

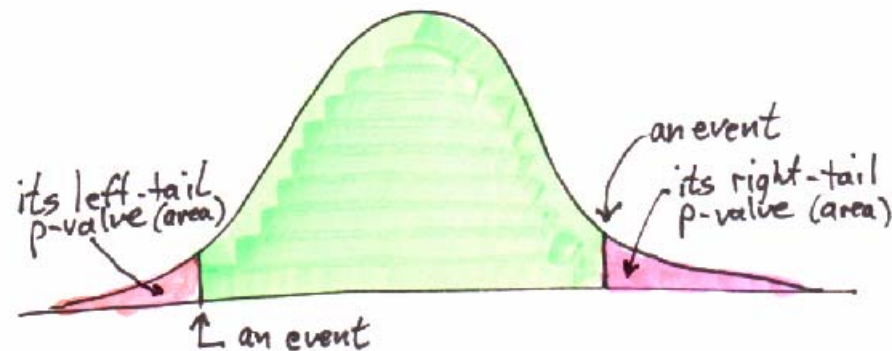
CS395T
Computational Statistics with
Application to Bioinformatics

Prof. William H. Press
Spring Term, 2009
The University of Texas at Austin

Unit 4: Distributions, P-values, CLT

The idea of p-value (tail) tests is to see how extreme is the observed data relative to the distribution of hypothetical repeats of the experiment under some “null hypothesis” H_0 .

If the observed data is too extreme, the null hypothesis is disproved. (It can never be proved.)



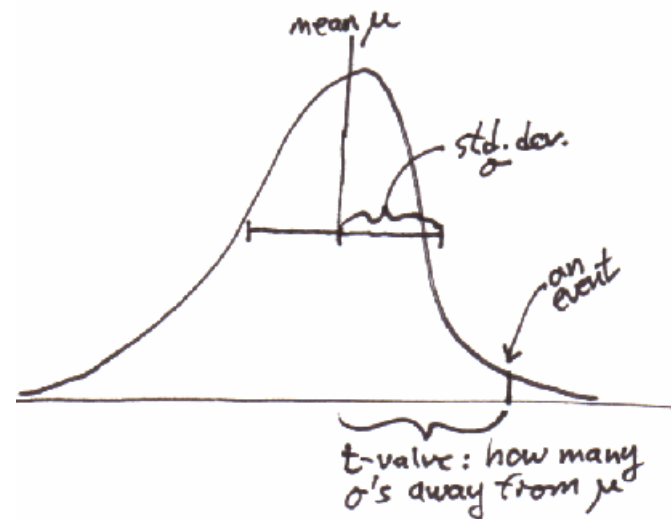
If the null hypothesis is true, then p-values are uniformly distributed in $(0,1)$, in principle exactly so.

There are some fishy aspects of tail tests, which we discuss later, but they have one big advantage over Bayesian methods: You don't have to enumerate all the alternative hypotheses (“the unknown unknowns”).



Don't confuse p-values with t-values (also sometimes named "Student")

t-value = number of standard deviations from the mean



Intentionally drawn
unsymmetric, not
just sloppy drawing!



It's much easier to compute are scores that depend only on the mean and standard deviation of the expected distribution. But, in general, these are interpretable as "likely" or "unlikely" only relative to a Gaussian (which may or may not be relevant). Often we are in an asymptotic regime where distributions are close to Gaussian. But beware of t-values if not!

The reason that Gaussian's often **are** relevant is, of course, the Central Limit Theorem, which we will now discuss.

Important to understand the Central Limit Theorem and where it might or might not apply.

The characteristic function of a distribution is its Fourier transform.

$$\phi_X(t) \equiv \int_{-\infty}^{\infty} e^{itx} p_X(x) dx$$

(Statisticians often use notational convention that X is a random variable, x its value, $p_X(x)$ its distribution.)

$$\phi_X(0) = 1$$

$$\phi'_X(0) = \int ixp_X(x) dx = i\mu$$

$$-\phi''_X(0) = \int x^2 p_X(x) dx = \sigma^2 + \mu^2$$

So, the coefficients of the Taylor series expansion of the characteristic function are the (uncentered) moments.

