Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 9: Introduction to Hypothesis Testing I

Jason Mezey

Feb 21, 2023 (T) 8:05-9:20

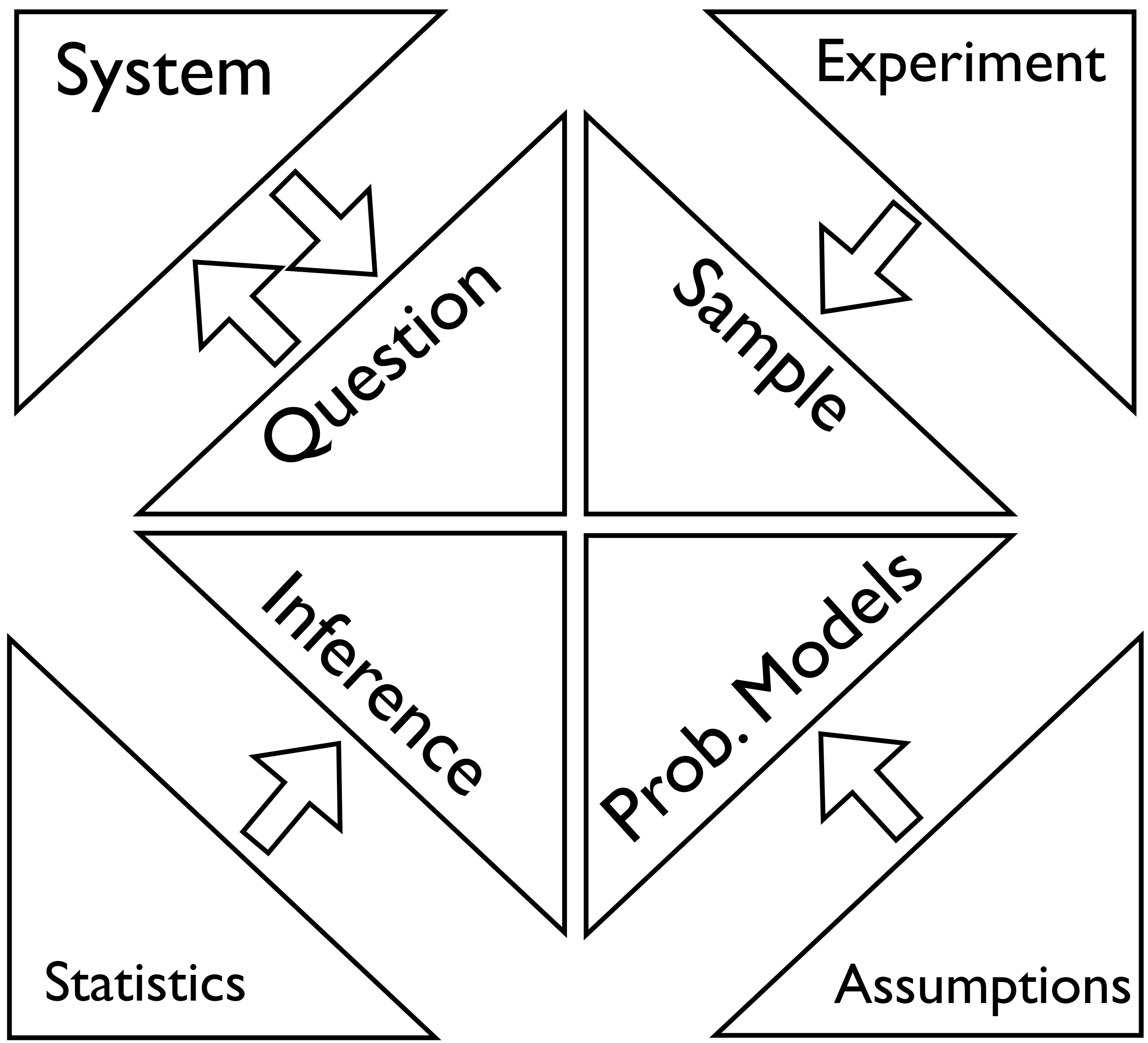
Announcements

- I will no longer respond to direct emails to me (only Piazza messages)
- CMS appears stable enough (those still having difficulties I will communicate with you directly on this)
- We will be back in the classroom Thurs (Feb 23)
- Homework #3 will be assigned Thurs (Feb 23)
- We will have office hours next week but TBD because of winter break (no office hours this week)

