Biom. J. 25 · 1983 · 5, 459–467

Georg-August-Universität Göttingen, Abteilung für Forstliche Biometrie, BRD

Stein's Two-Stage Test in Linear Models

J. SABOROWSKI

Abstract

After a brief presentation of STEIN's two-sample test there is shown an improved procedure for general linear hypotheses analogous to the improvement of the two-sample-t-test described by STEIN. The example of an analysis of variance in an one-way layout demonstrates the execution of the procedure as well as the problems that occur with the determination of the test parameters n_0 and z. For these problems there is finally suggested a practicable way of solution.

Key words: Two stage sampling, analysis of variance.

1. Introduction

One of the problems in planning experiments for testing a linear hypothesis consists in determining a suitable sample size. More precisely, "suitable" means that although the sample size is kept as small as possible we can obtain a sufficient test power that can be required at a relevant point in the class of alternatives. The power function of the F-test usually applied, however, is well known to depend on the residual variance σ^2, which in general is not known before the experiment.

Due to the fact that as σ^2 increases, the power function converges towards the error of the first kind, it is theoretically not possible to safeguard against unexpectedly high variances, even if very large samples are used. For instance, even if it is possible, on the basis of previous experiments, to state an interval $[\sigma^2_{\min}, \sigma^2_{\max}]$ for σ^2, one would have to recur to $\sigma^2_{\max}$ for determining the sample size.

By contrast, STEIN's two-stage test (STEIN, 1945) is independent of σ^2, i.e. for arbitrary variances it adheres to a required power. The test is carried out by estimating the variance σ^2 on the basis of n_0 initial samples and then calculating the number of samples that are going to be needed in addition. Determining the test parameters n_0 and z may appear to be a tedious job, but it can be done easily and fast with the help of electronic data processing. Contrary to the two-stage t-test, an application of this method to oneway or higher-way layouts has not, to our knowledge, been carried out up to now.

In later paragraphs, we give a comprehensive representation of STEIN's two-stage test for linear models, of Moshman's method to determine an optimal initial sample size when σ^2 is known, and finally we discuss an example from forestry research, using that as a representative of various other problems where choosing the appropriate sample size is similarly important. The initial sample size is here determined by using a loss function based on MOSHMAN (1958), thus not only taking into account the expectation of the total sample size N' (as suggested by SEELBINDER 1953), but also a quantile of the distribution of N'.

2. General description of procedure

The term "sample x_i" will henceforth be understood to mean the vector

$$x_i = (x_{i1}, \ldots, x_{im})'$$

consisting of m mutually independent observations, where x_{ij} are subject to a normal distribution, with the same variance σ^2 which is independent of i and j. The x_i are also regarded as independent, and we assume a mean vector a to be an element of a k-dimensional space L_k with $k \leqq m$. We then have the usual hypotheses (LEHMANN, 1959; WITTING, NÖLLE, 1970):

$$H: a \in L_h \qquad K: a \in L_k - L_h \qquad h < k\,.$$

From n_0 initial samples $x_1, \ldots, x_{n_0}$ we estimate σ^2 using

$$s^2 = \frac{1}{n_0 m - k} \sum_{i=1}^{n_0} \left| x_i - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_k(x_j) \right|^2$$

where $\hat{a}_k(x_j)$ is the projection of sample x_j onto L_k; then we calculate the total sample size N as a function of s^2 according to

$$N = \max \left(n_0 + 1, \left[\frac{s^2}{z} \right] + 1 \right)$$

($[x]$ is the largest integer smaller than x).

In the case of an error of the first kind, α, the hypothesis will be rejected if

$$\frac{n_0 m - k}{k - h} T = \frac{1}{k-h} \frac{|\hat{a}_k(\sum^N c_i x_i) - \hat{a}_h(\sum^N c_i x_i)|^2}{z} > F_\alpha\ (k-h,\ n_0 m - k)$$

$F_\alpha(\nu_1, \nu_2)$ means the α-fractile of an F-distribution with the degrees of freedom (ν_1, ν_2); the c_i are weights that depend on s and have the following properties:

$$\sum_{i=1}^{N} c_i = 1 \qquad \sum_{i=1}^{N} c_i^2 = z/s^2 \quad c_1 = \ldots = c_{N-1}\,.$$

The c_i exist because of $1/N \leqq z/s^2$. Through an improved procedure it will later on be possible to dispense with these weights.

