# CS395T
# Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

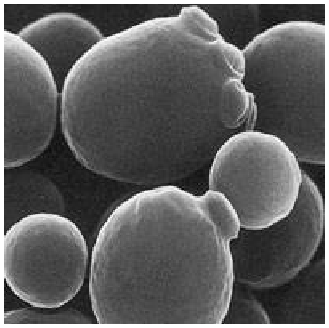
Lecture 7

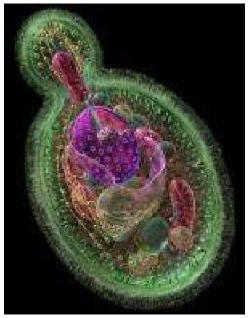
For practice with p- and t-values, let’s look at the Sac cer genome.

We’ll use as a data set all of Chromosome 4.

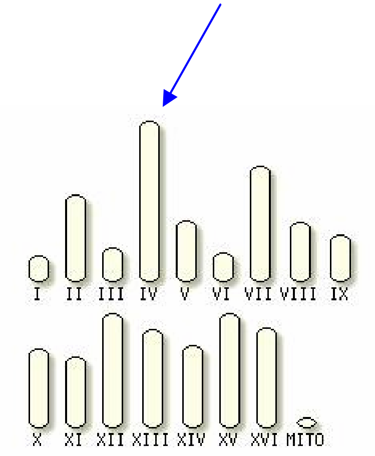
Yeast and Human are very close relatives in the great scheme of things.

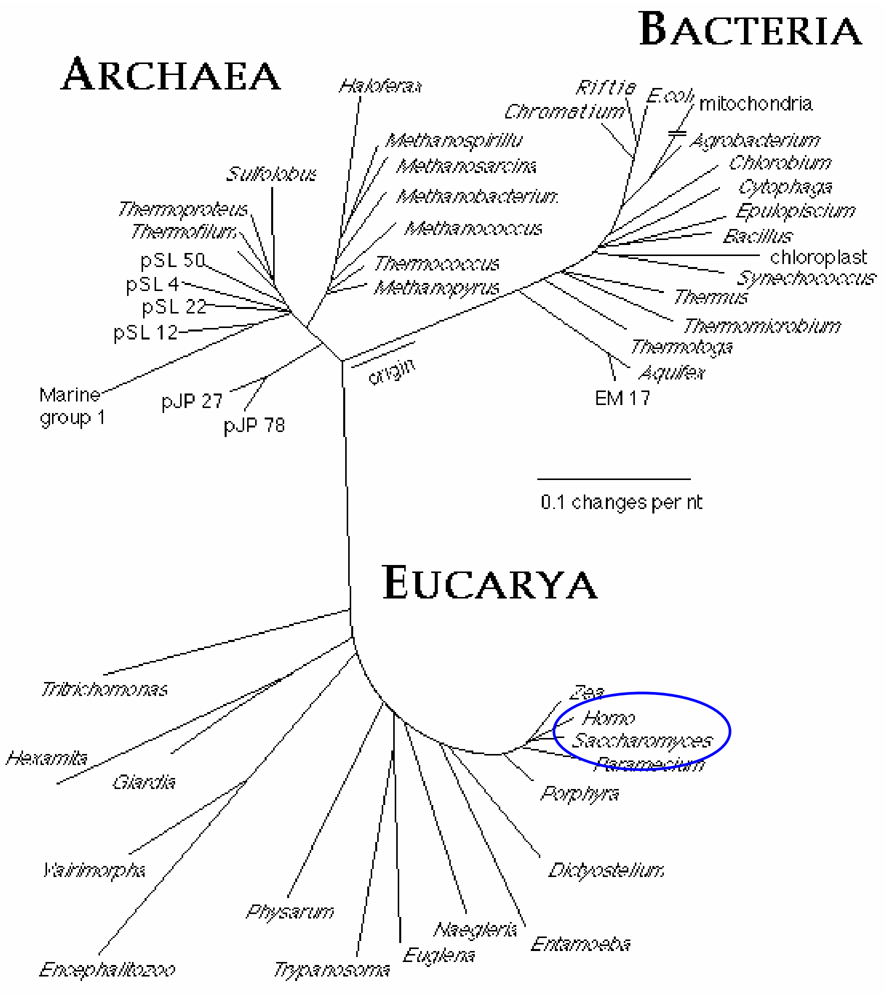
***Saccharomyces cerevisiae***
***= baker’s yeast***





Chromosome 4:
ACACCACACC…(1531894 omitted)…TAGCTTTTGG





Norm Pace, U.C. Berkeley

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

Take the file **SacSerChr4.txt** (on course web site).

