STAT 540
Class meeting 02
Wednesday, January 7, 2015

Dr. Gabriela Cohen Freue
Department of Statistics
(Preparation by Dr. Jenny Bryan)

Introduction to statistical inference, part 1
The Big Picture and some need-to-know probability

Course webpage:
http://stat540-ubc.github.io

Check the website for further information about:

- Instructors and TAs
- Lectures
- Seminars
- Announcements

Computing seminar kicks off today!

ESB 1042 (primary room) and 1046 (if needed), on main floor of this building

11am - 12pm: come work through R / RStudio installation and more; test drive it all where someone can hear you scream; Intro to Git(Hub); Exploration of small dataset

12pm - 1pm Crash course in molecular biology/ genetics

On your own this week: keep working on materials listed on webpage for seminar 01

First few seminars will have substantial overlap with some STAT 545A content; alums may wish to work through the STAT 540-specific bits independently (or come help others!)

Quick index into STAT 545A content:
http://stat545-ubc.github.io/topics.html

Today's lecture and Monday's may not offer much for the STAT students ... you have been warned

“Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid.”

“Rigor and clarity are not synonymous.”

-- Larry Wasserman in preface of “All of Statistics”

You are a prisoner and the only way to save your life is to work a math problem.

You can pick one of two related problems.

Here they are ....

There is a coin.

Problem #1

It comes up heads with probability $p_H = 0.5$.

The Executioner is going to conduct 10,000 trials, where each trial = counting the number of heads in 10 “regular” flips of the coin.

You must guess what proportion of the 10,000 trials will have outcome 7.

Let $p_{☹}$ be the difference between your guess and the actual observed proportion.

You will be executed with probability $p_{☹}$.

Problem #2

The Executioner is going to tell you the outcome of 10,000 trials, where each trial = counting the number of heads in 10 coin tosses.

You must describe the coin(s) and toss(es).

Let $p_{☹}$ be like so: If no difference between your description and the truth, then $p_{☹} = 0$. As difference grows, $p_{☹}$ tends to 1.*

You will be executed with probability $p_{☹}$.

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

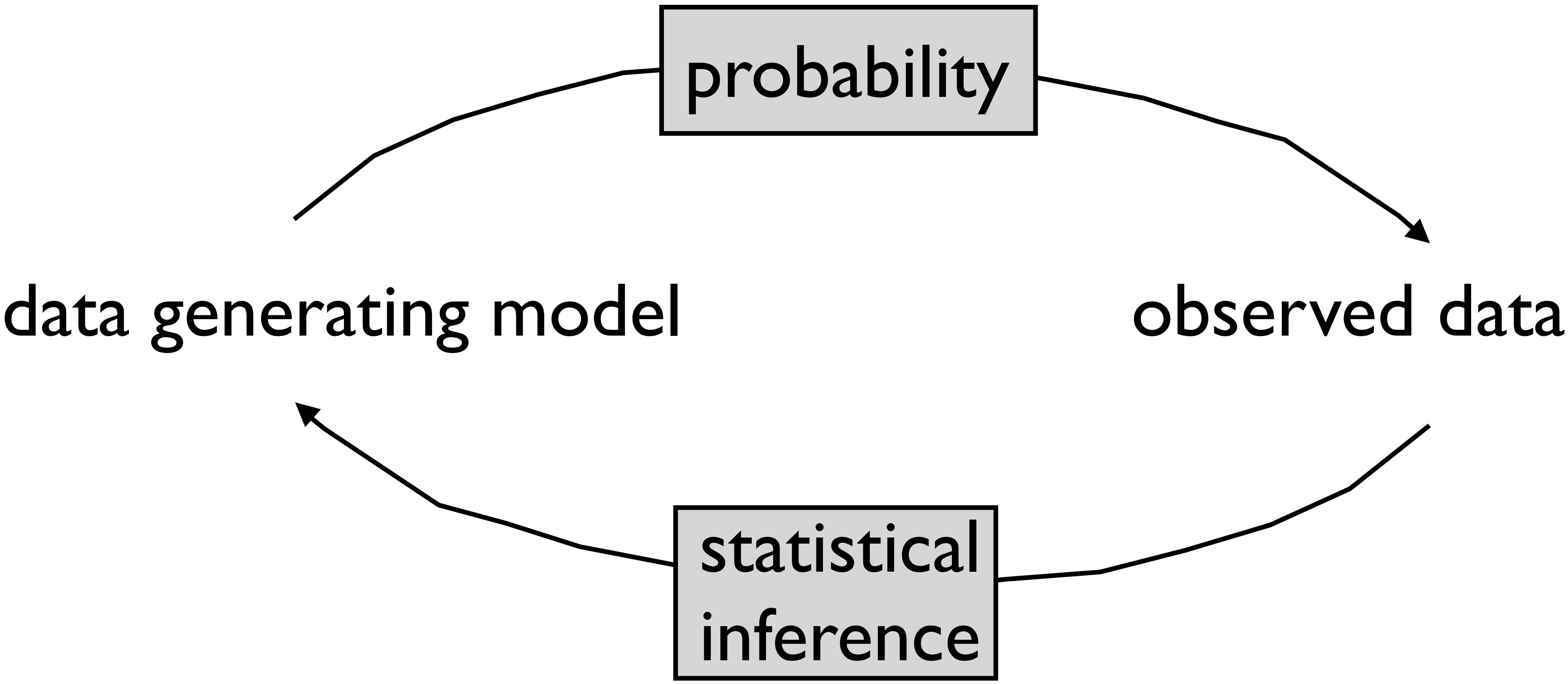
You will have ~30 mins to work the problem.

You will have a some basic computation and graphing capability but no internet, life lines, etc.

Which do you choose?

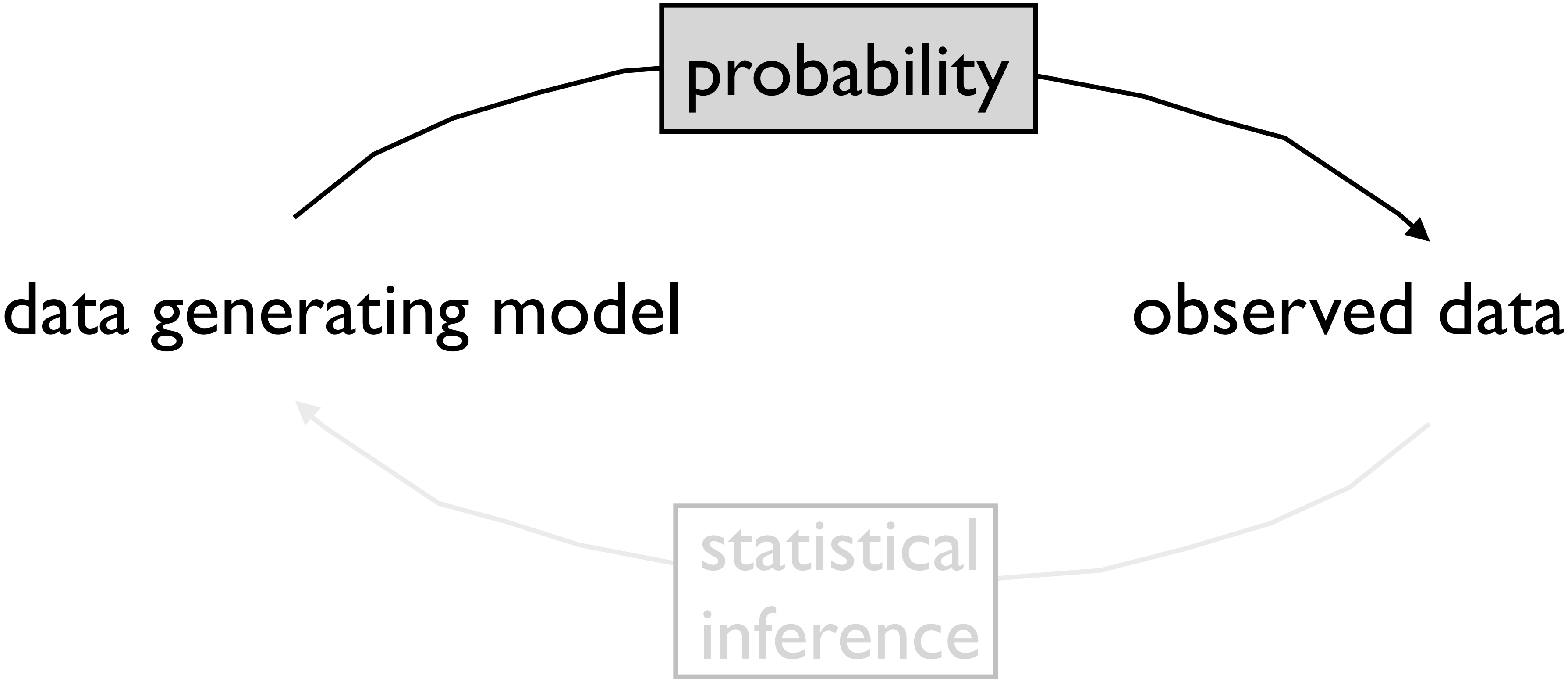
Why?

<lively discussion>



Adapted from Figure 1 of "All of Statistics".

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



probability

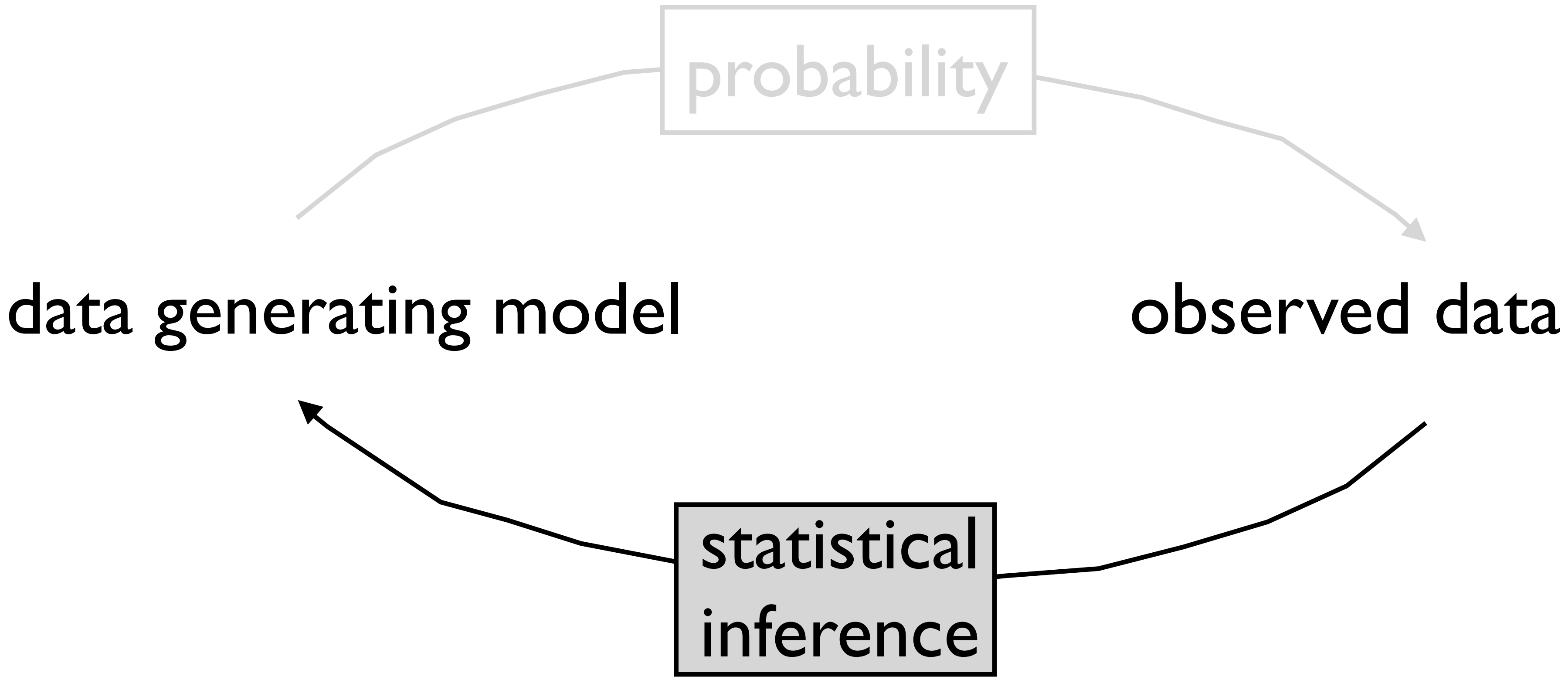
data generating model

observed data

statistical inference

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

“Given the observed data, what can we say about the model that generated the data?”



