Armidale Animal Breeding Summer Course 2011

Statistical methods and design in plant breeding and genomics

Course manual

Dr Ian Mackay
National Institute of Agricultural Botany
Cambridge, UK

Armidale Animal Breeding Summer Course 2011

Statistical methods and design in plant breeding and genomics

Course manual

Dr Ian Mackay
National Institute of Agricultural Botany
Cambridge, UK

Organized by the University of New England, Armidale, Australia

7-11 February 2011

ACKNOWLEDGEMENTS

This course is based on material I have previously presented in Cambridge to students who attended the NIAB course on Quantitative Genetics in Plant Breeding. The Cambridge course would not have been possible without the help, support and direct input of several friends and colleagues at NIAB. I remain grateful to their support over the last four years. Among these is my boss, Professor Andy Greenland, who deserves a special mention for allowing me to bunk off to Australia for two weeks.

That I was invited to give this course is a result of renewed contact with my old friend Professor John Gibson - I always knew he would get on - and it is pleasing to see that at least one of us has made something of himself. I am also grateful for the help and support of Professors Alan Kinghorn and Julius van der Werf.

Finally, thanks in advance to the attendees of Module 1 of the 2011 Armidale Animal Breeding Summer Course for turning up, and to my students of previous courses, not only for turning up, but surviving to provide feedback which I hope results in this course being a success.

Ian Mackay
National Institute of Agricultural Botany
Cambridge
UK

INTRODUCTION

These notes are to accompany module 1 of the 2011 Armidale Animal Breeding Summer Course: Statistical methods and design in plant breeding and genomics, to be held at the University of New England, Armidale in February March 2011.

Plant breeders are required to know a whole pile of stuff in an array of disciplines ranging from the macro to the micro: agronomy to molecular biology. Unfortunately, with the expansion of molecular genetics and genomics, opportunities for training in some these have reduced as universities and research institutes concentrate on the more popular new *omics* based disciplines.

This loss of training is particularly apparent for statistics and quantitative genetics: never popular subjects among biologists. This is unfortunate as the newer disciplines can only deliver improved varieties if integrated into practical plant breeding programmes. Such integration is most efficiently achieved through application of quantitative genetics and statistics. Meanwhile, at the heart of all breeding remain such traditional pursuits as designing and analysing yield trials to rank varieties in order of merit. Therefore, at a time when more than ever, plant breeding needs staff trained in quantitative methods, this training has all but vanished.

This course is an attempt to help stop the rot. I hope first to supply the statistical and genetical background to quantitative methods, old and modern, which is relevant to plant breeding. Secondly, I shall provide hands-on experience of these analyses. It is too much to hope that you actually enjoy the course, but I hope will find at least some of it useful in the future. Computers have democratised statistics and data handling. No longer do you need a PhD in Applied Mathematics to understand and implement many of the sophisticated methods of analysis available: they are within the grasp of all of us, as you shall see.

SOME MATHEMATICS FOR PLANT BREEDERS

Recommended books

Wikipedia is generally very good on mathematics and statistics.
CatchUp Maths & Stats for the life and medical sciences, Harris, Taylor & Taylor (£13.39 from Amazon), starts off at a very basic level and covers a lot.

Introduction

You won't need to know all that is in this section for this course, but I thought I'd start off by giving some formulae and results which will cover everything you'll need in the next two weeks (and more). I haven't given rigorous definitions or proofs, most of which are beyond me anyway, but I hope the content will enable you to read many mathematical and statistical formulae without quite such a feeling of exclusion or bewilderment at the black arts of mathematics and statistics.

To start with something easy:

Integers	These are whole numbers - 1, 2, 3…. and -1, -2, etc.
Rational numbers	Ratios (i.e. fractions) of numbers.
Irrational numbers	Numbers which cannot be expressed as rational numbers, although they can be approximated by them to any desired degree of accuracy. The classic examples are $\sqrt{2}$ and π.
Real numbers	Continuous. Include the above three classes.

Simple mathematical rules. These are well known and taken for granted, for example:

$3 \times 2 = 2 \times 3$

or algebraically:

$AB = BA$

but this is not the case for some other systems of calculation (to come).

Algebra

We let symbols and letters stand for numbers. Often the symbols can stand for any possible number, but sometimes we constrain the range of possible numbers to lie in some interval, for example between 0 and 1 or to include positive integers only.

Powers

a^x means a times a times a times a x times. eg 4^3 is 4 x 4 x 4 = 64.

In this expression, a is called the base and x is the exponent.

$a^{1/x}$ means the xth root of a. eg $64^{1/3}$ = 4 as 4 x 4 x 4 = 64

$1/a^x$ means $1/a$ times $1/a$ times $1/a$ x times. eg $1/4^2$ is ¼ x ¼ = 1/16

$1/a^x$ can also be written as a^{-x}.

As a consequence, a^x times $1/a^y$ is a^x times a^{-y} which can also be written a^{x-y}

For example 4^2 x 4^{-3} is 4^{2-3} is 4^{-1} = ¼. (16/64 = 1/4)

Thus, you can add and subtract the exponents to simplify the expression.

If $x = y$, $a^{x-y} = a^0 = a^x/a^y = 1$. Therefore

$a^0 = 1$

$a^1 = a$

There is no meaning that can be given to a number raised to the power of an irrational number. So for example:

9^π has no meaning, but can be approximated by $9^{22/7}$

There is no straightforward meaning which can be given to $(-a)^{1/2}$
since, for example, -3 x -3 = 9 and 3 x 3 = 9, so it isn't clear what $(-9)^{1/2}$
means. However, $(-a)^{1/3}$ is straight forward. For example -3 x -3 x -3 = -27.

Functions:

A function is a mathematical expression relating one or more quantities. So

$y = x^2$ is a function.

Functions can be continuous or discontinuous, that is to say the output, y, in the function could take any value or could be restricted to specific values, 0 and 1 only, or positive integers only. It is usually obvious what sort of a function we are dealing with.

A function of y is generally written as $y = f(x)$: the variable y is a function (as yet undefined) of the variable x.

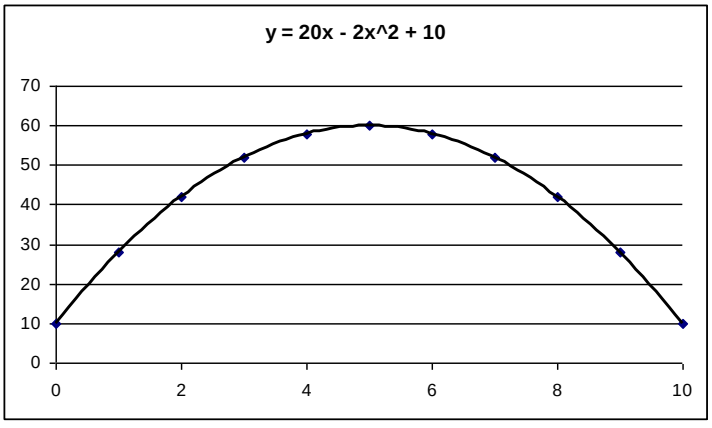
Here are some common functions you will encounter in this course:

Linear $y = b(x) + c$ or more commonly $y = bx + c$

The standard formula for a straight line; b is the gradient and c is the intercept – the value of y when x has the value zero. The point (y=0, x=0) is known as the origin.

Quadratic $y = b_1 x + b_2 x^2 + c$

This is often the first port of call in regression when we are attempting to fit a curve to data rather than a straight line and when we have no specific theoretical function we want to fit. Note that in the example below there is a maximum value of y (the peak). Depending on the parameter values (the values of b_1 , b_2 and c) there can be a minimum instead. Note too that parts of the line are quite curved and other parts are nearly straight. When fitted to real data, we generally use only a small portion of the possible range of x values so we can accommodate varying degrees of curviness. However, if we ever extrapolate beyond the range of our known x values (that is we predict y for values of x outside the range of our observed data) we need to be very careful – we may cross over the maximum (or minimum) and find values of y decreasing with increasing values of x, whereas biologically we might expect y to continue to increase. Extrapolation is always dangerous, even when the observed relationship is a straight line.



This form of function can still be viewed as a linear, in the sense that y is a linear function of the predictor variables. In this case we have two predictors x and x^2. The fact that one of our predictor variables is the square of the other doesn't matter.

Factorial $y = x!$

$x!$ is mathematical shorthand for $x(x-1)(x-2)(x-3)\ldots. 1$

x must be a positive integer. 0! is defined as 1

Factorials are discontinuous functions – the results are integers. There is an equivalent continuous function, the gamma function, written as $y = \Gamma x$ but we don't need it for this course.

Exponential. $y = e^x$

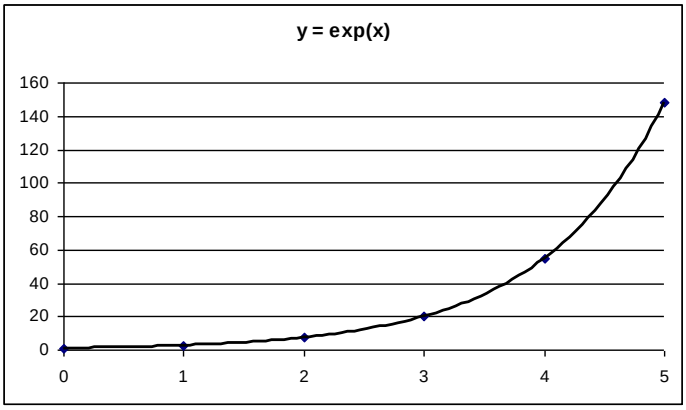
e is the mathematical constant, 2.718… e^x is often written as exp(*x*).

The exponential function, e^x is defined as :

$e^x = 1 + x + x^2/2! + x^3/3! + \ldots$

Then $x = 1$, this gives,

$e = 1/0! + 1/1! + 1/2! + 1/3! + \ldots = 1 + 1 + 1/2 + 1/6 + 1/24 + \ldots \sim 2.718$



The reason for the mathematical obsession with *e* will be explained later. Here we just note that

$y = e^x$

can be expressed as a more mundane power function

$y = a^z$

by searching for the appropriate values of *a* and *z*.

A more general form of the curve is

$y = ae^{bx}$

where choice of *a* and *b* provides curves which fit many forms of growth and decay in biology. Note that this function is non-linear: *y* is not a linear function of the predictor variable. However, it can be transformed to a linear scale:

$$f(y) = bx + f(a)$$

by use of the appropriate function for *y* and *a*. As *a* is a constant, so is *f(a).* Therefore *f(y)* is the formula for a straight line. The required function is the logarithmic, which comes next…

Logarithms. $y = log_b(x)$

Logarithms are all defined in terms of a base (subscript *b* above) commonly 10, 2, or *e*. The logarithm of a number is the power to which the base must be raised to get the number. That is to say, if

$$y = log_b(x)$$

then

$$b^y = x$$

Before the days of calculators, logarithms (logs) were important as a means of calculating the product of large numbers with reasonable accuracy quite quickly using slide rules or log tables. This use has vanished, but they remain important because of their role as a kind of inverse of power functions. Also, in statistical analysis, some characters with which we deal tend to be more amenable to analysis and manipulation if we work on the logs of the measurements rather than the original measurements themselves.

The logarithm (log) of a number is the power to which the base must be raised to get the number. Logs to the base ten have an easy interpretation and are often used as the scale of measurement of strength of evidence for genetic linkage. They are also the scale of measurement for pH.

$\log_{10}(x) = y$ means that $10^y = x$

So:

x	$\log_{10}(x)$	
0	undefined	there is no power of 10 which is zero.
1	0	since 10^0 is 1
10	1	since 10^1 is 10
100	2	since 10^2 is 100
1000	3	since 10^3 is 1000

also:

x	$\log_{10}(x)$	
0.1	-1	since 10^{-1} is 0.1
0.01	-2	since 10^{-2} is 0.01
0.001	-3	since 10^{-3} is 0.001

finally:

-100	undefined	since you cannot have the square root of a minus number.

In fact, no minus number has a corresponding logarithm.

Note 100 x 1000 = 100,000
2 + 3 = 5

Multiplication on the logarithmic scale is mirrored by addition of the corresponding logs (The inverse $\log_{10}$ of 5 is 100,000.) Inverse logs are described as antilogs. This useful property means that if:

$z = x^y$

then

log(z) = ylog(x).

This is true for logarithms to any base.

Logarithms to the base 10 are easy to understand, but are not the most commonly used. When mathematicians and statisticians refer to logs they generally mean logs to the base e, referred to as "natural" logs.

Natural logarithms are sometimes abbreviated to "logs" and sometimes to "ln."

Converting $\log_{10}$ to $\log_e$

Sometimes we must convert logs from base 10 to base *e* or and vice versa. For example, in linkage analysis both log of odds (LODs - base 10) and log likelihood ratios (LRT - base *e*) are used to measure the strength of a signal. There will be more on this when we discuss linkage. These conversions are carried out as follows:

$e^x = 10^y = z$ ie x is ln(z) and y is $\log_{10}(z)$

So $\ln(z) = \ln(10^y) = y.\ln(10)$

and

$\log_{10}(z) = \log_{10}(e^x) = x.\ \log_{10}(e)$

For example if z = 20

ln(20) = 2.996
$\log_{10}(20)$ = 1.301
ln(10) = 2.303
2.303 x 1.301 = 2.996

and

$\log_{10}(20)$ = 1.301
ln(20) = 2.996
$\log_{10}(e)$ = 0.434
2.996 x 0.434 = 1.300

This also explains how e^x can also be expressed as a^y

Transformation of exponential to linear functions.

For the exponential function:

$y = ae^{bx}$

then

$\ln(y) = \ln(a)+bx\ln(e)$

$= \ln(a) + bx$ since $\ln(e) = 1$

So taking logs of an exponential or power function gives a linear relationship on the log scale.

Binomial $(p+q)^n = p^n + np^{n-1}q + n(n-1)/2!\ p^{n-2}q^2 \ldots$

If $p + q = 1$ this gives the binomial probability distribution. The coefficients of succeeding terms in the binomial can be found by Pascal's triangle:

n	coefficients
1	1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1

The first term and last terms are 1 and the other terms are the sum of the pair of corresponding two terms on the line above. So for example:

$(p+q)^4$

coefficients	1	4	6	4	1
exponent of p	4	3	2	1	0
exponent of 1	0	1	2	3	4

Putting all this together gives:

$$1p^4p^0 + 4p^3q^1 + 6p^2q^2 + 4p^1q^3 + 1q^0q^4$$

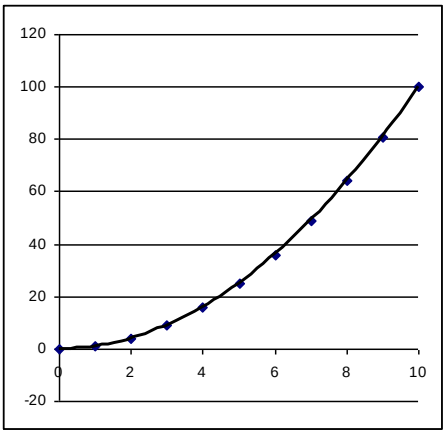
or

$$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$$

This provides an easy way of writing down genotype frequencies under Hardy-Weinberg equilibrium for polyploid species, which will be discussed in the section on Population Genetics.

Differential calculus

Differentiation is the process of finding how a function changes when its input values change. It is a means of studying rates of change. Consider the graph of $y=x^2$ below.



What is the rate of change at any point? That is to say, what is the gradient of the graph? Around $x = 0$, there is very little change in y as x changes, so we would expect the gradient to be close to zero. However, as x increases, the slope gets steeper and steeper.

If the graph is a straight line rather than a curve, the rate of change is constant over the whole range and can be found as:

$(y_2-y_1)/(x_2-x_1)$ which gives the value of b in $y= bx +c$

For more complex functions, although the gradient is not constant over the whole range, we can find the gradient at any particular point [x,y] by considering the most miniscule change in values to x (and therefore to y). This miniscule change can be written as a change from $[x,y]$ to $[x+\delta x, y+ \delta y]$. Then the gradient can be approximated by

$[(y+ \delta y)-y] / [(x+\delta x)-x] = \delta y/\delta x$

For example, take the function $y=x^2$.

Then as x moves to $x+\delta x$,

y moves from x^2 to $(x+\delta x)^2$

so the change in y is $x^2+2x\,\delta x + \delta x^2 - x^2 = 2x\delta x + \delta x^2$

The gradient is therefore $2x\,\delta x + \delta x^2 / \delta x = 2x + \delta x$

As δx gets smaller and smaller it effectively disappears ("tends to zero"), at which point the gradient is $2x$. Formally we say that

$\delta y / \delta x \sim 2x$

and as x tends to zero

$dy/dx = 2x$

Note the subtle difference in symbols δ and d. The differential of a function $f(x)$is some times written as $f\,'(x)$. Differentiation for a second time, is written as d^2y/d^2x or $f\,''(x)$.

Similar reasoning will give the differential (ie the gradient) for other functions. However, for most standard functions it is easier to look them up (eg in Wikipedia) or to remember some standard results:

Some standard differentials

$d(c)dx$	=	0	the differential of a constant is zero
$d(x^n)dx$	=	nxa^{n-1}	
$d(k+cx^n)dx$	=	nca^{n-1}	this is the one to remember
$d(e^x)/dx$	=	e^x	

That $d(e^x)/dx = e^x$ is one of the reasons that mathematicians have such an enthusiasm for the exponential function. It follows directly from the definition of *e*:

$de^x/dx = d(e^x = 1 + x + x^2/2! + x^3/3! + \ldots)/dx$

The differential of each term in the series is identical to the term to its left, apart from the first term which vanishes. As the series is infinite, differentiation leaves the whole function unaltered.

Not all functions are this simple to differentiate, but differentiation can often be accomplished by stringing together these results with two additional rules:

Differentiation of a product

If $y = uv$ and u and v are themselves functions of x,

Then

$dy/dx = udv dx + vdu dx$

eg

$y = x\ln(x)$

$dy/dx = x(1/x)+\ln(x).1 = 1+\ln(x)$

The trick is deciding how to split the function up into a *u* and a *v*.

Differentiation of a function of a function

If $y = f(x)$ can be viewed as a function of a function: $y = f(u)$ and $u=f(x)$, then

$dy/dx = dy/du \,.\, du/dx$

eg

$y = \log(x^3)$

Set $u = x^3$

$dy/du = 1/u = 1/ x^3$

$du/dx = 3x^2$

So $dy/dx = 3x^2 / x^3 = 3/x$

Differentiation to fit models to data.

An example of differentiation which is frequently encountered is in the estimation of the best set of parameters to explain observed data. For example, the data could be yield and the parameters could be variety performances. Any set of parameters, however obtained, can be used to produce predicted, or fitted, values of what the data would have been if those parameters provided a perfect explanation of the data. The best estimate of the parameters is then that which gives the closest fit to the observed values of the data.

If we have a function which measures the goodness of fit, then by varying the parameter values, the minimum or maximum value of the goodness of fit function (depending on the nature of the function) will give us our best set of parameter estimates. For any function $y = f(x)$, at the point at which y rises to a maximum, or falls to a minimum, the gradient is zero: y is neither increasing nor decreasing as x changes (infinitesimally) at this point. So the best estimate of our parameters is given by the value of the parameters at which the differential of our goodness of fit function, with respect to the parameters, is zero. More on this later.

Integration

Integration is the opposite of differentiation. That is integration is the *antiderivative* of differentiation. There is a slight complication:

$y = bx + 3$

$dy/dx = b$ There is a unique answer.

But $\int b\, dx = bx +$ any constant There are an infinite number of answers.

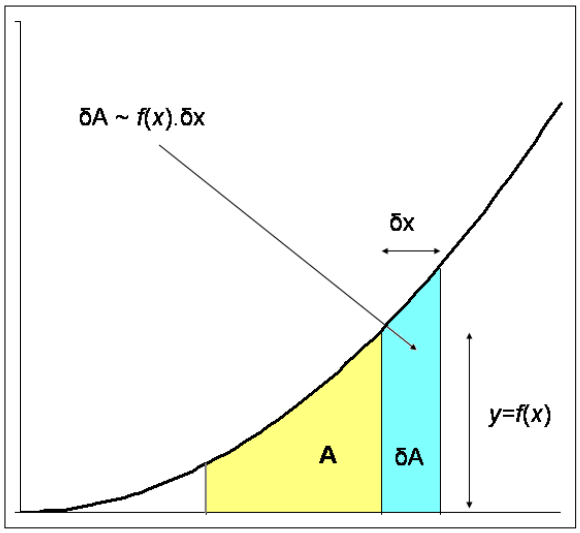
This is usually not a problem because the constant can often be found in other ways. It is always a problem if you forget about it, however.

(Note the symbol $\int$ to denote integration.)

Integration can also be viewed as the area under the curve: this is the way in which it is usually encountered. This sounds sensible but can be hard to see. Try this:

Consider the function plotted below

$y = f(x)$.



Let the area under the curve between two values of x be A. Then let's increase the area a tad (δA) Provided that δA is small enough, this increase can be approximated by adding a rectangle of size $f(x)\ \delta x$: a thin sliver of height $f(x)$ and width δx.

So $\delta A \sim f(x)\ \delta x$

Therefore $\delta A / \delta x \sim f(x)$

So, as for differentiation, as $\delta x \rightarrow 0$

$$dA/dx = f(x)$$

In other words, if we know the function which gives the area under some curve, differentiation of that function for the area gives the function for the curve. Therefore the *antiderivate*, or integral of a curve gives the area underneath the curve.

$$\int f(x)\ dx = A$$

Some standard results for integration are given below:

$$\int x^n\ dx = x^{(n+1)}/(n+1) + C \quad \text{except for } n = -1$$

$$\int x^{-1}\,dx = \int 1/x\,dx = \ln(x) + C$$

$$\int e^{x}\,dx = e^{x} + C$$

In practice, when we integrate, we usually want to find the area between some values of x (though these can sometimes range from $+\infty$ to $-\infty$).

For example suppose we want to find the area under the curve

$$y = x^2$$

between $x = 1$ and $x = 10$.

We write this as

$$\int_1^{10} x^2\,dx = \left[x^3/3 + c \right]_1^{10}$$

$$= (1000/3 + c) - (1/3 + c)$$

$$= 333$$

Note that the constant C cancels out when we are working with definite integrals.

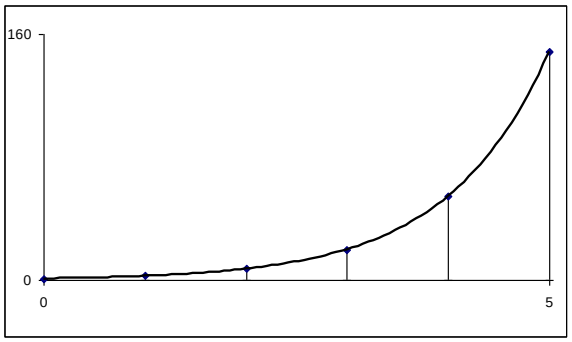
This business with areas is related to the *fundamental theorem of calculus*: "The integral of a continuous function always exists and integration is the inverse of differentiation."

For integration of more complex functions there are additional methods equivalent to the rules for differentiation of complex functions which we won't go into. In fact, most functions can't be integrated algebraically, but we can still calculate the areas numerically.

Numerical integration

At one extreme, numerical integration amounts to plotting the curve on graph paper and counting the squares under the curve. This can be quite accurate. Better is to use the *trapezoid rule*:

Split the area under the curve into n vertical strips - calculate the area of each vertical strip as average height times the width and add the areas.



This method has the advantage of simplicity – you can implement it in Excel for example. It is also clear that the greater the number of strips, the greater the accuracy. Methods are available to evaluate its possible error but we shall not go into those.

There are other methods. A popular one is Simpson's rule. Rather than splitting the area under the curve into multiple trapezoids, this considers three y values at a time, fits a quadratic curve to the three points and calculates the area under the curve by integration. Adjacent sets of three are then added up. This can also be managed in Excel and is more accurate that the trapezoid rule since it approximates the shape of the function with multiple curves rather than multiple straight lines.

Taylor's series

This is difficult; don't worry if you don't understand it. We are more interested in the results that the method can provide. Some of these are given at the end of this section.

Taylor's series approximates other functions as of "polynomials of infinite degree." You are unlikely to have to use it yourself but you will come across it and use its results. Its utility lies in that, although infinite, usually all but the first two or three terms can be ignored. Without proof, for the function $f(x) = f(a+h)$ (ie evaluate the function for a value of $x = a+h$.

$$f(a+h) \equiv f(a) + f'(a)h + f''(a)h^2/2! + f'''(a)h^3/3! \ldots .$$

f' and f'' etc stand for the first differential, the second differential and so on. The second differential is got by differentiating the function with respect to x twice. eg :

$$f''(x^3)/\mathrm{d}x = 6x.$$

You will also come across Maclaurin's series which is just a special case of Taylor's with 'a' set to zero

$$f(x) \equiv f(0) + f'(0)x + f''(0)x^2/2! + f'''(0)x^3/3! \ldots$$

Some examples:

Instead of defining *e* as the series given earlier, we could have defined the exponential function $y = e^x$ as the function for which $dy/dx = y$ – ie differentiation gives us the same function.

So $f(x) = f'(x) = f''(x)$ etc.

$f(0) = e^0 = 1$ since the value of any number raised to the power zero is 1.

So using Maclaurin's theorem we get:

$e^x = 1 + x + x^2/2! + x^3/3! + \ldots$ as before.

Some other useful approximations to come out of Taylor's series which we may call upon are:

$(1+x)^{-1} = 1 - x + x^2 - x^3 + x^4 \ldots$ Converges if $-1 < x < 1$

$\ln(1+x) = x - x^2/2 + x^3/3 - x^4/4 \ldots$. Converges if $-1 < x \leq 1$

In the last example, when x is small, $\ln(1+x) \sim x$

eg $x = 0.1$, $\ln(1.1) = 0.095$

Matrix algebra

Basics

Matrix notation is a means of writing down a compact summary of expressions and equations which would otherwise take up a lot of space. There are then rules for handling matrices algebraically. These provide easy ways of solving simultaneous equations, among other things. To manipulate matrices arithmetically by hand is steady work. However, they are readily handled by computer (and also within Excel for modest sized matrices). So in the treatment given here, I shall assume all the hard work will be done by a computer and just outline the notation and manipulations that are carried out.

Suppose we have a set of variables - a_1, a_2, a_3, a_4

We call $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$ a column vector. Vectors and matrices are usually written in bold.

a' = $\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix}$ is a row vector.

The **'** symbol means transpose the matrix – swap row and column positions.

A matrix with r rows and c columns has individual elements a_{ij}
and is referred to as

A = $[a_{ij}] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ Note the bold **A** to stand for the whole matrix

i indexes the rows and j the columns. That is, rows first, then columns.

If $a_{ij} = a_{ji}$ for all i and j – the matrix is symmetrical – the rows and columns can be interchanged.

In this case, **A'** = **A**

Matrices can be added if they have the same number of rows and columns. All that happens is that the corresponding elements of the two matrices are added.

Matrices can be multiplied if the number of columns of the first matrix is equal to the number of rows of the second matrix. The result is a matrix with row number equal to the number of rows in the first matrix and column number equal to the number of columns in the second matrix. The multiplication operation is complicated. The product of the elements in row i of the first matrix and of column i of the second matrix are added to produce element ij of the product. eg

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} -3 & 1 & 2 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 9 & 6 \\ -1 & 12 & 9 \\ -3 & 15 & 12 \end{bmatrix}$$

In matrix form this would be written as

AB = **C**

Beware, **AB** ≠ **BA**

Matrices can be multiplied by single numbers (called scalars). Here all elements of the matrix are multiplied by the same number. To distinguish them from matrices, scalars are not written in bold. Multiplication of a matrix by a scalar is the same as multiplying by a square matrix in which all elements are zero except those on the diagonal where every element has the value of the scalar. Row and column numbers must still conform to the

rules for matrix multiplication however. If the scalar is 1, then such a matrix is called the identity matrix, usually denoted by the letter **I:**

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}$$

If we have two square matrices such that

AB = **I**

Then **A** is said to be the inverse of **B** and vice versa. This is written as $\mathbf{A}^{-1} = \mathbf{B}$. In this case, **AB** = **BA**.

For square matrices of two or three rows and columns, the inverse can be calculated by hand. Larger matrices require a computer.

Some other useful results

ABC = **(AB)C** = **A(BC)**

(AB)' = **A'B'**

A(B+C) = **AB** + **AC**

Calculus for matrices

Results are given without proof – try a couple of simple numerical examples if you wish.

Suppose we have an equation:

$$y = x_1 + 2x_2 + 3x_3$$

The three variables x_1, x_2, x_3 could represent numbers of apples, oranges and pairs: anything.

In matrix form:

$$y = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

or

$y = \mathbf{a'x} = \mathbf{x'a}$ (as the answer is a scalar)

with $\mathbf{a'} = [1\ \ 2\ \ 3]$

If we were to differentiate y with respect to each variable separately:

$dy/dx_1 = 1$

$dy/dx_2 = 2$

$dy/dx_3 = 3$

In matrix form this is:

$dy/d\mathbf{x} = d(\mathbf{a'x})/d\mathbf{x}$

$= d(\mathbf{x'a})dx = [1\ 2\ 3] = \mathbf{a'}$

This is the matrix equivalent of the d(kx)/dx = k.

Suppose now there are multiple values of y, each with a different value of **a**.

For example:

$y_1 = x_1 + 2x_2 + 3x_3$
$y_2 = 2x_1 + 4x_2 + 5x_3$

or

$\mathbf{y} = \mathbf{Ax}$

where $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \end{bmatrix}$

d**y**/d**x** is taken to mean $\begin{bmatrix} \frac{dy_1}{dx_1} & \frac{dy_2}{dx_1} \\ \frac{dy_1}{dx_2} & \frac{dy_2}{dx_2} \\ \frac{dy_1}{dx_3} & \frac{dy_2}{dx_3} \end{bmatrix}$

As a result, d**y**/d**x** = d(**Ax**)/d**x** = **A**'

Note the result is **A**' and not **A** as one might expect. Fortunately, in the form in which differentiation of matrices if often met in statistics, **A** is often symmetrical, so this makes no difference.

Some other standard results for matrix differentiation are given below:

y = *f*(***x***) = k **y** is a constant whatever the values of **x**:

d*f*(***x***)/d(***x***) = **0**

This is the matrix equivalent of the differential of a constant being zero

y = *f*(***x***) = **a**'***x***

d*f*(**x**)/d(**x**) = d(**a'x**)/d**x** = d(**ax**')/d**x** = **a**

This is the matrix equivalent of the d(b*x*)/ b*x* = k, where k is a constant, as we have just seen.

y = *f*(**x**) = **x**'**Ax** where **A** is symmetrical, then

d*f*(**x**)/d(**x**) = d(**x**'**Ax**)d**x** = 2**xA**

This is the matrix equivalent of $d(x^2)dx = 2x$. We require this result in estimating parameters by least-squares as we shall see later.

Use of matrix notation and matrix algebra to solve simultaneous equations:

Suppose we have a set of equations:

$y_1 = b_1x_{11}+b_2x_{12}+b_3x_{13}$
$y_2 = b_1x_{21}+b_2x_{22}+b_3x_{23}$
$y_3 = b_1x_{31}+b_2x_{32}+b_3x_{33}$

The vector **y** could be the expected yield at each of three farms. **X** could represent rainfall, fertilizer and sunlight at each of the three farms. $[b_1\ b_2\ b_3]$ = **b'** are then (regression) coefficients relating these three environmental measurements to yield. (For this example, assume these have been established elsewhere in some experiment or other.) In matrix form, the whole set of equations can be written as

$$\mathbf{y} = \mathbf{Xb}$$

If we know **X** and **y** but not **b** (we lost them somehow), we can find it as:

$$\mathbf{X}^{-1}\mathbf{y} = \mathbf{X}^{-1}\mathbf{Xb} = \mathbf{Ib} = \mathbf{b}$$

This is just like normal algebra except we have to be careful about the order in which things get multiplied or divided: $\mathbf{yX}^{-1}$ wouldn't work, for example. We'll use this procedure extensively later on. Although it may not seem like it at the moment, manipulating complex sets of equations through matrix algebra is easier and simpler than manipulating the individual equations themselves. If we expanded this example to have 300 independent variables (the x's) rather than 3, the matrix form of the equations would remain unchanged.

Eigenvalues and eigenvectors – aka characteristics roots and vectors.

These are used in a common multivariate analysis called principal component analysis (PCA). PCA is often used in genetic diversity studies and to study relationships among multiple related phenotypic traits. It is worth knowing something about how this works, even if all you ever do is take output from a standard statistical package. The description I give below is not the standard one, but it fits in directly with this common statistical use and I find it easier to understand.

Suppose we have a square symmetrical matrix **Y**. We want to approximate this matrix by a vector such than

Y ~ kk'

For example, given

$$\mathbf{Y} = \begin{bmatrix} 5 & 6 & 4 \\ 6 & 10 & 1 \\ 4 & 1 & 15 \end{bmatrix}$$

we can use **k' =** [-1.89 -1.86 -3.29] as an approximation. Then **kk'** =

$$\begin{bmatrix} 3.58 & 3.52 & 6.22 \\ 3.52 & 3.46 & 6.12 \\ 6.22 & 6.12 & 10.83 \end{bmatrix}$$

Not fantastically good, but could be worse. Suppose we try to increase the precision of the approximation as:

Y ~ kk' + jj'

Then if **j'** = [-1.05 -2.53 2.04], **jj'** =

$$\begin{bmatrix} 1.11 & 2.67 & -2.15 \\ 2.67 & 6.42 & -5.17 \\ -2.15 & -5.17 & 4.16 \end{bmatrix}$$

and **kk'** +**jj'** =

$$\begin{bmatrix} 4.69 & 6.19 & 4.07 \\ 6.19 & 9.89 & 0.96 \\ 4.07 & 0.96 & 14.98 \end{bmatrix}$$

This isn't bad.

We could add a third term:

i' = [0.56 -0.34 -0.13]

in which case we find:

Y = kk' + jj' + ii'

So the original matrix has been decomposed into three vectors. The utility of this approach is that for much larger matrices, a few vectors, often only one or two, give an adequate approximation of the whole matrix and can greatly simplify the interpretation of the data.

The vectors **k**, **j** and **i** are orthogonal – they are independent of each other – the variability in the matrix accounted for by **j** is independent of that accounted for by **k**. Algebraically this means that if we form a matrix by butting up the columns

Z = [**k**:**j**:**i**]

then

Z'Z is a diagonal matrix - all the off-diagonal elements are zero.

The vectors are not normally presented in the form given above. Each vector is scaled up or down such that the sum of squares of the elements = 1. For the example above:

k' = [-1.89 -1.86 -3.29] = sqrt(17.865) [-0.447 -0.440 -0.778]
j' = [-1.05 -2.53 2.04] = sqrt(11.690) [-0.308 -0.741 0.597]
i' = [0.56 -0.34 -0.13] = sqrt(0.445) [0.839 -0.507 -0.196]

and $(0.447^2 + 0.440^2 + 0.778^2 = 1)$ etc.

In this case **Z'Z** = **I** **:** the unit matrix

The scalars are called eigenvalues or latent roots and the scaled vectors are called eigenvectors or latent vectors. Finally, we can put the whole thing together as

Y = ZLZ'

where

Z is the matrix of eigenvectors, (scaled so the sum of squares of each column = 1)
L is the diagonal matrix of eigenvalues.

The advantage of this way of writing down the vectors is that the eigenvalues give a measure of the importance of that vector in describing the matrix and that the elements of the eigenvector give a measure, for that eigenvalue, of the importance (or "loading") of that row or column of the matrix in the approximation. In the example above, there are two large and one small eigenvalues, and we found that use of the vectors associated with the two large values gave a good approximation to the whole matrix.

This dry mathematic interpretation can frequently be given some biological meaning. For example, the matrix **Y** could be a matrix of correlation coefficients among traits measured on a set of varieties. The eigenvectors define an alternative set of uncorrelated derived traits. (For each variety or individual, each trait value is multiplied by the

corresponding element in the eigenvector. Add these up to get the derived trait value associated with that eigenvalue for that individual.) A few of these derived traits, those with the largest eigenvalues, often approximate the original correlation matrix so well that we feel justified in simplifying our analysis by working just with these. Moreover, the magnitude and sign of the elements of an eigenvector may also be interpretable, indicating for example that the new variable it is a measure of, say, early growth. This is the crux of principle component analysis.

Singular value decomposition

Singular vectors and values are the equivalent of eigenvectors and values, but the matrices are not required to be square. The singular value decomposition underpins a popular form of analysis of genotype x environment interaction termed AMMI which we shall come to in due course. You don't need to understand svd to carry out this type of analysis but if may help demystify the process.

The singular value decomposition of a rectangular matrix **X** with r rows and c columns is

$$\mathbf{X} = \mathbf{USV}'$$

S is a diagonal matrix with dimensions = minimum (rows, cols) of **X**.
U has r rows and the same number of columns as **S**
V has c rows and the same number of columns as **S**. (but nb it is **V**' in the equation).

The singular values in **S** are analogous to eigenvalues and the vectors in **U** and **V** are analogous to eigenvectors. **S, U** and **V** are ordered in descending value of the singular values in **S**. Just as the sum of the eivenvalues^2 = 1 and the off diagonals of **ZZ**' were all zero, so too here, we have

$$\mathbf{U'U} = \mathbf{I}$$
$$\mathbf{VV'} = \mathbf{I}$$

X can be written as

$$\mathbf{X} = \mathbf{U}_1 s_1 \mathbf{V}_1' + \mathbf{U}_2 s_2 \mathbf{V}_2' \ldots \mathbf{U}_p s_p \mathbf{V}_p'$$

where the subscripts refer to column vectors extracted from **U** and **V** and to each singular value in turn. This is called singular value decomposition or spectral value decomposition.

An increasingly accurate approximation to **X** is then given by inclusion of successive terms in the sum.

All this is easiest to see in an example.

$$\mathbf{X} = \begin{bmatrix} 5 & 2 & 1 \\ 3 & 7 & 6 \\ 1 & 3 & 9 \\ 2 & 5 & 2 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} -0.246 & 0.608 & 0.753 \\ -0.668 & 0.186 & -0.332 \\ -0.605 & -0.675 & 0.333 \\ -0.358 & 0.374 & -0.460 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 14.353 & 0 & 0 \\ 0 & 5.661 & 0 \\ 0 & 0 & 3.155 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} -0.317 & 0.649 & 0.692 \\ -0.611 & 0.418 & -0.672 \\ -0.725 & -0.636 & 0.264 \end{bmatrix}$$

$$\mathbf{U}_1 s_1 \mathbf{V}_1' = \begin{bmatrix} -0.246 \\ -0.668 \\ -0.605 \\ -0.358 \end{bmatrix} 14.353 \begin{bmatrix} -0.317 & -0.611 & -0.725 \end{bmatrix}$$

$$= \begin{bmatrix} 1.121 & 2.159 & 2.563 \\ 3.040 & 5.854 & 6.948 \\ 2.754 & 5.302 & 6.293 \\ 1.631 & 3.141 & 3.728 \end{bmatrix}$$

Not a spectacularly good approximation but we can see that it's trying its best. However

$$\mathbf{U}_1 s_1 \mathbf{V}_1' + \mathbf{U}_2 s_2 \mathbf{V}_2' = \begin{bmatrix} 3.356 & 3.598 & 0.373 \\ 3.725 & 6.295 & 6.277 \\ 0.274 & 3.706 & 8.723 \\ 3.000 & 4.024 & 2.383 \end{bmatrix}$$

is better.

The column vectors in **U** can be viewed, for each singular value in turn, as showing the importance, for that singular value, of each row in the original matrix. The column vectors in **V** can be regarded as the loadings, for that singular value, of each column of the original matrix. In fact there is a close relationship between singular values and vectors and the eigenvalues and vectors of some closely related square matrices:

Using our rules for matrix algebra

XX' = [**USV**'][**USV**']'

= [**USV'**][(**US**)**V'**]'

= **[USV'][V(US)'**]

= **[USV'][VS'U']**

= **USIS'U'** since **VV' = I**

= **USS'U'**

This is just a decomposition of **XX**' into a matrix of eigenvectors **U** and a diagonal matrix of eigenvalues **SS**' (also a diagonal matrix).

Working the other way around we get

X'X = **VSS'V'**

So **V** is also the set of eigenvectors of **XX**' and **U** is the set of eigenvectors of **X'X**, the eigenvalues of **X'X** and **XX**' are identical and are given by **SS**.

All this is arid and complicated, but as stated earlier, we can use these results in the analysis of genotype x environment interaction (among other applications), in which rows will be varieties, columns sites, and the elements of the matrix are the g x e terms for each variety-site combination. Just as with principle component analysis, we are searching for simple patterns among varieties and among sites which may indicate some hidden structure and explanation of genotype x environment interaction.

INTRODUCTION TO STATISTICS

The subject of statistics concerns the collection, summary and analysis of data. If all plants and all plots of a variety always produced the same yield, no matter where they were grown or in what season, then the yield of one plant grown on one occasion would suffice to assess yield potential of the variety for evermore.

There is however, in all biological material, an inherent variability which cannot be attributed to any specific cause; so called 'chance' variation. This variability results in plant-to-plant differences in yield even under controlled environment conditions, and under normal farming conditions plant responses to environment also vary. Intuitive estimates of yield and comparisons of one variety or treatment with another are based on experience, i.e. on an accumulation of evidence. The precision of such estimates and confidence in them is linked with the level of consistency or repeatability of results. This existence of chance variation necessitates replication in order to provide evidence of the level of variation against which to draw conclusions.

Before considering ways of measuring variation and assessing the confidence that may be placed in experimental results, we will study the types of data and distributions that may be encountered.

1 Types of data

As scientists we need to examine many different types of data in the course of our work.

Some of the data (Examples 1 and 2) are not given in a numerical form and we may need to convert it into numbers before it can be processed.

Example 1 Has the seed germinated? Write: 1 for YES 0 for NO

Then the data become: 1 1 0 1 0 0 1 0 0 0 1 1

Example 2 Give ranks to the different letters (best = 1, worse = 6). Then data become:

	A	B	C	D	E	F
Taster 1	1	5	2	3	4	6
Taster 2	2	1	4	3	6	5

We must be careful to distinguish between numbers and numerals.

Numbers may be divided into 3 types:

i. Nominal - e.g. seed lot number, plot number.

ii. Ordinal - representing a position, e.g. colour rating, harvest date.

iii. Interval and ratio scales - quantities that may be added and subtracted (and for ratio scales - multiplied and divided) e.g. temperatures in $^{\circ}$C, heights above the ground in cm, dry matter yields in tonnes/ha.

The best data are measurements on an interval and ratio scale because when communicating the results, a much more precise description may be given.

When recording data it is important to record it to the correct accuracy. Data quality may always be downgraded but it is impossible to upgrade data. Data may be converted from interval to ordinal scale but in doing so information is thrown away.

E.g.	measurements	2.6	3.8	6.0
	ordinal scale	1	2	3

With the ordinal scale the difference between 2.6 and 3.8 (1.2) is given the same weight as the difference between 3.8 - 6.0 (2.2)

Data recorded on a nominal/ordinal scale may be collected in such a way that they may be used as interval scale data e.g. data from 10 plants scored for presence or absence of disease may be converted to an interval scale by calculating % of plants infected.

Type (iii) data (i.e. interval and ratio scale) is that which we most commonly need to process statistically. This may be subdivided into 2 types:

i. continuous

ii. discrete

Continuous data is that which can take any value within a given interval e.g. the diameter of a sweetcorn cob would be expected to be somewhere between 4.0cm and 5.5cm. No value in this range could be excluded. There are an infinite number of possible values that the diameter can take within this range. Discrete data may take only distinct discrete values e.g. the number of weed seeds in a sample of 50 seeds must be a whole number, and is therefore discrete.

NB Data that have been rounded to the nearest whole number may at first sight appear to be discrete, when in fact they are continuous.

2 Summarising Data

For large quantities of data it is helpful to summarise it graphically or by presenting some measures which describe some attribute of the data. This can be illustrated in the example below

Example: The following data represent moisture content percentages observed in different Winter Barley varieties in the same trial in 1997.

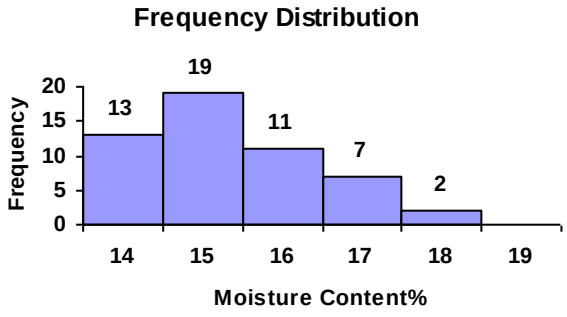
Original data values:

15.40	14.80	17.20	16.60	17.10	17.00
16.00	16.30	14.90	15.30	17.60	16.50
15.30	17.00	16.80	18.10	18.30	17.90
15.60	16.70	16.00	15.20	15.00	16.40
15.40	15.20	14.70	14.20	15.30	14.20
16.00	15.60	14.20	15.30	14.90	14.50
14.30	15.20	15.10	14.60	14.30	14.60
15.50	15.20	17.00	15.40	15.70	14.60
15.40	16.00	15.20	16.10		

Summary of data in groups:

Range of values	frequency
14.00 - 14.90	13
15.00 - 15.90	19
16.00 - 16.90	11
17.00 – 17.90	7
18.00 - 18.90	2

Graph of data as Histogram (or bar chart):



Measures of central tendency or location

a) Mode - this is the value or range which contains the most values.

(15.00 -15.90 in the above example).

b) Median - This is the middle value when all values are ranked in order of magnitude. In the above example there are 52 values. The middle value will lie between the 26th and 27th values when they are arranged in order. This value is 15.40.

c) Mean - this is obtained by summing all the values and dividing the total by the number of values.

$$\text{mean} = \bar{x} = \frac{\sum_i x_i}{n}$$

In above example

$$\bar{x} = 15.71$$

Measures of spread or dispersion

a) Range - this is the difference between the highest and lowest values.

In the above example the Range = 18.30 - 14.20 = 4.10

b) Semi-interquartile range - this is the difference between the values which define the upper and lower quarters of the data when all values are ranked.

The lower quarter is delineated by the 13th value = 14.90
The upper quarter is delineated by the 39th value = 16.40

Hence the semi-interquartile range = ½(16.40 - 14.90) = 0.75

c) Variance - this is a measure of dispersion about the mean and is estimated from

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = \frac{1}{(n-1)}\left\{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}\right\}$$

n-1 is referred to as the number of degrees of freedom ie. the number of independent data values required for the estimate. The disadvantage of the variance is that the units are not on the same scale as the original data.

The standard deviation s is the square root of the sample variance.

For the above example s = 1.060.

d) Coefficient of variation
In practice for many variables the variability increases with the mean (eg higher yields may mean higher standard deviations). The coefficient of variation expresses the standard deviation as a percentage of the mean.

$$CV = \frac{s}{\bar{x}} \times 100$$

3 Populations and samples

The total aggregate of observations that might occur as the result of performing a particular operation in a particular way is referred to as the population of observations, e.g. all crops of wheat in East Anglia in 1999 define a population of yields. In theory the population can be thought of, as an infinite number of observations but in practice is made up of N observations where N is very large.

In statistical experimentation it is normally impracticable to record the entire population so estimates would be based on a sample of the population. In statistical terms we observe the sample but wish to apply the conclusions to a population. It is important that the sample is collected in a way that achieves this objective.

Population - the total aggregate of all observations which might occur as the result of performing a particular operation in a particular way.

Sample - the actual observations taken which are usually a small proportion of the population.

The objective of statistical inference is to draw conclusions about the population based on the results from the sample. It is important that samples are randomly selected from the population of interest.

4 Distributions

Certain families of distributions are particularly useful.

4.1 Discrete distributions

A discrete distribution consists of a set of possible values together with associated probabilities of occurrence for each of these values.

a) Binomial

A binomial distribution typically arises when we are interested in the number of members of a randomly selected group of individuals that possess a certain characteristic eg.

- the number of heads, r, obtained in 10 tosses of a coin.
- a certain type of seed contains 1% off-types. Provided that the off-types are randomly distributed then the number of off-types in samples of size 100 seeds follows a binomial distribution.

For a sample of size n with constant probability p,
the mean is np and the variance np(1-p).

The model for a Binomial distribution is:

$$P(X = r) = ncr \times p^r \times (1-p)^{n-r}, \quad r = 0,1,2,\ldots,n$$

where $0 < p < 1$, and ncr is $n!/r!(n-r)!$, $r! = 1x2x\ldots xr$, $0!=1$

b) Poisson

If events are occurring randomly then the number of events occurring in a given time or space interval follows a poisson distribution.

For a Poisson random variable, the probability that X is some value x is given by the formula

$$P(X = x) = (e^{-\mu} \times \mu^x) / x!, \quad x=0,1,2,\ldots$$

where μ is the average number of occurrences in the specified interval.

Example

The number of false fire alarms in a suburb of Houston averages 2.1 per day. Assuming that a Poisson distribution is appropriate, the probability that 4 false alarms will occur on a given day is given by

$$P(X = 4) = (e^{-2.1} \times 2.1^4) / 4! = 0.0992$$

4.2 Continuous distributions

Continuous random variables can theoretically take any value on a continuous scale eg yields

Exponential

If events are occurring randomly, then the time (or distance) between successive events follows an exponential distribution and is characterised by a probability density function:

$$P(x) = \mu e^{-\mu x}$$

Example

In a seed lot, the times between the germination of successive seeds follows an exponential distribution.

5 The Normal Distribution

The family of normal distributions is particularly useful and has the following properties.

(i) the mean, mode and median of the distribution are equal.
(ii) the distribution is symmetrical about the mean.
(iii) the frequency distribution of values can be completely defined in mathematical terms if the mean and standard deviation of the values are known.
(iv) 95% of the values lie within 1.96 standard deviations of the mean.
(v) 99.8% of the values lie within 3.09 standard deviations of the mean.

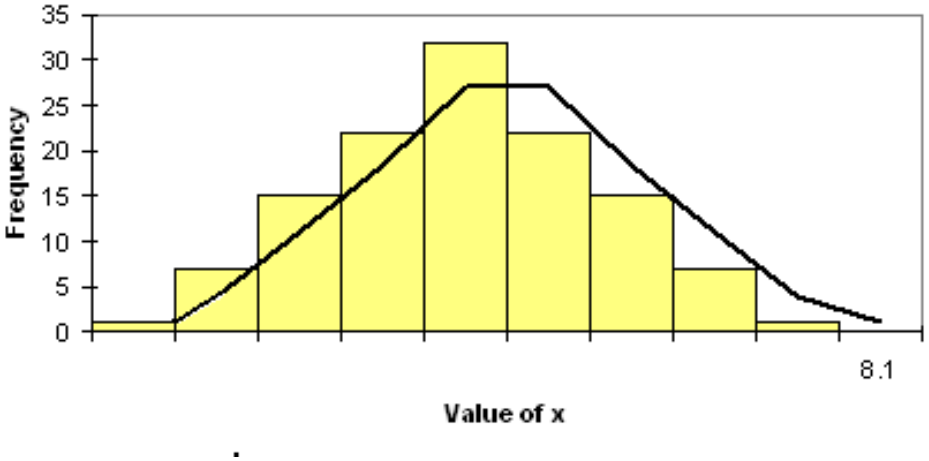
No simple formulae exist for probabilities associated with normal distributions. Tables can be used eg. Cambridge Elementary Statistical Tables - D.V. Lindley and J.C.P. Miller C.U.P.1952. These always refer to the $N(0,1)$ distribution - the so-called **standard normal distribution** with mean 0 and variance 1. More commonly and with greater flexibility, statistical packages or functions within spreadsheets provide these probabilities.

The importance of the normal distribution lies with an important result called **the Central Limit Theorem** which states that if a random sample of size n is taken from any distribution (not necessarily normal) with mean μ and variance σ^2 then if

n is reasonably large the sample means will follow a normal distribution with mean μ and variance σ^2/n (as $n \to \infty$) .

We say: $x \sim N(\mu, \sigma^2/n)$.

If this approximation holds then probabilities associated with the sample mean $\bar{x}$ can be evaluated approximately using the normal distribution tables. How large n needs to be depends on how near the underlying distribution is to a normal distribution.



6 Estimators

6.1 Point Estimates

Point estimates are single numbers estimating population parameters calculated from the sample data.

Each population distribution has one or two measures (parameters) associated with it that are needed to enable us to determine the probability density function and hence the theoretical frequencies of the distribution. We hope the statistics derived from the sample will be good estimates of the population statistics but as the sample statistics are subject to random variation they will not necessarily exactly equal the true (population) values.
Listed below are estimators for the distributions previously mentioned.

	Distribution	Parameter(s)	Estimator(s)
(i)	Binomial	n	sample size
		p	$\hat{p} = \frac{x}{n}$
(ii)	Poisson	μ	$\hat{\mu} = \bar{x}$
(iii)	Normal	μ	$\hat{\mu} = \bar{x}$
		σ^2	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
(iv)	Exponential	λ	$\hat{\lambda} = \frac{1}{\bar{x}}$

Although point estimates are useful it is often necessary to know how precise is the estimate.

6.2 Interval Estimates

An interval estimate is a statement that the population parameter lies within specified limits. A **confidence interval** is an interval estimate to which some specified level of confidence can be allocated. It has the property that in repeated sampling a known percentage of the confidence intervals will include the population parameter.

Using the properties of the normal distribution previously mentioned
95% of the $\bar{x}$ values lie within 1.96 standard errors of the mean
99.8% of the $\bar{x}$ values lie within 3.09 standard errors of the mean

However in reality we usually only have one value of $\bar{x}$ and we are trying to find out something about the population mean μ. The following confidence intervals can be calculated:

95% confidence interval for μ $\quad \bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

99.8% confidence interval for μ $\quad \bar{x} - 3.09\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 3.09\frac{\sigma}{\sqrt{n}}$

The 95% confidence interval means that if we calculate many such intervals, 95% will contain the true population value μ and 5% will not ie. we expect to be wrong 1 in 20 times. The reliability of the estimate can be improved by using the 99.8% confidence interval but this may be too wide to be useful.

This interval relies on an accurate assessment of the population variance σ^2. If this has to be estimated from the sample the confidence intervals are calculated as follows:

95% confidence interval for μ $\quad \bar{x} - t_{n-1,0.025}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,0.025}\frac{s}{\sqrt{n}}$

99% confidence interval for μ $\bar{x} - t_{n-1,0.005}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,0.005}\frac{s}{\sqrt{n}}$

The value t_{n-1} is the value of the t-statistic based on n-1 degrees of freedom and the required probability for the confidence interval. The t distribution differs according to the size of sample. The larger the sample the closer the approximation to the normal distribution.

7 Least significant differences between two sample means

In comparing the means of two populations, sample means x_1 and x_2 are used to estimate the population means μ_1 and μ_2. The difference between the 2 sample means is also normally distributed with mean μ_1-μ_2 and variance $\sigma_1^2 + \sigma_2^2$ equations.

The standard error of the difference between these 2 sample means is given by the equation:

se(difference) = $\sqrt{2}$ x se(mean)

Where the variance is equal for both populations and the number of observations is the same for each sample, the variance for the difference between the sample means is given by

$t_{n-1} \times$ se(difference).

The least significant difference (or LSD) calculated as t_α x se(difference) - usually the 5% level is the value which must be exceeded for the difference between the means to be considered different from 0 at the stated level of significance. The t value is taken from tables of the t distribution for the required level of significance and the degrees of freedom used in the calculation of the variance.

L.S.Ds should be used with care as misleading conclusions can be obtained if indiscriminantly used.

8 Hypothesis Testing

We have seen how we can use statistical theory to provide ourselves with confidence intervals for point estimates. Now we shall see how to use it to test assertions and scientific theories.

Example 1

A nutrition expert attached to the LEA claims that the average fat consumption in a canteen meal is 40g. A sample of 89 children gave a mean fat consumption of 52g with a standard deviation of 23g. Is the claim justified?

What can we say about the distribution of fat consumption?

What can we say about the distribution of the mean of a sample of size 89?

We are testing a hypothesis

H_0 (the null hypothesis) - the mean of the population $\mu = 40$g

against

H_1 (the alternative hypothesis) - the mean of $\mu \neq 40$g.

We can use our previous theory to find a confidence interval for μ.

The 95% confidence interval is

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

i e. $$52 - \frac{1.96 \times 23}{\sqrt{89}} \leq \mu \leq 52 + \frac{1.96 \times 23}{\sqrt{89}}$$

$$52 - 4.78 \leq \mu \leq 52 + 4.78$$

There are 2 conclusions we can come to:

(i) we have chosen one of the 5% of samples which give us a confidence interval not containing μ.

(ii) the assertion of the nutrition expert is incorrect.

If we calculate the 99.8% confidence interval, we find $\mu \leq 52 \pm 7.5$.

Hence, conclusion i) becomes

i) We have chosen one of the 0.2% of samples which give us a confidence interval not containing μ. - i.e. it seems even more unlikely.

We reject the null hypothesis (i.e. $\mu = 40$) at the 0.2% level of significance.

NB It is important to include the last part of the sentence since we are stating our chances of being wrong.

Example 2

The percentage of onions of a certain variety with a diameter 40mm or more is 25%. A farmer believes that his onions do better than this. Out of 110 bulbs chosen at random, 33 had a diameter greater than 40mm. Does this indicate that the farmer is correct?

H_0: $p = 0.25$

H_1: $p \neq 0.25$

95% confidence interval for p is-

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}$$

$$\frac{33}{110} - 1.96\sqrt{\frac{0.25 \times 0.75}{110}} \leq p \leq \frac{33}{110} + 1.96\sqrt{\frac{0.25 \times 0.75}{110}}$$

$$0.3 - 0.08 \leq p \leq 0.3 + 0.08$$

$$0.22 \leq p \leq 0.38$$

0.25 is included in this confidence interval, and therefore we have no reason to reject the null hypothesis.

i.e. The results do not provide evidence that the farmer is correct. It is often easier to rewrite the 95% confidence interval as

$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$ where the estimate of the mean = $\bar{x}$

OR $$-1.96 \leq \frac{p - p(hat)}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96$$ where the estimate of the proportion p(hat) = $\hat{\rho}$

To test the null hypothesis in case i), we calculate $\frac{x - \mu}{\sigma/\sqrt{n}}$, where μ is the value given by H_0.

H_0 is rejected (at 5% level of sig.) if this value is ≥ 1.96 or ≤ -1.96.

(at 0.2% level of sig.) if this value is ≥ 3.09 or ≤ 3.09.

To test the null hypothesis in case ii), we calculate

$\frac{(p - p(hat))}{\sqrt{\frac{p(1-p)}{n}}}$ where p is the value given by H_0.

H_0 is rejected (at 5%) level of significance if this value ≥ 1.96 or ≤ -1.96.
H_0 is rejected (at 0.2%) level of significance if this value $\geq \underline{3.09}$ or ≤ -3.09.

9 Comparison of Means

When carrying out an experiment or a survey to investigate the differences between samples, we should try and eliminate as much of the underlying variability as possible by good experimental design.

Example 1
In the comparison of the two diets, the diets were compared on mice which were genetically similar (i.e. from the same litter).

Example 2
In an experiment to compare the effects of methyl methacrylate and paraffin on the clotting times of human blood, one sample of blood was used for each pair of treatments. This was done to eliminate the variability between blood samples from different people.

The above examples are referred to as 'paired samples', since each value in one sample has a natural partner in the other sample.

Sometimes it is not possible to devise an experiment in this way and therefore we need to be able to test for differences between means when the samples are not paired.

Example 3
It is wished to compare the effectiveness of two different reading methods. A group of 100 children are randomly allocated to either method A or method B, and after 1 year their reading abilities are assessed by means of a test. The results are as follows:-

X is the score obtained using method A and Y using method B.

Method A $n_1 = 45$ $\sum_{i=1}^{45} x_i = 423$ $s_1 = 2.2$

Method B $n_2 = 55$ $\sum_{i=1}^{55} y_i = 567$ $s_1 = 1.9$

Since the samples are large we can assume that

$x \sim N(\mu_1, s_1^2/n_1)$ and $y \sim N(\mu_2, s_2^2/n_2)$

Since we are interested in the difference between $\bar{x}$ and $\bar{y}$, given the above, we can say

$$\bar{x} - \bar{y} \sim N(\ \mu_1 - \mu_2\ ,\ s_1^2/n_1 + s_2^2/n_2\ )$$

Using Normal theory, therefore

$$-1.96 \ \leq \ \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}} \ \leq \ +1.96$$

for 95% of samples.

In this case, the null hypothesis H_0 : $\mu_1 = \mu_2$
and the alternative hypothesis H_1 : $\mu_1 \neq \mu_2$

$$\text{therefore } z = \frac{\bar{x} - \bar{y}}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}} = \frac{423/45 - 567/55}{\sqrt{(2.2^2/45 + 1.9^2/55)}} = -2.18$$

Since the calculated z value lies outside the limits ± 1.96, we reject H_0 at 5%, and conclude that there is a difference in the effectiveness of the reading methods.

In the above example, the samples were sufficiently large to estimate accurately the standard deviations of the distributions from which they were taken. Sometimes this is not the case, but we may still proceed, provided that the samples are both taken from Normal populations with the same variance.

10 Comparison of variances

Sometimes we need to test whether two samples are taken from populations with the same variances.

If we are sure that the samples are taken from normally distributed populations, then

$$\frac{s_1^2}{s_2^2} = \frac{\Sigma(x_i - \bar{x})^2}{(n-1)} \Big/ \frac{\Sigma(y_j - \bar{y})^2}{(m-1)}$$

follows an F distribution with n-1 and m-1 degrees of freedom, and this can be used to test the significance of the ratio of the variances.

e.g. For the example given above the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$
the alternative hypothesis $H_1: \sigma_1^2 \neq \sigma_2^2$

If H_0 is correct, then $\frac{\sigma_1^2}{\sigma_2^2} = 1$

To test, calculate $\frac{s_1^2}{s_2^2} = \frac{2.2^2}{1.9^2} = 1.34$

(in order to use the tables you will need to ensure that the larger estimate is divided by the smaller).

This follows an F distribution with 44 and 54 degrees of freedom. According to the tables, for 95% of values,

$$(1 / 1.60) \leq F^{44}_{54} \leq (1.60).$$

The calculated value lies within these limits and we therefore accept H_0, i.e. the samples are taken from populations with the same variances.

11 The t distribution

In many scientific experiments it is impracticable to use large samples. Our previous theory is dependent upon the fact that we have a sufficiently large sample to estimate σ, if we are sampling from a Normal distribution. Suppose we only have a sample of size 5. When testing a

hypothesis about μ, we calculate the statistic $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and expect that 95%

of the time it will lie in the interval ($\pm$ 1.96).

What difference does it make if we calculate $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

Where $s = \frac{\sqrt{\Sigma(x_i - \bar{x})^2}}{n - 1}$

We can test this by simulation. The distribution of t obtained when a sample of size 5 is drawn from a Normal distribution and x and s are calculated from the sample,

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ can be simulated

The distribution looks almost Normal, although it can be seen to be rather more widely spread than the Normal. A calculation of the coefficient of kurtosis (skewness and Peakness of the data) would also reveal departures from Normality.

If we were to take a larger sample (e.g. 20), then the distribution of t would be closer to the Normal, since the estimation of σ is more accurate.

The exact distribution of t differs according to the size of the sample n. Probability points are tabulated against n - 1, known as the degrees of freedom.

The tables indicate that if we calculate the t statistic described above,

then 95% of the time it will lie in the interval $\pm$ 2.78 (for n = 5).

Example 1
The lengths of 8 birds eggs taken at random points in the colony of a particular species are as follows:

3.2 2.9 3.0 3.1 3.0 2.8 2.9 3.0

Give a 95% confidence interval for the mean length of all the eggs of this bird species.

$x = 2.9875$ $s = 0.1246$ $t_7 = 2.36$ (at 5%)

$$-2.36 \leq \frac{2.9875 - \mu}{0.1246/\sqrt{8}} \leq 2.36$$

$$2.9875 - 2.36 \times \frac{0.1246}{\sqrt{8}} \leq \mu \leq 2.9875 + 2.36 \times \frac{0.1246}{\sqrt{8}}$$

$2.88 \leq \mu \leq 3.09$ at the 95% confidence interval.

Example 2
Consider the results of an experiment to compare the effects of methyl methacrylate and paraffin on the clotting time of human blood. The following results were reported by Hirschboeck of the clotting times observed in 10 pairs of blood samples. One out of each pair was chosen at random and treated with methacrylate and the other with paraffin. The results are shown below.

Sample	1	2	3	4	5	6	7	8	9	10
Paraffin	10	27	11	18	19	16	16	18	22	26
Methacrylate	13	20	9	12	11	14	19	12	11	18
Difference d	-3	7	2	6	8	2	-3	6	11	8

What is the mean difference between the clotting times and is it likely to be due to chance?

$d = 4.4$ $s = 4.74$ $t_9 = 2.26$ (5%)

95% confidence limits $4.4 \pm \frac{2.26 \times 4.74}{\sqrt{10}}$

$= 4.4 \pm 3.39$

Therefore the true mean value for the difference in clotting times lies between 1.01 and 7.79 with 95% confidence, and is not zero.

12 Comparison of Means of Small Samples

In order to compare the means of 2 small independent samples (size n_1 and n_2) taken from Normal distributions which have the same variances we can use the fact that

$$\frac{\bar{x} - \bar{y}}{s\sqrt{(1/n_1 + 1/n_2)}}$$

follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

s^2 is called the pooled variance and is calculated from both samples by

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Example

A homogeneous block of land was sown with 9 plots each containing one of two varieties of winter wheat A and B.

A farmer needs to decide whether there is any real difference between the two varieties. What assumptions does she need to draw conclusions? What might she conclude? Give reasons for your answer. The dry matter yields in tonnes/ha were as follows

A	B
6.7	4.5
8.0	7.1
10.6	8.7
8.3	5.3
8.9	

$n_a = 5$ $\quad n_b = 4$

$\Sigma x_i = 42.5$ $\quad \Sigma y_j = 25.6$

$\Sigma x_i^2 = 369.35$ $\quad \Sigma y_j^2 = 174.44$

$\bar{x} = 8.5$ $\quad \bar{y} = 6.4$

$s_1^2 = 2.025$ $\quad s_2^2 = 3.533$

She needs to assume that the dry matter yields are normally distributed, and that the variances of each population A and B are equal.

The latter assumption may be tested using an F - distribution

$$F^3_4 = \frac{s_2^2}{s_1^2} = \frac{3.533}{2.025} = 1.74$$

which is not significant at 5% (cf tabulated value 4,3 degrees of freedom 0.05 is 9.12).

Therefore the samples come from populations with equal variances.

To test whether the means are the same, first calculate

$s^2 = 2.671$

Under H_0 this quantity has a t distribution with 7 degrees of freedom. Therefore

$$\frac{8.5 - 6.4}{\sqrt{2.671(1/5 + 1/4)}} = 1.92,$$ which is not significant at 5% (t_7 at 5% = 2.365)

Hence we conclude that variety A is not higher yielding than variety B at the 5% level of significance.

13 Non-Parametric Tests

In order to use the t-test to derive confidence intervals and test hypotheses about the mean, we have to assume that (i) the samples are taken from a normal distribution (ii) that the observations are independent and identically distributed and (iii) that if there is more than one sample the distributions from which they were taken have equal variances. If these conditions are not satisfied it may be possible to use non-parametric or distribution free tests. However, most of these tests do require the condition that observations are independent. (Yield of successive plots in a field trial may be correlated and therefore not independent) examples of the use of some of these tests are given below:

Wilcoxon signed-rank test (ref. Ridgeman 1975) **(for paired samples)**

In an experiment to assess the virulence of 2 strains of pea blight, 10 plants were taken at random and the first trifoliate leaflets inoculated with the bacteria, one race to each outer leaflet at random. After 10 days, the extent of the resulting chlorosis was estimated. The data obtained are given below:-

Plant	Race A	Race B	Diff (A-B)	Rank	Sign
1	15.6	14.3	+1.3	5	+
2	17.8	16.7	+1.1	4	+
3	13.2	14.1	-0.9	3	-
4	4.1	4.3	-0.2	1	-
5	5.5	2.2	+3.3	7	+
6	2.0	2.3	-0.3	2	-
7	8.7	4.0	+4.7	9	+
8	6.7	2.4	+4.3	8	+
9	7.5	1.7	+5.8	10	+
10	18.0	15.6	+2.4	6	+

Now sum separately the ranks which belong to the positive differences and those which belong to the negative differences.

$R_+ = 49 \qquad R_- = 6$

The normal approximation for this statistic is valid only for samples of size 8 or greater, and so the procedure calculates this approximation z, where

$$z = (WS - n(n+1)/4) / \sqrt{\{ n(n+1)(2n+1)/24 \}}$$

where WS = n x (n+1)/4 - (modulus total sum of signed ranks) /2

Where n is the number of observations

only use if the sample size is at least 8.

E.g. WS = (10 x (10+1)/4) – (43/2) = 6 (43 = 49 – 6)

$z = (6-(10 \times (10+1)/4) / \sqrt{\{10 \times (10+1) \times (2(10) + 1) / 24\}}$

= -21.5/9.81

=-2.19

Tables indicate the probability of the more extreme values of R_+ and R_- for non-zero differences when the two populations are the same. In the present case, with n = 10, the one-tail probability is 0.0143. Since the alternative hypothesis is that the two races simply differ in virulence, either positively or negatively, we need to double the probability (i.e. 0.0286). Hence the races are different at the 5% probability levels.

Alternatively, the smaller of R_+ and R_- can be compared directly with the appropriate table:

Critical value R, for U at P=0.05 where n =10 is 8. R_- is less than 8 and therefore we reject the hypotheses that the samples are the same.

The Sign Test (ref Ridgeman 1975)
(for paired samples)

This is even easier, but the test is not so powerful since it uses less of the information.

There are 7 plants for which Race A is more virulent than Race B and 3 plants for which Race A is less virulent than Race B. If the null hypothesis is true, these values should follow a binomial distribution. The probability of obtaining **exactly** 7 to 3 in favour of a particular race is 0.1172, so the probability of the difference is 2 x 0.1172 = 0.2344. Hence we cannot reject the null hypothesis using this particular test.

To find the probability of obtaining **at least** 7 to 3 in favour of a particular race, the probability of obtaining 0, the probability of obtaining 1, the probability of obtaining 2 and the probability of obtaining 3 would have to be found. Each of these probabilities would then have to be added together. The probability of obtaining at least 7 to 3 is 0.1719. The probability of the difference is 2 x 0.1719 = 0.3438. Again the null hypothesis cannot be rejected using this test.

The Mann-Whitney U Test (ref Box, Hunter & Hunter 1978)
(for unpaired samples)

An experiment was performed on a manufacturing plant by making in sequence 12 batches of a chemical using the standard production method (A), followed by 8 batches of a chemical using a modified method (B). The results from this plant trial are given below. What evidence do the data provide that method B is better than method A?

Method	B	A	A	A	B	B	A	B	A	A
Observation	79.3	79.7	81.4	81.7	82.6	83.2	83.7	83.7	84.5	84.5
Rank	1	2	3	4	5	6	7½	7½	9½	9½

Method	B	A	A	A	B	B	A	B	A	A
Observation	84.7	84.8	85.1	86.1	86.3	87.3	88.5	89.1	89.7	91.9
Rank	11	12	13	14	15	16	17	18	19	20

Calculate the sum of the ranks for each method separately.

For A, rank sum = 130½ For B, rank sum = 79½

For this test the smaller sum is used with appropriate "U" significance tables.

The model used is:

$$U_k = n_1 \times n_2 + n_k \times (n_k+1)/2 - R_k, \quad k=1,2$$

Where n_k is the sample size of sample k and R_k is the sum of ranks for sample k

E.g. using Sample from method B

$U_b = 8 \times 12 + 8(8+1)/2 - 79½ = 52½$

It seems that we cannot conclude that there are differences between the methods using this non-parametric test.

Other non-parametric tests exist for one-way and two-way analysis of variance.
For further reading see SIEGEL, S., 1956 Non parametric Statistics for the Behavioural Sciences McGraw-Hill, New York.

DESIGN AND ANALYSIS OF VARIETY TRIALS

1 Introduction

Because all biological material has inherent variability, replicated plots are needed to assess the performance of a variety. To compare the relative performance of several varieties under the same conditions, replicated plots of all the varieties are grown on a **contiguous** area. These plots are collectively referred to as a **variety trial**.

In addition to the data from the trial providing information about the relative performance of varieties under the same conditions, the analysis of the data is used to assess the confidence that we have in the results. More confidence can be placed in the results of a trial with a lower level of variation than one with a higher level of variation.

Trial designs suitable for variety trials are dealt with in the following sections.

2 Completely randomised design

Suppose that 6 varieties were to be compared in a design of 4 replicates.

This can be done by randomly assigning the varieties, A, B, C, D, E and F, to the 24 experimental plots as shown below.

C	A	E	C	B	D
F	C	F	B	F	C
A	D	D	F	E	B
A	E	E	A	D	B

The **random allocation** of varieties to the plots ensures that each variety has an equal chance of being assigned to each plot. Randomisation also protects against source of bias and is also necessary if significance tests are to be made.

One disadvantage of the design is that the allocation of varieties to plots may be advantageous to some varieties and not to others. Another disadvantage is that if the trial area is large, the trial could be subjected to positional effects such as fertility gradients or previous cropping which will produce inaccurate variety comparisons.

The design will be most useful in a controlled environment experiment or when the area covered by the experiment is small, for example in a growth chamber or in single plant experiments covering a small area.

Example of the above data is given below:

Treatment/Rep	1	2	3	4	Total	Mean
A	8.4	7.9	7.0	8.3	31.6	7.9
B	7.3	7.1	7.5	8.9	30.8	7.7
C	6.6	6.8	6.3	5.5	25.2	6.3
D	7.6	6.5	6.5	6.2	26.8	6.7
E	5.9	7.9	7.3	7.3	28.4	7.1
F	6.6	7.0	8.5	7.9	30.0	7.5
					172.8	7.2

When there are only two treatments, we tested if there was a difference between the population means by comparing the differences between the two sample means with the natural variation estimated from the variation between members having the same treatment. With more than two treatments we cannot get a single difference to represent the treatment effect, but it is not unreasonable to think that we could assess the variation between treatments as the square of the deviations of the treatment means from the general mean. Thus if there was no natural variation and no treatment effect, we might expect treatment means A, B, C, D, B and F all to have a mean of 7.2 and any deviation from that value would be a measure of the treatment effect.

It follows therefore that the effect for treatment A is 7.9 - 7.2 = 0.7. Similarly effects can be calculated for the other treatments.

The total variation of the 24 yields about the trial mean is:

$$\text{total SS} = (8.4\text{-}7.2)^2 + (7.9\text{-}7.2)^2 + \ldots\ldots + (8.5\text{-}7.2)^2 + (7.9\text{-}7.2)^2$$

$$= 8.4^2 + 7.9^2 + \ldots\ldots + 8.5^2 + 7.9^2 - \frac{(172.8)^2}{24}$$

This total variation in the trial may be split into two parts:

a) variation due to treatment differences
b) natural variation.

Assuming that the yield of each plot is the result of the additive effect of the factors, it follows that

plot yield = trial mean + treatment effect + residual.

