AD-A096 977 INDIANA UNIV AT BLOOMINGTON DEPT OF MATHEMATICS F/G 12/1
ON ESTIMATING INTER-SUBJECT VARIABILITY OF CHOICE PROBABILITIES--ETC(U)
FEB 81 R BARTOSZYNSKI, M L PURI AFOSR-76-2927
UNCLASSIFIED AFOSR-TR-81-0207 NL

1 OF 1
AD A096977

END
DATE
FILMED
4-81
DTIC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

1

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (18) AFOSR-TR- 81-0207 (19)	2. GOVT ACCESSION NO. AD-A096 977	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) On Estimating Inter-Subject Variability of Choice Probabilities under Observability Constraints.		5. TYPE OF REPORT & PERIOD COVERED (9) Interim technical rept.,
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Robert Bartoszyński ~~and~~ Madan L. Puri		8. CONTRACT OR GRANT NUMBER(s) AFOSR-76-2927
9. PERFORMING ORGANIZATION NAME AND ADDRESS Indiana University, Department of Mathematics, Bloomington, Indiana 47405		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (16) (17) 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research, Bolling Air Force Base, D.C. 20332		12. REPORT DATE (11) February 1981
		13. NUMBER OF PAGES 16 (12) 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (15) AFOSR-76-2927, NSF-MCS76-00951		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

LEVEL

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC SELECTED MAR 30 1981 S D A

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Choice probabilities, inter-subject variability, discrimination experiment.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper provides methods for estimating inter-subject variability of the probability of a given event defined in terms of subject's behavior (e.g., probability of a given choice in discrimination experiment). The constraints consist of using no more than two independent observations for each subject. Estimators are provided forassessing the inter-subject "variance" of the analyzed probabilities; also, a test variable is given for testing the hypothesis that the average probability is the same for two groups of subjects.

DD FORM 1 JAN 73 1473

402 5??

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AFOSR-TR- 81-0207

On Estimating Inter-Subject Variability of Choice Probabilities under Observability Constraints

Robert Bartoszyński* and Madan L. Puri**

Institute of Mathematics, Polish Academy of Sciences, Warszawa, Poland and Indiana University, Bloomington, Indiana

The paper provides methods for estimating inter-subject variability of the probability of a given event defined in terms of subject's behavior (e.g. probability of a given choice in discrimination experiment). The constraints consist of using no more than two independent observations for each subject. Estimators are provided for assessing the inter-subject "variance" of the analyzed probabilities; also, a test variable is given for testing the hypothesis that the average probability is the same for two groups of subjects.

* Work supported by the National Science Foundation Grant No. MSC 76-00951.

** Work supported by the Air Force Office of Scientific Research, AFSC USAF, Grant No. AFOSR-76-2927.

81 3 27 77

1. Introduction

A practical problem in which the results of the present paper may be useful is as follows. The experimenter is interested in a certain event, say A, defined in terms of behavior of the subject (e.g., choice of one of the two displayed stimuli which are to be discriminated).

The experimenter needs the probability $P_s(A)$, where s denotes the subject. To estimate this probability one should make a series of independent observations on the same subject, a procedure often not feasible for reasons such as memory or learning effects, or simply subject's boredom.

In such cases, one often makes the assumption that $P_s(A)$ does not depend on s, and proceeds to estimate $P(A)$ using data for large groups of subjects, each tested once or twice.

The crucial issue in such a procedure is the inter-subject variability of $P_s(A)$. The present paper gives, among others, methods of assessing this variability, and testing the hypothesis of equality of average $P_s(A)$ in two groups of subjects, under the constraint that at most two observations are taken for each subject. In a sense this paper constitutes a statistical counterpart of an earlier paper written by one of the authors (see Bartoszyński (1978)).

