Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 3: Introduction to probability basics

Jason Mezey
Feb 1, 2022 (T) 8:05-9:20
Announcements 1

• Website:
  • Still having a posting issue (we can put up videos / not decks) = we hope to resolved soon
  • Once resolved, we will also get up lectures from last year as well
• Everyone should now be up on Piazza (!!)
  • A test message + email was sent 10PM last night - please contact me ASAP if you did not receive it / are not up on Piazza yet!
  • Assuming everyone is now up, from now on, no direct emails (!!) = use Piazza!
• Everyone should be up on CMS (!!)
  • If you are not, contact me (by Piazza!) ASAP
• You need to be on CMS because your first homework is going to be assigned tomorrow (!!) - see next two slides
Announcements II

• Computer labs

• Remote again this week and OPTIONAL (but you must let your TA know you do NOT plan to attend if you are opting out!)

• Next week: computer labs will be in person (for students who are 100% remote, please stay tuned… we are working on how to allow to join the labs & lectures)

• This week, among other topics, your TA’s will introduce Latex (!!) which is extremely useful = if you’ve never used, even you planned to opt out - opt back in this week :)

• I will hold my first hours next week (!!)

• For the time being I will be holding office hours remotely = by zoom (if you have an issue with this, please message me and we can figure out a solution)

• Next week (due to my constraints) office hours will be Mon., Feb 7, 4:30-6:30PM

• This may not be the regular time… please stay tuned

• Stay tuned for the office hours zoom link, which will be distributed by Piazza message after class on Thurs.
Announcements III: Homework #1 (!!)

- Homework #1 will be posted on CMS tomorrow (Wed., Feb 2)
- Problems are divided into “easy”, “medium, and “hard”
- All 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 answers will not be accepted)
- Please feel free to attend office hours for help :)
- Answers must be typed (!!) including all equations - if this is a problem go to computer lab this week (= intro to Latex!)
- You must upload your answers to CMS (do not email!)
- You must upload your homework by 11:59PM, Tues., Feb 8 (otherwise it is late - no excuses!!)
Summary of lecture 3: Introduction to probability basics

• Last lecture, we introduced critical concepts about genetics and began our rigorous introduction to concepts in probability.

• In this lecture we will introduce more critical concepts in probability including rigorous definitions of sample spaces, sigma algebras and probability functions.
Review: 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

Inference

Experiment

Sample

Prob. Models

Statistics

Assumptions
Review: a system

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

- Example: Genetic contribution to height

\[
\text{Genome} \rightarrow \text{Height} \quad \begin{cases} \text{A} \rightarrow \text{Taller (on average)} \\ \text{T} \rightarrow \text{Shorter (on average)} \end{cases}
\]
Review: Experiments and Sample Outcomes

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe

- **Sample outcome** - a possible outcome of the experiment

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, …

(Note: we have not defined a **Sample** - we will do this later!)
Review: 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 sample space 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):
  - $\{H,T\} / \emptyset, \{H\}, \{T\}, \{H, T\}$
  - $\{HH, HT, TH, TT\} /$
    
    $\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{TH, HT, TT\}\{HH, TH, HT, TT\}$

  - $\mathbb{R}$ / more complicated to define the sigma algebra of interest (see next slide…)

- Note that the pair $(\Omega, \mathcal{F})$ is referred to as a measurable space
The (appropriate) Sigma Algebra on the Reals

- For probability, we need an appropriate Sigma Algebra on the Reals (remember there are many possible Sigma Algebra!)

- Interestingly, this Sigma Algebra does not include all subsets of the reals

- One problem is this would include “more sets than we need” for what we need in probability

- Another problem is these subsets include “non-measurable sets” such that if they were included, we could not define a probability measure (!!)

