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

Lecture 6: Inference, samples, statistics, and estimators

Jason Mezey
jgm45@cornell.edu
Feb. 13, 2018 (T) 8:40-9:55
Announcements

• Homework announcements:
  • Homework #2, problem - with problem 2h everyone will get credit
  • Homework #3 will be posted later today
  • Again please note the late policy: in to CMS past deadline, it is late and subject to a penalty

• In Ithaca, COMPUTER LAB will now be held THURS. 5-6PM FOR EVERYONE in WEILL HALL (!!!)
  • Your registration stays the same (!!!) i.e., if you registered for Fri. 8-9AM you do nothing - but show up for lab Thurs.5-6PM
  • The room in Weill Hall will change week to week (e.g., this week 226 Weill Hall) “Ithaca-NYC Classrooms”
  • Please see the room schedule posted on the class website (we will announce changes but check back often!)
Summary of lecture 6

• Last lecture, we discussed functionals of random variables that provided information about the “structure” or r.v. regardless of the probability distribution

• In this lecture, we will introduce specific probability models with the concept of parameterized probability distributions, where we also begin our discussion of inference, including the concept of a sample (and i.i.d.), the concept of a statistic, and the concept of estimators
Conceptual Overview

- System
- Experiment
- Question
- Sample
- Inference
- Prob. Models
- Statistics
- Assumptions
Probability models I

- We have defined \( \Pr(X) \), a probability model on a random variable, which technically we produce by defining \( \Pr(\mathcal{F}) \) and \( X(\Omega) \)

- So far, we have considered such probability models without defining them explicitly (except for an illustrative few examples)

- To define an explicit model for a given system / experiment, we are going to assume that there is a “true” probability model, that is, a consequence of the experiment that produces sample outcomes

- We place “true” in quotes since the defining a single true probability model for a given case could only really be accomplished if we knew every single detail about the system and experiment (would a probability model be useful in this case?)

- In practice, we therefore assume that the true probability distribution is within a restricted family of probability distributions, where we are satisfied if the true probability distribution in the family describes the results of our experiment pretty well / seems reasonable given our assumptions
• In short, we therefore start a statistical investigation *assuming* that there is a single true probability model that correctly describes the possible experiment outcomes given the uncertainty in our system.

• In general, the starting point of a statistical investigation is to make *assumptions* about the form of this probability model.

• More specifically, a convenient assumption is to assume our true probability model is specific model in a family of distributions that can be described with a compact equation.

• This is often done by defining equations indexed by *parameters*.
Probability models III

- **Parameter** - a constant(s) $\theta$ which indexes a probability model belonging to a family of models $\Theta$ such that $\theta \in \Theta$

- Each value of the parameter (or combination of values if there is more than one parameter) defines a different probability model: $\Pr(X)$

- We assume one such parameter value(s) is the true model

- The advantage of this approach is this has reduced the problem of using the sample to answer a broad question to the problem of using a sample to make an educated guess at the value of the parameter(s)

- Remember that the foundation of such an approach is still an assumption about the properties of the sample outcomes, the experiment, and the system of interest (!!!)
Discrete parameterized examples

• Consider the probability model for the one coin flip experiment / number of tails.

• This is the Bernoulli distribution with parameter \( \theta = p \) (what does \( p \) represent!?) where \( \Theta = [0, 1] \)

• We can write this \( X \sim \text{Bern}(p) \) and this family of probability models has the following form:

\[
Pr(X = x|p) = P_X(x|p) = p^x (1 - p)^{1-x}
\]

• For the experiment of \( n \) coin flips / number of tails, we can assume the Binomial distribution \( X \sim \text{Bin}(n, p) \):

\[
Pr(X = x|n, p) = P_X(x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}
\]

\[
\binom{n}{x} = \frac{n!}{x!(n-x)!}
\]

\[
n! = n \times (n-1) \times (n-2) \times \ldots \times 1
\]

• There are many other discrete examples: hypergeometric, Poisson, etc.
Continuous parameterized examples

- Consider the measure heights experiment (reals as approximation to the sample space) / identity random variable

- For this example we can use the family of normal distributions that are parameterized by \( \theta = [\mu, \sigma^2] \) (what do these parameters represent!?) with the following possible values: \( \Theta_\mu = (-\infty, \infty), \Theta_\sigma^2 = [0, \infty) \)

- We often write this as \( X \sim N(\mu, \sigma^2) \) and the equation has the following form:

\[
Pr(X = x | \mu, \sigma^2) = f_X(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

- There are many other continuous examples: uniform, exponential, etc.
Example for random vectors

- Since random vectors are the generalization of r.v.'s, we similarly can define parameterized probability models for random vectors.

