MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

MONOTONICITY PROPERTIES OF STATISTICAL PROCEDURES FOR AN INCREASING NUMBER OF OBSERVATIONS OR A DECREASING NUMBER OF PARAMETERS, ESPECIALLY ANOVA PROCEDURES

R. PINCUS

Academy of Sciences of G.D.R., Berlin, G.D.R.

The effect of different kinds of pooling additional observations in the analysis of variance is studied. A slight extension of a known result for the monotonicity of the power of the F-test in the degrees of freedom is shown.

1. Formulation of the problem

Let $(\mathscr{X}, \mathfrak{A}, P_\theta)$, $\theta \in \Theta$, be a family of probability spaces and $y_1, \ldots, y_n$ a sample of a P_θ-distributed random variable, or, more generally, let $y_1, \ldots, y_n$ have a common distribution $P_{\theta,n}$. For a given statistical decision problem, like testing, estimation, selection, etc., defined by the triplet (Θ, E, L), the parameter space, the decision space and the loss function $L(\theta, e)$ — disregarding sigma-algebras here — we usually search for decision functions δ mapping the sample space into E,

$$\delta\colon (y_1, \ldots, y_n) \to \delta(y_1, \ldots, y_n), \tag{1}$$

or their randomized versions with 'optimal' or at least satisfactory properties.

Having a few observations only, we cannot hope in general to have good procedures from the viewpoints of power, variance and error probabilities, respectively, although they may perhaps be optimal in a certain sense. So it is natural to try to improve the characteristics of a statistical procedure by involving some more observations and thus increasing the amount of information about the unknown parameter. Although decision making in most cases works without explicit costs of observations, one has in mind that they involve costs of course, and so intuitively one demands that

$$\delta_{n+1}(y_1, \ldots, y_{n+1}) \precsim \delta_n(y_1, \ldots, y_n), \tag{2}$$

i.e. that a decision function based on $n+1$ observations should be better than one based on n observations only. As an adequate definition of a 'better'-relation we take

(3) $$E_0 L(\theta, \delta_{n+1}(y_1, \ldots, y_{n+1})) \leqslant E_0 L(\theta, \delta_n(y_1, \ldots, y_n)) \quad \text{for all } \theta,$$

where strict inequality holds for some θ.

In this generality the question of improving any decision function δ_n is not difficult to answer, since $\mathscr{D}_{n+1}$, the space of functions from the $y_1, \ldots, y_{n+1}$ — sample space, contains $\mathscr{D}_n$, and, on the other hand, $(y_1, \ldots, y_n)$ is in general not a sufficient statistic of $(y_1, \ldots, y_{n+1})$. So under mild conditions we can construct a δ_{n+1} — possibly depending on δ_n — which is better than δ_n.

For practical purposes, however, one is more interested in comparing the risk of a given standard procedure for n and $n+1$. So, for example in estimating a location parameter we will compare $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ with $\bar{y}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} y_i$, or a median of $y_1, \ldots, y_n$ with one of $y_1, \ldots, y_{n+1}$, etc. If the $y_1, \ldots, y_{n+1}$ are not independently distributed, it is easy to see that $\bar{y}_{n+1}$ is not necessarily better than $\bar{y}_n$. This perhaps demonstrates that the monotonicity property is not a priori trivially fulfilled. Even in the i.i.d. case there exists no result according to which, say, a Maximum-Likelihood-Estimate based on $n+1$ observations has a smaller Mean-Square-Error than one based on $y_1, \ldots, y_n$. Only for the corresponding bounds of the Rao–Cramer type such monotonicity is shown. Obviously the non-monotonicity of a statistical procedure could have unpleasant consequences for experimental designs.

2. Additional information by parameter restrictions

Closely related to situations where we can use additional observations are those situations where the components of the unknown parameter vector θ are known to fulfil certain functional relations, or, after regular transformation of the parameter space, where some components of θ are known to be zero. Roughly speaking in a normal linear model

(4) $$y = X\theta + \varepsilon,$$

where y is a random $n \times 1$-vector, X is $n \times k$ and known, and θ is a fixed but unknown $k \times 1$-vector, the information $\theta_1 = 0$ leading to a new model

$$y = X^*\theta^* + \varepsilon,$$

where in X^* and θ^* the first columns or components of X and θ, respectively, are dropped, has partly the same effect as an additional observation

$$y_{n+1} = x'_{n+1}\theta + \varepsilon_{n+1}.$$

They both lead to a χ^2_{n+1}-distributed variable instead of a χ^2_n-variable for the estimation of σ^2.

