HAMBURGER FORSCHUNGSBERICHTE
AUS DEM ARBEITSBEREICH
SOZIALPSYCHOLOGIE
-HAFOS-

# The Stepwise Hybrid Statistical Inference Strategy: FOSTIS

Erich H. Witte & J. Kaufman

HAFOS 1997 NR. 18

Psychologisches Institut I der Universität Hamburg
Von-Melle-Park 6, 20146 Hamburg

# The Stepwise Hybrid Statistical Inference Strategy: FOSTIS

ERICH H. WITTE & J. KAUFMAN

August 12, 1997

Notwithstanding the body of venerable and formidable criticism (e.g., Birnbaum, 1962; Bakan, 1966; Meehl, 1967; Morrison & Henkel, 1970), demand remains unwavering for inferential argument articulated via *significance*. To date there are several possible grounds for this persistence; perhaps foremost of these being a certain objectivity obtained from algorithmic qualification of results. Relieved of the burden of result qualification, research requires only the null hypothesis prediction. In this paper we stand with previous objection to the current standard by arguing that its usage has undesirable effects on theory development and therefore should be modified so that prediction takes a more specific form. Having discussed this and in view of other considerations, an alternative is put forward and discussed.

Let us start with the notion that behavioral science has the job of determining whether or not relationships (orderly effects) exist between operationalized identities. For example, one particular situation might lead us to ask whether cognitive visualization is positively related to intelligence. If the reader generally agrees this to be a reasonable approach, we suggest that a feedback from the inference method is responsible. That is, our perspective to research thinking - call it meta-methodology, is just the other side of how we go about testing our theories. "Knowing" beforehand our method of reckoning, it shows up in the thinking; in the example above, the "whether" corresponds to a search for a "significant" relationship between visualization and intelligence[1].

It is natural, of course, that young research traditions require these questions of whether differences exist, as well as much subsequent sketching out and exploration. But practice as dictated by sole *significance* testing concentrates only on promising results without specification of the parameters to be tested. In one notable example, cognitive dissonance, even after forty years the level of measurement is never used in the specification of hypotheses. Is this a satisfactory development of the ideas being considered? Does our science become more precise?

With these last questions we have in mind two aspects of our science which particularly suffer from the non-specificity of current qualitative inferential methods based solely on the null hypothesis. First, consider the troubles involved in the communication of findings reported by significance difference. The comparison and contrast

[1] As has been pointed out elsewhere, *significance* doesn't even answer this question!

of experimental results is not straight forward, one has to work backwards from the reporting. The current methods afford an accumulation of knowledge at a qualitative level, one which is often incomparable and necessitate the obviously inaccurate and complicated subsequent dredging of meta-analysis. All to answer the question of how variables are related.

And this raises the second point; are we content to argue that a target exists to shoot at or do we have some basic relationships which might guide us around at different targets. That is, if we are seeking to answer how our variables are related, and we have in our possession pertinent findings - we are then ready to ask those more specific questions and develop more specific theories. In sum, while it is indeed inevitable that a mirage of alluring new variables and theories remain on the horizon, it is necessary to have more exact and precise statements and results with which to specify current knowledge.

**0.1. Fostis.** It has been argued that the incorporation of quantitative predictions into our inferential methods will result in more precise theory development. One alternative which facilitates this need is the FOur STep Inference Strategy or FOSTIS[2]. While FOSTIS indeed begins with the precise specification of an alternative hypothesis, as an inference strategy, it is something more.

FOSTIS is an orchestration of several well made contributions (not all of which are well known in the behavioral sciences) chosen to capitalize their strengths on particular tasks. Rather than getting bogged down in difficult arguments already made elsewhere, the plan for the remainder of the paper will be: 1) state guiding principles which offer a general idea for a theory testing strategy, 2) present FOSTIS and highlight its design, 3) consider a specific example with FOSTIS, 4) discuss and offer concluding remarks.

PRINCIPLES:

#1 **Evidential Inference** (likelihood):

Hypotheses are adjudged in light of their felicity with experimental data (evidence); it is the experimental result which has occurred. Determination is not made whether a certain data set is more or less probable under a hypothesis.

#2 **Relative Confirmation**:

Use of a measure of the relative confirmation of one hypothesis against another, not an absolute measure of single hypotheses. As is common, we might refer to a pair of hypotheses as $H_{(0)}$ and $H_{(1)}$.

