Multiple Testing When Many P-Values are Uniformly Conservative

Qingyuan Zhao, Dylan S. Small, and Weijie Su (2019)

Presented by: Giora Simchoni

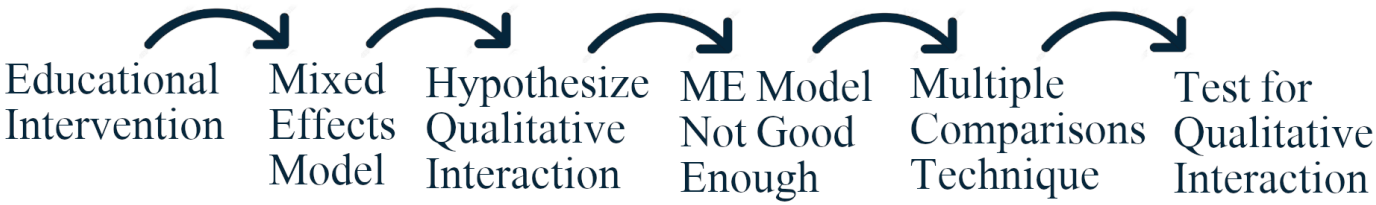
Multiple Comparisons and Selective Inference Sem. A 2020

Stat. and OR Department, TAU

2020-11-05

At Highest Level

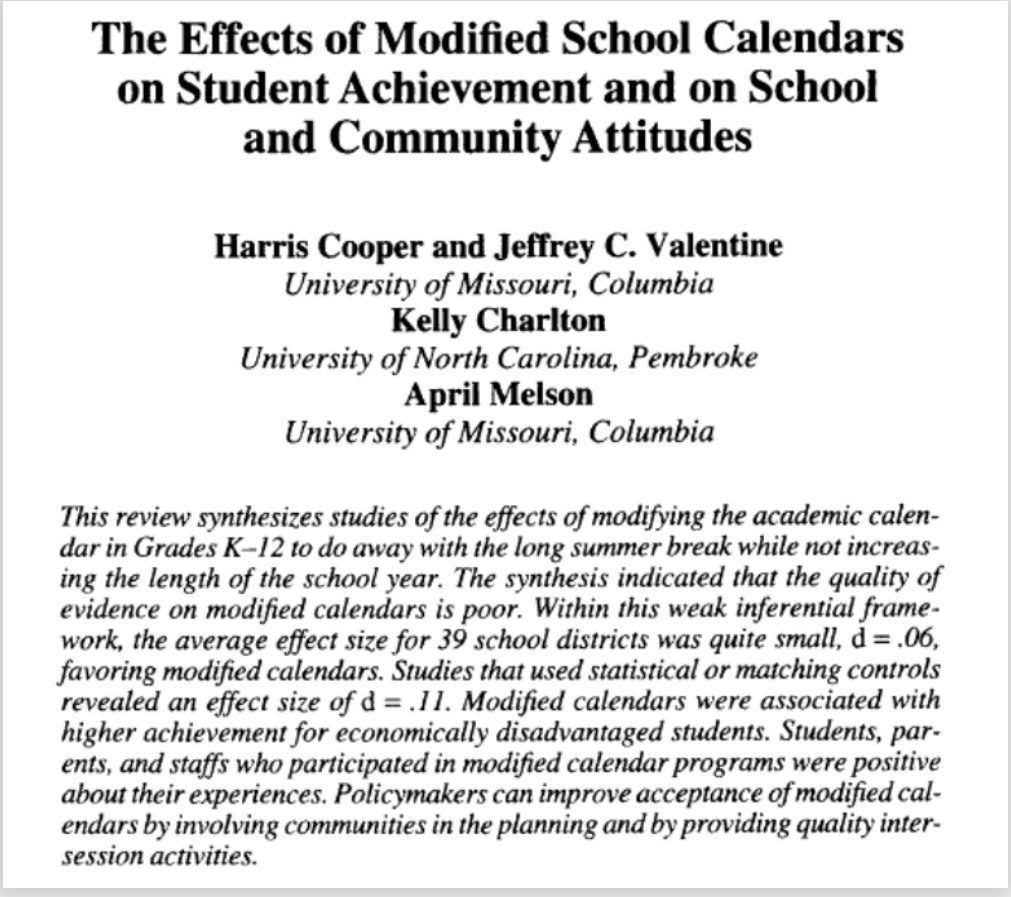
There's a Story



Motivation

Educational Intervention

Cooper et al. (2003) Meta-Analysis:



The Effects of Modified School Calendars on Student Achievement and on School and Community Attitudes

Harris Cooper and Jeffrey C. Valentine
University of Missouri, Columbia
Kelly Charlton
University of North Carolina, Pembroke
April Melson
University of Missouri, Columbia

This review synthesizes studies of the effects of modifying the academic calendar in Grades K–12 to do away with the long summer break while not increasing the length of the school year. The synthesis indicated that the quality of evidence on modified calendars is poor. Within this weak inferential framework, the average effect size for 39 school districts was quite small, d = .06, favoring modified calendars. Studies that used statistical or matching controls revealed an effect size of d = .11. Modified calendars were associated with higher achievement for economically disadvantaged students. Students, parents, and staffs who participated in modified calendar programs were positive about their experiences. Policymakers can improve acceptance of modified calendars by involving communities in the planning and by providing quality intersession activities.

A Typical (Simplified) Single-Study Model

$$y_j = \beta_0 + \beta_M X_j + \epsilon_j$$

Where:

$$y_j = \text{Student j "achievement"}$$

$$X_j = \begin{cases} 1, & \text{Student j in Modified Calendar} \\ 0, & \text{Otherwise} \end{cases}$$

$$\epsilon_j \sim N(0, \sigma^2)$$

And we're interested in:

$$H_0 : \beta_M \leq 0 \; vs. \; H_1 : \beta_M > 0$$

(Yes, in some studies this could literally be a t-test)

A Typical (Simplified) Meta-Study Model

- Observe n studies for which we only have the bottom line, such as treatment effect β_{M_i} or its p-value p_i
- Each β_{M_i} comes with its own scale or σ_i
- Is there a "global" effect?
- One approach is to take the *effect sizes* $d_i = \frac{\beta_{M_i}}{\sigma_i}$
- Let $\mu_d = E(d)$
- Use a single sample t/z-test to test the global null hypothesis:

$$H_0 : \mu_d \leq 0 \; vs. \; H_1 : \mu_d > 0$$

A (Still Simplified) Meta-Study Mixed Model

$$y_i = \mu + b_{S_i} + \epsilon_i$$

Where:

$$y_i = \text{Study i treatment effect or } \beta_{M_i}$$

$$b_{S_i} \sim N(0, \sigma_S^2) = \text{Study i random effect}$$

$$\epsilon_i \sim N(0, \sigma_i^2)$$

Where σ_i^2 is known from Study i.

