

Statistical Methods for High Dimensional Biology

Lecture 4 – Review of Probability and Statistics

January 15/2018

Sara Mostafavi

Who am I?

- Assistant professor, jointly appointed in Statistics and Medical Genetics, and associated faculty of Computer Science.
- Computational Biologist: interested in developing and using statistical machine learning methods for high-dimensional genomics data, human health and precision medicine.
- BSc, Queen's + UToronto, Computer Sci and Life Sci;
- PhD, UToronto, Computer Science, (Machine learning, and computational biology).
- Postdoc, Stanford/Harvard, Statistical Machine learning for human health

Announcements

- GitHub account + email from TAs
- Keep on submitting seminar deliverables!
- Crow sourcing experiment: keep track of frequently encounters issues in seminars

Outline

- Statistics: Philosophy, goals, and central concepts
- Basics: what you should know / terminology
- Hypothesis testing

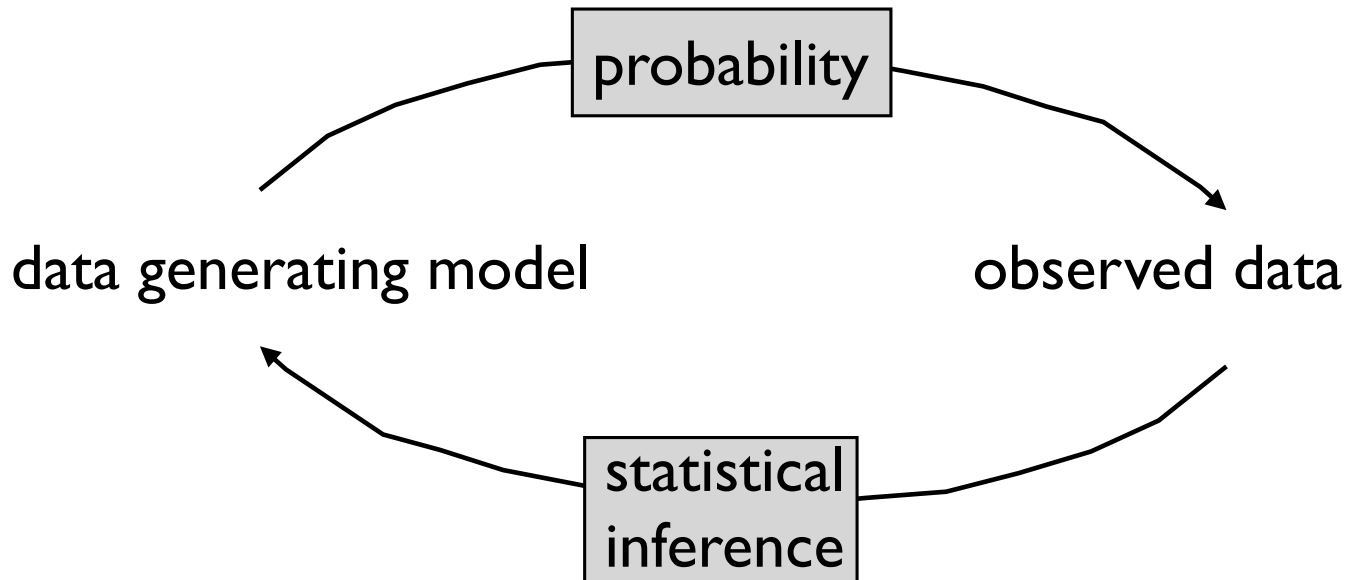
Your goals:

- Make sure a) you know the terminology (can't advance past this one if you don't know the language), b) you are not confused by any of the concepts

Statistics: philosophy and goals

- Data science, statistics, and machine learning all concern collecting and analyzing data.
- The field of statistics concerns the science of measuring and **learning** from data and **communicating uncertainty** about the results.
 - Data science and machine learning have enabled “on mass” application.
- Statistical and computational methods should not be used as “recipes” to follow → non robust science.
 - We emphasize: rigorous understanding to perform routine statistical analysis but also foundation to follow up on specific topics.

Probability and Inference



Learning (from data): The acquisition of knowledge through experience
Hence “information/interpretation” + “prediction”

- Language of probability enables us to make *predictions* and discuss *uncertainty*.
- Statistical inference enables us to *understand* the data.
- We need both to learn from data.

Review of terminology and basic concepts

- Random variable and its distribution
- Independent and identically distributed
- Models, parameters and their estimators
- Central Limit Theorem
- Hypothesis testing

RV and its distribution

- **Sample space**: set of all possible outcomes of an experiment.
- **Random Variable (RV)**: A variable whose value results from the measurement of a quantity that is subject to variation (outcome of an experiment)
 - RV = A *function* that maps any possible outcome of an experiment to a real number
 - E.g., outcome of coin flip, expression level of gene A
- **Probability** : A number assigned to an outcome, satisfying certain rules (for now okay to think of as *frequency* of an outcome)
- **Probability distribution** : A function that maps outcomes to probabilities

Example

Experiment: 2 coin tosses





Sample space $\Omega \rightarrow (TT, TH, HT, HH)$

Random variable $X(\omega) \rightarrow$ Number of heads

$\omega \rightarrow$ Greek letters for outcome of the experiment

$X(\omega) \rightarrow$ capital letters for Random variables

$\Omega \rightarrow$ sample space





	ω	$X(\omega)$
TT		0
TH		1
HT		1
HH		2

NB: We could define other rvs for tossing two coins.

Assigning probabilities to outcomes

ω = an outcome of the experiment

$X(\omega)$ = number of heads

			Probability distribution	
	probability	$X(\omega)$	$\frac{P(X=x)}{P_X(x)}$	x
	0.25	0	0.25	0
	0.25	1	0.5	1
	0.25	1	0.25	2
	0.25	2	<hr/> 1	
	<hr/> 1			

Each realization of our random variable corresponds to an event in the sample space, and we can associate a probability to each realization. SO an RV has an associated probability distribution.

Two types of random variables

- A **discrete** rv has a countable number of possible values
 - e.g. dice throwing outcome, genotype measured on a SNP chip
- A **continuous** rv takes on values in an interval of numbers
 - e.g., expression level of a gene, blood glucose level

Probability mass/density function

- **Probability distribution** is the mathematical function describing the possible values of random variables and their associated probabilities.
 - Discrete rv associated with probability mass function (pmf)
 - Continuous rv associated with probability density function (pdf)
- It's rare for outcomes and associated probabilities of a rv to be represented as a table of numbers (toy examples mostly!)
- Much more common and elegant (and necessary): we have a mathematical formula that gives the probability of $X=x$ for all x .

Random variables can be characterized by a distribution

Example: discrete RV

Binomial PDF

Following previous example...

X : number of heads in n tosses

The diagram illustrates the components of a binomial distribution. It features three main elements: a variable X , a probability distribution formula, and a parameter p . An arrow labeled "Variable" points from the text "Variable" to the X in the expression $X \sim \text{Bin}(n, p)$. Another arrow labeled "parameter" points from the text "parameter" to the p in the same expression. A third arrow labeled "probability distribution" points from the text "probability distribution" to the formula $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$.

$$X \sim \text{Bin}(n, p)$$
$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