- A way of describing the appropriate Sigma Algebra for the Reals is all open and closed intervals (where \( a \) and \( b \) may be any number) and all unions and intersections of these intervals:

  \[
  [a, b], (a, b], [a, b), (a, b)
  \]

- It seems like these should include all subsets of the Reals, but they don’t...
Probability functions

- **Probability Function** - maps a Sigma Algebra of a sample to a subset of the reals:

\[
Pr(\mathcal{F}) : \mathcal{F} \rightarrow [0, 1]
\]

- Not all such functions that map a Sigma Algebra to \([0, 1]\) are probability functions, only those that satisfy the following Axioms of Probability (where an axiom is a property assumed to be true):

1. For \(A \subseteq \Omega\), \(Pr(A) \geq 0\)
2. \(Pr(\Omega) = 1\)
3. For \(A_1, A_2, \ldots \in \Omega\), if \(A_i \cap A_j = \emptyset\) (disjoint) for each \(i \neq j\): \(Pr(\bigcup_{i=1}^{\infty} A_i) = \sum_{i}^{\infty} Pr(A_i)\)

- Note that since a probability function takes sets as an input and is restricted in structure, we often refer to a probability function as a *probability measure*

- Note that the pair \((\Omega, \mathcal{F})\) is referred to as a measurable space and the triple \((\Omega, \mathcal{F}, Pr)\) is referred to as a measure space
Probability function: example 1

• For “two coin flips” a probability function will assign a probability to each subset of the Sigma Field:

\[
\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{TH, HT, TT\}\{HH, TH, HT, TT\}
\]

• We could define a probability function as follows:

\[
Pr(\emptyset) = 0
\]

\[
Pr(\{HH\}) = 0.25, Pr(\{HT\}) = 0.25, Pr(\{TH\}) = 0.25, Pr(\{TT\}) = 0.25
\]

\[
Pr(\{HH, HT\}) = 0.5, Pr(\{HH, TH\}) = 0.5, Pr(\{HH, TT\}) = 0.5,
\]

\[
Pr(\{HT, TH\}) = 0.5, Pr(\{HT, TT\}) = 0.5, Pr(\{TH, TT\}) = 0.5,
\]

\[
Pr(\{HH, HT, TH\}) = 0.75, \text{ etc. } Pr(\{HH, HT, TH, TT\}) = 1.0
\]

• Not that this is one possible probability model - what other possible probability models could be assumed for this system / experiment?
Probability function: example II

- The following is (one example) of a probability function (on the sigma algebra) for the two coin flip experiment:

\[
Pr(\emptyset) = 0
\]

\[
Pr(\{HH\}) = 0.25, Pr(\{HT\}) = 0.25, Pr(\{TH\}) = 0.25, Pr(\{TT\}) = 0.25
\]

\[
Pr(\{HH, HT\}) = 0.5, Pr(\{HH, TH\}) = 0.5, Pr(\{HH, TT\}) = 0.5,
\]

\[
Pr(\{HT, TH\}) = 0.5, Pr(\{HT, TT\}) = 0.5, Pr(\{TH, TT\}) = 0.5,
\]

\[
Pr(\{HH, HT, TH\}) = 0.75, \text{ etc.} \quad Pr(\{HH, HT, TH, TT\}) = 1.0
\]

- The following is an example of a function (on the sigma algebra) of the two coin flip experiment but is not a probability function:

\[
\mathcal{F}(\emptyset) = 0
\]

\[
\mathcal{F}(\{HH\}) = 0.25, \mathcal{F}(\{HT\}) = 0.25, \mathcal{F}(\{TH\}) = 0.25, \mathcal{F}(\{TT\}) = 0.25
\]

\[
\mathcal{F}(\{HH, HT\}) = 0.5, \mathcal{F}(\{HH, TH\}) = 0.5, \mathcal{F}(\{HH, TT\}) = 1.0,
\]

\[
\mathcal{F}(\{HT, TH\}) = 0, \mathcal{F}(\{HT, TT\}) = 0.5, \mathcal{F}(\{TH, TT\}) = 0.5,
\]

\[
\mathcal{F}(\{HH, HT, TH\}) = 0.75, \text{ etc.} \quad \mathcal{F}(\{HH, HT, TH, TT\}) = 1.0
\]
That’s it for today

- Next lecture, we will continue our discussion of probability by introducing the concept of conditional probability, independence, and random variables!