Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 16: Introduction to GWAS

Jason Mezey
March 19, 2024 (T) 8:40-9:55

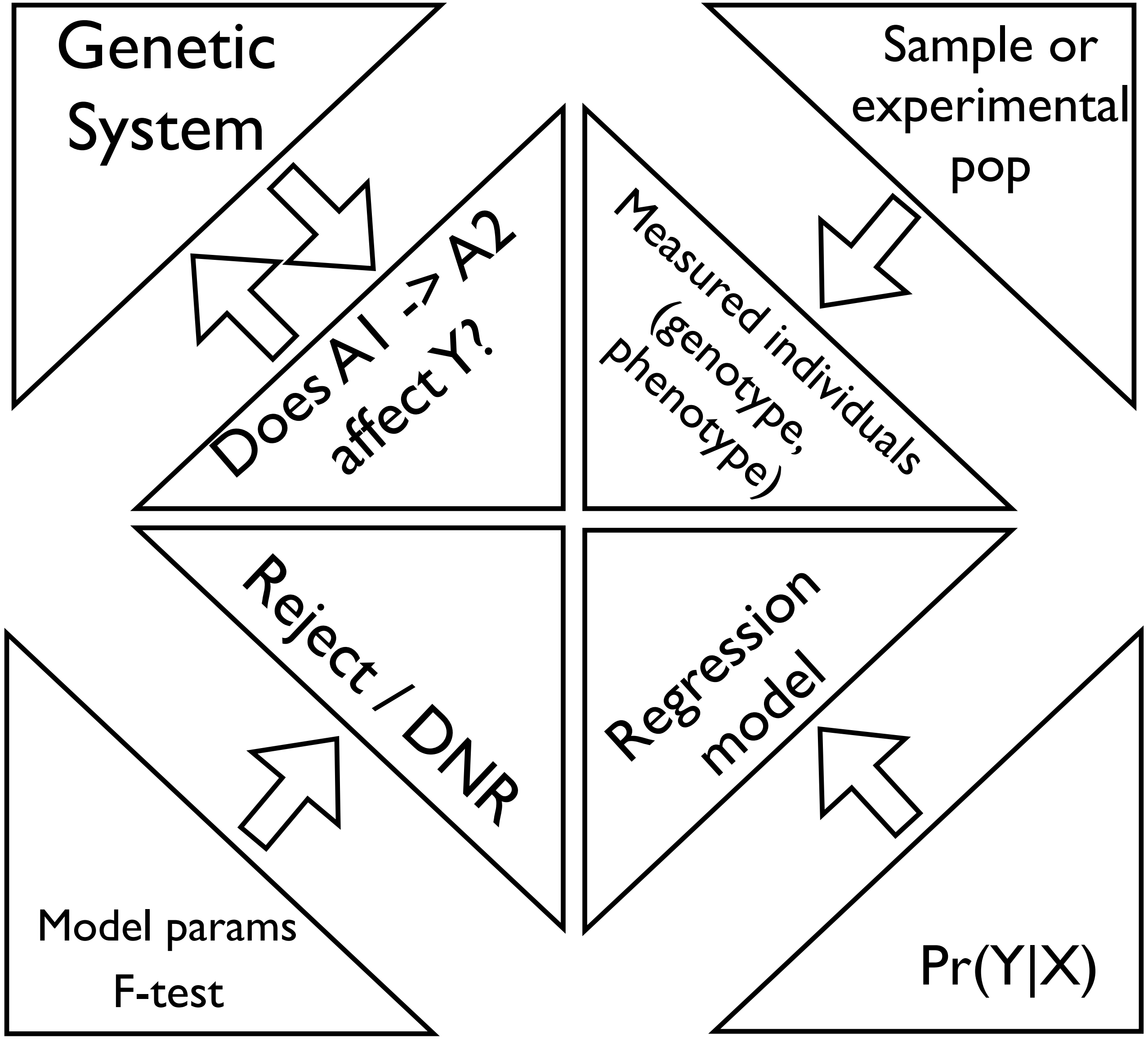
Announcements

- REMINDER: I will be lecturing by zoom this coming Thurs (March 21) and next week (March 26 and 28)
- There will be NO OFFICE HOURS THIS WEEK (March 20) (!!) but we will have office hours next week (March 25) as regularly scheduled
- Homework #4 (last homework!) has been posted and is due by 11:59PM Friday March 29
- Your midterm (!!) will be the week of April 8 (more information to follow in lecture this coming Thurs)

