Quantitative Genomics and Genetics

BIOCB 4830/6830; PBSB.5201.03

Lecture 4: Random Variables and Random Vectors

Jason Mezey

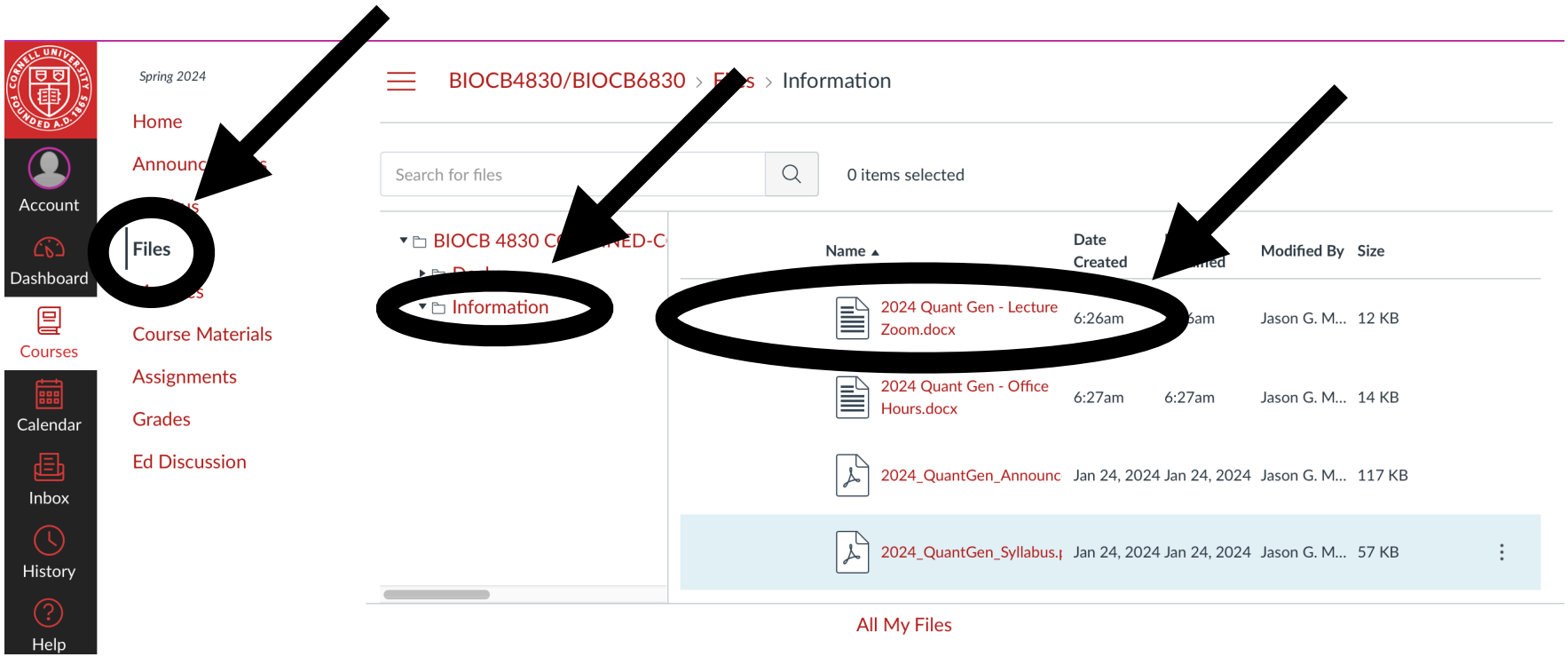
Feb 1, 2024 (Th) 8:40-9:55

Announcements I

- I am now only answering Canvas email (!!)
- PLEASE NOTE (!!): Lectures next week and the following week (Feb 6, Feb 8, Feb 13, Feb 15):
  - We currently do NOT have classrooms in NYC - I will update you by Canvas announcement / email if this changes
  - All NYC students will join by zoom for these lectures (see next)
  - We will have classrooms in Ithaca as per normal
- We are opening a Zoom lecture option:
  - This is for everyone in the class = anyone is we welcome to join any lecture going forward by zoom (i.e., whether in Ithaca / NYC)
  - I would still recommend coming to a classroom for lectures (IF YOU CAN!)

