RESPONSE ERRORS AND BIASED INFORMATION

CHIKIO HAYASHI

(Received Jan. 31, 1968)

0. Introduction

In the present paper, we shall discuss some problems concerning the bias in estimating a variable Y from a variable X, both of which are subject to errors or fluctuations in measurement and are expressed as random variables, in the case where biased information originates in disregarding the response errors or response fluctuations. Let X_i and Y_i be the measurements on the ith object, $i=1, 2, \cdots, n$, n being the size of sample or universe. X's and Y's may be quantitative or qualitative. We shall give, below, a method of evaluating the estimation bias of Y's by fixed X's which are known realized values of the random variable X and a method of correcting the distortion in the statistical analysis of cross-tabulated data.

These situations arise when we treat the relation between two variables, for example, the change between time t_1 and time t_2 from a cross tabulation (correlated pattern) of them, or the before-after analysis in a usual follow-up study. These ideas will be useful for us in appraising the validity of quantitative representation of our data.

I. QUALITATIVE CASE

Here we treat the case where the variables are qualitative, i.e. represented by item-category-response and the response error or fluctuation is represented by a probabilistic model.

1. The simplest case

First we treat the simplest case where variable Y is qualitative and subject to error, and we require an unbiased estimate without any reference to X. This is an introduction to the correlated case. An idea similar to that in this section is found in [1], [13], although I have already introduced my idea in [4], [7] and developed it along this line in [8], [9], [10].

The response categories are assumed to be dichotomous (+, −). The response probabilities are shown in Table 1.

Table 1. Response-probability

true \ response	+	−
+	p	$1-p$
−	$1-q$	q

$1-p$ and $1-q$ represent the response error probabilities.

Let n_+ be the number of true + responses, n_- be the number of true − responses, where $n_+ + n_- = n$, the total number of responses. Here we assume that true responses + and − exist. We may call it a structure. Let m_+ be the observed number of response +, and m_- be the observed number of response −, where $n = m_+ + m_-$. We must infer the true response pattern (n_+, n_-) from (m_+, m_-), because our aim is not to know the apparent response pattern (m_+, m_-), which does not give us any valid information concerning the true response pattern as it is [12]. p and q are assumed to be known. In this case, the estimates $\hat{n}_+$ and $\hat{n}_-$ of n_+ and n_- are given by

$$\begin{pmatrix} \hat{n}_+ \\ \hat{n}_- \end{pmatrix} = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}'^{-1} \begin{pmatrix} m_+ \\ m_- \end{pmatrix}$$

where ( )′ means transposed matrix, under the condition of the existence of inverse matrix. $\hat{n}_+$ and $\hat{n}_-$ are unbiased estimates of n_+ and n_-. The variances of $\hat{n}_+$ and $\hat{n}_-$ are easily calculated. For, example,

$$\sigma^2_{\hat{n}_+} = L^2\{n_+ p(1-p) + n_- q(1-q)\},$$

where

$$L = \frac{1}{p+q-1},$$

and of course $\sigma^2_{\hat{n}_+} = \sigma^2_{\hat{n}_-}$.

In the case where the number of response categories is k ($k \geqq 3$), we can also give a similar solution.

Let P be the response-probability matrix, an element of which is p_{ij}, $i=1, 2, \cdots, R$, $j=1, 2, \cdots, R$, R being the number of categories in an item, and $\sum_{j=1}^{R} p_{ij} = 1$ for any i. $\mathfrak{N}$ is a column vector of the numbers of true responses to the categories in an item, an element of which is n_i, where $\sum_{i=1}^{R} n_i = n$, n being the total number of responses. $\mathfrak{M}$ is a column vector of the observed numbers of responses to the categories in an item, an element of which is m_i where $\sum_{i=1}^{R} m_i = n$. Let $\hat{\mathfrak{N}}$ be an estimate of $\mathfrak{N}$. According to the same reduction, we have an unbiased estimate $\hat{\mathfrak{N}}$ of $\mathfrak{N}$ as follows,

$$\hat{\mathfrak{N}} = P'^{-1}\mathfrak{M}\,,$$

when the inverse matrix exists.

The variance-covariance matrix $\sigma(\hat{\mathfrak{N}})$ is calculated from the variance-covariance matrix $\sigma(\mathfrak{M})$. Here an element of $\sigma(\hat{\mathfrak{N}})$ is $\sigma_{ij}(\hat{\mathfrak{N}})$, $i=1, 2, \cdots, R$, $j=1, 2, \cdots, R$ and $\sigma_{ii}(\hat{\mathfrak{N}})$ is the variance of the ith element of $\hat{\mathfrak{N}}$ and an element of $\sigma(\mathfrak{M})$ is $\sigma_{kl}(\mathfrak{M})$, $k=1, 2, \cdots, R$, $l=1, 2, \cdots, R$ and these are calculated from the equations $\sum_{i}^{R} m_{ij} = m_j$ for $j=1, 2, \cdots, R$, where m_{ij} is the random variable of the number of those who belong to the ith category in "true response" but respond to the jth category in actual response, which occurs with probability p_{ij}. Hence, $(m_{i1}, m_{i2}, \cdots, m_{iR})$ is subject to a multinomial distribution with parameters $(p_{i1}, \cdots, p_{iR})$, and m_{ij} and $m_{i'j'}$ are independent for any i, j, i', j' except $i=i'$. Thus $\sigma_{kl}(\mathfrak{M}) = -\sum_{j}^{R} n_j p_{jk} p_{jl}$ for $l \neq k$, and $\sigma_{ll}(\mathfrak{M}) = \sum_{j}^{R} n_j p_{jl}(1-p_{jl})$. Thus we have,

$$\sigma(\hat{\mathfrak{N}}) = P'^{-1}\sigma(\mathfrak{M})(P^{-1})\,.$$

2. Estimation of response probability

If p and q are unknown, we can obtain the response probabilities by a test-retest method. This is similar to the estimation of the parameters in latent structure analysis (for example, [2]). In this case, we use the following model.