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

In statistical inference, it often feels like we are **working a math problem backwards**.

“Here’s a steaming pile of messy data .... where do you think it came from?”

It is often uncomfortable because you almost never know everything you need to know to “get it right”.

So we make simplifying assumptions, take intelligent guesses, do as many sanity / consistency checks as possible, and hope for the best!

“Math Solution” to Problem #1

$$X \sim Bin(n, p)$$

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

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

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

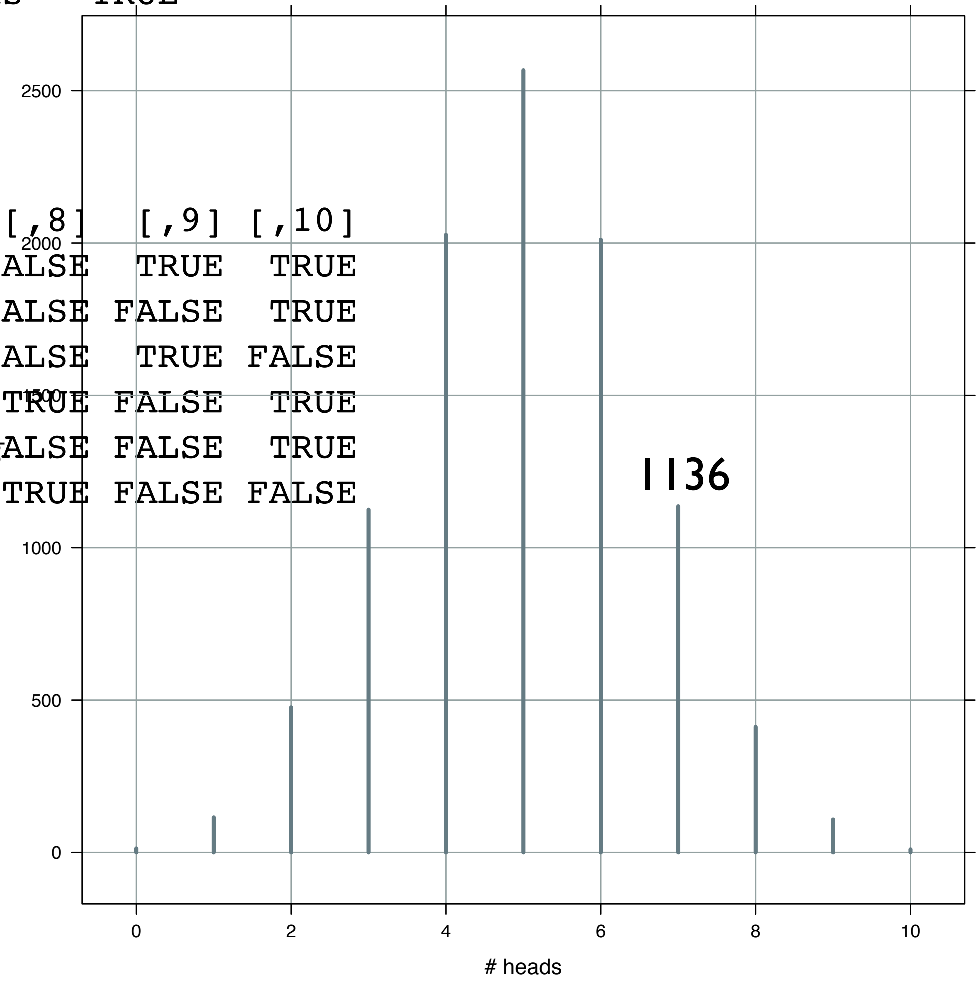
I'd guess that 1172 out of 10000 trials have an outcome of 7 heads.

```
> 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 Problem #1

```
> B <- 10000

> coinFlips <- runif(n * B) > 0.5          # heads = TRUE

> 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 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
```



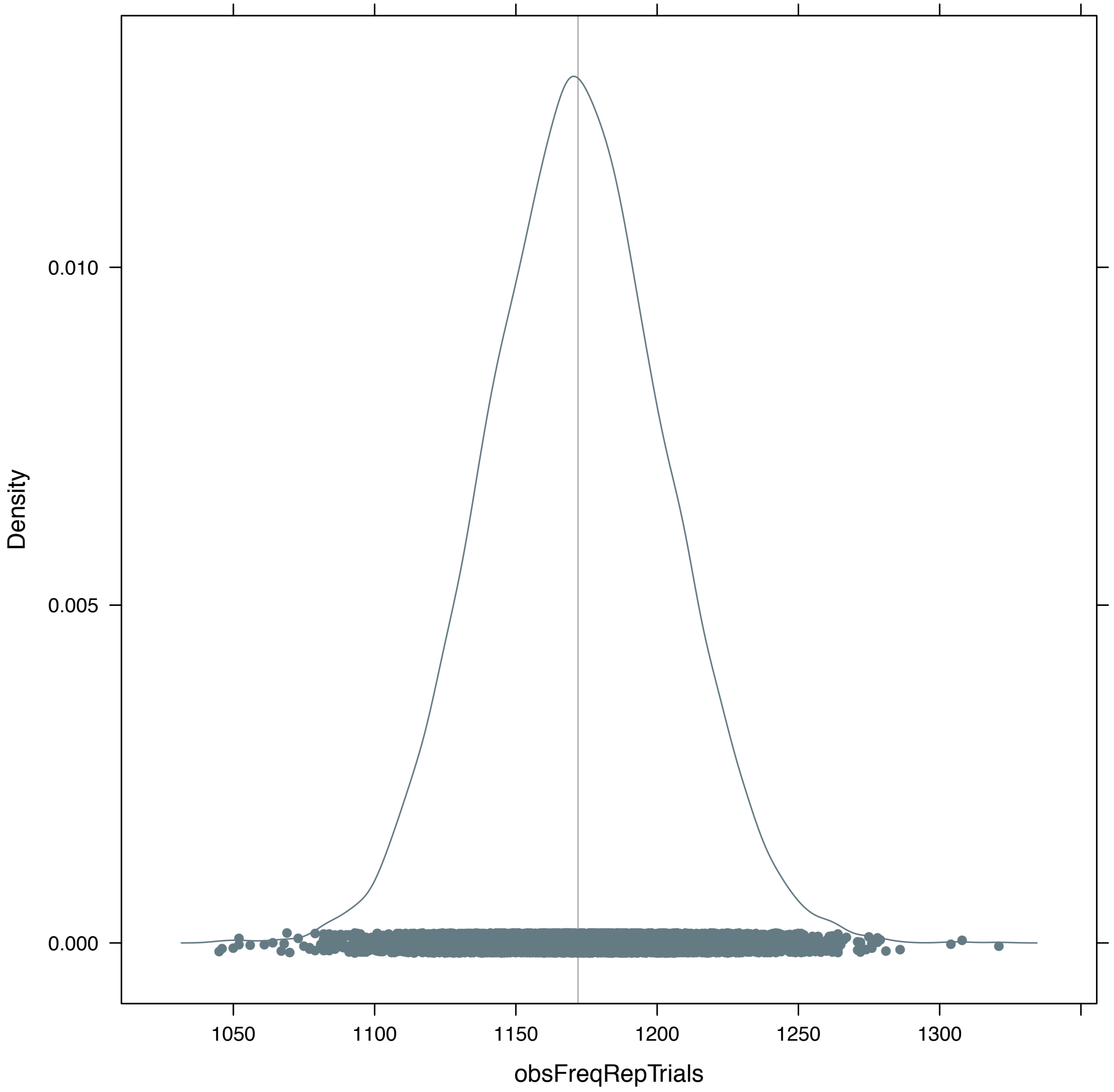
1136

Compare to the actual observed number of 7‘s, which was 1145.