2. The General Scheme

We shall consider the following situation. Let $G_n = \{G_1, \ldots, G_n\}$ be a system of independent experiments. Assume that each experiment may lead to a "success" or to a "failure", with (unknown) probability of success in experiment G_i equal p_i. Our problem will be to construct methods of inference about probabilities p_i in situations, when for some reasons one is allowed to make <u>at most two</u> independent observations of each experiment G_i.

Obviously, not much may be inferred about the individual values p_i; we shall therefore construct estimators of "moment-like" quantities

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i , \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} p_i^2 - \bar{p}^2 . \tag{1}$$

Let X_i be 1 or 0 depending on whether the first trial in experiment G_i leads to success or failure, and similarly, let Y_i be the random variable coding the outcome of the second trial on G_i.

We assume that different experiments and different trials on the same experiment are independent; moreover,

$$P(X_i = 1) = P(Y_i = 1) = p_i , \; i = 1,\ldots,n . \tag{2}$$

Denote

$$\xi_{ij} = X_i Y_j \tag{3}$$

and put

$$U = \frac{1}{n}\sum_{i=1}^{n} X_i , \; V = \frac{1}{n}\sum_{i=1}^{n} Y_i , \tag{4}$$

$$L = \tfrac{1}{2}(U + V) \ , \ Z = \frac{1}{n} \sum_{i=1}^{n} \xi_{ii} \ , \ W = Z - UV \quad . \tag{5}$$

The following proposition is an improvement of the similar proposition proved by Bartoszyński (1971):

PROPOSITION 1. <u>The random variables</u> L <u>and</u> W <u>are unbiased estimates of</u> $\bar{p}$ <u>and</u> σ^2 <u>respectively</u>. <u>Moreover</u>, $\operatorname{Var} L \le 1/8n$, $\operatorname{Var} W < 3/4n$ <u>and</u> $\operatorname{Var}(W|\sigma^2 = 0) \le 1/16n$.

<u>Proof</u>: By (2) , we have $EX_i = EY_i = p_i$. Consequently, $EU = EV = \frac{1}{n}\sum_{i=1}^{n} p_i = \bar{p}$, and also $EL = \bar{p}$. Since $E\xi_{ii} = EX_iY_i = EX_iEY_i = p_i^2$, we get, by independence of U and V

$$\begin{aligned} EW &= \frac{1}{n}\sum_{i=1}^{n} E\xi_{ii} - EUV = \frac{1}{n}\sum_{i=1}^{n} p_i^2 - EUEV = \frac{1}{n}\sum_{i=1}^{n} p_i^2 - \bar{p}^2 \\ &= \sigma^2 \end{aligned} \tag{6}$$

which proves the unbiasedness.

Next, $\operatorname{Var} X_i = \operatorname{Var} Y_i = p_i(1 - p_i) \le 1/4$, and using the assumption of independence, we can write $\operatorname{Var} U = \operatorname{Var} V = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var} X_i \le 1/4n$. Since U and V are independent, we have

$$\operatorname{Var} L = \tfrac{1}{4}(\operatorname{Var} U + \operatorname{Var} V) \le 1/8n \ . \tag{7}$$

It remains to evaluate the variance of W. We may write

$$\mathrm{Var}\, W = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \xi_{ii} - UV \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\, \xi_{ii} + \mathrm{Var}\, UV - \frac{2}{n} \mathrm{Cov}\left(\sum_{i=1}^{n} \xi_{ii} , UV \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\, \xi_{ii} + \frac{1}{n^4} \mathrm{Var} \sum_{i,j=1}^{n} \xi_{ij} - \frac{2}{n^3} \mathrm{Cov}\left(\sum_{i=1}^{n} \xi_{ii}, \sum_{k,m=1}^{n} \xi_{km} \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\, \xi_{ii} + \frac{1}{n^4} \sum_{i,j,k,m=1}^{n} \mathrm{Cov}\, (\xi_{ij}, \xi_{km})$$