The number of those belonging to true response $+$ is n_+, the num-

Table 2.

true \ response	$+$	$\pm$	$-$	Total
$+$	p_{++}	$p_{+\pm}$	p_{+-}	1
$\pm$	$p_{\pm+}$	$p_{\pm\pm}$	$p_{\pm-}$	1
$-$	p_{-+}	$p_{-\pm}$	p_{-+}	1

ber of true $\pm$ responses is $n_{\pm}$, and the number of true $-$ responses is n_{-}, where $n_{+}+n_{\pm}+n_{-}=n$. The numbers of response pairs $++$, $+\pm$, $\cdots$ etc. obtained by test-retest are m_{ij} $i=+, \pm, -$, $j=+, \pm, -$. The equations which hold in the mean (expectation) are:

$$\left\{\left(\begin{array}{c|c|c} P & 0 & 0 \\ \hline 0 & P & 0 \\ \hline 0 & 0 & P \end{array}\right)\left(\begin{array}{c|c|c} p_{++}I & p_{+\pm}I & p_{+-}I \\ \hline p_{\pm+}I & p_{\pm\pm}I & p_{\pm-}I \\ \hline p_{-+}I & p_{-\pm}I & p_{--}I \end{array}\right)\right\}' \begin{pmatrix} n_{+} \\ 0 \\ 0 \\ 0 \\ n_{\pm} \\ 0 \\ 0 \\ 0 \\ n_{-} \end{pmatrix} = \begin{pmatrix} m_{++} \\ m_{+\pm} \\ m_{+-} \\ m_{\pm+} \\ m_{\pm\pm} \\ m_{\pm-} \\ m_{-+} \\ m_{-\pm} \\ m_{--} \end{pmatrix} \qquad \ldots.(A)$$

where

$$P=\begin{pmatrix} p_{++} & p_{+\pm} & p_{+-} \\ p_{\pm+} & p_{\pm\pm} & p_{\pm-} \\ p_{-+} & p_{-\pm} & p_{--} \end{pmatrix}, \quad I=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad 0=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and $(\ \ )'$ means a transposed matrix.

If the m_{ij}'s are known by a test-retest, where $\sum_i \sum_j m_{ij}=n$ and the relations $m_{ij} \fallingdotseq m_{ji}$ for $i \neq j$ are required to hold, we can solve the equations and estimate n_{+}, $n_{\pm}$, n_{-}, p_{ij}'s, under the conditions $n_{+}+n_{\pm}+n_{-}=n$, $\sum_j^R p_{ij}=1$, $i=+, \pm, -$, and some assumptions with respect to the p_{ij}'s. This solution is given by the following steps:

$$\text{(i)} \quad \begin{array}{ll} p_{ij}={}^t p_{ij}(1+\Delta p_{ij}) & \text{for all } i, j \\ n_r={}^t n_r(1+\Delta n_r) & \text{for all } r \end{array} \qquad \ldots.(B)$$

where ${}^t p_{ij}$'s and ${}^t n_r$'s are the tth approximate values, Δp_{ij}'s and Δn_r's are correction terms and $\Delta^2, \Delta^3, \cdots$ are neglected. We estimate the ${}^t p_{ij}$'s and ${}^t n_r$'s.

(ii) We rewrite (A) by using (B), and obtain simultaneous linear equations *w.r.t.* Δp_{ij}'s and Δn_r's, ${}^t p_{ij}$'s and ${}^t n_r$'s being known.

(iii) We solve the simultaneous linear equations mentioned above and have Δp_{ij}'s and Δn_r's.

$$\text{(iv)} \quad \begin{array}{l} {}^t p_{ij}(1+\Delta^t p_{ij})={}^{t+1}p_{ij} \\ {}^t n_r(1+\Delta^t n_r)={}^{t+1}n_r\,. \end{array}$$

We take ${}^{t+1}p_{ij}$'s and ${}^{t+1}n_r$'s instead of ${}^t p_{ij}$'s and ${}^t n_r$'s. Then we repeat the process mentioned above.

(v) Thus, we can solve (A) by the method of successive approximation. However, the idea is deterministic, because we deal with an equation which holds only in expectation.

We can obtain estimates $\hat{p}_{ij}$'s and $\hat{n}_r$'s of p_{ij}'s and n_r's in a similar way, and also the mean square errors or the mean cross-product errors by the idea of (B), where Δ means error (deviation) from p_{ij} or n_r.

Example of assumption with respect to p_{ij}'s. We take $p_{++}=p$, $p_{+\pm}=\alpha p$, $p_{+-}=\alpha^2 p$, $p_{\pm+}=(1-b)\dfrac{n_+}{n_++n_-}$, $p_{\pm\pm}=b$, $p_{\pm-}=(1-b)\dfrac{n_-}{n_++n_-}$, $p_{--}=q$, $p_{-\pm}=\beta q$, $p_{-+}=\beta^2 q$, where $\alpha<1$, $b>0$ and $\beta<1$. The relations $p>\alpha p>\alpha^2 p$ and $q>\beta q>\beta^2 q$ hold. From $p+\alpha p+\alpha^2 p=1$ and $q+\beta q+\beta^2 q=1$, we have $p=\dfrac{1}{1+\alpha+\alpha^2}$, $q=\dfrac{1}{1+\beta+\beta^2}$. These mean that $+$ or $-$ is nearer to $\pm$ than to $-$ or $+$, and response error probability in true $+$ or $-$ is larger in $\pm$ than in $-$ or $+$ (the nearer the response categories are, the larger the response error probability is). The assumptions with respect to $p_{\pm *}$ mean that response probabilities for $+$ and $-$ in the neutral response group are proportional to the numbers n_+, n_- of true $+$ and $-$ responses, i.e. they follow the general trend of response except for the neutral response, and $p_{\pm\pm}$ corresponds to the proportion of intrinsic neutral response.

