Statistical testing

Samantha Kleinberg

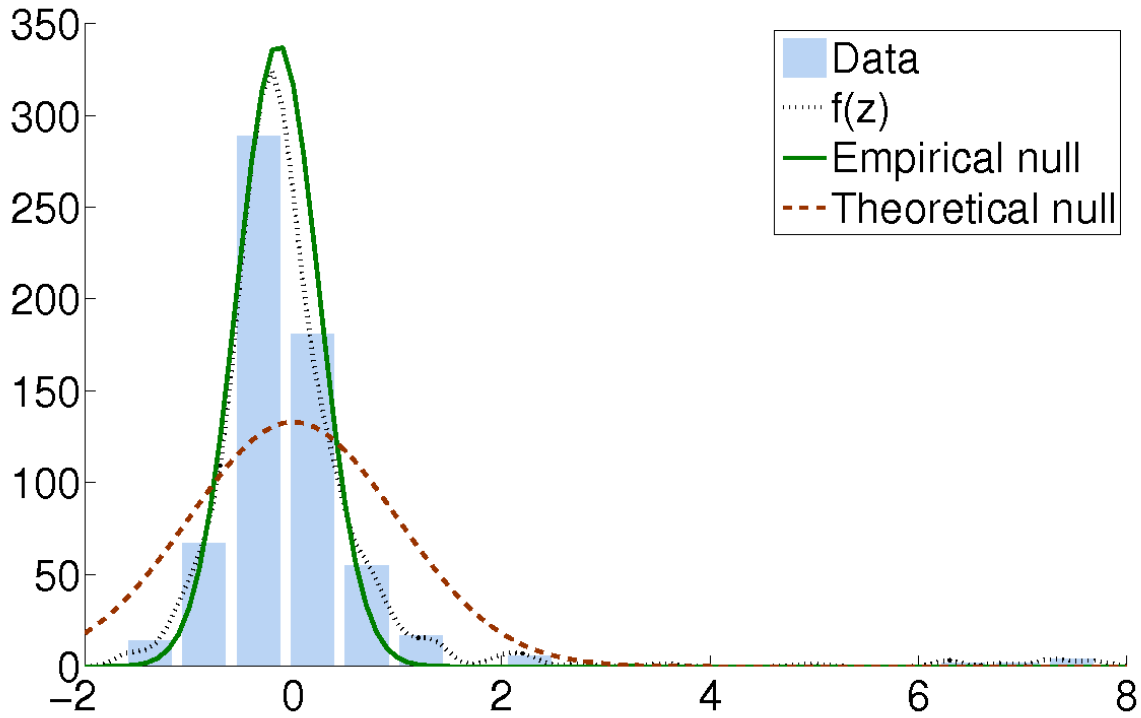
October 20, 2009

Significance testing and bioinformatics

- Gene expression: Frequently have microarray data for some group of subjects with/without the disease. Want to find which genes are different in patients with disease.
  i.e. which are different enough that they are significant?
- Epidemiology: People in a region seem to have a high rate of cancer. Is this rate significantly out of the ordinary?
- Etiology: Many factors seemingly associated with CFS, which are overr represented in the CFS population versus a control?

More motivation

We often have some statistics associated with our results and must choose a threshold. How should we do this?



Basic problem

- How can we tell if a result is significant?
- Example: flip a coin 10 times
  - Expect to see 5 heads, 5 tails
  - What if we see 9 heads and 1 tail?
  - If the coin is fair, probably of heads = probability of tails = $1/2$
  - If coin is fair, probability of 9 H 1 T is $(\frac{1}{2}^{10}) \times 10 = 0.010$
- Assuming a fair coin, this observation is extremely unlikely
- What if we're testing 100 coins?
- More chance of seeing anomalous outcomes, so must account for this

p-values

A p-value is:

the probability of getting a test statistic *at least as extreme* as what is observed, given that the null hypothesis is true.

