Applying Resampling to Analyze the Sensitivity of a Hypothesis Test to Confounding

William M. Goodman, Ph.D.[1]
[1]University of Ontario Institute of Technology, 2000 Simcoe Street North, Oshawa, Ontario, Canada L1H 7K4

Abstract
This paper demonstrates a resampling-based method for sensitivity analysis, to quantify the risk that a hypothesis test result may be in error, as a consequence of a confounder's influence. Due to confounding, test statistics that apparently occur on the tail of the null distribution, and so are "significant", may in fact be in non-critical regions—if the null distribution is corrected to reflect the confounder's impact on sample measurements. The proposed method is analogous to creating a power curve, based on the varying risks of Type II error depending on the relative proximity of the true population mean to the null mean. Similarly, we display varying risks of reaching false positive conclusions, due to confounding, depending on the relative severity of the bias. The proposed method is introduced in the context of a multi-stage experiment, which is intended to illustrate the general conditions for using the method, and showing how to apply it.

Key Words: Resampling, simulation, confounders, sensitivity analysis

1. Introduction and Model Description

This paper demonstrates how resampling techniques can be used to assess the potential impact of suspected confounding on the results of a hypothesis test. The focus of the proposed method is *a posteriori*. Where possible, good experimental designs (such as decisions on sample size and/or blocking) are clearly recommended, but controlled experiments are not always possible. Even where they are possible, a new or previously unknown confounding factor may be identified after the data are collected.

A preliminary version of the resampling-based method described here was developed in the context of research by the author on a certain commercial tool used for evaluating critical thinking. Resampling procedures were used to model the instrument's susceptibility to confounding, which might explain some seemingly contradictory results found in the literature where that instrument has been used (Goodman 2008). The present paper is intended to generalize and formalize the basic methods introduced in that cited case.

By applying the methods proposed here, one can generate a set of conditional probabilities, of the sort illustrated below. The following table and figure display the varying risks of reaching false positive conclusions, in consequence of varying possible severities of confounding effects that are suspected to be present. (It is presumed that the *exact* severities of these effects, if any, are not specifically known.) The type of output described here is analogous to creating a power curve, based on the varying risks of Type II error depending on the relative proximity of the true population mean to the null mean (although the true distance is probably unknown). Analogously, we here display the varying risks of reaching false positive conclusions, due to the relative severities of confounding effects that are suspected to be present.

For a given H_0, with a degree of confounding c_i = ...	P(*p*-value < α | H_0 is true, but Confounding = c_i)
0.00	5.0%
0.07	5.6%
0.15	9.0%
0.20	10.4%
⋮	⋮