Still the global null hypothesis would be:

$$H_0 : \mu \leq 0 \; vs. \; H_1 : \mu > 0$$

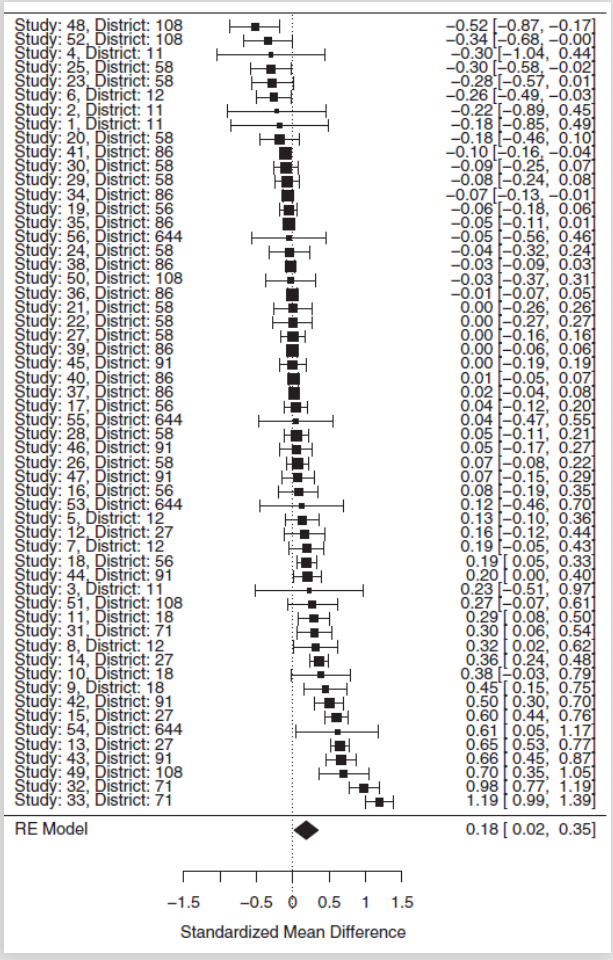
Result: "quite small"

Cooper et al.

TABLE 3

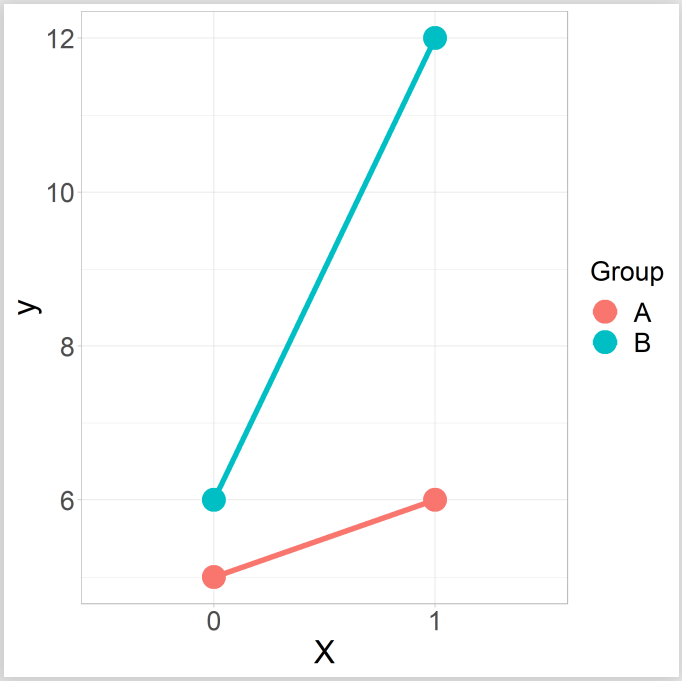
Stem and leaf display of unweighted achievement effect sizes

Stem	Leaf
+.7	8
+.5	34
+.4	3
+.3	2448
+.2	22359
+.1	002223
+.0	3479
−.0	033556779
−.1	00
−.2	014
−.3	7
−.4	
−.5	6

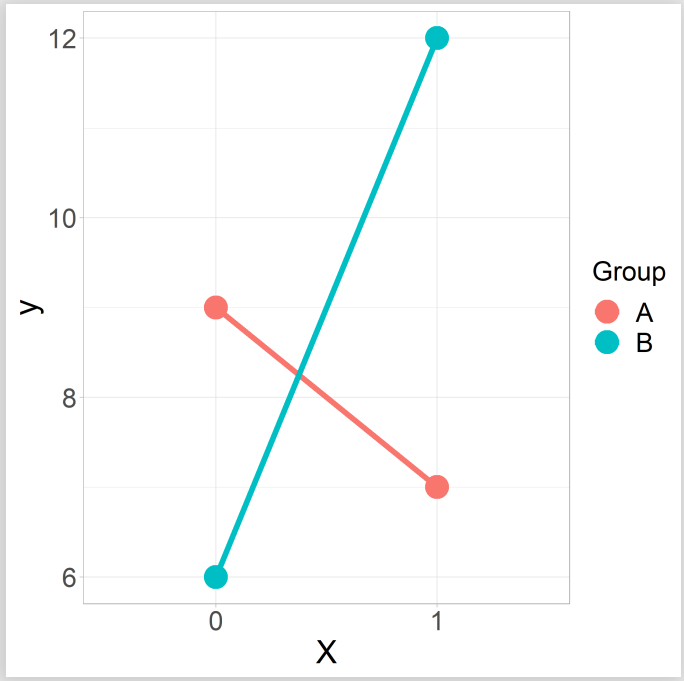


But could there be a (qualitative) interaction?

Quantitative (Ordinal) Interaction



Qualitative (Disordinal) Interaction



Why is the ME model unsuitable for testing qualitative interaction?

