Traditional Heritability Analysis

Suggested Reading

Introduction to Quantitative Genetics; Falconer and Mackay (4th edition, 1996; Chapters 7-10)

Data Files

This lecture includes a few examples in `R`. You can run these directly on your Desktop (no need to log into Hong's server) or personal laptop

To run these, you must first download the data file `mod7_data.R`

Having started `R`, you can then load these data in by typing
`load("mod7_data.R")`

```
> ls()
[1] "Galton"   "kinmice"  "pairs1"   "pairs2"   "squares1"
[6] "squares2" "Y1"   "Y2"
```

Covariance

Let X and Y be random variables

Their expected values (means) are $\mathbb{E}(X)$ and $\mathbb{E}(Y)$

The variance of X is

$$Var(X) = \mathbb{E}[X - \mathbb{E}(X)]^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

and the covariance of X and Y is

$$Cov(X, Y) = \mathbb{E}[X\text{-}\mathbb{E}(X)][Y\text{-}\mathbb{E}(Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Properties:

$$Cov(X, X) = Var(X)$$

$$Cov(A + B, C) = Cov(A, C) + Cov(B, C)$$

$$Var(A + B) = Var(A) + Var(B) + 2Cov(A, B)$$

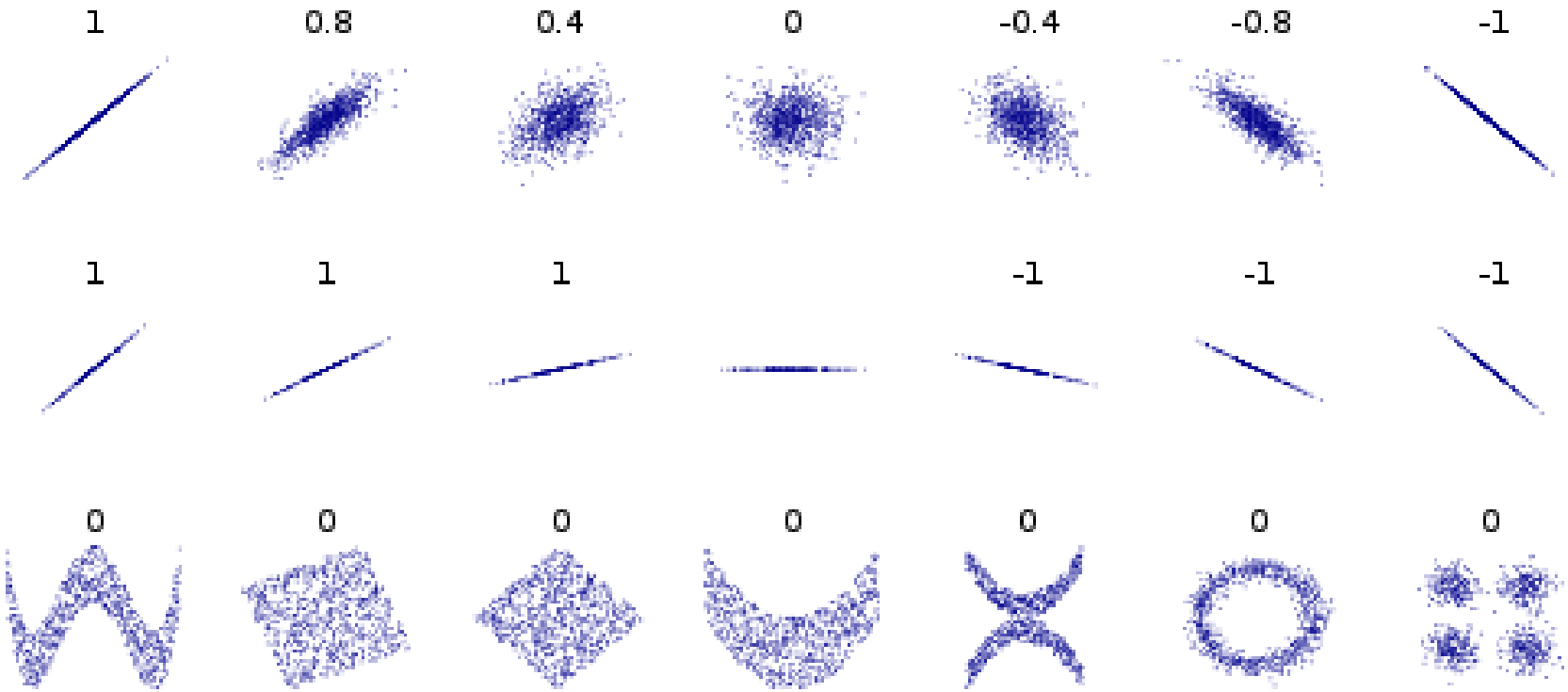
$$Var(cA) = c^2 Var(A)$$

Correlation

Covariance depends on units, so more usual to report correlation:

$$Cor(X, Y) = Cov(X, Y)/Var(X)^{.5}Var(Y)^{.5}$$

Takes values between -1 and +1



Linear Regression

Regress Y (dependent variable) on X (independent / explanatory variable) using the linear model

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

Want to see whether change in X effects change in Y ($\beta \neq 0$), and if so, whether X has a positive ($\beta > 0$) or negative ($\beta < 0$) influence and how large.

Solve by taking covariance of both sides with X:

$$Cov(Y, X) = Cov(\alpha, X) + \beta Cov(X, X)$$

Covariance between r.v. and constant is zero

$$\therefore \hat{\beta} = Cov(X, Y)/Var(X)$$

$$\text{or } \hat{\beta^2} = Cor(X, Y)^2 \times Var(Y)/Var(X)$$

Why are we Interested in Heritability?

Heritability is a fundamental property in quantitative genetics

It indicates to what extent genetic similarities imply phenotypic similarities

Important for human diseases:

determines how well we can predict individual risk

Important in plant & animal genetics:

determines potential of selective breeding

Often interested in broad comparisons, or in demonstrating a trait has positive heritability

Types of Phenotypes

The two major trait types are:

- Quantitative
  phenotypes are continuous measurements in the real interval
  i.e., take values within $(-\infty, \infty)$
- Binary / case-control
  phenotypes take one of two values
  typically recorded as 1 (case) and 0 (control)

Will focus on quantitative traits
  but most ideas can be applied to binary traits

Other trait types include count and survival data
  methods for analysing these are more complicated

Definition of Heritability

Broad sense heritability considers all genetic contributions

- Phenotype = Genetics + Environmental Noise
- $Y = G + E$
- $Var(Y) = Var(G) + Var(E)$ (assumes no $G \times E$ interactions)

Broad sense heritability: $H^2 = Var(G)/Var(Y)$

Narrow sense heritability considers only additive contributions

- Phenotype = Additive Genetics + Environmental Noise
- $Y = A + E$
- $Var(Y) = Var(A) + Var(E)$ (assumes no $A \times E$ interaction)

Narrow sense heritability: $h^2 = Var(A)/Var(Y)$

Definition of Heritability

Can divide the phenotype further. E.g.,

- Phenotype = Additive + Dominant + Common Environment + Noise
- $Y = A + D + C + E$
- $Var(Y) = Var(A) + Var(D) + Var(C) + Var(E)$

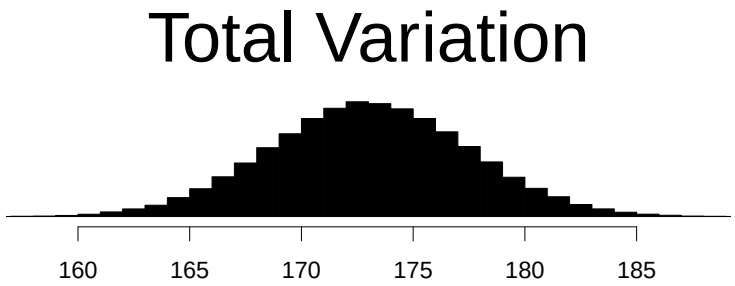
Broad sense heritability: $H^2 = (Var(A) + Var(D))/Var(Y)$

Narrow sense heritability: $h^2 = Var(A)/Var(Y)$

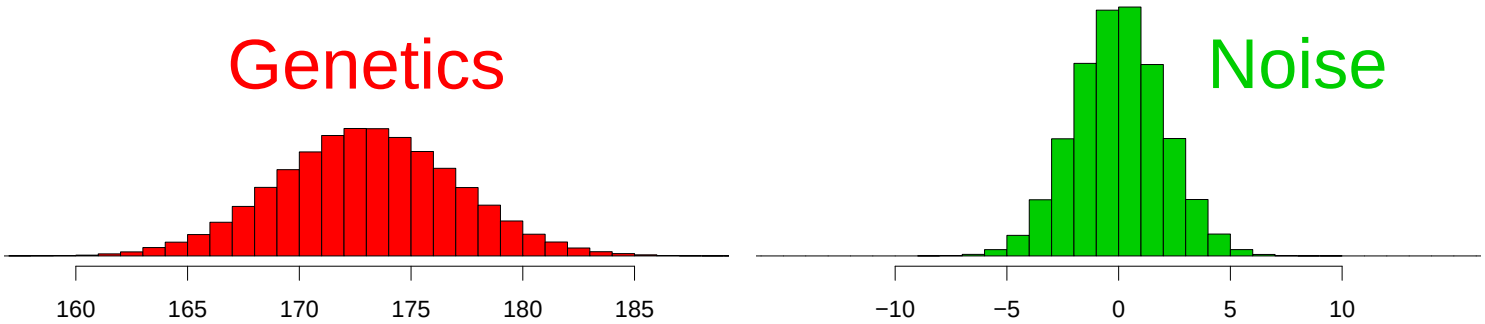
We typically focus on narrow-sense heritability

