

Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 3: Conditional Probability and Random Variables

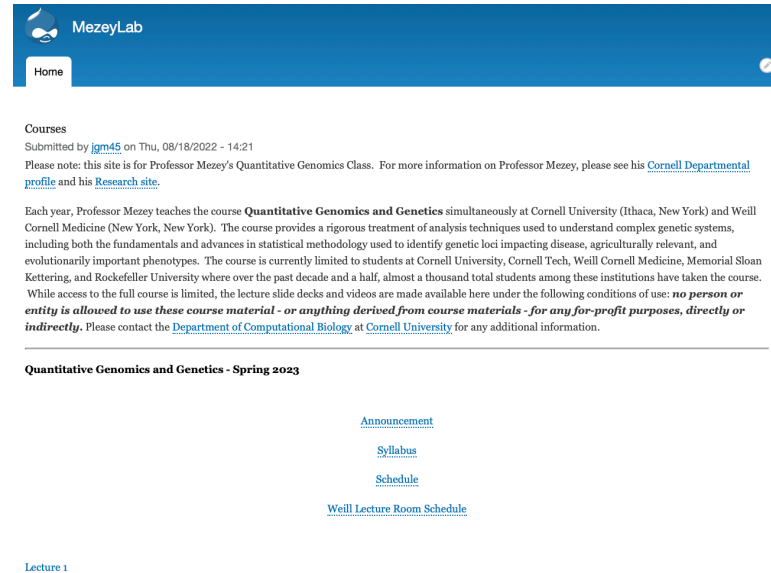
Jason Mezey
Jan 31, 2023 (T) 8:05-9:20

Piazza (!!)

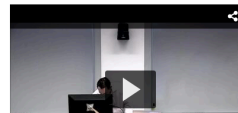
- MAKE SURE YOU ARE SIGNED UP ON PIAZZA whether you officially register or not = all communication for the course (!!)
- Class: <https://piazza.com/cornell/spring2023/btry4830btry6830>
- If you are not be able to sign up PLEASE EMAIL ME ASAP at [**jgm45@cornell.edu**](mailto:jgm45@cornell.edu) and I will get you on
- Note: we may be getting some annoying advertisements starting in a week or so... (we'll see how annoying these are and assess)

Class Resource: Website

- The website is now up (!!): <https://mezeylab.biohpc.cornell.edu>



The screenshot shows the MezeyLab website homepage. At the top is a blue navigation bar with the MezeyLab logo and a 'Home' button. Below the navigation bar, the page is titled 'Courses' and includes a submission date of 'Submitted by jgm45 on Thu, 08/18/2022 - 14:21'. A note states: 'Please note: this site is for Professor Mezey's Quantitative Genomics Class. For more information on Professor Mezey, please see his [Cornell Departmental profile](#) and his [Research site](#).' The main text describes the course 'Quantitative Genomics and Genetics' taught by Professor Mezey at Cornell University and Weill Cornell Medicine. It mentions that the course is limited to students at these institutions and that lecture slide decks and videos are available under specific conditions of use. Below the text are several links: 'Announcement', 'Syllabus', 'Schedule', and 'Weill Lecture Room Schedule'. At the bottom left, there is a link for 'Lecture 1'.



- This has the syllabus, calendar AND NYC room calendar (!!)
- This also has videos for this year so far (and last year) AND I try to post lecture decks before each lecture (!!)

Class Resource: CMS

- Assignments and computer labs (!!) will be posted on Cornell CMS (as BTRY 4830)
- Class CMS is up (!!): <https://cmsx.cs.cornell.edu/web/guest/>
- If you have a NetID you should be able to access and register for the class CMS site
- If you have a CWID (i.e., you are at Weill) we are adding you to CMS (stay tuned)
- We will be sending a Piazza message to ask you to login to CMS tomorrow (Weds., Feb 1) as a test
- We will be posting your homework #1 on CMS on Thurs (Feb 2)

Times and Locations I

- Lectures are every Tues. / Thurs. 8:05-9:20AM - see class schedule (to be posted)
- In-person lecture locations:
 - Ithaca: All in-person lectures in Weill Hall 226
 - NYC: Many different locations (!!) SEE POSTED SCHEDULE (on website)
- Zoom option:
 - Anyone may join by zoom for any lecture (I still encourage you to come to class...)
 - The zoom link has been shared with you by Piazza message
 - PLEASE DO NOT SHARE BEYOND THE CURRENT CLASS (e.g., if we get zoom-bombed, we may need to remove the option...)

