

Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 2 – Review of probability and
statistical inference part 1

Paul Pavlidis
January 9 2017

****Based on lecture originally prepared by Dr. Jenny Bryan****
Also thanks to Dr. Su-In Lee for some of the slides

Announcements

Announcements will be made in class and posted on the
Announcements page of the course website

([https://stat540-
ubc.github.io/subpages/announcements.html](https://stat540-ubc.github.io/subpages/announcements.html)).

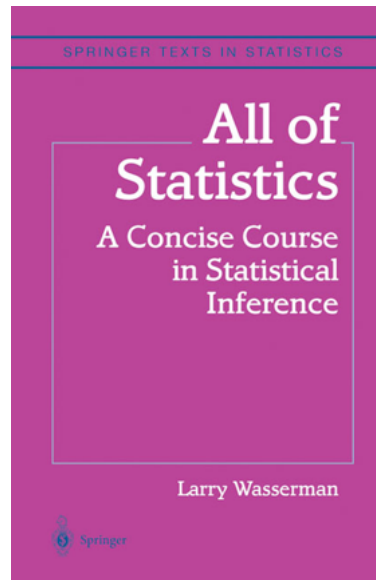
1. Students should go over Jennifer Bryan's tutorial on
Git, GitHub + R and RStudio
(<http://happygitwithr.com/>), prior to seminar 2, in
order to ensure a smooth running seminar session.
2. A course github repo will be made for you once all
students have completed the survey. Please fill it out
as soon as possible if you have not
already: <https://goo.gl/forms/NzMRW87Ccmfmc6x13>

Outline for lectures 2 & 3

- Central concepts (philosophy & goals) in statistics.
- Basic need-to-know stats/probability terminology.
- Intro to hypothesis testing.

Learning goals for these two lectures

- The role of **probability** [distributions]
- The role of a **model** in working with data
- The role of **inference** and **hypothesis testing**
- **Why** we need all these things (error, uncertainty)



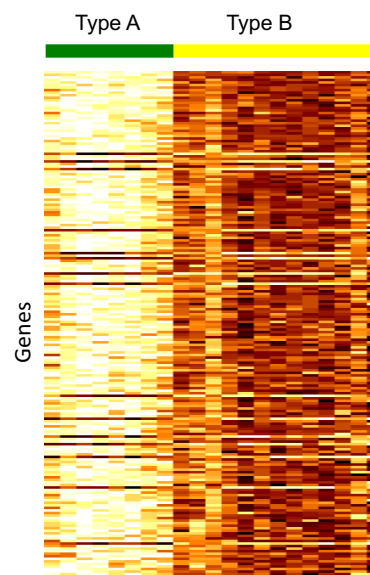
For more in-depth mathematical treatment of these topics

Especially chapters 1-3, 6, 9

Free to UBC on SpringerLink

We could just cut to the chase ...

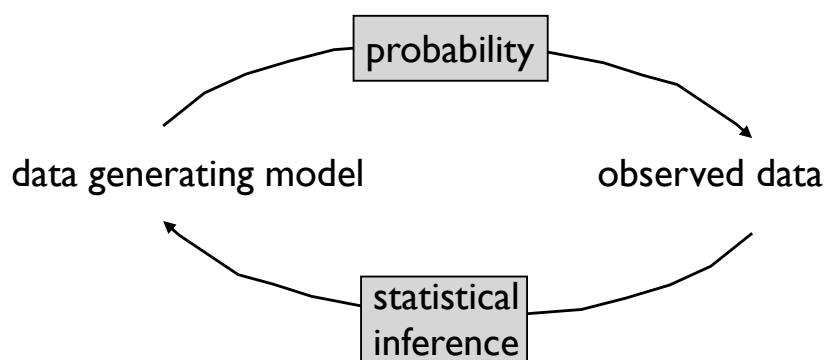
- In seminar 6, we show you how to get to this. You could just follow the recipe.
- Instead, we're going to build up from the fundamentals.



Why we build up from basic statistics?

- The fields of statistics, machine learning, and data mining are concerned with collecting and analyzing data.
- Language of uncertainty, error, and probability
- “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”



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^\oplus be the difference between your guess and the observed proportion. You'll be executed with probability p^\oplus .

