Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 15: Genetic Model Estimation and Hypothesis Testing

Jason Mezey

March 14, 2024 (Th) 8:40-9:55

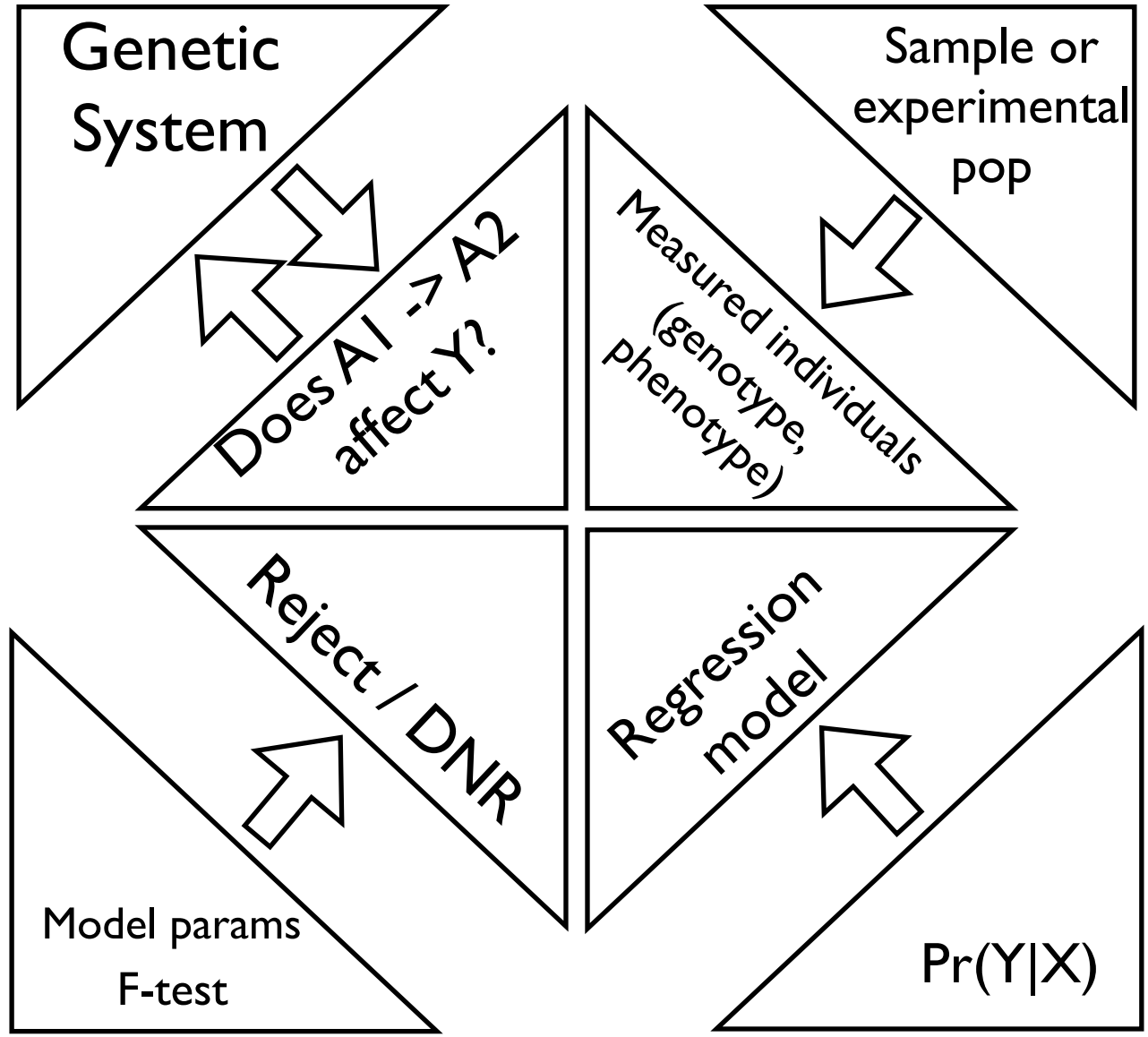
Announcements

- A Homework #2 Key "V2" now posted (an error has been corrected)
- I have updated the syllabus (!!) where please note
  - We will have one more homework (#4) that will be available next week
  - Your "midterm" will be AFTER Cornell, Ithaca Spring break (available April 9) - more to come on this in the coming weeks…
- The next two weeks (March 18 and March 25) I will be lecturing by zoom for BOTH Tues and Thurs lectures
  - You are welcome to join by zoom