If this additive model correctly explains how the plot yields arise,

total SS = treatment SS + residual SS

treatment SS = $4(7.9-7.2)^2 + 4(7.7-7.2)^2 + ... + 4(7.5-7.2)^2$

$$= \frac{3l.6^2 + 30.8^2 + \ldots\ldots + 30.0^2}{4} - \frac{(l72.8)^2}{24}$$

residual SS = total SS - treatment SS
The residual SS is a measure of the natural variation within treatments and alternatively is called error SS.

It may also be computed from the sum of the squares of the differences between the observed yields and the yields estimated from the additive model.

Degrees of freedom

The sum of squares measuring squared deviations of n sample values about the sample mean has n-1 degrees of freedom associated with it and gives rise to a variance calculated as:

variance = mean square = sum of squares / degrees of freedom.

The degrees of freedom are the number of independent observations of which a sum of squares is composed. For example the total sum of squares has been derived from the differences between the 24 observations and their mean. If we knew 23 of these and the mean, we could derive the twenty fourth. Consequently only 23 of these are independent, and we say that the sum of squares has 23 degrees of freedom associated with it.

Residual Mean Square

The residual mean square is assumed to estimate the random variation in the trial data. If the treatment means are all estimates of the same value of a population mean ie there is no significant difference between the treatments, then the variation amongst the recorded treatment means is a consequence of random variation and the variance calculated from the treatment mean square is estimating the same "true" value of random variation as is the residual mean square.

$$\frac{\text{treatment ms}}{\text{residual ms}} = \frac{1.52}{0.54} = 2.81$$

This ratio follows the F-distribution. The tabulated value of an F variate corresponding to P=0.05 (5, l8 df) is 2.77 and at P=0.01 is 4.25. The calculated value is therefore of such magnitude as would occur by chance with probability of between 5% and 1% and the assumption of no significant differences between the treatment means is discredited at a minimum of P=0.05. Having ascertained that real differences exist among the

treatments it is appropriate to apply the 't' test to assess the magnitude of differences necessary to achieve significance.

The variance of a single plot yield $s^2 = 0.54$.

The variance of a treatment mean of 4 plots = $s^2 \div 4 = 0.135$

se(treatment mean) = $\sqrt{0.135} = 0.367$

The variance of the difference between treatment means = $2 \times 0.135 = 0.27$

se(difference treatment means) = $\sqrt{0.27} = 0.520$
The variance, s^2, is estimated with 18 df and the corresponding value of t at the 5% probability level is 2.101.

The least significant difference between treatment means is therefore
se(difference treatment means) x $t_{0.05}$ = 0.53 x 2.101 = 1.1

Any two treatment means differing by more than 1.1 are said to differ significantly at the 5% probability level (P=0.05).

ie LSD(treatment means) = 1.1

Coefficient of variation for the trial is

$$\sqrt{\frac{\text{se per plot x 100}}{\text{trial mean}}} = \sqrt{\frac{0.54 \times 100}{7.2}} = 10.2$$

The disadvantage of the design is that an undesirable allocation of treatments can occur which is advantageous to some treatments and not others eg. if there was a fertility trend in the land with high fertility due at the right, treatment F would be at a considerable advantage compared to the other treatments. Similarly B might be at a disadvantage with 2 plots out of 4 being at one end of the trial area.
The design will be most useful in controlled environment experiments eg. growth chambers or when the area covered by the experiment is small eg. single plant experiments or when the land is uniform.
If we know, or anticipate that there are or might be positional effects, a much sounder experimental procedure is to use both blocking and randomisation.

Analysis of variance

Sources of variation	df	SS	ms	F
Between treatments	5	7.60	1.52	2.81*
Residual	18	9.72	0.54	
Total	23	17.32		

Total SS = $(8.4 - 7.2)^2 + (7.9 - 7.2)^2 + \ldots + (8.5 - 7.2)^2 + (7.9 - 7.2)^2$

= $8.4^2 + 7.9^2 + \ldots + 8.5^2 + 7.9^2 - \frac{(172.8)^2}{24}$

= 17.32

treatment SS = $4(7.9 - 7.2)^2 + 4(7.7 - 7.2)^2 + \ldots + 4(7.5 - 7.2)^2$

= $\frac{31.6^2 + 30.8^2 + \ldots + 30.0^2}{4} - \frac{(172.8)^2}{24}$

= 7.60

residual SS = $0.5^2 + 0.0^2 + (-0.9)^2 + \ldots + 1.0^2 + 0.4^2$

= total SS – treatment SS

= 17.32 – 7.60

= 9.72

Standard errors and least significant differences

se(treatment mean) = $\sqrt{\frac{0.54}{4}}$

= 0.37

LSD (treatment means) = $\sqrt{2}\, t_{18}$ x se(treatment mean)

= 1.414 x 2.101 x 0.37

= 1.1

3 Complete (or randomised) block design

If we know, or anticipate that there are or might be positional effects, or even if we know nothing about the trial area, a much sounder experimental procedure is to use **blocking**. This is a device used to minimise within block variation from sources other than the factors under test and utilises the fact that plots close to each other are more likely to be similar than those farther apart. The trial area is divided into homogeneous units known as blocks and in a complete block design all varieties are grown once in each of these blocks (replicates). A different randomisation is used for each replicate.

An example of a complete block design with 6 varieties, A, B, C, D, E and F, in 4 replicates is shown below. Each variety has been randomly assigned to one plot within each replicate.

Replicate 1

C	E	B	D	F	A

Replicate 2

F	D	E	B	C	A

Replicate 3

E	C	A	F	D	B

Replicate 4

D	A	F	E	B	C

Complete block designs are very commonly used in agricultural experimentation due to their simplicity and ease of analysis. In the absence of any knowledge about the trial area, the trial should be laid out so that the area covered by a replicate is approximately square, and within each replicate the plots are long and thin so that they sample the remaining block differences equally. Replicates can be used advantageously to spread work on a trial, for example by drilling/harvesting one replicate on one day, the others on another day. In this case replicate differences would reflect a combination of positional and time effects (if they exist).

Replicate size increases as the number of varieties increases. All plots in a replicate should be contiguous and therefore the larger the replicate the more likely it is that large variation between plots within the replicates will occur. Unless the trial site is known to be very uniform then an experiment containing a large number of varieties (usually 15 varieties or more) might be better conducted using an incomplete block design.

4 Incomplete Block designs

When there is a large number of varieties to test, say 25 instead of 6, each replicate of 25 plots is likely to cover a large area of land with an increasing risk of lack of uniformity within the replicate. To insure against such lack of uniformity, further blocking is carried out with each replicate being sub-divided into smaller units known as blocks. It is assumed that there is greater homogeneity within each block than can be expected between the blocks. Such designs are **called incomplete block designs**.

Included in these designs are specific incomplete block designs such as balanced, square and rectangular lattices but these designs are restrictive. The requirements of the statutory and non statutory variety led to the development of generalised lattice designs. (Patterson and Williams 1976) The original catalogue of designs provides designs suitable for any number of treatments up to 100, grown in 2,3 or 4 replicates. The experimenter is able to specify the size of the block and hence to fix the number of blocks per replicate. Additionally, if the total number of varieties does not factorise exactly, the design will permit some blocks to contain one more or one fewer plot than all the other

blocks. Thus a trial of 46 varieties may be grown with each replicate having 46 plots, subdivided into 8 blocks, 6 of these containing 6 plots and 2 containing 5 plots.

An example of an incomplete block design for 14 varieties (A, B, C, etc.) grown in 2 replicates each containing 3 blocks is shown below.

Replicate 1

H	I	E	L	B	K	D	G	N	A	F	M	J	C
2	2	2	2	2	1	1	1	1	1	3	3	3	3

Replicate 2

L	A	E	J	C	D	K	I	H	N	B	G	M	F
1	1	1	1	3	3	3	3	3	2	2	2	2	2

For incomplete block designs, the number of plots per block should be chosen so that the area covered by a block is approximately square. With the plot sizes used for the National List and Recommended List cereal variety trials, this results in a block size of between 4 and 8 plots. All plots within a block must be contiguous.

5 Other single treatment factor designs

Latin Square designs are also suitable for using for variety trials but again their restrictive nature makes them impractical for use for trials containing a large number of varieties.

Row and column designs may also be used and use blocking in two directions to reduce variation.

6 Analysis of variance

The analysis of variance is a technique which enables the estimation of components of variance from several sources, e.g. replicates, varieties and blocks, and provides a mean of simultaneously assessing whether there are significant differences amongst the variety means. Before applying this technique it is necessary to assume that the variety effects are additive and that experimental errors are independently and Normally distributed.

Most biological quantitative measurement data - yield, length - are suitable for application of the analysis of variance. Percentage data which covers a wide range of percentages is an example of a variable which may only be suitable for analysis in this way after transformation.

7 Analysis of variance for a complete block trial

With a complete block design, two specific factors, variety and replicates (reflecting positional effects) contribute to the overall variation in yield.

Assuming that the yield of each plot is the result of the additive effects of the factors, each plot value may be derived as follows

plot value = trial mean + variety effect + replicate effect + residual

The residual term is assumed to indicate the amount of random variation present and is the difference between the recorded plot value and the fitted plot value (= trial mean + variety effect + replicate effect).

If this additive model correctly explains how the plot values arise then:

Total ss = Replicate ss + Variety ss + Residual ss

The replicate and variety sums of squares are derived from the replicate and variety totals respectively, and the total sums of squares is derived from the individual data items. The residual ss is found by subtraction. The components of variance (mean squares) in the analysis of variance table are derived from these sums of squares (ss).

The residual mean square is assumed to estimate the variance arising from random variation in the trial data. If the variety means are all estimates of the same value of corresponding population means i.e. there is no significant difference between varieties, then the variation among the recorded variety means is a consequence of random variation and the variance calculated from the variety mean square is estimating the same "true" value of random variation as is the residual mean square.

Under the assumption of no significant difference between varieties the variance ratio, (variety ms)/(residual ms), has an F distribution. The calculated value of this ratio is compared with the tabled F variate and if the calculated value exceeds the tabled value then the assumption of no significant differences between the variety means is discredited.

Having ascertained that real differences exist amongst the variety means it is appropriate to apply the "t" test to assess the magnitude of difference necessary to achieve significance. This is done by computing the LSD (variety means). Any pair of variety means differing by at least the LSD (P=0.05) are said to differ significantly at the 5% probability level.

LSD (variety means) (P=0.05) = se (variety mean) x $\sqrt{2}$ x t_{α}, where α is the df of the residual mean square.

Example for a complete block trial

The trial is a complete block design with 10 varieties and 3 replicates.
Parameter: dry matter yield in t/ha.

Variety	Mean	%control (c)		Plot data		
A (c)	3.52	86	-	3.72	3.75	3.10
B (c)	4.72	115	+	4.86	4.73	4.56
C (c)	4.08	99		4.08	4.41	3.74
D	2.48	60	-	2.46	2.74	2.23
E	4.85	118	+	4.82	5.15	4.57
F	4.07	99		4.12	4.31	3.76
G	3.95	96		3.94	4.10	3.80
H	3.38	82	-	3.32	3.50	3.32
J	4.39	107	+	4.47	4.53	4.16
K	3.21	78	-	3.33	3.35	2.96
Trial mean	3.86		Mean	3.91	4.06	3.62
Control mean		4.11				
SE (variety mean)	0.067	1.60				
LSD (pairs)	0.201	4.9				
LSD (v control)	0.164	4.0				
V sig	0.1					
DF	18					
CV	3.0					

Analysis of variance table

Source of variation	df	sums of squares	mean squares	F ratio
Replicates	2	0.9913	0.4956	31.58***
Varieties	9	14.2735	1.5859	116.07***
Residual	18	0.2460	0.0137	
Total	29	15.5108	0.5349	

The calculated F ratio for varieties has a value of 116.07. The tabled value of the F variate corresponding to P=0.05 (9,18 df) is 2.46 and to P=0.001 is approximately 5.76. The calculated value is therefore of such magnitude as would occur by chance with probability of less than 0.1% and the assumption of no significant difference between the variety means is discredited at P=0.001.

The significant F ratio for replicates indicates that blocking has been worthwhile.

Residuals for the complete block data

Plot	Rep 1	Plot	Rep 2	Plot	Rep 3
1	-0.06	11	0.05	21	0.18
2	0.04	12	0.07	22	0.09
3	-0.11	13	-0.07	23	0.10
4	0.07	14	-0.05	24	-0.09
5	-0.05	15	-0.06	25	-0.03
6	0.01	16	0.04	26	0.02
7	-0.06	17	0.11	27	-0.06
8	0.09	18	0.14	28	-0.01
9	0.14	19	-0.04	29	-0.18
10	-0.07	20	-0.18	30	-0.01

The fitted value for plot 1 (variety G, rep 1) = 3.86 + (3.95 - 3.86) + (3.91 - 3.86) = 4.00

Residual = recorded value - fitted value = 3.94 - 4.00 = -0.06

A positive residual indicates that the variety is performing better than expected in that plot and a negative one that the variety is performing worse than expected. Because the residuals are Normally distributed with a mean 0 and a variance equal to the residual mean square, any plot with a residual that differs from 0 by at least 2 x sd has only a 5% chance of occurring and could indicate an atypical plot. Such plots should be investigated. Groups of residuals that all have the same sign, e.g. groups of large negative or large positive residuals, indicate lower/higher yielding areas in the trial and also may warrant further investigation.

8 Analysis of variance for an Incomplete Block Design

The analysis of variance for an incomplete block design is very similar to that for a complete block design except there is an additional component of variance due to the extra blocking factor.

Because the design is resolvable (i.e. each variety occurs once in each replicate) the data may be analysed as if it was from a complete block design. This is useful if for some reason some of the plots in the trial fail and invalidate the incomplete block design.

Incomplete block designs are usually more efficient (i.e. produce a lower residual variance) than complete block designs and this is seen from the last table in the example below.

Example for an Incomplete Block Design

Trial randomisation

Design for 21 varieties, 3 replicates and 3 blocks per replicate

Variety	Plot	Rep.	Blk	Vty.	Plot	Rep	Blk	Vty.	Plot	Rep	Blk	Vty.
A	1	1	2	J	22	2	3	P	43	3	3	J
B	2	1	2	AA	23	2	3	W	44	3	3	R
C	3	1	2	E	24	2	3	D	45	3	3	X
D	4	1	2	B	25	2	3	T	46	3	3	S
E	5	1	2	P	26	2	3	Z	47	3	3	AC
G	6	1	2	X	27	2	3	C	48	3	3	C
H	7	1	2	T	28	2	3	M	49	3	3	D
J	8	1	3	M	29	2	1	AA	50	3	1	E
M	9	1	3	R	30	2	1	E	51	3	1	M
N	10	1	3	Y	31	2	1	A	52	3	1	T
P	11	1	3	V	32	2	1	R	53	3	1	A
R	12	1	3	AC	33	2	1	H	54	3	l	Y
S	13	1	3	C	34	2	1	V	55	3	1	N
T	14	1	3	G	35	2	1	X	56	3	1	Z
V	15	1	1	D	36	2	2	AC	57	3	2	W
W	16	1	1	Z	37	2	2	S	58	3	2	B
X	17	1	1	W	38	2	2	G	59	3	2	P
Y	18	1	1	A	39	2	2	J	60	3	2	H
Z	19	1	1	N	40	2	2	B	61	3	2	G
AA	20	1	1	S	41	2	2	N	62	3	2	V
AC	21	1	1	H	42	2	2	Y	63	3	2	AA

Average efficiency factor 0.8658

Summary of results and data values for dry matter yield in t/ha

Variety	Mean (adj)	% control (c)		Plot data		
A (c)	10.55	103		10.58	10.54	10.85
B (c)	10.11	99		10.57	9.08	10.10
C (c)	10.94	107	+	11.14	10.72	11.23
D (c)	9.38	91		9.33	9.30	9.61
E (c)	10.34	101		10.34	10.33	10.82
G	10.27	100		10.77	9.88	9.59
H	9.44	92	-	9.60	9.76	8.88
J	9.46	92	-	9.86	8.40	9.92
M	10.56	103		11.19	10.38	10.40
N	10.62	103		11.34	9.14	11.03
P	9.48	92	-	10.25	8.92	9.13
R	10.26	100		10.99	10.02	10.28
S	10.15	99		10.50	9.54	10.06
T	10.45	102		10.80	10.25	10.56
V	9.70	95	-	10.29	9.76	9.17
W	10.09	98		10.02	9.52	10.42
X	8.74	85	-	9.40	8.37	8.90
Y	10.19	99		10.79	9.19	10.42
Z	9.30	91	-	9.71	9.14	9.15
AA	9.54	93	-	9.82	9.77	9.11
AC	9.25	90	-	9.19	9.05	9.34

Trial mean	9.94	
Control mean (t/ha)	10.26	
Se (variety mean)	0.220	2.15
LSD (pairs)	0.633	6.2
LSD (v Cont)	0.491	4.8
V sig	0.1	
DF	34	
CV%	3.6	

***** Analysis of variance *****

Variate: Data

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
Rep stratum	**2**	**5.6626**	**2.8313**	**15.03**	
Rep.*Units* stratum					
Variety	**20**	**20.4731**	**1.0237**	**5.43**	**<.001**
Residual	**40**	**7.5367**	**0.1884**		
Total	**62**	**33.6724**			

This shows there is a definite difference between variety yields

ie F pr <.001

*** MESSAGE: the following units have large residuals.**

Rep 2 *units* 20 -0.993 s.e. 0.346

******* Tables of means *******

Variate: Data

Grand mean 9.945

Variety	A(c)	AA	AC	B(c)	C(c)	D(c)	E(c)
	10.657	9.567	9.193	9.917	11.030	9.413	10.497

Variety	G	H	J	M	N	P	R
	10.080	9.413	9.393	10.657	10.503	9.433	10.430

Variety	S	T	V	W	X	Y	Z
	10.033	10.537	9.740	9.987	8.890	10.133	9.333

***** Standard errors of differences of means *****

Table	Variety
rep.	3
d.f.	40
s.e.d.	0.3544

9 Validation of trials data

Trial results may be invalid because either the variation in the trial as a whole is too great compared with the variation expected from previous experience or because the conditions under which the trial was conducted may be atypical of conditions to be expected in agricultural practice in the future. e.g. severely frosted.

The statistical parameters available from the analysis of variance are for assisting in the interpretation and validation of trials data and should not be used as strict guidelines for the omission of trials data. Data/trials should never be omitted just because the results seem peculiar.

10 Measures of internal variation

One practice sometimes used for omitting the results from a trial is to set an upper limit on the coefficient of variation (CV) and to consider omitting a trial if its CV exceeds this limit. The weakness of this criterion is that the CV is inversely proportional to the trial mean and hence exclusion of trials with high CVs may result in the exclusion of a higher proportion of low yielding trials. A more satisfactory method is to study the error variance and mean yield of the trial in question and to see if the residual variance is consistent with the other trials or not.

The majority of variety trials contain varieties with a wide range of yields and the analysis of variance should therefore give a highly significant **F ratio** for varieties. (significant at P=0.001) If this does not happen then this may indicate that the residual variance of the trial is unusually high (i.e. that there is a large variety x replicate interaction). A study of the plot **residuals** for the trial will highlight plots which have high residuals and hence are contributing most to the residual variance. Calculation of the range of yields for each replicate in the trial will show whether or not all replicates are equally variable.

11 Measures of external variation

Having studied the parameters from the analysis of variance for the individual trials' data, the trial series may then be considered as a unit. (A trial series is for example all the winter wheat Recommended List trials grown in 2001.)

Calculation of standardised residuals for the variety x site matrix will show in which trials a variety is yielding more than (or less than) expected. A positive residual indicates a variety performance higher than expected, a negative value that the variety is performing lower than expected at that site. Values greater than 2 indicate that the (observed - fitted) value differs significantly from 0 at the P=0.05 level of significance.

The **standard deviation ratio** (sd ratio) is derived from the standardised residuals for each trial/variety and is a measure of the relative variation of each trial/variety compared with the overall variation. A value of 1.0 indicates that the trial/variety has average variation and a value greater/less than 1.0 indicates above/below average variability. For trials with sd ratios greater than 1.2, the relationship between the standardised residuals and any agronomic records taken on the trial will be investigated, with the aim of explaining the high variation.

The **correlation coefficient** relating an individual trial's performance with the mean performance over all trials indicates how well the individual trial's results agree with the overall mean. Cereal variety trials' yield usually give correlation coefficients greater than 0.40 and therefore a trial with a correlation coefficient of less than this is an indication that the varieties are performing very differently at that site, i.e. that there is a variety x site interaction.

12 Over trials analysis

Variety trials are usually carried out at more than one site in a season and for more than one year and there is a requirement to amalgamate the results from all trials to produce an overall measure of performance. Because it is impossible to test all varieties in all trials in every year, the variety testing system gives rise to both complete and incomplete sets of data.

a) Reports based on complete sets of data.

Reports based on a complete data matrix, i.e. with all varieties grown at all sites, are relatively straightforward to analyse.

If the table has two factors such as varieties and sites, and if the requirement is to produce relative variety performances irrespective of whether there is a variety x site interaction, then it is reasonable to compute an analysis of variance using the variety means at each site, with the sites as replicates. The LSD (variety means) is computed from the residual variance (= variety x site variance).

A data set with more factors, such as years, sites and varieties is more complicated to analyse, but most statistical computer packages provide methods of doing this.

b) Reports based on incomplete data sets

A different statistical method is used to analyse the results from a set of trials in which all varieties are not grown. The method used is REML and this makes adjustment for the non-occurrence of varieties in some trials. The variety means produced are all comparable.

A simple example of the method is given over the page.

Variety	Year					Unadjusted mean	Adjusted mean
	1996	1997	1998	1999	2000		
A	5.10	4.79	3.74	3.27	5.16	4.42	4.42
B	4.53	4.42	3.52	3.27	4.68	4.08	4.08
C	-	-	4.26	4.05	5.52	4.61 *	4.86
D	-	-	3.95	3.72	5.04	4.24 *	4.49
E	4.58	4.95	4.07	3.64	5.34	4.52	4.52
F	4.39	4.60	3.55	3.20	5.05	4.16	4.16
G	4.39	4.42	3.42	3.23	5.12	4.12	4.12
H	4.79	4.42	3.42	3.27	4.86	4.16	4.16
I	-	5.20	4.20	3.75	5.40	4.64 #	4.74
J	4.44	4.10	3.24	3.18	-	3.74 #	3.94

* 3 year mean # 4 year mean

	l996-97	1998-2000	1996-2000
Mean of varieties A, B, E, F, G, H	4.62	3.99	4.24
Difference from 5 year mean	-0.38	+0.25	
Adjusted mean for variety D = 4.24 + 0.25 =	4.49		

Glossary of statistical terms

1 x_i is the ith value.

2 n is the number of values in a data set.

3 μ is the population mean.

$\hat{\mu}$ is the estimate of the population mean.

$\bar{x}$ is the sample mean $= \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

$= \dfrac{\sum_i x_i}{n}$

4 σ^2 is the population variance.

s^2 is the sample variance $= \dfrac{\sum_i \left(x_i - \bar{x}\right)^2}{n-1}$

5 s is the standard deviation = $\sqrt{}$variance

6 se is the standard error.

7 ms = mean square = variance.

8 se(mean) = standard error of the mean $= \dfrac{s}{\sqrt{n}}$

9 se(difference) = $\sqrt{2} \times$ se(mean)

10 LSD (or Sig.Diff.) = the least significant difference

$= \sqrt{2} \times t_{n-1} \times$ se(mean)

$= t_{n-1} \times$ se(difference)

where t_{n-1} = Student's t-value with n-1 degrees of freedom.

11 CV = coefficient of variation $= \dfrac{100 \times s}{mean}$

12 Significance levels: 5% = p□0.05 = *

1% = p□0.01 = **

0.1% = p□0.001 = ***

13 SS = sums of squares

14 CF is the correction factor $= \dfrac{\left(\sum x_i\right)^2}{n}$

STATISTICS: PROBABILITY, DISTRIBUTIONS, MEANS, VARIANCES, SIGNIFICANCE, POWER.

This is an attempt to put a little more rigour and theory into the background of the statistics that you probably all know. It might not help too much with the application of what you know, but hopefully it will give you some comfort as we move out of familiar territory. I hope there will be some new stuff too. These are advanced notes: they are not intended to show you how to calculate a correlation coefficient or carry out an analysis of variance. However, this is not a mathematically rigorous exposition: I may lie and cheat a little.

Some perceptions of statistics

"We see that the theory of probabilities is at bottom only commonsense reduced to calculation: it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct." PS Laplace

"The true ratio of the numbers can only be ascertained by an average deducted from the sum of as many single values as possible; the greater the number, the more are merely chance effects eliminated." G Mendel

"The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation." RA Fisher

"The equanimity of your average tosser of coins depends upon a law, or rather a tendency, or let us say a probability, or at any rate a mathematically calculable chance which ensures that he will not upset himself by losing too much, nor upset his opponent by winning too often. This made for a kind of harmony and a kind of confidence; it related the fortuitous and the ordained into a reassuring union which we recognised as nature. The sun came up about as often as it went down in the long run, and a coin showed heads about as often as it showed tails. Then a messenger arrived. We had been sent for. Nothing else happened. Ninety-two coins spun consecutively have come down heads ninety-two consecutive times…" T Stoppard (Rosencrantz & Guildenstern Are Dead)

Three men are in a hot-air balloon. The mist comes down and they are lost. One of the three men says, "I've got an idea. We can call for help." So they all lean out of the basket and shout: "Hellllooooo! Where are we?" Fifteen minutes pass. Then they hear a faint voice: " Hellllooooo! You're lost!!" One of the men says, "That must have been a statistician." Puzzled, one of the others asks, "Why do you say that?" The reply: "For three reasons. (1) he took a long time to answer, (2) he was absolutely correct, and (3) his answer was absolutely no use to anyone."

For me, statistics is a set of methods which can make sense of messy data and tease out signal from a noisy background. In addition, knowledge of these methods can lead to the design of better experiments and often prevent the generation of messy and uninterpretable data in the first place. Biological experiments, in particular, often require statistical design and analysis. Statistical methods provide me with a framework for thinking about scientific problems.

And finally, statistics is far too important to be left in the hands of statisticians.

Probability

If you do genetics you need to know something about probability – see the quote from Mendel above.

Some examples:

Ten coins are tossed and come down as heads six times and are left to rest. Then if I select one of these coins without looking, the probability that it is heads-up is 6/10.

Given a coin that we believe to be fair, the probability that it will come down heads if we toss it is 0.5: it is equally likely to be heads as tails.

Another coin is tossed a very large number of times, 1,000,000 say. It comes down heads 200,000 times. We now believe that the probability that the coin will come down as heads if we toss it again is 0.2.

Underlying the examples above is the concept of a population of events. There is also the concept of probability both in predicting the future and/or of describing or summarizing the past. In the first example, it is clear that we have a population of size 10 and that we are summarizing the past. In the second example we are predicting the future on the basis of prior knowledge – we believe on the basis of experience and what we know about physics and the Royal Mint, that the coin is unbiased. In this example, the population of events is effectively infinite – we believe that if the coin was tossed an infinitely large number of times that it would come down as heads ½ the time. In the final example, we are explicitly using historic events to predict the future.

Note that in the last example, we have modified out belief. Before tossing the coin, we had an a-priori probability that the coin would come down heads of ½. After many tosses, we modified this prior belief to a new value – the posteriori probability. It is fairly obvious in this case, but suppose we had tossed the coin only 10 times and observed two heads - would we modify our probability to 0.2, leave it at 0.5, of move it to somewhere in between the two?

Think of probability as a frequency or as a proportion: we count the number of times an outcome occurs (or we think it will occur) on a number of occasions. In some circumstances, we can count over a population of conceptually infinite size. We don't always have to define probability in terms of categorical outcomes (heads/tails or male/female). We can have the probability that a man is over 1.8m tall: what proportion of men exceeds this height? In this case, we would have to be careful about defining our population again – do we mean UK males, Dutch males (tend to be taller in my experience) or all males. What age range are we considering, etc?

Some rules and definitions for probability

Probability is usually written as p(x) = 0.05, meaning the probability that event x (eg being male and > 2m) occurs is 0.05.

If events are mutually exclusive, then probabilities can be added. For example, with a biallelic locus segregating in an F2:

p(AA) = 0.25
p(Aa) = 0.5
p(aa) = 0.25

So if allele A is dominant, the probability of observing the dominant phenotype is
p(dominant) = p(AA)+p(Aa) =0.75

Note that over all possible outcomes, probabilities add up to 1.
p(AA)+p(Aa)+p(aa) = 1

If two different outcomes are independent, the probability of both outcomes is the product of the probability of each outcome:

p(male) = 0.5
p(believe in Santa) = 0.1

P(male and believe in Santa) = p(male) x p(believe in Santa) = 0.05

Different outcomes are often not independent eg

p(male) = 0.5
p(like football) = 0.3 (I'm guessing).

But liking football is much more a male than a female thing. Perhaps the probability of liking football if you are male is 0.6. We write this as

p(football | male) = 0.6.

The '|' stands for "conditional on." In words:

p(likes football conditional on being male) = 0.6.

Note that p(football | male) is not the same as p(male | football) – which we'll check out to see if works as a diagnostic test for sex.

As a general rule of probability:

p(football | male).p(male) = p(male | football).p(football).

So we can work out p(male | football) :

0.6 x 0.5 = p(male | football) x 0.3

so

p(male | football) = 1.0

It follows that p(female | football) = 0, and also that p(male | don't like football) = 0.4

For this simple case, it is easy to see what is going on if we lay out the probabilities in a contingency table:

	male	female	total
like football	0.3	0.0	0.3
dislike football	0.2	0.5	0.7
total	0.5	0.5	1.0

Note that conditional probability is not necessarily the same as the probability that both events happen. So p(male | football) is 1.0 but p(male & football) is 0.3.

In this case, liking of football is very good at eliminating females – it has high *specificity* for males. However it has low *sensitivity* in that it doesn't detect them all. These definitions are important in medical statistics: maleness could be a disease and football a screening method.

p(A | B).p(B) = p(B | A).p(A) is better known as Bayes' theorem, more generally written as :

p(A | B) = p(B | A).p(A) / p(B)

In this context, p(A|B) is often called the posterior probability of A and p(A) the prior probability. Bayes theorem accounts for how our prior belief in A is modified by additional information coming from B.

Probability distributions.

Going back to our height example, we stated p(>1.8m) as 0.05. There must also be a p(>1.7m), a p(>1.6m^2) and so on. There is also a p(<1.8m and >1.6m). Every interval or subset will have a probability and the complete description of probabilities over all possible subsets is given by the probability distribution. Probability distributions are important for two reasons. Firstly they allow us to summarise data succinctly. For example, for height in males, we could tabulate the probability of men having heights between 1.5m and 1.6m then 1.6m and 1.7m and so on. However, if height tends to follow a particular frequency distribution, then we can summarise the distribution better by describing the properties of that distribution. In addition, if we know the distribution to which an observation or event is meant to belong, then we can assign a probability to that observation. If the probability is particularly low, we may reconsider our knowledge and decide that the observation is not quite what we thought.

Probability distributions can be split into those for outcomes which we can count – males, females, genotypes and so on, and those where the outcome is continuous, such as height and weight – most of the measurable phenotypes we come across in plant breeding. The former are called discrete distributions, and the latter continuous.

For discrete distributions, the function giving the probability for a particular outcome is called the probability mass function (p.m.f). For continuous distributions the equivalent function is the probability density function (p.d.f). This is more complex conceptually since it isn't really a probability at all. In the pedantry of probability, we can't attach a probability to someone having a height of exactly 2m, since no one is exactly 2m tall. All we say is that their height lies between two values, say 2m and 2+δm. To find the probability of someone's height being in this interval we have to integrate the p.d.f between these values.

Below are the some commonly encountered forms of both distributions, starting with discrete. There is more than you require here, but you should at least read about the binomial and normal distributions.

Bernoulli distribution.

This is the distribution of single events – something happens or it doesn't. A coin tossed once is the obvious example. A child could be a girl or a boy. I can win a race or not win. A positive outcome is given the value one and a negative the value zero. What you call positive or negative is arbitrary. The average of a Bernoulli distribution is just the

probability that the event happens: p. So for me to win a race, the probability is very small – both empirically and a priori. That is all there is to be said.

Binomial distribution

We observe an event a number of times, say *n*, and count the number of successes (*r*). The complete distribution is given by the binomial expansion :

$$\text{p}(r \text{ successes out of } n) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)}$$

This function for the distribution is called the "probability mass function" and gives the probability that the particular number of outcomes is observed.

If the probability of a success is *p*, then the mean number of successes is *np*. If we don't know what *p* is, then the best estimate is *r/n*.

We shall cover a lot of cases in which *p* is the probability of allele *A* being observed and (1-*p*) is the probability of allele *a* being observed.

The Bernoulli is a special case of the binomial with $n = 1$

Multinomial

We observe *n* events as before, but there are more possible outcomes – AA, Aa and aa. say, for three possible genotype classes.

In this case, p(AA)+p(Aa)+p(aa) = 1 where p(xx) is the probability of observing genotype xx. The expected number of AA individuals in our sample of *n* is therefore n.p(AA) with n.p(Aa) individuals and n.p(aa). For the three outcomes, the complete distribution is given by the multinomial expansion:

$$\text{p}(r_1, r_2, r_3 \text{ outcomes in } n) = \frac{n!}{r_1! r_2! r_3!} p_1^{r_1} p_2^{r_2} p_3^{r_3}$$

r_1, r_2 and r_3 are the observed numbers of each class. and p_1, p_2 and p_3 are the probabilities of observing each class. If these are unknown they can be estimated from the data as r_1/n, r_2/n and r_3/n.

The expansion to more than three outcomes is obvious.

The binomial is a special case of the multinomial distribution.

Poisson

This distribution is often followed by counts of events (usually rare events) which occur in an interval of time or space. A classic example is the number of soldiers killed by horse kicks in the Prussian army. Another might be the number of new mutations observed in a population, or the number of ergot infected ears of wheat in a plot. If *p* is the probability of the rare event occurring, then the average number of occurrences is generally given as λ. With *k* as the number of events, the complete distribution is given as:

$$p(k \text{ events}) = e^{-\lambda}\lambda^{k} / k!$$

Note that the Poisson distribution has a probability for every number of outcomes count from 0 to infinity. In practice, however, these probabilities become vanishingly small as the number of counts rises.

The Poisson distribution is close to the binomial distribution if *n* (for the binomial) is large and *p* is small. This is because, given the binomial distribution

$$\frac{n!}{r!(n-r)!} p^{r} (1-p)^{(n-r)}$$

with $p = \lambda/n$ is

$$\frac{n!}{r!(n-r)!} \frac{\lambda^{r}}{n^{r}} (1-p)^{(n-r)}$$

and

$n!/(n-r)!\, n^{r} \rightarrow 1$

$(1-p)^{(n-r)} = (1-p)^{n} \cdot (1-p)^{-r}$

with

$(1-p)^{n} \rightarrow e^{-np} = e^{-\lambda}$

and

$(1-p)^{-r} \rightarrow e^{-r\lambda/n} \rightarrow 1$

Put all this together:

$$\frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} \quad \sim \quad e^{-\lambda} \lambda^r / r!$$

This comes in handy sometimes.

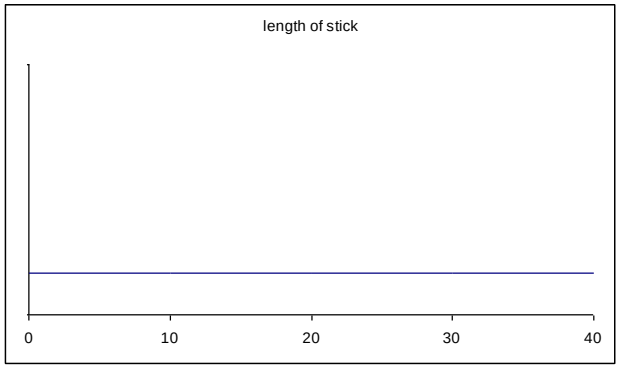
For large λ, the Poisson distribution can be approximated by a normal distribution (introduced below) with both variance and mean equal to λ.

Uniform distribution

Every outcome is equally likely. The uniform distribution can be either discrete or continuous. An example of a discrete uniform distribution would be the probability of a letter in the alphabet being chosen at random. An example of a continuous uniform distribution is the distribution of lengths of a broken stick, where the break has occurred at random along its length.

To my mind this is the easiest of the continuous distributions. The distribution of values is uniform between the range *a* and *b*. It is easy to see that the mean of a set of numbers drawn at random from the same uniform distribution is expected to have a mean equal to the mid point

mean = (*a*-*b*)/2

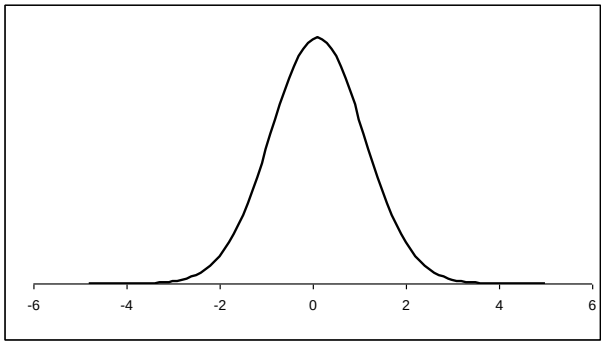


In the example, *a* = 0 and *b* 40.

A special case is the distribution of real numbers in the range 0 – 1. This is often the expected distribution of probability itself: a probability between 0.01 and 0.02 is just as likely as a probability between 0.5 and 0.51, but a probability of between 0.5 and 0 is much more common than a probability between 0.0001 and 0.

Normal distribution

The bell –shaped curve. Also called the Gaussian curve.



This is the probability distribution with which we shall be dealing with most in this course. Its probability density function is:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This complicated looking formula has two parameters. Mu (μ) is the mean, also the median (the midpoint of the distribution) and the mode. (For continuous distributions the mode is the value of x for which the probability density is a maximum. For discontinuous distributions it is, more simply, the most common class or value). Sigma squared (σ^2) is the variance, to which we shall return. The square root of the variance, σ, is called the standard deviation and is related to the spread of the distribution.

For continuous distributions, the probability of any single outcome is infinitesimally small. No-one is exactly 2m tall. As precision of measurement increases, we may be able to say that someone's height lies between 1.99999m and 2.00001m. But within this interval, for a continuous distribution, there are still an infinite number of other intervals – and so on. To assign probability we integrate the pdf to find the area under the curve between our desired limits. So although no one is 2m tall, we can do sums like

$$\int_{1.999}^{2.001} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

to find the probability that someone's height lies between 1.999m and 2.001m. This is good enough for basketball and the police. Of course, these calculations are usually carried out by statistical software or using spreadsheet functions.

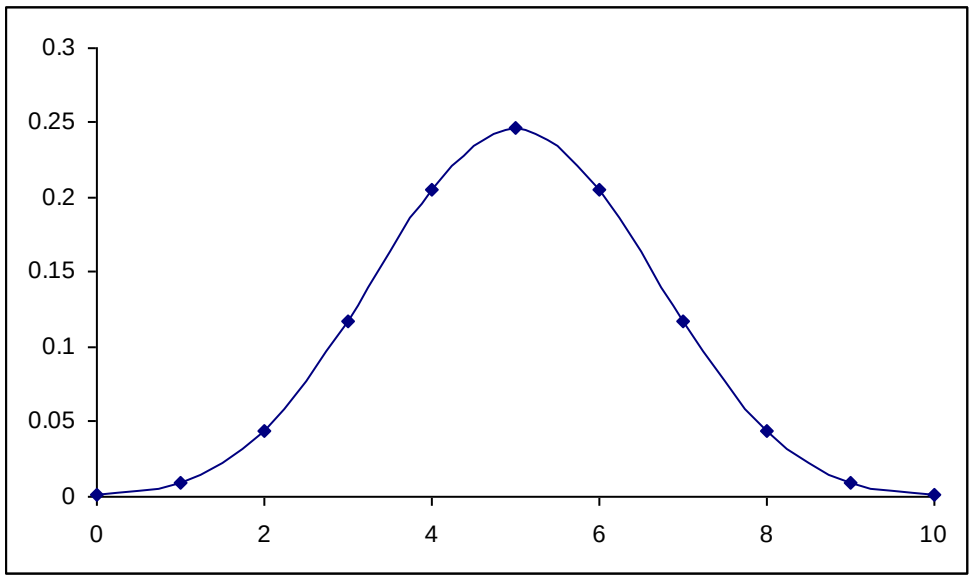
Most commonly we want to find the probability of an unusual observation. The normal distribution has a range of -∞ to +∞. To quantify how strange an unusually large observation is, we would calculate the probability

$$\int_x^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This is to say, we calculate the probability of finding an observation as large or greater than the one we actually have. ie p(outcome ≥ *x*). We can calculate the probability of an unusually small observation by integrating between -∞ and *x*.

Often we want to find the probability that an observation deviates from the population mean to the extent observed. Here we integrate between *x* and +∞ as before, but double the probability to account for the fact that deviations from the mean this large are equally likely to be positive as negative: the normal distribution is symmetrical about the mean as you can see in the figure.

The normal and binomial distributions are related – as the population size of the binomial gets larger and larger (ie *n* in the binomial formula), provided that *p* is not too close to zero, then the shape of the two distributions is pretty much the same. For example, the curve below is for a binomial distribution with n = 10 and p = 0.5.



The *standard normal distribution* is a normal distribution with a mean of zero and a standard deviation of 1. Data from any distribution can be standardized by subtracting from each observation the mean and then divide by the standard deviation:

$$z = \frac{x - \mu}{\sigma}$$

This is useful when it comes to comparing data - maybe for different traits or for different variety trials - since it puts all measurement on the same scale. Note however that if the original distribution is not normal, the standardised distribution will also not be normal.

Although the pdf for the normal is an ugly looking thing, the normal distribution has many useful properties, It is symmetrical for one. Importantly, many traits of interest in plant breeding – eg yield – are approximately distributed as normal. Even when traits are not distributed normally, it often makes very little difference to the outcome of our statistical analysis if we treat them as such, as we shall explain shortly. In addition, simple transformation to the raw data, for example taking logs, often generates transformed variables which are closer to normal.

To indicate that a variable X is normally distributed, we write $X \sim N(\mu,\sigma^2)$ so $\sim N(10,100)$ would indicate a distribution with a mean of 10 and a variance of 100.

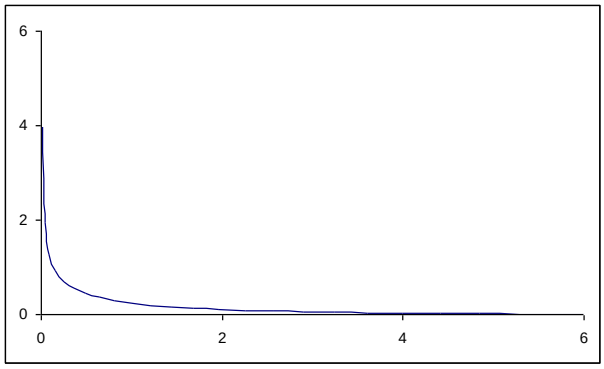
The chi distribution

If we take a standard normal distribution $N(0,1)$ and either ignore all the negative values, or take the absolute values (or square everything then take the square root) we get a chi (χ) distribution. For obvious reasons, this is also referred to as the half-normal distribution. The pdf is a complex looking thing that needn't concern us.

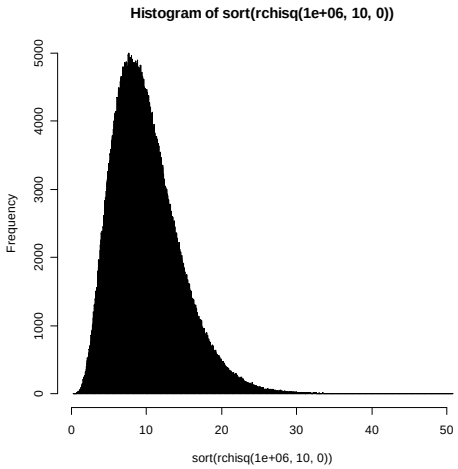
We give the chi distribution only because it leads to:

The chi square distribution

If we square a standardised normal distribution with a mean of zero, we get a chi-squared distribution with one degree of freedom (df – to be explained later). With one df the distribution looks like this:



If we have observations for a set of individuals from a number, n, of $N(0,1)$ distributions and the data in each are independent of each other (we say they are independently and identically distributed i.i.d.) and we square then sum the values of each observation for each individual, we get a chi-squared distribution with n degrees of freedom. You might ask why we would want to do this, but the chi-squared distribution comes up in a lot of significance testing. A chi squared distribution with 10 df is shown below



As the number of df increases, the chi-squared distribution can be approximated by the normal, although the approximation is not that good in this case.

The pdf is even more unpleasant than the normal distribution but fortunately we won't be required to use it directly in this course.

The mean of a chi-squared distribution is equal to its df.

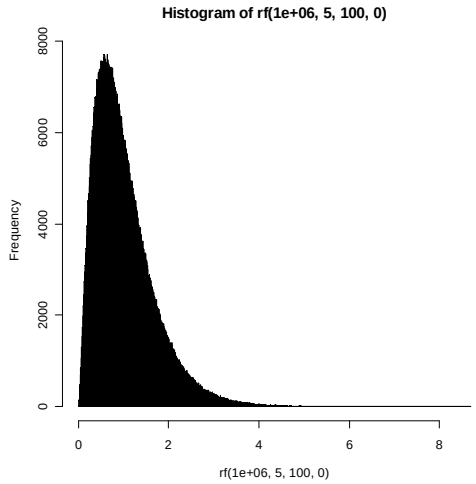
The F distribution

This also features heavily in significance testing, has a horrible p.d.f., but is related to chi sq and therefore ultimately to the normal distribution.

The F distribution is defined as the ratio of two chi-squared distributions, each first divided by its degrees of freedom:

$$(\chi^2_a/a)/ (\chi^2_b/b)$$

The pdf is therefore quite unpleasant. The shape of the distribution depends on the degrees of freedom of both numerator and denominator. An example for 5 and 100 df is shown below.



Although the pdf is horrible, we can summarise some properties of the distribution.

If the denominator df are large, and the numerator df are *a* then *a*F is approximately distributed as χ^2 with df equal to *a*.

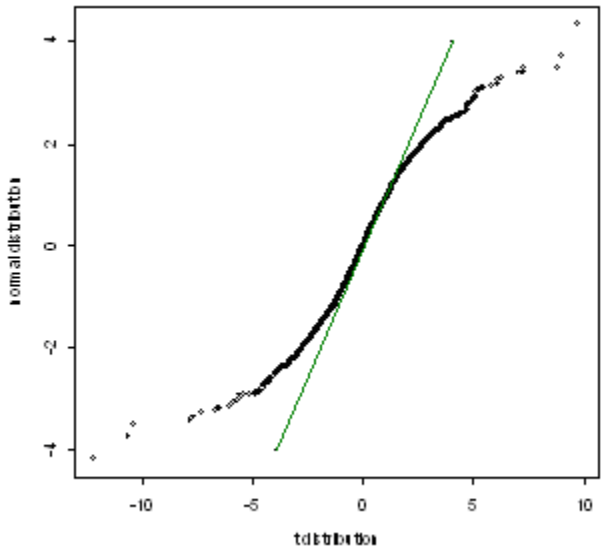
If the numerator df is 1 then F is distributed as (Students' t-distribution)^2, described below.

Student's t-distribution

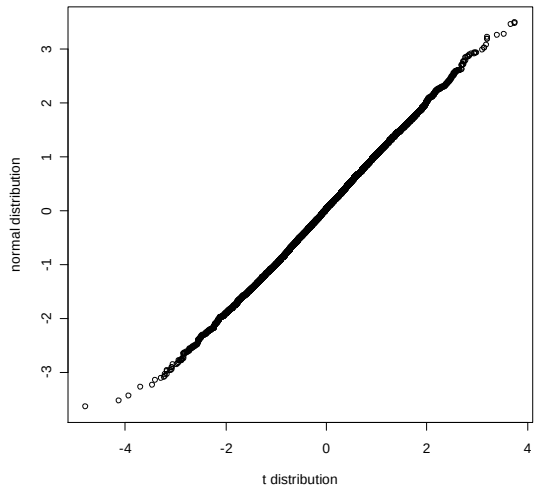
Named after Student (pseudonym of William Gossett).

This is an extension of the normal distribution to cases when we don't know what the standard deviation or variance is. However, we can estimate it from the data, as we shall see. If we used this estimate directly in the p.d.f. for a normal distribution to assign probabilities to our observations they would look slightly more extreme than they actually are. To avoid this, we ought not apply the p.d.f. for a normal distribution to calculate probabilities, we should use t-distribution instead. Again the pdf is quite a nasty looking item which need not concern us.

A plot of the t distribution looks very similar to the normal distribution, but the tails of the distribution are longer – there is more chance of observing extreme values. The graph below plots 10,000 sorted random normally distributed numbers against a corresponding set of 10,000 numbers from a t-distribution with 5 df. You can see that between values of about +2 and -2 there is very good agreement between the two, but after this values from the t-distribution become more extreme.



The mean of the t distribution is zero. As the number of observations in the sample increases, this distribution approximates to a standard normal distribution. Here is the plot corresponding to that above for t with 100 df.



The relationships is close to 1:1 over the whole range, so with this number of degrees of freedom, we will get very similar answers whether we treat the data as normally distributed or distributed as t.

It we square a variate distributed as t with n df, then the squared variate is distributed as F with 1 and n df.

The normal distribution again: variance, standard deviation, central limit theorem and standard error.

For any population, whatever the distribution, variance is defined as

$$E(x\text{-}Ex)^2$$

The 'E' stands for "expected value", ie the mean. E*x* is therefore the mean of the population. *x*-E*x* is the deviation of an observation from its mean. Obviously sometimes this deviation will be positive and sometimes it will be negative. By squaring it, it is always positive. $E(x\text{-}Ex)^2$ is thus the mean of the deviations squared. As such, for any probability distribution, it is a measure of the spread of the data around the mean.

For the normal distribution, $E(x\text{-}Ex)^2 = \sigma^2$. The square root of the variance, the standard deviation, σ, is the distance from the mean to the point of inflection of the distribution. The point of inflection is the point as you go up a slope where it stops getting steeper and starts to get flatter. Mathematically, it is the point where the second differential of a function (f '') is zero. Remember the first differential, f ', detects maxima and minima.

Generally, we work in variances rather than standard deviations: they are easier to manipulate but harder to understand. For example, variances but not standard deviations can often be added across datasets. However, the units of measurement of variance are those of the variable squared, whereas those of the standard error are identical to the raw data. So if I'm told that the mean salary of a plant breeder is £30k (say) but that the standard deviation among breeders is £15k (say) then that means more to me than being told that the variance is £225k.

Even if data are not normally distributed, it is worth calculating the variance, for two reasons.

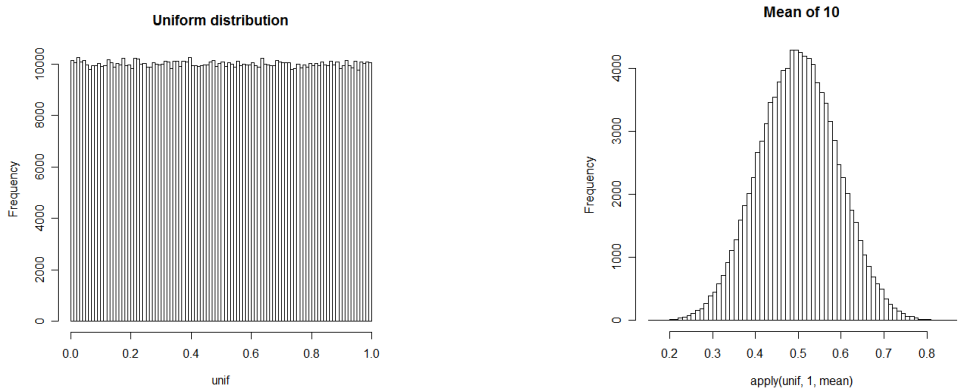
1) Often, we are not interested in the probability distribution of the raw data, we are interested in the distribution of the means of the data. The variance of the mean of a set of n observations is the variance of the observations divided by n. So if we had a set of 100 males and took their average height, the variance of the average is the variance among the males themselves divided by 100. By the variance of the average, what we mean is that if we repeated this experiment a very large number of times, preferably an infinite number of times, then calculated the variance of these means, we would get the same variance as if we divided the variance among the individual males by 100. This is important, because it means that we can use an estimate of the variance of the mean as an indication of how accurately we have measured that mean in the first place without the requirement of repeating the experiment many times. The square root of the variance of the means is called the standard error. Standard deviation refers to the population and its estimate is independent of sample size. Standard error refers to the mean and its estimate goes down as sample size goes up.

2) The central limit theorem – raised to almost religious status by Francis Galton (Darwin's cousin and early biometrician).

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error" [ie the central limit theorem]. The law would have been personified by the Greeks and deified if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason."F Galton Natural Inheritance (1889)

The central limit theorem states that for a given distribution, not necessarily normal, with mean μ and variance σ^2, then sample means of size n taken from this distribution will also have a mean equal μ (obviously) and a variance σ^2/n, but most amazingly the distribution of means will itself approach normality. This is true for almost all probability distributions, and certainly for all the ones that you will encounter in practice. (There are some distributions which have no mean or variance so the central limit theorem can't apply, but are still proper probability distributions: their integral = 1. Don't ask.) So whatever the original probability distribution, the distribution of the average of samples drawn from that distribution will tend to normal – with all the advantages of manipulation and interpretation that that gives. Of course, the closer to normal the original distribution,

the smaller the sample size need be to get a normal distribution of sample means, but even for horrible U-shaped distributions, it works eventually. In fact, in the days of less powerful computers, one method of deriving normally distributed random numbers was to take the mean of samples of uniformly distributed random numbers (which are easier to generate.) Here is an example.



The distribution on the left is of 1,000,000 uniformly distributed random numbers. That on the right is of the means of those numbers taken ten at a time.

In addition to allowing us to infer properties about estimates of means, in plant breeding the central limit theorem has another consequence. It probably accounts for why many phenotypes appear to be roughly normally distributed – they are the average of a whole host of underlying environmental and genetical factors. Each of these can have an unknown distribution but by the time they are averaged to generate a phenotype, the resultant distribution can look pretty normal.

Having raised the profile of the variance, we shall now back track quickly and look at what the variance is for the distributions we have considered so far:

	mean	variance	
Bernoulli	p	$p(1-p)$	
binomial	np	$np(1-p)$	
multinomial	p_i	$n\ p_i(1-\ p_i)$	for each class i in turn
Poisson	λ	λ	
uniform	$(a+b)/2$	$(a-b)^2/12$	
normal	μ	σ^2	
standardized normal	0	1	
chi-square	df	2df	
F	$df_2/(df_2-2)$	messy	mean is ~ 1.
t	0	df/(df-2)	

df = degrees of freedom.

Manipulating variances

If $y = kx$ where k is any constant, $Vy = k^2Vx$

This follows from the definition of a variance:

$$\begin{aligned} Vy &= E[y - Ey]^2 \\ &= E[kx - E(kx)]^2 \end{aligned}$$

The mean of a variable multiplied by a constant is the same as the constant multiplied by the mean of the variable: it makes no difference, other than convenience, whether we measure variety yield in tonnes per hectare or pounds per acre.

$$E(kx) = kE(x)$$

So

$$\begin{aligned} Vy &= E[kx - kEx]^2 \\ &= k^2\, E[x - Ex]^2 \end{aligned}$$

This is common sense – the variance is measured in units of x squared, so if we change the scale of measurement by a factor k, then scale of measurement of the variance is changed by a factor k^2.

We have already commented that variances are additive. So if we have a variate *y* and a variate *x*, the variance of (*x*+*y*) is V*x* + V*y*. This also follows from the definition of the variance as $E(x-Ex)^2$. For example, if $z = x + y$

$$E(z-Ez)^2 = E[x+y-E(x-y)]^2.$$

The mean of a sum is the same as the sum of the means so

$$Ez = E(x+y) = Ex + Ey.$$

So

$$\begin{aligned} E(z-Ez)^2 &= E[x+y-Ex-Ey)]^2 \\ &= E[\,(x-Ex)^2 + (y-Ey)^2 + 2E(x-Ex)(y-Ey)] \\ &= E(x-Ex)^2 + E(y-Ey)^2 + 2E(x-Ex)(y-Ey) \end{aligned}$$

For independent variables, the value of x is unrelated to the value of y so the third, cross product, term is zero. (It is actually the covariance – see below.) So

$$E(z-Ez)^2 = Vx + Vy + 0$$

If one variate is a function of another: $y = f(x)$ and

$$Vy \sim (dy/dx)^2 \, Vx.$$ (This can be proved by Taylor's theorem.)

If one variate is a function of two other variables: $z = f(x,y)$, and

$$Vz \sim (dy/dx)^2 \, Vx + dz/dx)^2 \, Vy$$

NB – these formulae are only correct for independent (uncorrelated) variates, although they can be modified to include correlation.

Correlation and covariation

Independent variables generally present no problem: the joint probability of an observation (x,y) say, is just $p(x).p(y)$.

If the variables are correlated, life is a little harder. We shall restrict the discussion to variables considered in pairs. This usually gets us most of what we require. The simplest way to study the relationship is to plot one variable against another. This is always worth doing.

To quantify the way in which the two variables vary together, we use the covariance, which is analogous to the variance for a single variable. It is defined as:

$E(y-Ey)E(x-Ex)$

cf – variance = $E(y-Ey)^2$ so the if $y = x$ the $Cov(x,y) = V(x)$.

Thus, rather than averaging the square of the deviations from the mean as in the calculation of variance, we average the product of the deviations for one trait with the deviations from the other.

In passing, and referring back to the previous section, note that for two correlated traits, x and y, the variance of the sum is:

$$V(x + y) = Vx + Vy + 2Cov(x,y)$$

Also, $V(x + x) = Vx + Vx + 2Cov(x,x)$

$= 4Vx$

as it should since V(x +x) = V(2x).

Note that the covariance can be negative.

Covariances can be calculated for any pair of traits, whatever their distribution, but the interpretation of covariance is most easy if the two distributions are themselves normally distributed.

If two distributions are standardized to have zero mean and variance of one (so they are $N(0,1)$ if they were normal on the original scale), then the covariance between the rescaled traits cannot be lower than -1 or greater than +1. The covariance between pairs of traits rescaled in this manner is called the correlation coefficient – it is easier to understand than the raw covariance because of the -1 … +1 scale. High values indicate a strong positive linear relationship; low values a strong negative linear relationship and values close to zero the absence of a relationship.

In fact, there is no need to go through the rigmarole of transforming the data as described above. The correlation coefficient is just:

$$\mathrm{cov}(x,y) / \sqrt{(\mathrm{V}(x).\mathrm{V}(y))}$$

The correlation will not pick up all relationships between pairs of traits. In extreme cases, where a plot of one trait against another is U shaped, then there is a clear relationship but the correlation is zero. I must say, however, that I've never come across a relationship between traits like this, but for example, if extreme phenotypes tend to die and phenotype is plotted against survival, then this relationship could arise. (This is called stabilizing selection: an intermediate phenotype has highest genetic fitness.) Also, when one or both traits have non-normal relationships, single data points can render the correlation close to +1 or -1 even though for the majority of the data the relationship is much weaker. The moral is always to plot your data, don't just look at the correlation coefficient.

Estimation

The discussion above has described distributions in terms of their known means, variances and other parameters. Most of the time, we have a set of data from which we wish to estimate parameters – most commonly the mean and variance. Below, without commentary, we state how these parameters are estimated. Then we try to justify this, and therefore show by extension how we can proceed to estimate parameters in more complex cases.

mean	$\hat{\mu}$	=	$\Sigma x_i / n$
variance	$\hat{\sigma}^2$	=	$\Sigma (x_i - \hat{\mu})^2 /(n-1)$
alternative formula easier for hand calculation		=	$(\Sigma x_i^2 - (\Sigma x_i)^2 /n) / (n-1)$
variance of mean	$V_{\hat{\mu}}$	=	$\hat{\sigma}^2 / n$
standard deviation	s.d.	=	$\sqrt{\sigma^2}$
standard error	s.e.	=	$\sqrt{\hat{\sigma}^2 / n}$
covariance	cov_{xy}	=	$\Sigma (x_i - \hat{\mu}_x)^2 (y_i - \hat{\mu}_y)^2 /(n-1)$
alternative formula easier for hand calculation			$(\Sigma x_i y_i - (\Sigma x_i \Sigma y_i) /n) / (n-1)$
correlation.coefficient	r_{xy}	=	$cov_{xy} / \sqrt{(V_x V_y)}$

There are various conventions for distinguishing between data, parameters and estimates of parameters. Those that are followed frequently depend (at least in my case) on the ease with which they can be typed in your favourite word processor.

Greek letters	parameters
Latin letters	data
^ sign over a letter	generally means “estimate of”
- sign over a letter	generally means “average of”

Why are these estimates the ones to use? Others are possible. For example, for the normal distribution, why not use the mode or median as an estimate of the mean rather than the arithmetic average? They all have the same expectation.

In this course, we shall rely predominantly on two commonly used estimation methods – least squares, which is particularly appropriate for parameters associated with the normal distribution, and maximum likelihood, which is more broadly applicable. The two give similar, but not necessarily identical answers. Historically, least squares came first, but we shall deal with maximum likelihood first. In addition I shall attempt to explain something about Bayesian estimation. This is used little in the course, but is becoming of increasing importance in complex estimation problems in genetics and bioinformatics so

you will inevitably come across it or end up using software which implements Bayesian methods.

Maximum likelihood

Maximum likelihood estimates have the following properties:

bias:	ML estimates can often be biased in small samples but:
consistent:	ML estimates home in on the true values as the sample sizes increases.
sufficient:	ML estimates use all the information available in the sample
efficient:	ML estimates have no unbiased competitors which are more precise (in large samples). Essentially ML estimates have the smallest variance.

It is unfortunate that ML estimates can sometimes be biased, but generally, for the samples sizes used in most sensible experiments, there is little need to worry.

The method of maximum likelihood searches for values of parameters which maximize the fit of the parameters to the data in our sample. In other words we search for values of the parameters which make the data more likely. In a sense, we select the parameter values which maximize the probability of observing our data. There is a problem here in that we are treating the parameters as if they themselves belong to a probability distribution. To avoid saying we want to maximize p(parameters | data) we say that we want to maximize the likelihood of the parameters given the data, l (parameters | data). What we actually work with, however is the probability of the data given the parameters, which we can calculate from the p.d.f. of the distribution describing the data.

$$l \text{ (parameters | data)} = \text{p (data | parameters)}$$

So to maximize l (parameters | data) we treat it just as if it were a probability. This is all a bit philosophical. If you are one of those happy folk who neither worry nor care about this, then I envy you. It is all easier to follow with an example.

We have a set of observations $[x_1, x_2 \ldots x_n]$ from a normally distributed population from which we wish to estimate the mean and the variance. What are the maximum likelihood estimates?

The mean first:

Using the pdf of the normal distribution,

$$l\,(x_i) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{\frac{-(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}}$$

We wish to find the values for $\hat{\mu}$ and $\hat{\sigma}^2$ which maximize *l*. Each observation is independent of the others, so likelihood, just like probability, can be multiplied together to get the total likelihood of the sample

$$l\,(x) \quad = \quad l\,(x_i) \,.\, l\,(x_2) \ldots\, l\,(x_n) \quad = \quad \Pi l_{x_i}$$

The Π symbol stands for multiplication of all the following terms.

This product, which involves working directly with the pdf is too hard. In general, with likelihood, it is easier to work with natural logs.

$$L_{(x)} \quad = \quad \Sigma \ln(1/(\hat{\sigma}\sqrt{2\pi}) + \Sigma\,(x_i - \hat{\mu})^2 / (2\hat{\sigma}^2)$$

We could solve this numerically by plotting $L_{(x)}$ against $\hat{\mu}$ and finding the value of $\hat{\mu}$ which maximizes L. In complex problems, this can be the only way we can proceed, although the search for the maximum is carried out by computer using algorithms designed for this purpose. In this case, however, we can solve the equation. The maximum value of $L_{(x)}$ is given when

$$d\,(L_{(x)})/d\,\hat{\mu} \quad = \quad 0$$

The differential is found using the "function of a function" rule:

$d\,(\Sigma \ln(1/(\hat{\sigma}\sqrt{2\pi}))\,/\,d(\hat{\mu})$ is zero because it has no terms in $\hat{\mu}$

For the second term, setting $(x_i - \hat{\mu}) = z$ and remembering $dy/d\hat{\mu} = dy/dz.\ dz/d\hat{\mu}$

$$d\,\Sigma z^2 / (2\hat{\sigma}^2) / d\,\hat{\mu} \quad = \quad \Sigma\, 2(x_i - \hat{\mu})\ / (2\hat{\sigma}^2)\ .\ d((x_i - \hat{\mu})\ / d\,\hat{\mu}$$

$$= \quad \Sigma\, -2(x_i - \hat{\mu})\ / (2\hat{\sigma}^2)\ .$$

Therefore:

$$d\,(L_{(x)})/d\,\hat{\mu} \quad = \quad -\,\Sigma\ (x_i - \hat{\mu})\ /\ \hat{\sigma}^2$$

Setting this to zero:

$$0 \quad = \quad -\,\Sigma\ (x_i - \hat{\mu})\ /\ \hat{\sigma}^2$$
$$0 \quad = \quad -\,\Sigma\ (x_i - \hat{\mu})$$
$$0 \quad = \quad -\Sigma\, x_i + n\,\hat{\mu}$$

$$\hat{\mu} \quad = \quad \Sigma\, x_i / n \qquad \text{as we would expect.}$$

For the variance, differentiation is harder but still only requires the "function of a function rule"

$$d\,(L_{(x)})/d\,\hat{\sigma}^2 \quad = \quad \Sigma\, -\hat{\sigma}\ (\sqrt{2\pi})\ \hat{\sigma}^{-2} / (\sqrt{2\pi}) \quad - \quad \Sigma\, [(x_i - \hat{\mu})^2 /2](-2\,\hat{\sigma}^{-3})$$

$$n\,\hat{\sigma}^{-1} \quad = \quad \Sigma\, [(x_i - \hat{\mu})^2 /2]2\,\hat{\sigma}^{-3}$$

$$n\,\hat{\sigma}^2 \quad = \quad \Sigma\, (y_i - \hat{\mu})^2$$

$$\hat{\sigma}^2 \quad = \quad \Sigma\, (y_i - \hat{\mu})^2 / n$$

So we have the maximum likelihood estimates. ML approaches will find the parameters for other p.d.f. too. An easier one to try is to find the ML estimate of *p* for the binomial distribution (you need the "function of a function rule" and to remember that d(log(*x*)/d*x* = 1/*x*). Often it is not possible to solve the ML equations algebraically as here. However, if you can write the likelihood down, then you can usually find the ML estimates numerically, sometimes even in Excel, as we shall see.

The ML estimate of the variance is biased. The bias arises because the ML solution to the variance treats the mean as known. But if the mean is unknown and has also been estimated from the data, then the unbiased variance is no longer

$$\hat{\sigma}^2 \quad = \quad \Sigma\, (y_i - \mu)^2 / n$$

but

$$\hat{\sigma}^2 \quad = \quad \Sigma\, (y_i - \hat{\mu})^2 / (n-1)$$

The sum of squared deviations is divided by the degrees of freedom. The degrees of freedom are n minus the number of parameters estimated: one here because we have estimated the mean. If n is large, the bias is small. Generally, when analyzing data, we fit more that just a mean and include parameters to account for varieties and other factors. The unbiased estimate of error remains the (sum of squares) / (degrees of freedom) or SSQ/df but the df for error is now (n-k) where k is the number of parameters estimated (including the mean). In this case SSQ are deviations from the fitted model, not just from the mean.

Even after dividing the variance by (n-1), there remains a very slight bias which is usually ignored but can occasionally crops up when working with very small populations.

Suppose that we sampled the whole population. We can now calculate the mean and variance exactly, rather than merely estimate them from a sample. In this case, the divisor for SSQ should be n, and not n-1. Now suppose we sampled all but a single member of the population. Intuitively, it seems unfair to reduce the df from n to n-1 just for a single observation – surely the bias in the estimate of variance can't be that great. This is correct. If the population size is N and the sample size is n, the unbiased estimate of error is

$$\hat{\sigma}^2 \quad = \quad (SSQ/n)\ [n/(n-1)][(N-1)/N]$$

Written this way, the unbiased estimate is the ML estimate (ie divide by n) multiplied by n/(n-1) to correct for the estimation of one parameter (the mean) then multiplied by (N-1)/N to correct for the proportion of the population we have sampled. N will usually be so large that this correction isn't worth bothering about. If N=n, then we have sampled the whole population and the ML estimate is unbiased.

Bayesian Estimation.

"Proof," I said, "is always a relative thing. It's an overwhelming balance of probabilities. And that's a matter of how they strike you." Raymond Chandler. Farewell, My Lovely

Another view of ML estimation is derived from Bayes' theorem. Bayesian statisticians have fewer hang-ups about talking of maximizing the probability of parameters. I describe the approach superficially here. We are not going to get into Bayesian statistics, but you will come across many references to it. Remember Bayes' theory is:

$p(A \mid B).p(B) = p(B \mid A).p(A)$

Suppose A are the parameters we want to estimate and B are the data.

$$p(\text{parameters} \mid \text{data}).p(\text{data}) = p(\text{data} \mid \text{parameters}).\ p(\text{parameters})$$

p(data) is fixed, it doesn't depend on anything else and is treated as a "normalizing constant" – a constant that balances the equation.

$$p(\text{parameters} \mid \text{data}) \quad \propto \quad p(\text{data} \mid \text{parameters}).\ p(\text{parameters})$$

$$p(\text{parameters} \mid \text{data}) \quad = \quad l\ (\text{parameters} \mid \text{data}).\ p(\text{parameters})$$

p(parameters | data) is called the *posterior probability* of the parameters. It is not the same as the likelihood of the parameters given data.

p(parameters) is the prior probability of the parameters.

l (parameters | data) is the likelihood of the parameters and can be viewed equal to, or at least proportional to p(data | parameters).

So the posterior probability of the parameters is the likelihood of the parameters given the data multiplied by the prior probability of the parameters. The data can be regarded as modifying our prior beliefs. The means of the parameters from the posterior distribution are the "Bayesian point estimators."

The contentious issue with Bayesian statistics is whether we should include p(parameters) or not. If we ignore it, or treat all possible parameter values as equally likely, then maximizing p(data | parameters) and p(parameters | data) give identical results so maximum likelihood estimation and Bayesian estimation is equivalent. If we have strong priors, then the results can be quite different. Much of the time, fortunately, they give near identical answers. (Part of the reason for this is that Bayesian statisticians, for all their enthusiasm, often use something called "weak priors.") The strength of Bayesian statistics is that it takes account of the fact that, to quote one its great advocates "it is impossible to know nothing about anything."

As an example consider the coin tossed 92 times and coming down heads each time in Tom Stoppard's play "Rosencrantz & Guildenstern are Dead." The maximum likelihood estimate of p(heads) is therefore zero. However, they started the coin tossing with some prior belief (seemingly different for each of the two characters) that the coin was fair. In this simple case, the strength of this belief can be quantified through "pseudocounts." These are (not necessarily imaginary) counts of heads and tails outcomes in previous coin tossing experiments. Suppose we select pseudocounts of 500 heads and 500 tails. A simple modified estimate of the probability that the coin turns up heads (the posterior probability) is

p(heads) = (500 + 0) / (1000 +92) = 0.46

If we were less confident a-priori that the coin was fair, we might set our pseudocounts to 5, in which case we the posterior probability is

p(heads) = (5 + 0) / (10 +92) = 0.05

The posterior probability is small but is still not zero. The elimination of such zero estimates can be viewed as a strength of the method. A frequentist (ie non-Bayesian) approach to avoiding an estimate of zero could be to search for the value of p(heads) which just rendered the observed result non-significant at some value. For a 5% level of significance this corresponds to p(heads) = 0.04.

The pseudocounts can also be used to define the prior probability distribution of p(heads) from which we think our coin has been sampled. If the prior was binomially distributed, the pseudocount of 500 would imply a standard deviation (sqrt(pq/N) of 0.0158: we expect that in nearly all cases that our coin has a true p(heads) close to 0.5. With the

pseudocounts of 5, the standard deviation is 0.158: so we should not be too surprised to find p(heads) departing from 0.5.

What is clear is that the final answer depends very strongly on the prior probability, in addition to the observed data. In the case of the "weak prior" (pseudocounts of 10) the estimate is not far removed from the simple maximum likelihood estimate of 0.

In fact, the prior distribution generally used in Bayesian estimation of binary outcomes is the Beta distribution, which we have not encountered before, though it is closely related to the binomial. It ranges from 0 to 1 but within those limits can take on a range of quite different shapes, from U shaped to near normally distributed. It is characterised by two parameters, a and b. For positive integer values of a and b, a-1 can be viewed as the pseudocount of heads and b-1 the pseudocount of tails. For Bayesian estimation of binary outcomes, it turns out that the use of a Beta prior means that the posterior estimate of p also has a Beta distribution, but with different values of a and b. These become

a = pseudocount of heads + observed heads – 1
b = pseudocount of tails + observed tails – 1

The mean of a Beta distribution is a/(a+b) so for our examples the posterior estimates are:

p(heads) = (501+92) / (1002 +92)
p(heads) = (11 +92) / (22 +92)

These are very similar but not identical to those calculated previously.

This simple (?) example shows one the strength of Bayesian analysis in that it can incorporate prior knowledge and a weakness in its sensitivity, at least in this case, to the distribution of the prior. A further weakness for many (including me) is that it is hard to understand. However, the approach is opening up avenues of analysis for very large datasets, complex problems and datasets with more parameters than observations, that are less amenable to other methods.

Least squares

Historically this method precedes ML, and is much used for data for which error can be approximated by the normal distribution. It is also probably the easiest of all methods to understand. Least squares does what it says on the tin – it minimizes the squared deviation between observed and expected values. So to estimate the mean, E(y) we want to minimize $\Sigma(y_i - Ey_i)^2$

Differentiate this, set the answer to zero and solving gives E (yi/n) as for ML estimation. The variance here is left to look after itself – it is the variance we are minimizing in estimating the mean.

Error and confidence limits

Now we know how to estimate parameters, we ought to know how accurately that mean is estimated. We know that the variance of the mean is just the variance of the sample/n. We can also demonstrate that for a normal distribution, 95% of observations will lie within +/- two (more precisely 1.96) standard errors of the mean.

So we can assign 95% confidence limits to the mean as:

$\hat{\mu}$ + 1.96 s.e and $\hat{\mu}$ - 1.96 s.e

Although everyone quotes confidence limits, there is confusion, if not controversy, about how they should be interpreted. We can't say that we are 95% certain that the true value lies in this interval, although that is very often the impression given. All we can really say is that if we repeated this experiment multiple times, we would expect 95% of the repeat estimates to lie in this interval. Don't worry about it.

Estimation in more complex cases – regression and the analysis of variance

We want to fit a straight line of the usual form y = ax +c.

For reasons which we become apparent shortly, we'll label this as:

$y = b_0 + b_1 x$.

Suppose y = farm yield, x = fertilizer added.

b_0 is then the expected yield with no fertilizer (the intercept)
b_1 is the regression coefficient to translate added N into the predicted increase in yield.

