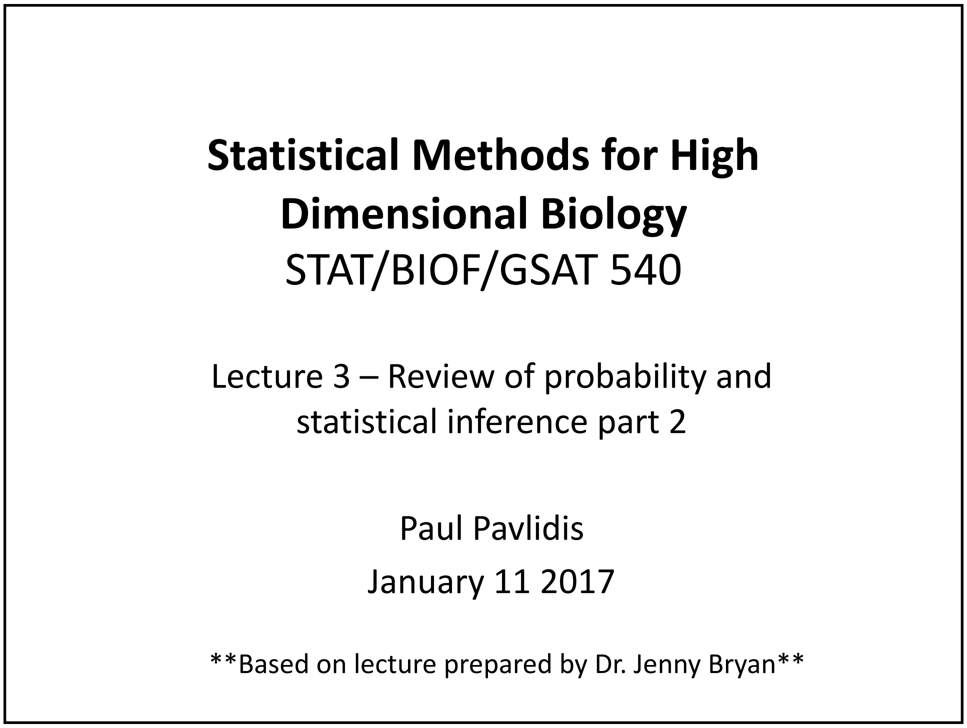


Statistical Methods for High Dimensional Biology

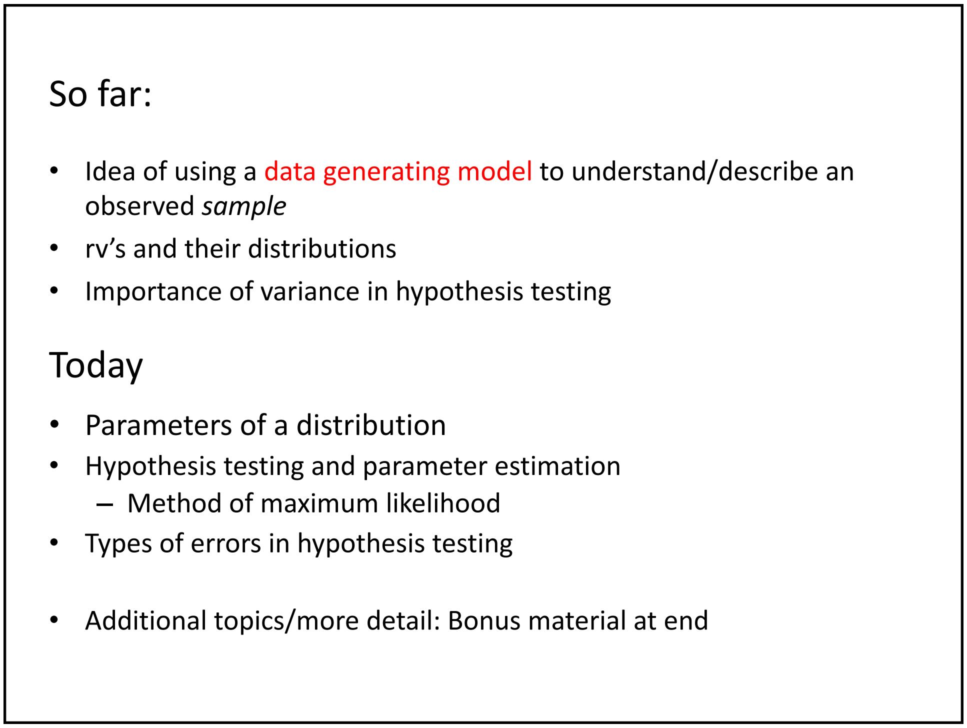
STAT/BIOF/GSAT 540

Lecture 3 – Review of probability and statistical inference part 2

Paul Pavlidis

January 11 2017

Based on lecture prepared by Dr. Jenny Bryan



So far:

- Idea of using a data generating model to understand/describe an observed *sample*
- rv's and their distributions
- Importance of variance in hypothesis testing

Today

- Parameters of a distribution
- Hypothesis testing and parameter estimation
  - Method of maximum likelihood
- Types of errors in hypothesis testing

- Additional topics/more detail: Bonus material at end

Random variables can be characterized by a distribution

Following previous example…

X : number of heads in n tosses

$X \sim Bin(n,p)$ ← n and p are parameters (Θ)

Variable

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

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

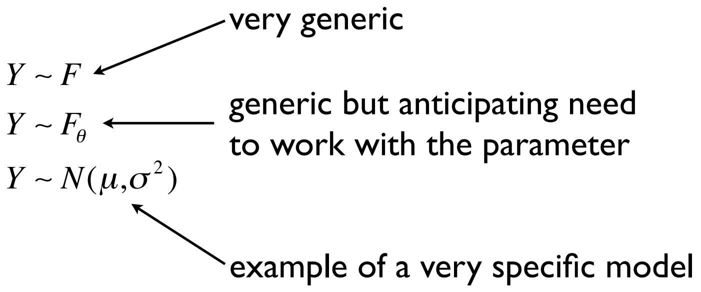
$\sum_{x=0}^{a} \begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x}$ ← cumulative distribution

Parameters determine distributions

- When sampling from a population described by a pmf/pdf f(X|θ), then knowledge of θ yields knowledge of the entire population.
  - This description is the "statistical model"
- This is why parameter estimation is useful:
  - If we are tossing a coin, we would like to estimate the parameter p

