Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 7: Introduction to Statistics and Estimators

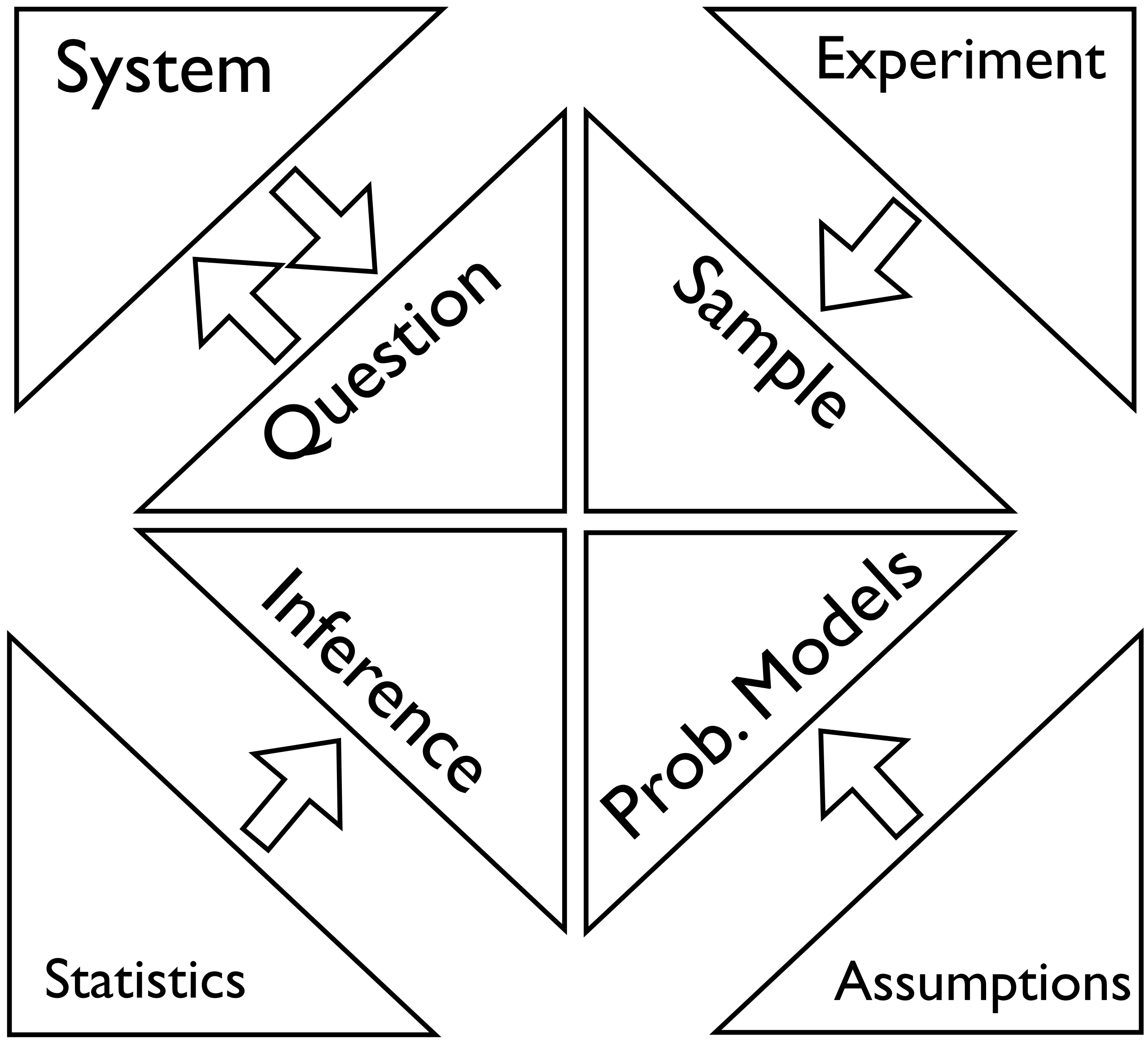
Jason Mezey

Feb 14, 2023 (T) 8:05-9:20

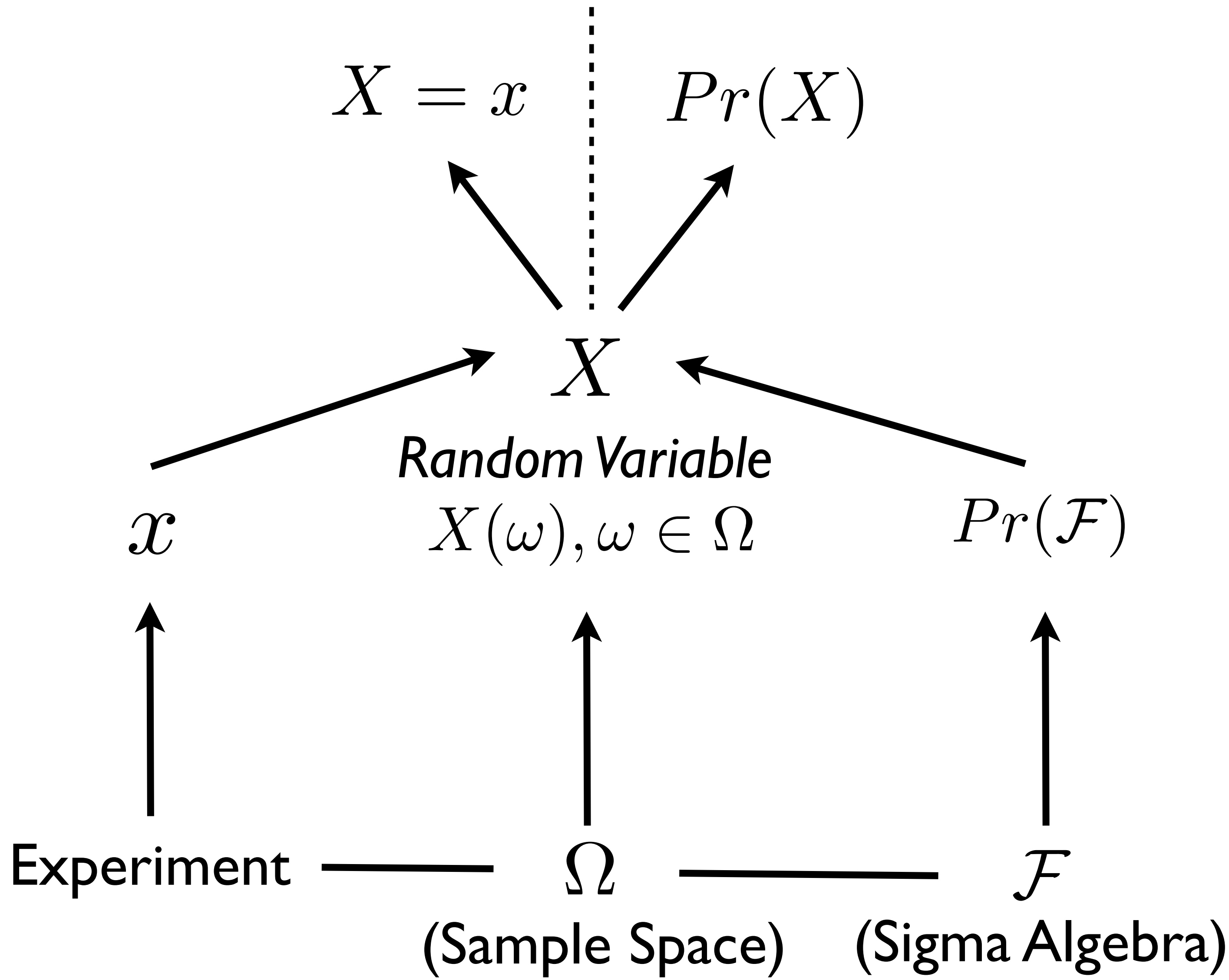
Summary of lecture 7: Statistics and Estimators

- Last lecture, we discussed inference and the critical concept of samples and sampling distributions (i.e., the probability distribution of the sample)
- Today we will begin our introduction to statistics and how we use these for one type of inference: estimation (and we will begin our introduction to likelihood…)

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

- 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

- **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: 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*):

