Hamburger Forschungsberichte

AUS DER ARBEITSGRUPPE

Sozialpsychologie (HaFoS)

A statistical inference strategy (FOSTIS): A non-confounded hybrid theory.

Erich H. Witte

HaFoS, 1994, Nr. 9
Psychologisches Institut I der Universität Hamburg
Von-Melle-Park 6, D-2000 Hamburg 13

Summary

Due to the many misconceptions surrounding the common significance tests, a catalogue of demands to be satisfied by statistical induction is developed. The result of such criteria is a four step hybrid theory of statistical inference (FOSTIS), which is organized hierarchically: Starting with a planning phase (Neyman-Pearson), going on with a loglikelihood-test (Bayes, Fisher, Wald), coming to a maximum likelihood-test (Edwards), and ending with an effect qualification.
Strengths and weaknesses of this approach are discussed. Generally, there is something to be learned: the more meaningful statistical induction should be the more precise theoretical deduction must be.

A statistical inference strategy (FOSTIS): A non-confounded hybrid theory.

Erich H. Witte

University of Hamburg, Germany

Do we need these endless discussions over which is the correct statistical inference strategy? As we learned from Gigerenzer and Murray (1987), they are theoretical models of cognition and decision. According to this perspective, however, the discussion should be limited to this particular area of research and should be further discussed and criticized without having a fundamental influence on other areas of psychological research. In connection with this research, different statistical inference theories are rediscovered as alternatives, sometimes with very different conclusions or probabilities of corroboration. If we take the entire body of psychological research as a kind of decision process guided by the wrong or only defective theory of inference, or that single specific hybrid theory of statistical inference that is dominant (Cohen,1990; Gigerenzer & Murray,1987), then all of our psychological knowledge evaluated by the use of this judgmental criterion might be defective. At best, we cannot decide whether or not our theories are corroborated by the empirical data, because the acceptance of a true alternative hypothesis , the power of a test, is much more interesting for theoretical development than the probability of rejection of a true null hypothesis, assuming that the alternative hypothesis is the theoretically relevant expression. Since it is known that the power of our statistical tests is on the average near $1-\beta = 0.50$ (Cohen, 1962,1990, Sedlmeier & Gigerenzer,1989, Witte,1980) the acceptance of the true alternative hypothesis turned out no better than had we flipped a coin. In general and without change, our experiments have been unacceptably underpowered since the publication of Cohen's handbook on power analysis (1969). If this could happen for 25 years without change, this criterion of evaluation is a massive hint that something must be

wrong with our theory testing procedure. What of our knowledge in the behavioral sciences is true, if almost all of what has been corroborated is evaluated by underpowered inference strategies?

What remains astounding is the inference revolution in the behavioral sciences in the observed way. How could this have happened? There is a combination of reasons. Most of them have been discussed before, so there is no need to discuss them here intensively again (Gigerenzer & Murray, 1987 ; Witte,1989):
a) the procedure is easy to teach, b) you always get a decision, c) you only need small samples, and d) the procedure is mechanical.
All these reasons are true, I think, but they cannot explain why the massive critique has had no influence at all.

In my opinion, the main reason is that this kind of inference strategy is just the other side of the coin of theory construction. They correspond to one another exactly. The kind of hypotheses and their tests filtered through an inferential test strategy rely on each other. If we are satisfied with the hypotheses usually formulated, then there is no better inference strategy than the hybrid test theory described in our books on statistics. Our discussion of the significance test , the significance test controversy (Morrison & Henkel,1970), is a surface-level discussion; the fundamental problems are deeply rooted in the kind of theories and hypotheses accepted as scientific by the scientific community. Thus, if the inference strategy is felt to be highly problematic, then the quality of the theory construction is insufficient, because the inferential strategy of the classical significance test is the ideal instrument for the scientific feeling:

a) the significance test controls the influence of a random factor, our scientific demand;
b) it ignores other data from earlier experiments, the simplicity of data analysis;
c) it leads to a high rate of reinforcement for the theoretician, because he/she needs only to assume that something happened; and if the sample size is not too small, a signi-

ficant difference will be observed, usually, a corroboration of a theoretical hypothesis;

d) to formulate a theory one needs only a prediction of a very rough qualitative difference without any idea of its amount found in earlier studies, theoretical economy.

On the one hand, this strategy is doubtless an improvement over strategies that preceded the inference revolution. However, it can also act as a stagnation if we continue to ignore the fundamental problems of the general strategy employed to prove our hypotheses. Since psychology and other behavioral sciences have developed, in a pre-paradigmatic state a unifying disciplinary submatrix through the use of an accepted inference strategy, it is nearly impossible to formulate more precise hypotheses, as they cannot be tested by our significance tests in the usual way. Thus, the state of the art in behavioral sciences requires vague hypotheses, because they have to be tested by significance tests. All theoretical progress depends upon our significance tests as a result of their dominance as the fundamental approach to evaluating theories (in our empirical research).

Under such a condition it is necessary to accept that a critique of the significance test is also a critique of those theoretical constructions formulated in such a way that they only can be tested by a significance test. A change of the inference strategy requires a change of the theoretical construction. Change must occur simultaneously in both. This is a long process, though, because it looks to some extent like a scientific revolution. If it is generally accepted at this point that our hypotheses cannot become more precise, then significance tests must be accepted as the ideal methods. However, if it is generally accepted that significance tests are insufficient, then a standard must be set for what is needed in terms of a more sufficient inference strategy. However, one consequence of such a standard would be, that formulation of our hypotheses must become more precise. It is logically impossible to have one part without the other.

Some historical remarks on statistical inference

This paper is not the place to discuss the historical development of statistics in detail. But it is necessary to give some indication of where the methods come from and why they have been used historically (see Krüger, Daston & Heidelberger,1987; Krüger, Gigerenzer & Morgan, 1987; Stigler, 1986; Witte,1980).

