with the biggest likelihood accepted.

It has been noticed that for real data sets, simplified evolutionary models like those that we have discussed can give rise to severe errors in the estimation of branch lengths. Increasingly complex evolutionary models (for instance Kimura's model is more complex than the Jukes-Cantor one, which is a special case of it) improve matters somewhat, but significant errors can still occur . Thus sometimes it is necessary to work with the more complex models.

The following example is taken from Ewens and Grant and illustrates the modelling of the evolutionary relationship between 14 different mammals. These are as follows:
Marsupial Mole, Wombat, Rodent, Elephant Shrew, Elephant, Whale, Dolphin, Pig, Horse, Bat, Insectivore, Human, Sea Cow, Hyrax.
The sequences for the 14 species are in Ewens and Grant, page 405. The distances between the species were found using the Kimura model of the previous section, and are given in the following table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 ma. mole	0	.11	.41	.44	.39	.37	.40	.37	.41	.36	.40	.37	.42	.39
2 wombat	.11	0	.39	.40	.36	.33	.35	.33	.35	.32	.33	.33	. 38	.34
3 rodent	.41	.39	0	.33	.30	.24	.25	.22	.28	.23	.25	.23	.32	.31
4 el. shrew	.44	.40	.33	0	.20	.26	.26	.25	.28	.28	.28	.26	.20	.21
5 elephant	.39	.36	.30	.20	0	.22	.23	.22	.25	.25	.24	.21	.11	.12
6 whale	.37	. 33	. 24	. 26	. 22	0	.03	.10	.16	.17	.18	.17	.22	.24
7 dolphin	.40	.35	.25	.26	.23	.03	0	.11	.16	.19	.18	.17	.22	.25
8 pig	.37	.33	.22	.25	.22	.10	.11	0	.17	.18	.19	.17	.24	.24
9 horse	.41	.35	.28	.28	.25	.16	.16	.17	0	.17	.21	.20	.25	.26
10 bat	.36	.32	.23	.28	.25	.17	.19	.18	.17	0	.15	.20	.27	.27
11 insectiv.	.40	.33	.25	.28	.24	.18	.18	.19	.21	.15	0	.19	.26	.26
12 human	.37	.33	.23	.26	.21	.17	.17	.17	.20	.20	.19	0	.22	.23
13 sea cow	.42	.38	.32	.20	.11	.22	.22	.24	.25	.27	.26	.22	0	.14
14 hyrax	.39	.34	.31	.21	.12	.24	.25	.24	.26	.27	.26	.23	.14	0

Four different methods were used to find the optimal phylogenetic tree using the distances from the above table. The methods were UPGMA, neighbour-joining, parsimony and maximum likelihood. The trees found using these methods are shown in Figure 18.1.

FIGURE 18.2

Some features are common to the trees formed by all four methods, whereas others vary from tree to tree. Thus the relationship between horse, pig, whale and dolphin is the same for all four trees. There are large differences in the placement of humans, however. In the UPGMA tree, 'human' is grouped with horse,pig, whale and dolphin; in the neighbour joining tree it is not that closely linked to those types. The parsimony tree puts humans in the same group as elephant, sea cow, hyrax and elephant shrew. The maximum likelihood tree places humans and rodents together.

19 Evolutionary and Genetic Algorithms

Evolutionary algorithms are stochastic searching models which are based upon the processes of evolution. They may be used to model actual evolutionary processes, and this is what we will assume in this section, but also to model completely different situations. For instance the classical travelling salesman problem can be approached using evolutionary algorithms.

Genetic algorithms are a special class of evolutionary algorithms and we will use the terms interchangeably from now on. The basic structure of a genetic algorithm is as follows. There will be a set of optimisation criteria in place to determine whether we have reached a suitable solution to our problem (or evolved to a stable population).

FIGURE 19.1

Step 1: Generate an initial population of individuals. This will usually be done completely randomly, with no thought given to the 'fitness' of individuals.

Step 2: Are the optimisation criteria met? If so we stop the process. This is extremely unlikely to happen at the start of the process. If they are not met we continue.

Step 3: Generate a new population. This is done in a number of steps. Not all have to be included in any particular process, but all are needed to replicate a real genetic population (although leaving some out may lead to broadly similar conclusions).

Step 3a: Selection occurs according to the fitness of the individual to decide which individuals reproduce.

Step 3b: Recombination. Parents are recombined to produce offspring.

Step 3: Mutation. all offspring will be mutated with a certain probability.

We can then compute the fitness of the offspring and replace the parents with the offspring as the members of the population.

Step 4: Are the optimisation criteria met? If so the process stops, as in step 2. if they are not, then step 3 is repeated with the 'new' population replacing the original at the start, and undergoing further selection, recombination and mutation.

Evolutionary algorithms thus operate in a similar way to the more traditional optimisation methods, but there are differences. The major differences are that evolutionary algorithms are probabilistic rather than deterministic (and information such as derivatives are not needed) and that at any point the process is not at a point but at a set of points (a population).

