USDA

United States Department of Agriculture

Deriving Statistical Verdicts from Null Findings: Why Null Hypothesis Statistical Testing Diminishes the Evidence Base to Inform Policy, and a Proposed Solution

Jason P. Brown[a], Dayton M. Lambert[b], and Timothy R. Wojan[c]
Federal Reserve Bank of Kansas City[a], University of Tennessee[b] and Economic Research Service/USDA[c]

Paper presented at the 2018 Federal Committee on Statistical Methodology Conference, Washington DC, March 8-11

The views expressed are those of the authors and should not be attributed to the Economic Research Service, USDA, the Federal Reserve Bank of Kansas City, the Federal Reserve System or the University of Tennessee.

Outline

- The problem of using null hypothesis statistical testing (NHST) to inform policy decisions
- Replicate 2004 ERS study using newly available methods: Does the Conservation Reserve Program adversely affect farm dependent rural economies?
- Original findings were convincing but expressed caution regarding unknown error probability
- Estimating the power of the test to derive a statistical verdict

Moneyball for Government: Direct Resources to Effective Programs, Away from Ineffective Programs

- Nonsignificant findings in program evaluation can have normative implications
- Strict adherence to Null Hypothesis Statistical Testing requires "suspending judgment" in the case of a null finding and may contradict a central tenet of Moneyball for Government
- The indeterminacy of null findings can diminish the evidence available to inform decision-making if statistical power is not estimated

Statistical Evidence of Ineffectiveness Requires Statistical Power Estimate

- Statistical Power (the probability of finding something that truly exists) a function of:
  - Sample size
  - Effect size
  - Variability of Treatment Effect

- p[Type II Error] = (1 – Power) is defined for a population, cannot be derived simply from sample

- If economists rarely estimate power quantitatively, is there evidence that all of its determinants are given due consideration in a qualitative assessment of power?

Quantitative Evidence Suggests Power Issues Not Routinely Addressed

- McCloskey and Ziliak (1996, 2004) show that determinants of power rarely discussed in survey of all *American Economic Review* articles by applied economists between 1980 and 2000.
- Ioannidis, et al. (2017) in a meta-analysis show that the median statistical power of 6,700 empirical studies is 18%.
- Katz, Kling, & Liebman (2001) assess Moving to Opportunity Boston Pilot (N = 540) but do not address Pr Type II error despite limited sample

But what if sample size is not the predominant determinant of power?

		d = 0.1 Cohen's Very Small Standardized Effect		
N	Power	10% ↑ Sample Size	10% ↑ Effect Size	10% ↑ Standard Deviation
100	0.114	6.14%	12.28%	-9.65%
300	0.251	7.97%	16.73%	-13.94%
500	**0.384**	**8.07%**	**16.67%**	**-14.84%**
800	0.558	7.17%	14.52%	-13.62%
1600	0.846	3.78%	7.09%	-8.63%
3200	0.988	0.51%	0.81%	-1.92%

Source: Power estimates and changes computed using SAS Proc Power pairedmeans option

Two changes in practice will make assessments of power more reliable

- The easier part: posit relevant effect sizes *prior* to data collection/compilation. The adequacy of N = 50, 200, 1500 or 30000 cannot be assessed without knowing how big a thing you are looking for

- The harder part: develop priors on the variability of treatment from prior studies of the phenomenon. If none, can compute power using Monte Carlo methods

Demonstration: Testing for Unintended Consequences of USDA's Conservation Reserve Program

- Did the CRP, which retired environmentally sensitive land from agricultural production, cause job losses in farm dependent counties?

- ERS difference-in-difference analysis of counties with high percentages of CRP matched with similarly situated low CRP counties

- Empirically challenging due to lack of plausible counterfactuals—differences in treatment and control counties persisted despite matching

Employment Growth Lagged in Treatment Counties, But No Evidence Attributable to CRP

- Average long-term employment growth in control counties exceeded treatment group by 5.8%
- Short-term ('85-'92) effect negative (lower employment in high CRP counties) and significant (at 10% level) in 7 out of 20 specifications
- Long-term ('85-'00) effect positive (higher employment in high CRP counties) and significant (at 10% level) in 3 out of 20 specifications. One negative estimate not significant
- "Some confidence" with a caveat:

  "consistent estimates provide some confidence that the absence of statistical significance can be interpreted as 'CRP has no effect,' even though we do not know the probability of a Type II or false negative error. Since the absence of evidence is not evidence of absence, this approach helps to corroborate the findings from the matched-pair analysis" --Sullivan, et al 2004, ERS AER-834 p. 31

Statistical Power in the 2004 CRP Study

- Sample size 190 matched pairs.
  - Small sample size opens up possibility of a low power test.

- Effect Size:
  - Not specified in original charge, power not defined
  - Large Effect: job losses equivalent to program benefits (-2.7%)
  - Moderate Effect: -1% attributable to CRP
  - Small Effect: -0.1% to -0.5%

- Variability of Treatment Effect
  - Derived from instrumental estimation of sample data

Computing Power (Complement of Probability of Type II Error) of the Test

- Using the study's empirical model:
  - Estimate original model
  - Capture vector of residuals
  - Draw boot strapped samples of covariates and residuals with resampling

- Simulated the distribution of the covariates and residuals according to pre-specified sample sizes (N = 50,100,150, 200, …350)

- Bootstrapping offers the advantage of not taking a stand on the distribution of the covariates
  - Implicitly we are using the joint empirical distribution of the design matrix

- Run the Monte Carlo simulation

Monte Carlo Simulation

- Various CRP effect and sample sizes were evaluated:
  - β = -0.001, -0.005, -0.01, -0.015, and -0.027
  - N = 50, 100, 190, 200, 250, 300, 350

