DOUBLE CHECKING FOR TWO ERROR TYPES

J.J.A. Moors*

March 8, 1999

Abstract

When auditors have to check large populations of recorded values, they use sampling methods nowadays. From the number of errors found in the sample, an upper confidence level for the fraction of errors in the population can be derived. Thereby, it is assumed that all auditor's checks were faultless.

Auditors may make mistakes, however: errors in the sample may remain unnoticed, a correctly recorded value may be seen as an error by the auditor. Consequently, it is important to check the auditing process itself. In this paper, this is done by checking a subsample of the checked values once more - now by an expert who is assumed to work flawlessly. The numbers of both types of auditor's error have to be combined with the number of errors found in the first sample; from these, an upper confidence limit for the population error fraction has to be derived.

As a first step, the maximum likelihood estimators for the parameters involved are presented here. Then, the desired upper limit can be calculated by similar methods as used in Moors et al. (1997).

Key words: auditing, confidence limit, double inspection, error types, inspection errors, quality control, repeated checks.

JEL codes: C13, C42, C63, M41

1 The model

In a very large population of recorded values, an unknown fraction p_1 is recorded incorrectly. To estimate p_1, an auditor draws a random sample of n values and checks these; he holds X of them for incorrect. (Random variables will be denoted by capitals.)

However, the auditor is not unfallible: with probabilities p_2 and p_4, respectively, he makes the following two errors:

- an incorrectly recorded value is considered correct;
- a correct value is viewed as erroneous.

Note that p_2 and p_4 are in fact conditional probabilities. To take into account these two error possibilities, a random subsample of size m is drawn from the already checked values. The number of values in this subsample, seen as erroneous by the auditor, is denoted by X_1.

An absolute expert now double checks these m values flawlessly. Among the X_1 values, labeled as erroneous by the auditor, he finds Z_1 values to be correct after all; Z_2 values are erroneous indeed. Among the remaining $m - X_1$ values, the expert finds Y_1 new errors - missed by the auditor; W values are correct indeed. The total number of errors found by the expert is $Y = Y_1 + Z_2$. Figure 1 shows the probability tree and the observed numbers, introduced here; Table 1 presents an even simpler overview.

Figure 1. Double checked (sub)sample.

Population	Auditor's check		Expert's check	Numbers	
	correct		correct		
		$1-p_4$		W	
correct					
$1-p_1$					
	incorrect		correct		
		p_4		Z_1	
					X_1
	incorrect		incorrect		
		$1-p_2$		Z_2	
incorrect					Y
p_1					
	correct		incorrect		
		p_2		Y_1	
				m	

Table 1. Double checked (sub) sample.

	Expert		
Auditor	Correct	Incorrect	Total
Correct	W	Y_1	$W + Y_1$
Incorrect	Z_1	Z_2	X_1
Total	$W + Z_1$	Y	m

The following binomial distribution Laws follow immediately:

$\mathrm{L(Y_1)} = B(m, p_1\ p_2)$

$\mathcal{L}(\mathcal{Z}_\infty) = \mathcal{B}[\Updownarrow, (\infty - \sqrt{}^{\infty})\sqrt{}^{\triangle}]$

$\mathcal{L}(\mathcal{Z}_\in) = \mathcal{B}[\Updownarrow, \sqrt{}^{\infty}(\infty - \sqrt{}^{\in})]$

A simpler representation of the joint distribution of the sextet $(W, X_1, Y_1, Y, Z_1, Z_2)$ is:

$$\begin{cases} \mathcal{L}(\mathcal{Y}) = \mathcal{B}(\Updownarrow, \sqrt{}^{\infty}) \\ \mathcal{L}(\mathcal{Y}_\infty|\mathcal{Y}) = \mathcal{B}(\dagger, \sqrt{}^{\in}) \\ \mathcal{L}(\mathcal{Z}_\infty|\mathcal{Y}) = \mathcal{B}(\Updownarrow - \dagger, \sqrt{}^{\triangle}) \\ Z_1, Z_2 independent, conditionallyonY \end{cases} \tag{1.1}$$

Of the original sample, $n - m$ values are checked only once; let X_2 denote the number of errors found by the auditor amoung these. The distribution of $X_2 = X - X_1$ satisfies

$$\begin{cases} \mathcal{L}(\mathcal{X}_\in) = \mathcal{B}[\backslash - \Updownarrow, \sqrt{}^{\infty}(\infty - \sqrt{}^{\in}) + (\infty - \sqrt{}^{\infty})\sqrt{}^{\triangle}] \\ \qquad X_2 independentof(W, X_1, Y_1, Y, Z_1, Z_2) \end{cases} \tag{1.2}$$

Now, (1) and (2) together represent the precise distribution of all random variables involved. In comparison with (3) in Moors et al. (1997), the distribution of Z_1 is added.