$$- \frac{2}{n^3} \sum_{i,k,m=1}^{n} \mathrm{Cov}\, (\xi_{ii}, \xi_{km}) .$$

Taking into account the fact that $\mathrm{Cov}\, (\xi_{ij}, \xi_{km})$ is zero unless $i = k$ and/or $j = m$, we obtain after some transformations the formula

$$\mathrm{Var}\, W = \frac{(n-1)^2}{n^4} \sum_{i=1}^{n} \mathrm{Cov}\, (\xi_{ii}, \xi_{ii})$$

$$- \frac{4(n-1)}{n^4} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{Cov}\, (\xi_{ii} , \xi_{ij})$$

$$+ \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{Cov}\, (\xi_{ij}, \xi_{ij})$$

$$+ \frac{2}{n^4} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{\substack{k \neq i \\ k \neq j}} \mathrm{Cov}\, (\xi_{ij}, \xi_{ik}) . \tag{9}$$

We have, if different letters stand for different indices,

$$\mathrm{Cov}\ (\xi_{ii}, \xi_{ii}) = p_i^2(1 - p_i^2) , \tag{10}$$

$$\mathrm{Cov}\ (\xi_{ii}, \xi_{ij}) = p_j p_i^2(1 - p_i) , \tag{11}$$

$$\mathrm{Cov}\ (\xi_{ij}, \xi_{ij}) = p_i p_j(1 - p_i p_j) , \tag{12}$$

$$\mathrm{Cov}\ (\xi_{ij}, \xi_{ik}) = p_j p_k p_i (1 - p_i) . \tag{13}$$

Thus, all terms in (9) are nonnegative, and all are bounded from above by $1/4$. We may therefore write, omitting the second sum in (9)

$$\begin{aligned}\mathrm{Var}\ W &\le \frac{(n-1)^2}{n^4} \cdot \frac{n}{4} + \frac{1}{n^4} \cdot \frac{n(n-1)}{4} + \frac{2}{n^4} \cdot \frac{n(n-1)(n-2)}{4} \\ &\le \frac{1}{4n} + \frac{1}{4n^2} + \frac{1}{2n}\end{aligned} \tag{14}$$

which gives the asserted bound.

Finally, if $\sigma^2 = 0$, then all p_i are equal. Let their common value be p . Substituting (10) - (13) to (9), we obtain, after reduction

$$\mathrm{Var}\ (W|\sigma^2 = 0) = \frac{n-1}{n^2}\ p^2\ (1-p)^2 . \tag{15}$$

The maximum is attained at $p = \frac{1}{2}$ and is asymptotically equal to $1/16n$, which completes the proof.

Thus, the bound for the conditional variance $\mathrm{Var}(W|\sigma^2 = 0)$ is sharp. One may conjecture that the general bound for $\mathrm{Var}\ W$ can be improved, even considerably. However, simple numerical data show that the maximum under constraint $\sigma^2 = 0$ is not the overall maximum. Indeed, if $n = 6$ and $p_1 = \dots = p_6 = 0.5$ (hence $\sigma^2 = 0$), the variance of W equals $1/16n = 8.68 \cdot 10^{-3}$. If $p_1 = 0.45$, $p_2 = \dots = p_5 = 0.5$, $p_6 = 0.55$ (hence $\sigma^2 = 8.3 \cdot 10^{-4}$) variance of W is $8.69 \cdot 10^{-3}$. If $p_1 = p_2 = p_3 = 0.45$, $p_4 = p_5 = p_6 = 0.55$ (hence $\sigma^2 = 2.5 \cdot 10^{-3}$) variance of W is $8.71 \cdot 10^{-3}$. Similarly, if $p_1 = p_2 = p_3 = 0.4$, $p_4 = p_5 = p_6 = 0.6$ (hence $\sigma^2 = 0.01$), we have $\operatorname{Var} W = 8.8 \cdot 10^{-3}$, etc.

To make Proposition 1 useful for practical applications, we shall complement it by specifying conditions under which the distribution of W is asymptotically normal

For that purpose, we assume that the system G_n is a beginning of a potentially infinite sequence of independent experiments $G_1, G_2, \dots$. In short, we assume that n may be chosen arbitrarily large.