```
> (pSad <- abs(myGuess - obsFreq)/B)
[1] 0.0009
```

Not too bad, as probability of death goes. Happens to outperform 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).



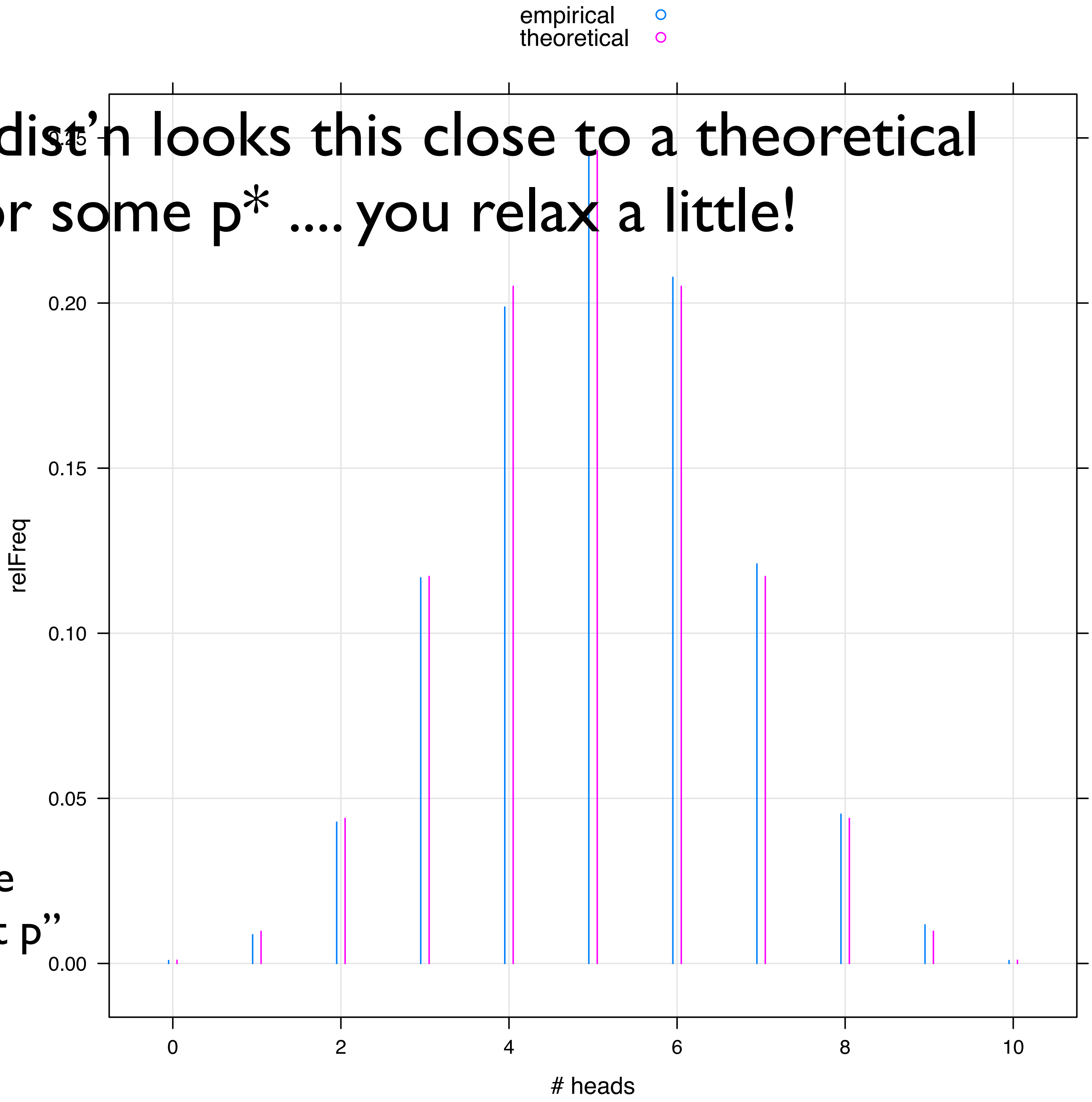
“Solution” to Problem #2

Assuming nothing ... is a probably a death sentence!

You’ll desperately hope that the same coin was flipped in each trial and that the 10 flips in each trail are “regular flips”.

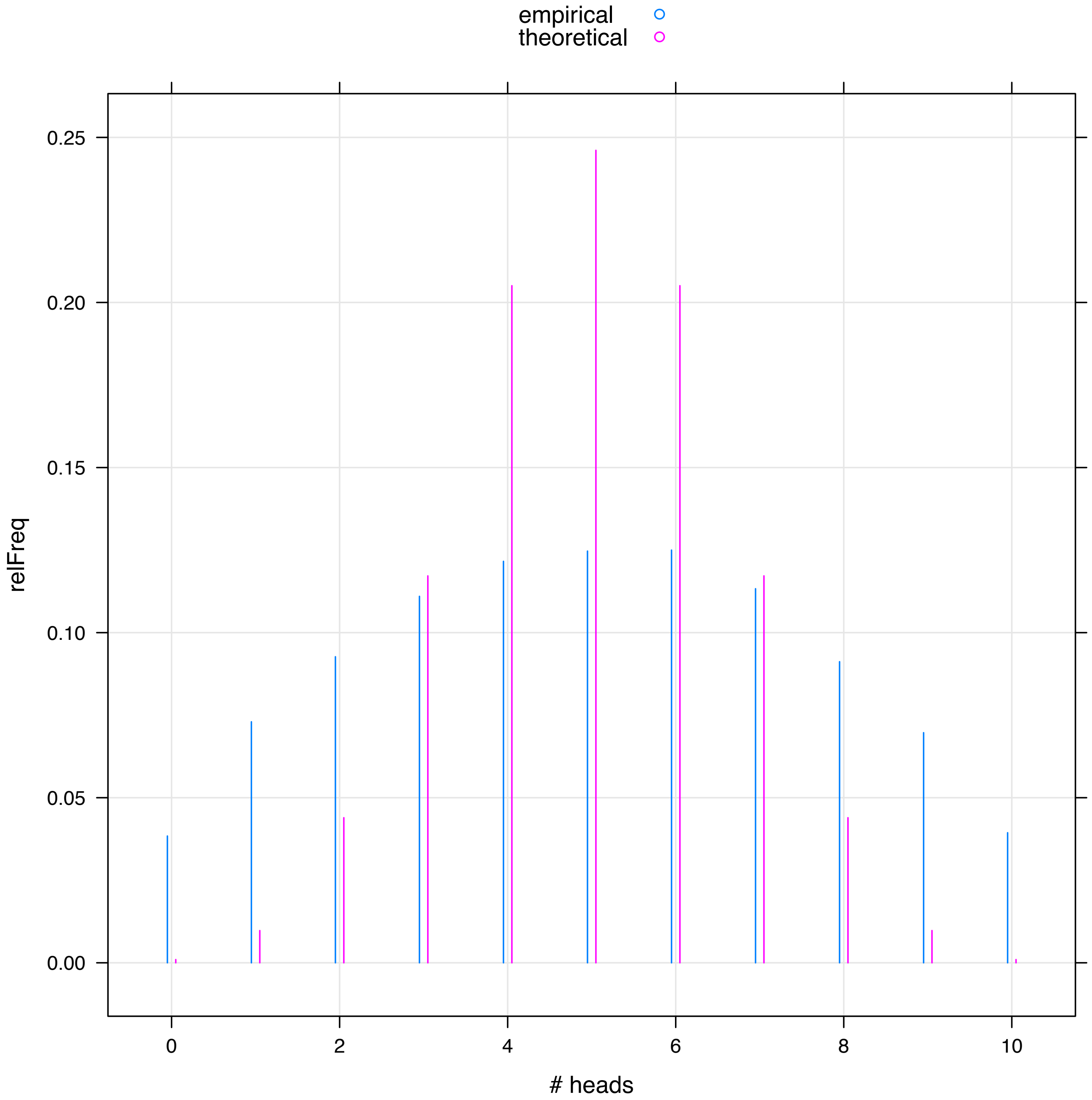
You might inspect the data to see if this is plausible ....

If the empirical dist’n looks this close to a theoretical Bin(n = 10, p) for some p* .... you relax a little!

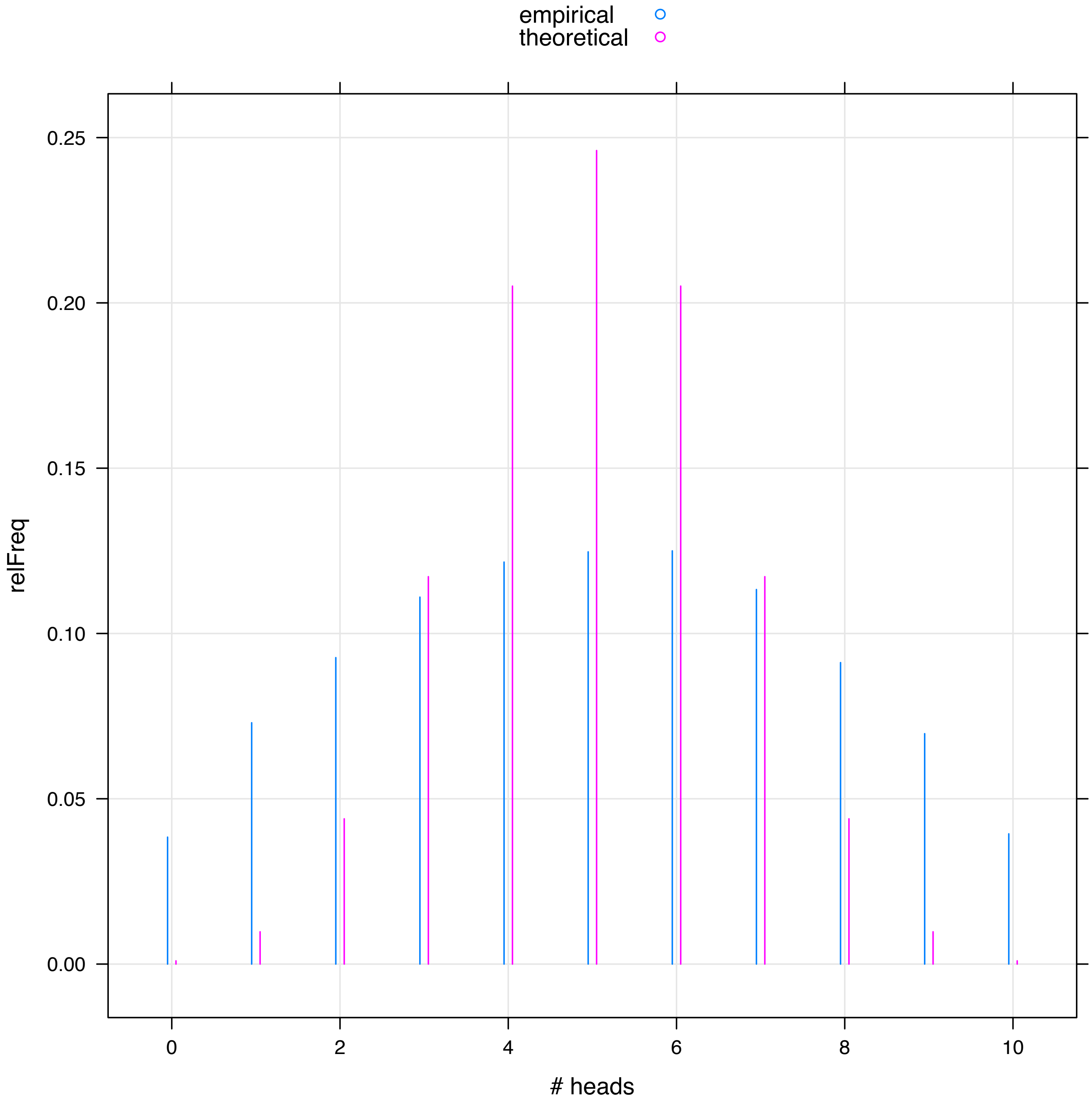


* Here I tried p = 0.5 and one could imagine ways to pick the “best p” from the observed data ....

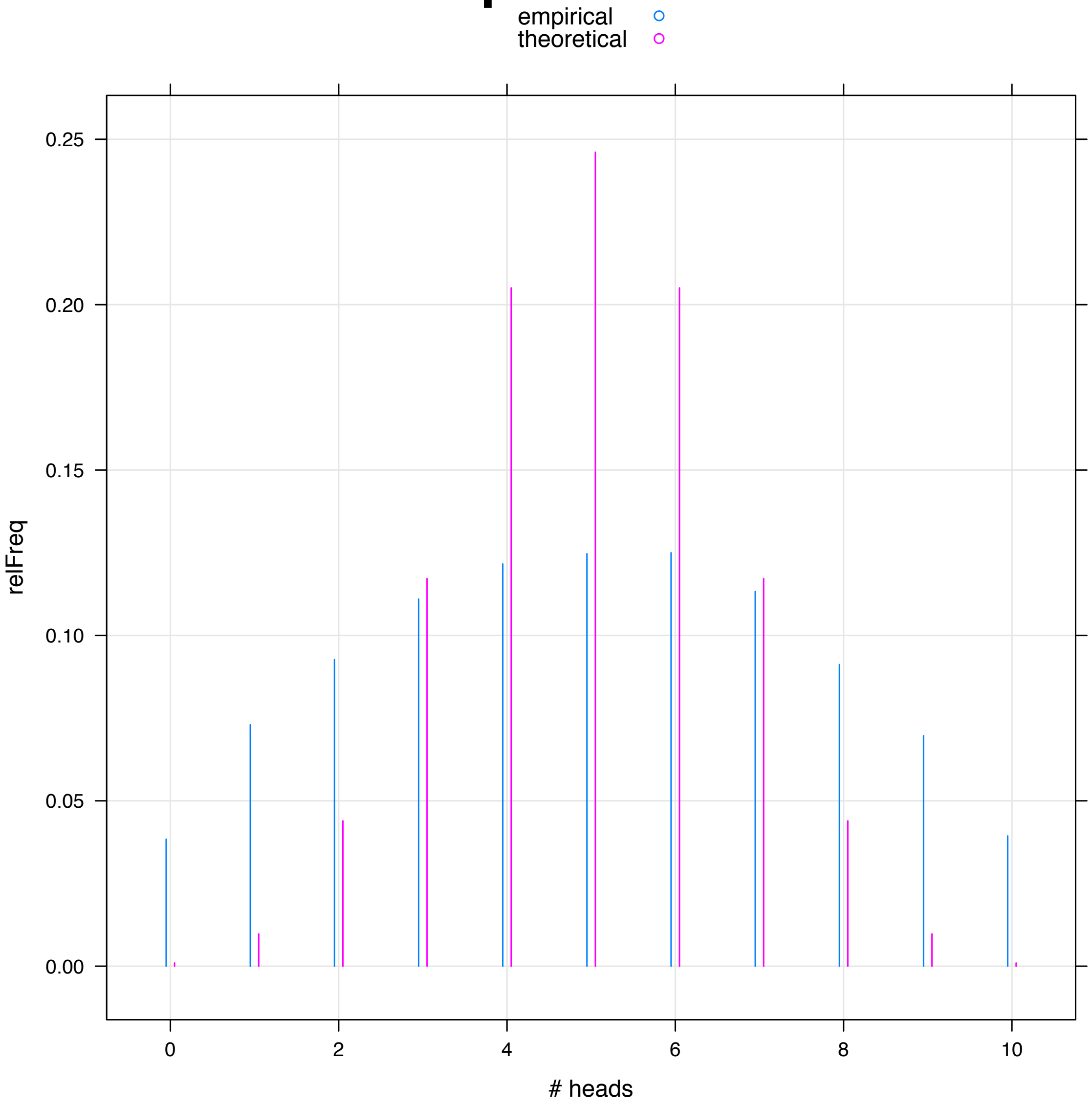
What if the theoretical distribution and the empirical look like this? Does anything concern you?



Our model can't be right. Empirical distribution much more "spread out".



Bottom line: though the empirical dist’n and theoretical dist’n have the same “typical” value -- which is 5 -- the empirical dist’n has WAY more variability than the binomial can “explain”.



“Solution” to Problem #2, cont’d

If data inspection is comforting, you might make the “default” assumption of one coin, “regular flips” ...

Then you just need to pick the value of p_H that is “most compatible” with the data.

If the data inspection is troubling, you must consider more complicated alternatives.

Maybe the coins are selected for each trial from some bucket of coins? Maybe you can assume the p’s themselves have some distribution and then try to infer that? Oh dear ....

What I hope the thought experiment has foreshadowed ...

The importance of knowing (or speculating) how the data was collected.

The breathtaking beauty of deliberate experimental design, which helps guarantee things are “plain vanilla”, e.g. same coin, independent tosses & trials.

The unavoidability of making educated guesses in statistical inference -- at best you try to *minimize and characterize your errors*. They can *never* be eliminated.

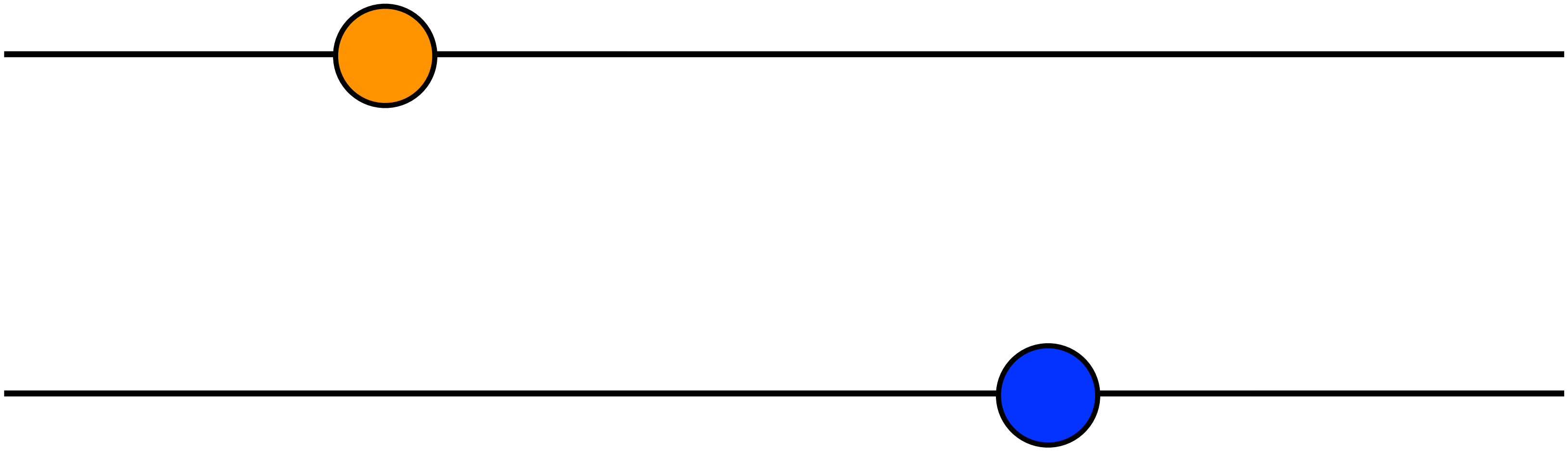
Hallmarks of sophisticated, mature thinking about statistical on inference:

You know there are no "right answers" (but realize there are some "wrong" ones).

You appreciate "statistical significance" as a useful concept but you don't take it too seriously or literally (see above).

