Rethinking multiple testing for replication and preregistration

Greg Francis

Open Science Conference
12 March 2019

Multiple testing

- Suppose your garden has an abundant supply of bryony
  - A poisonous weed
- You discover that bryony is used for eight different homeopathic treatments
  - ankle sprain, arthritis, backaches, painful breasts, broken bones, bruising, constipation, and coughs
- You are a bit skeptical about homeopathic approaches, so you decide to run a study to see whether bryony actually treats these problems

Multiple testing problem

- For a given set of data, multiple tests potentially inflate the probability of making at least one Type I error
- If you use a criterion of α for each of *k*=8 tests, then (if each null is actually true) the probability of a non-significant outcome is

$$(1-\alpha)$$

- The probability of 8 tests *all* tests being non-significant is

$$(1-\alpha)^8$$

- The probability of *at least one test* being significant is

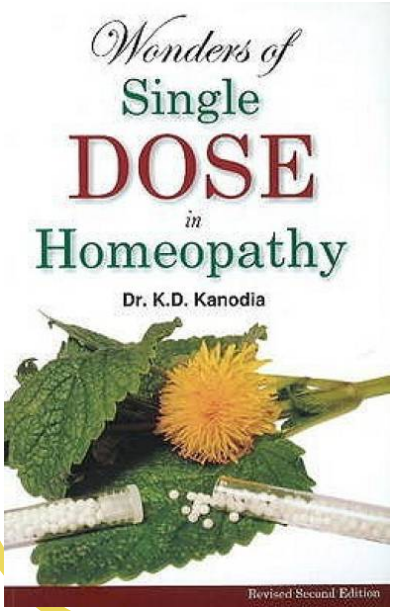
$$1-(1-\alpha)^8$$

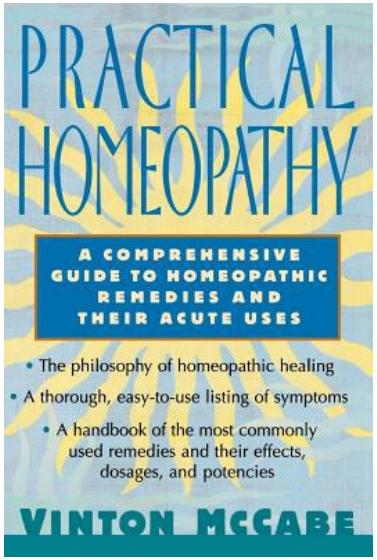
- If we use a 0.05 criterion, we have

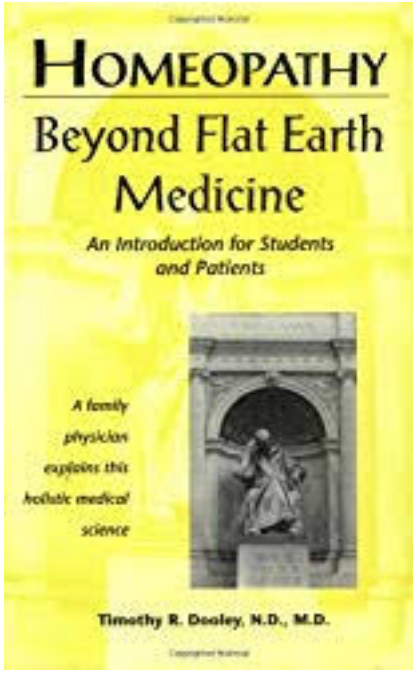
$$1-(1-0.05)^8 = 0.34$$

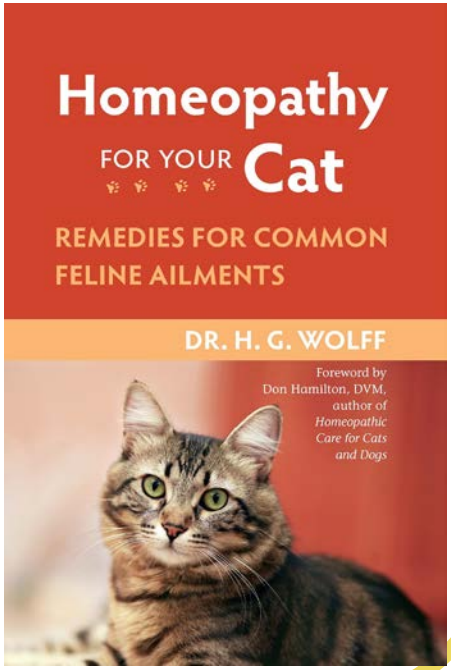
Multiple testing problem

- If we just keep testing bryony for problems where it does not actually work, we will sometimes incorrectly conclude that it does work
- If we then recommend that treatment, we will not actually be helping them
  - And (if the world is rational), your bryony business will fail









Bonferroni correction

- A common approach to dealing with the multiple testing problem is to adjust the criterion for significance
- We want an adjusted criterion, α_c, such that

$$1 - (1 - \alpha_c)^k = \alpha$$

- Šidák (1967)

$$\alpha_c = 1 - \sqrt[k]{1 - \alpha}$$

- To a very good approximation, this is almost the same as the Bonferroni correction

$$\alpha_c = \frac{\alpha}{k}$$

Bonferroni correction

- In our bryony example, we would use

$$\alpha_c = \frac{\alpha}{8} = \frac{0.05}{8} = 0.00625$$

- If the null is true for all (independent) tests, then the probability of generating at least one Type I error

$$1 - (1 - \alpha_c)^8 = 1 - (0.99375)^8 = 0.0489$$

- This kind of correction controls the Type I error rate as intended by taking into account the variety of relevant tests
  - Your samples are less likely to show bryony works as a treatment

ANOVA

- Similar issues pop up in lots of analyses
- For example, suppose you run a 2x2 independent ANOVA with $\alpha=0.05$ and you will make *some kind of conclusion* if you find any of the following:
  - A significant main effect for factor 1
  - A significant main effect for factor 2
  - A significant interaction
- What's the probability you will *make some kind of conclusion*, even if all population means equal each other?

ANOVA

- I don't know of an analytical solution, but you can run simulated experiments to discover that the probability of concluding *some effect* is 0.14
- In each simulation, you generate "data" by sampling from a normal distribution

```
Condition1A  <- rnorm(20, mean=0, sd=1)
Condition1B <- rnorm(20, mean=0, sd=1)
Condition2A  <- rnorm(20, mean=0, sd=1)
Condition2B  <- rnorm(20, mean=0, sd=1)
```

- And then running an ANOVA on the data

ANOVA

- For a given ANOVA, you simply check whether you find a significant:
  - Main effect for first factor
  - Main effect for second factor
  - Interaction
  - At least one test
- Repeat 10,000 times and see what proportion of tests produce significant outcomes

alpha = **0.05**
Type I error rate for main effect (first factor)= 0.0533
Type I error rate for main effect (second factor)= 0.0496
Type I error rate for interaction= 0.0464

Type I error rate for at least one test (main effect or interaction)= 0.1409

ANOVA

- Applying Bonferroni correction means using

$$\alpha_c = \frac{\alpha}{k} = \frac{0.05}{3} = 0.0167$$

- Which controls the Type I error rate for concluding *some* effect

alpha = **0.0166667**
Type I error rate for main effect (rows)= 0.017
Type I error rate for main effect (columns)= 0.0154
Type I error rate for interaction= 0.0163

Type I error rate for at least one test (main effect or interaction)= 0.0472

ANOVA

- Bonferroni is a bit conservative
  - The tests are not independent because they use the same data set
- With simulations, you could try different values of α until finding one that gives the desired Type I error for that design

alpha = **0.01725**
Type I error rate for main effect (first factor)= 0.0172
Type I error rate for main effect (second factor)= 0.0159
Type I error rate for interaction= 0.0175

Type I error rate for at least one test (main effect or interaction)= 0.0497

Correction cost

- If there *is* an effect, applying Bonferroni dramatically reduces experimental power
- Suppose one of the *k*=8 tests is based on a real effect
  - μ_0=0 and μ_a=1 with σ=2
  - And use *n*=40 observations for a one-sample t-test
- When using

$$\alpha_c = \frac{\alpha}{8} = \frac{0.05}{8} = 0.00625$$

- The power of **this** test is 0.61
- If you used α=0.05, the power would be 0.87

