Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 13: Intro to Genetic Probability Models (Regression)

Jason Mezey
March 7, 2024 (Th) 8:40-9:55

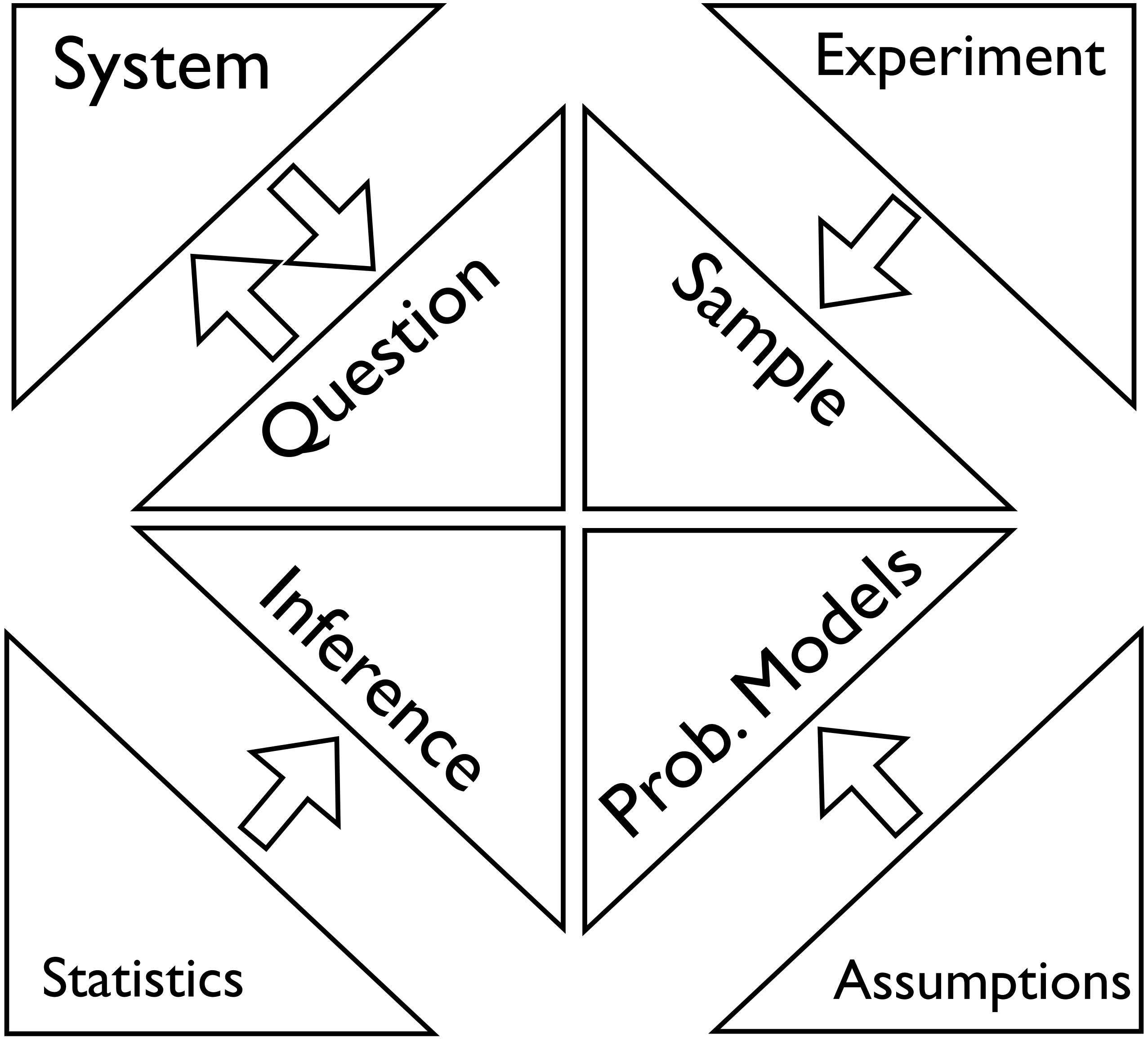
Announcements

- Typo in homework #3 - I WILL CORRECT THIS TODAY and I will announce when the correction is posted
- Homework #2 key will be up soon