Addition of independent r.v.'s:

$$\text{let } S = X + Y$$

$$p_S(s) = \int p_X(u)p_Y(s - u)du$$

$$\phi_S(t) = \phi_X(t)\phi_Y(t)$$

(Fourier convolution theorem.)

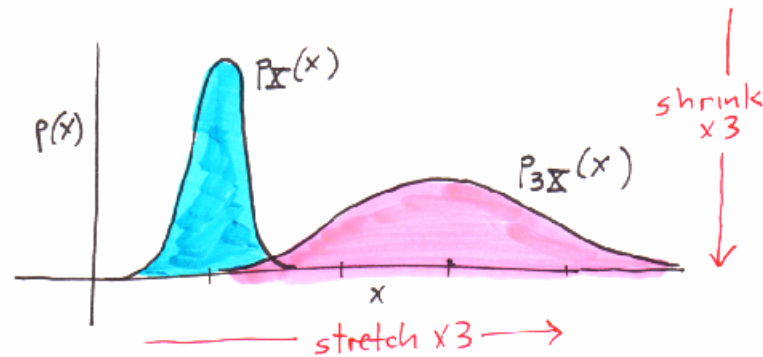
Proof of convolution theorem:

$$\left. \begin{aligned} \phi_X(t) &\equiv \int_{-\infty}^{\infty} e^{itx} p_X(x) dx \\ p_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(t) e^{-itx} dt \end{aligned} \right\} \text{Fourier transform pair}$$

$$\begin{aligned} p_S(s) &= \int_{-\infty}^{\infty} p_X(u) p_Y(s-u) du \\ &= \int_{-\infty}^{\infty} p_X(u) \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_Y(t) e^{-it(s-u)} dt \right] du \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_Y(t) e^{-its} \left[\int_{-\infty}^{\infty} p_X(u) e^{itu} du \right] dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_Y(t) \phi_X(t) e^{-its} dt \end{aligned}$$

So, $\phi_S(t) = \phi_Y(t) \phi_X(t)$

Scaling law for r.v.'s:



Scaling law for characteristic functions:

$$\begin{aligned}\phi_{aX}(t) &= \int e^{itx} \underline{p_{aX}(x)} dx \\ &= \int e^{itx} \underline{\frac{1}{a} p_X\left(\frac{x}{a}\right)} dx \\ &= \int e^{i(at)(x/a)} p_X\left(\frac{x}{a}\right) \frac{dx}{a} \\ &= \phi_X(at)\end{aligned}$$

What's the characteristic function of a Gaussian?

```
syms x mu pi t sigma
p = exp(-(x-mu)^2 / (2*sigma^2)) / (sqrt(2*pi)*sigma)
p =
1/2*exp(-1/2*(x-mu)^2/sigma^2)*2^(1/2)/pi^(1/2)/sigma
norm = int(p, x, -Inf, Inf)
norm =
1
cf = simplify(int(p*exp(i*t*x), x, -Inf, Inf))
cf =
exp(1/2*i*t*(2*mu+i*t*sigma^2))
```

```
In[14]:= $Assumptions = $Assumptions && (sig > 0)
```

```
In[15]:=
```

```
p = (1 / (Sqrt[2 Pi] sig)) Exp[-(1 / 2) ((x - mu) / sig) ^2]
```

```
Out[15]=
```

$$\frac{e^{-\frac{(-\mu+x)^2}{2 \text{sig}^2}}}{\sqrt{2 \pi} \text{sig}}$$

```
In[16]:= Integrate[p, {x, -Infinity, Infinity}]
```

```
Out[16]=
```

1

```
In[17]:= Integrate[p Exp[I t x], {x, -Infinity, Infinity}]
```

```
Out[17]=
```

$$e^{i \mu t - \frac{\text{sig}^2 t^2}{2}}$$

Tell Mathematica that sig is positive.
Otherwise it gives "cases" when taking
the square root of sig^2

Cauchy distribution has ill-defined mean and infinite variance, but it has a perfectly good characteristic function:

$$x \sim \text{Cauchy}(\mu, \sigma), \quad \sigma > 0$$
$$p(x) = \frac{1}{\pi\sigma} \left(1 + \left[\frac{x - \mu}{\sigma} \right]^2 \right)^{-1}$$