A p-value is NOT:

- Probability of the null hypothesis being true
- Something that can definitely say whether a hypothesis is true
- Able to show causality (a small p-value won't prove that smoking causes lung cancer). Correlation $\neq$ causation

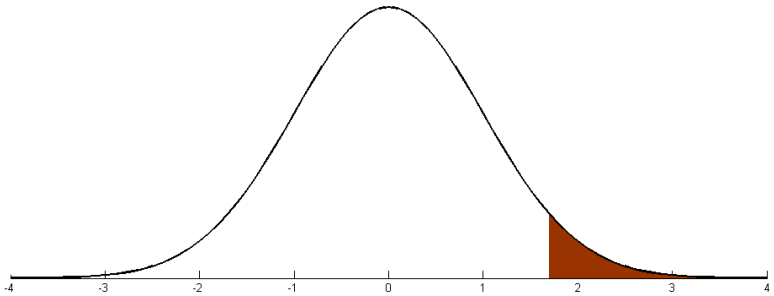
Example

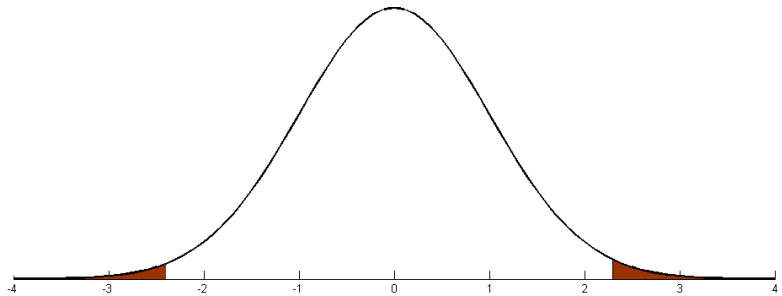
- This means that for the coin flipping case our p-value will be $P(9H1T) + P(10H) + P(10T) + P(9T1H)$

$$P(10H) = P(10T) = (1/2)^{10} = 0.001 \quad (1)$$
$$P(9H1T) = P(9T1H) = (1/2)^{10} \times 10 = 0.01 \quad (2)$$
$$\text{Total} = 0.001 + 0.001 + 0.01 + 0.01 \quad (3)$$
$$= 0.022 \quad (4)$$

- Frequent threshold is $\alpha = 0.05$ (Note, nothing special about 0.05, it's just a convention!)
- Since $p < \alpha$, we should say the coin is unfair $(0.022 < 0.05)$

One tail or two?





- Two tailed: test is biased for heads *or* tails, so we look at getting many more or many fewer heads
- One tailed: Test if coin biased just for tails or just for heads

Multiple tests

- Now what if we flip **100** coins 10 times
- Should we expect to see at least one run of $9H$ $1T$?
- If α_c is significance level for one test, and α_e is level for experiment, does $\alpha_c = 0.05$ guarantee $\alpha_e = 0.05$?

Let's check

Let's say x is the event of getting $9H$ and $1T$. Then, y is the event of getting a result at least as extreme as this (i.e. x, or $9T$ $1H$, or all H or all T).
Before we calculated $P(y) = 0.022$.
So,

$$P(\neg y) = 1 - P(y) = 1 - 0.022 = 0.978 \tag{5}$$

Now we want the probability of y at least once in 5 tries. That's:

$$1 - P(\neg y)^5 = 0.11 \tag{6}$$

What about 50 tries?

$$1 - P(\neg y)^{50} = 0.67 \tag{7}$$

100 tries? The probability is 0.89.

General case

Then, with $\alpha = 0.05$, the probability of a false positive due to chance is:

$$(1 - 0.95^{100}) = .994 \tag{8}$$

Why?

- If we test N with significance level α_c, will find:

$$\alpha_e = 1 - (1 - \alpha_c)^N = \text{ if tests independent} \tag{9}$$

$$\alpha_e \leqslant N \times \alpha_c = \text{ if dependent} \tag{10}$$

- In general, can approximate the experiment-wise significance level as
  $N \times \alpha_c$

Types of error

	Accept null	Reject null	totals
True null H	U	V $(F+)$	m_0
False null H	T $(F\text{-})$	S	m_1
Total	$m - R$	R	m