Count the letters **A,C,G,T**.

You should get:

***A = 476750***<br>
***C = 289341***<br>
***G = 291352***<br>
***T = 474471***

Are these counts consistent with the model

$$p_A = p_C = p_C = p_T = 0.25 \ ?$$

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

Are they consistent with the model

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

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

As always, the starting point is to write down a model. Bayesian: What is the probability of the data. Frequentist: What is the probability of a test statistic for a null hypothesis.

A possible model is multinomial: At each position an i.i.d. choice of A,C,G,T, with respective probabilities adding up to 1.

Almost equivalent (and simpler for now) is 4 separate binomial models: At each position an i.i.d. choice of A vs. not A with some probability $p_A$. Then do separately for $p_C$, $p_G$, $p_T$.

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

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

Why? 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*).

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

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

Let’s dispose of the silly (all p’s = 0.25):

The test statistic: the value of the observed count under the null hypothesis that it is binomially (or equivalent normally) distributed with p=0.25.

$$\mu = 0.25\,N$$

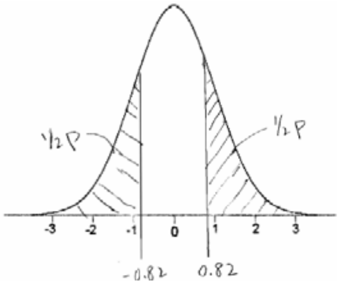
$$\sigma = \sqrt{0.25 \times 0.75\,N}$$

$$t = \frac{n - \mu}{\sigma}$$

$$p = 2[1 - P_{\text{Normal}}(|t|)]$$

t-value = number of standard deviations

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



| | t-value | p-value |
|---|---|---|
| A | 174.965 | ≈ 0 |
| C | –174.715 | ≈ 0 |
| G | –170.963 | ≈ 0 |
| T | 170.713 | ≈ 0 |

The null hypothesis is (totally, infinitely, beyond any possibility of redemption!) ruled out.

The not-silly model: A and T occur with identical probabilities, as do C and G.

The test statistic: Difference between A and T (or C and G) counts under the null hypothesis that they have the same p, which we will estimate in the obvious way (which is actually an MLE).

$$\hat{p}_{AT} = \tfrac{1}{2}(n_A + n_T)/N$$
$$\hat{p}_{CG} = \tfrac{1}{2}(n_C + n_G)/N$$

$$n_A \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$n_T \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$\Rightarrow n_A - n_T \sim \mathrm{Normal}(0, \sqrt{2N\hat{p}_{AT}(1-\hat{p}_{AT})})$$

the difference of two Normals is itself Normal

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

It makes Bayesians nervous to see parameters estimated by MLE, then re-used in estimating other parameters. People do this all the time, and it’s usually OK. But Bayesians feel more secure estimating the full posterior probability of all the parameters at once!

## In MATLAB the calculation now looks like this:

```
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))
```

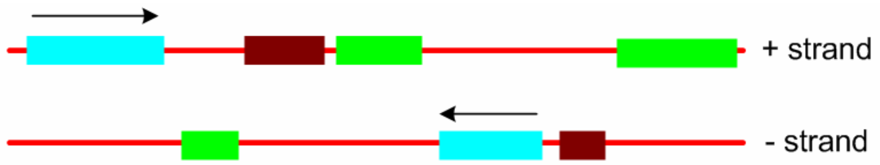
*A = 476750*
*C = 289341*
*G = 291352*
*T = 474471*

2-tailed

```
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
```

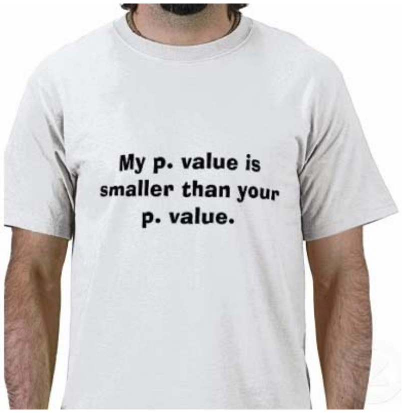
Surprise!
The model is ruled out
with high significance
(small p-value)!

Why? Because, we're discovering genes!



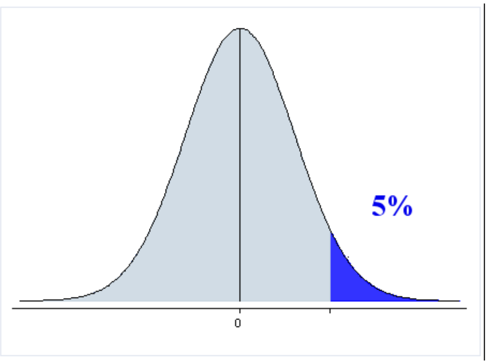
The fluctuating "units" are indeed not single bases. Rather, they are genes which, individually, do not have (or prefer) A=T, C=G. Their placement on one strand or the other is random.

