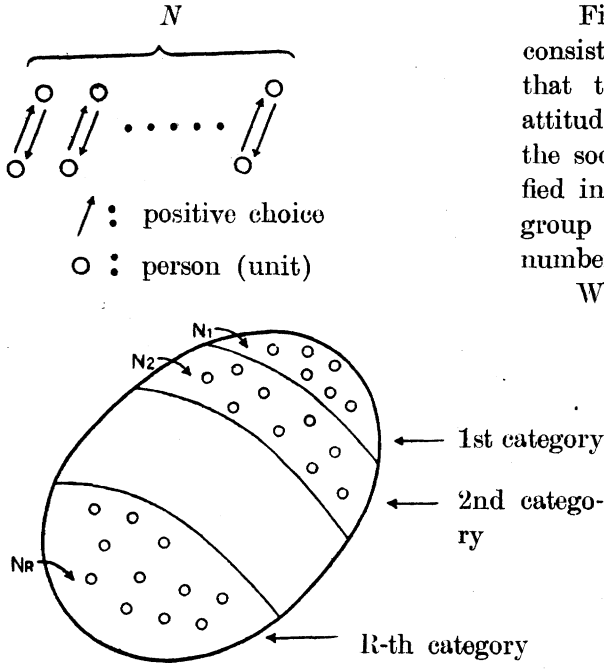
# On a Matching Problem

Chikio HAYASHI and Hirotugu AKAIKE

(Received January 20, 1953)

## § 1. Introduction

The problem we treat in this paper arose when the sociometric data were analysed. The data obtained by the so-called sociometry technique express the human relation and the method of analysis is not simple. But we think it is important to study the method of analysing sociometric data from the mathematico-statistical point of view. In the following the problem will be considered whether the pair who express positive (negative) choices in the sociometric data have the same (different) behavior or not. That is to say, the statistical problem as to the relation between human relations and the patterns of their behaviour will be discussed. Considering the group, the elements of which are pairs of size $N$, for example expressing mutual positive choice, we shall treat it as a matching problem.




First, notice that the pairs consist of two units and, suppose that their behaviour (opinion or attitude) has been measured before the sociometric test and are classified into several categories. In the group mentioned above, the total number of units is $2N$.

When $N_i$ denotes the number of units who are classified into the $i$-th category in opinion, and $R$ denotes the number of categories, we have

$$\sum_{i=1}^{R} N_i = 2N.$$

Now we construct the population by giving the same sampling probability to each unit. From this

population we draw two units and record whether they have the same opinion (atitude) or not, and define the random variable $X_1$, such that

$X_1 = 1$ ; when they have the same opinion, i.e. are classified into the same category in opinion

$X_1 = 0$ ; otherwise

Then we have

$$P_r\{X_1 = 1\} = \sum_{i=1}^{R} \frac{N_i(N_i - 1)}{2N(2N - 1)}$$

$$P_r\{X_1 = 0\} = 1 - P_r\{X_1 = 1\}$$

In the next step, we draw two units without replacement, and record whether they have the same opinion (attitude) or not, and define the random variable $X_2$, such that

$X_2 = 1$ ; when they have the same opinion

$X_2 = 0$ ; otherwise

Then

$$P_r\{X_2 = 1\} = \sum_{i=1}^{R} \frac{N_i(N_i - 1)}{2N(2N - 1)}$$

$$P_r\{X_2 = 0\} = 1 - P_r\{X_2 = 1\}$$

Similarily, we difine $X_3, X_4, X_5, \ldots, X_N$. $X_i$ is obviously correlated with $X_1, X_2, \ldots, X_N$. In the last step, i.e. the $N$-th step, only two units are remained. Wen ow put

$$w = \sum_{i=1}^{N} X_i$$

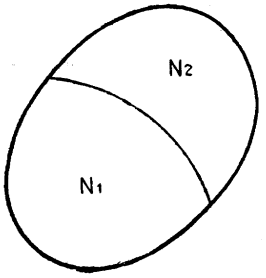
which is the random variable representing the number of pairs having the same opinion in the population above mentioend. We derive its distribution function.

Once the distribution function (consequently the moments) of $w$ is known, the test is carried out whether the number of pairs having the same opinion which are obtained in the survey is significant, i.e. whether the relation between human relations and behaviour is significant, or not.

## § 2. Distribution function and moments of $w$

(i) The case where $R = 2$, i.e. the opinion is measured by dichotomous category, yes or no.

