Emagnification:

A TOOL FOR ESTIMATING EFFECT SIZE MAGNIFICATION AND PERFORMING DESIGN CALCULATIONS IN EPIDEMIOLOGICAL STUDIES

David J. Miller
James T. Nguyen
Matteo Bottai

Working Paper

http://www.imm.ki.se/biostatistics/emagnification/working_paper_2019.pdf

KAROLINSKA INSTITUTE
Unit of Biostatistics
Institute of Environmental Medicine
Stockholm Sweden

August 2019

`emagnification`: A Tool for Estimating Effect Size Magnification and Performing Design Calculations in Epidemiological Studies

Karolinska Institute Working Paper

David J. Miller, James T. Nguyen, and Matteo Bottai

August 2019

ABSTRACT

Artificial effect size magnification (ESM) may occur in underpowered studies, where effects are only reported because they or their associated p-value have passed some threshold. Ioannidis (2008) and Gelman and Carlin (2014) have suggested that the plausibility of findings for a specific study can be evaluated by computation of ESM, which requires statistical simulation. In this paper, we present a new Stata package called -`emagnification`- that allows straightforward implementation of such simulations in Stata. The commands automate these simulations for epidemiological studies and enable the user to assess ESM on a routine basis for published studies using user-selected study-specific inputs that are commonly-reported in published literature. The intention of the package is to allow a wider community to use ESMs as a tool for evaluating the reliability of reported effect sizes and to put an observed statistically significant effect size into a fuller context with respect to potential implications for study conclusions.

Keywords: emagnification, replication, reproducibility, inflation, magnification, p-value, Type M error, effect size magnification, winners curse

David J. Miller
Office of Pesticide Programs
U.S. Environmental Protection Agency
Washington, DC, USA
miller.davidj@epa.gov

James T. Nguyen
Office of Pesticide Programs
U.S. Environmental Protection Agency
Washington, DC, USA
nguyen.james@epa.gov

Matteo Bottai
Unit of Biostatistics
Institute of Environmental Medicine , Karolinska Institutet
Stockholm, Sweden
matteo.bottai@ki.se

This paper is being made available for discussion and comment purposes and represents an earlier version of a manuscript that is currently under review by the *Stata Journal*. The analysis described in this article has been reviewed by the US EPA's Office of Chemical Safety and Pollution Prevention (OCSPP) and approved for release. Approval does not signify that the contents necessarily reflect the views, policies, or determinations of the Agency, nor does the mention of trade names of commercial products constitute endorsement or recommendation for use.

Note: Some material presented here was originally generated by two of the authors who served in various capacities on a European Food Safety Agency (EFSA) Panel on Plant Protection Products and their Residues (PPR) which, in turn, followed up on findings of the External Scientific report 'Literature review on epidemiological studies linking exposure to pesticides and health effects' (University of Ioannina Medical School, 2013) (EFSA-Q-2014-00481). As part of their work on the PPR, the authors contributed to the review and writing of "Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report Literature review of epidemiological studies linking exposure to pesticides and health effects' "and its Annex D where much of this material originally appeared. The PPR Panel report is published in the EFSA Journal (EFSA PPR Panel Report, 2017), an official publication of EFSA. The present work introduces the new Stata command `-emagnification-` and is the result of an expansion and extension of the original EFSA PPR Panel work as part of a post-PPR Panel collaboration by the authors.

Table of Contents

1. Introduction

There is increasing interest and concern in the scientific community in recent years on the "replication crisis" in science. Specifically, scientists are finding that the result from scientific experiments can be difficult to reliably replicate on subsequent investigations. Some have gone so far as to assert and provide support for a contention that most published research findings are false (Ioannidis, 2005). Others have pointed out that even the more modest goal of reproducing previous research – demonstrating that others can calculate using the same data and methods – is frequently difficult or impossible (ASA 2017).

Several ideas have been advanced with respect to the reasons for this increased difficulty in replicating scientific results; these have included "vibrational effects", which develop from the multitude of choices in the way the data are analyzed; increased pressures to publish; publication bias; and the prevalence of and emphasis in research on null-hypothesis-significance-testing. Several researchers, directly or indirectly, have at least partially ascribed the current replication issues in science to a combination of an emphasis on testing of novel hypothesis, a lack of power in the studies that are done, and an over-emphasis on the part of researchers and publishers on p-values and "achieving (statistical) significance". This latter reason arises at least partly from the fact that underpowered studies for discovering statistically significant effect sizes of interest lead to artificially magnified effect size estimates associated with any effect that might be "discovered", and has variously been termed effect size magnification, effect size inflation, truth inflation or, by Button et al., the "winners curse" (Button, 2013).

We will here use the term effect size magnification (ESM) to describe the phenomenon that a reported association may be artificially inflated when the very reporting only occurred because the effect attained a certain size or was statistically significant. This phenomenon is more likely to occur in underpowered studies, where random variation may mistakenly be interpreted as an important scientific discovery. Studies are underpowered when they are small or have comparatively large random variation.

As an example of this ESM concept and why it may come about, it is useful to imagine a thought experiment in which a trial is run thousands of times, each with variable sample sizes instead of what actually happens in practice, a one-time experiment or observation with one fixed sample size. In this thought experiment, there will be a broad distribution of observed effect sizes over the thousands of times the trial is run with varying sample sizes. While the observed medians of these estimated effect sizes are expected to be close to the true effect size regardless of sample size, the trials from smaller size studies from these simulations will necessarily systematically produce a wider variation in observed effect sizes than the larger trials; only a small proportion of the observed effects in these small size studies (i.e. low power) will pass any given statistical threshold of significance – and these will be only the ones with the greatest of effect sizes. Thus, when these low-powered (generally smaller) studies with greater random variation do indeed find a significance-triggered association as a result of passing a given statistical threshold (e.g., $p<0.05$), they are more likely to overestimate the size of that effect. What this means is that research findings of low-powered and statistically significant studies are systematically biased in favor of finding (artificially) inflated effects. Stated mathematically: conditional on a result passing some pre-determined threshold of statistical significance, test level, or magnitude, the estimated effect size is a biased estimate of the true effect size with the magnitude of this bias inversely related to power of the study.

For illustrative purposes and as an introduction to the issue, we draw on a concrete example from the work of John Ioannidis appearing in Table 2 of his article "Why Most Discovered Associations Are Inflated" (Ioannidis, 2008). His table is recreated here as Table 1. Ioannidis generated Table 1 from a series of

simulations designed to illustrate the ESM phenomenon in which a low study power will lead to exaggerated effect sizes for those results that are statistically significant.

Table 1: Simulations for effect Sizes Passing the Threshold of Formal Statistical Significance (p = 0.05) (Excerpted from Table 2 of Ioannidis, 2008, with correction[a])

True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations	
			Median (IQR[b])	***Median Fold Inflation***
1.10	30	1000	1.23 (1.20 – 1.29)	1.11
1.10	30	250	1.51 (1.49 – 1.55)	1.37
1.25	30	1000	1.29 (1.26 – 1.67)	1.03
1.25	30	250	1.60 (1.50 – 1.67)	1.28
1.25	30	50	2.73 (2.60 – 3.16)	2.18

[a] See explanatory note at footnote 2 in text
[b] IQR indicates interquartile range.

As can be seen in the first data row of Table 1, Ioannidis begins by assuming a true odds ratio for an association of 1.10 and that the proportion of exposed individuals in the control (or non-diseased) group is 30%. It follows then, mathematically, that the expected proportion of exposed individuals in the case group would be 0.3204.[1] Ioannidis then simulates a set of epidemiological studies in which (i) the control group in each simulated study includes 1000 subjects and the number of exposed subjects within the control group is randomly drawn from a binomial distribution with probability 0.3000 representing the control group proportion; and (ii) the case group in each simulated study includes 1000 subjects and the number of exposed subjects within the case group is randomly drawn from a binomial distribution with probability of 0.3204, representing the case group proportion. The observed odds ratio of each of many simulated studies in which "n" samples are drawn per group is then computed and stored. The median odds ratio of these simulated studies is expected to be equal to the true odds ratio value of 1.10, but we would expect that only a proportion of those observed odds ratios that happened to have large values would be statistically significant ($p < 0.05$). This is what is illustrated in Table 1, focusing on and highlighting the simulation results of the odds ratios *that happened to be found significant at p<0.05*. When looking at only those odds ratios that pass this $p<0.05$ statistical threshold, the medians among this subset of (statistically significant) odds ratios is observed to be 1.23, shown in the first row of data in Table 1 which is higher than the true odds ratio of 1.1 used to generate the simulation. In fact, Table 1 shows that a considerable fraction (>75%) of the simulated significant odds ratios are inflated compared to the true odds ratio of 1.10 since the interquartile range (IQR) of the medians were found in Ioannidis's simulations to be (1.20-1.29).[2] This phenomenon illustrates (via computer simulation) that when a researcher's or data user's focus is on statistically significant results, such statistically significant results will be systematically biased high (magnified or inflated) for underpowered studies: in this example in which the power can be calculated to be 27% (not shown) and where the actual or true odds ratio is 1.10, the median statistically significant odds ratio is estimated as 1.23, representing a systematic inflation of 11%, this. As the sample size gets smaller (say, from 1000 in each of the comparison groups to 250 as shown in the second line of data in Table 1), the magnification becomes greater, with a median odds ratio of 1.51 (IQR: 1.49-1.55) corresponding to a median inflation of 37%. As shown in the last row of Table 1, Ioannidis uses a still smaller sample size of 50 and increases the true odds ratio from 1.10 to 1.25 (producing only 15% power):

[1] P1 = (P0 x OR) / [(1 – P0) + (P0 x OR)] where: P1 = Expected proportion of exposed individuals among cases; P0 = Expected background or control group proportion; and OR = True odds ratio between exposed and control individuals
[2] Simulation with `-emagnification-` done by the authors suggests the 1.23 listed as p25 in Table 2 of Ioannidis (2008) is a typographical error, and the actual value for the p25 should be closer to 1.20.

in doing this, he produces a median odds ratio for statistically significant results of 2.73 representing a magnification factor of 2.18 -- more than double (118% inflation) the true odds ratio of 1.25 for those results that pass the p<0.05 significance threshold!

While Table 1 above is useful to illustrate the ESM phenomenon, demonstrate how it arises, and quantify it, the tabulated conditions and numbers in Table 1 are necessarily fixed and it would be useful to be able to generate such a table "on the fly" with inputs that are specific to a researcher's particular study of interest. This Stata Journal article introduces new Stata commands `-emagnification proportion-` and `-emagnification rate-` (hereinafter referred to generically as `-emagnification-`) to easily implement and examine the above described ESM phenomenon using simulations. Stata's `-emagnification-` command automates these simulations and enables the user without too much effort to compute the magnification factors numerically on a routine basis in specific settings or for specific studies. It is an outgrowth of work done by two of the authors as part of a European Food Safety Agency (EFSA) Panel on Plant Protection Products and their Residues (PPR) and is the result of an expansion and extension of the original EFSA PPR Panel work during post-PPR Panel collaboration by the authors.[3]

The remainder of this paper introduces this new Stata command, which facilitates performing ESM simulations. The general idea is that if researchers are interested in effect size estimates only if they cross some boundary of significance or magnitude, then the reported estimates are biased away from the null if they derive from low powered studies. The magnitude of bias can be expressed as the median of reported estimates (estimates exceeding a selected boundary such as p<0.05) relative to the true value, and the new Stata package implements simulations to quantify this for odds ratios and rate (or risk) ratios. In addition to the median of reported estimates, other percentiles (e.g., 25th and 75th , or 10th and 90th) can considered. Such Stata simulations can be useful when evaluating reported effect sizes in a published epidemiological paper and the new package makes this a simple numerical exercise.

2. The `emagnification` Command

2.1 Syntax

Effect Magnification for Proportions

```
emagnification proportion, p0(numlist) or(numlist) n0(numlist) n1(numlist)
   [other_options]
```

Effect Magnification for Rates

```
emagnification rate, r0(numlist) rr(numlist) n0(numlist) n1(numlist)
   [other_options]
```

The words 'proportion' and 'rate' after 'emagnification' can be abbreviated to any number of characters. For example, all the following three lines are allowed:

[3] As part of their work on the PPR, two of the authors (MB and DJM) contributed to the review and writing of "Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects' " (EFSA, 2017) and its Annex D where much of this material originally appeared. The effect size magnification calculations/simulations appearing in Annex D of that EFSA report were generated using custom-coded SAS program by one of authors (JTN) of this *Stata Journal* article. The results for the current Stata `-emagnification-` command introduced here were tested against this earlier SAS script and compared favorably.