It is possible to distinguish four origins of the probability concept:

a) practical statistics as the distribution of a population in different classes, e.g. men - women in old Egypt.
b) gambling in the Middle Ages;
c) decision making and the proof of the existence of God (Pascal);
d) logic and truth (Leibniz, Bayes).

Obviously, there are divergent origins of the concept of probability, and a mixture of these categorical differences into one concept must lead to certain inconsistencies in the interpretation of the results of significance tests, although the different angles each have specific conditions under which they are justified .

This leads to the conditions which were decisive for the construction of current test theories. As everybody knows, Fisher was interested in applied research of agricultural field experiments. This had, at first, nothing to do with scientific inference: However, this procedure of manipulating conditions resembles the variation of independent variables coming from theoretical models predicting an effect. (The no effect null hypothesis can be used as the prediction by a random influence and should be falsified, which is the probabilistic version of falsification in the Popperian sense.) Since philosophy of science in this regard teaches that there is no verification, there is no need to formulate an alternative hypothesis. This represents a convenient misunderstanding of falsification as theory test, because the theoretically relevant hypothesis should be tested, and the failure of disproving this hypothesis increases the credibility of the theory. But there is no theory

adequate to a precise prediction in these field experiments; therefore, the null hypothesis is taken as the only precise prediction from a trivial random effect. This theoretical distribution of expected data is determined before any empirical research. After an empirical result has been observed, the decision depends on the probability of this datum under the hypothetical distribution. The conclusion leading from the empirical result to the falsification of the random hypothesis is only indirect. There is no experimental planning and no idea of a theoretically precise prediction. This program of statistical inference only controls a random effect as an explanation of the results and takes the falsification of the explanation to be an increase in the credibility of the theory.

The next step was to accept that there has to be something like a theoretical alternative hypothesis to the random effect. Furthermore, the chosen inference test should have some characteristics comparable to those of other possible tests, so that the decision based on these tests is optimal in a mathematically defined sense. The test to be used should be uniformly the most powerful, unbiased, and consistent according to the set of standards in the inference theory developed by Neyman and Pearson. Thus, the Neyman-Pearson theory is an improvement of Fisher's original theory into a more consistent and mathematically defined strategy. (However, the alternative theory is only existent in an abstract sense not precised. It is therefore known by using significance tests that the power is maximal, but the real level of power can only be identified if the alternative hypothesis is also a precise parameter.) Thus, our significance tests deal with the expectation of certain empirical results under hypothetical assumptions. Under specific conditions, the distributions of the expected data are unique (t, F, normal). Thus, it is very well known before the occurrance of any empirical result what data, with what probability, are expected. With the determination of two hypothetical parameters, the experimental condition can be planned exactly, and with the errors to be accepted. After the precision of the alternative hypothesis the Neyman-Pearson theory is a theory of planning experiments without any empirical

data. This was the direction in which Cohen (1969,1977) developed the Neyman-Pearson approach. What is needed is the definition of an effect size. This is the minimal value of the alternative hypothesis, which now has been defined quantitatively as the lowest point of the interval defining the hypothetically acceptable parameters. With this quantification of the alternative hypothesis, the whole inference strategy becomes more informational, although Cohen's approach is based completely on Neyman-Pearson's theory, which is a planning and not a test theory.

There is another approach in the modern history of the inference revolution, namely the decision theory developed by Wald (1947) in his sequential tests. His concrete problem was to decide whether a set of manufactured products was defective or not. This decision should be made with as few observations as possible under tolerable error rates. Thus the number of observations is a variable rather than a constant. Although Wald is discussing the Neyman-Pearson theory, he uses a different decision criterion either for or against the null hypothesis, his sequential probability ratio test. This means there is no a priori determination of the sample size, but the decision to reject or accept the lot has to be determined beforehand, by fixing α- and ß-error and the minimal deviation. The consequence is that after each outcome is known, there is a new decision about the two hypotheses. With this sequential sampling there can be a very high reduction of sample size needed for a decision. This can be possible because the real error of the lot can be much greater than the minimal error, and hence detected earlier than the minimal error. However, this inferential strategy is not constructed for the purpose of testing scientific hypotheses. It is optimal under the particular conditions for which it is developed, namely accepting or rejecting a set of manufactured products. Under these circumstances whether the lot is error free or not is the only interesting detail. It is by no means relevant to know the real amount of the difference if this deviation is above a chosen threshold. To transform this condition into scientific inference means that the null hypothesis is the theoretically interesting

derivation and that the other hypotheses are irrelevant. They are taken as variables. Thus the theoretically relevant parameter - in using this test strategy for scientific purposes- is not deduced from assumptions; it varies randomly.

There is only one inference strategy directly concerned with testing a hypothetical parameter: the idea developed by Bayes (see e.g. Stigler,1986). Such a model is the only one that tries to solve the problem of "induction" or "inverse probability". This concept should be at the center of our inference tests, but it is almost totally ignored in comparison to the sequential tests and the power analytic planning concepts. These require only a partially specified alternative hypothesis, at least at the point of the lowest acceptable deviation. There is a consensus in the scientific community that specification of hypothetical parameters is nearly impossible (but see Tukey, 1969). The consequence of this consensus is state-of-the-art of significance testing. (For further discussion see the very informational article by Cohen, 1990 and Meehl, 1978.)

A possible set of criteria to be fulfilled by an inferential test strategy

The first and main step of a statistical inference strategy is the clarification of what is wanted. All statistical test theories have their own central kind of application, which can then be extended to scientific inference. If these demands are known, then a statistical procedure which is able to fulfil these standards should be developed. These standards should be based on three foundations: a) philosophy of science (e.g. Earman,1983; Stegmüller,1973; Maher,1992), b) mathematical statistics (e.g. Kendall & Stuart,1963; Silvey,1970), and c) empirical research conditions in the behavioral sciences (e.g. Cohen, 1977,1990).

