Multiple testing

Anne–Laure Boulesteix
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie

February 15th, 2017

– Typeset by FoilTEX –

Structure

1. Introduction and examples

2. Control of the FWER

3. Control of the FDR

Multiple tests

Many variables have to be tested simultaneously, for instance:

- Association of 100,000 SNPs with disease status (healthy vs. Parkinson)
- Association between 20,000 gene expression levels and survival time in leukemia patients
- . . .

These variables that have to be tested are from now on denoted as $V_1, \ldots, V_m$.

Multiple tests: example

Question: Test whether the means of the variables $V_1, \ldots, V_m$ are different in groups "bad" and "good".

Tested null-hypotheses:

$$H_0^{(j)} : \ \mu_1^{(j)} = \mu_2^{(j)},$$

where $\mu_1^{(j)}$ resp. $\mu_2^{(j)}$ stands for the mean of variable j in group 1 resp. 2.

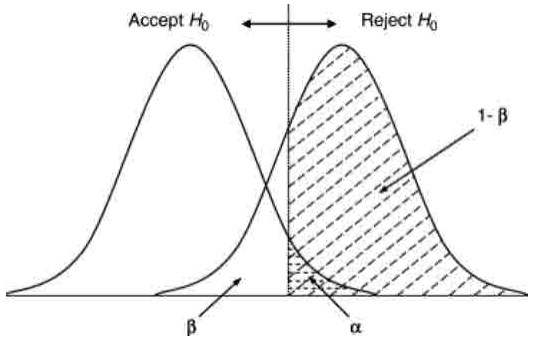
Test: Two-sample t-test

Significance level α is set to 0.05 (usual choice).

Testing

Univariate analyses:
Test whether the means of $V_1, \ldots, V_{1000}$ are different in groups "bad" and "good".

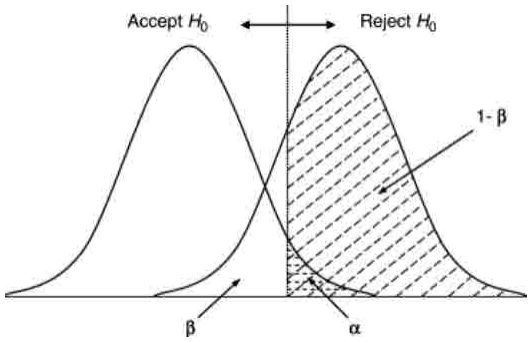
Naive approach: Perform a t-test at the level 0.05 for each variable, i.e. reject the null-hypothesis of equality of the means for all variables having a p-value < 0.05.



Multiple testing

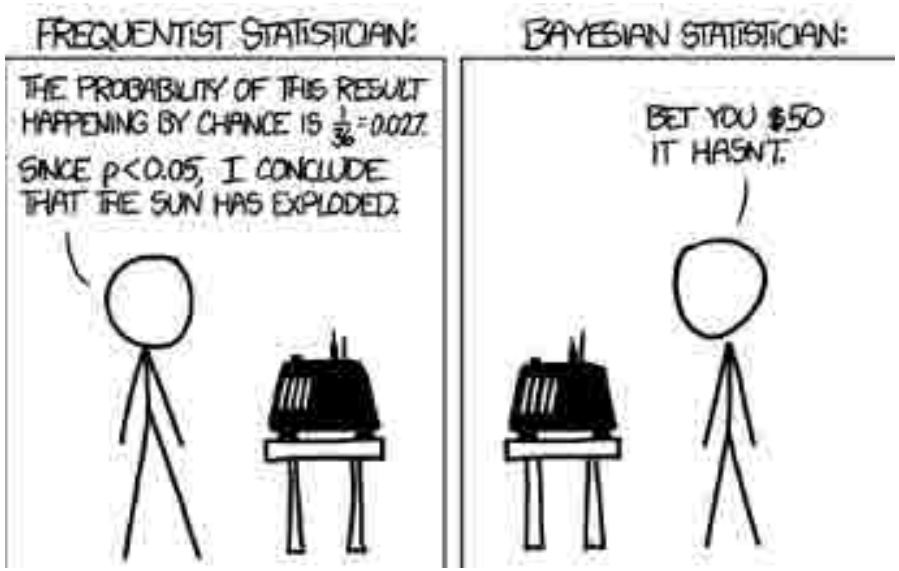
Naive approach: Perform a t-test at the level 0.05 for each variable, i.e. reject the null-hypothesis of equality of the means for all variables having a p-value < 0.05.