Definition of Heritability

For example, the variance of human height is about 20cm



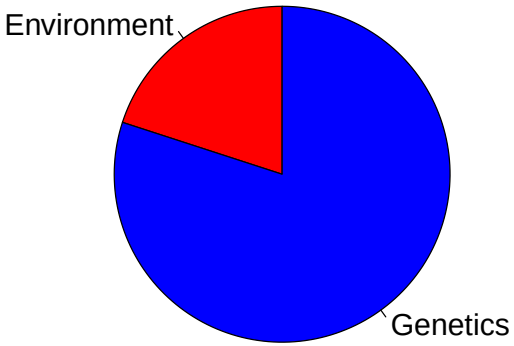
Of which about 16cm is due to genetics, 4cm due to other factors (noise)



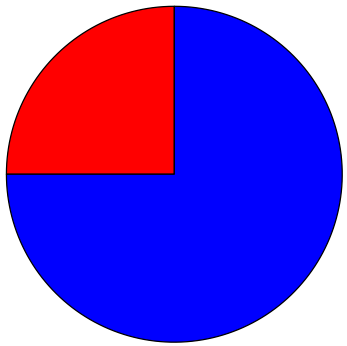
Therefore, the heritability of height is $16/20 = 80\%$

Some Heritabilities of Human Traits

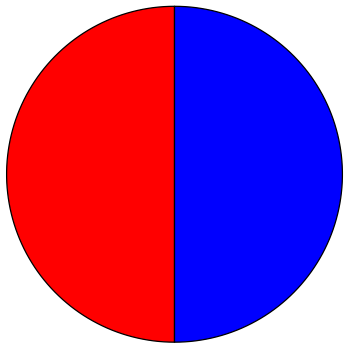
Human Height



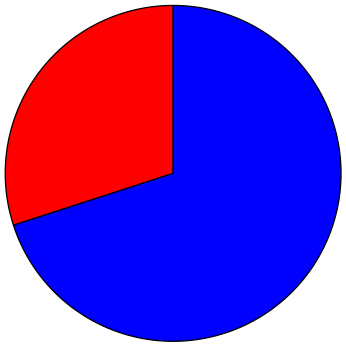
Schizophrenia



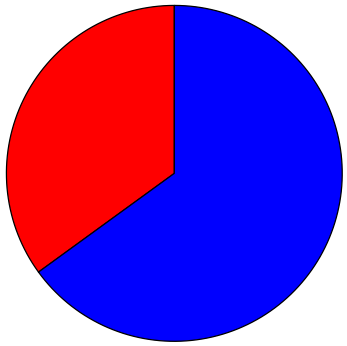
Obesity



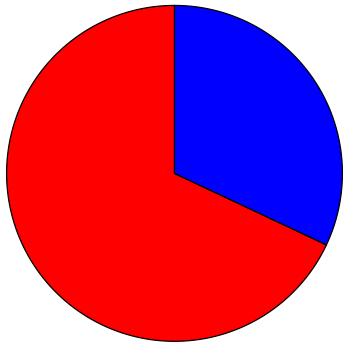
Crohn's Disease



Bipolar Disorder



Epilepsy

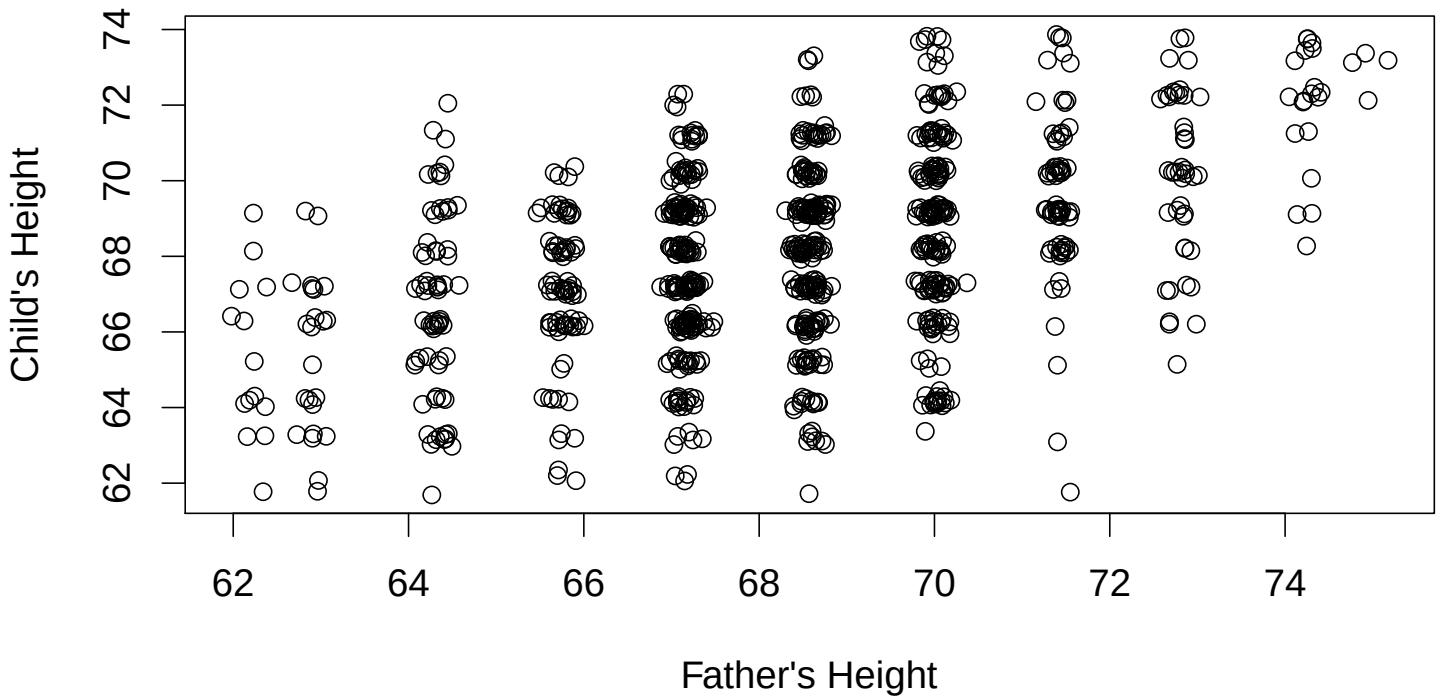


Example: Galton height data

```
> summary(Galton)
     father           child
 Min.   :62.22   Min.   :61.70
 1st Qu.:67.17   1st Qu.:66.20
 Median :68.58   Median :68.20
 Mean   :68.31   Mean   :68.09
 3rd Qu.:69.99   3rd Qu.:70.20
 Max.   :74.94   Max.   :73.70
> var(Galton[,1])
[1] 6.389121
> var(Galton[,2])
[1] 6.340029
```

Example: Galton height data

```
> plot(Galton[,1],Galton[,2])
> plot(Galton[,1]+rnorm(928,0,.1),Galton[,2]+rnorm(928,0,.1),
xlab="Father's Height",ylab="Child's Height")
> abline(v=mean(Galton[,1]),col=2,lty=3,lwd=3)
> abline(h=mean(Galton[,2]),col=2,lty=3,lwd=3)
> abline(a=0,b=1,col=2,lty=3,lwd=3)
```



Example: Galton height data

```
> lm(Galton[,2]~Galton[,1])

Call:
lm(formula = Galton[, 2] ~ Galton[, 1])

Coefficients:
(Intercept)  Galton[, 1]
     36.872        0.457
```

We are fitting the model $C = \alpha + \beta F$, where F contains fathers' heights, C contains childrens' heights. We estimate $\alpha = 36.9$ (intercept) and $\beta = 0.46$ (slope/gradient).

