UNIVERSITY OF GÖTEBORG
Department of Statistics

RESEARCH REPORT 1986:5

ISSN - 0349-8034

TESTING THE APPROXIMATE AGREEMENT WITH A HYPOTHESIS

by

Marianne Frısén

August 1986

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

TESTING THE APPROXIMATE AGREEMENT WITH A HYPOTHESIS

M. Frísén

Department of Statistics, University of Göteborg,
S-411 25 GÖTEBORG, SWEDEN.

Summary

When statistical tests are applied it is often known beforehand that the hypotheses would be rejected with sufficiently large sample sizes. This happens whenever hypotheses is not exactly true but only approximately true. Some attempts of solution of this dilemma are discussed and exemplified with test of bioequivalence. One of these, powerfunction analysis, is applied on preparatory tests. In that case the approximate agreement with some condition (e.g. normal distribution) for the main analysis (e.g. t -test) is tested.

Key words: Approximate validity; Powerfunction; Bioequivalence; Preparatory tests.

Supported by grants from the Swedish Council for Research in the Humanities and Social Sciences.

1. Introduction

The classical theory of tests of statistical hypotheses, as formulated by e.g. Lehmann (1959) , is generally well accepted among statisticians. However, the use of the theory for practical problems is often experienced as a logical dilemma. Martin-Löf (1974) described it as: "... with large sets of data our results are purely negative: no matter what model we try, we are sure to find significant deviations which force us to reject it". The above dilemma has been mentioned before (e.g. by Berkson 1938) but is still a problem.

Both the formulation of a simple hypotheses, and the practical consequences in some use of the test methods are absurd in many applications. The formulation is not appealing in situations where there is no reason to believe that the null hypothesis H_0 is exactly true but instead its approximate validity is of interest. Some examples of hypotheses where approximate validity might be of interest are given below:

H_{0A} : Treatment with vitamin C has no effect on the incidence of the common cold.

H_{0B} : Two alternative formulations of a drug are "bioequivalent" , that is equal amounts of them produce equal therapeutic effects.

H_{0C} : X is normally distributed.

If the test is applied without any concern about the power it is easy to misuse the test and actually wrong conclusions because of a stereotype application are quite common in practice. Many statisticians have reacted to this misuse of statistical tests and some have advocated that tests should be avoided and advocated other procedures. First some alternative methods are discussed, then the strict use of the powerfunction in the context of tests is discussed and applied to several problems.

2. A REVIEW OF SOME ATTEMPTS TO SOLVE THE DILEMMA

2.1. Confidence intervals

Often a test can be replaced by a confidence intervall. Then there is the general information about the location of a parameter and the uncertainty, without special reference to some specific value.

The close relation between test and confidence interval usually makes transformation of one to the other easy. Sometimes confidence intervals are used as a test e.g.: "Make the statement two formulations of a drug are not bioequivalent if a confidence interval for a measure of the difference does not contain zero". In those cases the procedures of test or confidence interval are equivalent and with the same need for care in the interpretation.

However, in their straightforward use tests and confidence intervals do not give the same kind of inference and they are suitable for different kinds of problems. The important difference is whether all values are of the same concern or not. Even when an exact value is of less importance but the same action would be taken as soon as the parameter is close to a specific value, there is a specific value of special concern. The criticism of hypothesis testing that it puts unduly high stress on one value has a correspondence as the inference from a confidence interval is (sometimes unduly) symmetrical with respect to parameter values.

2.2. Enlargement of the null-hypothesis

Hodges and Lehmann (1954) suggest that the size of the test should be fixed on the limit of an enlargement H_0' of H_0. For the hypothesis $H_0: \mu = 0$ the enlargement could be $H_0': |\mu| < m$, where m is a positive constant. At some hypotheses, for example about the expectation in a normal distribution, with known variance, this is a trivial change. In other cases, for example the corresponding hypothesis when the variance is unknown, substantial changes are necessary. For the latter situation it is possible to perform the test as a combination of two tests, namely:

$$H_{01}: \quad \mu \leq m \qquad \text{against} \qquad H_{11}: \quad \mu > m$$

and

$$H_{02}: \quad \mu \geq m \qquad \text{against} \qquad H_{12}: \quad \mu < m$$

The original hypothesis is rejected if either of these separate tests leads to rejection and "accepted" if none of them does. This way of testing has the advantage that it is simple to perform and that only existing tables are used. A disadvantage is that when a test of the size α is performed the power is only that of a t-test of size $\alpha/2$ if the variance is small. Hodges and Lehmann suggest therefore an unbiased, modified t -test. They also give diagrams of the critical levels of this test.

