Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 3 – Review of probability and statistical inference

Sara Mostafavi

January 11 2016

Lectures prepared by Dr. Jenny Bryan

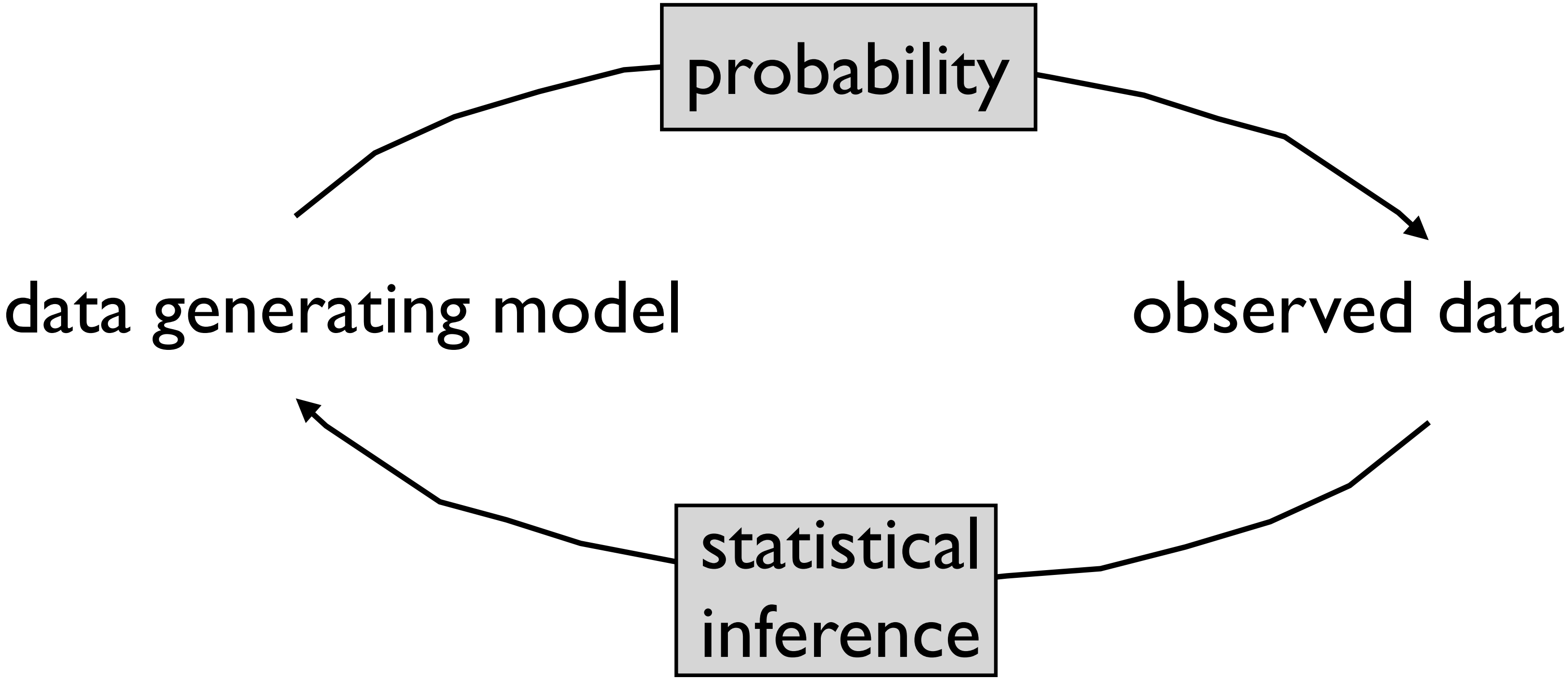
So far we have reviewed:

- estimate/use data generating model to understand/ describe an observed *sample*
- rv's and their distributions
- Importance of variance and hypothesis testing

Today

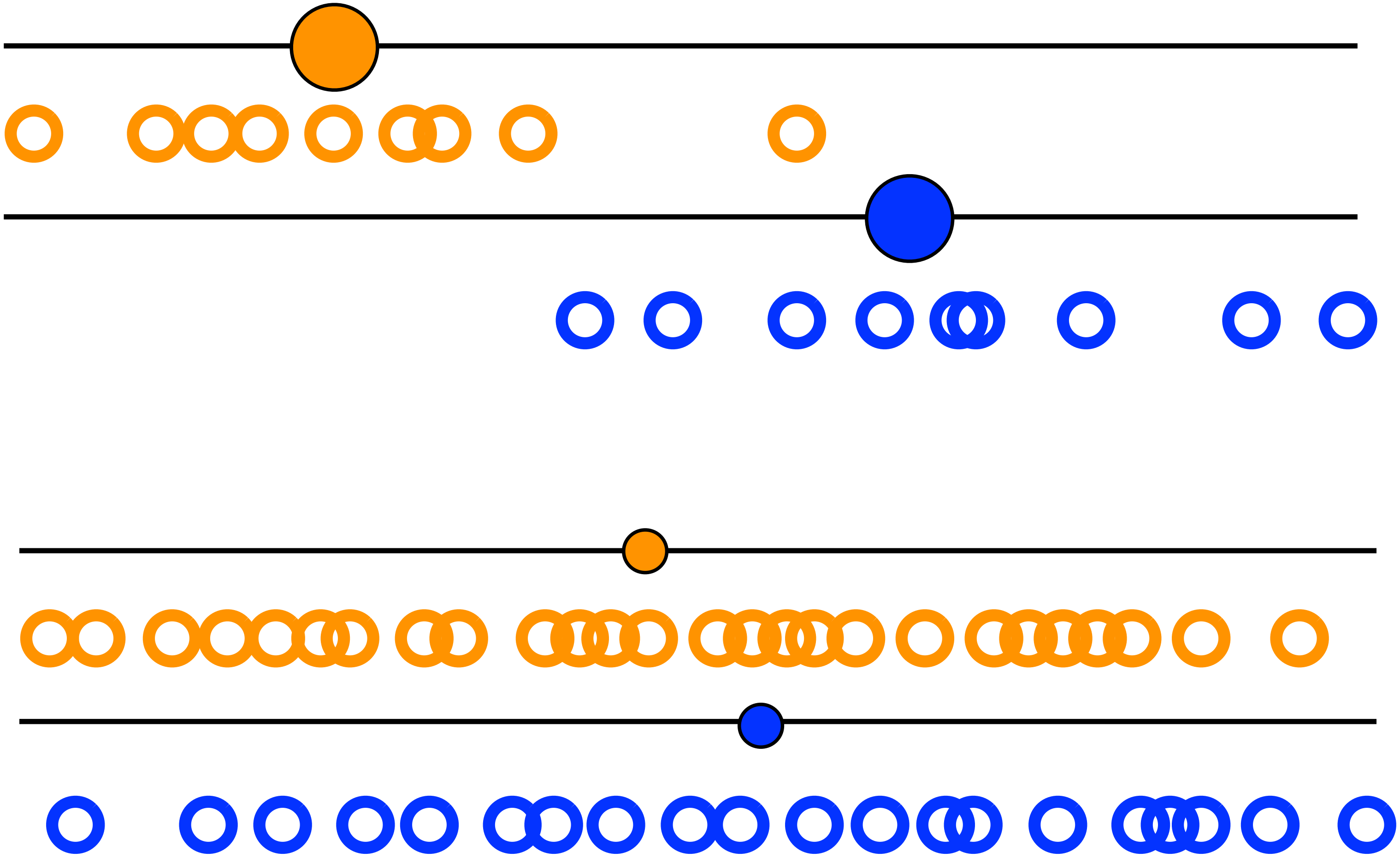
- IID
- Hypothesis testing and parameter estimation
  - Method of maximum likelihood
- Types of errors in hypothesis testing

Going from data to model (vs model to data) requires lots of assumptions and simplifications.

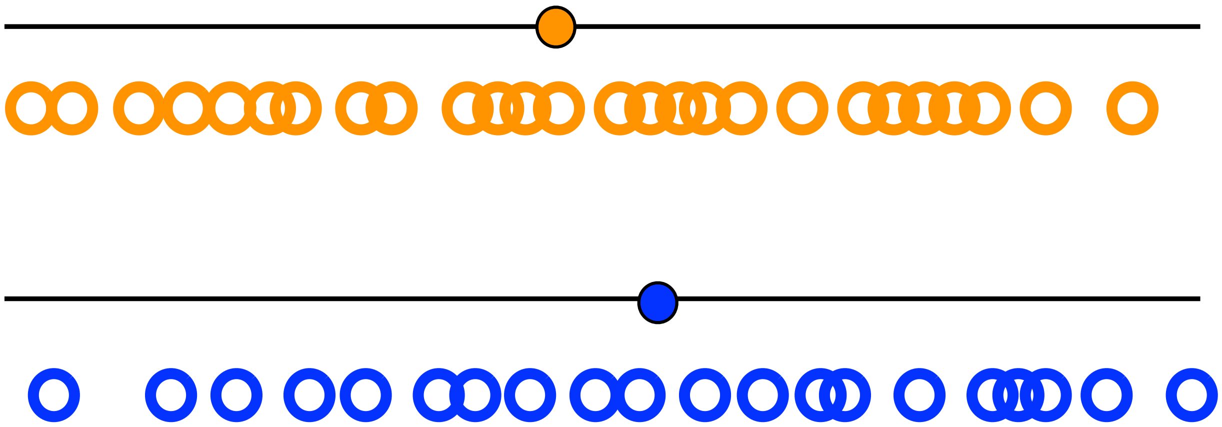


Adapted from Figure 1 of "All of Statistics".

