Joachim Krueger (1999) Do we Need Inferential Statistics?. Psycoloquy: 10(066) Social Bias (19)

Volume: 10 (next, prev) **Issue**: 066 (next, prev) **Article:** 19 (next prev first) **Alternate versions:** ASCII Summary

Topic: Social bias View Topic

Article: 19) Krueger 10(066) Do we Need Inferential Statistics? View Article

PSYCOLOQUY (ISSN 1055-0143) is sponsored by the American Psychological Association (APA).

DO WE NEED INFERENTIAL STATISTICS? Reply to Sriram on Krueger on Social-Bias

Joachim Krueger
Department of Psychology
Brown University, Box 1853
Providence, RI 02912
http://www.brown.edu/Departments/Psychology/faculty/krueger.html

Joachim_Krueger@Brown.edu

Abstract

The distinction between descriptive and inferential statistics is indeed artificial. Sriram (1999) stresses the descriptive value of test statistics and their associated p values. I support his proposal and present a Bayesian argument (and example) for the connection between p values and coefficients of replicability. I also agree with the view that much implicit Bayesianism can be detected in the day-to-day operation of research workers. Coaxing this ghost out of the closet, I think, would make social psychological research 'more positive'.

Keywords

Bayes' rule, bias, hypothesis testing, individual differences probability, rationality, significance testing, social cognition, statistical inference

1. Sriram (1999) offers many thoughtful and provocative insights into the use of Null Hypothesis Significance Testing (NHST) in social psychology and elsewhere. Ranging from research on social bias to the behavioural (un)reality of the hot hand in tennis to work on neuro-imaging, Sriram makes three major points (and several minor ones). First, the problem with applications of NHST, he says, lies primarily in the NH part, and not so much in the ST part. Second, test statistics, and the p values they represent, are best understood as descriptive rather than inferential statistics. Third, most researchers act as if they were Bayesians, although they avoid using explicit Bayesian methods. I agree with these views, and in this reply I elaborate them where I think elaboration is needed.

I. CRITIQUING 'NH' RATHER THAN 'ST'

2. In my original statement (Krueger 1998), I stressed that research on social-cognitive biases is itself biased because it identifies rationality with the point-specific null hypothesis of no bias (H0), whereas bias lies on either side of that point (H0; Krueger, 1998). Some of these points even represent incorrect norms for rational thinking. The use of appropriate norms and the admission of null intervals could remove some of the distortions inherent in this paradigm. Sriram's suggestions of how to test streakiness in tennis games illustrates the flexibility of this approach.

II. INFERENTIAL STATISTICS ARE DESCRIPTIVE

3. The distinction between descriptive and inferential statistics is indeed arbitrary. Inferential statistics produce decisions concerning the status of hypotheses that can only be made if a decision criterion is agreed upon. The conventional practice to reject H0 requires the use of a cliff value ($p = .05$) that might as well lie elsewhere. Descriptive statistics depend wholly on the data sampled. Researchers only decide which statistics to compute, but not what to infer from them. Bayesian analyses and the construction of confidence intervals are powerful inductive tools that work well without the inevitable simplifications involved in all-or-none rejection decisions.

4. Sriram urges that the p value produced by NHST (i.e., the probability of the observed data, or data more extreme, will occur under H0, [p(D|H0)]) be interpreted in continuous instead of categorical fashion. He notes that this is done routinely in neuro-imaging studies because it facilitates communication among researchers. He also notes that the ambiguities of p are well-known, which I take to mean that people realise that p is a confounded index. Its value depends on the raw effect size (e.g., the difference between two means), the variability among the sample observation (which reflects, in part, the precision of measurement), and the size of the sample (which reflects, in part, the determination and the resourcefulness of the researcher).

5. Following Greenwald, Gonzalez, Harris, and Guthrie (1996), Sriram suggests that the exact p value is highly informative because it forecasts the replicability of the observed phenomenon. In a simulation study, Greenwald et al. (1996) demonstrated that as p decreases, the chances that an exact replication study will reject the null hypothesis at $p = .05$ or better, increase. This finding is provocative and important, but the method requires a crucial modification. I will use a hypothetical coin tossing experiment to illustrate the Greenwald method and a necessary Bayesian modification of it.

6. The method assumes that two hypotheses are under consideration. H0 is the usual null hypothesis of no difference, whereas H1 is its alternative, which is unknown before study. After study, H1 is identified with the observed effect size. This post-hoc setting of H1 is a common device when the theory is too imprecise to specify an exact H1 a priori (Goodman, 1999; Hagen, 1997). Then, "replicability can be computed as the power of an exact replication study" (Greenwald et al., p. 179). In other words, replicability is understood as the probability of the data (or data more extreme) under the alternative hypothesis (i.e., p(D|H1)).

7. Suppose 10 coin tosses yield 8 heads. If the coin is fair (H0: p(heads) = .5), the probability of this or anything more extreme to happen is about .05. Standard NHST suggests that H0 be rejected (regardless of the how sure we are at the outset that the coin is fair). We now assume that the coin is badly biased (H1: p(heads) = .8). Under H1, the probability of getting at least 8 heads in the next series of 10 tosses is .68. This probability is the power of the replication study. To illustrate the association between p and replicability, assume the first 10 tosses yielded 9 heads. This state of affairs would have led to a more

decisive rejection of H0 (p(D|H0) = .01). The H1 suggested by these data (p = .9) would generate greater power to reject H0 (i.e., the probability of getting at least 8 heads under this H1 = .93) in the follow-up study.