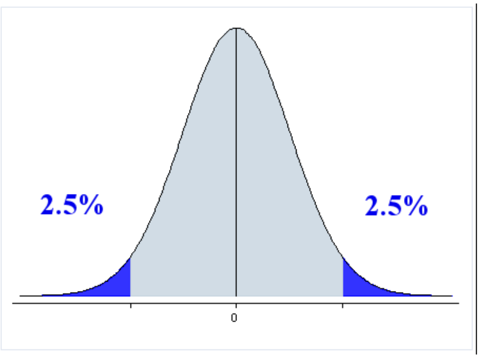
# The classic p-value (or tail-) test terminology:



t-shirt for sale on the Web

- “null hypothesis”
- “the statistic” (e.g., t-value or $\chi^2$)
  - calculable for the null hypothesis
  - intuitively should be “deviation from” in some way
- “the critical region” $\alpha$
  - biologists use 0.05
  - physicists use 0.0026 (3 $\sigma$)
- 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 $\alpha$ significance level
- after seeing the data you
  - may adjust the significance level $\alpha$
  - may not try a different statistic, because any statistic can rule out at the $\alpha$ level in $1/\alpha$ tries (“data dredging” for a significant result!)
- if you decided in advance to try N tests, then the critical region for $\alpha$ significance is $\alpha/N$ (Bonferroni correction).

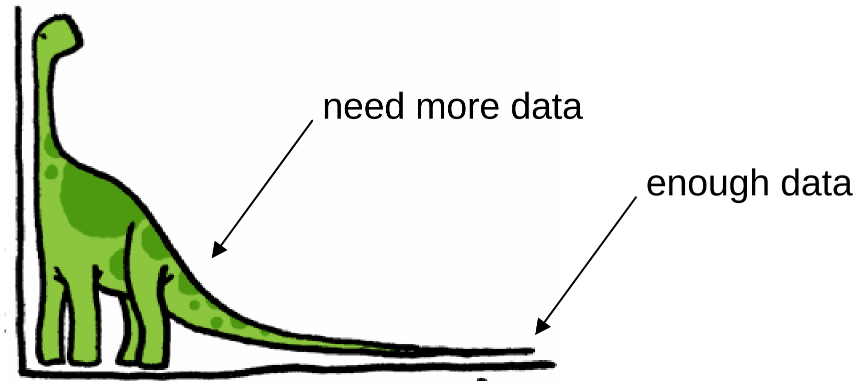




## Tips on tail tests:

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 three 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(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."*

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

$$p(\geq N) = 2^{-(N+1)}(1 + \tfrac{1}{2} + \tfrac{1}{4} + \cdots) = 2^{-N}$$

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

(*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.

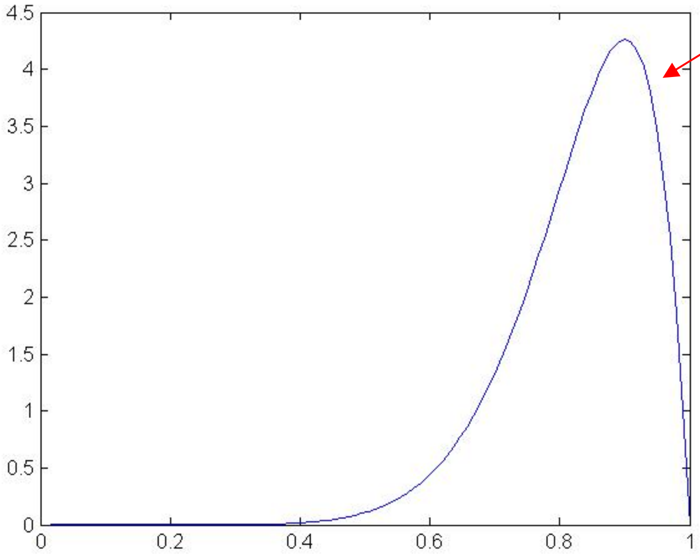
What would be a Bayesian approach?

$H_p$ is the hypothesis that prob = $p$.

$P(H_p)$ is its probability.

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

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



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

Likelihood (or posterior probability) ratio:

$$\frac{P(H_{0.5}|\text{data})}{P(H_{\max}|\text{data})} = \frac{0.1074}{4.2616} = 0.0252$$

Bayes tail probability:

$$\int_0^{0.5} P(H_p|\text{data})dp = 0.0059$$