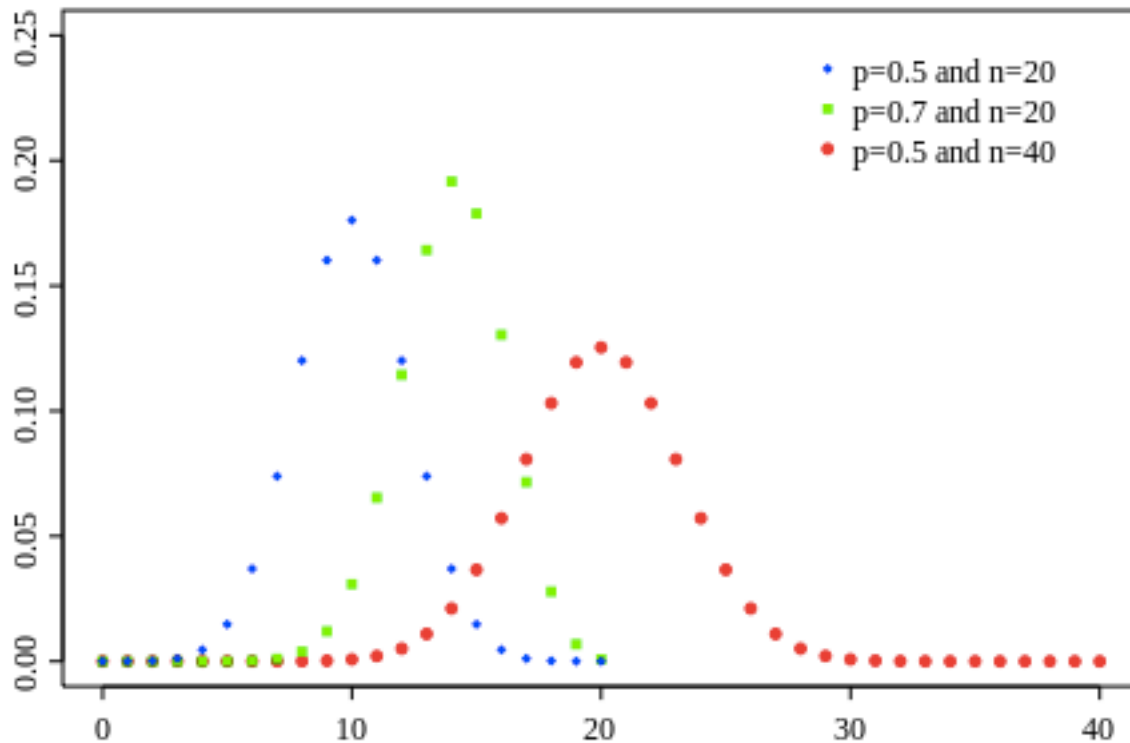
Variable

parameter

probability distribution

$X \sim \text{Bin}(n, p)$ ← parameter

$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$ ← probability distribution



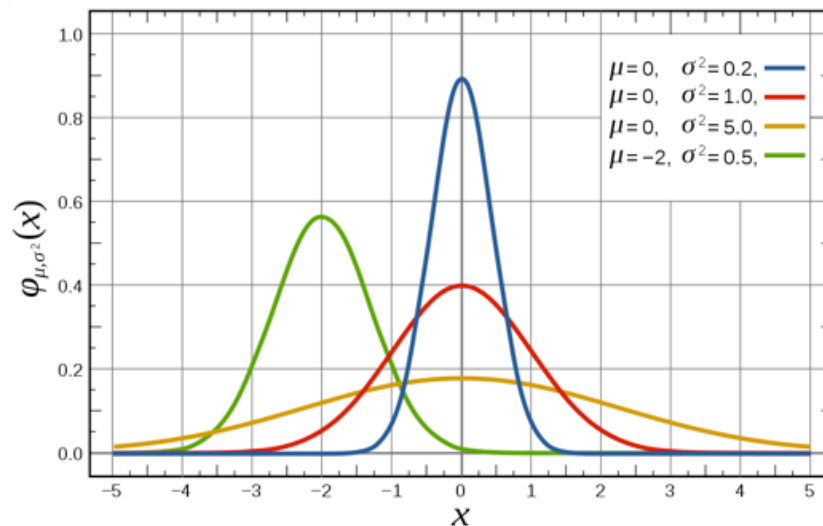
Example PDF: normal

normal, Gaussian

Parameters of
the distribution

$$X \sim N(\mu, \sigma^2)$$

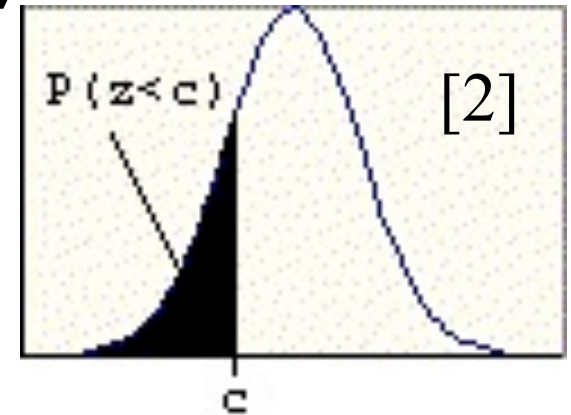
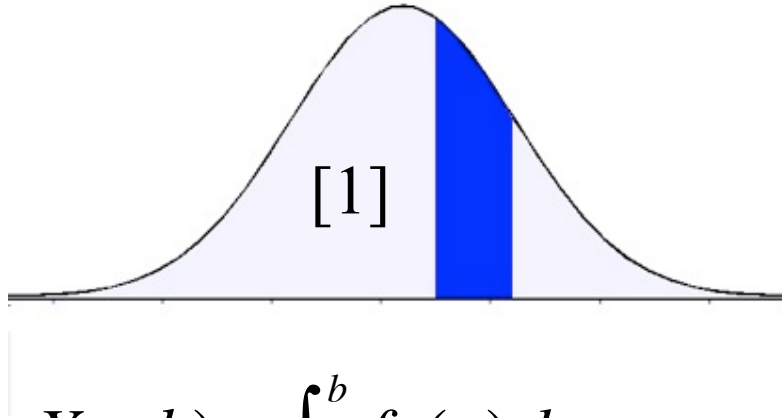
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



Probabilities for continuous RVs

- The density does not give you probabilities directly: $f(x)$ is not the probability that X takes the exact value x (in fact $P(X=x) = 0$)
- More “proof” $f(x)$ is not a probability: $f(x)$ can be greater than 1
- Probabilities are obtained from densities by integration.
- Integral of $f(x)$ over $(-\infty, \infty) = 1$ (probability of there being *some* outcome is 1)

how to get a probability from a density

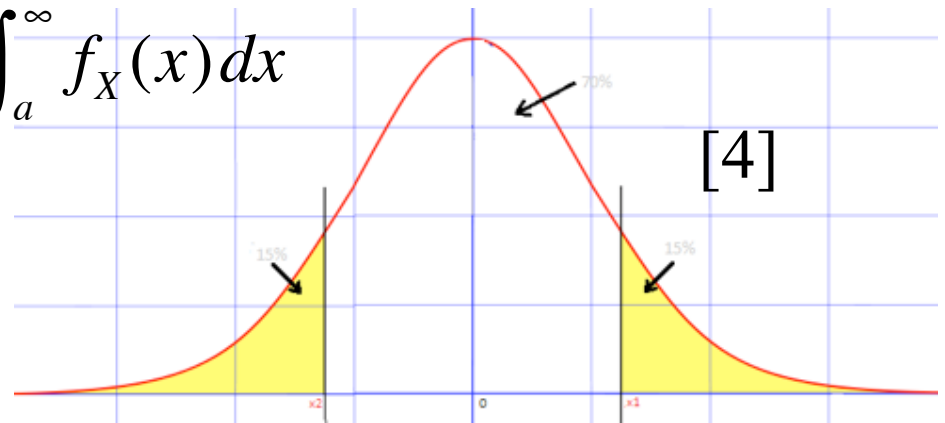
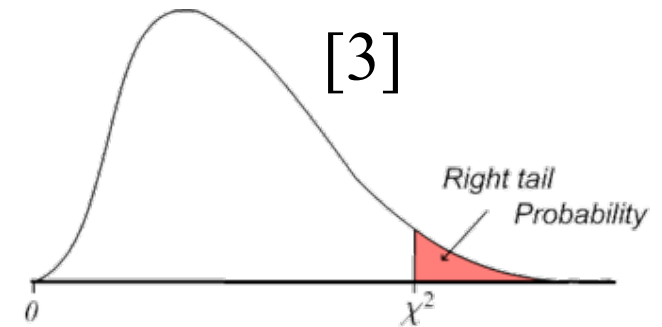


$$[1] P(a < X < b) = \int_a^b f_X(x) dx$$

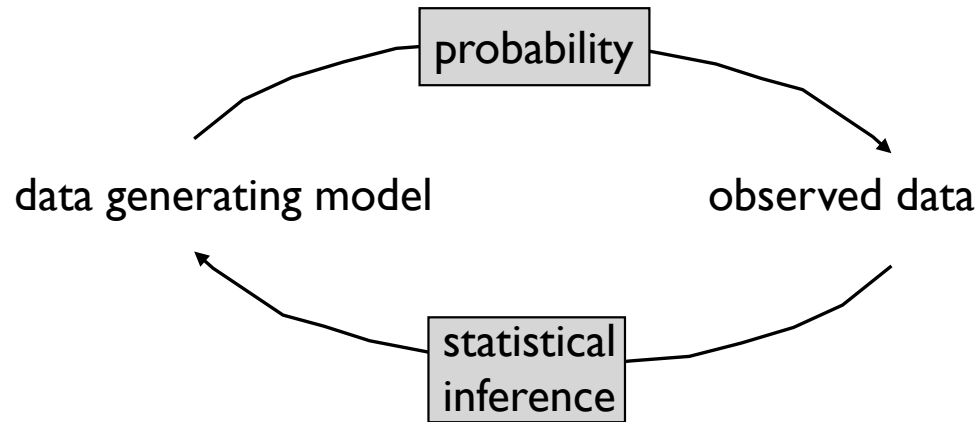
$$[2] P(X \leq a) = \int_{-\infty}^a f_X(x) dx$$

$$[3] P(X \geq a) = \int_a^{\infty} f_X(x) dx$$

$$[4] P(|X| \geq a) = \int_{-\infty}^{-a} f_X(x) dx + \int_a^{\infty} f_X(x) dx$$



Probability and Inference



Learning (from data): The acquisition of knowledge through experience
Hence “information/interpretation” + “prediction”

- Language of probability enables us to make *predictions*.
- Statistical inference enables us to *understand* the data.