The boundary of the enlargement of H_0 would often be rather arbitrary and the gain could be small in knowing that the power at this border is at most exactly 5% (for example). A lower power at this border compared with an ordinary test would mean a lower power also for important alternatives. It is thus necessary also for an enlarged hypothesis to judge the power function for several values and to adjust α and n to obtain an appropriate test procedure.

For some applications this enlargement of the nullhypothesis has proved a successful way to facilate the interpretation. Such a case is the tests for bioequivalence (H_{0B} above) which are made in order to get a new formula of a drug registered. The authorities sets limits for an enlarged hypothes s of equivalence and the manufacturer has to investigate this approximate equivalence.

2.3. Interchange between the null and alternative hypotheses

Ofter the desired and expected statement after an investigation is that a hypothesis is (approximately) valid. This is for example the case at control of bioequivalence of a new formulation of a drug or at the control of side-effects of a drug.

Statements that a hypothesis, say "$\mu = 0$" or "$|\mu| < d$" is true are often made on the base of a non-significant test of this hypothesis in spite of no control of the power of the test. To avoid this obvious misuse of statistical tests and to get a more appealing formulation, it has been suggested that the null and alternative hypotheses change places. Instead of testing $H_0: |\mu| < d$ against $H_A: |\mu| \geq d$ a significance test is made of H_A against H_0. Such tests are described by e.g. Lehmann (1959, p.88), Hauck and Andersson (1984) and Dahlbom and Holm (1986). Usually the same kind of limits of the same testcharacteristic is used as in an ordinary test of H_0 against H_A. For this situation and with identical limits for the example with $H_0: |\mu| < d$ och $H_A: |\mu| \geq d$ the difference will be described:

Let $P(\mu) = \Pr$ (reject H_0) be the powerfunction of an ordinary test of H_0. The misuse mentioned in the beginning av this section was that the statement "H_0 valid" was made when only the size of this test, $\sup\limits_{|\mu| < d} P(\mu)$, but nothing else of the powerfunction was controled. With the interchange of hypotheses the size of the test of H_A, that is $\sup\limits_{|\mu| \geq d}(1 - P(\mu))$, is controled. This is an important improvement but the best must be not to control only one measure of the powerfunction but to know as much as possible about the powerfunction. Mandallaz and Mau (1981) have in a simulation study illustrated how large the probability of falsely rejecting H_0 can be when methods with only control of $\sup\limits_{|\mu| \geq d}(1 - P(\mu))$ are used. There are several equivalent formulations of methods. In fact in the paper of

Mandallaz and Mau mentioned above the hypotheses are formulated in the ordinary way but methods which are formulated as rules based on confidence intervals and which are equivalent to tests of the alternative hypotheses are examined.

The size of the test of the interchanged hypothesis of e.g. $H_A: |\mu| \geq d$ has caused some confusion. A clear derivation is found in Dahlbom and Holm (1986) . According to Westlake (1981) the regulatory agencies might suspect that the use of the true size of a test of bio-equivalence would appear as a too relaxed standard. Partly because of this pedagogical reason he suggests a rule based on "symmetrical confidence intervals". The probability that such an interval covers the true value is not constant but depends on the parameter value.

2.4. Absolute index of deviation

Martin-Löf (1970, 1974) suggests "redundancy" as a measure of how good a model is. Redundancy is a measure used in information theory. Martin-Löf also presents a scale in which values of this characteristic correspond to "very bad" , "bad" , "good" and "very good" fit of the model. He has used this method in judgements of models used in traffic research (Jönrup & Svensson 1971) .

Such an index would be very useful, but this method is controversial as is obvious by the discussion that followed the paper of 1974 . It doesn't seem possible to state whether a model is "good enough" without considerations about the specific use of the model, i.e. what it is supposed to be good enough for. A numerical deviation from a hypothesis can be negligible for one purpose while it is of the greatest importance for another.

