Multiplicity Control

Introduction to Multiplicity Control

- When we test hypotheses, we usually do so with a specific nominal Type I error rate (e.g., α = .05)
- Whenever we conduct "multiple" tests of significance, we must consider what effect the ***multiplicity*** has on the error rates of the test statistics
    - More specifically, the 'overall' Type I error rate
        - The overall Type I error rate can be defined in many different ways
    - This applies to any type of multiplicity, such as contrasts in ANOVA, multiple predictors in regression, main effects and interactions in factorial models, multiple outcome variables, mixes of these types of effects, etc.

Forms of Multiplicity

- Single Omnibus Test
  - E.g., 4 levels of an IV and we conduct all 6 pairwise comparisons
- Multiple Tests of the Same Form
  - E.g., we might compare males and females on each of 7 different outcome variables (e.g., generalized anxiety, panic, agoraphobia, social anxiety, …)
- Multiple Tests of Different Form
  - E.g., if we conduct a multiple regression model with three predictors, one a 4-level categorical variable (6 pairwise comparisons), and 5 outcome variables, then we have C = 40 tests of statistical significance
    - C = (2 continuous predictors + 6 pairwise comparisons per model) X 5 outcome variables = 40 hypothesis tests

Effect of Multiplicity on Type I Errors

- The overall probability of Type I errors approaches $1-(1-\alpha)^C$ for *independent* comparisons (or approximately $C\alpha$)
  - For example, when we conduct C = 6 hypothesis tests (let's say there are 4 groups and we conduct all 6 pairwise comparisons), each at $\alpha = .05$, the overall Type I error rate approaches $1-(1-.05)^6 = .265$ or approximately $6(.05)=.30$
    - Note: most tests are not independent; when the tests are not independent the overall rate will actually be less than $1-(1-\alpha)^C$, but still much higher than α
- The important question is whether the calculated rate of a Type I error (in this case 26.5%) is acceptable

No Multiplicity Control

- If we choose to not control for multiplicity, the probability of making a Type I error for any given test (α_{PT}) is set equal to the nominal α level
  - $\alpha_{PT} = \alpha = .05$
- However, it becomes evident that as the number of tests to be conducted increases, so does the overall Type I error rate for the set of comparisons
  - E.g., overall Type I error rates ~ $C\alpha_{PT}$
- Is this acceptable??

Multiplicity Control: Familywise Error Rate (α_{FW})

- Here, we control the probability of making at least one Type I error across a <u>*set*</u> of tests
- If we maintain $\alpha_{FW} = \alpha$, then α_{PT} will necessarily be $< \alpha$
  - Thus, increasing the number of tests has no effect on the overall Type I error rate (α_{FW}), but α_{PT} will become increasingly smaller as the number of comparisons increases
- Is this acceptable??

Example

- Back to our example where we want to conduct all pairwise mean comparisons with 4 groups (*C* = 6) using α = .05
  - If we test each of the C = 6 pairwise comparisons at α_{PT} = .05 then:
    - $\alpha_{FW} \simeq 1\text{-}(1\text{-}\alpha_{PT})^{C} \simeq 1\text{-}(1\text{-}.05)^{6} \simeq .265$ (26.5%)
    - $\alpha_{FW} \simeq C\alpha_{PT} \simeq 6(.05) \simeq .30$ (30%)
  - If we test each of the C = 6 pairwise comparisons with α_{FW} = .05 then:
    - $\alpha_{PT} \simeq 1\text{-}\ (1\text{-}\alpha_{FW})^{1/C} \simeq 1 - .95^{1/6} \simeq .00852$ (.85%)
    - $\alpha_{PT} \simeq \alpha_{FW}/6 \simeq .05/6 = .0083$ (.83%)

Per-test vs Familywise Type I Error Control

- So … if you are conducting 6 pairwise comparisons in a one-way ANOVA, which option would you prefer?
  - No Multiplicity Control
    - $\alpha_{PT} = .05; \alpha_{FW} = .265$
  - Familywise Error Control
    - $\alpha_{PT} = .008; \alpha_{FW} = .05$
- Note: The exact same issues apply in other multiplicity testing situations (e.g., 6 outcome variables, 6 predictors, or a mix of types of tests)

