CONFIDENCE INTERVAL OF PREASSIGNED LENGTH FOR THE BEHRENS-FISHER PROBLEM

By Saibal Banerjee

Indian Statistical Institute, Calcutta

0. Summary. It is shown that by drawing multiple samples from $N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ it is possible to have confidence interval of preassigned length for the Behrens-Fisher problem.

1. Introduction. Given a normal population $N(\mu, \sigma^2)$ by drawing two samples as specified by Stein [5] it is possible to have confidence interval of preassigned length for the population mean μ. It is also possible [5] by adopting the same procedure to ensure that the probability of accepting the hypothesis $H_0(\mu = \mu_0)$ when an alternative hypothesis $H'(\mu = \mu')$ is true, is equal to some preassigned value $1 - \beta$ $(0 < \beta < 1)$. Given two independent samples of n_i units from two normal populations $N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ it is possible ([1], [2]) to have confidence interval for $c_1\mu_1 + c_2\mu_2$ (where $c_i(i = 1, 2)$ are known constants) in terms of sample estimates of population means and variances; it is also possible ([1], [2], [3]) to test the hypothesis $H_0(c_1\mu_1 + c_2\mu_2 = M_0)$ on the basis of the sample estimates of population means and variances with error of the first kind less than or equal to α $(0 < \alpha < 1)$. It is now shown that given two normal populations $N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ by drawing multiple samples (in all four samples, two samples from each population) it is possible to have confidence interval of preassigned length for $c_1\mu_1 + c_2\mu_2$ (where c_i $(i = 1, 2)$ are known constants) with confidence coefficient greater than or equal to some preassigned value $1 - \alpha$ $(0 < \alpha < 1)$. It is also shown that by adopting the same procedure it is possible to ensure that the probability of accepting the hypothesis $H_0(c_1\mu_1 + c_2\mu_2 = M_0)$, when an alternative hypothesis $H'(c_1\mu_1 + c_2\mu_2 = M')$ is true, is equal to or less than some preassigned value $1 - \beta$ $(0 < \beta < 1)$. The operational procedure as specified ensures selection of the final samples (or the second stage samples) from the two populations in such a way that the total cost of selecting the second stage samples is approximately minimized.

2. Procedure. Let there be two populations $N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ and suppose it is required to have a confidence interval of preassigned length for the linear function $c_1\mu_1 + c_2\mu_2$ (where c_1 and c_2 are known constants). (When $c_1 = 1$ and $c_2 = -1$ we get the Behrens-Fisher problem). Let $1 - \alpha$ and 2Δ denote respectively the preassigned confidence coefficient and the preassigned length of the confidence interval. The following procedure may be adopted in order to have a confidence interval of length 2Δ with confidence coefficient greater than or equal to $1 - \alpha$. Two samples x_{ij} $(i = 1, 2; j = 1, 2, \cdots, n)$ may be drawn from the two populations providing the estimates

Received 29 November 1966.

(2.1) $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n, \qquad s_i^2 = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2/(n-1).$

Let T_1 be the cost of sampling one unit from the first population and T_2 be the cost of sampling one unit from the second population. Also let t be some positive numerical value satisfying the relation

(2.2) $2\int_{-t}^{t} f(t, \nu)\, dt - \int_{-t'}^{t'} f(t, \nu + 1)\, dt = 1 - \alpha$

where $f(t, \nu)$ denotes frequency function of Student's t-variate with $\nu(= n - 1)$ df and $t' = t(\nu + 1)^{\frac{1}{2}}/\nu^{\frac{1}{2}}$. Now determine θ_1 and θ_2 so that

$$t^2c_1^2s_1^2/(n + \theta_1) + t^2c_2^2s_2^2/(n + \theta_2) = \Delta^2$$

subject to the restraint that T defined as $T = T_1\theta_1 + T_2\theta_2$ is a minimum. By applying Lagrange's multiplier the solution is given by

(2.3) $\theta_i = t^2 \, |c_i| \, s_i(T_1^{\frac{1}{2}} \, |c_1| \, s_1 + T_2^{\frac{1}{2}} \, |c_2| \, s_2)/\Delta^2 T_i^{\frac{1}{2}} - n, \qquad (i = 1, 2).$

Determine two integers n_1 and n_2

(2.4) $n_i = \max \{[\theta_i] + 1, 1\}, \qquad (i = 1, 2),$

where $[q]$ indicates the largest integer less than q. Draw two samples of n_1 and n_2 units respectively from the two populations. Let $\bar{x}_1'$ and $\bar{x}_2'$ be estimates of population means μ_i $(i = 1, 2)$. Define the combined estimates as

