On Asterisk Inflation

Joshua S. Goldstein, *American University*

ABSTRACT In recent years, many political science articles have changed the traditional and correct method for flagging statistically significant coefficients in tables. They now add extra asterisks at each significance level. This article describes the shift and why it should be reversed.

All scientific communities depend on the common use of symbols in published research. Research results from diverse locations, theoretical schools, and time periods must share a common language in order for knowledge cumulation to occur. In the social sciences, statistical-significance testing has played such a prominent role in theory development that the rigorous reporting of statistical results has become a core value in maintaining the integrity of the research process.

Over decades, in social science journals, a convention developed to flag significant coefficients in large statistical tables, usually of regression results, for the reader's convenience in interpreting those results. The convention uses asterisks after the results, with more asterisks showing higher significance levels, and a footnote at the end of the table explaining them. Although the American Psychological Association (APA) now recommends another method (see below), political scientists and other social science disciplines still rely on the asterisks as the standard. For example, in the October 2008 issue of the *American Journal of Political Science* (*AJPS*), all 15 articles reported statistical results in tables and 12 of these used asterisks after coefficients to indicate significance.

WHAT HAPPENED?

Traditionally, a single asterisk indicates significance at $p < .05$, two asterisks indicate $p < .01$, and three indicate $p < .001$. After decades of such usage, with style manuals indicating it as the preferred usage, and with other social sciences using the same system, one might consider the standard of one asterisk for .05 set in stone. However, in the past decade the use of asterisks has changed in political science. Today, frequently, one asterisk means $p < .10$. The magic level of $p < .05$ that used to receive an asterisk now receives two, and .01, which used to get two asterisks, now gets three. I call it asterisk inflation.

There is, of course, nothing magic about $p < .05$ in the first place. But a lot of social science history lies behind it. In assessing whether a hypothesis might be in error, the 5% tail of the bell curve serves as a shared benchmark of statistical significance, the relaxing of which loosens the rigor of hypothesis testing in the discipline.

Most importantly, after decades of standard usage, the .05 level is not arbitrary but a landmark by which readers scanning results in journals can assess what matters more and less. Had we used .06 or .04 the results would have come out fairly similarly and no less rigorously. The problem comes in changing the benchmark through time. Just as currency inflation erodes the value of money, asterisk inflation erodes the value of research findings. A result that would have been not significant 20 years ago could well be significant today. This important shift in disciplinary standards, however, does not appears to result from any disciplinary process of discussion and decision, but rather from individual decisions in a collective-goods dilemma where everyone must inflate their asterisks if enough others do so (or risk having their results underappreciated by readers).

In some inductive, theory-building applications of statistics, a level of $p < .10$ may matter enough to report (as, for example, has traditionally been done for vector autoregression results). In those cases another symbol, such as a + or † sign, can flag the .10 level, preserving the integrity of the asterisk system for the .05 to .001 levels. The recent shift, by contrast, seems to serve little purpose other than to make results look more impressive.

To confirm the changing usage of asterisks in political science journals, I counted articles in several issues. In the *American Political Science Review* (*APSR*) issue for September 1986, the four articles that used asterisks for statistical results (out of 13 articles in the issue) all used the standard of $p < .05$ for one asterisk. More than 11 years ago, in the September 1998 issue of the *APSR*, again all four articles with statistical tables (out of nine articles in the issue) used this traditional standard for asterisks. Six years ago, in the August 2003 issue, three out of 10 articles include statistical tables, and two of them use the .05 standard for one asterisk, but the third uses the new system, one asterisk for $p < .10$. The August 2008 issue does not contain any asterisks at all, but in the previous May 2008 issue, four out of seven articles report statistical results in tables, with three using the new .10 criterion and the fourth probably doing the same but with a typographical error on page 227 (two asterisks means both .01 and .05!).

Joshua S. Goldstein *is professor emeritus of international relations at American University in Washington, D.C., and research scholar at the University of Massachusetts, Amherst, where he lives. He can be reached through his Web site, joshuagoldstein.com.*

doi:10.1017/S104909650999059X

https://doi.org/10.1017/S1049096510990598 Published online by Cambridge University Press

Thus, over the past decade, the flagship journal of political science has changed the norm, without a decision or a discussion, for what an important symbol in social science means. This might be good for aspirants to tenure who seem to be finding more significant results than ever, but it is not good for the discipline.

