Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 11: Introduction to Genetic Modeling

Jason Mezey
March 2, 2023 (Th) 8:05-9:20

Announcements

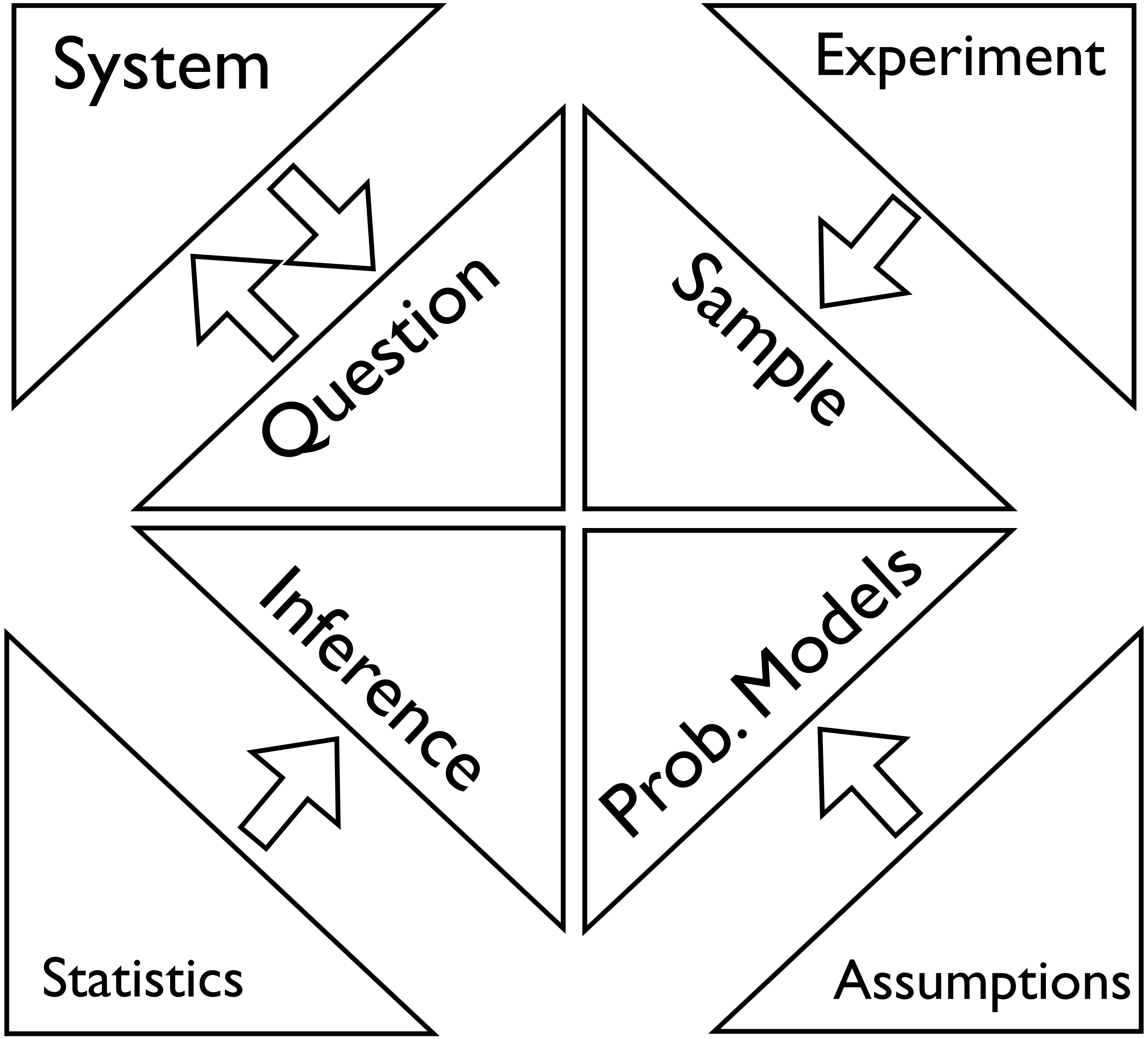
- Office hours this week will be TOMORROW (Fri, March 3) - zoom information will be set today
- Homework corrections (=I will post a correction!):
  - For instructions question 2 parts [f-j] in the "PLEASE NOTE THE FOLLOWING" section the equation: $\bar{X} = \sum_{i=1}^{n} X_i$ should be $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (!!)
  - 2g "under the null hypothesis in part [a]" - part [a] should be part [f]!
- Homework hint (!!) for (part of) 2f:

code to answer for $H_A > 0$: 'qnorm(0.95, 0, 1 / sqrt(20), lower.tail = TRUE)'

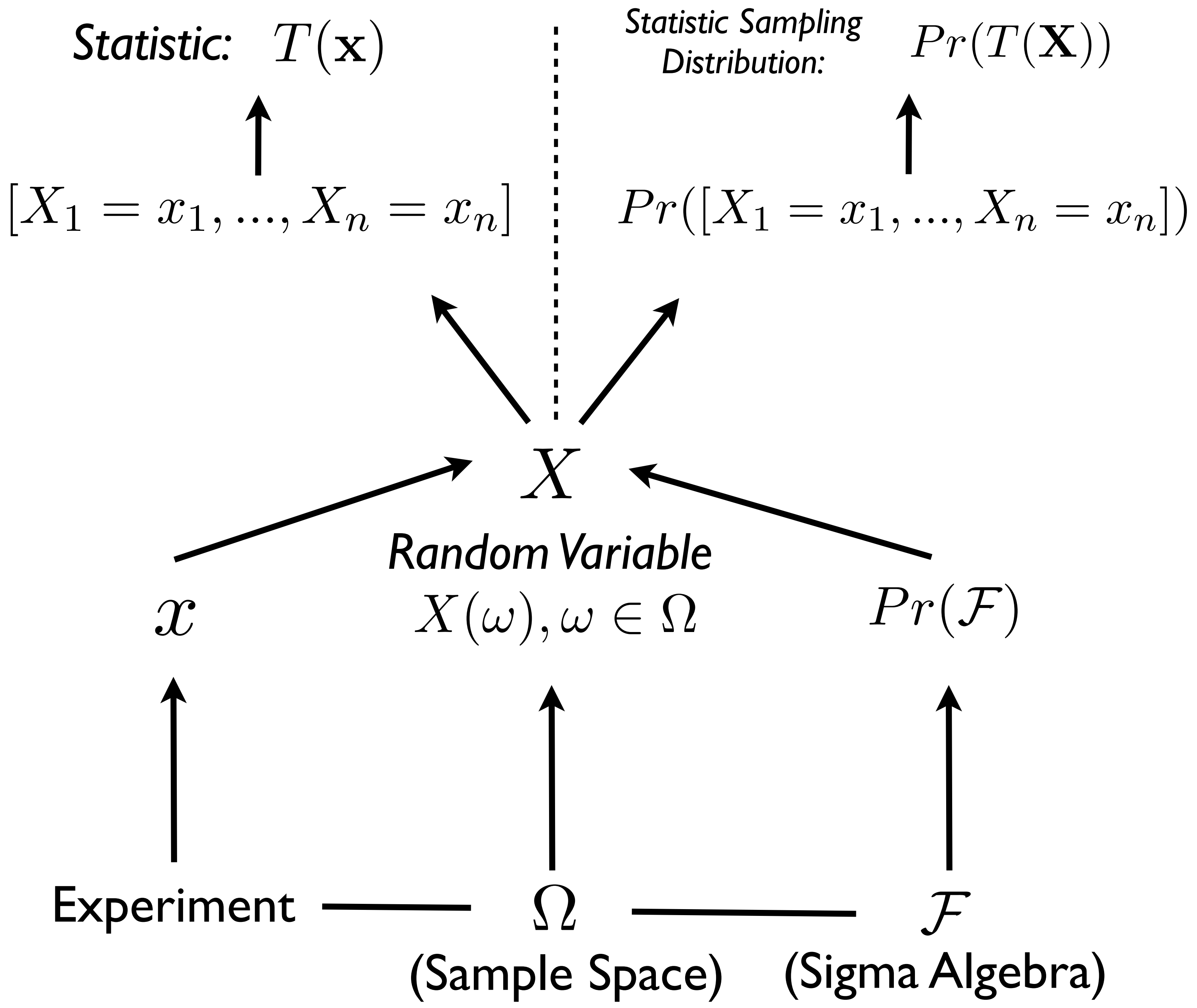
Summary of lecture 11: Introduction to Hypothesis Testing

- Last lecture, we almost completed our (general) discussion of hypothesis testing (!!)
- Today, we will complete the discussion of hypothesis testing and begin discussing genetic modeling!

