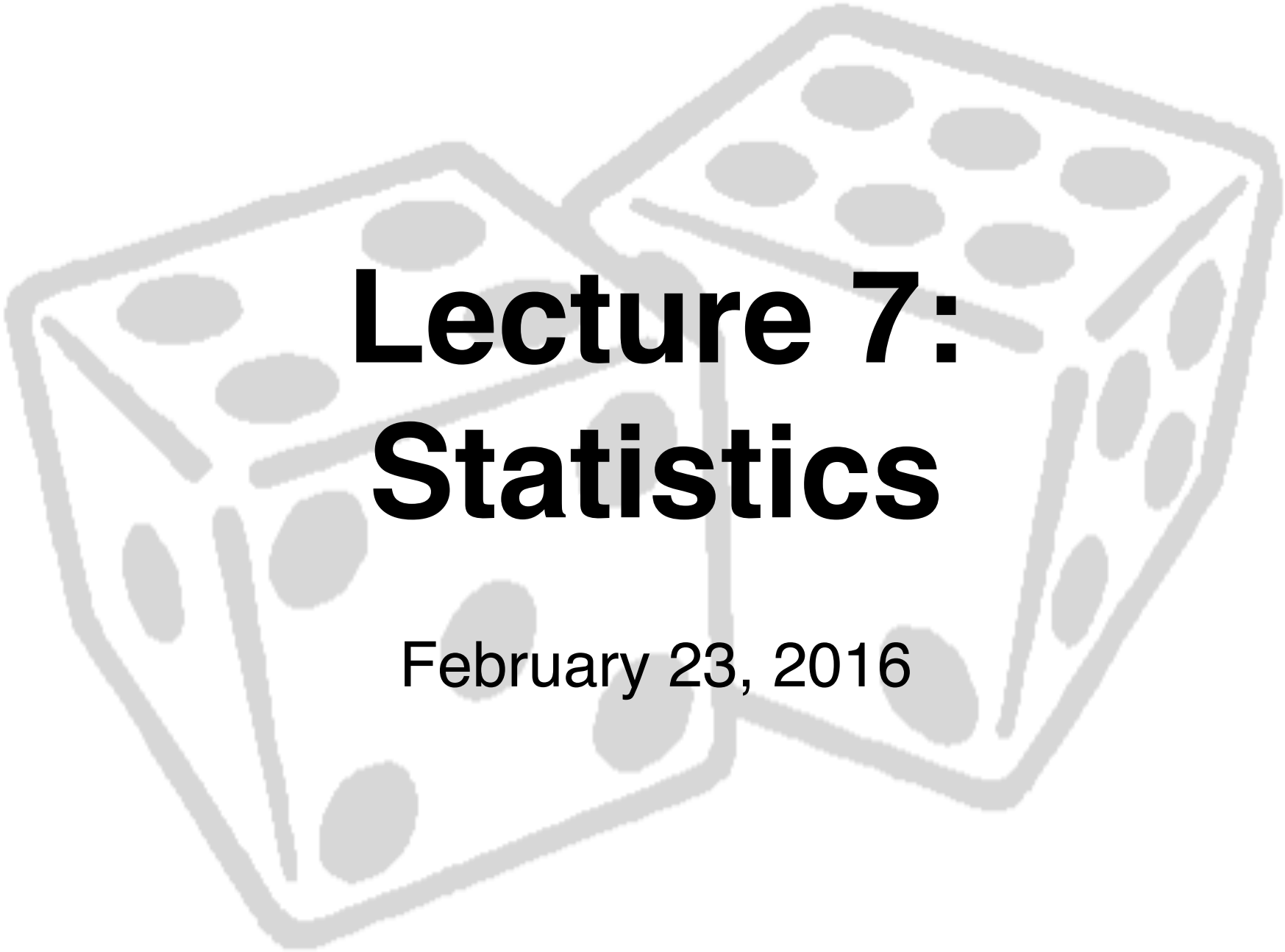


Lecture 7: Statistics

February 23, 2016

Overview

Introduction

Data statistics

- Random variables, common probability distributions
- mean, variance, standard deviation, standard error
- confidence intervals

Hypothesis testing

- parametric and nonparametric tests
- multiple comparisons
- bootstrapping and shuffling
- interpreting a p-value
- dependence and correlation
- common mistakes

Overview

Introduction

Data statistics

- Random variables, distributions
- mean, variance, standard deviation, standard error
- confidence intervals

Hypothesis testing

- parametric and nonparametric tests
- multiple comparisons
- bootstrapping and shuffling
- interpreting a p-value
- dependence and correlation
- common mistakes

Introduction

Currently – large issue with irreproducible results. Most of this is probably not fraud, but "sloppy science", including sloppy statistics

NATURE | COMMENT

Research methods: Know when your numbers are significant

David L. Vaux

Nature **492**, 180–181 (13 December 2012) | doi:10.1038/492180a
Published online 12 December 2012

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

PDF | Citation | Reprints | Rights & permissions | Article metrics

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Scientific method: Statistical errors

***P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.**

Regina Nuzzo

Error prone

Nature **487**, 406 (26 July 2012) | doi:10.1038/487406a
Published online 25 July 2012

PDF | Citation | Reprints | Rights & permissions | Article metrics

Biologists must realize the pitfalls of work on massive amounts of data.

NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

In the past couple of years, the *Nature* journals publishing biological research have started paying much more attention to statistics. We have appointed a statistical advisor, Terry Hyslop from Duke University, who has helped us assemble a panel of statisticians who act as consultants on certain papers.

Introduction

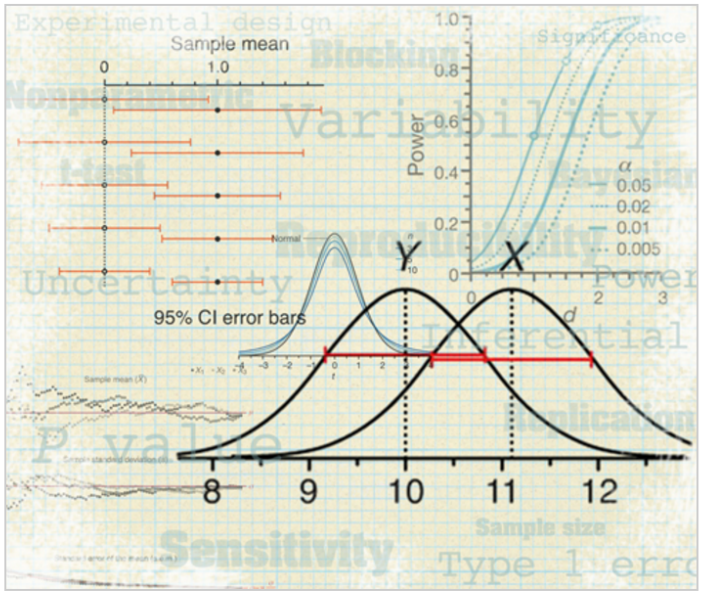
WEB COLLECTION

Search Go

Advanced search

Statistics for biologists

Home | Practical guides | Statistics in biology | Points of Significance | Other resources



There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late.

This collection highlights important statistical issues that biologists should be aware of and provides practical advice to help them improve the rigor of their work.

Nature Methods' **Points of Significance** column on statistics explains many key statistical and experimental design concepts. **Other resources** include an online plotting tool and links to statistics guides from other publishers.

Image Credit: Erin DeWalt

A lot of today's lecture materials comes from these columns!

Statistics in biology

Overview

Introduction

Data statistics

- Random variables, distributions
- mean, variance, standard deviation, standard error
- confidence intervals

Hypothesis testing

- parametric and nonparametric tests
- multiple comparisons
- bootstrapping and shuffling
- interpreting a p-values
- dependence and correlation
- common mistakes

What is a random variable?

heads (0)

lose $5

tails (1)

win $5

Coin flipping:

Bernoulli random variable

Definition of random variable: a variable whose value is subject to variations due to chance (from Wikipedia)

Building up a distribution from a random variable

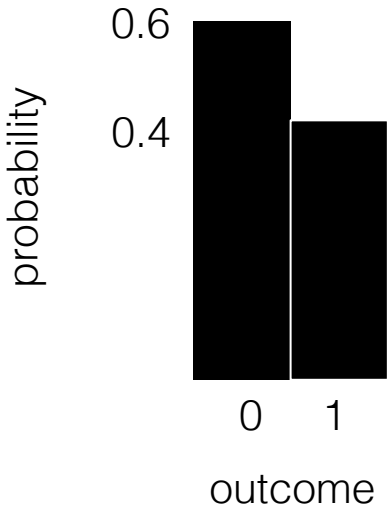
Outcomes of a bunch of flips

0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1

9 heads, 6 tails

heads (0)

tails (1)



Bernoulli distribution

Definition of probability distribution: a distribution that assigns a probability to each possible outcome from an experiment (from Wikipedia)

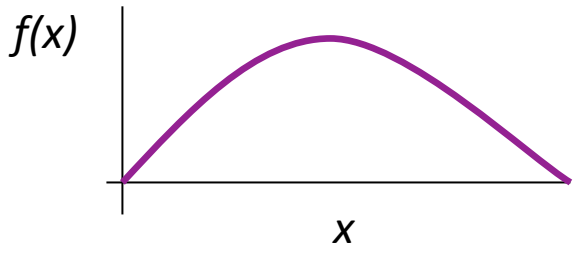
Bernoulli distribution is defined by two parameters: p (probability) and options (k = 0 or 1)

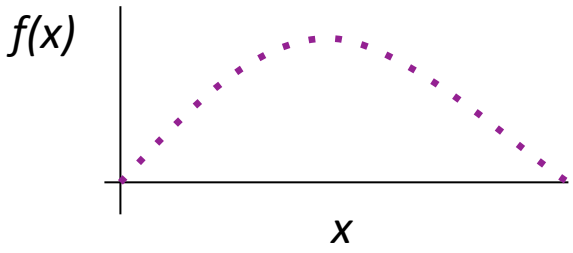
$$f(k,p) = p^k(1-p)^{1-k}$$

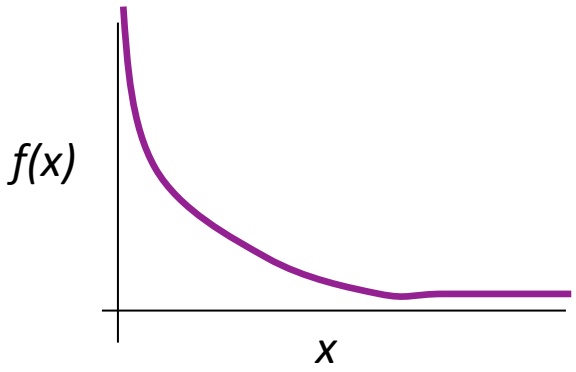
Probability distributions

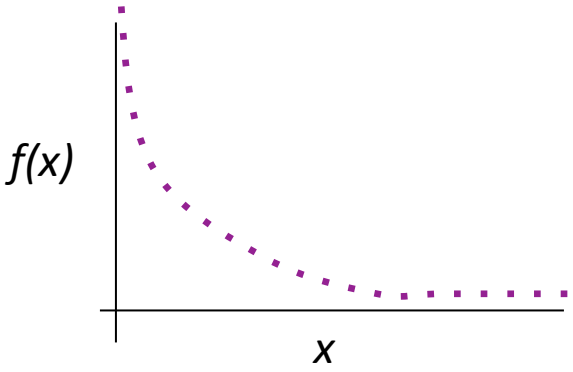
Probability distributions can look like anything

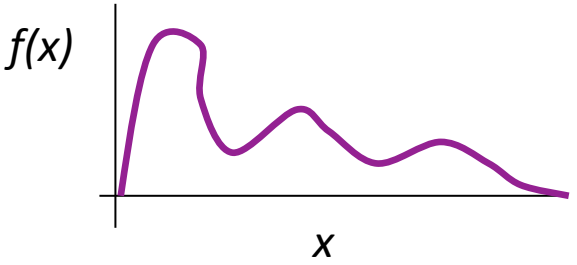
Can be continuous (Gaussian) or discrete (Poisson)

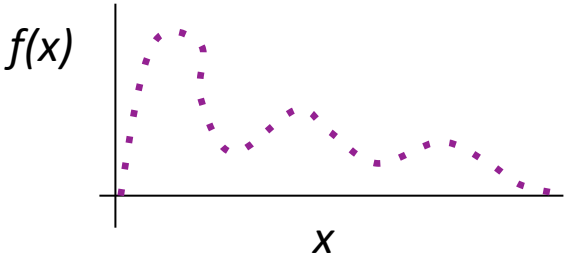










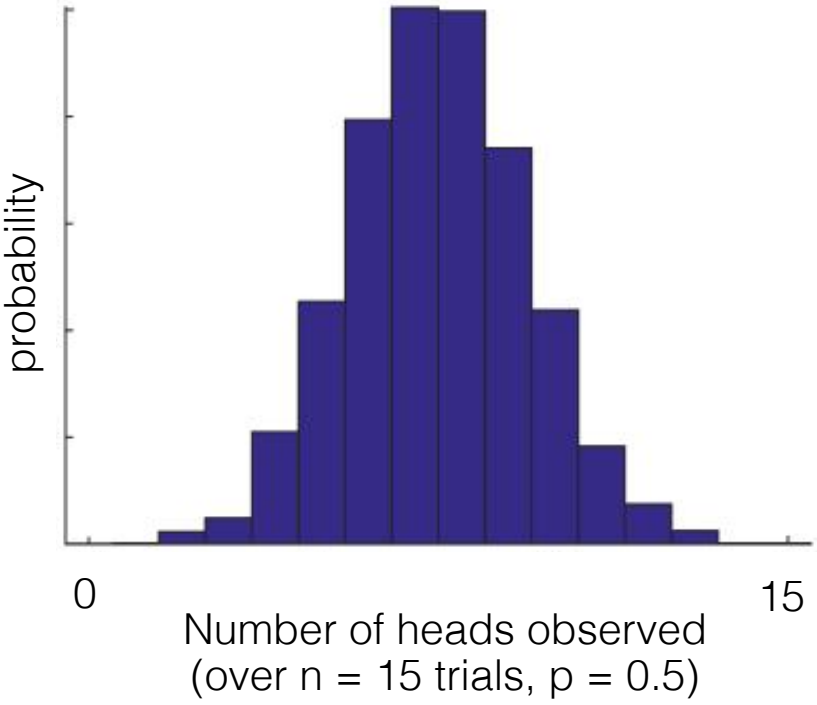


"Probability density function"

"Probability mass function"

Probability distributions: binomial distribution

How many heads are observed in 15 flips?



heads (0)

tails (1)

MATLAB: `binornd(15,0.5,1000,1)`

Binomial distribution is defined by three parameters: k (number of 'successes'), p (probability), n (number of trials)

$$f(k,p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

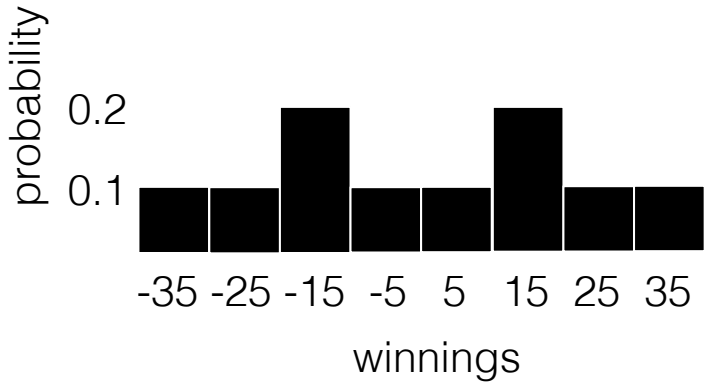
Probability of getting k successes given n and p

[Bernoulli distribution is the binomial distribution with n = 1]

Probability distributions: another distribution

But what if you make a histogram of your winning for each trial… what does this distribution look like? What does it look like for a lot of trials?