Other journals appear to be following suit but more slowly. In the October 2008 issue of *AJPS*, six articles use the old .05 standard, seven articles use the new .10 standard, and two simply list *p* values without asterisks. Similarly, of six total articles in the October 2008 issue of the *Journal of Conflict Resolution* (*JCR*), three use the .05 method and two use the .10 method. And in the September 2008 issue of *International Studies Quarterly*, of 10 articles, one uses asterisks by the .10 method and three by the .05 method, but with one of the latter noting on each table that three asterisks means "significance at the .000 level" (see page 475), a rare accomplishment indeed in social science!

This brings us to the APSA style manual, which follows explicitly the *Chicago Manual*'s fifteenth edition unless otherwise stated. The APSA manual states: "Probability levels for coefficients or tests of statistical significance should be keyed by *, **, and *** in order of increasing significance (e.g., .05, .01, .001). An alternative to flagging significance is to report probability levels for all coefficients in a separate column" (APSA 2006, 36). This guidance would be crystal clear except for that little "e.g." (for example), which probably should read "i.e." (that is). If, after all, the *p* levels given are only an example, then authors could claim a loophole to justify inflating their use of asterisks, especially if they have not read the *Chicago Manual*.

The first APSA style manual (APSA 1988, 6) mandated one to three asterisks "in order of decreasing probability" and then referred the reader to the *Chicago Manual* for details. At that time the latter had the unequivocal language on using the "specific levels" of .05, .01, and .001. The APSA manual also showed an example table, which used one asterisk for $p \leq .05$. The 1993 revised APSA style manual (APSA 1993, 21) introduced the "e.g." language—one to three asterisks "in order of increasing significance (e.g., .05, .01, .001)"—but still sent the reader to the *Chicago Manual* for details.

There is, of course, nothing magic about $p < .05$ *in the first place. But a lot of social science history lies behind it. In assessing whether a hypothesis might be in error, the 5% tail of the bell curve serves as a shared benchmark of statistical significance, the relaxing of which loosens the rigor of hypothesis testing in the discipline.*

This asterisk anarchy creates, to say the least, an inconvenience to the reader. Nor does the inconsistent usage appear to be explained by the use of one-tailed and two-tailed statistical tests (even when the tailedness is noted in the table notes at all). Even if it were, $p < .01$ in a one-tailed test (three asterisks by the new system) is not equivalent to $p < .001$ in a two-tailed test (three asterisks by the old system).

WHAT'S RIGHT?

So what is the right way to use asterisks? The answer is, clearly, the old way—one asterisk for $p < .05$.

The bible of style for political science journals is the *Chicago Manual of Style*. Most journals' instructions for contributors refer either to the *Chicago Manual* directly (as *JCR* does) or to the APSA style manual (as *APSR* and *AJPS* do), which in turn is based on Chicago style.[1]

The current fifteenth edition of the *Chicago Manual* (University of Chicago Press 2006, 512) provides a paragraph on significance levels. It states that "asterisks may be used as reference marks. If two or three standard significance levels are noted, a single asterisk is used for the lowest level of significance, two for the next higher, and so on. If values other than these three are given, however, footnote letters are preferable to asterisks, to avoid misleading the reader." It then lists what are presumably "these three," although several sentences of notes (such as italicizing *p*) come between: "* $p < .05$ ** $p < .01$ *** $p < .001$." The manual goes on to suggest, "on *p* values consult the *Publication Manual of the American Psychological Association*" (which I will discuss shortly).

The reference to "misleading the reader" makes clear that the three values listed are those to be used. An earlier, thirteenth edition of the *Chicago Manual* (University of Chicago Press 1982, 335) is explicit on the point, stating, "By convention asterisks are used … with the specific levels being given in the notes below" and then listing the same three *p* values with their asterisks.

In 1999, in an *APSR* style sheet posted online but not in print, editor Ada Finifter tightened the rules:

> Tables should report statistical significance with one, two, and three asterisks respectively for *p* .05, .01, and .001, and with a note indicating this usage, and whether a one- or two-tailed test was applied. Please do not describe probability levels greater than .05 as "statistically significant" or use special symbols to flag them in tables … Nevertheless, as an alternative to the asterisk system that identifies probabilities of .05 and lower, exact probability levels for all variables may always be given.[2]

In the 2001 print revision (APSA 2001, 42), however, the "e.g." reappeared and the mention of Chicago disappeared (although the entire manual remained founded on Chicago style). The reversion to earlier language used in the 1993 style manual appears inadvertent, resulting from an editorial transition and no doubt the use of the previous print version as the basis for the new one. The 2001 language remains unchanged in the 2006 revision.

