

STAT 540

Class meeting 03

Monday, January 12, 2015

Dr. Gabriela Cohen Freue

Department of Statistics

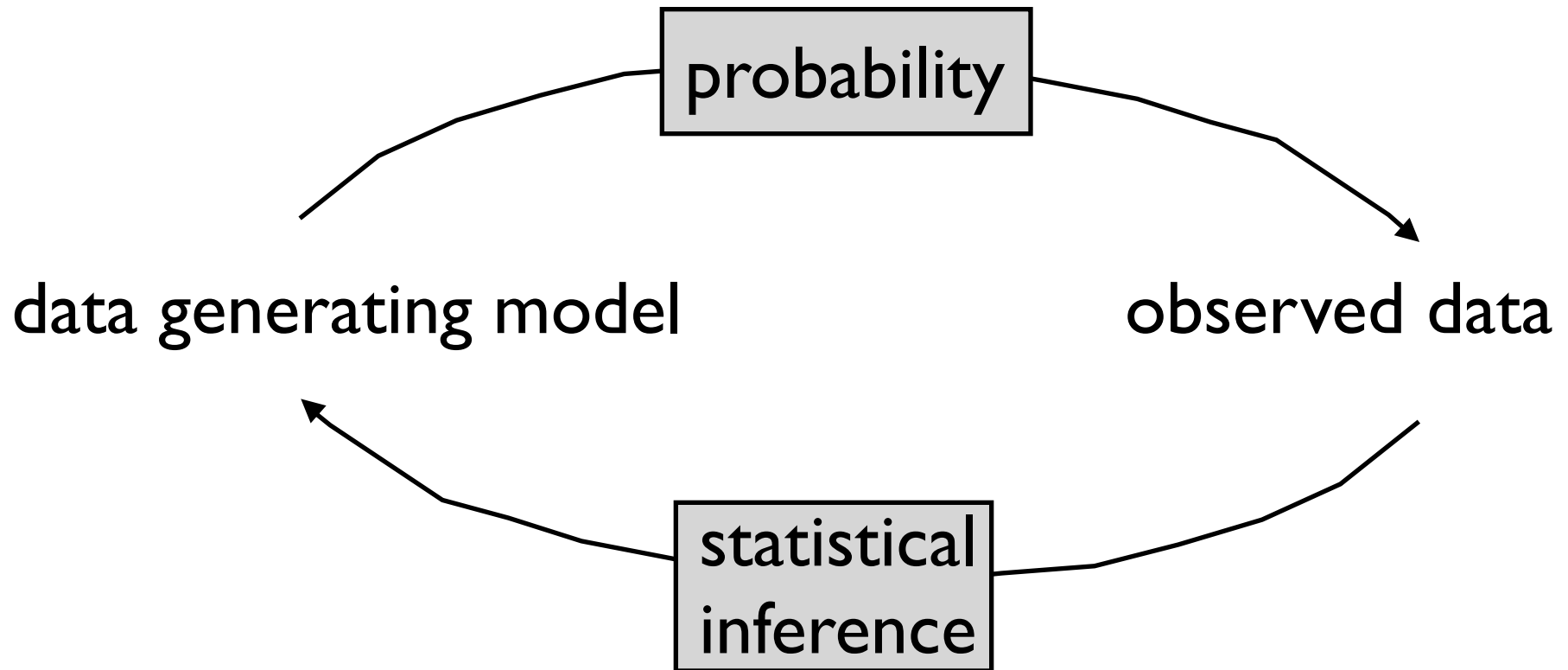
(adapted from Dr. Jenny Bryan's preparation)

Introduction to statistical inference, part 2

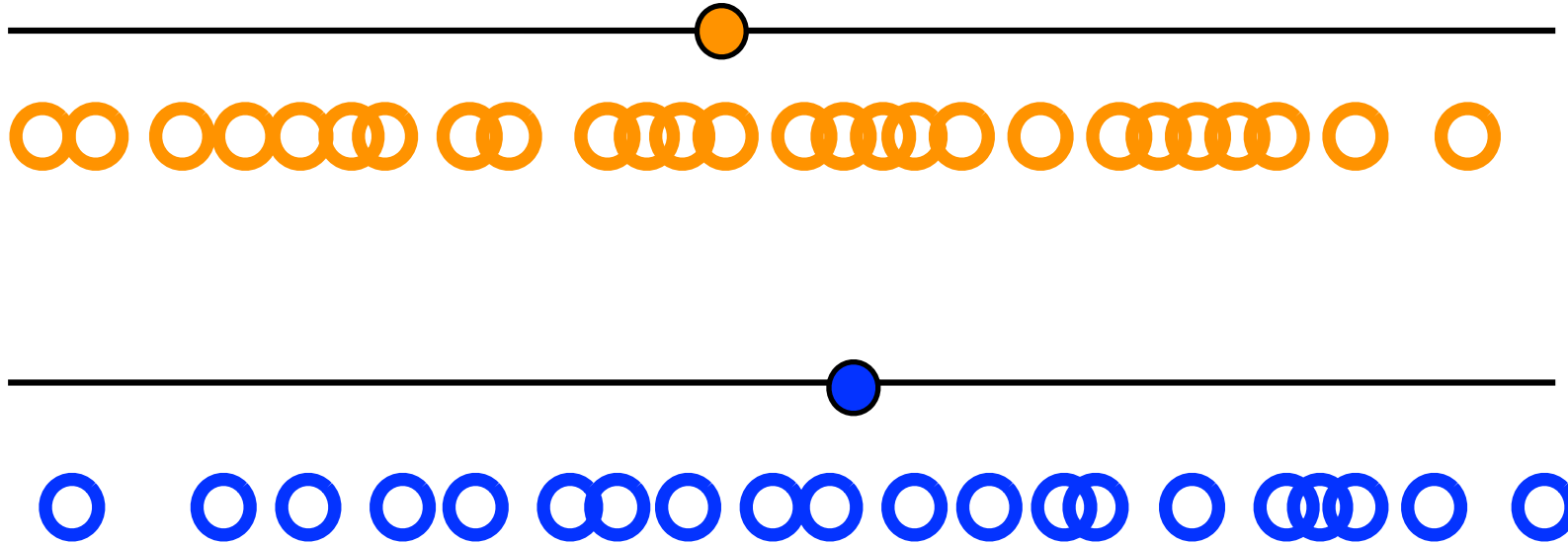
... end of basic probability

Hypothesis testing and estimation

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



From data to model...



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

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

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

What do we mean by iid?

THE THINGS LISTED BELOW ARE DIFFERENT!

THE DIFFERENCES MATTER VERY VERY MUCH!

events, e.g. a of a fair coin yields a tail

random variable, e.g. X

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

parameter, e.g. probability p of success in a binomial rv

Distinguish between what's random but attainable
(actual data) vs. the unknown but ultimately important
true state of nature (parameters).

iid

independent
identically
distributed

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

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

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, $j \in 1, 2, \dots, 10$ ^{“in”}

$$\begin{aligned} P(A) &= 1 - P(\text{not } A) \\ &= 1 - P(\text{all tosses yield tails}) \\ &= 1 - P(T_1 T_2 \dots T_{10}) \\ &= 1 - P(T_1) P(T_2) \dots P(T_{10}) * \\ &= 1 - 0.5^{10} \approx 0.999 \end{aligned}$$

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

$T_j = \text{toss } j \text{ yields tails}, j \in 1, 2, \dots, 10$

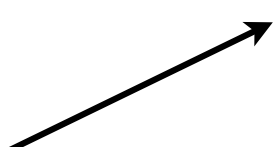


Subscripts will be used often to deal with groups of things that should be handled similarly.

In this case, handy for defining 10 events at once.

Pronounced “Tee jay” or “Tee sub jay”

Resist the temptation to tune out when math notation arises (if you’re prone to that).

$$\begin{aligned}
& P(\text{all tosses yield tails}) \\
&= P(T_1 T_2 \cdots T_{10}) \\
&= P(T_1) P(T_2) \cdots P(T_{10}) \\
&= \prod_{j=1}^{10} P(T_j) = \prod_j P(T_j) = 1 - 0.5^{10}
\end{aligned}$$


Subscripts and “mass production” symbols for addition (\sum) and multiplication (\prod) make a winning combination!

iid

independent
identically
distributed

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

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

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

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

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

What is a distribution? what are F and G ?

$$\begin{aligned}
& P(\text{all tosses yield tails}) \\
&= P(T_1 T_2 \cdots T_{10}) \\
&= P(T_1) P(T_2) \cdots P(T_{10}) \\
&= \prod_{j=1}^{10} P(T_j) = \prod_j P(T_j) = 1 - 0.5^{10}
\end{aligned}$$

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

$$\begin{aligned}
& P(\text{all tosses yield tails}) \\
&= P(T_1 T_2 \cdots T_{10}) \\
&= P(T_1) P(T_2) \cdots P(T_{10}) \\
&= \prod_{j=1}^{10} P(T_j) = \prod_j P(T_j) = 1 - 0.5^{10}
\end{aligned}$$

events $\longrightarrow T_j : \text{toss } j \text{ is a head}$

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

iid $X \sim \text{Bernoulli}(0.5)$

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

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

Increasing abstraction

Coin comes up heads with probability $p \leftarrow$ parameter

Toss it 10 times.

A = at least one head

T_j = toss j yields tails, $j \in 1, 2, \dots, 10$

$$P(T_j) = 1 - p$$

$$P(A) = 1 - P(\text{not } A)$$

$$= 1 - P(T_1 T_2 \cdots T_{10})$$

$$= 1 - \prod_{j=1}^{10} P(T_j)$$

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

another rv?

distribution?
parameters?

Increasing abstraction

Coin comes up heads with probability $p \leftarrow$ parameter

Toss it 10 times.