- $\sigma_S^2 > 0$ significantly, was found. What does that mean?
- So, go to a Fixed Effects model and make school/district a categorical variable with ~50 levels?...
- Can the ME model incorporate interaction? Can't guess apriori (and therefore put in the model) which groups (schools, districts) belong to positive/negative treatment effect

Meta-Analysis as a Multiple Testing Problem

Meta-Analysis as MTP

- n subgroups or independent studies
- $y_i \sim N(\mu_i, \sigma_i^2)$ [Study i treatment effect or β_{M_i}]
- Single study null hypothesis: $H_{0i} = \{\mu_i \leq 0\}$
- Global null hypothesis: $H_0 = \cap_{i=1}^{n} H_{0i} = \{\mu_i \leq 0, \forall i\}$

$\Longrightarrow$

- Forget the t-test or ME model
- Get the p-value p_i from each study for H_{0i}
- Treat $p_1, \ldots, p_n$ with your favorite MTP handler: Bonferroni, Fisher, BH, ...
- Test the global null H_0 accordingly
- E.g. with Bonferroni reject H_0 if $min(p_i) \leq \alpha/n$

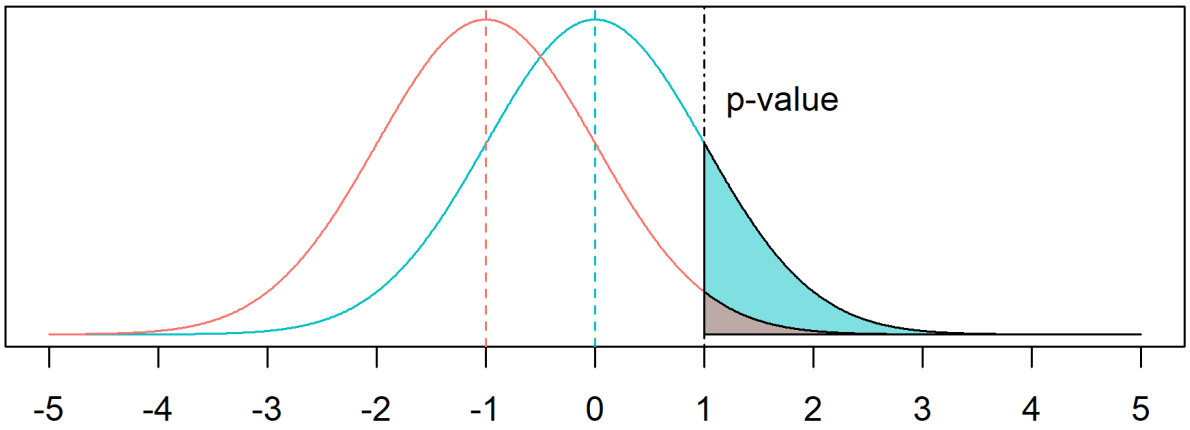
Testing a Global Null

- Think of any global test that is associated with a series of p-values adjusting function $p : [0,1]^n \to [0,1]$
- Bonferroni is really: $p^B(p_1, \ldots, p_n) = \min(n \cdot \min_i p_i, 1)$

```
p_global_bonferroni <- function(p_vals) {
  n <- length(p_vals)
  return(min(n * min(p_vals), 1))
}
```

- Looks good, right?

The Problem of Conservative Tests

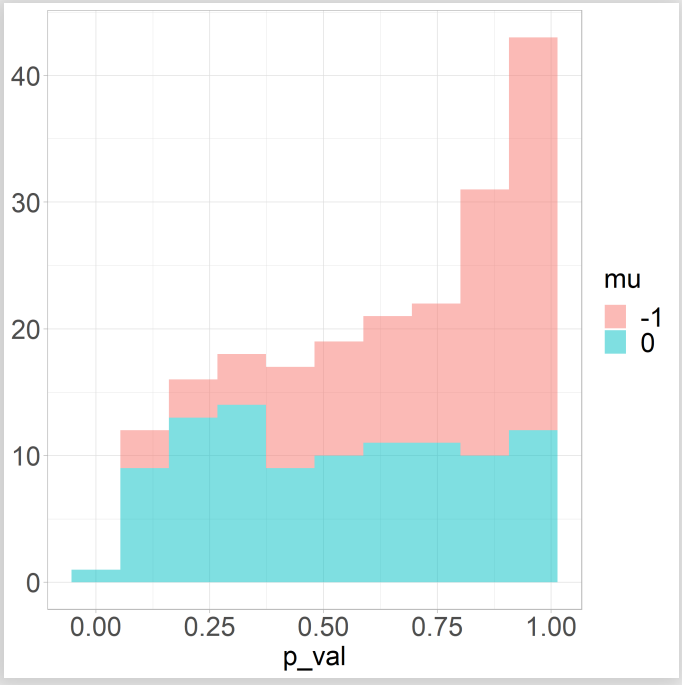


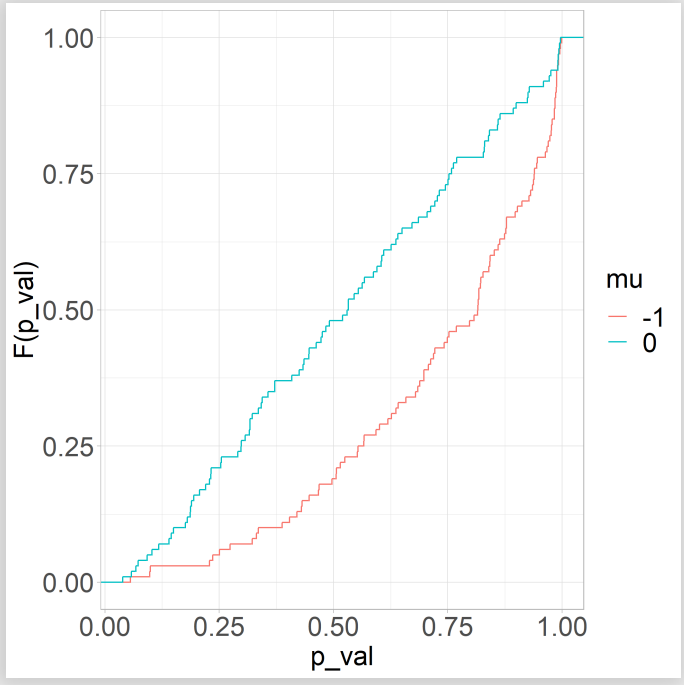
The Problem of Conservative Tests

- Suppose $Y_1, \ldots, Y_{100}$ are $N(\mu_i, 1)$ RVs
- Global null: $H_0 = \cap_{i=1}^{n} H_{0i} = \{\mu_i \leq 0, \forall i\}$
- Observe $y_1, \ldots, y_n$
- p-value would be: $p_i = P_{H_{0i}}(Y_i > y_i) = 1 - \phi(y_i)$
- Calculate $p_1, \ldots, p_n$
- If in reality $\{\mu_i = 0, \forall i\} \implies p_i \sim U(0, 1)$
- If in reality e.g. $\{\mu_i = -1, \forall i\} \implies p_i \succ U(0, 1)$
- a.k.a p-values have stochastically larger distribution than $U(0, 1)$
- a.k.a p-values are conservative, $P_{H_{0i}}(Y_i > y_i)$ "should be" smaller