3. The F-test

For illustration we will use as an example an unbalanced one-way classification model with observations

(5) $$\begin{array}{l} y_{11}, \ldots, y_{1n_1}, \\ y_{21}, \ldots, y_{2n_2}, \\ \vdots \qquad \vdots \\ y_{k1}, \ldots, y_{kn_k}, \end{array}$$

where y_{ij} is drawn from an $N(\mu_i, \sigma^2)$-distributed population with μ_i and σ^2 unknown, $i = 1, \ldots, k$.

In order to prove the hypothesis of homogeneity, namely that the expectations of the first r populations are equal, i.e.

$$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_r$$

against the alternative

$$H_1\colon \text{at least one inequality holds},$$

the F-statistic

(6) $$\frac{\frac{1}{r-1}\sum_{i=1}^{r} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}$$

with

$$N = n_1 + \ldots + n_k - k, \qquad \bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},$$

$$\bar{y}_{..} = \frac{1}{n_1 + \ldots + n_r}\sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij},$$

is commonly used. Its distribution is that of an $F_{r-1,N}$-distributed variable with noncentrality

(7) $$\psi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{r} n_i(\mu_i - \bar{\mu})^2,$$

where $\bar{\mu} = \frac{1}{r}\sum_{i=1}^{r}\mu_i$.

If we now have an additional observation in one of the k groups, the corresponding F-statistic will be $F_{r-1,N+1}$-distributed. If this group is one of the last $k-r$, then the noncentrality remains unchanged; otherwise under the alternative it is increased.

If additional information in the form of an equality of the means of two groups is present, i.e. $\mu_{i_0} = \mu_{i_1}$, then after pooling the corresponding observations the F-statistic has a $F_{r-2,N+1}$ or $F_{r-1,N+1}$-distribution, respectively, depending on whether μ_{i_0} and μ_{i_1} are involved in the hypothesis or not.

In all cases the degrees of freedom of the numerator have been reduced or those of the denominator increased and the noncentrality of the new F-statistic under the alternative is at least not smaller than that of the original $F_{r-1,N}$. 'Under the alternative' means here that part of the alternative which underlies the given restrictions of course.

Because of the monotone likelihood ratio the power of an F-test is increasing in ψ^2 (see Lehmann [3], p. 312). Therefore the question if we get better properties of an analysis of variance procedure by exploiting the information given by additional observations or restrictions is answered if one succeeds in proving that the power of an F-test, for noncentrality ψ^2 and significance level α fixed, is an increasing function of the degrees of freedom of the denominator and a decreasing function of the degrees of freedom of the numerator. If we denote by $\beta_{m,n}(\psi^2)$ the power function of an F-test for the hypothesis

$$H_0\colon \psi^2 = 0 \quad \text{against} \quad H_1\colon \psi^2 > 0 \tag{8}$$

with significance level α, where the corresponding test statistic has m and n degrees of freedom, it suffices to show for $\psi^2 > 0$

$$\beta_{m+1,n}(\psi^2) < \beta_{m,n}(\psi^2) < \beta_{m,n+1}(\psi^2). \tag{9}$$

This relation has only recently been proven by Das Gupta and Perlman [1].

In the following we extend this result slightly to the case where instead of (8) we test

$$H_0\colon \psi^2 \leqslant \psi_0^2 \quad \text{against} \quad H_1\colon \psi^2 > \psi_0^2, \tag{10}$$

which corresponds to the perhaps somewhat more realistic hypothesis that the group means are 'approximately' equal, in the sense that $\frac{1}{\sigma^2}\sum n_i(\mu_i-\bar{\mu})^2 \leqslant \psi_0^2$, ψ_0^2 being a suitably chosen constant. We need therefore the following lemma, the proof of which is omitted because of its obvious correlation with the Neyman–Pearson lemma.

LEMMA. *Let P_{θ_0} and P_{θ_1} be two probability measures on a measurable space $(\mathscr{X}, \mathfrak{A})$ with densities $p_{\theta_0}(x)$ and $p_{\theta_1}(x)$, respectively. If $\mathscr{A}$ and $\mathscr{B}$ are two sets with $P_{\theta_0}(\mathscr{A}\setminus\mathscr{B}) = P_{\theta_0}(\mathscr{B}\setminus\mathscr{A})$ and*

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < \frac{p_{\theta_1}(x^*)}{p_{\theta_0}(x^*)}$$

for all $x \in \mathscr{A}\setminus\mathscr{B}$ and $x^* \in \mathscr{B}\setminus\mathscr{A}$, then the test with acceptance region $\mathscr{A}$ is for θ_1 more powerful than the one with acceptance region $\mathscr{B}$, both having the same significance level.

