Appreciate Uncertainty around Statistical Estimates in Policy Research

Dr Akisato Suzuki

School of Politics and International Relations

University College Dublin

akisato.suzuki@ucd.ie

5 December 2020

Many pieces of policy research use statistical inference to estimate (i.e., do a "smart guess" about) the effects of policy interventions, based on data. As more and more data are becoming available and, accordingly, statistical modeling (aka data science) is becoming more common, it is crucial to understand the underlying assumption, or "nuance," of statistical estimates, for better evidence-based policymaking.

In this context, this short paper focuses on one of the reasons why caution is necessary when we interpret statistical estimates for the effects of policy interventions: uncertainty around these estimates. As the word "estimate" implies, there is always uncertainty around statistical estimates. In other words, estimates may or may not turn out to be correct, although it is possible to compute, and therefore evaluate, the *degree* of uncertainty for these estimates.

Scientific jargon and technicality often make it difficult for non-experts (and even experts) to see the full nuance of uncertainty around statistical estimates. Failure to see this nuance could give a false impression that a study is definite about the presence or absence of an effect, resulting in unwarranted overconfidence.

I discuss three common concepts that describe statistical estimates and are associated with uncertainty: (1) "statistical significance," (2) "statistical insignificance," and (3) the "mean." To make the discussion concrete, I use two studies in conflict research as examples. More detailed (but also more technical) discussion of mine can be found in Suzuki (2020a, 2020b).

1. "statistical significance" does not mean you can be sure that the effect of a policy intervention is real

The first example study is Nielsen et al. (2011). It examines the effect of a large sudden decrease in foreign aid on the likelihood of the onset of armed conflict in recipient countries. It provides a logical and detailed explanation of how a large sudden drop in

foreign aid can affect a power balance between the government and rebels in the recipient country, thereby increasing the likelihood of armed conflict (221-222). Then, the study analyzes data statistically to examine whether a large sudden decrease in foreign aid indeed has had such an effect. The analysis finds the effect of a large sudden drop in foreign aid "statistically significant" (225-229). The study concludes the statistical analysis supports its argument. The policy implication is that, to prevent greater likelihood of armed conflict, donor countries should not decrease foreign aid suddenly in a large scale.

Like Nielsen et al. (2011), many other studies in policy research argue that there is evidence for the effect of a policy intervention because the statistical estimate of the effect is "statistically significant." Statistical significance is determined based on a statistical measure of uncertainty, usually something called a p-value.[1] The p-value takes a value between 0 and 1. Conventionally, if a p-value for the statistical estimate of an effect is smaller than 0.05, the estimate is considered statistically significant.

The p-value itself is a continuous measure of uncertainty (Lew 2012). A smaller p-value implies lower uncertainty, but it is usually impossible to eliminate uncertainty entirely. In other words, even if a p-value for the statistical estimate of the effect of a policy intervention is smaller than 0.05 and therefore considered statistically significant, there is usually still some uncertainty left, over whether the effect is real. Indeed, it is rare that we can be sure the effect is real, only by looking at a p-value.

The American Statistical Association once issued a statement on p-values, saying "No single index [such as a p-value] should substitute for scientific reasoning" (Wasserstein and Lazar 2016, 132). In other words, to evaluate the credibility of an argument in policy research, it is insufficient to look only at the presence or absence of statistical significance. It is also necessary to evaluate the explanation of why and how a policy intervention is expected to have a stated effect. The argument by Nielsen et al. (2011) is plausible not just because they find the statistically significant effect of a large sudden decrease in foreign aid on the likelihood of armed conflict. It is also because Nielsen et al. (2011) provide a convincing explanation of why.

It is also important to note that a statistically significant effect, if indeed credible, does not necessarily mean that the effect is also *practically* significant. In other words, even if the uncertainty of the estimate of a policy intervention's effect is low and therefore considered statistically significant, the size of the effect might be too small to be practically useful. I elaborate on this point of effect sizes later.

2. "statistical insignificance" does not mean you can be sure that the effect of a policy intervention is null

Nielsen et al. (2011, 225) also report that they find no evidence for a large sudden *increase* in foreign aid causing greater likelihood of the onset of armed conflict in recipient

[1] An alternative measure to determine statistical significance is a confidence/credible interval, but my argument here equally applies to it (for greater detail, see Suzuki 2020b). Here, for simplicity, I focus on the p-value.

countries. The argument is partly based on the statistical estimate of the effect of such an increase being not statistically significant, but also accompanied by a detailed discussion of why a large sudden increase in foreign aid might not be conflict-provoking (Nielsen et al. 2011, 230).

