Lectures 7: null hypethesis tests II.

Null hypothesis testing is reductio ad absurdum argument:

- Null hypothesis testing is a reductio ad absurdum argument adapted to statistics: a hypothesis is shown to be valid by demonstrating the improbability of the consequence that results from assuming the counter-claim to be true (fair coin).

- The only hypothesis that needs to be specified in this test and which embodies the counter-claim is referred to as the null hypothesis.

- A result is said to be statistically significant if it allows us to reject the null hypothesis. That is, as per the reductio ad absurdum reasoning, the statistically significant result should be highly improbable if the null hypothesis is assumed to be true.

- The rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null hypothesis. However, unless there is a single alternative to the null hypothesis, the rejection of null hypothesis does not tell us which of the alternatives might be the correct one.

Null hypothesis testing:

- A statistical hypothesis refers to a probability distribution that is assumed to govern the observed data. If X is a random variable representing the observed data and H is the statistical hypothesis under consideration, then the notion of statistical significance can be quantified by the conditional probability P(X I H), which gives the likelihood of the observation if the hypothesis is assumed to be correct.
- The p-values should not be confused with probability on hypothesis (as is done in Bayesian Hypothesis Testing) such as P(H I X) , the probability of the hypothesis given the data, or P(H), the probability of the hypothesis being true, or P(X), the probability of observing the given data.

Null hypothesis testing:

The p-value is defined as the probability, under the assumption of hypothesis H, of obtaining a result equal to or more extreme than what was actually observed. Depending on how it is looked at, the "more extreme than what was actually observed" can mean
$\{ X \geq x \}$ (right-tail event) or
$\{ X \leq x \}$ (left-tail event) or the "smaller" of
$\{ X \leq x \}$ and $\{ X \geq x \}$ (double-tailed event).

Thus, the p-value is given by

- $P(X \geq x \mid H)$ for right tail event,
- $P(X \leq x \mid H)$ for left tail event,
- $2 * \min \{ P(X \leq x \mid H) , P(X \geq x \mid H) \}$ for double tail event.

The smaller the p-value, the larger the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation. The hypothesis H is rejected if any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value α, which is referred to as the level of significance.

Unlike the p-value, the α level is not derived from any observational data and does not depend on the underlying hypothesis; the value of α is instead determined by the consensus of the research community that the investigator is working in.

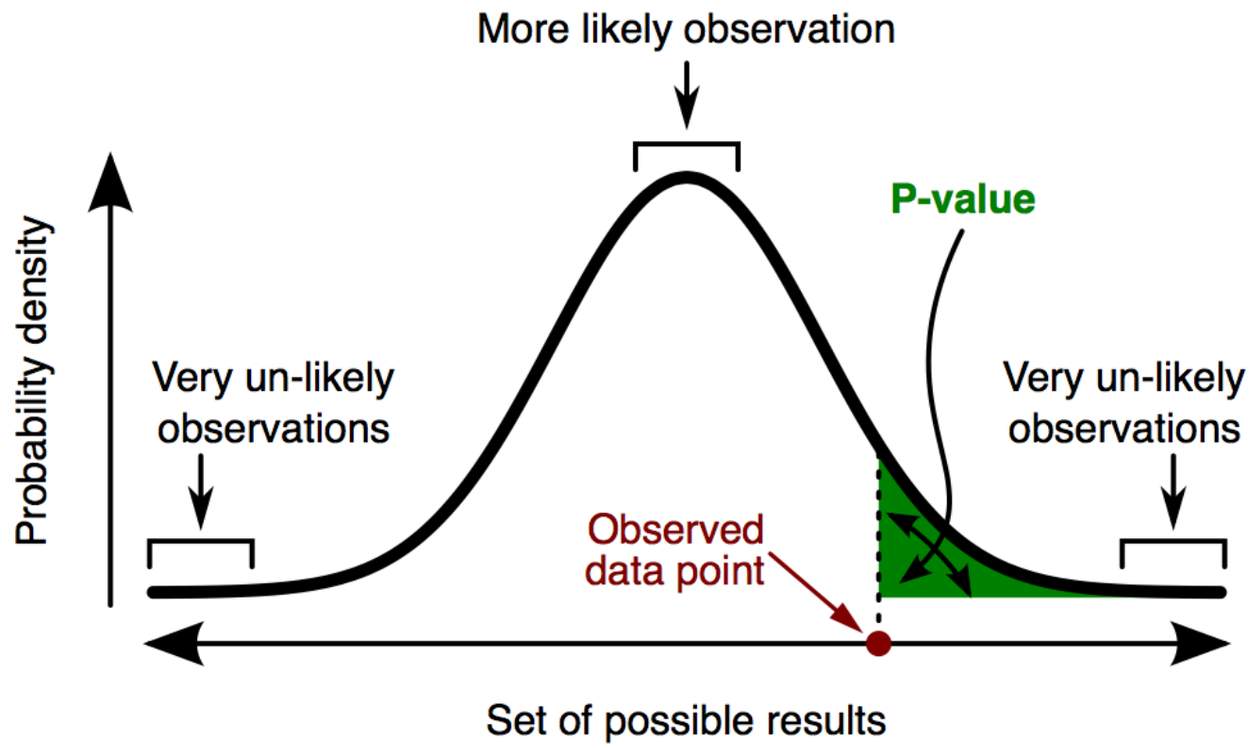
Null hypothesis testing:

Important:

Pr (observation I hypothesis) ≠ Pr (hypothesis I observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error:
the transposed conditional fallacy.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

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.

Coin flipping null hypothesis testing:

As an example of a statistical test, an experiment is performed to determine whether a coin flip is fair (equal chance of landing heads or tails) or unfairly biased (one outcome being more likely than the other).
Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips. The null hypothesis is that the coin is fair, and the test statistic is the number of heads. If a right-tailed test is considered, the p-value of this result is the chance of a fair coin landing on heads at least 14 times out of 20 flips. That probability can be computed from binomial coefficients as

$$\begin{aligned}&\text{Prob}(14\text{ heads}) + \text{Prob}(15\text{ heads}) + \cdots + \text{Prob}(20\text{ heads})\\ &= \frac{1}{2^{20}}\left[\binom{20}{14} + \binom{20}{15} + \cdots + \binom{20}{20}\right] = \frac{60{,}460}{1{,}048{,}576} \approx 0.058\end{aligned}$$

This probability is the p-value, considering only extreme results that favor heads. This is called a one-tailed test. However, the deviation can be in either direction, favoring either heads or tails. The two-tailed p-value, which considers deviations favoring either heads or tails, may instead be calculated. As the binomial distribution is symmetrical for a fair coin, the two-sided p-value is simply twice the above calculated single-sided p-value: the two-sided p-value is 0.116. In the above example:

null hypothesis (H_0) p(head) = 0.5

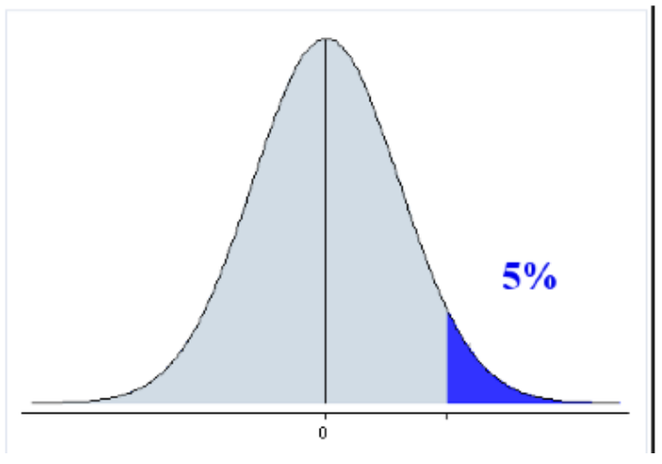
- Test statistic: number of heads
- Level of significance: 0.05
- Observation O: 14 heads out of 20 flips; and
- Two-tailed p-value of observation O given H_0 = 2*min(Prob(no. of heads ≥ 14 heads), Prob(no. of heads ≤ 14 heads))= 2*min(0.058, 0.978) = 2*0.058 = 0.116.

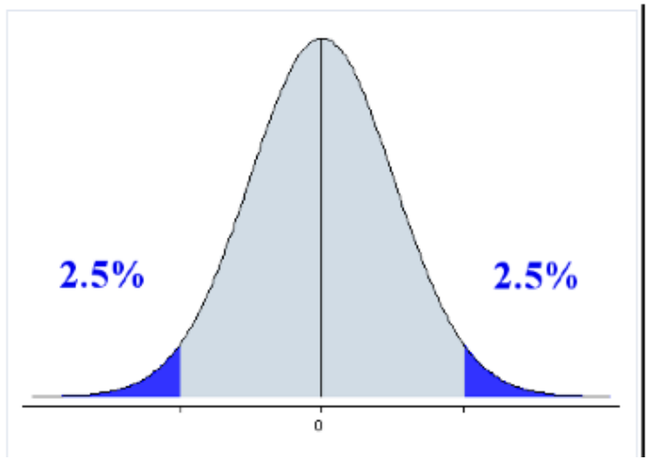
Note that the Prob(no. of heads ≤ 14 heads) = 1 - Prob(no. of heads ≥ 14 heads) + Prob(no. of head = 14) = 1 - 0.058 + 0.036 = 0.978; however, symmetry of the binomial distribution makes that an unnecessary computation to find the smaller of the two probabilities. Here, the calculated p-value exceeds 0.05, so the observation is consistent with the null hypothesis, as it falls within the range of what would happen 95% of the time were the coin is in fact fair. Hence, the null hypothesis at the 5% level is not rejected. Although the coin did not fall evenly, the deviation from expected outcome is small enough to be consistent with chance.
However, had one more head been obtained, the resulting p-value (two-tailed) would have been 0.0414 (4.14%). The null hypothesis is rejected when a 5% cut-off is used.

