Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 2 – Review of probability and statistical inference

Sara Mostafavi

January 6 2016

Lectures prepared by Dr. Jenny Bryan
Also thanks to Dr. Su-In Lee for some of the slides

Announcements:

- Course webpage:

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

- Check website for information about:
  - Instructors and TAs
  - Lectures; seminars
  - Announcements

Computing seminar kicks off today!

- ESB 1042—two sessions:
  - 11am-12pm: come work through R/R studio installations and basics
  - 12pm-1pm: crash course in molecular biology/genetics

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

- Make sure TAs have your email address and GitHub ID: we need this to create your repositories.

Who am I?

- Assistant professor, jointly appointed (50/50) in Statistics and Medical Genetics, and associated member of Computer Science.
- Computational Biologist: interested in developing and using statistical machine learning methods for high-dimensional genomics data.
- BSc, Computer Sci and Life Sci; PhD, Computer Science, (Machine learning, and computational biology).

outline

- Central concepts (philosophy & goals) in statistics.
- Basic (need-to-know) Stats/Probability terminology.
- Review/intro hypothesis testing.

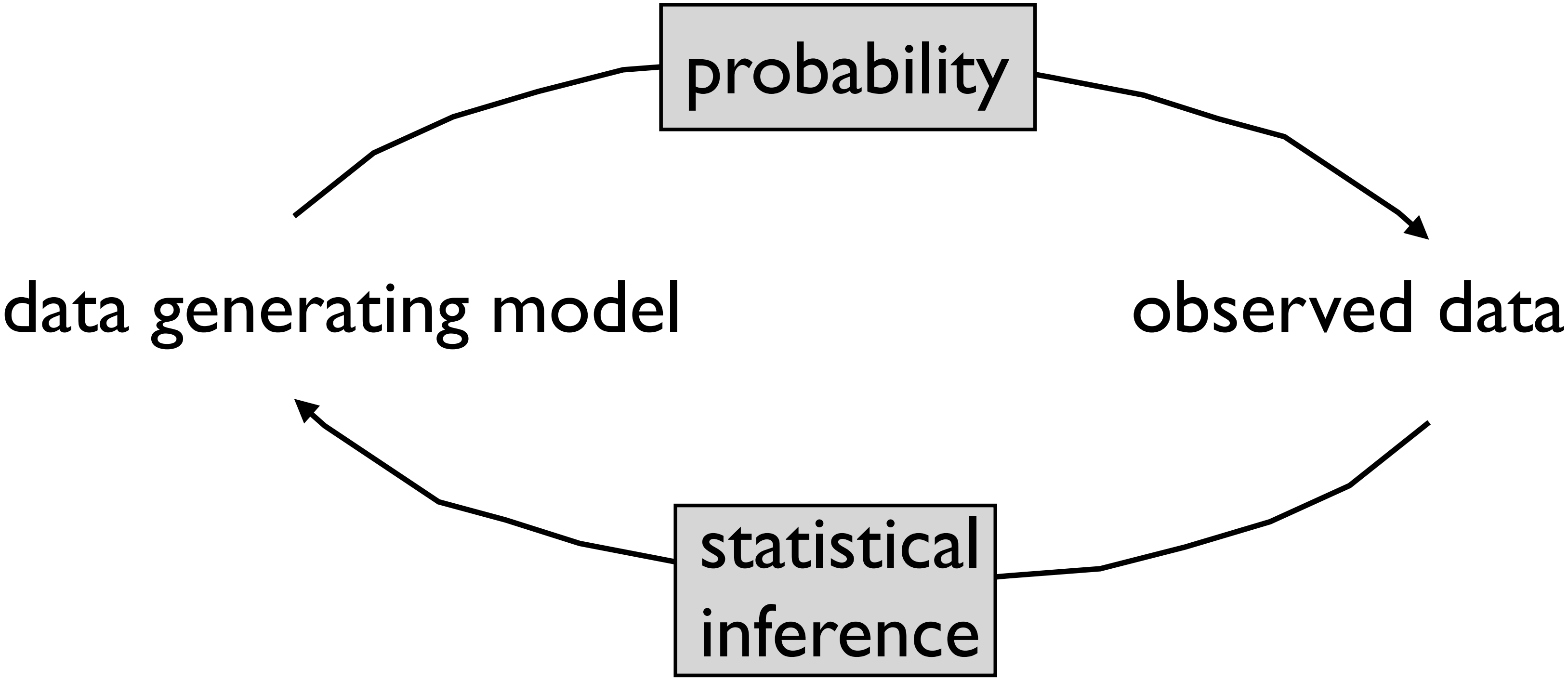
Why we build up from basic statistics?

- The fields of statistics, machine learning, and data mining are concerned with collecting and analyzing data.

- "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."

  Larry Wasserman in preface of "All of Statistics"

- Language of uncertainty, error, and probability
- But also see "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author."



probability

data generating model

observed data

statistical inference

Adapted from Figure 1 of “All of Statistics”

Motivating example….

You are a prisoner and the only way to save your life is to work out one 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☹ be the difference between your guess and the observed proportion. You'll be executed with probability p☹ .

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☹ 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 some basic computation and graphics capability, but no internet, life lines etc.

Which question do you choose to answer? And why?

Let’s come back to that after we review the basics.

Review of basic terminology/concepts you need to know

- Random variable (RV) 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

Review: Sample Space, random variable and its distribution

Random Variable

- rv is a variable whose value results from the measurement of a quantity that is subject to variation due to chance (i.e., outcome of a random process)
  - *Models* the outcome of an experiment with some randomness
  - e.g. dice throwing outcome, expression level of gene A

- More formally …

- Random variable = a function that maps outcomes of an experiment to a real number.

- e.g.,
  - random experiment: toss a coin twice
  - outcome space, sample space = all possible outcomes of the experiment.

    (HH, HT, TT, TH)
  - rv is a function defined on the sample space.
  - rv=number of heads in a given experiment.

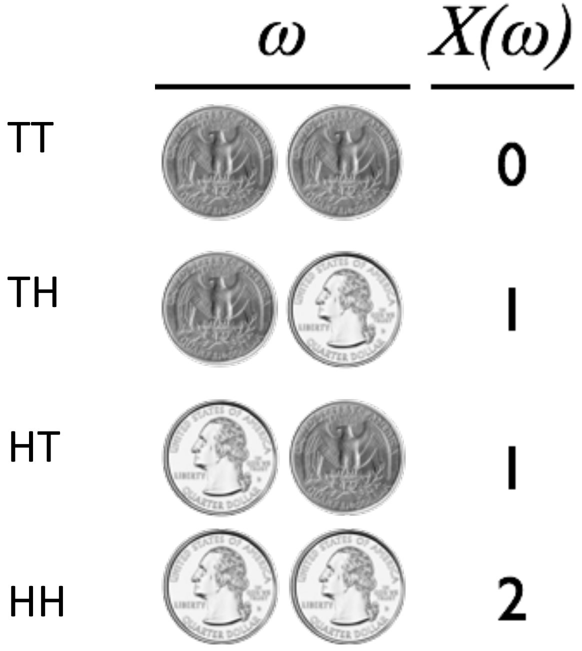
Example

Experiment: 2 coin tosses

ω → greek letters for outcome of the experiment
X(ω) → capital letters for Random variables
Ω → sample space

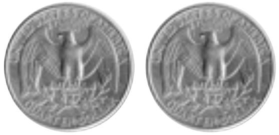
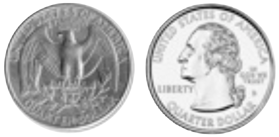
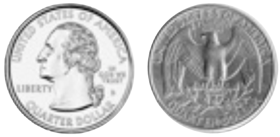
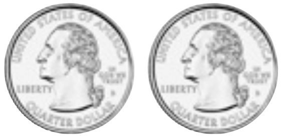
<u>For this example:</u>

X(ω) → Number of heads
Ω → (TT,TH, HT, HH)



TT
TH
HT
HH

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:

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

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

- Rare for outcomes and associated probabilities of a rv to be represented as a table of numbers.