- U: true null, we correctly accept null hypothesis
- S: false null, we correctly reject the null hypothesis
- V: false positive, null hypothesis is true, but we rejected it
- T: false negative, null hypothesis is false, but we accepted it (missed opportunity for discovery)
- Other terminology:
  Type I error: reject null when shouldn't (False +)
  Type II error: don't reject null when should (False -)

FDR/FNR

	Accept null	Reject null	totals
True null H	U	V $(F+)$	m_0
False null H	T $(F\text{-})$	S	m_1
Total	$m - R$	R	m

- FDR (false discovery rate): V/R
  proportion of falsely rejected nulls out of all rejected nulls
- FNR (false negative rate): $U/(m - R)$
  proportion of falsely accepted nulls out of all accepted nulls
- FWER: $P(V \geqslant 1)$
  probability of at least one false discovery out of all tests
- PCER (per comparison error rate) V/m

What to control?

- Could control Type I or Type II error: is it better to make a false discovery or miss a possible discovery? (We focus on FDR, since, for example it's "worse" to incorrectly say a gene is an oncogene when it's not, than to not find all oncogenes)
- Probability of even one error, or ratio of errors to real discoveries? (We'll look at methods for both)

What's a family of hypotheses?

- Previously slides referred to some group of m tests, but glossed over how we create this group
- Set of simultaneous tests
- But also assume that this family is from the same distribution
- Coin flipping: we assumed the same null hypothesis for all 100 coins, i.e. that they're fair. What if 50 are biased and 50 are fair?

Correcting for multiple tests

Bonferroni correction

- Controls probability of at least one false positive (FWER)
- May result in many false negatives. Why?
- Main idea: for overall (experiment-wise) α to be 0.05, need individual tests to be stricter

Bonferroni correction

- Recall that:

$$\alpha_e = \alpha_c \times N \tag{11}$$

- So, if we want a particular α_e we can rearrange this to find the correct α_c

$$\alpha_c = \frac{\alpha_e}{N} \tag{12}$$

- This means that if we want our significance level to be $\alpha_e = 0.05$, and we're doing $N = 100$ tests, each one needs to be conducted with:

$$\alpha_c = 0.05/100 = 0.0005 \tag{13}$$

More on the Bonferroni correction

If the tests are independent, this will give us an α of much less than our desired 0.05. Why? Recall that when tests are independent:

$$1 - (1 - \alpha)^N \tag{14}$$

But that the bonferroni correction uses:

$$\alpha \times N \tag{15}$$

For $\alpha = 0.05$ and $N = 100$, this gives 1 and 5 respectively. We want to control false discoveries, but don't want to overestimate these, leading to making few discoveries.

Controlling the FDR

- Bonferroni focused on probability of making *any* false discoveries (FWER)
- But compare:
  - 10 tests, 2 false discoveries
  - 100 tests, 2 false discoveries
- It's much more serious to have 20% FDR than 2% FDR
- Now, we focus on the proportion of false discoveries out of all discoveries: controlling the FDR.
- For large scale testing (such as with DNA microarrays), FDR is much better measure

Methods for controlling FDR: Benjamini Hochberg[1]

- Order the m p-values so $p_1 < p_2 < \ldots p_m$
- Then with k being the largest i such that:

$$P_{(i)} \leqslant \frac{i}{m}\alpha, \tag{16}$$

- We reject all $H_{(i)}, i = 1, 2, \ldots k$.
- This controls FDR at rate α when tests are independent or positively correlated.

[1] Benjamini and Hochberg. *Controlling the false discovery rate: a practical and powerful approach to multiple testing* (1995)

Benjamini-Hochberg correction example[2]

Let's say we have 15 comparisons, with the following ordered p-values:
0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000.

- If we control FWER at 0.05 with Bonferonni, we have 0.05/15=0.0033
  This means we should reject the first three null hypotheses
- Now using BH, start with $p_{(15)}$ and calculate:

$$\text{is } 1 \leqslant \frac{15}{15} 0.05 = 0.05 \tag{17}$$

[2] Taken from Benjamini & Hochberg (1995)

Benjamini-Hochberg correction example

Let's say we have 15 comparisons, with the following ordered p-values:
0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000.