Summary of lecture 16: Introduction to GWAS

- Last lecture, we completed our introduction to inference for genetic (Regression!) models - specifically MLE and Hypothesis Testing using F-statistics!
- Today we will begin our introduction to Genome-Wide Association Studies (and associated critical concepts)

Conceptual Overview



Review: Genetic system

- **causal mutation -** a position in the genome where an experimental manipulation of the DNA would produce an effect on the phenotype under specifiable conditions
- Formally, we may represent this as follows:

$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- Note: that this definition considers "under specifiable" conditions" so the change in genome need not cause a difference under every manipulation (just under broadly specifiable conditions)
- Also note the symmetry of the relationship
- Identifying these is the core of quantitative genetics/genomics (why do we want to do this!?)
- What is the perfect experiment?
- Our experiment will be a statistical experiment (sample and inference!)

Review: Genetic probability model

- Remember that we define the random variables we need for our genetic model by

$$Y : (*, \Omega_P) \rightarrow \mathbb{R}$$

$$X : (\Omega_g, *) \rightarrow \mathbb{R}$$

- Where we have three possible genotypes:

$$\Omega_g = \{A_1A_1, A_1A_2, A_2A_2\}$$

- The quantitative genetic model is a "multiple" regression model with the following TWO independent ("dummy") *X* variables:

$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

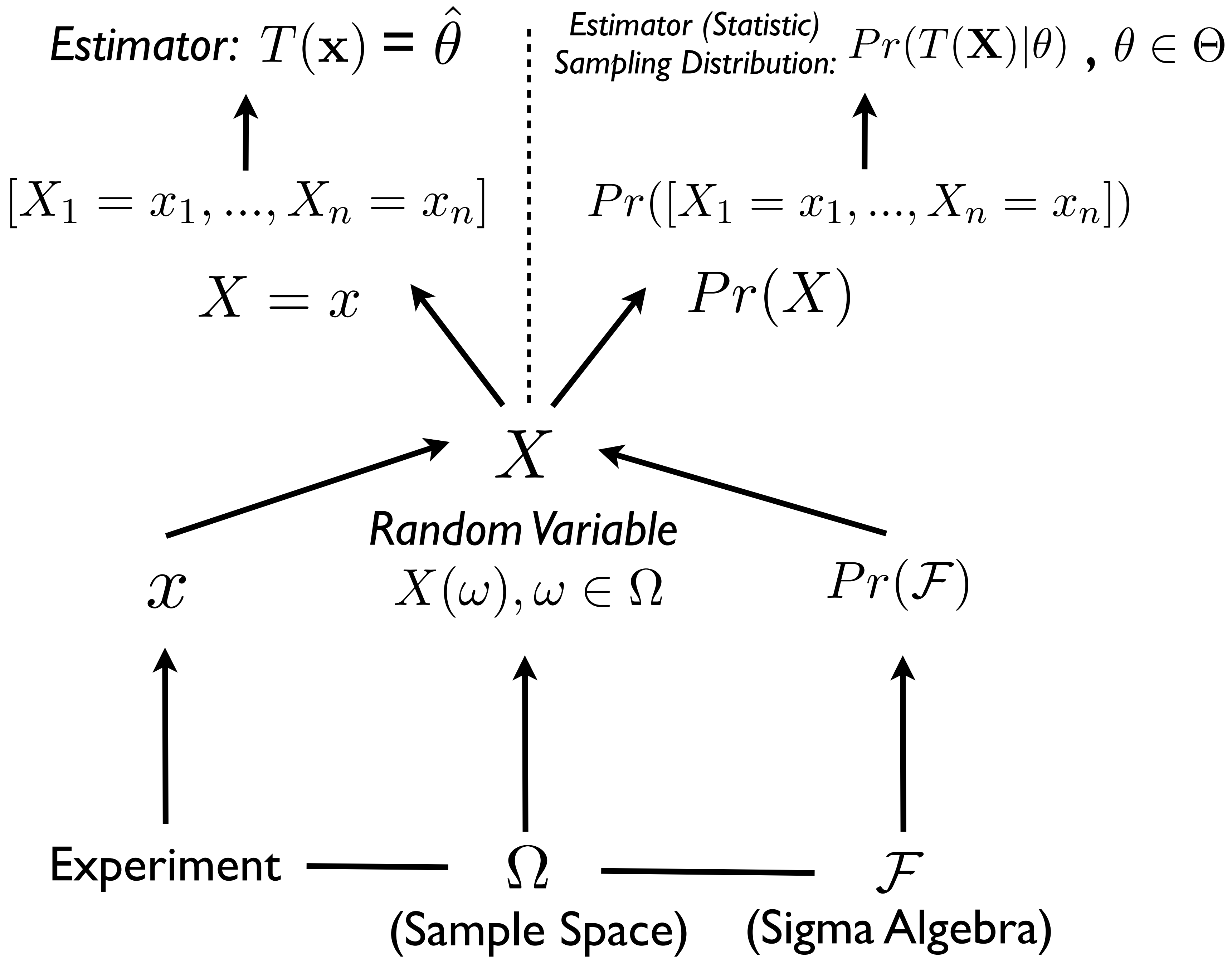
1		A_1A_2	
-1	A_1A_1		A_2A_2
	-1	0	1

- and the following "multiple" regression equation:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