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

- 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)

Examples of PMFs

- Bernoulli distribution:
  - Distribution for a random variable with two outcomes (e.g., “coin toss”)

$$X \sim Bernoulli(p)$$

$$P(X = x) = \begin{cases} X = 1 & p \\ X = 0 & 1 - p \end{cases}$$

- Binomial distribution:
  - Number of successes in a sequence of “binary” experiments.
  - Distribution for “Number of heads in a *n* coin tosses”

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

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1 - p)^{n-x}$$

Example PDF: uniform

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

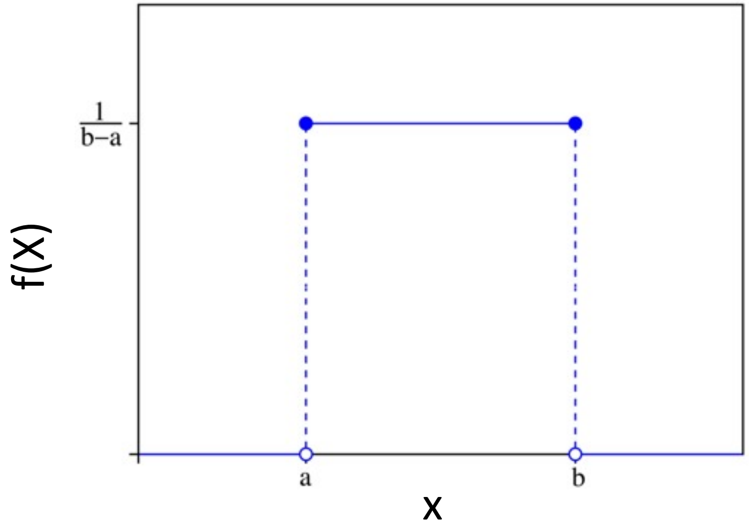
$$f(x) = 1, \text{ for } x \in [0,1]$$

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

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

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

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



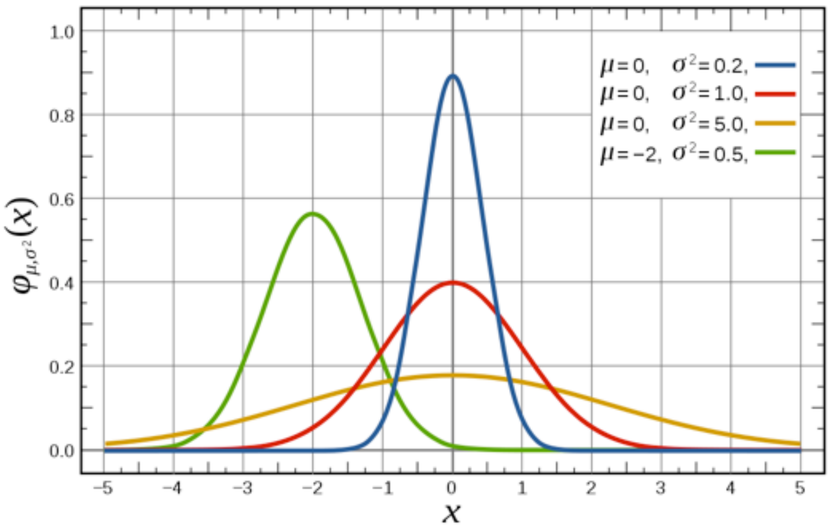
Constant probability over the possible values of X that fall in some range

Example PDF: normal

normal, Gaussian

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

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

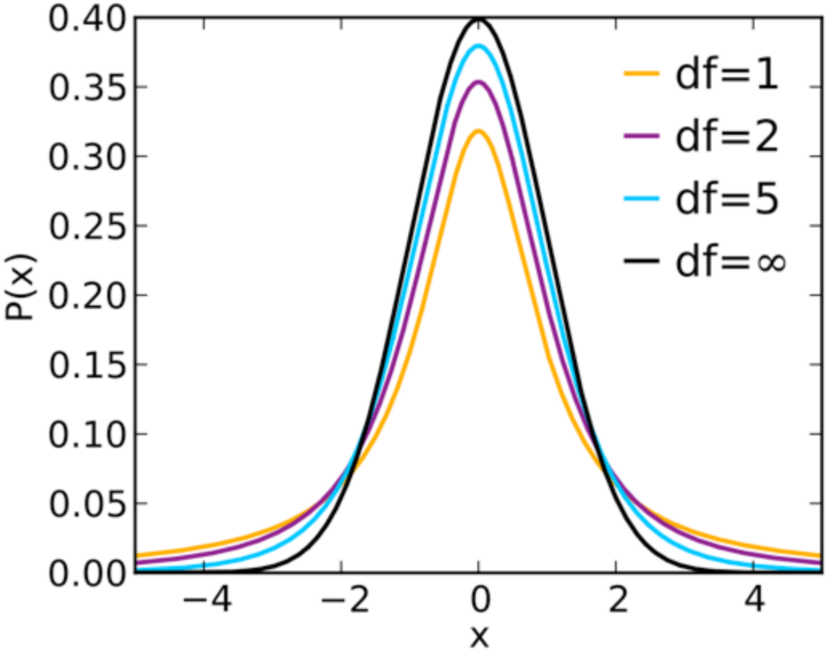


Example PDF: t-distribution

t, Student's t

$X \sim t_n$

$f(x) =$ <I will spare you that>



http://en.wikipedia.org/wiki/File:Student_t_pdf.svg to the dismay of many, a density does not give you probabilities directly