#3 **Fair Competition:**

[2]FOSTIS first appeared in print in german journals (Witte, 1977 and Witte,1989) as well as in the book *Signifikanz und statistische Inferenz*, Witte, 1980.

The a priori probabilities of hypotheses $H_{(0)}$ and $H_{(1)}$ hypotheses are made equal. The result of this demand is that $H_{(0)}$ and $H_{(1)}$ are of the same kind (e.g., hypotheses are both simple point values along a parameter, or both are intervals of equal lengths).

#4 **Predetermination of Error:**

The acceptable Type I ($\alpha$) and Type II ($\beta$) error rates for $H_{(0)}$ and $H_{(1)}$ are determined and stated at the beginning of research.

#5 **Error Symmetry:**

Type I and Type II error rates are set equal. As both hypotheses are being numerically fixed, either error leads research astray and any incorrect inference must be treated with the same degree of caution.

#6 **Hypothesis Scrutiny** (maximum likelihood):

When hypothesis $H_{(1)}$ has received empirical support relative to hypothesis $H_{(0)}$, the extent to which the confirmation of $H_{(1)}$ depends on the improbability of $H_{(0)}$ must be considered. This is to say, no other hypothesis relative to the alternative hypothesis should be better corroborated by the data than would be expected by predetermined errors.

#7 **Explained Variance Qualification**:

When substantial inferential faith can be placed in a theory and its predictions, it is still necessary to determine and report the amount of explained variance in the data.

Having betrayed the general perspectives in the work, we proceed to the proposal. FOSTIS consists of a plan of four steps, the combined sequence of which takes research on a course from start to finish. After meeting the requirements of the planning elements in step one, FOSTIS makes a *relative comparison* in step two and *scrutinizes* this comparison in step three. If reached, step four asks for *explained variance qualification.*

**Step I** (the empirical condition).

First and foremost we must have two competing hypotheses specified, the easiest case being simple point hypotheses. Exploratory data and/or past research offers a basis for the proposal of a specific point value along the parameter of interest for the hypotheses. It may also be natural to afford this determination by the specification of projected effect sizes (Levy, 1967; Cohen, 1969; McGraw & Wong, 1992).

Suppose that the null hypothesis $H_{(0)}$ and the alternative hypothesis $H_{(1)}$ have been specified. Given this information, we must now have some idea of a general yet reliable plan or condition under which the two hypotheses will be compared. As the theory of Neyman and Pearson offers a firm base, it is prescribed as planning theory.

As a matter of principle and based on the characteristics of our phenomena, we choose a value $\Lambda$ and set

$$\alpha = \Lambda = \beta.$$

The required minimum sample size $n$ is easy to determine (Cohen, 1977) and we are ready to collect the empirical evidence.

**Step II** (Wald's likelihood ratio test).

When the empirical condition is met in the first step and the experiment has been conducted, the two hypotheses are ready to be compared. To test the relative likelihood of the two hypotheses given the experimental result, we form the likelihood ratio and use Wald's criterion[3] (the ratio of power to Type I error):

$$\frac{\mathcal{L}\left(\frac{H_{(1)}}{x}\right)}{\mathcal{L}\left(\frac{H_{(0)}}{x}\right)} > \frac{(1-\beta)}{\alpha}.$$

A likelihood ratio conveys the extent to which one hypothesis is more or less likely than the other, given the experimental results. Likelihoods[4] are closely related to probabilities, though the interpretation of the critical value is somewhat different than in the Neyman Pearson sense. As proved by Wald, the criterion assures us that the alternative hypothesis will pass step two in (1-$\beta$)% of the cases that the alternative hypothesis is true. This test does not need a subjective probability estimation of a hypothesis, because it is a relative confirmation test which eliminates the subjective probability of the hypotheses by testing like hypotheses.

Should the ratio pass the criterion, substantial faith can then be placed in the alternative vis-a-vis the null hypothesis. From a larger perspective however, the alternative hypothesis must be passed on to the step three for scrutiny in terms of all other possible hypotheses. On the other hand, if the ratio fails to pass the test, the alternative hypothesis can not be accepted for further consideration and the analysis terminates by reporting the empirical sample's parameters (mean and variance).

