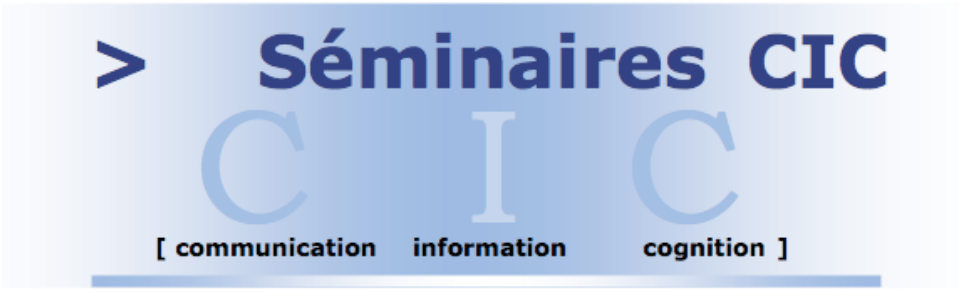

http://eradec.teluq.uquebec.ca/

Statistical decision and falsification in science: going beyond the null hypothesis !

Dominic Beaulieu-Prévost

Statistical decision and falsification in science : going beyond the null hypothesis !

Dominic Beaulieu-Prévost – Chercheur au Département de Psychologie
Université de Montréal
Contact: dbprevost@hotmail.com
Site: www.projetmemoire.info

Résumé

Le processus de décision est fondamental en sciences cognitives, non seulement en tant que thème de recherche, mais aussi comme élément central du processus de validation empirique des hypothèses. Dans la très grande majorité des recherches, la décision d'appuyer ou non une hypothèse s'argumente autour d'un modèle probabiliste de décision qui prend la forme d'un test de signification statistique. Ce modèle, basé sur le rejet ou non de l'hypothèse nulle, est utilisé depuis plus de 50 ans. Par contre, des critiques de plus en plus nombreuses (i.e. plus de 300 articles) soulignent des failles majeures l'invalidant comme modèle de décision. Et malgré ces critiques, le status quo semble toujours être maintenu… Cette présentation en deux parties tente de faire le point sur la problématique de la décision statistique en sciences cognitives. Dans la première partie, (a) le modèle actuel est explicité et présenté dans le contexte plus global de la validation d'hypothèses, puis (b) les critiques et les problèmes majeurs sont présentés, particulièrement en lien avec le principe de falsifiabilité qui est au coeur du processus de décision en science. La deuxième partie propose une variation très simple du modèle de décision statistique qui permet d'éviter les problèmes majeurs du modèle traditionnel et d'améliorer la pertinence des interprétations découlant des décisions statistiques, entre autres en remplaçant les hypothèses ponctuelles par des hypothèses d'intervalle.

The empirical evaluation of hypotheses is a crucial element of the research process in the social and cognitive sciences. In most published studies, this evaluation process is elaborated around a probabilistic decision model taking the form of a test of statistical significance. In a way, the statistical procedure IS the decision process. This model, based on the rejection or not of a null hypothesis, has been used for more than 50 years. However, a growing number of criticisms (i.e. more than 300 articles) pinpoint major problems that invalidate it as a decision model and question its capacity to answer most of our research questions. And despite these criticisms, the status quo was maintained, until recently... In fact, these criticisms started to bring changes in the scientific community about a decade ago.

Recently, the *APA Task Force on Statistical Inference* seriously thought about banishing tests of significance from all APA journals (Wilkinson and the Task Force on Statistical Inference 1999). Even though they decided not to do it, they still recommended a reduction of the use of tests of significance et suggested to use of more useful methods (e.g. effect sizes, confidence intervals, bayesian estimations,...). Some editors of scientific journals are also starting to follow these

recommendations (e.g. Memory & Cognition). This gradually makes it harder to publish articles using traditional statistical methods. Furthermore, a growing number of authors present and popularize alternatives to significance testing (e.g. Kline 2004; Tryon 2001). An indepth reform of the methods of statistical inference is really emerging and it can only improve the quality and pertinence of the scientific publications.

The purpose of this chapter is to provide a comprehensive summary of the problem of statistical decision in science and to present the necessary conceptual tools to face the demands of the reform.

