RV_cursci_example.Rmd

Peder M. Isager

4/6/2021

Contents

1 Preliminary evidence for the validity of equation 4 - the replication value of replicated studies.

Whether equation 4 is a valid estimator of expected utility gain is an empirical question. Validating such a measure of replication value will be challenging, since there does not exist an observable ground-truth measure of expected utility gain to compare equations against. However, validation studies can be conducted if one is willing to make certain assumptions. A preliminary validation study of this kind is provided here.

If the expected utility gain of a study makes it more likely that researchers will select that study for replication, then RVCn estimates and researchers' decisions of which study to replicate should be correlated. Studies that have been replicated in a scientific discipline should have a higher replication value than non replicated studies from the same discipline if replication study authors use expected utility as a selection criterion, since equation 4 is expected to track expected utility. We conducted an exploratory study to test this hypothesis.

1.1 Methods

1.1.1 Sample

We tested the hypothesis that replicated studies receive higher RVCn in the field of psychology. We began by collecting one dataset intended to represent the population of replicated studies in psychology, and one dataset intended to be representative of the general population of empirical studies in psychology.

The sample of replicated studies consisted of original studies listed in the Curate Science replication dataset (https://curatescience.org/app/replications). The Curate Science replication dataset contains information for 1127 replications of 149 original study articles (some articles contained more than one replicated study), primarily from the field of psychology. For these original studies we estimated RV_{Cn} using the sample size available with the replication dataset, and using publication year and citation count information from Crossref (citation counts were extracted from crossref 2020-10-20).

The comparison sample consisted of 9799 articles referenced in the tables of meta-analyses published in Psychological Bulletin between the years 1914 and 2017. Since all articles contained findings that have been referenced in meta-analysis tables in a general-topic psychology journal, we assumed this dataset forms a representative sample of published empirical psychology studies. We also assumed, given the generally low rate of replication in psychology [@Makel2012], that the sample would consist almost entirely of non-replicated original studies.

1.1.2 Statistical analyses

Our (non-preregistered) main hypothesis was that average RV_{Cn} would be higher in the sample of replicated studies than in the comparison sample. In addition, we analysed differences between samples in citation count, citations per year, and sample size, to better understand the causes of any potential differences in RV_{Cn} between the two groups. Because these variables are highly skewed, non-parametric methods were used for all analyses. Median value and interquartile range were calculated for each sample for each variable of interest. To compare differences between samples, we calculated Vargha and Delaney's A for each variable of interest, which represents the probability that a random observation from the sample of replicated studies has a higher value than a random observation from the sample of non-replicated studies [@Vargha2000]. Bootstrapped 99% confidence intervals are reported for each effect size A. All analyses were exploratory, and alpha cannot be adequately controlled. Thus, null-hypothesis significance testing is not used for inference, and the confidence intervals reported should be interpreted cautiously.

The data files and analysis script used to generate all results reported below are openly available on OSF (*link to OSF project*).

