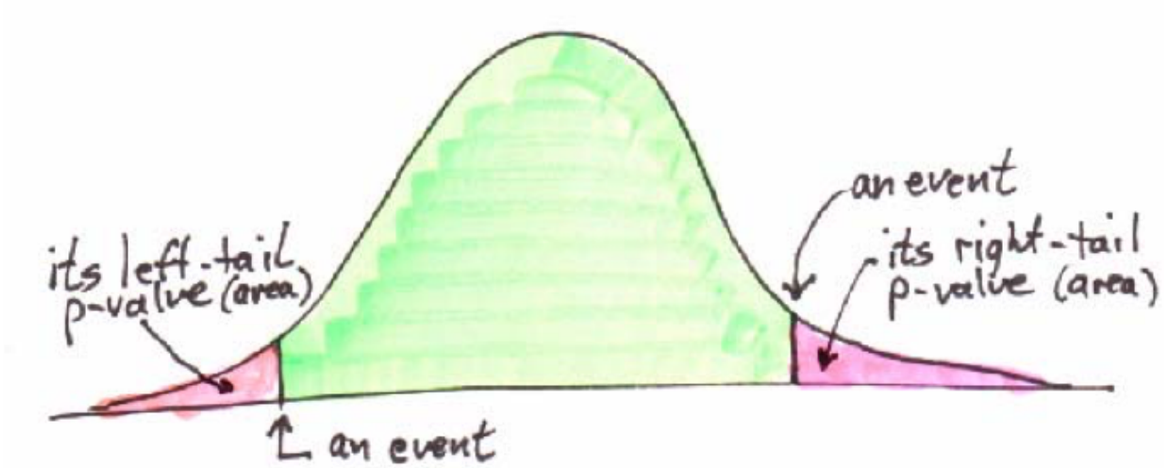
Lectures 6: null hypethesis tests I.

frequentist view of null hypothesis:

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.)



The idea is to pick a null hypothesis that is uninteresting, so that if you rule it out you have discovered something interesting.

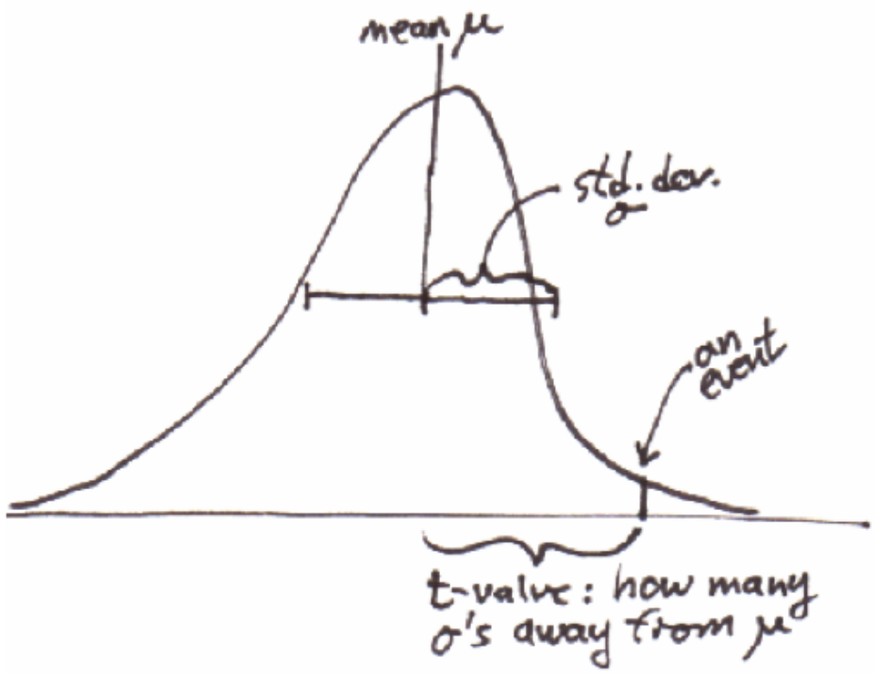
If the null hypothesis is true, then p-values are uniformly distributed in (0,1), in principle exactly so.

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").

frequentist view of null hypothesis:

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!

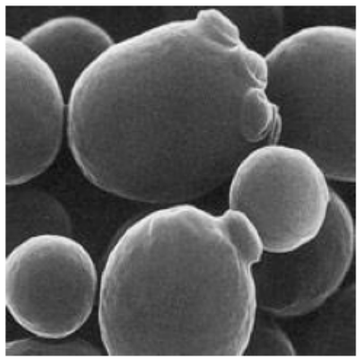
It's much easier to compute a score ("statistic") that depends only on the mean and standard deviation of the expected distribution. But, in general, this is interpretable as "likely" or "unlikely" only relative to a Gaussian (which may or may not be relevant). Often we are in an asymptotic regime where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

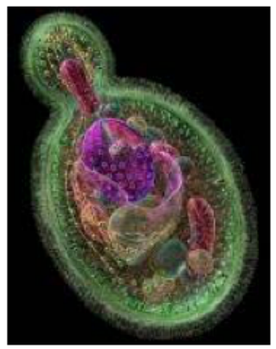
The reason that t-values often are relevant is, of course, the Central Limit Theorem, as we have seen.

frequentist view of null hypothesis:

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.

