

Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 3: Conditional Probability and Random Variables

Jason Mezey
Jan 30, 2024 (T) 8:40-9:55

Announcements I

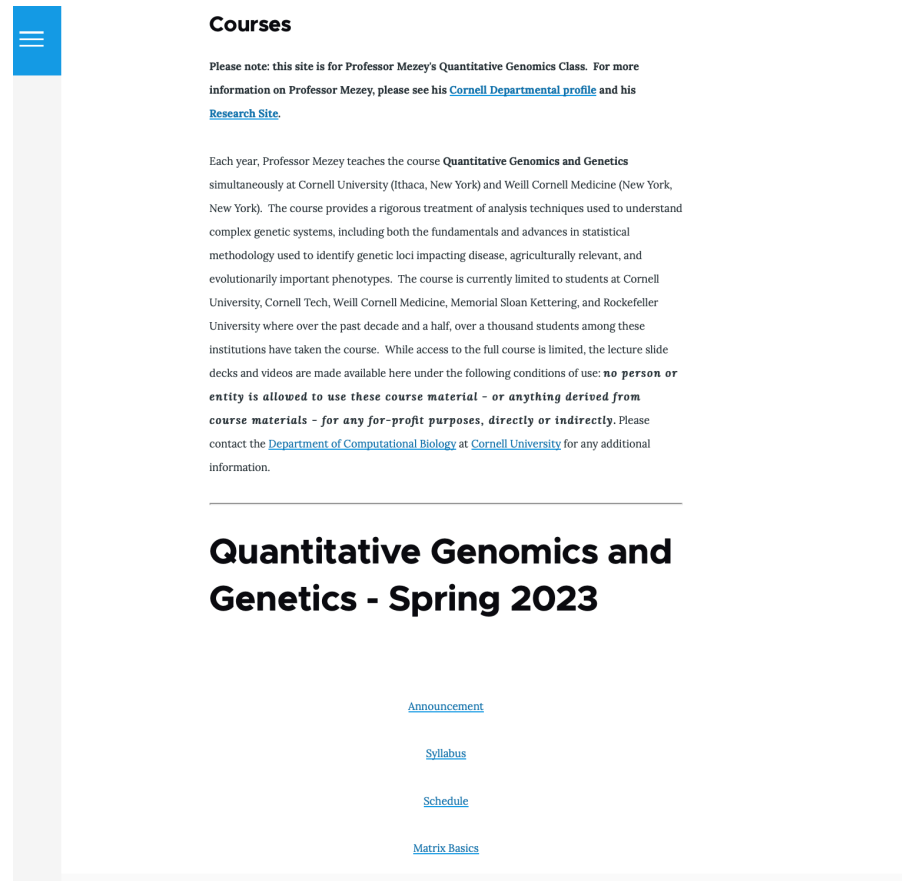
- Make sure you are up on Canvas / have activated your account (!!)
- Note that ALL EMAIL must be sent through Canvas (!!)
- Check out the “Ed Discussion” forum - we will be posting (and I encourage you to post if you have questions / things you’d like clarified etc.!)
- For those in NYC, we DO have a classroom for lecture on Thurs (Feb 1) but we may not have classrooms for the next two weeks (i.e., Feb 6-15) please stay tuned...
- We may put in a zoom option for everyone (again stay tuned = we will announce on canvas / by email)

Announcements II

- **FIRST COMPUTER LAB IS THIS WEEK (Thurs. Feb 1 / Fri. Feb 2)**
- For those IN ITHACA (= Labs with Beulah):
 - Lab 1: 3:35-4:25PM on Thurs. (Mann Library B30A)
 - Lab 2: 9:05-9:55AM on Fri. (Mann Library B30A)
- For those IN NYC (= Labs with Sam!):
 - FRIDAYS (only!) 9-10AM in A-950 Auditorium, 1300 York Ave (9th floor)
 - NO THURS LAB!