It’s the variability that really really matters. If you don’t acknowledge and get a handle on the variance, it’s not safe to draw inferences.



Why we care about IID observations?



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

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

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

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*.

Toss a fair coin 10 times.
A = at least one head

T_j = toss j yields tails, j ∈in 1, 2, ..., 10

What's the probability of A if you toss a fair coin 10 times?

Toss a fair coin 10 times.
A = at least one head

T_j = toss j yields tails, j ∈in 1, 2, ..., 10

P(A) = 1 - P(not A)
= 1 - P(all tosses yield tails)
= 1 - P(T_1 T_2 ... T_{10}) *
= 1 - P(T_1) P(T_2) ... P(T_{10})
= 1 - 0.5^{10} ≈ 0.999

*Independence of the events T_j is critical to making this such a simple calculation!

iid

independent
identically
distributed

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

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

Independence of events or rvs makes it much easier to write down the probability of joint events or the joint distribution. Be aware of assumptions.!

iid

independent
identically
distributed

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

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

What is a distribution? what are F and G?

Are T_j events or rvs? can you define a rv? are there any parameters?

$$
\begin{aligned}
&P(\text{all tosses yield tails}) \\
&= P(T_1 T_2 \cdots T_{10}) \\
&= P(T_1)\, P(T_2) \cdots P(T_{10}) \\
&= \prod_{j=1}^{10} P(T_j)
\end{aligned}
$$

events $\longrightarrow$ T_j : toss j is a head

rv $\longrightarrow$ X_j : number of heads in toss j

iid

$$X \sim \mathit{Bernoulli}\ (0.5)$$

$$P(X = 1) = 0.5$$

$$P(X = 0) = 1 - 0.5$$

Increasing abstraction ....

Coin comes up heads with probability p. ← parameter

Toss it 10 times.

A = at least one head

T_j = toss j yields tails, j ∈in 1, 2, ..., 10

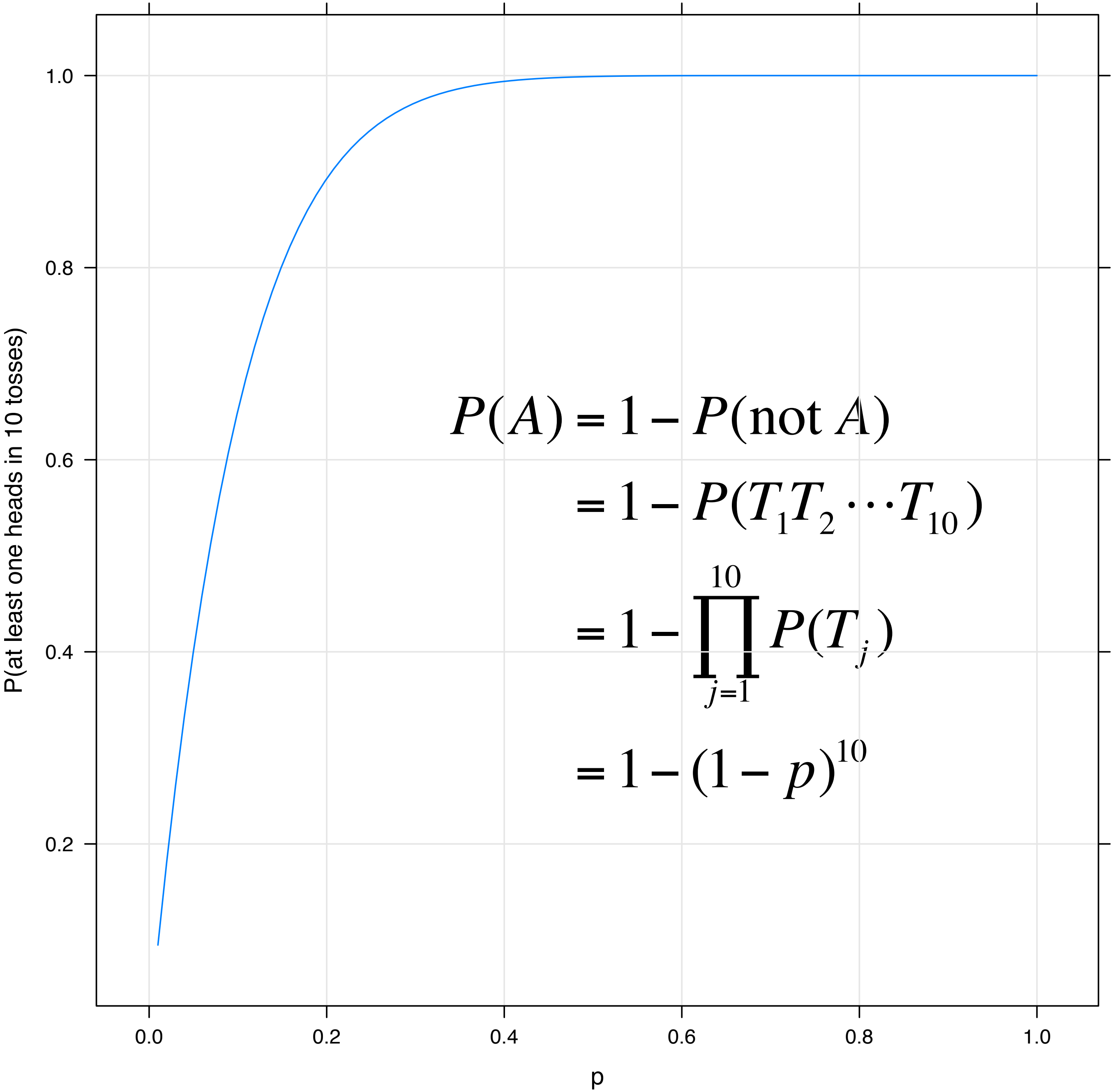
P(T_j) = 1 - p

$$P(A) = 1 - P(\text{not } A)$$
$$= 1 - P(T_1 T_2 \cdots T_{10})$$
$$= 1 - \prod_{j=1}^{10} P(T_j)$$
$$= 1 - (1 - p)^{10}$$

X : number of heads in 10 tosses

$$X \sim Bin(10, p)$$

$$P(X = 10) = (1 - p)^{10}$$

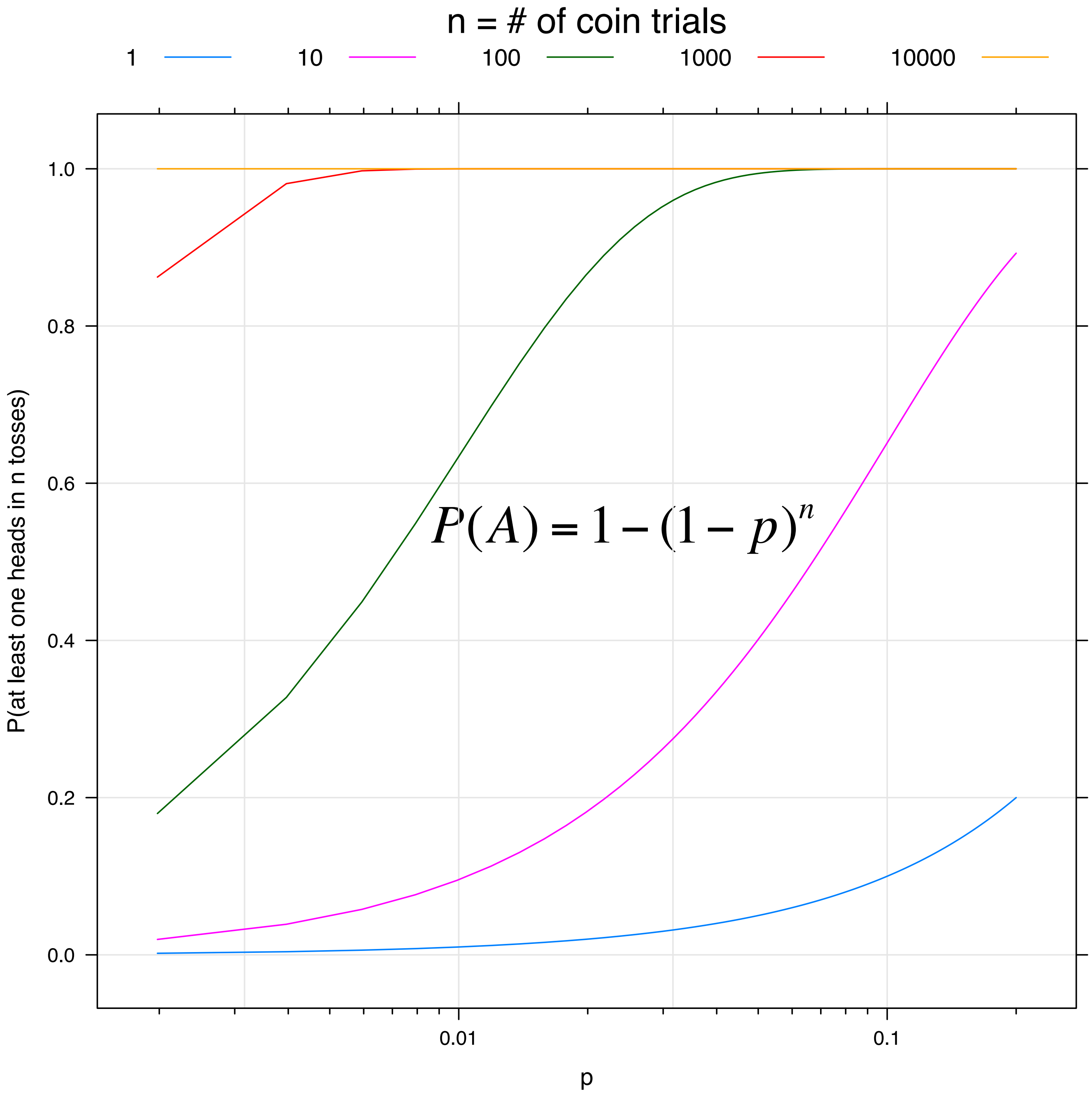


$$\begin{aligned} P(A) &= 1 - P(\text{not } A) \\ &= 1 - P(T_1 T_2 \cdots T_{10}) \\ &= 1 - \prod_{j=1}^{10} P(T_j) \\ &= 1 - (1-p)^{10} \end{aligned}$$

Increasing abstraction and sneaky foreshadowing of the incredible multiple testing problems faced in genomics........