(2.5) $z_1 = a_1\bar{x}_1 + (1 - a_1)\bar{x}_1', \qquad z_2 = a_2\bar{x}_2 + (1 - a_2)\bar{x}_2',$

where a_i $(i = 1, 2)$ satisfy the relation

(2.6)
$$a_1^2/n + (1 - a_1)^2/n_1 = 1/(n + \theta_1),$$
$$a_2^2/n + (1 - a_2)^2/n_2 = 1/(n + \theta_2).$$

It can be shown that it is possible to determine a_i $(i = 1, 2)$ satisfying (2.6). There will be two solutions a_{11} and a_{12} for a_1, and either solution may be used in (2.5). Also there will be two solutions a_{21} and a_{22} for a_2, and either solution may be used in (2.5).

Now z_1 and z_2 depend on s_1 and s_2 through the numbers n_1 and n_2. For given s_1 and s_2, $c_1 z_1 + c_2 z_2$ is distributed normally with mean $c_1 \mu_1 + c_2 \mu_2$ and variance $c_1^2\sigma_1^2/(n + \theta_1) + c_2^2\sigma_2^2/(n + \theta_2)$. For fixed s_1^2 and s_2^2 or χ_1^2 and χ_2^2 (where $\chi_i^2 = \nu s_i^2/\sigma_i^2$)

(2.7)
$$P\{(c_1z_1 + c_2z_2 - c_1\mu_1 - c_2\mu_2)^2 \leqq \Delta^2 \mid s_1^2, s_2^2\}$$
$$= P\{y^2 \leqq \omega_1 t^2\chi_1^2/\nu + \omega_2 t^2\chi_2^2/\nu \mid \chi_1^2, \chi_2^2\},$$

where y is distributed as $N(0, 1)$, $\omega_1 = \chi_2/(\chi_2 + \phi\chi_1)$, $\phi = |c_2| \, \sigma_2 T_2^{\frac{1}{2}}/|c_1| \, \sigma_1 T_1^{\frac{1}{2}}$ and $\omega_2 = 1 - \omega_1$. Denoting by $G(u)$ cumulative distribution function of a χ^2 variate with 1 df, it follows from (2.7) that for variation in s_1^2 and s_2^2

(2.8) $P\{(c_1z_1 + c_2z_2 - c_1\mu_1 - c_2\mu_2)^2 \leqq \Delta^2\} = E\{G(\omega_1b_1 + \omega_2b_2)\}, \qquad b_i = t^2\chi_i^2/\nu.$

It can be shown that $G(u)$ is an upward convex function of u, so that

$$G(\omega_1 b_1 + \omega_2 b_2) \geqq \omega_1 G(b_1) + \omega_2 G(b_2).$$

Now

$$(2.9)\qquad \begin{aligned} E\{\omega_1 G(b_1)\} &= E\{(1-\omega_2)G(b_1)\} = E\{G(b_1)\} - E\{\omega_2 G(b_1)\} \\ &= \int_{-t}^{t} f(t,\nu)\,dt - E\{\phi\chi_1 G(b_1)/(\chi_2 + \phi\chi_1)\}. \end{aligned}$$

Let $f(\chi^2, \nu)$ denote frequency function of a χ^2 variate with ν df. Since $1/(\chi_2 + \phi\chi_1)$ monotonically decreases and $G(b_1)$ monotonically increases with χ_1^2 it can be shown that

$$(2.10)\qquad \begin{aligned} &\int_0^\infty f(\chi_1^2, \nu)\{\phi\chi_1 G_1/(\chi_2 + \phi\chi_1)\}\,d\chi_1^2 \\ &= \int_0^\infty Kf(\chi_1^2, \nu+1)\{\phi G_1/(\chi_2+\phi\chi_1)\}\,d\chi_1^2 \\ &= \int_0^\infty Kf(\chi_1^2, \nu+1)\{\phi/(\chi_2+\phi\chi_1)\}\{G_1 - \lambda + \lambda\}\,d\chi_1^2 \\ &= \lambda \int_0^\infty Kf(\chi_1^2, \nu+1)\{\phi/(\chi_2+\phi\chi_1)\}\,d\chi_1^2 \\ &\quad + \int_0^\infty Kf(\chi_1^2, \nu+1)\{\phi/(\chi_2+\phi\chi_1)\}\{G_1 - \lambda\}\,d\chi_1^2 \\ &< \lambda \int_0^\infty Kf(\chi_1^2, \nu+1)\{\phi/(\chi_2+\phi\chi_1)\}\,d\chi_1^2 \\ &= \lambda \int_0^\infty \omega_2 f(\chi_1^2, \nu)\,d\chi_1^2, \end{aligned}$$

