STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 85–106, DOI 10.21307/stattrans-2019-016

APPLICATION OF THE STRATEGY COMBINING MONETARY UNIT SAMPLING AND THE HORVITZ–THOMPSON ESTIMATOR OF ERROR AMOUNT IN AUDITING – RESULTS OF A SIMULATION STUDY

Bartłomiej Janusz[1]

ABSTRACT

Auditors need information on the performance of different statistical methods when applied to audit populations. The aim of the study was to examine the reliability and efficiency of a strategy combining systematic Monetary Unit Sampling and confidence intervals for the total error based on the Horvitz-Thompson estimator with normality assumption. This strategy is a possible alternative for testing audit populations with high error rates. Using real and simulated data sets, for the majority of populations, the interval coverage rate was lower than the assumed confidence level. In most cases confidence intervals were too wide to be of practical use to auditors. Confidence intervals tended to become wider as the observed error rate increased. Tests disclosed the distribution of the Horvitz-Thompson estimator was not normal. A detailed analysis of the distributions of the error amount in the examined real audit populations is also given.

Key words: audit sampling, Monetary Unit Sampling, Horvitz-Thompson estimator, error distribution.

1. Introduction

Audit tests are often based on samples. Auditors can use statistical sampling methods in order to estimate or test hypotheses about the error or the correct (audit) value. In practice, when using statistical sampling, auditors usually estimate confidence intervals for the total error amount or the total audit value of the tested account. This practice allows one to control inference precision as well as sampling risk. Moreover, by comparing confidence bounds with tolerable error or category's value auditors can test hypothesis about the error amount or the correct value of a category.

Systematic Monetary Unit Sampling (MUS) scheme combined with confidence intervals based on the Horvitz-Thompson point estimator of the total error and an assumption of the estimator's asymptotic normality is one of the sampling strategies proposed in the audit literature (Statistical Models and Analysis

[1] TAURON Polska Energia S.A., Operational Audit Unit, Poland. E-mail: bartlomiej.janusz@tauron.pl. ORCID ID: https://orcid.org/0000-0002-9472-9833.

in Auditing, 1989). We will refer to this strategy as MUS HT strategy. In the literature MUS scheme with the Horvitz-Thompson estimator is often called MUS with the mean-per-unit estimator. Auditors select items with probabilities proportional to their book amount mainly because it is assumed that the risk of big error is higher for line items with a higher book value than for line items with smaller values (Arens and Loebbecke, 1981). Systematic sampling is used because of its simplicity and low sampling costs. Confidence intervals based on the Horvitz-Thompson estimator may be applied especially to audit populations characterized by non-trivial error rate, in the case of which other popular audit sampling strategies based on attribute methods do not yield useful outcomes. However, the performance of intervals based on the Horvitz-Thompson statistic for populations with low error rates, which dominate in audit practice, is questionable (Statistical Models and Analysis in Auditing, 1989). Another problem is the convergence to normality of the distribution of the Horvitz-Thompson estimator when systematic MUS is used.

The aim of the conducted simulation study was to verify reliability and effectiveness of the MUS HT strategy. The simulation was based on real data sets containing annual inventory results. The error rate in analysed sets was higher than for usual audit populations described in the literature. In the case of such populations, the examined sampling strategy should perform better than other popular audit sampling strategies, for example strategies based on attribute sampling. Additionally, we analysed the performance of MUS HT for generated populations with low error rates. Reliability was measured by comparing actual confidence levels to nominal confidence levels. Effectiveness was evaluated by comparing intervals' length to the population total book amount. We conducted our study for samples of size 50 and 100. For these sample sizes we examined normality of point estimator distribution.

The results of the simulation conducted on real audit data may be useful for audit practitioners as well as those who are occupied with applications of statistical methods in auditing. It can contribute to the identification when the strategy can yield positive outcomes and when it should not be used.

We also give an analysis of the distributions of the error amount in the examined populations. The majority of such analysis described in the literature is based on audit samples due to unavailability of accounting data. As our data sets come from a full study the results may also help auditors in better understanding the error amount distributions in accounting populations.

2. Error distribution in audit populations

The efficiency and reliability of the applied sampling strategy can be strongly influenced by characteristics of audited populations, especially by the distribution of errors. Arens and Loebbecke (1981) indicate that typical distribution of an absolute error value is characterized by high rate of error free (error value equal to 0) elements. Due to high rate of error free elements it is often assumed in the literature (for example (Statistical Models and Analysis in Auditing (1989))) that the error amount distribution can be modelled as a nonstandard mixture of the distributions. The error amount for *i*th element is treated as random variable of the following type:

$$D_i = \begin{cases} D'_i, & \text{with probabilit y } \nu, \\ 0, & \text{with probabilit y } (1 - \nu), \end{cases} \tag{1}$$

where:

D_i – the error amount for *i*th element – random variable,

$D_i = X_i - Y_i$,

X_i – the book amount for *i*th element – random variable,

Y_i – the audit amount for *i*th element – random variable,

D'_i – random variable different than zero and representing error value,

$\nu = \sum_{i=1}^{N} \frac{d_{0/1i}}{N}$ – error rate in the population,

$d_{0/1i}$ – dummy variable equal to 1 in the case *i*th element contains an error, and 0 in the case *i*th element is error free,

N – population size (line items).

Johnson, Leith, and Neter (1981) studied distribution of errors for accounts receivable and inventories of companies in the United States of America. The authors analysed audit files for 55 companies in the case of accounts receivable and 26 companies in the case of inventories. Their results show high variability of the error rate. Furthermore, the error rate increases with an increase in category value and with an increase in mean value of line items, which is contrary to the general auditors' beliefs that line items with high book amount are precisely verified and thus error probability should be lower. The median of error rate in their study equalled 0.024 for accounts receivable and 0.154 for inventories.

For accounts receivable overstatements (errors for which book value is higher than correct value) dominated significantly. In the case of inventories the number of overstatements and understatements (errors for which book value is lower than correct value) was similar.

Distribution of the error amount in the examined populations differed from normal distribution. High concentration around mean was observed by the authors. Moreover, the distributions were positively skewed and a big number of high value overstatements (exceeding value: mean + 3 x standard deviation) was observed.

A similar study was carried out by Ham, Losell, and Smielauskas (1985) who examined accounts receivable, inventories, accounts payable, sales and purchases. The data used by authors came from the audit files of 5 annual audits for each of 20 companies selected by an audit firm. The median error rate for different categories varied from 0.011 to 0.188 and in the case of inventories it equalled 0.041.

