New approaches to error control in multiple testing

Juliet Popper Shaffer

Fourth Lehmann Symposium

May 2011

Follow-up to talk in Lehmann memorial session

- This talk continues a theme started in my talk at the memorial for Erich. There will be some overlap, so I apologize if you were there.

Outline: Brief History

- Tukey (1953),
- Lehmann (1957)
- Advent of mass well-structured testing
- Benjamini-Hochberg (1995)
- Recent extensions of Tukey and BH approach
- More recent changes

Outline: More-Recent changes, relation to optimality

- Change in emphasis from control of Type I error to balance of Type I and Type II error (or power)
- Different level of optimality and relation to balance issues
- Brief comparison of two approaches

The multiplicity problem

- If m hypotheses are tested, each as if it is the only one, at some level α, and if all are true, the expected number of errors (Type I errors) will be m α and the probability of one or more errors will increase also substantially.

Tukey et al methods

- In 1953 Tukey wrote a book-length manuscript called the Problem of Multiple Comparisons. It was circulated to a limited group but unpublished until 1994 in his collected works.

Tukey described several possible criteria for controlling error

- Defined:
- Per-comparison error rate
- Per-family error rate
- Familywise error rate

Per-comparison:PCER

- Expected number of errors per comparison. Average level of error control for individual tests. Multiplicity issues do not affect procedures.

Per-Family error rate (PFER)

- The expected number of errors in the family of tests under consideration.
- This is m_0 times the PCER, where m is the number of hypotheses in the family and m_0 is the number of true hypotheses.

Deciding on a family is the main problem in many situations with a variety of hypotheses.

Family-wise (FWER)

- Probability of one or more Type I errors in the whole family of tests.
- This is a compromise between per-comparison and per-family error rates. With small α and small-to-moderate correlations among tests, usually close to PFER.

Universal PFER-FWER control: Bonferroni

- Test each hypothesis at level α/m.
- Universal control of PFER and thus of FWER.
- For exact FWER-control with independent tests can test at $1 – (1-\alpha)^{(1/m)}$
- If m_0 is estimated can use m_0 instead of m in equations above.

Bonferroni

- Although it has been typically used to control the FWER, note that it controls the PFER, in fact exactly at m_0 α/m.
- The FWER, on the other hand, is smaller than the PFER and the difference increases with the degree of positive correlation among the test statistics.
- Many more-powerful procedures have been developed to control the FWER.

Example: Stepwise tests

- Stepdown: Test the most significant hypothesis at α/m; if rejected, test the next most significant at α/(m-1) or higher in some structured situations, etc.
- Stepup: Test the least significant hypothesis at α, if accepted test the next least significant hypothesis at a smaller level, etc.
- Step-up-down: Generalization of both of the above.

Lehmann (1957)

- In two papers that year, Erich introduced a loss-function approach to multiple testing, applicable in a very general way.
- In testing a single hypothesis, if a is the loss for a false rejection (Type I error), and b for a false acceptance (Type II error), and the test is best unbiased and is carried out at level b/(a+b), the procedure has uniformly minimum risk among unbiased procedures.

Lehmann (1957)

- If the losses are additive over a number of such tests, the multiple procedure has uniformly minimum risk.

Pre- 1990s

- The Tukey approach of controlling Type I error at a suitably low level was dominant in the early applications. There were usually only a small number of hypotheses tested, and the consequences of a false rejection could be severe (e.g. comparing a number of treatments for a disease).

Ill-structured mass testing

- There were situations in which many hypotheses were of interest, but they were not of the same type and/or not of equal importance. They were typically divided into separate families for testing, and the decisions about family size were more important than the choice of error rate.

Many hypotheses:Family issues:Ex:Factorial Designs

- There were some cases in which many hypotheses were tested, but the main problem there was deciding on families. For example, in multifactor designs, should all tests for main effects, interactions of all levels, be treated as one big family? How should families be defined? How about followup tests on simple effects and interactions?

Many hypotheses:Family Issues: Ex. Surveys

- Many subgroups, possibly many characteristics of interest. How should families be defined? It isn’t even clear what the total number of hypotheses is.