The importance of the positive real number z lies in the fact that by specifying z, the non-centrality parameter δ^2 of the distribution of T, and hence the power, is determined, for, according to Stein, T possesses for $a \in L_k - L_h$ a so-called noncentral F'-distribution with $(k-h,\ n_0m-k)$ degrees of freedom; this distribution is specified through the density

$$f(x)=\frac{\Gamma\left(\frac{n+m}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m-1}{2}\right)}\int\limits_{-\sqrt{x}}^{\sqrt{x}}(x-y^2)^{\frac{m-3}{2}}\,(1+x+2y\delta+\delta^2)^{\frac{-(n+m)}{2}}\,dy$$

and

$$\delta^2=\frac{|\hat{a}_k(a)-\hat{a}_h(a)|^2}{z\,(n_0m-k)}\qquad n=n_0m-k\qquad m=k-h$$

If $\delta^2=0$, the distribution of $(n_0m-k)/(k-h)\cdot T$ is reduced to the usual central F-distribution with the same degrees of freedom. Given n_0, it is possible to calculate for any point a from the class of alternatives K, the power $\beta(a)$; according to

$$\beta(\vec{a})=1-\int\limits_0^c f(x)\,\mathrm{d}x$$

with

$$c=\frac{k-h}{n_0m-k}\,F_\alpha\,(k-h,\ n_0m-k)$$

so that $z=z(n_0,\ a)$ can, after relatively few trials, be determined in such a way that the test reaches the desired power at point a. Although there is no exact solution to the integral in $f(x)$, it can nevertheless be approximated numerically with sufficient precision by using a computer.

3. Improved procedure

If, in T, N is replaced by N', the weights c_i are replaced by the factor $1/\sqrt{N'}$, where

$$N'=\max\left(n_0,\ \left[\frac{s^2}{z}\right]+1\right)$$

and if furthermore z is replaced by s^2, the resulting statistic will be

$$T'=\frac{1}{n_0m-k}\,\frac{1}{N'}\,\frac{|\hat{a}_k\,(\sum^{N'} x_i)-\hat{a}_h\,(\sum^{N'} x_i)|^2}{s^2}$$

Presupposing hypothesis H, the distribution of $(n_0m-k)/(k-h)\cdot T'$ is reduced to $F(k-h,\ n_0m-k)$, with the result that in

$$\left\{\frac{n_0m-k}{k-h}\,T'>F_\alpha\,(k-h,\ n_0m-k)\right\}$$

a critical region with respect to level α arises (see appendix II). The power function $\beta'(a, \sigma^2)$ is now no longer independent of σ^2; however, the following holds (see appendix I):

$$\beta'(a, \sigma^2) \geqq \beta(a)$$

Thus, using the new improved critical region, it is in any case true to say that the power will not be lower than the required one.

4. Initial sample size n_0

As STEIN (1945) has pointed out, the distribution of N' depends, not only on z/σ^2, but also on n_0, for

$$\text{EN}' = \text{E}(N'|n_0) = n_0 P(\chi^2(n) < q) + \frac{\sigma^2}{z} P(\chi^2(n+2) > q) + \Theta P(\chi^2(n) > q)$$

$$0 \leqq \Theta \leqq 1 \qquad n = n_0 m - k \qquad q = n_0 n \frac{z}{\sigma^2}.$$

Furthermore the p-quantile of N' is the smallest number $N_p(n_0)$ for which

$$P(N' \leqq N_p) = P\left(\chi^2(n) \leqq N_p n \frac{z}{\sigma^2}\right) \geqq p$$

holds. In this formula, $\chi^2(n)$ is a random function with a χ^2-distribution with n degrees of freedom. Instead of choosing n_0 in such a way that EN′ is minimized, MOSHMAN (1958) suggested that additionally the behaviour of N_p as subject to n_0 should be taken into consideration, in order to keep down the probability of very large samples. This is relevant because it has been found that in regions where EN′ increases only slightly, N_p continues to decrease considerably (see table 1).