Let $N$ be the number of pairs, expressing a sort of relation in sociometric data, for example, mutual positive choice, positive choice, mutual




negative choice etc., and let $2N$ be the number of units. Then

$$2N = N_1 + N_2$$

$$w = \sum_{i=1}^{N} X_i$$

Now we calculate the moments of $w$ by taking account of covariance factors.

$$\overline{w} = E(w) = N\left(1 - \frac{2N_1N_2}{2N(2N-1)}\right) \equiv NP$$

$$\sigma_w^2 = E(w - E(w))^2 = NP(1-NP) +$$

$$N(N-1) \times \frac{N_1(N_1-1)(N_1-2)(N_1-3)+N_2(N_2-1)(N_2-2)(N_2-3)+2N_1N_2(N_1-1)(N_2-1)}{2N(2N-1)(2N-2)(2N-3)}$$

$$\mu_3 = E(w-E(w))^3 = \frac{N_2(N_2-1)\,(N_2-2)(N_2-3)(N_2-4)(N_2-5)}{(2N-1)(2N-3)(2N-5)}$$

$$+ 3\left(2 - \frac{N_2(N_2-1)}{2N-1}\right)\frac{N_2(N_2-1)(N_2-2)(N_2-3)}{(2N-1)(2N-3)}$$

$$+ \left(4 - 6\frac{N_2(N_2-1)}{2N-1} + 3\frac{N_2^2(N_2-1)^2}{(2N-1)^2}\right)\frac{N_2(N_2-1)}{2N-1} - \frac{N_2^3(N_2-1)^3}{(2N-1)^3}$$

$$\mu_4 = E(w - E(w))^4$$

$$= \frac{N_2(N_2-1)(N_2-2)(N_2-3)(N_2-4)(N_2-5)(N_2-6)(N_2-7)}{(2N-1)(2N-3)(2N-5)(2N-7)}$$

$$+ 4\cdot\left(3 - \frac{N_2(N_2-1)}{2N-1}\right)\frac{N_2(N_2-1)(N_2-2)(N_2-3)(N_2-4)(N_2-5)}{(2N-1)(2N-3)(2N-5)}$$

$$+ 2\cdot\left(14 - 12\frac{N_2(N_2-1)}{2N-1} + 3\frac{N_2^2(N_2-1)^2}{(2N-1)^2}\right)$$

$$\times \frac{N_2(N_2-1)(N_2-2)(N_2-3)}{(2N-1)(2N-3)} + 4\cdot\left(2 - 4\frac{N_2(N_2-1)}{2N-1}\right.$$

$$\left. + 3\frac{N_2^2(N_2-1)^2}{(2N-1)^2} - \frac{N_2^3(N_2-1)^3}{(2N-1)^3}\right)\frac{N_2(N_2-1)}{2N-1} + \frac{N_2^4(N_2-1)^4}{(2N-1)^4}$$

The exact distribution of $w$ is given by

$$P_r(w=X) = \frac{N!\,N_1!\,N_2!}{(2N)!} \cdot \frac{2^{N-X}}{\left(\frac{X-m}{2}\right)!\left(\frac{X+m}{2}\right)!\,(N-X)!}$$

where

$N_1$; number of units with the opinion of the first category
$N_2$; number of units with the opinion of the second category

$$2N = N_1 + N_2$$

$$m = \left|\frac{N_1 - N_2}{2}\right|$$

$$X = m,\ m+2,\ m+4,\ \ldots,\ N-\delta\ ;\quad \delta = \begin{cases} 1 & \text{when } N - m \equiv 1 \pmod 2 \\ 0 & N - m \equiv 0 \pmod 2 \end{cases}$$

Now putting

$n_1$ = number of pairs with the same opinion of the first category
$n_2$ = number of pairs with the same opinion of the second category
$n_3$ = number of pairs with the different opinions,

we have

$$N_1 = 2n_1 + n_3$$
$$N_2 = 2n_2 + n_3$$
$$w = n_1 + n_2 = 2n_2 + \frac{N_1 - N_2}{2}.$$

Using the probability distribution given above, we have

$$E(n_2(n_2-1)\ldots(n_2-i+1)) = \frac{N_2(N_2-1)(N_2-2)\ldots(N_2-2i+1)}{2^i\cdot(2N-1)(2N-3)\ldots(2N-2i+1)}$$

These relations give also the moments too. Putting $Nr$ instead of $N_2$ in these formulae, where

$$Nr = N_2 \quad \text{or} \quad r = \frac{N_2}{N} \quad (0 \leqq r \leqq 2),$$

we have

$$\bar{w} = N\left(1 - \frac{r(2-r)}{2}\right) + O(1)$$

$$\sigma_w^2 = N\cdot\frac{r^2(2-r)^2}{4} + O(1)$$

when $N\to\infty$.

