Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 10: Introduction to Hypothesis Testing II

Jason Mezey

Feb 23, 2023 (Th) 8:05-9:20

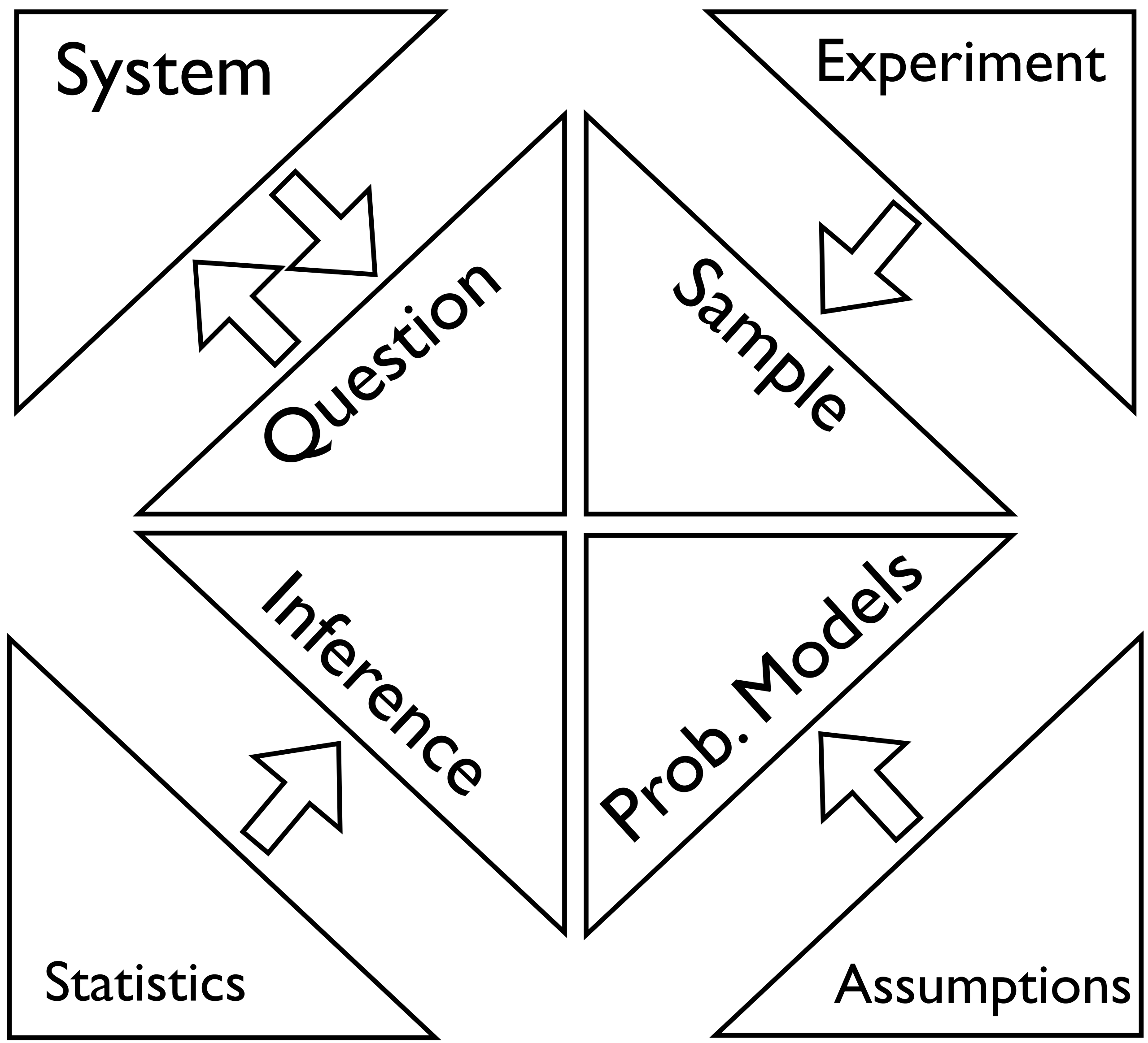
Announcements

- There will be NO CLASS THIS TUES (Feb 28 = Cornell, Ithaca winter break)
- Homework #3 will be assigned this evening (Feb 23)
- We will have office hours next week but day and time TBD (I will send a message about this next week)

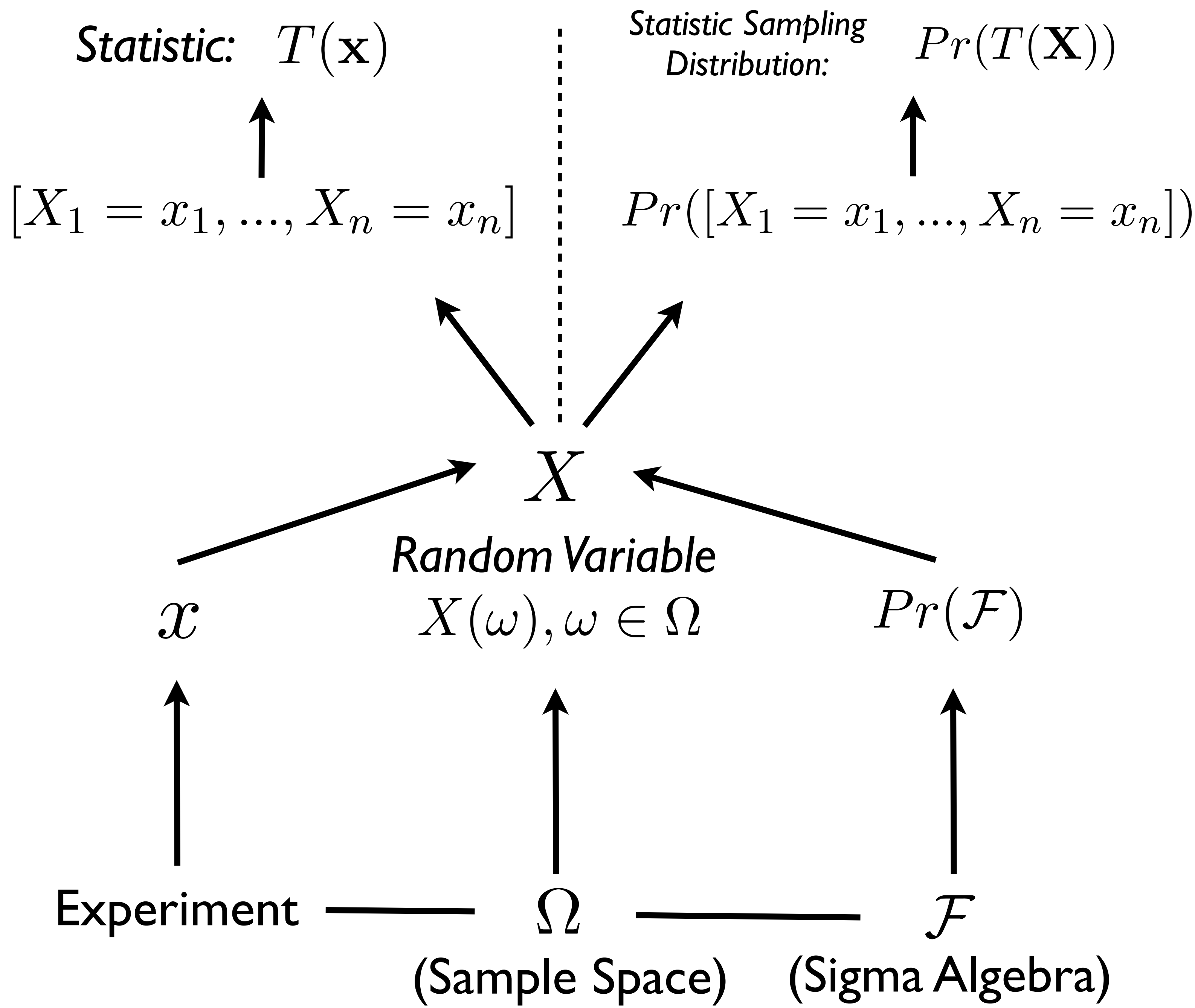
Summary of lecture 10: Introduction to Hypothesis Testing

