STATISTICAL RESEARCH REPORT No 7
Institute of Mathematics August 1977
University of Oslo

THE LOGIC OF STATISTICAL INFERENCE
SIGNIFICANCE TESTING AND DECISION THEORY

by

Erling Sverdrup

THE LOGIC OF STATISTICAL INFERENCE: SIGNIFICANCE TESTING AND DECISION THEORY

Erling Sverdrup
University of Oslo, Norway

1. INTRODUCTION

The title of this lecture contains the word "decision theory", which to some statistician would suggest that I am going to talk in very general and abstract terms. I may put some statisticians to ease and perhaps disappoint others by saying at once that so will not be the case.

My lecture will consist of two parts. The first part will discuss the meaning of the classical type of significance testing; which seems to have held its position among statisticians despite the heavy critisism it has sometimes been subject to. In the second part of my lecture I am going to deal with renewed attempts to construct test methods essentially from the distribution of relevant statistics under the null-hypothesis. These are ideas which have re-emerged after having received death-sentence many years ago when the Neyman-Pearson theory and the likelihood principle were widely accepted among statisticians.

*) To be presented to 41st session of the International Statistical Institute, New Dehli 5-15 December 1977. (22 pages + 3 appendices)

2. THE MEANING OF SIGNIFICANCE TESTING

So let me then first go into the problem of the meaning of significance testing, which sometimes has been the concern of statisticians; some of whom have even found it contradictory to test null-hypotheses which are known to be false and would have been rejected anyhow if the number of observations were large enough. I have been in some doubt if statements of this kind are meant to be taken seriously as an objection to significance testing in general, or just as a warning against misuse. In any case I hope that my considerations should have some relevance.

The purpose of many statistical investigations is to find important effects which depend upon the unknown parameters in the model. To fix the idea, let us think of $p = (p_1, p_2, \ldots)$ as having components of binomial probabilities. We are interested in <u>effects</u> $f(p)$. Thus we consider a class of functions f. An effect f "exists" if $f(p) > 0$ and is non-existent if $f(p) = 0$. (If $f(p) < 0$ then $-f$ exists.) Let now H_o be the set of all p for which $f(p) = 0$ for all interesting effects. Reversing this construction let $\mathcal{F}$ be the class of all f which are such that $f(p) = 0$ for $p \in H_o$. $\mathcal{F}$ is the set of <u>contrasts</u> relatively to H_o.

H_o is the "<u>null-hypothesis</u>", but should properly be called the <u>null-state.</u>

As an illustration consider the case where $p_1, p_2, \ldots$ are binomial probabilities at equidistant points of time. We may be interested in the ups and downs of p_i hence in effects $p_i - p_j$. This leads to

$$H_o : p_1 = p_2 = p_3 = \ldots$$

On the other hand our main interest could be in the curvature, i.e. in the "escalating" effects $p_{i+1} - 2p_i + p_{i-1}$, which leads to

$$H_o : p_i = \alpha + \beta i \quad ; \quad i=1,2,\ldots$$

with α and β unknown parameters under H_o.
We thus choose the null-hypothesis H_o, not because we have any a priori confidence in it, or are interested in the truth of it, but because we are interested in certain effects which are contrasts relatively to the hypothesis. We may even know in advance that the null-state H_o cannot be true. Hence the term null-state is more appropriate than null-hypothesis.

Now let us consider two situations treated in any decent text book of statistics. The first one is the one-way lay-out for normally distributed variables, the other one is homogeneity testing by multinomial distributions.

In the first case we observe X_{ij} ; $i=1,2,\ldots,n_j$; $j=1,2,\ldots,r$, which are independent and normally distributed with $\operatorname{var} X_{ij} = \sigma^2$ and $EX_{ij} = \xi_j$ (unknown). We are interested in comparing different ξ_i, for examples in pairs $\xi_i - \xi_j$, or if one group of ξ_j on the average is greater than another group, or if ξ_j is covariant with some quantity t_j, $\Sigma\, \xi_j(t_j-\bar{t}) > 0$, or if the influence of t_j on ξ_j is accelerating $(\xi_{i+1}-\xi_i)/(t_{i+1}-t_i) - (\xi_i-\xi_{i-1})/(t_i-t_{i-1}) > 0$ and so on. In short, we are interested in discovering contrasts $\Sigma_{j=1}^{r} f_j\xi_j$ (with $\Sigma\, f_j = 0$) which are > 0. Obviously the null-state is $\xi_1 = \ldots = \xi_r$. According to Scheffé's wellknown method it should be asserted that $\Sigma\, f_j\xi_j > 0$ if

$$\Sigma\, f_j\bar{X}_j > \sqrt{(r-1)c}\; S \sqrt{\Sigma\, f_j^2/n_j} \qquad (1)$$

where $\bar{X}_j$ is the class average, S^2 is the usual unbiased estimate of σ^2 with n-r degrees of freedom and c is $1-\epsilon$ fractile of the Fisher distribution with r-1 and n-r degrees of freedom ($n=\Sigma n_j$). In particular we state $\xi_i > \xi_j$ if

$$\bar{X}_i - \bar{X}_j > \sqrt{(r-1)c}\; S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \qquad (2)$$

Hence we use $\sqrt{(r-1)c}$ in place of Student's fractile t with n-r degrees of freedom.

Now, consider the case of r × s contingency tables with factor A having the levels $A_1,\ldots,A_r$ and factor B the levels $B_1,\ldots,B_s$. Among n independent trials, N_{ij} of them have level $A_i \wedge B_j$. We consider $n_i = \Sigma_{j=1}^{s} N_{ij}$ as fixed. For given A_i, p_{ij} is the probability of B_j in a single trial $\Sigma_{j=1}^{s} p_{ij} = 1$. We are interested in seeing how the distribution over $B_1,\ldots,B_s$ changes with A_i. Hence we are interested in contrasts $\Sigma_{i,j}\, p_{ij} f_{ij}$ where $\Sigma_i f_{ij} = 0$ for each j. This leads to the null-state of homogeneity

$$p_{ij} = p_j \text{ for all i and j}$$

where $p_1,\ldots,p_s$ are unknown parameters in the null-state. We do not believe, of course that this state could be true at all.

The maximum likelihood estimates a priori of p_{ij} are

$$p^*_{ij} = N_{ij}/n_i$$

whereas the maximum likelihood estimates in the null-state are

$$\hat{p}_{ij} = \hat{p}_j = \Sigma_i\, N_{ij}/n = N_j/n$$

The observed contrasts $\Sigma f_{ij} p^*_{ij}$ are to be compared with the standard deviation $\sigma_f(p)$ of $\Sigma f_{ij} p^*_{ij}$. In $\sigma_f(p)$ we may either use null-state estimates of p or a priori estimates of p, to obtain the two slightly different criterions for stating $\Sigma f_{ij} p_{ij} > 0$, viz.

$$\Sigma p^*_{ij} f_{ij} > \sqrt{c} \sqrt{\Sigma_i \frac{1}{n_i} (\Sigma_j f^2_{ij} \hat{p}_j - (\Sigma f_{ij} \hat{p}_j)^2)} \qquad (3)$$

or

$$\Sigma p^*_{ij} f_{ij} > \sqrt{c} \sqrt{\Sigma_i \frac{1}{n_i} (\Sigma f^2_{ij} p^*_{ij} - (\Sigma f_{ij} p^*_{ij})^2)} \qquad (4)$$

In both cases c is the $1-\epsilon$ fractile of the chi-square distribution.

Now, what relationships are there between the three methods which I have described and the classical tests used in those situations?

It is well known that (1) is true for at least one $f = (f_1, \ldots, f_r)$ if and only if

$$Z = \Sigma n_j (\bar{X}_j - \bar{\bar{X}})^2 / (r-1) S^2 > c \qquad (5)$$

(where $\bar{\bar{X}}$ is the total mean). Similarly (3) takes place for some contrast $\{f_{ij}\}$ if and only if the ordinary chi-square

$$Z_1 = \Sigma_{i,j} \frac{(N_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j} = n(\Sigma \frac{N^2_{ij}}{n_i N_j} - 1) > c \qquad (6)$$

In order to find a similar relationship in the case of (4), we have to find the minimum chi-square estimates of p_j under the hypothesis, using observed numbers in denominator. This gives the following estimates of p_j;

$$\hat{\hat{p}}_j = \bar{p}_j / \Sigma_{i=1}^r \bar{p}_i \quad \text{where } \bar{p}_j = n / \Sigma_i \frac{n_i}{p^*_{ij}}$$

Then (4) takes place for some contrasts if and only if

$$Z_2 = \sum_{i,j} \frac{(N_{ij} - n_i \hat{\hat{p}}_j)^2}{N_{ij}} = n(-1+1/\Sigma \bar{p}_j) > c \qquad (7)$$

Note that the relations between Z_1 and Z_2 and the corresponding multiple decisions procedures are strictly true algebraic relations. There is nothing asymptotical about them, as we are used to in the case of chi-square goodness of fit tests.

From these results it follows that under the null-hypothesis the probability of finding a false effect is (exactly or approximately) equal to ϵ. How interesting is this result? Of course it is of very little interest. This is obvious if you are sure in advance that the null hypothesis cannot be true. However, even if the hypothesis may be true, it is uninteresting. Because the mathematical result in itself says nothing about the probability of stating that $\Sigma\, p_{ij} f_{ij} > 0$ for any set of p not consistent with this statement. The true set of p's may be such that $\Sigma\, p_{ij} f_{ij} \leq 0$. What is then the probability of stating the opposite? Clearly the failure to say anything about that and instead just make a statement relatively to the null-hypothesis, lend support to the point of view that testing hypothesis known to be wrong is an absurdity.