Estimators



Estimator: $T(\mathbf{x}) = \hat{\theta}$

Estimator (Statistic) Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$

$[X_1 = x_1, ..., X_n = x_n]$

$Pr([X_1 = x_1, ..., X_n = x_n])$

$X = x$

$Pr(X)$

X

Random Variable

x

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

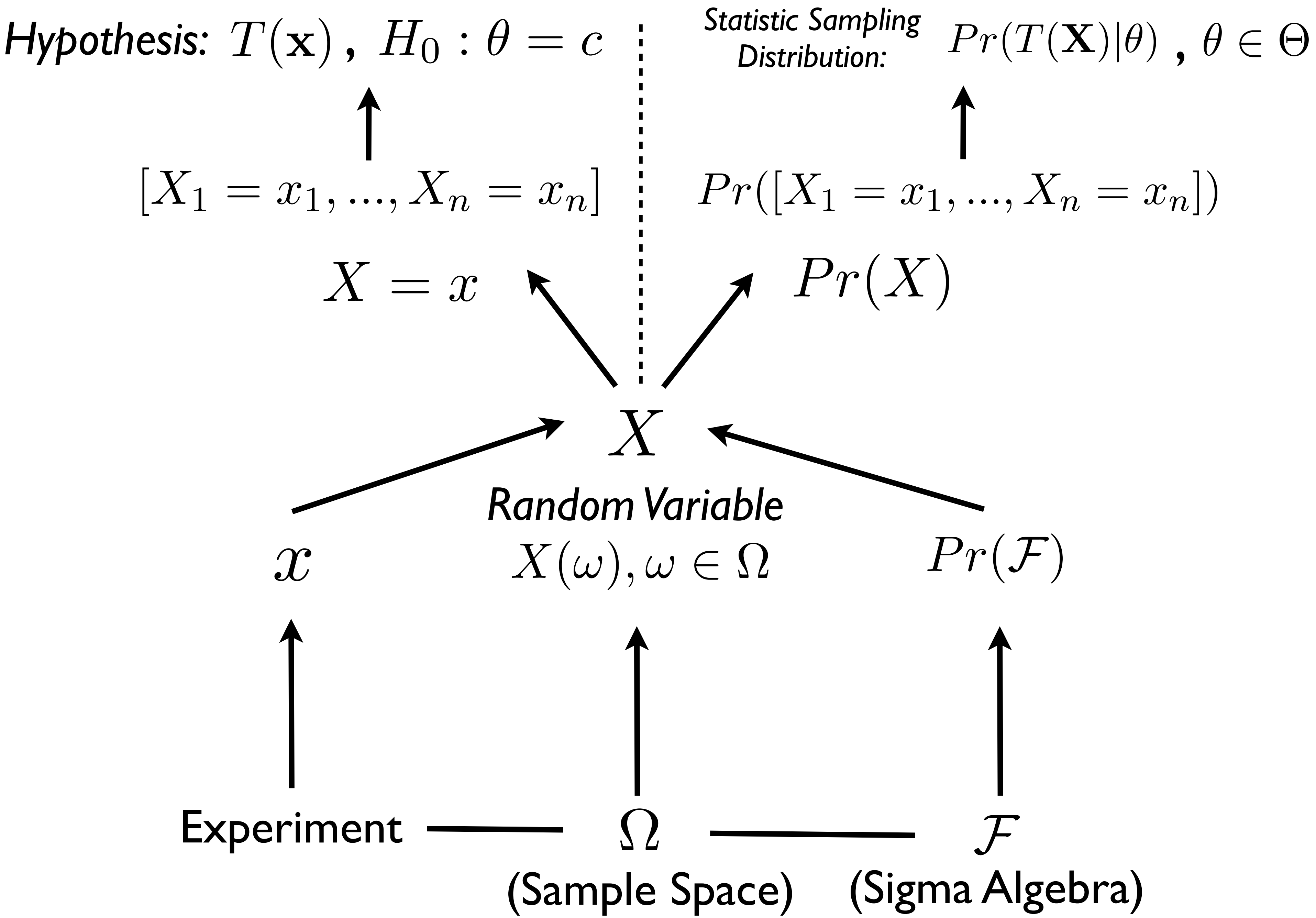
Ω

(Sample Space)

$\mathcal{F}$

(Sigma Algebra)

Hypothesis Tests



Hypothesis: $T(\mathbf{x})$, $H_0 : \theta = c$

Statistic Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$

$[X_1 = x_1, ..., X_n = x_n]$

$Pr([X_1 = x_1, ..., X_n = x_n])$

$X = x$

$Pr(X)$

X

Random Variable

x

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

Ω

(Sample Space)

$\mathcal{F}$

(Sigma Algebra)

Review: Genetic inference I

- Recall that inference (whether estimation or hypothesis testing) starts by collecting a sample and defining a statistic on that sample

- In this case, we are going to collect a sample of n individuals where for each we will measure their *phenotype* and their *genotype* (i.e. at the locus we are focusing on)

- That is an individual i will have phenotype y_i and genotype $g_i = A_j A_k$ (where we translate these into x_a and x_d)

- Using the phenotype and genotype we will construct both an *estimator* (a statistic!) and we will additionally construct a *test statistic*

- Remember that our regression probability model defines a sampling distribution on our sample and therefore on our estimator and test statistic (!!)

Review: Genetic inference II

- For notation convenience, we are going to use vector / matrix notation to represent a sample:

$$y_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d + \epsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_\mu + x_{1,a}\beta_a + x_{1,d}\beta_d + \epsilon_1 \\ \beta_\mu + x_{2,a}\beta_a + x_{2,d}\beta_d + \epsilon_2 \\ \vdots \\ \beta_\mu + x_{n,a}\beta_a + x_{n,d}\beta_d + \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,a} & x_{1,d} \\ 1 & x_{2,a} & x_{2,d} \\ \vdots & \vdots & \ddots \\ 1 & x_{n,a} & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

Review: Genetic estimation I

- We will define a MLE for our parameters:

$$\beta = [\beta_{\mu}, \beta_{a}, \beta_{d}]$$

- Recall that an MLE is simply a statistic (a function that takes a sample in and outputs a number that is our estimate)

- In this case, our statistic will be a vector valued function that takes in the vectors that represent our sample

$$T(\mathbf{y}, \mathbf{x}_a, \mathbf{x}_d) = \hat{\beta} = [\hat{\beta}_{\mu}, \hat{\beta}_{a}, \hat{\beta}_{d}]$$

- Note that we calculate an MLE for this case just as we would any case (we use the likelihood of the fixed sample where we identify the parameter values that maximize this function)

