Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 8: Introduction Maximum Likelihood Estimators

Jason Mezey

Feb 16, 2023 (Th) 8:05-9:20

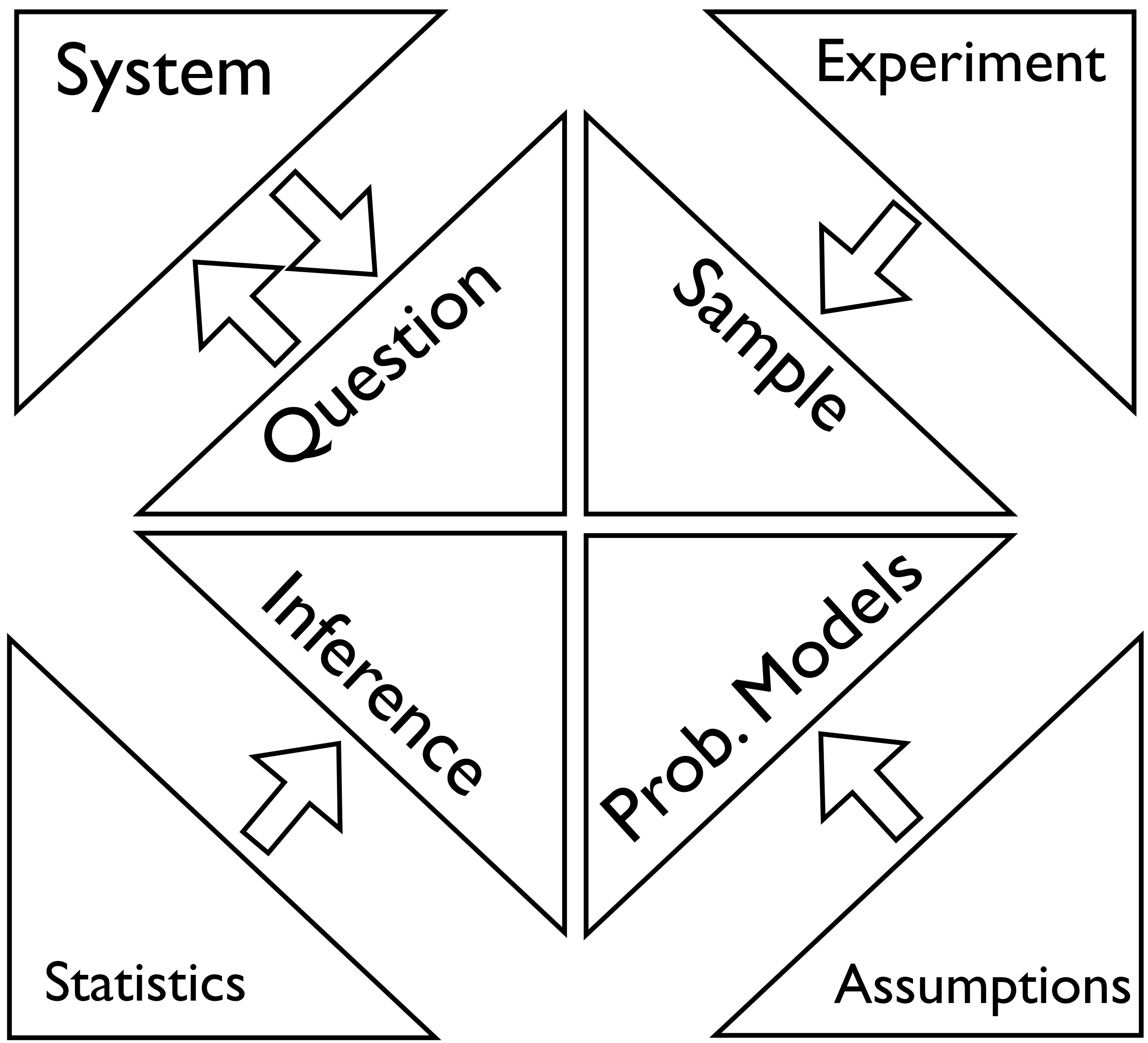
Announcements

- Homework #2 due 11:59pm tomorrow (Fri., Feb 17) (!!)
- Next lecture (Tues., Feb 21) will be entirely by zoom FOR ALL STUDENTS (details to follow by Piazza message)
- There will be NO office hours next week (!!)
- There will be an office hours before homework #3 is due (this will likely be assigned Feb 23)