Matlab and Mathematica both sadly fails at computing the characteristic function of the Cauchy distribution, but you can use fancier methods* and get:

$$\phi_{\text{Cauchy}}(t) = e^{i\mu t - \sigma|t|}$$

 note non-analytic at t=0

*If $t > 0$, close the contour in the upper $1/2$ -plane with a big semi-circle, which adds nothing. So the integral is just the residue at the pole $(x - \mu)/\sigma = i$, which gives $\exp(-\sigma t)$. Similarly, close the contour in the lower $1/2$ -plane for $t < 0$, giving $\exp(\sigma t)$. So answer is $\exp(-|\sigma t|)$. The factor $\exp(i\mu t)$ comes from the change of x variable to $x - \mu$.

Central Limit Theorem

$$\text{Let } S = \frac{1}{N} \sum X_i = \sum \frac{X_i}{N} \text{ with } \langle X_i \rangle \equiv 0$$

Can always subtract off the means, then add back later.

Then

$$\begin{aligned} \phi_S(t) &= \prod_i \phi_{X_i/N}(t) = \prod_i \phi_{X_i} \left(\frac{t}{N} \right) \\ &= \prod_i \left(1 - \frac{1}{2} \sigma_i^2 \frac{t^2}{N^2} + \dots \right) \quad \text{Whoa! It better have a convergent Taylor series around zero! (Cauchy doesn't, e.g.)} \\ &= \exp \left[\sum_i \ln \left(1 - \frac{1}{2} \sigma_i^2 \frac{t^2}{N^2} + \dots \right) \right] \\ &\approx \exp \left[-\frac{1}{2} \left(\frac{1}{N^2} \sum_i \sigma_i^2 \right) t^2 + \dots \right] \quad \text{These terms decrease with N, but how fast?} \end{aligned}$$

So, S is normally distributed

$$p_S(\cdot) \sim \text{Normal}(0, \frac{1}{N^2} \sum \sigma_i^2)$$

$$p_S(\cdot) \sim \text{Normal}(0, \frac{1}{N^2} \sum \sigma_i^2)$$

Moreover, since

$$NS = \sum X_i \quad \text{and} \quad \text{Var}(NS) = N^2 \text{Var}(S)$$

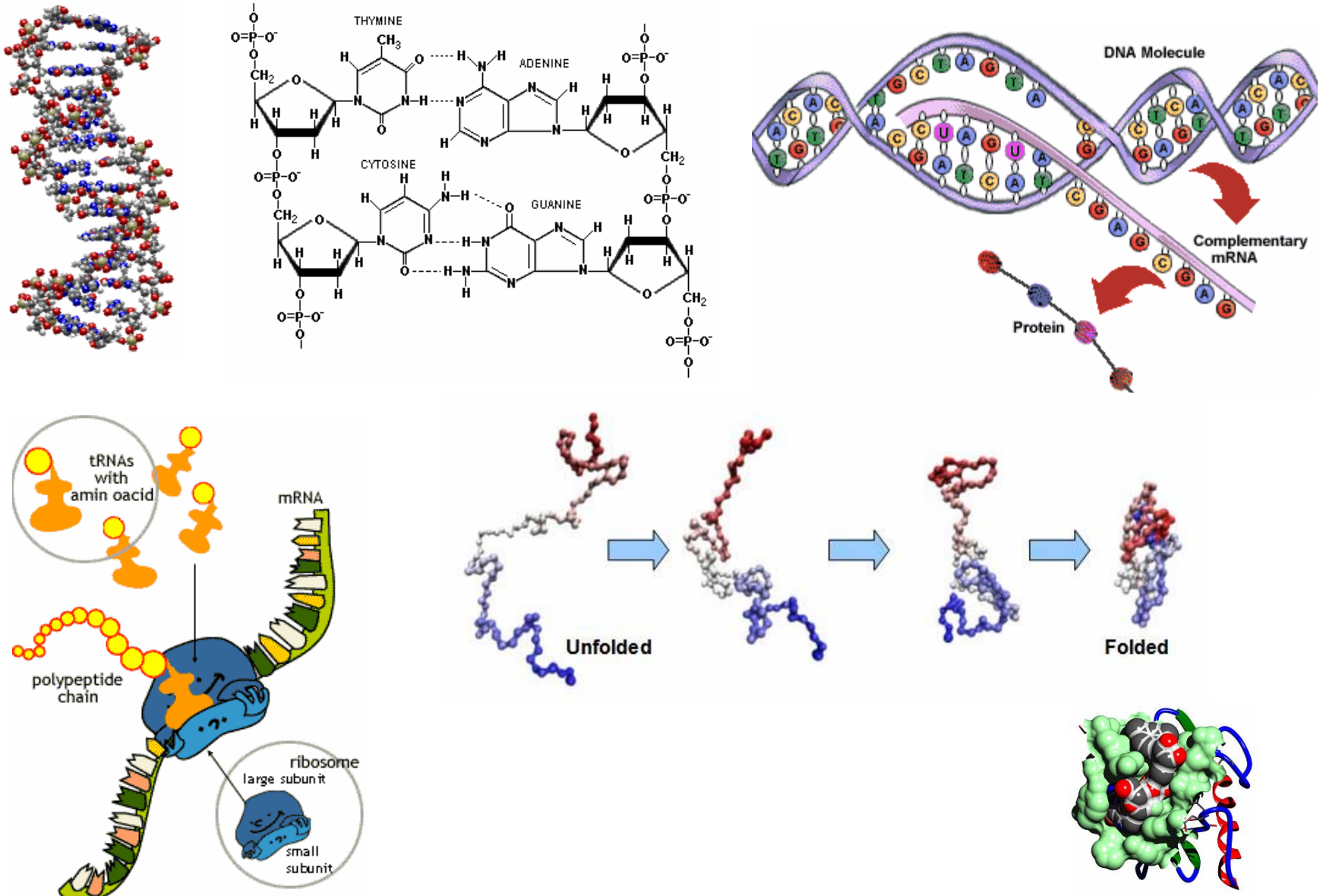
it follows that the simple sum of a large number of r.v.'s is normally distributed, with variance equal to the sum of the variances:

$$p_{\sum X_i}(\cdot) \sim \text{Normal}(0, \sum \sigma_i^2)$$

If N is large enough, and if the higher moments are well-enough behaved, and if the Taylor series expansion exists!

Also beware of borderline cases where the assumptions technically hold, but convergence to Normal is slow and/or highly nonuniform. (This can affect p-values for tail tests, as we will soon see.)

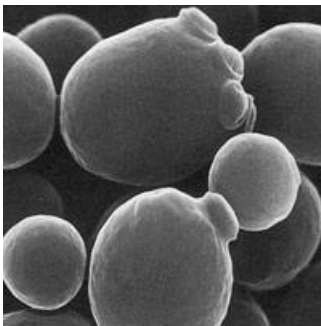
Quick review of all of modern molecular biology for those who missed it!



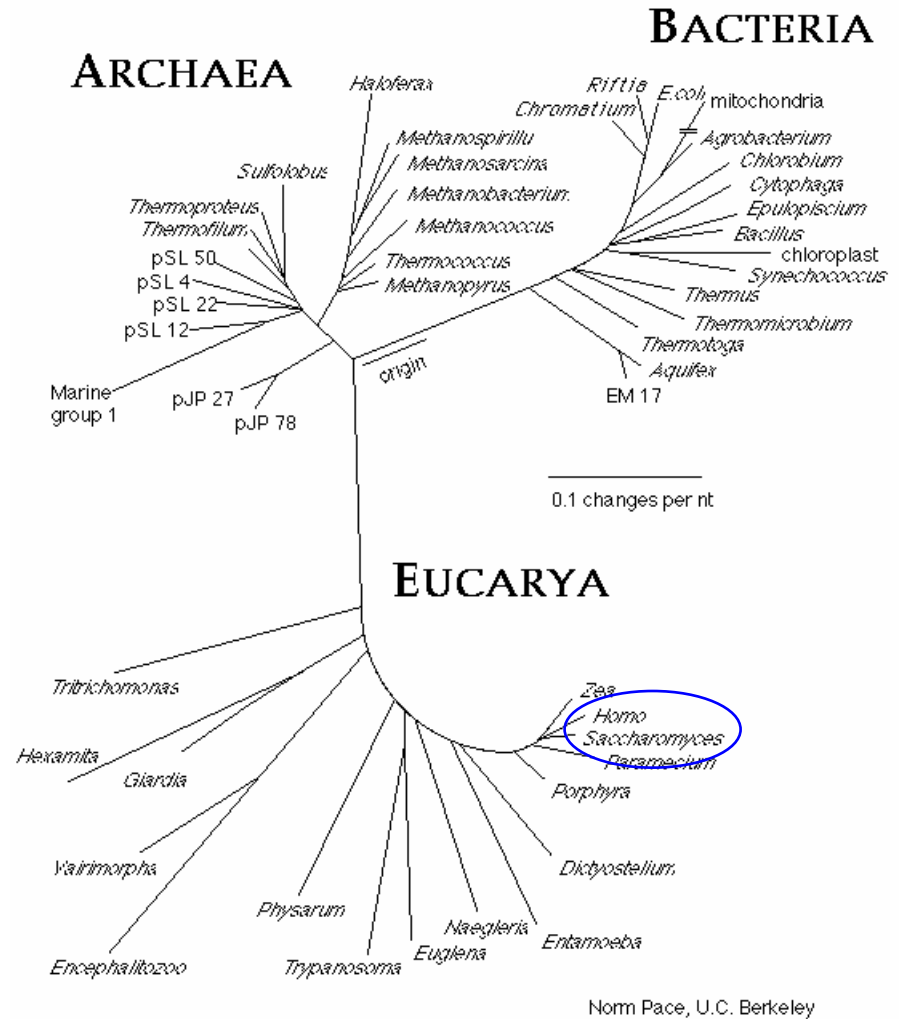
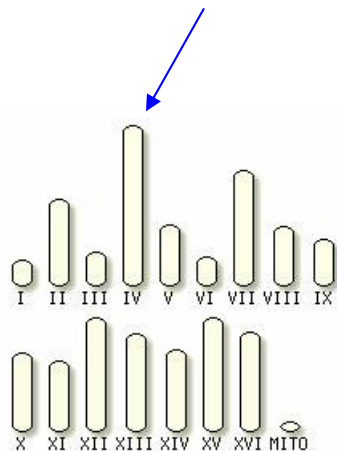
credits all web anon.

For practice with p- and t-values, let's look at the Sac cer genome.
 We'll use as a data set all of Chromosome 4.
 Yeast and Human are very close relatives in the great scheme of things.

Saccharomyces cerevisiae
 = *baker's yeast*



Chromosome 4:
 ACACCACACC...(1531894 omitted)...TAGCTTTTGG



Count nucleotides A,C,G,T on SacCer Chr4:

```
fi d = fopen(' SacSerChr4. txt' );
[chr, len] = fread(fi d, inf, ' int8' );
len
len =
    1531914
fclose(fi d);
