Behavior Research Methods
2006, 38 (1), 24-27

SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing

MICHAEL J STRUBE
Washington University, St. Louis, Missouri

The ease with which data can be collected and analyzed via personal computer makes it potentially attractive to "peek" at the data before a target sample size is achieved. This tactic might seem appealing because data collection could be stopped early, which would save valuable resources, if a peek revealed a significant effect. Unfortunately, such data snooping comes with a cost. When the null hypothesis is true, the Type I error rate is inflated, sometimes quite substantially. If the null hypothesis is false, premature significance testing leads to inflated estimates of power and effect size. This program provides simulation results for a wide variety of premature and repeated null hypothesis testing scenarios. It gives researchers the ability to know in advance the consequences of data peeking so that appropriate corrective action can be taken.

Imagine the following scenario. A researcher, heeding the advice to take statistical power seriously (see, e.g., Cohen, 1988, 1992, 1994; Wilkinson & Task Force, 1999), estimates in advance of his research study a sample size that will produce a .80 probability of correctly rejecting a false null hypothesis at the .05 level of significance. He is a bit disheartened at the size of the projected sample, perhaps because most published research is underpowered (see, e.g., Clark-Carter, 1997; Cohen, 1962; Dar, Serlin, & Omer, 1994; Finch, Cumming, & Thomason, 2001; Sedlmeier & Gigerenzer, 1989), and so suggests that fewer subjects are usually needed to find significant results. The situation is perhaps all the more frustrating given scarce subject pool resources and the time and money it will take to conduct the larger-than-expected study. But he sees a possible solution. Data collection is automated via computer, allowing each participant's responses to be quickly appended to the growing data file. Moreover, a simple statistical analysis is just a mouse-click away. Why not test the key hypothesis *as the data are collected* rather than waiting until the target sample is achieved? If a significant result emerges before the target sample size is reached, the study can be concluded early, saving valuable resources.

On the face of it, this plan seems sensible. Resources usually *are* quite scarce, and any savings on one study can be devoted to other research. Furthermore, the power analysis may have been based on a quite crude effect-size estimate in the first place. For example, there might be little similar research available from which to gather a confident empirical estimate of the effect size, perhaps leaving the researcher to use rather crude benchmarks for "small," "medium," and "large" effects (Cohen, 1992). Consequently, the target sample size might have a rather wide "confidence interval" around it, making the target a best guess, but perhaps not a very good one.

Although it might seem attractive to "peek" at the data to avoid collecting a larger sample size than necessary, those peeks are not free, in the statistical sense. They are similar to repeated null hypothesis testing after data collection is complete, with the attendant inflation of the Type I error rate. Important initial work on this problem was done by McCarroll, Crays, and Dunlap (1992), who conducted simulations to examine three kinds of peeking rules. Each of their simulations began with a basic two-group design ($n = 10$ per group) but then used one of three different decision rules regarding the collection of additional data points: (1) add a data point to each group if the current significance test is not statistically significant ($p > .05$), (2) add a data point to each group if the current significance test is not statistically significant but the F ratio is at least 1.00, and (3) add a data point to each group if the current significance test is not statistically significant but the probability level is at least less than a predetermined cutoff (which ranged from .06 to .10 in different simulations). In each simulation, up to 10 additional data points were added to each group, providing a maximum of 10 peeks at the data before data collection was suspended. Results clearly demonstrated Type I error inflation for each decision rule, with obtained Type I error rates as high as .39 for the third rule (i.e., "keep adding data if the current results are at least marginally significant").

Recently, Strube and Hanson (2004) extended McCarroll et al.'s (1992) work in several ways. They investigated a wide range of target sample sizes (e.g., of the sort

This program is available from the author via e-mail attachment, via FTP at www.artsci.wustl.edu/~socpsy (executable file name: snoop.exe, zipped files for Visual Basic, Version 5 Professional: snoop.zip), or on disk (send a self-addressed and stamped disk mailer to the author). Correspondence regarding this manuscript should be addressed to M. J Strube, Department of Psychology, Box 1125, One Brookings Drive, Washington University, St. Louis, MO 63130 (e-mail: mjstrube@wustl.edu).

a power analysis might suggest) and varied the sample size at which hypothesis testing was initiated so that in some cases it commenced quite early in the planned data-collection sequence, whereas in other cases it began quite late in the sequence. They also varied the number of data points added before a new test was conducted, in order to simulate cases in which batches of data were collected before each new hypothesis test was calculated. Hypothesis testing continued until a significant result was found or the target sample size was reached. Finally, Strube and Hanson examined the consequences of premature hypothesis testing when the null hypothesis was false as well as when the null hypothesis was true.