It is impossible to discuss any of these three foundations in detail. Therefore, the demands are given without a lenghty

explanation:

1) The central point is that a measure of the relative confirmation of one hypothesis against another under empirical evidence is needed (principle of relative confirmation). After the discussion of induction in philosophy of science it seems that empirical evidence can only result in a measure of relative confirmation between two hypotheses, and not in an absolute measure of a single hypothesis.
2) The information of the data should be exhausted in the description of the empirical evidence, so that the measure of the evidence contains all the information needed for the relative confirmation of the two hypothetical parameters (principle of sufficiency).
3) Allowed transformations of the empirical data must not have an influence on the confirmation of the hypotheses (principle of the independence between data transformation and hypotheses confirmation).
4) The kind of sampling (one step or sequential) should be independent of the confirmation (principle of the independence between sampling and confirmation). In the tradition of the Neyman-Pearson theory, the result of the sample as well as the more extreme results are taken into consideration: This means that the one-step sampling is fundamental for this theory.
5) As a technical demand, it is favorable that the global measure of confirmation from independent samples is the sum of each sample's measure (principle of additivity of the confirmation measures).
6) In general, the a priori probabilities of the two hypotheses should be equal, because the search for truth and the more pragmatic criteria for the consequences of a false decision should be seperated (principle of independence between truth supporting and consequence evaluation). The result of this demand is that the two hypotheses should be equivalent in kind (e.g. null and alternative hypotheses should be both simple point parameters or both should be intervall parameters) and the error rates(α, β) should be equal,too.

7) The error rates themselves should be known before one hypothesis is judged as more confirmed as the other (principle of error estimation).
8) The relative confirmation of one hypothesis against another also depends on the rejected hypothesis: Nothing is known about the direct relation between data and the chosen hypothesis. During the process of theory development it is necessary to evaluate the correspondence between empirical results and the chosen hypothesis, because a high correspondence means that the theoretical modification can be laid aside for the moment (principle of hypothesis qualification). This correspondence, however, must be evaluated in a probability model of confirmation and using the data themselves.
9) Until now, all of the evaluations of the hypothesis are based on an abstract model of confirmation. Using a theoretical explanation as a prediction in a theoretical or applied context it is necessary to know more about the amount of explained variance of the data by the theory. If this amount of variance is rather small, then the theoretical explanation works for the summarized statistic used in the test (e.g. the mean); however, the other uncontrolled factors (e.g. expressed by the variance) are too strong for a sufficient influence of the theory in the data measured, although it has been confirmed by the inference strategy (principle of effect qualification).
10) As a last and most important point, a general measure of confirmation, which evaluates the relative support of one hypothesis against another using the empirical evidence, is needed. This has been called inverse probability or likelihood (principle of likelihood).

These ten principles could be a catalogue of demands for an inference strategy. Of course, this catalogue is only a proposal, and can naturally be modified. However, it is a positive formulation of those standards that should be satisfied if the behavioral sciences are to show theoretical progress. Such depends on the deep connection between inference strategy and theoretical construction and development.

Of course, such criteria cannot be fulfilled at the beginning of a research tradition. At that time, methods for the exploration of data (Tukey,1977) should be used, and not the test of precise theories and their hypotheses. But our usual strategy looks like pseudo-testing at the very beginning of research, concentrating on promising results without specification of the parameters to be tested. Contrary to our quantitative data analysis, the formulation of the hypotheses are still simply qualitative, even after forty years of research (e.g. dissonance theory). The level of measurement (e.g. interval data) is almost never used for a specification of hypotheses.

A non-confounded inference test strategy: FOSTIS

With this catalogue in mind, the question becomes how we might satisfy the demands. The solution is a number of different approaches combined into one strategy, because the different approaches each have strengths and weaknesses (Witte,1977, 1980,1989). This combination is hierarchically organized into four steps of decision (called FOSTIS). At each step, there is a three-valued logic that entails either the acceptance of one of the two hypotheses, or a continuation with the sampling of data as it is well known from Wald's sequential tests.

The first step is the planning of the empirical condition under which two hypotheses should be tested. The best planning theory is that of Neyman and Pearson. However, two parameters have to be specified: $\theta(0)$ and $\theta(1)$ as point hypotheses, which is the easiest case. Furthermore, the error rates have to be specified as $\alpha = \beta$. The consequence of this specification is that minimal sample size can be determined. This represents the condition under which the test of the hypotheses is allowed and the first step of the inferential strategy is satisfied. If the minimal number of the sample size is not reached, then the results are to be reported without a test. Science is, after all, a cumulative process. Why do we need to test hypotheses? It

is interesting that Cohen (still 1992) proposes to increase ß to 0.20, because a smaller value "would result in a demand for N that is likely to exceed the investigator's resources" (p.156). This might be true, but then one ought to wait until other investigators include their resources, making the empirical condition. ß should never be increased as a general strategy to compensate for them. If there is something to be tested, then there is a need for minimal resources, else it looks like a pseudo-test with all the consequences of ignoring non-significant results and multiple tests. To some extent, inadequate resources can be fortunate, because different labs must combine their results, which is much more informative than confirmation from one lab. It might also be the case that the resources are greater than the minimum. Under these circumstances, a reduction of the Type I and Type II error is possible.

If the empirical condition passes the first step, then the two hypotheses can be tested. The test is a likelihood ratio test which in fact is similar to the Wald probability ratio test, but in a non-sequential form :

$$L[\theta(1)/x]/L[\theta(0)/x] \geq \epsilon = (1-\beta)/\alpha \qquad (1)$$

If the likelihood ratio exceeds a critical value determined by the ratio of the power and the Type I error, then the alternative hypothesis is sufficiently better confirmed than the null hypothesis. This test is based on the ideas developed by Bayes, conceptualized as a likelihood by Fisher, and used as a test by Wald. Thus the results of these theories can be used for the evaluation of the general test strategy.