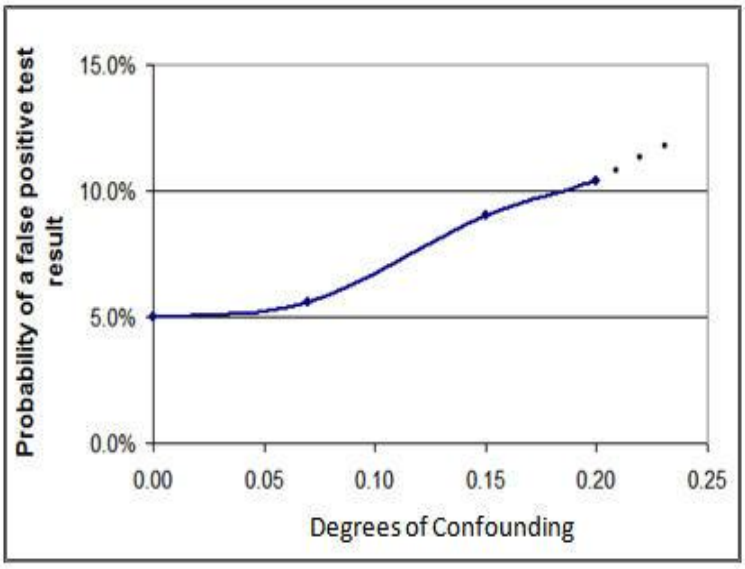

Figure 1: Possible Output Formats for Analyses of Sensitivity to Confounding

In the graph on the right, the particular *x* scale depends on the causal, confounding factors whose presence is suspected. There is no standard scale to fit all cases. Often, a bit of confounding can be tolerated—namely, when the probability of a false positive test result is just about what we expect (i.e., approximately equal to α). As the confounding severity increases, then the effective, true value of α will start to increase, meaning that there is a greater probability of a false positive result than the researcher may realize. For example, in the figure, if the extent of confounding is 0.15, then when the *p*-value is calculated (based on conventional—but flawed—assumptions about H_0) the calculated *p*-value will *appear* to be less than (α = 0.05) 9% of the time, so 9% is the true, effective value of alpha for the test.

2. Literature Note

Important issues of how (and why) to assess the sensitivity of test results to unmeasured confounders (or to "adjust" for their effects) have been raised in a long literature— particularly for research (e.g. observational and epidemiological) where rigid controls and randomization are not feasible. A classical example is found in Cornfield's 1959 work on smoking and lung cancer (cited in Steenland & Greenland, 2004). Addressing the possibility that the apparently strong correlation between those two variables could be due just to confounding, Cornfield determined that this probability is remotely small.

However, as lamented by Sander Greenland (a major contributor to this literature) (2005), the analysis of potential bias by confounding "has never taken root in basic statistics teaching and is hence uncommon" in many important applications. More often, a few relevant "study limitations" related to possible confounding are discussed informally and qualitatively in authors' papers. Possibly contributing to this state of affairs is the complexity of trying, in practice, to formally parameterize the nature or impact of a suspected confounder (let alone of multiple suspected confounders), and of applying some of the analytical models suggested in the literature.

This paper proposes a fresh, hands-on approach. While generally in the class of methods based on sensitivity analyses (compare Lin *et al*, 1998, Margolis *et al*, 1999, and Schneeweiss, 2006), it hopefully can be perceived as more readily adaptable, in practice, by researchers in various fields, to encourage more widespread analysis of possible confounder bias in one's research.

3. Purpose and Research Questions

The purpose of the research is to begin to formalize and generalize the procedures described in the introduction for measuring and reporting the sensitivity of particular hypothesis tests to confounding due to causal factors whose

existence and potential impacts are suspected. Based on real, experimental data that were collected for this purpose, it was examined (a) how well does the proposed model fit the data, and (b) what does the model tell us about the experimental results that might not be obvious in any case, without recourse to the new model.

4. Methods

4.1 Multi-stage Approach

The following multi-stage approach was applied: (1) An actual experiment leading to (2) a hypothesis test was conducted—but with a simulated "storyline" to model how confounding might enter, unknown, into the measurements, thus invalidating the test results. (3) Then a "meta-experiment" was conducted, to confirm that confounding did indeed bias the previous experimental results. (4) Given this confirmation of confounding, a procedure was developed to assess the test's sensitivity (in (2)) to precisely the type of confounding that was identified. That is, how much of the confounding effect would have been sufficient to bias the original experiment? The overall structure of the multi-stage experiment is illustrated below.