Several of Strube and Hanson's (2004) results are illustrated in Figures 1, 2, and 3 (which were generated using the program described later). Figure 1 shows the inflation of the Type I error rate when the population correlation is 0 for various target sample sizes and various sample sizes at which hypothesis testing starts with a testing increment of 1 (i.e., conduct a new hypothesis test with the addition of each new participant and stop only if the test is significant at $p < .05$ or the target sample size has been reached). Early starting points translate into more peeks at the data, which produce greater inflation of the Type I error rate, especially for large target sample sizes. For example, when (1) the target sample size is 25, (2) testing begins with a sample size of 5, and (3) a new test is conducted with the addition of each new participant, there are 21 potential peeks at the data before the target sample size is reached, and the Type I error is about .21. But if the target sample size is increased to 100 and the same starting point (5) is used, then there are 96 potential peeks at the data, and the Type I error is about .33.

Figure 2 illustrates the consequences of data peeking on power when the null hypothesis is false—in this case, when the population correlation is .20. Here, the estimated power is inflated when the data are tested repeatedly. This occurs because the extra hypothesis tests result in more rejections of the null hypothesis than if the data were tested only after the target sample size had been reached.

Finally, Figure 3 displays the estimated population correlation when the true population correlation is .20. Because the procedure is biased toward stopping only when significance is obtained or the target sample size is achieved, an inflated estimate of the underlying population correlation results because it takes a larger correlation to achieve significance at smaller sample sizes (when peeking occurs) than at the target sample size.

Of course, researchers may have a wide variety of effect sizes and target sample sizes and may want to know what would happen for a wide variety of starting points, testing increments, and α levels. The purpose of the present computer program is to provide a versatile and quick way to gauge the impact of premature and repeated null hypothesis testing under such widely varying conditions.

The SNOOP Program

The SNOOP program is written in Visual Basic (Version 5) and runs under the Windows operating system. SNOOP provides simulation results for a wide variety of premature and repeated null hypothesis testing scenarios defined by the user. The following simulation parameters can be specified using scroll bars on the program screen: (1) the number of iterations in the simulation (maximum = 10,000), (2) the magnitude of the population correlation being tested (i.e., the effect size in a power

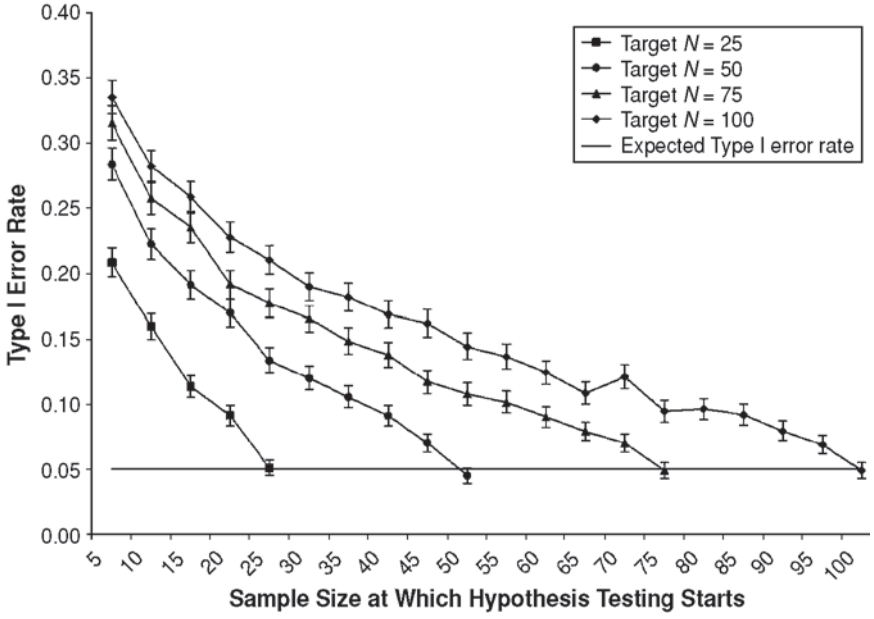

Figure 1. Estimated Type I error rates and 95% confidence intervals as a function of target sample size and sample size at which hypothesis testing starts when the testing increment is 1 ($\rho = 0$).