- As an example, if we consider an experiment where we measure “height” and “weight” and we take the 2-D reals as the approximate sample space (vector identity function), we could assume the bivariate normal family of probability models:

\[
f_X(x|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} \exp\left[ -\frac{1}{2(1-\rho^2)} \left( \frac{(x_1 - \mu_1)^2}{2\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_1)^2}{2\sigma_2^2} \right) \right]
\]

- Other continuous distributions that we will run into during this course are the Uniform, chi-squared, t, F, Gamma, and Beta. The former we will discuss in the context of the distribution of p-values, the middle three will come up in our discussion of sampling distributions of statistics, and we will discuss the last two during our lectures on Bayesian statistics.

- One final point to note: While we have considered parameterized statistical models for individual ‘univariate’ random variables, there are analogous forms of all of these distributions for random vectors with multiple elements, which are ‘multivariate’ random variables. Although the multivariate forms have additional parameters, we will consider some multivariate forms of these distributions in this class, e.g., the multivariate Normal distribution.

3 Introduction to inference

A major goal of the field of statistics is inference: the process of reaching conclusions concerning an assumed probability distribution, specifically the parameters, on the basis of a sample.
Recall that our eventual goal is to use a sample (generated by an experiment) to provide an answer to a question (about a system).

So far, we have set up the mathematical foundation that we need to accomplish this goal in a probability / statistics setting (although note we have not yet provided formalism for a sample!!).

Specifically, we have defined formal components of our framework and made assumptions that have reduced the scope of the problem.

With these components and assumptions in place, we are almost ready to perform inference, which will accomplish our goal.
Introduction to inference II

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

- There are two major types of inference we will consider in this course: estimation and hypothesis testing.

- Before we get to these specific forms of inference, we need to formally define: samples, sample probability distributions (or sampling distributions), statistics, statistic probability distributions (or statistic sampling distributions).
We are going to define a function from class). This concept is often introduced to us as output. Function to do this, we define a measure which map sample spaces to the real numbers. Before we consider the specifics of how we define a function, let's consider the intuitive definition of a function: a mathematical operator that takes an input and produces an output. Function $X = x$ and $Pr(X)$ is a function, which we could have written as $X(x)$. Then we introduce the ability function $Pr(Y)$. Then we consider the specifics of how we define a function. 

**Random Variable**

$X(s), s \in \Omega$  $Pr(F)$

**Experiment**

$\mathcal{X}$

$\Omega$ (Sample Space)

$\mathcal{F}$ (Sigma Algebra)

**So far**

$X = x$  $Pr(X)$
Samples

\[ [X_1 = x_1, \ldots, X_n = x_n] \quad \text{Pr}(X_1 = x_1, \ldots, X_n = x_n) \]

\[ X = x \quad \text{Pr}(X) \]

Random Variable

\[ X(\omega), \omega \in \Omega \]

\[ \mathcal{X} \]

Experiment

\( \Omega \)

(Sample Space)

\( \mathcal{F} \)

(Sigma Algebra)
Inference concepts

• Recall that our eventual goal is to use a sample (generated by an experiment) to provide an answer to a question (about a system)

• Inference (informally) is the process of using the output of the experiment = experimental trials (the sample) to answer the question

• For our system and experiment, we are going to assume there is a single “correct” probability function (which in turn defines the probability of our possible random variable outcomes, the probability of possible random vectors that represent samples, and the probability of possible values of a statistic)

• For the purposes of inference, we often assume a parameterized family of probability models determine the possible cases that contain the “true” model that describes the result of the experiment