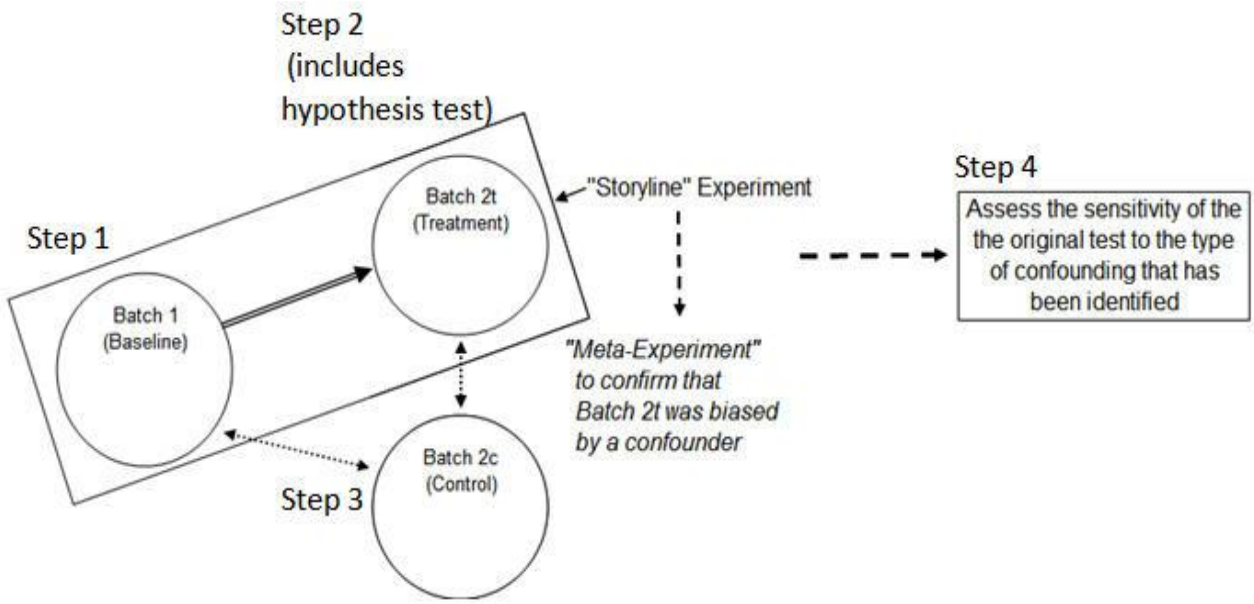

Figure 2: Basic Structure of the Multi-stage Experiment

4.2 Step One

Eighty objects (snowballs, comprising Batch 1 in the figure) were constructed with fresh, pliable snow, to match as closely as possible the dimensions of an exemplar. Each object's circumference was measured and recorded, using the method of wrapping a (non-stretchable) cord around the circumference, and then measuring the length of the used portion of the cord against a ruler. By the next day, the quality of the snow had changed, becoming less pliable. The "storyline" hypothesis was that, if snowballs were constructed with the day-old snow exactly as with the original, fresh snow in Batch 1, the measured circumferences should *not* be significantly changed.

4.3 Step Two

A new batch (Batch 2t) of 80 snowballs was then constructed, using the now-older and less pliable snow, to match as closely as possible the same dimensions as in Batch 1. Each object's circumference was measured and recorded as in Batch 1—except this time wrapping an *elasticized* cord (such as used in sewing) around the circumference of each object, and then measuring the length of the used portion of the cord against a ruler. For the storyline: Presume that the experimenter was unaware that the cord was different from the one used for Batch 1, or else did not realize that it might matter. A conventional test was used to compare the mean circumferences (as measured) for the two batches of

objects, Batch 1 and Batch 2t. Not surprisingly (stepping outside the storyline), an apparent difference in means *was* found between the two batches.

4.4 Step Three

Outside the storyline, we presume that the apparent difference in means between Batches 1 and 2t was really due to confounding introduced by the change in cords used for measuring. But it is conceivable that the mean circumferences *really did* change between the two measuring sessions. Batch 2c was therefore constructed as a control, to ensure that no real change did occur in the actual circumferences being measured. Namely, 80 more objects (snowballs) were created under similar conditions to Batch 2t, but this time the original, non-stretchable cord was used for measuring circumferences. Not surprisingly, the mean circumferences were the same as for the original batch.

4.5 Step Four

(a) Having now established that confounding likely caused the apparent difference in means between Batches 1 and 2t, external data were sought to help quantify the degree of confounding error that might have been introduced in the first ("storyline") experiment. Namely, a constant, snowball-size circumference was measured 10 times using the original cord method, and then 10 times with the elasticized cord method; the discrepancy between the average apparent measurements by each method was recorded.

