BTRY 4830/6830: Quantitative Genomics and Genetics

Lecture 2: Introduction to probability basics

Jason Mezey
jgm45@cornell.edu
Aug. 28, 2014 (Th) 8:40-9:55
Announcements

- Registration updates / reminders:
  - You must register for both the lecture and lab
  - If you are in Ithaca and could not register previously, you can now register for the lecture and lab - register for either lab (you will go to the same computer lab regardless of your registered lab)
  - In Ithaca, undergrads register for 4830 / grads for 6830
  - You may take the class for a grade, S/U | P/F, or Audit
  - You are also welcome to sit in this class, but if you choose to do this and are able to register for the class please register for an “Audit” (note that we will grade any work you turn in!)
Announcements

- Computer lab - FIRST LAB IS TODAY (!!):
  - Regardless of your registration, computer labs will be each week, Thurs. 5-6PM
  - If you have an unavoidable conflict with this time, please contact me ASAP
  - In Ithaca the lab will be in MNLB30A (Mann Library Basement)
  - At WCMC, the lab will be in the Dept. of Genetic Med. conference (= lecture)
  - In Ithaca, please bring your laptop this week (you will likely not have to in subsequent weeks), at WCMC bring your laptop every week (!!)
Announcements III

- Class office hours:
  - Jason - Thurs. 3-5PM in 101 Biotech Suite AND Dept. Genetic Med. conference (BOTH Ithaca and Weill)
  - Amanda - Tues. 3-5PM in 343 Weill Hall (Ithaca only)
  - Jin - no office hours
  - Unofficial office hours can be scheduled by appointment (!!)
  - These will start NEXT week = NO office hours this week!!

- Class email listserv: MEZENY-QUANTGENOME-L@cornell.edu
  - email Amanda to get on (or off) list: yg246@cornell.edu

- Test message will be sent tomorrow afternoon

- Class website: http://mezeylab.cb.bscb.cornell.edu/
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).

- The modeling framework for the field is developed from the fields of probability and statistics.
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

```
Genome → Height

? SNP { A, T } → Taller (on average), Shorter (on average)
```
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 / 5’, 5’3”, 5’3.5, ...
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 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 (e.g. the ZFC axioms).

- There is no perfect set of axioms for all mathematics (e.g. see Godel’s incompleteness theorems).
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 and Infinite:

- 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:**
  \[ \mathbb{R} = \{\leftarrow 0 \rightarrow\} \]

- Note that these sets are infinite (although they represent two different “sizes” of infinite!), where we often make use of the following symbols:
  \[ -\infty > x > \infty \]
Sample Spaces / Sigma Algebra

- **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” / $\{5', 5'3'', 5'3.5'', ...\}$

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

- **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. $A_1, A_2, ... \in \mathcal{F}$ 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):
  - $\{H, T\}$ / $\emptyset, \{H\}, \{T\}, \{H, T\}$
  - $\{HH, HT, TH, TT\}$ / see board
  - $\{5', 5'3'', 5'3.5'', ...\}$ / let’s table this one for the moment
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 element of a domain set to exactly one element of the codomain set.

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

\[ Y = X^2 \]
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 this subset are probability functions, only those that satisfy the following Axioms of Probability (where an axiom is a property assumed to be true):

1. For \( \mathcal{A} \subset \Omega \), \( Pr(\mathcal{A}) \geq 0 \)
2. \( Pr(\Omega) = 1 \)
3. For \( \mathcal{A}_1, \mathcal{A}_2, \ldots \in \Omega \), if \( \mathcal{A}_i \cap \mathcal{A}_j = \emptyset \) (disjoint) for each \( i \neq j \): \( Pr(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} Pr(\mathcal{A}_i) \)
Probability functions on the reals

- In any realistic case, our true sample outcomes will be discrete.
- However, we often model cases using the real numbers as a sample space, which has nice properties and mathematical tools available that we can take leverage.
- Example: we often use the real numbers as the sample space for an experiment where the sample outcome is human heights.

