Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 7: Inference and Samples

Jason Mezey

Feb 13, 2024 (T) 8:40-9:55

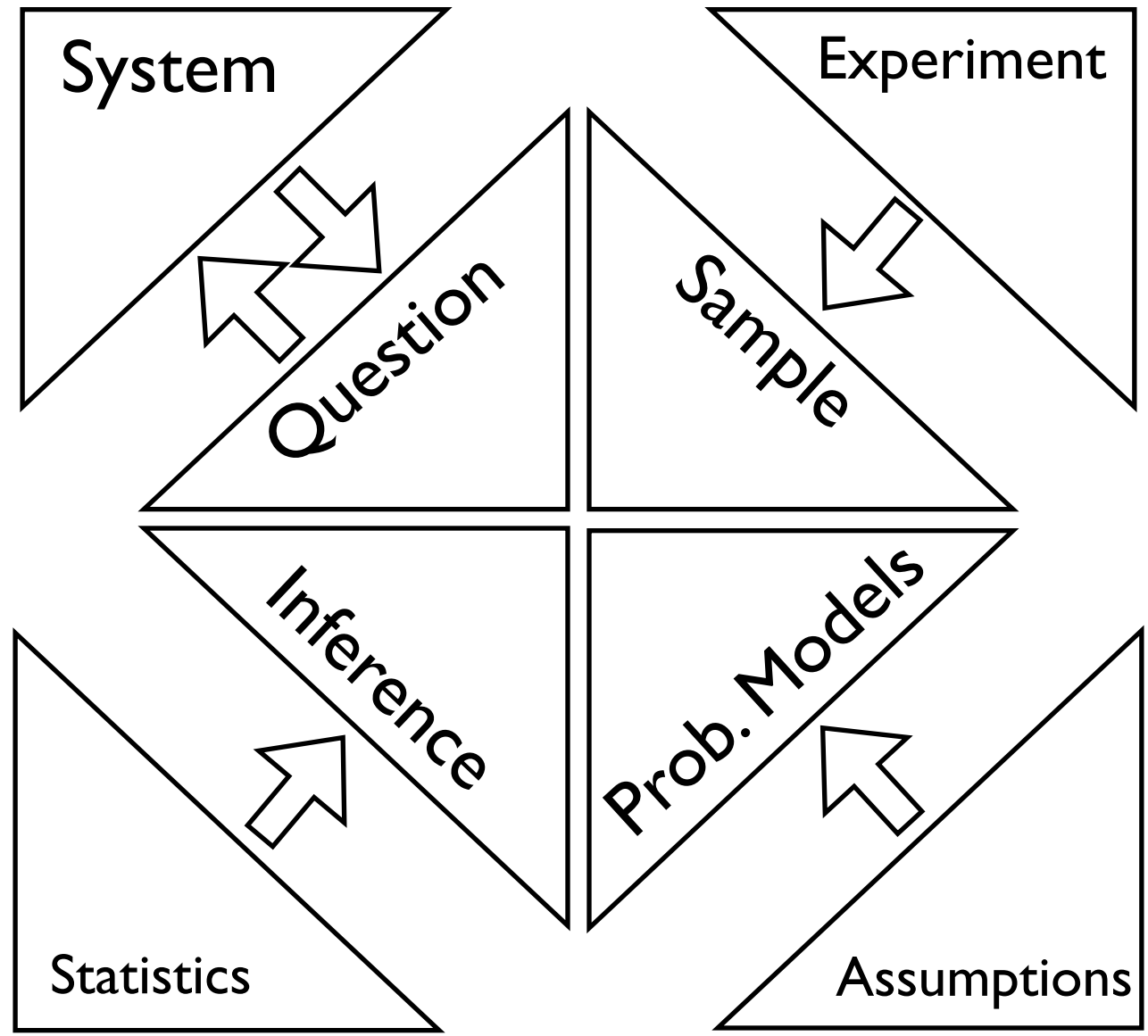
Announcements

- For Weill (NYC) students - we will have a lecture classroom for Thurs (Feb 15): A-250 (1300 York Ave, 2nd floor)
- 2nd homework will be available later today (Feb 13) and will be due Feb 23) by 11:59PM (!!)