chr(chr==65) = 3;
chr(chr==67) = 2;
chr(chr==71) = 4;
chr(chr==84) = 1;
count = accumarray(chr, 1, [4 1])
count =
    474471
    289341
    476750
    291352
pnuc = count ./ len
pnuc =
    0.3097
    0.1889
    0.3112
    0.1902
```

TCAG order (note: more
often we'll use ACGT)



Are these counts consistent with

$$p_A = p_C = p_G = p_T = 0.25 ?$$

(Of course not! But we'll check.)

Are they consistent with

$$p_A = p_T \approx 0.31 \quad p_C = p_G \approx 0.19 ?$$

That's a deeper question! You might think yes,
because of A-T and C-G base pairing.

The models are always binomial (sorting into bins by probabilities).

But the counts are all so large that the normal approximation is highly accurate:

$$\text{Bin}(n, p) \approx \text{Normal}(np, \sqrt{np(1-p)})$$

Let's dispose of the silly model (all 0.25):

`mu = 0.25*len;`

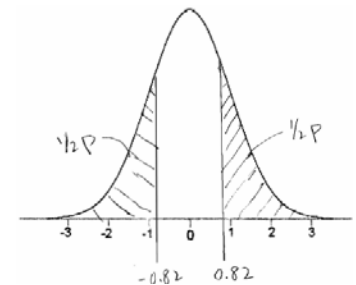
`sig = sqrt(0.25*0.75*len);`

`tval = (count-mu) ./ sig`

`pval = 2*(1-normcdf(abs(tval), 0, 1))`

t-value = number of standard deviations

p-value = tail probability (here, 2-tailed)