Announcements II

- Where to find the lecture zoom link:



- PLEASE DO NOT SHARE THE ZOOM LINK beyond the class (or we may need to shut it down if we have a problem…)

Announcements III

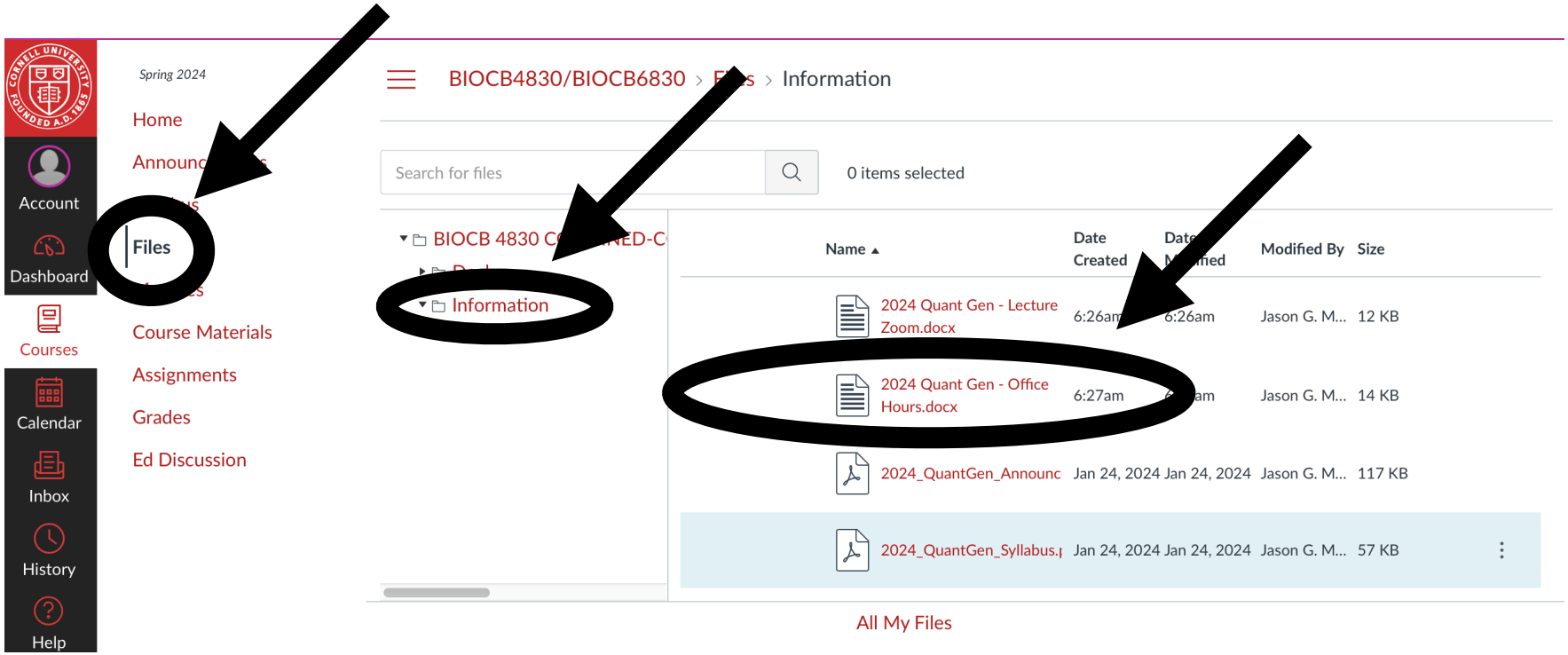
- Homework #1 (!!) will be posted later TODAY (Thurs., Feb 1) on CANVAS (I will Canvas announce / email when it is available):
  - Due 11:59PM, Fri., Feb 9 and MUST BE UPLOADED TO CANVAS (!!)
  - If you upload late (even by a minute…) you will get a penalty (note that no excuses will be accepted = you can always upload early…)
  - Homeworks are "open book" and you may work together but hand in your own work (!!)
  - You may use ChatGPT (or related) BUT you may not want to…
  - Answers must be typed (!!) including all equations - if this is a problem go to computer lab this week (= intro to Latex!)
- Problems are divided into "easy", "medium, and "difficult"
  - You can complete the "easy" and "medium" (make sure you give yourself enough time!)
  - For the "difficult" at least attempt (but note that you can get an "A" in the class even if you do not / cannot complete these problems!)
  - The "difficult" problems are NOT extra credit - they are part of the assignment (please attempt them!) BUT you can still get an "A" in the class even if you don't do them!
- Please feel free to attend office hours for help (!!) see next slide

Announcements IV

- I will hold office hours on WEDNESDAYS every week 12-2PM starting NEXT week (Feb 7)
- Please note: if this day and time turns out to be inconvenient for many, we may change it…
- Any give week… we may change if needed (I will Canvas announce / email any changes)
- We will hold office hours by ZOOM (see next slide)
- We will ALSO (this week ONLY!) have office hours10:30-12:30 on FRIDAY (Feb 1) - I will email the link for this later today…
- I will record office hours (and post them on Canvas)
- You may also set up individual sessions with me by appointment