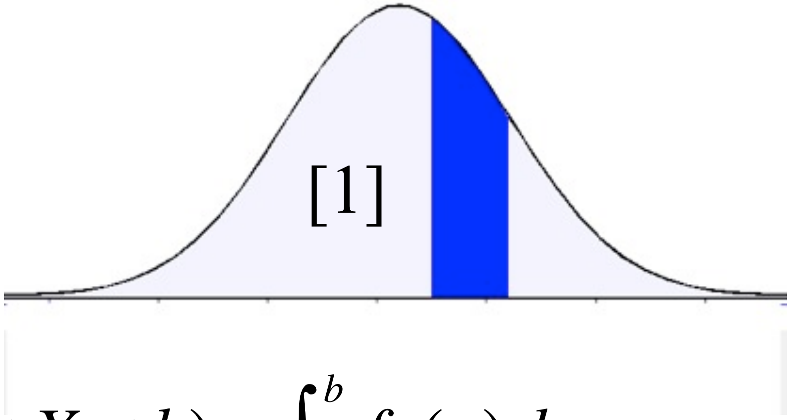
f(x) is NOT the probability that cont rv *X* takes on the value *x* ... which, by the way, is zero

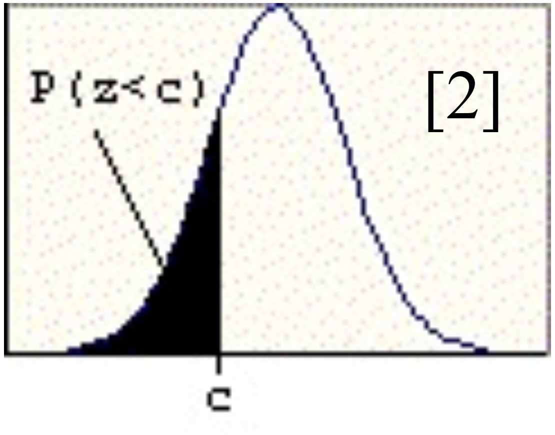
more "proof" f(x) is NOT a probability: f(x) can be greater than 1 (but still never less than 0)

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

how to get a probability from a density



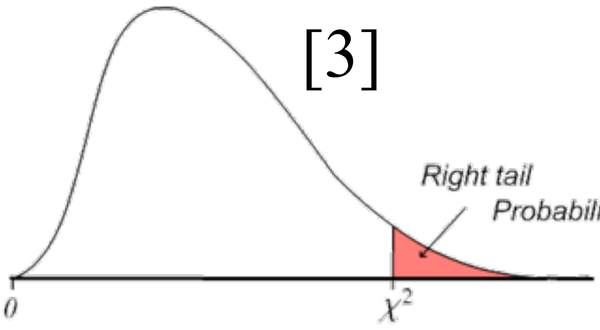


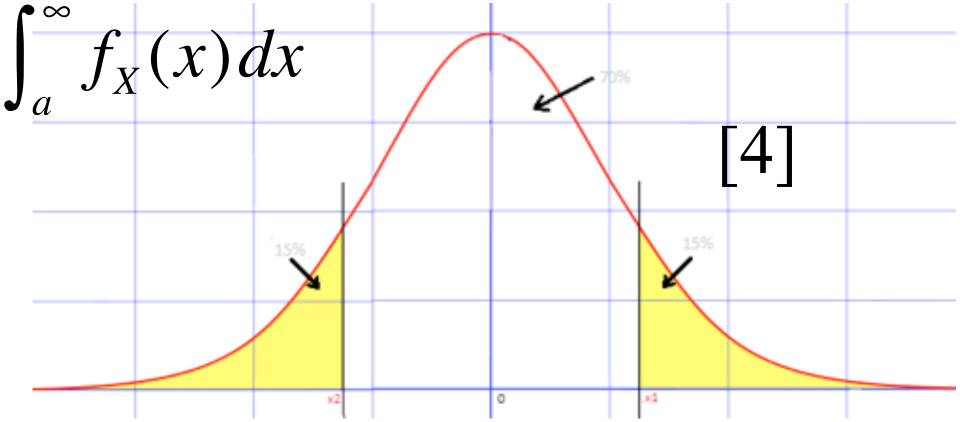
[1] $P(a < X < b) = \int_a^b f_X(x)\,dx$

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





- Complete specification of rv’s distribution seems to require:

1) The “family” (i.e., uniform, normal, binomial, etc)

2) One or more parameters.

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

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

Back to the prisoner Qs:

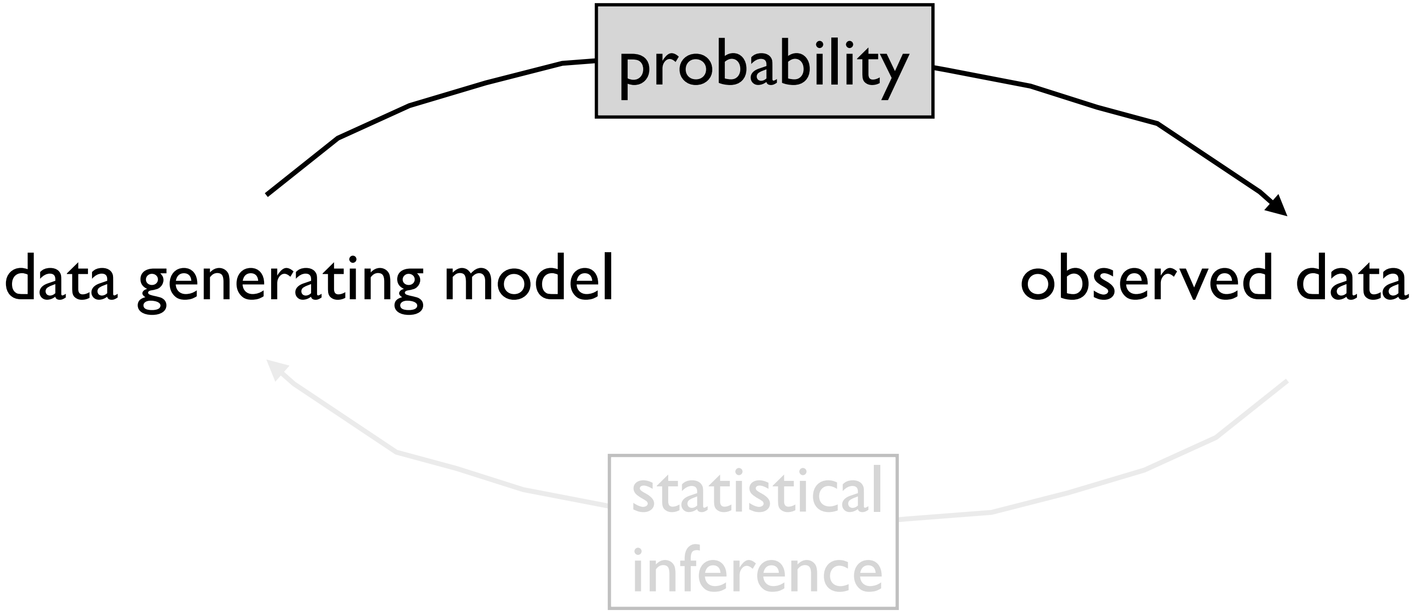
Q1:

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?