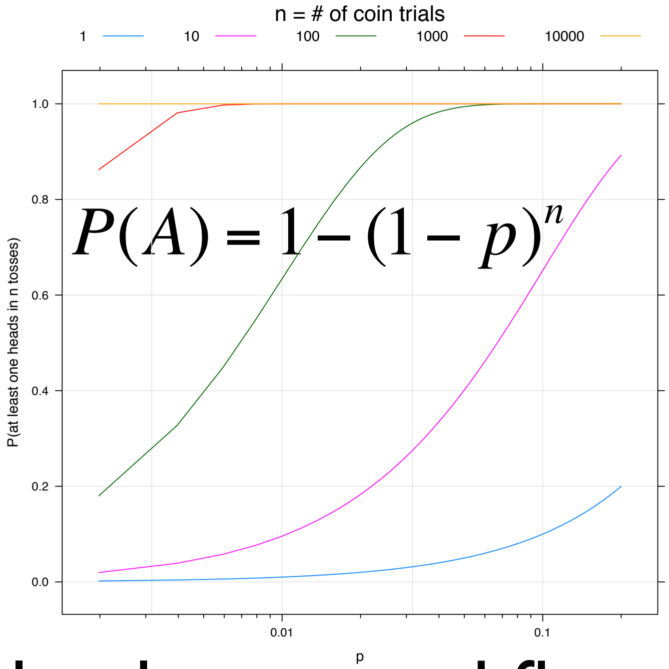
Coin comes up heads with probability p.
Toss it n times.
A = at least one head

$$P(A) = \text{<same stuff as before, really>}$$
$$= 1 - (1 - p)^n$$



n = # of coin trials
1
10
100
1000
10000
P(at least one heads in n tosses)
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.1
p

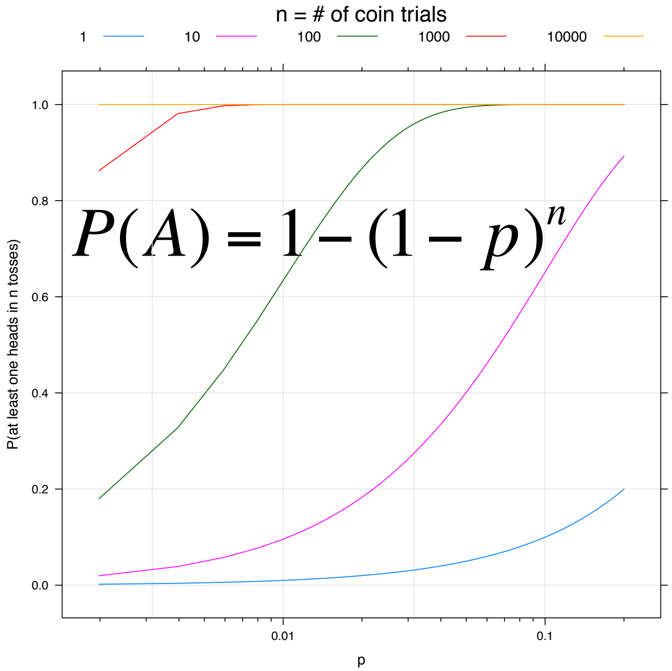
$$P(A) = 1 - (1 - p)^n$$



$$P(A) = 1 - (1 - p)^n$$

In a typical, entry-level stats workflow, test procedures will yield a false positive with a specified probability (e.g. the "alpha level").

Doing lots of tests today? Then I *guarantee* you'll get a false positive. In fact, you'll get *LOTS*.
This is the multiple testing problem and it is almost crippling in genomics. More on that later.



$$P(A) = 1 - (1 - p)^n$$

In a genomics experiment…

What if "head" = false positive = false "significant" gene

Doing lots of tests today? Then I *guarantee* you'll get a false positive. In fact, you'll get *LOTS*.
This is the multiple testing problem and it is almost crippling in genomics. More on that later.

Random variables can be characterized by a distribution

Following previous example…

X : number of heads in n tosses

$$X \sim Bin(n,p)$$ ← parameter

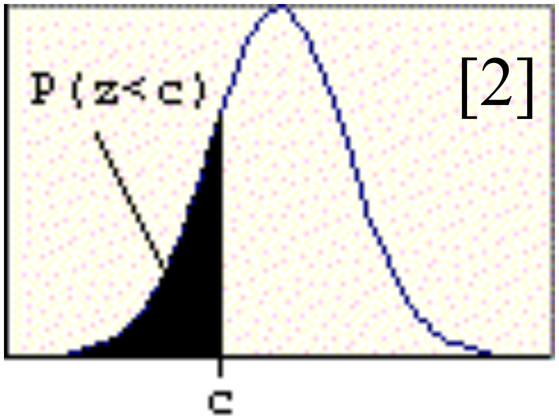
Variable

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

$$F_X(a) = P(X \leq a) = \sum_{x \leq a} p_X(x) \quad \text{(for a discrete } X\text{)}$$

$$\sum_{x=0}^{a} \binom{n}{x} p^x (1-p)^{n-x}$$ ← cumulative distribution

how to get a probability from a density



$$[1]\ P(a < Y < b) = \int_a^b f_Y(y)\,dy$$

$$[2]\ P(Y \le a) = \int_{-\infty}^a f_Y(y)\,dy$$

$$[3]\ P(Y \ge a) = \int_a^\infty f_Y(y)\,dy$$

$$[4]\ P(|Y| \ge a) = \int_{-\infty}^{-a} f_Y(y)\,dy + \int_a^\infty f_Y(y)\,dy$$

“cumulative distribution function”

“cumulative distribution function (CDF)”

$$F_Y(a) = P(Y \leq a) = \int_{-\infty}^{a} f_Y(y)\,dy \text{ (for a continuous Y)}$$

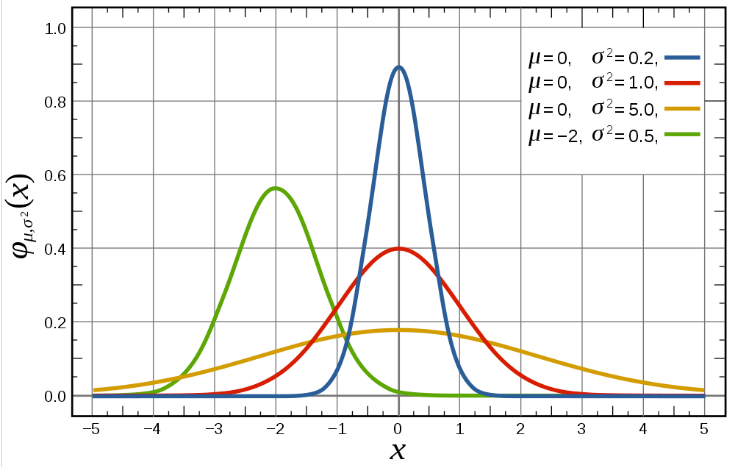
$$F_Y(a) = P(Y \leq a) = \sum_{y_i \leq a} p_Y(y_i) \text{ (for a discrete Y)}$$