(b) Given the assessment (in (a)) of possible measurement error, a resampling-based method was used to see how sensitive the original hypothesis test (in step two) would be to generating false positive results, given possible degrees of confounding due to the measurement error just discussed. In particular, it was assessed whether the apparent false positive in the storyline experiment would be a likely outcome, given the possible degrees of confounding..

Original	Resample	Case 0: Unadjusted Resample	Case 1: Adjust.. by -1%	Case 2: Adjust ..by -2%	Case 3: Adjust ..by -2.5%	Case 4: Adjust ...by −3%	*etcetera* ...	Summaries	Two-Tailed T-Test Results
27.5	27.00	24.00	23.74	23.48	23.35	23.21	...	MEANS	
25.5	28.50	26.00	25.74	25.48	25.35	25.21	...	Batch1 (Resampled)	
26.5	26.00	26.00	25.74	25.48	25.35	25.21	...	26.18	Batch1<=> 2:Case 0
30.0	27.00	27.00	26.74	26.48	26.35	26.21	...	Batch2 (Case 0)	0.06
26.0	25.00	25.00	24.74	24.48	24.35	24.21	...	26.53	Batches 1<=> 2:Case 1
26.0	26.00	27.50	27.24	26.98	26.85	26.71	...	Batch2 (Case 1)	0.63
27.0	27.00	27.50	27.24	26.98	26.85	26.71	...	26.26	Batches1<=> 2:Case 2
26.0	26.00	26.50	26.24	25.98	25.85	25.71	...	Batch2 (Case 2)	0.34
24.0	26.00	26.00	25.74	25.48	25.35	25.21	...	26.00	Batches1<=> 2:Case 3
26.0	28.50	27.00	26.74	26.48	26.35	26.21	...	Batch2 (Case	0.10

Figure 3: The Simulation Model

4.6 The Simulation Model

The basis for Step 4, above, was the author's simulation model, as partially illustrated in Figure 3. Data for the original, Batch 1 sample comprised the left column of the figure. New Batch 1 samples can be simulated based on sampling with replacement from the original data set (since the latter is presumed to be the best available indicator of the initial population distribution). The second column of the figure (partly) shows one such resample from the original data. On the null hypothesis for comparing the mean circumferences in Batch 1 and Batch 2, the Batch 2 sample could be simulated by another resampling (without adjustments) from the original data set. The "Case" columns correspond to possible degrees of confounding that could be introduced to "adjust" the non-adjusted resampling data, to better reflect what is likely to be *measured* in Batch 2 (due to confounding effects)—in spite of the fact that the distribution of the underlying resample followed the null hypothesis of "no change". The rightmost column shows the apparent *p*-value for the test of interest, for the different cases—e.g.., comparing the mean of the new, simulated Batch 1 sample with the mean of the new, simulated Batch 2 sample, as adjusted by varying degrees of confounding. The proposed

method does not restrict the test statistic to be calculated; for example in epidemiological studies, rate ratios or odds ratios might be calculated.

There is no constant rule for how the above "adjustments" columns should be constructed; this should follow from an experimenter's knowledge or suspicions of what causal factors might be leading to confounding, and how they might operate. In the case of the "storyline" experiment, it could be expected that a stretchable cord would expand when wrapped around each snowball, and then contract to a smaller length when subsequently placed against a ruler to record a measurement; so *negative* "adjustments" of measured results seem plausible. Possible percentages of adjustments were made with respect to the expected sample *mean* values under H_0, because the confounding effect was not expected to be systematically greater for sample measurements that happened to fall above the null mean than for those that happened to fall below the null mean. The precise model used for the confounding effects would need to be defended, in practice; but by explicitly showing the steps and formulas that embody the suspected biases, this could facilitate important discussions, and lead to improvements in the model.

Note that Figure 3 only shows possible outcomes of a single simulation of (in this case) one Batch 1 sample versus one Batch 2 comparison, with various possible levels of confounding. If thousands of such resamples are generated the results might look like Figure 4. Each row simulates the possible results of one experiment, under various possible scenarios for confounding. For each scenario, the resulting *p*-values shown at right are based on conducting an intended hypothesis test as if not acknowledging or realizing that the confounding scenario is biasing the results. If under a scenario, the proportion of samples that portray a *p*-value < α is actually greater than the proportion indicated by α itself, then the confounding scenario has introduced a notable bias in the test results.