False Discovery Rate
Type I Error Control

- The false discovery rate is the expected *proportion of false rejections to the total number of rejections*
  - This is in contrast to α_{FW}, which is the proportion of one or more false rejections out of the total number of hypothesis tests
- False discovery rate control represents a compromise between familywise error and no control
- FDR is becoming more popular, especially when the number of tests conducted is very large (e.g., fMRI and DNA microarray research)

FDR vs FWE

- Say we do 10 experiments, each with 10 tests (and somehow know which are true/false rejections)
- # of total rejections: 7,8,9,6,7,6,7,8,9,6
- # of false rejections: 2,1,3,0,0,2,0,1,4,0
  - FWE = # of experiments with at least one false rejection divided by the number of experiments = 6/10 = .6
  - FDR = average proportion of number of false rejections to the total number of rejections
    - FDR = (2/7 + 1/8 + 3/9 + 0 + 0 + 2/6 + 0 + 1/8 + 4/9 + 0)/10 = .165
- Thus, a more liberal (powerful) test would control the FDR at α (i.e., a less extreme adjustment would be required)

Which Type of Control Should You Choose??

- No Multiplicity Control
- Familywise Error Control
- False Discovery Rate Control

- Familywise/False Discovery Rate error control have been routinely recommended by statisticians, quantitative methodologists from psychology, statistics textbook authors, journal editors, etc.

Procedures for Controlling the *Familywise Error Rate*

- Bonferroni
  - α_{PT} is set at α/C (recall: C represents the number of tests)
  - Therefore, if we were to conduct 3 tests and $\alpha_{FW} = .05$, then:
    - $\alpha_{PT} = .05/3 = .0167$
  - Bonferroni is a *simultaneous* procedure; all *p*-values are compared against the same α level
- Extremely conservative in most situations (e.g., pairwise comparisons, correlation matrices, multiple predictors in a GLM)

Bonferroni Example

- $\alpha = .10$
- $C = 6$
- $p_A = .018$, $p_B = .005$, $p_C = .024$, $p_D = .141$, $p_E = .002$, $p_F = .055$
- $\alpha_{\mathrm{PT}} = \alpha/C = .10/6 = .017$
- Statistically significant ($p \leq \alpha$):
  - p_B, p_E

Procedures for Controlling the Familywise Error Rate

- Holm
  - A *stepwise* (not sequential) modified Bonferroni procedure that considers the number of possible null hypotheses remaining, given previously rejected null hypotheses
    - The p-values (p_c) are ordered from smallest to largest
      - $p_1, \ldots, p_C$; $c = 1, \ldots, C$
    - α_c, starting at $c = 1$ (smallest p-value), is set at $\alpha / (C-c+1)$ and if any $p_c > \alpha_c$ testing stops and all remaining p-values (i.e., p_c to p_C) are declared nonsignificant
  - Can be much more powerful than the simultaneous Bonferroni procedure

Holm Example

- $\alpha = .10$
- $p_A = .018$, $p_B = .005$, $p_C = .024$, $p_D = .141$, $p_E = .002$, $p_F = .055$
- C = 6
- Ordered: $p_1 = .002$, $p_2 = .005$, $p_3 = .018$, $p_4 = .024$, $p_5 = .055$, $p_6 = .141$
- $\alpha_{PT} = \alpha / (C\text{-}c+1)$
- $\alpha_1 = .017$, $\alpha_2 = .020$, $\alpha_3 = .025$, $\alpha_4 = .033$, $\alpha_5 = .050$, $\alpha_6 = .100$
    - *1)* $p_1 = .002 < \alpha_1 = .017$
    - *2)* $p_2 = .005 < \alpha_2 = .020$
    - *3)* $p_3 = .018 < \alpha_3 = .025$
    - *4)* $p_4 = .024 < \alpha_4 = .033$
    - *5)* $p_5 = .055 > \alpha_5 = .050$
    - *6) Also declare* $p_6 = .141$ not significant since p_5 *was not significant*
- Statistically significant ($p \leq \alpha$):
    - p_A, p_B, p_C, p_E

Procedures for Controlling the False Discovery Rate

- Benjamini-Hochberg
  - Stepwise (not simultaneous) modified Bonferroni procedure
  - Rank the *p*-values (p_c) from smallest to largest ($p_1 \ldots p_C$)
  - α_c, starting at $c = C$, is set at $\alpha(c/C)$
    - Thus, the largest *p*-value (p_C) is compared against α
      - i.e., $(c/C)\ \alpha = (C/C)\ \alpha = \alpha$
  - If any test is significant, reject the null for this test and all nulls associated with smaller *p*-values; if any test is not significant go to the next stage of testing