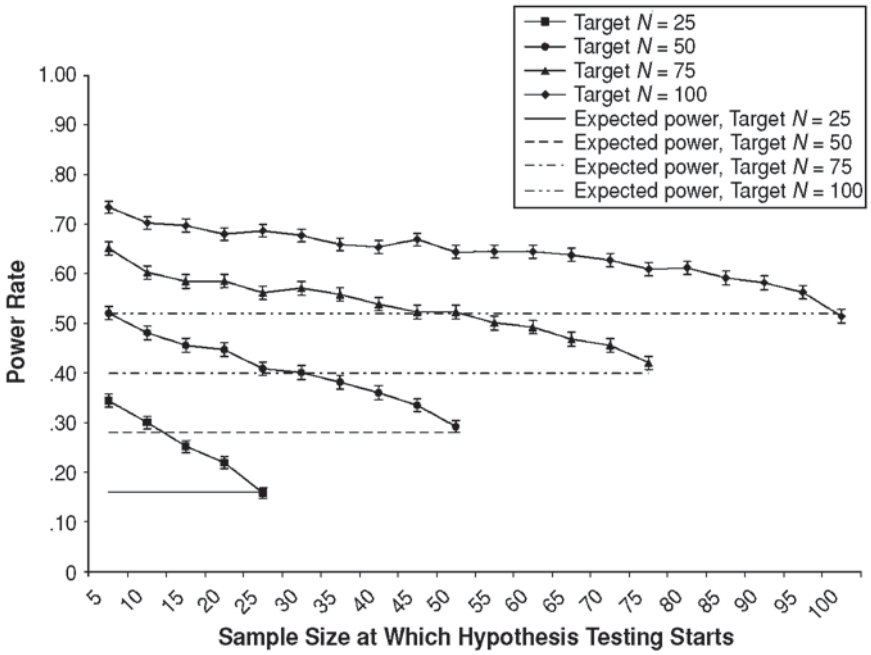

Figure 2. Estimated power rates and 95% confidence intervals as a function of target sample size and sample size at which hypothesis testing starts when the testing increment is 1 ($\rho = .20$).

analysis), (3) the target sample size for the study (perhaps from a power analysis, maximum = 250), (4) the sample size at which statistical testing is to begin (minimum = 5, maximum = target sample size), (5) the increment in sample size between successive statistical tests (minimum = 1), and (6) the nominal Type I error rate at which the empirical correlations are to be tested ($\alpha = .10$, .05, .025, or .01). After these parameters have been chosen, the program conducts a Monte Carlo simulation assuming bivariate normal distributions and displays the following results: (1) the average number of significance tests conducted per iteration, (2) the estimated population correlation based on the simulation, (3) the estimated Type I error rate (population $r = 0$) or estimated power rate (population $r \neq 0$), and (4) the 95% confidence limits for the Type I error rate or power rate.

For each iteration in a simulation, the program uses Wichmann and Hill's (1982) random number generator for producing uniform random numbers (for an evaluation of this algorithm, see Brysbaert, 1991) and the polar method (see Brysbaert, 1991) for transforming the uniform random numbers into standard normal random numbers. Pairs of correlated random numbers are created by first generating two independent standard normal variables (Z_1 and Z_2) and then creating a third standard normal variable (Z_3) using the following calculation:

$$Z_3 = \rho * Z_1 + \sqrt{1-\rho^2} * Z_2$$

where ρ is the population correlation. N pairs of variables Z_1 and Z_3 then are used to calculate the empirical correlations on which hypothesis testing is conducted (two-tailed tests are used).[1] Correlations are then tested for significance at the desired α level and at the desired starting point and testing increment. Testing proceeds for each iteration until a significant result is found or the target sample size is achieved. The tally of the number of iterations on which a significant result is found provides either the estimated Type I error rate (population $r = 0$) or the estimated power rate (population $r \neq 0$). The 95% confidence limits for the Type I error rate or the power rate are calculated using the binomial distribution.