The authors showed that for accounts receivable and sales overstatements prevailed. Understatements dominated for accounts payable and purchases. In the case of inventories the number of overstatements and understatements was similar. For the majority of cases the error amount distribution was not normal for accounts receivable, accounts payable and inventories.

Allen and Elder (2005) analysed 435 sampling applications collected from inventory and accounts receivable during 1994 and 1999. Authors found that in 49% of sampling applications from year 1994 and 46% from year 1999 auditors detected errors.

Durney, Elder, and Glover (2014) indicate that introduction of Sarbanes-Oxley Act in 2002 caused a decrease in error rates and error magnitudes in accounting data in the United States of America. Authors analysed data set of 160 audit sampling applications from audits conducted by a large auditing firm after SOX implementation. The mean misstatement rate (the sum of absolute values of difference between audit and the book amount for line items tested by auditors divided by the sum of the book amount for line items tested by auditors) across all sampling applications in their study was 0.002. In the case of 0.581 examined sampling applications misstatement rate was 0.

The presented representative results regarding the error distributions indicate potential problems with interval estimation mainly due to non-normality and rare error occurrence. Further on we present the results of our simulation study on interval estimation efficiency and reliability. The examination was based on annual inventory results conducted in the plants of international corporation as well as additional sets generated from real populations, characterized by lower frequency of errors.

3. Description of populations being basis for the simulation study

The basis for the examination were sets containing annual inventory results conducted in the 13 warehouses of 6 manufacturing plants of an international corporation. Our populations are not based on results of sample tests but come from a full study of all inventory items in a particular warehouse. The origin of populations is the reason for special character of errors – they were caused only by incorrect registration of stock quantity. The errors were detected by employees of a plant who conducted stock counting and were corrected before conducting audit procedures by auditors.

Stock taking results were in form of files and consisted of records that for each stock item (for example a specific type of springs or specific type of pipes) – line item – contained the following data: stock item description, warehouse, manufacturing plant, quantity according to inventory results, inventory correction, unit cost. For the purpose of the study we made an assumption that the quantities registered after stock taking were correct. We gave a denomination for each population according to the following convention: “plant_type of warehouse”. Plants were numbered from 1 to 6. The type of warehouse was coded using the following capital letters: M – warehouse of materials, PT – warehouse of work in progress, PG – finished goods, Z – all warehouses in a plant.

Distributions of the book amount are similar for all the examined populations. They are highly skewed right and contain outliers – a small number of stock items of a very big book amount. Observations of zero value occur – these are stock items for which quantity registered before inventory equalled zero but during stock taking they were identified in a warehouse. Moreover, distributions are strongly concentrated around values smaller than the mean book amount for stock items.

In the case of the examined warehouses the percentage of stock items containing errors varied from 0.428 to even 0.980 and were very high compared to the studies described in the literature. The conducted analysis did not reveal any relationship between the error rate and either plant or warehouse type. The error rate in the studied populations did not depend on warehouse book amount. No relationship between the book amount of the stock item and the error rate was observed.

In Table 1 we present characteristics of the error distribution in the studied populations.

Table 1. Characteristics of error distribution in studied populations.

Popula tion	Number of stock items	Error rate	Mean error (euro)	Total error amount (euro)	Coefficient of variation	Moment coefficient of skewness	total error amount / total book amount
1_M	2 370	0.931	12.54	29 713.44	4.81	-4.43	0.0103
1_PG	462	0.755	234.33	108 258.69	271.32	3.27	0.1274
1_Z	2 934	0.890	6.07	17 795.75	122.21	-7.44	0.0041
2_M	256	0.946	0.10	26.36	12.53	1.58	0.0000
2_PT	360	0.975	1.29	463.11	13.62	14.88	0.0003
2_Z	648	0.935	0.74	479.25	9.60	19.73	0.0001
3_PT	518	0.894	0.97	499.76	13.94	20.93	0.0003
4_PT	747	0.980	1 606.75	1 200 244.55	7.57	4.63	0.1228
5_M	501	0.603	-531.83	-266 448.65	-213.06	-6.22	-0.0240
5_PT	410	0.751	181.00	74 209.13	-15.76	2.87	0.0286
5_Z	2 615	0.428	-16.49	-43 129.91	15.70	-13.34	-0.0029
6_PT	430	0.837	16.19	6 962.08	58.21	-2.35	0.0050
6_Z	925	0.533	34.26	31 686.49	169.66	-2.54	0.0035

In the case of 11 populations the total error was positive and only in 2 cases it was negative. The number of overstatements exceeded the number of understatements for all the analysed sets. The rate of overstated line items among all items in the error ranged from 0.571 to 0.722. High coefficient of variation together with big differences between the minimum and maximum value of errors indicate high variability of the error amount.

In accordance with audit methodology, auditors are interested in material errors, i.e. errors that can influence the economic decisions taken on the basis of the financial statements. The materiality level, in a simplified way, can be established as a percentage, ranging usually from 0.5% to 2%, of the category's book amount. For all populations we calculated a ratio of the total error amount divided by the total book amount. The absolute value of this ratio in 6 cases exceeded 0.005 and in 4 cases exceeded 0.02. It means that only in the case of 4 populations the error would be material if auditors established the materiality level in a way described above and 2% multiplying factor was used. If auditors used 0.5% multiplying factor 6 populations would be assumed significantly in an error. Only for 2 out of 13 sets the absolute value of the total error divided by the total book amount ratio was higher than 5%. In the case of 4 populations the total error was lower than 500 euro while the book amount exceeded 1.5 million euro.

The conducted analysis did not reveal any significant relationship between the error amount and the book amount of the stock item.

Figures 1 and 2 present typical distributions of the error amount for the examined populations. For better presentation the outliers – stock items with a very big or a very small error amount were not included in the figures.

The distribution of the error amount was strongly concentrated around zero. The number of errors decreased with increasing absolute error amount. This property is typical for all the examined populations. In the case of sets for which the error rates were lower (for example warehouse 5_Z) the observed distribution of the error amount can be described as nonstandard mixture of the distributions given by Eq. (1). An example of such a distribution may be seen in Figure 1. Concentration of the error amount around zero caused low values of mean and median of the error amount. At the same time outliers – stock items with a very big absolute error amount – caused high level of variability of error value measured by standard deviation. Due to low mean values, strong concentration around zero and high values of standard deviations, 87% to 99% of observations laid in the interval: mean +/- standard deviation. In the case of 7 populations the error distribution was positively skewed and in the case of 6 populations it was negatively skewed. The absolute value of moment coefficient of skewness was high for all populations.