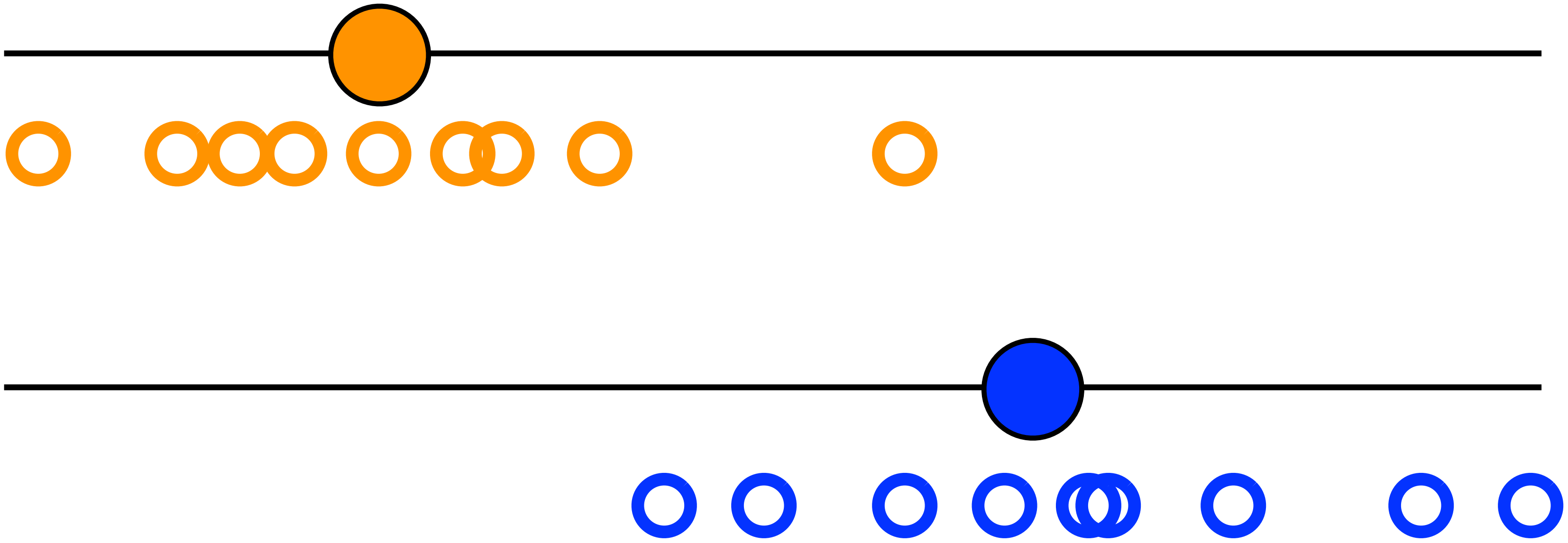
You are always working to get a handle on *variability* -- much more than worrying about the "average" (which is usually quite easy to see).

Does this constitute evidence that the “oranges” are meaningfully different from the “blues”?



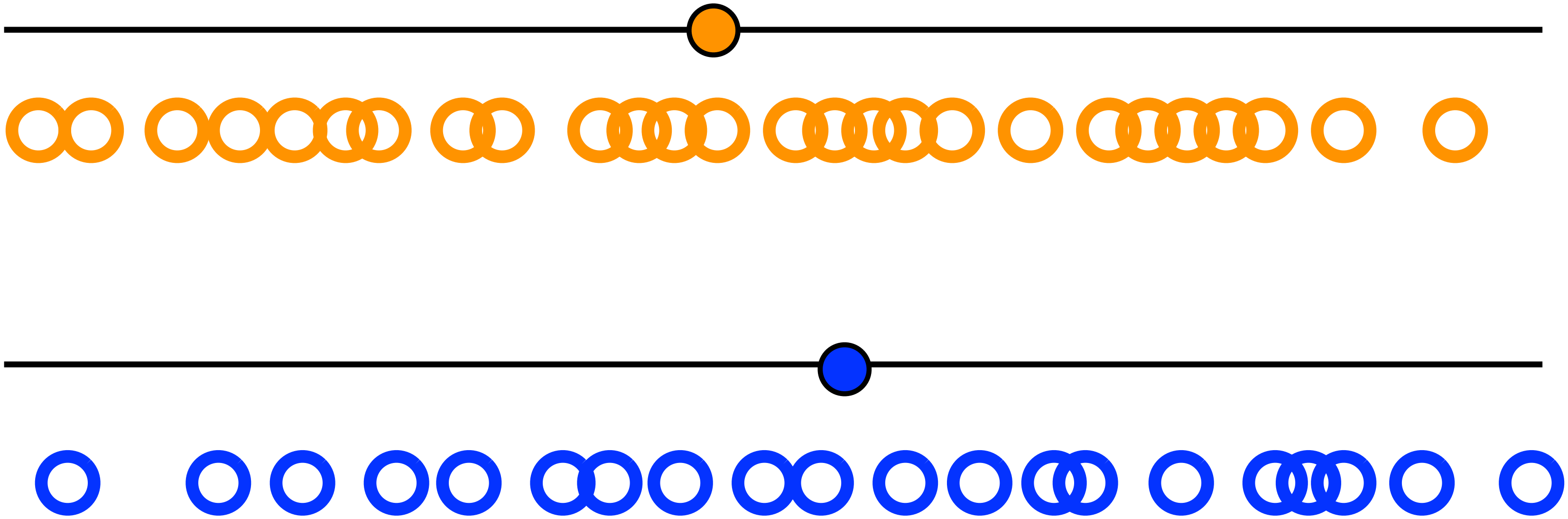
Orange circle = average of orange observations
Blue circle = average of blue observations

Does this constitute evidence that the “oranges” are meaningfully different from the “blues”?



Yeah, pretty compelling evidence to me.

Does this constitute evidence that the “oranges” are meaningfully different from the “blues”?



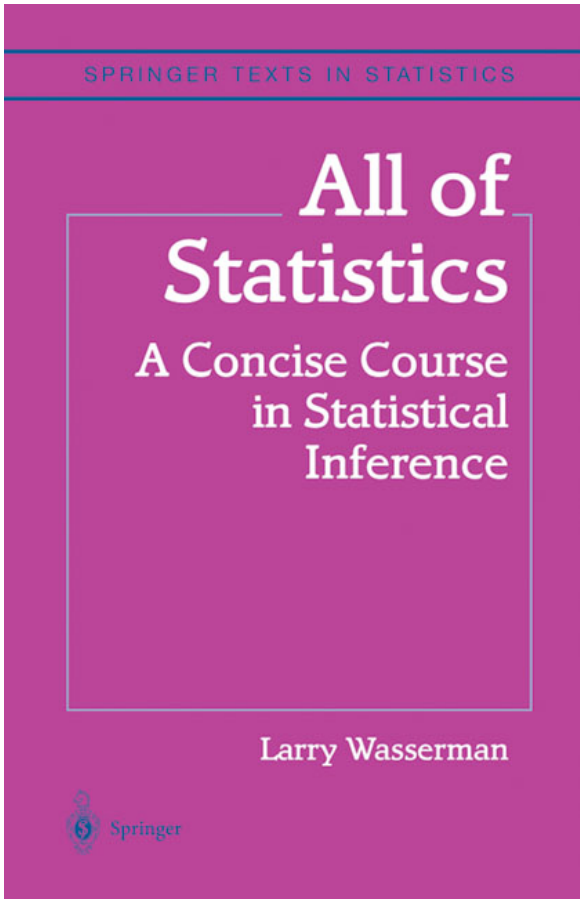
No, not so much.

Even if it’s “statistically significant”, is it big enough to matter in the orange / blue subject area?

More on the fundamental importance of *variance* in statistical inference

Understanding the "background variability" or "white noise" in some observable quantity is what allows us to attach meaning to observed phenomena

A priori, hard to know if an observed difference is "real" .... but knowing that the difference is 3 times the typical "spread" certainly suggests it might be!



SPRINGER TEXTS IN STATISTICS

All of Statistics

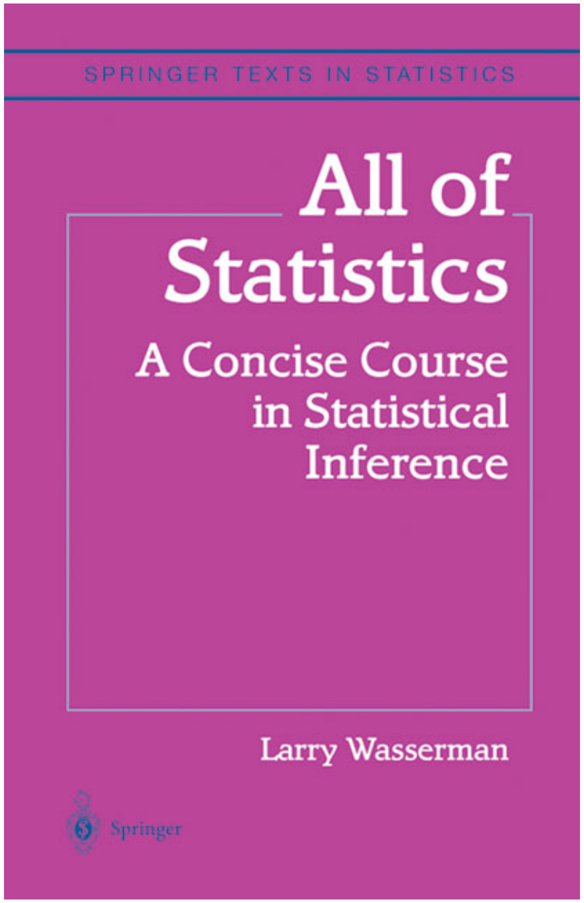
A Concise Course in Statistical Inference

Larry Wasserman

Springer

At a higher mathematical level than our course, but still relevant.

Goal: getting motivated, mature students up to speed on the core concepts of modern statistics.


SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer

Available via SpringerLink!

Other idiosyncratic recommendations:

Introduction to probability and statistics for engineers and scientists By Sheldon M. Ross
Google ebooks link (he seems to have several books that are intro stats)

Mathematical Statistics and Data Analysis
by John Rice
amazon link

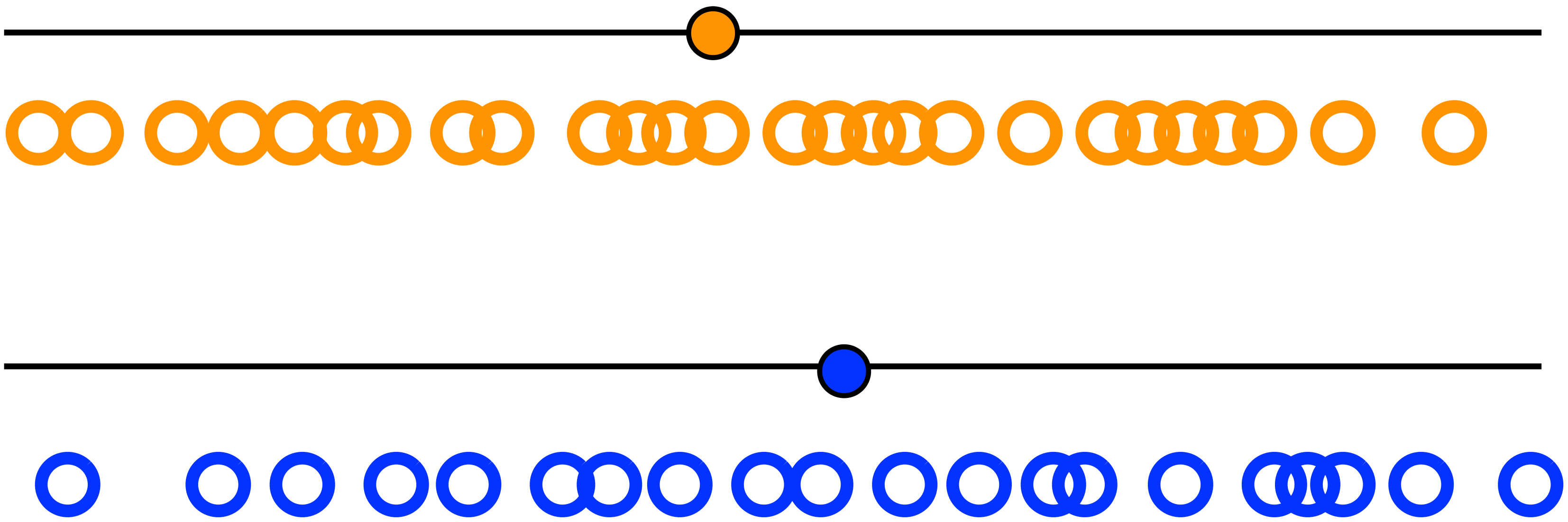
Both of these books have widely varying 'reviews', so your mileage may vary ....

http://www.xtranormal.com/watch/6878253/

that sort of complete failure to communicate and get science done is what we’re trying to avoid here

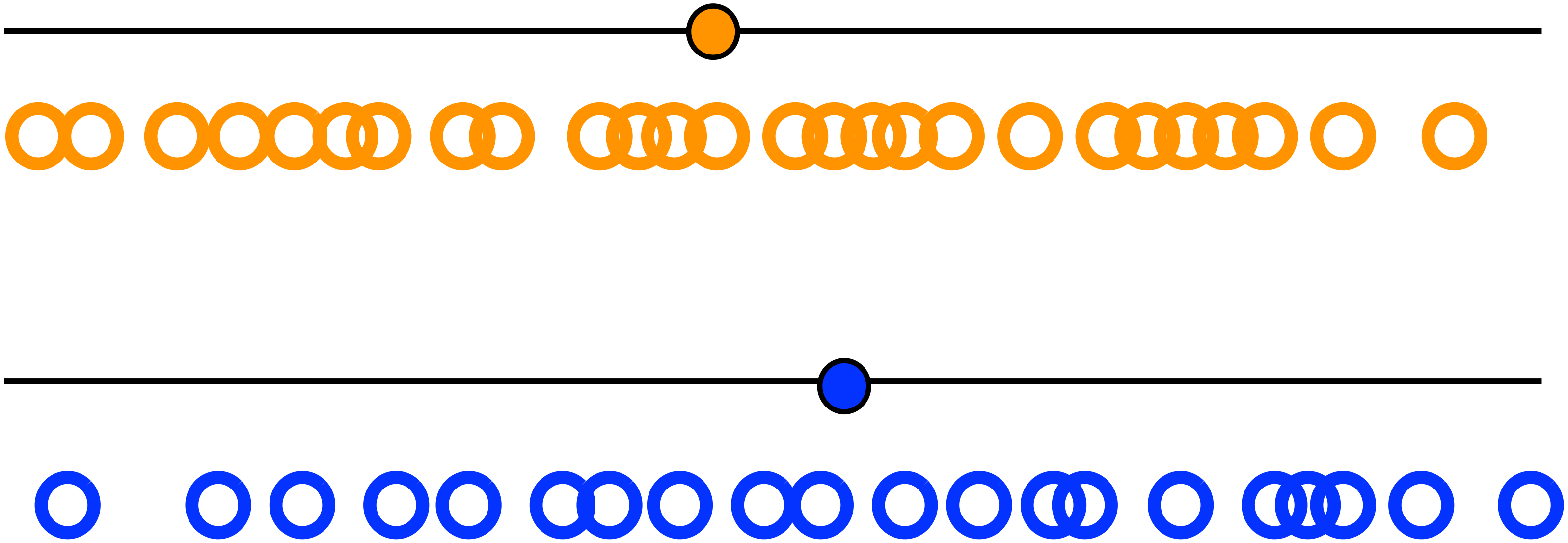
will review key concepts in probability and basic statistical inference on a “need to know” basis