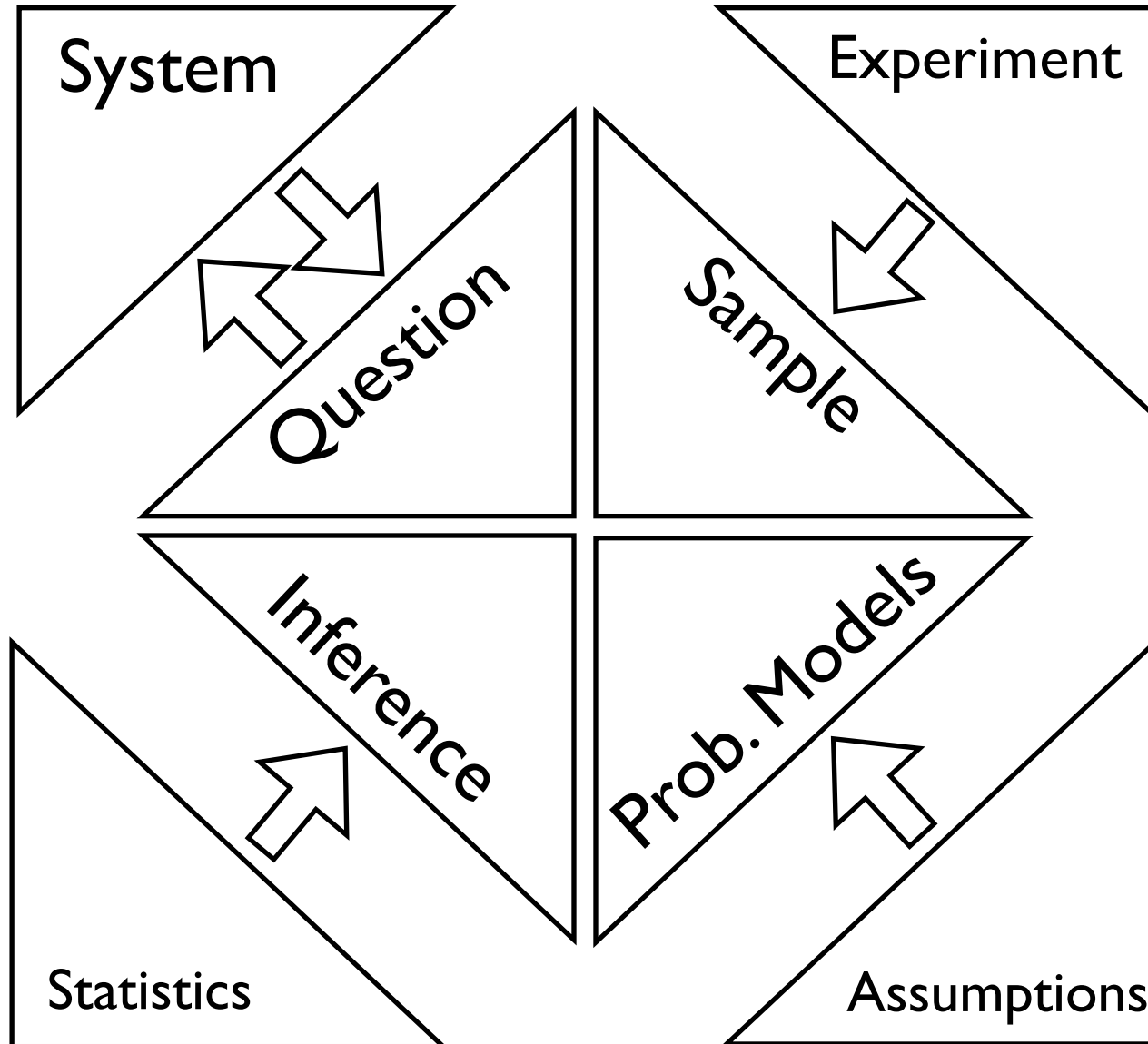
Times and Locations II

- **FIRST COMPUTER LAB IS THIS WEEK (Thurs. Feb 2 / Fri. Feb 3) - more information to come this week!**
- PLEASE NOTE THE LAB TIMES (!!)
- For those IN ITHACA (= Labs with Mitch!):
 - Lab 1: 5:30-6:30PM on Thurs. (Weill Hall 226)
 - Lab 2: 8-9AM on Fri. (Weill Hall 226)
 - Please go to the Lab you registered for (!!)
- For those IN NYC (= Labs taught by Sam!):
 - Lab 1: 4-5PM on Thurs. (In WCMCI 300 Classroom; G [B215], H [B217])
 - Lab 2: 9-10AM on Fri. (By zoom - Sam will distribute the invite)
 - PLEASE NOTE: if you are in HOUSTON or you are VERY EXPERIENCED with R, please join Fri (!!) - otherwise, join on Thurs!
- You may skip the first 2 labs without penalty BUT
 - If you are not VERY familiar with R programming you may want to go
 - If you do not already use Latex you may want to go (e.g., homeworks!)