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

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

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

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

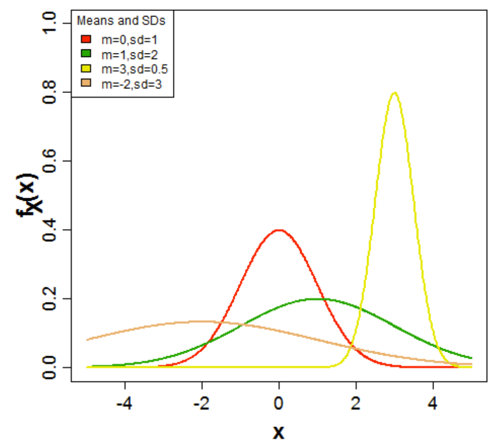
Review: Continuous parameterized examples

- 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:

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



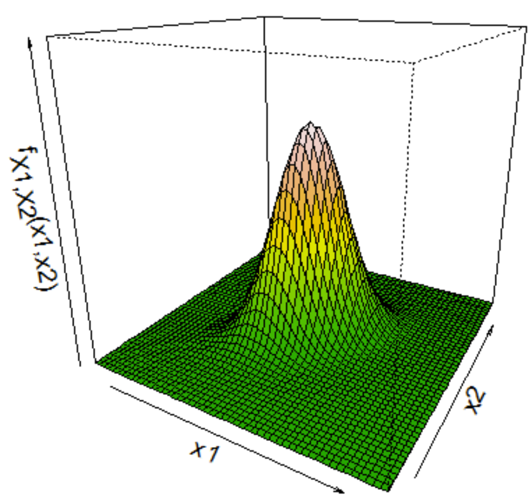
- There are many other continuous examples: uniform, exponential, etc.

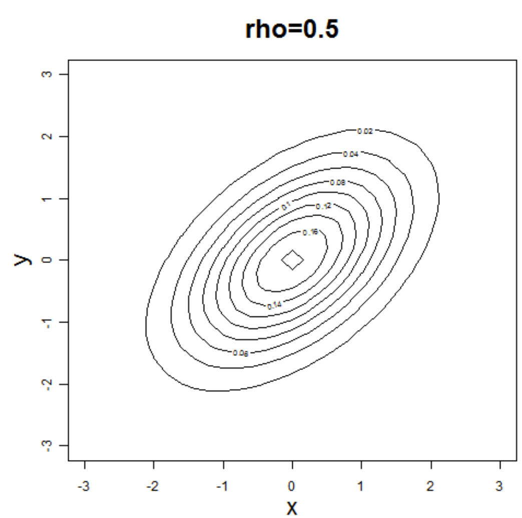
Review: Example for random vectors

- 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 a sample (generated by 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 (although note we have not yet provided formalism for a sample!!)
- 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

