p–values Had a Good Run: A Primer on the ‘New Statistics’

Rob Cribbie
Quantitative Methods Program
Department of Psychology
York University

Part 3: Multiplicity Issues and Modern NHST

- One of the most complex issues to battle when conducting modern NHST analyses is the effect of multiplicity

- *Multiplicity* refers to the fact that if you conduct multiple tests of significance, each at a nominal Type I error rate of α, then the overall probability of a Type I error (α_o) will be much greater than α if the H_0 associated with >1 tests is true

Multiplicity Control

Statistical Methods in Psychology Journals

Guidelines and Explanations

Leland Wilkinson and the Task Force on Statistical Inference
APA Board of Scientific Affairs

Multiplicities. *Multiple outcomes require special handling. There are many ways to conduct reasonable*

- When *Wilkinson and the Task Force on Statistical Inference* discussed issues with NHST in their famous 1999 paper, no other topic garnered more page space then the discussion of multiplicity control

Effect of Multiplicity

- The amount of inflation of α_o depends primarily on:
  - The number of tests conducted
    - The more tests conducted the greater the chance of making a Type I error
    - If all the tests were independent, the overall Type I error rate would be about $\alpha_0 = T\alpha$, where T represents the number of tests conducted
  - The correlations among the tests
    - Generally, the larger the correlation among the tests the greater the overall probability of a Type I error
      - In the extreme case where all tests were perfectly correlated the overall Type I error rate would equal α

In what situations is it recommended that we control for multiplicity?

- Pairwise comparisons from an ANOVA?
- Multiple tests of correlation?
- Multiple predictors in a regression?
- Multiple outcome variables?
- Number of parameters in a path analysis or structural equation model?
- Number of voxels in an fMRI analysis?
- Etc., Etc., …

Some Language Related to Multiplicity Control

- Familywise Error Rate Control
  - Controlling the rate of Type I errors across all related tests
    - By definition, the familywise rate is the probability of at least one Type I error across all T tests
      - E.g., all pairwise comparisons in ANOVA, all predictors in a regression
    - Examples of procedures that control this error rate are the Bonferroni, Scheffé, Holm
    - When familywise error rate control is imposed, $\alpha_o = \alpha$, however $\alpha_T < \alpha$, where α_T represents the error rate per test

Some Language Related to Multiplicity Control

- Per-test/Testwise Error Rate
  - Controlling for the rate of Type I error separately for each test α
    - The probability of a Type I error for a specific test (α_T) of interest is maintained at α
    - Analogous to no multiplicity control since each test is conducted at level α
  - In this case, $\alpha_T = \alpha$, however $\alpha_o > \alpha$

Why is Multiplicity Control Recommended?

- As was stated earlier, the overall Type I error rate (α_o) depends on the number of tests conducted and the correlation among the tests (if multiple H_0s are true)
- For independent tests, α_o approaches:
    - $1-(1-\alpha)^T$
    - Recall that T represents the number of tests conducted
  - For example, if 10 independent tests are each conducted at $\alpha = .05$:
    - $1-(1-\alpha)^T = 1 - (1-.05)^{10} = .40$
      - The rate for correlated tests will generally be lower

Why is Multiplicity Control Recommended?

- Therefore, since α_o increases above α any time multiple tests are conducted (when multiple H_0s are true), proponents of multiplicity control argue that this control should be imposed any time multiple tests of significance are conducted
  - This is common practice in some settings (e.g., pairwise comparisons in ANOVA), but not in others (e.g., multiple outcome variables), even though the issues are the same

Arguments AGAINST Multiplicity Control

- Consistency
  - When a researcher adopts multiplicity control, the Type I error rate for each test depends on the number of tests being conducted
    - E.g., if one researcher is comparing Protestant and Catholic students on anxiety, and another is comparing Protestant, Catholic, Jewish, Muslim and Atheist students on anxiety, the researchers will have a different α_T for the comparison of Protestant and Catholic students if multiplicity control is imposed

Arguments AGAINST Multiplicity Control

- Power
  - Since α_T is reduced when multiplicity control is imposed, power will also be reduced
- "Natural" unit of analysis
  - Many researchers have argued that each individual test (e.g., pairwise comparison) is the natural unit of analysis and therefore Type I error control should be imposed at the individual test level (i.e., per-test or testwise control)
- Simplicity
  - No complicated multiple comparison procedures

Arguments AGAINST Multiplicity Control

- Modern reasons for not adopting multiplicity control …
- 1) Replication!
  - The reason for multiplicity control is to eliminate errors of statistical inference, but that is naturally handled by repeating studies under similar conditions and comparing the magnitude of the effects
    - More on this to come
- 2) Reduced Role of *p*–values
  - If *p*–values play only a minor role in inference, then the need for multiplicity control is significantly reduced
- 3) H_0 is Always False
  - So … there is no such thing as a Type I error!

Some Twitter Discussion

Daniël Lakens
@lakens

Following

Just read Rothman (1990) "No need to correct for multiple comparisons" jstor.org/stable/20065622 The argument is completely flawed. 3000 citations. I am left to wonder if the illogical conclusion in the article is cited for its convenience instead of its analytical rigor.

Some Twitter Discussion

Ken Rothman @ken_rothman · Apr 8

I like adjustments for multiple comparisons when analyzing random numbers. I don't do much of that, but those doing GWAS or studying psychic phenomona may come close.

Genome Wide Association studies

Discussion Point

- Is there any reason to impose multiplicity control in practice (e.g., testing multiple hypotheses in applied behavioural science research)?