- For each combination of β and N, 1000 iterations were drawn from boot-strapped sample (with replacement) of covariates and OLS residuals
  - Reconstruct the dependent variable, y^*
  - Model re-estimated

- For α = 0.05, count the number of times, *r*, p-value on β is < α

- Power = *r* / iterations

Power of One-Tailed Test for Different Treatment Effect Sizes of Beta

	Sample Size							
Beta	50	100	150	190	200	250	300	350
-0.001	0.07	0.05	0.04	0.04	0.04	0.04	0.04	0.04
-0.005	0.10	0.15	0.22	0.29	0.30	0.39	0.47	0.53
-0.010	0.19	0.47	0.70	0.82	0.84	0.91	0.96	0.98
-0.015	0.33	0.77	0.94	0.98	0.98	1.00	1.00	1.00
-0.027	0.65	0.99	1.00	1.00	1.00	1.00	1.00	1.00

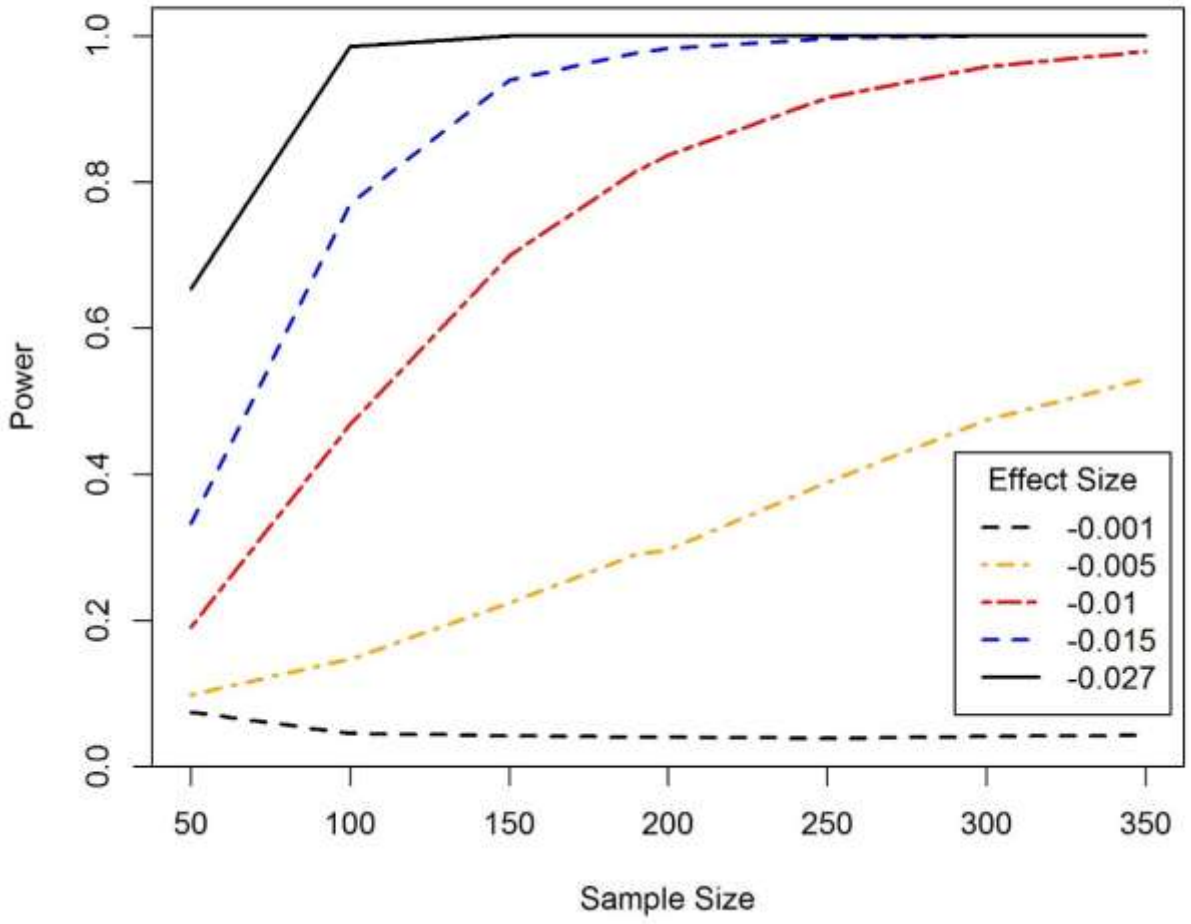
The Paradox of Power

- Putting a relevant constraint on the analysis (e.g., looking for an effect of -1%) can make it much more informative than an unconstrained analysis (e.g., looking for any effect).

- Although an *a priori* effect size is contrary to Fisher's principal concern (positivist scientific exploration), applied economists have an interest in testing for an explicit alternative hypothesis.

- *Ex ante* consideration of "the effect size that matters" only requires prioritizing economic significance over statistical significance

Simulated Power Curves



Conclusion

- 2004 CRP study correctly noted that failure to identify significant negative effects did not provide statistical evidence of no adverse effects.
- Ex post power analysis confirms original analysis had
  - a very high probability of detecting a large effect
  - adequate power for detecting a moderate effect
  - low power for detecting a small effect
- Our statistical procedure to simulate power can be applied to any empirical model as long as:
  - the data generating process can be replicated
  - the effect size of economic significance or policy relevance is stated
- Effectively playing Moneyball for Government will require resolving the indeterminacy of null findings.

Thank you!

Questions? Comments?

Tim Wojan
twojan@ers.usda.gov