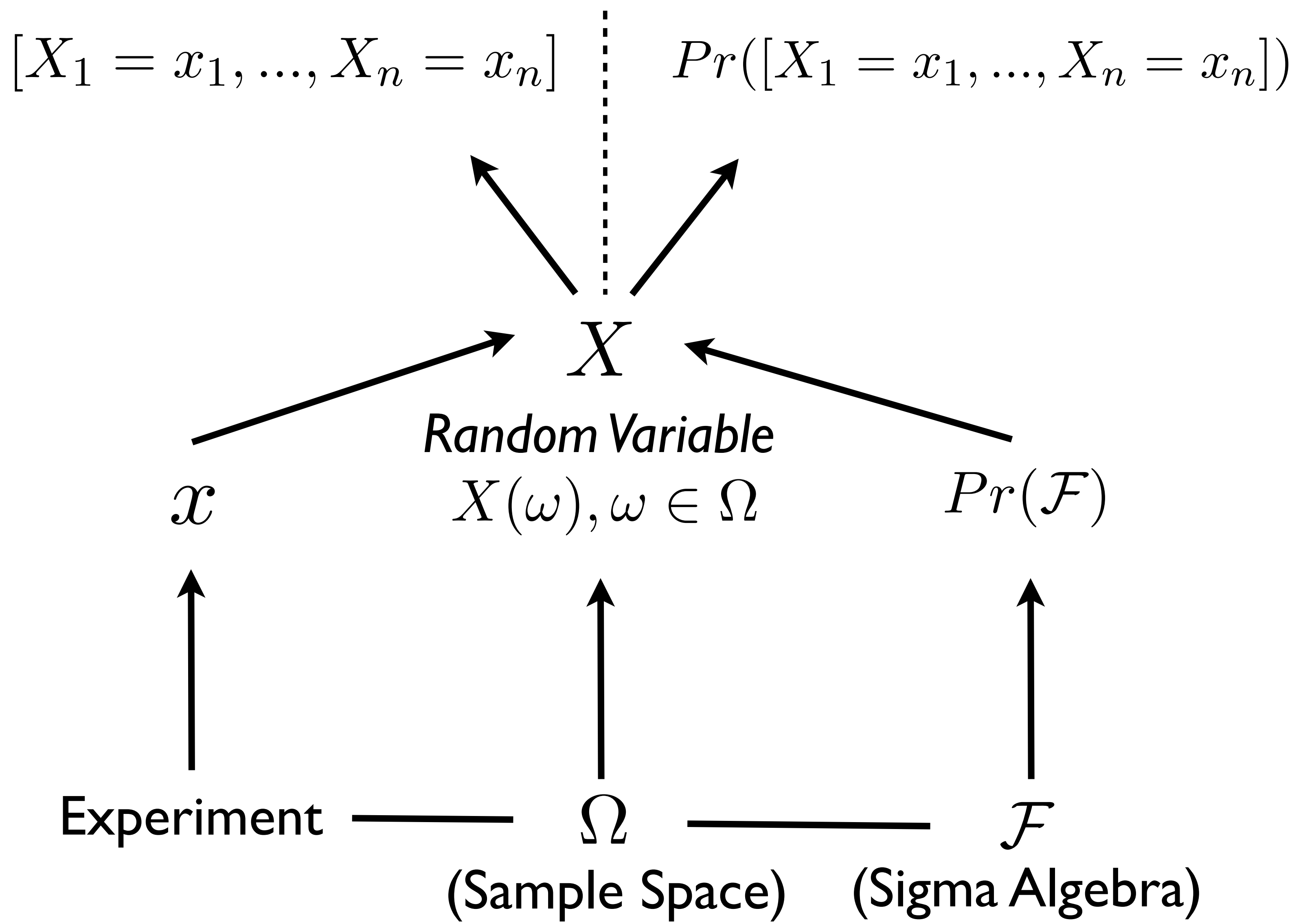
Summary of lecture 8: Maximum Likelihood Estimators

- Last lecture, we discussed statistics and how we use these for one type of inference: estimation
- Today, we will discuss the most important class of estimators: maximum likelihood estimators (MLE)
- Time permitting, we will also (briefly) discuss confidence intervals

