1

Guide to Interactive Simulations

In this document, I describe ways of using the interactive simulation programs. I consider each program separately, but a general approach for using the programs is described for the analysis of treatment effects on a mediator/outcome by testing group differences on means. Be sure to read it first. I show how to use the simulation to gain an appreciation for sampling error dynamics for the various links in RETs. I also show how to use the programs to conduct and gain perspectives on power analysis.

INTERACTIVE SIMULATION FOR TREATMENT EFFECTS: MEANS

The Basics of Evaluating Treatment Effects on Means

Consider the case of anorexic women who undergo a cognitive-behavioral therapy (CBT) program to help them gain weight. In an RET, I conceptualize a population of patients who have received treatment and a comparable population of patients who have not. Suppose the former population has a mean weight of 100 pounds (45.4kg) and the latter population has a mean weight of 90 pounds (40.8kg), so the treatment has a true effect of increasing weight by 10 pounds (4.5kg). Suppose also that the standard deviation for weight within each population is 15 pounds (6.8kg).

I conduct a study in which I randomly sample a total of 70 individuals, 35 from each population. The 35 individuals in the treatment group have been given CBT and the 35 individuals in the control group are a wait-list control. Suppose I find that that for this

sample of 70 individuals, the difference in mean weight was 8.4 pounds. This does not equal the true population difference of 10.0 and the disparity from 10.0 can be conceptualized as being due to sampling error, assuming all other aspects of the study were well implemented and controlled. To gain perspectives on the dynamics of such sampling error, I use the simulation program. I enter the population treatment mean as 100, the population control mean as 90, the within group standard deviation as 15, and the sample size as 70 in the Input box.

Upon execution of the generated R syntax, the program creates 10,000 "replications" of my study but each time selecting a different random sample of 70 cases. The program first reports a sub-group of 25 of the 10,000 studies. Here are the results:

	Treat mean	Ctrl mean	Difference	p Value
Study 01	98.7739	88.3323	10.4416	0.0038
Study 02	103.2865	89.7057	13.5807	0.0007
Study 03	100.5284	87.2374	13.2910	0.0008
Study 04	95.6223	87.2366	8.3857	0.0623
Study 05	101.6162	86.0007	15.6154	0.0002
Study 06	103.3272	89.7894	13.5377	0.0002
Study 07	100.9563	89.8447	11.1116	0.0015
Study 08	97.8659	87.5534	10.3125	0.0100
Study 09	102.0607	90.6470	11.4137	0.0031
Study 10	96.9584	92.1386	4.8198	0.1587
Study 11	102.0976	87.7208	14.3768	0.0003
Study 12	100.1510	90.8145	9.3365	0.0095
Study 13	99.1189	85.3180	13.8009	0.0008
Study 14	100.9913	90.5335	10.4578	0.0072
Study 15	101.0823	92.8728	8.2095	0.0427
Study 16	97.8462	88.8534	8.9929	0.0247
Study 17	99.1568	88.0959	11.0610	0.0059
Study 18	97.4825	88.6902	8.7923	0.0111
Study 19	99.4318	90.9628	8.4690	0.0192
Study 20	99.6610	93.0426	6.6184	0.0742
Study 21	99.8934	88.3110	11.5824	0.0032
Study 22	99.6057	91.1681	8.4376	0.0184
Study 23	98.7352	91.8702	6.8650	0.0814
Study 24	94.9329	85.5413	9.3916	0.0035
Study 25	101.3849	91.1361	10.2488	0.0130

My particular study could be any one of these 25 studies (perhaps Study 22). As I scan the results of the 25 studies, I see that each one produced a different result. Sometimes the mean difference in the sample/study was statistically significant and sometimes not. For one sample/study, the mean difference was as much as almost 16 pounds (Study 05) and for another, it was as little as 4.8 pounds (Study 10). These study-to-study (or sample-to-sample) fluctuations are bothersome and knowledge that they exist lead me to interpret the results for my single study with humility. In my study, I found a mean difference of 8.4

pounds ($p < 0.05$), but this is but one instantiation of the many possible results I could have obtained given the nature of sampling error. Note also that the difference of 8.4 pounds that I found in my study was statistically significant (Study 22) but the same difference was not statistically significant in another study (Study 04). This is because the estimated within group standard deviations differed in the two studies due to sampling error in them. Inspection of the 25 study results is sobering.

The true population Cohen's d was (100-90)/15 = 0.67 and the true proportion of explained variance due to the treatment condition was 0.10. The program also reports these statistics in each of the 25 studies. Here are the results:

	Cohen d	Eta sqr
Study 01	0.7214	0.1166
Study 02	0.8586	0.1575
Study 03	0.8486	0.1544
Study 04	0.4564	0.0502
Study 05	0.9493	0.1860
Study 06	0.9484	0.1858
Study 07	0.7942	0.1379
Study 08	0.6379	0.0935
Study 09	0.7395	0.1218
Study 10	0.3432	0.0290
Study 11	0.9094	0.1734
Study 12	0.6432	0.0950
Study 13	0.8481	0.1543
Study 14	0.6667	0.1013
Study 15	0.4974	0.0591
Study 16	0.5533	0.0720
Study 17	0.6839	0.1061
Study 18	0.6286	0.0911
Study 19	0.5774	0.0780
Study 20	0.4366	0.0461
Study 21	0.7362	0.1208
Study 22	0.5817	0.0790
Study 23	0.4259	0.0440
Study 24	0.7282	0.1185
Study 25	0.6146	0.0874

There also are sample-to-sample fluctuations in these statistics. For example, Cohen's d was 0.34 in one random sample (Study 10) but 0.95 in another sample (Study 05). The percent of explained variance was 2.9% in Study 10 but 18.6% in Study 5.

I address below ways of reducing such sample-to-sample fluctuations when designing studies, but let's first examine summary statistics for the results across all 10,000 studies. Here is the table the program reports:

```
Summary Statistics Across the 10000 Studies
              Average       SD       Min       Max       q10       q90
Mean diff     10.0653   3.5912   -3.6027   23.9101    5.4558   14.6175
Cohen d        0.6842   0.2515   -0.2405    1.8100    0.3669    1.0045
Eta sqr        0.1133   0.0670    0.0000    0.4538    0.0330    0.2038
SEmeandiff     3.5672   0.3080    2.4238    4.9276    3.1783    3.9652
```

Across all studies, the average mean difference between the treatment and control conditions was 10.07, which is quite close to the true population difference of 10.0. It turns out that had I conducted the simulation using many more random samples than 10,000, the average across the studies would have been 10.0. This is the essence of what we call an ***unbiased estimator***; across all possible random samples of a given size from a population (e.g., N=70), the average of the estimates equals the value of the true population parameter. In this case, the sample mean difference between the treatment and control conditions is an unbiased estimator of the true population mean difference. Note that when I state that an estimator is unbiased, I am not saying that the result from any one sample accurately describes the true population difference. It may or may not, as was evident when we examined the results for the 25 studies. When we speak of an unbiased estimator, we speak of a highly specialized concept, namely whether its average across all possible random samples of a given size equals the population parameter in question.

Note from the above table that one study (under `MIN`) found that people in the CBT condition gained, on average, *less* weight than people in the control condition (mean difference = -3.60), while in another study (under `MAX`) people in the CBT condition gained, on average, almost 24 pounds more than those in the control condition. Across the 10,000 studies, fully 10% of the studies observed mean differences less than 5.46 pounds (see the 10th quantile column, `q10`) while 10% of the studies observed mean differences greater than 14.6 pounds (see the 90th quantile column, `q90`). Appreciation of the existence of such arbitrary sample-to-sample fluctuations helps us keep the results for any one study in perspective.

A statistic in the table that is of particular interest is the standard deviation of the sample mean differences across the 10,000 studies. It indicates the "typical" disparity between a given sample result and the true population difference. In this case, it was 3.59 pounds. On average, sample mean differences were "off" by 3.59 pounds when estimating the true population mean difference of 10 pounds. This standard deviation has a special name in statistics. It is called the ***standard error of the difference between independent means***. An *estimate* of it is routinely reported on most computer output. Large standard errors suggest large sample-to-sample fluctuations for results and small standard errors suggest small sample-to-sample fluctuations. Indeed, a standard error of zero would indicate that every random sample yielded the same result, i.e., there is no sampling error.

Standard errors (or variants of them, such as margins of error) are our friends and should be embraced in reports of statistical analyses.[1]

In statistics, we often refer to the ***efficiency*** of an estimator. Informally, efficiency is the extent to which the estimator yields low standard errors or less sample-to-sample fluctuations than other estimators. We desire estimators that are unbiased and efficient.