If we take Δp, Δq, Δb instead of Δp_{ij}'s, following the procedure mentioned above, we have the following simultaneous linear equations (C) with respect to $\Delta\alpha$, $\Delta\beta$, Δb, Δn_+ and Δn_-, and we can solve (C) with respect to them.

Here we take $\alpha={}^0\alpha(1+\Delta\alpha)$, $\beta={}^0\beta(1+\Delta\beta)$, $b={}^0b(1+\Delta b)$, $n_-={}^0n_-(1+\Delta n_-)$ and $n_-={}^0n_-(1+\Delta n_-)$, where the symbol $_0$ means either the true value or the lower order approximation in the successive steps of numerical calculation. Thus we have n_+, n_-, α, β, b by a successive approximation method.

$$\begin{pmatrix} U_{11} & U_{12} & U_{13} & U_{14} & U_{15} \\ U_{21} & U_{22} & U_{23} & U_{24} & U_{25} \\ U_{31} & U_{32} & U_{33} & U_{34} & U_{35} \\ U_{41} & U_{42} & U_{43} & U_{44} & U_{45} \\ U_{51} & U_{52} & U_{53} & U_{54} & U_{55} \end{pmatrix} \begin{pmatrix} \Delta n_+ \\ \Delta n_- \\ \Delta\alpha \\ \Delta\beta \\ \Delta b \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \end{pmatrix}, \qquad \ldots.(C)$$

or $U\Delta=V$ for short, where the determinant of U is not zero generally and the elements of U and V are as follows:

$$U_{11}=(A_1^2-C_1^2)^0n_++2C_1T_1{}^0n_\pm$$
$$U_{12}=-2C_1T_1{}^0n_\pm+({}^0\beta^4B_1^2-C_1^2)^0n_-$$
$$U_{13}=-2A_1S_1{}^0n_+$$

$$U_{14}=2\,{}^0\beta^2 B_1 B_3\,{}^0n_-$$
$$U_{15}=-2C_1\,{}^0bw_1\,{}^0n_\pm$$
$$U_{21}=({}^0\alpha A_1^2-{}^0bC_1){}^0n_+ + {}^0bT_1\,{}^0n_\pm$$
$$U_{22}=-{}^0bT_1\,{}^0n_\pm+({}^0\beta^3B_1^2-{}^0bC_1){}^0n_-$$
$$U_{23}=A_1(A_2-{}^0\alpha S_1){}^0n_+$$
$$U_{24}={}^0\beta B_1({}^0\beta B_2+B_3){}^0n_-$$
$$U_{25}={}^0b(C_1-{}^0b{}^0w_1){}^0n_\pm$$
$$U_{31}=({}^0\alpha^2A_1^2-C_1C_2){}^0n_+ + (C_2T_1-C_1T_2){}^0n_\pm$$
$$U_{32}=(C_1T_2-C_2T_1){}^0n_\pm+({}^0\beta^2B_1^2-C_1C_2){}^0n_-$$
$$U_{33}=A_1(A_3-{}^0\alpha^2S_1){}^0n_+$$
$$U_{34}=B_1(B_3-{}^0\beta^2S_3){}^0n_-$$
$$U_{35}={}^0b(T_1+T_2){}^0n_\pm$$
$$U_{41}=({}^0\alpha^2A_1^2-{}^0b^2){}^0n_+$$
$$U_{42}=({}^0\beta^2B_1^2-{}^0b^2){}^0n_-$$
$$U_{43}=2\,{}^0\alpha A_1A_2\,{}^0n_+$$
$$U_{44}=2\,{}^0\beta B_1B_2\,{}^0n_-$$
$$U_{45}=2\,{}^0b^2\,{}^0n_-$$
$$U_{51}=({}^0\alpha^4A_1^2-C_2^2){}^0n_+-2C_2T_2\,{}^0n_\pm$$
$$U_{52}=2C_2T_2\,{}^0n_\pm+(B_1^2-C_2^2){}^0n_-$$
$$U_{53}=2\,{}^0\alpha^2A_1A_3\,{}^0n_+$$
$$U_{54}=-2B_1S_3\,{}^0n_-$$
$$U_{55}=-2\,{}^0bC_2{}^0w_2\,{}^0n_\pm$$
$$V_1=m_{++}-(A_1^2\,{}^0n_++C_1^2\,{}^0n_\pm+{}^0\beta^4B_1^2\,{}^0n_-)$$
$$V_2=m_{+\pm}-({}^0\alpha A_1^2\,{}^0n_++{}^0bC_1\,{}^0n_\pm+{}^0\beta^3B_1^2\,{}^0n_-)$$
$$V_3=m_{+-}-({}^0\alpha^2A_1^2\,{}^0n_++C_1C_2\,{}^0n_\pm+{}^0\beta^2B_1^2\,{}^0n_-)$$
$$V_4=m_{\pm\pm}-({}^0\alpha^2A_1^2\,{}^0n_++{}^0b^2\,{}^0n_\pm+{}^0\beta^2B_1^2\,{}^0n_-)$$
$$V_5=m_{--}-({}^0\alpha^4A_1^2\,{}^0n_++C_2^2\,{}^0n_\pm+B_1^2\,{}^0n_-)$$

