PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 15 (1995), pp. 67–71

NEYMAN AND THE BEHRENS–FISHER PROBLEM – AN ANECDOTE

BY

GEORGE A. BARNARD (BRIGHTLINGSEA)

Through disagreement clearly expressed we make progress.

J. Neyman

Abstract. An anecdote relating to the Behrens–Fisher problem illustrates the need for statisticians to be familiar with all four main approaches to the foundations of statistical inference, and with their relations to each other.

In 1974 the Indian Statistical Institute organised a memorial conference for two great contributors to statistical theory and practice – Prasanta Mahalanobis and Yuri Linnik. I was privileged to take part in the conference and to deliver a lecture which was attended by Jerzy Neyman. My topic was the *generalised Behrens–Fisher problem* (GBF):

We are given two samples of independent observations x_i, $i = 1, 2, \ldots, m$ and y_j, $j = 1, 2, \ldots, n$. The distributions of the two samples are specified by taking as the "basic pivotals"

$$p_i = (x_i - \lambda)/\sigma, \qquad q_j = (y_j - \lambda - \delta)/\varrho\sigma$$

with joint density

$$\prod_i \varphi(p_i) \times \prod_j \psi(q_j).$$

The parameters λ, δ, ϱ and σ are unknown and variation-independent of each other, while the densities φ and ψ are known only approximately. We wish to test the composite hypothesis H_0: $\delta = 0$ with risk of error of the first kind not exceeding ε. In the original, *normal Behrens–Fisher problem* (NBF) both densities are known to be normal.

Neyman and Pearson introduced the concept of "risk of error of the first kind" in their classic paper (1933). In the section devoted to composite hypotheses Neyman and Pearson wrote:

> In the first place it is *evident* [my stress – G.B.] that a necessary condition for a critical region, w, suitable for testing H_0 is that
>
> $$P_0(w) = \iiint \ldots \int_{\omega} p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n = \text{constant} = \varepsilon \qquad (72)$$
>
> for *every* simple hypothesis of the subset ω. [J. Neyman and E. S. Pearson, *Joint Statistical Papers*, p. 163]

The requirement thus imposed became known later as that of "Similarity" (S) and it was satisfied by the principal tests involving continuous observations in use in 1933. But in 1935 Fisher proposed a "solution" to the NBF problem which was shown by Bartlett to fail to satisfy S.

In the GBF problem, Neyman and Pearson's subset ω is the subset of points $(\sigma, \varrho, \lambda, \delta)$ in $W = R^+ \times R^+ \times R \times R$ in which $\delta = 0$. For the original NBF, a nonrandomised test corresponds to a function $\chi(s_x, r, \bar{x}, \bar{y})$ of the minimal sufficient statistic $(s_x, r = s_y\sqrt{m}/s_x\sqrt{n}, \bar{x}, \bar{y})$ taking only the values either 0 or 1, with the interpretation that $\chi(s_{x_0}, r_0, \bar{x}_0, \bar{y}_0) = 1$ implies rejection of the hypothesis. Here the suffix $_0$ appended to an observable means that it is to be replaced by its observed value. To satisfy the condition S, we must have

$$\mathrm{E}(\chi | H_0) \equiv \varepsilon$$

(where $\mathrm{E}(\cdot|H_0)$ denotes expectation on H_0) for all points in the subset ω. In 1962 Linnik showed that such χ's exist, but they necessarily have properties which rule them out from serving as test functions. For example, there would be values $(s_{x_0}, r_0, \bar{x}_0, \bar{y}_0)$ with the standardised observed difference of sample means $|\bar{y}_0 - \bar{x}_0|/\{s_{x_0}^2/m + s_{y_0}^2/n\}^{1/2}$ arbitrarily small, yet for these values $\chi(s_{x_0}, r_0, \bar{x}_0, \bar{y}_0) = 1$ so that H_0 is rejected. On the other hand, any test function which was not a function of the minimal sufficient statistic would be equivalent to a randomised test which would also be ruled out on common sense grounds. In short, the "similarity" requirement S was, in this problem, in conflict with common sense.