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100)
. emagnification prop, p0(.5) or(2) n0(100) n1(100)
. emagnification p, p0(.5) or(2) n0(100) n1(100)
```

2.2 Description

The -emagnification- command estimates effect magnification of proportions and rates through simulations.

The 'emagnification proportion' syntax estimates odds ratios with the 'logit' command.

The 'emagnification rate' syntax estimates relative risks with the 'poisson' command.

Iterations that do not converge (for example zero events are generated which may happen with small counts) are dropped, with the number of valid (completed) iterations shown at the end of the Stata run in the results table in the column labeled 'valid'. If all iterations are completed with valid results and no runs are dropped, this will equal the number of iterations requested by the user. The program will continue running even with invalid (and dropped) results.

2.3 Options

p0(numlist) specifies the proportions in the reference group when used with the -emagnification proportion- syntax. This is sometimes estimated in case-control studies using odds ratios as the number of exposed subjects in the reference (control) group divided by the number of subjects in the reference group

or(numlist) specifies the odds ratios of the case (comparison) group versus the reference group

r0(numlist) specifies the rates in the reference group when used with the -emagnification rate- syntax. This is sometimes estimated in cohort studies using rate ratios (or relative risks) as the number of diseased individuals in the reference (unexposed) group divided by the number of subjects in the reference group.

rr(numlist) specifies the risk ratios of the exposure (comparison) group versus the reference group

n0(numlist) specifies the sample size in the reference group

n1(numlist) specifies the sample size in the comparison group

pctile(numlist) specifies the percentiles of the distribution of significant effects sizes specified in numlist where what is significant is defined in the level( ) option; defaults to 10 50 90.

`ifactor(numlist)` specifies the inflation factors of the percentiles specified in numlist where the inflation (exaggeration) factor is equal to the relevant percentile divided by the true odd ratio. The set of percentiles specified in the 'ifactor' option may be different from that specified in the 'pctile' option. For example, both the following two lines are allowed:

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100) pctile(50)
   ifactor(50)

. emagnification proportion, p0(.5) or(2) n0(100) n1(100) pctile(50)
   ifactor(90)
```

`nsim(#)` specifies the number of simulated datasets; defaults to 10.

`level(#)` specifies the significance level of the test; defaults to 0.05.

`onesided` specifies a one-sided test; defaults to two-sided.

`exact` for proportion specifies the specifies the Fisher's exact test instead of the default chi-square test; for rates, specifies the exact Poisson regression instead of the default Poisson regression.

`seed(string)`specifies the seed for the pseudo-random number generator. The 'emagnification' command is based on simulated pseudorandom data. Therefore, the same command line can produce non-identical results, when run multiple times. If the 'seed' option is specified, the sequence of pseudorandom number generator starts at the specified seed, and the same command line produces identical results every time it is run. For example, the following two lines may not produce identical estimates:

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100)

. emagnification proportion, p0(.5) or(2) n0(100) n1(100)
```

Conversely, the following two lines produce identical estimates:

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100) seed(123)

. emagnification proportion, p0(.5) or(2) n0(100) n1(100) seed(123)
```

The pseudorandom-number generator seed is saved in the `r(seed)' macro, whether or not the 'seed' option is specified in the 'emagnification' command. The 'r(seed)'-saved macro can be used to replicate the results of the latest simulation. For example, the following two lines, run consecutively, produce identical estimates:

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100)

. emagnification proportion, p0(.5) or(2) n0(100) n1(100) seed(`=r(seed)')
```

`log` shows the simulation iterations. This is convenient to track Stata's progress on long runs with many iterations.

`clean` shows the results without separator lines.

2.4 Stored Results

`emagnification` stores the following in r():

Scalars
```
r(level)   the level of the tests
```

Macros
```
r(cmdline) command as typed
r(seed)    the seed used by the pseudo-random numbers generator
```

Matrices
```
r(table)   the table of the results
```

2.5 Example Inputs

The following examples illustrate the `-emagnification-` command

Estimate the effect magnification for a proportion

```
. emagnification proportion, p0(.5) or(2) n0(100) n1(100)
```

Estimate the effect magnification for a rate

```
. emagnification rate, r0(.5) rr(2) n0(100) n1(100)
```

Estimate the effect magnification for a proportion in multiple scenarios using 0.1 as the level of significance and showing the inflation factor for only the median statistically significant result

```
. emagnification proportion, p0(.5 .9) or(1.5 2) n0(100 200) n1(100 200)
.pctile(50 90) ifactor(50) nsim(100) level(.1) onesided seed(123) log
clean
```

Show the saved results of the latest estimation:

```
. return list
. matrix list r(table)
```

3. Applications

This section provides further details of the Ioannidis example described earlier in the Introduction and shows how to conduct this analysis with the new Stata command. It then goes on to present two examples using data taken from the epidemiological literature. The first epidemiological case example in this section is from a case control study using odds ratios published by Greenland et al. (1994) and dealing with resin exposures and lung cancer in a group of workers involved in the assembly of transformers. The second epidemiological case example uses rate ratios (aka relative risks) and introduces a publication from the U.S. Agricultural Health Study, a cohort study begun in the mid-1990's

which follows 90,000 pesticide applicators and their wives living in IA and NC which investigates diazinon exposure among applicators of this pesticide and lung cancer. The `-emagnification-` command is able to deal with both odds ratios (generally from case-control studies for which `-emagnification proportion-` is used) and rate (or risk) ratios (generally from cohort studies for which `-emagnification rate-` is used) and the two epidemiological examples illustrate each of these, respectively.

3.1 Ioannidis Effect Size Magnification Example Simulation

Before beginning with the two epidemiological examples in order to illustrate the utility of the `-emagnification-` command with case studies, it is useful to revisit the Ioannidis example illustrated in Table 1 and to use `-emagnification-` to replicate the simulation results the table. To do this, the following four input values are required:

1. the number of subjects in reference group;
2. the number of subjects in the comparison group;
3. the specific proportion or rate of interest in the reference group. Here, this is the proportion of exposed subjects in the control group since the Ioannidis example uses odds ratios; and
4. the assumed (true) odds ratios (here) or rate ratios of interest.

In the `-emagnification proportion-` syntax for odds ratios, we have the following inputs (all derived from Table 1):

- `n0(numlist)` is the number of subjects in reference group, here 1000 control subjects as per the "Sample n Per Group" column from Table 1;
- `n1(numlist)` is the number of subjects in comparison group, here 1000 case subjects as per the "Sample n Per Group" column from Table 1;
- `p0(numlist)` is the proportion of interest in the reference group; here in the Ioannidis example, this is the proportion of exposed subjects in the control group = 0.30 as listed under the "Control Group Rate (%)" column of Table 1;
- `or(numlist)` is the assumed true odds ratio(s), here 1.10 or 1.25, as per the "True OR" column in Table 1.

We insert these values into the `-emagnification proportion-` command with several additional options described in Part 3 to simulate the results from the first row of the Table 1 from Ioannidis:

```
emagnification proportion, p0(0.30) or(1.1) n0(1000) n1(1000) pctile(25 50 75)
ifactor(50) nsim(1000) level(0.05) onesided seed(123)
```

This generates the following Stata output table:

p0	p1	true_or	n0	n1	valid	power	p25	p50	p75	if_p50
.3	.3203883	1.1	1000	1000	1000	.274	1.202	1.235	1.289	1.123

As can be seen, the values for the p50 (i.e., median) of 1.235 and IQR of (1.202-1.289) approximate well those given in the simulation performed by Ioannidis in the first row of Table 1.

Using similar syntax except taking advantage of the ability of the `-emagnification-` command to use Stata `numlists`, we can replicate both the second and fourth rows of Table 1 with the following single Stata command, adding the log option to monitor Stata's progress in real time:

```
emagnification proportion, p0(0.30) or(1.10 1.25) n0(250) n1(250)
pctile(25 50 75) ifactor(50) nsim(1000) level(0.05) onesided seed(123)
log
```

This generates the following output since the log option was used:

```
Scenario 1: p0 = .3, or = 1.1, n0 = 250, n1 = 250
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 2: p0 = .3, or = 1.25, n0 = 250, n1 = 250
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%


The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p25	p50	p75	if_p50
.3	.3203883	1.1	250	250	1000	.114	1.423	1.481	1.563	1.346
.3	.3488372	1.25	250	250	1000	.298	1.440	1.519	1.623	1.215

Similarly, the values estimated here by Stata (medians of 1.481 and 1.519 and respective IQRs of (1.423-1.563) and (1.440-1.623) approximate reasonably well those provided by Ioannidis and appearing in Table 1. The remaining two simulations from Ioannidis (corresponding to the third and fifth rows in Table 1) can be recreated using the following two Stata commands:

```
emagnification proportion, p0(0.30) or(1.25) n0(1000) n1(1000) pctile(25
50 75) ifactor(50) nsim(1000) level(0.05) onesided seed(123) log

emagnification proportion, p0(0.30) or(1.25) n0(50) n1(50) pctile(25 50
75) ifactor(50) nsim(1000) level(0.05) onesided seed(123) log
```

[*results not shown*]

Similarly, these latter two Stata commands re-produce the simulation values generated by Ioannidis. Importantly, these illustrate – as he did – that the more underpowered (and generally smaller) a study is and the smaller the true effect size that the study is investigating, the greater the degree that observed effect sizes that pass some pre-established statistical threshold or are by other means "discovered" will be inflated. Here, we see the median inflations vary from 18% with a small true odds ratio and a large sample to a near doubling of the true odds ratio a smaller sample size of only 50, even with a more substantive odds ratio of 1.25.

The following two sections illustrate the use of the `-emagnification-` command with two epidemiological studies which appeared in the literature and are more typical of the situations that users of this command may encounter.

3.2 Odds Ratios: Greenland et al. (1994) Resin Worker Example

While the above example shows the `-emagnification-` command duplicating the Ioannidis table, these would not be typical uses of the command for the researcher who is faced with evaluating potential ESM in (already) published studies. The first epidemiological case example in this section is from a study published by Greenland et al. (1994). Greenland studied the exposure rates among lung cancer deaths and controls from occupational exposure to resins in a facility that assembled transformers. It is used to illustrate the `-emagnification proportion-` command and is relevant to case-control studies using odds ratios. The example here -- for ease of exposition -- focuses on the crude (as opposed to adjusted) estimates provided in that publication.[4] There are 45 exposed cases, 94 unexposed cases, 257 exposed controls, and 945 unexposed controls.

First, using Stata's `-cci-` command to estimate totals and odds ratios, we see from the Stata output below that there is a statistically-significant positive association between exposure to resins and lung-cancer deaths (OR = 1.76; 95% CI: (1.20, 2.58)).

```
. cci 45 94 257 945, woolf

                                                    Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |        45          94  |        139       0.3237
       Controls |       257         945  |       1202       0.2138
----------------+------------------------+------------------------
          Total |       302        1039  |       1341       0.2252
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.760286       |    1.202457    2.576898 (Woolf)
 Attr. frac. ex. |         .4319106       |    .1683693    .6119365 (Woolf)
 Attr. frac. pop |         .1398272       |
                +-------------------------------------------------
                             chi2(1) =     8.63  Pr>chi2 = 0.0033
```

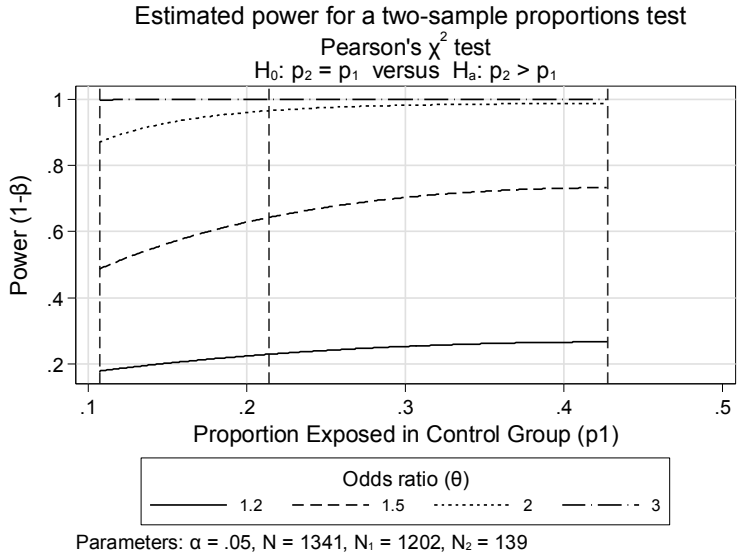
Before using the Stata `-emagnification-` command, it us useful to first use the graphing capabilities of Stata's built-in `-power twoproportions-` command along with the relevant estimated values provided in the above 2x2 table to estimate the power to detect various (assumed) true odds ratios that might be of interest (here, 1.2, 1.5, 2.0, and 3.0) at 0.5x the exposed proportion in control group, 1x the exposed proportion in control group, and 2x the exposed proportion in control group. Supplementing the user-written `-emagnification-` command with Stata's built-in `-power twoproportions-` command provides the user a broader and important view of power-related issues and can assist when evaluating a specific epidemiological study and the degree to which the power of that study might be sensitive to the assumed reference proportion (here for the odds ratio, the exposed proportion in the control group). Power is dependent in part on the true effect size as well as this reference proportion and selecting several

[4] The data is also provided in Rothman *et al.*'s *Modern Epidemiology*. See Table 19-1 (p. 349) in the third edition.

reference proportions in the command can be used as a sensitivity analysis. This sensitivity of the power to the proportion exposed in the control group is shown in Figure 1 and was generated using the following Stata command:[5]

```
. power twoproportions (`=0.5* 257/1202'(0.001) `=2 * 257/1202'),
test(chi2) oratio(1.2 1.5 2.0 3.0) n1(1202) n2(139)graph(recast(line)
xline(`=0.5* 257/1202' `= 257/1202'
`=2*257/1202',lpattern(dash))legend(rows(1)size(small) position(6))
ylabel(0.2(0.2)1.0) xtitle("Proportion Exposed in Control Group
(p1)")scheme(s1manual)) onesided
```

Figure 1. Graph showing estimated power for a (one-sided) two-sample proportions test evaluating power as a function of exposed proportion in control-group at true odds ratios of 1.2, 1.5, 2.0, and 3.0.