2.5. Bayesian inference

A comparison between posterior probabilities of H_0 and H_1 by the posterior odds depends on the sample size in quite a different way than in a significance test. For smaller sample sizes it gives smaller odds for H_0 in cases where the significance test would give the same evidence (e.g. just significant on the 5% --level). This coincides in many cases with the intuition. For large sample sizes we want smaller p -values to consider the difference to be important. However, the Baysian theory implies other difficulties and it is not generally accepted among applied statisticians.

Illuminating Bayesian formulations of some of the methods used for test of bioequivalence are given by Mandallaz and Mau (1981) and by the editor in a discussion of the paper by Kirkwood (1981) .

2.6. Decision theory

Sometimes you may have specified losses for actions based on H_0 and H_1 respectively. Then some decision rule, such as minimax, may be the appropriate solution. Often, however, it is not possible to specify the losses.

2.7. Powerfunction analysis

The powerfunction of a test contains the information about the properties of the test. These properties should be adjusted to meet the requirements of the application.

The powerfunction should be examined before a test is performed to make sure that the test procedure has reasonable characteristics. The powerfunction contains all information about the procedure. A few characteristics of the curve might be enough but attempts to characterize the procedure by only one measure leeds to difficulties as was seen above. The level and the slope of the power function are adjusted by the level of significance and the sample size. Too high power for alternatives close to H_0 can thus be avoided by letting α depend on n . Close alternatives to H_0 do often have nearly the same consequences as H_0 has. A low power is thus desirable for these alternatives. Construction of locally most powerful tests has the opposite aim, namely to maximize the power in the immediate surrounding of H_0 . However, tests with steep powerfunctions have good large-sample properties.

In some cases, for instance when there are nuisance parameters, the examination of the power function can be complicated. But by approximations, estimates and examples it will generally be possible to illuminate the power to some extent. The desired power function is of course dependent on the application. A discussion with the expert on the application about the power for some relevant examples of situations both close and far from H_0 would thus be necessary.

One kind of problem has a special status in these respects, namely those where the application is another statistical method. That is, a preparatory test is performed to decide about the main statistical analysis. This means that the statistician is the expert of the application and can settle the question of suitable power. There are thus possibilities of a unified and theoretical treatment of the problem. Some cases where preparatory tests for different main methods are relevant will be discussed below.

3. SOLUTIONS FOR PREPARATORY TESTS

3.1. General formulation

Let the test characteristic of the main test be T and the rejection region for T at test on the desired level α_T be R_T . A preparatory test of some assumption of the main test is often performed. Often, the implicit aim of this preparatory test is to make sure that the desired significance level α_T is not much exeeded. Let Q be the test characteristic of the preparatory test and R_Q the rejection region of Q . R_Q is thus the region of Q corresponding to the decision that the assumption of the main test was not a close enough approximation for the main test to be used on the nominal level α_T .

Let α'_T be a constant larger than α_T and A a set of models

$$A = \left\{ a : \Pr\,(T \in R_T \mid a) \geq \alpha'_T \right\}$$

Let P_Q be defined by

$$\sup_{a \in A} \Pr\,(Q \in R_Q \mid a)$$

Let α_Q be the probability of rejection in the preparatory test when the assumption M about the main test is exactly fulfilled, that is:

$$\alpha_Q = \Pr\ (Q \varepsilon R_Q \mid M)$$

The risk of wrongly judging the assumption as not fulfilled is thus given by α_Q . The risk of wrongly judging the assumption as approximately fulfilled is given by $1 - P_Q$. P_Q is depending on the set A . However, it is not necessary to specify this set explicitly in order to compute P_Q . The implicit definition by the relation between α_T and α'_T will be sufficient. It is important that the main test and the preparatory test correspond as is clear from the formulation above.