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

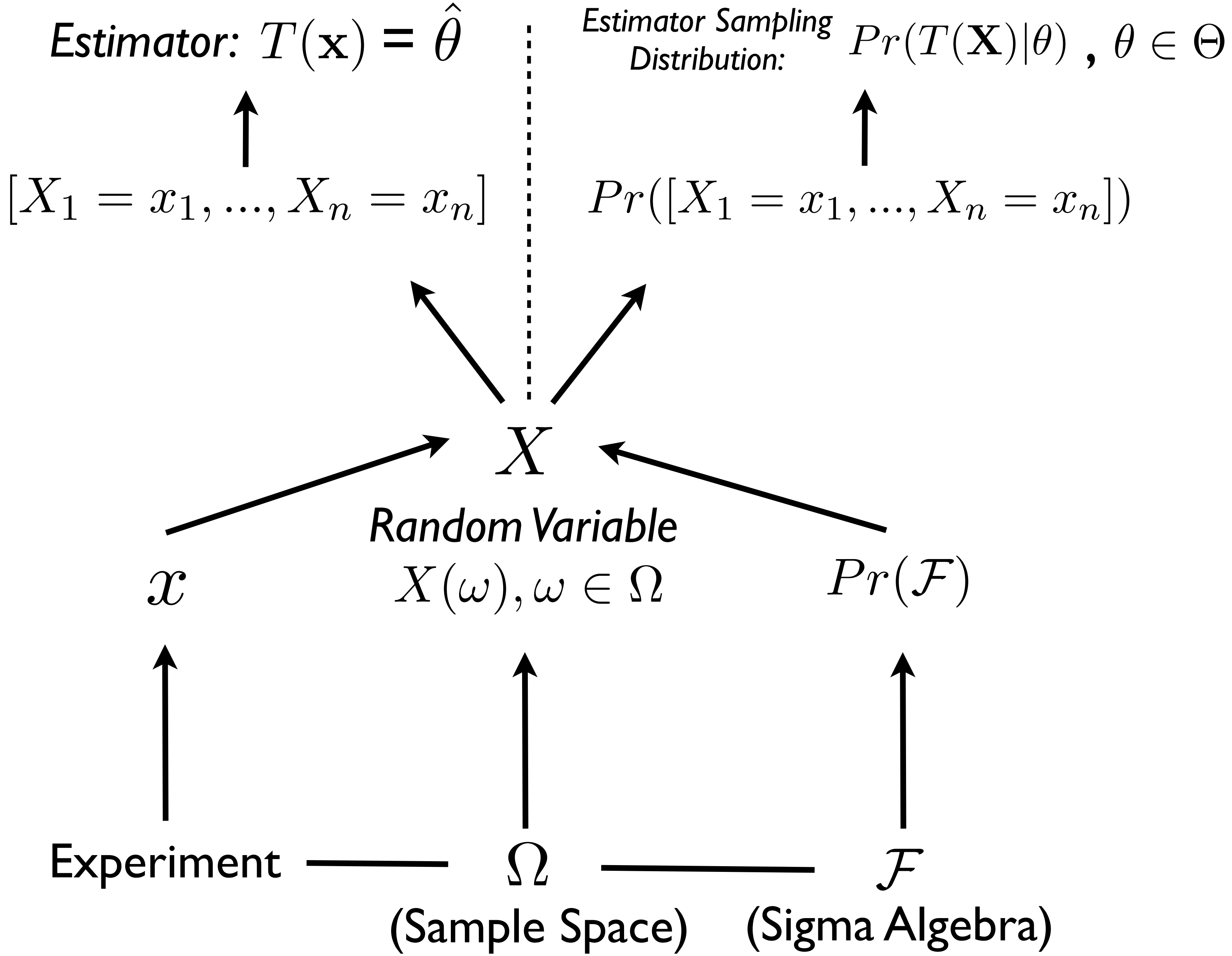
Ω

(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

Random Variable

x

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

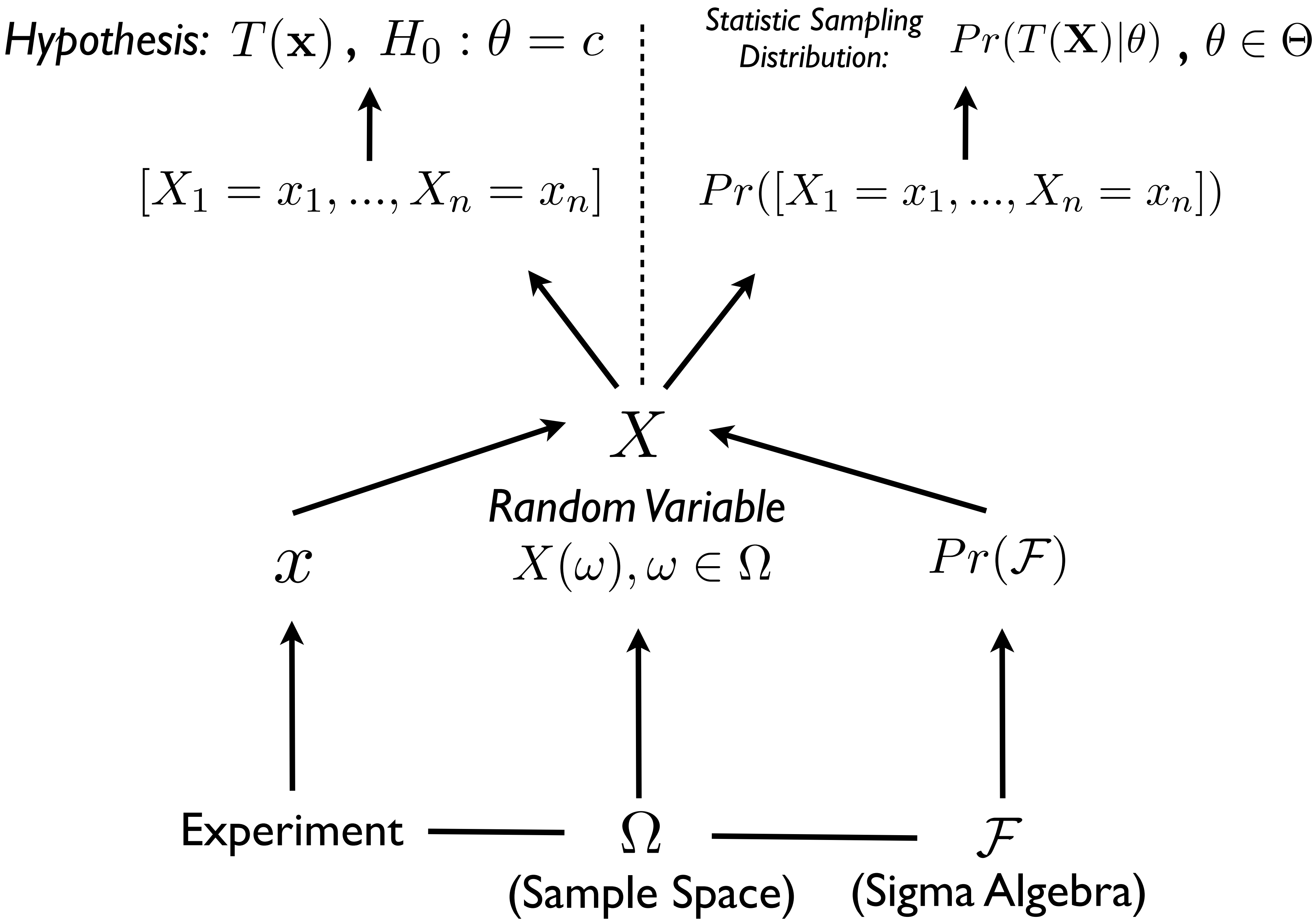
Ω

(Sample Space)