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

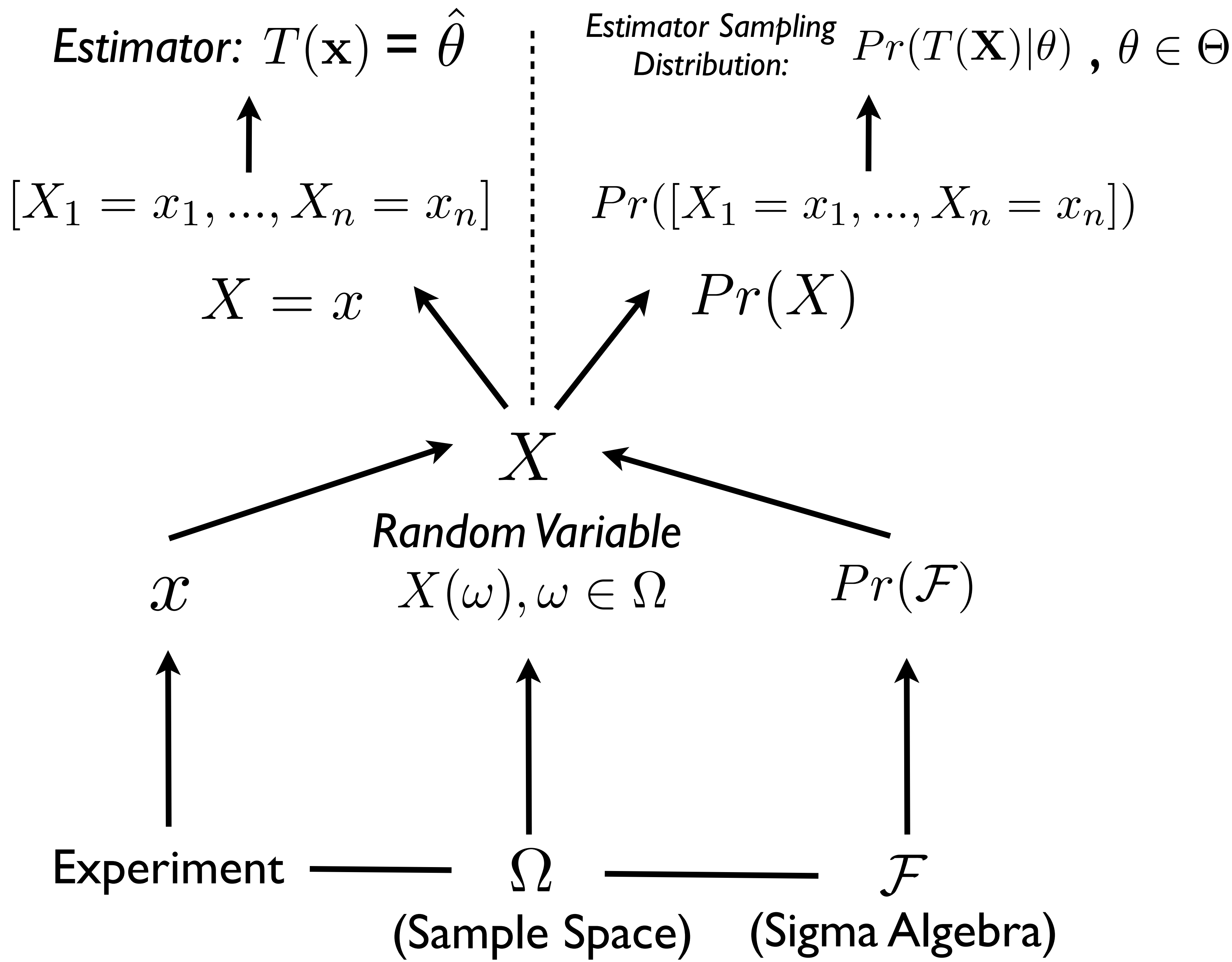
Conceptual Overview



Statistics

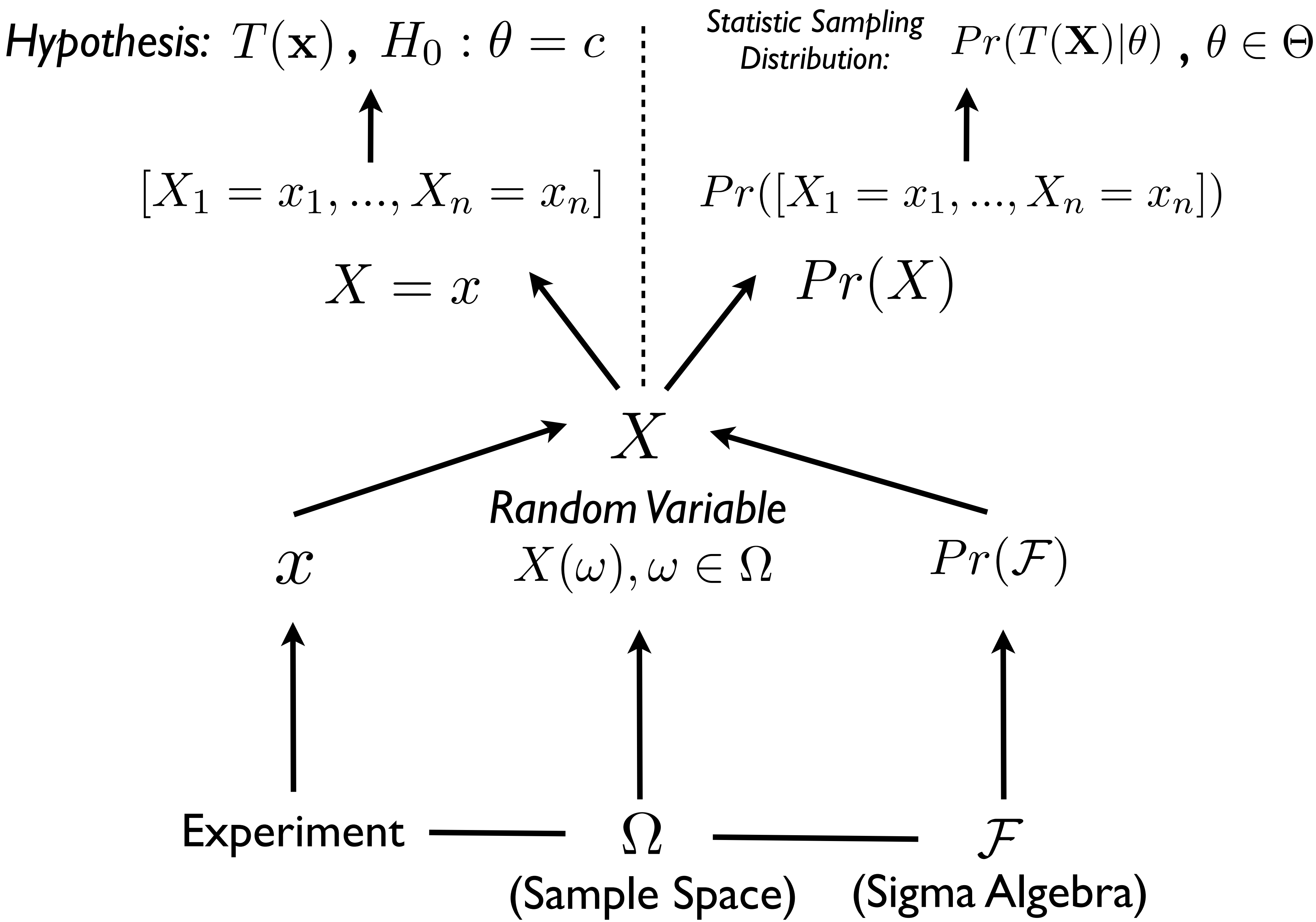

Statistic: $T(\mathbf{x})$
Statistic Sampling Distribution: $Pr(T(\mathbf{X}))$
$[X_1 = x_1, ..., X_n = x_n]$
$Pr([X_1 = x_1, ..., X_n = x_n])$
X
Random Variable
x
$X(\omega), \omega \in \Omega$
$Pr(\mathcal{F})$
Experiment
Ω
(Sample Space)
$\mathcal{F}$
(Sigma Algebra)