Turns out, $2 \times 0.46 = 0.92$ is an estimate of heritability; will see why later

Ways to measure heritability

There are many possible study designs, but the basic principle is to see whether closely related individuals tend to have similar phenotypes

If close relatives have more similar phenotypes $\Rightarrow$ higher heritability

1 - The study can either focus on collecting only one (or two) types of relations; e.g., recruiting father-son pairs, or only twins

Estimates of heritability follow from **The Covariance Equation**

2 - The study can recruit entire families

There will be a range of different relationships, so we estimate heritability using **The Mixed Model**

Measuring relatedness

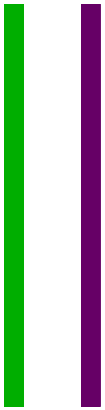
Relatedness can be measured via θ, r and f, the coefficients of coancestry (or kinship), of relatedness and of inbreeding

The coefficient of coancestry θ between two individuals is the probability that two homologous alleles, one chosen at random from each individual, are identical by descent (IBD) from a known common ancestor. If (k_0, k_1, k_2) denote the fraction of the genome "IBD0", "IBD1" or "IBD2"

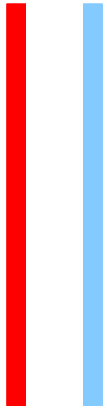
David will discuss these in more detail

Identity by Descent

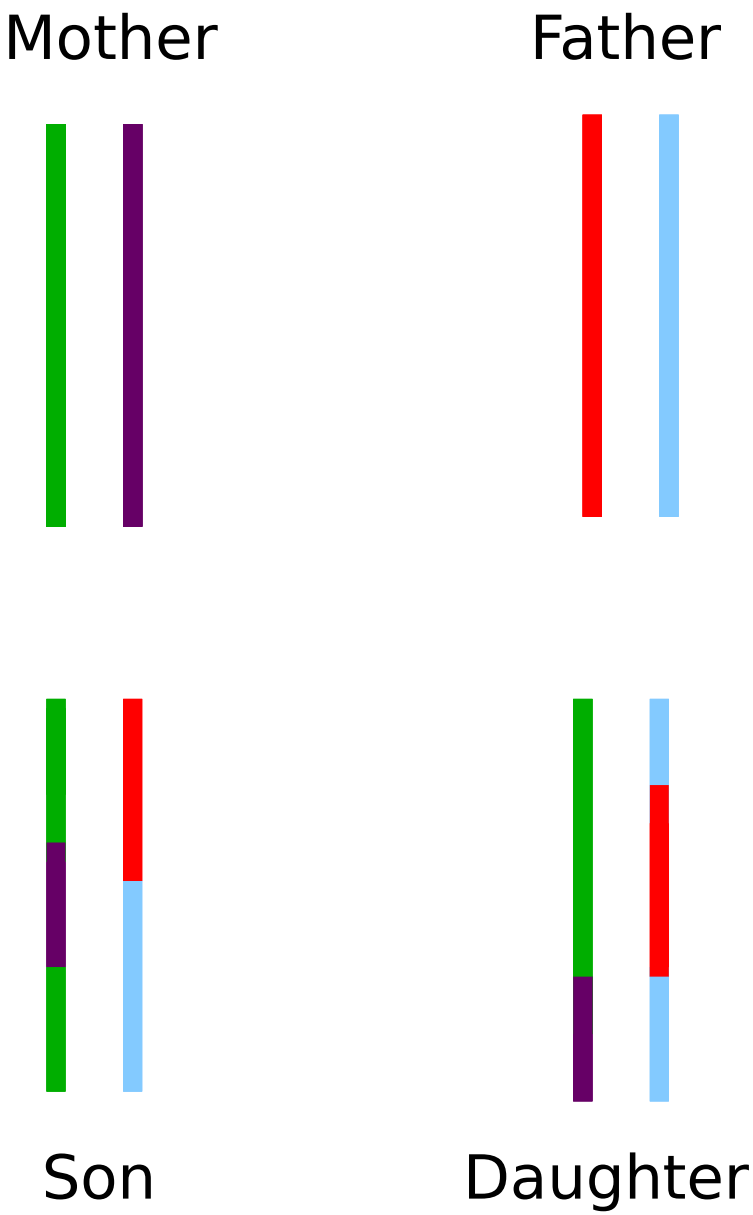
Mother



Father



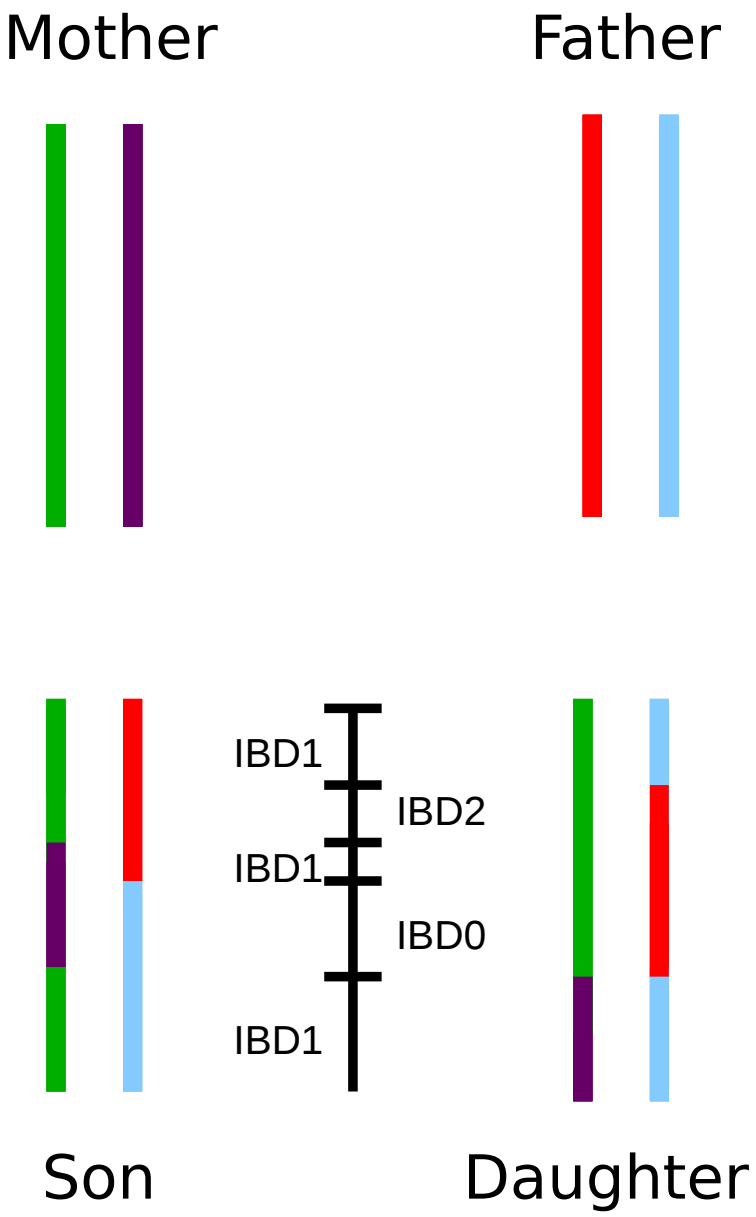
Identity by Descent



Each child inherits a set of maternal chromosomes and a set of paternal chromosomes

Matching colours indicate the son and daughter are IBD for that position

Identity by Descent



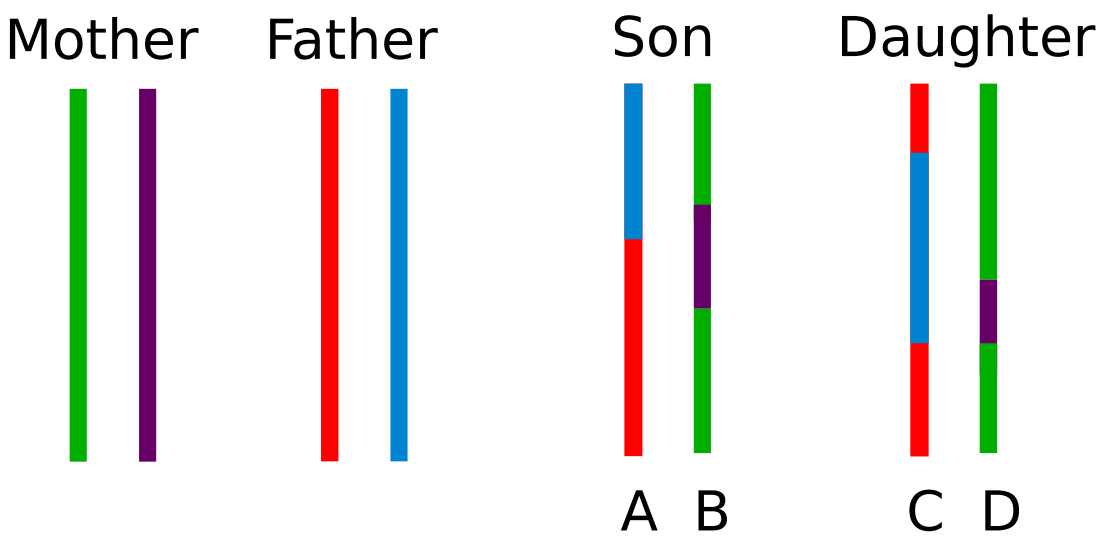
Each child inherits a set of maternal chromosomes and a set of paternal chromosomes

Matching colours indicate the son and daughter are IBD for that position

Alleles will match at positions of different colours, simply due to chance (not common ancestry)