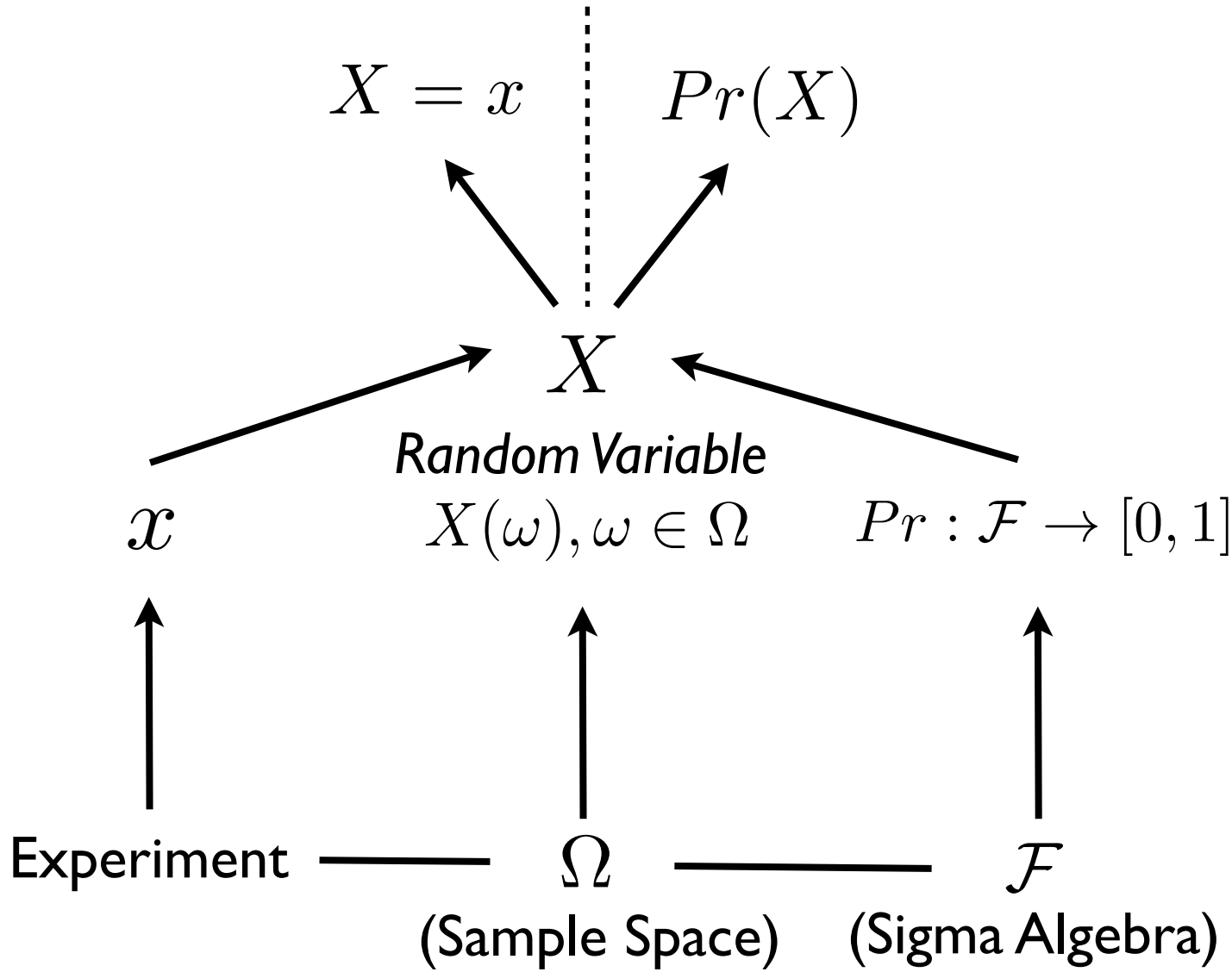
Summary of lecture 7: Introduction to samples (and statistics!)