**Step III** (maximum likelihood test).

When the alternative hypothesis has been judged satisfactory with respect to its counterpart, the question remains as to whether that confirmation has depended solely on the improbability of the null hypothesis. To do so, the ratio of *maximum likelihood* (the most likely hypothesis given the data) is tested with respect to the alternative hypothesis:

$$\frac{\mathcal{L}\left(\frac{H_{(\max)}}{x}\right)}{\mathcal{L}\left(\frac{H_{(1)}}{x}\right)} < \frac{(1-\beta)}{\alpha}.$$

Here we are further demanding that the hypothesis which is best supported by the empirical result be no more than $\frac{(1-\beta)}{\alpha}$ times as likely than the alternative hypothesis.

[3]Sequential Analysis, Wald, 1947.

[4]A detailed development of likelihood (a concept conceptualized by Fisher) and its relationship to probability is given in Edwards, 1972.

That is, even if we have good reason to believe in our alternative, we want to make sure that no other hypothesis is that much more likely.

**Step IV.**

The fourth step derives from the discussion about the observed effect size by Cohen (1977). There are many measures of effect size and it is not easy to choose an acceptable one. We might recommend a coefficient of determination of 10%, that is, if the hypothesized difference can determine at least 10% of the variance between variables then the theoretical explanation is precise enough to be an instance for the test of the theory. This criterion has nothing to do with the probability model of statistical inference. It takes into consideration the error of measurement rather than the error of wrong decision between two hypotheses.

It is unfortunate that this type of question is often only asked at the point of meta-analysis and not at each experiment. Normally, this criterion has been used indirectly if the difference, e.g., between the theoretically predicted means, was related to the empirically estimated error variance in a t-test planning strategy at the beginning of the inference procedure.

At the end of the test strategy, we ask whether this assumption is satisfied. One consequence of this criterion is that the hypotheses should be formulated in such a way at the beginning that they are strong enough to be differentiated from a random effect under the testing condition. That is, the measurement of the variables must be precise for the test. Should the enterprise fail to pass this criterion, it might be possible to seek the reduction of error variance and not the alteration of the alternative hypothesis.

**0.2. Example.** Many of our hypotheses are formulated as mean differences and tested by the t-test. First, one must estimate the standard deviation of the measurement. Second, the alternative hypothesis must be precisely stated. Let us consider an example where from past research it is predicted that the difference between null and alternative hypotheses is one half of the standard deviation, $d = 0.5$. As a matter of principle we set $\alpha = .05 = \beta$. With the specification of these parameters, the sample size can be determined in the planning of the experimental condition. For this see Table 2.3.2 in Cohen's useful handbook (pages 28 - 39). A rough approximation for the determination of the sample size is given by the formula

$$n = 2 \times \left( z(1-\alpha) + \frac{z(1-\beta)}{d} \right)^2 .$$

In our example, this produces (after rounding up) a sample size of 88, which corresponds to Cohen's table 2.4.1 (pages 54 - 55).

After the planning of the experiment it is carried out. Suppose for our example that the empirical value found corresponds to a $d = 0.53$, obviously very close to our

prediction. The likelihood ratio of the two hypotheses is determined with $d = 0.00$ (null) and $d = 0.50$ (alternative). There are very many ways to do this. One simple way would be to resolve all problems with the help of the normal distribution and standardized z-values. For this reason, the noncentral t-distribution is transformed into a normal distribution with a theoretically expected mean as a z-value (see Cohen, 1977, p. 456). All these z-values are on the same scale and can be added or subtracted for calculating the probability densities which are equivalent to the likelihoods if ratios are taken into consideration. This is the reason why the tabulated ordinates of the normal distribution are used to calculate the likelihood ratios.

At first the z-value of the theoretical expectation is calculated as the mean of the noncentral t-distribution:

$$z(H_{(1)} \text{ with } d = 0.50) = \frac{\left(0.50 * 87 * \sqrt{(2 * 88)}\right)}{(2 * 87 + 1.21 * (1.65 - 1.06))} = 3.30$$

$$\text{and } z(\text{empirical with } d = 0.53) = 3.50.$$

Now we have to determine the probability densities under the two distributions with $d(H_{(0)}) = 0.00$, $z(H_{(0)}) = 0.00$ and $d(H_{(1)}) = 0.50$ and $z(H_{(1)}) = 3.30$.The ordinate under the first distribution is $\varsigma(H_{(0)}) = 0.0009$ and under the second distribution $\varsigma(H_{(1)}) = 0.3910$. The likelihood ratio test leads to $\frac{3910}{9} = 434.44$. This is greater than $\frac{1-\beta}{\alpha} = \frac{95}{5} = 19$ which is the critical value of the second step. Thus, the alternative hypothesis is significantly confirmed.