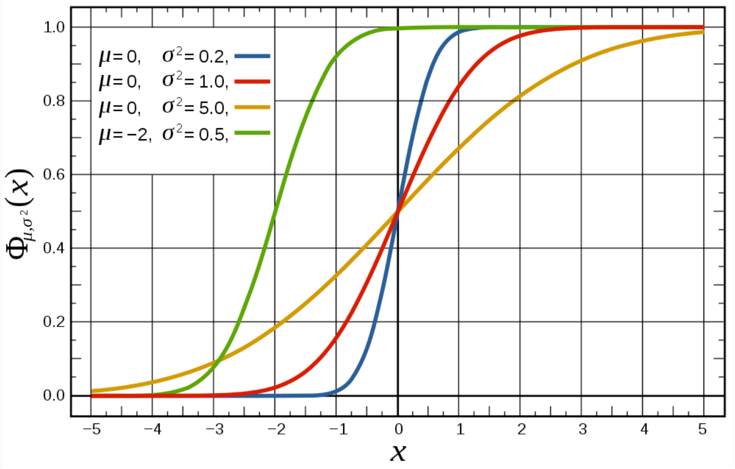
yes, we really do distinguish the density function (continuous rv) from the CDF with the deceptively subtle lowercase “*f*” vs. uppercase “*F*”

density or prob. mass function

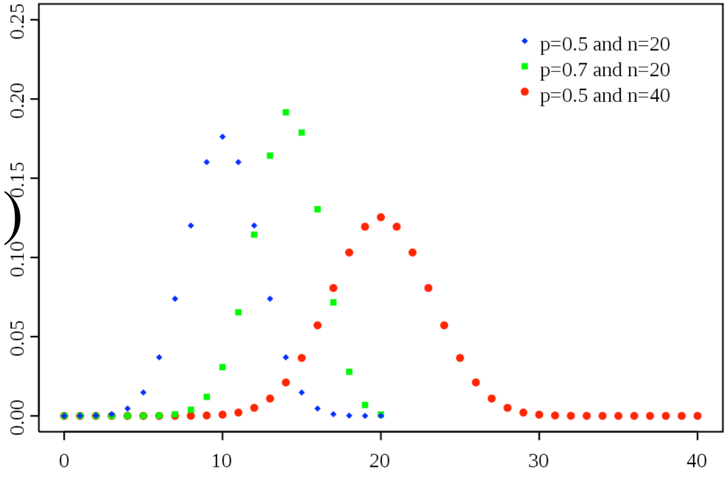
CDF

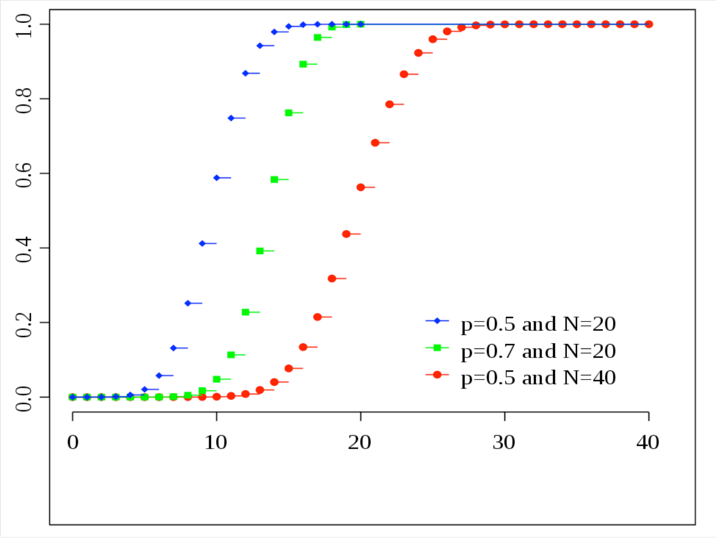
$N(\mu,\sigma^2)$





$Binom(n,p)$





sources of images on previous page

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

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

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

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

Why is it important to learn about probability distributions?

Given pmf/pdf of an r.v. X, we can:

- Compute the probability of various events, mean/variance of X, without having to perform experiments!
- We can simulate real systems and get the data.

we’re starting to leave basic probability and transition into statistical inference .....

First some vocabulary …

statistical model

very generic

$$Y \sim F$$

$$Y \sim F_\theta$$

generic but anticipating need to work with the parameter

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

example of a very specific model

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)

statistical model

the *parameter space* is the set of all possible values for the parameter

to say a model is "parametric" means the parameter space is a nice friendly Euclidean space

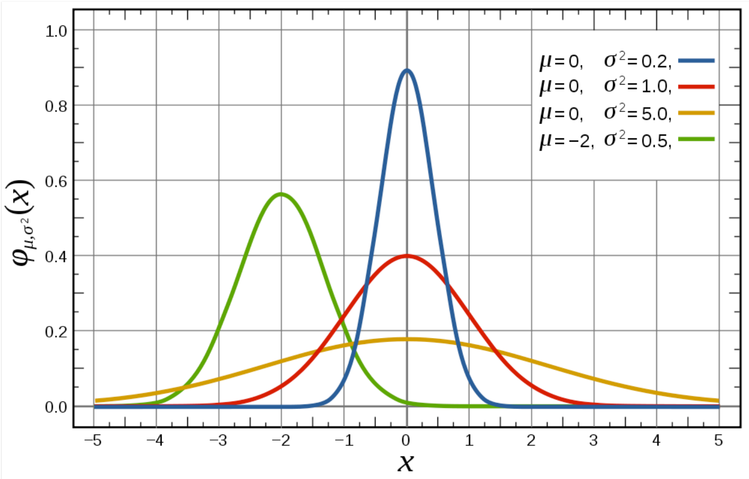
when we assume data is normally distributed about it's mean ... we're doing parametric inference; the parameter space is a nice friendly half-plane in R^2

world's favorite parametric model

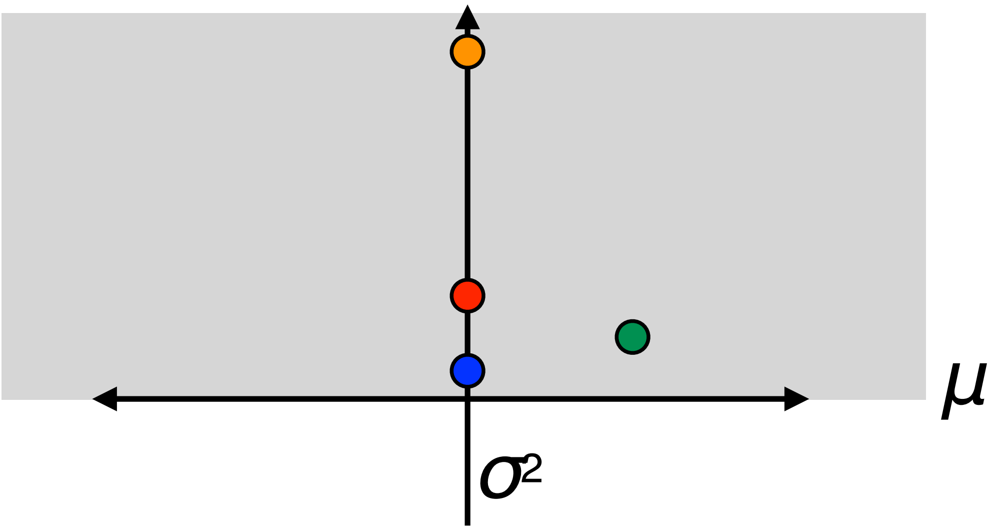
$$Y_1, \ldots Y_i, \ldots, Y_n \sim F_\theta = N(\mu, \sigma^2)$$

$$\theta = (\mu, \sigma^2)$$

the parameter space, i.e. all possible values of $\theta = (\mu, \sigma^2)$



Parameter space



parameter space = set of all possible values for the parameter

"model is parametric" ⇔ parameter space is a nice friendly Euclidean space

family	typical notation	parameter θ
<generic>	$Y \sim F_{\theta}$	θ
Bernoulli	Y ~ Bern(p)	θ = p
binomial	Y ~ Bin(n, p)	θ = (n, p)
uniform	Y ~ Unif[a, b]	θ = (a,b)
Normal	$Y \sim N(\mu, \sigma^2)$	$\theta = (\mu, \sigma^2)$
Student's t	$Y \sim t_{df}$	$\theta = df$

Parametric models we've reviewed ....

“semi-parametric” and “nonparametric” imply the parameter space isn’t a simple Euclidean space

means the parameter space is more exotic, e.g. at least partially a function space, an infinite dimensional space

BUT one does not have to feel comfortable with, say, function spaces, to *apply* nonparametric statistical methods (e.g. rank based procedures like the Wilcoxon test) responsibly

Parameter estimation

- Variables vs parameters
- Bard & Yonathan (1974) (Nonlinear Parameter Estimation. New York):
  - ...Usually a probabilistic model is designed to explain the relationships that exist among quantities which can be measured independently in an experiment; these are the variables of the model. To formulate these relationships, however, one frequently introduces "constants" which stand for inherent properties of nature. These are the parameters.

"Let $(Y_1, Y_2, \ldots, Y_n)$ be independent, identically distributed random variables."

or

"$\mathrm{Y}_i \sim F$"

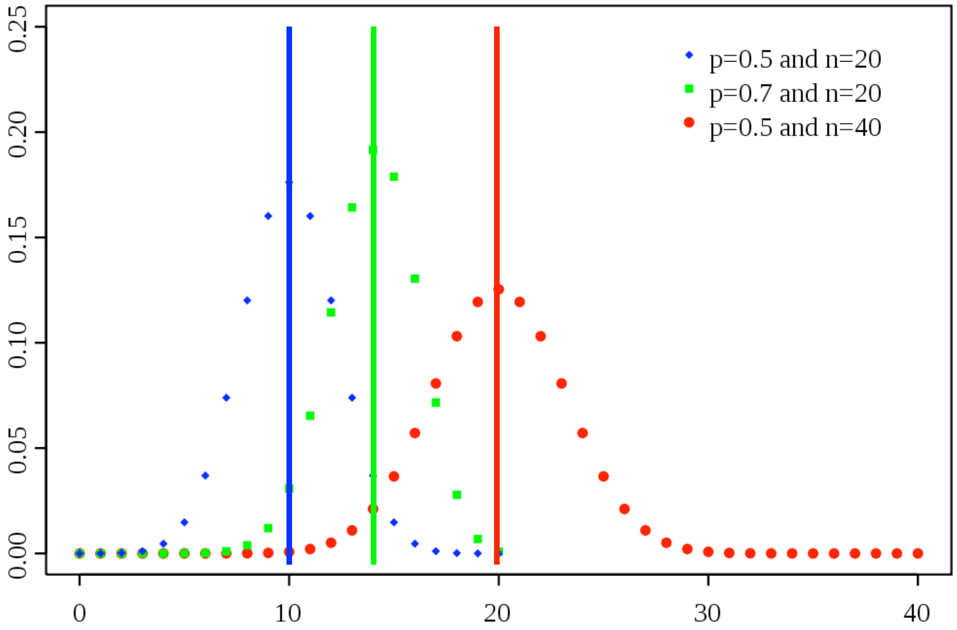
the two parameters of any distribution *F* you're mostly like to care about