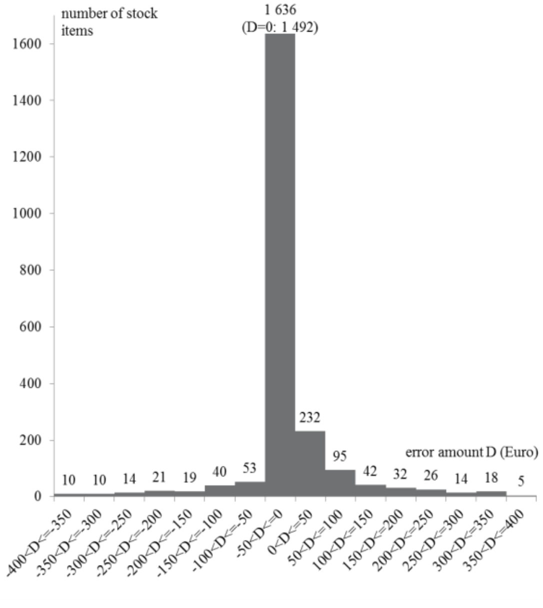

Figure 1. Exemplary distribution of error amount (D) – population 5_Z (outliers* excluded).

* Outliers include 154 items containing errors ranging from - 110 980.64 euro to - 400.14 euro and 194 items containing errors ranging from 406.40 euro to 34 957.03 euro.

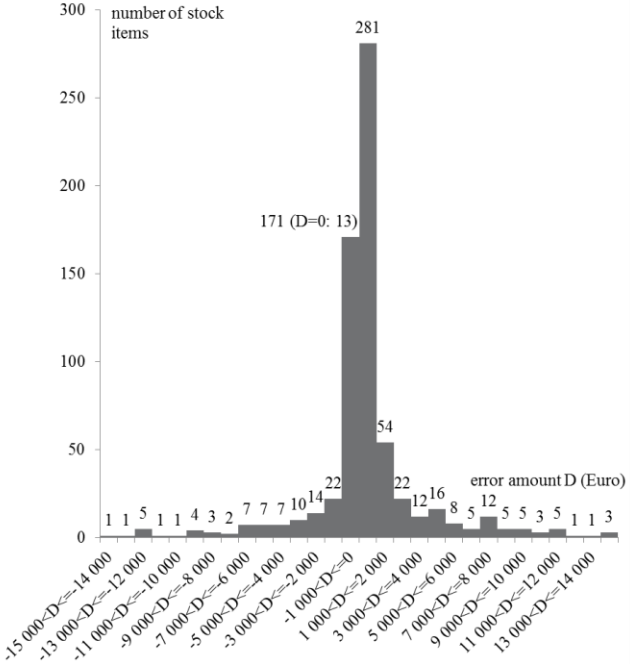

Figure 2. Exemplary distribution of error amount (D) – population 4_PT (outliers* excluded).

* Outliers include 20 items containing errors ranging from - 43 121.09 euro to - 15 321.31 euro and 38 items containing errors ranging from 15 119.99 euro to 127 057.86 euro.

The error rates for all the examined populations were much higher than the error rates identified in other studies on the distribution of errors in accounting populations. In order to test the performance of the sampling strategy for populations with low error rates we generated additional data sets based on selected original populations.

In order to choose populations for creating new data sets we determined the selection criteria taking into account to what extent these criteria allocate the original sets into homogenous, representative groups and what impact they may have on estimation results. We determined five selection criteria:

- variability of the book amount measured with coefficient of variation,
- negative or positive sign of the total error value,
- variability of the error amount measured with coefficient of variation,
- materiality of the total error,
- distribution characteristics of taints (stock item's error amount divided by its book amount).

Using these criteria we chose four populations as the basis for generating additional sets: 1_PG, 3_PT, 5_PT, 5_Z.

On the Basing on results described in the literature we selected three different target error rates: 0.02; 0.07; and 0.15. For each chosen original population we created three sets with different target rates.

For generating purposes we assumed that during stock taking all errors were identified and thus after making the adjustments all data is correct. We generated additional populations in such a way as if stock counting was not fully effective, that means, for randomly selected line items in error no adjustments resulting from stock count were made. Stock items for which errors were not corrected were chosen in such a way that for each erroneous item a number from 0 to 1 was randomly drawn with equal probability. If the number was smaller or equal to the value of quotient of target and the original (value in real set) error rate, the adjustment was not made and an error equal to the original error amount was assigned to line item. The original book value and the audit value were assigned to line item. If the randomly drawn number was higher than the value of quotient of target and the original (value in real set) error rate, the error was corrected and the line item's book value in generated set equalled audit value in the original population.

In such a way we generated 12 additional populations. For each population we used the following notation convention: base population and numbers 2, 7, or 15 depending on the target error rate. For example 1_PG_2 stands for the population generated from set 1_PG with the assumed target error rate equal to 0.02.

In the case of additional populations generated based on set 5_Z, for which the total error amount was negative, total book value was higher than in the case of the original population. For all other generated sets the total book amount decreased comparing to the original populations. Variability of the book amount in generated populations was similar to variability in base sets. For all populations the distribution of the book amount was highly skewed right. No relationship between the book amount of the stock item and the error rate was observed for new populations.

Due to random selection of stock items remaining in the error, the real error rates in the generated populations differed from the target rates. We present the error rates and characteristics of the error amount in Table 2.

Table 2. Characteristics of error in generated populations.

Population	Number of stock items	Error rate	Mean error (euro)	Total error amount (euro)	Coefficient of variation	Moment coefficient of skewness	total error amount / total book amount
1_PG_2	462	0.024	6.18	2 853.26	28.32	18.90	0.0038
1_PG_7	462	0.065	15.32	7 077.14	38.43	14.00	0.0095
1_PG_15	462	0.147	54.63	25 237.29	32.07	11.64	0.0329
3_PT_2	518	0.017	0.00	0.82	31.32	8.32	0.0000
3_PT_7	518	0.077	0.02	11.04	12.93	6.35	0.0000
3_PT_15	518	0.160	0.57	295.32	21.99	22.41	0.0002
5_PT_2	410	0.012	1.56	639.23	-714.10	18.59	0.0003
5_PT_7	410	0.071	59.77	24 504.89	11.96	8.43	0.0096
5_PT_15	410	0.180	169.71	69 581.19	8.26	15.31	0.0269
5_Z_2	2 615	0.022	14.34	37 486.48	23.08	27.92	0.0025
5_Z_7	2 615	0.065	-1.46	-3 830.84	345.56	-20.77	-0.0003
5_Z_15	2 615	0.137	3.81	9 972.03	11.24	-10.52	0.0007

The absolute total error for the majority of new populations decreased significantly in comparison with the base sets. For populations 5_Z_2 and 5_Z_15 the total error was positive in contrary to the original set 5_Z for which it was negative. The only population with the negative total error was 5_Z_7. Despite a significant decrease in the total errors, in the case of 4 generated sets the absolute value of the ratio: the total error divided by the total book amount was higher than 0.005 and for 2 of them it was higher than 0.02. Lower total errors caused lower mean errors and higher coefficient of variation. No significant relationship between the error amount and the book amount of the stock item was observed.