The third step of FOSTIS is the qualification of the confirmed hypothesis in light of the hypothesis best confirmed by the data. Thus, the likelihood ratio of the confirmed hypothesis and that hypothesis which is best supported by the data is calculated, as formulated in step three. The maximal ordinate of the normal distribution is $\varsigma(H_{(\max)}) = 0.3989$. The ratio is 1.02 which is less than 19 as the critical value and therefore passes that step.

Finally, the observed effect size of $d = 0.53$ which is greater than the medium effect defined in Cohen(1977), however, leads to a coefficient of determination $r = 0.06$ . Thus, the criterion of the fourth step has not been passed. Because of the hypothetically assumed $d = 0.50$ such a result is not astonishing. This last step gives us an impression of the errors of measurement in the test of our theories.

**0.3. Discussion.** After fixing the Type I and Type II errors in step I of FOSTIS, they are to serve as a base for all criteria. If the experimental condition of step one has been met, but the likelihood ratio in step two fails to reach the critical value, we can either assume that hypothesis was poorly made or that experimentation was flawed. If the alternative hypothesis is passed on to the third step only to fail there, possibly a very rare occurrence, then it would appear that a revision of the alternative

hypothesis and a new test of the revision is in order. This would be the case where the empirical results are much more different from the null hypothesis than predicted.

If all three steps of the testing procedure have been passed but the fourth has been failed, the question is generally whether the theory is useful to explain data in the given context. One consequence is to increase the precision of the measurement or to increase the theory to more specific conditions. The critical value recommended is an amount classified by Cohen (1977) between medium and large. This is a rather strong criterion, but we often forget that our theories are used to explain or predict complex daily events, or results in an experimental setting with complex influences.

Under a principle of parsimony, it is easier to assume that a theory has no influence, than to advance a more complicated explanation with little empirical evidence. This critical value has the technical function of being too limited to empirical results that are not too near to the null effect. If the theoretical effect size is only mediocre, and the empirical results are still smaller, then there comes a point at which the strength of the theory is so modest that it is more acceptable to ignore the theoretical influence postulated.

In general, the main strategy should predict no influence, and accept a theoretically postulated influence only if it can no longer be ignored under the empirical conditions. Our significance tests, however, implicitly follow the strategy of accepting each hypothesis should something not be able to be subsumed under a random effect. There is no limit to the smallness of such influence expressed in a measure of effect size. Cohen's guidelines are very lenient towards the theoretician; a small but acceptable effect explains 1% of the total variance, and what is called a large effect explains only 14% of the variance. This might be one reason why there are so many theories which pass this lenient criterion. It is only necessary to either await a significant result with an extremely high Type II error or to increase the sample size.

**0.4. Conclusion.** We have long seen unwanted consequences with the current significance methods and it is often stated that many are dissatisfied with it (e.g., Pitz, 1978; Rogers, Howard & Vessey, 1993). Here it will have been noticed, however, that this paper has avoided many of the aspects involved in the significance test controversy, which is as old as the discipline of inferential statistics itself. Rather, it is hoped that the principles which are stated will speak for themselves and that the discussion and search for better methods will be furthered by this proposal which stands on the belief that statistical induction depends on precise theoretical deduction.

## BIBLIOGRAPHY

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.

Birnbaum, A. (1962) . On the foundation of statistical inference. *Journal of the American Statistical Association*, 57, 269-306.

Cohen, J. (1969,1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Edwards, A.W.F. (1972). *Likelihood.* Cambridge, MA: Cambridge University Press.

Levy, P. (1967). Substantive significance of significant differences between groups. *Psychological Bulletin*, 67, 37-40.

McGraw, K.O. & Wong, S.P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.

Morrison, D.E. & Henkel, R.E. (Eds.) (1970). *The significance test controversy.* Chicago: Aldine.

Pitz, G.F. (1978). Hypothesis testing and the comparison of imprecise hypotheses. *Psychological Bulletin*, 85, 794-809.

