Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 10: Quantitative Genomics I

Jason Mezey
jgm45@cornell.edu
March 2, 2017 (Th) 8:40-9:55
Announcements

• Homework #4 will be posted later today
Summary of lecture 10

• Last lecture, we completed our general discussion of inference (i.e. estimation and hypothesis testing)

• Today, we will introduce critical concepts in genetics and introduce the model (a regression!) that is the basis of quantitative genetics/genomics and how to perform inference with this model
Conceptual Overview

System

Question

Sample

Prob. Models

Inference

Statistics

Assumptions
Conceptual Overview

Genetic System

Does A1 $\rightarrow$ A2 affect $Y$?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Pr($Y|X$)

Reject / DNR

Model params F-test
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.
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
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:

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

- 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!)
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
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, 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 \( \Omega_g \) is the genotype sample space at a locus and \( \Omega_P \) 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:
  \[ \Omega_g = \{ A_1 A_1, A_1 A_2, A_2 A_2 \} \]
- 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:

\[ E[Y, X] = [EY, EX] \]

- and the variance

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

- 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:

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

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

- 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:

$$Pr(Y, X) = Pr(Y|X)$$
Let’s review the structure of a linear regression (not necessarily a genetic model):

\[ Y = \beta_0 + X\beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2) \]
A *multiple regression* model has the same structure, with a single dependent variable $Y$ and more than one independent variable $X_i, X_j$, etc.
The genetic statistical model I

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

\[ X_a(A_1A_1) = -1, \quad X_a(A_1A_2) = 0, \quad X_a(A_2A_2) = 1 \]
\[ X_d(A_1A_1) = -1, \quad X_d(A_1A_2) = 1, \quad X_d(A_2A_2) = -1 \]

\[
\begin{array}{c|cc}
1 & A_1A_2 \\
-1 & A_1A_1 & A_2A_2 \\
-1 & 0 & 1 \\
\end{array}
\]

- and the following regression equation:

\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon^2) \]

- where for an individual we may write:

\[ y_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d + \epsilon_i \]
The genetic statistical model II

- The probability distribution of this model, is therefore:
  \[ Pr(Y|X) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d, \sigma_\epsilon^2) \]

- Which has four parameters:
  \( \beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 \)

- The three \( \beta \) parameters are required to model the three separate genotypes (A1A1, A1A2, A2A2)

- The \( \epsilon \) can be thought of as a random variable that describes the probability an individual will have a specific value of \( Y \), conditional on the genotype AiAj, where the probability is normally distributed around the value determined by the \( X \)'s and \( \beta \)'s
  \[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
The genetic statistical model III

- Let’s consider a specific example where we are interested modeling the relationship between a genotype and a phenotype (such as height) where the latter is well approximated by a normal distribution.

- For this case, the (unknown) conditions of the experiment define the true values of the parameters (unknown to us!), which we will say are the following (note these are the same for all individuals in the population since they are parameters of the probability distribution):

  \[ \beta_\mu = 0.3, \beta_a = -0.2, \beta_d = 1.1, \sigma^2_\epsilon = 1.1 \]

- Consider an individual \( i \) with \( g_i = A1A2 \) such that we have:

  \[ X_a(A1A2) = 0, X_d(A1A2) = 1 \]

- If this individual has a phenotype value \( y_i = 2.1 \) then we have the epsilon value \( \epsilon_i = 0.7 \) where the probability of this particular value (i.e. the interval surrounding this value) is defined by \( \epsilon \sim N(0, \sigma^2_\epsilon) \)

  \[ 2.1 - 0.3 + (0)(-0.2) + (1)(1.1) + 0.7 \]
Genetic probability model VI

• Note that, while somewhat arbitrary, the advantage of the $X_a$ and $X_d$ coding is the parameters $\beta_a$ and $\beta_d$ map directly on to relationships between the genotype and phenotype that are important in genetics:

  • If $\beta_a \neq 0, \beta_d = 0$ then this is a “purely” additive case
  • If $\beta_a = 0, \beta_d \neq 0$ then this is only over- or under-dominance (homozygotes have equal effects on phenotype)
  • If both are non-zero, there are both additive and dominance effects
  • If both are zero, there is no effect of the genotype on the phenotype (the genotype is not causal!)
That’s it for today

- Next week: quantitative genomics II (hypothesis testing and an introduction to GWAS!)