How does conservative look like?

```
y1 <- rnorm(n = 100, mean = 0)
p1 <- 1 - pnorm(y1, mean = 0)
y2 <- rnorm(n = 100, mean = -1)
p2 <- 1 - pnorm(y2, mean = 0)
```





Qualitative Interaction as a MTP

Qualitative Interaction as MTP

- n subgroups or independent studies
- $y_i \sim N(\mu_i, \sigma_i^2)$
- Single study "positive" null: $H_{0i}^+ = \{\mu_i \geq 0\}$
- Global "positive" null: $H_0^+ = \cap_{i=1}^n H_{0i}^+ = \{\mu_i \geq 0, \forall i\}$
- Global "negative" null: $H_0^- = \cap_{i=1}^n H_{0i}^- = \{\mu_i \leq 0, \forall i\}$

$\implies$ Null hypothesis of NO qualitative interaction:

$$H_0 = H_0^+ \cup H_0^-$$

Reject H_0 if both H_0^+ and H_0^- are rejected at level α.

> 💡 Why don't we need a multiple comparisons correction here?

So we're good?

- Forget the t-test or ME model
- For H_0^+:
  - Get the p-value p_i from each study for H_{0i}^+
  - Adjust $p_1, \ldots, p_n$ with your favorite MTP handler: Bonferroni, Fisher, BH, ...
  - Test the global null H_0^+ e.g. with Bonferroni if $p_+^B \leq \alpha$
- Repeat for H_0^-, reject if both H_0^+ and H_0^- are rejected
- Could also report a global p-value which is $\max(p_+^B, p_-^B)$
- Done.

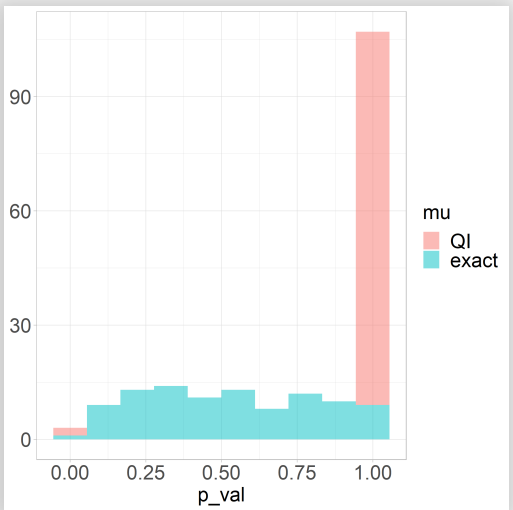
Qualitative Interaction Scenario

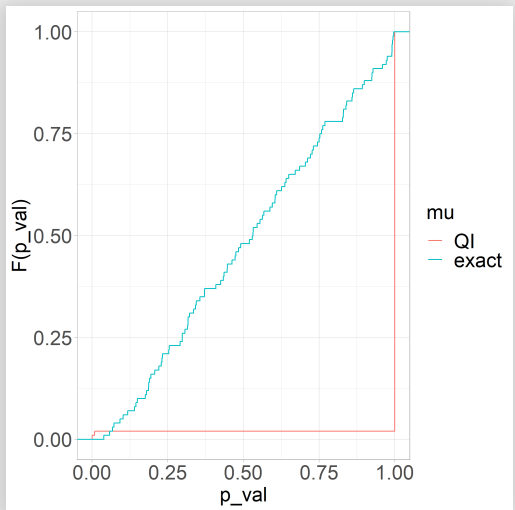
In reality $\mu_1 = \mu_2 = 3$ and $\mu_3 = \ldots = \mu_{100} = -10$

```
y3 <- rnorm(100, mean = c(3, 3, rep(-10, 98)))
p3 <- 1 - pnorm(y3, mean = 0)

signif(head(p3), digits = 2)
```

```
## [1] 0.00880 0.00073 1.00000 1.00000 1.00000 1.00000
```





Conservative p-values: particularly bad for QI

- H_0^- will never be rejected (Bonferroni (and friends) lose power!)
- H_0^+ will always be rejected

```
p_pos <- 1 - pnorm(y3, mean = 0)
p_neg <- pnorm(y3, mean = 0)

p_global_bonferroni(p_pos); p_global_bonferroni(p_neg)
```

```
## [1] 0.07271705
```

```
## [1] 1.297286e-32
```

```
max(p_global_bonferroni(p_pos), p_global_bonferroni(p_neg))
```

```
## [1] 0.07271705
```

- So H_0 of no QI will never be rejected! Can't "prove" QI when it clearly is the case.

To summarize

> Intuitively, if we do observe (p1, p2, p3 ... , p100) = (0.001, 0.001, 1, ..., 1), the first thing to be noticed is there are exceptionally many large p-values. This indicates many conservative tests. Naturally, we would like to "ignore" these large p-values and only use the two smaller ones, with which we can easily reject the global null. However, we cannot simply remove the large p-values because this would be data snooping and make the subsequent inference invalid.

Conditional Test

What are Zhao et al. suggesting?

- Given $p_1, \ldots, p_n$ independent p-values
- Set a fixed threshold parameter $0 < \tau \leq 1$
- Let $S_\tau = \{i | p_i \leq \tau\}$ (group of p-values smaller than τ)
- From basic probability: if p_i are exact and $p_i \sim U(0,1)$ then $p_i | p_i \leq \tau \sim U(0, \tau)$ then $p_i/\tau | \{i \in S_\tau\} \sim U(0,1)$
- Now take these *conditional* p-values $p_i/\tau | \{i \in S_\tau\}$ and perform your favorite MTP procedure p:

$$p(p_1, \ldots, p_n; \tau) = p(p_i/\tau | \{i \in S_\tau\})$$

- Where: $p(\emptyset) = 1$

What happens when $\tau = 1$?