Announcements III

- An additional class website: <https://mezeylab.biohpc.cornell.edu>



The screenshot shows a website page with a blue header containing a white hamburger menu icon. Below the header is a light gray vertical bar. The main content area is white and features the following text:

Courses

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](#).

Each year, Professor Mezey teaches the course **Quantitative Genomics and Genetics** simultaneously at Cornell University (Ithaca, New York) and Weill Cornell Medicine (New York, New York). The course provides a rigorous treatment of analysis techniques used to understand complex genetic systems, including both the fundamentals and advances in statistical methodology used to identify genetic loci impacting disease, agriculturally relevant, and evolutionarily important phenotypes. The course is currently limited to students at Cornell University, Cornell Tech, Weill Cornell Medicine, Memorial Sloan Kettering, and Rockefeller University where over the past decade and a half, over a thousand students among these institutions have taken the course. While access to the full course is limited, the lecture slide decks and videos are made available here under the following conditions of use: **no person or entity is allowed to use these course material - or anything derived from course materials - for any for-profit purposes, directly or indirectly.** Please contact the [Department of Computational Biology at Cornell University](#) for any additional information.

Quantitative Genomics and Genetics - Spring 2023

[Announcement](#)

[Syllabus](#)

[Schedule](#)

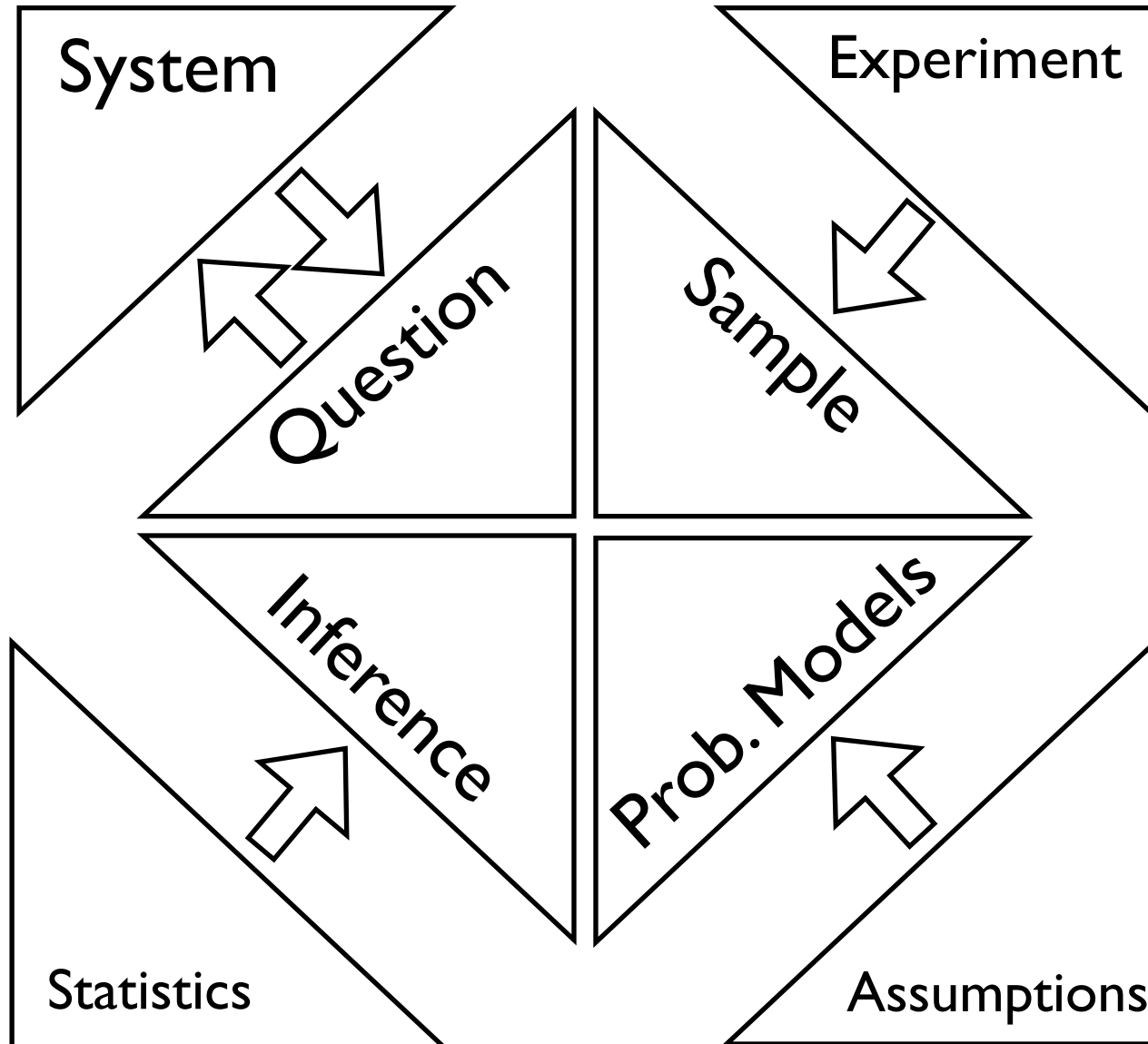
[Matrix Basics](#)