Review: Introduction to inference

- Our eventual goal is to use a sample (generated by an experiment) to provide an answer to a question (about a system)
- 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, the probability of possible random vectors that represent samples, and the probability of possible values of a statistic)
- 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

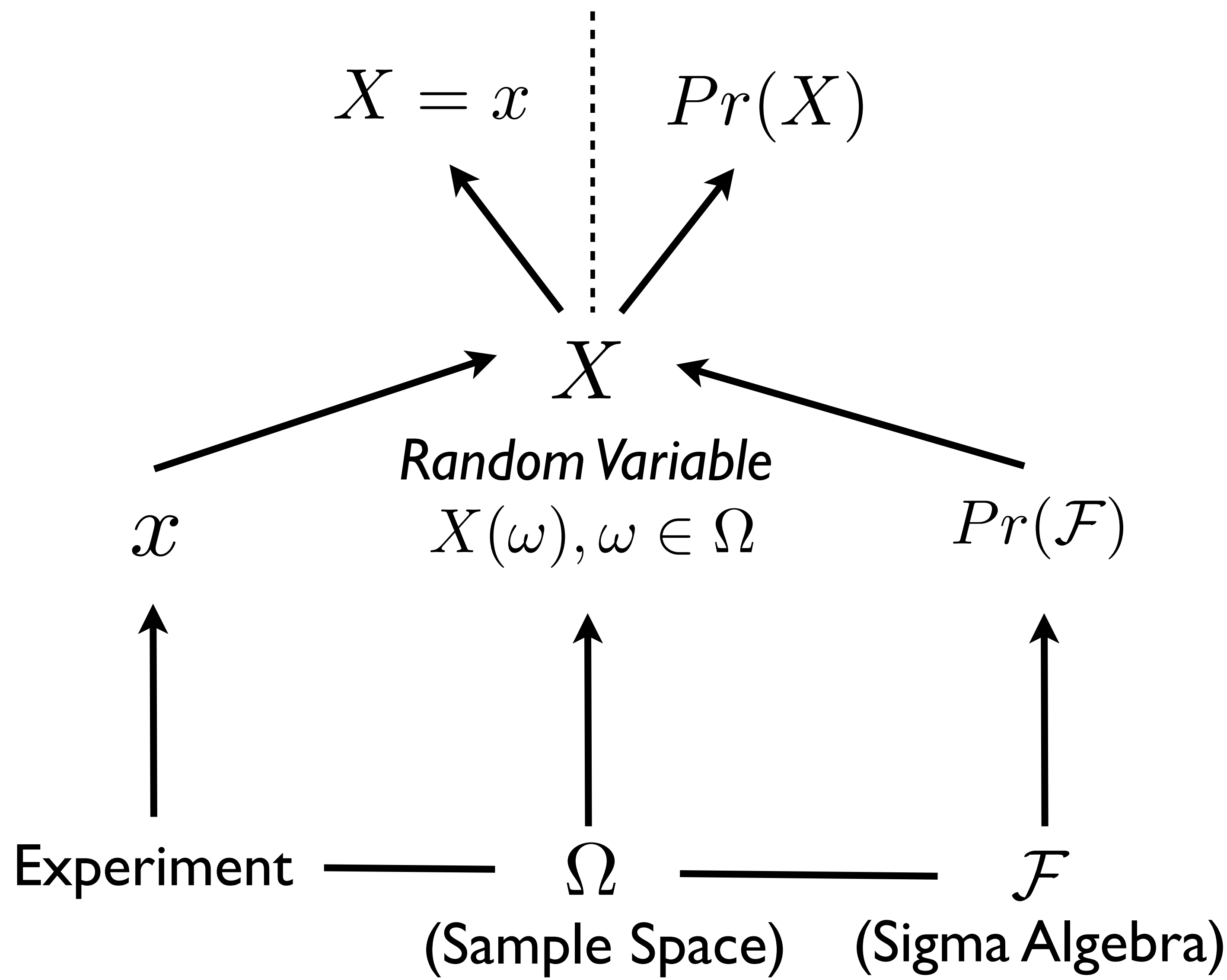
Review: Inference

- **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*

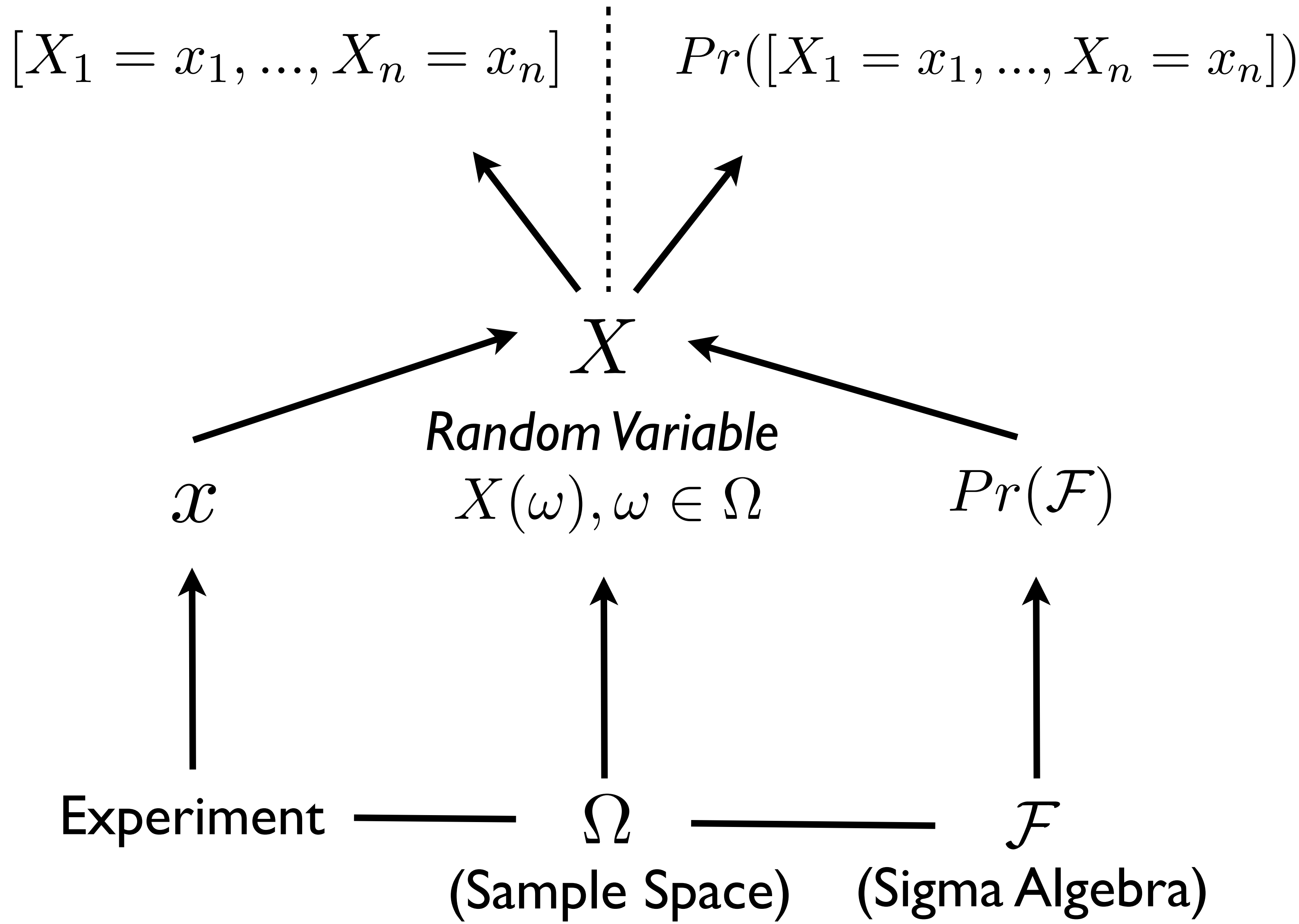
- Before we get to these specific forms of inference, we need to formally define: *experimental trials*, *samples*, *sample probability distributions* (or sampling distributions), *statistics*, *statistic probability distributions* (or statistic sampling distributions)

So far



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

Where we’re headed: Samples



Review: 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”

Review: 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:

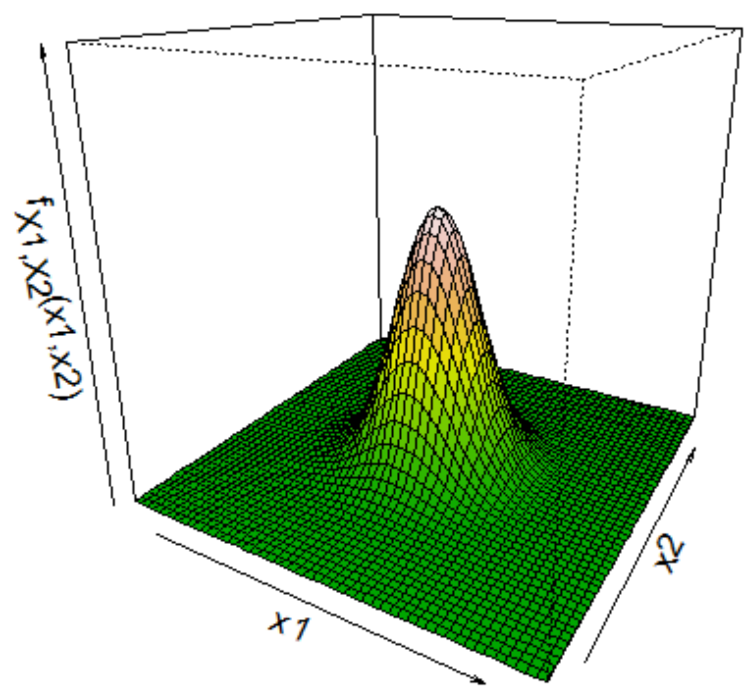
$$\mathbf{x} = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0]$$

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

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

Review: Example of sampling distributions

- 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 assume an i.i.d sample, each sample $X_i = x_i$ has a normal distribution with parameters $\theta = [\mu, \sigma^2]$ and each is independent of all other $X_j = x_j$
- For example, the sampling distribution for an i.i.d sample of $n = 2$ is:



Review: 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.

Samples: Technical Notes

- Technical note I: when considering a sample where each observation is independent, the actual sample space is actually a "product" of sample spaces (product space) where each random variable in the sample (the random vector of the sample) is a function on one of the sample spaces:

$$\Omega_{ind} = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$$

- If this is the case, why have we bothered considering a case where we define multiple random variables on a "single" sample space, e.g., $X_1(\Omega)$ and $X_2(\Omega)$?

- This largely for conceptual reasons, e.g., when considering samples we may want to consider each observation of the sample to contain two observations, such that each observation in the sample is a vector (!!) and the sample is a set of vectors (a matrix!)