Example: Conditional Bonferroni for a Global Null

- Reject a global null H_0 if $|S_\tau| > 0$ and:

$$\min_i (p_i/\tau) \leq \alpha/|S_\tau|$$

- Or as we put it "Conditional Bonferroni" p-value would be:

$$p^{CB}(p_1, \ldots, p_n; \tau) = \min(|S_\tau| \cdot \min_i (p_i/\tau), 1)$$

- In our earlier example for $\tau = 0.8$:

```
p_global_bonferroni(p3[p3 <= 0.8] / 0.8)
```

```
## [1] 0.001817926
```

Conditional Testing of Qualitative Interaction

```
tau <- 0.8
p_pos <- 1 - pnorm(y3, mean = 0)
p_neg <- pnorm(y3, mean = 0)

(p_pos_gl <- p_global_bonferroni(p_pos[p_pos <= tau] / tau))
```

```
## [1] 0.001817926
```

```
(p_neg_gl <- p_global_bonferroni(p_neg[p_neg <= tau] / tau))
```

```
## [1] 1.589175e-32
```

```
max(p_pos_gl, p_neg_gl)
```

```
## [1] 0.001817926
```

Back to Educational Intervention

Table 6. Combined p-values for qualitative interaction in the motivating applications in Section 1.1. Three versions of the Bonferroni's test and Fisher's combination test are used: the unconditional test ($\tau = 1$), the conditional test with threshold 0.5 and 0.8, and the conditional test with adaptively selected threshold τ.

		$\tau = 1$	$\tau = 0.8$	$\tau = 0.5$	τ adaptive
Modified calendar (school)	Bonferroni	0.044	0.034	0.039	0.033
	Fisher	0.224	0.004	0.005	0.003
	IBGA	0.042			
	LRT	0.011			
Modified calendar (district)	Bonferroni	0.347	0.158	0.189	0.245
	Fisher	0.788	0.088	0.113	0.281
	IBGA	0.274			
	LRT	0.351			
Writing-to-learn	Bonferroni	0.831	0.519	0.415	0.519
	Fisher	1	0.917	0.692	0.917
	IBGA	0.556			
	LRT	0.985			

Conditional Test: Assumptions

Defintion 1: Validity

A global test is *valid* if $P(p(p_1, \ldots, p_n) \leq \alpha) \leq \alpha$ for all $0 \leq \alpha \leq 1$ under the global null H_0

Is Conditional Bonferroni global p-value valid?

$$P(p^{CB}(p_1, \ldots, p_n) \leq \alpha) = P(|S_\tau| \cdot \min(p_i/\tau) \leq \alpha)$$
$$= P(|\{i|p_i \leq \tau\}| \cdot \min(p_i/\tau) \leq \alpha) = \ldots$$

Notice $|S_\tau|$ is a RV, so we're not in trivial land anymore.

though see *Theorem 2* for a proof the "conditional Bonferroni still controls the Type I error asymptotically when the test statistics `(Y_i)` are not independent but equally correlated"

Definition 2: Uniform Validity and Conservativity

A global test is *uniformly valid* if for all $0 < \tau < 1$ such that $P(p_i \leq \tau) > 0, p_i/\tau$ given $p_i \leq \tau$ is valid.

A p-value is called *uniformly conservative* if it is conservative and uniformly valid.

So, by definition:

Proposition 2: The conditional test using any fixed $0 < \tau \leq 1$ and any valid global test is also valid if $p_1, p_2, \ldots, p_n$ are independent and uniformly valid.

One conclusion is you can use the conditional test when your p-value test is unfirmly valid/conservative.

Wait, aren't all valid tests uniformly valid?

No.

$$P(p_i/\tau \leq \alpha | p_i \leq \tau) = \frac{P(p_i/\tau \leq \alpha \cap p_i \leq \tau)}{P(p_i \leq \tau)} = \frac{P(p_i \leq \tau\alpha)}{P(p_i \leq \tau)} \overset{?}{\leq} \alpha$$

- If p_i is exact as said $p_i | p_i \leq \tau$ is exact, this is worth α.
- If p_i is conservative then both numerator and denominator are smaller than $\tau\alpha$ and α respectively, and then what?
- E.g. if p_i is discrete and $P(p_i \leq \tau\alpha) = P(p_i \leq \tau)$
- And see section 7.2 for more interesting examples

Which Tests are Uniformly Valid/Conservative?

Any valid test p_i with CDF $F_i(x) = P(p_i \leq x)$ for which:

$$F_i(\tau x) \leq x F_i(\tau), \forall 0 \leq x, \tau \leq 1$$

Because:

$$P(p_i/\tau \leq x | p_i \leq \tau) = \frac{P(p_i \leq \tau x)}{P(p_i \leq \tau)} = \frac{F_i(\tau x)}{F_i(\tau)} \overset{?}{\leq} x$$

And whether it is uniformly conservative depends on $p_i/\tau | p_i \leq \tau$ being conservative or not.

Which means...

- Geometrically, this means that the function $F_i(x)$ is always below the segment from $(0, 0 = F_i(0))$ to $(\tau, F_i(\tau))$ if $0 \leq x \leq \tau$.