Note: Dashed vertical lines represent control group proportions at 0.5x of the observed exposed proportion in control group, 1x of that observed exposed proportion in control group, (i.e., 257/1202), and 2x that of the observed exposed proportion in control group and are used to illustrate the sensitivity of the estimated power to these exposed proportions in control group.

In the above command, the reference group count (here, controls) is 1202, the comparison group count (here, cases) is 139, and the observed exposure proportion among the control reference group ("Proportion Exposed in Control Group (p1)") is 257/1202, or 0.2138, all of which are taken from the Stata `-cci-` output shown above. The center vertical dashed line in Figure 1 represents this proportion. At this (baseline) proportion, the power to detect a true odds ratio of 1.2 is about 25%, the power to detect a true odds ratio of 1.5 is about 65%, and the power to detect true odds ratios of 2.0 and 3.0 are 96% and ca. 100%, respectively. The power to detect these odds ratios would differ at half the baseline proportion (leftmost vertical dashed line) and double the baseline proportion (rightmost vertical dashed line) and can also be read from the graph if such a sensitivity analysis was desired. Figure 1 suggests low power of this study

[5] The Stata command power twoproportions (`=0.5* 257/1202' `=257/1202' `=2 * 257/1202'), test(chi2) oratio(1.2 1.5 2.0 3.0) n1(1202) n2(139)onesided can be used to directly calculate power in a tabulated format without the graph. These results agree with those produced for the -emagnification- output.

to detect true odds ratios as low as 1.2 or 1.5 (25% power and 65% power, respectively) and, thus, that the study might be subject to ESM if the true odds ratios were of this size.

The `-emagnification-` command introduced here allows us to take the above `-power twoproportions-` analysis one step further and to quantitatively estimate the ESM that might be present; such an analysis will permit the user to evaluate whether the observed statistically significant "discovered" odds ratio of 1.76 is potentially consistent with a true odds ratio of 1.2 given the power/sample size of the study. As described earlier in Part 2, the `-emagnification-` command does this by repeatedly drawing values for the 2x2 epidemiological table that are consistent with (i) the given row marginals (here, total number of cases and total number of controls); (ii) any assumed (or various assumed if *numlist* is used) true odds ratios; and (iii) binomial draws from the population of cases and controls with a given control group proportion (here `257/1202`). The output includes various user-selected quantiles of the distribution of the odds ratio determined by the simulation to have crossed a user-selected statistical threshold (indicated by `level()`) and (optionally) the degree of inflation which these odds ratios represent.

Using the `-emagnification-` command for the Greenland (1994) data, we have[6]:

```
emagnification proportion, p0(`=257/1202') or(1.2 1.5 2.0 3.0) n0(1202)
n1(139) pctile(10 50 90) ifactor(50) nsim(1000) level(0.05) onesided
seed(123) log
```

Since the log option is selected here to show real-time Stata progress, Stata generates the following output:

```
Scenario 1: p0 = .21381032, or = 1.2, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 2: p0 = .21381032, or = 1.5, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 3: p0 = .21381032, or = 2, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 4: p0 = .21381032, or = 3, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p10	p50	p90	if_p50
.2138103	.2460507	1.2	1202	139	1000	.243	1.422	1.531	1.786	1.276
.2138103	.2897407	1.5	1202	139	1000	.653	1.453	1.647	2.012	1.098
.2138103	.3522961	2	1202	139	1000	.967	1.599	1.999	2.549	0.999
.2138103	.4493007	3	1202	139	1000	1	2.349	2.989	3.780	0.996

As can be seen, if the true odds ratio for the lung cancer-resin relationship were 1.2 and the non-diseased reference group had a true rate of lung cancer p0 equal to the observed rate of 0.21381 (=257/1202) from the study, we would expect the observed median of statistically significant odds ratios passing a p<0.05

[6] The command here generates effect magnifications tables for only one p0 value. If a sensitivity analysis around p0 is desired at 0.5x, 1x, and 2x the control proportion rate, the numlist option for p0 can be used to generate this larger set of scenarios: `emagnification proportion, p0(`=0.5* 257/1202' `=257/1202' `=2* 257/1202') or(1.2 1.5 2.0 3.0) n0(1202) n1(139) pctile(10 50 90) ifactor(50) nsim(1000) level(0.05) onesided seed(123) log`

threshold to be 1.53; this represents a median inflation of 28% (represented by `if_p50` in the above table) over the true odds ratio of 1.2 used in the simulation; by definition, half of the expected observed statistically significant odds ratios would be above this median value of 1.53 and half would be below. From the output above, we can also see that if the true odds ratios were alternatively 1.5, 2.0, or 3.0, half of the observed statistically significant odds ratios would exceed 1.65, 2.00, and 3.00, respectively, reflecting inflation rates of 10%, 0%, and 0%. Note that the estimated inflation amounts of 0% and 0% in the above output correspond to powers of 97% (to detect a true odds ratio of 2.0) and essentially 100% (to detect a true odds ratio of 3.0), while the (substantially larger) respective inflation factors of 28% and 10% for odds ratios of 1.2 and 1.5 correspond to lower powers of 24% (for the odds ratio of 1.2) and 65% (for the odds ratio of 1.5). This is as expected: all else being equal, statistically significant effect sizes arising from low powered studies attempting to detect a small true effect size show larger magnifications due to ESM than higher powered studies attempting to detect larger true effect sizes. The analysis presented in the table parallels that done by Ioannidis (2008) which was described earlier.

With respect to the study-observed resin worker odds ratio of 1.76 (95% CI: 1.20, 2.58) we saw generated by `cci 45 94 257 945, woolf`, the above Stata output suggests that if the true odds ratio in the transformer resin study were 1.2 and the study were to be repeated many times, we would expect to find most of the statistically significant results (the middle 80% of the distribution, in this case, since we used `pctile(10 50 90)`) to vary between 1.422 (p10) and 1.786 p(90), with a median odds ratio of 1.53 (p50). It follows from this that -- given the size of the study and other characteristic factors -- an odds ratio of 1.76 is (just) within the (middle 80%) bounds that would be expected to occur after considering effect size magnification if the true odds ratio was 1.2 and the observed odds ratio of 1.76 was "selected" for attention because it was significant. Thus, the observed odds ratio of 1.76 is reasonably consistent or compatible with a true odds ratio of as low as 1.2 if we define reasonable consistency as being within the middle 80% of the distribution (i.e., the p10 to p90 range illustrated above) that would be expected for the set of "discovered" statistically significant results. Said another way: if it were determined that an odds ratio of as low as 1.2 were of substantive interest, the power of this study was inadequate to distinguish the observed "discovered" odds ratio of 1.76 from a true odds ratio of 1.2:

Note as an illustrative hypothetical that had the sample size for this study been 10 times greater (and each cell in the 2 x 2 table been thus increased proportionately 10-fold), the (still) observed odds ratio of 1.76 would then be found to be incompatible with a true odds ratio of 1.2 or 1.5 (whose p90 values for statistically significant effects are 1.31 and 1.62, respectively) as shown below:

```
emagnification proportion, p0(`=2570/12020' ) or(1.2 1.5 2.0 3.0)
n0(12020) n1(1390) pctile(10 50 90) nsim(1000) level(0.05) onesided
seed(123) log
```

p0	p1	true_or	n0	n1	valid	power	p10	p50	p90
.2138103	.2460507	1.2	12020	1390	1000	.849	1.140	1.210	1.311
.2138103	.2897407	1.5	12020	1390	1000	1	1.380	1.495	1.623
.2138103	.3522961	2	12020	1390	1000	1	1.851	1.995	2.152
.2138103	.4493007	3	12020	1390	1000	1	2.792	3.004	3.241

One would conclude from this latter (10 times larger) example that if the true odds ratio were 1.2, effect size magnification would **not** account for an odds ratio as high as the observed 1.76; this contrasts (due

to its ten times larger sample size) with the conclusion reached for the earlier (and actual) "ten times smaller study" Greenland et al. (1994) resin study described above. Said differently: if it were determined that an odds ratio of as low as 1.2 were of substantive interest, the power of this larger (hypothetical) study would have been sufficient to conclude that the observed odds ratio of 1.76 would not likely be a result of ESM. Such a change in characterization of the observed odds ratio is not unexpected: increasing the sample size (here, 10-fold) will increase the power to detect a given effect size and this in turn will result in a smaller degree of ESM as demonstrated by Ioannidis's simulation exercises. Of course, this could likely have been determined based on the power of 85% to detect an odds ratio of 1.2 as shown in the above table which exceeds the 80% value for power that typically results in no or minimal ESM. The -emagnification- command, however, allows one to better characterize and quantitate potential ESM by showing a range of statistically significant odds ratios that are compatible with the one that was observed (or "discovered") in the study.

3.3 Rate Ratios: Agricultural Health Study Example

In the final case example, we consider ESM and *rate ratios* (or, equivalently, *relative risks)* as opposed to odds ratios as in the previous resin worker study. Specifically, we look at a case study of lung cancer and its putative association with diazinon exposure in pesticide applicators from the U.S. Agricultural Health Study (Jones *et al*., 2015). Here, a statistically significant rate ratio of 1.60 (95% CI: 1.11 to 2.31) with respect to lifetime days of use was reported by the study authors when comparing the top tertile of exposure to the no exposure reference category. The study authors concluded that this provided additional evidence of an association [of diazinon] with lung cancer risk. The number of subjects at each exposure level was provided in the publication (non-exposed group: N = 17710, and T(ertile)1, T2, and T3 were categorized based on the exposure distribution), and we obtain the required information from the publication to perform an ESM calculation; specifically: (i) the number of subjects in the reference non-exposed group = 17710; (ii) the number of subjects in each of the exposed groups (tertiles) = 1710[7]; and (iii) the number of diseased individuals (lung cancer) in the reference non-exposed group = 199 (from Table 3 of the cited publication).

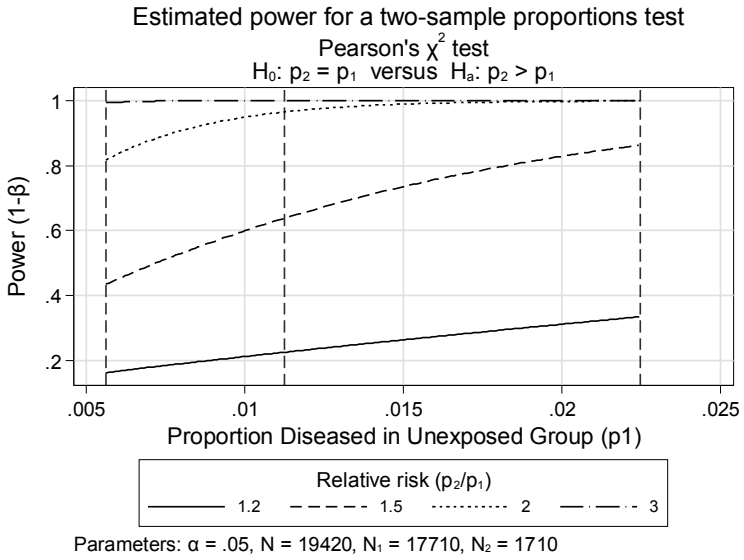
As with the Greenland et al. (1994) publication, we are interested prior to conducting an ESM calculation in evaluating the power of the study associated with the estimated background rate of 199/17710 (=0.011237) for detecting exemplar rate ratios of 1.2, 1.5, 2.0, and 3.0 among the subjects in each tertile of the diazinon exposed individuals. As a form of sensitivity analysis similar to that done in the Greenland et al. example, we are also interested in doing this power calculation assuming the true rate is one half of this background rate (or 0.005617), and twice this background rate (0.022473). As before, this analysis was performed using Stata's -power twoproportions- command and is shown below in Figure 2 for true rate ratios of 1.2, 1.5, 2.0, and 3.0 for 0.5x-, 1x- and 2x- the (observed) background rate of 199 diseased individuals/17,710 persons:

```
power twoproportions (`=0.5* 199/17710'(0.0001) `=2 * 199/17710'),
test(chi2) rrisk(1.2 1.5 2.0 3.0) n1(17710) n2(1710)graph(recast(line)
xline(`=0.5* 199/17710' `=199/17710' `=2 * 199/17710',lpattern(dash))
legend(rows(1)size(small) position(6))
```

[7] Specifically: N of each tertile= (2350+2770)/3=1710 from the publication's Table 1 where: the value 2350 represents the number in the lowest exposed level and the value of 2770 represents the number of the two highest exposed levels when the exposed subjects were dichotomously categorized.

