Application of the Hypergeometric Distribution In a Special Case of Rare Events

Yan Liu, Mary Batcher and Wendy Rotz, Ernst & Young LLP
Yan Liu, Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036

Abstract: Auditors are often faced with reviewing a sample of business invoices and estimating the number of error items and the total error amount for a rarely occurring event. One special case is dividing the dollars in a large population of invoices into two categories according to whether they meet or do not meet the requirements, are or are not in error. In other words, the error amount follows a nonstandard mixture distribution in which the error amount is either zero with a large probability or the original invoice amount with a very small probability. It is likely that the sample or some strata of the sample will include only zero error items. Under this scenario, the classical method will not produce satisfactory estimation, especially when a conservative estimate of the number of error items or error amount is needed. We will show some flexible applications using the hypergeometric distribution.

Key words: mixture distributions; hypergeometric distribution; audit sampling; rare events; stratified sampling

1. Introduction

Suppose the population includes N invoices and each has a known invoice amount. The invoices are divided into two classes – error class C and non-error class $\tilde{C}$. If an invoice is in error, then the error amount is equal to its invoice amount; otherwise the error amount is zero. The percentage of items in error is very small, i.e., most observations have a zero error amount. For example, if a sample, simple random sample or stratified random sample, contains no error, classical methods do not apply anymore. In this paper, estimation using the hypergeometric distribution is proposed. An application using the upper bound of the error rate for a simple random sample can be found in Cochran (1977) and Wilburn (1984). We extend the application of the hypergeometric distribution to stratified designs and to the estimation of the error amount. The proposed method is also illustrated using a simulation.

Let x be the known amount for each item and y be the unknown error amount in the population. Then the error amount is

$$y = \begin{cases} x, & \text{if the item is in Error Class } C \\ 0, & \text{if the item is in Non-error Class } \tilde{C} \end{cases}$$

2. Application To Simple Random Sample

Suppose that M items in the population are in error class C and the other $N-M$ items in the population are in the non-error class $\tilde{C}$. From the simple random sample of n items, the number of error items a in the sample given a total of M error items in the population follows the hypergeometric distribution:

$$\Pr(a \mid M) = \frac{\binom{M}{a}\binom{N-M}{n-a}}{\binom{N}{n}} \qquad (1)$$

In practice, the value of M is of interest but unknown. Instead, the sample result of a error items in the sample is known. Therefore, the probability of M error items in the population given that a out of n items in the simple random sample turn out to be in error is of interest. It can be derived as

$$\Pr(M \mid a) = \frac{\Pr(a \mid M)}{\sum_{M=a}^{N} \Pr(a \mid M)} \qquad (2)$$

Further, ***assuming that error items are evenly spread in the population***, which is the application condition for this method in estimating the error amount and also a reasonable assumption when fine stratification is used. Then an approximate of the

total error amount Y corresponding to the M error items in the population is

$$Y = \overline{X}M \quad [1] \tag{3}$$

where $\overline{X}$ is the known mean invoice amount of the population.

The probability of Y given the sample results of a error items is

$$\Pr(Y|a) = \Pr(M|a) \tag{4}$$

3. Application To Stratified Sample

To apply the above method in stratified samples, we need to find the probability distribution of $M = M_1 + M_2$. Here the subscript denotes the stratum number, i.e., M_1 and M_2 are the population number of error items in strata 1 and 2. Let a_1 and a_2 be the number of error items from the sample of size n_1 in stratum 1 and the sample of size n_2 in stratum 2 separately. Since the selections across strata are independent, the probability distribution of total error items $M = M_1 + M_2$ in the population given the sample result of a_1 error items in stratum 1 and a_2 error items in stratum 2 is

$$\Pr(M|a_1, a_2)$$

$$= \sum_{\substack{All\ M_1\ \&\ M_2\ Such \\ That\ M=M_1+M_2}} p(M_1|a_1)\,p(M_2|a_2) \tag{5}$$

where $\Pr(M_1|a_1)$ and $\Pr(M_2|a_2)$ have the same functional form of equation (2) except with subscripts.