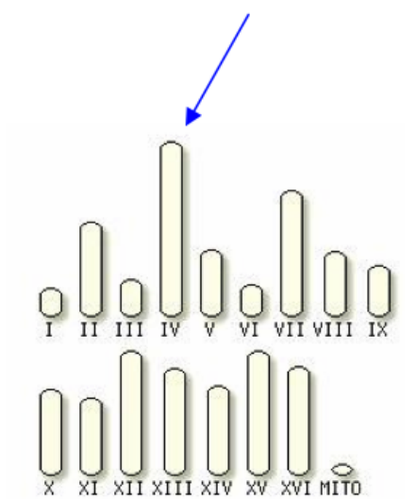
Saccharomyces cerevisiae
= baker's yeast



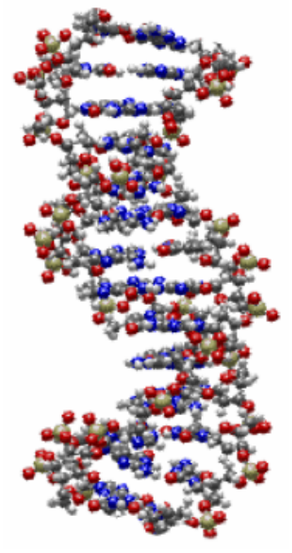


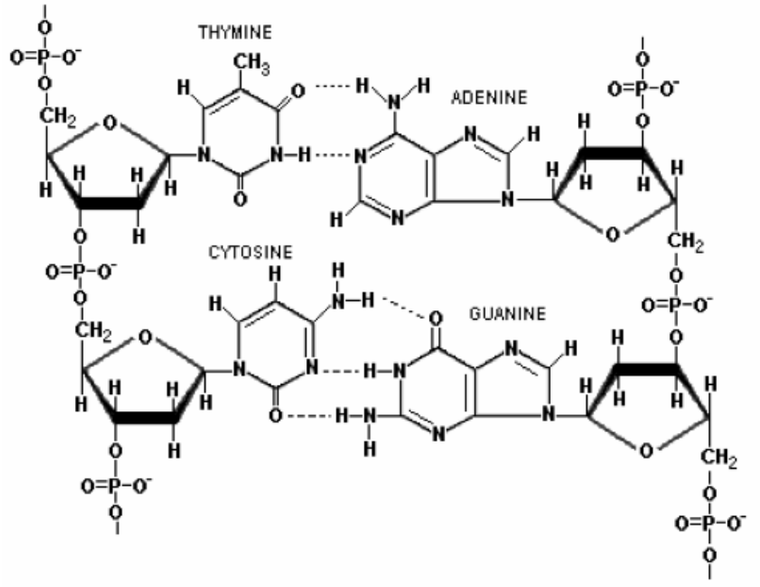
goal is to build probability models for chromosome 4 from four nucleobases ACGT and subject them to null hypothesis

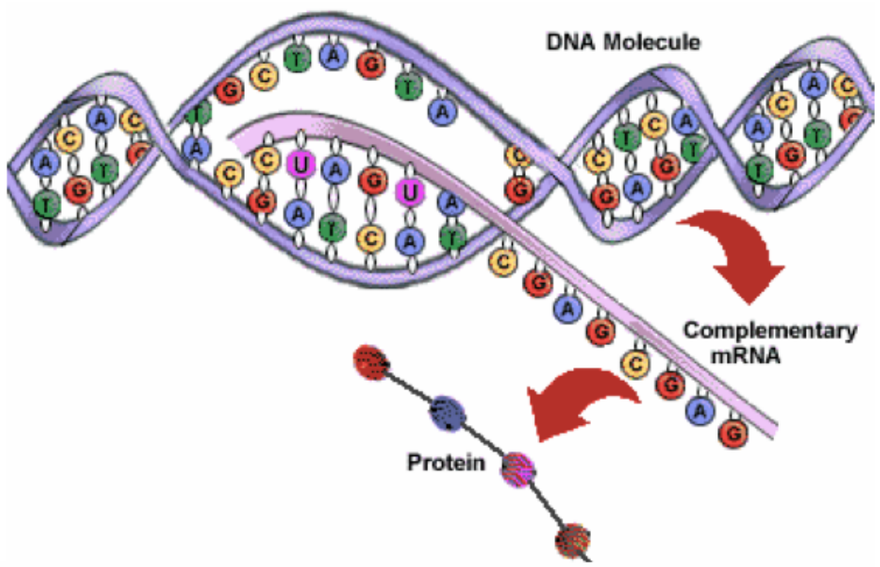
Chromosome 4:
ACACCACACC (1531894 omitted) TAGCTTTTGG