Statistical models

First some notation…



$Y \sim F$ — very generic

$Y \sim F_\theta$ — 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)

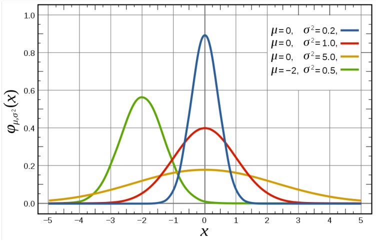
Statistical model

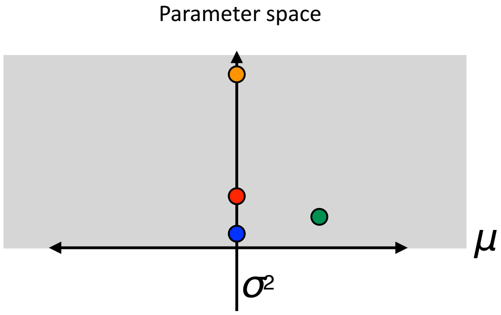
- The *parameter space* is the set of all possible values for the parameter
  - To say a model is “parametric” means the parameter space is a nice friendly Euclidean space
  - When we assume data is normally distributed about it’s mean … we’re doing parametric inference; the parameter space is a nice friendly half-plane in R^2
- A goal is to “*guess”* the parameter values: “fit the model to the data”
- The model is a representation that (we hope) approximates the data and (more importantly) the population that the data were sampled from.
- We can then use this model:
  - For hypothesis testing and other forms of inference
  - For prediction
  - For simulation

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 = 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~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 ~ N(μ, σ^2)	θ = (μ, σ^2)
Student's t	Y ~ t_{df}	θ = df

Parametric models we've reviewed ....

“Nonparametric”

- “Semi-parametric” and “nonparametric” imply the parameter space isn’t a simple Euclidean space
- The parameter space is more exotic, e.g. at least partially a function space, an infinite dimensional space
- You don’t have to feel comfortable with, say, function spaces, to *apply* nonparametric statistical methods responsibly
  – (e.g. rank-based procedures like the Wilcoxon test)

Working with distributions

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

or

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

The two properties of any distribution *F* you’re mostly like to care about (meaning: wanting to make inferences about):

1. Its expected value (aka expectation or mean)
2. Its variance

Expectation, expected value, the mean

- Often denoted *E(Y)* or μ or μ_Y
- A general property of random variables*
- 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”
- It can be a ***<u>parameter</u>*** we want to estimate

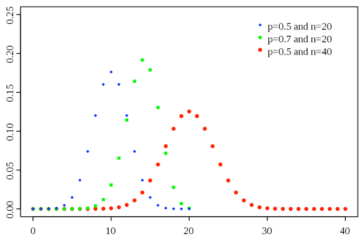
*There are distributions for which it can’t be computed, but “all” is close enough for us.

Definition of E(Y)

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

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

- Intuition: it’s a sum of the values that can be taken by Y weighted by how likely each value is ($p_Y(y)$ or $f_Y(y)$)
- A measure of “location” of the distribution
- Often is one of the parameters of the distribution (e.g. normal) or is easily computed from them (e.g. binomial)



binomial example:

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

$$E(Y) = np$$

Variance of a rv

- 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? A useful measure is the **variance**
- Like E(Y), is is a general property of rvs
- Usually denoted *V(X)* or σ^2 (variance) and σ (sd)

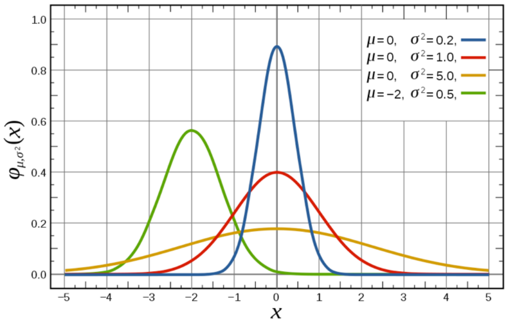
$$V(Y) = E[(Y\ - E(Y))^2]$$

- Long-run average of the squared differences between obs vals $Y = y$ and the true mean μ
- Standard deviation = √variance
- It can be a parameter (e.g. of the normal distribution)

Variance as a parameter of a probability distribution

Normal as example; bigger σ^2 ↔ bigger "spread"

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Wikipedia

Stopped here

Deep breath

- Random variables have probability distributions
- We can summarize some facts about a rv with the mean (location) and variance (spread)
- However, we usually aren't given these values. We have to estimate them from data
- The data are a sample from Y of size *n*
  - We compute a sample mean and variance (and other things, but let's stick with those) – this is *inference*
- Assuming *n* is "small", we have to be very concerned with how well our estimates match the "true" values of the mean (et al.) before we try to say anything important about Y.

The sample mean is a random variable

- Often denoted $\overline{Y}$ or $\overline{Y}_n$ or $\hat{\mu}$ or $\hat{\mu}_Y$
- Defined: $$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$ Where n is the sample size
- 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

Notational sidebar: statisticians LOVE to put hats on Greek letters as a 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 sample mean has an expected value

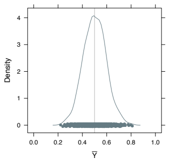
- Because it's a random variable …
- The expected value of the sample mean is the true mean: $$E(\overline{Y}_n) = \mu$$ Note! This is not the mean of Y
- The variance of the sample mean is: $$V(\overline{Y}_n) = \frac{\sigma^2}{n}$$ Note! This is not the variance of Y

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 *n*. More data: less uncertainty in sample mean

Variance of the sample mean

The variance of the sample mean is:



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

Where σ^2 is the variance of X (which we will detail in a bit) and *n* is how many samples from X we used to compute the mean.

- 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 square root of this is the **standard error** of the sample mean.

