On the Chi-Square Test for Weighted Samples

By Hirojiro AOYAMA

(Received April 5, 1953)

1. Introduction

When testing whether the sample distribution fits the given theoretical one, we usually make use of the χ^2-test, and it is assumed that the sample is got by the unrestricted random sampling. This assumption, however, is not satisfied in the ordinary sample survey. That is, we use the stratified unrestricted random sampling, the stratified subsampling and so on. Nevertheless in the analysis of the data we sometimes treat the sample as obtained by the unrestricted random sampling.

In this paper we shall discuss under what condition this treatment is valid when we adopt the stratified random sampling.

2. Chi-Square Test in the Stratified Sampling

We divide the universe of size N into R strata by some control, whose size are N_1, N_2,......, N_R, respectively. Here we take n_1, n_2,......, n_R samples from these strata by simple random sampling, respectively.

As for a certain characteristic, let n_{ij} be the number of samples with the j-th category of the characteristic in the i-th stratum $(i=1,2,..., R;\ j=1,2,...,M)$. They are illustrated in the following table. When we know the universe values N_{ij} and $N_{(j)}$, we can test as usual, whether these sample data are fitted for the universe distribution by means of the χ^2-test with M-1 degrees of freedom for

Cat. Stratum	1	2	...	j	...	M	Sum
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1M}	n_1
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2M}	n_2
⋮	⋮	⋮		⋮		⋮	⋮
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iM}	n_i
⋮	⋮	⋮		⋮		⋮	⋮
R	n_{R1}	n_{R2}	...	n_{Rj}	...	n_{RM}	n_R
Sum	$n_{(1)}$	$n_{(2)}$	...	$n_{(j)}$	...	$n_{(M)}$	n

stratum and $R(M-1)$ degrees of freedom for whole universe.

When we have no information of N_{ij} except for that of $N_{(j)}$, we have the following table.

Cat.	1	2	....	j	....	Sum
Estimated value	$\sum_i N_i \frac{n_{i1}}{n_i}$	$\sum_i N_i \frac{n_{i2}}{n_i}$	....	$\sum_i N_i \frac{n_{ij}}{n_i}$	....	N
Universe value	$N_{(1)}$	$N_{(2)}$	....	$N_{(j)}$	....	N

In this case we put degrees of freedom $M-1$ and

$$\chi^2 = \frac{n}{N}\sum_j \frac{\left(\sum_i N_i \frac{n_{ij}}{n_i} - N_{(j)}\right)^2}{N_{(j)}} = \frac{n}{N}\sum_j \frac{\left(\sum_i N_i \frac{n_{ij}}{n_i}\right)^2}{N_{(j)}} - n \qquad (1)$$

This procedure, however, is limited in the strict sense only to the criterion whether our whole samples are simple random. And even for the proportional sampling, in which the contribution of each sample is the same, the degrees of freedom are not just $M-1$.

As for the expected value of χ^2, neglecting the finite population correction, we have

$$E(\chi^2) = \frac{n}{N}\sum_i \sum_j \frac{N_{ij}(N_i - N_{ij})}{n_i N_{(j)}} \qquad (2)$$