Well-structured mass hypotheses

- It has always been difficult to convince investigators to use strict Type I error-controlling procedures due to the loss of power for individual tests.
- This became especially true with the advent of well-structured mass hypothesis testing: Testing with families of very large size.

Examples

- Microarrays: Thousands of tests
- Neuroimaging: individual pixels
- Astronomy: Millions of tests
- Also, in some of these cases, a small number of Type I errors could often be tolerated (e.g. microarrays) since results would be subject to other testing for confirmation.

Benjamini-Hochberg (1995)

- At a very fortunate time, Benjamini and Hochberg introduced a new criterion: control of the false discovery rate (FDR). The idea is to keep the proportion of false rejections among the rejections to a suitably small value.
- Accompanied by a test controlling FDR.

- The observed proportion of false discoveries, FDP, is the ratio:
- no. of false rejections /no. of rejections,

defined as zero if there are no rejections.

- The false discovery rate, FDR, is the expected value of FDP.
- Much recent work is devoted to methods for controlling alternative versions of the FWER, the FDP, or the FDR and other related measures. New criteria are constantly being defined. (Unfortunately the same terms are often defined differently.)

Alternative measures: Examples

- For Type I error control (rather than FDR or FDP)
- k-FWER: Prob. of k or more errors controlled (Lehmann and Romano, 2005; van de Laan et al, 2004).
- pFDR (Storey,2002,2003), k-FDP and k-FDR (Sarkar,2007), ERR-erroneous rejection ratio (Cheng, 2006), aFDR (Pounds and Cheng, 2005), cFDR (Tsai et al, 2003).

More-recently: new emphasis on both kinds of error

- The methods mentioned above all concentrate on some kind of Type I error control. If specified, it is usually .05, as for original FWER. More-recent approaches explicitly consider both Type I and Type II error control.

Balancing errors

- In scientific work there have to be conventions for deciding what hypotheses to accept/reject. The control of Type I error at a traditional level α has played that role. Can we develop alternative criteria taking both kinds of error into account?
- FDR is a start, although still based on Type I error control. More recent emphasis on explicit consideration of the other type of error.

New criteria are needed

- If both types of error (I and II) are to be taken into account, new criteria are needed. Decision-theoretic approaches are helpful in this regard. Given choices of weighting for different types of errors, optimal procedures are defined.
- Consider errors of rejection and acceptance jointly.

Genovese and Wasserman (2002)

- In this paper, G and W proposed consideration of the False Nondiscovery Rate (FNR), defined as (no. of false acceptances/no. of acceptances).
- Spawned literature on other measures related to Type II error.

Some alternatives

- For Type II-like error control (rather than FNR):
- NDR-non-detection ratio (Craiu and Sun, 2008) (called FNR by Pawitan et al, 2005, who use FNDR for what others call FNR, FNS-fraction of non-selection (Delongchamp et al, 2004), MR-Miss rate (Taylor, Tibshirani, and Efron, 2005).

- Genovese and Wasserman also consider risk functions combining the two rates: FNR + λ FDR.
- Several other authors consider various measures of false detections and false non-detections jointly, either fixing one and maximizing the other (Strimmer,2008; Chi,2008) and/or combining them in some way (Craui and Sun, 2008; Sarkar,2006; Pawitan et al, 2005).

- Each author defends a proposed alternate measure as more intuitively meaningful than other measures.
- Craiu and Sun (2008), for example, consider the NDR, the expected proportion of falsely-accepted hypotheses among the false hypotheses, to be a better measure than the FNR, the expected proportion of falsely-accepted hypotheses among the accepted hypotheses, and they and others explicitly compare the two in different situations.

Type I and Type II errors considered

- Note that all FDR-like measures are closely related to Type I errors and all FNR-like measures are closely related to Type II errors.

Loss function approaches balancing Type I and II errors

- In the two 1957 papers (A theory of some multiple decision problems I and II) Erich took a decision-theoretic loss function approach.
- Other early loss-function approaches are due to Duncan. More recently both Charlie and Peter have considered loss functions. See also Sarkar et al, 2008, Rice (2010), many others. Some are bayesian, some frequentist.
-

Relation of recent work to the Lehmann 1957 papers

- As noted, Erich’s 1957 papers used a decision-theoretic approach to multiple testing, where the criterion was minimizing a weighted combination of Type I and Type II errors. So a simple intuitive alternative approach to use of these more complex measures would be to minimize such a combination directly.

