Statistical Modelling and Inference

Barcelona GSE, Fall 2015

Prof: O. Papaspiliopoulos

Note the difficulty levels 1,2,3, which are grading the difficulty of each exercise from easy (1, necessary for pass), to medium (2, required for obtaining 70-80%) to harder (3, required for a first class mark).

Project work: you will receive in total 20 marks for project. This project should be handed in by Friday 28/10/16

Project: Multiple testing

1 References

1. Wasserman Chapter 1; Background on basic probability
2. Wasserman Chapter 10 on testing; especially 10.7 on multiple testing
3. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, Benjamini and Hochberg, J. R. Statist. Soc. B (1995) 57
4. Multiple Hypothesis Testing in Microarray Experiments, Dudoit, Shaffer and Boldrick, Statistical Science 2003, Vol. 18
5. Stability, Bin Yu, Bernoulli 2013
6. Why most published research findings are false, Ioannidis, J.P.A., PLoS Med., 2005

2 Hypothesis testing

2.1 Test statistics and calibration of tests

We consider a (reasonably a priori) scientific hypothesis, that we will call H throughout, about a phenomenon that generates data $\mathbf{X}$. The type of hypotheses we consider here imply certain properties for the distribution of the data. We will denote by $\mathbf{Y}$ a dataset randomly drawn from a distribution consistent with H; therefore, if H holds, $\mathbf{X}$ is a realisation of such a random dataset, but if H does not hold we would expect draws of $\mathbf{Y}$ to "look" different from $\mathbf{X}$.

For multi-dimensional datasets we need appropriate tools to measure how $\mathbf{X}$ "looks" relative to realisations of $\mathbf{Y}$ from H. We test for the support of the data in favour of H by computing a scalar test statistic $T(\mathbf{X})$ and fixing a set of values, say R_α, such that

$$\mathbb{P}_H[T(\mathbf{Y}) \in R_\alpha] = \alpha$$

where P_H denotes the probability distribution of the data $\mathbf{Y}$ when H is true. The above setting fixes the probability, under H, that $T(\mathbf{Y})$ takes a value in R_α to be α, where typically α is fairly small, say 0.05 or 0.01 (more serious Statistics sets α in relation to the impact that a false positive might have in a decision that relies on the hypothesis test).

Given a test statistic T and a region R_α as given above, we reject H at level α for the given dataset $\mathbf{X}$, if $T(\mathbf{X}) \in R_\alpha$. If we follow this procedure repeatedly (frequentist Statistics) with random datasets $\mathbf{Y}$ drawn from a distribution consistent with H, we would be rejecting H incorrectly (false positive) $100\alpha\%$ of the times.

α is known as the size of the test.

2.2 P-value

The p-value for a given test statistic T, hypothesis H and dataset $\mathbf{X}$ is the *smallest* level α for which H would be rejected:

$$p - value = \inf\{\alpha : T(\mathbf{X}) \in R_\alpha\}\,.$$

1. **Distribution of the p-value under the null**

   Suppose that the hypothesis H is tested, using data $\mathbf{X}$, a test statistic T, and a critical region $R_\alpha = (c_\alpha, \infty)$, that is, reject if $T(\mathbf{X}) > c_\alpha$, where c_α is chosen so that $\mathbb{P}_H[T(\mathbf{Y}) > c_\alpha] = \alpha$. Let F_H denote the distribution of $T(\mathbf{Y})$ under H. Then, c_α is chosen as the solution to the equation $\alpha = 1 - F_H(c_\alpha)$. In what follows we assume that F_H is a continuous distribution function. Since it is also increasing, the equation $\alpha = F_H(x)$ has the unique solution $x = F_H^{-1}(\alpha)$, for any $\alpha \in (0, 1)$, where F_H^{-1} denotes the inverse function.

   The aim is to show that the distribution of the associated p-value under H is uniform. We do this in various stages.

   1. [**1**] Show that for any α, $c_\alpha = F_H^{-1}(1 - \alpha)$
   2. [**1**] Show that the p-value of the test, as a function of the data $\mathbf{X}$ used, is given by $p(\mathbf{X}) = 1 - F_H(T(\mathbf{X}))$.
   3. [**2**] Show that for any univariate random variable y with continuous distribution function F, the random variables $F(y)$ and $1-F(y)$ follow the uniform distribution.
   4. [**1**] Using the above results, show that the p-value follows the uniform distribution under H.

Remark: convince yourselves that the result in 3 is not true if Y follows a discrete distribution. Consider for example $y \sim Bernoulli(1/2)$.