Table 1
Parameters of the distribution of N′
($z/\sigma^2 = .0856$ $k = m = 6$)

n_0	EN'	$N_{.95}$	Ψ
2	12.183	24.516	−0.203
3	12.1823	20.469	0.000
4	12.1824	18.737	0.087
5	12.183	17.725	0.137
6	12.184	17.046	0.171
7	12.188	16.549	0.194
8	12.20	16.167	0.208
9	12.24	15.861	0.209
10	12.33	15.610	0.185
11	12.52	15.397	0.117
12	12.88	15.216	−0.019
13	13.43	15.058	−0.236
14	14.17	14.919	−0.532
15	15.06	15.0	−0.894
16	16.01	16.0	−1.333

This phenomenon can be taken into account by choosing an initial sample size $n_0 = n_0^{\text{opt}}$ in such a way that

$$\Psi(n_0) = (1-p)\,(N_p(n_0^*) - N_p(n_0)) - (1 - P(n_0^*))\,(\mathrm{E}(N' \mid n_0) - \mathrm{E}(N' \mid n_0^*))$$

is maximized. Let

$$P(n_0^*) = P\,(N' \leqq \mathrm{E}(N' \mid n_0^*)) - P\left(\chi^2(n) \leqq \mathrm{E}(N' \mid n_0^*)\, n\, \frac{z}{\sigma^2}\right)$$

where $n_0 = n_0^*$ has been selected so as to minimize $\mathrm{E}(N' \mid n_0)$. Ψ is evidently going to be large if a $(1-p)$-weighted strong decrease of the p-quantile of N' goes alongside with a relatively slight increase in the expectation of N'—which increase is to be weighted the less strong, the smaller a probability we have of N' exceeding $\mathrm{E}\,(N' \mid n_0^*)$. Ψ is thus a measure for the quality of the distribution of N' with regard to a minimal total sample size.

One drawback of this objective procedure lies in the fact that Ψ is in itself a function of z/σ^2. Nevertheless it is extremely helpful in the planning of experiments as the following section will show.

5. An application example from forestry

In order to determine the effects of spacing douglas firs on the average ovendry density of the wood, an investigation was made of six differently spaced homogeneous stands (HAPLA, 1980). Because establishing this variable by taking bore chip samples requires a great deal of work and costs are caused as a result of the depreciation of the stems that undergo the boring, the number of chips to be taken is of major importance. If μ denotes the overall mean and ν_i the deviation from μ in the i-th spacing, there results an one-way analysis of variance for the hypotheses

$$H: \nu_1 = \ldots = \nu_6 = 0 \qquad K: \sum \nu_i^2 > 0 \qquad k = 6 \qquad h = 1\,.$$

Furthermore, if $m = k$, one sample will consist of six measurements, one for each spacing. As a difference in the average ovendry density of less than $\approx .04$ gr/cm³ is of no economic interest, the idea is to obtain a high power while keeping the range of means $\mu + \nu_i$ at .04 gr/cm³. To establish a corresponding distance $\sum \nu_i^2$ from the hypothesis H, we make the additional supposition that

$$\nu_1 = -0.020 \qquad \nu_2 = -0.012 \qquad \nu_3 = -0.004$$
$$\nu_4 = 0.004 \qquad \nu_5 = 0.012 \qquad \nu_6 = 0.020$$

and require a test power of approx. 95 % at the point

$$a = (\mu + \nu_1, \mu + \nu_2, \ldots, \mu + \nu_6)' \in L_k - L_h$$

i.e.

$$|\hat{a}_k(a) - \hat{a}_h(a)|^2 = \sum \nu_i^2 = 0.00112\,.$$

The standard deviation σ among the six subgroups, if we look at experiences from other experiments, can be vaguely expected to lie between .025 and .045. We therefore begin by establishing, for different σ in the above-mentioned interval with the help of an interactive programme using iteration, a common value for z and the optimal initial sample sizes (section 4)–these sizes being a function of z/σ^2; also, we establish the resulting test powers. For $z=.535 \cdot 10^{-4}$, we obtain table 2.

Table 2

σ	0.025	0.031	0.036	0.041	0.045
n_0^{opt}	9	14	19	26	31
$\beta(a)$	94.6	95.3	95.5	95.7	95.8

This means that in every case the desired test power is reached with sufficient accuracy, although the initial sample sizes vary between 9 and 31. The weak influence of n_0 on the test power is accounted for by the fact that as n_0 decreases, the decrease of the 2nd degree of freedom n_0m-k of the F'-distribution counteracts the increase of δ^2. As we lack furtherreaching information about σ^2, we pick as an a-priori distribution for σ^2 the rectangular distribution in the interval $[6.25 \cdot 10^{-4}, 2.025 \cdot 10^{-3}]$ and measure for each σ^2 and n_0 the decrease in quality of the distribution of N', as compared to the use of $n_0=n_0^{opt}$ (which according to section 4 is optimal), by establishing the difference

$$0 \leqq \Psi(n_0^{opt}) - \Psi(n_0) =$$
$$(1-p)\,(N_p(n_0)-N_p(n_0^{opt})) - (1-P(n_0^*))\,(\mathrm{E}(N' \mid n_0^{opt}) - \mathrm{E}\,(N' \mid n_0))$$