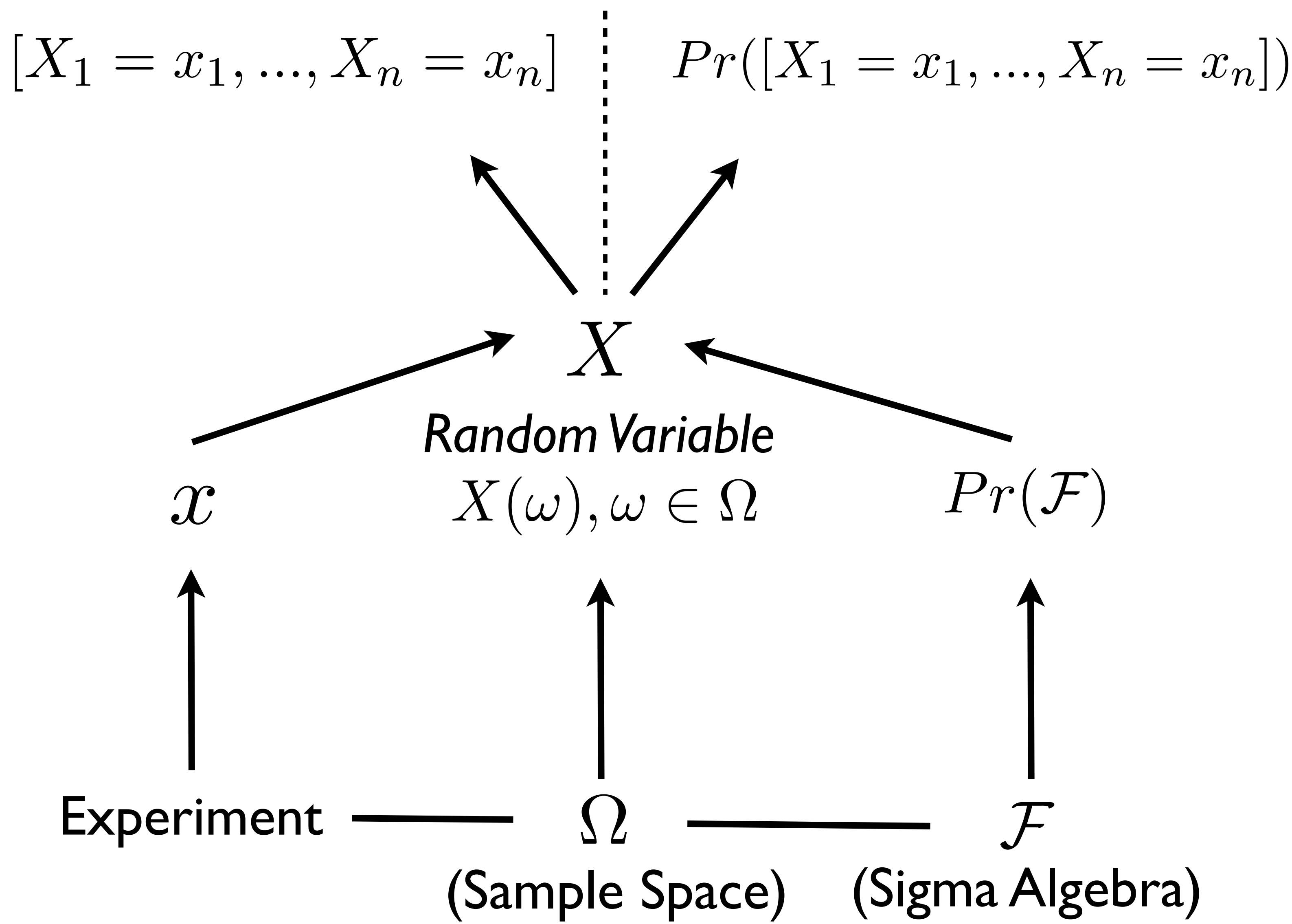
Summary of lecture 9: Introduction to Hypothesis Testing

- Last lecture, we (almost) completed our (general) discussion of estimators
- Today, we will (very) briefly discuss confidence intervals and begin our discussion of hypothesis testing (!!)

