Lecture 2: Introduction to probability basics

Jason Mezey
jgm45@cornell.edu
Jan. 30, 2018 (T) 8:40-9:55
Announcements

- Registration updates / reminders:
  - You must register for both the lecture and lab
  - In Ithaca, undergrads register for 4830 / grads for 6830 (please register if you can)!
  - For Post-docs, Continuing Education students - fill out, scan and send me any forms you need signed - I will sign, scan send back
  - For students at Weill: please register through “LEARN” system
  - For students at Rockefeller: email Kristen Cullen cullenk@mail.rockefeller.edu
  - For postdocs in NYC and students at MSK: Please fill out the “Application for Non-Degree Students” ASAP (!!) and get it to registrar@med.cornell.edu
  - You may take the class for a grade, S/U | P/F, or Audit (please register as an Audit if you can do so)
Announcements II

• Remember that the locations of the lecture in NYC and the computer lab can change lecture to lecture - week to week (!!) - see the class schedule now posted on the website (see next slide)

• First Computer lab is this week (!!):
  • In Ithaca, regardless of your registration, computer labs in Ithaca will be either Thurs. 5-6PM (!!) in MNLB30A (Mann Library Basement) OR Fri. 8-9AM in 224 Weill Hall
  • In Ithaca, you are welcome to go to either regardless of where you register (pending this working well for Manisha)
  • In NYC, computer lab is Thurs. 4-5PM in LC-504 - 5th Floor Conference Room, 1300 York Ave
  • Unless you are going to MNLB30A bring your laptop!
Announcements III

• Class website: http://mezeylab.cb.bscb.cornell.edu/
  • First 2018 class materials are now posted!
  • Check back often (!!)

• Jason will hold office hours every Thurs. 2-4PM (starting this Thurs.)!
  • We currently plan to use zoom: https://cornell.zoom.us/j/724550601

• You must therefore sign up for a Zoom account
• If you cannot do this, please contact me
Announcements IV

- MAKE SURE YOU SIGN UP ON PIAZZA whether you officially register or not = all course communication (!!): https://piazza.com/class/jckpr075ilk5n4?cid=7

- Step 1: Sign up on Piazza (if you don’t have an account already)!

- Step 2: Enroll in BTRY 6830 (regardless if you are grad or undergrad)

- If you cannot enroll email Manisha directly!!

- Question Posting Protocol:
  - Public posts (Let the community of students and instructors help out)
  - Private posts (To Jason, Manisha, Zijun)

- Please note that expected response times to questions will be minimum >24hrs (sometimes longer...) depending on the availability of the instructors

- We encourage public posts so that your classmates can help you out as well

- ONCE YOU ENROLL ON PIAZZA PLEASE DON’T email Jason / Manisha / Zijun’s on their direct email (unless it’s an emergency)
Announcements V

• MAKE SURE YOU SIGN UP ON CMS if you plan to do work for the course (registered or unregistered)

• If you do not have a NetID you will need to email Manisha and she will get you signed up

• Assignments will be posted on CMS (https://cms.csuglab.cornell.edu/) for BTRY 6830

• All submissions should be made through the CMS website - DO NOT email your homework directly to Jason / Manisha / Zijun (!!)
Announcements VI

• Homework #1 is posted (!!!) and will be due at 11:59PM, Feb. 5

• You must upload your homework by 11:59PM on Mon. 2/5 (otherwise it is late - no excuses!!)

• Answers must be typed (!!!) - please talk to us if this is a problem...

• Homeworks are “open book” and you may work together but you MUST hand in your own work (i.e., a copy of someone’s written answer will not be accepted)

• Problems will be divided into “easy”, “medium, and “hard”
Summary of lecture 2: Introduction to probability basics

- Last lecture, we provided a broad introduction to the field of *quantitative genomics and genetics*, which is concerned with *modeling* and the *discovery* of relationships between genomes (genotypes) and phenotypes.

- In this class, we will be concerned with the most basic problem of the field: how to identify genotypes where differences among individual genomes produce differences in individual phenotypes (e.g. genetic association studies).

- Today, we will discuss critical foundational concepts in biology and the modeling framework for the field, which is developed from the fields of probability and statistics.
Foundational biology concepts

• In this class, we will use statistical modeling to say something about biology, specifically the relationships between genotype (DNA) and phenotype.