I want you to feel comfortable with the standard statistical approaches for saying whether the oranges and blues are truly different. That's our near term goal.



first we blitz through fast
then we go back to review key concepts, notation, jargon

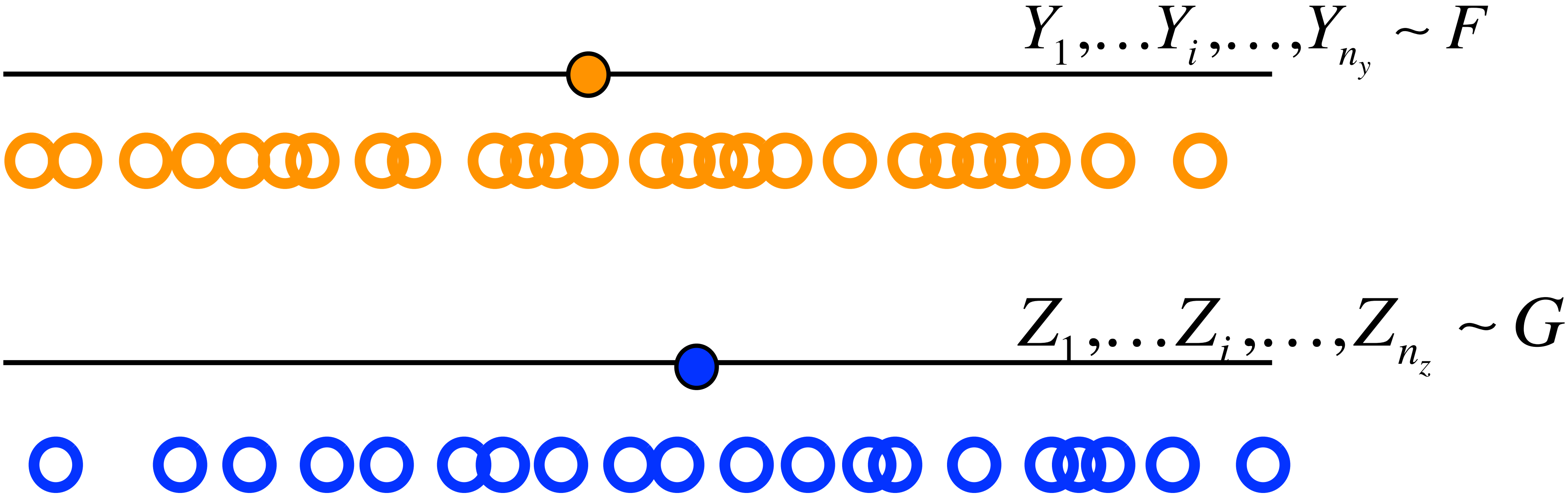
Observe data $(Y_1 = y_1, \ldots Y_i = y_i, \ldots Y_n = y_{n_y})$ and $(Z_1 = z_1, \ldots Z_i = z_i, \ldots Z_n = z_{n_z})$.



Regard the data as iid observations of random variables that have certain (unknown) distributions.

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

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



$$Y_1,\ldots Y_i,\ldots,Y_{n_y} \sim F$$

$$Z_1,\ldots Z_i,\ldots,Z_{n_z} \sim G$$

Ask a precise, answerable question.

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

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

Pick one answer -- usually the boring one -- and call it the null hypothesis H_0:

$H_0 : \mu_Y = \mu_Z$

Or, equivalently:

$H_0 : \mu_Y - \mu_Z = 0$

$$t = \frac{\bar{Y}_{n_y} - \bar{Z}_{n_z}}{s_{\bar{Y} - \bar{Z}}}$$

Pick a "test statistic" -- here I show the two sample t test statistic -- for which we know its distribution under H_0: $\mu_Y = \mu_{Z.}$

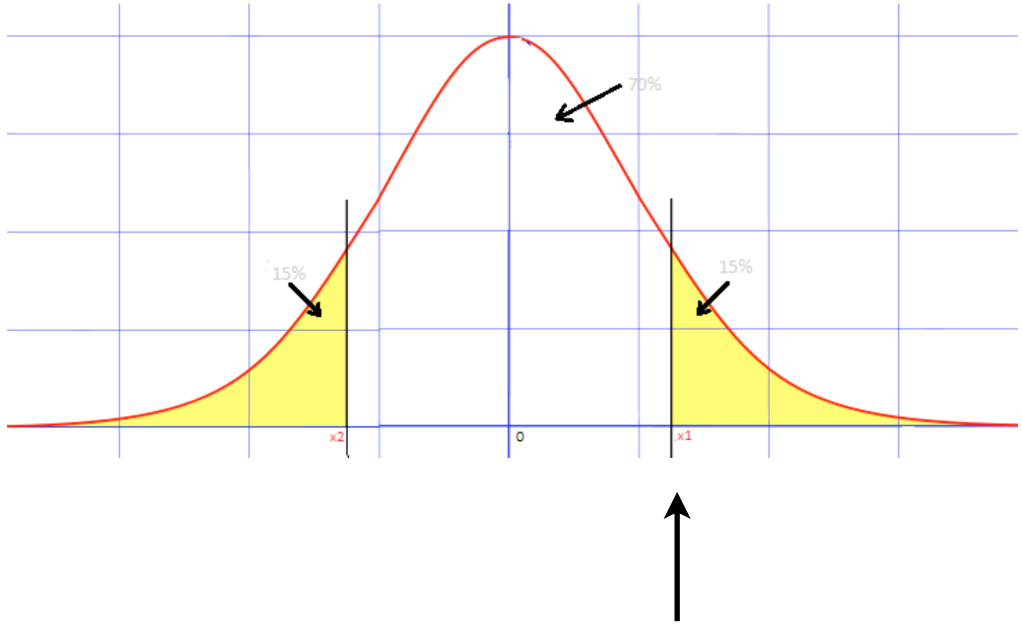
In this case, theory tells us that $t \sim t_{ny + nz - 2}$

Compute the actual observed value of test statistic t and convert to a p-value, the probability of seeing a value as or more extreme than the observed.

$$\text{p-value(obs. test stat.)} = P(|\text{test statistic rv}| \geq \text{obs. test stat.})$$

imagine this is a t distribution with $ny + nz - 2$ degrees of freedom

sum of the yellow areas = p-value



obs. value of test stat.

http://www.tutorvista.com/math/estimating-with-confidence

Partial inventory of things we need to review

random variable and its distribution

iid

parameters of a distribution

an estimator of a parameter

a parameter space

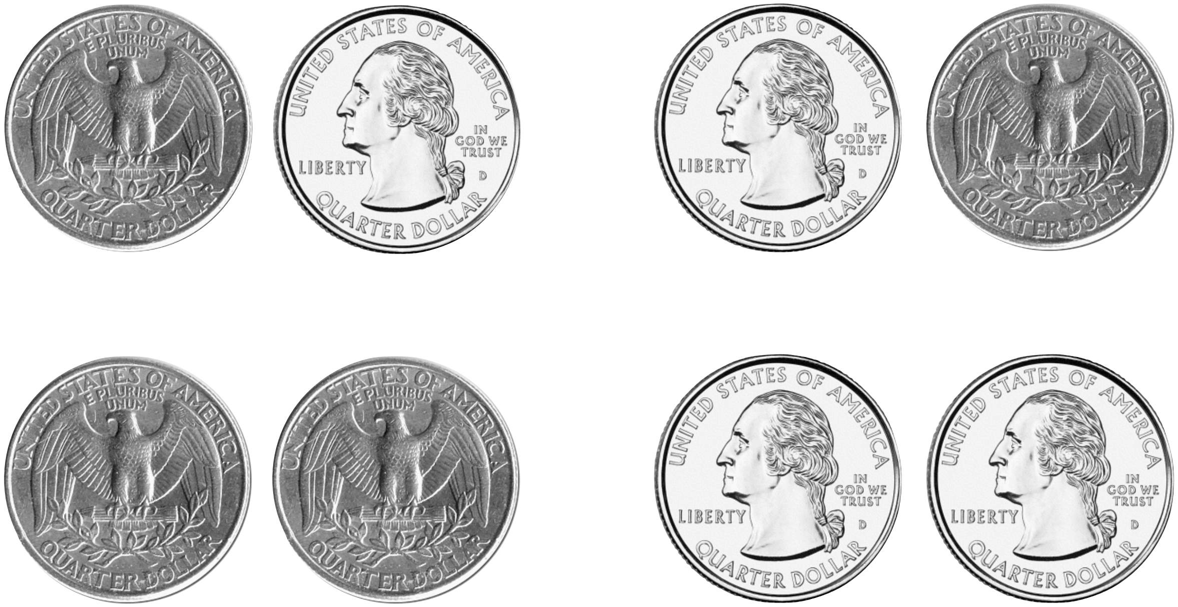
null and alternative hypotheses

the sampling distribution of an estimator

large sample results for averages

random experiment: toss a fair coin 2 times

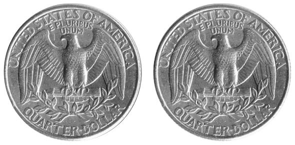
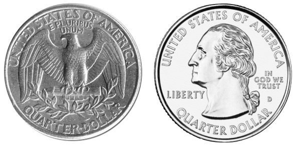
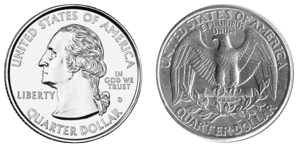
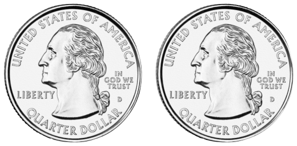
outcome space, sample space = all possible outcomes of the experiment



Flipping coint photo by Chris Sharp via http://www.freedigitalphotos.net/
Quarter sides from http://www.teacherfiles.com/clip_art_us_money.htm