`tval =`

`170.7134`

`-174.7157`

`174.9657`

`-170.9634`

`pval =`

`0`

`0`

`0`

`0`

A technical aside :

CLT applies to binomial because it's sum of Bernoulli r.v.'s: N tries of an r.v. with values 1 (prob p) or 0 (prob $1-p$).

$$\mu = p \times 1 + (1-p) \times 0 = p$$

$$\sigma^2 = p \times (1-\mu)^2 + (1-p) \times (0-\mu)^2 = p(1-p)$$

The not-silly model:

```

dif = [count(1)-count(3); count(2)-count(4) ]
pdiff = [pnuc(1); pnuc(2)]
mu = [0; 0];
sig = sqrt(2 .* pdiff .* (1 - pdiff) .* len)
tval = (dif - mu) ./ sig
pval = 2*(1-normcdf(abs(tval), 0, 1))
    
```

the difference of two Normals is itself Normal

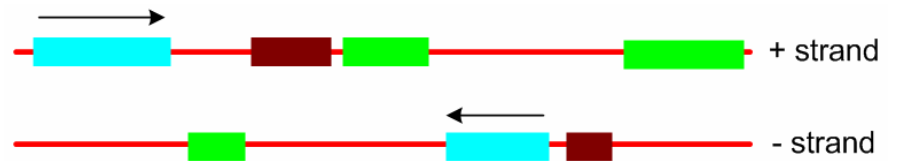
the variance of the sum (or difference) is the sum of the variances

```

dif =
    -2279
    -2011
pdiff =
    0.3097
    0.1889
mu =
    0
    0
sig =
    809.3402
    685.1154
tval =
    -2.8159
    -2.9353
pval =
    0.0049
    0.0033
    
```

Surprise!
The model is ruled out with high significance (small p-value)!

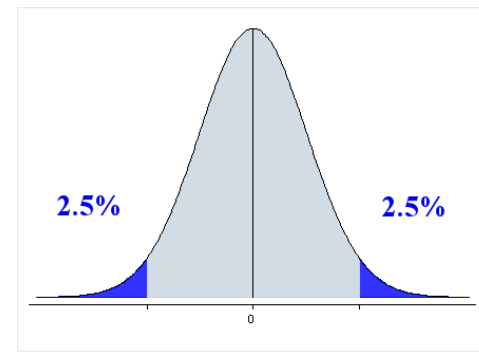
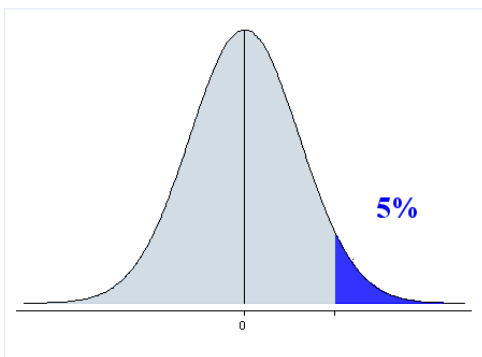
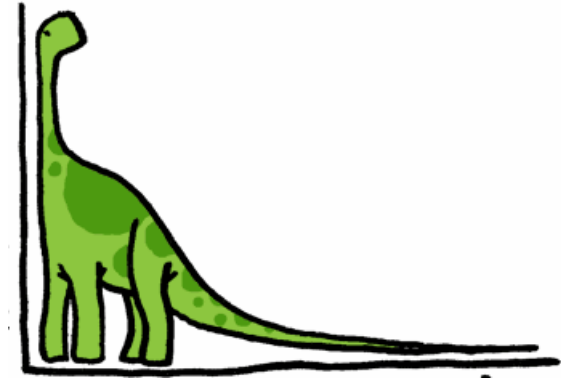
Why? Because, we're discovering genes!



The fluctuating “units” are indeed not single bases. Rather, they are genes which, individually, do not have A=T, C=G. Their placement on one strand or the other is random.