- If you can't see the board work for lecture 2 - check out the lecture 2 from 2023 (=less blurry)

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

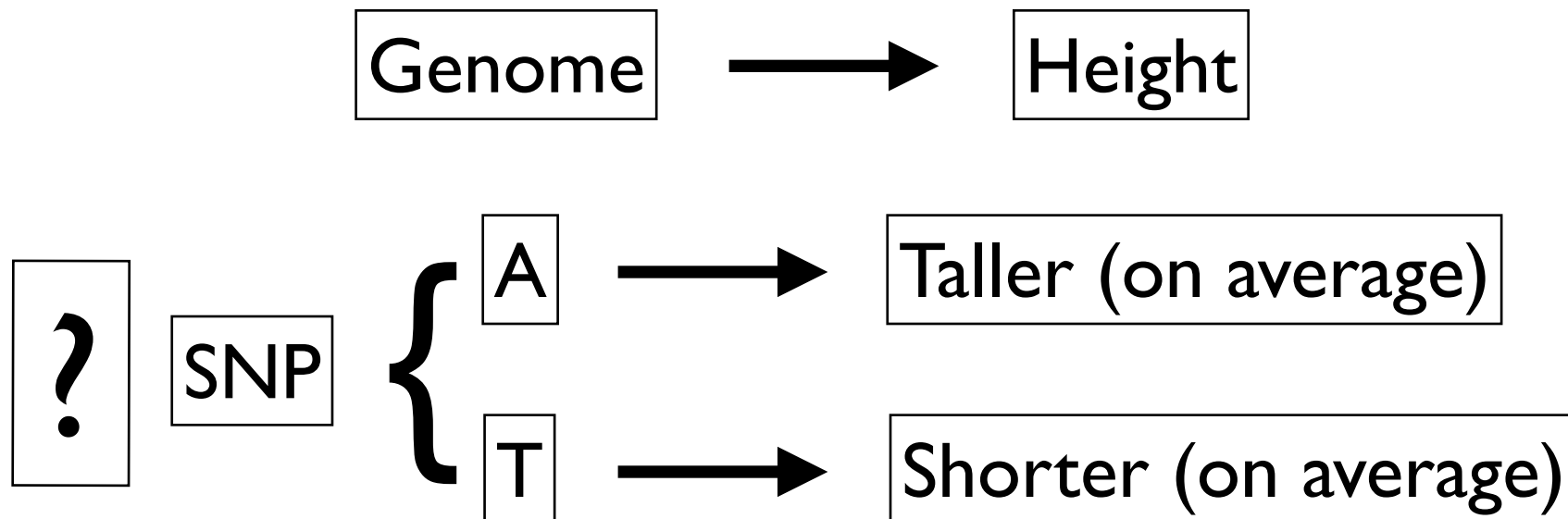
- Last class, we introduced the foundations needed to define / the definition of a probability function!
- Today we will discuss TWO critical concepts: conditional probability AND random variables (!!)