Fortunately, the situation is not so bad. The mathematics needed to solve it is very easy. From a statistical point of view it is very interesting that it can be proved that

$$\Pr\Big(\bigcup_{f:\Sigma \xi_j f_j \leq 0} (\text{stating } \Sigma\, \xi_j f_j > 0) \,|\, \xi\Big) \leq$$

$$\leq \Pr\Big(\bigcup_{f} (\text{stating } \Sigma\, \xi_j f_j > 0) \,|\, \xi \in H_o\Big) = \epsilon$$

in the analysis variance situations. A similar result is true in the general linear-normal situation.

It is now seen that we are in the happy and entirely new situation of having freed ourselves completely from any nullhypothesis. Using the rule (1) mentioned we know that the probability of committing at least one error is at most ε for any value of $(\xi_1,\ldots,\xi_r)$. We are not hampered by having to refer back to the nullhypothesis.

Nothing can, of course, prevent us from performing the test in the following manner. Ascertain first if the variance ratio Z is $> c$. If it is not, then drop the whole statistical analysis. If it is true, then we may look around for interesting effects. Numerical convenience may justify such a procedure. But then we have also justified "testing" the null-state. Thus we may test without having an hypothesis. The hypothesis or rather, null-state is just there to generate the class of contrasts. As the number of observations goes to infinity the probability that $Z > c$ goes to 1, and we will almost certainly "reject". That is as it should be, and it should not make statisticians unhappy.

In the case of being interested in only one scalar parameter β in the linearnormal situation, the set of all possible effects are $a\beta > 0$ for different a. Thus the null-state is a Student hypothesis $\beta = 0$ and we have just the choice between $\beta < 0$, $\beta > 0$ (according as $a < 0$ or > 0) or saying nothing. We have a three-decision problem. We are not interested in accepting the nullhypothesis (the null-state). We are not interested in rejecting it as false either, because we know that. We simply state that $\beta <$ or > 0 according as the estimate $\hat{\beta} < 0$ or > 0 after having obtained clearance by Student's significance testing. The

method uniformly maximizes the performance among all performance unbiased methods. That means that we want to avoid stating that $\beta > 0$ when $\beta < 0$ and vice versa and we want optimal chance of discovering that $\beta > 0$ (resp. $\beta<0$). That is the purpose of Student's test. We are completely relieved from the burden of any null-hypothesis.

Returning again to the multinomial trials let me first comment upon a technical point. It is true under very general assumptions with a null-state $p_{ij} = \varphi_{ij}(\theta_1,\ldots,\theta_t)$, that in order to obtain compatibility between multiple comparison and chi-square goodness of fit testing you should proceed as follows. With null-state estimated variance - which is the conventional manner of doing it - you should use maximum likelihood estimates and expected numbers in the denominator of the chi-square when testing. With a priori estimated variance you should use chi-square minimizing estimates with observed numbers in the denominator and the corresponding minimum of chi-square when testing.

Now it can be proved that the probability of making a false statement is asymptotically at most equal to the level ϵ regardless of the value of the p's, even if they vary with n = number of observations, perhaps in such a manner that they go to a value consistent with the null-state. There are, of course, some restrictions which I shall not go into. Thus again we are completely freed from the null-hypothesis and the test is just a <u>clearance test</u>, allowing us to say something.

In the special case of double dichotomy

	B	not-B	Sum
A	X	M-X	M
not-A	L-X		n-M
Sum	L	n-L	n

we would suspect positive or negative dependence between A and B according as X is $> \frac{M}{n}L$ or $< \frac{M}{n}L$ and we are permitted to make one of those statements if the well known chi-square $(LM-nX)^2/nL(n-L)$ with one degree of freedom is significant. We certainly must perform this test (or the exact hypergeometric test) to be able to infer dependence. We have no worries whatsoever because independence is excluded a priori.

If we use exact testing with cumulative hypergeometric distribution H(x), then we state negative and positive dependence according as $H(X) < \varepsilon$ or $> 1-\varepsilon$. If we add an unimportant randomization we have again a method which uniformly maximizes the performance among performance unbiased methods at a certain level.

As the benevolent audience would perhaps have realized, it has not been my purpose to advocate the use of multiple comparison procedures, even if I would be willing to do that also. My purpose has been to give a reasonable interpretation of a large class of significance tests, which have persisted to be in common use despite the doom that has been hanging over their heads. The classical Karl Pearson test, as we know it today, which to the old generation of statisticians was the very embodiment of statistical testing, has almost never been a two decision problem. Significance has always meant scrutinizing data. The progress that was made by Scheffé was to define the last part of the procedure in rigorous terms.

There may have been one exception viz., the testing that grouped independent and identically distributed variables have density of a certain form, e.g. one density in the Karl Pearson system. That is the kind of test that is seldom recommended today. It is out of date. This, I think, proves my point. It is the chi-square test as a clearance test wich has survived.

Returning to my main point, significance testings have been discredited because they have been interpretated by means of a wrong decision space. Any statistical situation must be described by means of three elements, the model, the decision space and (at least in principle) the loss function. It is the decision space that has become cripled in the usual presentation of significance testing. The desicion space contains neither rejection nor acceptance of the hypothesis. However, the decision space contains the decision not to state anything as a possible decision. The purpose of significance testing is to see if this possibility could be excluded. Perhaps clearance testing would be a better word than significance testing.

Somebody may perhaps find it peculiar that the construction of the test requires derivation of the sampling distribution in the null-state, that is, under an assumption that cannot be true. Perhaps that is the reason why somebody has felt compelled to attach credence to the hypothesis. However, the logic behind the presentation of significance tests as clearance tests should be clear enough.

Let me add parenthetically, deviating somewhat from my main theme, that the multiple study of contrasts by binary and multinary observations is a very basic problem in statistics.

It is really a problem of reading (contingencies) tables of the kind that are published in large quantities by goverment statistical bureaus in all countries. The reading is performed every day by statisticans, often by very crude methods, or without any method at all. The methodological problem is not easy.

You may discover interesting features and want to test if they are real. You cannot use the method which would have been adequate if you had suspected the relationship in advance. Hence you have to adopt a soul searching attitude of defining the state of your mind before you look at the data. Some may object to such a procedure. However, it is good to be reminded that statistical inference concerning historical observations is as subjective as just that. On the other hand, to discard historical data altogether is a too easy way out of the difficulties. They may contain important informations. A general admonition to exercise caution is not satisfactory. The warning must be worked out in rigorous terms. That is what one attempts to do by the multiple comparison approach. This approach amounts to additional insight into the performance of the method. You have knowledge of the all over probability of making an error. Perhaps some other properties of the performance should be studied, e.g. within a limited class of decisions, the expected number of errors.

Consider a 12 × 2 contingency table for "testing" homogeneity, or rather, discovering interesting deviations from homogeneity. Suppose you test a difference between two frequencies, which you find interesting, at a 5% level by conventionally using a critical value

of 1.64 for the ratio between difference of frequencies and standard deviation. Then the all-over level would be the awesome 99.4%. If you want an all over level of 5%, you must use 4.44 as critical value for the ratio just mentioned. I don't find that unreasonable. You have to pay heavily for "snooping" around in historical data.

3. A GENERAL APPROACH TO SIGNIFICANCE TESTING

It is natural to generalize my interpretation of significance testing in the following manner.

$X_1, X_2, \ldots, X_n$ are independent with the same density $p(x;\theta)$, where $\theta = (\theta_1, \ldots, \theta_r)$. The null-state is to the effect that $\varphi(\theta) = 0$; i.e. $\varphi_i(\theta) = 0$; $i=1,2,.,s$. Let $\mathcal{F}$ be the class of all functions $f(\theta)$ such that $\varphi(\theta) = 0$ implies $f(\theta) = 0$. We are interested in the existence of f-effects, i.e. $f(\theta) > 0$, and we declare this to be the case if

$$f(\theta^*) > \sqrt{c}\left\{\Sigma f_j^2(\theta^*)\gamma_{jj}(\theta^*) + 2 \sum_{j<i} f_j(\theta^*) f_i(\theta^*)\gamma_{ji}(\theta^*)\right\}^{\frac{1}{2}}$$

where θ^* is the maximum likelihood estimate for θ a priori, $f_j = \frac{\partial}{\partial \theta_j} f$, $\gamma_{ij} = \mathrm{cov}(\theta_i^*, \theta_j^*)$, c is 1-$\epsilon$-fractile of the chi-square distribution with s degrees of freedom. On the right hand side null-state maximum likelihood estimates may also be used. Then the probability of falsely stating at least one effect $f \in \mathcal{F}$ is at most ϵ. Furthermore stating at least one effect f is asymtotically in probability equivalent to

$$-2 \log(\text{Likelihood ratio}) > c$$

Thus we are back on the good old standard error computations in large samples, but now combined with the likelihood ratio.

This outline should be stated in rigorous terms and proved, if it has not already been done.

4. CURVE-FITTING BY MEANS OF A "FALSE" MATHEMATICAL EXPRESSION

Closely related to the situation of testing without an hypothesis is the situation of smoothing sequences of observations by means of a simple analytical expression. The demographers and the actuaries have been interested in this problem and they have been smoothing data as part of a statistical analysis long before exact probabilistic approach became common in statistics. The mortality as a function of age is smoothed by a very crude expression, disregarding many significant variations in the mortality with age, which for special purposes they do not want to be bothered with. Perhaps a common attitude has been that if you are really interested in describing the observations by means of such a crude expression, then just pretend that the expression represents expected values and derive the method accordingly. A better approach would obviously be to make no assumptions that are known to be wrong and construct the method under realistic assumptions.