Let (k_0, k_1, k_2) be expected fractions IBD0, IBD1 and IBD2
Then $\theta = k_1/4 + k_2/2$

Identity by Descent



Allele picked from	Probability common ancestor (same colour)
A & C	1/2
A & D	0
B & C	0
B & D	1/2
Average	1/4

Therefore, full-sibs also have $\theta = 1/4$ and $r = 1/2$

Measuring relatedness

Relatedness can be measured via θ, r and f, the coefficients of coancestry (or kinship), of relatedness and of inbreeding

The coefficient of coancestry θ between two individuals is the probability that two homologous alleles, one chosen at random from each individual, are identical by descent (IBD) from a known common ancestor. If (k_0, k_1, k_2) denote the fraction of the genome "IBD0", "IBD1" or "IBD2"

The coefficient of relatedness r is twice the coefficient of kinship

The coefficient of inbreeding (or of consanguinity) of an individual is the coefficient of kinship of their parents; an individual is outbred if their parents are unrelated

David will discuss these in more detail

Common coefficients

Relationship (S, T)	k_0, k_1, k_2	θ	r
MZ twins (0,1)	0, 0, 1	1/2	1
DZ twins / Full-sibs (2,2)	1/4, 1/2, 1/4	1/4	1/2
Parent-child (1,1)	0, 1, 0	1/4	1/2
Half-sibs (2,1)	1/2, 1/2, 0	1/8	1/4
Uncle-niece (3,2)	1/2, 1/2, 0	1/8	1/4
Grandparent-grandchild (2,1)	1/2, 1/2, 0	1/8	1/4
Cousins (4,2)	3/4, 1/4, 0	1/16	1/8

Handy formula: if we define each relationship by S, its degree (the number of links (meioses) between the pairs) and T, the number of common ancestors, then $\theta = T \times 0.5^{S+1}$

The Covariance Equation

Suppose the phenotype model $Y = A + D + C + E$

(additive and dominant effects, common environment and noise)

If Individuals i and j have IBD vector (k_0, k_1, k_2), then common to assume:

$$\begin{aligned} Cov(Y_i, Y_j) &= rVar(A) + k_2 Var(D) + \gamma Var(C), \\ &= 2\theta Var(A) + k_2 Var(D) + \gamma Var(C), \end{aligned}$$

where $\gamma = 1$ if the individuals share environment (else $\gamma = 0$)

Proof complicated (see end or Intro to QG), but can check makes sense:

If unrelated, $Cov(Y_i, Y_j) = 0$;

If MZ Twins, $Cov(Y_1, Y_j) = Var(A) + Var(D) + Var(C)$;

For DZ Twins, get $Cov(Y_1, Y_j) = Var(A)/2 + Var(D)/4 + Var(C)$

Using the Covariance Equation to Estimate Heritability

We want to estimate $h^2 = Var(A)/Var(Y)$

It's fairly easy to measure $Var(Y)$

Can estimate from any randomly picked individuals

We use the covariance equation to estimate $Var(A)$

If we measure the covariance between pairs of individuals, then The Covariance Equation tells us how this value is related to $Var(A)$

The Covariance Equation can also be used for binary traits (see end)

Parent-child studies

Suppose we have heights for fathers F and children C

1 - Estimate $Var(Y)$

Can get this from either $Var(F)$ or $Var(C)$

For Galton data, $Var(F) = 6.4$ (inches), while $Var(C) = 6.3$

2 - Estimate $Var(A)$

Putting $r = 0.5$, $k_2 = 0$ and $\gamma = 0$ into the covariance equation gives $Cov(F, C) = Var(A)/2$

For these data, Cov(F,C) = 2.9 (inches)

so our estimate of $Var(A)$ is 5.8

Parent-child studies

Suppose we have heights for fathers F and children C

3 - Estimate $h^2 = Var(A)/Var(Y)$

Therefore, our estimate of heritability is $\frac{2Cov(F,C)}{Var(F)} = 5.8/6.4 = 0.91$

When regressing C on F, the aim is to solve $C = \alpha + \beta F$

the least squares estimate of β is $Cov(F, C)/Var(F)$

Therefore, with parent-child data, you can estimate h^2 simply by regressing children's height on father's height (and multiplying by two)

MZ twins

Let M1 and M2 be MZ twins ($r = 1$, $k_2 = 1$, $\gamma = 1$)

The covariance equation gives $Cov(M1, M2) = Var(A) + Var(D) + Var(C)$

If we assume $Var(D) = Var(C) = 0$

then $Cov(M1, M2) = Var(A)$

Therefore $Cov(M1, M2)/Var(M1)$ is an estimate of h^2

Probably OK to suppose no dominance contributions, but assuming no effects of shared environment is unlikely to be accurate ...

MZ and DZ Twins: The Twin Method / ACE Model

Let D_1 and D_2 be DZ twins ($r = 0.5$, $k_2 = 0.25$, $\gamma = 1$)

The covariance equations are

$$Cov(M1, M2) = Var(A) + Var(D) + Var(C)$$
$$Cov(D1, D2) = Var(A)/2 + Var(D)/4 + Var(C)$$

If we assume $Var(D) = 0$, we get:

$$Cov(M1, M2) = Var(A) + Var(C)$$
$$Cov(D1, D2) = Var(A)/2 + Var(C)$$

Therefore, $2[Cov(M1, M2) - Cov(D1, D2)]$ is an estimate of $Var(A)$

If we divide through by Var(Y)=Var(M1)=Var(M2)=Var(D1)=Var(D2),

we find that $2[Cor(M1, M2) - Cor(D1, D2)]$ is an estimate of h^2

MZ and DZ Twins: The Twin Method / ACE Model

So to estimate heritability using the twin methods, simply compute correlation between MZ twins, then correlation between DZ twins, then twice the difference is an estimate of h^2

If the difference is zero, then all similarity must be due to shared environment

Meanwhile, 2 Cor(D1,D2) - Cor(M1,M2) provides an estimate of Var(C)

Some Caveats

All heritability values are estimates

so should be accompanied by standard errors / confidence interval

Larger sample size (or higher sample relatedness) $\Rightarrow$ higher precision

Important to realise that heritability is sample-specific; it depends on sample make-up and population / environment considered.

- h^2 for height lower in poor societies reflecting limited access to food & medical care
- h^2 for IQ is low in infants and rises with age, presumably due to changes in environmental effects.

More Caveats

An estimate's accuracy depends on accuracy of the model assumptions

Twin studies have consistently estimated high genetic contributions for a wide range of phenotypes. Some researchers are suspicious of these high estimates.

We have been ignoring cross terms and assuming additivity

We have made very simple assumption about environment: either entirely shared or absent. Environment may differ between MZ and DZ twins, e.g. different parental attitude or in utero environment (DZ twins mistaken for MZ were found to resemble MZ more than DZ twins in phenotype concordances for psychological traits

Gunderson et al., Twin Res. Hum. Genet. (2006)

Twins reared apart (adoption studies) can also be used to minimise the effect of common sibling environment

Limitations of Using the Covariance Equation

Can be difficult to recruit enough pairs of the desired type (e.g., twins, or father-son pairs

This approach is very wasteful, as it excludes other family members

Given related pairs of multiple types, in theory, we could perform separate analyses and combine results

But the more elegant solution is to use the Mixed Model

The Mixed Model generalizes the covariance equation, allowing us to simultaneously analyse multiple types of relatedness

Mixed model

The mixed model takes the form:

$$\begin{bmatrix} \quad Cov(Y) \quad \end{bmatrix} = \begin{bmatrix} \quad K \quad \end{bmatrix} Var(A) + \begin{bmatrix} \quad I \quad \end{bmatrix} Var(E)$$

K is (twice) the kinship matrix; each element of K indicates the coefficient of relatedness between a pairs of individuals

So if individuals 1 and 2 are parent-child, $K_{1,2} = 0.5$; if individuals 3 and 4 are twins, $K_{3,4} = 1$; if individuals 1 and 3 are unrelated, $K_{1,3} = 0$, etc.

I is an identity matrix (1 on diagonal, 0 otherwise)

Matrices

Matrices allow us to store many equations in an efficient manner

E.g.,[†]

$$\begin{bmatrix} 7 & -2 \\ -6 & 1 \end{bmatrix} = \begin{bmatrix} -3 & 2 \\ -5 & 8 \end{bmatrix} a + \begin{bmatrix} 3 & 5 \\ 8 & -4 \end{bmatrix} b,$$

stores 4 separate equations:

$$\begin{aligned} 7 &= -3a & &+3b \\ -6 &= -5a & &+8b \\ -2 &= 2a & &+5b \\ 1 &= 8a & &-4b \end{aligned}$$

[†]for demonstration only, these equations don't make sense mathematically!