• Let’s start with the biology by asking the following question: why DNA?

• The structure of DNA has properties that make it worthwhile to focus on...
It’s the same in all cells with a few exceptions (e.g. cancer, immune system...)
It’s passed on to the next generation
It has convenient structure for quantifying differences.
It’s responsible for the construction and maintenance of organisms.

Note: other regions of genomes can impact phenotypes...
• **Quantitative genomics** is a field concerned with the *modeling* of the relationship between genomes and phenotypes and using these models to *discover and predict*

• We will use frameworks from the fields of probability and statistics for this purpose

• Note that this is not the only useful framework (!!!) - and even more generally - mathematical based frameworks are not the only useful (or even necessarily “the best”) frameworks for this purpose
A non-technical definition of probability: a mathematical framework for modeling under uncertainty

Such a system is particularly useful for modeling systems where we don’t know and / or cannot measure critical information for explaining the patterns we observe

This is exactly the case we have in quantitative genomes when connecting differences in a genome to differences in phenotypes
Statistics and probability III

• We will therefore use a probability framework to model, but we are also interested in using this framework to discover and predict.

• More specifically, we are interested in using a probability model to identify relationships between genomes and phenotypes using DNA sequences and phenotype measurements.

• For this purpose, we will use the framework of statistics, which we can (non-technically) define as a system for interpreting data for the purposes of prediction and decision making given uncertainty.
Definitions: Probability / Statistics

- **Probability** (non-technical def) - a mathematical framework for modeling under uncertainty

- **Statistics** (non-technical def) - a system for interpreting data for the purposes of prediction and decision making given uncertainty

These frameworks are particularly appropriate for modeling genetic systems, since we are missing information concerning the complete set of components and relationships among components that determine genome-phenotype relationships.
Conceptual Overview

System

Question

Experiment

Sample

Inference

Prob. Models

Statistics

Assumptions
Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Example: Genetic contribution to height

```
? SNP
{ A T } Taller (on average)
{ A T } Shorter (on average)
```

Genome → Height
Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Examples: (1) coin, (2) heights in a population

  - Coin - same amount of metal on both sides?
  - Heights - what is the average height in the US?
Experiments (general)

• To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment.

• Examples of “ideal” experiments (System / Experiment):
  • SNP contribution to height / directly manipulate A -> T keeping all other genetic, environmental, etc. components the same and observe result on height
  • Coin / cut coin in half, melt and measure the volume of each half
  • Height / measure the height of every person in the US
Experiments (general)

• To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment.

• Examples of “non-ideal” experiments (System / Experiment):
  • SNP contribution to height / measure heights of individuals that have an A and individuals that have a T.
  • Coin / flip the coin and observe “Heads” and “Tails”.
  • Height / measure some people in the US.
Experiments and samples

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Experimental trial** - one instance of an experiment
- **Sample outcome** - a possible outcome of the experiment
- **Sample** - the results of one or more experimental trials

Example (Experiment / Sample outcomes):

- Coin flip / “Heads” or “Tails”
- Two coin flips / HH, HT, TH, TT
- Measure heights in this class / 1.5m, 1.71m, 1.85m, ...
Modeling the results of (non-ideal) experiments

- Mathematics (while not the only approach!) provides a particularly valuable foundation for describing or modeling a system or the outcomes of an experiment.

- The reason is that a considerable amount of mathematics is constructed (on purpose!) to provide a good representation of how we think about the world in a way that matches our intuition.

- Once constructed, we can use this modeling approach to formalize our intuition in a manner that has currency for others and develop deeper understanding.

- In general, mathematics useful for modeling (including probability) can be developed from foundations developed in set theory.