We will now consider the main parts of the process of the evolutionary algorithm in turn.

19.1 Selection

The selection step chooses which individuals generate new offspring in the next generation and is the most fundamental step in the process. To perform selection, we must have a measure of fitness of each individual. Each individual in the population has a reproduction probability based upon their own objective fitness value, and on that of all the other individuals in the population.

Thus the probability of any new individual being a copy of type i is $P_i(f_1, f_2, \ldots, f_n)$ for some function P_i such that

$$\sum_{i=1}^{n} P_i(f_1, f_2, \ldots, f_n) = 1$$

We talk in terms such as *selective pressure* the relative probability of the best individual being chosen compared to the average probability of selection, and *loss of diversity* which is the proportion of individuals in a population that are not selected at all (and so have no offspring) by the above method.

The simplest form of selection is what is known as *roulette wheel selection* or *stochastic sampling with replacement.* If there are n individuals in our current population with fitnesses $f_1, f_2, \ldots, f_n$ as above, and we wish to generate m new individuals for our *mating population* (i.e. that prior to recombination) in our next population we choose each of these m independently where each of a copy of individual i with probability

$$P_i(\mathbf{f}) = \frac{f_i}{f_1 + f_2 + \ldots + f_n}$$

Thus the number of individuals in the mating population who are copies of individual i follows a Binomial distribution with parameters m and $P_i(\mathbf{f})$.

There are many other types of selection; for example individuals can be placed spatially in an environment and selection occurs within local groups, so that characteristics can vary considerably between localities.

19.2 Recombination

The selection process has generated a mating population. Individuals are now paired according to some rule; completely at random, according to whether an individual has been designated male or female (and then at random) or within the locality for example.

When a pair is chosen, they have to generate offspring.

Let us suppose that individual 1 is described by the the sequence (we shall refer to this as the genome of the individual)

$$t_{11}, t_{12}, \ldots, t_{1k}$$

and individual 2 by

$$t_{21}, t_{22}, \ldots, t_{2k}$$

(for the purposes of our process k may be 1 or 2 or much more, typically each value representing an allele of some gene).

This recombination (random recombination) can be done simply by choosing the first element from the sequence for the offspring to be t_{11} or t_{21} each with probability 0.5, and similarly for each other value, all independently of each other. So our new individual is generated as in the following figure.

FIGURE 19.2

When our elements are continuous values, we can generate new elements by using a linear combination of the parental values, so that the ith element is

$$\alpha_i t_{1i} + (1 - \alpha_i)(1 - t_{2i})$$

where α_i is an observation from a uniform (0,1) distribution. There are other variants on this idea.

An alternative method is the *crossover* method, which takes into account the correlation between genes on the same chromosome. For instance, the single-point crossover chooses a value at random from the list of positions $0, 1, \ldots, k$ and then creates two new individuals. If the chosen value is j, then our two new individuals would be

$$t_{11}, t_{12}, \ldots, t_{1j}, t_{2j+1}, \ldots, t_{2k}$$

and

$$t_{21}, t_{22}, \ldots, t_{2j}, , t_{1j+1}, \ldots, t_{1k}$$

There are, again, many variants on this idea.

19.3 Mutation

We thus have a population of new individuals who are perfect copies of a combination of their two parents. There will be many genomes which can never be reached by selection and recombination alone; and they might be the fittest. Mutations prove very useful in helping to find the best genomes.

Mutations are random alterations in individuals. The mutation rate is related to the probability of a change in any given element of the genome of the individual. Generally to obtain good results, this mutation rate should decrease with the number of variables in the genome of our individuals; if there are n elements in the genome, a mutation rate of $1/n$ is sensible.

How mutations should occur, in the sense of which allele should be generated when a mutation occurs from a given allele, depends upon the form of alleles. If they are continuous values, or form a sequence of discrete values where there is an obvious ordering, then generally mutations should be small in size most of the time, but occasionally of large size.

For instance for continuous data, the mutation distance can be normally distributed, possibly centred on the original value (but possibly not). Thus if we start at allele t, then there is no mutation with probability $1-r$ and a mutation occurs with probability r which sends t to s where the value of s follows a normal distribution with mean t and variance σ^2, where σ^2 is a measure of the size of mutations, the larger σ^2, the larger mutation distances tend to be.

19.4 Fitness values

After a cycle of selection, recombination and mutation, we finally have a new generation. At the start of the section on selection, we talked about the fitnesses f_i of individuals. These are in reality properties of the genome of the individual and the should be a general function of the genome which is its fitness. It could be something simple like the sum of the genetic components $t_{11} + t_{12} + \ldots + t_{1k}$, for our first individual from before, so the larger the values of each t_{1i} the better, but will usually be something more complicated which it is difficult to see how to optimise.

The fitness of real genetic situations is of course very complex, but this is also true for the other types of problem that the genetic algorithm can be applied to, such as the travelling salesman problem.