Mixed model explained

The model: Cov(Y) = K Var(A) + I Var(E)

$$\begin{pmatrix} Var(Y_1) & Cov(Y_1, Y_2) & Cov(Y_1, Y_3) & \dots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & Cov(Y_2, Y_3) & \dots & Cov(Y_2, Y_n) \\ Cov(Y_3, Y_1) & Cov(Y_3, Y_2) & Var(Y_3) & \dots & Cov(Y_3, Y_n) \\ \vdots & & \ddots & & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & Cov(Y_n, Y_3) & \dots & Var(Y_n) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & K_{1,2} & K_{1,3} & \dots & K_{1,n} \\ K_{2,1} & 1 & K_{2,3} & & K_{2,n} \\ \vdots & & \ddots & & \vdots \\ K_{n,1} & K_{n,2} & K_{2,3} & \dots & 1 \end{pmatrix} Var(A) + \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} Var(E)$$

The mixed model represents a collection of $n +^n C_2 = n \times (n+1)/2$ equations, each describing the covariance between a pair of individuals:

$$Cov(Y_i, Y_j) = \begin{cases} K_{i,j} Var(A) & \text{when } i \neq j \\ Var(A) + Var(E) & \text{when } i = j \end{cases}$$

Instructions for using the mixed model

1 - collect individuals for one or more families of any relationship types

2 - record phenotypes Y

3 - construct kinship matrix K

Using software, solve to find estimate of Var(Y), Var(A) and Var(E)

The estimate of h^2 is $Var(A)/Var(Y)$

The usual method of solving is called REML; this finds Var(Y), Var(A) and Var(E) which maximise the REstricted Maximum Likelihood

Many software exist for this (e.g., ASREML, GCTA, LDAK)

Will cover GCTA and LDAK in later modules

Kinship matrix example 1

Suppose we have 3 parent-child pairs, (P1 & C1, P2 & C2, P3 & C3)

$$K = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1 \end{pmatrix}$$

(Individuals P1 C1 P2 C2 P3 C3)

Kinship is zero for pairs when relatedness not known

Kinship matrix example 2

Suppose we have a parent-child pair (P & C), two MZ twins (M1 & M2) and two DZ twins (D1 & D2). The corresponding kinship matrix is

$$K = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

(Individuals P C M1 M2 D1 D2)

Kinship matrix example 2

Suppose we have a parent-child pair (P & C), two MZ twins (M1 & M2) and two DZ twins (D1 & D2). The corresponding kinship matrix is

$$K = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1 \end{pmatrix}$$

(Individuals P C M1 M2 D1 D2)

Because we have siblings, we could include common environment Var(C) in the mixed model: $\text{Cov(Y)} = \text{K Var(A)} + \gamma \text{ Var(C)} + \text{I Var(E)}$

where the matrix γ contains pairwise shared environ. coefficients

Arrays and Matrices in R

Reminder: to construct arrays and matrices in R, use array() and matrix()

```
array(dim=3)
[1] NA NA NA

array(1,dim=3);array(c(1,2,3),dim=3);array(1:3,dim=3)

array(1:2,dim=3)
[1] 1 2 1

matrix(nrow=2,ncol=3)
     [,1] [,2] [,3]
[1,]   NA   NA   NA
[2,]   NA   NA   NA

matrix(1:6,nrow=2,ncol=3);matrix(1:6,nrow=2)

diag(2)
[1,]    1    0
[2,]    0    1

diag(3,2);diag(c(1,2,3));diag(c(1,2),6)
```

www.r-tutor.com/r-introduction/matrix

www.r-tutor.com/r-introduction/matrix/matrix-construction

Construct this matrix in R

$$K = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1 \end{pmatrix}$$

Construct this matrix in R

Can do in one step:

```
> kin1=matrix(c(1,.5,0,0,0,0,.5,1,0,0,0,0,
0,0,1,1,0,0,0,0,1,1,0,0,
0,0,0,0,1,.5,0,0,0,0,.5,1),byrow=T,nrow=6)
```

Or multiple steps, by first allocating an identity matrix

```
> kin1=diag(6)
```

then using square brackets to change elements

```
> kin1[1,2]=.5; kin1[3,4]=1
```

and so on

More general matrix construction

```
> pairs1
     Ind1 Ind2 Relatedness
[1,]    1    2  0.5
[2,]    3    4  1.0
[3,]    5    6  0.5

#fill kin1 using a loop - here is an example loop
>for(row in 1:3)
{
i=pairs1[row,1];
j=pairs1[row,2];
rel=pairs1[row,3]
cat(paste("Individuals ",i," & ",j,", Relatedness: ",
rel,"\n",sep=""))
}
```

Construct this matrix in R

```
> kin1=diag(6)
> for(row in 1:3)
{
i=pairs1[row,1];
j=pairs1[row,2];
rel=pairs1[row,3]

kin1[i,j]=rel;
kin1[j,i]=rel
}
```

Now make the kinships corresponding to pairs2

Hint: use nrow() to get number of rows, max() to find out how many individuals

Construct this matrix in R

```
> kin2=diag(100)
> for(row in 1:nrow(pairs2))
{
i=pairs2[row,1];j=pairs2[row,2];rel=pairs2[row,3]
kin2[i,j]=rel;kin2[j,i]=rel
}

#see what it looks like
> up=which(upper.tri(kin2,diag=F)) #plot only upper diagonal terms
> hist(kin2[up],n=100,xlab="Relatedness",axes=F)
> axis(1,at=.5^(0:5),lab=c(1,"1/2","1/4","1/8","1/16","1/32"))
> axis(2)

#plot only positive (upper diagonal) terms
> pos=intersect(up,which(kin2>0))
> hist(kin2[pos],n=100,xlab="Relatedness",axes=F)
> axis(1,at=.5^(0:5),lab=c(1,"1/2","1/4","1/8","1/16","1/32"))
> axis(2)
```

Solving the mixed model

The proper approach for estimating Var(A) and Var(E) is to use REML, but an alternative (Haseman Elston regression) is to regress across all pairs the squared difference between the two phenotypes (Z) on the kinship value (X). Minus twice the gradient is an estimate of Var(A)

1 - collect n individuals
2 - record phenotypes Y
3 - construct K
4 - for each pair of individuals, i & j, record relatedness ($X = K_{i,j}$) and squared difference in phenotypes ($Z = (Y_i - Y_j)^2$)
5 - regress Z on X (using model $Z = \alpha + \beta X$)

α has expected value 2 Var(Y), β has expected value -2 Var(A) (proof at end)