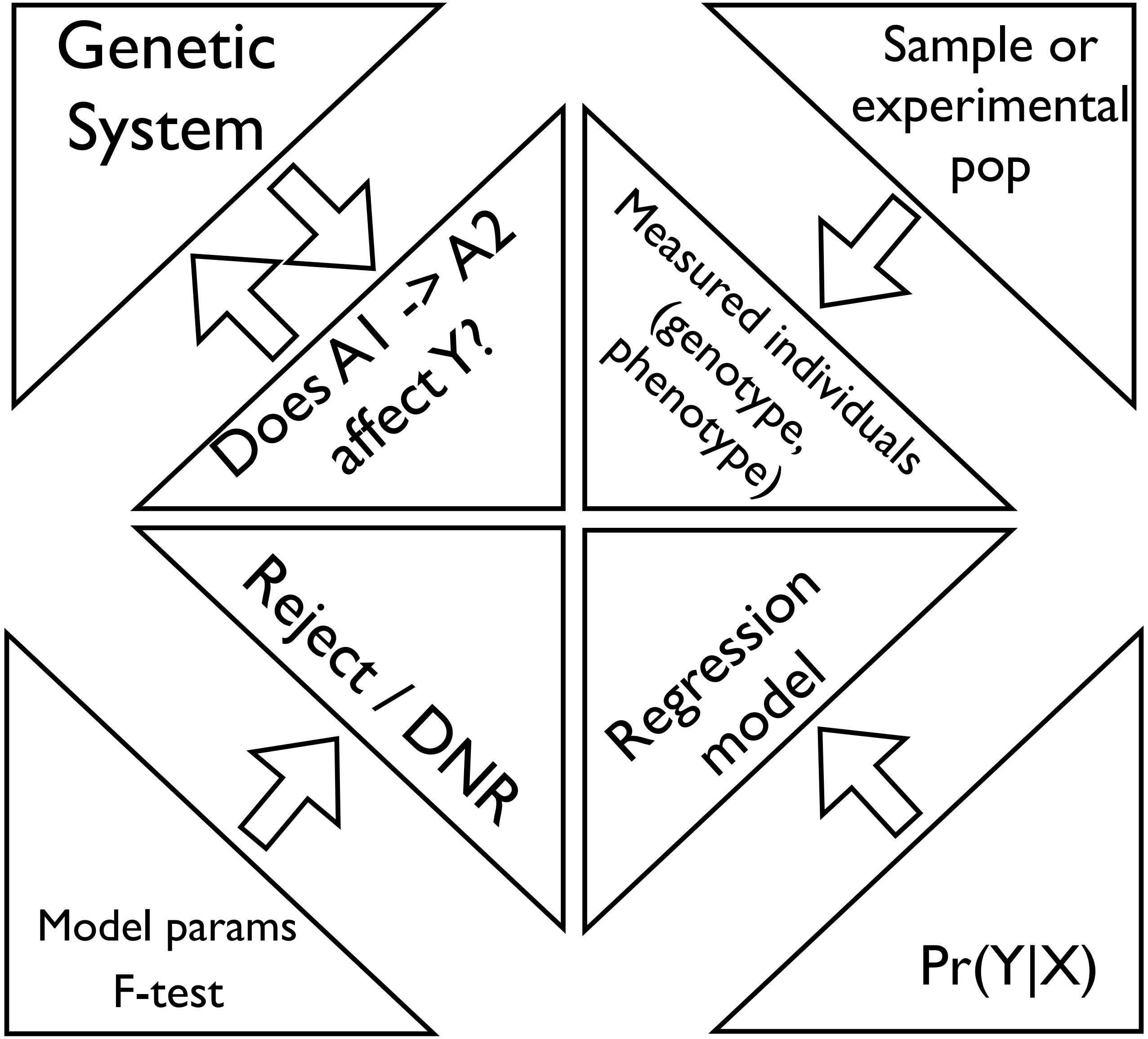
Summary of lecture 13: Genetic Probability Models

- Last lecture, we finished our discussion of Hypothesis Testing
- And began our introduction to Genetic Models (!!)
- Today we will continue the introduction with components needed for inference with genetic models!
- …which requires we introduce the Genetic Probability Models = Regressions (!!) - families of probability models!

Conceptual Overview



Conceptual Overview



Genetic System

Sample or experimental pop

Does A1 -> A2 affect Y?