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



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

rv has a binomial distribution

(number of successes in n independent “binary” experiments)

→

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

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

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

→

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

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

RV has a binomial distribution →

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

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

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

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

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

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

RV has a binomial distribution

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

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

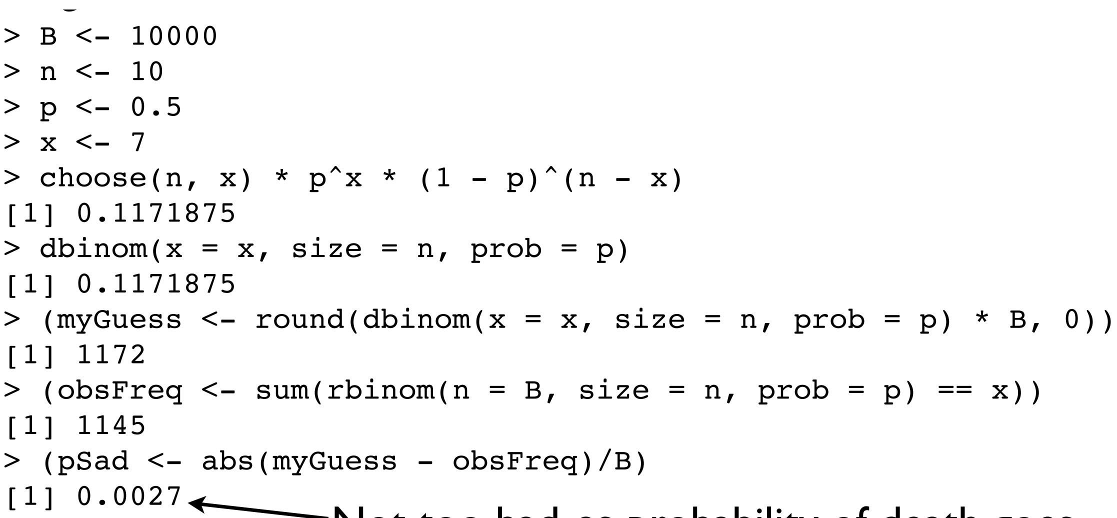
For Q1 we are given the parameters of the model (i.e., binomial distribution)

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

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

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

R code for computing the solution and visualizing the "data":

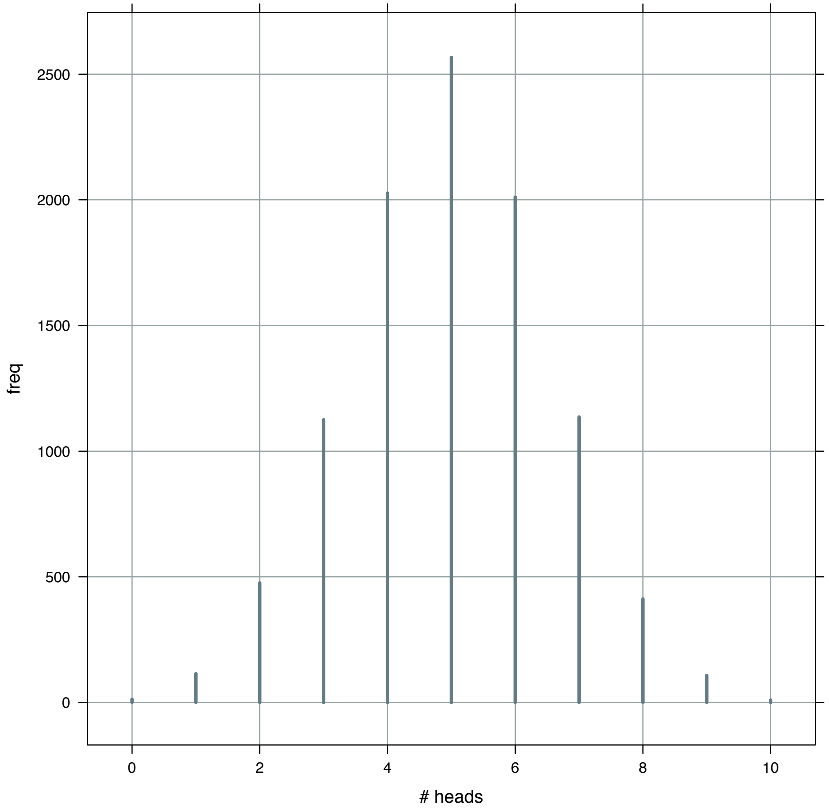


```
> B <- 10000
> n <- 10
> p <- 0.5
> x <- 7
> choose(n, x) * p^x * (1 - p)^(n - x)
[1] 0.1171875
> dbinom(x = x, size = n, prob = p)
[1] 0.1171875
> (myGuess <- round(dbinom(x = x, size = n, prob = p) * B, 0))
[1] 1172
> (obsFreq <- sum(rbinom(n = B, size = n, prob = p) == x))
[1] 1145
> (pSad <-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

Not too bad, as probability of death goes.

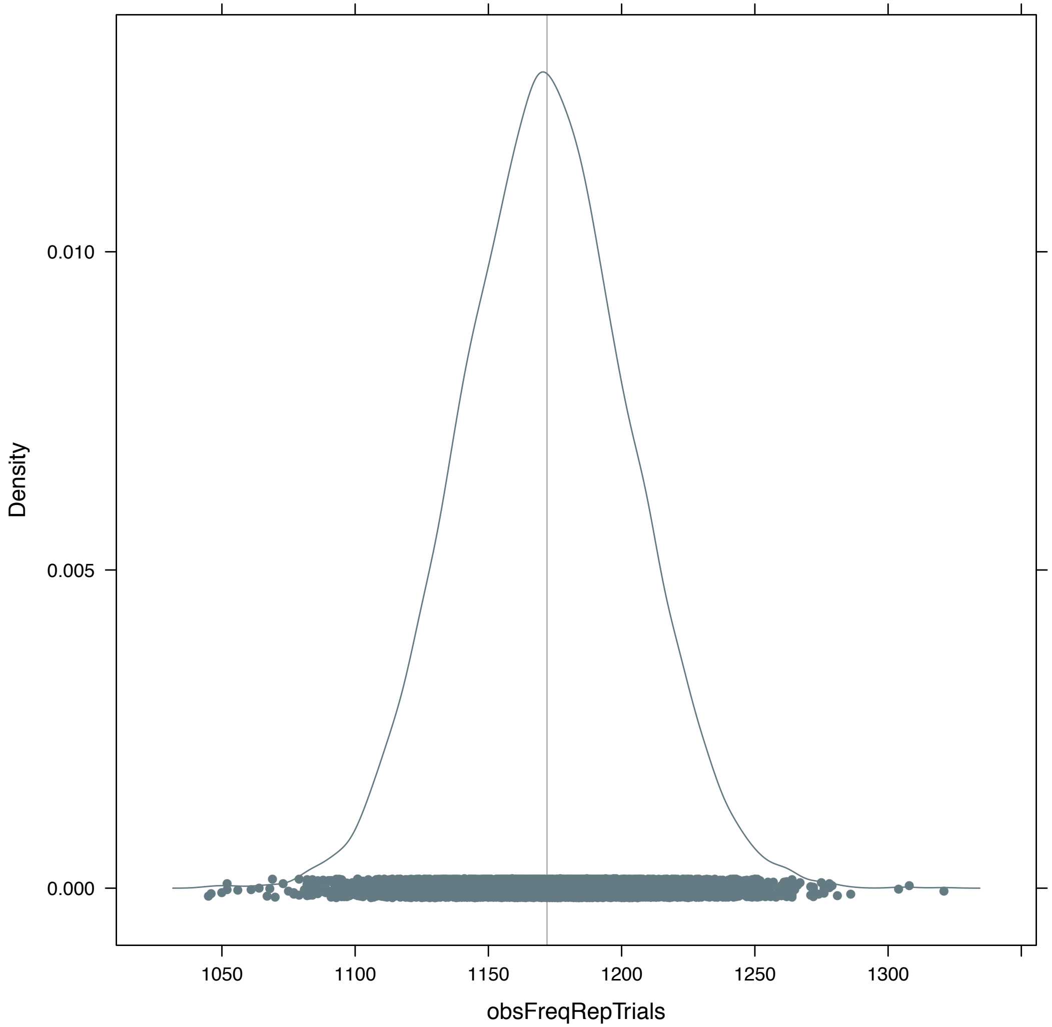
“Brute force” solution to Q1

