Lecture 14: Multiple testing

Rebecka Jörnsten, Mathematical Sciences

MSA220/MVE441 Statistical Learning for Big Data

12th May 2022

Statistical Testing - recap

Statistical testing

Every statistical test is associated with the risk that you false declare a finding (a false positive, a false rejection of a null hypothesis).

We pick the *level* of our statistical test to safe-guard this from happening at some acceptable level of risk.

Terminology:

- Data X which is random (e.g. a vector, two vectors, a summary statistic like a mean, an estimated coefficient,...)
- Test statistic $T(X)$ which is random through X (e.g. a z-score, t-value etc)
- Null hypothesis: You assume something about the data, e.g. that the mean is 0, that the true model coefficent is 0, ...

Statistical testing

Terminology:

- Test statistic $T(X)$ which is random through X (e.g. a z-score, t-value etc)
- Null hypothesis: You assume something about the data, e.g. that the mean is 0, that the true model coefficent is 0, ...
- Under the null we can work out the distribution for T explicitly (e.g t-distribution with associated degrees of freedom) OR...
- ... we can generate the null distribution through simulation.
  Example: Permutation test
  - Testing difference of mean of X between two classes
  - Permute the labels and re-compute the test-statistic
  - Repeat $B \sim 1000$ times and compare the distribution of test-statistics under permutation to the original.

Statistical testing

More terminology:

- Test statistic $T(X)$ which is random through X (e.g. a z-score, t-value etc)
- Distribution (CDF) of T under the null

$$P(T \leq t \mid H_0)$$

  where H_0 refers to the null hypothesis
- Alternative hypothesis H_1.
  - This is usually very open: e.g. $H_1 : \beta_j \neq 0$
  - It can refer to a subset of model coefficients being non-zero and some non-zero.

Statistical testing

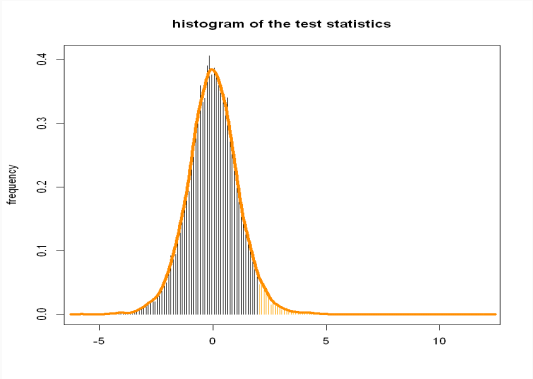
- Distribution (CDF) of T under the null

$$P(T \leq t \mid H_0)$$

  where H_0 refers to the null hypothesis
- Level of the test α: threshhold for the test statistic.
- If we observe T above this threshhold we reject H_0, otherwise we *fail to reject*
- Note: we can never prove or accept an alternative hypothesis since we have not worked out the distribution for the test statistic under this assumption.
- Note: I am using a one-sided test her for ease of presentation/visualization.

Statistical testing

- Level of the test α: threshhold for the test statistic.
- If we observe T above this threshhold we reject H_0, otherwise we *fail to reject*
- p-value: we compute the probability mass of the pdf $p_T(t)$ for observed T *or values even more extreme*