```
ylabel(0.2(0.2)1.0)xtitle("Proportion Diseased in Unexposed Group
(p1)") scheme (s1manual)) onesided
```

Figure 2. Graph showing estimated power for a (one-sided) two-sample proportions test evaluating power as a function of unexposed-group proportion at true rate ratios of 1.2, 1.5, 2.0, and 3.0.



Note: Dashed vertical lines represent control group proportions at 1x of that observed (here, 199/17710) as well as 1/2x of that observed and 2x of that observed and are used to illustrate the sensitivity of the power to these background rate assumptions.

As can be seen in Figure 2, this study had a power of about 23% at 1x the background rate (i.e., the unexposed group proportion, equal to 199 diseased individuals/17,710 subjects = 0.011237) to detect a rate ratio of 1.2. To detect a rate ratio of 1.5, there is about 64% power. Power is greater than 80% to detect rate ratios of 2.0 and 3.0 at a true rate ratio of 1.2. Sensitivity analyses show that at 0.5x the observed background rate, the power would be about 14% and 48% to detect rate ratios of 1.2 and 1.5, respectively, and greater than 80% to detect rate ratios of 2.0 and 3.0. Alternatively if the true background rate was actually twice the observed background rate, we would have about 86% power to be able to detect a rate ratio of 1.5 and essentially 100% power to detect a rate ratio of 2.0.

Given the above and the fact that this is an analysis looking at a rate ratio, `-emagnification rate-` was used to estimate the ESM given (assumed) true rate ratios of 1.2, 1.5, 2.0, and 3.0. Here, we also request Stata to perform these analysis at the observed background rate of `199/17710`, but also at 0.5x and 2x the observed proportion of disease among the unexposed:

```
emagnification rate, r0(`=0.5*199/17710' `=199/17710' `=2*199/17710')
rr(1.2 1.5 2 3) n0(17710) n1(1710) pctile(10 50 90) ifactor(50) nsim(1000)
level(.05) seed(123) onesided
```

r0	r1	true_rr	n0	n1	valid	power	p10	p50	p90	if_p50
.0056183	.006742	1.2	17710	1710	1000	.16	1.608	1.772	2.105	1.476
.0056183	.0084274	1.5	17710	1710	1000	.41	1.630	1.846	2.390	1.231
.0056183	.0112366	2	17710	1710	1000	.788	1.668	2.114	2.719	1.057
.0056183	.0168549	3	17710	1710	1000	.992	2.255	2.959	3.884	0.986
.0112366	.0134839	1.2	17710	1710	1000	.239	1.419	1.518	1.791	1.265
.0112366	.0168549	1.5	17710	1710	1000	.63	1.445	1.641	1.993	1.094
.0112366	.0224732	2	17710	1710	1000	.967	1.598	1.997	2.466	0.999
.0112366	.0337098	3	17710	1710	1000	1	2.488	3.012	3.672	1.004
.0224732	.0269678	1.2	17710	1710	1000	.311	1.298	1.395	1.552	1.163
.0224732	.0337098	1.5	17710	1710	1000	.839	1.345	1.525	1.789	1.016
.0224732	.0449464	2	17710	1710	1000	.999	1.709	2.007	2.325	1.004
.0224732	.0674195	3	17710	1710	1000	1	2.596	2.999	3.412	1.000

As before, the analysis presented in the above table parallels that done by Ioannidis (2008) to illustrate the ESM concept, except that this uses various (assumed) rate ratios rather than the odds ratios used by Ioannidis in his example and thus uses the -`emagnification rate`- version of the Stata command. From the Stata output above, for example, when r0 = 0.0112366 (= 199/17710), a true rate ratio of 1.2 would result in instead – were the study to be repeated numerous times – statistically significant rate ratios that would vary from 1.42 (at the 10th percentile) to 1.79 (at the 90th percentile). The median expected value of the rate ratio is equal to 1.52, higher by 26% than the true rate ratio of 1.2 used in the simulation, as shown in the above output under if_p50; as before, half of the expected statistically significant rate ratios would be above this median of 1.52 and half would be below.

With respect to the rate ratio of 1.6 (95% CI: 1.11, 2.31) that was observed in the Jones et al. (2015) study, the above table suggests that such a rate ratio is not inconsistent with a true rate ratio of 1.2 because the p90 for "discovered" statistically significant results when the true rate ratio is 1.2 is 1.79. Said another way: if it were determined that a rate ratio of 1.2 were of substantive interest and the observed rate ratio of 1.60 were selected for review on account of its passing a $p<0.05$ threshold and achieving statistical significance, the Jones et al. (2015) study power was not sufficient to distinguish the observed rate ratio of 1.60 from a true rate ratio of 1.2.

4. Discussion

The above examples have demonstrated that ESM has the potential to be considerable when the power of a study is low. From a practical perspective, these simulation results demonstrate that ESM should be of interest to those evaluating statistically significant results from low powered studies and that any large effect sizes observed from such studies should be interpreted cautiously.

One question the reader may ask is how these estimated e-magnification intervals differ from or relate to the typical confidence intervals around point estimates that populate much of the literature. In addition, the reader may ask what advantages there are to considering **both** the (classic) 95% confidence interval around the effect size and any estimated effect magnification interval as derived through -`emagnification`-. Further, the reader may wonder about extent to which the (classic) confidence interval and the effect size magnification interval are expected to be similar and how these two intervals should best be interpreted by the practitioner with respect to specific study results.

First, the (classic) confidence interval around an effect size (such as a mean or mean difference, or an odds ratio, rate ratio, or hazard ratio) suggests -- loosely speaking -- a plausible range for that effect size estimate if (i) the experiment or study were to be repeated many times (ii) if all errors in the study were random (no systematic errors/biases); and (iii) if the underlying assumed statistical model (e.g., binomial in the case of an odds ratio or rate ratio) was correct.[8] For example, for the resin worker case study presented earlier the odds ratio was estimated to be 1.76 with a (classic) 95% confidence interval of (1.20, 2.58). This confidence interval can be interpreted as a plausible range of estimated odds ratios were the observational study to be repeated many times in the exact same way and if all differences in results in those study replications could be ascribed entirely to the random nature of the (here, Bernoulli/binomial) data generation process.

What `-emagnification-` does is address a separate and distinct question from that addressed by the classic confidence interval. This separate and distinct question was originally suggested by Ioannidis and later more directly by Gelman and Carlin (2014) for the case of continuous data (rather than categorical data that is the focus here). The question that the emagnification approach addresses is what would happen if the study were repeated many times given a range of user-assumed (true) effect sizes *if one were interested in and focused on only statistically significant results*. This may be particularly useful to the user who is studying (expected) small effects using noisy measurements with small sample sizes since it is this user that is most likely to experience (or be bitten by) ESM. This is due in part to the fact that the low power of the study to detect small effect sizes leads unstable to p-values.[9] The Greenland et al. (1994) resin worker case study presented earlier, for example, estimated an odds ratio of 1.76 that was shown to be statistically significant with "plausible" bounds for the classic 95% CI of 1.20 to 2.58 which assumes all errors are random. E-magnification addresses an alternative but still very important and relevant question: if the study were to be repeated thousands of times and all errors were (still) random, what is a plausible range for (other) statistically significant results given the size/power of the actual study, the exposed proportion among the control reference group, and any (assumed series of) true effect sizes. For the resin worker study, the `-emagnification-` command allows one to say that if the true odds ratio were only 1.20 (instead of the 1.76 estimated by the study) then 80% of the statistically significant results if the study were to be repeated many times would be between 1.42 and 1.79 (with half below the median estimated OR of 1.53 for statistically significant results and half above).[10] Corresponding (80%) emagnification intervals if the true (yet in reality unknown) odds ratios were 1.5, 2.0 and 3.0 would be (1.45, 2.01), (1.60, 2.55) and (2.35, 3.78), respectively. How might these emagnification results be interpreted by the data user?: Although Greenland et al. (1994) estimated a statistically significant odds ratio for the relationship between lung cancer and exposure to resins of 1.76 (95% CI: 1.20 to 2.58), it would not be unusual to observe this high an odds ratio if the true OR were in fact as low as 1.2 if one were to select (i.e., condition

[8] A stricter and more technically accurate description would indicate that a confidence interval is an interval that is expected to contain the true parameter (or effect size) over an infinite number of repetitions of the study with a frequency no less than the confidence level, provided that the underlying statistical model is correct and there is no bias (Rothman et al. (2008)). This is an indirect interpretation of what is likely the real question of interest when interpreting a single study of interest: "does the confidence interval so constructed contain the true parameter (or effect size) of interest?" Unfortunately – and contrary to some interpretations by non-statisticians – the confidence interval is unable address this question under a frequentist interpretation. The best that can be said and remain true to the definition of a confidence interval and how it is calculated would be to say the interval was calculated in such a way so that if the study were repeated many times, the true effect size would be expected to be contained in the calculated interval with a frequency no less than the confidence level used to calculated the confidence bounds, given the provisos that the underlying statistical model is correct and that all errors are random with none systematic (i.e., no bias).

[9] See for example Geoff Cumming's the "dance of the p values" video at https://www.bing.com/videos/search?q=dance+of+the+p+values&view=detail&mid=6D48A4D9F8A653BA10496D48A4D9F8A653BA1049&FORM=VIRE which illustrates how the p-value – particularly for low powered studies – can be very imprecise.

[10] Using `emagnification proportion, p0(`=257/1202') or (1.2 1.5 2.0 3.0) n0(1202) n1(139) pctile(5 50 95) nsim(1000) level(0.05) onesided seed(123) log`

on) only those odds ratios that were statistically significant. More specifically, odds ratios for those results *that were statistically significant* would be expected to range from 1.42 to 1.79 in the middle 80% of the distribution of (those) observed statistically significant odds ratios. What this means is that the observed/reported statistically significant odds ratio of 1.76 may have been substantively affected by ESM if observed odds ratio had been "called out" due to it passing the $p<0.05$ threshold. What does this mean for the data user? The data user should understand that the study might not be adequately powered and that the "discovered" statistically significant effect size (here, a statistically significant odds ratio of 1.76 with 95% CI of 1.20, 2.58) might be reasonably attributable to ESM even if the true yet unknown odds ratio were as low as 1.2. In short, consideration of ESM allows the user to view the data from a different angle from that of the classic 95% confidence interval, one that allows "what if" scenarios to be played out to examine (particularly) the effects that low powered studies with imprecise effect sizes might have on statistically significant results.

Some readers may question these ESM calculations which focus on and emphasizes the power of a study and consider them to be simply a variant of (discredited) *post-hoc* power calculations. They are not[11]. Instead, ESM calculations can be considered to be calculations related to the "design calculations" or "post-data design analysis" advocated by Gelman and Carlin in their article in *Perspectives in Psychological Science* entitled "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors" (Gelman and Carlin, 2014)[12,13] and discussed further more recently in greater mathematical detail in Lu et al (2019). Gelman and Carlin advocate using power calculations -- re-emphasized and named "design calculations " to focus on errors in magnitude and sign instead of declarations of statistical significance -- **after** the data has been collected to help inform a statistical data summary.[14] Although they focus on continuous outcomes rather than the categorical/contingency table outcomes focused on here with `-emagnification-`, Gelman and Carlin state that such design calculations are intended to address the relevant post-data collection question of not 'What is the power of a test?', but instead the more relevant *post-hoc* question of 'What might be expected to happen in studies of this size?'. This is what was done here in this article with `-emagnification-`: a variety of plausible (odds or rate) ratios were selected to cover a broad range of plausible underlying true effect sizes, and the question 'What might be expected to happen in studies of this size if the researcher focus is on discovered, statistically significant effect sizes?' was addressed. The answer for the Greenland et al. (2004) resin example as generated by `-emagnification-` would be that the subset of repeated studies of this size, power, and background rate

[11] The `-emagnification-` calculations are *not post-hoc* power calculations because they do not use the effect size estimates estimated by the study but instead estimate the observed effect size distributions for statistically-significant observed effects found assuming different, user-selected potential effect sizes.

[12] The article includes an R program on design calculations for experiments whose outcomes a continuous and that would be more typical in research in psychology (e.g., effect sizes measured as standardized mean differences) than epidemiology (e.g., effects sizes measured as odds, rate, or hazard ratios). Although written about and discussed in their article in terms of the design calculations they advocate, the underlying concepts and their implications are the same as applied here with `-emagnification-`.

[13] Importantly – and of relevance here – the R code published as part of the Gelman and Carlson article has been recently translated to Stata in the Stata command `-rdesigni-` written and recently updated by Daniel Klein at the University of Kassel and available for download from SSC (Klein, 2019). This Stata command implements the design analysis approach discussed in Gelman and Carlin (2014) and -- as is true for the Gelman and Carlin publication – approaches the issue from a design analysis/calculation perspective for *continuous* outcome data which is not necessarily easily adapted to the odds and rate ratios considered more typical in epidemiology and discussed here in the context of `-emagnification-`. A more recent user-written Stata program which is similar to Klein's `-rdesigni-` is `-retrodesign-` has been written by Arial Linden and is also available for download from SSC (Linden, 2019). Like Klein, this user-written program also is specific for continuous outcomes.

[14] See "Yes, it makes sense to do design analysis ("power calculations") after the data have been collected" at http://andrewgelman.com/2017/03/03/yes-makes-sense-design-analysis-power-calculations-data-collected/

that generated statistically significant results would be expected 80% of the time to produce odds ratios of between 1.42 and 1.79 when the true odds ratio is 1.2. The median odds ratio found among such statistically significant results would be 1.53, and reflect an inflation or magnification factor (termed a Type M error or "Exaggeration ratio" by Gelman and Carlin) of 28%. The Stata ESM calculations/simulations discussed here can be considered "sister" calculations specific for categorical data to the (*post-hoc*) design calculations for continuous data advocated by Gelman and Carlin since they derive from the same principles and address the same issues.

As we do here, Gelman and Carlin specifically recommend that such ESM-like *post-hoc* design calculations be done when strong statistically significant evidence for non-null effects have been found because "a *[discovered statistically*] significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect when researchers study small effects using noisy measurements and small sample sizes". Gelman and Carlin continue and state that such calculations may be even *more* relevant for findings that are found to be statistically significant because the interpretation of a statistically significant result can change quite substantially depending on the researchers belief in a plausible size of the underlying effect. Such an analysis will help better characterize study results and allow more meaningful interpretation of how unexpected a result might be in any discovery phase of research given a series of (user-defined) plausible true underlying effect sizes.

While this paper has focused on the issues associated with effect size magnification, it is important nevertheless for the researcher to recognize that there can be very legitimate and reasonable countervailing or counteracting forces that tend to drive effect sizes in the other direction (toward the null) or work toward effect size "deflation" or "suppression". Perhaps the most well-known of these is non-differential misclassification bias whereby non-differential misclassification of exposures (or disease) can result in a suppression of effect size, thereby leading under certain circumstances to a systematic under-prediction of the effect size (Rothman et al., 2008; Ioannidis, 2008). The oft-cited statement that "non-differential misclassification biases effects toward the null" is not always true and (particularly) for underpowered studies there can be counteracting forces that drive (significant) effect size estimates away from the null. Recognizing this is an important step forward. How these competing forces – biasing results *toward* the null in the case of non-differential misclassification and *away* from the null with effect size magnification when one focuses on statistically significant results -- "play out" will be very situation-dependent and cannot be predicted in advance; the `-emagnification-` command should help researchers characterize and begin to quantify to some extent the latter contributor. With respect to the potential contribution of misclassification biases toward overall bias, quantitative bias analysis (QBA) – as implemented by Stata's `-episens-` command (Orsini et al., 2008), for example -- would serve as a useful adjunct to `-emagnification-` in that it allows quantification of misclassification (and other) biases in a flexible and easy-to-use Stata software tool. Specifically, `-episens-` is able to assess the uncertainty of exposure–disease associations due to misclassification of the exposure or disease, due to selection bias, and due to unmeasured confounding: as such, it is able to quantitatively evaluate a variety of these potentially countervailing forces.[15] Used in conjunction with or as a 'second step" after evaluation of the potential for effect size magnification with `-emagnification-` and `-power twoproportions-`, the `-episens-` command can supplement and better define other potential biases, some of which may operate in the opposite or "effect deflation" direction. Unmeasured (residual) confounding is potentially another concern

[15] `-biasepi-` is a recently released user-written Stata command that also performs quantitative bias analysis and is available for download from the SSC ideas website by typing ssc install biasepi (Wu, 2019). Probabilistic bias analysis, however, is not supported.

in observational epidemiological studies, and a recent user-written Stata program that performs sensitivity analysis for unmeasured confounding is `-evalue-` which uses the methodology proposed by Vanderweele and Ding (2017) (Linden et al., 2019).

Finally, it should be remembered that although the discussion and examples used here focused on epidemiology, the ESM phenomenon is a principle applicable to discovery science in general and is not a specific affliction or malady of epidemiology (Button (2013); Button *et al.* (2013); Lehrer (2010); Ioannidis (2005); Ioannidis (2008); Reinhart (2015)). As indicated earlier, it is often seen in studies in pharmacology, in gene studies, in psychological studies, and in oft-cited medical literature. Such truth inflation would be expected to be more characteristic in fields where studies are small and underpowered because such studies have widely varying results or where many researchers are performing similar studies and compete to publish "new" or "exciting" results (Ioannidis (2005); Reinhart (2015)).

5. Summary and Conclusion

While it is generally widely-known that small, low-powered studies can result in false negatives since the study power may be inadequate to reliably detect an effect size deemed to be meaningful by the investigator, it is less well known that these studies can result in inflation of estimates of effect size if those estimated effects are required to pass a statistical threshold (e.g., the common $p<0.05$ threshold used for statistical significance) to be judged important, relevant, or "discovered". More specifically: low powered studies tend to produce greater degrees of ESM in results that are found to be statistically significant (or pass other threshold criteria) than higher powered studies. The phenomenon is not specific to epidemiology and is applicable to any science in which studies tend to be underpowered and emphasize the use of p-values to "discover" an effect, and it is important that users of statistical study results recognize this issue and its potential interpretational consequences. Specifically: <u>any discovered associations from an underpowered study that are highlighted or focused upon on the basis of passing a statistical or other similar threshold will be systematically biased away from the null</u>.

The potential degree of this inflation or bias away from the null will depend on a number of issues including: the background rate of the outcome of interest; the sample size of the study; and the effect size of interest. It follows that low powered epidemiological studies investigating small or weak effects in populations that have a low background rate of the (health) outcome of interest will tend toward the greatest degree of ESM. It is important to recognize that this is an issue related to how studies are interpreted by users, and not one that is intrinsic to or the fault of the study design, nor one that is related to good scientific principles or practices.

This article introduces the Stata command `-emagnification-` which follows directly from work of John Ioannidis appearing in 2008 in the journal *Epidemiology,* "Why most discovered true associations are inflated" (Ioannidis, 2008) and to a lesser extent that of Gelman and Carlin (2014). In the 2008 Ioannidis article, Ioannidis illustrates by simulation the ESM phenomenon, and these ideas were incorporated into `-emagnification-` such that the calculations similar to those done by Ioannidis to estimate the degree of potential ESM can be easily performed in Stata. The take home message from these simulations and the original work by Ioannidis and extensions by Gelman and Carlin (2014) is that a study should be not only suitably powered to avoid a false negative (Type II error) but also be suitably powered to avoid substantial ESM if it is statistically significant or other threshold-crossing criteria that are of interest. It is important to note that if a study is suitably powered (oftentimes 80% or more), there is no systematic risk inflation

In sum, the effect size magnification phenomenon is real, is important for more appropriately interpreting underpowered studies, and in many ways is under-recognized and under-appreciated in the research community and among regulators and decision-makers. The new `-emagnification-` Stata command introduced here is a tool that permits reported statistically-significant effect size estimates from possibly underpowered epidemiological studies to be better evaluated and judged. As such, it can assist individuals reviewing such studies to put an observed statistically significant effect size into a fuller context that allows better judgments regarding adequacy of sample size *vis-a-vis* the observed effect size. In doing so, users will gain a better understanding of power and sample size issues and in interpreting their potential implications with respect to study conclusions.

6. References

American Statistical Association (ASA). 2017. "Recommendations to Funding Agencies for Supporting Reproducible Research." [accessed 13 August 2019 at https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf]

Button, K., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flink, Emma S.J. Robinson, and M.R. Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365-376. [accessed 13 August 2019 at http://www.nature.com/nrn/journal/v14/n5/full/nrn3475.html]

Button, K. 2013. "Unreliable neuroscience? Why power matters". *The Guardian* newspaper (UK). 10 April [Accessed 13 August 2019 at https://www.theguardian.com/science/sifting-the-evidence/2013/apr/10/unreliable-neuroscience-power-matters]

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues). 2017. Ockleford C, Adriaanse P, Berny P, Brock T, Duquesne S, Grilli S, Hougaard S, Klein M, Kuhl T, Laskowski R, Machera K, Pelkonen O, Pieper S, Smith R, Stemmer M, Sundh I, Teodorovic I, Tiktak A, Topping CJ, Wolterink G, Bottai M, Halldorsson T, Hamey P, Rambourg M-O, Tzoulaki I, Court Marques D, Crivellente F, Deluyker H and Hernandez-Jerez AF, 2017. Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'. *EFSA Journal* 2017:15(10):5007, 101 pp. [accessed 13 August 2019 at https://doi.org/10.2903/j.efsa.2017.5007]

Gelman, A. and J. Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*. Vol 9(6): 641-651. [accessed 13 August 2019 at http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf]

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). E124.doi:10.1371.pmed.0020124. [accessed 13 August 2019 at http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124]

Ioannidis, J. P. A. 2008. Why most discovered true associations are inflated. *Epidemiology* 19: 640-648. [accessed 13 August 2019 at http://www.dcscience.net/ioannidis-associations-2008.pdf]

Jones R. R., F. Barone-Adesi, S. Koutros, C. C. Lerro, A. Blair, J. Lubin, S. L. Heltshe, J. A. Hoppin, M. C. Alavanja, and L. E. Beane-Freeman. 2015. Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study: an updated analysis. *Occup Environ Med* 72: 496-503.

Klein, D. (2019). RDESIGNI: Stata module to perform design analysis. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458423.html

Lehrer, J. 2010. "The Truth Wears Off: Is there something wrong with the scientific method". *New Yorker*. 13 December. [Accessed 13 August 2019 at http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off]

Linden A. (2019). RETRODESIGN: Stata module for computing type-S (Sign) and type-M (Magnitude) errors. Statistical Software Components, Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s458631.html

Linden A, Mathur M. B., VanderWeele, T. J. (2018). EVALUE: Stata module for conducting sensitivity analyses for unmeasured confounding in observational studies. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458592.html

Lu, J., Qiu, Y. and Deng, A. 2019. A note on Type S/M errors in hypothesis testing. British Journal of Mathematical and Statistical Psychology 72(1):1-17. [Accessed 13 August 2019 at https://onlinelibrary.wiley.com/doi/pdf/10.1111/bmsp.12132

Orsini, N., Bellocco, R., Bottai, M, Wolk, Alicja, and Greenland, S. 2008. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *The Stata Journal* 8(1): 29-48. [accessed 13 August 2019 at http://www.stata-journal.com/article.html?article=st0138]

Orsini, N., Bellocco, R., Bottai, M. and Greenland S. (2006). EPISENS: Stata module for Deterministic and probabilistic sensitivity analysis. Statistical Software Components, Boston College Department of Economics. Revised 13 August 2019. https://ideas.repec.org/c/boc/bocode/s456792.html

Reinhart, A. 2015. Statistics Done Wrong: the woefully complete guide. No Starch Press (San Francisco, CA).

Rothman, KJ, Greenland, S and Lash, TL. Modern Epidemiology. 2008. 3rd ed. Lippincot, Williams, and Wilkins. Philadelphia.

VanderWeele, T. J., and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. Annals of Internal Medicine, 167(4): 268-274.

Wu, Chunsen (2019). BIASEPI: module to perform simple bias analysis, multidimensional bias analysis, and multiple bias modeling. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458617.html

Yarkoni, T. 2009. "Ioannidis on effect size inflation, with guest appearance by Bozo the Clown." 21 November. [Accessed on 13 August 2019 at http://www.talyarkoni.org/blog/2009/11/21/ioannidis-on-effect-size-inflation-with-guest-appearance-by-bozo-the-clown/]

About the authors

David Miller is formerly a senior statistician and currently a supervisory chemist/branch chief in the Health Effects Division of US EPA's Office of Pesticide Programs in Washington, DC.

James Nguyen is a mathematical statistician in the Health Effects Division of US EPA's Office of Pesticide Programs in Washington, DC.

Matteo Bottai is a professor of biostatistics in the Unit of Biostatistics at the Institute of Environmental Medicine at Karolinska Institutet in Stockholm, Sweden.

ANNEX A: Stata Code for -emagnification-