where

$$A_1=\frac{1}{1+{}^0\alpha+{}^0\alpha^2}\qquad A_2={}^0\alpha(A_1-S_1)\qquad A_3={}^0\alpha^2(2A_1-S_1)$$
$$B_1=\frac{1}{1+{}^0\beta+{}^0\beta^2}\qquad B_2={}^0\beta(B_1-S_3)\qquad B_3={}^0\beta^2(2B_1-S_3)$$
$$S_1=A_1^2({}^0\alpha+2\,{}^0\alpha^2)\qquad S_3=B_1^2({}^0\beta+2\,{}^0\beta^2)$$
$$C_1=(1-{}^0b){}^0w_1\qquad C_2=(1-{}^0b){}^0w_2$$
$$T_1=C_1\,{}^0w_2\qquad T_2=C_2\,{}^0w_1$$
$${}^0w_1=\frac{{}^0n_+}{{}^0n_++{}^0n_-}\qquad {}^0w_2=\frac{{}^0n_-}{{}^0n_++{}^0n_-}=1-{}^0w_1\,.$$

And then we have p_{ij}, $i,j=+,\pm,-$, and n_+, $n_\pm$, n_-. Furthermore the matrix of mean square errors and mean cross-product errors of both p's and n's are calculated from the matrix of mean square errors and mean cross-product errors of n_+, n_-, α, β, b. The matrix of mean square errors and mean cross-product errors L of n_+, n_-, α, β, b is approximately calculated from (C). We take $\varDelta\mathfrak{M}$ as a column vector of deviations of a column vector $\mathfrak{M}$, an element of which is m_{ij} ($i,j=+$, $\pm$, $-$, i.e. m_{++}, $m_{+\pm}$, m_{+-}, $m_{\pm\pm}$, m_{--}), and then we have,

$$L=E\left\{\begin{pmatrix}\Delta n_+\\ \Delta n_-\\ \Delta\alpha\\ \Delta\beta\\ \Delta b\end{pmatrix}(\Delta n_+, \Delta n_-, \Delta\alpha, \Delta\beta, \Delta b)\right\}$$

$$=U^{-1}E\left\{\begin{pmatrix}\Delta m_{++}\\ \Delta m_{+\pm}\\ \Delta m_{+-}\\ \Delta m_{\pm\pm}\\ \Delta m_{--}\end{pmatrix}(\Delta m_{++}, \Delta m_{+\pm}, \Delta m_{+-}, \Delta m_{\pm\pm}, \Delta m_{--})\right\}U^{-1\prime}$$

where Δm_{ij} (for some $i, j=+, \pm, -$ as mentioned above) means a sampling fluctuation of m_{ij} from the mean value $E(m_{ij})$, which is expressed by ^{0} symbols on n_+, n_-, α, β, b and equal to the second term of the corresponding constant term V in (C).

$$E\left\{\begin{pmatrix}\Delta m_{++}\\ \Delta m_{+\pm}\\ \Delta m_{+-}\\ \Delta m_{\pm\pm}\\ \Delta m_{--}\end{pmatrix}(\Delta m_{++}, \Delta m_{+\pm}, \Delta m_{+-}, \Delta m_{\pm\pm}, \Delta m_{--})\right\}$$

is theoretically calculated in the same way as was $\sigma(\mathfrak{M})$ in Section 1 and expressed in terms of p's and n's, i.e. $^0\alpha$, $^0\beta$, ^0b and $^0n_+$, $^0n_-$. Thus we have L, and then

$$E\left\{\begin{pmatrix}\Delta n_+\\ \Delta n_\pm\\ \Delta n_-\\ \Delta p_{++}\\ \vdots\\ \Delta p_{--}\end{pmatrix}(\Delta n_+, \Delta n_\pm, \Delta n_-, \Delta p_{++}, \cdots, \Delta p_{--})\right\}.$$

3. Correlated case

We give the following examples.

Suppose that in a measurement the item has three categories $+$, $\pm$, $-$, and the response probabilities are as shown in Table 3, where $p+2q=1$, and the true numbers of those belonging to $+$, $\pm$, $-$ are n_+, $n_\pm$, n_- respectively, with $n_++n_\pm+n_-=n$.

Then we assume that p and q are known and n_+, $n_\pm$, n_- are unknown. An example of cross tabulation in test-retest is shown below in expectation, where $n_+=100$, $n_\pm=1000$, $n_-=100$, $p=0.8$, $q=0.1$.

Table 3.

true \ response probability		+	±	−
n_+	+	p	q	q
$n_\pm$	±	q	p	q
n_-	−	q	q	p

Table 4.

test \ retest	+	±	−	total
+	75	89	26	190
±	89	642	89	820
−	26	89	75	190
total	190	820	190	1200

We have the same marginal distribution for test and retest; however, it may not reveal the true distribution which is obtained by the method of Section 1. If we have only the cross-tabulation without knowing the existence of response error, we might conclude that the subjects who responded + in the test tend to give ± or − responses in the retest, and those who responded − in the test incline toward ± or +.