- Technical note II: regardless of the size of *n*, there is a sampling distribution although as $n \rightarrow \infty$ this becomes a probability distribution that only assigns a non-zero value (one!) to only the entire sample space element of the Sigma Algebra

The 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:

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

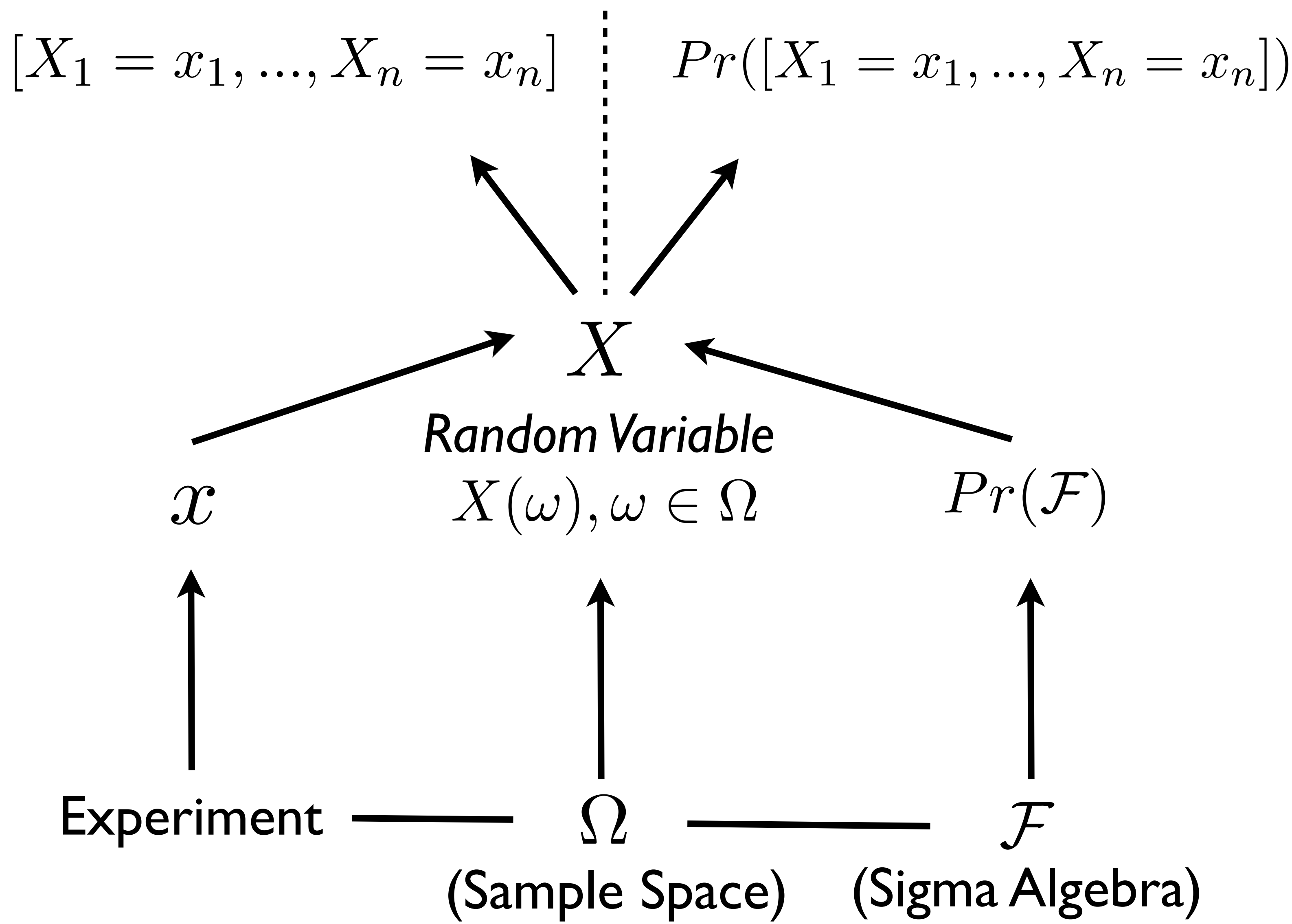
- As another example, for our measure heights / identity r.v., we could produce a sample of n=10 experimental trails, which might look like:

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

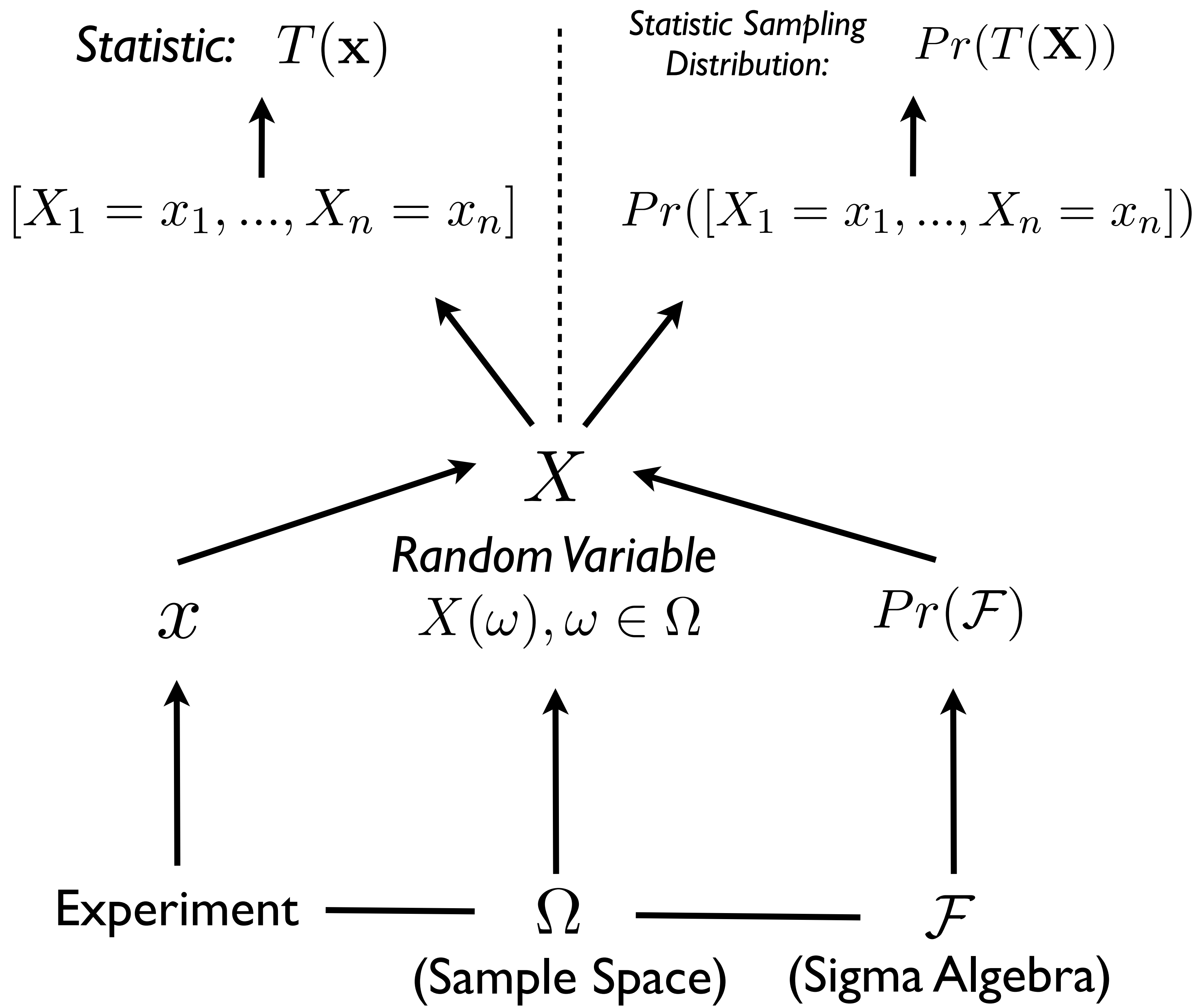
- In each of these cases, we would like to use these samples to perform inference (i.e. say something about our parameter of the assumed probability model)

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

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)

Statistics I

- **Statistic** - a function on a sample
- Note that a statistic *T* is a function that takes a vector (a sample) as an input and returns a value (or vector):

$$T(\mathbf{x}) = T(x_1, x_2, ..., x_n) = t$$

- For example, one possible statistic is the mean of a sample:

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

- It is critical to realize that, just as a probability model on X induces a probability distribution on a sample, since a statistic is a function on the sample, this induces a probability model on the statistic: the *statistic probability distribution* or the *sampling distribution* of the statistic (!!)

Statistics II

- 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

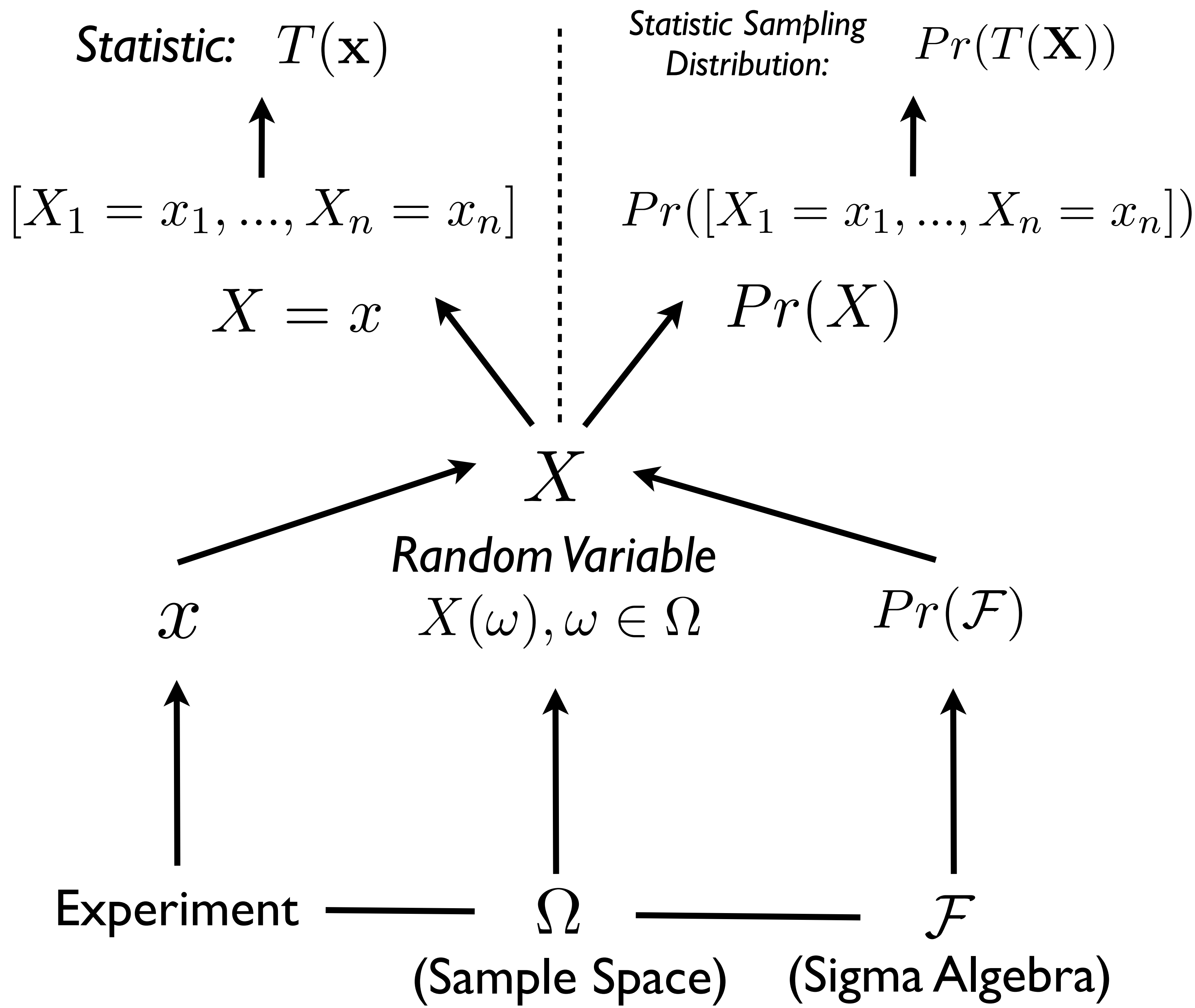
- If we calculate the following statistic:

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

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

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

Statistics


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

Statistics and estimators I

- Recall for the purposes of inference, we would like to use a sample to say something about the specific parameter value (of the assumed) family or probability models that could describe our sample space

- Said another way, we are interested in using the sample to determine the "true" parameter value that describes the outcomes of our experiment

- An approach for accomplishing this goal is to define our statistic in a way that it will allow us to say something about the true parameter value

- In such a case, our statistic is an estimator of the parameter: $T(\mathbf{x}) = \hat{\theta}$

- There are many ways to define estimators (we will focus on maximum likelihood estimators in this course)

- Intuitively, an estimator is the value for which we have the best evidence for being the true value of the parameter (our "best guess") based on the sample, given uncertainty and our assumptions

- Note that without an infinite sample, we will never know the true value of the parameter with absolute certainty (!!)

Statistics and estimators II

- **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

Statistics and estimators III

- Since our underlying probability model induces a probability distribution on a statistic, and an estimator is just a statistic, there is an underlying probability distribution on an estimator:

$$Pr(T(\mathbf{X} = \mathbf{x})) = Pr(\hat{\theta})$$

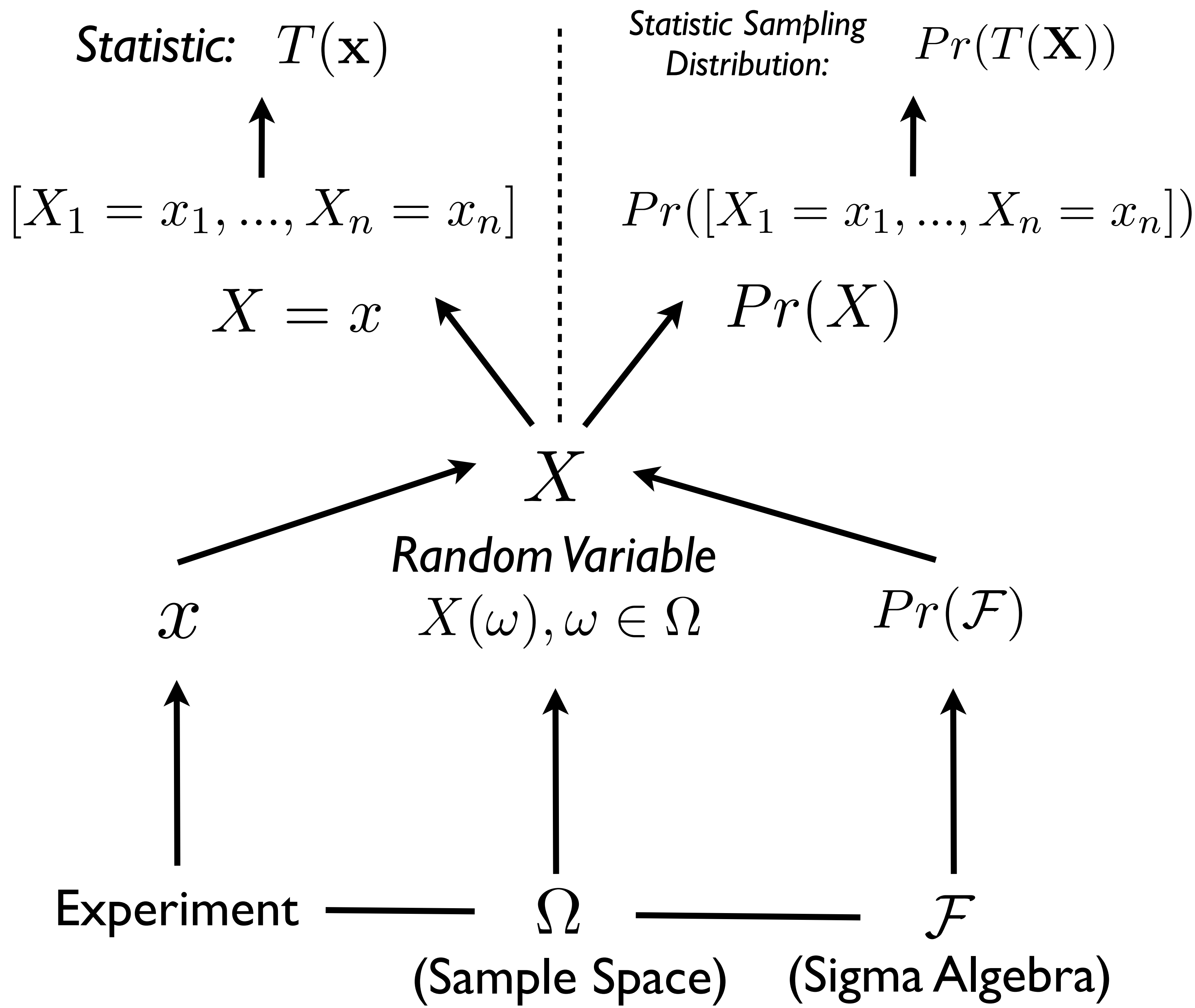
- Our estimator takes in a vector as input (the sample) and may be defined to output a single value or a vector of estimates:

$$T(\mathbf{X} = \mathbf{x}) = \hat{\theta} = \left[\hat{\theta}_1, \hat{\theta}_2, ...\right]$$

- We cannot define a statistic that always outputs the true value of the parameter for every possible sample (hence no perfect estimator!)

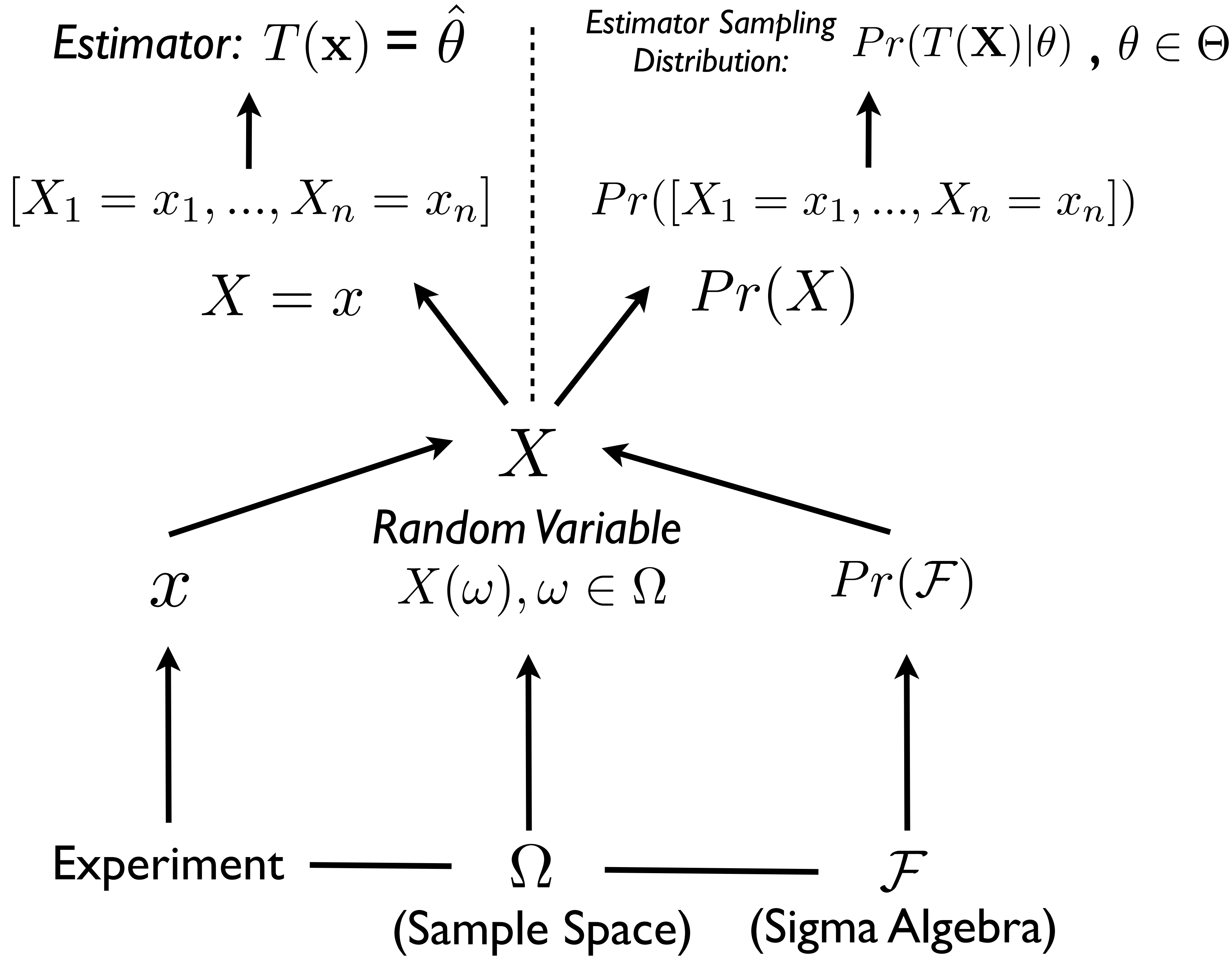
- There are different ways to define "good" estimators and lots of ways to define "bad" estimators (examples?)

Statistics



Statistic: $T(\mathbf{x})$
Statistic Sampling Distribution: $Pr(T(\mathbf{X}))$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
$X = x$
$Pr(X)$
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)

That’s it for today

- Next lecture, we will continue our discussion of estimators (and Maximum Likelihood Estimators)!