A = at least one head

T_j = toss j yields tails, $j \in 1, 2, \dots, 10$

$$P(T_j) = 1 - p$$

$$P(A) = 1 - P(\text{not } A)$$

$$= 1 - P(T_1 T_2 \cdots T_{10})$$

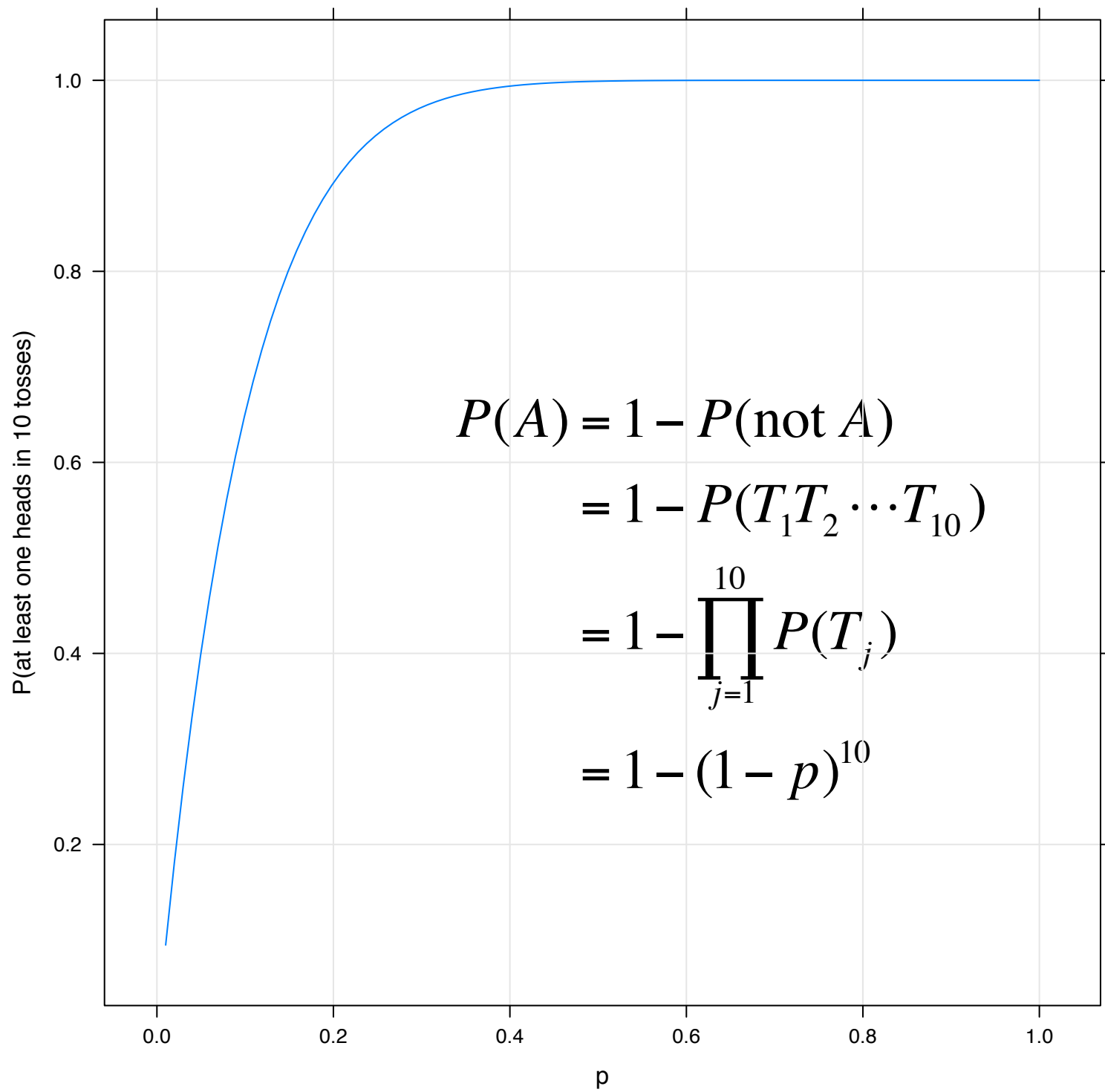
X : number of heads in 10 tosses

$$= 1 - \prod_{j=1}^{10} P(T_j)$$

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

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

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



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

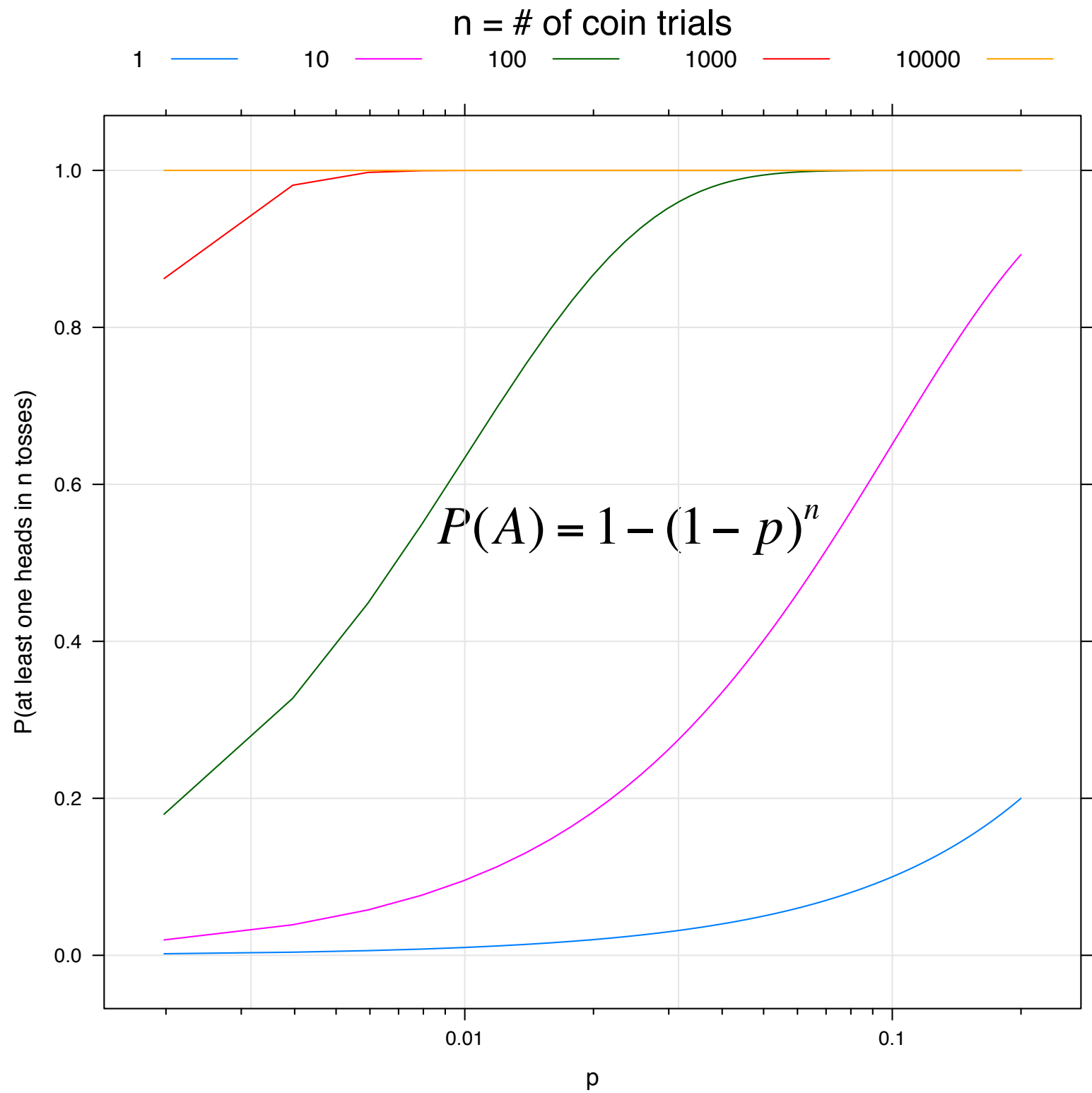
Coin comes up heads with probability p .

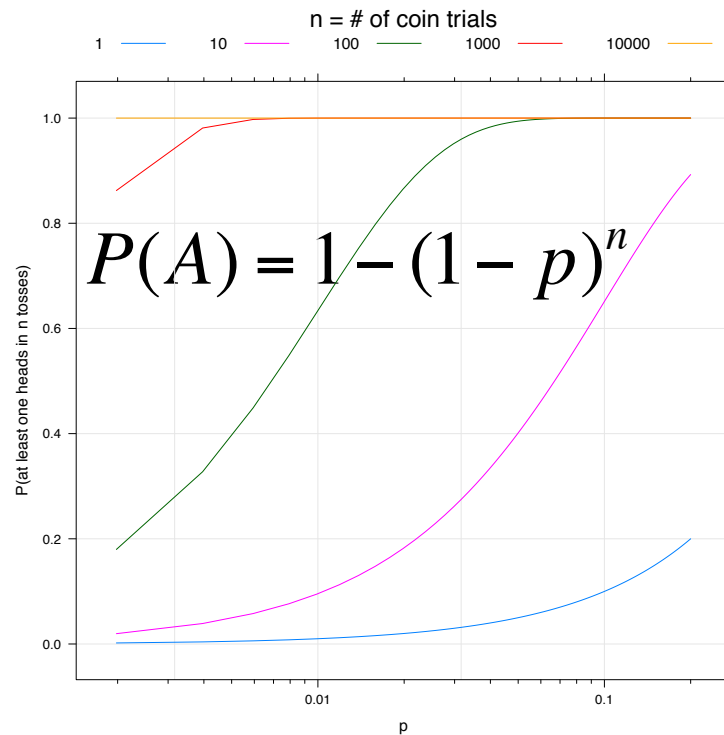
Toss it n times.

A = at least one head

$$P(A) = \text{<same stuff as before, really>}$$

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





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.

iid

independent
identically
distributed

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

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

What is a distribution? what are F and G ?

Random variables can be characterized by a distribution

Following previous example...

X : number of heads in n tosses

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

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

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

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

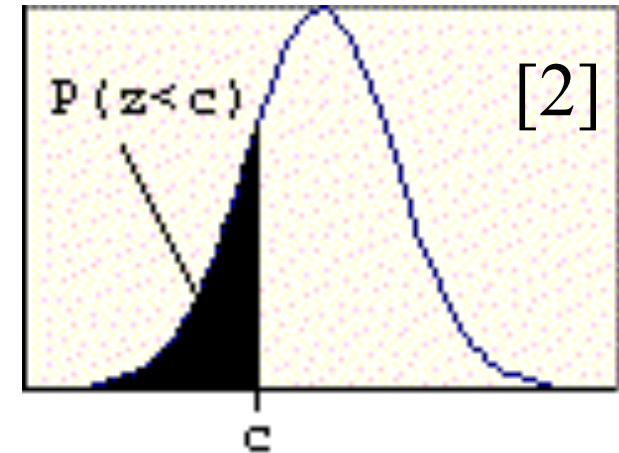
how to get a probability from a density

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

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

$$[3] P(Y \geq a) = \int_a^{\infty} f_Y(y) dy$$

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



“cumulative distribution function”

“cumulative distribution function (CDF)”

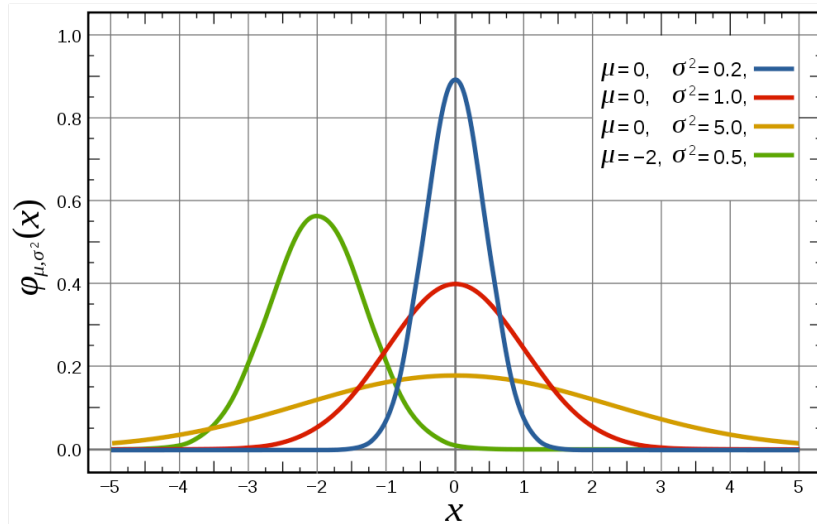
$$F_Y(a) = P(Y \leq a) = \int_{-\infty}^a f_Y(y) dy \text{ (for a continuous Y)}$$