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^\oplus be like so: If no “difference” between your description and the truth, then $p^\oplus = 0$. As “difference” grows, p^\oplus tends to 1.*
- You will be executed with probability p^\ominus .

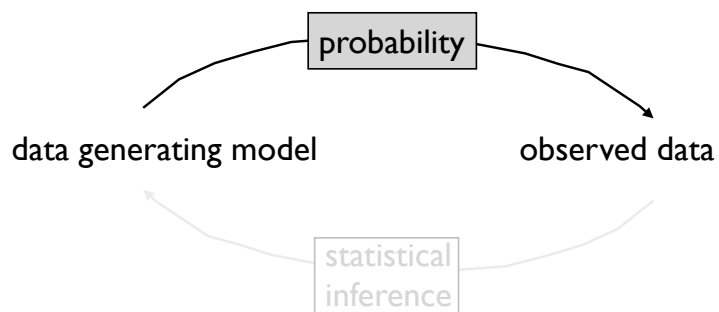
* Sorry this is so vague but I can't do better without getting bogged down in details. Go with me.

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

Which question do you choose to answer? And why?

Problem #1

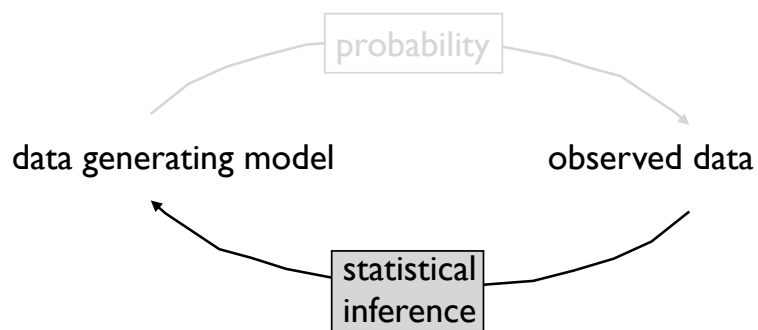
“Given the data generating model, what are some properties of the observed data?”



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.

Review of basic terminology/concepts you need to know

- Random variable (RV) and its distribution
- Independent and identically distributed (IID)
- Large sample results for averages (CLT)
- Parameters of a distribution
- An estimator of a parameter
- A parameter space
- The sampling distribution of an estimator
- Null and alternative hypotheses

Random Variable (rv)

- 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. coin flip outcome, expression level of gene A
- More formally ...

- Random variable = a function that maps any possible **outcome** of an **experiment** to a real number.
 - Generally, this is only worth doing if the experiment can have more than one possible outcome ... that range of outcomes is the **sample space**.
- Probability = A number assigned to an outcome satisfying certain rules (for now okay to think of as *frequency* of an outcome)
- Probability distribution = Function that maps outcomes to probabilities
(more precise definitions in a bit)

Example

Experiment: 2 coin tosses

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





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

General notation (following Wasserman)

$\omega \rightarrow$ 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	1	
	<hr/>			
	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

- **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 .

Examples of PMFs

- Bernoulli distribution:

- Distribution for a random variable with two outcomes (e.g., “coin toss”)
- p is the (sole) parameter for this distribution
- For a “fair” coin, $p = 0.5$

$$X \sim \text{Bernoulli}(p)$$

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

- Binomial distribution:

- Number of successes in a sequence of “binary” experiments.
- Distribution for “Number of heads in a n coin tosses”
- This distribution has two parameters

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

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

Example PDF: uniform

$$X \sim \text{Unif}(0,1)$$

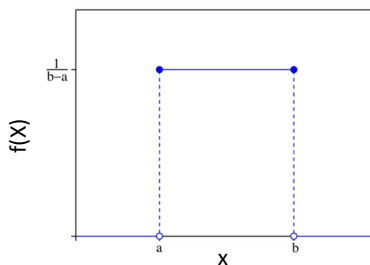
$$f(x) = 1, \text{ for } x \in [0,1]$$

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

$$X \sim \text{Unif}(a,b)$$

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

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



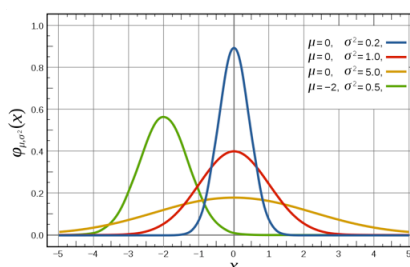
Constant probability over the possible values of X that fall in some range