- In the linear regression case (just as with normal parameters) this has a closed form:

$$MLE(\hat{\beta}) = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

Review: Genetic estimation II

- Let's look at the structure of this estimator:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,a} & x_{1,d} \\ 1 & x_{2,a} & x_{2,d} \\ \vdots & \vdots & \ddots \\ 1 & x_{n,a} & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$MLE(\hat{\beta}) = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

$$MLE(\hat{\beta}) = \begin{bmatrix} \hat{\beta}_\mu \\ \hat{\beta}_a \\ \hat{\beta}_d \end{bmatrix}$$

Review: hypothesis testing I

- We now have everything we need to construct a hypothesis test for:

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- This is equivalent to testing the following:

$$H_0 : Cov(X, Y) = 0$$

- For a linear regression, we use the F-statistic for our sample:

$$F_{[2,n-3]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}) = \frac{MSM}{MSE}$$

- We then determine a p-value using the distribution of the F-statistic under the null:

$$pval(F_{[2,n-3]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}))$$

Review: hypothesis testing II

- To construct our LRT for our null, we will need several components, first the predicted value of the phenotype for each individual:

$$\hat{y}_i = \hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d$$

- Second, we need the "Sum of Squares of the Model" (SSM) and the "Sum of Squares of the Error" (SSE):

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \qquad SSE = \sum_{n=1}^{n} (y_i - \hat{y}_i)^2$$

- Third, we need the "Mean Squared Model" (MSM) and the "Mean Square Error" (MSE) with degrees of freedom (df) $df(M) = 3 - 1 = 2$ and : $df(E) = n - 3$

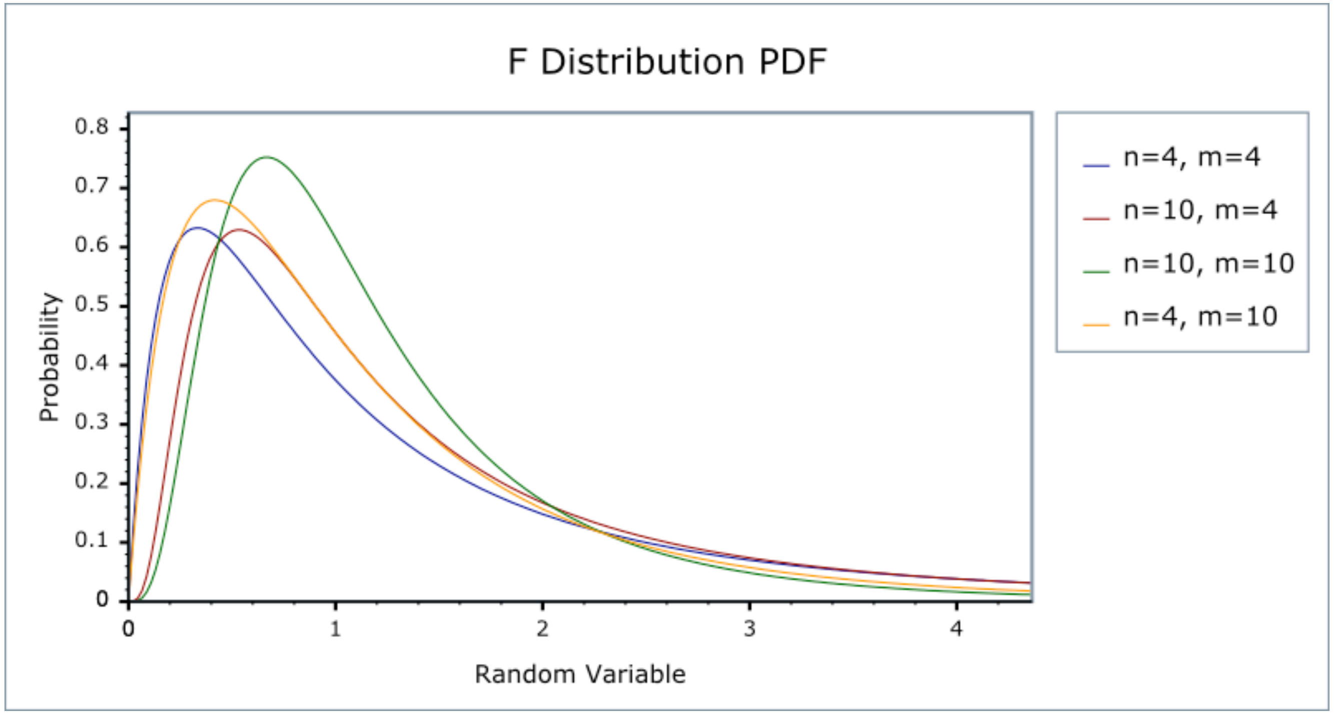
$$MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \qquad MSE = \frac{SSE}{df(E)} = \frac{SSE}{n-3}$$