2 Point estimators

From the expectations

$E(Y_1) = mp_1p_2,\ E(Z_1) = m(1 - p_1)p_4,\ E(X) = n[p_1(1 - p_2) + (1 - p_1)p_4]$

the moment estimators for p_1, p_2 and p_4 can be found immediately. The moment estimator F_1 for p_1 reads

$$F_1 = \frac{X}{n} + \frac{Y_1 - Z_1}{m} \tag{2.3}$$

It has the curious property that the numbers of the two different errors may compensate each other: if $Y_1 = Z_1$, the estimator reduces to the usual sample fraction of errors. This is not very satisfactory.

To find the maximum likelihood (ML) estimator, the loglikelihood function is derived from (1) and (2). Introduce the probability p_3 that a correct value is found correct indeed in both checks, and the probability p_5 that an incorrect value is considered incorrect indeed throughout:

$$\begin{cases} p_3 = p_1(1 - p_2) \\ p_5 = (1 - p_1)p_4 \end{cases}$$

Then the loglikelihood reads

$$\begin{aligned} &\log L(p_1, p_3, p_5) = c + y_1 \log(p_1 - p_3) + z_2 \log p_3 \\ &+z_1 \log p_5 + w \log(1 - p_1 - p_5) + x_2 \log(p_3 + p_5) \\ &+(n - m - x_2) \log(1 - p_3 - p_5) \end{aligned}$$

It will be assumed first that w, y_1, z_1 and z_2 are positive. Equating the three partial derivatives to 0 leads to the equations for the ML estimates g_i for p_i $(i = 1, 3, 5)$:

$$\left.\begin{aligned} &(a) \quad y_1 g_1 {-} g_3 = w 1 - g_1 {-} g_5 \\ \\ &(b) \quad y_1 g_1 {-} g_3 - z_2 g_3 = x_2 g_3 {+} g_5 - n - m - x_2 1 - g_3 {-} g_5 \\ \\ &(c) \quad w 1 - g_1 {-} g_5 - z_1 g_5 = x_2 g_3 {+} g_5 - n - m - x_2 1 - g_3 {-} g_5 \end{aligned}\right\} \tag{2.4}$$

This system can be solved as follows. First of all, (4a)-(4b)+(4c) reduces to

$$z_2 \; g_5 = z_1 \; g_3 \tag{2.5}$$

while (4a) is equivalent to

$$y_1(1 - g_3 - g_5) = (w + y_1)(g_1 - g_3) \tag{2.6}$$

Substitution of (5) and (6) in the right-hand side of (4b) gives after some simplification

$$x_1 y_1 (n - x) g_3 = x(w + y_1) z_2 (g_1 - g_3) \tag{2.7}$$

Using (5), (4a) can be rewritten as

$$y_1(z_2 - x_1 \; g_3) = (w + y_1) z_2 (g_1 - g_3) \tag{2.8}$$

Finally, combination of (7) and (8) gives

$$g_3 = \frac{x \; z_2}{n \; x_1}$$

This expression even holds for $y_1 = 0$; the only exception is of course the case $x_1 = 0$. Excluding this exception for the moment, the ML estimators for the auxiliary variables become

$$G_3 = XZ_2nX_1, \quad G_5 = XZ_1nX_1 \tag{2.9}$$

In principle, the central estimators can be simply derived:

$$\begin{cases} G_1 &= nX_1Y_1{+}X(WZ_2{-}Y_1Z_1)nX_1(W+Y_1) \\ G_2 &= (n-X)X_1Y_1nX_1Y_1{+}X(WZ_2{-}Y_1Z_1) \\ G_4 &= X(W+Y_1)Z_1nX_1W - X(WZ_2{-}Y_1Z_1) \end{cases} \tag{2.10}$$

Note that for $Z_1 = 0$, the formulae for G_1 and G_2 reduce to expression (6) in Moors et al (1997).