- A lot of assumptions, called axioms, are at the foundation of set theory put in place so that set theory produces logical constructions that match our intuition.
Sets / Set Operations / Definitions

- **Set** - any collection, group, or conglomerate
- **Element** - a member of a set
- **Set Operations:**
  - **Union** \( (\cup) \) \( \equiv \) an operator on sets which produces a single set containing all elements of the sets.
  - **Intersection** \( (\cap) \) \( \equiv \) an operator on sets which produces a single set containing all elements common to all of the sets.
- **Important Definitions:**
  - **Element of** \( (\in) \) \( \equiv \) an object within a set, e.g. \( H \in \{H, T\} \)
  - **Subset** \( (\subset) \) \( \equiv \) a set that is contained within another set, e.g. \( \{H\} \subset \{H, T\} \)
  - **Complement** \( (A^c) \) \( \equiv \) the set containing all other elements of a set other than \( A \), e.g. \( \{H\}^c = \{T\} \).
  - **Disjoint Sets** \( \equiv \) sets with no elements in common.
- **A Special Set:** **Empty Set** \( (\emptyset) \) \( \equiv \) the set with no elements (the empty set is unique and is sometimes represented as \( \{\} \)).
Some Special Sets

- The following sets have properties that align with our intuitive conception about how we represent and use groups

- The **Natural Numbers** and the **Integers**:
  \[\mathbb{N} = \{1, 2, 3, \ldots\}\]
  \[\mathbb{Z} = \{\ldots -3, -2, -1, 0, 1, 2, 3, \ldots\}\]

- The **Reals**:
  \[\mathbb{R} = \mathbb{Q} \cup \mathbb{R}\]

- Note that these sets are infinite (although they represent two different “sizes” of infinite: countable and uncountable), where we often make use of the following symbols in both cases:
  \[-\infty > x > \infty\]
Sample Spaces

- **Sample Space** ($\Omega$) - set comprising all possible outcomes associated with an experiment

- Examples (Experiment / Sample Space):
  - “Single coin flip” / \{H, T\}
  - “Two coin flips” / \{HH, HT, TH, TT\}
  - “Measure Heights” / any actual measurement OR we could use $\mathbb{R}$

- **Events** - a subset of the sample space

- Examples (Sample Space / Examples of Events):
  - “Single coin flip” / $\emptyset$, \{H\}, \{H, T\}
  - “Two coin flips” / \{TH,\}, \{HH, TH,\}, \{HT, TH, TT\}
  - “Measure Heights” / \{1.7m\}, \{1.5m, ..., 2.2m\} OR \[1.7m, (1.5m, 1.8m)\]
Functions

- Now that we have formalized the concept of a sample space, we need to define what “probability” means.

- To do this, we need the concept of a mathematical function.

- **Function** (formally) - a binary relation between every member of a domain to exactly one member of the codomain.

- **Function** (informally) - ?
Example of a function

\[ Y = X^2 \]
Probability functions (intuition)

• **Probability Function** (intuition) - we would like to construct a function that assigns a number to each event such that it matches our intuition about the “chance” the event will happen (as a result of an experiment)

• To be useful, we need to assign a number not just to each individual element of the set but to EVERY event

• To accomplish this, we will need the concept of a **Sigma Algebra** (or **Sigma Field**)

• What’s more, we need to make sure the function that we use to assign these numbers adheres to a specific set of “rules” (axioms)
Sample Spaces / Sigma Algebra

- **Sigma Algebra** ($\mathcal{F}$) - a collection of events (subsets) of $\Omega$ of interest with the following three properties: 1. $\emptyset \in \mathcal{F}$, 2. $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, 3. $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

  *Note that we are interested in a particular Sigma Algebra for each sample space...

- Examples (Sample Space / Sigma Algebra):
  - \{H,T\} / $\emptyset$, \{H\}, \{T\}, \{H,T\}
  - \{HH, HT, TH, TT\} /
    \begin{align*}
    \emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\} \\
    \{HH \cup HT\}, \{HH \cup TH\}, \{HH \cup TT\}, \{HT \cup TH\}, \{HT \cup TT\}, \{TH \cup TT\} \\
    \{HH \cup HT \cup TH\}, \{HH \cup HT \cup TT\}, \{HH \cup TH \cup TT\}, \{HT \cup TH \cup TT\} \\
    \{HH \cup HT \cup TH \cup TT\}
    \end{align*}
  - $\mathbb{R}$ / more complicated to define the sigma algebra of interest...
  - Note that the pair $(\Omega, \mathcal{F})$ is referred to as a measurable space
That’s it for today

- Next lecture, we will introduce random variables, random vectors, and parameterized probability models