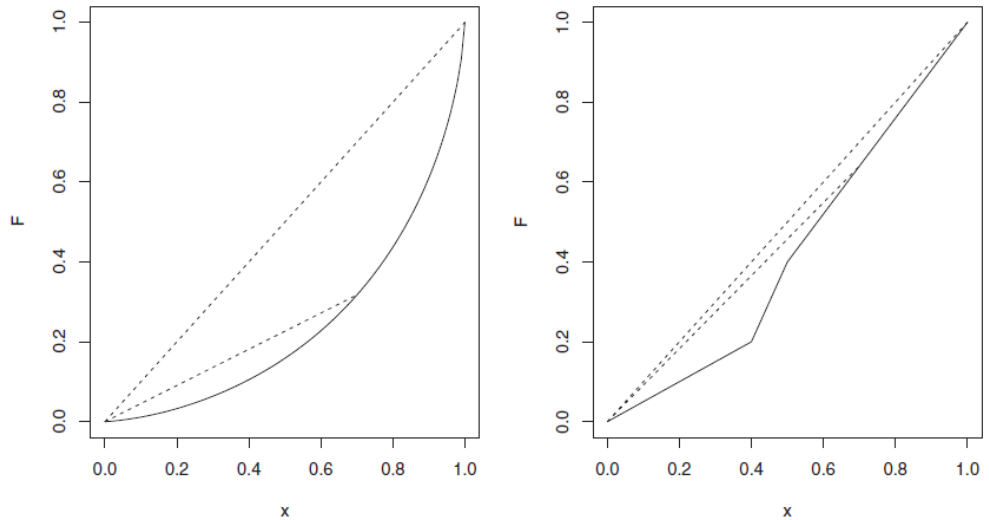
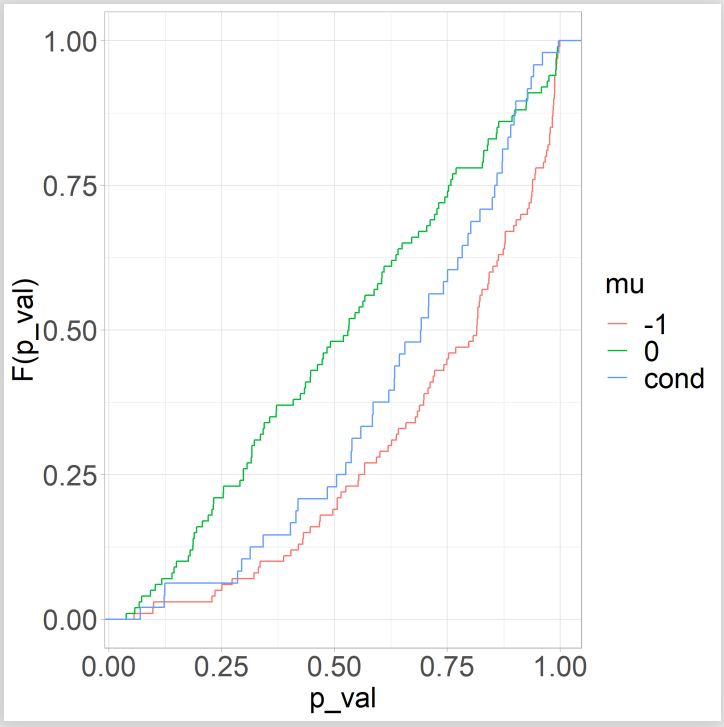


Figure 2. Two examples of uniformly conservative CDFs. The left plot is the distribution of $\Phi(Y)$ where $Y \sim \mathrm{N}(-1, 1)$. The right plot corresponds to a piecewise constant density function: $f(x) = 0.5 \cdot I(0 \leq x \leq 0.4) + 2 \cdot I(0.4 < x \leq 0.5) + 1.2 \cdot I(0.5 < x \leq 1)$. Both CDFs satisfy the condition $F(x\tau) \leq xF(\tau)$ for all $0 \leq x, \tau \leq 1$ so they are uniformly conservative. The geometric interpretation of this condition is illustrated by the two dashed lines corresponding to $\tau = 0.7$ and 1. The right plot suggests that convexity of CDF is not necessary for uniform conservativeness.

Our conditional test looks Ok



Which means...

- A sufficient condition (but not necessary) for uniform conservativeness is convexity of the CDF
- From Calculus: When the CDF $F(x)$ is differentiable, convexity of $F(x)$ is equivalent to the density $f(x)$ being monotonically increasing
- Where would we get a statistic $T(Y)$ with monotonically increasing f?
- From Statistical Theory: The one dimensional exponential family with parameter θ has monotone likelihood ratio (MLR) in statistic $T(Y)$, meaning for every $\theta_2 > \theta_1$ the likelihood ratio $f_{\theta_2}(y)/f_{\theta_1}(y)$ is a non-decreasing function of $T(Y)$, and the uniformly most powerful test at level α would be to reject $H_0 : \theta \leq \theta_0$ if $T(Y) \geq C$ and $P_{\theta_0}(T(Y) \geq C) = \alpha$

Which means...

- **Proposition 3:** When the true $\theta < \theta_0$, the UMP one-sided test of $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ in the one-dimensional exponential family is uniformly conservative
- So for example for $Y_i \sim N(\mu_i, \sigma_i^2)$ **where σ_i is known** and we're interested in the one sided test $H_{0i}^- : \mu_i \leq 0$. If the true $\mu_i < 0$ the UMP one-sided test using Y_i is uniformly conservative and we can go ahead and use the conditional test!

Conditional Test: Power Simulations

Will the global one-sided null be rejected?

"1 strong 99 very conservative"

```
n_sim <- 10000
reject <- numeric(n_sim); reject_cond <- numeric(n_sim)

for (j in 1:n_sim) {
  y <- rnorm(100, mean = c(4, rep(-4, 99)))
  p <- 1 - pnorm(y, mean = 0)
  reject[j] <- as.integer(p_global_bonferroni(p) < 0.05)
  reject_cond[j] <- as.integer(p_global_bonferroni(p[p <= 0.8]/0.8
}
mean(reject); mean(reject_cond)
```

```
## [1] 0.758
```

```
## [1] 0.9856
```

Will the global one-sided null be rejected?

Table 2. Power (in %) in the eight simulation settings in which the global tests (Bonferroni, Fisher, Tukey, and TruncatedP) are applied to the unconditional and conditional *p*-values. The truncation threshold τ is 0.5 or chosen adaptively as described in Section 3.3. We also considered the "union" two-step procedure as described in Section 2.3.