Null hypothesis testing summary:

- "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 (3 σ)
- 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 ("data dredging" 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).

Phys. Rev. Lett. discovery threshold:
5 σ (0.000057 percent)





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.

$$\mu = 0.25\,N$$

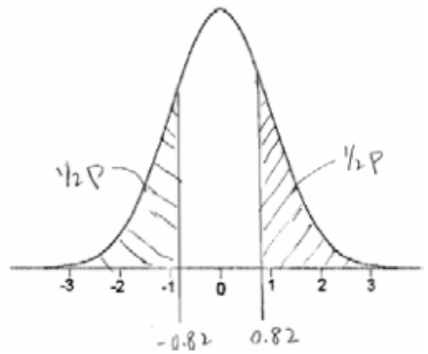
$$\sigma = \sqrt{0.25 \times 0.75\,N}$$

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

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

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).

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

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

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))
```

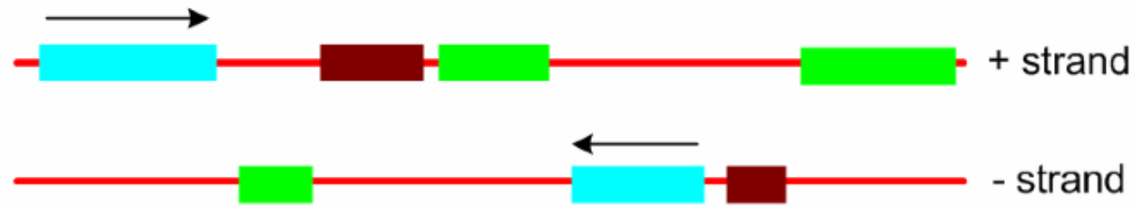
2-tailed

```
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.

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.

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

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):

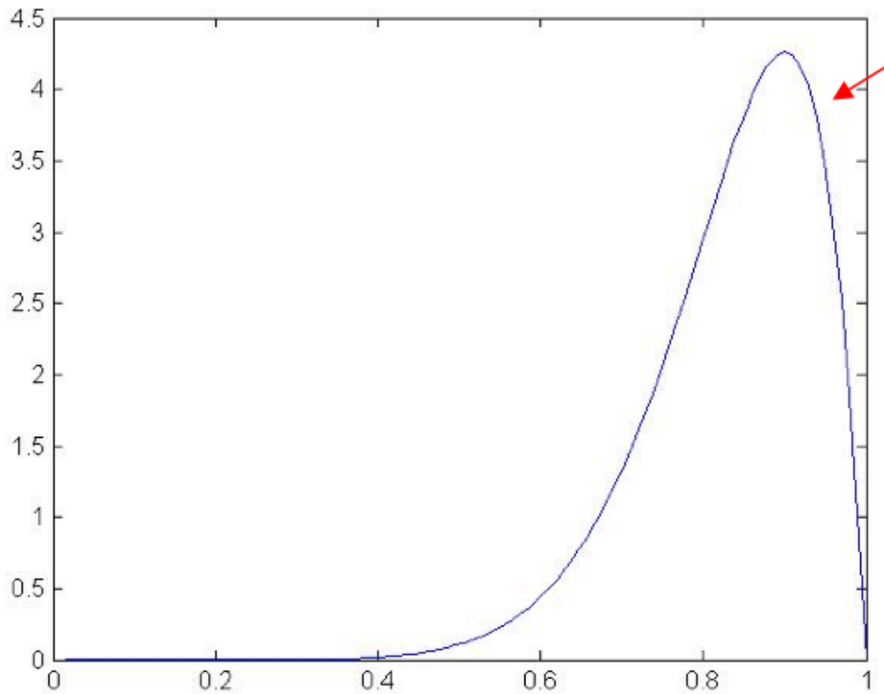
What would be a Bayesian approach?

H_p is the hypothesis that prob = p.

$P(H_p)$ is its probability.

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

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



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 <u>not</u> 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.