Conclusions matter

- With multiple tests, scientists need to protect against Type I error for *at least one comparison*
- The reasoning is that if a scientist finds even one significant result, they will conclude that there is support for an effect

- What if we turn it around?
- Sometimes scientists *require* multiple significant results before concluding support for an effect

Means and proportions

- Nairne, Thompson & Pandeirada (2007): Survival processing
  - Thinking about a word in terms survival leads to better memory than thinking about the word in terms of moving
  - Within subjects design
- Each subject provides a score for how many words are recalled in each condition
- Two (one-tailed) tests that are related to the same effect (survival vs. moving processing)
  - Test for mean differences across conditions (within subjects *t*-test)
  - Test the proportion of subjects that have a higher score for the survival than for the moving scenario (higher than 0.5?)

Means and proportions

- Suppose the null hypothesis is true
  - No difference in population means for survival and moving processing
- With (50,000) simulations, we find that the probability of having *both* tests produce a significant result using the α=0.05 criterion is only **0.02282**
  - Very conservative!
- Bonferroni correction makes things worse, **0.01202**

$$\alpha_c = \frac{\alpha}{k} = \frac{0.05}{2} = 0.025$$

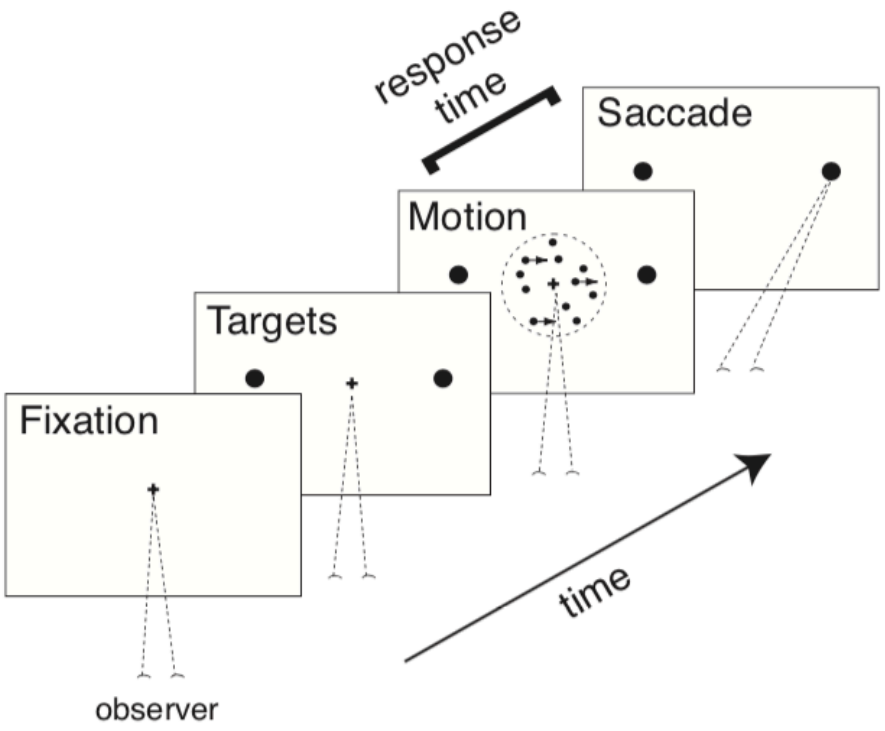
Reverse correction

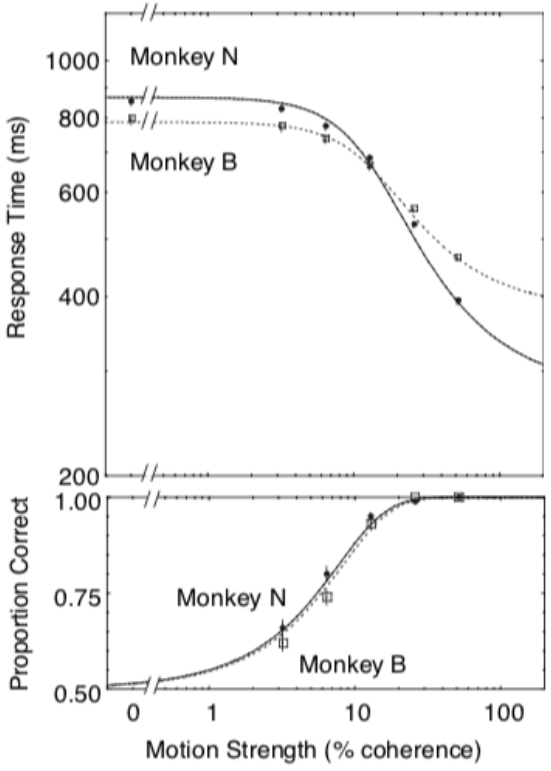
- To control the Type I error rate based on both tests, we should *relax* the significance criterion for individual tests
- Simulations show that using α=0.085 leads to a Type I error rate of **0.05058**
- Where Type I error means that *both tests* produced significant results at the 0.085 level

- An extra constraint on the data must be harder to achieve

RT and accuracy

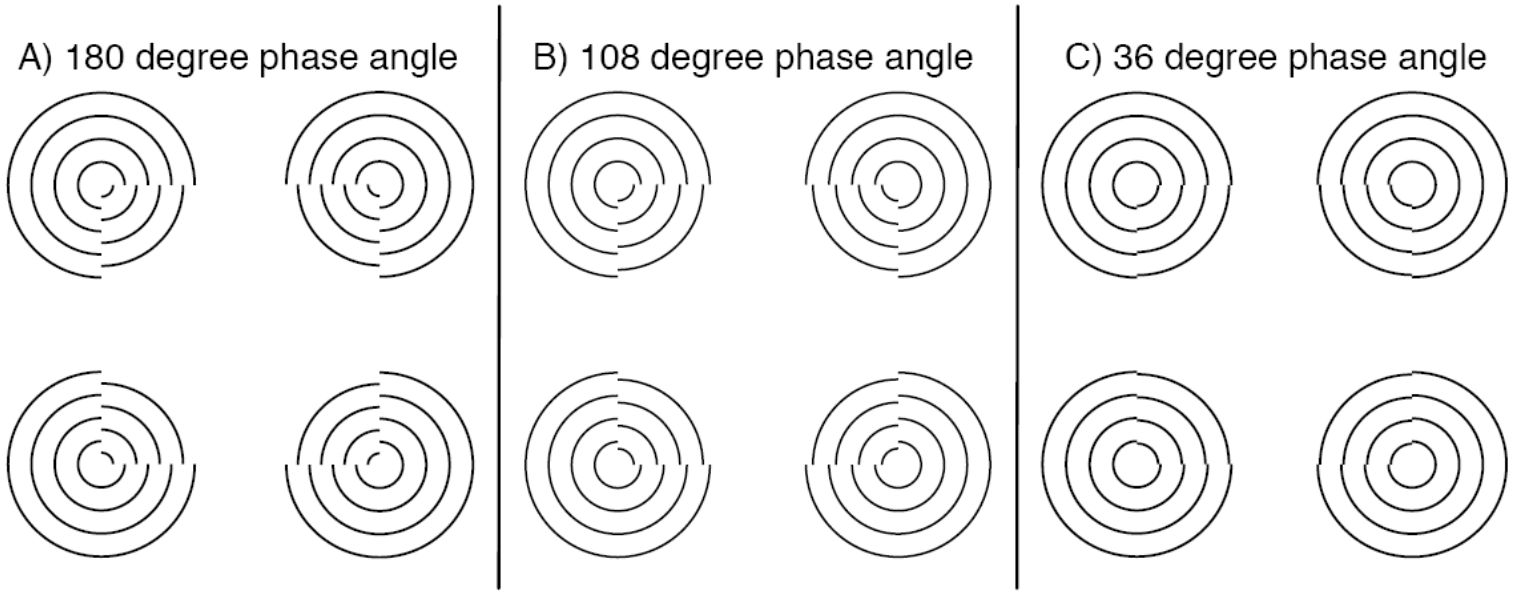
- There are situations where reaction time and accuracy are expected to go together
- As stimulus “strength” (e.g., contrast or coherency) increases, reaction time decreases and percent correct identification increases
- Palmer, Huk & Shadlen (2005)