I. The traditional model of statistical decision

According to the traditional model of statistical decision, two competing hypotheses have to be defined: The null hypothesis (H0), which states that there is no effect, and the alternate (and generally more interesting) hypothesis (H1), which states that there is an effect. Depending on the research context, an effect might take the form of a difference between two groups or of a correlation between two variables. As Fisher (1925) taught us, we cannot prove a theory but we can refute it. Consequently, the goal of the procedure is to evaluate whether or not H0 can be confidently rejected. To represent the degree of statistical uncertainty that one agrees to deal with for the decision, a probabilistic decision criterion (i.e. the alpha) is chosen. Traditionally, the alpha is fixed at 0.05 (i.e. 5%). The decision to reject or not H0 is then made by comparing the p value (calculated from the data) to the alpha. Ideally, the risk of making a type II error (i.e. not rejecting H0 when it should have been rejected) is also estimated through an evaluation of the statistical power of the design. Because of space constraints, readers interested in the details of the procedure are encouraged to consult an introductory book on statistics in social sciences (e.g. Gravetter and Wallnau 2003).

This procedure, called significance testing, is actually used in most studies as the heart of the process by which scientific hypotheses gain (or loose) credibility from an empirical test. Indeed, it is still probably taught to most undergraduate students in social science as THE standard procedure that has to be used to empirically evaluate a hypothesis. However, recent surveys (e.g. Lecoutre, Poitevineau and Lecoutre 2003) have shown that even researchers do not agree on the correct way to interpret tests of significance. Some of the most popular beliefs are that it allows an evaluation of the probability:

a) that H0 is false;

b) that the data are the result of chance (i.e. that H0 is true);

c) of doing a type I error if H0 is rejected (i.e. rejecting H0 when it should not have been rejected);

d) that an experimental replication produces statistically significant results (by calculating 1-*p*);

e) that the decision (to reject H0 or not) is correct;

f) to obtain results as extreme as these if H0 is true.

As surprising as it might seem to many users of significance testing, only (f) is a valid statement. Using similar statements, it was found that only 11% of a sample of academic psychologists interpreted tests of significance adequately (Oakes 1986). One of the reasons why many people mistakenly believe that some of the first five statements are true is that they confound the probability of the data given the hypothesis (p(D/H)) and the probability of the hypothesis given the data (p(H/D)). While the p value gives the probability of the data given that the null hypothesis is true (p(D/H0)), many researchers are tempted to interpret it as the probability that the null hypothesis is true given the data (p(H0/D)) which corresponds to statement (b). However, p(D/H) and p(H/D) are not equivalent. Basically, statements (a), (c), (d) and (e) can be considered as variations of the same mistake. Although the non-equivalence of p(D/H) and p(H/D) can be demonstrated mathematically from Bayes' theorem, the following examples might provide a fast and easy demonstration. Although the probability that a man is dead given that he was hung (p(Dead/Hung)) is very high, the probability that a man was hung given that he is dead (p(Hung/Dead)) is not. Using a similar logic, the probability that a human is a female given that s/he is pregnant is not equivalent to the probability that a human is pregnant given that she is a female.

II. Major problems with the model

Although misinterpretations of significance testing seem to be quite common among researchers, it is still not a strong argument against the use of significance testing. Indeed, as some authors have argued, this problem is a "human factor" problem, not a methodological one. However, it is not the only major problem related to the use of significance testing. Three of the most important problems of significance testing, i.e. (1) the relation to sample size, (2) the logical improbability of H0 and (3) the lack of plausibility, will be explained in the following paragraphs.

II.I. The relation to sample size

One of the major problems of significance testing is that the *p* value is not only related to effect size, but also to sample size. More specifically, the *p* value becomes smaller as the effect size increases and as the sample size increases. A direct consequence of this relation is that for a specific effect size, the *p* value is an index of sample size. A more problematic consequence is that a statistically significant result will ALWAYS be obtained if the sample is big enough, unless the effect size is EXACTLY zero. An irrelevant effect can thus be highly significant just because of sample size. Politically speaking, the *p* value is also an indirect measure of funding simply because highly funded research teams have more resources available to insure that their samples are big enough to produce statistically significant results.