Summary of lecture 3: Introduction to conditional probability and random variables

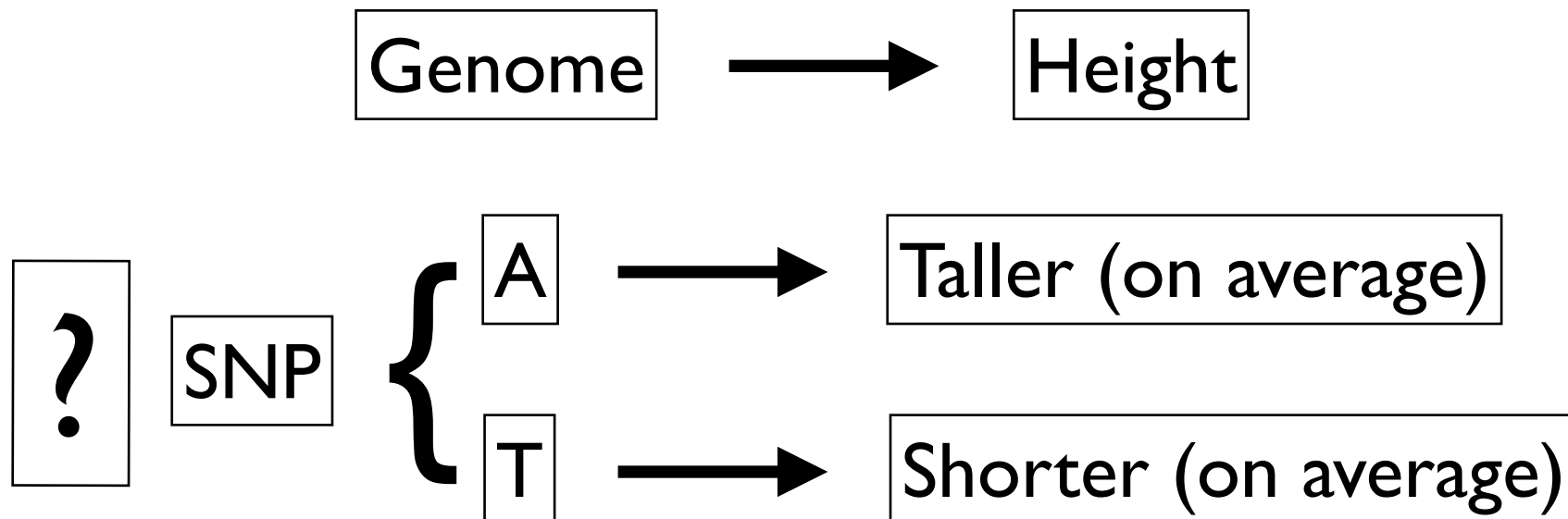
- Last class, we introduced the foundations needed to define / the definition of a probability function (=model)!
- Today we will discuss TWO critical concepts: conditional probability AND random variables (!!)

Conceptual Overview



Review: a system

- **System** - a process, an object, etc. which we would like to know something about
- Example: Genetic contribution to height



Review: Experiments and Outcomes

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Experiment Outcome** - a possible result of the experiment
- Example (Experiment / Outcomes):
 - Coin flip / “Heads” or “Tails”
 - Two coin flips / HH, HT, TH, TT
 - Measure heights in this class / 1.5m, 1.71m, 1.85m, ...