The foregoing derivation breaks down in several cases; they are studied in detail below. Cases (a) and (b) apply to the situation that the complete subsample consists either of correct or incorrect values. Cases (c) and (d) apply to the auditor finding the complete subsample correct or incorrect, respectively. In (a)-(d) is it assumed that exactly two of the four variables W, Y_1, Z_1, Z_2 have value 0; the cases that three of them have value 0 can be derived from these.

$$Case y_1 = z_2 = 0a \tag{2.11}$$

In this situation all values in the subsample are correct; consequently, there is no way to obtain information on p_2. Indeed, the expression for G_2 in (10) does not hold any longer, while G_1 and G_4 can be simplified to

$$G_1 = G_3 = 0, \ G_4 = G_5 = \frac{X}{n}$$

The interpretation is that errors found are considered to be auditor's mistakes.

$$Case w = z_1 = 0b \tag{2.12}$$

Now, no information on p_4 is obtainable; (10) reduces to

$$G_1 = 1 - G_5 = 1, \ G_2 = 1 - G_3 = 1 - \frac{X}{n}$$

The auditor only finds correct values by mistake.

$$Case z_1 = z_2 = 0c \tag{2.13}$$

In this case the expression for G_1 in (10) breaks down. Using the reparametrization

$$\begin{cases} p_6 = p_1 p_2 \\ p_7 = p_3 + p_5 \end{cases}$$

the loglikelihood may be simplified to

$\log L(p_6, p_7) = c + y \log p_6 + (m-y)\log(1-p_6-p_7)$
$+x_2 \log p_7 + (n-m-x_2)\log(1-p_7)$ So, not all parameters p_i can be estimated separately. The ML-equations become

$$\frac{y}{g_6} = \frac{m-y}{1-g_6-g_7} = \frac{x_2}{g_7} - \frac{n-m-x_2}{1-g_7}$$

with the solution

$$G_6 = \frac{(n-X_2)Y}{nm}, \; G_7 = \frac{X_2}{n} \tag{2.14}$$

Some heuristics will be used now to find an estimator for p_1 nevertheless. Since the auditor judges all values - correct or not - in the subsample to be correct, p_2 should be large and p_4 small. Hence, we make the additional assumption

$$p_2 = 1 - p_4 \tag{2.15}$$

Then, (11) leads to

$$G_1 = \frac{Y_1}{m}$$

In this case, only the subsample of size m is used to estimate p_1.

$$Case w = y_1 = 0d \tag{2.16}$$

Now, the loglikelihood may be written as $\log L(p_3, p_5) = c + y \log p_3 + (m-y)\log p_5$
$+ x_2 \log(p_3 + p_5) + (n-m-x_2)\log(1-p_3-p_5)$ The ML-equations become

$$\frac{y}{g_3} = \frac{n-m-x_2}{1-g_3-g_5} - \frac{x_2}{g_3+g_5} = \frac{m-y}{g_5}$$

with the solution

$$G_3 = \frac{(m+X_2)Z_2}{nm}, \; G_5 = \frac{(m+X_2)(m-Z_2)}{nm} \tag{2.17}$$

All values in the subsample are seen as incorrect by the auditor: p_2 should be small and p_4 large. Using assumption (12) once more, this leads to

$$G_1 = \frac{Z_2}{m}$$

Again, uncertainty about p_2 and p_4 leads to discarding the $n-m$ auditor's observations.

It may even occur that three variables of the quartet (W, Y_1, Z_1, Z_2) are zero; such a case may be seen as the pairwise occurrence of (a)-(d). Note that the foregoing solutions are consistent in the sense that both members of such a pair lead to the same solution.

$$Case w = me \tag{2.18}$$

This is (a) $\cap$ (c) with the solution $G_1 = 0$.

$$Case z_1 = mf \tag{2.19}$$

Case (a) $\cap$ (d), $G_1 = 0$.

$$Case z_2 = mg \tag{2.20}$$

Case (b) $\cap$ (d) with solution $G_1 = 1$.

$$Case y_1 = mh \tag{2.21}$$

Case (b) $\cap$ (c) with $G_1 = 1$.