If the alternative hypothesis has passed both steps, then there is still the question of whether the confirmation depends solely upon the improbability of the null hypothesis, or whether the empirical data corroborates this hypothesis to some extent. For this kind of evaluation a maximum likelihood test proposed by Edwards (1972) is used:

$$L[\theta(1)/x]/\ \max\ L[\theta(i)/x] \geq Q = 1 - \sqrt{(1-\beta)}*\alpha \qquad (2)$$

This test and its criterion are strange, so some experience is needed to make it more plausible. The idea is that the confirmed

hypothesis should not deviate from that hypothesis best supported by the data more than one standard deviation of a binomial distribution with the parameters p = (1-ß) and q = α. The acceptance of the alternative hypothesis is based on the probability that it itself is true (1-ß) and that the null hypothesis is true (α). Thus, this decision of acceptance by the data varies with the truth of both in light of this data as empirical evidence. Only under this condition can the correspondence between hypothetical assumption and empirical evidence postulated be acceptable.

The fourth criterion comes from the discussion about the observed effect size by Cohen (1977). There are many measures of effect size (e.g. McGraw & Wong,1992), and it is not easy to choose one. My preference is the coefficient of determination. If the theoretical assumption can explain at least 10 % of the variance, then the empirical evidence for the 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.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 can ask whether this assumption is satisfied. One consequence of this criterion is such 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, and that the measurement of the variables are precise enough for such a test. One consequence of a failure to pass this criterion might be the reduction of the error variance and not the alteration of the hypothesis.

At best, the meaning of each inference theory's step can be explained by its failure to pass the critical value. After the fixation of the Type I and Type II error, these errors are to serve as a base for all critical values. Also, the likelihood

test of the second step 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 which further means equal a priori probabilities. This is one reason why the kind of hypotheses should be similar and not dissimilar, as in the usual significance test, in which a simple point hypothesis is tested against a complex alternative hypothesis.

If the critical value of the first step is not passed, then the empirical basis is not sufficient for a test of the hypotheses. Under such conditions a description of the data is given such that a later researcher can use this data as one kind of empirical evidence, to be combined with his own results so as to make possible a test of hypotheses. This means taking seriously science as a process in which data from different researchers can be integrated. How can this be done with the proposed inference strategy ? (I do not want to discuss the current meta-analytic methods, which are fundamentally based on the classical significance tests; see below.) At first, the number of the needed sample size which satisfies chosen Type I and Type II errors could be given. The value of the logarithm of the likelihood ratio is then calculated and published. The next researcher can merely add this to his own logarithm of the new likelihood ratio, because the combination of these independent loglikelihood ratios is simply the sum of the two. Such a data description without test clearly demonstrates the insufficient base of the empirical evidence. Such a procedure cannot lead to Type II errors around $\beta=0.50$.

If the emprirical evidence is sufficient, but the test of the hypotheses does not reach the critical value, then the resulting assumption is that either the hypothesis is incorrect or the experimental setting was defective. In the first case a reformulation of the hypothesis would be required, and in the second a new experimental setting.

If during the third step the critical value has not been passed then it would appear to be a revision of the concrete

theoretical parameter consistent with the idea that basic theoretical assumptions are true. This is the case if the empirical results are more different from the null hypothesis than assumed. Under theoretical conditions, an empirical result can also have an effect which is too large. Usually, it is implicitly assumed that the greater the deviation from null hypothesis the more the theory is confirmed. But this is only true if all deviations are seen as a corroboration of the theory. This assumption only holds if the alternative hypothesis is unspecified and only qualitative.

If all three steps of the testing procedure have been passed but the fourth has been failed, then the question is generally whether the theory is useful to explain data in the given empirical context. One consequence is to increase the precision of the measurement or to restrict the theory to more specific conditions. The critical value used 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 very little empirical evidence. This critical value has the technical function of being 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 parsimonious to ignore the theoretical influence postulated. In general, the main strategy should be to 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 so-called theories which pass this lenient criterion. It is only necessary either to await a significant result with an extremely high Type II error, or to increase the sample size. The theoretician will almost always win. This cannot be an acceptable inference strategy, although it is better than not controlling a random effect at all. This more or less simple control, however, stands at the beginning of research, and must be refined with the development of more precise hypotheses and a more complex inference strategy in the course of its process. One example of what could have been followed is provided by the four step inference strategy (FOSTIS).

An example of the test theory FOSTIS

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 precised. From past research it is predicted that the difference should be one half of the standard deviation $d = 0.50$. From this theoretical assumption, it is known beforehand that if the alternative hypothesis is true, only about 6% of the variance is determined. Thus, the last step will not be passed, and the measurement error will have to be reduced in the future. Until now there has been no alternative to this prediction. Furthermore, the Type I and Type II error should be equal to $\alpha = \beta = 0.05$. With the specification of these parameters, the sample size can be predicted as $N = 88$. For this planning of the experimental condition, Table 2.3.2 in Cohen (1977, pages 28 - 39) is very useful. A rough approximation for the determination of the sample size is given by the formula

$$N = 2[z(1-\alpha) + z(1-\beta)]^2/d^2 \qquad (3)$$

N : sample size of each group

$z(1-\alpha)$: standard z-value at $(1-\alpha)$

$z(1-\beta)$: standard z-value at $(1-\beta)$

d : hypothetical difference of the means standardized by the common standard deviation.

For the example, there is d =0.50, and $z(1-\alpha)=z(1-\beta)=1.65$. If the formula above is used after rounding to the next higher natural number, we get N = 88 as demanded sample size. This corresponds to Cohen's table 2.3.2. His table 2.4.1 (page 54-55) in which the sample sizes are estimated, gives a size of N=87. This means that the rounding procedure is not always necessary. It is very simple to use this approximation formula, which is correct enough in most cases.

