Suspect Research and Statistical Inference

Jacob Schauer, Northwestern University

Background

Especially over the past two decades, the emphasis in education has been to implement policy backed by rigorous research, which often involves randomized trials. Simultaneously, a replication crisis has arisen in science, including in psychology and economics (Camerer et al., 2016; Ioannidis, 2005; Open Science Collaboration, 2015). A prevailing opinion is that suspect research practices—e.g., p-hacking—contribute significantly to this crisis (Lindsay, 2015). This has eroded trust in published findings, and it raises questions about how to form policy given questionable evidence (Hedges, 2017). Such practices also have implications for moves within education to make research more transparent. These are important issues with no easy solutions, but it is clear that as a field, we must understand the impact of suspect research practices (SRP) on statistical inferences and the policymaking calculus.

Focus of Study

This paper considers the practice of conditional data collection (CDC), and how it can induce bias in statistical inferences. CDC occurs when the decision to recruit additional participants in an experiment depends on some observation about data that has already been collected. For instance, a researcher who fails to find a statistically significant result may add additional subjects to their experiment. In a survey of psychology researchers John, Loewenstein, & Prelec (2012) report that over half admitted to doing exactly this. However, we show in this paper shows that researchers need not actually run significance tests to bias inferences. Thus, we examine the bias of estimators based on conditionally collected data, where the decision to collect more data depends on either complete or partial knowledge of the interim results. As described below, we model complete knowledge of interim results deterministically and partial knowledge probablistically.

Research Design

Suppose an initial experiment is conducted with n subjects in each of a treatment and control arm. Observations are assumed independent and normal with known variance σ^2. The results that follow are exact, but when the variance is unknown, they are only approximate. The true effect θ is estimated by a mean difference T. The estimate from the first $2n$ subjects is $T_0 \sim N(\theta, 2\sigma^2/n)$.

Based on some observation O about T_0, m subjects are added per arm, leading to a new estimate T_1. This can be repeated an arbitrary number of times, however to simplify presentation, we focus on the case where this is done only once. Researchers might report T_1 as the study result, which assumes all data were collected independently. However, we only observe $T_1|O$.

CDC can induce bias for a broad class of O that contain information about T_0. We focus on two possible conceptions of O that involve the null hypothesis test $H_0 : \theta = 0$. First, we assume O is an observed outcome of the test involving T_0; if H_0 is maintained, then more data is collected. This corresponds to the classic definition of CDC, and it can greatly inflate Type I error rates (Simmons, Nelson, & Simonsohn, 2011).

However, researchers can distort inferences without observing test results. Instead, they may collect more data based on some idea about what the test's outcome may be. That is, given O, they know that T_0 will not be statistically significant with some probability η. If η is large, then more data is collected. This is an approximation of "passive data snooping". It may occur, for example, if a researcher observes data collection and gets the impression that the treatment group is not responding as expected. It is important to note that this can affect inferences *even if the researcher does not actually conduct the hypothesis test.*

Results and Conclusions

We present results here for scenarios where additional data has been collected. While a more complete set of results are available in the paper, what follows demonstrates that even collecting one addtional wave of data can greatly bias statistical inferences.

If data collection depends on any observation O that carries information about T_0, treatment effect estimators can be biased. In order to see this, note that we may write

$$T_1 = \frac{n}{n+m}T_0 + \frac{m}{n+m}X_1$$

where X_1 is the mean difference among the $2m$ new subjects. Since inference involving T_1 depends on O, the expectation of $T_1|O$ is

$$E[T_1|O] = \frac{n}{n+m}E[T_0|O] + \frac{m}{n+m}\theta$$

Note that $E[T_1|O] = \theta$ only if $E[T_0|O] = E[T_0] = \theta$. That is, $T_1|O$ can be biased whenever O contains information about T_0.

When O corresponds to the results of a hypothesis test (i.e, $O \equiv |T_0| \leq 1.96\sqrt{2\sigma^2/n}$), then $T_0|O$ will have a truncated normal distribution. The bias of $T_1|O$ can be derived analytically:

$$B_1 = \sigma\frac{\sqrt{2n}}{n+m}\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}$$

where α and β correspond to rejection regions for the hypothesis test. This expression depends on $\sqrt{2\sigma^2/n}$ and the initial power π_0 of the design with $2n$ subjects. Table 1 shows the percent bias of $T_1|O$ for a range of θ, π_0, and m/n. Note that these values are large (between 20-50%), but also negative, as CDC attenuates estimates toward zero if T_1 is reported regardless of statistical significance. However, if the researcher only reports T_1 when it is significant, the bias has a more complex expression, and can be positive (see Table 2).

We can distort inferences even if we do not conduct an interim analysis. Suppose the researcher knows that T_0 will be nonsignificant with probability η, and that η is large enough that s/he decides to collect more data. Estimates T_1 in this case have a more complex sampling distribution, but analytical results are still possible. The bias may be written as

$$B_\eta = \sigma\frac{\sqrt{2n}}{n+m}[\phi(\alpha) - \phi(\beta)]\left(\frac{\eta}{\Phi(\beta) - \Phi(\alpha)} - \frac{1-\eta}{1 - [\Phi(\beta) - \Phi(\alpha)]}\right)$$

which depends on η, π_0, m, and n. When $\pi_0 = 1 - \eta$, the bias is zero, since O provides no new information about T_0. However, when that is not the case, the bias can be substantial. Table 3 shows the percent bias of $T_1|O$ when $\eta = 0.5$, which can be as high as 20%.