Outcomes	Winnings
0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1	-$15
0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0	$5
1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1	$35
1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0	-$15
1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0	$25
0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1	-$5
0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0	$5
0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1	$15

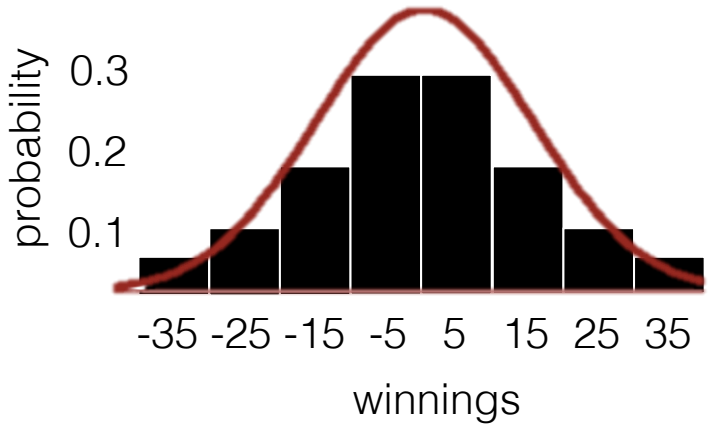


Probability distributions: normal distribution

But what if you make a histogram of your winning for each trial… what does this distribution look like?

A normal distribution!

Outcomes	Winnings
0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1	-$15
0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0	$5
1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1	$35
1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0	-$15
1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0	$25
0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1	-$5
0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0	$5
0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1	$15
…	



Probability distributions: normal distribution

This is because of the central limit theorem:

The average of many samples will tend to be normally distributed, regardless of the underlying distribution

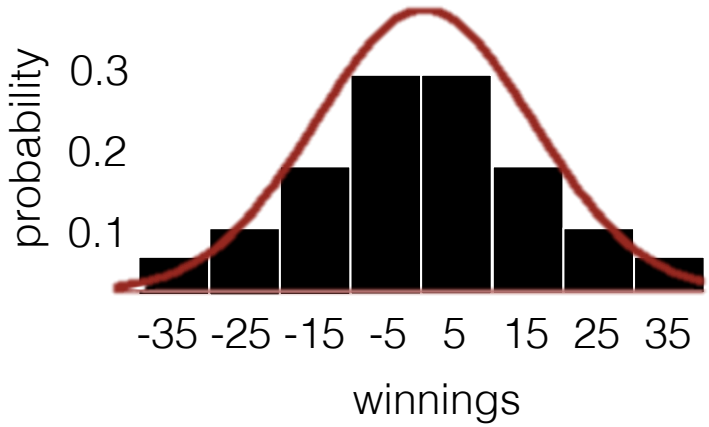
Samples have to be “independent and identically distributed” (i.i.d.) – each sample is independent of the ones before and after it

Binomial distribution with large n will converge to a normal distribution

<u>Examples of times when CLT will pop up:</u>

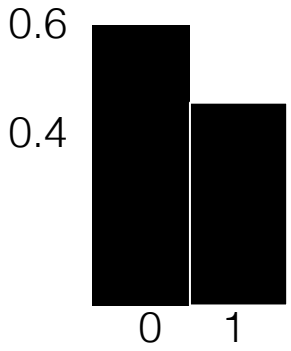
Membrane potential noise from many ion channels opening and closing randomly

Average french fry length in a batch of fries



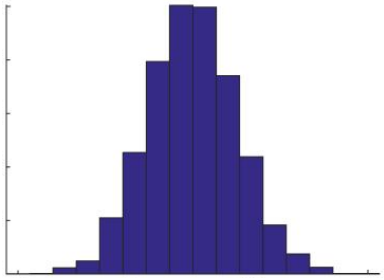
Common probability distributions in biology

Bernoulli distribution



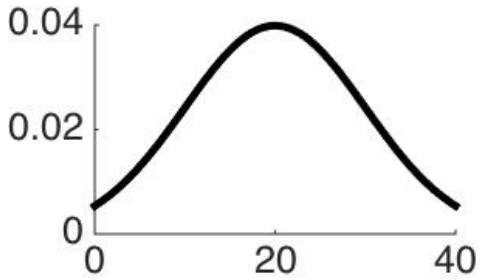
$$f(k,p) = p^k(1-p)^{1-k}$$

Binomial distribution



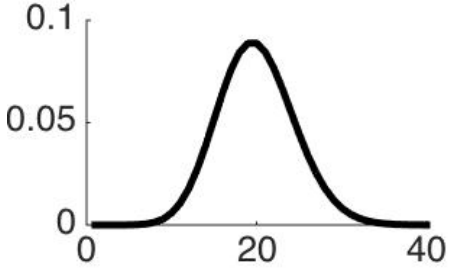
$$f(k,p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Normal distribution



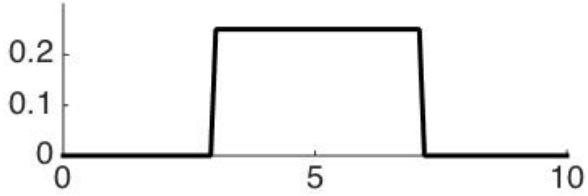
$$f(y,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\mu-y)^2}{2\sigma^2}}$$

Poisson distribution



$$f(y,\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

Uniform distribution



$$f(a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

New ones…

Exponential distribution

Gamma distribution

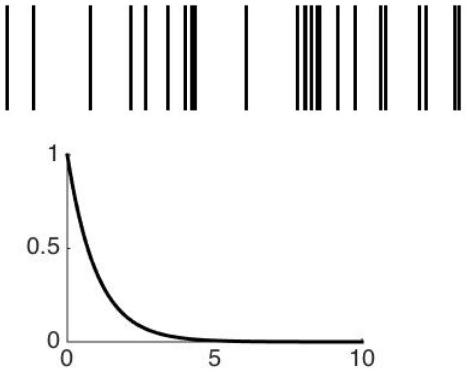
Negative binomial distribution

Common probability distributions in biology

Exponential distribution

A continuous probability distribution that describes the waiting time between events in a Poisson process

Example:
Distribution of ISI's in a spike train:

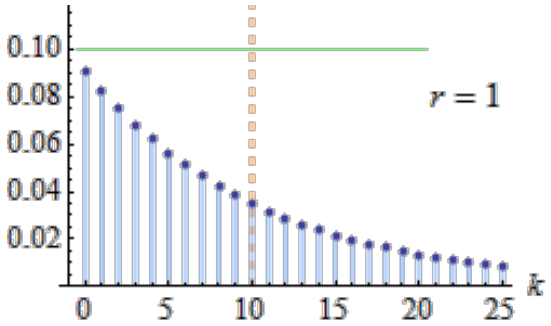


$$f(y,\mu) = \mu e^{-\mu y}$$

Negative binomial distribution

A discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a specific number of failures occurs

Example: Number of heads you get when you flip a coin before getting 5 tails

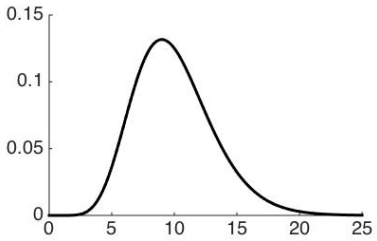


$$f(k,p,r) = \binom{k+r-1}{k} p^k (1-p)^r$$

Gamma distribution

A continuous probability distribution that describes the waiting time until the r^{th} event occurs for a Poisson process

Example: Time until r number of DNA strands are produced during PCR



$$f(y,\mu) = \mu \frac{(\mu y)^{r-1} e^{-\mu y}}{(r-1)!}$$

Common probability distributions in biology

Finding what distribution matches your data helps characterize the mechanism underlying your data

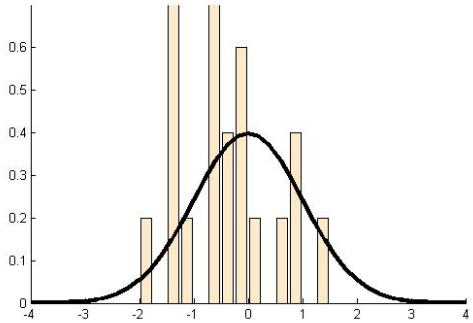
Identifying your distribution is important for choosing a statistical test when you are testing a hypothesis

Common probability distributions in biology

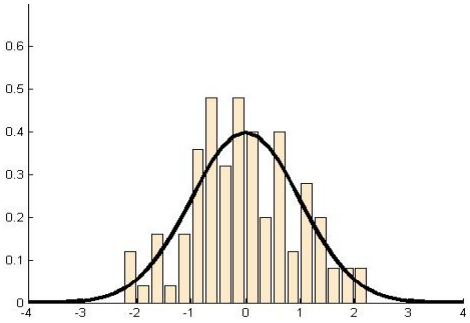
After collecting your data, how do you know its probability distribution?

Sometimes, you'll have a good guess based on the data (e.g. that it follows a binomial)

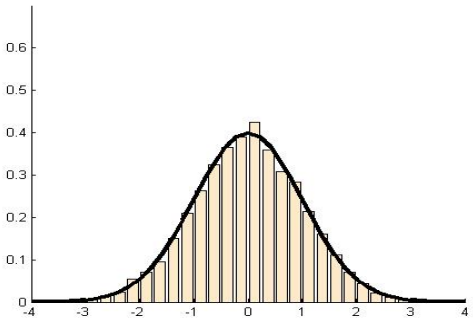
Other times, you'll have to actually make a plot of the probability distribution



n=20



n=100

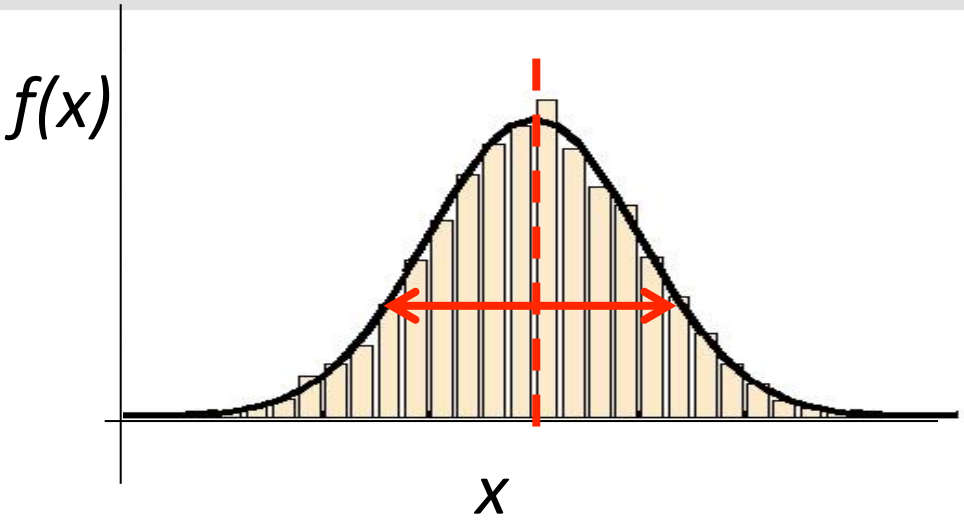


n=1000

To do this: Just bin your data by making a histogram.
The more data, the better approximation.

MATLAB:
```
[count, bins] = hist(data)
probs = count/sum(count)
bar(bins, probs)
```

Finding values of interest in probability distribution



There are a couple of things that we typically have to find once we have a probability distribution

"Point statistics"

Mean $$E[x] = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} P(x_i)x_i$$

Together: characterize location and spread

Median – center of all observations

Variance / standard deviation $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = E[(x - \bar{x})^2]$$

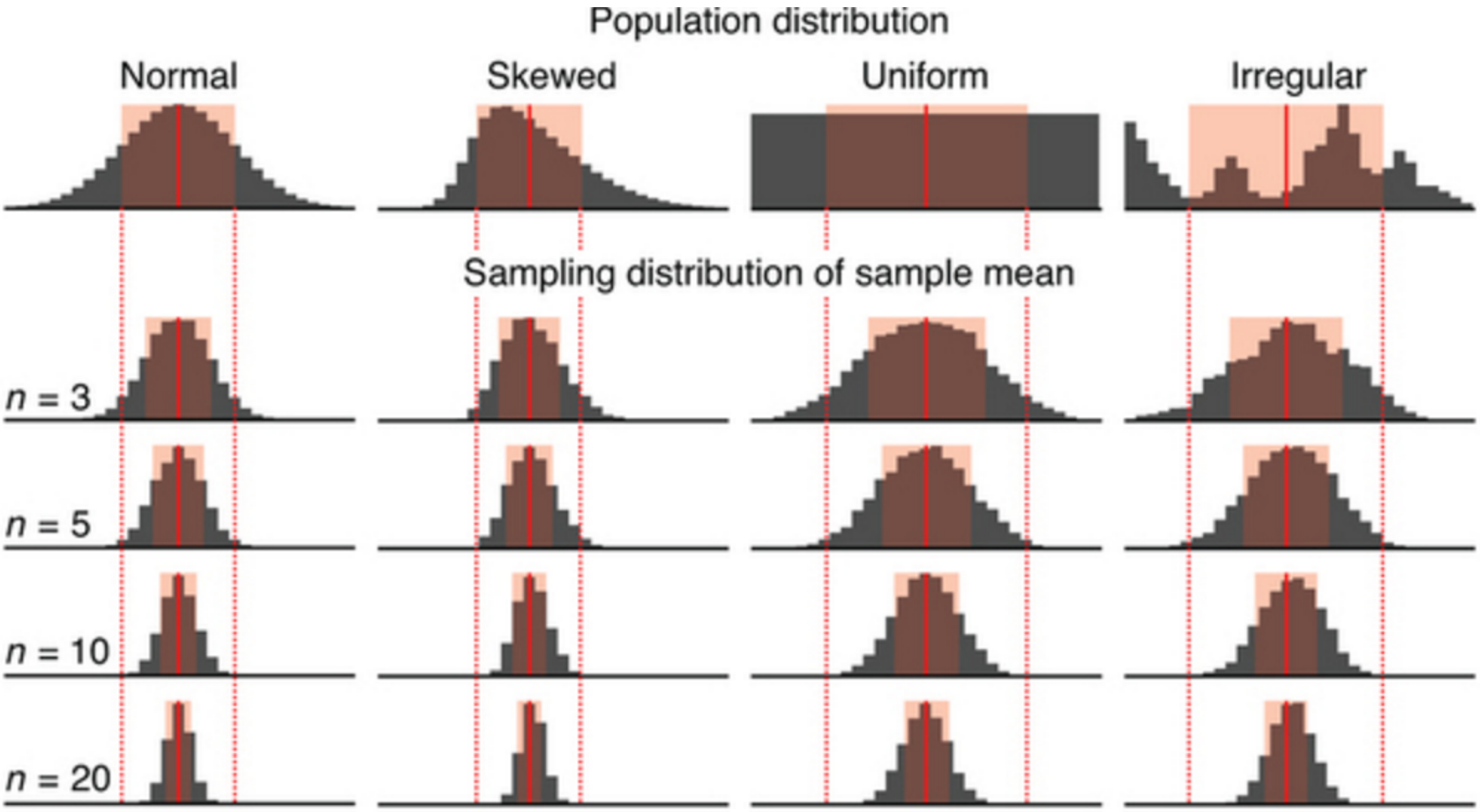
We usually cannot access the true mean (or variance) – instead, we sample it

Estimate s.d. in our estimate of the mean: Standard error of the mean $\frac{\sigma}{\sqrt{n}}$

Standard error of the mean

Standard error of the mean is the standard deviation of the distribution of our estimate of the mean

[How can we get this distribution?]