Problem: A p-value of 0.05 means that the probability to observe this value of the test statistic or a more extreme value is 0.05.



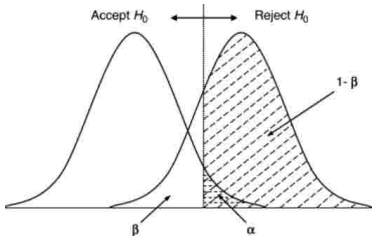
0.05 = 5% is low, but not zero!

0.05 = 5% is low, but not zero!



Frequentist vs. Bayesian

Multiple testing



- 0.05 = 5% is low, but not zero!

- If we consider 1000 tests and assume that all null-hypotheses are true (i.e. should not be rejected), the null-hypothesis will (wrongly) be rejected for 5%× 1000=50 variables!

→ The naive approach yields 50 *false positives* in this case and gives the impression that there are "interesting variables" in the data although it is not the case.

Type I error

- If we test at the level $\alpha = 0.05$ a single hypothesis that is true, the probability that it will be (wrongly) rejected by our test is $\alpha = 0.05$. This is called the **type I error**.

- If we test at the level $\alpha = 0.05$ several hypotheses that are true, the probability that at least one of them will be (wrongly) rejected by our test is **larger than** $\alpha = 0.05$.

Testing only one hypothesis

	fail to reject H_0	reject H_0
H_0 true	$1-\alpha$	α
H_0 false	β	$1-\beta$

α: Type I error = Probability to reject H_0 given that it is true

β: Type II error = Probability to fail to reject H_0 given that it is false

$1-\beta$: Power = Probability to reject H_0 given that it is false

Testing m hypotheses simultaneously

	fail to reject H_0	reject H_0	Total
H_0 true	...	V	m_0
H_0 false	...	...	$m - m_0$
Total	$m - R$	R	m

$m_0 =$ Number of true hypotheses

$R =$ Number of rejected hypotheses

$V =$ **Number of false positives**

Type I error and FWER

- The **probability** $P(V \geq 1)$ **that at least one true hypothesis will be wrongly rejected** by our test can be seen as a **generalization of the concept of type I error** to the case of multiple testing.

- It is called the **Family-Wise-Error-Rate** (FWER).

- The purpose of classical adjustment procedures is to control the FWER, i.e. to ensure that it is not larger than a fixed level α.

The FWER with the naive approach: a simple example

- Let us consider two null-hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ that are independent of each other.
- Let us further suppose that the two null-hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ are true.
- We apply the naive approach, i.e. we do both tests at the level α.
- Then the FWER (probability that at least one of the hypotheses is wrongly rejected) is

$$1 - (1 - \alpha)^2 = 1 - 1 + 2\alpha - \alpha^2 = 0.0975 \text{ for } \alpha = 0.05.$$

The FWER with the naive approach: a simple example

Then the FWER (probability that at least one of the hypotheses is wrongly rejected) is

$$1 - (1 - \alpha)^2 = 1 - 1 + 2\alpha - \alpha^2 = 0.0975 \text{ for } \alpha = 0.05.$$

This is much more than $\alpha = 0.05$!

We want to apply multiple testing procedures to make the FWER smaller than $\alpha = 0.05$.

Controlling the FWER

- We have to be **"more strict", i.e. to reject less hypotheses** in order to control the FWER.

- "Being more strict" means:
  - considering a threshold α^* smaller than $\alpha = 0.05$,
  - or equivalently: transforming the p-value p (that is compared to α) into a larger $adjusted$ p-value p^*.

- There are several possible ways to do that, i.e. several $adjustment\ procedures\ for\ multiple\ testing$.