Measured individuals (genotype, phenotype)

Reject / DNR

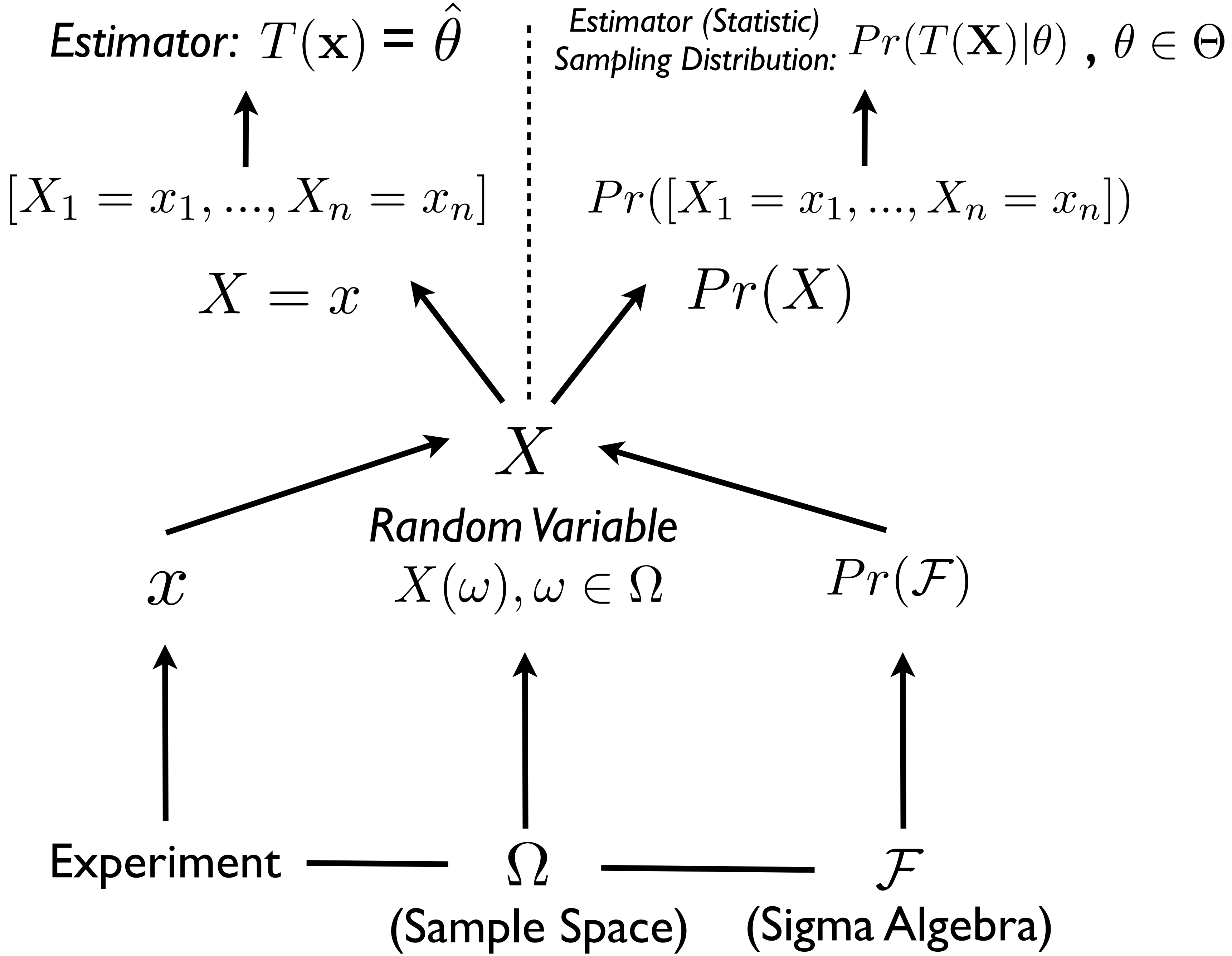
Regression model

Model params

F-test

Pr(Y|X)