#1: it's expected value (or expectation or mean)
#2: it's variance

expectation, expected value, the mean

$$E(Y) = \sum_y y p_Y(y) \text{ for discrete rv Y}$$

$$E(Y) = \int y f_Y(y)\, dy \text{ for continuous rv Y}$$



binomial example:

$$Y \sim Binom(n, p)$$

$$E(Y) = np$$

the mean is a measure of "location"
often is one of the "obvious" parameters (e.g. normal)
or is easily computed from them (e.g. binomial)

variance
standard deviation = √variance

it is a parameter

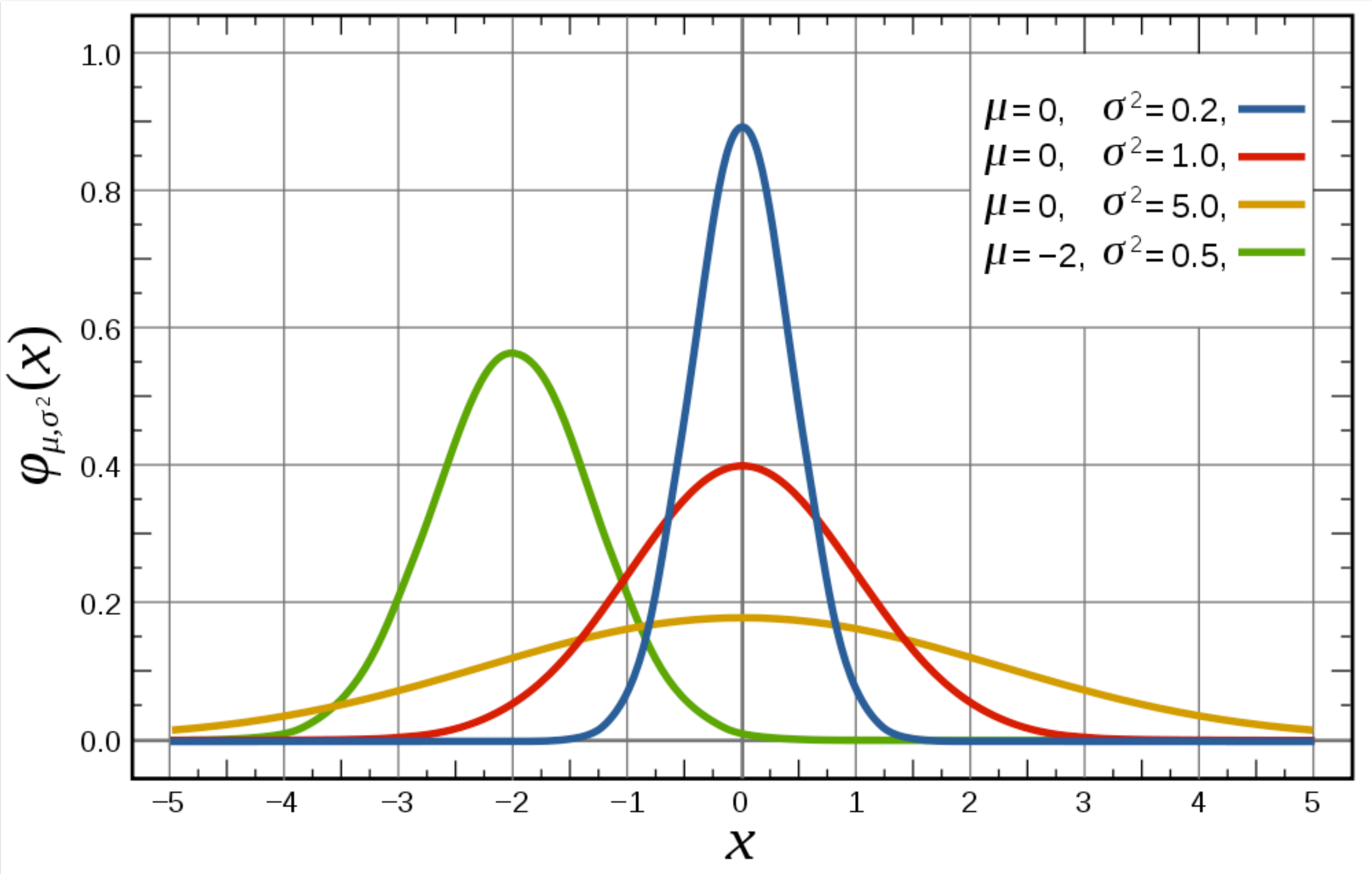
usually denoted *V(X)* or σ^2 (variance) and σ (sd)

$$V(Y) = E(Y - \mu)^2$$

common sense "definition" of variance:
a long-run average of the squared differences
between obs vals *Y* = *y* and the true mean μ

variance

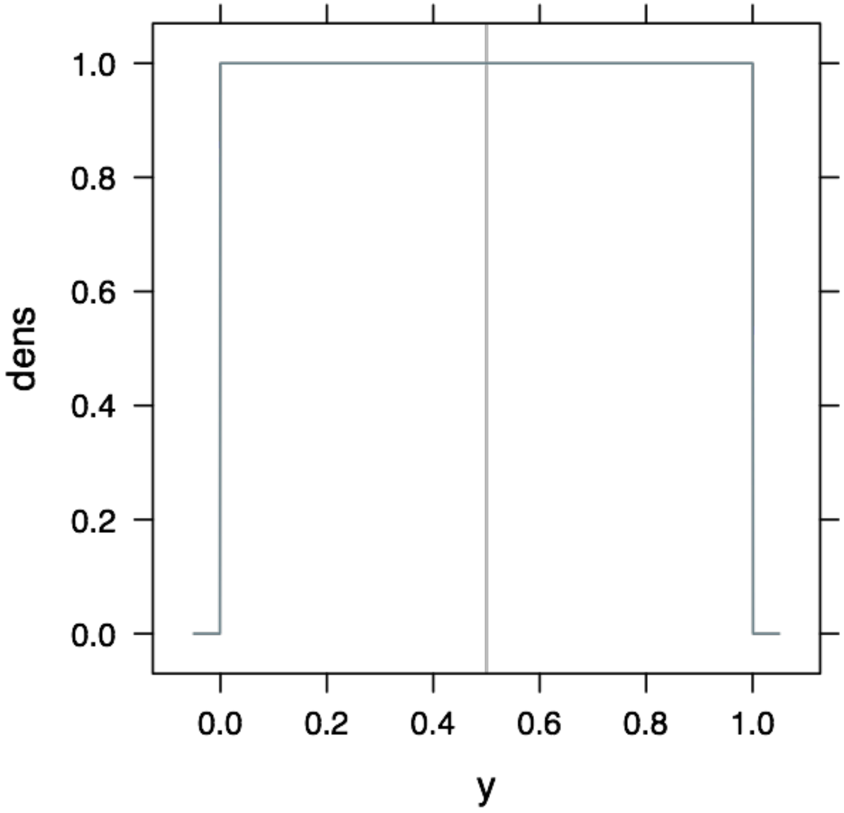
standard deviation = √variance



normal as example; bigger σ^2 ↔ bigger “spread”

Exercise: solve for mean and variance of uniform distribution Unif(a,b)

consider Y ~ Unif(0,1)
E(Y) = 0.5



First: recall that

$$\begin{aligned}\operatorname{Var}(X) &= \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2\end{aligned}$$

Answers:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \int_a^b x\frac{1}{b-a}\mathrm{d}x = \frac{1}{2(b-a)}\left[x^2\right]_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{b+a}{2}\end{aligned}$$

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \int_a^b x^2.\frac{1}{b-a}\mathrm{d}x - \left(\frac{b+a}{2}\right)^2 = \frac{1}{3(b-a)}\left[x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

we have completely arrived at statistical inference now (vs. building our probability foundation) canonical breakdown of typical statistical inference activities:

hypothesis testing vs. estimation

in either case, you are trying to say something intelligent about a parameter

hyp testing: does the true value of the parameter lie in an exciting or boring part of the parameter space?

estimation: what's your best guess at the true value of the parameter?

Parameters determine Distributions

- When sampling is from a population described by a pdf or pmf $p(x|\theta)$, knowledge of θ yields knowledge of the entire population.

