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

Lecture 4: Introduction to probability basics II

Jason Mezey
jgm45@cornell.edu
Feb. 18, 2021 (Th) 8:05-9:20
Announcements I: Computer Labs (!!)

Computer Lab Information: Zoom links

Hi All,

A reminder that your first computer lab is this week (!) tomorrow (Thurs., Feb. 18) and Fri. (Feb. 19).

Please remember that if you are physically located in ITHACA you need to attend one of the two Ithaca labs (taught by Beulah - see zoom links [1] below) and if you are physically located in NYC you need to attend one of the two NYC labs (taught by Scott - see zoom links [2] below). Beyond this restriction, which of the two computer labs open to you that you choose to attend is up to you - and please contact your TA if you have additional questions (e.g., policy on switching any given week etc.).

[1] ZOOM LINKS FOR ITHACA LABS:

Topic: Quant Gen Lab (Thurs - Ithaca)
Time: Feb 18, 2021 05:00 PM Eastern Time (US and Canada)
Every week on Thu, until May 13, 2021, 13 occurrence(s)
Join Zoom Meeting
https://cornell.zoom.us/j/94925518722?pwd=UGd2R0ZkNGI1YVdacShVNFAvTExhdxz09

Meeting ID: 949 2551 8722
Passcode: 583217

AND-----------------------------------------------

Topic: Quant Gen Lab (Fri - Ithaca)
Time: Feb 19, 2021 08:00 AM Eastern Time (US and Canada)
Every week on Fri, until May 14, 2021, 13 occurrence(s)
Join Zoom Meeting
https://cornell.zoom.us/j/94042593474?pwd=cz83U01aOHFMOMyVSnUSQkeNSkM3dz09

Meeting ID: 940 4259 3474
Passcode: 285230

[2] ZOOM LINKS FOR NYC LABS:

Topic: Quant Gen Lab (Thurs - NYC)
Time: Feb 18, 2021 04:00 PM Eastern Time (US and Canada)
Every week on Thu, until May 13, 2021, 13 occurrence(s)

Thursday: https://weillcornell.zoom.us/j/98525965972
passcode: gene

AND-----------------------------------------------

Topic: Quant Gen Lab (Fri - Ithaca)
Time: Feb 19, 2021 09:00 AM Eastern Time (US and Canada)
Every week on Fri, until May 14, 2021, 13 occurrence(s)
Friday: https://weillcornell.zoom.us/j/97957447110
passcode: gene
Announcements II: CMS (and Homework #1)

• Not everyone is up on CMS yet… We hope to have this resolved today (we will message you when everyone is up)

• In general, you will download your homeworks (etc.) from CMS and upload your work to CMS

• Your 1st Homework is assigned / will be available today (!!) - but since not everyone is on CMS, this homework ONLY will be available on the class website (where the lectures are posted…)

• Homework #1 will be due in a week and you will submit on CMS
Announcements III: Homework #1 (!!)

• Homework #1 will be posted on the class website TODAY (Thurs.)

• 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 answer will not be accepted)

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

• We will ask you to upload your answers to CMS

• You must upload / email your homework by 11:59PM, Thurs., Feb 25 (otherwise it is late - no excuses!!)
Summary of lecture 4: Introduction to probability basics II

• Last lecture, we began our introduction to probability by discussing sample spaces and sigma algebras.

• Today we will continue adding fundamental probability concepts including probability functions and conditional probability (and independence!)
Review: 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.
Review: Sets / Set Operations

- **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 \( \{ \} \)).
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)
Review: 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**)
Review: 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:

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

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

- 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:

\[
Pr(\emptyset) = 0
\]

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

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

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

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

- 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:

\[ Pr(\emptyset) = 0 \]
\[ Pr(HH) = 0.25, Pr(HT) = 0.25, Pr(TH) = 0.25, Pr(TT) = 0.25 \]
\[ Pr(HH \cup HT) = 0.5, Pr(HH \cup TH) = 0.5, Pr(HH \cup TT) = 0.5 \]
\[ Pr(HT \cup TH) = 0.5, Pr(HT \cup TT) = 0.5, Pr(TH \cup TT) = 0.5 \]
\[ Pr(HH \cup HT \cup TH) = 0.75, \text{ etc.} \ Pr(HH \cup HT \cup TH \cup TT) = 1.0 \]