II.II. The logical improbability of H0

A second major problem is related to the non-equivalence of H0 and H1. Indeed, while there is only one specific value associated with H0 (i.e. zero), there is a range of possible values associated with H1 (i.e. anything except zero). H0 is thus said to be a point hypothesis while H1 is a range hypothesis. The main problem with point hypotheses is that they are logically improbable on a continuous scale. For example, if I want to test the hypothesis that I will get a "1" the next time I roll a six-sided die, there is a logical probability of one chance out of six that the hypothesis is true because there are six possible events. However, the number of possible values on a continuous scale is infinite. Consequently, there is only one chance out of the infinite that a specific point hypothesis is true. When defining point hypotheses as a special case of range hypotheses (i.e. hypotheses with the smallest possible range), the problem can be summarized by the following statement: *The precision of a hypothesis limits its logical probability of being true*. Indeed, the hypothesis that I will get an odd number the next time I roll a six-sided die has three times more chances of being true than the hypothesis that I will get a one. Thus, restricting the range of possible values for a hypothesis reduces its probability of being true. If the logic is applied to significance testing, the nonsensical nature of the approach becomes obvious. As a point hypothesis on a continuous scale, H0 is ALWAYS false, since $1/\infty$ can clearly be considered a negligible probability. Indeed, the probability that an intervention has an effect size of EXACTLY zero is infinitesimal. Therefore, H1 is ALWAYS true and the concepts of type I and type II errors are totally useless!

II.III. The lack of plausibility of H0

A third major problem with significance testing is the lack of plausibility of the null hypothesis, especially in the "soft" sciences. This notion called the crud factor (Meehl 1990), can be summarized by the following statement: *In the sciences of the living (i.e. from biology to sociology), almost all of the variables that we measure are correlated to some extent*. H0 is thus rarely plausible. It is important to specify that the crud factor does not refer to random sampling error nor to measurement error. As Meehl (1997) states it:

> The crud factor consists of the objectively real causal connections and resulting statistical correlations that we would know with numerical precision if we always had large enough samples (e.g. a billion cases) or if we had measured all of the members of the specified population so that no sampling errors [...] remained.

For example, almost all of the items of the MMPI are statistical correlates of gender when a sufficiently large sample is used (Pearson, Swenson and Rome 1965). A resulting consequence of the situation is that the emergence of a statistically significant effect cannot be claimed as support for a specific theory because a whole class of theories could also explain such an effect. This is the empirical equivalent of what is known in formal logic as the error of affirming the consequent. Even if, according to hypothesis X, there is a positive correlation between variables A and B, the fact that there is indeed a positive correlation between A and B does not lead to the logical conclusion that X is true.

II.IV. The consequences…

Using significance testing to appraise the validity of a scientific hypothesis implies using a decision criterion that confounds effect size and sample size to test a hypothesis that we already know is false and unrealistic. And when we successfully reject this false hypothesis, we wrongly infer that this test improves the plausibility/credibility of our "scientific" hypothesis. It's nothing more than trying to boost our confidence in our cherished hypothesis by rejecting an unrealistic and "known to be false" hypothesis.

III. Is there a way out?

As suggested earlier, significance testing is a dead end as a decision model. The purpose of this next section is to propose a way out of this dead end without having to relearn everything from scratch.

III.I. The falsification principle

We can never prove a theory although we can refute it. This statement that summarizes the limits of inductive inference is used since Fisher to justify the logic of significance testing. Since we cannot prove H1, we'll do our best to refute H0. And it could have been an interesting idea if H0 was not already known to be false and unrealistic! It is basically correct to argue that a statement cannot be inductively proven but that it can be refuted, but it is useless to empirically test a statement's truth value when it is already known. To understand how the logic of inductive inference can be adequately applied to the empirical evaluation of hypotheses, one has to go back to Popper's (1963) falsification principle. Popper's approach to theory appraisal basically argues that science proceeds by setting up theories and then attempting to disprove or falsify them. Theories that are continually resistant to falsification are accepted as "possibly true" while those that are falsified at one point are rejected as false. Thus, it is not so much the falsification (or not) of a theory that makes it scientifically interesting but its continual resistance to attempts to falsify it. However, for the empirical test to be valid, the tested theory (or hypothesis) has to be falsifiable, i.e. hypothetical situations that would falsify the theory's predictions have to exist.

The falsification principle was first used to criticize psychoanalysis and marxism as unfalsifiable because they could explain every possible situation (Popper 1963). A similar criticism can also be applied to significance testing. As we have seen above, H1 is always true because it includes the whole continuum of possible results (except one point). Indeed, if we fail to reject H0, we can always claim that the sample was not big enough. H1 is thus unfalsifiable, which makes it a scientifically worthless hypothesis. By the same token, significance testing as a model of theory appraisal can only be seen as a scientifically useless procedure. However, the problem of significance testing is not so much in the

statistical principles used to evaluate the probability of an event but in the specific hypotheses that are systematically tested (i.e. H0 and H1).