Now we can prove the

THEOREM. *For the test problem*

$$H_0\colon \psi^2 \leqslant \psi_0^2 \quad \textit{against} \quad H_1\colon \psi^2 > \psi_0^2$$

we have, provided the tests based on the three test statistics discussed above all have the same significance level,

$$\beta_{m+1,n}(\psi^2) < \beta_{m,n}(\psi^2) < \beta_{m,n+1}(\psi^2)$$

for all $\psi^2 > \psi_0^2$.

Proof. Let $x_i \sim N\left(\frac{\psi}{\sqrt{m}}, 1\right)$, $i = 1, \ldots, m$, and $v_j \sim N(0,1)$, $j = 1, \ldots, n+1$, all independently distributed. Further, let

$$\begin{aligned} z_1 &= \frac{x_1^2+\ldots+x_m^2}{x_1^2+\ldots+x_m^2+v_1^2+\ldots+v_{n+1}^2},\\ z_2 &= \frac{v_{n+1}^2}{x_1^2+\ldots+x_m^2+v_1^2+\ldots+v_{n+1}^2}. \end{aligned} \tag{11}$$

Then, if we denote by $p_{\psi^2}(z_1, z_2)$ the density of the variable (z_1, z_2), the likelihood ratio $p_{\psi_1^2}(z_1,z_2)/p_{\psi_0^2}(z_1,z_2)$ for $\psi_1^2 > \psi_0^2$ is a monotone increasing function of z_1 alone. This may be derived from the fact that $\frac{x_1^2+\ldots+x_m^2}{v_1^2+\ldots+v_{n+1}^2}$ and $\frac{v_1^2}{v_1^2+\ldots+v_m^2}$ are independently distributed, the latter not influenced by ψ, and from the existence of a monotone likelihood ratio for the F-distribution.

Since the $F_{m,n+1}$-test corresponds to a test with acceptance region $z_1 \leqslant c_\alpha$, and the $F_{m+1,n}$-test to a test with acceptance region $z_1+z_2 \leqslant c'_\alpha$, while the $F_{m,n}$-test corresponds to the acceptance region $\frac{z_1}{c''_\alpha}+z_2 \leqslant 1$, with appropriately chosen constants c_α, c'_α and c''_α which ensure the significance level α, it immediately follows that the $F_{m,n+1}$-test, as the Neyman–Pearson test, is the most powerful one among the three.

It remains to show that $F_{m,n}$ is better than $F_{m+1,n}$. Looking at Figure 1, we see, however, that the lemma can be applied if we denote by $\mathscr{A}$ the acceptance region of the $F_{m,n}$-test and by $\mathscr{B}$ that of the $F_{m+1,n}$-test. The z_1-coordinates and therefore

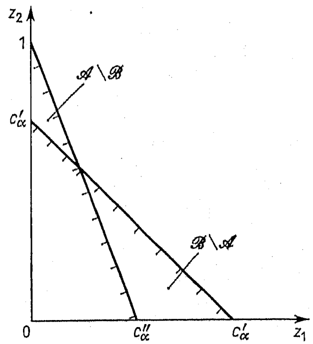

Fig. 1

the likelihood ratio are smaller on $\mathscr{A}\setminus\mathscr{B}$ than on $\mathscr{B}\setminus\mathscr{A}$. Thus the theorem is proved. ■

In their remark 2.2 Das Gupta and Perlman [1] already briefly discussed the geometric approach of the proof of the theorem, without using it, however, for the comparison of $F_{m+1,n}$ and $F_{m,n}$.

4. Multivariate analysis of variance

(It immediately follows from the preceding section that, if we have a normally distributed p-vector $x \sim N_p(\mu;\Sigma)$ and a Wishart-distributed variable with n degrees of freedom $S \sim W_p(n,\Sigma)$, Hotelling's α-test for $H_0\colon \mu = 0$ against $H_1\colon \mu \neq 0$ for $H_0\colon \mu'\Sigma^{-1}\mu \leqslant \psi_0^2$ against $H_1\colon \mu'\Sigma^{-1}\mu > \psi_0^2$), based on the F-distributed up to a factor statistic $x'S^{-1}x$, has increasing power for raising n. This situation arises as a special case of MANOVA models which we will describe in its canonical form as follows.