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: 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: Sigma Algebra

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

Review: Probability functions I

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

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

- 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 $\mathcal{A}_1, \mathcal{A}_2, \dots \subset \Omega$, 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 I

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

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

- When we define a probability function, this is an assumption (!!), i.e. what we believe is an appropriate probabilistic description of our system / experiment
- 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:

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

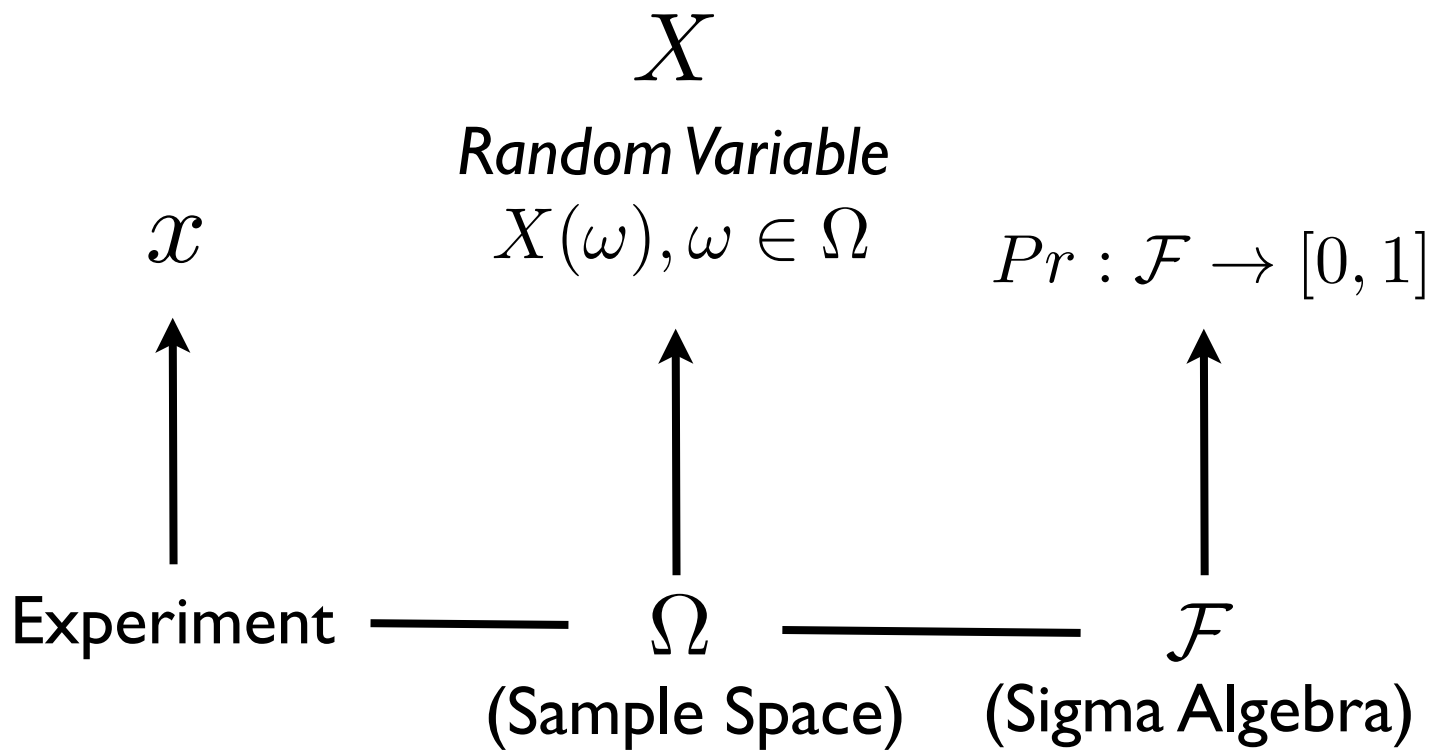
- Intuitively:

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

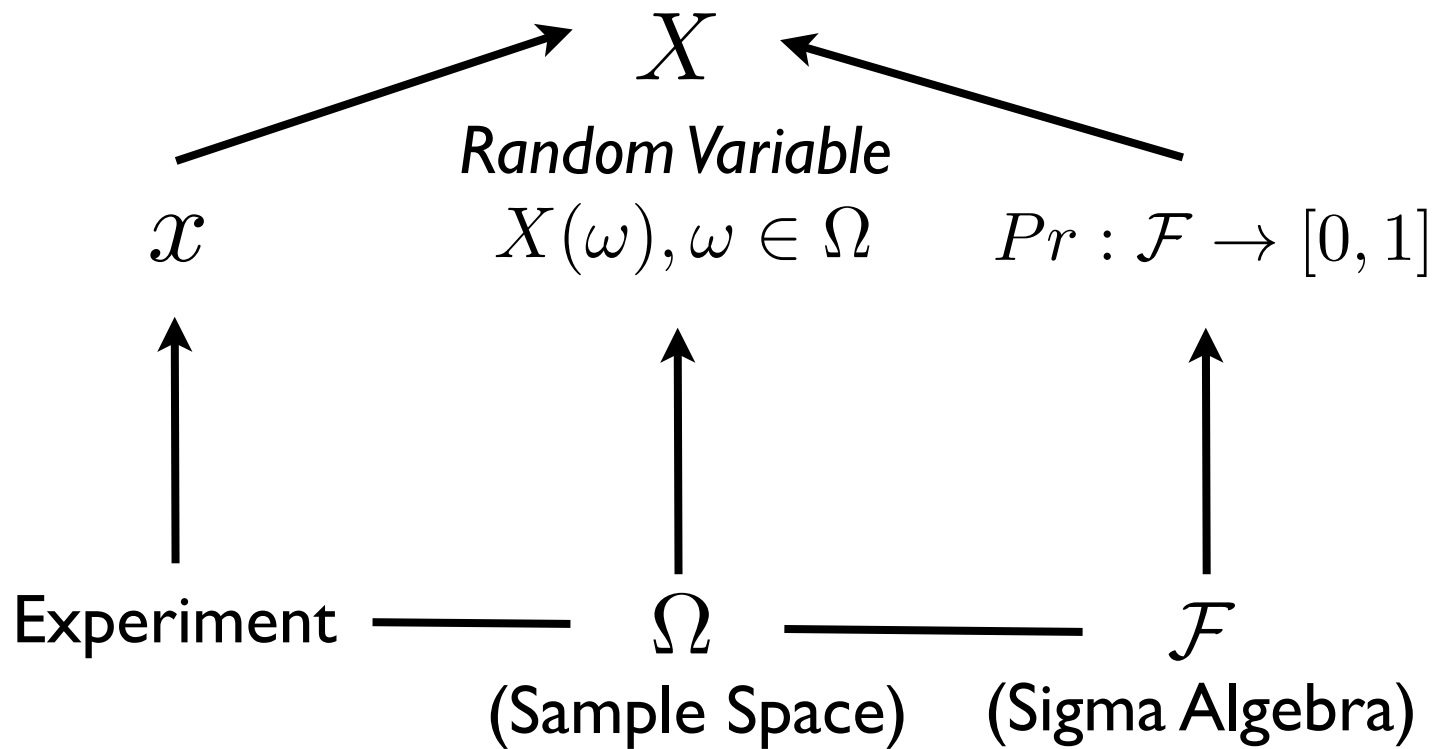
- Note that these functions are not constrained by the axioms of probability, e.g. not constrained to be between zero or one (although they must be measurable functions and admit a probability distribution on the random variable!!)
- We generally define them in a manner that captures information that 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:

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

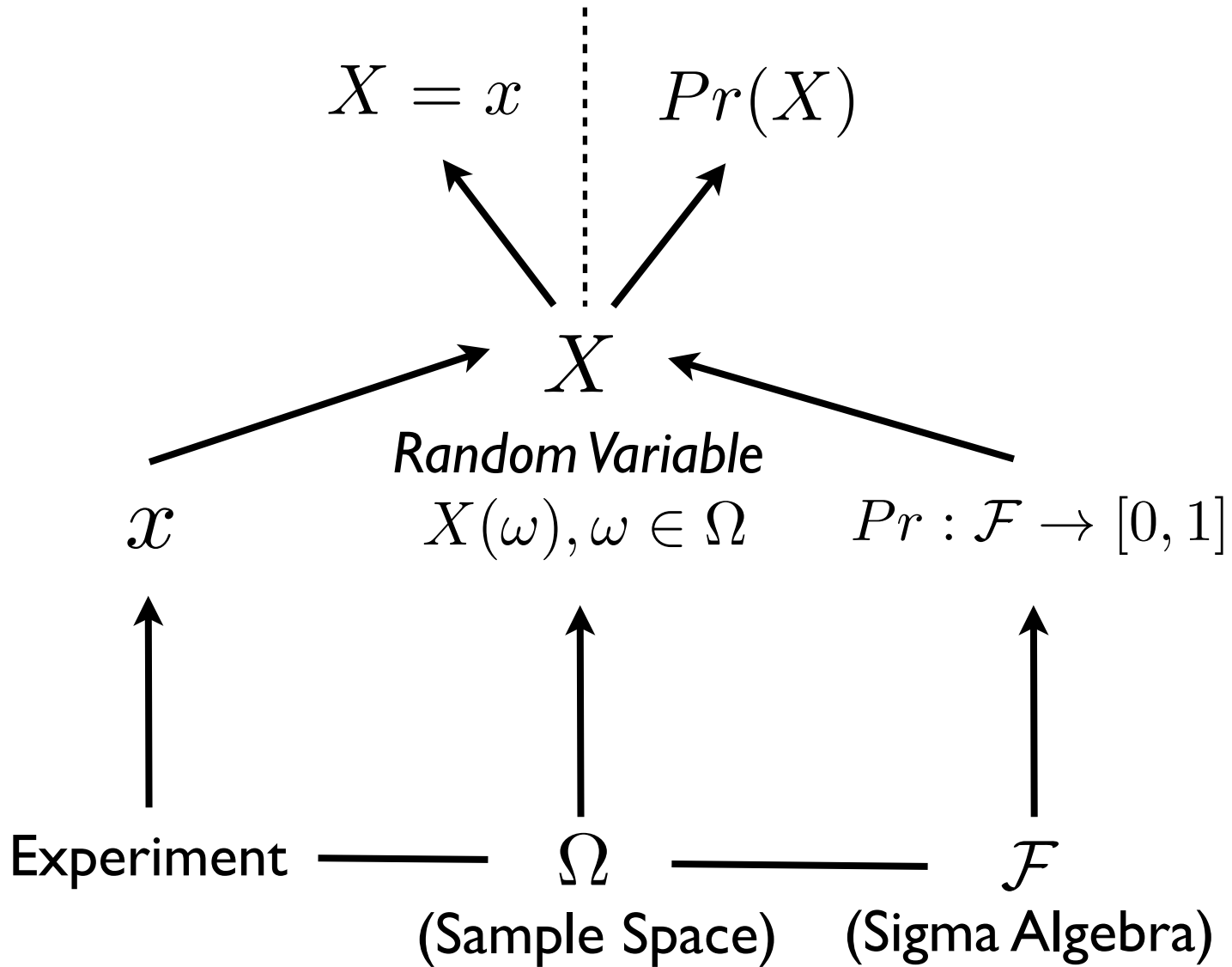
Random Variables



Random Variables



Random Variables



That's it for today

- Next lecture, we will continue our random variables and random vectors!