We shall illustrate the idea by returning to the example of analysis of variance of a one-way lay-out.

Assume that we have discovered, by looking at the group means $\bar{X}_i$; $i=1,2,\ldots,r$, that ξ_i varies roughly linearly with some quantity t_i, $i=1,2,\ldots,r$. Then we might be interested in joint confidence intervals for all

$$\eta(t) = \alpha + \beta(t-\bar{t})$$

when t varies continuously. Here

$$\bar{t} = \frac{1}{n}\Sigma n_j t_j,\ \alpha = \frac{1}{n}\Sigma n_j \xi_j,\ \beta = \Sigma n_j (t_j-\bar{t})\xi_j / \Sigma n_j (t_j-\bar{t})^2 .$$

This is the famous problem of Working and Hotelling (1929). Let

$$M = \Sigma n_j (t_j-\bar{t})^2,\ a = \frac{1}{n}\Sigma n_j \bar{X}_j,\ b = \Sigma n_j (t_j-\bar{t})\ \bar{X}_j / M,$$

$$\hat{\eta}(t) = a+b(t-\bar{t}),\ c_j(t) = \sqrt{n_j}\left[\frac{1}{n} + \frac{t_j-\bar{t}}{M}(t-\bar{t})\right].$$

$$K_t^2 = rf \sum_{j=1}^{p} c_j^2(t)$$

where f is the $(1-\varepsilon)$-fractile of the Fisher distribution with r and n-r degrees of freedom. Then by using Scheffé's multiple comparison method we find that

$$\hat{\eta}(t) - K_t S < \eta(t) < \hat{\eta}(t) + K_t S$$

defines a $(1-\varepsilon)$-confidence band for the regression values.

Note that we do <u>not</u> assume $EX_{ij} = \alpha+\beta(t_j-\bar{t})$. The statement above that "ξ_j varies roughly linearly with t_j" is just a motivation, it is not a basis for the mathematical derivation of the method. Hence our method is completely rigorous. If we had assumed that $\xi_j = \alpha+\beta(t_j-\bar{t})$, then we could get a confidence band,

$$|\eta(t)-\hat{\eta}(t)| < S\sqrt{2f\left(\frac{1}{n} + \frac{(t-\bar{t})^2}{M}\right)}$$

where f is now the $(1-\varepsilon)$-fractile of the Fisher distribution with 2 and n-2 degrees of freedom and S^2 is the usual estimate of σ^2. This was Hotelling and Working's solution.

The statistician may adopt the same attitude in the case of a, say, three-way lay-out in the analysis of variance. He may be interested in estimating the representation of the means by main effects and first order interactions, but without assuming the second order interactions to be 0.

5. TESTS DERIVED FROM THE NULL-DISTRIBUTION

In the second part of my lecture I shall briefly go into the question of the basic ideas of constructing efficient test procedures. This question was discussed very thoroughly in the 1930's and the 1940's when the Neyman-Pearson theory was founded. We have been used to think that an epoch-making contribution was rendered at that time.

The question that has been raised lately is the following. Could we really reject a statistical hypothesis after having observed the most probable outcome under the hypothesis? One might perhaps be captivated by this leading question and answer it in the negative. However, after a second thought one would make a complete turn-about.

The question would of course be answered in the affirmative by any statistician. Rejecting an hypothesis after having observed the most probable outcome under the hypothesis is done by statisticians every day in his run-of-the-mill statistical work. If the hypothesis is to the effect that in n Bernoulli trials the probability of success is 0.6, then the probability of a given sequence of events is $0.6^x \ 0.4^{n-x}$, where x is the number of successes. This has its largest value for $x = n$. But certainly with $x = n = 1000\ 000$

successes you would reject 0.6, even if this is the most probable outcome.

Perhaps the following example is useful when faced with operation research people who don't like statistical inference and are satisfied when a very likely sequence of events has occurred under the assumptions about the process. The number of telephone calls during each quarter of an hour is observed. Let X_t be number of calls during the t-th quarter. The hypothesis states that the expected distance between two calls is ½ hour. Thus the traffic intensity is $\lambda=2$ calls per hour. Under the Poisson assumption the probability of x calls during $T = \frac{1}{4}$ hour is

$$\Pr(X_t=x) = \frac{(\lambda T)^x}{x!} e^{-\lambda T} = e^{-\frac{1}{2}}/2^x x!$$

Thus the most probable outcome of the time series $X_1, X_2, \ldots$ is $(0,0,0,\ldots)$. But certainly if calls never occur one would reject that average distance between two calls is ½ hour.

Now, the idea has been advanced that if you consider only "relevant" statistics, then the principle of rejecting when an unlikely event occurs and accepting when a likely event occurs, is basicly sound. It seems clear that "relevant" means minimal sufficient a priori. I am myself not able to see why this should be a basicly convincing principle when referring to relevant statistics but not when referring to the original observations, but shall go along with it anyhow.

The basic principle is then the following. (Martin-Löf 1974). Consider the relevant, i.e. minimal sufficient statistic T a priori and the minimal sufficient statistic U under the hypothesis. Thus U is a function of T. Find the conditional density of T given U under the hypothesis and reject when this density is small adjusting it to a level in the traditional manner. Note that this density

would be independent of the nuisance parameters.

The two examples which I have given do not fit this prescription. But the following example serves as counter example. It is suspected that there are far more non-paying (cheating) passengers on a tramcar line A than on a tramcar line B. Hence an inspection is made and the inspector finds the first nonpaying passenger on line A after X inspections. On line B the first non-paying passenger is found after Y_1 inspections and the second after additional Y_2 inspections. Let the probability that a passenger is nonpaying be p_A and p_B, respectively, on the two lines. Then we have

$$\Pr((X=x) \cap (Y_1=y_1) \cap (Y_2=y_2))$$

$$= p_A(1-p_A)^{x-1}p_B^2(1-p_B)^{y_1+y_2-2}$$

The sufficient statistic a priori is $T = (X, Y_1+Y_2)$.

Set $Y = Y_1 + Y_2$. We find

$$\Pr((X=x) \cap (Y=y)) = (y-1)p_B^2(1-p_B)^{y-2}p_A \times (1-p_A)^{x-1}$$

Under the null-hypothesis $p_A = p_B = p$ and $U = X + Y_1 + Y_2 = X + Y$ is a sufficient statistic

$$\Pr(U=u) = \binom{u-1}{2}p^3(1-p)^{u-3}$$

Hence under the hypothesis

$$\Pr(X=x \mid U=u) = 2(u-x-1)/(u-1)(u-2);$$

$$x=1,2,\ldots,u-2$$

which decrease from $2/(u-1)$ to $2/(u-1)(u-2)$.

We should of course reject when X is small, but get rejection for large X by the new principle. This example can be made two-sided. Then we ought of course to reject when X is close to 1 or u-2, but by the new principle we get rejection only when X is close to u-2. [If we had inspected to obtain a and b non-paying passengers respectively, we would have got

$$\Pr(X=x \mid U=u) = \binom{x-1}{a-1}\binom{u-x-1}{b-1}\Big/\binom{u-1}{a+b-1}$$

which could be called the "inverse hypergeometric distribution".]

Of course, the geometric distribution $p(1-p)^{x-1}$: x=1,2,... is a much simpler example. Regardless of p, the most probable value of x is x=1. But that does not mean that no value of p could be rejected if x=1.

Using Fisher's F to test equality of the variances in samples from two normal populations, the density of F under the hypothesis is monotonic if the number of observations are m = 3 and n > 3 respectively, resulting in the most likely result being that the three observations in the first sample are very close together. Certainly that should result in rejection. No statistician would warn against using the F-test when the numbers of degrees of freedom are small, at least not because the density is monotonic or perhaps U-shaped.

An interesting example in my opinion is the case of the dealer in bridge who gives himself 13 spades, or perhaps the best no-trump hand. Why are we not willing to ascribe 13 spades to chance despite the fact that the event 13 spades is not less probable than any other hand? The null-hypothesis that the dealer is not cheating is in this case that all the $N = 52!/(13!)^4$ combinations of the 4 hands are equally likely. If we reject this hypothesis in the case of

13 spades allotted to the dealer, it must be because other circumstances than those which follow from the density under the hypothesis are taken into account.

To be more precise let

$$G_1, G_2, \ldots, G_r$$

be r groups of four-hands-combinations which are such that all combinations x belonging to a fixed G_i are equally favourable to the dealer, whereas if $x \in G_i$ and $x' \in G_j$ with $j > i$ then x' is more favourable than x to the dealer. It certainly would be a formidable (really prohibitive) task to determine $G_1, \ldots, G_r$ but in principle they are given. Let

$$N_1, N_2, \ldots, N_r, \quad (\sum_1^r N_i = N)$$

be the number of four-hands-combinations in $G_1, \ldots, G_r$ respectively. We now let the test statistic be T(x), where T(x) is defined by $T(x) = t$ if $x \in G_t$. By the classical test principle the hypothesis (of no cheating) should be rejected if T(X) is large. This seems rather obvious from the definition of T(x). On the other hand it seems to be irrelevant whether $N_{T(x)}/N$ is large or small, i.e. which G_i contains few or many combinations. Thus it is the rules of the game, and a thorough knowledge of the game of bridge, which are required to determine the test. The test cannot be constructed from the hypothesis alone. It would be a poor statistician who would neglect the facts about bridge when constructing the test, just as it would be a poor statistician who neglect biological facts when devising a test for analyzing some biological observations. That is just the idea of the a priori spesification, to be able to mould the known facts of the

matter into the test construction. Before the Neyman-Pearson era it was recommended to do that by using statistic measuring the distance from the hypothesis. This is of course still a sound procedure. The main point is that you must bring in something which cannot be derived from the null-hypothesis alone.