- Last class, we completed our discussion of expectations and variances of random variables and vectors!
- We will also began our discussion of probability models
- Today we will complete our discussion of probability models and introduce samples

Conceptual Overview



Review: Random Variables



Review: Random vectors

- We are often in situations where we are interested in defining more than one r.v. on the same sample space
- When we do this, we define a **random vector**
- Note that a vector, in its simplest form, may be considered a set of numbers (e.g. [1.2, 2.0, 3.3] is a vector with three elements)
- Also note that vectors (when a vector space is defined) ARE NOT REALLY NUMBERS although we can define operations for them (e.g. addition, "multiplication"), which we will use later in this course
- Beyond keeping track of multiple r.v.'s, a *random vector* works just like a r.v., i.e. a probability function induces a probability function on the random vector and we may consider discrete or continuous (or mixed!) random vectors
- Note that we can define several r.v.'s on the same sample space (= a random vector), but this will result in one probability distribution function (why!?)

Review: Probability models I

- We have defined Pr(*X*), a probability model (=probability function!) on a random variable, which technically we produce by defining Pr function on the sigma algebra and the X (random variable function) on the sample space
- So far, we have generally considered such probability models / functions without defining them explicitly (except for a illustrative few examples)
- To define an explicit model for a given system / experiment we are going to assume that there is a "true" probability model, that is a consequence of the experiment that produces sample outcomes
- We place "true" in quotes since the defining a single true probability model for a given case could only really be accomplished if we knew every single detail about the system and experiment (would a probability model be useful in this case?)
- In practice, we therefore assume that the true probability distribution is within a restricted family of probability distributions, where we are satisfied if the true probability distribution in the family describes the results of our experiment pretty well / seems reasonable given our assumptions

Review: Probability models II

- In short, we therefore start a statistical investigation *assuming* that there is a single true probability model that correctly describes the possible experiment outcomes given the uncertainty in our system
- In general, the starting point of a statistical investigation is to make *assumptions* about the form of this probability model
- More specifically, a convenient assumption is to assume our true probability model is specific model in a family of distributions that can be described with a compact equation
- This is often done by defining equations indexed by *parameters*

Review: Probability models III

- **Parameter** - a constant(s) θ which indexes a probability model belonging to a family of models Θ such that $\theta \in \Theta$
- Each value of the parameter (or combination of values if there is more than on parameter) defines a different probability model: Pr(*X*)
- We assume one such parameter value(s) is the true model
- The advantage of this approach is this has reduced the problem of using results of experiments to answer a broad question to make an educated guess at the value of the parameter(s)
- Remember that the foundation of such an approach is still an assumption about the properties of the the experiment, and the system of interest (!!!)

Discrete parameterized examples

- Consider the probability model for the one coin flip experiment / number of tails.
- This is the Bernoulli distribution with parameter θ = *p* (what does *p* represent!?) where $\Theta = [0, 1]$
- We can write this *X* ~ Bern(*p*) and this family of probability models has the following form:

$$Pr(X = x|p) = P_X(x|p) = p^x(1-p)^{1-x}$$

- For the experiment of *n* coin flips / number of tails, *one possible* family Binomial distribution *X* ~ Bin(*n*, *p*):

$$Pr(X = x|n, p) = P_X(x|n, p) = \binom{n}{x} p^x(1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

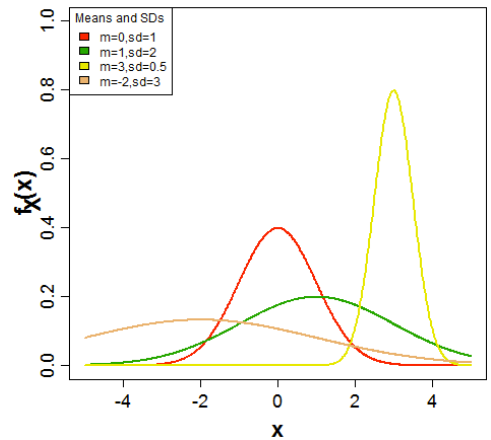
$$n! = n * (n-1) * (n-2) * ... * 1$$

- There are many other discrete examples: hypergeometric, Poisson, etc.