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)

Hypothesis Tests



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:

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

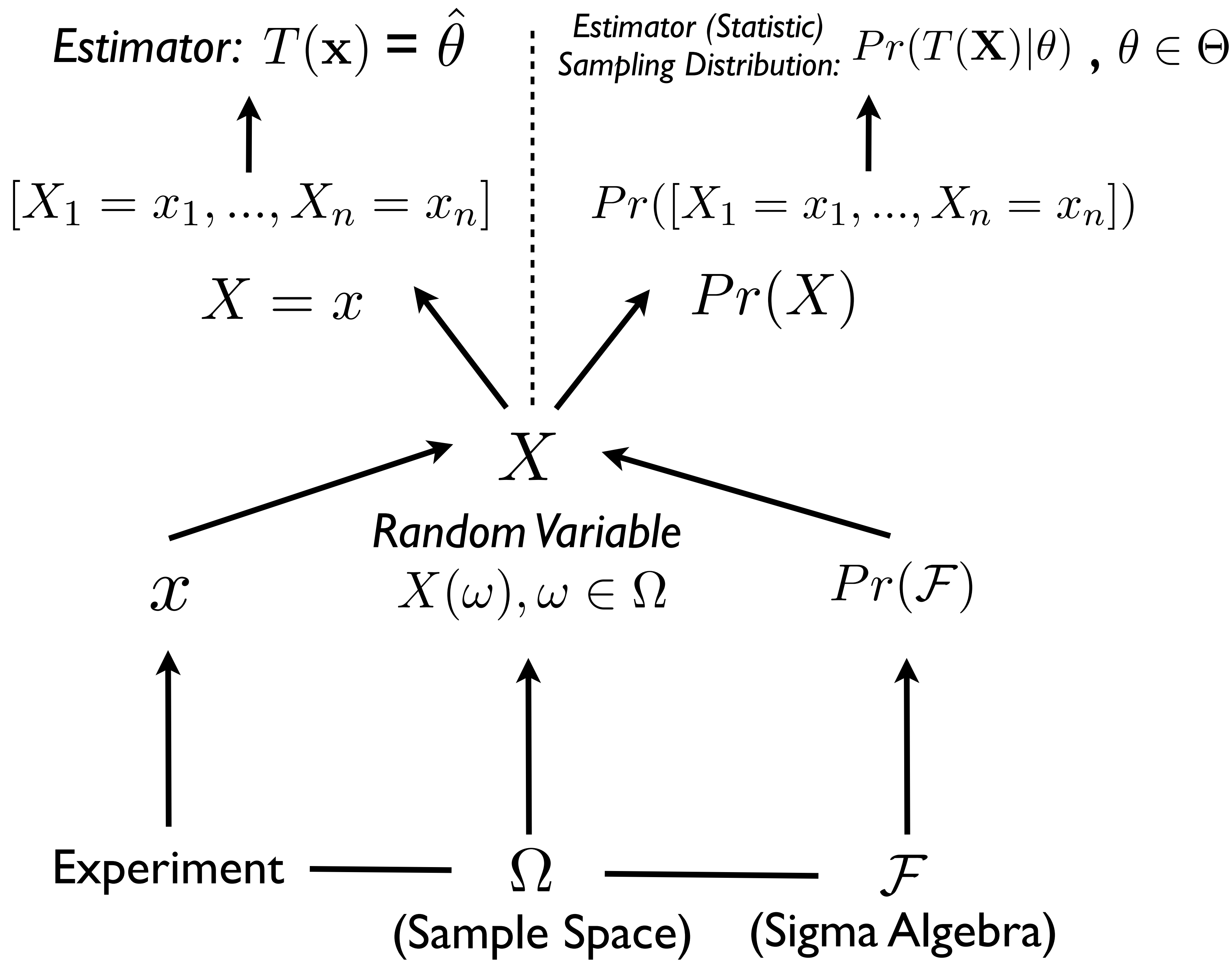
what is $\Pr(T(\mathbf{X}))$?

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

Estimation and Hypothesis Testing

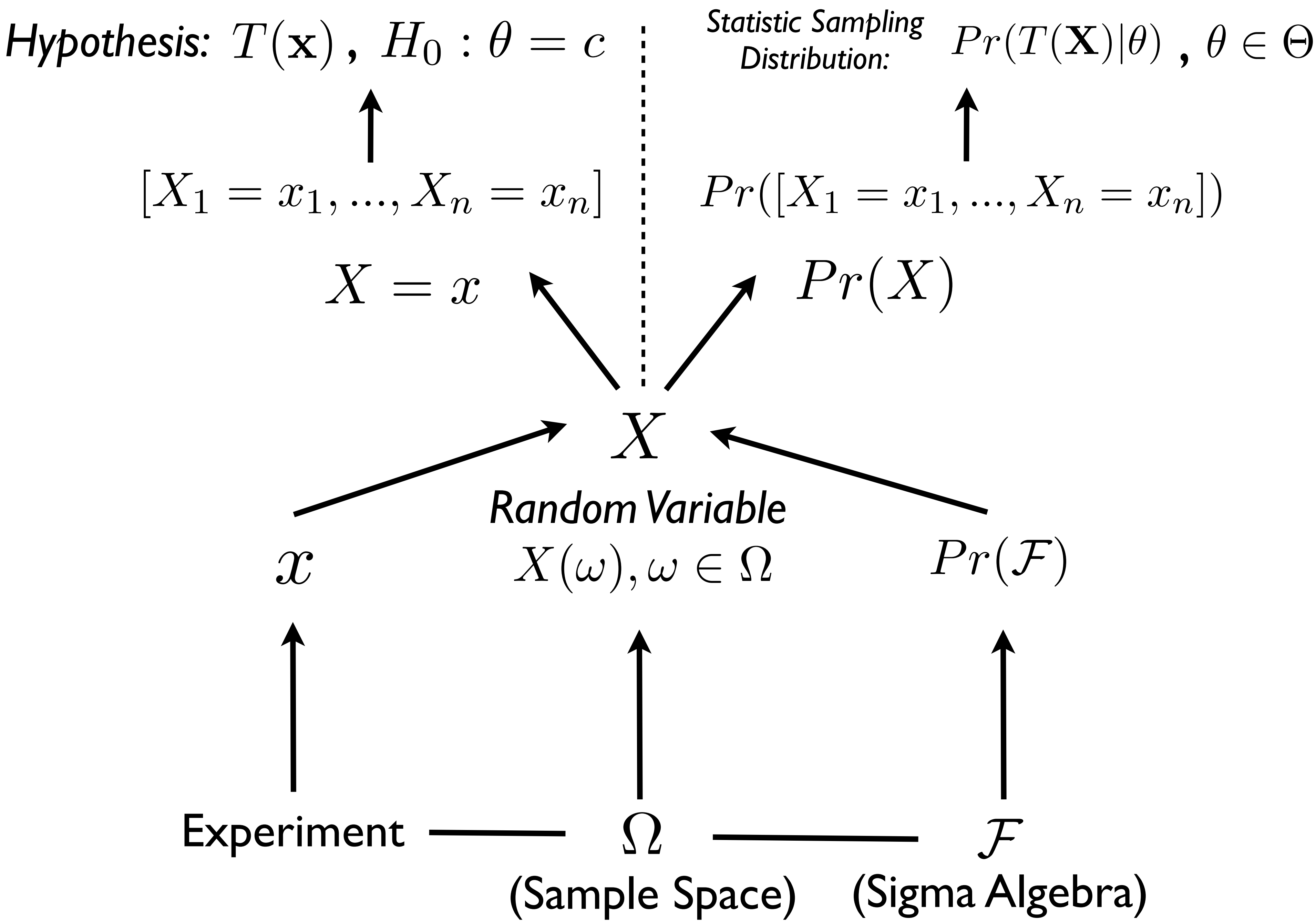
- Thus far we have been considering a “type” of inference, *estimation*, where we are interested in determining the actual value of a parameter
- We could ask another question, and consider whether the parameter is NOT a particular value
- This is another “type” of inference called *hypothesis testing*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



