What Statistical Significance Means

Siu L. Chow
UNIVERSITY OF REGINA

ABSTRACT. Sohn (1998) presents a good argument that neither statistical significance nor effect size is indicative of the replicability of research results. His objection to the Bayesian argument is also succinct. However, his solution of the `replicability belief' issue is problematic, and his verdict that significance tests have no role to play in empirical research is debatable. The strengths and weaknesses of Sohn's argument may be seen by explicating some of his assertions.
KEY WORDS: chance, control, reliability, replication, statistical significance.

Sohn (1998) observes correctly that (1) the mathematical basis of statistical significance is the sampling distribution of the test statistic, (2) research prediction is based on theory, not on statistical significance, (3) a statistically significant result indicates neither the truth of the research hypothesis nor the replicability of the research result, and (4) the distinction need be made in methodological discussions between the truth of a hypothesis and the validity of the attempt to substantiate empirically the hypothesis (called *the truth-validity distinction* henceforth). He also rejects correctly, albeit with no explication, the Bayesian alternative on the grounds that (1) the Bayesian approach is predicated on a data-collection situation that is not typical of psychological research, and (2) contrary to the Bayesian claim, empirical research is indeed conducted to ascertain the truth of the hypothesis (see Chow, 1996; Mayo, 1996, for some critiques of the Bayesian approach).

The *mathematical basis, theory-dependent prediction* and *what statistical significance cannot do* observations (i.e. the aforementioned Observations (1), (2) and (3), respectively) lead Sohn to conclude that statistical significance has no role to play in empirical research because statistical significance is not an index of the replicability of

research results. Moreover, he concludes that a statistically significant effect is neither a clinically important nor a genuine effect. Instead, a genuine effect is found if the treatment effect is `clearly discernible' for the individual on a continuous basis. Another reason for denying a role for the null hypothesis significance test procedure (NHSTP) in empirical research seems to be that prediction and control are not predicated on statistical significance.

Sohn's argument invites the following questions: (1) When does the effect become discernible? (2) How does a theory become well substantiated? (3) Is it warranted to identify the aim of scientific research with prediction and control in the `forecast and shape' sense? (4) Do descriptive statistics suffice to render meaningful or genuine the research outcome? An explication of Sohn's *mathematical basis, theory-dependent prediction,* and *what statistical significance cannot do* observations will provide the answers to Questions (1) through (4). The explication is implicit in Sohn's truth-validity distinction observation. It will be shown that not making good use of the distinction leads Sohn to characterize incorrectly Ho, Ht and the meaning of `statistical significance'. His dismissal of NHSTP is unwarranted.

The truth-validity distinction may be illustrated with the study of the antithrombotic effect of aspirin mentioned by Sohn with reference to Table 1. Consider first the pharmacological phenomenon that the synthesis of prostaglandins forms blood platelets. Blood circulation is hindered when the cardiovascular system is blocked by these platelets. Being acetylsalicylic acid, aspirin inhibits the synthesis of prostaglandins (see Row 1 of Table 1). This account is `well substantiated' in Sohn's terms, but only at the physio-pharmacological level. The issue is whether or not it is also well substantiated at the clinical level.

It is necessary to test clinically the pharmacological account of aspirin's efficacy because other chemical agents, various life-style variables or psychological factors may act on either the acetylsalicylic acid or the prostaglandins at the neuropsychopharmacological or clinical level. That is, the substantive hypothesis is the

yet-to-be-substantiated clinical hypothesis `Aspirin reduces the blockage by blood platelets of the cardiovascular system ` (see Row 2 of Table 1). Note that this hypothesis is one about aspirin's antithrombotic effect on the cardiovascular system in general, not about the state of health of the cardiovascular system of any particular individual. Seen in this light, it is not clear how the `discernible effects for individual' solution to the *replicability belief* issue may be justified. It is also important to note that the clinical hypothesis is not tested directly. Instead, an implication of the clinical hypothesis is first derived, namely `Aspirin promotes the health of the cardiovascular system' (see the consequent of [P 1] in Row 3 of Table 1).