Review: Set Definitions

- **Set** - any collection, group, or conglomerate
- **Element** - a member of a set
- **A Special Set:** **Empty Set** (\emptyset) \equiv the set with no elements (the empty set is unique and is sometimes represented as $\{ \}$).
- **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** (\mathcal{A}^c) \equiv the set containing all other elements of a set other than \mathcal{A} , e.g. $\{H\}^c = \{T\}$.
 - Disjoint Sets** \equiv sets with no elements in common.

Review: Sample Spaces

- **Sample Space** (Ω) - set comprising all possible outcomes associated with an experiment
- (Note: we have not defined a **Sample** - we will do this later!)
- 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” / $\{1.7m\}, \{1.5m, \dots, 2.2m\}$ OR $[1.7m], (1.5m, 1.8m)$

Review: Probability functions I

- **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 $\mathcal{A} \subset \Omega$, $Pr(\mathcal{A}) \geq 0$

2. $Pr(\Omega) = 1$

3. For $\mathcal{A}_1, \mathcal{A}_2, \dots \in \Omega$, if $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ (disjoint) for each $i \neq j$: $Pr(\bigcup_i^\infty \mathcal{A}_i) = \sum_i^\infty Pr(\mathcal{A}_i)$

- 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*

Review: Probability functions II

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

$$Pr(\emptyset) = 0$$

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

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

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

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

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

$$\cancel{Pr}(\emptyset) = 0$$

$$\cancel{Pr}(\{HH\}) = 0.25, \cancel{Pr}(\{HT\}) = 0.25, \cancel{Pr}(\{TH\}) = 0.25, \cancel{Pr}(\{TT\}) = 0.25$$

$$\cancel{Pr}(\{HH, HT\}) = 0.5, \cancel{Pr}(\{HH, TH\}) = 0.5, \cancel{Pr}(\{HH, TT\}) = 1.0,$$

$$\cancel{Pr}(\{HT, TH\}) = 0, \cancel{Pr}(\{HT, TT\}) = 0.5, \cancel{Pr}(\{TH, TT\}) = 0.5,$$

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

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!!):

The formal definition of the conditional probability of \mathcal{A}_i given \mathcal{A}_j is:

$$Pr(\mathcal{A}_i|\mathcal{A}_j) = \frac{Pr(\mathcal{A}_i \cap \mathcal{A}_j)}{Pr(\mathcal{A}_j)}$$

- 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: $Pr\{HH\} = Pr\{HT\} = Pr\{TH\} = Pr\{TT\} = 0.25$

	H_{2nd}	T_{2nd}
H_{1st}	HH	HT
T_{1st}	TH	TT

	H_{2nd}	T_{2nd}	
H_{1st}	$Pr(H_{1st} \cap H_{2nd})$	$Pr(H_{1st} \cap T_{2nd})$	$Pr(H_{1st})$
T_{1st}	$Pr(T_{1st} \cap H_{2nd})$	$Pr(T_{1st} \cap T_{2nd})$	$Pr(T_{1st})$
	$Pr(H_{2nd})$	$Pr(T_{2nd})$	

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

$$Pr(T_{1st}) = Pr(\{TH\} \cup \{TT\}) \quad Pr(T_{2nd}) = Pr(\{HT\} \cup \{TT\})$$

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):

	H_{2nd}	T_{2nd}
H_{1st}	HH	HT
T_{1st}	TH	TT

	H_{2nd}	T_{2nd}	
H_{1st}	0.25	0.25	0.5
T_{1st}	0.25	0.25	0.5
	0.5	0.5	

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):

	H_{2nd}	T_{2nd}
H_{1st}	HH	HT
T_{1st}	TH	TT

	H_{2nd}	T_{2nd}	
H_{1st}	0.25	0.25	0.5
T_{1st}	0.25	0.25	0.5
	0.5	0.5	

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):

	H_{2nd}	T_{2nd}
H_{1st}	HH	HT
T_{1st}	TH	TT

	H_{2nd}	T_{2nd}	
H_{1st}	0.25	0.25	0.5
T_{1st}	0.25	0.25	0.5
	0.5	0.5	

$$Pr(H_{2nd}|H_{1st}) = \frac{Pr(H_{2st} \cap H_{1st})}{Pr(H_{1st})} = \frac{Pr(\{HH\})}{Pr(\{HH\} \cup \{HT\})} = \frac{0.25}{0.5} = 0.5$$