Announcements V

- Where to find the lecture zoom link:

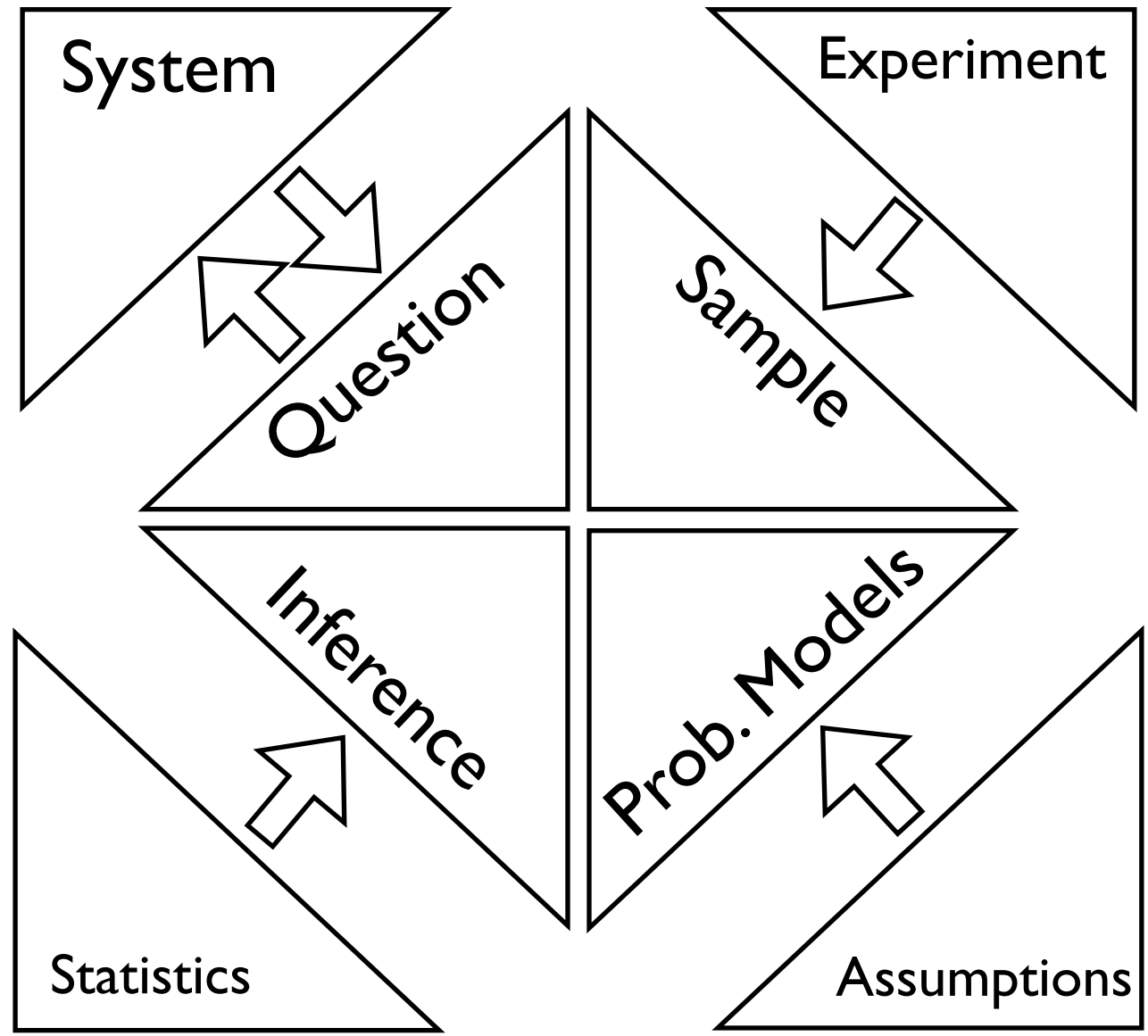


- PLEASE DO NOT SHARE THE ZOOM LINK beyond the class (or we may need to shut it down if we have a problem…)

Summary of lecture 4: Introduction to random variables (and vectors)

- Last class, we introduced conditional probability (and independence!)
- Today we will discuss random variables (and random vectors)

Conceptual Overview



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, ..., 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**. $\mathcal{A} \in \mathcal{F}$ then $\mathcal{A}^c \in \mathcal{F}$, **3**. $\mathcal{A}_1, \mathcal{A}_2, ... \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}$