As noted, when you conduct a study of mean differences, the t test or regression analysis will typically report an estimated standard error for the difference based on the sample data. However, the reported standard error is merely an estimate and it too will be subject to sampling error. In the last row in the above table, for the term labeled *SEmeandiff*, I provide descriptive statistics for the estimated standard error for the difference between the two means from each study across the 10,000 studies. The average standard error was 3.57. This value should be close to the true standard error, which, as noted above, was 3.59 (see the SD value in the row for `Mean diff`). If the average of the estimated standard errors is considerably smaller than the true standard error, this suggests the test likely will yield too many Type I errors.

Another statistic reported by the simulation program is the proportion of studies out of the 10,000 that found a statistically significant result ($p < 0.05$). In the current example, it was 0.79; 79% of the studies observed a statistically significant result whereas 21% of the studies failed to reject the null hypothesis of no group mean difference in weight gain. Given that the true population mean difference in weight is non-zero, the 0.79 statistic reflects the power of the statistical test when we use a sample size of 35 per group for these particular populations. Although this level of power approaches standards that many researchers seek (power of 0.80 or greater), it does not alter the fact that there are non-trivial sample-to-sample fluctuations at play, fluctuations that I would prefer to minimize in the bigger scheme of things. There is more to study design than having adequate statistical power.

A final piece of information generated by the simulation program is a density plot of the mean differences across the 10,000 studies (see my book for a description of density plots; they are like probability plots or histograms that show the fundamental shape of a distribution of scores). The generated plot is shown in Figure 1 with a normal distribution superimposed on it. The distribution is given a special name in statistics, the ***sampling distribution of the difference between two independent means***. Note that the distribution is roughly normally distributed. Had I used a very large number of replications instead of just 10,000, the approximation would be quite close. Statisticians make use of this fact to derive confidence intervals and p values for statistical tests by knowing the properties of

[1] Technically, standard errors usually are estimated based on mathematical derivations rather than simulations, but the results of the simulation here yields a reasonable proxy.

sampling distributions, such as if it is normally distributed, t distributed, or chi square distributed.

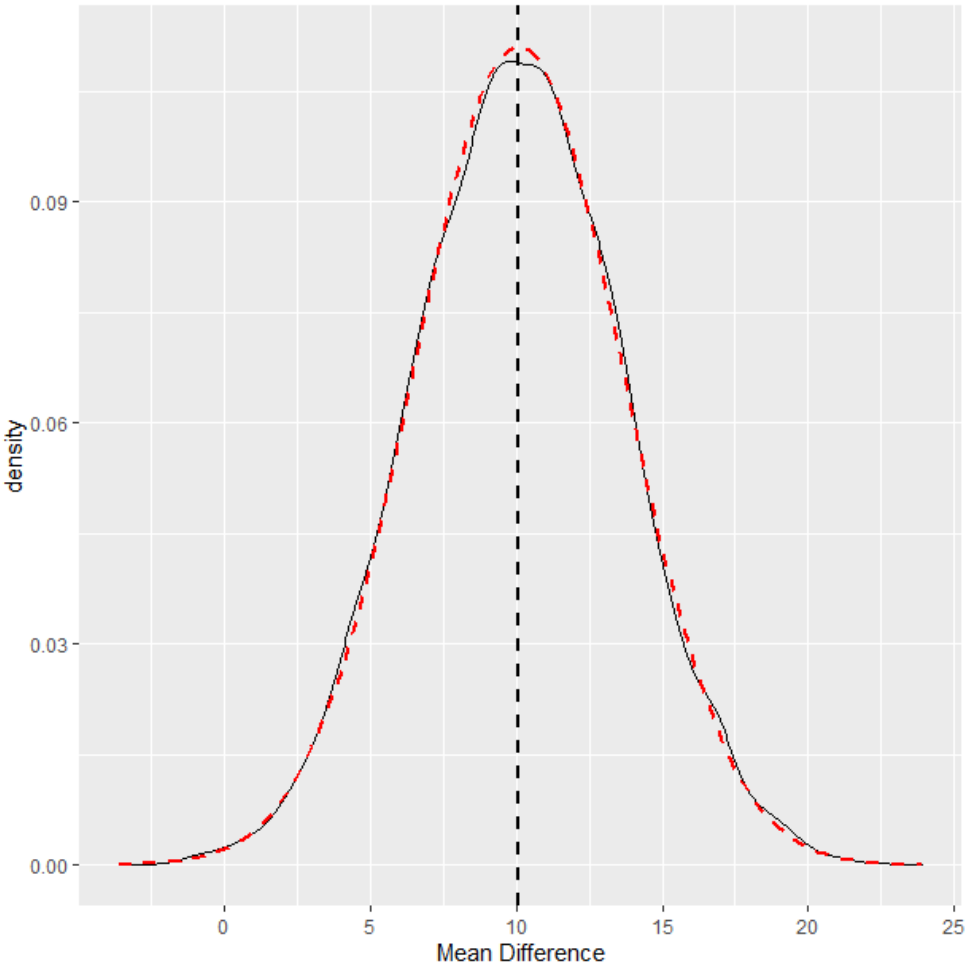

FIGURE 1. Sampling distribution of the difference between independent means

In sum, the program illustrates in concrete ways the important concepts of sampling error, standard errors, unbiased estimators, efficient estimators, statistical power and sampling distributions.

Type I Errors for Group Mean Comparisons

Suppose in the simulation program I specified the two populations as having identical population means, say both equal to 100 and both having a within group SD of 15. In this

case, the number of null hypothesis rejections across the 10,000 studies reflects Type I errors, i.e., the proportion of times the null hypothesis was incorrectly rejected. When I modeled this scenario in the program, the result was 0.049, which maps well onto an alpha level of 0.05 (the default alpha level in the simulation). This result is not surprising because the simulation was set up to satisfy the assumptions of the statistical procedure used to analyze the data. However, you can alter the input values where the statistical assumption of equal within group SDs is violated and examine the impact it has on Type I (and for that matter, Type II) errors. For example, I might set the means for the treatment and control groups to each be 100, the SD for the treatment group to be 10, the SD for the control group to be 15, and the sample sizes to be 35 in each group. The analysis would then provide perspectives on the effects of variance heterogeneity on the Type I error rate. When I did so, the proportion of rejections of the null hypothesis was 0.052. Violation of the assumption did not make much difference in this case. In my book, I advocate for the use of robust estimators that do not make homogeneity of variance assumptions.

Reducing Sampling Error (Sample-to-Sample Fluctuations)

One way to reduce sampling error and the resultant sample-to sample fluctuations that occur is to increase one's sample size. Suppose instead of a sample size of 35 per group, I quadruple it and use a sample size of 140 per group, or a total N of 280. I re-ran the simulation using this sample size and the effects of increasing the N were notable. Here are the side-by-side values of the 25 randomly selected studies in the two simulations:

	n=35 per grp Difference	n=140 per grp Difference
Study 01	12.3632	12.5144
Study 02	13.8309	9.8685
Study 03	15.8767	9.6031
Study 04	8.7051	12.6547
Study 05	13.0683	9.0801
Study 06	8.5263	12.3443
Study 07	12.6382	9.0227
Study 08	9.3578	10.8584
Study 09	3.0968	8.8043
Study 10	8.6984	10.0666
Study 11	12.7581	6.9597
Study 12	8.6774	10.0840
Study 13	4.1755	13.3679
Study 14	15.1318	11.7767
Study 15	5.1337	9.7332
Study 16	10.1320	9.0430
Study 17	9.9999	6.6325
Study 18	8.1195	12.6563

```
Study 19      14.6196            9.9760
Study 20       9.9632           12.7057
Study 21       4.9350           13.1071
Study 22       9.5551           10.1379
Study 23      10.2517            8.5934
Study 24       8.3547            6.8240
Study 25       6.5791           10.9827
```

If you scan the two columns of numbers, you see that the numbers in the second column tend to be closer to one another than the numbers in the first column; there is less sample-to-sample variability in them. For the n=35 per group simulation, the minimum and maximum mean differences for the 10,000 studies were -3.60 and 23.91, respectively; for the n=140 per group simulation, they were 3.58 and 16.55. For the n=35 per group simulation, the 10^{th} and 90^{th} quantiles were 5.46 and 14.62; for the n=140 per group simulation, they were 7.70 and 12.31. The standard error for the mean difference for the n=35 per group simulation was 3.59; for the n=140 per group simulation, it was 1.79, a reduction of about 50%. Clearly, we are better served by the larger sample size.

There is another way of reducing sampling error other than increasing sample size that some researchers overlook in RETs. The strategy is particularly useful for research in which it is difficult to increase N due to cost or practical constraints. I can illustrate the general principle using a simplistic example. Consider the case of two populations each with 5 observations. Here are the scores in each of them:

Population A	Population B
2	4
3	4
4	4
5	4
6	4

The mean in each population is 4, but the two populations obviously differ in the variability of scores. Suppose I do not know the value of the population means and I am told I can randomly sample two cases from each population to estimate the mean. In Population A, I might end up sampling the scores 4 and 6 when I select the two observations randomly, and the average of them is 5. Absent any other information, 5 is my best guess about the population mean and, as it turns out, I am "off" by 1 unit because of sampling error. For Population B, I randomly select the scores of, say, the third and fifth persons, which are 4 and 4 and the mean is 4. There is no sampling error. Note that because there is no variability in the scores in Population B, it does not matter which two cases I happen to sample because

they all equal 4 and, when averaged, will yield the value of the Population mean. In Population A, where the scores are quite variable, there are many combinations of two that I could sample, some of which will yield means that are quite discrepant from the population mean. The more variability there is in scores in the population, the more sampling error there will be in sample means, everything else being equal. This is a core principle in sampling theory.