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

Wald, A. (1947, 1967). *Sequential Analysis.* New York: Wiley.

Witte, E.H. (1977). Zur Logik und Anwendung der Inferenzstatistik. *Psychologische Beiträge*, 19, 290-303.

Witte, E.H. (1980). *Signifikanztest und statistische Inferenz: Analysen, Probleme, Alternativen.* Stuttgart: Enke.

Witte, E.H. (1989). Die "letzte" Signifikanztestkontroverse und daraus abzuleitende Konsequenzen. *Psychologische Rundshau*, 40, 76-84.

# HAMBURGER FORSCHUNGSBERICHTE
# -HAFOS-

| | |
|---|---|
| HaFoS Nr. 1<br>1992 | Witte, E.H.: The extended group situation theory (EGST), social decision schemes, models of the structure of communication in small groups, and specific effects of minority influences and selfcategorization: An integration. |
| HaFoS Nr. 2<br>1992 | Witte, E.H. & Schwerm, M.: Technikfolgenabschätzung und Gentechnologie - Die exemplarische Prüfung eines Expertenberichts auf psychologische Konsistenz und Nachvollziehbarkeit. |
| HaFoS Nr. 3<br>1992 | Witte, E.H.: Dynamic models of social influence in small group research. |
| HaFoS Nr. 4<br>1993 | Witte, E.H. & Sonn, E.: Trennungs- und Scheidungsberatung aus der Sicht der Betroffenen: Eine empirische Erhebung. |
| HaFoS Nr. 5<br>1993 | Witte, E.H., Dudek, I. & Hesse, T.: Personale und soziale Identität von ost- und westdeutschen Arbeitnehmern und ihre Auswirkung auf die Intergruppenbeziehungen. |
| HaFoS Nr. 6<br>1993 | Hackel, S., Zülske, G., Witte, E.H. & Raum, H.: Ein Vergleich 1993 berufsrelevanter Eigenschaften von "ost- und westdeutschen" Arbeitnehmern am Beispiel der Mechaniker. |
| HaFoS Nr. 7<br>1994 | Witte, E.H., The Social Representation as a consensual system an correlation analysis. |
| HaFoS Nr. 8<br>1994 | Doll, J., Mentz, M. & Witte, E.H., Einstellungen zur Liebe und Partnerschaft: vier Bundungsstile. |
| HaFoS Nr. 9<br>1994 | Witte, E.H.: A statistical inference strategy (FOSTIS): A non-confounded hybrid theory. |
| HaFoS Nr. 10<br>1995 | Witte, E.H. & Doll, J.: Soziale Kognition und empirische Ethikforschung: Zur Rechtfertigung von Handlungen. |

| | |
|---|---|
| HaFoS Nr. 11<br>1995 | Witte, E.H.: Zum Stand der Kleingruppenforschung. |
| HaFoS Nr. 12<br>1995 | Witte, E.H. & Wilhelm, M.: Vorstellungen über Erwartungen an eine Vorlesung zur Sozialpsychologie. |
| HaFoS Nr. 13<br>1995 | Witte, E.H.: Die Zulassung zum Studium der Psychologie im WS 1994/95 in Hamburg: Ergebnisse über die soziodemographische Verteilung der Erstsemester und die Diskussion denkbarer Konsequenzen. |
| HaFoS Nr. 14<br>1995 | Witte, E.H. & Sperling, H.: Wie Liebesbeziehungen den Umgang mit Freunden geregelt wünschen: Ein Vergleich zwischen den Geschlechtern. |
| HaFoS Nr. 15<br>1995 | Witte, E.H.: Soziodemographische Merkmale der DoktorandInnen in Psychologie am Hamburger Fachbereich. |
| HaFoS Nr. 16<br>1996 | Witte, E.H.: Wertewandel in der Bundesrepublik Deutschland (West) zwischen 1973 bis 1992: Alternative Interpretationen zum Ingelhart-Index. |
| HaFoS Nr. 17<br>1996 | Witte, E.H & Silke Lecher: Systematik von Beurteilungskriterien für die Güte von Gruppenleistungen. |

Die Hamburger Forschungsberichte werden herausgegeben von
Prof. Dr. Erich. H. Witte
Psychologisches Institut I der Universität Hamburg