Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 2: Introduction to probability basics

Jason Mezey

Jan 26, 2023 (Th) 8:05-9:20

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 (!!) - schedule to be posted (i.e., you will have to check every lecture!)
- Zoom option:
  - Remote (to both Ithaca / NYC) students are joining by zoom now (please mute / unmute to ask questions)
  - By next week, we will have a zoom option for everyone (we will discuss)
- Lectures will be recorded:
  - These will be posted along with slides / notes
  - I encourage you to come to class...

Times and Locations II

- There is a REQUIRED computer lab
- **FIRST COMPUTER LAB WILL BE NEXT WEEK (Thurs. Feb 2 / Fri. Feb 3) - more information to come next week!**
- PLEASE NOTE (!!): LAB TIMES ARE DIFFERENT THAN LISTED
- For those IN ITHACA (= Labs Mitch!):
  - Lab 1: 5:30-6:30PM on Thurs. (Weill Hall 226)
  - Lab 2: 8-9AM on Fri. (Weill Hall 226)
- For those IN NYC (= Labs taught by Sam!):
  - Lab 1: 4-5PM on Thurs. (In WCMC1300 Classroom; G [B215], H [B217])
  - Lab 2: 9-10AM on Fri. (By zoom - please stay tuned for invite)
- You may skip the first 2 labs without penalty

Times and Locations IV

- I (Jason) will hold office hours for both campuses by zoom
- No office hours this week OR next week
- My first office hours will be Feb 6 (Mon)
- You may also set up individual sessions with me by appointment

Class Resources: Piazza

- MAKE SURE YOU SIGN UP ON PIAZZA whether you officially register or not = all communication for the course (!!)
- Class: https://piazza.com/cornell/spring2023/btry4830btry6830
- If you received the Piazza test message / email last night - you should be good to go
- If you are having an issue getting on Piazza, please see me after class / email me directly at **jgm45@cornell.edu** and I will get you on
- EVERYBODY PLEASE GET REGISTERED ASAP (!!)

Class Resources: website and CMS

- The class website: https://mezeylab.biohpc.cornell.edu
- This has not yet been updated (we are still working on it…)
- Assignments and computer labs (!!) will be posted on Cornell CMS (as BTRY 4830)
- This is not yet setup… - please DO NOT TRY to register yet (and stay tuned for information on how to register)

Summary of lecture 2: Introduction to probability basics

- In this class, we will be concerned with the most basic problem of quantitative genomics: how to identify genotypes where differences among individual genomes produce differences in individual phenotypes (i.e. genetic association studies)
- Today we will discuss the rigorous conceptual set-up of probability and essential math concepts