- We test each in turn:

$$\text{is } p_{(5)} = 0.0201 \leqslant \frac{5}{15} 0.05 = 0.017$$

- Finally, the first that satisfies the constraint:

$$p_{(4)} = 0.0095 \leqslant \frac{4}{15} 0.05 = 0.013$$

- So, we reject the null hypotheses corresponding to the first 4 tests. With Bonferroni, rejected only first 3.

q-values[3]

- Introduces new measure, *q*-value, focusing on the fact that we expect many positives in such large studies
- Examples:
  - Detecting differentially expressed genes: use microarrays to find genes differentially expressed between tumor types
  - Genetic dissection of transcriptional regulation: find relationship between markers and gene expression

[3] Storey & Tibshirani. *Statistical significance for genomewide studies* (2003)

Observation

Since we test so many hypotheses $0.05m$ is much too large. To get around this, people frequently use much lower values for α, and still receive many positives, likely still allowing many false discoveries. FDR is much more useful than FWER, but want a measure of significance associated with *each* feature

q-value

- Order p-values, then if reject null for some p', reject all with $p \leqslant p'$
- q-value for a particular feature is expected proportion of false positives if that feature is called significant
- Calculate q for each feature, then thresholding $q = \alpha$
- Main idea is that we're assessing each feature individually, so we can compare how significant each is

p versus q:

- p-value: probability of a null feature being at least as extreme as observation
- q-value: expected proportion of false positives among all features at least as extreme as observed one. Or: the minimum FDR when we call this feature significant. At $q = 0.05$, this means that of all the features with q less than the current feature, 5% are false positives
- However. . . a gene near the edge of null/not-null will be seen as less likely than it should to be a false positive (since the more significant ones are so unlikely, they keep down the FDR). For some test with $q = 0.05$, that particular q has a higher than 5% chance of being false, since the ones with smaller q-values are likelier to be true positives.

Local FDR[4]

- Can we use a similar approach as for q-values, with FDRs?
- While q-values were specific to each test, the results still considered the entire tail
- What we really want is to look at each individual result and see how much it differs from our expectations
- We can do this by calculating the fdr locally: probability of a hypothesis being null, conditional on its test statistic
- Caveat: assume N at least in hundreds, but don't need independent tests

[4] Bradley Efron. *Local false discovery rates* (2005)

Definition

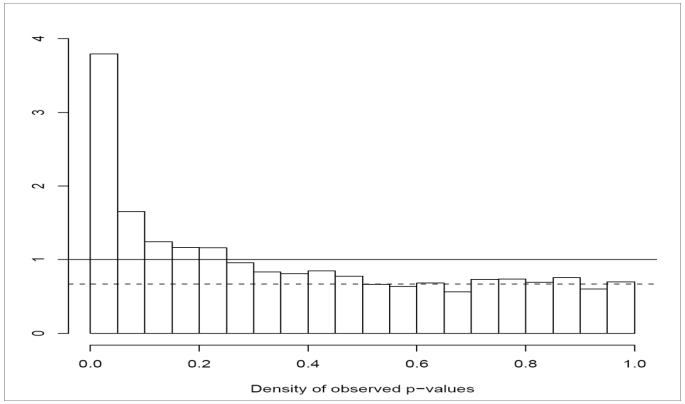
- The local false discovery rate, fdr, is defined as:

$$fdr(z) \equiv P\{null|z\} \tag{18}$$

- Relation to q-value:
  fdr will generally be larger than q, assuming fdr decreases as z increases.

Where do nulls come from?