We shall prove

PROPOSITION 2. *If*

$$\sum_{i=1}^{\infty} p_i^2(1 - p_i^2) = \infty \tag{16}$$

then the random variable W *defined by* (5) *is asymptotically normal*.

Proof. Since $p_i^2(1 - p_i^2) \leq 2p_i(1 - p_i)$, we have also

$$\sum_{i=1}^{\infty} p_i(1 - p_i) = \infty \tag{17}$$

and it follows that all three random variables U, V and Z

defined by (4) and (5) are asymptotically normal (see e.g. Fisz 1961).

We shall prove that the joint distribution of (Z, U, V) is asymptotically normal. Consider a linear combination

$$aZ + bU + cV \tag{18}$$

which may be written as

$$\frac{1}{n} \sum_{i=1}^{n} [aX_i Y_i + bX_i + cY_i] = \frac{1}{n} \sum_{i=1}^{n} \eta_i \ . \tag{19}$$

Each of the random variables η_i assumes the values $0, b, c,$ and $a + b + c$ with probabilities respectively equal to $(1 - p_i)^2$, $p_i(1 - p_i)$, $p_i(1 - p_i)$ and p_i^2 . Consequently

$$E\eta_i = (a + b + c)p_i^2 + (b + c)p_i(1 - p_i) \tag{20}$$

and also, after some reductions,

$$\text{Var } \eta_i = a^2 p_i^2 (1 - p_i^2) + 2a(b + c)p_i^2(1 - p_i) + (b^2 + c^2)p_i(1 - p_i) \ . \tag{21}$$

Since the random variables η_i are bounded, it suffices to show that $\sum \text{Var } \eta_i$ diverges for any a, b, c which do not vanish simultaneously. If $a \neq 0$, this is true in view of (16), while if $a = 0$, we must have $b^2 + c^2 > 0$, and divergence is ensured by (17).

Using now theorem from Rao (1965), p. 319-340, we conclude that the random variable W is asymptotically normal, with mean σ^2 and variance bounded by $3/4n$.

3. Comparison of systems

Let us now consider two disjoint (hence independent) systems $G^{(1)}$ and $G^{(2)}$, consisting respectively of n_1 and n_2 experiments. Let $n = n_1 + n_2$ and let $G = G^{(1)} \cup G^{(2)}$ be the combined system.

In the sequel, we shall denote by $\sum^{(1)}$ and $\sum^{(2)}$ the summations extended over the indices in $G^{(1)}$ and $G^{(2)}$ respectively.

Let

$$\bar{p}^{(1)} = \frac{1}{n_i}\sum{}^{(1)} p_i \ , \ \bar{p}^{(2)} = \frac{1}{n_2}\sum{}^{(2)} p_i \ , \qquad (22)$$

$$\sigma_1^2 = \frac{1}{n_i}\sum{}^{(1)} (p_i - \bar{p}^{(1)})^2 \ , \ \sigma_2^2 = \frac{1}{n_2}\sum{}^{(2)} (p_i - \bar{p}^{(2)})^2 \ , \qquad (23)$$

and let $\bar{p}$ and σ^2 have the same meaning as before.

Under the constraints of the preceding section, namely that one can have at most two independent trials on each experiment, we shall construct a test for the hypothesis that $\bar{p}^{(1)} = \bar{p}^{(2)}$.

Let W, W_1 and W_2 be the random variables defined by (5) for systems G, $G^{(1)}$ and $G^{(2)}$.

We shall prove

PROPOSITION 3. <u>The random variable</u>

$$K = W - \frac{n_1}{n} W_1 - \frac{n_2}{n} W_2 \qquad (24)$$

<u>satisfies the condition</u> $EK \geq 0$, <u>with</u> $EK = 0$ <u>if, and only if</u>

$\bar{p}^{(1)} = \bar{p}^{(2)}$. Moreover, if $n \to \infty$, then for any n_1, n_2 we have $\operatorname{Var} K \le c_n \sim \frac{2n_1n_2}{n^3}$.