After the planning of the experiment the next step is the obtaining of a result. It is d(emp) = 0.53 observed. The likelihood ratio of the two hypotheses must be determined with d=0.00 (null hypothesis) and d=0.50 (alternative hypothesis). There are 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 theoretical 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. At first the z-value of the theoretical expectation is calculated as the mean of the noncentral t-distribution:

$$z(hyp) = [d(hyp)*(N-1)*\sqrt{2N}]/[2(N-1)+1.21*(z(1-\alpha)-1.06)] \quad (4)$$

$$z(hyp) = [0.50*87*\sqrt{2*88}]/[2*87+1.21*(1.65-1.06)] = 3.30$$

z(emp) = 3.50 with d(emp) = 0.53 .

What must be determined at this point is the ordinate of the probability density at the points of the empirical results for the expectation of the null hypothesis[1]:

z(null) = 0 - z(emp) = -3.50 with y(null) = 0.0009, and the alternative hypothesis: z(alt) = 3.30 - 3.50 = -0.20 with y(alt) = 0.3910. The loglikelihood ratio test then is the following: log [y(alt) / y(null)] = log [434.44] = 2.64.

The critical value of this second step is:

$$\epsilon = \log [(1-\beta)/\alpha] = \log [19] = 1.28.$$

The data significantly confirms the alternative hypothesis.

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 best empirically supported must be

determined, which is y(alt) = 0.3910 divided by y(max) = 0.3989, or the maximal ordinate of the normal density function. The result is: Q=0.98. The critical value to be passed is $Q=1-\sqrt{\alpha(1-\beta)}=0.78$. In this way is the third step of the inference strategy passed.

The fourth step goes back to the empirical result itself and evaluates its effect not in a probability model, but in the measurement itself. An effect size of d = 0.53 has been observed. If the sample size of both experimental groups are equal, the coefficient of determination is $r^2=0.06$ (see Cohen, 1977, p.22-24). Thus, the criterion of the fourth step has not been passed. In the context of FOSTIS, this means that the error of measurement has to be reduced, and not the Type I or Type II error. Because of the hypothetically assumed d= 0.50, such a result was expected. At this point the empirical conditions have to be modified so that a more profound effect can be observed. This leads to restricted conditions of the theory's validity. In general, the observed effects of our theories are not very promising if mean effect sizes from many studies are given in the meta-analyses. The other way to broaden a theory's application to its still existing but minimal influence is another strategy of theory development (Prentice & Miller, 1992) but it is necessary to use both procedures in finding conditions in which theoretically assumed effects are optimized and minimized without triviality. But the amount of the theoretically determined variance must be known. One possible consequence of this is the prediction of the null hypothesis under specific conditions, because the influence of the theory should be eliminated under the experimentally manipulated conditions. However, without the demand of the fourth step, there will be no discussion of these errors of measurement.

Extreme conditions for the relative confirmation of a hypothesis

This is the technical demonstration of extreme conditions only when the alternative hypothesis is corroborated by FOSTIS. If the intention is to pass the fourth step, one must choose an

experimental condition that allows for 10% of the variance to be determined. Of course, this will be the prediction of the theory which will vary with the empirical test condition as a result of the error of measurement. An $r^2=0.10$ with a $d=0.67$ will be predicted. With this theoretical prediction, the t-distribution tells us the sample size of $N = 49$ of each group if $\alpha = \beta = 0.05$.

The next question concerns the empirical value that minimally confirms the alternative hypothesis. Such is the case if the ordinates' relation of the two hypotheses at the point of the empirical result d(emp) equals $(1-\beta)/\alpha = 19$. To pass the critical value it is necessary to get an empirical result which is no less than d(emp,min) = 0.50 (see Appendix).

The third step of FOSTIS is the qualification of the alternative hypothesis through the maximum likelihood of the data. The alternative hypothesis is accepted only if d(emp,max) is not greater than 0.81 (see Appendix).[2]

The fourth step, the effect qualification, is determined by d(emp,mineff) = 0.67, which was the alternative hypothesis. In such a case the percentage of the determined variance is 10% . If an empirical value was observed which is less than d=0.50, the alternative hypothesis cannot be accepted, for the empirical evidence gives no basis for a decision between both hypotheses if $\alpha=\beta=0.05$. If, however, an empirical value greater than d=0.81 is observed, the alternative hypothesis should be corrected in the direction of a greater expected difference. The fourth step, the effect qualification, has to do with the measurement error rather than the decision error. Until now, there has been no such criterion in which this point of measurement error has been integrated in our test theories. In the last few years, a more intensified discussion about effect sizes has started, yet still without the prospect integration into the test theory[3].

Critical comments on FOSTIS

FOSTIS has been published in German journals (Witte,1977, 1989) and in a more comprehensive book (Witte,1980). Of course, it has been criticized intensively (Bredenkamp,1983; Diepgen,1991; Hager,1991; Kleiter,1991; Wendel,1991) and very seldom praised (Brocke, Iseler, Holling & Liepmann,1983). Because of the intensive discussion surrounding it, it now seems appropriate to enter into a discussion of particular points.