The approximate distribution of $w$, when $N$ is large, is given by using the Stirling's approximation formula.

$$\log P(w - \bar{w} = y) = \log\left(\frac{1}{\sqrt{2\pi}\left(\frac{r(2-r)}{4}\cdot\sqrt{N}\right)}\right)$$

$$-\frac{1}{2}\frac{1}{\left(\frac{r(2-r)}{4}\sqrt{N}\right)^2}\left\{\left(\frac{y+(1-r)^2}{2}\right)^2-\frac{(1-r)^4}{4}\right\}+Q(yN)$$

$$Q(yN)=\frac{1}{N}\left\{\frac{11}{12}-\frac{1}{r(2-r)}-\frac{1}{3}\left(\frac{1}{r^2}+\frac{1}{(2-r)^2}\right)-\frac{1}{r^2(2-r)^2}\cdot\frac{y^2}{N}\right.$$
$$\left.+\frac{16}{3}\cdot\frac{(1-r)^2}{r^4(2-r)^4}\cdot\frac{y^3}{N}-\frac{1}{3}\left(\frac{r^3+(2-r)^3}{r^3(2-r)^3}\right)^2\frac{y^4}{N^2}\right\}+O\left(\frac{1}{N^{3/2}}\right)$$

Let $Z$ be a continuous random variable with the probability density function $f(z)$ defined as follows.

$$\log f(z)=\log P_r\left\{\frac{w-\bar{w}+(1-r)^2}{2}\in[z-\frac{1}{2},z+\frac{1}{2})\right\}$$

Then, putting

$$\frac{w-\bar{w}+(1-r)^2}{2}=\zeta$$

and $\zeta_z$=the value of $\zeta$ which belongs to $[z-\frac{1}{2},z+\frac{1}{2})$, we have

$$\log f(z)=\log P_r\{\zeta=\zeta_z\}$$
$$=\log\frac{1}{\sqrt{2\pi}\left(\frac{r(2-r)}{4}\sqrt{N}\right)}-\frac{1}{2}\frac{1}{\left(\frac{r(2-r)}{4}\sqrt{N}\right)^2}\left\{\zeta_z^2-\frac{(1-r)^4}{4}\right\}$$
$$+Q(2\zeta_z-(1-r)^2,N).$$

By this formula we can see that when $r=1$ or $N_1=N_2$, $\frac{Z}{\sqrt{N}}$ is approximately normally distributed with mean zero and variance $(1/4)^2$, or $w$ is approximately normally distributed with mean $\bar{w}$ and variance $\left(\frac{r(2-r)}{2}\right)^2N$.

Evaluation of $Q(y,N)$ for $y=2\cdot\left(\frac{r(2-r)}{2}\right)\cdot\sqrt{N}$ shows that in the test of significance of $w$ at significance level 5% it will be reasonable to treat $w$ as if it were a random variable obeying the Gaussian distribution with mean $\bar{w}$ and variance $\sigma_w^2$ for $N\geqq 80$ and $r=1$. Generally for any $N$, when $r\fallingdotseq 1$, we shall be able to put much more confidence in the confidence interval of the type $[\bar{w}-k\sigma_w,\bar{w}+k\sigma_w]$ than that assured by the Tchebycheff's inequality. We will tabulate a part of the distribution of $w$ for the case where $r=1$ and $N=N_1=N_2=40$ with the approximate value by normal curve.

| $X$ | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_r(w = X)$ | 0.002 | 0.013 | 0.050 | 0.126 | 0.215 | 0.248 | 0.195 | 0.104 | 0.037 | 0.009 | 0.001 |
| Approximate value | 0.002 | 0.013 | 0.048 | 0.124 | 0.217 | 0.252 | 0.196 | 0.102 | 0.036 | 0.008 | 0.001 |

Generally when $N_1 \neq N_2$, $w$ is not regarded as tending to Gaussian distribution with $N \to \infty$.

(ii) The case where $R > 2$.

Calculating the moments of $w$, we have

$$E(w) = N\left(1 - \frac{\sum_{j\neq k}^{R}\sum N_j N_k}{2N(2N-1)}\right) \equiv NP$$

$$\sigma_w^2 = E(w - E(w))^2$$

$$= NP(1-NP) + N(N-1)\left\{\sum_{i=1}^{R}\frac{N_i(N_i-1)(N_i-2)(N_i-3)}{2N(2N-1)(2N-2)(2N-3)}\right.$$

$$\left. + \sum_{j\neq k}^{R}\sum\frac{N_j(N_j-1)N_k(N_k-1)}{2N(2N-1)(2N-2)(2N-3)}\right\}$$

By means of these formula, for the significance test of $w$, we can use the followings. But we must take the corresponding critical regions to the alternative hypotheses and interpret the results.