  - I will be projecting to the Cornell (Ithaca) classrooms as usual
  - I will be projecting to Weill Cornell Med (NYC) classrooms as (now updated) on the syllabus (PLEASE NOTE: we do not have an NYC classroom March 19 - please join by zoom!)

Summary of lecture 15: Genetic Model Hypothesis Testing

- Last lecture, we completed our introduction to Regression models (=families of probability models!)
- Today we will discuss how to do inference for these models - specifically MLE and Hypothesis Testing using F-statistics!

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

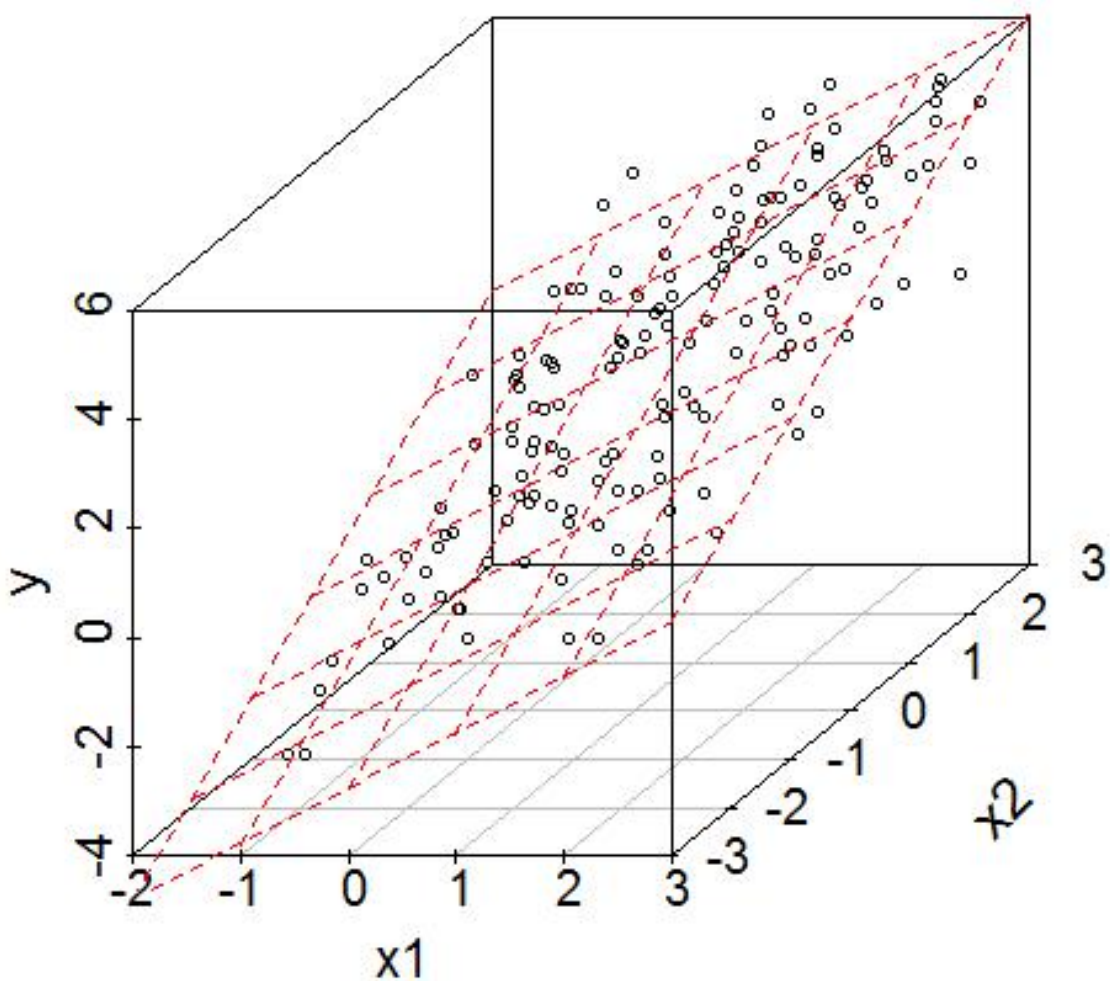
- and the following "multiple" regression equation:

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

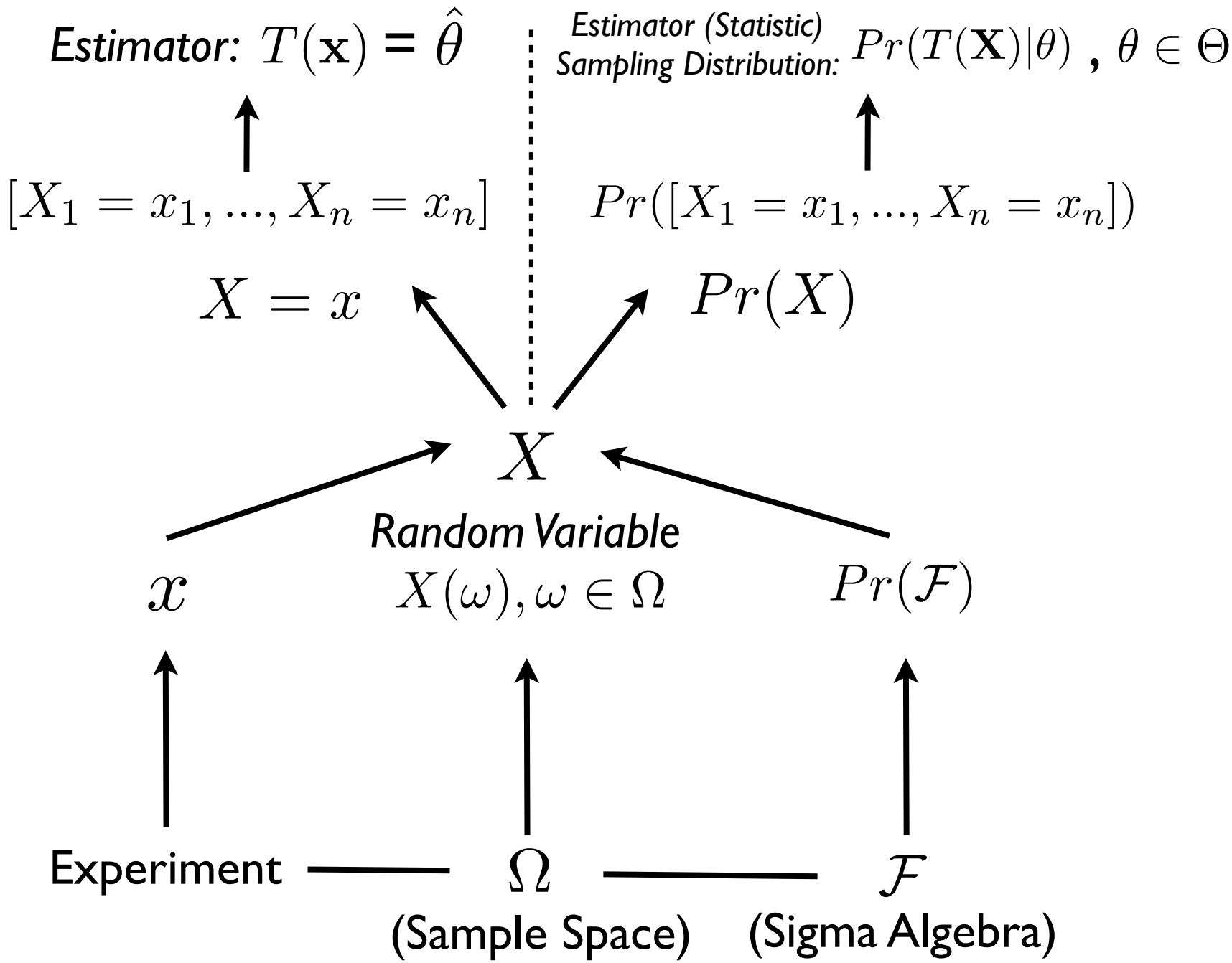
$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

