Contributions to Testing Statistical Hypotheses in Auditing

Janusz L. Wywiał

This monograph was published
by PWN Scientific Publishers in 2016
ISBN 978-83-01-18534-3

Contents

Introduction

Applications of statistical methods in auditing make it possible to save costs of controlled management and accounting systems. Due to this, the number of applications of statistical methods in auditing is gradually increasing. Moreover, it is expected that audit reports are prepared as quickly as possible, and statistical methods can also help with this.

Applications of statistical methods in auditing are considered e.g. in the following monographs: Arens and Lobbecke (1981), *Artificial Intelligence...* (1995), Guy et al. (1986), Leslie et al. (1979), Neter and Loebbecke (1975), Robert (1978). In Polish literature this subject is presented in books by Hołda and Pociecha (2004, 2009), Karliński (2005), Klima (2005), Przybycin and Rojek (1966) and Wywiał (2014a).

However, although the already large amount of statistical literature dealing with auditing problems is constantly expanding, it seems that many auditing problems still need more efficient solutions. This is the main reason why new applications of statistics in auditing are still being searched for. This book contains the results of the author's research into new applications of statistics in auditing. The main purpose of this book is to present the author's new contributions to applying statistical methods in auditing. Due to this, the book is not a classical monograph about statistical auditing which shows all former theories and applications of statistical methods in auditing. Instead, the author focuses on presenting his own solutions.

All kinds of statistical inference approaches are considered in this book. These approaches are outlined in the first chapter, where some basic models of statistical inference are presented. Moreover, the first chapter also gives an explanation as to why the decision-making process in auditing is treated as a problem of verifying appropriately formulated statistical hypotheses.

The second chapter is devoted to compliance tests. This topic is considered as a way of verifying hypotheses on the total number of errors or the fraction of errors. The classical Neyman-Pearson and Bayesian approaches to testing statistical hypotheses are taken into account. We present formal methods of evaluating the necessary sample size under assumed auditing risks.

Usually these risks are called the risk of incorrect rejection (of the auditing system) and the risk of incorrect acceptance. When the theory of testing statistical hypotheses is taken into account, risks are represented by the significance level of the test or the probability of an occurrence of type II error. Several iteration methods as well as the Monte-Carlo method are applied to evaluate the necessary sample size.

The process of making auditing decisions based on so-called substantive tests seems to be more complicated than using compliance tests. That is why this problem is considered in chapters 3-5. Statistical inference methods are supported by the distribution of book values. Those values are treated as observations of the auxiliary variable in the whole population. The audit procedures based on substantive tests are considered as the problem of verifying statistical hypotheses about total (accounting) error.

In the third chapter, the randomization and model approaches are taken into account, but statistical inference is only based on simple random samples or on a set of random samples drawn from the stratified population. Usually, the distribution of variables observed in accounting is not known. Hence, the exact distributions of test statistics are also not known. In this situation, inference has to be based on asymptotic distributions of statistics. More details about this problem are presented in subsections 3.3.2-3.3.4. The well-known Berry-Esséen inequality as well as Monte-Carlo the simulation are used to evaluate the sample size to assure sufficient convergence of test statistics to normal distribution. The proposed simulation algorithms and some results of Monte-Carlo inference are reported in subsection 3.3.5. Under the assumed normal distribution of a test statistic, the necessary sample size is evaluated in subsection 3.3.6 under the risk of incorrect rejection or incorrect acceptance. Regression and ratio estimators, well known in survey sampling, are used as test statistics. In section 3.4 their variances are evaluated under the assumption that accounting error and true observation are independent. The well-known optimal stratification procedures of the support of an auxiliary variable are presented in subsection 3.5.2. They are based on the method of evaluating stratified sample sizes to assure sufficient convergence to the normality of the test statistic based on the stratified sample mean, proposed in the subsection 3.5.3.

The chapter fourth generalizes the considerations from chapter three but uses more complex cases of data observed in the samples drawn by means of so-called monetary (dollar) sampling designs. Monetary sampling designs are characterized by first-order inclusion probabilities, which are approximately proportionate to book values. In this book the monetary-type sampling designs are treated as particular sampling designs characterized by first-order inclusion probabilities that are non-decreasing functions of a positive auxiliary variable.

Some specific sampling designs are presented in section 4.1. Their first-order inclusion probabilities are approximately proportional to book values, as was explained e.g. in the case of sampling designs that are dependent on the function of the auxiliary variable order statistics presented in subsection 4.1.8 and 4.1.9. A sampling scheme based on the gradual left truncation of a positive and discrete auxiliary variable is proposed in subsection 4.1.7. In practical auditing analysis, a systematic sampling scheme with probabilities proportional to book values is preferred. This is

equivalent to the Hartley-Rao sampling scheme as shown in subsection 4.1.4. The important issue of the convergence of the studentized Horvitz-Thompson statistic to (standard) normal distribution is considered in section 4.2. It is shown that such convergence is possible on the basis of the central theorems mainly of Hájek (1964) and Berger (1998). Moreover, the simulation procedures proposed in subsection 3.3.5 can be used to evaluate the necessary sample size to assure sufficient convergence to the normality of the test statistic. The statistical inference based on the Horvitz-Thompson (1952) statistic applied by Cordy (1993) to a continuous population is considered in section 4.3. In this case, the sampling design is defined as the appropriate multivariate density function of continuous observations of an auxiliary variable. In subsection 4.3.2, three original sampling schemes and designs are proposed. The sampling schemes are defined by means of some conditional density functions of the gradually left truncated distributions of an auxiliary variable. The appropriate conditional functions of sampling schemes as well as inclusion probabilities are derived for the cases where the auxiliary variables have exponential and Pareto distribution functions.

The fifth chapter deals with the special issue of inference based on the model approach. The observations of audited values as well as of values contaminated with accounting errors are treated as values of a two-dimensional random variable. Moreover, book values are treated as values of the random variable with a distribution function that is a mixture of the distributions of the two variables previously defined. The model approach allows to apply statistical inference procedures based on the well-known method of moments or the likelihood ratio test for verifying hypotheses on the total error amount. Moreover, the Bayesian approach is also taken into account. In detail, statistical inference is based on book values being observations of a random variable whose distribution is a mixture of two Poisson distributions. In subsection 5.2.2, we show how to obtain estimators of the parameters of the mixture using the method of moments. The proposed test statistics are functions of those estimators. The likelihood ratio test is considered in detail in subsection 5.2.3. Finally, in subsection 5.2.4, the appropriate posteriori distribution of total error is derived. The randomization and model approach are jointly considered in section 5.3. The generalized likelihood function of the sample drawn according to a complex sampling design is defined. The test statistic is determined as a function of the estimators of appropriate parameters derived on the basis of the likelihood function.

The proofs of the theorems or details of the derivations of some expressions are presented in the first section of the Appendix. The author's computer programs, written in R Core Team (2015) programming language are presented in the second section of the Appendix.

The book should be useful for auditors dealing with applications of statistics in auditing. Moreover, it should be valuable for statisticians or students interested in applying statistical methods in practice.

I am very grateful to reviewer Stanisław Heilpern for his valuable comments. Moreover, I am very thankful to Tomasz Żądło for proofreading the manuscript.

This project is supported by the grant from the Polish National Scientific Center DEC-2012/07/B/HS4/03073.

Chapter 1
Basic definitions and notation

1.1 Fixed population approach

Let U be an identifiable population that can be treated as the following sequence of integers: $U = (1, 2, ..., k, ..., N)$. The index k is attached to a k-th considered object, for instance an accounting document or accounting subsystem. In this book three well-known statistical inference procedures will be considered. The first one is the design-based approach, also known as the fixed and finite population (or randomization) approach. The second one is the super-population (model) approach. Finally, the Bayesian inference methods will be studied.

In the randomization approach, the size N of the population U is fixed. Let z be fixed variable whose z_k value is observed on a k-th population element, so $k \in U$. We assume that $z_k = 1$ when the k-th population element is wrong (e.g. when there is an error in the k-th accounting document). The observed k-th population element is free from errors when $z_k = 0$. Hence, the population U is divided into the two following sub-populations (domains): $U_0 = \{k : \ z_k = 0, k \in U\}$ and $U_1 = \{k : \ z_k = 1, k \in U\}$, so $U = U_0 \cup U_1$. The sizes of the domains U_1 and U_0 are $0 \leq N_1 = \sum_{k \in U} z_k \leq N$, $0 \leq N_0 = N - N_1 \leq N$, respectively, and $N_0 + N_1 = N$.

On the basis of the paper: *Statistical Models...* (1989) or the textbook by Wywiał (2014a) we introduce the following notation. The book (or recorded) amounts are observed on all population elements and are denoted by x_k, $k \in U$. These observations are treated as values of the variable x. The audited (corrected) amounts are denoted by y_k, $k \in U$. The error amount (absolute error amount) of the k-th item are denoted by

$$d_k = x_k - y_k, \qquad d_{A,k} = |x_k - y_k|, \qquad k \in U. \tag{1.1}$$

The fraction of error is defined as follows:

$$e_k = \frac{d_k}{x_k}, \quad \text{provided} \quad x_k \neq 0, \quad k \in U \tag{1.2}$$

This is called the tainting (or taint) of the k-th item. Hence:

$$d_k = e_k x_k, \qquad k \in U. \tag{1.3}$$

$$d_k = \begin{cases} 0, & \text{if} \quad k \in U_0, \\ x_k - y_k, & \text{if} \quad k \in U_1, \end{cases}$$

Hence:

$$x_k = \begin{cases} y_k, & \text{if} \quad k \in U_0, \\ y_k + d_k, & \text{if} \quad k \in U_1. \end{cases}$$

The above defined error index can be written as follows:

$$z_k = \begin{cases} 0, & \text{if} \quad d_k = 0, \\ 1, & \text{if} \quad d_k \neq 0, \end{cases} \qquad k \in U. \tag{1.4}$$

The vectors: $\mathbf{x}^T = [x_1 x_2 ... x_N]$, $\mathbf{y}^T = [y_1 y_2 ... y_N]$ and $\mathbf{z}^T = [z_1 z_2 ... z_N]$ are the parameters of the population. Usually, $\mathbf{x} \in R_+^N$ and $\mathbf{y} \in R_+^N$. The vector of error amounts is: $\mathbf{d} = \mathbf{x} - \mathbf{y}$. The population (total) error amount is defined as the following sum:

$$d_U = \sum_{k \in U} d_k$$

The population (total) absolute error amount is defined as the following sum:

$$d_{A,U} = \sum_{k \in U} |d_k|$$

The population mean error amount and the population mean of absolute mean error are defined, respectively, as follows

$$\bar{d}_U = \frac{1}{N} \sum_{k \in U} d_k, \quad \bar{d}_{A,U} = \frac{1}{N} \sum_{k \in U} |d_k|.$$

Let us underline that it is possible that $\bar{d}_U = 0$, when $\bar{d}_{A,U} \neq 0$. But if $\bar{d}_U \neq 0$, then $\bar{d}_{A,U} \neq 0$.

In many practical research studies we can expect that $d_k = x_k - y_k \geq 0$ for all $k \in U$ or $y_k - x_k \geq 0$ for all $k \in U$. The latter case can be met in tax auditing. In both defined cases $d_U = \sum_{k \in U} (x_k - y_k)$ or $d_U = \sum_{k \in U} (y_k - x_k)$, respectively. In general, the k-th error amount d_k can be non-negative or negative. That is why in this book, **in order to simplify consideration we will usually assume that $x_k \geq y_k$ for all $k \in U$**. This lets us write the following:

$$d_U = x_U - y_U$$

where

$$x_U = \sum_{k \in U} x_k, \qquad y_U = \sum_{k \in U} y_k.$$

Totals x_U, y_U, d_U and $m_U = N_1$ will be called the population book amount and the population audited amount, respectively. Moreover, the total (population) number of errors is defined as follows:

$$m_U = \sum_{k \in U} z_k = N_1.$$

Under the above assumption the following total values will also be considered:

$$x_{U_h} = \sum_{k \in U_h} x_k, \quad y_{U_h} = \sum_{k \in U_h} y_k, \quad h = 0,1, \quad x_{U_0} = y_{U_0},$$

Hence, the population error amount can be rewritten as follows:

$$d_U = d_{U_1} = \sum_{k \in U_1} d_k = x_{U_1} - y_{U_1}. \tag{1.5}$$

The population mean book amount, the mean audited amount, the mean error amount and fraction of errors are defined by

$$\bar{x} = \bar{x}_U = x_U/N, \quad \bar{y} = \bar{y}_U = y_U/N, \quad \bar{d} = \bar{d}_U = d_U/N = \bar{x}_U - \bar{y}_U, \quad p = \frac{m_U}{N} = \frac{N_1}{N},$$

respectively. Moreover, the following means will be considered:

$$\bar{y}_{U_h} = y_{U_h}/N_h, \, \bar{x}_{U_h} = x_{U_h}/N_h, \quad \bar{d}_{U_h} = d_{U_h}/N_h, \quad h = 0,1, \quad \bar{d}_{A,U_1} = \frac{1}{N_1} \sum_{k \in U_1} |d_k|, \tag{1.6}$$

$$\bar{d}_{U_0} = \bar{d}_{A,U_0} = 0, \quad \bar{d}_{U_1} = \bar{x}_{U_1} - \bar{y}_{U_1}, \quad \bar{d}_{A,U} = p\bar{d}_{A,U_1}$$

When $\bar{d}_{A,U_h} = 0$, then $\bar{d}_{U_1} = 0$. But it is possible that $\bar{d}_{U_1} = 0$, if $\bar{d}_{A,U_1} \neq 0$.

The decomposition of the population mean values are as follows:

$$\bar{y}_U = (1-p)\bar{y}_{U_0} + p\bar{y}_{U_1}, \quad \bar{x}_U = (1-p)\bar{x}_{U_0} + p\bar{x}_{U_1}, \quad \bar{y}_{U_0} = \bar{x}_{U_0}.$$

Moreover, let us note that

$$d_U = x_U - y_U = \sum_{k \in U_1} d_k = Np\bar{d}_{U_1}, \quad \bar{d}_U = p\bar{d}_{U_1}$$

and

$$g_U = \frac{d_U}{y_U} = p\frac{\bar{d}_{U_1}}{\bar{y}_U}.$$

The following variances will be taken into account:

$$v(x,y) = v_U(x,y) = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{x})(y_k - \bar{y}), \quad v(x) = v(x,x), \quad v(y) = v(y,y)$$

or

$$\sigma(x,y) = \sigma_U(x,y) = \frac{1}{N}\sum_{k\in U}(x_k - \bar{x})(y_k - \bar{y}), \quad \sigma^2(x) = \sigma(x,x) = \frac{N-1}{N}v(x).$$

The correlation coefficient is:

$$\rho(x,y) = \frac{v(x,y)}{\sqrt{v(x)v(y)}} = \frac{\sigma(x,y)}{\sigma(x)\sigma(y)}.$$

The variance of book amounts can be decomposed as follows:

$$\begin{aligned}\sigma^2(x) &= \sigma^2(y) + 2p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d) + p\sigma^2_{U_1}(d) + \\ &\quad + p(1-p)\left(\bar{d}_{U_1} + 2\left(\bar{y}_{U_1} - \bar{y}_{U_0}\right)\right)\bar{d}_{U_1} = \\ &= p\sigma^2_{U_1}(y) + (1-p)\sigma^2_{U_0}(y) + 2p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d) + p\sigma^2_{U_1}(d) + \\ &\quad + p(1-p)\left(\bar{d}_{U_1} + \bar{y}_{U_1} - \bar{y}_{U_0}\right)^2\end{aligned}$$

where

$$\rho_{U_1}(y,d) = \frac{\sigma_{U_1}(y,d)}{\sigma_{U_1}(y)\sigma_{U_1}(d)}$$

$$\sigma_{U_1}(d,y) = \frac{1}{N_1}\sum_{k\in U_1}(y_k - \bar{y}_{U_1})(d_k - \bar{d}_{U_1}),$$

$$\sigma^2_{U_1}(y) = \sigma_{U_1}(y,y), \quad \sigma^2_{U_1}(d) = \sigma_{U_1}(d,d).$$

Hence, the above decomposition shows how the variance of book amounts depends on the parameters of variables y and d. The decomposition of the covariance is as follows:

$$\sigma(x,y) = \sigma^2(y) + p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d) + p(1-p)\left(\bar{y}_{U_1} - \bar{y}_{U_0}\right)\bar{d}_{U_1}.$$

Particularly, if $\bar{d}_{U_1} = 0$, then

$$\sigma^2(x) = \sigma^2(y) + 2p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d) + p\sigma^2_{U_1}(d),$$

$$\sigma(x,y) = \sigma^2(y) + p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d)$$

On the basis of the considered decompositions we have:

$$\begin{aligned}&\rho(x,y) = \\ &= \frac{\sigma^2(y) + p\sigma_{U_1}(y)\sigma_{U_1}(d)\rho_{U_1}(y,d) + p(1-p)\left(\bar{y}_{U_1} - \bar{y}_{U_0}\right)\bar{d}_{U_1}}{\sigma(y)\sqrt{\sigma^2(y) + 2p\sigma_{U_1}(y,d) + p\sigma^2_{U_1}(d) + p(1-p)\left(\bar{d}_{U_1} + 2\left(\bar{y}_{U_1} - \bar{y}_{U_0}\right)\right)\bar{d}_{U_1}}}\end{aligned}$$

It is difficult to infer using the obtained coefficient without additional assumptions. When we assume that $\rho_{U_1}(y,d) = 0$ and $\bar{d}_{U_1} = 0$, then

$$\rho(x,y) = \frac{\sigma(y)}{\sqrt{\sigma^2(y) + p\sigma^2_{U_1}(d)}}.$$

Hence, $0 < \rho(x,y) \leq 1$. When observations of variable y are not contaminated by errors, then $\rho(x,y) = 1$. A large variance $\sigma^2_{U_1}(d)$ and a close to one fraction p results in coefficient ρ^2 close to zero. Moreover, let us note that the parameter

$$\zeta = \rho^2(x,y) = \frac{\sigma^2(y)}{\sigma^2(y) + p\sigma^2_{U_1}(d)} \tag{1.7}$$

is called the reliability coefficient, see Guilford (1971). In the case of a small spread of errors or a small fraction p the coefficient ζ is close to one.

Let us go back to expressions (1.1) - (1.4). They define the vectors **x**, **y** and **z** for the population U. Moreover, **x=y+d**, where **d** is the vector of errors. From a practical point of view we are able to observe only the vector **x** in the whole population, while in general we are interested in the parameters y_U, d_U and p. Inference on these parameters can be based on the joint distribution of the variables (**x**, **y**). The properties of this distribution should be recognized based on an appropriately selected sample from the population. The data observed in the sample lets us infer the parameters y_U, d_U and p.

In general we can say that statistical analysis in auditing is focused on inference on the frequency of the errors denoted by p or the total error amount d_U. Statistical inference on these parameters is developed using one of two approaches. The randomization approach is considered if we assume that the vectors **x**, **y** and **z** are fixed (i.e. non-random). The model approach is taken into account when **x**, **y** or **z** are treated as values of some multidimensional random variable.

1.2 Model approach

In the previous section, the parameter of the population was denoted by the matrix $[\mathbf{y}\ \mathbf{x}]$ and consisted of non-random column vectors: **y** and **x**. In our case $\mathbf{y} \in R^N_+$ and $\mathbf{x} \in R^N_+$. Let us assume that those vectors are appropriate values of the following two random vectors: $\mathbf{Y}^T = [Y_1 ... Y_k ... Y_N]$ and $\mathbf{X}^T = [X_1 ... X_k ... X_N]$. Therefore, the population book amount x_k is the outcome of the random variable X_k and the audited (or true) population amount y_h is the value of Y_k, $k = 1,...,N$. Let $F(\mathbf{y},\mathbf{x})$ be the distribution function of the random matrix $[\mathbf{Y}\ \mathbf{X}]$. Hence, the matrix $[\mathbf{y}\ \mathbf{x}]$ is the outcome of $[\mathbf{Y}\ \mathbf{X}]$.

The conditions for defining a class of distribution functions to which $F(\mathbf{y},\mathbf{x})$ is assumed to belong is called the super-population or population model (see Cassel et al. (1977) or Fuller (2009) or Särndal et al. (1992)). Usually, the following simplified version of the model is defined:

$$F(\mathbf{y},\mathbf{x}) = \prod_{k=1}^{N} F(y_k, x_k) \tag{1.8}$$

where $F(y_k, x_k)$ is the distribution function of the variables $[Y_k\ X_k]$ attached to the k-th population element, $k \in U$.

The basic moments of the random variables are:

$$E(Y_k) = \mu(Y_k) = \begin{cases} \sum_i y_{k,i} P(Y_k = y_{k,i}), \\ \int_{R_+} y_k f(y_k) dy_k, \end{cases}$$

$$\sigma(Y_k, X_k) = \begin{cases} \sum_i (y_{k,i} - \mu(Y_k))(x_{k,i} - \mu(X_k)) P(X_k = x_{k,i}, Y_k = y_{k,i}), \\ \int_{R_+} \int_{R_+} (y_{k,i} - \mu(Y_k))(x_{k,i} - \mu(X_k)) f(y_k, x_k) dy_k dx_k \end{cases}$$

where $f(y_k, x_k)$, $f(y_k)$ are density functions. Moreover: $\sigma(Y_k, Y_k) = \sigma^2(Y_k)$, $k = 1, ..., N$.

The particular case for the model defined by expression (1.8) is as follows:

$$\begin{cases} Y_k = a x_k + b + \varepsilon_k, \\ E(\varepsilon_k) = 0, \quad V(\varepsilon_k) = \sigma^2 c_k \\ \text{for} \quad k = 1, ..., N, \end{cases} \tag{1.9}$$

where $E(Y_k | X_k = x_k) = a x_k + b$ and $c_k > 0$, $k = 1, ..., N$. This model is called the regression model. Usually it is assumed that the parameters c_k and $E(Y_k | X_k = x_k)$ are positively correlated.

Now let us take into account the index of error denoted previously by z_k, $k = 1, ..., N$ and assume that it is the value of the binary random variable Z_k with the following probability distribution:

$$P(Z_k = z_k) = \begin{cases} 1 - p_k, & \text{for} \quad z_k = 0 \\ p_k, & \text{for} \quad z_k = 1 \quad k = 1, ..., N. \end{cases} \tag{1.10}$$

Now let us assume that the values of the independent random variables X_k, $k = 1, ..., N$ are generated as follows:

$$X_k = (1 - Z_k) Y_k + Z_k W_k = Y_k + Z_k (W_k - Y_k) \tag{1.11}$$

where the value w_k of the random variable $W_k = Y_k + D_k$ is the accounting amount contaminated with accounting error D_k. Hence, some of the observations of X are contaminated with errors and all values of W are contaminated with errors. The above equation can be rewritten as follows:

$$X_k = Y_k + Z_k D_k. \tag{1.12}$$

The above model is usually considered under the assumption that Y, D and Z are independent, see also the last chapter.

Frequently, in practical research it is assumed that the random matrix $[\mathbf{X}\ \mathbf{Y}\ \mathbf{Z}]$ defined above is a simple random sample drawn from the distribution $F(x,y,z)$, where $P(Z=1)=p$ and $P(Z=0)=1-p$. Hence, the elements of the rows of the matrix $[\mathbf{X}\ \mathbf{Y}\ \mathbf{Z}]$ are independent and each of them is distributed according to $F(x,y,z)$ (see the classical definition of the simple random sample, Wilks (1962), p. 195).

Under the assumption that Y, D and Z are independent, the following equation is proved in Appendix 6.1.1:

$$F_x(x) = (1-p)F_y(x) + pF_w(x), \tag{1.13}$$

where in the discrete case of random variables Y and D:

$$F_w(x) = \sum_{\{d\}} F_y(x-d)P(D=d).$$

When Y and D are continuous:

$$F_w(x) = \int_{-\infty}^{\infty} F_y(x-u)f_d(u)du,$$

where $F_w(.)$ is the distribution function of the random variable W and $f_d(.)$ is the density function of the random variable D.

Particularly, when we assume that $p = P(Z=1) = 1$ then $X = W = Y + D$. In this case some parameters considered in the previous subsection can be simplified as follows:

$$\sigma^2(x) = \sigma^2(y) + \sigma^2(d) + 2\sigma(d)\sigma(y)\rho(y,d), \tag{1.14}$$

$$\rho(x,y) = \frac{\sigma(y) + \sigma(d)\rho(y,d)}{\sqrt{\sigma^2(y) + \sigma^2(d) + 2\sigma(d)\sigma(y)\rho(y,d)}}, \tag{1.15}$$

If $\rho(d,y) = 0$, then

$$\rho(x,y) = \frac{\sigma(y)}{\sqrt{\sigma^2(y) + \sigma^2(d)}}$$

and Gilford's reliability coefficient takes the following form:

$$\zeta = \rho^2(x,y) = \frac{\sigma^2(y)}{\sigma^2(y) + \sigma^2(d)}.$$

Instead of the distribution $F(\mathbf{y},\mathbf{x})$ or $G(\mathbf{d},\mathbf{x})$, more general distribution denoted by $F(\mathbf{y},\mathbf{x}|\Theta)$ or $G(\mathbf{d},\mathbf{x}|\Theta)$, where $\Theta \in R^p$ is a vector of parameters, can be considered. We are interested in inference on Θ or its real function. For instance, the hypothesis on Θ can be tested by means of the likelihood ratio test, (see Chapter 5).

Moreover, the parameter Θ can be treated as a random variable and its probability distribution is called a priori. In this case, the well-known Bayesian inference rule can be considered (see the section 2.4 and Chapter 5).

Statistical models for auditing are considered e.g. by Kaplan (1973), Cox and Snell (1979), Ijiri and Kaplan (1971), McCray (1984), Sorensen (1969), Stringer (1963) or in *Statistical models ...* (1989).

1.3 Outline of statistical inference in auditing

Statistical inference in auditing is usually reduced to the problem of evaluating the appropriate confidence intervals which assess the total accounting error d_U with assumed confidence at the level $\gamma \in [0;1]$. When the hypothetical value d_0 of the parameter d_U is inside the confidence interval, then the audited accounting system is considered to be working properly. In the case when d_0 is outside of the interval the system is wrong. As is well known, this decision rule is equivalent to the appropriate rule considered in testing statistical hypotheses (see e.g. Beck and Solomon (1985) or Wilks (1962)). More aspects connected with choosing the appropriate statistical inference procedure in auditing are considered e.g. by Lobbecke (1995), Lobbecke and Nether (1975).

In this book, all auditing problems are only considered in the context of testing appropriately formulated hypotheses usually using the value of the parameters d_U or p. It seems that such an approach is not only very convenient, but also natural, because the risks considered in auditing are compatible with errors of the first as well as of the second kind in testing statistical hypotheses. Moreover, some well-known statistical inference procedures, such as the likelihood ratio test, can be easily adapted for solving auditing problems.

Usually, as considered above, the sample s consists of two mutually disjoint subsamples s_0 and s_1 so that $s = s_0 \cup s_1$. The sizes of the samples s, s_0 and s_1 are n, n_0 and n_0, respectively. The set s is an outcome of the random sample (set) S. Hence, the capital letter S denotes only the sub-set of some population elements. The small letter s denotes the concrete set of elements drawn from the population. Similarly, s_0 and s_1 denote outcomes of the random samples S_0 and S_1, respectively. For instance $n_1 = m_s$, where m_s is the number of observations contaminated with errors and m_s is the value of the estimator m_S of the earlier introduced parameter m_U. Similarly, $p_s = \frac{m_s}{n}$ is an outcome of random variable $p_S = \frac{m_S}{n}$. Inference on parameter $d_U = x_U - y_U$ can be conducted in two directions. Firstly, the parameter d_U can be estimated on the basis of the statistic $d_S = x_U - y_S$, where y_S is at least an asymptotically unbiased estimator of y_U. The next estimator $d_{S_1} = \sum_{k \in S_1} d_k$ of the parameter d_U is based only on observations of errors $d_k = x_k - y_k$ in the sample S_1. In this case, the size of the sample S_1 and its distribution depends on the sampling scheme. The current consideration deals with the design-based approach. In the case of the model approach e.g. values $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ are treated as values of the following sequences of random variables $(X_1, ..., X_n)$ and $(Y_1, Y_2, ..., Y_n)$, respectively. That is why e.g. the above considered statistic d_{S_1} will be denoted by $D_{S_1} = \sum_{k \in S_1} (X_k - Y_k)$. Hence, $D_{s_1} = \sum_{k \in s_1} (X_k - Y_k)$ means the statistic determined on the basis of fixed sample s_1. Moreover d_{s_1} is an outcome of D_{s_1}.

Statistics D_S and D_{S_1} d_S and d_{S_1} can be treated as test statistics of a hypothesis on the parameter d_U. Inference based on the statistic d_{S_1} can be based on well-known small area estimation methods (see e.g. Domański and Pruska (2001), Ghosh and Rao (1994), Pfefferman (2002, 2013), Rao (2003), Särndal et al. (1992), Żądło (2008, 2015)). The considerations developed in the last chapter of this book are in some sense connected with ideas of statistical inference based on small area estimation methods. In general, we prefer classical statistical inference rules because it seems that they can be applied more simply in statistical auditing.

Chapter 2
Compliance tests

2.1 Testing hypotheses on fraction of accounting errors

According to the notation introduced in section 1.1, we can formulate several hypotheses. In general, the auditor's purpose is to test the following hypothesis:

$$H_0 : m_U \leq m_0, \qquad H_1 : m_U \geq m_1 > m_0, \tag{2.1}$$

where m_U is the number of non-proper accounting documents, and m_0 and m_1 are admissible and inadmissible numbers of incorrect accounting documents, respectively.

The formulated hypotheses are equivalent to the following:

$$H_0 : p_U \leq p_0, \qquad H_1 : p_U \geq p_1 > p_0, \tag{2.2}$$

where $p_U = m_U/N$ is the fraction of non-proper e.g. accounting documents, $p_0 = m_0/N$ and $p_1 = m_1/N$ are admissible and un-admissible fractions (shares) of wrong accounting documents, respectively.

The hypotheses are tested based on the data observed in a sample and by means of an appropriate test statistic. Let α be the significance level of a test equal to the risk η of the incorrect rejection of a properly working system, so $\alpha = \eta$. An error of the second kind occurs with probability ν and it is equal to the risk of incorrect acceptance of a system not working properly, so $\nu = \kappa$. Let us note that the power of a test is denoted by $\beta = 1 - \nu$.

The theory of statistics provides several tests to verify the formulated hypotheses. Testing can be based on a simple random sample drawn with (or without) replacement or on a non-simple sample. The sample will be denoted by S and its outcomes by s. The details about sampling designs as well as sampling schemes can be found in sections 3.2 and 4.1. Here we only outline the problem. Under the formulated hypotheses, the auditor has to determine the admissibility parameters m_0 and m_1 or p_0 and p_1. Moreover, the risk of incorrect rejection η and the risk of incorrect acceptance κ have to be assigned. Let us consider the simple test statistic m_S, equal to the

number of identified elements with errors in a sample S of size n. The significantly large values of the statistic m_S suggest that hypothesis H_0 should be rejected. Hence, the rejection (or critical) region of the test is upper-tailed and it is given by:

$$c_\alpha = [m_\alpha;\infty), \qquad P(m_S \in c_\alpha | H_0) = \alpha$$

When a value of the test statistic falls into the rejection region of the test, the hypothesis H_0 is rejected with the risk $\eta = \alpha$. When a value of the test statistic does not fall into the rejection region, then hypothesis H_0 is accepted with the risk $\kappa = \nu = P(m_S \in c_\alpha | H_1)$.

When the sample is a simple random sample drawn without replacement, then the test statistic has the well-known hypergeometric probability distribution. In section 2.2 and under the above hypotheses and the assigned risks of both types, the sample size is evaluated. Moreover, in section 2.3 we can find several methods of approximating the test statistic distribution when the value of the parameter p_U is close to zero.

Usually, in auditing, the size of a sample is limited by costs and it is therefore impossible to increase it. In this situation, the following hypotheses are considered, which are similar to the above:

$$H_0 : m_U \leq m_0, \qquad H_1 : m_U > m_0, \tag{2.3}$$

or

$$H_0 : p_U \leq p_0, \qquad H_1 : p_U > p_0.$$

In this case, the significance level is equal to the risk η of incorrect rejection of a properly working system. The risk κ of incorrect acceptance of a system not working properly is not controlled. Hence, when the value of the test statistic falls into the critical region, hypothesis H_0 is rejected with risk $\eta = \alpha$. When the value of the test statistic does not fall into the rejection region, hypothesis H_0 is not rejected. However, in this case, H_0 cannot be accepted because the level of risk κ cannot be assessed.

In some sense, the above hypotheses given by (2.1) and (2.2) have the following dual forms:

$$H_0' : m_U \geq m_1, \qquad H_1' : m_U \leq m_0 < m_1$$

or

$$H_0' : p_U \geq p_1, \qquad H_1' : p_U \leq p_0 < p_1. \tag{2.4}$$

Now, the significance level of the test is equal to the risk of incorrect acceptance of a system not working properly, denoted by, so κ, so $\alpha = \kappa$. An error of the second kind occurs with probability ν and is equal to the risk η of incorrect rejection of a properly working system, so $\nu = \eta$. Small values of the test statistic m_S lead to a rejection of hypothesis H_0'. In the considered case, the power of the test is $\beta = 1 - \eta$.

In practice, auditors pay special attention to the risk of incorrect acceptance of a system not working properly. That is why the following simplified hypotheses are considered:

$$H_0' : m_U \geq m_1, \qquad H_1' : m_U < m_1$$

or

$$H_0' : p_U \geq p_1, \qquad H_1' : p_U < p_1. \tag{2.5}$$

In this case, the significance level α is equal to the risk κ of incorrect acceptance of a system not working properly. The risk η of incorrect rejection of a properly working system is not controlled. Hence, when the value of the test statistic falls into the lower-tailed rejection region, hypothesis H_0 is rejected with risk $\kappa = \alpha$. When the value of the test statistic does not fall into the rejection region, hypothesis H_0 is not rejected, but we cannot accept it because the power of the test is not assessed. In practice, in the above formulated hypotheses, the parameters m_1 and p_1 can be replaced with m_0 and p_0, respectively.

2.2 Verification using an exact test

The construction of compliance tests is based on the well-known hypergeometric probability distribution. Some aspects of this approach are considered e.g. by Talens (2005). Let m_s be the number of errors identified in a simple random sample of size n, drawn without replacement. According to the notation introduced in section 1.1 we can write:

$$m_S = \sum_{k \in S} z_k.$$

It is well known that the statistic m_S has the following hypergeometric probability distribution:

$$P(m_S = m) = \frac{\binom{m_0}{m}\binom{N-m_0}{n-m}}{\binom{N}{n}}.$$

Hence, hypothesis H_0 defined by expressions (2.1) or (2.3) is rejected when the value of the test statistic m_s is significantly high. The critical value of the test denoted by m_η is determined by the expression:

$$\eta \geq \alpha = P(m_S \geq m_\eta | H_0)$$

where

$$m_\eta = min_{m=0,1,2\ldots}\{m\} \quad \text{and} \quad P(m_S \geq m | H_0) \leq \eta.$$

η is the assumed risk of incorrect rejection of a properly working system and α is the significance level of the test. If $m_s \geq m_\eta$, then the system is not accepted as working properly, with the risk of improper decision equal to η. When $m_s < m_\eta$, then the system is accepted as working properly with the risk of improper decision not greater than the risk of incorrect acceptance of the system not working properly denoted by κ, where

$$\kappa \geq \nu = P(m_S \leq m_\eta | H_1)$$

where ν is the probability of type II error and $\nu = 1 - \beta$ where β is the power of the test.

Of course, the above inequality will be fulfilled when the sample size is sufficiently large. This is evaluated as the solution of the following system of inequalities:

$$\begin{cases} \eta \geq P(m_S \geq m|H_0) = 1 - \sum_{k=0}^{m-1} \frac{\binom{m_0}{k}\binom{N-m_0}{n-k}}{\binom{N}{n}}, \\ \\ \kappa \geq P(m_S < m|H_1) = \sum_{k=0}^{m-1} \frac{\binom{m_1}{k}\binom{N-m_1}{n-k}}{\binom{N}{n}}, \end{cases} \tag{2.6}$$

Let $m_{\eta,\kappa}$ and $n_{\eta\kappa}$ be the solution of the system. Hence, under the assumed risks η and κ, the necessary sample size should be equal to $n_{\eta\kappa}$ and the critical value of the test equal to $m_{\eta,\kappa}$.

The above system of inequalities can be solved using a computer program written in the R programming language. The program is presented in Appendix 6.2.1.

2.3 Approximation of the test statistic distribution

In practice the distribution of the statistic m_S is approximated in order to simplify the evaluation of the sample size. Some usually used methods of approximation are presented below.

2.3.1 Binomial approximation

In practice, hypergeometric probability distribution can be approximated by means of the well-known binomial probability distribution (see Colopper and Pearson (1934), Gerstenkorn and Śródka (1974), Johnson et al. (1992), Krzyśko (2000)). Approximation is considered when the probability $p = \frac{m_u}{N}$ is close to zero or close to one. Usually, it is taken into account when $N \geq 1000$ and $p \leq 0.04$ or $p \geq 0.96$. In this situation the hypergeometric probability distribution function is replaced with the following:

$$P(m_S = m) = \binom{n}{m} p^m (1-p)^{n-m}. \tag{2.7}$$

System (2.6) is exchanged for the following:

$$\begin{cases} \eta \geq P(m_S \geq m|H_0) = 1 - \sum_{k=0}^{m-1} \binom{n}{k} p_0^k (1-p_0)^{n-k}, \\ \\ \kappa \geq P(m_S < m|H_1) = \sum_{k=0}^{m-1} \binom{n}{k} p_1^k (1-p_1)^{n-k}. \end{cases} \tag{2.8}$$

The solution of the above system denoted by $(m_{\eta,\kappa}, n_{\eta\kappa})$ under the assumed risks η and κ determines the necessary sample size $n_{\eta\kappa}$ and the critical value of the test $m_{\eta,\kappa}$. The above system of inequalities can be solved using a computer program written in the R programming language. This program is presented in Appendix 6.2.2.

Jowett (1963) proposed to evaluate binomial distribution by means of the well-known F-distribution (see Seber (2013), too). This lets us rewrite the above inequality system as follows:

$$\begin{cases} \eta \geq P(m_S \geq m|H_0) = 1 - P\left(W_0 \leq \frac{(1-p_0)(m+1)}{p_0(n-m)}\right), \\ \\ \kappa \geq P(m_S < m|H_1) = P\left(W_1 \leq \frac{(1-p_1)m}{p_1(n-m+1)}\right), \end{cases}$$

where the random variable W_0 has central F-distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom while W_1 has central F-distribution with $2(n-m+1)$ and $2m$ degrees of freedom.

Finally, let us note that Agresti and Coull (1998), Brown, Cai and DasGupta (2002), Chen (1990) and Wilcox (2012) also considered problems of approximation in binomial distribution.