- Finally, we calculate our (LRT!) statistic, the F-statistic with degrees of freedom [2, n-3]:

$$F_{[2,n-3]} = \frac{MSM}{MSE}$$

Review: hypothesis testing IV

- In general, the F-distribution (continuous random variable!) under the H0 has variable forms that depend on d.f.:



- Note when calculating a p-value for the genetic model, we consider the value of the F-statistic we observe or more extreme towards positive infinite (!!) using the F-distribution with [2,n=3] d.f.

- However, also this is actually a two-tailed test (what is going on here (!?)

Review: hypothesis testing V

- An F-statistic is a Likelihood Ratio Test (LRT) statistic after a simple (monotonic) transformation

$$F\text{-statistic} = f(\Lambda)$$

- Note that an F-statistic has an exact pdf under many conditions (note that we do not always produce a LRT that has an exact pdf that we can state easily)
- Also note that a t-test is actually an F-statistic (and therefore a transformed LRT) for a case where we are comparing the means of just two groups (when might this apply in genetic testing!?), similarly for a t-test of the slope of a regression)

Side-topic: Alternative (ANOVA) formulation I

- Note that we can construct an equivalent formulation to our *linear regression* using an *ANOVA* coding
- ANOVA stands for *ANalysis Of VAriance* and, despite the name, it is really a test of whether "means" of groups are different
- A genetic ANOVA model is the same as our linear regression, except the "dummy" variables are coded differently (everything else is the same!)

Side-topic: Alternative (ANOVA) formulation II

- Remember the independent (dummy) variable coding for a regression is:

$$X_{\mu}(A_1A_1) = 1, X_{\mu}(A_1A_2) = 1, X_{\mu}(A_2A_2) = 1$$

$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

- The ANOVA coding is the following:

$$X_{A_1A_1}(A_1A_1) = 1, X_{A_1A_1}(A_1A_2) = 0, X_{A_1A_1}(A_2A_2) = 0$$

$$X_{A_1A_2}(A_1A_1) = 0, X_{A_1A_2}(A_1A_2) = 1, X_{A_1A_2}(A_2A_2) = 0$$

$$X_{A_2A_2}(A_1A_1) = 0, X_{A_2A_2}(A_1A_2) = 0, X_{A_2A_2}(A_2A_2) = 1$$

- The models corresponding to a linear regression and ANOVA are:

$$Y = X_{\mu}\beta_{\mu} + X_a\beta_a + X_d\beta_d + \epsilon$$

$$Y = X_{A_1A_1}\beta_{A_1A_1} + X_{A_1A_2}\beta_{A_1A_2} + X_{A_2A_2}\beta_{A_2A_2} + \epsilon$$

Side-topic: Alternative (ANOVA) formulation III

- For the ANOVA formulation, the parameters are:

$$\theta = [\beta_{A_1A_1}, \beta_{A_1A_2}, \beta_{A_2A_2}]$$

- And we test the null hypothesis:

$$H_0 : \beta_{A_1A_1} = \beta_{A_1A_2} = \beta_{A_2A_2}$$

$$H_A : \beta_{A_jA_k} \neq \beta_{A_lA_m} \qquad jk \neq lm$$

- Note that estimation (MLE) and the hypothesis test (F-test) construction are the same (=same equations)!!

- Why would we use an ANOVA formulation (what is the difference)?

