ON THE CONSISTENCY OF THE TWO-SAMPLE EMPTY CELL TEST

M. Csorgo* and Irwin Guttman

(received August 15, 1962)

1. Introduction. This paper considers the consistency of the two-sample empty cell test suggested by S. S. Wilks [2]. A description of this test is as follows: Let a sample of n_1 independent observations be taken from a population whose cumulative distribution function $F_1(x)$ is continuous, but otherwise unknown. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n_1)}$ be their order statistics. Let a second sample of n_2 observations be taken from a population whose cumulative distribution function is $F_2(x)$, assumed continuous, but otherwise unknown.

Define cells $I_1, \ldots, I_{n_1+1}$ by

$$(1.1) \qquad I_i = (X_{(i-1)}, X_{(i)}], \quad i = 1, \ldots, n_1 + 1 ,$$

where $X_{(0)} = -\infty$ and $X_{(n_1+1)} = +\infty$.

Let $r_1, \ldots, r_{n_1+1}$ be the number of observations of the second sample that lie in $I_1, \ldots, I_{n_1+1}$ respectively. Let S_o be the number of I_i, $i = 1, \ldots, n_1 + 1$ which are

* Work supported by a National Research Council of Canada Studentship.

such that $r_i = 0$, that is, the number of empty cells. Under the hypothesis that $F_1 = F_2$, Wilks in [2] and [3] gives a somewhat complicated analytic derivation of the probability function of S_o and obtains the result

$$(1.2) \qquad P(S_o = s_o) = \frac{\binom{n_1+1}{s_o}\binom{n_2-1}{n_1-s_o}}{\binom{n_1+n_2}{n_1}} = p(s_o)$$

where the sample space of S_o is given by

$$\mathscr{S} = [k,\ k+1,\ \ldots,\ n_1] \text{ and } k = \max[0,\ n_1+1-n_2].$$

A simplified proof of (1.2) may be found in [4].

Using (1.2), it can be easily shown that

$$E(S_o) = \frac{n_1(n_1+1)}{n_1+n_2}$$

$$(1.3) \qquad \sigma^2(S_o) = \frac{n_1^2(n_1^2-1)}{(n_1+n_2)(n_1+n_2-1)} + \frac{n_1(n_1+1)}{n_1+n_2} - \frac{n_1^2(n_1+1)^2}{(n_1+n_2)^2}$$

(For these results see Wilks [2] and [3] where the method of factorial moments is used to obtain them.)

If we let $n_2 = \rho n_1 + O(1)$, $\rho > 0$, these reduce to

$$(1.4) \qquad \begin{aligned} E(S_o) &= n_1\left(\left(\frac{1}{1+\rho}\right) + O\left(\frac{1}{n_1}\right)\right) \\ \sigma^2(S_o) &= n_1\left(\frac{\rho^2}{(1+\rho)^3} + O\left(\frac{1}{n_1}\right)\right) \end{aligned}$$

https://doi.org/10.4153/CMB-1964-008-x Published online by Cambridge University Press

which in turn imply that

$$E\left(\frac{S_o}{n_1+1}\right) \to \frac{1}{1+\rho}$$

and

$$\sigma^2\left(\frac{S_o}{n_1+1}\right) \to 0$$

as $n_1, n_2 \to \infty$, and by Tchebychev's inequality, these results imply that S_o/n_1+1 converges in probability to $\frac{1}{1+\rho}$, if $F_1 = F_2$.

We can use these results to make a test of the hypothesis $F_1 = F_2$ at the approximate 100α% level. This is given by

(1.5)
$$\begin{cases} \text{Reject if } s_o \geq b \\ \text{Accept otherwise} \end{cases}$$

where b is such that

(1.6)
$$P(S_o \geq b) = \sum_{s_o=b}^{n_1} p(s_o) \leq \alpha$$

$$P(S_o \geq b-1) = \sum_{s_o=b-1}^{n_1} p(s_o) > \alpha .$$

Tables of (1.6) have been tabulated by the authors for α = .01 and .05 and published in Technometrics [4].

2. Consistency. The form of the test (1.5) follows from the following considerations. Let G_o be the class of pairs of

continuous cumulative density functions $(F_1(x), F_2(x))$ such that $F_1(x) = F_2(x)$. Let $F_1^{-1}(u)$ be the inverse of the c.d.f. $F_1(x)$ and let G_1 be the class of pairs of continuous c.d.f.'s $(F_1(x), F_2(x))$ satisfying:

(i) $F_2(F_1^{-1}(u))$ has a derivative, say $g(u)$, for all u on $(0, 1)$ except possibly for a set of probability measure zero.

(ii) The derivatives of $F_2(F_1^{-1}(u))$ and $F_1(F_1^{-1}(u)) = u$ with respect to u on $(0, 1)$ differ over a set of positive probability.

In [3] Wilks states the following

THEOREM. The test defined by (1.5) and (1.6) is consistent for testing any $(F_1, F_2) \epsilon G_o$ against any $(F_1, F_2) \epsilon G_1$ as $n_1, n_2 \to \infty$ so that $n_2 = n_1 \rho + O(1)$, where $\rho > 0$.