Motivating example

You are a prisoner and the only way to save your life is to work out one of two math problems.

You can pick which of the two related problems you'd like to solve.

Here they are ...

Problem #1

- There is a coin that comes up *heads* with probability $p_H = 0.5$
- The executioner is going to conduct 10,000 **experiments** (or **trials**), where each experiment = counting the number of heads in 10 “regular” flips of the coin. **Outcome** of each experiment = number of heads in 10 coin flips.
- Q1: what’s the proportion of experiments where the **outcome** is 7?
- Let p_{\odot} be the difference between your guess and the observed proportion. You’ll be executed with probability p_{\odot} .

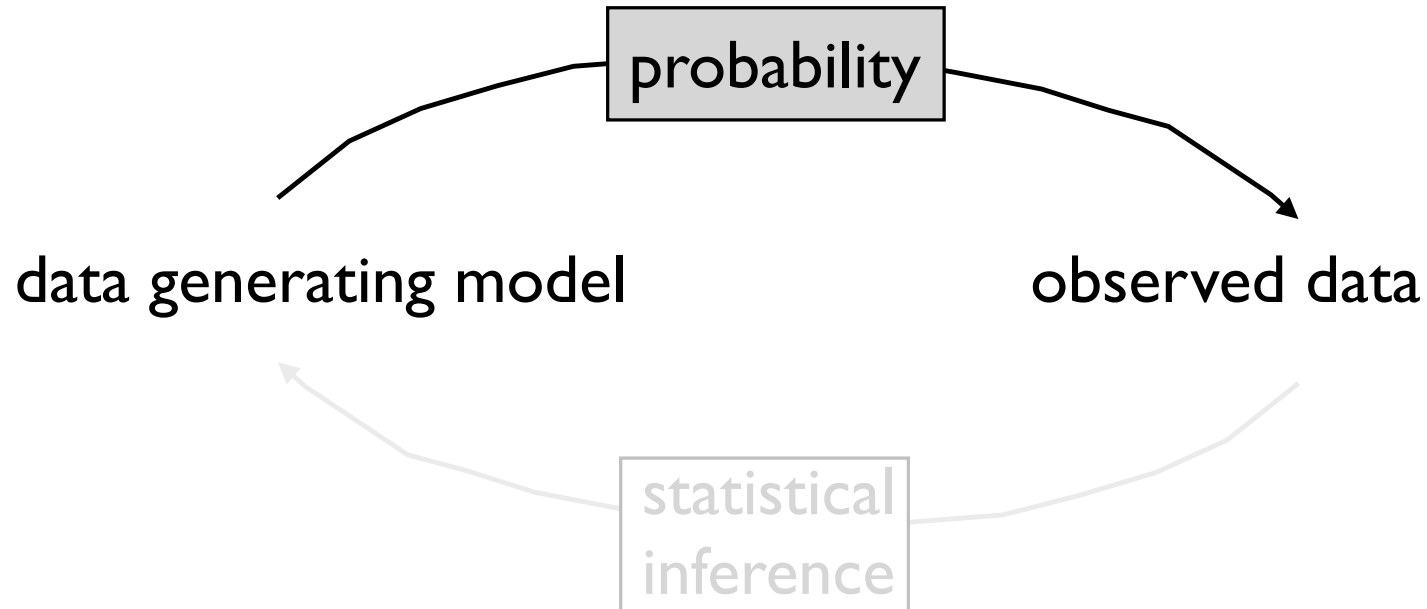
Problem #2

- The executioner is going to tell you the outcome of 10,000 experiments, where each experiment=number of heads in 10 coin flips.
- You must describe the coin(s) and toss(es).
- Let $p_{\text{☹}}$ be like so: If no “difference” between your description and the truth, then $p_{\text{☹}} = 0$. As “difference” grows, $p_{\text{☹}}$ tends to 1.*
- You will be executed with probability $p_{\text{☹}}$.

* Sorry this is so vague but I can't do better without getting bogged down in details. Go with me.

Problem #1

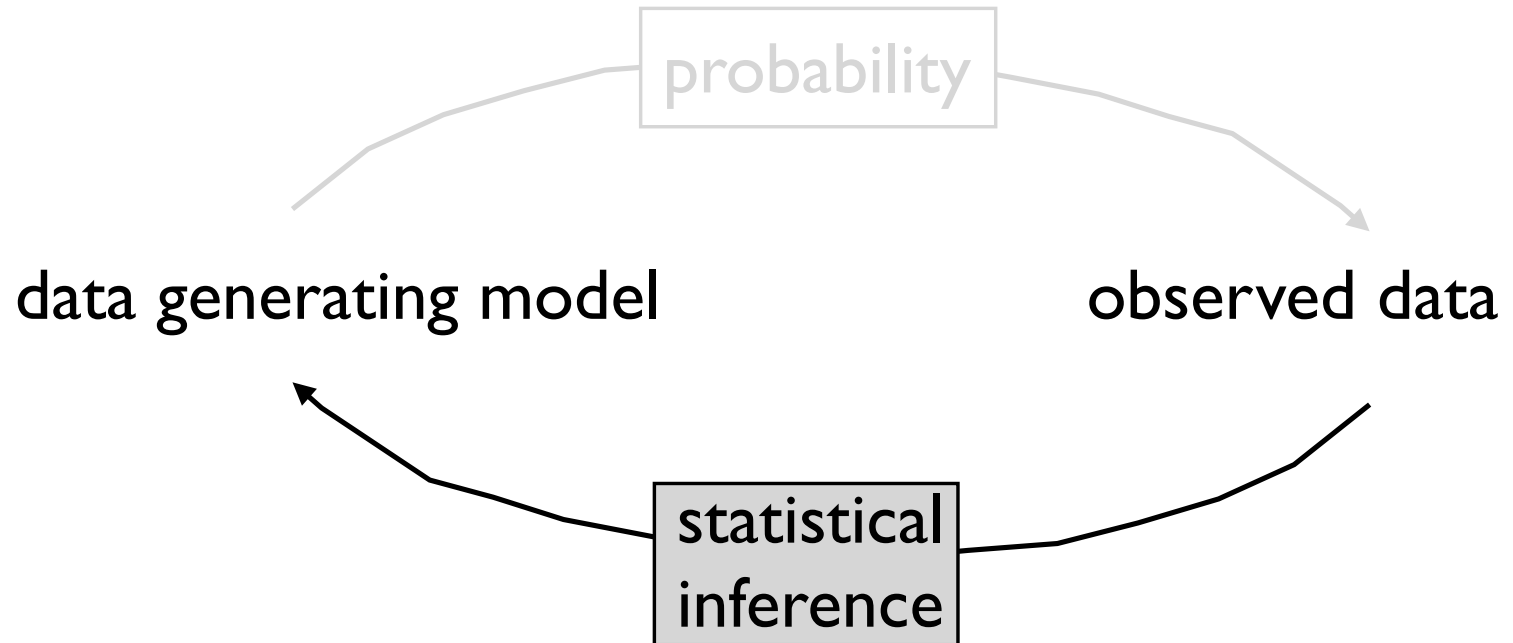
“Given the data generating model, what are some properties of the observed data?”



Adapted from Figure 1 of “All of Statistics” and associated text.

Problem #2

“Given the observed data, can we describe the model that generated the data?”



“Statistical inference is the process of deducing properties of an underlying distribution by analysis of data”

Adapted from Figure 1 of “All of Statistics” and associated text.

Back to the prisoner Qs:

Q1:

There is a coin that comes up *heads* with probability $p_H = 0.5$

The executioner is going to conduct 10,000 **experiments** (or **trials**), where each experiment/trial = counting the number of heads in 10 “regular” flips of the coin.