Review: 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

$$H_0 : \theta = c$$

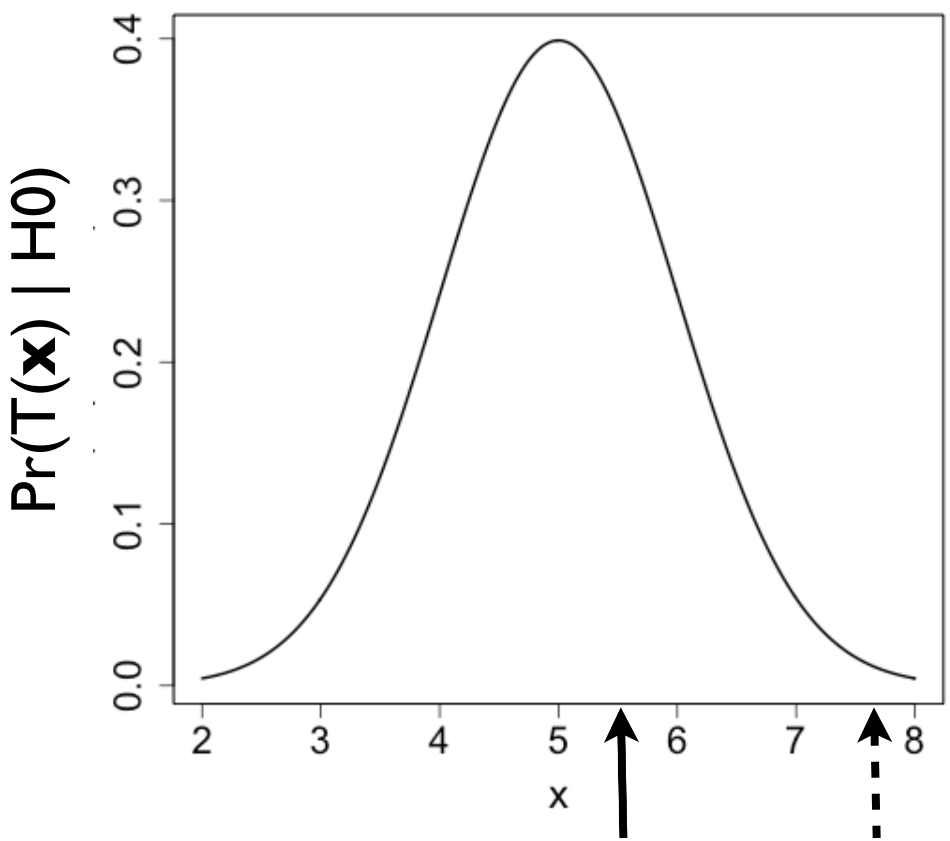
- For example, for our height experiment / identity random variable, we have $Pr(X|\theta) \sim N(\mu, \sigma^2)$ and we could consider the following null hypothesis:

$$H_0 : \mu = 0$$

Review: 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:



Hypothesis testing III

- Our goal in hypothesis testing is to use a sample to reach a conclusion about the null hypothesis
- To do this, just as in estimation, we will make use of a statistic (a function on the sample), where recall we know the sampling distribution (the probability distribution) of this statistic
- More specifically, we will consider the probability distribution of this statistic, assuming that the null hypothesis is true:

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

- Note that this means we have a probability distribution of the statistic given the null hypothesis!!
- We will use this distribution to construct a *p-value*

p-value 1

- We quantify our intuition as to whether we would have observed the value of our statistics given the null is true with a *p-value*
- **p-value** - the probability of obtaining a value of a statistic $T(\mathbf{x})$, *or more extreme*, conditional on H0 being true
- Formally, we can express this as follows:

$$pval = Pr(|T(\mathbf{x})| \geqslant t | H_0 : \theta = c)$$

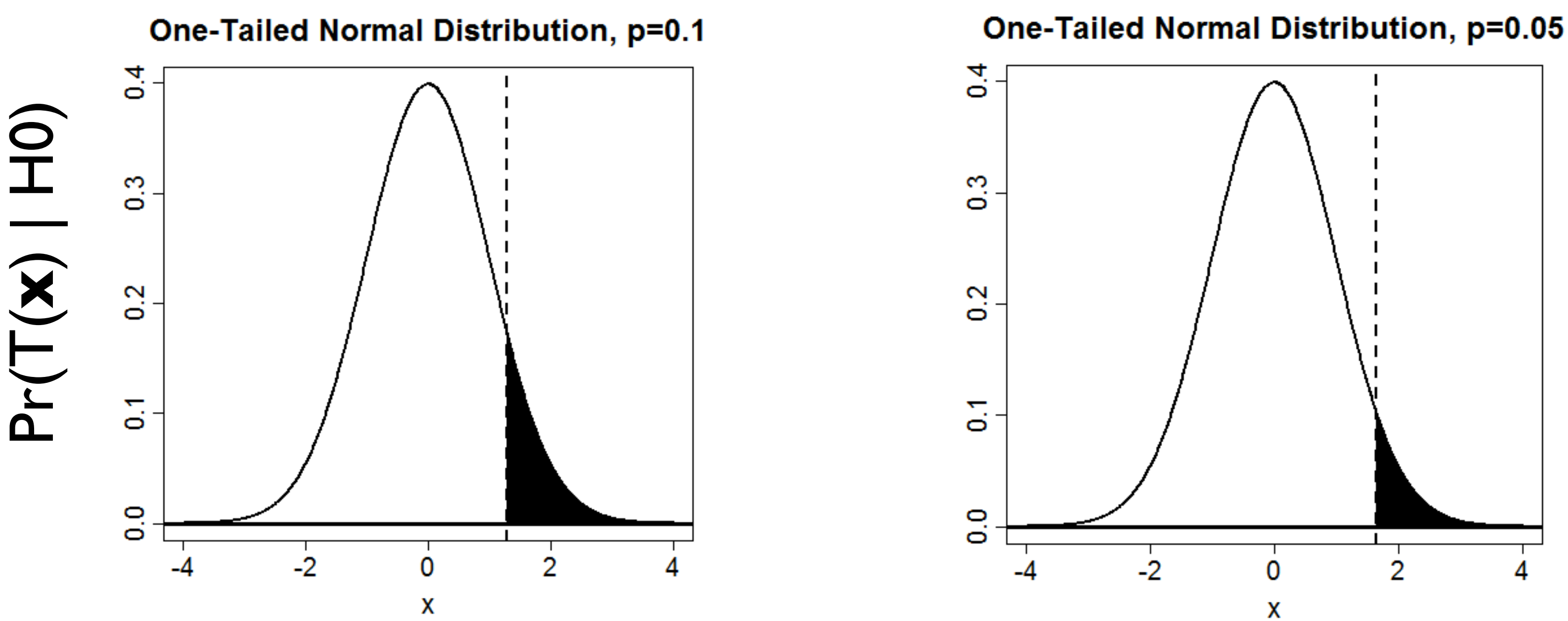
- Note that a p-value is a function on a statistic (!!) that takes the value of a statistic as input and produces a p-value as output in the range [0, 1]:

$$pval(T(x)) : T(x) \rightarrow [0, 1]$$

p-value II

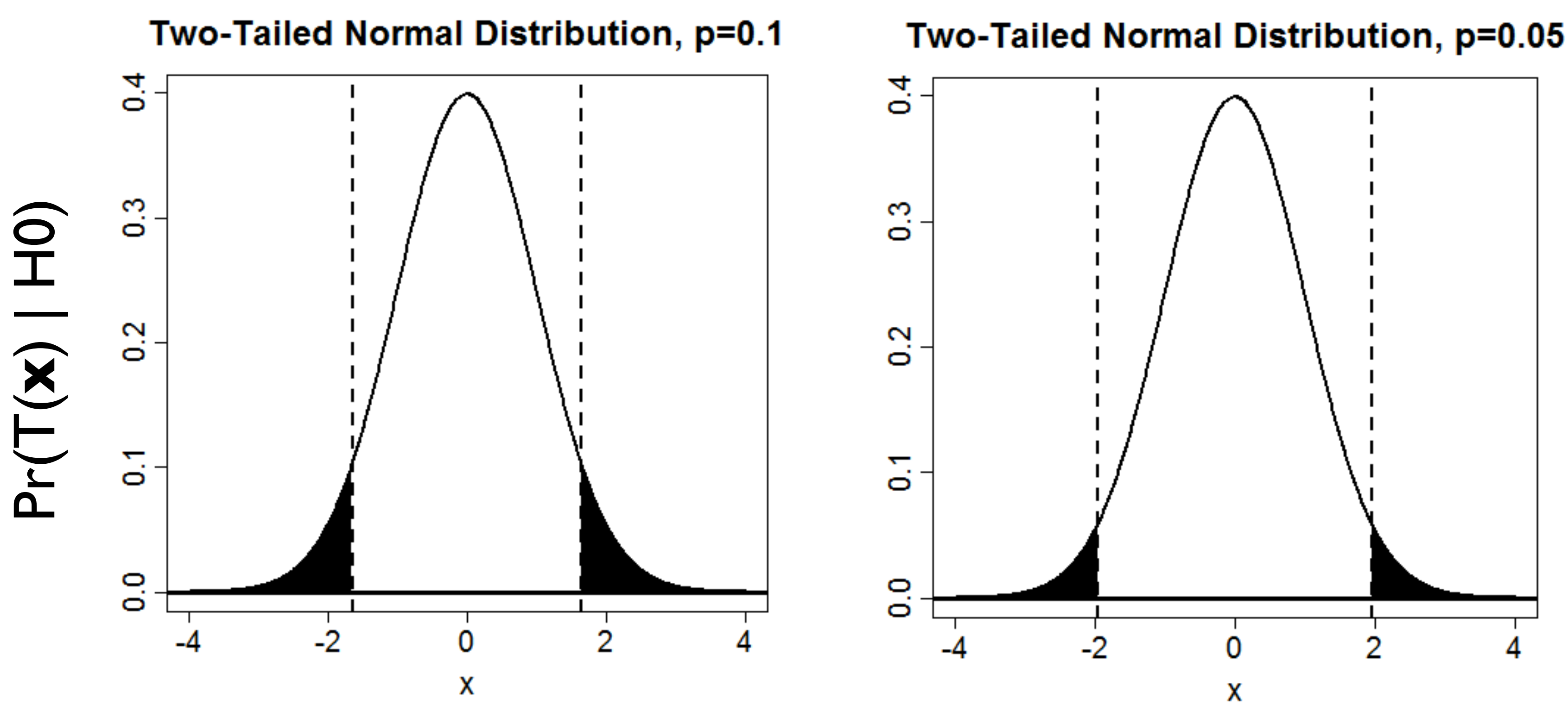
- As an intuitive example, let’s consider a continuous sample space experiment / identify r.v. / normal family / *n*=1 sample / identity statistic, i.e. $T(x) = x$

- Assume we know $\sigma^2 = 1$ (is this realistic?), let’s say we are interested in testing the null hypothesis $H_0 : \mu = 0$ and let’s say that we assume that if we are wrong the value of μ will be greater than zero (why?)



p-value III

- Same example: let's consider a continuous sample space experiment / identify r.v. / normal family / n=1 sample / identity statistic, i.e. T(X) = X / assume we know $\sigma^2 = 1$ / we test the null hypothesis $H_0 : \mu = 0$ and let's assume that if we are wrong the value of μ could be in either direction (again, why?)



p-value IV

- More technically a p-value is determined not just by the probability of the statistic given the null hypothesis is true, but also whether we are considering a "one-sided" or "two-sided" test

- For a one-sided test (towards positive values), the p-value is:

$$pval(T(\mathbf{x})) = \int_{T(\mathbf{x})}^{\infty} Pr(T(\mathbf{x})|\theta = c) dT(\mathbf{x})$$

$$pval(T(\mathbf{x})) = \sum_{T(\mathbf{x})}^{max(T(\mathbf{X}))} Pr(T(\mathbf{x})|\theta = c)$$

- For a two-sided test, the p-value is:

$$pval(T(\mathbf{x})) = \int_{-\infty}^{-|T(\mathbf{x}) - median(T(\mathbf{X})|} Pr(T(\mathbf{x})|\theta = c) dT(\mathbf{x}) + \int_{|T(\mathbf{x})| - median(T(\mathbf{X})|}^{\infty} Pr(T(\mathbf{x})|\theta = c) dT(\mathbf{x})$$

$$pval(T(\mathbf{x})) = \sum_{min(T(\mathbf{X}))}^{-|T(\mathbf{x}) - median(T(\mathbf{X})|} Pr(T(\mathbf{x})|\theta = c) + \sum_{|T(\mathbf{x}) - median(T(\mathbf{X})|}^{max(T(\mathbf{X}))} Pr(T(\mathbf{x})|\theta = c)$$

Hypothesis Testing IV

- To build a framework to answer a question about a parameter, 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

$$H_0 : \theta = c$$

- Once we have assumed a null hypothesis, we know the probability distribution of the statistic, assuming the null hypothesis is true:

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

- **p-value** - the probability of obtaining a value of a statistic *T*(**x**), *or more extreme*, conditional on H0 being true:

$$pval = Pr(|T(\mathbf{x})| \geqslant t|H_0 : \theta = c)$$

$$pval(T(x)) : T(x) \rightarrow [0, 1]$$

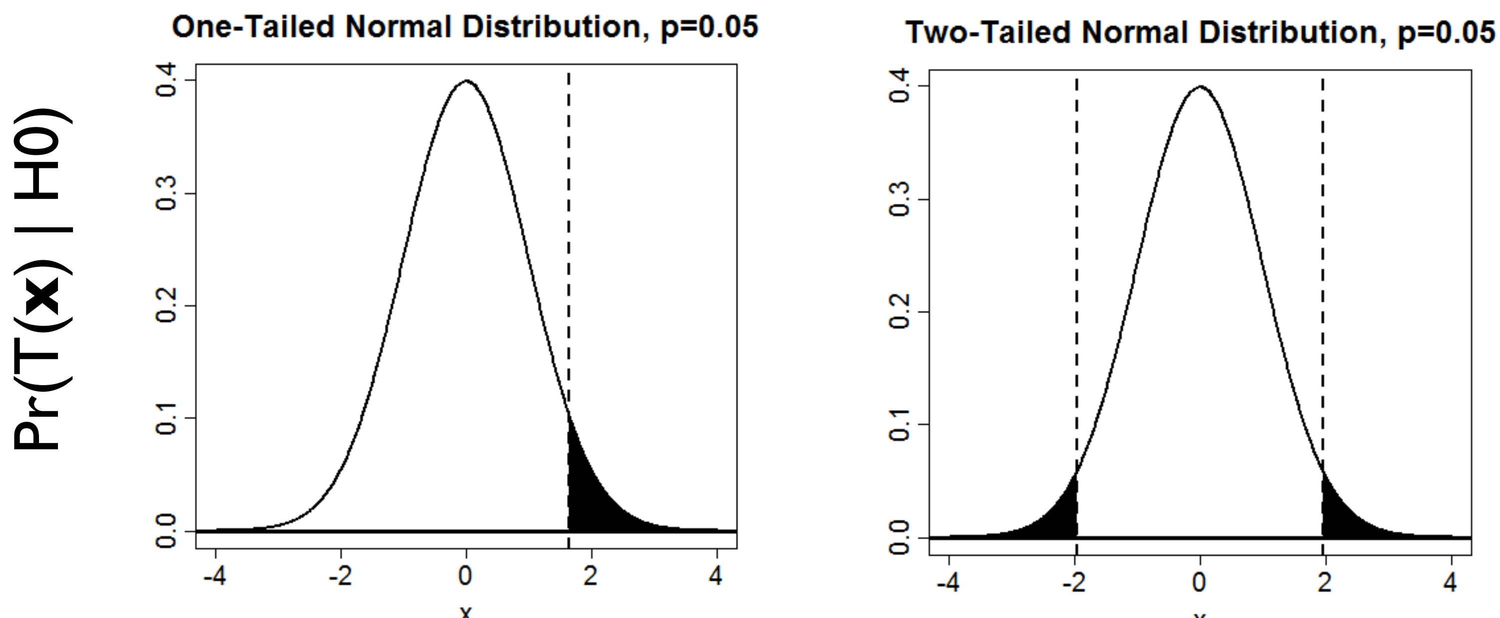
- *Note that a p-value is a function of a statistic* (!!)