Multiple Testing

Multiple testing

Example: the South African heart disease data

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.0760913  1.3404862  -5.279  1.3e-07 ***
sbp          0.0065040  0.0057304   1.135 0.256374
tobacco      0.0793764  0.0266028   2.984 0.002847 **
ldl          0.1739239  0.0596617   2.915 0.003555 **
adiposity    0.0185866  0.0292894   0.635 0.525700
famhist      0.9253704  0.2278940   4.061  4.9e-05 ***
typea        0.0395950  0.0123202   3.214 0.001310 **
obesity     -0.0629099  0.0442477  -1.422 0.155095
alcohol      0.0001217  0.0044832   0.027 0.978350
age          0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5
```

Here, there are 5 significant coefficients but you are actually performing 9 tests (9 features in total).

- Using level $\alpha = 0.05$ means each test as a probability of 5% of generating a false rejection.
- Across the 9 features, the probability of making at least one false positive is

$$P(\text{At least one false positive}) =$$

$$= 1 - (1 - \alpha)^9 \simeq 0.37$$

Multiple testing

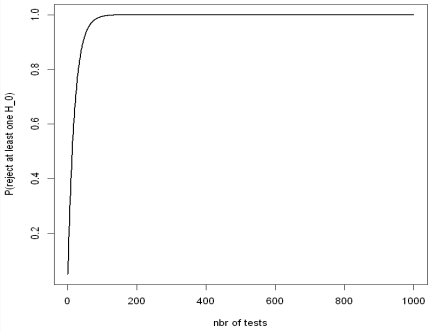
Multiple testing problem: If I test n true null hypotheses at level α, then on average we can expect to falsely reject $\alpha \times n$ of them.

Common problem:

- Test whether a gene's expression is linked to disease across 10000+ genes.
- Detection of a server attack in a large network (anomaly detection)
- fMRI - detection of "active" regions (pixel level test)
- Really: most studies involve multiple testing but perhaps at the modest scale of 10 tests like the heart disease example.
- Already with 10 tests you are very likely to encounter at least one false positive and .....

Multiple testing

- Most studies may involve multiple testing but perhaps at the modest scale of 10 tests like the heart disease example.
- Already with 10 tests you are very likely to encounter at least one false positive and .....
- once we reach 100 tests this probability reaches 99%!



Type I and II errors

	H_0 true	H_0 false	Total
Reject H_0	V	S	R
"Accept H_0	U	T	$n - R$
	n_0	$n - n_0$	n

- R = number of rejected H_0 (our "findings")
- V = number of type I errors (our false rejections)
- T = number of type II errors (our missed detections)

Family wise error rate, FWER

	H_0 true	H_0 false	Total
Reject H_0	V	S	R
"Accept H_0	U	T	$n - R$
	n_0	$n - n_0$	n

- $FWER = P(V \geq 1)$
- This is what was illustrated in the figure on the previous slide
- How can we reduce this risk?
- What if we adjust the level of the test to reflect that we are performing multiple tests?

FWER

	H_0 true	H_0 false	Total
Reject H_0	V	S	R
"Accept H_0	U	T	$n-R$
	n_0	$n-n_0$	n

- $FWER = P(V \geq 1)$
- Let us adjust the level α to α/n
- This is called the Bonferroni correction and controls FWER at level α regardless of the number of true null hypotheses n_0