Proof. Let us write

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum(p_i - \bar{p})^2 \\
&= \frac{1}{n}\sum^{(1)}(p_i - \bar{p}^{(1)} + \bar{p}^{(1)} + \bar{p})^2 + \frac{1}{n}\sum^{(2)}(p_i - \bar{p}^{(2)} + \bar{p}^{(2)} - \bar{p})^2 \\
&= \frac{1}{n}\sum^{(1)}(p_i - \bar{p}^{(1)})^2 + \frac{1}{n}\sum^{(2)}(p_i - \bar{p}^{(2)})^2 \\
&\quad + \frac{n_1}{n}(\bar{p}^{(1)} - \bar{p})^2 + \frac{n_2}{n}(\bar{p}^{(2)} - \bar{p})^2 \\
&= \frac{n_1}{n}\sigma_1^2 + \frac{n_2}{n}\sigma_2^2 + \frac{n_1}{n}(\bar{p}^{(1)} - \bar{p})^2 + \frac{n_2}{n}(\bar{p}^{(2)} - \bar{p})^2 .
\end{aligned}
\tag{25}
$$

By Proposition 1 the random variables W , W_1 and W_2 are unbiased estimators of σ^2 , σ_1^2 and σ_2^2 respectively. Consequently,

$$EK = EW - \frac{n_1}{n}EW_1 - \frac{n_2}{n}EW_2 = \frac{n_1}{n}(\bar{p}^{(1)} - \bar{p})^2 + \frac{n_2}{n}(\bar{p}^{(2)} - \bar{p})^2 \ge 0 , \tag{26}$$

with equality holding if and only if $\bar{p}^{(1)} = \bar{p}^{(2)}$, since then $\bar{p}^{(1)} = \bar{p} = \bar{p}^{(2)}$.

To evaluate the variance of K , let us write

$$W = \frac{1}{n}\sum\xi_{ii} - UV = \frac{1}{n}\sum\xi_{ii} - \frac{1}{n^2}\sum\xi_{jk} , \tag{27}$$

and similar formulas hold for W_1 and W_2.

Using (27) we may write

$$K = \frac{1}{n}\left\{\frac{1}{n_1}\sum_{j,k}^{(1)}\xi_{jk} + \frac{1}{n_2}\sum_{j,k}^{(2)}\xi_{jk} - \frac{1}{n}\sum_{j,k}\xi_{jk}\right\} . \tag{28}$$

Consequently, we obtain

$$\begin{aligned} n^2 \operatorname{Var} K = {} & \frac{1}{n_1^2}\operatorname{Cov}\left(\sum^{(1)}\xi_{jk}, \sum^{(1)}\xi_{jk}\right) \\ & + \frac{1}{n_2^2}\operatorname{Cov}\left(\sum^{(2)}\xi_{jk}, \sum^{(2)}\xi_{jk}\right) + \frac{1}{n^2}\operatorname{Cov}\left(\sum\xi_{jk}, \sum\xi_{jk}\right) \\ & - \frac{2}{nn_1}\operatorname{Cov}\left(\sum^{(1)}\xi_{jk}, \sum\xi_{jk}\right) - \frac{2}{nn_2}\operatorname{Cov}\left(\sum^{(2)}\xi_{jk}, \sum\xi_{jk}\right) , \end{aligned} \tag{29}$$

since $\sum^{(1)}\xi_{jk}$ and $\sum^{(2)}\xi_{jk}$ are independent.