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

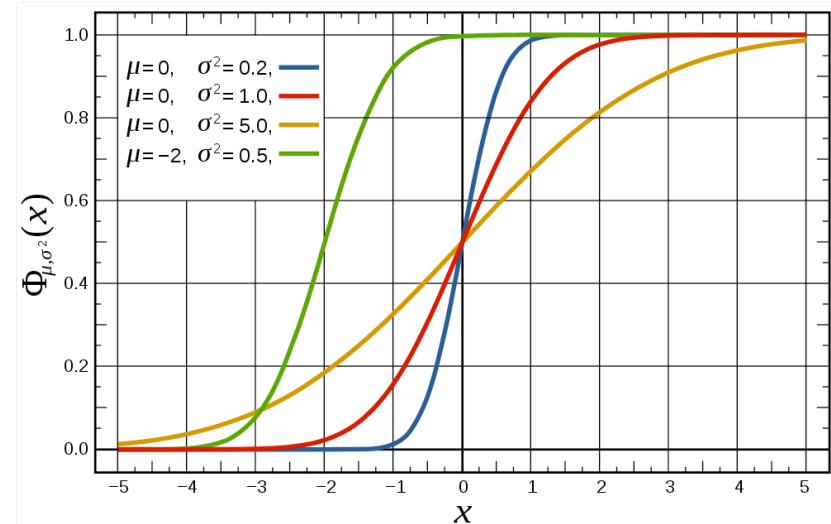
yes, we really do distinguish the density function (continuous rv) from the CDF with the deceptively subtle lowercase “ f ” vs. uppercase “ F ”

density or prob.
mass function

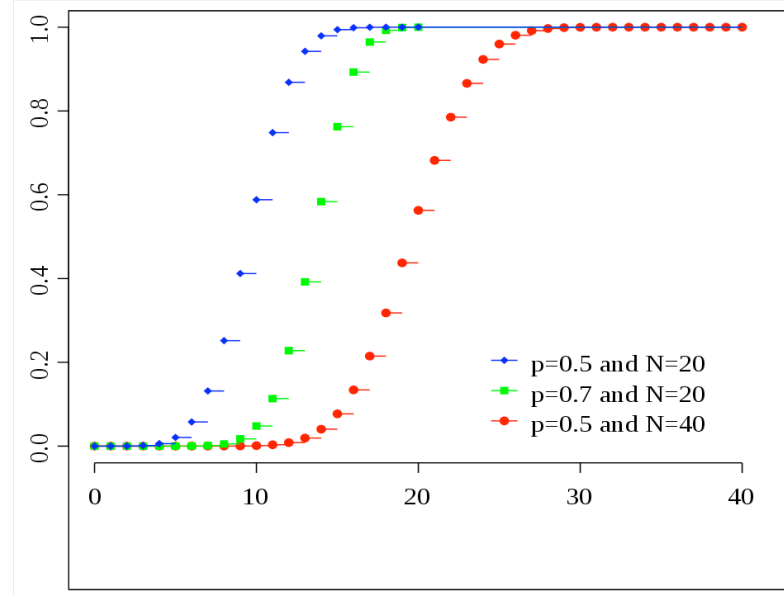
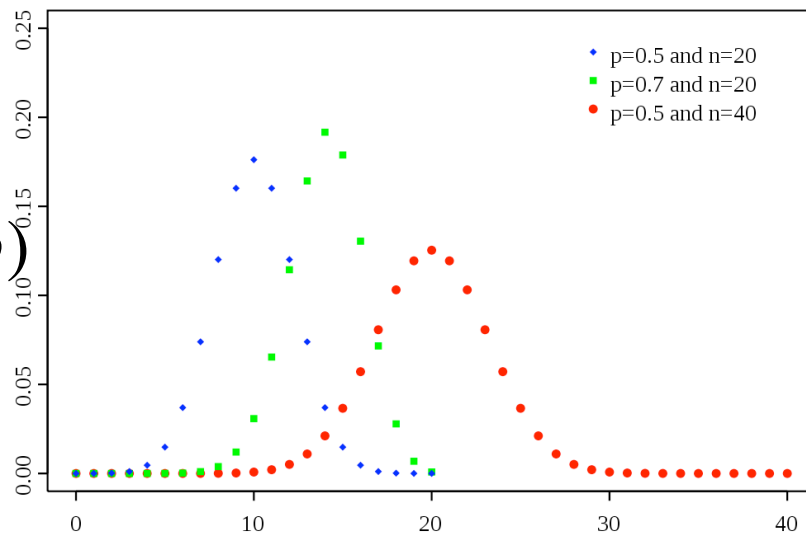
$$N(\mu, \sigma^2)$$



CDF



$$Binom(n, p)$$



sources of images on previous page

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

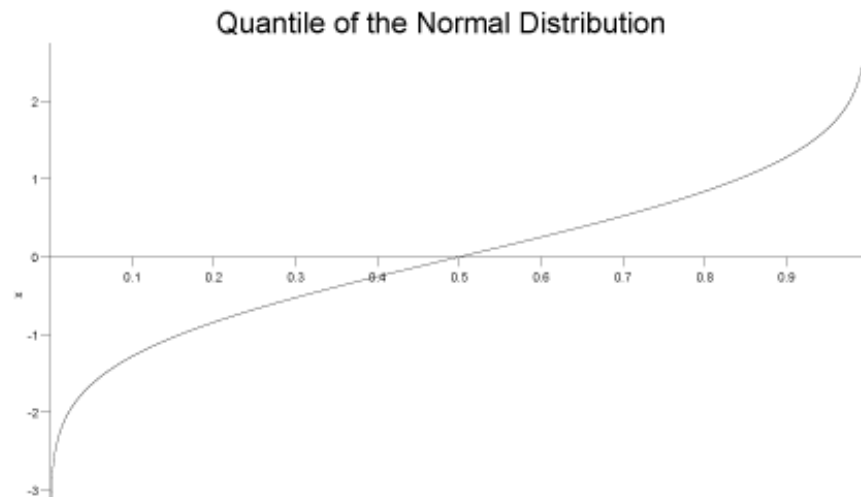
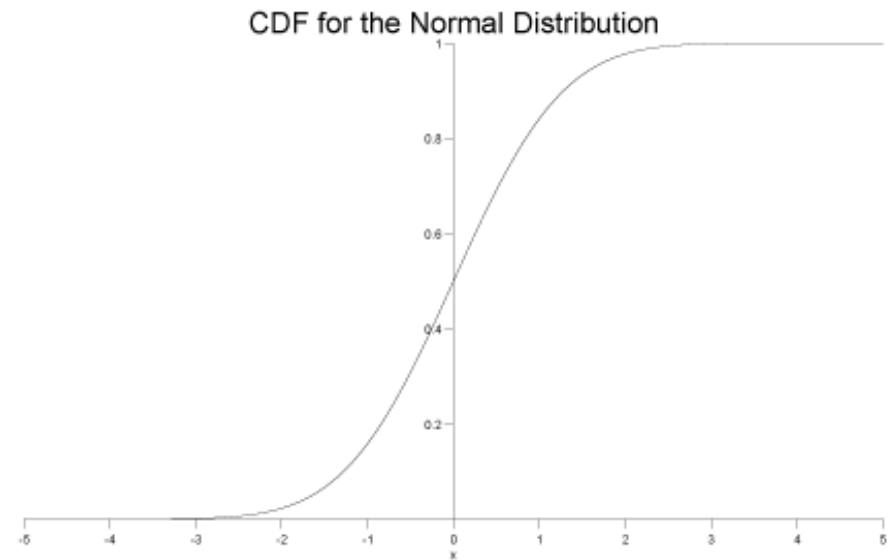
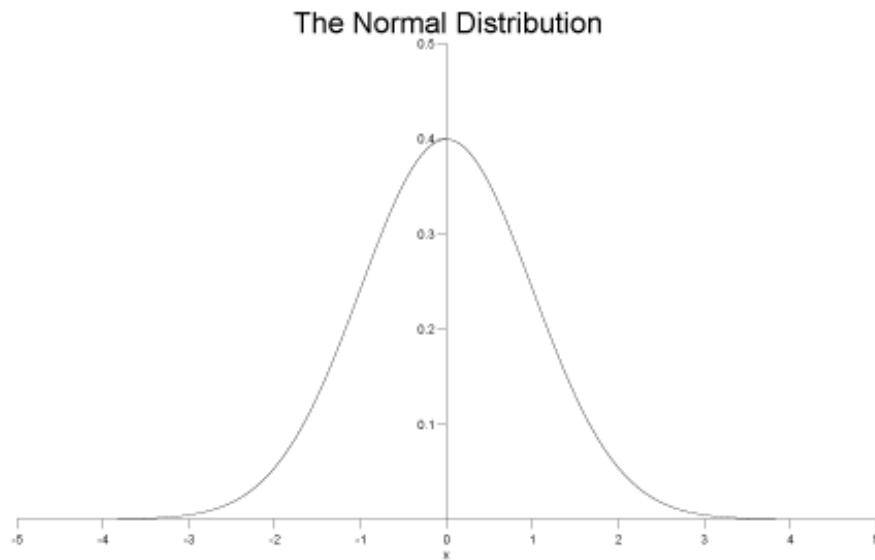
http://en.wikipedia.org/wiki/File:Normal_Distribution_CDF.svg

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

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

inverse CDF, quantile function

$$F_Y^{-1}(q) = \text{smallest } y \text{ such that } F_Y(y) > q$$



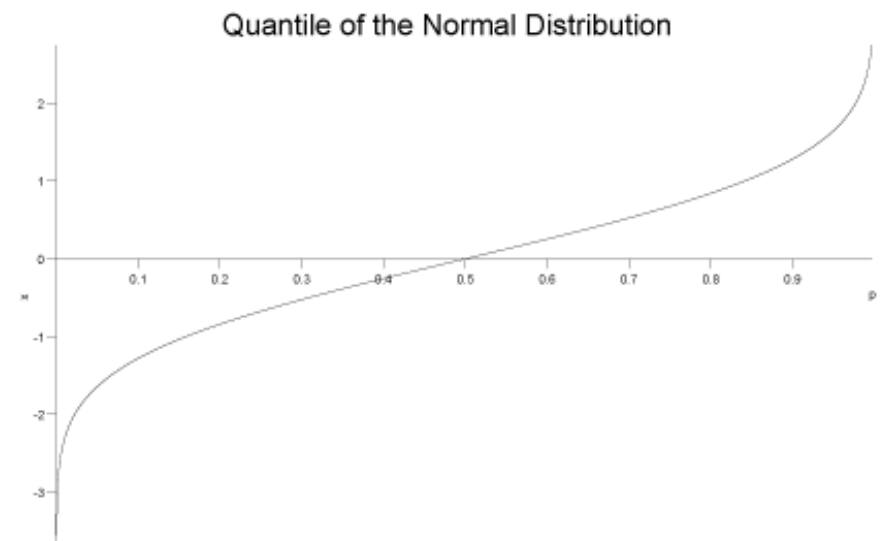
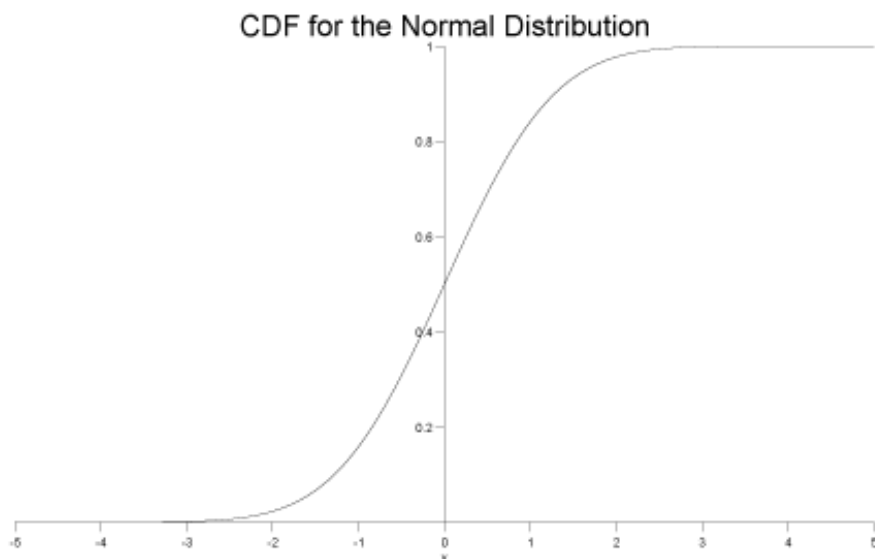
inverse CDF, quantile function