- Two questions:
  - What approximations are we making?
  - Why are these approximations reasonable?
The Sigma Field on the reals

- To define a probability function, we need the appropriate Sigma Field on the reals.
- Interestingly, we cannot use all of the subsets of the reals. The problem is these subsets include “non-measurable sets” such that if they were included, we could not define a probability measure.
- It turns out the appropriate Sigma Field for the reals includes all open and closed intervals (where \( a \) and \( b \) may be any number):
  \[
  [a, b], (a, b], [a, b), (a, b)
  \]
- It seems like these should include all subsets of the reals, but they don’t...
Thoughts about what probability is modeling

- We are attempting to model the results of a non-ideal experiment to understand a system.
- Such experiments include extensive amounts of uncontrolled aspects (important for the system!) that we usually cannot specify.
- What we may be able to do is provide a reasonable model of how these uncontrolled aspects impact the results of the experiments.
- More specifically, we assume that the impact of the uncontrolled aspects are random but where certain outcomes are more probable than others (note the assumption!!)
- This is what a probability function is built to model (= to provide the probability of random outcomes of an experiment).
- Note that while random is intuitive, it’s a problematic concept...
An essential concept: conditional probability (and independence)

- As well as having an intuitive sense of what it means for something we observe to be random (within definable rules) we also have an intuitive sense about how the rules change once we observe specific outcomes or assume certain possibility applies.
- This intuition is captured in conditional probability.
- This is the essential concept in any area of probabilistic modeling, where the concept of independence directly follows.
- In fact, almost anything we are doing in statistics, machine learning, etc. is really attempting to identify or leverage conditional probabilities.
- As an example, we could consider the conditional probability that someone will be taller or shorter if they have a “T” at a particular position in the genome.
Conditional probability

- We have an intuitive concept of *conditional probability*: the probability of an event, given another event has taken place.
- We will formalize this using the following definition (note that this is still a probability!!):

\[
Pr(A_i|A_j) = \frac{Pr(A_i \cap A_j)}{Pr(A_j)}
\]

- While not obvious at first glance, this is actually an intuitive definition that matches our conception of conditional probability.
An example of conditional prob.

- Consider the sample space of “two coin flips” and the following probability model: $Pr\{HH\} = Pr\{HT\} = Pr\{TH\} = Pr\{TT\} = 0.25$

<table>
<thead>
<tr>
<th></th>
<th>$H_{2nd}$</th>
<th>$T_{2nd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1st}$</td>
<td>$HH$</td>
<td>$HT$</td>
</tr>
<tr>
<td>$T_{1st}$</td>
<td>$TH$</td>
<td>$TT$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$H_{2nd}$</th>
<th>$T_{2nd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1st}$</td>
<td>$Pr(H_{1st} \cap H_{2nd})$</td>
<td>$Pr(H_{1st} \cap T_{2nd})$</td>
</tr>
<tr>
<td>$T_{1st}$</td>
<td>$Pr(T_{1st} \cap H_{2nd})$</td>
<td>$Pr(T_{1st} \cap T_{2nd})$</td>
</tr>
<tr>
<td></td>
<td>$Pr(H_{2nd})$</td>
<td>$Pr(T_{2nd})$</td>
</tr>
</tbody>
</table>

$Pr(H_{1st}) = Pr(HH \cup HT)$, $Pr(H_{2nd}) = Pr(HH \cup TH)$

$Pr(T_{1st}) = Pr(TH \cup TT)$, $Pr(T_{2nd}) = Pr(HT \cup TT)$
An example of conditional prob.

- Intuitively, if we condition on the first flip being “Heads”, we need to rescale the total to be one (to be a probability function):

<table>
<thead>
<tr>
<th></th>
<th>( H_{2nd} )</th>
<th>( T_{2nd} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{1st} )</td>
<td>( HH )</td>
<td>( HT )</td>
</tr>
<tr>
<td>( T_{1st} )</td>
<td>( TH )</td>
<td>( TT )</td>
</tr>
</tbody>
</table>

Here is an intuitive way to think about what is going on. If we know that the first flip is a ‘Heads’, we are now dealing with a ‘new’ sample space that contains two elements: $\text{HH}$ and $\text{HT}$. Conceptually of defining the first flip to be ‘Heads’ we now are in the first row of the original sample space.