Bonferroni correction

Consider testing n different null hypotheses H_0^j, $j = 1, \cdots, n$, all of which are, in fact, true. We want to control

$$P(\text{reject at least (any) hypothesis}) \leq \alpha$$

Bonferroni method:

- Perform each test at significance level α/n, instead of level α.

$$P(\text{reject any null hypothesis} = P(V \geq 1) =$$

$$= P(\text{reject } H_0^1 \cup \cdots \text{ reject } H_0^n) = \leq P(\text{reject } H_0^1) + \cdots + P(\text{reject } H_0^n) =$$

$$= \alpha/n + \cdots + \alpha/n = \alpha$$

Bonferroni correction

Bonferroni controls the FWER regardless of how many hypothesis are in fact true nulls, n_0.

- Perform each test at significance level α/n, instead of level α.

$$P(\text{reject any null hypothesis} = P(V \geq 1) =$$

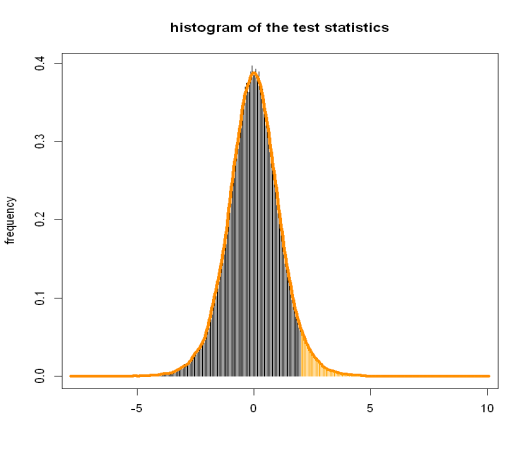
$$= P(\text{reject } H_0^1 \cup \cdots \text{ reject } H_0^{n_0} =\leq P(\text{reject } H_0^1) + \cdots + P(\text{reject } H_0^{n_0}) =$$

$$= n_0 \alpha/n \leq \alpha$$

- Adjust level of the test: $\alpha_n = \alpha/n$
- Adjusted p-value: $p.adj = n \cdot p.val$

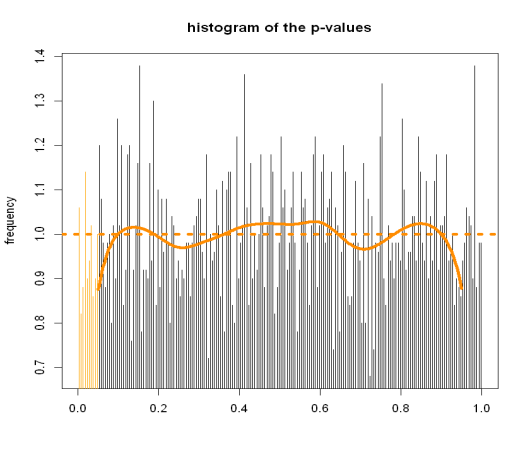
More about p-values

- Reject hypothesis j if p-value $p_j \leq \alpha/n$
- What do p-values from a set of tests look like?
- Fact: a p-values under the null is distributed as $U[0, 1]$



More about p-values

- Reject hypothesis j if p-value $p_j \leq \alpha/n$
- What do p-values from a set of tests look like?
- Fact: a p-values under the null is distributed as $U[0,1]$



More about p-values

- Reject hypothesis j if p-value $p_j \leq \alpha/n$
- What do p-values from a set of tests look like?
- Fact: a p-values under the null is distributed as $U[0,1]$

Let's say we reject a hypothesis if the test statistic T is large. The p-value is the upper tail of the CDF F of T

$$P(\text{p-value } \leq t) = P(1 - F(T) \leq t) = P(1 - t \leq F(T)) =$$

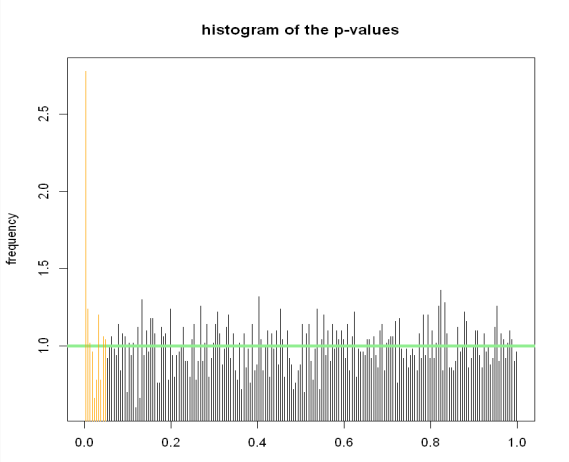
$$= P(F(T) \geq 1 - t) = 1 - P(F(T) \leq 1 - t) =$$

$$= 1 - P(T \leq F^{-1}(1 - t)) = 1 - F(F^{-1}(1 - t)) = 1 - (1 - t) = t$$

I.e. the p-value is uniformly distributed.

Back to testing

- Reject hypothesis j if p-value $p_j \leq \alpha/n$
- What do p-values from a set of tests look like when some of the nulls are false?



Back to testing

	H_0 true	H_0 false	Total
Reject H_0	$V = 477$	$S = 100$	$R = 577$
"Accept H_0	$U = 9423$	$T = 0$	$n - R = 9423$
	$n_0 = 9900$	$n - n_0 = 100$	$n = 10000$

- The good news is I found every non-null
- The bad news is that I made lots of false rejections

Back to testing

- What if we use the Bonferroni correction
- Use level $\alpha/10000$ here

	H_0 true	H_0 false	Total
Reject H_0	$V = 1$	$S = 16$	$R = 17$
"Accept H_0	$U = 9899$	$T = 84$	$n - R = 9983$
	$n_0 = 9900$	$n - n_0 = 100$	$n = 10000$