• This reduces the problem of inference to identifying the “single” value(s) of the parameter that describes this true model
• Recall that we have defined experiments (= experimental trials) in a probability / statistics setting where these involve observing individuals from a population or the results of a manipulation

• We have defined the possible outcome of an experimental trial, i.e. the sample space $\Omega$

• We have also defined a random variable $X$, where the random variable maps sample outcomes to numbers, the quantities in which we are ultimately interested

• Since we have also defined a probability model $\Pr(X)$, we have shifted our focus from the sample space to the random variable
Samples II

- **Sample** - repeated observations of a random variable $X$, generated by experimental trials

- We will consider samples that result from $n$ experimental trials (what would be the ideal $n = \text{ideal experiment}$?)

- We already have the formalism to represent a sample of size $n$, specifically this is a random vector:

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

- As an example, for our two coin flip experiment / number of tails r.v., we could perform $n=3$ experimental trials, which would produce a sample = random vector with three elements
Samples III

- Note that since we have defined (or more accurately induced!) a probability distribution \( \Pr(X) \) on our random variable, this means we have induced a probability distribution on the sample (!!):

\[
\Pr(X = x) = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P_x(x) \text{ or } f_x(x)
\]

- This is the sample probability distribution or sampling distribution (often called the joint sampling distribution)

- While samples could take a variety of forms, we generally assume that each possible observation in the sample has the same form, such that they are identically distributed:

\[
\Pr(X_1 = x_1) = \Pr(X_2 = x_2) = \ldots = \Pr(X_n = x_n)
\]

- We also generally assume that each observation is independent of all other observations:

\[
\Pr(X = x) = \Pr(X_1 = x_1)\Pr(X_2 = x_2)\ldots\Pr(X_n = x_n)
\]

- If both of these assumptions hold, then the sample is independent and identically distributed, which we abbreviate as i.i.d.
We are going to define a function \( f \) (or \( \text{model} \)) that maps the values taken by \( X \) to the real numbers. To use sample spaces in probability, we need a way to map these sets to the real numbers.

To do this, we define a probability function \( X \) (intuitive def.) \( \Pr(X) \) (1)

This concept is often introduced to us as an ability function \( \mathbb{R} \rightarrow \mathbb{R} \).

\[ X = x \quad \Rightarrow \quad \Pr(X) \]

Let's consider the intuitive definition of a function:

\[ \mathcal{X} \rightarrow \mathbb{R} \]

\[ X(\omega), \omega \in \Omega \]

Before we consider the specifics of how we define a function, we need to understand the concept of a random variable.

\[ X = x \quad \Rightarrow \quad \Pr(X) \]

\[ \mathcal{F} \quad \Omega \quad \mathcal{X} \]

(Sigma Algebra) (Sample Space) (Random Variable)
Example of sampling distributions

- As an example, consider our height experiment (reals as approximate sample space) / normal probability model (with true but unknown parameters $\theta = [\mu, \sigma^2]$ / identity random variable

- If we assume an i.i.d sample, each sample $X_i = x_i$ has a normal distribution with parameters $\theta = [\mu, \sigma^2]$ and each is independent of all other $X_j = x_j$

- For example, the sampling distribution for an i.i.d sample of $n = 2$ is:
Samples IV

- Technical note I: when considering a sample where each observation is independent, the actual sample space is actually a “product” of sample spaces (product space) where each random variable in the sample (the random vector of the sample) is a function on one of the sample spaces:

\[ \Omega_{ind} = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n \]

- If this is the case, why have we bothered considering a case where we define multiple random variables on a “single” sample space, e.g., \( X_1(\Omega) \) and \( X_2(\Omega) \)?

- This largely for conceptual reasons, e.g., when considering samples we may want to consider each observation of the sample to contain two observations, such that each observation in the sample is a vector (!!) and the sample is a set of vectors (a matrix!)

- Technical note II: regardless of the size of \( n \), there is a sampling distribution although as \( n \to \infty \) this becomes a probability distribution that only assigns a non-zero value (one!) to only the entire sample space element of the Sigma Algebra
Samples V

• It is important to keep in mind, that while we have made assumptions such that we can define the joint probability distribution of (all) possible samples that could be generated from $n$ experimental trials, in practice we only observe one set of trials, i.e. one sample.