If we condition on the first flip being “Heads”, we need to rescale the total to be one (to be a probability function): $\pi = \frac{\{HH,HT\}}{p}$. For example, in a fair coin example, $p = \frac{1}{2}$, so $\pi = \frac{1}{2}$.

This conditional probability is therefore: $\pi = \{HH,HT\}$, which we can do by defining $p = \{\text{HH}\}$.

The probability of the first row in the original sample space is $\{HH\}$, which has probability $\pi = \{HH,HT\}$.

Here is an intuitive way to think about what is going on. If we know that the first flip is a ‘Heads’, we need to rescale the total to be one (to be a probability function): $\pi = \frac{\{HH,HT\}}{p}$. For example, in a fair coin example, $p = \frac{1}{2}$, so $\pi = \frac{1}{2}$.

This conditional probability is therefore: $\pi = \{HH,HT\}$, which we can do by defining $p = \{\text{HH}\}$.

The probability of the first row in the original sample space is $\{HH\}$, which has probability $\pi = \{HH,HT\}$.
Independence

• The definition of *independence* is another concept that is not particularly intuitive at first glance, but it turns out it directly follows our intuition of what “independence” should mean and from the definition of conditional probability

• Specifically, we intuitively think of two events as “independent” if knowing that one event has happened does not change the probability of a second event happening

• i.e. the first event provides us no insight into what will happen second
Independence

- This requires that we define independence as follows:

  If $A_i$ is independent of $A_j$, then we have:

  $$Pr(A_i|A_j) = Pr(A_i)$$

- Why is this? It follows from the definition of conditional prob.:

  $$Pr(A_i|A_j) = \frac{Pr(A_i \cap A_j)}{Pr(A_j)} = \frac{Pr(A_i)Pr(A_j)}{Pr(A_j)} = Pr(A_i)$$

- This in turn produces the following relation for independent events:

  $$Pr(A_i \cap A_j) = Pr(A_i)Pr(A_j)$$
Example of independence

Consider the sample space of “two coin flips” and the following probability model:

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

\[
\begin{array}{c|c|c|c}
 & H_{2nd} & T_{2nd} & Pr(H_{1st}) \\
\hline
H_{1st} & Pr(H_{1st} \cap H_{2nd}) & Pr(H_{1st} \cap T_{2nd}) & Pr(H_{1st}) \\
\hline
T_{1st} & Pr(T_{1st} \cap H_{2nd}) & Pr(T_{1st} \cap T_{2nd}) & Pr(T_{1st}) \\
\hline
& Pr(H_{2nd}) & Pr(T_{2nd}) &
\end{array}
\]

\[
\begin{array}{c|c|c|c}
 & H_{2nd} & T_{2nd} \\
\hline
H_{1st} & 0.25 & 0.25 & 0.5 \\
\hline
T_{1st} & 0.25 & 0.25 & 0.5 \\
\hline
& 0.5 & 0.5 &
\end{array}
\]

In this model, \( H_{1st} \) and \( H_{2nd} \) are independent, i.e. \( Pr(H_{1st} \cap H_{2nd}) = Pr(H_{1st})Pr(H_{2nd}) \).
Example of non-independence

- Consider the sample space of “two coin flips” and the following probability model:

<table>
<thead>
<tr>
<th></th>
<th>$H_{2nd}$</th>
<th>$T_{2nd}$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1st}$</td>
<td>$Pr(H_{1st} \cap H_{2nd})$</td>
<td>$Pr(H_{1st} \cap T_{2nd})$</td>
<td>$Pr(H_{1st})$</td>
</tr>
<tr>
<td>$T_{1st}$</td>
<td>$Pr(T_{1st} \cap H_{2nd})$</td>
<td>$Pr(T_{1st} \cap T_{2nd})$</td>
<td>$Pr(T_{1st})$</td>
</tr>
<tr>
<td></td>
<td>$Pr(H_{2nd})$</td>
<td>$Pr(T_{2nd})$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$H_{2nd}$</th>
<th>$T_{2nd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1st}$</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>$T_{1st}$</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

In this model $H_{1st}$ and $H_{2nd}$ are not independent, i.e. $Pr(H_{1st} \cap H_{2nd}) \neq Pr(H_{1st})Pr(H_{2nd})$.
That’s it for today

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