Continuous parameterized examples

- Consider the measure heights experiment (reals as approximation to the sample space) / identity random variable
- For this example we can use the family of normal distributions that are parameterized by $\theta = [\mu, \sigma^2]$ (what do these parameters represent!?) with the following possible values: $\Theta_\mu = (-\infty, \infty)$, $\Theta_{\sigma^2} = [0, \infty)$
- We often write this as $X \sim N(\mu, \sigma^2)$ and the equation has the following form:

$$Pr(X = x|\mu, \sigma^2) = f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

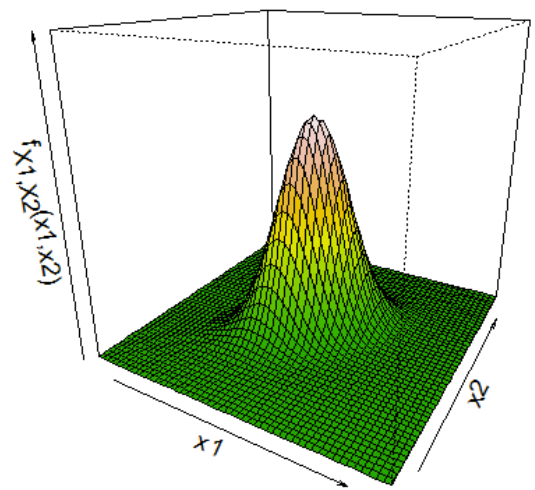


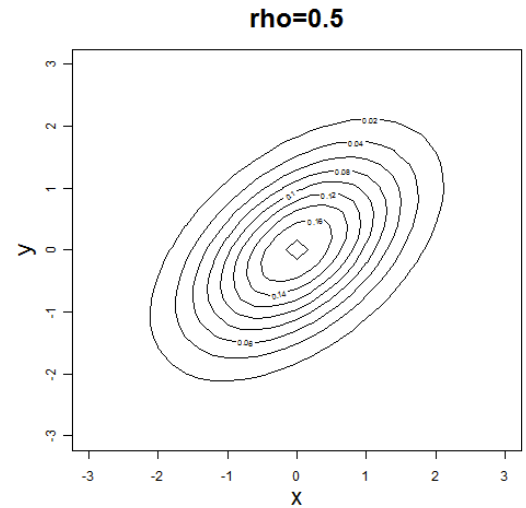
- There are many other continuous examples: uniform, exponential, etc.

Example for random vectors

- Since random vectors are the generalization of r.v.'s, we similarly can define parameterized probability models for random vectors
- As an example, if we consider an experiment where we measure "height" and "IQ" and we take the 2-D reals as the approximate sample space (vector identity function), we could assume the bivariate normal family of probability models:

$$f_{\mathbf{X}}(\mathbf{x}|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_1)^2}{2\sigma_2^2}\right)\right]$$





Introduction to inference I

- Recall that our eventual goal is to use an experiment to provide an answer to a question (about a system)
- So far, we have set up the mathematical foundation that we need to accomplish this goal in a probability / statistics setting
- Specifically, we have defined formal components of our framework and made assumptions that have reduced the scope of the problem
- With these components and assumptions in place, we are almost ready to perform *inference*, which will accomplish our goal