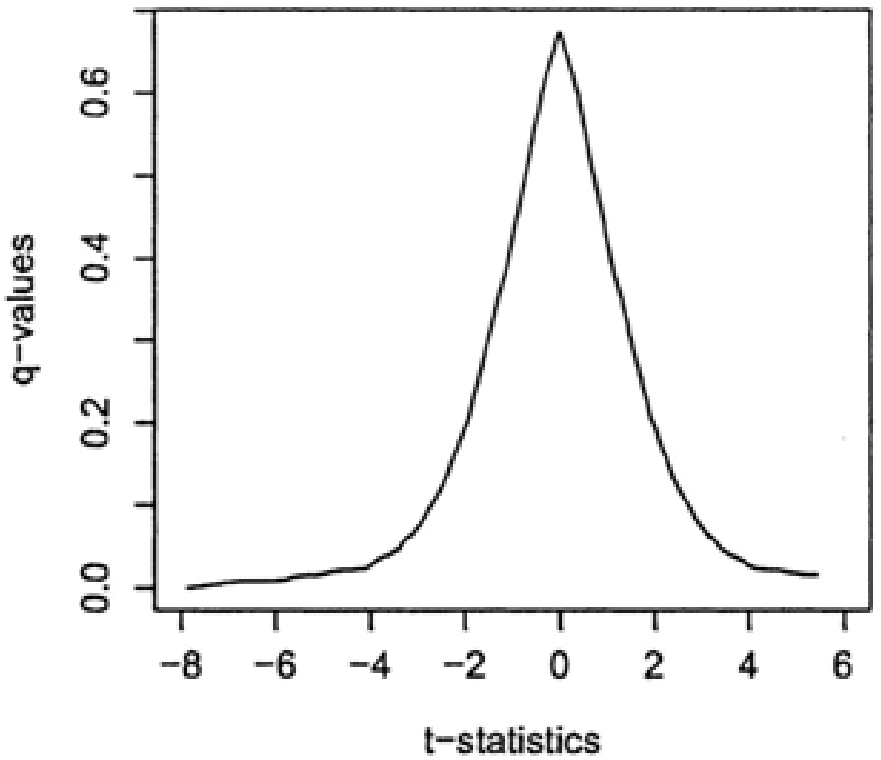
- Coin flipping: clear what should happen if coin is fair
- Microarrays, testing whether gene activities are correlated: not so clear what should happen
- Storey & Tibshirani: assume nulls uniformly distributed



More on nulls

- Frequently permute data (scramble the data between two tumor types, then compute test statistics) - but this is computationally very expensive - imagine thousands of genes and multiple microarrays. Also, if there is dependence between any of the microarrays, this won't work.
- New method: get the null from the data, empirically

T-statistics vs q-values:



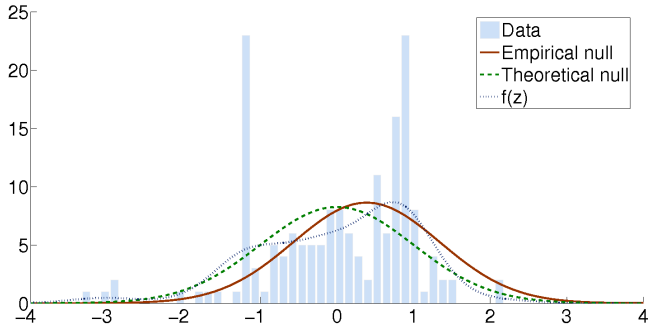
(From Storey & Tibshirani, 2003)

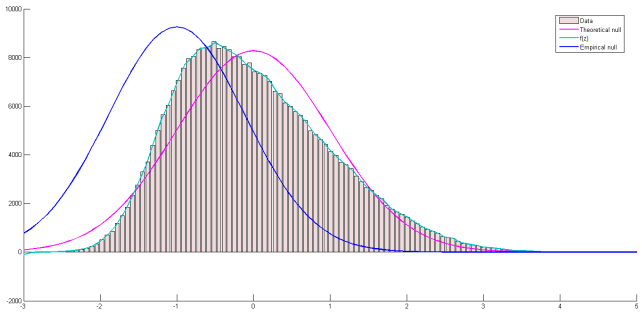
The empirical null hypothesis

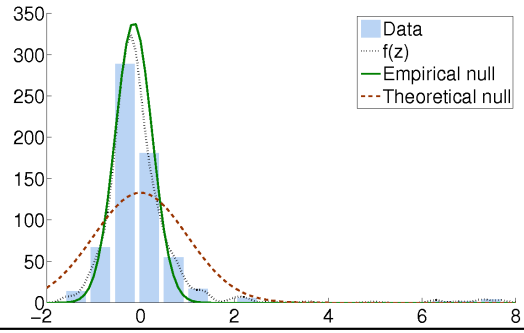
- Basic assumption: if all hypotheses are null, our test statistics should follow a normal distribution
- Deviations from this null indicate significant results
- When there are some non-nulls, then our observation is the mixture of two distributions: One normal, giving the nulls, and one other distribution for the non-null results.
- Find, from the results, what the null should be, then compare results to that
- Where there is a large deviation from what is expected with the null, call those results significant (reject the null hypothesis)