random variable = a function that maps outcomes of the experiment into a real number

ω = an outcome of the experiment
$X(\omega)$ = number of heads

ω	$X(\omega)$
	0
	1
	1
	2

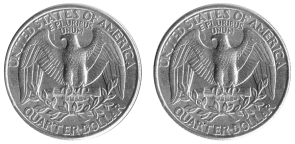
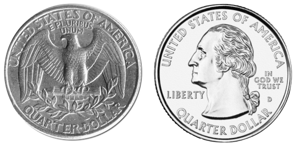
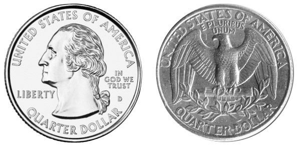
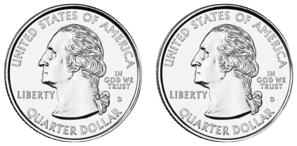
reality check:

in real life, especially this class, the random experiment (e.g. coin toss) and associated outcome space will fade from view and real-valued random variables will be the main focus

from now, on I will blur the two and you'll probably follow just fine

ω = an outcome of the experiment

$X(\omega)$ = number of heads

	probability	$X(\omega)$
	0.25	0
	0.25	1
	0.25	1
	0.25	2
	1	

$P(X = x)$ $P_X(x)$	x
0.25	0
0.5	1
0.25	2
1	

notational sidebar:

Capital letters *X*,*Y* etc very popular for rvs

same letter, *but in lower case*, used to represent the outcomes or observed values

This is not a typo, it actually means something:

$X = x$ “the event that rv *X* takes on the value *x*”

So you’ll see things like this, depending on context:

$$P(X = x), P_X(x), P(x), p(x)$$

rare for the outcomes and associated probabilities of a rv to be represented first and foremost as a table of numbers

much more common and elegant:
we distill that into a function, i.e. we have a formula that gives the probability of $X = x$ for arbitrary x

this is what people are referring to when they say "probability distribution"

let's look at some famous distributions

Bernoulli

the “coin toss” distribution
-- or, more generically, a random variable with two outcomes

$$X \sim Bernoulli(p)$$

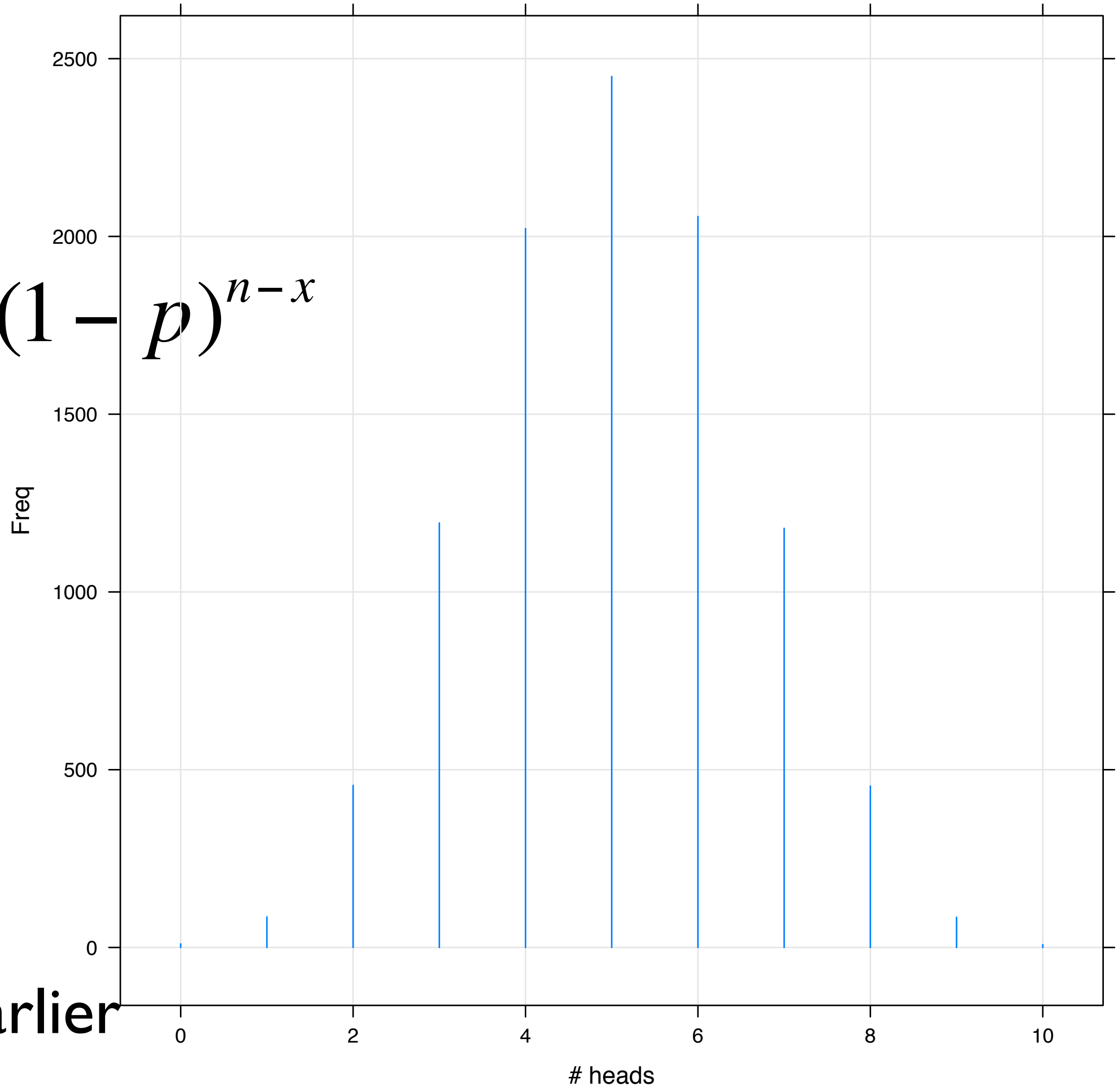
$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

Binomial
the "# of heads in *n* coin tosses" distribution

$$X \sim Bin(n, p)$$

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



Bin(n = 10, p = 0.5) shown earlier

complete specification of the rv's distribution seems to require:

the "family", e.g. Bernoulli, binomial

AND

one or more parameters

$$X \sim Bernoulli(p)$$

$$X \sim Bin(n,p)$$

THE THINGS LISTED BELOW ARE DIFFERENT!

THE DIFFERENCES MATTER VERY VERY MUCH!

random variable, e.g. X

observation of a random variable, e.g. $X = 5$

parameter, e.g. the single-trial success probability p of a binomial rv notational sidebar: meet the “twiddle”

$$X \sim Bin(n, p)$$