We meet a similar situation in the case of numerical variables as mentioned later on. We must necessarily construct error models. We have also a similar feature in the cross-tabulation of two items, I, II. See Tables 5 and 6, and suppose that p_{ij}'s and q_{kl}'s were known (or estimated), with

$$\sum_j p_{ij}=1,\ \sum_l q_{kl}=1;\ i=+,\pm,-,\ k=+,\pm,-;$$

$$\sum_i \sum_j n_{ij}=n,\ \sum_k \sum_l m_{kl}=n\,.$$

Table 5.

item I

true \ response probability	+	±	−
+	p_{++}	$p_{+\pm}$	p_{+-}
±	$p_{\pm+}$	$p_{\pm\pm}$	$p_{\pm-}$
−	p_{-+}	$p_{-\pm}$	p_{--}

item II

true \ response probability	+	±	−
+	q_{++}	$q_{+\pm}$	q_{+-}
±	$q_{\pm+}$	$q_{\pm\pm}$	$q_{\pm-}$
−	q_{-+}	$q_{-\pm}$	q_{--}

From m_{kl}'s, the n_{ij}'s are to be estimated by the following equations which are quite similar to (A) mentioned previously.

Valid discussions of the cross-tabulation must be based on estimates of n_{ij}'s instead of the observed m_{ij}'s found in the table.

Table 6.

true

I \ II	+	±	−
+	n_{++}	$n_{+\pm}$	n_{+-}
±	$n_{\pm+}$	$n_{\pm\pm}$	$n_{\pm-}$
−	n_{-+}	$n_{-\pm}$	n_{--}

observation

I \ II	+	±	−
+	m_{++}	$m_{+\pm}$	m_{+-}
±	$m_{\pm+}$	$m_{\pm\pm}$	$m_{\pm-}$
−	m_{-+}	$m_{-\pm}$	m_{--}

$$\begin{pmatrix} \hat{n}_{++} \\ \hat{n}_{+\pm} \\ \hat{n}_{+-} \\ \hat{n}_{\pm+} \\ \hat{n}_{\pm\pm} \\ \hat{n}_{\pm-} \\ \hat{n}_{-+} \\ \hat{n}_{-\pm} \\ \hat{n}_{--} \end{pmatrix} = S^{-1} \begin{pmatrix} m_{++} \\ m_{+\pm} \\ m_{+-} \\ m_{\pm+} \\ m_{\pm\pm} \\ m_{\pm-} \\ m_{-+} \\ m_{-\pm} \\ m_{--} \end{pmatrix}$$

$$\left\{ \left(\begin{array}{c|c|c} Q & 0 & 0 \\ \hline 0 & Q & 0 \\ \hline 0 & 0 & Q \end{array} \right) \left(\begin{array}{c|c|c} p_{++}I & p_{+\pm}I & p_{+-}I \\ \hline p_{\pm+}I & p_{\pm\pm}I & p_{\pm-}I \\ \hline p_{-+}I & p_{-\pm}I & p_{--}I \end{array} \right) \right\}' = S$$

where

$$Q = \begin{pmatrix} q_{++} & q_{+\pm} & q_{+-} \\ q_{\pm+} & q_{\pm\pm} & q_{\pm-} \\ q_{-+} & q_{-\pm} & q_{--} \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad 0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and ( )′ means transposed matrix, under the condition of the existence of S^{-1}. $\hat{n}_{ij}$'s are unbiased estimates of n_{ij}'s and the variance-covariance matrix of them is easily calculated by the same method in Section 1 in the case where S is known. In the case of p's and q's being estimated, the mean square errors and mean cross-product errors of $\hat{n}$'s are approximately calculated.

4. Numerical example

We use two questions.

(i) If you have no children, do you think it necessary to adopt a child in order to continue the family line, even if there is no blood relationship? Or do you think this is not important?

〔answer〕 would adopt; would not adopt; depends on circumstances, and others.

(ii) Which political party do you support?

〔answer〕 liberal democratic; socialist and communist; no party and don't know.

Using the data of the panel-surveys in 1963 and 1965 [12], we show the cross-tabulations for these two questions.

(i)

65 \ 63	would adopt	would not adopt	depends on circumstances	Total
would adopt	391	147	55	593
would not adopt	158	201	53	412
depends on cir.	53	52	46	151
Total	602	400	154	1156

(ii)

65 \ 63	liberal	no party and D.K.	social	Total
liberal	330	83	60	473
no party and D.K.	136	186	76	398
social	58	57	170	285
Total	524	326	306	1156

The marginal distributions are not so different, especially in (i).

Marginal distributions in percent

	(i)			(ii)		
	would	would not	depends	liberal	no party	social
63	52	35	13	45	28	27
65	51	36	13	41	34	25

Thus, we take the probabilistic response model and estimate the response probabilities p's and frequency distribution n's in true response. As $m_{ij} \fallingdotseq m_{ji}$, $(i, j=+, \pm, -)$ are to hold, we use the adjusted cross tabulations as below, in which $(m_{ij}+m_{ji})/2$ is used for both m_{ij} and m_{ji}.

(i)

65 \ 63	would adopt	would not adopt	depends on circumstances	Total
would adopt	391	152.5	54	597.5
would not adopt	152.5	201	52.5	406
depends on cir.	54	52.5	46	152.5
Total	597.5	406.0	152.5	1156.0

(ii)

65 \ 63	liberal	no party and D.K.	social	Total
liberal	330	109.5	59	498.5
no party and D.K.	109.5	186	66.5	362
social	59	66.5	170	295.5
Total	498.5	362.0	295.5	1156.0

We calculate according to the formulae given above and obtain the response-probability matrix and frequency distribution in true response as below:

$$\begin{pmatrix} 0.75 & 0.20 & 0.05 \\ 0.18 & 0.76 & 0.06 \\ 0.25 & 0.32 & 0.43 \end{pmatrix}, \quad \begin{pmatrix} 657 \\ 264 \\ 235 \end{pmatrix} \quad \text{for (i)}$$

and

$$\begin{pmatrix} 0.73 & 0.21 & 0.06 \\ 0.02 & 0.96 & 0.02 \\ 0.10 & 0.25 & 0.65 \end{pmatrix}, \quad \begin{pmatrix} 620 \\ 138 \\ 398 \end{pmatrix} \quad \text{for (ii)}.$$