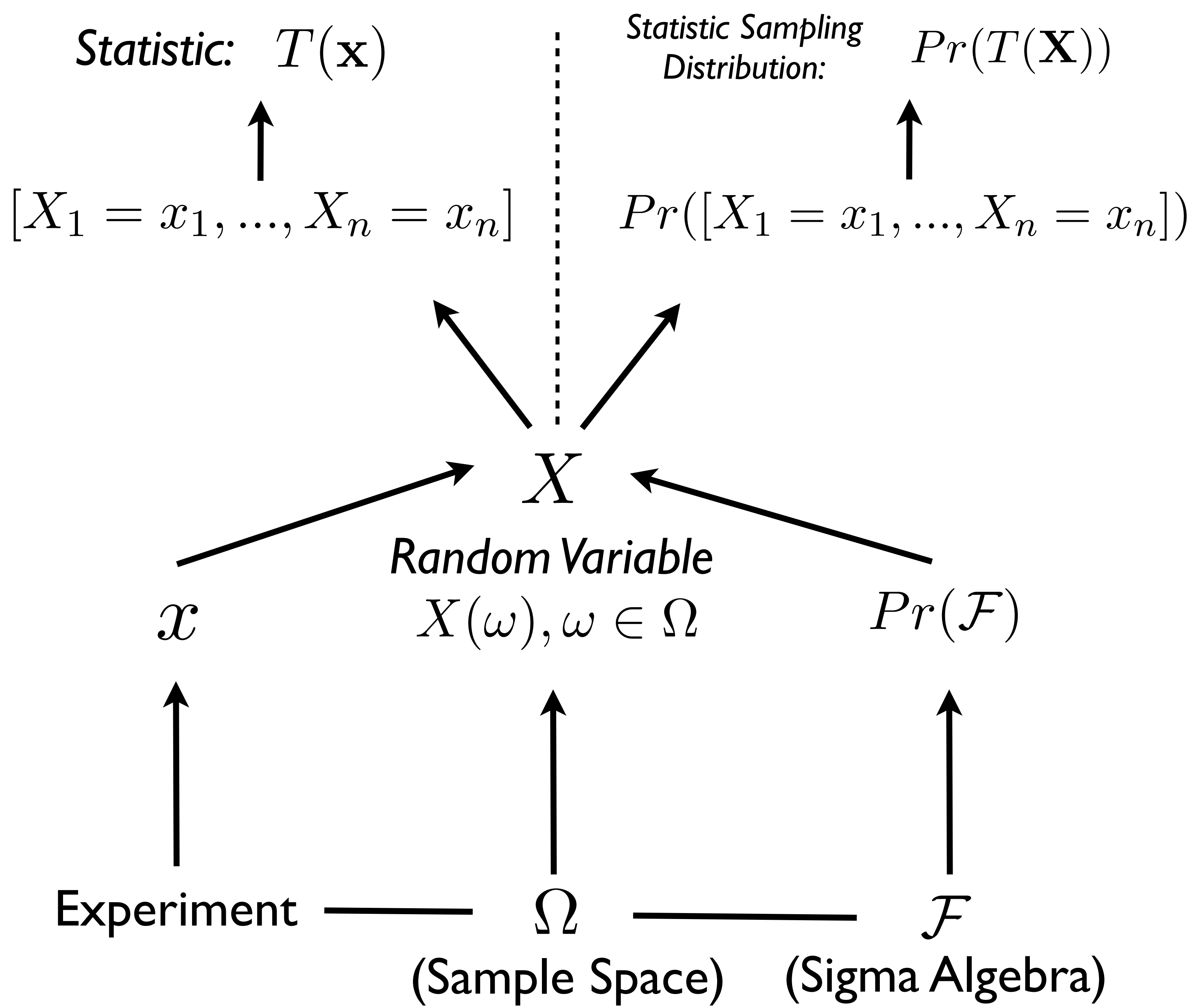
Conceptual Overview



Samples

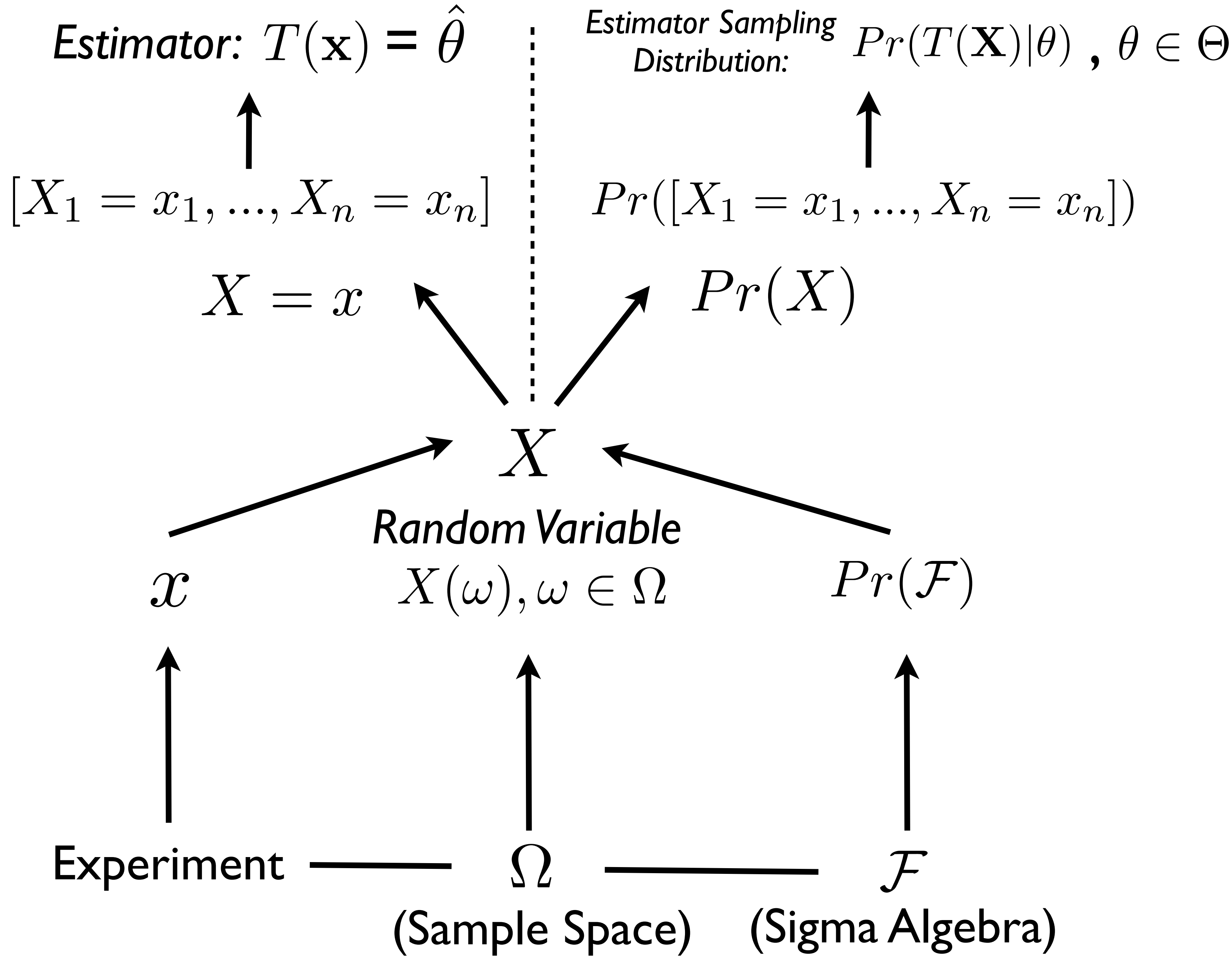

$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

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

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Sample Space** (Ω) - set comprising all possible outcomes associated with an experiment
- **Sigma Algebra** or **Sigma Field** ($\mathcal{F}$) - a collection of events (subsets) of the sample space of interest
- **Probability Measure (=Function)** - maps a Sigma Algebra of a sample to a subset of the reals
- **Random Variable** - (measurable) function on a sample space
- **Probability Mass Function / Cumulative Mass Function (pmf / cmf)** - function that describes the probability distribution of a discrete random variable
- **Probability Density Function / Cumulative Density Function (pdf / cdf)** - function that describes the probability distribution of a continuous random variable