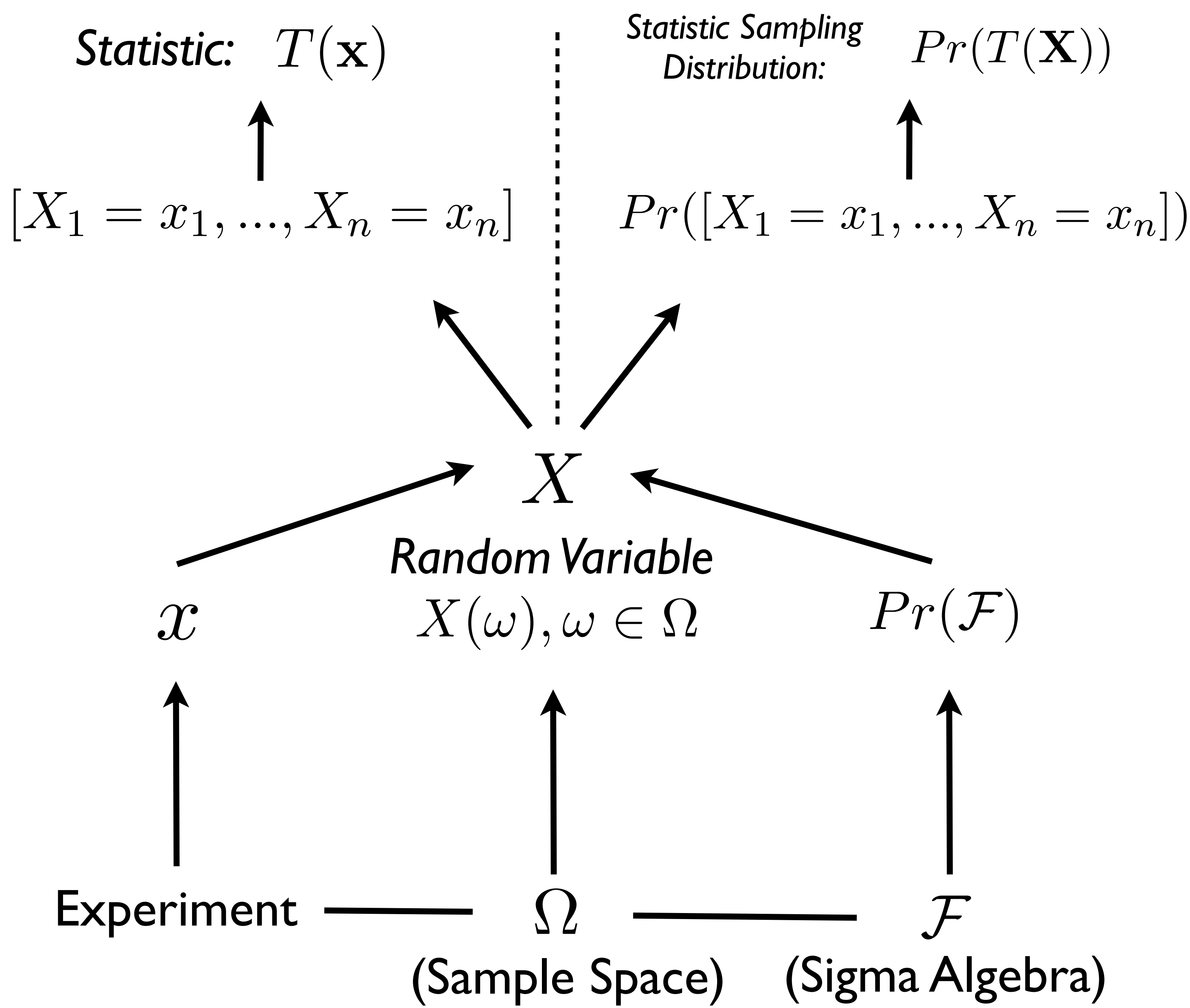
Conceptual Overview



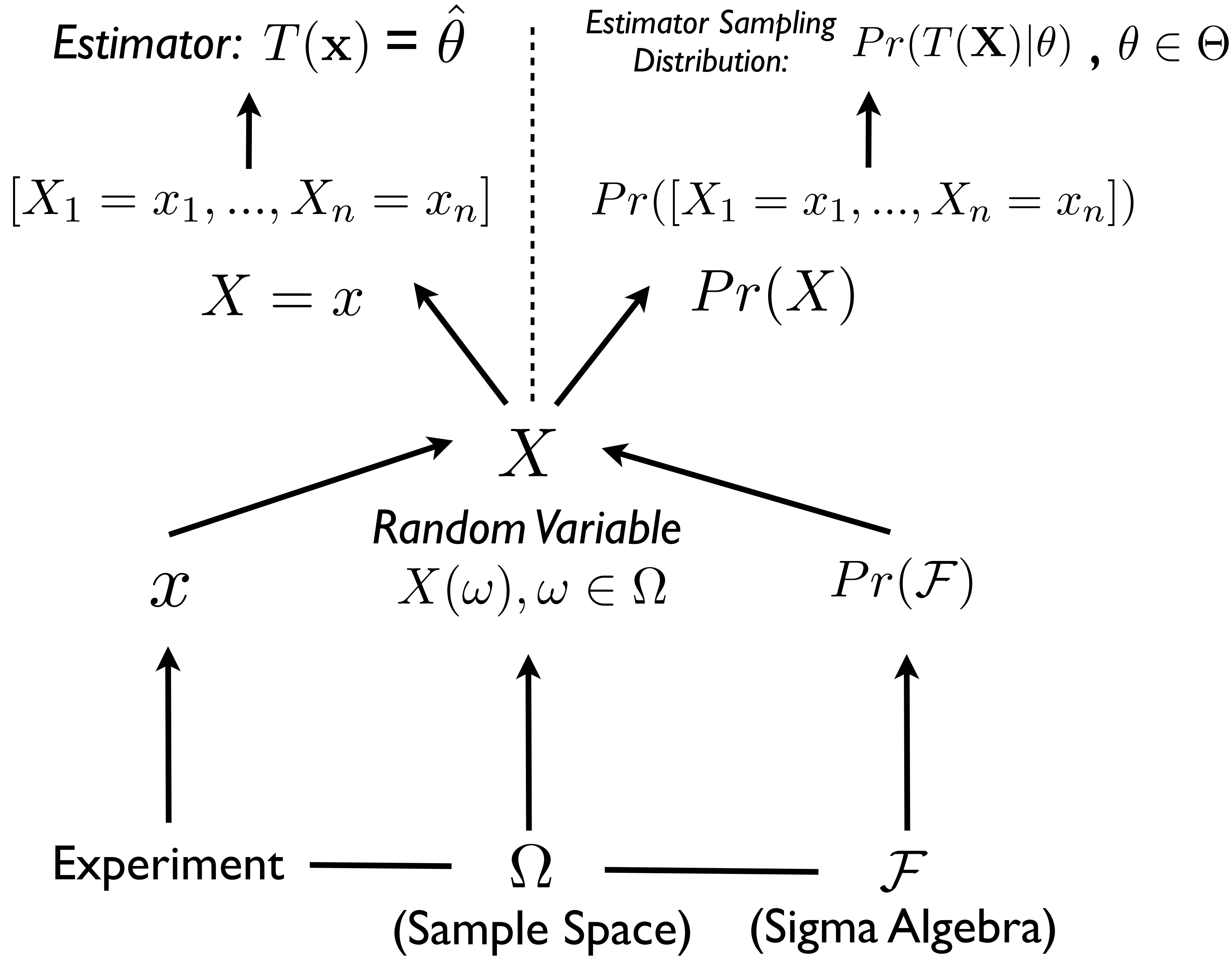
Samples


$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Statistics


Statistic: $T(\mathbf{x})$
Statistic Sampling Distribution: $Pr(T(\mathbf{X}))$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Estimators



Estimator: $T(\mathbf{x}) = \hat{\theta}$
Estimator Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
$X = x$
$Pr(X)$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Review: Probability models

- **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 the problem of using a sample 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 sample outcomes, the experiment, and the system of interest (!!!)

Review: Inference

- **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)

Review: Samples

- **Sample -** repeated observations of a random variable *X*, generated by experimental trials

- 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

- 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 (!!):

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

Review: Observed Sample

- It is important to keep in mind, that while we have made assumptions such that we can define the joint probability distribution of (all) possible samples that could be generated from *n* experimental trials, in practice we only observe one set of trials, i.e. one 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:

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

- In each of these cases, we would like to use these samples to perform inference (i.e. say something about our parameter of the assumed probability model)