As for the variance of χ^2 we have

$$\begin{aligned} V(\chi^2) = {} & \frac{n^2}{N^2}\sum_j \frac{1}{N_{(j)}^2}\Bigg[\sum_i \frac{1}{n_i^2}\Big\{(N_i - N_{ij})(-3N_{ij}^3 + 11N_i N_{ij}^2 - 8N_{ij}N_{(j)} \\ & -4N_i N_{ij} N_{(j)})\Big\} + 3\sum_{i\neq i}\sum \frac{N_{ij}N_{i'j}}{n_i n_{i'}}(N_i - N_{ij})(N_{i'} - N_{i'j})\Bigg] \\ & + \frac{n^2}{N^2}\sum_{j\neq j'}\sum \frac{1}{N_{(j)}N_{(j')}}\Bigg[\sum_i \frac{1}{n_i^2}\Big\{3N_{ij}^2 N_{ij'}^2 - N_i N_{ij}^2 N_{ij'} - N_i N_{ij'}^2 N_{ij} \\ & + N_i^2 N_{ij} N_{ij'} + 4N_{ij}N_{ij'}^2 N_{(j)} - 2N_i N_{ij} N_{ij'} N_{(j)} + 4N_{ij'}N_{ij}^2 N_{(j')} \\ & -2N_i N_{ij} N_{ij'} N_{(j')}\Big\} + \sum_{i\neq i'}\sum \frac{1}{n_i n_{i'}}\Big(N_i N_{i'} N_{ij} N_{i'j'} + N_{ij}^2 N_{i'j'}^2 \\ & - N_i N_{ij} N_{i'j'}^2 - N_{i'} N_{i'j'} N_{ij}^2 + 2N_{ij}N_{ij'}N_{i'j}N_{i'j'}\Big)\Bigg] + \frac{n^2}{N^2}\sum_j \frac{1}{N_{(j)}^2} \\ & \times \sum_i \frac{N_{ij}}{n_i^3}(N_i - N_{ij})(N_i^2 - 6N_i N_{ij} + 6N_{ij}^2) + \frac{n^2}{N^2}\sum_{j\neq j'}\sum \frac{1}{N_{(j)}N_{(j')}}\sum_i \frac{1}{n_i^3} \\ & \times\left(-6N_{ij}^2 N_{ij'}^2 + 2N_i N_{ij}^2 N_{ij'} + 2N_i N_{ij} N_{ij'}^2 - N_i^2 N_{ij} N_{ij'}\right) \end{aligned}$$

$$= I_1\left(\text{terms of } \frac{1}{n_i^2} \text{ or } \frac{1}{n_i n_{i'}}\right) + I_2\left(\text{terms of } \frac{1}{n_i^3}\right) \qquad (3)$$

For the proportional sampling we get

$$E(\chi^2)_{\text{prop.}} = M - \sum_i \sum_j \frac{N_{ij}^2}{N_i N_{(j)}} \qquad (4)$$

and only for $R=1$

$$E(\chi^2) = M-1 \qquad (5)$$

If we can assume that the stratification into R strata is carried out at random, that is, the stratification of $N_{(j)}$ into N_{ij} $(i=1, 2, \dots, R)$ is carried out at random, the expectation $\mathcal{E}$ of χ^2 with respect to this randomization is

$$\mathcal{E}E(\chi^2)_{\text{prop.}} \doteqdot M-1-\frac{M}{N}(R-1) \qquad (6)$$

Hence for large N we have the same result as (5) in the sense of expectation. Here we must take care of the fact that $\sum_i \sum_j \frac{N_{ij}^2}{N_i N_{(j)}}$ attains 1 as the minimum value.

In general case we have

$$\mathcal{E}E(\chi^2) \doteqdot \frac{n(M-1)}{R^2} \sum_i \frac{1}{n_i} \qquad (7)$$

$$\mathcal{E}V(\chi^2)_{\text{prop.}} = \frac{2(M-1)}{R} + 2(R-1)^2 - \frac{2(R-1)}{R} + O\left(\frac{1}{N}\right) + O\left(\frac{1}{n}\right) \qquad (8)$$

For $R=1$ we can easily get approximately from (3)

Table of $\mathcal{E}V(\chi^2)_{\text{prop.}}$ for various M and R

R \ M	2	3	4	5	10	15	20
2	2.0	3.0	4.0	5.0	10.0	15.0	20.0
3	7.3	8.0	8.7	9.3	12.7	16.0	19.3
4	17.0	17.5	18.0	19.5	21.0	23.5	26.0
5	30.8	31.2	31.6	32.0	34.0	36.0	38.0
10	160.4	160.6	160.8	161.0	162.0	163.0	164.0
15	390.3	390.4	390.6	390.7	391.3	392.0	392.7
20	720.2	720.3	720.4	720.5	721.0	721.5	722.0

$$V(\chi^2) \doteqdot 2(M-1)\left(1-\frac{1}{n}\right) \qquad (9)$$

Therefore, except for $R=1$ we cannot consider that this statistic χ^2 is distributed according to the χ^2-law with $M-1$ degrees of freedom and for the proportional sampling we can only state that χ^2 obeys approximately the χ^2-law with $M-1$ degrees of freedom in the sense of the expectation under the condition $M=R(R-1)$ and $R \fallingdotseq 1$.