Due to applied generation method, the distribution of the error amount in additional populations can be described as nonstandard mixture of the distributions given by Eq (1). In Figure 3 we present a typical distribution of the error amount for the generated populations. For better presentation outliers were excluded.

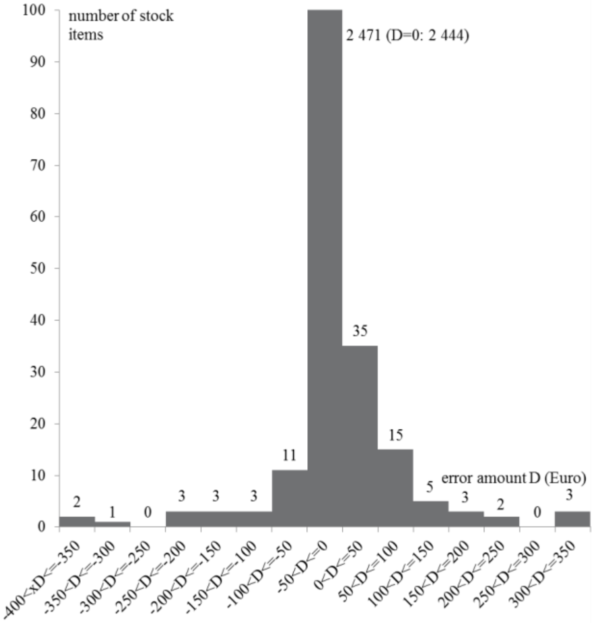

Figure 3. Exemplary distribution of error amount (D) – population 5_ Z_7 (outliers* excluded).

* Outliers include 22 items containing errors ranging from -41 257.73 euro to -485.64 euro and 36 items containing errors ranging from 406.40 euro to 15 031.86 euro.

In addition to a very high number of error free stock items, relatively numerous elements with low error amount occurred. The majority of these errors were overstatements. In the case of 10 populations the number of overstatements was higher than the number of understatements. For the majority of the generated sets the distribution of the error amount was highly positively skewed. For two populations the distribution of the error amount was highly skewed left.

4. Description of the simulation study

The subject of the simulation study was the efficiency and reliability of MUS HT strategy. It is believed that this strategy may work well when the error rate in the audited population is high as it is in the case of our data sets. The actual confidence levels compared to the nominal confidence levels as well as the average length of confidence intervals compared to population the total book amount were used as the main evaluation criteria. Auditors use the total book amount to determine the materiality levels and thus it is useful to compare interval's length with the total book amount for judging estimation efficiency. We calculated the actual confidence levels as a ratio of the number of intervals that contained the total error divided by the number of all estimated intervals.

We conducted the study for samples of size 50 and 100. For each sample size 1,000 samples, from each population (original and additionally generated), were drawn. For each sample confidence interval was calculated.

When conducting substantive testing, that is testing the accuracy of the registration of transactions in book of accounts and correctness of book balances of accounts, auditors usually know book values of line items that make up the account, population of transactions. It allows one to apply such random sampling designs that elements are selected with probabilities proportional to their book amount.

This way of selecting a sample is very useful to auditors because line items with bigger values have bigger probabilities of being selected. Auditors want to test such elements for two reasons. First of all, one of the evaluation criteria of audit works is the ratio of the value of the tested elements to the value of all elements. The higher the value of this ratio, the more complete and accurate the audit is concerned to be.

Furthermore, it is assumed, that the risk of a big error is higher for line items with the higher value than for line items with the small value (Arens and Loebbecke (1981)). Even if this relationship does not hold it was showed (for example Jonhson, Leitch, and Neter (1981) and Neter, Jonhson, and Leitch (1985)) that in the case of audit populations the variability of the error value measured by standard deviation increases for line items with bigger book values. It seems to justify the concentration of works on elements with big book values.

A review of different sampling designs in the case of which first-order inclusion probabilities are proportional or approximately proportional to the book amount can be found in (Tillé (2006), Wywiał (2016)). Because of simplicity and low selection costs auditors often use a systematic sampling scheme, without replacement, proposed by Madow (Sarndal, Swensson, and Wretman (1992)), (Arens and Loebbecke (1981)). We will refer to this sampling scheme described below as systematic MUS (Monetary Unit Sampling). Let population elements be listed in a random order,

$T_{x0}=0,$

$T_{xi}=T_{x(i-1)}+x_i,$

where:

x_i – the book amount of the i^{th} line item, we assume that $x_i>0$ for each $i=1,...,N$;

T_{xi} – the sum of the book amount of line items numbered from 1 to i;

N – the number of elements in a population (line items).

Let the sample size be equal n line items, then sampling interval a equals: $a=T_x/n$, where $T_x = T_{xN}$ – population the total book amount. Number b is drawn with equal probability from the interval (0,a). Sample s consists of the following line items:

$$s = \{i: T_{x(i-1)} < b + (k - 1)a <= T_{xi} \text{ for } k = 1,...,n\}. \tag{2}$$

If for each $i=1,...,N$ $x_i< T_x/n=a$, then we obtain a sample scheme without replacement. Because N is usually significantly bigger than n, and elements with big book values are usually tested separately with probability equal to 1, the condition $x_i< T_x/n$ is assumed to be easily fulfilled in the case of audit populations.

For systematic MUS the sampling interval does not have to be an integer. That is why the number of elements in a sample is fixed and equals n. First-order inclusion probability for the i^{th} element equals (Wolter (1985)):

$$\pi_i = \frac{nx_i}{T_x}. \tag{3}$$

One can see that the second-order inclusion probabilities depend not only on the book value of elements but also on their order in the population. One of the ways to solve this problem is to replace exact values of second-order inclusion probabilities with their approximations. Wolter (1985) proposes, after Hartley and Rao (1962), the following approximation of second-order inclusion probabilities:

$$\pi'_{ij} = \frac{n-1}{n}\pi_i\pi_j + \frac{n-1}{n^2}(\pi_i^2\pi_j + \pi_i\pi_j^2) - \frac{n-1}{n^3}\pi_i\pi_j\sum_{k=1}^{N}\pi_k^2 . \tag{4}$$

The approximation is correct to order $O(N^{-3})$, if the following two conditions hold:

(i) elements are listed in random order; (ii) π_i is order $O(N^{-1})$, (Wolter (1985)).