The classic p-value (or tail-) test terminology:

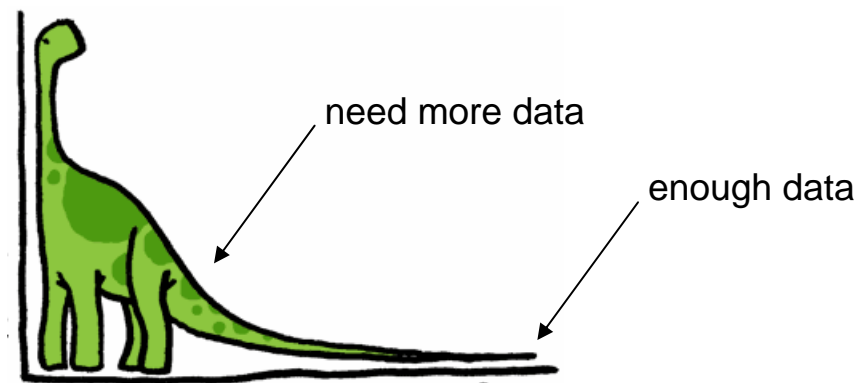
- “null hypothesis”
- “the statistic” (e.g., t-value or χ^2)
 - calculable for the null hypothesis
 - intuitively should be “deviation from” in some way
- “the critical region” α
 - biologists use 0.05
 - physicists use 0.0026 (3σ)
- one-sided or two?
 - somewhat subjective
 - use one-sided only when the other side has an understood and innocuous interpretation
- if the data is in the critical region, the null hypothesis is ruled out at the α significance level
- after seeing the data you
 - may adjust the significance level α
 - **may not try a different statistic**, because any statistic can rule out at the α level in $1/\alpha$ tries (shopping for a significant result!)
- if you decided **in advance** to try N tests, then the critical region for α significance is α/N (Bonferroni correction).



Tips on tail tests:

Don't sweat a p-value like 0.06. If you really need to know, the only real test is to get significantly more data. Rejection of the null hypothesis is exponential in the amount of data.

In principle, p-values from repeated tests s.b. exactly uniform in $(0,1)$. In practice, this is rarely true, because some "asymptotic" assumption will have crept in when you were not looking. All that really matters is that (true) extreme tail values are being computed with moderate fractional accuracy. You can go crazy trying to track down not-exact-uniformity in p-values. (I have!)



Here are two Bayesian criticisms of tail tests:

(1) Their result depends on the choice of test or (more argumentatively) what was in the mind of the experimenter

These are called “stopping rule paradoxes”.

Hypothesis H_0 : a coin is fair with $P(\text{heads})=0.5$

Data: in 10 flips, the first 9 are heads, then 1 tail.

Analysis Method I. Data this extreme, or more so, should occur under H_0 only

$$\frac{1 + 10 + 10 + 1}{2^{10}} = 0.0214$$

(you lose: referee wants $p < 0.01$ and tells you to get more data)



Analysis method II.

“I forgot to tell you,” says the experimenter, “my protocol was to flip until a tail and record $N (=9)$, the number of heads.”

$$\text{Under } H_0 \quad p(N) = 2^{-(N+1)}$$

$$p(\geq N) = 2^{-(N+1)} \left(1 + \frac{1}{2} + \frac{1}{4} + \dots\right) = 2^{-N}$$

$$P(\geq 9) = 2^{-9} = 0.00195$$

(Nature hold the presses!)

Stopping rule effects are a serious methodological issue in biomedical research, where for ethical reasons stopping criteria may depend on outcomes in complicated and unpredictable ways, or be ad hoc after the experiment starts (and rightly so – see next slide!)

April 8, 2006

British Rethinking Rules After Ill-Fated Drug Trial

By [ELISABETH ROSENTHAL](#),
International Herald Tribune

In February, when Rob O. saw the text message from Parexel International pop up on his cellphone in London — "healthy males needed for a drug trial" for £2,000, about \$3,500 — it seemed like a harmless opportunity to make some much-needed cash. Parexel, based in Waltham, Mass., contracts with drug makers to test new medicines.

Just weeks later, the previously healthy 31-year-old was in intensive care at London's Northwick Park Hospital — wires running directly into his heart and arteries, on dialysis, his immune system, liver, kidneys and lungs all failing — the victim of a drug trial gone disastrously bad.