Linear regression IV

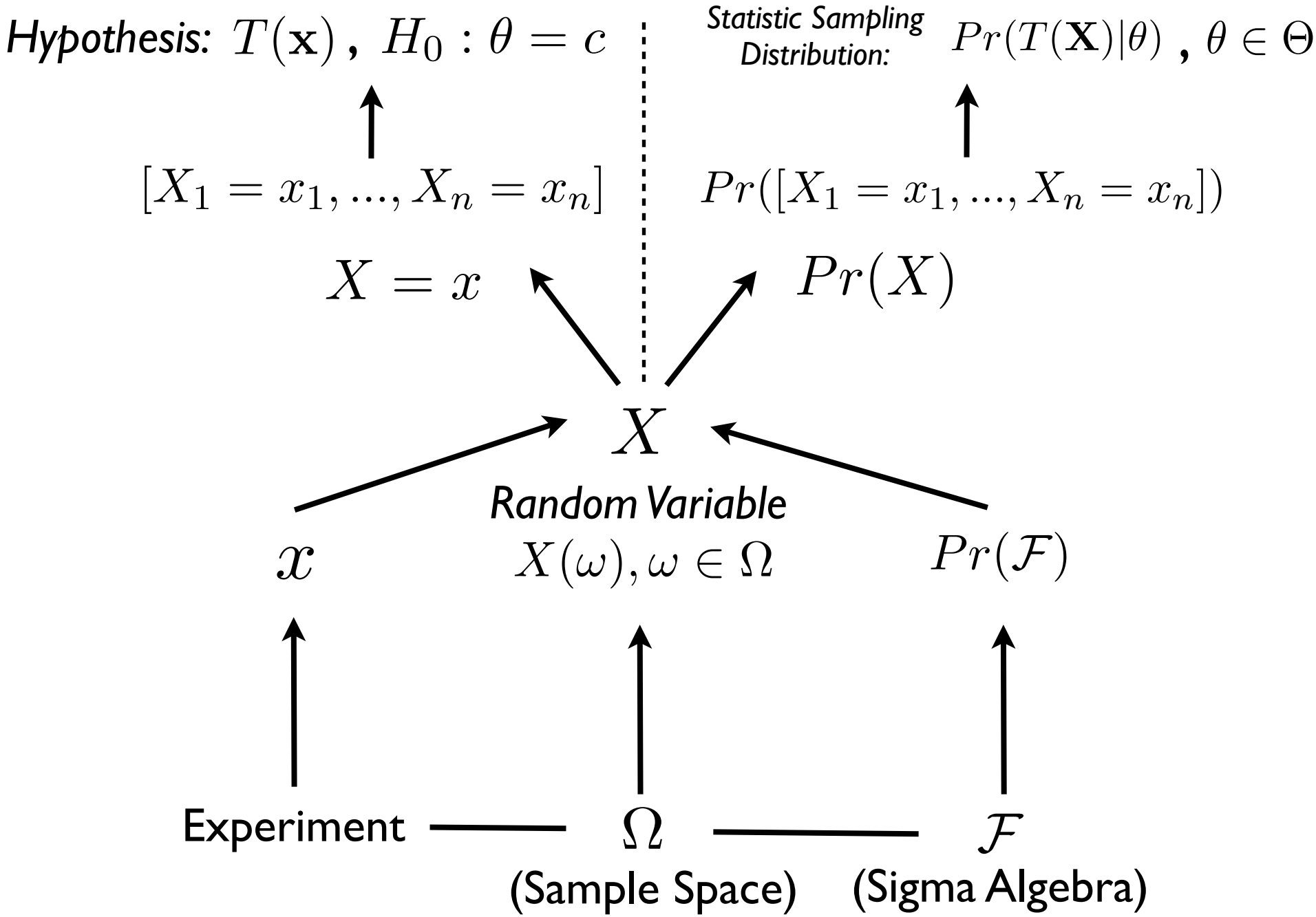
- A *multiple regression* model has the same structure, with a single dependent variable *Y* and more than one independent variable X_i, X_j, e.g.,