  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 $\mathcal{A} \subset \Omega, Pr(\mathcal{A}) \geqslant 0$
2. $Pr(\Omega) = 1$
3. For $\mathcal{A}_1, \mathcal{A}_2, ... \subset \Omega$, 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: 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:

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

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

Review: 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	

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

Review: 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.:

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

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

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

Review: 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. $Pr(H_{1st} \cap H_{2nd}) = Pr(H_{1st})Pr(H_{2nd})$

Review: 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:

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

- Intuitively:

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

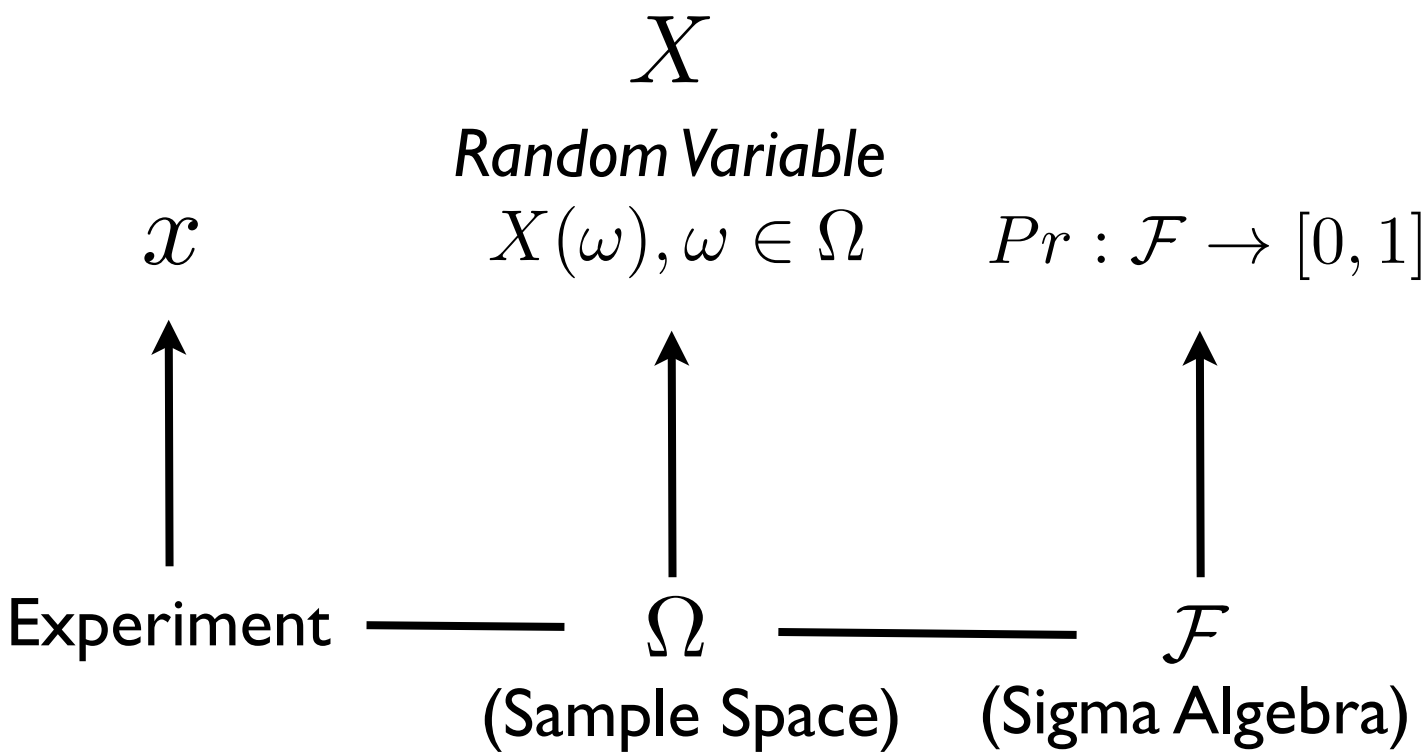
- Note that these functions are not constrained by the axioms of probability, e.g. not constrained to be between zero or one (although they must be measurable functions and admit a probability distribution on the random variable!!)