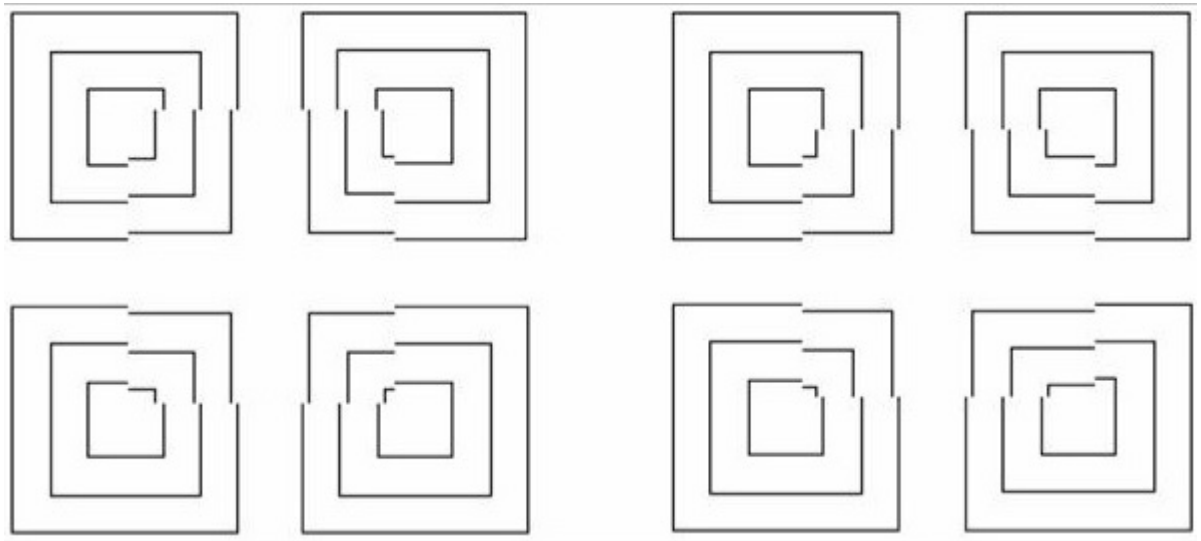
RT and accuracy

- Suppose a scientist wants to investigate the “strength” of illusory contours (Francis & Wede, 2010)



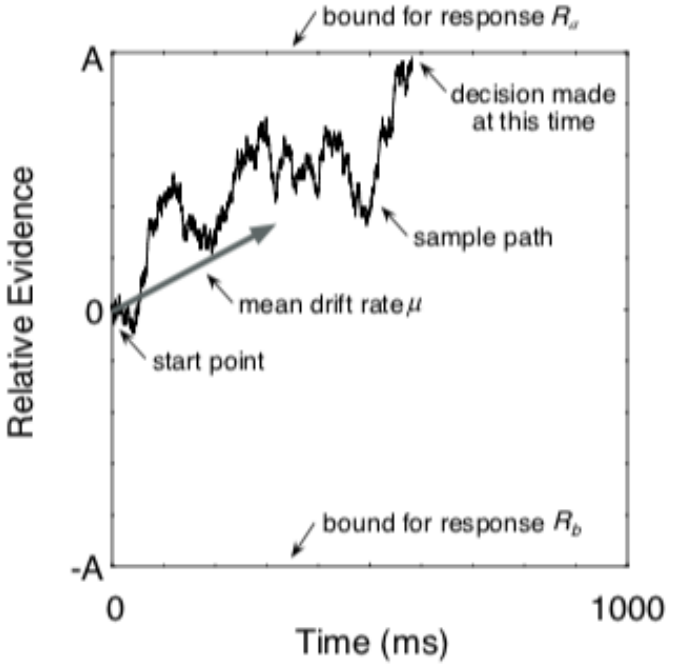
RT and accuracy

- Suppose a scientist wants to investigate the “strength” of illusory contours
- Discriminate vertical from horizontal illusory rectangle
- Vary the strength of the illusory rectangle