- Using the entire sample is unwieldy, so we do this by defining a *statistic*

Review: Statistics

- As an example, consider our height experiment (reals as approximate sample space) / normal probability model (with true but unknown parameters $\theta = [\mu, \sigma^2]$ / identity random variable

- If we calculate the following statistic:

$$T(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

what is $\Pr(T(\mathbf{X}))$?

- Are the distributions of $X_i = x_i$ and $\Pr(T(\mathbf{X}))$ always the same?

Review: Estimators

- **Estimator** - a statistic defined to return a value that represents our best evidence for being the true value of a parameter
- In such a case, our statistic is an *estimator* of the parameter: $T(\mathbf{x}) = \hat{\theta}$
- Note that ANY statistic on a sample can in theory be an estimator.
- However, we generally define estimators (=statistics) in such a way that it returns a reasonable or "good" estimator of the true parameter value under a variety of conditions
- How we assess how "good" an estimator depends on our criteria for assessing "good" and our underlying assumptions

Review: Estimator example I

- As an example, let's construct an estimator
- Consider the single coin flip experiment / number of tails random variable / Bernoulli probability model family (parameter p) / fair coin model (assumed and unknown to us!!!) / sample of size *n*=10
- We want to estimate p, where a perfectly reasonable estimator is:

$$T(\mathbf{X} = \mathbf{x}) = \hat{\theta} = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- e.g. this statistic (=mean of the sample) would equal 0.5 for the following particular sample (will it always?)