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

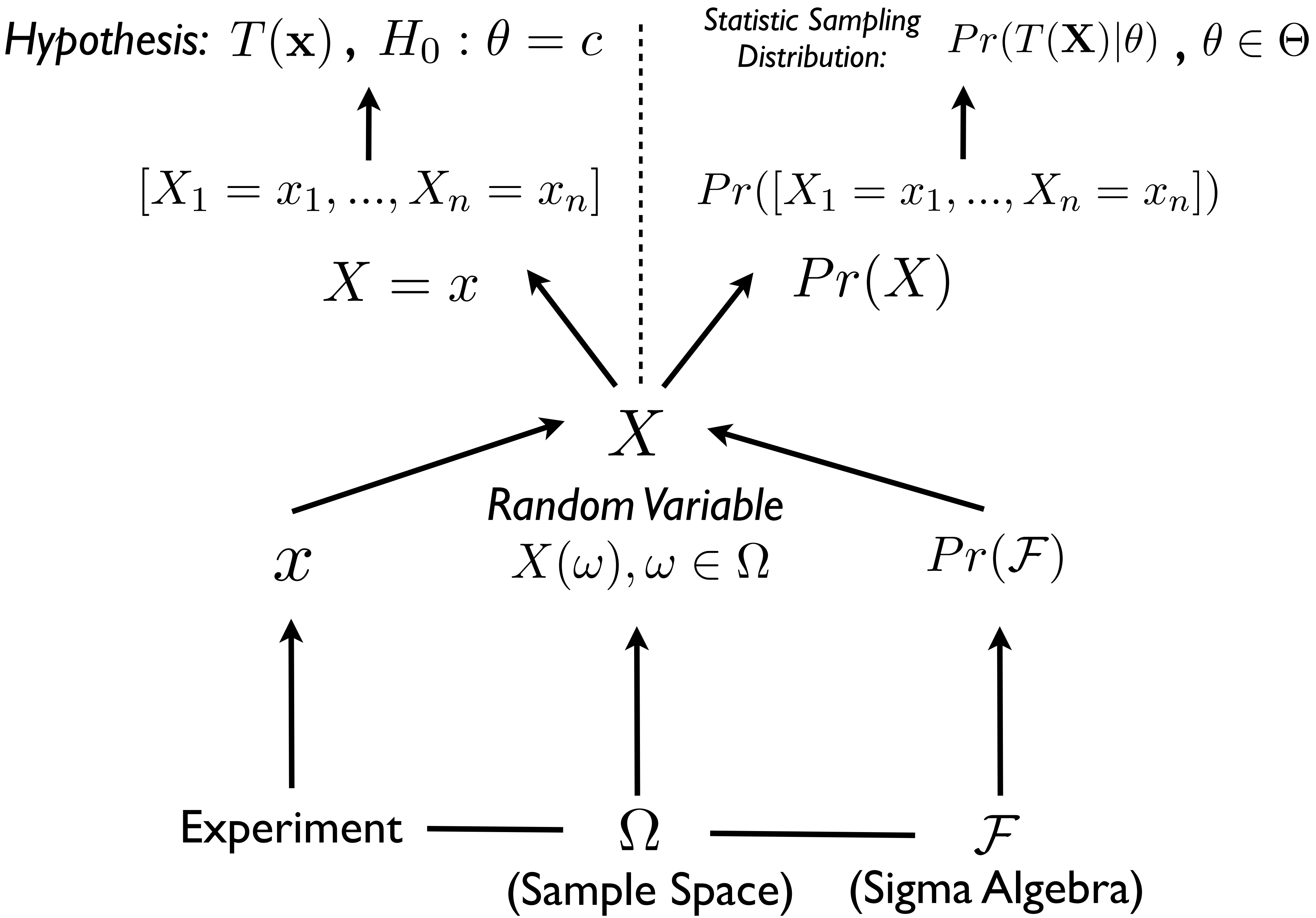
Ω

(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 system 1

- We will reduce the complexity of a genetic system to two components: the *genome* (the inherited DNA possessed by an individual) and the *phenotype* (an aspect we measure)
- In quantitative genetics we are interested in positions in the genome where differences produce a difference in phenotype
- These differences were originally a result of a *mutation*

Review: Genetic system II

- **mutation -** a change in the DNA sequence of a genome
- In a population of individuals (broadly defined), all differences in the genomes among the individuals were originally due to mutations
- Note: for our purposes, regardless of the cause of a mutation, we consider any difference produced in a genome that is passed on (or could be passed on) to the next generation to be a mutation
- For example, a SNP (Single Nucleotide Polymorphism; = A, G, C, T difference), Indels, microsatellites, etc.
- Also note that we will ignore the physical structure of a mutation (e.g. SNP, Indel, etc.) and quantify differences as A_i, A_j, etc.
- More specifically, we will be concerned with causal mutations, cases where the difference in genome is responsible for a difference in phenotype

Review: Genetic system III

- **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:

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

- 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: The statistical model 1

- We will make the following assumptions about the system:
  - At least one causal mutation affecting the phenotype of interest has occurred during the history of the population
  - At the locus (position) where the mutation occurred, there are at least two alleles (states of DNA) among individuals in the population (i.e. one is the original state, the other is the mutation)
- **polymorphism** - the existence of more than one allele at a locus
- These differences were originally a result of a *mutation*

Review: The statistical model II

- For most of this class, we will be discussing *diploid* systems (i.e. cases where individuals have two copies of a chromosome), which are *sexual* (i.e. offspring are produced that have a genome that is a copy of half of the mother's and half of the father's genome), and we will be considering polymorphisms that only have two alleles (e.g. A_1 and A_2)
- However, note that the formalism easily extends to ANY genetic system (bacteria, tetraploids, cancer, etc.)
- We are also largely going to consider a *natural experiment* (i.e. our sample will be selected from an existing set of individuals in nature), although again, the formalism extends to *controlled experiments* as well (!!)

The statistical model III

- As with any statistical experiment, we need to begin by defining our sample space
- In the most general sense, our sample space is:

$$\Omega = \{ \text{ Possible Individuals } \}$$

- More specifically, each individual in our sample space can be quantified as a pair of sample outcomes so our sample space can be written as:

$$\Omega = \{\Omega_g \cap \Omega_P\}$$

- Where Ω_g is the genotype sample space at a locus and Ω_P is the phenotype sample space
- Note that genotype $g_i = A_j A_k$ is the set of possible genotypes, where for a diploid, with two alleles:

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

- For the phenotype, this can be any type of measurement (e.g. sick or healthy, height, etc.)

The statistical model IV

- Next, we need to define the probability model on the sigma algebra of the sample space ($\mathcal{F}_{\{g,P\}}$):

$$Pr(\mathcal{F}_{\{g,P\}})$$

- Which defines the probability of each possible genotype and phenotype pair:

$$Pr\{g,P\}$$

- We will define two (types) or random variables (* = state does not matter):

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

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

- Note that the probability model induces a (joint) probability distribution on this random vector (these random variables):

$$Pr(Y, X)$$

The statistical model V

- The goal of quantitative genomics and genetics is to identify cases of the following relationship:

$$Pr(Y \cap X) = Pr(Y, X) \neq Pr(Y)Pr(X)$$

- Remember that, regardless of the probability distribution of our random vector, we can define the expectation:

$$\mathrm{E}\left[Y, X\right] = \left[\mathrm{E}Y, \mathrm{E}X\right]$$

- and the variance:

$$Var\left[Y, X\right] = \begin{bmatrix} Var(Y) & Cov(Y, X) \\ Cov(Y, X) & Var(X) \end{bmatrix}$$

- The goal of quantitative genomics can be rephrased as assessing the following relationship:

$$Cov(Y, X) \neq 0$$

The statistical model VI

- We are going to consider a parameterized model to represent the probability model of *X* and *Y* (that is the true statistical model of genetics!!!)

- Specifically, we will consider a *regression model*

- For the moment, let's consider a regression model with normal error:

$$Y = \beta_0 + X\beta_1 + \epsilon$$

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

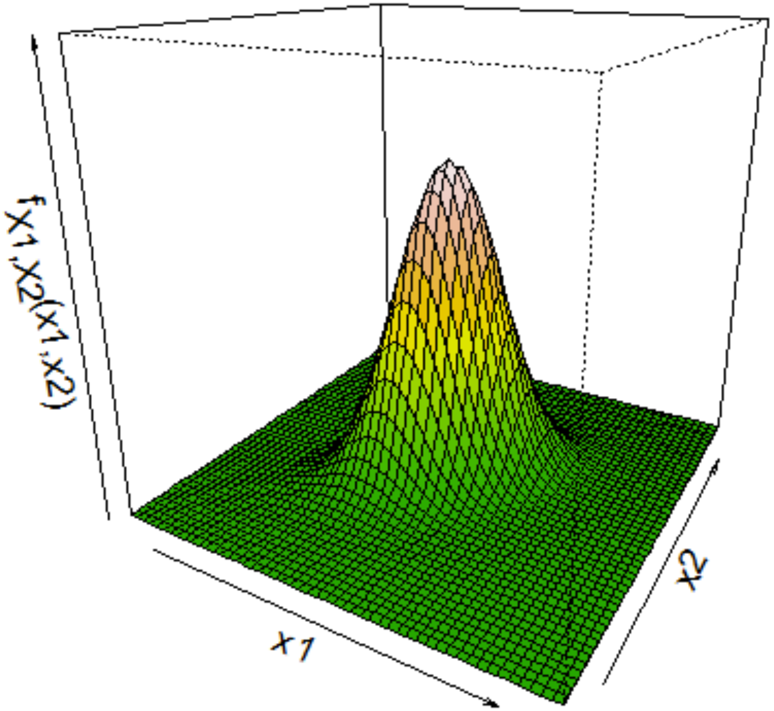
- Note that in this model, we consider Y to be the *dependent* or *response* variable and X to be the *independent* variable (what are the parameters!?)