MEANS							p-values for TTest				
Batch 1 (Resampled)	Batch 2 (Case 0)	Batch 2 (Case 1)	Batch 2 (Case 2)	Batch 2 (Case 3)	Batch 2 (Case 4)	Batch 2 (Case 5)	Batchs 1 <=> 2:Case 0	Batchs 1 <=> 2:Case 1	Batchs 1 <=> 2:Case 2	Batchs 1 <=> 2:Case 3	Batchs 1 <=> 2:Case 4
26.606250	26.100000	25.833938	25.567875	25.434844	25.301813	22.641188	0.011716	0.000147	0.000001	0.000000	0.000000
26.381250	26.175000	25.911188	25.647375	25.515469	25.383563	22.745438	0.271371	0.012882	0.000128	0.000007	0.000000
26.412500	26.275000	26.010875	25.746750	25.614688	25.482625	22.841375	0.508303	0.054587	0.001608	0.000173	0.000014
26.156250	26.393750	26.132188	25.870625	25.739844	25.609063	22.993438	0.229658	0.902914	0.148970	0.036060	0.006126
26.168750	26.375000	26.113312	25.851625	25.720781	25.589938	22.973063	0.327107	0.791954	0.132672	0.034298	0.006489

Figure 4: Outcomes of Many Resamples

5. Results and Discussion

For the "storyline" experiment, the mean circumferences as measured were 26.3cm and 20.7 cm, respectively, for batches 1 and 2t. Based on a conventional t test, the difference was highly significant (p-value ≈ 0.000). (Both samples were large and the distributions were both reasonably normal; the results were the same whether equal, or unequal, variances were assumed.) For the meta-experiment step (see Step 3 in Figure 2), the mean circumferences as measured were 26.3cm and 26.7 cm, respectively, for batches 1 and 2c. This difference was *not* significant—suggesting that the apparent difference between batches 1 and 2t was spurious; i.e. not due to a real change in circumferences.

Using the procedure described in Step 4 (see Section 4.5), it was estimated that the effect of switching from using the original cord (in Step 1) to using the elasticized cord (in Step 2) was likely to decrease the (apparent) measured value of the mean circumference by as much as 13%. (The estimate is not precise, since the measured constant circumference did not include such features of actual snowballs as bumps and indentations, and potentials for deforming or breaking, etc.)

Based on the output of the proposed resampling method, the original hypothesis test described in Step 2 (see Section 4.3) was *highly* sensitive to confounding. (See Figure 5.) Of course, one might have *expected* that a conventional test for differences in means between batch 1 and batch 2t circumferences would yield false positive results, on learning in the previous paragraph that confounding has affected one of the two group's measurements by about 13%. However, it may not be as obvious that even a far smaller elasticity of the measuring tool could have grossly inflated the risk of Type I error. E.g. a measuring-tool elasticity of just 1% would have rendered the true, effective $\alpha \approx 27\%$; a measuring-tool elasticity of even ½ % would have rendered the true, effective $\alpha \approx 15\%$. In short, the test used was *highly* sensitive to virtually *any* degree of confounding, due to introducing elasticity into the measuring instrument.

For a given H_0, confounded to degree c_i = ...	P(*p*-value < α | H_0 is true, but Confounding = c_i)
0.0%	5.0%
1.0%	26.8%
2.0%	75.2%
2.5%	90.7%
3.0%	97.5%
•	•
•	•
•	•
13.0%	≈ 100%