Non-Intuitive Hypothesis Testing Concepts I

- We do not know what the true model is (=parameter values are) in a real case!
- We assess a null hypothesis that we define!
- We assess this null hypothesis by calculating a p-value which assumes that the null hypothesis is true!
- We assess this null hypothesis by calculating a p-value from a single sample!
- We make one of two decisions: cannot reject or reject!
  - We decide on the value p-value that allows us to decide
  - If we reject, we interpret this as strong evidence against the null hypothesis being correct but we do not know for sure!
  - If we cannot reject, we cannot say anything (i.e., we have no evidence that the null is wrong and we cannot say that the null is right)!

Hypothesis decisions I

- We use the p-value to make a decision about the null hypothesis
- Specifically, we use the p-value for our sample to decide whether we "accept" (or better stated: "cannot reject") the null hypothesis or "reject" the null hypothesis
- To do this, we use a value α such that if the p-value is below this value we "reject", if it is above we "cannot reject"
- Note that this value of α corresponds to a critical value ("threshold") of the test statistic c_α
- For example for a value $\alpha = 0.05$ we have the following for our previous examples:

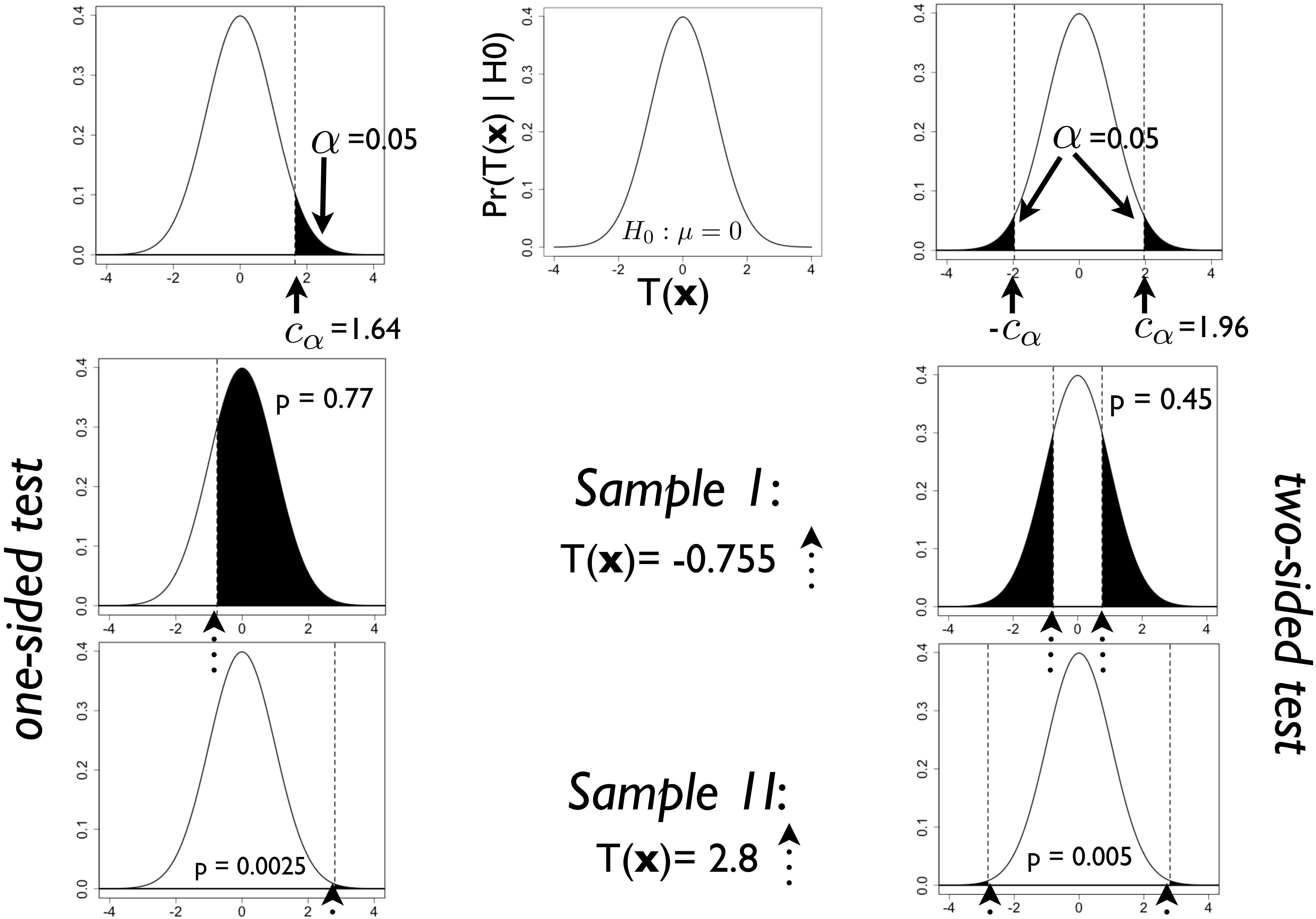


$$\alpha = \int_{c_\alpha}^{\infty} f_X(x)dx \qquad \alpha = \int_{-\infty}^{-c_\alpha} f_X(x)dx + \int_{c_\alpha}^{\infty} f_X(x)dx$$

Hypothesis decisions II

- Note that there are two possible outcomes of a hypothesis test: we reject or we cannot reject
- We never know for sure whether we are right (!!)
- If we cannot reject, this does not mean H0 is true (why? What if our p-value is 0.99?)
- The value α is called the type I error, the probability of incorrectly rejecting H0 when it is true
- The value $1 - \alpha$ is the probability of making a correct decision not to reject H0
- Note that we can control the level of type I error because we decide on the value of α

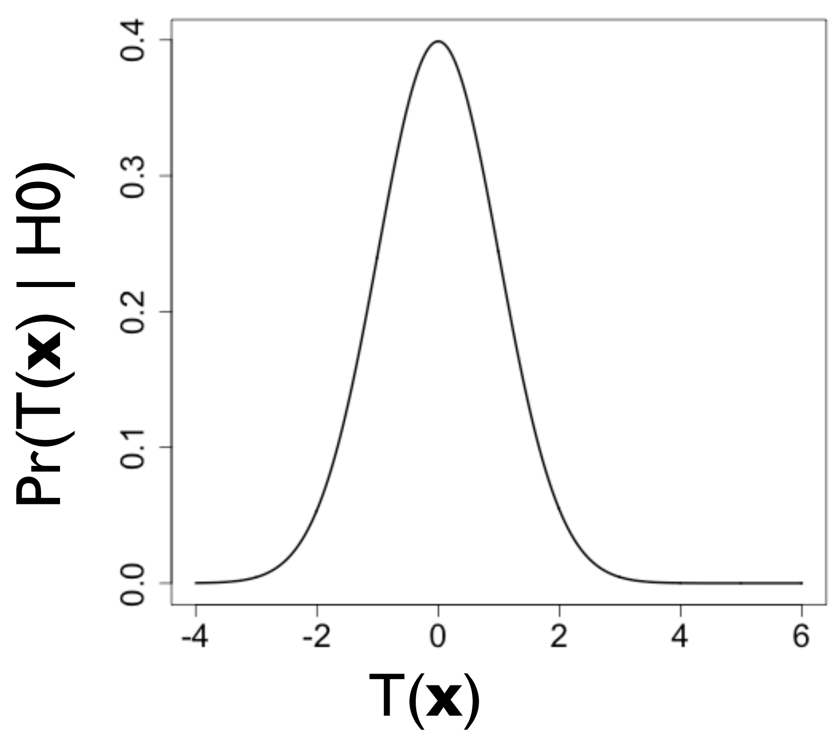
Assume H0 is correct (!): $\mu = 0$



Results of hypothesis decisions I: when H0 is correct (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