Using a more realistic example, suppose my RET is focused on reducing depression. There are well documented differences in depression as a function of biological sex; females tend to have higher levels of depression than males. In an RET, the within group standard deviation of depression at posttest reflects, in part, these gender differences. The treatment group has both males and females in it and this fact creates variability in depression scores within that group, i.e., it increases the withing-group SD. Similarly, the control group has both males and females in it and this also creates variability of depression scores within that group. If I measure biological sex and statistically control for it in my statistical analysis, the result will be less within cell variability in depression and, in turn, less sampling error. In short, I can reduce sampling error not only by increasing sample size but also by including strategically chosen covariates in the analysis. See my book for elaboration of this strategy.

For the anorexia RET, the within group standard deviation for weight was 15 pounds. One viable candidate for use as a covariate is the baseline outcome measure, namely baseline weight. The baseline measure of weight will be uncorrelated with the treatment condition (given random assignment) yet it likely is strongly related to the posttest weight. Indeed, it is not uncommon for baseline and posttest measures of a construct to be correlated between 0.50 and 0.70. Suppose I re-run the simulation with the original N = 70 (n = 35 per group), but now I use the baseline as a covariate. I might find that the posttest within cell variability of weight is cut in half, from 15 to 7.5. Here is the summary table for the results across the 10,000 “studies”:

```
Summary Statistics Across the 10000 Studies
              Average       SD       Min       Max       q10       q90
Mean diff     10.0327   1.7956    3.1987   16.9550    7.7279   12.3087
Cohen d        1.3643   0.2709    0.4271    2.6331    1.0249    1.7135
Eta sqr        0.3185   0.0837    0.0442    0.6375    0.2104    0.4268
SEmeandiff     1.7836   0.1540    1.2119    2.4638    1.5891    1.9826
```

The results are about the same as the case where I quadrupled the sample size to an n of 140 per group, but the “cost” of measuring and covarying out the baseline outcome is likely much less compared to that of quadrupling the sample size. The estimated power of the baseline adjusted analysis in the new simulation was greater than 0.99 as compared with

0.79 in the original simulation. There obviously is something to be said for identifying baseline variables that cause variability in the outcome and then statistically controlling for them to increase power and reduce sampling error. You can use the simulation program to explore the consequences of varying sample size and covariate inclusion to reduce within cell SD, a topic I discuss in more depth in my book.

A third way of reducing sampling error is by experimental design, such as by restricting the study to a homogeneous population (e.g., only females) and by ensuring uniform implementation of study and treatment protocols. The more heterogeneous the population and the more subtle variability there is in the implementation of study protocols, the more unintended variability in mediators and outcomes there can be.

Power Analysis for Group Mean Comparisons

You can use the interactive simulation program to conduct power analyses when planning studies and to explore the consequences of different sample size decisions. To conduct a power analysis, you must specify the minimum effect size (MIES) that you want to be sure to detect because it is deemed as being clinically or substantively important. In studies of mean differences, this is often expressed in terms of Cohen's d and a common MIES is a population *d* of 0.50. In the simulation program, you can mimic this case by setting the population treatment mean to 0.50, the control group mean to 0.00, and the within group SD to 1.0. This produces a population *d* of (0.50–0.00)/1 = 0.50, the results of which will generalize to any combination of numbers that produce a *d* of 0.50, such as (100-90)/20. I can evaluate approximate statistical power for any given sample size by specifying the sample size of interest under this scenario. As an example, I ran the simulation for a *d*=0.50 using n = 65 per group. The power estimate from the simulation program was 0.81.

Some researchers allocate a smaller number of cases to the control group to allow for a larger sample size in the treatment condition for additional within-group analyses on that group. I repeated the above power analysis but with 100 individuals in the treatment group and 30 individuals in the control group (note that the overall N remains the same at 130). The statistical power to detect a population *d* of 0.50 was reduced from 0.81 to 0.68. It turns out that the group with the smaller sample size lowers power because of the increased sampling error associated with it's mean; this decrease is not offset by the increased precision for the group with the larger N. There are trade-offs of non-1:1 allocations.

INTERACTIVE SIMULATION FOR TREATMENT EFFECTS: PERCENTS

The Basics of Evaluating Treatment Effects on Percents

Consider the case of a program that encourages people to obtain a new vaccination against

a deadly virus. In an RET, I conceptualize a population of individuals who have been exposed to the program and a comparable population of individuals who have not. Suppose the former population has a vaccination rate of 50% and the latter population has vaccination rate of 40%, so the treatment has a true effect of increasing the percent of people obtaining the vaccination by 10%.

I conduct a study in which I randomly sample a total of 250 individuals, 125 from each population. The 125 individuals in the treatment group have been given the program and the 125 individuals in the control group have not. Suppose I find that that for this sample of 250 individuals, the percent difference was 12.8%. This does not equal the true population difference of 10.0% and the disparity from it can be conceptualized as being due to sampling error, assuming all aspects of the study were well implemented and controlled. To use the simulation program, I enter the population treatment percent as 50% and the population control percent as 40%. I enter samples sizes of 125 per group.

After executing the generated R syntax, the program conducts 10,000 "replications" of my study but each time it selects a different random sample of 125 cases from each population. As with the prior simulation, the program first reports a sub-group of 25 of the 10,000 studies. Here are the results:

	Treat pct	Ctrl pct	Difference	p Value
Study 01	44.0	32.8	11.2	0.0680
Study 02	46.4	32.0	14.4	0.0192
Study 03	51.2	44.8	6.4	0.3112
Study 04	51.2	46.4	4.8	0.4480
Study 05	46.4	38.4	8.0	0.2003
Study 06	40.0	34.4	5.6	0.3598
Study 07	55.2	40.0	15.2	0.0156
Study 08	44.8	45.6	-0.8	0.8990
Study 09	53.6	36.8	16.8	0.0073
Study 10	51.2	40.8	10.4	0.0984
Study 11	50.4	43.2	7.2	0.2538
Study 12	51.2	34.4	16.8	0.0069
Study 13	61.6	40.0	21.6	0.0006
Study 14	54.4	33.6	20.8	0.0008
Study 15	52.8	36.0	16.8	0.0072
Study 16	46.4	44.8	1.6	0.7997
Study 17	51.2	40.8	10.4	0.0984
Study 18	45.6	34.4	11.2	0.0701
Study 19	49.6	36.8	12.8	0.0404
Study 20	52.8	41.6	11.2	0.0755
Study 21	51.2	37.6	13.6	0.0299
Study 22	56.8	42.4	14.4	0.0222
Study 23	52.0	40.0	12.0	0.0563
Study 24	60.8	32.8	28.0	0.0000
Study 25	47.2	36.0	11.2	0.0718

My particular study could be any one of these 25 studies (perhaps Study 19). Scanning the results of the 25 studies, you can see each one produced a different result. Sometimes the percent difference for the sample/study was statistically significant and sometimes not. For one sample/study, the percent difference was 28% (Study 24) and for another study, it was negative, favoring the control condition over the treatment condition (Study 8). These study-to-study (or sample-to-sample) fluctuations are disconcerting and lead me to interpret the results for my one study cautiously. In my study, I found a percent difference of 12.8% ($p < 0.05$), but this is but one instantiation of the many possible results I could have obtained given the nature of sampling error.

Here is the table of summary statistics for the percent differences across all 10,000 studies:

```
Summary Statistics Across the 10000 Studies
                Average      SD      Min      Max      q10      q90
Percent diff     9.9942  6.2672 -14.4000  32.0000   1.6000  18.4000
SEdiff           6.2359  0.0612   5.8709   6.3244   6.1512   6.3041
```

Across all studies, the average percent difference was 9.99%, which is quite close to the true population difference of 10.0%. This is because the sample percent difference is an unbiased estimator of the population percent difference. Across the 10,000 studies, one study found that people exposed to the program were *less* likely to obtain a vaccination by 14.4% relative to individuals who did not receive the program (see `MIN`), while in another study (under `MAX`) 32.0% more people exposed to the program obtained the vaccination than people who were not exposed to the program. 10% of the studies observed percent differences less than 1.6% (see the 10th quantile column, `q10`) while 10% of the studies observed percent differences greater than 18.4% (see the 90th quantile column, `q90`). The standard error of the percent difference was 6.26, suggesting that the "typical" disparity between a sample result and the true population difference was 6.26%. The mean of the standard errors reported in each study (6.24) was close to the true standard error of 6.27.