Setting	Method	$\tau = 1$	$\tau = 0.8$	$\tau = 0.5$	Adaptive
1. All null	Bonferroni	5.1	5.0	5.0	5.0
	Fisher	5.3	5.5	5.2	5.7
	Tukey	4.9	5.3	4.9	5.5
	TruncatedP	5.3	5.3	5.2	5.3
	Union	4.0			
2. 1 strong 99 null	Bonferroni	76.7	76.7	76.6	76.7
	Fisher	25.4	28.2	34.8	27.5
	Tukey	7.0	7.2	7.7	7.1
	TruncatedP	22.7	25.2	31.2	24.8
	Union	74.7			
3. 1 strong 99 conservative	Bonferroni	76.4	81.4	85.2	84.0
	Fisher	0.0	0.1	19.8	44.1
	Tukey	0.0	0.0	0.1	0.1
	TruncatedP	0.0	0.2	21.0	44.2
	Union	74.3			
4. 1 strong 99 very conservative	Bonferroni	75.9	98.7	98.0	98.7
	Fisher	0.0	98.5	98.0	98.3
	Tukey	0.0	0.0	0.0	0.0
	TruncatedP	0.0	98.6	98.0	98.5
	Union	80.2			
5. 1 strong 99 extremely conservative	Bonferroni	75.4	98.8	97.8	98.9
	Fisher	0.0	98.8	97.8	98.9
	Tukey	0.0	0.0	0.0	0.0
	TruncatedP	0.0	98.8	97.8	98.9
	Union	98.8			
6. 20 weak 80 null	Bonferroni	22.7	22.1	21.0	22.5
	Fisher	73.6	67.5	56.9	70.2
	Tukey	59.0	52.3	39.6	53.8
	TruncatedP	69.2	63.9	52.7	66.1

Will the QI null be rejected?

"50 positive 50 negative"

```
reject <- numeric(n_sim); reject_cond <- numeric(n_sim)

for (j in 1:n_sim) {
  y <- rnorm(100, mean = c(rep(1, 50), rep(-1, 50)))
  p_pos <- 1 - pnorm(y, mean = 0)
  p_neg <- pnorm(y, mean = 0)
  p_pos_gl <- p_global_bonferroni(p_pos)
  p_neg_gl <- p_global_bonferroni(p_neg)
  p <- max(p_pos_gl, p_neg_gl)
  reject[j] <- as.integer(p < 0.05)
  p_pos_gl <- p_global_bonferroni(p_pos[p_pos <= 0.8] / 0.8)
  p_neg_gl <- p_global_bonferroni(p_neg[p_neg <= 0.8] / 0.8)
  p <- max(p_pos_gl, p_neg_gl)
  reject_cond[j] <- as.integer(p < 0.05)
}
mean(reject); mean(reject_cond)
```

```
## [1] 0.1817
```

```
## [1] 0.2111
```

Will the QI null be rejected?

Table 4. Power (in %) of testing qualitative interaction in the seven simulation settings. The proposed methods—global tests (Bonferroni, Fisher, Tukey, and TruncatedP) applied to the unconditional and conditional *p*-values—are compared with two existing methods, the interval-based graphical approach (IBGA; Pan and Wolfe 1997) and the likelihood ratio test (LRT; Gail and Simon 1985). For the conditional global tests, the truncation threshold τ is 0.5 or chosen adaptively as described in Section 3.3.

Setting	Method	$\tau = 1$	$\tau = 0.8$	$\tau = 0.5$	Adaptive
1. 1 positive 99 null	Bonferroni	3.7	3.8	3.8	3.8
	Fisher	0.1	1.9	1.8	1.9
	Tukey	0.0	0.5	0.3	0.7
	TruncatedP	0.6	2.0	1.7	2.0
	IBGA	3.9			
	LRT	1.3			
2. 1 positive 1 negative	Bonferroni	60.0	60.1	59.9	60.1
	Fisher	1.1	8.4	12.0	8.3
	Tukey	0.1	0.9	0.4	1.0
	TruncatedP	2.7	7.6	9.4	7.8
	IBGA	60.4			
	LRT	12.6			
3. 1 positive 99 negative	Bonferroni	50.4	49.8	44.4	55.0
	Fisher	0.0	0.1	20.0	44.4
	Tukey	0.0	0.0	0.1	1.6
	TruncatedP	0.0	0.2	21.1	44.5
	IBGA	51.4			
	LRT	0.0			
4. 20 positive 80 negative	Bonferroni	11.6	14.5	13.8	16.0
	Fisher	0.0	13.8	47.8	39.5
	Tukey	0.6	6.7	27.6	21.9
	TruncatedP	0.2	24.9	49.7	44.1
	IBGA	12.0			
	LRT	3.9			
5. 50 positive 50 negative	Bonferroni	17.9	20.7	18.0	20.5
	Fisher	69.9	98.2	96.7	98.2
	Tukey	73.9	94.1	90.5	93.9
	TruncatedP	92.1	98.1	95.0	97.9

Conditional Test: Adaptive Threshold

How to choose τ without sacrificing the validity of the test?

- **Proposition 4:** Let $F_x = \sigma(\{p_i | p_i \geq x\})$ be the backward filtration for $0 \leq x \leq 1$. If $\tau = \tau(p_1, \ldots, p_n)$ is a backward stopping time in the sense that $\{\tau \geq x\}$ is F_x-measurable for any $0 \leq x \leq 1$, then Proposition 2 still holds.

- Which basically means (see proof) you're good if you (say, interactively):

  - decide on a decreasing sequence $\tau_1 > \ldots > \tau_K$
  - at stage k base your criterion of stopping **ONLY** on $\{p_i | p_i > \tau_k\}$ [the p-values you've discarded so far]
  - apply global test on $\{p_i / \tau | p_i \leq \tau\}$ if τ is the τ_k you've stopped at

What would be a good stopping criterion?

- Look again at "Conditional Bonferroni":