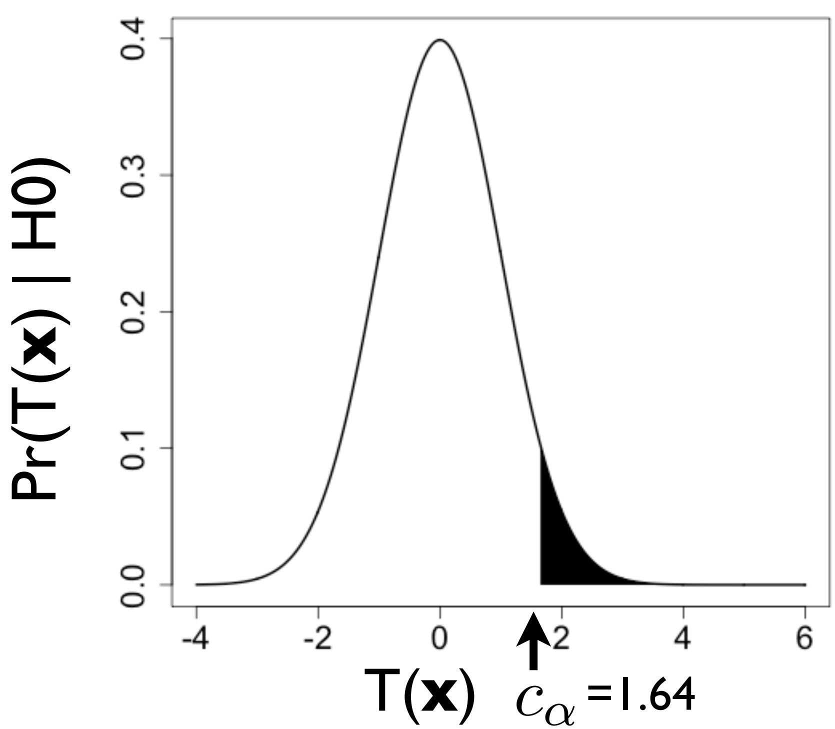
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



Results of hypothesis decisions I: when H0 is correct (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

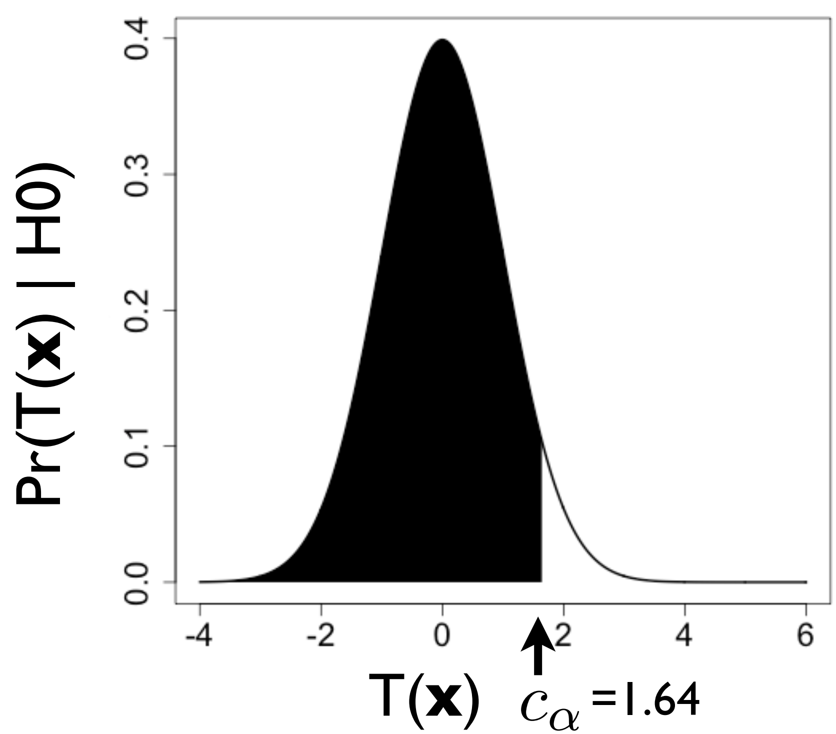
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	$\boxed{\alpha\text{, type I error}}$



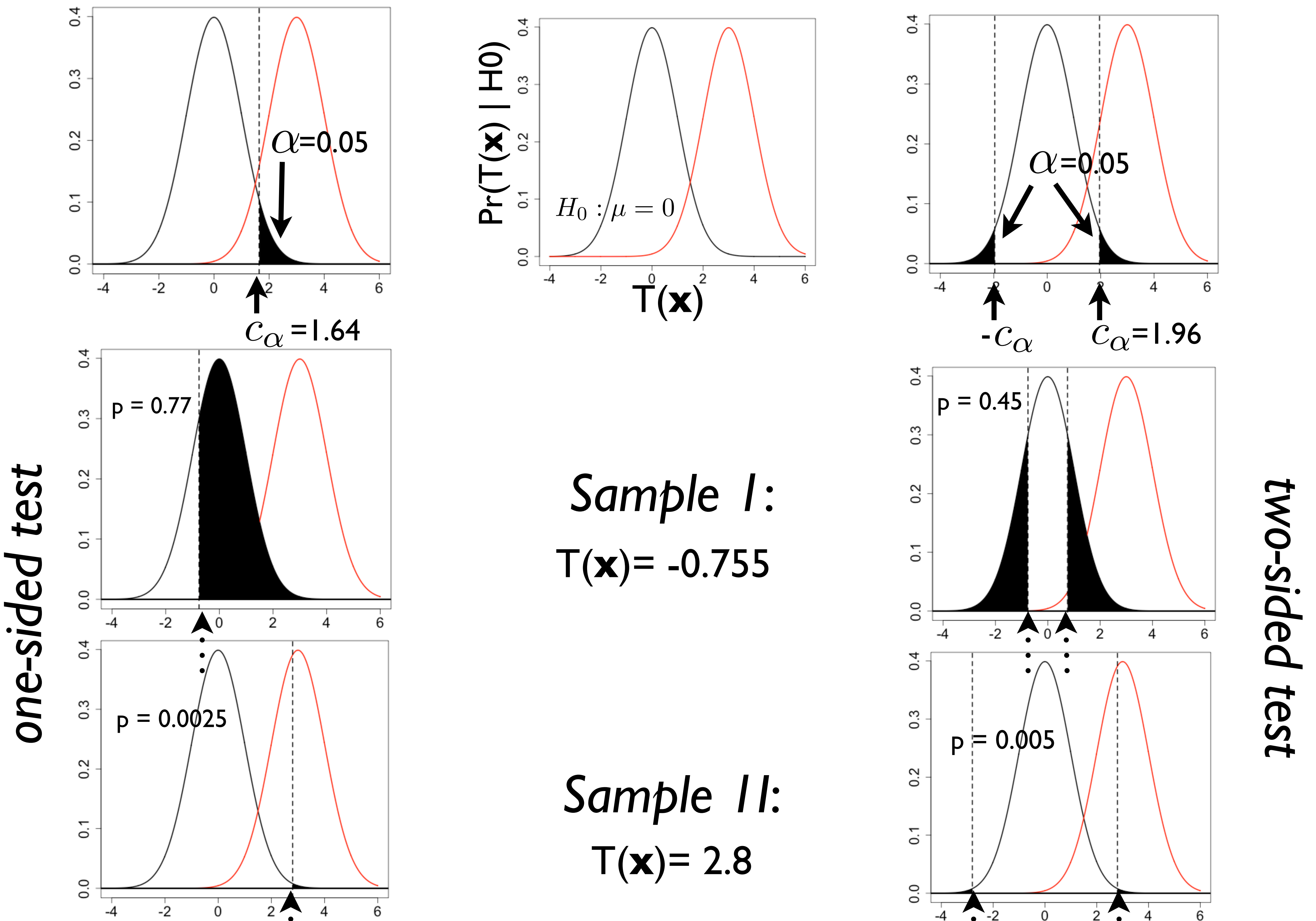
Results of hypothesis decisions I: when H0 is correct (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

	H_0 is true
cannot reject H_0	$\boxed{1\text{-}\alpha\text{, (correct)}}$
reject H_0	α, type I error