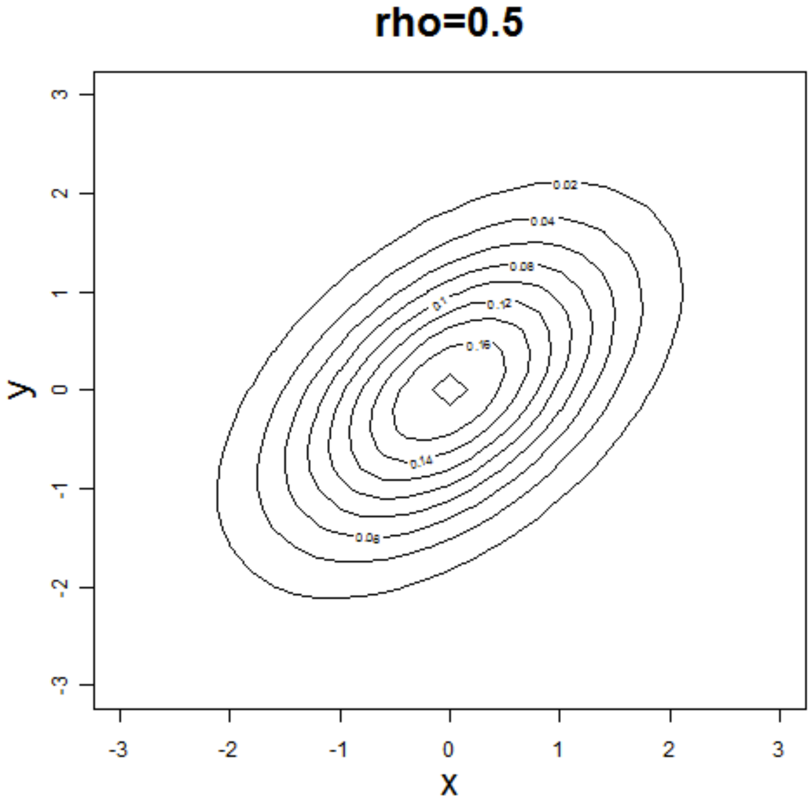
- Also note implicitly assumes the following:

$$Pr(Y, X) = Pr(Y|X)$$

Linear regression is a bivariate distribution

- We've seen bivariate (multivariate) distributions before:

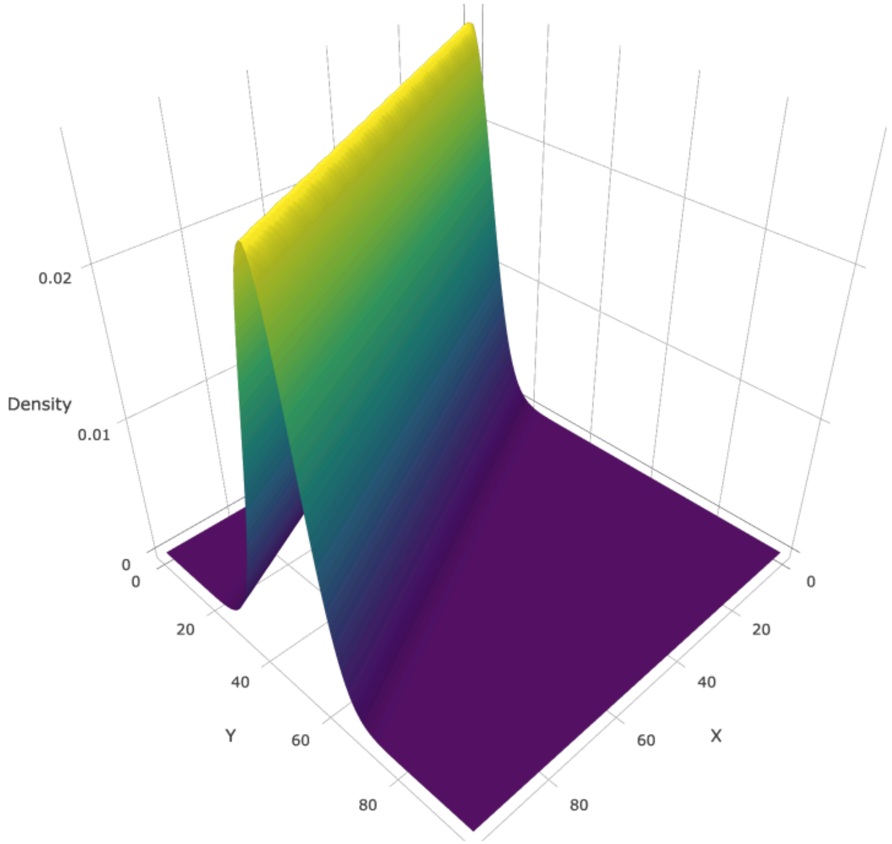


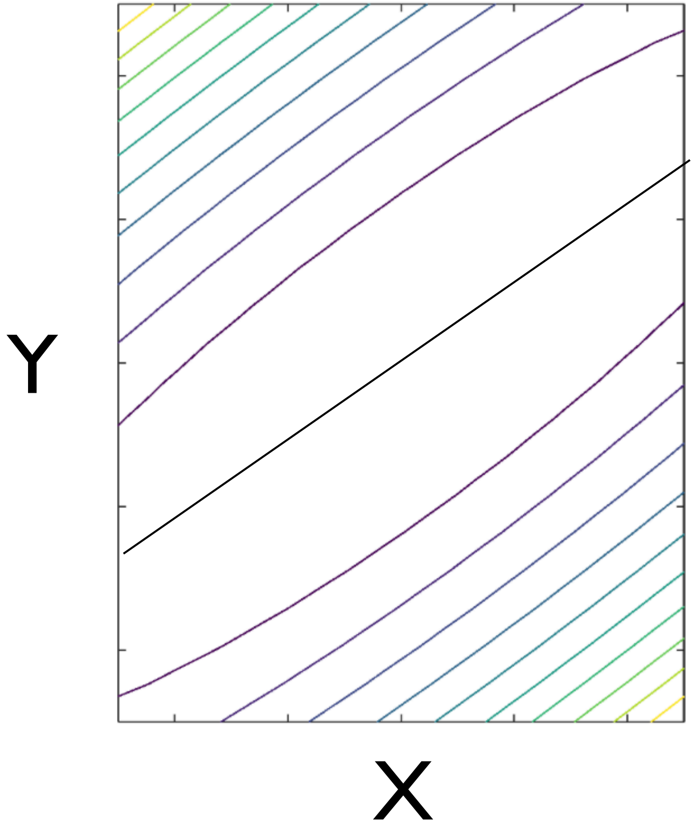


Linear regression I

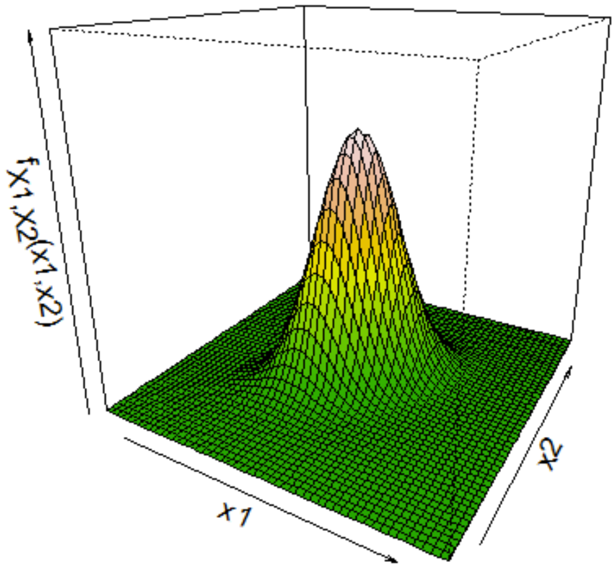
- Let’s review the structure of a linear regression (not necessarily a genetic model):