TABLE 1. The instigating phenomenon, clinical, research and statistical hypotheses that underline the aspirin study

	Level of discourse	**What is said at the level concerned**
1	The physio-pharmacological phenomenon	Acetylsalicylic acid inhibits the formation of blood platelets brought about by the synthesis of prostaglandins.
2	The clinical hypotheses	Aspirin reduces the blockage by blood platelets of the cardiovascular system.
3	The clinical hypotheses and its implication	If the clinical hypothesis is true, then aspirin promotes the health of the cardiovascular system. [P1]
4	Statistical hypotheses	(a) Null (H_O): The proportion of MI is the same for both the experimental and control groups (b) Alternative *(H_1):* The proportion of MI is lower in the experimental than in the control group.
5	The chance and null hypotheses	If chance influences are responsible for the data, then H_O. [P2]
6	The sampling distribution	If H_O, then the test statistic (χ^2) has a sampling distribution that is approximated by the chi-square distribution with *df = I . [P3]*

It is understandable why psychologists take the *replicability belief* seriously when they (a) characterize `Aspirin promotes the health of the cardiovascular system' as the *prediction* of the clinical hypothesis, and (b) conduct research in order to discover the means to predict (i.e. in the sense of forecasting) and control phenomena (in Skinner's [1938] sense of shaping or constraining phenomena). It is not possible to forecast or shape future events if the probability of repeating the earlier result is not known. However, it is

problematic to treat `predict and control' as though it is synonymous with `forecast and shape' in methodological discussions.

The number of myocardial infarction (MI) incidents may be used as both (a) an index of the health of the cardiovascular system, and (b) the criterion of rejection for the clinical hypothesis. As may be seen from the conditional proposition [P1] in Row 3 of Table 1, the relationship between the clinical hypothesis and `Aspirin promotes the health of the cardiovascular system' is not predictive (the common misleading characterization notwithstanding), but prescriptive in the theoretical sense. That is, the researcher is not forecasting what will happen in the future on the basis of the clinical hypothesis. Instead, the researcher uses `Aspirin promotes the health of the cardiovascular system' as the criterion to decide whether or not the clinical hypothesis is tenable. Specifically, it prescribes that the clinical hypothesis cannot be **true if the** health of the cardiovascular system is not promoted by using aspirin.

If the correct `prescription' characterization is used instead of `prediction', the *replicability belief* issue becomes irrelevant to the tenability of the clinical hypothesis. What is important for substantiating a theory or hypothesis is not replicability, but Lykken's (1968) constructive replications (as noted by Sohn) or Garner, Hake and Efiksen's (1956) converging operations. Specifically, the theory is true only to the extent to which it has survived concerted attempts to falsify it by a properly designed and executed series of converging operations (Chow, 1989, 1991, 1996).

What is said about the health of the cardiovascular system in the consequent of [Pl] is not amenable to quantitative description, let alone statistical treatment. Consequently, it has to be expressed in the appropriate form with reference to how the data are collected in the study. Specifically, in the case of the aspirin study in question (Steering Committee of the Physicians' Health Study Research Group [SCPHSRG], 1988), a group of physicians was divided randomly into the experimental and control groups. Physicians in the experimental group received an aspirin tablet every other day and those in the control group received a placebo tablet under an identical regime over a

five-year period. The dependent variable was the proportion of participants who suffered from myocardial infarction.

To formulate the statistical alternative hypothesis *(H_1)* is to represent the implication of the clinical hypothesis (viz. the health of the cardiovascular system) in terms of the data collection conditions and the dependent variable (see (b) in Row 4 of Table 1). It may be seen readily that H_1 is not the clinical hypothesis even though it is derived from the clinical hypothesis (see Chow, 1996, for a more detailed discussion). In view of the direction of the clinical hypothesis, the statistical alternative hypothesis *(Hi)* reads as follows:

H_1: MI Proportion$_{aspirin}$ < MI Proportion$_{placebo}$.

This hypothesis may be assessed with the test statistic χ^2. However, H_1 is not specific enough for arriving at the χ^2 statistic because the proportion of participants suffering from MI to the proportion of participants not suffering from MI in the `Aspirin' condition is not specified in the clinical hypothesis. (See Chow, 1996, for the reason why this is not a liability of the hypothesis.) It is for this reason that the appeal is made to the null hypothesis that the proportional frequencies in question are determined by chance influences. In the context of the `Aspirin' study, `chance influences' means that the independent variable (viz. *Medication)* and outcome variable (i.e. whether or not an individual suffered from MI) are independent. Consequently, equal proportions are expected for the experimental and control conditions as follows:

H_O: MI Proportion$_{aspirin}$ ≥ MI Proportion$_{placebo}$.