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)

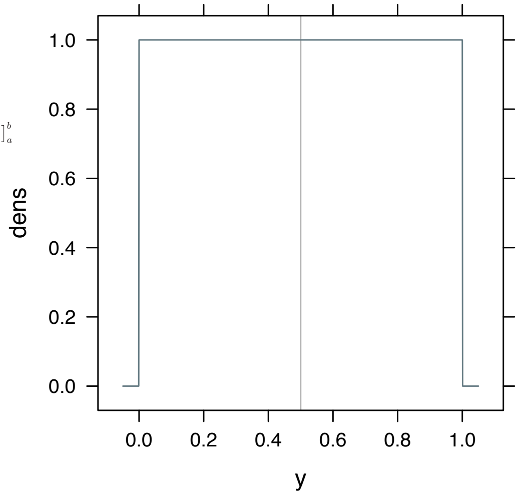
- Note that the variance of the sample mean involves σ^2 which we generally don't know … We have to estimate it

the average, the sample mean: an empirical case

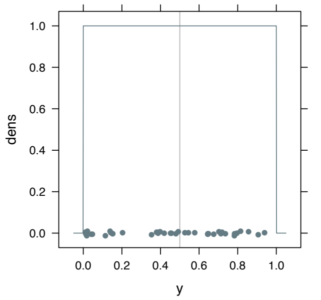
consider Y ~ Unif(0,1)
E(Y) = 0.5

$$E(X) = \int x f_x(x)\,dx$$

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



Sampling the sample mean

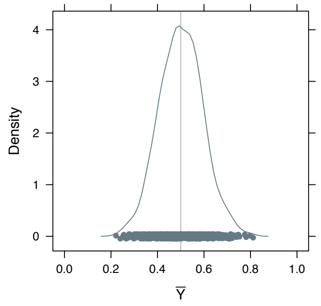


take a sample of size n and compute the mean

e.g. (0.3365 , 0.1733 , 0.0861, 0.3933 , 0.8044 , 0.0111, 0.2331, 0.9339, 0.2268, 0.7859)

$$\bar{Y} = 0.3984$$

… now do that lots of times …



The distribution of values is the **sampling distribution of the sample mean** This isn't as exotic as it sounds.

Visually confirms that E(Y) = 0.5

Note that this distribution is not uniform.

In fact the Central Limit Theorem tells us that this distribution tends towards **normality** (so long as sample is IID)

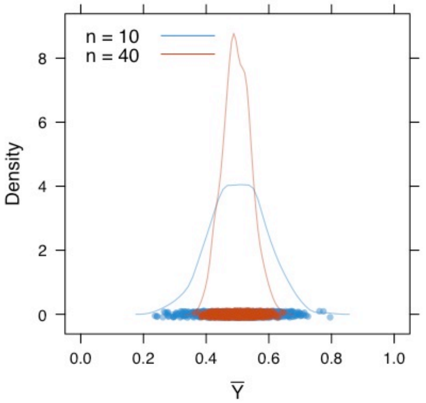
Importantly, in real life we only get one sample mean, but we can use the fact that it comes from this distribution.