Reads like so:

“rv *X* is distributed as a binomial rv ...”

twiddle is NOT being used to evoke “approximation” so far, our examples limited to “discrete” rvs

= rvs that take on a countable number of possible values

but ... we must get more general and talk about “continuous” rvs

= rvs that take on an infinite number of possible values the “probability distribution” is more user friendly for discrete rvs than continuous

for discrete rvs, *P(X = x)* is technically called the *probability mass function*

for continuous rvs, the analagous thing is the *density*

typical ways you’ll see density denoted:
$f_X(x), f(x)$

to the dismay of many, a density does not give you probabilities directly

f(x) is NOT the probability that cont rv *X* takes on the value *x* ... which, by the way, is zero

more "proof" f(x) is NOT a probability: f(x) can be greater than 1 (but still never less than 0)

probabilities can only be obtained from densities by taking an integral

luckily you can often get a computer to do this for you; longer ago, we used tables

famous continuous rvs

uniform

$$X \sim Unif(0,1)$$

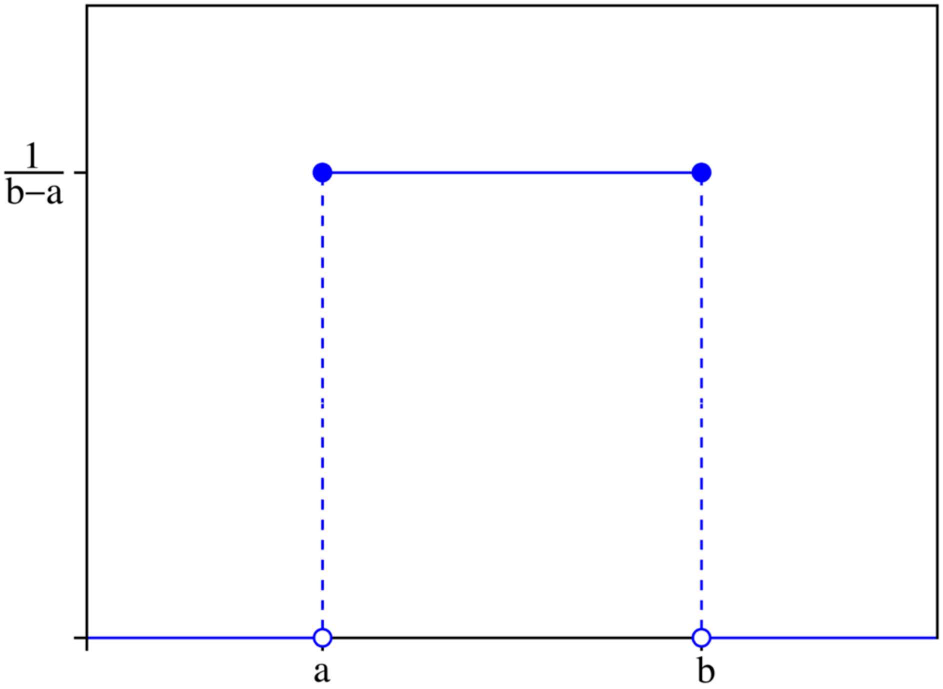
$$f(x) = 1, \text{ for } x \in [0,1]$$

$$f(x) = 0, \text{ otherwise}$$

$$X \sim Unif(a,b)$$

$$f(x) = \frac{1}{b-a}, \text{ for } x \in [a,b]$$

$$f(x) = 0, \text{ otherwise}$$

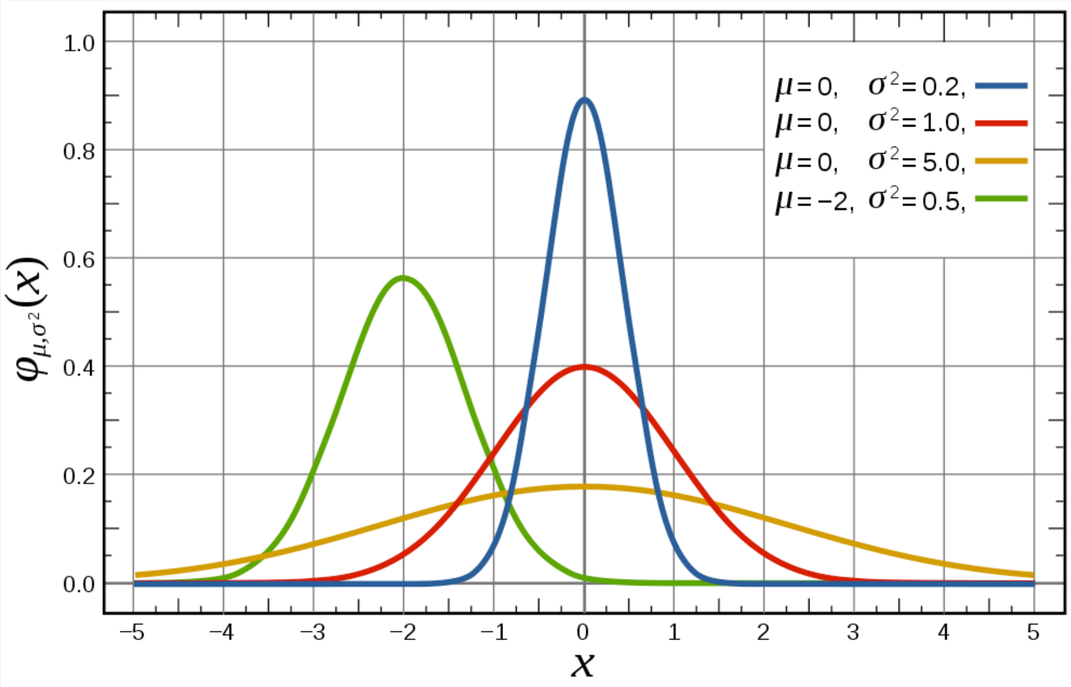


http://en.wikipedia.org/wiki/File:Uniform_distribution_PDF.png

normal, Gaussian

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

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

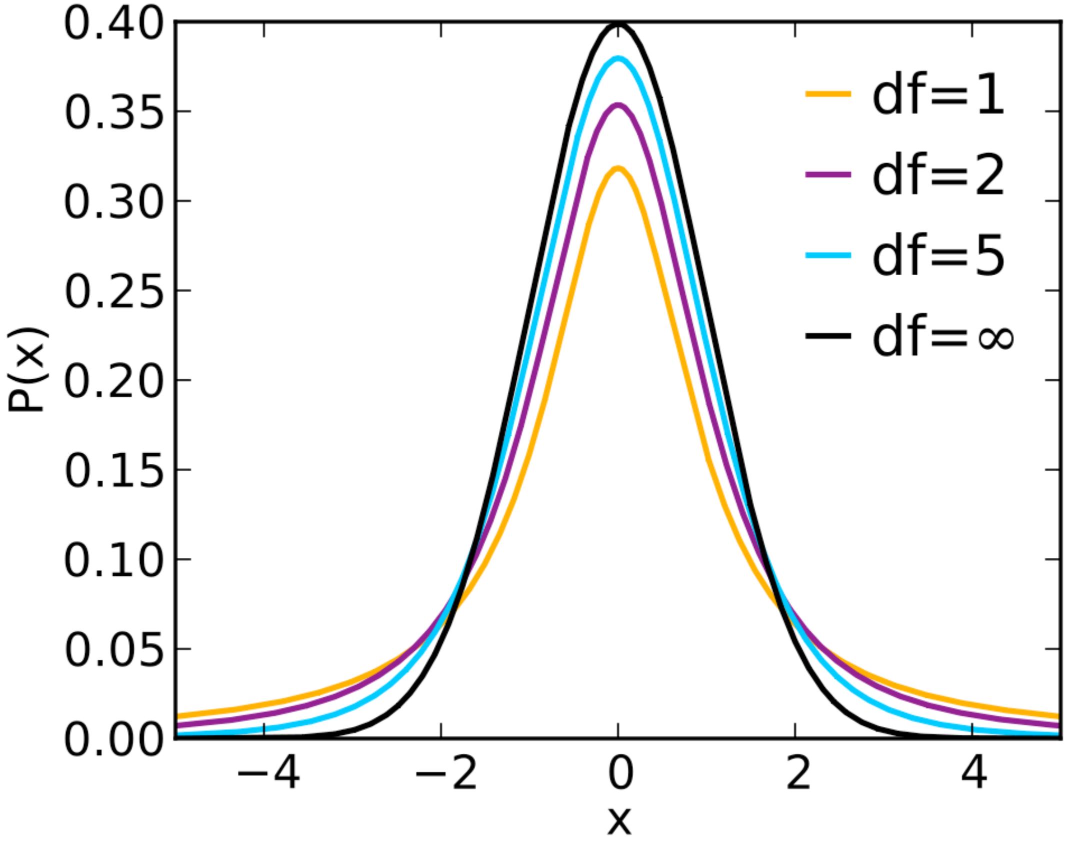


http://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg

t, Student's t

$$X \sim t_n$$

$$f(x) = \text{<I will spare you that>}$$



http://en.wikipedia.org/wiki/File:Student_t_pdf.svg

Déjà vu?
complete specification of the rv's dist'n seems to require:

the "family", e.g. uniform, normal, Student's t

AND

one or more parameters

$$X \sim Unif(a,b)$$

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

$$X \sim t_n$$

notational sidebar

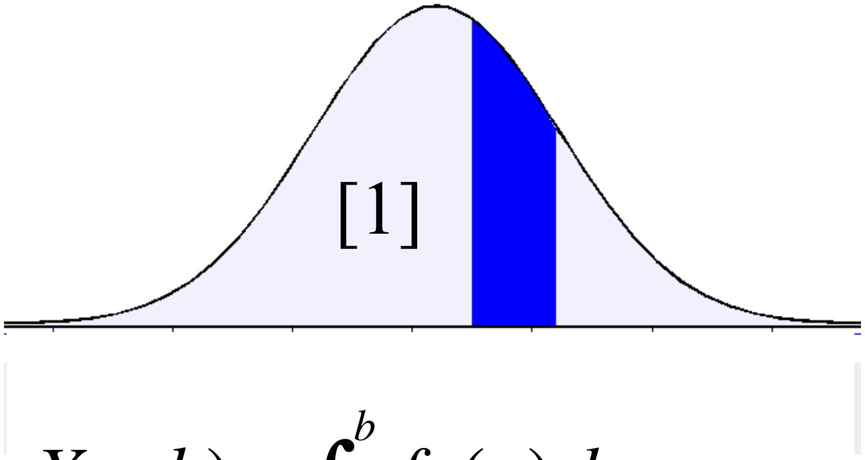
statisticians LOVE to use Greek letters for parameters

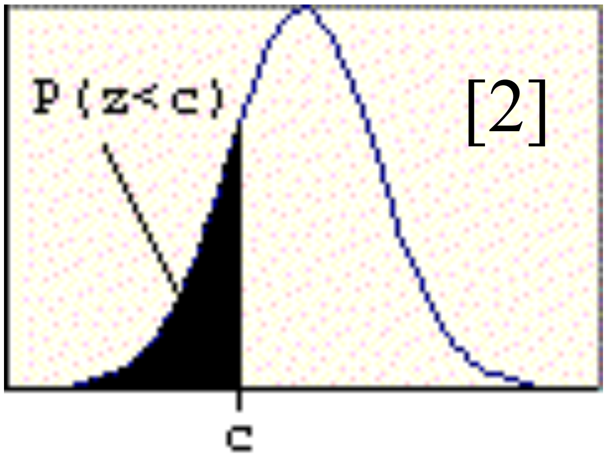
helps to reinforce what's a random variable (e.g. X) and what's a parameter (e.g. μ or σ^2)

this well-intentioned convention probably unsettles some people ("It's all Greek to me!" ≈ "I don't understand.") but I think the pros outweigh the cons