We have an observed yield and a yield predicted or estimated from our regression equation. The observed yield will not generally be perfectly predicted by the regression. To account for this we add an error term, almost always designated as *e*.

observed: $y = b_0 + b_1 x + e$
expected $\hat{y} = b_0 + b_1 x$

We can use maximum likelihood or least squares (for normally distributed variables) to find the best fitting line. Since most of the time we can get away with treating our data as normally distributed we shall use least squares here. We wish to minimize the error sums of squares, totalling over all our observed data:

$$\Sigma(y - \hat{y})^2 \quad = \quad \Sigma (y - b_0 - b_1x)^2 \quad = \quad \Sigma e^2$$

We want the estimates of b_0 and b_1 which minimize Σe^2

$$d [\Sigma (y - b_0 - b_1x)^2] / d b_0 = 0$$
$$d [\Sigma (y - b_0 - b_1x)^2] / d b_1 = 0$$

Start with b_0, using the "function of a function" rule again (or you can multiply out the brackets).

$$-2\Sigma (y - b_0 - b_1x) = 0$$

$$\Sigma b_0 = \Sigma y - \Sigma b_1x$$

$$n b_0 = \Sigma y - b_1 \Sigma x$$

$$b_0 = \Sigma y/n - b_1 \Sigma x /n$$

We need to know b_1 to estimate b_0

b_1:

$$-2\Sigma x (y - b_0 - b_1x) = 0$$

$$-2\Sigma xy + 2b_0 \Sigma x + 2b_1\Sigma x^2 = 0$$

substitute for b_0

$$-\Sigma xy + (\Sigma y - b_1 \Sigma x)/n \, . \, \Sigma x + b_1\Sigma x^2 = 0$$

$$-\Sigma xy + \Sigma y \Sigma x/n - b_1 \Sigma x \Sigma x/n + b_1\Sigma x^2 = 0$$

$$b_1\Sigma x^2 - b_1 \Sigma x \Sigma x /n = \Sigma xy - \Sigma y \Sigma x /n$$

$$b_1(\Sigma x^2 - \Sigma x \Sigma x /n) = \Sigma xy - \Sigma y \Sigma x/n$$

$$b_1 SSx = SPxy$$

$$b_1 = SPxy / SSx$$

The linear regression coefficient, b_1, is the sum of products / sum of squares of x, or equivalently, Cov(xy) / Var (x). Once we have estimated b_1 we can estimate b_0 by substitution. In fact, if this substitution is made, the regression equation can be rearranged as:

$$y - \bar{y} = b_1(x - \bar{x}) + e$$

a regression of $y - \bar{y}$ on $x - \bar{x}$ which has identical slope but passes though the origin.

Note that for least squares estimation, it is the distribution of *e* that matters and not of *y*. In fact, depending on the magnitude of the *b*s and *x*s, the distribution of *y* could be bimodal, or anything. This is important and often causes confusion – it is the distribution of errors that we are concerned with and on which the assumptions of our estimation methods are based. (The same is true for ML estimation – the p.d.f. we adopt is that for the errors, not for the effects we are interested in estimating.

In matrix form, the regression can be written as:

$$\mathbf{y} = \mathbf{Xb+e}$$

y' is the observed data $[y_1, y_2, y_3 \ldots y_n]$

$$\mathbf{b}' = [b_0, b_1]$$

X is a matrix of two columns with the independent variables

$$\begin{matrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{matrix}$$

etc.,

The first column of 1's is the coefficient by which the value of the intercept (b_0) will be multiplied, just as *x* is the value of the coefficient for the regression (b_1).If we set the first column to **0** or left b_0 out of the model, we would be fitting a regression line which is forced through the origin of the graph (the point 0,0 – forcing yield to be zero when no N is added). In this example we don't want to do that but there are instances when it can be biologically meaningful to do just this.

We want to minimize:

$$\mathbf{(y- Xb)^2} \quad = \quad \mathbf{e^2}$$

nb – $\mathbf{e^2}$ = **e'e** is the error sum of squares

Multliply out

$$\mathbf{y'y-b'X'Xb-}2\mathbf{b'X'y} \quad = \quad \mathbf{e^2}$$

differentiate with respect to **b** and set to zero

$$-2\mathbf{'X'Xb-}\ 2\mathbf{X'y} \quad = \quad \mathbf{0}$$

$$\mathbf{X'\ Xb} \quad = \quad \mathbf{X'y}$$

$$\mathbf{(X'X)^{-1}X'Xb} \quad = \quad \mathbf{(X'X)^{-1}X'y}$$

$$\mathbf{b} \quad = \quad \mathbf{(X'X)^{-1}X'y}$$

For just one or two parameters in **b** this looks no more simple than writing out the equations long hand. However, with several parameters (b_0 , b_1 … b_n) each with an associated column in **X**, the matrix method is more simple and concise.

Once we have fitted our model, we can get the estimate of error variance by

$$\mathbf{(y- Xb)^2} \ / \ \mathrm{df}$$

Note, this also gives us a procedure for testing how well our model fits – we can add or drop columns to **X,** refit the model, and compare the change in sum of squares of the error variance before or after adding columns. This is the basis of the analysis of variance table. Note that the columns of **X** do not have to be uncorrelated – the correlation is taken into account in the **X'X** term. However, if a pair of columns is identical, the model fitting will fail – you are trying to fit the same thing twice. This is a special case of linear dependence – one column is a simple linear function (involving addition and multiplication only) of the other columns. In this case the model fitting will fail too. That aside, you can't have more columns in **X** than you have data in **y.** Each column accounts for 1 df. The first column, all 1's, fits the mean effect. You do not get off scot-free if the columns of **X** are correlated, however. In this case, the order in which a column is added makes a difference to the change in sums of squares associated with the corresponding parameter. If b_1 is added before b_2, then the sum of squares associated with b_2 is that *after* the effect of b_1 has been accounted for and can be smaller than the sum of squares for b_2 if it is fitted before b_1. As a result, the parameters fitted first in a model can sometimes appear more important than they should. However, the total change in sums of squares due to fitting b_1 and b_2 is identical whatever order they are fitted. There is an example of

this in the R tutorial. When **X** represents a design matrix from a planned experiment, then the columns are frequently independent and this problem does not arise – an advantage of a designed experiment.

If, instead of an independent variable like N application, we want to study the effect of colour (of seed say) or the effect of variety, we still proceed in exactly the same way. Suppose we have two varieties and want to model the difference in yield between them. Our first attempt at the design matrix, **X,** might be:

mean	V1	V2
1	1	0
1	0	1
1	1	0
…	…	…

In the first three rows we have two observations on variety 1 and one on variety 2. Any observation is going to be on either V1 or V2. This doesn't work - the columns of **X** are not independent: the 'mean' column is the sum of columns V1 and V2. We know that if the variety isn't V1 then it must be V2. So one column is redundant. We can delete it column or more conventionally, we fit:

mean	V1
1	1
1	-1
1	1

V2 is now estimated as - V1.

With three varieties we would start with:

mean	V1	V2	V3
1	1	0	0
1	0	1	0
1	1	0	0
1	0	0	1

which also won't fit as V1+V2+V3 = mean. So we estimate V3 as -V1-V2:

mean	V1	V2
1	1	0
1	0	1
1	1	0
1	-1	-1

In the same manner, with more varieties we would equate the last variety to -(sum of all the other variety effects).

The reduction in the error sums of squares attributable to the inclusion of varieties in the model is the difference in fit between including the modified columns of **X** for V1 and V2 and dropping them, with 2 df accounting for the effect of three varieties.

Although for balanced designs and in standard cases, one would not normally analyse the data in this way, for non-standard analyses this approach can be best, especially if you are carrying out the calculations yourself. Statistical packages have very general analysis algorithms which are more sophisticated than the method outlined here, but they are much closer to these matrix methods than to the standard methods for hand calculation given in old statistical text books.

Power, significance, and multiple testing

Power, in the statistical sense, is a word frequently used and little understood. It is related to statistical significance, which is often also misunderstood. Hopefully these notes will help.

Significance

"Old statisticians never die they just become non-significant."

The null hypothesis
Classical (frequentist) statistics assumes – at the start of the experiment – that there is no effect of the treatment(s) for which you designed the experiment to detect. This is a convenient fiction. If you really believed this, you wouldn't bother to do the experiment.

This assumption of no effect is called "the null hypothesis" (NH), sometimes abbreviated to just "the null". In a variety trial, the NH would be that all the varieties are the same. In genetical experiments it could be that some binary trait (eg short/tall or susceptible / resistant) is under simple genetic control – so we would expect a 3:1 ratio in an F2.

After we have analysed the results of our experiment, we may decide to reject the NH – and conclude, for example, that several of our new varieties yield more than the existing standards, or we may accept the null-hypothesis – the ratio of short : tall plants is in agreement with a 3:1 segregation ratio. Owing to noisy measurement and assay, particularly in biological experiments, unexpected results do sometimes occur. A fair coin tossed ten times will sometimes come down heads all ten times (1 in 1024 times in fact). In this particular example, most people, after five or six tosses coming down heads, would begin to smell a rat and would reject the null hypothesis and believe the coin to be biased. However, even after 10 such occurrences, they may be wrong: they were just unlucky. On average, with a fair coin, they would be unlucky 1 in 1024 times. That is to say, they would have falsely rejected the NH that the coin was fair. (For the Bayesian view of biased coins, see the earlier section on Bayesian estimation.)

Significance is just the probability of rejecting the null hypothesis when it is true. (1/1024 in the example above). It is the probability of a false positive when the null hypothesis is true.

A false positive is also called a type I error or an error of the first kind.

Note that significance doesn't mean that over a lifetime of coin tossing experiments that we should on average observe 1/1024 ten-times-heads-up experiences. This would only be the case if all the coins we encounter were fair. Depending on the company we keep, some proportion of them could be biased, so what we observe over our lifetime could be greater than 1/1024.

The value of 1/1024 in this coin tossing exercise is called the significance level of the experiment. We can choose whatever level we wish - it depends entirely on how desperate we are to find a genuine difference (on which more below) and how damaging we feel rejecting the null-hypothesis falsely might be. However, conventional levels are 5% (significant), 1% (highly significant) and 0.1% (very highly significant). These levels are abbreviated by convention as *, ** and ***. These standards were adopted in pre-computer days. Nowadays, probability can be calculated directly in Excel, or in statistical software. It is better to report the exact probability of rejecting the null hypothesis when true (abbreviated to p-value). This allows the reader to decide for him/herself how much faith to put on the results. An easy way to annoy a statistician is to insist that "there is no effect" when the p-value is 0.055 but that "there is an effect" at 0.045.

To get their revenge, the significance level is sometimes referred to by statisticians, but by nobody else, as "size" – as in "the size of the test" – which is easily confused with the size of the experiment, which is something else entirely.

Note that there is no requirement to describe in any way what the alternative to the null hypothesis might be. When the null hypothesis is rejected, all we need to say is that it seems improbable. Not only is the choice of significance level entirely in the hands of the experimenter, but so too is the interpretation of the result after the NH is rejected.

Power

Statistical power is generally understood to represent how likely it is that an experiment will detect the effect you are searching for: how likely are you to discover that a set of varieties are genuinely different in yield. This interpretation is broadly correct, but the definition is quite precise:

Power is the probability of accepting the alternative hypothesis when true.
It is (1 – probability of rejecting the alternative hypothesis when it is false.)
It is (1 - the probability of a false negative.)

A false negative is termed a “type II error” or an error of the second kind.

Note we are now talking about an “alternative hypothesis.” To make any statement about power, we must first have in mind an alternative hypothesis in addition to the null hypotheses and have fixed a desired significance level. The alternative hypothesis is usually expressed as the magnitude of the difference or differences that we are hoping to find.

For example, we may want to detect a difference between a candidate variety and a control variety of 8% in yield. A variety showing this much improvement in yield will earn our employer top dollar, so we want our proposed experiment to be 95% certain to detect such a variety. We are not too worried about rejecting the NH falsely - in which case we would believe our candidate variety to be much improved over the control when in truth it has much the same yield – so we will select a significance level of 10%. This means that of the subset of varieties which are not different to the control, we are prepared to submit 10% to the official testing authority. (We can only do this because the fees charged by the testing authority are so reasonable.) So we have selected a 10% significance level and 95% power. We can then design an experiment on a sufficiently large scale which meets these objectives. If we drop the desired power, we would get away with a smaller experiment. If we increased the significance level to 1%, we would need a larger experiment: power cannot be calculated or expressed without an explicit statement of the adopted significance level.

Note that we cannot make statements in advance of the experiment about the likelihood of any selected line genuinely being 8% better than the control. This depends on the distribution of lines to be tested, which we usually do not know. If sufficient lines are tested, we may be able to make statements after the experiment is completed about the proportion of selected lines which are genuine improvements and the proportion for which we falsely rejected the null hypothesis (a “false discovery”). This is discussed later.

Note that examples of power above all refer to the planning of experiments, and it is in this context that power is best talked about – what size of effect are you interested in detecting, and therefore how large an experiment you should conduct. In general, plant breeders rarely do this sort of thing. This could be because they don’t know how to, or it could be that after virtually a century of scientific testing of improved varieties, there is an accurate but largely empirical feel for the correct scale of testing. In contrast, in research in medical genetics, much effort is put into consideration of power before an experiment is started. After an experiment has been carried out, then the results are statistically significant or not and power calculations are less interesting. An exception to this is when a statistician is called in to carry out a post-mortem – why did it fail to detect the hoped for effect. This can be very instructive – no significant result (at the 5% level say) was found because in this experiment the candidate variety would have to yield 35% more than the control.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." RA Fisher

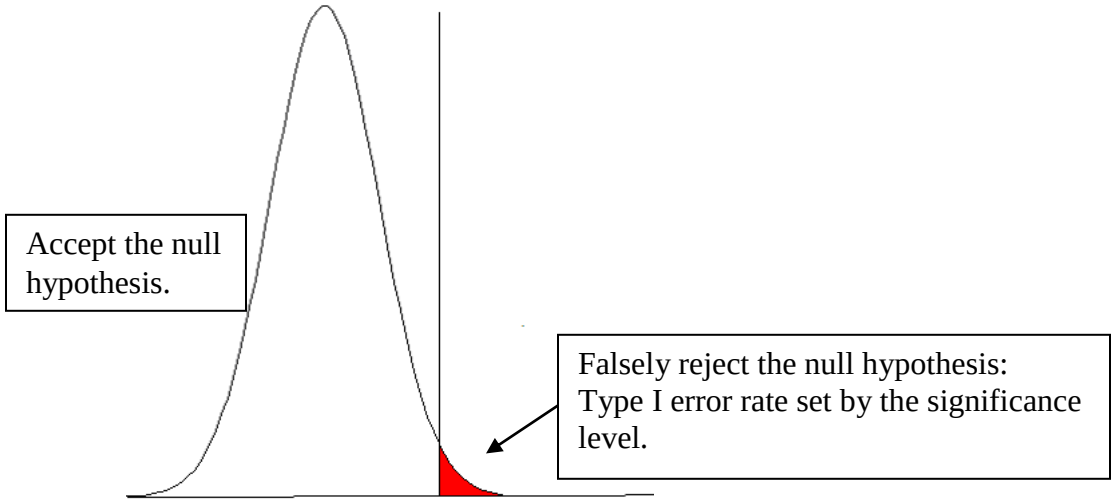
A good on-line account of power and significance, together with an excellent interactive display of their interrelationship is given at:

http://www.intuitor.com/statistics/T1T2Errors.html

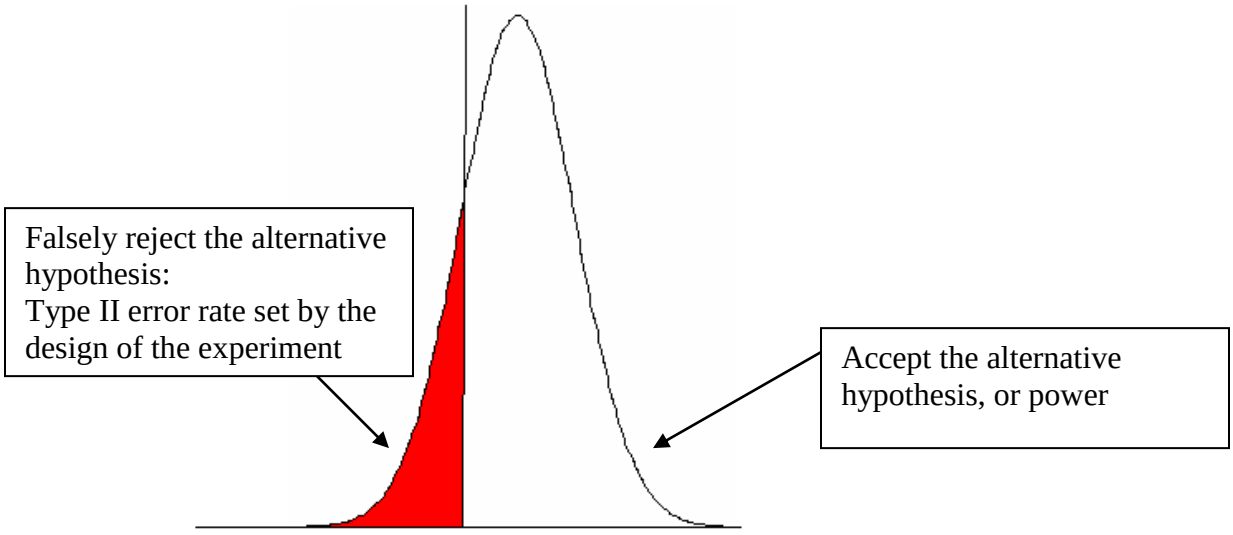
http://www.intuitor.com/statistics/CurveApplet.html

A more simple illustration version is given below.

Distribution of trait under the null hypotheses.



Distribution of trait under the alternative hypotheses.



In this example, the two distributions might represent the yield of plants carrying alternative forms of a QTL. We wish to select plants carrying the increasing allele. In the absence of QTL or marker information, a plant is classified as carrying the increasing allele if it falls to the right of the vertical line and as carrying the decreasing allele if it falls to the left. Probabilities of rejecting the null hypothesis and the alternative hypothesis when they are true and given by the shaded areas. The location of the vertical line is entirely in the experimenter's hands.

The use of power

The principal use of power should be in the design of experiments. For example, in a mapping experiment, we could calculate the size of the F2 population required to detect a QTL accounting for 10% of the phenotypic variation, at a genome wide significance level of 0.01 with 80% power. If we established that we needed to grow 50,000 plants, we would consider the experiment uneconomic and abandon it. If we established we would need to grow only 50 plants, we would move forward with confidence. Somewhere in the middle, we might find that we needed 1000 plants, but if we accepted a lower significance level, maybe 0.05, then power would remain acceptably high.

In practice, for any experiment, once any two of significance, power, and magnitude of effect are fixed, the third can be calculated. We may also require an estimate of the magnitude of error – most probably from similar previous experiments. For yield trials, this is not usually a problem. We can therefore use power calculations to design experiments more rationally – in terms of the size of populations required for mapping populations or the number of replications that need to be grown in a field trial.

Power calculation can also be used after an experiment has been completed (typically after it has failed to detect any significant effect). Here, retrospective calculations, using the observed error variation can give some idea about how large an effect we might reasonably have expected to see. If this effect is very large, then the experiment was too small – it was under-powered and may need to be repeated.

Calculation of power

For tests comparing two normally distributed means, power can be fairly easily calculated using a spreadsheet. The example below will calculate the power to detect a difference in the two means given in the second and third columns of the top row, with a variance given in the second row. The significance level is also entered by the user and power is calculated – in this case it is 52% for a 2.5% significance level. The formulae that do the work are also displayed. Although these look complicated, all they are doing is computing the appropriate areas of the normal distributions shown in the diagrams on the previous page.

	null	alternative	
mean	0	2	
variance of mean	1		
significance value	0.025		
cut-off	1.960		NORMINV(1-B4,B2,SQRT(B3))
power		0.516	1-NORMDIST(B6,C2,SQRT(B3),TRUE)

This spreadsheet is strictly only correct for large sample sizes – such that testing for significance using a normal distribution is valid. Strictly, we should be using a t-distribution. The effect that we wish to detect – here the difference between the two means, is called the "non centrality parameter." In R, this can be fed directly into "pt": the command for calculating probabilities under the t distribution. (If you are unfamiliar with R, consult the R tutorial.) Using the example above, but first assuming a very large number of observations:

1) Calculate the threshold for significance for our desired significance level:

```
qt(0.0250,df=1000000,lower.tail=F)
[1] 1.959966
```

2) Calculate the probability of exceeding this threshold in a t test with a true difference between means of 2.0 (ie the non centrality parameter is 2.0)

```
> pt(1.96,df=100000,ncp=2,lower.tail=F)
```

```
[1] 0.5159552
```

Thus we get the same answer as in Excel, as we should.

We can now drop the degrees of freedom to something more realistic:

```
> pt(qt(0.0250,20,lower.tail=F),df=20,ncp=2,lower.tail=F)
[1] 0.4775688
```

In this case, there is little difference in power between using the t distribution with 20 degrees of freedom, and using a normal distribution.

The chi-squared distribution also has a non centrality parameter. This can be used to calculate power in contingency chi-squared tests. If the usual chi-squared test is calculated using the made-up numbers that you expect under the alternative hypothesis, the value is the non-centrality parameter and can be used to calculate power. For example, suppose we suspect segregation distortion at a locus and want to test for this in a backcross of 100 individuals. Under the null-hypothesis we expect a 1:1 segregation. Suppose the true segregation pattern were 6:4 rather than 1:1. Then we would expect to observe 60 offspring of one type and 40 of the other. This gives a chi-squared value of $(60\text{-}40)^2/100 = 4$ In R:

```
> pchisq(4,1,0,lower.tail=F)
[1] 0.04550026

> pchisq(4,1,4,lower.tail=F)
[1] 0.5000317
```

The first call to `pchisq` returns a p-value close to 0.05, the correct p-value for a chi-sq of 4.0 with 1df. The third term in the `pchisq` function is the non-centrality parameter, zero by default but declared explicitly here.

The second call to `pchisq` returns a p-value of 0.5, calculated from a chi-squared distribution with a non-centrality value of 4. This makes sense: the expected segregation distortion gives a chi-squared statistic of 4, so it seems reasonable that when carrying out such experiments for real, half the time we would expect our result to exceed this threshold and half the time to be smaller, giving us the p-value of 0.5. (Note that this is not the mean chi- squared statistic, but the median. The mean is the non-centrality parameter plus the degrees of freedom: 5 for our alternative hypothesis and 1 under the null-hypothesis.)

R is even more helpful, since commands are supplied – `power.t.test` and `power.prop.test` which can be used directly to calculate power, significance and sample size. Consult the R manuals or type `help(power.t.test)` or `help(power.prop.test)` for details. The “odds-and-sods” spreadsheet also has a workbook to calculate power for 2x2 contingency chi-squared tests.

A conceptually simple way to calculate power is by computer simulation, and increasingly this approach is used. For more complex situations, for example establishing significance levels for genome wide linkage analyses or for cases where the distribution of the test statistic itself is not known, this may be the only accurate method. A simple simulation example is given below.

Consider rogue observations from a normally distributed population with a mean of 2 and a variance of 1. We wish to discriminate these from a normally distributed population with a mean of 0 and variance of 1. With a one-tailed significance level of 2.5%, the threshold to be exceeded is 1.96 and this will occur for close to 50% of the rogue observations. What is the effect of power if the rouge observations are log-normally distributed with the same mean and variance?

1) Generate 100,000 log-normal random numbers with a mean of two and a variance of 1.

```
> a<-(rnorm(100000,0,1))                    generate N(0,1) observations
> b<-exp(a)                                 b is log normal
> c <- (b-mean(b))/(sqrt(var(b))) +2        set mean =0, var = 1
> mean(c)
[1] 2                                       correct
> var(c)
[1] 1                                       correct
```

2) Count the proportions which are equal to, or exceed the threshold:

```
> length(a[a>=1.96])/100000          significance  threshold is correct
[1] 0.02502

> length((a+2)[a+2>=1.96])/100000    power  if rogue distribution is normal
[1] 0.51719

> length(c[c>=1.96])/100000          power if log normal
[1] 0.32833
```

So power is reduced from about 50% to 33%.

Suppose the null distribution was log normal too. The 2.5% significance level, for a log-normal population with a mean of zero and a variance of one can be found from our simulated data as:

```
> quantile(c-2,0.975)
   97.5%
2.481569
```

(Using c-2 to recycle our log-normal random numbers by adjusting the mean to zero.)

Power is the proportion of times this threshold is exceeded under the alternative hypothesis:

```
> length(c[c>=2.544928])/100000
[1] 0.14846
```

The answer is quite different!

Conventions for calculation

By convention, 5% significance levels are often chosen to declare statistical significance. Equally, 80% power is often selected as a suitable threshold for calculating sample sizes in experiments. The choice is entirely yours, however, and depends on a compromise between the costs to you of selecting a false positive and of rejecting a genuine result.

Multiple testing

Significance levels are thresholds whose choice is entirely in the hands of the experimenter and reflect the risk that he or she is prepared to make in rejecting the null hypothesis. A 5% significance level means that the null hypothesis, if true, will be falsely rejected 1 in 20 times. Thus, if one is looking for genetic association between a candidate marker and 20 independent phenotypes, none of which are genuinely associated with the marker, then the null hypothesis will be rejected, and a false association declared, on average, for one of these traits. In fact the number of falsely accepted associations will follow a binomial distribution, such that the probability of getting no significant results over all 20 tests is 0.95^{20} or 0.358. Thus, over the whole experiment, the probability of falsely rejecting the null hypothesis at least once is 1-0.358 or 0.642. To establish a 5% significance level over the whole experiment, we divide our single test significance level by the number of independent tests we are to carry out. In this example 0.05 / 20 = 0.0025. This is called the *Bonferroni correction*, after its discoverer. Thus, treating 0.0025 as the significance level for any single phenotype-marker association, then over all 20 phenotypes, there is only a $1-(1-0.0025)^{20}$ = 5% chance of finding one or more significant results. In other words, were we to repeat the experiment 20 times, we would expect to observe an experiment with at least one significant result only once. In fact, this is a slight approximation. A more exact adjustment is:

$$p_{test} = 1-(1-p_{expt})^N$$

p_{test} is the desired significance level over the whole experiment.
p_{expt} is the significance level to be calculated for a single test.
N is the number of independent tests.

In the present example, if P_{test} is set at 0.05, then P_{expt} turns out to be 0.002561.

This is the *Šidák* test, often falsely referred to as the Bonferroni, though as can be seen they give very similar results.

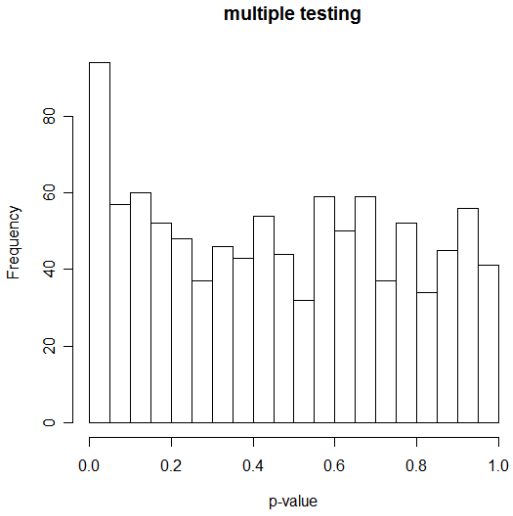
Although use of the Bonferroni correction will certainly guard against the generation of false positives, it has the unfortunate effect of making it extremely difficult for any single test to achieve significance, even if genuine.

The increased stringency required to declare any individual result significant greatly reduces the power of any single test. Moreover, in many cases in genetics, phenotypes and/or genotypes are not independent. In such cases, the Bonferroni correction can become too conservative to be of any use. For fear of declaring a false positive we find nothing. ("It's only those who do nothing that make no mistakes." Conrad.) This is particularly true in gene expression experiments where there can be thousands of tests, but there are correlations in expression among many pairs of genes. Equally, linkage disequilibrium among candidate polymorphisms, even if the phenotypes themselves are independent, has the same effect. There are two solutions to this. Firstly, with regard to correlations between traits and/or genotypes, we rely once more on simulation or permutation tests: in this case to establish the correct significance levels. The exact details will vary from case to case. For example, with a set of candidate genes and multiple phenotypes, the genetic data would be randomised over records (individuals or lines), while keeping the phenotypes fixed. (Equally phenotypes could be randomised and genotypes fixed). Note that the multiple genotypes for each record are maintained intact: the randomisation is of the complete set of genotypes across records. Thus the correlations among phenotypes and among genotypes are maintained, but the correlation between genotypes and phenotypes (if present) is broken by the randomisation. After each randomisation, tests for association across phenotype-genotype pairs are carried out, and the p-values saved. After multiple such randomisations, the empirical probability distribution of finding one or more significant result in a randomisation can be estimated, and compared to the observed results. To give an extreme example, if we have two perfectly correlated SNPs and two perfectly correlated phenotypes, then our Bonferroni adjusted p-value for a 5% significance level is 0.05/4. However, in essence we have only a single association test, replicated four times. The randomisation test would generate at least one significant result (in fact it would always generate four), in 5% of randomisations. Therefore, our empirical significance level for the whole experiment would be correctly adjusted back to 0.05. This is essentially the approach used in setting significance levels for genome-wide linkage analysis using randomisation tests. These can take a lot of computer time to run, but are often worth the effort.

Although randomisation tests overcome the problem of correlated data, they can do nothing about the loss of power which arises from multiple independent tests. Recent approaches have concentrated on learning to live with these through the concept of the *false discovery rate* (FDR). The FDR is the expected proportion of falsely rejected null hypothesis (type I errors) among all rejected hypothesis. If the null hypothesis is true for all the tests you have carried out in an experiment, then the false discovery rate will be 1.0 – all results accepted as significant are false. If the null hypothesis is truly false for a proportion of tests, then some proportion of the results accepted as significant will be

false positives. The FDR attempts to control this proportion, after the experiment is complete, by adjusting the p-value threshold for the proportion of results accepted.

The development of methods to measure FDR has been driven in part by the massive multiple testing problem generated in gene expression analysis using microarrays, and in part by the ease with which very large numbers of SNPs can be tested for association with phenotypes in human genetics. Essentially, these methods study the empirical distribution of p-values over all tests. If the null hypothesis is true for all tests, p-values should follow a uniform distribution: so a histogram of p-values should show bars of roughly the same height. An excess of low p-values would indicate the null hypothesis to be significant for at least some of the tests. The histogram below shows the p-values from 1000 1 df chi-squared tests, for 900 of which the null hypothesis is true. For the remaining 100 values, the null-hypothesis was false, with a power of 50% for rejection of the null-hypothesis at the 5% significance level. (These data were simulated in R.)



It is clear that there are excess results at low p-values. In this case, because we are using simulated data, we expect 45 of the 900 tests for which the null-hypothesis is true to be significant at the 5% level together with 50 of the 100 tests for which it is false. We observe 94 in total. If we were to increase our significance threshold to 0.01, then we would expect 9 false positive results and 27 genuine positives: a FDR of 25%.

In the absence of knowledge about how the data were generated, we can still see, from the plot above, that at high p-values, when the null hypothesis is most likely to be true in all cases, that the average number of observations in each 0.05 probability interval is about 45. We can therefore predict, at low p-values, that the number of false positives should also be about 45 in each interval and that any observed excess is due to genuine discovery. In this way, we could generate our own empirical FDR (also called the q-

value). Fortunately, there is an R package "qvalue" available to do this in a more quantifiable and exact manner. The manual describes the method in more detail and contains useful references: http://cran.r-project.org/web/packages/qvalue/qvalue.pdf

The use of FDR has allowed experimenters to select significance thresholds after an experiment is completed and therefore to control the proportion of false positive results that they select. This can remove the arbitrary nature of specifying significance thresholds. However, FDR requires that sufficient statistical tests are carried out in an experiment to estimate FDR accurately. Moreover, it does not avoid the desirability of assessing power prior to undertaking an experiment. (To assess FDR prior to an experiment, one would require knowledge of the distribution of the null, the alternative, and the mixing proportions of the two.)

An example in a QTL trait mapping context is given by Benjamini and Yekutieli (2005). This includes a simple description of one method of calculating the FDR: as originally described by Benjamini and Hocheberg in 1995.

Benjamini Y, Yekutieli D. 2005. Quantitative Trait Loci Analysis Using the False Discovery Rate. *Genetics* **171:**783-790

Another simple account is in: Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. and Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research* 125: 279–284

Benjamini, Y, Hochberg T. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**:289–300.

Type III errors

The original work and discussion of significance power and type I and type II errors was later extended to include many other error-type definitions. None of these have caught on, but one which seems to me to be particularly relevant to genetics, medical research, and even to the public understanding of science is the Type III error or error of the third kind. A type III error is the correct rejection of the null hypothesis but the acceptance of the wrong alternative hypothesis. This seems such a useful term to me, and to occur so often, that I don't understand why it hasn't been taken up more. Some examples:

An association between a marker and phenotype is attributed to close linkage of the marker to a QTL when the true reason is to do with population substructure.

The unlikely occurrence of two cot deaths in a family is attributed to infanticide.

And many more.

These sorts of error are still often referred to as false positives, but to my mind they are true positives, it is just the interpretation that can go wrong.

Final comments

A statistician shouts up to three men in a balloon to tell them they are lost. After hearing them grumble about his advice he works out they must be management:
1) They work out what information they need to get themselves sorted out.
2) They ask someone else to get it for them.
3) Now that they have the information, they are still lost but it's someone else's fault.

"Thinking that this single value
For the level in his serum
Might not be sufficient data
To establish without question
What the normal value should be,
Hiawatha with his cunning
Took a logarithmic table,
Photographed a page at random
For a lantern slide of figures,
Showed it very confidently
With his back towards the audience
Talking fast and very softly
At the figures thus projected
Which were very small and many
Like the sands upon the seashore;
And the audience, not hearing
What he spoke towards the blackboard
Very softly, very swiftly
Like the gentle brook in springtime,
Thought him wise and very clever
To have got so many figures
And their standard deviations,
Arithmetical progressions,
Geometrical regressions,
And regression coefficients;
Praised his industry, his brilliance,
And applauded his statistics,
For they had not understood him
Nor could read his logarithms."

From "Hiawatha's Lipid" Hugh Sinclair, 1958.

THE DESIGN AND ANALYSIS OF VARIETY TRIALS REVISITED

Estimation of variety performance under field conditions is the most important component of any plant breeding programme. Organisations that efficiently allocate their resources to identify varieties which are better than those of their competitors will generally last, organisations which don't will deservedly fail.

It is possible to carry out this process without any recourse to statistical methods. In practical terms, a breeder who identifies one or more experimental fields in which variety performance is indicative of performance in the target market or environment, and who grows a large enough area of each candidate variety together with the best currently available varieties, will make progress. The care and precision which are put into growing good trials are more important as the bells and whistles that are added by improved statistical design and analysis. Nevertheless, there are things that statistics can add:

An estimate of precision.

Control and adjustment for unforeseen (and foreseen) problems in the field.

An efficient way of allocating resources between area (or plot number) per variety and number of candidate varieties.

Experimental design: the three Rs

In an ideal world, one could grow a single plot of each variety and, on harvesting, be confident that the performance of each variety had been assessed accurately. This may sometimes occur, but we never know. Scientists who think they can assess experimental material with a single measurement can generally be found in physics, chemistry and engineering. They study dead things. The study of living things is not so easy but is a lot more interesting.

To overcome the problems inherent in measuring biological material, a series of sensible and intuitive, though sometimes subtle principles was put forward in the 1920s primarily by R A Fisher, then working at Rothamsted. These principles still stand. Fisher and others, in addition to developing standards for experimental design, also elucidated principles for the analysis of experiments. At the time, in the absence of readily available computers and calculators, a lot of effort was placed in creating protocols for designing experiments which could subsequently be analysed by hand. So successful were they that these designs are still frequently used. However, cheap and powerful computers have permitted the development of more flexible experimental designs and alternative methods of analysis. In this section we shall review the principles of experimental design and introduce some of these new methods.

The basic three principles of experimental design are replication, randomisation and restraint: the three Rs.

1) Replication.

Suppose we want to compare the yield of two wheat varieties. Starting by ignoring variability in the trial ground; we grow variety A and B side by side, harvest them and weigh the grain. If the yield of A is greater than the yield of B we have an answer. But the answer means little since we don't know how much of the difference in our measurements is due to field effects and how much is due to the varieties themselves. We need replication.

Suppose we harvest half of plot A and take its weight and then the second half of plot A and take a weight and then do the same for plot B. Now that we have two results for each variety, we have the basis of a crude assessment of how variable the measurements are between replicate plots for each variety. We can now assess (crudely) if the difference in average yield between the two varieties is greater than that between replicate plots of the same variety..

So the reason for replication is that we need an estimate of plot to plot variation which we can compare to the difference between the two variety means. In practice we would do this through statistical methods such as the t-tests or the analysis of variance. This is impossible with just two large plots: formally all degrees of freedom are taken up in the variety comparison and none are left to assess variability between replicate plots. Even with two plots of each variety we would have only two degrees of freedom with which to estimate error. Generally we need a minimum of around ten, and more are better.

2) Randomisation

"This is a glass house, it's uniform, there is no need to randomise." (A UK Ministry of Agriculture scientist, predictably but sadly subsequently promoted to a position of power and influence.)

Suppose we have split the field into 12 equal area strips running the length of the field. If we plant them as AAAAAABBBBBB, we may detect a genuine difference between the left hand side and right hand side of the field, over and above the error seen between replicate plots within each variety. But any difference in yield could equally be attributable to a difference in fertility across the field or to differences in variety. Planting the two varieties in this pattern is little better than having no replication at all.

Alternatives are to plant the varieties as ABABABABABAB, or to plant them in some random pattern – drawing letters from a virtual hat I got BBAAAABBABAB. We all know that the thing to do is randomise, but it isn't that clear why arranging the plots systematically as ABABAB… is wrong.

There is a subtle and a not so subtle explanation. The not so subtle explanation is that we don't know in advance what the pattern of environmental variation across the field is and we are kidding ourselves if we think we can guess it. There may be land drains, differences in soil compaction attributable to the direction of ploughing up and down the field, etc. It is unlikely that we could find a systematic arrangement of plots which could account for everything.

The more subtle argument is that although replication will provide us with an estimate of error, we need randomisation to ensure that the estimate of both variety effect and of error is unbiased. For example, suppose there were differences in fertility that followed the same pattern ABABABABABAB. If there was no true difference between the two varieties, we would still appear to find one, and the estimate of error would be too low (because the fertility effects have been sucked up into the estimate of variety effect). So not only would we falsely conclude that there was a difference between the two varieties, but we would be really confident that the difference was genuine because the error was so low. Equally, if we adopted AAAAAABBBBBB as a layout for the varieties, but the fertility pattern was ABAB… and we estimated error from the plot to plot variation within each variety, then we would overestimate the error, which is unfortunate because in this instance the estimate of the difference between the two varieties could be very accurate. To see this, consider an analysis of variance to test the difference between the two means – with 10 df for the error term. The expected value of F if there is no difference between varieties is 1.24 (Calculated by simulation in R. With sufficient df, the expected value would be 1.) But we have designed our experiment so the expected value of F is zero. The total sum of squares (SS) in the experiment is constant whatever the arrangement of plots and varieties. So this partitioning has been accomplished by pushing the SS for "between varieties" back into the error term, which is therefore inflated.

The only solution is to randomise – which provides an unbiased estimate of error.

3) Restraint

Also known as blocking or as local restraint or as restricted randomisation or as local control. In variety trials, and most agricultural trials, it is generally called blocking.

Staying with the example above, if we had some reasonable expectation that the two halves of the field might differ in fertility, we could treat the plots which lie in one half as one block, and the others as a second block. We could allocate the two varieties to plots so that they are equally represented in the two blocks. Within the blocks, however, the varieties must still be allocated to plots at random. We could analyse this experiment exactly as before, ignoring the blocks, and estimate variety effects and error. However, we can also calculate the difference between the blocks. Since each variety is equally represented in each block, this is an unbiased estimate of fertility effect between the two halves of the field. It will have a SS associated with it: the larger the block effect, the larger the SS. Comparing the two methods of analysis, for the same experiment the total SS must remain constant. Also, the variety means are identical in the two forms of analysis so the SS for varieties also must be the same. To balance the total, therefore, the

SS associated with plot to plot error must be reduced. As plot to plot error is the method by which we assess the precision of our variety effects, the introduction of blocking has increased precision by controlling error. The experimental design has partitioned sources of error into those between blocks, which have no influence on the precision of the estimates of variety effects, and the remainder or residual variation, which does. Blocking, therefore, is an experimental method for increasing the precision with which our effects of interest (here varieties) are estimated by partitioning field effects into a component between blocks, which has no effect on variety precision, and a residual which does. It doesn't necessarily work: we could have been wrong in our belief that the two halves of the field differed in fertility. In this case, there will be little or no reduction in the residual SS and consequently the precision with which variety effect are estimated will not change much either. (Precision can actually fall, since it is determined by the error variance and not the error SS. Blocking reduced the degrees of freedom available for error so the estimated error variance can rise even though the error SS falls.)

Formally, when we analyse our data, we now include a term to account for blocks, a term to account for varieties, and estimate the error variance by the deviations from predicted values, where the predicted value includes an effect for the block: we are fitting the model:

$$y_{ij} = \mu + v_i + b_j + e_{ij}$$

y is the yield of a given plot; μ is the mean for the whole experiment; v_i is the difference between the mean for the ith variety and the grand mean; b_j is the difference between the mean for the jth block and the grand mean and e_{ij} is the residual error specific to plot y_{ij}.

The expected error variance is $E(e_{i)}^2)$

In analysis without blocking, the model fitted is

$$y_{ij} = \mu + v_i + z_{ij}$$

and the expected error variance is $E(z_{i)}^2)$

Equating the two,

$$b_j + e_{ij} = z_{ij}$$

so it is clear, if block effects are genuine, that successfully incorporating them into the experimental design and analysis will reduce the error variance.

Although in this hypothetical experiment, the two blocks are of equal size, there is no requirement for this: we could have one block containing only four plots (two A varieties and two B) and one containing five. In fact, in these examples, because we have several

plots of each variety within each block, not only can we fit and test for effects of blocks and of varieties, we can also fit an interaction term – to assess whether the difference between the two varieties is consistent over the two blocks. We won't go into that.

Generally, we are uncertain when we look at a field where "natural" blocks should occur – where are the high and low fertility sections of the field. Also, there are rarely sharp discontinuities to indicate precisely where blocks should start and finish. An extremely common experimental design is to set up blocks such that each block contains each treatment (or variety) once only. This is a randomised complete block design. A small example of the design and analysis for six varieties in four replace blocks is given below.

Randomised Complete Block Design – example analyses

I've analysed this in four different ways. Firstly in a spreadsheet. Secondly, by matrix methods, and then using two different computer packages – R and GenStat – just to show you everything gives the same answer, although the formats tend to differ:

field layout and data

rep 1	rep 2	rep 3	rep 4
C	F	E	D
E	D	C	A
B	E	A	F
D	B	F	E
F	C	D	B
A	A	B	C

Unrandomised design with yield data:

Treatment/Rep	1	2	3	4
A	9.4	8.9	7	7.3
B	9.3	9.1	8.5	8.9
C	7.6	7.8	6.3	4.5
D	7.6	6.5	5.5	4.2
E	6.9	8.9	7.3	6.3
F	7.6	8	8.5	6.9

Excel:

Treatment/Rep	1	2	3	4	var. means
A	9.4	8.9	7	7.3	8.15
B	9.3	9.1	8.5	8.9	8.95
C	7.6	7.8	6.3	4.5	6.55
D	7.6	6.5	5.5	4.2	5.95
E	6.9	8.9	7.3	6.3	7.35
F	7.6	8	8.5	6.9	7.75
rep means	8.07	8.20	7.18	6.35	7.45

Item	df	SS	MS	F	P
Replicates	3	13.343	4.448	7.040	0.004
Variety	5	23.600	4.720	7.471	0.001
Error	15	9.477	0.632		
total	23	46.420	2.018		

Matrix analysis

Design matrix **X** and data matrix **y.**

Note the reduced design matrix, **X**, with
variety F = -(A+B+C+D+E) and rep 4 = -(rep 1 + rep 2 + rep 3)

X is columns 2..9 **y** is column 10.

plot	mean	A	B	C	D	E	rep 1	rep 2	rep 3	yield
1	1	1	0	0	0	0	1	0	0	9.4
2	1	0	1	0	0	0	1	0	0	9.3
3	1	0	0	1	0	0	1	0	0	7.6
4	1	0	0	0	1	0	1	0	0	7.6
5	1	0	0	0	0	1	1	0	0	6.9
6	1	-1	-1	-1	-1	-1	1	0	0	7.6
7	1	1	0	0	0	0	0	1	0	8.9
8	1	0	1	0	0	0	0	1	0	9.1
9	1	0	0	1	0	0	0	1	0	7.8
10	1	0	0	0	1	0	0	1	0	6.5
11	1	0	0	0	0	1	0	1	0	8.9
12	1	-1	-1	-1	-1	-1	0	1	0	8
13	1	1	0	0	0	0	0	0	1	7
14	1	0	1	0	0	0	0	0	1	8.5

15	1	0	0	1	0	0	0	0	1	6.3
16	1	0	0	0	1	0	0	0	1	5.5
17	1	0	0	0	0	1	0	0	1	7.3
18	1	-1	-1	-1	-1	-1	0	0	1	8.5
19	1	1	0	0	0	0	-1	-1	-1	7.3
20	1	0	1	0	0	0	-1	-1	-1	8.9
21	1	0	0	1	0	0	-1	-1	-1	4.5
22	1	0	0	0	1	0	-1	-1	-1	4.2
23	1	0	0	0	0	1	-1	-1	-1	6.3
24	1	-1	-1	-1	-1	-1	-1	-1	-1	6.9

X’ (omitted)

X’X

24	0	0	0	0	0	0	0	0
0	8	4	4	4	4	0	0	0
0	4	8	4	4	4	0	0	0
0	4	4	8	4	4	0	0	0
0	4	4	4	8	4	0	0	0
0	4	4	4	4	8	0	0	0
0	0	0	0	0	0	12	6	6
0	0	0	0	0	0	6	12	6
0	0	0	0	0	0	6	6	12

$(X'X)^{-1}$

0.04	0	0	0	0	0	0	0	0
0	0.21	-0	-0	-0	-0	0	0	0
0	-0	0.21	-0	-0	-0	0	0	0
0	-0	-0	0.21	-0	-0	0	0	0
0	-0	-0	-0	0.21	-0	0	0	0
0	-0	-0	-0	-0	0.21	0	0	0
0	0	0	0	0	0	0.13	-0	-0
0	0	0	0	0	0	-0	0.13	-0
0	0	0	0	0	0	-0	-0	0.13

X'y

179
1.6
4.8
-4.8
-7.2
-1.6
10.3
11.1
5

$(X'X)^{-1}X'y$

mean	7.45
A	0.7
B	1.5
C	-0.9
D	-1.5
E	-0.1
R1	0.62
R2	0.75
R2	-0.3

Aside from the mean, the other effects are given as deviations from the mean. Adding the mean back:

A	8.15
B	8.95
C	6.55
D	5.95
E	7.35
F	7.75
R1	8.07
R2	8.2
R3	7.18
R4	6.35

The error SS is given as $\mathbf{(y-Xb)}^2$ I won't print this out, but it gives the same answer as before.

The sums of squares for varieties are found as follows.

1) Take **X'X** and delete all rows and columns except those relating to the five independent variety effects, then invert it. (In this case, because the experiment is balanced, this gives exactly the same answer as striking out the same rows and columns of $\mathbf{(X'X)}^{-1}$ directly. With missing data this is not the case.)

$\mathbf{X}_{vars}'\mathbf{X}_{vars}^{-1}$

0.21	-0	-0	-0	-0
-0	0.21	-0	-0	-0
-0	-0	0.21	-0	-0
-0	-0	-0	0.21	-0
-0	-0	-0	-0	0.21

Then calculate $\mathbf{b}_{vars}'(\mathbf{X}_{vars}'\mathbf{X}_{vars}^{-1})\mathbf{b}_{vars}$

		-	-	-							
1.6	4.8	4.8	7.2	1.6		0.21	-0	-0	-0	-0	1.6
						-0	0.21	-0	-0	-0	4.8
						-0	-0	0.21	-0	-0	-4.8
						-0	-0	-0	0.21	-0	-7.2
						-0	-0	-0	-0	0.21	-1.6

= 23.6 as before.

A corresponding procedure will give the blocks SS and the ANOVA table can be constructed.

Genstat:

Data were pasted from excel into the GenStat spreadsheet, then all commands were selected from the pull-down menus.

GenStat Release 9.1 (PC/Windows XP) 30 January 2008 22:14:22
Copyright 2006, Lawes Agricultural Trust (Rothamsted Experimental Station)
Registered to: Nat. Institute of Agricultural Botany

GenStat Ninth Edition
GenStat Procedure Library Release PL17

```
1  %CD 'C:/Documents and Settings/x991006/My Documents'
2  "Data taken from unsaved spreadsheet: New Data;1"
3  DELETE [REDEFINE=yes] _stitle_: TEXT _stitle_
4  READ [PRINT=*; SETNVALUES=yes] _stitle_
7  PRINT [IPRINT=*] _stitle_; JUST=left
```

Data imported from Clipboard
on: 30-Jan-2008 22:18:19

```
   8  DELETE [REDEFINE=yes] treatment,rep,yield
   9  UNITS [NVALUES=*]
  10  FACTOR [MODIFY=yes; NVALUES=24; LEVELS=6;
LABELS=!t('A','B','C','D','E','F')\
  11  ; REFERENCE=1] treatment
  12  READ treatment; FREPRESENTATION=ordinal
```

Identifier	Values	Missing	Levels
treatment	24	0	6

```
  14  FACTOR [MODIFY=yes; NVALUES=24; LEVELS=4;
LABELS=!t('r1','r2','r3','r4')\
  15  ; REFERENCE=1] rep
  16  READ rep; FREPRESENTATION=ordinal
```

Identifier	Values	Missing	Levels
rep	24	0	4

```
  18  VARIATE [NVALUES=24] yield
  19  READ yield
```

Identifier	Minimum	Mean	Maximum	Values	Missing
yield	4.200	7.450	9.400	24	0

```
  22
  23  "One-way design in randomized blocks"
  24  DELETE [REDEFINE=yes] _ibalance
  25  A2WAY [PRINT=aovtable,information,means,%cv;
TREATMENTS=treatment; BLOCKS=rep; FPROB=yes;\
  26   PSE=diff,lsd; LSDLEVEL=5; PLOT=*; EXIT=_ibalance] yield;
SAVE=_a2save
```

Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
rep stratum	3	13.3433	4.4478	7.04	
rep.*Units* stratum					
treatment	5	23.6000	4.7200	7.47	0.001
Residual	15	9.4767	0.6318		
Total	23	46.4200			

Information summary

All terms orthogonal, none aliased.

Tables of means

Variate: yield

Grand mean 7.45

treatment	A	B	C	D	E	F
	8.15	8.95	6.55	5.95	7.35	7.75

Standard errors of differences of means

Table	treatment
rep.	4
d.f.	15
s.e.d.	0.562

Least significant differences of means (5% level)

Table	treatment
rep.	4
d.f.	15
l.s.d.	1.198

Stratum standard errors and coefficients of variation

Variate: yield

Stratum	d.f.	s.e.	cv%
rep	3	0.861	11.6
rep.*Units*	15	0.795	10.7

R:

For small amounts of data such as this, it is possible to use the "PopTools/Rscripts/Range to Dataframe" option of the Excel add-in PopTools (introduced in the Excel tutorial) to format and copy the data from Excel to the clipboard and then paste the data directly into the R Console. Otherwise, the data can be saved as a text file and read into an R table.

```
> treatment <-
c("A","B","C","D","E","F","A","B","C","D","E","F","A","B","C","D","E","F","A","B","C","D"
,"E","F")
> rep <-
c("r1","r1","r1","r1","r1","r1","r2","r2","r2","r2","r2","r2","r3","r3","r3","r3","r3","r
3","r4","r4","r4","r4","r4","r4")
> yield <-
c(9.4,9.3,7.6,7.6,6.9,7.6,8.9,9.1,7.8,6.5,8.9,8,7,8.5,6.3,5.5,7.3,8.5,7.3,8.9,4.5,4.2,6.3
,6.9)
> trial <- data.frame(list("treatment"=treatment,"rep"=rep,"yield"=yield))
> str(trial)
'data.frame':   24 obs. of  3 variables:
 $ treatment: Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2 3 4 ...
```

```
 $ rep      : Factor w/ 4 levels "r1","r2","r3",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ yield    : num  9.4 9.3 7.6 7.6 6.9 7.6 8.9 9.1 7.8 6.5 ...


> rcb<-lm(yield~factor(rep)+factor(treatment))


Response: yield
          Df  Sum Sq Mean Sq F value   Pr(>F)
rep        3 13.3433  4.4478  7.0401 0.003540 **
treatment  5 23.6000  4.7200  7.4710 0.001066 **
Residuals 15  9.4767  0.6318
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(rcb)

Call:
lm(formula = yield ~ rep + treatment)

Residuals:
      Min        1Q    Median        3Q       Max
-1.066667 -0.525000  0.008333  0.450000  1.050000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.7667     0.4867  18.011 1.43e-11 ***
repr2         0.1333     0.4589   0.291  0.77538
repr3        -0.8833     0.4589  -1.925  0.07342 .
repr4        -1.7167     0.4589  -3.741  0.00197 **
treatmentB    0.8000     0.5620   1.423  0.17509
treatmentC   -1.6000     0.5620  -2.847  0.01225 *
treatmentD   -2.2000     0.5620  -3.914  0.00138 **
treatmentE   -0.8000     0.5620  -1.423  0.17509
treatmentF   -0.4000     0.5620  -0.712  0.48758
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7948 on 15 degrees of freedom
Multiple R-Squared: 0.7958,     Adjusted R-squared: 0.687
F-statistic: 7.309 on 8 and 15 DF,  p-value: 0.0005211
```

Compared with the other methods, it is no wonder non-statisticians like Excel. However, its problem is that although simple and clear, the method of analysis is only correct if there are no missing data. There are tricks and approximations one can use to get around the problem of missing data, but the other more computationally intensive methods generally take missing data in their stride. And of course, once experimental designs increase in complexity, use of Excel becomes harder and harder.

Balanced Incomplete Block designs

Randomised complete blocks designs work well, and nothing much can go wrong with them, provided you stick to the randomisation. However, although the validity of the blocking is guaranteed by the randomisation within each replicate, for large variety trials especially, one pair of varieties can end up adjacent or very close to each other but be in different blocks, whereas another pair will lie in the same block but could be 100m apart. According to our model, the difference between two plots lying in different blocks will be:

$$v1+b1-v2-b2 \qquad (1)$$

whereas the difference between two plots in a single block is:

$$v1+b1-v2-b1$$

$$= v1-v2 \qquad (2)$$

(1) could still be taken as an estimate of the difference between v1 and v2 if we are prepared to accept that, on the whole, for a random pair of blocks, b1-b2 is expected to be zero. More on this later. However, because of the inclusion of terms for blocks, the standard error of (1) will be calculated as larger that (2) even though the biological reality is that plots which are close together tend to be more similar. This doesn't introduce bias - our randomisation procedure eliminates that, but it implies that for large variety trials, there is room for improvement.

To account for this, "incomplete block" designs were introduced by Yates in 1936 - again at Rothamsted - early on in the evolution of field trial design. In these, the block size is smaller than the number of varieties or treatment combinations. As a result not all varieties can occur in every block – hence the name – but those varieties that do occur in the same block are physically always close to one another. As described above, the comparison between plots in different blocks is less precise than the comparison between plots in the same block. However, varieties are now allocated to blocks such that every pair of varieties occur together in blocks an equal number of times. As a result all comparisons between pairs of varieties are estimated with equal precision but with a reduction in error variance as a result of having smaller block sizes.

Here is an example of an incomplete block design for 16 varieties grown in 5 replicates each containing 4 blocks.

rep	block				
1	1	1	2	3	4
1	2	5	6	7	8
1	3	9	10	11	12
1	4	13	14	15	16
2	5	1	5	9	13
2	6	6	2	14	10
2	7	11	15	3	7
2	8	16	12	8	4
3	9	1	6	11	16
3	10	12	15	2	5
3	11	14	9	8	3
3	12	7	4	13	10
4	13	1	7	12	14
4	14	8	2	13	11
4	15	10	16	3	5
4	16	15	9	6	4
5	17	1	8	10	15
5	18	2	7	9	16
5	19	3	6	12	13
5	20	4	5	11	14

Each variety concurs with three others in one block in each replicate – 15 concurrences in total. Among 16 varieties there are only 15 possible concurrences for any selected variety. If you search you will find that each pair of varieties only appear within a block once.

For incomplete block designs, the number of plots per block should be chosen so that the area of a block is approximately square. With the plot sizes used for the National List and Recommended List cereal variety trials, this results in a block size of between 4 and 8 plots. All plots within a block must be contiguous.

The first incomplete block designs could still be analysed by hand. To achieve this, there were severe restrictions on the number and combination of replicates, varieties, and blocks. For example, for the design above, it is impossible to add or delete replicates while maintaining the balance between variety concurrences (unless you increase the number five at a time). Square numbers of varieties: 16, 25, 36 etc. tend to feature heavily in the designs. The bible for these designs is “Experimental Designs” (Cochran and Cox, 1957) which provided recipes for the most common, references to many others, details of how to analyse them, and how to cope with the inevitable problems of missing data. They also include designs for blocking in two dimensions. In fact, the example above has 16

varieties arranged in a 4x4 square for each replicate. You will find that the variety concurrences are balanced within column blocks as well as within row blocks. Note this design is not randomised. Randomisation (to get an unbiased estimate of error) is carried out by first shuffling rows within a rep then by shuffling columns within a rep. Replicates within the field should be allocated at random too.

The model for analysis is:

$$y_{ij} = \mu + v_i + b_j + r_{jk} + c_{jl} + e_{ij}$$

This extends the model we used for complete blocks before to include a term for row blocks within replicates, *r*, and for column blocks within replicates, *c*.

Note that although the incomplete blocks in this design, and in many others, are still clustered into complete replicates (they are said to be "resolvable" - there are some statistical properties that go with this property which needn't concern us here), this isn't a strict requirement of incomplete block designs. Resolvable incomplete block designs are also called lattice designs. Balanced designs exist which are not resolvable. However, the clustering of plots into complete replicates has practical advantages– you may not have resources to measure all phenotypes on all replicates for example. There is also a statistical advantage. It turns out that if the design is resolvable, then one is still justified in analysing it as if it were a randomised complete block experiment – the estimates of error and of variety effects are still valid. This may seem like common sense, but requires statistical theory beyond my comprehension to demonstrate.

These designs have been extremely successful and are still in use. Their TARDIS like property of testing more varieties that you can fit in any single block has resulted in better control of experimental error – because variety comparisons are made by within block comparisons so no plots being compared are ever that far from each other (but see below). However, they are now overused – the availability of computers for both design and analysis, while not rendering them obsolete, means that there are frequently more suitable alternatives available. This is particularly true in plant breeding where we are typically dealing with large numbers of varieties and cookbook designs for testing these numbers may not be available.

An example of this misuse is found in the obsession with testing varieties in lots of twenty five prevalent in some plant breeding organisations and lamentably with some official testing authorities too. (The latter often have access to card-carrying statisticians and should know better.) Their standard protocol is to test varieties in sets of 25 in 6 replicates in a "balanced lattice design". Balanced lattice designs are excellent, and if I had 25 varieties I wished to test, and felt that six was the appropriate replication number, then I would be very happy to use one. I cannot recall this ever having occurred. There are typically more than 25 varieties; some new candidates and some which have been in trial for longer. To accommodate this, several trial series are created. For every series, four common control varieties are typically selected. Comparisons within series are made by the standard trial analysis, but comparisons between series can only be made by

comparison through the four common standards. This approach to trial design has been called "Procrustean" by Roger Mead ("The Design of Experiments") meaning that the experimental material is made to fit into a design for which it was never intended. The statistical tail is wagging the biological dog. This system of testing is wrong because:

1) It is wasteful - 16% of resources are devoted to controls. Moreover if you wanted that degree of resource allocated to control varieties, surely you would be better off with more control varieties rather than more plots of the same controls.
2) It is imprecise - comparisons between varieties in different series can only be made through the control varieties and these are not themselves estimated perfectly.

In fact it can get worse. I've seen the same randomisation plan used at every site, and the first replicate is often not randomised at all - varieties arranged alphabetically for example, to make it easy for visitors to walk the trial. (This replicate often has better weed control too.)

Partially balanced incomplete block designs

Following the development of balanced incomplete block designs, partially balanced designs were developed. These come in many forms. These were introduced early too (1939) and relaxed some of the restrictions on the way in which varieties were allocated to blocks. However, at least among plant breeders, these did not catch on until a form of resolvable partially balanced incomplete block design termed an alpha design was developed by Patterson and Williams in 1976 and Patterson, Williams and Hunter in 1978. For these designs, there is little limitation on block size or replicate number, but software is required both to produce the designs and for their analysis. Their flexibility comes at a slight cost. The lack of balance means that all pairs of varieties are no longer compared with equal precision, though the difference in precision is usually slight. Designs can be created for large numbers of varieties - variety numbers of 500 say, but with block sizes of 5 or 10 (or anything else). The designs are resolvable: the blocks are arranged into complete replicates. In the more commonly used designs, varieties either concur once in any block, or never. These are described as (0,1) designs. Sometimes it is not possible to create (0,1) designs and some variety pairs concur twice (0,1,2) designs. Generally a (0,1) design is available provided the block size is $< \sqrt{(\text{no. vars})}$.

They are called alpha designs because there was planned to be a second series of designs – beta designs, but these never emerged, essentially because alpha designs alone do a good enough job.

Alpha designs are most easily produced by computer. There is a procedure AGALPHA for this in GenStat, not available from the menu system, which will produce plans for up to 100 varieties, with some minor restrictions on block size and replicate number. We shall examine an example alpha design in the tutorial.

Another source of good flexible incomplete block designs is from http://biometrics.hri.ac.uk/DesignOfExperiments/ . which will be discussed more in the section on row and column designs.

Recovery of inter block information in incomplete block designs.

Most incomplete block designs for plant breeding contain around 5-10 entries per block. In the discussion so far, these have been treated as if they were experimental factors with the same status as variety effects: we fit a simple model to the data to account for both factors. There is a degree of dishonesty in this approach – generally we don't care about the block effects themselves. Nor have the blocks been positioned to sit on top of some particular patch of the field knowingly to control for environmental variables specific to that patch. In statistical jargon, we say our blocks are "random effects" rather than "fixed effects." Fixed effects are the easiest to define. A fixed effect is an experimental treatment we are interested in: it could be nitrogen levels, it could be the performance of a new variety. Random effects are generally samples from some real or hypothetical population, which will have its own p.d.f. In some cases we may be more interested in assessing properties of the population rather than in performance of the samples themselves, or we may not care about them at all. This distinction can be fuzzy. Varieties, for example, can sometimes be regarded as samples from a population – a sample of lines from an F2 for example. Equally, on occasions, what started life as a random effect may end up as a fixed effect: if particularly high variation among blocks in our yield trial turned out to be attributable to nematode infection, we might get quite exercised about estimating the effect of nematodes on yield. In general, if you can conceive of your experimental treatments as being sampled from some population, you are justified to treat them as random, otherwise they are fixed. There are ambiguities. Plant breeders treat year effects as random. Animal breeders treat them as fixed. (In fact in plant breeding it doesn't matter - really large random effects, like years, tend to behave like fixed effects in their influence on variety performance. More on this later.) Even more confusing, animal breeders treat their sires and dams as random effects but plant breeders treat their varieties as fixed. Bayesian statistical approaches tend not to have this problem – everything is treated as a random effect as all parameters to be estimated are viewed as samples from prior distributions.

An extreme view of the difference between random effects and fixed effect is that fixed effects are something we care about and something that we have specifically set the experiment to measure. A random effect is something we don't care about: we just want to eliminate the problems they cause from our attempts to study the fixed effects.

To return to our incomplete blocks. Generally, we are justified in treating these as random - not only do we not care about them, but we can view them as samples from a hypothetically infinite population of similar blocks which we could have used in our field.

Now return to our comparison of two varieties in different blocks

v1+b1-v2-b2

but now we shall treat the blocks as random effects, drawn from a population with a mean of zero and an unknown variance. (The mean is zero because the block effects are expressed as deviations from the experiment mean.) If we have lots of differences such as given in (1) we can fit variety effects by minimising the variance attributable to blocks. For example, suppose we have a block size of two, and two blocks, one containing varieties v1 and v2, the other v2 and v4. Then the difference between the two block means has expectation v1 – v4.

Generally, unless block sizes are very small, the estimation of variety effects is more precise when using within block information (minimising **e'e**) than using between block information (minimising **b'b**). However the estimate of variety effects with the smallest variance is then a weighted mean of estimates from the two sources (weighted by 1/[error variance of the variety estimates]). The analysis minimising the within blocks analysis is called the intra-block analysis. The analysis using only the between blocks analysis is called the interblock analysis. The combined analysis is called just that, or the analysis with recovery of interblock information, or more generally the mixed model – mixed in the sense that the model contains both fixed and random effects (in addition to the error term). In practice you use a computer to do the analyses. We shall restrict ourselves to some observations about the process and ignore details of how the estimates of the fixed and ranomd effects are made.

1) If differences between blocks are very large, then the blocks variance is very large compared to the within blocks error and the difference between the combined analysis and the intra block analysis is small – all the kafuffle about recovering interblock information achieves little. In this sense treating blocks as random or fixed makes no difference if the blocks variance is much larger than the error variance.

2) If differences between blocks are negligible - we happen to have a very uniform experiment - then analysing the experiment as an incomplete block design is pointless and we might as well analyse the thing as a RCB - which is a valid analysis because the design is resolvable.

3) It follows from 1 and 2 that the greatest gain in recovering interblock information is when differences between blocks are modest. In fact, for most trials, at least in the UK, that is generally the case.

4) A very general method to analyse incomplete block designs was developed by Patterson and Thompson. Patterson is retired but still active. Thompson, his PhD student, went on to become head of statistics at Rothamsted – a position originally held by Fisher, then by Yates, and also by Nelder (who was very instrumental in the development of the statistical package GenStat) and by Gower (who made

important contributions to multivariate distance analyses). The method, caller Residual (or Restricted) Estimation by Maximum Likelihood (REML) has become the standard for estimation from data in which there are multiple random effects – each with their own associated variance structure. We shall come across it many times. It is useful in genetic experiments for estimating components of genetic variation and heritability.

The calculation of efficiency

It may seem a lot of effort to go to design and analyse trials in this way. Was it worth it? For resolvable designs, we have seen that we are justified in analysing the data both as an incomplete block design and as a randomised complete block. We can define the efficiency of our incomplete block design as the ratio of the variance of a variety difference in the RCB analysis to the variance of a difference in the incomplete block design. ie as:

$$Vdiff_{rcb}/Vdiff_{\text{incomplete blocks}}$$

There is no particular requirement to define efficiency in this way, we could work on standard errors or standard errors of differences. Historically, this is the one that has been used. An interesting alternative for plant breeding trials would be to define efficiency in terms of the expected response to selection when selecting on means produced by the two systems. This is the ratio of the heritabilities from the two systems. It therefore takes into account not only error variance but also the magnitude of the genetic variance, which has nothing to do with the experiment. If genetic variation is very low compared to environmental variation, then the measure of efficiency is the same. If genetic variation is very large, then it doesn't matter how you test your varieties. If heritability in the incomplete block design is 0.5, then the efficiency measured by expected response to selection turns out to be (normal measure of efficiency/2 + ½). So for an efficiency (normal measure) of 1.5, which isn't unusual in a large trial, the ratio of heritabilities is only 1.25.

In fact it is worse than this – we need to consider response to selection under an RCB when the selection itself is on the trait measured in incomplete blocks. This is an example of indirect selection - selection on one trait for response in another - to which we shall return. Taking this into account, the efficiency of the incomplete block design is reduced further to 1.12. So although alpha designs (or other equivalent incomplete block designs) represent an improvement over RCB designs, the improvement to the breeding programme is less than we might think. Nevertheless, you have nothing to lose: if you are planning a large RCB experiment, you should use a resolvable incomplete block design instead. It provides insurance in case things go badly wrong. You can still analyse it as an RCB if you insist.

Efficiency factor.

The "efficiency factor" is different to "efficiency" with which it is can easily be confused. The "efficiency factor" is important in the production of the design. It is a criterion by which alternative designs can be judged. "Efficiency" is more important when we review the experiment results. Suppose we have a homogeneous field in which the variance between two plots is unrelated to the distance apart of our plots. In this case, the error variance within incomplete blocks will be identical to the error variance for the RCB– there is no variation between blocks over and above that resulting from error within blocks. In this case, the RCB will out perform the intra block analysis – because with the latter not all comparisons between varieties are made within the same block. In these idealised circumstances of complete homogeneity, the efficiency of the incomplete block design, calculated as before, will always be <1. The expected efficiency of a design when the field is completely homogeneous is its efficiency factor. Different designs will have different efficiency factors and the best design for a given block size is the one with the highest efficiency factor. We can calculate the efficiency factor algebraically (difficult). However, there is a way we can calculate it numerically which is useful if you ever produce your own trial design or have a design with an unknown efficiency factor. Some statistical packages, GenStat included, will allow you to fix the error variance, and then estimate effects with this variance fixed. If you make up some data up, fix the error to some value, 1 is the obvious choice, then analyse the data as an RCB and as an incomplete block (with no recovery of interblock information), you will be provided with variances of differences between varieties from each of the designs. The ratio of these is the efficiency factor. When an incomplete block design is analysed with real data, the design must first recover from its innate inefficiency (efficiency factor <1) before gaining from the biological reality that the trial field isn't homogenous.

Deciding on Block size.

An examination of the experimental field may indicate some natural pattern of blocking and suggest allocating blocks or replicates to cover some specific observed feature - eg changes in soil type. However, in practice, at least in Europe, if faced with something like this, most breeders and trials managers would choose a different field. The ideal block size - that is the number of entries in a block - depends on the efficiency factor of the design and the pattern of fertility and other effects in the field. Generally, for most crops, there is some history which will indicate the sort of block size which has worked in the past. If historic data are available, then "post blocking" can be used to assess the effect of changing the number of entries per block. Here, one imposes on the existing trial data a new experimental design with a different block size, analyses the results, and tabulates block and error variances. By doing this for multiple trials and multiple block sizes, one can arrive at a reasonably objective decision for a block size that should work in practice. Note that in addition to observing the partition of error variation, we must remember that the post-blocking design is unlikely to be the optimal one for that block size because we are stuck with the historical arrangement of varieties in the field. Other designs with better allocation of varieties to blocks will exist with higher efficiency factors. All these

things can be quantified, however. In the absence of any information at all, a good place to start is to go for square blocks. The rationale for this is that the blocks are intended to control patchy fertility effects in the field, so for a given area, the more compact they are, the better. We end up back with Fisher's original approach – with several plots adjacent to each other within the block, and long(ish) plots running the length of the block. It is also probably not worth having blocks with more entries than √(total no. of entries) – in which case a (0,1) alpha design is available for the modest replicate numbers used in practice. For plant breeding experiments with large numbers of entries, this objective isn't difficult.

Deciding on Plot shape size

Statistics has had very little to say about the shape and size of plots. Fisher stated that plots should be long and thin so that the length of the plot would adequately sample environmental heterogeneity in one direction of the field, while blocking would control for error in the other direction. This is more or less how things have remained, although the availability of trial designs which control for heterogeneity in both directions (row and column designs) has reduced the need for this requirement.

The principal constraint on plot shape and size is practical: plot drills and harvesters generally work on a fixed width. The cost of buying new or altering existing equipment, even if possible, means that the cost of changing plot dimensions is generally prohibitive. Plot length can be changed more easily than plot width, but there are restrictions here too. Firstly there is a minimum length - the equipment can't function accurately if plots size is too small. Secondly, length may be restricted by the width of equipment which works across the plot. For example, best practice is to spray pesticide across plots so that any overlap or gaps between passes is spread equally over all plots. If spraying down the plots, certain plots may receive substantially more (or less) spray than others. So plot length is restricted to integer multiples of the sprayer width. All cultivations should be carried out across plots if possible. Nonetheless, there remain some opportunities to vary plot dimensions. In row crops, such as sugar beet, one can always vary how many rows to include in a plot.

In deciding on plot dimensions, there are two conflicting requirements:

Firstly, because of inter plant competition, traits such as yield are not so much properties of a single plants, but of a collection of plants. How large this collection must be before we achieve sensible assessments depends on the biology of the particular crop. Problems arise from both intra and inter genotype competition. Differential height between cultivars can mean that the plants in one plot can out-compete plants in neighbouring plots, but intra genotype (intra plot) competition between plants will also affect yield. There are various statistical / ecological / genetic treatments of competition which include corrections for its effect, but we aren't going to go into those in this course. In general, plant breeders try to avoid the problem by adjusting plot width until it is sufficiently wide that the effects of competition can be ignored. Sometimes, border strips around plots are

discarded - these contain the plants that are typically most affected by competitive effects. Fisher advocated that for sugar beet, plots should be four rows wide and the outer two rows should be discarded. However, this sort of operation is costly in terms of space and effort, so generally breeders and official testing bodies prefer to increase plot dimensions until the effects of competition are judged unimportant. In cereals, because of the gap between adjacent plots, border plants tend to grow taller and more lushly: a result of reduced inter plant competition. However, a simple increase or decrease in plant size at the edges of plots may look bad, but the effect is unlikely to be as serious as it looks. If the effect of absence of competition is simply to add a constant to the yield of each plot, estimates of differences in variety performance will be unaffected. Only if there is any interaction in performance between the edges and insides of plots is there a problem. This is a problem of genotype x environment interaction and can be treated as such. In sugar beet for example, across Europe three row plots have become the norm. A series of experiments in which yield from the inner and outer rows of plots was measured separately and variety performance compared, confirmed that this compromise worked well. Nevertheless, questions about plot width and the effect of interplot competition are still raised from time to time; usually by a breeder who thinks his/her variety is being unfairly treated.

Secondly, from the statistical point of view, if it wasn't for the problems of plant interference, then for a given area, the best plot would contain just one plant. Consider a plot of two plants. The plot performance is the total yield of both plants. Call the yield of the two plants x and y. Then (x+y) has variance Vx +Vy +2cov(xy) where the variance and covariance terms represent environmental error. If there is no inter plant competition, the covariance term will be positive: because fertility effects tend to occur as patches or gradients. Now, if we had two single plants located some distance apart, then the variance of (x+y) would be Vx+Vy + 2cov'(xy) – with a new and likely much smaller covariance terms. For the sake of simplicity, we'll assume a large trial so that x and y are quite a distance apart. As a first approximation, cov'(xy) = 0. So because of the fertility patterns we find in practice, small plots will give smaller standard errors of variety means than large plots containing the same number of plants.

There is therefore, a compromise between the increased precision afforded by small plots and the bias in estimation that one gets due to plant to plant interactions. Layered on top of this are the costs of running trials, in which a few large plots will generally be cheaper to manage than many small plots.

There is an upper limit on plot size too. Assuming a constant experimental area, larger plots means fewer replications. If each variety is represented by only a single replication, then there is no estimate of experimental error. As a rule of thumb, you require 10 df for a decent estimate of error. So in an (improbable) experiment to compare only two varieties, you would need 6 replicates of each.

A discussion of the factors to consider in determining plot shape and size can be found in: "Working rules for determining the plot size and numbers of plots per block in field experiments." Lin & Binns J Ag Sci 1984 **103**:11, 11-15..