CLT says that the distribution of our estimate of the mean is normal!

Mean of sample mean is the population mean

Need to quadruple amount of data to get double the precision

Ways to make error bars

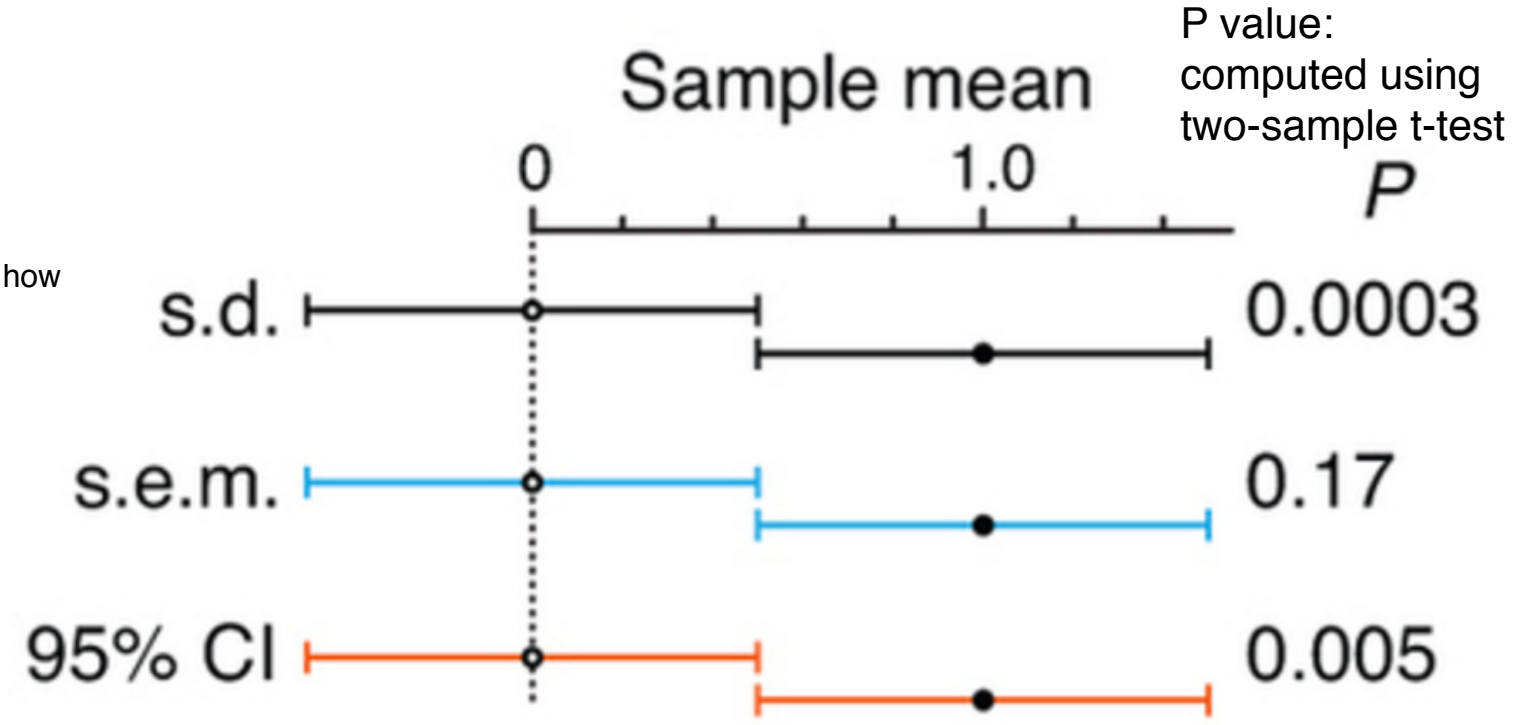
In most papers, authors choose to plot error bars using standard deviation or standard error (about half and half). Side note – only about 2/3 of papers use error bars when they should

A third way to make error bars: confidence intervals

S.D. – to investigate full population distribution – how much overlap?

S.E.M– to investigate estimate of mean – are these significantly different?

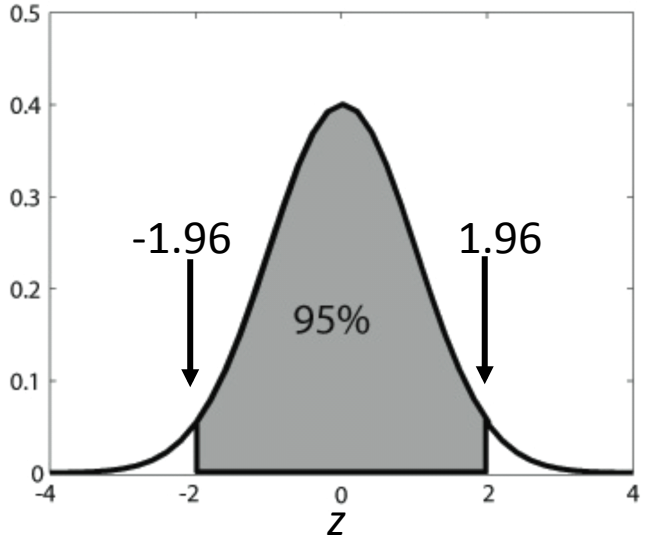
C.I. – how reliable is this measurement?



P value: computed using two-sample t-test

Calculating a confidence interval

The standard normal distribution (Z): mean 0 and standard deviation 1.



P(-1.96 < Z < 1.96) = 0.95

95% of the time, a sample from the Z distribution is between -1.96 and +1.96 (CI)

Relate other distributions back to this distribution

Transform a sample distribution to the standard normal distribution

$$Z = \frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}}$$

$$P(-1.96 < \frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

$$P\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

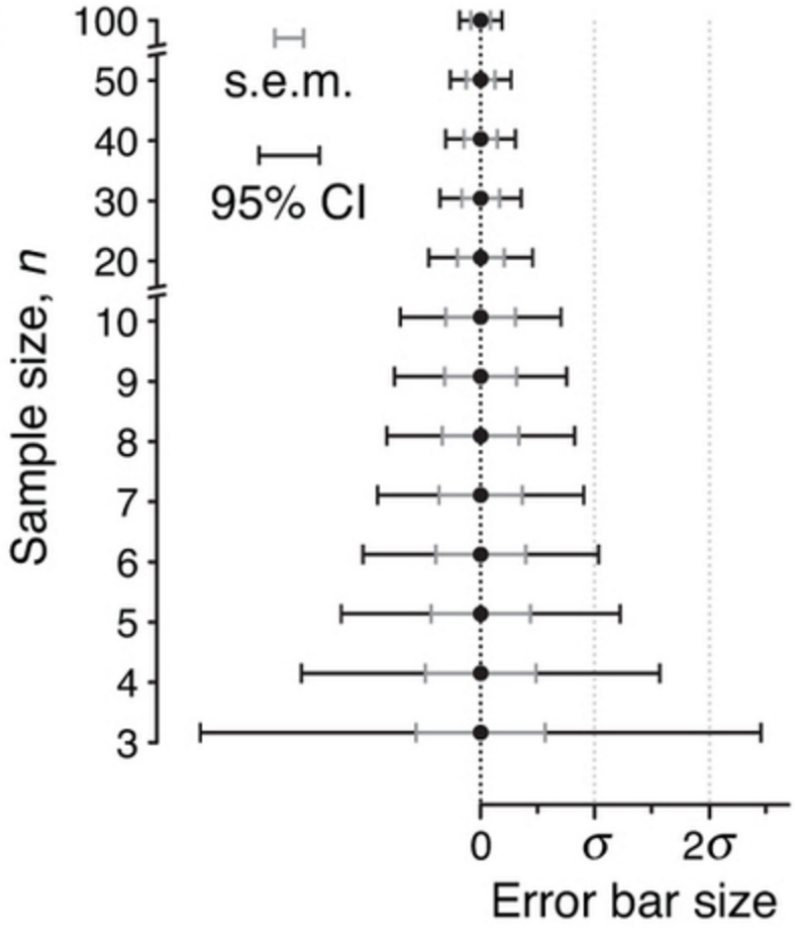
$$P\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\hat{\mu} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