The definite initial sample size is then obtained by minimizing the expectation

$$D = \mathrm{E}\,(\Psi(n_0^{opt}) - \Psi(n_0))$$

according to table 3, turning out to be $n_0=13$. As was to be expected, this is a figure ranging in the lower part of the interval [9, 31], because in particular $\mathrm{E}\,(N' \mid n_0)$, which is more heavily weighted in Ψ, increases much more rapidly if n_0 is large relative to n_0^{opt}, than it does if n_0 is small (see table 1).

Table 3

n_0	9	10	11	12	13	14	15
D	.1820	.1571	.1387	.1284	.1282	.1409	.1684

The indicated values of D have here been obtained by approximation, using an equidistant partition of the interval of variances into 50 subintervals.

Because of $\hat{a}_k(x_i)=x_i$, the samples $x_1, \ldots, x_{13}$ immediately yield

$$s^2 = \frac{1}{n_0m-k} \sum_{i=1}^{13} \sum_{j=1}^{6} \left(x_{ij} - \frac{1}{13} \sum_{l=1}^{13} x_{lj}\right)^2 = 0.793 \cdot 10^{-3} \quad N' = 15\,,$$

which means that only two more samples have to be taken.

From

$$|\hat{a}_k(\sum^{N'} x_i) - \hat{a}_h(\sum^{N'} x_i)|^2 = N'^2 \sum_j (x_{.j} - x_{..})^2$$

it follows that

$$\frac{n_0 m - k}{k-h} T' = \frac{1}{k-h} N' \frac{\sum (x_{.j} - x_{..})^2}{s^2} = 10.308$$

which amounts to the rejection of hypothesis H, as

$$F_{0.05}(5,72) = 2.342 .$$

For the sake of comparison we would like to draw the reader's attention to the fact that if using the one-stage F-test, based on an upper limit of $\sigma = .045$, a sample size of 35 would have been required in order also to ensure a test power of 95 %. It is only for $s = .043$ that, because of the somewhat lower efficiency of STEIN's procedure, N' becomes 35 and $s = .045$ yields $N' = 38$. But because of

$$\text{Var } s^2 = \frac{2}{n_0 m - k} \sigma^4 = 1.14 \cdot 10^{-7}$$

one can hardly expect any considerably larger values for s, even when $\sigma = .045$.

Test statistics for higher-way layouts can also easily be derived from the general representation in sections 2 and 3. In the case of regression analysis, however, the assumptions of section 2 imply a repeated application of m fixed sets of predictors.

Zusammenfassung

Nach einer kurzen Darstellung von STEIN's Zweistufentest wird analog zu der von STEIN angegebenen Verbesserung des Zweistufen-t-Tests auch ein verbessertes Verfahren für allgemeine lineare Hypothesen angegeben. Das Beispiel einer einfaktoriellen Varianzanalyse demonstriert die Durchführung des Verfahrens sowie die Probleme, die bei der Bestimmung der Testparameter n_0 und z auftreten. Hierfür wird schließlich ein praktikabler Lösungsweg vorgeschlagen.

References

BÖNNER, N., 1976: Das Steinsche Zweistufenverfahren. Unveröffentlichtes Manuskript, Freiburg.

DANTZIG, G. B.: On the non-existence of tests of 'Student's' hypothesis having power functions independent of the variance.

HAPLA, F., 1980: Untersuchung der Auswirkung verschiedener Pflanzverbandsweiten auf die Holzeigenschaften der Douglasie. Dissertation, Göttingen.

LEHMANN, E. L., 1959: Testing statistical hypotheses. New York, 203–204, 265–286.

MOSHMANN, J., 1958: A method for selecting the size of the initial sample in Stein's two sample procedure. Ann. Math. Statist. **29**, 1271–1275.

SEELBINDER, B. M., 1953: On Stein's two-stage sampling scheme. Ann. Math. Statist. **24**, 640–649.

STEIN, C., 1945: A two-sample test for a linear hypothesis whose power is independent of the variance. Ann. Math. Statist. **16**, 243–258.