Quantitative genomic analysis I

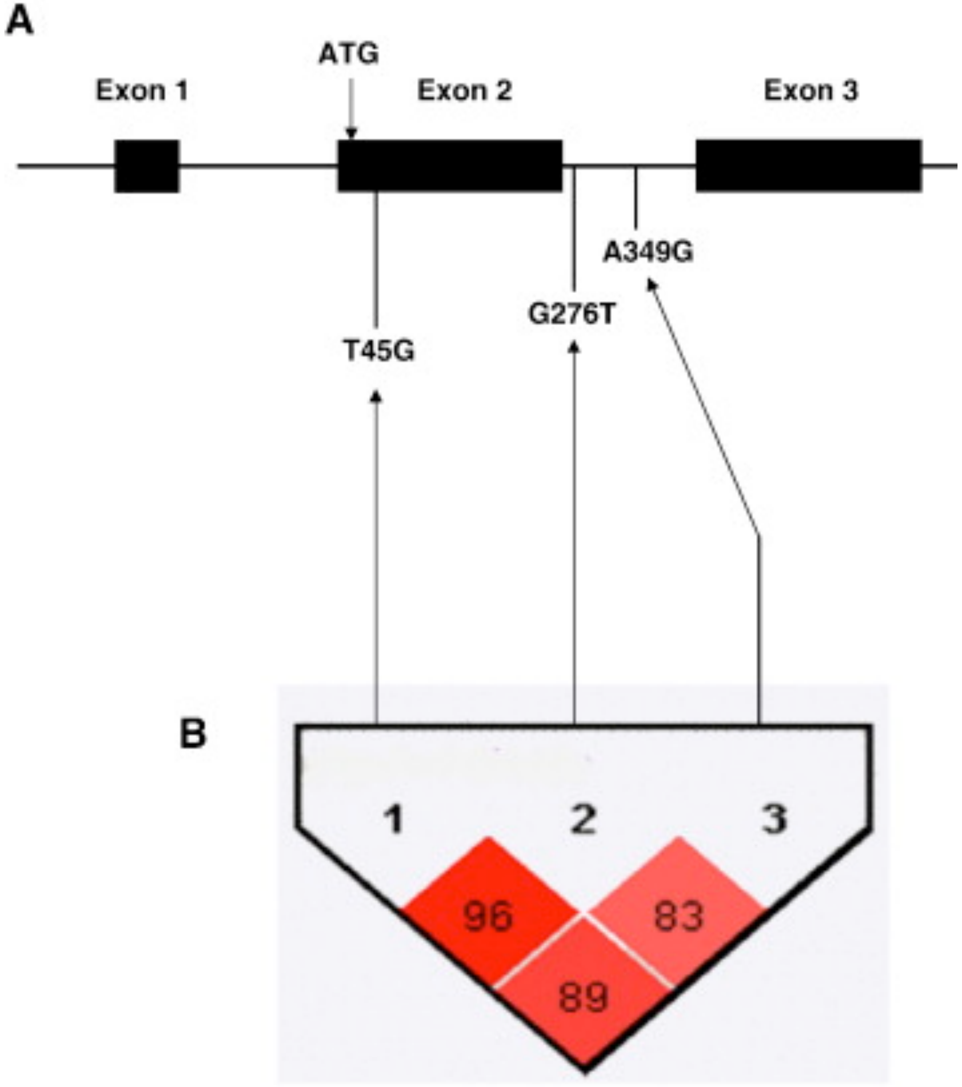
- We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype
- Occasionally we will assess this hypothesis for a single genotype
- In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome
- We therefore perform a hypothesis test of *many genotypes throughout the genome*
- This is a genome-wide association study (GWAS)

Quantitative genomic analysis II

- Analysis in a GWAS raises (at least) two issues we have not yet encountered:
  - An analysis will consist of many hypothesis tests (not just one)
  - We often do not test the causal polymorphism (usually)
- Note that this latter issue is a bit strange (!?) - how do we assess causal polymorphisms if we have not measured the causal polymorphism?
- Also note that causal genotypes will begin to be measured in our GWAS with next-generation sequencing data (but the issue will still be present!)

Correlation among genotypes

- If we test a (non-causal) genotype that is correlated with the causal genotype AND if correlated genotypes are in the same position in the genome THEN we can identify the genomic position of the casual genotype (!!)
- This is the case in genetic systems (why!?)
- Do we know which genotype is causal in this scenario?