where $G_1 = G(b_1)$, $\lambda = \int_0^\infty f(\chi_1^2, \nu+1)G_1\,d\chi_1^2 = \int_{-t'}^{t'} f(t, \nu+1)\,dt$, $t' = t(\nu+1)^{\frac{1}{2}}/\nu^{\frac{1}{2}}$, $K = 2^{\frac{1}{2}}\{\Gamma(\nu/2+\frac{1}{2})\}/\Gamma(\nu/2)$. From (2.9) and (2.10) it follows that

$$(2.11)\qquad E\{\omega_1 G(b_1)\} > \int_{-t}^{t} f(t,\nu)\,dt - E(\omega_2)\int_{-t'}^{t'} f(t,\nu+1)\,dt.$$

Also similarly it can be shown that

$$E\{\omega_2 G(b_2)\} > \int_{-t}^{t} f(t,\nu)\,dt - E(\omega_1)\int_{-t'}^{t'} f(t,\nu+1)\,dt$$

so that

$$(2.12)\qquad P\{(c_1z_1 + c_2z_2 - c_1\mu_1 - c_2\mu_2)^2 \leqq \Delta^2\} > 1 - \alpha$$

by virtue of (2.2). The length of the confidence interval

$$c_1z_1 + c_2z_2 - \Delta \leqq c_1\mu_1 + c_2\mu_2 \leqq c_1z_1 + c_2z_2 + \Delta$$

is 2Δ as preassigned.

Apart from the question of having a confidence interval of preassigned length 2Δ, in Stein's theory by making a suitable choice of the numerical value of Δ, it can be ensured that if μ' be the true value of the mean then the error of the second kind (i.e., the probability of accepting the hypothesis $\mu = \mu_0$) has a preassigned value. For the two-means case as well by suitably choosing numerical value of Δ it can be ensured that if M' be the true value of the mean then the error of the second kind (i.e., the probability of accepting the hypothesis $M = M_0$), has a value not greater than some preassigned value $1 - \beta$. Let M' be the true value of $c_1\mu_1 + c_2\mu_2$ and M_0 be the value by the hypothesis. Let P_1 denote the error of the second kind so that for fixed χ_1^2 and χ_2^2

$$\begin{aligned} P_1 &= P\{(c_1z_1 + c_2z_2 - M_0)^2 \leqq \Delta^2 \mid \chi_1^2, \chi_2^2\} \\ &= P\{|y/tp^{\frac{1}{2}} - d| \leqq 1 \mid \chi_1^2, \chi_2^2\}, \end{aligned} \tag{2.13}$$

where y is distributed as $N(0, 1)$ independently of χ_i^2, $d = (M_0 - M')/\Delta$ and $p = \omega_1 b_1 + \omega_2 b_2$. (2.13) for clarity of exposition may be considered under two headings (i) $|d| \leqq 1$ and (ii) $|d| > 1$. For (i) P_1 is equal to

$$\tfrac{1}{2}E\{G((A_1^2p) + G(A_2^2p)\}, \tag{2.14}$$

where $A_1 = t(1 + |d|)$ and $A_2 = t(1 - |d|)$. Now it can be shown that

$$\omega_1\chi_1^2 + \omega_2\chi_2^2 \leqq \chi_1^2/(1 + \phi) + \phi\chi_2^2/(1 + \phi),$$

so that from [1] and [4]

$$\begin{aligned} P_1 &< \tfrac{1}{2}E\{G(A_1^2q) + G(A_2^2q)\} \\ &\leqq \tfrac{1}{2}E\{G(A_1^2q') + G(A_2^2q')\} \\ &= \textstyle\int_0^{A_1} f(t, 2\nu)\, dt + \int_0^{A_2} f(t, 2\nu)\, dt, \end{aligned} \tag{2.15}$$

where $q = (\chi_1^2 + \phi\chi_2^2)/\nu(1 + \phi)$ and $q' = (\chi_1^2 + \chi_2^2)/2\nu$.
For (ii) it can be shown that

$$\begin{aligned} P_1 &= \tfrac{1}{2}E\{1 - G(B_1^2p)\} - \tfrac{1}{2}E\{1 - G(B_2^2p)\} \\ &< \tfrac{1}{2}E[\textstyle\sum_{i=1}^2 \omega_i\{1 - G(B_1^2\chi_i^2/\nu)\}] - \tfrac{1}{2}E\{1 - G(B_2^2q')\} \\ &= \tfrac{1}{2} - \tfrac{1}{2}E\{\textstyle\sum_{i=1}^2 \omega_i G(B_1^2\chi_i^2/\nu)\} - \tfrac{1}{2}E\{1 - G(B_2^2q')\} \\ &< 2\textstyle\int_{B_1}^\infty f(t, \nu)\, dt - \int_{B_1'}^\infty f(t, \nu + 1)\, dt - \int_{B_2}^\infty f(t, 2\nu)\, dt, \end{aligned} \tag{2.16}$$

where $B_1 = t(|d| - 1)$, $B_1' = B_1(\nu + 1)^{\frac{1}{2}}/\nu^{\frac{1}{2}}$ and $B_2 = t(|d| + 1)$. From (2.15) and (2.16) it therefore follows that given M_0 and M', Δ can be so determined so that P_1 is less than some preassigned value.