• For example, for our one coin flip experiment / number of tails r.v., we could produce a sample of $n=10$ experimental trials, which might look like:

$$\mathbf{x} = [1, 1, 0, 1, 0, 0, 1, 1, 0]$$

• As another example, for our measure heights / identity r.v., we could produce a sample of $n=10$ experimental trials, which might look like:

$$\mathbf{x} = [-2.3, 0.5, 3.7, 1.2, -2.1, 1.5, -0.2, -0.8, -1.3, -0.1]$$

• In each of these cases, we would like to use these samples to perform inference (i.e. say something about our parameter of the assumed probability model).

• Using the entire sample is unwieldy, so we do this by defining a statistic.
Statistics I

- **Statistic** - a function on a sample

- Note that a statistic $T$ is a function that takes a vector (a sample) as an input and returns a value (or vector):

$$T(x) = T(x_1, x_2, \ldots, x_n) = t$$

- For example, one possible statistic is the mean of a sample:

$$T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- It is critical to realize that, just as a probability model on $X$ induces a probability distribution on a sample, since a statistic is a function on the sample, this induces a probability model on the statistic: the *statistic probability distribution* or the *sampling distribution* of the statistic (!!)
Statistics II

• As an example, consider our height experiment (reals as approximate sample space) / normal probability model (with true but unknown parameters $\theta = [\mu, \sigma^2]$ / identity random variable

• If we calculate the following statistic:

$$T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

what is $\Pr(T(X))$?

• Are the distributions of $X_i = x_i$ and $\Pr(T(X))$ always the same?
Statistics

Statistic: $T(x)$

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

$X = x$

Random Variable

$\mathcal{X}$

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

$Pr(\mathcal{F})$

Experiment

$\Omega$

(Sample Space)

$Pr(T(X))$

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

$Pr(X)$

$Pr(\mathcal{F})$
Statistics and estimators I

• Recall for the purposes of inference, we would like to use a sample to say something about the specific parameter value (of the assumed) family or probability models that could describe our sample space.

• Said another way, we are interested in using the sample to determine the “true” parameter value that describes the outcomes of our experiment.

• An approach for accomplishing this goal is to define our statistic in a way that it will allow us to say something about the true parameter value.

• In such a case, our statistic is an estimator of the parameter $T(x) = \hat{\theta}$.

• There are many ways to define estimators (we will focus on maximum likelihood estimators in this course).

• Each estimator has different properties and there is no perfect estimator.
Statistics and estimators II

- Estimation is a “type” of inference, i.e. where we use a sample to reach a conclusion about a parameter.
- Specifically, estimation is the process of saying something about the specific value of the true parameter.
- Again, as a reminder, we do this by defining an estimator $\hat{\theta}$, which is a function on our sample.
- Intuitively, an estimator is the value for which we have the best evidence for being the true value of the parameter (our “best guess”) based on the sample, given uncertainty and our assumptions.
- Note that without an infinite sample, we will never know the true value of the parameter with absolute certainty (!!).
Estimator - a statistic defined to return a value that represents our best evidence for being the true value of a parameter.

In such a case, our statistic is an estimator of the parameter: $\hat{T}(x) = \theta$

Note that ANY statistic on a sample can in theory be an estimator.

However, we generally define estimators (=statistics) in such a way that it returns a reasonable or “good” estimator of the true parameter value under a variety of conditions.

How we assess how “good” an estimator depends on our criteria for assessing “good” and our underlying assumptions.
Statistics and estimators IV

- Since our underlying probability model induces a probability distribution on a statistic, and an estimator is just a statistic, there is an underlying probability distribution on an estimator:
  \[ Pr(T(X = x)) = Pr(\hat{\theta}) \]

- Our estimator takes in a vector as input (the sample) and may be defined to output a single value or a vector of estimates:
  \[ T(X = x) = \hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, ...] \]

- We cannot define a statistic that always outputs the true value of the parameter for every possible sample (hence no perfect estimator!)

- There are different ways to define “good” estimators and lots of ways to define “bad” estimators (examples?)
That’s it for today

• Next lecture, we will begin our discussion of maximum likelihood estimators