In pivotal inference, when interest is focussed on a parameter function $\alpha(\lambda, \delta, \varrho, \sigma)$, we attempt to find functions Q, Q', Q'' of the basic pivotals such that Q is the maximal pivotal function involving only observables, Q' is the maximal pivotal function involving only observables and α, while Q'' does not involve α. It is only when this is possible that we can rigorously infer a "pivotal distribution" for α. In the GBF problem we can set

$$\boldsymbol{p} = s(t_x \boldsymbol{1} + \boldsymbol{a}_1), \qquad \boldsymbol{q} = zs(t_y \boldsymbol{1} + \boldsymbol{a}_2)$$

subject to

$$\boldsymbol{1}^{\mathrm{T}}\boldsymbol{a}_1 = 0, \quad \boldsymbol{1}^{\mathrm{T}}\boldsymbol{a}_2 = 0, \quad \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 = m(m-1), \quad \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 = n(n-1).$$

We then find, with a little algebra, that $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2)$ and $\boldsymbol{b} = (s, z, t_x, t_y)$, where, with the usual sample notation,

$$\boldsymbol{a}_1 = (\boldsymbol{x} - \bar{x}\boldsymbol{1})\sqrt{m/s_x}, \quad \boldsymbol{a}_2 = (\boldsymbol{y} - \bar{y}\boldsymbol{1})r\sqrt{m/s_x},$$

and

$$s = s_x/\sigma\sqrt{m}, \quad z = r/\varrho, \quad t_x = (\bar{x} - \lambda)\sqrt{m/s_x}, \quad t_y = (\bar{y} - \lambda - \delta)r\sqrt{m/s_x}.$$

Obviously, $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ is the maximal Q. Then with $\alpha = (\lambda + \delta, \varrho\sigma)$, for example, we note that $\lambda + \delta$ occurs only in t_y, while $\varrho\sigma$ will occur in the product zs. Thus for this α we take $Q' = (t_y, zs)$, leaving $Q'' = (t_x, s)$ or, equivalently, (t_x, z). Then, if the conditional density $pf(Q', Q'')$ given $Q = Q_0$ is approximately $\xi(Q', Q'')$, since α is variation-independent of (λ, σ), the parameters involved in Q'', our information about α is contained in Q'_0 together with the marginal density $\int \xi(Q', Q'')dQ''$ of Q'. Assuming, with Fisher, that Q'_0, now a function of α, retains this marginal density when the observations are known, we find an approximate pivotal density for α which we can use to assess the relative plausibility, in the light of the data, of various propositions about α. If, as in the NBF problem, the pivotal densities φ and ψ were known exactly, this pivotal density would be exact, and it would coincide with the Fisher fiducial density for α. It often happens that a smoothing effect operates to make the accuracy with which the marginal density of Q' is known considerably exceed the accuracy with which φ and ψ are known.

But when, as in the GBF problem, the parameter of interest is δ, we note that δ occurs only in t_y, and to remove its companion in t_y we must subtract rt_x. Since $r = \varrho z$, we are led to consider

$$t_y - \varrho z t_x \equiv (\bar{y} - \bar{x} - \delta)r\sqrt{m/s_x} \equiv (\bar{y} - \bar{x} - \delta)\sqrt{n/s_y},$$

and this is essentially the only function of pivotals which involves only observables and δ. Except that it is not a function of our pivotals! It involves ϱ.

If we knew something, apart from the data, about ϱ, expressible in the form of an approximate density $\zeta(\varrho)$ for ϱ, we could express this by taking the basic pivotals to be $\boldsymbol{p}$, $\boldsymbol{q}$, and ϱ with joint density:

$$\left[\prod_i \varphi(p_i) \times \prod_j \psi(q_j) | \varrho\right] \times \zeta(\varrho).$$

Then the above transformations would apply to the distributions within the [ ] brackets. But the pivotal function $\varrho z = r$ would also be parameter-free, and so the maximal Q would extend to $[\boldsymbol{a}_1, \boldsymbol{a}_2, r]$. Then, conditioning the joint density of s, r, t_x, t_y within the [ ] brackets on $r = r_0$, we could form the pivotal function

$$v = r_0 t_y - t_x = (\bar{y} - \bar{x} - \delta)\sqrt{m/s_x}$$

and the marginal density of this, averaged over $\zeta(\varrho)$, will be:

$$\zeta(v|\boldsymbol{a}_0, r_0) = K\iiiint[(s^{m+n-1}/\varrho^{n-1})\prod_i \varphi(s(t_x+a_{1i_0})) \times \prod_j \psi((r_0/\varrho)s(t_y+a_{2j_0}))]\zeta(\varrho)d\varrho dt_x dt_y ds$$

from which we can derive the pivotal distribution of δ based on the available data.