- **Probability Distribution Function / Cumulative Distrbution Function (pdf / cdf)** - function that describes the probability distribution of a discrete OR continuous random variable

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

- **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 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

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:

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

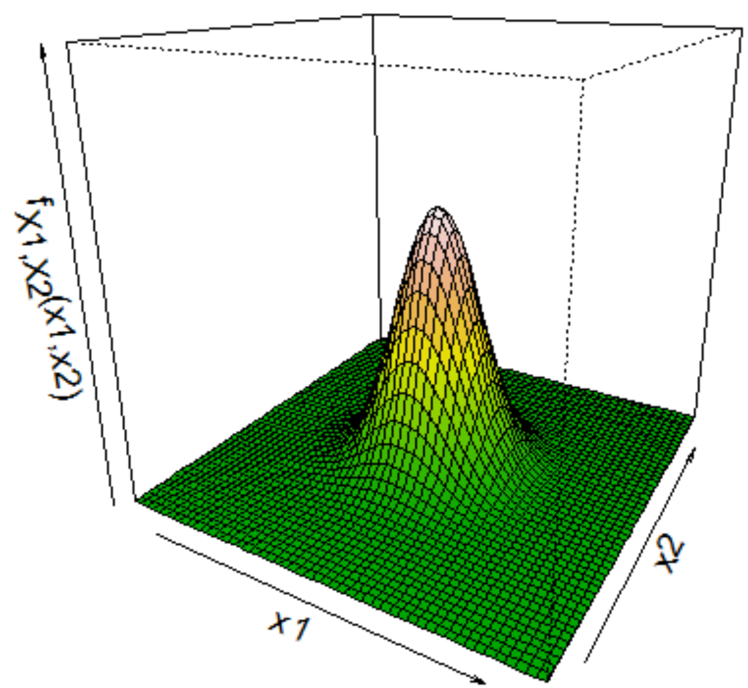
- We also generally assume that each observation is independent of all other observations:

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

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

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: 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 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 (!!)