```
> B <- 10000

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

> 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

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

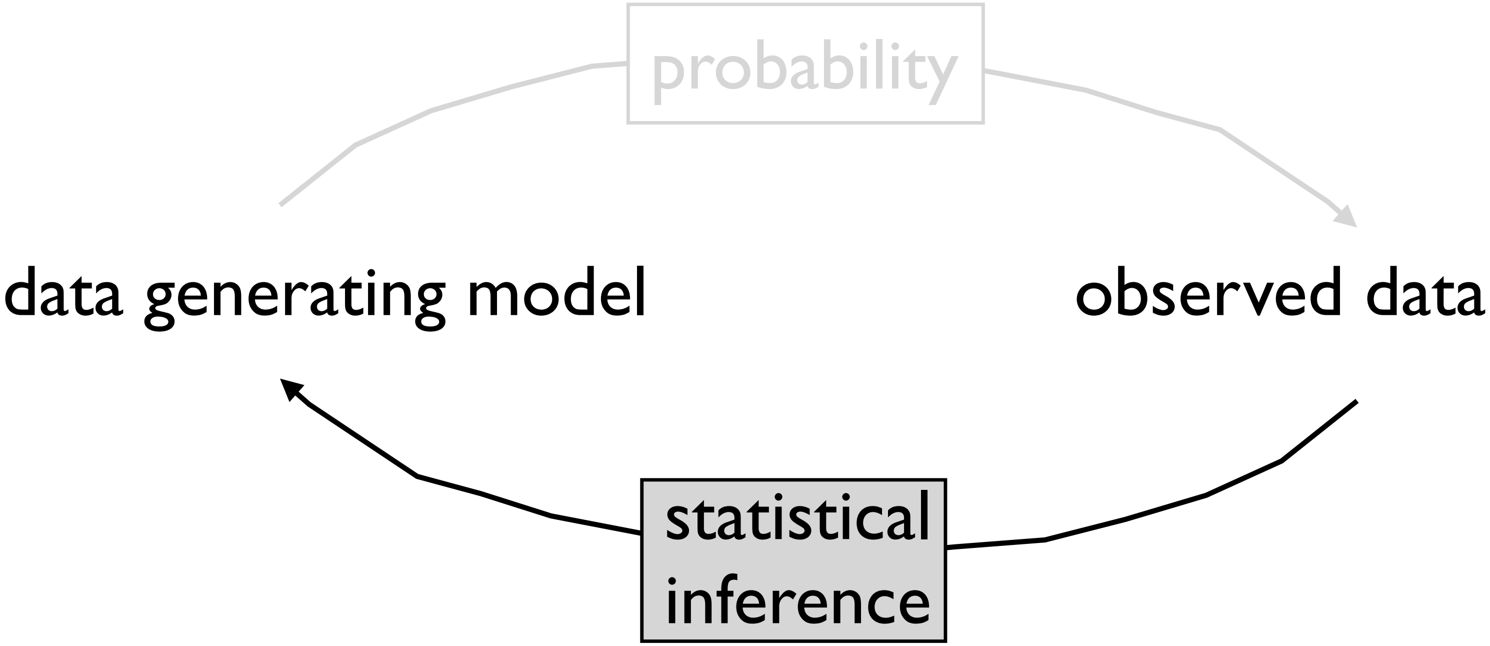


Not too bad, as probability of death goes. Happens to outperform the math solution but that’s not a general fact.

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).



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



data generating model

observed data

probability

statistical inference

"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.

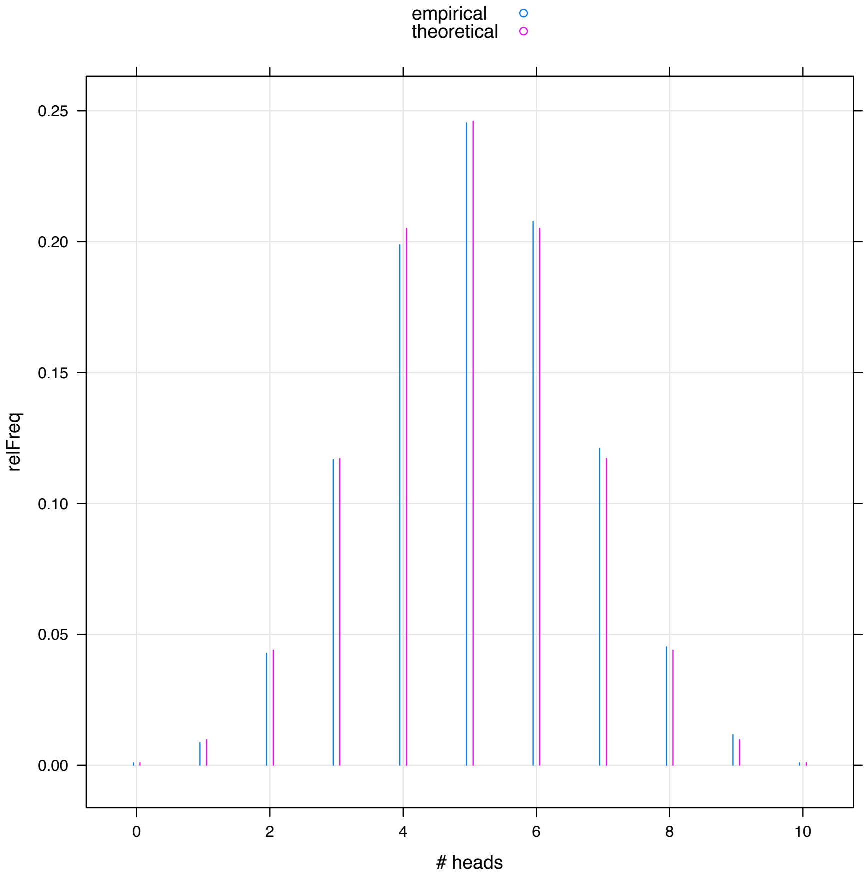
“Solution” to Q2

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”.