To properly appreciate **NHSTP** in general, and the meaning of `statistical significance' or Ho in particular, it is necessary to note that two conditional propositions, [P2] and [P3], are implicated in the appeal to H_O, namely

[P2]: If chance influences are responsible for the data, then H_O is true. (Row 5 of Table 1)

[P3]: If H_O is true, then the test statistic (χ^2) has a sampling distribution that is approximated by the chi-square distribution with $df = 1$. (Row 6 of Table 1)

It may be seen from Proposition [P2] that H_O is the implication of the hypothesis about the putative influences of chance factors on the data collection procedure, and that [P3] stipulates the sampling distribution to use in carrying out NHSTP. In other words, Propositions [P2] and [P3] are used jointly to assess whether or not the data can be explained solely in terms of chance influences. Hence, that NHSTP has nothing to do with the health of the cardiovascular system or the replicability of research results does not mean that it has no role to play in empirical research. To say that the result is statistically significant is to say that, given the pre-determined level of stringency (viz. the α level), there is sufficient reason to exclude chance influences as an explanation for the data.

The condition in which NHSTP is valid may be seen by considering the methodological requirement that the experimental and control groups be identical in all aspects but one (viz. the research treatment they received). Crucial to this requirement for the two-group design used in the SCPHSRG (1988) report is that participants were assigned randomly to the experimental and control groups. This procedural feature is necessary for the statistical assumption that measurement errors are distributed equally between the two groups *in the long run* (or `infinitely many replications' in Sohn's terms). That is, this is the key assumption that underlies Propositions [P2] and [P3].

Be that as it may, measurement errors may not be distributed equally in any particular study, as a result of chance influences. The issue is how much departure from the equal distribution of errors between the two groups can be tolerated before the *chance influences* hypothesis ceases to be credible. It is for this reason that the sampling distribution of the test statistic predicated on chance influences is used to determine the associated probability of the outcome of the particular study (i.e. [P2] and [P3]). Specifically, the convention is adopted that chance influences cannot be excluded if the associated probability of the test statistic is larger than the predetermined a value. In such

an event, the effect of the manipulation is not discernible in the sense that it is not possible to say whether the result is due to the research manipulation or chance influences. On the other hand, a significant result indicates a discernible effect in the sense that chance influences are excluded with reference to a well-defined criterion. By the same token, a significant effect is a genuine effect in the sense that it is not brought about by chance factors. Important to the present discussion is that it is not clear how this function of excluding the *chance influences* explanation can be achieved by using descriptive statistics.

As Ho is the implication of the *chance influences* hypothesis, *H,* is the implication of the hypothesis that some non-chance factor is responsible for the data. This characterization of *H,* is different from the common practice of identifying it with the substantive hypothesis (or the clinical hypothesis in the present discussion). For this reason, statistical significance is not informative as to which non-chance factor it may be. At the same time, although the significant result (in the correct direction) affirms the consequent of Proposition [PI] in Table 1, deductive logic does not permit any definite conclusion about the clinical hypothesis (COPI, 1982). This is the case because a healthy cardiovascular system may be brought about by some mechanism or factor other than the research manipulation (e.g., a healthy diet). That is, there are alternative explanations for the data, either methodological or conceptual. How can the clinical hypothesis be substantiated under such circumstances?

The simple answer is to exclude as many of these recognized alternative explanations as possible. Some of them are excluded by the inductive principle that underlies the design of the experiment (see Chow, 1991, 1996). For example, the personal history of MI may be excluded as an explanation because it was a control variable in the `Aspirin' study in the form of a participant-selection criterion. The more alternative explanations are excluded, the clearer the meaning of the data. The important point is that NHSTP is not meant to be used to serve this function of excluding alternative theoretical explanations.

Other alternative explanations may have to be excluded by conducting additional studies. In such an event, different independent, dependent and control variables, different experimental tasks, and different sets of Propositions [P1], [P2] and [P3] may be implicated in these converging operations. Nevertheless, NHSTP is used in every one of these additional studies to exclude the *chance influences* explanation (Chow, 1989). The hypothesis becomes substantiated when it survives the converging operations. That is to say, research data owe their meanings to the theoretical foundation of the research. At the same time, the theoretical foundation is outside the domain of statistics. In other words, questions about the truth or meaning of the hypothesis are not statistical, but conceptual and logical in nature. That NHSTP has nothing to do with truth or the meaning of the data should not detract from its important role in empirical research (of excluding chance influences on the data collection procedure).