Review: 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?

Review: Estimators I

- **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: Estimators II

- 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?)

Estimator example 1

- 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?)

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

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. (!!)

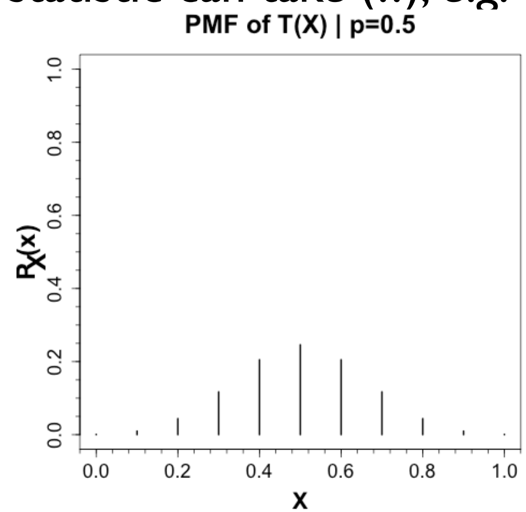
$$[X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}] \sim p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} ... p^{x_{10}}(1-p)^{1-x_{10}}$$

Estimator example II

- Define a statistic *T*(**X**)

$$T(\mathbf{X} = \mathbf{x}) = T(\mathbf{X}) = \bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$$

- Note the values the statistic can take (!!), e.g. with true *p*=0.5



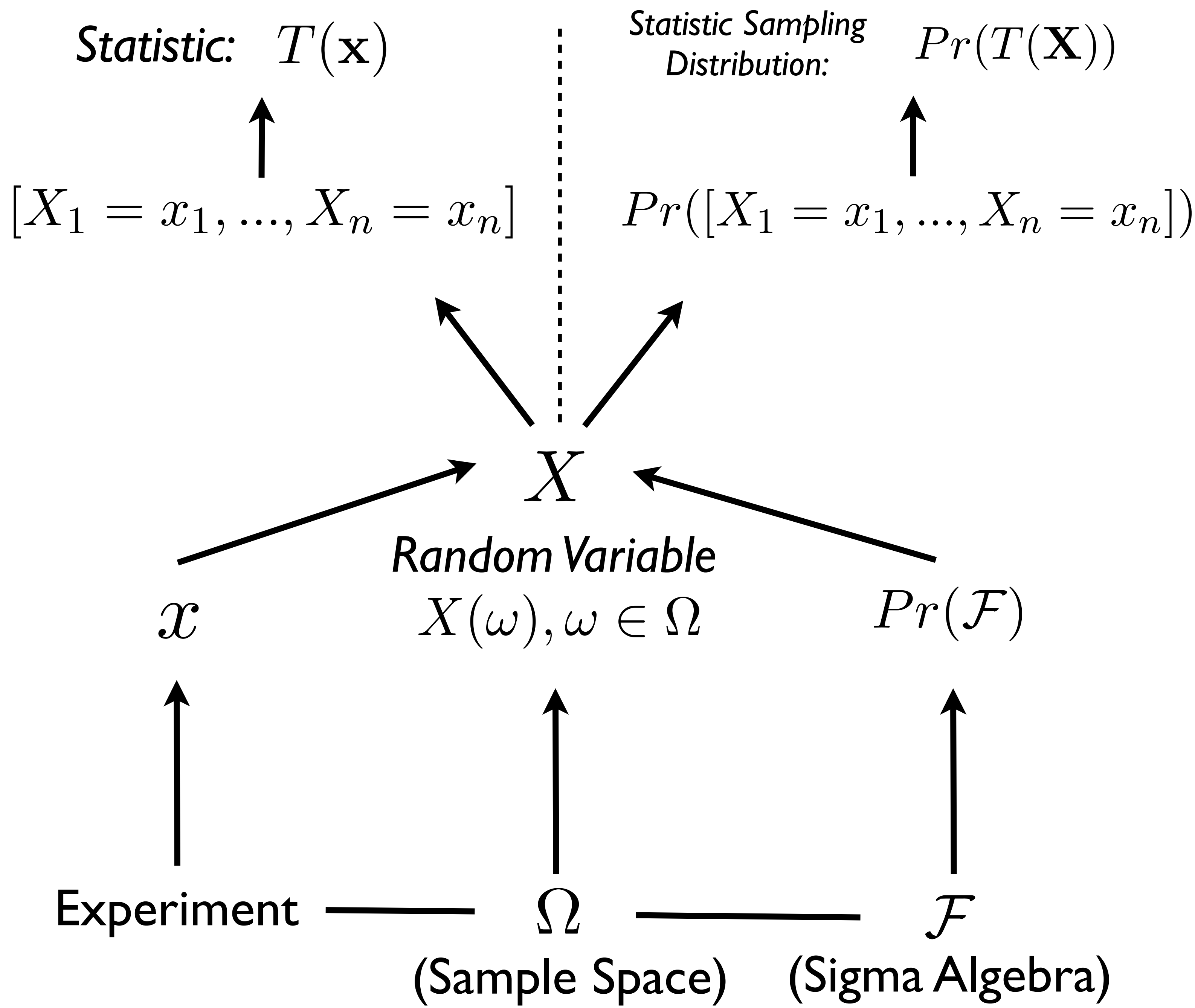
- Side note: we can write the sampling distribution (pmf) of the statistic as