$$F_Y^{-1}(q) = \text{smallest } y \text{ such that } F_Y(y) > q$$

$$F_Y^{-1}(0.5) = \text{"the median"}$$

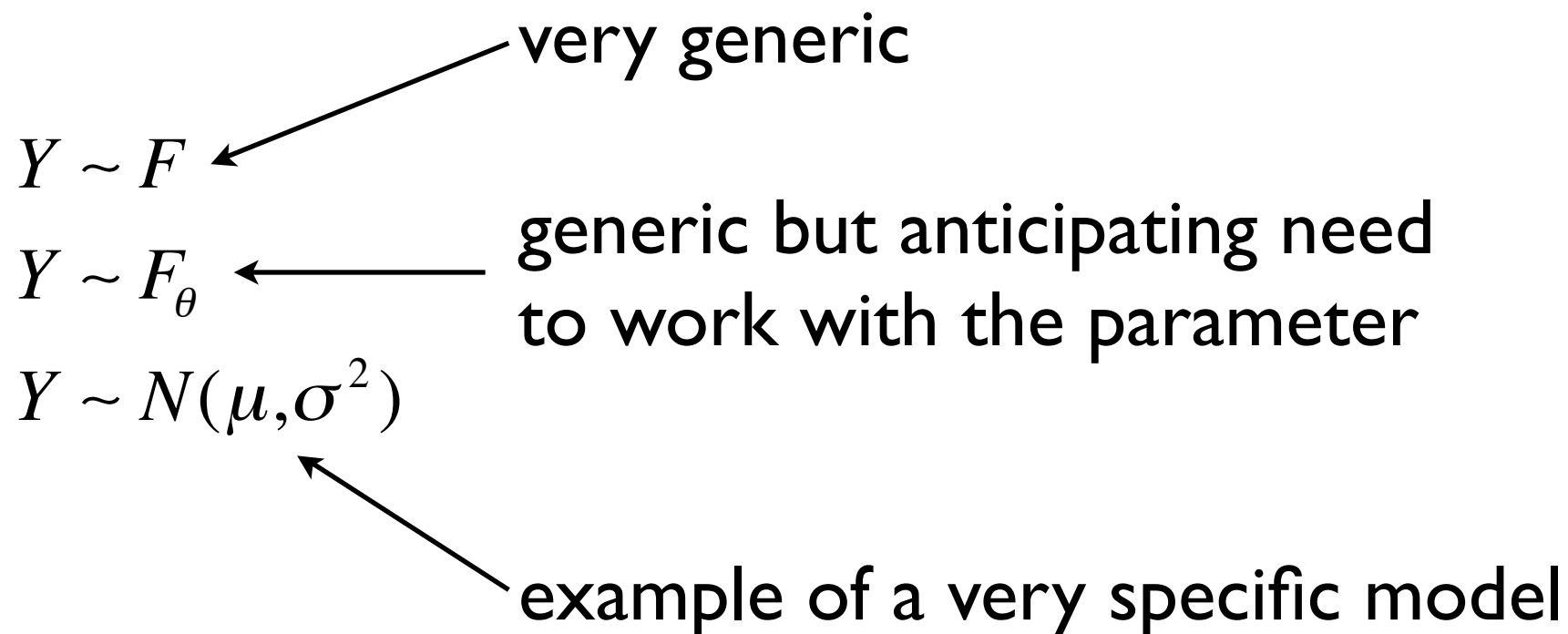
$$F_Y^{-1}(0.25) = \text{"the first quartile"}$$

$$F_Y^{-1}(q) = \text{value that traps probability } q \text{ to the left and } 1 - q \text{ to the right}$$



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

statistical 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

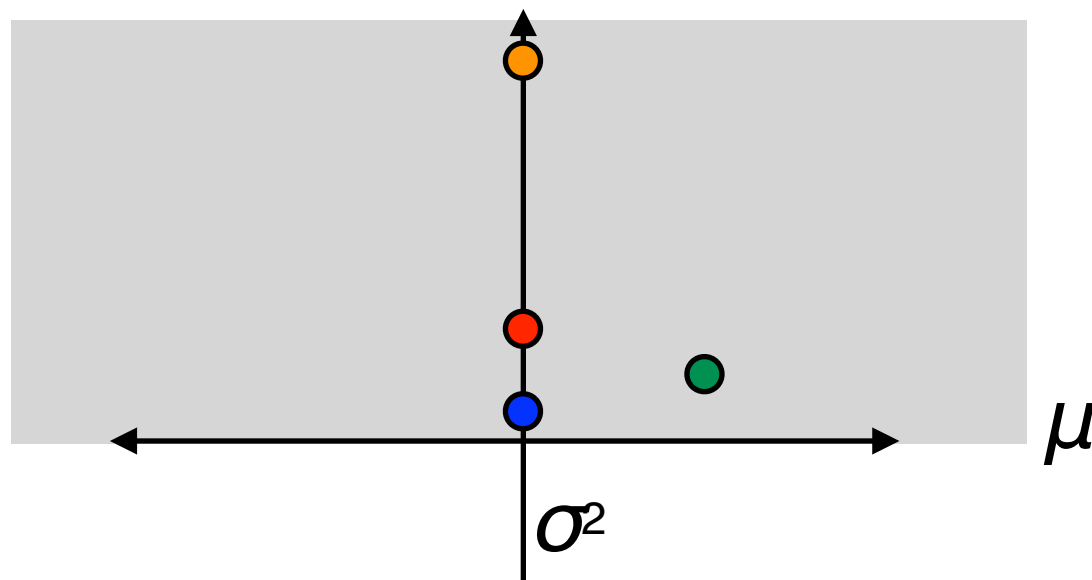
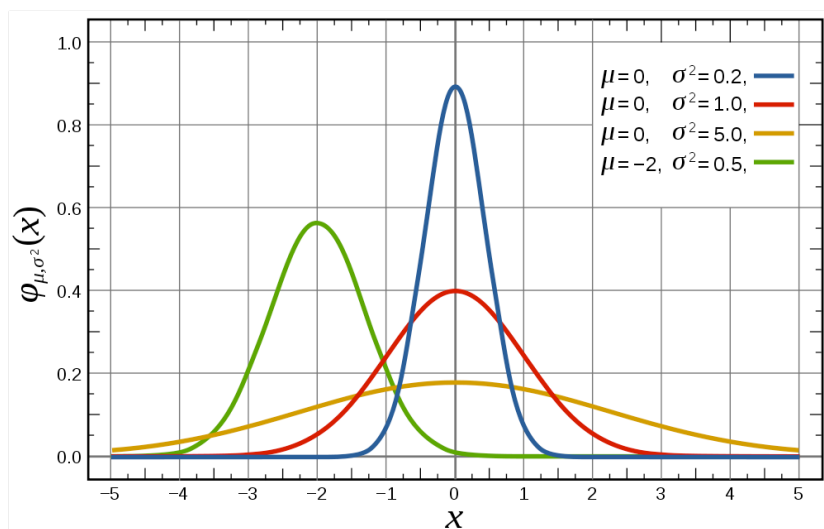
when we assume data is normally distributed about its mean ... we’re doing parametric inference; the parameter space is a nice friendly half-plane in \mathbb{R}^2

world's favorite parametric model

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

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

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



Parametric models we've reviewed

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

“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

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

or

" $Y_i \sim F$ "

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

it is a parameter

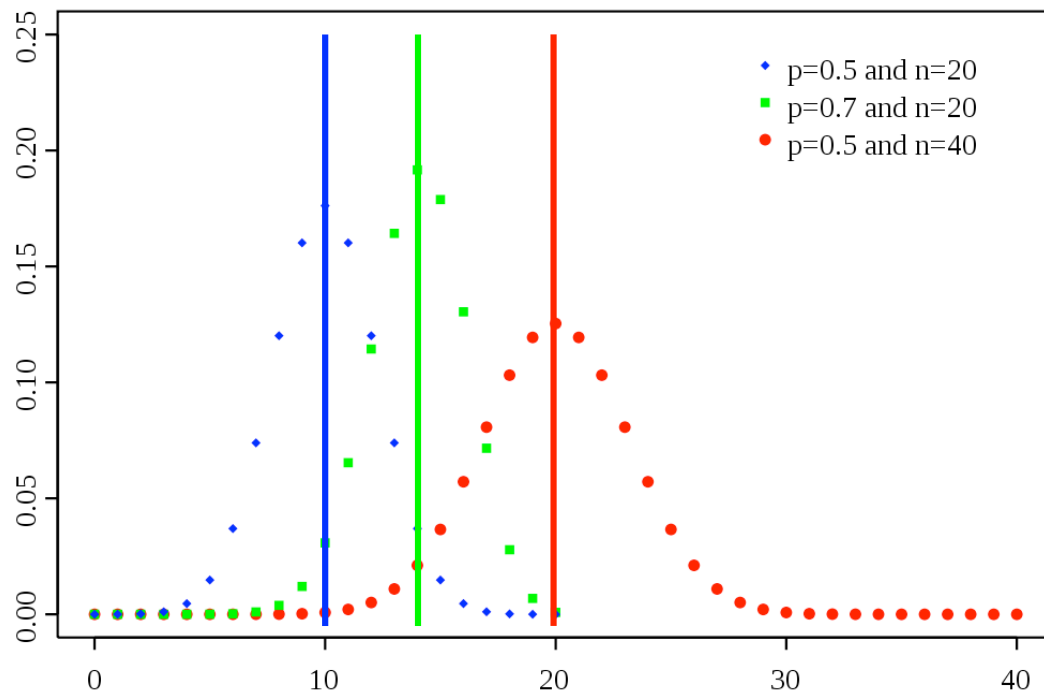
often denoted $E(Y)$ or μ or μ_Y

common sense “definition”: a long-run average
 $E(Y)$ approx equal to (sum of Y_i 's)/ n
the bigger n is, the better the “approximation”

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:

$$Y \sim \text{Binom}(n, p)$$

$$E(Y) = np$$

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)

and now something related *but different*

the average, the sample mean

it is a random variable!!!! not a parameter!!!!

often denoted \bar{Y} or \bar{Y}_n or $\hat{\mu}$ or $\hat{\mu}_Y$

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

main usage:

as a point estimator of the true mean or as a test statistic -- or part of a test statistic -- for hypothesis tests re: the mean

$$\bar{Y} \text{ or } \bar{Y}_n \text{ or } \hat{\mu} \text{ or } \hat{\mu}_Y$$

notational sidebar:

statisticians LOVE to put hats on Greek letters

a constant visual reminder of what's random (the thing with the hat) and which parameter it is an estimator for (the Greek letter without the hat)

sometimes we put the sample size n in the subscript to reinforce that something is random and that its distribution depends on the sample size

the expected value of the sample mean is the true mean:

$$E(\bar{Y}_n) = \mu$$

“the sample mean is unbiased”

the variance of the sample mean is:

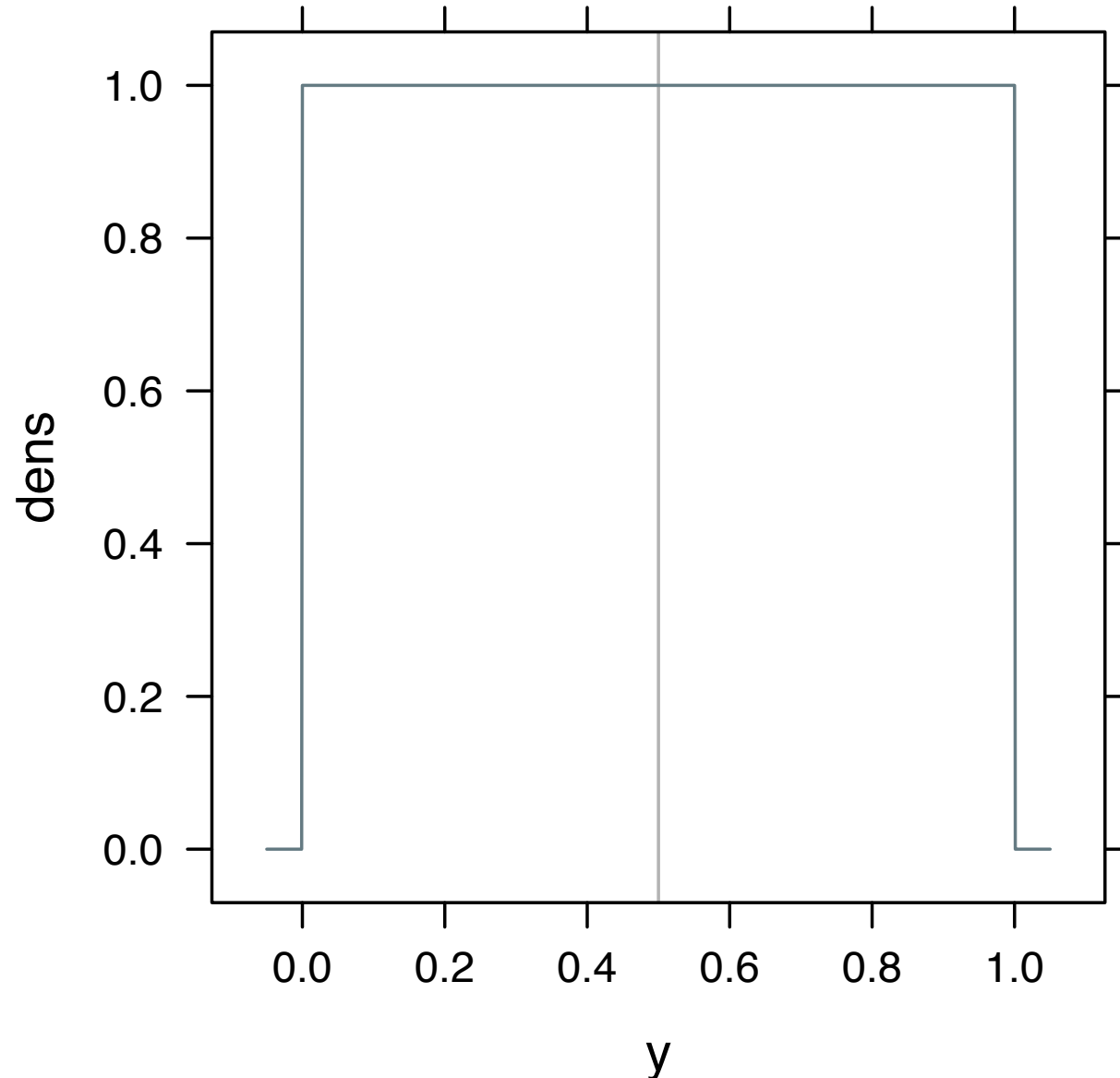
$$V(\bar{Y}_n) = \frac{\sigma^2}{n}$$

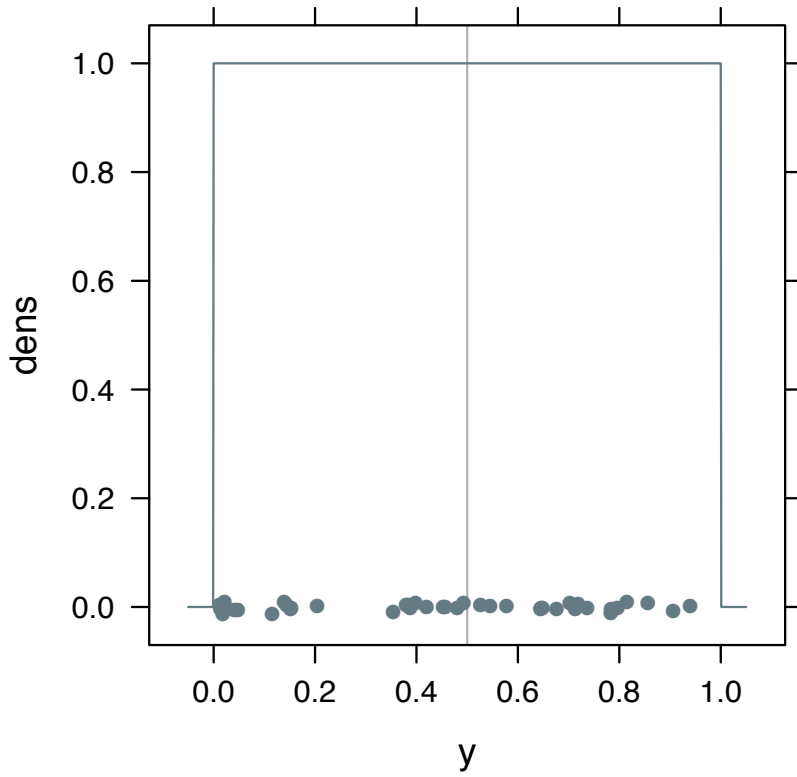
i.e., the variance of the sample mean is fundamentally determined by the underlying variance of the data --it is also affected by the sample size

the average, the sample mean: an empirical case

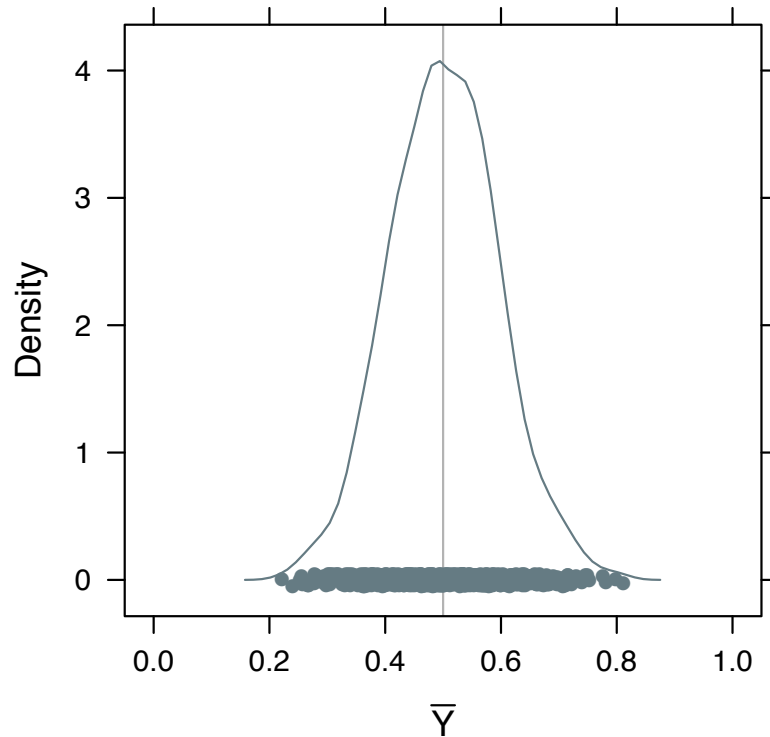
consider $Y \sim \text{Unif}(0, 1)$

$$E(Y) = 0.5$$





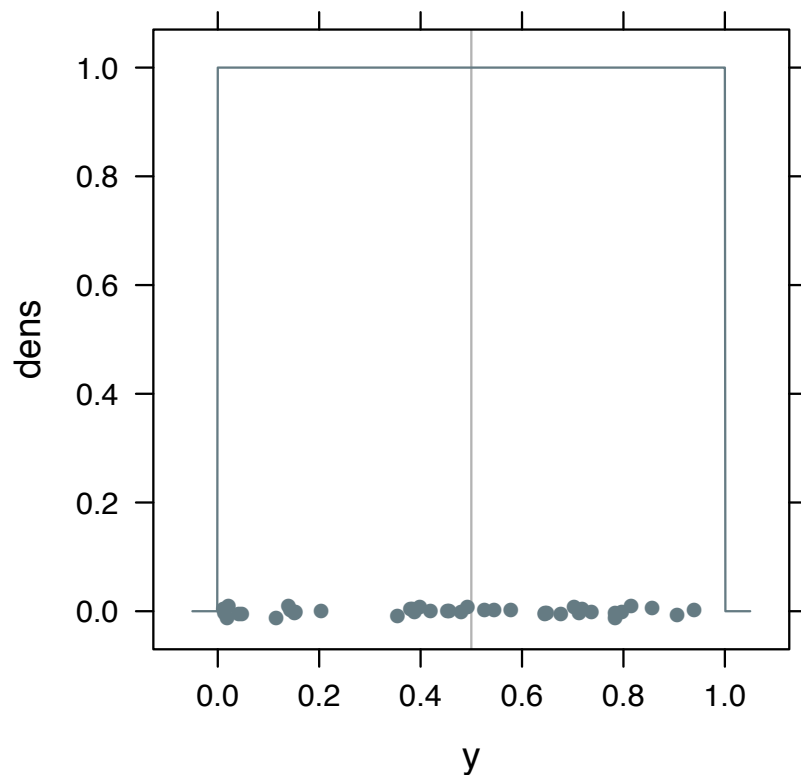
take a sample of size n



take the average

... now do that lots of times ...

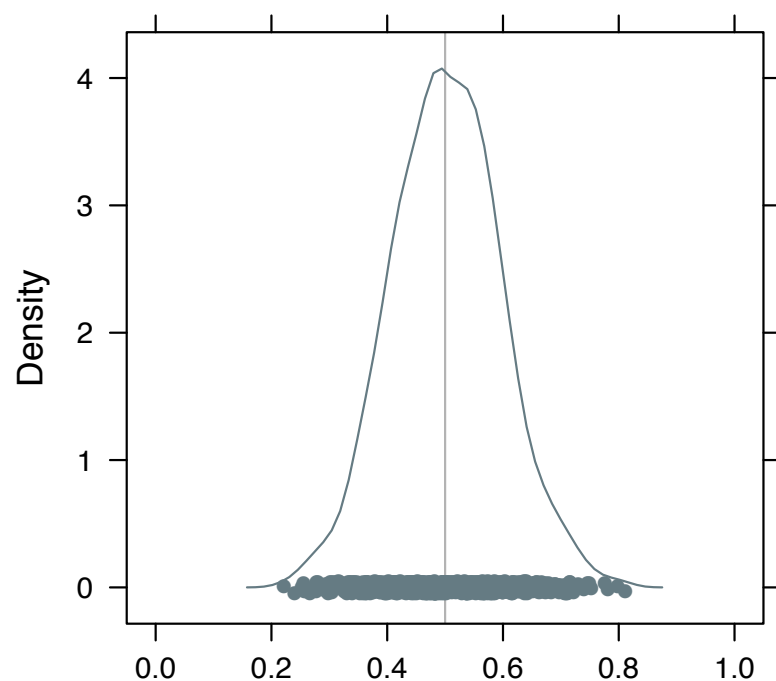
what does that distribution look like?