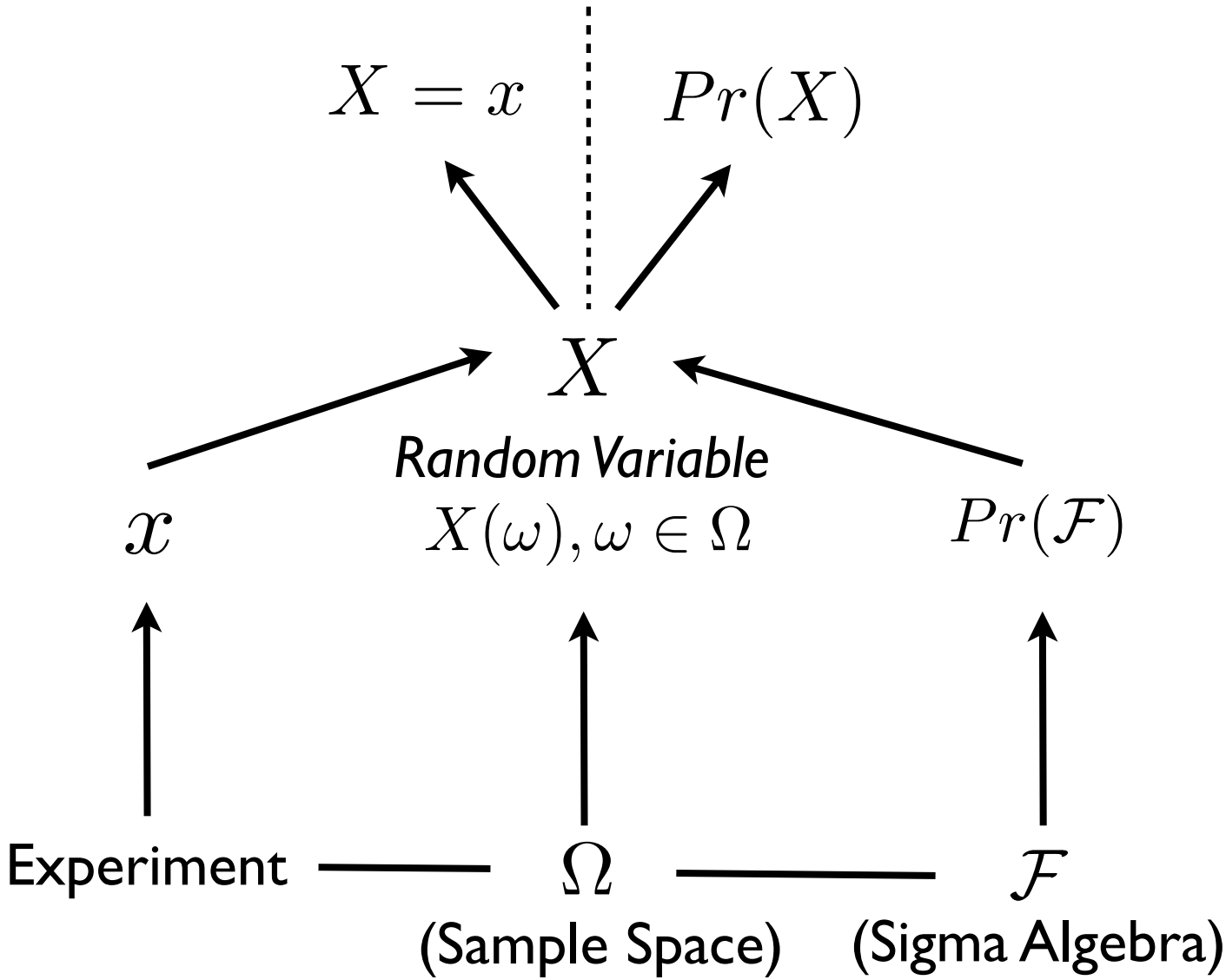
Introduction to inference II

- For our system and experiment, we are going to assume there is a single "correct" *probability function* (which in turn defines the probability of our possible random variable outcomes)
- For the purposes of inference, we often assume a *parameterized* family of *probability models* determine the possible cases that contain the "true" model that describes the result of the experiment
- This reduces the problem of inference to identifying the "single" value(s) of the parameter that describes this true model
- Inference (informally) is the process of using the output of an experiment to answer the question
- Our eventual goal is to use a **sample** (generated by experiment trials) to provide an answer to a question (about a system)

