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

Lecture 7:
Samples, statistics, and estimators

Jason Mezey
jgm45@cornell.edu
Feb. 11, 2019 (T) 8:40-9:55
Announcements

• Correction:
  
  \[ Pr(\{HH, HT\}) = Pr(\{HH\} \cup \{HT\}) \]
  
  \[ Pr(X_1 = 1, X_2 = 1) = Pr(X_1 = 1 \cap X_2 = 1) \]

• Homework #2 due 11:59PM Fri. (Feb 14)

• Office hours Weds. (Feb. 12) FROM 3-5PM (!!) (NOT 4-6PM!)

• NO OFFICE HOURS MON.
Summary of lecture 7

- Last lecture, we introduced specific probability models with the concept of parameterized probability distributions, where we also begin our discussion of inference.

- Today, we will continue 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
- Question
- Inference
- Prob. Models

- Experiment
- Sample
- Assumptions
- Statistics
• 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
• **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)
To use sample spaces in probability, we need a way to map these sets to the real numbers.

### Review: So far

1. **Random Variable**
   - $X = x$
   - $Pr(X)$
   - $X(\omega), \omega \in \Omega$

2. **Experiment**
   - $\mathcal{X}$

3. **Sample Space**
   - $\Omega$

4. **Sigma Algebra**
   - $\mathcal{F}$

5. **Probability Function**
   - $Pr(\mathcal{F})$
As of today (!!): Samples

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

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

\[\mathcal{X} \quad X(\omega), \omega \in \Omega \quad \text{Pr}() \]

Experiment \quad \Omega \quad \mathcal{F}

(Sample Space) \quad (Sigma Algebra)
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:

\[
\begin{bmatrix}
X = x
\end{bmatrix} = 
\begin{bmatrix}
X_1 = x_1, \ldots, X_n = x_n
\end{bmatrix}
\]

- 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, than the sample is independent and identically distributed, which we abbreviate as i.i.d.
To use sample spaces in probability, we need a way to map these sets to the real numbers.

We are going to define a probability function $P_r(X)$ which maps sample spaces to numbers:

$$X = x \quad P_r(X)$$

$X$ represents a random variable, $X(\omega), \omega \in \Omega$.

$\Omega$ is the sample space, $F$ is the sigma algebra.

Experiment $\rightarrow$ Sample Space $\rightarrow$ Random Variable $\rightarrow$ Probability Function $\rightarrow$ Numbers.
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

- 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:

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

• 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):
  \[ T(x) = T(x_1, x_2, \ldots, x_n) = t \]
- 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?
This concept is often introduced to us as output. To use sample spaces in probability, we need a way to map these sets to the real numbers. Thus we define a function $A$ as $A(x) = X$, where $X$ is a measure of the experiment $A$. This function $A$ maps the sample space $S$ to the real numbers. If $X$ is a random variable, then $A$ is a random variable function. A probability function is then defined as $Pr(X)$.
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: \( \hat{\theta} \).

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 (!!).
Statistics and estimators III

- **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: $\hat{T}(x) = \hat{\theta}$

- 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:

\[ T(X = x) = \hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \ldots] \]

- 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?)
Method of moments estimator I

- As an example of how to construct estimators, let’s construct a method of moments estimator.

- Consider the single coin flip experiment / number of tails random variable / Bernoulli probability model family (parameter $p$) / fair coin model (assumed and unknown to us!!!) / sample of size $n=10$.

- What is the sampling distribution (of the sample) in this case?

- We want to estimate $p$, where a perfectly reasonable estimator is:
  \[ T(X = x) = \hat{\theta} = \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

- What is the probability distribution of this statistic in this case?