To prove this theorem it is sufficient to show that if $(F_1, F_2) \epsilon G_1$, $S_o/(n_1+1)$ converges in probability to a number greater than $1/(1+\rho)$ as $n_1, n_2 \to \infty$ with $\frac{n_2}{n_1} \to \rho > 0$, for it will be recalled from (1.4) that $1/(1 + \rho)$ is the quantity to which $S_o/(n_1 + 1)$ converges in probability if $(F_1, F_2) \epsilon G_o$.

We recall that $r_1, \ldots, r_{n_1+1}$ denote the number of observations of the second sample that lie in the $n_1 + 1$ cells $I_1, \ldots, I_{n_1+1}$ respectively. For each non-negative integer r, let $Q_{n_1}(r)$ be the proportion of values among $r_1, \ldots, r_{n_1+1}$ which are equal to r. Then, in particular, we have

$Q_{n_1}(0) = \frac{S_o}{n_1+1}$, the proportion of empty cells.

Under the conditions (i) and (ii) of this section, J. R. Blum and L. Weiss in [1] prove that

$$(2.1) \qquad P\left[\lim_{(n_1, n_2;\rho)} \sup_{r \geq 0} \left| Q_{n_1}(r) - Q(r) \right| = 0\right] = 1$$

where $\lim_{(n_1, n_2;\rho)}$ denotes the limit as $n_1 \to \infty$, $n_2 \to \infty$ in such a way that $n_2/n_1 \to \rho$, $\rho > 0$, and

$$(2.2) \qquad Q(r) = \rho^r \int_0^1 \frac{g^2(u)}{[\rho + g(u)]^{r+1}} \, du$$

where $g(u)$ is the derivative of $F_2(F_1^{-1}(u))$, satisfying conditions (i) and (ii) of this section.

As a special case of (2.1) we have that

$$(2.3) \qquad P\left[\lim_{(n_1, n_2;\rho)} \left| Q_{n_1}(0) - Q(0) \right| = 0\right] = 1$$

if $(F_1, F_2) \in G_1$, where we have now that

$$(2.4) \qquad Q(0) = \int_0^1 \frac{g^2(u)}{[\rho + g(u)]} \, du .$$

It is also implied by (2.3) that

$$(2.5) \qquad \lim_{(n_1, n_2;\rho)} P(|Q_{n_1}(0) - Q(0)| \geq \epsilon) = 0$$

for any $\epsilon > 0$, however small, if $(F_1, F_2) \in G_1$; that is $Q_{n_1}(0) = \frac{S_o}{n_1+1}$ converges in probability to $Q(0)$ (expression (2.4)). Therefore, the test defined by (1.5) and (1.6) is consistent for testing any $(F_1, F_2) \in G_o$ against any $(F_1, F_2) \in G_1$ if

https://doi.org/10.4153/CMB-1964-008-x Published online by Cambridge University Press

$$(2.6) \qquad \int_0^1 \frac{g^2(u)}{[\rho + g(u)]} \, d\,u > \frac{1}{1+\rho} \quad ,$$

where we recall from (1.4) that $1/1+\rho$ is the quantity to which $Q_{n_1}(0) = \frac{S_o}{n_1+1}$ converges in probability if $(F_1, F_2) \in G_o$.

The inequality of (2.6) is proved as follows. We have by Schwarz's inequality that

$$(2.7) \quad \int_0^1 \frac{g^2(u)d\,u}{\rho + g(u)} \int_0^1 (\rho + g(u))\, d\,u > \left\{ \int_0^1 \frac{g(u)}{\sqrt{\rho + g(u)}} \sqrt{\rho + g(u)} \; d\,u \right\}^2$$

that is

$$\left(\int_0^1 \frac{g^2(u)d\,u}{\rho + g(u)} \right) (\rho + 1) > 1$$

which gives

$$\int_0^1 \frac{g^2(u)d\,u}{\rho + g(u)} > \frac{1}{1+\rho} \quad ,$$

if $g(u)$ differs from unity over a set of positive probability. This condition obtains if $(F_1, F_2) \in G_1$, since the derivatives of $F_2(F_1^{-1}(u)) = g(u)$ and u are assumed to differ over a set of positive probability on $(0, 1)$, and under this condition the above strict Schwarz inequality (2.7) holds. This completes the proof of the above theorem.

https://doi.org/10.4153/CMB-1964-008-x Published online by Cambridge University Press

REFERENCES

1. J. R. Blum and L. Weiss, Consistency of Certain Two-Sample Tests. Ann. Math. Stat., Vol. 28 (1957), pp. 242-246.

2. S. S. Wilks, A Combinatorial Test for the Problem of Two Samples from Continuous Distributions. Proceedings of the Fourth Berkeley Symposium, Vol. I (1961), pp. 707-717.

3. S. S. Wilks, Mathematical Statistics. (Wiley, 1962).

4. M. Csorgo and Irwin Guttman, On the Empty Cell Test. Technometrics, Vol. 4, (1962), pp. 235-247.

McGill University and University of Wisconsin

https://doi.org/10.4153/CMB-1964-008-x Published online by Cambridge University Press