Q: what's the proportion of experiments where the **outcome** is 7?

RV is “# of heads in 10 tosses”.

RV has a binomial distribution



$$X \sim \text{Bin}(n, p)$$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

For Q1 we are given the parameters of the model (i.e., binomial distribution)



$$X \sim \text{Bin}(n = 10, p = 0.5)$$

$$P(X = 7) = \binom{10}{7} 0.5^7 0.5^3 \approx 0.1172$$

I'd guess that 1172 of 10000 experiments/trials will have an outcome of 7 heads.

$$X \sim \text{Bin}(n = 10, p = 0.5)$$

$$P(X = 7) = \binom{10}{7} 0.5^7 0.5^3 \approx 0.1172$$

R code for the computer experiment and visualization of the “data” :

```
> B <- 10000
> n <- 10
> p <- 0.5
> x <- 7
> choose(n, x) * p^x * (1 - p)^(n - x)
[1] 0.1171875
> dbinom(x = x, size = n, prob = p)
[1] 0.1171875
> (myGuess <- round(dbinom(x = x, size = n, prob = p) * B, 0))
[1] 1172
> (obsFreq <- sum(rbinom(n = B, size = n, prob = p) == x))
[1] 1145
> (pSad <- abs(myGuess - obsFreq)/B)
[1] 0.0027
```

← Not too bad, as probability of death goes.

“Brute force” solution to Q1

```
> B <- 10000
```

```
> coinFlips <- runif(n * B) > 0.5      # heads = TRUE (Imagine the executioner gave you time to flip coins yourself to try it out)
```

```
> coinFlips <- matrix(coinFlips, nrow = B)
```

```
> head(coinFlips)
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
```

```
[1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE
[2,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE
[3,] TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE
[4,] TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE TRUE
[5,] TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE
[6,] FALSE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE
```

```
> y <- rowSums(coinFlips)
```

```
> head(y)
```

```
[1] 2 6 4 7 5 5
```

```
> head(y == 7)
```

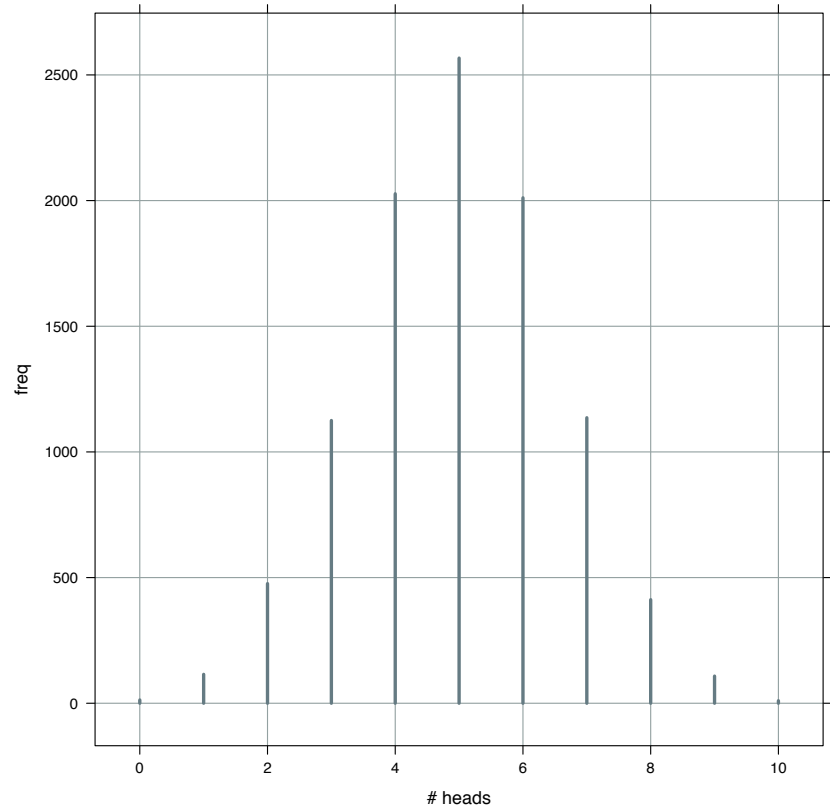
```
[1] FALSE FALSE FALSE TRUE FALSE FALSE
```

```
> (myGuess <- sum(y == 7))
```

```
[1] 1136
```

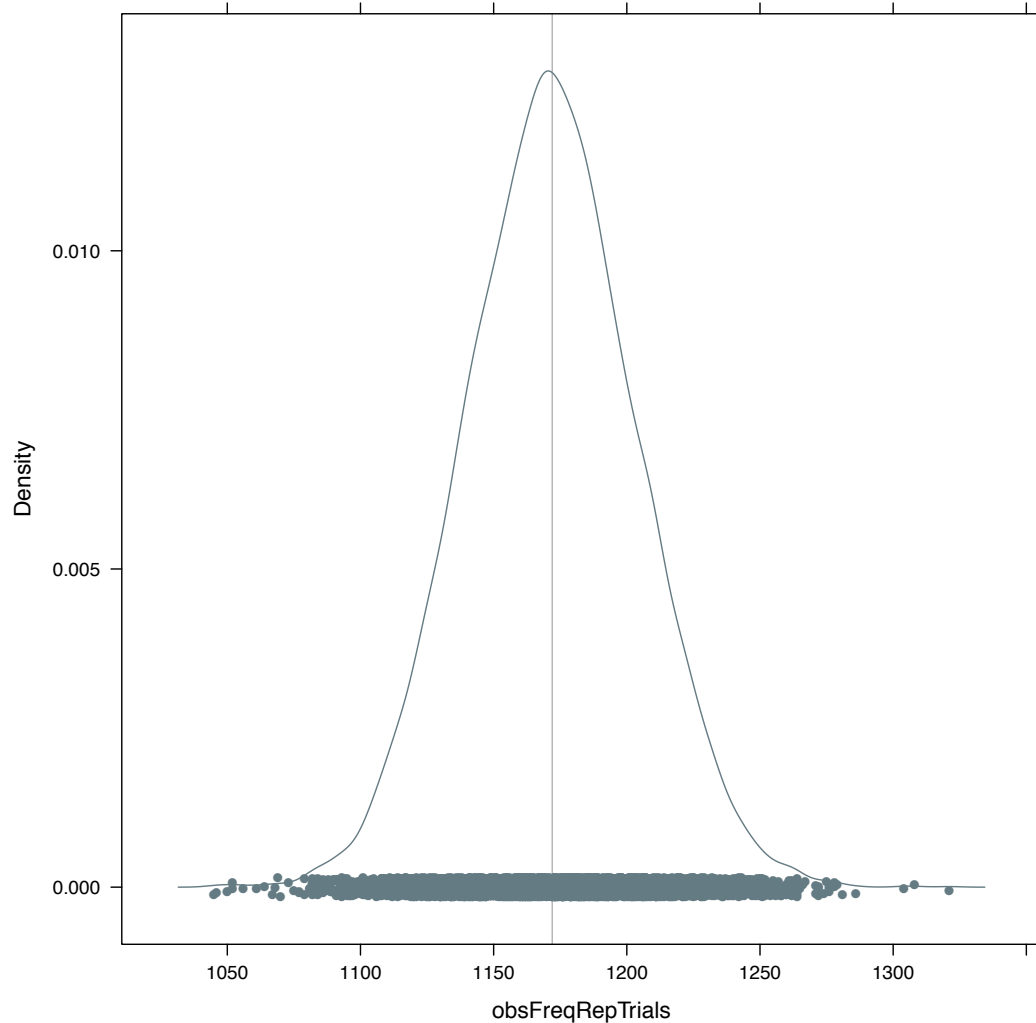
```
> (pSad <- abs(myGuess - obsFreq)/B)
```

```
[1] 0.0009
```



Not too bad, as probability of death goes. In this instance outperforms the math solution but that's not a general fact.

Empirical dist'n of many “brute force solutions” ... on average, gets the “math solution”, i.e. guessing that 1172 of 10000 trials will result in 7 heads (vertical line).