Varying nulls

Theoretical null: results will fall within a normal distribution with mean 0 and standard deviation 1[5].
Empirical null: Inferred from data







[5] $std = \sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/N}$

Multiple testing with an empirical null[6]

Assume two classes, with prior probabilities:

$$p_0 = P(\text{null}) \tag{19}$$
$$p_1 = P(\text{non} - \text{null}) \tag{20}$$

- Densities $f_0(z)$ and $f_1(z)$ describe the distribution of these classes
- When using theoretical null $f_0 = N(0, 1)$
- Assume p_0 much larger than $p_1 = 1 - p_0$, perhaps 0.90

[6] Bradley Efron. *Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis* (2003).

Defining the FDR

- With both classes together we have the mixture:

$$f(z) = p_0 f_0(z) + p_1 f(z) \tag{21}$$

- False discovery rate is prob of case being null, given its test-statistic:

$$fdr(z) = P(i = null | z_i = z), \tag{22}$$

which is:

$$p_0 f_0(z)/f(z) \tag{23}$$

- Since p_0 assumed close to one, can use:

$$f_0(z)/f(z) \tag{24}$$

(continued)

- Then we will calculate $f(z)$ from observations (by fitting to the data, for example with a spline fit) and now "only" need to estimate $f_0(z)$
- Reject null for:

$$f_0(z)/f(z) \leqslant \alpha \tag{25}$$

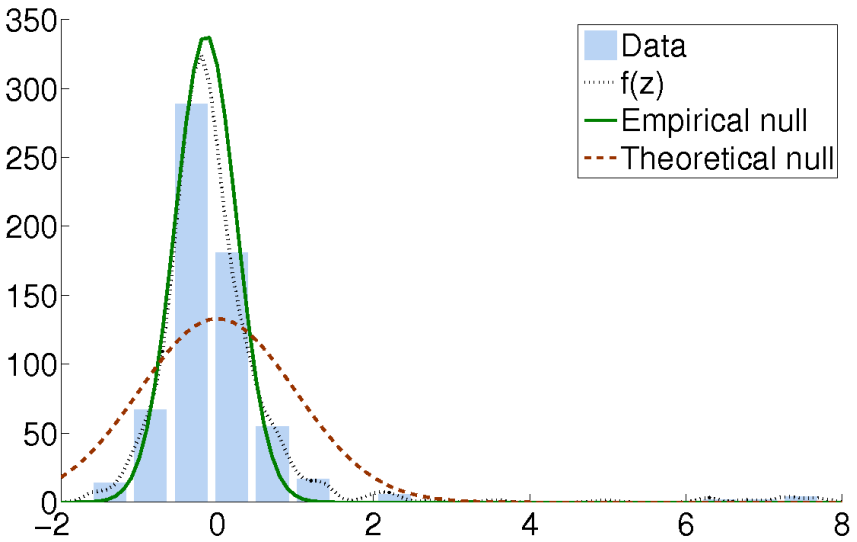
- Note that what we're computing is the fdr for each z. This is the local fdr.
- As number of features tends toward ∞, fdr approaches FDR

Inferring the null

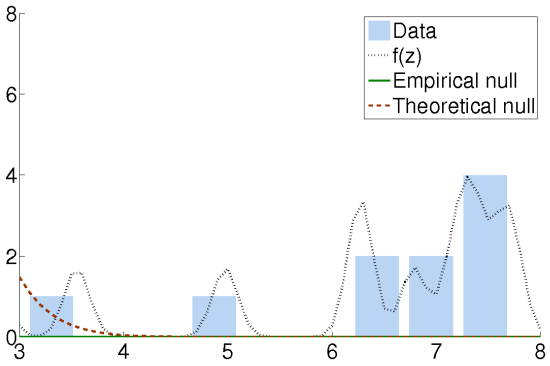
- Observation: If z's normally distributed, then there's a central peak
- Assume f_0 is given by $N(\mu, \sigma)$, so we must find μ and σ
- Most methods look at area around $z = 0$, testing density of results to find the peak.