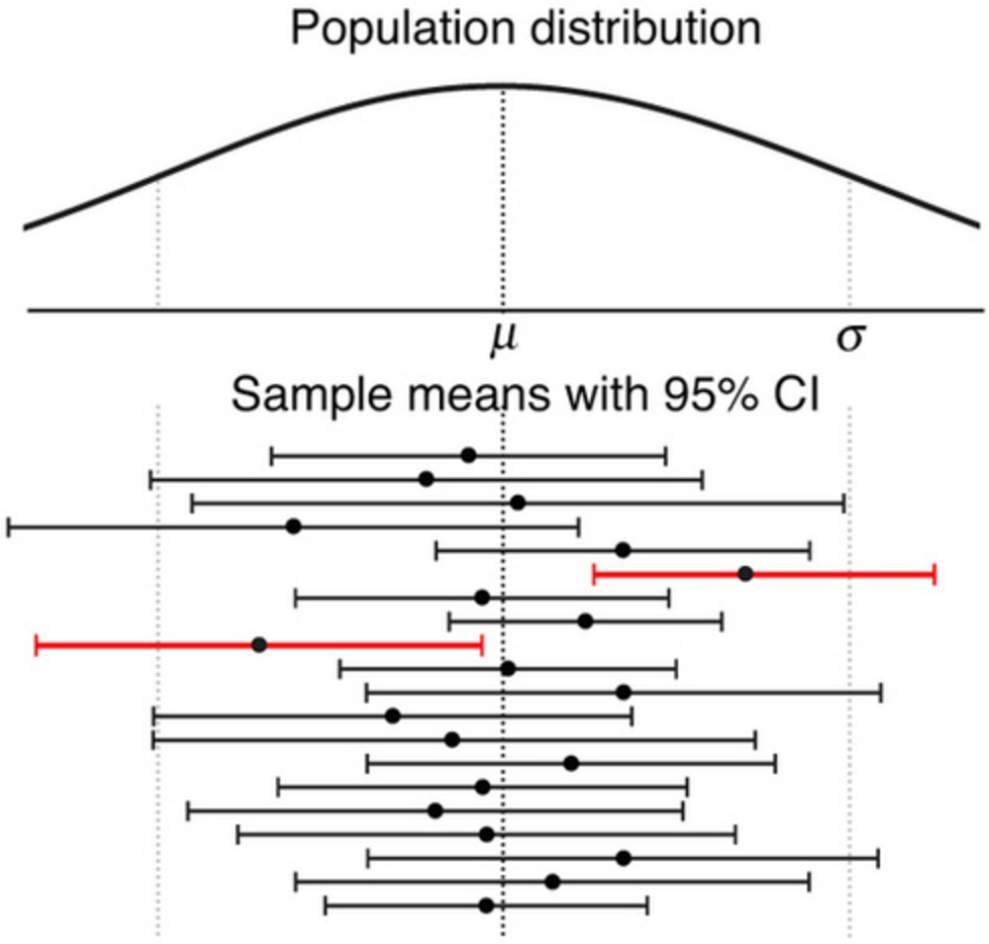
Without the 1.96, this is SEM – 67% CI

Confidence interval vs SEM

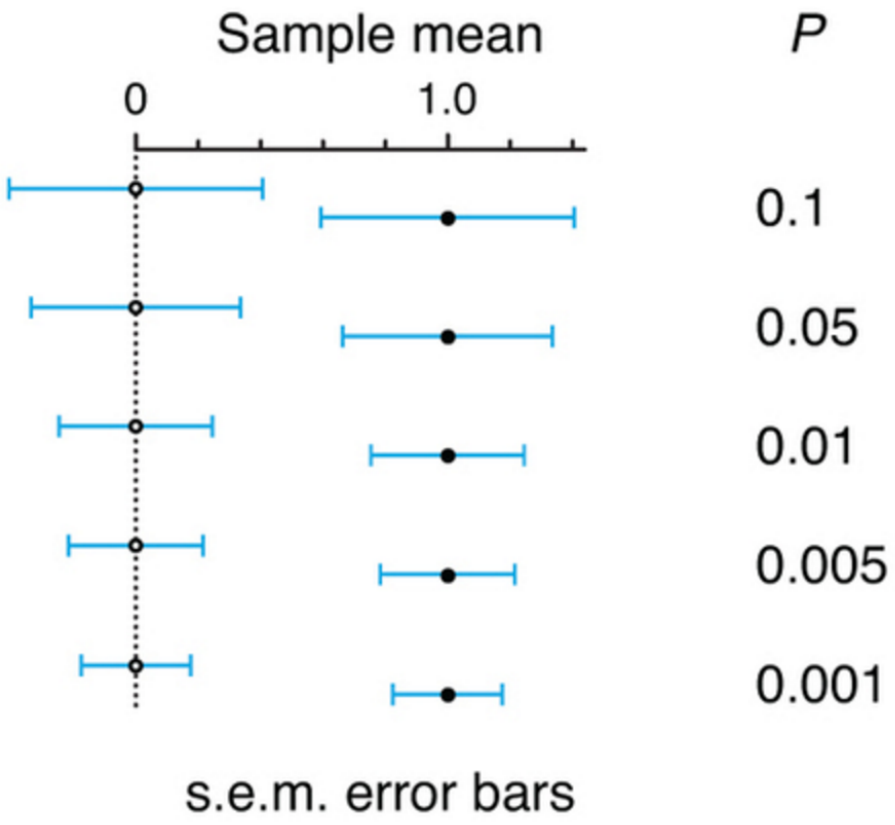
Compare SEM and CI

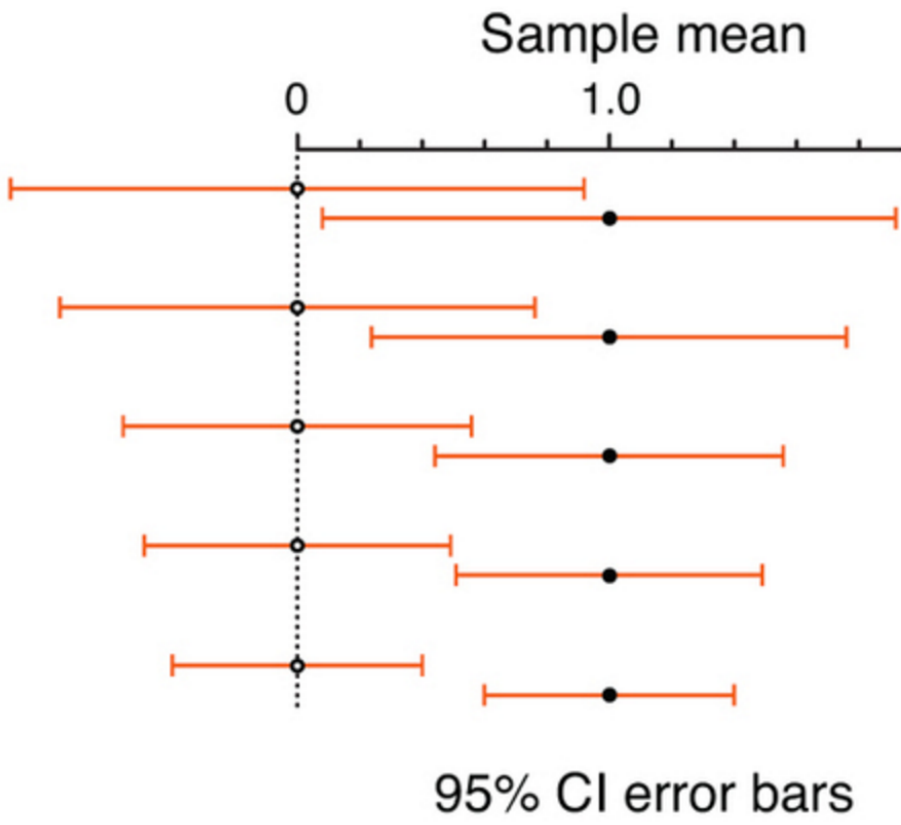


Interpretation of a confidence interval: if this experiment was repeated for a total of 100 times, 5 of the times the interval wouldn't contain the real mean



Eyeballing significance





s.e.m. error bars

95% CI error bars

Overview

Introduction

Data statistics

- Random variables, distributions
- mean, variance, standard deviation, standard error
- confidence intervals

Hypothesis testing

- parametric and nonparametric tests
- multiple comparisons
- bootstrapping and shuffling
- interpreting a p-value
- dependence and correlation
- common mistakes

Hypothesis testing

Is this a fair game? What if the coin isn't fair?

Consider the outcome of several games. If confidence interval doesn't overlap with 0, you might be suspicious..

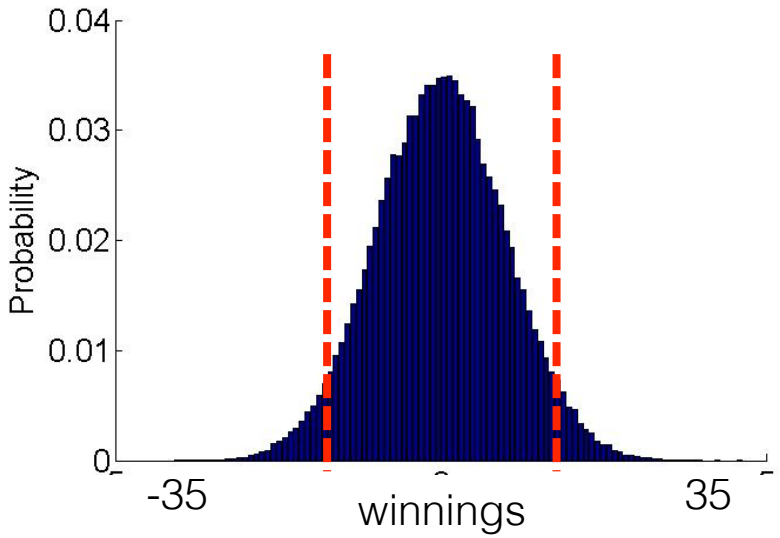
But how do you reject the hypothesis that the coin (and betting game) is far?

Null hypothesis: Game is fair

Alternative hypothesis: Game is not fair

Select cut-off such that the probability observing data greater/ less than the cutoff is < 0.05

Null distribution

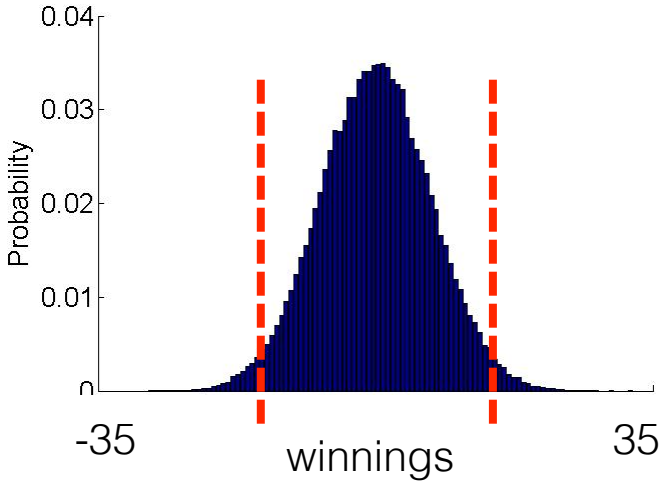


Hypothesis testing

STEPS:

1. Formulate the null hypothesis
2. Select a statistical test
3. Identify the test statistic under the null hypothesis
4. Determine rejection region of the test statistic
5. Calculate test statistic
6. Determine whether or not the test statistic falls within the rejection region – compute a p-value

One MATLAB command!

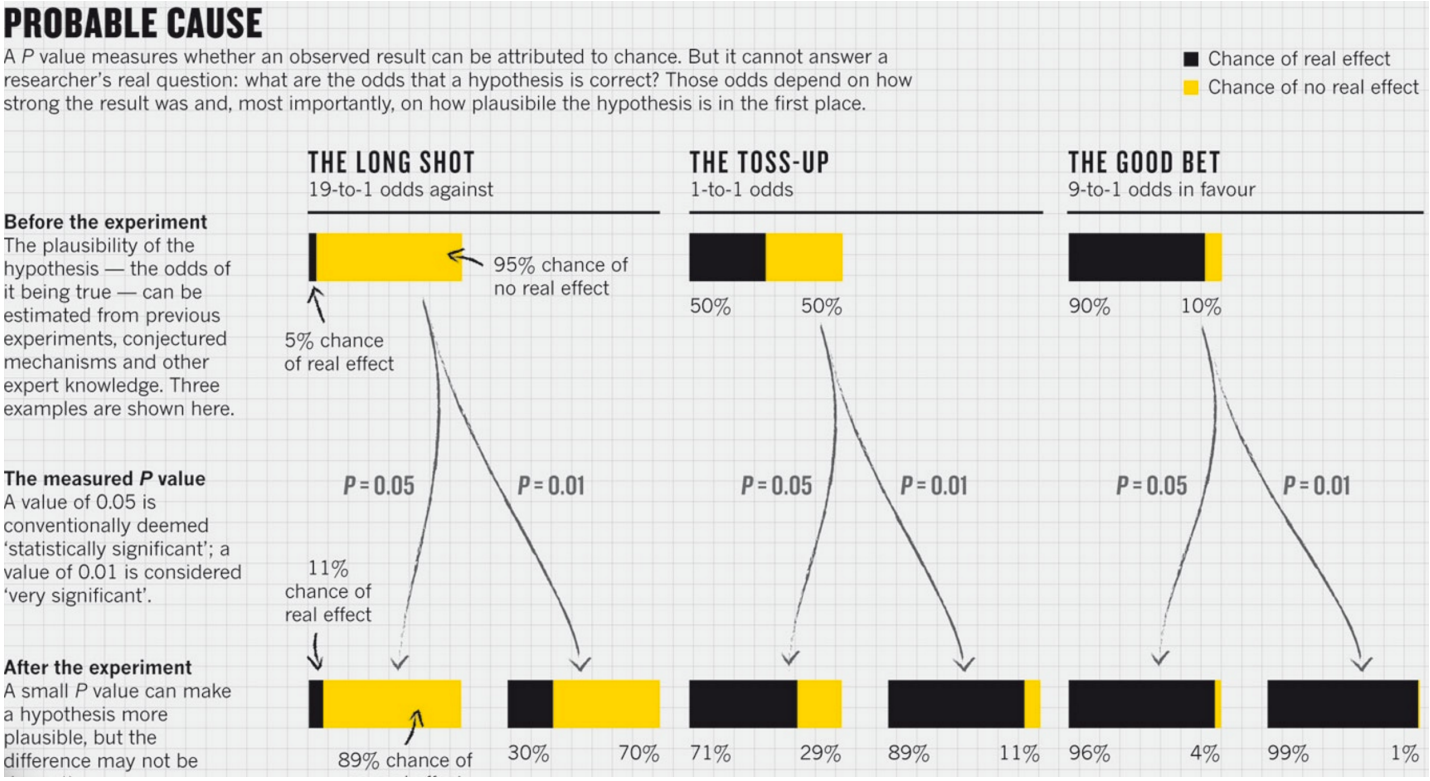


Cautionary notes on the p-value

Definition of a p-value: the chances of finding this result if the null hypothesis is true

It does not tell you whether the null hypothesis is not true, or if your alternative hypothesis is true

In order to say this, you must also incorporate how likely the effect is.



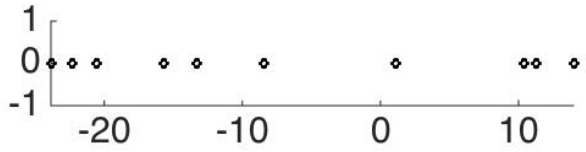
What p-values calculate: $\mathrm{P}(x \geq \mathrm{Data}|\mathrm{H}_0)$

What we want to know: $\mathrm{P}(\mathrm{H}_1|\mathrm{Data}) = \dfrac{\mathrm{P}(\mathrm{Data}|\mathrm{H}_1)\mathrm{P}(\mathrm{H}_1)}{\mathrm{P}(\mathrm{Data})}$

Hypothesis testing: is the coin fair?

Procedure:

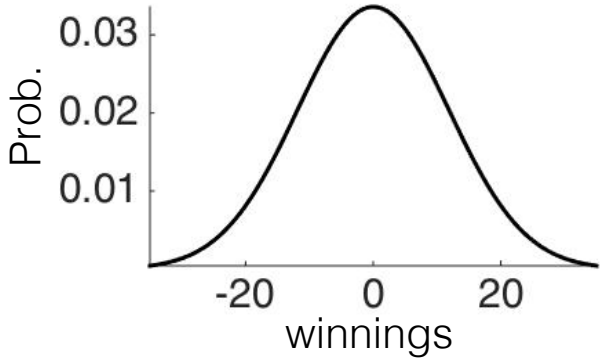
1. Play 10 games and record the winnings for each game



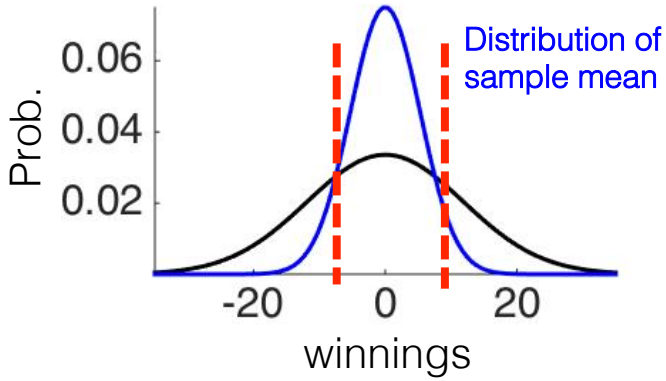
2. Compute the mean and standard deviation of your winnings (assumed to be normal).

Mean = -3.7 S.D. = 3.4

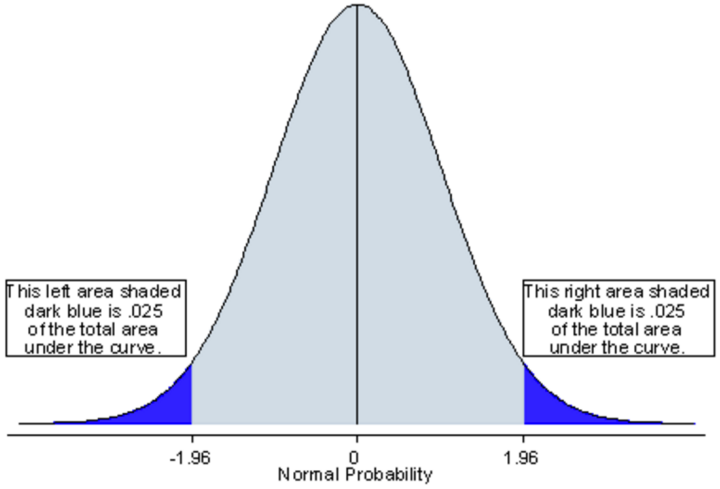
3. Assume that the null distribution has 0 mean (in this case), but the same standard deviation. Then, form the null distribution (also assumed to be normal).