One of six healthy young men to receive TGN1412, a novel type of immune stimulant that had never before been tried in humans, Rob O. took part in a study that is sending shock waves through the research world and causing regulators to rethink procedures for testing certain powerful new drugs.

Although tests of TGN1412 in monkeys showed no significant trouble, all six human subjects nearly died. One is still hospitalized and the others, though discharged, still have impaired immune systems, their future health uncertain.

On Wednesday, after releasing its interim report on the trial as well as previously confidential scientific documents that were part of the application for a trial permit, the British government announced it was convening an international panel of experts to "consider what necessary changes to clinical trials may be required" for such novel compounds.

The outcome "could potentially affect clinical trials regulation worldwide," the announcement said. In statements this week, both Parexel and **the drug's manufacturer, TeGenero, emphasized that they had complied with all regulatory requirements and conducted the trial according to the approved protocol.** But they declined to answer questions e-mailed to them about the specifics of the science involved.

"The companies have worked according to strict standards applicable for such type of studies," said Kristin Kaufmann, a spokeswoman for TeGenero.

What would be a Bayesian approach?

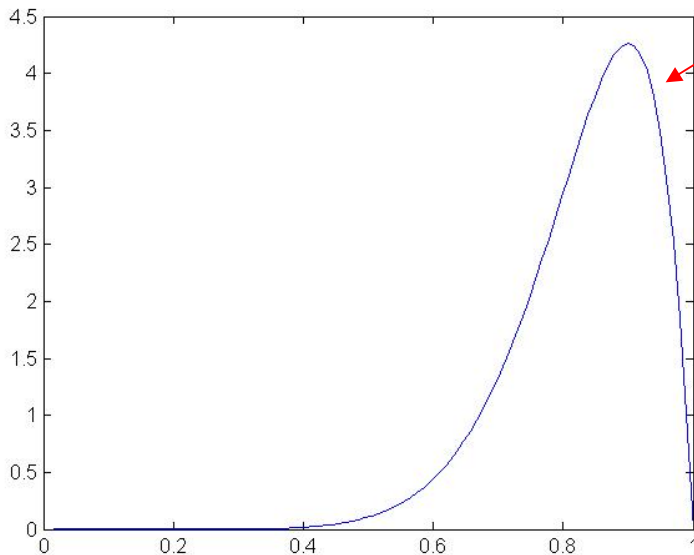
H_p is the hypothesis that prob = p .

$P(H_p)$ is its probability.

$$P(H_p|\text{data}) \propto P(\text{data}|H_p)P(H_p) \propto p^9(1-p)$$

$$P(H_p|\text{data}) = \frac{p^9(1-p)}{\int_0^1 p^9(1-p)dp}$$

```
xx = linspace(0, 1);  
plot(xx, betapdf(xx, 10, 2), '-');
```



The curve is the answer.
We might, however, summarize it in various ways:

```
y1 = betapdf(.5, 10, 2)  
y1 =  
    0.1074  
y2 = betapdf(.9, 10, 2)  
y2 =  
    4.2616  
quad(@(x)betapdf(x, 10, 2), 0, .5)  
ans =  
    0.0059
```

For an example in which we might use a more sophisticated prior, suppose the data is **10 heads in a row**.

“Hmm. When people make me watch them flip coins, 95% of the time it’s a (nearly) fair coin [A], 4% of the time it’s a double-headed [B] or double-tailed coin [C], and 1% of the time something else weird is happening [D].”

Case A:	$0.95 \times (0.5)^{10} = 0.00093$	0.043
Case B	$0.02 \times 1^{10} = 0.02$	0.915
Case C	$0.02 \times 0^{10} = 0$	0.000
Case D	$0.01 \times \int_0^1 p^{10} dp = 0.00091$	0.042

This kind of analysis is not usually publishable, unless you can justify your choice of prior on the basis of already published data. (In such a case it is dignified by the term “meta-analysis”.) However, it is a good way to live your life, especially if you are a person who likes to make bets!

(Can you remember that we were listing two Bayesian criticisms of tail tests?)

(2) Not suitable for comparing hypotheses quantitatively. Best you can do is rule one out, leaving the other viable. Ratio of p-values is not anything meaningful!

you should go learn about Likelihood Ratio tests, but I personally think that Bayes odds ratio is easier to compute and easier to interpret