- e.g. this statistic (=mean of the sample) would equal 0.5 for the following particular sample (will it always?)
  \[ x = [1, 1, 0, 1, 0, 0, 1, 1, 0] \]
Method of moments estimator II

- Let's continue with our example of constructing a method of moments estimator

- Consider the single coin flip experiment / number of tails random variable

  \[
  \Omega = \{H, T\} \quad X : X(H) = 0, X(T) = 1
  \]

- Bernoulli probability model family (parameter \( p \))

  \[X \sim p^X (1 - p)^{1-X}\]

- Sample of size \( n=10 \)

  \[[X = x] = [X_1 = x_1, X_2 = x_2, \ldots, X_{10} = x_{10}]\]

- Sampling distribution (pmf of sample) if i.i.d. (!!)

  \[[X_1 = x_1, X_2 = x_2, \ldots, X_{10} = x_{10}] \sim p^{x_1} (1 - p)^{1-x_1} p^{x_2} (1 - p)^{1-x_2} \ldots p^{x_{10}} (1 - p)^{1-x_{10}}\]
Method of moments estimator III

- Define a statistic $T(X)$

$$T(X = x) = T(x) = \bar{X} = \frac{1}{10} \sum_{i=1}^{10} x_i$$

- Note the values the statistic can take (!!), e.g. with $p=0.5$

$$[T_{\text{min}}, \ldots, T_{\text{max}}] = [0, 0.1, \ldots, 1] \longleftrightarrow [0, 1, \ldots, 10]$$

- We can therefore write the sampling distribution (pmf) of the statistic as

$$Pr(T(x)) \sim \binom{n}{nT(x)} p^{nT(x)} (1 - p)^{n-nT(x)}$$
Method of moments estimator IV

- We are going to use the statistic (mean) of the sample as an estimator of the parameter - and it follows the estimator has the same distribution (!!)

\[ T(x) = \hat{\theta} = \hat{p} \]

\[ Pr(\hat{p}) \sim \binom{n}{nT(x)} p^{nT(x)}(1 - p)^{n - nT(x)} \]

- Also note that the expected value of this estimator is the true value of the parameter (do we ever know this true value!?)

\[ E\hat{p} = p \]

- In practice, one sample and we estimate a single value for the parameter

\[ x = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0] \]
Method of moments estimator V

- As another example consider the heights experiment / identity random variable / Normal probability model family / with true parameters unknown to us (!!) / sample of size n=10
- A perfectly reasonable estimator $\hat{\mu}$ is:
  $$T(X = x) = \bar{X} = \hat{\theta} = \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
- Note that the sampling distribution of this statistic is also normal, where this statistic is the expected value of the statistic sampling distribution (why might this be a good thing?):
  $$T(X = x) = \bar{X} = \hat{\mu} \sim N(\mu, \sigma^2/n)$$
  $$E(X) = \int_{-\infty}^{\infty} X f_X(x) dx = \mu$$
- e.g. this statistic (=mean of the sample) would equal 0.01 for the following particular sample (will it always?)
  $$x = [-2.3, 0.5, 3.7, 1.2, -2.1, 1.5, -0.2, -0.8, -1.3, -0.1]$$
Method of moments estimator VI

• For this same example, we could similarly define the following estimator:

\[ T(X = x) = Var(X = x) = \hat{\theta} = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2 \]

• What is the sampling distribution of this statistic?

• This is the variance of the sample and is justified because:

\[ Var(X) = \int_{-\infty}^{\infty} (X - \mu)^2 f_X(x) dx = \sigma^2 \]

• In sum, since we are calculating means and variances of samples, and these are “moments” when applied to random variables with a probability distribution, these are method of moments estimators.
Introduction to maximum likelihood estimators (MLE)

- We will generally consider maximum likelihood estimators (MLE) in this course.

- Now, MLE’s are very confusing when initially encountered...

- However, the critical point to remember is that an MLE is just an estimator (a function on a sample!!),

- i.e. it takes a sample in, and produces a number as an output that is our estimate of the true parameter value.

- These estimators also have sampling distributions just like any other statistic!

- The structure of this particular estimator / statistic is complicated but just keep this big picture in mind.
Likelihood I

- To introduce MLE’s we first need the concept of *likelihood*

- Recall that a probability distribution (of a r.v. or for our purposes now, a statistic) has fixed constants in the formula called *parameters*

- For example, for a normally distributed random variable

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

- However, we could turn this around and fix the sample and let the parameters vary (this is a likelihood!)

- For example, say we have a sample \(n=1\), where \(x=0.2\) then the likelihood is (if we just set \(\sigma^2 = 1\) for explanatory purposes):

\[
L(\mu | x = 0.2) = \frac{1}{\sqrt{2\pi}} e^{- (0.2-\mu)^2}
\]
That’s it for today

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