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 provides us no insight into what will happen second

Independence

- This requires that we define independence as follows:

If \mathcal{A}_i is independent of \mathcal{A}_j , then we have:

$$Pr(\mathcal{A}_i|\mathcal{A}_j) = Pr(\mathcal{A}_i)$$

- This implies the following from the definition of conditional prob.:

$$Pr(\mathcal{A}_i|\mathcal{A}_j) = \frac{Pr(\mathcal{A}_i \cap \mathcal{A}_j)}{Pr(\mathcal{A}_j)} = \frac{Pr(\mathcal{A}_i)Pr(\mathcal{A}_j)}{Pr(\mathcal{A}_j)} = Pr(\mathcal{A}_i)$$

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

$$Pr(\mathcal{A}_i \cap \mathcal{A}_j) = Pr(\mathcal{A}_i)Pr(\mathcal{A}_j)$$

Example of independence

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

	H_{2nd}	T_{2nd}	
H_{1st}	$Pr(H_{1st} \cap H_{2nd})$	$Pr(H_{1st} \cap T_{2nd})$	$Pr(H_{1st})$
T_{1st}	$Pr(T_{1st} \cap H_{2nd})$	$Pr(T_{1st} \cap T_{2nd})$	$Pr(T_{1st})$
	$Pr(H_{2nd})$	$Pr(T_{2nd})$	

	H_{2nd}	T_{2nd}	
H_{1st}	0.25	0.25	0.5
T_{1st}	0.25	0.25	0.5
	0.5	0.5	

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

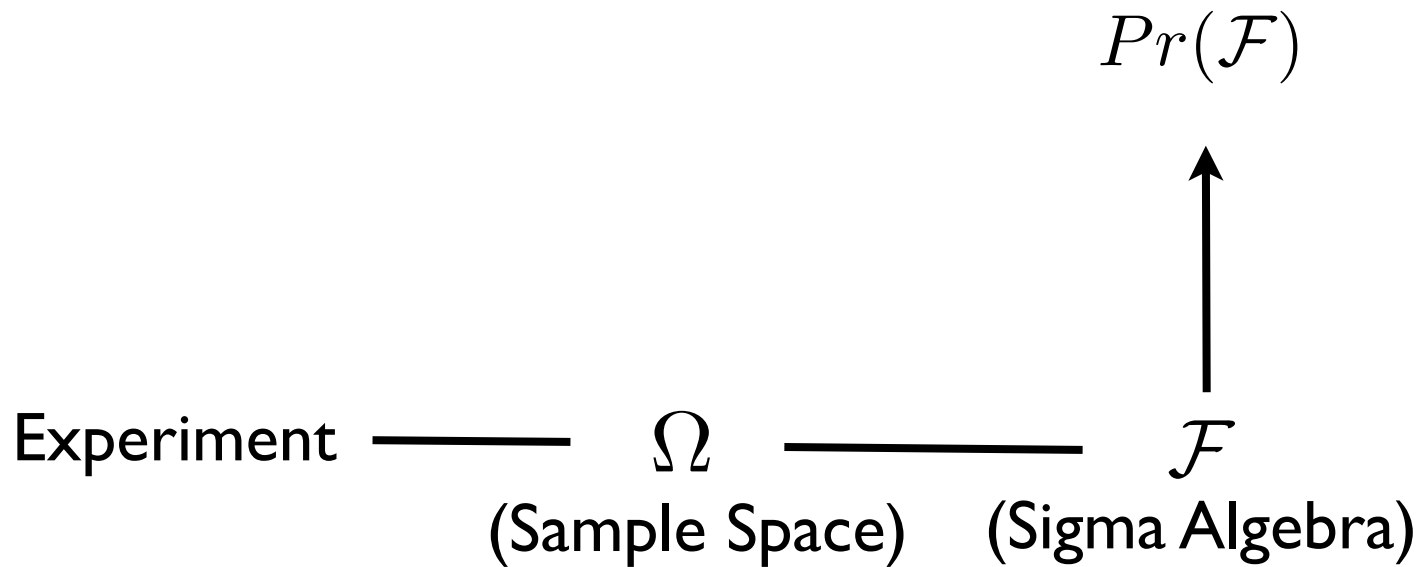
- Consider the sample space of “two coin flips” and the following probability model:

	H_{2nd}	T_{2nd}	
H_{1st}	$Pr(H_{1st} \cap H_{2nd})$	$Pr(H_{1st} \cap T_{2nd})$	$Pr(H_{1st})$
T_{1st}	$Pr(T_{1st} \cap H_{2nd})$	$Pr(T_{1st} \cap T_{2nd})$	$Pr(T_{1st})$
	$Pr(H_{2nd})$	$Pr(T_{2nd})$	

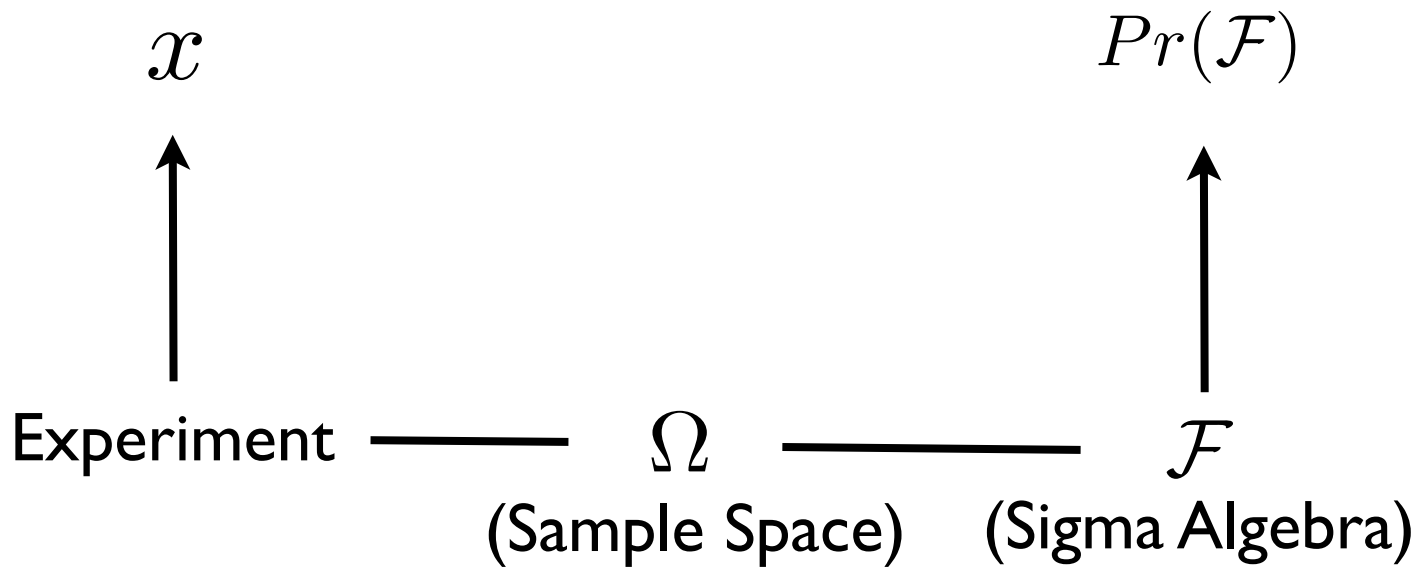
	H_{2nd}	T_{2nd}	
H_{1st}	0.4	0.1	0.5
T_{1st}	0.1	0.4	0.5
	0.5	0.5	