so $-\beta/\alpha$ is an estimate of h^2

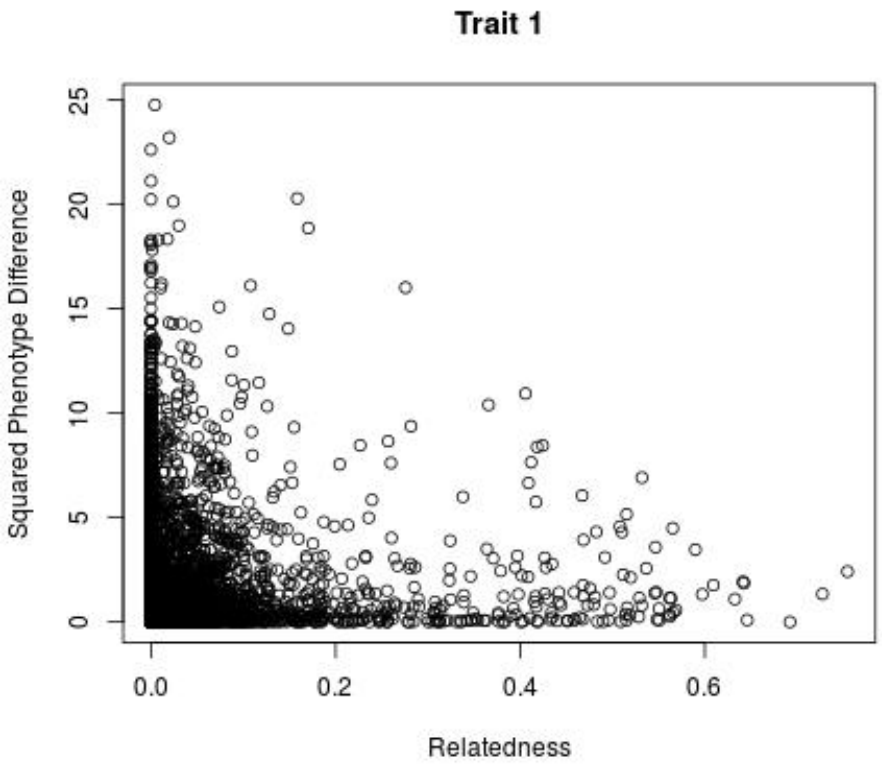
Mice example

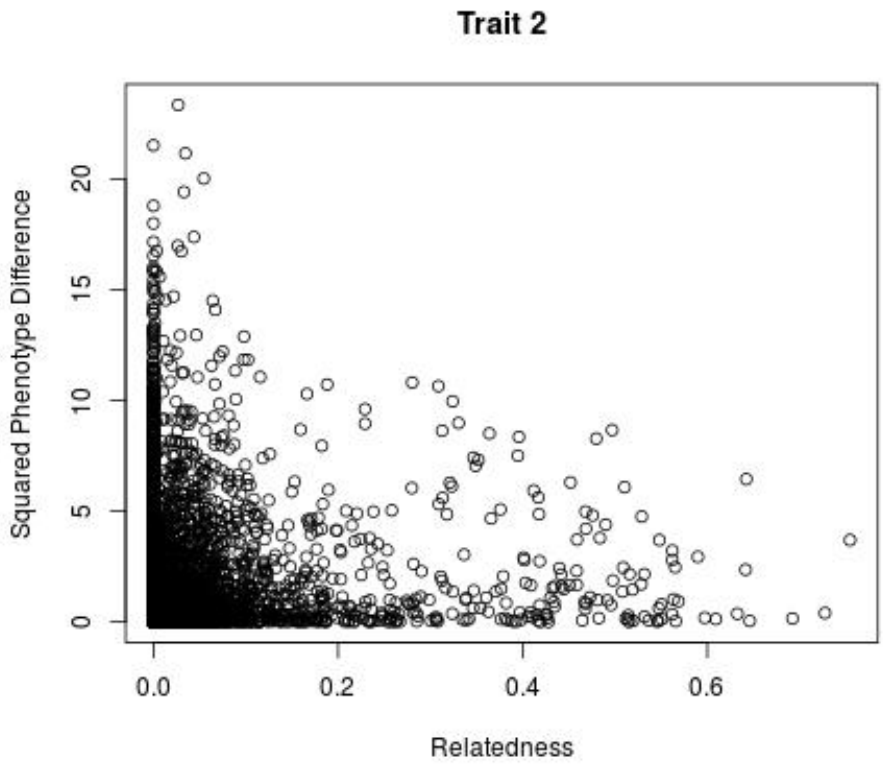
```
#kinmice is a kinship matrix for 100 highly related mice
> hist(as.numeric(kinmice), n=100)

#Y1 and Y2 are two phenotypes
#squares1 and squares2 contain the squared differences

#hint, to make these, could use a loop within a loop
> squares1=NULL;squares2=NULL
> for(i in 1:99)
{
for(j in (i+1):100)
{
squares1=rbind(squares1,c(i,j,kinmice[i,j],(Y1[i]-Y1[j])^2))
squares2=rbind(squares2,c(i,j,kinmice[i,j],(Y2[i]-Y2[j])^2))
}}
```

(Ignore for moment that kinships are continuous)

Which trait is most heritable?

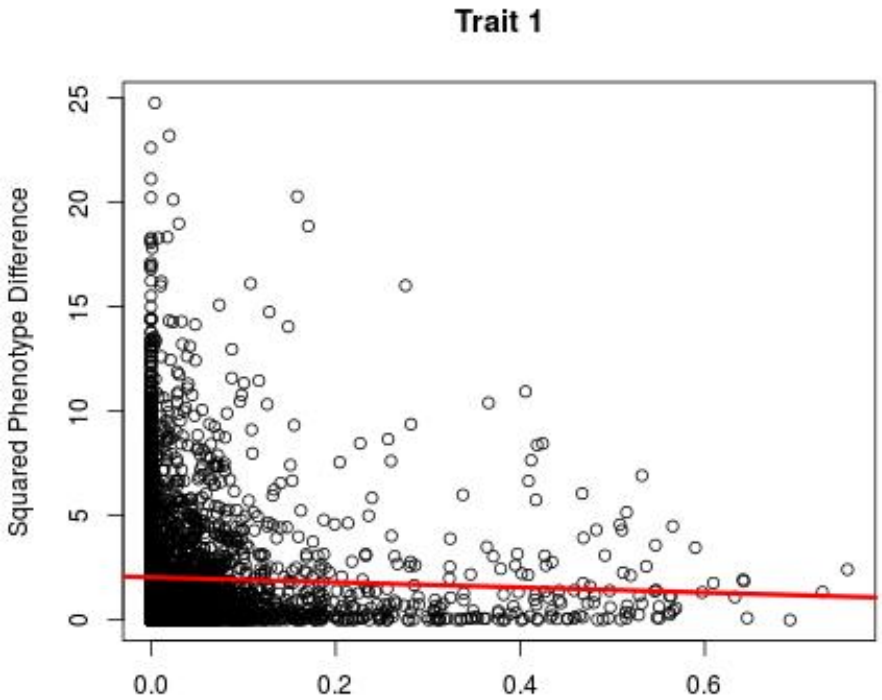
```
> par(mfrow=c(1,2))
> plot(squares1[,3],squares1[,4],xlab="Relatedness",
ylab="Squared Phenotype Difference",main="Trait 1")
> plot(squares2[,3],squares2[,4],xlab="Relatedness",
ylab="Squared Phenotype Difference",main="Trait 2")
```

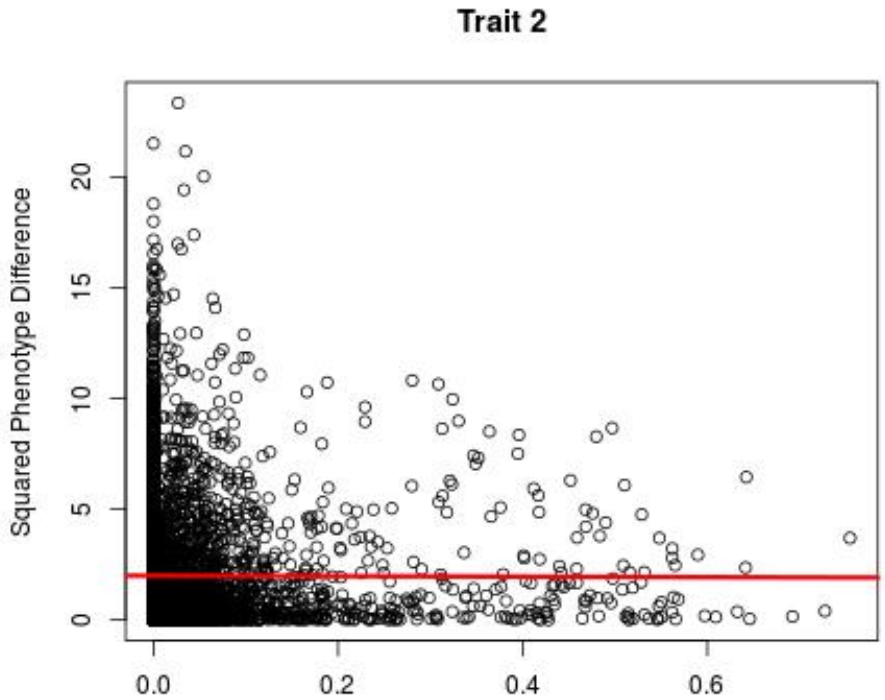




Hard to tell? Add regression lines

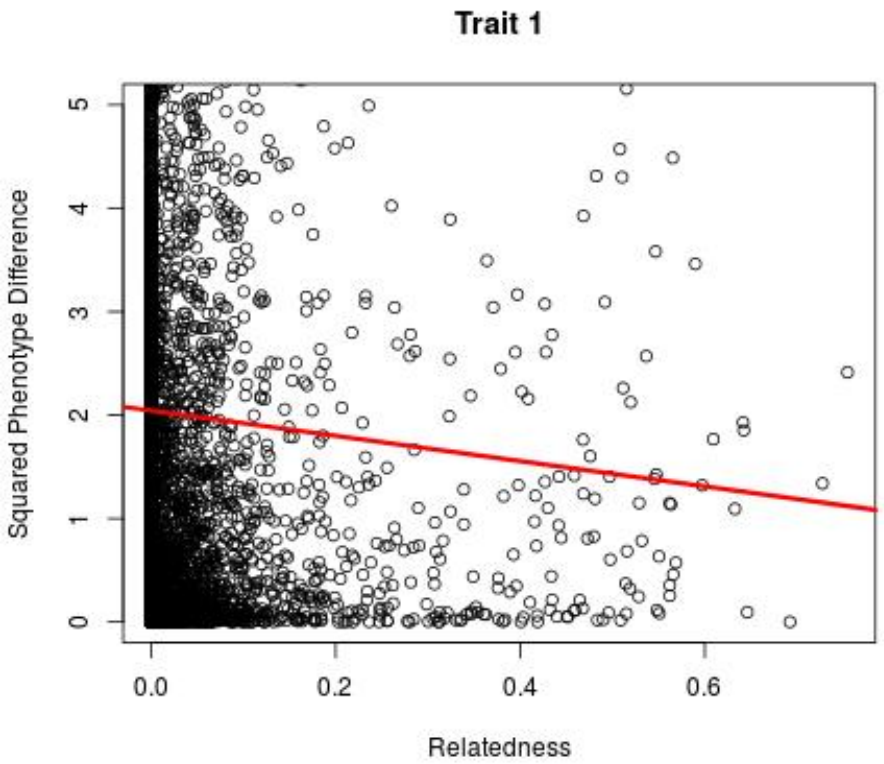
```
> reg1=lm(squares1[,4]~squares1[,3])
> reg2=lm(squares2[,4]~squares2[,3])
> plot(squares1[,3],squares1[,4],main="Trait 1")
> abline(reg1,col=2,lwd=3)
> plot(squares2[,3],squares2[,4],main="Trait 2")
> abline(reg2,col=2,lwd=3)
```