Example PDF: normal

normal, Gaussian

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

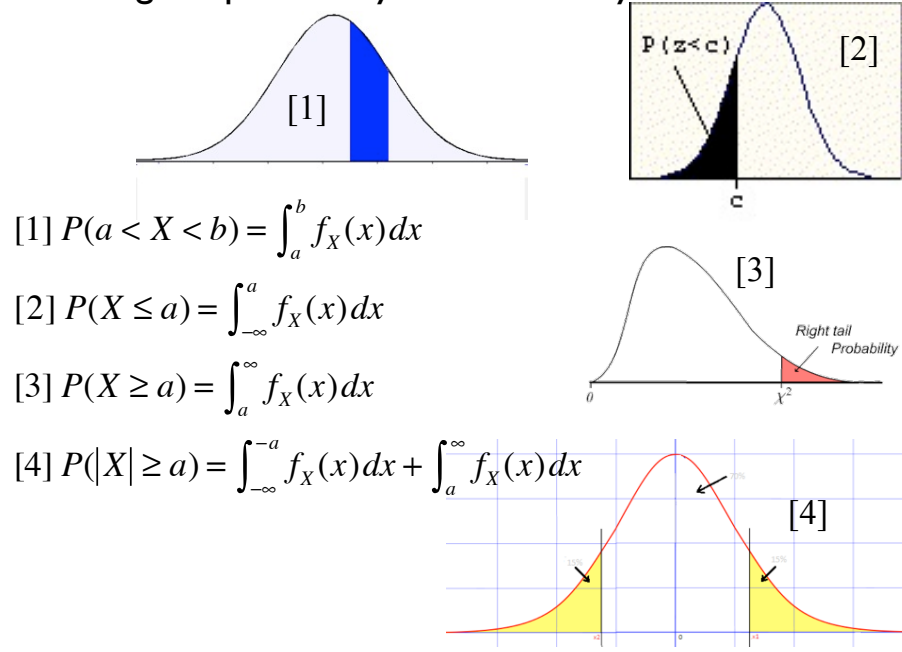
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



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



Back to the prisoner Qs:

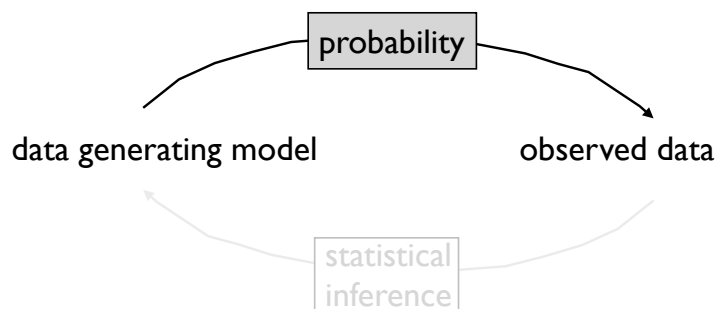
Q1:

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 is “# of heads in 10 tosses”.

RV has a binomial distribution \longrightarrow $X \sim \text{Bin}(n, p)$

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

For Q1 we are given the parameters of the model (i.e., binomial distribution) \longrightarrow $X \sim \text{Bin}(n = 10, p = 0.5)$

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

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 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 (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

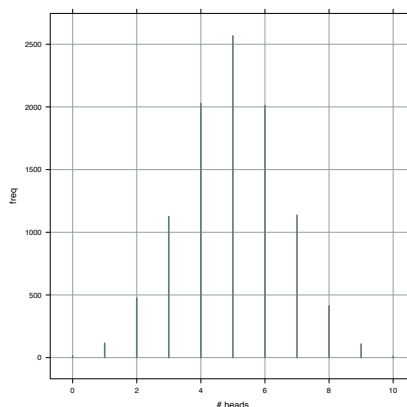
> y <- rowSums(coinFlips)