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
$X(\omega), \omega \in \Omega$
x
$Pr(\mathcal{F})$
Experiment
Ω
$\mathcal{F}$
(Sample Space)
(Sigma Algebra)

Hypothesis Tests



Review: Genetic probability model

- For set of genotypes at a position taking states A1A1, A1A2, A2A2 (given our *X*a and *X*d coding) for the TRUE model (!!):
  - If $\beta_a \neq 0$ AND / OR $\beta_d \neq 0$ then the genotypes (polymorphism, alleles, mutation) are causal (!!)
  - If β_a = 0 AND β_d = 0 there is no effect of the genotype on the phenotype (the genotype is not causal!)
- There our H0 of interest is β_a = 0 AND β_d = 0 (!!)

Review: Genetic inference I

- For our model focusing on one locus:

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

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

- We have four possible parameters we could estimate:

$$\theta = \left[\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon\right]$$

- However, for our purposes, we are only interested in the genetic parameters and testing the following null hypothesis:

$$H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0$$
$$H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0$$

OR

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$
$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

Review: Genetic inference II

- 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 (!!)

Matrix Basics

$$\mathbf{v} = \vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \mathbf{M}_1 = \bar{M}_1 = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \qquad \mathbf{M}_2 = \bar{M}_2 = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

We will also follow statistics convention where the first subscript will index rows and the second will index columns (note this is usually reversed in mathematics literature).

Matrix sum: $\mathbf{M}_1 + \mathbf{M}_1 = \begin{bmatrix} m_{11} + m_{11} & m_{12} + m_{12} \\ m_{21} + m_{21} & m_{22} + m_{22} \end{bmatrix}$

Matrix transpose: $\mathbf{M}_2^{\mathrm{T}} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$

Scalar times a matrix: $c\mathbf{M}_1 = \begin{bmatrix} cm_{11} & cm_{12} \\ cm_{21} & cm_{22} \end{bmatrix}$

Matrix multiplication:

$$\mathbf{M}_1\mathbf{M}_1 = \begin{bmatrix} m_{11}m_{11} + m_{12}m_{21} & m_{11}m_{12} + m_{21}m_{22} \\ m_{21}m_{11} + m_{22}m_{21} & m_{21}m_{12} + m_{22}m_{22} \end{bmatrix} \quad \mathbf{M}_2\mathbf{M}_1 = \begin{bmatrix} am_{11} + dm_{21} & am_{12} + dm_{22} \\ bm_{11} + em_{21} & bm_{12} + em_{22} \\ cm_{11} + fm_{21} & cm_{12} + fm_{22} \end{bmatrix}$$

$$\mathbf{v}\mathbf{v}^{\mathrm{T}} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} v_1v_1 & v_1v_2 \\ v_2v_1 & v_2v_2 \end{bmatrix}, \quad \mathbf{v}^{\mathrm{T}}\mathbf{v} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1v_1 + v_2v_2$$

If the following holds: $\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = \begin{bmatrix} v_1 & v_2 \end{bmatrix}\begin{bmatrix} v_3 \\ v_4 \end{bmatrix} = 0$ then $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal.

The identity matrix is defined as follows: $\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, i.e. diagonal elements are “1” and all other elements are “0”.

The inverse of a matrix $\mathbf{M}^{-1}$ has a structure such that is satisfies the following relationship (for a “square”, *k* x *k* matrix): $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$ and $\mathbf{M}^{-1}\mathbf{M} = \mathbf{I}$.