```
. which emagnification
c:\ado\plus\e\emagnification.ado
*! v.1.0.0 Matteo Bottai 12aug2019
**************************************************
. type "c:\ado\plus\e\emagnification.ado"

*! v.1.0.0 Matteo Bottai 12aug2019
cap program drop emagnification
program emagnification, rclass
version 14
syntax namelist , *
if strpos("proportion","`namelist'")==1 {
syntax namelist , p0(numlist >0 <1) or(numlist >0) ///
        n0(numlist integer >0) n1(numlist integer >0) ///
        [ pctile(numlist integer >0 <100) ifactor(numlist integer >0 <100) nsim(numlist max=1
integer >0) ///
        level(numlist max=1 >0 <1) onesided EXACT seed(string) log clean format(string) ]
if "`nsim'"=="" {
        local nsim 10
        }
if "`seed'"=="" {
        local seed = c(rngstate)
        }
if "`level'"=="" {
        local level .05
        }
if "`pctile'"=="" {
        local pctile = "10 50 90"
        }
if "`format'"=="" {
        local format %4.3f
        }
if "`exact'"=="" {
        local test "chi2"
        local retp = cond("`onesided'"=="","r(p)","r(p)/2")
        }
        else {
        local test "exact"
        local retp = cond("`onesided'"=="","r(p_exact)","r(p1_exact)")
        }
preserve
qui drop _all
set seed `seed'
qui set obs `=(`nsim')*(`: word count `p0'')*(`: word count `or'')*(`: word count `n0'')*(`: word
count `n1'')'
qui gen p0 = .
qui gen p1 = .
qui gen true_or = .
qui gen n0 = .
qui gen n1 = .
qui gen sim = .
qui gen or = .
qui gen p = .
local c = 0
local s = 0
qui foreach p_0 of local p0 {
foreach o of local or {
local p_1 = (`o'*`p_0'/(1-`p_0'))/(1+(`o'*`p_0'/(1-`p_0')))
foreach n_0 of local n0 {
foreach n_1 of local n1 {
local ++s
if "`log'"!="" {
        noi di _new as txt "Scenario " as res `s' as txt ": p0 = " as res `p_0' as txt ", or = "
as res `o' as txt ", n0 = " as res `n_0' as txt ", n1 = " as res `n_1'
        noi di as txt "Completed: " _cont
        }
```

```
cap set obs `=`n_0'+`n_1''
gen x = _n>`n_0' in 1/`=`n_0'+`n_1''
forv i = 1/`nsim' {
if "`log'"!="" & mod(`i',round(`nsim'/10))==0 {
        noi di as txt int(`i'/round(`nsim'/10))*10 "% " _cont
        }
local ++c
replace p0 = `p_0' in `c'
replace p1 = `p_1' in `c'
replace true_or = `o' in `c'
replace n0 = `n_0' in `c'
replace n1 = `n_1' in `c'
replace sim = `i' in `c'
gen y = (runiform()<(cond(x==0,`p_0',`p_1')))
tab y x, matcell(f) `test'
replace or = f[1,1]*f[2,2]/f[1,2]/f[2,1] in `c'
replace p = `retp' in `c'
drop y
}
drop x
}
}
}
}
// qui replace or = c(maxfloat) if sim!=. & p!=. & or==.
qui drop if sim==. | p==.
qui gen if_or = or/true_or
qui egen group = group(p0 true_or n0 n1)
qui egen valid = count(p), by(group)
qui egen power = mean((p<`level') & ((or>1)==(true_or>1) | "`onesided'"==")), by(group)
qui keep if (p<`level') & ((or>1)==(true_or>1) | "`onesided'"==")
qui foreach p of local pctile {
        egen p`p' = pctile(or), p(`p') by(group)
        format `format' p`p'
        }
qui foreach p of local ifactor {
        egen if_p`p' = pctile(if_or), p(`p') by(group)
        format `format' if_p`p'
        }
qui egen tag = tag(group)
qui keep if tag==1
drop or p sim group tag if_or
di _newline(2) as txt "The tests are " as res cond("`onesided'"==", "two-sided", "one-sided") as
txt " with level = " as res `level'
list , noobs sepby(p0) `clean'
}
else if strpos("rate","`namelist'")==1 {
syntax namelist , r0(numlist >0) rr(numlist >0) ///
        n0(numlist integer >0) n1(numlist integer >0) ///
        [ pctile(numlist integer >0 <100) ifactor(numlist integer >0 <100) nsim(numlist max=1
integer >0) ///
        level(numlist max=1 >0 <1) onesided EXACT seed(string) log clean format(string) ]
if "`nsim'"=="" {
        local nsim 10
        }
if "`seed'"=="" {
        local seed = c(rngstate)
        }
if "`level'"=="" {
        local level = .05
        }
if "`pctile'"=="" {
        local pctile = "10 50 90"
        }
if "`format'"=="" {
        local format %4.3f
        }
preserve
qui drop _all
set seed `seed'
```

```
qui set obs `=(`nsim')*(`: word count `r0'')*(`: word count `rr'')*(`: word count `n0'')*(`: word
count `n1'')'
qui gen r0 = .
qui gen r1 = .
qui gen true_rr = .
qui gen n0 = .
qui gen n1 = .
qui gen sim = .
qui gen rr = .
qui gen p = .
local c = 0
local s = 0
qui foreach r_0 of local r0 {
foreach r of local rr {
local r_1 = `r'*`r_0'
foreach n_0 of local n0 {
foreach n_1 of local n1 {
local ++s
if "`log'"!="" {
        noi di _new as txt "Scenario " as res `s' as txt ": r0 = " as res `r_0' as txt ", rr = "
as res `r' as txt ", n0 = " as res `n_0' as txt ", n1 = " as res `n_1'
        noi di as txt "Completed: " _cont
        }
cap set obs `=`n_0'+`n_1''
gen x = _n>`n_0' in 1/`=`n_0'+`n_1''
forv i = 1/`nsim' {
if "`log'"!="" & mod(`i',round(`nsim'/10))==0 {
        noi di as txt int(`i'/round(`nsim'/10))*10 "% " _cont
        }
local ++c
qui replace r0 = `r_0' in `c'
qui replace r1 = `r_1' in `c'
qui replace true_rr = `r' in `c'
qui replace n0 = `n_0' in `c'
qui replace n1 = `n_1' in `c'
qui replace sim = `i' in `c'
gen y = rpoisson(cond(x==0,`r_0',`r_1'))
cap poisson y x
if _rc==0 {
        replace rr = exp(_b[x]) in `c'
        if "`exact'"=="" {
                replace p = normal(-abs(_b[x]/_se[x]))*2/(1+("`onesided'"!="")) in `c'
                }
                else {
                cap expoisson y x
                if _rc==0 {
                        mat p = e(p_sufficient)
                        replace p = p[1,1]/(1+("`onesided'"!="")) in `c'
                        }
                }
        }
drop y
}
drop x
}
}
}
}
qui drop if sim==. | p==.
qui gen if_rr = rr/true_rr
qui egen group = group(r0 true_rr n0 n1)
qui egen valid = count(p), by(group)
qui egen power = mean((p<`level') & (sign(rr-1)==sign(true_rr-1))), by(group)
qui keep if (p<`level') & (sign(rr-1)==sign(true_rr-1))
qui foreach p of local pctile {
        egen p`p' = pctile(rr), p(`p') by(group)
        format `format' p`p'
        }
qui foreach p of local ifactor {
        egen if_p`p' = pctile(if_rr), p(`p') by(group)
        format `format' if_p`p'
```

```
        }
qui egen tag = tag(group)
qui keep if tag==1
drop rr p sim group tag if_rr
di _newline(2) as txt "The tests are " as res cond("`onesided'"=="", "two-sided", "one-sided") as
txt " with level = " as res `level'
list , noobs sepby(r0) `clean'
}
else {
        di as err "Only 'proportion' and 'rate' are allowed"
        exit
        }
tempname table
mkmat _all, matrix(`table')
return matrix table `table'
return scalar level = `level'
return local seed `seed'
return local cmdline "emagnification `0'"
restore
end
```

ANNEX B: Stata Help File for -emagnification-

Title

emagnification -- Effect magnification

Syntax

Effect magnification for proportions

emagnification proportion, p0(numlist) or(numlist) n0(numlist) n1(numlist) [other_options]

Effect magnification for rates

emagnification rate, r0(numlist) rr(numlist) n0(numlist) n1(numlist) [other_options]

Description

emagnification estimates effect magnification of proportions and rates through simulations.

Options

p0(numlist) specifies the proportions in the reference group

or(numlist) specifies the odds ratios of the exposure group versus the reference group

r0(numlist) specifies the rates in the reference group

rr(numlist) specifies the risk ratios of the exposure group versus the reference group

n0(numlist) specifies the sample size in the reference group

n1(numlist) specifies the sample size in the exposure group

Other options

pctile(numlist) shows the percentiles specified in numlist; defaults to 10 50 90

ifactor(numlist) shows the inflation factors of the percentiles specified in numlist

nsim(#) number of simulated datasets; defaults to 10

level(#) significance level of the test; defaults to 0.05

onesided specifies a one-sided test

```
      exact for proportion specifies the Fisher's exact test instead of the default
chi-square test; for rates specifies the exact Poisson regression instead of the
default Poisson regression

      seed(string)     specifies the seed for the pseudo-random numbers generator

      log              shows the simulation iterations

      clean            shows the results without separator lines

      format(format)   specifies the display format for the percentiles
```

Examples

```
    Estimate the effect magnification for a proportion:

        emagnification proportion, p0(.5) or(2) n0(100) n1(100)

    Estimate the effect magnification for a rate:

        emagnification rate, r0(.5) rr(2) n0(100) n1(100)

    Estimate the effect magnification for a proportion in multiple scenarios:

        emagnification proportion, p0(.5 .9) or(1.5 2) n0(100 200) n1(100 200)
pctile(50 90) ifactor(50) nsim(20) level(.1) onesided seed(123) log clean



    Show the saved results of the latest estimation:

        return list

        matrix list r(table)

Saved results

    Scalars

      r(level)   the level of the tests

    Macros

      r(cmdline) command as typed

      r(seed)    the seed used by the pseudo-random numbers generator

    Matrices

      r(table)   the table of the results
```

Also see

```
    [PSS] power

    [R] seed
```

Reference

Miller, D. J., Nguyen, J. T., and Bottai, M. emagnification: A tool for estimating effect size magnification and performing design calculations in epidemiological studies. Stata Journal (submitted)

Author

Matteo Bottai

Unit of Biostatistics

Institute of Environmental Medicine, Karolinska Institutet

Stockholm, Sweden

ANNEX C: Sample Test files for `-emagnification-`

Testing of `-emagnification-` was done by comparing native Stata and SAS code for power calculations (`-power twoproportions-` and PROC POWER, respectively) to the power simulations resulting from `-emagnification-` and also comparing the output of `-emagnification-` to output from SAS code originally developed for use in Annex D of 'Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report Literature review of epidemiological studies linking exposure to pesticides and health effects' published by the European Food Safety Agency's Panel on Plant Protection Products and their Residues (PPR) in the EFSA Journal (EFSA PPR Panel Report, 2017). Below are three representative test cases which illustrate some of the analyses and comparisons which were performed.

Illustrative Test Case #1

We selected a study investigating the association between malathion and non-Hodgkin's lymphoma (NHL) (Waddell et al. (2001), [link]). Here, we have i) the number of subjects in the reference non-exposed group = 1018 (from Table 1 of the Wadell et al. article: non-farmers = 243 cases + 775 controls); ii) the number of subjects in the exposed group = 238 (from Table 4 of Wadell et al. : malathion users = 91 cases + 147 controls); and iii) the number of cases in the reference non-exposed group =243 (from Table 1: 243 cases in non-farmer or non-exposed group). While the authors reported an adjusted OR = 1.6, 95% CI = 1.2 – 2.2, we can compute the crude OR = 1.97, 95% CI = 1.46 – 2.66:

```
. cci 91 243 147 775, woolf

                                                      Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |        91         243  |        334       0.2725
       Controls |       147         775  |        922       0.1594
----------------+------------------------+------------------------
          Total |       238        1018  |       1256       0.1895
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.974329       |    1.464792    2.661111 (Woolf)
 Attr. frac. ex. |         .4934988       |    .3173092    .6242172 (Woolf)
 Attr. frac. pop |         .1344562       |
                +-------------------------------------------------
                             chi2(1) =    20.39  Pr>chi2 = 0.0000
```

We can calculate the power of the comparisons between the ever vs. never exposed, given the assumption that any true OR = 1.2, 1.5, 2.0, etc. using Stata's standard power commands:

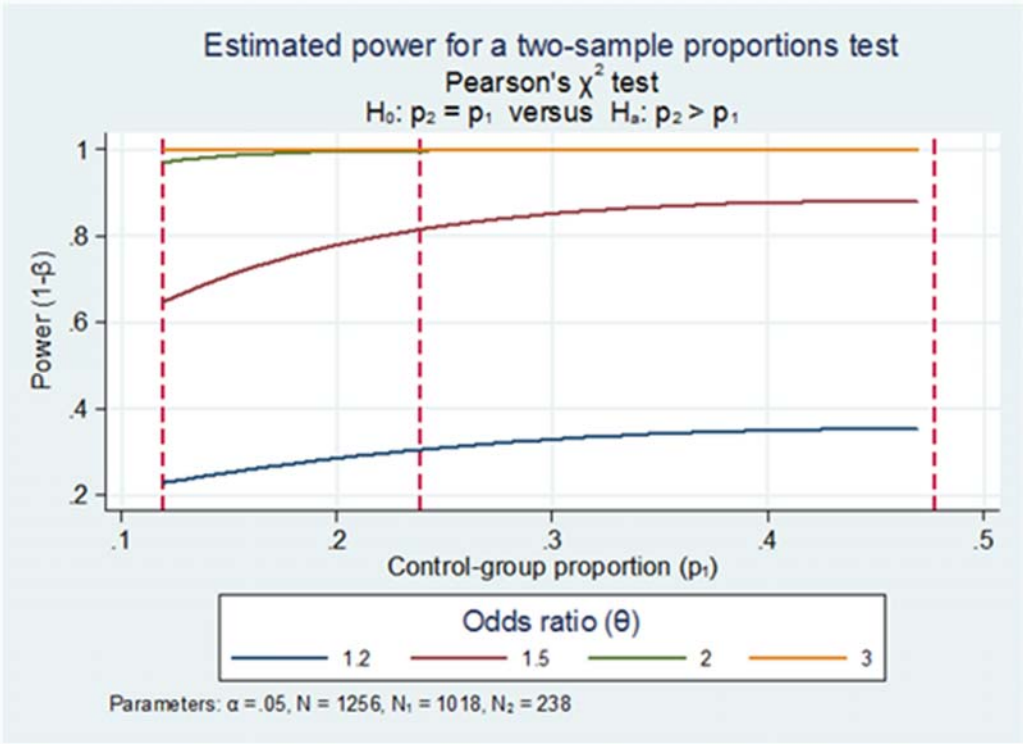
```
. power twoproportions (`=0.5* 243/1018' `=243/1018' `=2 * 243/1018'),
test(chi2) oratio(1.2 1.5 2.0 3.0) n1(1018) n2(238) onesided

Estimated power for a two-sample proportions test
```

```
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 > p1
```

alpha	power	N	N1	N2	delta	p1	p2	oratio
.05	.2279	1256	1018	238	1.2	.1194	.1399	1.2
.05	.647	1256	1018	238	1.5	.1194	.1689	1.5
.05	.9693	1256	1018	238	2	.1194	.2133	2
.05	1	1256	1018	238	3	.1194	.2891	3
.05	.3047	1256	1018	238	1.2	.2387	.2734	1.2
.05	.8149	1256	1018	238	1.5	.2387	.3199	1.5
.05	.9971	1256	1018	238	2	.2387	.3854	2
.05	1	1256	1018	238	3	.2387	.4847	3
.05	.3522	1256	1018	238	1.2	.4774	.523	1.2
.05	.8779	1256	1018	238	1.5	.4774	.5781	1.5
.05	.9992	1256	1018	238	2	.4774	.6463	2
.05	1	1256	1018	238	3	.4774	.7327	3

Such power relations are graphed below with the middle dotted line in the graph showing power at the NHL proportion of 0.2387 among non-farmers/non-exposed and the left-hand and right-hand vertical dashed lines representing a form of sensitivity analysis at one-half and twice the NHL proportion among non-farmers/non-exposed, respectively.

```
power twoproportions (`=0.5* 243/1018'(0.01) `=2 * 243/1018'),
test(chi2) oratio(1.2 1.5 2.0 3.0) n1(1018) n2(238)graph(recast(line)
xline(`=0.5* 243/1018'  `=243/1018' `=2 * 243/1018',lpattern(dash))
legend(rows(1)size(small)) ylabel(0.2(0.2)1.0)) onesided
```



Using -emagnification-, we see that the predicted powers are similar to those produced above by Stata's -power twoproportion- command:

```
emagnification proportion, p0(`=0.5* 243/1018' `=243/1018' `=2 *
243/1018') or(1.2 1.5 2.0 3.0) n0(1018) n1(238) pctile(10 50 90)
nsim(1000) onesided seed(123)

The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p10	p50	p90
.1193517	.139883	1.2	1018	238	1000	.228	1.417	1.524	1.784
.1193517	.1689455	1.5	1018	238	1000	.625	1.445	1.650	2.034
.1193517	.2132514	2	1018	238	1000	.976	1.610	1.997	2.518
.1193517	.2890563	3	1018	238	1000	1	2.406	3.013	3.780
.2387033	.2733921	1.2	1018	238	1000	.3	1.319	1.422	1.616
.2387033	.3198771	1.5	1018	238	1000	.817	1.361	1.568	1.868
.2387033	.3854084	2	1018	238	1000	.996	1.646	1.988	2.416
.2387033	.4847074	3	1018	238	1000	1	2.464	2.988	3.633
.4774067	.5229555	1.2	1018	238	1000	.321	1.287	1.367	1.565
.4774067	.5781126	1.5	1018	238	1000	.866	1.331	1.535	1.819
.4774067	.6462766	2	1018	238	1000	.999	1.665	1.971	2.448
.4774067	.7326633	3	1018	238	1000	1	2.419	2.998	3.724

Further, we also see that the p10, p50, p90 and the power values from Stata's -emagnification- command above are similar to those generated by the corresponding separately-developed SAS code as illustrated in the summary table below.

SAS Simulation Results Illustrating Effect Size Magnification Given *True* Odds Ratios of 1.2, 1.5, 2.0, and 3.0[a]

True values		N analyzed datasets	Power[b]	Distribution of Observed Significant ORs			
Proportion of diseased individuals in non-exposed group	OR			N	10th Percentile	Median (% inflation)	90th Percentile
0.1194 (1/2 background)	1.2	1000	0.22	220	1.4	1.5 (25%)	1.8
	1.5	1000	0.66	661	1.5	1.7 (13%)	2.0
	2	1000	0.97	972	1.6	2.0 (0%)	2.5
	3	1000	1.0	1000	2.4	3.0 (0%)	3.7
0.2387 (1x background)	1.2	1000	0.32	323	1.3	1.4 (17%)	1.6
	1.5	1000	0.81	812	1.4	1.6 (7%)	1.8
	2	1000	1.0	997	1.6	2.0 (0%)	2.4
	3	1000	1.0	1000	2.5	3.0 (0%)	3.6
0.4774	1.2	1000	0.34	337	1.3	1.4 (17%)	1.6

SAS Simulation Results Illustrating Effect Size Magnification Given *True* Odds Ratios of 1.2, 1.5, 2.0, and 3.0[a]							
True values		**N analyzed datasets**	**Power[b]**	**Distribution of Observed Significant ORs**			
Proportion of diseased individuals in non-exposed group	**OR**			**N**	**10th Percentile**	**Median (% inflation)**	**90th Percentile**
(2x background)	1.5	1000	0.87	872	1.3	1.5 (0%)	1.8
	2	1000	1.0	1000	1.6	2.0 (0%)	2.5
	3	1000	1.0	1000	2.4	3.0 (0%)	3.7

NOTE: The logistic regression model was used to compute the odds ratios for the two groups. The EXACT Test was used in the analysis of some datasets when the maximum likelihood estimate did not exist (perhaps due to a zero cases in one of the groups).

[a]: One-sided test, α = 0.05, N non-exposed=1018, N malathion exposed = 238, N iterations = 1000 (datasets)

[b]: the power resulting from this simulation may be close but not match exactly with the power calculated from built-in procedures in statistical software such as SAS (PROC POWER) or Stata (power twoproportion). This may be due to number of datasets simulated being of insufficient size. However, 1000 iterations is sufficient to adequately estimate the power and to illustrate the degree of effect size magnification given a statistically significant result (here alpha ≤ 0.05).

Illustrative Test Case #2

The following is an excerpted table from an article that looks at odds ratios for semen quality for men exposed to elevated pesticide levels. Looking specifically at atrazine in MO, we see an OR of 11.3 (95% CI: 1.3, 98.9).

Table 4. ORs for low semen quality for men exposed to elevated pesticide levels.

	Missouri				Minnesota			
Pesticide	Level (μg/g creatinine)	Cases	Controls	OR (95% CI)	Level (μg/g creatinine)	Cases	Controls	OR (95% CI)
Alachlor	< 0.15	3	15	Reference	< 0.15	6	21	Reference
	0.15–0.7	10	8	6.3 (1.3–29.4)	≥ 0.15	3	6	1.8 (0.3–9.2)
	> 0.7	12	2	30.0 (4.3–210)				
IMPY	< 0.1	6	20	Reference	< 0.1	3	12	Reference
	0.1–3.0	9	3	10.0 (2.0–49.2)	0.1–3.0	3	9	1.3 (0.2–8.2)
	> 3.0	10	2	16.7 (2.8–98.0)	> 3.0	3	6	2.0 (0.3–13.1)
Atrazine	< 0.1	17	24	Reference	< 0.1	9	25	
	≥ 0.1	8	1	11.3 (1.3–98.9)	≥ 0.1	0	2	
Metolachlor	< 0.15	5	11	Reference	< 0.15	5	17	Reference
	0.15–0.3	11	8	3.0 (0.7–12.2)	≥ 0.15	4	10	1.4 (0.3–6.3)
	> 0.3	9	6	3.3 (0.8–14.5)				
2,4-D	< 0.1	20	19	Reference	< 0.1	9	27	
	≥ 0.1	5	6	0.8 (0.2–3.0)	≥ 0.1	0	0	
1-Naphthol	< 1.5	9	2	Reference	< 2.0	3	8	Reference
	> 1.5	12	1	2.7 (0.2–34.2)	2.0–4.0	2	4	1.3 (0.2–11.5)
					> 4.0	4	15	0.7 (0.1–4.0)
3,5,6-Trichloropyridinol	< 0.5	5	2	Reference	< 1.25	1	3	Reference
	≥ 0.5	16	1	6.4 (0.5–86.3)	1.25–5.0	2	11	0.5 (0.04–8.3)
					> 5.0	6	13	1.4 (0.1–16.2)
4-Nitrophenol	< 0.1	20	3		< 0.1	5	14	Reference
	≥ 0.1	1	0		≥ 0.1	4	13	0.9 (0.2–3.9)

Here is the Stata run which reproduces the observed values from the above table for the Missouri data set:

```
. cci 8 17 1 24, woolf

                                                          Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |         8          17  |         25       0.3200
       Controls |         1          24  |         25       0.0400
----------------+------------------------+------------------------
          Total |         9          41  |         50       0.1800
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |        11.29412        |     1.289901    98.88906 (Woolf)
Attr. frac. ex. |        .9114583        |    .2247467    .9898877 (Woolf)
Attr. frac. pop |        .2916667        |
                +-------------------------------------------------
                             chi2(1) =     6.64  Pr>chi2 = 0.0100
```