WETHERILL, G. B., 1966: Sequential methods in statistics. London, New York, 198–199.

WITTING, H., NÖLLE, G., 1970: Angewandte mathematische Statistik. Stuttgart, 11–22.

Manuscript received: 27. 8. 1981

Author's address:

Dipl.-Math. Joachim SABOROWSKI

Abteilung für Forstliche Biometrie

Büsgenweg 5

D - 3400 Göttingen, BRD

Appendix I

We reduce the problem by an orthogonal transformation C to the canonical form (LEHMANN (1959)):

$$Y_i = C \cdot X_i \qquad b = C \cdot a \qquad b_{k+1}, \ldots, b_m = 0$$

that is

$$Y_i \sim N(b, \sigma^2 I_m) \quad \text{und} \quad |\hat{a}_k(a) - \hat{a}_h(a)|^2 = \sum_{j=h+1}^{k} b_j^2$$

(Y_i, b, a, X_i are vectors).
Now the general linear hypothesis can be expressed by

$$H: b_{h+1} = \ldots = b_k = 0$$

and STEIN (1945) shows that

$$T = \sum_{j=h+1}^{k} \left[\frac{1}{\sqrt{zn}} \left(\sum_{i=1}^{N'(S)} c_i Y_{ij} - b_j \right) + \frac{b_j}{\sqrt{zn}} \right]^2$$

$$\frac{n}{\sigma^2} S^2 \sim \chi^2(n) \qquad n = n_0 m - k$$

$$\frac{1}{\sqrt{N'}} \sum_{i=1}^{N'} (Y_{ij} - b_j) \sim N(0, \sigma^2)$$

and

$$T' = \frac{1}{nS^2} \frac{1}{N'(S)} \sum_{j=h+1}^{k} \left(\sum_{i=1}^{N'(S)} Y_{ij} \right)^2 = \sum_{j=h+1}^{k} \left[\frac{\frac{1}{\sqrt{N'}} \sum_{i=1}^{N'} (Y_{ij} - b_j)}{\sqrt{n}\, S} + \frac{\sqrt{N'}\, b_j}{\sqrt{n}\, S} \right]^2$$

As $S^2, \sum_{i=1}^{n_0} Y_{ij}, Y_{n_0+1,j}, \ldots$ are independent random functions $\sum_i (Y_{ij} - b_j)$ and S have the same property. From STEIN's definition of the F'-distribution follows

$$\sum_{j=h+1}^{k} \left[\frac{\frac{1}{\sqrt{N'}} \sum_i (Y_{ij} - b_j)}{\sqrt{n}\, S} + \frac{b_j}{\sqrt{nz}} \right]^2 \sim F'(k-h, n) \left(\frac{\sum b_j^2}{nz} \right)$$

and $\frac{N'}{s^2} \geqq \frac{1}{z}$ implies $\frac{N' b_j^2}{n s^2} \geqq \frac{b_j^2}{nz}$ and

$$\beta'(a, \sigma^2) = \text{Prob} \left(T' > \frac{k-h}{n} \mathrm{F}_\alpha(k-h, n) \right)$$
$$\geqq \text{Prob} \left(F'(k-h, n) \left(\frac{\sum b_j^2}{nz} \right) > \frac{k-h}{n} \mathrm{F}_\alpha(k-h, n) \right) = \beta(a).$$

Appendix II

Assumed $V_{h+1}, \dots V_k$, S are independent random functions and $V_i \sim N(0, \sigma^2)$, then

$$\sum_{j=h+1}^{k}\left(\frac{1}{\sqrt{n}}\frac{V_j}{S}\right)^2 \sim F(k-h, n)$$

and hypothesis H implies

$$P\left(\frac{\frac{1}{\sqrt{N'}}\sum_i^{N'(s)} Y_{ij}}{S}\,\middle|\, S=s\right) = P\left(\frac{\frac{1}{\sqrt{N'}}\sum_i^{N'(s)} Y_{ij}}{s}\right) = N\left(0, \frac{\sigma^2}{s^2}\right)$$

$$= P\left(\frac{V_j}{s}\right) = P\left(\frac{V_j}{S}\,\middle|\, S=s\right)$$

due to the independence of $\sum Y_{ij}$ and S.
($P(Z \mid V=v)$ denotes the conditional distribution of Z given $V=v$.)
This completes the proof of

$$T' \sim F(k-h, n).$$