RT and accuracy

- We can generate simulated data using a version of the diffusion model (*rtdists*, Singmann et al., 2018)
- We assume two stimulus "strengths": strong and weak
- If the null is true, both conditions sample from a model with the same parameters
- To conclude an "effect" of stimulus strength, we require both
  - Significant (one-tailed) decrease in RT for strong compared to weak stimulus
  - Significant (one-tailed) increase in accuracy for strong compared to weak stimulus



RT and accuracy

- Suppose the null hypothesis is true
  - No difference in populations for strong or weak stimuli
- With (50,000) simulations, we find that the probability of having *both* tests produce a significant result using the α=0.05 criterion is only **0.00212**
  - Very conservative!
- Bonferroni correction makes things worse, **0.0006**

$$\alpha_c = \frac{\alpha}{k} = \frac{0.05}{2} = 0.025$$

Reverse correction

- To control the Type I error rate based on both tests, we should *relax* the significance criterion for individual tests
- Simulations show that using $\alpha=0.21$ leads to a Type I error rate of **0.04648**
- Where Type I error means that both tests produced significant results at the 0.21 level

- An extra constraint on the data must be harder to achieve

Within and between

- Facial feedback hypothesis
- Controversy about replication studies
  - No effect
  - Different methods?
- Marsh, Rhoads & Ryan (2018)
- Within subjects design, between subjects ordering of condition
- Argued there is a facial feedback effect on the basis of two tests from the same data
  - Significant within-subjects comparison of conditions
  - Significant between-subjects comparison of conditions for 1st trial only (more similar to original study)

Within and between

- Suppose the null hypothesis is true
  - No difference in populations for strong or weak stimuli
- With (50,000) simulations, we find that the probability of having *both* tests produce a significant result using the α=0.05 criterion is only **0.01952**
  - Very conservative!
- Bonferroni correction makes things worse, **0.00774**

$$\alpha_c = \frac{\alpha}{k} = \frac{0.05}{2} = 0.025$$

Reverse correction

- To control the Type I error rate based on both tests, we should *relax* the significance criterion for individual tests
- Simulations show that using α=0.103 leads to a Type I error rate of **0.04886**
- Where Type I error means that both tests produced significant results at the 0.103 level

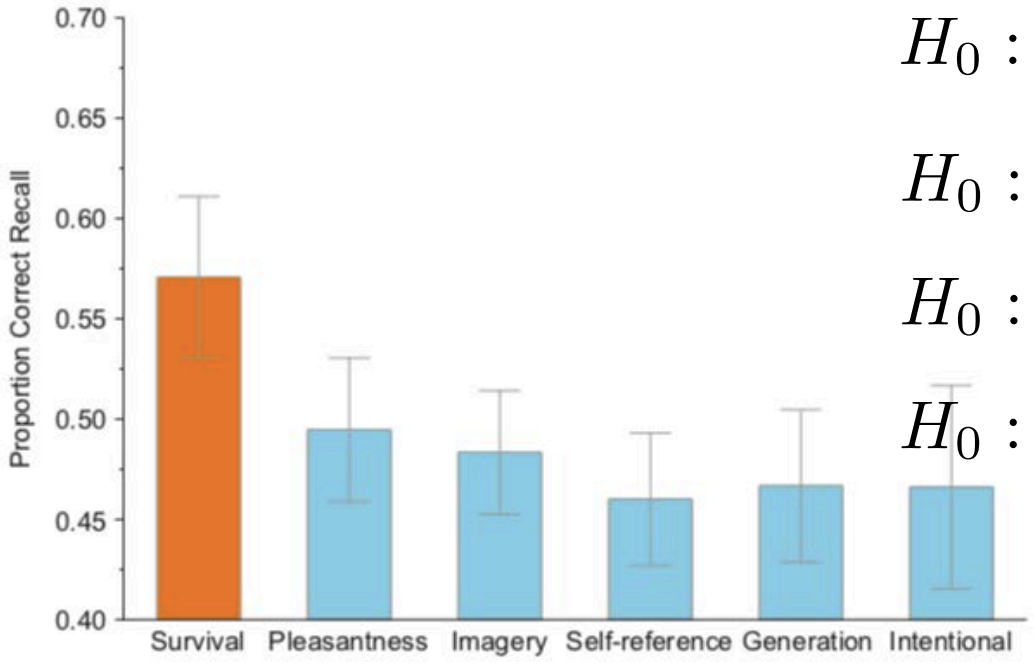
- An extra constraint on the data must be harder to achieve