When I examined on the output the proportion of studies out of the 10,000 that found a statistically significant result ($p < 0.05$), it was 0.35. This is an estimate of the statistical power of the test, and it was bleak. Despite a 10% true population percent difference, only 35% of the samples yielded a statistically significant result. The same level of statistical power (more or less) would be evident in logit or probit regression.

Finally, the plot of the sampling distribution showed a pattern that was roughly normal (see Figure 2).

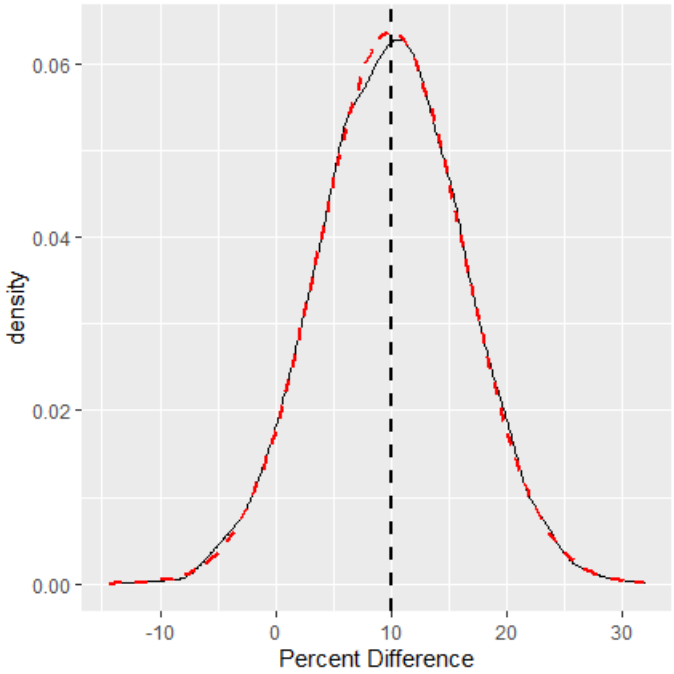

FIGURE 2. Sampling distribution of the difference between independent percents

Some Properties of Power for Percent Differences

There are interesting properties of analyzing percent differences relative to sampling error and statistical power. The outcome variable is essentially a set of 0s and 1s. If the percent of 1s is 50%, there is much more variability in the scores than if the percent of 1s is 90%, in which case most of the scores are 1s. Similarly, if the percent of 1s is 10%, most of the scores are 0 and there again is little variability. With larger variability in scores, there will be more sampling error per my discussion of the role of population variability for mean differences. It follows that sampling error will be less as percentages move away from 50% and towards the extremes. To illustrate this, I used the simulation program and specified a 10% difference between the treatment and control conditions but instead of using 50% and 40% to define the population percents, I used 20% and 10%. The statistical power went from 0.35 in the original simulation to 0.62. The standard error of the percent difference went from 6.27% to 4.51%. The further the percents are from 50%, the less sampling error you have and the greater will be the statistical power, everything else being equal. When designing an RET that will analyze percent differences, your guesses about the level of the percents involved can be crucial for making sample size decisions.

Type I Errors for Group Percent Differences

You also can use the simulation program to explore Type I errors by setting the two population percents to equal values. For example, when I set the percents to 50% in the treatment and control populations and used a sample size of 125 per group, the proportion of null hypothesis rejections was 0.05. This equals the theoretical 0.05 alpha level. Parenthetically, for logit, probit, and MLPM frameworks, there can be slight to moderate inflations of Type I errors with smaller sample sizes. For example, when I used a sample size of 40 per group and set both groups percent levels at 50%, the proportion of null hypothesis rejections was 0.059. See my book for elaboration.

Reducing Sampling Error (Sample-to-Sample Fluctuations)

As with the analysis of means, the primary methods for reducing sampling error for the analysis of percents are (1) increase sample size, (2) incorporate strategically selected covariates that impact the binary outcome but that are unrelated to treatment condition, and (3) restrict the study to a homogeneous population and ensure uniform implementation of study and treatment protocols. The use of covariates is straightforward for the MLPM approach but not so for logit and probit modeling. See my book for details. A third way of reducing sampling error is by experimental design and implementation uniformity.

Sample Size Demands When Analyzing Percents

As you work with the percent difference simulation program, you will discover that comparing percentages between groups is much more sample size demanding than comparing means. For example, to achieve power of 0.80 to detect a population difference of 10% for a two tailed alpha level of 0.05 where the population percents are 30% and 20%, I need a sample size of about 300 per group. By contrast, to detect a mean difference corresponding to a Cohen's *d* of 0.50 with power of 0.80, the required sample size is about 65 per group. In some disciplines, researchers are quick to dichotomize continuous outcome variables because they find it easier to interpret percents or because it is a way of dealing with non-linearity. As I discuss in Chapter 3 of my book, unless there is strong theoretical or substantive justification for such dichotomization, it usually is not a good practice to pursue. Dramatic loss in statistical power is but one reason not to do so.

INTERACTIVE SIMULATION FOR MEDIATOR EFFECTS ON OUTCOMES

The Basics of Evaluating Mediator Effects on Outcomes

Consider an RET where I have three continuous mediators that are presumed to

independently impact a continuous outcome. For example, the outcome might be the strength of a person's intention to support policies favorable to reducing climate change (Y) and the mediators are (a) beliefs in the long term negative consequences of climate change (M1), (b) beliefs in the short term negative consequences of climate change (M2), and (c) beliefs that it is economically feasible to enact the policies (M3). The classic analysis of the effects of the mediators on policy support regress Y onto M1, M2, and M3 and then evaluate the statistical significance and magnitude of the regression coefficients associated with each mediator. This might take the form of traditional OLS regression, a form of maximum likelihood regression in an SEM context, or robust regression, such as MM regression or quantile regression.

In the mediation effect simulation program, I apply OLS multiple regression with the idea that the sampling error dynamics observed in that context likely generalize to other analytic contexts. When setting up the structure of the program, I decided to have you input the population correlations between Y, M1, M2, and M3 to give you control over predictor redundancy via the magnitude of the correlations between mediators, as well as how strongly each mediator is related to the outcome. From this information, I create a population covariance matrix in which I assume the variances of Y, M1, M2, and M3 are each 1.0. As such, the regression coefficients that derive from the input correlations can be interpreted much like standardized regression coefficients although, technically, they are not standardized. This is because I ultimately treat the input matrix as a covariance matrix. Setting up the program in this way puts each of the mediators on a common metric, making comparisons between their regression coefficients more straightforward. The mean structure of the four variables does not impact the values of the regression coefficients, which are of primary interest. I also create the population variables so that they are multivariately normally distributed.

Standards for interpreting standardized-like coefficients vary and, as discussed in my book, can be fraught with difficulties. However, I structured input into the simulation program so that you know the magnitude of the population correlation for each mediator with the outcome. This can help put results in context, as I illustrate with the worked example below. Some researchers argue that standardized-like coefficients near or greater than 0.20 are generally meaningful, but this standard is subject to controversy. Given the input variables all have a population standard deviation of 1, you can use this standard as a rough guide. A coefficient of 0.20 means if M increases by 1 unit in the population (i.e., 1 SD), the mean of Y increases by a fifth of a standard deviation (i.e., 0.20).

Literature reviews have found that a typical correlation between a wide range of social psychological variables is about 0.35. For the worked example, I set the population correlations among all of the variables (Y, M1, M2 and M3) to 0.35. In this sense, all of

the mediators are equally important in predicting the outcome because each is correlated 0.35 with it. As well, each mediator shares some common variance with the others.

After executing the generated R syntax, the program begins by providing the population equation that results from the input correlations. Here is the equation:

```
y = 0 + 0.2059 m1 + 0.2059 m2 + 0.2059 m3    ; R square = 0.2162
```

Each mediator has the same regression coefficient, namely 0.2059. For every one unit that a given mediator increases, the mean Y is predicted to increase by 0.2059 units, holding constant the other mediators. The population squared multiple correlation is 0.2162. It is against these values that the program examine sampling error dynamics.