> head(y)
[1] 2 6 4 7 5 5

> head(y == 7)
[1] FALSE FALSE FALSE TRUE FALSE FALSE

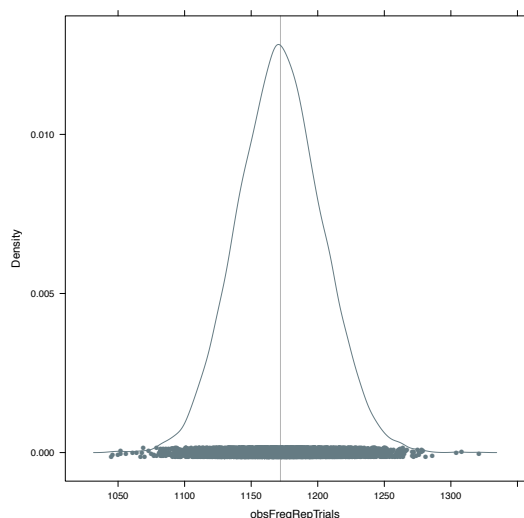
> (myGuess <- sum(y == 7))
[1] 1136

> (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).

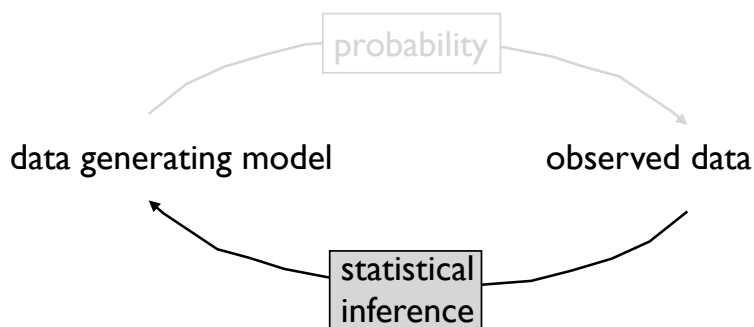


Recall Problem #2

- 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^\oplus be like so: If no “difference” between your description and the truth, then $p^\oplus = 0$. As “difference” grows, p^\oplus tends to 1.*
- You will be executed with probability p^\ominus .

* 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

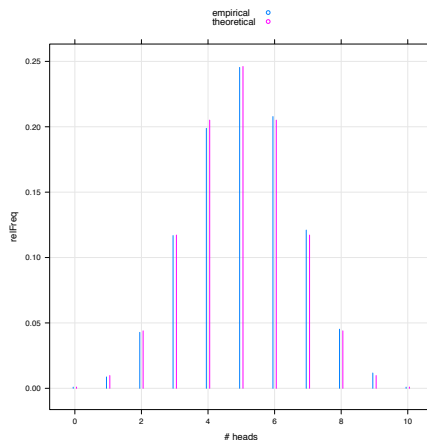
Assuming nothing ... is probably a death sentence!

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?

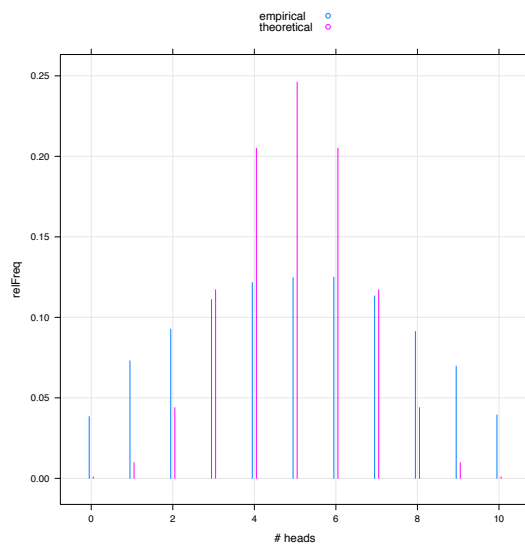
Comparison of empirical distribution to the theoretical distribution with $\text{Bin}(n=10, p)$ for some p (in this case $p=0.5$)

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 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

IID

- A {requirement, assumption} in numerous settings is that the data are IID: Independent and Identically Distributed.
- **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 “identically distributed”

- Toy example: Imagine the executioner is using different coins for each toss.
- In experiments we can try to avoid violations: “keep conditions constant” e.g., use animals that are the same age, use the same temperature.
- However, I.D. is often an assumption we are more or less forced to live with e.g., that all people with a particular diagnosis are “equivalent”.