Suppose that we have got the p's and n's. Next, we use the cross-tabulation data of (i)×(ii), and estimate the true cross-tabulation of (i)×(ii). As the data of (i)×(ii), we use the matrix whose elements are arithmetic means of the corresponding elements of the observed (i)×(ii) in the two years,

$$\text{(i)} \left\{ \overbrace{\begin{pmatrix} 275 & 184.5 & 138 \\ 160.5 & 122.5 & 123 \\ 63 & 35 & 34.5 \end{pmatrix}}^{\text{(ii)}} \right. .$$

From this matrix, we estimate the true cross-tabulation matrix by the method mentioned in Section 3. Thus we have

	(ii)			Total
(i)	406.8	70.8	179.8	657.4
	101.7	10.9	149.9	262.5
	121.2	55.3	59.6	236.1
Total	629.7	137.0	389.3	

This shows a more reasonable feature than does the cross-tabulation of raw data, and reveals a clearer structure. Those who support the conservative (liberal-democratic) party are quite in favour of "would adopt", and respond both to "would not adopt" and "depends on circumstances" in about the same proportion. Those of the "don't know" group vote predominantly for "would adopt" and "depends on circumstances". This is quite different from the cross-tabulation of raw data. We are aware of the fact that the marginal frequency distribution in true response is quite different from that of the data in both cases, and we know that the frequency distributions in the data lead us to an invalid interpretation without taking response error into consideration.

The reproduced cross tables obtained by using these calculated parameters are as shown below.

$$\begin{pmatrix} 390.9 & 252.6 & 53.9 \\ 152.6 & 201.5 & 52.4 \\ 53.9 & 52.4 & 45.8 \end{pmatrix}, \quad \begin{pmatrix} 597.4 \\ 406.5 \\ 152.1 \end{pmatrix} \quad \text{for (i)},$$

and

$$\begin{pmatrix} 332.1 & 107.7 & 52.5 \\ 107.7 & 180.2 & 75.1 \\ 62.5 & 75.1 & 173.1 \end{pmatrix}, \quad \begin{pmatrix} 492.3 \\ 363.0 \\ 300.7 \end{pmatrix} \quad \text{for (ii)}.$$

It is seen that they fit the data fairly well, especially in the case of (i).

II. QUANTITATIVE CASE

1. Fundamental theory

We assume that

$$x = x_0 + \varepsilon$$

$$y = y_0 + \eta,$$

where x_0 and y_0 are true values, and ε and η are error terms, represented by random variables; $E(\varepsilon)=0$, $E(\eta)=0$, $E(\varepsilon^2)=\sigma_\varepsilon^2$, $E(\eta^2)=\sigma_\eta^2$, and $E(\varepsilon\eta)=0$. x_0 and y_0 are of course random variables too, and $E(x_0)=M_x$, $E(y_0)=M_y$, $E(x_0^2)-E(x_0)^2=\sigma_{x_0}^2$, $E(y_0^2)-E(y_0)^2=\sigma_{y_0}^2$, and we also assume $E[\{x_0-E(x_0)\}\varepsilon]=0$, $E[\{x_0-E(x_0)\}\eta]=0$, $E[\{y_0-E(y_0)\}\varepsilon]=0$, and $E[\{y_0-E(y_0)\}\eta]=0$.

Now, we take $x_0=y_0$, $\sigma_\varepsilon^2=\sigma_\eta^2$ for simplicity. We imagine that x and y stand for measurements at time t and time $t+1$, respectively, and $x_0=y_0$ for the same object holds essentially; however, $x\neq y$ may be observed. To be exact, the ith object has (x_i, y_i), where $x_i=x_{0i}+\varepsilon_i$ and $y_i=y_{0i}+\eta_i=x_{0i}+\eta_i$. The conditions of mutual independence mentioned above hold for every element.

If x_0 and y_0 are random variables which follow a density function other than the uniform—for example, a Gaussian distribution—and the errors ε and η are random variables which follow a Gaussian distribution, the linear regression of y on x is clearly not L but L_x in Fig. 1. That

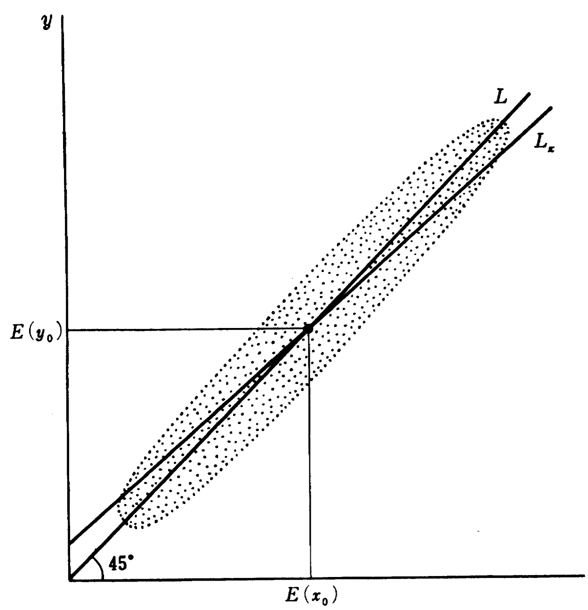

Fig. 1. Response errors

is, the expectation of y corresponding to x, smaller than $E(x_0)$, is larger than x, while that corresponding to x, larger than $E(x_0)$ is smaller than x, and $E(x_0)=E(y_0)$. Without taking this into account, we would have incorrect conclusions.

