BTRY 4830/6830: Quantitative Genomics and Genetics

Lecture 5: Parameterized probability models, inference, samples, estimators

Jason Mezey
jgm45@cornell.edu
Sept. 9, 2014 (T) 8:40-9:55
Announcements

• Reminder: two supplemental readings posted
• Reminder: Homework #1 was due 11:59PM last night (!!)
• We will have this graded one week from today
• Homework #2 will be posted tomorrow (I will send email announcement)
Summary of lecture 5

• Last lecture, we discussed random vectors and functions that take both random variables AND probability models as input to produce useful “summary” (and more!) outputs useful for random variables / probability models in general (expectations, variances)

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

- System
- Experiment
- Question
- Sample
- Inference
- Prob. Models
- Statistics
- Assumptions
Random Variables

\[ X = x, \ Pr(X) \]

- \( \mathcal{E}(\Omega) \)
- \( X(\Omega) \)
- \( \Omega \)
- \( \mathcal{F} \)
- \( \Pr(\mathcal{F}) \)

Experiment (Sample Space) (Sigma Algebra)
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 a 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
- 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 outcomes of our experiment.

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 a 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 on 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 $X \sim \text{Bin}(n, p)$:

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

- 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:

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

• 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 defined 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)
Samples I

- 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 $X(\Omega)$, 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.
The values taken by $X$ are numbers:

This concept is often introduced to us as the probability function from a set $X$. To do this, we define a function $F_r$ which maps sets to the real numbers. For example, we can have the function $F_r = \mathbb{1}_{H,T}$ where $r$ is the function that maps a sample to the real line.

Random Variable

$X = x$, $Pr(X)$

Sample of size $n$

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

Sampling Distribution

$X \sim \mathbb{1}_{H,T}$

Experiment

$\mathcal{E}(\Omega)$

Sample Space

$\Omega$

Sigma Algebra

$\mathcal{F}$

Probability

$Pr(\mathcal{F})$
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 (!!):

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

• 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 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 sample is independent of all other samples:

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

• If both of these assumptions hold, than the sample is independent and identically distributed, which we abbreviate as i.i.d.

• Technical note: 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.
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 $X_i = x_i$.

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

• 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:

\[
x = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0]
\]

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

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

• 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):

  $$T(x) = T(x_1, x_2, ..., x_n) = t$$

- 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 (!!)
To use sample spaces in probability we need a way to map these sets to the real numbers. This concept is often introduced to us as the intuitive definition of a function: a mathematical operator that takes an input and produces an output. For example, we can have the function $T(X), Pr(T(X))$ which maps a sample to its statistic.

A random variable $X$ can be seen as a function $X(\Omega)$ from the sample space $\Omega$ to the real numbers. The sampling distribution $Pr(X)$ gives the probability of each possible value of $X$. The statistic $T(X)$ and its distribution $Pr(T(X))$ are derived from the sample space and the random variable. The MLE $\hat{\theta}$ is an estimator of the parameter $\theta$, where $\hat{\theta}$ is chosen to maximize the likelihood of the data $x$.

The diagram shows the relationship between the experiment $\mathcal{E}(\Omega)$, the sample space $\Omega$, the random variable $X$, the statistic $T(X)$, and the sampling distribution $Pr(T(X))$.
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 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.
That’s it for today

- Reminder (!!) Amanda has office hours today (please attend if additional discussion of these concepts would be helpful)
- Next lecture, we will introduce likelihood and maximum likelihood estimators