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

Review: Estimator example II

- Let's continue with our example constructing the probability model
- Consider the single coin flip experiment / number of tails random variable

$$\Omega = \{H, T\} \qquad X : X(H) = 0, X(T) = 1$$

- Bernoulli probability model family (parameter p)

$$X \sim p^X (1-p)^{1-X}$$

- Sample of size *n*=10

$$[\mathbf{X} = \mathbf{x}] = [X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}]$$

- Sampling distribution (pmf of sample) if i.i.d. (!!)

$$[X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}] \sim p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} ... p^{x_{10}}(1-p)^{1-x_{10}}$$

Review: Introduction to maximum likelihood estimators (MLE)

- We will generally consider *maximum likelihood estimators* (MLE) in this course
- Now, MLE's are very confusing when initially encountered...
- However, the critical point to remember is that an MLE is just an estimator (a function on a sample!!),
- i.e. it takes a sample in, and produces a number as an output that is our estimate of the true parameter value
- These estimators also have sampling distributions just like any other statistic!
- The structure of this particular estimator / statistic is complicated but just keep this big picture in mind

Review: Introduction to MLE's

- A maximum likelihood estimator (MLE) has the following definition:

$$MLE(\hat{\theta}) = \hat{\theta} = argmax_{\theta \in \Theta} L(\theta|\mathbf{x})$$

- Recall that this statistic still takes in a sample and outputs a value that is our estimator (!!) Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)

- Sometimes these estimators have nice forms (equations) that we can write out

- For example the maximum likelihood estimator when considering a sample for our single coin example / number of tails is:

$$MLE(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- And for our heights example:

$$MLE(\hat{\mu}) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad MLE(\hat{\sigma}^2) = \frac{1}{n}\sum_{i}^{n}(x_i - \overline{x})^2$$

Brief Introduction: Properties of estimators I

- Remember (!!) for all the complexity in thinking about, deriving, etc. MLE's these are still just estimators (!!), i.e. they are statistics that take a sample as input and output a value that we consider an estimate of our parameter

- MLE in general have nice properties (and we will largely use them in this class!), but there are many other estimators that we could use

- This is because there is no "perfect" estimator and each estimator that we can define has different properties, some of which are desirable, some are less desirable

- In general, we do try to use estimators that have "good" properties based on well defined criteria

- In this class, we will briefly consider two: *unbiasedness* and *consistency*

Properties of estimators II

- We measure the bias of an estimator as follows (where an unbiased estimator has a bias of zero):

$$Bias(\hat{\theta}) = \mathrm{E}\hat{\theta} - \theta$$

- We consider an estimator to be consistent if it has the following property

$$lim_{n \rightarrow \infty} Pr(|\hat{\theta} - \theta| < \epsilon) = 1$$

- Note that one can have an estimator that is consistent but not unbiased (and vice versa!)

- As an example of the former, the following MLE is biased but consistent

$$MLE(\hat{\sigma^2}) = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2$$

- An unbiased estimator of this parameter is the following:

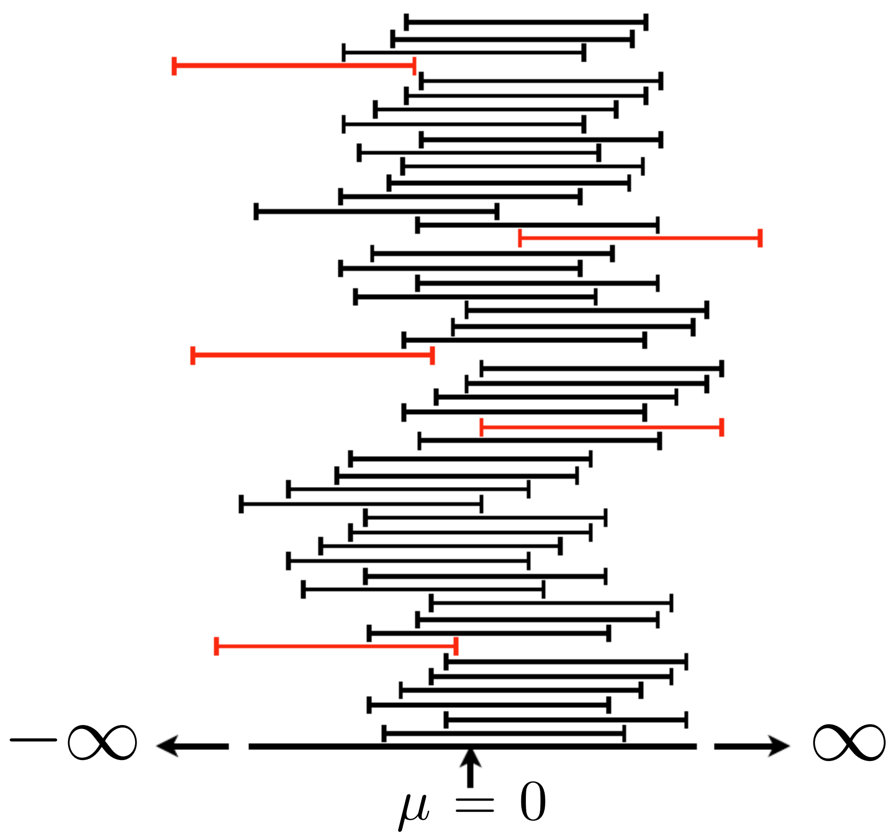
$$\hat{\sigma^2} = \frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2$$