Research papers occasionally say that there is no evidence for the effect of a policy intervention because it is "statistically insignificant" (a synonymous expression is "not statistically significant"). Conventionally, if a p-value is not smaller than 0.05, it is considered statistically insignificant. However, again, the p-value is a continuous measure of uncertainty. Thus, even if a p-value is not smaller than 0.05, it does not mean we can be sure that there is no effect. It is rare that we can be sure the effect is null, only by looking at a p-value. A p-value greater than 0.05 only means there is greater uncertainty over whether the effect is real, than when a p-value is smaller than 0.05. And how "greater" the uncertainty is also depends on how large a p-value is. For example, if a p-value is 0.51, the degree of uncertainty is barely different from when a p-value is 0.49, although the former means statistical insignificance while the latter means statistical significance. Indeed, some scholars have raised objections to the simplistic dichotomy of statistical significance vs. insignificance (e.g., Amrhein, Greenland, and McShane 2019).

It is also important to note that the absence of evidence does not mean evidence for the absence of an effect. It is possible that the effect of a policy intervention is real but statistical analysis fails to find its statistical significance, at least for two reasons. First, if the effect were small, more data would be necessary to identify such a small effect. Second, if the effect were heterogeneous, i.e., positive in some cases and negative in other cases, the averaged effect could be nearly zero. In such a case, separately analyzing these two types of cases would be necessary to identify each type of effect independently.

3. The "mean" is not the only plausible value

Now, I introduce the second example study to explain the remaining statistical concept I discuss. Ruggeri, Dorussen, and Gizelis (2017) study the effect of peacekeeper deployment in a particular locality on the chance of the termination of local conflict. Its statistical analysis estimates that the deployment of 500 United Nations (UN) peacekeepers in a particular locality decreases the likelihood of the continuation of local conflict, in comparison with their absence, on average from 0.9 to 0.48, i.e., by 0.42 (Ruggeri, Dorussen, and Gizelis 2017, 179).

Like Ruggeri, Dorussen, and Gizelis (2017), when considering the size of the effect of a policy intervention rather than merely the presence or absence of the effect, research papers often focus on the average, or the mean of statistical estimates, as the reference point. The mean should not be considered, however, as the only plausible value. Indeed, these papers also usually report an interval estimate (the so-called "confidence" or "credible" interval) to present other plausible values.

The interval estimate covers all plausible values around the mean under a specific "confidence" or "credible" level; the conventional level is 95%. Ruggeri, Dorussen, and

Gizelis (2017) also report a 95% confidence interval; it ranges approximately between 0.36 and 0.45 (Ruggeri, Dorussen, and Gizelis 2017, 179, Figure 4). In short, the estimate suggests that the deployment of 500 UN peacekeepers in a particular locality is likely to decrease the likelihood of the continuation of local conflict, by between 0.36 and 0.45. Table 1 presents these values in a compact format.

Table 1: Summary of the mean and interval estimates reported in Ruggeri, Dorussen, and Gizelis (2017)

		95% confidence interval	
	mean	lower bound	upper bound
Effect size	0.42 decrease	0.36 decrease	0.45 decrease

I will not elaborate here on the detailed definition and interpretation of the confidence/credible interval (for these, see, for example, Amrhein, Greenland, and McShane 2019; Kass 2011; Kruschke and Liddell 2018). Rather, I emphasize the point that the mean should not be the only reference point when the size of the effect of a policy intervention is discussed.

The mean may *sometimes* be the most plausible value within the interval estimate. In the aforementioned example, the deployment of 500 UN peacekeepers in a particular locality might indeed be most likely to decrease the likelihood of the continuation of local conflict by the mean 0.42, rather than any other value in the interval of 0.36 and 0.45. Even so, some other values around the mean may often be just slightly less plausible than the mean.

The values close to the upper and lower bounds of the interval are usually the least plausible values within the interval. But it is sometimes possible that the mean is not the most plausible value and a value closer to the upper or lower bound is more plausible than the mean (Kruschke 2015, 342–43).

It is also important to note that the plausibility of the most plausible value may not be as high as the sum of the plausibility of all other values within the interval. For example, in the aforementioned study, the plausibility that the mean 0.42 is the true effect size, may be smaller than the plausibility that any value between 0.36 and 0.45 other than 0.42 is the true effect size.