The (ii) condition holds as

$$\pi_i = \frac{nx_i}{T_x} \le \frac{nx_i}{Nx_{\min}} \le \frac{1}{N}\frac{nx_{\max}}{x_{\min}} \quad \text{and} \quad \frac{nx_{\max}}{x_{\min}} = const \tag{5}$$

where:

$x_{min} = min(x_i,$ for $k = 1,...,n)$ - minimum book amount over all population elements,

$x_{max} = max(x_i,$ for $k = 1,...,n)$ - maximum book amount over all population elements.

In audit practice, unless contrary evidence exists, the natural order of the population is accepted as one possible random ordering. We assumed that the auditor would not randomize the analysed populations before sampling as there is no evidence that a relationship between the population original order and the error amount exists. Thus, the sets were not randomized before the sample selection. Such an approach may significantly reduce the sampling space. Sampling distribution for systematic selection without randomization differs from distributions used for evaluation purposes, which may lead to differences between actual and the assumed confidence level (Hoogduin, Hall, Tsay, Pierce (2015)).

We used the following confidence interval based on an assumption of asymptotic normality of point estimator proposed by Horvitz and Thompson (1952):

$$\left(t_{d\pi} - u_{\alpha/2} v^{1/2}\left(t_{d\pi}\right); t_{d\pi} + u_{\alpha/2} v^{1/2}\left(t_{d\pi}\right)\right) \tag{6}$$

where:

$t_{d\pi} = \sum_{i \in s} \frac{d_i}{\pi_i}$ – the Horvitz-Thompson estimator of the total error amount (T_d), (7)

d_i – the error amount for i^{th} element, $d_i = y_i – x_i$; y_i - audit amount for i^{th} element,

$$v\left(t_{d\pi}\right) = \frac{1}{n-1} \sum_{i}^{n} \sum_{i<j}^{n} \left(1 - \pi_i - \pi_j + \sum_{k=1}^{N} \frac{\pi_k^2}{n}\right)\left(\frac{d_i}{\pi_i} - \frac{d_j}{\pi_j}\right)^2$$

– estimator of variance of $t_{d\pi}$, (8)

$u_{\alpha/2}$ – number for which: $\Phi(u_{\alpha/2}) = 1 – \alpha/2$,

Φ – cumulative distribution function of standardized normal distribution N(0,1).

The estimator given by Eq. (8) was obtained by Wolter (1985) from the well-known estimator of variance of the Horvitz-Thompson total value estimator due to Yates and Grundy (1953):

$$v_{wk}\left(t_{d\pi}\right) = -\frac{1}{2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{d_i}{\pi_i} - \frac{d_j}{\pi_j}\right)^2 , \tag{9}$$

by replacing π_{ij} with their approximations given by Eq. (4). Tillé (2006) proposes several variance estimators requiring knowledge of only first inclusion probabilities. These variance estimators are constructed on the basis of variance approximations.

The study conducted by Christensen, Elder, and Glover (2015) revealed that the confidence level required in the case of substantive tests ranges from 30% to 96%. For all firms the high end of confidence range was consistently at or near 95%. In order to check how the examined strategy performs in the most demanding situation, we applied 95% nominal confidence level for all confidence intervals computed in the simulation study.

In the case of systematic MUS the scheme elements with 0 book amount have zero inclusion probability and are excluded from sampling population. Furthermore, as mentioned above, elements with “very big book amount”, which means elements for which $x_i \geq T_x / n$ are rejected from the sampled population and usually form a stratum which is being subject to full testing. Rejection of elements with "very big book amount" causes a decrease in the total book amount T_x of sampled population and thus further elements may not fulfil the "new" condition $x_i < T_x / n$ and must be rejected from sampled population. It did not allow to apply systematic MUS in the case of the following 14 populations:

- sample size 50 – 2_PT
- sample size 100 – 1_PG, 1_PG_2, 1_PG_7, 1_PG_15, 2M, 2_PT, 2_Z, 5_M, 5_PT, 5_PT_2, 5_PT_7, 5_PT_15, 6_PT.

The subject of estimation in the simulation study was the total error amount of the "modified" populations obtained after excluding elements with 0 book amount and elements with "very big book amount".

When selecting samples from populations with low error rates it is possible that for some samples no error occurs. For such samples the length of interval given by Eq (6) equals zero. It is perceived by auditors as a serious drawback of the analysed strategy. For all populations and sample sizes the rate of intervals with zero length was calculated. We took into consideration such intervals while computing coverage percentage and excluded them for calculation of mean length of confidence interval.

5. Simulation study results

In Table 3 we present the true level of confidence, mean length of confidence interval, mean distance between point estimator and the total error value and the rate of zero length intervals (corresponding to the rate of error free samples) for the sample size 50.

For 10 out of 24 populations coverage percentage was higher than the nominal value (95%). For 1 population (1_PG_7) coverage percentage equalled 1. The lowest coverage ratio amounted to 0.035 in the case of population 3_PT_15.

Only for 6 populations the mean length of confidence interval was lower than 0.005 of the book amount. In the case of 6 sets the mean length of interval was higher than the population book amount. Such wide confidence intervals are in practice useless for auditors.

The value of the mean distance between the point estimator and the total error value divided by the population book amount only in the case of 8 sets was lower than 0.02, among these, in the case of 6 sets it was lower than 0.005. For 7 populations the value of this quotient exceeded 0.1. In the case of all the examined sets the value of the mean distance between the point estimator and the total error value was lower than the mean length of the confidence interval.

In the case of 8 populations zero length intervals occurred, which means that samples with no errors were present. All cases related to the generated populations with lower error rates. The highest percentage of such intervals – 0.769 – occurred for set 1_PG_2. No zero length intervals occurred for populations with the non-trivial error rate.

Table 3. True level of confidence, mean length of confidence interval, mean distance between point estimator and total error value, rate of zero length intervals – sample size 50.