- We generally define them in a manner that captures information that is of interest

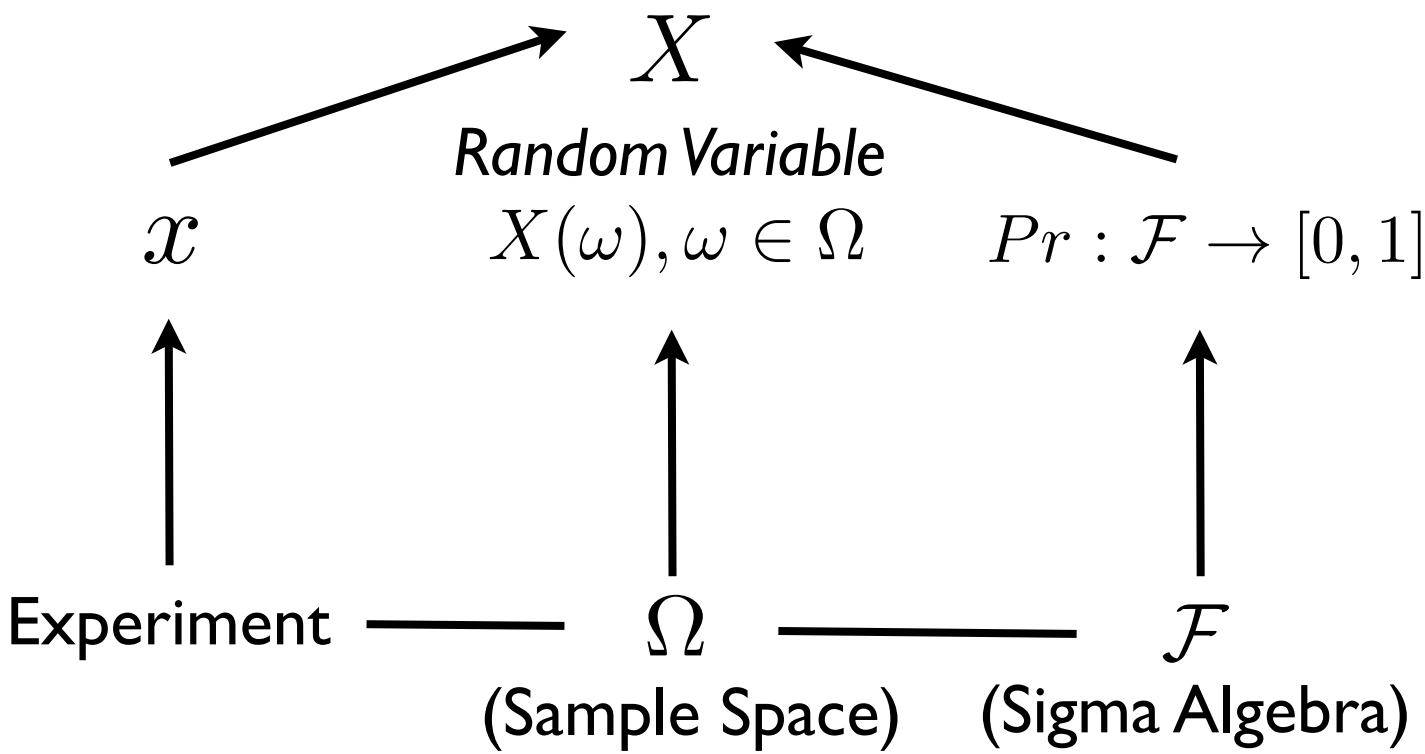
- As an example, let's define a random variable for the sample space of the "two coin flip" experiment that maps each sample outcome to the "number of Tails" of the outcome:

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

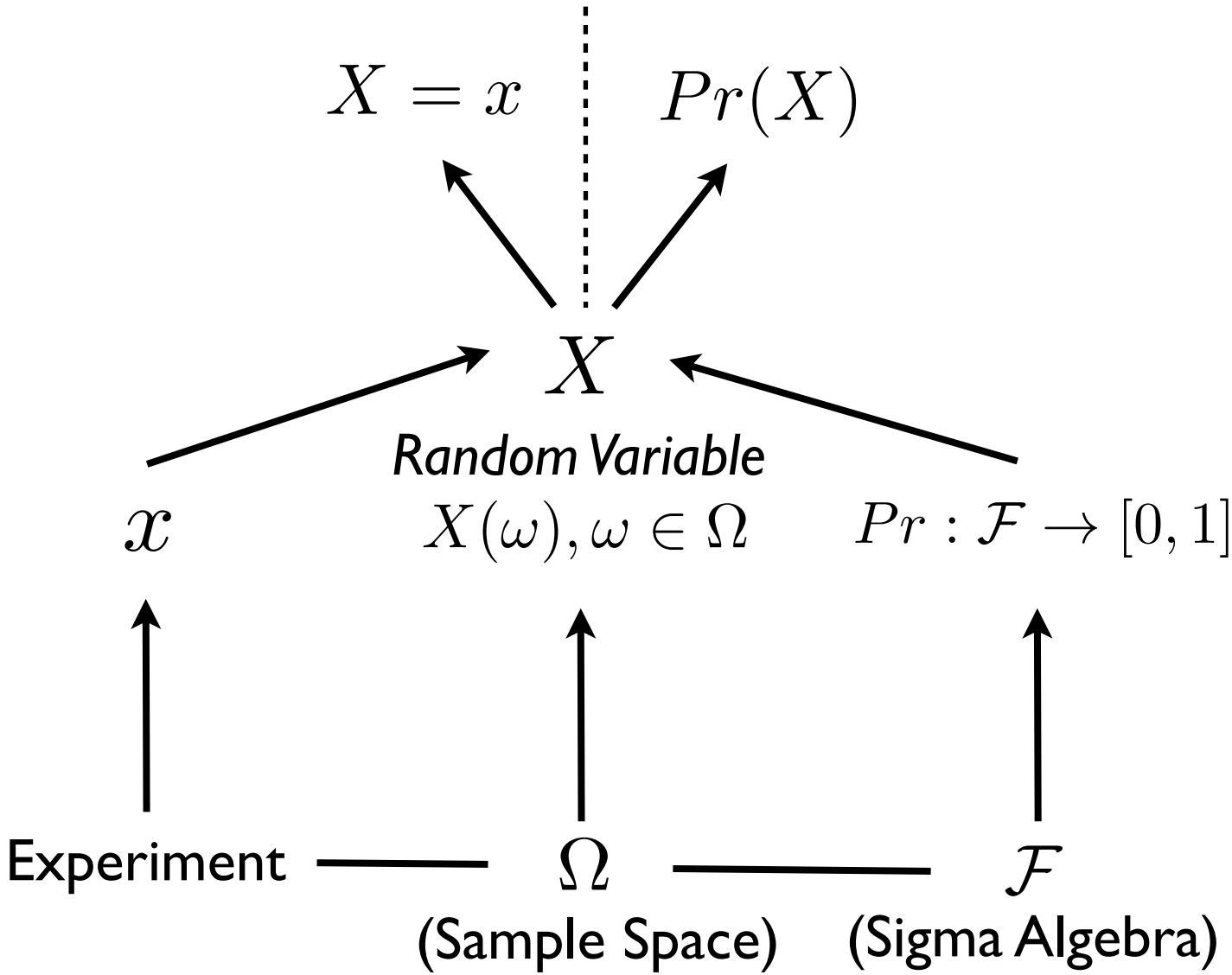
Random Variables



Random Variables



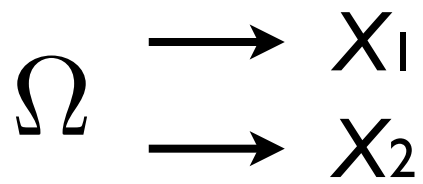
Random Variables



Random variables III

- Why we might want a concept like *X*:
  - This approach allows us to handle non-numeric and numeric sample spaces (sets) in the same framework (e.g., {H, T} is non-numeric but a random variable maps them to something numeric)
  - We often want to define several random variables on the same sample space (e.g., for a "two coin flips" experiment "number of heads" and "number of heads on the first of the two flips"):

$$X_1 : \Omega \to \mathbb{R}$$
$$X_2 : \Omega \to \mathbb{R}$$



  - A random variable provides a bridge between the abstract sample space that is mapped by *X* and the actual outcomes of the experiment that we run (the sample), which produces specific numbers *x*
  - As an example, the notation *X* = *x* bridges the abstract notion of what values could occur *X* and values we actually measured *x*

Random variables IV

- A critical point to note: because we have defined a probability function on the sigma algebra, this "induces" a probability function on the random variable *X*:

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

- In fact, this relationship allows us to "start" our modeling with the random variable and the probability on this random variable (i.e. the Sample Space, Sigma Algebra, and original probability function on random variable are implicit - but remember these foundations are always there!!)

- To bridge probability of an occurrence and what actually occurs in the experiment e often use an "upper" case letter to represent the function and a "lower" case letter to represent the values we actually observe:

$$Pr(X = x)$$

- We will divide our discussion of random variables (which we will abbreviate r.v.) and the induced probability distributions into cases that are discrete (taking individual point values) or continuous (taking on values within an interval of the reals), since these have slightly different properties (but the same foundation is used to define both!!)