Suppose I conduct an RET with reference to the above population with a sample size of 150 and I obtain the following regression results: b for M1 = is 0.23, b for M2 = 0.15, b for M3 = 0.24, with the squared R being 0.23. The p values for the coefficients for M1 and M3 turned out to be statistically significant ($p < 0.05$) but not for M2. My results are different than the true population coefficients, a fact that is attributable to sampling error, assuming the study is well designed and implemented. The simulation program reports 10,000 "replications" of my study but for each study, it selects a different random sample of 150 cases. As with the prior simulations, the program first reports a sub-group of 25 of the 10,000 studies. Here are the results for the three regression coefficients:

	b1	b2	b3	pval b1	pval b2	pval b3
Study 01	0.2520	0.1922	0.2606	0.0011	0.0062	0.0003
Study 02	0.2500	0.0999	0.2899	0.0052	0.2737	0.0020
Study 03	0.2528	0.1502	0.1389	0.0031	0.0468	0.1115
Study 04	0.3539	0.2545	0.0210	0.0000	0.0002	0.7754
Study 05	0.2228	0.1540	0.2275	0.0230	0.0619	0.0050
Study 06	0.1537	0.1039	0.3510	0.0617	0.2365	0.0000
Study 07	0.1976	0.2468	0.2294	0.0153	0.0018	0.0078
Study 08	0.3092	0.0594	0.2120	0.0000	0.4235	0.0068
Study 09	0.1962	0.1409	0.2818	0.0117	0.0909	0.0019
Study 10	0.2882	0.1781	0.1673	0.0011	0.0280	0.0281
Study 11	0.2079	0.3245	0.1397	0.0118	0.0000	0.0703
Study 12	0.1180	0.0186	0.3008	0.1045	0.8161	0.0001
Study 13	0.2136	0.1272	0.3607	0.0028	0.0938	0.0000
Study 14	0.2157	0.1441	0.1663	0.0090	0.0868	0.0454
Study 15	0.1478	0.2812	0.1229	0.0379	0.0000	0.1225
Study 16	0.2334	0.1499	0.2426	0.0022	0.0644	0.0031
Study 17	0.2018	0.1782	0.1873	0.0099	0.0127	0.0218
Study 18	0.1402	0.2816	0.2789	0.0972	0.0005	0.0013
Study 19	0.0872	0.2189	0.4198	0.2534	0.0027	0.0000

Study 20	0.0706	0.2229	0.3429	0.3510	0.0042	0.0000
Study 21	0.1827	0.0985	0.3100	0.0317	0.2890	0.0010
Study 22	0.1209	0.3512	0.2377	0.1499	0.0001	0.0033
Study 23	0.0440	0.2874	0.1705	0.5894	0.0001	0.0310
Study 24	0.2227	0.2800	0.1990	0.0048	0.0001	0.0106
Study 25	0.2713	0.0244	0.2374	0.0010	0.7688	0.0045

My particular study could have been be any one of these 25 studies. Scanning the results of the 25 studies, you can see that each one produced a different result. Examining column 1 for b1, one study found the coefficient to be as low as 0.04 and statistically non-significant (Study 23) and another study found it to be as high as 0.35 (Study 04). Such sample-to-sample fluctuations also are evident for the two other regression coefficients. These study-to-study (or sample-to-sample) fluctuations are disconcerting and, as before, lead me to interpret the results for my one study cautiously. Note also that some of the studies found that all three coefficients were significant but other studies found that only two of them were statistically significant. Still others found that only one of the coefficients was significant. This is despite the fact that all three of the mediators in the population are relevant and each is equally important. Such are the evils of sampling error.

Here are the results for the 25 studies for the squared multiple correlations and the adjusted squared multiple correlations. The latter applies an adjustment to the squared multiple correlation because the squared multiple correlation is a positively biased estimator of the true population squared multiple correlation:

	R sqr	Adj R sqr
Study 01	0.3146	0.3005
Study 02	0.1762	0.1593
Study 03	0.1603	0.1431
Study 04	0.3167	0.3026
Study 05	0.1876	0.1709
Study 06	0.2197	0.2037
Study 07	0.2234	0.2074
Study 08	0.2637	0.2486
Study 09	0.2638	0.2487
Study 10	0.2763	0.2614
Study 11	0.2454	0.2299
Study 12	0.1690	0.1519
Study 13	0.3367	0.3230
Study 14	0.1607	0.1434
Study 15	0.2420	0.2264
Study 16	0.2309	0.2151
Study 17	0.1950	0.1784
Study 18	0.2717	0.2568

```
Study 19   0.3670       0.3540
Study 20   0.2822       0.2674
Study 21   0.1965       0.1800
Study 22   0.2552       0.2399
Study 23   0.1910       0.1743
Study 24   0.3111       0.2969
Study 25   0.1879       0.1712
```

These statistics also show considerable sample-to-sample fluctuations. In one study, R^2 was 0.16 (Study 03), representing 16% explained variance, whereas in another study it was 0.37 (Study 19), representing 39% explained variance.

Here is the table of summary statistics for each parameter across all 10,000 studies:

```
               Average      SD      Min      Max        q10      q90
b1              0.2061  0.0811  -0.1304   0.5404 0.10190000   0.3087
b2              0.2045  0.0811  -0.1364   0.5356 0.10050000   0.3091
b3              0.2060  0.0817  -0.1414   0.5415 0.10187526   0.3105
R square        0.2297  0.0593   0.0528   0.4789 0.15370000   0.3069
Radj square     0.2138  0.0605   0.0334   0.4682 0.13630000   0.2927
SEb1            0.0809  0.0067   0.0597   0.1108 0.07250000   0.0896
SEb2            0.0809  0.0067   0.0572   0.1102 0.07260000   0.0897
SEb3            0.0809  0.0068   0.0585   0.1103 0.07252812   0.0898
```

Note that for all of the statistics except the R square, the average of the estimates across the 10,000 replication studies are quite close to values of their true population counterparts. This is because they are unbiased estimators in the technical sense of the term. Examination of the minimum and maximum values across the 10,000 studies as well as the 10^{th} and 90^{th} quantiles gives a sense of the fluctuations in sample results. The standard errors for each of the regression coefficients (the first three rows in the `SD` column) indicate that the "typical" disparity between the sample estimate of the coefficient and the true population coefficient value was about 0.081. The means of the standard errors for each coefficient (see the last three rows of the table) are close in value to the standard errors.

When I examined on the output the proportion of studies out of the 10,000 that found a statistically significant result ($p < 0.05$) for each coefficient, here is what I found:

```
Analysis of Null Hypothesis Rejections Across the 2000 Studies
   Proportion of nulls rejected
b1                       0.7183
b2                       0.7089
b3                       0.7227
```

These represent power estimates and all were about 0.71.

Finally, the plot of the sampling distribution for each of the regression coefficients across the 10,000 studies showed a pattern that was roughly normal (see Figure 3).

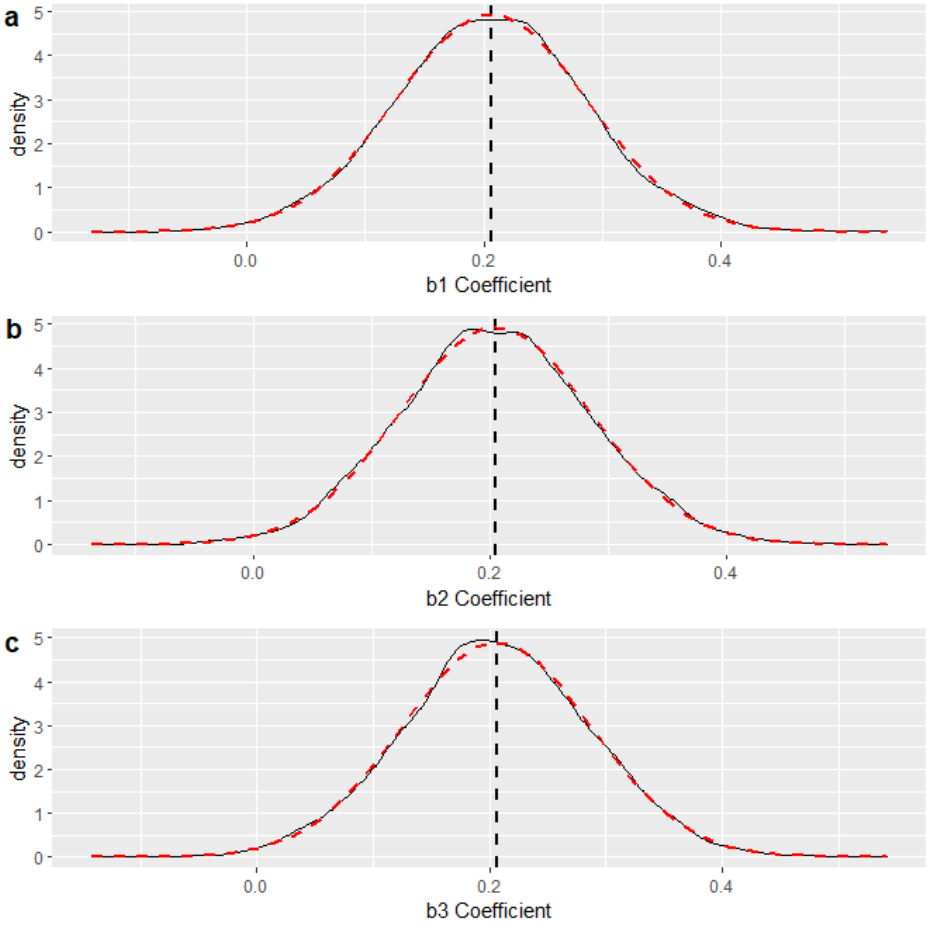

FIGURE 3. Sampling distribution of the three regression coefficients

Multivariate Patterns of Statistical Significance

Although the statistical power for a given coefficient was near 0.71, a result that I am particularly interested in is how often all three mediators were statistically significant given that they all have the exact same non-zero population coefficient. How often does a given study conclude that this is the case in terms of statistical significance? The simulation program performs a pattern analysis that provides perspectives on this question. Here is the relevant table of results:

```
                Proportion
1-1-1               0.3069
1-1-0               0.1769
1-0-1               0.1859
1-0-0               0.0486
0-1-1               0.1793
0-1-0               0.0458
0-0-1               0.0506
0-0-0               0.0060
all three sig       0.3069
only two sig        0.5421
only one sig        0.1450
```