$\mathcal{F}$

(Sigma Algebra)

Hypothesis Tests



Hypothesis: $T(\mathbf{x})$, $H_0 : \theta = c$

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

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?

Hypothesis Tests

$$\begin{tikzcd}
\textit{Hypothesis:}\ T(\mathbf{x})\ ,\ H_0 : \theta = c & {} & \begin{array}{c}\textit{Statistic Sampling}\\ \textit{Distribution:}\end{array}\ Pr(T(\mathbf{X})|\theta)\ ,\ \theta \in \Theta \\
{[X_1 = x_1, ..., X_n = x_n]} \arrow[u] & & Pr([X_1 = x_1, ..., X_n = x_n]) \arrow[u] \\
X = x & & Pr(X) \\
& X \arrow[ul] \arrow[ur] \arrow[uuu, no head, dotted] & \\
x \arrow[ur] & \begin{array}{c}\textit{Random Variable}\\ X(\omega), \omega \in \Omega\end{array} & Pr(\mathcal{F}) \arrow[ul] \\
\text{Experiment} \arrow[u] \arrow[r, no head] & \underset{\text{(Sample Space)}}{\Omega} \arrow[u] \arrow[r, no head] & \underset{\text{(Sigma Algebra)}}{\mathcal{F}} \arrow[u]
\end{tikzcd}$$

Review: Hypothesis testing

- 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: p-value

- 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*(**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]:

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

Review: p-value II

- 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:

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

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

Review: Hypothesis Testing

- 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:

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

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

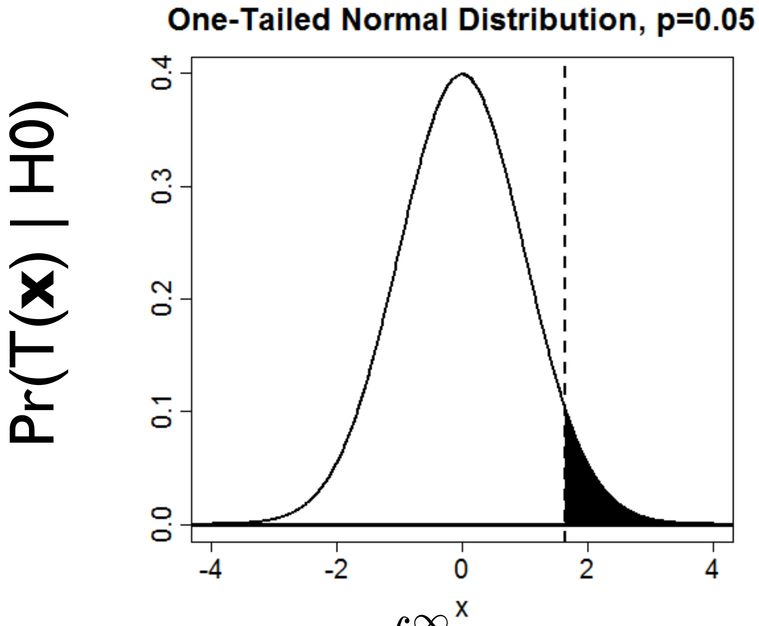
- *Note that a p-value is a function of a statistic* (!!)

Review: Non-Intuitive Hypothesis Testing Concepts

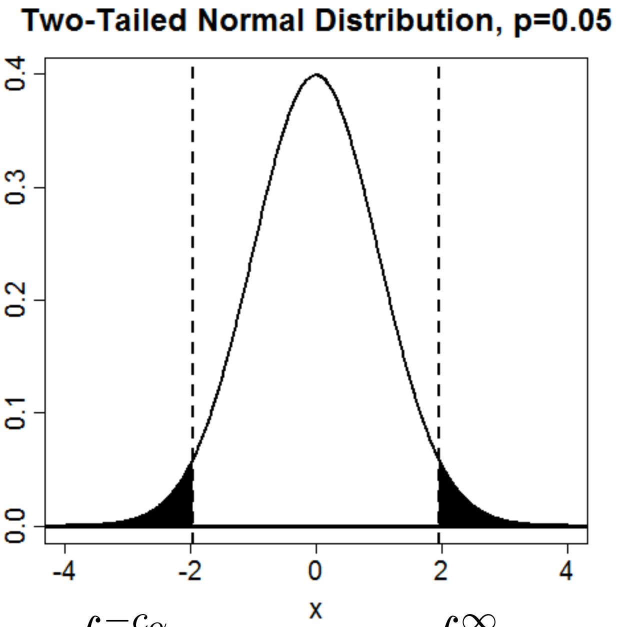
- We do not know what the true model is (=parameter values are) in a real case!
- We assess a null hypothesis that we define!
- We assess this null hypothesis by calculating a p-value which assumes that the null hypothesis is true!
- We assess this null hypothesis by calculating a p-value from a single 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)!

Review: 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:



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

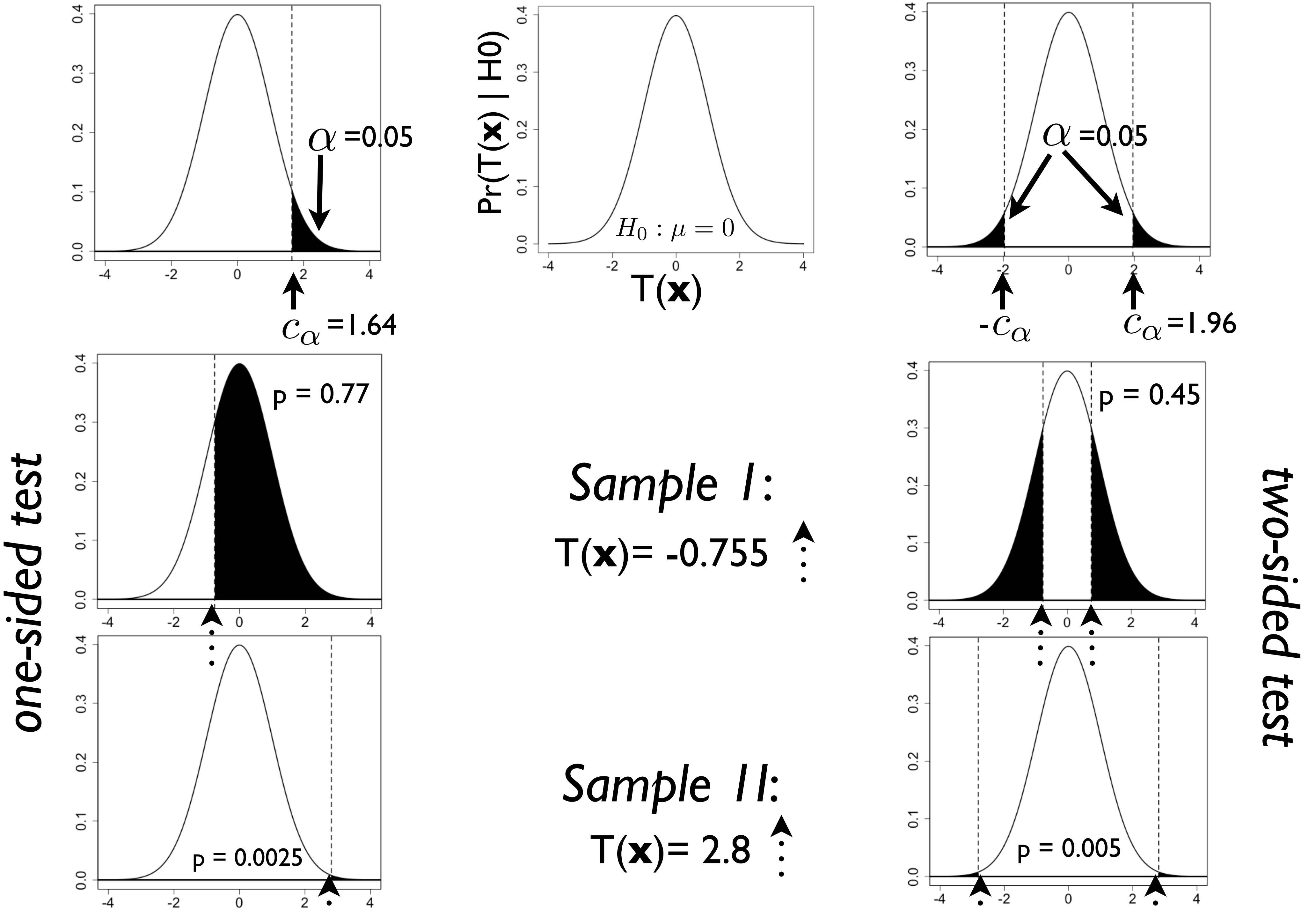


$$\alpha = \int_{-\infty}^{-c_\alpha} f_X(x)dx + \int_{c_\alpha}^{\infty} f_X(x)dx$$

Review: 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

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



one-sided test

two-sided test

Sample I:

T(**x**)= -0.755

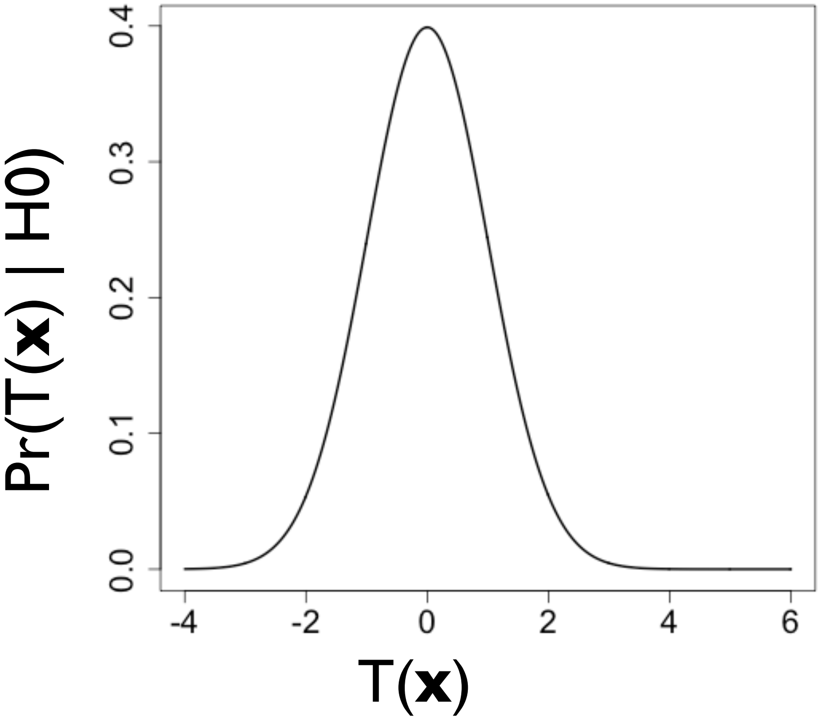
Sample II:

T(**x**)= 2.8

Review: 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*

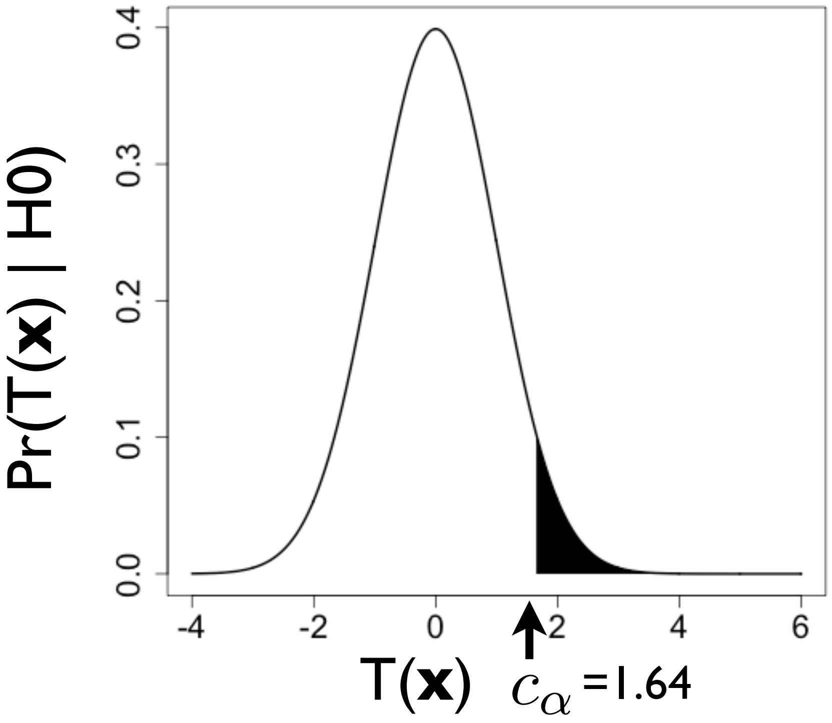
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



Review: 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*

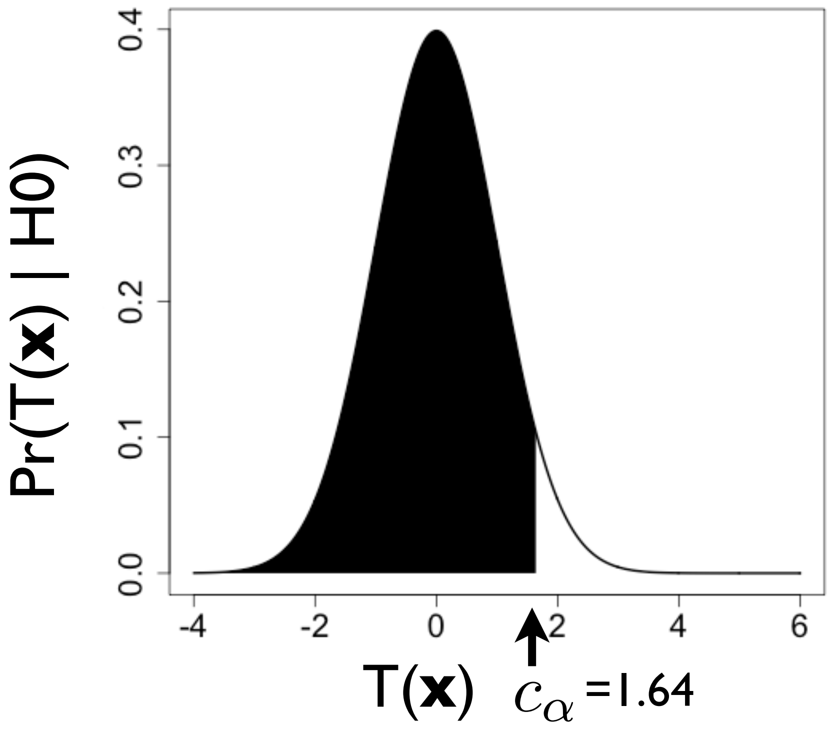
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