Using reverse correction

- In every case, the two tests work together to draw a coherent conclusion
- If an “effect” is present for one test it must also logically be present for the other test (except for sampling variability)
  - Difference of means across conditions and difference of proportion of subjects showing advantage for one condition
  - Difference of RT and difference of accuracy
  - Within and between subjects comparisons for the same effect
- This requirement prevents use of reverse correction in many situations

Don’t use reverse correction

- Nairne, Pandeirada & Thompson (2008): Survival processing is the **best** of a variety of memory techniques (independent samples)
  - Must reject 5 hypothesis tests



$$H_0 : \mu_{\text{Survival}} = \mu_{\text{Pleasantness}}$$

$$H_0 : \mu_{\text{Survival}} = \mu_{\text{Imagery}}$$

$$H_0 : \mu_{\text{Survival}} = \mu_{\text{Self-reference}}$$

$$H_0 : \mu_{\text{Survival}} = \mu_{\text{Generation}}$$

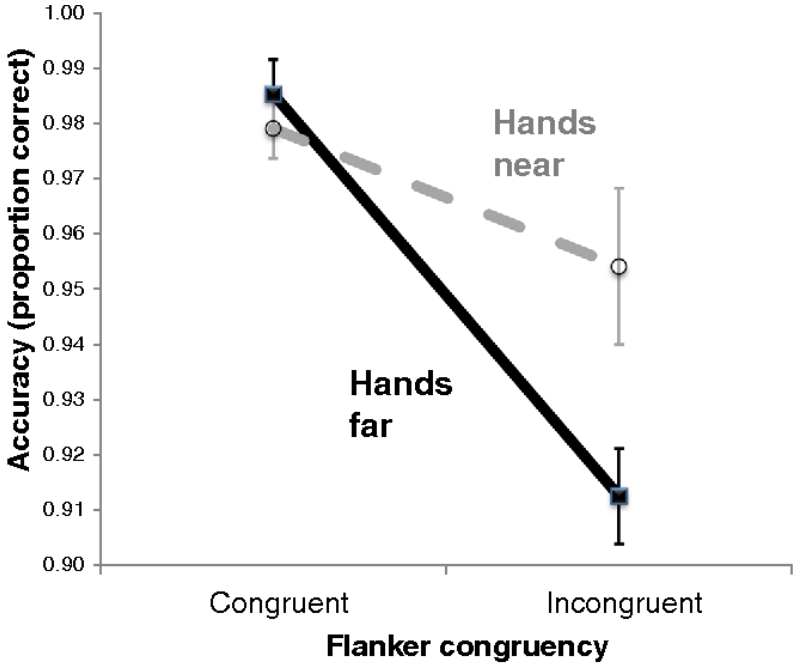
$$H_0 : \mu_{\text{Survival}} = \mu_{\text{Intentional}}$$

Don’t use reverse correction

- As before, it is the case that if all the null hypotheses are true, the Type I error rate for *all tests* is very low: 0.00156
- You can set α=0.285 to make the Type I error rate for this situation be approximately 0.05
- However, it is quite plausible that some null hypotheses are true and other null hypotheses are false
- If even only 1 null hypothesis is true, then the overall conclusion is an error
- Using the adjusted criterion, the Type I error rate could be as high as 0.285!