Discrete vs Continuous Random Variables

- There are TWO broad categories of random variables: "Discrete" and "Continuous"
  - If the values the random variable can take can be "counted" then the random variable is DISCRETE
  - If the values the random variable cannot be "counted" (e.g., the random variable can take any values on the REALs) then the random variable is CONTINUOUS
- We need to treat the (mathematical) mechanics of these two categories differently…
- *Technical points: (A) discrete random variables may be finite or infinite as long as they take "countable" states (e.g., the naturals are countable while the reals are uncountable), (B) a continuous random variable can only be defined on an uncountable sample space (usually the reals), but a discrete (or mixed) random variable may be defined in a continuous sample space*

Discrete random variables / probability mass functions (pmf)

- If we define a random variable on a discrete sample space, we produce a discrete random variable. For example, our two coin flip / number of Tails example:

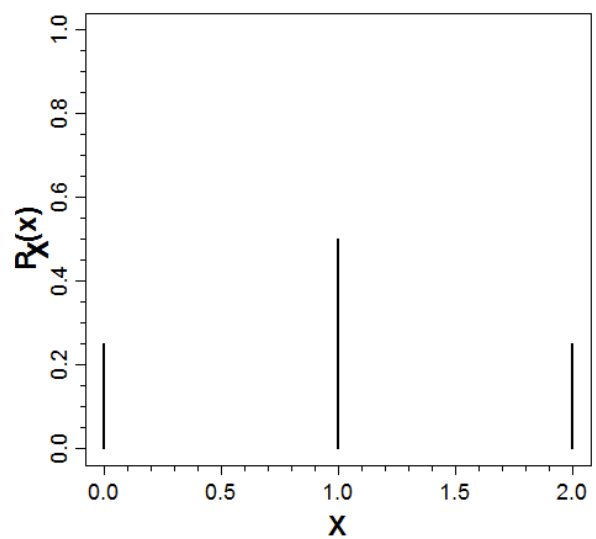
$$X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2$$

- The probability function in this case will induce a probability distribution that we call a **probability mass function** which we will abbreviate as pmf

- For our example, if we consider a fair coin probability model (assumption!) for our two coin flip experiment and define a "number of Tails" r.v., we induce the following pmf:

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

$$P_X(x) = Pr(X = x) = \begin{cases} Pr(X = 0) = 0.25 \\ Pr(X = 1) = 0.5 \\ Pr(X = 2) = 0.25 \end{cases}$$



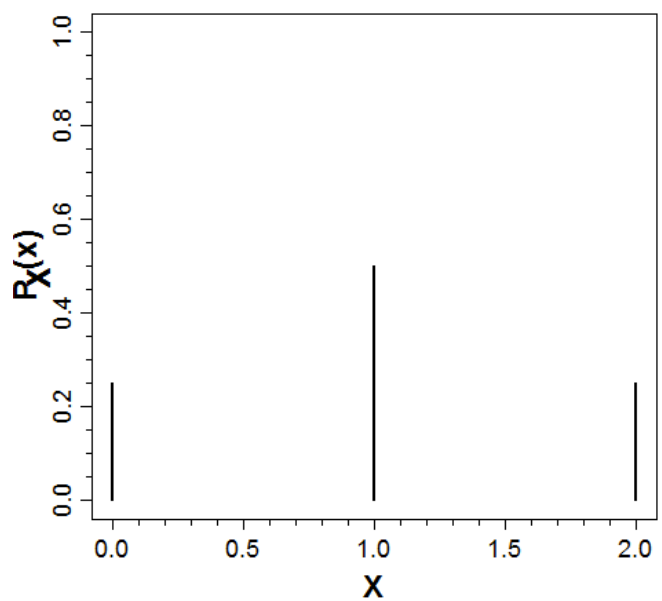
Discrete random variables / cumulative mass functions (cmf)

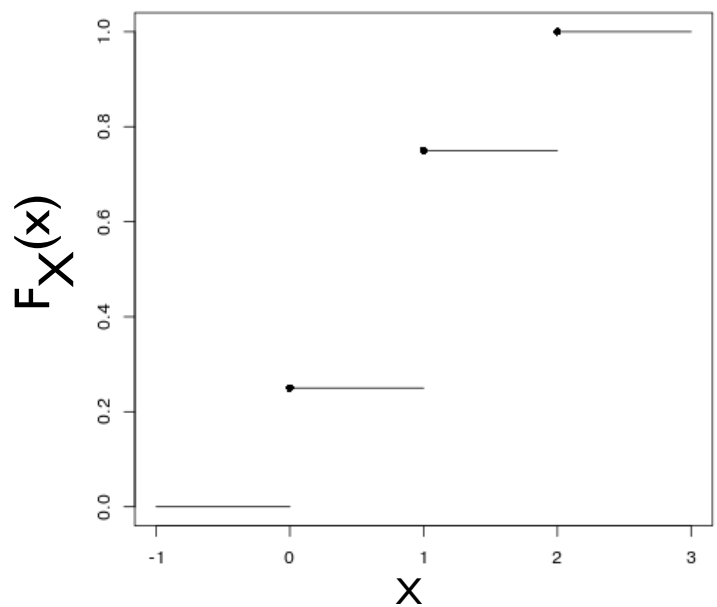
- An alternative (and important!) representation of a discrete probability model is a **cumulative mass function** which we will abbreviate (cmf):

$$F_X(x) = Pr(X \leqslant x)$$

where we define this function for X from $-\infty$ to $+\infty$.