2.3.2 Poisson approximation

Hypergeometric probability distribution can be approximated by means of the well-known Poisson probability distribution when probability p decreases and sample size n increases. Like in the previous section, approximation is usually considered when $N \geq 1000$ and $p \leq 0.04$ or $p \geq 0.96$. In this situation, system (2.6) based on the hypergeometric probability distribution function is replaced with the following:

$$\begin{cases} \eta \geq P(m_S \geq m|H_0) = 1 - \sum_{k=0}^{m-1} \frac{(np_0)^k}{k!} e^{-np_0}, \\ \\ \kappa \geq P(m_S < m|H_1) = \sum_{k=0}^{m-1} \frac{(np_1)^k}{k!} e^{-np_1}. \end{cases} \tag{2.9}$$

The solution of the above system denoted by $(m_{\eta,\kappa}, n_{\eta\kappa})$ under the assumed risks η and κ can be solved using a computer program written in the R programming language. This program is presented in Appendix 6.2.3.

The Poisson distribution function can be approximated by means of the well-known chi-square distribution (see e.g. Johnson et al. (1992) and Seber (2013), p. 12). This lets us rewrite the above inequality system as follows:

$$\begin{cases} \eta \geq P(m_S \geq m|H_0) = 1 - P\left(\chi^2_{2(m+1)} \leq 2np_0\right), \\ \\ \kappa \geq P(m_S < m|H_1) = P\left(\chi^2_{2m} \leq 2np_1\right), \end{cases}$$

where the random variable χ^2_r has the central chi-square distribution with r degrees of freedom.

2.3.3 Computer simulation approximation

Binomial or Poisson approximation does not seem to be sufficiently accurate in the case of very small or very close to one probabilities, e.g. $p < 0.001$ or $p > 0.999$. In this case, the computer simulation procedure can be used to evaluate the necessary sample size and critical value under the assumed risks η and κ. The tested hypothesis, defined in section 2.1 by means of expression (2.2), is $H_0 : p \leq p_0$ and the alternative one is $H_1 : p \geq p_1$.

The algorithm is two-stage. Firstly, the critical value is evaluated. In order to do this we assume some starting level of the sample size. Let us denote it by n_0. The sample s_1 of size n_0 consists of values 0 or 1. The value 1 is generated with probability p_0, so the value 0 is generated with probability $1 - p_0$. Next, the frequency of the 1 value is calculated. Let this be denoted by p_{s_1}. The samples s_k, $k = 1, ..., r$ are drawn independently a large number of times. Usually, the number of repetitions of the samples should not be lower than $r = 100000$. In this way we independently calculate r sample frequencies p_{s_k}, $k = 1, ..., r$. Based on the ordered sequence of the frequencies, the quantile of the order $1 - \eta$ is evaluated. This quantile is treated as the critical value of the test under the assumed sample size at the starting level n_0. Let this be denoted by p_{η,n_0}.

In the second stage of the algorithm, the power of the test is evaluated under the assumed sample size n_0 and the determined critical value p_{η,n_0}. Next, series of independent samples each of size n_0 are generated but now the value 1 is generated with probability p_1. The number of repetitions of the sample is equal to r. On the basis of each generated sample s_k of size n_0, the frequency p_{s_k}, $k = 1, ..., r$ is calculated. Let b_{η,n_0} be the frequency of values p_{s_k} exceeding the critical value p_{η,n_0}. The frequency b_{η,n_0} assesses the power of the test $\beta = 1 - \kappa$.

If $b_{\eta,n_0} \geq \beta$, then the sample size n_0 is treated as sufficiently large. It means that under the size n_0 and simulated critical value p_{η,n_0} the test reaches the assumed risks η and κ. If $b_{\eta,n_0} < \beta$, then the simulation procedure starts again from assuming a level of the sample size larger than n_0. Let the new sample size be equal to $n_1 = n_0 + c$, where c is an integer, for instance $c = 10$. In general, if $b_{\eta,n_t} \geq \beta$, then the sample size n_t is treated as sufficiently large. This means that under the size n_t and simulated critical value p_{η,n_t} the test reaches the assumed risks η and κ. If $b_{\eta,n_t} < \beta$, then the simulation procedure is started again from assuming a larger sample size

$n_{t+1} = n_t + c$. Finally, let us note that Ryan (2013) and Chernick and Liu (2002) discuss some problems connected with the approach considered above.

The described algorithm can be solved using a computer program written in the R programming language. This program is in presented Appendix 6.2.4.

2.3.4 Normal approximation

We are still considering testing hypothesis H_0 against H_1 defined in expression (2.2). As was earlier considered earlier, the test statistic is the number of errors identified in the sample and it is denoted by m_S or the sample fraction of the number of errors denoted by $p_S = m_S/n$. In the case of a simple random sample drawn without replacement the following standardized form of the test statistic is considered:

$$z_S = \frac{p_S - p_0}{\sqrt{\frac{N-n}{N-1}\frac{p_0(1-p_0)}{n}}}.$$

On the basis of well-known of de Moivre-Laplace central limit theorem (see e.g Gerstenkorn et al. (1974)), we can approximate the probability distribution of the test statistic z_S by the standard normal distribution for a sample size not smaller than 100 but only for $0.04 \leq p \leq 0.96$.

Under the assumed significance level α equal to risk η and the probability of the appearance of type II error ν equal to risk κ, we have (see e.g. Wywiał (2014a)):

$$\eta = \alpha = P(p_S \geq p_\eta | H_0) = P(z_S \geq z_\eta | H_0) = P(Z \geq z_\eta) = 1 - \phi(z_\eta),$$

$$\kappa = \nu = P(p_S < p_\eta | H_1) = P(z_S < z_\kappa | H_1) = P(Z < z_\kappa) = \phi(z_\kappa),$$

where

$$p_\eta = p_0 + z_\eta \sqrt{\frac{N-n}{N-1}\frac{p_0(1-p_0)}{n}},$$

$$z_\kappa = \frac{p_0 - p_1}{\sqrt{\frac{N-n}{N-1}\frac{p_1(1-p_1)}{n}}} + z_\eta \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}},$$

where $\phi(z) = P(Z < z)$ is the distribution function of the standard normal random variable: $Z \sim N(0;1)$.

The above results let us derive the necessary sample size:

$$n_* = N \left(1 + \frac{(N-1)(p_1 - p_0)^2}{\left(z_\eta \sqrt{p_0(1-p_0)} - z_\kappa \sqrt{p_1(1-p_1)} \right)^2} \right)^{-1}, \tag{2.10}$$

where $\phi(z_\kappa) = \kappa = \nu$, $\phi(z_\eta) = 1 - \eta = 1 - \alpha$.

In the case of a large population size or a simple random sample drawn with replacement, the above expression simplifies to the following form:

$$n_{**} = \frac{\left(z_\eta \sqrt{p_0(1-p_0)} - z_\kappa \sqrt{p_1(1-p_1)}\right)^2}{(p_1-p_0)^2}. \tag{2.11}$$

Finally, let us consider the problem of testing the following hypotheses (see expressions (2.4)):

$$H_0' : \ p_U \geq p_1, \qquad H_1' : \ p_U \leq p_0 < p_1. \tag{2.12}$$

The above hypotheses are in some sense dual to the hypotheses defined by expression (2.2). In this case we have the following test statistic:

$$z_S = \frac{p_S - p_1}{\sqrt{\frac{N-n}{(N-1)n} p_1(1-p_1)}}, \tag{2.13}$$

$$\kappa = \alpha = P(p_S \leq p_\kappa | H_0) = P(z_S \leq z_\kappa | H_0) = \phi(z_\kappa),$$
$$\eta = \nu = P(p_S > p_\kappa | H_1),$$

where

$$p_\kappa = p_1 + z_\kappa \sqrt{\frac{N-n}{N-1} \frac{p_1(1-p_1)}{n}}, \tag{2.14}$$

and $\phi(z_\kappa) = P(Z < z_\kappa) = \kappa = \alpha$, $Z \sim N(0;1)$. We can show that the necessary sample size evaluated under the assumed risks $\kappa = \alpha$ and $\eta = \nu$ is determined by expressions (2.10).

In the statistical literature another normal approximation of the probability distribution of the test statistic p_S is considered. The probability distribution of the transformation:

$$\hat{p_S} = 2arcsin(\sqrt{p_S})$$

is well approximated by the normal probability distribution with variance equal to $1/n$ (see Ryan (2013), p. 104). This lets us approximately evaluate the necessary sample size on the basis of the expression:

$$n_{***} = \frac{1}{4} \left(\frac{z_\nu + z_\alpha}{arcsin(\sqrt{p_0}) - arcsin(\sqrt{p_1})} \right)^2$$

This expression is valid when the hypotheses defined by expression (2.2) are tested.

2.4 Bayesian approach

We are interested in testing the following hypothesis (see Wywiał (2014)):

$$H_0 : \ p \leq p_1, \qquad H_1 : \ p > p_1.$$

Now the parameter $p \in (0;1)$ is treated as a random variable. We take into account the estimator $p_S = m_S/n$ from the simple random sample drawn with replacement, where m_S has Bernoulli distribution defined by expression (2.7). Usually, it is assumed that the parameter p has the well-known beta distribution $B(a,b)$. This is called the prior distribution. Its density function is given by:

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad a > 0, \quad b > 0 \tag{2.15}$$

where $\Gamma(\nu) = \int_0^\infty t^{\nu-1}e^{-t}dt$, $\nu > 0$. The expected value, the variance and the skewness coefficient are:

$$E(p) = \frac{a}{a+b}, \quad V(p) = \frac{ab}{(a+b)^2(a+b+1)}, \quad \beta_1(p) = 2\frac{(a+b)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}.$$

The posterior distribution of p is given by (see e.g. Domański et al. (2014), Ghosh and Meeden (1997), Krzyśko (2004) or Santer and Duffy (1989)):

$$h(p|m_s) = \frac{\Gamma(a+b+n)}{\Gamma(a+m_s)\Gamma(b+n-m_s)} p^{a+m_s-1}(1-p)^{b+n-m_s-1}$$

Let us suppose that c_0 is the loss that deals with the situation when hypothesis H_0 is accepted when H_1 is true. This means that the auditor accepts a system when it does not work properly. The parameter c_1 is the loss generated by rejecting H_0 when it is true. This means that a properly working system is not accepted.

The general Bayesian rule of testing statistical hypotheses leads to the evaluation of the following posterior probabilities:

$$P(p < p_1|m_s) = \int_0^{p_1} h(p|m_s)dp, \qquad P(p \geq p_1|m_s) = \int_{p_1}^1 h(p|m_s)dp$$

The risk of accepting hypothesis H_1 when H_0 is true is equal to $c_1P(p < p_1|m_s)$. The risk of accepting H_0 when H_1 is true is equal to $c_0P(p \geq p_1|m_s)$. The decision rule is as follows. Hypothesis H_0 is rejected when:

$$P(p < p_1|m_s) \leq \frac{c_0}{c_0+c_1} = \omega. \tag{2.16}$$

Hypothesis H_0 is accepted if:

$$P(p < p_1|m_s) > \omega. \tag{2.17}$$

Usually, $c_0 = c_1$. Hence, $\omega = 0.5$.

The next decision rule involves the following Bayes factor (see, e.g. Jeffreys (1961) or Robert (2007), p. 227):

$$B = \frac{P(p < p_1|m_s)P(p \geq p_1)}{P(p \geq p_1|m_s)P(p < p_1)} \qquad l_B = ln(B). \tag{2.18}$$

Its value is explained as follows (Lodewyckx et al. (2011) after Raftery (1995)):
$0 < l_B \leq 1$, is weak support for H_0,
$1 < l_B \leq 3$, is positive support for H_0,
$3 < l_B \leq 5$, is strong support for H_0,
$l_B > 5$, is very strong support for H_0.
The more popular decision rule is as follows (see Kass, and Raftery (1995)):
If $0 < l_B \leq 2$, evidence against H_0 is not worth more than a bare mention.
If $2 < l_B \leq 6$, evidence against H_0 is positive.
If $6 < l_B \leq 10$, evidence against H_0 is strong.
If $l_B > 10$, evidence against H_0 is very strong.

Santer et al. (1989) write that usually it is assumed that $a+b=n$ and $E(m_S) = p_0$ where p_0 is the mean value of the admissible (or expected) level of probability that the audited system does not work properly. This lets us assess that $a = np_0$ and $b = n - a = n(1-p_0)$. Finally, we have:

$$V(p) = \frac{p_0(1-p_0)}{n+1} \approx V(p_S).$$

The parameters a and b can be estimated by means of the empirical Bayes procedure (see e.g. Copas (1972), Griffin and Krutchkoff (1971) or Walter and Hamdani (1987)).

Example 2.1. (Wywiał (2014)): An auditor controls 40 accounting documents. He has found that two of them contain errors. It is assumed that the internal control system works properly when $p_0 = 0.03$. The auditor states that the internal control system does not work properly when $p \geq 0.08 = p_1$ where p_1 is the inadmissible probability of finding documents with errors. Moreover, the auditor assumes that $c_0 = c_1$.

Our hypothesis is as follows:

$$H_0 : p \leq p_1 = 0.08, \qquad H_1 : p > p_1 = 0.08.$$

On the basis of the previous results we have: $m = 2$, $n = 40$, $p_0 = 0.03$, $a = 1.2$, $b = 38.8$.

The posterior distribution is:

$$h(p|2) = \frac{\Gamma(80)}{\Gamma(3.2)\Gamma(76.8)} p^{2.2}(1-p)^{75.8} \sim B(3.2, 76.8).$$

In this case, $E(p|2) = 0.04$, $V(p|2) = 0.0005$, $\beta_1(p|2) = 1.1202$. Using the R-function $pbeta(0.08, 3.2, 76.8)$ we have:

$$P(p < p_1|2) = 0.9463 > \omega = \frac{1}{2}.$$

Hence, the auditor should accept the internal control system as working properly.

The prior distribution of p is:

$$h(p) = \frac{\Gamma(40)}{\Gamma(1.2)\Gamma(38.8)} p^{0.2}(1-p)^{37.8} \sim B(1.2, 38.8).$$

In this case, $E(p|2) = 0.03$, $V(p|2) = 0.0007$, $\beta_1(p|2) = 1.7874$. Let us note that the variance is larger in the case of the prior distribution than in the posterior distribution. The skewness coefficients satisfy a similar relationship. So, the prior distribution of p is more asymmetric than the posterior one.

Using the $pbeta(0.08, 1.2, 38.8)$ function we have: $P(p < p_1 = 0.08) = 0.9435$, $P(p \geq p_1|2) = 0.0537$, $P(p \geq p_1) = 0.0566$.

$$B = \frac{1.0031}{0.949} = 1.0569, \qquad l_B = 0.0241.$$

Concluding, the evidence against the quality of the internal control system (against H_0) is poor because $0 < l_B \leq 0.5$.

The hypothesis considered here can be tested using a random sample drawn from a stratified population. This case of inference is considered e.g. by Wendell and Schmee (1996), Meeden (2003) and Wywiał (2014).

Chapter 3
Substantive tests based on basic random samples

3.1 Testing hypotheses on total accounting amount errors

Some substantive audit procedures are connected with testing the following hypothesis:

$$H_0: \quad |d_U| \leq d_0, \qquad H_1: \quad |d_U| > d_0 \tag{3.1}$$

or

$$H_0: \quad |\bar{d}| \leq \bar{d_0}, \qquad H_1: \quad |\bar{d}| > \bar{d_0}$$

where d_0 is the admissible total accounting error and $\bar{d_0}$ is the admissible mean population accounting error. The hypotheses formulated by the above expressions deal with cases when an accounting error can be positive or negative.

In order to simplify this consideration, let us assume that all errors are not negative. That is why the following hypotheses are taken into account:

$$H_0: \quad d_U \leq d_0, \qquad H_1: \quad d_U \geq d_1 > d_0 \tag{3.2}$$

or

$$H_0: \quad \bar{d} \leq \bar{d_0}, \qquad H_1: \quad \bar{d} \geq \bar{d_1} > \bar{d_0}$$

where d_0, $\bar{d_0}$ is the admissible total (mean) accounting error and, d_1, $\bar{d_1}$ is the inadmissible total (mean) accounting error. In practice the parameter d_0 ($\bar{d_0}$) is stated on the level of mean value. When the observed accounting amount is not less than the appropriate true accounting amount, then the accounting error can be defined as $d = x - y$. If the observed accounting amount is not greater than the appropriate true accounting amount, the non-negative accounting error can be defined as $d = y - x$.

As an example, the hypotheses shown in expression (3.2) are tested on the basis of data observed in a sample and by means of an appropriate test statistic d_S. The significance level of the test will be denoted by α and it is equal to the risk of incorrect rejection (e.g. the accounting report) denoted by η, so, $\alpha = \eta$. The power of the test is assumed to be at level β. Hence, the probability that type II error will

occur (hypothesis H_0 is accepted when H_1 is true) is $\nu = 1-\beta$ and it is equal to the risk of incorrect acceptance, so $\nu = \kappa$.

In general, according to the above hypotheses the rejection region of the test is upper-tailed:

$$c_\alpha = [d_\alpha;\infty), \qquad P(d_S \in c_\alpha | H_0) = \alpha = \eta. \tag{3.3}$$

When $d_s \in c_\alpha$, hypothesis H_0, given by expression (3.2) is rejected. This decision is wrong with risk $\eta = \alpha$. If the value of d_S does not fall into the critical region, the hypothesis H_0 is accepted. This decision is wrong with risk $\kappa = \nu$.

When parameter d_1 is not defined, the above hypotheses can be simplified to the following form:

$$H_0: \quad d_U \le d_0, \qquad H_1: \quad d_U > d_0 \tag{3.4}$$

or

$$H_0: \quad \bar{d} \le \bar{d}_0, \qquad H_1: \quad \bar{d} \ge \bar{d}_0$$

In this case only the risk of incorrect rejection can be controlled. The critical region is shown by expression (3.3).

The dual versions of the hypotheses formulated by expression (3.2) are as follows:

$$H_0': \quad d_U \ge d_1, \qquad H_1': \quad d_U \le d_0 < d_1 \tag{3.5}$$

or

$$H_0': \quad \bar{d} \ge \bar{d}_1, \qquad H_1': \quad \bar{d} \le \bar{d}_0 < \bar{d}_1.$$

In this case, the significance level α of the test is equal to the risk of incorrect acceptance (e.g. of the accounting report) denoted by κ, so $\alpha = \kappa$. The probability ν, that type II error occur, is equal to the risk of incorrect rejection, so $\nu = \eta$. According to the hypotheses in expression (3.5), the lower-tailed rejection region of the test is:

$$c_\alpha = (-\infty;d_\alpha], \qquad P(d_S \le d_\alpha | H_0) = \alpha = \kappa. \tag{3.6}$$

When $d_s \in c_\alpha$, hypothesis H_0, given by expression (3.5) is rejected. This decision is wrong with risk κ. If the value of d_S does not fall into the rejection region, hypothesis H_0 is accepted. This decision is wrong with risk η. Finally, let us note that when d_0 is not known hypotheses (3.5) can by simplified to the following:

$$H_0': \quad d_U \ge d_1, \qquad H_1': \quad d_U < d_1 \tag{3.7}$$

or

$$H_0': \quad \bar{d} \ge \bar{d}_1, \qquad H_1': \quad \bar{d} < \bar{d}_1.$$

In this case, only the risk of incorrect acceptance $\kappa = \alpha$ can be controlled.

Sometimes the total (mean) absolute error amount $d_{A,U}$ ($\bar{d}_{A,U}$) defined in Subsection 1.1 is taken into account. In this case the hypotheses are formulated as follows:

$$H_0: \quad d_{A,U} \le d_0, \qquad H_1: \quad d_{A,U} \ge d_1 > d_0$$

or

$$H_0: \quad \bar{d}_{A,U} \leq \bar{d}_0, \qquad H_1: \quad \bar{d}_{A,U} \geq \bar{d}_1 > \bar{d}_0$$

where d_0, ($\bar{d}_0$) is the total (mean) admissible accounting error and, d_1, ($\bar{d}_1$) is the total (mean) inadmissible accounting error.

The dual hypotheses for the above hypotheses are:

$$H_0': \quad d_{A,U} \geq d_1, \qquad H_1': \quad d_{A,U} \leq d_0 < d_1$$

or

$$H_0': \quad \bar{d}_{A,U} \geq \bar{d}_1, \qquad H_1': \quad \bar{d}_{A,U} \leq \bar{d}_0 < \bar{d}_1$$

Now, let us consider the model approach defined in section 1.2. We assume that the independent random variables are denoted by (Y_k, X_k), $k = 1,...,N$ and $E(Y_k) = \mu_k(y), E(X_k) = \mu_k(x)$, $k = 1,...,N$. The expected values of the totals are $E(Y_U) = E\left(\sum_{k\in U} Y_k\right)$ and $E(X_U) = E\left(\sum_{k\in U} X_k\right)$. Let $\delta = E(X_U) - E(Y_U)$ and $\bar{\delta} = E(\bar{X}_U) - E(\bar{Y}_U)$. The following hypothesis will be considered:

$$H_0: \quad \delta \leq d_0, \qquad H_1: \quad \delta \geq d_1 > d_0 \tag{3.8}$$

or

$$H_0: \quad \bar{\delta} \leq \bar{d}_0, \qquad H_1: \quad \bar{\delta} \geq \bar{d}_1 > \bar{d}_0.$$

This hypothesis is similar to the hypothesis defined by (3.2). Let us note that it is also possible to formulate hypotheses similar to the hypotheses defined before expression (3.8). The same is true for the rejection regions.

3.2 Basic sampling designs and schemes

Let s be the set of sample elements. Hence, $s = \{k_i : k_i \in U, i = 1,...,n\}$. If the population elements are not replicated in s then $s \subseteq U$. This is the outcome of drawing a sample s without replacement from the population U of size $N \geq n$. If elements are drawn with replacement from the population, the elements in the sample can be replicated. Only samples of the fixed size n will be considered in this book.

The set of all possible samples of size n drawn without or with replacement from the population U will be denoted by $\mathbf{S}$ and called the sample space (support). For samples drawn without replacement, the set $\mathbf{S}$ consists of $\binom{N}{n}$ different samples.

The function $P(s)$ on $\mathbf{S}$, satisfying

$$P(s) \geq 0 \text{ for all } s \in \mathbf{S} \text{ and } \sum_{s\in\mathbf{S}} P(s) = 1.$$

is called the sampling design.

Equivalently, the above probability $P(s)$ can be replaced with $P(S = s)$, where S is a random set of indices drawn from the population and s is its outcome.

The probability of selecting the fixed unit k in a sample s is called the first-order inclusion probability and denoted by π_k. It is determined by the following expression:

$$\pi_k = \sum_{s:k\in s} P(s), \quad k = 1,\ldots,N.$$

Similarly, the second order inclusion probability is as follows:

$$\pi_{k,l} = \sum_{k,l\in s,k\neq l} P(s), \quad k \neq l, \quad k = 1,\ldots,N, \quad l = 1,\ldots,N.$$

In the case of samples of a fixed size:

$$\sum_{k\in U} \pi_k = n, \qquad \sum_{k,l\in s,k\neq l} \pi_{k,l} = n(n-1).$$

More properties of different sampling designs can be found e.g. in the monograph by Tillé (2006).

3.2.1 Simple random samples

Simple random sample drawn without replacement

The sampling design of a simple random sample drawn without replacement is as follows:

$$P_0(s) = \binom{N}{n}^{-1} \qquad \text{for all} \qquad s \in \mathbf{S}. \tag{3.9}$$

The inclusion probabilities for the sampling design P_0 are as follows:

$$\pi_k = \frac{n}{N} \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)}$$

. The set of probabilities implementing the sampling design P_0 is defined as described hereafter. The probability of firsts electing fixed population element k_1 in a sample is as follows:

$$p(k_1) = N^{-1} \text{ for } k_1 = 1,\ldots,N.$$

The conditional probability of selecting the fixed population element k_i to a sample, provided that the elements $k_{i-1},\ldots,k_1$ have been selected to the sample, is as follows:

$$p(k_i|k_{i-1},\ldots,k_1) = \tfrac{1}{N-i+1} \text{ for } i = 2,\ldots,n \text{ and } k_i = 1,\ldots,N.$$

Let us note that Rao T.V.H. (1962) proved that for any given design $P(s)$ there exists at least one sampling scheme that implements $P(s)$.

Simple random sample drawn with replacement

The sampling design of a simple random sample drawn with replacement of size n is as follows (see. Tillé (2006), p. 54):

$$P_1(\mathbf{s}) = \frac{n!}{N^n \prod_{k\in U} n_k!}, \quad \text{where} \quad s \in \underline{\mathbf{S}}, \quad 0 \le n_k \le n, \quad \sum_{k=1}^{N} n_k = n \tag{3.10}$$

and n_k is the number of the replication of the k-th population element in the sample s.

The inclusion probabilities are as follows:

$$\pi_k = 1 - \left(\frac{N-1}{N}\right)^n, \qquad \pi_{k,l} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$$

where $k = 1,...,N$, $l = 1,...,N$, $k \neq l$.

3.2.2 Simple systematic sample

Let $N = nq$ where n is the sample size and q is the integer. The sampling space is defined as the following sequence of possible samples: $\mathbf{S} = (s_e; e = 1,...,q)$ where $s_e = (i_{e+(j-1)q}; j = 1,...,n)$. The simple systematic sampling design is as follows:

$$P_2(s) = \frac{1}{q} \qquad for \qquad s \in \mathbf{S}.$$

The inclusion probabilities are given by:

$$\pi_i = \frac{1}{q} \qquad for \qquad i = 1,...,N,$$

$$\pi_{i,j} = \begin{cases} \frac{1}{q} & for \quad i \neq j, \quad i \in s, \quad j \in s, \\ 0 & for \quad i \neq j, \quad i \notin s, \quad or \quad j \notin s. \end{cases}$$

Sampling scheme: let u be the value of a random variable uniformly distributed on the interval $[0;1]$. The population element (which is the first element of the sample s_e, $e = 1,...,q$) denoted by i_e has to fulfil the following inequalities:

$$\frac{i_e - 1}{q} < u \le \frac{i_e}{q}.$$

3.2.3 Stratified sampling

Let the population U be divided into non-empty and mutually disjoint subsets U_h, $h = 1,..,H$, called strata, so $U = \bigcup_{h=1}^{H} U_h$. The size of the stratum U_h will be denoted by $N_h \geq 1$, $h = 1,...,H$ and $N = \sum_{h=1}^{H} N_h$. Let $w_h = N_h/N$. The stratified sample will be denoted by $s = (s_1,...,s_H)$ where $s_h \subseteq U_h$ for $h = 1,...,H$. Let $0 < n_h \leq N_h$ be the size of the sample s_h, $h = 1,...,H$. When the sampling designs of the samples s_h, $h = 1,...H$ are denoted by $P_{3,h}(s_h)$, respectively, then the sampling design of the stratified sample is as follows:

$$P_3(s) = \prod_{h=1}^{H} P_{3,h}(s_h) \tag{3.11}$$

where

$$P_{3,h}(s_h) = \binom{N_h}{n_h}^{-1}, \qquad h = 1,...,H.$$

$P_3(s)$ is called the simple stratified random sample. Its inclusion probabilities are as follows:

$$\begin{cases} \pi_k = \frac{n_h}{N_h}, & \text{for} \quad k \in U_h \\ \pi_{k,l} = \frac{n_h}{N_h}\frac{n_t}{N_t}, & \text{for} \quad k \in U_h, \quad l \in U_t, \\ \pi_{k,l} = \frac{n_h(n_h-1)}{N_h(N_h-1)}, & \text{for} \quad k,l \in U_h \end{cases}$$

where $k = 1,..,N$, $l = 1,...,N$, $l \neq k$, $h = 1,...,H$, $t = 1,...,H$, $h \neq t$.

For simple random samples drawn with replacement we have:

$$P_4(s) = \prod_{h=1}^{H} P_{4,h}(s_h).$$

$$P_{4,h}(s_h) = \frac{n_h!}{N_h^{n_h} \prod_{k \in U_h} n_{h,k}!}, \quad \text{where} \qquad 0 \leq n_{h,k} \leq n_h, \quad \sum_{k=1}^{N_h} n_{h,k} = n_h.$$

3.3 Simple random sample mean

3.3.1 Basic definitions

Let T_s be the real function of data on y or x observed in the sample s drawn according to the sampling design $P(s)$. The sampling strategy is the pair $(T_S, P(s))$ (see Cassel at al. (1977)). In our case T_S will be a test statistic. The basic parameters of the strategy are:

$$E(T_S, P(s)) = \sum_{s \in \mathbf{S}} t_s P(s), \quad V(T_S, P(s)) = \sum_{s \in \mathbf{S}} (t_s - E(T_S, P(s)))^2 P(s).$$

When T_S is an estimator of a parameter $\theta \in \Theta$, the following mean square error is taken into account:

$$MSE(T_S, P(s)) = \sum_{s \in \mathbf{S}} (t_s - \theta)^2 P(s) = V(T_S, P(s)) + b^2(T_S)$$

where $b(T_S) = E(T_S, P(s)) - \theta$ is bias of estimator T_S.

Let us consider testing the hypotheses on the population error amount defined by expressions (3.2). The test statistic based on the unbiased estimator of the parameter d_U is denoted by:

$$d_S = N(\bar{x}_U - \bar{y}_S), \qquad \bar{y}_S = \frac{1}{n} \sum_{k \in S} y_k.$$

The statistic $\bar{y}_S$ is the mean determined on the basis of the simple random sample drawn without replacement. Its parameters are:

$$E(d_S) = d_U, \qquad V(d_S) = \frac{N(N-n)}{n} v(y).$$

Hence, the statistic d_S is the unbiased estimator of the population total d_U. The unbiased estimator of the variance is:

$$V_S(d_S) = \frac{N(N-n)}{n} v_S(y), \qquad v_S(y) = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y}_S)^2. \tag{3.12}$$

For the simple random sample drawn with replacement the statistic d_S is the unbiased estimator of d_U but its variance is as follows:

$$V(d_S) = N^2 \frac{\sigma^2(y)}{n}. \tag{3.13}$$

The unbiased estimator of the variance is obtained through replacing parameter $\sigma^2(y)$ with $v_S(y)$.

The hypothesis defined by expression (3.2) can be tested by means of the following equivalent statistics:

$$z_S = \frac{d_S - d_U}{\sqrt{V_S(d_S)}} = \frac{d_S - d_U}{\sqrt{V_S(y_S)}} = \frac{y_U - y_S}{\sqrt{V_S(y_S)}} = \frac{\bar{d}_S - \bar{d}_U}{\sqrt{V_S(\bar{d}_S)}} = \frac{\bar{y}_U - \bar{y}_S}{\sqrt{V_S(\bar{y}_S)}}. \tag{3.14}$$

where

$$\bar{d}_S = \frac{d_S}{N} = \bar{x}_U - \bar{y}_S, \qquad V_S(\bar{d}_S) = V_S(\bar{y}_S) = \frac{N-n}{Nn} v_S(y).$$

When we substitute the parameters d_0, y_0, $\bar{d}_0$, $\bar{y}_0$ for d_U, y_U, $\bar{d}_U$, $\bar{y}_U$, respectively, the random variable z_S becomes the test statistic of the hypothesis formulated by expressions (3.2) or (3.4).

3.3.2 Asymptotic normality of test statistics

When we consider the statistical inference based on the randomization approach, the exact distribution of the statistic z_S is dependent on the population parameters $\mathbf{d}$ or $\mathbf{z}$ (see Section 1.1). However, these can be observed only in the sample S. Hence, the exact distribution of z_S cannot be evaluated without additional assumptions. It is only possible to consider the convergence of distribution to normal distribution.

For a simple random sample drawn with replacement and under the assumption that $n \to \infty$ the probability distribution converges to standard normal distribution (see e.g. Cramér (1946)). That conclusion is valid for statistical inference based on random samples considered in some particular cases of the model approach.

In the case of sampling without replacement, the convergence problem is a little more complicated and it was considered e.g. by Madow (1948), Erdös and Rényi (1959) and Hajek (1960). The following additional condition should be considered (see e.g. Cochran (1977) or Bellhouse (2001)). The data are observed in populations $U_1,...,U_k,....,U_a$ with appropriately increasing sizes $N_1 < ... < N_a$. In the population U_k, $k = 1,...,a$ variables $y_{k,i}$, $k = 1,...,a$ and $i = 1,...,N_k$ can be observed. Let S_k be the simple random sample of size n_a drawn from the population U_k, $k = 1,...,a$ and $n_1 < ... < n_a$. Finally, let Q_a be the set of units in the a-th population that:

$$|y_{k,i} - \bar{y}_a| \geq \sqrt{n_a(1 - n_a/N_a)}v_{S_a}(y).$$

Under the introduced notation Lindeberg's condition is:

$$\lim_{a\to\infty} \frac{\sum_{i\in Q_a}(y_{k,i} - \bar{y}_a)^2}{(N_a - 1)v_{S_a}(y)} = 0.$$

When the above condition is fulfilled, the statistic z_S defined by expressions (3.14) has asymptotically standard normal distribution. Hence, we can expect that when a simple random sample of a large size is drawn without replacement from a large-sized population without outlier data, then the probability distribution of z_S can be approximated by means of standard normal distribution.

3.3.3 Bootstrap approach

The well-known bootstrap procedure (see e.g. Efron (1979)) is used to approximate the distribution of statistics by means of normal distribution. Firstly, let us consider the simple random sample of fixed size n drawn with replacement, which we denote by $s = \{k_1,...,k_n\}$. Observations of the variable under study in the sample s will be denoted by $\mathbf{y}_s = [y_{k_1},...,y_{k_n}]$. Let S_j, $j = 1,...,B$ be an independent simple random sample drawn with replacement from the set s. On the basis of the observation $\mathbf{y}_{s_j}$, $j = 1,...,B$ the following statistics are determined:

$$\bar{y}_{S_j} = \frac{1}{n}\sum_{k\in S_j} y_k, \quad j = 1,...,B, \quad \bar{\bar{y}}_S = \frac{1}{B}\sum_{j=1}^{B}\bar{y}_{S_j}, \quad v_{B,S}(y) = \frac{1}{B-1}\sum_{j=1}^{B}(\bar{y}_{S_j} - \bar{\bar{y}}_S)^2. \tag{3.15}$$

The distribution of the statistic defined by expression (3.14) can be approximated by the distribution of the following statistic:

$$z_{B,S} = \frac{\bar{\bar{y}}_S - \bar{y}_S}{v_{B,S}(y)}. \tag{3.16}$$

In the case of a simple random sample drawn without replacement, Gross (1980) and Chao and Lo (1985) consider the following procedure. Let us suppose that the simple sample $s = (k_1,...,k_i,...,k_n)$ of size n is drawn without replacement. We assume that N/n is the integer. Now, we construct an artificial population denoted by U_*, which consists of the sample elements replicated N/n times. Hence, the population U_* is of size N. Next, the simple random samples, denoted by s_J, $j = 1,...,B$, are independently selected without replacement from the population U_*. Then, the statistics defined by expression (3.15) are computed. Moreover, as with sampling with replacement, the statistics defined by expression (3.16) can also be used to test hypothesis (3.2).

Usually, in practice it is stated that the probability distributions of the statistics defined by expression (3.16) are sufficiently well approximated by means of normal distribution. Approximation deals with cases when statistics are evaluated on the basis of the data observed in a simple random sample drawn without replacement as well as a simple random sample drawn with replacement.

Let us note that the above results are valid in the case of a model approach inference based on a simple random sample. More about bootstrap tests can be found in the monographs by Domański et al. (2014) or Hall (1992).

3.3.4 Series expansions

We continue to analyse the possibilities of approximating probability distributions of test statistics. In moderate sizes of samples or populations it is possible to approximate the probability distribution of a statistic using series expansion. Usually, the well-known Edgeworth (1907) series is considered. In this case, the approximated distribution of a considered statistic has to be asymptotically normally distributed.

The standardized central moments and absolute central moments of a one-dimensional random variable Z are defined as follows:

$$\lambda_r(z) = \frac{E(Z-E(Z))^r}{\sigma^r(z)}, \quad \tau_r(z) = \frac{E|Z-E(Z)|^r}{\sigma^r(z)}, \quad r = 1,2,... \quad r = 1,2,...$$

where $\lambda_3(z)$ is the well-known skewness coefficient and $\lambda_4(z)$ is the kurtosis coefficient.

The well-known Berry-Esséen inequality, following Krzyśko (2000), p. 255, is as follows:

$$sup_z|F_s(z)-\Phi(z)| \leq \varsigma \frac{\tau_3(z)}{\sqrt{n}}$$

where $\frac{1}{\sqrt{\pi}} \leq \varsigma < 0.8$, s is a simple random sample of size n, and $F_s(z)$ and $\Phi(z)$ are sample and standard normal distributions, respectively.

In order to simplify our considerations we will take into account the model approach. This means that our sample will be treated as the classical simple random sample from a population with the probability distribution function $F(d,y) = F_1(d)F_2(y)$ considered in Section 1.2 and under the assumption that $p = 1$. According to the model $X = Y + D$, where the values of X, Y and D are observations of book, audited and error amounts, respectively. Under the assumption that Y and D are independent the following expression can be derived:

$$\lambda_4(x)\sigma^4(x) = \lambda_4(y)\sigma^4(y) + 6\sigma^2(y)\sigma^2(d) + \lambda_4(d)\sigma^4(d).$$

The well-known moment inequality (see e.g. Fisz (1967), p. 88) leads to the following:

$$\tau_3^2(z) \leq \lambda_4(z).$$

The above results and the Berry-Esséen inequality lead to the following:

$$\tau_3^2(x) \leq \lambda_4(x) = \lambda_4(y)\zeta^2 + 6\zeta(1-\zeta) + \lambda_4(d)(1-\zeta)^2,$$

where

$$\zeta = \frac{\sigma^2(y)}{\sigma^2(y)+\sigma^2(d)} \in (0;1]$$

is called the reliability coefficient (see expression (1.7)).

We are going to consider the properties of the probability distributions of the following statistics (see expression (3.14)):

$$z_S(x) = \frac{\bar{x}_U - \bar{x}_S}{\sqrt{v_S(x)}}\sqrt{n}, \qquad z_S(y) = \frac{\bar{y}_U - \bar{y}_S}{\sqrt{v_S(y)}}\sqrt{n}.$$

Based on the Berry-Esséen inequality we have:

$$sup_x|F_s(z_s(x)) - \Phi(z_s(x))| \leq \varsigma \frac{\tau_3(x)}{\sqrt{n}} \leq \varsigma\sqrt{\frac{\lambda_4(x)}{n}} = \\ = \varsigma\sqrt{\frac{\lambda_4(y)\zeta^2 + 6\zeta(1-\zeta) + \lambda_4(d)(1-\zeta)^2}{n}}. \quad (3.17)$$

When the reliability coefficient is close to one (which is equivalent to a very small variance of error), then we have approximately $\lambda_4(x) \approx \lambda_4(y)$ and

$$sup_y|F_s(z_s(y)) - \Phi(z_s(y))| \leq \varsigma \frac{\tau_3(y)}{\sqrt{n}} \leq \varsigma \sqrt{\frac{\lambda_4(y)}{n}} \approx \varsigma \sqrt{\frac{\lambda_4(x)}{n}}. \tag{3.18}$$

Hence, when $\zeta \to 1$, the right sides of the above two inequalities are close to each other. Therefore, we can expect that the distribution of both statistics $z_S(x)$ and $z_S(y)$, are approximately close to each other when the reliability coefficient is close to one. Hence, the particular parameters of the distribution of the test statistic $z_S = z_S(y)$, defined by expression (3.14), can be approximated by the parameters of the statistic $z_S(x)$, which are known in advance.