The first column of the table lists the different significance patterns, with a 0 indicating a statically nonsignificant result and a 1 indicating a statistically significant result. The first digit is for b1, the second digit is for b2, and the third digit is for b3. For example, the 1-1-1 pattern refers to the case where all three coefficients were statistically significant. The pattern 1-1-0 refers to the case where b1 and b2 were statistically significant but not b3. The entry in the second column is the proportion of times across the 10,000 studies that the pattern occurred. The last three rows combine some of the patterns and shows (a) the proportion of times all three coefficients were statistically significant, (b) the proportion of times two and only two of the coefficients were statistically significant, and (c) the proportion of times one and only one of the coefficients was statistically significant.

In the worked example, despite the fact that each coefficient had statistical power of 0.71, in only 31% of the replication studies were all three coefficients statistically significant. The statistical power for the multivariate pattern of 1-1-1 was only 0.31. What sample size would I need to bring this power for the 1-1-1 pattern up to 0.80? With some trial and error using the simulation program, I found it to be approximately 275.

The required sample size for multivariate patterns of coefficient magnitudes will differ depending on the intercorrelations among the predictors and the uniformity and magnitude of the correlations between the mediators and the outcome. You can explore these dynamics using the simulation program.

Type I Errors for the Regression Coefficients

To explore Type I errors with the simulation program, you can set the correlation between a target mediator and the outcome to zero. This will translate into a zero regression coefficient for the mediator in the population, unless a suppressor dynamic is induced. Suppression occurs when a predictor has a zero or near zero correlation with the outcome but a moderate to large correlation with one or more of the other predictors (see Conger &

Jackson, 1972). In such cases, the coefficient for the suppressor variable can be large despite its zero correlation with the outcome. You will be able to identify if suppression occurs by examining the population coefficients for the population model that are shown on the program output. If the mediator whose zero order correlation coefficient with the outcome you set to zero has a non-zero coefficient, then suppression likely occurred.

Reducing Sampling Error (Sample-to-Sample Fluctuations)

The same three methods for reducing sampling error described earlier apply to the analysis of mediator effects on outcomes. These include (1) increasing sample size, (2) incorporating strategically selected covariates that increase the squared multiple correlation of the outcome, and (3) restricting the study to homogenous populations and ensuring uniformity of study protocol implementation. You can easily explore the impact of increasing sample size on sampling error in the simulation program. You can explore the role of covariates in the simulation program by conceptualizing the study as having only two mediators (M1 and M2) and then treating M3 as if it were a covariate that has specified correlations with the outcomes and the two mediators (i.e., mentalize the M3 label in the program as having the label COV). The coefficients and statistical power for the two mediators will be impacted by their correlation with COV as well as the correlation of COV with the outcome. For example, if you set the correlation between the two mediators and the "covariate" to 0 and you set the correlation between the covariate and the outcome to 0.35, you will see the power of the regression coefficients increase and their sample-to-sample fluctuations decrease.

The Use of Adjusted versus Unadjusted R Squares

As noted, the sample R^2 is a positively biased estimator of the true population R^2 in that, on average, it tends to overestimate the population R^2. As a result, statisticians have suggested a correction factor or adjustment to the sample R^2 to yield an unbiased estimator. In the worked example, the population R^2 was .2162. Across the 10,000 replication studies, the average sample R^2 was 0.23 and the average adjusted R^2 was 0.21. These values reflect the relative bias in the squared R and the unbiasedness in the adjusted R squared. Despite this, the use of the adjusted R squared is controversial.

First, the amount of bias in R^2 tends to be less with larger sample sizes, larger population R squares, and smaller numbers of predictors. When I repeated the simulation using a sample size of 500, the average sample R^2 across the 10,000 replications was 0.2196 as compared to the average adjusted R2 of 0.2149. The difference between these values is trivial. More importantly, when sample sizes are smaller, the correction suggested by statisticians sometimes results in negative squared multiple correlations, a result that is

nonsensical. For example, when I re-ran the worked example but with a sample size of 50, the minimum adjusted R^2 reported on the output was -0.05. If I adopt the practice of setting negative adjusted R^2s to zero, then the adjusted R^2 is no longer unbiased; it becomes a positively biased estimator. Finally, cases occur where the standard error for R^2 is lower than the standard error for the adjusted R^2. This means that although the adjusted R square provides an estimator that is less biased, there is a trade-off because it also is less efficient, i.e., it can exhibit larger sample-to-sample fluctuations than the unadjusted R^2. The simulation program can be used to explore such dynamics.

Additional Explorations

The mediator effect simulation program can be used to explore a wide range of issues, only some of which I have illustrated here. These include the effects of predictor collinearity, differential correlation structures of mediators with outcomes, multivariate power analysis, the role of covariates, and the use of statistical corrections for bias.

INTERACTIVE SIMULATION FOR OMNIBUS MEDIATION

The Basics of Evaluating Omnibus Mediation

This simulation evaluates the sampling error for the product of coefficients in a two-link mediational chain in an RET, from treatment to mediator (p1) and from the mediator to the outcome (p2). It also evaluates the Sobel test of this coefficient product (see my book for elaboration of the test). There are many versions of the Sobel test. The current program uses the normal theory version, which also is used in the Hayes (2018) PROCESS framework for mediation analysis. I focus on the case of a continuous mediator (M) and a continuous outcome (Y). Note that I do not recommend the general use of the normal theory Sobel test; it works well in some cases (bootstrapped based approaches usually work better). The analytic strategy I use in the program is to conduct two separate OLS regressions, one to isolate p1 and the other to isolate p2 and then I invoke the Sobel test using the standard errors for p1 and p2. I assume equal n in the two groups defining T. Independent of the test, the simulation program documents and provides you with an appreciation of sampling error when evaluating p1*p2 products and this is my primary focus; to help you grasp sampling error dynamics.

In addition to the Sobel test, I apply the joint significance test in the simulation. This test declares mediation if both p1 and p2 are jointly significant. If either p1 or p2 is not significant, the mediation chain is "broken" and mediation is called into question.

Finally, I provide information on p3, the direct effect of T on Y independent of the mediator. I create the scenario where the population value of p3 is zero.

The program requires you to provide population values of the two path coefficients. To make interpretation easier, I scale M and Y to have overall means of zero and standard deviations of 1.0. You first enter the population effect (p1) of the dummy variable for the treatment versus control conditions, T, on the mediator. The value of p1 is the population mean M for the treatment condition minus the population mean M for the control condition. If p1 equals 0.50, this indicates that the mean on M for the treatment group minus the mean on M for the control group is 0.50, or about half a standard deviation of M. You can think of p1 as roughly analogous to a Cohen's d but where the standardizer is the SD of M rather than the pooled within group SD of M. You will not want to exceed a value of 1.75 for p1, which represents an extremely strong effect (the mean difference is one and 3/4 SDs of M).

. Next, you enter the population effect of M on Y, p2. Because I set the mean and variance of both M and Y to 0 and 1.0, p2 is simply the correlation between the mediator and the outcome. The square of this correlation is the proportion of variation in Y that M accounts for in Y. If you enter p2=0.50, this implies that M accounts for 25% of the variance in Y. Technically, p2 is interpreted as a regression coefficient as it is not formally standardized - it just so happens that both M and Y have population standard deviations of 1.0. If p2 equals 0.20, this indicates that for every one unit M increases (which represents one standard deviation on M), the mean of Y is predicted to increase by 0.20 Y units (which is a fifth of a standard deviation of Y).

Finally, you enter the total sample size in the study. I assume half the N is in the treatment group and half in the control group. As an example, for the simulation program, I set the p1 = 0.30, p2 = 0.40 and I use a total sample size of 130 (65 per group). The overall population omnibus effect in the population is p1*p2 = (0.30)(0.40) = 0.12. This coefficient reflects the Y mean difference between the treatment and control groups through this particular mediational chain.