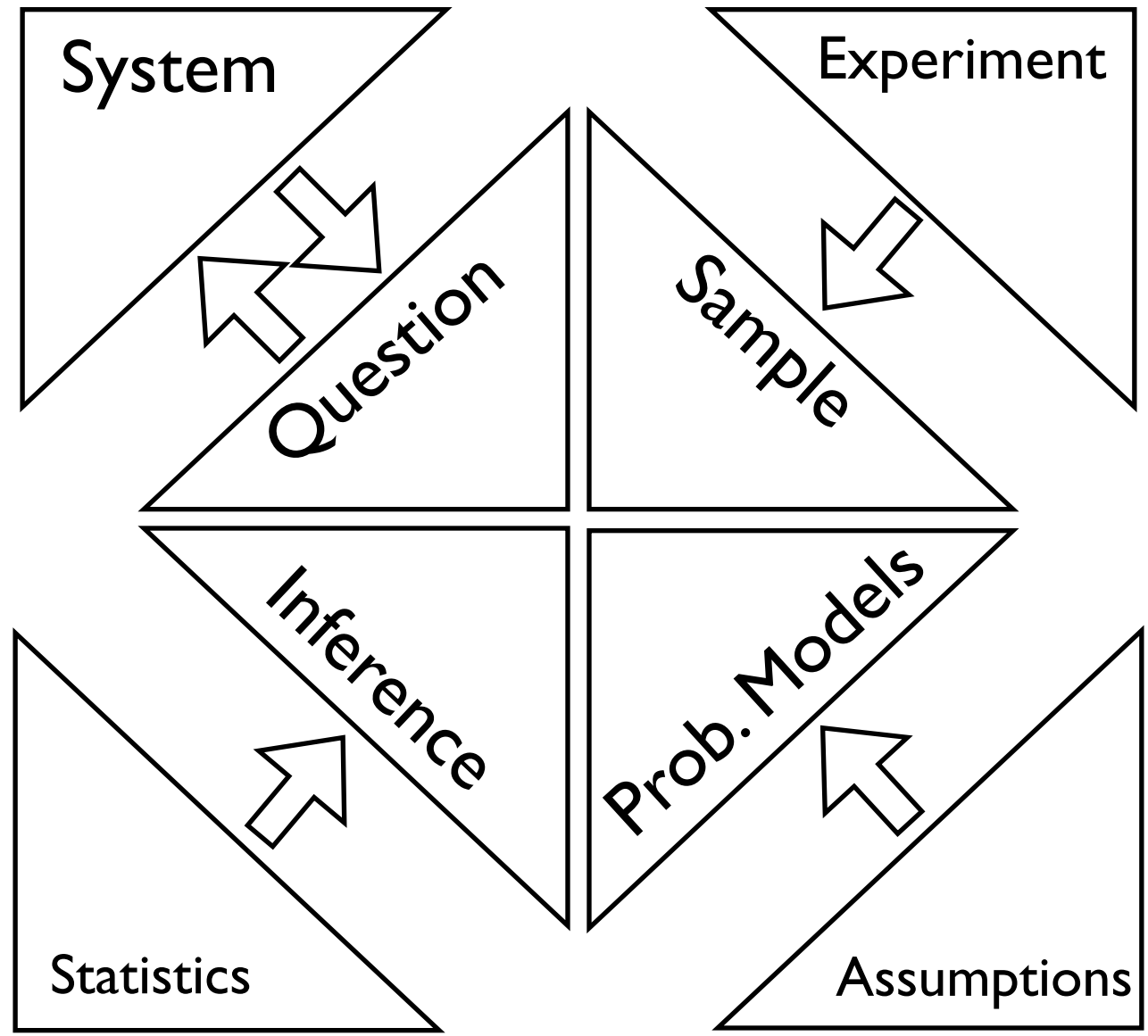
Definitions: Probability / Statistics

- **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

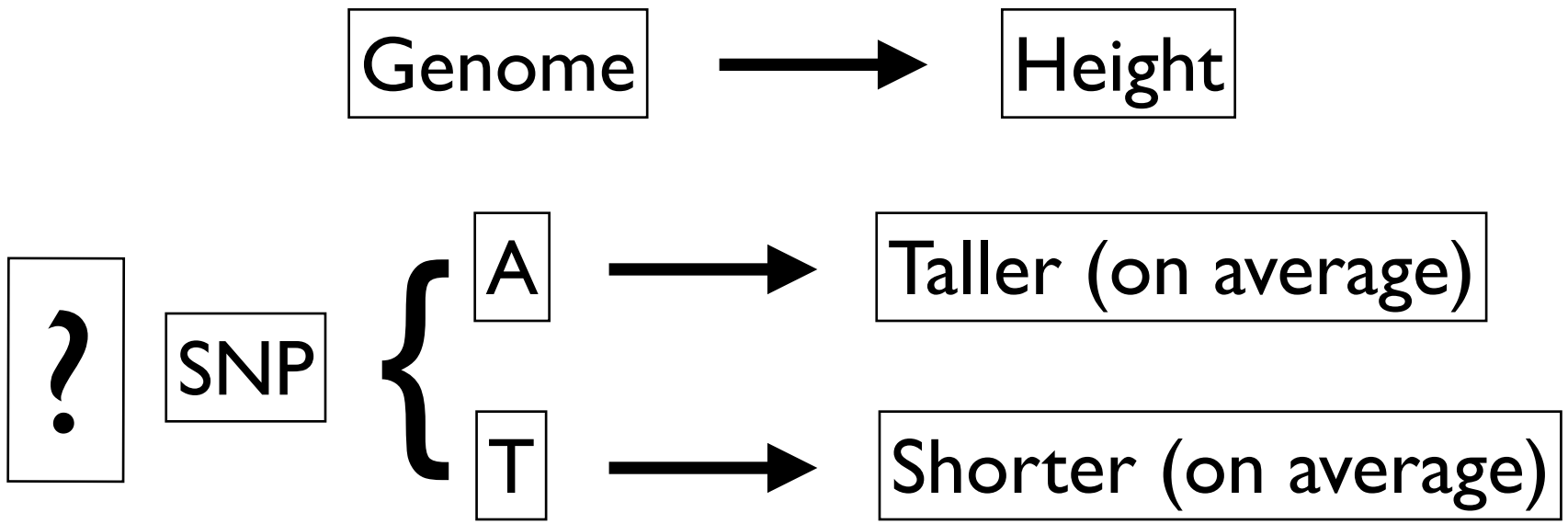
Conceptual Overview



Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Example: Genetic contribution to height



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 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, …

Sets / Set Operations / 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.

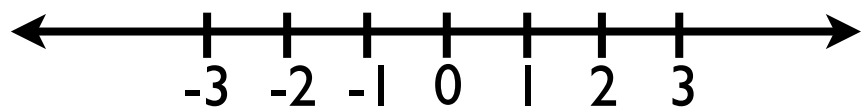
Some Special Sets

- The following sets have properties that align with our intuitive conception about how we represent and use groups
  - The **Natural Numbers** and the **Integers:**

$$\mathbb{N} = \{1, 2, 3, ...\}$$

$$\mathbb{Z} = \{... - 3, -2, -1, 0, 1, 2, 3, ...\}$$

  - The **Reals:**

$\mathbb{R} =$ 

- Note that these sets are infinite (although they represent two different "sizes" of infinite: countable and uncountable), where we often make use of the following symbols in both cases:

$$-\infty > x > \infty$$

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)