Population	Coverage percentage – true level of confidence	mean length of confidence interval / population book amount	mean distance (total error value – point estimator value) / population book amount	Rate of intervals length = 0
1_M	0.920	0.316	0.080	0.000
1_PG	0.438	2.011	0.374	0.000
1_PG_2	0.231	0.122	0.045	0.769
1_PG_7	1.000	0.824	0.156	0.000
1_PG_15	0.999	1.146	0.183	0.000
1_Z	0.927	0.313	0.068	0.000
2_M	0.998	0.002	0.000	0.000
2_Z	0.728	0.001	0.000	0.000
3_PT	0.355	0.001	0.000	0.000
3_PT_2	0.597	0.000	0.000	0.370
3_PT_7	0.621	0.000	0.000	0.077
3_PT_15	0.035	0.001	0.000	0.002
4_PT	0.884	0.308	0.060	0.000
5_M	0.997	0.727	0.111	0.000
5_PT	0.589	3.830	1.117	0.000
5_PT_2	0.892	0.209	0.031	0.108
5_PT_7	0.998	1.258	0.197	0.001
5_PT_15	0.999	1.727	0.212	0.000
5_Z	0.987	0.322	0.073	0.000
5_Z_2	0.752	0.040	0.008	0.203
5_Z_7	0.919	0.086	0.019	0.002
5_Z_15	0.991	0.148	0.033	0.000
6_PT	0.995	3.942	0.555	0.000
6_Z	0.989	0.195	0.033	0.000

Taking into account both evaluation criteria: the actual confidence level and the mean length of interval, the only population for which the analysed strategy performed well was set 2_M. It should be also noted that in the case of this population there were no samples free of error. For all other data sets the performance of MUS HT strategy was unacceptable because of either too low coverage ratio or too long intervals.

As discussed above, the reasons for applying sampling schemes for which selection probability is proportional to the book amount are among others alleged growth of risk of a big error as well as an increase in variability of the value of errors with an increase in the book amount of line items.

For the examined populations no relationship between the error amount and the book amount of the stock item was observed. We carried out an additional analysis in order to verify if the variability of the value of errors increases for elements with bigger book values. For this purpose, we ordered stock items in each examined population with growing book amount. We divided population size (N) by 10 and rounded the obtained result to the nearest integer ($N/10_{rounded}$). Next, we divided the set into 10 strata in such a way that to the first stratum $N/10_{rounded}$ stock items with highest book amount were assigned, to the second stratum next $N/10_{rounded}$ stock items with highest book amount were allocated and so on until the ninth stratum. The tenth stratum contained elements with lowest book amount that were not assigned to previous nine strata. For each stratum we calculated standard deviation of the error amount.

For all sets for which the coverage percentage was greater than or equal to the nominal confidence level, the variability of the error amount generally increases with an increase in the book amount. However, in some of these populations this trend is not so obvious. Furthermore, in the case of, 4 out of 5 populations for which the true confidence level was lower than 0.5 such relationship did not occur. For 2 data sets for which true level of confidence was lower than the assumed level of confidence - populations 5_PT and 5_Z_2 - the variability of the error amount generally increases with an increase in the book amount.

5.1. Increase in sample size effect

In Table 4 we present the true level of confidence, the mean length of confidence interval, the mean distance between point estimator and the total error value and the rate of zero length intervals for samples size 100.

Table 4. True level of confidence, mean length of confidence interval, mean distance between point estimator and total error value, rate of zero length intervals – sample size 100.

Population	Coverage percentage – true level of confidence	mean length of confidence interval / population book amount	mean distance (total error value – point estimator value) / population book amount	Rate of intervals length = 0
1_M	0.922	0.275	0.059	0.000
1_Z	0.929	0.276	0.055	0.000
3_PT	0.936	0.017	0.004	0.000
3_PT_2	0.926	0.000	0.000	0.035
3_PT_7	1.000	0.001	0.000	0.000
3_PT_15	0.180	0.015	0.003	0.000
4_PT	0.914	0.976	0.625	0.000
5_Z	0.997	0.587	0.121	0.000
5_Z_2	0.924	0.104	0.015	0.045
5_Z_7	0.999	0.282	0.048	0.000
5_Z_15	1.000	0.318	0.046	0.000
6_Z	0.991	0.676	0.096	0.000

An increase in sample size from 50 to 100 caused a higher coverage percentage for all the studied populations. Only for one set the true confidence level was lower than 0.9 and amounted to 0.180 (3_PT_15). For this population no increase in the error amount variability with an increase in the book amount was observed. In the case of 5 sets coverage ratio was greater than or equal to nominal confidence level.

An increase in sample size caused a decrease in estimator variance but still “very long” intervals occurred. For 3 populations mean length of confidence interval was higher than 0.5 of the population book amount. In the case of 4 out of 12 sets the mean length of interval was lower than 0.02 of population the total book amount, among these, in the case of 2 sets it was lower than 0.005 of population total book value.

For 7 out of 12 populations an increase in the sample size caused an increase in the ratio: the mean distance between point estimator and the total error divided by the population book amount. In the case of 5 sets this quotient was lower than 0.02, among these, in the case of 4 populations it was lower than 0.005. For all the examined sets from which 100 item samples were drawn, the value of mean distance between the point estimator and the total error value was lower than the mean length of the confidence interval.

In the case of 2 populations: 3_PT_2 and 5_Z_2, zero length intervals occurred.

Taking into account both evaluation criteria: actual confidence level and mean length of interval, the only population for which the analysed strategy performed well was set 3_PT_7. It should be also noted that in the case of this population there were no error free samples.

5.2. Error rate effect

We did not observe a relationship between the error rate and the true confidence level. For groups of populations 3_PT_50 and 1_PG_50 a sharp decrease in the coverage percentage with an increase in the nominal error rate from 0.07 to 0.15 can be observed. Furthermore, in the case of group of populations 5_PT_50 the true confidence level for the original set is the lowest, for group of populations 1_PG_50 it is significantly lower than for sets with the nominal error rate 0.07 and 0.15. Finally, in the case of group of populations 3_PT_50 the coverage percentage for the original set is much lower than for sets with the nominal error rate 0.02 and 0.07. Figure 4 presents changes in coverage percentage with change of error rates for these groups of populations.

For group 3_PT_100, not presented in Figure 4, changes in actual confidence level with changes in the error rate had the same pattern as in the case of group 3_PT_50. An upward trend of the true confidence level with an increase in the error rate occurs only in the case of 2 groups of populations: 5_Z_50 and 5_Z_100 – groups not presented in Figure 4.

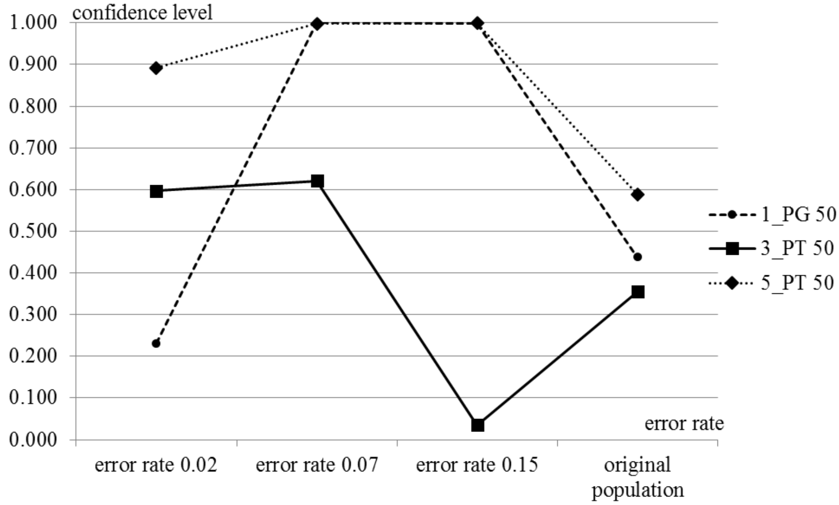

Figure 4. Relationship between error rate and true confidence level.