One of the general problems handled is the usability of our science to build precise point hypotheses. Only the hypothetical prediction of qualitative differences is possible; no precise theoretical construction is possible. Thus an inference strategy which demands such precise theoretical modelling is unrealistic. Such a rebuke is correct only if new research is started and we are more or less exploring our data. Under such conditions, after the first phase of playing with data, the question is whether something which cannot be explained by a random effect occured. In the tradition of the commonly used hybrid theory, the researcher begins with a nondirectional test, continues with a directional test, and then tries to become more precise in the theory. The last step never happens, however, not because it is impossible to measure reactions on an interval scale, but because the theories do not contain a mathematical kernel. That can predict new or initially old results precisely in the same way as they are measured. This process of theoretical development in combination with data as its basis does not represent a common strategy in behavioral sciences (see also Meehl,1978). Each collection of experimental data is seperated from all the others (other than those in meta-analyses, which are mainly used to combine effects in the light of a null hypothesis, and not as a construction of a precised theory). If there is an alternative to the classical hybrid theory of statistical inference, which is the optimal transformation of the theoretical construction, then it might be possible that the theoretical construction itself can also be influenced in such a way as to continue with specification of theoretical predictions in the first few years after the beginning of

research (Tukey, 1969).

Another solution of the problem might be the formulation of interval hypotheses, rather than point hypotheses. Then, however, it is necessary to accept an interval null hypothesis, because on logical grounds the hypotheses to be tested should be of the same kind. The consequence of this demand is that the theoretician has to formulate four points, the end points of each interval, and this does not seem to be any easier than the specification of two point hypotheses. The test, however, is comparable to the formulation of point hypotheses.

A second critical point is the introduction of the third step: validation of an accepted hypothesis on the hypothesis most supported by the data. From the usual test against the null hypothesis this is a critical point, because a confirmed hypothesis can be rejected even if it is much more evident than the null hypothesis. However, this argumentation derives from the classical view of rejecting a random effect and not accepting a point hypothesis. If theory-building has progressed, then it is also a question of the alternative hypothesis' confirmation by the data itself and not exclusively as it relates to the null hypothesis. The chosen criterion is based on the likelihood principle, using the same base-line as the test itself. The critical value might be too easy a convention to be accepted. In my opinion, it is consistent.

After making decisions under the perspective of a probability model, the fourth step of FOSTIS suggests that there is a need to go back to the measurement and to the explained variance, which is sometimes small though statistically significant. The idea is that two thirds of the standard deviation is necessary to ensure that the two theoretically relevant parameters are different in the experimental condition under which the theory is tested. Lacking a clear separation, there are so many possible explanations of the observed effect that no clear-cut interpretation is possible. This is, of course, not a demand to be made at the beginning of a research

tradition; it should rather be satisfied at a fully developed phase.

At this point the main critical point from the view of the Neyman-Pearson theory should be discussed: How does the combination of step one (planning of the empirical test condition) and step two (using the likelihood test for the decision between two hypotheses) work? Is it true that the decision made at the second step is based on the Type I and Type II errors on the first step? At first glance, the combination of both test theories seems to be inconsistent, for the critical value of rejecting the null hypothesis with α=0.05 (one-tailed test) and N=49 is d(crit) = 0.34 (see Cohen,1977, p.30-31). This value is much less than the d(min)=0.50. If such an effect size of d=0.34 is the most frequently reported empirical difference in the publications, then the power of this effect is around 1-ß=0.50, as the mean power of the usual significance tests observed in the publications. Compared with the likelihood test criterion, which led to a difference of at least d=0.50, there is a leniency effect for the acceptance of the alternative hypothesis in the significance-testing strategy of rejecting the null hypothesis.

If we now take the minimal critical difference of d=0.50, then the focus of our attention should be its Type II error, under the assumption that the alternative hypothesis is true with d=0.67. Using the approximation of the non-central t-distribution to the normal distribution, we get a z-value of z=0.83. This means that the Type II error is about ß=0.20. It seems that the decision rule, using likelihood ratios, and the planning rule, using probabilities, do not lead to the same result. Since our inference strategy is symmetrical concerning the Neyman-Pearson theory, the whole difference between the two hypothetical parameters d(null)=0.00 and d(alt)=0.67 should be averaged to get the critical value d(crit)=0.335. This is the value in which α=ß=0.05, as planned with the help of the Neyman-Pearson theory. The FOSTIS' decision rule of the second step leads to d=0.50, which is nearer to the alternative hypothesis' parameter than the one predicted with ß=0.05 by the Neyman-

Pearson theory. (This was the critical point of the arguments against FOSTIS). Fortunately, FOSTIS' step two is a non-sequential Wald test, and the whole argumentation is based on his findings without any new ideas. It is necessary that these two testing strategies are not confounded on the level of the empirical evidence. Under step one, the Neyman-Pearson theory is used to predict the empirical data from the view of two hypothetical parameters. Under step two, a decision has to be made between two hypothetical values in the event that a specific datum has been observed. These decisions should be wrong to the same extent that the predicted data gives wrong information about their hypothetical distribution. For any single sample value, the likelihood of the confirmed hypothesis should be at least A times as large under the confirmed hypothesis as under the rejected, in which A depends on the ratio of the power and the Type I error. Since there is only one empirical observation to decide between the hypotheses, this decision should be at most as far off the mark as is predicted by the planning of the testing condition. This is the principle used by Wald to construct his tests. FOSTIS's combination of the two different situations before and after the knowledge of the experimental data has been labelled its chief inconsistency. Before knowledge we must plan the sample size, assuming theoretical parameters and predicting empirical values. After knowledge of the data, we must decide between the hypotheses. What is not allowed from a theoretical point of view is the regression from step two to step one in a comparison of the empirical data. Both conditions are only comparable in a more abstract, general description of a test. This description is given by four parameters with three degrees of freedom: d,N,α,β. Using these four parameters of a statistical test, the planning and the decision conditions are equivalent, for the accepted hypothesis is more probable than the ratio of the power to the Type I error. However, they do not lead to the same empirical value. This might be irritating at first glance, but the knowledge conditions before and after the empirical results are very different. The test strategy proposed by FOSTIS is thus hierarchically ordered, and the decision of the second step must not be evaluated from the perspective of Neyman-Pearson theory

as the first step. However, the possible errors of the experimental condition as predicted by the Neyman-Pearson theory are used to determine the decision after the knowledge of the result, so that each single decision reflects the ratio of the Type I to the Type II error insofar as the likelihood ratio passes the critical value used in the Wald test (see Appendix).