3.2. Evaluation of a medical diagnostic method

The problem to test whether a hypothesis is approximately true was present in an investigation of the visual field of the eye (Frisén 1974). This investigation was initiated by a hypothesis that normal (healty) persons, in contrast to people with certain diseases, have elliptical isopters. An isopter is a representation of the locus of points on the retina with the same visual capacity. If the above hypothesis is true or approximately true the ellipticity of isopters might be a diagnostic aid. As the isopters are observed with stochastic error, a statistical test, the main test, was constructed, on the basis of the characteristics of the ellipse, so that the power against those departures from elliptical shape which are present in diseases was high. The test characteristic in this test was named T . The hypothesis that normal isopters satisfy elliptical shape well enough for T to be of diagnostic value was tested in a clinical trial by a preparatory test with the test characteristic Q .

R_T is the rejection region for T at test at the significance level α_T . R_Q is the region for Q corresponding to the decision that normal isopters differ too much from elliptical shape for T to be useful as a test characteristic. A is the set of models for which

$$\Pr\ (T \varepsilon R_T) \geq \alpha_T'$$

where α_T' is a constant larger than α_T . P_Q is the lower boundary of $\Pr\ (Q \varepsilon R_Q)$, when the model is a member of A . α_Q is the value of $\Pr\ (Q \varepsilon R_Q)$ when the model is an ellipse.

The term "elliptical enough" above can be specified by α_T and α_T' . The risk to wrongly judge the normal isopter elliptical enough for T to be useful is then specified by $1 - P_Q$. The risk to wrongly judge the normal isopters not elliptical enough is specified by α_Q .

This means that any alternative to elliptical shape which has the probability of at least α_T^1 to be detected in future examination of a patient, by test on the significance level α_T , has at least the probability P_Q to be detected by the present experiment. On the other hand, the probability to reject the test characteristic T, when normal isopters are exact ellipses except for stochastic variation, is α_Q . By medical judgements numerical values were allotted to α_T , α_T' , P_Q and α_Q .

The resulting test procedure was examined by calculation of P_Q for several values of α_T' .

3.3. Test of homoscedasticity

Analysis of variance requires homoscedasticity. It is sometimes recommended that one should begin with a preparatory test of homoscedasticity and proceed with the analysis of variance in unchanged form when - and only when - the first test does not lead to rejection. The usual test of homoscedasticity, Bartlett's test, is very sensitive for departures from the assumtion of normality. The procedure has therefore been compared to a trip in a rowboat to check whether the sea is calm enough for a steam ship. However, there are other problems than the possible departures from normality. Even if a test demonstrates that there are departures from homoscedasticity, this does not imply that analysis of variance should not be performed. The departure might still be so small that the effect of the analysis of variance is negligible for the practical purpose. On the other hand there might be departures which invalidate the analysis in spite of no rejection in a test of homoscedasticity.

There are two possible consequences of an error in the condition (homoscedasticity) for the main test (analysis of variance). The first is that the probability of rejection when H_0 is true might be larger than the nominal significance level. The second is that the power for alternatives in H_1 might be less than what it had been if the condition was fulfilled.

The first consequence usually causes concern while the second one can be neglected. If this is the case, then the formulation given above can be used directly. T would be the test characteristic in the main test (analysis of variance) while Q would be the test characteristic of the preparatory test (e.g. Bartlett's test). If also the second consequence is to be considered, the formulation is somewhat more complicated, but follows the same lines.

3.4. Choice of parametric or non-parametric methods

A widely accepted and used procedure is to use a test of goodness-of-fit (on a conventional level of significance). Parametric methods are then chosen according to whether H_0 is accepted or rejected.

An examined situation (Frisén 1982) , is the common one where a t -test is considered and chosen if a Kolmogorov-Smirnov test on the 5% level "accepts" the hypothesis of normal distribution.

It was demonstrated that with a fixed level of significance the deviation will not be detected when the sample size is small, and thus the effect is serious, but will be detected when the sample size is large, and the deviation doesn't matter.

A more reasonable procedure in this respect was achieved with a constant critical value.

3.5. Choice of prognosis model

The choice of model for prognosis is often guided by a preparatory test. The hypothesis that a tentative model is true is tested. If the hypothesis is "accepted" the model is used for prognosis but if the hypothesis is rejected on some (often arbitrary) significance level another model is tried. This procedure of testing models is often done systematically on a large number of models. An example of this is the widely spread use of the standard programs of stepwise regression. The most commonly used versions of this method does not take into account the multiple test situation. Modifications to ensure that the test really has the claimed significance level have been suggested e.g. by Mohn & Volden (1972) . However, a correct specification of the size of the test does not solve the problems connected with the dilemma discussed in this paper. For a fixed level of significance the complexity of the resulting model will be strongly dependent on the size of the sample used for the preparatory test. This is a warning against the uncritical use of the procedure.