4. We can also form the distribution of sample means, and compute how likely it is that we find a mean of a given magnitude (compute area under curve)

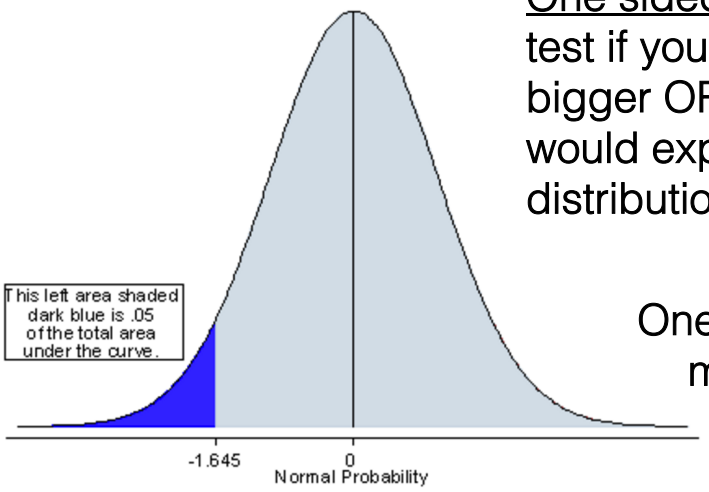


One-sided versus two-sided tests

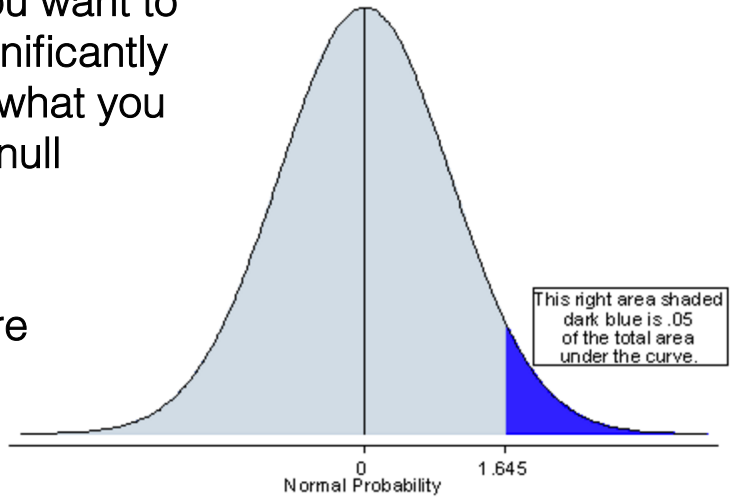


Two sided-test: when you want to test if your sample is significantly different (could be larger, could be smaller) than what you would expect given the null distribution

One sided-test: when you want to test if your sample is significantly bigger OR smaller than what you would expect given the null distribution



One-sided tests are more sensitive



Hypothesis testing and t-statistic

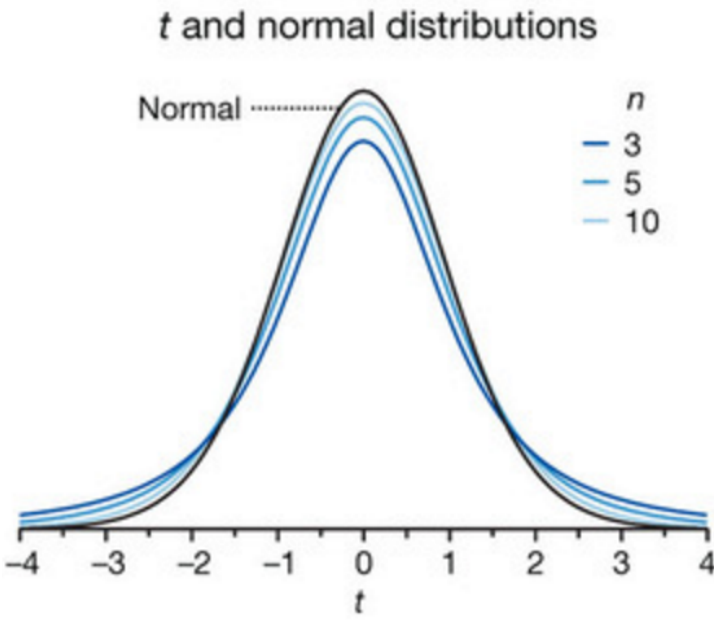
Instead of doing it this way, you can also just compute the t-statistic:

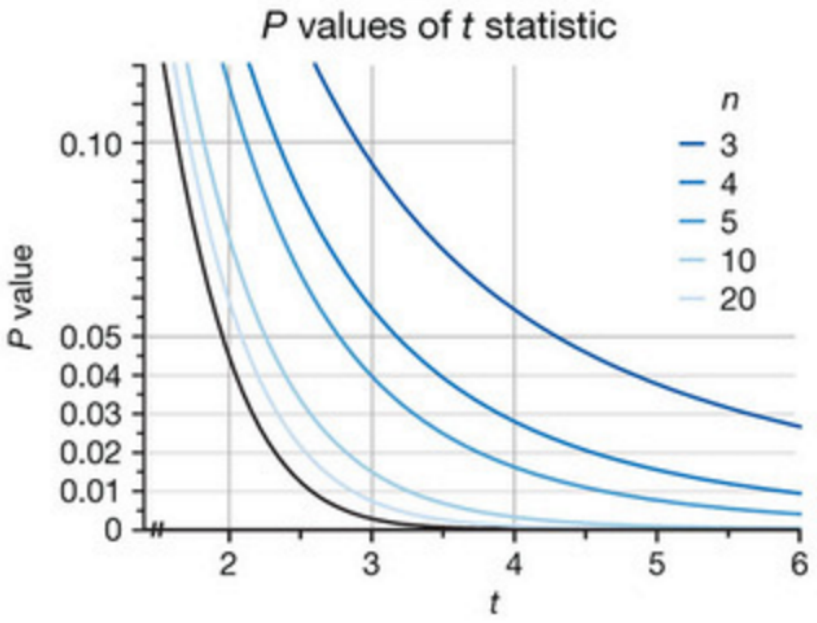
(sample mean – null mean)

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}$$

Sample variance ___underestimates___ the actual variance of the null distribution

So, the distribution isn't quite normal, and follows the t-distribution:





Lots of tests for lots of different scenarios

Handout on website:

What you're testing	Assumptions	Test Name	Hypotheses	Notes
Something about μ	Normal distribution Variance known	z-test	H_0: $\mu_{test} = \mu_{null}$ H_1: $\mu_{test} \neq \mu_{null}$ or $\mu_{test} > \mu_{null}$ or $\mu_{test} > \mu_{null}$	Use a two-tailed t-test if H_1 is an inequality. Use a one-tailed test if H_1 is directional (i.e. you have previous evidence which lets you rule out one direction).
Something about μ	Normal distribution Variance unknown	t-test	H_0: $\mu_{test} = \mu_{null}$ H_1: $\mu_{test} \neq \mu_{null}$ or $\mu_{test} > \mu_{null}$ or $\mu_{test} > \mu_{null}$	Use a two-tailed t-test if H_1 is an inequality. Use a one-tailed test if H_1 is directional (i.e. you have previous evidence which lets you rule out one direction). # degrees of freedom = n-1
How well does an observed frequency distribution of certain events fit the distribution predicted by the null hypothesis?	*n* is larger than 30. Expected frequencies are all larger than 5. Error is normally distributed.	Pearson's chi-square test	H_0: observed distribution comes from predicted distribution H_1: observed distribution doesn't come from predicted distribution	# degrees of freedom = # categories – 1
Are all of the means from several distributions the same?	Everything is normal and has equal variance	ANOVA	H_0: all μ's are equal H_1: at least one pair of μ's are not equal (but DOESN'T tell you which one)	Based on idea that observed value = average + variation within group + variation across groups. Null hypothesis is that variation across groups is 0 Degrees of freedom within groups = sample size within a group -1 Degrees of freedom between groups = # groups -1

So far, we have only dealt with situations where the data satisfies a normality constraint…

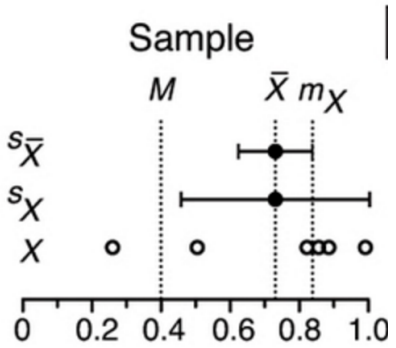
… but what if this isn't the case?

Parametric versus non-parametric tests

Parametric tests: rely on assumptions about shape of underlying distribution (like normality, or large sample sizes and CLT). Computes tests based on things like mean.

Nonparametric tests: rely on fewer assumptions. Computes tests based on things like median

Non-parametric version of the one-sample t-test we just did - the 'sign test'



Null hypothesis: the sample median is not significantly different from the reference median.

The test statistic = number of samples greater than the null median.

Calculate test statistic

One-sample *t*-test

$t = (\bar{X} - M)/s_{\bar{X}}$

$= (0.72 - 0.40)/0.11$

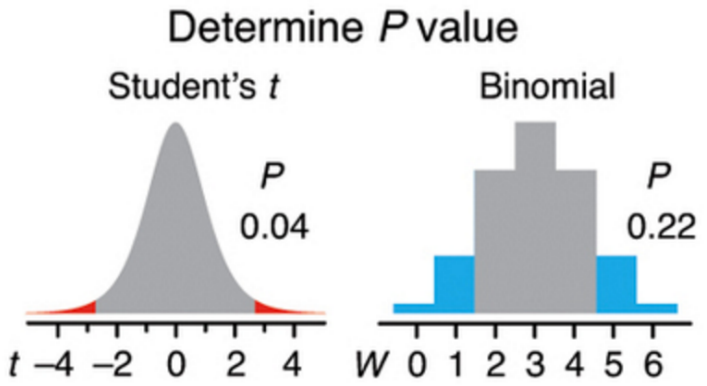
$= 2.84$

Sign test

$W = \text{count}(X_i > M)$

$= 5$

Under the null hypothesis, we would expect to see just as many above and below the median, so this statistic follows a binomial distribution.



The p-value of the sign test is much higher than the t-test p-value… non-parameteric tests are less sensitive.

Only measures how many are above – need more info to rule out null hypothesis (95% as efficient)

Parametric versus non-parametric tests

For just about every parametric test, there is a similar test of a non-parameteric flavor

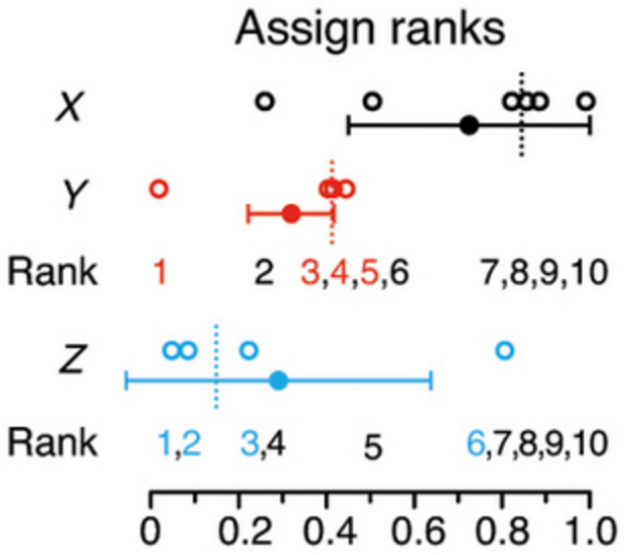
Analysis	Parametric procedure	Nonparametric procedure
Compare means between groups	Two-sample t-test	Wilcoxon rank-sum test
Compare paired means between groups	Paired t-test	Wilcoxon signed-rank test
Compare means between 3+ groups	ANOVA	Kruskal-Wallis test
Estimate degree of association between variables	Pearson coefficient of correlation	Spearman's rank correlation

http://www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf

Wilcoxon rank-sum test

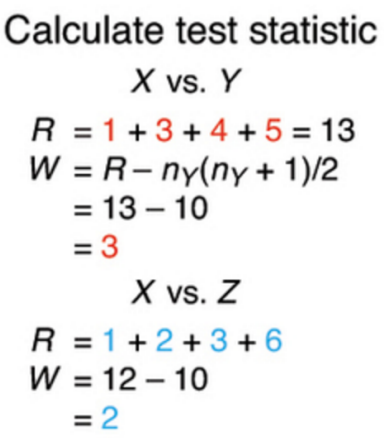
To compare two samples, Wilcoxon rank-sum test assumes that the samples come from distributions of the same shape (doesn't have to be normal)

To compare the medians of two samples:

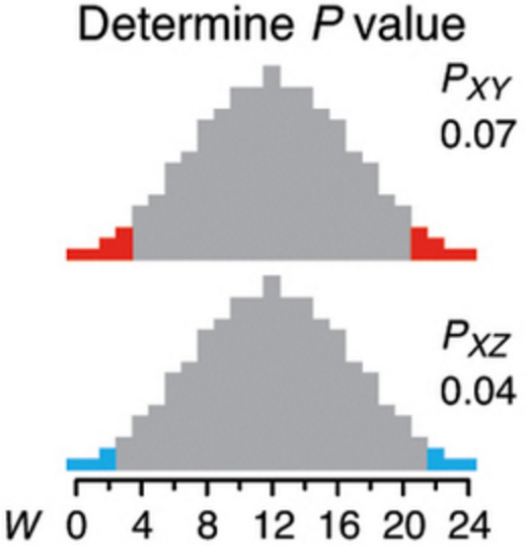


Step 1: Assign ranks to the pooled data

A smaller median = a smaller set of ranks



Step 2: Compute test statistic (W), which is the degree to which the sum of ranks is larger than the lowest possible in the sample with the lower ranks



Step 3: Compare W to the null distribution of all possible combinations of rank-orders

(For small samples, hard to reach small p-values)

(For large samples, this distribution becomes normal, and we can use the z-test)

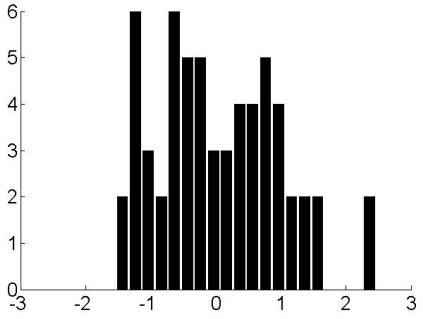
A lot of these methods have a lot of assumptions, and assume that we have quite a bit of data…

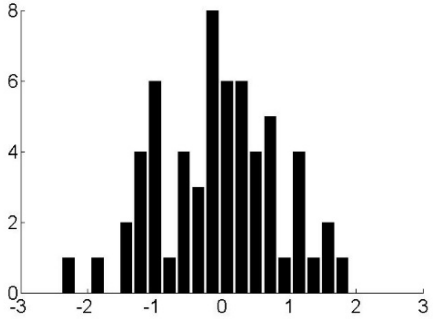
… what if that isn't the case?