Don’t use reverse correction

- Examine enhanced cognition near the hands (Weidler & Abrams, 2014)
  - Eriksen flanker task as a measure of cognitive function
- Two tests:
  - Significant interaction
  - Significant contrast for incongruent trials



Don’t use reverse correction

- If all population means are equal to each other, using α=0.05 for both the test of interaction and the contrast is going to have a quite small Type I error rate
- You could adjust α to be larger and establish a 0.05 Type I error rate
- However, it could be that there *is* an interaction (thus, the test would have high power) but not the contrast of interest (or vice-versa)
- Your Type I error would be close to the adjusted α value

Replication

- One response to the replication “crisis” is a call for more replications
  - Often to double-check a finding
- E.g. Chiang, Shivacharan, Wei, Gonzales-Reyes & Durand (2018): ephaptic coupling for neurotransmission in the brain

The review committee at *The Journal of Physiology* – in which the research has been published – insisted the experiments be completed again before agreeing to print the study.

Durand et al. dutifully complied, but sound pretty understanding of the cautiousness, all things considered, given the unprecedented weirdness of the observation they're reporting.

Replication

- One response to the replication "crisis" is a call for more

Participants from the political left, right, and center (N = 1,979) completed a perceptual judgment task in which words were presented in different shades of gray. Participants had to click along a gradient representing grays from near black to near white to select a shade that matched the shade of the word. We calculated accuracy: How close to the actual shade did participants get? The results were stunning. Moderates perceived the shades of gray more accurately than extremists on the left and right ($p = .01$).

We conducted a direct replication while we prepared the manuscript. We ran 1,300 participants, giving us .995 power to detect an effect of the original effect size at $\alpha = .05$. The effect vanished ($p = .59$).

Our immediate reaction was **"why the #&@! did we do a direct replication?"** Our failure to replicate does not make definitive the conclusion that the original effect is false, but it raises enough doubt to make reviewers recommend against publishing.

Multiple testing

- For these types of replications, there are multiple tests, with the requirement that *both* experiments must produce a significant result
  - Non-significance in *either* experiment would force researchers to not conclude evidence for an effect
- This interpretation has dramatic impacts on the properties of the hypothesis tests

Type I error

- Occurs when a random sample of data produces a "significant" outcome even though the null hypothesis is true
- For many people, the point of hypothesis testing is to control the rate of Type I errors
  - $p<0.05$
- Suppose I require a significant outcome for two independent samples: an original study and a replication
- If the null is true, then the probability of getting two significant outcomes is 0.05 x 0.05 = 0.0025
- So, requiring replication success is very stringent in terms of Type I error

Reverse Bonferroni

- Suppose you wanted the Type I error rate across two studies to be 0.05
- You could require that *each* independent study produces a significant outcome with a criterion of

$$\alpha_r = \sqrt{0.05} \approx 0.22$$

- If either study produces a non-significant result, you do not conclude there is an effect
- With μ_0=0, μ_a=1, and σ=2 with *n*=25, each test has power=0.89
- The probability of both tests producing a significant result is 0.89 × 0.89= 0.79
- Note, you would still do better running a single test with *n*=50 and α=0.05 (Power=0.93)

Reverse Bonferroni

- If you require *k* independent tests to produce a significant result in order to conclude that there is “an effect”, then to have the Type I error rate for that conclusion to be α set the criterion for each independent test to be

$$\alpha_r = \sqrt[k]{\alpha}$$

Power

- Consider one-sample *t*-tests with null and alternative populations having $\mu_0=0$ and $\mu_a=1$ with $\sigma=2$
- If you use two smaller samples, $n=25$, and increase the criterion to $\alpha=0.05$, the probability of a significant out come for *each* test is 0.670, but the probability of both tests being significant is $0.670 \times 0.670 = 0.449$
- With $n=50$ and $\alpha=0.0025$, the power is 0.636
  - One test with a criterion of $\alpha=0.0025$ has more power than requiring two significant tests each with a criterion of $\alpha=0.05$
- Neyman-Pearson lemma

Conclusions

- Multiple testing for at least one significant outcome leads to an increase in Type I error
  - Resolved by reducing the significance criterion
- Multiple testing for a set of outcomes leads to a decrease in Type I error
  - Resolved by increasing the significance criterion
  - Requires test to address the same theoretical conclusion
- Multiple testing across replications behaves similarly
  - Resolved by increasing criterion used for each test
  - Oftentimes there are better choices
- Reversing Bonferroni