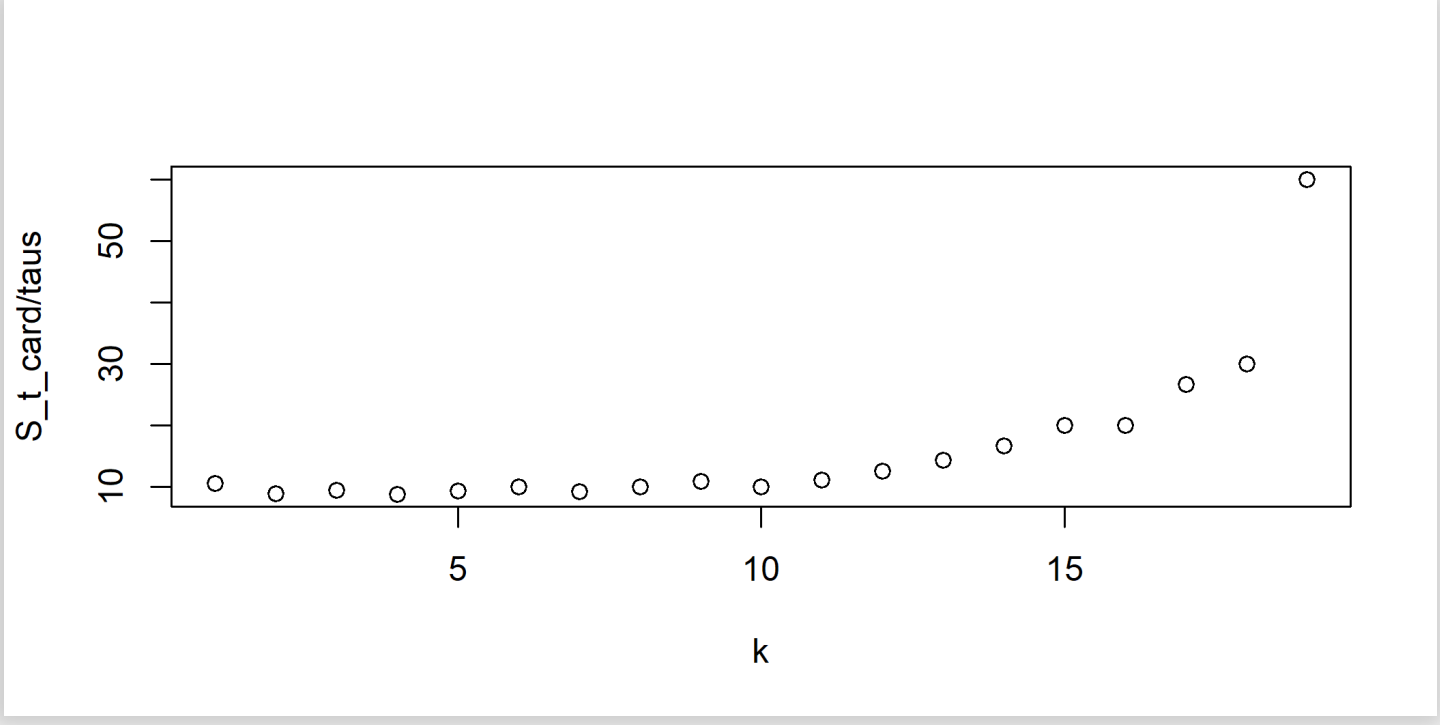
$$p^{CB}(p_1, \ldots, p_n; \tau) = \min(|S_\tau| \cdot \min_i (p_i/\tau), 1)$$

$$= \min(\frac{|S_\tau|}{\tau} \cdot \min_i (p_i), 1)$$

- We'd want to minimize $\frac{|S_\tau|}{\tau}$!
- Intuitively we'd want to discard as many p_i (i.e. decrease $|S_\tau|$) in the minimum amount of steps (i.e. increase τ)
- But how will we know we've reached the minimum $\frac{|S_\tau|}{\tau}$ "in real time"?

```r
p_values <- c(1.0, 1.0, 0.95, 0.9, 0.8, 0.7, 0.5, 0.2, 0.1, 0.01,
taus <- seq(0.95, 0.05, -0.05)
K <- length(taus)
S_t_card <- numeric(K)
for (k in 1:K) {
  S_t_card[k] <- length(p_values[p_values <= taus[k]])
}

plot(1:K, S_t_card / taus, xlab = "k")
```



A good step to stop in is a step in which the "discarding" is slowing down - $|S_\tau|$ stays the same while τ is decreasing. Or the derivative is not increasing.

Making sure the derivative doesn't increase:

- Let F be the (average) CDF of the p-values: $F(x) = (1/n)\sum F_i(x)$
- Then, $|S_\tau|/\tau = |i : p_i \leq \tau|/\tau \approx [nF(\tau)]/\tau$
- The derivative would be:

$$\frac{d}{d\tau}\frac{F(\tau)}{\tau} = \frac{f(\tau)\tau - F(\tau)}{\tau^2}$$

- So let's stop when there's no strong evidence that $f(\tau)\tau - F(\tau) > 0$

- We can only estimate $f(\tau)$ and $F(\tau)$ though, with window size $0 < \omega \leq 1 - \tau_1$:

$$\hat{F}(\tau) = \frac{|S_\tau|}{n}; \hat{f}\,(\tau) = \frac{|i|\tau \leq p_i \leq \tau + \omega|}{n\omega}$$

- The quantity $n\omega\hat{f}\,(\tau_k)$ counts the number of p-values in window size ω.
- If the p_i are uniform and independent [and we should stop!] this is a $Binomial(n, q\omega)$ variable where $q < \hat{F}(\tau_k)/\tau_k$
- Also this would mean $E(n\omega\hat{f}\,(\tau_k)) = nq\omega$ which would mean $E(\hat{f}\,(\tau_k)) = q < \hat{F}(\tau_k)/\tau_k$ and our derivative isn't increasing.

- final stopping criterion: if we fail to reject that $n\omega\hat{f}\left(\tau_k\right) \sim Binomial(n, q\omega)$ with $q < \hat{F}(\tau k)/\tau k$ with some significance level β

- The authors use the p-value of a Binomial test explicitly:

```
binom.test(f.hat[k] * n * width, n, width * F.hat[k] / tau.seq[k],
           alternative = "greater")$p.value
```

Algorithm 1

Algorithm 1: An adaptive algorithm to select the threshold τ.

Data: Uniformly valid p-values: $p_1, p_2, \ldots, p_n$.
Parameter: Window size w; decreasing sequence $1 - w \geq \tau_1 > \cdots > \tau_K > 0$; stopping criterion $0 < \beta < 1$.
Result: τ.

$k = 1$;
repeat
  $\hat{f}_k = |\{i | \tau_k \leq p_i \leq \tau_k + w\}|/(nw)$;
  $\hat{F}_k = |\mathcal{S}_{\tau_k}|/n$;
  if *failed to reject* $nw\,\hat{f}_k \sim \text{Binomial}(n, qw)$ *with* $q < \hat{F}_k/\tau_k$ *at level* β **then**
    break
  end
  $k \leftarrow k + 1$;
until $k = K$;
return τ_k.

Other Stuff

- Testing for Practical Importance
- When the p-values are dependent
- Power of Conditional Bonferroni over Bonferroni
- Beyond global testing