The asymptotic normality of the probability distribution of the above statistics $z_S(x)$ and $z_S(y)$ can be proved similarly to how the normality of the statistics considered in section 3.3.2 can be proved.

Now we consider that observed sample s is drawn from a finite population of size N. In this case, the Edgeworth expansion of the probability distribution of the test statistic $z_S(y)$, given by expression (3.14), is defined by Babu and Singh (1985). It is as follows:

$$P(z_S(y) < z) = \Phi(z) + \frac{1}{6\sqrt{n}}\lambda_3(y)\left(3z^2 - \frac{N-2n}{N-n}(z^2-1)\right)\phi(z)\sqrt{\frac{N-n}{n}} + o(n^{-1/2}) \tag{3.19}$$

where $\phi(z)$ is the density function of standard normal distribution.

The skewness coefficient $\lambda_3(x)$ can be decomposed as follows:

$$\lambda_3(x) = \lambda_3(y)\zeta^3 + \lambda_3(d)(1-\zeta)^3.$$

When the reliability coefficient $\zeta \to 1$, then $\lambda_3(x) \to \lambda_3(y)$. Hence, when N, n and $N-n$ are sufficiently large, the probability given by (3.19) can be approximated by the following:

$$\begin{aligned} &P(z_S(y) < z) \approx P(z_S(x) < z) = \\ &= \Phi(z) + \frac{1}{6\sqrt{n}}\lambda_3(x)\left(3z^2 - \frac{N-2n}{N-n}(z^2-1)\right)\phi(z)\sqrt{\frac{N-n}{n}} + o(n^{-1/2}). \end{aligned} \tag{3.20}$$

According to the model assumption introduced in Chapter 1, the skewness coefficient $\lambda_3(x)$ is known in advance because it can be evaluated on the basis of all observations of the book amount observed in a population, which are treated as auxiliary data. Hence, the above formula lets us evaluate the p-values of tests based on the statistics defined by expression (3.17) under the assumption that the sample size is sufficiently large.

Babu and Singh (1985) suggested that the bootstrap technique allows for approximation using thr Edgeworth series expansion defined by expression (3.19). In order to do this, the statistic $z_S(y)$ should be approximated by the bootstrap type statistic defined by expressions (3.16) and the skewness coefficient $\lambda_3(y)$ should be replaced by its estimator based on sample S. This leads to the following formula:

$$P(z_{B,S}(y) < z) \approx \Phi(z) + \frac{1}{6\sqrt{n}}\lambda_{3,S}(y)\left(3z^2 - \frac{N-2n}{N-n}(z^2-1)\right)\phi(z)\sqrt{\frac{N-n}{n}}. \tag{3.21}$$

More properties of the above type of approximation are considered by Babu and Singh (1984).

Finally, let us consider the Edgeworth series expansion for a studentized sample mean from the well-known simple exponential distribution with the density function $f(y) = \beta \exp\{-\beta y\}$ for $y > 0$ and $f(y) = 0$ for $y \leq 0$. Its expected value and variance are: $\mu = E(Y) = 1/\beta$ and $V(Y) = 1/\beta^2$. Exponential distribution is frequently considered in financial auditing for modelling the book amount and the audited amount. The test statistic can be formulated as follows (see expression (3.14)):

$$z_S = \frac{\mu - \bar{y}_S}{\sqrt{v_S(y)}}\sqrt{n}.$$

Hall (1992), pp. 70-71 and 116 derived the following series expansion:

$$P(z_S < z) = \Phi(z) + n^{-1/2}q_1(z)\phi(z) + n^{-1}q_2\phi(z) + ... \tag{3.22}$$

where

$$\begin{cases} q_1(z) = \frac{1}{6}\lambda_3(y)(z^2+1), \\ q_2(z) = z\left(\frac{1}{12}(\lambda_4(y)-3)(z^2-3) - \frac{1}{18}\lambda_3(y)(z^4+2z^2-3) - \frac{1}{4}(z^2+3)\right) \end{cases}$$

In the case of exponential distribution we have: $\lambda_3(y) = 2$ and $\lambda_4(y) = 6$. Therefore, the above system of equations becomes as follows:

$$\begin{cases} q_1(z) = \frac{1}{3}(z^2+1), \\ q_2(z) = -\frac{z}{3}\left(\frac{1}{3}z^4 + \frac{2}{3}z^2 + \frac{7}{2}\right). \end{cases} \tag{3.23}$$

Hence, (3.22) and (3.23) let us approximate the p-value of the test defined by (3.2) or (3.4). On the basis of (3.14), we can calculate the value z_s of the test statistic:

$$z_S = \frac{d_S - d_0}{\sqrt{V_S(y_S)}} = \frac{\bar{y}_0 - \bar{y}_S}{\sqrt{V_S(\bar{y}_S)}} = \frac{\bar{y}_0 - \bar{y}_S}{\sqrt{v_S(y)}}\sqrt{n}.$$

Hence, the p-value is determined by the following expression:

$$p = 1 - P(z_S < z_s|H_0)$$

where the probability $P(z_S < z_s|H_0)$ is evaluated by means of the right side of equation (3.22) for $z = z_s$.

More about the applications of series expansion for approximating the probability distribution of the mean value can be found in the paper by Helmers (2000), where the critical value of the test is evaluated using the Cornish-Fisher formula. He also considers the bootstrap variants of series expansion.

3.3.5 Evaluating necessary sample size to fit statistic distribution with normal distribution

In section 3.3.2, on the basis of the central theorem, we concluded that with a sufficiently large sample size the distribution of the statistics defined by expression (3.14) are well-approximated by standard normal distribution. Our purpose now is to assess that sample size.

Evaluating necessary sample size using the Berry-Esséen inequality

On the basis of expression (3.17) and (3.18), we concluded that under the assumption that the reliability coefficient is close to one, the distributions of the statistics $z_S(y)$ and $z_S(x)$ are approximately close to each other. Let ε be the assumed accuracy of approximation of the distribution of the statistic. Expression (3.18) lets us write the following:

$$sup_y|F_s(z_s(y)) - \Phi(z_s(y))| \leq \varepsilon \approx \varsigma\sqrt{\frac{\lambda_4(x)}{n}}. \tag{3.24}$$

This leads to the following approximation of the necessary sample size:

$$n_\varepsilon \approx \frac{0.64\lambda_4(x)}{\varepsilon^2}. \tag{3.25}$$

The derived approximation of the necessary sample size cannot be treated as accurate, but it lets us at least assess an order of magnitude of the value n_ε.

Simulation evaluation of necessary sample size to test hypothesis on normality

As in the previous paragraph, we assume that the distributions of the statistics $z_S = z_S(y)$ and $z_S = z_S(x)$ are very close to each other. The next procedure for assessing the necessary sample size is to test the normality of statistics evaluated based on the empirical distribution of the statistic $z_S(x)$. The proposed procedure is as follows.

Let s_j, $j = 1,...,a$ be simple random samples (each of size n_l) independently drawn from a population where all book amounts are observed. Next, the following values are calculated:

$$z_{s_j}(x) = \frac{\bar{x}_{s_j} - \bar{x}}{v_{s_j}(x)}\sqrt{n_l}, \qquad \bar{x}_{s_j} = \frac{1}{n_l}\sum_{k\in s_j} x_k, \quad v_{s_j}(x) = \frac{1}{n_l - 1}\sum_{k\in s_j}(x_k - \bar{x}_{s_j})$$

where $j = 1,...,a$, $n_l > 2$. The observed data $\{z_{s_j}(x)\}$, $j = 1,...,a$ lets us test the hypothesis on the normality of the distribution of the statistic $z_S(x)$. More precisely, we test the hypothesis that the distribution of $z_S(x)$ is the same as standard normal distribution. When the hypothesis on normality is not rejected then we treat the

distribution of $z_S(x)$ as sufficiently close to standard normal distribution under the sample size n_l, $l = 1,...$ When the hypothesis is rejected then we increase the sample size to the level $n_{l+1} = n_l + c$ (where $c \geq 1$ is the integer) and the hypothesis on the normality of $z_S(x)$ is tested again. Of course the hypothesis is rejected under the assumed significance level α and it is accepted under the power of the test denoted by β. As is well known, the test can reach the assumed levels α and β under the appropriate number a of observations of $z_S(x)$. Hence, before testing normality we should first determine the number a. This depends on the construction of the considered test of goodness of fit.

It seems that the well-known chi-square test for goodness of fit will be convenient in our case to test normality. Let $\omega_g = P(z_{g-1} < Z \leq z_g) = \Phi(z_g) - \Phi(z_{g-1})$ where $\Phi_g = \Phi(z_g) = P(Z < z_g)$, $Z \sim N(0;1)$, $z_0 = -\infty$, $z_{G+1} = \infty$ and $g = 0,1,...,G+1$, $G \geq 2$. Let $\Phi_0 = [\Phi_{0,1}...\Phi_{0,G}]$, $\Phi_1 = [\Phi_{1,1}...\Phi_{1,G}]$ and $\omega_0 = [\omega_{0,1}...\omega_{0,G+1}]$, $\omega_1 = [\omega_{1,1}...\omega_{1,G+1}]$.

Our hypotheses on normality are as follows:

$$H_0 : \Phi = \Phi_0, \qquad H_1 : \Phi = \Phi_1 \neq \Phi_0. \tag{3.26}$$

These are equivalent to the following:

$$H_0 : \omega = \omega_0, \qquad H_1 : \omega = \omega_1 \neq \omega_0.$$

Hence, testing the hypothesis on normality is reduced to testing the hypothesis on the vector of probabilities ω. We expect that the tails of the distribution of the test statistic $z_S(x)$ and the appropriate tails of standard normal distribution are well fitted. That is why, in practice, the vectors Φ and ω will be, for instance, constructed as follows:

$$\Phi_0^{(1)} = [0.01\ \ 0.05\ \ 0.1\ \ 0.9\ \ 0.95\ \ 0.99],$$

$$\omega_0^{(1)} = [0.01\ \ 0.04\ \ 0.05\ \ 0.8\ \ 0.05\ \ 0.04\ \ 0.01],$$

$$\omega_1^{(1)} = [0.012\ \ 0.048\ \ 0.06\ \ 0.76\ \ 0.06\ \ 0.048\ \ 0.012],$$

$$\omega_1^{(2)} = [0.011\ \ 0.044\ \ 0.055\ \ 0.78\ \ 0.055\ \ 0.044\ \ 0.011],$$

or

$$\Phi_0^{(2)} = [0.01\ \ 0.05\ \ 0.1], \qquad \omega_0^{(2)} = [0.01\ \ 0.04\ \ 0.05\ \ 0.9],$$

$$\omega_1^{(3)} = [0.012\ \ 0.048\ \ 0.06\ \ 0.88], \qquad \omega_1^{(4)} = [0.011\ \ 0.044\ \ 0.055\ \ 0.89].$$

The chi-square test statistic is:

$$T_a = a \sum_{g=1}^{G+1} \frac{(W_g - \omega_g)^2}{\omega_g} \tag{3.27}$$

where

$$W_g = \frac{M_g}{a}, \qquad M_g = \sum_{j=1}^{a} I_{g,j}, \qquad a = \sum_{g=1}^{G+1} M_g,$$

if $z_{g-1} < z_{s_j}(x) \leq z_g$ then $I_{g,j} = 1$ and otherwise $I_{g,j} = 0$, $j = 1,...,a$, $g = 1,...,G+1$. The vector $\mathbf{M} = [M_1...M_g...M_{G+1}]$ has multinomial probability distribution with parameters ω and a. When the parameter a is large, the statistic T_a has chi-square probability distribution with G degrees of freedom and the non-centrality parameter:

$$\Delta_\omega = a \sum_{g=1}^{G+1} \frac{(\omega_{1,g} - \omega_{0,g})^2}{\omega_{0,g}},$$

so $T_a \sim \chi^2_G(\Delta_\omega)$. When hypothesis H_0 is true and the sample size is large, then the statistic has central chi-square distribution with G degree of freedom, so $T_a \sim \chi^2_G$. According to Cochran (1952), the statistic T_a is sufficiently well approximated by the chi-square distribution when the sample size fulfils the inequality: $min_{g=1,...,G+1}\{a\omega_g\} \geq 5 = E_0$. Santner and Duffy (1989), p. 65, wrote: $E_0 = 1$. For instance, according to Cochran's recommendation, when $min\{\omega\} = 0.01$ then $a \geq 500$.

The necessary sample size for testing the above hypotheses under the assumed significance level and the power of the test is partially evaluated on the basis of the algorithm proposed in Chapter 2.3.3, where the simulation technique is used to assess the necessary sample size needed to test the hypotheses defined by (2.1) or (2.2).

Let us assume that the above hypothesis will be tested using the statistic T_a under the significance level α and the power β. Our purpose is to evaluate number $a_{\alpha,\beta}$ so that under the stated hypotheses H_0 and H_1 the test based on statistic T_a has a significance level equal to α and a power equal to β.

Let each of the independently distributed random vectors $\mathbf{M}_h$, $h = 1,...,A$, have multinomial distribution with parameters ω_1 and a_l, $l = 1,....$ The generated value of the variable $\mathbf{M}_h$ will be denoted by $\mathbf{m}_h$. Let $t_{a_l,h}$ be the value of $T_{a_l,h}$ evaluated on the basis of independently generated values $\mathbf{m}_h$, $h = 1,....A$. The critical value of the test based on the statistic T_{a_l} will be denoted by t_α.

This leads to evaluating the set of values: $S_{\#l} = \{t_{a_l,h}, h = 1,...,A\}$. The power of the test is assessed according to the following expression:

$$b_l = \frac{1}{A} \sum_{h=1}^{A} I(t_{a_l,h}),$$

where $I(t_{a_l,h}) = 0$, if $t_{a_l,h} < t_\alpha$ or $I(t_{a_l,h}) = 1$, if $t_{a_l,h} \geq t_\alpha$. When $b_l \geq \beta$ then the algorithm is stopped and the sample size $a_{\alpha,\beta} = a_l$ is treated as sufficient for testing the above hypothesis under the assumed significance level α and power β. In the case when $b_l < \beta$, we assume that $n_{l+1} > n_l$ and the algorithm is started again. The presented algorithm is implemented by the program in Appendix 6.2.5.

The critical value of the test can be determined based on the limit distribution of the chi-square test statistic or through simulation. In the former case, large number

values of the test statistic are evaluated based on the independently replicated samples of fixed size n from multinomial distribution with the parameters ω_0. That set of values is ordered from the smallest to the largest. Finally, based on this sequence, the value t_α is determined as the $1-\alpha$ sample quantile. The critical value determined in this way will be called the simulated critical value of the chi-square test of goodness of fit.

Table 3.1 shows the necessary sample size evaluated by means of that program on the basis of $A = 100000$ replications of the test statistic values.

Table 3.1: The number of sample sizes evaluated on the basis of the limit distribution of the chi-square test statistic.

		$\omega_0^{(1)}$	$\omega_0^{(1)}$	$\omega_0^{(1)}$	$\omega_0^{(1)}$	$\omega_0^{(2)}$	$\omega_0^{(2)}$	$\omega_0^{(2)}$	$\omega_0^{(2)}$
		$\omega_1^{(1)}$	$\omega_1^{(1)}$	$\omega_1^{(2)}$	$\omega_1^{(2)}$	$\omega_1^{(3)}$	$\omega_1^{(3)}$	$\omega_1^{(4)}$	$\omega_1^{(4)}$
α	β	*limit*	*simulated*	*limit*	*simul.*	*limit*	*simul.*	*limit*	*simul.*
0.1	0.9	1470	1410	5880	5860	2660	2750	10620	10760
0.05	0.95	2090	2110	8350	8390	3870	4040	15460	15760
0.01	0.99	3510	3630	14010	14120	6720	7050	26850	27270
0.005	0.995	4110	4230	16420	16650	7950	8450	31790	32080
0.001	0.999	5490	5720	21960	22080	10830	10040	43290	44090
	Δ_ω/a :	0.01	0.01	0.0025	0.0025	0.0044	0.0044	0.0011	0.0011

An analysis of Table 3.1 lets us say that the necessary sample size increases when the significance level of the chi-square test of goodness of fit decreases or its power increases. The necessary sample size increases when the distance coefficient Δ_ω decreases. The sample sizes evaluated under the critical values determined on chi-square distribution are very similar to the appropriate sample sizes determined under the simulated critical value of the test.

The next iteration of the algorithm leads to the evaluation of the sample size n_{nor}, under which the sample distribution of the statistic $z_S(x)$ fits with standard normal distribution in the sense of the hypotheses specified by expression (3.26). It is as follows. Let n_l be the sample size obtained during l-th iteration ($l = 1,2,...$) of the algorithm. The elements of the series $(s_{l,t})$, $t = 1,...,a_{\alpha,\beta}$ are simple random samples of size n_l independently drawn without replacement from the population U where values of the variable x are observed. On the basis of each sample the values $z_{s_{l,t}}(x)$, $t = 1,...,a_{\alpha,\beta}$ are calculated. Finally, the value t_{a_l} of the chi-square test statistic T_{a_l} is evaluated.

When $P(T_{a_l} \geq t_{a_l}|H_0) > \alpha$ then we accept the hypothesis that the distribution of statistic $z_{S_l}(x)$ from the sample of size $n_l = n_{nor,\alpha,\beta}$ is sufficiently close to the standard normal distribution. This decision can be wrong with probability $(1-\beta)$. The algorithm is stopped.

If $P(T_{a_l} \geq t_{a_l}|H_0) \leq \alpha$ then we reject the hypothesis that the distribution of statistic $z_{S_l}(x)$ from the sample of size n_l is sufficiently close to the standard normal distribution. This decision can be wrong with probability α. In this situation we start the $(l+1)$-th stage of the algorithm from assigning a new sample size $n_{l+1} > n_l$.

Example 3.1. Let us use the above algorithm to evaluate the necessary sample size for testing the hypothesis on the normality of statistic $z_S(x)$ under an assumed significance level and power of the chi-square test of goodness of fit. The samples were drawn from a population of size $A = 100000$, consisting of generated pseudo-values from gamma distribution with a scale parameter equal to 1 and shape parameter equal to 1. For instance, let us consider the following hypotheses:

$$\begin{cases} H_0: \ \omega = [[0.01\ 0.04\ 0.05\ 0.8\ 0.05\ 0.04\ 0.01], \\ H_1: \ \omega = [0.011\ 0.044\ 0.055\ 0.78\ 0.055\ 0.044\ 0.011]. \end{cases}$$

The computer evaluation leads to the following results. Under $\alpha = 0.05$ and $\beta = 0.95$ and a simple random sample drawn without replacement, the necessary sample sizes were $n_{0.05,0.95} = 8350$, $n_{nor,0.05,0.95} = 1040$ or $n_{0.1,0.9} = 5880$ and $n_{nor,0.1,0.9} = 930$. In the case of sampling with replacement, the necessary sample sizes were $n_{nor,0.05,0.95} = 1060$ and $n_{nor,0.1,0.9} = 920$. Hence, the necessary sample sizes for sampling with or without replacement were similar.

The described algorithm can be implemented using some computer programs written in the R programming language. These can be found in Appendix 6.2.5. Finally, let us note that some other ideas of evaluating the necessary sample size are considered by Lalu and Krishnan (1978) or Pekasiewicz (2010).

3.3.6 Evaluating necessary sample size to test hypotheses on the total amount of error under assumed auditing risks

Evaluating necessary sample size under assumed risk of incorrect rejection and risk of incorrect acceptance

Let us consider testing the following hypotheses:

$$H_0: \ d_U \leq d_0, \qquad H_1: \ d_U \geq d_1 > d_0 \geq 0, \tag{3.28}$$

where d_0 is the admissible level of the total book value and d_1 is the inadmissible level.

The formulated hypotheses can be tested by means of the following statistic (see expression (3.14)):

$$z_S = \frac{d_S - d_U}{\sqrt{V_S(y_S)}} \tag{3.29}$$

where d_S is the asymptotically unbiased estimator of d_U and $V_S(d_S)$ is the consistent estimator of $V(d_S)$. We assume that z_S is well approximated by means of normal distribution under the sufficiently large sample size n. Under the assumed significance level α equal to the risk of incorrect rejection η we have:

$$\eta = P\left(d_S \geq z_\eta \sqrt{V_S(y_S)} + d_0 \middle| H_0\right). \tag{3.30}$$

It is equivalent to the following:

$$\eta = P(Z_S \geq z_\eta | H_0) \tag{3.31}$$

The risk of incorrect acceptance is:

$$\kappa = P\left(d_S \leq z_\eta \sqrt{V_S(y_S)} + d_0 \middle| H_1\right). \tag{3.32}$$

It can be transformed into the following form:

$$\kappa = P\left(z_S = \frac{d_S - d_1}{\sqrt{V_S(y_S)}} \leq z_\eta + \frac{d_0 - d_1}{\sqrt{V_S(y_S)}} \middle| H_1\right),$$

Let us suppose that the statistic $V_S(y_S)$ can be rewritten in the following way:

$$V_S(y_S) = \frac{V_S}{n}, \tag{3.33}$$

where $V_S > 0$. Under assumed normality of Z_S we have:

$$\kappa = \phi\left(z_\eta + \frac{d_0 - d_1}{\sqrt{V_S(y_S)}}\right) = \phi\left(z_\eta + \frac{d_0 - d_1}{\sqrt{V_S}}\sqrt{n}\right) = \phi(z_\kappa). \tag{3.34}$$

This leads to the conclusion that under the assumed risk of incorrect rejection η and fixed v_s, the risk of incorrect acceptance κ converges to zero, when $n \to \infty$.

Therefore, if $z_s \geq z_\eta$, then we reject hypothesis H_0 with the probability of a wrong decision equal to η. If $z_s < z_\eta$, we accept hypothesis H_0 with risk equal to κ.

Expression (3.34) leads to the following:

$$z_\kappa = z_\eta + \frac{d_0 - d_1}{\sqrt{V_S(y_S)}}. \tag{3.35}$$

In the case of a simple random sample drawn without replacement, $V_S(d_S) = \frac{N-n}{Nn} v_{*S}(y)$. Hence, the above expression lets us evaluate the following:

$$n \geq n_{\eta,\kappa} = \left(\frac{1}{N} + \frac{(d_0 - d_1)^2}{(z_\kappa - z_\eta)^2 N^2 v_{*s}(y)}\right)^{-1} \approx \frac{N^2 v_{*s}(y)(z_\kappa - z_\eta)^2}{(d_0 - d_1)^2} = n_a, \tag{3.36}$$

where $\phi(z_\kappa) = \kappa$, $\phi(z_\eta) = 1 - \eta$ and v_s is assessed e.g. on the basis of a pilot sample or some previous survey. Therefore, a sample size not less than $n_{\eta,\kappa}$ leads to testing the hypotheses defined by (3.28) with the risk of incorrect rejection equal to η and risk of the incorrect acceptance equal to κ.

Now let us assume that the sample is drawn using the stratified sampling design under proportionally determined sizes of sub-samples (to the size of the strata) selected from the strata. Similarly, as above, we can derive the following necessary sample size:

$$n \geq n_{\eta,\kappa} = \left(\frac{1}{N} + \frac{(e_0 - e_1)^2}{(z_\kappa - z_\eta)^2 N^2 v_{*prop}(y)} \right)^{-1} \approx$$
$$\approx \left(\frac{z_\kappa - z_\eta}{e_0 - e_1} \right)^2 N^2 v_{*prop}(y) = n_{ap}, \quad (3.37)$$

where (see section 3.2.3):

$$v_{*prop}(y) = \sum_{h=1}^{H} w_h v_{*s_h}(y), \qquad n_{\eta,\kappa} < n_{ap}.$$

In practice, the variance v_{*s_h} can be evaluated on the basis of a pilot or some previous survey.

The results of this sub-section can be generalized into the following hypotheses:

$$H_0 : |x_U - y_U| = |d_U| \leq d_0, \qquad H_1 : |d_U| \geq d_1 > d_0,$$

where d_0 is the admissible level of difference between total book values and the true total, and the inadmissible level is denoted by d_1. The stated problem is considered e.g. by Wywiał (2014a).

Testing hypotheses about total amount under assumed risk of incorrect rejection

In practice parameter d_1 in (3.28) does not have to be specified. In this situation, those hypotheses can be reduced to the following:

$$H_0 : d_U \leq d_0, \qquad H_1 : d_U > d_0, \quad (3.38)$$

where the admissible level of total book value $d_0 > 0$. In this situation, it is not possible to control the risk of incorrect acceptance κ. Now it is only possible to make a decision regarding the rejection of hypothesis H_0 under the risk of incorrect rejection $\eta = \alpha$. On the basis of expression (3.29) the value of the test statistic z_S is evaluated. The critical value of the test is defined by expression (3.31). Hence, if $z_s \geq z_\eta$, then hypothesis H_0 is rejected with risk of incorrect rejection equal to η. When $z_s < z_\eta$, hypothesis H_0 cannot be rejected but it cannot be accepted either, because the risk κ cannot be assessed.

Testing hypotheses about total amount under assumed risk of incorrect acceptance

In order to control the risk of incorrect acceptance the following hypotheses (in some sense dual to those defined in (3.38)) can be specified:

$$H_0' : d_U \geq d_1, \qquad H_1' : d_U < d_1 \tag{3.39}$$

where the inadmissible level of total book value $d_1 > 0$. Now the rejection of hypothesis H_0' by a test means accepting the accounting report as correct with risk equal to κ. If the test does not reject hypothesis H_0', this does not mean that the accounting report is wrong, because it is not possible to assess the risk of its incorrect rejection $\eta = 1 - \beta$. Hence, in this case, the assumed risk of incorrect acceptance is $\kappa = \alpha$, where α is the significance level of the test. The critical value z_κ of the test is determined by the equation:

$$\kappa = P(z_S \leq z_\kappa | H_0').$$

Similarly, on the basis of expression (3.29) we can evaluate the value z_s of the test statistic. Therefore, if $z_s \leq z_\kappa$, then hypothesis H_0' is rejected under significance level $\alpha = \kappa$. This means, for example, that the accounting report is accepted with the risk of incorrect acceptance equal to $\kappa = \alpha$. When $z_s > z_\kappa$, hypothesis H_0' is not rejected. This only means that the accounting report could be incorrect. In this case we cannot decide that the report is wrong because the risk η cannot be assessed.

The idea can be used e.g. for the problem of auditing tax revenues. In this case, d_1 can be treated as the inadmissible total of unpaid taxes.

3.4 Ratio and regression statistics

In auditing, book amounts and audited amounts are usually highly correlated. That is why regression and ratio estimators are useful in auditing research. In this case, the parameters of these estimators take a specific form. Our considerations are based on the assumption that $X = Y + D$ (see subsection 1.2). The simple random sample is from the distribution of the random variables $[X, Y]$.

The ratio estimator of the mean audited amount is as follows:

$$\bar{y}_{I,S} = \frac{\bar{y}_S}{\bar{x}_S} \bar{x}_U . \tag{3.40}$$

It is the asymptotically unbiased estimator of $\bar{y}$ and its variance under a simple random sample drawn without replacement is given by (see, e.g. Cochran (1977) or Konijn (1973)):

$$V(\bar{y}_{I,S}) = \frac{1}{n}\sigma^2(y)\left(1+\left(\frac{\gamma(x)}{\gamma(y)}\right)^2 - 2\frac{\gamma(x)}{\gamma(y)}\rho(x,y)\right) + O(n^{-3/2}), \tag{3.41}$$

where $\gamma(x) = \sigma(x)/\bar{x}_U$, $\gamma(y) = \sigma(y)/\bar{y}_U$, $O(n^{-r}) = \frac{c}{n^r}$, $r > 0$ and c is constance. The bias of $\bar{y}_{I,S}$ equals:

$$b_I = E(\bar{y}_{I,S}) - \bar{y} = \frac{1}{n}\sigma(y)\gamma(x)\left(\frac{\gamma(x)}{\gamma(y)} - \rho(x,y)\right) + O(n^{-2}).$$

When $\rho(x,y) > 0$, then:

$$\frac{|b_I|}{\sqrt{V(\bar{y}_{I,S})}} \leq \gamma(x).$$

The unbiased estimator of $V(y_{I,S})$ is as follows:

$$V_S(\bar{y}_{I,S}) = \frac{1}{n(n-1)}\sum_{k\in S}^{N}(y_k - h_S x_k)^2, \qquad h_S = \frac{\bar{y}_S}{\bar{x}_S}. \tag{3.42}$$

When the random variables Y and D are not correlated and $X = Y + D$, then expression (3.41) simplifies to:

$$V(\bar{y}_{I,S}) = \frac{1}{n}\sigma^2(y)\left(1+h^2\left(1+\frac{\sigma^2(d)}{\sigma^2(y)}\right) - 2h\right) + O(n^{-3/2}),$$

where $h = \frac{\bar{y}}{\bar{x}}$. The relative efficiency coefficient is:

$$deff(\bar{y}_{I,S}) = \frac{V(\bar{y}_{I,S})}{V(\bar{y}_S)} = 1 + h^2\left(1+\frac{\sigma^2(d)}{\sigma^2(y)}\right) - 2h,$$

$deff(\bar{y}_{I,S}) < 1$, if $h/2 < \rho^2(x,y) = \frac{\sigma^2(y)}{\sigma^2(y)+\sigma^2(d)}$. Hence, the ratio estimator is more precise than the simple random sample when the squared reliability coefficient is greater than $\frac{\bar{y}}{2\bar{x}}$. Particularly, $deff(\bar{y}_{I,S}) < 1$, when when $h = 1$ and $\sigma^2(y) > \sigma^2(d)$.

The well-known regression estimator of population mean $\bar{x}$ is as follows:

$$\bar{y}_{reg,S} = \bar{y}_S - b(\bar{x}_S - \bar{x}),$$

where

$$b = \rho(x,y)\frac{\sigma(y)}{\sigma(x)} = \frac{\sigma(x,y)}{\sigma^2(x)}.$$

The variance is:

$$V(\bar{y}_{reg,S}) = \frac{\sigma^2(y)}{n}\left(1-\rho^2(x,y)\right).$$

The relative efficiency is:

$$def(\bar{y}_{reg,S}/\bar{y}_S) = \frac{V(\bar{y}_{reg,S})}{V(\bar{y}_S)} = 1 - \rho^2(x,y).$$

When the random variables Y and D are not correlated and $X = Y + D$ then the above expression defining the slope coefficient b reduces to Gilford's reliability coefficient (see Chapter 1):

$$b = \rho^2(x,y) = \zeta = \frac{\sigma^2(y)}{\sigma^2(y) + \sigma^2(d)}.$$

Usually, the value of the coefficient is not known. That is why the following version of the regression estimator is considered in practice:

$$\hat{y}_{reg,S} = \bar{y}_S - b_S(\bar{x}_S - x_U) \tag{3.43}$$

where

$$b_S = \frac{v_S(x,y)}{v_S(x)}. \tag{3.44}$$

The statistic $\hat{y}_{reg,S}$ is the asymptotically unbiased estimator of $\bar{y}$. The variance of $\hat{y}_{reg,S}$ is estimated by means of the statistic:

$$V_S(\hat{y}_{reg,S}) = \frac{1}{n} v_S(y)\left(1 - \rho_S^2(x,y)\right), \tag{3.45}$$

where

$$\rho_S(x,y) = \frac{v_S(x,y)}{\sqrt{v_S(x)v_S(y)}}.$$

The bias of the regression estimator $\hat{y}_{reg,S}$ is approximately given by (see. Wywiał(1992, 2003)):

$$bias(\hat{y}_{reg,S}) = \frac{\sigma(y)\sqrt{\lambda_4(x) - 1}}{n}\left(\theta_{2,1} - \theta_3(x)\rho(x,y)\right)$$

where:

$$\theta_{2,1}(x,y) = \frac{\sigma_{2,1}(x,y)}{\sigma(y)\sqrt{\sigma_4(x) - \sigma^4(x)}}, \quad \theta_3(x) = \frac{\sigma_3(x)}{\sigma(x)\sqrt{\sigma_4(x) - \sigma^4(x)}},$$

$$\lambda_4(x) = \frac{\sigma_4(x)}{\sigma^4(x)}, \quad \sigma_{a,b}(x,y) = E(X - E(X))^a(Y - E(Y))^b, \quad \sigma_a(x) = \sigma_{a,0}(x,y),$$

$$\sigma_a(x) = \sigma_{a,0}(x,y), \quad \sigma(x) = \sqrt{\sigma_2(x)}.$$

Moreover, $\theta_{2,1}(x,y) \in [-1;1]$ and $\theta_3(x) \in [-1;1]$ is the normalized skewness coefficient (see Wywiał, (1981, 1982)).

When the random variables Y and D are not correlated and $X = Y + D$ then the above expression defining the slope coefficient b can be estimated by means of (3.44) or by:

$$\zeta_S = \frac{v_S(y)}{v_S(y) + v_S(d)}. \tag{3.46}$$

Moreover, in this case, the variance of the regression estimator can be estimated as follows:

$$V_S(\hat{y}_{reg,S}) = \frac{1}{n} v_S(y)\,(1-\zeta_S)\,. \tag{3.47}$$

Expressions (3.40), (3.42) and (3.43) and (3.45)-(3.47) let us construct the following test statistics:

$$Z_{I,S} = \frac{d_{I,S}-d}{\sqrt{V_S(d_{I,S})}}, \quad Z_{reg,S} = \frac{d_{reg,S}-d}{\sqrt{V_S(d_{reg,S})}} \tag{3.48}$$

where $d_{I,S} = x_U - N\bar{y}_{I,S}$, $V_S(d_{I,S}) = N^2 V_S(\bar{y}_{I,S})$ and $d_{reg,S} = x_U - N\bar{y}_{reg,S}$, $V_S(d_{reg,S}) = N^2 V_S(\bar{y}_{reg,S})$. Hence, the two above statistics can be used to test the hypotheses formulated by (3.1).

The variances $V(\bar{y}_{reg,S})$ and $V(\bar{y}_{I,S})$ can also be estimated by means of the bootstrap method. Finally, let us note that it is possible to generalize the obtained results into a case where the sample is selected by means of more complex sampling schemes than simple random sampling using the design-based approach.

3.5 Mean from stratified random sample

3.5.1 Basic properties

The stratified sampling design was presented in section 3.2.3. The estimator of the total value of the population y_U is:

$$y_{w,S} = N\bar{y}_{w,S}, \tag{3.49}$$

where

$$\bar{y}_{w,S} = \sum_{h=1}^{H} w_h \bar{y}_{S_h}, \quad \bar{y}_{S_h} = \frac{1}{n_h}\sum_{k\in S_h} y_k. \tag{3.50}$$

The estimator of d_U is as follows:

$$d_{w,S} = x_U - y_{w,S}. \tag{3.51}$$

The statistics $d_{w,S}$ and $y_{w,S}$ are the unbiased estimators of the parameters d_U and y_U, respectively.

For simple random samples drawn without replacement from the strata the variance of $y_{w,S}$ is as follows:

$$V(y_{w,S}, P_3(s)) = N^2 V(\bar{y}_{w,S}, P_3(s)) = \sum_{h=1}^{H} \frac{N_h(N_h-n_h)}{n_h} v_{*U_h}(y), \tag{3.52}$$

where sampling design $P_3(s)$ is defined in Subsection 3.2.3 and

$$v_{*U_h}(y) = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2, \qquad \bar{y}_{U_h} = \frac{1}{N_h - 1} \sum_{k \in U_h} y_k.$$

For sampling with replacement the variance is as follows:

$$V(y_{w,S}, P_4(s)) = N^2 V(\bar{y}_{w,S}, P_4(s)) = \sum_{h=1}^{H} N_h^2 \frac{v_{*U_h}(y)}{n_h}, \tag{3.53}$$

where $P_4(s)$ is explained in Subsection 3.2.3 and

$$v_{*U_h}(y) = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2, \qquad \bar{y}_{U_h} = \frac{1}{N_h - 1} \sum_{k \in U_h} y_k.$$

In the case of sampling without replacement, the unbiased estimator of the variances $V(y_{w,S}, P_3(s))$ is:

$$V_S(y_{w,S}, P_3(s)) = V_S(d_{w,S}, P_3(s)) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} v_{*S_h}(y), \tag{3.54}$$

where

$$v_{S_h}(y) = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_{S_h})^2.$$

In sampling with replacement, the unbiased estimator of the variance $V(y_{w,S}, P_4(S))$ is:

$$V_S(y_{w,S}, P_4(s)) = V_S(d_{w,S}, P_4(s)) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} v_{S_h}(y). \tag{3.55}$$

Usually, the sample sizes are evaluated proportionally to the size of the strata according to the expression:

$$n_h = n w_h = n \frac{N_h}{N} \qquad h = 1, ..., H. \tag{3.56}$$

In this case and for sampling without replacement, the variance and its unbiased estimator take the following forms:

$$\begin{cases} V(y_{w,S}, P_3(s)) = V(d_{w,S}, P_3(s)) = \frac{N(N-n)}{n} \sum_{h=1}^{H} w_h v_{*U_h}(y), \\ V_S(y_{w,S}, P_3(s)) = \frac{N(N-n)}{n} \sum_{h=1}^{H} w_h v_{*S_h}(y). \end{cases} \tag{3.57}$$

In the case of sampling with replacement, we have:

$$\begin{cases} V(y_{w,S}, P_4(s)) = V(d_{w,S}, P_4(s)) = \frac{N}{n} \sum_{h=1}^{H} w_h v_{*U_h}(y), \\ V_S(y_{w,S}, P_4(s)) = V_S(d_{w,S}, P_4(s)) = \frac{N}{n} \sum_{h=1}^{H} w_h v_{*S_h}(y). \end{cases} \tag{3.58}$$

In the case of sampling with replacement the well-known Neyman (1934) allocation of sample sizes in the strata is as follows (see also Tschuprow (1923)):

$$n_h = n\frac{w_h\sqrt{v_{*U_h}(y)}}{\sum_{i=1}^{H} w_i\sqrt{v_{*U_h}(y)}}, \quad h = 1,...,H, \tag{3.59}$$

where n is the postulated sample size of the stratified sample. In this case the variance takes the following minimal value:

$$V(y_{w,S}, P_4(s)) = \frac{N^2}{n}\left(\sum_{h=1}^{H} w_h\sqrt{v_{*U_h}(y)}\right)^2. \tag{3.60}$$

Moreover, the following methods of sample allocation are popular. Allocation proportionate to stratum aggregates is defined as follows:

$$n_h = n\frac{w_h\bar{y}_{U_h}}{\sum_{i=1}^{H} w_i\bar{y}_{U_i}} = n\frac{y_{U_h}}{y_U}, \quad h = 1,...,H. \tag{3.61}$$

When the variable under study is binary, expressions (3.59) and (3.61) reduce to the following, respectively:

$$n_h = n\frac{w_h p_h(1-p_h)}{\sum_{i=1}^{H} w_i p_i(1-p_i)}, \quad h = 1,...,H,$$

$$n_h = n\frac{w_h p_h}{\sum_{i=1}^{H} w_i p_i} = n\frac{m_{U_h}}{m_U}, \quad h = 1,...,H.$$

Moreover, when p_h is close to zero then the last expression approximates the previous one.