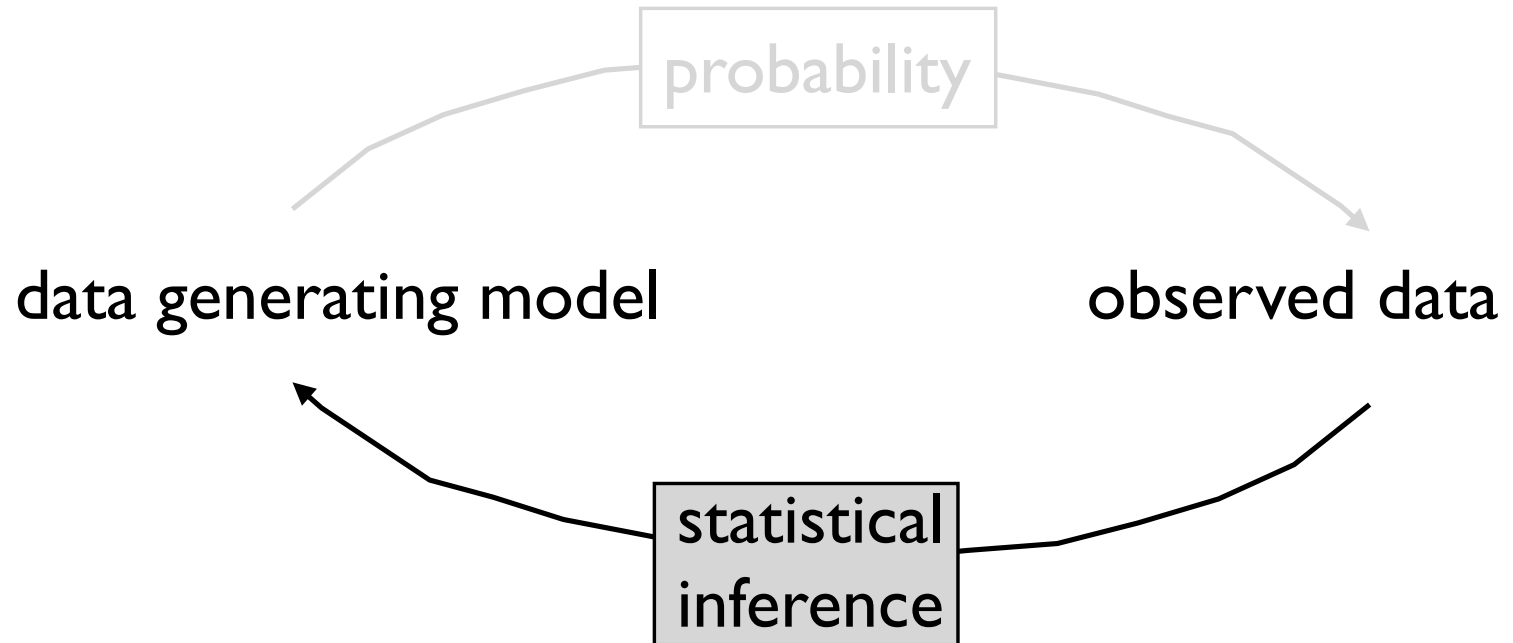


Recall Problem #2

- The executioner is going to tell you the outcome of 10,000 experiments, where each experiment=number of heads in 10 coin flips.
- You must describe the coin(s) and toss(es).
- Let $p_{\text{☹}}$ be like so: If no “difference” between your description and the truth, then $p_{\text{☹}} = 0$. As “difference” grows, $p_{\text{☹}}$ tends to 1.*
- You will be executed with probability $p_{\text{☹}}$.

* Sorry this is so vague but I can't do better without getting bogged down in details. Go with me.

“Given the observed data, can we describe the model that generated the data?”



“Statistical inference is the process of deducing properties of an underlying distribution by analysis of data”

Adapted from Figure 1 of “All of Statistics” and associated text.

“Solution” to Q2

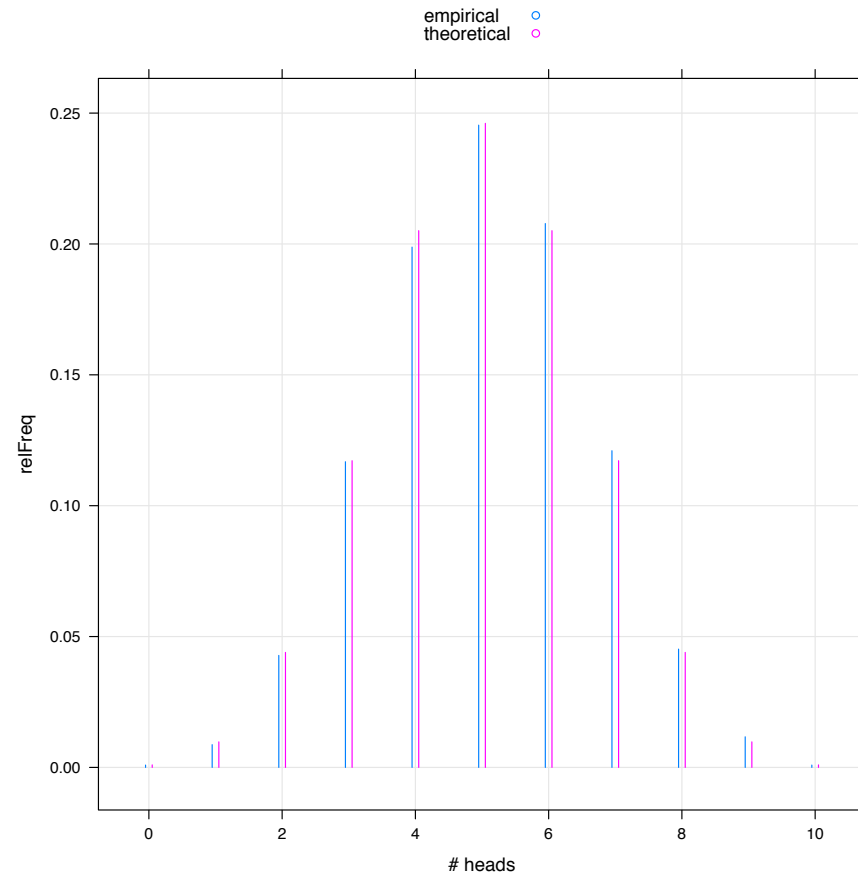
Assuming nothing ... is probably a death sentence!

You'll hope that: a) same coin was flipped in each experiment, b) the flips in each experiments are “regular flips”.

What would you do? Maybe inspect the data to see if it looks plausible under the binomial/regular flip model?

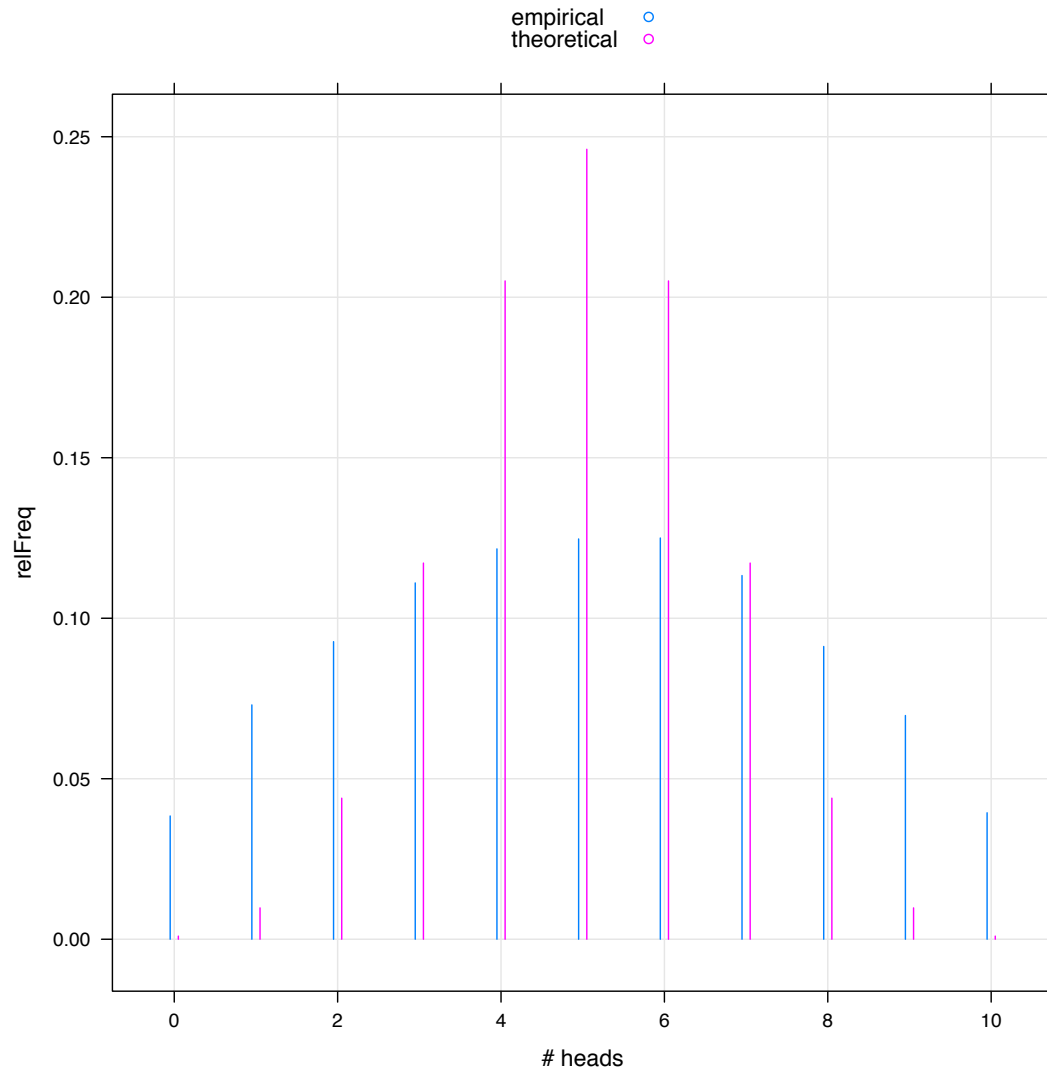
Comparison of empirical distribution to the theoretical distribution with $\text{Bin}(n=10, p)$ for some p (in this case $p=0.5$)