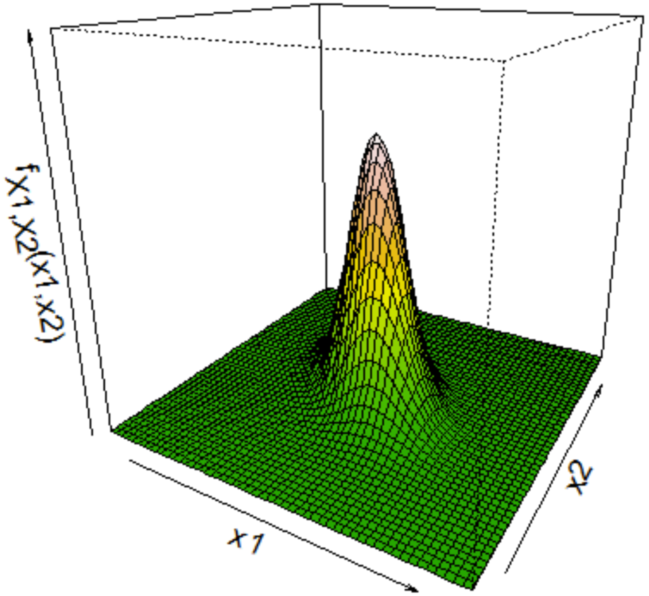
$$Y = \beta_0 + X\beta_1 + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

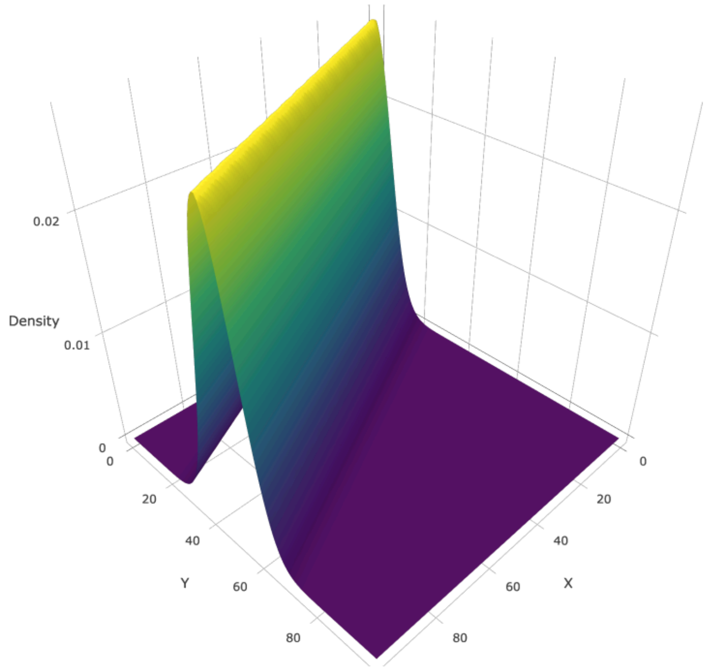


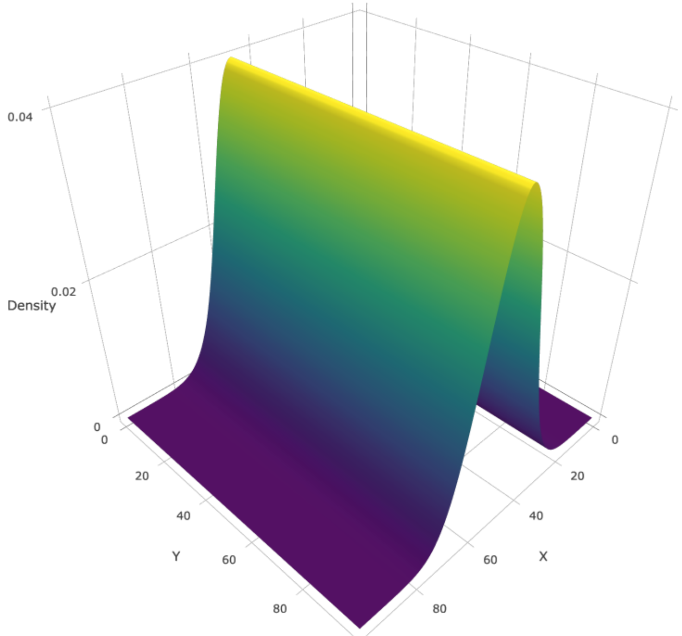


Linear regression II









That’s it for today

- Next lecture, we will discuss inference for Genetic Models (!!)