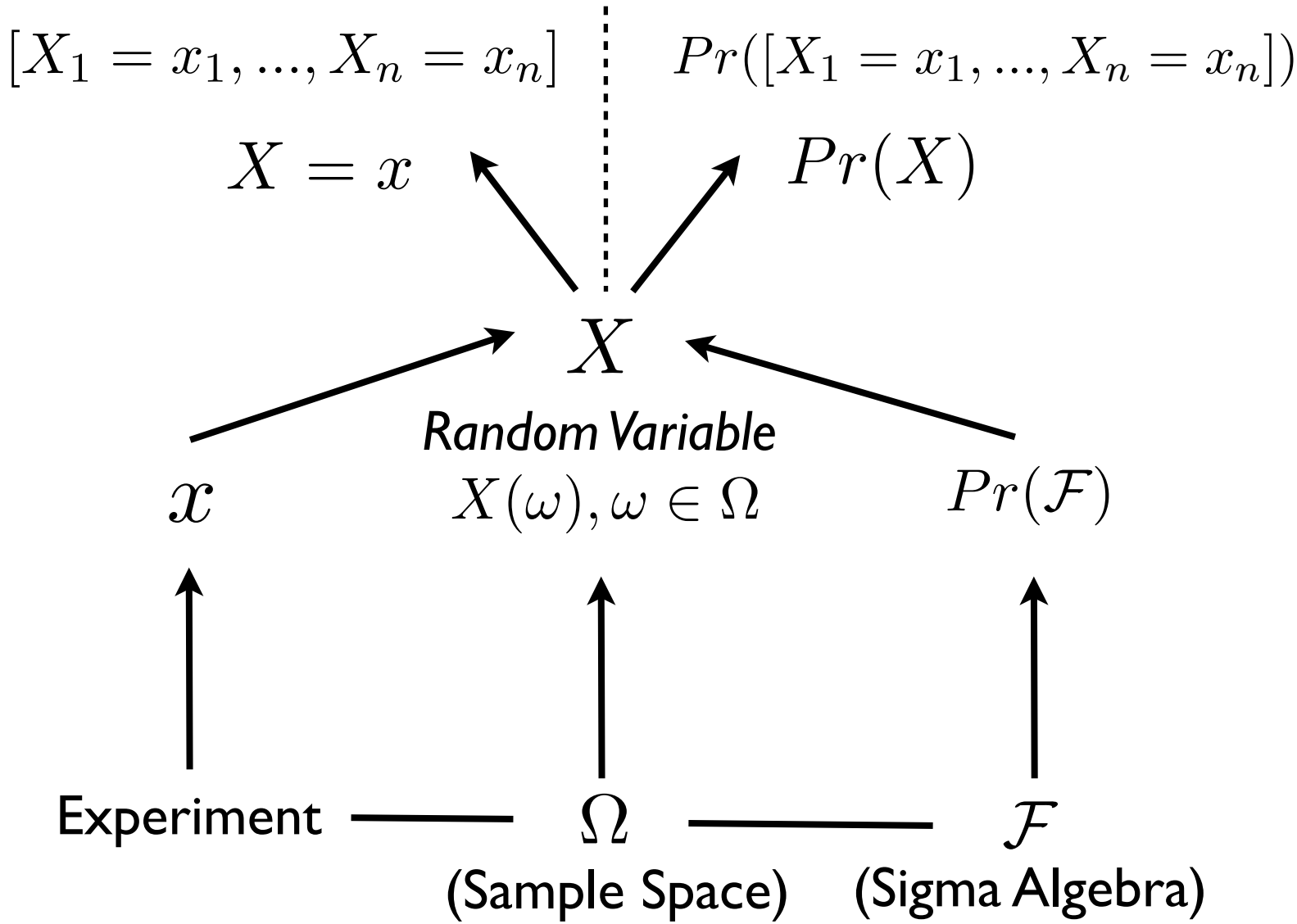
Introduction to inference III

- **Inference -** the process of reaching a conclusion about the true probability distribution (from an assumed family probability distributions, indexed by the value of parameter(s)) on the basis of a sample
- There are two major types of inference we will consider in this course: *estimation* and *hypothesis testing*
- Before we get to these specific forms of inference, we need to formally define: *experimental trials*, *samples*, *sample probability distributions* (or sampling distributions), *statistics*, *statistic probability distributions* (or statistic sampling distributions)