Non-parametric tests: permutation test

Randomization (or permutation) tests:

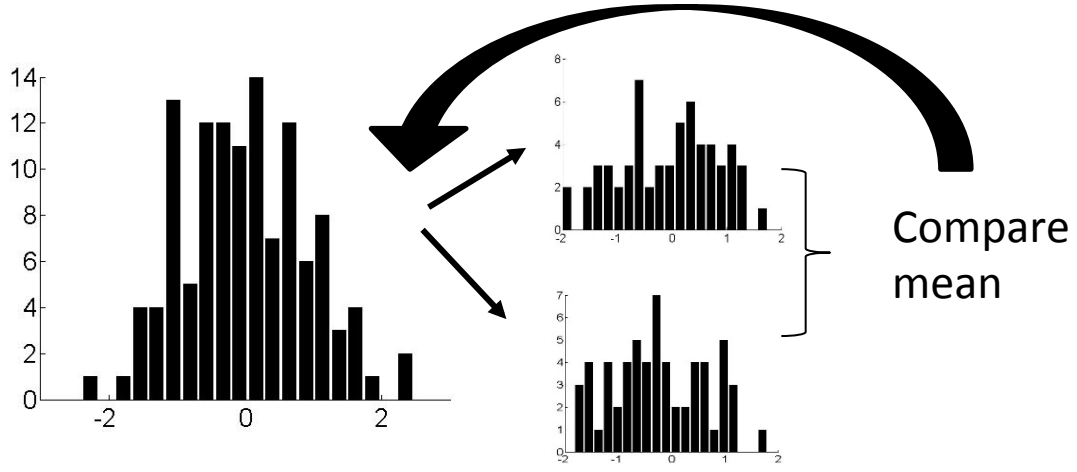
Consider two sets of data:

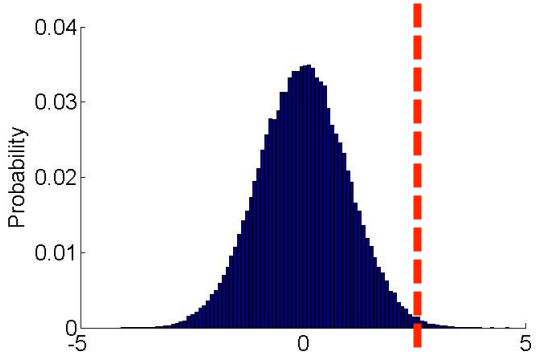




Null hypothesis: two sets of data from same distribution (or have the same mean)

If we aggregate data, randomly split it, and then compare the differences in the mean, we will get a null distribution of mean differences that we can compare to



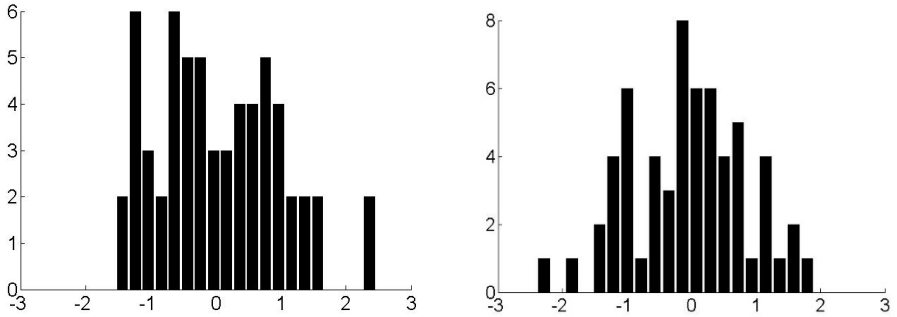


http://white.stanford.edu/~knk/Psych216A/Psych216ALecture2.pdf

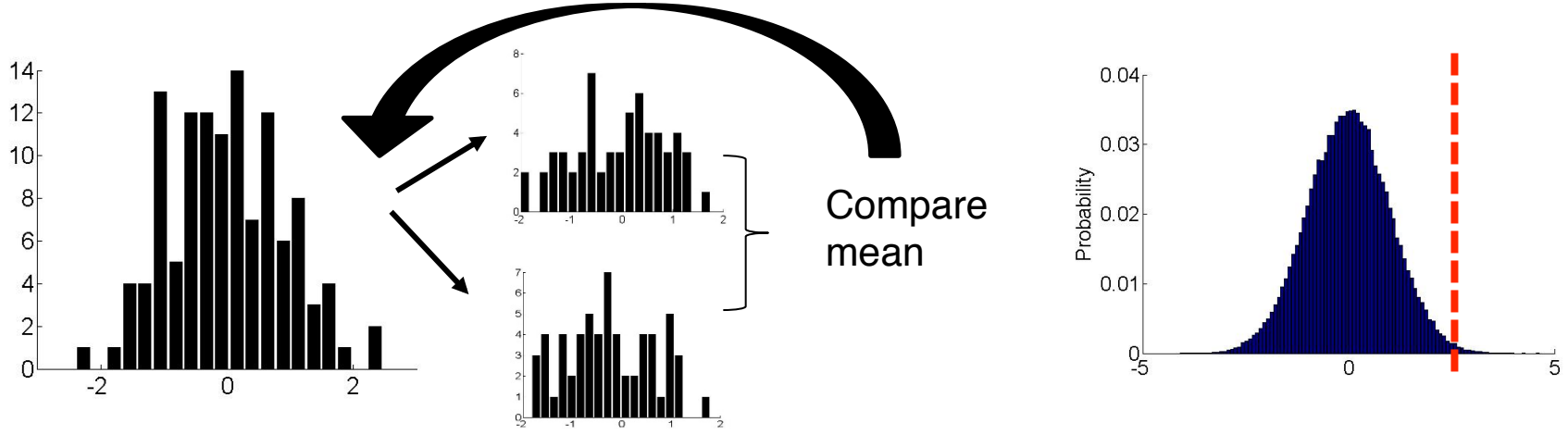
Non-parametric tests: bootstrapping

Bootstrapping tests:

Same data sets:



Same procedure, but sample **with replacement**!



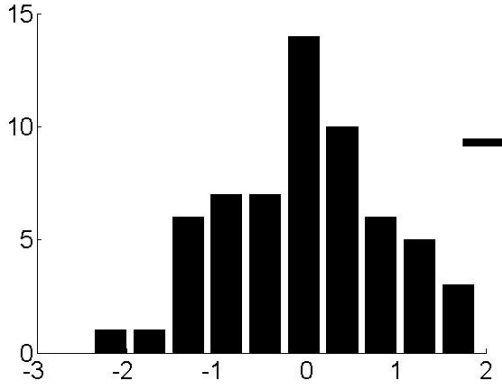
http://white.stanford.edu/~knk/Psych216A/Psych216ALecture2.pdf

Non-parametric tests: bootstrapping

Another common method using bootstrapping is to estimate distribution of the mean.

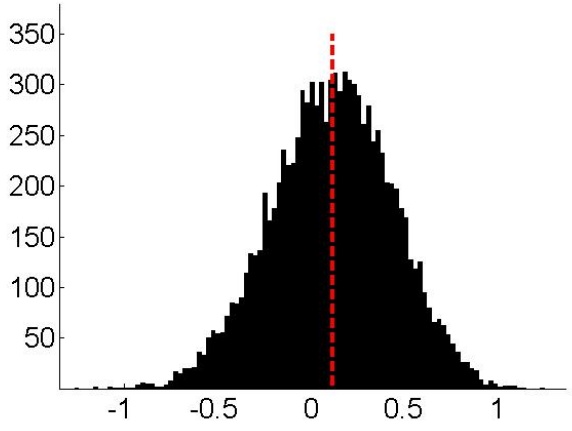
We mentioned this briefly earlier when discussing generating a distribution of the sample mean

Sample population:



Resample and compute mean 10,000 times:

Distribution of sample mean



Can use this method to determine confidence interval on estimate of the mean

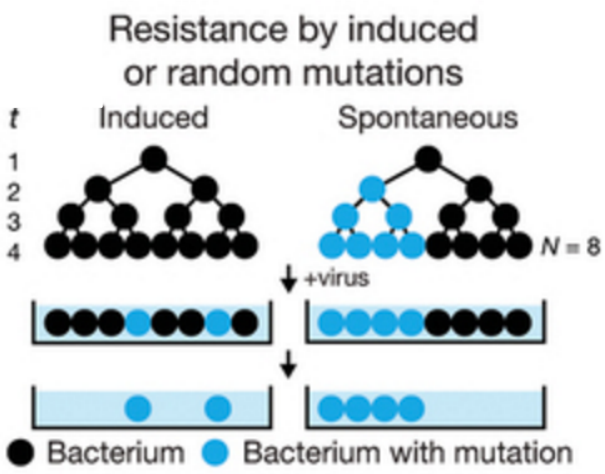
With two populations, can generate the distributions of the sample mean for each population and then look for overlap

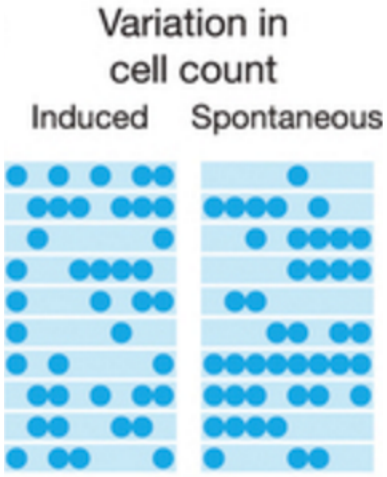
http://www.stat.rutgers.edu/home/mxie/rcpapers/bootstrap.pdf

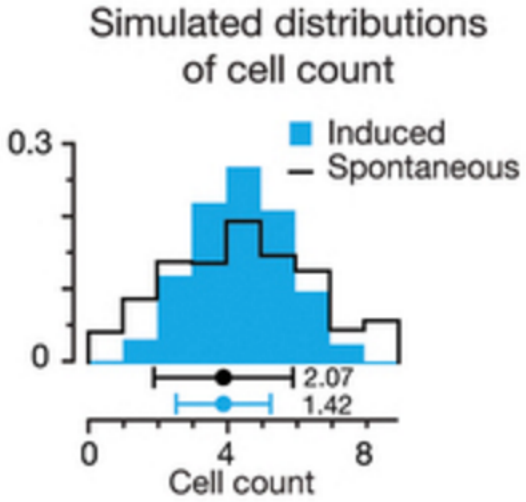
Bootstrapping

Sometimes, figuring out the shape of the sample statistic is difficult – but bootstrapping can help!

An example from the literature:







Investigating mutations that confer viral resistance onto bacteria

Question: are mutations spontaneous or induced by exposure to a virus?

To answer: grow bacteria, plate it onto a medium with virus, and then look at how many bacteria survive

Repeat this experiment many times

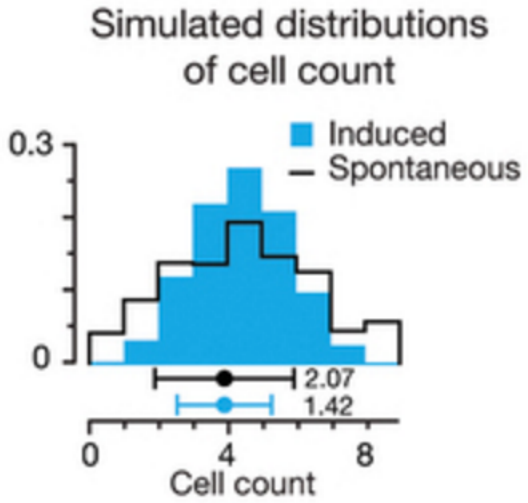
Look at difference in distribution

Variance will be way higher if the mutation is spontaneous

(Virus infection bacteria is Poisson-distributed)

http://www.nature.com/nmeth/journal/v12/n6/full/nmeth.3414.html

Bootstrapping



Compute metric: VMR (variance-mean ratio)

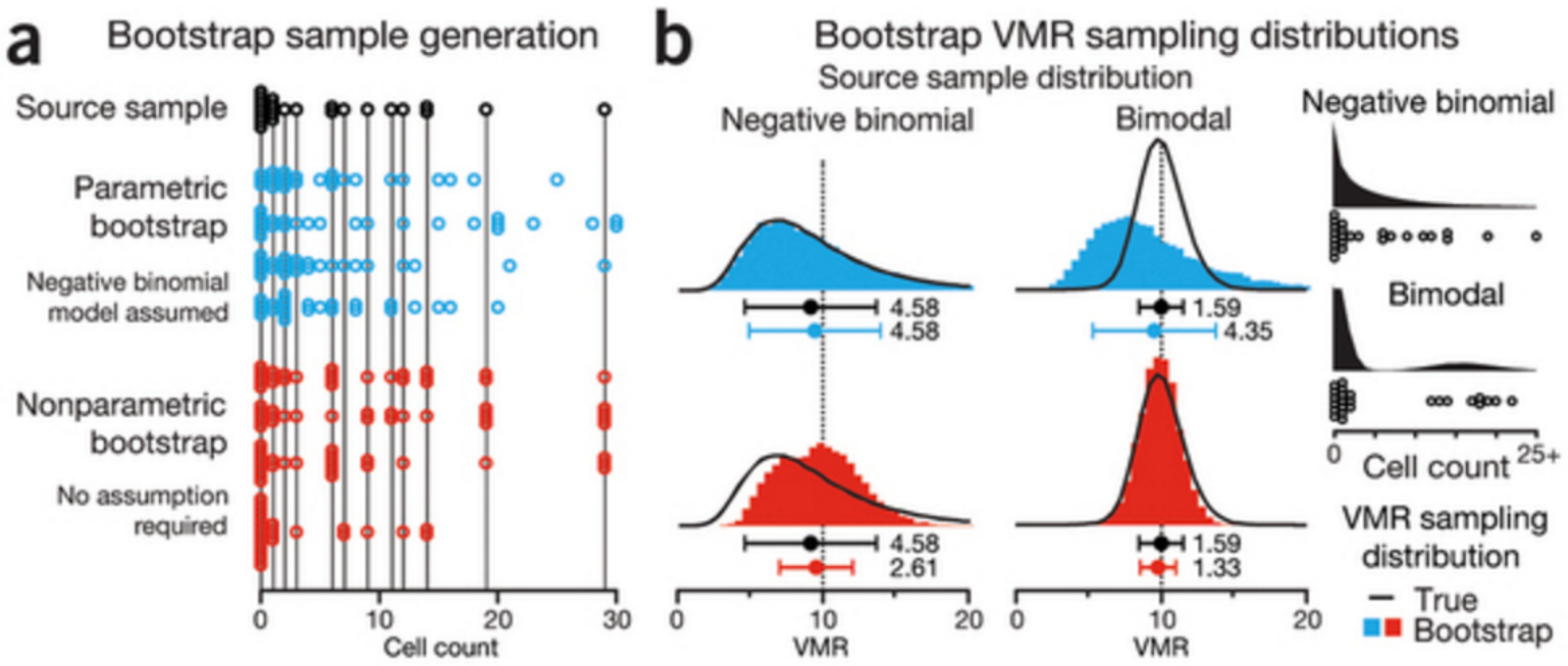
VMR = 1 for Poisson

Question – is VMR greater than 1?

Back in the day, have to fill out the VMR distribution with LOTS of experiments

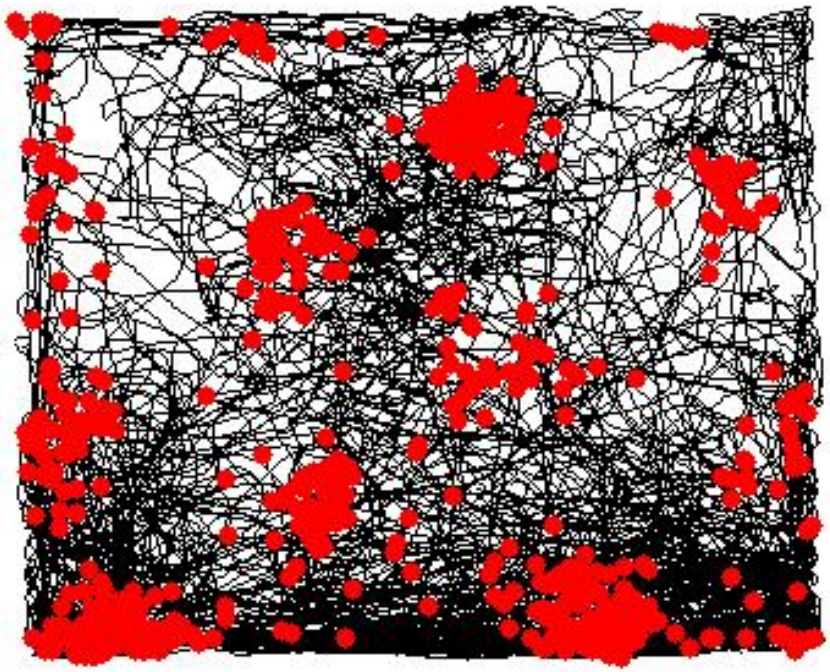
Now – can use computational approaches with modest sample sizes to estimate noise in our VMR estimate

Re-sample cell counts (with replacement), and then re-compute VMR



Shuffling tests

How likely is this to occur by chance?