Confidence intervals I

- For the estimation framework we have considered thus far, our goal was to define an estimator that provides a "reasonable guess given the sample" of the true value of the parameter
- This is called "point" estimation since the true parameter has a single value (i.e. it is a point)
- We could also estimate an interval, where our goal is to say something about the chances that the true parameter (the point) would fall in the interval
- **confidence interval** (CI) - an estimate of an interval defined such that if it were estimated individually for an infinite number of samples, a specific percentage of the estimated intervals would contain the true parameter value
- Don't worry if this concept seems confusing (it is!) let's first consider an example and then discuss some basics

Confidence intervals II

- As an example, assume the standard normal r.v. $X \sim N(0,1)$ correctly describes our sampling distribution if we were to produce 50 independent samples, each of size n=10 and we were to estimate a CI for each one, we would expect to get the following:



Confidence intervals III

- A CI is therefore calculated from a sample (and reflects uncertainty!)
- A CI is an estimate of an *interval*, as opposed to an estimate of a parameter, which is a *point* estimate (more technically, the CI is an estimate of the endpoints of the interval)
- This estimated interval of a CI (generally) includes the estimate of the parameter in the "middle"
- In general, a CI provides a measure of "confidence" in the sense that the smaller the interval, the more "confidence" we have in our estimate (if this seems circular, it is meant to be!)
- In general, we can make the CI smaller with a larger sample size *n* and by decreasing the probability that the interval contains the true parameter value, i.e. a 95% CI is smaller than a 99% CI
- NOTE THAT A 95% CI estimated from one sample does not contain the true parameter value with a probability of 0.95 (!!!) - the definition of a CI says if we performed an infinite number of samples, and calculated the CI for each, then 95% of these intervals would contain the true parameter value (strange?)