In summary, the ML estimator for p_1 is given by

$$G_1 = \begin{cases} Y_1 m & for X_1 = 0 \\ (n-X)Y_1 n(m-X_1) + X(Y-Y_1)nX_1 & for 0 < X_1 < m \\ Y - Y_1 m & for X_1 = m \end{cases} \tag{2.22}$$

3 Example

To evaluate the behaviour of the estimators F_1 and G_1 the following nummerical example was considered:

$n = 50, \quad m = 20, \quad p_1 = 0.15, \quad p_2 = 0.2, \quad p_4 = 0.1$

From the binomial distribution in (1) and (2), 100000 replications of the vector

$(x_2, \quad y_1, \quad y, \quad z_1, \quad z_2)$

were obtained. For each combination of values, the moment estimate f_1 and the ML-estimate g_1 were calculated. Figure 2 shows a picture of the observed frequency distributions. Table 2 presents some distributional characteristics; the measures for skewness and kurtosis are the third and fourth standardized moments, respectively.

Figure 2a. Simulated distribution of moment estimator F_1.

Figure 2b. Simulated distribution of ML estimator G_1.

Table 2. Characteristics of simulated distributions.

Estimator	Mean	Variance	Skewness	Kurtosis
F_1	0.15013	0.005909	0.1145	3.068
G_1	0.14924	0.005395	0.2373	2.953

Both distributions appear to be quite similar. The ML estimator has a lower variance, mostly due to the negative values that F_1 can take.

The simulations were repeated for $p_1 = 0.15$ with other values of p_1 and p_4, now with 50000 replications. The average values of F_1 and G_1 appeared to be quite close to p_1: for all combination of (p_1, p_2, p_4)-values, the average of F_1 deviated from p_1 at most $4 * 10^{-4}$, reflecting the unbiasedness of F_1. The average G_1-value per simulation run fell short of p_1 throughout, the maximum difference being $5.6 * 10^{-3}$. Table 3 shows the variances of both estimators.

Table 3. Simulated variance of G_1 (and F_1) $* 1000$: $p_1 = 0.15$.

p_4	0	0.05	0.10	0.15	0.2
p_2					
0	3.214	4.048	4.618	4.990	5.304
	(2.538)	(3.733)	(4.890)	(5.894)	(6.776)
0.05	3.487	4.275	4.806	5.170	5.448
	(2.771)	(4.016)	(5.105)	(6.164)	(7.068)
0.1	3.730	4.455	5.024	5.378	5.647
	(2.970)	(4.190)	(5.409)	(6.481)	(7.472)
0.15	3.971	4.725	5.189	5.564	5.762
	(3.210	(4.511)	(5.659)	(6.769)	(7.664)
0.2	4.211	4.910	5.360	5.662	5.857
	(3.427)	(4.730)	(5.868)	(6.872)	(7.912)

For $p_4 \geq 0.1$, G_1 has a lower variance than F_1; for $p_4 \leq 0.05$ on the other hand, F_1 is more accurate.

4 Discussion

Moors et al. (1997) discussed a model for double checking, where the first investigator (the auditor) could only make one possible mistake: missing an error. Of the many possible generalizations mentioned there, one was considered here: the additional possibility is taken into consideration that the auditor finds fault with a correct value.

Both the moment and the ML estimators for the three parameters involved were derived. Note that the ML estimators deviate from the expressions found by Ter Steeg (1998); the explanation is that Ter Steeg based his derivation on the distribution of (X, Y_1, Z_1) only.

The logical next step of course is to find upper confidence levels for the crucial parameter p_1. We plan to do so in the near future.

Acknowledgement

I am much indebted to Leo Strijbosch, both for his thorough reading of the manuscript and for doing the simulations.

References

Moors, J.J.A., B.B. van der Genugten and L.W.G. Strijbosch (1997), Repeated audit controls, CentER Discussion Paper 97113.

Ter Steeg, G.J. (1998), Steekproefcontrole op steekproefcontrole, afstudeerscriptie RUG.