The accuracy of the program increases with the number of iterations, but the program also takes longer to run the simulation. SNOOP runs quite quickly on computers equipped with Pentium III (or faster) processors, completing even the most intensive simulations (10,000 iterations, $N = 250$, starting point = 5, increment = 1) within 30 sec. With slower processors, reasonably accurate results can be obtained with as few as 1,000 iterations. When the starting value is set to the sample size, and the population correlation does not equal zero, the program can be used to estimate the power for different sample sizes.

The program is written to estimate the consequences of data peeking on the testing of correlations between two variables. Of course, the program applies equally well to simple two-group experimental designs, in which case the effect size is a point-biserial correlation. Similarly, the program can be used for simple two-group repeated measures designs, in which case the dependent measure represents a single degree of freedom transformation (e.g., linear or quadratic trend) of the original repeated measures.

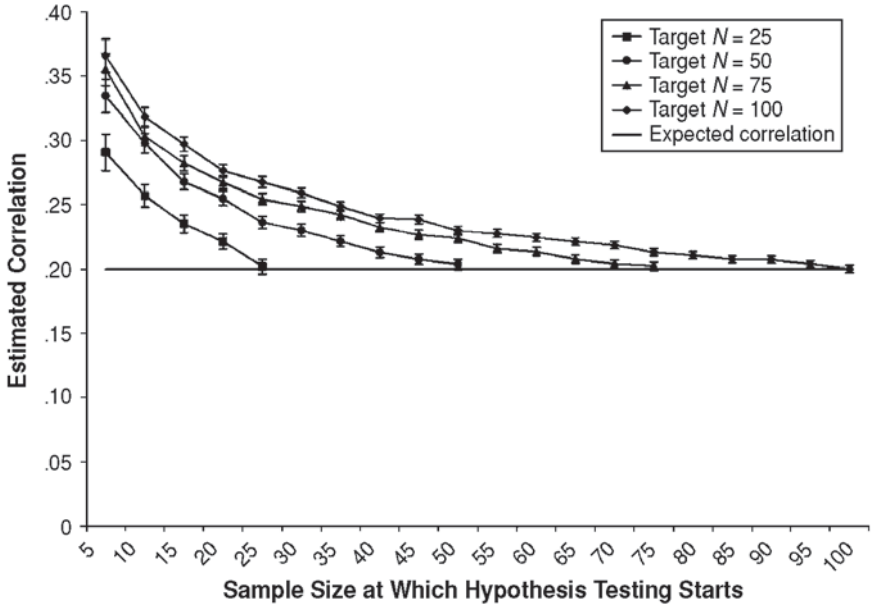

Figure 3. Estimated effect sizes and 95% confidence intervals as a function of target sample size and sample size at which hypothesis testing starts when the testing increment is 1 ($\rho = .20$).

Availability

An executable version of the program can be obtained as an e-mail attachment from the author at mjstrube@wustl.edu. The program can also be obtained by sending a disk and a self-addressed, stamped mailer to M. J Strube, Department of Psychology, Box 1125, One Brookings Drive, Washington University, St. Louis, MO 63130.

REFERENCES

Brysbaert, M. (1991). Algorithms for randomness in the behavioral sciences: A tutorial. *Behavior Research Methods, Instruments, & Computers*, **23**, 45-60.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, **88**, 71-83.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal & Social Psychology*, **65**, 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155-159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting & Clinical Psychology*, **62**, 75-82.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational & Psychological Measurement*, **61**, 181-210.

McCarroll, D., Crays, N., & Dunlap, W. P. (1992). Sequential ANOVAs and Type I error rates. *Educational & Psychological Measurement*, **52**, 387-393.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, **105**, 309-316.

Strube, M. J, & Hanson, J. S. (2004). *The perils of peeking: Consequences of premature and repeated null hypothesis testing*. Manuscript submitted for publication.

Wichmann, B. A., & Hill, J. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Applied Statistics*, **31**, 188-190.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594-604.

NOTE

1. Obtained correlations in the program are transformed using Fisher's procedure,

$$Z_r = .5 \ln \frac{1+r}{1-r},$$

and tested for significance using

$$Z = \frac{Z_r}{\frac{1}{\sqrt{N-3}}}.$$

The Fisher values are then averaged over trials in a particular simulation and then back-transformed to correlations for ease of interpretation:

$$r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}.$$

(Manuscript received September 17, 2004;
revision accepted for publication January 31, 2005.)