III.II. Constructing scientifically useful hypotheses

If we summarize, a scientifically useful hypothesis has to be probable, plausible and falsifiable. All point hypotheses (e.g. H0) are thus scientifically useless since they are improbable to the point of being false. Hypotheses that include every possible result except one (e.g. H1), are also scientifically useless since they are unfalsifiable. Consequently, the only scientifically useful hypotheses are range hypotheses that both include and exclude a significant amount of possible results.

Even though an infinity of possible range hypotheses could be constructed, most scientifically meaningful hypotheses can be summarized by of one of the following types: (1) There is (or not) a substantial effect, (2) There is (or not) a harmful effect and (3) There is (or not) a trivial effect.

To understand the meaning of these types of hypotheses, the notion of substantial effect has first to be clarified. Basically, the concept of "substantial effect" is the equivalent of "clinically significant effect" although it is not limited to clinical settings. A substantial effect is simply an effect whose size is large enough to be of interest. However, it is important to mention that the minimal value of a substantial effect is always context-dependent. To adequately quantify the minimal value of a substantial effect (or the maximal value of a trivial effect), one has to assess the important aspects of the study such as the theoretical importance of the effect, the practical purpose of the phenomenon, the potential cost of an intervention and, minimally, the sensitivity of the scale. For example, if the effect of an intervention on depression is measured with a depression scale from 1 to 10, it might be decided that an effect size of one would be the smallest interesting value since it is the smallest possible difference that can be detected by the scale. However, it the intervention is extremely costly, it might be decided that the effect size would need to be of at least 2.5 for the intervention to be interesting. Two different minimal values can often be quantified for the same study: The minimal value to consider that an effect is theoretically interesting and the minimal value to consider that an effect has practical applications. For example, if you are interested to investigate a potential link between self-esteem and school performance, you might be satisfied with correlations of 0.09 (i.e. 1% of explained variance) or more, but if you plan to increase school performance through a large-scale self-esteem intervention, you might evaluate that only correlations of at least 0.30 (i.e. 9% of explained variance) are deemed to be interesting. An advantage of having to define the minimal value of a substantial

effect is that it forces researchers to take into account the purpose of their study because such a value cannot be defined for meaningless studies.

As soon as the minimal value of a substantial effect is defined, the three possible types of hypotheses can automatically be defined:

1) *The hypothesis of a substantial effect* evaluates whether or not the effect is at least equal to the minimal value of the substantial effect.
2) *The hypothesis of a harmful effect* is defined as the opposite of the hypothesis of a substantial effect. It can be used to evaluate the possibilities of a harmful or counter-intuitive effect of substantial value.
3) *The hypothesis of a trivial effect* evaluates whether or not the effect is between the minimal substantial effect and the minimal harmful effect. When this hypothesis is tested for a comparison between two means, it is also called a test of equivalence (see Rogers, Howard and Vessey 1993) since it evaluates whether or not the means are substantially different.

III.III. Testing the hypotheses

Although significance testing is a highly inadequate procedure to evaluate the scientific validity of a hypothesis, most of what is normally taught in an undergraduate statistics course is still valid and extremely useful. In fact, what most critics of significance testing reject is not the validity of the statistical principles but the pertinence of automatically and exclusively testing the null hypothesis (Kline 2004; Cohen 1994). Indeed, nearly all of the proposed alternatives to significance testing are based on the same probabilistic model. The solution presented in this chapter, which uses confidence intervals instead of tests of significance, is no exception.

Understanding confidence intervals

Confidence intervals are mathematically equivalent to tests of significance. Indeed, for every test of significance, an equivalent confidence interval can be constructed. However, instead of providing a *p* value to evaluate if an effect is statistically different from zero, confidence intervals provide information about the effect size in the sample and the precision of the parametric estimation of the effect size. The basic model of an effect size is CI = S +/- SE * Cv, where the confidence interval (CI) is constructed by adding and subtracting from a statistics (S) the product of its standard error (SE) and the two-tailed critical value at the

chosen alpha level of statistical significance (Cv). Every value around the effect size and between the upper and lower limits of the interval is included in the confidence interval. When the CI excludes zero, the equivalent test of significance is statistically significant and vice versa.