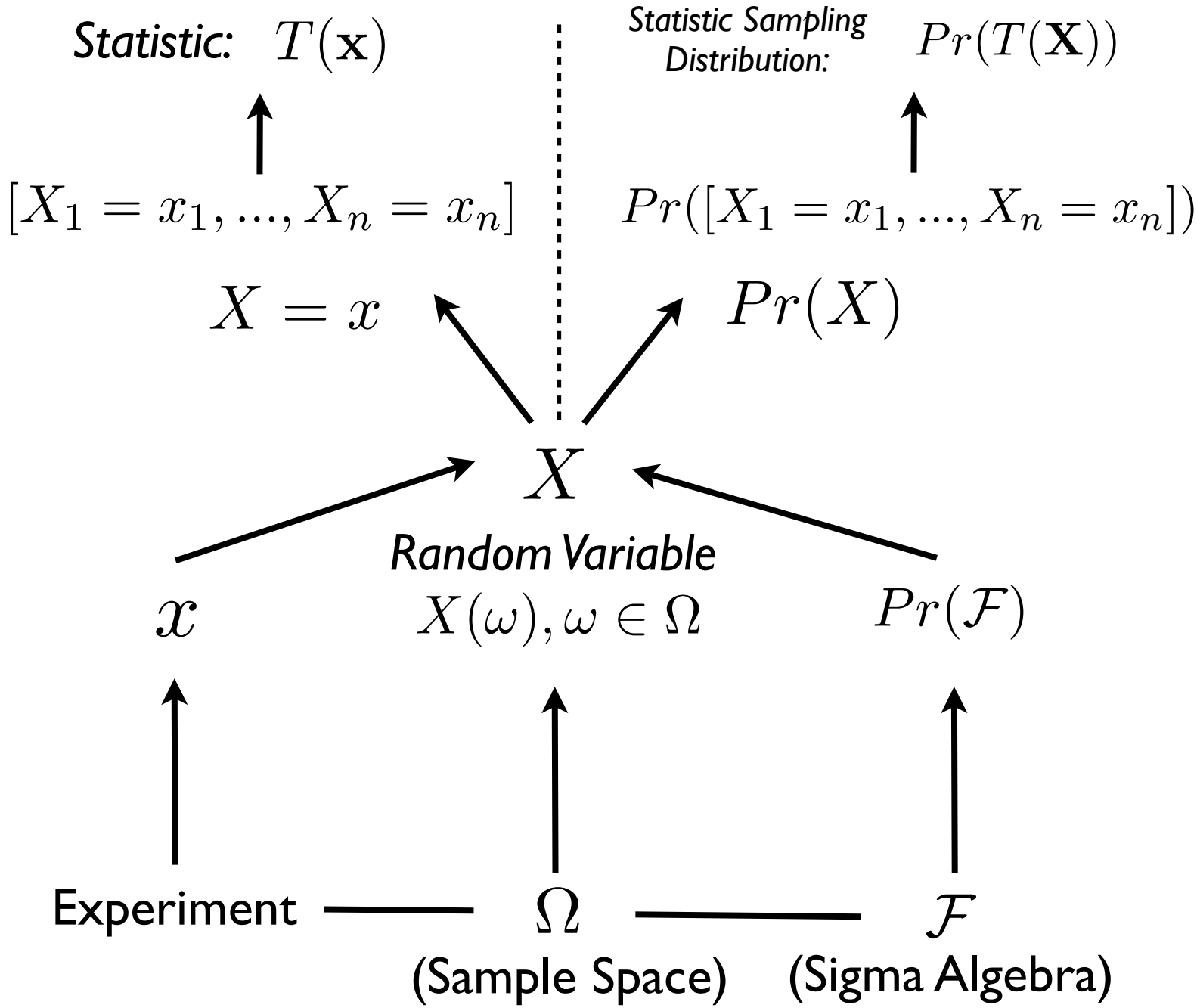
So far


$X = x$
$Pr(X)$
X
Random Variable
$X(\omega), \omega \in \Omega$
x
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Where we're headed: Samples