The empirical distribution seems plausible ... you can relax a little.



* Here I used $p=0.5$, but one could imagine varying p , and picking the one that “best” matches the empirical data.

But what if the empirical distribution looked like this?



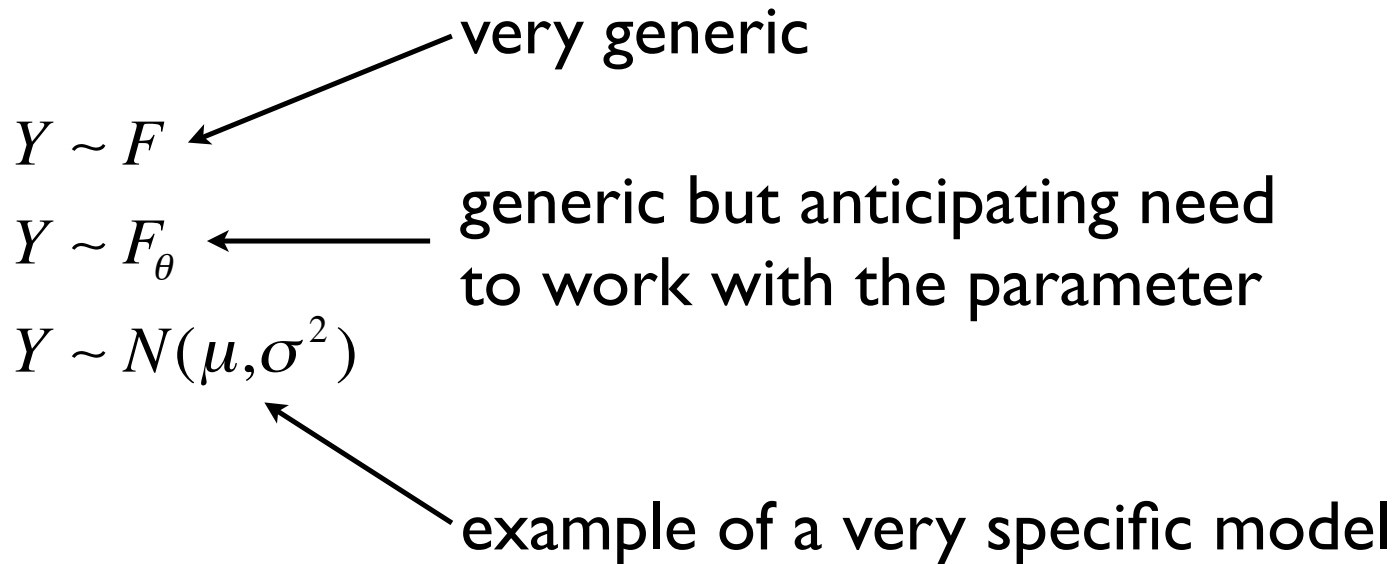
The binomial model can't be right (even with varying p), the empirical distribution is much more spread out.

“Solution” to Q2 continued

- Based on inspection, you need to decide if you believe these data were generated by one coin (“regular flips”).
- Then you need figure out value of the parameter.
- This example highlights: the importance of knowing how data was generated/collected.
- Inference involves “judgment calls”, and there are no “right answers” (however, there are “wrong” ones).

Statistical models

First some notation...



a statistician doesn't mean much when they say
“model” ... nothing terribly specific or mechanistic ...
just specifying a probability distribution and, optionally,
more details about the parameter(s)

Parameters determine distributions

- When sampling from a population described by a pmf/pdf $f(X|\theta)$, then knowledge of θ yields knowledge of the entire population.
 - This description is the “statistical model”
 - Think back to the prisoner problem
- This is why parameter estimation is useful:
 - If we are tossing a coin, we would like to estimate the parameter p

Statistical model & inference

- The *parameter space* is the set of all possible values for the parameter
- A goal is to “*guess*” (i.e., estimate) the parameter values: “fit the model to the data”
- The model is a representation that (we hope) approximates the data and (more importantly) the population that the data were sampled from.
- We can then use this model:
 - For hypothesis testing and other forms of inference
 - For prediction
 - For simulation

IID

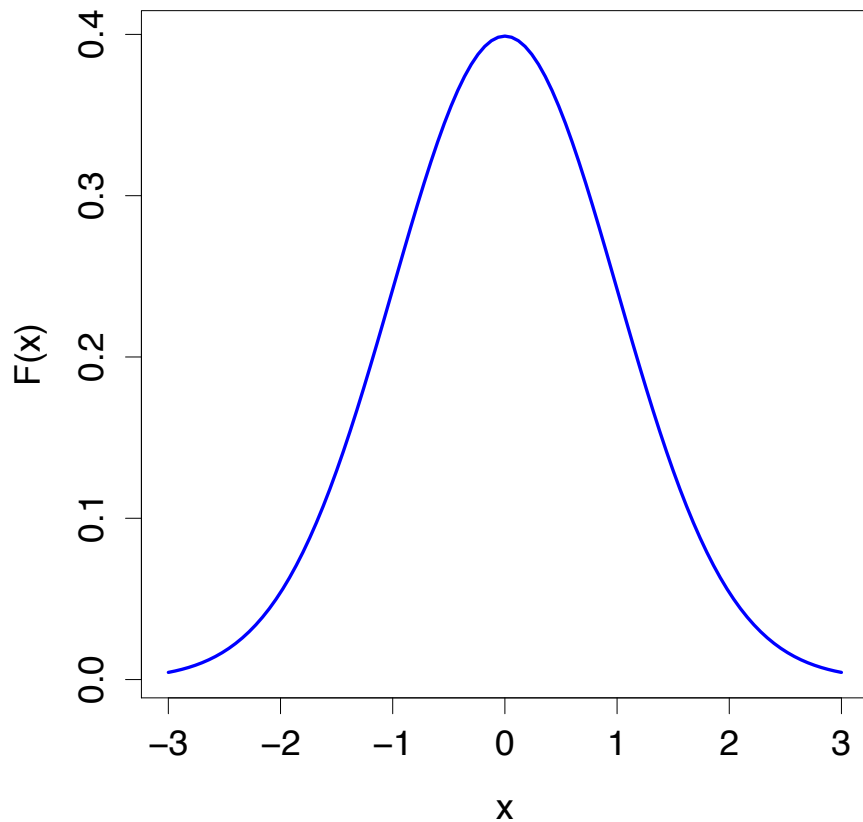
- A {requirement, assumption} in numerous settings is that the data are IID: **I**ndependent and **I**dentically **D**istributed.
- **Identically Distributed**: a set of observations (events) are from the same population (that is, they have the same underlying probability distribution)
 - E.g. a t-test assumes that under the null, all observations come from the same normal distribution
- **Independent**: all samples satisfy the condition $P(A,B) = P(A)P(B)$ where A and B are events (without loss of generality for any number of events) – that is, the joint probability is the product of the individual event probabilities.

Violations of Independence

- Toy example: imagine executioner is using just one coin, but each toss breaks a piece off of the side that landed down.
- Experimental design is in part about trying to avoid unwanted dependence
- Example of a violation that we will encounter later in the course: **batch effects**.