visual confirmation of

$$E(\bar{X}_n) = \mu$$

= 0.5 in this case

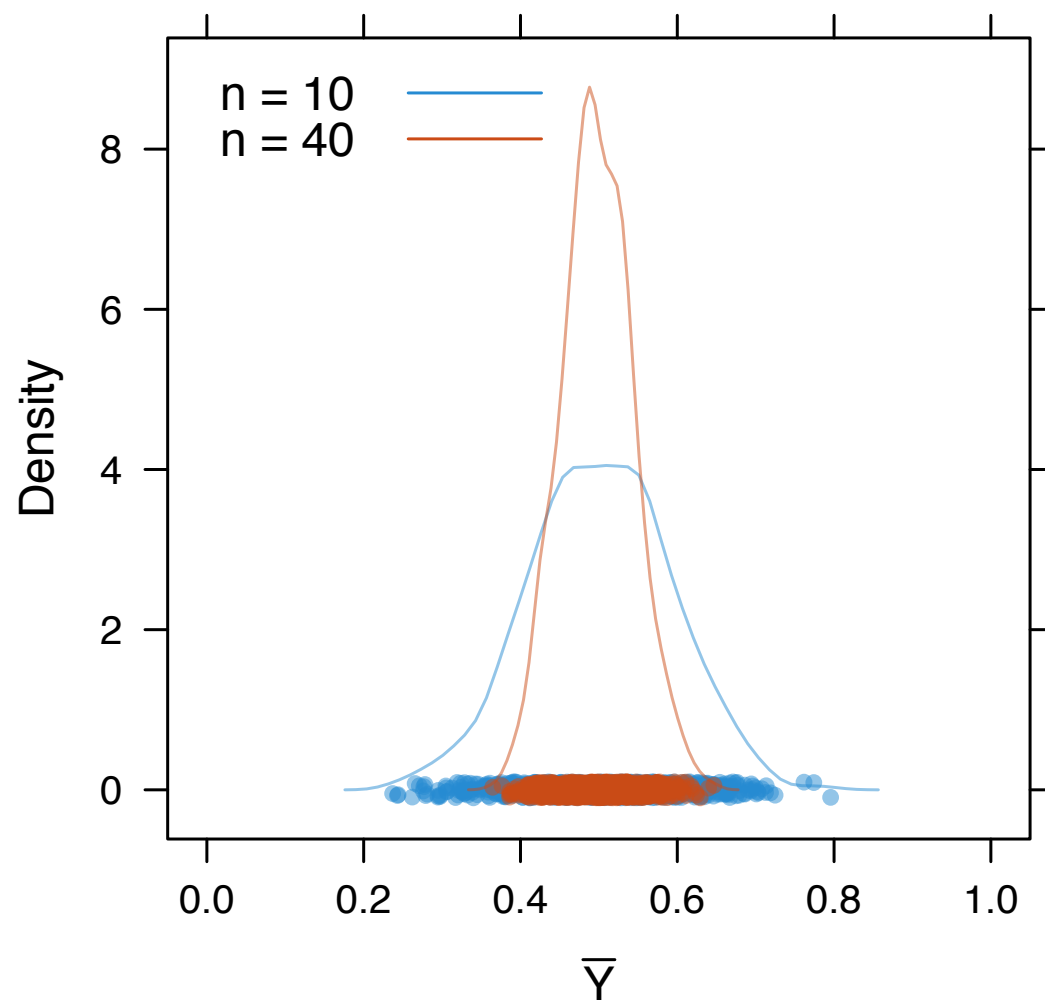


notice that the distribution of sample means doesn't look uniform Central Limit Theorem coming soon!

```
> n <- 10  
> numSamp <- 1000  
> xBar <- rowMeans(matrix(runif(n * numSamp), nrow = numSamp))  
> n2 <- 40  
> xBar2 <- rowMeans(matrix(runif(n2 * numSamp), nrow = numSamp))  
> densityplot(~ xBar + xBar2, ...)
```

visual confirmation of

$$V(\bar{X}_n) = \frac{\sigma^2}{n}$$



the variance of the sample mean is

$$V(\bar{X}_n) = \frac{\sigma^2}{n}$$

i.e., the variance of the sample mean is fundamentally determined by the underlying variance of the data --it is also affected by the sample size

this is why it is nonsensical to ask if a sample size of $n = 3$ (or 20 or whatever) is “enough” to perform statistical inference, in the absence of some info on σ^2 (and specific discovery goals)

discouraging that the variance of the sample mean involves σ^2 which we generally don't know

what if you want to know more about Y ?

the mean gives a sense of what's “typical” or where the center of observed values of Y will lie, but what sort of spread will those observed values have?

the sample mean is obviously a good guess at $E(Y) = \mu$, but how good is it?

variance

standard deviation = $\sqrt{\text{variance}}$

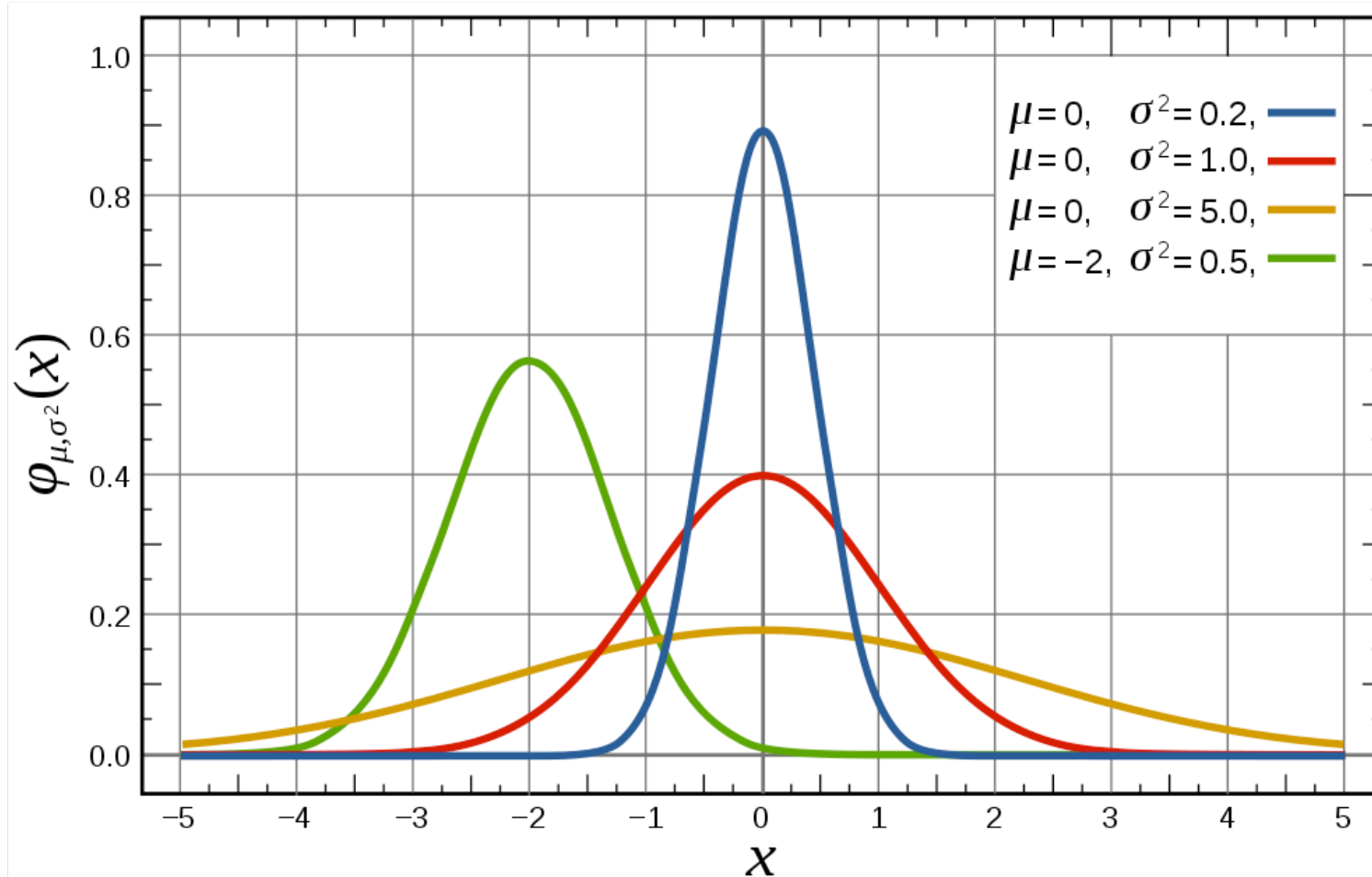
it is a parameter

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



normal as example; bigger $\sigma^2 \leftrightarrow$ bigger “spread”

and now something related *but different*

the sample variance

it is a random variable!!!! not a parameter!!!!

often denoted s^2 or $\hat{\sigma}^2$ *

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \qquad E(s^2) = \sigma^2$$

main usage:

I'm using the sample mean to infer something about the true mean and, to my horror, the quality of that guess depends on the variance. So I'm forced to worry about the variance. (“nuisance parameter”)

* n vs. $n - 1$ sidebar

a “statistic” is a rv that’s a function of the data

classic examples:

- the sample mean
- the sample variance

two main reasons we love them:

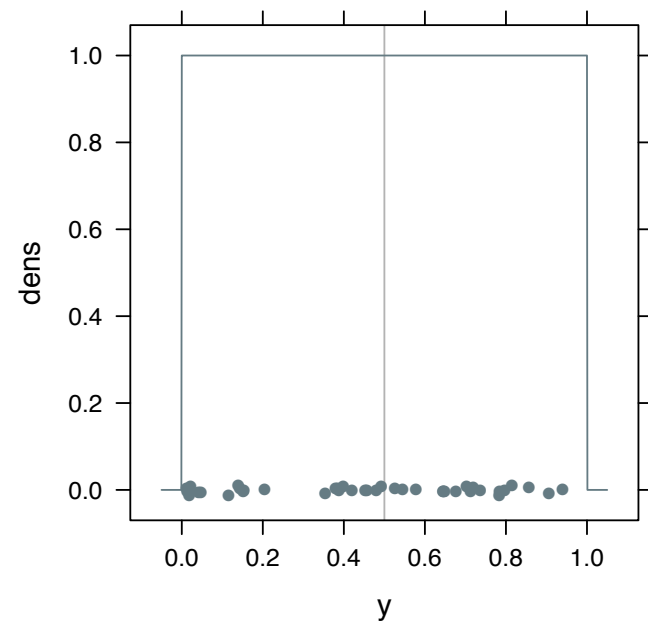
[1] sometimes they are estimators for parameters we care about

[2] sometimes they are the basis for a hypothesis test

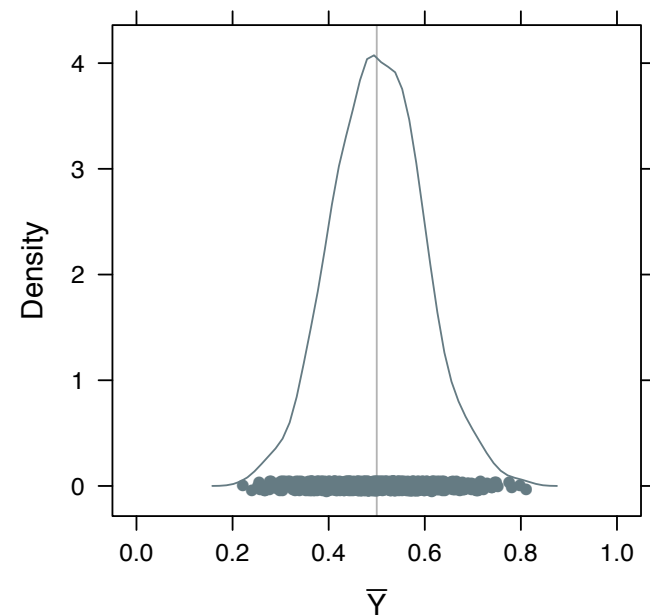
the distribution of a statistic is called its sampling distribution. it’s related to the distribution of the data but it is not the same

visual confirmation that distribution of a statistic -- such as sample mean here -- is NOT the same as the distribution of the underlying data

dist'n of data
 $\text{Unif}(0,1)$



dist'n of the sample
average
... NOT $\text{Unif}(0,1)$



... looks kind of like the normal dist'n, no?

we generally know more about a statistic's sampling distribution as n gets large

“large sample theory”, “limit theory”, “asymptotic theory”

the law of large numbers

common sense “statement”:

the average of a large, iid sample will be close to the true mean

the law of large numbers (formally)