Similarly, the approximate total error amount $Y = Y_1 + Y_2$ corresponding to M_1 and M_2 is

$$Y = Y_1 + Y_2 = \overline{X}_1 M_1 + \overline{X}_2 M_2 \tag{6}$$

where $Y_1 = \overline{X}_1 M_1$ is the approximate total error amount corresponding to M_1 in stratum 1 and $\overline{X}_1$ is the known population mean invoice amount of stratum 1. Similarly, $Y_2 = \overline{X}_2 M_2$ is the approximate total error amount corresponding to M_2 in stratum 2 and $\overline{X}_2$ is the known population mean invoice amount of stratum 2. The probability distribution of Y given the sample results (a_1, a_2) is calculated by

$$P(Y|a_1, a_2)$$

$$= \sum_{\substack{All\ Y_1\ \&\ Y_2\ Such \\ That\ Y=Y_1+Y_2}} p(Y_1|a_1)\,p(Y_2|a_2) \tag{7}$$

where

$$\Pr(Y_1|a_1) = \Pr(M_1|a_1)$$

and

$$\Pr(Y_2|a_2) = \Pr(M_2|a_2).$$

4. Estimation

Using the probability distribution of M given a by equation (2), there are two point estimators, the mode and the mean $E(M|a)$. The classical method is to treat the mode as a function of some average, where $\frac{a}{n}$ is the average number of error items in the sample. This method is well established in Cochran (1977) and appropriate if a is not close to zero. For a simple random sample, the mode $\frac{a}{n}N$ is an unbiased estimator of M in the sense of repeated sampling. But for a very small, the mode may not be a good estimate. The corresponding confidence interval from Cochran (1977) works well only if a is reasonably large. Further, when there are no error items in the sample, this method does not apply anymore.

Instead, $E(M|a)$ is a good choice especially when a conservative estimate is desired. Since the

---

[1] To take into account the error amount in the sample, we can use $Y = Y_n + \overline{X}_{(N-n)}(M-a)$ to approximate the total error amount, where Y_n is the known total error amount from the sample of n items and $\overline{X}_{(N-n)}$ is the known mean invoice amount from the $N-n$ non-sampled items.

conditional probability distribution of M given a is positively skewed in the rare error situation, the mean estimator is more conservative than the mode estimator. When there are no error items in the simple random sample and the mode is zero, we would use $E(M|a)$ as the point estimator.

The exact confidence interval of the population number of error items can be obtained using the conditional probability distribution of M given a by equation (2). The confidence intervals can be determined from the probability distribution by the criterion of Minimum Expected Length (Bain & Engelhardt, 1991).

Similarly, the estimate about the total error amount is obtained from the probability distribution of Y given a by equation (3).

The proposed estimator and confidence interval can also be used in a stratified sample using the probability distributions defined by (5) and (7).

The estimation process is illustrated in the following simulation section.

5. Simulation and Estimation Example

We typically have stratified designs in practice. The population of 600 records is first sorted by the value of x and then divided into two equal-sized strata. Figure 1 is the histogram of the design variable x. The population summary is given in Table 1.

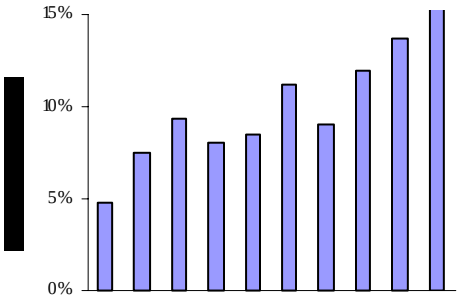

Figure 1. Histogram of the Simulation Population (x)

As shown in Table 1, there are 15 (5%) error items in stratum 1 and 30 (10%) error items in stratum 2. The error items are evenly spread out over the population. Two thousand stratified random samples of size 70 are selected, with 30 items from stratum 1 and 40 items from stratum 2. Table 1 and Table 2 give the distribution of the 2000 repeated samples for stratum 1 and stratum 2. Theoretically, the number of errors in the samples from stratum 1 could be 0 – 15. But the probabilities on the high end are extremely small. As shown in Table 1, none of the 2000 samples have more than 7 errors. The number of errors per sample from stratum 2 is no more than 10.

