Opinionated Lessons in Statistics

by Bill Press

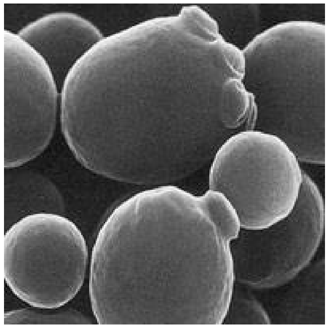
#13 The Yeast Genome

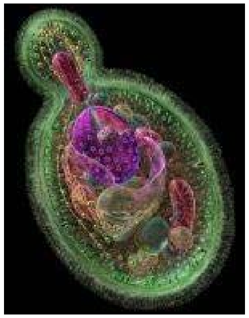
For practice with p- and t-values, let’s look at the Sac cer genome.

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

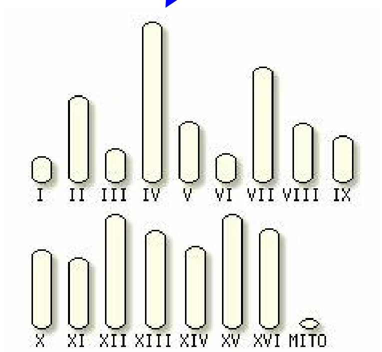
Yeast and Human are very close relatives in the great scheme of things.

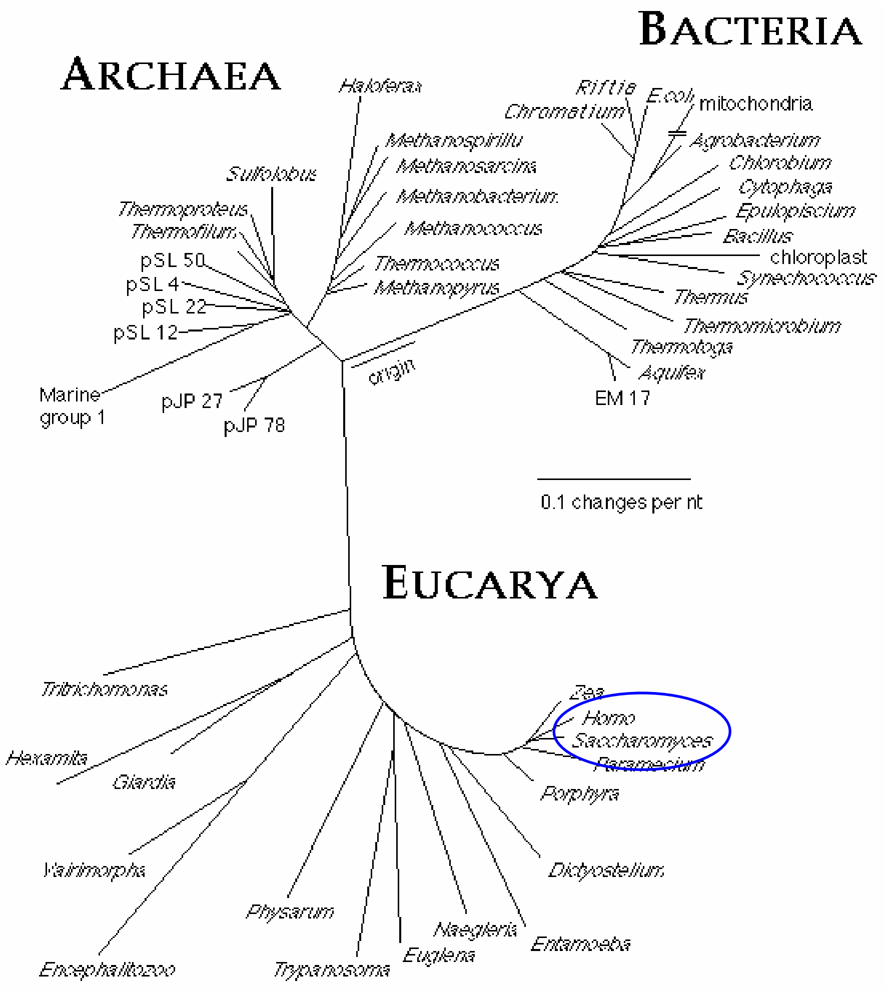
Saccharomyces cerevisiae
= baker’s yeast





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
C = 289341
G = 291352
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*).

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

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

Can we rule out the silly hypothesis that 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.

$$\mu = 0.25\,N$$

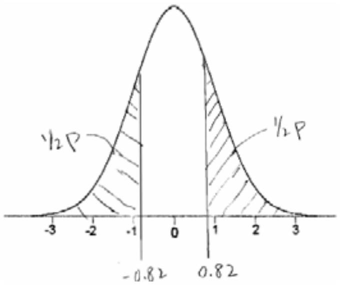
$$\sigma = \sqrt{0.25 \times 0.75\,N}$$

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

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

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. We estimate p by its maximum likelihood estimate (see Bernoulli Trials lecture).

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

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

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!

Calculate the answer (here in MATLAB):

```
dif = [count(1)-count(3); count(2)-count(4) ]
pdiff = [0.5*(count(1)+count(3))/n; 0.5*(count(2)+count(4))/n]
mu = [0; 0];
sig = sqrt(2 .* pdiff .* (1 - pdiff) .* len)
tval = (dif - mu) ./ sig
pval = 2*(1-normcdf(abs(tval),0,1))
```

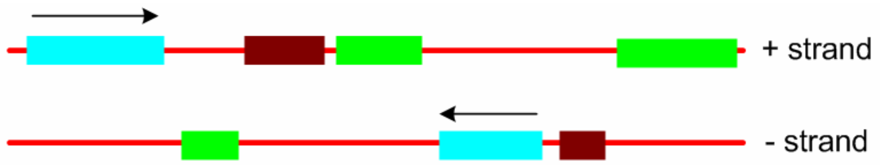
A = 476750
C = 289341
G = 291352
T = 474471

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.