Evaluating Treatment Effects and Replicability

- work in progress -

Victor Gonzalez-Jimenez
University of Vienna

Karl H. Schlag
University of Vienna

November 4, 2019

1 Motivation

Replication: some want to emphasize validity of experiments, some worry that results will change if experiment run again (design problems, false positive e.g. due to publication bias)

$\longrightarrow$ recent call for replicating existing experiments

This paper:
- how to statistically design and evaluate replication studies
(methodological study)
- reinvestigate data sets in leading paper Camerer etal (2016)
(operational study)

2 Replication Studies

What is a replication study?
- run identical design (use same tests) with sufficiently many subjects

Q: What does it mean that a result does not replicate when running experiment again and:
A (traditional): failure to find a significant effect
$\longrightarrow$ relies on sample size of replication
A (adjusted): find that treatment effect does not exist or cannot be substantial
$\longrightarrow$ relies on subjective assessment of what substantial means

Input: Choose replication sample size.

Dilemma: How to find sample size if the test does not have power formula, if the test is not even correct for the original study?

3 Sources of Errors in Replication Studies

When designing replication study:
- compute sample size using estimated effect in original study
- for power calculations use asymptotic distributions (of some test)
- desire that estimate of original study is in the CI of replication study
- use some other statistics (without inference) to undermine claims

(no need to introduce new methodology, need to follow statistical methodology correctly)

When interested in original study and not in how original study was analyzed:
- make claims about lack of reproducability when there was no evidence of an effect originally
- why be interested in reproducing false claim

(need to remember: - that samples are finite
- that estimates are only guesses
- difference between existence of an effect and evidence of an effect)

4 Methodological Contributions

Remind community

- of some correct tests and their formulae for power
- to use boundaries of CI to determine size of replication study
- to not directly base inference on power calculations for replication sample size

Classify replication results into three categories
(maintained hypothesis is that magnitudes are allowed to differ)

- both studies show significant effect in same direction (good)
- at least one of the two studies has insignificant effect (inconclusive)
- the two studies show significant effect in opposite directions (bad)

variations:

- impose lower bound on magnitudes
- identify significant changes in magnitude of treatment effect

5 Summary of Camerer etal (2016)

18 studies in AER and QJE

focus here on the 8 studies that involve test of means

5/8 replicate (significant treatment effect in original and replicated)
3/8 do not replicate (have insignificant effects in replicated data set)
(7/8 have significant effect in combined data set)

6 Correct and Incorrect Tests

a test with size 5% is correct if it can be proven that it rejects the null hypothesis in at most 5% of the data sets for a given sample size

WMW correct for $H_0 : F_{X_1} \equiv F_{X_2}$
t test correct for $H_0 : \{EX_1 = EX_2\} \cap \{X_i \sim N(\mu_i, \sigma)\}$

but neither correct for
$H_0 : EX_1 = EX_2, med(X_1) = med(X_2), P(X_2 > X_1) = P(X_1 < X_2)$

correct tests for identifying (signed) treatment effects given independent samples:
- binary valued tests (binomial, McNemar, z test, Boschloo, (Fisher))
- mean tests of Schlag (2008) if variables have known bounds
- median test of Schlag (2015)
- stochastic inequality test of Schlag (2008)

7 Stochastic Inequality Test

"brave to change your hypothesis and gain the power"

X, Y independent rv

$H_0 : P(Y > X) \leq P(Y < X)$ vs $H_1 : P(Y > X) > P(Y < X)$

H_1 in words: "Y tends to be larger than X"

type II error given size 5% equals $1.25 = \frac{1}{1-0.2}$ times the type II error of binomial test with size 1% ($0.01 = 0.2 \cdot 0.05$)

(it is an ordinal test)

CI: find closure of all $d \in \mathbb{R}$ such that $H_0 : P(Y > X + d) \leq P(Y < X + d)$ is not rejected

8 Overview of Studies

Study	Variable	Range	Method	N	Replicates in Camerer etal (2016)?
3	Surplus	[6,23]	WMW test	216	✓
6	Payoff	[0, 1.75]	t test	158	✓
7	Efficiency Ratio	[0,1]	WMW test	54	No
8	Efficiency	[0,1]	WMW test	168	✓
9	Gap WTA-WTP	[0,10]	t test	112	No
11	Median Cooperation rate	[0,1]	WMW test	78	✓
13	Worker Earnings	[0,120]	WMW test	120	No
16	RAD Fundamental Value	[0,1]	WMW test	120	✓