1.2 Results

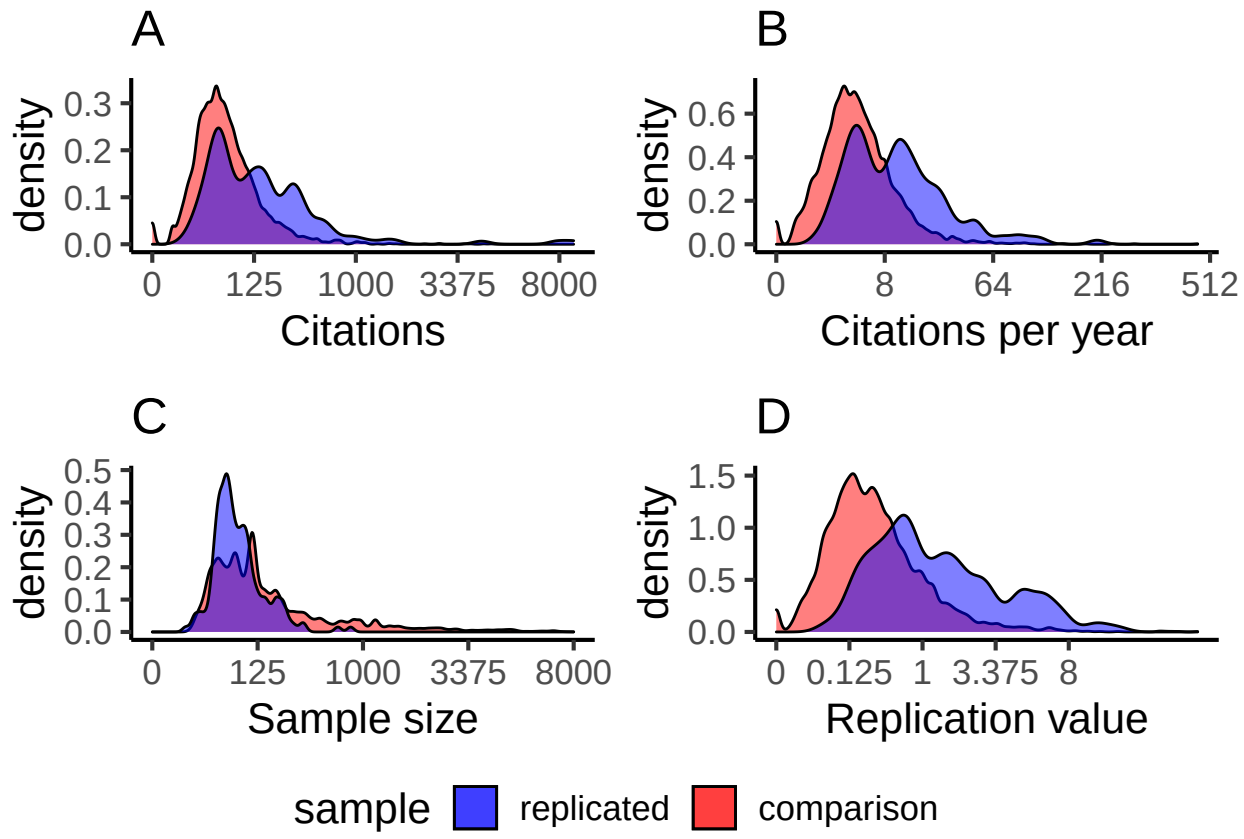

Figure 1: Distributions of various parameters in the comparison sample of psychological findings (red) and a sample of replicated findings in psychology (blue). The scale in all plots have been transformed by taking the cube root of the true values, which preserves the overall shape of the distribution but compresses the scale towards 1. A) Distribution of citation counts. B) Distribution of average citations per year. C) Distribution of sample size. The scale limit is set at 8000, which excludes less than 1% of values. D) Distribution of RV_{Cn} replication value estimates calculated from equation 4.

Statistical results are presented in table 1. Cube-root transformed distributions of the variables of interest are

Table 1: Summary statistics for variables of interest in the replicated (Curate Science) and non-replicated (Psychological Bulletin) samples.

variable	group	n	median	Q1	Q3	A
citation count	replicated	145	112.000	37.000	301.000	0.711 99%CI[0.652, 0.767]
citation count	comparison	4071	39.000	17.000	82.500	
citations per year	replicated	145	10.091	3.545	20.000	0.746 99%CI[0.686, 0.802]
citations per year	comparison	4071	2.882	1.200	6.211	
sample size	replicated	167	57.000	36.000	99.000	0.334 99%CI[0.294, 0.379]
sample size	comparison	4071	113.000	49.000	297.000	
replication value	replicated	167	1.006	0.461	2.700	0.797 99%CI[0.753, 0.841]
replication value	comparison	4071	0.243	0.090	0.580	

presented in figure 1. In summary, RV_{Cn} was, on average, substantially greater in the sample of replicated studies than in the comparison sample (figure 1D). This difference seemed to be driven both by differences in citations per year, and by differences in sample size. Replicated studies received a greater number of citations per year (figure 1B). Conversely, replicated studies had substantially lower sample size, on average (figure 1C).