The `control' component of Sohn's `predict and control' view of science is predicated on interpreting `predict' to mean making a forecast. It has been shown that the `forecast' meaning is incorrect because the implication of the to-be-tested hypothesis is a theoretical prescription of what *should* happen if the hypothesis is true, not a forecast of what will happen. It follows that the `control' component of Sohn's view is also debatable because, as has been noted by Boring (1954, 1969), `control' does not mean to shape or to constrain the to-be-studied phenomenon in methodological discussions.

The three correct meanings of `control' are (a) a valid comparison baseline line, (b) two means of achieving constancy of condition, and (c) the provision for excluding a procedural artifact. Moreover, the two means of achieving the constancy of condition are (a) using properly chosen control variables (i.e. variables that are being held constant), and (b) using all the
levels of the independent variable consistently throughout the study as prescribed in the design. To institute these control measures is to fulfill the formal requirement of an inductive principle in order to exclude alternative explanations (Chow, 1991, 1996). This is the reason why control is important to scientific research.

The result of the `Aspirin' study is not convincing if the issue of controls is given due consideration. The numerical difference between 189 (the placebo group) and 104 (the aspirin group) cases of MI is ambiguous because (1) the number of participants in each of the two groups is not known at the time the report was made, (2) there is no indication that the two groups were comparable in terms of life style because life style was neither a participant-selection criterion nor a control variable in the study (e.g. diet, physical condition, smoking or drinking habit, etc.), (3) it is not known whether or not the two groups were comparable in terms of age (which ranged from 40 to 84), and (4) the effectiveness of the double-blind control is suspect because the nature of the control tablet is not known.

Consider the double-blind procedure used in the SCPHSRG study. It is not clear from the report whether or not the control and aspirin tablets produced the same side-effects if taken with an empty stomach. In the absence of such a control feature, the participants would know whether they were in the experimental or control group. They might behave differently as a result of the reactivity effects. To complicate the issue even more, what was actually used was not the two-group design, but the 2 x 2 factorial design. However, the SCPHSRG report was written as though the two-group design was used, and there was no report as to whether or not there was any interaction between the two types of drug used at the clinical level (viz. aspirin and beta carotene). It is also debatable whether or not it is legitimate to terminate the study in the fortuitous manner reported.

It is true that the placebo and aspirin groups had 11,037 and 11,034 participants, respectively, at the start of the study. However, what the exact numbers were at the conclusion of the study is not known. All is known is that `[after] an average of 57.0 months of follow-up, 87.6 percent of the **participants are still . . . reporting that** they take at least one of their types of study pills, and 83.0 percent that they take both types of pills' (SCPHSRG, 1988, p. 262). It does not follow that the two groups had the same number of participants at the end of the study. For all we know, the difference between

104 and 189 MI cases for the aspirin and placebo groups, respectively, may be very small (if not zero) in proportion terms.

In conclusion, Sohn succeeds in showing what statistical significance does not mean and what NHSTP cannot do, notably that it is not an index of the replicability of the research results. However, his conclusion that statistical significance has no role to play in empirical research is debatable because (1) NHSTP is the indispensable tool for assessing whether or not chance influences can be excluded as an explanation of the data, and (2) it is unreasonable to use NHSTP to deal with non-statistical issues.

References

Boring, E.G. (1954). The nature and history of experimental control. *American Journal of Psychology, 67,* 573-589.

Boring, E.G. (1969). Perspective: Artifact and control. In R. Rosenthal & R.L. Rosnow (Eds.), *Artifact in behavioral research (pp. 1-11).* New York: Academic Press.

Chow, S.L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin, 106,* 161-165.

Chow, S.L. (1991). Rigor and logic: A response to comments on `Conceptual Rigor'. *Theory & Psychology, 1,* 389-400.

Chow, S.L. (1996). *Statistical significance: Rationale, validity and utility.* London: Sage.

Copi, I.M. (1982). *Introduction to logic* (6th ed.). New York: Macmillan.

Garner, W.R., Hake, H.W., & Eriksen, C.W. (1956). Operationalism and the concept of perception. *Psychological Review, 63,* 149-159.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-159.

Mayo, D.G. (1996). *Error and the growth of experimental knowledge.* Chicago, IL: University of Chicago Press.

Skinner, B.F. (1938). *The behavior of organisms.* New York: Appleton-Century.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology, 8,* 291-311.

Steering Committee of the Physicians' Health Study Research Group (1988). Preliminary report: Findings of the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 262-264.