Here are 100,000 iterations in Stata, first *not* using the "exact" option and then using it:

```
. emagnification proportion, p0(`=1/25') or(1.2) n0(25) n1(25) pctile(10 50 90)
ifactor(50) nsim(100000) level(0.05) onesided seed(123) log

Scenario 1: p0 = .04, or = 1.2, n0 = 25, n1 = 25
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

The tests are one-sided with level = .05

  +-------------------------------------------------------------------------------------+
  |  p0        p1   true_or   n0   n1   valid      power     p10     p50     p90   if_p50 |
  |-------------------------------------------------------------------------------------|
  | .04   .047619       1.2   25   25   89403   .0481192   6.000   6.000   7.579    5.000 |
  +-------------------------------------------------------------------------------------+


. emagnification proportion, p0(`=1/25') or(1.2) n0(25) n1(25) pctile(10 50 90)
ifactor(50) nsim(100000) level(0.05) onesided seed(123) exact log

Scenario 1: p0 = .04, or = 1.2, n0 = 25, n1 = 25
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

The tests are one-sided with level = .05

  +-------------------------------------------------------------------------------------+
  |  p0        p1   true_or   n0   n1   valid      power     p10     p50     p90   if_p50 |
  |-------------------------------------------------------------------------------------|
  | .04   .047619       1.2   25   25   89403   .0028746   7.579   7.579   9.333    6.316 |
  +-------------------------------------------------------------------------------------+
```

… and here are the SAS results for the corresponding SAS code for effect size magnification:

SAS code results for 200K iterations:

NC=25, NE=25, NSim=200000

The MEANS Procedure

Analysis Variable : OddsRatio

CRate	OR	N	Power	N Obs	10th Pctl	Median	90th Pctl
0.04	1.2	178698	0.0119195514	2130	5.89	5.89	7.83

Since both Stata and SAS have built-in procedures to calculate (one-sided) power (i.e., -power twoproportion- and PROC POWER, respectively), we looked at the results of each software's built-in procedures for power calculations and how they compare, both to each other and to the power estimates provided above.

As highlighted below and using a reference group proportion of 0.04, a true odds ratio of 1.2, and n1=n2=25 along with a one-sided test (Ho: p2 = p1 versus Ha: p2 > p1), both Stata's native -power twoproportion- and SAS's PROC POWER estimate the power to be 0.065:

Stata -power twoproportions- results using the chi2 test:

```
. power twoproportions 0.04, test(chi2) oratio(1.2) n1(25) n2(25)
onesided

Estimated power for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 > p1

Study parameters:

        alpha =    0.0500
            N =        50
           N1 =        25
           N2 =        25
        delta =    1.2000  (odds ratio)
           p1 =    0.0400
           p2 =    0.0476
   odds ratio =    1.2000

Estimated power:

        power =    0.0651
```

Stata `-power twoproportions-` results using Fisher's exact test:

```
. power twoproportions 0.04, test(fisher) oratio(1.2) n1(25) n2(25)
onesided

Estimated power for a two-sample proportions test
Fisher's exact test
Ho: p2 = p1  versus  Ha: p2 > p1

Study parameters:

        alpha =    0.0500
            N =        50
           N1 =        25
           N2 =        25
        delta =    1.2000  (odds ratio)
           p1 =    0.0400
           p2 =    0.0476
   odds ratio =    1.2000

Estimated power and alpha:

        power =    0.0025
 actual alpha =    0.0011
```

Illustrative Test Case #3

The following is an excerpted from article on Parkinson's Disease (PD) in individuals exposed to paraquat (Lee et al., 2012). Here is the table from the Lee et al (2012) article; the adjusted OR (aOR) for exposure to "ambient residential and workplace exposures" is a statistically significant 1.36 (95% CI: 1.02-1.81) and the corresponding crude OR (cOR) is also statistically significant at 1.43 (95% CI: 1.11-1.84):

Table 2 Associations between traumatic brain injury, ambient paraquat exposure, and Parkinson disease

	Cases n = 357, n (%)	Controls n = 754, n (%)	Crude model	Adjusted model[a]
Head trauma[b]				
No	315 (88.2)	681 (93.2)	1.0	1.0
Yes	42 (11.8)	50 (6.8)	1.82 (1.18-2.80)	2.00 (1.28-3.14)
Ambient residential and workplace exposures				
No	188 (52.7)	463 (61.4)	1.0	1.0
Yes	169 (47.3)	291 (38.6)	1.43 (1.11-1.84)	1.36 (1.02-1.81)

[a] Adjusted for age (continuous), gender, ever smoked, race, county, education (school years).
[b] Missing head trauma information for 23 subjects.

Performing this in Stata recognizing that the Number of Controls =754, Number of Cases =357, and Number of Exposed Individuals in Control group =291:

```
. cci 169 188 291 463, woolf

                                                      Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |       169         188  |        357       0.4734
       Controls |       291         463  |        754       0.3859
----------------+------------------------+------------------------
          Total |       460         651  |       1111       0.4140
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.430266       |    1.109128    1.844387 (Woolf)
Attr. frac. ex. |         .3008294       |    .0983908    .4578144 (Woolf)
Attr. frac. pop |         .1424094       |
                +-----------------------------------------------
                          chi2(1) =     7.64  Pr>chi2 = 0.0057
```

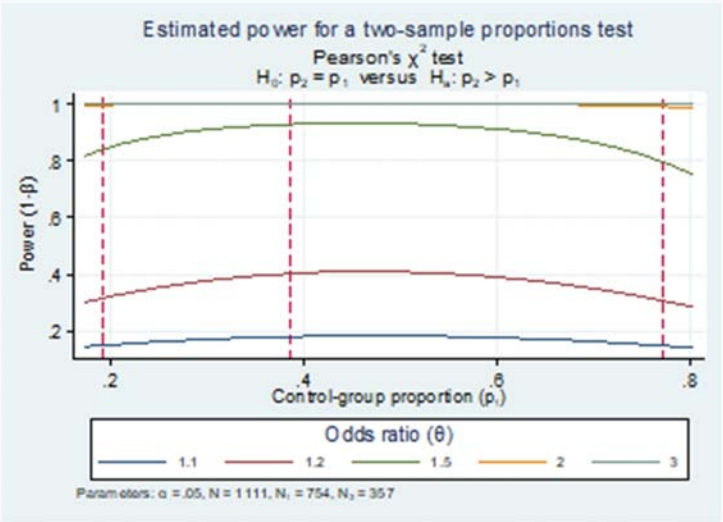
Using Stata's -powertwoproportions- command,

```
. power twoproportions (`=0.5* 291/754' `=291/754' `=2 * 291/754'), test(chi2)
oratio(1.1 1.2 1.5 2.0 3.0) n1(754)n2(357) onesided

Estimated power for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 > p1
```

alpha	power	N	N1	N2	delta	p1	p2	oratio
.05	.1498	1111	754	357	1.1	.193	.2082	1.1
.05	.3175	1111	754	357	1.2	.193	.223	1.2
.05	.8426	1111	754	357	1.5	.193	.264	1.5
.05	.9986	1111	754	357	2	.193	.3235	2
.05	1	1111	754	357	3	.193	.4177	3
.05	.1802	1111	754	357	1.1	.3859	.4088	1.1
.05	.403	1111	754	357	1.2	.3859	.4299	1.2
.05	.9309	1111	754	357	1.5	.3859	.4853	1.5
.05	.9999	1111	754	357	2	.3859	.5569	2
.05	1	1111	754	357	3	.3859	.6534	3
.05	.1483	1111	754	357	1.1	.7719	.7882	1.1
.05	.3068	1111	754	357	1.2	.7719	.8024	1.2
.05	.7969	1111	754	357	1.5	.7719	.8354	1.5
.05	.9929	1111	754	357	2	.7719	.8713	2
.05	1	1111	754	357	3	.7719	.9103	3

And then graphing this:

```
power twoproportions (`=0.45* 291/754'(0.01) `=2.1 * 291/754'),
test(chi2) oratio(1.1 1.2 1.2 1.5 2.0 3.0) n1(754) n2(357)
graph(recast(line) xline(`=0.5* 291/754'  `=291/754' `=2 *
291/754',lpattern(dash)) legend(rows(1)size(small))
ylabel(0.2(0.2)1.0)scheme(s2color)) onesided
```



And finally using the -emagnification- command:

```
emagnification proportion, p0(`=0.5* 291/754' `=291/754' `=2 * 291/754')
or(1.1 1.2 1.5 2.0 3.0) n0(754) n1(357) pctile(10 50 90)ifactor(50)
nsim(1000) level(0.05) onesided seed(123)


The tests are one-sided with level = .05
```

p0	p1	true_or	n0	n1	valid	power	p10	p50	p90	if_p50
.1929708	.2082493	1.1	754	357	1000	.164	1.310	1.376	1.574	1.251
.1929708	.22296	1.2	754	357	1000	.304	1.308	1.401	1.603	1.167
.1929708	.2639855	1.5	754	357	1000	.865	1.342	1.543	1.826	1.029
.1929708	.3235131	2	754	357	1000	.998	1.660	1.999	2.422	0.999
.1929708	.4177034	3	754	357	1000	1	2.485	3.000	3.580	1.000
.3859417	.4087601	1.1	754	357	1000	.169	1.249	1.314	1.446	1.194
.3859417	.4299434	1.2	754	357	1000	.383	1.252	1.330	1.497	1.108
.3859417	.4852696	1.5	754	357	1000	.951	1.329	1.523	1.780	1.015
.3859417	.5569378	2	754	357	1000	1	1.694	2.009	2.359	1.004
.3859417	.6534431	3	754	357	1000	1	2.503	2.989	3.560	0.996
.7718833	.7882295	1.1	754	357	1000	.155	1.315	1.384	1.531	1.258
.7718833	.8023897	1.2	754	357	1000	.322	1.323	1.426	1.633	1.188
.7718833	.8354067	1.5	754	357	1000	.821	1.375	1.581	1.917	1.054
.7718833	.8712575	2	754	357	1000	.995	1.579	1.989	2.556	0.995
.7718833	.9103233	3	754	357	1000	1	2.350	3.053	3.928	1.018

Again, we see the output with respect matches that of Stata's -power twoproportion- command.

Finally, we can also see that the p10, p50, p90 and if_p50 values also match those produced by SAS and summarized in the table below:

SAS Simulation Results Illustrating Effect Size Magnification Given *True* Odds Ratios of 1.1, 1.2, 1.5, 2.0, and 3.0[a]							
True values		**N analyzed datasets**	**Power[b]**	**Distribution of Observed Significant ORs**			
Proportion of exposed individuals in control group	**OR**			**N**	**10th Percentile**	**Median**	**90th Percentile**
0.1929708 (1/2 background)	1.1	1000	0.153	153	1.31	1.38	1.55
	1.2	1000	0.302	302	1.32	1.41	1.62
	1.5	1000	0.842	842	1.34	1.54	1.83
	2	1000	1	1000	1.67	2.00	2.39
	3	1000	1	1000	2.53	3.04	3.66
0.3859417 (1x background)	1.1	1000	0.164	164	1.25	1.30	1.45
	1.2	1000	0.373	373	1.25	1.33	1.51
	1.5	1000	0.932	932	1.31	1.51	1.76
	2	1000	1	1000	1.71	2.01	2.38
	3	1000	1	1000	2.52	3.00	3.60
0.7718833 (2x background)	1.1	1000	0.149	149	1.31	1.39	1.58
	1.2	1000	0.312	312	1.32	1.42	1.64
	1.5	1000	0.803	803	1.37	1.56	1.93
	2	1000	0.992	992	1.61	2.01	2.54
	3	1000	1	1000	2.38	3.04	4.04

NOTE: The logistic regression model was used to compute the odds ratios for the two groups. The EXACT Test was used in the analysis of some datasets when the maximum likelihood estimate did not exist (perhaps due to a zero exposed in one of the groups).

[a] One-sided test, α = 0.05, N control=754, N cases = 357, N iterations = 1000 (datasets)

[b] The power resulting from this simulation may be close but not match exactly with the power calculated from built-in procedures in statistical software such as SAS (PROC POWER) or Stata (power twoproportion). This may be due to number of datasets simulated being of insufficient size. However, 1000 iterations is sufficient to adequately estimate the power and to illustrate the degree of effect size magnification given a statistically significant result (here alpha ≤ 0.05).