The new principle may perhaps be called the null-distribution criterion. It has been called "the exact test", presumably because it is so convincing a priori. It must be taken seriously, because so many statisticians have shown interest in it. It is indeed revolutionary and violates most of what I have been used to teach my students. It is sweepingly more general than the Neyman-Pearson principle because it disregards the alternatives to the hypothesis except for the purpose of constructing a minimal sufficient statistics under the a priori assumptions.

The new principle contradicts the Neyman-Pearson point of view and the likelihood idea. Those are ideas which many statisticians have considered fundamental.

In all of the examples which I have mentioned the obvious test procedure is uniformly most powerful unbiased. I have seen no example where the null-distribution criterion leads to an obviously acceptable procedure whereas power-optimizing or the likelihood criterion does not.

I still feel that the idea of Neyman and Pearson from 1933 of judging a procedure from its power or performance is fundamental. It created a revolution in the logic of statistical inference. I also feel that the set of implications

unbiasedness in power => similarity
<=> conditional testing

is still important, both because it represents

fundamental ideas of inference and because it is a useful construction. The reasons why it is a useful construction are; first; modern computational outfit makes it possible to perform exact tests; second; statistical analysis focussed on one parameter in a regular Darmois-Koopman class of distributions will always be important; third; the construction can be generalized to situations of inference about many paramers either as stepwise procedures (T.W. Anderson) or multiple decision procedures (Erich Lehmann). I find myself repeatedly making use of conditioning, justified by power unbiasedness, in concrete practical applications. The objections that can be advanced to power unbiasedness I find less serious than those which can be made to the widely accepted concept of unbiasedness in the mean by point estimation.

BIBLIOGRAPHY

[1] Anderson, T.W. (1962). The choice of degree of polynomial regression as a multiple decision problem. Ann. Math. Stat. 33, 255-265.

[2] Bahadur, R.R. (1952). A property of the t-statistics. Sankhya 12, 79-88.

[3] Goodman, L.A. (1964). Simultaneous confidence interval for contrasts among multinomial populations. Ann. Math. Stat. 35, 716-725.

[4] Lehmann, E. (1957). A theory of multiple decison problems. Ann. Math. Stat. 28, Part I, 1-25, part II, 547-571.

[5] Loeve, M. (1955). Probability Theory. D. van Nostrand Company, Inc. New York.

[6] Martin-Löf, P. (1974). The notion of rendundancy and its use as a quantitative measure of discrepancy between a statistical hypothesis and observational data. Scand. J. Stat. 1, 3-18.

[7] Reiersøl, O. (1961). Linear and non-linear multiple comparison in logit analysis. Biometrika 48, 359-365.

[8] Scheffé, H. (1953). A method of judging all contrasts in the analysis of variance. Biometrika 40, 87-104.

[9] Sverdrup, E. (1975). Multiple comparisons by binary and multinary observations. Artikler fra Statistisk sentralbyrå, nr. 75.

[10] Sverdrup, E. (1975). Tests without power. Scand. J. Statist. 2, 158-160.

[11] Sverdrup, E. (1976). Significance testing in multiple statistical inference, Scand. J. Stat. 3, 73-78.

Appendix I

SIGNIFICANCE TESTING AND MULTIPLE COMPARISON BY LINEAR-NORMAL MODELS

The following results are perhaps well known, but is stated for the sake of completeness.

1. Suppose that X_{ij}; $i=1,2,\ldots,n_j$; $j=1,2,\ldots,r$ are independent and normally distributed with $\mathrm{var}X_{ij} = \sigma^2$, and $EX_{ij} = \xi_j$. We state that $\Sigma_{j=1}^{r} f_j\xi_j > 0$ ($\Sigma f_j=0$), if

$$\Sigma f_j\bar{X}_j > \sqrt{(r-1)c}\, S\sqrt{\Sigma f_j^2/n_j} \tag{1}$$

It is well-known (Scheffé) that (1) is true for some f ($\Sigma f_j=0$) if and only if

$$Z = \Sigma n_j(\bar{X}_j-\bar{X})^2/(r-1)\, S^2 > c \tag{2}$$

(We use the same notations as in section 2.)
If $\xi_1=\ldots=\xi_r$ then the probability of making a wrong statement is the same as the probability of stating (1) for some f. It follows that the probability of a false statement is exactly ϵ. Now, let $\xi_1,\ldots,\xi_r$ be arbitrary. The probability of making a false statement is then

$$\begin{aligned} &\Pr\Big(\bigcup_{f:\Sigma f_j\xi_j\le 0}(\Sigma f_j\bar{X}_j > K_f)\Big) \\ &= \Pr\Big(\bigcup_{f:\Sigma f_j\xi_j\le 0}(\Sigma f_j(\bar{X}_j-\xi_j)+\Sigma f_j\xi_j > K_f)\Big) \end{aligned} \tag{3}$$

where K_f denotes the right hand side of (1). However, since $\Sigma f_j\xi_j \le 0$, (3) is

$$\leq \Pr\Big(\bigcup_{f:\Sigma f_j\xi_j\leq 0}(\Sigma f_j(\bar{X}_j-\xi_j) > K_f\Big)$$
$$\leq \Pr\Big(\bigcup_{\text{all } f}(\Sigma f_j(\bar{X}_j-\xi_j) > K_f\Big) \tag{4}$$

By what we have stated above, the last term is equal to ϵ since all $X_{ij}-\xi_j$ have the same expectation.

2. In the general case let $X = (X_1,\ldots,X_n)'$ have independent and normally distributed components with variance σ^2 and

$$\xi = EX = y\beta \tag{5}$$

where y is a known $(n\times s)$ matrix of rank s and $\beta = (\beta_1,\ldots,\beta_s)'$. Let

$$\varphi = (\beta_1,\ldots,\beta_r)' \ , \ r < s \tag{6}$$

According to the general multiple comparison rule we shall state that

$$f'\varphi = \sum_{j=1}^{r} f_j\beta_j > 0 \tag{7}$$

if

$$f'\hat{\varphi} = \sum_{j=1}^{r} f_j\hat{\beta}_j > \sqrt{rc}\,S_f \tag{8}$$

where c is the $1-\epsilon$ fractile for the Fisher distribution with r and n-s degrees of freedom,

$$S_f^2 = f'g^{-1}f\hat{\sigma}^2 \tag{9}$$

$\hat{\beta}_j$ and $\hat{\sigma}^2$ are the usual estimates of β_j and σ^2, and g is such that $g^{-1}\sigma^2$ is the covariance matrix of $\hat{\varphi} = (\hat{\beta}_1,\ldots,\hat{\beta}_r)'$. It is well known that (8) takes place for some f if and only if

$$F = \hat{\varphi}'g\hat{\varphi}/r\hat{\sigma}^2 > c$$

However, this is the usual Fisher's F for testing $\beta_1=\ldots=\beta_r=0$.

Hence the probability of making a false statement if $\beta_1=\ldots\beta_r=0$ is ϵ.

Consider now for arbitrary $\varphi = (\beta_1,\ldots,\beta_r)'$ the probability of making a false statement

$$\Pr\Big(\bigcup_{f:\Sigma\beta_i f_i\leq o}\Big(\sum_{i=1}^{r} f_i\hat{\beta}_i > \sqrt{rc}\, S_f\Big)\Big) \qquad (10)$$

Let us introduce $\hat{\gamma}_i = \hat{\beta}_i-\beta_i$; $i=1,2,\ldots,r$.
Then (10) may be written

$$\Pr\Big(\bigcup_{f:\Sigma\beta_i f_i\leq o}\Big(\sum_{i=1}^{r} f_i\hat{\gamma}_i+\Sigma f_i\beta_i > \sqrt{rc}\, S_f\Big)\Big) \qquad (11)$$

Hence by the same reasoning as above (11) is

$$\leq \Pr\Big(\bigcup_{f}(\Sigma f_i\hat{\gamma}_i > \sqrt{rc}\, S_f)\Big) \qquad (12)$$

where the union is taken over all f. However, since $\hat{\beta}_i$, is least square estimate of β_i, then $\hat{\gamma}_i = \hat{\beta}_i- \beta_i$ is least square estimate of $\gamma_i= \beta_i-\beta_i = 0$; $i=1,2,\ldots,r$. Hence, by what has been stated above, (12) has probability ϵ.

Appendix II

STATEMENT AND PROOF OF THE PROPERTIES OF THE MULTIPLE COMPARISON RULE BY BINARY AND MULTINARY OBSERVATIONS

We give in this appendix a more detailed and somewhat more well-organized proof of the properties of the multiple comparison rule by multinary observations, than that which was presented by Sverdrup (1975). A correction is also made, see Propositions 2 and 3 with footnote.

1. Statement of the Properties of the General Multiple Comparison Rule.

A. The result of n independent multinary trials are observed. The series of trials may be divided into s sequences such that there are n_a trials in the a-th sequence; $a=1,2,\ldots,s$; $\Sigma n_a = n$. Each of the trials in the a-th sequence may result in one of r_a mutually exclusive events

$$A_{a1},\ldots,A_{ar_a}$$

with probabilities

$$p_{a1},\ldots,p_{ar_a}, \quad \sum_{j=1}^{r_a} p_{aj} = 1 \qquad (1)$$

respectively. We assume a priori that all p_{aj} are between 0 and 1. The observed number of times the r_a events occur are

$$N_{a1},\ldots,N_{ar_a}, \quad \sum_{j=1}^{r_a} N_{aj} = n_a , \qquad (2)$$

respectively. Let $R = \Sigma_{a=1}^{s} r_a$.