Overview of procedure

- Main idea: Histogram of test statistics, for each bin figure out if it's bigger than expected
- Here there are 642 hypotheses, with the empirical null $N(0.39, 0.96)$



Up close



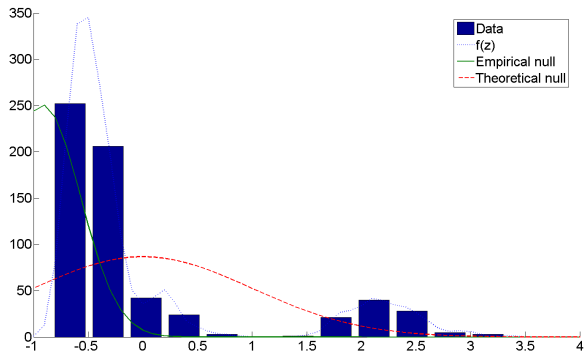
$$fdr(3) = 0/2 = 0 \tag{26}$$

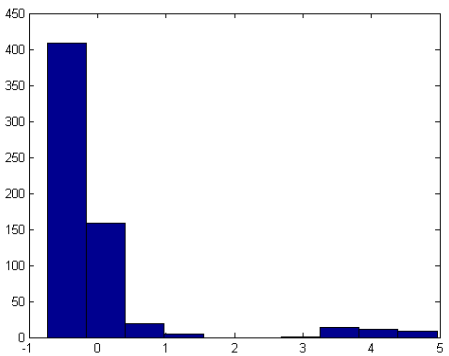
There were no nulls expected with $z = 3$
Expected count for z is: $bin_w \times f_0(z) \times N$ (N is number of hypotheses tested, $f_0(z)$ is a norm pdf from the inferred mean/std

More on empirical null

- The good: tests don't need to be independent, don't need to know the null
- The bad: if the underlying distribution is not normal, you're out of luck, also falls apart when true positives are a not-insignificant fraction of all hypotheses tested

These are not normally distributed and are fit poorly:





Recap

- When you choose a procedure, be sure it's controlling what you want to control: false positives or false negatives, overall all tests or probability of at least one
- Be aware of the assumptions: if method controls when all tests independent, be sure your tests are independent!

Probabilities and frequencies

- Probability: number between 0 and 1 that tells how likely an outcome is
- For the set of all (mutually exclusive) outcomes, the probability adds to one:
  e.g. A coin can be heads or tails $P(H) = P(T) = 1/2$.
  $P(H) + P(T) = 1$
  Mutually exclusive means we can't have both H and T at the same time.
- This corresponds to how often we will observe each outcome
- If we flip a coin 10,000 times, roughly $1/2$ the flips should be heads and $1/2$ should be tails

Conditional probability and independence

- Probability of B conditional on A:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \tag{27}$$

- Independence:

$$P(A \wedge B) = P(A)P(B) \tag{28}$$

- Then, if A and B are independent:

$$P(B|A) = \frac{P(A)P(B)}{P(A)} = P(B) \tag{29}$$

- This means that A doesn't tell us anything about B. A coin coming up heads on the previous flip doesn't change the probability that it will come up tails on the next flip (unless the coin is biased)

More on probability

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \tag{30}$$

if A and B mutually exclusive (i.e. H and T), $P(A \wedge B) = 0$ so:

$$P(A \vee B) = P(A) + P(B) \tag{31}$$

$$P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A) \tag{32}$$

Multiple trials

Now what is the probability of getting at least one H in N coin flips? Since each flip is independent, we can calculate:

$$P(>1H) = 1 - P(\text{no } H \text{ in } N \text{ flips}) \quad = 1 - P(\text{no } H \text{ in one flip})^N \tag{33}$$

Probability of not getting heads is $1/2$, so this is:

$$1 - (1/2)^N \tag{34}$$

If $N = 2$, $P = 0.75$, but if $N = 10$, $P \approx 1$