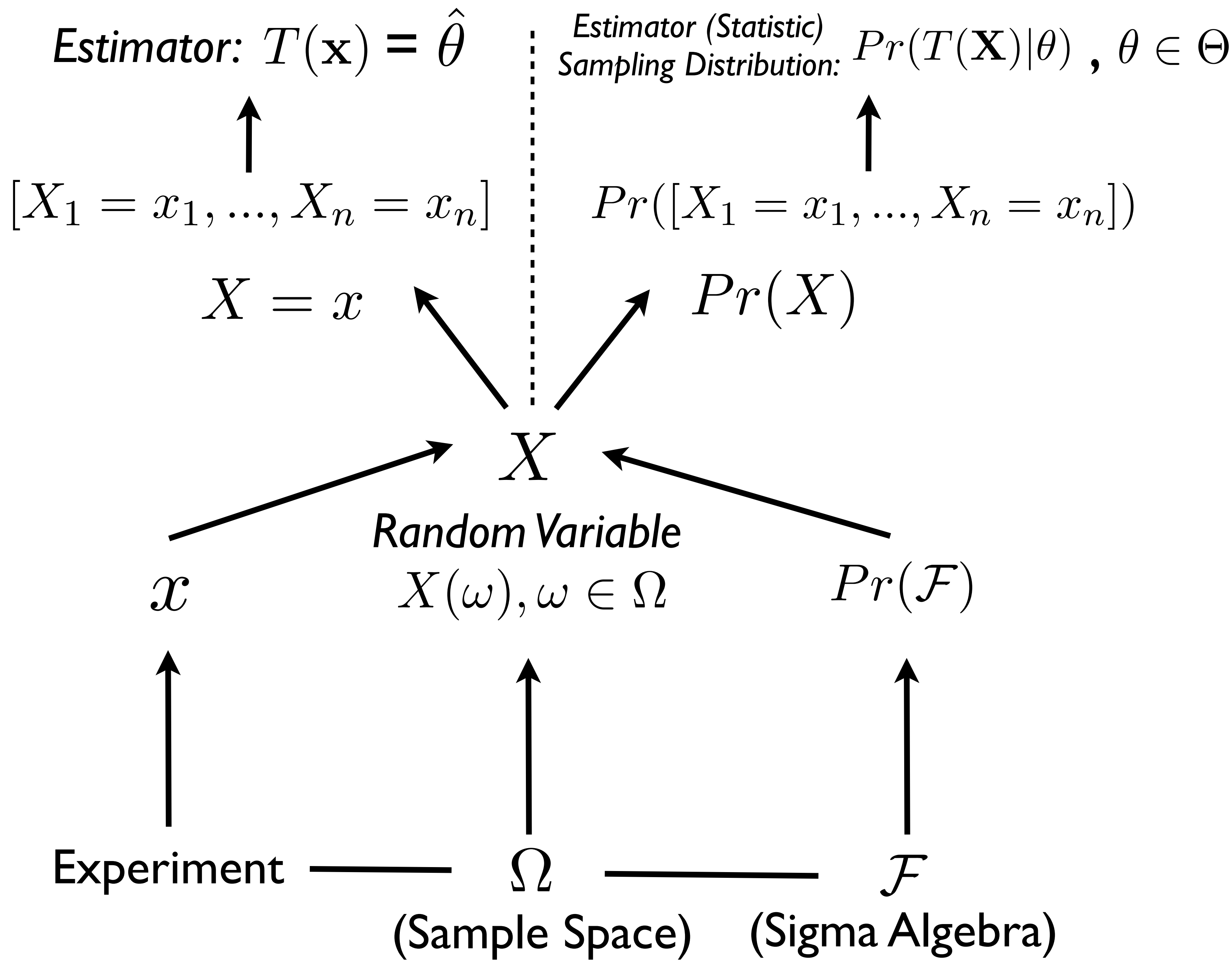
Review of essential concepts

- **Inference -** the process of reaching a conclusion about the true probability distribution (from an assumed family of probability distributions indexed by parameters) on the basis of a sample
- **System, Experiment, Experimental Trial, Sample Space, Sigma Algebra, Probability Measure, Random Vector, Parameterized Probability Model, Sample, Sampling Distribution, Statistic, Statistic Sampling Distribution, Estimator, Estimator Sampling distribution**