Then: Statistics!



Experiments to Samples (what we observe!)

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Experiment Outcome** - a possible outcome of the experiment
- **Sample Space** - set comprising all possible outcomes of an experiment
- **Experimental Trial** - one instance of an experiment
- **Sample** - (informal) results of one or more experimental trials
- Example (Experiment / Sample Space / Sample):
  - Coin flip / {H,T} / T, T, H, T, H
  - Two coin flips / {HH, HT, TH, TT} / HH, HT, HH, TH, HH
  - Measure heights in this class / Reals / 5'9", 5'2", 5'1", 6'0", 5'7"

Samples I

- **Sample** - repeated observations of a random variable *X*, generated by experimental trials
- We will consider samples that result from *n* experimental trials (what would be the ideal *n* = ideal experiment!?)
- Since a set of actual experimental outcomes may not be numbers (e.g., a set of H and T's) we want to map them to numbers…
- We already have the formalism to do this and represent a sample of size *n*, specifically this is a random vector:

$$[\mathbf{X} = \mathbf{x}] = [X_1 = x_1, ..., X_n = x_n]$$

- As an example, for our two coin flip experiment / number of tails r.v., we could perform *n*=2 experimental trials, which would produce a sample = random vector with two elements

Example: Observed Sample!

- For example, for our one coin flip experiment / number of tails r.v., we could produce a sample of n = 10 experimental trials, which might look like:

$$\mathbf{x} = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0]$$

- As another example, for our measure heights / identity r.v., we could produce a sample of n=10 experimental trails, which might look like:

$$\mathbf{x} = [-2.3, 0.5, 3.7, 1.2, -2.1, 1.5, -0.2, -0.8, -1.3, -0.1]$$

Samples II

- Recall that we have defined experiments (= experimental trials) in a probability / statistics setting where these involve observing individuals from a population or the results of a manipulation
- We have defined the possible outcome of an experimental trial, i.e. the sample space Ω
- We have also defined a random variable *X*, where this can take values representing the outcomes of our experimental trials, i.e., $X = x$
- Since the random variable X also has an induced probability distribution associated with it, we can also consider Pr(X), i.e., the probability of each possible outcome of an experiment or the entire sample!
- Since this defines a probability model Pr(*X*), we have shifted our focus from the sample space to the random variable

Sample Probability Distribution

- Note that since we have defined (or more accurately induced!) a probability distribution Pr(X) on our random variable, this means we have induced a probability distribution on the sample (!!):

$$Pr(\mathbf{X} = \mathbf{x}) = Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P_{\mathbf{X}}(\mathbf{x}) \text{ or } f_{\mathbf{X}}(\mathbf{x})$$

- This is the sample probability distribution or sampling distribution (often called the joint sampling distribution)

- While samples could take a variety of forms, we generally assume that each possible observation in the sample has the same form, such that they are identically distributed:

$$Pr(X_1 = x_1) = Pr(X_2 = x_2) = ... = Pr(X_n = x_n)$$

- We also generally assume that each observation is independent of all other observations:

$$Pr(\mathbf{X} = \mathbf{x}) = Pr(X_1 = x_1)Pr(X_2 = x_2)...Pr(X_n = x_n)$$

- If both of these assumptions hold, than the sample is independent and identically distributed, which we abbreviate as i.i.d.

That's it for today

- Next lecture, we will continue our discussion of inference by introducing statistics (and estimators)!