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

\[ Pr(\emptyset) = 0 \]
\[ Pr(HH) = 0.25, Pr(HT) = 0.25, Pr(TH) = 0.25, Pr(TT) = 0.25 \]
\[ Pr(HH \cup HT) = 0.5, Pr(HH \cup TH) = 0.5, Pr(HH \cup TT) = 0 \]
\[ Pr(HT \cup TH) = 1.0, Pr(HT \cup TT) = 0.5, Pr(TH \cup TT) = 0.5 \]
\[ Pr(HH \cup HT \cup TH) = 0.75, \text{ etc.} \ Pr(HH \cup HT \cup TH \cup TT) = 1.0 \]
Probability functions on the Reals

• In any realistic case, our true sample outcomes will be discrete

• However we often model cases where the Reals have nice properties and mathematical tools available that we can leverage

• An intuitive way to conceptualize how we make use of the Reals is when we use them for the sample space (!!)

• Note: statisticians sometimes conceptualize sample spaces as abstract and a mapping to the Reals for random variables… we’ll discuss later

• 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?
Probability functions on the Reals

- For a sample space that is the Reals, recall that a probability function has to assign (appropriate) probabilities to the Sigma Algebra: open and closed (and half open-closed) intervals and unions, intersections

- Taking just the one to one mapping of a function on the reals would not assign intervals (subsets of the Sigma Algebra) a probability

- ...unless we are careful about the construction and how we make use of a function (see “Random Variables” next lecture!)
Additional thought about what probability is modeling

• Again, note that our eventual objective is to 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...
Essential concepts: 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>

\[
\begin{array}{ccc}
  \text{\(H_{1st}\)} & \text{\(Pr(H_{1st} \cap H_{2nd})\)} & \text{\(Pr(H_{1st} \cap T_{2nd})\)} & \text{\(Pr(H_{1st})\)} \\
  \text{\(T_{1st}\)} & \text{\(Pr(T_{1st} \cap H_{2nd})\)} & \text{\(Pr(T_{1st} \cap T_{2nd})\)} & \text{\(Pr(T_{1st})\)} \\
  \text{\(H_{2nd}\)} & \text{\(Pr(H_{2nd})\)} & \text{\(Pr(T_{2nd})\)} & \text{\(Pr(T_{2nd})\)} \\
\end{array}
\]

\[
Pr(H_{1st}) = Pr(HH \cup HT), \quad Pr(H_{2nd}) = Pr(HH \cup TH) \\
Pr(T_{1st}) = Pr(TH \cup TT), \quad Pr(T_{2nd}) = Pr(HT \cup TT)
\]
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>

<table>
<thead>
<tr>
<th></th>
<th>$H_{2nd}$</th>
<th>$T_{2nd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1st}$</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$T_{1st}$</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
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):

\[
\begin{array}{c|cc}
 & H_{2nd} & T_{2nd} \\
\hline
H_{1st} & HH & HT \\
T_{1st} & TH & TT \\
\end{array}
\]

Here is an intuitive way to think about what is going on. If we know that the first flip is a ‘Heads’, this limits the outcomes to

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

Thus in our fair coin example, where the probability of ‘Heads’ or ‘Tails’ is

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

\[
\begin{array}{|c|c|c|}
\hline
\text{H}_{1st} & \text{H}_{2nd} & \text{T}_{2nd} \\
\hline
\text{H}_{1st} & \text{HH} & \text{HT} \\
\text{T}_{1st} & \text{TH} & \text{TT} \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
\text{H}_{1st} & \text{H}_{2nd} & \text{T}_{2nd} \\
\hline
0.25 & 0.25 & 0.5 \\
0.25 & 0.25 & 0.5 \\
0.5 & 0.5 & \text{---} \\
\hline
\end{array}
\]

\[
Pr(H_{2nd}|H_{1st}) = \frac{Pr(H_{2nd} \cap H_{1st})}{Pr(H_{1st})} = \frac{Pr(HH)}{Pr(HH \cup HT)} = \frac{0.25}{0.5} = 0.5
\]
• Next lecture, we will continue our discussion of probability by introducing random variables!