Let Y_1 and Y_2 be random matrices with independently normally distributed p-dimensional row vectors, each having the same covariance matrix Σ. We assume $EY_1 = M$ and $EY_2 = 0$, i.e.

$$Y_1 \sim N_{r\times p}(M,\Sigma),\quad Y_2 \sim N_{n\times p}(0,\Sigma). \tag{12}$$

With $S_H = Y_1'Y_1$ and $S_E = Y_2'Y_2$ the usual MANOVA procedures for testing $H_0\colon M = 0$ against $H_1\colon M \neq 0$ are based on the characteristic roots of $S_H(S_H+S_E)^{-1}$ such as Pillai's $V = \operatorname{tr} S_H(S_H+S_E)^{-1}$, Wilks' $\Lambda = |S_E|/|S_H+S_E|$, Roy's $\lambda_{\max} = \lambda_{\max} S_H(S_H+S_E)^{-1}$ and Hotelling's $T_0^2 = \operatorname{tr} S_H S_E^{-1}$. Having now a random (row-)vector

$$Y_3 \sim N_{1\times p}(0,\Sigma) \tag{13}$$

and denoting $S_G = Y_3'Y_3$, we can express the monotonicity property we wished to show in accordance with the arguments used in the first two sections in the following way:

Each of the above-mentioned tests is better than the corresponding one with S_H replaced by S_H+S_G, but worse than the one with S_E replaced by S_E+S_G, provided they all have the same level of significance. Transforming

$$Z_1 = Y_1 S_E^{-1/2},\quad Z_2 = Y_3 S_E^{-1/2},\quad W = S_E,$$

we get after some manipulations the density

$$p_M(Z_1,Z_2) = c\cdot|Z_1'Z_1+Z_2'Z_2+I|^{-n/2}\times \tag{14}$$
$$\times \iiint e^{\operatorname{tr} W/2}|W|^{1/2(n-p-1)}\exp\left(\operatorname{tr}(Z_1'Z_1+Z_2'Z_2+I)^{-1/2}Z_1'CMV\Gamma\right)d(C)\,d(\Gamma)\,dW,$$

where $d(C)$ and $d(\Gamma)$ denote the Haar-measures on the orthogonal groups $\mathcal{O}_r$ and $\mathcal{O}_p$, respectively. The second factor of (14) depends for fixed M on the characteristic roots of

$$Z_1'Z_1(Z_1'Z_1+Z_2'Z_2+I)^{-1} = Y_1'Y_1(Y_1'Y_1+Y_3'Y_3+S_E)^{-1} = S_H(S_H+S_G+S_E)^{-1},$$

$\lambda_1,\ldots,\lambda_s$, $s = \min(r,p)$ only; so they form a sufficient statistic and the tests based on them form an essentially complete class among the tests based on Z_1 and Z_2. On the other hand, from (14) via the monotonicity and convexity of the likelihood ratio $p_M(Z_1,Z_2)/p_0(Z_1,Z_2)$ in the characteristic roots $\lambda_1,\ldots,\lambda_s$ it follows that an admissible test has an acceptance region which is convex and monotone in $\lambda_1,\ldots,\lambda_s$ (see also Schwartz [4]).

For each test φ based on the characteristic roots of $S_H(S_H+S_E)^{-1}$ — where we do not use S_G — or of $(S_H+S_G)(S_H+S_G+S_E)^{-1}$ — where S_H is replaced by S_H+S_G — we can construct another one φ_0 depending only on $\lambda_1,\ldots,\lambda_s$ with the same power function, namely its conditional expectation $E[\varphi(Z_1,Z_2)|\lambda_1,\ldots,\lambda_s]$. The test φ_0, however, is strictly randomized and so contradicts the above-mentioned conditions of an admissible test. Applying for instance Pillai's V to the model (12), (13), we see that among the competitors $\operatorname{tr} S_H(S_H+S_E)^{-1}$, $\operatorname{tr}(S_H+S_G)(S_H+S_G+S_E)^{-1}$ and $S_H(S_H+S_G+S_E)^{-1}$ only the last one can be admissible. (For its admissibility see e.g. Schwartz [4].) However it has not been proved yet that this one really dominates the other two. The only result which seems to exist in that direction in MANOVA is due to Das Gupta and Perlman [1], who proved that for the Wilks test $|S_E|/|S_H+S_E|$ leads to a more powerful test than $|S_E|/|S_H+S_G+S_E|$, provided the rank of the unknown M is one.

References

[1] S. Das Gupta and M. D. Perlman, *On the power of Wilks' U-test for* MANOVA, J. Mult. Analysis 3 (1973), pp. 220–225.
[2] —, —, *Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2-test*, J. Amer. Statist. Assoc. 69 (1974), pp. 174–180.
[3] E. L. Lehmann, *Testing statistical hypotheses*, Wiley, New York 1959.
[4] R. Schwartz, *Admissible tests in* MANOVA, Ann. Math. Statist. 38 (1967), pp. 698–710.

Added in proof: Only after this lecture the paper by B. K. Ghosh came to knowledge of the author, which overlaps the second inequality in the theorem, but not the first one.

[5] B. K. Ghosh, *Some monotonicity theorems for χ^2, F and t distributions with applications*, J. Royal Statist. Soc., Ser. B, 35 (1973), pp. 480–492.

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976