#### Quantitative Genomics and Genetics BioCB 4830/6830; PBSB.5201.03

#### Lecture 8: Statistics and Estimators

Jason Mezey Feb 15, 2024 (Th) 8:40-9:55

### Announcements

- NEXT WEEK I: 2nd homework will is due Feb 23) by II:59PM (!!)
- NEXT WEEK 2: Lecture on Tues (Feb 20) will be by zoom for BOTH Ithaca and NYC!
- NEXT WEEK 3: I will need to reschedule office hours (stay tuned!)

# Summary of lecture 8: Statistics and Estimators

- Last lecture, we discussed probability models and introduce samples
- Today we are going to introduce statistics and estimators

## **Conceptual Overview**



#### **Review: Random Variables**



#### Review: Random vectors

- We are often in situations where we are interested in defining more than one r.v. on the same sample space
- When we do this, we define a **random vector**
- Note that a vector, in its simplest form, may be considered a set of numbers (e.g. [1.2, 2.0, 3.3] is a vector with three elements)
- Also note that vectors (when a vector space is defined) ARE NOT REALLY NUMBERS although we can define operations for them (e.g. addition, "multiplication"), which we will use later in this course
- Beyond keeping track of multiple r.v.'s, a random vector works just like a r.v., i.e. a probability function induces a probability function on the random vector and we may consider discrete or continuous (or mixed!) random vectors
- Note that we can define several r.v.'s on the same sample space (= a random vector), but this will result in one probability distribution function (why!?)

#### Review: Probability models I

- 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*

#### Review: Probability models II

- **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 results of experiments to answer a broad question 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 the experiment, and the system of interest (!!!)

#### Review: Introduction to inference

- 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)
- 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 (informally) is the process of using the output of an experiment to answer the question
- Our eventual goal is to use a **sample** (generated by experiment trials) to provide an answer to a question (about a system)

#### **Review: Inference**

- **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: experimental trials, samples, sample probability distributions (or sampling distributions), statistics, statistic probability distributions (or statistic sampling distributions)



#### Samples



#### Review: Experiments to Samples (what we observe!)

- **Experiment** a manipulation or measurement of a system that produces an outcome we can observe
- **Experiment Outcome** a possible outcome of the experiment
- Sample Space set comprising all possible outcomes of an experiment
- **Experimental Trial** one instance of an experiment
- **Sample** (informal) results of one or more experimental trials
- Example (Experiment / Sample Space / Sample):
  - Coin flip / {H,T} / T, T, H, T, H
  - Two coin flips / {HH, HT, TH, TT} / HH, HT, HH, TH, HH
  - Measure heights in this class / Reals / 5'9", 5'2", 5'1", 6'0", 5'7"

#### **Review: Samples I**

- **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* = ideal experiment!?)
- Since a set of actual experimental outcomes may not be numbers (e.g., a set of H and T's) we want to map them to numbers...
- We already have the formalism to do this and represent a sample of size *n*, specifically this is a random vector:

$$[\mathbf{X} = \mathbf{x}] = [X_1 = x_1, \dots, X_n = x_n]$$

As an example, for our two coin flip experiment / number of tails r.v., we could perform n=2 experimental trials, which would produce a sample = random vector with two elements

#### Example: Observed 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:

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

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

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

#### Review: Samples II

- 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 this can take values representing the outcomes of our experimental trials, i.e., X = x
- Since the random variable X also has an induced probability distribution associated with it, we can also consider Pr(X), i.e., the probability of each possible outcome of an experiment or the entire sample!
- Since this defines a probability model Pr(X), we have shifted our focus from the sample space to the random variable

#### Sample Probability Distribution

• 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 (!!):

 $Pr(\mathbf{X} = \mathbf{x}) = Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P_{\mathbf{X}}(\mathbf{x}) \text{ or } f_{\mathbf{X}}(\mathbf{x})$ 

- 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:

$$Pr(X_1 = x_1) = Pr(X_2 = x_2) = \dots = Pr(X_n = x_n)$$

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

$$Pr(\mathbf{X} = \mathbf{x}) = Pr(X_1 = x_1)Pr(X_2 = x_2)...Pr(X_n = x_n)$$

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

#### The Observed Sample!

- 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:

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

• As another example, for our measure heights / identity r.v., we could produce a sample of n=10 experimental trails, 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*

#### Samples



#### Statistics



#### 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):

$$T(\mathbf{x}) = T(x_1, x_2, ..., x_n) = t$$

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

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

 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  $\theta = \left[\mu, \sigma^2\right]$  / identity random variable
- If we calculate the following statistic:

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

what is  $\Pr(T(\mathbf{X}))$ ?

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

#### Statistics



#### 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:  $T(\mathbf{x}) = \hat{ heta}$
- There are many ways to define estimators (we will focus on maximum likelihood estimators in this course)
- 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 II

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

• 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:

$$Pr(T(\mathbf{X} = \mathbf{x})) = Pr(\hat{\theta})$$

 Our estimator takes in a vector as input (the sample) and may be defined to output a single value or a vector of estimates:

$$T(\mathbf{X} = \mathbf{x}) = \hat{\theta} = \begin{bmatrix} \hat{\theta_1}, \hat{\theta_2}, \dots \end{bmatrix}$$

- 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?)

#### Statistics



#### Estimators



#### Estimator example I

- As an example, let's construct an 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
- We want to estimate p, where a perfectly reasonable estimator is:

$$T(\mathbf{X} = \mathbf{x}) = \hat{\theta} = \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

• e.g. this statistic (=mean of the sample) would equal 0.5 for the following particular sample (will it always?)

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

#### Estimator example II

- Let's continue with our example constructing the probability model
- Consider the single coin flip experiment / number of tails random variable

$$\Omega = \{H, T\} \qquad X : X(H) = 0, X(T) = 1$$

Bernoulli probability model family (parameter p)

$$X \sim p^X (1-p)^{1-X}$$

• Sample of size *n*=10

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

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

 $[X_1 = x_1, X_2 = x_2, \dots, X_{10} = x_{10}] \sim p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\dots p^{x_{10}}(1-p)^{1-x_{10}}$ 

#### Estimator example II

• Define a statistic  $T(\mathbf{X})$ 

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

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



• Side note: we can write the sampling distribution (pmf) of the statistic as

$$Pr(T(\mathbf{X})) \sim {\binom{n}{nT(\mathbf{X})}} p^{nT(\mathbf{X})} (1-p)^{n-nT(\mathbf{X})}$$

• Remember for our sample, the value of our statistic for our observed sample (!!) would equal 0.5 (will it always?)

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

#### Statistics



#### 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

#### That's it for today

• Next lecture, we will continue our discussion of estimators by introducing Maximum Likelihood Estimators (MLE)!