FDR Example

- $\alpha = .10$
- $p_A = .018$, $p_B = .005$, $p_C = .024$, $p_D = .141$, $p_E = .002$, $p_F = .055$
- $C = 6$
- Ordered: $p_1 = .002$, $p_2 = .005$, $p_3 = .018$, $p_4 = .024$, $p_5 = .055$, $p_6 = .141$
- $\alpha_{PT} = \alpha(c/C)$
- $\alpha_1 = .017$, $\alpha_2 = .033$, $\alpha_3 = .050$, $\alpha_4 = .067$, $\alpha_5 = .083$, $\alpha_6 = .100$
  - *1)* $p_6 = .141 > \alpha_6 = .100$
  - *2)* $p_5 = .055 < \alpha_5 = .083$
  - *4)* Also declare p_4, p_3, p_2, and p_1 significant since p_5 was significant
- Statistically significant ($p \leq \alpha$):
  - p_A, p_B, p_C, p_E, p_F

Myths Regarding Multiplicity Control

- Planned or pre-registered tests don’t require multiplicity control
- Multiplicity control is only required in ANOVA (e.g., pairwise comparisons), but does not apply to multiple predictors in regression, multiple outcome variables, combinations of these types of tests, etc.
- Multiplicity control is only required if you do *lots* of tests
- If my tests are correlated, then I don’t need multiplicity control

Do We Need Multiplicity Control At All?

- Earlier it was stated that familywise or false discovery rate Type I error control are typically recommended within the field of psychology
- However, the case for NEVER imposing multiplicity control is pretty strong and is gaining momentum

Reason 1 : More Power

- If we don’t adjust for multiplicity, we have more power for testing our hypotheses
- TRUE … But …
  - Although this is one of the most common reasons provided for dumping multiplicity control, it has no theoretical justification
    - The very easy counter-argument is simply to power your study taking into account multiplicity control

Reason 2: Simplicity

- Hancock & Klockars (1996)
  - "If [multiplicity control was abandoned], virtually all multiple comparisons would be easily conducted with *t*-tests using liberal critical values, and the MCP researcher would be unemployed".
- Great point … researchers despise having to control for multiplicity because it's complicated, lowers power, etc.
- However, like power, simplicity is not a valid reason for letting go of multiplicity control

Reason 3: Subjectivity in Analysis

- Multiplicity control is, at best, sporadically applied
- Researchers, reviewers, editors, etc. cannot agree on when and how multiplicity control should be applied
- This subjectivity reinforces the need for more clarity regarding when (if ever) multiplicity control should be imposed,
- But, like power and simplicity, is not a strong justification for not adopting multiplicity control

Reason 4: Consistency

- Researcher A
  - Explores differences between Arts and Science students on Perfectionism ($C = 1$, $p = .03$)
  - $\alpha = \alpha_{FW} = \alpha_{PT} = .05$ (statistically significant, $p < \alpha_{PT}$)
- Researcher B
  - Explores pairwise differences between Arts, Science, Engineering, Nursing, Health and Humanities students on Perfectionism (C = 15, $p_{\text{Arts,Science}} = .03$)
    - $\alpha = \alpha_{FW} = .05$
    - $\alpha_{PT} = \alpha_{FW} / C = .05 / 15 = .003$ (Bonferroni)
      - Not statistically significant, $p_{\text{Arts,Science}} > \alpha_{PT}$
- These researchers have the same *p*-value, but come to different conclusions regarding the difference between Arts and Science students – inconsistent!

Reason 5: The Test of Interest is the Natural Unit of Analysis

- When discussing familywise/false discovery rate Type I error control, how do you decide upon a *family* of tests over which control will be applied?
  - Imagine a researcher who is evaluating the relationship between each of the Big 5 personality factors and blood flow in 7 brain regions ($C = 35$)
  - Further, the researcher is going to run this study in three different samples (5-10 year old kids, intro psych students, seniors) (C = 105)
  - The researcher is also going to collect blood flow at different times of day (morning, noon, afternoon, evening) (C = 420)
- So far, the researcher is conducting over 400 tests, and this might only be Study 1!