After executing the generated R syntax, the program conducts 10,000 “replications” of the study but each time it selects a different random sample of 130 cases from the population. As with the prior simulation, the program first reports a sub-group of 25 of the 10,000 studies. Here are the results:

```
25 Example Studies from the Same Population
                 p1         p2         p1*p2    p1*p2 p Value
Study 01     0.5556     0.5158        0.2866          0.0076
Study 02     0.1761     0.4953        0.0872          0.3003
Study 03     0.0093     0.3317        0.0031          0.9557
Study 04     0.1875     0.4827        0.0905          0.2377
Study 05     0.4790     0.4519        0.2164          0.0123
Study 06     0.1627     0.4636        0.0754          0.3491
Study 07     0.4286     0.4586        0.1966          0.0419
```

```
Study 08     0.3524     0.4797        0.1691          0.0454
Study 09     0.2498     0.4574        0.1143          0.1792
Study 10     0.3661     0.2980        0.1091          0.0784
Study 11     0.3778     0.4255        0.1608          0.0434
Study 12     0.3076     0.3917        0.1205          0.0933
Study 13     0.5917     0.3451        0.2042          0.0078
Study 14     0.4009     0.4078        0.1635          0.0388
Study 15     0.1095     0.3644        0.0399          0.5504
Study 16     0.2259     0.3932        0.0888          0.2412
Study 17     0.4179     0.3900        0.1630          0.0360
Study 18     0.0348     0.4234        0.0147          0.8399
Study 19     0.4052     0.5119        0.2075          0.0169
Study 20     0.2909     0.3877        0.1128          0.1276
Study 21     0.3268     0.3664        0.1197          0.0657
Study 22     0.2251     0.3872        0.0872          0.2145
Study 23     0.4854     0.4117        0.1999          0.0161
Study 24     0.3170     0.4893        0.1551          0.0939
Study 25     0.2527     0.4594        0.1161          0.1660
```

Scanning the results of the 25 studies, each one produced a different result. Sometimes the omnibus mediation effect was statistically significant and sometimes not. For one sample/study, the omnibus mediation effect (mean difference on Y between the treatment and control conditions through the mediator M) was a meager 0.015 (Study 18) and for another study, it was 0.29, about a third of a standard deviation of the outcome. The true Y mean difference as a function of M is (0.40)(0.30) = 0.12, which is just over a tenth of a standard deviation of M. These sample-to-sample fluctuations are disconcerting and lead me to interpret the results for my one study cautiously.

Here is the table of summary statistics across all 10,000 studies:

```
Summary Statistics Across the 10000 Studies
           Average      SD      Min      Max      q10      q90
p1          0.3010  0.1740  -0.3920   0.9505   0.0765   0.5221
p2          0.3997  0.0823   0.1177   0.7413   0.2946   0.5049
p1*p2       0.1203  0.0752  -0.1815   0.4265   0.0284   0.2183
SEp1*p2     0.0763  0.0142   0.0285   0.1319   0.0581   0.0948
```

Across the studies, the average omnibus effect was 0.12, which maps well onto the true population effect. This is because the sample coefficient product is an unbiased estimator of the population coefficient product. One study in the group of 10,000 studies found that the control group scored better on the outcome through the M mediational chain, yielding a negative p1 of -0.392 (see `MIN`), while in another study (under `MAX`) the omnibus mediation effect was 0.95, almost a full standard deviation of Y. Across the 10,000 studies,

10% of the studies observed a mean difference through the mediator less than 0.076 (see the 10th quantile column, `q10`) while 10% of the studies observed differences greater than 0.522 (see the 90th quantile column, `q90`). The latter is much stronger than the true effect of 0.12. The standard error of the coefficient product was 0.0752, suggesting that the "typical" disparity between a sample result and the true population omnibus effect was 0.0752. The mean of the standard errors reported in each study (0.0763) was reasonably close to the true standard error of 0.0752.

Here is the output for the proportion of studies out of the 10,000 that found a statistically significant result ($p < 0.05$), for different parameters:

```
Analysis of Null Hypothesis Rejections Across the 10000 Studies
                      Power
p1*p2                0.3165
Joint signif test    0.4112
                           Proportion of null rejections
p1                                                0.4119
p2                                                0.9977
p3 (direct effect T to Y                          0.0491
```

The statistical power for the Sobel test of the omnibus mediation effect was only 0.32, which is bleak. Despite a 0.12 true population effect, only 32% of the samples yielded a statistically significant result. The power for the joint significance test was higher, at 0.41. The statistical power for p1 and p2 individually (0.41 and 0.99) was better, with the joint significance test having power roughly equal to the product of the power estimates for the individual paths. The direct effect from T to Y (p3) was zero in the population, so the result for p3 should be near the alpha level of 0.05. It was 0.049.

Finally, the plot of the sampling distribution showed a pattern that is slightly non-normal (see Figure 4).

Type I Errors for Omnibus Mediation Effects

You can use the simulation program to explore Type I errors by setting one or both of the population path coefficients equal to 0. For example, when I set the population p1 to 0, thereby producing a product coefficient of 0, and I re-ran the analysis, the proportion of null hypothesis rejections for the Sobel test was 0.016 and for the joint significance test it was 0.048. The Sobel test is known for being conservative for Type I errors when the true p1*p2 = 0 and this is reflected in the simulation.

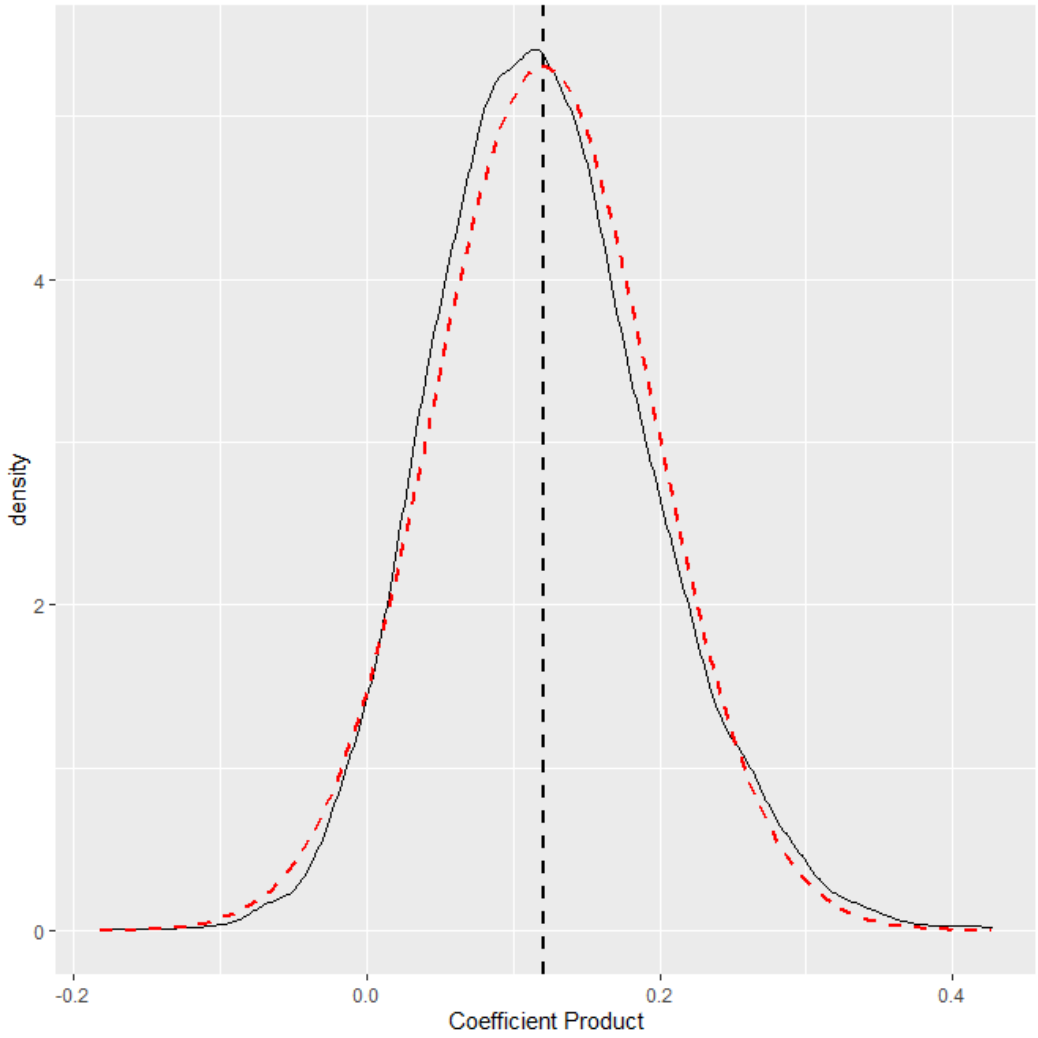

FIGURE 4. Sampling distribution of the product of coefficients

Reducing Sampling Error (Sample-to-Sample Fluctuations)

As with the prior simulations, the primary methods for reducing sampling error for the analysis of omnibus mediation are (1) increase sample size, (2) incorporate strategically selected covariates that impact the mediator or outcome, and (3) restrict the study to a homogeneous population and ensure uniform implementation of study and treatment protocols.