- Hmm... a bit too cautious perhaps?

Controling False Discovery Rate

FPR and FDR

	H_0 true	H_0 false	Total
Reject H_0	V	S	R
"Accept H_0	U	T	$n-R$
	n_0	$n-n_0$	n

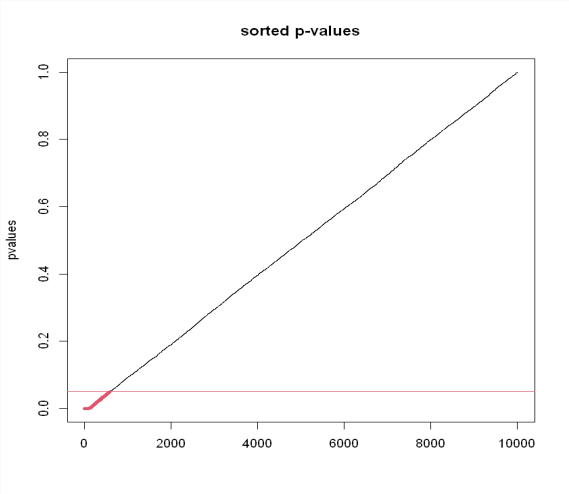
- $FPR = V/n_0$ false positive rate
- $FDP = \frac{V}{R}1[R \geq 1]$ false detection proportion
- $E(FDP) = FDR$ is the false discovery rate
- Benjamini-Hochberg (BH) procedure compares the sorted p-values to a diagonal cutoff line with a slope q, finds the largest p-value that still falls below this line, and rejects the null hypotheses for the p-values up to and including this one.

FPR and FDR

- The FDR (false discovery rate) has gained a lot of traction because pracitioners have found Bonferroni to be too conservative
- (There are also alternative FWER controlling methods that are less conservative)
- The Benjamini-Hochberg (BH) procedure compares the *sorted p-values* to a diagonal cutoff line with a slope q.
  We find the largest p-value that still falls below this line, and rejects the null hypotheses for the p-values up to and including this one.

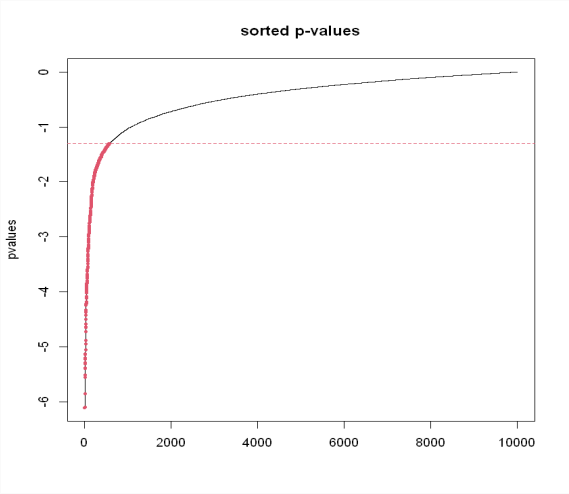
Back to testing

- Sorted p-values with the 0.05 threshhold



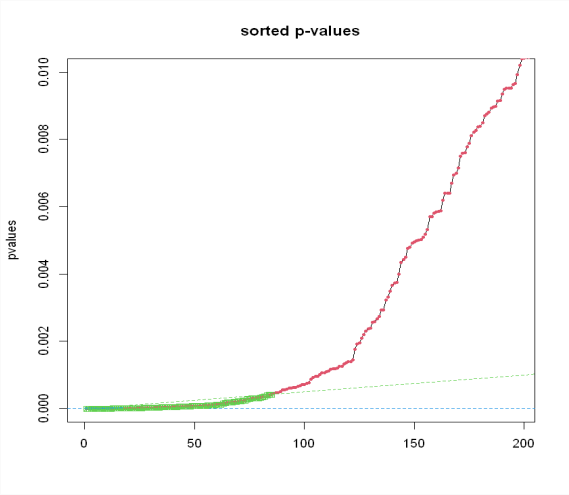
Back to testing

- It is common to plot p-values on a log10 scale since the thresholds we often use (0.01, 0.001) correspond to levels of the log10 plot.



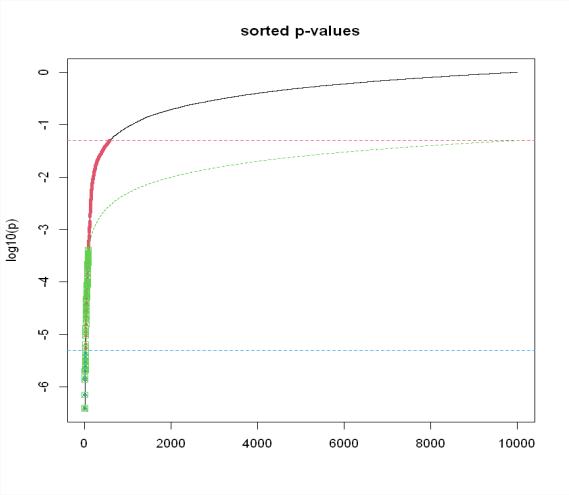
FDR control

- Sorted p-values with the 0.05 threshhold in red, Bonferroni i blue
- BH in green (slope $q = 0.05$ cutoff)



FDR control

- Sorted p-values with the 0.05 threshhold in red, Bonferroni i blue
- BH in green (slope $q = 0.05$ cutoff)



FDR control

Formally, the BH procedure at level q is defined as follows:

- Sort the p-values. Call them $P_{(1)} \leq ... \leq P_{(n)}$
- Find the largest r such that $P(r) \leq q(r/n)$
- Reject the null hypotheses $H_{(1)}, ..., H_{(r)}$.

Benjamini and Hochberg (1995)): Consider tests of n null hypotheses, n_0 of which are true. If the test statistics (or equivalently, p-values) of these tests are independent, then the FDR of the above procedure satisfies $FDR \leq \frac{n_0 q}{n} \leq q$.

Note: FDR control is not guaranteed if the test statistics are dependent.
q is thus our acceptable level of the false discovery rate. This might be higher than a common choice for α. Think of this in terms of follow-up experiments. How many uninformative follow-up experiments are you willing to run?

Back to testing example

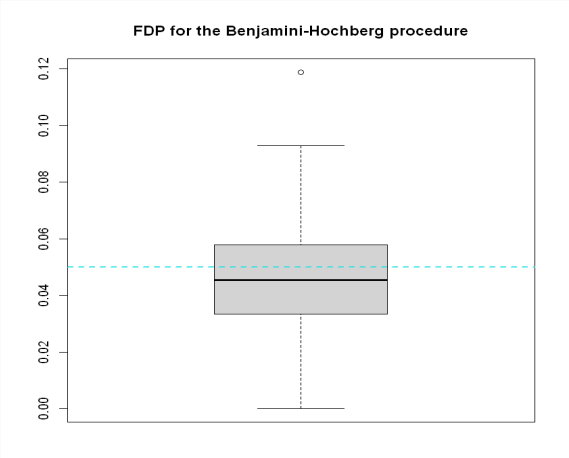
- What if we use the BH correction in our example from above?

	H_0 true	H_0 false	Total
Reject H_0	$V = 3$	$S = 82$	$R = 85$
"Accept H_0	$U = 9897$	$T = 18$	$n - R = 9915$
	$n_0 = 9900$	$n - n_0 = 100$	$n = 10000$

- The observed FDP is 0.035