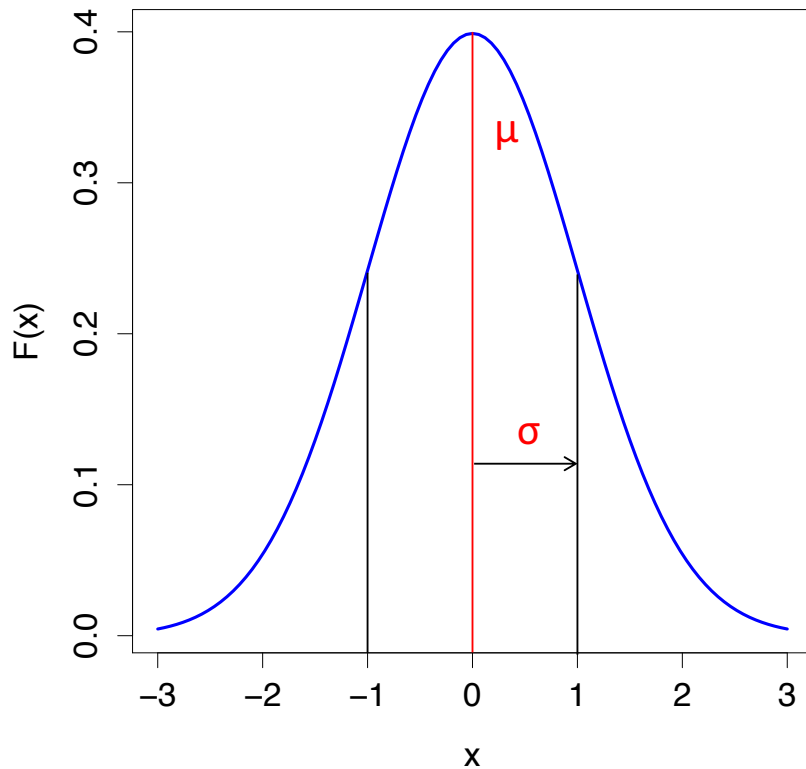
Parameter Estimation

Example: Estimators for normally distributed data



What two parameters define a normal distribution?

Estimators for normally distributed data



What two parameters define a normal distribution?

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / 2\sigma^2}$$

mean = μ

standard deviation = σ

Parameter Estimation

- **Estimator:** Statistic whose calculated value is used to estimate a parameter, θ
 - ***statistic*:** a single measure/scalar that is computed from data (thus it's an RV itself)
- **Estimate:** A particular realization of an estimator
- Types of estimates:
 - **Point estimate:** single number, most plausible value of θ , given the data
 - **Interval estimate:** a range of numbers (aka confidence interval) that informs us about the quality of the estimate

Estimators for normally distributed data

- Given a sample from a normally distributed population, what estimator would you use for μ , σ ?

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

Estimator for normally distributed data

- Let's say that we collected a sample from our normal looking population.
- We estimated the mean from our sample, but how good is the estimate?
- What would it depend on?

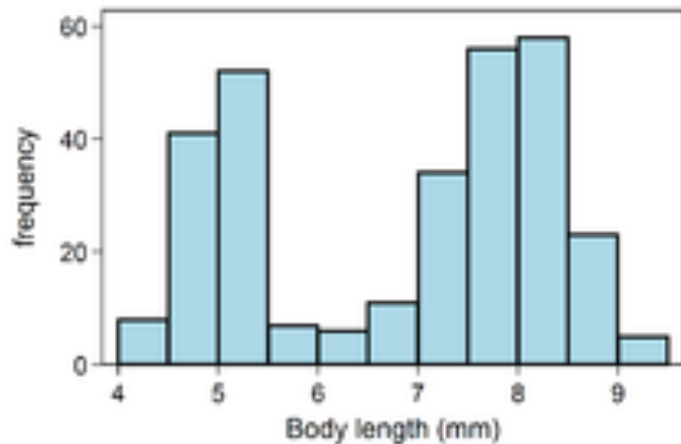
Estimator for normally distributed data

- Let's say that we collected a sample from our normal looking population.
- We estimated the mean from our sample, but how good is the estimate?
- What would it depend on:
 - Sample size
 - Variability of the population (hence variance)

Central Limit Theorem

- Any function (statistics) of a data/observations is a random variable
- Thus, any statistic, because it's random, has a probability distribution function – this is called the sampling distribution
- Let's focus on the sampling distribution of the sample mean

Sampling distribution of the Sample Mean



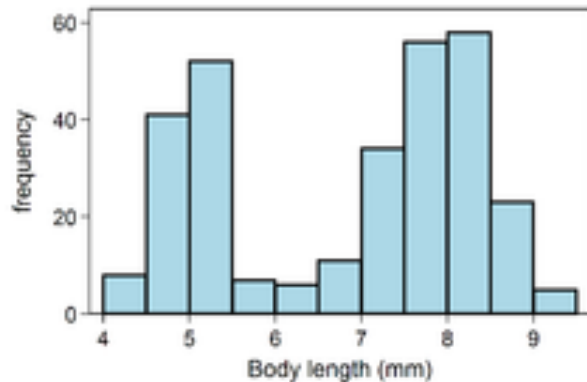
We have a population distribution.