B. The null-state H is to effect that the p_{aj} are specified functions

$$p_{aj} = \varphi_{aj}(\theta);\ \theta\in\Theta\ ;\ j=1,2,\ldots,r_a;\ a=1,2,\ldots,s$$

of a parameter $\theta = (\theta_1,\ldots,\theta_t)$, where θ varies in an open set Θ in the t-space, $t < R-s$. We assume that the φ_{aj} have continuous second order derivatives, and that the rank of the R×t matrix

$$\left\{\frac{\partial\varphi_{aj}(\theta)}{\partial\theta_i}\right\}\ (a,j) = (1,1),\ldots,(s,r_s),\ i=1,\ldots,t, \qquad (4)$$

is t.

A function $f(p)$ of $p = \{p_{11},\ldots,p_{sr_s}\}$ is a contrast relatively to H if $f(\varphi) = 0$ for all θ. It will be called smooth if it has continuous first order derivatives

$$f_{aj}(p) = \frac{\partial f}{\partial p_{aj}} \qquad (5)$$

We shall consider a class $\mathcal{F}$ of smooth contrasts f with no stationary points for $p = \varphi$. Two cases will be treated.
Case (i). $\mathcal{F}$ is the set of all (or some) linear contrasts

$$f = \Sigma\, f_{aj}p_{aj} + f_o \qquad (6)$$

Thus in this case the f_{aj} are independent of p.
Case (ii). $\mathcal{F}$ is such that the class of all f_{aj} obtained by varying f in $\mathcal{F}$ is equicontinuous.

C. The statistical method can be described as follows. First the maximum likelihood estimates $\hat{\theta}$ under H are found as solutions of

$$\sum_{a,j} \frac{N_{aj}}{\varphi_{aj}(\hat{\theta})} \cdot \frac{\partial\varphi_{aj}(\hat{\theta})}{\partial\theta_i} = 0;\ i=1,2,\ldots,t \qquad (7)$$

Alternatively $\hat{\theta}$ is a modified minimum chi-square estimator, given by

$$\sum_{a,j} \frac{N_{aj}-n_a\varphi_{aj}(\hat{\theta})}{N_{aj}} n_a \frac{\partial\varphi_{aj}(\hat{\theta})}{\partial\hat{\theta}_i} = 0, \quad i=1,2,\ldots,t, \tag{8}$$

We do not care whether $\hat{\theta}$ actually maximizes the likelihood

$$L = \prod \varphi_{aj}(\theta)^{N_{aj}}$$

or, alternatively, minimizes,

$$\chi^2 = \Sigma \frac{(N_{aj}-n_a\varphi_{aj}(\theta))^2}{N_{aj}}$$

We assume that for all $\{N_{aj}\}$, (7) (or(8)) has either one or no solution. We shall let all n_a go to infinity in such a manner that $n_a/n = g_a > 0$. We assume that the probability that (7) (or (8)) has one solution goes to 1 for any p. When (7) (or (8)) has no solution we can let $\hat{\theta}$ have any value (e.g. such that it actually maximizes L). It can then be proved that plim $\hat{\theta}=\theta$, if $p=\varphi$.

Let $p^*_{aj} = N_{aj}/n_a$ and $\hat{\varphi}_{aj} = \varphi_{aj}(\hat{\theta})$. We find in case (i),

$$\sigma_f^2(p) = \text{var } f(p^*) = \sum_a n_a^{-1}[\sum_j f_{aj}^2 p_{aj} - (\Sigma f_{aj}p_{aj})^2] \tag{9}$$

In case (ii), $\sigma_f^2(p)$ can be found by linearizing $f(p^*)$ with respect to p^*-p. We then get the same expression (9) but with $f_{aj} = f_{aj}(p)$, depending on p. Now define

$$\hat{\sigma}_f = \sigma_f(\hat{\varphi}) \ , \ \sigma_f^* = \sigma_f(p^*) \tag{10}$$

These two quantities will be called respectively the null-state estimated and a priori estimated standard deviation.

The rule consists in stating that $f(p) > 0$ for all those $f \in \mathcal{F}$ for which

$$f(p^*) > \sqrt{z}\,\hat{\sigma}_f \tag{11}$$

where z is the 1-ε fractile of the chi-square distribution with R-s-t degrees of freedom. Alternatively we may use σ_f^* on the right hand side of (11).

It should be noted that if we want to test $f(p) \leq 0$ with a specified form f selected in advance, then we would have used the 1-ε fractile for the normal distribution instead of $\sqrt{z}$ (one degree of freedom for z).

D. We shall prove,

Proposition 1: In case (i) and (ii) with a priori estimated variances the limit of the probability of making a false statement; i.e. stating that $f > 0$ for some $f \in \mathcal{F}$ for which $f \leq 0$; is asymptotically $\leq \varepsilon$. More precisely

$$\limsup_{n \to \infty} \Pr\Big(\bigcup_{f(p) \leq o} (f(p^*) > \sqrt{z}\,\sigma_f^*) \Big) \leq \varepsilon \tag{12}$$

if $p = p^{(n)}$ approaches some $p^{(o)}$ as $n \to \infty$.

This also holds with null-state estimated variances $\hat{\sigma}_f^2$ instead of σ_f^{*2} provided $p^{(o)}$ equals some φ.||

Of course the case when $p^{(n)}$ is kept constant is included. However, it is desirable to let p go to some φ (e.g. in such a manner that all $\sqrt{n}\,(p_{aj} - \varphi_{aj})$ are kept constant. However, this special kind of convergence is not needed in the mathematical derivation.)

We shall also prove,

Proposition 2: In case (i) with $\mathcal{F}$ consisting of all linear contrasts and all φ_{aj} being linear *) in θ, we have if $p = \varphi$ (and hence all $f(p) > 0$ false),

$$\lim \Pr(\bigcup_f (f(p^*) > \sqrt{z}\,\hat{\sigma}_f)) = \epsilon \qquad (13)$$

This is also true with $\hat{\sigma}_f$ replaced by σ_f^* .||

E. We shall prove,

Proposition 3: Suppose case (i) with $\mathcal{F}$ consisting of all linear contrasts and all φ_{aj} linear *) in θ. Then if null-state estimated variances are used some contrast will be declared positive if and only if

$$Z = \sum_{a,j} \frac{(N_{aj} - n_a\hat{\varphi}_{aj})^2}{n_a\hat{\varphi}_{aj}} > z \qquad (14)$$

where the $\hat{\varphi}_{aj} = \varphi_{aj}(\hat{\theta})$ are maximum likelihood estimates. If a priori estimated variances are used, then some contrast will be declared positive if and only if

$$Z = \sum_{a,j} \frac{(N_{aj} - n_a\hat{\varphi}_{aj})^2}{N_{aj}} > z \qquad (15)$$

where the $\hat{\varphi}_{aj} = \varphi_{aj}(\hat{\theta})$ now are modified minimum chi-square estimates.||

Proposition 4: In case (i), (14) is a necessary condition for significant contrasts when null-state estimated variances are used and (15) is a necessary condition for significant contrasts when a priori estimated variances are used.||

Note that these relations between the multiple comparison rules on the one hand side and (14) and (15) on the other hand side are purely algebraic. They are strictly true, there are no approximations involved and

*) The assumption that φ should be linear was incorrectly left out in Sverdrup (1975).

they are not probability statements.

Proposition 5: In case (ii), (14) (resp. (15)) is asymptotically necessary condition for significant contrasts in the sense that if S_1 = "some significant contrast" and S_2 = "(14)(resp. (15)) is true", then $Pr(S_1-S_2) \to 0$ as $n \to \infty$, and $p^{(n)}$ converges in the manner described in Proposition 1.

Propositions 3-5 suggest that both in case (i) and case (ii) one might first check if (14) (or (15)) is true and only if such is the case go on to apply (11). Thus the test proposed is a refinement of the classical Karl Pearsson's significance test.

Proposition 6: A simultaneous confidence interval for all contrasts follows from,

$$\text{limsup } Pr(\bigcup_f \{\sqrt{n}\, f(p^*) - \sqrt{n}\, f(p) \geq \sqrt{z}\, \sigma_f^*\}) \leq \epsilon$$

Under the assumption of Proposition 2, we have

$$\lim Pr(\bigcup_f \{\sqrt{n}\, f(p^*) - \sqrt{n}\, f(p) \geq \sqrt{z}\sigma_f^*\}) = \epsilon \quad \|$$

Note that if the a priori estimated variance is used, then the estimate $\hat{\theta}$ is not needed in connection with the multiple comparison rule. It is only needed for checking (15).