Row and column designs

We briefly alluded to designs which block in two directions. There are alpha design versions of row and column designs, called alpha-alpha designs which we shall briefly describe. Software to produce them used to be available, for a fee, but I can no longer locate it. However, especially for large numbers of varieties, one can create reasonable designs by starting with an alpha design, but then rearranging entries within blocks to get a more uniform distribution of variety concurrences within columns. This sounds more onerous than it actually is. The efficiency factor of the design can be calculated as described earlier. Randomisation of the design is by first randomising the order of rows within replicates, then randomising the order of columns within replicates.

Fortunately, a reputable alternative source of one and two dimensional block designs is found here: http://biometrics.hri.ac.uk/. Currently, information on how the designs are produced is sparse, though help pages are promised. The site will generate arbitrary block designs for any arbitrary number of treatments with arbitrary replication and arbitrary block sizes. These are not alpha or alpha-alpha designs. They are constructed using a computer swapping algorithm and are D-optimal or near D-optimal. (Don't ask. They are good designs.)

As outlined earlier, the analysis of a two dimensional incomplete block design with no recovery of inter block information is straight forward (on a computer): we just have additional factors in the model. The analysis with recovery of inter block information has three error strata; between row blocks within replicates, between column blocks within replicates and a base error which can't really be regarded as within blocks any more, because there is only a single entry at the intersection of each row and column. In practice, the additional dimension of blocking generally doesn't usually give much of an increase in precision: the long thin plots and selection of the initial direction for blocking usually control most fertility effects. However, the additional dimension of blocking acts as additional insurance. Nothing is lost if there is no increase in precision and things can sometimes go badly wrong. This is particularly true in row crops, where sometimes there is no alternative but to carry out husbandry operations down rows. Tractor hoeing is the worst, with a serious risk of inducing "cultivator blight" in the USA or "steel worm" in the UK. There is no known genetic resistance.

Spatial analysis

An alternative method of analysis has become popular in the last decade or so, particularly in Australia (driven by Australian statisticians) and with some uptake now in Europe, though more by researchers than by official testing authorities and plant breeders (some of whom are still stuck in the 1930s remember). This is spatial analysis. It is worth stating that it is a method of analysis and not of experimental design.

Although the method, in its current form, is recent, it has a long history. A good review of the history of trial design is given by Edmondson ("Past developments and future opportunities in the design and analysis of crop experiments." J Ag Sci 2005, **143**:27–33). We have made much of the observation that plots which are close together tend be more correlated that plots which are further apart. A first effort to quantify this relationship was made by Fairfield Smith in 1938. The "law of Fairfield Smith" is an empirical statement that the correlation between plots decreases with the log of their distance apart. In other words it decays exponentially. Independently Papadapikis (1937) proposed a method of analysing trial data in which the central plot of each group of three was regressed on the mean of the outer two. The deviation of each plot yield from this regression was then used as the input data in a standard trial analysis. The rationale for this approach is the attempt to model environmental trends within the field in a continuous manner rather than in the discrete units of incomplete blocks. There are questions over the statistical validity of the approach which we won't go into and it never caught on in a big way, though was discussed from time to time over the next fifty years. An interesting but unpublished example is the work of Sydney Ellerton. Sydney was a plant breeder at the Plant Breeding Institute in Cambridge in the initial stages of WWII but was moved to manage a Polish sugar beet breeding station in Essex to ensure security of supply of seed for the crop during the war. He remained in the post until he retired and developed a method of trial analysis in which an initial adjustment for fertility trends was made by subtracting from plot number (n + 12) the mean of plots (n + 1..10, n + 14..23). That is, each plot was adjusted by the mean of the ten neighbouring plots on each side, but ignoring the two nearest neighbours. This adjustment by subtraction is equivalent to assuming a regression coefficient of 1. Then an additional adjustment was made in a Papadapikis manner by regressing the (adjusted) performance of the central plot on the (adjusted) mean of two nearest neighbours. A final estimation of variety effects was made on deviations from the regression analysis. The point of the two stage adjustment is that sugar beet, especially in the single row trials operated by Ellerton, is greatly affected by interplot competition. The second adjustment generally yielded negative correlation coefficients: interpreted as resulting from competition, but they could, in part, result from an over adjustment for fertility effects resulting from the assumed regression coefficient of 1 from the first adjustment on the 20 nearly-nearest neighbouring plots. There are other statistical question marks over this design, but I would like Sydney to have a mention. I once asked him how he had validated the method. He said he had asked Yates at Rothamsted and Mather, the Birmingham based biometrical geneticist, and they though it reasonable. Who am I to argue? Today, a little computer based modelling and re-analysis of historical data could easily validate and improve or reject the method, but life and Sydney have moved on. After the war, he came to be joint owner of the company,

which was highly successful until he sold it to Shell for about £4m in about 1977 to enjoy a long and affluent retirement. Lucky bastard.

At the heart of current methods of modelling fertility effects is the idea of autocorrelation - the correlation within a single variable between one observation and its neighbours. It is "autocorrelation" because it is not correlation between *x* and *y* but of *x* with other values of *x* some constant distance away. These methods originally took off with the analysis of time series, where an observation at one time is correlated with an observation at another. Outside the agricultural field (literally) a good example is in meteorology - a good predictor of tomorrow's weather is that it will be much the same as today's. Autocorrelation is calculated by taking a copy of the data and pasting it out of phase alongside the original data, then treating the two data copies as if they were independent variables and calculating the correlation in the normal manner. Note that you could paste 1,2… steps out of phase, so you can calculate autocorrelation for data 1,2,… steps out of phase. These correlations will be related – one would expect the autocorrelation for adjacent plots to be higher than that for plots two positions apart and so on (ignoring the sugar beet problem of interplot competition.). A simple function relating the various correlations could be:

$r_i = k^i$ i is distance (measured here in plots numbers)

In this model, the correlation of a plot with itself is 1, the correlation of neighbouring plots is k, the correlation of next-but-one plots is k^2 and so on. (nb what we really want is the correlation of error terms but we can't calculate these directly on plot data, because there is a variety component to plot yields which will act to reduce the observed correlation.) This simple model is described as an autoregressive model of order 1 (AR1). There are more complicated structures which include independent correlations for observations located more than one plot apart. However, we'll stick to order 1 since most analyses do too.

To date, all our analyses have treated the error terms (within blocks) as being independent. In matrix form, the error term for each plot is

$$\sigma^2\mathbf{I}$$

That is, each plot has the same error variance, σ^2 and the errors for each plot are independent: the correlation or covariance is zero. With autocorrelation our error structure becomes:

$$\sigma^2\,\mathbf{R}$$

R is made up of the elements defining the correlation between pairs of plots as described above (or in some other way). Note, so far we have just thought of autocorrelation as running in a single direction, just as for one dimensional blocking: the fertility effects we most want to adjust for are those running across plots. However, we can also consider autocorrelation in two dimensions. In this case, assuming we were fitting an AR1 model

in both directions, we would have to estimate two different ‘k’ values – one for row autocorrelation and one for columns, The elements of the matrix **R** are then just the product of the elements for rows and columns considered separately, which makes life easy.

In addition to the autocorrelated error, we can also choose to include an additional error term unique to each plot. (This can be viewed as a failure of the AR1 or other model to adequately describe the error structure.)

Finally, to analyse the data, we have a model with variety effects, correlated plot error terms. Other terms, fixed or random (eg blocks) could be fitted too. Everything must be estimated. Fortunately, our software will do all this for us, but the onus is on us to ensure that we have set up the model correctly and are interpreting and testing the results correctly too. We shall have a go in the tutorial.

The uptake of spatial analysis in Europe has been limited so far. Possibly the major reason for this is the conservatism of everyone involved – if it ain’t broke don’t fix it. Studies have been undertaken of the merits of spatial analysis compared to classic analysis methods. Generally, the improvement over analysis as a RCB is large, but the improvement over incomplete block designs – with blocking in either one or two dimensions - is slight. It seems that the approximation of continuous changes in fertility by stepwise changes as we move from block to block is pretty good in practice. There is also the risk, as highlighted by the heuristic Ellerton method, that spatial analysis may be distorted by inter-plot competition. Nevertheless, it is gaining ground. Note, however, this is only a method of analysis. I am not aware of work on what trial design is best if spatial analysis is planned in advance. It remains important that replicates of the same variety are well separated from each other and that pairs of varieties are distributed evenly in some sense. Much of this is achieved by use of incomplete block designs already. So for the moment, best practice is to lay out trials as two-dimensional incomplete block designs, consider analysing them using spatial analysis, but tread cautiously if you think there are variety differences in inter-plot competitiveness. DiGGer, Australian trial design software may be more suitable, but as I write this I have no knowledge of this software, other than that it is available from here: http://www.austatgen.org/files/software/downloads/

Unreplicated and partially replicated trials

Breeders often do not have the luxury of sufficient seed for a replicated trial, particularly in the early stages of a breeding programme. In addition, if genetic variation is sufficiently large, then the best strategy may be to select more intensely from a large number of unreplicated varieties rather than from half the number in two replicates each. However, there are still things we can do, and consideration must be given to the methods and statistical properties of unreplicated trials. Possibilities include:

1.) The addition of one or more check varieties at regular intervals. Data can then be expressed relative to the check, or the mean of the nearest check, or the weighted mean of checks by distance etc. etc. Typically >1 check varieties are included. Analysis in this manner assumes something about the error structure over the field. It is quite similar to running an incomplete block design with most entries unreplicated but with at least one common variety in each block. It would be better - in the sense of providing an unbiased estimate of error, if the location of the checks within each block was randomised, but I doubt if breeders engaged in their summer-time recreational activity of "scoring" every phenotype they can think of would like this. There is also a risk that in a uniform field, the adjustment by the nearest checks will introduce error rather than eliminate it (as the check varieties are also estimated with error). Potentially, spatial analysis may prove to be more efficient.

2.) Following on from (1), produce an incomplete block design in which some entries are replicated once, and some not at all. Designs like this with variable replication are called "augmented." The design is easily constructed from an alpha design for, say n entries in two replicates: x of the entries are allocated to 2x unreplicated varieties and the remaining (n-x) slots are filled with the controls. Note that block size must be adjusted such that there is a replicated entry in every block. Because of the way alpha designs are constructed, this is fairly easy to achieve. The design can be analysed as an incomplete block design or by spatial analysis. In this instance spatial analysis is probably best.

3.) Rely on pedigree relationships between candidates. Varieties from the same cross perform similarly. If they don't then either Mendelian genetics or the variety trial have gone very badly wrong and we should all go home. If our candidates are from several crosses, then it is easy to see that with an appropriate design and randomisation pattern, we can get unbiased estimates of the cross means, so at a minimum we could select between crosses. The within cross deviations from the cross means are a result of within cross genetic variation and of plot error. If we have an estimate of error from somewhere – from some replicated check varieties maybe, then we can decide how much weight we should place on the cross mean and how much on the deviation and select on an index of the two. There are more sophisticated methods too. If the between and within cross genetic variances have known expectations, then within cross error can also be estimated by subtraction. Without expanding here, there is clearly information in the pedigree of the varieties we are testing which can be incorporated into the analysis.

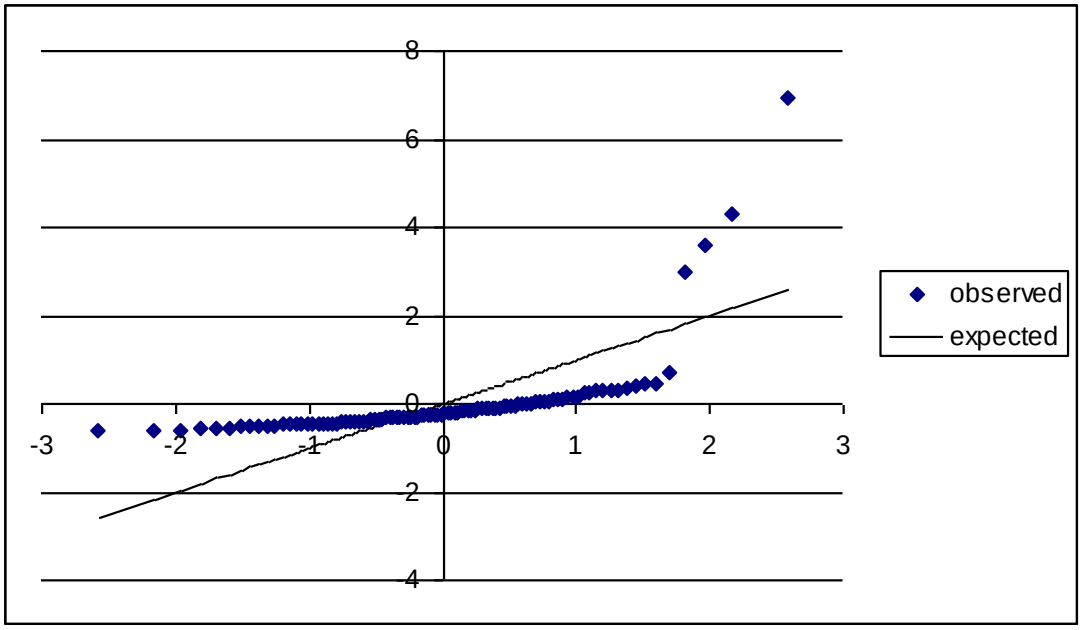
4) “Gridding” or blocking. Yields are expressed as deviations from a local average. Selection is then on the deviations rather than the raw data. Efficiency of this process will depend on heritabilities and the coarseness of field heterogeneity. Again, in a uniform field it is possible to do more harm than good.

Any others anyone? The consensus at the moment seems to be to combine (1) and (2). That is, to have both systematically placed check varieties and some varieties (they could be the same) replicated and randomised with the experimental material. I would always advocate at least some randomized control varieties: it is good to have available an unbiased estimate of error. That said, spatial analysis seems a reasonable way of analysing the data. With experience, one could chose to reduce or increase the number of replicated and of check varieties. (You can test the effect by analysing the trial including checks or with checks removed.) My prejudice is that one will be better off with fewer checks and more replicated controls. An example of the design and its analysis is given in the ASREML manual. (ASREML is an implementation of REML closely related to that in GenStat. The manual is free, the software is expensive.)
http://www.animalgenome.org/bioinfo/resources/manuals/ASReml/UserGuide.pdf

A final consideration is whether varieties from the same cross should be grouped together or randomised. If they are kept together, one will have better precision when comparing lines within-crosses but worse precision between crosses. Again, personally I would randomise. It may make life more inconvenient when taking field observations, but will guard against bias towards particular crosses when selecting on yield. If it is an unreplicated yield trial, it should be designed to give you the most efficient assessment of yield from unreplicated data. Inconvenience in making additional observations is of secondary importance.

Inspecting residuals – fertility plots

One outcome of the data analysis, however carried out, is that we have a predicted or fitted performance for each plot. We can inspect the difference between the observed data and the fitted values: the residuals. It is often informative to plot these. They should be normally distributed. This can be checked in a histogram or a “Q:Q” plot. In this, the ranked residuals, possibly after standardizing, are plotted against their expected value. The plot should lie close to a straight line passing through the origin with a gradient of zero. Exceptionally large or small values, or more general failure of the model will show up a deviations from this expected pattern. An example is shown below.



In this plot of 100 data points you can see that there are some clear outliers and that the observed gradient is substantially less than expected. In fact, in this made-up example, there are 90 random observations from a $N(0,1)$ distribution and 10 from a t distribution with one degree of freedom. Q:Q plots are becoming increasing commonplace in genetics, especially in gene expression experiments and in association testing.

Residuals should also be independent of the fitted effect for that trial plot – so a graph of residual against the fitted value should show no pattern. Most statistical software will generate this graph, the Q:Q plot and others for you: see the tutorials. In the event of gross failure - extreme non-normality or residuals increasing with fitted value - the data ought to be reanalysed. Reanalysis could use a different error distribution (opening the field of generalised linear models into which we shall not venture here) or more simply first transform the data in some way to make the error distribution better behaved. The most common transformation for plant breeders is to take logs of the data before analysis. In fact much biological data conforms better to normality on the log scale than on a linear scale. This is presumably because much biology is based on growth and factors interacting multiplicatively (ie additively on the log scale). However, in practice it is rare to get results which are substantially different from the initial analysis.

Individual residuals which are very large in absolute value are also worth inspection. As a rule of thumb, large means >3 error standard deviations. These are better revealed in Q:Q plots. Once identified, the first thing to do is to check the data for typing errors, misplaced decimal points and so on, and the field records for comments about bird / rabbit damage etc. Obviously transcription errors can be corrected. A high residual plus a comment about rabbits means that plot should be eliminated from the analysis. (There remains the risk that rabbits preferentially graze some varieties.) If there is no good cause to eliminate a plot other than its extreme value, then you are stuck. There are statistical methods for deciding whether to eliminate or keep these plots in the analysis, but in

practice they achieve little more than simple elimination of plots with residuals greater than some threshold. The non-statistician has a tendency to eliminate too many values. It is in the nature of extreme values from normal distributions that they look extreme. A simple procedure is to analyse the data with the values in and out and if it makes little difference then leave them in.

Residuals can also be calculated by subtracting from the observed yield the estimated variety effect only (ie leave the block effects in). Such residuals are estimates of the environmental effects on that plot. It is often informative to map these on their position in the field. These resulting fertility patterns can be very revealing. They can show effects of cultivations, irregular irrigation, fertiliser spillage and so on. On occasions, they reveal the presence of old field boundaries, hedge courses, buildings, drainage patterns, Saxon villages and so on. These are not merely of archaeological interest. They show up previously unknown sources of error in the field. If you are lucky, the block structure will have adequately accounted for their effect. In clear cut cases, one may consider reanalysing the data with the inclusion of an additional factor to account for the revealed effect. You need to exercise some caution over this, however, since it is always possible to find patterns in the clouds or dirty pictures in ink blots, or so I'm told. By overcorrecting, you will underestimate error and have a mistaken belief in the accuracy of your variety assessment.

Analysis across multiple sites and genotype x environment interaction

This will be discussed in the notes on Quantitative Genetics.
.

INTRODUCTION TO POPULATION GENETICS

Recommended text book: Theoretical Evolutionary Genetics by Joe Felsenstein is available free from http://evolution.genetics.washington.edu/ It doesn't do badly as a text book for quantitative genetics either, though is not as easy a read as Falconer & Mackay or Kearsey & Pooni.

Also Genetic Data Analysis 2nd ed. by Bruce Weir. The software Powermarker is based around this book. As the title implies this book is focused on data analysis.

Population genetics is the study of gene or allele flow with time and space through populations. Although generally regarded as of more interest to natural populations than domesticated crops, it has grown in importance in plant breeding with the increasing availability of molecular marker data and their use in diversity studies and linkage disequilibrium (LD) mapping. The basic principals also underpin much of quantitative genetics too. We shall therefore cover, at a superficial level, some introductory populations genetics.

Single loci: The Hardy-Weinberg Law

Before we describe it mathematically, we shall state why it is important. It states that genotype frequencies don't change in a population without the intervention of some external force. It is kind of the genetical equivalent of Newton's (first?) Law of motion – a body's motion is constant until some force acts to change it. For HW, these forces can be:

mutation
selection
sampling variation (drift)
migration
non-random mating

It can also be regarded, less fundamentally, but with more utility, as a means of predicting genotype frequencies from allele frequencies.

We shall derive the law for a locus with two alleles in a diploid organism in which mating is at random.

genotype	AA	Aa	aa
frequency	X	2Y	Z

X, 2Y and Z can have any frequency, provided X + 2Y + Z = 1.
Using 2Y as the frequency of the heterozygotes is just a trick to make the derivation easier. If it offends, you can substitute 2Y = H, say, and we'll still get the same result.

If mating is at random between genotypes, and union of gametes is random within each mating, then genotype frequencies in the next generation can be predicted from the product of the allele frequencies in the current generation.

genotype	AA	Aa	aa
frequency	X	2Y	Z
alleles	all A	½ A, ½ a	all a

Frequency of A gamete	X + ½ 2Y	=	p say		
Frequency of a gamete	Y + ½ 2Y	=	1-p	=	q say

with $p + q = 1$

	female gamete (freq)	
	A (p)	a (q)
male gamete (freq)		
A (p)	AA (p^2)	Aa pq
a (q)	Aa pq	aa (q^2)

giving genotype frequencies in the next generation

AA	Aa	aa
p^2	2pq	q^2

with allele frequencies:

A: $p^2 + ½\ 2pq = p(p+pq) = p$

so allele frequencies are unchanged and therefore the genotype frequencies in the next generation will also be AA p^2 Aa 2pq aa q^2.

Provided mating is at random these allele frequencies are reached after a single round of random mating. These genotype frequencies are therefore a stable equilibrium and the HW law is often referred to as the HW equilibrium, and the frequencies as the HW equilibrium frequencies.

Another way of representing the genotypes and their frequencies is as

$(pA + qa)^2$

provided we understand that after multiplying out, A^2 and a^2 represent the homozygous classes AA and aa. This representation offers an easy way to remember some extensions to HW. For more than two alleles at a locus, the HW frequencies are:

$(p_1A_1 + p_2A_2 + \ldots p_nA_n)^2 \quad =$

$p_1^2A_1A_1 + p_2^2A_2A_2 + p_3^2A_3A_3 + 2A_1A_2 + 2A_1A_3 + 2A_2A_3$

Extending further, for autopolyploids, the HW equilibrium genotype frequencies are given as

$(p_1A_1 + p_2A_2 + \ldots p_xA_x)^n$

where n is the ploidy level.

For example, for an autotetraploid (eg Medicago sativa) with two alleles at a locus:

$(pA + qa)^4 \quad =$

$p^4AAAA + 4p^3qAAAa + 6p^2q^2AAaa + 4pq^3Aaaa + q^4aaaa$

Note that

$(pA + qa)^4 \quad = \quad (pA^2 + 2pq\ Aa + qa^2]^2$

illustrating that HW equilibrium in autotetraploids is equivalent to the random union of diploid gametes, within which alleles are themselves in HW equilibrium proportions for a diploid.

Note that the coefficients of each genotype are given by the binomial expansion, of more easily through Pascal's triangle..

For diploids, HW proportions are generated after a single round of random mating. However, for autopolyploids this is not the case – the proportions are reached more slowly. For markers carried on X chromosomes too, HW proportions (in the heterogametic sex) are not attained immediately on random mating.

Non-random mating.

We shall define non-random mating as a system of breeding which does not alter allele frequencies in the population and does not involve any differential fertility between genotypes (that is, we exclude selection). This means we are considering inbreeding selfing, and forced outcrossing. We shall restrict our treatment to a pair of alleles in a diploid. Extension to multiple alleles is easy, to higher ploidy levels harder.

Departures from HW expectation. With X, 2Y and Z standing for frequencies:

	AA	Aa	aa
frequency with non-random mating	X	2Y	Z
frequency with random mating	p^2	2pq	q^2

We are stating there is no selection, so a reduction (increase) in frequency of heterozygotes must be balanced by an increase (reduction) in the frequency of homzygotes. Each heterozygote requires 1 A allele and 1 a allele, so a change in number of heterozygotes by 2 requires a change in numbers of AA genotypes by 1 and in numbers of aa genotypes by 1. So a reduction (increase) of n heterozygous genotypes must be matched by an increase (reduction) of n/2 in both homozygous classes. With a total of N individuals, if we set

$n/N = 2pqf$

we get

	AA	Aa	aa
frequency with non-random mating	p^2+pqf	2pq(1-f)	q^2+pqf

f is a measure of the departure of the population from HW equilibrium frequencies. Its maximum value is 1 in which case there are no heterozygotes, and its minimum is max(-p/q, -q/p) in which case one of the homozygous classes is missing. f is often described as the inbreeding coefficient of the population, though this isn't strictly true. f = 1 represents complete inbreeding, f = 0 gives HW proportions.

Suppose an initial population was selfed in all subsequent generations. Heterozygotes in one generation give rise to AA, Aa and aa genotypes in proportions ¼ : ½ : ¼. So:

generation	AA	Aa	aa
0	p^2	2pq	q^2
1	$p^2+pq/2$	pq	q^2+pq2
2	$p^2+pq3/4$	pq/2	$q^2+pq3/4$
3	$p^2+pq5/8$	pq/4	$q^2+pq5/8$
∞	$p^2+pq = p$	0	$q^2+pq = q$

The frequency of heterozygotes is halved each generation. In generation 1, f has the value of ½ and when inbreeding is complete, f has the value 1. A special case is an F2 population, in which case p = q = ½ and the frequency of heterozygotes declines as $½^n$ over n generations. Note however, that in this case, we are defining the F2 population, which is certainly in HW equilibrium, as being outbred (f=0). But the F2 is inbred compared to the F1, and any particular F1 may be a cross between two parental lines which are themselves related. So the F1 may have fewer heterozygotes than expected when compared to crosses among unrelated inbred lines from the same population. The important point is that the inbreeding coefficient is defined relative to a reference population. Changing the reference population will change the value of f.

An important case is for species (almost always of plants) in which some seed is set by random mating and some by selfing (eg oilseed rape). Suppose t is the proportion of seed set by outcrossing so (1-t) is the proportion set by selfing. Because there is always some outcrossing, the frequency of heterozygotes will never decline to zero. Equally, because there is always some outcrossing, even if we start with no heterozygotes in the initial generation, they will be generated. So heterozygosity rises if it is too low and falls if it is too high and there is a stable equilibrium. (This isn't really a proof, it could oscillate between limits which are >0 and <1, but it will do for us.) At equilibrium, the frequency in successive generations will be the same. Therefore:

freq in generation	AA	Aa	aa
n	P	2Q	R
n+1 from outcrossing	tp^2	2pqt	tq^2
n+1 from selfing	(1-t)(P+Q/2)	(1-t)Q	(1-t)(R+Q/2)

At equilibrium, the frequencies in the two generations are equal so the difference between them is zero

AA: 0 = $Pt - tp^2 - Q/2+Qt/2$
aa: 0 = $Rt - tq^2 - Q/2 +Qt/2$
Aa: 0 = $Q – 2pqt + Qt$

Solving for the heterozygotes:

2pqt = Q(1+t)
Q = 2pqt / (1+t)

The frequency of heterozygotes at equilibrium is therefore:

4pqt / (1+t)

This can be equated to a population with coefficient f:

2pq(1-f) = 4pqt / (1+t)

from whence:

$$f = (1-t) / (1+t)$$

This is also sometimes expressed in terms of the proportion of seed set by selfing

$$s = (1-t)$$

so

$$f = s / (2-s).$$

Note again that f need not be positive, although it is hard to regard it as a coefficient of inbreeding when it is negative. Although f has a maximum value of 1, its minimum value is <0 and depends on allele frequency (it cannot take a value such that the expected frequency of one of the homozygous classes would be less than zero). When allele frequencies are equal and all members of the population are heterozygous, the f = -1. This happens: crossing two inbred lines together then inbreeding, the inbreeding coefficient goes from +1, -1, 0, ½, ¾…as we pass from the inbred parents to the F1 to the F2, F3 and so on, where f is defined with reference to the F2.

Less extreme cases happen too. If our population is subdivided into two equal subpopulations with divergent allele frequencies but which are randomly mating within themselves, then there is an average reduction in heterozygosity compared to that expected from allele over the whole population.

Let frequency in population 1 = $p_1 = p+x$
frequency in population 2 = $p_2 = p-x$

Average heterozygosity $= (2p_1q_1 + 2p_2q_2) / 2$
$= (p+x)(1-p-x) + (p-x)(1-p+x)$
$= 2pq - 2x^2$

So there is always a deficiency of heterozygotes.

In addition, if the populations are crossed, then the frequency of heterozygotes in the hybrid population is:

$= (p+x)(1-p+x) + (1-p-x)(p-x)$
$= 2pq + 2x^2$

So there is an excess of heterozygotes in the hybrid population. This excess of heterozygotes is termed the Wahlund effect. If a trait is studied which shows some hybrid vigour, then a population intercross can show an increased performance. This is the basis of composite varieties. It may also explain at least in part the reason why, in crops where

population subdivision is particularly high, that the best hybrid varieties come from crosses between inbreds derived from different populations (eg maize).

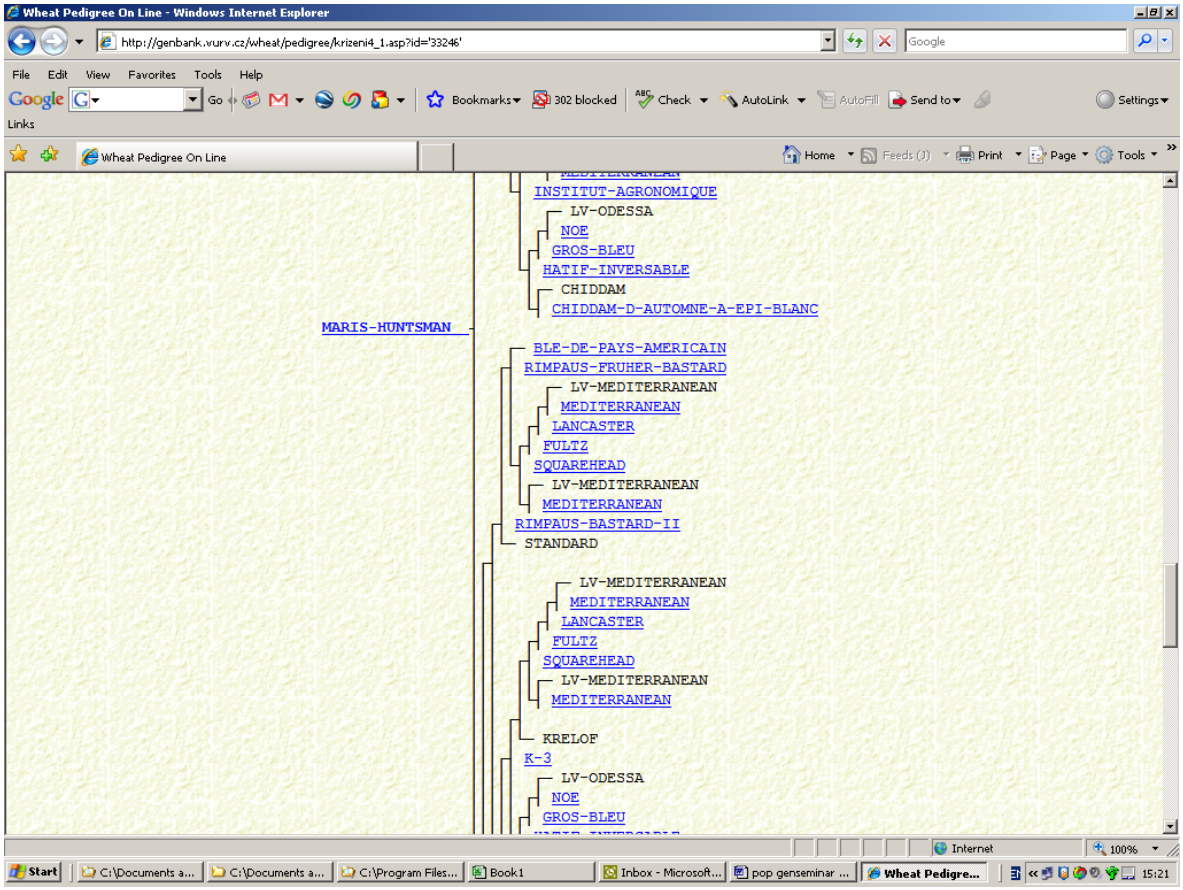
Divergence between populations can also be characterised in terms of Wright's F statistics, which can be viewed as equivalents of the inbreeding coefficient. See later.

Regular systems of inbreeding

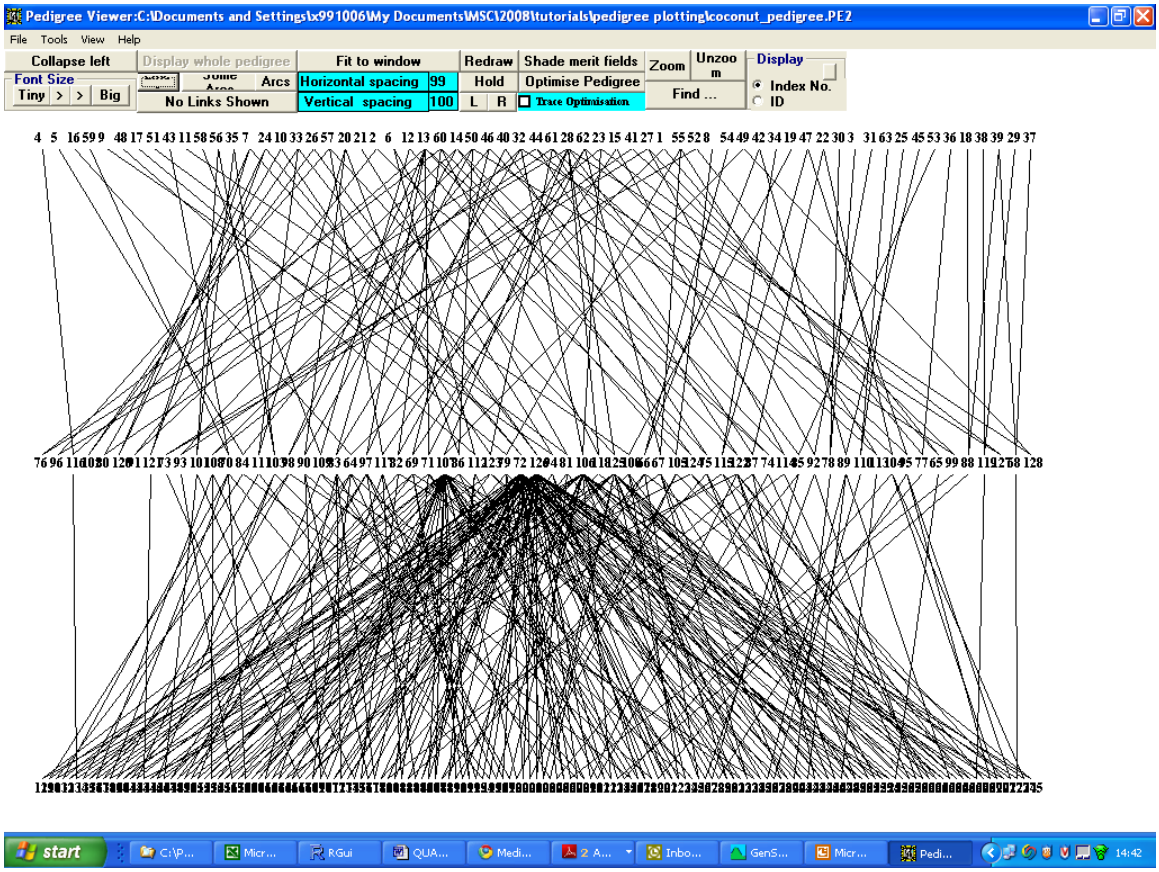
In breeding programmes, we are often more concerned with the inbreeding coefficient of a single individual or line rather than of the whole population. When dealing with inbred lines, F=1, we are often concerned to know how related pairs of lines are, The coefficient of kinship measures this and is closely related to F.

First however, a comment on drawing pedigrees. There are two ways of representing pedigrees: the animal breeding way, and the human genetics way. Plant breeders seem to prefer the human way. Writing software to draw pedigrees automatically is not easy. There is not much problem if there are no inbreeding loops, but if there are, then there is trouble. Some software which produces human genetics type pedigrees copes well, provided the pedigrees and loops are not too tortuous, which they usually are in crops of course, but much software takes the easy route out and breaks the inbreeding loops and duplicates the individual at the break point. This unfolding of pedigree means you can end up with the same cultivar appearing many times, which I find misleading. The animal style pedigrees are easier to draw by hand but don't look that pretty. Take your pick. I prefer the animal style because inbreeding is clearer. We can have a look at software for this in the tutorial.

Here is a small portion of the pedigree of the wheat Maris Huntsman, displayed in human genetics style, with inbreeding loops broken. You can see one of the ancestor lines, Squarehead, features several times. (With an inbreeding species like wheat, these are not strictly inbreeding loops. The loops result from crosses between related parents but no additional inbreeding is introduced to the progeny lines as these all fully inbred by successive rounds of selfing or doubled-haploid production.)


Wheat Pedigree On Line - Windows Internet Explorer
http://genbank.vurv.cz/wheat/pedigree/krizeni4_1.asp?id='33246'
Wheat Pedigree On Line
INSTITUT-AGRONOMIQUE
LV-ODESSA
NOE
GROS-BLEU
HATIF-INVERSABLE
CHIDDAM
CHIDDAM-D-AUTOMNE-A-EPI-BLANC
MARIS-HUNTSMAN
BLE-DE-PAYS-AMERICAIN
RIMPAUS-FRUHER-BASTARD
LV-MEDITERRANEAN
MEDITERRANEAN
LANCASTER
FULTZ
SQUAREHEAD
LV-MEDITERRANEAN
MEDITERRANEAN
RIMPAUS-BASTARD-II
STANDARD
LV-MEDITERRANEAN
MEDITERRANEAN
LANCASTER
FULTZ
SQUAREHEAD
LV-MEDITERRANEAN
MEDITERRANEAN
KRELOF
K-3
LV-ODESSA
NOE
GROS-BLEU

Here is an example of an animal breeding style pedigree for a three generation commercial coconut population. The very high contribution of a few trees in generation two to generation three is very clear.



F for a single individual or line is still defined with reference to a base population in which F = 0: all individuals are regarded as outbred. Often, in extended pedigrees, we view the founders (those for whom we do not know the parents) as being outbred: there is little else we can do, so we are looking at inbreeding which within the pedigree only.

F for a diploid individual is defined as the probability that the two alleles it carries are identical by descent (ibd). The coefficient of kinship, aka coefficient of consanguity, aka the coancestry is the expected inbreeding coefficient of the progeny of the cross between two individuals. It is therefore the probability that an allele picked from one individual and an allele picked from another individual are identical by descent. The multiple terms come from different translations of the French "consanguinité" into English. The French geneticist Malécot did much work in this area.

We shall follow Falconer &Mackay in referring to the inbreeding coefficient of an individual x as F_x and the coefficient of kinship of its parents as f_{ab}.

There are two methods of calculating F and f. One is best suited to hand calculation on animal style pedigrees, the other to calculation by computer on human style pedigrees. The computer method first.

Consider this pedigree

A x B C x D

P x Q

Y

We rely on the relationship

$F_Y = f_{PQ} = ¼ (f_{AC} + f_{AD} + f_{BC} + f_{CD})$

This can be seen to be true by considering the probabilities of drawing alleles ibd from P and Q.

This rule can be extended and modified as required. So

$f_{PQ} = ½ (f_{PC} + f_{PD})$

(either consider the probabilities directly or redraw the pedigree as (PxP) x (CxD))

The relationship between parent and offspring is

$f_{PA} = ½ (f_{AB} + f_{AA}) = 1/2$

(estimate p(ibd) or redraw as (AxA)x(AxB))

Selfing

$f_{AA} = ½(1+F_A) = ½$ if A is not inbred

These rules can be used to work from ancestors through to descendents in a pedigree, computing the inbreeding coefficients and kinships as you go. They can also be used to compute the inbreeding coefficient and the rate of approach to homozygosity in regular systems of inbreeding. Of these, the most common in crops is selfing, in which the inbreeding coefficient increases as $1- ½^n$ each generation.

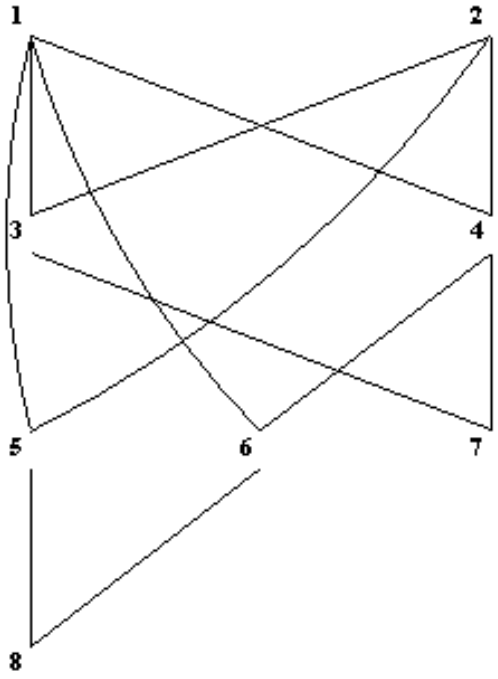
The coefficient of kinship is often confused with the coefficient of relationship. The coefficient of relationship is the correlation in additive genetic values between pairs of a specified type or relative. The coefficient of relationship is 2x the coefficient of kinship if the relatives are themselves not inbred. In general, the coefficient of relationship is:

$$r = 2f_{ab} / \sqrt{[(1+F_a)(1+F_b)]}$$

Confused? Try F&M for a better explanation.

The second way of establishing the inbreeding coefficient of an individual or the coefficient of kinship of the parents works best on pedigrees written down in the animal breeding manner

This is well explained in F&M. It is also better understood from a diagram. Firstly note that if there are no loops in a pedigree, then provided the founders are non inbred or are assumed to be non inbred (the usual case) then no individual in the pedigree is inbred.



In this simple case, there are several loops. For example, ID6 is a descendent of ID1 through both lines of descent. ID8 a descendent of ID2 through both parents and ID8 is a descent of ID1 through three different paths.

If there are multiple lines of descent connecting any ancestor to any other individual, then that individual must be inbred – there is a probability that the same gene copy has passed down each side of the loop. The probability that a selected gene copy in a parent is shared by one of the offspring gene copies is ½. The probability that this gene copy is passed on another generation is ½ x ½ and so on. So the probability of a particular gene copy passing down a line of descent from ancestor to a particular gene copy in the inbred individual is $½^{r}$, where r is the number of links in the paths connecting the individual to its ancestor.

Now consider a second line of descent linking the same ancestor and descendant. This may also have originated from the same copy of the gene from the same ancestor by passing down the other line of descent, this time with probability $½^s$ where s is the number of links in this line of descent. So the probability that the two copies in the descendant originate from the same ancestral copy is $½^{(r+s)}$. This is the probability for a single selected ancestral allele. As their are two ancestral alleles, the probability that the descendant alleles are identical for either of these is $2(½^{(r+s)})$ or $½^{(r+s-1)}$. In addition, there is a probability that the descendant alleles came from the same ancestor, but each from one of the two different ancestral allele copies. This probability is also $½^{(r+s-1)}$. If the ancestor was inbred, there is a probability that the two ancestral copies of the gene are already identical by descent, with is just F_A. So there is an additional probability that the descendant has inherited different alleles from the ancestor, but they happen to be ibd anyway. This is just $F_A ½^{(r+s-1)}$. Setting r + s - 1 = n and putting all this together we get:

$$F_x = ½^n (F_A + 1)$$

where n is also the number of individuals in the loop, ignoring the descendant whose inbreeding coefficient we are calculating, or equivalently it is the number of steps in the path minus 1.

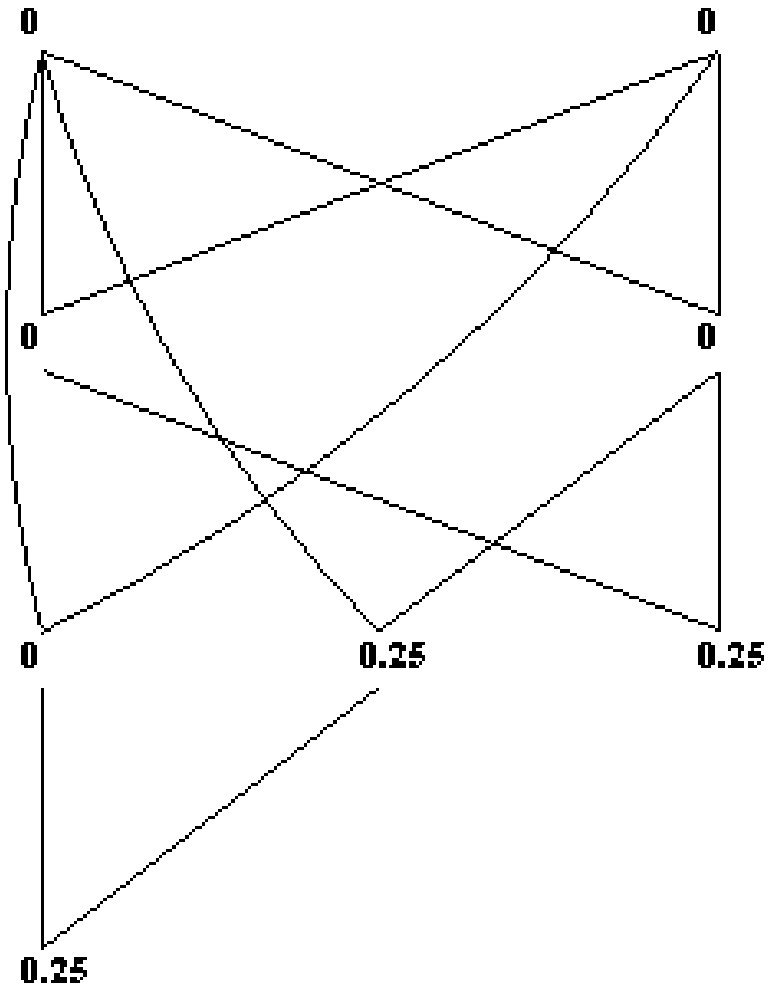
If there is more than one loop, involving one or more common ancestors, we simply sum over all possible paths

$$F_x = \Sigma ½^n (F_A + 1)$$

An example of this is in Falconer & Mackay. Make sure you count all paths, and don't include some that you shouldn't. This can be error prone. It is best to use a computer to calculate F if you can find some software. In the example above, ID8 is involved in the following loops:

5-2-4-6
5-1-6
5-1-4-6

Assuming the ancestors are all outbred, F_8 is $½^4 + ½^3 + ½^4 = ¼$.
Fortunately, we can get our software to check:



Neither of these methods is ideal for inbred crops, where every individual in the pedigree is inbred and the real interest is in coefficients of relationship. We do not want to include in the pedigree each generation of inbreeding before the next set of crosses are made. We only wish to include the parents and recombinant inbred lines (or doubled haploids) in every pedigree. If we assume inbreeding is complete, then F is always 1. However, the probability of a single copy of a parental inbred passing to the recombinant inbred remains a half even though many generations of selfing occur. We account for the fact that both parental copies are identical through the inbreeding coefficient of the parent. As a result, we can write down our pedigree using the inbred parents and progeny only, set F to 1 for all individuals then calculate coefficients of kinship as normal (ie as the inbreeding coefficient of the cross between any pair of parents). For pedigrees with no consanguineous loops this amounts to calculating the kinships as if all founders were outbred, then doubling the answer.

Sampling variation: genetic drift

Genetic drift is the process by which allele frequencies change over time, solely through chance sampling effects without any effects of selection, mutation, migration and so on. It is a consequence of finite population sizes. Felsenstein, in his free book, gives an interesting example. You have 2 parents, 4 grandparents, 8 great grandparents. You don't have to go too far back before you require more ancestors than there are people available. If you go back 40 generations, ~ 1000 years, you need more than a trillion (10^{12}). The only way to reconcile this problem is if some of the available ancestors were used more than once in your pedigree: your ancestors and therefore your parents are related and you are inbred. The result of inbreeding is increased homozygosity, so you are more homozygous than you might expect, or would like. This form of inbreeding through restricted population size affects genotype and allele frequencies too. It affects all populations of finite size and we'll also see that the smaller the population, the more important the effect. We'll try to quantify it.

Firstly, note that moving a population along one generation with random mating is like drawing a random sample from the existing population to create the next one. Although this sampling is at random, it is *with replacement* - a parent (or more exactly a parental allele) can be used more than once. A consequence of sampling with replacement is that the sampling variance is binomial. Imagine that we draw a sample from the current generation. This sample could be larger than the actual population size, which is why the sampling must be with replacement. This sample constitutes the next generation. Given allele frequencies in the current generation, we can use the binomial distribution to work out the full probability distribution of alleles in our sample (aka the next generation) and calculate the probability that it contains 0, 1 ,2 … 2N alleles of type A. (N is the diploid population size, so there are 2N alleles). Whatever the population's size, there is a finite probability that the population will be fixed for either the A or a allele. The probability that the number of A alleles in the next generation exactly equals the number of A alleles in the current generation can also be calculated and will be small, even for small population (sample) sizes. We know therefore, that our new allele frequency, p_1, will most probably differ from that in the current generation, p_0, and that the variance of this difference is just the variance of the allele frequency in the sample. For a binomial distribution with reasonably large sample size this is :

$$p_0q_0/2N$$

This is therefore the variance of the change in allele frequency from p_0 to p_1 between generations 0 and 1.

(Note, if sampling were without replacement, the sampling follows a hypergeometric distribution – something that we don't come across much in genetics).

Expected allele frequency changes over a single generation under drift are easily dealt with using the binomial distribution. The problem is predicting the consequences over many generations. After the initial generation, we don't have a definitive value of allele frequency from which we can estimate the variance of the change in the subsequent generation. Computing the variance of change in allele frequency over generations and estimation of statistics such as the time for allele frequency to change by a specified amount becomes very complex, especially once we include the effects of selection in addition to those of drift. We shall not attempt to derive these, but will give some results shortly.

Another way of considering the consequences of drift is through its effect on heterozygosity. Treating the starting population as completely heterozygous, we pull out a random sample, with replacement, of 2N gametes. These are paired at random to constitute the next generation. Under this sampling scheme, there is a chance 1/2N that an individual carries the same parental allele - is *identical by descent* - so the probability that an individual is heterozygous is now (1-1/2N). In the next generation (generation 2) the probability that we pull out a pair which are copies originating from the same parental allele is again 1/2N. Of the (1-1/2N) pairs which originate from different parental copies, a proportion 1/2N will be from the same grandparental copies, and so will still be ibd but 1-1/2N will be from different grandparental copies. So

$$H_2 = (1-1/2N)^2$$

or

$$H_2 = (1-1/2N)H_1$$

so that

$$H_n = (1-1/2N)H_{n-1}$$

and

$$H_n = (1-1/2N)^n$$

See Falconer & Mackay and Felsenstein for more details and a better explanation.

Although we have argued this from a starting point of a population in which all alleles are different, this isn't a requirement.

The probability that a pair of alleles is identical by descent is also the inbreeding coefficient for a suitably defined population. This business of requiring a suitable reference population is a nuisance. In one sense, one copy of allele A must be ibd with another if we define our reference population far enough back in time - assuming there was only a single mutation from a→A, which is usually the case at the single nucleotide level. In this case, all homozygous genotypes are ibd. However, in another sense, we have large outbred populations which we do not regard or treat as inbred, but in which homozygous genotypes do occur. To avoid confusion, not that successfully in my case, we also talk about identify by state (ibs) and identity by descent (ibd). Ibd implies ibs but ibs does not require ibd. Not very satisfactory really.

This leads us to the relationship between H and the inbreeding coefficient. F = 1-H. The inbreeding coefficient is generation n is

$$F_n = 1-(1-1/2N)^n$$

Dropping the requirement for the initial population to be all heterozygous and returning to our biallelic standard case, this also gives expected genotype frequency of heterozygotes as

$$2pq(1-F) = 2pq(1-1/2N)^n$$

As n increases, heterozyogtes are reduced in frequency. Ultimately there are none. On average, over multiple possible outcomes of random sampling over a very large number of generations, only AA and aa genotypes will be found, with frequencies p and q over all outcomes but in any particular instance only AA or aa genotypes will be present As a result:

The probability of fixation of an allele through drift is just the frequency in the initial population: rare alleles are most likely to be lost through drift.

Variation between population isolates.

If populations are split into subpopulations which are isolated, allele frequencies will diverge over time as a result of drift. The expected average allele frequency over all populations will remain the same as in the initial founder population. We have already seen that this divergence will result in increased heterozgosity when these subpopulations are intermated and that this can be exploited in hybrid breeding programmes. The divergence of populations isolated over time, the consequences of migration between subpopulations and the relationship between genetic divergence and the physical distance apart of the populations are much studied in population and evolutionary genetics (where models of speciation often require populations to be isolated). In plant breeding, it is important in terms of classifying and quantifying sources of novel genetic variation. This is largely assessed using molecular markers, which beggars the question about the relationship between these markers, largely assumed to be neutral, and the genes of

interest to the breeder which have usually had a history of selection. Here we shall restrict ourselves to a consideration of divergence using genetic markers only.

To measure genetic variation at a single locus within a population we use diversity:

$$\text{diversity} = 1-\Sigma p_i^2$$

Note that this is just the expected heterozygosity under HW, but by defining it this way we have a statistic that can equally be applied to inbreeding species. Over multiple loci, we can just take the average diversity. Comparing populations, you need to take care that you include a common set of loci across all populations, including those which are monomorphic in some populations. There is a slight bias in this estimate which is sometimes corrected for. The Expected variance of allele frequency is just half the heterozygosity of that allele. The maximum likelihood estimate of this is:

$$V_{(p)} = n/N - n^2/N^2 = p - p^2 = p(1-p)$$

where n is the count the allele p and N is the number of chromosomes. As a maximum likelihood estimate of variance, this is biased and should be adjusted by N/(N-1). If the variance is biased, then so is diversity which should be corrected in the same manner.

$$\text{unbiased diversity} = (1-\Sigma p_i^2)N/(N-1)$$

A rival measure of diversity is polymorphic information content, PIC. This is defined as:

$$\text{PIC} = 1 - \Sigma p_i^2 - \Sigma 2 p_i^2 p_j^2$$

where the second sum is across the n(n-1)/2 pairs of n different alleles. PIC was originally used to assess the utility of markers for human genetic linkage analysis. In practice diversity and PIC are very closely correlated and to my mind it is only the ignorant who use PIC - presumably driven by the cool acronym. Even worse, individuals sometimes will refer to PIC when they mean, and calculate, expected heterozygosity. I expect that the majority of people who use PIC could not define it.

A full treatment of the estimation of diversity and the variance of the estimates is given in Weir.

Genetic distance and Fst

In addition to quantifying variation within populations, we often need to quantify variation between populations. The most commonly used statistic to quantify this is Fst. Unfortunately, since its introduction is has shattered into many slightly different versions of the same thing, and it is frequently difficult to understand which version is being used, how it has been calculated and to what it refers. Once this has been decided, there is an additional problem of computing the variance of the statistic, which can be difficult, to say the least. The account below is the best I can do.

Following usual practice, we consider two alleles in a diploid. Assume we have a parental population which has split into subpopulations among which allele frequencies have diverged. For the purpose of defining Fst the divergence can be due to anything, for purposes of subsequent interpretation, the divergence is often assumed to be by drift alone. We compare observed and expected genotype frequencies in any subpopulation as:

	AA	Aa	aa
observed	obs(AA)	obs(Aa)	obs(aa)
with inbreeding	p^2+pqF	$2pq(1-F)$	q^2+pqF

This parameter set will give a perfect fit to any dataset if estimated from the data. However, the parameters can also be taken from the subpopulation or from the ancestral population (or equivalently from the average over all the subpopulations). Let

$\bar{p}$ = estimate over all subpopulations
p' = estimate within a subpopulation
F_{IT} = estimate of inbreeding coefficient over everything
F_{IS} = estimate within the subpopulation

then

$$1 - F_{IT} = obs(Aa)/[2\bar{p}(1-\bar{p})]$$

and

$$1 - F_{IS} = obs(Aa)/[2p'(1-p')]$$

from which we can estimate F_{IT} and $F_{IS.}$

Generally, because of the Wahlund effect we discussed earlier, $F_{IS} < F_{IT.}$

The ratio $(1 - F_{IT}) / (1 - F_{IS}) = [2p'(1-p')] / [2\bar{p}(1-\bar{p})]$

can be used as a measure which decreases as the populations diverge. We would prefer to have a measure which increases as populations diverge, so we take:

$$F_{ST} = 1 - [2p'(1-p')] / [2\bar{p}(1-\bar{p})]$$

This then gives the classic relationships between these F values, first introduced in the 1920s by Sewall Wright.

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

With multiple populations and alleles, we take averages over alleles within loci, then over populations. In estimating parameters, if the samples are of unequal size, we can take weighted averages as Weir suggests, or we may prefer not to if concerned that this will give undue weight to a small number of large samples. If the subpopulations themselves are of unequal size, then that can also gives us problems of interpretation. The computation of variances and confidence intervals is also problematic, and generally we are better off permuting or bootstrapping estimates, but must decide whether to bootstrap over markers, individuals, populations or what. Expert opinion (not mine) is that all of this is a problem. In many publications using Fst to measure divergence, it is not clear what has been done, or why, so caution is required. To quote David Balding, many researchers are at risk of "confusing familiarity with understanding." A recent review is Holsinger and Weir (2009) *Nat Rev Genet* **10**:639-650.

F_{ST} can also be estimated, or defined even, as the variance in allele frequencies over populations divided by the value expected from the average allele frequency:

$$F_{ST} = \mathrm{var}(p'_i) / \bar{p}(1-\bar{p})$$

Here is an example:

population	p	q	pq	Fst
1	0.031	0.969	0.030	0.847
2	0.222	0.778	0.173	0.119
3	0.040	0.960	0.038	0.804
4	0.787	0.213	0.168	0.145
5	0.259	0.741	0.192	0.021
average	0.268	0.732		0.387

Fst for each population is calculated as

$$F_{ST} = 1 - [2p'(1-p')] / [2\bar{p}(1-\bar{p})]$$

For example, for the first population:

$$\text{Fst} = 1 - 0.030 / (0.268 \times 0.732)$$

$$= 0.847$$

The average Fst over all five populations is 0.387.

Estimated from the variance of allele frequencies across populations:

The (maximum likelihood) estimate of var(p'_i) is 0.076

$$F_{ST} = var(p'_i)/ \bar{p}(1-\bar{p}) = 0.076 / (0.268 \times 0.732) = 0.387$$

This account of F_{ST} has been developed for diploids in terms of their inbreeding coefficients. When defined as a ratio of variances, it applies equally well to all ploidy levels and to fully inbred collections of crop varieties which for most practical purposes can be treated as haploids. F_{ST} is sometimes written (eg by Weir) as θ when discussing the haploid case.

F_{ST} also has an explicit population genetics expectation for subpopulations which are diverging through drift alone:

$$F_{ST} = 1 - (1-1/2N)^t$$

where t is the time since the populations split and N is the diploid population size.

There are many other related methods of quantifying diversity and distance. Several are discussed in Weir. We discuss briefly only one.

Alleles may be coded as 1 and 0, for presence and absence, and a hierarchical analysis of variance carried out to partition variation into components for between populations, within populations, and within individuals within populations. Because of the nature of the data, and the accumulation of results across markers and alleles, significance of the resulting F ratios is usually carried out empirically. This method has been labelled the Analysis of Molecular Variance or AMOVA and is available in much software, notably Arlequin, whose authors were at the forefront of developing and promoting this method. It would presumably be easy to adopt REML to molecular data which might allow better treatment of missing and unbalanced data, but I am not aware that this has been done.

We finish this section with a quote from Weir. “Care is needed to match the distance to the intended scope of inference.” In other words, don’t use a distance measure for a purpose for which it is not intended. For example, another commonly used measure, derived by Nei, is appropriate for studying long term evolution with populations diverging through both drift and mutation. This probably isn’t appropriate for most plant breeding applications. F_{ST} is a safer measure since, as defined here, it assumes no mutation but that populations are changing through drift. One can apply geometric

distances too, which assume nothing about genetics. Whether this is a good thing or a bad thing probably depends on the data set and the application.

Effective population size.

Often, in discussion of population genetics, diversity and so on, the term effective population size (written as Ne) is encountered. Population genetics theory has generally been worked out using something called the Fisher-Wright model. This refers to a particularly well behaved sort of population in which mating is at random (so selfing can occur), generations are discrete, population sizes are constant from year to year, each progeny gamete is equally likely to originate from any parental gamete (so that in diploids each individual contributes two gametes to the next generation on average, but the probability of contributing is binomial or Poisson. All this simplifies the maths, but can be very far from biological reality. Sewall Wright introduced the concept of effective population size to bring the Fisher-Wright model back in line with reality. He showed that correct results could be produced for many more complex situations by replacing the true population size or sizes with an effective population size which gave the correct answer if used in the standard Fisher-Wright model. What is more, it turns out that for many commonly found departures from the Fisher-Wright model; population sizes varying over generations, very variable family sizes, separate sexes etc, that the appropriate value for Ne could be quite easily calculated.

The study of Ne and how to estimate it has subsequently taken on a life of its own. We shall only report one result of importance to breeders, namely that if parents are forced to contribute equal numbers of gametes to the next generation, Ne is 2N. This is quite an improvement if one is concerned about loss of variation through drift. Moreover, it is often the norm in breeding programmes: selected parents may be inter-crossed but equal numbers of progeny are taken from each cross. This is routine in recurrent selection programmes for example. It needs to be taken into account when deciding on intensities of selection. To my mind, the breeders I have known have a tendency not to select sufficiently hard in the interests of conserving genetic variation. Implicitly they want to avoid fixing disadvantageous alleles by drift though they would not usually describe their concerns in those terms.

The equal contribution of gametes to the next generation is also important in the conservation of genetic variation in gene banks. Here, on seed multiplication, the greatest care should be made to ensure that parents contribute equally. The gold standard for this is through making pair-crosses and then taking equal numbers of progeny from each cross. Taking equal numbers of seed from each parent after open pollination is better than nothing, but the contribution of male gametes is uncontrolled and can vary greatly from plant to plant. One proviso to this argument is that it may be more costly to generate 25 full sib families say (Ne = 100) than it is to sow out 1000 plants, let them open pollinate, then harvest the bulk. However, even here the open pollinated bulk does not necessarily have the higher Ne. The unequal contributions to the next generation, the correlation between male and female gamete contributions (arising from selfing, if it occurs, and also

from the tendency for large plants to produce more seed and more pollen) can all combine to drive Ne down. In an extreme case, in a natural population of a poppy species, Ne was estimated as 1% of N. (JS Gale, Theoretical Population Genetics 1990).

This is as far as we're going to go in our consideration of pure drift: after this the mathematics gets heavy. We shall just give some results of interest.

Mutation

All variation ultimately originates from mutation, which can range from single base pair changes through gene duplication/deletion, chromosome inversions and translocations up to whole genome duplications. Here, we consider the effect of mutations which have no effect on viability or fertility.

In a diploid population of size N, suppose mutations occur from allele *A* to allele *a* at a frequency u per generation per copy of *A*. If the frequency of *A* in the population is p in generation n, then in generation n + 1 the frequency of *A* will be p(1-u) and the frequency of *a* will rise to q + pu. If mutation from *a* to *A* also occurs at a frequency v, then the net change in a over a generation is pu - qv. At equilibrium:

$$pu = qv \quad \text{giving}$$

$$q = u/(u+v)$$

This seems reasonable - the equilibrium frequency of the mutant allele is the relative rate of forward mutation over the forward + background rate. It isn't very interesting however. At a single nucleotide, the forward and reverse rates of mutation may be fairly similar, but over a whole gene, the rate from the functional form to the non-functional form is generally much greater that the non functional to functional so it is hardly worth worrying about reverse mutation. Also, within a gene, it is more likely that a second mutation will affect a different base or amino acid rather than reverse the initial mutation. Recent population genetics models of molecular evolution have therefore tended to rely on two models of mutation called the "infinite alleles" model and the "infinite sites" model. The names are reasonably self explanatory.

Mutation and drift.

The fate of mutations tends to be determined by selection and drift and of these, at least initially, drift is much more important.

When a new mutation occurs, it will have an initial frequency of 1/2N. If N is large, the probability that a mutation is lost will be high. Conversely, if N is small, the probability that the mutation is lost in a single generation is small, but the probability that the mutation occurs in the first place is much lower. Over a single generation, these two

effects are exchangeable. With a mutation rate u in a diploid population of size N, the frequency of the new mutation is 2Nu so the probability of loss in the generation immediately following mutation is, from the Poisson distribution, e^{-2Nu}.

Mutation is creating variation, but drift is disposing of it. For the infinite alleles model, there is an equilibrium:

$$F \sim 1/(1+4N_eu)$$

F is the probability that a pair of alleles are ibd which, for the infinite alleles model, is the probability that an individual is homozygous. At this point there is no change in allele frequency from generation to generation. The equivalence between Ne and u is maintained, see Felsentstein for details. The quantity $(1+4N_eu)$ is, for reasons given in Felsenstein, the *effective number of alleles*. This is the number of alleles at a locus, of equal frequency, that might be expected in a population for a particular value of $4N_eu$. For reasonable values of N_e and u, we expect a fair number. In practice, this means that we ought not be too surprised to find numbers of DNA variants within genes or other stretches of sequence of modest length.

Substitution rates

Although most new mutation get lost by drift, some get lucky and increase in frequency to fixation. This is the whole idea behind the neutral theory of molecular evolution: most of the observed change in DNA sequence is a result of random fixation of neutral mutations. The probability of fixation of a new mutation is 1/2N. 2Nu neutral mutations occur per generation (we need to be careful to define u as the neutral mutation rate). So the expected number of neutral mutations arising in a generation which will ultimately be fixed is 2Nu x 1/2N = u. So the rate of substitution – the frequency of neutral mutations which are fixed per generation is u and is independent of N.

Selection

Ignoring drift, selection can be accounted for as follows. We define the contribution (in gametes) of each genotype to the next generation as its *fitness* and quantify this through a *coefficient of selection* s, which will vary depending on the model or form of selection. Given genotype frequencies before selection, we can then compute genotype and allele frequencies after selection and therefore the allele and genotype frequencies in the next generation. Assuming random mating, then immediately after mating, but before any selection:

	AA	Aa	aa
initial frequency	p^2	2pq	q^2
relative viability	1	1-hs	$(1-s)q^2$
after selection	p^2	(1-hs)2pq	$(1-s)q^2$
freq. after selection	$p^2/\bar{\omega}$	$(1-hs)2pq/\bar{\omega}$	$(1-s)q^2/\bar{\omega}$

$\bar{\omega} = (1+s)p^2+(1+hs)2pq+q^2$ is the mean population fitness, which takes on a life of its own in some parts of population genetics in the same way as Ne.

After much algebra, the frequency of *a* after selection is:

$$(q-hspq-sq^2)/ (1-2hspq - sq^2)$$

and the change in allele frequency over a single generation is:

$$spq[q+h(p-q)] / (1-2hspq - sq^2)$$

h is a factor to account for any different effect of selection on heterozygotes compared to the homozygote.

These formulae are messy but different fitness models can be accounted for by specifying the value of h:

h	= 1	allele *a* dominant in fitness
h	= 0	allele *a* recessive in fitness
h	> 1	genotype *aa* is overdominant in fitness (heterozygous advantage)
h	= ½	allele *a* is additive in fitness
$(1-hs)^2$	= 1-s	allele *a* is multiplicative in fitness.

Substituting for h gives some simplification to the equations but they remain complicated. F&M gives a table (2.2), though the fitness model is not always defined exactly as here.

Other simplifications can be made if allele *a* is rare (q small) and/or the coefficient of selection is small (which it usually is, especially for each locus of a polygenic trait). A favourite is the case of multiplicative fitness. This is because, if we assign fitnesses to alleles rather than genotypes as :

	A	a
frequency	p	q
fitness	1	(1-s)

$\bar{\omega}$ = (1-sq)

then just as $(pA + qa)^2$ will give the genotype frequencies for a diploid, then

$$[pA + (1-s)a]^2$$

gives the fitnesses of the diploid with $\bar{\omega} = (1-sq)^2$

Putting this together, the frequencies of the diploid, after selection are given as

$$[pA + q(1-s)a]^2 \ / (1-sq)^2$$

or :

	AA	Aa	aa
initial frequency	p^2	2pq	q^2
relative viability	1	1-s	$(1-s)^2$

(Note the modification of the definition of fitness: s here was hs in the original diploid model).

This makes life particularly easy. Allele frequency changes and much else can be calculated for the simpler haploid model and applied directly to the diploid case. The allele frequency after selection is:

$$q(1-s) / (1-sq)$$

and the change in frequency is:

$$qs(q-1) / ((1-sq)$$

Some care is required in reading accounts of these changes in different text books, since the terms are sometimes defined in changes in frequency of allele *a* (F&M) and sometimes in terms of changes in *A* (Fe). Sometimes the default value of s is negative (F&M) and sometimes positive (Fe).

The multiplicative model can also frequently stand in for the additive model:

	AA	Aa	aa
multiplicative	1	1-s	$(1-s)^2$
additive	1	1-s	1-2s

If s is small, $(1-s)^2 \sim 1-2s$.

Multiplicative fitnesses also have the property of leaving the population in HW equilibrium after selection (but before random mating). Other forms of selection leave the population out of HW equilibrium until random mating restores is (but with a new allele frequency). This means that selection cannot be detected by comparing observed and expected genotype frequencies if fitnesses are multiplicative, and they are going to be hard to detect if they are small but additive too.

Stable polymorphisms.

Some forms of selection maintain variation. The most well known of these is heterozygous advantage or overdominance. There is no haploid equivalent for this model. Following Felsenstein's definition of terms

	AA	Aa	aa
fitness	1-s	1	(1-t)

$$\bar{\omega} = 1\text{-}sp^2 - st^2$$

Frequency of *A* after selection

$$p(1\text{-}sp) / \bar{\omega}$$

Frequency of *a* after selection

$$q(1\text{-}tq) / \bar{\omega}$$

and at equilibrium

$$p/q = p(1\text{-}sp) / q(1\text{-}sq)$$

so $p = t / (s + t)$

.

This equilibrium is stable. If p deviates from t/(s+t) selection will act to return the allele frequency to its equilibrium value. This is true provided 1-s and 1-t are both <1 - the heterozygote is the most fit genotype. If the heterozygote is the least fit genotype, then the equilibrium is not stable, and selection will hasten the fixation of the most common allele.

Overdominance as a means of maintaining variation (and as an explanation for heterosis) has a continuing band of enthusiasts, at least partly because of the simplicity of its mathematics. There are some counter arguments. Firstly, at the molecular level, there is no need to routinely invoke selection to explain the presence of large amounts of polymorphism: drift alone does a good job in many (some would say virtually all) cases. Secondly, you can't have heterozygous advantage in haploids, yet they seem to be as polymorphic as diploids. Equally, in the presence of inbreeding, rather than random mating, the conditions under which the equilibrium is stable are more stringent. The same problems of stability apply to multiple alleles: there are constraints on the fitness values of the genotype classes under which a stable polymorphism will result. Finally, there remain very few cases in which the heterozygous advantage of a polymorphism has been demonstrated experimentally. The most well known and best case is still that of sickle cell anaemia in man.

Occasionally, one comes across cases where an excess of heterozyogtes in a population believed to be randomly mating is taken as providing evidence of heterozygous advantage. However, this is not the case. The fallacy arises because many selection schemes leave the population out of HW equilibrium (the exception being when fitnesses are multiplicative). A trite example is selection against a recessive homozygote. If one homozygous class is completely eliminated, then under HW expectations, with new, post selection allele frequencies, we expect at least some homozygotes even though we observe none. In fact, if we compute the expected frequencies, we find that there is an equal deficiency in both homozygous classes. (The deficiency will always be equal. In terms of the inbreeding coefficient, as discussed earlier, it is pqF). This fallacious interpretation was very popular in the 1970s.

Frequency dependent selection

There are many other possible mechanisms which might maintain polymorphisms; fitnesses which fluctuate over time and space for example. Discussion of these can be found in Fe. Here we are going to discuss one other: *frequency dependent selection.* In this, the fitness of a genotype or an allele is inversely proportional to its frequency. We shall consider only selection acting on alleles, but the conclusions are essentially the same for selection acting on genotypes. A rare allele will be selected for and rise in frequency, but if it rises too far, its fitness will drop below that of other, rarer alleles and will decline. Since we expect this form of selection to maintain multiple alleles, we'll assume n alleles from the start. In haploids:

	A_1	A_2	…	A_n
frequency	p_1	p_2	…	p_n
fitness	$1-p_1s$	$1-p_2s$	…	$1-p_ns$

$$\bar{\omega} = 1- s\Sigma p_i^2$$

Frequencies after selection can be calculated in the usual way. At equilibrium, there will be no change in allele frequency so

$$0 = p_1(1- p_1s) / \bar{\omega} - p_1$$

$$0 = p_1 (1-p_1s) - p_1(1- s\Sigma p_i^2)$$

$$0 = p_1^2s - p_1s\Sigma p_i^2)$$

$$p_1^2s = p_1s\Sigma p_i^2)$$

$$p_1 = \Sigma p_i^2$$

If there are k alleles of equal frequency, then

$$p_1 = 1/k$$

and

$$\Sigma p_i^2 = k\,(1/k)^2 = 1/k$$

These equilibrium frequencies can be shown to be stable. Under this simple model, frequency dependent selection maintains alleles at equal frequency. What is more, new mutations will be at an immediate selective advantage. As a result, frequency dependent selection is more effective at maintaining variation than drift alone, although because the equilibrium frequency with multiple alleles is low, the effects of drift in distorting allele frequencies from their equilibrium frequencies can be quite large.

A model for diploids can be generated by assuming multiplicative fitnesses, so the fitness of heterozygotes is $(1-sp_i)(1-sp_j)$ and of homozygotes is $(1-sp_i)^2$. This will also produce a stable equilibrium of equal allele frequencies. Other models also exist (see Fe) with differing equilibrium frequencies.

Frequency dependent selection tends to be implicated when we are dealing with sex, disease or both. (And there is a theory that the evolution of sex has been driven by pressure of disease.) An observation of higher allele numbers than expected under drift, especially if they are functional, is viewed as circumstantial evidence of frequency dependent selection. Definitive evidence requires the cause of the selection to be identified too. The MHC complex in vertebrates (HLA in human) is a good example - a large set of highly polymorphic genes involved in the immune response, but also implicated in sexual selection in species as diverse as stickleback and human. The best example of all comes from plants, however, where the large number of alleles found at loci determining self incompatibility systems are maintained by frequency dependent selection. Disease resistance genes in plants also tend to have high frequencies of functional polymorphism which have been interpreted as evidence of frequency dependent selection.

Selection and drift

We return again to directional selection. Clearly allele frequencies change as a result of selection, but the deterministic equations given earlier cannot accurately predict this change except for intermediate allele frequencies in large populations. In general we must also consider the effect of drift. That this must be so is easily seen by considering the fate of an advantageous but recessive mutation. No selective change in allele frequency can occur until mutant homozygotes appear in the population. This requires the allele frequency to first move from 1/2N to around $\sqrt{(1/2N)}$, a process which can only take place by drift. Qualitatively, we might expect drift to hinder the spread of an advantageous allele if it is rare, since it is more likely to be lost by chance. At high allele

frequencies, drift might assist in fixing an advantageous allele. In small populations, an advantageous allele may behave as if it were neutral – the noise of sampling overcoming the signal of selection. In large populations, if fitness differences are large enough, we may get away with the deterministic treatment. This area of population genetics is mathematically demanding, to say the least, and we are not going to go into it in any detail.

We'll stick to the multiplicative model for the reasons we gave earlier. A complete solution, (slightly approximate) was given by Kimura in 1957. The probability of fixation of an allele with frequency p is

$$U(p) \sim (1-e^{-4Nsp})/(1-e^{-4Ns})$$

Just to cause confusion here, the genic model used is with fitnesses (1+s) and 1 (as in Fe) and not 1 and (1-s).

Remember, the diploid population size is N so we are dealing with 2N chromosomes. Sometimes, in discussion of the haploid case, the population size, and therefore the number of chromosomes is given as N and the formula requires altering accordingly.