Multiple testing terminology

- **controlling** the type I error
- **correcting** p-values, **correcting** for multiple testing, **correction** procedure
- **adjusting** p-values, **adjusting** for multiple testing, **adjustment** procedure

Bonferroni procedure

- Consider the larger threshold $\alpha^* = \alpha/m$
- or equivalently transform the p-value p into $p^* = \min(p \times m, 1)$

It can be shown mathematically that by doing that we control the FWER, i.e. we have

$$\text{FWER} < \alpha.$$

Bonferroni procedure: example

- We test 3 null-hypotheses and obtain the p-values $\mathbf{0.023}$ (for $H_0^{(1)}$), $\mathbf{0.784}$ (for $H_0^{(2)}$) and $\mathbf{0.004}$ (for $H_0^{(3)}$), respectively.
- With the naive approach, we would reject $H_0^{(1)}$ and $H_0^{(3)}$.
- With Bonferroni adjustment the threshold is $\alpha^* = 0.05/3 \approx 0.017$ instead of 0.05, and we reject only $H_0^{(3)}$.
- Equivalently, we can transform the p-values into $0.023 \times 3 = 0.069$, 1 and $0.004 \times 3 = 0.012$ and we also immediately see that only $H_0^{(3)}$ is rejected.

Bonferroni is conservative

- **Problem of Bonferroni procedure**: It is too conservative, i.e. it "conserves" (accepts) null-hypotheses too often.
- This leads to a **poor power**, i.e. some hypotheses that are false are not rejected although they should be rejected.
- **Improvement:** Holm procedure

The Holm procedure

- Order the p-values $p_1, \ldots, p_m$ from the smallest to the largest:

$$p_{(1)} < \cdots < p_{(m)}.$$

- Compare $p_{(k)}$ to the threshold $\alpha^* = \frac{\alpha}{m+1-k}$ (that is smaller than α).

- If k_0 denotes the smallest k for which $p_{(k)} > \alpha^*$, reject the hypotheses corresponding to the smaller p-values $p_{(1)}, \ldots, p_{(k_0-1)}$.

- If we never have $p_{(k)} > \alpha^*$, reject all null-hypotheses.

The Holm procedure: same example

- We test 3 null-hypotheses and obtain the p-values $\mathbf{0.023}$ (for $H_0^{(1)}$), $\mathbf{0.784}$ (for $H_0^{(2)}$) and $\mathbf{0.004}$ (for $H_0^{(3)}$), respectively.

- With Bonferroni adjustment, we would reject only $H_0^{(3)}$.

- With Holm:
  - We order the p-values: $0.004 < 0.023 < 0.784$.
  - $H_0^{(3)}$ (with $p = 0.004$) is rejected because $0.004 < 0.05/3$.
  - $H_0^{(1)}$ (with $p = 0.023$) is rejected because $0.023 < 0.05/2$.
  - $H_0^{(2)}$ (with $p = 0.784$) is not rejected.

Holm vs. Bonferroni

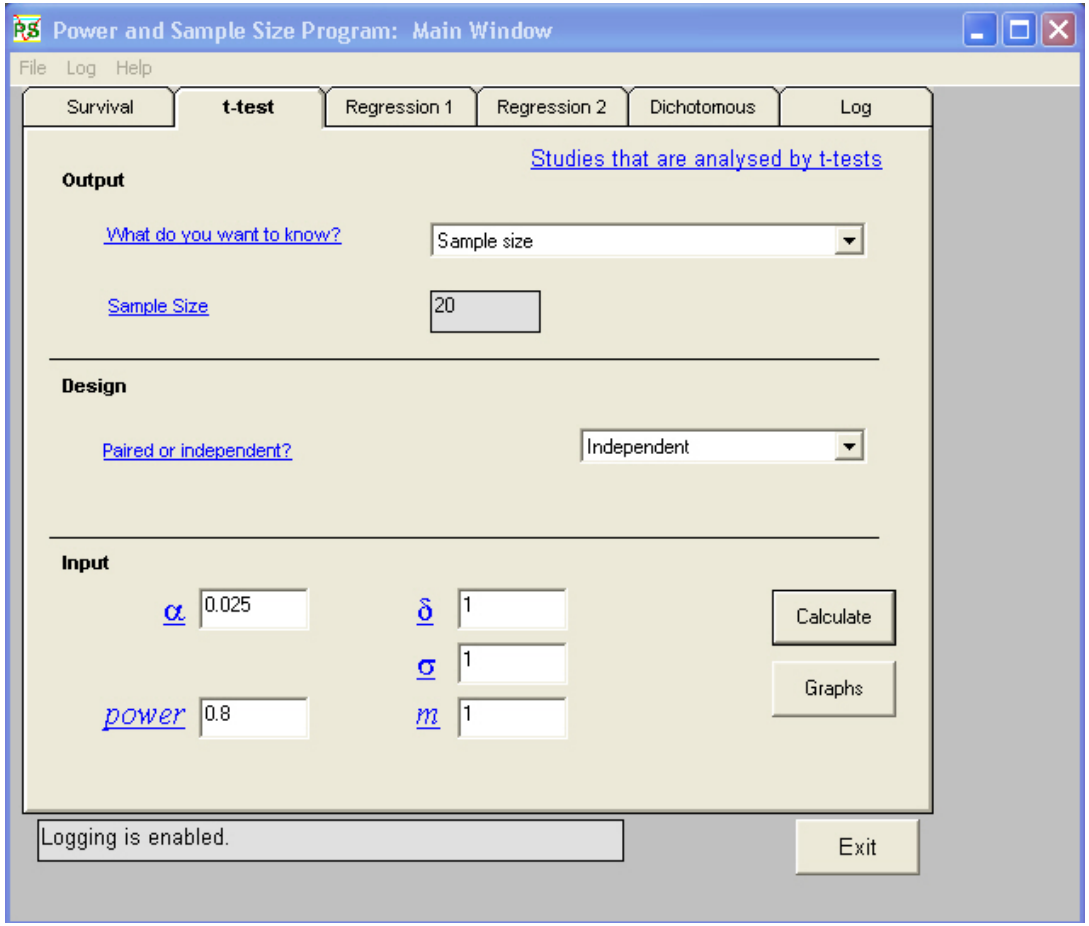
- Holm also controls the FWER, i.e. after adjustment with Holm's procedure we have FWER$< \alpha$.