Equal allocation is defined as follows:

$$n_h = \frac{n}{H}, \quad h = 1,...,H. \tag{3.62}$$

Let us note that the literature regarding survey sampling provides other methods of optimally evaluating sample sizes (see e.g. Cochran (1977), Wywiał (1992, 2003, 2013a)).

3.5.2 Stratification

The above analysis leads to the obvious conclusion that the efficiency of inference also depends on the stratification of the population. Reasonable methods of stratification were considered by e.g. Bühler and Deutler (1975), Cochran (1977), Dalenius (1950, 1957, 1959), Kish (1965), Kozak (2004, 2004a, 2011), Lavallée and Hidiroglou (1988), Lednicki and Wieczorkowski (2003), Niemiro (1999) and Wywiał (2003). A larger review of statistical literature about stratification is shown by Gamrot (2014) or Khan et al. (2008). In the case of auditing problems, book values treated as auxiliary variable observations are usually highly correlated with

audited values treated as observations of the variable under study. Therefore, the optimal stratification of the population based on auxiliary variables should be approximately equivalent to the optimal stratification based on the variable under study.

11111111111111111111111111111111111111111

Let us assume that the data (x_k, y_k), $k = 1,...,N$ are values of a two-dimensional continuous random variable with the marginal density functions $f_1(x)$ and $f_2(y)$. Moreover, let the correlation coefficient $\rho(x,y)$ be close to the value 1 and the interval $I = [a;b]$ be support for the densities $f_1(x)$ and $f_2(y)$. The length of the interval is denoted by $\Delta = b - a$. The support is divided into the intervals $I_h = [y_{(h-1)}; y_{(h)})$, $h = 1,...,H+1$ so that $I = \bigcup_{h=1}^{H+1} I_h$ and $y_{(0)} = a$, $y_{(H+1)} = b$. The length of I_h is denoted by Δ_h, so $\Delta = \sum_{h=1}^{H+1} \Delta_h$. Let $F(y) = \int_{-\infty}^{y} f(t)dt$ and

$$\mu_h = \frac{1}{F(y_{h+1}) - F(y_h)} \int_{y_h}^{y_{h+1}} y f(y) dy,$$

$$\sigma_h^2 = \frac{1}{F(y_{h+1}) - F(y_h)} \int_{y_h}^{y_{h+1}} (y - \mu_h)^2 f(y) dy.$$

Dalenius (1950) shows that the solution $y_{(1)},...,y_{(h)},...,y_{(H)}$ of the following equality system

$$\frac{\sigma_{y,h}^2 + (y_h - \mu_{y,h})^2}{\sigma_{y,h}} = \frac{\sigma_{y,h+1}^2 + (y_{h+1} - \mu_{y,h+1})^2}{\sigma_{y,h+1}}, \quad h = 1,...,H-1 \tag{3.63}$$

leads to minimisation of the stratified sample mean variance under the Neyman allocation (see expressions (3.59) and (3.60)).

Let the variable under study and the auxiliary variable be highly correlated and

$$Y = a_1 X + a_0 + \varepsilon. \tag{3.64}$$

Delenius (1957) shows that in this case, equation system (3.63) can be rewritten as follows:

$$\frac{a_1^2(\sigma_{x,h}^2 + (x_h - \mu_{x,h})^2) + 2\sigma_{\varepsilon,h}^2}{a_1 \sigma_{x,h} \sqrt{1 + \sigma_{\varepsilon,h}^2 / a_1^2 \sigma_{x,h}^2}} = \frac{a_1^2(\sigma_{x,h+1}^2 + (x_{h+1} - \mu_{x,h+1})^2) + 2\sigma_{\varepsilon,h+1}^2}{a_1 \sigma_{x,h+1} \sqrt{1 + \sigma_{\varepsilon,h+1}^2 / a_1^2 \sigma_{x,h+1}^2}},$$

for $h = 1,...,H-1$. When $\sigma_{\varepsilon,h}^2 / a_1^2 \sigma_{x,h}^2$ is close to zero then the above system reduces to the following (see Cochran (1977), p. 131):

$$\frac{\sigma_{x,h}^2 + (x_h - \mu_{x,h})^2}{\sigma_{x,h}} = \frac{\sigma_{x,h+1}^2 + (x_{h+1} - \mu_{x,h+1})^2}{\sigma_{x,h+1}}, \quad h = 1,...,H-1. \tag{3.65}$$

Hence, the solution of the above equation system dependent on the parameters of the auxiliary variable (book values) can approximate the solution of equation (3.63).

In the case of proportional allocation given by (3.56), the solution of the equation system:

$$y_h = \frac{\bar{y}_h + \bar{y}_{(h+1)}}{2}, \quad h = 1,...,H-1$$

leads to the minimization of (3.57) (see Hess, Sethi and Balakrishnan (1966)).

When the sample sizes are equal (see equation (3.62)), then the stratification points $(y_{(1)},...,y_{(h)},...,y_{(H)})$ are the solution of the following equation system:

$$w_h(\sigma^2_{y,h} + (y_h - \mu_{y,h})^2) = w_{h+1}(\sigma^2_{y,h+1} + (y_{h+1} - \mu_{y,h+1})^2), \quad h = 1,...,H-1$$

The solution of the above system minimizes the variance of the stratified sample size when the sample sizes drawn from the strata are equal to $\frac{n}{H}$ (see Cochran (1961) and Sethi (1963)).

When allocation is proportional to stratum aggregates (see expression (3.61)), the stratification points satisfy the following equation system:

$$\frac{\mu_{y,h}(y_h - \mu_{y,h})^2 - \sigma^2_{y,h}(y_h - \mu_{y,h})}{\mu^2_{y,h}} = \frac{\mu_{y,h+1}(y_{h+1} - \mu_{y,h+1})^2 - \sigma^2_{y,h+1}(y_{h+1} - \mu_{y,h+1})}{\mu^2_{y,h+1}}$$

where $h = 1,...,H-1$.

The solutions of the above equation systems are only approximated. Dalenius and Hodges (1959) consider the solution of the equation system given by the expression (3.63) when the density $f(y)$ is exponential or normal. Moreover, they proposed the rule of determining the strata boundaries based on $\sqrt{f(y)}$. The iteration solutions of that system are considered e.g. by Cochran (1961), Hess et al. (1966), Mahalanobis (1952), Serfling (1968) and Sethi (1963).

3.5.3 Approximation of test statistic distribution

The hypotheses defined in subsection 3.1 can be tested on the basis of the following statistic:

$$z_{w,S} = \frac{d_{w,S} - d_U}{\sqrt{V_S(d_{w,S})}} \tag{3.66}$$

where $d_{w,S} = x_U - y_{w,S}$. Based on the appropriate central limit theorems and under some additional general conditions (see Fuller (2009)), we can infer that if $h = 1,...,H$ $n_h \to \infty$, $N_h \to \infty$, $N_h - n_h \to \infty$, then $z_{w,S} \to Z \sim N(0,1)$.

$$z_{w,S} = \frac{d_{w,S} - d_U}{\sqrt{V_S(d_{w,S})}} \to Z \sim N(0,1).$$

In the case of sampling with replacement, the above property results from the central limit theorem and the well known theorem of Sludzki (see e.g. Cramér (1946)).

Now let us adapt the method of evaluating sample size that was considered in the subsection 3.3.5, which is necessary to approximate the distribution of the statistic

by normal distribution. The marginal density $f_1(x)$ can be written as the following mixture of truncated densities:

$$f_1(x) = \sum_{h=1}^{H} w_h f_{1,h}(x)$$

where

$$f_{1,h}(x) = \begin{cases} \frac{f_1(x)}{w_h} & for \quad x \in (x_h; x_{h+1}], \\ 0 & for \quad x \notin (x_h; x_{h+1}] \end{cases}$$

$$w_h = F(x_{h+1}) - F(x_h), \quad F(x) = \int_{-\infty}^{x} f_1(t)dt \quad h = 0, ..., H-1,$$

$$x_0 = -\infty, \quad x_{H+1} = \infty.$$

From a practical point of view we can assume that $x_0 = 0$ and that $x_{H+1} = m$ is a fixed value.

When the number of strata is large, the densities $f_{1,h}(x)$ can be approximated by the uniform distribution (see e.g. Dalenius and Hodges (1959)):

$$g_{1,h}(x) = \begin{cases} \frac{1}{\Delta_h} & for \quad x \in (x_h; x_{h+1}], \\ 0 & for \quad x \notin (x_h; x_{h+1}] \end{cases}$$

where

$$\Delta_h = x_{h+1} - x_h, \quad h = 0, ..., H+1.$$

Let us assume that the simple sample S_{*h} of size n_h is drawn with replacement from the uniform distribution on the interval $(x_h; x_{h+1}]$, $h = 1, ..., H$. The moments of the uniform distribution are as follows:

$$\mu_h = \frac{x_h + x_{h+1}}{2}, \quad \sigma_h^2(x) = \frac{\Delta_h^2}{12}, \quad \sigma_{3,h}(x) = 0, \quad \sigma_{4,h}(x) = \frac{\Delta_h^4}{80}, \quad \tau_{3,h}(x) = \frac{\Delta_h^3}{32}$$

where $\sigma_{r,h}(x)$ is the central moment of order r (particularly $\sigma_{2,h}(x) = \sigma_h^2(x)$) and $\tau_{3,h}(x)$ is the absolute central moment of order r.

Let $S = (S_1, ..., S_h, ..., S_H)$ where $S_h = [Y_{h,1}, ..., Y_{h,n_h}]$ is the sample from the uniform distribution determined by the density $g_{1,h}(x)$ treated as a model of the h-th stratum, $h = 1, ..., H$. In this case the stratified sample mean is written as follows:

$$\bar{Y}_{w,S} = \sum_{h=1}^{H} w_h \bar{Y}_{S_h}, \qquad \bar{Y}_{S_h} = \frac{1}{n_h} \sum_{k \in S_h} Y_k$$

The central moments of the mean from the stratified sample $\bar{Y}_{w,S}$ (see Appendix 6.1.2) are:

$$\begin{cases} \sigma^2(\bar{Y}_{w,S}) = \frac{1}{12} \sum_{h=1}^{H} w_h^2 \frac{\Delta_h^2}{n_h}, \\ \sigma_3(\bar{Y}_{w,S}) = 0, \\ \sigma_4(\bar{Y}_{w,S}) - 3\sigma_h^2(\bar{Y}_{w,S}) = -\frac{1}{120} \sum_{h=1}^{H} w_h^4 \frac{\Delta_h^2}{n_h} \end{cases} \tag{3.67}$$

Particularly, if $n_h = nw_h$ for all $h = 1, ..., H$, then

$$\sigma^2(\bar{Y}_{w,S}) = \frac{1}{12n} \sum_{h=1}^{H} w_h \Delta_h^2.$$

If $n_h = n\frac{w_h\Delta_h}{\sum_{i=1}^{H} w_h\Delta_i}$ for all $h = 1, ..., H$, then

$$\sigma^2(\bar{Y}_{w,S}) = \frac{1}{12n} \left(\sum_{h=1}^{H} w_h \Delta_h \right)^2.$$

In Appendix 6.1.2 we derived the following:

$$\tau_3(\bar{Y}_{w,S}) \leq \frac{3\sqrt{3}}{4} \frac{\sum_{h=1}^{H} \frac{w_h^3\Delta_h^3}{n_h^2}}{\left(\sum_{h=1}^{H} \frac{w_h^2\Delta_h^2}{n_h} \right)^{3/2}}. \tag{3.68}$$

Particularly, if $n_h = m$ for all $h = 1, ..., H$,

$$\tau_3(\bar{Y}_{w,S}) \leq \frac{3\sqrt{3}}{4\sqrt{m}} \frac{\sum_{h=1}^{H} w_h^3\Delta_h^3}{\left(\sum_{h=1}^{H} w_h^2\Delta_h^2 \right)^{3/2}}.$$

When $\Delta_h = \Delta$ for all $h = 1, ..., H$,

$$\tau_3(\bar{Y}_{w,S}) \leq \frac{\sqrt{12^3}}{32} \frac{\sum_{h=1}^{H} \frac{w_h^3}{n_h^2}}{\left(\sum_{h=1}^{H} \frac{w_h^2}{n_h} \right)^{3/2}}.$$

If $n_h = nw_h$ for all $h = 1, ..., H$, then

$$\tau_3(\bar{Y}_{w,S}) \leq \frac{3\sqrt{3}}{4\sqrt{n}} \frac{\sum_{h=1}^{H} w_h\Delta_h^3}{\left(\sum_{h=1}^{H} w_h\Delta_h^2 \right)^{3/2}}.$$

If $n_h = n\frac{w_h\Delta_h}{\sum_{i=1}^{H} w_h\Delta_i}$ for all $h = 1, ..., H$,then

$$\tau_3(\bar{Y}_{w,S}) \leq \frac{3\sqrt{3}}{4\sqrt{n}}. \tag{3.69}$$

The above inequality is also true when $n_h = nw_h$ and $\Delta_h = \Delta$ or if $w_h = 1/H$ and $n_h = n\frac{\Delta_h}{\sum_{i=1}^{H} \Delta_i} = \frac{n}{b-a}\Delta_h$ for all $h = 1, ..., H$, where $[a;b]$ is the support of the density $f_2(y)$.

Based on the results we can approximate the distribution of the stratified sample mean in a similar way to sections 3.3.4 and 3.3.5.

Chapter 4
Substantive tests based on complex sampling designs

4.1 Monetary sampling designs

An important class of sampling designs and sampling schemes are dependent on auxiliary variables. They can lead to significant improvement in the quality of statistical inference on the population parameters of design variables under study. In auditing accounting amounts observed in all elements of the population are treated as values of an auxiliary variable.

The monetary sampling design is characterised by first-order inclusion probabilities proportional to appropriate book values (see Stringer (1963), Leslie et al. (1979), Särndal at all (1992), Fienberg et al. (1997)). In practice the inclusion probabilities are usually approximately proportional to the book values (see e.g. *Statistical models...* (1989)). It seems that the more important problem than the proportionality property is constructing efficient methods of testing hypotheses. Here, sampling designs with inclusion probabilities defined as non-decreasing functions of an auxiliary variable are treated as approximations of the monetary sampling design. Reviews of such sampling designs can be found e.g. in the monographs by Brewer and Hanif (1983), Chaudhuri and Stenger (2005), Chaudhuri and Vos (1988), Tillé (2006) and the *International Encyclopedia of Statistical Science* (2011).

In practice we can meet situation when audit costs are, in some sense, proportional to complexity of accounting units. Hence, sampling schemes with inclusion probabilities dependent on costs of the units observations can be considered. Such kind of sampling schemes are considered under constrained total cost of the unit observation (see e.g. Pathak (1976) and Gamrot (2014)).

4.1.1 Lahiri - Midzuno- Sen's sampling design proportional to sample mean

In the previous sections, accounting amounts were denoted by $x_1,...,x_N$. Let us assume that all these values are positive. The sample and population means of the auxiliary variable are denoted by $\bar{x}_s = \frac{1}{n}\sum_{k\in s} x_k$ and $\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$, respectively. The size n of a sample is the effective sample size. The sampling design of a sample s proportional to the sample mean of the auxiliary variable is as follows (see Lahiri (1951), Midzuno (1952) and Sen (1953)):

$$P_5(s) = \frac{1}{\binom{N}{n}} \frac{\bar{x}_S}{\bar{x}} \tag{4.1}$$

The first and second order probabilities are as follows (see e.g. Rao (1977) or Wywiał (1992)):

$$\pi_k = \frac{N-n}{(N-1)N}\frac{x_k-\bar{x}}{\bar{x}} + \frac{n}{N}, \qquad \frac{n-1}{N-1} < \pi_k < \frac{n}{N},$$

$$\pi_{k,t} = \frac{n(n-1)}{N(N-1)} + \frac{n-1}{N-2}\left(\pi_k + \pi_t - \frac{2n}{N}\right), \qquad \frac{n(n-1)}{N(N-1)} < \pi_{k,t} < \frac{n^2}{N^2}$$

where $k \neq t = 1,\ldots,N$.

The sampling scheme implementing the sampling design is as follows. Let $p_k = \frac{x_k}{x_U}$, $k = 1,...,N$. The first element is drawn from the population into the sample with probability p_k, $k = 1,...,N$. The next $n-1$ elements of the sample are drawn without replacement from the remaining $N-1$ of the population as the simple sample of size $n-1$.

The ordinary ratio estimator considered in subsection 3.4 is the unbiased estimator of the population mean. Wywiał (1992, 1995, 2003) considered more properties of the sampling design defined above. Finally, let us note that in the considered case the first-order inclusion probabilities are not exactly proportional to the appropriate values of the auxiliary variables.

4.1.2 Inclusion probabilities proportional to book values

The first-order inclusion probabilities proportional to the values of accounting amounts are defined in the following way:

$$\pi_k = \frac{nx_k}{\sum_{i\in U} x_i} = n\frac{x_k}{x_U}, \qquad k = 1,...,N. \tag{4.2}$$

We have to underline that sometimes it is possible that $\pi_h > 1$ for some $h \in U$. In this situation, the following procedure is proposed (see e.g. Tillé (2006) or Ardilly

and Tillé (2006)). According to the above expression we evaluate the quantities $\pi_k^{(0)}$, $k = 1,...,N$. Next we determine the maximal value:

$$\pi_h^{(0)} = \text{maximum}_{k\in U}\{\pi_k^{(0)}\},$$

If $\pi_h^{(0)} \leq 1$, then $\pi_k = \pi_k^{(0)}$, for $k = 1,...,N$. However, when $\pi_h^{(0)} > 1$, we assume that $\pi_h = 1$, and we evaluate the following:

$$\pi_k^{(1)} = \frac{(n-1)x_k}{\sum_{i=1}^N x_i - x_h}, \qquad k = 1,...,N; \quad k \neq h.$$

Next, we determine the index g in such a way that

$$\pi_g^{(1)} = \text{maximum}_{k\in\{U-h\}}\{\pi_k^{(1)}\}$$

If $\pi_g^{(1)} \leq 1$, then we state $\pi_h = 1$, $\pi_k = \pi_k^{(1)}$ for $k = 1,...,N$, and $k \neq h$. This completes the algorithm. In the opposite case when $\pi_g^{(1)} > 1$, we state that $\pi_h = 1$, $\pi_g = 1$, and we follow the above procedure once again.

Generalizing the outlined algorithm let us suppose that in the e-stage ($e = 0,1,...,r < n$) of the procedure, $\pi_{k_u} = 1$ ($u = 1,...,e$) where

$$\pi_k^{(e)} = \frac{(n-e)x_k}{\sum_{i=1}^N x_i - \sum_{h=1}^e x_{k_h}}, \qquad k \in U - \{k_1,...,k_e\} \tag{4.3}$$

and $\{k_0\} = \varnothing$. Next, we determine the index k_{e+1} in such a way that

$$\pi_{k_{e+1}}^{(e)} = \text{maximum}_{k\in\{U-\{k_1,...,k_e\}\}}\{\pi_k^{(e)}\}.$$

If $\pi_{k_{e+1}}^{(e)} \leq 1$ we state that $\pi_{k_i} = 1$ for $i = 1,...,e$ and $\pi_k = \pi_{k_i}^{(e)}$ for $k \in \{U - \{k_1,...,k_e\}\}$. The algorithm is completed.

In the opposite case, when $\pi_{k_{e+1}}^{(e)} > 1$ it is stated that $\pi_{k_{e+1}} = 1$ and

$$\pi_k^{(e+1)} = \frac{(n-e-1)x_k}{\sum_{i=1}^N x_i - \sum_{h=1}^{e+1} x_{k_h}}, \qquad k \in \{U - \{k_1,...,k_{e+1}\}\}. \tag{4.4}$$

The algorithm is then continued.

The R - code of the program that implements the above procedure is presented in Appendix 6.2.6.

4.1.3 Inclusion probabilities determined by the distribution of the order statistic of an auxiliary variable

Let us consider a simple random sampling with replacement design with unequal probabilities of drawing population elements. The probability of selecting the k-th population element in each draw be denoted by $p_k = f(x_k)$. Usually, those probabilities are determined as follows:

$$f(x_k) = \frac{x_k}{\sum_{i \in U} x_i}, \quad k \in U.$$

Let X_i, $i = 1, ..., n$ be a random variable whose values are the observations of an auxiliary variable in the i-th draw. So, its probability function is: $P(X_i = x_k) = f(x_k) = p_k$ for $i = 1, ..., n$ and $k = 1, ..., N$. Hence, the random variables X_1, X_2,....,X_i,...,X_n are independent and have the same probability function as defined by $P(X = x_k) = f(x_k)$, $k = 1, ..., N$. Therefore, the sequence $((X_i),\ i = 1, ..., n)$ can be treated as data observed in the sample drawn with replacement from the population U with unequal probabilities $f(x_k)$, $k = 1, ..., N$.

Let the sample $((X_i),\ i = 1, ..., n)$, ordered by the values of X_i, be denoted by $((X_{r:n}), r = 1, ...n)$, where $X_{r:n}$ is the r-th order statistic (so, $X_{r:n} \leq X_{r+1:n}$, $r = 1, ..., n-1$). The distribution function of the r-th order statistic is as follows (see Arnold, Balakrishnan and Nagaraja (2008), p. 42):

$$P(X_{r:n} = x_k) = \sum_{i=0}^{r-1} \sum_{j=0}^{n-r} \frac{n!(F(x_{k-1}))^{r-1-i}(1-F(x_k))^{n-r-j}(f(x_k))^{i+j+1}}{(r-1-i)!(n-r-j)!(i+j+1)!}$$

where $F(x_k)$ is the distribution function:

$$F(x_k) = P(X_i < x_k) = \sum_{\{x_i:\, x_i < x_k;\, i \in U\}} p_i.$$

Wywiał (2012a) proposed to select a k-th $k = 1, ..., N$ population element with replacement to the sample s of size n in a single draw with the probability $p_k(r:n) = P(X_{r:n} = x_k)$. Hence, the sampling design is as follows:

$$P_{6,r}(s) = \prod_{k \in s} p_k(r:n).$$

The described sampling design prefers selection (with replacement) of the population elements close to the mode of the distribution of order statistic $X_{r:n}$. Hence, for instance, the sampling design $P_{6,1}(s)$ ($P_{6,n}(s)$) prefers drawing small (large) values of the auxiliary variable. The efficiency of estimation based on these sampling designs is considered by Wywiał (2012a).

4.1.4 Systematic sampling with varying inclusion probabilities

Subsection 3.2.2 presented simple systematic sampling, which is characterized by constant probabilities of sample selection as well as constant inclusion probabilities. Systematic sampling designs with non-constant inclusion probabilities were considered by Hartley and Rao (1962) and Rao et. al. (1962). Let the sample be the following sequence of successively drawn population elements $s = (i_1, i_2, ..., i_n)$. Firstly, we evaluate the following cumulative sums:

$$V_e = \sum_{k=1}^{e} \pi_k,$$

where $V_0 = 0$ and $V_N = n$. Next, the random value u is generated from the uniform distribution on the interval $(0;\ 1)$. The population element identified by the index i_1 is selected to the sample s when

$$V_{i_1-1} < u \leq V_{i_1}.$$

The next element drawn into the sample is labelled i_2, which fulfils the inequality:

$$V_{i_2-1} < u+1 \leq V_{i_2}.$$

More generally, $i_z \in s$, if and only if:

$$V_{i_z-1} < u+z-1 \leq V_{i_z}, \quad z = 1, ..., n. \tag{4.5}$$

A sampling scheme similar to the above is as follows. We determine the following cumulative sums:

$$x_{c,e} = \sum_{j=1}^{e} x_j, \qquad e = 0, 1, ..., N,$$

where $x_{c,e} = 0$, $x_{c,N} = x_U$. Next, we calculate the ratio:

$$I_n = z\left(\frac{x_U}{n}\right),$$

where $z(a)$ is the value a rounded to the nearest integer. The parameter I_n is sometimes called the interval of the systematic sampling scheme. Now the number k is randomly drawn from the sequence $(1, 2, ..., I_n)$. The population element indexed by i_1 is selected to the sample s if

$$x_{c,i_1-1} < k \leq x_{c,i_1}$$

The next population element, indexed by i_2, is selected to s when

$$x_{c,i_2-1} < k + I_n \leq x_{c,i_2}$$

Generally the algorithm selects the element i_z to s if and only if

$$x_{c,i_z-1} < k + (z-1)I_n \leq x_{c,i_z}, \qquad z = 1,...,n. \tag{4.6}$$

If $z\left(\frac{x_U}{n}\right)$ is an integer and $\pi_k = \frac{nx_k}{x_U} \leq 1$ for $k = 1,...,N$, then the appropriate inequalities given by expressions (4.5) and (4.6) are equivalent. Each inequality in expression (4.5) is obtained from the appropriate inequality, given by (4.6) through multiplying the former by $\frac{1}{I_n}$. Hence, in this case the dollar sampling scheme and the sampling scheme of Hartley and Rao et al. are equivalent.

More properties of the various versions of the systematic sampling design are considered e.g. by Tillé (2006). The computer program that implements the Hartley-Rao sampling scheme and inclusion probabilities is presented in Appendix 6.2.7.

4.1.5 Rejective sampling

Let π_k, $k \in U$ be the postulated inclusion probabilities, e.g. defined as proportional to book value, see section 4.1.2. A sample of size n is drawn with replacement according to a preassigned sampling scheme. If, in the selected sample, at least one population element appears more than once, then the sample is rejected and a new sample is drawn. This operation is replicated until no population element is repeated in the sample. This type sampling scheme implements the following sampling design:

$$P_7(s) = c\prod_{k\in s} q_k, \qquad s \in S,$$

where S is the sample space of all samples of the same size which can be selected from a population of size N. The constant c is determined in such a way that $\sum_{s\in S} P(s) = 1$. Hájek (1981) evaluated the following approximation for the probabilities q_k (see also Berger (1998)):

$$q_k = \lambda \frac{\pi_k}{1-\pi_k}\left(1 + \frac{\hat{\pi} - \pi_k}{d_\pi}\right), \quad k = 1,...,N,$$

where

$$d_\pi = \sum_{k=1}^{N} \pi_k(1-\pi_k), \qquad \hat{\pi} = \frac{1}{d_\pi}\sum_{k=1}^{N} \pi_k^2(1-\pi_k),$$

and λ is evaluated in such a way that

$$\sum_{k=1}^{N} q_k = 1.$$

The more precise explanation of the sampling scheme is as follows. The sample s of size n is drawn with replacement from a population of size N. The kth population element is selected from a population with probability q_k at each draw. If, in the

obtained sample, at least one element, is repeated then the sample is rejected and a new sample is selected. This operation is replicated until no elements are repeated in the selected sample.

Let $H(P) = \sum_{s \in \mathbf{S}} P(s) log(P(s))$ be the entropy function of the sampling design $P(s)$. It is assumed that $0log(0) = 0$. Hájek (1959) proves that the rejective sampling design $P_7(s)$ maximizes entropy in the class of all fixed sample size sampling designs drawn without replacement with first-order inclusion probabilities defined by $(\pi_k,\ k \in U)$.

The divergence (distance) of the sampling design $P(s)$ from the rejective sampling design $P_7(s)$ is defined as follows (see Berger (1998)):

$$D(P) = \sum_{s \in \mathbf{S}} P(s) log\left(\frac{P(s)}{P_7(s)}\right) \tag{4.7}$$

$D(P) = 0$ if and only if $P(s) = P_7(s)$ for all $s \in \mathbf{S}$. For $k = 1, ..., N$ let π_k be the first-order inclusion probabilities of the sampling design $P(s)$ and $\pi_k^{(r)}$ be the fist-order inclusion probabilities of the rejective sampling design $P_7(s)$. Berger (1998) proves that

$$|\pi_k - \pi_k^{(r)}| \leq \sqrt{2D(P)} \quad \text{for all} \quad k = 1, ...N.$$

4.1.6 Rao-Sampford sampling scheme

Rao (1965) and Sampford (1967) considered the following sampling design:

$$P_{RS}(s) = c_1 \left(\sum_{k \in U-s} \pi_k\right) \prod_{k \in s} \frac{\pi_k}{1-\pi_k}, \qquad s \in S,$$

where c_1 has to be evaluated in such a way that $\sum_{s \in S} P_{RS}(s) = 1$. The first-order inclusion probabilities of the sampling design are exactly equal to preassigned values π_k, $k = 1, ..., N$.

The sampling scheme implementing the above sampling design is as follows. The first population element is drawn into the sample with probability π_k/n, $k = 1, ..., N$. The remaining $n-1$ elements are drawn with replacement according to the probabilities proportional to $\frac{\pi_k}{1-\pi_k}$, $k = 1, ..., N$. If, in the selected sample, there is at least one repeated element then the sample is drawn again. This algorithm is replicated until there is no repetition in the selected sample.

The program that implements the above sampling scheme is presented e.g. in the textbook by Wywiał (2014a).

4.1.7 Truncated sampling scheme

Let population elements be ordered according to the increasing values of an auxiliary variable. So, $i < j$ if and only if $x_i < x_j$, $i, j = 1, ..., N$ and $i \neq j$. Let us consider the following sampling scheme. Let L_i be a random variable whose value is equal to the number (label) of a selected population element into the sample s of size n during the ith draw, $i = 1, ..., n$. The first population element is drawn from the population with the following probability:

$$p_{l_1} = P(L_1 = l_1) = \frac{1}{N-n+1} \quad \text{for} \quad l_1 = 1, ..., N-n+1.$$

The next elements are selected with probabilities:

$$p_{l_2/l_1} = P(L_2 = l_2 | L_1 = l_1) = \frac{1}{N-n+2-l_1} \quad \text{for} \quad l_1 < l_2 \leq N-n+2,$$

$$p_{l_3/l_2} = P(L_3 = l_3 | L_2 = l_2) = \frac{1}{N-n+3-l_2} \quad \text{for} \quad l_2 < l_3 \leq N-n+3,$$

Generalizing the introduced selection we have:

$$p_{l_i/l_{i-1}} = P(L_i = l_i | L_{i-1} = l_{i-1}) = \frac{1}{N-n+i-l_{i-1}} \tag{4.8}$$

for $l_0 = 0$, $l_i = l_{i-1} + 1, ..., N-n+i$, $i = 1, ..., n$ and $p_{l_1/l_0} = p_{l_1}$.

Hence, the defined sampling scheme leads to selection of the ordered sample: $s = (l_1, ..., l_i, ..., l_n)$ where $l_{i-1} < l_i$.

The sampling design is as follows:

$$P_{l.trunc}(s) = \prod_{i=1}^{n} p_{l_i/l_{i-1}} = \frac{1}{\prod_{i=1}^{n}(N-n+i-l_{i-1})}. \tag{4.9}$$

Let us note that the conditional probabilities of the sampling design are evaluated through gradual left truncation of the sequence $(1, 2, ..., l_i, ..., N)$. Hence, we can name the sampling scheme and design as gradually left truncated. Moreover, let us note that, similarly to above, we can define the right side (or both side) truncated sampling scheme and design.

In particular, when $n = 2$, we have

$$P_t(s) = P(L_1 = l_1, L_2 = l_2) = \frac{1}{(N-1)(N-l_1)}$$

for $1 \leq l_1 \leq N-1$ and $l_1 < l_2 \leq N$. In this case:

$$p_{l_1} = P(L_1 = l_1) = \frac{1}{N-1} \quad \text{for} \quad l_1 = 1, ..., N-1.$$

$$p_{l_2/l_1} = P(L_2 = l_2 | L_1 = l_1) = \frac{1}{N - l_1} \quad \text{for} \quad l_1 < l_2 \leq N,$$

The inclusion probabilities, as derived in Appendix 6.1.3, are as follows:

$$\pi_k = \begin{cases} \frac{1}{N-1} & \text{for} \quad k = 1, \\ \frac{1}{N-1} + \frac{1}{N-1} \sum_{h=1}^{k-1} \frac{1}{N-h} & \text{for} \quad k = 2, ..., N-1, \\ \frac{1}{N-1} \sum_{h=1}^{N-1} \frac{1}{N-h} & \text{for} \quad k = N. \end{cases} \tag{4.10}$$

We can show that the inclusion probabilities π_k derived above are values of the increasing function of the indexes $k = 1, ..., N-1$.

The sampling scheme and sampling design defined above can be generalized into the following way by involving the values of the positive auxiliary variable (book amounts). Let $x_{i-1} \leq x_i$ for $i = 2, ..., N$. Based on expression (4.8), the conditional probabilities of the new sampling design can be defined by the equations:

$$p_{l_i/l_{i-1}}(x) = P(L_i = l_i | L_{i-1} = l_{i-1}) = \frac{x_{l_i}}{\sum_{j=l_{i-1}+1}^{N-n+i} x_j} \tag{4.11}$$

for $l_i = l_{i-1} + 1, ..., N - n + i$, $i = 1, ..., n$

$$p_{l_1}(x) = p_{l_1/l_0}(x) = P(L_1 = l_1) = \frac{x_{l_1}}{\sum_{j=1}^{N-n+1} x_j}. \tag{4.12}$$

This leads to the following sampling design:

$$P_{l.trunc,x}(s) = \prod_{i=1}^{n} p_{l_i/l_{i-1}}(x). \tag{4.13}$$

In the case of sample size $n = 2$ the derivation presented in Appendix 6.1.3 and expressions (4.11) (4.12) let us evaluate the inclusion probabilities based on the following equations:

$$\pi_k = \begin{cases} \sum_{h=1}^{k-1} P(L_2 = k | L_1 = h) P(L_1 = h) + \\ + P(L_1 = k) \sum_{t=k+1}^{N} P(L_2 = t | L_1 = k) & \text{for} \quad k = 1, ..., N-1, \\ \sum_{h=1}^{N-1} P(L_2 = N | L_1 = h) P(L_1 = h) & \text{for} \quad k = N. \end{cases} \tag{4.14}$$

Let us note that in section 4.3 we consider the continuous version of a gradually truncated sampling scheme.

4.1.8 Sampling design proportional to the function of one order statistic

Let $(X_{r:n}, r = 1,...,n)$ be the sequence of the order statistics of observations of the auxiliary variable in the simple random sample s drawn without replacement. Let $G(r,i) = \{s : X_{r:n} = x_i\}$ be the set of all samples whose r -th order statistic of the auxiliary variable is equal to x_i where $r \leq i \leq N-n+r$. Moreover, $\bigcup_{i=r}^{N-n+r} G(r,i) = S$. The size of the set $G(r,i)$ is denoted by $g(r,i) = Card(G(r,i))$ and

$$g(r,i) = \binom{i-1}{r-1}\binom{N-i}{n-r} = \\ = \frac{\Gamma(i)\Gamma(N-i+1)}{\Gamma(r)\Gamma(i-r+1)\Gamma(N-i-n+r+1)\Gamma(n-r+1)} = \\ = \frac{1}{(r-1)(n-r)B(r-1,i-r+1)B(N-i-n+r+1,n-r)},$$

$$\sum_{i=r}^{N-n+r} g(r,i) = \binom{N}{n} = \frac{\Gamma(n+1)\Gamma(N-n+1)}{\Gamma(N+1)} = \frac{1}{nB(n,N-n+1)}.$$

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$, $\Gamma(a+1) = a\Gamma(a)$ and $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

The probability distribution of the order statistic $X_{r:n}$ is as follows, see Wilks (1962):

$$P(X_{r:n} = x_i) = P(s \in G(r,i)) = \frac{g(r,i)}{\binom{N}{n}}.$$

The relationship between the order statistic and the sample quantile of order $\alpha \in (0;1)$ is explained by the following equation (see e.g. Fisz (1967)):

$$Q_\alpha = X_{r:n}$$

where $r = [n\alpha]+1$ is the integer part of the value $n\alpha$, $r = 1,2,...,n$ and $X_{r:n} = Q_\alpha$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$.

The considered sample can be denoted by $s = \{s_1, i, s_2\}$ where $s_1 = \{i_1,...,i_{r-1}\}$, $s_2 = \{i_{r+1},...,i_n\}$, $i_j < i$ for $j = 1,...,r$, $i_r = i$ and $i_j > i$ for $j = r+1,...,n$.

Let $f(x_i)$ be a positive function of value x_i of the order statistic $X_{r:n}$ of the auxiliary variable. In particular, $f(x_i) = x_i$.

Wywiał (2008) proposed the following sampling design proportional to $f(x_i)$:

$$P_{q1}(s) = \frac{f(x_i)}{\sum_{j=r}^{N-n+r} g(r,j)f(x_j)} \qquad \text{for} \quad i \in s \in G(r,i) \tag{4.15}$$

Hence, the probability of selecting a sample s is proportional to the value $f(x_i)$ where x_i is the observation of the order statistic $X_{r:n}$.

Let

$$P_{q1}(s: f(x_u) \le f(X_{r:n}) \le f(x_v)) = \sum_{i=u}^{v} P_{q1}\left(s: f(X_{r:n}) = f(x_i)\right) = \frac{\sum_{i=u}^{v} f(x_i)g(r,i)}{\sum_{j=r}^{N-n+r} f(x_j)g(r,j)}$$

The sampling design defined above is a particular case of the following conditional sampling design:

$$P_{q1}(s|u,v) = \frac{P_{q1}(s)}{P_{q1}(f(x_u) \le f(X_{r:n}) \le f(x_v))} = \frac{f(x_i)}{z_r(u,v)} \tag{4.16}$$

for $i \in s \in G(r,i)$ and $r \le u \le i \le v \le N-n+r$ where:

$$z_r(u,v) = \sum_{j=u}^{v} f(x_j)g(r,j)$$

The sampling design $P_{q1}(s|u,v)$ depends only on those values of the auxiliary variable which are not less than x_u and not greater than x_v. More about the concept of the conditional sampling design can be found in Tillé (1998, 2006).

In particular, if $f(X_{r:n}) = X_{r:n}$, $r = u < v = N-n+r$, $P_{q1}(s|r,N-n+r) = P_{q1}(s)$. Moreover, when $x_i = c$ for all $i = 1,...,N$, the sampling design $P_{q1}(s)$ can be simplified to the simple random sampling without replacement $P_0(s)$. Other particular cases of the conditional sampling design are considered by Wywiał (2008, 2014a, 2015).

Let us assume that if $x \le 0$, then $\delta(x) = 0$ and when $x > 0$, then $\delta(x) = 1$. Hnce, $\delta(x)\delta(x-1) = \delta(x-1)$. The first- and second-order inclusion probabilities are defined the following two theorems proved by Wywiał (2008, 2015).