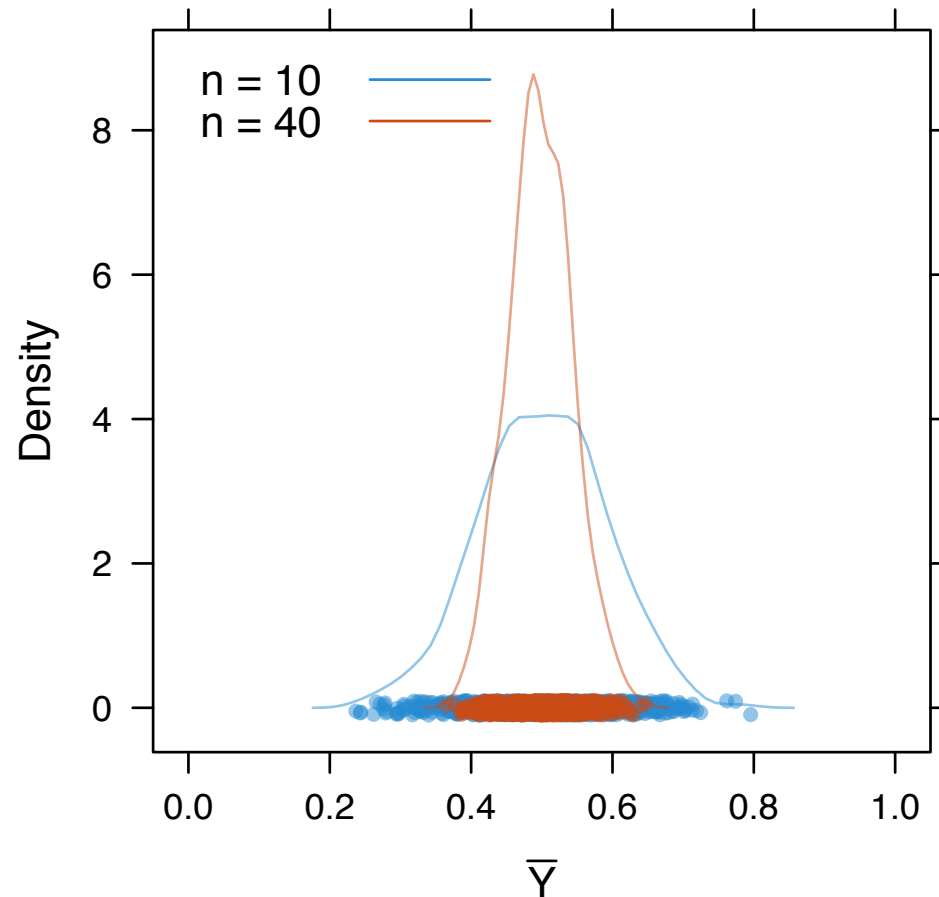
Let X_1, X_2, \dots be an IID sample, let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1)$. Recall that the sample mean is defined as $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and that $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathbb{V}(\bar{X}_n) = \sigma^2/n$.

5.6 Theorem (The Weak Law of Large Numbers (WLLN)).³

If X_1, \dots, X_n are IID, then $\bar{X}_n \xrightarrow{P} \mu$.

Interpretation of the WLLN: The distribution of \bar{X}_n becomes more concentrated around μ as n gets large.

Imagine this trend continuing as n gets bigger and bigger the sample mean sampling dist'n gets more and more concentrated around $\mu = 0.5$



the central limit theorem

common sense “statement”:

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

the central limit theorem (formally)

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

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

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

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

rookie misconception re: law of large numbers:

“If I can just make my sample big enough, I won’t have to worry about error.”

there is no sample that is “big enough” in an unqualified sense

in stats, there are precious few fundamental constants, like there are in math (think: π and e) or physics (think: speed of light)

context and goals always matter

rookie misconception re: central limit theorem:

“I can average any large-ish bunch of numbers and divide by the sd and call it a z-score. Then I can compare it to a $N(0,1)$ to determine statistical significance. I’ve got a hit if the number’s greater than 1.96!”

the CLT assumes you’re averaging observations that are iid!

averaging gene expression for 1 gene across exchangeable subjects ... yeah, CLT applies

averaging gene expression for 1 subject across genes ... no, CLT does not apply

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?

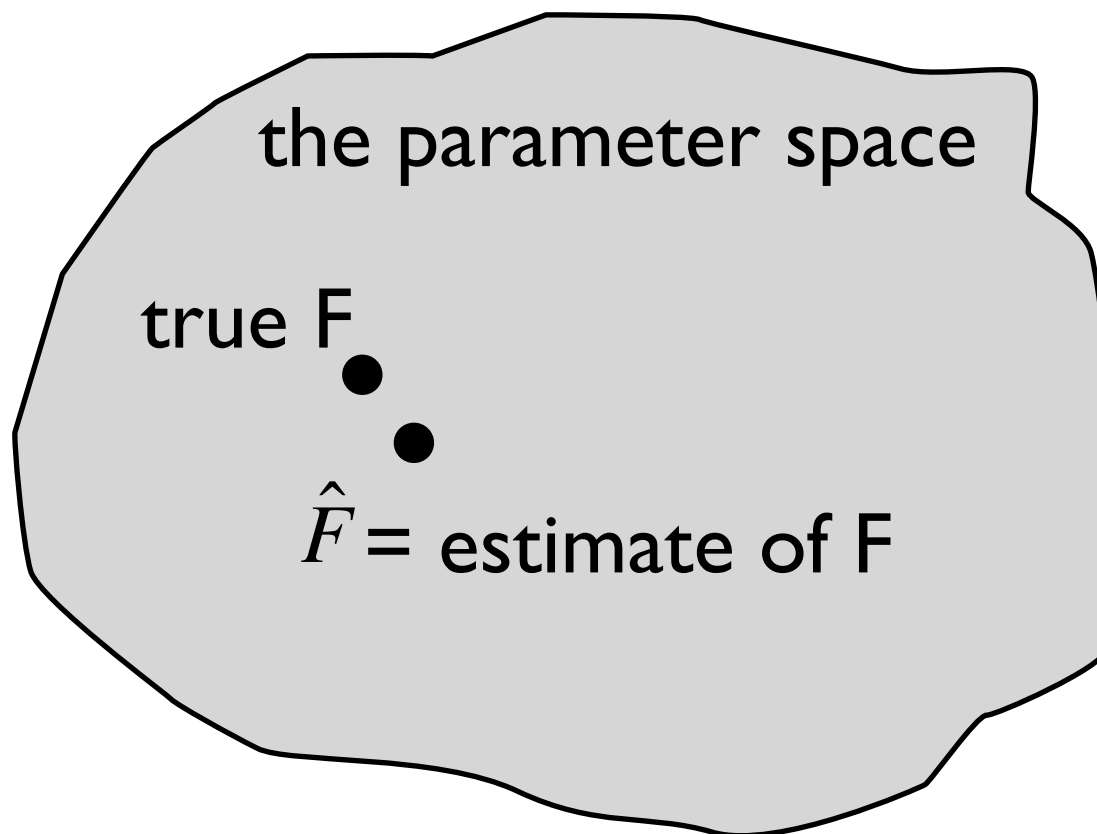
estimation: what's your best guess at the true value of the parameter?

estimation in generic statistical model

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

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

Estimate F with \hat{F} .



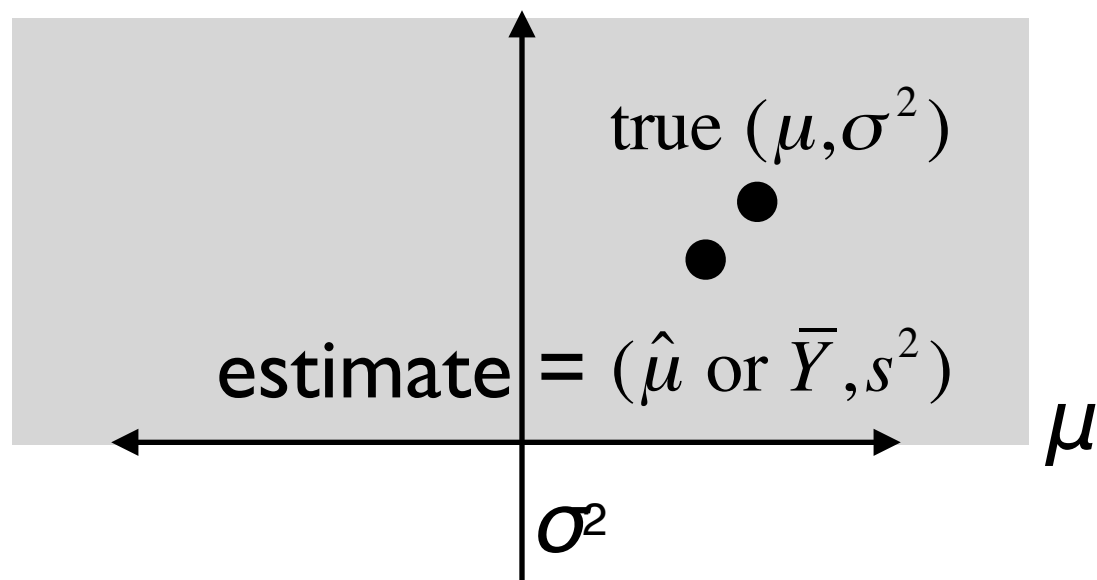
estimation in very specific statistical model

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

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

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

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

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

testing

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

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

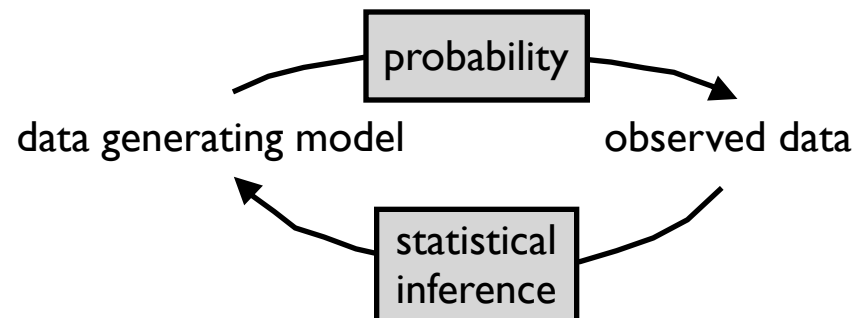
does $E_F(Y) = \mu_Y = \mu_Z = E_G(Z)$?

Call this statement the null hypothesis H_0 :

$$H_0 : \mu_Y = \mu_Z$$

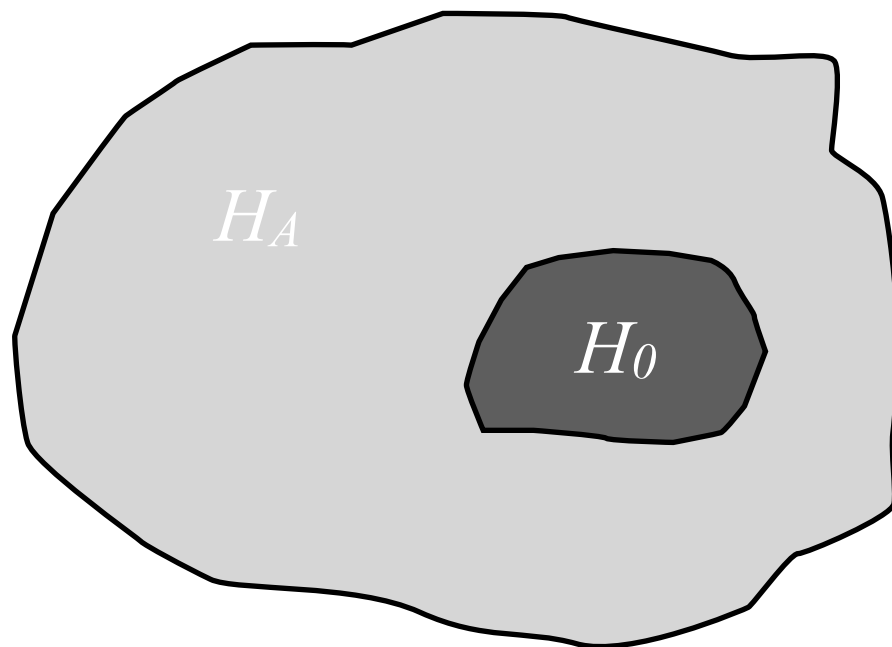
Or, equivalently:

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



statistical model

the parameter space



In formal hypothesis testing:

Define a “null (boring) region” for the parameter -- the dark gray area.