Functions

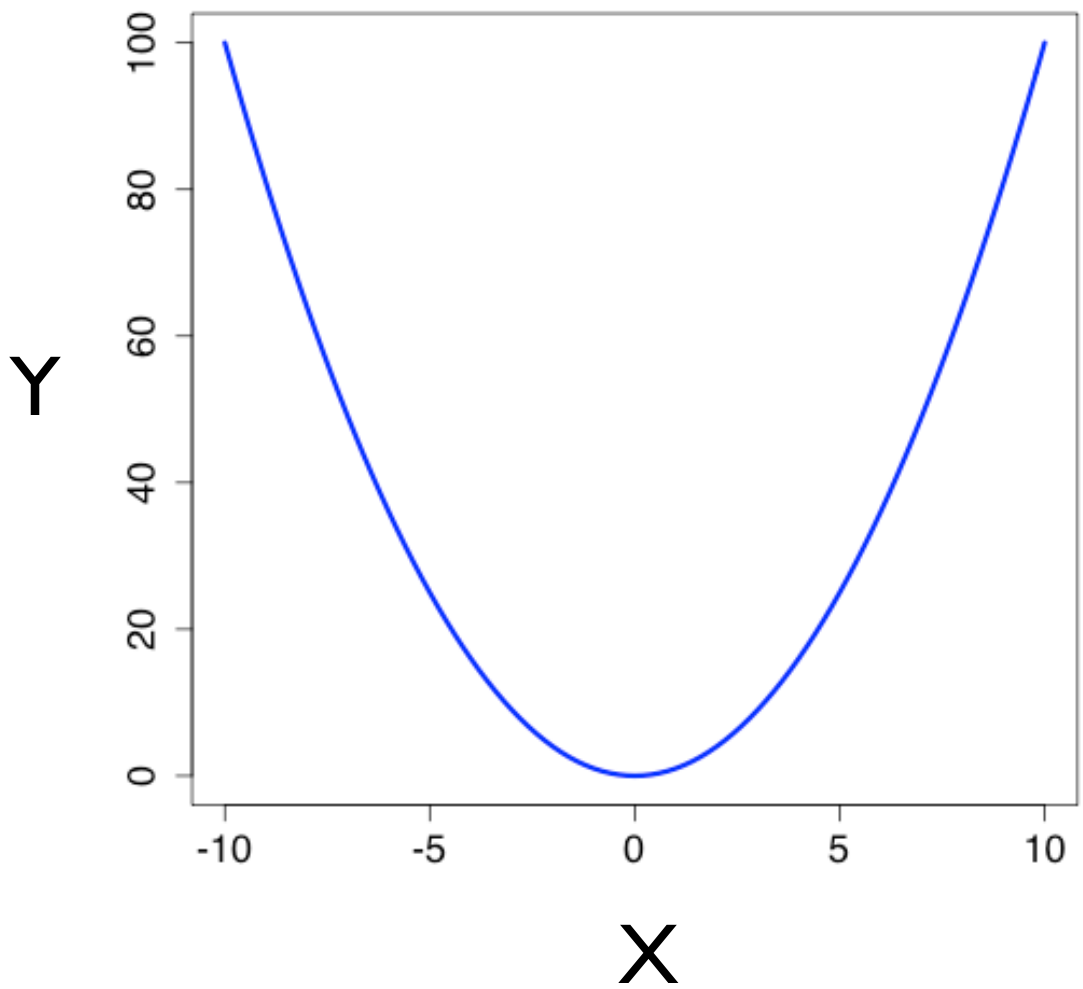
- Now that we have formalized the concept of a sample space, we need to define what "probability" means
- To do this, we need the concept of a mathematical function

- **Function** (formally) - a binary relation between every member of a domain to exactly one member of the codomain

- **Function** (informally) - **?**

Example of a function

$$Y = X^2$$



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 sample space but to every EVENT
- To accomplish this, we will need the concept of a **Sigma Algebra** (or **Sigma Field**)
- What's more, we need to make sure the function that we use to assign these numbers adheres to a specific set of "rules" (axioms)

Sample Spaces / 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…)

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:

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

- 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 $\mathcal{A}_1, \mathcal{A}_2, ... \in \Omega$, 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*

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:

$$Pr(\emptyset) = 0$$

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

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

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

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

- 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:

$$Pr(\emptyset) = 0$$

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

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

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

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

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

$$\not{Pr}(\emptyset) = 0$$

$$\not{Pr}(\{HH\}) = 0.25, \not{Pr}(\{HT\}) = 0.25, \not{Pr}(\{TH\}) = 0.25, \not{Pr}(\{TT\}) = 0.25$$

$$\not{Pr}(\{HH, HT\}) = 0.5, \not{Pr}(\{HH, TH\}) = 0.5, \not{Pr}(\{HH, TT\}) = 1.0,$$

$$\not{Pr}(\{HT, TH\}) = 0, \not{Pr}(\{HT, TT\}) = 0.5, \not{Pr}(\{TH, TT\}) = 0.5,$$

$$\not{Pr}(\{HH, HT, TH\}) = 0.75, \quad \text{etc.} \quad \not{Pr}(\{HH, HT, TH, TT\}) = 1.0$$

That's it for today

- Next lecture, we will continue our discussion of probability by introducing concept of conditional probability and random variables!