INSTITUTE OF STATISTICAL MATHEMATICS

ERRATA

These Annals, Vol. II, No. 1, 1950. P. 18 insert after last line of section 2

"In these cases we have not the maximum value but only the stationary value just as the minimax solution. If we want to obtain the maximum value, we must estimate the rational rate k_1 and k_2 from experiences in the past time. This fact holds also in the following sections."

Vol. V, No. 1, 1953

Page line

27, 9, read $M-1-\frac{(M-1)(R-1)}{N}$ instead of the right hand side of (6)

27, 12, insert under the assumption after "we have"
$N_i=N/R$

27, 14, read $2(M-1)+O\left(\frac{1}{N}\right)+O\left(\frac{1}{n}\right)$ instead of the right hand side of (8)

27, last, read (strike off the table)

28, 5-6, read (strike off the sentence "under the condition $M=R(R-1)$ and $R \mp 1$"

Vol. VI, No. 1, 1954

Page	line				
13,	12,	read	$\binom{M}{Mp_i}p^{Mp_i}q^{Mq_i}$	instead of	$\binom{M}{Mp_i}p^{Mp_i}p^{Mq_i}$
14,	3,	read	$0.96\,N$	instead of	$096\,N$
15,	6,	read	$\ldots k\sqrt{\varepsilon^* D^2(X)}\} \leqq \frac{1}{k^2}$	instead of	$\ldots k\sqrt{\varepsilon^* D^2(X)} \leqq \frac{1}{k^2}$
15,	23,	read	$X_{(i)}$	instead of	$X_{i)}$
24,	7,	read	$-\mu_{11}(2)\mu_{20}(2)\ldots$	instead of	$-\mu_{11}(1)\mu_{20}(2)\ldots$
24,	10,	read	$\frac{N_1^2N_2}{N^3}((\bar{X}_1-\bar{X}_2)\ldots$	instead of	$\frac{N_1^2N_2}{N^3}(\bar{X}_1-\bar{X}_2)\ldots$
25,	9,	read	$\frac{2N_1N_2}{N^3}(\bar{Y}_1-\bar{Y}_2)^2\cdots$	instead of	$\frac{2N_1N_2}{N^2}(\bar{Y}_1-\bar{Y}_2)^2\cdots$
			$+\frac{N_1N_2}{N^5}(N_1^3+N_2^3)\cdots$	instead of	$+\frac{N_1N_2}{N^5}(N_1^2+N_2^3)\cdots$
28,	2 from the bottom,		$+O(n^{-3/2})$	instead of	$+O(^{-3/2})$
30,	11,	read	$-\frac{4\mu_{31}}{\mu_{11}\mu_{20}}-\frac{4\mu_{13}}{\mu_{11}\,\mu_{02}}+\cdots$	instead of	$-\frac{4\mu_{31}}{\mu_{11}\mu_{21}}-\frac{4\mu_{13}}{\mu_{11}\mu_{12}}+\cdots$
36,	3 from the bottom,		the coming issue	instead of	this issue

Page	line				
54,	6,	read	[20], Lemma	instead of	[20 , Lemma
68,	28,	read	$e^{I\varepsilon(t)}$	instead of	$eI^{\varepsilon(t)}$
97,	6,	read	(X_j, Y_j) has	instead of	$(X_j, Y_j$ has
98,	2,	read	$\lim_n \frac{D(S_{nj})}{D_n}$	instead of	$\lim_n \sum_l \frac{D(S_{nj})}{D_n}$