Theorem 4.1. *The first-order inclusion probabilities for the conditional sampling design $P(s|u,v)$ are as follows: if $k < u$,*

$$\pi_k(u,v) = \frac{\delta(r-1)\delta(v-1)\delta(u-1)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} f(x_i);$$

if $u \le k \le v$,

$$\pi_k(u,v) = \frac{1}{z_r(u,v)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}f(x_i) + \right.$$
$$\left. + \binom{k-1}{r-1}\binom{N-k}{n-r}x_k + \delta(r-1)\delta(v-k)\sum_{i=k+1}^{v}\binom{i-2}{r-2}\binom{N-i}{n-r}f(x_i)\right);$$

if $k > v$,

$$\pi_k(u,v) = \frac{\delta(n-r)\delta(N-v)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} f(x_i),$$

Theorem 4.2. *The second-order probabilities for the conditional sampling design $P(s|u,v)$ are as follows: if $k < u$, $t < u$ and $t \ne k$,*

$$\pi_{k,t}(u,v) = \frac{\delta(r-2)\delta(v-2)\delta(u-2)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-3}{r-3} \binom{N-i}{n-r} f(x_i);$$

if $k > v$, $t > v$ *and* $t \neq k$,

$$\pi_{k,t}(u,v) = \\ = \frac{\delta(n-r-1)\delta(N-v-1)\delta(N-u-1)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-1}{r-1} \binom{N-i-2}{n-r-2} f(x_i);$$

if $k < u$ *and* $t > v$ *or* $t < u$ *and* $k > v$,

$$\pi_{k,t}(u,v) = \frac{\delta(r-1)\delta(n-r)\delta(u-1)\delta(N-v)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-2}{r-2} \binom{N-i-1}{n-r-1} f(x_i);$$

if $k < u$ *and* $u \leq t \leq v$ *or* $t < u$ *and* $u \leq k \leq v$,

$$\pi_{k,t}(u,v) = \\ = \frac{\delta(r-1)}{z_r(u,v)} \left(\delta(n-r)\delta(t-u)\delta(t-2) \sum_{i=u}^{t-1} \binom{i-2}{r-2} \binom{N-i-1}{n-r-1} f(x_i) + \right. \\ + \delta(t-1) \binom{t-2}{r-2} \binom{N-t}{n-r} x_t + \\ \left. + \delta(r-2)\delta(v-t)\delta(v-2)\delta(t-1) \sum_{i=t+1}^{v} \binom{i-3}{r-3} \binom{N-i}{n-r} f(x_i) \right);$$

if $u \leq k \leq v$ *and* $t > v$ *or* $u \leq t \leq v$ *and* $k > v$,

$$\pi_{k,t}(u,v) = \frac{\delta(n-r)}{z_r(u,v)} \left(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1) \cdot \right. \\ \cdot \sum_{i=u}^{k-1} \binom{i-1}{r-1} \binom{N-i-2}{n-r-2} f(x_i) + \delta(N-k) \binom{k-1}{r-1} \binom{N-k-1}{n-r-1} x_k + \\ \left. + \delta(r-1)\delta(v-k)\delta(N-v)\delta(v-1)\delta(N-k-1) \sum_{i=k+1}^{v} \binom{i-2}{r-2} \binom{N-i-1}{n-r-1} f(x_i) \right);$$

if $u \leq k < t \leq v$ *or* $u \leq t < k \leq v$,

$$\pi_{k,t}(u,v) = \frac{\delta(v-u)}{z_r(u,v)}\Bigg(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1)\cdot$$
$$\cdot\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2}f(x_i)+\delta(n-r)\delta(N-k)\binom{k-1}{r-1}\binom{N-k-1}{n-r-1}x_k+$$
$$+\delta(r-1)\delta(n-r)\delta(t-k-1)\delta(t-2)\delta(N-k-1)\sum_{i=k+1}^{t-1}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}f(x_i)+$$
$$+\delta(r-1)\delta(t-1)\binom{t-2}{r-2}\binom{N-t}{n-r}x_t+$$
$$+\delta(r-2)\delta(v-t)\delta(v-2)\delta(t-1)\sum_{i=t+1}^{v}\binom{i-3}{r-3}\binom{N-i}{n-r}f(x_i)\Bigg).$$

The above expressions are also true for non-distinct values of the auxiliary variable.

The first- and second-order inclusion probabilities can be evaluated using the computer program presented in Appendix 6.2.8.

The conditional sampling design $P_{q1}(s|u,v)$ is implemented by means of the following sampling scheme. Firstly, population elements should be ordered according to non decreasing values of the auxiliary variable. Next, the i-th population element where $i = u, u+1, ..., v$, is drawn with probability

$$P_{*,q1}(i|u,v) = \frac{f(x_i)g(r,i)}{\sum_{j=u}^{v} f(x_j)g(r,j)}.$$

Finally, two simple samples s_1 and s_2 are drawn without replacement from the subpopulations $U_1 = \{1, ..., i-1\}$ and $U_2 = \{i+1, i+2, ..., N\}$, respectively. The sample s_1 is of size $r-1$ and the sample s_2 is of size $n-r$. The sampling designs of these samples are independent and

$$P_{1,q1}(s_1) = \frac{1}{\binom{i-1}{r-1}}, \quad P_{2,q1}(s_2) = \frac{1}{\binom{N-i}{n-r}}.$$

The selected sample $s = \{s_1, i, s_2\}$ fulfils the equation:

$$P_{*,q1}(i|u,v)P_{1,q1}(s_1)P_{2,q1}(s_2) = P_{q1}(s|u,v)$$

where $r = u, u+1, ..., v$.

The computer program that implement the above sampling scheme is in Appendix 6.2.9. More properties of the introduced sampling scheme in the context of estimating accuracy of the population mean are considered by Wywiał (2007, 2015). He shows that sampling designs proportional to one of the last order statistics of the auxiliary variable usually leads to first-order inclusion probabilities that are approximately proportional to the appropriate values of the auxiliary variable.

4.1.9 Sampling design proportional to the function of two order statistics

Let $X_{r:n}$ and $X_{u:n}$ be two order statistics of a positive auxiliary variable observed in a simple random sample drawn without replacement. Let $G(r,u,i,j) = \{s : X_{r:n} = x_i, X_{u:n} = x_j\}$, $r = 1,..,n-1$; $u = 2,...,n$, $r < u$ be the set of all samples whose r -th and u -th order statistics of the auxiliary variable are equal to x_i and x_j, respectively, where $r \leq i < j \leq N-n+u$. The sets $G(r,u,i,j)$ fulfil the following equation:

$$\bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j) = S.$$

$g(r,u,i,j)$ is the size of the set $G(r,u,i,j)$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}.$$

Let $x_i \leq x_j$ for $i < j$ and $i, j = 1,...,N$ as was previously assumed. Moreover, let $f(x_i,x_j,C)$ be the following positive function of the values x_i and x_j of the order statistics $X_{n:r_1}$ and $X_{n:r_2}$, respectively,

$$f(x_i,x_j,C) = \begin{cases} f(x_i,x_j) & for \quad f(x_i,x_j) \in C, \\ 0 & for \quad f(x_i,x_j) \notin C \end{cases} \tag{4.17}$$

where $C \subseteq R_+$ and $r_1 \leq i \leq N-n+r_1$ and $r_1 < r_2 \leq j \leq N-n+r_2$.

The conditional sampling design proportional to the function $f(x_i,x_j,C)$ is as follows:

$$P_{r_1,r_2}(s|C) = \frac{f(x_i,x_j,C)}{z(r_1,r_2,C)} \qquad (i,j) \in s \in G(r_1,r_2,i,j) \tag{4.18}$$

where

$$z(r_1,r_2,C) = \sum_{i=r_1}^{N-n+r_1} \sum_{j=i+r_2-r_1}^{N-n+r_2} f(x_i,x_j,C)g(r_1,r_2,i,j).$$

If $C = R_+$, the conditional sampling design $P_{r_1,r_2}(s|C)$ reduces to the unconditional sampling design $P_{r_1,r_2}(s)$.

Wywiał (2009, 2015) derived the following first-order inclusion probabilities of the sampling design $P_{r_1,r_2}(s|C)$:

Theorem 4.3.

$$\pi_k(r_1,r_2,C) =$$

$$= \frac{1}{z(r_1,r_2,C)}\Bigg(\delta(r_1-1)\sum_{i=r_1\delta(r_1-k)+(k-1)\delta(k+1-r_1)\delta(N-n+r_1-k)}^{N-n+r_1}\sum_{j=i+r_2-r_1}^{N-n+r_2}\binom{i-2}{r_1-2}$$

$$\binom{j-i-1}{r_2-r_1-1}\binom{N-j}{n-r_2}f(x_j,x_i,C)+\delta(k-r_1)\delta(N-n+r_2-k)\delta(r_2-r_1-1)$$

$$\sum_{i=r_1}^{min(k-1,N-n+r_1)}\sum_{j=max(i+r_2-r_1,k+1)}^{N-n+r_2}\binom{i-1}{r_1-1}\binom{j-i-2}{r_2-r_1-2}\binom{N-j}{n-r_2}f(x_j,x_i,C)+$$

$$+\delta(k-r_2)\delta(n-r_2)\delta(N-n+r_2-k+1)$$

$$\sum_{i=r_1}^{k-r_2+r_1-1}\sum_{j=i+r_2-r_1}^{k-1}\binom{i-1}{r_1-1}\binom{j-i-1}{r_2-r_1-1}\binom{N-j-1}{n-r_2-1}f(x_j,x_i,C)+\delta(n-r_2)$$

$$\delta(k-N+n-r_2)\sum_{i=r_1}^{N-n+r_1}\sum_{j=i+r_2-r_1}^{N-n+r_2}\binom{i-1}{r_1-1}\binom{j-i-1}{r_2-r_1-1}\binom{N-j-1}{n-r_2-1}f(x_j,x_i,C)+$$

$$+\delta(k+1-r_1)\delta(N-n+r_1-k+1)\binom{k-1}{r_1-1}$$

$$\sum_{j=k+r_2-r_1}^{N-n+r_2}\binom{j-k-1}{r_2-r_1-1}\binom{N-j}{n-r_2}f(x_j,x_k,C)+\delta(k-r_2+1)\delta(N-n+r_2-k+1)$$

$$\binom{N-k}{n-r_2}\sum_{i=r_1}^{k-r_2+r_1}\binom{i-1}{r_1-1}\binom{k-i-1}{r_2-r_1-1}f(x_k,x_i,C)\Bigg)$$

The second-order inclusion probabilities are derived by Wywiał (2009, 2015).

Let us assume that the population elements are ordered according to increasing values of the auxiliary variable. The sample s is divided into three samples, s_1, s_2 and s_3 in such a way that

$$s = s_1 \cup \{i\} \cup s_2 \cup \{j\} \cup s_3, \qquad s_1 = \{k : k \in U, x_k < x_i\},$$

$$s_2 = \{k : k \in U, x_j > x_k > x_i\} \qquad \text{and} \qquad s_3 = \{k : k \in U, x_k > x_j\}.$$

The population U is divided in the following way:

$$U = U(1,i-1) \cup \{i\} \cup U(i+1,j-1) \cup \{j\} \cup U(j+1,N)$$

where

$$U(1,i-1) = (1,...,i-1), \qquad U(i+1,j-1) = (i+1,...,j-1),$$

$$U(j+1,N) = (j+1,...,N).$$

The sample space of sample s_1 of size r_1-1 will be denoted by $\mathbf{S}_1 = S(U(1,i-1);s_1)$. Moreover, let $\mathbf{S}_2 = S(U(i+1,j-1);s_2)$ be the sample space of sample s_2 of size r_2-r_1-1 and ${}_3 = S(U(j+1,N);s_3)$ be the sample space of sample s_3 of size $n-r_2$.

The sampling scheme of the sampling design $P_{r_1,r_2}(s|C)$ proposed by Wywiał (2009, 2015) is explained by the following probabilities:

$$P_{r_1,r_2}(s|C) = P_{1a}(s_1)p_{r_1,r_2}(i|C)P_{1b}(s_2)p'_{r_1,r_2}(j|C)P_{1c}(s_3)$$

where

$$P_{1a}(s_1) = \binom{i-1}{r_1-1}^{-1}, \quad P_{1b}(s_2) = \binom{j-i-1}{r_2-r_1-1}^{-1}, \quad P_{1c}(s_3) = \binom{N-j}{n-r_2}^{-1}$$

for $s_1 \in \mathbf{S}_1, \qquad s_2 \in \mathbf{S}_2, \qquad s_3 \in \mathbf{S}_3,$

$$p_{r_1,r_2}(i|j,C) = P\left(X_{(r_1)} = x_i | X_{(u)} = x_j, C\right) = \frac{P_{r_1,u}\left(X_{(r_1)} = x_i, X_{(u)} = x_j, C\right)}{P_{r_1,u}\left(X_{(u)} = x_j, C\right)},$$

$$P_{r_1,r_2}\left(X_{(r_1)} = x_i, X_{(r_2)} = x_j, C\right) = \sum_{s \in G(r_1,r_2,i,j)} P_{r_1,r_2}(s|C) = \frac{f(x_j,x_i,C)g(r_1,r_2,i,j)}{z(r_1,r_2,C)},$$

$$p'_{r_1,r_2}(j|C) = P_{r_1,r_2}\left(X_{(r_2)} = x_j, C\right) = \frac{1}{z(r_1,r_2,C)} \sum_{i=r_1}^{N-n+r_1} f(x_j,x_i,C)g(r_1,r_2,i,j).$$

In order to draw the sample s, firstly, the j-th element of the population is selected according to the probability function $p'_{r_1,r_2}(j|C)$. In the next step, the i-th element of the population is drawn according to the probability function $p_{r_1,r_2}(i|j,C)$. Finally, the samples s_1, s_2 and s_3 are drawn according to the sampling designs $P_{1a}(s_1)$, $P_{1b}(s_2)$ and $P_{1c}(s_3)$, respectively.

Wywiał (2015) showed that the sampling design proportional to the sum of two order statistics is approximately proportionate to the values of the auxiliary variable. This approximation is good especially when the order statistics are of order 2 and $n-1$. More properties of the sampling design dependent on two order statistics is considered by Wywiał(2011, 2013b, 2013c).

The computer program for the computation of the first-order inclusion probabilites is presented in Appendix 6.2.10. The computer program shown in Appendix 6.2.11 implements the sampling scheme.

Wywiał (2015) defined the conditional sampling design dependent on three order statistics and derived the first-order inclusion probabilities. The inclusion probabilities in this case are very complicated. That is why, in Appendix 6.2.12 we have only included the computer program for evaluating the first-order inclusion probabilities. Appendix 6.2.13 contains the program that implements the sampling scheme. Moreover, let us note that the general case of the sampling design dependent on more than three order statistics is considered by Wywiał (2009a).

4.2 Horvitz-Thompson statistic

Our purpose is still to test hypotheses on the total population error formulated in section 3.1. The inference is based on complex monetary samples. Firstly, the construction of the well-known Horvitz-Thompson statistic including some of its asymptotic properties is shown. Next, a review of the sampling designs or schemes proportional to book values is presented. Finally, continuous sampling designs and schemes are considered.

4.2.1 Basic properties

We assume that the considered accounting population is finite and fixed. The design-based approach is used. This means that observed variables are treated as fixed (not random). Our purpose is to make inference on the total population error amount denoted by $d_U = x_U - y_U$ (see the section 1.1). Now, the considered test statistic will be based on estimators of d_U. The total population book amount denoted by x_U is known in advance. Hence, estimation of the parameter d_U reduces to estimation of the total audited amount y_U.

The well-known Horvitz-Thompson (1952) statistic is as follows (see also Rao (2004)):

$$y_{HTS} = \sum_{k=1}^{N} \frac{y_k I_k}{\pi_k} = \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{4.19}$$

where $I_k = 1$ when a kth population element is in a sample S and $I_k = 0$ if a kth population element is not drawn to the sample. Moreover, $E(I_k) = \pi_k$, $E(I_k, I_l) = \pi_{k,l}$, $V(I_k, I_l) = \pi_{k,l} - \pi_k \pi_l$ and $V(I_k) = V(I_k, I_k) = \pi_k(1-\pi_k)$, $k, l = 1, ..., N$, $k \neq l$.

When $\pi_k > 0$ for all $k = 1, ..., N$, then y_{HTS} is an unbiased estimator of y_U, and its variance is:

$$V(y_{HTS}) = \sum_{k=1}^{N} \left(\frac{y_k}{\pi_k} \right)^2 \pi_k(1-\pi_k) + \sum_{k=1}^{N} \sum_{i=1, i \neq k}^{N} \frac{y_k y_i}{\pi_k \pi_i} (\pi_{k,i} - \pi_k \pi_i).$$

or

$$V(y_{HTS}) = \sum_{k=1}^{N} \frac{y_k^2}{\pi_k} + \sum_{k=1}^{N} \sum_{i=1, i \neq k}^{N} \frac{y_k y_i}{\pi_k \pi_i} \pi_{k,i} - y^2.$$

When the effective sample size is fixed, Yates and Grundy (1953) derived:

$$V(y_{HTS}) = \sum_{k=1}^{N} \sum_{i=1, i \neq k}^{N} \left(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i} \right)^2 (\pi_{k,i} - \pi_k \pi_i).$$

The variance $V(y_{HTS})$ is estimated by means of the following statistics provided $\pi_{k,i} > 0$ for all $i \neq k = 1, ..., N$:

$$V_S(y_{HTS}) = \sum_{k=1}^{N} \left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k) I_k + \sum_{k=1}^{N} \sum_{i=1, i\neq k}^{N} \frac{y_k y_i}{\pi_k \pi_i} \frac{\pi_{k,i} - \pi_k \pi_i}{\pi_{k,i}} I_k I_i, \tag{4.20}$$

or

$$V_S(y_{HTS}) = \sum_{k=1}^{N} \sum_{i=1, i\neq k}^{N} \left(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i}\right)^2 \frac{\pi_{k,i} - \pi_k \pi_i}{\pi_{k,i}} I_k I_i. \tag{4.21}$$

Both statistics above are unbiased estimators of the variance $V(y_{HTS})$

The estimator of the mean value $\bar{y}$ is as follows:

$$\bar{y}_{HTS} = \frac{1}{N} y_{HTS} = \frac{1}{N} \sum_{k=1}^{N} \frac{y_k I_k}{\pi_k}.$$

The estimator of the population size number is:

$$N_{HTS} = \sum_{k=1}^{N} \frac{I_k}{\pi_k}. \tag{4.22}$$

Its variance is:

$$V(N_{HTS}) = \left(\sum_{k=1}^{N} \frac{1}{\pi_k} + \sum_{k=1}^{N} \sum_{i=1, i\neq k}^{N} \frac{\pi_{k,i}}{\pi_k \pi_i} \right) - N^2.$$

The unbiased estimator of $V(N_{HTS})$ is:

$$V_S(N_{HTS}) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k^2} I_k + \sum_{k=1}^{N} \sum_{i=1, i\neq k}^{N} \left(\frac{1}{\pi_k \pi_i} - \frac{1}{\pi_{k,i}} \right) I_k I_i. \tag{4.23}$$

On the basis of the Horvitz-Thompson statistic ratio and regression type estimators are constructed. The ratio estimator is as follows:

$$y_{rHTS} = y_{HTS} \frac{x_U}{x_{HTS}},$$

This is an approximately unbiased estimator of the population total and its variance is approximately as follows:

$$V(y_{rHTS}) \approx V(y_{HTS}) + h^2 V(x_{HTS}) - 2hV(x_{HTS}, y_{HTS})$$

where

$$h = \frac{y_U}{x_U}, \qquad V(x_{HTS}) = V(x_{HTS}, x_{HTS}),$$

$$V(x_{HTS}, y_{HTS}) = \frac{1}{N^2} \sum_{k=1}^{N} \frac{x_k y_k (1-\pi_k)}{\pi_k} + \frac{2}{N^2} \sum_{i=1}^{N} \sum_{k>i}^{N} \frac{x_k y_i}{\pi_k \pi_i} (\pi_{ki} - \pi_i \pi_k).$$

The estimator of the variance $V(y_{rHTs})$ is the following statistic:

$$V_S(y_{rHTS}) = V_S(y_{HTS}) + h_S^2 V_S(x_{HTS}) - 2h_S V_S(x_{HTS}, y_{HTS})$$

where $h_S = \frac{y_{HTS}}{x_{HTS}}$,

$V_S(x_{HTS}, y_{HTS}) = \frac{1}{N^2}\sum_{k=1}^{N} \frac{x_k y_k (1-\pi_k)}{\pi_k^2} I_k + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k\neq i}^{N} \frac{x_k y_i}{\pi_k \pi_i} \frac{\pi_{ki}-\pi_i\pi_k}{\pi_{ki}} I_k I_i.$

The Horvitz-Thompson-regression type estimator of population total is defined as follows:

$$y_{regHTS} = y_{HTS} + a_{HTS}(x_U - x_{HTS})$$

where

$$a_{HTS} = \frac{v_{HTS}(x,y)}{v_{HTS}(x)}$$

$$v_{HTS}(x,y) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - x_{HTS})(y_k - y_{HTS})\frac{I_k}{\pi_k}, \qquad v_{HTS}(x) = v_{HTS}(x,x).$$

y_{regHTS} is the approximately unbiased estimator of the population mean and its variance is as follows:

$$V(y_{regHTS}) \approx V(y_{HTS}) + a^2 V(x_{HTS}) - 2aV(x_{HTS}, y_{HTS})$$

where $a = \frac{c_*(x,y)}{v_*(x)}$.

The unbiased estimator of the variance $V(y_{regHTS})$ is as follows:

$$V_S(y_{regHTS}) = V_S(y_{HTS}) + a_{HTS}^2 V_S(\bar{x}_{HTs}) - 2a_{HTS}^2 V_S(x_{HTS} y_{HTS}).$$

Let us note that more properties of the Horvitz-Thompson estimator are considered e.g. by Berger (1998), Patel and Patel (2010), Tillé (2006) and Wywiał (1992, 2003). Barbiero and Mecatti (2010) consider the bootstrap-type estimator of the variance of the regression statistic.

4.2.2 Estimators of variance

Estimation of approximated variance

In sampling without replacement the variance of the Horvitz-Thompson statistic depends on first- and second-order inclusion probabilities. The second-order inclusion probabilities are usually very complicated functions and it is difficult to derive them. Moreover, the estimators of the variance can take non-positive values with probabilities greater than zero. In this situation, some approximations of the variance that are dependent only on first-order inclusion probabilities are proposed. A review of those approximations can be found in the monograph by Tillé (2006). In the case of an exponential sampling design (see, Deville and Tillé (2005), Tillé (2006)) the variance of Thorvitz-Thompson statistic can be approximated by means of the fol-

lowing expression:

$$V_{approx}(y_{HTS}) = \sum_{k \in U} b_k (\hat{y}_k - \hat{y}_*)^2 \tag{4.24}$$

where $\hat{y}_k = \frac{y_k}{\pi_k}$ and

$$\hat{y}_* = \frac{\sum_{t \in U} \hat{y}_k b_k}{\sum_{t \in U} b_k}.$$

Hájek (1981) proposed to evaluate the coefficients b_k, $k \in U$ by means of the following expression:

$$b_k = \frac{\pi_k (1 - \pi_k) N}{N - 1}, \qquad k \in U.$$

Others methods of determining of these coefficients are shown by Tillé (2006).

The estimator of the variance $V_{approx}(y_{HTS})$ is as follows:

$$V_{approx,S}(y_{HTS}) = \sum_{k \in S} c_k (\hat{y}_k - \hat{y}_{*S})^2 \tag{4.25}$$

where

$$\hat{y}_{*S} = \frac{\sum_{t \in S} \hat{y}_k c_k}{\sum_{t \in S} c_k}.$$

The coefficients c_k, $k \in U$ are determined in several ways (see the short review in the book by Tillé (2006)). The Deville (1993) proposition is as follow:

$$c_k = \frac{(1 - \pi_k) n}{n - 1}, \qquad k \in U.$$

Jackknife estimation

Rao et. al. (1992) considered the jackknife-type estimator of the variance for a two-stage sampling from stratified population. A sampling design with unequal first-order inclusion inclusion probabilities is a particular case of that sampling design (see also Berger and Skinner (2005)). Under sampling design, the jackknife-type estimator of the variance $V(y_{HTS})$ is as follows:

$$V_{J1,S}(y_{HTS}) = \frac{n - 1}{n} \sum_{i \in S} (y_{HTS_{-i}} - y_{HTS})^2 \tag{4.26}$$

where

$$y_{HTS_{-i}} = \frac{n}{n - 1} \sum_{k \in S_{-i}} \frac{y_k}{\pi_k}, \quad i = 1, ..., n,$$

S_{-i} is S without the ith unit.

Berger and Skinner (2005) proposed the following jackknife-type estimator of the variance of the Horvitz-Thompson's statistic (see also Campbell (1980)):

$$V_{J2,S}(y_{HTS}) = N^2 \sum_{i\in S}\sum_{j\in S} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \varepsilon_{(i)}\varepsilon_{(j)} \tag{4.27}$$

where

$$\varepsilon_{(i)} = (1-w_i)(\bar{y}_{HTS} - \bar{y}_{HTS_{-i}}), \qquad w_i = \frac{1}{N_{HTS}\pi_i}, \quad i \in S,$$

$$\bar{y}_{HTS} = \frac{1}{N_{HTS}} \sum_{k\in S} \frac{y_k}{\pi_k}, \qquad N_{HTS} = \sum_{k\in S} \frac{1}{\pi_k},$$

$$\bar{y}_{HTS_{-i}} = \frac{1}{N_{HTS_{-i}}} \sum_{k\in S_{-i}} \frac{y_k}{\pi_k}, \qquad N_{HTS_{-i}} = \sum_{k\in S_{-i}} \frac{1}{\pi_k}.$$

Berger and Skinner (2005) considered in detail the consistency of the estimator $V_{jackk,S}(y_{HTS})$.

Bootstrap estimation

A review of bootstrap-type estimators of the variance of the Horvitz-Thompson statistic is presented e.g. by Barbieri and Mecatti (2010). The naive bootstrap estimator of the variance $V(y_{HTS})$ is as follows. Let s be the observation of a sample of size n drawn from a fixed and finite population according to a sampling design $P(s)$ with the first order inclusion probabilities π_k, $k \in U$. On the basis of s, the statistic y_{HTS} is calculated. On the resampling step, the independently drawn simple random samples, denoted by $S_{\#b}$, $b = 1,...,B$, each of size n, are drawn with replacement from the original sample s. On the replication step the statistics

$$y_{HTS_{\#b}} = \sum_{i\in S_{\#b}} \frac{y_i}{\pi_i}, \quad b = 1,...,B, \tag{4.28}$$

are evaluated. Next, the bootstrap variance is determined according to the following expression:

$$V_{S_\#} = \frac{1}{B-1} \sum_{b=1}^{B} (y_{HTS_{\#b}} - y_{HTS_\#})^2, \quad y_{HTS_\#} = \frac{1}{B} \sum_{b=1}^{B} y_{HTS_{\#b}}. \tag{4.29}$$

Let $E_\#(t_{S_\#})$ be the expected value of the statistic $t_{S_\#}$ determined on the basis of the resampling design. Barbieri and Mecatti (2010) show that $E_\#(y_{HTS_{\#b}}) = y_{HTS}$ and

$$E_\#(V_{S_\#}) = \sum_{i\in S} \left(\frac{y_i}{\pi_i} - \frac{y_{HTS}}{n} \right)^2 = v_{naive}.$$

Holmberg's (1998) bootstrap algorithm is recommended. For this algorithm, let s be the sample of fixed size n drawn from the population U of size N according to the sampling design $P(s)$. Let $1/\pi_k = c_k + r_k$ where $c_k = [1/\pi_k]$, $0 \le r_k < 1$, $k \in s$. Let $A_k = c_k + \varepsilon_k$ where $\varepsilon_k = 1$ with probability r_k and $\varepsilon_k = 0$ with probability $1 - r_k$.

The elements of the sequence $(\varepsilon_1,..,\varepsilon_n)$ are independent. Hence, the elements of $(A_1,...,A_n)$ are also independent. The values of the random sequence $(A_1,...,A_n)$ are denoted by $(a_1,...,a_n)$. Let $N_\# = \sum_{k\in s} a_k$ be the size of a new artificial population $U_\#$. In the set $U_\#$ the pair (y_k,x_k) is replicated a_k times. On the basis of the auxiliary variable observations x_k where $k \in U_\#$, the sampling design $P_\#(s)$ is constructed according to the same rule as the original sampling design $P(s)$. Therefore, the inclusion probabilities $\pi_{\#k}$, $k \in U_\#$ are evaluated on the basis of $P_\#(s)$. Next, the sample $s_\#$ of size n is drawn from the artificial population $U_\#$ according to the sampling design $P_\#(s)$. The described procedure is repeated independently B-times. This leads to the independent samples: $(s_{\#1},...,s_{\#b},...,s_{\#B})$. Next, the Horvitz-Thompson statistic as well as the bootstrap estimator of its variance are evaluated according to expressions (4.28) and (4.29), respectively.

Barbieri and Mecatti (2010) proposed a calibrate version of Holmberg's method. They constructed an algorithm leading to the minimization of $|x_U - x_{U_\#}|$, where $x_U = \sum_{k\in U} x_k$ and $x_{U_\#} = \sum_{k\in U_\#} x_k$. Their other modification leads to a simplification of Holmberg's method. More details about bootstrap methods of variance estimation in complex sampling are considered e.g. by Antal and Tillé (2011), Chauvet (2007) and Kuk (1989).

4.2.3 Normal approximation of test statistic distribution

Our purpose is to test the hypotheses formulated in section 3.1. Usually the test statistic is defined as follows:

$$\hat{z}_{1S} = \frac{d_{HTS} - d_U}{\sqrt{V_S(d_{HTS})}}$$

where $d_{HTS} = x_U - y_{HTS}$. In the case when the randomization approach is taken into account, the total book value x_U is scalar. Hence, the above statistic can be rewritten as the following studentized Horvitz-Thompson statistic:

$$\hat{z}_{1S} = \frac{d_{HTS} - d_U}{\sqrt{V_S(y_{HTS})}} = \frac{y_U - y_{HTS}}{\sqrt{V_S(y_{HTS})}}. \tag{4.30}$$

Under quite complicated additional assumptions about the homogeneity of distribution of the statistic y_{HTS}, there are central theorems (see e.g. Fuller (2009), Hájek (1964, 1981)) which let us infer that if $n \to \infty$, $N \to \infty$, $N - n \to \infty$ then

$$\hat{z}_{1S} \to Z \sim N(0,1).$$

In these cases, convergence to the normality of the studentized Horvitz-Thompson statistic should be considered under a specific sampling design.

Usually, convergence to the normality of the statistic $\hat{z}_{1S}$ is considered with Sludzky's well-known theorem.

Let

$$z_{1S} = \frac{d_{HTS} - d_U}{\sqrt{V(y_{HTS})}} = \frac{y_U - y_{HTS}}{\sqrt{V(y_{HTS})}}.$$

Berger and Skinner (2005) infer on the basis of Sludzky's theorem that if, under specific assumptions (considered below), the statistic z_{1S} converges to standard normal distribution and the statistic $V_S(y_{HTS})$ is a consistent estimator of the variance $V(y_{HTS})$, then the statistic $\hat{z}_{1S}$ converges to normality, too. Hence, in order to prove the normality of the statistic $\hat{z}_{1S}$, we have to assign conditions under which the statistic $V_S(y_{HTS})$, is consistent estimator of $V(y_{HTS})$ and z_{1S} has asymptotically normal distribution.

Normality of the Horvitz-Thompson statistic under rejective sampling

The standardized Horvitz-Thompson statistic is denoted as follows:

$$Z_{HTrS} = \frac{y_U - y_{HTrS}}{V(y_{HTrS})} \tag{4.31}$$

where

$$y_{HTrS} = \sum_{k \in S} \frac{y_k}{\pi_k(r)},$$

$$V(y_{HTrS}) = \sum_{k \in U} z_k^2 \frac{1 - \pi_k(r)}{\pi_k(r)},$$

$$z_k = y_k - G(r,y)\pi_k(r), \quad k \in U,$$

$$G(r,y) = \frac{1}{d_{r,\pi}} \sum_{k \in U} y_k(1 - \pi_k(r)), \qquad d_{r,\pi} = \sum_{k=1}^{N} \pi_k(r)(1 - \pi_k(r)).$$

Moreover, let

$$L(\varepsilon) = \sum_{\{k:\, |e_k| > \varepsilon\}} e_k^2 \pi_k(r)(1 - \pi_k(r)), \quad \varepsilon > 0,$$

$$e_k = \frac{z_k}{\pi_k(r)\sqrt{V(y_{HTrS})}};$$

$$\varepsilon_* = Inf\{\varepsilon :\, L(\varepsilon) \leq \varepsilon\}.$$

The quantity $L(\varepsilon)$ can be treated as the "standardized amount" of outlier observations of size ε of the variable under study. Hence, ε_* is the minimal possible level of this amount.

The distribution function of the random variable z_{HTrS} is given by

$$F_S(z) = P(z_{HTrS} < z) = \sum_{\{s:\, z_{HTrs} < z\}} P_7(s).$$

Let $\varphi(z)$ be distribution function of standard normal variable. Hájek (1964) proves the following central theorem:

Theorem 4.4. *If* $\varepsilon_* \to 0$ *and* $d_{r,\pi} \to \infty$ *then*

$$|F_S(z) - \varphi(z)| \to 0.$$

Finally, we can conclude based on Sludky's theorem that when the rejective sampling design fulfils the assumptions of the above theorem and we can propose an unbiased estimator of the variance $V(y_{HTs})$ then the distribution of the statistic $\hat{z}_{HTrS}$ defined by expression (4.31) can be approximated by means of standard normal distribution.

Under the rejective sampling design and the inclusion probabilities $\pi_k \propto x_k, k \in U$ Hájek (1964, 1981), proposed the following estimator of the variance:

$$V_S(y_{HTs}) = \frac{n}{n-1} \sum_{k \in S} \left(\frac{y_k}{\pi_k(r)} - \frac{x_U}{n} I_{y:x,S} \right)^2 (1 - \pi_k(r)), \tag{4.32}$$

where

$$I_{y:x,S} = \frac{\sum_{k \in S} \frac{y_k}{x_k}(1 - \pi_k(r))}{\sum_{k \in S}(1 - \pi_k(r))} \approx \frac{1}{n} \sum_{k \in S} \frac{y_k}{x_k}.$$

Hence, when $d_{r,\pi}$ is sufficiently large and ε_* is sufficiently close to zero, the distribution of the statistic $\hat{z}_{HTrS}$ can be approximated by means of standard normal distribution.

The next estimator is as follows:

$$\hat{V}_S(y_{HTS}) = \frac{1}{2} \sum_{k \in S} \sum_{j \in S} \left(\frac{y_k(r)}{\pi_k(r)} - \frac{y_j}{\pi_j(r)} \right)^2 (1 - \pi_k(r))(1 - \pi_j(r)), \tag{4.33}$$

When the assumptions of the Hájek (1964) theorem are fulfilled Wywiał (2012) proved that $\hat{V}_S(y_{HTS})$ is the consistent estimator of the variance $V(\bar{y}_{HTs})$. Hence, the statistic

$$\hat{z}_{2S} = \frac{y_U - y_{HTS}}{\sqrt{\hat{V}_S(y_{HTS})}}$$

has approximately normal distribution when $d_{r,\pi}$ is sufficiently large and ε_* is sufficiently close to zero.

Normality of the Horvitz-Thompson statistic under Rao-Sampford sampling

Berger (1998) using the results of Prášková (1982) proved the following theorem:

Theorem 4.5. *If B_1, B_2 are such positive constance that*

$$\sum_{k=1}^{N} \left(\frac{y_k}{\pi_k} \right)^4 \leq B_1 N,$$

$$V(y_{HTS}) = B_2 N,$$

then there are positive numbers N_0 and k_1 such that for $N \geq N_0$,

$$|F_S(z) - \varphi(z)| < \frac{k_1}{\sqrt{N}} + \sqrt{2D(P)}$$

where $D(P)$ is the divergence coefficient defined in subsection 4.1.5 by (4.7).

The theorem lets us assess the degree of convergence of the distribution of the standardized Horvitz-Thompson statistic z_{HTS} to normal distribution under any sampling design $P(s)$.

Finally, Berger (1998) proves the following generalization of Hájek's theorem 4.4 for any sampling design which is close to the rejective sampling design in the sense of the divergence coefficient:

Theorem 4.6. *If the design $P(s)$ is implemented, the statistic z_{HTS} has asymptotic normal distribution, i.e.*

$$|F_S(z) - \varphi(z)| \to 0,$$

if and only if $\varepsilon_* \to 0$, $d_\pi \to \infty$ and $D(P) \to 0$ where $d_\pi = \sum_{k \in U} \pi_k(1-\pi_k)$.

Finally let us note that the simulation procedures proposed in subsection 3.3.5 can be applied to evaluating necessary sample size to assure sufficient convergence to the normality of the test statistic. Of course, in this case the bootstrap version of the studentized Horvitz-Thompson statistic can be taken into account.

More problems connected with convergence to normality statistics that are dependent on complex samples in the randomization approach were also considered e.g. by Berger (1998), Fuller (2009), Prášková (1985), Rosén (1972), Sen (1995), Wywiał (2013) and Víšek (1979).

4.3 Continuous population approach

4.3.1 Basic definition

In the considered case values of book amounts as well as audited amounts can be treated as values of a two-dimensional continuous random variable. The problem of selecting not only simple random samples from a continuous population is considered e.g. in geology, ecology and geography. We will briefly refer to Cordy's (1993) results about the extensions of well-known properties of the design-based approach for inference about a continuous population. Simplifying the problem, the basic definitions are as follows.

Let the population $U \subset R$. For instance, $U = R_+ - \{0\}$. The sample space, denoted by $S_n = U^n$, is the set of the ordered sample denoted by $\mathbf{x} = (x_1, ..., x_n)$, $x_k \in U$, $k = 1, ..., n$. Let $\mathbf{x}$ be a value of the n-dimensional random $\mathbf{X} = (X_1, ..., X_n)$ with density function $f(\mathbf{x}) = f(x_1, ..., x_n)$. Let $f_i(x)$ and $f_{i,j}(x, x')$ be marginal density functions of X_i and (X_i, X_j), respectively, $j > i = 1, ..., n$. The inclusion probabilities are defined as follows:

$$\pi(x) = \sum_{i=1}^{n} f_i(x), \qquad \pi(x,x') = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} f_{i,j}(x,x'), \quad x,x' \in U. \tag{4.34}$$

Let $f(x_i|x_{i-1},x_{i-2},...,x_1)$, $i = 1,...,n-1$ be the conditional density function of the randomly selected x_i value in the i-th draw (provided that the values $(x_{i-1},x_{i-2},...,x_1)$ were drawn earlier). Hence, the density function of the sampling design can be written as follows:

$$f(x_i,x_{i-1},...,x_1) = \prod_{i=2}^{n} f(x_i|x_{i-1},x_{i-2},...,x_1)$$

The generalization of the well-known Horvitz-Thompson (1952) statistic into the continuous case is as follows (see also Cordy (1993)). Let $y(x)$ be the integrable function $y: U \to R$. The parameter of the populations is defined by:

$$g_y = \int_U y(x)dx.$$

The estimator of g_y is:

$$g_{y,\mathbf{X}} = \sum_{\mathbf{X}=\{X_1,...,X_i,...,X_n\}} \frac{y(X_i)}{\pi(X_i)} \tag{4.35}$$

Cordy (1993) proved the following two theorems:

Theorem 4.7. *The statistic $g_{y,\boldsymbol{X}}$ is the unbiased estimator for g_y, if the function $y(x)$ is either bounded or non-negative and $\pi(x) > 0$ for each $x \in U$.*

Theorem 4.8. *If the function $y(x)$ is bounded, $\pi(x) > 0$ for each $x \in U$, and $\int_U (1/\pi(x))dx < \infty$, then*

$$V(g_{y,\boldsymbol{X}}) = \int_U \frac{y^2(x)}{\pi(x)} + \int_U \int_U y(x)y(x') \frac{\pi(x,x') - \pi(x)\pi(x')}{\pi(x)\pi(x')} dxdx'.$$

When, in addition, $\pi(x_i,x_j) > 0$ for all $x_i,x_j \in U$, $i \neq j = 1,...,n$, the unbiased estimator of the above variance is:

$$V_{\boldsymbol{X}}(g_{y,\boldsymbol{X}}) = \sum_{\boldsymbol{X}=\{X_1,...,X_i,...,X_n\}} \frac{y^2(X_i)}{\pi^2(X_i)} + \\ + \sum_{\boldsymbol{X}=\{X_1,...,X_i,X_j...,X_n\}} y(X_i)y(X_j) \frac{\pi(X_i,X_j) - \pi(X_i)\pi(X_j)}{\pi(X_i,X_j)\pi(X_i)\pi(X_j)}$$

is the unbiased estimator of $V(g_{y,\boldsymbol{X}})$.