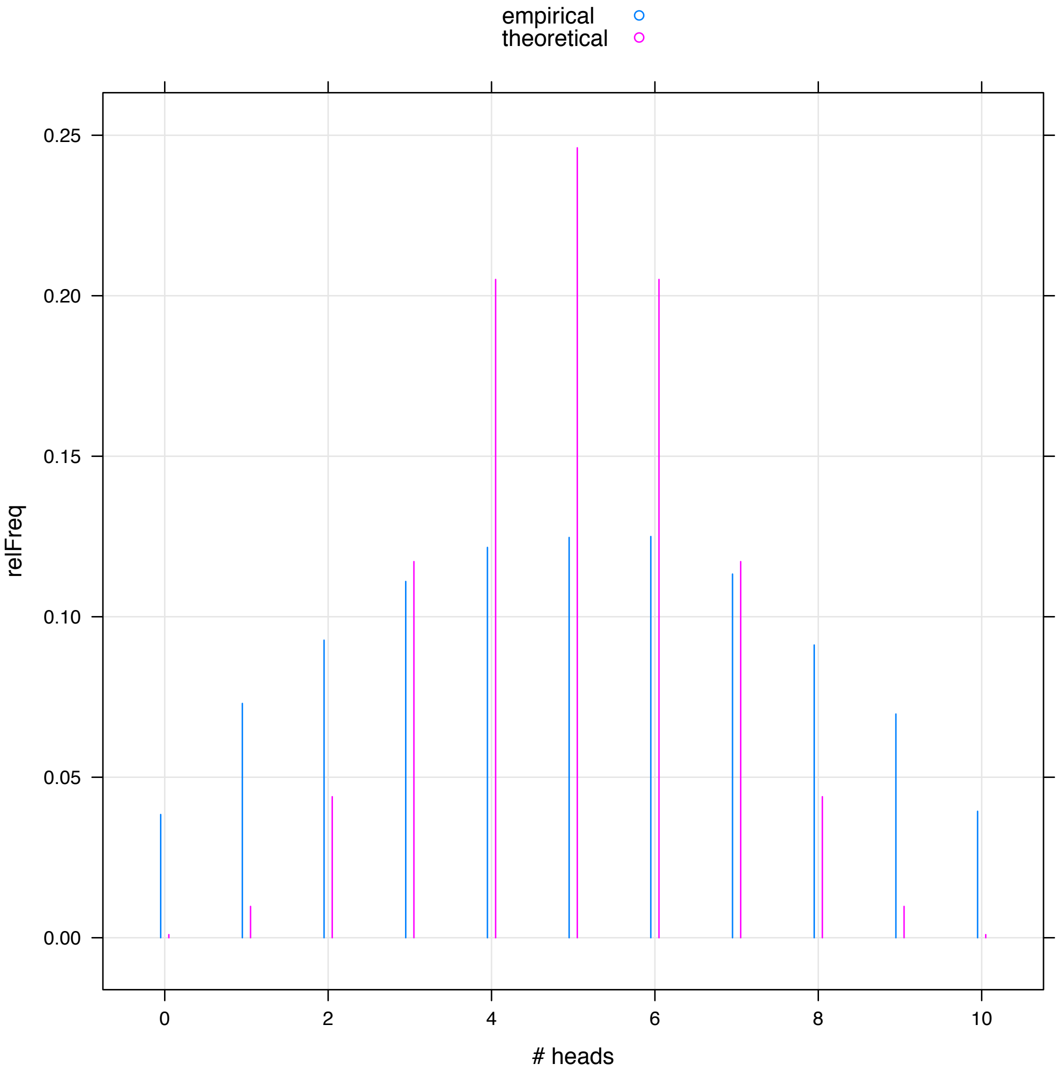
What would you do? Maybe inspect the data to see if it looks plausible under the binomial/regular flip model?

Comparison of empirical distribution to the theoretical distribution with 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 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).

Brief intro to hypothesis testing

Hypothesis testing

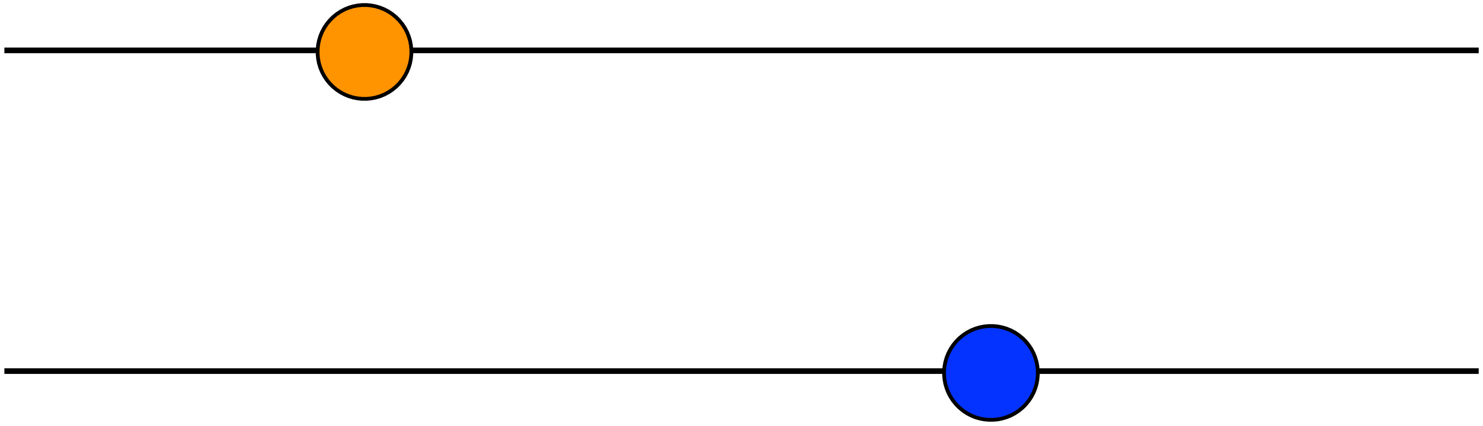
- Example: consider expression level of gene A in some disease (case) and some healthy (control) samples. Is the expression level different in disease compared to healthy samples?

- We can formulate evidence – collect and analyze sample information – for the purpose of determining which of the two hypotheses is false.