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**.

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).

Quick dip into hypothesis testing

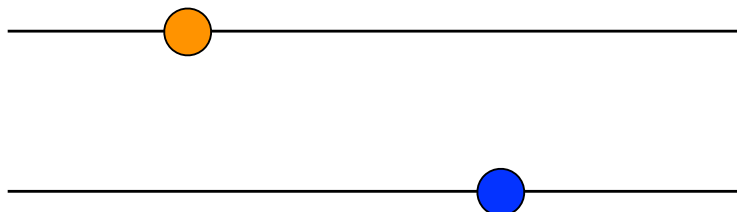
- Example: consider expression level of gene A in some disease (case) and some healthy (control) samples. Is the expression level different in disease compared to healthy samples? Or the same?
- These two hypotheses are **mutually exclusive** – only one can be true
- We can formulate evidence – collect and analyze sample information – for the purpose of determining which of the two hypotheses is false.
- Hypothesis test: Formally examine two opposing conjectures (hypotheses), H_0 and H_A
- A hypothesis test tries to assess which one is true (and how confident we are in that assessment)

Steps of hypothesis testing

1. Get some data (a sample)
2. Ask a precise and answerable question which has a mutually exclusive yes/no answer.
3. Define a test-statistic that corresponds to the question. You should know the distribution of the test-statistic under the null hypothesis (so you can evaluate how likely the test-statistic is)
4. Compute the p-value associated with the observed test-statistic. The p-value expresses how surprised we would be to see a test statistic like this if the null hypothesis was true.

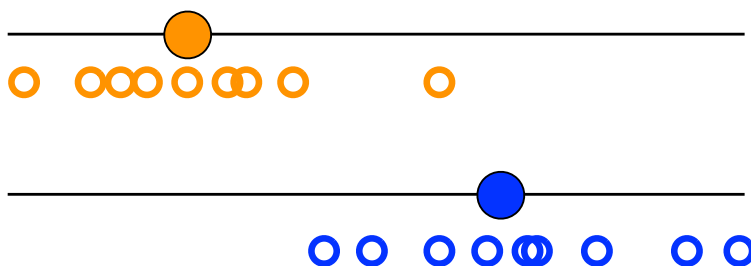
More precise definitions/explanations later!

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



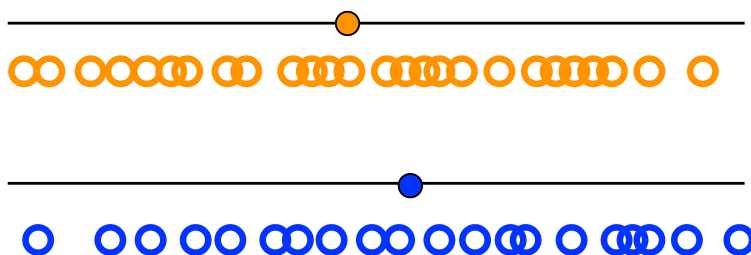
Orange circle = average of orange observations
Blue circle = average of blue observations

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



Yeah, pretty compelling evidence to me.

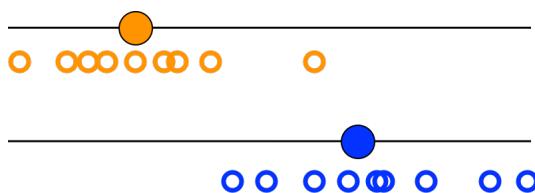
Does this constitute evidence that the “oranges” are meaningfully different from the “blues”?



No, not so much.

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

Some intuition about why we like to see the actual data points



- We realize that the sample mean (big dots) has uncertainty – if we took a new sample, our estimate would change.
- The spread of the data (smaller circles) lets us see how sketchy (or not) our guesses are.
- Having more data lets us be more certain.

Formalization of this intuition is key to **statistical inference**
We'll develop this further later