The process continues until we find a satisfactory stopping point (perhaps the difference between the fittest and least fit individual is so small that all diversity is lost, barring mutations). The nature of the process means that this will generally occur at a point of high population fitness (but of course not necessarily the highest possible such point).

20 Dynamic programming

20.1 Introduction

Dynamic programming is a method of solving problems that are apparently complex but highly structured, by making use of the structure. The general approach is to break the problem down into subproblems and build the final solution using the solutions to each of the subproblems.

This type of approach is ideal for computers as it typically involves the repetition of many fairly straightforward calculations. One application of this idea is in games where there are sequential decisions to be made and where it is possible to identify all of the end-points of the game. The best strategy at each point depends upon what the best reply is; the game can be solved by working backwards from all of the possible final positions (backwards induction). *Noughts and crosses* can be completely solved in this way, but theoretically so could far more complex games like chess, if computers were powerful enough.

We shall consider one example of this approach that we met briefly in Section 13, which deals with the alignment of two DNA or protein sequences.

20.2 The Needleman Wunsch algorithm

The input for the Needleman Wunsch algorithm consists of two sequences

$$x = X_1X_2 \dots X_m, \quad y = Y_1Y_2 \dots Y_n$$

of lengths m and n respectively. The elements of these sequences belong to some given alphabet of N symbols ($N = 4$ for DNA sequences and $N = 20$ for protein sequences). In addition there is a linear gap penalty d, and a given substitution matrix S (e.g. for DNA data the score of an alignment between an a and a g would be the element in row 1, column 2 of this matrix).

We find a highest scoring alignment for two sequences, based upon the highest-scoring alignments of smaller subsequences of x and y.

Denote the segment of the first i elements of x by $x_{1,i}$, and the segment of the first j elements of y by $y_{1,j}$ i.e.

$$x_{1,i} = X_1X_2 \dots X_i \quad y_{1,j} = Y_1Y_2 \dots Y_j$$

Label a highest scoring alignment of $x_{1,i}$ and $y_{1,j}$ by $B(i, j)$ for $i = 0, 1, \dots, m$ and $j = 0, 1, \dots, n$. When $j = 0$ (or $i = 0$) this is just the alignment of a sequence of elements against the appropriate number of gaps (each scoring $-d$), so that

$$B(i,0) = -id, \quad B(0,j) = -jd$$

We also set $B(0,0) = 0$.

if we can find all of the $B(i,j)$ we have an $(m+1) \times (n+1)$ matrix, where the entry in the last row and column, $B(m,n)$ is a highest-scoring alignment of the two full sequences.

From above we already know $B(i,0)$ and $B(0,j)$ for all i and j, and we can thus proceed recursively to find $B(i,j)$ in terms of the three elements directly "behind" it,
$B(i-1,j-1), B(i-1,j)$ and $B(i,j-1)$.

Firstly note that a highest scoring alignment can end in three ways. Either
(i) X_i is paired with Y_j,
(ii) X_i is paired with a gap $-$, or
(iii) a gap $-$ is paired with Y_j.

In (i) $B(i,j)$ equals the score of the best alignment between $x_{1,i-1}$ and $y_{1,j-1}$ plus an extra term for the match between X_i and Y_j which we label $s(i,j)$, the element from the substitution matrix which corresponds to the elements X_i and Y_j (thus for DNA if $X_i = Y_j = a$ $s(i,j)$ is the element in row 1 and column 1 of the matrix). Thus we would have
$B(i,j) = B(i-1,j-1) + s(i,j)$.
In (ii) $B(i,j)$ equals the score of the best alignment between $x_{1,i-1}$ and $y_{1,j}$ plus an extra term $-d$ for the alignment of X_i with a gap, so that $B(i,j) = B(i,j-1) - d$.
In (iii) similar reasoning gives
$B(i,j) = B(i-1,j) - d$.

The best alignment is of course the best of these three possibilities, and so

$$B(i,j) = max\{B(i-1,j-1) + s(i,j), B(i-1,j) - d, B(i,j-1) - d\}$$

Using this method we obtain the value of $B(m,n)$. Thus we can find the best possible score with certainty.

The problem is that the running time for this algorithm is of the order mn, so that for very large sequences the time involved can be unrealistically large.

We shall conclude this section (and the course) with a simple example of how the process works.

EXAMPLE

Example: Let $x = gaatct$, $y = catt$ so that $m = 6$ and $n = 4$. We use the simple scoring system of +1 for a match, -1 for a mismatch (the matrix S is 4×4 with every element on the leading diagonal 1, and all others -1) and $d = -2$.

the application of the Needleman Wunsch algorithm is shown in the following figure. Arrows show where each element comes from (in each case there are three possibilities).

FIGURE 20.1

Following the bold arrows gives a highest-scoring alignment, shown below.

g a a t c t
c − a t − t

Note that sometimes more than one arrow comes from a cell, indicating that two possibilities give the same score. By following alternative paths, other highest-scoring alignments can be found, e.g.

g a a t c t
− c a t − t