The book values can be treated as observations of a continuous and positively valued random variable. The sampling designs based on the continuous auxiliary variables considered below and the estimator above let us estimate the total error

amount. Moreover, the estimator lets us construct the statistic in order to verify appropriately formulated hypotheses about the total error amount. The problem of the distribution of the test statistics can be solved using the appropriate central theorems.

4.3.2 Gradually truncated continuous sampling design

Gradually truncated uniform sampling design

Let us assume that the auxiliary variable is uniformly distributed on the interval $[a;b]$ where $0 \leq a < b$. The first value of the auxiliary variable is drawn (generated) according to the density function of the well-known uniform distribution:

$$f(x_1) = \begin{cases} \frac{1}{b-a}, & x_1 \in (a;b), \\ 0, & x_1 \notin (a;b). \end{cases} \tag{4.36}$$

The next observation is generated from the following truncated uniform distribution:

$$f(x_2|x_1) = \begin{cases} \frac{1}{b-x_1}, & x_2 \in (x_1;b), \\ 0, & x_2 \notin (x_1;b). \end{cases}$$

In general, the observation x_i is drawn according to the following density function:

$$f(x_i|x_{i-1}) = \begin{cases} \frac{1}{b-x_{i-1}}, & x_i \in (x_{i-1};b), \\ 0, & x_i \notin (x_{i-1};b), \end{cases} \quad i = 1,...,n, \quad x_0 = a. \tag{4.37}$$

This leads to the sampling design defined by the following density function:

$$f(x_n, x_{n-1}, ..., x_1) = \begin{cases} \frac{1}{\prod_{i=0}^{n-1}(b-x_i)}, & a = x_0 < x_1 < x_2 < ... < x_n < b, \\ 0, & otherwice. \end{cases} \tag{4.38}$$

The expression (4.37) leads to the following conditional distribution function:

$$F(x_i|x_{i-1}) = P(X_i < x_i|x_{i-1}) = \begin{cases} 0, & x_i \in (-\infty;x_{i-1}], \\ \frac{x_i - x_{i-1}}{b-x_{i-1}}, & x_i \in (x_{i-1};b), \\ 1, & x_i \in [b;\infty), \end{cases} \quad i = 1,...,n. \tag{4.39}$$

Hence, the sequence $\{F(x_i|x_{i-1})\}$ defines the sampling scheme that implements the sampling design given by the density function $f(x_n, x_{n-1}, ..., x_1)$. The random variable X_i is uniformly distributed on the interval (x_{i-1}, b), so $X_i \sim U(x_{i-1}, b)$, $i = 1,...,n$. The observation $(x_1,...,x_n)$ is generated according to the following inverse distribution function:

$$x_i = F^{-1}(x_i|x_{i-1}) = x_{i-1} + (b - x_{i-1})u_i, \quad i = 1,...,n, \quad x_0 = a \tag{4.40}$$

where u_i, $i = 1,...,n$ are values of the independently distributed uniform random variables $U_i \sim U(0,1)$, $i = 1,...,n$. The introduced algorithm can be called as gradually left truncated uniform sampling scheme.

In Appendix 6.1.4 we present proof of the following theorem:

Theorem 4.9. *The marginal densities of the sampling design $f(x_1,...,x_n)$ defined by expression (4.38) are as follows:*

$$f_i(x) = \frac{1}{(b-a)\Gamma(i)} \left(ln(b-a) - ln(b-x)\right)^{i-1} = \\ = \frac{1}{(b-a)\Gamma(i)} \sum_{j=0}^{i-1} (-1)^j \binom{i-1}{j} ln^{i-j-1}(b-a) ln^j(b-x), \tag{4.41}$$

$$f_{i,j}(x,x') = \frac{(ln(b-a) - ln(b-x))^{i-1} (ln(b-a) - ln(b-x'))^{j-i-1}}{\Gamma(i)\Gamma(j-i)}, \quad x' > x, \tag{4.42}$$

for $i, j = 1,...,n, i \neq j$.

Now let us assume that the first population element is drawn according to the density function defined by expression (4.36). The second element is drawn from the following truncated uniform distribution:

$$f_2(x_2|x_1) = \begin{cases} \frac{1}{x_1 - a}, & x_2 \in (a; x_1), \\ 0, & x_2 \notin (a; x_1). \end{cases}$$

The observation x_i is generated according to the following density function:

$$f_2(x_i|x_{i-1}) = \begin{cases} \frac{1}{x_{i-1} - a}, & x_i \in (a; x_{i-1}), \\ 0, & x_i \notin (a; x_{i-1}), \end{cases} \quad i = 1,...,n, \quad x_0 = a. \tag{4.43}$$

Hence, the sampling design defined by the following density function:

$$f_2(x_n, x_{n-1}, ..., x_1) = \begin{cases} \frac{1}{\prod_{i=0}^{n-1} (x_i - a)}, & b = x_0 > x_1 > x_2 > ... > x_n > a, \\ 0, & otherwise. \end{cases} \tag{4.44}$$

The conditional distribution functions, denoted by $F_2(x_i|x_{i-1}) = P(X_i < x_i|x_{i-1})$, are as follows:

$$F_2(x_i|x_{i-1}) = \begin{cases} 0, & x_i \in (-\infty; a], \\ \frac{x_i - x_{i-1}}{x_{i-1} - a}, & x_i \in (a; x_{i-1}), \\ 1, & x_i \in [b; \infty), \end{cases} \quad i = 1,...,n. \tag{4.45}$$

Hence, the sequence of the above conditional probability distribution function is the sampling scheme implementing the sampling design defined by the density function $f_2(x_n, x_{n-1}, ..., x_1)$. Now the value x_i is the observation of the random variable X_i uniformly distributed on the interval $(a; x_{i-1})$, so $X_i \sim U(a; x_{i-1})$, $i = 1, ..., n$. The observation $(x_1, ..., x_n)$ can be generated according to the following inverse distribution function:

$$x_i = F_2^{-1}(x_i|x_{i-1}) = x_{i-1} + (x_{i-1} - a)u_i, \quad i = 1, ..., n, \quad x_0 = a \tag{4.46}$$

where u_i, $i = 1, ..., n$ are values of independently distributed uniform random variables $U_i \sim U(0,1)$, $i = 1, ..., n$. The constructed algorithm can be called as gradually right truncated uniform sampling scheme.

Theorem 4.10. *The marginal densities of the sampling design $f_2(x_1, ..., x_n)$ defined by expression (4.44) are as follows*

$$f_{2,i}(x) = \frac{(ln(b-a) - ln(x-a))^{i-1}}{(b-a)\Gamma(i)} = \\ = \frac{1}{(b-a)\Gamma(i)} \sum_{j=0}^{i-1} (-1)^j \binom{i-1}{j} ln^{i-j-1}(b-a) ln^j(b-x), \tag{4.47}$$

$$f_{2,i,j}(x, x') = \frac{(ln(b-a) - ln(x-a))^{i-1} (ln(b-a) - ln(x'-a))^{j-i-1}}{(b-a)(x-a)\Gamma(i)\Gamma(j-i)} \tag{4.48}$$

for $x' < x$, $i, j = 1, ..., n, i \neq j$.

This theorem can be proved similarly to Theorem 4.9.

The above both theorems enable inference on the basis of the Horvitz-Thompson statistic redefined in the previous subsection.

Gradually truncated exponential sampling design

Let us assume that the first population element is drawn according to the density function of the well-known simple exponential distribution:

$$f(x_1) = \begin{cases} \lambda e^{-(x_1 - c)}, & x_1 \geq c > 0, \\ 0, & x_1 < c. \end{cases} \tag{4.49}$$

The second element is drawn from the following truncated exponential distribution:

$$f(x_2|x_1) = \begin{cases} \lambda e^{-\lambda(x_2 - x_1)}, & x_2 \geq x_1, \\ 0, & x_2 < x_1. \end{cases}$$

Generalizing this process, we have:

$$f(x_i|x_{i-1}) = f(x_i|x_{i-1},...,x_1) = \begin{cases} \lambda e^{-\lambda(x_i - x_{i-1})}, & x_i \geq x_{i-1}, \\ 0, & x_i < x_{i-1}. \end{cases} \quad i = 2,...,n. \quad (4.50)$$

Hence, the above expression leads to the conclusion that the density function of the sampling design is as follows:

$$\begin{cases} f(x_n, x_{n-1},...,x_1) = \lambda^n e^{-\lambda(x_n - c)}, & c \leq x_1 \leq x_2 \leq ... \leq x_n, \\ 0 \quad \text{otherwise}. \end{cases} \quad (4.51)$$

In Appendix 6.1.5 the following theorem is proved:

Theorem 4.11.

$$\begin{cases} f_i(x) = \frac{\lambda^i}{\Gamma(i)}(x-c)^{i-1}e^{-\lambda(x-c)}, & x \geq c \\ 0, & x < c. \end{cases} \quad i = 1,...,n. \quad (4.52)$$

$$\begin{cases} f_{j,i}(x,z) = \frac{\lambda^j}{\Gamma(i)\Gamma(j-i)}(z-c)^{i-1}(x-z)^{j-i-1}e^{-\lambda(x-c)}, & x > z > c \\ 0, & otherwise. \end{cases} \quad (4.53)$$

$for \quad j > i = 1,...,n.$

Hence, the function $f_i(x)$ is the well-known density function of the well-known gamma probability distribution with a shape parameter equal to i and rate equal to λ. In this case, $E(X) = \frac{i}{\lambda}$, $V(X) = \frac{i}{\lambda^2}$, and the skewness and kurtosis coefficients are equal to $\frac{2}{\sqrt{\lambda}}$ and $\frac{6}{\lambda}$, respectively.

The sampling scheme is determined by the conditional probabilities: $F(x_i|x_{i-1})$, $i = 1,...,n$ where

$$F(x_i|x_{i-1}) = \int_{x_{i-1}}^{x_i} f(x_i|x_{i-1})dx_i = -e^{\lambda x_{i-1}} \int_{x_{i-1}}^{x_i} e^{-\lambda x}dx =$$
$$= -e^{\lambda x_{i-1}} \int_{-\lambda x_{i-1}}^{-\lambda x_i} e^z dz = 1 - e^{-\lambda(x_i - x_{i-1})}.$$

Finally we obtain:

$$F(x_i|x_{i-1}) = \begin{cases} 0 & \text{for} \quad x < c, \\ 1 - e^{-\lambda(x_i - x_{i-1})} & \text{for} \quad x \geq c. \end{cases} \quad (4.54)$$

The inverse function to the distribution $u_i = F(x_i|x_{i-1})$ is as follows:

$$F^{-1}(x_i|x_{i-1}) = x_i = x_{i-1} - \frac{ln(1-u_i)}{\lambda}, \quad i = 1,...,n. \quad (4.55)$$

where u_i, $i = 1,...,n$ are values of the independently distributed uniform random variables $U_i \sim U(0,1)$, $i = 1,...,n$. The introduced algorithm can be called as gradually left truncated exponential sampling scheme.

Gradually truncated Pareto sampling design

Let us suppose that the first population element is drawn according to the density function of the well-known Pareto distribution:

$$f(x_1) = \begin{cases} \frac{ac^a}{x_1^{a+1}}, & x_1 \geq c, \\ 0, & x_1 < c. \end{cases} \tag{4.56}$$

The second element is drawn from the following truncated Pareto distribution:

$$f(x_2|x_1) = \begin{cases} \frac{ax_1^a}{x_2^{a+1}}, & x_2 \geq x_1, \\ 0, & x_2 < x_1. \end{cases}$$

Generalizing this process, we have:

$$f(x_i|x_{i-1}) = f(x_i|x_{i-1},...,x_1) = \begin{cases} \frac{ax_{i-1}^a}{x_i^{a+1}}, & x_i \geq x_{i-1}, \\ 0, & x_i < x_{i-1}, \end{cases} \quad i = 1,...,n. \tag{4.57}$$

where $x_0 = c$, $f(x_1|c) = f(x_1)$. Therefore, the above expression leads to the conclusion that the density function of the sampling design is as follows:

$$f(x_n, x_{n-1},...,x_1) = f(x_1)\prod_{i=2}^{n} f(x_i|x_{i-1},...,x_1) = \\ = \frac{a^n c^a}{x_n^{a+1}\prod_{i=1}^{n-1} x_i}, \quad c \leq x_1 \leq x_2 \leq ... \leq x_n. \tag{4.58}$$

In Appendix 6.1.6 the following theorem is proved:

Theorem 4.12.

$$\begin{cases} f_i(x) = \frac{a^i}{\Gamma(i)c}\left(\frac{c}{x}\right)^{a+1} ln^{i-1}\left(\frac{x}{c}\right), & x \geq c \\ 0, & x < c, \end{cases} \quad i = 1,...,n, \tag{4.59}$$

$$\begin{cases} f_{j,i}(z,x) = \frac{a^j c^a}{\Gamma(i)\Gamma(j-i)z^{a+1}} ln^{(i-1)}\left(\frac{x}{c}\right) ln^{j-i-1}\left(\frac{z}{x}\right), & z \geq x \\ 0, & otherwise \end{cases} \tag{4.60}$$

for $j > i = 1,...,n.$

The above theorem and expression (4.34) lead to the following:

$$\pi(x) = \frac{c^a}{x^{a+1}} \sum_{i=1}^{n} \frac{a^i}{\Gamma(i)} ln^{i-1}\left(\frac{x}{c}\right). \tag{4.61}$$

The sampling scheme is determined by the sequence of the conditional probability distribution functions: $F(x_i|x_{i-1})$, $i = 1,...,n$ where

$$F(x_i|x_{i-1}) = \int_{x_{i-1}}^{x_i} f(x_i|x_{i-1})dx_i = ax_{i-1}^a \int_{x_{i-1}}^{x_i} x^{-a-1}dx = \left(1 - \left(\frac{x_{i-1}}{x_i}\right)^a\right).$$

Finally:

$$F(x_i|x_{i-1}) = \begin{cases} 0 & \text{for} \quad x_i < x_{i-1}, \\ 1 - \left(\frac{x_{i-1}}{x_i}\right)^a & \text{for} \quad x \geq x_{i-1}, \end{cases} \qquad i = 1,...,n. \tag{4.62}$$

The inverse function to the distribution $u_i = F(x_i|x_{i-1})$ is as follows:

$$F^{-1}(x_i|x_{i-1}) = x_i = x_{i-1}\left(1-u_i\right)^{-\frac{1}{a}}, \quad i = 1,...,n \quad x_0 = c. \tag{4.63}$$

where u_i, $i = 1,...,n$ are values of the independently distributed uniform random variables $U_i \sim U(0,1)$, $i = 1,...,n$.

The defined algorithm can be called as gradually left truncated Pareto sampling scheme.

Example 1. Let us consider the following functions $y_1(x) = x^2$, $y_2(x) = x$ and $y_3(x) = \sqrt{x}$ defined on the interval $U = [0;1]$. Our purpose is the estimation of the following parameters:

$$g_1 = \frac{1}{3}, \quad g_2 = \frac{1}{2}, \quad g_2 = \frac{2}{3} \quad \text{where} \quad g_i = \int_0^1 y_i(x)dx \quad i = 1,2,3.$$

Firstly let us suppose that the simple random sample of the size n is selected according to the uniform distribution on the interval $(0;1)$. Its distribution function we denote by $F_0(x)$ and it is shown by expression (4.39) for $x_{i-1} = 0$ and $b = 1$. The distribution functions $F_i(y)$ of the random variables $Y_i = y_i(X)$ are as follows:

$$F_1(y) = \begin{cases} 0 \quad \text{for} \quad y \leq 0, \\ \sqrt{y} \quad \text{for} \quad y \in (0;1), \\ 1 \quad \text{for} \quad y \geq 1, \end{cases} \qquad \text{and} \quad E(Y_1) = \tfrac{1}{3}, \quad V(Y_1) = \tfrac{4}{45},$$

$$F_2(y) = F_0(y) \quad \text{where } F(.) \text{ is defined above and} \quad E(Y_2) = \frac{1}{2}, \quad V(Y_2) = \frac{1}{12},$$

$$F_3(y) = \begin{cases} 0 \quad \text{for} \quad y \leq 0, \\ y^2 \quad \text{for} \quad y \in (0;1), \\ 1 \quad \text{for} \quad y \geq 1. \end{cases} \qquad \text{and} \quad E(Y_3) = \tfrac{2}{3}, \quad V(Y_3) = \tfrac{1}{18},$$

Let $\mathbf{x} = (x_1,...,x_j,...,x_n)$ be the simple random sample of size n drawn from the distribution $F_0(x)$. The above results lead to conclusion that the parameters g_i, $i = 1,2,3$, can be estimated on the basis of the following simple random sample means:

$$\bar{y}_{i,n} = \frac{1}{n}\sum_{j=1}^{n} y_i(x_j), \quad i = 1,2,3.$$

It is easy to show that

$$E(\bar{y}_{i,n}) = g_i, \quad V_{0,i} = V(\bar{y}_{i,n}) = \frac{1}{n}V(y(x)), \quad i = 1,2,3. \tag{4.64}$$

Now let us take into account the sampling designs and schemes defined by the expressions (4.38) and (4.40) under the assumption that $a = 0$, $b = 1$. The parameters g_i, $i = 1,2,3$, are estimated by means of the above defined Horvitz-Thompson statistic, given by (4.35). The variances of this statistic under the considered sampling designs can be calculated by means of Theorems 4.7 - 4.9.

In our case we asses the variances on the basis of the following simulation experiment. Let $\mathbf{x}$ be samples of size n drawn according to gradually left truncated uniform sampling schemes defined by the conditional distribution functions $F_k(x_i|x_{i-1})$, defined by (4.40).

The sample $\mathbf{x}$ is independently replicated $t = 1,...,10000$-times. Next, values of the estimator $g_{y,i}(\mathbf{x})$, given by (4.35), are calculated and the variance is assessed in the following way:

$$V_{*i} = V_*(g_{y,i}(\mathbf{x})) = \frac{1}{10000}\sum_{k=1}^{10000}(g_{y,i}(\mathbf{x}) - \mathbf{g_i})^2 \tag{4.65}$$

Finally, the evaluated variances are compared with the variance of the simple random sample by means of the following well known relative efficiency coefficient:

$$e_i = \frac{V_{*i}}{V_{0,i}}, \quad i = 1,...,5.$$

Table 4.1 The relative efficiency (%) of the estimation under the gradually left truncated uniform sampling schemes.

n	g_1	g_2	g_3
2	37	57	29
3	31	321	81
4	63	718	203

Source: own calculations.

Analysis of Table 4.1 let us conclude that the left gradually truncated sampling design is more efficient than the ordinary simple random sampling only in the case of the sample size equal to 2 in the case of all considered parameters. Moreover, the relative efficiency coefficient decreases when the sample size increases. Hence, the left truncated sampling scheme can be useful on the second stage of the two stages sampling design.

Chapter 5
Substantive tests based on mixture distributions

In this chapter a set of items with non-zero error amounts is a domain in the accounting population. Similarly like in Subsection 1.2, book amounts are treated as values of a random variable whose distribution is a mixture of the distributions of correct (true) amount and the distribution of the amount contaminated by the error. The mixing coefficient is equal to the proportion of the items with non-zero error amounts. A mixture of two Poisson distributions is taken into account. The well-known method of moments is proposed to estimate the mixtures of probability distributions. This lets us construct some statistics in order to test the outlined hypotheses. Moreover, the well-known ratio likelihood method of inference is considered. The main results of the chapter are evaluated using the model approach, including the Bayesian rules of statistical inference. Moreover, the design-based approach as well as the mixed model-design approach are taken into account.

5.1 Model of accounting observations

Let $F_0(y|\theta_0)$ be the probability distribution function of the random variable Y, whose values are true accounting values (amounts) and $\theta_0 \in \Theta_0$ where Θ_0 is the parameter space. The values of the random variable W can be treated as accounting amounts contaminated by the errors. The distribution function of W is denoted by $F_1(w|\theta_1)$, where $\theta_1 \in \Theta_1$. Moreover, let $\Theta = \Theta_0 \cup \Theta_1$.

Let us suppose that an accounting error appears with probability p. Formally, we can write $Z = 1$ when an accounting error occur and $Z = 0$ when it does not occur. We assume that $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$.

Let the values of the random variable X be the observations of the accounting amounts generated as follows. $X = Y$ when $Z = 0$ and $X = W$ when $Z = 1$, see Subsection 1.2. Therefore, $F(x|Z = 0) = F_0(x|\theta_0)$ and $F(x|Z = 1) = F_1(w|\theta_1)$. Finally, according to the well-known total probability theorem we have:

$$F(x) = F(x|Z = 0)P(Z = 0) + F(x|Z = 1)P(Z = 1)$$

$$\begin{aligned} F(x) = P(X < x) &= P(X < x, Z = 0) + P(X < x, Z = 1) = \\ &= P(X < x|Z = 0)P(Z = 0) + P(X < x|Z = 1)P(Z = 1) = \\ &= P(Y < x|Z = 0)P(Z = 0) + P(W < x|Z = 1)P(Z = 1) = \\ &= F(x|Z = 0)P(Z = 0) + F(x|Z = 1)P(Z = 1) \end{aligned}$$

and finnally

$$F(x|\theta) = (1-p)F_0(x|\theta_0) + pF_1(x|\theta_1) \tag{5.1}$$

where $\theta = \theta_0 \cup \theta_1$ and $\theta \in \Theta = \Theta_0 \cup \Theta_1$ is the parameter space. Hence, the probability distribution of observed accounting amounts is a mixture of the distribution function $F_0(x|\theta_0)$ of the true amounts and the distribution function $F_1(x|\theta_1)$ of the amounts contaminated by the errors. Let us note that the above model for generating values of the random variable X in some sense reminds us of *Schrödinger's cat* (1935), which is used to explain of observations in quantum physics.

When the random variables Y and W are continuous, by differentiating both sides of equation (5.1) we have:

$$f(x|\theta) = (1-p)f_0(x|\theta_0) + pf_1(x|\theta_1). \tag{5.2}$$

Therefore, the probability density of observed accounting amounts is the mixture of the density $f_0(x|\theta_0)$ of the true amounts and the density $f_1(x|\theta_1)$ of the amounts contaminated by the errors. Another model based on mixture distributions considered by Kaplan (1973).

In the case of discrete probability distribution we have:

$$\begin{aligned} P(X = x|\theta) &= (1-p)P(X = x|Z = 0, \theta_0) + pP(X = x|Z = 1, \theta_1) = \\ &= (1-p)P_0(Y = x|\theta_0) + pP_1(W = x|\theta_1) \end{aligned} \tag{5.3}$$

Let D and Y be independent where D is the accounting error. Hence:

$$W = Y + D, \qquad X = Y + ZD, \qquad X = (1-Z)Y + ZW.$$

Let us suppose that an accounting error $d = 0$ with probability $1-p$. When $d \neq 0$, then the accounting error is distributed according the density function denoted by $f_2(d|\theta_2)$. Hence, the distribution of the accounting error can be defined as follows:

$$f_3(d|\theta_2, p) = pf_2(d|\theta_2)I(d) + (1-p)(1-I(d))$$

where $I(d) = 1$ when $d \neq 0$ and $I(d) = 0$ when $d = 0$. The above density (in a more general context) was considered by Kvanli et al. (1998) and Chen et al. (1998).

The basic moments of the random variable X are, see Appendix 6.1.7:

$$E(X) = pE(X|Z = 1) + (1-p)E(X|Z = 0) = pE(W) + (1-p)E(Y), \tag{5.4}$$

$$V(X) = p(1-p)\left((E(X|Z=1)-E(X|Z=0))^2 + \right.$$
$$+ pV(X|Z=1) + (1-p)V(X|Z=0) =$$
$$= p(1-p)(E(W)-E(Y))^2 + pV(W) + (1-p)V(Y), \quad (5.5)$$

$$C_3(X) = p(E(X|Z=1)-E(X))^3 + (1-p)(E(X|Z=0)-E(X))^3 +$$
$$+ 3p(E(X|Z=1)-E(X))V(X|Z=1)) +$$
$$+ 3(1-p)(E(X|Z=0)-E(X))V(X|Z=0)) +$$
$$+ pC_3(X|Z=1) + (1-p)C_3(X|Z=0) =$$
$$= p(1-p)(1-2p)(E(W)-E(Y))^3 - 3p(1-p)(E(W)-E(Y))V(Y) +$$
$$+ 3p(1-p)(E(W)-E(Y))V(W) + pC_3(W) + (1-p)C_3(Y), \quad (5.6)$$

$$C_4(X) = p(E(X|Z=1)-E(X))^4 + (1-p)(E(X|Z=0)-E(X))^4 +$$
$$+ 6p(E(X|Z=1)-E(X))^2V(X|Z=1)) +$$
$$+ 6p(1-p)(E(X|Z=0)-E(X))^2V(X|Z=0)) +$$
$$+ 4p(E(X|Z=1)-E(X))C_3(X|Z=1)) +$$
$$+ 4p(1-p)(E(X|Z=0)-E(X))C_3(X|Z=0)) +$$
$$+ pC_4(X|Z=1) + (1-p)C_4(X|Z=0) =$$
$$= p(1-p)(3p^2-3p+1)(E(W)-E(Y)) + 6p(1-p)^2(E(W)-E(Y))^2V(W) +$$
$$+ 6p^2(1-p)V(Y) - 4p(1-p)(E(W)-E(Y))C_3(W) +$$
$$+ 4p(1-p)(E(W)-E(Y))C_3(Y) + pC_4(W) + (1-p)C_4(Y) \quad (5.7)$$

where $C_r(X|Z=k) = E((X-E(X|Z=k))^r|Z=k)$, $k=0,1$, $r=0,1,2,...,C_2(...) = V(...)$.

We assume that the elements of the random vector $\mathbf{X} = \mathbf{X}_U = [X_1...X_N]$ are attached to the appropriate population elements. Before auditing, we assume that the elements of the vector $\mathbf{X} = [X_1, X_2, ..., X_N]$ are independent and identically distributed random variables. Let an auditor arbitrarily select the sample s of size n where the sub-vector $\mathbf{X}_s$ of $\mathbf{X}_U$, $n \leq N$ is observed. The random vector X_s is observed in s where the objects are controlled. After the auditing process, the sample s is split into two disjoint sub-samples s_0 and s_1, where $s_0 \cup s_1 = s$. The set s_1 is of size $n_1 = k$ and the set s_0 is of size $n_0 = n-k$. In the sub-sample s_0, there are accounting amounts observed without errors. These are values of the random variables denoted by $\{X_i = Y_i, i \in s\}$ or $\mathbf{X}_{s_0} = \mathbf{X}_{s_0}$. In the sub-sample s_1, accounting amounts contaminated by the errors are observed as values of the random variables $\{W_i, i \in s_1\}$ or $\mathbf{X}_{s_1} = \mathbf{X}_{s_1}$.

Therefore, after the auditing process we have observations of the following data:

$$\mathbf{X}_U = \{X_i : i \in U\} = (X_s, X_{U-s})$$

where

$$\mathbf{X}_s = \{X_i : i \in s\}, \qquad \mathbf{X}_{U-s} = \{X_i : i \in U - s\}$$

and

$$\mathscr{X}_U = (\mathscr{X}_s, \mathbf{X}_{U-s}) \qquad \text{where} \qquad \mathscr{X}_s = \{(X_i, Z_i) : i \in s\}.$$

Finally, let D_i be the accounting errors observed in the sub-sample s_1.

Using the model approach, our purpose is to test the hypothesis about the expected value of the following difference between the sum observed in the population of accounting amounts and the sum of the true values of those amounts. On the basis of equation (5.4) we have:

$$\begin{aligned}\tau(\theta) &= E\left(\sum_{i\in U} X_i - \sum_{i\in U} Y_i\right) = N(E(X|\theta) - E(Y|\theta_0)) = \\ &= N(E(X|\theta) - E(X|Z=0,\theta_0)) = Np(E(X|Z=1,\theta_1) - E(X|Z=0,\theta_0)) = \\ &= Np(E(X|\theta_1) - E(Y|\theta_0)) = Np(E(W|\theta_1) - E(Y|\theta_0)).\end{aligned}$$

Hence:

$$\tau(\theta) = N\bar{\tau}(\theta) \quad \text{where} \quad \bar{\tau}(\theta) = E(X|\theta) - E(Y|\theta_0) = p(E(W|\theta_1) - E(Y|\theta_0)). \tag{5.8}$$

The hypotheses considered in chapter 1 can be rewritten as follows

$$H_0: \quad \bar{\tau}(\theta) = \bar{\tau}_0, \qquad H_1: \quad \bar{\tau}(\theta) \neq \bar{\tau}_0 \tag{5.9}$$

On the basis of expression (5.2) and the above results, we can derive the following likelihood function:

$$L(\mathscr{X}_U|\theta) = L(\mathscr{X}_s|\theta)\, L(\mathbf{X}_{U-s}|\theta)$$

where

$$\begin{aligned}L(\mathscr{X}_s|\theta) &= \prod_{i\in s}[P(Z=1)f_1(x_i|\theta_1)]^{z_i}[P(Z=0)f_0(x_i|\theta_0)]^{1-z_i} = \\ &= \prod_{i\in s} p^{z_i} f_1^{z_i}(x_i|\theta_1)(1-p)^{1-z_i} f_0^{1-z_i}(x_i|\theta_0) = \\ &= \prod_{i\in s} p^{z_i} f_1^{z_i}(w_i|\theta_1)(1-p)^{1-z_i} f_0^{1-z_i}(y_i|\theta_0),\end{aligned}$$

$$L(\mathbf{X}_{U-s}|\theta) = \prod_{i\in U-s} f(y_i|\theta).$$

Hence:

$$L(\mathscr{X}_U|\theta)) = \prod_{i\in s} p^{z_i} f_1^{z_i}(w_i|\theta_1)(1-p)^{1-z_i} f_0^{1-z_i}(y_i|\theta_0) \prod_{i\in U-s} f(y_i|\theta). \tag{5.10}$$

Its logarithmic version is:

$$l(\mathscr{X}_U|\theta)) = ln(p)\sum_{i\in s} z_i + ln(1-p)\sum_{i\in s}(1-z_i) + \sum_{i\in s} z_i ln(f_1(x_i|\theta_1)) + \\ + \sum_{i\in s}(1-z_i)ln(f_0(x_i|\theta_0)) + \sum_{i\in U-s} ln(f(x_i|\theta)) = \\ = kln(p) + (n-k)ln(1-p) + \sum_{i\in s_1} ln(f_1(x_i|\theta_1)) + \sum_{i\in s_0} ln(f_0(x_i|\theta_0)) + \\ + \sum_{i\in U-s} ln(f(x_i|\theta)). \quad (5.11)$$

The hypotheses expressed by equation (5.9) can be verified by means of the well-known likelihood ratio test on the basis of the following statistic:

$$\lambda_s = \frac{\sup_{\theta\in\Theta,\bar{\tau}(\theta)=\bar{\tau}_0} L(\mathscr{X}_U|\theta)}{\sup_{\theta\in\Theta} L(\mathscr{X}_U|\theta)}. \quad (5.12)$$

We can expect that when hypothesis H_0 is true and N, n, $N-n$, n_0 and $n-n_0$ are sufficiently large then the statistic $Q_s = -2ln(\lambda_s)$ is well approximated by the chi-square distribution with one degree of freedom (see e.g. Silvey (1959)). Hypothesis H_0 is rejected if the value of the test statistic Q_s is significantly large.

The well-known maximum likelihood method or method of moments can be used to estimate the function $\bar{\tau}(\theta)$. The estimator of that function let us construct test statistics to verify the above hypotheses. This problem is considered in details in the next subsections.

5.2 Inference on the basis of the Poisson distribution

5.2.1 Basic properties

In the previous subsection it was assumed that values of the random variable Y are treated as true accounting amounts. We assume that the variable $Y \sim Pois(a)$ (the true accounting amount) and $D \sim Pois(b)$ (the accounting error) where $Pois(\lambda)$ is the Poisson distribution with the following probability function:

$$f(k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad \text{for} \quad k = 0,1,2,...$$

Under the assumption that Y and D are independent, the random variable $W = Y + D \sim Pois(a+b)$ where values of the random variable W are treated as accounting amounts contaminated by the errors. According to the results of the previous subsection the observed value before the auditing process is defined as the value of the mixture of the distributions of the variables Y and W. Equation (5.3) leads to the following results:

$$f(x|a,b,p) = pf_1(x|a,b) + (1-p)f_0(x|a) \quad (5.13)$$

where

$$f_1(x|a,b) = \frac{(a+b)^x}{x!}e^{-a-b}, \quad f_0(x|a) = \frac{a^x}{x!}e^{-a} \tag{5.14}$$

for x=0,1,2,... Therefore, the probability density of observed accounting amounts is the mixture of the density $f_0(x|a)$ of the true amounts and the density $f_1(x|a,b)$ of the amounts contaminated by the errors. Let us suppose that any sample s has not been selected form the population in order to audit its elements. So, in this case, $s = \varnothing$. In this situation, it is possible to test the hypothesis stated by expression (5.9) about the total audit error on the basis of the estimators of the parameters a, b and p. They can be derived using the well-known method of moments or maximum likelihood method.

5.2.2 Inference on the basis of sample moments

As is well-known, the moments of $\Psi \sim Pois(\lambda)$ are as follows:

$$E(\Psi) = V(Z) = C_3(Z) = \lambda, \quad E(\Psi^2) = \lambda(1+\lambda), \quad E(\Psi^3) = \lambda(1+\lambda)^2. \tag{5.15}$$

Firstly, let us assume that the sample s is empty. In this situation expressions (5.3)-(5.5) and (5.14)-(5.15) lead to the following results:

$$\begin{cases} E(X) = a + pb, \\ V(X) = a + pb + p(1-p)b^2, \\ C_3(X) = a + pb + 3p(1-p)b^2 + p(1-p)(1-2p+2p^2)b^3. \end{cases} \tag{5.16}$$

The moments are estimated as follows:

$$\bar{X}_{e,s} = \frac{1}{n}\sum_{i\in s} X_i^e, \qquad C_{e,s}(X) = \frac{1}{n}\sum_{i\in s}(X_i - \bar{X}_U)^e, \qquad s \subseteq U.$$

Particularly when $s = U$:

$$\bar{X}_{1,U} = \bar{X}_U, \qquad C_{2,U}(X) = V_U(X).$$

The well-known method of moments and the above expressions lead to the following system of equations:

$$\begin{cases} \bar{X}_U = a + pb, \\ V_U(X) = a + pb + p(1-p)b^2, \\ C_{3,U}(X) = a + pb + 3p(1-p)b^2 + p(1-p)(1-2p+2p^2)b^3. \end{cases}$$

The solution of this system provides the consistent estimators of the parameters a, b and p. In Appendix 6.1.8 the following solution of the above system is derived:

$$\begin{cases} p_{1,U} = \frac{1-\sqrt{1-4z_U}}{2} \quad \text{or} \quad p_{2,U} = \frac{1+\sqrt{1-4z_U}}{2}, \quad \text{provide} \quad z_U \leq \frac{1}{4}, \\ a_{i,U} = \bar{X}_U - p_{i,U} b_U, \quad i = 1,2, \\ b_{i,U} = \sqrt{\frac{V_U(X) - \bar{X}_U}{p_{i,U}(1-p_{i,U})}}, \quad i = 1,2 \end{cases} \tag{5.17}$$

where

$$z_U = \frac{1}{2} + \frac{1-\sqrt{8A_U+1}}{8A_U}, \qquad \text{for} \quad 0 < A_U < 1,$$

$$A_U = \frac{(V_U(X) - \bar{X}_U)^3}{(C_{3,U}(X) - 3V_U(X) + 2\bar{X}_U)^2} > 0.$$

On the basis of equations (5.8), (5.15) and (5.16) we have $\bar{\tau} = E(X|a,b,p) - E(Y|a) = pb$. This and expression (5.17) let us construct the following test statistic for hypothesis (5.9):

$$G_{i,U} = \frac{p_{i,U} b_{i,U} - \bar{\tau}}{\sqrt{V_s(p_{i,U} b_{i,U})}} \tag{5.18}$$

where $V_s(p_{i,U} b_{i,U})$ for $i = 1,2$ are consistent estimators of the variance $V(p_{i,U} b_{i,U})$. The estimator $V_s(p_{i,U} b_{i,U})$ can be constructed by means of the well-known bootstrap method (see Efron (1979)). Moreover, let us note that expression (5.17) leads to the conclusion that the statistics $p_{i,U}$ and $b_{i,U}$ are the functions of the moments $\bar{X}_U$, $V_U(X)$ and $C_3(X)$, which can be denoted by $f(\bar{X}_U, V_U(X), C_3(X))$. So, the estimator of the variance $V(p_{i,U} b_{i,U})$ can be evaluated by means of the well-known method of moments and the Taylor expansion of the function $f(\bar{X}_U, V_U(X), C_3(X))$ (see e.g. Cramér (1946) or Rao (1973)). The central theorems let us show that under sufficiently large N, $N-n$ and n, the test statistics $Z_{i,U}$, $i = 1,2$ have asymptotically normal distributions which let us verify hypothesis (5.9).

Now let us consider a situation where the sample s is not empty. Firstly, we assume that all accounting amounts observed in the sample s are free from errors. Hence, $k = 0$. In this case, the parameters a, b and $0 < p < 1$ can be estimated on the basis of the following equations:

$$\begin{cases} E(Y) = a, \\ E(X) = a + pb, \\ V(X) = a + pb + p(1-p)b^2. \end{cases}$$

After replacing the moments of the random variable with the appropriate sample moments, we have:

$$\begin{cases} \bar{Y}_s = a_s, \\ \bar{X}_{U-s} = a_s + pb, \\ V_{U-s}(X) = \bar{X}_{U-s} + p(1-p)b^2 \end{cases}$$

where $\bar{Y}_s = \frac{1}{n}\sum_{i \in s} X_i^e$. The solution of the above system is as follows:

$$\begin{cases} a_s = \bar{Y}_s, \\ b_s = \frac{(\bar{X}_{U-s}-\bar{Y}_s)^2+V_{U-s}(X)-\bar{X}_{U-s}}{\bar{X}_{U-s}-\bar{Y}_s}, \\ p_s = \frac{(\bar{X}_{U-s}-\bar{Y}_s)^2}{(\bar{X}_{U-s}-\bar{Y}_s)^2+V_{U-s}(X)-\bar{X}_{U-s}}, \end{cases} \tag{5.19}$$

provided that $\bar{X}_{U-s} > \bar{Y}_s$ and $V_{U-s}(X) \geq \bar{X}_{U-s}$.