- But it has more power, i.e. when a null-hypothesis is false, it is more likely to be rejected by Holm than by Bonferroni.

$\rightarrow$ Holm should be preferred to Bonferroni.

- But Holm is more complicated. In some cases, it is more practical to consider Bonferroni adjustment, e.g. for computing a sample size.

Sample size and adjustment



Inconvenience of FWER in high-dimensional settings

- Suppose that we test as many as $m = 1000$ hypotheses simultaneously.
- The FWER (probability that at least one null-hypothesis is wrongly rejected) is not very relevant: one false positive out of 1000 tests would not be so dramatic.
- The **proportion of false positives** within the rejected hypotheses is a more relevant feature.

False Discovery Rate (FDR): example

- **Example:** We test 1000 hypotheses and reject 65 of them. 55 of these rejected hypotheses are truly false, but 10 are not false and should actually not have been rejected.

- The proportion of false positives among the rejected hypotheses is $\mathbf{10/65 = 15.4\%}$.

- This is called **false discovery rate**.

FDR: formal definition

- Let Q be zero if no null-hypotheses are rejected.
- Let Q denote the proportion of false positives within the rejected null-hypotheses if at least one null-hypothesis is rejected.
- Then the FDR is defined as the mean of Q: $FDR = E(Q)$.

To put it simply: when many hypotheses are tested, it is unlikely that none of them is rejected. Hence the FDR can roughly be thought of just as the **proportion of false positives within the rejected null-hypotheses**.

Controlling the FDR

- One might want to rather control the FDR instead of the FWER.
- This makes sense in the case of many null-hypotheses (large m).
- Just as the FWER can be seen as a generalization of the type I error to the case of multiple testing, the FDR can be seen as an alternative generalization.

Benjamini-Hochberg procedure

- Benjamini and Hochberg (JRSS B, 1995) suggested a procedure to control the FDR, i.e. to ensure that FDR$<$0.05.
- It can be applied when at least a few hundreds of hypotheses are tested.
- It controls the FDR only if the hypotheses are independent (unrealistic assumption) and in some special cases of dependence.

Benjamini-Hochberg procedure

- Order the p-values $p_1, \ldots, p_m$ from the smallest to the largest:

$$p_{(1)} < \cdots < p_{(m)}.$$

- Compare $p_{(k)}$ to the threshold $\alpha^* = \frac{\alpha \cdot k}{m}$ (that is smaller than α).

- If k_0 denotes the largest k for which $p_{(k)} \leq \alpha^*$, reject the hypotheses corresponding to the smaller p-values $p_{(1)}, \ldots, p_{(k_0)}$.