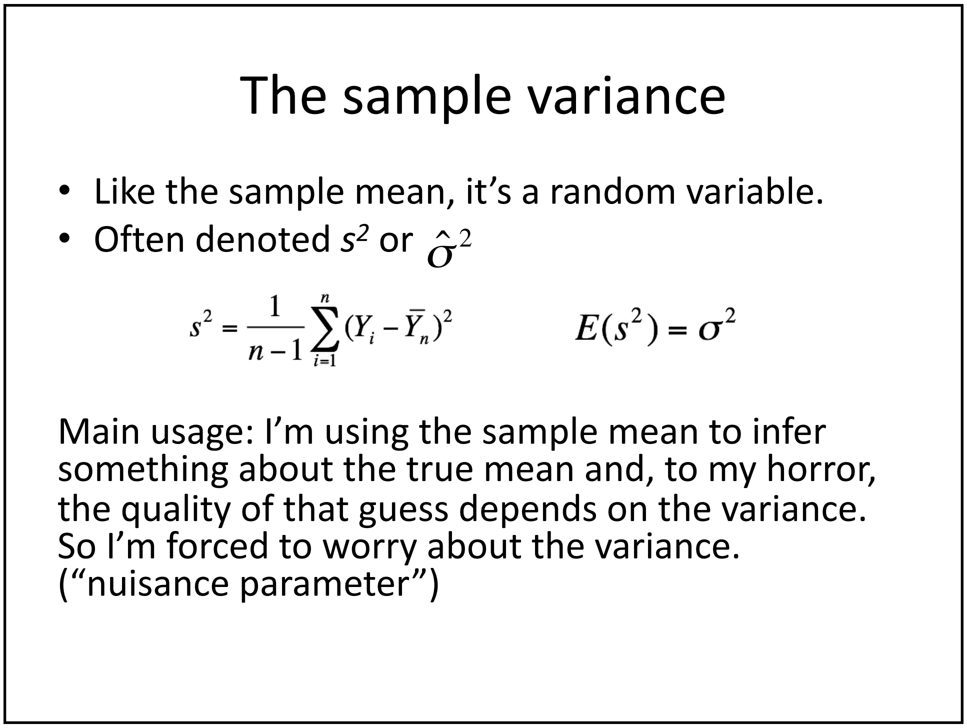
Effect of sample size on $V(\bar{X}_n)$

```
> 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

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



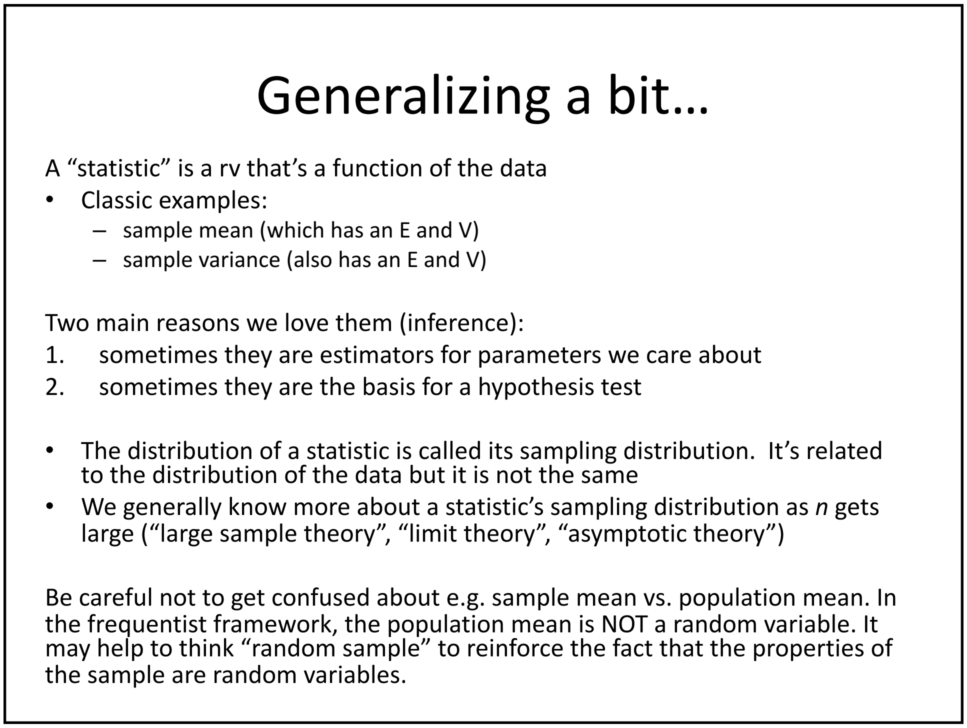


The sample variance

- Like the sample mean, it's a random variable.
- Often denoted *s²* 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")



Generalizing a bit…

A "statistic" is a rv that's a function of the data

- Classic examples:
  - sample mean (which has an E and V)
  - sample variance (also has an E and V)

Two main reasons we love them (inference):

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
- We generally know more about a statistic's sampling distribution as *n* gets large ("large sample theory", "limit theory", "asymptotic theory")

Be careful not to get confused about e.g. sample mean vs. population mean. In the frequentist framework, the population mean is NOT a random variable. It may help to think "random sample" to reinforce the fact that the properties of the sample are random variables.

Summarizing

- We have data (a sample) that we want to use to make **inferences** about the “process that produced that data” (loosely: the “population”)
- We use statistics like the sample mean as **estimators** of the population distribution properties.
- Theory (CLT et al.) tells us a lot about the properties of these estimators (as long as certain assumptions like IID hold).
- These let us make statements about how accurate we think our estimators are … and much more.

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?

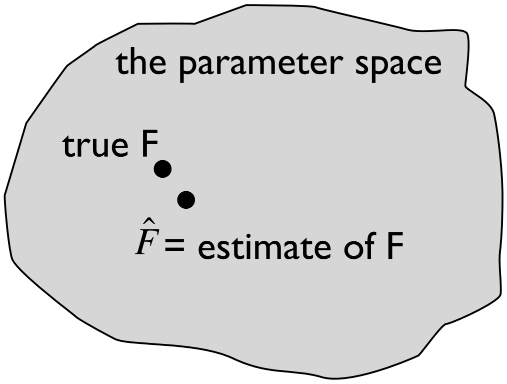
estimation: what's your best guess at the true value of the parameter?

estimation in generic statistical model

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

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

Estimate F with $\hat{F}$.



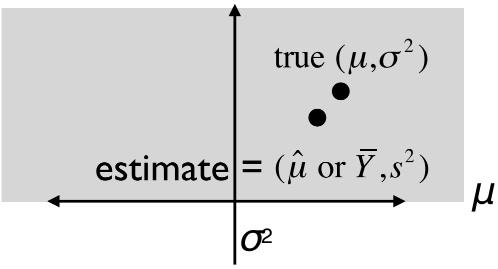
estimation in very specific statistical model

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

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” 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)

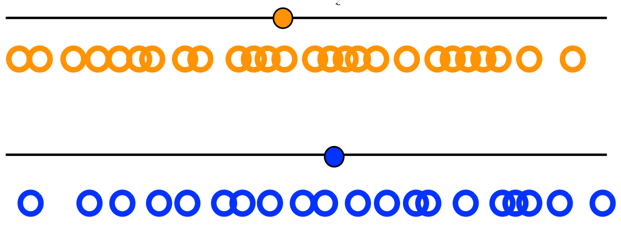
Must begin with a clear understanding of how to do this for one gene and two conditions

Then extend to more genes, more conditions

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.

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

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

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

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

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 ...

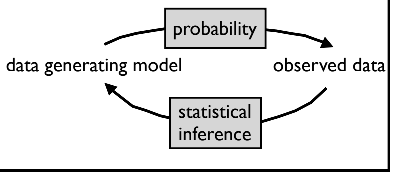
does $E_F(Y) = \mu_Y = \mu_Z = E_G(Z)$?

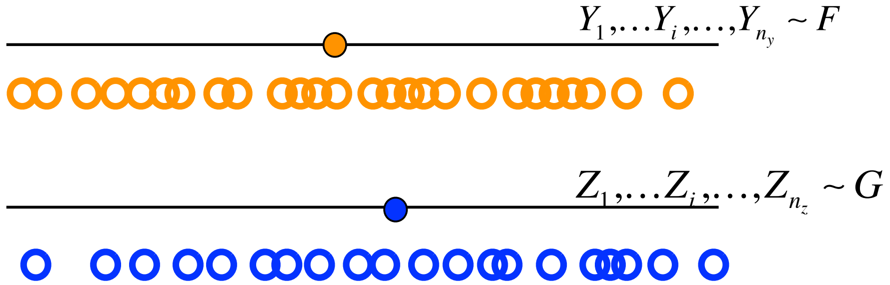