Expression (5.8), (5.15), (5.19) lead to the following estimator of parameter $\bar{\tau} = E(X|a,b,p) - E(Y|a) = pb$:

$$\bar{T}_s = p_s b_s = \bar{X}_{U-s} - \bar{Y}_s.$$

The statistic $\bar{T}_s$ is the unbiased estimator of the parameter $\bar{\tau}$. Finally, we have the following test statistic of the hypothesis (5.9):

$$G_{1,s} = \frac{\bar{T}_s - \bar{\tau}}{\sqrt{V_s(\bar{T}_s)}} \tag{5.20}$$

When hypothesis H_0 is true and the sizes N and $N-n$ are sufficiently large the statistic $G_{1,s}$ has asymptotically normal distribution.

When $0 < k \leq n < N$ and $0 < p < 1$ we have:

$$\begin{cases} \bar{\tau} = E(X) - E(Y) = pb, \\ E(X) = a + pb, \\ V(X) = a + pb + p(1-p)b^2. \end{cases}$$

This leads to the following:

$$\begin{cases} \bar{T}_s = \bar{X}_s - \bar{Y}_s = pb, \\ \bar{X}_U = a + pb, \\ V_U(X) = a + pb + p(1-p)b^2, \end{cases}$$

$$\begin{cases} \bar{T}_s = \bar{X}_s - \bar{Y}_s = pb, \\ \bar{X}_U = a + pb, \\ V_U(X) = \bar{X}_U + \frac{1-p}{p}\bar{T}_s^2. \end{cases}$$

Hence, the estimators of the parameters are:

$$\begin{cases} a_s = \bar{X}_U - p_s\bar{T}_s, \\ b_s = \frac{\bar{T}_s}{p_s}, \\ p_s = \frac{\bar{D}_s^2}{\bar{D}_s^2+V_U(X)-\bar{X}_U}, \end{cases} \tag{5.21}$$

provided that $V_U(X) \geq \bar{X}_U$.

Finally, when $0 < k \leq n \leq N$, we have the following test statistic for hypothesis (5.9):

$$G_{2,s} = \frac{\bar{T}_s - \bar{\tau}}{\sqrt{V_s(\bar{T}_s)}}. \tag{5.22}$$

When hypothesis H_0 is true and the sizes N, n and $N-n$ are sufficiently large the test statistic $G_{2,s}$ has asymptotically normal distribution.

5.2.3 Inference on the basis of the likelihood function

Expressions (5.13) - (5.14) lead to the following log-likelihood function:

$$l(\mathscr{X}_U, a, b, p) = kln(p) + (n-k)ln(1-p) - na - kb + ln(a)\sum_{i\in s_0} x_i + \\ + ln(a+b)\sum_{i\in s_1} x_i - \sum_{i\in s} ln(x_i!) + \sum_{i\in U-s} ln(f(x_i|a,b,p)). \tag{5.23}$$

The derivatives of this function are as follows:

$$\frac{\partial l(\mathscr{X}_U, a, b, p)}{\partial a} = \frac{1}{a}\sum_{i\in s_0} x_i + \frac{1}{a+b}\sum_{i\in s_1} x_i - n + \\ + \sum_{i\in U-s} \frac{(1-p)(\frac{x_i}{a}-1)f_0(x_i|a,b,p) + p(\frac{x_i}{a+b}-1)f_1(x_i|a,b,p)}{f(x_i|a,b,p)},$$

$$\frac{\partial l(\mathscr{X}_U, a, b, p)}{\partial b} = p\sum_{i\in U-s} \frac{(\frac{x_i}{a+b}-1)f_1(x_i|a,b,p)}{f(x_i|a,b,p)} + \frac{1}{a+b}\sum_{i\in s_1} x_i - k,$$

$$\frac{\partial l(\mathscr{X}_U, a, p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} + \sum_{i\in U-s} \frac{f_1(x_i|a,b,p) - f_0(x_i|a,b,p)}{f(x_i|a,b,p)}.$$

Particularly, when $U = s$ we have:

$$l(\mathscr{X}_s, a, b, p) = kln(p) + (n-k)ln(1-p) - na - kb + ln(a)\sum_{i\in s_0} x_i + \\ + ln(a+b)\sum_{i\in s_1} x_i - \sum_{i\in s} ln(x_i!). \tag{5.24}$$

In Appendix 6.1.9 it is proved that the following statistics:

$$\hat{a}_s = \bar{X}_{s_0} = \bar{Y}_{s_0}, \qquad \hat{b}_s = \bar{X}_{s_1} - \bar{X}_{s_0} = \bar{X}_{s_1} - \bar{Y}_{s_0}, \qquad \hat{p}_s = \frac{k}{n} \tag{5.25}$$

are the maximum likelihood estimators of the parameters a, b and p, respectively provided $0 < n < N$, $\bar{X}_{s_1} > \hat{p}_s^2(1-\hat{p}_s)$ and $\bar{X}_{s_1} - \bar{Y}_{s_0} > 0$.

Hypothesis (5.9) can be tested by means of the following statistic:

$$G_{3,s} = \frac{\hat{\tau}_s - \bar{\tau}}{\sqrt{V_s(\hat{\tau}_s)}}$$

where

$$\hat{\tau}_s = \hat{p}_s(\bar{X}_{s_1} - \bar{Y}_{s_0}),$$

On the basis of the appropriate central theorems we can expect that under the sufficiently large sub-sample sizes k and $n-k$ the statistic $G_{3,s}$ is well approximated by normal distribution.

Let us suppose that the values of the parameters b and p are specified separately according to the following hypotheses:

$$H_0' : E(R) = b = b_0 \quad \text{and} \quad p = p_0, \qquad H_1' : b \neq b_0 \quad \text{or} \quad p \neq p_0.$$

In this case, assuming that $U = s$, expressions (5.12) (5.24) lead to the following likelihood ratio test statistic:

$$Q_{1,s} = -2ln(\lambda_{1,s})$$

where

$$\lambda_{1,s} = \frac{l\left(\mathscr{X}_U | a_{2s}, b_0, p_0\right)}{l\left(\mathscr{X}_U | \hat{a}_s, \hat{b}_s, \hat{p}_s\right)},$$

$$a_{2,s} = \frac{\bar{X}_s - b_0}{2} + \frac{1}{2n}\sqrt{n^2(\bar{X}_s - b_0) + 4n(n-k)\bar{X}_{s_0} b_0}.$$

and $\hat{a}_s, \hat{b}_s, \hat{p}_s$ are given by (5.25). $a_{2,s}$ is obtained as a solution of the equation $\frac{\partial l(a, b_0, p_0)}{\partial a} = 0$. The test statistic $Q_{1,s}$ has approximately chi-square distribution with two degrees of freedom when hypothesis H_0 is true and the sample size is sufficiently large. The hypothesis H_0' is rejected when value of $Q_{1,s}$ is significantly large.

In the considered case of the Poisson model the ratio likelihood test statistic defined by (5.12) takes the following form:

$$\lambda_{2,s} = \frac{sup_{\{0<p<1, a>0, b>0, 0<p<1, pb=\tau_0\}} l\left(\mathscr{X}_U | a, b, p\right)}{l\left(\mathscr{X}_U | \hat{a}_s, \hat{b}_s, \hat{p}_s\right)}.$$

When the hypothesis (5.9) is true and N, $N-n$ and n are sufficiently large, then the statistic

$$Q_{2,s} = -2ln(\lambda_{2,s})$$

has chi-square distribution with one degree of freedom. The hypothesis defined by (5.9) is rejected when $Q_{2,s}$ takes significantly large value.

5.2.4 Bayesian approach

The Bayesian approach is considered in order to model the parameters of the distribution of auditing errors (see e.g. Sorensen (1969) or *Statistical models...* (1989)). Let us assume that the parameter b of Poisson distribution $f_1(x|a,b)$ defined by expression (5.14) has gamma distribution $\Gamma(\nu,\theta)$ with the following density function:

$$g(b|\nu,\theta) = \frac{\theta^\nu}{\Gamma(\nu)} b^{\nu-1} e^{-\theta b} \tag{5.26}$$

where $\nu > 0$ and $\theta > 0$ are known parameters. Therefore, $\Gamma(\nu,\theta)$ is the prior distribution of the parameter b. Moreover, let the well-known beta distribution $B(\eta,\kappa)$ be the prior distribution of probability p. Hence, the density function of the random variable p is:

$$t(p) = \frac{\Gamma(\kappa+\eta)}{\Gamma(\eta)\Gamma(\kappa)} p^{\eta-1}(1-p)^{\kappa-1} \tag{5.27}$$

where

$$\Gamma(\eta) = \int_0^\infty u^{\eta-1} e^{-u} dt.$$

We assume that p and b are independently distributed.

Let us consider the particular case when $U = s$. In this situation the likelihood function (see the expression (5.10)) is as follows:

$$L(\mathscr{X}_s|a,b,p) = p^k (1-p)^{n-k} \prod_{i\in s_0} f_0(y_i|a) \prod_{j\in s_1} f_1(w_j|a,b). \tag{5.28}$$

The posterior distribution of b and p under the fixed observation $\mathscr{X}_s$ (which is evaluated in the Appendix 6.1.10) is as follows:

$$q(b,p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa) = q_1(b|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa) q_2(p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa) \tag{5.29}$$

where

$$q_1(b|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa) = \frac{b^{\nu-1} e^{-(k+\theta)b} (a+b)^m}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} =$$

$$= \frac{e^{-(k+\theta)b} \sum_{h=0}^m \binom{m}{h} a^h b^{m+\nu-h-1}}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}}, \tag{5.30}$$

$$q_2(p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa) = \frac{\Gamma(n+\eta+\kappa)}{\Gamma(k+\eta)\Gamma(n-k+\kappa)} p^{k+\tau-1}(1-p)^{n-k+\kappa-1}. \tag{5.31}$$

Hence, the posterior distributions of b and p under the fixed $\mathscr{X}_s$ are independent.

According to the framework of Bayesian inference the hypotheses are formulated as follows:

$$H_0 : \bar{\tau} \le \bar{\tau}_1, \qquad H_1 : \bar{\tau} > \bar{\tau}_1 \tag{5.32}$$

where $\bar{\tau} = pb$. Let c_0 be the loss generated by accepting H_0, when H_1 is true. Let c_1 be the loss generated by rejecting H_0, when it is true. This means that the observed admissible total accounting error is not accepted. In our case, according to the general Bayesian rule of testing statistical hypotheses (see e.g. Krzyśko (2004)) the following posterior probabilities are evaluated:

$$P(\bar{\tau} = pb \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa). \tag{5.33}$$

So, $c_1 P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa)$ is the risk of accepting hypothesis H_1 when H_0 is true. Moreover, $c_0 P(\bar{\tau} > \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa)$ is the risk of accepting H_0 when H_1 is true. The decision rule is as follows. Hypothesis:
H_0 is rejected when $c_1 P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) \leq c_0 P(\bar{\tau} > \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa)$,
H_0 is accepted when $c_1 P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) > c_0 P(\bar{\tau} > \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa)$.
This decision rule is equivalent to the following:
H_0 is rejected when $P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) \leq \frac{c_0}{c_0 + c_1} = r$,
H_0 is accepted when $P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) > r$. Let us note that if $c_0 = c_1$, then $r = 0.5$.

The next decision rule is based on the following Bayes factor (see, e.g. Robert (2007), pp. 227):

$$B_f = \frac{P(\bar{\tau} \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) P(\bar{\tau} > \bar{\tau}_1 | a, \nu, \theta, \eta, \kappa)}{P(\bar{\tau} > \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) P(\bar{\tau} \leq \bar{\tau}_1 | a, \nu, \theta, \eta, \kappa)}. \tag{5.34}$$

Let $l = log_e(B_f)$. The decision rule is as follows (see Lodewyckx et al. (2011), compare Raftery (1995)):

If $0 < l \leq 1$, there is weak support for H_0,
if $1 < l \leq 3$, there is positive support for H_0,
if $3 < l \leq 5$, there is strong support for H_0,
if $l > 5$, there is very strong support for H_0.
Other scales are considered e.g. by Kass and Raftery (1995).

Finally, let us note that $P(\bar{\tau} = pb \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa)$ is evaluated (see Appendix 6.1.10) as follows:

$$\begin{aligned} P(\bar{\tau} = pb \leq \bar{\tau}_1 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) &= P\left(b \leq \frac{\bar{\tau}_1}{p} | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa\right) = \\ &= \frac{\Gamma(n+\eta+\kappa)}{\Gamma(k+\eta)\Gamma(n-k+\kappa)\sum_{h=0}^{m}\binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} \\ \sum_{h=0}^{m}\binom{m}{h} c_h \int_0^1 \Gamma\left(m+\nu-h, \frac{\bar{\tau}_1(k+\theta)}{p}\right) p^{k+\eta-1}(1-p)^{n-k+\kappa-1} dp \end{aligned} \tag{5.35}$$

where $\Gamma\left(m+\nu-h, \frac{\bar{\tau}_1(k+\theta)}{p}\right)$ is the incomplete gamma function and

$$c_h = \frac{a^h}{(k+\theta)^{m+\nu-h-1}}.$$

In order to evaluate the above integrals the appropriate numerical methods have to be applied.

5.3 Model-design approach

The design-based approach to testing the hypothesis about total audit error is widely considered in Chapter 4. Here, the so-called model-design approach is taken into account.

Similarly to the previous sections, let the amount value x_i be the value of the random variables X_i, $k = 1,...,N$. Moreover, let all assumptions stated about the probability distribution of the random vector $\mathbf{X}_U$ be still valid. Let the sub-vector $\mathbf{X}_s$ of $\mathbf{X}_U$ be selected in such a way that s is the sample of size n drawn from the population U according to the sampling design $P(s)$.

Let us consider a simple random sample drawn without replacement. Under the assumed model hypothesis (5.9) can be tested on the basis of the following statistic:

$$G_{4,S} = \frac{X_U - Y_S - \tau}{\sqrt{V_S(X_U - Y_S)}} \tag{5.36}$$

where

$$X_U = \sum_{i \in U} X_i, \qquad Y_S = \sum_{i \in S} Y_i,$$

$$V_S(X_U - Y_S) = (N-n)V_S(y) + NV_S(x-y), \tag{5.37}$$

$$V_S(y) = \frac{1}{n-1}\sum_{i \in s}(Y_i - \bar{Y}_S)^2, \qquad V_S(x-y) = \frac{1}{n-1}\sum_{i \in s}(X_i - Y_i - \bar{X}_S + \bar{Y}_S)^2.$$

The expression (5.37) is derived in Appendix 6.1.11.

Usually, under some quite weak assumptions and appropriate central theorems we can infer that the statistic $G_{4,S}$ has asymptotically normal distribution when $N \to \infty$, $n \to \infty$ and $N - n \to \infty$.

Let $\varepsilon = [\varepsilon_1 ... \varepsilon_i ... \varepsilon_N]$ be an N-dimensional binary random variable where $\varepsilon_i = 1$ or $\varepsilon_i = 0$ when $i \in S$ or $i \notin S$, respectively. Hence, $S = \{i : \ \varepsilon_i = 1, \ i = 1,...,N\}$. Moreover, $E(\varepsilon_i) = \pi_i$, $E(\varepsilon_j \varepsilon_i) = \pi_{j,i}$, $j \neq i$, $j,i = 1,...,N$.

The data involving sampling of the population and auditing the selected sample can be written as follows (see the section 5.1):

$$\mathscr{X}_{U,\pi} = (\mathscr{X}_{S_0,\pi}, \mathscr{X}_{S_1,\pi}, \mathscr{X}_{U-S,\pi})$$

where $S = S_0 \cup S_1$ and

$$\mathscr{X}_{S_1,\pi} = \{(X_i, Y_i) : \ \varepsilon_i = 1, Z_i = 1, i \in U\},$$

$$\mathscr{X}_{S_0,\pi} = \{X_i = Y_i : \ \varepsilon_i = 1, Z_i = 0, i \in U\},$$

$$\mathscr{X}_{U-S,\pi} = \{X_i : \ \varepsilon_i = 0, i \in U\}.$$

The likelihood function of the data $\mathscr{X}_{U,\pi}$ is as follows:

$$L(\mathscr{X}_{U,\pi}|\theta) = \prod_{i\in U}\left[p^{z_i}f_1^{z_i}(x_i|\theta_1)(1-p)^{1-z_i}f_0^{1-z_i}(x_i|\theta_0)\right]^{\frac{\varepsilon_i}{\pi_i}} f^{\frac{1-\varepsilon_i}{1-\pi_i}}(x_i|\theta).$$

The log-likelihood function is:

$$\begin{aligned} l(\mathscr{X}_{U,\pi}|\theta) = ln(p)\sum_{i\in U}\frac{\varepsilon_i z_i}{\pi_i} + \sum_{i\in U}\frac{\varepsilon_i z_i}{\pi_i}ln(f_1(x_i|\theta_1)) + \\ + ln(1-p)\sum_{i\in U}\frac{\varepsilon_i(1-z_i)}{\pi_i} + \sum_{i\in U}\frac{\varepsilon_i(1-z_i)}{\pi_i}ln(f_0(x_i|\theta_0)) + \\ + \sum_{i\in U}\frac{(1-\varepsilon_i)}{1-\pi_i}ln(f(x_i|\theta)) \end{aligned}$$

or

$$\begin{aligned} l(\mathscr{X}_{U,\pi}|\theta) = ln(p)\sum_{i\in S_1}\frac{1}{\pi_i} + \sum_{i\in S_1}\frac{ln(f_1(x_i|\theta_1))}{\pi_i} + ln(1-p)\sum_{i\in S_0}\frac{1}{\pi_i} + \\ + \sum_{i\in S_0}\frac{ln(f_0(x_i|\theta_0))}{\pi_i} + \sum_{i\in U-S}\frac{ln(f(x_i|\theta))}{1-\pi_i}. \end{aligned}$$

Let us note that it is easy to show that $E_p(l(\mathscr{X}_{U,\pi}|\theta)) = l(\mathscr{X}_U|\theta)$ (see the expression (5.11)).

The log-likelihood function can be used to test hypothesis (5.9). In particular, in the case of the Poison distribution considered in subsection 5.1, the log-likelihood function takes the following form:

$$\begin{aligned} l(\mathscr{X}_U,a,b,p) = (ln(p)-b)\sum_{i\in S_1}\frac{1}{\pi_i} + ln(1-p)\sum_{i\in S_0}\frac{1}{\pi_i} - \sum_{i\in S}\frac{1}{\pi_i}a + ln(a)\sum_{i\in S_0}\frac{x_i}{\pi_i} + \\ + ln(a+b)\sum_{i\in S_1}\frac{x_i}{\pi_i} - \sum_{i\in S}\frac{ln(x_i!)}{\pi_i} + \sum_{i\in U-S}\frac{ln(f(x_i|a,b,p))}{1-\pi_i}. \end{aligned}$$

Its expected value is equal to the right side of the expression (5.23).

When $s = U$ the last sum of the above expression disappears. Moreover, the maximum likelihood estimators of the parameters a, b and p are derived similarly as in subsection 6.1.9. The estimators are as follows:

$$A_{HT,S_0} = \frac{\sum_{i\in S_0}\frac{X_i}{\pi_i}}{\sum_{i\in S_0}\frac{1}{\pi_i}} = \frac{\sum_{i\in S_0}\frac{Y_i}{\pi_i}}{\sum_{i\in S_0}\frac{1}{\pi_i}},$$

$$B_{HT,S} = \frac{\sum_{i\in S_1}\frac{X_i}{\pi_i}}{\sum_{i\in S_1}\frac{1}{\pi_i}} - a_{HT,S_0} = \frac{\sum_{i\in S_1}\frac{W_i}{\pi_i}}{\sum_{i\in S_1}\frac{1}{\pi_i}} - a_{HT,S_0}, \qquad P_{HT,S} = \frac{\sum_{i\in s_1}\frac{1}{\pi_i}}{\sum_{i\in S}\frac{1}{\pi_i}}.$$

It is easy to show that the statistics $A_{HT,s}$, $B_{HT,s}$ and $P_{HT,s}$ are unbiased estimators of the parameters a, b and p.

The test statistic of hypothesis (5.9) is constructed in the following way:

$$G_{5,S} = \frac{\tilde{\tau}_{HT,S} - \bar{\tau}}{\sqrt{V_S(\tilde{\tau}_{HT,S})}}$$

where

$$\tilde{\tau}_{HT,S} = P_{HT,S} B_{HT,S} \qquad \bar{\tau} = pb.$$

The estimator of the variance $V(\tilde{\tau}_{HT,S})$ denoted by $V_S(\tilde{\tau}_{HT,S})$, can be constructed using the bootstrap method. As in section 5.2.2, it is possible to explain that the test statistic $G_{5,S}$ has asymptotically normal distribution under some additional assumptions.

Chapter 6
Appendix

6.1 Proofs of theorems or derivations of expressions

6.1.1 Derivation of expression (1.13)

$$X = (1-Z)Y + ZW, \qquad X = Y + ZD.$$

Hence:

$$\begin{aligned} F(x) = P(X<x) = P(Y+ZD<x) = P(Z=0)P(Y<x) + P(Z=1)P(Y+D<x) = \\ = (1-p)P(Y<x) + pP(Y+D<x) = (1-p)F_y(x) + pP(W<x) = \\ = (1-p)F_y(x) + pF_w(x) \end{aligned}$$

where in the case of the independent discrete random variables Y and D:

$$\begin{aligned} F_w(x) = P(W<x) = P(Y+D<x) = \sum_{\{(d,y):d+y<x\}} P(D=d)P(Y=y) = \\ = \sum_{\{d\}} P(D=d) \sum_{\{y:y<x-d\}} P(Y=t-d) = \sum_{\{d\}} F_y(x-d)P(D=d) \end{aligned}$$

When Y and D are continuous and independent,

$$\begin{aligned} F_w(x) = P(W<x) = P(Y+D<x) = \int_{\{(u,y):u+y<x\}} f_d(u)f_y(y)dydu = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{x-u} f_d(u)f_y(y)dydu = \int_{-\infty}^{\infty} f_d(u)\left(\int_{-\infty}^{x-u} f_y(y)dy\right)du = \\ = \int_{-\infty}^{\infty} F_y(x-u)f_d(u)du. \end{aligned}$$

The derivation is based on the well-known rules for transforming random variables (see e.g Gerstenkorn and Śròdka, (1974) pp. 292 or Krzyśko (2000), pp. 132-133).

6.1.2 Derivation of expressions (3.67) and (3.68)

$$\sigma_3(\bar{Y}_{w,s}) = E(\bar{Y}_{w,s} - \mu_y)^3 = E\left(\sum_{h=1}^{H}(\bar{Y}_{s_h} - \mu_{h,y})w_h\right)^3 = E\left(\sum_{h=1}^{H}\bar{U}_{s_h}w_h\right)^3$$

where $\bar{U}_{s_h} = \bar{Y}_{s_h} - \mu_{h,s}$, and $E(\bar{U}_{s_h}) = 0$, $V(\bar{U}_{s_h}) = \frac{\sigma^2_{h,y}}{n_h}$. Moreover, the random variables $(\bar{U}_{s_1}, ..., \bar{U}_{s_H})$ are independent. Therefore:

$$\sigma_3(\bar{Y}_{w,s}) = E(\bar{Y}_{w,s} - \mu_y)^3 = E\Bigg(\sum_{h=1}^{H}\bar{U}^3_{s_h}w^3_h + 3\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\bar{U}^2_{s_h}w^2_h\bar{U}_{s_k}w_k + \\ + \sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\sum_{t=1,t\neq h,t\neq k}^{H}\bar{U}_{s_h}w_h\bar{U}_{s_k}w_k\bar{U}_{s_t}w_t\Bigg) = \sum_{h=1}^{H}w^3_h\frac{\sigma_{3,h,y}}{n^2_h}.$$

$$\sigma_4(\bar{Y}_{w,s}) = E(\bar{Y}_{w,s} - \mu_y)^4 = E\left(\sum_{h=1}^{H}(\bar{Y}_{s_h} - \mu_{h,y})w_h\right)^4 = \\ = E\left(\sum_{h=1}^{H}\bar{U}_{s_h}w_h\right)^3\sum_{h=1}^{H}\bar{U}_{s_h}w_h = E\Bigg(\sum_{h=1}^{H}\bar{U}^4_{s_h}w^4_h + 4\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\bar{U}^3_{s_h}w^3_h\bar{U}_{s_k}w_k + \\ + 3\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\bar{U}^2_{s_h}w^2_h\bar{U}^2_{s_k}w^2_k + 6\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\sum_{t=1,t\neq h,t\neq k}^{H}\bar{U}^2_{s_h}w^2_h\bar{U}_{s_k}w_k\bar{U}_{s_t}w_t + \\ + \sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}\sum_{t=1,t\neq h,t\neq k}^{H}\sum_{l=1,l\neq h,l\neq k,l\neq t}^{H}\bar{U}_{s_h}w_h\bar{U}_{s_k}w_k\bar{U}_{s_t}w_t\bar{U}_{s_l}w_l\Bigg) = \\ = \sum_{h=1}^{H}E(\bar{U}^4_{s_h})w^4_h + 3\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}E(\bar{U}^2_{s_h})w^2_hE(\bar{U}^2_{s_k})w^2_k = 3\sum_{h=1}^{H}w^4_h\frac{\sigma^4_{h,y}}{n^2_h} + \\ + \sum_{h=1}^{H}w^4_h\frac{\sigma_{4,h,y} - 3\sigma^4_{h,y}}{n^3_h} + 3\sum_{h=1}^{H}\sum_{k=1,k\neq h}^{H}w^2_h\frac{\sigma^2_{h,y}}{n_h}w^2_k\frac{\sigma^2_{k,y}}{n_k} = \\ = 3\left(\sum_{h=1}^{H}w^2_h\frac{\sigma^2_{h,y}}{n_h}\right)^2 + \sum_{h=1}^{H}w^4_h\frac{\sigma_{4,h,y} - 3\sigma^4_{h,y}}{n^3_h}$$

When in the h-th stratum, the random variable U_h is uniformly distributed on the interval with length Δ_h, $h = 1, ..., H$, then we can show that:

$$\sigma^2_{h,y} = \frac{\Delta^2_h}{12}, \quad \sigma_{4,h,y} = \frac{1}{\Delta_h}\int_{-\Delta_h/2}^{\Delta_h/2}u^4du = \frac{\Delta^4_h}{80}, \quad \sigma_{4,h,y} - 3\sigma^4_{h,y} = -\frac{\Delta^4_h}{120}$$

This and the result of the above derivation lead to expression (3.67).

$$\delta_3(U_h) = E|U_h|^3 = \frac{1}{\Delta_h}\int_{-\Delta_h/2}^{\Delta_h/2}|u|^3du = \frac{1}{\Delta_h}\left(-\int_{-\Delta_h/2}^{0}u^3du+\int_0^{\Delta_h/2}u^3du\right) = \frac{\Delta_h^3}{32}.$$

The third absolute central moment of the statistic $\bar{Y}_{w,s}$ is:

$$\delta_3(\bar{Y}_{w,s}) = E|\bar{Y}_{w,s}-\mu_y|^3 = E|\bar{U}_{w,s}-\mu_y|^3 = E\left|\sum_{h=1}^{H}\bar{U}_{s_h}w_h\right|^3 \le \sum_{h=1}^{H}E\left|\bar{U}_{s_h}\right|^3 w_h^3 \le$$
$$\sum_{h=1}^{H}\frac{w_h^3}{n_h^3}\sum_{i\in s_h}E|U_i|^3 = \sum_{h=1}^{H}\frac{w_h^3}{n_h^2}E|U_h|^3 = \frac{1}{32}\sum_{h=1}^{H}\frac{w_h^3}{n_h^2}E|\Delta_h|^3$$

Therefore, the standardized third central moment is:

$$\tau_3(\bar{Y}_{w,s}) = \frac{\delta_3(\bar{Y}_{w,s})}{\sigma^3(\bar{Y}_{w,s})} = \frac{E|\bar{Y}_{w,s}-\mu_y|^3}{\sigma^3(\bar{Y}_{w,s})} \le \frac{\sqrt{12^3}}{32}\frac{\sum_{h=1}^{H}\frac{w_h^3}{n_h^2}E|\Delta_h|^3}{\left(\sum_{h=1}^{H}\frac{w_h^2}{n_h}E|\Delta_h|^2\right)^{3/2}}$$

This result directly leads to expression (3.68).

6.1.3 Derivation of expression (4.10)

On the basis of expressions (4.8) and (4.9), the derivation of the inclusion probabilities in the case of $n = 2$ is as follows:

$$\pi_1 = \sum_{k=2}^{N}P(L_1=1,L_2=k) = P(L_1=1)\sum_{k=2}^{N}P(L_2=k|L_1=1) =$$
$$= \frac{1}{N-1}\sum_{k=2}^{N}\frac{1}{N-1} = \frac{N-1}{(N-1)^2} = \frac{1}{N-1},$$

$$\pi_2 = P(L_1=1,L_2=2)+\sum_{k=3}^{N}P(L_1=2,L_2=k) = \frac{1}{(N-1)^2}+$$
$$+P(L_1=2)\sum_{k=3}^{N}P(L_2=k|L_1=2) = \frac{1}{(N-1)^2}+\frac{1}{N-1}\sum_{k=3}^{N}\frac{1}{N-2} =$$
$$= \frac{1}{(N-1)^2}+\frac{1}{N-1} = \frac{N}{(N-1)^2},$$

$$\pi_3 = P(L_1=1,L_2=3)+P(L_1=2,L_2=3)+\sum_{k=4}^{N} P(L_1=3,L_2=k) =$$
$$= \frac{1}{(N-1)^2}+\frac{1}{(N-1)(N-2)}+P(L_1=3)\sum_{k=4}^{N} P(L_2=k|L_1=3) =$$
$$= \frac{1}{(N-1)^2}+\frac{1}{(N-1)(N-2)}+\frac{1}{N-1}\sum_{k=4}^{N}\frac{1}{N-3} =$$
$$= \frac{1}{(N-1)^2}+\frac{1}{(N-1)(N-2)}+\frac{1}{N-1},$$

After generalizing the above derivation, we have:

$$\pi_k = \sum_{h=1}^{k-1} P(L_1=h,L_2=k)+\sum_{t=k+1}^{N} P(L_1=k,L_2=t) =$$
$$= \sum_{h=1}^{k-1} P(L_2=k|L_1=h)P(L_1=h)+P(L_1=k)\sum_{t=k+1}^{N} P(L_2=t|L_1=k) =$$
$$= \frac{1}{N-1}\sum_{h=1}^{k-1}\frac{1}{N-h}+\frac{1}{N-1}\sum_{t=k+1}^{N}\frac{1}{N-k} = \frac{1}{N-1}\sum_{h=1}^{k-1}\frac{1}{N-h}+\frac{1}{N-1}$$

for $k = 1,...,N-1$. Finally, we have:

$$\pi_N = \sum_{h=1}^{N-1} P(L_1=h,L_2=N) = \sum_{h=1}^{N-1} P(L_2=N|L_1=h)P(L_1=h) =$$
$$= \frac{1}{N-1}\sum_{h=1}^{k-1}\frac{1}{N-h}.$$

This result directly leads to the expression (4.10).

6.1.4 Proof of theorem 4.9

On the basis of expressions (4.37) and (4.38) we obtain:

$$f_1(x_j,...,x_1) = \prod_{i=0}^{j-1} f_1(x_{i+1}|x_i,...,x_1) = \prod_{i=0}^{j-1} f_1(x_{i+1}|x_i) = \frac{1}{\prod_{i=0}^{j-1}(b-x_i)},$$
$$j = 1,...,n-1, \quad a < x_1 < x_2 < ... < x_n < b.$$

In order to derive the marginal density function we make the following evaluations:

$$f_1(x_j,...,x_2) =$$
$$= \int_a^{x_2} f_1(x_j,...,x_2,x_1)dx_1 = \frac{1}{(b-a)\prod_{k=2}^{j-1}(b-x_k)} \int_a^{x_2} \frac{dx_1}{b-x_1} =$$
$$\frac{1}{(b-a)\prod_{k=2}^{j-1}(b-x_k)} ln(b-x_1)\Big|_a^{x_2} = \frac{ln(b-a)-ln(b-x_2)}{(b-a)\prod_{k=2}^{j-1}(b-x_k)}.$$

$$f_1(x_j,...,x_3) = \int_a^{x_3} f(x_j,...,x_2)dx_2 =$$
$$= \frac{1}{(b-a)\prod_{k=3}^{j-1}(b-x_k)} \int_a^{x_3} \frac{ln(b-a)-ln(b-x_2)}{(b-x_2)}dx_2 =$$
$$= \frac{1}{(b-a)\prod_{k=3}^{j-1}(b-x_k)} \int_0^{ln(b-a)-ln(b-x_3)} ydy = \frac{(ln(b-a)-ln(b-x_3))^2}{2(b-a)\prod_{k=2}^{j-1}(b-x_k)}$$

where $y = ln(b-a)-ln(b-x_2)$, $dy = \frac{dx_2}{b-x_2}$.

$$f_1(x_j,...,x_4) = \int_a^{x_4} f_1(x_j,...,x_3)dx_3 =$$
$$= \frac{1}{2(b-a)\prod_{k=4}^{j-1}(b-x_k)} \int_a^{x_4} \frac{(ln(b-a)-ln(b-x_3))^2}{(b-x_3)}dx_3 =$$
$$= \frac{1}{2(b-a)\prod_{k=4}^{j-1}(b-x_k)} \int_0^{ln(b-a)-ln(b-x_4)} y^2dy =$$
$$= \frac{(ln(b-a)-ln(b-x_4))^3}{2\cdot 3(b-a)\prod_{k=4}^{j-1}(b-x_k)}.$$

Generalizing the above derivation, we obtain:

$$f_1(x_j,...,x_i) = \int_a^{x_i} f_1(x_j,...,x_{i-1})dx_{i-1} =$$
$$= \frac{1}{(b-a)(i-2)!\prod_{k=i}^{j-1}(b-x_k)} \int_a^{x_i} \frac{(ln(b-a)-ln(b-x_{i-1}))^{i-2}}{(b-x_{i-1})}dx_{i-1} =$$
$$= \frac{1}{(b-a)(i-2)!\prod_{k=i}^{j-1}(b-x_k)} \int_0^{ln(b-a)-ln(b-x_i)} y^{i-2}dy =$$
$$= \frac{(ln(b-a)-ln(b-x_i))^{i-1}}{(b-a)\Gamma(i)\prod_{k=i}^{j-1}(b-x_k)}$$

where $y = ln(b-a)-ln(b-x_{i-1})$, $dy = \frac{dx_{i-1}}{b-x_{i-1}}$.

Particularly, for $i = j$ we have:

$$f_{1,i}(x_i) = f_1(x_i) = \int_a^{x_i} f(x_i, x_{i-1})dx_{i-1} =$$
$$= \frac{1}{(b-a(i-2))} \int_a^{x_i} \frac{(ln(b-a) - ln(b-x_{i-1}))^{i-2}}{b-x_{i-1}} dx_{i-1} =$$
$$= \frac{(ln(b-a) - ln(b-x_i))^{i-1}}{(b-a)\Gamma(i)}, \quad i = 1, ..., n.$$

This result directly leads to expression (4.41).

Let

$$A_{ij} = \frac{(ln(b-a) - ln(b-x_i))^{i-1}}{(b-a)(b-x_i)\Gamma(i)} \tag{6.1}$$

$$f_1(x_j, ..., x_{i+2}, x_i) = \int_{x_i}^{x_i} f_1(x_j, ..., x_{i+2}, x_{i+1}, x_i)dx_{i+1} =$$
$$= \frac{A_{ij}}{\prod_{k=i+2}^{j-1}(b-x_k)} \int_{x_i}^{x_{i+2}} \frac{dx_{i+1}}{b-x_{i+1}} = -\frac{A_{ij}}{\prod_{k=i+2}^{j-1}(b-x_k)} ln(b-x_{i+1})\Big|_{x_i}^{x_{i+2}} =$$
$$= \frac{A_{ij}(ln(b-x_i) - ln(b-x_{i+2}))}{\prod_{k=i+2}^{j-1}(b-x_k)}.$$

$$f_1(x_j, ..., x_{i+3}, x_i) = \int_{x_i}^{x_{i+3}} f(x_j, ..., x_{i+3}, x_{i+2}, x_i)dx_{i+2} =$$
$$= \frac{A_{ij}}{\prod_{k=i+3}^{j-1}(b-x_k)} \int_{x_i}^{x_{i+3}} \frac{ln(b-x_i) - ln(b-x_{i+2})}{b-x_{i+2}} dx_{i+2} =$$
$$= \frac{A_{ij}}{\prod_{k=i+3}^{j-1}(b-x_k)} \int_0^{ln(b-x_i)-ln(b-x_{i+3})} ydy =$$
$$= \frac{A_{ij}(ln(b-x_i) - ln(b-x_{i+3}))^2}{2\prod_{k=i+3}^{j-1}(b-x_k)},$$

$$f(x_j, ..., x_{i+h}, x_i) = \int_{x_i}^{x_{i+h}} f_1(x_j, ..., x_{i+h}, x_{i+h-1}, x_i)dx_{i+h-1} =$$
$$= \frac{A_{ij}}{(h-2)!\prod_{k=i+h}^{j-1}(b-x_k)} \int_{x_i}^{x_{i+h}} \frac{(ln(b-x_i) - ln(b-x_{i+h-1}))^{h-2}}{b-x_{i+h-1}} dx_{i+h-1} =$$
$$= \frac{A_{ij}}{(h-2)!\prod_{k=i+h}^{j-1}(b-x_k)} \int_0^{ln(b-x_i)-ln(b-x_{i+h})} y^{h-2}dy =$$
$$= \frac{A_{ij}(ln(b-x_i) - ln(b-x_{i+h}))^{h-1}}{(h-1)!\prod_{k=i+h}^{j-1}(b-x_k)}.$$

When $h = j - i$:

$$f_1(x_j,x_i)=\int_{x_i}^{x_j} f(x_j,x_{j-1},x_i)dx_{j-1}=$$
$$=\frac{A_{ij}}{(j-i-2)!}\int_{x_i}^{x_j}\frac{(ln(b-x_i)-ln(b-x_{j-1}))^{j-i-2}}{b-x_{j-1}}dx_{j-1}=$$
$$=\frac{A_{ij}}{(j-i-2)!}\int_0^{ln(b-x_i)-ln(b-x_j)} y^{h-2}dy=\frac{A_{ij}(ln(b-x_i)-ln(b-x_j))^{j-i-1}}{(j-i-1)!}$$

The derived result and the expression (6.1) after appropriate changing of the notation lead to the expression (4.42).