The failure to pass each step has a different and specific meaning for the development of a theory. Thus, it is necessary to plan the future steps of a theory depending on that theoretical development (tested by FOSTIS) that has reached up to this point.

Combining results from different studies

Our common inference strategy leads to an isolated test of each study against a random effect. This, however, was felt to be unsatisfactory, so that the methods called meta-analyses have been used. These methods enable the researcher to combine several empirical studies into one test against a random effect. Because all effects are evaluated solely from the view of the null hypothesis these meta-analyses are a natural extension of the common hybrid theory of significance testing.

This combination rule is not comparable with those principles formulated as the foundation for FOSTIS. The FOSTIS strategy requires a combination of loglikelihood ratios deriving from both hypotheses into a new measure of hypothesis, one that tests from different samples. Then, however, the critical value must also be changed, for an increasing of the sample size might lead to a reduction of the Type I and Type II error over the first fixed critical size. The critical value is determined by the Neyman-Pearson theory, and due to the change of α and β, it changes with increasing N. The combination rule of the loglikelihoods is simply the sum of each sample's loglikelihood ratio into the total measure, as is well known from sequential testing (Wald, 1967). This kind of integration needs two hypotheses, and represents the way to become more precise in

theory construction. In my view, this is the alternative to the meta-analytic methods, but what is needed is theoretical preciseness.

For the third step, the integration is different, as a result of the need for an estimation of the empirical value for all samples. This problem has to do with estimating a parameter from different samples. What should be found is the maximum likelihood estimation of the empirical differences between the two means from different samples. This is the weighted average difference from the sample differences, as weighted by the sample size. This parameter is maximally supported by the data, and should be used as the critical value against which the confirmed hypothesis can be qualified.

For the fourth step this maximum likelihood estimation from different samples is used for the estimation of the empirical effect, which is transformed into a percentage of determined variance and compared with a critical value.

This kind of integration from different samples is a consequence of science as a process. Necessary, however, is the view provided by two hypotheses, not only one, as is preferred in the hybrid theory of significance testing.

Concluding Remarks

The discussion about the common inference strategy was oriented at the insufficient combination of Fisher's and Neyman-Pearson's ideas. Due to the many misconceptions surrounding this inference strategy, many researchers contemplated alternatives to this null hypothesis testing (e.g. Cohen, 1990). What is initially needed is a catalogue of demands to be satisfied by what might be called statistical induction. These demands need to be based on a combination of philosophy of science, mathematical statistics, and the assumptions of the active investigators. One result of the discussions was that the common hybrid theory is the statistical pendent of the formulated

hypotheses. It is not possible to improve the inference strategy without precisation of the theories. Thus the whole significance test controversy revolves around a fundamental discussion of the theoretical status of the disciplines using this inference strategy. Statistical consequences of this inference strategy include the power of the observed effects (around $1-\beta=0.50$), the observed effect sizes, which are less than medium, and that each empirical study is tested in an isolated fashion. A more serious consequence is that the theoretical models remain vague even after a long tradition of research with only qualitative hypotheses. The significance test is a perfect strategy to employ at the beginning of a research tradition, at which point nearly any experience of the data might be expected. After a time, however, the point arrives at which the theory has to be precised with a clear quantitative prediction, in the same manner in which the data were measured for the usual tests (Tukey, 1969). Such a theory, which is more precise, is not testable under the usual significance test strategy, because the testing procedure depends only on one point null hypothesis. The alternative hypotheses remain un-specified. If there was a concrete alternative hypothesis, then it would be possible to inspect both confidence intervals. Both, however, depend only on one hypothesis each. There is no relative confirmation of one hypothesis against the other.

This significance test controversy, which is as old as the significance tests themselves, should be taken into consideration by a catalogue of demands that addresses those things that a more satisfactory inference strategy should accomplish. Such a set of demands has been given, and a specfic new hybrid theory of statistical inference (FOSTIS) derived. Such hybridity is always necessary, for there are different phases of an empirical investigation that must be integrated into such an inference strategy. For each phase, however, a different statistical approach must be regarded as optimal. Four such phases have been identified, with a different meaning for the construction of a theory should the respective critical value not have been passed.

Generally, there is something to be learned: the more meaningful statistical induction should be, the more precise theoretical deduction must be.

References

Bredenkamp, J. (1983). Übersicht. In: Bredenkamp,J. & Feger,H.(Eds.). Hypothesenprüfung. Enzyklopädie der Psychologie. (B, Bd.5, S.1-23).Göttingen: Hogrefe.

Brocke,B.;Iseler,A.;Holling,H. & Liepmann,D. (1983). Kritik und Anwendung sozialwissenschaftlicher Statistik. Zur Relevanz der Leiserschen Statistikkritik für die Forschungspraxis. Anwendung sozialwissenschaftlicher Statistik. Zur Relevanz der Leiserschen Statistikkritik für die Forschungspraxis. Zeitschrift für Sozialpsychologie, 14, 189-196.

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

Cohen,J. (1969,1977). Statistical power analysis for the behavioral sciences. (1st ed.; rev.ed.). New York: Academic Press.

Cohen,J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

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

Diepgen,R. (1991). Inkonsistentes zur Signifikanztest-problematik. Ein Kommentar zu Witte (1989). Psychologische Rundschau, 42, 29-33.

Earman,J. (Ed.). (1983). Testing scientific theories. Minnesota Studies in the Philosophy of Science. (Vol.X). Minneapolis: University of Minnesota Press.

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

Gigerenzer,G. & Murray,D.J. (1987). Cognition as intuitive statistics. Hillsdale : Erlbaum.