Call this statement the null hypothesis H_0:

$H_0 : \mu_Y = \mu_Z$

Or, equivalently:

$H_0 : \mu_Z - \mu_Y = 0$





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

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

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:

$H_0 : \mu_Y = \mu_Z$

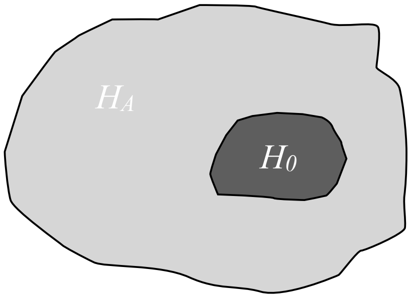
(Occam's razor =))

Or, equivalently:

$H_0 : \mu_Y - \mu_Z = 0$

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

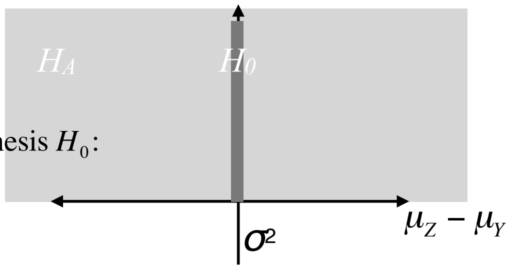
$$Y_1,\ldots Y_i,\ldots,Y_{n_y} \sim F = N(\mu_Y,\sigma^2)$$

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

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

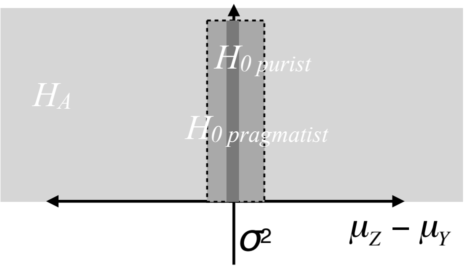


reality check re: null and alternative regions/hypotheses

"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

- **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”?
- Idea: we attempt to find the values of the parameters which would **most likely** produced the data that we in fact observed.
- A good estimator does this without bias, efficiently, and converges to the “true answer” as our sample size increases.

"Most likely": *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)
  - Read "Probability of the data given parameter values"

- 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:

$$D_1 = \{HTHHHTHHHT\} \qquad P(D \mid \theta) = p^7(1-p)^3$$
$$D_2 = \{HTH\} \qquad P(D \mid \theta) = p^2(1-p)$$
$$D_3 = \{TTTH\} \qquad P(D \mid \theta) = p^3(1-p)$$

.

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)

Note that this probability might be really small for any given choice of Θ; what we usually care about is how large it is relative to other choices of Θ.

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

The coin example (Bernoulli)

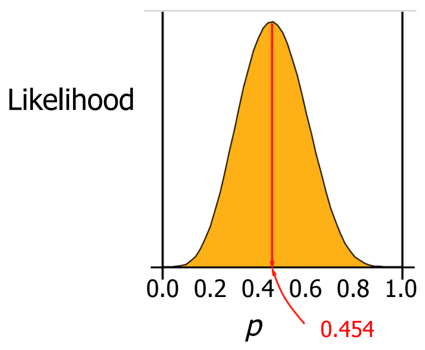
- We have data from 11 tosses of a coin (we don’t know *p* the probability of head)
  - RV = outcome is head
- Outcome of the experiment: {HHTHTTTHTTH}
- Probability of the outcome of the experiment:

  $pp(1-p)p(1-p)(1-p)(1-p)p(1-p)(1-p)p$
- The likelihood is L(Data| p)=$p^5(1-p)^6$
- We can now ask which choice of p maximizes this.

We can plot the data against its likelihood to figure out when we reach the maximum of the likelihood function.

Likelihood function

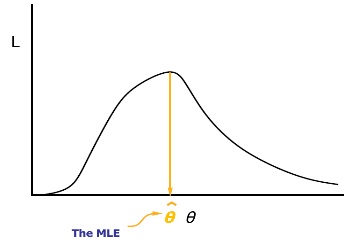
The likelihood is L(Data| p)=$p^5(1-p)^6$



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.

(When there are more than one parameter, likelihood function is multi-dimensional)



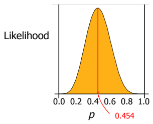
- Often easier to deal with **log** of the likelihood function
  - Partly because of tiny values → computer round-off errors, but also because products turn into sums etc. – math is easier.
  - Log (L) achieves its maximum at the same parameter values as L