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

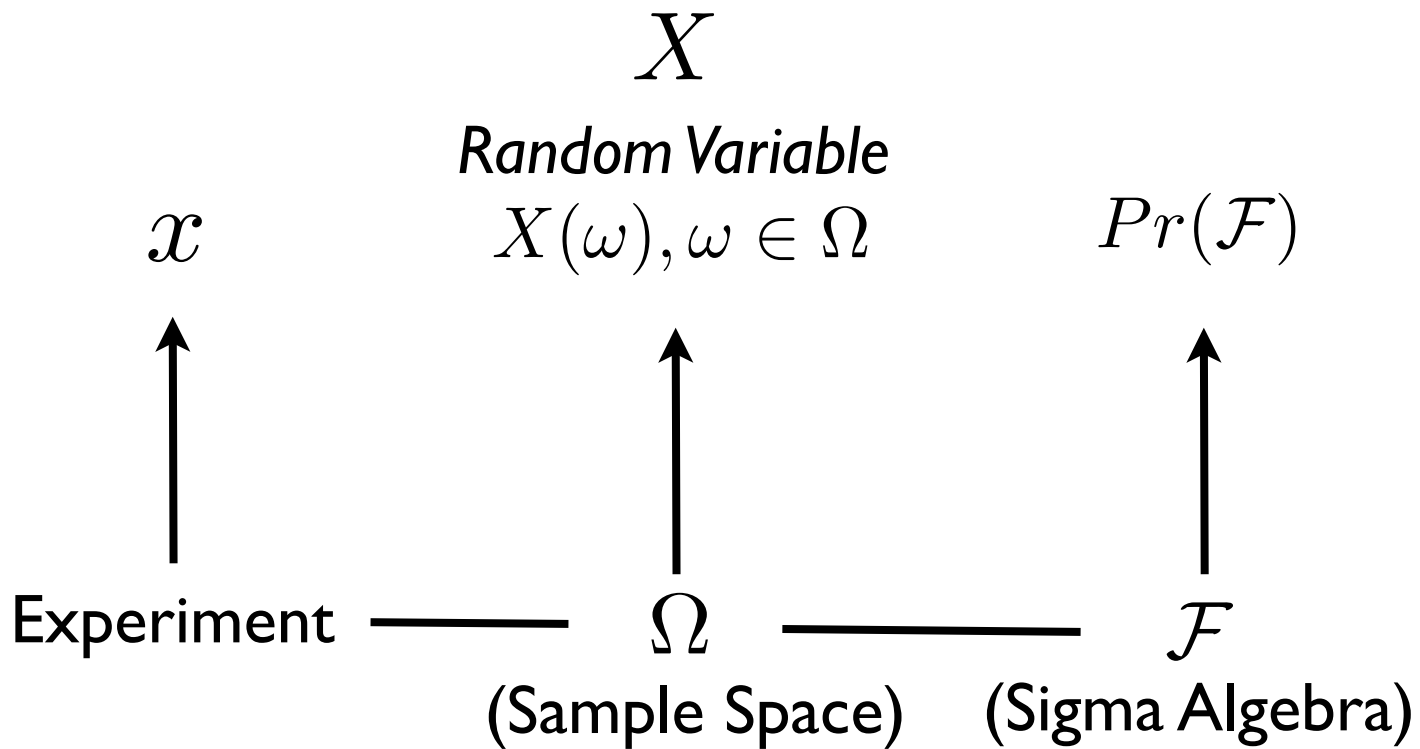
Next Essential Concept: Random Variables



Next Essential Concept: Random Variables



Next Essential Concept: Random Variables



Random variables I

- A probability function / measure takes the Sigma Algebra to the reals and provides a model of the uncertainty in our system / experiment:

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

- When we define a probability function, this is an assumption (!!), i.e. what we believe is an appropriate probabilistic description of our system / experimen
- We would like to have a concept that connects the *actual* outcomes of our experiment to this probability mode
- What's more, we are often in situations where we are interested in using numbers to represent the outcomes, e.g., “Heads” and “Tails” accurately represent the outcomes of a coin flip example but they are not numbers (e.g., we may be interested in “number of heads”)
- In addition, many of the mathematical tools we use in probability and statistics require the outcomes being represented within the reals
- We therefore are often interested in a function of the original sample space that maps this space to the reals
- We will define a *random variable* for this purpose
- In general, the concept of a random variable is a “bridging” concept between the actual experiment and the probability model, this provides a numeric description of sample outcomes that can be defined many ways (i.e. provides great versatility)

Random variables II

- **Random variable** - a real valued function on the sample space:

$$X : \Omega \rightarrow \mathbb{R}$$

- Intuitively:

$$\Omega \longrightarrow \boxed{X(\omega), \omega \in \Omega} \longrightarrow \mathbb{R}$$

- Note that these functions are not constrained by the axioms of probability, e.g. not constrained to be between zero or one (although they must be measurable functions and admit a probability distribution on the random variable!!)
- We generally define them in a manner that captures information that is of interest
- As an example, let's define a random variable for the sample space of the "two coin flip" experiment that maps each sample outcome to the "number of Tails" of the outcome:

$$X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2$$

That's it for today

- Next lecture, we will continue our random variables and introduce expectations, variances, and related!