Further reading:

1. *The result in 3 has a reverse which is the key in simulating random variables. If $u \sim U(0,1)$, and F is some distribution function (not necessarily continuous), then $F^{-1}(u) \sim F$. See Computing Lab.*

The result suggests one rule for obtaining a test of size α: reject when p-value is less than α

3 Multiple testing (MT)

Consider now the situation of many separate tests H_i, for $i = 1, \ldots, m$. Several modern scientific applications involve testing several hypotheses: e.g. testing whether elements in a covariance (or inverse covariance) matrix are 0; testing whether regression coefficients are 0; testing for difference between measurements in variables in case-control studies. In modern applications where thousands of different variables, hence hypotheses, are considered, multiple hypotheses is more the norm than the exception.

The *complete null* hypothesis is that all H_i's are true.

3.1 MT under independence assumptions

2. **Multiple testing under independence assumptions**

Recall that two events, A_1 and A_2, are independent (under an assumed probability model $\mathbb{P}$) if $\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1]\mathbb{P}[A_2]$. Similarly, recall that if A^c is the complementary to A event, $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$. Finally, recall that under the null hypothesis, the distribution of the p-value is uniform.

Our aim is to show that when $m \geq 1$ tests are carried out independently, each with significance level α, the overall probability of rejecting at least one of the hypotheses, when all of them are true, is $1 - (1-\alpha)^m$. We then discuss to find how to set each test to control this probability. We do this in various stages.

1. **[1]** Let $y_1, \ldots, y_m$ be m independent uniform random variables. Show that the probability that they are all larger than α is $(1-\alpha)^m$, that is, show that $\mathbb{P}[\cap_{i=1}^m \{y_i > \alpha\}] = (1-\alpha)^m$.
2. **[2]** Apply the above result to show that when we perform m independent tests with datasets $\mathbf{X}_1, \ldots, \mathbf{X}_m$, which yield p-values $p_1(\mathbf{X}_1), \ldots, p_m(\mathbf{X}_m)$, under the complete null the probability of rejecting at significance level α at least one of them is $1 - (1-\alpha)^m$.

3. [**1**] Under the above assumption, show that if we wish that the overall type I error is α, each independent test should be rejected at significance level $1-(1-\alpha)^{1/m}$.
4. [**1**] Plot on the same axes the functions $f_1(\alpha) = 1-(1-\alpha)^{1/m}$, $f_2(\alpha) = \alpha/m$ and $f_3(\alpha) = \alpha$, for $\alpha \in [0,1]$ and check that $f_2 \leq f_1 \leq f_3$.
5. [**(optional)**] Show mathematically that $f_2 \leq f_1 \leq f_3$.
6. [**1**] Conclude that, under the independence assumptions, the level needed for each test can be determined and it is smaller than the approach that makes no correction for multiple testing.

The conclusions from the previous exercise are two:

1. When several tests are carried out simultaneously the level of each will have to be adjusted to something smaller than α is we wish the overall probability of rejecting incorrectly at least one hypothesis to be α. If we do not make each test stricter, i.e. reduce its own significance level, we would be committing a multiple testing error, which means that the probability of rejecting incorrectly at least one hypothesis will be (potentially much) bigger than α. If we associate rejection with discovery (usually the null is the status quo and its rejection means evidence for something new) then multiple testing error amounts to false discoveries, only due to the multiple testing practice, not to real evidence.
2. If we make the assumption that the tests are independent of each other then we know precisely how to correct each to achieve an overall level α

The assumption of independence is OK as a working assumption to get started but it is entirely unrealistic. Often all test refer to the same dataset $\mathbf{X}$; e.g. test of regression coefficients, or correlation coefficients. Therefore, we should work harder to obtain a multiple testing correction that does not make this assumption.

3.2 A conservative but robust test: Bonferroni

In what follows we will write $C-H$ for the complete null hypothesis. Additionally, $p_i(\mathbf{X}_i)$ will denote the p-value of the i'th test based on data $\mathbf{X}_i$, although we might have that $\mathbf{X}_i = \mathbf{X}_1$ for all i, hence we will not be making assumptions about the joint distribution of the datasets, hence the p-values.

The probability of at least one rejection under the complete null is

$$\mathbb{P}_{C-H}\left[\cup_{i=1}^{m}\{p_i(\mathbf{Y}_i) < \alpha\}\right].$$

3. **Upper-bounding the probability of at least one rejection** [**2-3**] Show that

$$\alpha \leq \mathbb{P}_{C-H}\left[\cup_{i=1}^{m}\{p_i(\mathbf{Y}_i) < \alpha\}\right] \leq m\alpha$$

The above exercise leads to an important method: the Bonferroni correction for multiple testing. If m tests are carried out and we wish the overall significance to be α (or smaller),

we should reject each individual test at level α/m. Therefore, we become much stricter for each test, the stricter the more tests are being considered. In plain words, the more projects we consider that could lead to a new discovery, the more evidence we require for the results of each such project to be classified as discovery.