There are some special cases we can evaluate with this formula. Firstly,

s very small

$$\begin{aligned} U(p) &\sim [1-(1-4Nsp)]/[1-(1-4Ns)] \\ &\sim 4Nsp/4Ns \\ &\sim p \end{aligned}$$

just as for drift. For s to be small enough for this approximation to be valid,

$$s < 1/16N$$

Small selective advantages make a difference to fixation probabilities, even though evaluation of the Kimura formula shows that most rare advantageous mutations will still be lost through drift, particularly in small populations. Felsenstein comes up with a rule of thumb that natural selection is effective in the face of drift provided at least one individual every other generation dies (or is sterile). This is an impressively low amount of death, suggesting there may be something in the basic premise of the Darwin awards http://en.wikipedia.org/wiki/Darwin_Awards.

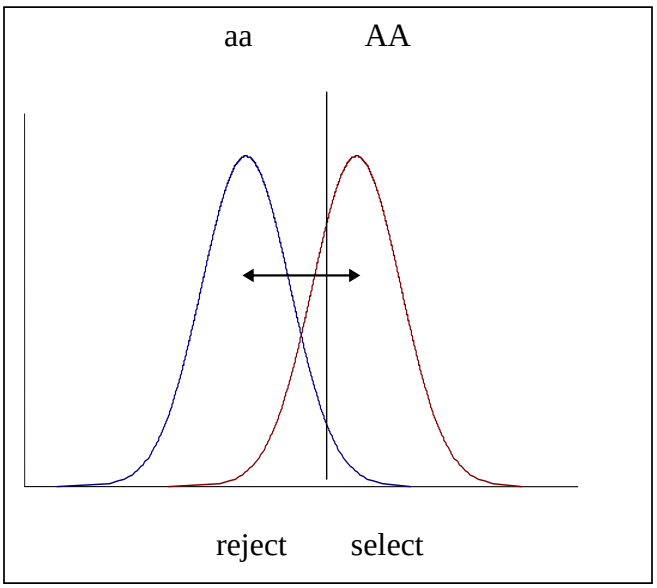
new mutation = p = 1/2N

$$\begin{aligned} U(p) &= (1-e^{-4Nsp})/(1-e^{-4Ns}) \\ &= (1-e^{-2s})/(1-e^{-4Ns}) \\ &\sim (1-e^{-2s}) && \text{for } 4Ns >> 1 \\ &\sim 2s && \text{for small s.} \end{aligned}$$

For example, with s = 0.01 and N = 1,000, the probability of fixation calculated by the Kimura formula is 0.0198 – very close to 2s. Although small, this is still forty times the probability of fixation under drift alone of 0.0005.

Selection on a quantitative trait

Clearly, is should be possible to relate selection on a quantitative trait to selection on single loci.



For a normally distributed quantitative trait with additive gene action, the difference in means between the homozygote and heterozygote at each biallelic locus is *a*, or 2*a* between both homozygotes. Under truncation selection, the proportion of selected individuals carrying the selected genotype is the fitness of the genotype. Provided we known the value of *a* and the phenotypic variance of the population (more strictly the residual variation within each class, but if *a* is small we can ignore that subtlety) then these proportions can be calculated. The relative proportions then give the relative fitnesses of each class and can be used to estimate the selective advantage at each locus. F&M show that, provided *a* is small:

$$s \sim i2a/Vp$$

where *i* is the intensity of selection.

This formula is correct for dominant, recessive, or additive traits (where the difference in fitness between the homozygote and heterozygote is defined as ½ *s*. If *s* is small, then additive and multiplicative models are equivalent, provided we take care in the

parametisation of the fitnesses. In our case, we treat the difference in fitness between homozygote and heterozygote as *s* rather than *s/2* so

$$s \sim ia / \sqrt{Vp}$$

Given this we can then insert values of s into the formula for the probability of fixation of an allele:

$$U(p) \sim (1-e^{-4Nsp}) / (1-e^{-4Ns})$$

We can then study the probability of fixing advantageous alleles with varying effect on our phenotype.

So for example, we can calculate that with 50 loci of equal effect, initial allele frequency of 0.5 (so you would expect about 25 to be segregating in the average cross between inbred lines), heritability of 0.1, selecting 10 out of 100 diploid lines (or 20 out of 200 doubled haploids), s is 0.097 and the probability of fixation of a locus is 0.875. In other words, roughly 6 of the loci will be fixed for the wrong allele. If we increase heritability to 100%, then the probability of fixation increases to near 1. This isn't strictly correct, since as selection progresses, Vp will change as allele frequencies change and loci get fixed. For such small numbers, one may as well simulate the whole thing.

More than one locus.

The practical importance of considering more than one locus has been driven by the advent of association mapping in populations. It is important to have some understanding of the forces which influence frequencies of multiple-locus genotypes. Fortunately, in practice, we rarely need to consider more than two loci at a time.

Linkage equilibrium

The Hardy Weinberg equilibrium allows us to predict genotype frequencies from allele frequencies. The multilocus equivalent is that we can predict gamete frequencies, or haplotype frequencies, from the allele frequencies at the individual loci making up the gamete or haplotype. From these haplotype frequencies, we can go on to predict multi-locus genotype frequencies by treating each haplotype as if it were an allele at a multi-allelic locus.

Most of our discussion will centre on two loci with two alleles: A a and B b. These are separated by recombination frequency θ. Allele frequencies are:

A	p
a	q = (1-p)

B r
b s = (1-r)

There are four possible haplotypes (more strictly gamete types if the loci are on different chromosomes, but we'll ignore this distinction):

AB
Ab
aB
ab

The equilibrium frequencies of these gamete types are:

$f_{AB} = pr$
$f_{Ab} = ps$
$f_{aB} = qr$
$f_{ab} = qs$

In tabular form:

	r B	sb
p A	pr AB	ps Ab
q a	qr aB	qs ab

Just as we introduced F to account for departure of single locus genotype frequencies from these expected values, here we introduce D, the coefficient of disequilibrium

$+D = f_{AB} - pr$
$-D = f_{Ab} - ps$
$-D = f_{aB} - qr$
$+D = f_{ab} - qs$

or

	r B	sb
p A	pr +D	ps - D
q a	qr - D	qs +D

It is easy to verify that these frequencies total to 1. In passing we note that only a single parameter is required to be added to the 2 x 2 contingency chi-squared table to give a

perfect fit, which is why a 2x2 contingency chi squared test has only 1df. This arrangement of haplotype frequencies in a contingency table gives us a clear steer about how to test for the presence of linkage disequilibrium.

Note that the sign of D is arbitrary. If we had defined fAB-pr as -D, then the absolute value of D would be unchanged but the sign would switch.

The interpretation of D

The coefficient of linkage disequilibrium, D, is not easy to interpret. Its range depends on allele frequency and is not symmetrical about zero. It has an absolute maximum value of |0.25| when allele frequencies at both loci are equal. To aid interpretation, two transformations are commonly used: D' and Δ^2 or r^2.

D' If D < 0, D' = D/ minimum $\{p_A p_B, (1-p_A)(1-p_B)\}$

If D > 0, D' = D/ minimum $\{p_A(1-p_B), (1-p_A)p_B\}$

This looks complicated, but all it is doing is acknowledging that if D increases without constraint, then eventually the frequency of AB or ab will become zero. A haplotype can't have a frequency less than zero, so this sets an upper limit on D, determined by which of AB or ab happens to have the lowest observed frequency. Similarly the lowest (negative) value of D is determined by the value at which Ab or aB is zero. D' is just D scaled by its maximum value if it is positive and by its minimum negative value if it is negative.

D' ranges from -1 to +1. Generally therefore, it is the absolute value of D' that is quoted. |D'| will take a value of one when, of the four possible haplotypes, only three are observed. When a new mutation occurs, it creates a new haplotype: a single copy carrying the mutant and one of the alleles at the other locus. So where there were initially only two haplotypes there are now three, but there could be four. To create the fourth haplotype we require an identical mutation on a chromosome carrying a different allele at the other locus (very unlikely) or we require recombination. Chronologically, the stages are:

1)	AB	Ab			initial population
2)	AB	Ab	aB		third haplotype created by mutation, D' = 1.
3)	aB x Ab		→ Ab		recombination
4)	AB	Ab	aB	Ab	four haplotypes present, D' <1

Following recombination, D' is <1. The value of D' can therefore serve as a test for the occurrence of historical recombination between two loci: if |D'| is less than one, then

recombination must have taken place in the time since the second polymorphism arose through mutation. This is referred to as the four gamete test.

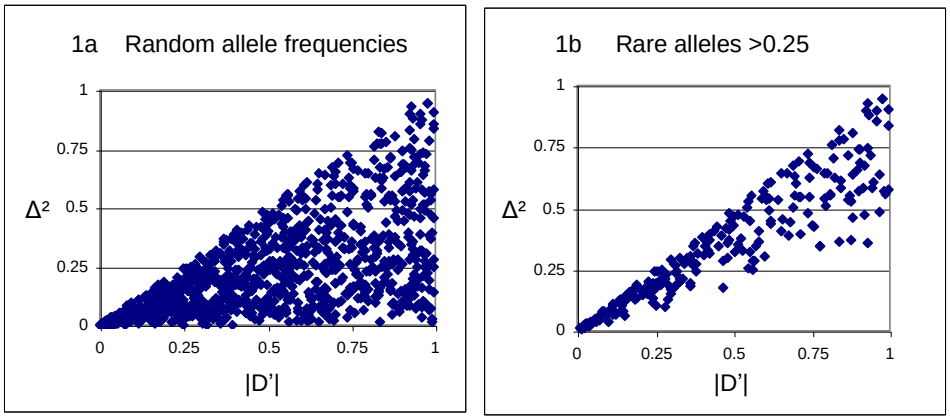
Δ^2 or r^2 $\Delta^2 = D^2/(pqsr)$

Δ^2 also ranges from 0 to 1. For a randomly mating population, it is the correlation coefficient squared between the two loci if the alleles are given numeric codes. It therefore has the advantage that it is very easy to calculate. Δ^2 will take a value of one when, of the four possible haplotypes, only two are observed. This is easy to see, since the correlation coefficient is +/- 1 only when one variable perfectly predicts the other. This is only possible if the first locus is a perfect predictor of the second, in which case allele states at the two loci must match. Since Δ^2 is a measure of predictability, it is useful for deciding appropriate marker densities and in studying the power of association to detect QTL where power is proportion to Δ^2 .

At extreme allele frequencies, D' and Δ^2 can take quite different values. At intermediate allele frequencies, their values tend to correlate. |D'| is never smaller than Δ^2. The graph below shows a plot of D' against Δ^2 for some simulated arbitrary values of p_A, p_B and D. Note that for any value of |D'|, Δ^2 ranges from zero up to that value. Δ^2 is never greater than |D'|. |D'| is more likely to take high values at extreme allele frequencies. This effect can be seen more clearly in figure 1b, which plots the data from figure 1a after removing loci with allele frequencies less than 0.25. It can be seen that at intermediate allele frequencies, |D'| and Δ^2 measure much the same thing.

Figure 1

Comparison of measures of LD measures



Measures of disequilibrium for multi-allelic loci

Generally, with markers such as SSRs, D, D', and Δ^2 are estimated allele by allele: comparing a single allele at one locus with a single allele at the other with all the remaining alleles lumped together into a single class for each marker. With n_1 and n_2 alleles at the two loci we have $n_1.n_2$ measures of LD. A single multi-allelic measure of LD is the average, weighted by allele frequencies, of these $n_1.n_2$ measures. This isn't entirely satisfactory: estimates of LD with rare alleles tend to be inflated as it is very easy to get a population sample which is missing many of the multiple possible haplotypes and this drives the estimate up, often dramatically. Sometimes rare alleles are lumped together before LD estimation and this can improve things. Sometimes resampling methods are used to estimate the magnitude of the bias empirically and this bias can be subtracted from the observed estimate.

The decay of linkage disequilibrium with time

LD decays through recombination. Recombination can only occur between the doubly heterozygous individuals:

AB/ab and aB/Ab

Such individuals will occur at a frequency of:

$2(pr + D)(qs + D)$ for AB/ab

and

$2(ps-D)(qr-D)$ for Ab/aB

in a randomly mating population.

We can assess the change in D by following the change in frequency of any gamete. We'll choose AB.

From AB/ab individuals, the frequency of AB gametes is $(1-\theta)/2$ (the non-recombinants)

From Ab/aB individuals, the frequency of AB gametes is $\theta /2$ (the recombinants

The frequency of AB gametes produced by these two main types is therefore:

$$2(pr + D)(qs + D)(1-\theta)/2 + 2(ps-D)(qr-D)\theta/2$$

This involves terms in θ and terms which do not involve θ. If θ is zero, there is no change in gamete frequencies. The change in gamete frequencies is therefore given by the terms in this formula which involve θ. These are

$$[- (pr + D)(qs + D) + (ps - D)(qr - D)]\theta$$

$$= - \theta [(pr + D)(qs + D) - (ps - D)(qr - D)]$$

$$= - \theta D$$

So the change in D over a generation is $-\theta D$. The value of D in the next generation is therefore

$$D - \theta D$$

$$= D(1 - \theta)$$

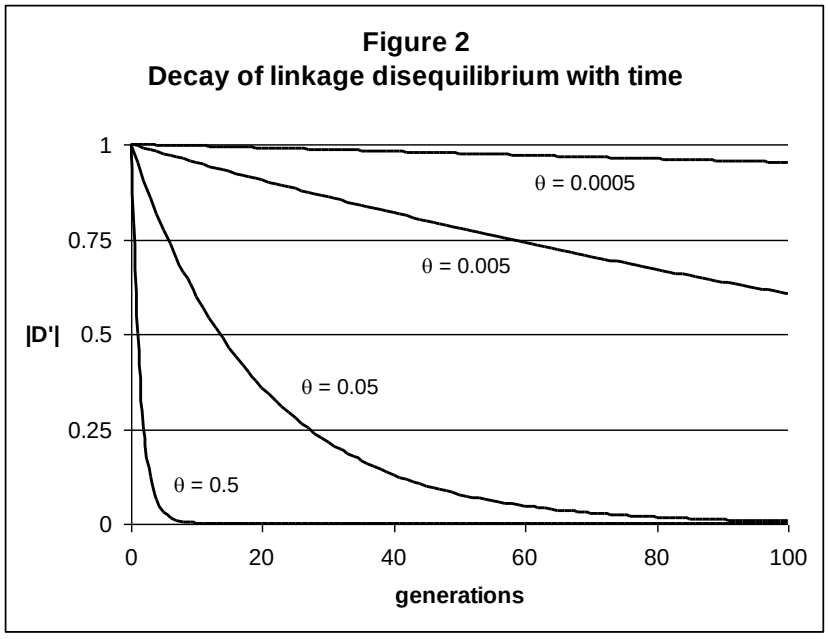
and over t generations

$$D_0 = D_0(1 - \theta)^t$$

which to a good approximation (Taylor's series) is

$$D_0 = D_0 e^{-\theta t}$$

for small θ and large t. This latter form shows that in the longer term, time and recombination are roughly equivalent – a halving of recombination fraction is compensated for by doubling the number of generations. Figure 1 shows the decay in linkage disequilibrium over time at a series of recombination fractions.



Linkage disequilibrium decays very rapidly in the absence of linkage but persists for a very long time with very tight linkage.

The effect of inbreeding

For inbreeding species, the decay in linkage disequilibrium over time is reduced. In the most extreme case, if the population consists of a set of inbred lines with no intercrossing, there is no opportunity for recombination and linkage disequilibrium is fixed. If some outcrossing occurs however, linkage disequilibrium will decay although at a slower rate. The effect of inbreeding in pedigree breeding programmes is an interesting example. Assuming that all varieties are fully inbred, the formula for the rate of decay of LD

$$D_t = D_0 (1-\theta)^t$$

will still apply provided the definitions of θ and t are modified. t is no longer the generation time, but the cycle time: the time taken to produce a set of progeny lines from a set of parents. θ is no longer the recombination fraction per generation, but the cumulative proportion of recombinants occurring from one cycle to another. For fully inbred lines this is 2r/(1+2r) where r is the true, generation-wise recombination rate (Haldane & Waddington; 1931). For closely linked markers (<2 cM say), 2r/(1+2r) ~ 2r. With a cycle time of eight years (this should be poor in a well run modern breeding programme but is probably reasonably accurate historically), the rate of decay of LD per generation is then roughly:

$$D_t \sim D_0 e^{-\theta t/4}$$

LD decays at about a quarter of the rate found in a truly randomly mating population with the same generation time. Of course this figure will be perturbed by the overlapping generation structure that breeders impose but it acts as a guide: in spite of the inbreeding nature of many crop plants, LD will be decaying among cultivated varieties as a result of recombination, though it may be generated by other forces within the breeding programme as we shall see.

Causes of linkage disequilibrium

Mutation

Consider a single polymorphism with two alleles, A and a, segregating in any reasonably large population. Suppose a new mutation, B → b say, occurs somewhere on a chromosome carrying the A allele. In the population as a whole there will be three haplotypes:

AB with a frequency very close to p_A
aB with a frequency very close to $1-p_A$
Ab the new mutant, carried on a single chromosome

There are four possible haplotypes in total, but only these three are observed, so |D'| = 1. In successive generations, assuming that the new b mutation is not lost from the population by drift but ultimately rises in frequency, the missing haplotype, ab, will be created by recombination. As we saw in the preceding graph, this can take a very long time for closely linked markers. For the majority of markers available for genotyping, mutation must have occurred a long time ago as many generations are required for allele frequencies to rise from a single copy to a frequency which makes genotyping worthwhile. The levels of linkage disequilibrium attributable to mutation will therefore only be high among very closely linked markers (or markers and QTL). Provided a sufficiently high marker density can be achieved, this situation is very favourable for association mapping.

In humans, it is common to find values of |D'| equal to 1 among very closely linked markers, often accompanied by high values of Δ^2. This indicates that little or no recombination has occurred among these markers. The pattern of LD in crop plants is less clear. Data are beginning to accumulate however. Among wild populations of *Arabidopsis*, an extensive survey has revealed that LD decays quickly – within 50 kb - even though this is an inbreeding species. (Nordborg et al. 2005).

Population bottlenecks, founder effects and drift.

A population bottleneck is an extreme reduction in population size. This might occur as a result of disease nearly wiping the population out, an environmental disaster or some other catastrophic event. A particular form of population bottleneck, a founder effect, occurs when a species colonizes a new niche or environment. Initially the population size can be extremely small. For a wild species only a few seeds might be carried to an island. For a crop species, only a few seeds or transplants may have been introduced to establish the crop in a new country. Any restriction in population size will generate LD. An F2 can be regarded as an extreme case: the population is established from two gametes in the preceding generation. As a result, levels of LD are at a maximum. However because linkage analysis occurs within a generation of the founding event, there has been little opportunity for LD to decay and it is hard to locate QTL accurately. Generally, the magnitude of LD generated by a bottleneck or founder effect is less extreme, but is still sufficient for association mapping. In crop plants, the activities of plant breeders themselves can result in population bottlenecks - the advent of a new disease or desired agronomic trait such as reduced height may result in a period of breeding in which only a small number of parental lines are used, or one or two lines are used very extensively for introgression.

In fact, any finite population size generates some degree of LD, in the same way that genetic drift always causes some change in allele frequency, whatever the population

size. For a population of constant size, a steady state is set up in which the expected value of Δ^2 is:

$$E(\Delta^2) = 1/(1+4N_e\theta)$$

Note the similarity between this equation and that for expected homozygosity under drift and mutation in the infinite alleles model.

Selection

Selection on a trait will change allele frequencies at QTL determining the trait. In addition, allele frequencies will change at markers closely linked to the QTL. This is called hitchhiking. Its effect is to generate LD among markers around the region of selection. A region of increased LD, often accompanied by a reduced amount of polymorphism compared to other genomic regions, can be a signature of selection – a sign that a particular region has been subjected to selection pressure. Such regions have been identified in many species; in plant most notably in maize and *Arabidopsis.*

Migration and population admixture

If two populations, formerly isolated, are brought together, LD can be created. This is a result of allele frequency differences between the two source populations, which may have arisen through drift or through selection. For example:

haplotype	pop 1	pop 2	combined	expected	difference
AB	0.04	0.64	0.34	0.25	0.09
Ab	0.16	0.16	0.16	0.25	-0.09
aB	0.16	0.16	0.16	0.25	-0.09
ab	0.64	0.04	0.34	0.25	0.09

In population 1, $p_A = 0.2$ and $p_B = 0.8$. In population 2, the frequencies are reversed. Within each population there is no linkage disequilibrium (for example $p_{aB} = p_a.p_B = 0.2x0.8 = 0.16$ in population 1).

If the two populations are intermixed, without any crossing, the haplotype frequencies are just the average of the separate population frequencies. However, the allele frequencies are averaged too, such that $p_A=p_B=0.5$ and linkage disequilibrium is generated. In fact, $D = 0.09$, $|D'| = 0.36$ and $\Delta^2 = 0.13$.

With more modest rates of migration or gene flow from one population to another, the generation of disequilibrium is less severe. Provided migrants intermate with the host population, the disequilibrium will decay in successive generations.

Migration can be both an asset or a problem in association mapping. If population admixture is known to have occurred and if markers are available which discriminate, even imperfectly, between the two parental populations, then these markers can be used to map traits for which the populations differ. This is "admixture mapping". It is the population based equivalent of mapping in an F2: instead of two parental inbred lines, there are two parental populations. In human genetics there is considerable interest in this method, particularly in the USA: Afro-Americans are known to have about 10% European ancestry and are therefore a suitable group in which to map traits for which Africans and Europeans differ. Suitable populations for admixture mapping in plants may exist, for example in crosses between Flint and Dent maize or in hybrid zones of Populus.

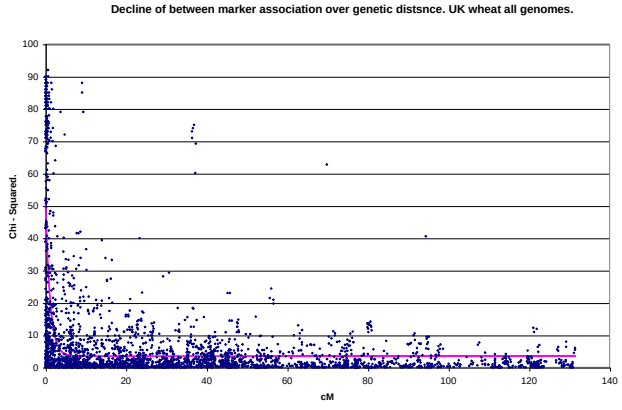
Generally, migration is a problem. If we are trying to exploit linkage disequilibrium arising from mutation or an ancient bottleneck, recent migration introduces long range LD which can mask the marker-trait associations arising from close linkage which we wish to fine map.

Summary of causes of LD

Linkage disequilibrium can arise from many causes. Current evidence shows that LD is generally higher between closely linked loci and that it declines with distance. However, instances of longer range LD do occur. There is therefore a major risk that associations between a QTL and a marker are not the result of close proximity but may arise from other causes which have not been taken into account. In practice, in any population, forces generating new LD and the decay of existing LD will both be occurring. Patterns of LD can therefore be complex. The requirement for successful association mapping is to detect and correct for long range associations arising from recent events while locating close range LD arising from mutation and historical population bottlenecks.

Plotting and modelling LD.

Plots of pair-wise measures of LD against genetic distance generally show a pattern something like the one in the figure below



LD generally decays with the distance separating the markers, but there is a lot of noise around the decay: LD is sometimes observed between markers which are a long distance apart, and there are often many pairs of markers showing little or no LD in spite of their proximity. So although the patterns almost always show that that LD decays with genetic distance – and this is why LD patterns can be used to map QTL – there is often a lot of noise. We shall make a few observations to keep in mind when reading accounts of LD decay and when modelling ones own data.

1) Rare alleles. When a new mutation has recently occurred or been introduced into a population, it is generally at very low allele frequency and there will have been little opportunity for recombination. In general, most rare alleles are young. In addition, rare alleles tend to give biased estimates of LD. The pattern of LD decay will therefore often appear much cleaner if LD is expressed as Δ^2 rather than D', or if markers with rare alleles are first excluded. In both cases, greater weight is being placed on older alleles, so LD will appear to decay more quickly – there has been more time for recombination.

2) Genetic and physical distance. LD patterns can be plotted against chromosome location, measured as a genetic distance (eg in cM) and/or physical distance. In humans, where data exist on a very fine scale, LD has been found to occur in blocks. There are small runs, of the order of 10s or 100s of kilobases, within which little or no recombination has occurred, followed by small gaps within which there is much recombination. Such blocks of LD are formed by selection, statistical artefact (the sampling distribution of markers along the chromosomes), and by recombination hot spots – confirmed by extremely fine linkage mapping using sperm. The functional basis of these recombination hotspots is not clear, and they are not conserved between humans and chimps, suggesting that they are not inherited as DNA. It is not hard to find papers in crop genetics however, which refer to linkage blocks, recombination hotspots and the like and interpret their data in this way. This may be correct; there is certainly limited recombination around the centromere, but generally I think it reflects an unjustified over-enthusiasm to transfer findings from humans to crops. Firstly, in crops, the marker density is almost always much lower than in the typical human study, so the sorts of blocks resulting from recombination hotspots in man cannot be the same as those identified in crops, even if they are genuine. There are several aspects by which patterns of chiasmata in plants differ from those in human, so we should be wary of translating results from one to the other too readily. This is not to say that reported LD blocks in crops are not genuine, or that they may not relate to recombination patterns. It is worth plotting pairwise LD against chromosome location, but there is a need for caution in interpretation, I feel.

3) There are two types of curve with a simple genetic expectation which can be fitted to LD data.

$$E(\Delta^2) = 1/(1+4N_e\theta)$$

which is appropriate for the long term equilibrium between drift and recombination. More sophisticated forms incorporate mutation.

$$E(D') = H+(H-L)(1-\theta)^t$$

This is the "Malecot model," named after the dead French population geneticist, which is more appropriate for decay of LD in the more recent past. H is the value of D' at zero recombination fraction, and L is the value of D' for unlinked markers or markers distantly placed on the chromosome. In an idealized population H is 1 and L is 0, giving the formula for decay of LD with time that we derived earlier. L can be viewed as background LD arising from a population admixture and migration.

In human genetic data, the first equation and other more sophisticated modelling methods are favoured, although the Malecot model is championed robustly by Newton Morton (the inventor of the LOD score in 1955 and still active at the University of Southampton). Both these formulae will give a curve which shows some form of decay in LD with distance and can give a fair fit to the data. Note that one method fits a curve to D or D' and the other to Δ^2. In practice the Malecot model can also be fitted to Δ^2 ,or to D' after filtering on allele frequency, though the interpretation of the parameters is less clear. Both are really only appropriate to idealized populations of the sort we will be very luck to meet when analyzing real data. However, the parameters we can estimate do have some sort of genetical meaning which will help in our understanding and they allow comparison of patterns between chromosomes and populations. When both seem to fit the data equally well, the parameter estimates from one may be silly. For example, setting L and M to 0 and 1 for simplicity, a value T= 18 with the Malecot curve gives a similar shaped curve to Ne = 10. However, whereas T = 18 seems plausible, Ne = 10 seems far too low, even for UK winter wheat. For decay of LD among modern crop varieties I feel the Malecot model is more appropriate. If nothing else, the Malecot curve is more flexible since LD is not forced to start at a value of 1.

There are also more complicated methods of modelling LD involving coalescent theory, but these are beyond me and are not generally applied to modelling at the whole chromosome level.

LD in a biparental cross

In most linkage analysis experiments, we start with two inbred parents - two gametes essentially. D' is therefore 1 among all possible pairs of gametes. Yet in the F2, when we start mapping, the expected LD among unlinked pairs is zero. It does not decay to a value of ½ as we might expect. This is because we have not mated the two parents at random. Half the mating should have been selfing of the parents. If this had occurred, then the second generation, instead of being 100% heterozygous, would segregate at 1:2:1 and be in HW equilibrium at each marker, but D' values would now be ½ among unlinked markers. Such a population would generate many more false positive results were it used for linkage analysis. So there is a hidden benefit of non-random mating in this case. Whether such a benefit exists in more diverse, but non-random mating populations for association mapping, I do not know.

Haplotypes

A haplotype is a set a set of genetic markers located on the same chromosome that are sufficiently closely linked to be inherited as a unit. If recombination occurs between markers within a haplotype, two new haplotypes are created. There are sometimes advantages in considering variation within a sequence or region of the genome in terms of its constituent haplotypes rather than by analyzing the constituent markers independently. Often, there will be fewer haplotypes than there are marker-allele combinations (2^n for n biallelic markers). This reduction in numbers can provide increased power. It may be possible to reconstitute evolutionary relationships from the haplotypes and this can provide more information in association analysis too. The construction of the evolutionary tree is not that difficult if there has been no recombination and there have been no duplicate mutations, in which case a "perfect phylogeny" can exist, but with recombination, gene conversion and repeat mutations this is difficult: again beyond this course and my abilities.

Identifying the two constituent haplotypes carried by an outbreeding diploid individual presents its own problems: AaBb individuals can be carrying AB ab haplotype pairs or Ab aB. Over short ranges they can be distinguished by sequencing. If family or pedigree information is available they can often be uniquely determined too. There are several statistical genetics methods available to do this too and we may be able look at these in the tutorial sessions.

QUANTITATIVE GENETICS / BIOMETRICAL GENETICS

Books: Kearsey and Pooni.
Falconer and Mackay (no relation)
Felsenstein has a chapter on quantitative genetics.

Most of the variation breeders are interested in is continuous. Quantitative genetics is the study of characters which are measured rather than classified (red, green, blue for example). Historically, there have been two approaches. The first concentrated on analysing variation in experimental crosses between two inbred parents and in generations derived from this cross, the F2, backcrosses and so on. This approach has generally been used by breeders and researchers of plants, particularly of inbreeding crops. The second approach concentrated on describing variation in populations and was adopted by breeders and researchers of animals, in which the production of inbred lines to generate large experimental F2s and related populations is much harder. There are strengths and weaknesses to both approaches. Generally, I feel breeders, particularly of inbred crops, would benefit from more knowledge of the quantitative genetics of populations. We shall therefore largely follow a population approach.

Means and variances

Consider a single major gene in HW equilibrium with two alleles:

A_1A_1	A_1A_2	A_2A_2
p^2	$2pq$	q^2

The average phenotype of each genotype can be written as:

m+a	m+d	m-a

m is a base effect, defined as the mean of the two homozygous classes.

A_1A_1	A_1A_2	A_2A_2
	m	
+a	d	-a

.

+/- a is then the deviation of each homozygous genotype from m.
d is the deviation of the heterozygous class from m: the dominance deviation.

d = 0 means there is no average heterozygous effect - the locus has additive inheritance.
d = a represents complete dominance of the A_1 allele over A_2
d = -a represents complete dominance of the A_2 allele over A_1 or equivalently A_1 is recessive.

d >|a| represents over or under dominance

The population mean for this locus is

$m + p^2a + 2pqd - q^2a$

$= m + (p-q)a + 2pqd$

since $p^2 - q^2 = (p+q)(p-q) = (p-q)$

Note that the population mean depends on allele frequency, not surprising really; we expect selection to change allele frequency and thus increase the mean for the traits we are selecting for. More on this later.

Over several independent (biallelic) loci, the population mean is the sum of the effects at each locus:

$\Sigma [m + (p-q)a + 2pqd]$ where summation is over all loci affecting the trait.

$= \Sigma(p-q)a + 2 \Sigma pqd$, ignoring the sum of the constants.

The variance is:

$$p^2a^2 + 2pqd^2 + q^2a^2 - [(p-q)a + 2pqd]^2$$

$$= p^2a^2 + 2pqd^2 + q^2a^2 - p^2a^2 - q^2a^2 + 2pq\,a^2 - 4p^2q^2d^2 - 4pq(p-q)ad$$

$$= 2pqd^2 + 2pq\,a^2 - 4p^2q^2d^2 + 4pq(q-p)ad$$

which turns out to be:

$$2pq[a+d(q-p)]^2 + 4p^2q^2d^2$$

$2pq[a+d(q-p)]^2$ is referred to as the additive genetic variance Va.

$4p^2q^2d^2$ is referred to as the dominance variance Vd.

The total genetic variation (on summing over loci) is then $Vg = Va + Vd$.

Of course, the phenotypic variance, Vp, includes environmental variation among individuals: even if there is no genetic variation a trait can vary. Treating the environmental and genetic variation as independent (this assumption can be dropped if required):

$$Vp = Vg + Ve = Va + Vd + Ve$$

While Vd has some intuitive appeal - it is the contribution of dominance to the population mean all squared $(2pq)^2$ - Va is a mess. However, as we shall see shortly, it is the genetical description of the component of total variance which is responsible for response to selection, and we can use this to justify its description as the additive variance. Moreover, these rather complicated terms reduce to much more simple expressions in the case of equal allele frequencies or purely additive gene action.

1) $p = q = ½$ an F2

mean = m (defined with reference to the F2 only)
$Va = ½ a^2$
$Vd = ¼ d^2$

In this case, the mean and variance are more simply derived by writing down the expectations for the F2 directly.

2) No dominance: d = 0

$Vg = Va = 2pqa^2$

Interaction terms can also be described in a similar manner but we shall restrict these notes to additive and dominance effects only.

Effect of inbreeding on the mean and variance

Assuming the original population is randomly mated, the effect of inbreeding can be quantified by the population inbreeding coefficient.

A_1A_1	A_1A_2	A_2A_2
$p^2+pq(1+f)$	$2pq(1-f)$	$q^2+pq(1+f)$

The average phenotype of each genotype can be written as before as:

m+a	m+d	m-a

The population mean is:

$= m + (p-q)a + 2pqd(1-f)$

Thus, in comparison to an outbred population, the mean is changed by an amount 2pqdf. If d is positive, then we shall observe a reduction in the mean - inbreeding depression - but the extent of the reduction is dependent on allele frequencies.

The effect of inbreeding on the genetic variance is harder to quantify. Here, we shall assume that the contribution of dominance to the genetic variance is negligible to make things easier. This assumption is not as bad as it may sound, especially for higher levels of inbreeding: there are then fewer heterozygotes in the population, so the contribution to the total variance coming from dominance effects will be smaller. Note that although we are assuming that dominance *variance* will diminish in importance with inbreeding, we are not assuming that inbreeding depression cannot be severe: when inbreeding is complete there is no dominance variation but inbreeding depression is at a maximum. For the purposes of describing the genetic variance under inbreeding therefore, we assume no dominance variation. For a population in HW equilibrium:

$Vg = Va = 2pq^2$

With inbreeding, it is easy to show:

$Vg = Va(1+f)$

Of note is when inbreeding is complete: f = 1 and

$Vg = 2Va.$

For equal allele frequencies: the case for homozygous lines derived from an F2:

$Vg = 2Va = a^2$

As before, all these expectations are for a single locus. To get the total genetic variance we sum over all loci.

In all these cases, terms for genetic variances can be extended to include expectations for interactions between loci, but we have no time to go into those here. In practice, the inclusion of interaction terms in genetic modelling makes little difference to the conclusions of relevance to plant breeding.

These formulae for genetic variance are all fine, but where to they get us? First we need to extend them to compare variances between and within different family types and across generations. We can then use them to do useful things like predict responses to selection from selecting among different family types and with different selection schemes.

Parent offspring regression

The most important genetic relationship is that between parent and offspring. In terms of statistics alone, if we regress offspring performance on the mean of the parents, then we shall have a regression coefficient which permits the prediction of offspring from the performance of parents. We can then use this relationship to calculate response to selection, measured on the offspring, from performance of the parents. However, this regression also has a simple genetic relationship which allows the prediction of response to selection without the requirement to directly compare progeny with parents.

The offspring-parent regression coefficient is covariance (offspring-parent) / variance(parent). We shall follow the derivation of the genetic expectation of these variances and covariances given in Falconer and Mackay Chapter 9, p150. We shall derive the regression for progeny mean on the mean of both parents (the mid-parent). We assume the population is mating at random, so the genotypes of the possible parental combinations are the product of the corresponding genotype frequencies under HW:

mating type	freq	parental mean	progeny type	progeny mean	progeny x parent
$A_1A_1xA_1A_1$	p^4	a	A_1A_1	a	a^2
$A_1A_1xA_1A_2$	$4p^3q$	½(a+d)	$½A_1A_1$ $½A_1A_2$	½(a+d)	$¼(a+d)^2$
$A_1A_1xA_2A_2$	$2p^2q^2$	0	A_1A_2	d	0
A_1A_2 xA_1A_2	$4p^2q^2$	d	$¼A_1A_1$ $½A_1A_2$ $¼A_2A_2$	½ d	$½d^2$
A_1A_2 xA_2A_2	$4pq^3$	½(-a+d)	$½A_2A_2$ $½A_1A_2$	½(-a+d)	$¼(a+d)^2$
A_2A_2 xA_2A_2	q^4	-a	A_2A_2	-a	a^2
mean		(p-q)a+2pqd		(p-q)a+2pqd	$[(p-q)a+2pqd]^2$

$$cov_{(o/p)} = p^4a^2 + 4p^3q¼(a+d)^2 + 2p^2q^20 + 4p^2q^2½d^2 + 4pq^3¼(a+d)^2 + q^4a^2 - [(p-q)a+2pqd]^2$$

After some basic, but potentially error prone, algebra, this simplifies to:

$$pq[[a+d([q-p]^2 = ½\ Va$$

The phenotypic variance among the parents is Vp. We are regressing onto the mid-parent, so the phenotypic variance among the (mean-of-two-parents-chosen-at-random) is just

$$Vp/2 = (Va + Vd + Ve)/2$$

The covariance we have derived is entirely genetic - we are assuming there is no environmental or error covariance between progeny and parents. This is often, but not always, true in plants since parents and offspring are often raised in different environments or years or, if raised in the same environment, are laid out in a suitably randomised trial design which guarantees the covariance is zero. The worst thing one could do is lay out parents and progeny in adjacent plots (it has been done). In animals and humans, the error covariance between parents and offspring (and also between members of the same family) is often not zero - families tend to be raised in the same

environment. Such common environment effects are also referred to as household effects (in humans), in utero effects and litter effects, depending on the case in hand. Seed quality effects can be important in crops, especially row-crops where plant emergence and early vigour can have a huge effect on final yield. In sugar beet, seed quality effects can dwarf the genetical contribution to the determination of variety performance, making variety assessment difficult, to say the least.

The regression of offspring on mid-parent is therefore:

$$\tfrac{1}{2} Va \;/\; \tfrac{1}{2} (Va + Vd + Ve) \;=\; Va/(Vp) = h^2_n$$

The ratio Va/Vp is called the heritability, or more correctly the narrow sense heritability and represents the proportion of the phenotypic variance which is attributable to additive genetic variation. There is another heritability, the broad sense heritability

$$h^2_b \quad = \; Vg/Vp \; = (Va + Vd)/Vp$$

which represent the proportion of the phenotypic variation which is genetic, whatever the cause.

We have derived the covariance of offspring on mid-parent. The covariance of offspring on single parent also turns out to be ½ Va. (This equality is a result of the random mating of parents, it does not apply if mating is not at random, see Falconer & Mackay on assortative mating). However, the variance among single parents is Vp and not ½ Vp. The regression of offspring on single parents is therefore $\tfrac{1}{2} h^2_n$ - half that for regression on mid-parent.

Heritabilty and the prediction of response to selection

Now that we know the regression coefficients of progeny on their parents, we can use these to predict the performance of the progeny. Take the mean performance of the parents of a single plant to be *x*. Then the predicted performance of the progeny *y* is:

$$y = bx + c$$

Since this relation is linear, the equation is also valid for any subgroup of parents and progeny. Let s and p refer to the mean of a selected group.

$\bar{y} = b\,\bar{x} + c$ where $\bar{x}$ and $\bar{y}$ are the means without selection

$y_s = b\,x_s + c$

Define response to selection as the increase in mean performance of the selected progeny compared to the mean if there were no selection. Then the response to selection is:

$$(y_s - \bar{y}) = b(x - x_s)$$

$x - \bar{x}$ is termed the selection differential S: it measures how hard we select among the parents.

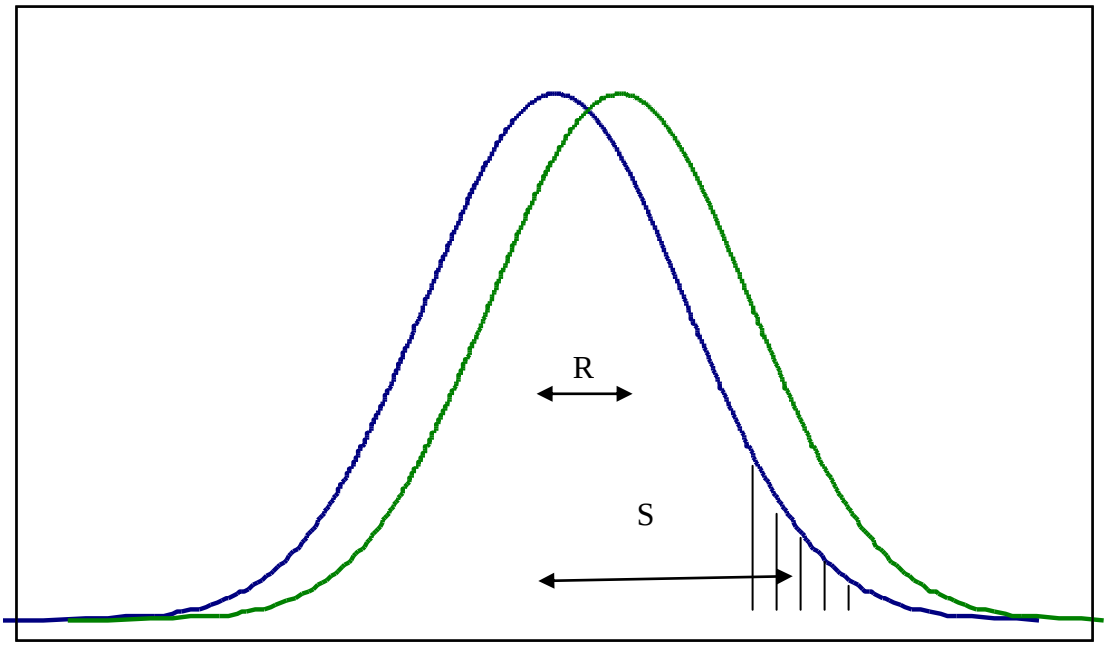
$y - \bar{x}$ is the response to selection R. If the selected parents have been pollinated by the population as a whole, then

$$R = \tfrac{1}{2} h^2 S.$$

For selection based on both parents, we have:

$$R = h^2 S.$$

This follows from the results of the previous section.



$R = h^2S$ is "the breeders equation." It is the most fundamental equation in all breeding, whether of plants or animals. If some proposed new technique or breeding scheme cannot be shown to increase R, or some conditional measure of R (eg R adjusted for cost and generation time), then it should be rejected. If it has no effect on R, it is not part of breeding.

There are some provisos on the use of this equation. Fortunately, most don't matter (see Falconer and Mackay).

Response to selection clearly depends on the selection differential. This in turn depends both on how hard you select and on the variability within the parental generation. Selecting the top few percent of individuals is termed truncation selection: it is equivalent to selecting all individuals above some threshold value of phenotype, say z. For a standardised normal distribution, $N(0,1)$, S is termed the intensity of selection and is tabulated for corresponding proportions selected in, for example Falconer & Mackay and Kearsey & Pooni. Or it can be calculated as:

$$i = \Phi(z)/p$$

where i is the intensity of selection, p is the proportion selected, and $\Phi(z)$ is the probability density function for a standard normal distribution at the truncation point z.

$$\Phi(z) = (2\pi)^{-1/2}\, e^{-\frac{1}{2} z^2}$$

That is, take the truncation point z, corresponding to the proportion selected (eg 1.96 for 2.5% selection), calculate the point probability density, $\Phi(z)$, then divide this by the proportion selected.

For distributions with different variances, S is calculated by taking the value for the standardised normal distribution and multiplying by $\sqrt{Vp}$ ie σ_p. So we have

$$R = ih^2\sigma_p$$

where σ_p is the phenotypic variation.

This in turn is often reparametised as

$$R = ih\sigma_g$$

This makes explicit that response depends on how hard you select, the precision with which your trait is assessed, h, and the available genetic variation. Different breeding schemes sometimes have different generation times, so response may need to be defined as response per year rather than response per generation:

$$R = ih\sigma_g / y$$

There remains a slight problem with i - its value has a slight dependency on the size of the population in which you select. For populations of size 400 or more, it doesn't matter, but for populations less than this, i declines even though the proportion selected remains constant. A good approximation is given by redefining the proportion selected, p, as

$$p = (k + \tfrac{1}{2}) / (n + k/2n)$$

where k = the number selected
n = the total population size.

The value can also be calculated by taking the average of the top k expected normal deviates from a population of size n. Methods have been provided to calculate i in "genetics odd and sods.xls".

The principle of predicting response to selection from the regression of progeny on parents can be easily extended to selection among entities other than outcrossed individuals, although we end up using the term heritability more loosely: to describe the proportion of genetic or additive genetic variation expressed between those entities. For example they could be averages over replicate plots of inbred lines. This is fine, provided it is made explicit to what h^2 refers. Unfortunately, in plant breeding literature, this is often not the case. With plants, we may be selecting between diverse family types, inbred lines, clones, hybrids and so on, measured on single plants or multiple plots. When studying estimates of h^2 and when comparing estimates between different studies take care.

We shall take as a contrasting example to that of selecting individuals, the selection among a set of inbred lines for submission to a second stage of testing, recommended list trials say. Here, there is no recombination so the "progeny" are merely the selected sample of inbred lines selected for testing in the next stage. Suppose the inbred lines are first assessed in r replicate plots, and the error variance among the replicate plots is Ve.

Then the phenotypic variance among all inbred lines:

$$Vp_{(inbred)} = Vg_{(inbred)} + Ve/r$$

The 'heritability' of the mean on the inbred lines is

$$h^2_{inbred} = Vg_{(inbred)} / (Vg_{(inbred)} + Ve/r)$$

Note that this heritability can be increased or decreased by increasing or decreasing r and can approach a value of 1 if r is sufficiently large. In practice, r will be limited by seed availability.

The selected lines are assessed the following season in a separate experiment. The number of replicate plots here are irrelevant. The only source of covariance between the lines in the two experiments is genetic, so

$$cov_{(offspring\text{-}parent)} = Vg_{(inbred)}$$

The response to selection is therefore:

$$R = i.h^2_{inbred}.Vp_{(inbred)}$$

We can then compare differences in response to selection arising from differences in allocation of resources to increased replication or to increasing the number of lines tested. Changing these will alter heritability and intensity of selection and there will be an

optimum balance between the two. Note that the genetic expectation of $Vg_{(inbred)}$ has not entered into the calculations. For samples derived from a randomly mated population this can easily be found to be

$$Vg_{(inbred)} = 4pqa^2$$

which is 2x the additive variation in the outbred population (assuming no dominance variation). The reason we can get away without knowledge of the genetic expectation here is that we are not passing through a sexual cycle and measuring response to selection in the following generation. In fact, the response to selection by intermating the selected inbreds to create the immediate next generation (a set of F1s) is more complex - the covariance cannot be easily expressed in terms of Va unless there is no dominance variation.

In summary therefore, we note that in plant breeding we are often interested in prediction of two types of response to selection, the first a prediction from one generation to the next (and beyond, see later) and the second, more easily estimated, is within a generation.

Genetic variances and covariances from other family types.

We shall not derive genetic variance components for other family types: the procedures follow those outlined above for the offspring-parent covariance. We may derive one or two in the tutorial for practice. I've given the most commonly encountered family types below. These allow comparisons of responses from different selection schemes – examples will be studied in the practical sessions. There is one additional complication: many of the family types have an expectation for genetic variance within families in addition to between families. These can be derived by subtraction of the between family variance from the total or directly by calculation of the within family variance and taking the sum weighted by the frequency of family types. Nb for compactness, we're only listing the genetic variance, we must remember that both between and within family variances will be affected by error.

Note that the error term attached to a between family variance will include a term reflecting the genetic variance within families:

$$Vg_{within} /n$$

where n is the number of plants contributing to the family mean. This term can be important in animal breeding where the family size is often quite small, but in plants the family size is usually large (think of the number of plants in the typical breeder's plot of any crop) and so can fortunately be ignored. These complications are dealt with thoroughly by F&M. None the less, it is worth while knowing the magnitude and nature of genetic variation within families since some selection schemes can include a component of single plant selection within families in addition to selection between families (often these two types of selection are for different traits). All expectations are

for lines derived from an initial randomly mated population. This is perhaps not that realistic for F1 hybrids, where crosses are often made between different populations which have diverged to some extend, nevertheless it will do for now. F1s could perhaps be treated using the Wahlund effect.

	between	within	mean
individuals	Va + Vd	N/A	Σ [m+(p-q)a+2pqd]
clones	Va + Vd	0	Σ [m+(p-q)a+2pqd]
full sibs	½ Va + ¼ Vd	½ Va + ¾ Vd	Σ [m+(p-q)a+2pqd]
half sibs	¼ Va	¾ Va + Vd	Σ [m+(p-q)a+2pqd]
S1 progenies *	1½ Va	¼ Va	Σ [m+(p-q)a+pqd]
S2 progenies *	1¾ Va	1/8 Va	Σ [m+(p-q)a+½pqd]
fully inbred lines *	2Va	0	Σ [m+(p-q)a]
DH lines *	2Va	0	Σ [m+(p-q)a]
F1s	Va + Vd	0	Σ [m+(p-q)a+2pqd]
4-way crosses.	½ Va + ¼ Vd	½ Va + ¾ Vd	Σ[m+(p-q)a+2pqd]

$Va = \Sigma 2pq[a+d(q-p)]^2$
$Vd = \Sigma 4p^2q^2d^2$

* Variance components for selfed families are not easily expressed in terms of Va and Vd. Those given here are under the assumption of no dominance. As discussed above, this is reasonable since with inbreeding dominance variation is of less importance.

Note the equivalence of many of the terms. As is well know, the genetic variance of doubled haploids and fully inbred lines is identical. This is true because we are considering the genetic variance to be the sum of effects over independent loci. More recombination is involved in producing inbred lines than doubled haploids and we shall study the effect of this see later. If QTL are unlinked, there is no difference between inbred and doubled haploid lines.

Less well recognised is the equivalence of genetic variance components between single plants, clones and F1 hybrids. This is worth dwelling on. In the absence of inbreeding, the genetic variation among single plants is as great as it can get for non-inbred families. For equivalent heritabilities and intensities of selection, therefore, the response to selection by selecting single plants (within a generation) is as great as anything. Of course, the single plant heritability for traits such as yield in crops is generally extremely low - zero is a good first approximation. Clonal propagation provides offers the opportunity through replication to increase response to selection by increasing heritability but is often not an economic way of selling or distributing a variety and it is not an option for seed crops (unless viable systems of apomixis are developed; a sensitive subject for many commercial breeding companies). F1 hybrids also offer the opportunity to select and fix the equivalent of the best single plants within a generation. Viewed in this way, F1 hybrids are not a mechanism for exploiting heterosis but rather a mechanism for achieving the equivalent of clonal propagation in a seed crop. Note in addition that the response to selection among F1s within a generation will exploit both additive and dominance variation, but that the response across generations will only exploit the

additive variation, since the covariance between parental F1s and progeny F1s has no dominance component. Depending on the relative magnitudes of Va, Vd and Ve therefore, it is possible for response to selection within a generation to be greatest among F1s or clones, but response over generations to be maximised by selecting among inbreds. Of course inbreeding depression, or its mirror image heterosis (they are one and the same) complicate the interpretation, but the point remains: a set of F1 hybrids are genetically equivalent to a randomly mated population.

These components of variation can be used directly to predict response to selection within a generation in exactly the same way. To use them to predict across generations is slightly harder, since we need to know the covariance between the family types being selected and the progeny. However, if neither parents nor offspring are inbred, this is just the additive genetic component of between families variation, ½Va for example for full-sib families. It is true only if the selected units alone contribute to the next generation. If, for example, single plants were tagged for selection after seed is set, then those single plants will have been pollinated at random and the response to selection is reduced by half: the response to selection is made up of a response due to selecting females and a response due to selecting males. In this case, only the females have been selected.

Assuming no dominance variation, the covariance between an inbred line and its F1 progeny is Va, If we pair selected inbred lines to produce these progeny, the genetic variance among the parental pairs is 2Va/ 2 = Va. The regression of F1 performance on the mean of the two inbred parents can therefore approach 1 provided dominance variation is not large and environmental variation (controlled by replication) is low. Equally, the covariance between a parental inbred line and recombinant inbred progeny produced is also Va. As a result, the response to selection among the recombinant inbreds is predicted to be the same as the response to selection within the current generation.

All this stuff on the prediction of response to selection from the heritability may seem pointless After all, in pure statistical terms once we have the equation for the regression of offspring on parent, we can predict response directly from this equation without any knowledge of the genetic expectation. This is true. However, because we know the genetic expectation we can estimate Va, Vd, Ve and heritabilities by any method and use these components directly to compare response to different selection schemes without the need to implement the scheme itself. This is the fundamental principle in the design of breeding schemes. A common example would be to compare response to selection among half-sib families with response among full-sibs. Full-sib families will generally have a higher heritability, but half-sib families are cheaper to produce so more can be grown and intensities of selection can be higher. To quantify these sorts of comparisons we require methods of estimating genetic and environmental components of variation which can then be used to consider alternative breeding and selection schemes.

Estimating genetic variances and means – F2 derived populations

The modelling of genetic effects is easiest among lines and generations derived from a biparental cross. A thorough account of this approach is in K&P. A glib summary of the approach is that means and variances are fitted to combinations of parental, F1, F2, BC1 and BC2 generations and variances are estimated among F2 plants and/or doubled haploid or inbred lines derived from the F2. Statements for that cross can then be made about the importance of additive and dominance variation, the magnitude of inbreeding depression, transgressive segregation, the expected phenotype of the best and worst lines that can be extracted from the cross, and so on. This approach has provided a lot of knowledge about the genetic architecture of quantitative traits in both plants and animals and has therefore been influential in discussions about breeding strategy and the most appropriate type of variety to produce - inbred or hybrid. It has also provided a number of quantitative geneticists with training which allowed them to enter careers in genetics and breeding with varying degrees of success. However, its direct impact on the course of practical plant breeding has been quite limited. Part of the problem is that when working within a cross among lines derived from a population, one is only working with roughly half the available genetic variation, the rest is expressed between crosses. (On deriving inbred lines by selfing, Vg between cross means = Vg within crosses for a population in equilibrium.) Possibly more of a problem, for practical application, was that predictions for each cross could not be made quickly enough. By the time the appropriate crosses had been made and experiments carried out, then lines had already been developed and selections made for the next cycle of crossing. (If they weren't, the breeding programme was not very efficient.) For completeness, below, we give expected means and variances for the most commonly encountered generations. Remember, all variances are relative to the F2, with the advantage that there is no dominance component in the estimate of additive variation (= ½ Σa^2)

	mean			variance within			
	m	[a]	[d]	V_a	V_d	V_{ad}	Ve
P1	1	1	0	0	0	0	1
P2	1	-1	0	0	0	0	1
BC1	1	½	½	½	1	1	1
BC2	1	- ½	½	½	1	-1	1
F1	1	0	1	0	0	0	1
F2	1	0	½	1	1	0	1
F3	1	0	¼	1½	¾	0	1
F∞	1	0	0	2	0	0	1

The use of [] around a and d requires some explanation. In the P1, some alleles which increase performance will be fixed, and some which decrease performance will be fixed. The same is true of P2. The difference between P1 and P2, summed over all loci which are segregating in the cross is therefore a net effect. The [] symbolises this net effect. It is quite possible that the P1, P2 have identical means, yet additive genetic variation within the cross is still detected, because the additive genetic variance is gross not net: ½Σa^2 . In fact this difference between the expectation for the means and variances can be exploited

to provide estimates of the degree of dispersion of increasing and decreasing alleles between the two parents – see K&P for details. Note however, although the dominance effect is written as [d], this is unaffected by the degree of dispersion of increasing alleles between the two parents, it doesn't matter which parent donated which allele, the F1 is always going to be heterozygous.

The term Vad, which only involves the backcross generations, is a summed cross product of additive and dominance terms, with a value which depends on the degree of dispersion. It occurs because in the backcross, allele frequencies are no longer half. See K&P for details.

Estimation and testing of significance of these parameters is also described in K&P. From the table alone, it is clear how some may be estimated (eg [d] = F1 – ½P1 – ½ P2), but these are not necessarily the best estimates. Data from multiple families may give better estimates and different families mean are generally known with differing degrees of precision.

We will mention one crude but cheap and cheerful estimate of genetic variation which is sometimes available, often for new or minor crops and/or newly introduced traits rather than for major crops. This is that Ve can be estimated directly from variation within the non-segregating generations: the parents and F1. Subtraction of Ve from variation among single F2 plants then gives an estimate of (Va +Vd) and thus the broad sense heritability.

Note also that the variance components listed above are for the total variance for each generation. Variance in the F3, for example, could be partitioned into variance within F2 family groupings and variation between F2 groupings. Details are to be found in K&P.

Risks of over fitting models

The model fitting approach introduced here can be extended to include means and variance components of epistasis (interactions between loci). There is a risk however, that one can get carried away with this model fitting process. My favourite example comes from an experiment in barley involving a subset of the families listed above. Specifically, the F1 was not included. (This is reasonable, the experiment was conducted in plots, and it is impractical to carry out sufficient hand crossing to generate the required quantity of F1 seed.) A simple model including [a] and [d] failed to fit the data so higher order interaction terms were included. The resulting model fitted the data well. However, the estimates of the parameters allowed a prediction of the F1 performance. The predicted yield was negative. That is to say, not merely lower than the low yielding parent but actually negative. Needless to say, this was not pointed out in the paper. It is highly likely that the consequences of competition between plants within plots differed between generations and distorted the estimates of genetic parameters. This is a Type III error: the initial model failed not because of epistasis, as the authors assumed, but for some other reason. None the less, it illustrates perfectly the risk with model fitting, especially as the number of parameters is increased, in explaining away the data in a manner divorced

from biological reality. This is true of much statistical and genetical modelling: the practice is probably quite prevalent and most statisticians/data hacks, including me, are likely to be guilty of it on occasions. This case is rare in that it is quite explicit that something had gone badly wrong.

"The glitter of the *t* table diverts attention from the inadequacies of the fare." Sir Austin Bradford Hill. The environment and disease: association or causation? Proc R Soc Med 1965 **58** 295-300

(Bradford Hill was the epidemiologist who first demonstrated the link between smoking and cancer in 1950.)

Cross prediction.

One attempt to make the biometrical approach more immediately relevant to plant breeding was cross prediction. Here, on the basis of estimates of the mean and additive genetic variation within a cross, predictions of responses to selection within crosses could be made. If means and variances were estimated over a series of crosses, then selection could be made between crosses for those showing the greatest potential to develop improved varieties. The problem in implementing this scheme is that Va must be assessed very quickly. If a population of lines has to be produced and raised to estimate Va, then why not just select among the lines you have already got and get on with it? To overcome this, it was proposed that Va could be estimated approximately from the variance among F3 family means. However, to my mind there is a more fundamental problem. Suppose our set of crosses is among inbred lines derived from a randomly mated population. Then at each locus segregating in the population, there will be a proportion

p lines of genotype AA
q lines of genotype aa

Heterozygous F1s will occur with frequency 2pq at each locus and only the progeny of these crosses will be genetically variable for that locus. If there are n independent loci segregating in the population, then the number of segregating loci will follow a binomial distribution with variance 2npq(1-2pq). For modest numbers of QTL segregating in the whole population at intermediate frequency, the variation in number of heterozygous loci (and therefore in Va) from cross to cross is too small to be detectable. For example, n = 30, p = 0.5 gives 15 loci segregating per cross on average, but the standard deviation of this number is only 2.7. Only if QTL of particularly large effect are segregating, or if allele frequencies are extreme, are differences in variance likely to be detectable and therefore worth including in a selection strategy. (Differences in variance generally require quite large experiments to detect.) Nowadays, in such cases, QTL detection followed by MAS is more likely to be effective.

An interesting example of this comes again from barley. A small set of crosses between and among 2 and 6 row barley were produced. Cross prediction showed that the variances of the 2 x 2 row crosses were all similar but that those between 2 and 6 row barleys were larger. Most routine barley breeding in the UK is between 2 row barley. Estimation of variances to improve selection of crosses has nothing to offer here. Crosses between different populations can generate increased variation, but that is hardly a surprise.

The most robust parameter to estimate from a cross is the mean m, and in the absence of epistasis this can be predicted as the mean of the two parents. So the approach of crossing the best with the best would appear to be the most effective, at least in the absence of additional information. Quantitative genetic methods have little to add here, although there remain breeders who will argue against this approach.

Heterosis

Under simple genetic models, for an F1 to exceed the performance of the best performing parent, there must either be overdominance at some of the QTL involved in the trait, and/or there must be some dominance, not necessarily complete, but with the increasing alleles dispersed to some extent between the two parents. Of these two explanations, quantitative genetic analysis favours dispersion. To some extent, this corresponds to common sense too. Dispersion of QTL between parents is nearly always found – even when extreme lines are crossed, including crosses between wild and cultivated forms. In fact, any response to selection would be impossible if QTL were not dispersed, otherwise the best parent would be as good as it is ever going to get.

Combining ability

In F1 hybrid evaluation, different crosses frequently share parents. It can be useful in such cases to estimate the average effect of each parental line. When applied to sets of crosses, these average effects are called general combining abilities. Depending on the quality of the phenotypic data, it can be more efficient to select on general combining ability (GCA) than on the performance of the lines themselves. GCA is more commonly used to predict the performance of crosses which have yet to be made; to suggest novel hybrids worth creating and testing.

The deviation of the performance of a hybrid from its value predicted from GCA is termed Specific Combining Ability (SCA). Aside from experimental error, SCA measures the interaction between the GCA of the two parental lines. Although hybrids with high SCA (and high GCA) are best, SCA cannot easily be predicted in advance; usually the cross must be created. Attempts are continuing to be made to find methods of predicting SCA, using marker data for example, but as far as I am aware these have not been particularly successful so far. The relative magnitudes of variation in GCA and SCA are of interest however, in the design of efficient breeding and testing programmes. With lines derived from a randomly mated population, the variance in GCA results from

additive genetic variance and that in SCA from dominance variance (ignoring complications due to epistasis).

Among F1 hybrids the genetic expectations of GCA are and SCA are

Vgca = ½ Va
Vsca = Vd

The total variation among all hybrids is Va + Vd. Randomly mating the inbred parents reconstitutes a randomly mated, non-inbred population as we discussed earlier. In combining ability terms this is made up of:

2Vgca +Vsca

with half the GCA variation contributed by the male parents, and half from the females. Note that those hybrids which contribute to the GCA of a parent are related as half-sibs. However, Vgca is ½ Va and not ¼ Va: the variance is inflated by a factor of two because these half-sib lines have inbred parents.

As the expected variance among GCA is all additive, why bother estimating GCA, why not just select on inbred performance? Although the inbred parental lines may have average performance substantially different from the average of crosses, the variances are

$\Sigma pq[a+d(q-p)]^2$ for GCA and

$4\Sigma pqa^2$ for the inbreds (ie 2Va)

The covariance between inbred and GCA is:

$2pq[a^2 + (q-p)ad]$

The correlation between inbred and GCA is therefore

$2pq[a^2 + (q-p)ad] \; / \; [\, 4pqa^2 \,.\, 2pq[a^2 + (q-p)d] \,]^{½}$

which = 1.

It is clear therefore, that under simple genetic models, parent and hybrid should have high correlation, so that selection may as well proceed on the inbred performance. This is the view taken by K&P. An alternative view can be found in much of the maize breeding literature. In practice, correlations are frequently quite low, but there are many non-genetical reasons why this may be the case too. Maize breeders continue to argue the toss with epistasis and overdominance being invoked to explain the low correlations. Simulations by OS Smith ("Covariance between Line per se and Testcross Performance"

Crops Science 1986 **26**:540-543) indicated that there is nothing mysterious about the low correlations often seen between inbred and hybrid performance and that they are to be expected even under simple genetic models. More recently, (Troyer & Wellin *Crop Sci***.2009**:49:1969-1976) a review of heterosis studies in maize, coupled with insider knowledge of the Pioneer Hi-Bred breeding programme has concluded that much more emphasis should be placed on direct selection among inbreds for yield.

***Estimation of GCA*.**

Consider first the case where male and female inbred parents are different. If the crossing scheme is complete, then analysis and estimation can proceed exactly as for a two-way analysis of variance. The GCAs are estimated as simple averages across all crosses with a common parent. Generally, in practical breeding programmes, the crossing scheme is incomplete. Not all pairwise crosses have been made. However, GCAs and variance components for males and females can be estimated by standard methods for incomplete crossing schemes.

When males and females comprise the same sets of lines, then this simple procedure must be modified somewhat, even when all n(n-1)/2 crosses have been made. The reason for this is as follows. Suppose we have four lines and all six crosses have been made. Summing across the (n-1) crosses involving parental line 1:

$T_1 = 3\mu + 3gca_1 + gca_2 + gca_3 + gca_4$.

GCA is the deviation from the overall mean so $\Sigma gca_i = 0$

$T_1 = 3\mu + 2gca_1$

So we must divide ($T_1 - 3\mu$) by n-2 and not by n-1, to get the correct estimate of GCA. There is more on this in F&M.

The set of all n(n-1)/2 crosses among the n parents is termed a half diallel. A full diallel involves all n^2 possible crosses (ie crosses, reciprocal crosses and the parents selfed. The detailed full analysis of the diallel (the Hayman analysis "The analysis of variance of diallel tables." Genetics 1954 **39**:789-809.) provides for estimates of variance components, the detection of gene interaction, dominance effects, reciprocal differences and so on, but is rarely carried out. K&P state that the number of parents is generally too small to allow confident inferences about the population from which the lines were drawn. A related concern is that the lines cannot often be regarded as a random sample from the population. A way around this is to regard the n parents as the founder lines of a new population (AJ Wright "Diallel designs, analyses, and reference populations." Heredity 1985 **54**:307–311) . The circumstances in which one would wish to do this within a plant breeding programme are limited, however.

The analysis of diallels can be difficult with standard statistical software so it is worth searching out procedures written specifically for them. (There is one in GenStat, I've not come across one for R.) Because the male parents and female parents are a common set, one cannot introduce males and females simply as two column vectors to feed into your favourite statistical software. Instead, you must have a column for each of the n parents, with a 0, 1 or 2 in each cell to indicate whether the parent is represented in that particular cross 0, 1 or 2 times. (2 times is for inbred parents). Then you would need to reduce the n columns to n-1 independent columns (subtract the last one from all the others). This is tedious to say the least. A simple way around the problem, proposed by R Thompson ("The use of multiple copies of data in forming and interpreting analysis of variance." In: K. Hinkelmann (Ed.), Experimental Design, Statistical Model and Genetic Statistics, 1984 Ch 11, pp155–171.), is to take two copies of the data, treat the male and female parents as if they were independent, but swap them over in the second copy. Then carry out the analysis across both sets. The total SS will be 2x too large. The males SS will equal the females SS and will also be two times to large. The residual SS will be two times too large. From the standard computer output, the correct analysis of variance can therefore be constructed.

Estimation of variances in populations

Generally, data will consist of measurements on sets of full sib families, half sib families, and sometimes selfed progenies. The simplest way to proceed is to estimate variance components using REML, and equate these with the expectations given in the table given earlier. Variance components can also be estimated by equating mean squares with expectations in an analysis of variance of these data, but if you have access to as statistical package which implements REML, this is easier. The only pitfalls to be aware of are those in analysing any set of trial data. If using variance components to compare breeding programmes (see the practical sessions) then we often end up assuming that dominance variation is negligible. For example we might estimate variance components empirically from S1 families but then use those components to predict the effectiveness of selection on full-sib families. We can generally get away with this since most of the programmes we wish to compare are based on family selection and the component of dominance variation in the family means is generally quite small. We are also fortunate, in plant breeding, to be working with large family sizes so the contribution of within family genetic variation to the family mean (V_{within} / family size) can be ignored too. Another pitfall is to ignore common environment effects at your peril. These will inflate between family variance components. Seed quality effects are important in crops. In crops where plants are grown as spaced plants, and in crops where vegetative rather than seed organs are harvested, the effect can be enormous. However, even is grain crops, there can be an effect.

Response to selection in the longer term – the Bulmer effect

We have discussed how components of variation and means can be used to predict the response to selection over a single generation and how this can be used to compare and evaluate alternative breeding programmes. Unfortunately response to selection in the long term is not a simple matter of multiplying the response predicted for the first generation by the number of generations. A theory of everything has been developed which takes into account mutation, drift and changes in allele frequency. It is not easily applied to plant breeding programmes and may not be necessary; prediction over modest numbers of generations is probably all that is required. For example, in wheat, the most rapid breeding will give a cycle time of about four years (including DH production, seed multiplication and yield assessment). Ten generations is then 40 years, within which time period it is likely that disease pressure, climate, consumer preferences and economics will all impact on breeding objectives to the extend that the predictions made at the start, even if accurate, become less relevant. Breeding objectives, environment and germplasm change.

1) Selection without recombination.

Suppose we have a set of inbred lines, select 10% according to some criterion, then retest the remaining lines and select again. How do we predict the response to selection in the second cycle. This is a statistical problem rather than a genetic problem, and was first studied by Pearson (I think) and revived in the 1960s by Curnow, Finney and Young, who were considering the problem of sequential selection: how to allocate resources over a series of two or three years worth of testing in which some selection is carried out at each stage. How intensely should we select at each stage? What is the optimum replicate number at each stage? Are you better off testing all varieties in only a single replicate, or would you be better off throwing half of them out so that the remainder can be tested in two replicates? This is all of great relevance to breeders, arguably more so than anything else in quantitative genetics / statistics.

Suppose we have data for a normally distributed trait on a set of lines from which we select the best x%. The difference in mean between the selected group and the whole population (ie S, the selection differential) is $i\sigma_p$. The selected group will have a reduced variance too, given (without proof) as:

$$Vp' = [1-i(i-z)]\, Vp$$

We've already discussed i. z is just the truncation point for selection on a standardised normal distribution. It can be looked up in tables of the cumulative normal distribution, or calculated directly using functions like normsinv in Excel.

However, what we need is not the reduction in phenotypic variance, but the reduction in the genetic variance on retesting. On retesting, since environmental deviations for the original data and the retested set are independent, the only reduction in variance must be attributable to the reduction in genetic variance. If the heritability on initial testing was 1,

then one would expect all the reduction in variance to be carried forward. If the heritability was zero, then there will be no reduction. It turns out that the expected reduction in variance is given as:

$$Vg' = [1-i(i-z)]\ h^2\ ]Vg$$

That is, the factor by which the phenotypic variance is reduced [1-i(i-z)] is multiplied by the heritability. As ever, we need to make sure that h^2 and Vg are the correct values for the family means or individuals among which we are selecting.

Once we have Vg' , we can predict response to selection from a second cycle of selection by calculating h^2 for this second round, which if replicate numbers and plot sizes remain the same will be lower than in the first generation.

This approach assumes that the distribution of genetic effects, after selection, is normal. This isn't true, but it has been shown, nevertheless, that the theory works well. Even better, we can apply this method another one or two times to give predictions of response over three or four (at a pinch) cycles of selection. This approach has been used to study optimum allocation of resources in sequential selection schemes and led to the Finney rule-of-thumb: the optimum allocation of resources comes from selecting equally intensely at each stage of selection and allocating equal resources at each stage of selection. For examples, with 200 varieties tested over two cycles of selection, we could select 20 for retesting before selecting a final set of two. We would test the 200 in single replicate trials, but the 20 in 10 replicates each.

This procedure assumes no GxE, or equivalently that Vg applies to all environments and not just the initial testing environment. Design of testing programmes in the presence of extensive GxE is better studied using computer simulation. Nevertheless, this little bit of theory provides a quick and easy method of evaluating alternative testing programmes. In practice, most breeding programmes allocate too little resource to testing in the later stages of selection.

Suppose our selection process is among a set of lines, derived initially from a randomly mated population. We can consider what has happened, in genetical terms, to the variance. First think about a pair of loci, A and B, with alleles at equal frequency, acting additively:

genotype	AB	Ab	aB	ab	coeff of disequilibrium
score	1.2	1.1	1.1	1	
freq	¼	¼	¼	¼	0

Assume selection response is a direct function of the score.

Unadjusted freq after selection:

1.2/4	1.1/4	1.1/4	1/4	

divide by total:

0.272	0.25	0.25	0.227	-0.00052

As a result of selection, there is a change in allele frequency, but linkage disequilibrium has also been generated between the two loci. Although this disequilibrium effect is slight, with many loci affected the trait, there are very many pairs of loci. With many loci, it is the generation of this disequilibrium rather than change in allele frequency which accounts for the reduction of the genetic variance after selection.

This can be formalised and made more explicit using something called the infinitesimal model, which is, in fact, the basis for much quantitative genetics theory. This assumes that a trait is governed by an infinitely large number of QTL, each of infinitesimally small effect. Over a few generations, for traits governed by only modest numbers of loci (30 say) it works remarkably well. An advantage of this model is that it can generate a finite response to selection with no change in allele frequency.

More formally, over pairs of loci, the total genotypic variance is:

$Vg = Vg_A + Vg_B + 2cov_{(AB)}$

Under the infinitesimal mode, allele frequencies do not change as a result of selection so Vg_A and Vg_B remain constant. Directional selection forces cov(AB) to be negative and Vg is therefore reduced. Other types of non-random mating and selection - assortative mating and disruptive selection for example - can cause $2cov_{(AB)}$ to become positive and Vg to increase.

This genetical interpretation has no impact on our consideration of the effects of selection within a single generation but is important for consideration of the effects of selection across generations.

2) Selection over several generations.

Consider a population maintained by pairing individuals at random in each generation, with no selection. The additive genetic variation among the progeny in each generation can by partitioned into a between parental pairs component with value ½Vg and a within pairs component, also ½Vg.

Moving on another generation, we can partition the variation into between grandparents (¼Vg), between parents within grandparents (¼Vg) and within parents (½Vg). Another generation would give the series 1/8, 1/8, ¼, ½ and if we went far enough back we could get a contribution to the current genetic variance of $½^n$ from the nth generation ancestors. The contribution of each ancestral generation is eroded at a rate of ½ per generation, to be replaced by variation within families - termed segregation variation - which emerges at a rate of ½Vg per generation and keeps Vg constant.

$$Vg = ½\ Vg_p + 1/2\ Vg_w$$

Now add in selection among the parents. After selection, but before generation of progeny, the genetic variation among the progeny is Vg'. Half of this variation is passed forward to the next generation so that:

$$Vg = ½\ Vg_p' + 1/2\ Vg_w$$

So half the reduction in genetic variance is carried forward to the next generation. Going forward another generation, without any more selection, half the parental contribution is passed forward to become the grandparental contribution:

$$Vg = ¼\ Vg_{gp}' + ¼\ Vg_p + 1/2\ Vg_w$$

and so over successive generations the reduction in Vg due to that single generation of selection will decay at a rate of a half per generation. However, if selection continues in each generation, then using subscript t to denote generations:

$$Vg_{(t+1)} = ½\ Vg_t[1-i(i-z)\ h^2\ ] + ½Vg_0$$

This recurrent formula can be used to compute the decline in genetic variance over successive generations, to an equilibrium value at which

$$Vg_{(t+1)} = Vg_{(t)}$$

Note that the decline in variance can be severe, reducing Vg to up to half its initial value. However, once selection stops, it is reversible since i, and therefore i(i-z) is zero, although the recovery may take a few generations to work itself through the system.