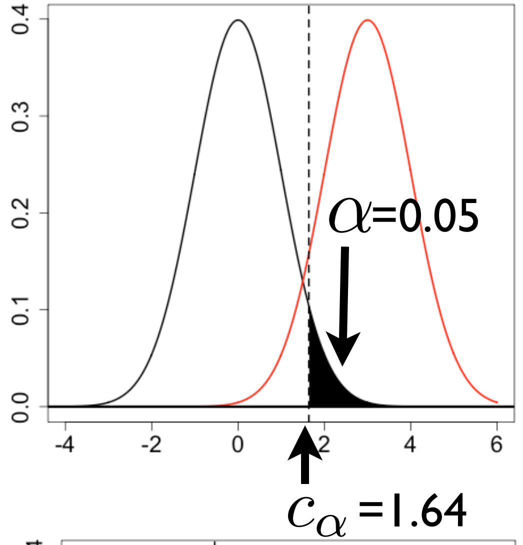
Review: Results of hypothesis decisions I: when H0 is correct

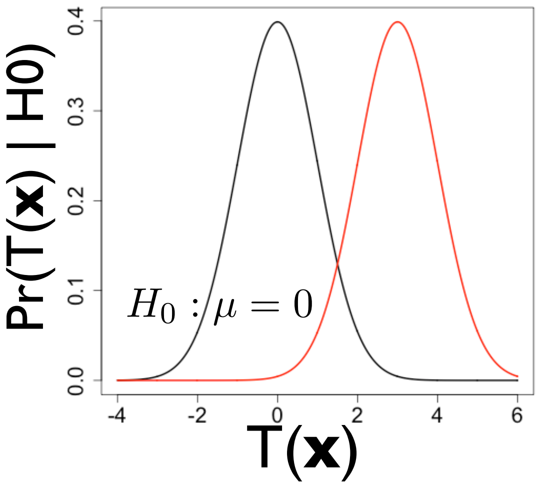
- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

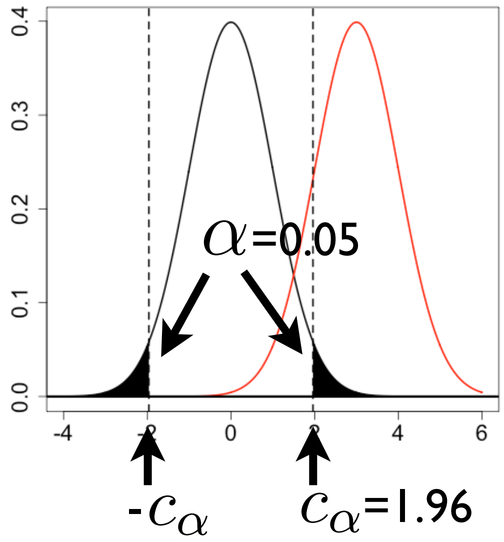
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



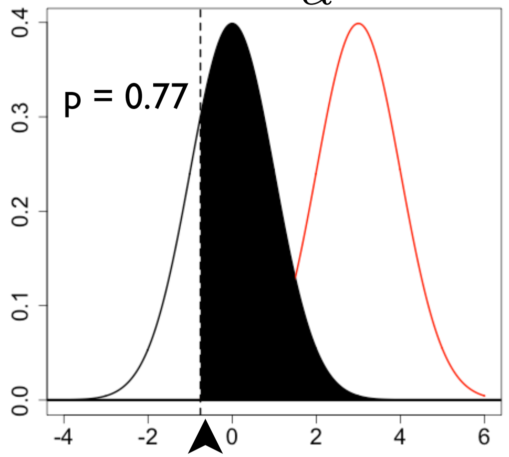
Review: Assume H0 is wrong (!): $\mu = 3$





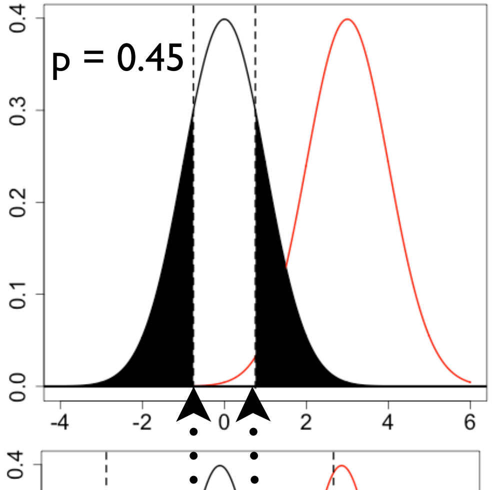


one-sided test

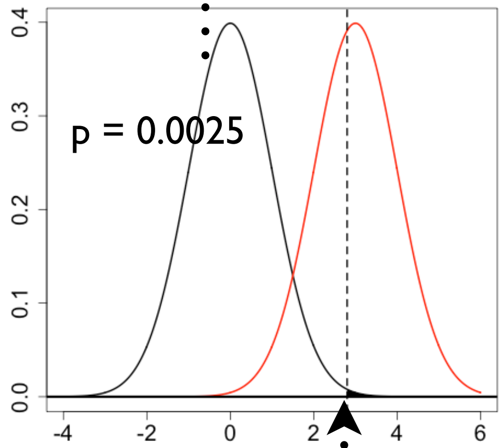


Sample I:

T(**x**)= -0.755

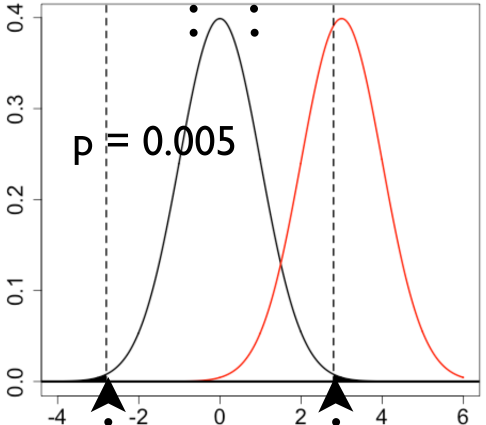


two-sided test



Sample II:

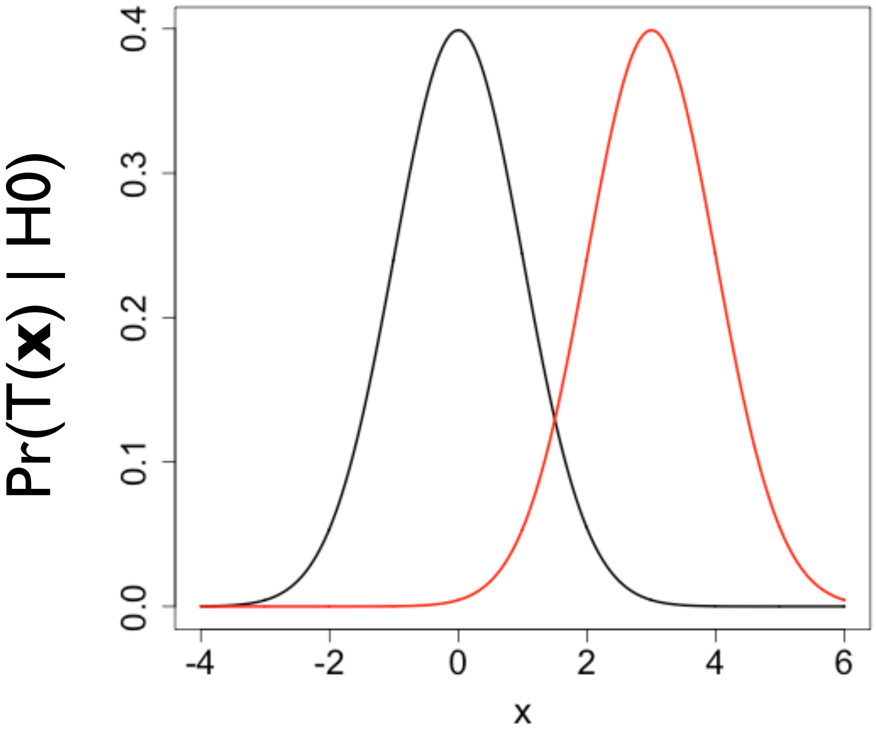
T(**x**)= 2.8



Review: 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*

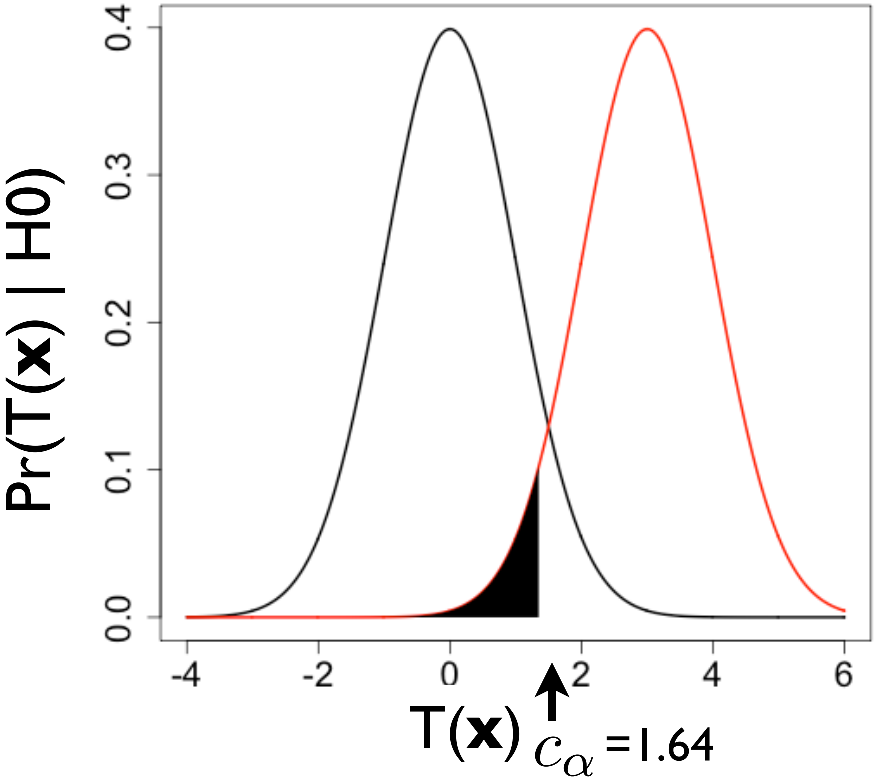
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Review: 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*

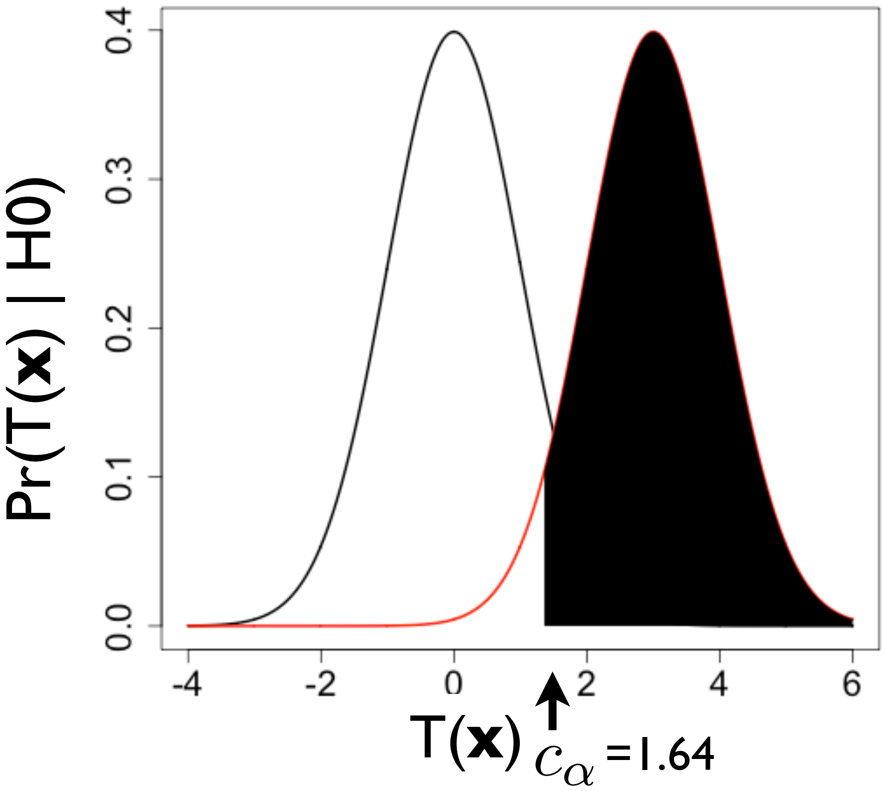
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Review: 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 I

- 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

Important concepts II

- Unlike type I error α , which we can set, we cannot control power directly (since it depends on the actual parameter value)
- However, since power $1-\beta$ depends on how far the true value of parameter is from the H0, we can make decisions to increase power depending on how we set up our experiment and test:
    - Greater sample size = greater power $1-\beta$
    - Greater the value of α that we set = greater power $1-\beta$ (trade-off!)
    - One-sided or two-sided test (which is more powerful?)
    - How we define our statistic (a more technical concept...)

Final general concept

- We need one more concept to complete our formal introduction to hypothesis testing: the *alternative hypothesis* (HA)
- This defines the set (interval) of values that we are concerned with, i.e. where we suspect our true parameter value will fall *if our H0 is incorrect*, i.e. for our example above:

$$H_A : \mu > 0 \qquad\qquad H_A : \mu \neq 0$$

- A complete hypothesis testing setup includes both H0 and HA
- HA makes the concept of one- and two-tailed explicit
- REMINDER (!!): If you reject H0 you cannot say HA is true (!!)

What if we did an infinite number of experiments to test our null?

- Note that since we have induced a probability model on our r.v. -> sample -> statistic, and a p-value is a function on a statistic, we also have a probability distribution on our p-values
- This is the possible p-values we could obtain over an infinite number of different samples (sets of experimental trials)!
- This distribution is always (!!) the uniform distribution on [0,1] when the null hypothesis is true (!!) regardless of the statistic or hypothesis test:

$$Pr(pval) \sim U[0,1]$$

Understanding p-values...

- **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**
  Null Hypothesis, Sampling Distribution Conditional on the Null, p-value, One-or-Two-Tailed,
  Type I Error, Critical Value, Reject / Do Not Reject
  1 - Type I, Type II Error, Power, Alternative Hypothesis

Likelihood ratio tests I

- Since there are an unlimited number of ways to define statistics, there are an unlimited number of ways to define hypothesis tests
- However, some are more "optimal" than others in terms of having good power, having nice mathematical properties, etc.
- The most widely used framework (which we will largely be concerned with in this class) are *Likelihood Ratio Tests* (LRT)
- Similar to MLE's (and they include MLE's to calculate the statistic!) they have a confusing structure at first glance, however, just remember these are simply a statistic (sample in, number out) that we use like any other statistic, i.e. with the number out, we can calculate a p-value etc.