- Formally examine two opposing conjectures (hypotheses), H_0 and H_A

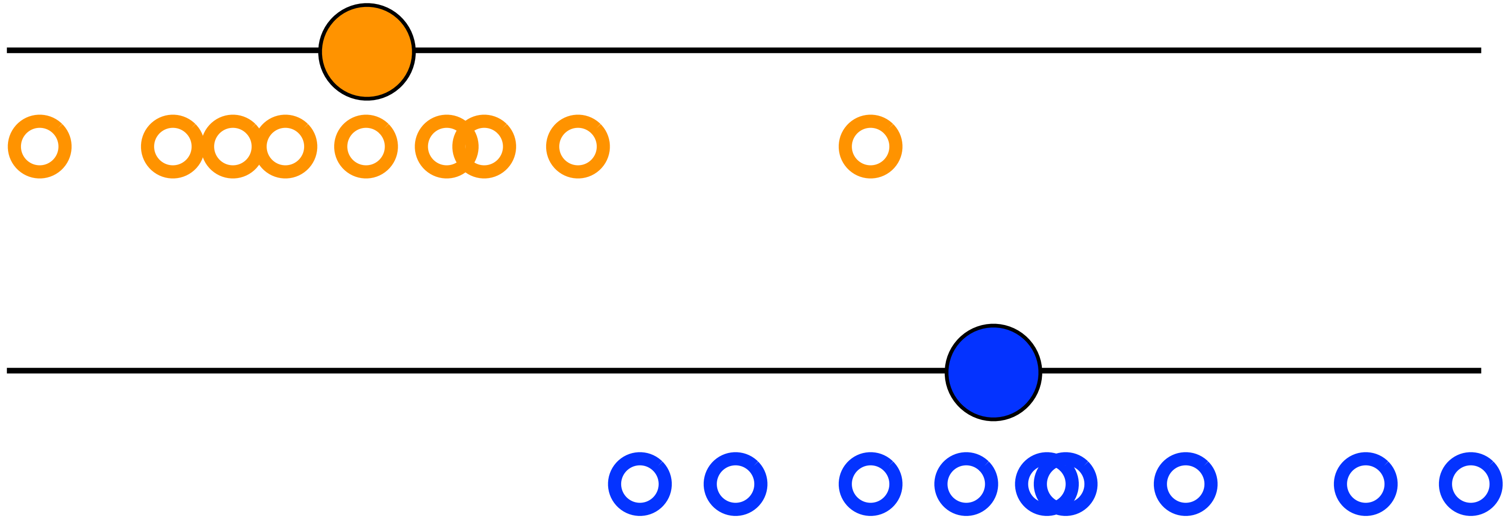
- These two hypotheses are mutually exclusive, so that one is true to the exclusion of the other.

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



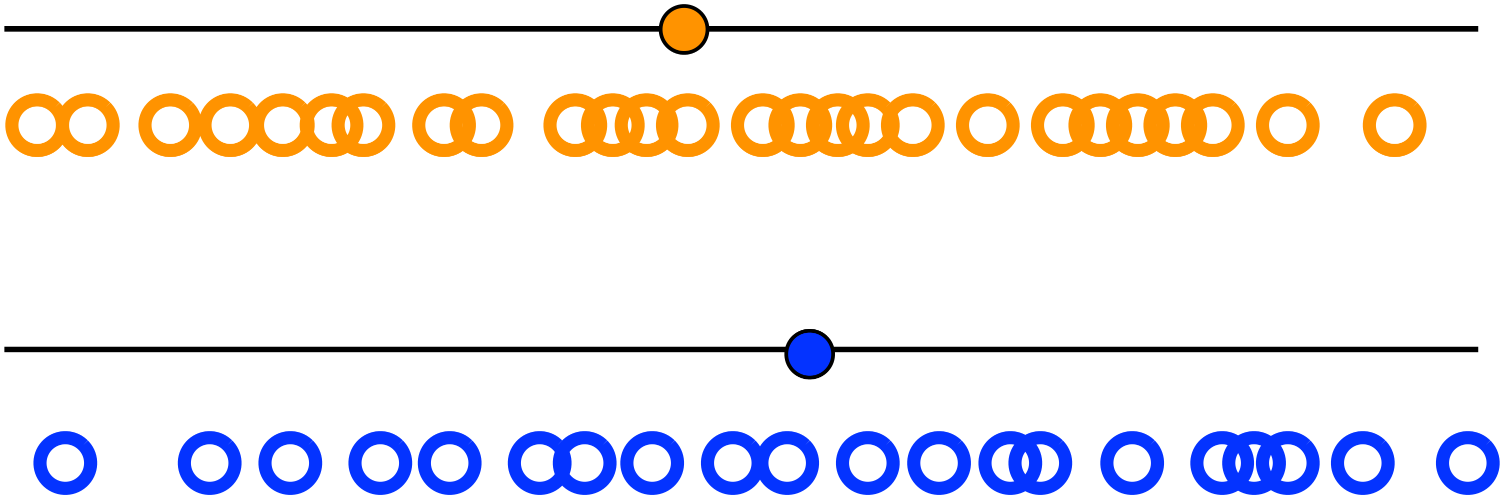
Orange circle = average of orange observations
Blue circle = average of blue observations

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"?



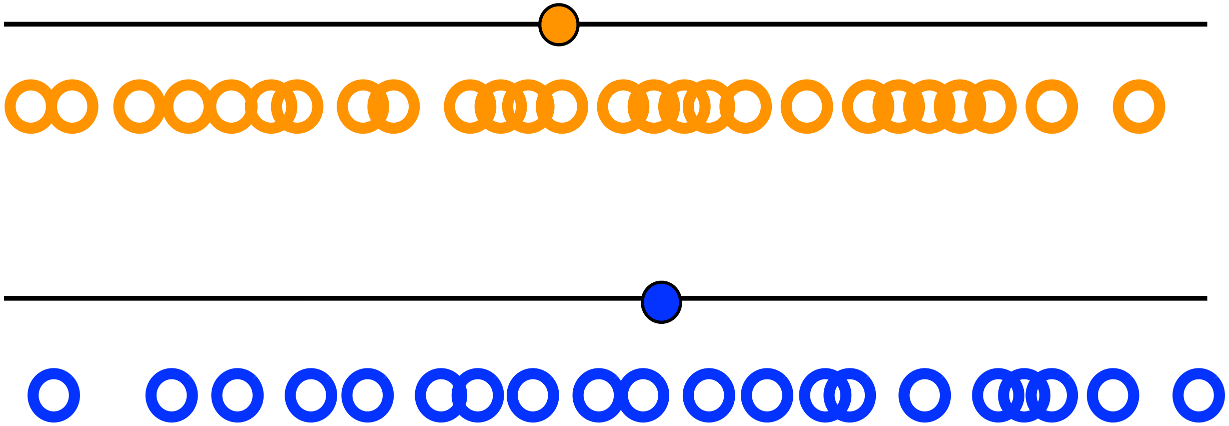
No, not so much.