This reduction in additive genetic variance on selection as a result of the generation of negative disequilibrium, is termed the Bulmer effect after its discoverer Bulmer in 1976. The theory can be used to compare alternative schemes for recurrent selection over several generations. It can also be used to consider if an additional generation of random mating, without selection, could improve response to selection per year or per generation by increasing Vg. This has been proposed and implemented by some breeders. We'll try it in the tutorial for inbred crops, where a cycle of selection might take 4 years, but two additional rounds of random mating might add only a year.

The result above considers that selection acts to reduce the additive variance only. However, it is good enough even in the presence of dominance variation and can be extended to include linked loci. Ultimately these more complicated scenarios are probably better treated through computer simulation. It is clear however, that starting from a position of no loss of Vg through disequilibrium, linked loci will make things worse since the decay of disequilibrium through random mating is slower. If loci are smeared at random over chromosomes, and the number of chromosomes is reasonably large, then most pairs of loci will be unlinked and the additional effect of linkage on the reduction in genetic variance may be slight. If loci are located in linked clusters (eg gene families) the effect could be greater. I am not aware that this has been studied in the light of current knowledge of genome organisation.

We shall briefly discuss two applications of the above theory. Firstly, selection between full sib families. Here the variance between families is ½ Va. This component of variance will be reduced by selection.

	before seln	after seln	next generation
between families	½ Va	½ Va$[1-i(i-z)h^2)]$	¼ {Va+Va$[1-i(i-z)h^2)]$}
within families	½ Va	½ Va	½ Va

This is less messy than it seems – only the between family component of variation is reduced by selection, and half of this reduction is passed on to the next generation. Within families variation is due to segregation and is unaffected by selection. A similar approach works equally well for other family types and can be used to consider, for example, complicated schemes of recurrent selection involving alternative cycles of full-sib family selection and S1 progeny testing within selected families (Mackay and Gibson – that's me: hooray).

This approach is also useful in predicting the response to selection within pedigree breeding programmes (Cornish). In the absence of selection, at each generation of inbreeding from an F1, the additive variance has the following expectation

	total	contribution from				
		F2	F3	F4	F5	F>5
F2	Va	Va				
F3	3Va/2	Va	Va/2			
F4	7Va/4	Va	Va/2	Va/4		
F5	15Va/8	Va	Va/2	Va/4		
F∞	2Va	Va	Va/2	Va/4	Va/8	Va/8

The proportional reduction in genetic variance as a result of selection within any generation can be predicted as $i(i-z)h^2$ as before. This reduction is passed on unchanged to the next generation, when additional variation is generated by segregation from residual heterozygosity within the line of descent. An expected additive variance component for the next generation can therefore be calculated and used to predict selection within this generation too. In this way, differing pedigree breeding schemes can be compared. In particular, the benefits of early generation selection, when heritabilities are typically low (small plots, low replication number) can be assessed. [Class exercise] In addition the whole pedigree breeding edifice can be compared with single seed descent and doubled haploid programmes, where selection is deferred until inbreeding is complete or near complete and cycle time is generally reduced. Remember, however, that this procedure is only optimising selection within a single cross. With multiple crosses, for an additive trait, half the variance is expected to be between crosses (initially), and for a complete treatment we must account for this, and for the reduction in between cross variance as a result of selection too.

Selection limits and changes in allele frequency at a single locus

A sophisticated theory of the consequences of long term selection on quantitative traits has been developed, largely by WG Hill and colleagues. See F&M for much discussion and details. Predictions of this theory have been compared to results of selection experiments. Most of these long term selection experiments have been carried out in animals, but the best experiment, still running, is the Illinois selection experiment for increased and decreased oil content in maize, started in 1896. It is a pity that no similar resources exist in other plant species. They could easily be generated in Arabidopsis within the time course of the standard government funded grant, for example. To initiate a long term selection experiment in a crop such as wheat, however, would require a commitment to long term funding which is not currently available. Broadly speaking, the predictions from these experiments are poor but the failures are for a diverse set of interesting reasons: natural selection countering the effects of artificial selection for example. Nevertheless, the available theory makes explicit some of the consequences of selection which might otherwise be neglected and is of relevance to plant breeders. Most crops were domesticated over 1000 years ago; 10,000 in the case of many. Domestication

of crops and animals are the longest term and most successful selection experiments of them all.

Long term response to selection must take into account:

1) Drift: favourable alleles can be lost through drift, especially if initial allele frequencies are low.

2) Population size.

3) Number of loci affecting the trait. For the same initial genetic variance, large numbers of loci of small effect have the potential to give a greater response than small numbers of larger effect.

4) Mutation. Ultimately, all genetic variants will be fixed by drift of selection. Unless new variation is generated by mutation, there will be no further response.

We shall give two formulae:

In the absence of mutation, the selection limit is expected to be

$$R_{lim} = 2Ne\ iVa/\sigma_p$$

As the intensity of selection gets larger, the effective population size gets smaller. The maximum response is expected when 50% of the population are selected each generation. The time to 50% of this maximum is expected to be between 1.4Ne and 2Ne generations.

With mutation, the steady state response to selection is expected to be

$$R_{ss} = 2Ne\ iVm/\sigma_p$$

Vm is the mutational variance - the additive variation arising from mutation per generation. This can be small, experiments in Drospohila suggested a value of Ve/1000, but the response to selection can be large if Ne is large. Ne is determined by the population size after selection, not before.

F&M describe the response to long term selection as being of initial response coming from additive variation in the base population, diminishing gradually as variation is deleted through selection and drift. Then response due to new mutations becomes of increasing importance and this response should be maintained at a lower rate. The relevance of this theory and its consequences are important to plant breeding but have, as far as I am aware, not been discussed explicitly in this context. In particular, what proportion of QTL contributing to the improved performance of today's elite crop varieties originate from the wild progenitors 10,000 years ago, and what proportion was contributed by mutations post-domestication. For example, at the ppd gene in barley, the same day-length insensitive variant, which has allowed the range of barley to expand into

northern Europe, has been found within wild barley, whereas the several mutations changing ear conformity from 2 row to a 6 row ear, (viewed, with some uncertainly, to have an effect on grain yield) are believed to have originated much more recently, post domestication. It would be informative to extend these analyses to more general traits, yield and quantity essentially. How widely should we search within wild germplasm for new sources of favourable variation? There is no doubt that wild forms generally show more variation at the DNA level, but this need not necessarily translate to traits for adaptation to agriculture. Equally, wild sources of variation have been important for introducing novel disease and stress resistances. But what about yield per-se? What proportion of the increase in yield in major crops is the result of mutations which arose after domestication rather than from the selection of variants present at low frequency in wild progenitors? This seems to me to be an area worthy of additional research - treating domestication as a 10,000 year selection experiment. Surveys of variation within domesticated and wild forms have been carried out, but these could be augmented by multiple crosses between wild and domesticated forms followed by genetic analysis using quantitative and marker based methods. This approach has been very successful in the impressive but more modest Illinois long term selection experiment (Laurie et al 2004. The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics **168**:2141-2155; reviewed by Hill. 2005. A century of corn selection. Science 307: 683-684.)

Multiple traits and environments.

Breeders generally select for multiple traits. If the traits are unrelated, then there is no problem; we can select independently for each, possibly weighting effort or selection intensity by their importance. If the traits are completely correlated there is no problem either, except to ask why we are measuring two traits in the first place when one will do. The problem is how to treat traits for which the correlation is less than perfect but too high to ignore. In some instances, it is useful to treat selection in different environments for the same trait, as if we were selecting for separate traits too. This approach can have advantages yet is not something that plant breeders generally do. (Animal breeders tend to work this way more often.) Other treatments of GxE are also possible and will be discussed later.

First, some notes of caution about correlations between traits:

Correlations are not necessarily causal. Type III errors again.

Correlations can arise from linkage or pleiotrophy.

Correlations can be generated by selection: a bivariate version of the Bulmer effect. I don't think this has been studied, but it could account for the near ubiquitous negative correlation observed between yield and quality in crops and livestock, however yield and quality are defined.

Zero correlation does not necessarily imply no pleiotrophy or no linkage. Two traits can be completely pleiotrophic but show no correlation. At some loci the increasing allele for one trait may also be increasing for the other, whereas at other loci, the increasing allele for one trait is decreasing for the other. It is possible to write down very simple biochemical pathways which generate this pattern. The assumption that no correlation implies no pleiotrophy or linkage is termed "consistency of gene action" (Gale & Eaves 1972 Heredity **29:**135-149). Its prevalence and the consequences for quantitative genetic analysis and selection have not really been discussed as far as I'm aware. Equally, linkage between loci can generate positive, negative, or zero correlation, depending on whether there is an excess of coupling or repulsion linkages in the population.

The principle behind the handling of multiple traits is that, just as phenotypic variation can be partitioned into genetic and environmental effects, so too can phenotypic covariation. This covariation can, in addition, be partitioned into additive and non-additive covariation, but we won't go there. As a consequence, we can derive genetic and environmental correlations between traits. These are often worth studying in their own right. They need not be of the same magnitude, or even of the same sign, and can on occasions give insight into the relationships between traits which might otherwise be missed. A disadvantage of genetic correlations is that they are generally not estimated with great precision: their estimated values can frequently be larger in magnitude than one. Moreover, the standard error for genetic correlation is not straightforward to estimate (see F&M). A resampling (bootstrap) or simulation method may help but would require care over what to resample, especially in unbalanced datasets.

Estimation.

Expectations of covariance between traits, in terms of environmental and genetic effects, are exactly the same as for variances of a single trait, so all the results given earlier still apply. For simplicity here, we shall assume we are dealing with sets of large families which are assessed in plots. Between replicate plot (co)variation can therefore be regarded as all environmental, and variation between family means as genetic and partly environmental (depending on the number of replicate plots). In practice, components of variation or covariation can be estimated from an analysis of covariance (which is just like an anova except you work with sums of products rather than sums of squares) by equating observed and expected terms in the analysis. An easy way of working with pairs of traits is to analyse the traits separately, then analyse the difference or sum of the two traits. Since

$V(a+b) = Va + Vb + 2cov(ab)$

the covariance terms can be extracted by difference.

However, variances and covariances can now be estimated directly by multivariate reml - though personally I have had trouble getting convergence - with the advantage that the reml estimation can be over >2 two traits at a time, which should be more accurate. Furthermore, GenStat also gives standard errors for the variance components which could be combined to give a standard error for the genetic and environmental correlation coefficients, though this is not that straightforward either.

Correlated response to selection.

Once we have estimates of (co)variance components, of particular interest is the prediction of the correlated response to selection. This is important for two reasons. Firstly, selection on one character (yield say) may adversely effect another (quality say) and we would like to know by how much. Secondly, it may prove more economic to select a second easily measured character, possibly of high heritability, to improve the key character, but we would like to quantify the effect of this indirect selection. Either way, for a pair of characters, the prediction is made by linear regression. We select on trait x to improve trait y. The direct response to selection on trait x is:

$$ih_x\sigma_{(gx)}$$

The response measured in trait y, the indirect response, is the direct response multiplied by the regression of genetical values of y on genetical values of x. Logic dictates that we require the genetic regression because any component of the regression relating to environmental values will not be carried over into response. So the correlated response is

$$ih_x\sigma_{(gx)}\, b_{g(yx)}$$

$$ih_x\sigma_{(gx)}\, cov_{g\,/}\,\sigma^2_{(gx)}$$

This can be re-expressed in various ways, of which F&M's favourite is:

$$ir_g h_x h_y \sigma_{(py)}$$

because this is analogous to single locus response with $r_g h_x h_y$ equivalent to h^2 for the single trait. With appropriate parameter estimates, the correlated response can be compared to direct response to decide which is best.

Index selection

This is a means of optimally selecting across a set of multiple traits. Each trait is assigned an "economic value" – the cash value that a unit increase in the trait would be worth to the breeders' paymasters. Nb, economic value can be negative for a trait such as % disease infection. If we knew the genetic value of each variety we were testing, we could calculate the worth of the variety as the sum of products of the genetic and economic values:

$$\Sigma e_i g_i$$

where e_i is the economic value and g_i is the genetic value of the i^{th} trait.

However, we know only the phenotypic values, but we can construct a similar weighted score on the basis of these values:

$$\Sigma b_i p_i$$

The best set of values for b is given, in matrix form as

$$\mathbf{b'} = \mathbf{e'GP}^{-1}$$

b is the vector of weights, or regression coefficients for the phenotypic scores
e is the vector of economic values.
G is the variance covariance matrix of genetic effects for the lines under test.
P is the variance covariance matrix of phenotypic effects for the lines under test.

Note the correspondence between $\mathbf{GP}^{-1}$ and the single trait heritability Vg/Vp. For a single trait, index selection reduces to the equivalent of selection on the expected response to selection (or breeding value) rather than on the trait itself. Where all lines or individuals are known with equal precision, this makes no difference other than to reduce each variety effect proportionally, but if variety means have been estimated from varying numbers of sites and years, or from varying replicates or measurements, this approach can make selection easier even for a single trait. For this, we multiply each phenotypic deviation from the mean by its personal heritability and select on the product. This product is a Bayesian estimate of the merit of the variety which incorporates the observed data with our prior knowledge, based on the population mean and the genetic variance. We shall come across this again and in more detail in the discussion of the mixed model and association mapping. Although we often end up estimating the mean and variance from the data too, this can be justified. The advantage of this process it that the phenotypes of varieties with limited data, which are know less precisely, are shrunk towards the mean to a greater extent than those with more replication. As a result, the ranking of varieties can change such that that a seemingly fantastic variety with little data will score lower than a more modest variety with lots of data.

Note too, that if we are only interested in selecting for one trait, all values of **e** except that for the trait would be zero. This provides an easy method of selecting to improve a single trait (say yield), taking into account other measurements only insofar as they influence yield. More sophisticated selection indices can be constructed, in which trait means have optimum values, rather than merit increasing or decreasing without limit. Correlated responses to selection for each trait can also be predicted as a result of selection on the index. See F&M for details. Indices can also be developed in which an increase in trait mean has no value, but a decrease is deleterious.

Selection index methodology has been applied primarily by animal breeders. They have generally been more interested in incorporating information from relatives rather than on incorporating information across multiple traits to improve response to selection. (You can't replicate animals in the same way as varieties of crops.) More recently animal breeders have moved to estimating breeding values using reml. Once the additive breeding values for each animal (or variety) have been predicted, these can be multiplied directly by the economic value. In crops, apart from the occasional failed sugar beet breeder, selection indices have been little used. Part of the lack of use is ignorance but part may be that heritabilities of variety means in crops are generally quite high by the time data on multiple traits are available. (That is not to say that these schemes are optimal however. Maybe selecting across more varieties but with lower replication would be better.) However, there is increasing interest in using reml to estimate variety performance, and in incorporating information from related varieties too.

The consideration of correlated responses also offers a simple way of dealing with genotype x environment interaction in variety trials. This approach is old, but to the best of my knowledge has not been adopted in any plant breeding programme. Performance at each trial site is treated as a separate trait. Response to selection in any environment can then be improved by incorporation of performance data from the other environments. (Set the economic value to 1 for the environment you are interested in and to 0 for all the others.) Often, breeders select for average performance across a range of sites and are not so interested in improving precision in any particular site. There are instances where this is not so, however. A breeder based in the UK may be interested in selling in a new market, Erewhon say. S/he cannot afford an extensive series of trials within Erewhon, a single trial on its own may not give sufficient precision, but a single trial augmented with information from the home country may be just the ticket.

The estimation of components of genetic and environmental (co)variance for analysis of data across sites in this way is simplified since the covariance across sites is all genetic – there is no environmental covariance *provided* different randomisation patterns have been used at each site. If this is not the case, interplot competitive effects and experimental edge effects will be correlated. This will inflate the covariance and make variety performance across sites appear more consistent that it really is. The effect may be slight, but this is another reason for using different randomisation plants at separate sites. Not to do so is sloppy.

The disadvantage of this approach to the analysis of GxE is that it is entirely statistical: there is no incorporation of any genetical or physiological knowledge into the analysis. An advantage is that it is simple.

A cynically practical, but less formal method of dealing with GxE was adopted by Ellerton (he of the strange and unpublished trial design) for testing sugar beet in the UK. He simply selected sites for recommended list trials on the basis of their correlation with the mean of the official test sites (run at that time by NIAB). I'm sure this isn't a unique approach, but once again Sidney was ahead of his time.

Genotype x Environment interaction – adding in some genetics and physiology.

If performance of a particular variety in a particular environment cannot be predicted from the marginal, or average effects of the variety and environment, then there is deemed to be GxE. In statistical terms. if the model

$y_{ijk} = m + g_i + s_j + e_{ijk}$

adequately describes the data, there is no GxE. If the model needs to be extended as

$y_{ijk} = m + g_i + s_j + ge_{ij} + e_{ijk}$

where the term ge represents the interaction of genotype g in environment s then we have genotype x environment interaction. The significance of ge can be assessed in an analysis of variance or other model fitting exercise. The s term can be partitioned into differences between years, regions, specific trial farms and so on. Equivalently the ge term can be partitioned further; into interactions of genetic effects with years and regions. The merit of this approach is that if components of variation are estimated (using eg reml) then they can be used to optimise selection schemes for testing over years and sites within years to maximise response to selection on average performance. Optimisation can take into account time and cost if required. This approach treats G x E as a nuisance to be coped with in selection for average performance. It is this way, at least within Europe, that most breeders operate: breeders select in the environment in which they wish to sell or distribute their varieties, generally by selecting on mean performance across a set of sites and years, and that is the end of the matter. It is hard to argue against this approach. The general problem is that although it is easy to detect GxE in the manner described above, attempts to predict variety performance in differing environments has been less successful, except in trivial and uninteresting cases: varieties of cereals bred for Spring sowing generally perform worse than varieties bred for Autumn sowing when both are sown in the Autumn, for example. Nevertheless, an extensive and growing quantitative genetic literature attempts to make these predictions. We outline some of the approaches below.

1) Include physiological measures as covariates in the analysis. Varieties may differ in their response to rainfall, for example. A regression coefficient measuring the response of each variety to rainfall at each testing location can be estimated and variety recommendations could be made for wet and dry areas independently. Of course, to estimate the regression coefficients you have to grow the varieties in wet and dry areas in the first place. Nevertheless, it is possible that selecting for predicted yield at a certain precipitation level may be more accurate than selecting directly on yield under that rainfall pattern (because the regression uses all the data). Other measures can be taken too; fertilizer levels, sunshine, temperature and so on. Moreover, the regression coefficients for each variety can themselves be treated as traits for selection – to select for responsiveness or stability. The direction of selection depends on your point of view, the trait, and the environmental factor under consideration. For example selection could be on a combination of average performance and for stability in the absence of rain, or for average performance and response to irrigation. The nature of selection depends on the breeding objectives. On the genetical front, the regression coefficients can themselves be subject to additional analysis.

2) Rather than regress variety performance on an independent variable, regress it on the average performance of all varieties at a site. There are some statistical difficulties with this approach, but they can be overcome. The approach originated with Yates, but is better known as the Finlay and Wilkinson regression. It has been popular because it removes the requirement to identify and measure the components of the environment which are responsible for the variability in variety performance. Finlay and Wilkinson analysed data that had first been log transformed, and there is often a strong case for working on log transformed data for these types of cross sites analyses: though if all you are interested in is average performance across all sites, this matters less. As a consequence of the way the regression is carried out, the expected regression coefficient in the absence of any effect is 1. Varieties which perform relatively better in high yielding sites will have regression coefficients >1 and varieties which perform more consistently across all sites will have regression coefficients <1. Again the regression coefficients can themselves be treated as traits. Quantitative genetics has been extended to incorporate GxE treated in this manner and is described in K&P.

3) Ecovalence. Variety stability is assessed as the contribution of a variety to the interaction SS. This assumes, probably more realistically that the Finlay and Wilkinson regression, that the stability of variety performance across sites is not related to average site performance. In fact (2) and (3) can be combined, partitioning the GxE variation for each variety into a linear regression component and a remainder.

4) Use of genetic covariance between sites to improve the precision of variety estimation at any particular site, as described in the earlier sections on “Multiple traits and environments” and “Index selection.”

5) AMMI Additive Main Effect, Multiplicative Regression. This carries out a singular value decomposition on the matrix of estimated g x e components with elements:

$$ge_{ij} = y_{ij} - m + g_i + s_j$$

We shall describe this in more detail.

AMMI

Remember that we can approximate a rectangular matrix as the sum of a series of terms. (By singular value decomposition or spectral decomposition – see the maths notes.) In each term, there is a vector for row effects, a vector for column effects and a scale factor (the singular value). The bigger the singular value, the more variation it accounts for. In AMMI, the matrix of ge terms is approximated by one or two singular values and their corresponding vectors. The idea is that the included singular values and vectors account for the majority of the G x E interaction, with the remainder attributable to noise. It is possible, but not guaranteed, that the singular vectors, the first giving loadings for varieties and the second for environments, may be interpretable in terms of knowledge of the environments and germplasm. They may discriminate between different origins of varieties on the one hand say, and on the other between different sets of environments in terms of rainfall, geography, whatever. This could then inform choice of parents and testing regimes for subsequent breeding. The full model is:

$$y_{ijk} = m + g_i + s_j + \sum_{n}^{N} u_n w_n v_n' + r_{ij} + e_{ijk}$$

Subscipts and terms:

i for varieties
j for sites or environments
k for replicates within sites (only one if we working on site means)
n for the columns of the spectral decomposition matrices (see maths notes)
u is the singular vector for genotypes
v is the singular vector for sites
w is the diagonal matrix of singular values
r is the matrix of residual GxE terms not accounted for by inclusion of the first N singular values.
e is the error – will be confounded with r if there is no replication of varieties within environments

Although we have described this analysis in terms of the spectral value decomposition of the GxE matrix, if we return to the relationships between eigenvalues and eigenvectors

and spectral value decomposition, it may become clearer what is going on. We'll call the matrix of variety x sites interactions **X**. Since these are interactions, formed by subtracting variety and site effects from the means, the rows sum to zero and the columns sum to zero (assuming no missing data).

X'X is then the matrix of sums of squares and sums of products between varieties at pairs of sites – it measures the similarity between sites by the agreement in variety interaction terms.

XX' is the matrix of sums of squares and sums of products between sites for pairs of varieties – it measures the similarity between varieties by the agreement in site interaction terms.

X'**X** and **XX**' have identical eigenvalues equal to **ww**' – the square of the individual singular values.

Eigenvectors of **X**'**X** are the same as the matrix of singular vectors **v** and eigenvalues of **XX**' are the same as the matrix of singular vectors **u**. Now the eigenvalues and eigenvectors of **X'X** are the raw results of a principal component analysis of sites (interaction terms) and the eigenvalues and vectors of **XX**' are the results of a PCA of varieties (interaction terms). So AMMI is doing the equivalent (for interaction) of taking the most important one or two principal components from a PCA on sites, and also the most important one or two principal components from a PCA on varieties..

There is an increasing amount of literature associated with this method of analysing G x E. An analysis of variance can be used to test the significance of these terms. An example is given in K&P. Although the estimate of the overall mean across sites for each variety is unchanged, the estimates of variety means at each site are changed and may be more accurate that the simple observed means. A comparison of this method of predicting means at a site with the selection index method discussed earlier is given by Piepho (TAG 1994 **89**:647-654): the selection index method won. However a combined method which outperforms either alone is described in Piepho (TAG 1998 **97** 195-201). Note that the AMMI approach can also be applied to the raw dataset: to a matrix of variety means at each site. This is not routinely carried out as far as I'm aware. It would treat variety and site effects as multiplicative and assume there was no independent G x E terms. This isn't as mad as it seems - after all, transforming data from raw scores to logs before analysis also has the effect of transforming multiplicative effects to additive effects, so analysing the data with a multiplicative model may not be as crazy as it at first seems. A similar approach, in which site mean effects are subtracted from the matrix and analysis is on the matrix of g+ge terms has been proposed and argued over.

AMMI is available as a procedure within GenStat.

Summary of G x E.

The analysis and interpretation of GxE interactions has had as much effort put into it as any other field (ha ha) in crop quantitative genetics. In spite of this, it seems to me that the impact of the alternative methods of analysis on breeding programmes has been slight. This is because breeders have, probably quite sensibly, stuck to selecting in their target environment, in which case problems to do with GxE melt away. This will remain the case until the analysis of GxE changes from being descriptive to predictive. However, even if this happens, at some stage, as GxE effects increase in importance relative to genetic main effects, then separate breeding programmes become necessary for each environment. Once this occurs, the problem goes away again.

MAPPING GENETIC MARKERS

Introduction

"Molecular Markers in Plant Genetics and Biotechnology" by Dominique de Vienne is an excellent introduction to the use of molecular markers in plant breeding. Science Publishers, 2003 ISBN 1578082390, 9781578082391

Why do we need to map genetic markers?

Ultimately, we want to use markers to tag loci determining traits and develop systems of marker assisted selection. We can do this marker by marker, of course, and if the marker and the trait are closely linked, then this may be sufficient: we don't need a map. However, it needs high densities of markers to work like this, stringent significance thresholds to compensate for multiple testing and therefore large population sizes. In addition it is easier to integrate trait loci into breeding programmes if we know where on the chromosome our loci are. If genetics maps are available, comparisons of QTLs can be made across populations and sometimes across species. Finally, detecting QTL using a map is more powerful than detecting them without one (unless marker density is very high) because we can use information from multiple linked markers to gain in both power to detect and in precision to locate QTL. Therefore, with few exceptions, it is best to have a map.

The mapping process can be broken down into three stages though these are not particularly independent:

Are markers linked?
What is marker order?
What are inter-marker distances?

We'll go through these in some detail.

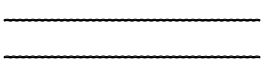
To start with, we'll assume we're mapping in a set of doubled haploid lines derived from an F1. This is the easiest case. We'll extend to other populations later.

We'll start with some definitions:

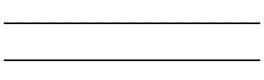
The recombination fraction, r (or sometimes θ) is the frequency of recombinant chromosomes. At meiosis:

one chromosome pair (a homologue)

→ chromosome doubling (before meiosis starts)

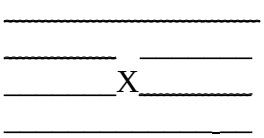


two chromatids



two chromatids

→



cross over between chromatids at a chiasma

non-recombinant gamete

recombinant gamete

recombinant gamete

non-recombinant gamete

Recombination is a result of chiasmata. A single chiasma will leave two recombinant and two non recombinant chromosomes as the product of breakage rejoining during meiosis. We can therefore define a (cytological) map distance m as the expected number of crossovers = ½ the number of chiasmata in a length of chromosome. Map distance measured in this way is linear – total map length is the sum of map length of separate sections of chromosome. A further consequence of the relationship between crossovers, recombination and chiasmata is that the recombination fraction can never be >0.5. Mather's formula relates chiasmata to recombination fraction:

$$\text{recombination fraction} = (1-p_{(\text{zero chiasmata})})/2$$

For markers which are very closely located on a chromosome, only zero or one single chiasma are likely to occur in a meiosis. For a small map distance m, $p_{(one\ chiasma)}$ will be 2m and $p_{(no\ chiasmata)}$ will be (1-2m). Substitute this into Mather's formula and we get

recombination fraction = map distance.

So for small distances, recombination and map distance are equivalent, which means recombination fractions are additive too.

This is not the case at longer distances. If we assume that in any chromosome interval chiasmata follow a Poisson distribution, then for a map distance m we require 2m chiasmata so the probability of no chiasmata is e^{-2m}. Substituting this into Mather's formula, we get:

$$r = 0.5\ (1- e^{-2m})$$

and

$$m = -0.5\ (\ln(1-2r)$$

This is the Haldane mapping function for relating recombination frequency to map distance on which we shall say more later.

Are markers linked?

The Bruce Weir book is good here.

Under random assortment (Mendel's second law, markers are on separate chromosomes) the pattern of segregation at one marker is independent of that at another. This is easily tested in a chi sq by comparing observed to expected:

Example: the F1 from the cross (AABB x aabb) is backcrossed to aabb to give:

	AB	Ab	aB	ab	total
observed	27	22	19	32	100
expected	N(1-r)/2	Nr/2	Nr/2	N(1-r)/2	
expected	25	25	25	25	

Chi-squared = 3.92 (3 df) p = 0.270

This chi-squared with 3 degrees of freedom can be partitioned into three 1 df tests, one for comparing the frequencies A:a to 1:1, one for B:b to 1:1 and one for linkage, which can be calculated by difference or by comparing (AB+ab):(Ab +aB) to 1:1. That is, we compare the frequency of non recombinants to recombinants. In this case the values are:

A:a	0.04	(p-value 0.841)
B:b	0.64	(p-value 0.424)
linkage	3.24	(p-value 0.072)

However, we are not just testing if the number of recombinants is significantly different from expected, we are testing if it is significantly *less* than expected - because the recombination fraction cannot be >0.5. If we were testing with a t-test or similar, we would use a 1 tailed test rather than a 2-tailed test. In practice we get the usual t-test probability then half it. For chi-sq, we can proceed in exactly the same manner – look up the usual p-value then halve it. Formally, the test statistic is stated to be distributed as a 50:50 mix of a distribution with a point probability mass of zero and a chi-sq distribution with 1 df. This is sometimes erroneously referred to as a one-tailed chi-sq test. This is incorrect. The usual significance tests we carry out with the chi-sq distribution are all one tailed: we test whether our statistic is greater than some threshold (eg 3.84 for 5% significance with 1 df). Testing if values are significantly lower than some chi-sq value is a test if the data fit significantly *better* than expected. (It has been used in this manner in the long running debate about whether Mendel cooked his data: do the data look too good to be true?)

If we reject the null hypothesis and decide that the markers are linked, the obvious (and ML estimate) of the recombination fractions is (number of recombinants / total). We can add confidence intervals to this if we wish using the formula for the variance of a binomial dist.

$$r(1-r) /N$$

This is simple in crosses when direct counting of recombinants and non-recombinants is possible. This is the case in backcrosses and populations of DH or SSD lines derived from an F2. For the F2 itself, this is not possible and both estimation and confidence limits are more complex.

We can also use maximum likelihood (ML) for estimation, and the likelihood ratio test (LRT) for significance testing. This is often the only option in more complex cases.

In the example, the estimate of r is 41/100 = 0.41

The LRT is -2ln(likelihood at r = 0.5 / likelihood at r = ML estimate)

These likelihoods are equal to the probabilities of the observed classes for the two values of r. The distribution is binomial. Ignoring the factorial part of the binomial distribution which cancels out in the LRT:

log likelihood at r = 0.5	=	100 ln(0.5)	= -69.315
log likelihood at r = 0.41	=	41 ln (0.41) + 59 ln(0.59)	= -67.686
LRT	=	2*(69.315 – 67.686)	= 3.258

which is very similar to the test for linkage from the chi-squared test and therefore gives a very similar p-value.

If neither marker fits Mendelian expectations very well, we can still test for linkage using a contingency table test. For our example:

	observed			expected	
	B	b		B	b
A	27	22	A	22.54	26.46
a	19	32	a	23.46	27.54

chi-sq = 3.20 which is very close to the previous value, not surprisingly in this case given the very close agreement to Mendelian expectations. This test is valid provided that the cause of the distortion at one locus is independent of that at the other. Non-independence of segregation distortion can mimic linkage and the test is no longer valid.

Although we can test for linkage in this manner, estimating recombination frequency is more complicated. We must fit a model which estimates the distortion at each locus together with the recombination frequency between them. These three parameters take up the three degrees of freedom available and give a perfect fit to the data. This is most easily done numerically (we may have a go). In this case, the estimate is 0.4102: there is virtually no distortion in this example so the estimate is little changed.

In backcrosses, segregation distortion at a single locus does not affect the test for linkage or the estimation of recombination frequency. If both loci are distorted, but the distortions are independent, then the estimation of r should be modified as above. If the distortion is not independent you are in trouble. Distorted segregation in the F2 presents more of a problem too. Most mapping programs ignore segregation distortion in their estimation of recombination fraction. The effect of distortion in DH or inbred lines is similar to that for the backcross.

Suppose we have the progeny but we don’t know the genotypes of the parents, so we don’t know which alleles concur on the same chromosome: we don’t know the phase of linkage. Then the tests for linkage can proceed exactly as before, but now we have a two tailed test rather than a one tailed test.

Assigning markers to linkage groups linked markers → linkage groups

This is simple enough in principle: merely identify pairs of markers which are judged to be linked and cluster these pairs so that all markers in a cluster are linked to each other. Some programs identify triplets of markers which are linked and then cluster these.

Mapmaker was the first freely available (and free) software for mapping markers in experimental crosses. From the mapmaker manual:

"If the LOD score is greater than some threshold, and if the distance is less than some other threshold, then the markers will be considered *linked*. By default, the LOD threshold is 3.0, and the distance threshold is 80 Haldane cM."

That is to say, the criteria for clustering markers are a combination of LOD score (or p-value) and distance apart. 80 cM corresponds to a recombination fraction of 0.4 – so we disregard loose linkages however strong the statistical significance. Sir Austin Bradford Hill would have approved.

For n markers there are n(n-1)/2 pairs. With 100 markers (not that many if evenly spaced over a whole genome), there are nearly 5000 pairs. Applying a simple Bonferroni correction, for an experiment-wide 5% significance level we should apply a pair-wise significance level of 0.05/4950 = 0.00001. Significance, or evidence for linkage is frequently reported in LOD terms. This would be a LOD score of about four (remember the test is 1 sided). As the tests are not independent (if pairs AB and BC are linked, then AC must be linked as well) this value is probably a bit high (for 100 markers). There are more sophisticated ways of deciding what the correct LOD should be. In practice a LOD of three is often used, partly for historic reasons, partly because it is about correct for human genetics. There is more on selecting significance thresholds in the section on QTL mapping.

It is possible that results from other mapping populations or a consensus map exist already. If your population has markers in common, these can help in assigning problematic markers within your own population.

Ordering markers

Consider three markers again.

AB: r = 0.2
BC: r = 0.1
AC: r = 0.25

B and C are closest, A seems closer to B and C, so the order we would select is:

CBA

What we have done is select the order which gives the lowest sum of adjacent recombination coefficients (SAR). We could also use SAL (sum of adjacent likelihoods), or LOD scores or whatever.

This is simple, but to extend the approach to more markers in the same linkage group is harder because the number of possible orders is n!/2 which for five markers is 60 and for 10 is 1,814,400. So unless the number of markers is very small it is difficult to examine all orders and select that which fits best.

Alternatives:

Seriation

Start with any pair of linked markers. Next, select the marker showing the closest linkage to one of these two and slot it into one of the three possible positions on the map following one of the strategies described above. Select another marker and slot it into one of the four possible positions. Repeat for all m markers. This requires (m-2)(m+3)/3 evaluations - 52 for ten markers, a more manageable number. However, this provides an initial order, not the final order. Starting with different pairs or selecting different markers to add may result in a different order. One can try several then pick the best using SAL, SAR or likelihood. This process, or a version of it, is called seriation. The finished order can be perturbed a bit to see if local changes improve the fit. The perturbation usually takes the form of "rippling" or "flipping" or both. As the names suggests, a pair of adjacent markers (or maybe three) have their order reversed or flipped. If the fit improves, the new order is kept. Rippling can proceed through all markers, and be repeated several times. At this stage in the process, likelihoods of the more limited numbers of alternative orders being considered can be calculated and compared.

Branch and bound

Start with an initial reasonably good order of all markers. Then challenge this with a order created by adding markers one at a time. As markers are added, the likelihood will decrease. If at any stage the likelihood of the new order becomes lower than the likelihood of the initial order, then the new order is discarded. In addition, all descendant orders are eliminated from consideration. For example, if the true order is ABCDEFGH but the order CDFA is found to have a lower likelihood, then there is no point in considering CBDFA or CDHFA etc.etc. If a complete order is found which has higher likelihood than the initial order, then it is substituted and the process is repeated. As a result, the total number of orders to be considered is greatly reduced (but can still be too big to be practical).

.

The Joinmap approach

Joinmap is the gold standard software for mapping markers but it is expensive and we are therefore not going to use it. It works on the matrix of pairwise recombination fractions as observations which it compares with a matrix of predicted values. Any of the possible orders of markers, with intermarker distances, can be used to create a matrix of predicted recombination fractions among all markers. The best order is taken to be the one which

minimises the sum of squares between observed and predicted values. Although more computationally intensive than the other methods, in a simple form this is possible in Excel for quite large numbers of markers. However, Joinmap is more sophisticated than this, in that weighted least squares rather than ordinary least squares is used. Here, the recombination fractions are weighted by the precision with which they are estimated: smaller recombination fractions are estimated more precisely than larger.

This method has the advantage of easy extension to combine maps from different populations: in fact it was this application that first led to its development.

PCA

Principal component analysis or equivalent methods can also be used to create an order for maps, working on the matrix of recombination coefficients or (additive) map distances. If the latent vectors of the two largest PCA are plotted, a horseshoe arrangement will be revealed. Joining-up the dots with give the order. Software for this approach is now available http://cbr.jic.ac.uk/threadmapper/ This is more sophisticated than the simple description given here and will cluster markers into linkage groups in addition to ordering within groups.

Simulated annealing

As a liquid cools, the molecules within it move about less and eventually crystallise into a solid. Simulated annealing mimics this process. We start off with an order of markers, and estimate a function, the likelihood say. The order is changed at random and the likelihood estimated again. If the likelihood is improved, the new order is accepted. If the new order is worse, it is accepted with some probability; otherwise the original order is kept. This process is repeated many times, but as time passes the probability of accepting a worse order is reduced; the equivalent of the liquid cooling down.

There are a lot of control parameters that can be adjusted: temperature, the extent of the perturbation of marker order, and so on. However it seems to work in practice and is encoded in the software GMendel. The method is similar in principle to "genetic algorithms" in which improved solutions to very general problems, not just genetic ones, evolve by mutation and selection. (Sometimes recombination is added too.)

You may come across this approach being used in other problems in genetics too. For example the in the program simwalk, which is used for mapping traits in large human pedigrees (but is not, as far as I'm aware used for ordering markers).

The three marker case in more detail

We shall map three markers by maximum likelihood – partly to show that we can and partly because it provides another way of looking at mapping functions.

Suppose the backcross AaBbCc x aabbcc gives observed genotypes:

			recombinant or non-recombinant		
	observed	No.	A_B	B_C	A_C
1	ABC	77	NR	NR	NR
2	ABc	3	NR	R	R
3	AbC	11	R	R	NR
4	aBC	9	R	NR	R
5	abc	70	NR	NR	NR
6	abC	5	NR	R	R
7	aBc	15	R	R	NR
8	Abc	10	R	NR	R

total 200

Chromosomes are labelled as recombinant or non-recombinant for each of the three recombination fractions. Estimate of these are

AB 0.225
BC 0.170
AC 0.135

Under the assumption that recombination in one interval is independent of that in another, then for any order, the overall likelihood is just the product of the likelihoods in the two intervals. Suppose the order is A_B_C. Then the likelihood of observing an ABC individual or an abc individual is:

p(no recombination in A_B) x p(no recombination in B_C) = (1-0.225)(1-0.170)

The log likelihood of observing 147 such individuals is therefore

147 ln[(1-0.225)(1-0.170)] = -64.86

The likelihood of observing an Abc or aBC individual is:

p(recombination in A_B) x p(no recombination in B_C), which for 19 individuals is

19ln[0.225(1-0.170)] = -31.881

and so on.

For all possible orders this gives the following likelihoods:

		order			
		A_B_C	*A_C_B*	*B_A_C*	unlinked
r in 1st interval		0.225	0.135	0.225	0.5
r in 2nd interval		0.17	0.17	0.135	0.5
ABC	77	-33.974	-25.514	-30.794	-106.745
ABc	3	-6.081	-11.323	-6.772	-4.159
AbC	11	-35.900	-21.087	-18.003	-15.249
aBC	9	-15.102	-19.699	-31.447	-12.477
abc	70	-30.886	-23.195	-27.994	-97.041
abC	5	-10.134	-18.872	-11.287	-6.931
aBc	15	-48.954	-28.755	-24.550	-20.794
Abc	10	-16.780	-21.888	-34.941	-13.863
LL		-197.810	-170.334	-185.789	-277.259
LRT		158.898	213.850	182.939	
LOD		34.504	46.437	39.725	

For completeness, I've included the likelihood under no linkage and therefore constructed a likelihood ratio test and the corresponding LOD score. (These have two degrees of freedom: two recombination fractions have to be estimated. We can also calculate expected numbers of recombinants and non-recombinants under each order. For example, if the true order is A_C_B: then the classes Acb and aCB result in equal frequency from a recombination in the first interval but not in the second. Their expected numbers are therefore 0.135*(1-0.17)*200 /2 = 11.205. The complete table of expected numbers is:

		A_B_C	*A_C_B*	*B_A_C*	unlinked
1	ABC	64.3	71.8	67.0	25
2	ABc	13.2	2.3	10.5	25
3	AbC	3.8	14.7	19.5	25
4	aBC	18.7	11.2	3.0	25
5	abc	64.3	71.8	67.0	25
6	abC	13.2	2.3	10.5	25
7	aBc	3.8	14.7	19.5	25
8	Abc	18.7	11.2	3.0	25
	χ^2_2	71.1	5.3	42.2	255.6

Whether comparing observed with expected in a χ^2_2 goodness of fit test as above (we require non-significance or low values) or using LODs (we are looking for the highest value), the best order is ACB, as we inferred from the two locus recombination fractions. Note that the comparison of LODs between different orders is not a formal significance test: all models (locus orders) have the same degrees of freedom: there is no nesting of hypotheses here. The comparison with the unlinked LOD is valid however.

The observed recombination fraction between A and B is not the sum of the two component intervals A_C and C_B. Recombination between A and C will only be observed if there is a recombination in AB but not in AC, or a recombination in BC but

not in AB. Recombination in both AB and BC - a double recombination - will result in no apparent recombination between A and C. Thus we have:

$$r_{ab} = r_{ac}(1-r_{bc}) + (1-r_{ac})r_{bc}$$

$$= r_{ac} + r_{bc} - 2r_{ac}r_{bc}$$

The relationship is not linear. It can be re-expressed as:

$$(1-2r_{ab}) = (1-2r_{ac})(1-2r_{bc})$$

which is additive on the log scale:

$$\ln(1-2r_{ab}) = \ln(1-2r_{ac}) + \ln(1-2\, r_{bc})$$

ln(1-2r) could be used as a mapping function. But when r is small:

$$\ln(1-2r) \sim -2r$$ (remember the maths revision)

Therefore, if we rescale the function by multiplying through by -0.5

$$-1/2 \ln(1-2r_{ab}) = -1/2 \ln(1-2r_{ac}) + -1/2 \ln(1-2\, r_{bc})$$

This is just the Haldane mapping function again and at small distances

$$-1/2 \ln(1-r) \sim r$$

The Haldane mapping function is a linear function of recombination fraction based on the assumption that recombination in adjacent intervals is independent. In fact it is frequently found that the relationship

$$r_{ab} = r_{ac} + r_{bc} - 2\, r_{ac}\, r_{bc}$$

does not hold. Generally, the presence of one chiasma greatly reduces the probability of another occurring close by in the same meiosis. This interference is quantified by the "coefficient of coincidence" c.

$$r_{ac} = r_{ab} + r_{bc} - 2c r_{ab}\theta_{bc}$$

This relationship will always hold: there are three parameters on the right hand side to fit the three observed recombination frequencies. LRT or chi-sq tests of the significance of c, given the marker order, are easily made.

Setting c = 1 gives the Haldane mapping function as we have seen.

Setting c = 0 at very close distances and to 1 at r close to 0.5 yields the Kosambi mapping function which is much used and generally gives a better fit than using the Haldane function. The Kosambi function, not derived here, is:

$m = ¼ \ln[(1+2r)/(1-2r)]$

$r = 1/2[(e^{4m} – 1) /(e^{4m} + 1)]$

There are many other mapping functions, which approximate reality better than Kosambi or Haldane. However, in practice, for QTL mapping, it makes little or no difference what function is used, especially at high marker densities when map distance and recombination fractions converge. More important is to get the order correct. It is also important to know what function is being used when maps are presented, since we may wish to convert from map distance to recombination fraction or vice-versa, depending on the analysis software we intend to use. Genetics odds and odds.xls provides a spreadsheet which allows inter-conversion from Haldane, Kosambi and recombination fraction.

As an alternative to working out marker order on the basis of pairs of markers, some mapping programmes will map all sets of three markers (or sometimes four) in a linkage group and use these to construct the full map. This can give a better final order yet still avoids evaluation of all possible orders.

The effect of errors

Genotype error is the bane of mapping. Molecular biologists believe all data they generate are error free. It is an uphill task to put them right with any degree of diplomacy. I once had an argument with one of the breed. He couldn't understand that most of his marker genotypes appearing 100% heterozygous among individuals in a randomly mating population indicated a problem: the selective death of homozygotes required to maintain this level of heterozygosity would amount to genocide. Fortunately he has moved on to become a physician, a profession where belief in your own absolute authority is de rigueur.

Genotype errors act to increase estimates of recombination frequency. Small error rates can have big effects. As a result of errors, total map length is increased. In fact, as genotyping quality has improved, so map length has decreased to the extent that map length estimated from markers generally agrees well with that estimated by counting chiasmata. There are two approaches to dealing with genotype error and it is possible to use both.

Firstly errors can be detected and eliminated or corrected (by re-scoring gels and by re-genotyping using a different method "a different chemistry" as the jargon has it). A well established detection method is to search for double recombinants. These should be rare, especially with closely linked markers, and are always worth checking. K&M suggest treating any double recombinants within a distance of 15cM with suspicion. As with any

data cleaning exercise, a decision to remove a double recombinant (by switching off a marker) is a judgement. Some double recombinants are expected. Multiple individuals with double recombinations involving the same interval are an indication of an error in marker order.

A more quantitative approach to detecting possible errors is to drop one marker from one individual and recalculate the likelihood. If the likelihood drops substantially, then this implies that the specific genotype is wrong. Many packages have routines that do this for you. Note that the test is not a LRT: you are dropping data not parameters so you cannot directly assign a probability in this way. For example, suppose we have three linked markers with a recombination frequency of 0.1 in both intervals and a recombination frequency between the two outermost markers of 0.18 (assuming no interference). Then from a single inbred line, the contribution to the likelihood with no recombination is 0.9^2. If the central marker is falsely scored, resulting in an apparent double recombination, the contribution is 0.1^2. If the central marker is dropped from the analysis the contribution is 0.82. The improvement in log likelihood as a result of dropping the central marker is $\ln(0.82 / 0.81) = 0.013$ when the marker is correctly called and 4.41 ($\ln(0.82 / 0.1^2)$) with the genotyping error: a considerable effect from a single data point. If we expressed this in terms of a pseudo LRT (8.8) or a pseudo LOD (1.9) the improvement looks even more impressive for a single marker (not that we can assign any p-value to these values).

It is worth keeping an eye on single marker segregation distortion and on markers with low calling rates: these are also indications of problems.

The second approach to dealing with genotype errors is not to detect and correct them but to include parameters for genotype error rates with the estimation of recombination fraction. The likelihood of the observations depends not only on whether recombination has occurred but also on whether each marker has been correctly called. Some packages can do this. The effect on map length and LRT can be studied for different assumed error rates. To estimate error rate directly, either repeat genotyping is required or pedigree structure must be exploited to detect illegal inheritance patterns (not all errors will be detected, but provided some are, this can be modelled). Crosses between inbred lines are not particularly good for this, however.

The best approach, of course, is to maintain high standards of genotyping: stains from bad data are never completely removed in the statistical wash.

Populations

Choice of a suitable population for mapping markers depends on more than mere statistics: selection of parents for subsequent QTL mapping, or breeding, the availability of dominant / codominant markers and so on can be more important.

The most commonly used populations for mapping are the F2, the backcross, or inbred lines (or DH) derived from the F1.

Backcross
An advantage of the backcross over the F2 is that it is easy to understand. A disadvantage is that we can only map markers for which the recurrent parent is fixed for the recessive allele, or for markers which are codominant. So with AFLPs and the like, we would expect only ½ the markers differing between the parents to segregate. (We could always work with backcrosses to both parents.)

F2 population
Contains more recombination than the backcross: 2x the number of meioses. The estimation of recombination fraction is not simple, but is possible through maximum likelihood.

Inbred lines and doubled haploids derived from F2 or F1.
The absence of homozygosity makes life straight forward. The expectations and analysis, after a bit of rescaling of recombination fraction for inbred lines (see below) are the same as for the backcross.

Inbred lines and DH derived from a backcross.
I'm not aware that these have been used but there is no reason why they shouldn't be. Although homozygous, markers will be segregating in a 3:1 ratio.

Full-sib families (genetically equivalent to 4 way crosses among inbred lines).
Not much used in plant genetics, where homozygous lines are often available. Used in animal breeding, and in outbred crops. Up to four alleles will be segregating. Loci may be linked in coupling or repulsion and this may be unknown. We can write down expectations for each observed class and estimate r by maximum likelihood. There are no new principles, but it is harder. It is much easier if the phase is known.

Extended / mixed pedigrees
Hardly ever used in plant breeding. There are some crops where this may be the only option eg long generation time outbreeding perennials (trees). This is also the only option in human genetics. It is difficult, but fortunately there is software to take the strain. Human geneticists initially used CEPH families (Centre d'Etudes du Polymorphisme Humain). These were large nuclear families (ie full-sibs) in which the grandparents were known. Known in the sense of being alive and available for genotyping. As a result the phase of the linkage in the grandchildren was known. These families are now available as immortalised cell lines though the software for detection and estimation of linkage has improved so that mapping in humans is no longer restricted to these more simple pedigree structures.

Map expansion

Different mapping populations can give different estimates of recombination frequency, even though the true map order is identical. For example, doubled haploid lines derived from an F1 will have undergone only a single meiosis. DH lines derived from an F2 will have undergone three meioses (one in the paternal and maternal gametes producing the F2, and then an additional one in generating the DH line). Inbred lines derived by selfing an F2 will have undergone many meioses, but as selfing proceeds there are fewer and fewer heterozygotes (and therefore double heterozygotes) so that there is less and less recombination. As a result, inbred lines and F2 derived doubled haploids will have a greater map length - larger recombination fractions for each interval - than DH lines derived from the same cross. This phenomenon is called map expansion. For the example given, if the recombination fraction, as estimated from the DH lines, is r, then the estimated recombination fraction from the inbred line is

$$R = 2r/(1+2r)$$

so

$$r = R/2(1-R)$$

If r is small, then R~2r:

there is roughly 2x as much recombination in small intervals for inbred lines. Following original work by Haldane, this has been extended and simplified for multiple markers more recently. (Teuscher & Broman "Haplotype probabilities for multiple strain recombinant inbred lines" Genetics 2007, 175:1267-1274)

None of this matters much except, when comparing maps or using maps derived from one population for QTL analysis in another, it is important to be sure where the maps came from and whether they are published in units of R or of r. Different software packages will handle this in different ways too.

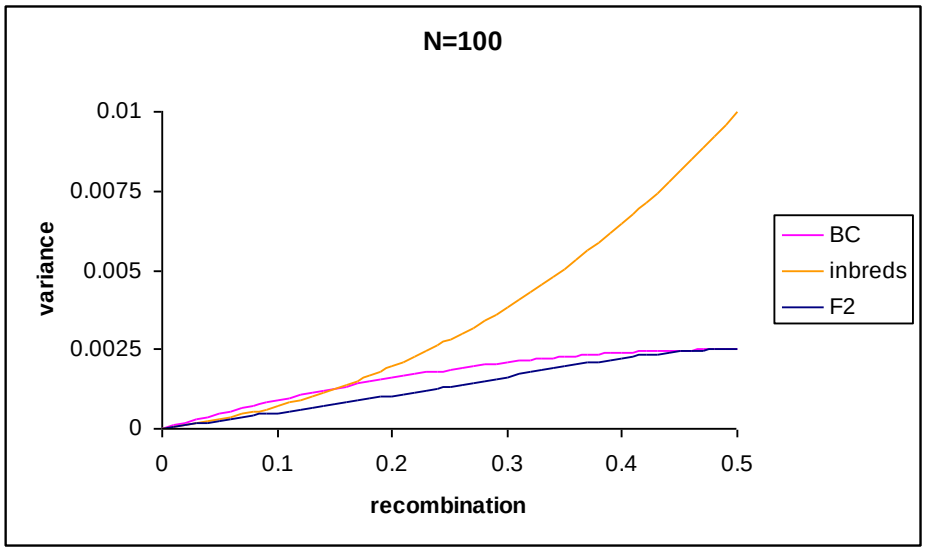
Scale and precision

Different types and sizes of populations can be compared by comparing the expected standard errors of recombination fractions. Formulae for variance of DH (= BC) inbred lines and the F2 are given below, for codominant markers

DH & BC $V_r = r(1-r)/N$

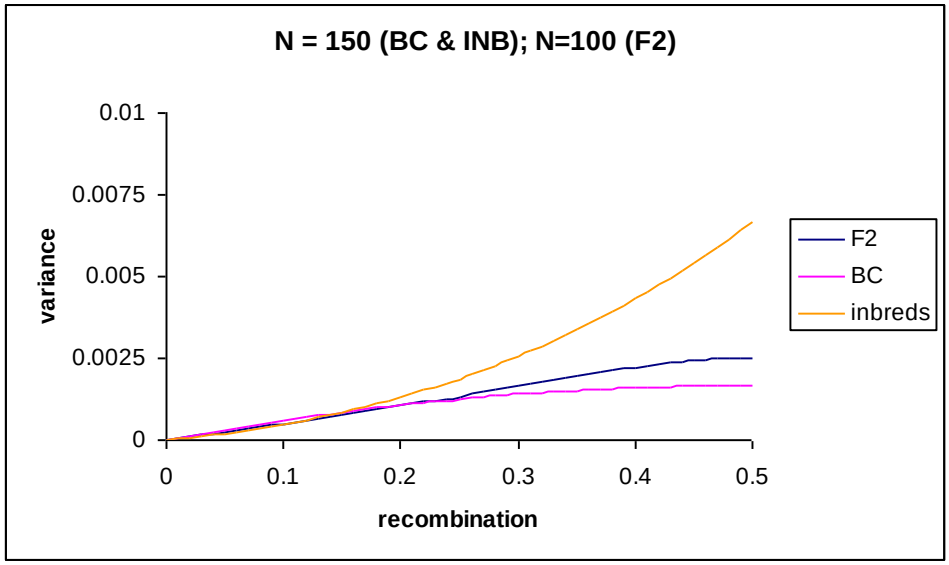
inbreds $V_r = r(1+2r)^2 / 2N$

F2 $V_r = r(1-r)(1-2r+2r^2) /2N(1-3r+3r^2)$



The F2s are best. The inbreds start well then fade away. These are all for constant population size.

If we make the BC and the inbred populations 1.5 x the size of the F2 we get this:



So over distances of <0.15 the three populations are equivalent if we raise 1.5 times more DH or inbred lines than F2 individuals.

A couple of additional points. Firstly, F2s are really cheap to produce and can be made on a grand scale. For mapping markers why don't people use F2s more often? Of course you can't maintain the plants indefinitely but so what: you can generate a lot of DNA from each plant and keep it. The answer is presumably partly that, for QTL mapping, inbred or

SSD lines will have to be genotyped anyway and partly that for dominant markers the F2 is no longer such a good population for mapping. When linked in repulsion, the variance of estimates is very high at *low* recombination frequency. Nevertheless, with high density SNP genotyping chips being developed, it seems to me that large F2 populations provide an opportunity to generate accurate fine scale maps which should not be overlooked.

Secondly, although recombinant inbred lines look bad at high recombination frequency, they are more accurate than DH lines for recombination frequencies ≤ 0.15, which is the most useful range for mapping both markers and QTL.

Finally, an interesting consequence of using inbred lines is that they make the Haldane mapping function fit the data better: because recombination can occur in adjacent intervals in different generations this reduces the amount of apparent interference when recombination is estimated in the final generation.

How many markers do we need?

To map a QTL we require about four evenly spaced markers per chromosome. However, to identify such markers we need to start with many more.

The simplest approach is to assume that we have a single circular chromosome - this gets rid of end effects - and to assume that markers are uniformly distributed over the chromosome. Suppose we want to be 95% certain to have a marker every 10cM. Suppose further the total map length is 3,000cM – a bit on the large size for a plant but smaller than wheat. There are 300 10cM intervals. We can treat the number of markers falling into an interval as following a Poisson distribution. Then with m markers, the average number in an interval will be m/300. So the probability of none in an interval is $e^{-m/300}$ from which it turns out that we would need about 900 markers to be 95% certain of having a marker in each 10cM interval. Equivalently, we could ask how many markers do we need for any genome location to be within 10cM of a marker. In this case, a pair of markers 20cM apart would cover an interval, so we would have the equivalent of 150 intervals and we could halve the number of markers. However you look at it, you need a lot.

More sophisticated approaches are possible, which account for the numbers of chromosomes and for chromosome ends. For example, Bishop et al. (1983) derived the following formula for the probability of at least one marker within a specified map distance of a major gene:

$$\mathrm{P} \geq 1 \text{ marker } = 1 - \frac{2C}{(N+1)}\left(\frac{X^{N+1}}{2L} - \frac{X^{N+1}}{L}\right) - \left(1 - \frac{CX}{L}\right)\left(1 - \frac{X}{L}\right)^{N}$$

in which C is the haploid number of chromosomes, L is the total map length, $X/2$ is the desired distance within which a marker must fall, and N is the number of segregating markers. This formula is valid only if X is less than the length of the shortest chromosome

in the genome. Assuming that, in general, at least one chiasma must occur per chromosome, then the minimum map length of a chromosome will be 50 cM, and this formula will be valid for desired intervals up to 25 cM. Since linkages in excess of 25 cM will be of little use in selecting for a linked QTL, effectively this formula covers the values of linkage of practical interest to plant breeders.

Finally

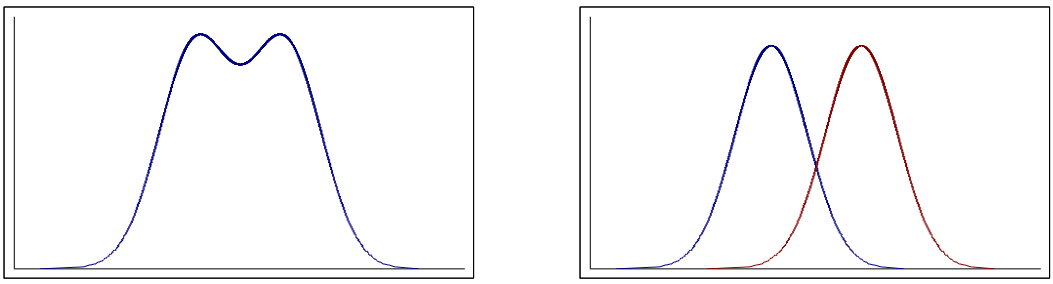
Remember, the finished map is an approximation. The markers are unlikely to be ordered correctly, and intermarker distances are estimates only.

DETECTING MAJOR GENES AND MARKER-QTL LINKAGE.

No markers, qualitative trait.

Segregation patterns can be directly tested for agreement with Mendelian expectations by a chi-squared test.

No markers, quantitative trait, possibly showing bimodality. Use of mixture models.



observed distribution → is a mixture of two underlying distributions

We need to fit a model to estimate the parameters of the two constituent distributions, and also to test if the two-distribution model fits better. Use maximum likelihood.

For a single normally distributed trait, Ø, the probability density function (pdf) of an observation, z_i is

$$\phi_{z_i} = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(z_i-\mu)^2}{2\sigma^2}}$$

where the mean (μ) and variance (σ^2) are known. If we treat the mean and variance as parameters to be estimated, then the likelihood of the observation is

$$l_{z_i} = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(z_i-\mu)^2}{2\sigma^2}}$$

The likelihood of the whole set of n observations is the product (Π) of the likelihood of each observation:

$$l_{\underline{z}} = l_{z_i} l_{z_i} \ldots l_{z_n} = \prod_{i=1}^{n} l_{z_i}$$

This equation can be solved to find the estimates of the mean and variance for which it is maximised: the maximum likelihood (ML) estimates. Of course, the likelihood can be solved algebraically: the ML estimate of the mean is just the sample mean and the ML

estimate of the variance is the sum of squares divided by the number of observations i.e. SS/n. The ML estimate of the variance is slightly biased: it should be SS/(n-1).

With a distribution composed of two underlying normal distributions, we need to estimate the proportion of individuals coming from each distribution, and the mean and variance of each distribution. We can write down the likelihood by relying on conditional probability:

Probability of observation z_i = (probability that z_i is in group 1) x (the pdf of z_i given that it is in group 1) + (probability that z_i is in group 2) x (the pdf of z_i given that it is in group 2)

In symbols:

$$l_{z_i} = p_1\phi_{1z_i} + p_2\phi_{2z_i}$$

and

$$l_{\underline{z}} = \prod_{i=1}^{n} l_{z_i}$$

p_1 = probability of belonging to group 1.
p_2 = probability of belonging to group 2.
$Ø_1$ = pdf for group 1, with mean and variance μ_1, σ_1^2
$Ø_2$ = pdf for group 2, with mean and variance μ_2, σ_2^2

p_1 and p_2 are the proportions of the whole population coming from each of populations 1 and 2. $p_1 + p_2 = 1$.

This set of likelihoods is harder to solve. It is still possible in Excel though (use Solver).

Not only does the method of maximum likelihood provide us with estimates for p_1, p_2, μ_1, μ_2, σ_1, σ_2, but we can use the maximum likelihoods themselves to provide significance tests. We compare the maximum likelihood for the full model given above with the maximum likelihoods of models with more restrictive assumptions or with fewer parameters. For example, if we set $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ then we are essentially fitting a model with just a single normal distribution. The likelihood from this restricted model can be compared with the likelihood from the full model in a likelihood ratio test (LRT). One interesting model, applicable in experimental crosses, is to compare the full model with a model in which p = 0.5, in say a backcross population or among a set of F2 derived inbred lines. Here, non-significance of the LRT compared to the full model, and significance compared to a model with only a single distribution can be taken as evidence that a major QTL is segregating in a Mendelian manner.

These models are called mixture models. The can be fitted in Excel without too much trouble (see class practical). Software is also available – eg emmix. (Expectation-maximisation mixture geddit?)

Fittings a model to an inheritance pattern is called segregation analysis. It is easy enough for F2 populations, but can also be carried out over multiple generations and families. This is sometimes done in human genetics - absence of experimental populations means there isn't an alternative. In my experience, the software (eg PAP - pedigree analysis package) is not easy to use, to say the least.

Single markers

We have a quantitative trait, as before, but instead of fitting a mixture distribution directly to the phenotype, we fit the effects of a single marker which we believe is linked to our QTL. Consider an F2. For each marker state (MM, Mm, mm) we want the probability of observing the possible QTL genotypes (QQ, Qq, qq). These can be written down, following the normal rules of segregation as:-

$P_{QQ|MM} = (1\text{-}r)^2$
$P_{Qq|MM} = 2r(1\text{-}r)$
$P_{qq|MM} = r^2$

$P_{QQ|Mm} = r(1\text{-}r)$
$P_{Qq|Mm} = (1\text{-}r)^2 + r^2$
$P_{qq|Mm} = r(1\text{-}r)$

$P_{QQ|mm} = r^2$
$P_{Qq|mm} = 2r(1\text{-}r)$
$P_{qq|mm} = (1\text{-}r)^2$

Here, $P_{QQ|MM}$ means the probability of the genotype QQ at the QTL, conditional on (|) having marker genotype MM. r is the recombination fraction between the QTL and the marker.

We can apply these probabilities to each individual in turn, given its marker genotype and its phenotype, to derive the likelihood. For example, for an MM individual:-

$$l_{z_i} = (1-r)^2 \phi_{QQ} + 2r(1-r)\phi_{Qq} + r^2 \phi_{qq}$$

where:

$Ø_{QQ}$ = pdf for the QTL genotype QQ
$Ø_{Qq}$ = pdf for the QTL genotype Qq
$Ø_{qq}$ = pdf for the QTL genotype qq

The likelihood over all individuals is again:

$$\prod_{i=1}^{n} l_{z_i}$$

and this can be solved to provide estimates of the parameters for the QTLs.

More simply, for single marker analysis, we can carry out an analysis of variance between the three marker classes with a significant result taken as evidence of a closely linked QTL. In fact, in this case:

$$\frac{\mu_{MM} - \mu_{mm}}{2} = a(1-2r) = a'$$

$$\mu_{Mm} - \frac{(\mu_{MM} + \mu_{mm})}{2} = d(1-2r)^2 = d'$$

Here a is the additive effect of the cross
d is the dominance effect of the cross.

The marker means alone cannot distinguish between a closely linked QTL of small effect and a loosely linked QTL of large effect.

Single marker analysis (sometimes called single-point analysis and sometimes called two-point analysis in human genetics) is always worth carrying out. It is conceptually easy, but in addition the loss of power from genotyping errors is not as great as in multiple marker methods.

Selective genotyping and bulked segregation analysis

To save money on genotyping, sets of lines from a mapping population can be selected such that there is little loss in power but a great saving in genotyping. The simplest case is to select the extreme phenotypes. Essentially, some proportion, the top 10% and bottom 10% say, are selected on phenotype and then genotyped. Differences in marker frequency between the two groups are then treated as evidence of linkage to a QTL. This approach is much easier to apply than to understand theoretically. The expected average allele frequency in the selected set was given by Hill ("A note on the theory of artificial selection in finite populations and application to QTL detection by bulk segregant analysis" Genet Res 1998;**72**:55-58) as:

$$P(\mathbf{n}) = \frac{M!}{(M-N)!}\left(\prod_j \frac{q_j^{n_i}}{n_j!}\right) \int_{-\infty}^{\infty} \left\{\prod_j [1 - \Phi(x - a_j)]^{n_j}\right.$$

$$\times \left[\sum_j q_j \Phi(x - a_j)\right]^{M-N} \left\{\sum_j \frac{n_j \phi(x - a_j)}{1 - \Phi(x - a_j)}\right\} dx$$

$P(\boldsymbol{n})$ is the probability $P(n_1...,n_j...,n_k)$ that there are n_j individuals of genotype j out of a total of N individuals in the pool selected from a total of M individuals in the population. Genotype j has a frequency in the population q_j and a genotypic value, measured in phenotypic standard deviation units a_j. $\Phi(y)$ denotes the distribution function and $\emptyset(y)$ the density function of the standardized normal distribution. This complex formula can be used in estimates of power calculation of selective phenotyping. For example, it has been used in studies of bulked segregation analysis and in choice of population and marker system for BSA (Mackay & Caligari *Crop Science* 2000, **40**:626-630 (2000); me again - hooray. In bulked segregation analysis, additional savings in genotyping are made by bulking the selected individuals and genotyping the bulk. Depending on the marker system, allele frequencies may be estimable from the bulks, or alternatively, we can search for markers which are fixed in the selected group. The efficiency of this process depends on the penetrance of the markers in the bulks.

The formula above can be simplified a lot if we assume large sample sizes (after selection). In this case, simple mixture models could be fitted by ML to (truncated) normal distributions. Unfortunately, in BSA, sample size is often quite small.

Multiple Marker Methods: Maximum Likelihood

To improve precision of QTL analysis, methods were extended to locate QTL within intervals flanked by pairs of markers. Writing down the likelihood in this case is more complicated, but involves nothing new. For example, consider an F2 segregating for two markers M_1 and M_2 with a QTL located between them. Assuming the parents were M_1QM_2 and m_1qm_2 homozygotes, the *gamete* types from the F1 have probabilities:-

$$p_{M_1QM_2} = (1-r_1)(1-r_2)/2$$
$$p_{M_1Qm_2} = (1-r_1)r_2/2$$
$$p_{M_1qM_2} = r_1r_2/2$$
$$p_{M_1qm_2} = r_1(1-r_2)/2$$
$$p_{m_1QM_2} = r_1(1-r_2)/2$$
$$p_{m_1Qm_2} = r_1r_2/2$$
$$p_{m_1qM_2} = (1-r_1)r_2/2$$
$$p_{m_1qm_2} = (1-r_1)(1-r_2)/2$$

r_1 is the (unknown) recombination fraction between M_1 and Q
r_2 is the (unknown) recombination fraction between Q and M_2

From these F1 gamete types, the probabilities of the F2 genotypes can be composed. For example:

$$P_{M_1M_1QQM_2M_2} = \left[(1-r_1)(1-r_2)/2\right]^2$$

These probabilities can then be used to derive the likelihood of each individual, conditional on its two locus marker genotype. Maximizing this likelihood will give more accurate estimates of effects for the QTL, and also of the recombination fraction r_1, between the QTL and M_1. Only a single recombination fraction is needed, because we rely on a genetic map of the markers having been previously supplied (or created). Given the recombination fraction between M_1 and M_2, r_2 can be calculated from r_1.

This is essentially what Mapmaker/QTL and R/QTL do. ML estimates of QTL effects are calculated at regular intervals between the two marker loci, and the results, as LOD scores, are plotted against chromosome location.

Again, we can work simply on means, but now estimates of both QTL effect and of location can be obtained. For example:-

$$\frac{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}{2} \approx a(1-2r_1r_2) \approx a$$

Since r_1r_2 will be small, even if M_1 and M_2 are quite a distance apart, the difference between the two non-recombinant homozygous marker classes can be taken as an estimate of the additive effect of the F2. After which,

$$r_1 = \frac{1}{2}\left(1 - \frac{\mu_{M_1M_2} - \mu_{m_1m_2}}{2a}\right)$$

taking means over marker over marker 1 only.

The advantage of QTL detection via maximum likelihood is that it is generally easy to write the likelihood down, even for quite complicated crossing schemes, phenotypes and marker types. Numerical methods have been developed to solve these likelihoods and are available in software such as Mapmaker/QTL and R/QTL. A disadvantage of these methods can be that they take a lot of computer time. This is less of a problem now than it was, but for repeated running of genome scans on randomised data, to derive genome wide empirical significance levels for example, it can still be limiting. Other methods of mapping are also available, however, which do not work directly with likelihoods.

Multiple Marker Methods: Kearsey & Hyne 1994

Kearsey and Hyne ("QTL analysis: a simple 'marker-regression' approach" TAG 1994,**89**:698-702) introduced a marker based regression method, probably the easiest of all, which uses all markers located on a chromosome simultaneously.

Recall that for a single marker, M_i say, with a single QTL

$$\frac{\mu_{M_iM_i} - \mu_{m_im_i}}{2} = a(1-2r_i)$$

$$\mu_{M_im_i} - \frac{(\mu_{M_iM_i} + \mu_{m_im_i})}{2} = d(1-2r_i)^2$$

If the QTL is correctly located on a chromosome with multiple markers, a regression of difference between the homozygous classes on $(1\text{-}2r_i)$ for each marker will give a straight line with slope 2a, passing through the origin.

If the QTL is located incorrectly and this regression is carried out, the goodness of fit of the line will not be so good. Therefore, for a set of markers on a chromosome, we can slide the putative position of the QTL from one end to the other and identify the position at which the error sum of squares is minimized as the most likely location of the QTL.

This approach is less simple, potentially, than it appears. Firstly a simple regression treats the markers as independent, whereas they are not, they are linked. This is corrected for by relying on empirical significance levels obtained by simulation. Secondly, the error variances around each marker will vary depending on the distance of the marker from the QTL. This can be accounted for by using a modified regression method – general least-squares regression, rather than ordinary least squares though I am unaware how much difference this makes in practice.

Also note that the regression can be carried out on

$$\mu_{Mm} - \left(\frac{\mu_{MM} + \mu_{mm}}{2}\right)$$

to detect and locate the dominance effect, d, although the test has less power.

The method can be extended to consider multiple QTL located on a chromosome by multiple regression on pairs (or more) of putative QTL locations, with a grid search of possible locations.

This regression method is encoded in the software QTL café (http://www.biosciences.bham.ac.uk/labs/kearsey/applet.html). Note it will not work simply for dominant markers in an F2.

Interval mapping by least squares regression: Haley & Knott 1992

This has nearly identical power to ML estimation, but is simpler computationally. The method works by estimating the additive and dominance effects, a and d as regression coefficients in a regression of the phenotype on a function of the marker genotypes. The trick is to find this function.

As usual,

$\mu_{QQ} = \mu + a$
$\mu_{Qq} = \mu + d$
$\mu_{qq} = \mu - a$

We have two flanking markers M_1, M_2, each with alleles M and m.

We are looking for a regression:-

$$z_i = \mu + ax_1 + dx_2 + e_i$$

Taking the mean over any flanking multiple marker genotype, for example $M_1M_1M_2M_2$:-

$$\mu_{M_1M_1M_2M_2} = \mu + ax_1 + dx_2$$

We need to calculate the mean of this marker class directly and equate terms involving QQ and qq with x_1 and terms involving Qq with x_2.

For an F2, write down the conditional probabilities of the QTL, for this marker class. (Hint – write down probabilities of gametes from the F1 first – already done for ML estimation).

$$p_{M_1M_1QQM_2M_2} = [(1-r_1)(1-r_2)/2]^2$$
$$p_{M_1M_1QqM_2M_2} = 2[(1-r_1)(1-r_2)/2][r_1r_2/2]$$
$$p_{M_1M_1qqM_2M_2} = [r_1r_2/2]^2$$

r_1 = recombination fraction between M_1 and Q
r_2 = recombination between Q and M_2
r_{12} = recombination fraction between M_1 and M_2

The addition of these three equations gives:-

$$p_{M\ M_1M_2M_2} = \left(\frac{1-r_{12}}{2}\right)^2$$

as could be written down directly by considering M_1 and M_2 only, ignoring the QTL

We can now write down the probability of each QTL class, conditional on the marker genotype $M_1M_1M_2M_2$. The conditional equations are just those above divided by $[(1-r_{12})/2]^2$

$$p_{QQ|M_1M_1M_2M_2} = (1-r_1)^2(1-r_2)^2/(1-r_{12})^2$$
$$p_{Qq|M_1M_1M_2M_2} = 2r_1r_2(1-r_1)(1-r_2)/(1-r_{12})^2$$
$$p_{qq|M_1M_1M_2M_2} = (r_1r_2)^2/(1-r_{12})^2$$

We can now write down the expected mean for the marker class:

$$\mu_{M_1M_1M_2M_2} = \mu + a\left[\frac{(1-r_1)^2(1-r_2)^2 - r_1^2r_2^2}{(1-r_{12})^2}\right] + d\left[\frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r_{12})^2}\right]$$

The terms in square brackets are the x-values for the regression equation, for the marker class $M_1M_1M_2M_2$.

Terms can be written down for the other marker classes, and also for crossing schemes other than an F1. As you can see, this is fiddly and error prone. Fortunately, it has already been done.

Armed with these sets of coefficients, all that remains is to regress the phenotype upon them. This is done for a range of values of r_1 varying the QTL location from M_1 to M_2, just as in ML interval estimation. The best placement for the QTL is the location which gives the smallest error sum of squares, or the largest value for $SS_{regression}/SS_{total}$. This latter quality can be converted to a LOD score equivalent for plotting purposes, if desired.

The method has become widely used, and can easily be extended for multiple QTL, the inclusion of covariates, and so on. It is available in virtually all software for QTL mapping.