- However, you can observe values over α
- You only control the FDR in *expectation*

E(FDP) for the BH procedure

- We repeat the simulation several times and record the observed FDP values
- We observe that the expected value of the FDP is below the threshhold 0.05 (dashed line)



The Benjamini-Hochberg procedure

- For each $\alpha \in (0, 1)$, let $M(\alpha)$ be the number of p-values $\leq \alpha$.
- Using a level of α and rejecting all hypotheses with p-values $\leq \alpha$ means we can expect to falsely reject $n_0 \cdot \alpha$ null hypotheses, since the null p-values are distributed as $U(0, 1)$.
- We estimate the false discovery proportion as

$$FDP = n_0 \cdot \alpha / M(\alpha)$$

Hang on! We don't actually know n_0.
However, we can obtain a conservative upper-bound from

$$FDP < n \cdot \alpha / M(\alpha)$$

The Benjamini-Hochberg procedure

- We estimate the false discovery proportion as

$$FDP = n_0 \cdot \alpha / M(\alpha)$$

- A conservative upper-bound n_0 is n
- We set $\alpha = P_{(r)}$, the r-th largest p-value. Then

$$FDP < n \cdot \alpha / M(\alpha) \leq q$$

with equality when

$$P_{(r)} \leq q \cdot r / n$$

Another way of thinking about this:

- So the BH procedure chooses α (in a data-dependent way) so as to reject as many hypotheses as possible, subject to the constraint

$$FDP < n \cdot \alpha / M(\alpha) \leq q$$

Adjusted p-values

We mainly talked about how to utilize the adjustments to test at a level α. However, the procedures we talked about can also be used to *adjust* the p-values to be used with a level selected later.

- Bonferroni: $p.adj^B = min(1, p.raw * n)$
- Benjamini-Hochberg:
  - sort the p.values: $p.raw(j), j = 1, \cdots, n$
  - BH procedure states we should reject hypothesis j if $p.raw(j) < \alpha(j/n)$ where j denotes the rank (lowest to highest)
  - That means we reject if $(np.raw(j))/j < \alpha$
  - Adjusted p-value

$$p.adj^{BH}(j) = \frac{p.raw(j)n}{j}$$

You can report the adjusted p-values instead or with the raw ones for later inference with a chosen α.

Take-home message

- If you perform many tests you all but guaranteed to get false positives
- If your study leads to follow-up experiments or studies you may need to control these false positives - use multiple testing corrections
- Sometimes it's more important to control the proportion of false positives among your detection - use a less aggressive adjustment and metric, the FDR (false discovery rate).
- Caveats:
  - Are you using the right test?
  - Where did the p-values come from? Perhaps you need to use non-parametric approaches like permutations or bootstraps to obtain them
  - In regression and anova there are also other post-processing procedures for pairwise comparisons etc.
  - Careful when the sample size is large... upcoming lectures