An alternate procedure is possible whereby computations of a_i as defined in (2.6) can be avoided. This procedure, however, is less powerful in detecting deviations from the hypothesis $H_0(c_1\mu_1 + c_2\mu_2 = M_0)$ when an alternative hypothesis $H'(c_1\mu_1 + c_2\mu_2 = M')$ is true for $|M_0 - M'| \leqq \Delta$. After drawing the initial samples determine second stage samples n_i' by

$$n_i' = \max\{[\theta_i] + 1, 0]. \tag{2.17}$$

Denoting by $\bar{x}_i''$ estimates of μ_i based on n_i' units combined estimates may be defined as

$$z_i' = (n\bar{x}_i + n_i'\bar{x}_i'')/(n + n_i').$$

Now for fixed χ_1^2 and χ_2^2

$$(2.18) \quad P\{(c_1z_1' + c_2z_2' - c_1\mu_1 - c_2\mu_2)^2 \leqq \Delta^2 \mid \chi_1^2, \chi_2^2\} = P\{y^2 \leqq \omega_1' t^2 \chi_1^2/\nu + \omega_2' t^2 \chi_2^2/\nu \mid \chi_1^2, \chi_2^2\},$$

where y is distributed as $N(0, 1)$, $\omega_i' = c_i^2\sigma_i^2/(n + \theta_i)l$ and $l = c_1^2\sigma_1^2/(n + n_1') + c_2^2\sigma_2^2/(n + n_2')$. As ω_i' is greater than ω_i $(i = 1, 2)$, $G(\omega_1'b_1 + \omega_2'b_2)$ is greater than $G(\omega_1 b_1 + \omega_2 b_2)$ and the confidence interval $c_1z_1' + c_2z_2' - \Delta \leqq c_1\mu_1 + c_2\mu_2 \leqq c_1z_1' + c_2z_2' + \Delta$ will have confidence coefficient greater than $1 - \alpha$. Let P_2 denote the probability of accepting the hypothesis $H_0(c_1\mu_1 + c_2\mu_2 = M_0)$ when an alternative hypothesis $H'(c_1\mu_1 + c_2\mu_2 = M')$ is true. It can be shown that for fixed values of χ_1^2 and χ_2^2

$$(2.19) \quad P_2 = P\{|y/tr^{\frac{1}{2}} - d| \leqq 1 \mid \chi_1^2, \chi_2^2\},$$

where y is distributed as $N(0, 1)$ and $r = \omega_1'b_1 + \omega_2'b_2$. From (2.16) and (2.19) it therefore follows that for variation in χ_1^2 and χ_2^2 for $|d| > 1$,

$$\begin{aligned} P_2 &< \tfrac{1}{2}E\{1 - G(B_1^2 r)\} - \tfrac{1}{2}E\{1 - G(B_2^2 r)\} \\ (2.20) \qquad &< \tfrac{1}{2}E\{1 - G(B_1^2 p)\} \\ &< 2\int_{B_1}^{\infty} f(t, \nu)\, dt - \int_{B_1'}^{\infty} f(t, \nu + 1)\, dt. \end{aligned}$$

From (2.20) it follows that given M_0, M' and $1 - \beta$, by making Δ small so that $|(M_0 - M')/\Delta|$ is greater than 1, P_2 can be made less than $1 - \beta$.

REFERENCES

[1] BANERJEE, S. (1961). On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t-table. *Sankhyā Ser. A* **23** 359–378.

[2] HÁJEK, J. (1962). Inequalities for the generalized Student's distribution. *Select. Transl. Math. Statist. Prob.* **2** 63–74.

[3] LAWTON, WILLIAM H. (1965). Inequalities for central and non-central distributions. *Ann. Math. Statist.* **36** 1521–1525.

[4] MICKEY, M. RAY and BROWN, MORTON B. (1966). Bounds on the distribution functions of the Behrens-Fisher statistic. *Ann. Math. Statist.* **37** 639–642.

[5] STEIN, C. M. (1945). Two sample test of a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16** 243–258.