In my Calcutta lecture I suggested that to demand "similarity" of tests was analogous to making the other demand sometimes made by Neyman – that a "point estimate" of a parameter should be "unbiased." In many cases to require unbiasedness of the estimate means that the only values which the estimate can take are impossible values for the parameter. Analogously, in the GBF and the NBF problems the "similarity" requirement rules out all reasonable solutions.

I was very nearly 60 years old at the time of my Calcutta lecture – a dangerous age, when one tends to think one has absorbed what is good in the teaching of one's predecessors, while rejecting their mistakes. Thinking my forthright criticisms might have upset the 80-year old Neyman, I approached him after the lecture to assure him that, in spite of my expressed disagreement, I held him personally in the very highest regard. He gave me his always warm smile, and said what I have placed in the epigraph to this paper.

I am now nearly as old as Neyman was then. After 70 one gains in hindsight, and even, perhaps, a little in wisdom. Fisher's mistake concerning his fiducial argument can now be seen to have arisen from a confusion between a "random variable" in the sense of Kolmogoroff and a "random variable" in the sense of "a quantity to which a probability distribution can be attached"; a footnote added to volume 3, p. 395, of his *Collected Papers* suggests that, when he died at 72, Fisher had begun to perceive the possible confusion.

In attacking Neyman's condition S, I went too far. I ignored the fact that there are two stages in statistical inference – the planning stage and the inferential stage. It is often essential to make some advance estimate of the number of observations required in an experiment to achieve a specified low risk ε of drawing the wrong conclusion. And for this purpose a serious attempt needs to be made to secure the condition S. Furthermore, it is important that the plans should be followed so far as possible lest bias in treatment allocation or in assessment of outcome should arise. But once the data begin to come in, we know more than we did when we were planning, and our final inferences must take account of this additional knowledge as far as possible.

In both the GBF and the NBF problems the analysis above shows that the unknown value of ϱ is important. Thus in practice we must make an advance guess about the value of ϱ, expressible in the form of an approximate prior $\zeta(\varrho)$. Then for a given $m+n$ (or, more generally, given cost $c_1 m + c_2 n$) we can adjust

m/n so as to give the best chance of obtaining a z value which gives maximal sensitivity.

Fisher's solution to the NBF problem arises if we take φ and ψ to be normal and $\log \varrho$ to be (approximately) uniformly distributed. The condition S will then be satisfied "on the average," in repeated sampling provided the ϱ values follow the assumed distribution. But in practice $\log \varrho$ is unlikely to be uniformly distributed; in the recovery of interblock information, for example, the between block variance is usually larger than that within blocks. And, more generally, taking $\log \varrho$ to be uniformly distributed *independently* of $\log \sigma$ suggests a degree of ignorance about the x and the y errors which is most exceptional.

Nowadays we have four broad approaches to statistical inference, one due to Fisher, another due to Neyman and Pearson, another due to Ramsey and de Finetti, and a fourth due to Jeffreys. Each can be seen to have its appropriate sphere of application. In planning a trial we must make guesses which should be checked for coherence in the sense of de Finetti; in applying the guesses to our choice of sample sizes etc. we need to use the Neyman–Pearson concept of the power function and to aim at the condition S; and finally, if we wish to make inferences with maximal objectivity, we need the Fisher–Jeffreys approach, conditional on the data. Finally, if we are prepared to accept some degree of subjectivity in our personal conclusions, we can condition not only on the data but also on our personal priors and in that way complete the circuit. What Neyman said about differences clearly expressed is true, though clear expression is often hard to achieve.

George A. Barnard
Mill House, 54 Hurst Green
Brightlingsea
Colchester, Essex
England, 3070 EH

Received on 3.9.1994