Let's take random samples from it

$$\bar{x}_1 = 7.54 \quad \text{Sample 1}$$

$$\bar{x}_2 = 9 \quad \text{Sample 2}$$

Sampling distribution of the Sample Mean

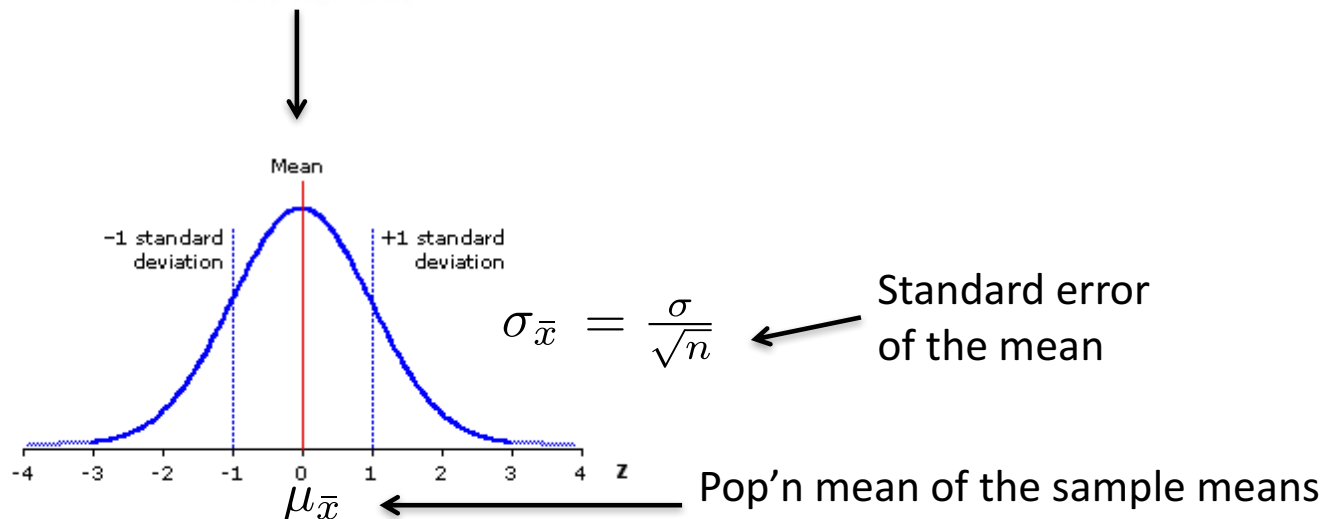


A population distribution

$\bar{x}_1 = 7.54$ A sample mean

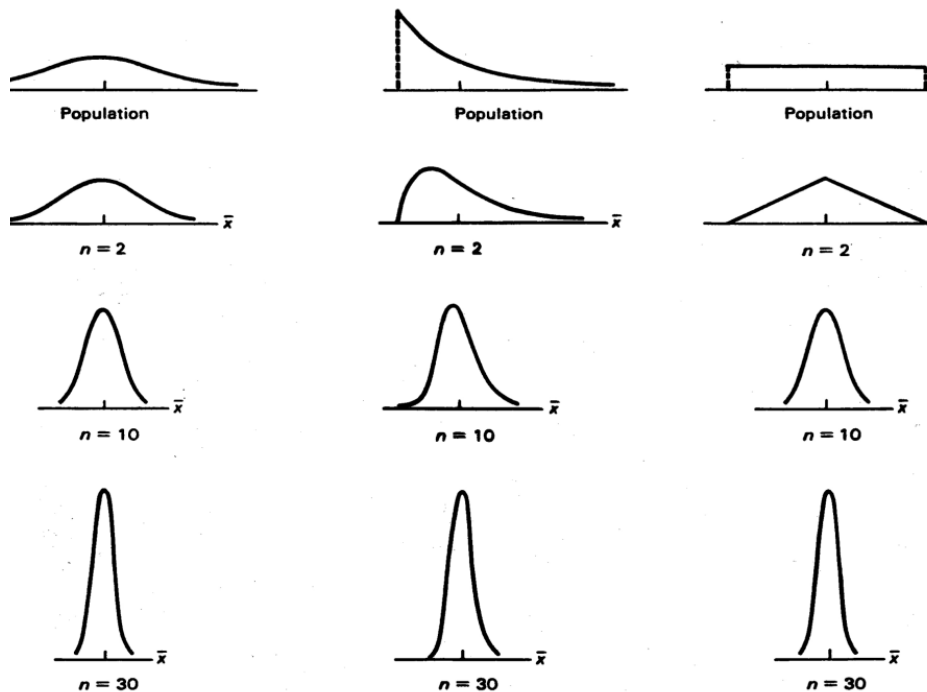
$\bar{x}_2 = 9$ Another

... Many more



Central Limit Theorem

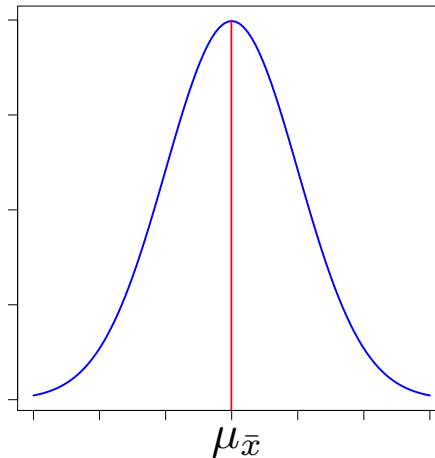
Let X_1, X_2, \dots be an iid random sample from some population with non-normal distribution. If the sample size is sufficiently large, then the sampling distribution will be normal.



$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sampling distribution

- Recall that the sample mean, \bar{x} is an RV, and hence has an associated distribution
- By CLT, the sampling distribution of the mean is normal:

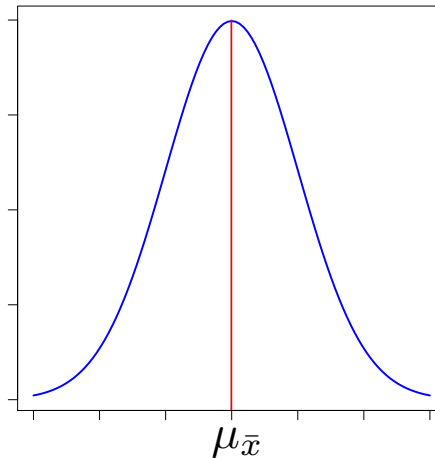


$$\mu_{\bar{x}} = \mu = \bar{x}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

Sampling distribution

- Recall that the sample mean, \bar{x} is an RV, and hence has an associated distribution
- By CLT, the sampling distribution of the mean is normal:



Standard Error (SE)

$$\mu_{\bar{x}} = \mu = \bar{x}$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

An arrow points from the text "Standard Error (SE)" to the variable s in the second equation.

Standard Error of the Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

- SE is the standard deviation of the sampling distribution of the mean
- Often get's confused in literature as “standard deviation” – pay attention to this, given that SE is smaller than SD..
- SE reflects the uncertainty about where the population mean be located, given a sample.
- When sample size ~ 30 , then the normal distribution is a good approximation for the sampling distribution of the sample mean. With smaller samples, the SE $\frac{s}{\sqrt{n}}$ is an underestimate.

the central limit theorem

common sense “statement”:

the sampling distribution for the average of a large, iid sample will be approximately a normal distribution

5.8 Theorem (The Central Limit Theorem (CLT)). *Let X_1, \dots, X_n be IID with mean μ and variance σ^2 . Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then*

$$Z_n \equiv \frac{\bar{X}_n - \mu}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0, 1)$. In other words,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Hypothesis testing

- “A statistical test examines a set of sample data, and on the basis of the expected distribution of the data, leads to a decision about whether to accept hypothesis underlying the expected distribution or to reject the hypothesis and accepts an alternative one.”
- Motivating example: given the expression level of gene A in some disease (case) and some healthy (control) samples, determine if gene A is differentially expressed in disease vs. healthy.
- **Mutually exclusive** hypotheses:
 - H_0 : the expression level is the same.
 - H_A : the expression level is different.

$$Y_1, \dots, Y_i, \dots, Y_{n_y} \sim \text{iid } F$$

$$Z_1, \dots, Z_i, \dots, Z_{n_z} \sim \text{iid } G$$

testing

Observe data $(Y_1 = y_1, \dots, Y_i = y_i, \dots, Y_{n_y} = y_{n_y})$ and
 $(Z_1 = z_1, \dots, Z_i = z_i, \dots, Z_{n_z} = z_{n_z})$.

Does $F = G$? OK, I'll settle for ...

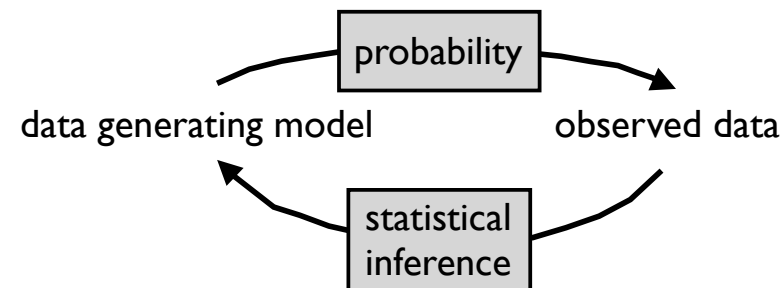
does $E_F(Y) = \mu_Y = \mu_Z = E_G(Z)$?

Call this statement the null hypothesis H_0 :

$$H_0 : \mu_Y = \mu_Z$$

Or, equivalently:

$$H_0 : \mu_Z - \mu_Y = 0$$



Steps in hypothesis testing

1. Collect some data (i.e., sample)
2. Ask a precise and answerable question which has a mutually exclusive yes/no answer.
3. Define a test-statistics that corresponds to the question. You typically know the expected distribution of the test-statistics under the null.
4. Compute the p-value associated with the observed test-statistics under the null distribution.

Hypothesis testing: T-test

- Expression level of gene A measured for n disease and m healthy samples:

$$- z_1, z_2, \dots, z_n \text{ \& } y_1, y_2, \dots, y_m$$

- Unpaired t-test with equal variance :

$$t = \frac{\bar{z} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

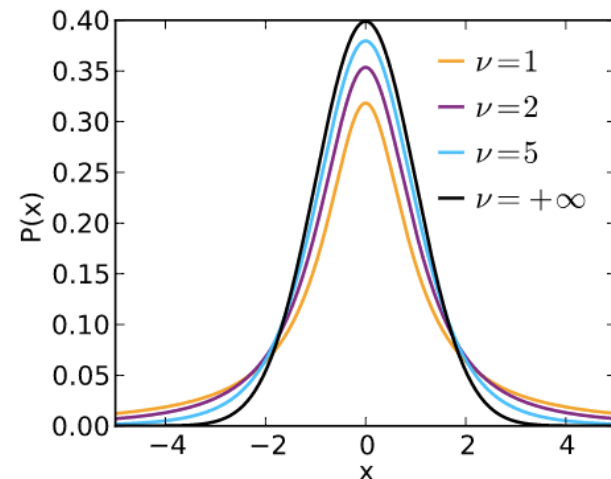
$$S_p^2 = \frac{\sum_{i=1}^n (z_i - \bar{z})^2 + \sum_{i=1}^m (y_i - \bar{y})^2}{(n-1) + (m-1)}$$

Hypothesis testing: T-test

- From theory we know the distribution of our test-statistics, if we are willing to make some assumptions:
 - Assuming normal distribution for X and Y , with equal variance then

$$t \sim t_{n+m-2}$$

Degrees of freedom



Hypothesis testing: T-test

- Plug in the observed t-statistic to find the probability of observing a value as larger or larger than the one observed.

e.g., t-stat = 2 ; two sided