By taking into account only those covariances $\operatorname{Cov}(\xi_{jk}, \xi_{rm})$ which are not zero, we arrive, after considerable algebra, at the formula

$$\begin{aligned} n^2 \operatorname{Var} K = {} & \left(\frac{n_2}{nn_1}\right)^2 \sum_i^{(1)} p_i^2(1 - p_i^2) + \left(\frac{n_1}{nn_2}\right)^2 \sum_i^{(2)} p_i^2(1 - p_i^2) \\ & + 4\left(\frac{n_2}{nn_1}\right)^2 \sum_i^{(1)} \sum_{j\neq i}^{(1)} p_j p_i^2(1 - p_i) \\ & + 4\left(\frac{n_1}{nn_2}\right)^2 \sum_i^{(2)} \sum_{j\neq i}^{(2)} p_j p_i^2(1 - p_i) \\ & - \frac{4n_2}{n^2 n_1} \sum_i^{(1)} \sum_j^{(2)} p_j p_i^2(1 - p_i) - \end{aligned} \tag{30}$$

(cont.)

$$- \frac{4n_1}{n^2 n_2} \sum_i^{(2)} \sum_j^{(1)} p_j p_i^2 (1 - p_i)$$

$$+ \left(\frac{n_2}{nn_1}\right)^2 \sum_i^{(1)} \sum_{j \neq i}^{(1)} p_i p_j (1 - p_i p_j)$$

$$+ \left(\frac{n_1}{nn_2}\right)^2 \sum_i^{(2)} \sum_{j \neq i}^{(2)} p_i p_j (1 - p_i p_j)$$

$$+ \frac{1}{n^2} \sum_i^{(1)} \sum_j^{(2)} p_i p_j (1 - p_i p_j) + \frac{1}{n^2} \sum_i^{(2)} \sum_j^{(1)} p_i p_j (1 - p_i p_j)$$

$$+ 2\left(\frac{n_2}{nn_1}\right)^2 \sum_i^{(1)} \sum_{j \neq i}^{(1)} \sum_{\substack{k \neq i \\ k \neq j}}^{(1)} p_j p_k p_i (1 - p_i)$$

$$+ 2\left(\frac{n_1}{nn_2}\right)^2 \sum_i^{(2)} \sum_{j \neq i}^{(2)} \sum_{\substack{k \neq i \\ k \neq j}}^{(2)} p_j p_k p_i (1 - p_i)$$

$$+ \frac{2}{n^2} \Big\{ \sum_i \sum_{j \neq i} \sum_{\substack{k \neq i \\ k \neq j}} p_j p_k p_i (1 - p_i)$$

$$- \sum_i^{(1)} \sum_{j \neq i}^{(1)} \sum_{\substack{k \neq i \\ k \neq j}}^{(1)} p_j p_k p_i (1 - p_i) - \sum_i^{(2)} \sum_{j \neq i}^{(2)} \sum_{\substack{k \neq i \\ k \neq j}}^{(2)} p_j p_k p_i (1 - p_i) \Big\} .$$

Using now the estimates $p_i^2(1 - p_i^2) \le \frac{1}{4}$, $p_j p_i^2 (1 - p_i) \le 4/27$, $p_i p_j (1 - p_i p_j) \le \frac{1}{4}$, and $p_j p_k p_i (1 - p_i) \le \frac{1}{4}$, we may write, omitting negative terms in (30):

$$n^2 \text{ Var } K \le \frac{n_2^2}{4n^2 n_1} + \frac{n_1^2}{4n^2 n_2}$$

$$+ \; 4 \cdot \frac{4}{27} \left(\frac{n_2}{nn_1}\right)^2 n_1(n_1 - 1) + 4 \cdot \frac{4}{27} \left(\frac{n_1}{nn_2}\right)^2 n_2(n_2 - 1)$$

$$+ \; \frac{1}{4} \left(\frac{n_2}{nn_1}\right)^2 n_1(n_1 - 1) + \frac{1}{4}\left(\frac{n_1}{nn_2}\right)^2 n_2(n_2 - 1)$$

$$+ \; \frac{n_1 n_2}{4n^2} + \frac{n_2 n_1}{4n^2} + 2 \cdot \frac{1}{4} \left(\frac{n_2}{nn_1}\right)^2 n_1(n_1 - 1)(n_1 - 2)$$

$$+ \; 2 \cdot \frac{1}{4} \left(\frac{n_1}{nn_2}\right)^2 n_2(n_2 - 1)(n_2 - 2)$$

$$+ \; \frac{1}{4} \cdot \frac{2}{n^2} [3n_1(n_1 - 1)n_2 + 3n_2(n_2 - 1)n_1] \quad . \tag{31}$$