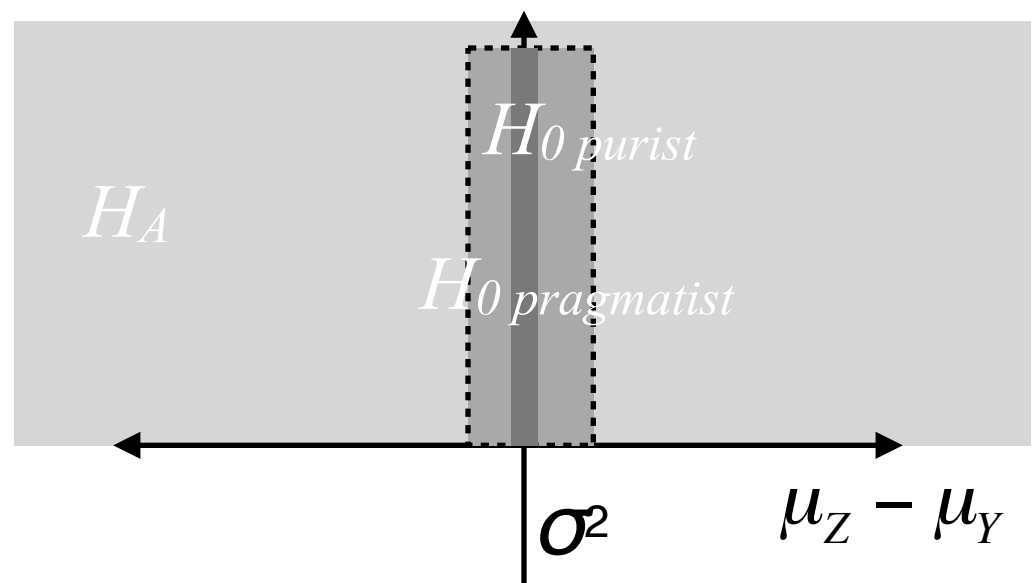
Ask whether the true value lies in that region or outside, in the “alternative (interesting) region” -- the light gray area.

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

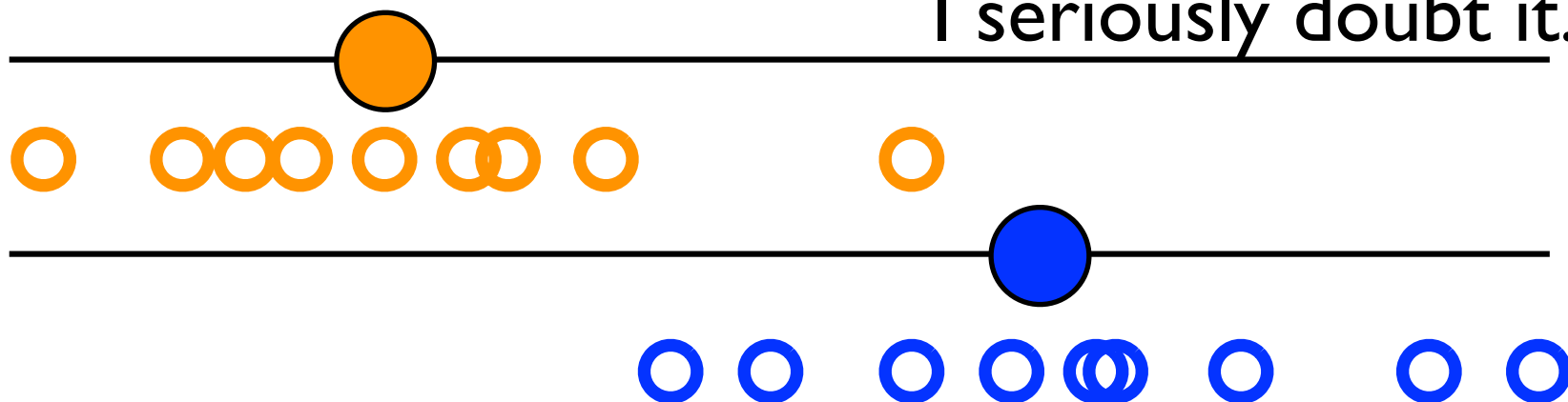
“pragmatist” usually defines the null region like the “purist”, because the math is so much more tractable and then accounts for concerns of “realist” when interpreting results (or, e.g., does a post hoc filter on observed difference in sample means)



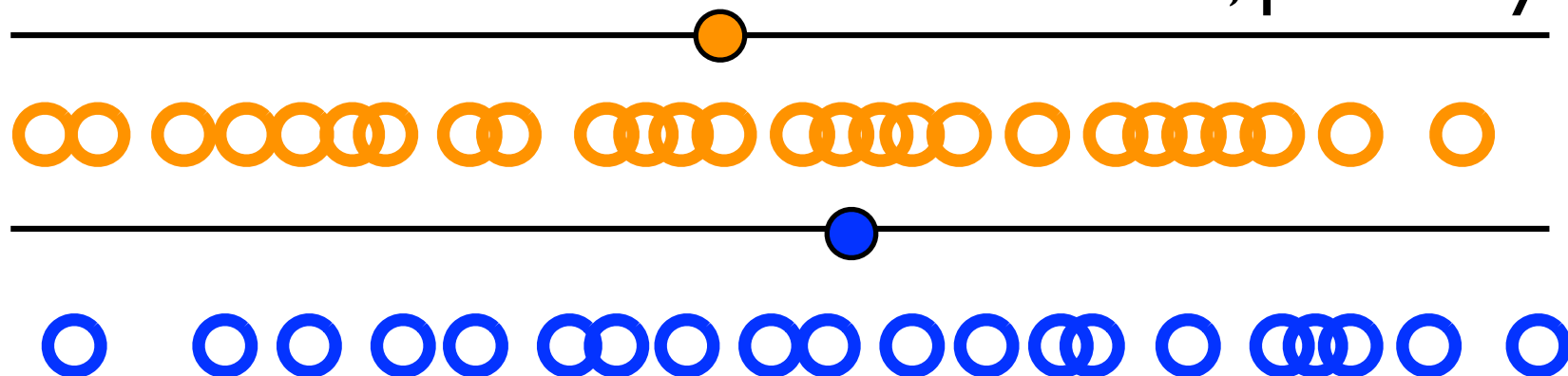
how do we use the
observed data to say
something specific and
quantitative about H_0 and
 H_A , i.e. which one appears
to be true?

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

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 reasons we love them:

[1] sometimes they are **estimators** for parameters we care about

[2] sometimes they are **test statistics**, i.e. the basis for a hypothesis test

examples of *statistics*: various functions of the observed data you will likely compute often in your stats life

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

the sample mean
most famous for being a great estimator for the true mean

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$$

the sample variance
most famous for being a great estimator for the true variance

$$t = \frac{\bar{Y}_n}{s / \sqrt{n}}$$

the one-sample t statistic
most famous for testing
 $H_0: \mu_Y = 0$

general idea for a test statistic:

when “big” or “extreme”, suggests that the observed data is very unexpected under the null hypothesis H_0

a p-value quantifies this incompatibility between the data and H_0 -- specifically, it's a tail probability

so to get a p-value, you must know or approximate the probability distribution of the test statistic under the null H_0

general idea for an estimator:

point estimate is your single best guess at the parameter

interval estimate, i.e. confidence interval, provides a set of possible values that are “compatible” with the data

construction of an interval estimate requires knowledge of the estimator’s distribution

therefore ...

to complete a hypothesis test or convey the precision of a point estimate, we need the statistic's or estimator's sampling distribution

“sampling” here should invoke “long-run”,
“hypothetical repeats of the experiment”

for an estimator, the standard deviation of the sampling distribution is called the standard error

for the sample mean, recall the standard error or “standard error of the mean” (“SEM”):

$$sd(\bar{Y}_n) = \frac{\sigma}{\sqrt{n}} \text{ often estimated by } \hat{sd}(\bar{Y}_n) = \frac{s}{\sqrt{n}}$$

but realize that this is just a specific standard error -- it's a more general concept and they won't always have this exact form

p-value...

is the probability under the null H_0 of observing a test statistic value as or more extreme than that computed from the observed data

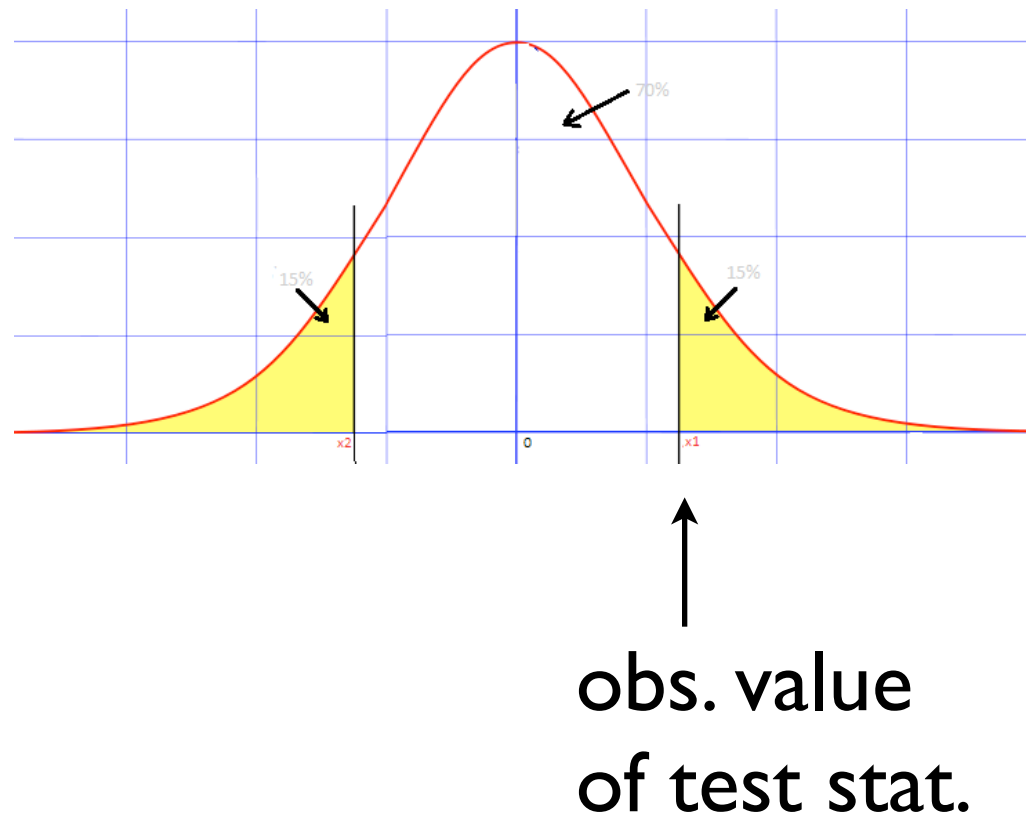
example in a two-sided test, i.e. when both very small and very large values of test stat are “extreme”:

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

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

imagine this is a
sampling distribution

sum of the yellow areas
= p-value



musings on p-values

in some sense, it's supreme laziness (cleverness?) to work this way: easy on the analyst only need to characterize dist'n of the test stat under the null

downside: an indirect, nonspecific 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

another way in which we “work backwards”

Bayesian critique

p-values will be eventually be thresholded --
whether we admit it or not -- to produce simple
calls:

p-value exceeds threshold	... 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

“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

false positive rate = P(false positive) for one truly null gene
alpha level used to threshold p-values

“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

family-wise error rate (FWER) = $P(\geq 1 \text{ false positive})$
controlled by Bonferroni correction

“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	true positives	# alts
		discoveries	# genes

“power” is probability of a true positive (requires more info to fully specify)

“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	true positives	# alts
		discoveries	# genes

false discovery rate = $E(\# \text{ false pos} / \# \text{ discoveries})$

end of general review and introduction
for statistical

ready to begin coverage of
comparing two groups