Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 14: Introduction to Statistics IV

Jason Mezey
jgm45@cornell.edu
April 1, 2021 (Th) 8:05-9:20
Summary of lecture 14

- Today we will complete our discussion of LRTs
- And begin our discussion of (Quantitative) Genetics (!!!)
Hypothesis Tests

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

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

$[X_1 = x_1, \ldots, X_n = x_n]$

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

$X = x$

$Pr(X)$

Random Variable

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

$Pr(\mathcal{F})$

Experiment

$(\text{Sample Space})$

$\Omega$

$(\text{Sigma Algebra})$

$\mathcal{F}$
Review: Understanding p-values...

- **Inference** - the process of reaching a conclusion about the true probability distribution (from an assumed family of probability distributions indexed by parameters) on the basis of a sample.

- **System, Experiment, Experimental Trial, Sample Space, Sigma Algebra, Probability Measure, Random Vector, Parameterized Probability Model, Sample, Sampling Distribution, Statistic, Statistic Sampling Distribution, Estimator, Estimator Sampling distribution, Null Hypothesis, Sampling Distribution Conditional on the Null, p-value, One-or-Two-Tailed, Type I Error, Critical Value, Reject / Do Not Reject I - Type I, Type II Error, Power, Alternative Hypothesis
Since there are an unlimited number of ways to define statistics, there are an unlimited number of ways to define hypothesis tests.

However, some are more “optimal” than others in terms of having good power, having nice mathematical properties, etc.

The most widely used framework (which we will largely be concerned with in this class) are Likelihood Ratio Tests (LRT).

Similar to MLE’s (and they include MLE’s to calculate the statistic!) they have a confusing structure at first glance, however, just remember these are simply a statistic (sample in, number out) that we use like any other statistic, i.e. with the number out, we can calculate a p-value etc.
Likelihood ratio tests II

- Likelihood Ratio Tests use a statistic with the following structure:

\[ \Lambda = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}_1|\mathbf{x})} \]

- \( L(\theta|\mathbf{x}) \) is the likelihood function

- \( \hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} L(\theta|\mathbf{x}) \) is the parameter that maximizes the likelihood given the sample restricted to the set of parameters defined by \( \mathcal{H}_0 \), which we symbolize by \( \Theta_0 \)

- \( \hat{\theta}_1 = \arg\max_{\theta \in \Theta_1} L(\theta|\mathbf{x}) \) is the parameter that maximizes the likelihood given the sample restricted to the set of parameters defined by \( \mathcal{H}_A \Theta_1 = \Theta_A \) or more usually the values \( \Theta_1 = \Theta_A \cup \Theta_0 \)

- We will assume the following for the alternative set of hypotheses, for example:

\[ H_0 : \mu = c \text{ then } H_A : \mu \neq c \]
Likelihood ratio tests III

- Again, consider our simplified normal r.v. with sample \( n \)
- The likelihood is:
  \[
  L(\theta|x) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}}
  \]
- and the LRT statistic for \( H_0 : \mu = c \) is:
  \[
  \Lambda = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}_1|x)}
  \]
  \[
  LRT = \Lambda = \frac{1}{(2\pi*\text{MLE}(\hat{\sigma}^2)^{n/2}}} e^{\sum_{i=1}^{n} \frac{-(x_i-H_0(\mu))^2}{2*\text{MLE}(\hat{\sigma}^2)}}
  \]
  \[
  \frac{1}{(2\pi*\text{MLE}(\hat{\sigma}^2)^{n/2}}} e^{\sum_{i=1}^{n} \frac{-(x_i-\text{MLE}(\hat{\mu}))^2}{2*\text{MLE}(\hat{\sigma}^2)}}
  \]
- where we have:
  \[
  H_0(\mu) = c
  \]
  \[
  \text{MLE}(\hat{\mu}) = \text{mean}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i
  \]
  \[
  \text{MLE}(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \text{mean}(x))^2
  \]
Likelihood ratio tests IV

- Remember, to calculate a p-value, we need to know the sampling distribution under the null (note likelihood ratio tests are two-sided tests!)

- If we consider the following transformation:

\[
LRT = -2 \ln(\Lambda) = -2 \ln \left( \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}_1|x)} \right)
\]

- It turns out that, under conditions that often apply, as the sample size \( n \to \infty \) the sampling distribution of this statistic under the null approaches (in the specific case on the last slide, the d.f. = \( k = 1 \)):

\[
Pr(LRT|H_0: \theta = c) \to \chi^2_{d.f.}
\]
Likelihood ratio tests V

• There is a difference between a sampling distribution (under the null) that approaches a distribution as \( n \to \infty \) and a case where we know the exact distribution for any size \( n \) (i.e., for the former, the null distribution is approximate).

• Why use a test statistic where the distribution under the null is approximate (since we need to know this distribution to do the hypothesis test!)?

• The approximation is very close even for moderate sized \( n \).

• An LRT is a very versatile way of constructing a hypothesis test with “good” properties for many types of cases.

• Even better, for some specific tests, the sampling distribution under the null for ANY sample size \( n \) is known exactly for a specified transformation of the likelihood ratio statistic.

• Note that this is the case for many of the tests you are familiar with (t-tests, F-tests, tests of the linear regression slope, etc.), that is, these tests are forms of likelihood ratio test statistic!!!
Conceptual Overview

System

Question

Inference

Experiment

Sample

Prob. Models

Statistics

Assumptions
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Reject / DNR

Pr(Y|X)

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, 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 \( \Omega_g \) 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):
  \[
  Y : (\ast, \Omega_P) \rightarrow \mathbb{R}
  \]
  \[
  X : (\Omega_g, \ast) \rightarrow \mathbb{R}
  \]
- 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 IV

- 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:

\[
Pr(Y, X) = Pr(Y | X)
\]
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):

\[ Y = \beta_0 + X\beta_1 + \epsilon \quad \epsilon \sim N(0, \sigma^2_\epsilon) \]
That’s it for today

• Next: continue with genetic analysis foundations!