F. It is of interest to consider the special case of homogeneity testing. Then $r_1 = \ldots = r_s = r$ and we choose as a null-state that $p_{a1},\ldots,p_{ar}$ are independent of a. This can be written

$$\varphi_{aj} = \theta_j;\ j=1,2,\ldots,r-1,\ \varphi_{ar} = 1-\theta_1-\ldots-\theta_{r-1} = \theta_r$$

We then get from (7) the maximum likelihood estimates

$$\hat{\varphi}_{aj} = \hat{\theta}_j = \sum_a N_{aj}/n = N_j/n = \hat{p}_j \tag{16}$$

and from (8) the minimum modified chi-square estimates

$$\hat{\varphi}_{aj} = \hat{\theta}_j = \bar{p}_j / \sum_1^r \bar{p}_i \tag{17}$$

where $\bar{p}_j$ is the harmonic mean of the $p^*_{aj} = N_{aj}/n_a$; a=1,2,...,s.

$$\bar{p}_j = n / \sum_a \frac{n_a}{p^*_{aj}} \tag{18}$$

The chi-square statistics are respectively for null-state estimated and a priori estimated variances.

$$Z = n\left(\sum \frac{N_{aj}^2}{n_a N_j} - 1\right), \quad Z = n(-1+1/(\Sigma \bar{p}_j)) \tag{19}$$

They are found from (14) and (15) respectively. Now it is seen that $\Sigma p_{aj} f_{aj}$ is a contrast if and only if $\sum_a f_{aj} = 0$ for j=1,2,...,r. According to the general rule with a priori estimated variances this should be declared > 0 if

$$\Sigma\, p^*_{aj} f_{aj} > \sqrt{z}\, \sqrt{\sum_a (\Sigma f_{aj}^2 p^*_{aj} - (\Sigma f_{aj} p^*_{aj})^2) n_a^{-1}} \tag{20}$$

where z is determined with (r-1) (s-1) degrees of freedom. (20) will take place for some f_{aj} if and only if

$$\Sigma \bar{p}_j < (1 + \frac{z}{n})^{-1} \tag{21}$$

where $\bar{p}_j$ is given by (18). (Note that $\bar{p}_j \leq \hat{p}_j$ and hence $\Sigma \bar{p}_j \leq 1$ with equality only if all p^*_{aj} are strictly independent of a. Thus heterogeneity is measured by the degree to which the harmonic means fall short of the arithmetic means.)

With null-state estimated variances we should state $\Sigma p_{aj} f_{aj} > 0$ if

$$\Sigma p^*_{aj} f_{aj} > \sqrt{z} \sqrt{\Sigma_a (\Sigma f^2_{aj} \hat{p}_j - (\Sigma f_{aj} \hat{p}_j)^2) n_a^{-1}} \qquad (22)$$

and this takes place for some f_{aj} **if and only if**

$$\Sigma_{a,j} p^{*2}_{aj} / \hat{p}_j > 1 + \frac{z}{n} \qquad (23)$$

2. Proof of the Assertions about the General Rule.

A. In sections A-F we shall treat the case (i) when $\mathcal{F}$ consists of linear contrasts and null-state estimate variances are used. We introduce

$$Y_{aj} = \frac{N_{aj} - n_a \hat{\varphi}_{aj}}{\sqrt{n_a \hat{\varphi}_{aj}}} \qquad (24)$$

and it is well known that in the limit, when $n \to \infty$; with $n_a = n g_a$, $g_a > 0$; then $Z = \Sigma_{a,j} Y^2_{aj}$ has chi-square distribution with R-s-t degrees of freedom. We now have in case (i) (see 1.B), since $f(\hat{\varphi}) = 0$,

$$f(p^*) = \Sigma f_{aj} \left(\frac{N_{aj}}{n_a} - \hat{\varphi}_{aj} \right)$$

We introduce $h_{aj} = f_{aj} \sqrt{\frac{\hat{\varphi}_{aj}}{g_a}}$ and get

$$f(p^*) = \frac{1}{\sqrt{n}} \Sigma h_{aj} Y_{aj} \qquad (25)$$

Note that if $\varphi(\theta)$ is linear and $\mathcal{F}$ consists of all linear contrasts, then the set of $(h_{11}, \ldots, h_{sr_s})$ forms a (R-t)-dimensional space.

B. Let us now consider,

$$\sigma_f^2(p) = \text{var } f(p^*) = \sum_1^s \frac{1}{n_a} \text{var} \sum_1^{r_a} f_{aj} N_{aj}/\sqrt{n_a} \tag{26}$$

We have

$$\text{cov } (N_{aj}/\sqrt{n_a},\ N_{ak}/\sqrt{n_a}) = \begin{cases} p_{aj}(1-p_{aj}) & \text{if } k=j, \\ -p_{aj}p_{ak} & \text{if } k\neq j. \end{cases} \tag{27}$$

We then get for $\hat{\sigma}_f^2 = \hat{\sigma}_f^2(\hat{\varphi})$

$$\hat{\sigma}_f^2 = \frac{1}{n} \sum_{a=1}^{s} \sum_{j,k} (\delta_{jk} - \sqrt{\hat{\varphi}_{aj}}\sqrt{\hat{\varphi}_{ak}}) h_{aj}h_{ak} \tag{28}$$

(where we have made use of the Krönecker δ).

C. Below we shall, in order to facilitate the introduction of matrix notations, replace (a,j) by a single letter i, such that i=1,2,...,R represents (a,j) in lexical ordering. Hence $N_{aj} = N_i$, $p_{aj} = p_i$, $\varphi_{aj}(\theta) = \varphi_i(\theta)$, $f_{aj} = f_i$, $h_{aj} = h_i$. We write also n_i and g_i in place of n_a and g_a. Thus n_i and g_i are constants on sections of lenghts $r_1 r_2,\ldots,r_s$, respectively, We denote the sections by $S_1,\ldots,S_a$ respectively, and have

$$\sum_{i\in S_a} N_i = n_a\ ,\quad \sum_{i\in S_a} p_i = 1$$

We can now write (24)

$$Y_i = \frac{N_i - n_i\hat{\varphi}_i}{\sqrt{n_i\hat{\varphi}_i}}\ ;\ i=1,2,\ldots,R \tag{29}$$

Now, let b denote a matrix of order R×s, the a-th column of which is

$(0,\ldots 0,\sqrt{\hat{\varphi}_{a1}},\ldots,\sqrt{\hat{\varphi}_{ar_a}},0\ldots 0)$. (The column starts with $\sum_1^{a-1} r_i$ zeros.) We see that

$$b'b = I \tag{30}$$

and get from (25)

$$f(p^*) = \frac{1}{\sqrt{n}} \sum_1^R h_i Y_i = \frac{1}{\sqrt{n}} h'Y \tag{31}$$

and from (28)

$$\hat{\sigma}_f^2 = \frac{1}{n} h'(I-bb')h \tag{32}$$

D. From the contrast property of f we have,

$$f_o + \Sigma f_i \varphi_i = 0, \tag{33}$$

hence

$$\sum_1^R f_i \frac{\partial \varphi_i(\theta)}{\partial \theta_j} = 0;\ j=1,2,\ldots,t \tag{34}$$

for any θ. We introduce

$$B_{ij} = \sqrt{\frac{g_i}{\hat{\varphi}_i}} \frac{\partial \varphi_i(\hat{\theta})}{\partial \hat{\theta}_j};\ i=1,2,\ldots,R;\ j=1,2,\ldots,t \tag{35}$$

It is seen that $B = \{B_{ij}\}$ is the matrix $\frac{\partial \varphi_i(\hat{\theta})}{\partial \hat{\varphi}_j}$ multiplied by a diagonal non-singular matrix. Hence, by 1.B, B must have rank t. We can write (34) with $\theta = \hat{\theta}$

$$h'B = 0 \tag{36}$$

From

$$\sum_{i \in S_a} \varphi_i(\theta) = 1;\ a=1,2,\ldots,s, \tag{37}$$

we get by derivation with respect to θ_j; $j=1,2,\ldots,t$, and setting $\theta = \hat{\theta}$,

$$b'B = 0 \tag{38}$$

E. Since B has full rank, the space V_t spanned by the columns of B is a t-dimensional subspace of the R-dimensional vector space V_R. Let H be a R×t matrix such that its columns constitute an orthonormal basis for V_t. Then of course $H'H = I$ and since by (36) and (38) h and all columns of b are perpendicular to V_t we have

$$h'H = 0; \tag{39}$$

$$b'H = 0 \tag{40}$$

From (40) it is seen that the matrix (H, b) of order R×(t+s) has orthogonal columns. We complete it and obtain an orthogonal matrix

$$K = (G, H, b) \tag{41}$$

of order R×R. G is of order R×(R-t-s).

Let us now introduce

$$d = K'h; \tag{42}$$

$$V = K'Y \tag{43}$$

If φ is linear and $\mathcal{F}$ consists of all linear contrast, then $\{d \mid f \in \mathcal{F}\}$ is a (R-t)-dimensional subspace. We have from (31)

$$\sqrt{n}\, f(p^*) = h'Y = d'V \tag{44}$$

(29) reduces to

$$0 = h'H = d'K'H = (d_{R-s-t+1}, \ldots, d_{R-s})$$

Hence,

$$d_{R-s-t+1} = \ldots = d_{R-s} = 0 \tag{45}$$

From equation (32) we get $n\hat{\sigma}_f^2 = h'h - h'bb'h = d'd - d'K'bb'Kd$.