- Note that “simple” (i.e., convex) likelihood 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.
- Specific example of 5 heads in 11 fair coin tosses:

  Log L = Log ($p^5(1-p)^6$) = 5 log(p) + 6 log(1-p)

- Differentiate the log L function and set derivative to zero to find maximum.
  - We will arrive at $\hat{p}$ = 5/11 = 0.4545
  - Matches our intuition: observed fraction of heads.



- Note interesting problem: what if we were "unlucky" and got 0 heads in our experiment? Is our estimate of p still reasonable?

World view according to Bayesians

- Frequentist statistics ("classic" statistics) assumes that parameters are ***fixed*** quantities that we want to estimate as precisely as possible (from data)
- Bayesian perspective: parameters are random variables; probability assigned to values of parameters to reflect the degree of belief in a value.
  - Even without any data, we still have some belief about the distribution of the parameter values (the **prior**)
  - Data simply adjusts our beliefs; more data=more adjustment
  - This approach addresses many philosophical (and at least quasi-practical) problems with the frequentist approach

Bayesian estimation

- In order to make probability statements about θ we make use of Bayes' rule:

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

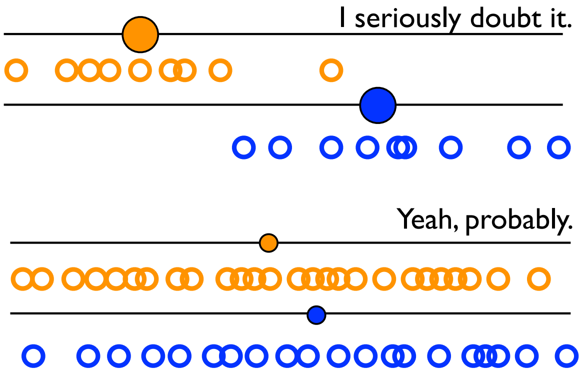
$$\boxed{P(\theta \mid D) \propto P(\theta)P(D \mid \theta)}$$

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

- Each P(.) is a distribution
- Prior is chosen (e.g. empirically, or for convenience, etc.)
- Data adjusts estimates (might be "away" from the prior)
- Possible point estimate of Θ: mode of the posterior
- We will encounter Bayesian methods when we discuss limma (lect 10)

Back to hypothesis testing …

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



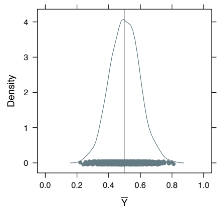
Properties of a test statistic

- 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 p-value 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



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

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 $t \sim t_{ny + nz - 2}$

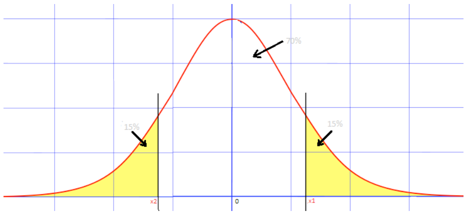
$t_{ny+nz-2}$ is the **sampling distribution** of the t statistic obtained for two sample means from the same normal distribution (IID).

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

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

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

sum of the yellow areas = p-value



obs. value of test stat.

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

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 statistic 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

- p-value …

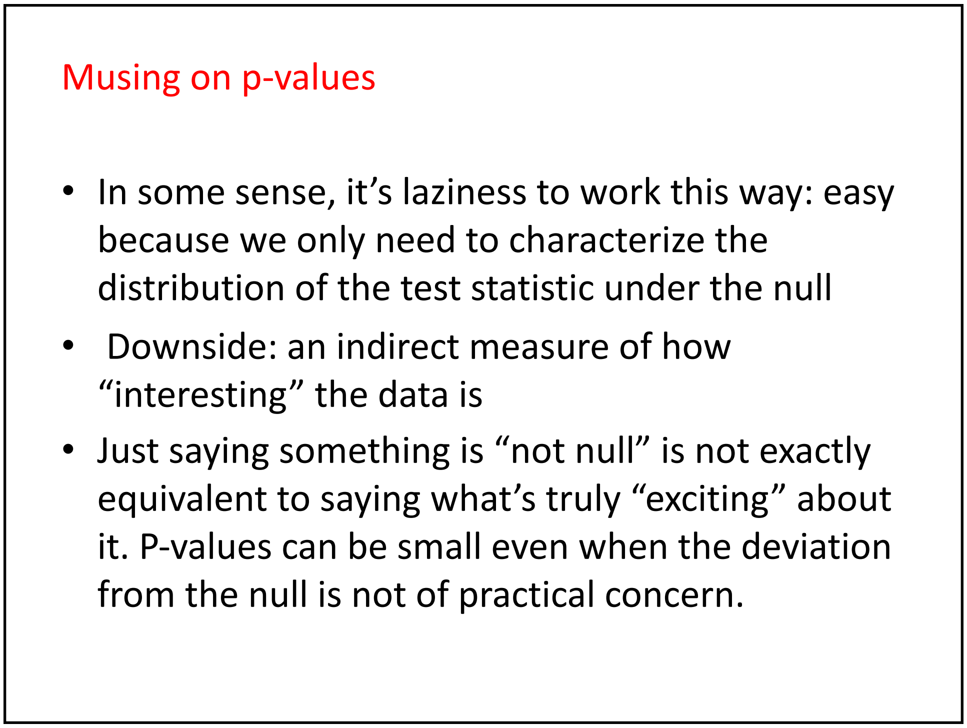