The problem could be approached as suggested in Section 2.6. by specifying the desired power for important alternatives. The "desired" power could be derived from a specification of the optimality criterion of the prognosis model. This specification of what, exactly, is meant by a "good" prognosis will be a valuable step in all construction of methods for prognosis, anyhow. A criterion which seems to be relevant for a vast number of applications is a minimum mean square deviation between the prognostic and true value.

The problems of choosing variables in a linear regression model (Frisén & Palm 1981) and of choosing the order of an AR model (Frisén 1979) can be treated in this way.

3.6. Preparatory test for estimation

The estimator to be used in a specific application is dependent on the conditions. Often a preparatory test on some condition (e.g. normality) is made.

There is an interesting case of how a condition has influence on the choice of estimator in the theory of aggregation. Chipman (1976) has derived a criterion for the choice between two estimators. This criterion is of the type where the estimator β is preferred over β^* when and only when a parameter λ of the model exceeds a numerical value, say λ_k . A statistical test of the null hypothesis that $\lambda = \lambda_k$ is also given by

Chipman (1976) . In order to take full advantage of Chipmans new result, it is of value to analyse the relation between the preparatory test (of $\lambda = \lambda_k$) and the statistical features of the estimators. Conditionally on the result of the test these features are not the same as unconditionally. Also, the sample size will strongly influence the symmetry of the procedure with respect to the two estimators β and β^* . This dilemma is again solved by a proper choice of the power of the preparatory test. These considerations are very similar to those described for prognoses in Section 3.5 .

REFERENCES:

Berkson, J. (1938). Some difficulties of interpretation in the chi-square test. J. Amer. Statist. Ass. 33 , 526-536.

Chipman, J.S. (1976). Statistical problems arising in the theory of aggregation. Mimeograph. Dayton.

Dahlbom, U. and Holm, S. (1986). Parametric and non-parametric tests for bioequivalence trials. Research Report 1986:2 , Dept. of Statistics, University of Göteborg.

Frisén, M. (1974). Stochastic deviation from elliptical shape. Almqvist & Wiksell, Stockholm.

Frisén, M. (1979). Some comments on the choice of method for time series analysis. Proc. SEAS Anniv. Meeting 1980 .

Frisén, M. (1982). On the choice between parametric and non--parametric methods. Proc. 15 th European Meeting of Statisticians.

Hauck, W.W. and Anderson, S. (1984). A New Statistical Procedure for Testing Equivalence in Two-Group Comparative Bioavailability Trials. Journal of Pharmacokinetics and Biopharmaceutics, 12 , 83-91 .

Hodges, J.L. & Lehmann, E.L. (1954). Testing the approximate validity of statistical hypotheses. J. Roy Statist. Soc. Ser. B. 16 , 261 .

Jönrup, H. & Svenson A. (1971). Effekten av hastighetsbegränsningar utanför tätbebyggelse. Statens trafiksäkerhetsråd. Meddelande 10 .

Kirkwood, T.B.L. (1981). Bioequivalence Testing - a need to rethink. Biometrics 37 , 589-594 . (With response by W.J. Westlake and the editor).

Lehmann, L. (1959). Testing statistical hypotheses. Wiley, New York.

Mandallaz, D. & Mau, J. (1981). Comparison of different methods for decision-making in bioequivalence assessment. Biometrics, 37 , 213-222 .

Martin-Löf, P. (1970). Statistiska modeller. Mimeograph. Department of Mathematical Statistics, Stockholm.

Martin-Löf, P. (1974). The notation of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. Scand. J. Statist. 1 : 3-12 . (Discussion pages 13-18) .

Westlake, W.J. (1976). Symmetrical confidence Intervals for Bioequivalence Trials. Biometrics, 32 , 741-744 .