$$Pr(T(\mathbf{X})) \sim \binom{n}{nT(\mathbf{X})} p^{nT(\mathbf{X})}(1-p)^{n-nT(\mathbf{X})}$$

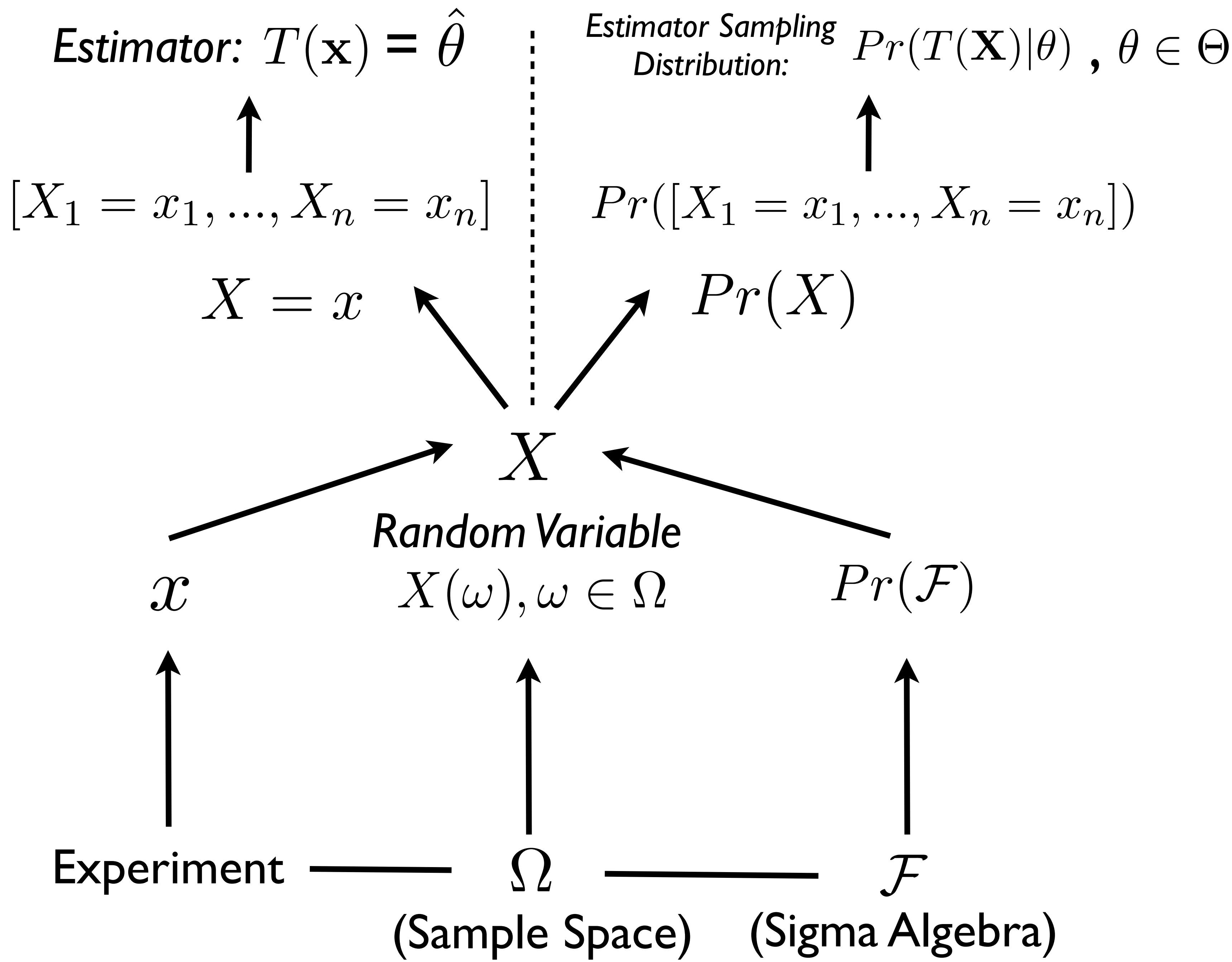
- Remember for our sample, the value of our statistic for our observed sample (!!) would equal 0.5 (will it always?)