Finally, it is even possible that a value outside an interval estimate would turn out to be the true value when we eventually found out the truth. In the case of the aforementioned study, although the confidence interval ranges between 0.36 and 0.45, it is not entirely impossible that the true effect size is something outside the interval, for example, 0.34 or 0.48.

Of course, because we cannot know the truth *a priori* (or ever), we do statistical estimation to do a smart guess. What is important to remember is that an interval estimate, even if correctly used, only implies a value outside the interval is *less plausible* to be the true value than the values within the interval, when we do not know the truth.

4. Alternative approaches

Given these nuances of the common statistical concepts that are associated with uncertainty, one might wonder whether there is a clearer way to express uncertainty. I have been working to develop such a way. One possible approach is to present the probability that a policy intervention has a certain minimum size of effect (Suzuki 2020b). Probability is an intuitive continuous scale of uncertainty, as often used in everyday life (e.g., in a weather forecast). For example, one could compute the probability that peacekeeper deployment will reduce the likelihood of armed conflict at least by 0.05.

A more elaborated approach is to compare an expected cost when a policy intervention is done, and the one when it is not. These expected costs can be computed based on the practically useful size of the effect of a policy intervention, the probability that the policy intervention has such an effect, and the cost of the policy intervention and the cost of leaving the status quo untouched (Suzuki 2020a). Such an approach enables the direct evaluation of statistical estimates in light of practical consideration, i.e., a cost-benefit perspective. For example, one could compare whether it is overall optimal to deploy peacekeepers in a particular locality, given the statistical estimates of its effect size and probability to reduce the risk of armed conflict, and the cost of peacekeeper deployment and the cost of leaving the risk of armed conflict unchanged.

A brief summary of these methods and their applications in a computer program is available in a blog post in the Connected_Politics Lab at the School of Politics and International Relations, University College Dublin (Suzuki 2020c).

5. Conclusion

Statistics is the science of uncertainty and allows us to do a smart guess based on statistical theory. Such a smart guess may be more likely to be right than a random guess. But it is just “more likely” and not guaranteed to be right. Caution is necessary.

Acknowledgement

I am grateful for helpful comments from the editor. I would like to acknowledge the receipt of funding from the Irish Research Council (the grant number: GOIPD/2018/328) for the development of this work. The views expressed are my own unless otherwise stated, and do not necessarily represent those of the institutes/organizations to which I am/have been related.

References

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. “Retire Statistical Significance.” *Nature* 567: 305–7.

Kass, Robert E. 2011. “Statistical Inference: The Big Picture.” *Statistical Science* 26 (1): 1–9.

Kruschke, John K. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. London: Academic Press.

Kruschke, John K., and Torrin M. Liddell. 2018. “The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective.” *Psychonomic Bulletin and Review* 25 (1): 178–206.

Lew, Michael J. 2012. “Bad Statistical Practice in Pharmacology (and Other Basic Biomedical Disciplines): You Probably Don’t Know P.” *British Journal of Pharmacology* 166 (5): 1559–67.

Nielsen, Richard A., Michael G. Findley, Zachary S. Davis, Tara Candland, and Daniel L. Nielson. 2011. “Foreign Aid Shocks as a Cause of Violent Armed Conflict.” *American Journal of Political Science* 55 (2): 219–32.

Ruggeri, Andrea, Han Dorussen, and Theodora-Ismene Gizelis. 2017. “Winning the Peace Locally: UN Peacekeeping and Local Conflict.” *International Organization* 71 (1): 163–85.

Suzuki, Akisato. 2020a. “Policy Implications of Statistical Estimates: A General Bayesian Decision-Theoretic Model for Binary Outcomes.” arXiv: 2008.10903 [stat.ME]. https://arxiv.org/abs/2008.10903.

———. 2020b. “Presenting the Probabilities of Different Effect Sizes: Towards a Better Understanding and Communication of Statistical Uncertainty.” arXiv: 2008.07478 [stat.AP]. https://arxiv.org/abs/2008.07478.

———. 2020c. “New R Packages to Evaluate the Statistical Uncertainty of Causal Effects More Informatively.” *The Connected Politics Blog*, September 14. https://www.ucd.ie/connected_politics/blog/newrpackagestoevaluatethestatisticaluncertaintyofcausaleffectsmoreinformatively/.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. “The ASA’s Statement on p-Values: Context, Process, and Purpose.” *The American Statistician* 70 (2): 129–33.