Lectures 6: null hypethesis tests II.

Null hypothesis testing:

- Computing a p-value requires a null hypothesis, a test statistic (together with deciding whether the researcher is performing a one-tailed test or a two-tailed test), and data. Even though computing the test statistic on given data may be easy, computing the sampling distribution under the null hypothesis, and then computing its cumulative distribution function (CDF) is often a difficult computation. Today, this computation is done using statistical software and computational power.

- When the null hypothesis is true, the probability distribution of the p-value is uniform on the interval [0,1]. By contrast, if the alternative hypothesis is true, the distribution is dependent on sample size and the true value of the parameter being studied.

The distribution of p-values for a group of studies is called a p-curve. The curve is affected by four factors: the probability that a study is examining a true hypothesis rather than a false hypothesis, the power of the studies investigating true hypotheses, the Type 1 error rates, and publication bias. A p-curve can be used to assess the reliability of scientific literature, such as by detecting publication bias or p-hacking.

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 hypothesis. 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.

$$\mu = 0.25\, N$$

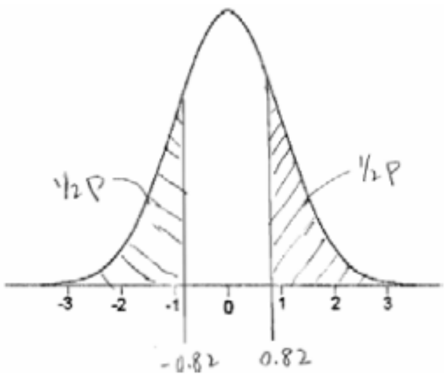
$$\sigma = \sqrt{0.25 \times 0.75\, N}$$

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

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

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).

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

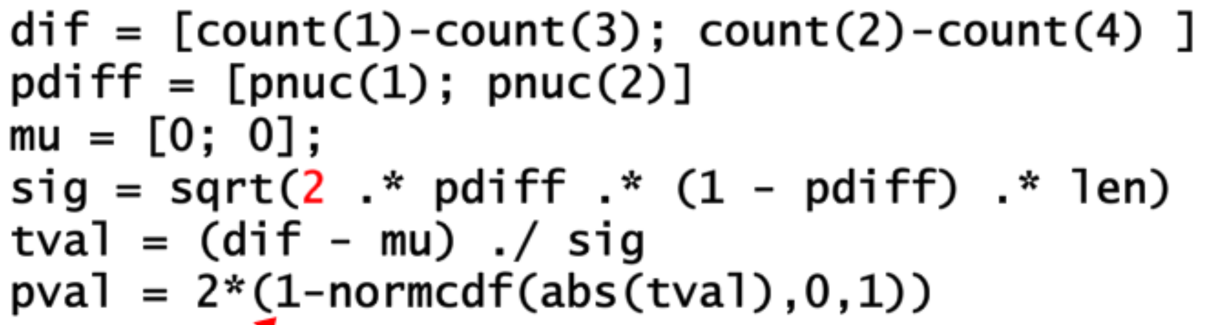
$$n_A \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$n_T \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$\Rightarrow n_A - n_T \sim \mathrm{Normal}(0, \sqrt{2N\hat{p}_{AT}(1-\hat{p}_{AT})})$$

the difference of two Normals is itself Normal

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

frequentist view of null hypothesis (DNA example):

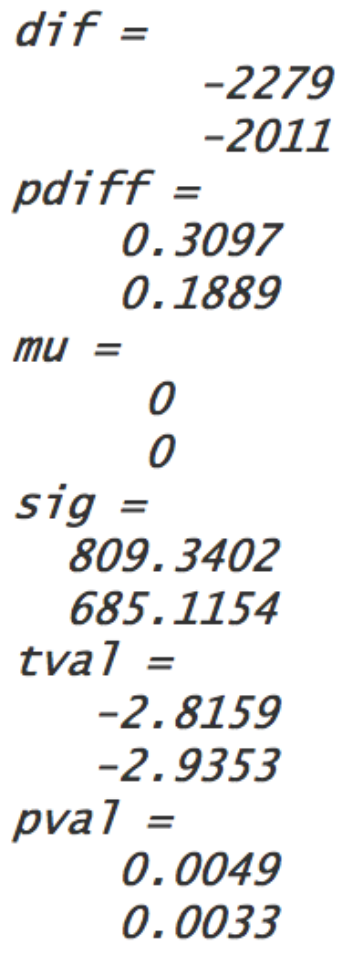
In MATLAB the calculation now looks like this:



```
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))
```