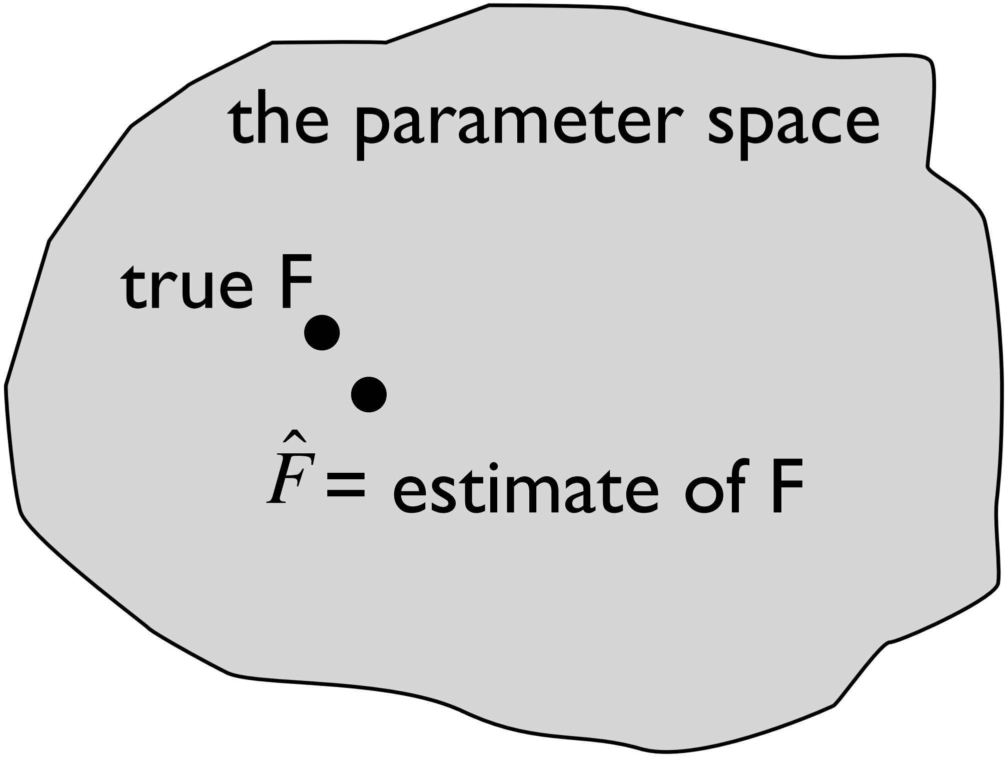
- This is why parameter estimation is useful:
  - e.g. if we are tossing a coin we would like to estimate the parameters p

estimation in generic statistical model

$Y_1, \ldots Y_i, \ldots, Y_n \sim F$

Observe data $(Y_1 = y_1, \ldots Y_i = y_i, \ldots Y_n = y_n)$.

Estimate F with $\hat{F}$.



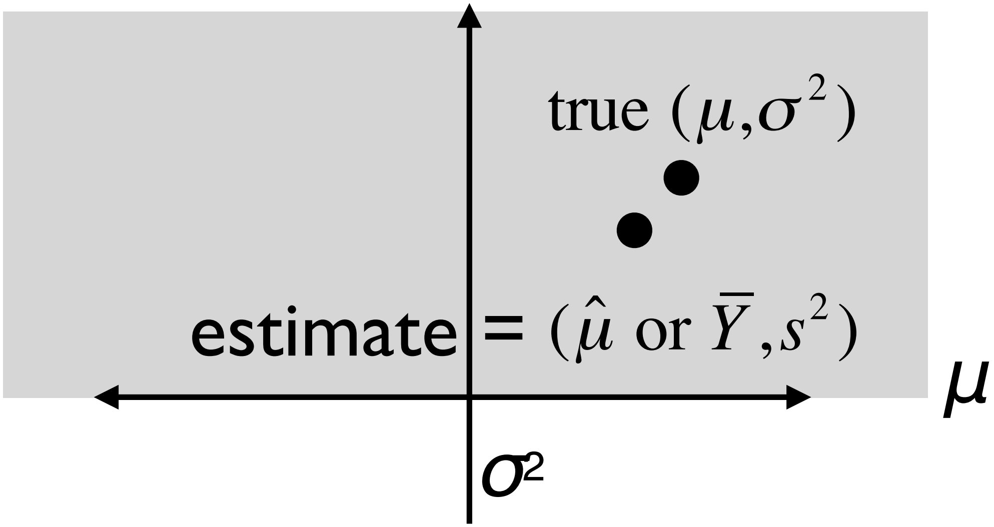
estimation in very specific statistical model

$Y_1,\ldots Y_i,\ldots,Y_n \sim F = N(\mu,\sigma^2)$

Observe data $(Y_1 = y_1,\ldots Y_i = y_i,\ldots Y_n = y_n)$.

Estimate F, i.e. estimate the mean μ and the variance σ^2 .

the parameter space



hypothesis testing in high-throughput experiments

~thousands of individual “cases” being studied in a massively parallel fashion

e.g., expression level of each individual gene in a genome under two different conditions, A and B

some genes -- presumably a small minority -- are truly “interesting” (Efron) or “alternative”, i.e. expression levels are different in condition A vs. condition B

the rest -- presumably most genes -- are truly boring (?) or “null”

hypothesis testing in high-throughput experiments

typical analytical goal:
based on observed, messy data, guess which genes are interesting and which are not *and characterize the quality of your guessing*

there's no magic from the "high-throughput" nature of this data (hurts more than helps, actually)

must begin with a clear understanding of how to do this for one gene and two conditions

then ... extend to more genes, more conditions

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

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

testing

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

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

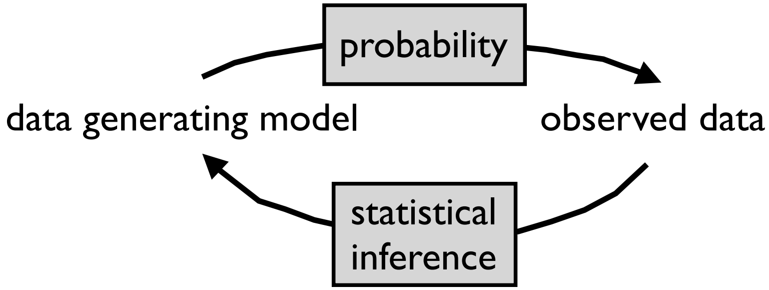
does $E_F(Y) = \mu_Y = \mu_Z = E_G(Z)$?

Call this statement the null hypothesis H_0:

$$H_0 : \mu_Y = \mu_Z$$

Or, equivalently:

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



probability

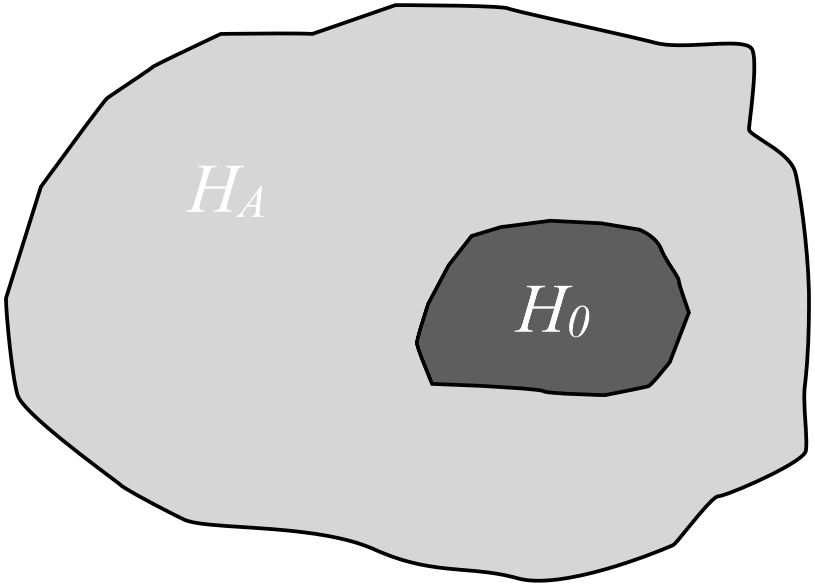
data generating model

observed data

statistical inference

statistical model

the parameter space



In formal hypothesis testing:
Define a “null (boring) region” for the parameter -- the dark gray area.
Ask whether the true value lies in that region or outside, in the “alternative (interesting) region” -- the light gray area.

testing in world's favorite statistical model

$Y_1,\ldots Y_i,\ldots,Y_{n_y} \sim F = N(\mu_Y,\sigma^2)$

$Z_1,\ldots Z_i,\ldots,Z_{n_z} \sim G = N(\mu_Z,\sigma^2)$

Observe data $(Y_1 = y_1,\ldots Y_i = y_i,\ldots Y_n = y_{n_y})$ and

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

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

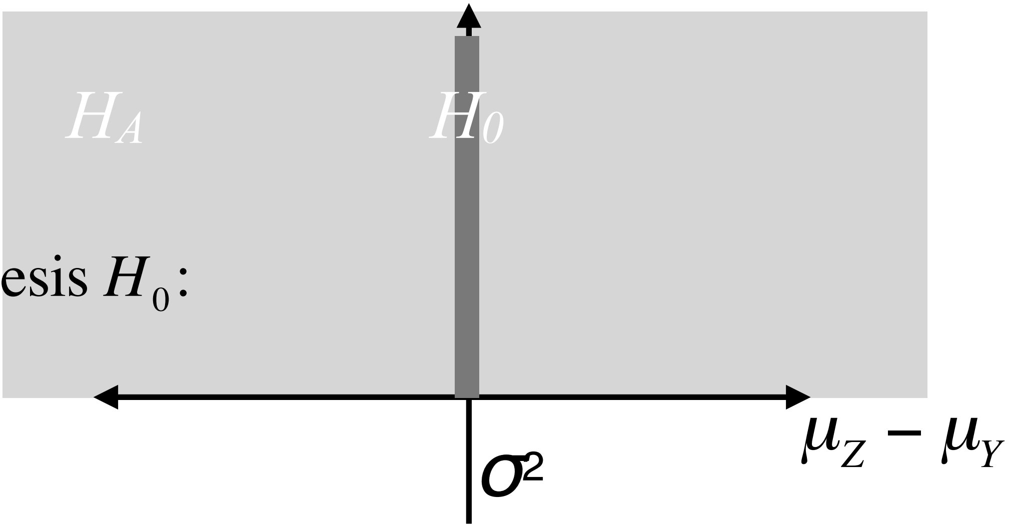
does $E_F(Y) = \mu_Y = \mu_Z = E_G(Z)$?