In contrast, it can be observed that together with the growth of the error rate the mean length of confidence interval increases. This relationship is presented in Figure 5.

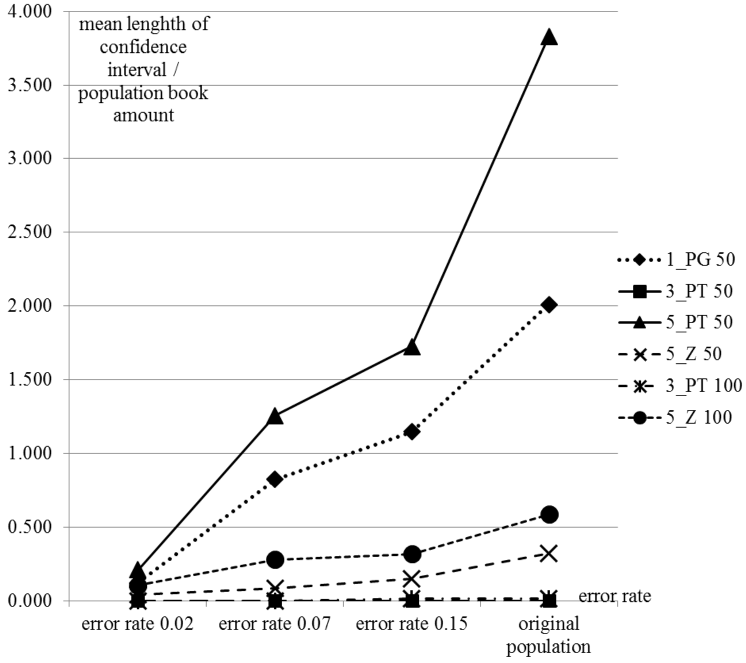

Figure 5. Relationship between error rate and length of confidence interval.

5.3. Normality of the distribution of point estimator

Examined confidence intervals were based on an assumption of asymptotic normality of the Horvitz-Thompson point estimator. Hájek (1964) derived conditions for asymptotic normality of the Horvitz-Thompson estimator in the case of rejective sampling under assumptions that n → ∞ and N - n → ∞. The author proposed two estimators of the Horvitz-Thompson estimator variance. Berger (1998) showed that if divergence between a given sampling design and rejective sampling design goes to zero then the Horvitz-Thompson estimator has an asymptotic normal distribution. The author gave also the rate of convergence of the Horvitz-Thompson estimator for any kind of sampling.

On the basis of the simulation results we tested if the assumption of normality of the Horvitz-Thompson estimator holds for statistic

$$st_{d\pi} = \frac{t_{d\pi} - T_d}{v^{1/2}\left(t_{d\pi}\right)}. \tag{10}$$

We did not conduct evaluation of the necessary sample size to ensure the sufficient convergence of the estimator to the normality. Such evaluation was discussed by Wywiał (2016). We verified normality of statistic given by Eq (10) with the Shapiro – Wilk W-test. It should be taken into account that this test has very high power.

In order to calculate the test statistic W we used approximation for the required coefficients proposed by Royston (1992). According to the author the approximation is accurate to ±1 in the fourth decimal place. Furthermore, we applied Royston's (1992) normalizing transformation for the W statistic.

Only in the case of population 5_Z_100 p – the value was higher than 0.05. For the remaining populations p – the values were very low: only in 2 cases they differed substantially from zero but were much lower than 0.05. One reason for this may be the observed high skewness of the analysed populations (Statistical Models and Analysis in Auditing (1989)). The conducted analysis showed that for different sets $st_{d\pi}$ statistic had very different distributions that cannot be attributed to one or specified group of the distribution types. One way of solving the problem of non-normality may be applying the bootstrap procedure.

6. Results of other studies

Our results are similar to other studies' results. Sampling strategy using systematic MUS scheme and the confidence interval based on the Horvitz-Thompson point estimator was subject of the simulation conducted by Neter and Loebbecke (description and results of the study are given following (Statistical Models and Analysis in Auditing (1989))). Four audit populations, for which the error rate was 0.3 were used in this study. Additionally, based on the original sets, 16 populations with lowered error rates from 0.3 to: 0.005, 0.01, 0.05, and 0.1 were generated. The examination was conducted for 14 of these sets. Six hundred samples of size 100 were drawn from each examined population. The true confidence level was calculated as percentage of intervals that covered the total error.

For none of the populations the true confidence level reached nominal value of 95.4%. The lowest coverage percentage equalled 5.2% for one of sets with the error rate 0.005. For two populations with error rates: 0.1 and 0.3 the true confidence level reached the highest value equal to 94.5.

These results are consistent with results obtained by Dworin and Grimlund (1984), who compared the reliability of the proposed new method of interval estimation called moment bound with the mean-per-unit estimator combined with MUS scheme. The performance of one-sided confidence interval (upper bound) for 128 inventory populations was analysed in the study. Their results show that only in the case of 13 populations the coverage rate for the mean-per-unit method reached or exceeded 95% nominal confidence level. The lowest coverage rate equalled 72.6%.

Kim Neter and Godfrey (1987) analysed reliability and efficiency of upper bound based on the mean-per-unit estimator and MUS Cell Sampling - a two-stage sampling scheme. According to the authors the MUS Cell Sampling scheme can be treated as an easy to use alternative to systematic sampling of monetary units. The reliability was measured by coverage ratio while efficiency was measured by relative tightness - mean bound in the replications expressed relative to the total error amount in a sampled population. The average coverage over all 64 study populations for bound based on the mean-per-unit estimator and MUS Cell Sampling equalled 76.7% while the minimum coverage was 23% and the maximum 94.2%. Average relative tightness equalled 1.75. The minimum and maximum value of this measure was 1.29 and 2.71 respectively.

Marazzi and Tillé (2017) conducted a simulation study in which MUS with the Horvitz-Thompson point estimator was compared with other sampling strategies. Authors did not analysed the confidence intervals but only the estimators' mean standard error. Their results show a relatively high empirical mean standard error in the case of MUS with the Horvitz-Thompson point estimator strategy.

