Multiple testing

General setup:

Description	notation	distribution
Data	$\mathcal{X} = (X_i, i = 1, \ldots, n)$	$P \otimes \ldots \otimes P$
Test-statistics	$T_n = (T_n(m), m = 1, \ldots, M) \in \mathbb{R}^M$	Q_n
Null-test statistics	$Z = (Z(m), m = 1, \ldots, M)$	Q_0

Often assumed (and we do that throughout) that we consider one-sided test; critical regions

$$\mathcal{C}_n = (\mathcal{C}_n(m), m = 1, \ldots, M), \quad \mathcal{C}_n(m) = (c_n(m), \infty)$$

for $c_n(m) \in \mathbb{R}$.

Multiple testing

We reject the m'th hypothesis if

$$T_n(m) > c_n(m).$$

A multiple testing procedure (MTP) is a choice of a vector of cut-offs $c_n = (c_n(m), m = 1, \ldots, M) \in \mathbb{R}^M$.

The objective is to detect among the M hypothesis the subset $\mathcal{H}_0 \subseteq \{1, \ldots, M\}$ of true hypotheses and $\mathcal{H}_1 = \mathcal{H}_0^c$ of false hypotheses.

Errors

Type I error (reject a true hypothesis): $m \in \mathcal{H}_0$ and $T_n(m) > c(m)$.

Type II error (accept a false hypothesis): $m \in \mathcal{H}_1$ and $T_n(m) \leq c(m)$.

$$
\begin{aligned}
V_n &= \sum_{m \in \mathcal{H}_0} I(T_n(m) > c(m)) \\
S_n &= \sum_{m \in \mathcal{H}_1} I(T_n(m) > c(m)) \\
R_n &= V_n + S_n
\end{aligned}
$$

Test statistics

The test statistic $T_n(m)$ is chosen so that it is “well behaved” if $m \in \mathcal{H}_0$ and large if $m \in \mathcal{H}_1$.

Often a form of asymptotic separation for $n \to \infty$ is assumed/imagined. Ex: t-test-types (single parameter), F-test (multiple parameters).

Trivial insight:

- The larger we choose the cut-offs the more type II errors we make (more conservative procedure).
- The smaller we choose the cut-offs the more type I errors we make (less conservative procedure).

Objectives

The objective of a MTP is to choose the least conservative (smallest) cut-offs that meet a criterion in terms of type I errors – a user defined parameter of interest.

General: Parameter of interest, Θ_n, of the distribution of (V_n, R_n) (alt. (V_n, S_n)).

More specific:

$$\Theta_n = \mathbb{E}G(V_n, R_n)$$

with

- $G(v,r) = I(v > 0)$ FWER
- $G(v,r) = I(v > k)$ gFWER(k)
- $G(v,r) = \frac{v}{\max\{r,1\}}$ FDR
- $G(v,r) = I(\frac{v}{\max\{r,1\}} > q)$ TPPFP(q)

Procedures

For a user defined level α select cut-offs such that

$$\Theta_n \leq \alpha.$$

Two essentially different approaches.

Deterministic procedures: The cut-offs do not depend upon the data – especially not on the observed test-statistics.

Non-deterministic procedures: The cut-offs are allowed to depend upon the data – often through the observed test-statistics.

Null-distributions

We do not know the distribution of T_n. We do not know $\mathcal{H}_0$. Hence,

We certainly do not know the distribution of (V_n, R_n).

Stochastic null-domination: We may be able to estimate a distribution Q_0 on $\mathbb{R}^M$ such that:

If Z has distribution Q_0 then $(Z(m), m \in \mathcal{H}_0)$ stochastically dominates $(T_n(m), m \in \mathcal{H}_0)$.

Intuition: Using Q_0 to construct the MTP at level α leads to a more conservative MTP than using Q_n, thus $\Theta_n \leq \alpha$.

Formal results

- For deterministic procedures stochastic null-domination (precisely jtNDT) implies that the procedure is conservative.
- In general, there are non-deterministic procedures where stochastic null-domination does not provide a conservative MTP,
- but for all procedures in the book, the conditions needed follow from stochastic null-domination.

Some proofs of Chapter 3 requires a modification to see the last result.

Estimated null-distributions

In practice, Q_0 is unknown but we have a consistent estimator Q_{0n} of Q_0.

- The book emphasizes a non-parametric bootstrap based estimator of Q_0.
- For single parameter hypothesis an alternative is the asymptotic multivariate normal distribution.

The book emphasizes two choices of Q_0

- The null shift and scale-transformed null distribution.
- The null quantile-transformed null distribution.

Control – Asymptotics

A MTP controls the error rate Θ_n at level α if $\Theta_n \leq \alpha$. This is actual control.