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

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)

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

Likelihood I

- To introduce MLE's we first need the concept of *likelihood*
- Recall that a probability distribution (of a r.v. or for our purposes now, a statistic) has fixed constants in the formula called *parameters*
- For example, for a normally distributed random variable

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

- However, we could turn this around and fix the sample and let the parameters vary (this is a likelihood!)
- For example, say we have a sample *n*=1, where x=0.2 then the likelihood is (if we just set $\sigma^2 = 1$ for explanatory purposes):

$$L(\mu|\mathbf{x} = 0.2) = \frac{1}{\sqrt{2\pi}} e^{-(0.2-\mu)^2}$$

Likelihood II

- **Likelihood** - a function with the form of a probability function which we consider to be a function of the parameters θ for a fixed the sample $[\mathbf{X} = \mathbf{x}]$
- The form of a likelihood is therefore the sampling distribution (the probability distribution!) of the i.i.d sample but there are (at least) three major differences:
  - We have parameter values as input and the sample *we have observed* as a parameter
  - The likelihood function does not operate as a probability function (they can violate the axioms of probability)
  - For continuous cases, we can interpret the likelihood of a parameter (or combination of parameters) as the likelihood of the point

Likelihood III

- Again, Likelihood has the form of a probability function which we consider to be a function of the parameters NOT the sample
- Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)
- They have the appealing property that for an i.i.d. sample

$$L(\theta|x_1, x_2, ..., x_n) = L(\theta|x_1)L(\theta|x_2)...L(\theta|x_n)$$

- They have other appealing properties, including they are sufficient statistics, the invariance principal, etc.

Normal model example I

- As an example, for our heights experiment / identity random variable, the (marginal) probability of a single observation in our sample is x_i is:

$$Pr(X_i = x_i|\mu, \sigma^2) = f_{X_i}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- The joint probability distribution of the entire sample of n observations is a multivariate (n-variate) normal distribution

- Note that for an i.i.d. sample, we may use the property of independence

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

to write pdf of this entire sample as follow:

$$P(\mathbf{X} = \mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- The likelihood is therefore:

$$L(\mu, \sigma^2|\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Normal model example II

- Let's consider a sample of size n=10 generated under a standard normal, i.e.

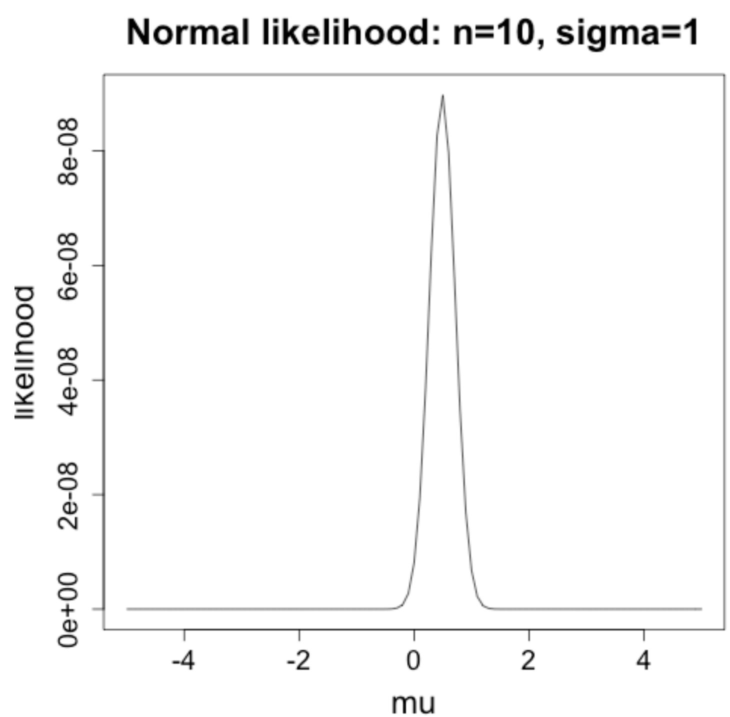
$$X_i \sim N(\mu = 0, \sigma^2 = 1)$$

```
[1] -1.0013985  1.0968952  0.4398448  0.7402079  1.5576818 -0.7619734  0.6158720  0.2738087  0.2182059  1.7288007
```