- If we never have $p_{(k)} \leq \alpha^*$, reject nothing.

Benjamini-Hochberg procedure: remarks

- The Benjamini-Hochberg procedure is less conservative than the Bonferroni or Holm procedures, i.e. it rejects more null-hypotheses.

- However, if all null-hypotheses are true, FDR=FWER. So do not expect too much from Benjamini-Hochberg if (almost) all null-hypotheses are true.

- But it sometimes happens that BH rejects null-hypotheses although Bonferroni does not: even if $p_{(1)} > \alpha/m$, we might have $p_{(k)} < \alpha k/m$ for some larger k.

Multiple testing with R: example

The ALL data set:

- publicly available from Bioconductor platform
- $n = 128$ patients with ALL leukemia
- $\approx$ 20 demographical and clinical variables (sex, age, date of diagnosis, remission, etc)
- expression levels of $m = 12625$ genes measured using Affymetrix microarrays

Multiple testing with R: example

```
> library(ALL)
> data(ALL)
> Y<-pData(ALL)$BT
> Y<-as.numeric(is.element(Y,c("B","B1","B2","B3","B4")))
> Y
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> X<-t(exprs(ALL))
> X[1:4,1:7]
       1000_at  1001_at 1002_f_at 1003_s_at  1004_at   1005_at  1006_at
01005 7.597323 5.046194  3.900466  5.903856 5.925260  8.570990 3.656143
01010 7.479445 4.932537  4.208155  6.169024 5.912780 10.428299 3.853979
03002 7.567593 4.799294  3.886169  5.860459 5.893209  9.616713 3.646808
04006 7.384684 4.922627  4.206798  6.116890 6.170245  9.937155 3.874289
> pval<- apply(X, MARGIN=2, FUN=function(x,y) t.test(x[y==0],x[y==1])$p.value,y=Y)
> sort(pval)[1:6]
     37988_at     39389_at     38242_at     41609_at   36773_f_at     38319_at
3.552271e-44 1.424983e-43 5.499754e-42 7.098067e-41 2.003368e-39 1.484983e-38
```

Multiple testing with R: example (ctd.)

```
> p.adjust(sort(pval),method="bonferroni")[1:6]
   37988_at     39389_at     38242_at     41609_at   36773_f_at     38319_at
4.484742e-40 1.799041e-39 6.943440e-38 8.961309e-37 2.529252e-35 1.874791e-34

> p.adjust(sort(pval),method="holm")[1:6]
   37988_at     39389_at     38242_at     41609_at   36773_f_at     38319_at
4.484742e-40 1.798899e-39 6.942340e-38 8.959180e-37 2.528450e-35 1.874049e-34

> p.adjust(sort(pval),method="BH")[1:6]
     37988_at     39389_at     38242_at     41609_at   36773_f_at     38319_at
4.484742e-40 8.995207e-40 2.314480e-38 2.240327e-37 5.058503e-36 3.124652e-35
```

Another example

```
> Y<-pData(ALL)$relapse
> pval<- apply(X, MARGIN=2, FUN=function(x,y) t.test(x[y=="TRUE"],x[y=="FALSE"])$p.value,y=Y)
> sort(pval)[1:6]
    36912_at     37458_at      1584_at     41222_at     36041_at   37238_s_at
8.687728e-05 9.966280e-05 1.600307e-04 1.797820e-04 2.632425e-04 3.146477e-04

> p.adjust(sort(pval),method="bonferroni")[1:6]
 36912_at   37458_at    1584_at   41222_at   36041_at 37238_s_at
        1          1          1          1          1          1

> p.adjust(sort(pval),method="holm")[1:6]
36912_at   37458_at    1584_at   41222_at   36041_at 37238_s_at
       1          1          1          1          1          1

> p.adjust(sort(pval),method="BH")[1:6]
 36912_at   37458_at    1584_at   41222_at   36041_at 37238_s_at
 0.5674368  0.5674368  0.5674368  0.5674368  0.6620711  0.6620711
```

Conclusion

- Do not ignore multiple testing issues.
- Adjust p-values when looking at statistical significance in univariate analyses.
- Consider using an FDR-based adjustment method like the Benjamini-Hochberg procedure when testing many hypotheses simultaneously.