Call this statement the null hypothesis H_0:

$H_0 : \mu_Y = \mu_Z$

Or, equivalently:

$H_0 : \mu_Z - \mu_Y = 0$

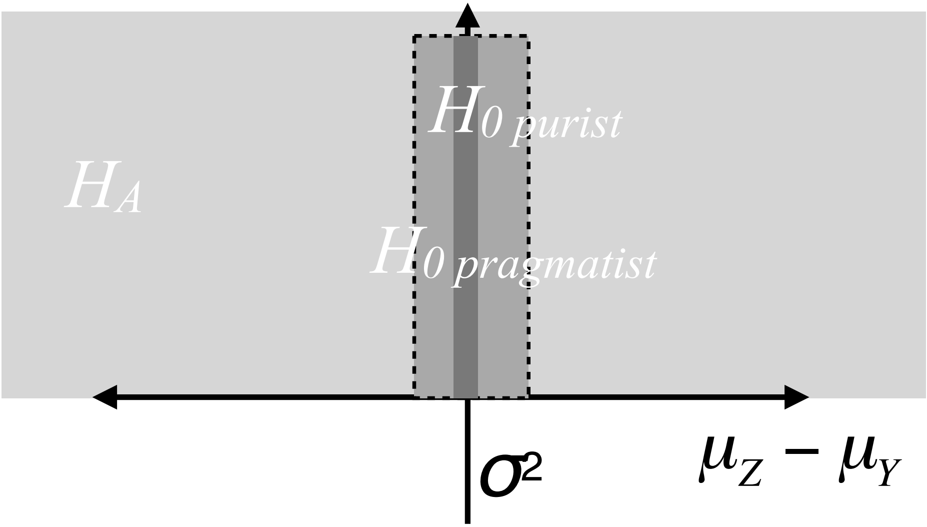


reality check re: null and alternative regions/hypotheses

"purist" defines null region as half-line where $\mu_Z - \mu_Y$ equals exactly zero

"realist" knows that the null region is a *neighborhood* around zero -- there are some differences too small to care about

"pragmatist" usually defines the null region like the "purist", because the math is so much more tractable and then accounts for concerns of "realist" when interpreting results (or, e.g., does a post hoc filter on observed difference in sample means)



Parameter estimation

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 is why parameter estimation is useful:
  - If we are tossing a coin, we would like to estimate the parameter p

Parameter estimation

- **Estimator**: rule/function whose calculated value is used to estimate the parameter
- **Estimate**: A particular realization of the estimator
- **Types of estimators:**
  - Point estimate: single number that can be regarded as the most plausible value of the parameter
  - Interval estimate: range of numbers, likely contain the true value of the parameter

Methods of point estimation

- (Methods of moments)
- Maximum likelihood estimation (MLE)
- Bayesian Inference

What are the properties of a good estimator?

- How well does the resulting estimate *explain* the "real world"?
- Proposed by geneticists/statisticians: Sir Ronald A Fisher in 1922
- Idea: we attempt to find the values of the parameters which would most likely produced the data that we in fact observed.

What is *Likelihood?*

- **Before** we perform an experiment, the outcome is unknown. Probability density function allows us to predict the probability of any outcome based on known parameters:
  - P(Data | θ)

- For example, say we know the probability of getting a head in a coin toss is *p*=0.6
  - Then we can calculate the probability of any outcome:

$$D_1 = \{HTHHHTHHHT\} \qquad P(D \mid \theta) = p^7(1-p)^3$$

$$D_2 = \{HTH\} \qquad P(D \mid \theta) = p^2(1-p)$$

$$D_3 = \{TTTH\} \qquad P(D \mid \theta) = p^3(1-p)$$

What is *Likelihood?*

- **After** the experiment is done, we know the outcome. Now we want to know the *likelihood* that a given parameter value would generate the outcome:

  L (Data | θ): p(Data | θ)

- **Estimate** θ by finding the value of θ that makes the data most *likely* (our estimate: $\hat{\theta}$)

The coin example

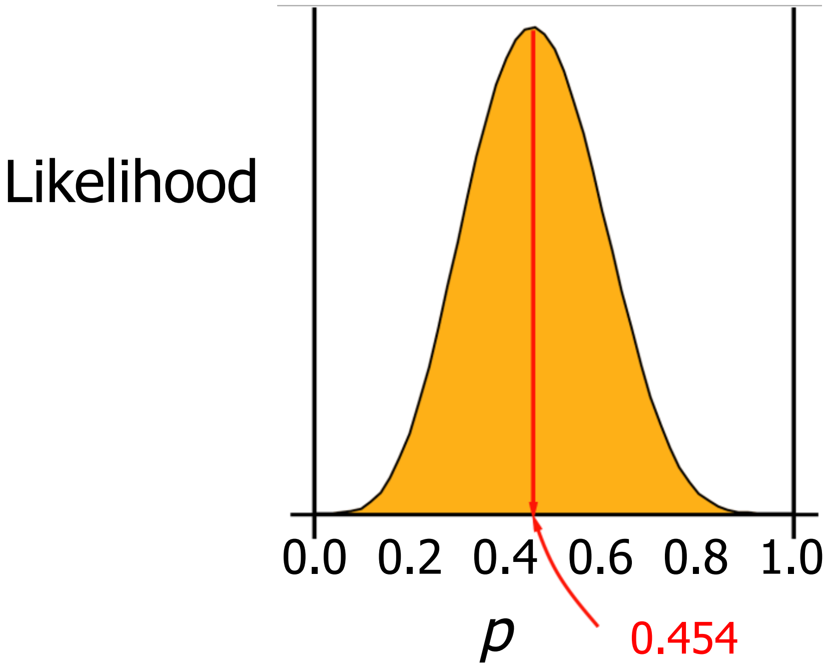
- We have data from 11 tosses of a coin (we don't know *p* the probability of head)
  - RV = outcome is head
- Outcome of the experiment: {HHTHTTTHTTH}
- Probability of the outcome of the experiment:

  $pp(1-p)p(1-p)(1-p)(1-p)p(1-p)(1-p)p$
- The likelihood is L(Data| *p*)=$p^5(1-p)^6$

- We can plot the data against it's likelihood to figure out when we reach the maximum of the likelihood function.

Likelihood function

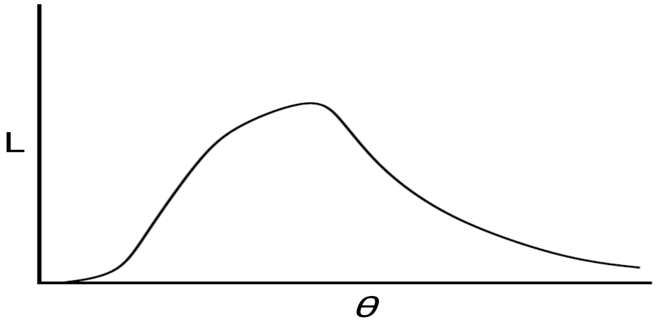
The likelihood is $L(\text{Data} \mid p) = p^5(1-p)^6$

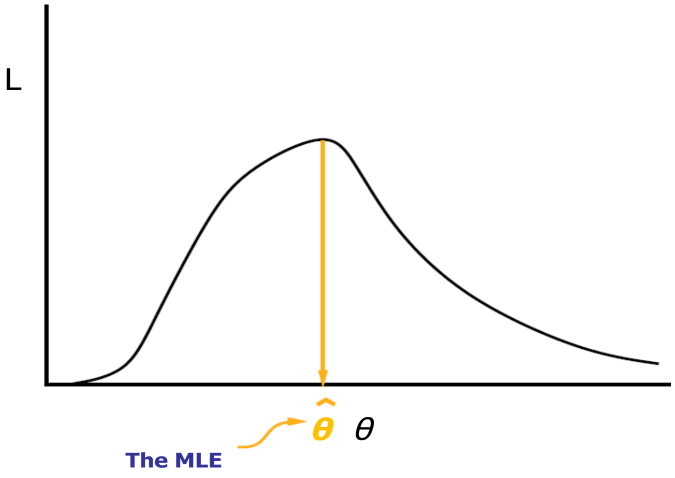


The likelihood function

- A function of the parameter(s) of our model for the observed data.

- We want to find parameters that result in the maximum of the likelihood function.





- Often the math is easier to deal with if we take the log of the likelihood function
  - Log (L) achieves its maximum at the same parameter values as L

- Note that “simple” (i.e., convex) likelihood function achieve their maximum at one parameter setting; non-convex likelihood functions have multiple local maxima

Solving for the solution of the maximum likelihood problem:

- General problem: we want to find the parameter settings that maximize some function given our data.
  - Log L = Log ($p^5.(1-p)^6$) = 5x log(p) + 6 log(1-p)
- Differentiate the log L function and set derivative to zero.
- We will arrive at p = 5/11

World view according to Bayesians

- Classic philosophy (frequentist) assumes that parameters are ***fixed*** quantities that we want to estimate as precisely as possible.