Likelihood ratio tests II

- Likelihood Ratio Tests use a statistic with the following structure:

$$\Lambda = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}_1|\mathbf{x})}$$

- $L(\theta|\mathbf{x})$ is the likelihood function
- $\hat{\theta}_0 = argmax_{\theta \in \Theta_0} L(\theta|\mathbf{x})$ is the parameter that maximizes the likelihood given the sample restricted to the set of parameters defined by H0, which we symbolize by Θ_0
- $\hat{\theta}_1 = argmax_{\theta \in \Theta_1} L(\theta|\mathbf{x})$ is the parameter that maximizes the likelihood given the sample restricted to the set of parameters defined by HA $\Theta_1 = \Theta_A$ or more usually the values $\Theta_1 = \Theta_A \cup \Theta_0$
- We will assume the following for the alternative set of hypotheses, for example:

$$H_0 : \mu = c \text{ then } H_A : \mu \neq c$$

Likelihood ratio tests III

- Again, consider our simplified normal r.v. with sample *n*
- The likelihood is:

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- and the LRT statistic for $H_0 : \mu = c$ is:

$$\Lambda = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}_1|\mathbf{x})} \qquad LRT = \Lambda = \frac{\frac{1}{(2\pi * MLE(\hat{\sigma}^2))^{\frac{n}{2}}} e^{\sum_{i=1}^{n} \frac{-(x_i - H_0(\mu))^2}{2*MLE(\hat{\sigma}^2)}}}{\frac{1}{(2\pi * MLE(\hat{\sigma}^2))^{\frac{n}{2}}} e^{\sum_{i=1}^{n} \frac{-(x_i - MLE(\hat{\mu}))^2}{2*MLE(\hat{\sigma}^2)}}}$$

- where we have:

$$H_0(\mu) = c$$

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

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

Likelihood ratio tests IV

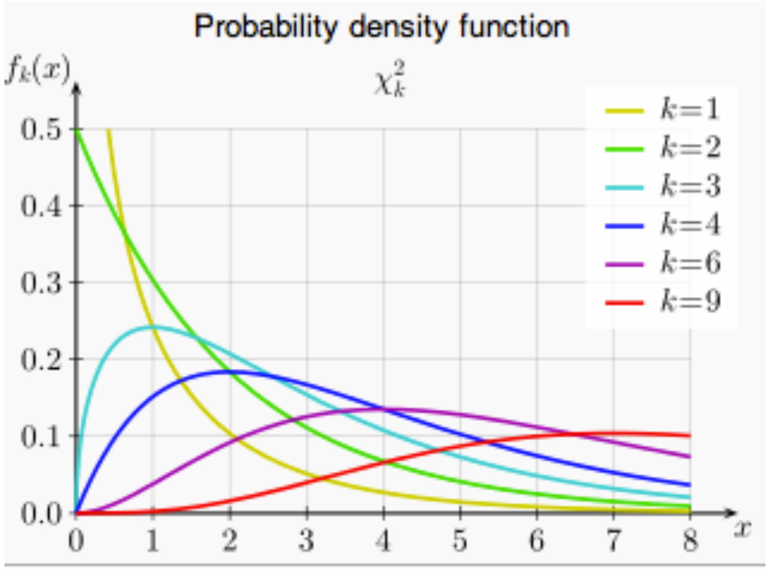
- Remember, to calculate a p-value, we need to know the sampling distribution under the null (NOTE likelihood ratio tests are two-sided tests!)

- If we consider the following transformation:

$$LRT = -2ln(\Lambda) = -2ln\left(\frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}_1|\mathbf{x})}\right)$$

- It turns out that, under conditions that often apply, as the sample size $n \rightarrow \infty$ the sampling distribution of this statistic under the null approaches (in the specific case on the last slide, the *d.f.* = *k* = 1!!):

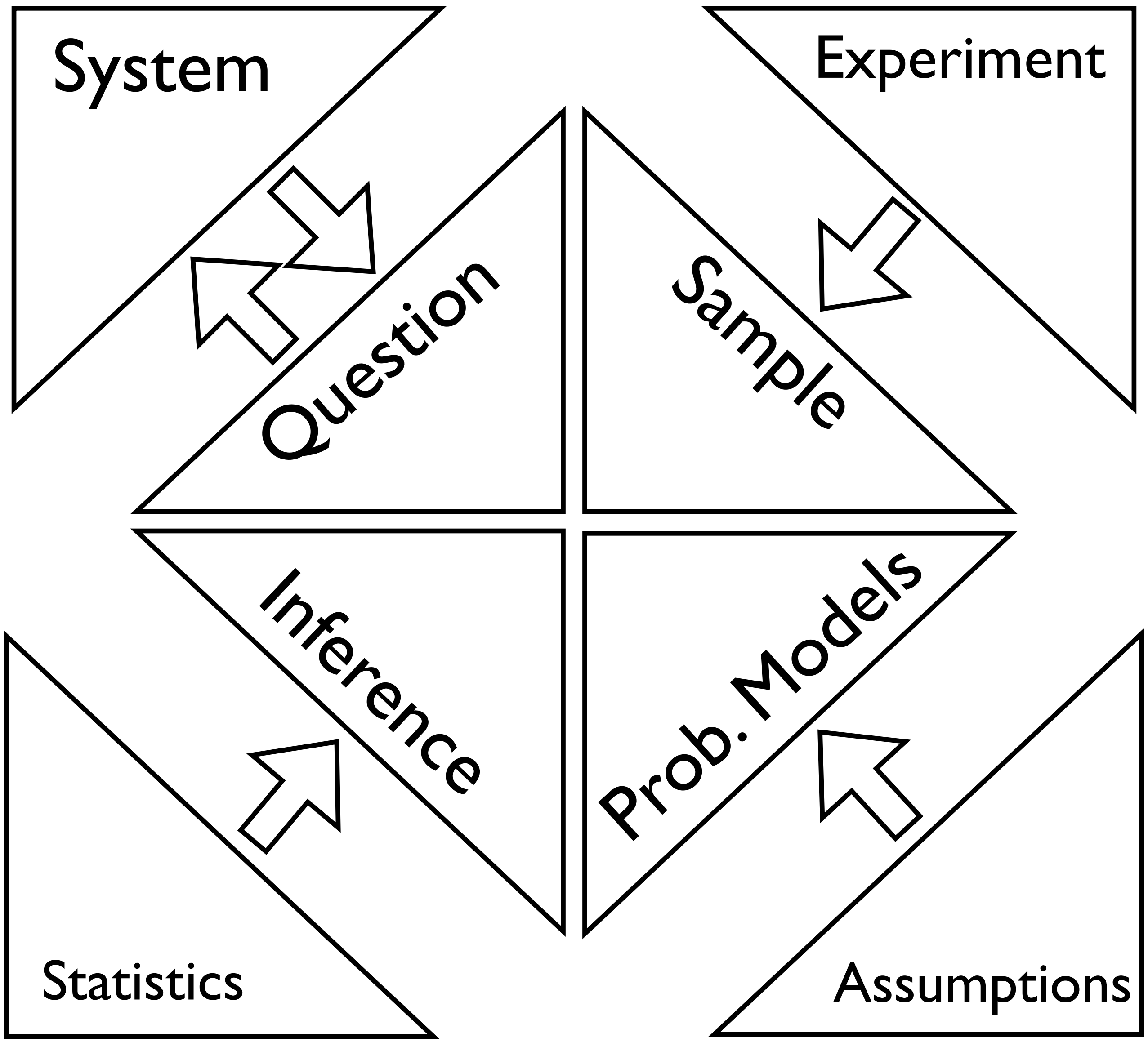
$$Pr(LRT|H_0 : \theta = c) \rightarrow \chi^2_{d.f.}$$