The decision to add data based on any observation about potential results can bias treatment effect estimators. And while the tables presented here show this occurs if data are collected in two waves, they can be generalized to any number of waves, and bias in those cases can be quite large. This suggests that explicit data snooping should be avoided, but also that behaviors that seem more reasonable can also be harmful.

References

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Hedges, L. V. (2017). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, 00–00. https://doi.org/10.1080/19345747.2017.1375583

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Table 1: Percent bias for T_1 | T_0 not significant. This table displays the percent bias of T_1 when T_0 is not statistically significant (H_0: $\theta = 0$, 2-sided $\alpha = 0.05$ level test). It is organized into three panels representing true effects that are small ($\theta = 0.2$), medium ($\theta = 0.5$), and large ($\theta = 0.8$) in Cohen's *d* scale. Columns correspond to the initial power ($\pi_0 = 0.4, 0.6, 0.8$) of the test involving T_0, and rows reflect adding *m* new data points where *m*/*n* ranges from 1% to 50%. Values reported are percent bias, e.g., for a small effect ($\theta = 0.2$), if the initial study had 80% power, T_1 has between -33% and -50% bias, depending on how many additional observations are collected.

	$\theta = 0.2$			$\theta = 0.5$			$\theta = 0.8$		
m/n	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$
0.01	-0.37	-0.43	-0.50	--	-0.44	-0.50	--	--	--
0.05	-0.36	-0.42	-0.48	-0.37	-0.42	-0.48	--	-0.43	-0.48
0.10	-0.34	-0.40	-0.45	-0.35	-0.40	-0.46	-0.36	-0.41	-0.46
0.25	-0.30	-0.35	-0.40	-0.31	-0.35	-0.40	-0.32	-0.36	-0.41
0.50	-0.25	-0.29	-0.33	-0.26	-0.29	-0.33	-0.26	-0.30	-0.34

Table 2: Percent bias for T_1 | T_0 not significant, and T_1 is significant. This table displays the percent bias of treatment effect estimates when T_0 is not significant (H_0: $\theta = 0$, 2-sided $\alpha = 0.05$ level test), and T_1 is only reported if it is statistically significant. It is organized into panels representing true treatment effects that are small ($\theta = 0.2$), medium ($\theta = 0.5$), and large ($\theta = 0.8$) in Cohen's *d* units. Columns correspond to the initial power ($\pi_0 = 0.4, 0.6, 0.8$), and rows reflect adding *m* new data points where *m*/*n* ranges from 1% to 50%. Values reported are percent bias. For instance, for a small effect ($\theta = 0.2$), if the initial study had 80% power, T_1 has between -23% and -28% bias, depending on how many additional observations are collected, given that it is significant.

	$\theta = 0.2$			$\theta = 0.5$			$\theta = 0.8$		
m/n	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$
0.01	0.18	-0.09	-0.28	--	-0.10	-0.28	--	--	--
0.05	0.20	-0.07	-0.26	0.16	-0.08	-0.27	--	-0.11	-0.29
0.10	0.20	-0.06	-0.25	0.17	-0.08	-0.26	0.10	-0.11	-0.28
0.25	0.19	-0.06	-0.24	0.16	-0.08	-0.24	0.10	-0.11	-0.26
0.50	0.15	-0.07	-0.23	0.12	-0.09	-0.24	0.06	-0.11	-0.25

Table 3: Percent Bias for $T_1 \mid O$. This table displays the percent bias of treatment effect estimates when we only know that T_0 will not reject the null hypothesis that $\theta = 0$ with probability 0.5 based on some observation O. It is organized into vertical panels representing true treatment effects that are small ($\theta = 0.2$), medium ($\theta = 0.5$), and large ($\theta = 0.8$) in Cohen's *d* units. Columns correspond to the initial power ($\pi_0 = 0.4, 0.6, 0.8$), and rows reflect adding m new data points where m/n ranges from 1% to 50%. Values reported are percent bias. For instance, if the true effect is small ($\theta = 0.2$) and the initial design had 80% power to detect it, T_1 has bias that is between -13% and -19%. Note that we need not observe the hypothesis test involving T_0 to induce this bias.

	$\theta = 0.2$			$\theta = 0.5$			$\theta = 0.8$		
m/n	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$
0.01	0.09	-0.07	-0.19	--	-0.08	-0.19	--	--	--
0.05	0.09	-0.07	-0.18	0.07	-0.08	-0.18	--	-0.10	-0.19
0.10	0.08	-0.07	-0.17	0.06	-0.08	-0.17	0.03	-0.09	-0.18
0.25	0.07	-0.06	-0.15	0.06	-0.07	-0.15	0.02	-0.08	-0.16
0.50	0.06	-0.05	-0.13	0.05	-0.06	-0.13	0.02	-0.07	-0.13