If Θ_n is estimated by Θ_{0n} – based on Q_{0n} among other things – we have nominal control at level α if $\Theta_{0n} \leq \alpha$.

In general we can only hope for asymptotic actual control, that is,

$$\limsup_{n\to\infty} \Theta_n \leq \limsup_{n\to\infty} \Theta_{0n}.$$

Better results are available for marginal MTPs if we know the marginal distribution of $T_n(m)$ if $m \in \mathcal{H}_0$ (Chapter 3).

Unknowns

A priori we do not know Q_0 – we estimate it as Q_{0n}.

A priori we do not know $\mathcal{H}_0$ either. Multiple hypothesis testing is all about "estimating" $\mathcal{H}_0$.

- If $\Theta_n = \mathbb{E}G(V_n)$ – a function of the distribution of V_n alone. Then if G is increasing, the procedure becomes more conservative if we replace $\mathcal{H}_0$ a priori by $\{1, \ldots, M\}$ (Chapters 4 and 5).
- In general, the "gain in power" by focusing on other error rates like FDR is lost if we simply replace $\mathcal{H}_0$ by $\{1, \ldots, M\}$.
- In particular, control of FDR reduces to control of FWER if $\mathcal{H}_0$ is $\{1, \ldots, M\}$.

p-values

For one-sided tests the p-value is noting but the monotonely decreasing transformation

$$P_n(m) = 1 - F_{n,m}(T_n(m))$$

where $F_{n,m}$ denotes the marginal distribution function for the distribution of $T_n(m)$.

We reject the m'th hypothesis if $P_n(m) < p(m)$ if and only if $T_n(m) > c(m)$ where $p(m) = 1 - F_{n,m}(c(m))$.

In practice we compute p-values using the null-distribution with marginals $F_{0n,m}$, thus

$$P_{0n}(m) = 1 - F_{0n,m}(T_n(m)).$$

Null domination implies larger p-values.

Common cut-offs

A MTP is called common cut-off if all $c(m), m = 1, \ldots, M$ are equal.

A common cut-off approach is only sensible if the test statistics $T_n(m)$ for $m = 1, \ldots, M$ all follow the same marginal distribution under the null.

Common quantile is simply a common cut-off based on the test statistics $(1 - P_{0n}(m), m = 1, \ldots, M)$.

Adjusted p-values

$$\tilde{P}_{0n}(m) = \inf\{\alpha \,|\, T_n(m) > c(m,\alpha)\}.$$

where it is made explicit that the cut-off, $c(m,\alpha)$, depends on the nominal level α.

Given the vector of adjusted p-values and a nominal level α we reject the m'th hypothesis if and only if $\tilde{P}_{0n}(m) \leq \alpha$.

Adjusted p-values can in many cases be computed simply as a kind of inflation of the original p-values.

Marginal or multivariate

Classical MTPs are marginal (Bonferroni, Holm, ...). They control FWER and are not too conservative if the test statistics are independent or close to independent.

Some modern MTPs (Benjamini-Hochberg/FDR) are marginal and requires assumptions on dependence structure to work.

Bootstrap based methods attempt to capture the complete dependence structure among the test statistics.

Deterministic or non-deterministic

Single step common cut-off or common quantile procedures where the cut-off is chosen according to Q_{0n} alone are regarded deterministic.

Two types of non-deterministic, sequential procedures:

- Step-down starts from the marginally most significant hypothesis and rejects using gradually less and less conservative cut-offs.
- Step-up starts from the marginally least significant hypothesis and accepts using gradually more and more conservative cut-offs.

General insight: Non-deterministic procedures are less conservative than deterministic procedures. In some cases they require assumptions on the dependence structure.

Road map

- Θ_n a parameter of the dist. of V_n only:
  - Know marginal distributions, close to independence: Simple corrections like Holm can be used. Chapter 3.
  - Do not know marginals or expect strong dependence: Single step procedures (Chapter 4) in general and for FWER step-down (Chapter 5).
- Θ_n a parameter of the dist. of (V_n, R_n):
  - Know marginal distributions, close to independence: Simple corrections like Benjamini and Hochberg or Lehmann and Romano can be used. Chapter 3.
  - Do not know marginals or expect strong dependence: Augmentation procedures based on a suitable FWER procedure (Chapter 6) or empirical Bayes (Chapter 7).

Points of view

- The MTP is a post-hoc correction; after discussing the appropriate modeling and all other issues. It does not bring anything new into the model or the test statistics.
- Asymptotic $n \to \infty$ justification may be hard appreciate for many practical problems where n is quite small.
- Accurate bootstrap marginal p-values for M large requires large B – time and memory consuming.
- Most important insight: Type I error rate does not need to be FWER. Often FDR or TPPFP(q) are much more appropriate for practical purposes (screening experiments, explorative data analysis, ...)