7. Conclusion

The purpose of the simulation study was to examine the efficiency and reliability of interval estimation for the MUS HT sampling strategy. The main evaluation criteria included actual confidence levels compared to nominal confidence levels as well as the average length of confidence intervals compared to the population total book amount. The basis for the examination were sets containing annual inventory results as well as additional, generated populations with lower error rates.

For the majority of populations the percentage of intervals that covered the total error amount was lower than the nominal confidence level. The obtained results show that for all populations for which the coverage percentage was greater than or equal to the nominal confidence level, the variability of the error amount increases with an increase in the elements' book value. The sample size growth had a positive effect on the coverage percentage. No relationship between the error rate and the actual confidence level was found. The observed non-normality of the Horvitz-Thomson point estimator standardized by its estimator of standard deviation, for applied sample sizes, may be one of the reasons for lower than the assumed true confidence levels.

For most cases, the length of the btained confidence intervals make them useless for auditors. In the case of some populations, the mean length of interval was higher than the population book amount. An increase in sample size caused a decrease in estimator variance but still "very long" intervals occurred. We observed that together with growth of the error rate the mean length of confidence interval increased.

Taking into consideration both evaluation criteria: the actual confidence level and the mean length of interval, the MUS HT strategy performed well only in two cases. Taking into account that for the majority of the used populations the error rates were very high, our results are in contrary to the belief that the analysed sampling strategy may be useful for populations with high error rates. The fact that the length of intervals increases with growth of the error rate seems to strengthen this conclusion.

The obtained results are consistent with results of other simulation studies on MUS HT strategy.

It must be, however, stressed that the applied approach, consistent with the typical auditors' way of using systematic MUS, assuming lack of randomization of populations before sample selection might significantly reduce the sampling space and thus might have a substantial impact on the obtained results.

One disadvantage of the systematic MUS revealed by the study is inability to apply this scheme to populations composed of elements with "very big book amount". Rejection of elements with "very big book amount" causes a decrease in the total book amount of the sampled population and thus further elements must be rejected because their book value is higher than the "new" sampling interval. In the case of 14 populations it did not allow for the application of the systematic MUS scheme.

The analysis of real accounting data sets showed that the distribution of the book amount is strongly concentrated around values smaller than the mean book amount, highly skewed right and contains outliers.

We did not observe a significant relationship between either the book amount of the stock item and the error rate or the book amount and the error amount of the stock item. The distribution of the error amount was strongly concentrated around zero. With increasing absolute error amount the number of errors decreased. The outliers caused a high level of variability of the error value measured by standard deviation. The number of overstatements exceeded the number of understatements for all the analysed sets. The absolute value of moment coefficient of skewness was high for all populations.

REFERENCES

ALLEN, R. D., ELDER, R. J., (2005). A Longitudinal Investigation of Auditor Error Projection Decisions, Auditing: A Journal of Practice & Theory, 24 (2), pp. 69–84.

ARENS, A. A., LOEBBECKE, J. K., (1981). Applications of Statistical Sampling to Auditing, Englewood Cliffs: Prentice – Hall, Inc.

BERGER, Y. G., (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. Journal of Statistical Planning and Inference, 67 (2), pp. 209–226.

CHRISTENSEN, ELDER, GLOVER, (2015). Behind the Numbers: Insights into Large Audit Firm Sampling Policies, Accounting Horizons, 29 (1), pp. 61–81.

DURNEY, M., ELDER, R. J., GLOVER, S. M., (2014). Field Data on Accounting Error Rates and Audit Sampling, Auditing: A Journal of Practice & Theory, 33 (2), pp. 79–110.

DWORIN, L., GRIMLUND, R. A., (1984). Dollar Unit Sampling for Accounts Receivable and Inventory, The Accounting Review, LIX(2), pp. 218–241.

HAM, J., LOSELL, D., SMIELIAUSKAS, W., (1985). An Empirical Study of Error Characteristics in Accounting Populations, The Accounting Review, LX (3), pp. 387–406.

HARTLEY, H. O., RAO, J. N. K., (1962). Sampling with Unequal Probabilities and without Replacement, The Annals of Mathematical Statistics, 33 (2), pp. 350–374.

HÁJEK, J., (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population, The Annals of Mathematical Statistics, 35 (4), pp. 1491–1523.

HOOGDUIN, L. A., HALL, T. W., TSAY, J. J., PIERCE, B. J., (2015). Does Systematic Selection Lead to Unreliable Risk Assessments in Monetary – Unit Sampling Applications? Auditing: A Journal of Practice & Theory, 34 (4), pp. 85–107.

HORVITZ, D. G., THOMPSON, D. J., (1952). A Generalization of Sampling Without Replacement From a Finite Universe, Journal of the American Statistical Association, 47 (260), pp. 663–685.

JOHNSON, J. R., LEITCH, R. A., NETER J., (1981). Characteristics of Errors in Accounts Receivable and Inventory Audits, The Accounting Review, LVI (2), pp. 270–293.

KIM, H. S., NETER, J., GODFREY, J. T., (1987). Behavior of Statistical Estimators in Multilocation Audit Sampling, Auditing: A Journal of Practice & Theory, 6 (2), pp. 40–58.

MARAZZI, A., TILLÉ, Y., (2017). Using past experience to optimize audit sampling design, Review of Quantitative Finance and Accounting, 49 (2), pp. 435–462.

NETER, J.,. JOHNSON, J. R, LEITCH, R. A., (1985). Characteristics of Dollar – Unit Taints and Error Rates in Accounts Receivable and Inventory, The Accounting Review, LX (3), pp. 488–499.

ROYSTON, P., (1992). Approximating the Shapiro – Wilk W-test for non-normality, Statistics and Computing, 2, pp. 117–119.

SARNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, Springer – Verlag New York, Inc.

STATISTICAL MODELS AND ANALYSIS IN AUDITING, (1989). Panel on Nonstandard Mixtures of Distributions, Statistical Science, 4 (1), pp. 2–33.

TILLÉ, Y., (2006). Sampling Algorithms, Springer Science+Business Media, Inc.

WOLTER, K. M., (1985). Introduction to Variance Estimation, Springer – Verlag New York, Inc.

WYWIAŁ, J. L., (2016). Contributions to Testing Statistical Hypotheses in Auditing, Warszawa: Wydawnictwo Naukowe PWN SA.

YATES, F., GRUNDY, P. M., (1953). Selection without Replacement from Within Strata with Probability Proportional to Size, Journal of the Royal Statistical Society. Series B (Methodological), 15 (2), pp. 253–261.