But

$$K'b = \begin{pmatrix} G' \\ H' \\ b' \end{pmatrix} b = \begin{pmatrix} 0 \\ 0 \\ b'b \end{pmatrix} \tag{46}$$

which combined with (30) and (45) gives

$$n\hat{\sigma}_f^2 = \sum_{1}^{R-s-t} d_i^2 \tag{47}$$

For V given by (43) we have

$$V = \begin{pmatrix} G' \\ H' \\ b' \end{pmatrix} Y$$

But by (29), the a-th component of b'Y is

$$\sum_{i \in S_a} (N_i - n_i \hat{\varphi}_i)/\sqrt{n_i} , \tag{48}$$

which from (37) and since n_i is constant, equals 0. Thus

$$V_{R-s+1} = \ldots = V_R = 0 \tag{49}$$

and by (44)

$$\sqrt{n}\, f(p^*) = \sum_{1}^{R-s-t} d_i V_i \tag{50}$$

By (50) and (47) the criterion (11) for stating that $f(p) > 0$ reduces to

$$\sum_{1}^{R-s-t} d_i V_i > \sqrt{z \sum_{i=1}^{R-s-t} d_i^2} \tag{51}$$

for all d. But for given Σd_i^2, an upper bound of the left hand side is, by Schwartz inequality,

$$\sqrt{\sum_{1}^{R-s-t} d_i^2 \sum_{1}^{R-s-t} V_i^2} \tag{52}$$

Thus we make a statement only if

$$\sum_{1}^{R-s-t} V_i^2 > z \tag{53}$$

Now we make use of the fact that $\hat{\theta}$ is maximum likelihood estimate in the null-state, i.e. satisfies (7), which can be written

$$\sum_{i=1}^{R} \frac{N_i}{\hat{\varphi}_i} \frac{\partial \varphi_i(\hat{\theta})}{\partial \varphi_j} = 0; \quad j=1,2,\ldots,t \tag{54}$$

By derivation of (37) with respect to θ_j, setting $\theta = \hat{\theta}$, multiplying by n_a, summing over all a, and subtracting from (54), we get

$$B'Y = 0 \tag{55}$$

Hence $H'Y = 0$ and $V_{R-s-t+1}=\ldots=V_{R-s}= 0$. Thus (53) is the same as

$$Z = \sum_{1}^{R} Y_i^2 > z \tag{56}$$

Thus we have proved the first part of Proposition 4. Since under the assumption of Proposition 3, $(d_1,\ldots,d_{R-s-t},d_{R-s+1},\ldots,d_R)$ varies freely (see remark after (42)) then (52) is the maximum of the left hand side of (51). Then (53) is also a sufficient condition for making a statement and the first part of Proposition 3 is proved.

F. Since in the null-state Z is chi-square distributed with R-s-t degrees of freedom, we have proved the first part of Proposition 2.

G. We shall still consider case (i), but we now assume that we use a priori estimated variances in the multiple comparison rule. The derivation in A-F can then be repeated with the following changes.

h_{aj} is now defined $= f_{aj}\sqrt{\frac{p^*_{aj}}{g_a}}$ and (24) is replaced by

$$Y_{aj} = \frac{N_{aj} - n_a \hat{\varphi}_{aj}}{\sqrt{N_{aj}}} \qquad (24)'$$

with the corresponding change in (29). In the definition of b, $\hat{\varphi}_{aj}$ is replaced by p^*_{aj}. The definition of B_{ij} in (35) is replaced by

$$B_{ij} = \sqrt{\frac{g_i}{p^*_i}}\,\frac{\partial \varphi_i(\hat{\theta})}{\partial \hat{\theta}_j} \qquad (35)'$$

From (8) we get (55) with Y defined by (24)'. Hence we get the last parts of Propositions 2, 3 and 4.

H. Now let $p \neq \varphi$ in case (i). We consider the multiple comparison rule when a priori estimated variances are used. Let

$$X_i = \frac{N_i - n_i p_i}{\sqrt{N_i}} \qquad (57)$$

Then

$$\sqrt{n}\, f(p^*) = \sqrt{n} \sum_1^R f_i \frac{N_i}{n_i} + f_o = \sum_1^R h_i X_i + \sqrt{n}\, f(p) \qquad (58)$$

where $h_i = f_i\sqrt{\frac{p^*_i}{g_i}}$. Hence the probability of a false statement can be written,

$$\Pr\Big(\bigcup_{f(p)\leq 0}\{\sum_{1}^{R} h_i X_i + \sqrt{n}\, f(p) \geq \sqrt{zn}\,\sigma_f^*\}\Big) \qquad (59)$$

where the union is taken over all $f \in \mathcal{F}$ such that $f(p) \leq 0$, for given p. It is seen that (59) is

$$\leq \Pr\Big(\bigcup_{f(p)\leq 0}\{h'X \geq \sqrt{zn}\,\sigma_f^*\}\Big) \leq$$

$$\leq \Pr\Big(\bigcup_{f}\{h'X \geq \sqrt{zn}\,\sigma_f^*\}\Big) = P_n, \qquad (60)$$

where in the last expression the union is taken over all linear contrasts.

We shall show that limsup $P_n \leq \epsilon$ when p is replaced by $p^{(n)}$ and $\lim_{n\to\infty} p^{(n)} = p$.

Below we shall refer to equations (24)',(28)',(29)', (30)',(35)', etc. They are equations (24),(28),(29), (30),(35) etc. above with $\hat{\varphi}$ replaced by p* on the proper places such as described in section G above. We use the same notations B,h,d,b,K,G,H for the modified concepts. We shall also need $\bar{B},\bar{h},\bar{d},\bar{b}$ etc. which are the modified B,h,d etc. with $\hat{\theta}$ replaced by θ and p_i^* by p_i. Thus

$$\bar{B}_{ij} = \sqrt{\frac{g_i}{p_i}}\,\frac{\partial\varphi_i(\theta)}{\partial\theta_j}$$

(see (35)' page 35).

Consider now

$$X_i = \frac{N_i - n_i p_i^{(n)}}{\sqrt{N_i}} = \frac{N_i - n_i p_i^{(n)}}{\sqrt{n_i p_i^{(n)}}}\sqrt{\frac{p_i^{(n)}}{p_i^*}}$$

Then it can be proved by means of a result generalizing Loeve (1955) p.295, that $(X_1,\ldots,X_R)$ converges in distribution to a vector $(\bar{X}_1,\ldots,\bar{X}_R)$ which is multinormal with

expectation 0 and covariance matrix

$$\text{covm}\ (\bar{X}_1,\ldots,\bar{X}_R) = I - \bar{b}\,\bar{b}' \tag{61}$$

We have used the fact that plim $p^* = p$, which follows because $p_i^* = p^*_i - p^{(n)} + p^{(n)}$ and plim $(p_i^*-p_i^{(n)}) = 0$ by Chebyshev's inequality.

Now, it is seen from (8) that

$$\sum_j \frac{p_j^*-\varphi_j^*(\hat{\theta})}{p^*_j}\, g_j\, \frac{\partial\varphi_j(\hat{\theta})}{\partial\hat{\theta}_i} = 0\ ;\ i=1,2,\ldots,t \tag{62}$$

Hence from the assumption of continuity, uniqueness etc. (see 1,C page 28), plim $\hat{\theta} = \theta$, where θ is the unique solution of

$$\sum_j \frac{p_j-\varphi_j(\theta)}{p_j}\, g_j\, \frac{\partial\varphi_j(\theta)}{\partial\theta_i} = 0\ ;\ i=1,2,\ldots,t \tag{63}$$

It follows that

$$\text{plim}\ B = \bar{B} \tag{64}$$

H in 2,E can be constructed from B in such a manner that the elements H_{ij} are continuous functions of B, hence of $\hat{\theta}$. We write $H_{ij}(\hat{\theta})$. Similarly $G_{ij}(\hat{\theta})$ can be constructed such that they are continuous. Then

$$\text{plim}\ K = \bar{K} = (G(\theta),\ H(\theta),\ \bar{b}) \tag{65}$$

Returning to (60), we introduce

$$W = K'X \tag{66}$$

and get

$$b'X = d'KK'W = d'W \tag{67}$$

by (42). We also have

$$W=\begin{pmatrix} G' \\ H' \\ b' \end{pmatrix} X \tag{68}$$

But the a-th component of b'X is

$$\sum_{i\in S_a} \sqrt{p_i^*}\, \frac{N_i - n_i p_i^{(n)}}{\sqrt{N_i}} = 0$$

Hence

$$W_{R-s+1} = \ldots = W_R = 0 \tag{69}$$

We have from (69) and (45)'

$$h'X = \sum_{i=1}^{R-s-t} d_i W_i \tag{70}$$

Consider now σ_f^{*2} in (58). We have from (32)'

$$n\sigma_f^{*2} = h'(I-bb')h = hh' - h'bb'h =$$
$$= d'd - d'K'bb'Kd = \sum_{1}^{R-s-t} d_i^2 \tag{71}$$

since b'K = (0,0,b'b) = (0,0,I).

We get from (58), (70), (71)

$$P_n = \Pr(A_n)$$

where $$A_n = \bigcup_f \left(\sum_{i=1}^{R-s-t} d_i W_i > \sqrt{z \sum_{i=1}^{R-s-t} d_i^2} \right) \tag{72}$$

But

$$\sum_i d_i W_i \leq \sqrt{\Sigma d_i^2 \Sigma W_i^2} \tag{73}$$

Thus A_n implies $\sqrt{\Sigma W_i^2} > \sqrt{z}$

and

$$P_n \leq \Pr\left(\sum_i^{R-s-t} W_i^2 > z \right) \tag{74}$$

From the convergence property of X, (66) and (65) it follows that W converges in distribution to $\bar{K}'\bar{X} = \bar{W}$ (say). Of course $\bar{W}$ is multinormal with covariance matrix

$$\text{covm } (W) = \bar{K}'(I-\bar{b}\bar{b}')\bar{K} \tag{75}$$

This matrix has zero elements except for the first R-s elements of the diagonal which are equal to 1. Hence $\bar{W}_{R-s+1}=\ldots=\bar{W}_R = 0$ with probability 1 and $\bar{W}_1,\ldots,\bar{W}_{R-s}$ are independent normal (0,1).

It follows that the right hand side of (74) converges to ϵ. Hence limsup $P_n \leq \epsilon$. Combining this result with (57) and (58) we obtain that the probability of a false statement has a probability the limsup of which is $\leq \epsilon$.

This proves Proposition 1 in case (i) with a priori estimated variances. The assertion in Proposition 1 with null-state estimated variances is proved in a similar manner.

I. Note that under the assumptions of Proposition 2 we have lim $P_n = \epsilon$. From this result (58) and limsup $P_n \leq \epsilon$ under general assumptions, the assertion in Proposition 6 follows in case (i). In case (ii) the assertion follows from the development below.