(a) the relation

$$P_r\{|w - E(w)| \geqq k\sigma_w\} \leqq \frac{1}{k^2},$$

for $N$ small and large,

(b) the relation, as an example the case where the two sided critical region is taken,

$$P_r\{|t| \geqq t_0\} = 1 - \frac{1}{\sqrt{2\pi}}\int_{-t_0}^{t_0} e^{-\frac{t^2}{2}}dt = \alpha,$$

where $\frac{w - E(w)}{\sigma_w} = t$, and $\alpha$ is the significance level, when $N_1 = N_2$

and $N$ is large,

and

(c) the exact distribution when $N$ is small for dichotomous case.

We tabulate below the 5% points $\overline{w}_{0.05}$, $\underline{w}_{0.05}$ of the distribution function of $w$ for some cases of dichotomous type, where

$$\overline{w}_\alpha \equiv \min (X;\ P_r(w \geq X) \leq \alpha)$$
$$\underline{w}_\alpha \equiv \max (X:\ P_r(w \leq X) \leq \alpha).$$

Case 1, $N_1 = N_2 = N$

| $N$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\overline{w}_{0.05}$ | 10 | 16 | 20 | 26 | 32 | 38 | 44 | 50 | 54 | 60 |
| $\overset{*}{w}_{0.05}$ | 7.3 | 13.4 | 19.2 | 24.9 | 30.6 | 36.1 | 41.6 | 47.1 | 52.5 | 58.0 |
| $\underline{w}_{0.05}$ | 0 | 4 | 8 | 12 | 16 | 22 | 26 | 30 | 34 | 40 |
| $\underset{*}{w}_{0.05}$ | 2.1 | 6.0 | 10.2 | 14.5 | 18.8 | 23.3 | 27.8 | 32.3 | 36.9 | 41.4 |

; $\overset{*}{w}_{0.05} = E(w) + 1.65 \times \dfrac{\sqrt{N}}{2}$

$\underset{*}{w}_{0.05} = E(w) - 1.65 \times \dfrac{\sqrt{N}}{2}$

$\dfrac{1}{\sqrt{2\pi}} \int_0^{1.65} e^{-\frac{t^2}{2}} dt \fallingdotseq 0.45$

These results show that the approximation by Gaussian distribution is of practical use in this case for even small $N$'s.

Case 2, $N_1 = 2N_2$

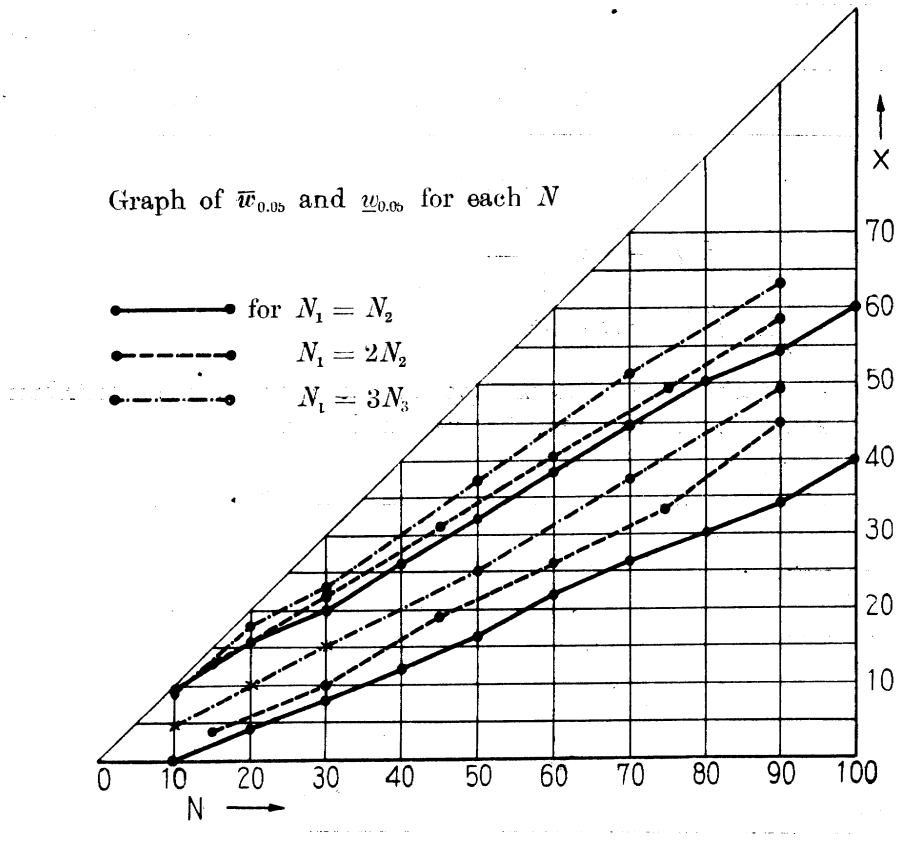
| $N$ | 15 | 30 | 45 | 60 | 75 | 90 |
|---|---|---|---|---|---|---|
| $N_1$ | 20 | 40 | 60 | 80 | 100 | 120 |
| $N_2$ | 10 | 20 | 30 | 40 | 50 | 60 |
| $\overline{w}_{0.05}$ | 13 | 22 | 31 | 40 | 49 | 58 |
| $\underline{w}_{0.05}$ | 4 | 10 | 19 | 26 | 33 | 44 |

