



Multiple Inference in Early Childhood Program Evaluations

David Deming
Harvard University



-
- i EC interventions have impacts in a number of domains
 - 1 Education, earnings, crime, health.....

 - i Also heterogeneity by subgroup (gender, race)

 - i If only interested in one outcome for one group (i.e. arrests for males), no need to adjust
 - 1 But researchers are rarely this disciplined!



-
- i With y outcomes and g groups, there can be $y * g$ hypothesis tests
 - 1 Tradition dictates the $p < .05$ standard
 - 1 But as the number of tests multiplies, the probability of a false rejection (Type I error) becomes non-trivial

 - i Will use Anderson's (JASA 2008) reanalysis of three early interventions to frame the discussion
 - 1 Particularly important for EC, where there are hypothesized effects in many domains



Confirmatory vs. Exploratory Hypotheses

- i Confirmatory hypotheses should be stated before randomized trial is conducted or data analysis begins
 - 1 Ex: Anderson looks at separate male and female results only, but wouldn't the pooled results also be of interest ex ante?

- i When hypothesis testing is exploratory (i.e. ex post), we should note this explicitly



2 Approaches to Multiple Inference in Anderson (2008)

- i Reduce the number of tests using a summary index
- i Adjust the p-values to correct for the increasing probability of a false rejection
 - 1 FWER and FDR



Summary Indices

- i Combine many outcomes into one (standardized and weighted) index
 - 1 Pr(false rejection) does not increase as outcomes are added
 - 1 Could be multiple measures of some latent variable (i.e. human capital)

- i Anderson organizes indices by life stage (Preteen, Teen, Adult) and intervention (ABC, PPP, ETP)
 - 1 MTO by outcome domain (employment, health, etc.)

Table 2. Summary index components

Project	Stage	Summary index components
ABC	Preteen	IQ (5, 6.5, 12), Retained in Grade (12), Special Education (12)
Perry	Preteen	IQ (5, 6, 10), Repeat Grade (17), Special Education (17)
ETP	Preteen	IQ (5, 7, 10), Retained in Grade (17), Special Help (17)
ABC	Teen	IQ (15), HS Grad (18), Teen Parent (19)
Perry	Teen	IQ (14), HS Grad (18), Unemployed (19), Transfers (19), Teen Parent (19), Arrested (19)
ETP	Teen	IQ (17), HS Dropout (18), Worked (18)
ABC	Adult	College (21), Employed (21), Convicted (21), Felon (21), Jailed (21), Marijuana (21)
Perry	Adult	College (27), Employed (27, 40), Income (27, 40), Criminal Record (27), Arrests (27), Drugs (27), Married (27)
ETP	Adult	College (21), Receive Income (21), On Welfare (21)



P-value adjustment: FWER

- ⌋ Familywise Error Rate (FWER)
 - 1 FWER is the probability that at least one true hypothesis is rejected
 - 1 Simplest method is Bonferroni (adjust p by $p * M$ where M is the number of hypotheses tested)
 - ⌋ Ex: 3 tests with t-stats 1.96, 1.4 and 1.0
 - ⌋ $\Pr(\text{joint null is true}) = 0.05 * 3 = 15\%$
 - 1 Anderson uses a more efficient method called free step-down resampling
 - 1 Well-suited to primary hypothesis testing (min $\Pr(\text{type I error})$)



		Female			
Project	Age	Effect	Naive p value	FWER p value	n
ABC	Preteen	.445 (.194)	.026	.125	54
Perry	Preteen	.537 (.177)	.004	.028	51
ETP	Preteen	.362 (.251)	.160	.349	30
ABC	Teen	.422 (.202)	.042	.156	53
Perry	Teen	.613 (.156)	0	.003	51
ETP	Teen	.456 (.299)	.138	.349	29
ABC	Adult	.452 (.144)	.003	.024	53
Perry	Adult	.353 (.150)	.022	.125	51
ETP	Adult	-.069 (.186)	.714	.701	29



P-value adjustment: FDR

i False Discovery Rate

- 1 FDR is the expected proportion of rejections that are Type I errors
- 1 If no hypotheses are rejected, $\text{FWER} = \text{FDR}$
- 1 True rejections make FDR less stringent than FWER
- 1 Suitable for exploratory analysis



Anderson (JASA 2008)

- i Adjustment for multiple inference can make a big difference
 - ┆ In a standalone reanalysis of Perry, 13 effects are significant using unadjusted p-values, but only 2 (5) reject at $p < 0.05$ ($p < .1$) when adjusted
- i But the assumptions are important
 - ┆ Particularly when choice of tests and outcomes is ex post



Anderson (JASA 2008)

- i Weighting: are all outcomes equal?
 - 1 Weights by covariance structure is great for statistical inference, but is agnostic about relative benefits

- i Aggregation across separate studies
 - 1 “Successful” studies get published and/or reanalyzed more often
 - 1 Could weight by treatment intensity (ABC > PPP > ETP)
 - i Make predictions based on this (effect sizes should be increasing in intensity)



Further Issues

- i How to weigh statistical inference against benefits and costs?
 - 1 For Perry and crime, B/C ratio is very high
 - 1 Skeptical view: Crime was not an ex ante outcome of interest (and would we have even known about it if effects were zero?)
 - 1 Sympathetic view: Given the B/C ratio, maybe we don't need to be 95% confident
 - i Also matches the pattern of results in early life to some degree
- i BOAPW – Lay out an ex ante theoretical framework that would predict a particular pattern of results, then test using new data



Gender Differences

- i Impacts on girls are more consistent across outcomes
 - 1 Effect size 0.27, se 0.09 for adult outcomes
 - 1 We can say with a high degree of confidence that these interventions benefit girls
- i Impacts on boys are less robust, but potential benefits are higher
 - 1 Effect size -0.05, se 0.11
 - 1 But estimated benefits are \$417k for males and \$100k for females (Table 7.8 in Belfield et al JHR)
- i Both are important for policy
 - 1 The confidence interval on the crime reduction benefits in Perry includes both \$500k and \$0
 - 1 But the expected benefits of intervention are much greater than zero



-
- i For more details, see IES Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations, by Peter Schochet:
<http://ies.ed.gov/ncee/pubs/20084018.asp>