As shown in Table 2, when the error occurrence is 5%, 386 out of 2000 samples include no non-zero error items and we cannot use classical methods. The chance of getting only 1 or 2 non-zero items in the sample is very high--again a situation where classical methods do not provide a good estimate and confidence interval.

Tables 2 and 3 show that the conditional mean is larger than the mode for each given a.

Table 1. Population Summary

Stratum	Population Size N	Number of Error Items M	Total Invoice Amount X	Total Error Amount Y
1	300	15	279,024	14,080
2	300	30	510,506	51,024
Total	600	45	789,530	65,104

Table 2. Distribution of 2000 Simple Random Samples of Size 30 from stratum

Number of Errors in the Sample a	Number of Samples	Mode of the Probability Distribution of Y Given a	Estimated Number of Errors $\hat{M} = E(M\|a)$
0	386	0	8.44
1	689	10	17.86
2	562	20	27.31
3	251	30	36.75
4	96	40	46.19
5	12	50	55.63
6	3	60	65.06
7	1	70	74.50
Total	**2000**	**15.2**	**22.8**

Table 3. Distribution of 2000 Simple Random Samples of Size 30 from stratum 2

Number of Errors in the Sample a	Number of Samples	Mode of the Probability Distribution of Y Given a	Estimated Number of Errors $\hat{M} = E(M\|a)$
0	81	0	8.44
1	253	10	17.87
2	440	20	27.31
3	492	30	36.75
4	395	40	46.19
5	215	50	55.62
6	79	60	65.06
7	36	70	74.50
8	8	80	83.94
9	0	90	93.37
10	1	100	102.81
Total	**2000**	**30.3**	**37.1**

To illustrate the estimation process, we will look at one of the 2000 samples. Suppose this sample includes zero error items from stratum 1 and 4 error items from stratum 2.

For stratum 1, we calculate $\Pr(M_1 | a_1 = 0)$ using equation (2) and $\Pr(Y_1 | a_1)$ using equations (3) and (4). The probability distribution is presented in Figure 2 and Table 4. For stratum 2, the probability distribution is presented in Figure 3 and Table 5.

From Table 4, it is easy to calculate the estimated number of errors in the population:

$$\hat{M}_1 = E(M_1 | a_1 = 0) = 8.44$$

The confidence interval at a 90% confidence level corresponding to a_1=0 is (0, 20) based on the criterion of minimum expected length.

Correspondingly, the estimated error amount is

$$\hat{Y}_1 = \hat{M}_1 \bar{X}_1 = 8.44 \times (279{,}024 \div 300) = \$7{,}847$$

The confidence interval for the error amount at a 90% confidence level is ($0, $18,602).

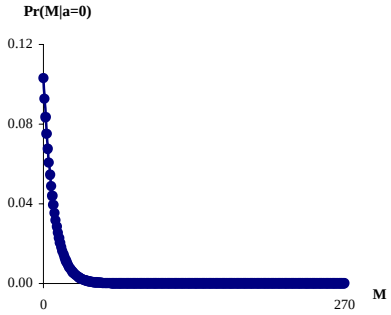

Figure 2. Probability Distribution of M Given a=0 (Stratum 1)

Table 4. Probability Distribution of Number Of Errors/Error Amount in the Population Given None Error in the Sample (STRATUM 1)