Even if we took the "e.g." as a genuine loophole, does a loophole in the APSA style manual language justify overturning decades of social science convention? One reason to answer no is that thoughtful political science guidance on the topic warns against it. The methodologically oriented journal *Political Analysis* provides authors this guidance:

> for those authors who wish to report "statistical significance," statistics with probability levels of less than .001, .01, and .05 may be flagged with 3, 2, and 1 asterisks, respectively, with notes that they

https://doi.org/10.1017/S1049096510990598 Published online by Cambridge University Press

> are significant at the given levels. Exact probability values may always be given. *Political Analysis* follows the conventional usage that the unmodified term "significant" implies statistical significance at the 5% level. Authors should not depart from this convention without good reason and without clearly indicating to readers the departure from convention.[3]

As with the *Chicago Manual*'s warning about "misleading the reader," the instructions from *Political Analysis* indicate the potential harm caused by changing the norm in asterisk usage.

Another reason the loophole approach does not make sense is the use of the .05 standard in other social science disciplines. Commensurability with other disciplines is a benefit in political science as it supports interdisciplinary research. In sociology, the traditional asterisk system with one for .05 is the norm (Leahey 2005). In the widely used statistical package SPSS, one asterisk means .05 and two mean .01 (Leech, Barrett, and Morgan 2004, 216). And the World Health Organization's style manual (WHO 1993, 32) prescribes the traditional usage of asterisks for the .05, .01, and .001 levels.

Finally, the APA, which used to use the traditional method, now recommends avoiding asterisks altogether:

> The preferred method of reporting P-values is to use an exact number . . . rather than as a range (e.g., $p > .05$ or $p < .05$). The principle here is that not everyone takes a strictly Pearsonian view of probabilities as strictly significant or strictly non-significant. Sir R. A. Fisher advocated the position that probabilities can be interpreted with varying degrees of significance. By providing an exact p-value rather than a range, readers may adopt either approach to evaluating probabilities. Additionally, when evaluating the strength of an effect, in the absence of other measures of effect size, the p-value can convey the strength of the finding. For instance, a p-value of 0.054 is more encouraging as a line for further study (say, with a larger sample size) than a p-value of 0.67. Additionally, a p-value of 0.012 indicates a stronger effect than a p-value of 0.049. Many journal editors and reviewers continue to insist upon the range method of reporting p-values for tests to make the a priori significance level more obvious, but because most researchers do not, in actual practice, select their supposedly a priori criterion for how unlikely a result has to be to be deemed significant, it is more honest to report the exact probability. (APA 2001, 170)

WHAT TO DO?

Political science may wish to follow psychology in the future and forgo asterisks in order to achieve "more honest" reporting. In the meantime, however, inflating the number of asterisks we use certainly falls in the category of *less* honest reporting.

Three measures would serve to bring political science back from this detour and restore the integrity of our statistical publications. First, the APSA style manual can substitute "i.e." for "e.g." and close the loophole once and for all. Second, journal editors can insist on correct usage and instruct reviewers accordingly. Third, authors themselves can try harder to do the right thing.

Even after we fix the problem, the discipline will be left with a legacy of articles that used too many asterisks in the first decade of the twenty-first century. Perhaps, like baseball fans coming to terms with records set during the steroids era, we will need to mark these suspect articles with, what else, an asterisk. ■

NOTES

1. ISQ, however, has no style rules for authors concerning statistical tables.
2. Ada Finifter, personal communication, December 2008.
3. *Political Analysis* (published by Oxford University Press), "Information for Authors," http://www.oxfordjournals.org/our_journals/polana/for_authors/general.html.

REFERENCES

American Psychological Association. 2001. *Publication Manual of the American Psychological Association*. 5th edition. Washington, D.C.: APA.

APSA. 1988. *Style Manual for Political Science*. Washington, D.C.: APSA.

———. 1993. *Style Manual for Political Science*. Revised ed. Washington, D.C.: APSA.

———. 2001. *Style Manual for Political Science*. Revised ed. Washington, D.C.: APSA.

———. 2006. *Style Manual for Political Science*. Revised ed. Washington, D.C.: APSA.

Leahey, Erin. 2005. "Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces* 84 (1): 1–24.

Leech, Nancy L., Karen Caplovitz Barrett, and George A. Morgan. 2004. *SPSS for Intermediate Statistics: Use and Interpretation*. New York: Routledge.

University of Chicago Press. 1982. *Chicago Manual of Style*. 13th ed. Chicago: University of Chicago Press.

———. 2006. *Chicago Manual of Style*. 15th ed. Chicago: University of Chicago Press.

World Health Organization. 1993. *WHO Editorial Style Manual*. Geneva: WHO.

https://doi.org/10.1017/S1049096510990598 Published online by Cambridge University Press