- So what does the likelihood for this sample "look" like? It is actually a 3-D plot where the x and y axes are μ and σ^2 and the z-axis is the likelihood:

$$L(\mu, \sigma^2 | \mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

- Since this makes it tough to see what is going on, let's set just look at the marginal likelihood for $\sigma^2 = 1$ when using the sample above:



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:

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

Getting to the MLE

- To use a likelihood function to extract the MLE, we have to find the maximum of the likelihood function $L(\theta|\mathbf{x})$ for our observed sample
- To do this, we take the derivative of the likelihood function and set it equal to zero (why?)
- Note that in practice, before we take the derivative and set the function equal to zero, we often transform the likelihood by the natural log (*ln*) to produce the log-likelihood:

$$l(\theta|\mathbf{x}) = ln[L(\theta|\mathbf{x})]$$

- We do this because the likelihood and the log-likelihood *have the same maximum* and because it is often easier to work with the log-likelihood
- Also note that the domain of the natural log function is limited to $[0, \infty)$ but likelihoods are never negative (consider the structure of probability!)

MLE under a normal model 1

- Recall that the likelihood for a sample of size *n* generated under a normal model has the following likelihood

$$L(\mu, \sigma^2 | \mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

- By remembering the properties of *ln*, we can derive the log-likelihood for this model

$$l(\mu, \sigma^2 | \mathbf{X} = \mathbf{x})) = -nln(\sigma) - \frac{n}{2} ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i}^{n} (x_i - \mu)^2$$

1. $ln\frac{1}{a} = -ln(a)$
2. $ln(a^2) = 2ln(a)$
3. $ln(ab) = ln(a) + ln(b)$
4. $ln(e^a) = a$
5. $e^a e^b = e^{a+b}$

- To obtain the maximum of this function with respect to μ we can then take the partial (!!) derivative with respect to μ and set this equal to zero, then solve (this is the MLE!):

$$\frac{\partial l(\theta | \mathbf{X} = \mathbf{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i}^{n} (x_i - \mu) = 0$$

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

MLE under a normal model II

- How about the σ^2? Use the same approach:

$$l(\mu, \sigma^2 | \mathbf{X} = \mathbf{x})) = -nln(\sigma) - \frac{n}{2} ln(2\pi) - \frac{1}{2\sigma^2} \sum_i^n (x_i - \mu)^2$$

$$\frac{\partial l(\theta | \mathbf{X} = \mathbf{x})}{\partial \sigma^2} = 0$$

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

- This equation will give us the maximum of the log-likelihood with respect to this parameter
- Will this produce the true value of σ^2 (!?)

A discrete example I

- As an example, for our coin flip / number of tails random variable
- The probability distribution of one sample is:

$$Pr(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

- The joint probability distribution of an i.i.d sample of size n is is an n-variate Bernoulli

$$Pr(\mathbf{x}|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- A TRICK (!!): it turns out that we can get the same MLE of p for this model by considering x = *total number of tails in the entire sample*:

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

- Such that we can consider the following likelihood:

$$L(p|\mathbf{X} = \mathbf{x}) = \binom{n}{x} p^{x}(1-p)^{n-x}$$

A discrete example II

- To find the MLE, we will use the same approach by taking the log-likelihood:

$$L(p|\mathbf{X} = \mathbf{x}) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$l(p|\mathbf{X} = \mathbf{x}) = ln\binom{n}{x} + xln(p) + (n-x)ln(1-p)$$

- taking the first derivative set to zero, then solve (again *x*=number tails!)

$$\frac{\partial l(p|\mathbf{X} = \mathbf{x})}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p}$$

$$MLE(\hat{p}) = \frac{x}{n}$$

- Question: in general, how do we know this is a maximum?

- We can check by looking at the second derivative and making sure that it is always negative (why?):

$$\frac{\partial^2 l(p|\mathbf{X} = \mathbf{x})}{\partial p^2} = -\frac{x}{p^2} + \frac{x-n}{(1-p)^2}$$

Last general comments (for now) on maximum likelihood estimators (MLE)

- In general, *maximum likelihood estimators* (MLE) are at the core of most standard "parametric" estimation and hypothesis testing (stay tuned!) that you will do in basic statistical analysis
- Both likelihood and MLE's have many useful theoretical and practical properties (i.e. no surprise they play a central role) although we will not have time to discuss them in detail in this course (e.g. likelihood has strong connections to the concept of sufficiency, likelihood principal, etc., MLE have nice properties as estimators, ways of obtaining the MLE, etc.)
- Again, for this course, the critical point to keep in mind is that when you calculate an MLE, you are just calculating a statistic (estimator!)

That's it for today

- Next lecture, we will (briefly) discuss confidence intervals and begin our discussion of hypothesis testing!