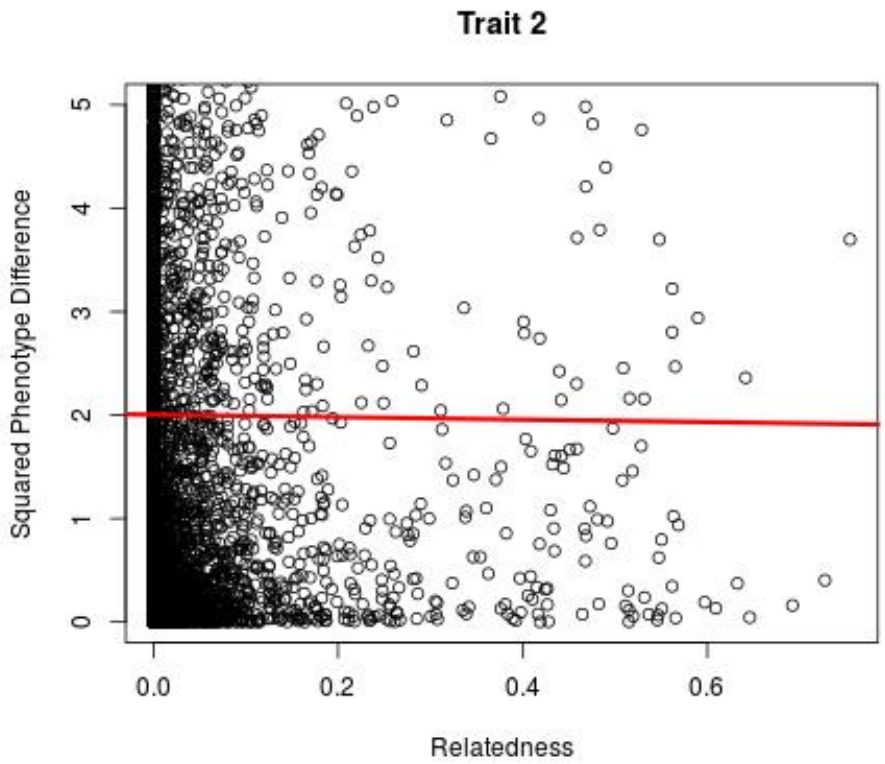




Still can't tell? Zoom in

```
> plot(squares1[,3],squares1[,4],main="Trait 1",ylim=c(0,5))
> abline(reg1,col=2,lwd=3)
> plot(squares2[,3],squares2[,4],main="Trait 2",ylim=c(0,5))
> abline(reg2,col=2,lwd=3)
```





Estimate h^2 by Haseman Elston Regression

```
> reg1=lm(squares1[,4]~squares1[,3])
> reg1
Call:
lm(formula = squares1[, 4] ~ squares1[, 3])
Coefficients:
  (Intercept)  squares1[, 3]
        2.043         -1.223

> reg1$coeff[1]
(Intercept)
   2.043432
> -reg1$coeff[2]/reg1$coeff[1]
    0.5986112

> -reg2$coeff[2]/reg2$coeff[1]
   0.05966459
```

Problems with pedigree-based heritability estimation

Traditional heritability analyses require only pedigree and phenotypic data. But relying on pedigree information is problematic:

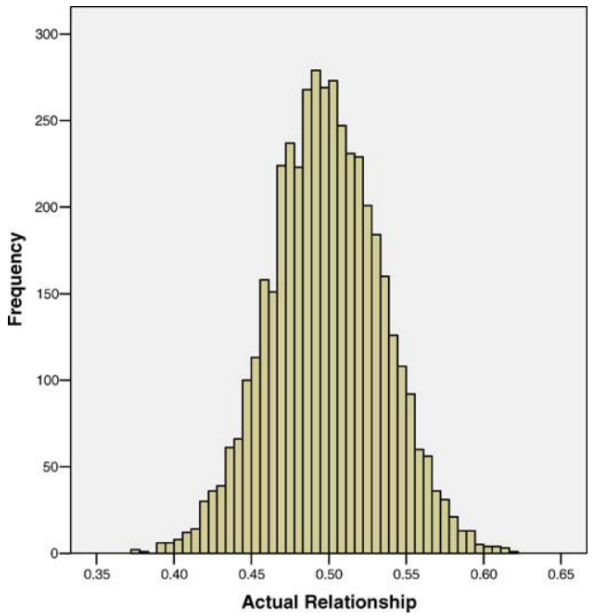
Estimates of heritability depend on the pedigree information available

- There is no such thing as a complete pedigree
- Discovering a new common ancestor will change estimates
- In absence of information we must assume individuals "unrelated"

Moreover, even a "complete" pedigree will provide only expected IBD fractions / coefficients of kinships. Now we have full-genome SNP data, we can measure IBD exactly

Theoretical distribution of θ

Full-siblings are expected to have $\theta = -.25$ ($r = 0.5$). But actual fraction of genome they share could be 25% higher or 25% less



From Assumption-Free Estimation from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Gen. 2006

Allelic Correlations

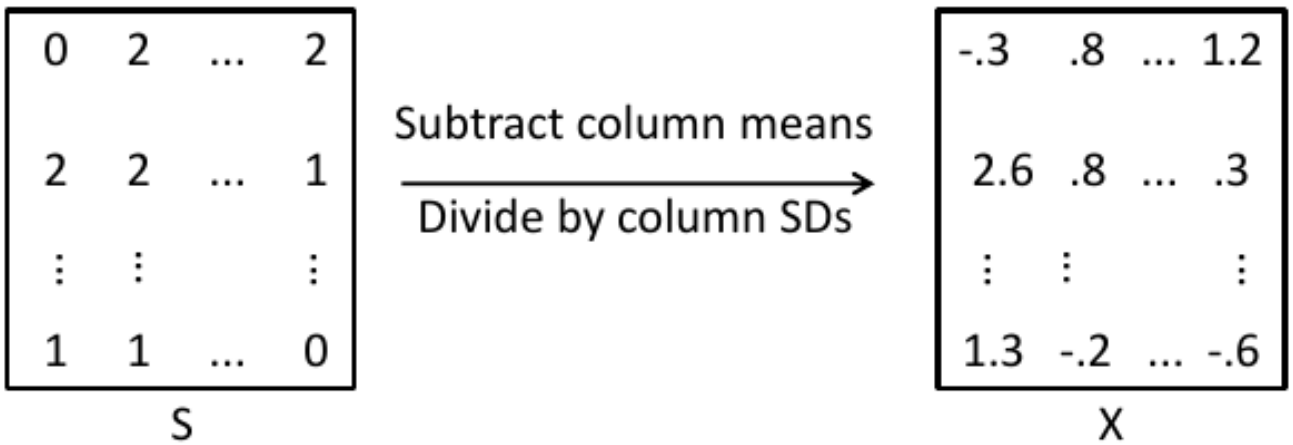
Now genome-wide genotyping of samples is widespread; we can easily record genotypes (0,1,2 allele counts) for over 1M SNPs.

Therefore, we now construct K using allelic correlations, which measure genomic similarity between pairs of individuals. These look similar to coefficients of relatedness (higher values indicate closer relatedness) but are continuous-valued (rather that restricted to 1s, 1/2s, 1/4s, etc), and while most values are within 0 and 1, values can lie outside (a negative value means two individuals are less genetically similar than two individuals picked at random)

(The mice kinships a few slides ago were estimated by allelic correlations)

Allelic Correlations

Suppose the matrix S (size n x N) contains for n individuals the genotypes for each of N SNPs. First, for each SNP, we standardise the genotypes so they have mean zero and variance one



Then we use XX^T/N ("allelic correlations") as an estimator of relatedness

Allelic Correlations

XX^T/N measures pairwise IBS (identity by state) how similar the genotype values are for each pair of individuals

For example, to calculate $K_{1,2}$, correlations between individuals 1 and 2, you can imagine laying their two genomes side by side