Reason 5: The Test of Interest is the Natural Unit of Analysis

- How do we break up these tests into *families*, in order to impose multiplicity control
  - Each brain area is a different family?
    - $\alpha_{PT} = \alpha_{FW}/C = .05/60 = .0008$
  - Each sub-group (kids, etc.) is a different family?
    - $\alpha_{PT} = \alpha_{FW}/C = .05/140 = .0003$
  - Each study is a separate family?
    - Study 1: $\alpha_{PT} = \alpha_{FW}/420 = .05/420 = .0001$
  - Number of tests the researcher conducts this year?
    - $\alpha_{PT} = \alpha_{FW}/C = \alpha_{FW}/? =$ really small!

Reason 5: The Test of Interest is the Natural Unit of Analysis

- *Careerwise* Type I Error Rate Control
  - O'Keefe (2003), and others, have suggested that if multiplicity control is the standard, then what is needed is *careerwise* control, in order to ensure that the number of tests a researcher conducts in his/her career is not related to the probability of a Type I error
- Although absurd, it follows logically from the premise of multiplicity control

Reason 6: There is No Such Thing as a Type I Error

- The whole premise of multiplicity control is that we need to control for situations in which we falsely reject a *true null hypothesis*
- Try to imagine a relationship being investigated in psychology where the true effect is null
    - E.g., $\rho = 0$, $\mu_1 - \mu_2 = 0$
- If you can, try to imagine a family that contains MULTIPLE null effects

Reason 6: There is No Such Thing as a Type I Error

- Cohen (1990):
  - *"The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), <u>is always false in the real world</u>. It can only be true in the bowels of a computer processor running a Monte Carlo study."*

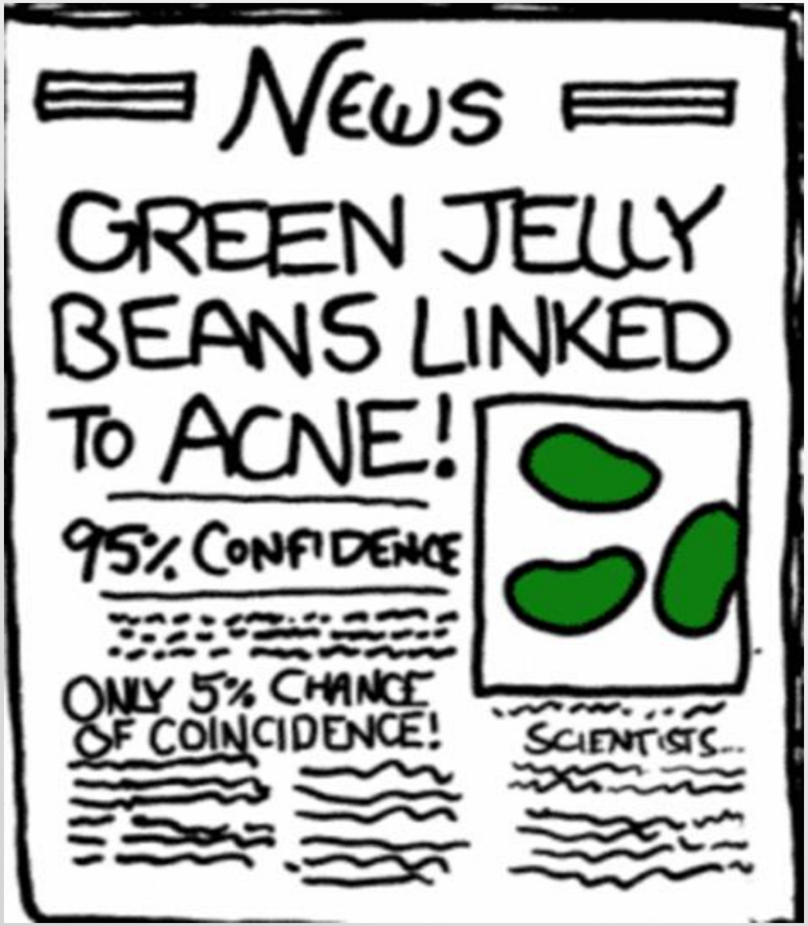
Reason 6: There is No Such Thing as a Type I Error