- The idea of putting weights on the two types of errors (I and II) and then considering minimum risk may be a more natural way of approaching a balance of errors than combining more indirect measures like FDR and FNR. Furthermore, it is more flexible in that each hypothesis can have different weights.

Lehmann 1957

- If a is the loss for a Type II error and b is the loss for a Type I error, a minimum-risk procedure uses the α level a/(a+b).
- Scale is arbitrary. Make a+b = 1. Then a is the minimum-risk level for the procedure, used for each test.
- Test each hypothesis at significance level a.

Equating BH-FDR and Lehmann-FDR

- Genovese and Wasserman showed that asymptotically the Benjamini-Hochberg (1995) method can be equated to a test of each hypothesis at a fixed level independent of the number of hypotheses.
- This is true also non-asymptotically with the level depending on the number of hypotheses.

Relation to FWER, PFER, and PCR.

- FWER is the probability of one or more errors, while PFER and PCR are both expected values: of family error rate and individual error rate, respectively.
- Both the latter involve testing each hypothesis individually using a specified significance level, although in one case (PFER) the level varies with m.

Gordon et al (2007)

- If Bonferroni is considered in terms of PFER control rather than FWER control, and the level is not fixed at a conventional level α, additional possibilities arise.
- Limit of 1 is not necessary.
- Gordon et al: Equate the number of Type I errors for Bonferroni and an FDR-controlling method, and then compare the two on power.

PFER and PCER

- If α is allowed to vary and be greater than one, for a given m there is no difference between PFER and PCER. So equating the Bonferroni and BH (1995) procedures to make some function of Type I error equivalent can also be expressed at equating PCER to the BH procedure.
- Erich’s 1957 proposal is to use the PCER (additive losses).

- The issue of weights: Even if measures of both discoveries and non-discoveries are considered, how can one decide, in using a weighted combination of FDR and FNR or other measures, what the weights should be?

Choosing weights

- Here, large numbers of hypotheses can provide useful information to apply to choosing weights using the Lehmann 1957 approach.
- When applied to mass data situations, it is possible to estimate the number of true hypotheses and some aspects of the distribution of the false hypotheses; these in turn can be useful in deciding on the weights to use in comparing Type I and Type II errors.

PCER equivalents to FDR

- The numbers in the table to follow give some approximate weights of Type 1 (a) and Type 2 (1-a) for Lehmann PCER to be equivalent to BH-FDR at $\alpha = .05$.
- A multiple test with these weights minimizes expected loss, given the specific tests used, if these individual tests are unbiased.

PCER – BH-FDR equivalents

	m = 4	**m = 20**	**m = 100**	**Asymptotic**
Alt. = 0	0125	0025	0005	---
Alt = 1				
Prop.m1=.05		0026	00055	1 e -08
Prop.m1=.10		0026	00057	1 e -08
Prop.m1=.25	0135	0028	0006	1.89 e -06
Prop.m1=.50	014	0031	0007	4.699 e -05
Alt. = 2				
Prop.m1=.05		003	0007	0001284
Prop.m1=.10		0034	001	0005310
Prop.m1=.25	0165	005	0033	002915
Prop.m1=.50	021	01	0095	009266
BHFDR jProp.	Prop.m$_1$ = .05 .0475	Prop.m$_1$=.10 .045	Prop.m$_1$=.25. .0375	Prop.m$_1$=.5 .025

Equating other measures to Type I-Type II comparisons

- It may be possible to equate many of the more complex criteria to equivalent measures of Type I-Type II balance. This could help in understanding the meaning of these criteria and in deciding how to use them in practice.

Summary

Multiple testing research has evolved from primarily considering Type I error to an interest in balancing Type I and Type II error in some fashion, direct or indirect.

Equating some of the more complicated criteria to the simple criterion of the balance between Type I and Type II error may help in deciding on the level of balance that is desirable.