Number of Errors In the Population M_1	Appropriate Error Amount Corresponding to M_1 Y_1	Probability $\Pr(M_1 \| a_1 = 0)$ or $\Pr(Y_1 \| a_1 = 0)$	Cumulative Probability $\sum_{M_1} \Pr$ or $\sum_{y_{11}} \Pr$
0	0	0.1030	0.1030
1	930	0.0927	0.1957
2	1,860	0.0834	0.2791
3	2,790	0.0750	0.3541
4	3,720	0.0674	0.4215
5	4,650	0.0606	0.4821
6	5,580	0.0544	0.5365
7	6,511	0.0489	0.5854
8	7,441	0.0439	0.6292
9	8,371	0.0394	0.6686
10	9,301	0.0353	0.7039
11	10,231	0.0317	0.7356
12	11,161	0.0284	0.7639
13	12,091	0.0254	0.7893
14	13,021	0.0228	0.8121
15	13,951	0.0204	0.8325
16	14,881	0.0182	0.8507
17	15,811	0.0163	0.8670
18	16,741	0.0146	0.8816
19	17,672	0.0130	0.8946
20	18,602	0.0116	0.9062
⋮	⋮	⋮	⋮
270	251,122	0.0000	1.0000

Similarly, for stratum 2, as summarized in Figure 3 and Table 5, we have

$$\hat{M}_2 = E(M_2 \mid a_2 = 4) = 46.19$$

$$\hat{Y}_2 = \hat{M}_2 \, \overline{X}_2 = 46.19 \times (510{,}506 \div 300) = \$78{,}597$$

The 90% confidence interval is (4, 70) for M_2 and ($6,806, $119,118) for Y_2.

To obtain the overall estimate of $M = M_1 + M_2$ and $Y = Y_1 + Y_2$ across all strata, we calculate the distributions M and Y given $a_1 = 0$ and $a_2 = 4$ using equations (5) and (7). Then it is straightforward to find the mean and the confidence interval using the probability distribution.

6. Conclusion

In rare error situations, it is likely that no or only a few non-error items appear in the sample. In this situation, classical methods either do not apply or can not provide an appropriate estimate and confidence interval. The hypergeometric method described above is a good solution, especially when a conservative estimate and confidence interval is desired.

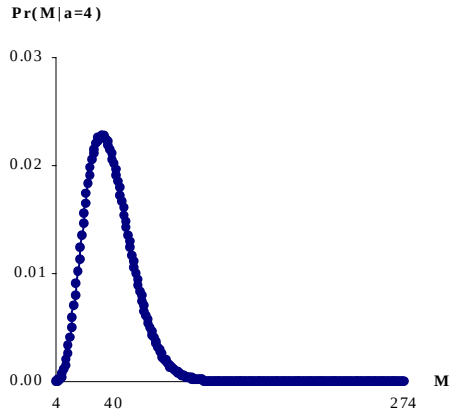

Figure 3. Probability Distribution of M Given a=4 (Stratum 2)

Table 5. Probability Distribution of Number Of Errors/Error Amount in the Population Given Four Error in the Sample (STRATUM 2)

Number of Errors In the Population M_2	Appropriate Error Amount Corresponding to M_2 Y_2	Probability $\Pr(M_2 \mid a_2 = 4)$ or $\Pr(Y_2 \mid a_2 = 4)$	Cumulative Probability $\sum_{M_2} \Pr$ or $\sum_{y_2} \Pr$
4	6,807	0.0000	0.0000
5	8,508	0.0000	0.0000
6	10,210	0.0001	0.0002
7	11,912	0.0002	0.0004
8	13,613	0.0004	0.0008
9	15,315	0.0007	0.0015
10	17,017	0.0010	0.0025
⋮	⋮	⋮	⋮
66	112,311	0.0100	0.8651
67	114,013	0.0094	0.8745
68	115,715	0.0089	0.8835
69	117,416	0.0084	0.8919
70	119,118	0.0079	0.8998
⋮	⋮	⋮	⋮
274	466,262	0.0000	1.0000

Reference

Cochran, W.G. (1977) *Sampling Techniques*, Wiley, New York.

Wilburn, A.J (1984) Practical Statistical Sampling for Auditors. New York & Basel: MARCEL DEFFER, Inc.

Bain, L.J & Engelhardt, M (1991). *Introduction to Probability and Mathematical Statistics*. Boston: PWS-KENT Publishing Company.