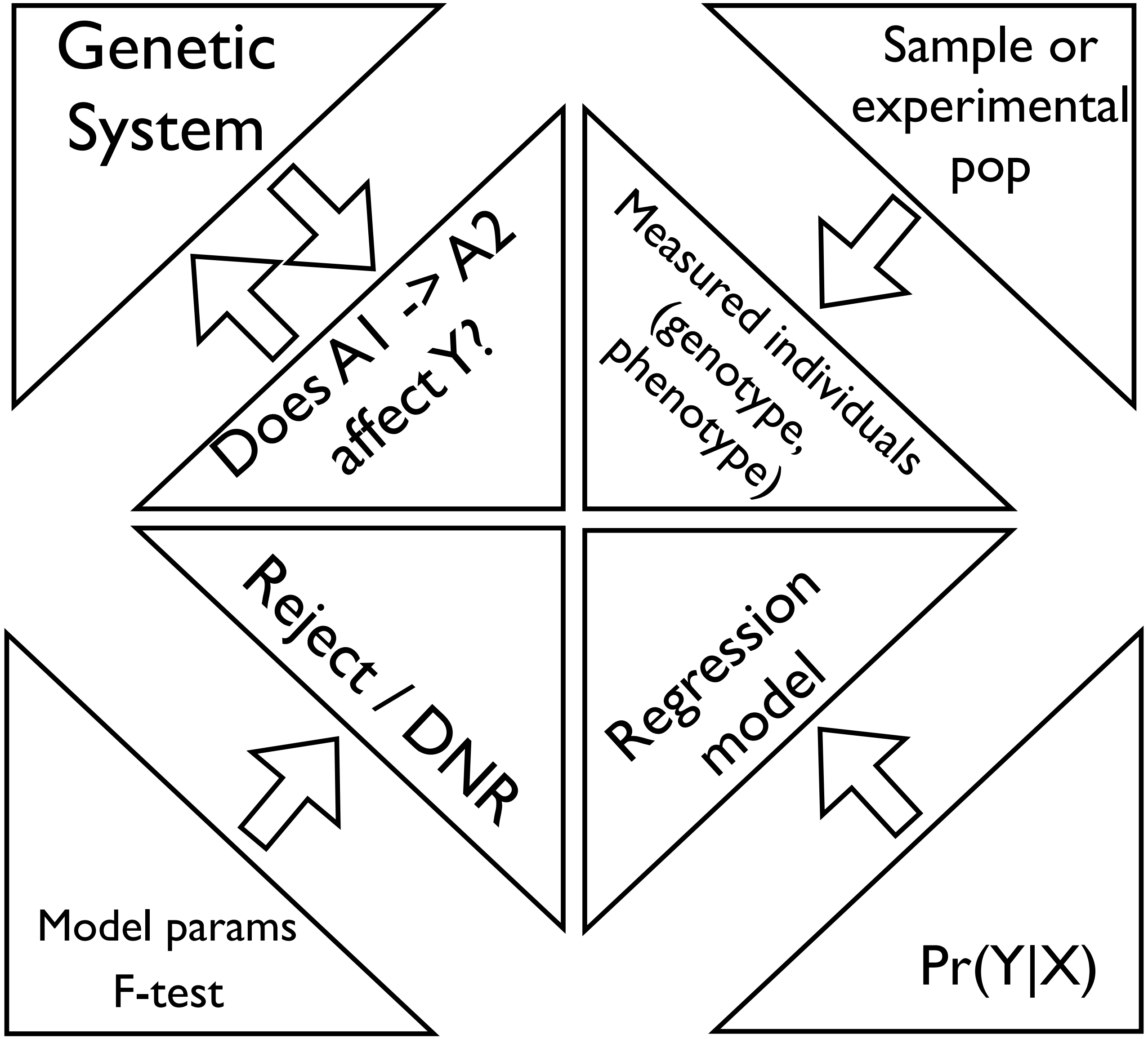
Likelihood ratio tests V

- There is a difference between a sampling distribution (under the null) that *approaches* a distribution as $n \rightarrow \infty$ and a case where we know the *exact* distribution for any size *n* (i.e., for the former, the null distribution is approximate)
- Why use a test statistic where the distribution under the null is approximate (since we need to know this distribution to do the hypothesis test!)?
    - The approximation is very close even for moderate sized *n*
    - An LRT is a very versatile way of constructing a hypothesis test with "good" properties for many types of cases
- Even better, for some specific tests, the sampling distribution under the null for ANY sample size *n* is known exactly for a specified transformation of the likelihood ratio statistic
- Note that this is the case for many of the tests you are familiar with (t-tests, F-tests, tests of the linear regression slope, etc.), that is, these tests are forms of likelihood ratio test statistic!!!

Conceptual Overview



Conceptual Overview



Genetic System

Sample or experimental pop

Does A1 -> A2 affect Y?

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Model params F-test

Pr(Y|X)

Genetic system 1

- We will reduce the complexity of a genetic system to two components: the *genome* (the inherited DNA possessed by an individual) and the *phenotype* (an aspect we measure)
- In quantitative genetics we are interested in positions in the genome where differences produce a difference in phenotype
- These differences were originally a result of a *mutation*

Genetic system II

- **mutation -** a change in the DNA sequence of a genome
- In a population of individuals (broadly defined), all differences in the genomes among the individuals were originally due to mutations
- Note: for our purposes, regardless of the cause of a mutation, we consider any difference produced in a genome that is passed on (or could be passed on) to the next generation to be a mutation
- For example, a SNP (Single Nucleotide Polymorphism; = A, G, C, T difference), Indels, microsatellites, etc.
- Also note that we will ignore the physical structure of a mutation (e.g. SNP, Indel, etc.) and quantify differences as A_i, A_j, etc.
- More specifically, we will be concerned with causal mutations, cases where the difference in genome is responsible for a difference in phenotype

Genetic system III

- **causal mutation -** a position in the genome where an experimental manipulation of the DNA would produce an effect on the phenotype under specifiable conditions

- Formally, we may represent this as follows:

$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- Note: that this definition considers "under specifiable" conditions" so the change in genome need not cause a difference under every manipulation (just under broadly specifiable conditions)

- Also note the symmetry of the relationship

- Identifying these is the core of quantitative genetics/genomics (why do we want to do this!?)

- What is the perfect experiment?

- Our experiment will be a statistical experiment (sample and inference!)

The statistical model 1

- We will make the following assumptions about the system:
  - At least one causal mutation affecting the phenotype of interest has occurred during the history of the population
  - At the locus (position) where the mutation occurred, there are at least two alleles (states of DNA) among individuals in the population (i.e. one is the original state, the other is the mutation)
- **polymorphism** - the existence of more than one allele at a locus
- These differences were originally a result of a *mutation*

The statistical model II

- For most of this class, we will be discussing *diploid* systems (i.e. cases where individuals have two copies of a chromosome), which are *sexual* (i.e. offspring are produced that have a genome that is a copy of half of the mother's and half of the father's genome), and we will be considering polymorphisms that only have two alleles (e.g. *A*1 and *A*2)

- However, note that the formalism easily extends to ANY genetic system (bacteria, tetraploids, cancer, etc.)

- We are also largely going to consider a *natural experiment* (i.e. our sample will be selected from an existing set of individuals in nature), although again, the formalism extends to *controlled experiments* as well (!!)

That’s it for today

- Next lecture (Tues, March 7), we will begin our discussion of quantitative genetic inference!