Denoting $n_1/n = x$, $n_2/n = 1 - x$, we obtain finally from (31):

$$\text{Var } K \sim \frac{1}{n} \cdot 2x(1 - x) + \frac{1}{n^2} \left\{\frac{1}{4} + \frac{16}{27} [x^2 + (1 - x)^2] \right\}$$

$$+ \; \frac{1}{n^3} \left\{ \frac{(1 - x)^2}{4x} + \frac{x^2}{4(1 - x)} \right\} ,$$

which yields the asserted asymptotic bound $2n_1n_2/n^3$, valid for all sufficiently large n .

Maximizing with respect to x we obtain

COROLLARY. As $n \to \infty$, we have regardless of n_1 and n_2

$$\mathrm{Var}\ K \le \alpha_n \sim 1/2n . \tag{32}$$

Similarly as in the preceding section, we have

PROPOSITION 4 . Under condition (16), the asymptotic (as $n \to \infty$) distribution of K is normal.

The proof is analogous to the proof of Proposition 2 and will be omitted.

Propositions 3 and 4 provide means of testing the hypothesis that $\bar{p}^{(1)} = \bar{p}^{(2)}$: one may take the random variable K as the test variable, and use one-sided critical region, large values suggesting rejection.

4. Applications

A typical application, for assessing the inter-subject variability of choice probabilities, has already been described in the Introduction. Somewhat more generally, one may consider the following situation.

Suppose that the subjects are trained to perform some classification task, e.g., classify some objects into binary categories. Taking one of these categories as "success", and classifying the same set of objects twice by the same group of subjects, one obtains a variety of possibilities of applying the results of this paper. Let p_{ij} be the probability that i-th subject will classify j-th object to the chosen category, and let X_{ij} , Y_{ij} be 0 or 1 depending on how the j-th object was classified by the i-th subject on the first and second trial.

Fixing i , and taking a large set of objects, one can estimate the variability of p_{ij} under changes of j ; if this variability is large, one may want to identify objects which are easier, or more difficult, to classify.

On the other hand, fixing j , and varying i , one can test whether there exists sufficient homogeneity among classifying subjects; one can also test the hypothesis, using Propositions 3 and 4, that the two groups of subjects are equally well trained, etc.

If there are more than two categories in the classification, one can apply the same technique by fixing one category. However, if one can ensure not two, but four independent classifications for each person, then one can define "success" as the event "identical classifications on two successive trials".

Letting p_{ijk} , $k = 1,2,...$ denote the probability that i-th subject will classify j-th object into k-th category, the probability of success in this case is

$$\sum_k p_{ijk}^2 = v_{ij} .$$

Again, one can test variability of v_{ij} under changes of subjects and objects.

Finally, observe that the above ways of testing the quality of classification by human classifiers is independent of the concept of "true category" of the object, and applies equally well to classification schemes in which such concept makes no sense (e.g. grading students' papers, etc.).

REFERENCES

BARTOSZYŃSKI, R. On the construction and evaluation of subjective classifications. Applications Mathematicae, 1971, XII, 1-21.

BARTOSZYŃSKI, R. On linear classifications under varying choice probabilities. Journal of Mathematical Psychology, 1978, 18, 249-259.

FISZ, M. Probability Theory and Mathematical Statistics. New York: Wiley, 1961.

RAO, C.R. Linear Statistical Inference and Its Applications. New York: Wiley, 1965.

LMED

8