Hager,W. (1991). Die Problematik eines "Theorien-Cocktails" in der statistischen Methodenlehre: Die letzte Signifikanz-testkontroverse, zu der das letzte Wort noch nicht gesprochen ist. Psychologische Rundschau, 42, 213-215.

Kendall,M.G. & Stuart,A. (1963). The advanced theory of statistics. London: Griffin.

Kleiter,G.D. (1991). Goldvermehrung in der Statistik? Zur Kontroverse zwischen Diepgen und Witte. Psychologische Rundschau, 42, 215-217.

Krüger,L.; Daston,L.J. & Heidelberger,M. (Eds.). (1987). The probabilistic revolution. Ideas in history. (Vol.I). Cambridge, MA: Bradford.

Krüger,L.; Gigerenzer,G. & Morgan,M.S. (Eds.). (1987). The probabilistic revolution. Ideas in sciences. (Vol.II). Cambridge, MA: Bradford.

Maher,P. (1992). Betting on theories. Cambridge, MA: Cambridge University Press.

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

Meehl,P.E. (1978). Theoretical risks and tabular asteriks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical psychology, 46, 806-834.

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

Prentice,D.A. & Miller,D.T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160-164.

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

Silvey,S.D. (1970). Statistical inference. Harmondsworth: Penguin Books.

Stegmüller,W. (1973)."Jenseits von Popper and Carnap": Die psychologischen Grundlagen des statistischen Schließens. Berlin: Springer.

Stigler,S.M. (1986). The history of statistics. Cambridge, MA: Belknap.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91.

Tukey,J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

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

Wendel,M. (1991). ... und Diepgen hat doch recht ! Ein Kommentar zur Witte-Diepgen-Kontroverse. Psychologische Rundschau, 42, 211-212.

Witte,E.H. (1977). Zur Logik und Anwendung der Inferenz-statistik. 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 Rundschau, 40, 76-84.

Footnotes

1 The way to determine the loglikelihood ratios takes into consideration that ratios of probability densities and of likelihoods are equivalent. This is the reason why the tabulated ordinates of the normal distribution are used to calculate the likelihood ratios.

2 These two values determine something like the confidence interval used in parameter estimation. However, the end-points of the interval are not symmetrical around the hypothetical parameter because they depend on different criteria.

3 To pass the critical values of all four steps of FOSTIS leads to the range of empirical observed values from 0.67 till 0.81. This range can be compared with a classical t-test (one-sided α=0.05), under which all values greater than 0.33 would lead to a corroboration of theoretical assumptions.

4 Is this a principle of least effort in our empirical research that the empirical effect sizes just pass the critical value on the average?

Appendix

The first question is about the minimal empirical value - d (emp, min) - so that the second step of FOSTIS will be passed?
At first, the d-value must be transformed into a z-value:

$$z = [0.50 \cdot 48 \cdot \sqrt{2 \cdot 49}]/[2(48)+1.21 \cdot (1.65-1.06)] = 2.46$$
$$z = 3.10 \text{ with } d(hyp) = 0.63.$$

The ordinates for the z-values $z=2.46-3.10=0.64$ and $z=2.46$ are $y=0.325$ and $y=0.019$.
The ratio of these two values is 17.11. This is just less than the critical value of

$$\frac{1-\beta}{\alpha} = \frac{0.95}{0.05}$$

For $d=0.51$, however, the relation of the ordinates is 20, just slightly more than the critical value.
Thus, the empirical value should be greater than $d=0.50$.

The second question is about the maximal empirical value - d (emp, max) - so that the third step of FOSTIS will be passed?
The critical value is $Q=1-\sqrt{\alpha(1-\beta)}=0.78$. Furthermore, the maximal value of the ordinate is known: $y(max)=0.3989$. Thus, the value of the ordinate and therefore the z-value can be found: $0.78 \cdot 0.3989 = y(emp,max) = 0.31$. The corresponding z-value is 0.72. Now it is possible to retransform this z-value into a d-value by the use of formula (4):

$$d = 0.72 \cdot 88.71/475.2 = 0.1344$$

The hypothetical d-value of the alternative hypothesis was $d=0.67$ so that the empirical value should not be greater than $d=0.8044 \approx 0.81$.

For the determination of the two extreme values, or what might be called a corroboration intervall, an indirect method is used because the direct method is very complicated.

A third question concerns the relationship between the size (α) and power ($1-\beta$) of the Neyman-Pearson theory and FOSTIS. To explain these relationships we go back to Wald's probability ratio test, which is equivalent to the likelihood ratio, as is known from the Bayes-theorem:

$$L[\theta(1)/x]/L[\theta(0)/x] = p[\theta(1)]/p[\theta(0)] \cdot p[x,\theta(1)]/p[x,\theta(0)].$$

In FOSTIS the left hand side is used and Wald introduced the right hand side without the probabilities of the hypotheses. This ratio, however, is 1 because of the equivalence of the hypotheses. Thus the decision criterion is the same. The main difference is the sequential decision process used by Wald's test and the non-sequential decision in FOSTIS. In my opinion,

the sequential testing is a confusing of planning and testing. The economy in the number of observations by the use of the sequential probability ratio test instead of the Neyman-Pearson test is not acceptable for a scientific (not applied) strategy because both theories are not comparable. Thus it is as senseless to criticize the decision under step two (Wald-test) from step one (Neyman-Pearson test) as the critique of the planning under step one from the decision under step two. Both steps have their own condition with an optimally adapted theory. Under step two the minimal value of the corroboration interval has been fixed by the possible error decisions α and β. This is a definition, and nothing else. It is not, however, comparable with the Neyman-Pearson theory, which is only optimal under the condition of step one. These are the main reasons for the hierarchical order of FOSTIS, which in itself is a modelling of the research process.