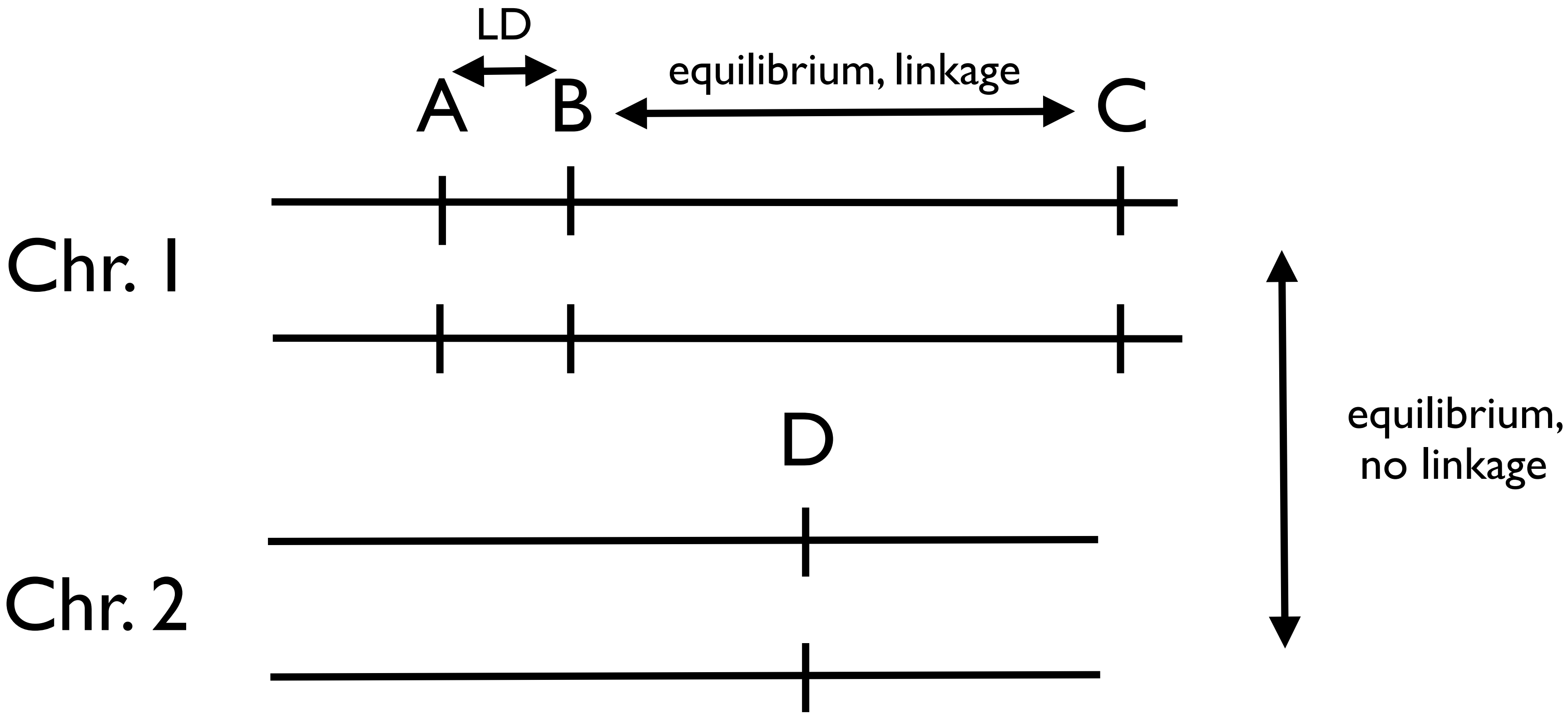


Copyright: Journal of Diabetes and its Complications; Science Direct;Vendramini et al

Linkage Disequilibrium

- Mapping the position of a causal polymorphism in a GWAS requires there to be LD for genotypes that are both physically linked and close to each other AND that markers that are either far apart or on different chromosomes to be in equilibrium

- Note that dis*equilibrium* includes both *linkage disequilibrium* AND other types of dis*equilibrium* (!!), e.g. gametic phase disequilibrium



That’s it for today

- Next lecture, we will continue our discussion of GWAS analysis (!!)