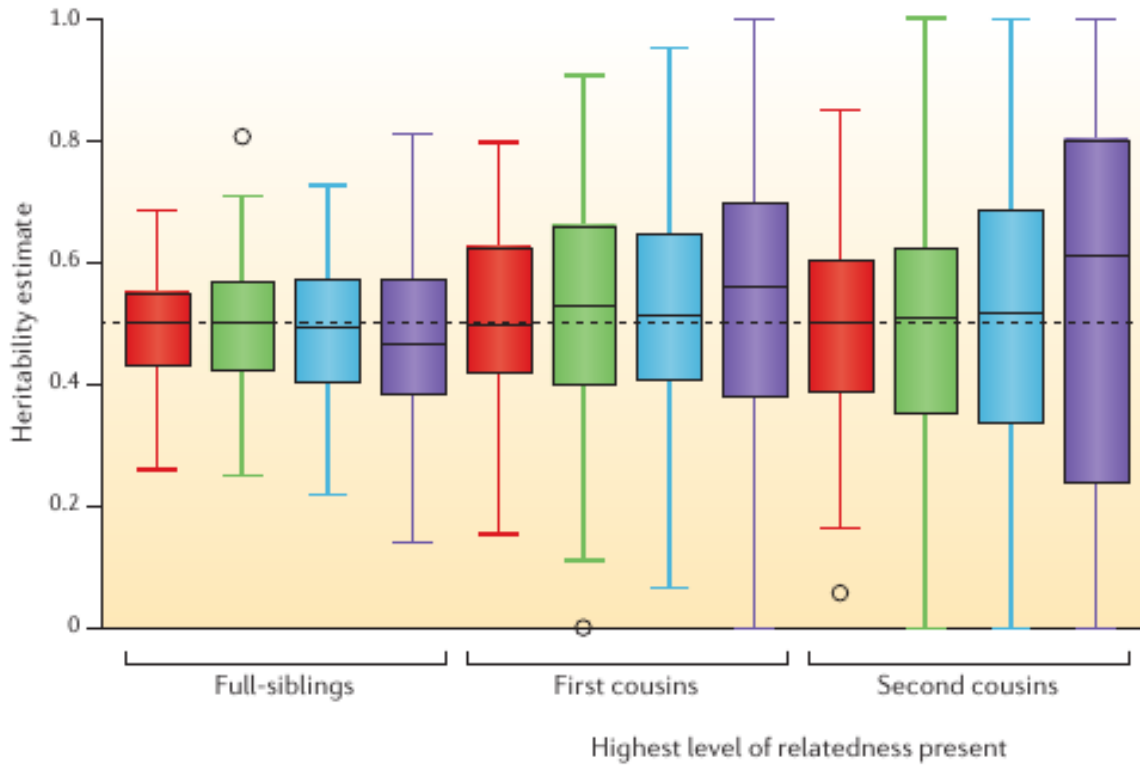
S_1:	0	2	2	1	0	2
S_2:	2	2	0	2	0	1
Effect on K_{12}	–	+	–	...	+	...

X_1:	-.3	.8	.9	.8	-.3	1.2
X_2:	2.6	.8	-.5	1.6	-.3	.3

K_{12} = (–0.78 +0.64 –0.45 +1.28 +0.09 +0.36) / N

Using actual relatedness improves accuracy

By measuring actual relatedness, rather than relying on expected relatedness, we can obtain (slightly) more precise estimate of h^2



Purple boxes are estimates using expected relatedness; red use actual relatedness (green, blue use less accurate measures of actual relatedness)

Formal Proof for Additive Covariance

Suppose the model $Y_i = A_i + D_i + E_i$
$= A_i(G_i) + D_i(G_i) + E_i$

where G_i contains genotypes of Individual i

$$Cov(Y_i, Y_j) = Cov(A_i + D_i + E_i, A_j + D_j + E_j)$$
$$= Cov(A_i, A_j) + Cov(D_i, D_j)$$

Must calculate $Cov(A_i, A_j)$ and $Cov(D_i, D_j)$; can do separately
Use the identity
$$Cov(A_i, A_j) = \sum_{j=0}^{2} \mathbb{P}(IBD = j)[\mathbb{E}(A_i A_j | IBD = j) - \mathbb{E}(A_i)\mathbb{E}(A_j)]$$
$$= \sum_{j=0}^{2} k_j[\mathbb{E}(A_i A_j | IBD = j) - \mathbb{E}(A)^2]$$

Suppose a single causal SNP with $\mu = 0$, so the SNP genotypes (0, 1, 2) have effects (-a, d, a) and minor (major) allele frequency p ($q = 1 - p$).

Formal Proof for Additive Covariance

$\mathbb{E}(A_i) = \mathbb{E}(A_j) = \mathbb{E}(A) = a(p^2 - q^2) = a(p - q)$
$Var(A_i) = Var(A_j) = Var(A) = a^2(p^2 + q^2) - a^2(p - q)^2 = 2a^2pq$

	IBD=0		$G_j(A_j)$		
G_i	$\mathbb{P}$	A_i	0 (-a)	1 (0)	2(a)
0	q^2	$-a$	q^2	$2pq$	p^2
1	$2pq$	0	q^2	$2pq$	p^2
2	p^2	a	q^2	$2pq$	p^2

	IBD=1		$G_j(A_j)$		
G_i	$\mathbb{P}$	A_i	0 (-a)	1 (0)	2(a)
0	q^2	$-a$	q	p	0
1	$2pq$	0	$q/2$	$1/2$	$p/2$
2	p^2	a	0	q	p

	IBD=2		$G_j(A_j)$		
G_i	$\mathbb{P}$	A_i	0 (-a)	1 (0)	2(a)
0	q^2	$-a$	1	0	0
1	$2pq$	0	0	1	0
2	p^2	a	0	0	1

$Cov(A_i, A_j) = k_0 \times 0 + k_1 Var(A)/2 + k_2 Var(A) = 2\theta Var(A)$

$$\begin{aligned}
&\mathbb{E}(A_i, A_j|IBD = 0) \\
&= q^2q^2a^2 - q^2p^2a^2 - p^2q^2a^2 + p^2p^2a^2 \\
&= a^2(q^4 - 2p^2q^2 + p^4) = a^2(p^2 - q^2)^2 \\
&= a^2(p - q)^2 \\
&\mathbb{E}(A_i, A_j|IBD = 0) - \mathbb{E}(A)^2 = 0
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}(A_i, A_j|IBD = 1) \\
&= q^2qa^2 + p^2pa^2 \\
&= a^2(p^3 + q^3) \\
&\mathbb{E}(A_i, A_j|IBD = 1) - \mathbb{E}(A)^2 \\
&= a^2(p^3 + q^3 - (p - q)^2) = pq = V(A_i)/2
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}(A_i, A_j|IBD = 2) \\
&= q^2a^2 + p^2a^2 \\
&= a^2(p^2 + q^2)\ \mathbb{E}(A_i, A_j|IBD = 2) - \mathbb{E}(A)^2 \\
&= a^2(p^2 + q^2 - (p - q)^2) = 2pq = V(A_i)
\end{aligned}$$

The Covariance Equation for Binary Traits

We can use the same strategy for binary traits.

Let $Y = 0$ indicate a control (healthy individual), $Y = 1$ a case (diseased individual)

The prevalence of the disease (population average) is

$$K = \mathbb{E}(Y) = \mathbb{P}(Y = 1)$$

We still use the same model for covariance:

$$Cov(Y_i, Y_j) = 2\theta Var(A) + k_2 Var(D) + \gamma Var(C)$$

Now, $Cov(Y_i, Y_j) = \mathbb{E}(Y_i Y_j) - K^2$

$$= \mathbb{P}(Y_i = 1, Y_j = 1) - K^2$$

Recurrence risk ratio

Given a diseased individual, the recurrence risk ratio λ_r is the probability a family member with relationship r is also diseased.

Most common is sibling relative risk, λ_S

Cancer	1st Degree	2nd Degree	3rd Degree	4th Degree	5th Degree
Breast	2.02	1.36	1.21	1.13	1.05
Prostate	1.89	1.36	1.19	1.10	1.10
Lung	2.00	1.39	1.10	1.02	1.04
Kidney	2.30	1.31	1.30	1.08	1.11

Amundadottir et al. PLoS Med. (2004)

Covariance between individuals

$$\lambda_r = \frac{\mathbb{P}(Y_i = 1 | Y_j = 1)}{K}$$

$$= \frac{\mathbb{P}(Y_i = 1, Y_j = 1)}{K^2}$$

Leads to:

$$\lambda_r = \frac{Cov(Y_i, Y_j) + K^2}{K^2}$$

$$= 1 + \frac{2\theta Var(A) + k_2 Var(D) + \gamma Var(C)}{K^2}$$

When $k_2 = 0$ and $\gamma = 0$:

$$\lambda_r = 1 + r \times constant$$

So with every extra degree of relationship, distance from 1 halves

Proving Haseman Elston Regression

Construct the vectors $Z = (Y_i - Y_j)^2$ and $X = K_{i,j}$. Each has length ${}^nC_2$

$$\begin{aligned}\mathbb{E}(Z) &= \mathbb{E}((Y_i - Y_j)^2) \\ &= \mathbb{E}(Y_i^2 + Y_j^2 - 2Y_iY_j)\end{aligned}$$

Without loss of generality, assume $\mathbb{E}(Y) = 0$, then

$$\begin{aligned}\mathbb{E}(Z) &= Var(Y) + Var(Y) - 2Cov(Y_i, Y_j) \\ &= 2Var(Y) - 2K_{i,j}Var(A)\end{aligned}$$

Therefore, when fitting the model $Z = \alpha + \beta X$,

$$\alpha = 2Var(Y) \text{ and } \beta = -2Var(A)$$

Therefore, $-\alpha/\beta$ is an estimate of $h^2 = Var(A)/Var(Y)$