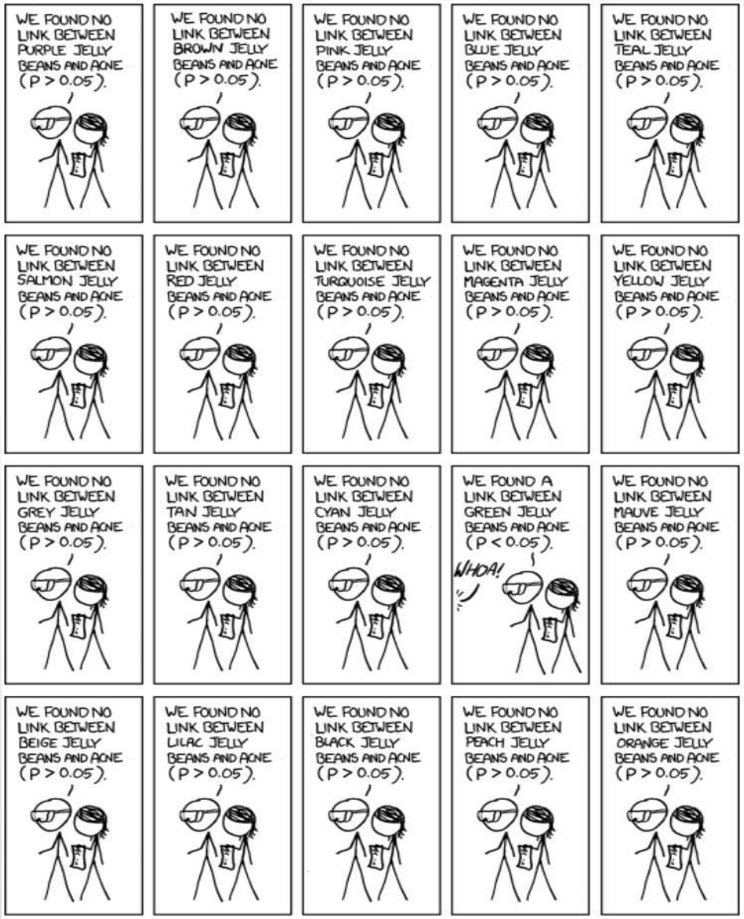
- Meehl (1990)
  - “Everything correlates to some extent with everything else”
  - Meehl also explains that he has found no competent psychologist that disputes this claim
- If there is no such thing as a Type I error, then what on earth are we controlling for?
- Researchers’ big worry should be power/Type II errors

Reason 7: Multiplicity Control is Not a Substitute for Replication

- Multiplicity control was designed to minimize the number of false positives that exist in the research literature
  - However, isn’t that the job of replication?
- Replication has solid theoretical support and is one of the key pillars of scientific enquiry

Reason 7: Multiplicity Control is Not a Substitute for Replication


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

- Would the green jelly beans be linked to acne in a replication?

Reason 8: Effect Sizes are the Primary Outcome of Research

- Researchers in the field of Psychology now treat effect sizes (with their accompanying confidence intervals) as the primary outcome of research studies
  - In other words, the focus is on the magnitude of the effects
  - Thus, null hypothesis significance testing now plays a reduced role in summarizing effects
- There is no need for multiplicity control in such a framework

Reason for Not
Abandoning Multiplicity Control

- There is one defense of multiplicity control that is worth discussing
  - Universal null hypothesis

Universal Null Hypothesis

- A clinical psychologist is conducting a general mental health check-up on a potential pilot, evaluating their status on depression, anxiety, bipolar disorder, personality disorders, etc.
  - Thus, individual null hypotheses are tested for depression, anxiety, etc., but there is also a *universal* null hypothesis that relates to general mental health
  - Imagine a completely healthy pilot with no mental health issues
    - Any Type I error for an individual hypothesis means a Type I error for the universal null hypothesis

Universal Null Hypothesis

- This sounds like a situation in which it is absolutely necessary to impose multiplicity control
- However, consider the following:
    - Imposing multiplicity control would violate the principle of consistency
    - There is no such thing as a Type I error
        - Find me anyone, not just a pilot, who is perfectly *normal (*if you can even define *normal)*
    - Replication … test the pilot regularly and look at the 'meta-diagnosis'