- This definition is not particularly intuitive, so it is often helpful to consider a graph illustration. For example, for our two coin flip / fair coin / number of Tails example:





Continuous random variables / probability density functions (pdf)

- For a continuous sample space, we can define a discrete random variable or a continuous random variable (or a mixture!)
- For continuous random variables, we will define analogous "probability" and "cumulative" functions, although these will have different properties
- For this class, we are considering only one continuous sample space: the reals (or more generally the multidimensional Euclidean space)
- Recall that we will use the reals as a convenient approximation to the true sample space

Mathematical properties of continuous r.v.'s

- For the reals, we define a probability density function (pdf): $f_X(x)$
- The pdf of X, a continuous r.v., does not represent the probability of a specific value of X, rather we can use it to find the probability that a value of X falls in an interval [a,b]:

$$Pr(a \leqslant X \leqslant b) = \int_a^b f_X(x)dx$$

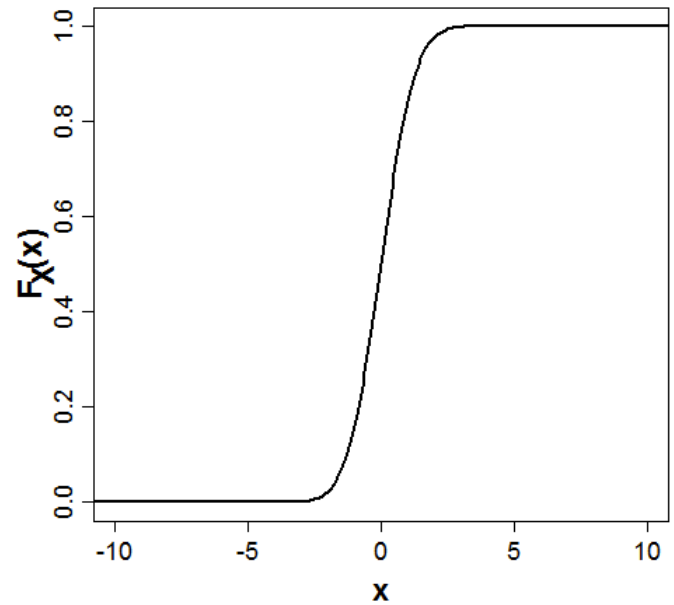
- Related to this concept, for a continuous random variable, the probability of specific value (or point) is zero (why is this!?)
- For a specific continuous distribution the cdf is unique but the pdf is not, since we can assign values to non-measurable sets
- If this is the case, how would we ever get a specific value when performing an experiment!?

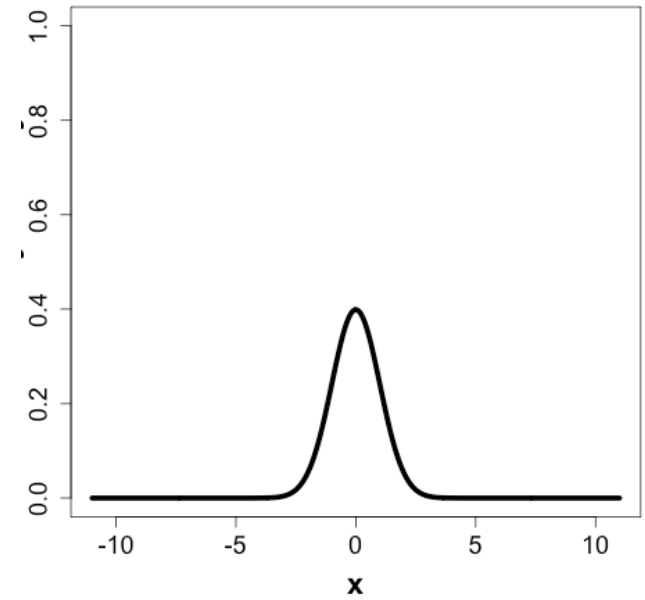
Continuous random variables / cumulative density functions (cdf)

- For continuous random variables, we also have an analog to the cmf, which is the **cumulative density function** abbreviated as cdf:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

- Again, a graph illustration is instructive
- Note the cdf runs from zero to one (why is this?)





That's it for today

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