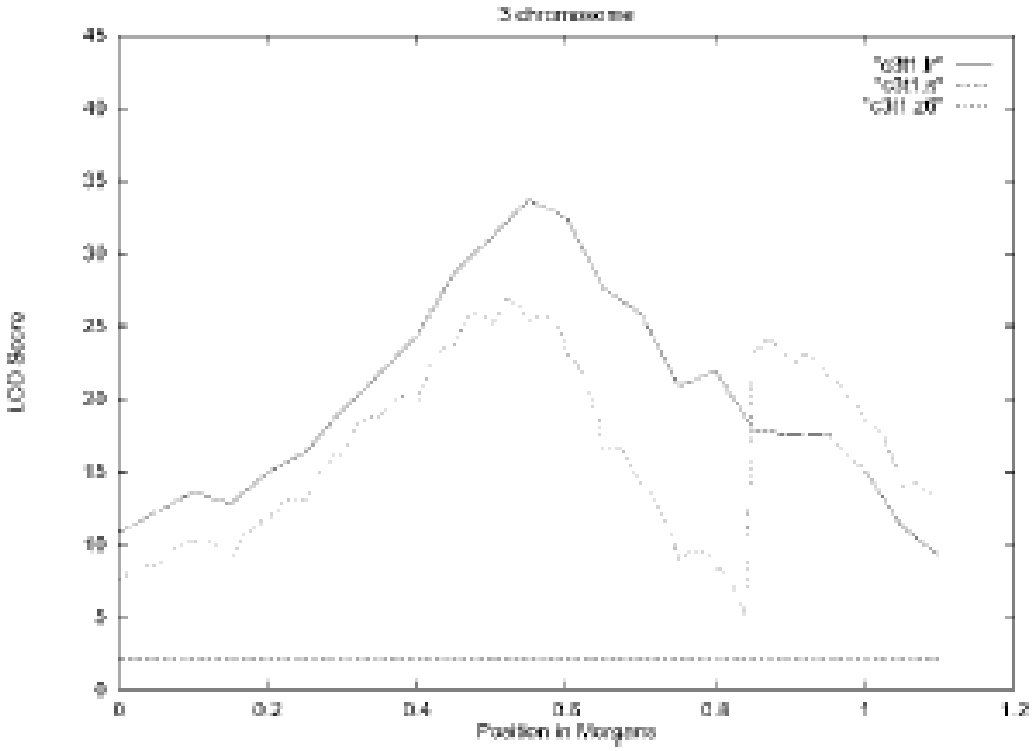
How many QTL might we detect?

Noor et al. (Genetics 2001,**159**:581–588) studied the effect of variation in gene density and recombination frequency through a simulation study in Drosophila in which 50 QTL were distributed over the genome in proportion to known gene density.

Regions of high gene density – expect more QTL
Regions of low recombination – expect more QTL

The number of genes per cM is the driver.

There should be no large QTL but:



In the same theme:

Kearsey & Farquhar Heredity 1998 **80**:137-142. Mapping studies rarely detect more than 12 QTL and most detect many fewer.
http://www.nature.com/hdy/journal/v80/n2/abs/6885001a.html

Hyne and Kearsey TAG 1995 **91**:471-476 The upper limit to the number of QTL that one can reasonably expect to detect in a mapping experiment is 12.
http://www.springerlink.com/content/v21062881u7270w5/

The harder you look the more you find? In the Illinois long term selection experiment fine mapping in a derived highly recombined population detected 50 QTL and predicted there to be 100. Laurie et al. Genetics 2004, **168**:2141-2155
http://www.genetics.org/cgi/content/abstract/168/4/2141

Estimates of number of QTL from mapping experiments are minimum estimates and there could be many more which are undetected, or for which we only detect their combined effect when linked in coupling.

Ghost QTL

This term relates to a grosser version of the effect described above. If a pair of QTL are linked in coupling, then linkage analysis may detect a single large QTL located between the two. Equally, if the QTL are linked in repulsion, both may avoid detection. Composite interval mapping (below) can reduce the occurrence of these effects.

The Beavis effect

The is the QTL version of "the winner's curse." Suppose we have 20% power to detect a QTL as significant at p = 0.05 in a population of DH lines. Suppose we have 101 genes of equal effect a=1, segregating in an F2-derived set of inbred lines for a trait with a heritability of 100%. Testing for significance of the difference in means between the two classes (assuming a normal distribution), the expected value of the test statistic is

$$E(t) = \frac{2a}{\sqrt{\frac{Vg}{50}+\frac{Vg}{50}}} = 1$$

since the expected difference between the two homozygous classes is 2a and the residual genetic variation within classes is $\Sigma a^2 = 100$

The 5% significance level, the two-tailed threshold is 1.96 (treating t as normally distributed.

The probability that one of our tests exceeds 1.96 is (in R)

```
1-pnorm(1.96,1,1) =  0.1685276
```
This is our power.

That is, we expect to find significant 17 loci out of the 101 total, and their minimum estimated effect must be 1.96 (solving E(t) = 2 for *a* in the equation above) showing there is very large bias in this case. In fact we can go a bit further:

The mean difference between classes for loci which exceed the threshold can be calculated following the methods for computing the intensity of selection as

$$i = \Phi(z)/p$$

This is correct for a standard normal distribution, $N(0,1)$, only. In our case, the variance of E(t) is 1 and the mean is also 1. We require

$$i = \Phi(1.96-1)/0.1685276 \quad = \quad 1.49$$

Adding back in the mean E(t) gives the mean of the selected group as 2.49 – a large bias. The calculation is easy in R:

```
dnorm(1.96-1)/0.1685276
```

or the whole exercise can be simulated

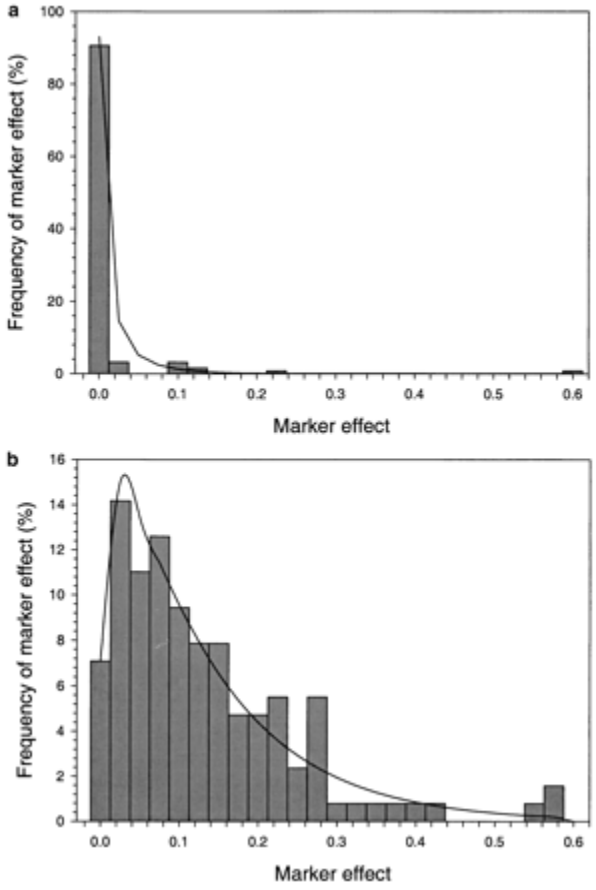
```
x<-rnorm(100000,1,0)        produce 100,000 N(1,1) numbers
y<-subset(x,abs(x)>1.96)    select those < - 1.96 and > +1.96
> length(y)/length(x)       how many are significant (= power)
[1] 0.1714
> mean(y)                   the average of the significant subset
[1] 2.454452
```

The answer is very close to that which we calculated. The advantage of simulating is that we could build on this to draw our gene effects from more realistic distributions rather than keeping them all equal as here.

This bias is named the Beavis effect after its discoverer. It describes the upward bias in the estimation of QTL effects in mapping experiments. For a given effect, as power increases, the bias is reduced, but in practice in most realistically sized mapping experiments, some QTL of minor effect will be detected and their effect will be grossly overestimated.

What is the distribution of QTL effects?

Xu (Genetics 2003 **163:**789-801) used a Bayesian method to estimate marker effects at all markers simultaneously over the whole genome (barley). The unbiased distribution of gene effects is L shaped whereas single marker regression (not the best of methods to select for comparison) show more apparent QTL of larger effect:



(a) Multiple-marker Bayesian analysis; (b) individual-marker regression analysis

The debate over the number of genes contributing to variation in the typical quantitative trait has had a long history. Even before marker based analysis, there were strong divisions into those who believed the number was large, and those who believed it was small. As the results from QTL mapping experiment first emerged, it seemed as though a small number of QTL commonly accounted for most of the genetic variation. This position has been eroded and the consensus view now is that the number of genes affecting the typical trait is large, closer to 100 than 10, and that most of these genes have small effects. Moreover, though a small number of genes do have large effects, there is more caution that formerly about immediately concluding that seemingly large QTL are the result of a single gene. Nevertheless, QTL of large effect do undoubtedly exist.

Detecting multiple QTLs: Composite Interval Mapping

So far we have only considered attempts to find QTL one at a time – we compare likelihoods for the presence of a QTL at a location with no QTL at the location. As a result of QTL segregating elsewhere in the genome, the error variance will be increased and the power to detect any specific QTL will be reduced. A simple way of increasing power is to reduce the error variance (talking in ANOVA terms) by including covariates. You can do this anyway - the covariate could be flowering time, disease incidence, whatever. But take care: if the covariates themselves are genetically correlated with your trait, you are redefining the trait you are trying to map. Equally, a covariate could be another marker linked to a known QTL. An example might be in mapping yield or quality in wheat, if the population is segregating for dwarfing genes or for the photoperiod response gene, *ppd*, you would wish to include genotypes at these loci as covariates.

If we don't know what marker covariates to include, we can carry out an initial genome scan, then include the marker closest to the peak QTL as a covariate and repeat the analysis, then add in the marker closest to the second largest QTL and so on. There would need to be stopping rules - AIC or similar -and there would need to be different significance thresholds at each scan, but this would work. Composite Interval Mapping is simply a method of doing just this.

Which markers to include? Always included the two markers flanking the interval being considered (ie we consider four markers in total). This is sufficient to absorb the effects of all the linked QTL on the chromosome except those in the intervals adjacent to the one being tested.

Number of unlinked markers? No more that $2\sqrt{n}$ where n is the number of individuals. We require some method of selection – select those which come up significant on interval mapping. or select those which show the largest single marker effects.

Multiple QTL mapping

CIM is still only searching for a single QTL at a time, with added covariates.

A two locus equivalent of interval mapping is to scan every pair of intervals simultaneously. For each putative QTL pair there are four genotype classes (assuming we're working with inbred lines). giving 1 df for the effect at each locus and one for their interaction. It is possible to maximize the likelihood for every pair of positions in every pair of intervals and report the results. This is implemented in R/QTL. Because so many positions are scanned, this is very slow especially when getting significance values by permutation.

Note that we also need permutation values for each LOD. There are three basic LODs:

full model with 3df
partitioned into
additive 2df
interaction 1df

There are also the two individual LODs for each QTL considered separately. R/QTL calculates five LODs, following a slightly different partition:

full model	3df
additive	2df
interaction	1df
(additive minus largest of the single LODs)	1df
(full minus largest of the single LODs)	2df

These last two are tests for the additional effect of a second QTL given the first (largest).

R/QTL does much of its multiple QTL analysis by imputation. Here, rather than modelling the probability of a QTL at specific positions within intervals, pseudo-marker data are first generated by imputation. In this, the probability that a specific chromosome location has originated from one or other of the parents is used to sample at random a progeny chromosome. Over the whole genome, this then gives a set of pseudo-markers which can be used in the QTL detection and location analysis, rather than including probabilities within the mapping exercise. So for example, if the probability at a particular locus is 0.8 that it is inherited from parent A, then samples of chromosomes would be drawn of which 0.8 would be A and 0.2 would be from the other parent. The initial imputation is slow, but subsequent analyses are fast. Multiple imputations are required to get significance levels, otherwise we would generate an excess of false positive results by treating the pseudo-markers markers as real. This approach, which I think is much used in Bayesian analysis, is good for multiple QTL analyses, since we end up carrying out the equivalent of single marker analyses which are faster.

The Advanced Intercross

This approach to fine mapping was introduced by Darvasi and Soller in 1995. Rather than mapping in an F2, or in inbred lines derived from an F2, one first intermates the F2 for several additional generations.

If θ is the recombination fraction in the F2, the expected frequency of a recombinant gamete in the Fn generation is:

$$\theta' = [\, 1-(1- \theta)^{n-2} \; (1-2\theta) \,] /2$$

For small values of θ

$$(1- \theta)^{n-2} \; (1-2\theta) \sim e^{-\theta(n-2)} e^{-2\theta} \sim e^{-\theta(n-2) - 2\theta} \sim e^{-\theta(n-2) - 2\theta} \sim e^{-\theta n}$$

and provided θn is small too

$\theta' \sim (1- e^{-\theta n})/ 2 \sim (1-(1- \theta n))/2 = \theta n/2$

So as the number of generations is increased, the proportion of recombinant gametes is increased. This "map expansion" gives greater precision in mapping. Mapping with lines derived from an advanced intercross proceeds exactly as for lines derived from an F2, but using the expanded map. For example, a QTL which is mapped to a 0.2 Morgan interval in the F10 has actually been mapped to a 0.05 Morgan interval on the F2 map. This approach requires more markers for mapping and more time. There is a loss of power to detect QTL - in effect we are carrying out more multiple tests. However, there is nothing to prevent one mapping as normal in the F2 or with F2 derived lines, and then fine mapping in the advanced intercross.

A simple extension to this approach has been implemented by Mott *et al* (PNAS 2000 **97**:12649-12654). In this, a population with >2 founders is established, intermated for many generations and then used for mapping. Mott *et al* were fortunate enough to access a mouse population with eight founders which had undergone 60 generations of random mating. The advantage of having multiple founders is that it increases the number of QTL segregating within the population and the number of traits for which the population is likely to be informative. Analysis methods become more complex, though single marker analysis remains straight forward. Simulations studies have demonstrated the power and precision of this method and it is being very successful in mouse: mapping multiple QTL for multiple traits to intervals of a few cM. In crops, the same approach has been called the multiparent advanced generation intercross (MAGIC) (Mackay and Powell TIPS 2007 **12**:57-63) and populations are being established in wheat, with some uptake now into other crops too.

Doubled haploid lines and single seed descent lines

These are not identical for mapping purposes. Doubled haploid lines have only a single round of meiosis between the F1 and their creation. SSD lines have many sexual cycles, but as homozygosity is approached, the opportunity for recombination declines with the frequency of double heterozygotes. As a result, the linkage map is expanded in SSD lines compared to the map in the F2 or in doubled haploids. This was first worked out by Haldane and for closely linked markers the apparent recombination fraction is doubled. So mapping in SSD lines could give twice the precision, at some loss of power, compared to doubled haploids. This will be offset, partly, by the fact that the SSD lines are never completely homozygous.

This is something to be wary of in mapping experiments. If you are creating your own marker map from the same population in which you plan to map QTL, then there is no problem. However, if you plan to use a map from a different population, or a consensus map, you need to take care that it is on the correct scale for your own use: a map

produced on doubled haploids could be half the length of the map you require for your population of SSD lines. A simple way around this it to keep the marker order the same but re-estimate the recombination fraction using your own data. Also note that QTL mapping programs have different conventions for how they read map distances. Some may apply the map directly to the data, some may expand the map for you if you declare that you are mapping SSD lines. You need to take care not to map SSD lines as if they were equivalent to DH lines or vice versa. This is aside from whether Kosambi or Haldane is treated as the default mapping unit.

Significance

Firstly a note on LODs and LRTs. You must not use these interchangeably. They are different. The likelihood ratio test, in the context of QTL analysis, is two times the natural logarithm of (likelihood there is a QTL linked to the marker or chromosome location / likelihood there isn't a linked QTL). For the sorts of sample sizes usually dealt with in linkage analysis, it is distributed as a chi-squared test. We are testing whether $\theta < 0.5$ against $\theta = 0.5$. We are not interested in testing $\theta > 0.5$. For this reason, we must half the usual probability from the chi-squared test. Usually, but not always, the test has 1df. The logarithm of odds is the $\log_{10}$ of the (likelihood there is a QTL at point x / likelihood there isn't a QTL at point x). It was introduced by Morton in 1955 to simplify the reading of results from linkage analysis experiments, particularity of humans. A LOD of 3 corresponds to an odds ratio of 1000 – the odds of their being a linked QTL is 1000 times greater that the odds that there isn't one. The mistake that is often made is to equate this to a probability of 10^{-3}. To convert a LOD to a LRT we multiply by 2ln(10) or 4.605. A LOD of 3 gives a LRT of 13.82. Treating a LRT as if it were a LOD will give a very optimistic view of the strength of evidence in favour of linkage. The probability associated with chi-squared = 13.82 (1 df) is 0.0002. Halving this, we get 0.0001. So a LOD of 3 is equivalent to a p-value of 10^{-4} and not 10^{-3}.

In fact, a LOD of three isn't a bad threshold to use for declaring statistical significance in many species. If we test only a few markers, we can use the Bonferroni correction to assign a significance level to the whole experiment. With interval mapping, as a first approximation we can use the Bonferroni correction to adjust for the number of intervals tested. Linkage between markers means that the tests among adjacent markers are not independent so the Bonferroni correction can be far too stringent. This can be taken into account. Lander and Bostein give the following formula for calculating an appropriate LRT threshold:

$$LRT = (C+2G\chi)p_{(\chi)}$$

C is the number of chromosomes.
G is the total genetic length (in Morgans not cM).
χ is the value of chi squared

The LOD threshold is then just LRT / 4.605

There is a workbook in "genetics odd and sods.xls" which does this. For a genome of 10 chromosomes, each 1 M long, this gives a genome wide LOD threshold of 3.07. That is, one genome scan in 20 will give a LOD greater than this under the null hypothesis. This corresponds to a nominal per-test significance level of 1.71×10^{-4}. In fact, the Lander and Botstein method of calculating thresholds is too stringent too. Significance levels are often determined empirically; by permuting the phenotype against the marker data and reanalysing. This works well, but can take an awful amount of computer time. An advantage of permutation testing is that it also accounts for effects of segregation distortion and missing data which are otherwise hard to deal with.

Support intervals for QTL location.

The convention is to provide the 1 LOD support interval; that is the interval of chromosome on either side of the peak LOD which is within one LOD unit of the peak. This is sometimes helpfully extended outwards to give the pair of flanking markers which encompass this interval. Although they cannot be translated exactly, this isn't too dissimilar from a 95% confidence interval for a mean: a 1 LOD support interval corresponds to a 4.61 LRT support interval. $p(\chi^2_1 = 4.61) = 0.016$ (1 tailed test), not too far of 0.025 for a conventional 95% c.i.

Sample sizes, marker numbers and power

Power depends more on sample size than on marker density. Once a mean density of 20-25cM is obtained, it is better to increase the population size than to add markers. There is little point in increasing marker density below 10 cM intervals. K&P are very clear – four or five well spaced markers per chromosome are adequate. For advanced intercrosses and mapping in more highly recombined populations, more markers will be required.

The typical mapping population consists of 100-200 SSD or DH lines. An incomplete survey of populations available with the EU (for a grant application) gave:

	no.	average	minimum	maximum	
wheat	18	153	65	241	
OSR	7	270	101	1100	(one large population)
maize	4	149	70	236	
durum wheat	9	196	100	384	

These numbers seem on the low side to me.

For a single locus test, sample size for any chosen levels of significance and power can be calculated from the properties of the normal distribution, just as for a standard test between two means (in this case the mean of the two markers classes). Following F&M, this simplifies to:

$n \geq 2 \ (z_\alpha + z_\beta)^2 / (d/\sigma)^2$

n is the number in each markers class, so the mapping population should be ≥2n

z_α is the cut-off corresponding to the (two-tailed) significance level ie 1.96 for $\alpha = 0.05$

z_β is the cut-off corresponding to the desired power of 1- β (nb-one tailed).

d is the difference between the two means

σ is the standard deviation of the mean of a marker class.

for a very modest $\alpha = 0.05$, $\beta = 0.1$ (90% power) $d/\sigma = 1$ (QTL heritability of 0.2 among DH or SSD lines this gives a desired population size is 42.

If we have a target LOD of 3, (p= 0.0001) then the population size increases to 107, and if we have a QTL heritability of 0.1, then the population size is 240. These calculations assume perfect markers for the trait so in practice these are lower estimates of sample size, though multiple marker methods of mapping will help.

A single locus heritability of 0.1 is still a large effect. Inbred lines are cheap to produce. I do not understand why, with a few exceptions, mapping populations are as small as they are. There is no requirement to genotype or phenotype all lines immediately but at some stage it is guaranteed that you will need more; you become interested in epistasis, the detection of which requires large sample sizes or you want to map for something like drought in a subset of lines with very similar maturity (this is still epistasis). Or you just want more precision. Whatever. You need a larger mapping population. Knowledge is power (Francis Bacon).

More accurate power calculations can, as usual, be determined by simulation. Simulations have the advantage that they can use real genotypes, complete with patterns of missing data for your specific mapping population.

Combining data across populations

The LOD score was originally introduced into human genetics by Newton Morton in 1955. The idea was that it is easy to understand, but also that it is additive across pedigrees. Over time, LODs on separate pedigrees could be added to give better support for linkage. So combining data across mapping experiments should be straight forward, yet is little used.

Firstly, I'll make explicit what is meant. We'll assume we have two inbred mapping populations with the same markers. We fit a QTL with an identical effect, *a*, at the identical chromosome position, *p*, in both populations. Then the LOD with 1 df (or equivalently the LRT) for *a* at *p* over both populations is just the sum of the LODs for *a*

and *p* in the two populations analysed separately. This is only true if *a* and *p* are kept constant in both populaions. If *a* is maximized independently in each population (still holding *p* the same) then the sum of LODs is now a two 2df test (we have fitted two parameters). This is acceptable, but the increase in df will lose power. The difference between the two LODs is a 1 df test for the difference in QTL effect between the two populations.

The frustrating thing about this simple approach is that not all our populations are likely to be segregating for the same set of QTLs, so by adding in populations we may just be adding noise. However, one cannot select populations where *a* seems large for a joint analysis– we would get very strongly biased effects and spurious significance levels. More complex models have therefore been developed in which the probability that a population is segregating is included in the model, or where *a* is treated as a random effect which varies from population to population. There are also methods and software such as *MetaQTL* (Veyrieras *at* al MBC Bioinformatics 2007 8:49) which do not require reanalysis of the raw data; they can work on published maps and estimates of location and QTLs. This could allow meta-analysis of several decades of QTL mapping experiments. Is there a risk here of publication bias? Probably not too great because most QTL mapping experiments seem to find something somewhere and therefore get published, usually quoting the negative results as well as the positive.

Beyond inbreds: full sibs, half -sibs, complex pedigrees

I shall say little about these because they are not much used in plant breeding. In animal genetics large half-sib families are common. In human genetics mapping has been mainly in multiple small full-sib families, especially of sib-pairs for which special methods exist. However there are some species where the generation of mapping populations is difficult or mapping must be attempted in near-natural populations. Coconut is my favourite.

For mapping within full-sib families and half-sib families, single marker analysis based on a t test remains straight forward and is described in more detail below.

Half-sibs

Mapping with half-sibs compares the difference in phenotype between alternative alleles from the common parent (typically the male in animal genetics and typically the female in plant genetics). We are mapping using those loci for which the common parent is heterozygous. We may have multiple alleles, for an SSR for example. The expected difference in means for a pair of markers is:

$(1-2\theta)(a_1-a_2)$ where a_i is the additive effect ith QTL allele.

Different half-sib families can have different linkage phases. Working on the square of the difference between the two alleles avoids this problem and has the expectation:

$(1-2\theta)^2\sigma_A^2 + V_{error}$

which no longer depends on phase.

If we carry out a t test for each of n families, square them and add them up, this is approximately a chi sq test with n df provided samples sizes are reasonably large. More effectively we can carry out a hierarchical ANOVA.

The problem with this approach is that rare QTLs, even of large effect, will be missed because most common parents will not be heterozygous. Though perhaps one could select on parental phenotype first.

Full-sib families

Again, we can compare differences between genotype classes using an ANOVA. There may be up to four alleles segregating. Expectations are more complicated. To combine over multiple families, we can carry out a hierarchical analysis of variance.

Mapping in phase known full-sib families (equivalent to a four-way cross) is easier. R/QTL will analyse these. R/QTL cannot currently analyse phase unknown full-sib families. For small family sizes, *QTL express* http://qtl.cap.ed.ac.uk/ will work. For complex pedigrees, *Merlin* or *Simwalk* are good starting points. If anyone needs more information about these approaches, best to contact me directly.

Power to detect QTL in these more complex population structures is weak compared to mapping in a population of SSD or DH lines.

Mapping traits with a non normal distribution

In order of increasing complexity:

1) Single marker analysis is always possible in general statistics packages.
2) Just analyse as usual and hope for the best.
3) Transform the data to try to make it normal.
4) Analyse as usual and use permutation to get significance.
   This will be OK for type I error but may lose power.
5) If you know the distribution, work directly on maximizing the likelihood.

Option 5 may be available in some mapping packages already. It is promised (long term) in R/QTL.

Genetical Genomics

I wanted to mention this somewhere. This term was introduced by Jansen and Nap (TIG 2001 17:388-391) to describe the marrying of genomics and proteomics with genetic analysis. They had in mind in particular the increasing availability of large quantities of gene expression data. By taking the expression profile of each individual in a mapping population, expression levels at each gene can be treated as a phenotype and QTL located. Sometimes the QTL are found to map to the same location as the gene, implying, but not absolutely demonstrating, variation in cis acting regulatory factors. QTL mapping to different chromosomes can only be a result trans acting regulators. One can compare expression patterns at different genes with each other and with their map position and infer biochemical pathways. Correlations with phenotypes of more direct relevance to breeders can also be incorporated. These mapping experiment have big multiple testing problems so stringent significance levels need to be set. Example analyses can be found in yeast, Arabidopsis, barley, and maize and I am sure there are others. The approach has been extended too, to use association mapping rather than linkage mapping, permitting application to humans and other species in which linkage analysis is difficult.

METHODS FOR LINKAGE DISEQUILIBRIUM MAPPING IN CROPS.

This section has been published as:

Mackay, I.J., Powell, W. (2007) The Significance and Relevance of Linkage Disequilibrium and Association Mapping in Crops. *Trends Plant Sci*. 12: 53-53.

See also
Cavanagh C, Morell,M, Mackay I, Powell W (2008) From mutations to MAGIC; resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology* 11:215-221

Abstract

Linkage disequilibrium (LD) mapping detects and locates quantitative trait loci by the strength of the correlation between a trait and marker. It offers greater precision in QTL location than family based linkage analysis and will therefore lead to more efficient marker assisted selection, facilitate gene discovery and help meet the challenge of connecting sequence diversity with heritable phenotypic differences. Unlike family based linkage analysis, LD mapping does not require family or pedigree information and can be applied to a range of experimental and non-experimental populations. However, care must be taken during analysis to control for the increased rate of false positive results arising from population structure and variety interrelationships. We discuss the suitability for crops of the alternative methods of LD mapping which have recently been developed. We advocate the development of multiparent advanced generation intercrosses in crop plants to facilitate future fine mapping projects. In addition, fine mapping of existing linkage regions and candidate genes should be attempted in existing sets of phenotyped cultivars, using statistical methods to control the rate of false positives.

Linkage disequilibrium mapping: methods developed for human genetics find applications in crops.

Linkage disequilibrium (LD) mapping, also known as association mapping or association analysis, detects and locates quantitative trait loci (QTL) based on the strength of the correlation between mapped genetic markers and traits. It relies on the decay of LD, initially present in a population, at a rate determined by the genetic distance between loci and the number of generations since it arose (Box 1). Over a series of generations, in an unstructured population (a randomly mating population with no complicating factors such as population subdivision and immigration), only correlations between QTL and markers closely linked to the QTL will remain, facilitating fine mapping. However, most populations have some degree of structure or subdivision and the simple relationship between strength of correlation and meiotic distance does not apply: correlations between unlinked loci often occur. Recently, methods of LD mapping which adjust marker-trait associations for these spurious associations have been

introduced. Originally developed for human genetics [1, 2], these methods and their derivates are now being applied to crops; driven by the development of cheaper, higher density molecular markers. Successful use will lead to more efficient marker assisted selection, facilitate gene discovery and help meet the challenge of connecting sequence diversity with heritable phenotypic differences.

In this review, we first describe the relationship between family based linkage analysis and LD mapping. We then outline the methods currently available before discussing the opportunities and challenges of LD mapping in crops. A review of practical results in crops can be found be found in [3].

Family based linkage mapping and LD mapping compared

Family based linkage (FBL) mapping can be regarded as a special case of LD mapping in which LD is generated by establishing a population from a very small number of founders in the very recent past. An F2 population, for example, is derived from a single F1 plant. The meiotic process and an appropriate experimental design ensure that the strength of the correlation between a marker and trait is proportional to the genetic distance of the marker from the QTL, with the correlation between unlinked loci being zero. Precision of QTL location depends on the detection of differences in recombination fraction (θ) between QTL and adjacent markers. In an F2 with markers located 0, 1 and 10 cM away from a QTL, the proportion of non-recombinant chromosomes is roughly 1, 0.99 and 0.9 respectively. Detecting a difference in signal strength between these markers requires a large experimental population. If the F2 was randomly mated for 100 generations, then the non-recombinant chromosomes are at frequencies of 1, 0.68 and 0.5 respectively [4] and QTL could be located more precisely. In natural populations of crop plants, or among collections of cultivars, there have often been many rounds of historical recombination. LD mapping exploits this historical recombination and provides opportunities for fine mapping that are difficult to achieve through family based linkage analysis. However, for QTL detection, rather than location, FBL mapping will generally be more powerful. In this case, the lack of recombination between a QTL and linked marker increases power of detection. For these reasons, it is unlikely that LD mapping will supersede FBL mapping: the two approaches are complimentary.

FBL and LD mapping also differ in their dependency on allele frequency in the population being mapped. In populations of plants derived from an F2, QTL are either not segregating, or are segregating at a frequency of 0.5 (ignoring selection and drift). Careful choice of parents, for example selecting phenotypic extremes, is therefore required to ensure that the population is segregating for most QTL for the trait of interest. LD mapping generally samples lines from a pre-existing population with multiple founders. The greater range of genetic material in such a population makes it more likely that multiple QTL will be segregating for multiple traits. However, allele frequencies at QTL and markers will also vary. Power of detection of QTL depends strongly on QTL allele frequency; rare alleles have low power of detection. Detection is also more likely if QTL and marker allele frequencies match. In LD studies therefore, it is wise to ensure that the full range of marker allele frequencies is covered. Moreover, if prior knowledge suggests that QTL allele frequencies are rare (for example a rare trait may show

Mendelian inheritance) then LD mapping is unlikely to be successful and FBL mapping is preferred.

For LD mapping to be possible, LD must be present in the population under study. Causes of LD are outlined in box 2.

Methods for LD mapping

1. The Multiparent Advanced Generation Intercross

In the Advanced Intercross [5], F2 individuals are intermated for several generations before mapping. The successive rounds of recombination cause decay of LD and the precision of QTL location to increase. This approach has now been extended to include populations with multiple parents, to take into account information from multiple linked markers [6, 7], and to prioritise candidate polymorphisms [8]. Its resolution and power are reviewed in [9]. The multiparent advanced generation intercross (MAGIC) was first proposed and applied to mice [6] where it is described as "heterogenous stock." Recent successes are described in [10]. In both crops and animals, an advantage of the method is that a population can be established containing lines which capture the majority of the variation available in the gene pool. Although it may take several years before these populations are suitable for fine mapping, they are cheap to set up and their value as mapping resources increase each generation. In plants, MAGIC can be used to combine coarse mapping with low marker densities on lines derived from an early generation, with fine mapping using lines derived from a more advanced generation of crossing and a higher marker density. If such populations were established now, they would be well placed to exploit the advances in genomics technology and reduction in genotyping and sequencing costs predicted to occur in the next few years [11-13].

2. The Transmission Disequilibrium Test and derivatives.

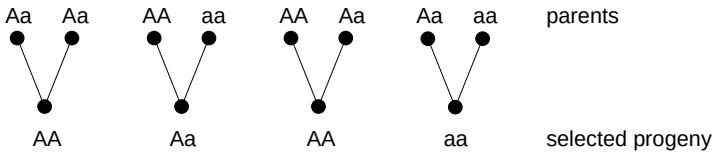
The ability to map QTL in collections of breeders' lines, old landraces or samples from natural populations has great potential. In these populations LD often decays more rapidly than in controlled crosses. Also, phenotypic data often already exist, saving time and money. The challenge is to distinguish QTL–marker associations arising from LD between closely linked markers from spurious background associations. The first and most robust method of achieving this was the transmission disequilibrium test (TDT) introduced by Spielman et al. in 1993 [14].

The TDT provides a way of detecting linkage in the presence of disequilibrium [14]. Neither linkage alone nor disequilibrium alone (i.e. between unlinked markers) will generate a positive result so the TDT is an extremely robust way of controlling for false positives. At its simplest, multiple families consisting of two parents and a single progeny are collected, as shown in Figure 1.

Figure 1. The transmission disequilibrium test. In the simplest case, progeny are selected for an extreme phenotype and transmissions to the progeny from heterozygous parents counted. In the case show, there are four heterozygous parents from which allele 'A' is transmitted three times and allele 'a' once. This frequency is compared to

the 1:1 ratio expected in the absence of linkage disequilibrium between the marker and linked QTL.

Figure 1



The single progeny in each family is usually selected for an extreme phenotype. In human genetics this typically means they are affected by the disease under study. Parents and progeny are genotyped, but only parents heterozygous at the marker locus are included in the analysis. From each parent, one allele must be transmitted to the progeny and one is not transmitted. Over all families, a count is made of the number of transmissions and non-transmissions. In the absence of linkage between QTL and marker, the expected ratio of transmission to non-transmission is 1:1. In the presence of linkage it is distorted to an extent which depends on the strength of LD between the marker and QTL. The distortion is tested in a chi-squared test. Power depends on the strength of LD and on the effectiveness of selection of extreme progeny in driving segregation away from expectation.

This elegant test is extremely robust to the effects of population structure, but is very susceptible to an increase in false positive results generated by genotype error and biased allele calling [15]. This risk can be reduced by modelling genotype errors and missing data in the analysis [16-18], or by comparing the transmission ratio for extreme phenotypes to that for control individuals or for the opposite extreme. The TDT has been extended to study haplotype transmissions, quantitative traits, the use of sib pairs rather than parents and progeny, and information from extended pedigrees. A review of the TDT and other family based association tests is given in [19].

In crops, parental and progeny lines are usually separated by several generations of gametogenesis rather than by one. In this case the TDT is still valid, but might no longer be so robust: the process of breeding might itself distort segregation patterns. A family-based association test based applicable to plant breeding programs has recently been proposed [20]. The authors point out that for candidate gene studies, this method is more cost effective than the alternative methods described below since no additional control markers are required. Some power will be lost, however, since only progeny derived from F1s known to have a heterozygous marker genotype are informative.

3. Genomic control

Population structure arising from recent migration and population admixture will generate LD between a trait and markers distributed over the whole genome. This can be detected by studying whether the distribution of the test statistic for association, estimated empirically from a set of genome-wide distributed markers, differs from the expected null distribution. This is the basis of genomic control (GC) [21, 22]. To estimate the empirical distribution accurately would require many markers. However, all that is required is to estimate the mean test statistic and compare with its expected value (1.0 for a 1 degree of freedom chi-squared test) for which only approximately 50 markers are required [23]. If the average chi-squared at a set of 50 control markers is much greater than one, population structure is indicated.

For any candidate marker, the null-hypothesis is now no longer absence of association between it and the trait. Rather, it is that there is no association above the background level resulting from population structure. To test for this, we simply divide the observed chi-squared between the candidate and trait by the average chi-squared at the control markers and look up the p-value of the adjusted chi-squared in the usual manner.

$$\chi^2_{genomiccontrol} = \frac{\chi^2_{observed}}{\sum \chi^2_{nullmarkers} / n}$$

GC is valid for any single degree of freedom test. Preferably, the control markers should loosely match the test marker in allele frequency, but this is not critical [22].

For quantitative traits, the difference between trait means for each marker class is usually tested in a t-test. Provided the number of observations is reasonably large, t^2 is distributed as a 1 degree of freedom chi-squared and GC can still be carried out. More recent work has suggested that greater accuracy is achieved by treating the test statistic as an F test with one degree of freedom (df) in the numerator and degrees of freedom in the denominator equal to the number of control loci [24].

More sophisticated versions of GC are available. With large numbers of candidate polymorphisms to test, the majority are not expected to be genuinely associated with the trait. In this case, procedures and software are available in which, in effect, the candidate markers act as their own controls. GC has also been extended to control for bias in accuracy of genotyping between DNA samples from different origins [25] and to tests with greater than one df [26].

GC also corrects for unknown kinship among collections of lines [21]. The presence of related lines can greatly increase the frequency of false positives. For many crop datasets this will be the greatest source of bias.

The correction of the false positive rate using GC comes at a cost: power is always decreased. This loss of power can be great in cases of extreme population subdivision [27]. Also, since loci can vary in their differentiation between populations, the uniform adjustment of GC might be insufficient for some candidate polymorphisms and overcorrect at others [28].

4 Structured association

Structured association (SA) provides a sophisticated approach to detecting and controlling population structure [29-31]. Again, additional markers are required, randomly distributed across the genome. Just as for GC, recent migration and population admixture are assumed to generate LD among unlinked and loosely linked markers which has yet to decay fully. However, we expect the parental populations themselves to be in linkage equilibrium. By trial and error one could allocate the individuals in our sample to parental populations such that disequilibrium within populations was minimised. One could then include information on population membership in the test of association. This is the approach taken for SA. First individuals are allocated to populations, then this information is used to control for population membership in test of association [29-31].

To allocate individuals to populations we need to know in advance how many populations there are. If unknown, this can be estimated: the allocation process is repeated for different possible numbers and the best fitting selected. Nevertheless, deciding on population number can be problematic.

The computer program STRUCTURE [29] uses computationally intensive methods to partition individuals into populations. Many individuals or lines will not belong uniquely to one, but will be the descendents of crosses between two or more ancestral populations. STRUCTURE also estimates the proportion of ancestry attributable to each population.

Following allocation of individuals to populations, the test for association is carried out in a model fitting exercise. Here, the principle is that variation attributable to population membership is accounted for first, using estimates of population membership from STRUCTURE, and then the presence of any residual association between the marker and phenotype is tested. For example, to test for association between a quantitative trait and a microsatellite, the trait is first regressed on the estimated coefficients of population membership and then on the marker – coded as a factor as if in an analysis of variance [32].

SA is effective in detecting and adjusting for the presence of population structure, but does not deal with consanguinity within populations. Recently, the Buckler group introduced a method in which population membership is estimated using STRUCTURE and kinship among varieties is estimated empirically from a second set of control markers [33]. The analysis takes into account both population structure and the correlation between individuals which results from their relationships. This method is implemented in the software TASSEL [29, 34]

5. Logistic regression

Recent simulations suggest that multiple stepwise logistic regression may be robust to the effect of population structure in its own right [27]. Here disease status (affected or unaffected) was used as the outcome variable in a logistic regression on multiple null and candidate markers. Stepwise multiple logistic regression gave false positive rates close to the desired significance level with little loss of power. The authors propose logistic regression with null markers as covariates as a less conservative (fewer false negatives) method than GC, but with a lower requirement for additional markers than SA. To date, the method has not been tested on crops and has not been adapted for quantitative traits.

Multiple regression with stepwise selection has been applied to barley however, to consider the joint effect of multiple marker-trait associations [35].

6. Principal component analysis

Recently a method termed EIGENSTRAT has been proposed, based on principal component analysis (PCA) across a large number of bialleleic control markers with a genome wide distribution [28]. The PCA summarises the variation observed across all markers into a smaller number of underlying component variables. These can be interpreted as relating to separate, unobserved, sub-populations from which the individuals in the dataset (or their ancestors) originated. The loadings of each individual on each principal component describe the population membership or the ancestry of each individual. These estimates are not ancestral proportions however (values can be negative) in the same way that estimates of ancestry from STRUCTURE are. The loadings are used to adjust individual candidate marker genotypes (coded numerically) and phenotypes for their ancestry. The adjusted values are independent of estimated ancestry so a statistically significant correlation between an adjusted candidate marker and adjusted phenotype is therefore evidence of close linkage of a trait locus to the marker.

The approach in EIGENSTRAT is similar to that of SA, but is less dependent on assessing the number of ancestral populations. Although each principal component is attributed to a separate population, the analysis is robust to the number included in the analysis, provided this is sufficiently large to capture all true population effects.

EIGENSTRAT was developed for application to human datasets with high density genotyping and low levels of population differentiation. Many crops have much higher levels of population differentiation and often only low densities of markers are available. In addition, EIGENSTRAT does not cope with close kinships. The authors suggest identifying these by other means and then selecting the largest subset of unrelated individuals. However, they also suggest combining EIGENSTRAT with GC to control for residual confounding. It is possible that such use of GC would also account well for kinship. EIGENSTRAT, unlike SA, will not readily handle multiallelic markers. However, a microsatellite with 10 alleles could be coded as 10 biallelic loci, all in complete LD. An analysis of human data showed EIGENSTRAT was little affected by LD among over three million SNPs. It is possible therefore, but remains to be demonstrated, that EIGENSTRAT will be applicable to more modest numbers of microsatellite genotypes, suitably coded. The method therefore shows great promise, but additional research is required to establish its suitability for crops.

Haplotype analysis

LD mapping can be extended to consider multiple markers simultaneously. For closely linked markers, haplotype analysis can offer advantages over single marker-by-marker analysis [36]. There are many possible approaches and methods and research in this area is continuing. Within the scope of this review, it is not possible to discuss these. The simplest approaches are:

1) Test each haplotype in turn against a pool of all others. This converts a system of n haplotypes to one of n biallelic loci. Analysis is then straightforward, but adjustment for multiple testing is required.

2) Ignore haplotypes, but analyse the constituent markers and their interactions jointly. A significant interaction is evidence of a haplotype effect over and above any effect attributable to the single markers.

Recommendations and conclusions

The substantial quantities of phenotype data already in existence from the variety trials of breeders and variety testing organisations are valuable resources for LD mapping. For example, a genome wide survey of associations with yield and yield stability components has been carried out in barley [35] using historic data. To generate novel phenotypic data for mapping traits such as stability of yield would usually be prohibitively expensive. Moreover, QTL are detected in germplasm of direct relevance to the crop. Unfortunately, all methods currently available for controlling population structure in such collections have weaknesses. For ease of application and low marker requirement we favour GC, even though it can be conservative: in the long run, false negative results are less damaging than false positives. With higher marker densities, the more intensive methods of SA and EIGENSTRAT should have greater power. However, even here GC can have a role: to confirm that these more sophisticated approaches have worked.

The resolving power of LD mapping depends on how rapidly LD decays with genetic distance. This varies between populations of landraces, wild progenitors and modern cultivars as a result of the diverse history to which crop plants have been subjected since their domestication [37]. In some populations, LD will decay so rapidly that they are best suited for fine mapping, whereas in others the decay might be so slow that whole genome scans are practical. In crops where collections of contemporary, historical, and wild material exist, selection of different sets of lines may permit both fine and coarse mapping [37]. However, in most crops, marker density is currently too low for genome scans. Before attempting these, power calculations should demonstrate that, given the rate of decay of LD in the population to be studied, the density of markers and their allele frequency distribution are adequate to detect linked QTL accounting for specified proportions of the phenotypic variation. Population size is also important. An LD mapping experiment will almost always have lower power than a FBL mapping experiment of equivalent size: if 100 lines are just sufficient for a FBL study, they will be too few for LD mapping.

For these reasons we believe that the best use of LD mapping is to refine the location of QTL identified in FBL and candidate gene studies. Longer term, prospects for high throughput genotyping and resequencing may make whole genome scans by LD mapping more feasible. The challenge is to identify and create the appropriate populations so that computational, analytical and profiling advances can be rapidly harnessed by the crop science community. For this purpose, the MAGIC is ideal: highly diverse, no population structure, and suitable for both fine and coarse mapping. We believe MAGIC populations should be established now in all crops.

Box 1 Linkage disequilibrium.

1. Principles of detecting and quantifying linkage disequilbrium

Linkage disequilibrium is the non-random association of alleles at separate loci located on the same chromosome. If one locus has alleles A and a with frequencies p_A and 1- p_A, and a second has alleles B and b with frequencies p_B and 1- p_B, then at equilibrium, even though the loci are linked, the expected haplotype frequencies are the product of the constituent allele frequencies. To take the AB haplotype for example:

$$p_{AB} = p_A \cdot p_B$$

We define any departure from this state of linkage equilibrium as:

$$D = p_{AB} - p_A \cdot p_B$$

At equilibrium, D = 0.

D is the coefficient of linkage disequilibrium. It can be difficult to interpet: its range depends on allele frequency and it is not symmetrical about zero. It is therefore usually rescaled to give it a range from 0 to 1.

2. The decay of linkage disequilibrium with time

Recombination causes gamete and haplotype frequencies to change towards their equilibrium values. Following random mating, in the absence of mutation, selection and chance effects, the value of the coefficient of linkage disequilibrium, D, in successive generations is:

$$D_{t+1} = D_t(1-\theta)$$

and therefore

$$D_t = D_0(1-\theta)^t$$

θ is the recombination fraction between the two loci.
t is the number of generations of random mating since the start.
D is the coefficient of linkage disequilibrium.

LD decays quicker at higher recombination frequencies. For unlinked loci, the decay is at a rate of ½ per generation.

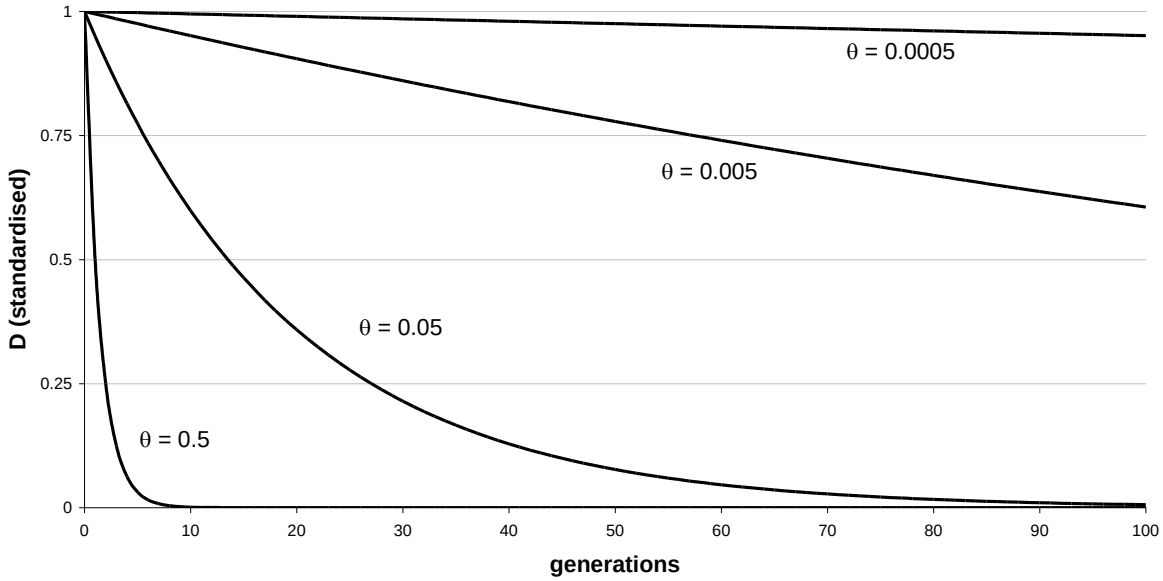
For close linkage and larger values of t:

$$D_t \sim D_0 e^{-\theta t}$$

Thus recombination frequency and time are interchangeable – a halving of recombination fraction is compensated for by doubling the number of generations. Figure 2 shows the decay in LD over time at a series of recombination fractions.

Figure 2. Decay of linkage disequilibrium with time for four different recombination fractions (θ). For unlinked loci, θ = 0.5 and LD decays rapidly within a small number of generation. For very closely linked loci, the decay in LD is extremely slow.

Figure 2.



LD decays very rapidly in the absence of linkage but persists for a very long time with very tight linkage.

Box 2. Causes of linkage disequilibrium

Mutation

Immediately after a mutation occurs, it is in LD with all other loci: the new mutation only occurs on a single haplotype. In successive generations, recombination causes LD to decay as new haplotypes are created, but this process takes a long time for closely linked markers. Most polymorphisms we observe are old: many generations are required for allele frequencies to rise to a frequency at which we detect them. Therefore, most pairs of polymorphic loci show little LD originating from mutation unless closely linked.

Population bottlenecks, founder effects and drift.

A population bottleneck is an extreme reduction in population size. It causes loss of variation and increased LD. A founder effect is a special case, occurring when a species colonizes a new environment. The number of founders can be extremely small - only a few seeds may be introduced to establish the crop. Most crop plants underwent at

least one bottleneck during domestication. The activities of plant breeders themselves can result in bottlenecks - the introduction of a new disease resistance or agronomic trait may result in a period of breeding in which a small number of parental lines are used extensively. In fact, any finite population size generates some degree of LD, just as genetic drift changes allele frequencies.

Selection

Directional selection changes allele frequencies at QTL determining the selected trait. Allele frequencies will also change at closely linked markers. This process, called hitchhiking, generates LD among markers around the selected locus [38, 39]. A region of increased LD, often accompanied by reduced polymorphism, can indicate a history of directional selection. Similarly, a region of increased LD and increased polymorphism can result from balancing selection. Such regions have been identified in maize and *Arabidopsis* for example [40, 41].

Migration and population admixture

If two populations, differing in allele frequency, are brought together, LD is created. Less extreme population admixture or migration also generates LD. If population admixture is known to have occurred and if markers are available which discriminate, even imperfectly, between the parental populations, then these markers can be used to map traits for which the populations differ. This is "admixture mapping" [42, 43] and has been applied in plants [44, 45].

More typically, migration and admixture are a problem for LD mapping. The long range LD they introduce mask the marker-trait associations arising from the close linkage which we wish to detect.

References

1 Bodmer, W.F. (1986) Human genetics: the molecular challenge. *Cold Spring Harbor Symp. Quant. Biol.* 51, 1-13

2 Duncan, C. *et al.* (2005) Recent Developments in Genomewide Association Scans: A Workshop Summary and Review. *Am. J. Hum. Genet.*, 77, 337-345.

3 Gupta, P.K. *et al.* (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol. Biol.* 57, 461-485

4 Winkler, C.R. *et al.* (2003) On the determination of recombination rates in intermated recombinant inbred populations. *Genetics* 164, 741-745

5 Darvasi, A. and Soller, M. (1995) Advanced Intercross Lines, an Experimental Population for Fine Genetic Mapping. *Genetics* 141,1199-1207

6 Mott, R. *et al.* (2000) A method for fine mapping quantitative trait loci in outbred animal stocks *Proc. Natl. Acad. Sci. U.S.A.* 97, 12649-12654

7 Mott, R. and Flint, J. (2002) Simultaneous Detection and Fine Mapping of Quantitative Trait Loci in Mice Using Heterogeneous Stocks. *Genetics* 160, 1609-1618

8 Yalcin, B. *et al.* (2005) Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* 171, 673-681

9 Valdar, W. *et al.* (2006) Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*. 172, 1783-1797
10 Valdar, W. *et al.* (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet.* 38, 879-887
11 Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.* 15, 1767-1776
12 Syvänen, A-C. (2005) Toward genome-wide SNP genotyping. *Nat. Genet.* 57, S5-S10
13 Macdonald S.J. *et al.* (2005) A low-cost open-source SNP genotyping platform for association mapping applications. *Genome Biol*. 6: R105
14 Spielman, R.S. *et al.* (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM) *Am. J. Hum. Genet.* 52, 506-516
15 Mitchell, A.A. and Chakravarti, A. (2003) Undetected Genotyping Errors Cause Apparent Overtransmission of Common Alleles in the Transmission/Disequilibrium Test. *Am. J. Hum. Genet.* 72, 598-610
16 Gordon, D. *et al.* (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet* 69, 371-80
17 Gordon, D. *et al.* (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* 12, 752-61
18 Allen, A.S. *et al.* (2003) Informative missingness in genetic association studies: case-parent designs. *Am. J. Hum. Genet.* 72, 671-80
19 Laird, N.M. and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*. 7, 385-394
20 Stich, B. (2006) A new test for family-based association mapping with inbred lines from plant breeding programs. *Theor Appl Genet*. 113, 1121-1130
21 Devlin, B. and Roeder, K. (1999) Genomic control for association studies, *Biometrics* 55, 997-1004
22 Reich, D.A. and Goldstein, D.B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* 20, 4-16
23 Bacanu S-A. *et al.* (2002) Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* 22, 78-93
24 Devlin, B. *et al.* (2004). Genomic control in the extreme, *Nat. Genet.* 36, 1129-1130
25 Clayton, D.G. *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet*. 37, 1243-1246
26 Zheng, G. *et al.* (2006) Robust Genomic Control for Association Studies. *Am. J. Hum. Genet.* 78, 350-356
27 Setakis, E. *et al.* (2006) Logistic regression protects against population structure in genetic association studies. *Genome Res*. 16, 290-296.
28 Price, A.L. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 38, 904-909
29 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959

30 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170-181
31 Falush, D. *et al.* (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567-1587
32 Aranzana, M.J. *et al.* (2005) Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. *PLoS Genet*. 1, e60
33 Yu, J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 38, 203-208
34 Zhang, Z. *et al.* (2006) TASSEL 2.0: a software package for association and diversity analyses in plants and animals. *Plant & Animal Genomes XIV Conference.*
35 Kraakman, A.T. *et al.* (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168, 435-446.
36 Buntjer, J.B. *et al.* (2005) Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci.* 10, 477-471
37 Caldwell, K.S. *et al.* (2006) Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, *Hordeum vulgare*. *Genetics* 172, 557-567
38 Maynard Smith, J. and Haig, J. (1974) The hitchhiking effect of a favourable gene. *Genet Res*. 23, 23-35
39 Barton, N. (2000) Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sc.i* 355, 1553-62
40 Palaisa, K. et al. (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci*. U.S.A. 101, 9885-9890
41 Tian, D. et al. (2002) Signature of balancing selection in Arabidposis. *Proc. Natl. Acad. Sci*. U.S.A. 99, 11525-11530
42 Darvasi, A. and Shifman, S. (2005) The beauty of admixture. *Nat. Genet.* 37, 118-119
43 Smith, M.W. and O'Brien, S.J. (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* 6, 623-632
44 Hu, Z.-M. (2005) Detection of linkage disequilibrium QTLs controlling low-temperature growth and metabolite accumulations in an admixed breeding population of Leymus wildryes. *Euphytica*. 141, 263 – 280
45 Buerkle, C.A. *et al.* (2006) Admixture in European Populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity* [Epub ahead of print]

Glossary

admixture: intermingling of individuals from genetically different populations.

analysis of variance: a method to test the statistical significance of differences among several categories, rather than just two; in which case a t-test is usually used.

candidate polymorphisms: polymorphisms which have not been chosen at random to test for trait association, but for which prior knowledge exists: they may be in a known linkage region or in a gene predicted to affect the phenotype for example.

centiMorgan (cM): a measure of genetic distance, additive over loci. At small values, distance in cM and recombination fraction (x100) are nearly identical.

chi-squared test: a widely used test of statistical significance.

consanguinity / kinship: close genetic relationships between individuals.

drift: the change in allele frequency over time which results from sampling variation from generation to generation.

false negative: the declaration of an outcome as statistically non-significant, when the effect is, in fact, genuine.

false positive: the declaration of an outcome as statistically significant, when there is no true effect.

family based linkage analysis: a method of mapping in which the co-inheritance of markers and traits is related to known genetic relationships between members of the same family or pedigree.

haplotype: a set of genetic markers located on the same chromosome, sufficiently closely linked to tend to be inherited as a unit.

landrace: an old cultivated form of a crop, potentially adapted to local growing conditions, but unimproved by contemporary plant breeding.

linkage disequilibrium (LD): the non-random association of alleles at separate loci located on the same chromosome (see Box 1).

logistic regression: a form of regression analysis in which the dependent variable is either 1 or 0, denoting presence or absence. In human genetics and epidemiology it is commonly used with 1 denoting diseased individuals and 0 healthy or control individuals. It can also be used to regress the presence/absence of a particular allele at a locus onto phenotype, as an alternative to the t-test.

mapping: the process of locating a genetic variant on a chromosome. Coarse mapping will only locate a variant within a broad interval. Fine mapping increases precision, ultimately allowing the identification of the functional polymorphism(s) responsible.

mapping population: a set of individuals or lines, typically derived from an F2 or a backcross, which are used to construct genetic maps and to detect and locate QTL on those maps by family based linkage analysis.

marker: an identifiable location on a chromosome.

microsatellite: repetitive lengths of short DNA sequences used as genetic markers.

multiple regression: regression analysis in which there are multiple independent variables. In LD mapping, these could be multiple markers, within the same or different genes.

multiple testing: in an experiment involving many candidate polymorphisms, many statistical tests will be carried out. A consequence of this multiple testing is that it is more likely that a false positive result will be declared by chance. Modified methods of significance testing can control the expected number of false positive results.

non-experimental population: a population not established specifically to map markers or QTL. It is not necessarily a natural population. For example, it could be a collection of breeders' lines.

population structure: the non-random distribution of genotypes among individuals within a population.

population subdivision : the partition of a population into subgroups such that most mating occurs within subgroups.

quantitative trait locus (QTL): a polymorphic site contributing to the genetic variability of a quantitative trait.

recombination fraction: the fraction of meiotic events that show recombination between a pair of loci.

single nucleotide polymorphism (SNP): a polymorphism involving a change in only a single nucleotide.

stepwise selection: a set of methods in which the best subset of all independent variables available for multiple regression is selected. Ideally, only those variables which have an effect on the dependent variable are selected, and all others are rejected. In LD mapping this approach attempts to separate markers affecting a trait from those which do not.

structured population: a population in which mating does not occur at random.

t-test: a test for the statistical significance of a difference between two means.

THE MIXED MODEL AND ASSOCIATION GENETICS

No good model ever accounted for all the facts, since some data was bound to be misleading if not plain wrong. James Watson

Introduction

This is an attempt to explain the mixed model, particularly in the context of association genetics. This has become the default method of controlling kinship and population structure in association mapping in plants. All methods fit models to the data. Different models have different weaknesses. In the absence of an experimental crossing scheme or a prior understanding of genetic relationships within the dataset, the mixed model seems to be the best compromise, but it is not perfect. Where possible, I've highlighted relationships between the mixed model and these alternative approaches.

I have laboured discussion about the form of the genetic relationship matrix, which has a central role in genetic applications of the mixed model. In my view this is quite difficult to understand, and different accounts are often ambiguous. Moreover, although software exists to solve the mixed model equations, often it is left to the user to provide their own kinship matrix, or to choose between alternatives and it is quite easy to get this wrong.

A model is described as mixed if, in addition to a base error or residual term, it contains a mixture of fixed and random terms. Remember, fixed effects are generally the things that you are interested in, for example fertilizer treatments, and random effects are generally things that get in the way of accurate estimation of the fixed effects, for example fertility effects in yield trials. This distinction is not always precise: blocks in a variety trial, usually treated as random variables, can sometimes be better considered as fixed if they are placed to account for specific known fertility effects in the experimental field. However, I've never come across an experiment in which fertilizer treatments could justifiably be regarded as random, though I expect one could be devised.

The Bayesian approach to statistics has no truck with these distinctions: all effects are random. The only thing that differs among different effects is their prior information.

In association mapping, the mixed model contains two random effects: the base error and an additional term to account for genetic variation among individuals (or lines). These individuals are often related. In such a case, to account for the genetic correlations among individuals we must incorporate into the model something called the numerator additive genetic relationship matrix (A). This is often referred to as the relationship matrix, which confusingly, isn't quite the same thing. The model also contains at least one fixed effect: the marker for which association with the trait is being tested.

A simple example

First we'll consider the difference between treating a line as fixed or random in a very simple genetic experiment. Suppose we have a set of *n* inbred lines derived from an F1, which we test in a completely randomised experiment with each line being tested in *r* replicates.

Whether treated as fixed or random, the model we are fitting is:

$$y_{ij} = \mu + g_i + e_{ij}$$

where y_{ij} is the jth observation on the ith line
μ is the mean
g_i is the effect of the ith line
e_{ij} is the error for the jth replicate of the ith line.

If we treat the lines as random, the analysis of variance has the following structure:

Item	df	expected mean square
Between lines	n-1	Ve + nVg
Within lines	n(r-1)	Ve

We can test for statistical significance among the lines with an F test in the usual manner. Vg and Ve can be estimated by equating the observed mean squares with their expectation or we could use an algorithm like REML to deliver us the variance components directly. In this case ANOVA and REML estimates would agree exactly. Heritability can be estimated as Vg/(Ve+Vg). Remember this is the heritability for a line tested in a single plot. For a line tested in *r* plots the heritability of a line mean would be Vg/(Ve/*r* +Vg).)

If we treat the lines as fixed – perhaps they are not a random sample from the cross but were selected for some other reason – then the analysis of variance, expected mean squares and significance test remain unchanged. However, Vg can no longer be regarded as the genetic variance of the population of lines which could potentially be produced, but merely a measure of the variance among the fixed set of lines which we have selected in this study. (NB If the lines were selected in a known manner, say only lines exceeding some phenotypic threshold were included, then in principle the selection process could be included in the model and an estimate of the population Vg could still be made.) The difference between treating the lines as fixed or random is in the estimation of their means. Treating the lines as fixed, the estimate of each mean is just the average over all replicates:

$$\sum_j y_{ij} \; / \, r \; = \; \hat{\mu} \; + \; \hat{g}_i \qquad \text{with variance Ve/r}$$

This follows since e ~ N(0,Ve) and we average over r e_{ij} terms in estimating the mean. This estimate is the Best Linear Unbiased Estimate, or BLUE.

The estimate of a variety mean, when treating varieties as random is more complicated. Each line is regarded as a random draw from a population of lines which has a distribution:

$$g \sim N(0,Vg)$$

In this case, even if there are no observations for any particular line, its effect can still be estimated:

$$\hat{g}_i = 0 \qquad \text{with variance Vg}$$

Therefore the estimate of the mean of the line with no observations is just the estimate of the population mean, $\hat{\mu}$.

In Bayesian terms, the distribution g ~ N(0,Vg) is the prior distribution of g.

If we have data we also have a direct estimate of the effect - $\Sigma y_{ij} / r - \hat{\mu}$, just as in the fixed case. We therefore have two sources of information, from the prior and from the data. These can be combined to produce a best estimate of g. Intuitively, as the number of replicates goes up and up, we would expect to place greater weight on the data and rely less on the prior information. In the limit we would ignore the prior information entirely and rely on the estimate from the data alone. To formalise this, we take a weighted mean of the two estimates; weighting by the reciprocal of the variances:

w_f = r/Ve for the fixed effect

w_p = 1/Vg for the prior:

As with all weighting procedures, we need to scale the weights to add up to 1 by dividing by their sum. This is:

$$r/Ve + 1/Vg = (rVg + Ve)/ (rVe.Vg)$$

Labelling the estimates of g as g_f and g_r for the fixed and random effects and the prior as g_p :

$$g_r = (w_p g_p + w_r g_r)/ (w_p + w_r)$$

$$= 0 \ . \ 1/Vg + g_f (r/Ve)] / [(rVg +Ve) / (rVe.Vg)]$$

$$= g_f (r/Ve) rVe.Vg / (rVg +Ve)$$

$$= g_f rVg ./ (rVg +Ve)$$

Dividing the numerator and denominator of the last term by r gives:

$$g_r = g_f (Vg) / (Vg + Ve/r)$$

Vg/(Vg+Ve/r) is just the heritability of a line mean based on r replicates and so we have:

$$g_r = g_f h^2$$

where h^2 is the heritability of the line mean.

The mean of the variety is then:

$$\hat{\mu} + g_f h^2$$

So the estimate of the random effect is the deviation of the line mean from the population mean scaled by the heritability. In Bayesian terms, this estimate has a term for the prior information (with mean 0) and a term coming from the data (g_f) which is weighted by the heritability. For very high heritabilities, means from the random and fixed effects models are very similar. With very low heritabilities, the contribution from the data is reduced and the estimate is shrunk back towards the population mean. "Shrinkage" is the term generally used to describe this difference between means from random effects models compared to fixed effects models.

There is a simple, non Bayesian, way of viewing the random effects estimate. Remember the breeders' equation:

$$R = h^2 S$$

The response to selection (measured as a deviation from the population mean) is just the mean of the selected group (also measured as a deviation) multiplied by the heritability. When selecting a single variety, S is the difference between the variety mean and the population mean (ie g_f) and R or g_r is its predicted deviation in the future, on retesting in a second experiment. So in the context of breeding, fixed effects for lines are the observed effects from the data, and random effects for lines are the predicted future effects of selected lines.

Means for the random terms in the model are called Best Linear Unbiased Predictors or BLUPs.

Note, in passing, that for the purposes of selection in breeding programmes, provided all lines are equally replicated, whether they are treated as fixed or random effects makes no difference to the line ranking. This is because the heritability of all line means is identical – all estimates are shrunk by the same constant. However, if replication varies from line to line, then individual line means will be shrunk by varying proportions as the heritability varies from line to line. In this case, the ranking can change. Treating the lines

as random is of use here to create a ranking for selection purposes which takes into account the varying degrees of replication. Other methods of selection among lines with unequal amounts of information can get awfully messy. There is much to recommend treating varieties as random even if they cannot be regarded as a truly random sample from a population. (Piepho 2007).

Association analysis with fixed and random effects: simple example.

Consider the same example as above, but suppose there is a marker, m, segregating which we wish to test for association with the trait. As we are working with lines derived from an F2, provided these lines are a random sample from the cross, or provided the selection of lines was independent of the marker genotype, the test of significance between marker classes is a direct test for linkage. The lines are fully inbred so there are only two marker classes and the analysis of variance, treating lines as random, takes the form:

Item	df	expected mean square
marker alleles	1	Ve + nVg + kVm
Between lines (within alleles	n-2	Ve + nVg
Within lines	n(r-1)	Ve

In this example, the marker is treated as a fixed effect. In most linkage and association analyses, marker effects are treated as fixed, but this is not always the case. In this, therefore, the variance Vm associated with the marker SNP cannot be used to make inferences about some population of markers. With no missing data and if each marker allele is carried by half the number of lines, the coefficient of Vm (k) will be nr/2. However, Mendelian sampling makes it unlikely that exactly half the lines carry each allele and k will be slightly less than nr/2

From the expected mean squares, the appropriate F test for marker association is obviously to test the marker alleles item against the between lines item. This will give the same p-value as a t-test between the marker alleles, working on line *means:* standard practice for single marker linkage analysis. However, if lines were treated as fixed, the analysis of variance has the following form:

Item	df	expected mean square
marker alleles	1	Ve + kVm
Between lines (within alleles)	n-2	Ve + nVg
Within lines	n(r-1)	Ve

The marker alleles item is now tested against the within lines item. Assuming there is any genetic variation at all within the cross, treating lines as fixed will inevitably give rise to an increased frequency of significant results compared to treating the lines as random.

A more complex family structure

Suppose we have a set of 200 F1 crosses with five inbred lines derived from each cross: 1000 inbreds in total. For the time being, we'll assume that the 400 parents of these F1s are unrelated. Therefore the five lines within each cross are related but lines from different crosses are not. More accurately, we should say that the lines from different crosses have very low relatedness: there is no such thing as an unrelated pair of lines. If one goes back far enough there will be a common ancestor.

Taking this family structure and a phenotype, we can analyse the phenotypic data in an analysis of variance:

Item	df	expected mean square
Between crosses	199	(Ve + Vg) + 5 Vg
Within crosses	800	(Ve + Vg)

As there is no replication of individual inbred lines, we do not have an independent estimate of error (Ve). However, we can still easily estimate Vg and Ve from the expected mean squares. This is for a random effects model: differences between F1s crosses are treated as random. Vg is therefore an estimate of the genetic variation in the population from which the parental lines were selected at random.

If we were to treat differences between F1s as fixed effects, the analysis of variance remains the same, but the expected means squares change:

Item	df	expected mean square
Between crosses	199	(Ve + Vg) + 5 Vg*
Within crosses	800	(Ve + Vg)

Vg* is the estimate of genetic variation among crosses within the sample used in this particular experiment. Vg is the genetic variation expressed within crosses and must be assumed to be identical within all crosses. Note that Vg* can be estimated, but Vg cannot be separated from Ve. However, the F ratio to test for differences between crosses is valid (and identical) whether crosses are treated as fixed or random.

Estimates of cross means themselves differ depending on whether they are treated as fixed or random. The model is:

$$y_{ij} = \mu + c_i + g_{ij} + e_{ij}$$

where y_{ij} is the jth observation on the ith cross
μ is the mean
c_i is the ith cross effect
g_{ij} is the effect of the jth line in the ith cross
e_{ij} is the residual effect for the jth line in the ith cross.

The g_{ij} and e_{ij} are completely confounded, so we can simplify the model by ignoring g_{ij}:

$$y_{ij} = \mu + c_i + e_{ij}$$

In effect we have now redefined the e term. What was formerly g + e is now regarded as just e, distributed as e ~N(0,Ve)

For fixed effects, the estimate of a cross mean is just the arithmetic mean of the individual lines within that cross:

$$\hat{\mu} + \hat{c}_i = \sum_j y_{ij} / 5$$

For random effects, the estimate of a cross mean is shrunk to:

$$\hat{\mu} + \hat{c}_i (Vg) /(Vg + Ve/5)$$

Now suppose there is biallelic marker segregating in the population at a frequency of 0.5 which we wish to test for association with the trait. We expect ½ the crosses to be segregating for the marker, ¼ to be fixed for one allele and ¼ to be fixed for the other. This makes the analysis of variance highly unbalanced. Treating the crosses as fixed, the significance of the fixed terms (markers and crosses) depends on the order in which they are fitted. Each term fitted is tested against residual variation after fitting that item. This will include not only error variation but also variation due to the remaining terms. Equally, the lack of balance introduces correlations between the two fixed terms such that some variation is common to both and is claimed by the first to be fitted. For example, with simulated data and an associated marker, I get the following F test statistics:

	Marker fitted first		Crosses fitted first
Marker	306.13	Crosses	3.37
Crosses	2.56	Marker	145.01

Treating crosses as fixed, fitting the effect for crosses first in the analysis of variance is analogous to "structured association" (SA). Generally with SA, covariates which indicate the membership of each line to each subpopulation are fitted first. In this example, the different crosses are equivalent to different subpopulations. We see that the F ratio for the marker is much reduced if it is fitted after the term for crosses: test of significance of the marker has been adjusted for spurious association which arises from the distribution of the SNP genotypes over families.

In crop association genetics, SA is less effective than the mixed model in controlling false positive results. This is partly because it is hard to assign the ancestry of each line accurately to appropriate subpopulations and partly that multiple separate ancestral subpopulations never existed in the first place. A symptom of this is the very common complaint among crop scientists using the software STRUCTURE that it is impossible to decide how many subpopulations there are, or to get stable and repeatable results from

replicate runs of the program. The extended pedigree on which all members of any collection of lines or individuals lie results in kinships among individuals varying in a near continuous manner. In contrast, in most human association mapping datasets, virtually all pairs of individuals can be treated as unrelated.
An analysis of variance is also possible if families are treated as random, but no longer has a simple form. Because between family variation is treated as a source of error, an estimate of the SNP effect can be made from comparisons between crosses in addition to the comparisons within crosses on which the fixed effects model relies (if family effects are fitted first). These two estimates have different amounts of information. Using the same data as above, the random effects anova is:

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
cross stratum					
SNP	1	125.2934	125.2934	63.59	<.001
Residual	198	390.1028	1.9702	2.57	
cross.*Units* stratum					
SNP	1	111.3178	111.3178	145.01	<.001
Residual	799	613.3670	0.7677		
Total	999	1240.0810			

The analysis has two separate strata and two separate significance tests. These can have slightly different interpretations (see below), but each provides an estimate of the SNP effect. A mean across the two strata, weighted by their relative information is also given. Increasingly, with unbalanced data like this, the analysis is carried out using REML which can handle more complex data sets than possible with anova. REML gives tests the significance of the fixed effects and estimates of their means (identical to the anova weighted mean in this example) in addition to estimating variance components. In essence, the significance test for the SNP is adjusted to account for the differing kinship relationships among lines. In this example there are only two: pairs of lines within crosses are more related than pairs of lines from different crosses. In this case, the simple structure of the data mean that it isn't necessary to specify what these relationships are, but this is not generally the case in association mapping.

To illustrate further the different consequences of treating crosses as fixed and random, I'll give results from some more simulations. To recall, there are 200 F1s with 5 inbred lines derived from each. Vg = 0.25, Ve = 0.5. Ve refers to the error of an inbred mean. Assuming the F1s are unrelated, and that we are dealing with inbred lines, the genetic variation between F1s = 0.25 and the genetic variation between lines within F1s is also 0.25 . So the total expressed genetic variation = Ve = 0.5 and Vp = 1.

I've simulated two SNPs:

SNP1: half the families are fixed for allele A, the other half for allele a.

SNP2: alleles segregate at random among the population of parents at a frequency of 0.5, so half the families are expected to be Bb, the others either BB or bb.

SNP1 partitions the data into two subgroups. Such a SNP could result from a genuine linkage with a trait locus. For example it could be a vernalisation gene splitting the data into Spring and Winter subgroups. The associated phenotype could be flowering time. Alternatively, the SNP could be one of many markers which are not linked to the trait but which happen to be associated (through drift or selection) with the partition of the lines into Spring or Winter types.

There are three phenotypes, simulated from Vg, Ve, and the SNPs. These are:

Vg and Ve only
Vg and Ve plus 1 (for AA inbreds) or plus 0 (for aa inbreds): a SNP1 effect
Vg and Ve plus 1 (for BB inbreds) or plus 0 (for bb inbreds): a SNP2 effect

Data were analysed with no family effect, with family effect fixed, and with family effect random. The table below gives results for all 18 tests. The test statistic is the Wald statistic. This is frequently used in mixed models. It can be treated as a chi-squared statistic, here with 1 df so a value >3.84 corresponds to a p-value < 0.05.

		Wald statistics	
phenotype	family effect	SNP1	SNP2
Vg + Ve	none	5.66	2.62
Vg + Ve	fixed	0	0.04
Vg + Ve	random	2.97	0.76
Vg+Ve+SNP1	none	187.47	0
Vg+Ve+SNP1	fixed	0	0.04
Vg+Ve+SNP1	random	98.44	0.02
Vg+Ve+SNP2	none	0.41	205.07
Vg+Ve+SNP2	fixed	0	112.4
Vg+Ve+SNP2	random	0.2	173.35

The first three lines of results are for the null phenotype. Anything significant is a false positive. Without accounting for family structure, SNP1 is statistically significant and SNP2 has quite a large test statistic too. This is a result of "double counting." The analysis assumes all lines are unrelated so we think there are more independent data than there actually is, error is underestimated and significance rises. However, with families as fixed, any test signal is completely wiped out. With families treated as random, SNP1 still gives quite high results.

The next three lines give the results for the phenotype influenced by SNP1. In this case, all three test statistics for association with SNP2 are very small. Association with SNP1 is very strong with no family effect included. The signal is roughly halved with families treated as random, and is again completely wiped out with families treated as fixed. SNP1 could be a genuine effect or it could be indirectly associated with subpopulation structure. Although association mapping methods act to control the problem of population structure and kinship, there is always a risk that they also remove genuine affects. Caveat emptor.