Case 3, $N_1 = 3N_2$

| $N$ | 10 | 20 | 30 | 50 | 70 | 90 |
|---|---|---|---|---|---|---|
| $N_1$ | 15 | 30 | 45 | 75 | 105 | 135 |
| $N_2$ | 5 | 10 | 15 | 25 | 35 | 45 |
| $\bar{w}_{0.05}$ | 9 | 18 | 23 | 37 | 51 | 63 |
| $\underline{w}_{0.05}$ | * | * | * | 25 | 37 | 49 |

; * shows that $P(w = m) \geq 0.05$.

In the above, $\bar{w}_{0.01}$ and $\underline{w}_{0.01}$ are in most cases equal to $\bar{w}_{0.05}+2$ and $\underline{w}_{0.05}-2$ respectively.



## § 3. Examples

In this section we shall refer, as an example, to the results of attitude survey towards American culture and French culture in a group, which was recently performed. Two groups, one which consisted of pairs of units in relation of mutual like ($0 \rightleftarrows 0$) and the other which consisted of pairs of units in relation of like-neutral ($0 \rightarrow 0$) were obtained. They recieved the attitude test before the sociometric test. They did not know the attitudes of others. Their opinions were classified into two categories in the first case where favourable and unfavourable scale was used, and into four categories in the second case where the grades of scale were used. The results are as follows.

1-st case.

| Sociometric relation \ Categories of opinion | 1 ($N_1$) | 2 ($N_2$) | Total ($2N$) | $w$ (realized) | $\bar{w}$ | $\sigma_w^2$ |
|---|---|---|---|---|---|---|
| mutual like ($0 \rightleftarrows 0$) | 48 | 68 | 116 | 24 | 29.6 | 14.19 (14.49) |
| like neutral ($0 \rightarrow 0$) | 161 | 241 | 402 | 96 | 104.2 | 46.32 (50.18) |

; ( ) of the last column means the value of $\sigma_w^2$ approximated by $NP(1-P)$.

Now for mutual-like group we have

$$|w - \bar{w}| = 1.48\sigma_w \quad ,$$

and for like-neutral group we have

$$|w - \bar{w}| = 1.20\sigma_w \quad .$$

From the exact distribution we have for mutual-like group

$$\underline{w}_{0.05} = 22,$$

$$\underline{w}_{0.01} = 20.$$

In this case we can not find the significant relation between sociometric status and opinion.

2-nd case.

| Sociometric relation \ Categories of opinion | 1 | 2 | 3 | 4 | Total | $w$ (realized) | $\bar{w}$ | $\sigma_w^2$ |
|---|---|---|---|---|---|---|---|---|
| mutual like ($0 \rightleftarrows 0$) | 9 | 39 | 52 | 16 | 116 | 13 | 19.3 | 11.05 (12.89) |
| like nutral ($0 \rightarrow 0$) | 21 | 140 | 185 | 56 | 402 | 66 | 71.1 | 39.41 (45.94) |

; ( ) of the last column means the value of $\sigma_w^2$ approximated by $NP(1-P)$

We have $|w-\bar{w}| < 2\sigma_w$. In this case too, we can not find the significant relation.

*Institute of Statistical Mathematics*

## REFERENCE

C. Hayashi: On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, vol III No. 2, 1952