We often meet this situation in data analysis when a treatment (including no treatment) is given between time t and $t+1$, and the effect of treatment is discussed. When the treatment has no effect, i.e. $x_0=y_0$ for every person, we might conclude, by some statistical test *w.r.t.* x_0 and y_0, that the treatment brings about an increase of x if $x<E(x_0)$ and a decrease of x if $x>E(x_0)$ if we were to neglect the error terms ε and η. A well-known theory exists for discussing the relation between x and y in the case when x and y are both subject to errors. However this theory only reveals the structure, and is not of use for prediction. Prediction of y is to be made on the basis of information on x even though x may include errors. For this purpose, the idea of regression of y on x becomes indispensable.

We shall show an illustrative example as below. Suppose that $x_0=y_0$ for the same object, and they are random variables which follow the Gaussian distribution, $\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(z-M)^2/2\sigma^2}$, and ε, η are random variables both of which follow the Gaussian distribution, $\frac{1}{\sqrt{2\pi}\,s}e^{-u^2/2s^2}$, s being a constant. Then the probability density function of the observed value $x=x_0+\varepsilon$ is $\frac{1}{\sqrt{2\pi}\,s}e^{-(x-x_0)^2/2s^2}$, x_0 being fixed.

Then we have, by easy calculation in probability theory,

$$P\{y\geqq w+d\mid x=w\}=\frac{1}{\sqrt{2\pi}}\int_{A+Bd}^{\infty}e^{-t^2/2}dt\,,$$

$$E(d)=-A/B$$

$$E(d-E(d))^2=1/B^2$$

where

$$A=(w-M)s/\sqrt{(s^2+\sigma^2)(s^2+2\sigma^2)}$$

$$B=\left(1+\frac{\sigma^2}{s^2}\right)s/\sqrt{(s^2+\sigma^2)(s^2+2\sigma^2)}\,.$$

Thus, we see that

$$P\{y\geqq w\mid x=w\}\begin{cases}>\frac{1}{2} & \text{if } w<M\\ <\frac{1}{2} & \text{if } w>M\\ =\frac{1}{2} & \text{if } w=M.\end{cases}$$

Also, since $E(d)$ is clearly equal to $-(w-M)/(1+\sigma^2/s^2)$, we see that

$$E(d)\begin{cases}>0 & \text{if } w<M\\ <0 & \text{if } w>M\\ =0 & \text{if } w=m\,.\end{cases}$$

For example, let $\sigma^2=200$ and $s^2=50$. Then,

$$E(d)=-(w-M)/(1+4)$$
$$=-(w-M)/5$$

and the effect of the error term is rather large, if s^2 is not negligible compared to σ^2.

It is noted that we always have $P\{y\geqq w\,|\,x=w\}=1/2$, if the x_0 and y_0 mentioned above follow a uniform distribution.

We shall show some cases as below.

(i) Suppose that $y_0=\alpha+\beta x_0$ for the same object where α and β are constants. x_0 follows the Gaussian distribution, the mean and variance of which are M and σ respectively. y_0 follows the Gaussian distribution, the mean being $\alpha+\beta M$ and the variance, $\beta^2\sigma^2$. The assumptions concerning ε and η are the same as those in the case mentioned above.

The regression line of y on x is

$$y=(\alpha+\beta w)-\beta(w-M)/(1+\sigma^2/s^2)\ ,$$

where x is observed as w.

(ii) Suppose that $y_0=\alpha+\beta x_0$, as mentioned above. ε follows the Gaussian distribution, the mean being 0 and the variance, s_1^2, whereas η follows the Gaussian distribution with mean 0 and variance s_2^2 ($\neq s_1^2$).

In this case, let d be the deviation from the regression line of y on x.

$E(d)=-\beta(w-M)/(1+\sigma^2/s_1^2)$ which is independent of s_2^2. The variance of d is $1/B'^2$, where $B'^2=(1+\sigma^2/s_1^2)(s_1/s_2)s_1/\sqrt{(s_1^2+\sigma^2)\{s_1^2+\sigma^2(1+\beta^2s_1^2/s_2^2)\}}$ which is influenced by s_2. If $s_1^2=0$, the mean deviation from the regression line of y on x is, of course, 0.

(iii) In a more general case, $x_0=y_0$ for the same subject, and these follow the same distribution $\Phi(x_0)$ ($\Phi(y_0)$) which is not Gaussian. For simplicity, we assume that ε follows the Gaussian distribution with mean 0 and a variance which is a function of x_0, $s^2=g(x_0)$, and that η follows the same Gaussian distribution. In this case we can calculate the distribution of d by numerical computation using the following formula:

$$P\{y\geqq w+d\,|\,x=w\}$$
$$=\int_{-\infty}^{\infty}\varphi(w,x_0)\Phi(x_0)dx_0\int_{w+d}^{\infty}\frac{1}{\sqrt{2\pi}\,s(x_0)}\exp\left\{-\frac{(z-x_0)^2}{2s(x_0)^2}\right\}dz\Big/C(w)\ ,$$

where $$C(w)=\int_{-\infty}^{\infty}\varphi(w,x_0)\Phi(x_0)dx_0\,,$$

$$\varphi(w,x_0)=\frac{1}{\sqrt{2\pi}\,s(x_0)}\exp\left\{-\frac{(w-x_0)^2}{2s(x_0)^2}\right\}\ .$$

Thus we can obtain the mean and variance of d.

It seems natural to assume that the distributions of errors follow a Gaussian distribution. If we can assume the $s(x_0)$, it is easy in practice to carry out the calculation, estimating $\Phi(x_0)$, for example, by estimating the parameters assuming the functional form of $\Phi(x_0)$ or generally estimating approximately from the data, because the observed frequency distribution in the data is realized by the compound distribution of $\Phi(x_0)$ with the error distribution. This is shown in [10] with examples in medical research.