9 Stochastic Inequality: Original Data

Study	Range	Effect Size Original Study	N	WMW	Replicates in Camerer etal (2016)?
3	[6,23]	-5.9**	216	-2.7***	✓
6	[0, 1.75]	0.907***	158	4.7***	✓
7	[0,1]	0.16***	54	2.4***	No
8	[0,1]	0.66***	168	3.9***	✓
9	[0,10]	0.89	112	1.6*	No
11	[0,1]	0.29***	78	15.5***	✓
13	[0,120]	51**	120	2.9***	No
16	[0,1]	-0.34*	120	-2.4***	✓

Note: *** 1% level, ** 5% level, * 10% level.

10 Stochastic Inequality: Both Data Sets

Study	Range	Effect Size Original	N	Effect Size Replication	N	Replicates in Camerer etal (2016)?
3	[6,23]	-5.9**	216	-7.1	312	✓
6	[0, 1.75]	0.907***	158	1.67***	153	✓
7	[0,1]	0.16***	54	-0.026	86	No
8	[0,1]	0.66***	168	0.61***	117	✓
9	[0,10]	0.89	112	0.65	250	No
11	[0,1]	0.29***	78	0.165**	19	✓
13	[0,120]	51**	120	-13	151	No
16	[0,1]	-0.34*	120	-0.11***	219	✓

Note: *** 1% level, ** 5% level, * 10% level.

11 Stochastic Inequality: Confidence Intervals

Study	Range	Effect Size Original	CI Original	Effect Size Replication	CI Replication	Over-lap?	Repl in Camerer?
3	[6,23]	-5.9**	[-8.4,-2.1]	-7.1	[-12.2,-1.8]	✓	✓
6	[0, 1.75]	0.907***	[0.907,0.908]	1.67***	[1.68,1.69]	No	✓
7	[0,1]	0.16***	[0.11,0.27]	-0.026	[-0.2,0.12]	✓	No
8	[0,1]	0.66***	[0.41,0.87]	0.61***	[0.33,0.87]	✓	✓
9	[0,10]	0.89	[-0.4,2]	0.65	[-0.2,1.4]	✓	No
11	[0,1]	0.29***	[0.33,0.52]	0.165**	[0.01,0.26]	No	✓
13	[0,120]	51**	[26,84]	-13	[-40,10]	No	No
16	[0,1]	-0.34*	[-0.72,0.01]	-0.11***	[-0.23,-0.01]	✓	✓

Note: 95% CI, *** 1% level, ** 5% level, * 10% level.

12 Summary of our Analysis

good news for 4 studies:

4 studies have significant effects in both data sets when using a correct test
1 of these 4 has a significantly smaller treatment effect in the replication study
(4/4 replicate according to Camerer etal)

some evidence of a treatment effect for 2 studies:

2 studies have at least one significant effect in the two data sets
(1/2 replicate according to Camerer etal)

not clear what is going on in 1 study:

1 study has insignificant effects in both data sets
(0/1 replicates according to Camerer etal)

there might be concern in 1 study:

1 study has insignificant and significantly smaller treatment effect in the replication study
(0/1 replicates according to Camerer etal)

13 Meta Analysis

Q: how to combine data to understand if there is overall a treatment effect?

Literature: run regression with fixed effects for each study, so $Y_{1i} = \alpha_i + \beta + \varepsilon$

Our approach: randomly draw a study proportional to study size and compare random treated to random non treated where study size defined as minimum of number treated and number not treated

H_0 :
$\frac{n_1}{n_1+n_2} P\left(Y_{11} > Y_{01}\right) + \frac{n_2}{n_1+n_2} P\left(Y_{12} > Y_{02}\right)$
$\leq$
$\frac{n_1}{n_1+n_2} P\left(Y_{11} < Y_{01}\right) + \frac{n_2}{n_1+n_2} P\left(Y_{12} < Y_{02}\right)$

so no common level and magnitudes may differ

14 Conclusion: What Should We Learn?

On the power of correct tests:

- good news: find treatment effect in 7/8 orginal studies despite fact that sample sizes designed for WMW test and t test
- bad news: only 4/8 maintain treatment effect in both studies (but replication sample sizes eradic)

On the evidence uncovered by correct tests:

- bad news: 2/8 have significantly smaller effect in replication studies

On the usefulness of correct tests:

- meta analysis will allow to compute average treatment effect without imposing structure between designs