- The probability under the null H_0 of observing a test statistic *value* as or more extreme than the one computed from the data.

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

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

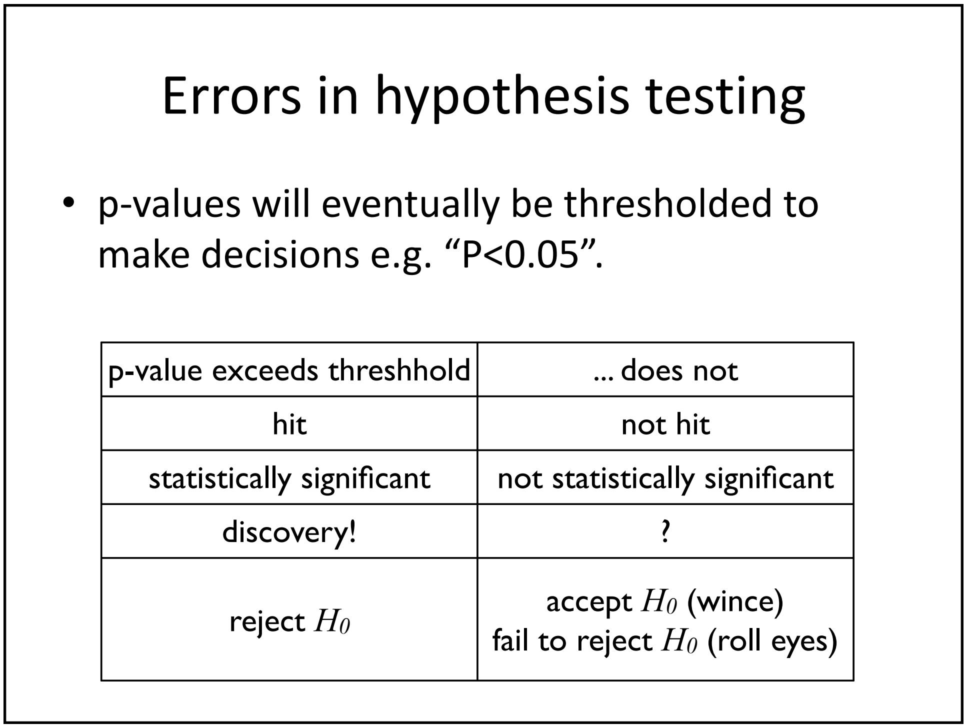
Properties of p-values under the null

- If we treat p-values as an rv: under the null, f(p) is uniform on (0,1)
- What is P(p<0.01)?
- Consider the implications of this; more on this later (and especially in the lecture on multiple test correction)



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” is not exactly equivalent to saying what’s truly “exciting” about it. P-values can be small even when the deviation from the null is not of practical concern.



Errors in hypothesis testing

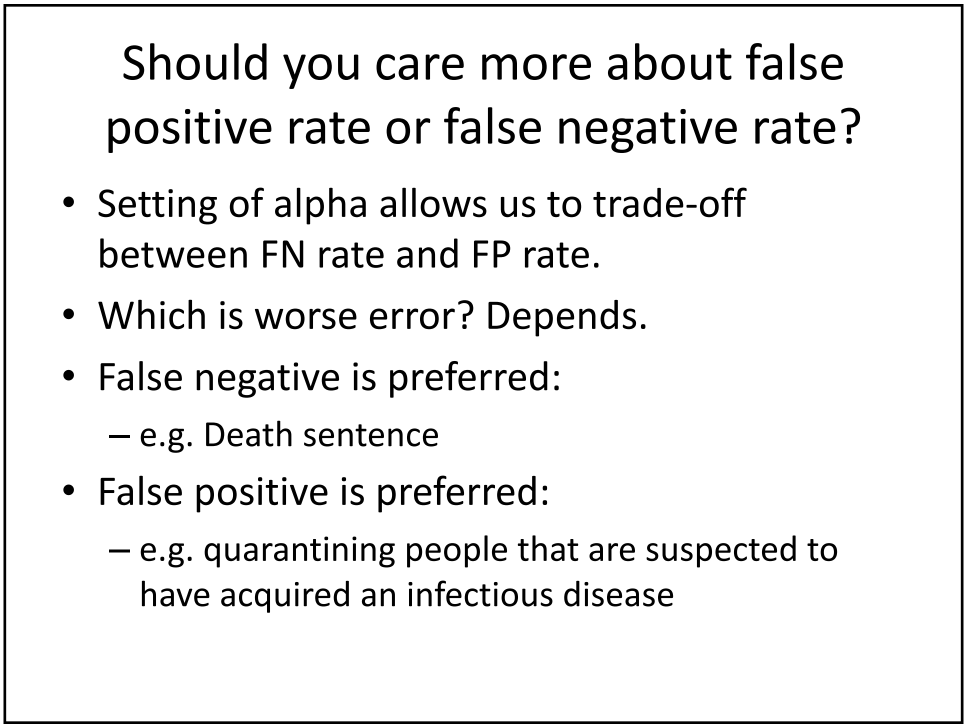
- p-values will eventually be thresholded to make decisions e.g. “P<0.05”.

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 more about false positive rate or false negative rate?

- Setting of alpha allows us to trade-off between FN rate and FP rate.
- Which is worse error? Depends.
- False negative is preferred:
  - e.g. Death sentence
- False positive is preferred:
  - e.g. quarantining people that are suspected to have acquired an infectious disease

Bonus material

- Exercise about E(X) and V(X) for uniform
- Cumulative density functions
- Law of large numbers and central limit theorem
- More about IID

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

First: recall that

$$E(X) = \int x f_x(x)\,dx$$

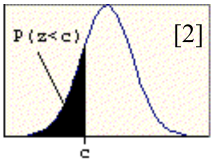
$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2\end{aligned}$$

Answers:

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

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

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)"

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

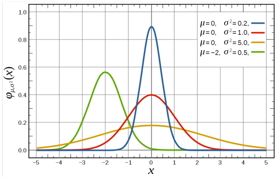
$$F_Y(a) = P(Y \le a) = \sum_{y_i \le a} p_Y(y_i) \text{ (for a discrete Y)}$$

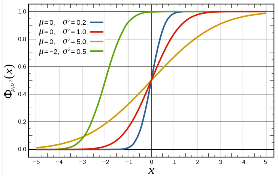
yes, we really do distinguish the density function (continuous rv) from the CDF with the deceptively subtle lowercase "*f*" vs. uppercase "*F*"

density or prob.
mass function

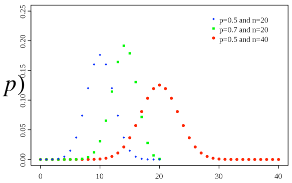
CDF

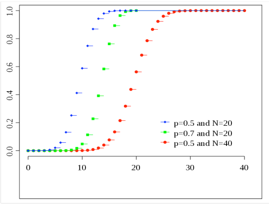
$N(\mu,\sigma^2)$





$Binom(n,p)$

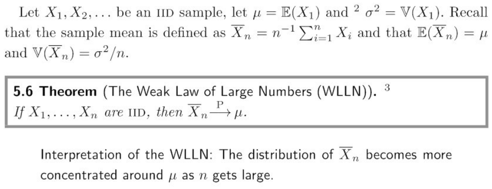




the law of large numbers

common sense "statement":
the average of a large, iid sample will be close to the true mean

the law of large numbers (formally)



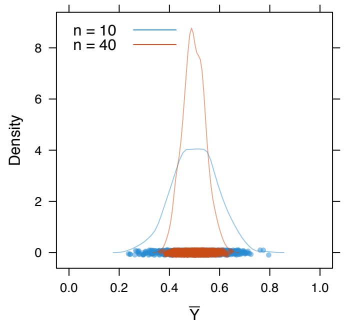
Let $X_1, X_2, \ldots$ be an IID sample, let $\mu = \mathbb{E}(X_1)$ and [2] $\sigma^2 = \mathbb{V}(X_1)$. Recall that the sample mean is defined as $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ and that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n$.