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

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

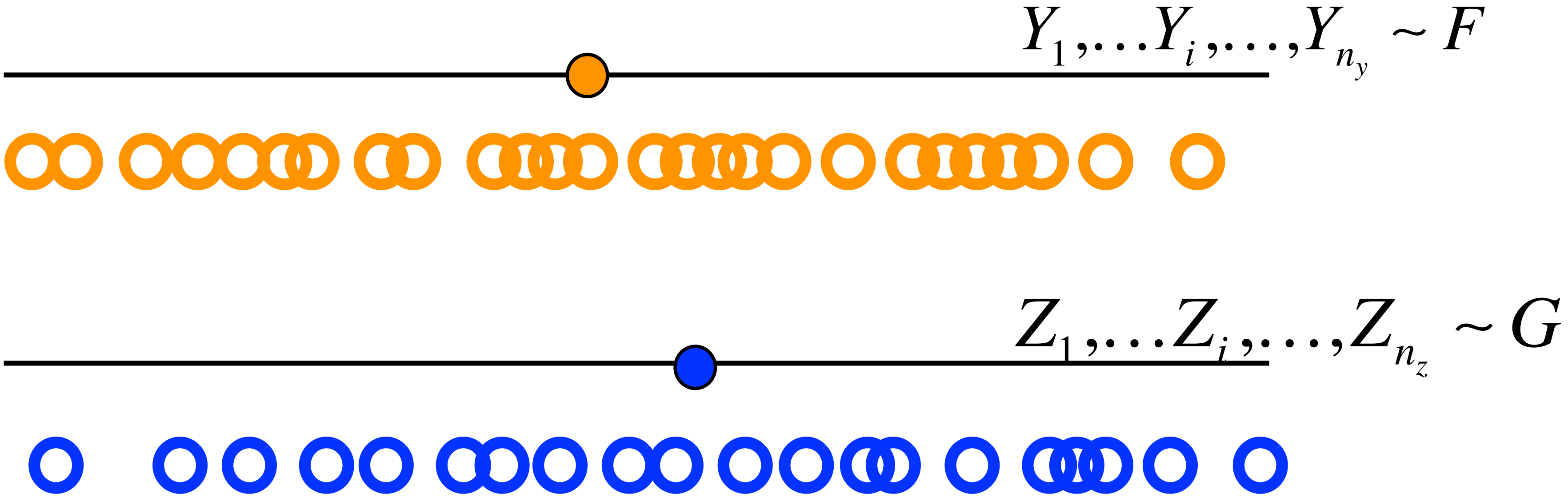
$(Z_1 = z_1, \ldots Z_i = z_i, \ldots Z_n = z_{n_z})$.



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

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

$H_0 : \mu_Y = \mu_Z$

(Occam's razor =))

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: $\mu_Y = \mu_Z$.

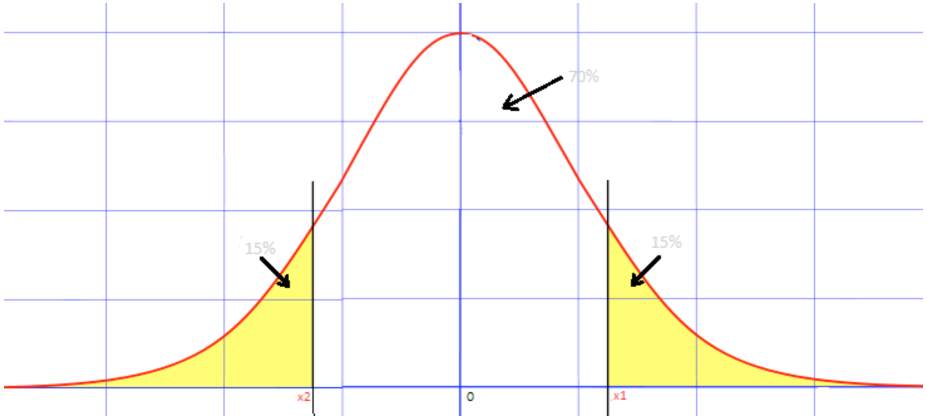
In this case, theory tells us that $t \sim t_{ny + nz - 2}$

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 $n_y + n_z - 2$ degrees of freedom

sum of the yellow areas = p-value

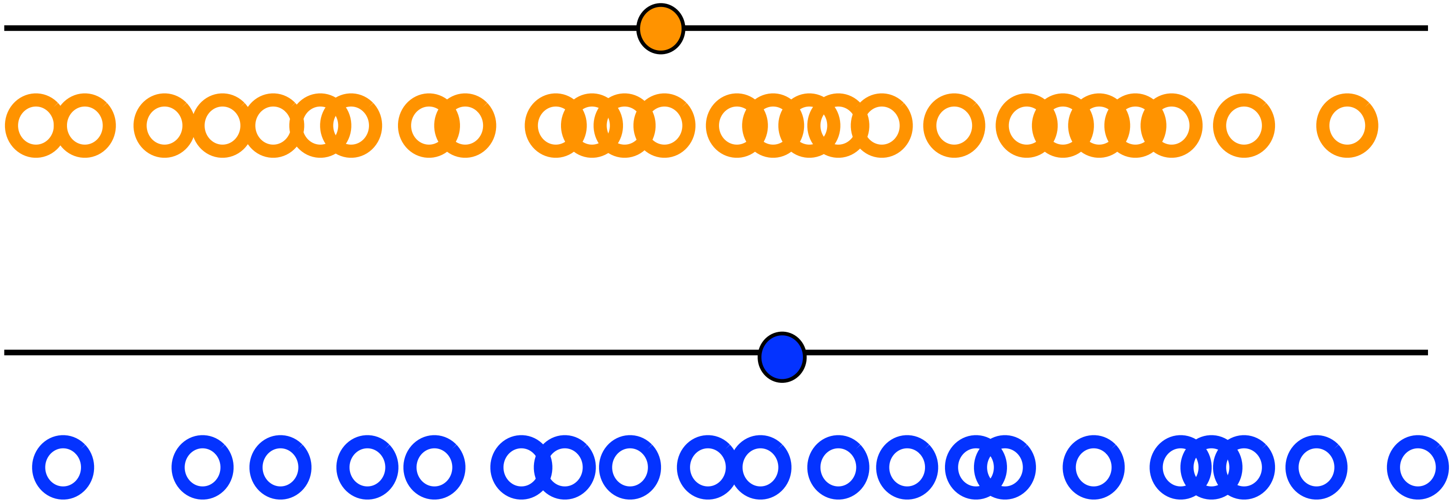


obs. value of test stat.

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

P-value

- After calculating the test statistic we convert it to a p-value by comparing the observed value to distribution of test statistic’s under the null hypothesis.
- P-value quantifies how likely the test statistics value is under the null hypothesis.
  - P-value <= alpha → reject H_0 at alpha level
  - P-value > alpha → Do not reject H_0 at alpha level



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*.

Steps of hypothesis testing (“the algorithm”)

1. Ask a precise and answerable question which as a mutually exclusive yes/no answer.
2. Define a test-statistic that corresponds to the question. You should know the distribution of the test-statistic under the null hypothesis.
3. Compute the p-value associated with the observed test-statistic.

Errors in hypothesis testing

Actual Situation “Truth”

Decision	**H_0 True**	**H_0 False**
Don Not Reject H_0		
Reject H_0		

Errors in hypothesis testing

	Actual Situation "Truth"	
Decision	H_0 True	H_0 False
Don Not Reject H_0	Correct Decision $1-\alpha$	Incorrect Decision Type II Error B
Reject H_0	Incorrect Decision Type I Error α	Correct Decision $1-\beta$

α = P(Type I Error) β = P(Type II Error)

Power = $1 - \beta$

Computing seminar kicks off today!

- ESB 1042—two sessions:
  - 11am-12pm: come work through R/R studio installations and basics
  - 12pm-1pm: crash course in molecular biology/genetics

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

- Make sure TAs have your email address and GitHub ID: we need this to create your repositories.