GÖTEBORGS UNIVERSITET
STATISTISKA INSTITUTIONEN

GRÖNA SERIEN RESEARCH REPORT

1975:1	Högberg, Per	Estimation of parameters in models for traffic prediction - A new approach
1975:2	Frisén, Marianne	The use of conditional inference in the analysis of a correlated contingency table
1975:3	Högberg, Per	Planning of traffic counts
1975:4	Jonsson, Robert	A branching poisson process
1975:5	Wold, Herman	Modelling in complex situations with soft information
1975:6	Areskoug, B., Lyttkens, E and Wold, H.	Six models with two blocks of observables as indicators for one or two latent variables
1976:1	Blomqvist, Nils och Svärdsudd, Kurt	Om sambandet mellan blodtryckets tillväxthastighet och nivå.
1976:2	Blomqvist, Nils	On the relation between change and initial value
1976:3	Wold, Herman	On the transition from pattern cognition to model building
1976:4	Blomqvist, Nils	Skattning av imprecision vid samtidig jämförelse av flera mätmetoder,
1977:1	Klevmarken, N. A.	A comparative study of complete systems of demand functions
1977:2	Eriksson, Bo	An approximation of the variance of counts for a stationary stochastic point process

1978:1	Eriksson, Bo	Approximation of the variance for the estimated mean in a stationary stochastic process
1979:1	Klevmarken, Anders	Utjämning av lönekurvor
1979:2	Klevmarken, Anders	On the complete systems approach to demand analysis
1979:3	Jonsson, Robert	A branching poisson process model for the occurrence of miniature endplate potentials
1980:1	Flood, L. och Klevmarken, A.	Prognosmodeller för fördelning av den totala privata konsumtionen på 65 varugrupper
1980:2	Creedy, J., Hart, P.E., Jonsson, A and Klevmarken, A.	The distribution of cohort incomes in Sweden 1960-1973
1980:3	Klevmarken, A.	Age, qualification and promotion supplements. A study of salary formation for salaried employees in Swedish Industry
1980:4	Jonsson, A.	A general linear model approach for separating age, cohort and time effects
1980:5	Westberg, Margareta	Kombination av oberoende statistiska test
1981:1	Arvidsen, Nils och Johnson, T.	Variance reduction through negative correlation, a simulation study
1981:2	Eriksson, Sven	Kommunurval, väljarurval och analysansatser
1981:3	Westberg, Margareta	The combination of independent statistical test. A comparison between two combination methods when the test statistics either are normally or chi-square distributed.
1981:4	Frisén, Marianne	Evaluation of a stochastic model for visual capacity by two observational studies.

1981:5	Arvidsen, N. och Johnsson, T.	Sampling. An interactive computer program for survey sampling estimation.
1982:1	Klevmarken, A.	Age, Period and Cohort analysis: A survey.
1982:2	Johnsson, T.	Household market and non-market activities - design issues for a pilot study.
1982:3	Klevmarken, N.A.	Household market an non-market activities.
1982:4	Klevmarken, N.A.	Pooling incomplete data sets.
1983:1	Flood, L.	Time allocation to market and non-market activities in Swedish households.
1983:2	Eriksson, S.	Analys av kategoriska data. En metodstudie i anslutning till statsvetenskaplig forskning.
1983:3	Klevmarken, N.A.	Asymptotic properties of a least-squares estimator using incomplete data.
1984:1	Klevmarken, N.A.	Econometric inference from survey data.
1985:2	Guilbaud, Olivier	Stochastic order relations for one-sample statistics of the Kolmogorov-Smirnov type.
1985:3	Frisén, M.	Unimodal regression.
1985:4	Frisén, M. och Holm, S.	Nonparametric regression with simple curve characteristics.
1985:5	Jonssson, R.	Methods for discriminating betwwen children with the fetal alcohol syndrome and control children on the basis of measurements of ocular fundi. - Some procedures for explorative analysis, tests and individual discrimination.
1985:6	Westberg, M.	An adaptive method of combining independent statistical tests.
1986:1	Johnsson, T.	Multiple comparison tests based on the bootstrap.
1986:2	Dahlbom, U. Holm. S.	Parametric and nonparametric tests for bioequivalence trials.

1986:3	Westberg, M.	A Tippett-adaptive method of combining independent statistical test.
1986:4	Jonsson, R.	Point- and interval estimation of the normal tail probability.
1986:5	Frisén, M.	Testing the approximate agreement with a hypothesis.