The final three lines give the results for the SNP2 phenotype. There is no significant SNP1 effect and everything is significant for SNP2, with random families giving a signal intermediate between the two other tests. The reason the fixed effect model has less power here than the random effects model is that the fixed effect model only uses information from the within family stratum of the analysis. With families treated as random, an estimate of the difference between SNP alleles is also made from differences between family means, and the two estimates are optimally combined. This is analogous to the recovery of inter-block information in trial designs: variety comparisons are also possible by making comparisons between blocks if the blocks are treated as random effects.

Thus, a major difference between the mixed model and structured association - in fact the only one in this example – is that the mixed model treats population structure (family membership in this example) as random and structured association treats it as fixed.

A significant result from the between family stratum in the random effects model could result from population subdivision (eg SNP1) rather than LD. However, the within family comparison from either fixed or random effects models is completely robust to the cause of any differences between families. This is because it is a direct test for linkage of the marker with the trait. Pooled over a large number of families it is a 1 df test (for a bi-allelic marker) for linkage disequilibrium rather than just for linkage. This is because genetic linkage between the QTL and marker will give a significant result within a single family, but when averaged over a large number of segregating families, the magnitude of the average association will be reduced. The direction of the association will change from cross to cross, depending on whether the QTL and marker are linked in coupling or in repulsion. Over many crosses repulsion and coupling linkages will cancel each other out. However, the closer the marker to the QTL, the stronger the LD is likely to be, so that either coupling or repulsion linkages will come to dominate. This is a plant equivalent of the transmission disequilibrium test in humans, with the same advantages and disadvantages: it is extremely robust to the presence of population structure but is very wasteful of data: no information from the non- segregating families is used.

It is possible to develop this further. For n segregating families, there would be n 1 df tests for linkage. Pooling these across families would give an n degree of freedom test for linkage which could be partitioned into a 1 df test for the net effect of linkage (ie for LD), and an n-1 df test for heterogeneity of association (ie for residual linkage). This is in effect a joint regression analysis on SNP genotypes coded as 0 and 1. The joint regression

term and heterogeneity of regression term correspond to the test for linkage disequilibrium and residual linkage respectively. One would expect that as markers get closer and closer to the QTL, the association test would get stronger and the heterogeneity test smaller. This is the basis of the quantitative trait disequilibrium test, or QTDT, used in human genetics (Abecasis et al 2000). A more general procedure has also been proposed for crops (Stich et al 2006).

In our example, the families fixed (SA) analysis does striking well in adjusting for false positives – all signal is wiped out except for the one clearly genuine case: (SNP2 phentype : SNP2 genotype). However, power has been lost compared to the mixed model for reasons discussed. The mixed model (families as random) doesn't do badly - with more power than SA but not quite such good control of false positives. This is the opposite of the current view of methods for association mapping in plants where the mixed model is held to be superior. There are two reasons for this. Firstly, with real datasets there can be great problems in allocating individuals to families or subpopulations. Secondly, the simple hierarchical subdivision of the data into discrete families or subpopulations modelled here is virtually never appropriate. Relationship among variety pairs is effectively a continuous variable whereas our simulated example has only two values. The most effective way to treat this is to incorporate relationships into the model. This is done through an extension to the mixed model which incorporates a relationship matrix into the analysis. However, the mixed model doesn't always guarantee perfect control of all genealogical sources of false positives, as we've seen in the simple example here. This can be viewed as a failure of randomisation: in trial designs, the randomisation process – of varieties within blocks and of blocks over the field, guarantees unbiased estimates. In association genetics, you have to work as best as you can with what you are given.

The role of the relationship matrix

In matrix form, a fuller form of the full mixed model is:

$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{e}$

$\mathbf{Y}$ = phenotype data	
$\boldsymbol{\beta}$ = the fixed effects	includes the marker(s) tested for association
$\mathbf{g}$ = the random effects	the genetic effects (families in our example)
$\mathbf{X}$ = design matrix for fixed effects	
$\mathbf{Z}$ = design matrix for random effects	aka **A**, the numerator additive relationship matrix
$\mathbf{e}$ = residual error terms	

The residual errors terms are, as usual, treated as unrelated and are represented as:

$\text{Var}(\mathbf{e}) = \text{Var}(\mathbf{e}) = \sigma^2_e \mathbf{I}$

where **I** is an identity matrix. (A square matrix with ones down the leading diagonal zeros elsewhere.)

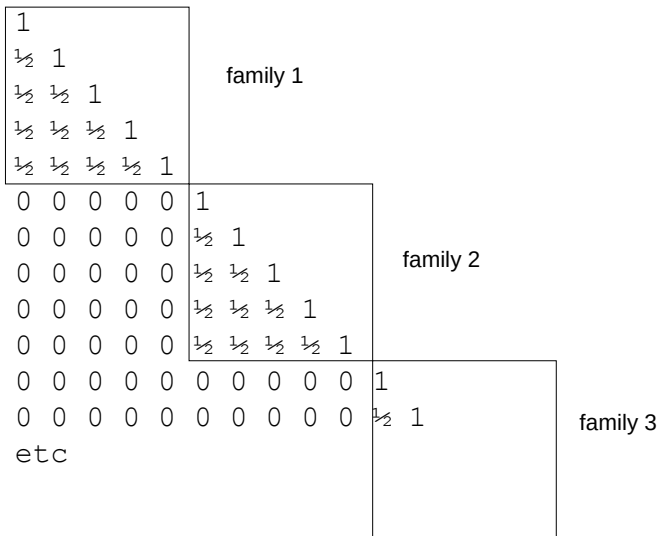
The random genetic effects are no longer treated as independent of each other. They have a variance/covariance structure given as

$\text{Var}(g) = \sigma^2_g \mathbf{K}$

K is the additive relationship matrix. If individuals are outbred and unrelated this is also an identity matrix.

The relationship matrix is symmetrical. Part of this matrix for the example we used in the last section is below:

Example numerator additive relationship matrix



The composition of this matrix is unambiguously explained by CR Henderson (1963, 1976) who was pivotal in developing and promoting the use of the mixed model in animal breeding as a means of estimating breeding values. Its elements are the coefficients of the additive genetic (co)variance between the individuals in the dataset. In the context of animal breeding, the matrix generally includes all founder individuals and ancestors of those in the dataset: the pedigree is extended backwards. Founders are assumed to be non-inbred. This makes the prediction of breeding values robust to the effects of selection within the pedigree. Strictly, the additive numerator relationship is applicable to additive traits only. However, the mixed model can be extended to include dominance and epistatic sources of genetic covariance too, by including relationship matrices for these effects.

The diagonals of the matrix are the coefficients of additive genetic variance for the individuals themselves. The off-diagonal entries in the table are coefficients of relationship, or twice the coefficients of kinships, both introduced earlier in the population genetics section of this manual. The coefficient of kinship between two individuals is the probability than an allele picked at random from one individual is identical by descent to an allele picked at random from the other, or p(ibd). These are 1/4 for full-sibs and 1/8 for half-sibs. For a population with no inbreeding, the diagonal entries are also coefficients of relationships or 2x the coefficients of relationship. The p(ibd) of an non-inbred individual with itself is a half (ie the inbreeding coefficient of its progeny), so the coefficient of relationship of an outbred individual with itself is one. Thus, in the absence of inbreeding, the complete relationship matrix required in the mixed model equations is 2x the kinship matrix.

With inbreeding, the off-diagonal elements are still genetic relationships or 2x the coefficients of kinship. The diagonal elements, however, are no longer relationships but 1+F; the coefficient of the additive genetic variance for an individual with inbreeding coefficient F. The diagonals will thus have a maximum value of 2 and a minimum of 1.

For a population of inbred lines, or doubled haploids, the diagonal of the relationship matrix required for the mixed model equations is 2. The inbred line analogue of a full-sib family is a set of progeny lines derived from the same cross. In this case, p(ibd) (aka kinship) is ½ so the off diagonal elements will be 1. The whole matrix is therefore 2x the equivalent matrix for an outbred population. However, as we shall see, this simple scalar adjustment to the matrix makes no difference to the association test and is therefore often ignored. The additive relationship matrix for the mixed model equations is therefore commonly treated as if it were merely the kinship matrix or 2x the kinship matrix. If the population under study is all fully inbred and there are no crosses between related lines this is not a problem. If this is not the case, this is dangerous. How dangerous in practice, I do not know.

We shall return to methods to construct the relationship matrix shortly. First, we'll describe some results analysing the simple example given below.

```
id  pa ma  phen  SNP
  5 1  2   0.665  1
  6 1  2   2.113  0
  7 1  2   0.667  1
  8 1  2  -0.161  0
  9 1  2   0.170  1
  10 3  4   1.669  0
  11 3  4   1.957  1
  12 3  4   0.108  0
  13 3  4   1.789  1
  14 3  4   2.881  0
```

This example has two full sib families, each with five individuals, a single phenotype and a single SNP. The first family has father no. 1 and mother no. 2. The second family has father 3 and mother 4. We'll treat the parents as unrelated. (It may seem odd that there are only two SNP classes – 0 and 1 for an outbreeder. We'll assume that one allele is

sufficiently rare that no homozygotes were observed. If they were, they would be coded as 2.)

First, analysing the data ignoring the pedigree information gives

Error variance Ve	= 1.161
Test for association (Wald statistic)	= 0.16

The significance of the SNP association is tested by the Wald statistic. This is the standard test statistic used by the GenStat implementation of REML. Other software may produce a likelihood ratio statistic, which also behaves like a chi-squared test. In mixed models however, the LRT is no longer simply twice the difference in likelihoods between models.

As the data are nicely balanced, we can fit a mixed model just by including a random term for families without explicitly defining a kinship matrix. In fact the kinship matrix used is **I**. This gives:

Variance between families	Vb = 0.266
Variance within families	Vw = 1.001
Wald stat	= 0.06

Thus, incorporating the family structure has reduced the Wald statistic. Also, with no inbreeding and assuming an additive trait,

$$\begin{aligned} Vb &= \tfrac{1}{2} Vg \\ Vw &= \tfrac{1}{2} Vg + Ve \end{aligned}$$

So

$$\begin{aligned} Vg &= 0.532 \\ Ve &= 0.735 \\ Vg/Ve = \gamma &= 0.724 \end{aligned}$$

In mixed model methodology, rather than reporting Vg, often the ratio $\gamma = Vg/Ve$ is given; here 0.724. This is because:

$$\begin{aligned} Var(y) = Var(g) + Var(e) &= \sigma^2_g K + \sigma^2_e I \\ &= \sigma^2_e I (\gamma K + I) \end{aligned}$$

It is γ rather than σ^2_g which the computer algorithms tend to work with and so this often gets reported by default. Note that heritability is

$$\sigma^2_g / (\sigma^2_g + \sigma^2_e) = \gamma / (1 + \gamma)$$

To adjust for kinship requires that the numerator relationship matrix is be inverted. For small datasets, this is no problem, but for very large datasets, national dairy herds for

example, this is computationally intensive. CS Henderson developed a method whereby the inverse of a kinship matrix could be written down directly from the pedigree. This is encoded as a special routine in some statistical software, VPEDIGREE in GenStat for example. However, for this example, we shall provide the matrix and incorporate it directly. I've given the GenStat code for this below.

```
"Simple association analysis using the mixed model in GenStat."

"In the results, the between family component is expressed as gamma = Vg/Ve."
"Heritability is then gamma / (1 + gamma)."
"Beware - gamma and therefore Vg are very specific to the relationship matrix used."
"They refer to a conceptual population with an identity matrix"
"(1's on the diagonal, 0's off.) so if the marker based relationships show all lines"
"are related, then Vg can be surprisingly large."
"This doesn't affect the validity of the test for association, however."

"Clear out the old data."

endjob


"Read in data"

"id - unique identifier for each entry."
" ma and pa - you don't need - only if using GenStat's VPEDIGREE command to get kinship
estimates from the pedigree."
"phen is the phenotype."
"SNP is the marker to be tested for association."

read id,ma,pa,phen,SNP
  5 1 2   0.665  1
  6 1 2   2.113  0
  7 1 2   0.667  1
  8 1 2  -0.161  0
  9 1 2   0.170  1
  10 3 4   1.669  0
  11 3 4   1.957  1
  12 3 4   0.108  0
  13 3 4   1.789  1
  14 3 4   2.881  0
 :

"Note: if using VPEDIGREE parents precede progeny in id number"

"Convert from variates to factors."

group [re=yes]id,ma,pa,SNP


"Read relationship matrix"
"The example is for two unrelated full sibs families with five individuals each."

SYMM [r=10] relationships
read relationships
 1.0
 0.5 1.0
 0.5 0.5 1.0
 0.5 0.5 0.5 1.0
 0.5 0.5 0.5 0.5 1.0
 0.0 0.0 0.0 0.0 0.0 1.0
 0.0 0.0 0.0 0.0 0.0 0.5 1.0
 0.0 0.0 0.0 0.0 0.0 0.5 0.5 1.0
 0.0 0.0 0.0 0.0 0.0 0.5 0.5 0.5 1.0
 0.0 0.0 0.0 0.0 0.0 0.5 0.5 0.5 0.5 1.0
 :


"You need to create a dummy variable to partition G and E within each individual"
```

```
factor [lev=10; val=1...10] id2


"Do the analysis"

 VCOMPONENTS [SNP] RANDOM=id2+id; INITIAL=1; CONSTRAINTS=none
 VSTRUCTURE [TERMS=id]  MODEL=fixed;matrix=relationships

"A warning will be printed that we have specified two residual terms."
"This is OK - ignore it."

 REML [PRINT=model,components,WALD] phen
```

This code isn't that intuitive, which is why I have given it here in full. In particular, in GenStat, it is necessary to define a second index variable (id2), identical to the first (id1) and explicitly fit variance components to both. One indexes the Ve terms, the other Vg. Any attempt to do otherwise leads to madness. On running, this code will give exactly the same answer as before. We can now study the effect of varying kinships. Suppose we had the more complex relationship matrix given below:

```
 1.0
 0.5   1.0
 0.5   0.5   1.0
 0.5   0.5   0.5   1.0
 0.5   0.5   0.5   0.5   1.0
 0.125 0.125 0.125 0.125 0.125 1.0
 0.125 0.125 0.125 0.125 0.125 0.25 1.0
 0.125 0.125 0.125 0.125 0.125 0.25 0.25 1.0
 0.125 0.125 0.125 0.125 0.125 0.25 0.25 0.25 1.0
 0.125 0.125 0.125 0.125 0.125  0.0 0.0  0.0  0.0 1.0
```

Note there are very few pairs of unrelated individuals. The analysis now gives:

Ve	=	0.305
γ	=	3.615
Wald stat	=	0.10

The relationship matrix is not totally flexible. It must be positive semi-definite: it must have an inverse. If this is not the case, this implies that there is something wrong with the specification of the relationship matrix. Identical (or even very closely related individuals) can cause problems. With our simple example, the following relationship matrices give *exactly* the same Wald statistic for association with the SNP.

	diagonal	within families	between families	Ve	gamma	Vg
1	1	0.5	0	0.735	0.725	0.533
2	1	0.25	0	0.203	5.259	1.068
3	1	0.5	0.25	0.469	2.272	1.066
4	1	0.5	-0.5	0.868	0.307	0.266
5	2	1	0	0.735	0.362	0.266
6	2	0.5	0	0.203	2.629	0.534

The relationship matrix specifies the form of the variances and covariances. With regard to the test for association, the absolute values don't matter: simple addition or multiplication of all the elements of the matrix has no effect on the test for association. Even negative values on the off diagonal have no effect. This is discussed below. The form of the relationship matrix does, however, affect the interpretation of the variance components.

For the top three rows of the table, the relationship matrix is appropriate for two full-sib families, two half sib-families and two full sib-families nested within a half-sib (something akin to the Nested Association Mapping design of Ed Buckler) respectively. Vg represents the variance among unrelated non-inbred individuals from the population (ie a set of individuals with a relationship matrix of **I**). The estimate of Vg in the second line is twice that in the first line, as it should be: the coefficient of Vg in the between families mean square is ½ assuming full-sib families and ¼ for half sibs. Inevitably the estimate of Vg doubles.

The fourth row has negative relationships among the individuals classified previously as unrelated. Ignoring how these relationships might be estimated, we shall first consider what they mean. As mentioned earlier, relationships among individuals are based on a fiction: that there is such a thing as a population in which all individuals are unrelated. (It is for this reason that RA Fisher eschewed the use of Sewall Wright's inbreeding coefficients and ploughed his own furrow to quantify inbreeding through the theory of junctions, or so my old supervisor and guru JS Gale, one of RA Fisher's last PhD students, told me.) Once we have defined a reference population, we can estimate relationships and inbreeding coefficients of any individuals derived from that population. All relationships and inbreeding coefficients are relative to that reference. An extreme case is encountered when dealing with individuals derived from an F2. The F2 can be treated as the base population in which all individuals are defined as unrelated and non-inbred. However, the two parental lines are less related than individuals in the F2 and the F1 is less inbred than the F2. The only way to quantify this is to allow relationships and inbreeding coefficients to be < 0. (We've come across something similar to inbreeding coefficients <0 before: the term used to describe the departure of genotype frequencies from Hardy-Weinberg expectation can be either positive or negative depending on whether a deficiency or excess of heterozygotes is observed. In the wholly made up relationship matrix on line four, therefore, pairs of individuals with negative relationships are less related than are pairs in the (conceptual) reference population.

The remaining two lines of the table are cases where the diagonal elements are no longer one, yet even in these cases, the Wald statistic remains unchanged. Case 5 represents inbred lines derived from two unrelated F1s. Case 6 is the inbred line equivalent of two half sib families: inbred lines within each family have one common inbred parent and one unrelated parent. Cases 5 and 6 have estimates of Vg which are half that for the outbred analogous cases 1 and 2. This is correct. The estimate of Vg refers to the conceptual outbred population of unrelated individuals. The additive genetic variation among these outbred individuals is expected to be half that seen among unrelated inbred lines. We

could use relationship matrices of the form given in cases 1 and 2 to analyse the inbred data given in cases 5 and 6. This is common practice, but we must also take care over the interpretation of Vg (or γ) and understand the estimate produced is for a population of inbred lines and not of outbred individuals.

Occasionally, one may have a mixed collection of inbred and outbred individuals to analyse. For example, in maize one may be working on inbreds and F1s. (In this case, a simple additive model of genetic variation is unlikely to be acceptable. We would need to account for dominance. This can be done, but not here.) In this case, the diagonal elements would be mixed – some would be 1 and some 2, or more generally (1 + F). In our example, if we treat the second family as inbred – diagonals 2 and off diagonals 1, and the first as outbred – diagonals 1 and off diagonals ½ then we get:

Ve	=	0.672
Vg	=	0.412
γ	=	0.614
Wald	=	0.10

These estimates seem reasonable.

The estimate of the additive genetic relationship matrix

Traditionally, particularly in animal breeding, the mixed model has been used for association mapping. For this purpose, the pedigree of all animals in a breeding herd is tracked. The genetic relationship matrix is calculated directly from the pedigree. To speed up computation, Henderson developed simple methods which allowed kinships to be written down sequentially, starting with the founder individuals, who were assumed to be unrelated and non-inbred. He also developed a method which allows the inverse of the matrix to be written down in the same manner, without the need to first produce the relationship matrix or to invoke matrix inversion routines, which otherwise take a lot of computer power (Henderson 1976). Several computer programmes will compute the additive genetic relationship matrix and its inverse from the pedigree. I am most familiar with GenStat which provides the command VPEDIGREE to generate the inverse relationship matrix directly from the supplied pedigree. Its use on the simple example above does indeed give the same answer but beware. These procedures can only be applied to inbreeding crops if there is no consanguinity – no breeding loops – within the pedigree. In this case, the computed relationship matrix will be exactly ½ its true value. If there are breeding loops, that is to say if there are some lines derived from related parents, this is no longer the case. This is because the progeny of related lines are partially inbred and will have diagonal elements in the relationship matrix which are >1. Doubling the relationship matrix to make it applicable to a set of inbred lines would give diagonal elements >2 which implies a p(ibd) >1 for an inbred line. As p(ibd) is a probability, this should be impossible. I am uncertain what the consequences of this are, but it would be relatively easy to test with simple example pedigree structures.

More recently, molecular markers have been used to estimate kinships. This change has only been possible as high densities of informative markers become available. Marker based estimates of relationship have come to prominence for two reasons.

Firstly, no pedigree information may be available. This is particularly so for wild species. Marker based estimates of kinship and inbreeding coefficients now allow estimation of variance components and heritability of wild species measured in-situ. This has opened up new opportunities for research in population and ecological genetics studies. K Ritland is a pioneer of this approach (Ritland 2000). In domesticated species too, pedigree information is often incomplete, even for very recent sets of cultivars. Breeders often wish to keep this information secret, or the pedigree may involve proprietal lines whose origin is not public, or a line assumed to be derived from an F2 may in fact originate from a backcross.

Secondly, even when pedigrees are known perfectly, marker based estimates can be more accurate. The relationship between two full-sibs, ½, is an expected value: the mean over a large number of full-sib pairs. In practice, Mendelian sampling within a family results in any particular pair deviating a little from this mean. With sufficient markers, this deviation can be measured accurately, in which case it is better to use the estimate rather than the expected value. (There will be a Bayesian estimate which incorporates the prior, ½, with the estimate from the markers. The more markers, the lower the weight given to the prior.) This approach has been used in human genetics to estimate heritabilities within full-sib families by correlating covariance in phenotype with within family variation in relationships. This provides estimates of heritability from variation *within* families. These estimates are not confounded with common environment effects which can otherwise be a problem in human genetics. (For what it's worth, the new estimates agree well with the original estimates from twin studies and the like). This approach was developed by Peter Visscher (2009).

As far as I'm aware, the question about how many markers is sufficient to estimate kinship with accuracy adequate for use in mixed modelling has not been thoroughly studied for plants. It will be a function of the extent of LD within the population (the less extensive LD, the more markers are required), the true variability in relationships and the informativeness of the markers / diversity of the population studied (which amount to the same thing). One sees published work which relies only on tens of SSR markers. Whether this is adequate I don't know.

There are several methods to estimate kinship, and hence the additive genetic relationship matrix, from markers. This is an area of active research still. Discussion is usually couched in terms of the probability of identity by state of alleles, p(ibs), which is what you observe at a locus, and the probability of identity by descent, p(ibd), which is what you wish to infer. For association mapping, I suspect the result for the candidate markers themselves don't differ too much among methods but estimates of Vg can differ enormously.

I shall describe two simple methods. More information will be found in the documentation for the tutorial.

1. Simple allele sharing.

For each marker, compute the probability that a randomly drawn allele from one individual is shared (ie is ibs) with a randomly drawn allele from the other. For haploids and inbreeders such as wheat, this is 1 if the two alleles are identical and zero if not. This can be averaged over all loci to give the estimate of kinship. The estimate of kinship of an individual with itself is its inbreeding coefficient F. The additive relationship matrix should therefore have elements 2 times kinship on the off diagonal and 1+ kinship on the diagonal. For fully inbred lines, F will be correctly estimated as 1, and the raw kinship matrix can be used directly in the mixed model (with care over the interpretation of variance components).

For diploid outbreeding individuals, the estimate of F will be <1. Strictly, the diagonal elements should be set to 1+ F and the off-diagonals to 2 times kinship. Frequently, however, I strongly suspect that the kinship matrix is used directly as input, with diagonal elements set to 1.

For both inbreeders and outbreeders, off-diagonal elements are very unlikely to ever be ≤ 0 with this method. With this simple measure of allele sharing, all individuals usually end up explicitly related. The estimate of heritability and genetic variance will apply to some Arcadian population in which every individual carried unique alleles. Multiallelic markers such as SSRs will generally give lower kinships than biallelic markers and therefore also give lower estimates of Vg.

To overcome some of the limitations of simple allele sharing we can take a weighted average over markers such that the more informative markers have more influence. An SSR with a high number of alleles will be more informative than a SNP as ibs for the SSR markers is more likely to arise through relatedness than through random sampling. One such weighting scheme is:

$$\text{Kinship} = [\, p(\text{ibs}) - x \,] / (1-x)$$

where x is the average p(ibs) for two random alleles at that locus drawn from the population It is usually estimated using the average allele frequencies in the dataset. This isn't strictly correct but is judged good enough. For a biallelic marker with frequencies p and q (=1-p), x = 1-2pq giving kinship estimates of of:

$[\, 1 - (1-2pq) \,] / (2pq) = 1$ for identical homozygotes

and

$1 - 1/4pq$ with one allele in common

$1 - 1/2pq$ with no alleles in common

Kinship is averaged over all loci as before. This scheme has the effect that the estimated inbreeding coefficient of an individual is 1 only if all loci in that individual are homozygous. For an inbred species this gives a relationship matrix with all leading diagonal values of 1 and off-diagonal elements which can be negative. For a diploid, negative estimates of both kinship and F can arise. Diagonal elements of the matrix should be replaced with 1 + F and the off diagonals with twice the kinship.

Other methods of increasing sophistication are also in use, all of which attempt to translate p(ibs) to p(ibd) which is what we are really after, for example Lynch (1988) and Melchinger et al (1991). Several are available within the software SPAGEDi

2. Correlation and excess allele sharing.

This is the preferred method at NIAB

In any particular dataset, all additive traits have the same numerator relationship matrix: for a single locus, the contribution to the genetic variance of an individual is $(1+F)2pqa^2$ and to the covariance $2fpqa^2$. If there was an additive trait, of heritability one, with known genetic variance, and controlled by a known large number of loci of known allele frequency, distributed uniformly with respect to the genetic map, we could get an estimate the numerator relationship matrix from each locus and average over loci for greater accuracy. We do not have such an ideal dataset, but we can synthesize something quite similar from the marker data.

We'll assume our markers are all bi-allelic, though the process could be extended to SSRs. We start by giving each marker equal genetic variances. To do this we give genotypes scores of 0, 1 and 2 (diploids) or 0,1 (haploids and inbreeders) as usual. We then standardize these scores by subtracting the mean number of alleles carried by individuals in the dataset (2p for diploids, p for haploids) and dividing by the standard deviation of allele numbers. (This is $\sqrt{2}(p(1-p)$ for diploids and $\sqrt{pq}$ for haploids). Each marker score now has a mean of zero and a variance of 1 across all the individuals in the dataset.

The average score over markers gives us the synthesized trait for an individual. The contribution each marker makes to the genetic variance of that individual is just the square of its score. The contribution to the covariance between individuals is the cross product.
We are interested in the mean of these contributions per individual, or per pair of individuals, over all loci. However, first we standardize again, so that each marker contributes equally to this average. To do this, we convert the scores over individuals within markers to standardised normal deviates by subtraction of the mean and division by the standard error (estimated within markers across individuals this time). The mean of the square and the mean of the cross product of these new scores will give us the result we want. However, it is not necessary to do this last standardisation explicitly as a matrix of these mean squares and mean cross products is just the variance / covariance matrix for individuals across all the markers. The elements of this matrix are estimates of the

coefficients of the numerator relationship matrix multiplied by a constant. The constant is the genetic variance, Vm, of the synthesised additive trait. There are now several ways we can proceed:

1. If all individuals in the dataset are outbred, we can convert the variance/covariance matrix to a correlation matrix. This is equivalent to standardising the marker trait (again) to Vm = 1 and zero mean. The elements of the correlation matrix are therefore the desired values for the numerator relationship matrix and the correlation matrix can be used directly.

2. If all individuals in the dataset are fully inbred, we can again convert the variance/covariance matrix to a correlation matrix. This is equivalent to standardising the marker trait to give it unit variance (over unrelated inbreds) and zero mean. The elements of the correlation matrix are now all half the desired values for the numerator relationship matrix, as discussed previously. We can use this matrix directly, or double it if we wish.

3. The variance covariance matrix can be divided by an estimate of Vm to give the numerator relationship matrix. For example, there may be a set of individuals or lines which can be used to define a reference population with known kinships and inbreeding coefficients. Vm would be estimated directly from this dataset.

4. As Vm is a constant, we can use the variance covariance matrix directly. The value of Vm will make no difference to any test of association, but it will give misleading estimates of variance components.

Options 1, 2 and 3 could probably be interpreted in a Bayesian manner as different extremes of methods to incorporate prior knowledge of inbreeding into the estimates. There will therefore be a Bayesian estimation method which incorporates such prior knowledge with the observed data.

This is not the usual way of explaining this estimate of the numerator additive relationship matrix, but to my mind relating the procedure to an additive polygenic trait based on the markers makes some properties of the estimates more clear. It is more commonly described in terms of excess allele sharing. The excess is the deviation from the mean.

The concept of treating the markers as if they are component genes of a quantitative trait is also easily extended to other ploidy levels. It may not work so well for dominant markers. Again simple examples may be telling.

Role of allele frequencies

Markers are first adjusted by the mean and variance of allele frequencies. Inevitable, these are estimated from the data. As the individuals in the dataset can be quite inbred / related / selected, these estimates are not ideal. Data could first be standardised using means and variances estimated from a reference population if one was available.

Alternatively, once we have a relationship matrix, marker allele frequencies could be re-estimated taking these into account and the whole process iterated. For large datasets, neither of these processes is likely to make much difference.

Relationship of the mixed model with Structured Association (SA).

Structured association corrects for population structure by including covariates to account for sub-population membership in a multiple regression of the trait on the marker to be tested. As the covariates are fitted first, the association of the marker is with the trait residuals after adjustment for variation in population membership. The covariates to account for population membership come from analysis with the program STRUCTURE to detect and quantify cryptic population membership using genome-wide marker data. However, other covariates can equally be used – known population membership is the most obvious choice. STRUCTURE does not work well on many crop populations: it is difficult to decide how many cryptic subpopulations exist, and much of the population structure is due to close kinship rather than the more gross subdivisions which STRUCTURE can detect. Yu et al (2006), in their initial advocacy of the mixed model in association mapping, proposed using both the population membership vectors estimated by STRUCTURE (the Q matrix) together with a marker-estimated kinship matrix (the K matrix). Q and K are commonly estimated using the same set of markers. Even if estimated using different markers, given that both sets ideally sample the whole genome, the data are correlated. As a result, there is a concern that use of both Q and K amounts to double counting the data.

The documentation for TASSEL suggests that one should set any negative estimates of the kinship matrix to zero. These negative relationships are more likely to be between pairs from different sub-population: the least related individuals. Replacing negative kinships with values of zero in K, but including the Q matrix should therefore have the effect of reducing the genetic distance between the most distantly related pairs to the population average (0) in K, but then accounting for large genetic distances by the allocation of individuals to different subpopulations in Q. This should give much the same answer as ignoring Q and analysing the data with K, including negative values. In UK wheat and barley association mapping studies this does seem to be the case (JonWhite, pers, comm.).

In inbreeding crops, structured association alone leaves a very high rate of false positive results, though still substantially less than analysis with no adjustment, There is an emerging view that analysing crop data with K alone (without tampering with negative values) is as good as anything.

Relationship of the mixed model with EIGENSTRAT

EIGENSTRAT (Price et al, 2006) works on an excess allele sharing kinship matrix, created in the same manner described above. The eigenvectors associated with the largest eigenvalues are then used as covariates in a regression on phenotype. In principle, candidate loci could be included directly in this regression to give a procedure very similar to SA; with eigenvectors used as covariates rather than vectors from STRUCTURE. For gross population divisions, which are generally detected in the first two or three eigenvectors, such an approach should give very similar results to SA. However, EIGENSTRAT correlates the residuals from the regression of phenotype on eigenvectors with a similar vector of residuals from a regression of the candidate marker on the same eigenvectors. The correlation between the two sets of residuals is then the tested for statistical significance. It isn't clear why both the phenotype and candidate marker are adjusted in this manner. In UK wheat and barley EIGENSTRAT behaves similarly to SA: less effective in controlling false positives than the mixed model but better than nothing (John White, pers. comm.). As with STRUCTURE, EIGENSTRAT is not good at controlling for false positives generated by close relationships among individuals.

A proposed combined approach

This is more speculative. The use of the kinship matrix in EIGENSTRAT suggests the following approach. The largest eigenvalues are selected for inclusion as covariates in a regression with phenotype, just as for structured association. However, the residual eigenvalues and eigenvectors can be used to construct a residual kinship matrix. This represents the residual genetic relationships not accounted for by the largest eigenvalues. This residual kinship matrix is used instead of the full kinship matrix to account for residual genetic variances/covariances in the mixed model. This partitioning ensures that data are only used once but that major population subdivisions are adjusted by treating them as fixed effects in a regression (which we have seen can be very effective) but residual kinship effects are still accounted for as random effects through the relationship matrix.

The number of eigenvalues to be included as covariates could be varied. If none are used, we have the simple mixed model. As the number increases, the analysis approaches that of EIGENSTRAT. The optimum number is likely to be low – most major population groups are apparent in the first two or three axes of a PCO plot. Additional eigenvalues are unlikely to account for much phenotypic variation in a regression analysis, and are better left to account for covariation within the residual kinship matrix. For any fixed number of eigenvalues removed, significance of the regression of each eigenvector can be tested. Thus it may be possible to remove eigenvectors sequentially, until one is removed which does not account for a significant proportion of the phenotypic variation. The non-significant eigenvector could then be added back into the kinship matrix to give a final model. One would expect the eigenvectors of the largest eigenvalues to be most important in accounting for phenotype, but this needn't be so. For example, in a dataset

consisting of Spring and Winter barley, the largest eigenvector will be associated with the spring-winter partition. However, not all phenotypes are associated with this split, but there may be some other gross subpopulation partition which is. Two row versus six row barley for example. So it could be that the largest eigenvalue should be left in the residual kinship matrix – we still need to account for it contribution to (co)variance - but the second principle component (say) should be included as a covariate. Some form of stepwise regression procedure might assist in selection. This is something we would like to study in the future.

Genomic control

Genomic control (GC) was the first statistical method to be used extensively in association mapping. In crops, apart from at NIAB, it is hardly ever used. Although it was introduced to account for differences between subpopulations, the assumptions on which it is based are more valid for controlling for close relationships between relatives. In this respect it is can be regarded as a poor man's substitute for the kinship matrix. However, it acts as a gross average adjustment – the test statistic for association is adjusted downwards by a constant proportion in all cases. There will be a form of the mixed model which will give an identical result to GC. It will have an error variance and a fixed value for γ to adjust for the inflation in variance due to cryptic population structure. For a collection of inbreds, I believe a mixed model equivalent to GC could be constructed as follows. The effect of subgroups is defined using the control SNPs. Each control SNP partitions variance into a between SNP term and an error term. Each SNP therefore gives an estimate of γ. The mean of these is used in the analysis of the candidate SNP as a fixed ratio of γ. With γ fixed, the relationship matrix in this analysis will have elements of 1 down the diagonal, and off-diagonal elements of zero if lines carry different alleles and a constant >0 if they carry the same allele. (This needs confirming.)

The mixed model is clearly more flexible as it accounts not just for average relationships, but for variation in relationships among pairs of individuals. As a result, depending on the distribution of a candidate marker among individuals, the test statistic can, in principle, rise as well as fall for the mixed model. Nevertheless, we have found genomic control to work well when used in conjunction with EIGENSTRAT or with SA, when it acts to mop-up the residual kinship effects not accounted for directly. GC with EIGENSTRAT, in particular, is therefore a very quick method of analysing a large dataset and can act as a sanity-check for more elaborate analyses.

More complex cases

We have only discussed data in which a single measure of a phenotype is analysed. More complex analyses which incorporate replicate observations on each line and even on multiple correlated phenotypes are possible, but we shall not go into those here. Multiple candidate markers may themselves be treated as random effects rather than fixed. As far as I'm aware, bespoke programs for the mixed model in association genetics, such as

Tassel and Emma do not provide these options. The mixed model for these more complex cases needs to be fitted in standard statistical software such as ASREML, GenStat, or SAS.

Currently most analyses tend to be carried out on line means, derived from an initial analysis of the data. The big advantage of replicated data is that is can provide an independent estimate of error variance. This improves the accuracy of the subsequent estimate of genetic variance. Statistical packages which implement the mixed model should allow you to fix the values of some variance components (or their ratios) in advance of the analysis. (Remember to scale by the number of replicates.) I know GenStat allows this very generally, and Tassel allows heritabilities to be fixed.

In conclusion.

Use the mixed model.

In our experience, which is largely on inbred cereal crops, this offers the best compromise between power to detect loci and control of false positive results.

Estimate your kinship matrix by excess allele sharing.

For more information on software implementation, see Zhang et al (2009), though this is somewhat biased towards Tassel, not surprisingly given the authorship.

References

Abecasis GR, Cardon LR, CooksonWO (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279-292

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**:69–93

Henderson, C. R. *1963. Selection* index and expected genetic advance. *In* W. D. Hanson and H. F. Robinson (Ed.). Statistical Genetics and Plant Breeding. National Academy of Sciences-National Research Council, Washington, DC, Pub. 982

Falush, D. et al. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587

Lynch, M., (1988) Estimation of relatedness by DNA Fingerprinting. *Mol. Biol. Evol.* **5**: 584-599.

Lynch, M. and Ritland, K. (1999). Estimation of Pairwise Relatedness With Molecular Markers. *Genetics* **152**:1753–1766.

Melchinger, A. E., M. M. Messmer, M. Lee, W. L. Woodman, and K. R. Lamkey, (1991) Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. *Crop Sci.* **31**: 669-678.

Piepho HP, Möhring J, Melchinger AE, Büchse A (2007) BLUP for phenotypic selection in plant breeding and variety testing *Euphytica* **161**:209–228

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909

Pritchard, J.K. et al. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959

Pritchard, J.K. et al. (2000) Association mapping in structured populations. *Am. J. Hum. Genet*. **67**, 170–181

Stich B, Melchinger AE, Piepho H-P, Heckenberger M, Maurer HP, Reif JC (2006). *Theor. Appl. Genet*. 113:1121–1130).

Visscher PM (2009) Whole genome approaches to quantitative genetics. *Genetica* **136**:351-358

Yu JM, Pressoir MG, Briggs WH, Bi IV, Yamasaki MG. *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**:203-208.

Zhang Z, Buckler ES, Casstevens TM, Bradbury PJ (2009) Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics* **10**:664-675

THE ROLE OF MOLECULAR MARKERS IN PRACTICAL PLANT BREEDING SCHEMES. MARKER ASSISTED SELECTION IN PRACTICE.

Introduction

"QTL analysis has produced great advances in plant breeding" recent review.

This is a quote from a 2007 review of QTL analysis in rice. It is a bit unfair to pick on this particular paper: many others would do to illustrate this supremely optimistic assessment of the value of QTL mapping to plant breeding. It has been the justification for many a grant application too: the deliverables from the proposed research are arrogantly stated to be a set of markers suitable for use in marker assisted selection by "the breeders." No wonder this gets up their noses. I'm a glass-half-empty sort of person, so in this presentation I shall go out of my way to point out the problems with application of molecular markers to breeding programmes. There is no limit to the amount of freely available literature which will point out the advantages. In passing I shall mention some of them too.

I should also be explicit that my own experience of applying marker based methods to breeding programmes is zero. You must form your own opinion of their merit. The best people to talk to would be controllers of breeding programmes in which markers are routinely used. Unfortunately, most such individuals are in the private sector and like to keep this sector private.

Given these views, what was the point of learning all about markers and QTLs? They do have a place and are used successfully in some cases. More importantly, things are getting better. In particular the costs of genotyping are getting cheaper and the availability of high densities of markers and of sequence data (for which, after all, markers are a mere surrogate) is increasing. Meanwhile, phenotyping costs increase. So the direct application of molecular markers and marker assisted selection to practical breeding programmes will increase.

Marker assisted selection and the breeders equation.

$$R=ih\sigma_g$$

Like everything else in breeding, the role of marker assisted selection (MAS) must be judged by its impact on this equation. If MAS cannot increase response to selection per year or per unit of cost, then it should not be part of the breeding programme. Looking at each of the terms in the equation in turn (in reverse order):

Genetic variance

MAS can have no effect on σ_g^2. Identifying a QTL, however large, does not alter the value of σ_g value. You haven't identified anything new. If phenotypic selection is already efficient in burning this fuel to drive genetic gain MAS will do no better.

Heritability

A consideration of the effect of heritability can be framed in the context of indirect selection or index selection. If the heritability of the phenotype is low, then selection on markers closely linked to a number of QTL can increase response to selection. We can be explicit about the conditions in which MAS will be more successful than phenotypic selection:

select on phenotype alone $R = ih_p^2\sigma_p$

select on markers alone $R = ir_g h_m h_p \sigma_p$

where the subscripts m and p stand for marker and phenotype and r_g is the genetic correlation between the index of a score based on markers and the phenotype. These equations are quite general. All inaccuracies in genotyping are accounted for by h_m and imprecision in prediction of phenotype by r_g.

For MAS to give a greater response than phenotypic selection

$$ir_g h_m h_p \sigma_p > ih_p^2\sigma_p$$

or

$$r_g h_m > h_p$$

but since $h_m^2 = 1$ (assuming no genotype errors)

$$r_g > h_p$$

This is better expressed as

$$r_g^2 > h_p^2$$

The genetic correlation coefficient squared between marker index and genotype must be higher that the heritability of the phenotype.

For a suitably defined marker index, weighting individual markers by the magnitude of their QTL effect, if *all* QTL are tagged perfectly and if QTL effects are known without error, then r_g^2 will be one and MAS will be impossible to beat. However, if only QTL of large effect are tagged and these only account for 50% of the genetic variation, say, then

MAS will have its work cut out: a heritability of 50% is not difficult to achieve which some modest replication.

This simple analysis shows that the argument in favour of MAS is not cut and dried. There are additional problems too. Before MAS is considered we need to detect marker-trait associations through linkage or LD mapping. QTL of large effect can be detected easily with little bias in their estimated effect and with high precision in their chromosome location. However, for traits of low heritability and/or for QTL of minor effect, detection and estimation is harder. Those minor QTL which are fortunate enough to be detected will necessarily have estimated effects which are biased upwards: from many minor QTL only those which, through sampling variation, appear to have a large effect in the mapping experiment will be detected. This is sometimes described as the Beavis effect after one of its discoverers. It reduces the efficiency of MAS by reducing r_g, the genetic correlation between marker score and the trait phenotype..

Index selection should give an improvement over selection purely on markers or purely on phenotype. Unfortunately it can still be let down by the bias and precision with which marker-QTL effects are estimated. For example, Bernardo, *Crop Sci* 2001, 41:1-4 simulated MAS in hybrid maize breeding in which markers tagging *all* QTL were included in the index but in which QTL effects were estimated from the phenotypic data. Because of the bias in the estimation of these effects, index selection frequently performed more poorly than phenotypic selection. Increasing the population size from which marker effects were estimated and increasing heritability improved the efficiency of MAS but in these circumstances phenotypic selection is more effective too. He suggested, at least for hybrid crops, that QTL mapping may be better restricted to identifying genes for which novel variation could be screened or induced. He also stated that MAS may be better suited to animal breeding or to inbred crops. Subsequent work by Bernardo and Charcosset has shown that for a trait under the control of 40 or more loci, it is often best to ignore minor QTL, even when their location is known exactly.

The idea of a marker index, or "molecular score" to be incorporated with phenotypic selection was first developed in a benchmark paper by Lande and Thompson (1990). Rather than mapping QTL in an independent experiment, their work assumed that marker associations were detected by regression in the population under selection. The prediction of genotype by linear regression on marker score is then incorporated into the selection process. This approach has the advantage that there is no requirement to map the QTL. Index selection gave the greatest improvement over phenotypic selection at low heritability, but large population sizes were needed to detect the marker trait associations - which partly defeats the point of MAS in the first place. Various ways of addressing the problem of bias in assessing QTL effect were considered in this and subsequent work, for example using half the population to detect QTL and the other half to estimate their effect. A related difficulty is deciding how many makers to include. One could select only those which achieve some predetermined level of significance for example, or one could choose to include them all. There are also more complicated alternatives.

The Lande and Thompson approach, as far as I'm aware, hasn't caught on in crops. The most likely use would be in outbreeding species. There is a requirement for the population to be closed - no population substructure - otherwise we would most likely end up selecting for a particular population type. There is also a requirement for LD to be sufficiently extensive for marker-trait correlations to be detected. None the less the method has a role, in that it places MAS in a quantitative genetics framework which then provides an objective means for its evaluation.

In animal breeding a more recent development following the same approach is "genomic selection" (Meuwissen *et al*. 2001 Genetics **157**:1819-1829). This looks ahead to the use of high density markers to cover the whole genome, selecting on a score accumulated over marker intervals of roughly 1 cM. This is now being discussed in plant breeding too. (Bernardo & Yu 2007 Crop Science **47**:514-621, Zhong et al 2009 Genetics **182**:355-364; Piepho 2009 Crop Sci **49**:1165-1176). This is an area of active research and shows a lot of promise. Genomic selection is now being exploited commercially by some animal breeding companies. It is discussed in more detail in the next chapter.

Much of the effectiveness of MAS for quantitative traits depends on the distribution of gene effects. The general view is that the distribution will be exponential, with most QTL having minor effect but a small number having a large effect. This is difficult to confirm experimentally because of the problems of bias in estimating the minor effects but is in line with predictions from evolutionary and population genetics theory. There is also the related question of the distribution of allele frequencies. This is thought most likely to be exponential too, in which case finding marker associations with the rare variants is going to be a problem.

Intensity of selection

If markers allow the screening of single plants for traits that would otherwise require multiple trial plots to measure, then the use of markers is likely to be cost effective: large populations can be screened and intensity of selection can be high. However, if the trait itself can be measured directly on single plants, phenotyping may be cheaper. This can only be decided on a case by case basis. For traits like yield, the problem again falls back on the difficulty of establishing unbiased validated marker trait associations which are close enough to use in MAS. If sufficient of these are found, then higher intensities of selection are possible because more single plants can be grown than plots. For one or two major genes, however, intensity of selection isn't a problem: whether selecting in plants or plots, you don't need to grow many to recover what you want. A two stage process in which single plants are first screened followed by selection among replicated plots could be advantageous, provided the cycle time of the selection scheme is not increased. This could work for crops bred through pedigree breeding, SSD or DH production, with optimum allocation of resources between MAS and phenotypic selection established using approaches similar to those discussed earlier for multi-stage selection.

In general, increasing intensity of selection is not particularly cost effective provided the population available for selection is reasonably large. Selecting the best line from 10,000

gives only 1.2 times the response to selecting the best line from 1000 and 1.5 times the response to selecting the best from 100.

Breeders equation summary

To summarise so far. For traits under the control of many genes, MAS will not substitute for phenotypic selection but can complement it, most usefully if heritabilities are low. However, in such circumstances precise location of QTL and accurate estimation of effect are hard to establish. Overestimates of effect and poor location can result in MAS reducing response rather than increasing it. For MAS to become routine in plant breeding for polygenic traits, estimates of QTL location and effect need improving.

Major genes, time and money

Major genes are often mapped to small intervals and functional markers within genes are sometimes available. Examples in wheat include markers for dwarfing genes, some disease resistances, and flowering time response to day length. In such cases there are clear opportunities for molecular markers to increase response to selection per unit cost. It may be easier to screen markers for resistance than to grow and artificially inoculate plants. Clearly in these cases MAS can be beneficial.

MAS can also save time. The area where this is most easily achieved is in speeding up backcrossing, and we shall consider this in some detail shortly. Markers can also reduce cycle time however, especially for traits in which selection cannot take place until after sexual maturity. Examples of these are yield in most perennial species or, for that matter, any trait which cannot be effectively measured on a single plant (because we need additional generations to bulk up seed for testing). We have already seen in the tutorials that reducing cycle time is hugely effective in increasing response to selection compared to increasing population size. Any prediction of performance before sexual maturity will therefore allow crosses for the next cycle of selection to be made early and increase response per year. Selection schemes based purely on phenotypic selection can also be devised to reduce cycle time, but usually at considerable expense. An extreme case is to produce the next generation without any selection but then to apply selection retrospectively. For example, a simple recurrent selection scheme might select 10 individuals from 45, then intermate those 10 in a ½ diallel (45 crosses) to form the next generation. Suppose crosses can be made after one year, but phenotype information is not available for two years. If crossing is delayed until after selection, this scheme takes two years per cycle of selection. However, if all 45 individuals are mated in a ½ diallel (990 crosses) at the end of the first year and grown on, then once phenotype information is available in the parental generation, the crosses among the 10 selected parents are already available too, as a subset, size 45, of the 990. This is hugely extravagant in seed production, but for a halving of cycle time it may be worthwhile. In schemes like this, MAS could be used to reduce the number of crosses to be made, but final selection could be postponed until phenotype information was available. Once again, the optimum approach is a combination of MAS and phenotyping.

Marker assisted backcrossing

A good review from the guru of MAB: HOSPITAL, F. (2003) Marker-assisted breeding. In H.J. Newbury (ed.) Plant Molecular Breeding. Blackwell Scientific Publishers, London, UK, pp30-56 http://fhospital.free.fr/fred/work/publications.html

The objective of most backcrossing programmes is to introgress one or more loci from the recurrent to the non-recurrent parent, but otherwise to recover the genome of the non-recurrent parent. Selection on markers can help in two ways.

1) Select for markers linked to the trait to be introgressed rather than for the trait itself. This is termed foreground selection. It can be cheaper, and in addition individuals heterozygous for the introgressed locus can be identified for the next cycle of backcrossing. For recessive traits, there is a 50% chance that any backcross plant will not be carrying the desired locus. Without markers, several crosses with different individuals must be made to ensure that at least one is carrying the locus. With markers, probable heterozygotes at the trait locus can be selected.

2) Aside from the chromosome region around the trait locus, we can select for homozygosity of marker alleles from the recurrent parent to speed up recovery of the recurrent parent genome. This is termed background selection.

Foreground selection

Selecting on a single marker rather than the phenotype, in each backcross generation we want to know the probability that the locus to be introgressed is still present. Hospital defines this probability as the "target control rate." For a single generation, it is just $(1-\theta)$ where θ is the recombination fraction between marker and trait locus. Over n generations of backcrossing it is:

$$TCR = (1-\theta)^n$$

Of course, this probability is 1 if the marker is completely linked to the trait but is only 0.81 after four generations of backcrossing for a marker at $\theta = 0.05$. Not unexpectedly then, for MAB with a single marker, we require quite close linkage. To increase the TCR, we can use flanking markers, one on each side of the trait locus, and select for heterozygotes at both flanking markers.

With recombination fractions θ_1 and θ_2 between each marker and the QTL and θ between the markers themselves, we select the carriers of both markers. These occur at frequency $(1-\theta)$. But these individuals include a proportion which are non recombinant $(1-\theta_1)(1-\theta_2)$

(assuming no interference) and a proportion of double recombinants which no longer carry the QTL. So the probability of carrying the QTL when we select on both markers is

$$TCR = [(1-\theta_1)(1-\theta_2) /(1- \theta)]^n$$

For example, with $\theta_1 = \theta_2 = 0.05$, then $\theta = 0.095$ and TCR = 0.99 after four generations, a considerable improvement. However, if your flanking markers are some distance apart, you will be forcing the introgression of quite a length of chromosome around the trait locus.

These probabilities can also be used to calculate the sample size required at each stage of backcrossing to reach a specified probability of the trait locus still being present at any generation of backcrossing. If the non-recurrent parent carries the desired haplotype M_1QM_2 , then the probability that an individual in the next generation of backcrossing also carries it is

$$½(1-\theta_1)(1-\theta_2)$$

(the ½ is because the individual might have inherited the recurrent parent haplotype.

So in a sample of size n, the probability that there is no individual carrying the desired haplotype is

$$r = [1-½(1-\theta_1)(1-\theta_2)]^n$$

The sample sizes are quite small for quite high probabilities that at least one individual does carry the desired haplotype, even with quite loose linkage.

On taking logs, this equation can be solved to calculate n for a given risk, r. We need to be confident that the QTL really is where we think it is. The formulae above assumes that the location of the QTL is known exactly. If we have a pdf for the location, the simple approach above can be extended by integrating over the whole interval, or even over the whole chromosome. This can be useful for considering introgression of a locus identified in a mapping experiment, where locations are typically not known precisely and there is a chance that the locus lies outside the selected flanking markers.

Background selection

In the absence of selection, or for chromosomes not carrying loci for selection, the proportion of the non-recurrent parent is $½^n$ where n is the number of generations of backcrossing plus 1 (the F1 is already 50% non-recurrent parent). This is an average of course. In practice, the introgressed genome will consist of segments of chromosome of variable length which decrease over successive generations of backcrossing.

Background selection is harder to evaluate, and computer simulations are often used. I can give some general conclusions. For the typical (100cM) chromosome, 2-4 markers

are sufficient. They should be evenly spaced but the first and last markers are not optimally placed at the ends of the chromosomes. No chromosome should be left unmarked. Background selection saves about two generations compared to backcrossing with no selection. So the BC4 generation with markers is equivalent to BC6 without. Sample sizes required for this are quite small. If resources are limiting, it is better to save MAB until the final generation.

On the chromosome carrying the trait to be introgressed, background selection is more complicated. The background selection is to reduce "linkage drag." That is it aims to reduce the length of the introgressed chromosome segment which contains the trait. A simple and crude treatment of selection against linkage drag is to reverse the argument for foreground selection given earlier. Suppose we have a marker within the QTL, or equivalently that we are screening for the QTL by phenotype. We have the same two flanking markers as before, but now we want to select for carriers of the haplotype, namely m_1Qm_2 .

Among backcrossed individuals homozygous for m_1 and m_2 , a proportion

$$\theta_1\theta_2 /(1- \theta)$$

will be carrying the QTL.

The probability of not finding such an individual in a population of size n is

$$[1-\theta_1\theta_2) /(1- \theta)]^n$$

which can again be solved for n on taking logs.

This is exactly the case for introgression of a transgene where we ought to have a marker for the transgene itself.

If we do not have a perfect marker for the trait, we can work with four flanking markers, two on each side of the trait locus. At the outer pair we select for the recurrent parent and at the inner pair for the non-recurrent parent. The free software, *popmin* is specifically written for this task. It searches for the minimum population size at each stage of the backcrossing crosses to achieve a desired probability of success. We shall have a look at it in the tutorial.

HOSPITAL, F., DECOUX, G. (2002) Popmin: a program for the numerical optimization of population sizes in marker-assisted backcross programs. *J. Hered*, **93**: 383-384.

Note again, that background selection also requires the location of the QTL to be known with some precision. In the worst case, if flanking markers were selected around a linkage peak which was actually a ghost resulting from two adjacent QTL linked in

coupling, then attempts at introgression, with accompanying selection for background makers could end up selecting against the QTL.

Some miscellaneous uses of MAS

Pyramiding genes

If there are several major loci conferring resistance to the same disease, it can be very difficult (but not impossible) to fix all genes in a single line using phenotypic selection alone. Selection using markers makes this process much easier. The expectation is that once fixed, the resistance is less likely to break down since the pathogen must mutate at multiple loci to overcome it.

Non-random mating

The statement made earlier that the availability of markers cannot alter the genetic variance is strictly only true for a random mating population in the absence of selection. We have already seen that genetic variance is altered by selection and by inbreeding. Assortative mating also affects genetic variation. If individuals are not mated at random, but are paired like with like, then the genetic variance in the next generation is increased. The increase is not great however. Details are in F&M (beware there is a misprint in table 10.6). For example, with an initial heritability of 0.5, and perfect correlation between the male and female phenotype, the heritability is 0.56 in the next generation.

Molecular markers provide an opportunity to exploit non-random mating in selecting parents. There are two alternatives. One is to select parents which are as diverse as possible, where diversity is measured by the markers. No direct correlation between markers and trait is assumed. In this case, phenotypic selection must still take place, since otherwise you end up crossing the best with the worst, with a predictably disappointing outcome. The second is to use known associations of markers with phenotype to build a genotype in the absence of any phenotype data.

The second strategy is the more interesting, and has been employed successfully in wheat breeding programmes in Australia. The process will take several generations. For n QTLs segregating in an F2, the probability of finding a single individual homozygous at all loci is $\frac{1}{2}^n$ which is less than one in a thousand for only ten loci. To guarantee that all 10 loci were fixed would require an even larger population size. Also it is unlikely that all ten loci would be segregating in the desired cross. So although single QTL can be very easily or rapidly fixed in a population, larger numbers require impossibly large population sizes. If QTL are linked in dispersion, then the numbers go up even more.

The first strategy is problematic. There is an assumption that marker diversity correlates with diversity for loci determining the traits of interest. This is plausible, but not necessarily correct. Marker diversity arises predominantly from drift and founder effects but variation at trait loci will be more influenced by selection. There need not be any

great correlation between the two. Even if exactly the same forces of drift and selection have acted on all loci, it is possible that the correlation between the 50 (say) loci determining most of the trait variation and a sample of 200 (say) markers with a genome wide distribution will still not be high. Moreover, even if the premise is true, as stated earlier, maximum diversity is likely to be between very divergent parents which are not adapted : Vg would rise but the mean could tumble down. Phenotypic data must still be included and the correct weighting of the two sources of information is required.

This approach has been pursued with more success in predicting which parents to cross to produce hybrid varieties. In this case, the problem of how to include phenotype information is reduced, since one is relying on dominance variation to cover the sins of the parents. In maize and some other hybrid crops, the best hybrids generally come from crosses between lines drawn from known, different, subpopulations (heterotic groups). Molecular markers are very successful in assigning lines to populations and markers can be used to select lines from different subpopulations as hybrid parents. But if we already know the origins of the lines this has achieved nothing. As far as I'm aware, markers have had little success in routinely predicting good parental combinations within populations. Research continues. There is now some interest in using linkage analysis to detect heterotic QTL, which seems to me to be something of an admission of failure.

The advantage and promise of association genetics

Bias and precision are the two big problems in moving from QTL mapping experiments in experimental populations to MAS in breeders' germplasm. Association genetics has the potential to reduce both these. A population of cultivars offers a readily available replication set in which to get an unbiased assessment of a QTL effect and in which to improve the precision of its chromosome location. Moreover, by working in elite cultivars, results are more likely to be of immediate relevance to breeders since the collection should contain lines which are already present in their crossing schemes.

It must be remembered however, that association genetics panels require LD to decay sufficiently slowly for mapping to be possible with the available marker density. As a consequence, precision may not be as high as hoped for, or if the precision is available power may be lost because marker density is too low. The population of elite lines available, at least in the public domain, is restricted also, especially in minor crops. For example, in UK winter wheat we have only managed to collect 175 modern cultivars, about the same as for a typical biparental mapping population, though over France, Germany and the UK we have collected about 700.

For these reasons, I think it is an error to place too heavy a reliance on association genetics. My view is that the development of diverse mapping populations, specifically for fine mapping and replication of linkages established elsewhere, has a major role to play here and will be even more important in the future. Approaches such as the advanced intercross and nested association mapping will become more important too.

A further advantage of these fine mapping techniques (including association genetics) is that they should also allow better resolution of closely linked QTL. In most mapping experiments QTL linked in repulsion can cancel each other out while linkages in association can result in an apparent single QTL of large effect. At best these scenarios will merely reduce response to selection but they could result in MAS doing more harm that good. In addition, a region tagged by a maker or flanking markers may contain variants affecting other traits and the locus itself may be pleiotropic. Fine mapping in diverse populations will increase the chance that these problems are avoided or detected. I suspect they have been underestimated as risks in the application of MAS; they don't seem to get talked about much. Perhaps as new technologies for mapping are applied the extent of these problems will become apparent.

An example of the potential for these problems to arise is given by the study of the effect of gene density in Drosophila on mapping. Because the DNA sequence is known, simulations of mapping experiments can randomly allocate QTL in proportion to gene density; a more realistic approach than allocating QTL at random over the genetic map. Regions of low recombination are likely to carry the strongest apparent QTL as a result of multiple independent QTL clustering in these regions. Simulations show that as a result of clustering of genes and known variation in recombination frequency, one detects apparently few QTL of large effect. I am not aware of similar studies in crops yet they would be possible in rice now. The effect will be more important in species with low chromosome numbers. It will also be most prevalent in studies between extreme crosses. These are more likely to be segregating for multiple QTL so the opportunity for clustering and the emergence of QTL of apparent large effect is greater. A way to avoid these problems may be to use sequence information to provide gene density estimates, take these into account when testing, and to map in populations with rapid decay of LD.

Novel crops and the importance of maintaining phenotyping

The enthusiasm for MAS in minor and novel crops is no lower than in the major crops although one often finds that basic genetic questions remain unanswered: does the trait show any genetic variation? What is the mating system? It is important to master the basic game before attempting drop shots. Investment in molecular methods should not come until an effective system of phenotypic selection has been put in place and there is some idea about how best to breed the crop.

Revision of estimates

The Lande and Thompson approach was explicit about the requirement to regularly re-evaluate marker trait correlations. Other approaches have followed this lead. That this is required for single marker-trait combinations identified in the population being selected is no surprise. However, there is also concern that QTL tagged through mapping experiments, even with flanking markers, can seemingly lose their effect quite quickly; before the QTL or flanking markers are fixed. This effect has been seen in some recurrent selection experiments. Explanations include inaccurate mapping (the QTL could lie outside the flanking markers), ghost QTL (same effect), the Beavis effect, GxE, epistasis

(changing the genetic background as a result of selection alters the effect of the QTL) and changing allele frequencies. Whatever the causes, the effect demonstrates further practical difficulties of implementing MAS for complex traits in practice.

GENOMIC SELECTION

The lecture notes of Dr Ben Hayes provide an excellent source of free information. Search on the web for the latest version.

The basic idea is that all markers or marker intervals are included simultaneously in a model to predict genetic merit or breeding value, commonly referred to as GEBV; genomic estimated breeding value. Because there is no selection of a subset of markers which are significant, there is no bias: no Beavis effect. However, there are generally more markers than genotyped individuals available on which to estimate the marker effects. As a result there are too fewer degrees of freedom available to fit the full model using standard regression methods. Below I outline two approaches (the easiest) to estimating breeding value for genomic selection. We shall also study these in the tutorial. I then describe some of the factors which I feel are important in considering if GS is right for your crop. As usual, I'll be miserable.

Ridge Regression

This was first proposed as a method by Whittaker et al (2000, *Genet. Res*. **75**249-252), in the context of the Lande and Thomson approach to marker assisted selection, as a means of avoiding the problem of marker selection. In effect this proposed genomic selection in all but name, prior to Meuwissen *et al*. in 2001.

In matrix form, to assess marker effects, ordinary least squares regression solves the equation

$$\mathbf{Y} = \mathbf{Xb}$$

as

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$\mathbf{b} = [\ b_0\ b_1\ b_2 \ldots b_n\ ]$ is a vector of fixed marker effects with b_0 the mean and $b_1 \ldots b_n$ the effects for each marker.
Y is a vector of phenotypes.
X is the design matrix for markers and assigns alleles at each locus to the individual phenotypes in **Y**.

Once the marker regression coefficients, **b,** are estimated in the initial generation, on a set of phenotyped and genotyped individuals, these can be used to predict the breeding value of any genotyped individual in successive generations. Selection than then proceed over several generations solely on these marker based predictions.

However, as there are usually more columns in **X** than there are rows in **Y**, there are insufficient degrees of freedom to fit all markers simultaneously.

Ridge regression modifies the ordinary least squares estimates as

$\mathbf{b} = (\mathbf{X}'\mathbf{X}+\mathbf{I}\lambda)^{-1}\mathbf{X}'\mathbf{Y}$

I is a unit matrix (all 1's down the leading diagonal) with the same dimensions as **X**.

λ is a positive number which acts to shrink the estimates of elements of **b** back towards zero. If λ is zero, the ridge regression elements reduce to the ordinary least squares solution (which will fail if there are too few df). As λ gets larger, the estimates of **b** move towards zero. The addition of the penalty term **I**λ to **X'X** allows estimates of b to be made for all markers simultaneously. It is necessary to find a suitable value for λ. One suggestion is to use Ve/Vm where Ve is the error variation of the trait and Vm is genetic variation associated with each marker in **b**. With a total trait heritability of 50%, and markers assumed to account equally for the total genetic variation, this gives λ = number of markers. This seems a reasonable place to start. In practice, accuracy changes very little with λ (see the tutorial).

With λ set to Ve/Vm, ridge regression is equivalent to BLUP of the marker effects themselves (as opposed to BLUP of the breeding values of the individuals). The markers are in effect treated as random effects, drawn from a normal distribution with a variance of Vm.

Because of the shrinkage towards zero, ridge regression is usually carried out after **Y** is first adjusted to a mean of zero: it doesn't usually make sense to want to shrink the mean towards zero too. Also, the columns of **X** are often standardized to zero mean and unit variance, though in my hands this seems to have little effect. A suitable standardisation, which also substitutes appropriate values for missing data, is to standardise the marker data exactly as for excess allele sharing estimation of the relationship matrix, as described in the preceding chapter. This approach has the advantage of "filling in" missing marker genotypes which otherwise cause problems.

This regression procedure is easy to code in R and can give good predictions within generations on experimental data (see class exercise).

BLUP

Best linear unbiased prediction of breeding values has been the bread and butter of animal breeding for decades; relying on the known pedigree relationships among animals to form the numerator relationship matrix. As discussed in the previous chapter, marker based estimates of kinship can be more accurate than those from pedigree and these may therefore improve the accuracy of breeding value estimation. In plants, there has been little use of BLUP to estimate breeding values and hardly any in inbreeding crops like wheat. This is partly because the pedigrees are unknown or inaccurate, and partly because direct phenotypic assessment of breeding value is sufficiently accurate that little improvement is possible though inclusion of information from relatives. Genomic

selection could change this, since if will give a more rapid cycle of selection for traits like yield than is possible through phenotyping. In addition, the marker based assessment of kinship opens up opportunities for the incorporation of genetic relationships in estimating breeding value which are not possible from pedigree alone. For example, marker based estimates among lines within a cross can be used to select between those lines. This is not possible with pedigree based estimates; within a cross all lines would be are treated as equally related.

In practice, estimation of BLUPs using marker data follows the same procedure outlined in the last chapter for association mapping under the mixed model. Here, however, we are no longer interested in the estimation of the fixed effects for one or a few markers, but of the breeding values of the individuals or lines. These are random effects. With standard software such as GenStat, we include the additive relationship matrix in the model exactly as for association mapping, but request that random effects are reported. We shall have a go in the tutorial. The model we are fitting is:

$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{e}$

Y = phenotype data
β = the fixed effects. These could just be the mean or could include known major genes, reps etc.
X = design matrix for fixed effects
Z = design matrix for random effects
e = residual error term with (co)variance/ **R**
R = the variance covariance matrix of error term: often just $\mathbf{I}\,\sigma^2_e$
g = the breeding values: random effects with (co)variance **Z**
Z = variance covariance matrix of the random effects, generally the numerator additive relationship matrix, σ^2_g **A,** estimated here from the marker data.

The BLUEs of the fixed effects and the BLUPs of the breeding values are given by the solution to Henderson's mixed model equations:

$$\begin{bmatrix} X^{'}R^{-1}X & X'R^{-1}Z \\ Z^{'}R^{-1}X & Z^{'}R^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{matrix} X^{'}R^{-1}y \\ Z^{'}R^{-1}y \end{matrix}$$

Obviously these are solved using specialised statistical software. Again, we'll have a go in GenStat. (They are not easily solved in R, unless one uses the commercial software ASREML, which is expensive. The solution will provide estimates of breeding value both for individuals with phenotype information (a value in **Y**) and also for individuals with no phenotype – the predictions we are interested in.

Other methods

More complex, generally Bayesian methods, have been developed for genomic selection, and research in this area is continuing. (The most commonly referred to methods are informatively called "Bayes A" and Bayes "B.") Ridge regression treats the variance associated with each marker as constant. This isn't true; a very small number of markers will be associated with large effects (and variances) but the bulk of markers should have little or no effect. More complex methods model the expected distribution of gene effects more realistically. However, the current view seems to be that BLUP works acceptably well in comparison to these other methods. However, this view is largely based on the results of simulation studies – where the answer you get depends very much on the model you simulate in the first place. There are no empirical, across generation, studies published that I am aware of.

In my hands, with wheat data, I get the very similar results from BLUP as from ridge regression. However, I have no experience of cross-generation studies either.

The problem of kinship

BLUP uses kinship explicitly to make predictions and works well. However, a potential problem is that no individual can be predicted to have a higher breeding value than any which has already been phenotyped. The closer the kinship of two individuals, the closer their breeding values will be. Suppose that the breeding values of one set of lines are known perfectly. The prediction of breeding value for other lines will be on the basis of their genetic similarity to members of this set. A line which is similar to the most elite line with a known breeding value will be predicted to have a similar breeding value itself. If the lines are not identical, however, its breeding value will be shrunk back towards the breeding values of the other lines (to an extent way which depends on its genetic relationship to them). As a result, the estimated breeding value can never exceed that of the best known breeding value.

This is a potential problem for genomic selection as we need to predict several generations ahead. Consider simple pedigree based prediction for an additive trait with a heritability of 100%. Offspring are predicted to have the same breeding value as the mid-parental value. The variance of mid-parent values is Vg/2. The remaining Vg/2 of genetic variation comes from segregation within families. However, the prediction accounts for half the genetic variation among the progeny, which is clearly worthwhile exploiting. (In fact most breeders do this already – they don't make crosses at random but cross the best with the best etc.etc. In this, they implicitly make a prediction of progeny performance.) If we predict two generations ahead on the basis of the average breeding value of the grandparents, the prediction will account for ¼ of the genetic variance among the grand-progeny. Great-grandparent prediction accounts for $1/8^{th}$ and so on. Such pedigree based predictions are thus of decreasing worth over generations and are unlikely to have any merit over more than three generations. The substitution of marker based relationships for pedigree relationships will improve accuracy, but not by much.

What we require are methods which transcend kinship relationships and allow prediction of breeding value which exceed those seen in the current generation. Methods other than BLUP can provide this, but they are not necessarily free from the gravitational pull of kinship. For example, in my hands ridge regression appears to work, at least in part, because the markers predict breeding value on the basis of kinship rather than by directly tagging multiple minor QTL (see the class exercise). This is, I suspect, particularly so when marker numbers are low and the variability in kinship among lines is large. In such cases, there may be too few markers to tag multiple trait loci directly, while the number is adequate to estimate kinship. There is a risk therefore than prediction forward over several generations of selection is reduced in accuracy because marker-trait regression coefficients are too influenced by kinship rather than by the direct effects of trait loci.

Recalibration of markers against phenotype

As selection on GEBV proceeds over generations, allele frequencies and markers and trait loci will change. In addition, recombination will act to reduce linkage disequilibrium among some pairs of markers, while selection (and drift) will act to increase it among others. As a result, over generations, the estimates of GEBV will reduce in accuracy as the calibration of markers against phenotypes becomes outmoded. This calibration must therefore be repeated regularly. This has been simulated, but mainly in the context of animal breeding. It seems likely that recalibration will be required every three or four generations at least. If it were required more frequently than this, the merit of GS would fade.

The accuracy of the GEBV will only be as good as the accuracy of the phenotypes used in the calibration set. This not only relies on the quality of yield trials, but also on the relevance of the population used in this exercise to breeders' germplasm. To take an extreme example, calibration among lines from one cross is unlikely to be of much use for GEBV in an unrelated cross. If the two crosses come from the same population, there is a maximum probability of 0.25 that both will be segregating at any specified bi-allelic locus. This probability reduces if allele frequencies are not equal. There are therefore likely to be few segregating loci for whom predictions from one cross can be transferred to the other. In addition, if the dataset used for calibration has large population structure effects or great variability in kinship, the calibration exercise is likely to be strongly influenced by this. Predictions of GEBV may therefore be very accurate within one or two generations but may fade quickly.

Numbers of markers and size of calibration set.

The number of markers will depend on the rate at which linkage disequilibrium decays within the population used to calibrate the markers and the breeders germplasm (not necessarily the same thing., and the extend to which one is willing to accept predictions which are strongly influenced by kinship. In a population in which LD decays very

rapidly, a large number of markers, >1000, are likely to be required to capture a substantial proportion of Vg by tagging trait loci. However, within the same population, a number of the order of 100 or so, could give accurate predictions based on kinship. In the context of animal breeding, where most research to date has concentrated, numbers of the order of ~10,000 markers are talked about. (This is still better than in humans where one might expect to need 500,000 – as used in the typical genome wide association study.)

The size of the calibration set

One hears discussion of very small numbers of markers and of individuals, <100 of both, being all that is required for genomic selection. To my mind this can only be correct if there is substantial LD within the population and if one is happy that predictions are dominated by kinship relationships. I could believe that these numbers might be acceptable for selecting within an F2 population, or among lines derived from it, but that is all. Again, animal breeding studies seem to indicate that very large numbers of animals are required, though they do not have the luxury of replication as a means of raising heritability.

Genomic selection in inbreeding species

In crops such as wheat and barley, Selection for yield is usually among inbred lines. Selection on GEBV is most effective if the breeding cycle is much quicker that through phenotypic selection. This requires that selection occurs among outbred individuals who are immediately intermated to create the next outbred generation for selection. However, the genetic variance among inbreds is twice that among outbreed individuals from the same population. For equivalent heritabilities (which is probably reasonable as a first approximation, since the markers must be calibrated against the phenotype, response to selection *per cycle* will be $\sqrt{2}$ or about 1.4 times greater for phenotypic than for genomic selection. Response to selection per year may still be greater, however. However, in addition, new lines must be derived from the outbred selection to sell and to recalibrate the markers against the phenotype. This divorcing of the unit of selection (the outcrossed individual) from the genotypes used to establish the marker index (the inbreds) will increase the gap between the generation of selection and the generation of calibration, which will also act to reduce the efficiency of genomic selection. An extreme alternative would be to carry out GS among inbred lines only, but this could increase cycle time to the extent that the gains from GS are not worth the effort. Another possibility is to use GS within crosses only – to increase the efficiency of within family selection. This may not require much additional time and may be worth the effort. There are thus a number of additional considerations to have to do when thinking about applying GS to inbreeding species.

The breeders' equation and genomic selection

Selection on GEBV does not change the genetic variation available for selection. It might increase heritability, but this depends on the quality of phenotypic data used in calibrating markers. Intensity of selection could increase – it may be practical to screen >10,000 individuals for genomic selection but it would be very expensive to screen 10,000 new lines in a yield trial. However, as we have discussed, increasing intensity of selection is not a very efficient way of increasing response to selection.

The great advantage that genomic selection offers is to reduce cycle time. In some crops, it will be possible to get through one or two generations of selection per year, whereas a selection scheme based on phenotype may require several years. In perennial species, trees for example, sexual maturity may occur many years before phenotyping is complete, so the increase in response to selection per year could be very large.

Summary

For the first time since the advent of QTL mapping with molecular markers, around 35 years ago, there is a realistic chance that we can select for polygenic traits like yield using markers. This may greatly increase the rate of response to selection per year. However, we must not get carried away by hype and enthusiasm. There are many issues, probably crop specific, which must be considered before deciding that GC is worth attempting. The best way to evaluate the merit of GS is through basic quantitative genetics principles. There will be many publications in this area in the coming year.

CONCLUSION

"The merging of quantitative and population genetics, driven by data generated by large-scale high-throughput genomics platforms, offers new approaches to classical problems in quantitative genetics."

Whole genome approaches to quantitative genetics
PM Visscher Genetica 2008
DOI 10.1007/s10709-008-9301-7

One must learn by doing the thing; for though you think you know it, you have no certainty until you try.

Sophocles ca.450 BC

GOODBYE

If you are still alive when you read this,
close your eyes. I am
under their lids, growing black.

Bill Knott