A confidence interval can be conceptually defined as a range of plausible values for the corresponding parameter. We could also say that conclusions that a parameter lies within a CI will err in [corresponding alpha] of the occasions. However, to interpret CIs beyond these simple definitions, one has to clarify what is meant by the notion of probability. Indeed, there are two radically different ways to interpret CIs that are related to two different interpretations of probability. The most commonly taught (but least understood) interpretation comes from the *frequentist approach*. It is indeed the approach on which traditional CIs are based. According to this approach, probability represents a long-term relative frequency. More explicitly, if CIs could be calculated for an infinity of random samples coming from the same population, the parameter of the population would be included in [1-alpha] of them. However, when a single CI is interpreted, it is inadequate to say that there is a probability of 95% that the parameter is included in the CI. From a frequentist point of vue, it makes no sense to speak about probabilities for a specific CI, it either includes the parameter or it doesn't. The only meaning that can be given to a specific CI is as a representation of the amount of sampling error associated with that estimate within a specified level of uncertainty. It is thus said that all the values included in a CI can be considered to be equivalent with a level of confidence of [1-alpha].

Researchers and decision makers are often more interested to know the probability that a specific CI includes the related parameter than to measure the sampling error of their study. What they crave for is the probability from a *subjective approach* or, more simply, a reasonable estimation of the odds of being correct if they conclude that the parameter is included in a specific CI. Using that definition, probability takes place in the eye of the beholder, not in the empirical world. The subjective approach to probability is generally called the *bayesian approach* because it is mathematically based on Bayes' theorem. It is indeed possible to calculate a bayesian CI for which it can be reasonably assumed that there is [1-alpha] chances that the parameter is included. However, to adequately calculate such a CI, one as to take into account both the experiment's data and all the previous knowledge one has about that parameter. It is a process extremely similar to a meta-analysis. There is still one case for which a bayesian CI coincides with its frequentist counterpart: It is when the bayesian CI is based on an agnostic prior, i.e. a judgment that one has no prior knowledge or belief about a parameter's possible value. It can thus be said that when only the experiment's data are taken into account to estimate a parameter (i.e. when an agnostic prior is postulated), a traditional CI represents an interval for which it is reasonable to assume that there is [1-alpha] chances that the parameter is included. By

extension, the distribution related to the CI can be understood as the distribution of the probable values of the parameter according to an agnostic prior.

Testing hypotheses with confidence intervals

As soon as adequate range hypotheses are defined and the CI is calculated, hypothesis testing can be done at a glance! You just have to see if the CI is either (1) totally included within the range of the hypothesis, (2) totally excluded from the range of the hypothesis or (3) partly included within the range of the hypothesis. If the CI is totally included, the hypothesis is *corroborated* (i.e. $p > 0.95$ if alpha = .05), if it is totally excluded, the hypothesis is *falsified* (i.e. $p < 0.05$ if alpha = .05) and if it is partly included, the hypothesis is *undetermined* (i.e. $0.05 < p < 0.95$ if alpha = .05). The notion of undetermination answers the question of statistical power: If a hypothesis is undetermined, it simply means that the sample is not large enough to let the test provide a clear answer. The exact subjective probability associated to a hypothesis can also be calculated although the demonstration is beyond the scope of this chapter.

IV. Conclusion

Significance testing systematically quantifies the plausibility of a "known-to-be-false" hypothesis (H0) to evaluate the validity of an unfalsifiable "known-to-be-true" alternate hypothesis (H1). It is thus useless as a basis for the scientific evaluation of hypotheses. However, researchers can easily evaluate scientific hypotheses if they operationalize them as falsifiable range hypotheses, estimate the minimal value of a substantial effect and construct confidence intervals from their data.

References

Fisher, R. 1925. *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Gravetter, F. J. and Wallnau, L. B. 2003. *Statistics for the behavioral sciences, 6th ed.* Belmont, CA: Wadsworth Publishing.

Kline, R. B. 2004. *Beyond significance testing.* Washington: American Psychological Association

Lecoutre, M.-P., Poitevineau, P. and Lecoutre, B. 2003. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Test. *International Journal of Psychology 38* (1): 37-45.

Meelh, P. E. 1990. Why Summaries of Research on Psychological Theories Are Often Uninterpretable. *Psychological Reports 66*: 195-244.

Meelh, P. E. 1997. The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In *What if there were no significance tests?*, edited by L. L. Harlow, S. A. Mulaik and J.H. Steiger, 393-425. Mahwah, NJ: Erlbaum.

Oakes, M. 1986. *Statistical inference*. New York: Wiley.

Pearson, J. S., Swenson, W. M. and Rome, H. P. 1965. Age and sex differences related to MMPI response frequency in 25,000 medical patients. *American Journal of Psychiatry 121* (10): 988-995.

Popper, K. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.

Rogers, J. L., Howard, K. I. and Vessey, J. T. 1993. Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin 113* (3): 553-565.

Tryon, W. W. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychoogical Methods 6*: 371-386.

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