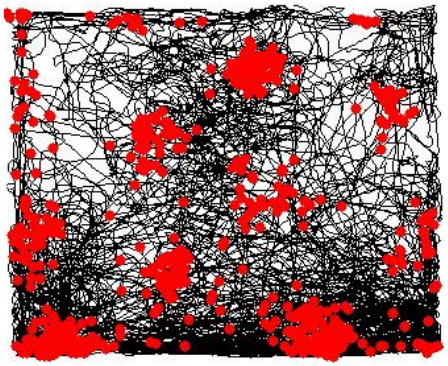
Ask this question while preserving spiking dynamics and behavior of the animal

Langston et al., 2010. Science.

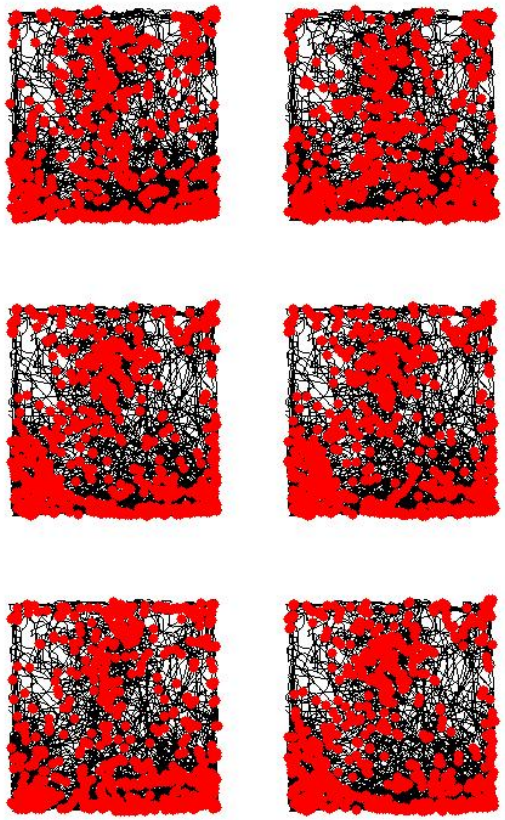
Shuffling tests

Shuffling tests:

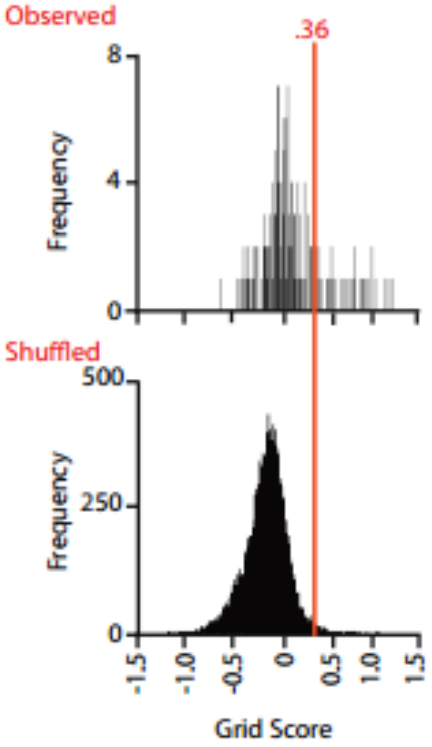
Quantify the “grid score”:
degree of 60 degree symmetry



Randomly shift spike train along the path of the animal and re-compute the grid score



Analyze null distribution of grid scores

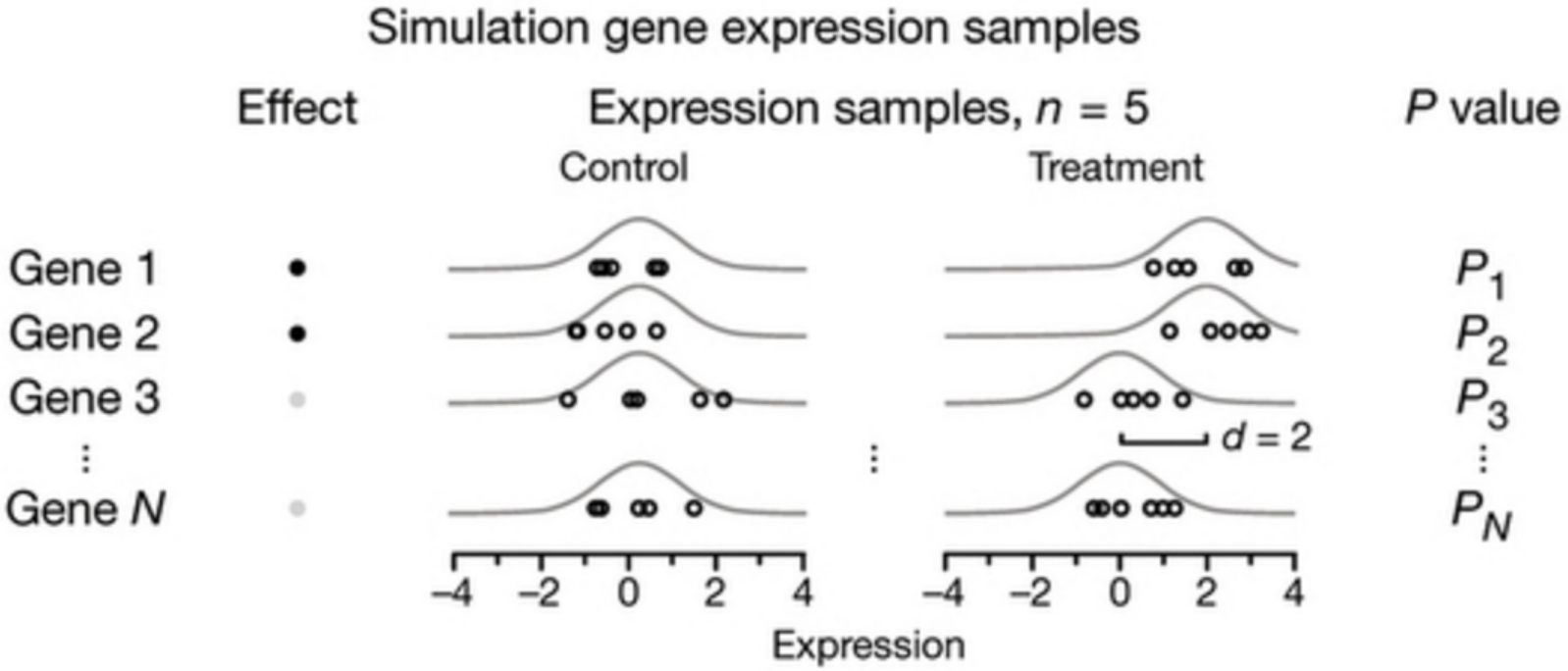


Giocomo et al., 2011. Cell.

So far, we have only dealt with testing one hypothesis at a time…

… but what if you have multiple?

Multiple comparisons



Compare effect of treatment for 10,000 genes
With a rejection level of 0.05, 500 genes could be incorrectly associated

If there is a 10% chance of observing an effect, there is a 1/3 of the ‘discoveries’ will be false

http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2900.html

Multiple comparisons: Bonferroni correction

One easy thing to do:

Bonferroni correction! Divide significance level by the number of tests

Divide by number
of genes tested

This will keep the proportion of false positives low (because the p-value will be low)

But - it does so at the expense of false negatives – so Bonferroni will be bad if the number of comparisons that you make is really large

For example, if you do 10,000 tests, then your p-value after Bonferroni correction will be p = 0.05/10,000, and only the strongest of strong effect will exhibit a p-value that large.

Multiple comparisons: BH procedure

What to do if you want to cut down on the rate of false negatives?

Benjamini Hochberg (BH) procedure

Procedure:

1. Rank-order p-values from smallest to largest. Then, the first p-value has rank 1, the second has rank 2, etc.

2. Define the BH critical value: (i/m)*Q, where i is the rank, m is the number of tests, and Q is the false discovery rate (false positives over all positives).

3. Then, find the largest p-value for which p < (i/m)*Q. All p-values less than this are considered significant.

Dietary variable	*P* value	Rank	(i/m)Q
Total calories	<0.001	1	0.010
Olive oil	0.008	2	0.020
Whole milk	0.039	3	0.030
White meat	0.041	4	0.040
Proteins	0.042	5	0.050
Nuts	0.060	6	0.060
Cereals and pasta	0.074	7	0.070
White fish	0.205	8	0.080
Butter	0.212	9	0.090
Vegetables	0.216	10	0.100
Skimmed milk	0.222	11	0.110
Red meat	0.251	12	0.120
Fruit	0.269	13	0.130
Eggs	0.275	14	0.140
Blue fish	0.34	15	0.150
Legumes	0.341	16	0.160
Carbohydrates	0.384	17	0.170
Potatoes	0.569	18	0.180
Bread	0.594	19	0.190
Fats	0.696	20	0.200
Sweets	0.762	21	0.210
Dairy products	0.94	22	0.220
Semi-skimmed milk	0.942	23	0.230
Total meat	0.975	24	0.240
Processed meat	0.986	25	0.250

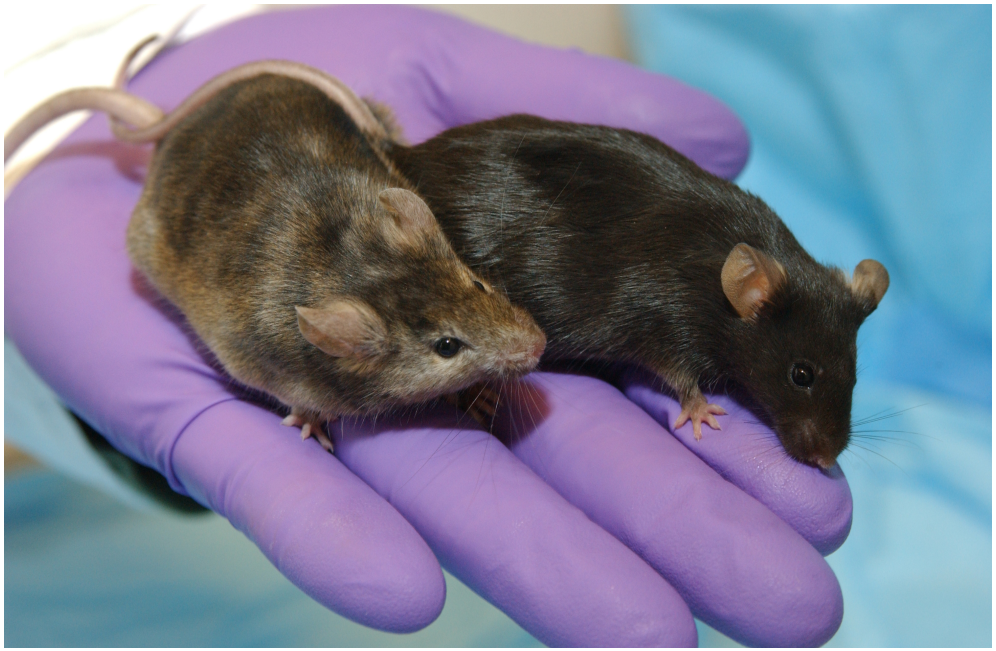
Larger picture with multiple comparisons

You make a knockout mouse.
You expect that this knockout mouse will show a specific phenotype: brain tumors.

Being a good and thorough graduate student, you test a bunch of things– food consumption, weight, sleep, cognitive tests, etc., but only find differences in brain tumors.

After correcting for multiple comparisons, the effect "goes away."

What do you do?



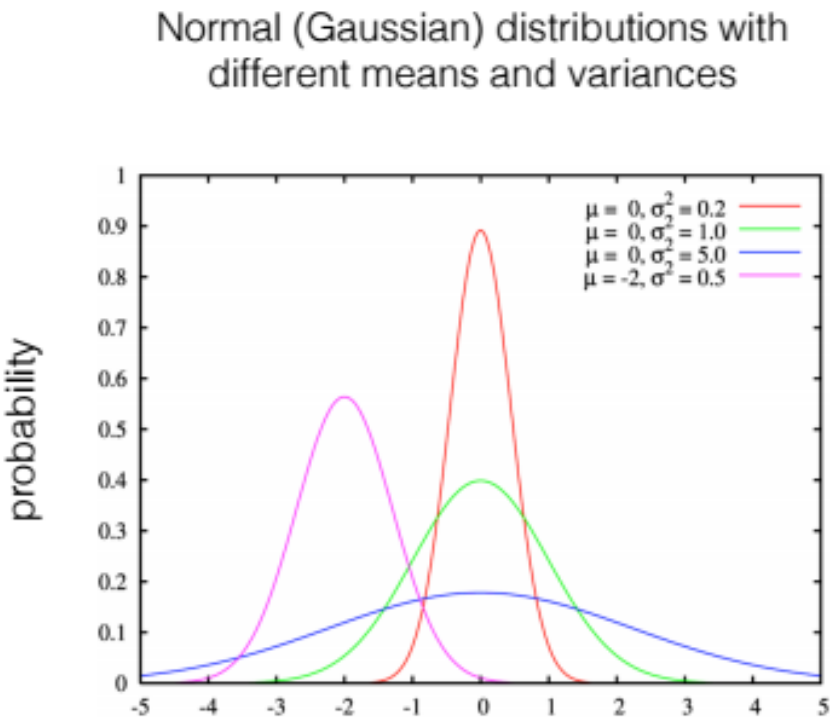
http://www.biostathandbook.com/multiplecomparisons.html

So far, we have assumed that the samples within a group are independent…

… but what if they aren’t?