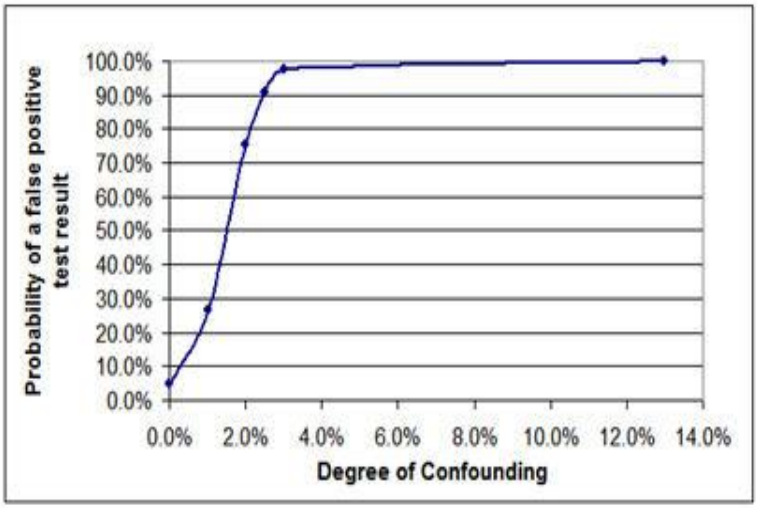

Figure 5: Output of the Analysis

Had this been a real case, any *future* researchers about snowballs would now be informed of the need for to better control for elasticity of the instrument, in the experimental design, and they would realize how sensitive the test is to that factor. Note that if, alternatively, someone had approached the unexpected results of batch two in the above case as coming from a simple, one-time "measurement error", then he or she might not be as cognizant of the need to control for confounding errors based on tool elasticity involving far less severity of this factor. In this way, the above output contributes new, useful information.

One consideration for future research is how to apply this method if, in turn, one uses a resampling-based method to determine the original *p*-values-as-measured. The models in Figures 3 and 4 were relatively easy to automate, since for each new simulated sample, software could directly calculate and record the both the revised test statistic of interest, and the corresponding *p*-value. But if the *p*-values for each pass are in turn determined by resampling, would the result be a "nested resampling" process that is onerous to perform?

It might also be useful to explore a method that starts from a test statistic, as generated from a conventional test, and asks: What is the true p-value for that test statistic, in light of the confounding factor? This is an alternative from the method proposed here, which assigns the *p*-value based on the *original* test assumptions, and then determines the true probability of obtaining such *p*-values that appear to be less than α. The relative advantages of these two methods would also be worth exploring.

Acknowledgements

I would like to thank the reviewers who invited me to present on this paper's topic at the 2008 Joint Statistical Meetings in Denver, and also to thank all colleagues who attended and/or co-presented at my session, providing

wonderful opportunities for feedback. I would also like to thank my two young Research Assistants, Rachel and Alek, for their great work building all the snowballs required for the experiments described above.

References

Goodman, W.M. 2008. Critical-thinking assessment: A case applying resampling to analyze the sensitivity of a hypothesis test to confounding. In *Case Studies in Business, Industry and Government Statistics*. 2(1).

Greenland, S. 2005. Multiple-bias modeling for analysis of observational data. In *Journal of the Royal Statistical Society A*. 168(Part 2): 267-306.

Lin, D.Y., Psaty, B.M., & Kronmal, R.A. 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. In *Biometrics*. 54(9): 948-963.

Margolis, J., Berlin, J.A., & Strom, B.L. 1999. A comparison of sensitivity analyses of the effect of wound duration on wound healing. In *Journal of Clinical Epidemiology*. 52(2): 123-128.

McCandless, L.C., Gustafson, P., & Levy, A. 2007. Bayesian sensitivity analysis for unmeasured confounding in observational studies. In *Statistics in Medicine*. 26: 2631-2347.

McCandless, L.C., Richardson, S., & Best, N. 2008. Adjustment for unmeasured confounding using propensity scores. (Under review; accessed in August 2008 at URL http://www.stat.ubc.ca/~lawrence/SABSA.pdf.)

Schneeweiss, S. 2006. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. In *Pharmacoepidemiology and Drug Safety*. 15: 291-303.

Steenland, K. & Greenland, S. 2004. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a sudy of silica and lung cancer. In *American Journal of Epidemiology*. 160(4): 384-392.