Note (as discussed earlier) that

$$\alpha/m \leq 1 - (1-\alpha)^{1/m}$$

hence the Bonferroni correction is more conservative than the one based on the assumption of independence tests: makes sense, Bonferroni makes less assumptions hence sets a lower threshold. But it is also robust to situations where independence does not hold.

3.3 Ordered p-values, family-wise error rate and a new MT correction

The approach we have followed so far has been based on some conventions that unavoidably lead to conservative corrections:

1. We treat all p-values the same way; we compare all of them to the *same* threshold, and we have investigated how to set that threshold. But it seems reasonable to be more exigent to larger p-values than with smaller ones.
2. We have tried to make no mistake, that is we compute the probability of at least one false rejection. Maybe instead we should try to obtain a good ratio between false positive and negatives.
3. We have made the complete null assumption: all H_i are true. But even a priori we believe that at least a small number will not be correct and the challenge is to find those in the haystack of large m.
4. The Bonferroni level is too small since it is based on a probability upper bound that is too large.

In this Section we will obtain a less conservative bound than Bonferroni by relaxing 1 above, while working with the assumption of independent tests. We will work with ordered p-values and set different thresholds for each one of them. In the sequel,

$$p_{(1)} \leq p_{(2)} \leq \cdots p_{(m)}$$

will denote the ordered p-values; we have dropped $\mathbf{X}_i$ from the notation to make it lighter.

A key result that is not too hard to prove is the following: if $y_i \overset{iid}{\sim} Uni(0,1)$, and we define

$$l_i = i\alpha/m$$

then

$$\mathbb{P}[\cap_{i=1}^{m}\{y_{(i)} > l_i\}] = 1 - \alpha$$

This observation already suggests a multiple testing correction under the working assumption of independent tests.

4. **[2]** Suppose that we carry out m independent tests with ordered p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots p_{(m)}$$

and we reject test i if $p_{(i)} < i\alpha/m$. Then, show that under the complete null the probability of at least one false rejection is α.

The method suggested by the previous exercise is demonstrated in the following figure, which compares with those by Bonferroni, the level obtained by independence assumptions and the level that does not account for MT.

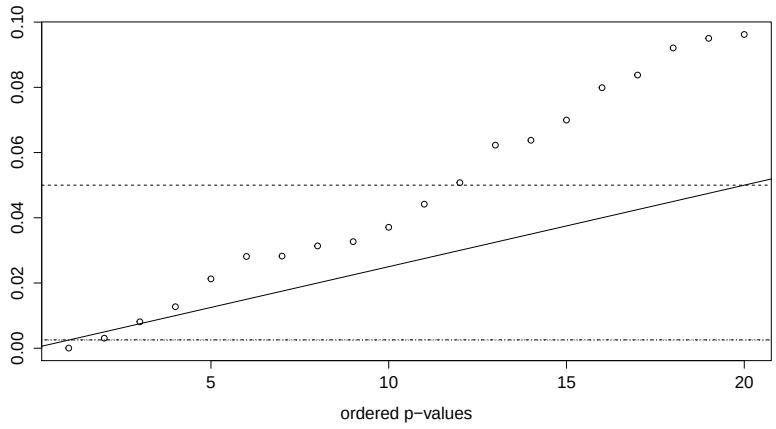

Figure 1: Bottom two horizontal lines are indistinguishable and correspond to level by Bonferroni and by independence assumptions. Top dashed horizontal is nominal α. The line is the varying level suggested by the previous exercise. Dots correspond to hypothetical p-values from 20 tests. The new technique would reject the tests with the 2 smallest eigenvalues.

3.4 False discovery rate and the Benjamini-Hochberg approach

The solution suggested by the previous exercise can actually go a long way into really solving the MT problem, by modifying slightly the rejection criterion. We wish to move away from the very conservative target of not making not even one false rejection and the unrealistic assumption that all hypotheses are a priori correct.

We define the false discovery proportion:

$$FDP = \frac{V}{R} = \frac{\text{falsely rejected}}{\text{total rejected}}$$

	H_0 Not Rejected	H_0 Rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m-R$	R	m

TABLE 10.2. Types of outcomes in multiple testing.

if $R > 0$ and 0 otherwise. Then $FDR = \mathbb{E}[FDP]$, where FDR stands for false discovery rate. This table from Wasserman summarises the situation:

Then, we have the following result, which I copy below from Wasserman's book, which in turns has summarised the result from the Benjamini and Hochberg paper (but should add as an assumption that test be independent, which is assumed in the BH paper).

10.26 Theorem (Benjamini and Hochberg). *If the procedure above is applied, then regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,*

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

In the theorem, the "procedure above" refers to a small modification of what we discussed in the previous section: find the first test whose p-value $p_{(i)} < i\alpha/m$. Then, reject this and **all** other tests with ordered p-values smaller than $p_{(i)}$.