8. Greenwald's method overestimates the replicability of a rejection decision because it assumes for sure that the post-hoc H1 is true. If one could be sure of that, a replication study would be pointless. Indeed, H1 is only a guess of the true state of affairs given the data obtained. Although this H1 is more probable after the first study than before it, its probability is not 1. The original H0, which has become less probable, has not been ruled out. To estimate replicability, it is necessary to multiply the probability of the data under each hypothesis, p(D|H), with the posterior probability that the respective hypothesis is true, p(H|D), and to sum the products. The posterior probability of each hypothesis depends on its prior and the data.

9. In the case of two exhaustive and mutually exclusive hypotheses, the simplest assumption is that both hypotheses are equiprobable a priori. We would thus begin by assuming that p(H0) = p(H1) = .5. Eight heads may be observed in 10 tosses because either H0 or H1 is true. Because we know that p(D|H0) = .05 and that p(D|H1) = .68, p(D) = .37 (i.e., .5*.05 + .5*.68). According to Bayes's Rule, the posterior probability of H0 is .07 (p(H0|D) = p(H0)*p(D|H0)/p(D)). Therefore, the posterior probability of H1 is .93. Now, entering the follow-up study, we no longer consider the two hypotheses equiprobable. What we need to know is the probability that at least 8 out of 10 tosses will be heads. This can happen under H0 with p(H0)*p(D|H0) = .03*.05, or under H1 with p(H1)*p(D|H1) = .93*.68. The sum of the two products is .63.

10. In this example, Greenwald's method overestimates the replicability of null rejection by .05 (.68 - .63). The difference may be small, but this is due, in part, to the convenient assumption that the two hypotheses are equiprobable at the outset. In a coin tossing experiment, H0 may initially enjoy an advantage. If, for example, p(H0) were .9 at the outset, the probability of getting at least 8 heads in study 2 after 8 heads turned up in study 1 would be .42. In other words, replicability cannot be estimated without making assumptions about the prior probabilities of hypotheses. The riskier that initial experiment is (with p(H0) being high), the harder it is to replicate a rejection decision.

III. IMPLICIT BAYESIANISM

11. Predictions of replicability thus require Bayesian priors. Sriram observes that "while researchers have not settled on Bayesian procedures for adjusting posteriors given the data, their actual behavior may reflect just that." I agree. It is indeed curious that the most orthodox application of NHST cannot function without implicit Bayesian assumptions. When a researcher rejects a hypothesis because the observed data are improbable under that hypothesis, he or she discloses a belief that the hypothesis has become improbable given the data. This is, of course, the problem of inverse conditional probabilities (Falk & Greenbaum, 1995). When p(D|H0) is small, p(H0|D) is probably small as well, but its exact value is indeterminate if H0 has no prior probability. Across simulations, the two conditionals are highly correlated, but the probability needed for the rejection of H0 (p(H0|D) tends to remain larger than the probability offered by the significance test (p(D|H0).

IV. POSITIVE PSYCHOLOGY

12. Sriram notes that there is already a great deal of research on "human abilities and all that is good in us." Again, I agree. But as Sriram himself points out, most of this research is in the areas of sensation and perception. Indeed, this contrast highlights the peculiar focus of social psychological research on

objectionable behaviour and flawed thinking. Even the most celebrated research on altruism, for example, shows how people fail to help (Latan & Darley, 1968). Recent efforts to raise research interest in 'positive psychology' may restore some balance to the soft wing of psychology. Reaching a better understanding of well-being, happiness, and optimal functioning seems to be worthy goal (Seligman, 1998).

REFERENCES

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard. Theory & Psychology 5: 75-98.

Goodman, S. N. (1999).Toward evidence-based medical statistics. 2: The Bayes factor. Annals of Internal Medicine 130: 1005-1013.

Greenwald, A. G., Gonzalez, R., Harris, R. J. & Guthrie, D. (1996). Effect sizes and values: What should be reported and what should be replicated? Psychophysiology 33: 175-183.

Krueger, J. (1998). The bet on bias: A foregone conclusion? PSYCOLOQUY 9(046). ftp://ftp.princeton.edu/pub/harnad/Psycoloquy.1999.volume.9/ psyc.98.9.46.social-bias.1.krueger http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46

Latan, B & Darley, J. (1968). Group inhibition of bystander intervention. Journal of Personality and Social Psychology 10: 215-221.

Seligman, M. E. P. (1998). Building human strength: psychology's forgotten mission. http://www.apa.org/monitor/jan98/pres.html

Sriram N. (1999). Inferential statistics are descriptive. Commentary on Krueger on social-bias. PSYCOLOQUY 10(46) ftp://ftp.princeton.edu/pub/harnad/Psycoloquy.1999.volume.10/ psyc.99.10.046.social-bias.18.sriram http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?10.046

Volume: 10 (next, prev) **Issue**: 066 (next, prev) **Article:** 19 (next prev first) **Alternate versions:** ASCII Summary

Topic: Social bias View Topic

Article: 19) Krueger 10(066) Do we Need Inferential Statistics? View Article