Genetic inference III

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

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

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

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

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

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:

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

Genetic estimation II

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

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

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

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

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

Genetic hypothesis testing I

- We are going to test the following hypothesis:

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

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

- To do this, we need to construct the following test statistic (for which we know the distribution!):

$$T(\mathbf{y}, \mathbf{x}_a, \mathbf{x}_d | H_0 : \beta_a = 0 \cap \beta_d = 0)$$

- Specifically, we are going to construct a likelihood ratio test (LRT)
- This is calculated using the same structure that we have discussed (i.e. ratio of likelihoods that take values of parameters maximized under the null and alternative hypothesis)
- In the case of a regression (not all cases!) we can write the form of the LRT for our null in an alternative (but equivalent!) form
- In addition, our LRT has an exact distribution for all sample sizes *n* (!!)

Genetic hypothesis testing II

- 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:

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

Genetic hypothesis testing III

- 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 : $df(E) = n - 3$

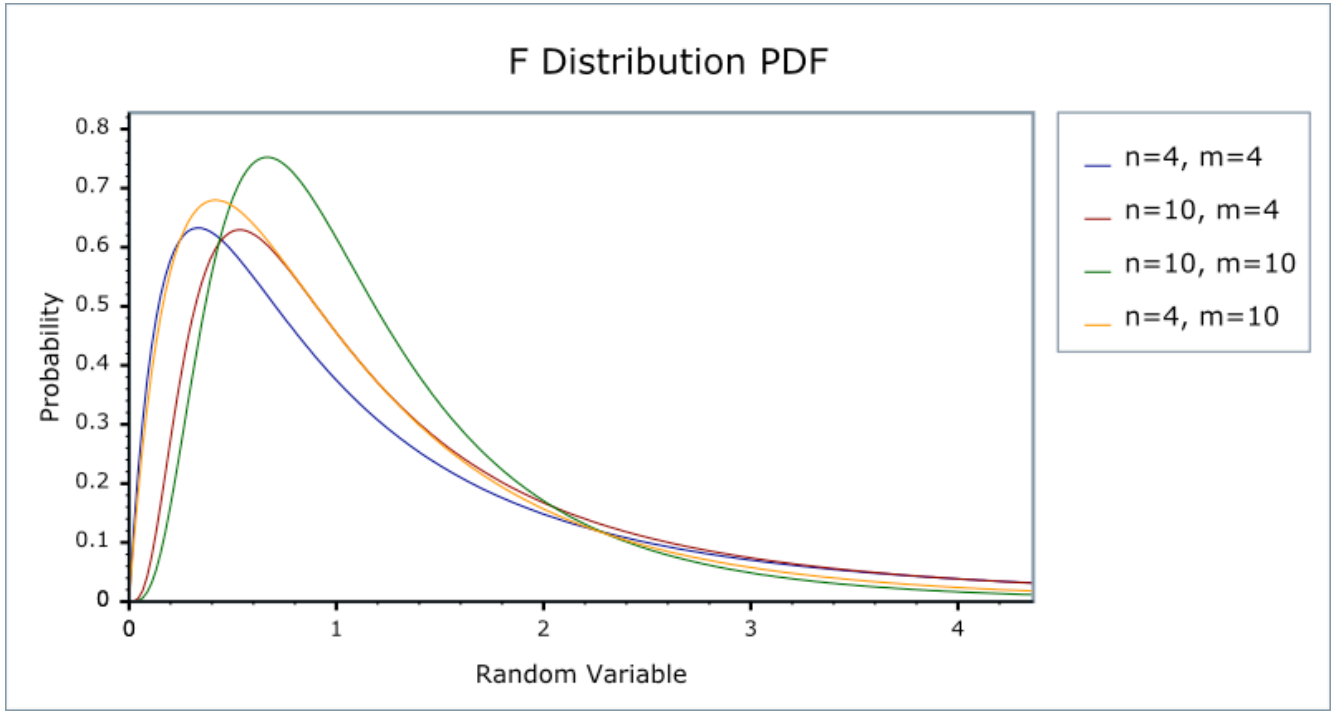
$$MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \qquad MSE = \frac{SSE}{df(E)} = \frac{SSE}{n-3}$$

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

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

Genetic 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 (!?)

Genetic 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)

That's it for today

- Next lecture, we introduce GWAS analysis (!!)