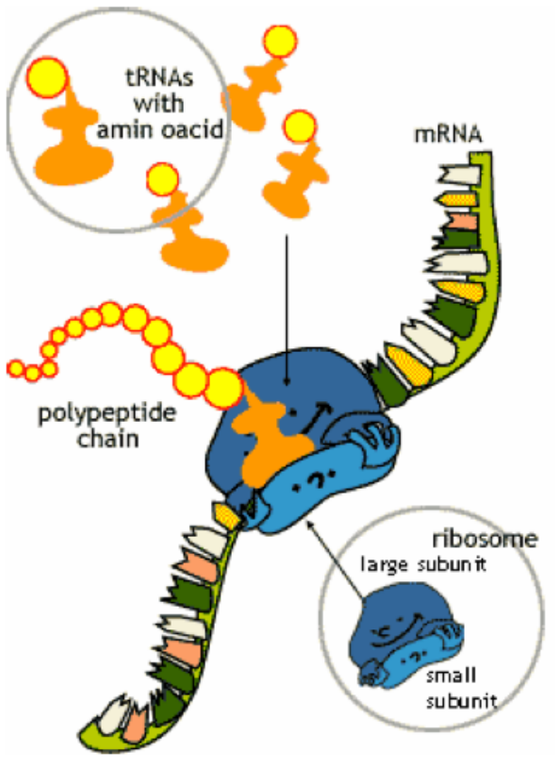


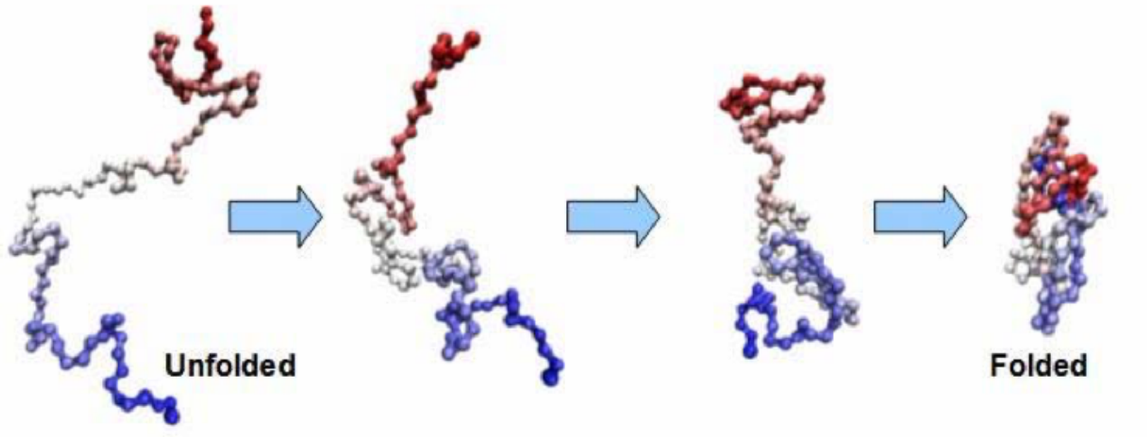
molecular biology on one slide:

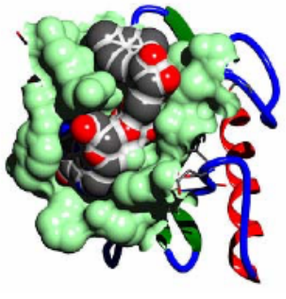












frequentist view of null hypothesis (DNA example):

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

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

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

frequentist view of null hypothesis (DNA example):

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*).

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

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

frequentist view of null hypothesis (DNA example):

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.

$$\mu = 0.25\,N$$

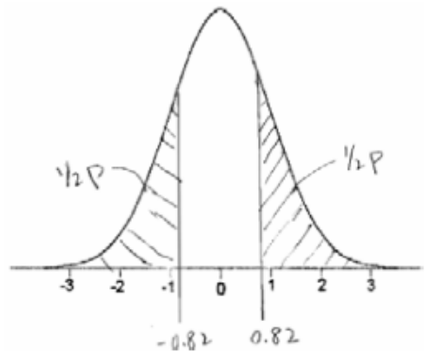
$$\sigma = \sqrt{0.25 \times 0.75\,N}$$

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

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

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.

frequentist view of null hypothesis (DNA example):

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).

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

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

the difference of two Normals is itself Normal

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

frequentist view of null hypothesis (DNA example):

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))
```

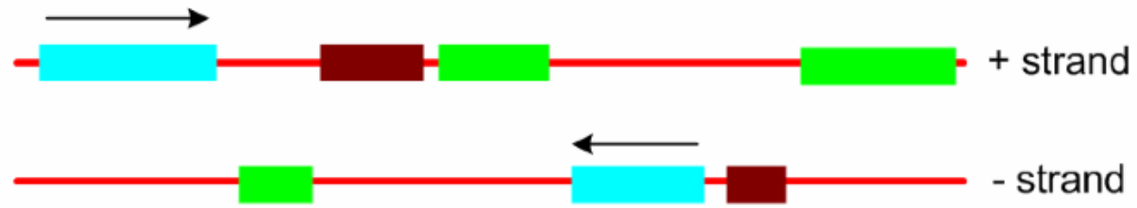
2-tailed

```
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)!

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

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.