5.6 Theorem (The Weak Law of Large Numbers (WLLN)). [3]
If $X_1, \ldots, X_n$ are IID, *then* $\overline{X}_n \xrightarrow{\mathrm{P}} \mu$.

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

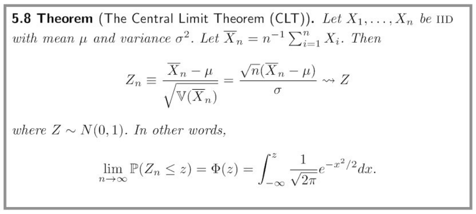
"All of Statistics" page 76

Imagine this trend continuing as n gets bigger and bigger ..... the sample mean sampling dist'n gets more and more concentrated around μ = 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



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

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

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

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

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:

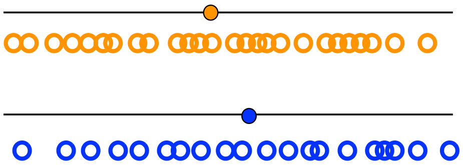
"I can average any large-ish bunch of numbers and divide by the sd and call it a z-score. Then I can compare it to a N(0,1) to determine statistical significance. I've got a hit if the number's greater than 1.96!"

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 (or, at least, you'll have to convince me)

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

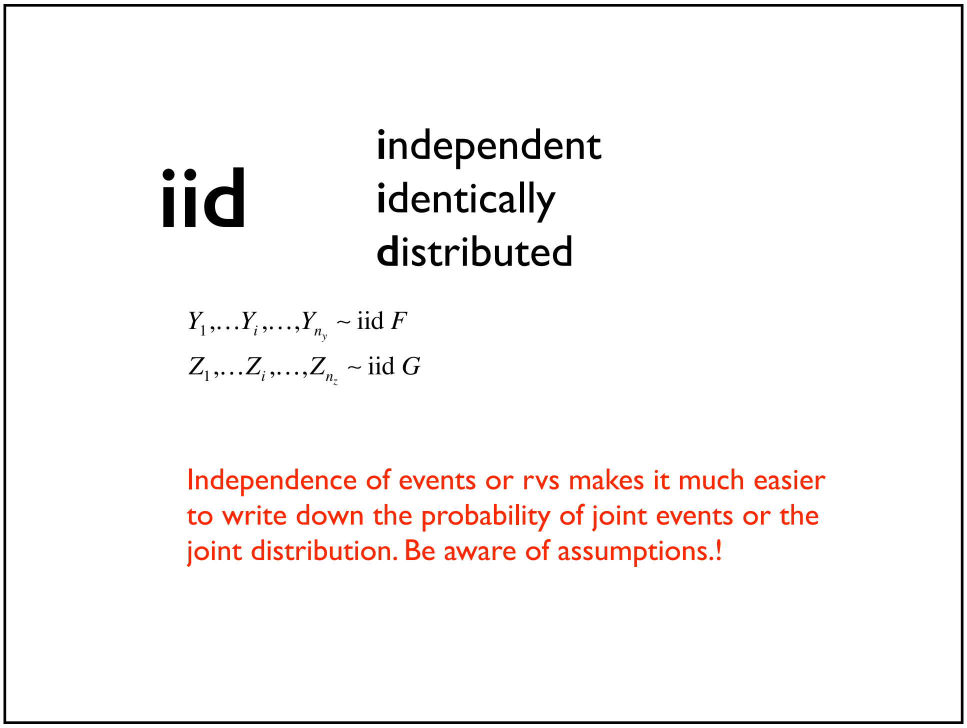
What's the probability of A if you toss a fair coin 10 times?

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

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

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

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



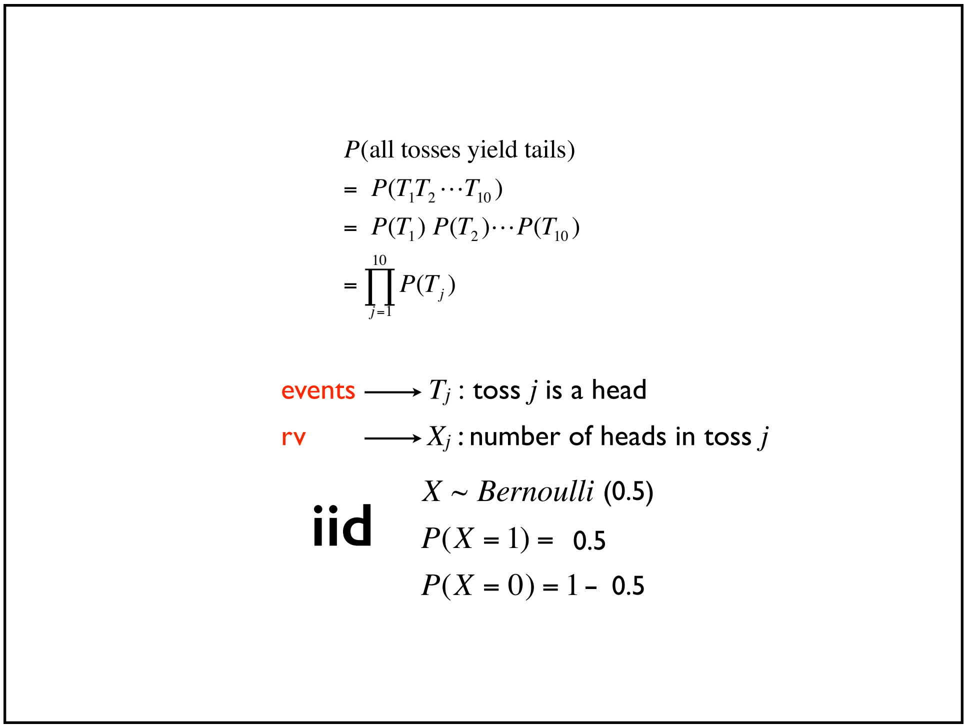
iid

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.!



$$\begin{aligned} &P(\text{all tosses yield tails}) \\ &= P(T_1T_2\cdots T_{10}) \\ &= P(T_1)\,P(T_2)\cdots P(T_{10}) \\ &= \prod_{j=1}^{10} P(T_j) \end{aligned}$$

events ⟶ T_j : toss j is a head

rv ⟶ 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

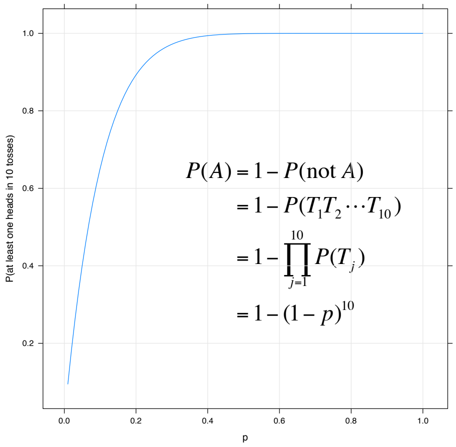
P(T_j) = 1 - p

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

X : number of heads in 10 tosses

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

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



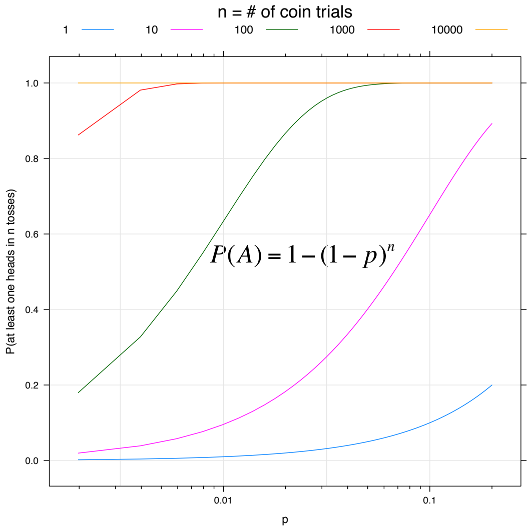
$$
\begin{aligned}
P(A) &= 1 - P(\text{not } A) \\
&= 1 - P(T_1 T_2 \cdots T_{10}) \\
&= 1 - \prod_{j=1}^{10} P(T_j) \\
&= 1 - (1-p)^{10}
\end{aligned}
$$

Parameter value

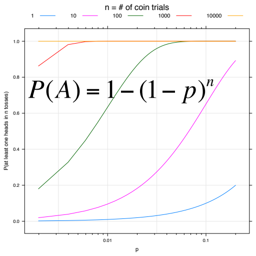
Increasing abstraction and sneaky foreshadowing of the incredible multiple testing problems faced in genomics……..

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

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



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



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

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.