- Bayesian perspective is different: parameters are random variables with probability assigned to particular values of parameters to reflect the degree of evidence for that value.

Properties of a good estimator

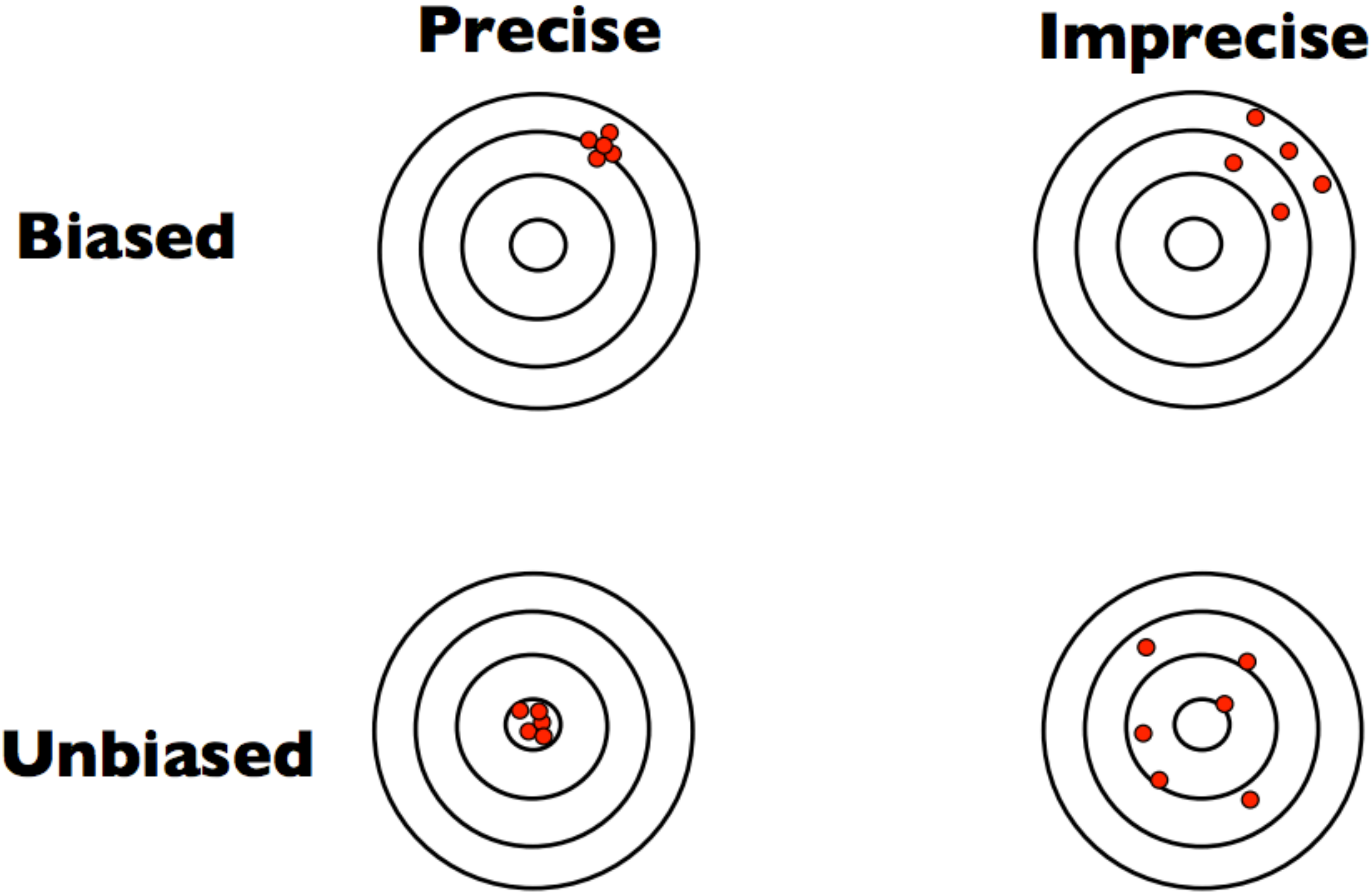
- Consistent: as sample size increases estimate approaches true parameter

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

- Unbiased $E[\hat{\theta}] = \theta$

- Precise: sampling distribution of estimate should have small standard error.

Bias vs. Precision



Bayesian estimation

- In order to make probability statements about θ we make use of Baye’s rule:

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$

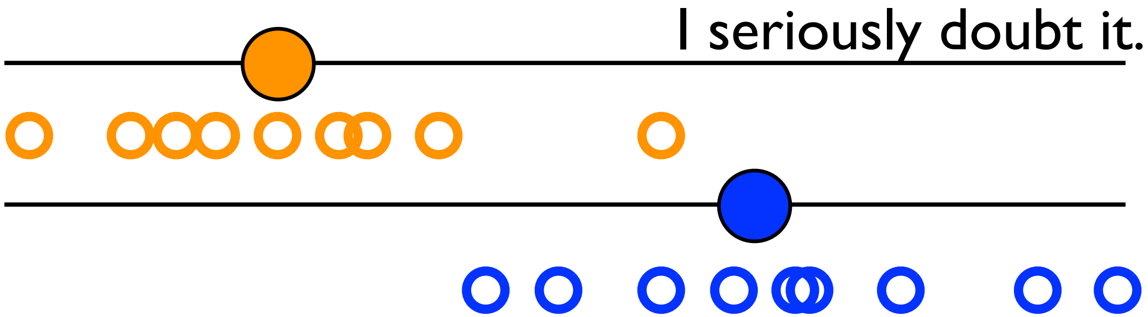
$$\boxed{P(\theta \mid D) \propto P(\theta)P(D \mid \theta)}$$

Posterior ∝ **Prior** × **Likelihood**

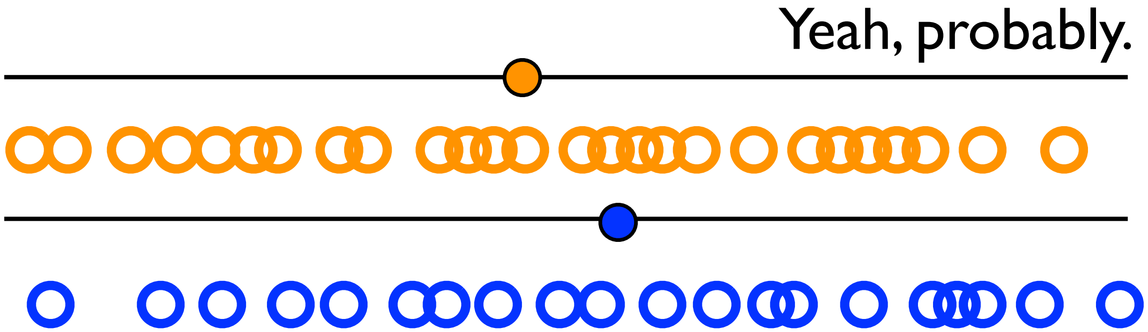
- Find θ, such as posterior is maximized

Back to hypothesis testing …

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



I seriously doubt it.



Yeah, probably.

- A *statistic* is a rv that's a function of the data.

- Classic examples:
  - The sample mean
  - The sample variance

- Two main reason why we love them:
  - Sometimes they are *estimators* for parameters of a model we care about (i.e., trying to model)
  - Sometimes they are *test statistics.* i.e., the basis for a hypothesis test

Properties of a test statistic

- When observed value (based on our sample) is “big” or “extreme”, suggests that observed data is very unexpected under the null hypothesis H_0

- We know the distribution of the test statistic under the null model: so we can compute a pvalue quantifying the incompatibility between observed value of test statistic and H_0

- Point estimate: single best guess of the parameter

- Interval estimate (e.g., confidence interval) provides a range of possible values for the parameters.

- Constructing the interval estimator requires knowledge of the estimator's distribution

- Therefore …

- To complete a hypothesis test, we need a statistic's *sampling distribution*
- "sampling" -- "hypothetical long repeats of the experiment"
- *Standard error*: standard deviation of the sampling distribution of an estimator.

- E.g., The standard error of the mean (SEM) (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population

- p-value …

- The probability under the null H_0 of observing a test statistic *value* as or more extreme than the one computed from the data.

- Two-sided test: both very small and very large values are considered extreme

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

- Musing on p-values

- In some sense, it’s laziness to work this way: easy because we only need to characterize the distribution of the test statistic under the null

- downside: an indirect measure of how “interesting” the data is

- Just saying something is not “null” or not “boring” is not exactly equivalent to saying what’s truly “exciting” about it.

Errors in hypothesis testing

- p-values will eventually be thresholded to make decisions

p-value exceeds threshhold	... does not
hit	not hit
statistically significant	not statistically significant
discovery!	?
reject H_0	accept H_0 (wince) fail to reject H_0 (roll eyes)

confusion matrix

“call” based on obs. data true state of nature	“not hit”	reject H_0 “hit”	
H_0 holds	true negatives	false positives	# nulls
H_A holds “interesting”	false negatives	true positives	# alts
		discoveries	# genes

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	true negatives	false positives Type I errors	# nulls
H_A holds "interesting"	false negatives Type II errors	true positives	# alts
		discoveries	# genes

Should you care about false positive rate or false negative rate?

- setting of alpha allows us to trade-off between FN rate and FP rate.
- False negative is preferred over false positive:
  - e.g., legal proceeding
- False positive is preferred over false negative:
  - E.g., quarantining people that are suspected to have acquired an infectious disease.