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

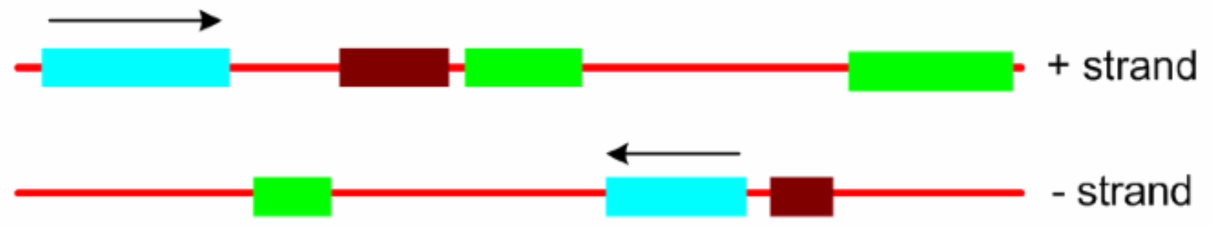
2-tailed



```
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.

Null hypothesis testing (Bayesian):

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.

regardless of the order in the outcome

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

2-sided tail test of 9 or more identical

9 heads or more

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

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

Null hypothesis testing (Bayesian):

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!)

Null hypothesis testing (Bayesian):

What would be a Bayesian approach?

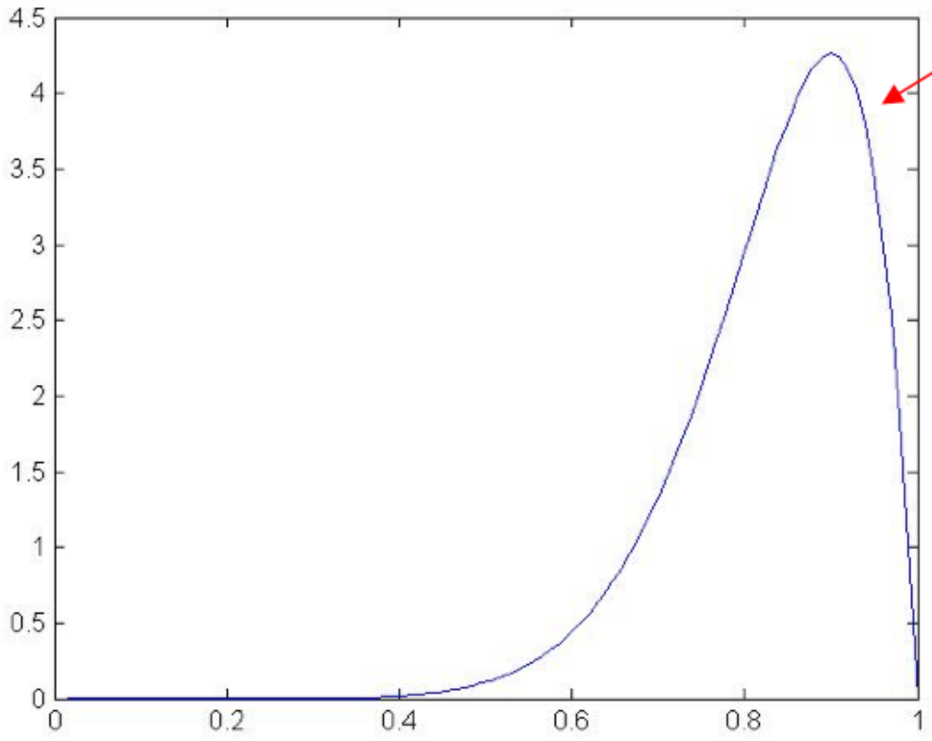
H_p is the hypothesis that prob = p.

$P(H_p)$ is its probability.

flat prior

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

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



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:

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

Null hypothesis testing (Bayesian):

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!

Null hypothesis testing (Bayesian):

(Can you remember that we were listing three 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

(3) The sanctification of certain p-values (e.g., the magic p=0.05 value) is naïve and misleading.

(on the one hand) 1 in 20 results are wrong! Imagine if we built nuclear power plants to this low a standard.

(on the other hand) the large majority of results with p=0.10 are in fact correct. These could sometimes be acted on.