2. Application

We meet the same situation in the follow-up study and this often leads us to invalid conclusions. Suppose that the functional relation between an outside variable y and factors $x_1, x_2, \cdots, x_R$ is determined in an experiment as $y=f(x_1, x_2, \cdots, x_R)+\varepsilon$, ε being the error term which is represented by a random variable independent of $x_1, x_2, \cdots, x_R$, with $E(\varepsilon)=0$ and $E(\varepsilon^2)=\sigma_\varepsilon^2$. We use this stochastic functional relation (whose precision is represented by ε) to estimate y from factors $x_1, x_2, \cdots, x_R$.

In the follow-up experiment, we have y's and x_1's, x_2's, $\cdots$, x_R's and we assume that y's have measurement errors. We get the estimated value y' from $x_1, x_2, \cdots, x_R$ by the functional relation $f(\ , \ , \cdots, \ )$ obtained in past experiment. y' is a random variable including an error in estimation. Suppose that the functional relation is to be verified by the follow-up study. Put y here as the y in the foregoing discussion, and y' here as the x in the foregoing discussion. The regression line of y on y' may not be L, i.e. the 45°-line but a straight line the slope of which is less than 45°. This contradicts our naive expectation, but it is generally true that the regression line is not the 45° line, even though the functional relation holds in the two experiments. The actual slope depends on the spacing of the chosen experimental conditions.

If this is ignored, we can not draw any valid conclusion in the follow-up study.

III. THEORY OF QUANTIFICATION AND RESPONSE ERRORS

Response errors are treated as follows in the theory of quantification, details of which are discussed in [5], [6] and their references. We take the response patterns $[\{\delta_i(j,k);\ j=1,2,\cdots,R,\ k=1,2,\cdots,K_j\}\ i=1,2,\cdots,N]$ where $\delta_i(jk)=1$ if the ith subject makes the kth category response in the jth item and $\delta_i(jk)=0$, otherwise, R being the number of items, K_j being the number of categories in the jth item and N the size of sample. If response errors exist, we assume that responses are represented by a probabilistic model, i.e. $\delta_i(jk)$ is represented by $\delta_i(jk) = {}^s p_{jk}$ if $i \in s$ (i belongs to the sth class) where $\sum_{k=1}^{K_j} {}^s p_{jk} = 1$ and $s=1, 2, \cdots, S$, and $\delta_i(jk)$ and $\delta'_{i'}(j'k')$ are independent for every i, i' including $i=i'$ if $j \neq j'$ ($j, j'=1, 2, \cdots, R$), R being the number of items. If j and j' are not independent, make a new combined item $(j \times j')$ and take $\delta_i(j \times j'\ k \times k') = {}^s p_{j \times j'\ k \times k'}$ where $k \times k'$ is the number of categories in the new item $(j \times j')$. If ${}^s p_{jk}$'s have been determined previously, they are used. However, ${}^s p_{jk}$'s must be estimated from the data in some cases. Bayes' theorem will be sometimes useful in the estimation of response probabilities from the data at hand. We have given a method of estimation of those probabilities in [9].

In the case where no outside variable exists, we are seriously misled if any response error is disregarded. We have also shown such examples in [9].

Acknowledgements

The author wishes to thank Professor M. M. Tatsuoka, University of Illinois for his valuable suggestions which considerably improved the presentation of the material and Mr. K. Takahasi, Institute of Statistical Mathematics for his careful reading of this manuscript.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] A. A. Abul-Ela, B. G. Greenberg and D. G. Horvitz, "A Multi-proportions randomized response model," *J. Amer. Statist. Ass.*, 62 (1967), 990-1008.

[2] T. W. Anderson, "On estimation of parameters in latent structure analysis," *Psychometrika*, 19, 1 (1954), 1-10.

[3] L. Guttman, "Deviation theory for dichotomies," *Ann. Inst. Statist. Math.*, 16, 1 (1964), 69-78.

[4] C. Hayashi, "Response errors and sampling design," *Proc. Inst. Statist. Math.*, 5, 1 (1957), 11-26.

[5] C. Hayashi, "Fundamental concepts of the theory of quantification and prediction," *Proc. Inst. Statist. Math.*, 7, 1 (1959), 43-64.

[6] C. Hayashi, "Sample surveys and theory of quantification," *Bull. Int. Statist. Inst.*, 38, IV (1961), 505-514.

[7] C. Hayashi, "Measurement problem in the light of statistical mathematics," *Journal of Japan Association of Philosophy of Science*, 6, 4 (1964), 143-152.

[8] C. Hayashi, "Statistical model analysis for philosophy of sciences," *Journal of Japan Association of Philosophy of Science*, 8, 2 (1967), 70-79.

[9] C. Hayashi, "Response errors and statistical analysis in social surveys," Collected Papers in Commemoration of the 20th Anniversary, Radio and TV Culture Research Institute and Public Opinion Research Institute of NHK, (1967), 471-546.

[10] C. Hayashi, Y. Fukuda, R. Hosoya and F. Hayashi, "Response errors and correlation analysis—Some problems and data analysis in medical science—," *Proc. Inst. Statist. Math.*, 15, 2 (1967), 107-125.

[11] K. Noda, "Reliability of responses in social surveys," *Proc. Inst. Statist. Math.*, 14, 2 (1966), 87-95.

[12] T. Suzuki, "A study of the Japanese national character—Part III—The third national-wide survey," *Ann. Inst. Statist. Math.*, Supplement IV, (1966), 15-64.

[13] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Ass.*, 60 (1965), 63-69.