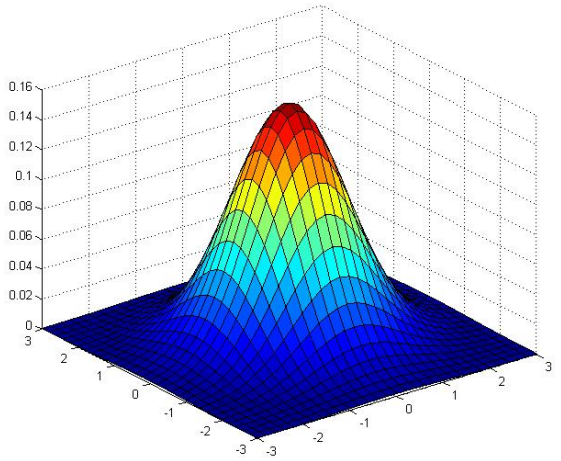
Dependent variables

Probability distribution for 1 variable:

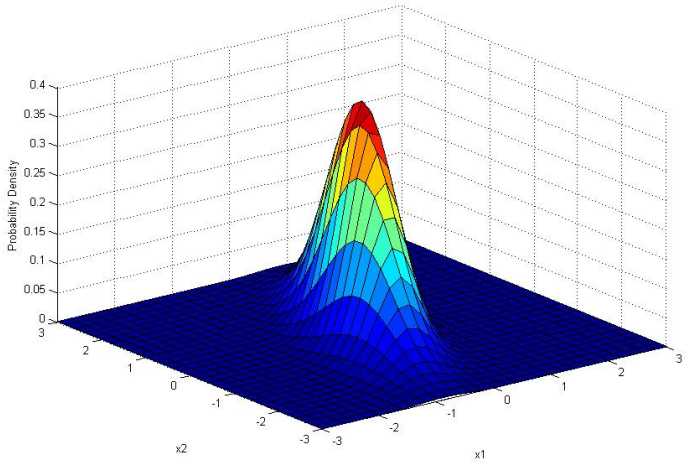


Probability distribution for 2 variables:

X and Y are **independent**:



X and Y are **dependent**:
If you know Y, that will change the probability distribution for X



Dependent variables

Conditional probability: $$\mathrm{P}(X|Y) = \frac{\mathrm{P}(X \cap Y)}{\mathrm{P}(Y)}$$

Events/variables are independent if: $$\mathrm{P}(X \cap Y) = \mathrm{P}(X)\mathrm{P}(Y)$$

$$\mathrm{P}(X|Y) = \mathrm{P}(X)$$

Examples:

Independent spike trains:

Dependent spike trains:

Probability of drawing a red card after black card from the deck

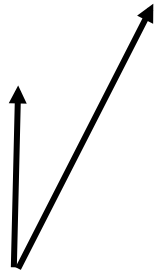
Dependent variables: correlation and covariance

Describe linear relationship between two variables

Covariance: how much two random variables change together

$$\operatorname{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

This is the dot product:



Can be rewritten as: $\operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y]$

Uncorrelated if: $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$

Correlation: Normalized version of covariance:

$$r = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

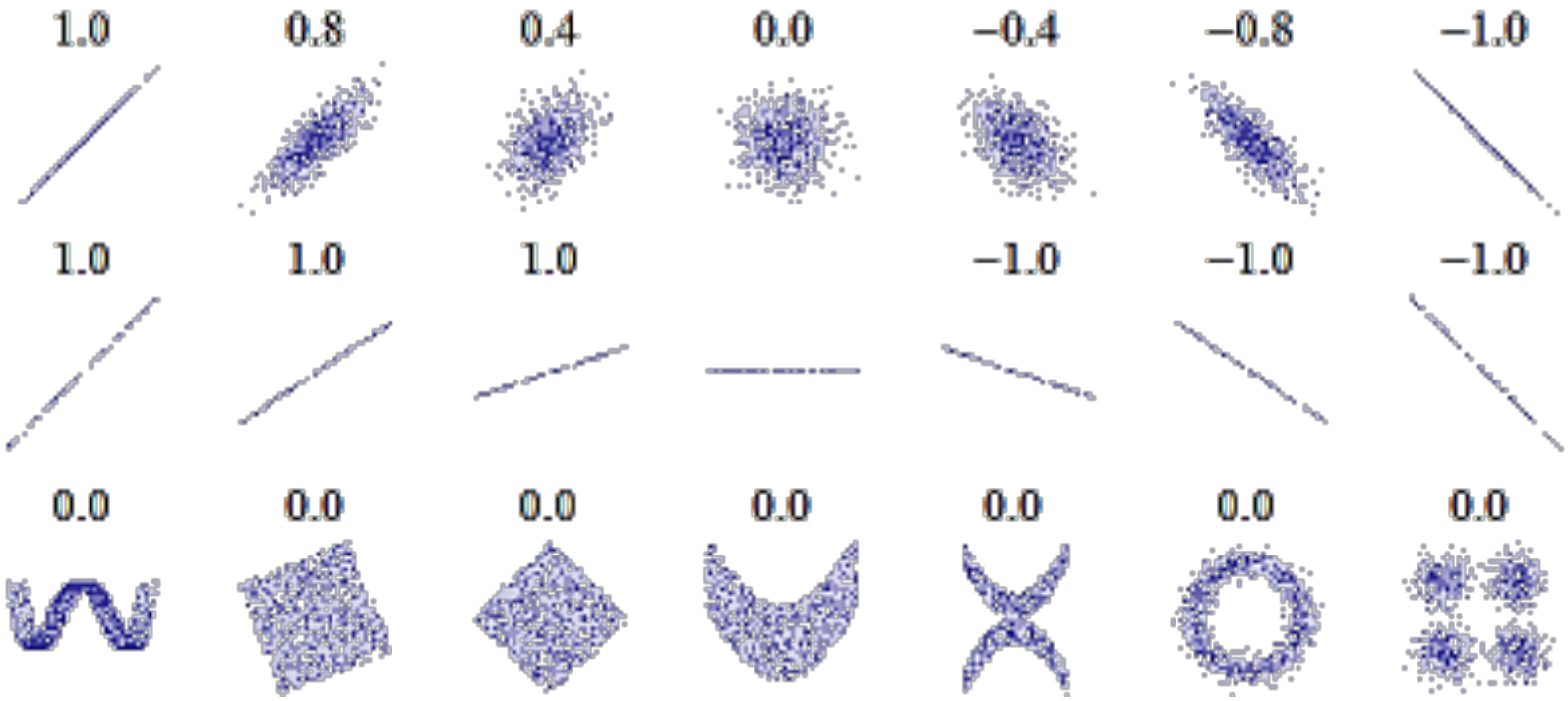
Compare unit vectors (cosine of angle between vectors:

Another way to say it: z-score each variable (subtract off mean and normalize by the standard deviation) and then compare their means

MATLAB: cov(data)
corrcoeff(data)

R-value goes between -1 and 1
R-squared goes between 0 and 1, gives fraction of variance explained

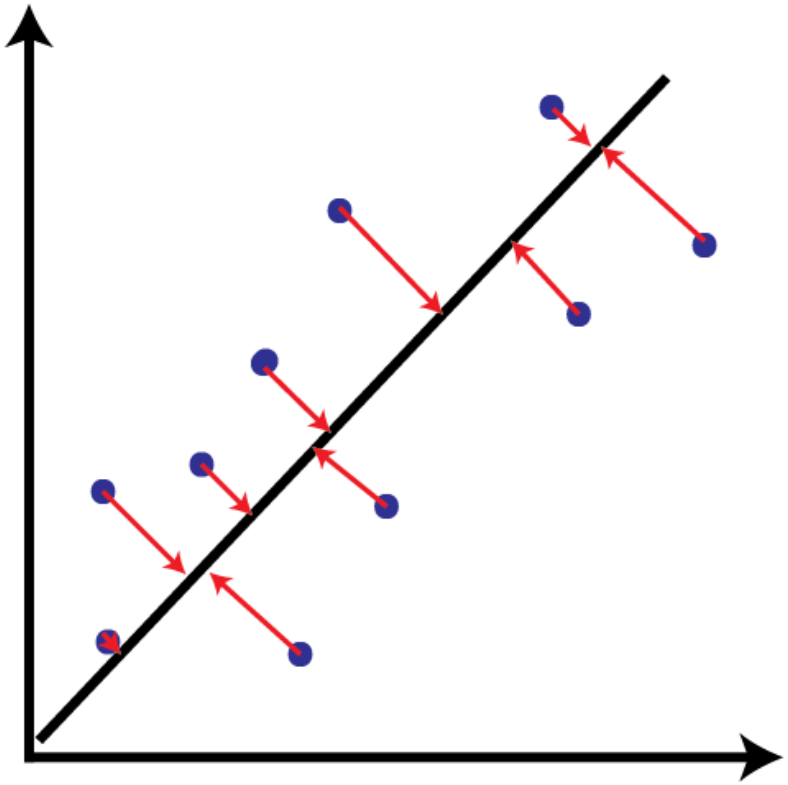
Correlation coefficient



If your data looks like the bottom row, use another method!
One option: mutual information (next lecture!)

Relationship between correlation and linear fit

Remember linear regression:



Square of correlation coefficient (coefficient of determination) is also the variance explained when doing linear regression!

MATLAB: polyfit(data(:,1),data(:,2),1))

Game Show Problem

(This material in this article was originally published in PARADE magazine in 1990 and 1991.)

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?
Craig F. Whitaker
Columbia, Maryland

Yes; you should switch. The first door has a 1/3 chance of winning, but the second door has a 2/3 chance. Here's a good way to visualize what happened. Suppose there are a million doors, and you pick door #1. Then the host, who knows what's behind the doors and will always avoid the one with the prize, opens them all except door #777,777. You'd switch to that door pretty fast, wouldn't you?

You blew it, and you blew it big! Since you seem to have difficulty grasping the basic principle at work here, I'll explain. After the host reveals a goat, you now have a one-in-two chance of being correct. Whether you change your selection or not, the odds are the same. There is enough mathematical illiteracy in this country, and we don't need the world's highest IQ propagating more. Shame!
Scott Smith, Ph.D.
University of Florida

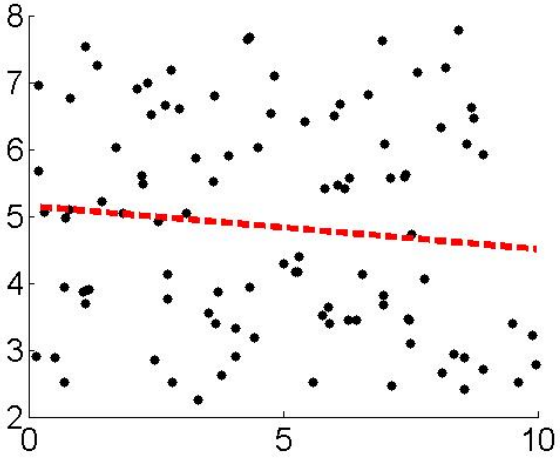
May I suggest that you obtain and refer to a standard textbook on probability before you try to answer a question of this type again?
Charles Reid, Ph.D.
University of Florida

You made a mistake, but look at the positive side. If all those Ph.D.'s were wrong, the country would be in some very serious trouble.
Everett Harman, Ph.D.
U.S. Army Research Institute

	DOOR 1	DOOR 2	DOOR 3	RESULT
GAME 1	AUTO	GOAT	GOAT	Switch and you lose.
GAME 2	GOAT	AUTO	GOAT	Switch and you win.
GAME 3	GOAT	GOAT	AUTO	Switch and you win.
GAME 4	AUTO	GOAT	GOAT	Stay and you win.
GAME 5	GOAT	AUTO	GOAT	Stay and you lose.
GAME 6	GOAT	GOAT	AUTO	Stay and you lose.

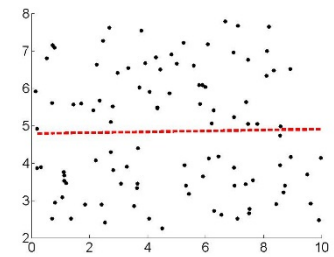
Shuffling + linear regression

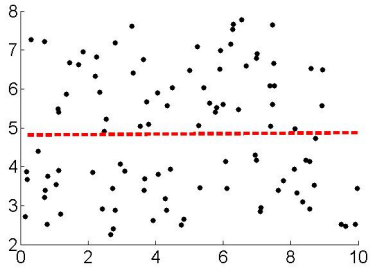
Is this slope significantly different from 0?

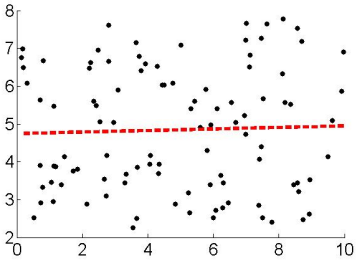


Analyze distribution of slope values!

Create null distribution of slopes by shuffling points along x-axis:

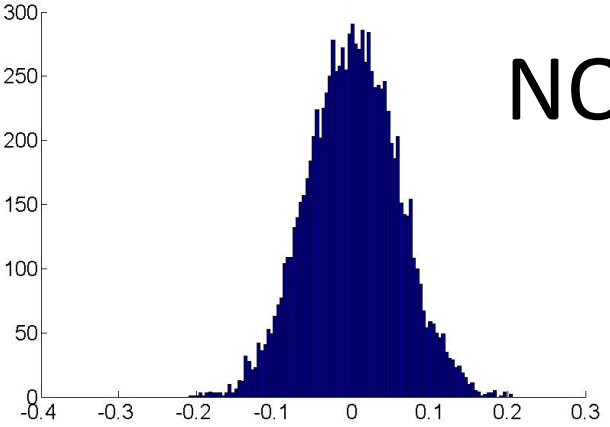






X 10,000

Null distribution

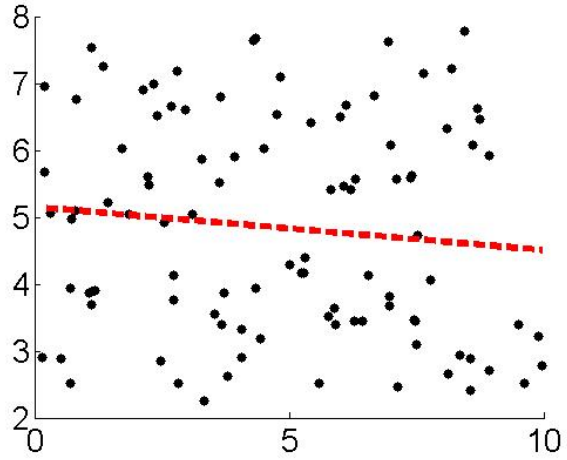


NO

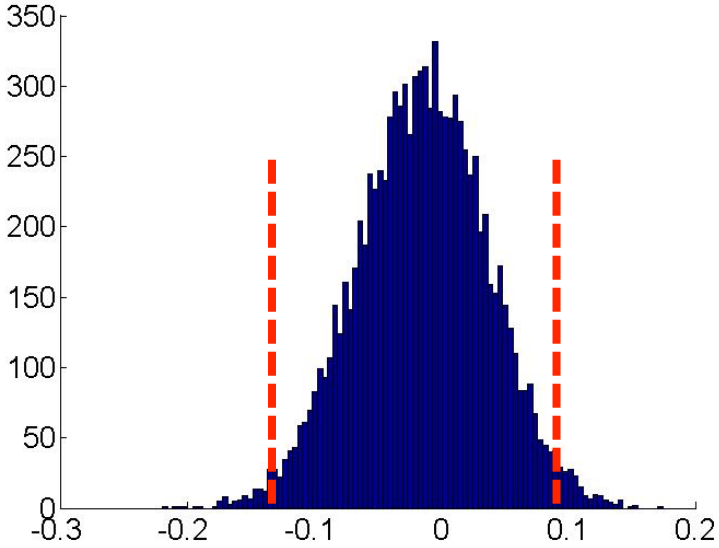
MATLAB: randperm(numel(data(:,1)))

Bootstrap + linear regression

Obtain a confidence interval on the slope



Resample from population and calculate the slope for each bootstrap iteration:



MATLAB: datasample(data(:,1),n)