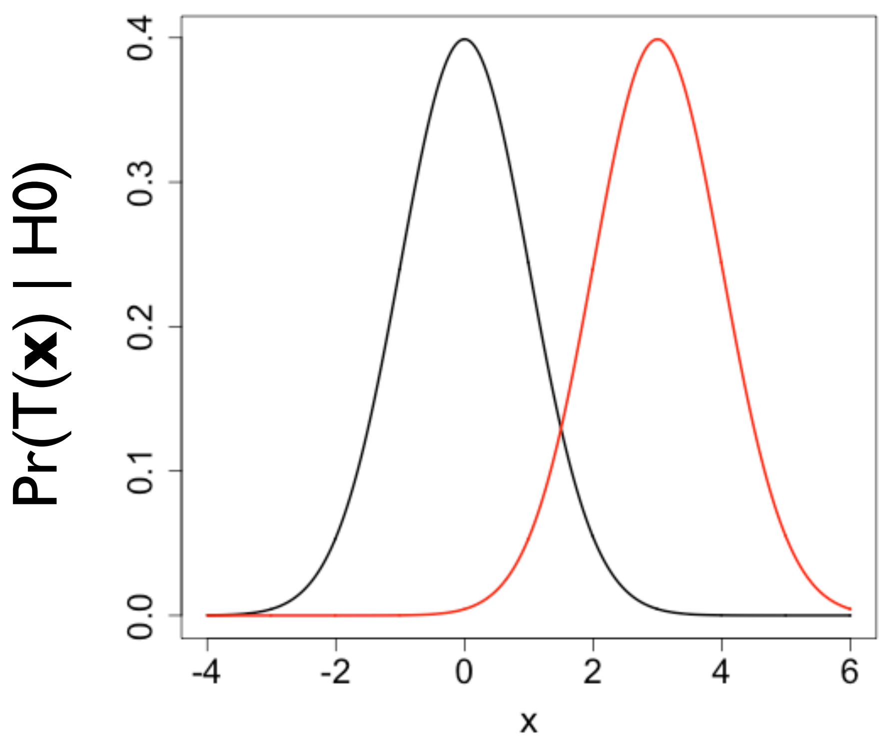
Assume H0 is wrong (!): $\mu = 3$



Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

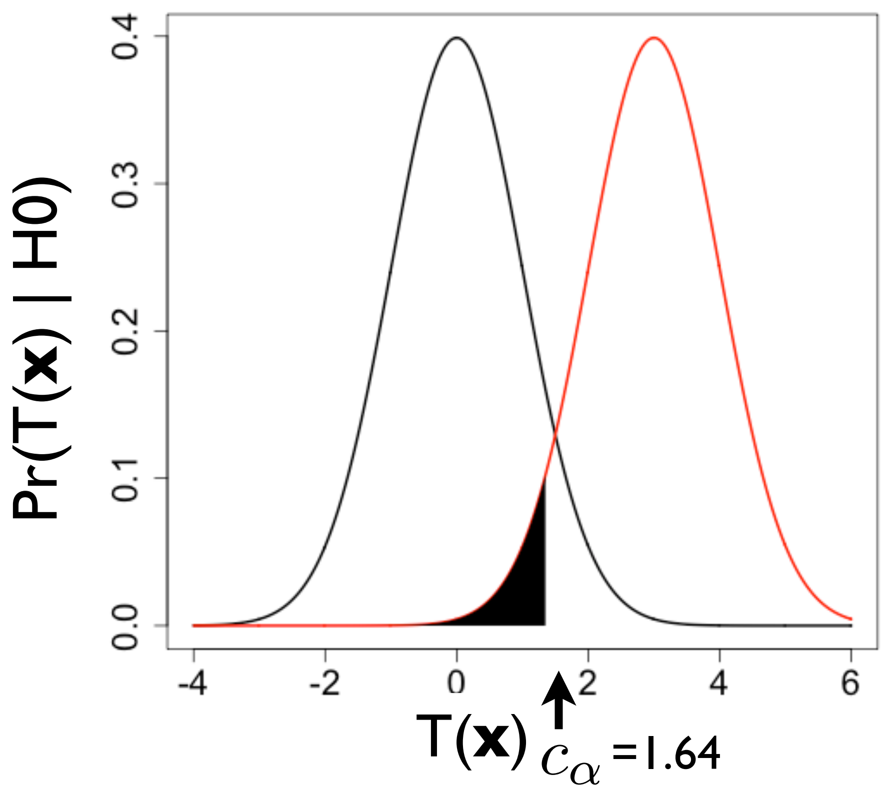
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

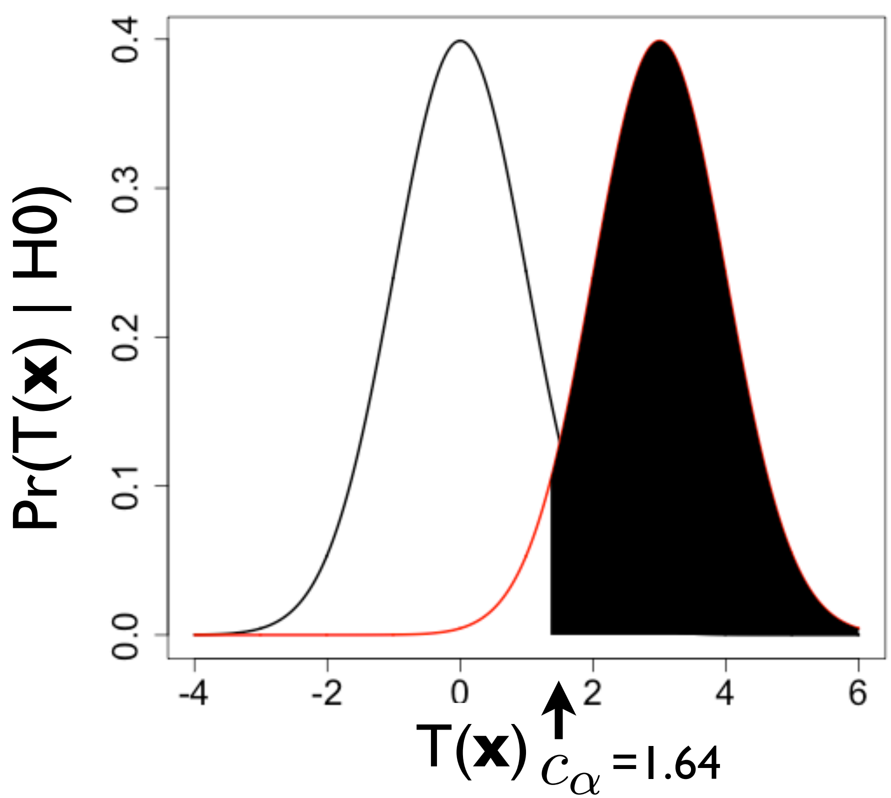
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	$\boxed{\beta\text{, type II error}}$
reject H_0	α, type I error	$1-\beta$, power (correct)



Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Technical definitions

- Technically, correct decision given H0 is true is (for one-sided, similar for two-sided):

$$1 - \alpha = \int_{-\infty}^{c_\alpha} Pr(T(\mathbf{x})|\theta = c)dT(\mathbf{x})$$

- Type I error (H0 is true) is (for one-sided):

$$\alpha = \int_{c_\alpha}^{\infty} Pr(T(\mathbf{x})|\theta = c)dT(\mathbf{x})$$

- Type II error given H0 is false is (for one-sided):

$$\beta = \int_{-\infty}^{c_\alpha} Pr(T(\mathbf{x})|\theta)dT(\mathbf{x})$$

- Power is (for one-sided):

$$1 - \beta = \int_{c_\alpha}^{\infty} Pr(T(\mathbf{x})|\theta)dT(\mathbf{x})$$

Important concepts

- REMEMBER (!!): there are two possible outcomes of a hypothesis test: we reject or we cannot reject
- We never know for sure whether we are right (!!)
- If we cannot reject, this does not mean H0 is true (why?)
- Note that we can control the level of type I error because we decide on the value of α

That's it for today

- Next lecture (Thurs, March 2), we will begin our discussion of quantitative genetics (and genomics)!