J. Let us now consider the case (ii) of <u>non-linear</u> contrasts f(p) and let us use <u>linearized null-state estimated variances.</u> We then state that $f(p) > 0$ if

$$f(p^*) > \sqrt{z}\,\hat{\sigma}_f \tag{76}$$

where $\hat{\sigma}_f = \sigma_f(\hat{\varphi})$,

$$\sigma_f^2(p) = \sum_a \frac{1}{n_a}\left[\sum_j f_{aj}^2(p)p_{aj} - \left(\sum_j f_{aj}(p)p_{aj}\right)^2\right] \tag{77}$$

and $f_{aj}(p) = \frac{\partial}{\partial p_{aj}} f(p)$.

We shall first study asymptotic properties of this rule if $p = \varphi$.

We now set

$$\sqrt{n}\, f(p^*) = \sqrt{n} \sum f_i(p')\left(\frac{N_i}{n_i} - \hat{\varphi}_i\right) =$$

$$= \sqrt{n} \sum f_i(\varphi)\left(\frac{N_i}{n_i} - \hat{\varphi}_i\right) + \sqrt{n}\, A'_f \qquad (78)$$

where p' is "between" p^* and $\hat{\varphi}$ and A'_f goes to 0 uniformly in f as p^* and $\hat{\varphi}$ go to φ. (78) may also be written

$$\sqrt{n}\, f(p^*) = \sum h_i Y_i + \sqrt{n}\, A'_f \qquad (79)$$

where Y_i is defined as before and

$$h_i = f_i(\varphi)\sqrt{\frac{\hat{\varphi}_i}{g_i}} \qquad (80)$$

By (77) we have

$$\sqrt{n}\,\hat{\sigma}_f = \left\{\sum g_i^{-1}\left[f_i^2(\hat{\varphi})\hat{\varphi}_i - \left(\sum f_j(\hat{\varphi})\hat{\varphi}_j\right)^2\right]\right\}^{\frac{1}{2}}$$

$$= \left\{\sum_i g_i^{-1}\left[f_i^2(\varphi)\hat{\varphi}_i - \left(\sum_j f_j(\varphi)\hat{\varphi}_j\right)^2\right]\right\}^{\frac{1}{2}} + A''_f =$$

$$= \sqrt{n}\,\tilde{\sigma}_f + A''_f \qquad (81)$$

where A''_f goes to 0 uniformly in f as $\hat{\varphi}$ goes to φ. We see that (32) above still holds with $\hat{\sigma}_f$ replaced by $\tilde{\sigma}_f$ and h_i defined by (80). We can now go through the same development as in 2.D - 2.H above. Equation (34) with $f_i = f_i(\hat{\varphi})$ is derived from the contrast property

$$f(\varphi(\theta)) = 0 \tag{82}$$

Thus (34) is rigorously true. We finally obtain in place of (50)

$$\sum_{i=1}^{R-s-t} d_i V_i + A_f > \sqrt{z \sum_{1}^{R-s-t} d_i^2} \tag{83}$$

where $A_f = A'_f + A''_f / \sqrt{n}$ goes to 0 uniformly in f as p* and $\hat{\varphi}$ go to φ.

Now, let S_1 denote the statement that (83) takes place for some f, and S_2 the statement that

$$\sum_{i=1}^{R-s-t} d_i V_i > \sqrt{z \sum_{i=1}^{R-s-t} d_i^2} \tag{84}$$

for some f. Then S_1 asymptotically implies S_2 in the sense that the probability of S_1-S_2 goes to zero. Because S_1-S_2 means that (83) is true for some f and the reverse inequality of (84) is true for all f. Hence S_1-S_2 implies that

$$A_f > 0 \tag{85}$$

is true for some f. Thus

$$\Pr(S_1\text{-}S_2) \leq \Pr(\bigcup_f (A_f > 0)) \tag{86}$$

(Note that by separability the union in (86) can be made countable.) The right hand side of (86) is the limit of

$$\Pr(\bigcup_f (A_f > \eta)) \tag{87}$$

as $\eta > 0$ goes to 0. (87) is equal to

$$1 - \Pr(\cap_f (A_f < \eta)) \tag{88}$$

However, we can find a δ such that $|p^*-\varphi| < \delta$ and $|\hat{\varphi}-\varphi| < \delta$ implies $A_f < \eta$ for all f, by the equicontinuity property. Hence (88) is less than

$$\Pr(|p^*-\varphi| \geq \delta) + \Pr(|\hat{\varphi}-\varphi| \geq \delta) \tag{89}$$

Now, first choose η so small that the difference between the right hand side of (86) and (87) is less than $\frac{\rho}{2}$. Afterwards we can find the corresponding δ. Then for fixed δ choose n so large that (89) is less than $\frac{\rho}{2}$. Then $\Pr(S_1-S_2) < \rho$ and $\Pr(S_1-S_2) \to 0$.
However, since

$$\Sigma\, d_i V_i \leq \sqrt{\Sigma\, d_i^2\ \Sigma\, V_i^2}$$

we have that (84) implies

$$\Sigma\, V_i^2 > z \tag{90}$$

Combining (90) and $\lim \Pr(S_1-S_2) = 0$ we obtain Proposition 5. We also have from $S_1 \subset (S_1-S_2) \cup S_2$ that

$$\Pr(S_1) \leq \Pr(S_2) + \Pr(S_1-S_2)$$

and hence

$$\text{limsup } \Pr(S_1) \leq \lim \Pr(\Sigma\, V_i^2 > z) \tag{91}$$

Hence we have proved that in the null-state
lim sup Pr(false statement) $\leq \epsilon$. The statement about limsup Pr(false statement) $\leq \epsilon$ for any $p^{(n)} \to p$ is now proved in essentially the same manner as when f is linear.

Instead of (58), we start out from

$$\sqrt{n}\, f(p^*) =$$

$$= \sqrt{n} \sum f_i(p^{(n)})\left(\frac{N_i}{n_i} - p_i^{(n)}\right) + \sqrt{n}\, f(p) + \sqrt{n}\, A_f' \qquad (92)$$

where $A_f' \to 0$ uniformly in f as p^* and p^n go to p, and use the same kind of arguments as in H and this section. Thus we have proved the assertions in Propositions 1 and 5 in case (ii).

Appendix III

PERFORMANCE OPTIMAL TESTING OF A SINGLE PARAMETER

Suppose that an observed random vector has probability distribution depending on parameters (ρ,τ), where ρ is scalar. We want to decide whether $\rho <$ or > 0, i.e. we have a choice between three decisions "$\rho < 0$","$\rho > 0$", "no inference". Let the acceptance regions for the three decisions be B_1, B_2, B_3, respectively, where $B_1 \cup B_2 \cup B_3$ = sample space.

The performance of the method (B_1, B_2, B_3) has two branches

$$\beta_1(\rho,\tau) = \Pr(B_1) \ , \ \beta_2(\rho,\tau) = \Pr(B_2) \qquad (1)$$

which are respectively the probabilities of stating that $\rho < 0$ and $\rho > 0$.

(B_1, B_2, B_3) is <u>performance optimal</u> with level ϵ if

$$\text{(i)} \quad \beta_1(\rho,\tau) \leq \epsilon \quad \text{for} \quad \rho \geq 0$$
$$\beta_2(\rho,\tau) \leq \epsilon \quad \text{for} \quad \rho \leq 0 \qquad (2)$$

$$\text{(ii)} \quad \beta_1(\rho,\tau) \geq \epsilon \quad \text{for} \quad \rho < 0$$
$$\beta_2(\rho,\tau) \geq \epsilon \quad \text{for} \quad \rho > 0 \qquad (3)$$

(iii) Among all methods satisfying (i) and (ii) it maximizes $\beta_1(\rho,\tau)$ for all (ρ,τ) with $\rho < 0$ and maximizes $\beta_2(\rho,\tau)$ for all (ρ,τ) with $\rho > 0$.

Suppose now that there exist a uniformly most powerful unbiased test for testing $\rho \geq 0$ and also for testing $\rho \leq 0$. Let the rejection regions be B_1 and B_2 respectively and suppose that $B_1 \cap B_2 = \emptyset$. Then (B_1, B_2, B_3), where $B_3 = \overline{(B_1 \cup B_2)}$, has optimal performance.

From this it follows that the Student test described in section 2 is performance optimal.

In the double dichotomy case (see section 2), let p_1 and p_2 be the probabilities of B under A and not-A respectively. Then there are positive or negative dependences according as

$$\rho = \log \frac{p_1}{1-p_1} \Big/ \log \frac{p_2}{1-p_2} \tag{4}$$

is positive or negative.

The uniformly most powerful test for testing $\rho \leq 0$ against $\rho > 0$ is the following. Reject if $H(X) > 1-\epsilon$. Reject with probability γ_2 if $X = c_2$. Here

$$h(x) = \binom{M}{x}\binom{n-M}{L-x}/\binom{n}{L} \ , \ H(x) = \sum_{y=0}^{x} h(y)$$

and c_2, γ_2 are determined such that $0 \leq \gamma_2 < 1$

$$1 - H(c_2) + \gamma_2 h(c_2) = \epsilon$$

Similarly we reject the null-hypothesis $\rho \geq 0$ if $H(X) < \epsilon$. We reject with probability γ_1 if $X = c_1$, where $0 \leq \gamma_1 < 1$ and

$$H(c_1-1) + \gamma_1 h(c_1) = \epsilon$$

Thus the optimum three-decision procedure consists in stating negative or positive dependence according as $H(x) < \epsilon$ or $> 1-\epsilon$, and making randomized decisions if $X = c_1$ or c_2. (Of course the randomization is never carried out in practice.)