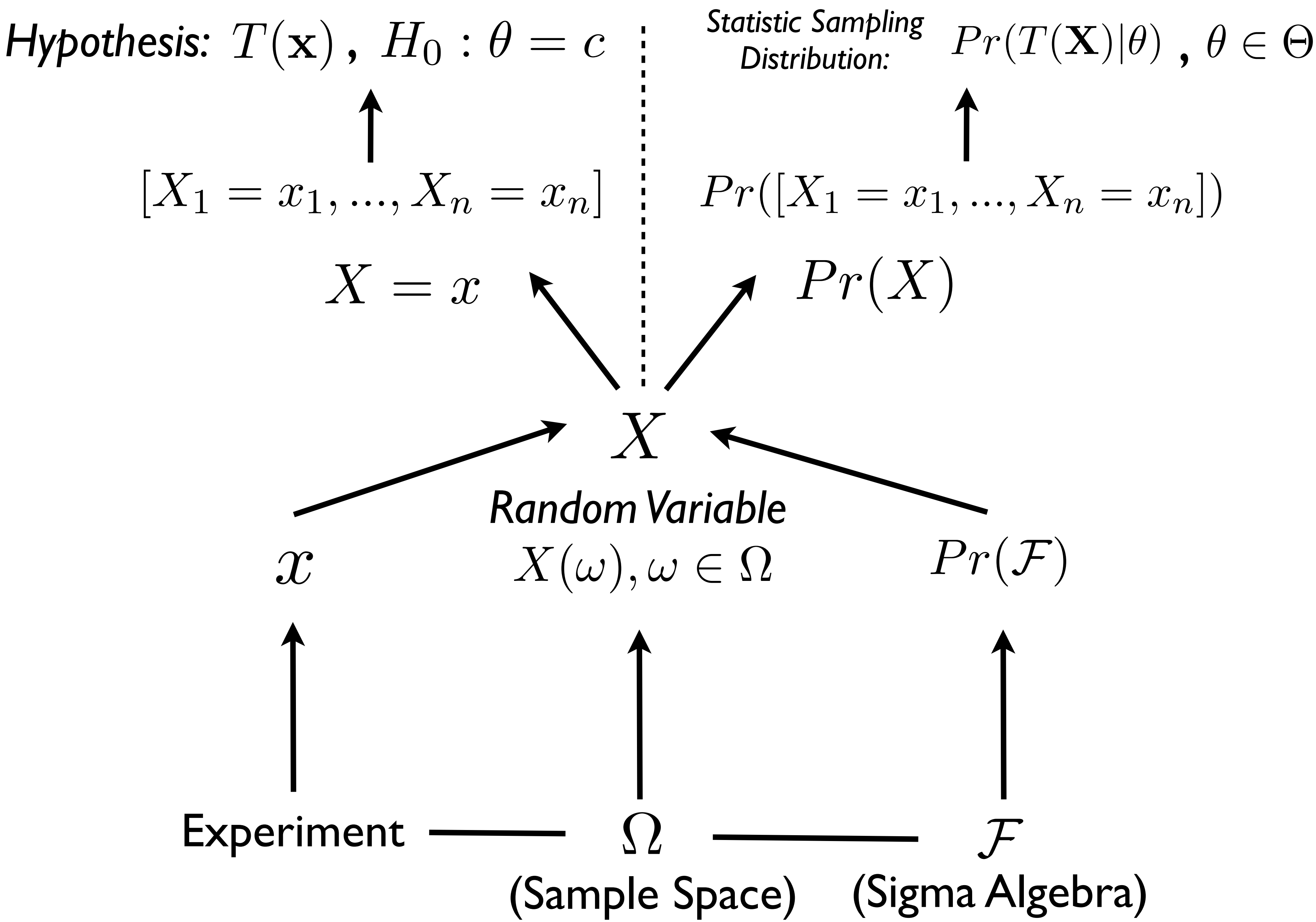
Estimation and Hypothesis Testing

- Thus far we have been considering a “type” of inference, *estimation*, where we are interested in determining the actual value of a parameter
- We could ask another question, and consider whether the parameter is NOT a particular value
- This is another “type” of inference called *hypothesis testing*
- We will use hypothesis testing extensively in this course

Estimators


Estimator: $T(\mathbf{x}) = \hat{\theta}$
Estimator (Statistic) Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
$X = x$
$Pr(X)$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

Hypothesis Tests



Hypothesis testing I

- To build this framework, we need to start with a definition of hypothesis
- **Hypothesis -** an assumption about a parameter
- More specifically, we are going to start our discussion with a *null hypothesis*, which states that a parameter takes a specific value, i.e. a constant

$$H_0 : \theta = c$$

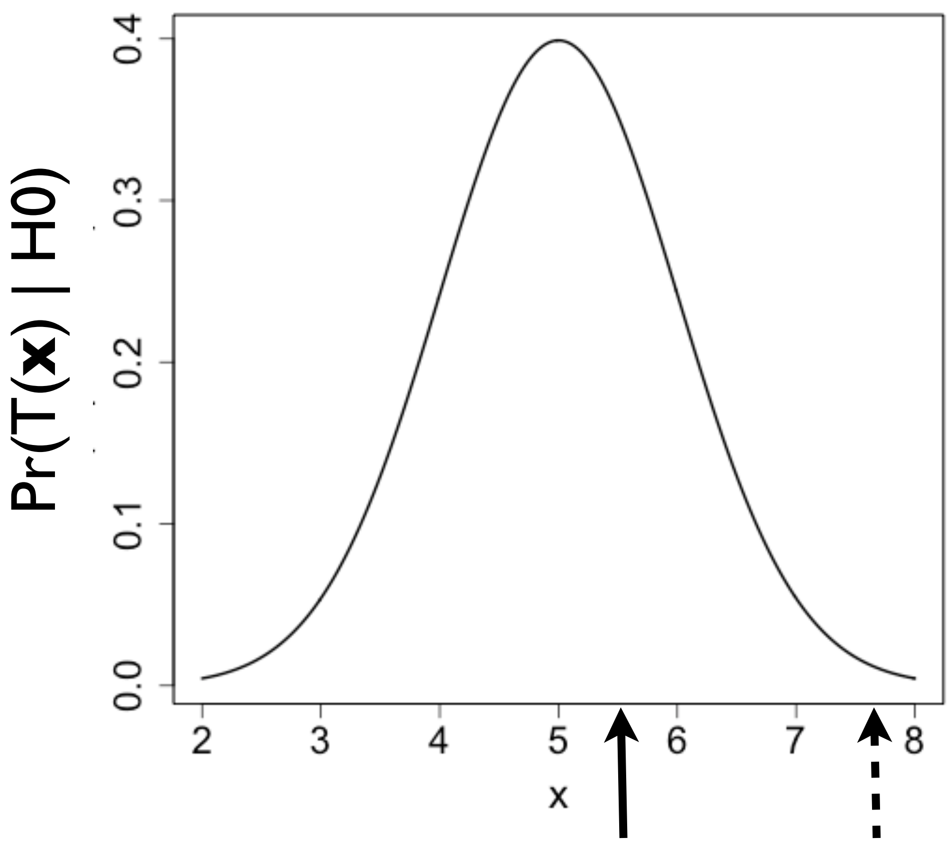
- For example, for our height experiment / identity random variable, we have $Pr(X|\theta) \sim N(\mu, \sigma^2)$ and we could consider the following null hypothesis:

$$H_0 : \mu = 0$$

Hypothesis testing II

- As example, consider our height experiment (reals as sample space) / identity random variable X / normal probability model $\theta = [\mu, \sigma^2]$ / sample *n*=1 (of one height measurement) / identity statistic $T(x) = x$ (takes the height measured height)

- Let's assume that $\sigma^2 = 1$ and say we are interested in testing the following null hypothesis $H_0 : \mu = 5.5$ such that we have the following probability distribution of the statistic under the null hypothesis:



That's it for today

- Next lecture, we will continue our discussion of hypothesis testing!