get comfortable with the use of Greek letters to denote parameters

how to get a probability from a density

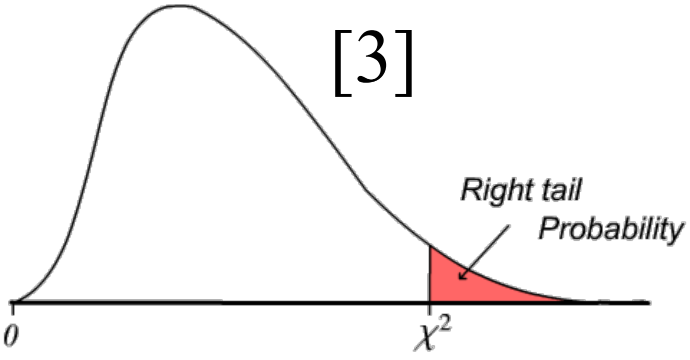




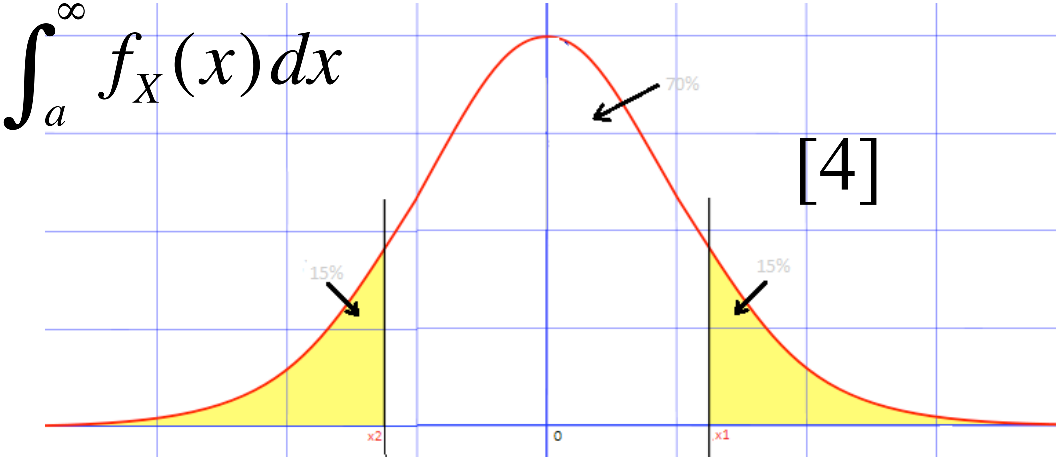
$$[1]\ P(a < X < b) = \int_a^b f_X(x)\,dx$$

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

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



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



sources of images on previous page

http://math.hope.edu/newsletter/2007-08/pdf-2.gif

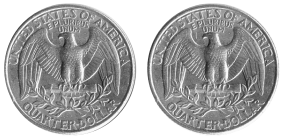
http://onlinecourses.science.psu.edu/stat100/sites/onlinecourses.science.psu.edu.stat100/files/lesson07/chi-square.gif

http://dsearls.org/courses/M120Concepts/ClassNotes/Statistics/520_standard_normal.htm

http://www.tutorvista.com/math/estimating-with-confidence probability of an event = a number between 0 and 1

two main ways to think about probability

[1] relative frequency: If I conduct the two-coin-toss experiment over and over again, the relative frequency of "tails, tails" will tend to 0.25.

P(

) = 0.5 * 0.5 = 0.25

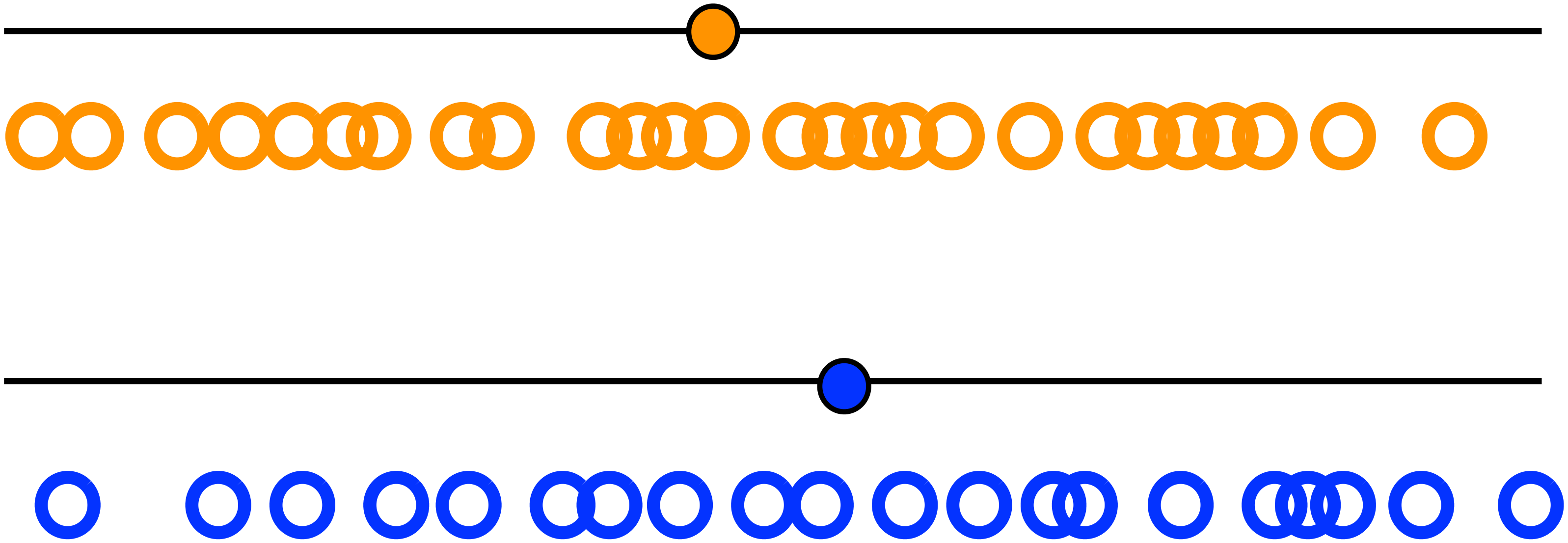
[2] degree of belief: <Not so easy to demo in a toy example but I bet you know what I mean.>

two main interpretations of probability and ties to major schools of statistical inference

[1] relative frequency ~ frequentist

[2] degree of belief ~ Bayesian, subjectivist

to be clear, Bayesians make use of both notions, but frequentists rely on the former



Regard the data as iid observations of random variables that have certain (unknown) distributions.

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

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

What do we mean by iid?

iid

independent
identically
distributed

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

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

The identically distributed part is straightforward: e.g. assume the Y's all have distribution F, whatever it is.

The *independence* is crucial and worth dwelling on.

Independence is a notion defined for events and for random variables.

In a more long-winded introduction, I would carefully distinguish.

But let's cut to the chase: independence of events or rvs makes it much easier to write down the probability of joint events or the joint distribution. It allows you to write these as a *simple product*.