Sample Size Demands When Analyzing Omnibus Mediation Effects

As you work with the omnibus mediation simulation program, you will discover that the

tests can be sample size demanding. In my book, I argue that omnibus tests often are not central to the program evaluations using RETs; that analysis of the individual links in a mediational chain is more informative. Having said that, statements about mediation often require taking into account multivariate patterns of statistical significance across multiple paths in a mediational chain (vis-à-vis the joint significance test) and there are occasions where one will want to make statements about omnibus mediation coefficients. As noted, roughly speaking, the statistical power of the omnibus effect via a joint significance test will be a multiplicative function of the statistical power for each individual path comprising the mediational chain, so you can use this as a rough guide to inform sample size needs; if you power your study so that p1 and p2 both power of 0.90, then the power for the multivariate pattern of statistical significance for both paths via the joint significance test will roughly be (0.90)(090) = 0.81, with qualifications introduced based on the dependency structure between p1 and p2.

INTERACTIVE SIMULATION FOR CORRELATIONS

The Basics of Evaluating Pearson Correlations

The final simulation is for a single Pearson correlation. You specify the population correlation and the sample size. The program generates a very large population with normally distributed X and Y scores with a correlation of the magnitude you specify and then performs 10,000 replication studies that randomly sample cases of size N. The simulation program allows you to appreciate the dynamics of the operative sampling error across the samples.

Suppose the population correlation is 0.25 and the sample size is 100. After executing the generated R syntax, the program conducts 10,000 "replications," each time selecting a different random sample of 100 cases from the population. As with the other simulation programs, the program first reports a sub-group of 25 of the 10,000 studies. Here are the results:

```
25 Example Studies from the Same Population
            Correlation    p Value
Study 01         0.3478     0.0004
Study 02         0.4167     0.0000
Study 03         0.2972     0.0027
Study 04         0.1835     0.0676
Study 05         0.1662     0.0985
Study 06         0.2888     0.0036
Study 07         0.2211     0.0271
Study 08         0.1995     0.0466
```

```
Study 09         0.3312     0.0008
Study 10         0.0720     0.4766
Study 11         0.3308     0.0008
Study 12         0.2307     0.0209
Study 13         0.2579     0.0096
Study 14         0.1999     0.0462
Study 15         0.3754     0.0001
Study 16         0.3634     0.0002
Study 17         0.4286     0.0000
Study 18         0.2357     0.0182
Study 19         0.2245     0.0247
Study 20         0.1312     0.1931
Study 21         0.2799     0.0048
Study 22         0.2443     0.0143
Study 23         0.2012     0.0448
Study 24         0.2993     0.0025
Study 25         0.3205     0.0011
```

Scanning the results of the 25 studies, each one produced a different result. Sometimes the sample correlation was statistically significant and sometimes not. For one sample, the correlation was 0.07 (Study 10) and for another study, it was 0.43 (Study 17). If you were to conduct a study using this population, your study could have produced any of these 25 results, or indeed, any of the 10,000 results in the broader simulation. The sample-to-sample fluctuations lead me to interpret the results for my one study with humility.

Here is the table of summary statistics across all 10,000 studies:

```
Summary Statistics Across the 10000 Studies
               Average      SD      Min      Max      q10      q90
Correlation     0.2483  0.0948  -0.1191   0.5878   0.1242   0.3676
```

Across studies, the average correlation was 0.248, which maps well onto the true population effect. Despite this, the sample correlation tends to be a biased estimator of the true population correlation, but the bias tends to be trivial for sample sizes larger than 20 and it can virtually be ignored.[2] One study in the group of 10,000 studies found a correlation of -0.11 (see `MIN`), while another study (under `MAX`), the correlation was 0.59. Across the 10,000 studies, 10% of the studies observed a correlation less than 0.12 (see the 10th quantile column, `q10`) while 10% of the studies observed a correlation greater than

[2] This is not to say that the Pearson correlation is always a good index of association. It suffers from shortcomings, as outlined in Wilcox (2017).

0.37 (see the 90th quantile column, `q90`). The standard error of the correlation was 0.10 (see the `SD` column), suggesting that the "typical" disparity between a sample result and the true population correlation was 0.10. I do not provide the mean of the standard errors from each study because there is no formula for estimating the standard error of a correlation (it usually is accomplished through bootstrapping).

When I examined the proportion of studies out of the 10,000 that found a statistically significant result ($p < 0.05$), the proportion was 0.72. This is the statistical power of the test of significance with sample sizes of 100 where the population correlation is 0.25.

Finally, the plot of the sampling distribution yielded a distribution that was relatively normal (see Figure 5). This will not always be the case, as the distribution is affected by the magnitude of the correlation. I leave it as an exercise for you to explore this.

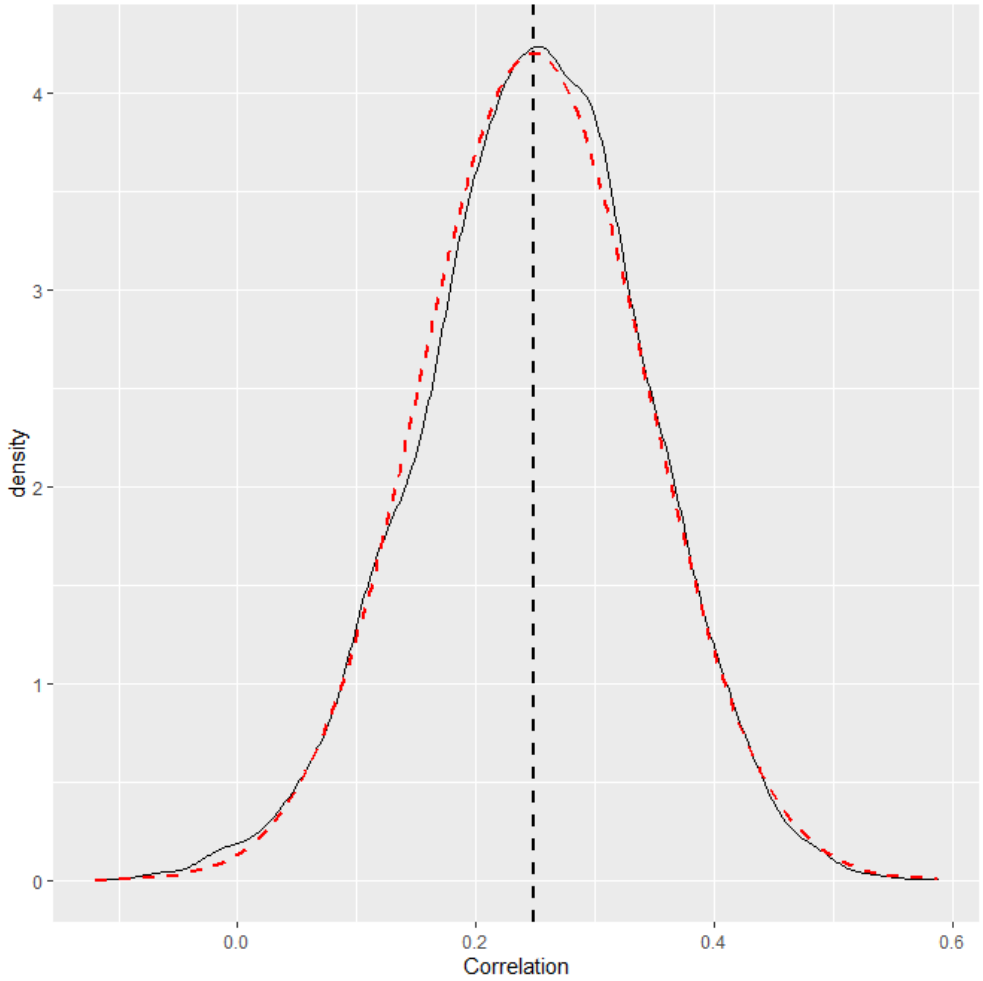

FIGURE 5. Sampling distribution of a correlation

CONCLUDING COMMENTS

I designed the five simulation programs to help give you a sense of sampling error dynamics for population parameters that often are of interest in RETs. My intent is for you to use the programs to explore how variations in study design and analysis (e.g., sample size, covariate control) can affect sampling error. Too often, researchers focus primarily on matters of Type I and Type II errors (statistical power), but such foci does not do justice to fully appreciating the implications of the nemesis of sampling error. My own orientation is to try to reduce sampling error as much as possible in the RETs I conduct, usually by seeking large sample sizes, by the strategic use of covariates, and by careful attention to the implementation of study protocols. I obsess about keeping my standard errors low and my margins of error narrow (since MOEs are directly related to standard errors), just as much as I worry about Type I and Type II errors. Hopefully you will find the simulation programs useful in deepening your understanding of sampling error.

REFERENCES

Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement*, 32, 579–599.

Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing*. Academic Press (Fourth edition).