6.1.5 Proof of theorem 4.11

On the basis of expression (4.51) we have:

$$f(x_j,...,x_2)=\int_c^{x_2} f(x_j,...,x_2,x_1)dx_1=\lambda^j e^{-\lambda(x_j-c)}\int_c^{x_2}dx_1=\lambda^j(x_2-c)e^{-\lambda(x_j-c)},$$

$$f(x_j,...,x_3)=$$
$$=\int_c^{x_3} f(x_j,...,x_3,x_2)dx_2=\lambda^j e^{-\lambda(x_j-c)}\int_c^{x_3}(x_2-c)dx_2=\frac{\lambda^j}{2}(x_3-c)^2e^{-\lambda(x_j-c)},$$

$$f(x_j,...,x_4)=$$
$$=\int_c^{x_4} f(x_j,...,x_4,x_3)dx_3=\frac{\lambda^j}{2}e^{-\lambda(x_j-c)}\int_c^{x_4}(x_3-c)^2dx_3=\frac{\lambda^j}{2\cdot 3}(x_4-c)^3e^{-\lambda(x_j-c)}.$$

After generalizing the above derivations, we have:

$$f(x_j,...,x_i)=\int_c^{x_i} f(x_j,...,x_i,x_{i-1})dx_{i-1}=$$
$$=\frac{\lambda^j}{(i-2)!}e^{-\lambda(x_j-c)}\int_c^{x_i}(x_{i-1}-c)^{i-2}dx_{i-1}=\frac{\lambda^j}{(i-1)!}(x_i-c)^{i-1}e^{-\lambda(x_j-c)}=$$
$$=\frac{\lambda^j}{\Gamma(i)}(x_i-c)^{i-1}e^{-\lambda(x_j-c)}. \quad (6.2)$$

Particularly for $i=j$:

$$f(x_j)=\int_c^{x_j} f(x_j,x_{j-1})dx_{j-1}=\frac{\lambda^j}{(j-2)!}e^{-\lambda(x_j-c)}\int_c^{x_j}(x_{j-1}-c)^{j-2}dx_{j-1}=$$
$$=\frac{\lambda^j}{(j-1)!}(x_j-c)^{j-1}e^{-\lambda(x_j-c)}=\frac{\lambda^j}{\Gamma(j)}(x_j-c)^{j-1}e^{-\lambda(x_j-c)}.$$

This, after changing the notation, leads to expression (4.52).

Let

$$A_{ij} = \frac{\lambda^j}{\Gamma(i)}(x_i - c)^{i-1} e^{-\lambda(x_j - c)}$$

On the basis of (6.2) we have:

$$f(x_j,...,x_{i+2},x_i) = \int_{x_i}^{x_{i+2}} f(x_j,...,x_i)dx_{i+1} = A_{ij}\int_{x_i}^{x_{i+2}} dx_{i+1} = A_{ij}(x_{i+2} - x_i),$$

$$f(x_j,...,x_{i+3},x_i) = \int_{x_i}^{x_{i+3}} f(x_j,...,x_{i+2},x_i)dx_{i+2} = A_{ij}\int_{x_i}^{x_{i+3}} (x_{i+2} - x_i)dx_{i+2} =$$
$$= A_{ij}\int_0^{x_{i+3}-x_i} zdz = \frac{A_{ij}}{2}(x_{i+3} - x_i)^2$$

$$f(x_j,...,x_{i+4},x_i) = \int_{x_i}^{x_{i+4}} f(x_j,...,x_{i+3},x_i)dx_{i+3} = \frac{A_{ij}}{2}\int_{x_i}^{x_{i+4}} (x_{i+3} - x_i)^2 dx_{i+3} =$$
$$= \frac{A_{ij}}{2}\int_0^{x_{i+4}-x_i} z^2 dz = \frac{A_{ij}}{2\cdot 3}(x_{i+4} - x_i)^3,$$

After generalizing of the above derivation, we have:

$$f(x_j,...,x_{i+k},x_i) = \int_{x_i}^{x_{i+k}} f(x_j,...,x_{i+k-1},x_i)dx_{i+k-1} =$$
$$= \frac{A_{ij}}{(k-2)!}\int_{x_i}^{x_{i+k}} (x_{i+k-1} - x_i)^{k-2} dx_{i+k-1} =$$
$$= \frac{A_{ij}}{(k-2)!}\int_0^{x_{i+k}-x_i} z^{k-2} dz = \frac{A_{ij}}{(k-1)!}(x_{i+k} - x_i)^{k-1}.$$

This result for $k = j - i - 1$ reduces to the following form:

$$f(x_j,x_{j-1},x_i) = \frac{A_{ij}}{(j-i-2)!}(x_{j-1} - x_i)^{j-i-2}.$$

Finally, we have:

$$f_{ji}(x_j,x_i) = f(x_j,x_i) =$$
$$= \int_{x_i}^{x_j} f(x_j,x_{j-1},x_i)dx_{j-1} = \frac{A_{ij}}{(j-i-2)!}\int_{x_i}^{x_j} (x_{j-1} - x_i)^{j-i-2} dx_{j-1} =$$
$$= \frac{A_{ij}}{(j-i-2)!}\int_0^{x_j-x_i} z^{j-i-2} dz = \frac{A_{ij}}{(j-i-1)!}(x_j - x_i)^{j-i-1} =$$
$$= \frac{\lambda^j}{\Gamma(i)\Gamma(j-i)}(x_i - c)^{i-1}(x_j - x_i)^{j-i-1} e^{-\lambda(x_j - c)}.$$

After changing the notation, the obtained result leads to expression (4.53).

6.1.6 Proof of theorem 4.12

Expression (4.58), leads to the following (see also the proof of the theorem 4.11):

$$f(x_j,x_{j-1},...,x_1) = \frac{a^j c^a}{x_j^{a+1}\prod_{i=1}^{j-1} x_i}, \quad c \le x_1 \le x_2 \le ... \le x_j.$$

$$f(x_j,x_{j-1},...,x_2) = \int_c^{x_2} f(x_j,x_{j-1},...,x_1)dx_1 = \frac{a^j c^a}{x_j^{a+1}\prod_{k=2}^{j-1} x_k}\int_c^{x_2}\frac{dx_1}{x_1} =$$
$$= \frac{A_j}{\prod_{k=2}^{j-1} x_k}(ln(x_2) - ln(c)) = \frac{A_j}{\prod_{k=2}^{j-1} x_k} ln\left(\frac{x_2}{c}\right).$$

where

$$A_j = \frac{a^j c^a}{x_j^{a+1}}$$

Similarly, we derive the following:

$$f(x_j,x_{j-1},...,x_3) = \int_c^{x_3} f(x_j,x_{j-1},...,x_2)dx_2 = \frac{A_j}{\prod_{k=3}^{j-1} x_k}\int_c^{x_3}\frac{1}{x_2} ln\left(\frac{x_2}{c}\right)dx_2.$$

Let $z = \frac{x_2}{c}$, so $dx_2 = cdz$ and

$$f(x_j,x_{j-1},...,x_3) = \frac{A_j}{\prod_{k=3}^{j-1} x_k}\int_1^{\frac{x_3}{c}} z^{-1} ln(z)dz = \frac{1}{2}\frac{A_j}{\prod_{k=3}^{j-1} x_k} ln^2(z)\Big|_1^{\frac{x_3}{c}} =$$
$$= \frac{1}{2}\frac{A_j}{\prod_{k=3}^{j-1} x_k} ln^2\left(\frac{x_3}{c}\right).$$

$$f(x_j,x_{j-1},...,x_4) = \int_c^{x_4} f(x_j,x_{j-1},...,x_3)dx_3 = \frac{1}{2}\frac{A_j}{\prod_{k=4}^{j-1} x_k}\int_c^{x_4}\frac{1}{x_3} ln^2\left(\frac{x_3}{c}\right)dx_3 =$$
$$= \frac{1}{2}\frac{A_j}{\prod_{k=4}^{j-1} x_k}\int_1^{\frac{x_4}{c}} z^{-1} ln^2(z)dz.$$

Let $u = ln(z)$, so $du = \frac{dz}{z}$ and

$$f(x_j,x_{j-1},...,x_4)=\frac{1}{2}\frac{A_j}{\prod_{k=4}^{j-1}x_k}\int_0^{ln\left(\frac{x_4}{c}\right)}u^2du=\frac{1}{2\cdot 3}\frac{A_j}{\prod_{k=4}^{j-1}x_k}ln^3\left(\frac{x_4}{c}\right).$$

$$f(x_j,x_{j-1},...,x_5)=\int_c^{x_5}f(x_j,x_{j-1},...,x_4)dx_4=$$
$$=\frac{1}{2\cdot 3}\frac{A_j}{\prod_{k=5}^{j-1}x_k}\int_c^{x_5}\frac{1}{x_4}ln^3\left(\frac{x_4}{c}\right)dx_4=\frac{1}{2\cdot 3}\frac{A_j}{\prod_{k=5}^{j-1}x_k}\int_1^{\frac{x_5}{c}}z^{-1}ln^3(z)dz=$$
$$=\frac{1}{2\cdot 3}\frac{A_j}{\prod_{k=5}^{j-1}x_k}\int_0^{ln\left(\frac{x_5}{c}\right)}u^3du=\frac{1}{2\cdot 3\cdot 4}\frac{A_j}{\prod_{k=5}^{j-1}x_k}ln^4\left(\frac{x_5}{c}\right).$$

Generalizing the above derivation, we have:

$$f(x_j,x_{j-1},...,x_i)=\frac{1}{(i-1)!}\frac{A_j}{\prod_{k=i}^{j-1}x_k}ln^{i-1}\left(\frac{x_i}{c}\right)=\frac{a^jc^a}{\Gamma(i)x_j^{a+1}\prod_{k=i}^{j-1}x_k}ln^{i-1}\left(\frac{x_i}{c}\right). \tag{6.3}$$

When $i=j-1$, we can evaluate the following marginal density function:

$$f_j(x_j)=\int_c^{x_j}f(x_j,x_{j-1})dx_{j-1}=\frac{a^jc^a}{\Gamma(j-1)x_j^{a+1}}\int_c^{x_j}\frac{1}{x_{j-1}}ln^{j-2}\left(\frac{x_{j-1}}{c}\right)dx_{j-1}=$$
$$=\frac{a^jc^a}{\Gamma(j-1)x_j^{a+1}}\int_1^{\frac{x_j}{c}}z^{-1}ln^{j-2}(z)dz=\frac{a^jc^a}{(j-2)!x_j^{a+1}}\int_0^{ln\left(\frac{x_j}{c}\right)}u^{j-2}du=$$
$$=\frac{a^jc^a}{(j-1)!x_j^{a+1}}u^{j-1}\Big|_0^{ln\left(\frac{x_j}{c}\right)}=\frac{a^jc^a}{(j-1)!x_j^{a+1}}ln^{(j-1)}\left(\frac{x_j}{c}\right).$$

Finally, we have:

$$f_j(x_j)=\frac{a^jc^a}{\Gamma(j)x_j^{a+1}}ln^{(j-1)}\left(\frac{x_j}{c}\right).$$

This result directly leads to expression (4.59)

On the basis of expression (6.3) we make the following derivation:

$$f(x_j,x_{j-1},...,x_{i+2},x_i)=\int_{x_i}^{x_{i+2}}f(x_j,x_{j-1},...,x_{i+1},x_i)dx_{i+1}=$$
$$=\frac{a^jc^a}{(i-1)!x_j^{a+1}\prod_{k=i+2}^{j-1}x_k}ln^{(i-1)}\left(\frac{x_i}{c}\right)\int_{x_i}^{x_{i+2}}\frac{dx_{i+1}}{x_{i-1}}=\frac{B_{ji}}{\prod_{k=i+2}^{j-1}x_k}ln\left(\frac{x_{i+2}}{x_i}\right)$$

where

$$B_{ji}=\frac{a^jc^a}{(i-1)!x_j^{a+1}}ln^{(i-1)}\left(\frac{x_i}{c}\right).$$

$$f(x_j,x_{j-1},...,x_{i+3},x_i)=\int_{x_i}^{x_{i+3}} f(x_j,x_{j-1},...,x_{i+2},x_i)dx_{i+2}=$$
$$=\frac{B_{ji}}{\prod_{k=i+3}^{j-1}x_k}\int_{x_i}^{x_{i+3}}\frac{1}{x_{i+2}}ln\left(\frac{x_{i+2}}{x_i}\right)dx_{i+2}=\frac{B_{ji}}{\prod_{k=i+3}^{j-1}x_k}\int_1^{\frac{x_{i+3}}{x_i}}\frac{1}{z}ln(z)dz=$$
$$=\frac{B_{ji}}{2\prod_{k=i+3}^{j-1}x_k}ln^2(z)\Big|_1^{\frac{x_{i+3}}{x_i}}=\frac{B_{ji}}{2\prod_{k=i+3}^{j-1}x_k}ln^2\left(\frac{x_{i+3}}{x_i}\right).$$

where $z=\frac{x_{i+2}}{x_i}$ and $dx_{i+2}=x_i dz$.

$$f(x_j,x_{j-1},...,x_{i+4},x_i)=\int_{x_i}^{x_{i+4}} f(x_j,x_{j-1},...,x_{i+4},x_{i+3},x_i)dx_{i+3}=$$
$$=\frac{B_{ji}}{2\prod_{k=i+4}^{j-1}x_k}\int_{x_i}^{x_{i+4}}\frac{1}{x_{i+3}}ln^2\left(\frac{x_{i+3}}{x_i}\right)dx_{i+3}=\frac{B_{ji}}{2\prod_{k=i+4}^{j-1}x_k}\int_1^{\frac{x_{i+4}}{x_i}}\frac{1}{z}ln^2(z)dz=$$
$$=\frac{B_{ji}}{2\prod_{k=i+4}^{j-1}x_k}\int_0^{ln\left(\frac{x_{i+4}}{x_i}\right)}u^2du=\frac{B_{ji}}{2\cdot 3\prod_{k=i+4}^{j-1}x_k}ln^3\left(\frac{x_{i+4}}{x_i}\right).$$

where $u=ln(z)$ and $du=\frac{dz}{z}$. Hence, for $1\le h\le j-i$ we have:

$$f(x_j,x_{j-1},...,x_{i+h},x_i)=\int_{x_i}^{x_{i+h}} f(x_j,x_{j-1},...,x_{i+h},x_{i+h-1},x_i)dx_{i+h-1}=$$
$$=\frac{B_{ji}}{(h-2)!\prod_{k=i+h}^{j-1}x_k}\int_{x_i}^{x_{i+h}}\frac{1}{x_{i+h-1}}ln^{h-2}\left(\frac{x_{i+h-1}}{x_i}\right)dx_{i+h-1}=$$
$$=\frac{B_{ji}}{(h-2)!\prod_{k=i+h}^{j-1}x_k}\int_1^{\frac{x_{i+h}}{x_i}}\frac{1}{z}ln^{h-2}(z)dz=$$
$$=\frac{B_{ji}}{(h-2)!\prod_{k=i+h}^{j-1}x_k}\int_0^{ln\left(\frac{x_{i+h}}{x_i}\right)}u^{h-2}du=\frac{B_{ji}}{(h-1)!\prod_{k=i+h}^{j-1}x_k}ln^{h-1}\left(\frac{x_{i+h}}{x_i}\right)=$$
$$=\frac{B_{ji}}{(h-1)!\prod_{k=i+h}^{j-1}x_k}ln^{h-1}\left(\frac{x_{i+h}}{x_i}\right).$$

For $h=j-i$ we have:

$$f_{j,i}(x_j,x_i) = f(x_j,x_i) = \int_{x_i}^{x_j} f(x_j,x_{j-1},x_i)dx_{j-1} =$$

$$= \frac{B_{ji}}{(j-i-2)!}\int_{x_i}^{x_j}\frac{1}{x_{j-1}} ln^{j-i-2}\left(\frac{x_{j-1}}{x_i}\right)dx_{j-1} = \frac{B_{ji}}{(j-i-2)!}\int_1^{\frac{x_j}{x_i}}\frac{1}{z} ln^{j-i-2}(z)dz =$$

$$= \frac{B_{ji}}{(j-i-2)!}\int_0^{ln\left(\frac{x_j}{x_i}\right)} u^{j-i-2}du = \frac{B_{ji}}{(j-i-1)!} ln^{j-i-1}\left(\frac{x_j}{x_i}\right) =$$

$$= \frac{B_{ji}}{(j-i-1)!} ln^{j-i-1}\left(\frac{x_j}{x_i}\right) = \frac{B_{ji}}{\Gamma(j-i)} ln^{j-i-1}\left(\frac{x_j}{x_i}\right).$$

Finally:

$$f_{j,i}(x_j,x_i) = \frac{a^j c^a}{\Gamma(i)\Gamma(j-i)x_j^{a+1}} ln^{(i-1)}\left(\frac{x_i}{c}\right) ln^{j-i-1}\left(\frac{x_j}{x_i}\right).$$

This leads to expression (4.60).

6.1.7 Derivation of expression (5.5)

On the basis of expressions (5.2)-(5.4) we have:

$$\begin{aligned}
V(X) = E(X-E(X))^2 = pE(X-E(X)|Z=1)^2 + (1-p)E(X-E(X)|Z=0)^2 = \\
= pE((X-E(X|Z=1)) + (E(X)-E(X|Z=1))|Z=1)^2 + \\
+ (1-p)E((X-E(X|Z=0)) + (E(X)-E(X|Z=0))|Z=1)^2 = \\
= pE((X-E(X|Z=1))^2|Z=1) + p(E(X)-E(X|Z=1))^2 \\
+ (1-p)E((X-E(X|Z=0))^2|Z=1) + (1-p)(E(X)-E(X|Z=0))^2 = \\
= p\left((E(X|Z=1)-E(X))^2 + V(X|Z=1)\right) + \\
+ (1-p)\left((E(X|Z=0)-E(X))^2 + V(X|Z=0)\right) = \\
= p(1-p)^2(E(X|Z=1)-E(X|Z=0))^2 + pV(X|Z=1) + \\
+ (1-p)p^2(E(X|Z=0)-E(X))^2 + (1-p)V(X|Z=0) = \\
= p(1-p)\left((E(X|Z=1)-E(X|Z=0))^2 + \right. \\
+ pV(X|Z=1) + (1-p)V(X|Z=0) = \\
= p(1-p)(E(W)-E(Y))^2 + pV(W) + (1-p)V(Y),
\end{aligned}$$

Epression (5.6) and (5.7) are derived in a similar way to the above expression.

6.1.8 Derivation of expression (5.17)

System (5.16) is equivalent to the following system:

$$\begin{cases} \bar{X}_U = a + pb, \\ V_U(X) = \bar{X}_U + p(1-p)b^2, \\ C_{3,U}(X) = -2\bar{X}_U + 3V_U(X) + p(1-p)(1-2p+2p^2)b^3 \end{cases} \tag{6.4}$$

The second and the third equations of the above system lead to the following equivalent expressions:

$$\left(\frac{V_U(X)-\bar{X}_U}{p(1-p)}\right)^{1/2} = \left(\frac{C_{3,U}(X)-3V_U(X)+2\bar{X}_U}{p(1-p)(1-2p+2p))}\right)^{1/3},$$

$$\left(\frac{V_U(X)-\bar{X}_U}{p(1-p)}\right)^{3} = \left(\frac{C_{3,U}(X)-3V_U(X)+2\bar{X}_U}{p(1-p)(1-2p(1-p))}\right)^{2}$$

$$A_U = \frac{z}{(1-2z)^2}$$

where

$$z = p(1-p), \qquad A_U = \frac{(V_U(X)-\bar{X}_U)^3}{(C_{3,U}(X)-3V_U(X)+2\bar{X}_U)^2} > 0$$

Finally, we have:

$$4A_U z^2 - (4A_U+1)z + A_U = 0 \tag{6.5}$$

The second and third equations of system (6.4) lead to inequalities:

$$\begin{cases} V_U(X) - \bar{X}_U = p(1-p)b^2 > 0, \\ C_{3,U}(X) + 2\bar{X}_U - 3V_U(X) = p(1-p)(1-2p+2p^2)b^3 > 0. \end{cases}$$

Hence, $A_U > 0$. After appropriate algebraic transformation, equation (6.5) is equivalent to the following:

$$4A(z-z_{1,U})(z-z_{2,U}) = 0$$

where:

$$z_{1,U} = \frac{1}{2} + \frac{1-\sqrt{8A_U+1}}{8A_U}, \qquad z_{2,U} = \frac{1}{2} + \frac{1+\sqrt{8A_U+1}}{8A_U}.$$

The solutions of equation $z = p(1-p)$ fulfil inequalities $0 < p < 1$, when $0 < z < \frac{1}{4}$. The inequalities $0 < z_{2,U} < \frac{1}{4}$ are not fulfilled for any $A > 0$. We can show that $0 < z_{1,U} < \frac{1}{4}$, if and only if $0 < A_U < 1$. Finally, equation $z_{1,U} = p(1-p)$ have the following solutions for $0 < z_{1,U} < \frac{1}{4}$:

$$p_{1,U} = \frac{1-\sqrt{1-4z_{1,U}}}{2} \quad \text{or} \quad p_{2,U} = \frac{1+\sqrt{1-4z_{1,U}}}{2}, \quad \text{provide} \quad 0 < A_U < 1.$$

Hence, the second and the third expressions of the system (5.16) let us derive the estimators of the parameters p, a and b which are written in expression (5.17).

6.1.9 Derivation of expression (5.25)

When $U = s$ the derivatives of the likelihood function defined by expression (5.24) are as follows:

$$\frac{\partial l(\mathscr{X}_s,a,b,p)}{\partial a} = \frac{1}{a}\sum_{i\in s_0} x_i + \frac{1}{a+b}\sum_{i\in s_1} x_i - n$$

$$\frac{\partial l(\mathscr{X}_s,a,b,p)}{\partial b} = \frac{1}{a+b}\sum_{i\in s_1} x_i - k$$

$$\frac{\partial l(\mathscr{X}_s,a,b,p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p},$$

$$\frac{\partial^2 l(\mathscr{X}_s,a,b,p)}{\partial a^2} = -\frac{1}{a^2}\sum_{i\in s_0} x_i - \frac{1}{(a+b)^2}\sum_{i\in s_1} x_i,$$

$$\frac{\partial^2 l(\mathscr{X}_s,a,b,p)}{\partial b^2} = \frac{\partial^2 l(\mathscr{X}_s,a,b,p)}{\partial a\partial b} = -\frac{1}{(a+b)^2}\sum_{i\in s_1} x_i.$$

$$\frac{\partial l^2(\mathscr{X}_s,a,b,p)}{\partial p^2} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2}, \qquad \frac{\partial l^2(\mathscr{X}_s,a,b,p)}{\partial a\partial p} = \frac{\partial l^2(\mathscr{X}_s,a,b,p)}{\partial b\partial p} = 0.$$

All of the first derivatives of the likelihood function $l(\mathscr{X}_s,\hat{a}_s,\hat{b}_s)$ are equal to zero when $a = \hat{a}_s$ and $b = \hat{b}$ where

$$\hat{a}_s = \bar{X}_{s_0} = \bar{Y}_{s_0}, \qquad \hat{b}_s = \bar{X}_{s_1} - \bar{X}_{s_0} = \bar{X}_{s_1} - \bar{Y}_{s_0}.$$

The Hessian of the likelihood function is negative defined for $(\hat{a}_s,\hat{b}_s)$. Hence, the statistics $\hat{a}_s$ and $\hat{b}_s$ are the maximum likelihood estimators of the parameters a, b and p, respectively.

6.1.10 Derivation of the posterior distribution

On the basis of expressions (5.26)-(5.28) we have the following joint distribution of the likelihood function of the data and the beta and gamma independent prior distributions:

$$q(\mathscr{X}_s,b,p|a,\nu,\theta,\eta,\kappa)=L(\mathscr{X}_s|a,b,p)g(b|\nu,\theta)t(p,\eta,\kappa)=$$
$$=p^k(1-p)^{n-k}t(p,\eta,\kappa)\prod_{i\in s_0}f_0(y_i|a)\prod_{j\in s_1}f_1(x_j|a,b)g(b|\nu,\theta)=$$
$$=\frac{\Gamma(\eta+\kappa)}{\Gamma(\eta)\Gamma(\kappa)}p^{k+\eta-1}(1-p)^{n-k+\kappa-1}\left(\prod_{i\in s_0}f_0(y_i|a)\right)\left(\prod_{i\in s_1}(x_i!)^{-1}\right)$$
$$\frac{\theta^\nu}{\Gamma(\nu)}b^{\nu-1}e^{-ka}e^{-(k+\theta)b}(a+b)^m$$

where $m=\sum_{i\in s_1}x_i$.

Therefore, the posterior distribution of b and p under the fixed $\mathscr{X}_s$ is:

$$q(b,p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)=\frac{q(\mathscr{X}_s,b,p|a,\nu,\theta,\eta,\kappa)}{\int_0^1\int_0^\infty q(\mathscr{X}_s,b,p|a,\nu,\theta,\eta,\kappa)dp\,db}=$$
$$=\frac{p^{k+\eta-1}(1-p)^{n-k+\kappa-1}b^{\nu-1}e^{-(k+\theta)b}(a+b)^m}{\int_0^1 p^{k+\eta-1}(1-p)^{n-k+\kappa-1}dp\int_0^\infty b^{\nu-1}e^{-(k+\theta)b}(a+b)^m db}=$$
$$=\frac{p^{k+\eta-1}(1-p)^{n-k+\kappa-1}b^{\nu-1}e^{-(k+\theta)b}(a+b)^m}{\frac{\Gamma(k+\eta)\Gamma(n-k+\kappa)}{\Gamma(n+\eta+\kappa)}\sum_{h=0}^m\binom{m}{h}a^h\int_0^\infty b^{m-h+\nu-1}e^{-(k+\theta)b}db}=$$
$$=\frac{p^{k+\eta-1}(1-p)^{n-k+\kappa-1}b^{\nu-1}e^{-(k+\theta)b}(a+b)^m}{\frac{\Gamma(k+\eta)\Gamma(n-k+\kappa)}{\Gamma(n+\eta+\kappa)}\sum_{h=0}^m\binom{m}{h}a^e\frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}}$$

Hence,

$$q(b,p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)=q_1(b|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)q_2(p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)\qquad(6.6)$$

where

$$q_1(b|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)=\frac{b^{\nu-1}e^{-(k+\theta)b}(a+b)^m}{\sum_{h=0}^m\binom{m}{h}a^h\frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}}$$

$$q_2(p|\mathscr{X}_s,a,\nu,\theta,\eta,\kappa)=\frac{\Gamma(n+\eta+\kappa)}{\Gamma(k+\eta)\Gamma(n-k+\eta)}p^{k+\eta-1}(1-p)^{n-k+\kappa-1}$$

Hence, expressions (5.30) and (5.31) have been proved.

The derivation of equation (5.35):

$$P(\bar{\tau} = pb \leq \bar{\tau}_0 | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa) = P\left(b \leq \frac{\bar{\tau}_0}{p} | \mathscr{X}_s, a, \nu, \theta, \eta, \kappa\right) =$$

$$= \int_0^1 q_2(p|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) dp \int_0^{\bar{\tau}_0/p} q_1(b|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) db =$$

$$= \int_0^1 q_2(p|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) dp \frac{\sum_{h=0}^m \binom{m}{h} a^h \int_0^{\bar{\tau}_0/p} b^{m+\nu-h-1} e^{-(k+\theta)b} db}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} =$$

$$= \int_0^1 q_2(p|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) dp \frac{\sum_{h=0}^m \binom{m}{h} \frac{a^h}{(k+\theta)^{m+\nu-h-1}} \int_0^{\bar{\tau}_0(k+\theta)/p} u^{m+\nu-h-1} e^{-u} du}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} =$$

$$= \int_0^1 q_2(p|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) dp \frac{\sum_{h=0}^m \binom{m}{h} \frac{a^h}{(k+\theta)^{m+\nu-h-1}} \Gamma(m+\nu-h, \frac{\bar{\tau}_0(k+\theta)}{p})}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} =$$

$$= p^{k+\tau-1}(1-p)^{n-k+\kappa-1}$$

$$\int_0^1 q_2(p|\mathscr{X}_s, a, \nu, \theta, \eta, \kappa) dp \frac{\sum_{e=h}^m \binom{m}{h} \frac{a^h}{(k+\theta)^{m+\nu-h-1}} \Gamma(m+\nu-h, \frac{\bar{\eta}_0(k+\theta)}{p})}{\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}} =$$

$$= \frac{\Gamma(n+\eta+\kappa)}{\Gamma(k+\eta)\Gamma(n-k+\kappa)\sum_{h=0}^m \binom{m}{h} a^h \frac{\Gamma(m+\nu-h)}{(k+\theta)^{m+\nu-h}}}$$

$$\sum_{h=0}^m \binom{m}{h} c_h \int_0^1 \Gamma\left(m+\nu-h, \frac{\bar{\tau}_0(k+\theta)}{p}\right) p^{k+\eta-1}(1-p)^{n-k+\kappa-1} dp$$

where $\Gamma\left(m+\nu-h, \frac{\bar{\tau}_0(k+\theta)}{p}\right)$ is the incomplete gamma function and

$$c_h = \frac{a^h}{(k+\theta)^{m+\nu-h-1}}.$$

6.1.11 Derivation of expression (5.37)

Firstly, let us derive the following variance $V_{mP}(X_U - Y_S) = NV_{mP}(\bar{X}_U - \bar{Y}_S)$,where $V_{mP}(Q)$ means the variance determined on the basis of the model assigned briefly to m and the sampling design $P = P(s)$.

$$V_{mP}(X_U - Y_S) = V_{mP}(N\bar{X}_U - n\bar{Y}_S) = E_m(V_P(N\bar{X}_U - n\bar{Y}_S)) + V_m(E(N\bar{X}_U - n\bar{Y}_S)) =$$

$$= n^2 E_m(V_P(\bar{Y}_S)) + V_m(N\bar{X}_U - n\bar{Y}_U) = \frac{n(N-n)}{N} E_m(V(Y)) + V_m\left(\sum_{k\in U}\left(X_k - \frac{n}{N}Y_k\right)\right) =$$

$$= \frac{n(N-n)}{N}\sigma^2(y) + NV_m\left(X_k - \frac{n}{N}Y_k\right) = \frac{n(N-n)}{N}\sigma^2(y) + N\sigma^2(x-y) +$$

$$+ N\left(1 - \frac{n}{N}\right)^2 \sigma^2(y) = N\sigma^2(x-y) + (N-n)\sigma^2(y).$$

Now it is easy to show that the statistic given by expression (5.37) is an unbiased estimator of the above derived variance $V_{mp}(X_U - Y_S)$.

6.2 Computer programs

6.2.1 Evaluation of sample size and critical value of the test under the assumed risks. Exact solution

The program implementing the solution of system 2.6 is as follows:

#Size of population:
$N < -200$
#Hypothesis H0:
$M0 < -60$
#Hypothesis H1:
$M1 < -100$
#Ryzyka:
$eta < -0.1;\ kappa < -0.2$
$pk < -1;\ n < -1$
while $((pk > kappa)\&(n < N - M0))$
$\{wk < -qhyper(1 - eta, M0, N - M0, n)$
$pk < -phyper(wk, M1, N - M1, n)$
$n < -n + 1\}$
#necessary sample size:
$n < -n - 1;\ n$
#critical value:
wk

6.2.2 Evaluation of sample size and critical value of the test under the assumed risks. Binomial approximation

The program implementing the solution of system 2.8 is as follows:
#Size of population:
$N < -2000$
#Hypothesis H0:
$p0 < -0.2$
#Hypothesis H1:
$p1 < -0.4$
#Risks:
$eta < -0.05;\ kappa < -0.1$
$pk < -1;\ n < -1$
while $((pk > kappa)\&(n < N))$
$\{wk < -qbinom(1-eta, n, p0)$
$pk < -pbinom(wk, n, p1)$
$n < -n+1\}$
#Necessary sample size:
$n < -n-1;\ n$
#Critical value:
wk

6.2.3 Evaluation of sample size and critical value of the test under the assumed risks. Poisson approximation

The program implementing the solution of system 2.9 is as follows:
#the population size:
$N < -19800$
#the hypothesis H0:
$p0 < -0.005$
#the hypothesis H1:
$p1 < -0.02$
#the risks:
$eta < -0.1;\ kappa < -0.05$
$pk < -1;\ n < -1$
while $((pk > kappa)\&(n < N))$
$\{wk < -qpois(1-eta, n*p0)$
$pk < -ppois(wk, n*p1)$
$n < -n+1\}$
#the necessary sample size:
$n < -n-1;\ n$

#the critical value:
wk

6.2.4 Evaluation of sample size and critical value of the test under the assumed risks. Monte Carlo solution

The program implementing the algorithm described in section 2.3.3 is as follows:

For a simple random sample drawn without replacement

```
#the size of population:
  N=10000
  #the start sample size:
  n=1000
  # the increase of sample size:
  dn=1
  # the upper limit of the sample size
  ng=2000
  # the hypothesis H0:
  p0=0.001
  # the hypothesis H1:
  p1=0.002
  # the significance level:
  a=0.1
  # the power of the test:
  b=0.95
  pop0=matrix(0,N,1)
  pop1=matrix(0,N,1)
  for (i in 1:round(p0*N)) pop0[i]=1
  for (i in 1:round(p1*N)) pop1[i]=1
  # number of replications:
  r=100000
  u0=matrix(0,r,1)
  d=c(1,0)
  pp0=c(p0,1-p0)
  u1=matrix(0,r,1)
  pp1=c(p1,1-p1)
  repeat
  for (t in 1:r)
  u0[t]=sum(sample(pop0,n))/n
  u0=sort(u0)
  pk=u0[floor(r*(1-a))+1]
```

```
uk=(pk-p0)/sqrt(p0*(1-p0)/n)
for (t in 1:r)
u1[t]=sum(sample(pop1,n))/n
u1=sort(u1)
bs=0
m=r
while(u1[m] ≥ pk) m=m-1
bs=(r-m)/r
n=n+dn
if ((n≥ng)||(bs≥b)) break
# the necessary sample size:
n
# the critical values:
pk; uk
```

For a simple random sample drawn with replacement

```
# the star size of the sample:
n=1050
# the increase of the sample size:
dn=1
# the upper bound of the sample size:
ng=2000
# the hypothesis H0:
p0=0.01
# the hypothesis H1:
p1=0.02
# the significance level of the test:
a=0.1
# the power of the test:
b=0.95
# the number of the replications:
r=10000
u0=matrix(0,r,1)
d=c(1,0)
pp0=c(p0,1-p0)
u1=matrix(0,r,1)
pp1=c(p1,1-p1)
repeat
for (t in 1:r)
u0[t]=sum(sample(d,n,replace=TRUE,pp0))/n
u0=sort(u0)
pk=u0[floor(r*(1-a))+1]
uk=(pk-p0)/sqrt(p0*(1-p0)/n)
```

```
for (t in 1:r)
u1[t]=sum(sample(d,n,replace=TRUE,pp1))/n
u1=sort(u1)
bs=0
m=r
while(u1[m]≥pk) m=m-1
bs=(r-m)/r
n=n+dn
if ((n≥ng)||(bs≥b)) break
# the sufficient sample size:
n
# the critical values:
pk; uk
```

6.2.5 *Evaluation of sample size to ensure convergence distribution of the statistic to normality*

The program lets us evaluate the necessary sample size to ensure convergence of the studentized statistic to standard normal distribution. The program implementing the algorithm presented in section 3.3.5 is as follows:

Programme evaluating necessary sample size for the chi-square test of goodness of fit under an assumed significance level and power

The critical value of the chi-square goodness of fit test is determined on the basis of its asymptotic distribution

```
# a - significance level, b - power of the test,
# ls - number of simulation,
# dn - increase of sample size n=2000 dn=10 ls=10000 a=0.1 b=0.9
# H0: p=p0; H1: p=p1
k=11
g=c(0.01,0.05,0.1,0.68,0.1,0.05,0.01)
p0=matrix(g,k,1);
g=c(0.405,0.1,0.099,0.088,0.077,0.066,0.055,0.044,0.033,0.022,0.011)
# g=c(0.009,0.045,0.09,0.712,0.09,0.045,0.009)
g=c(0.011,0.055,0.11,0.648,0.11,0.055,0.011)
p1=matrix(g,k,1)
qs=matrix(0,ls,1)
ws=matrix(0,ls,1)
fitchi=function(n,p,w)
n*sum((w-p)*(w-p)/p)}
```

```
fitchi(n,p0,p1)/n
bs=0; it=0
while ((bs < b)&(it <= 10000))
{n=n+dn
for (t in 1:ls) {ws=rmultinom(1,n,p0)/n; qs[t]=fitchi(n,p0,ws)}
qs=sort(qs)
wk=qs[floor((1-a)*ls)+1]
bs=0
for (t in 1:ls)
{ws=rmultinom(1,n,p1)/n;if(fitchi(n,p1,ws) >= wk)bs = bs+1}
bs=bs/ls
it=it+1
}
n; bs; wk;
```

Program evaluating necessary sample size to test normality under an assumed significance level and power

The procedure below tests the hypothesis (defined by expression (3.26)) on the normality distribution of the statistic $z_S(x)$ by means of the test statistic expressed by (3.27) and evaluated based on data observed in a simple random sample drawn with (or without) replacement. Moreover, the program evaluates the sample size "n" of a simple random sample (drawn without replacement) to ensure convergence of the distribution of the studentized sample mean with standard normal distribution. Determining the necessary sample size is done using the chi-square test by successively increasing the size n until the test does not reject the hypothesis on normality. The sample size "ls" of the test is determined separately on the basis of Monte Carlo experiments under an assumed significance level and power of the test, e. g. on the basis of the program: $n_c hi_l imit.txt$

```
# reading of data:
x=read.table("data.txt"); x=as.matrix(x)
#population size:
N=nrow(x)
# p - significance level:
pw=0.1
# size of sample:
n0=100; n=n0
# number of simulation:
lp=10000
#increase of the sample size:
dp=10
# number of the sample replication;
ls=5870
```

```
t=matrix(0,ls,1)
x1=mean(x); sq=sqrt(var(x))
H_0 : ω = pist where
pist=as.matrix(c(0.01,0.05,0.1,0.9,0.95,0.99))
a=nrow(pist)
pt=matrix(0,a+1,1)
z=matrix(0,a,1)
nr=matrix(0,n,1)
z=qnorm(pist)
pt[1]¡-pist[1]
if (a > 1) for (i in 2:a) pt[i]=pist[i]-pist[i-1]
pt[a+1]=1-pist[a]
if(n0 > n) n=n-n0
ilp=0;pws=0
while ((ilp <= lp)&(pws <= pw)){
n=n+dp
nq=sqrt(n)
nn=matrix(0,a+1,1)
for (i in 1:ls) {
# in the case of sampling without replacement:
nr=sample(1:N,n)
# in the case of sampling with replacement:
nr=sample(1:N,n,replace=TRUE)
sq=sqrt(var(x[nr]))
if (sq > 0) t[i]=(mean(x[nr])-x1)*nq/sq
if (t[i] < z[1]) nn[1]¡-nn[1]+1
else if (t[i] <= z[2]) nn[2]=nn[2]+1
else if (t[i] <= z[3]) nn[3]=nn[3]+1
else if (t[i] <= z[4]) nn[4]=nn[4]+1
else if (t[i] <= z[5]) nn[5]=nn[5]+1
else if (t[i] <= z[6]) nn[6]=nn[6]+1
else nn[7]=nn[7]+1
}
wt=chisq.test(nn,p=pt);
pws=wt$p.value
ilp=ilp+1
}
n
```

6.2.6 Evaluation of inclusion probabilities proportional to auxiliary variable values

The program lets us evaluate the first-order inclusion probabilities proportional to the values of a positive auxiliary variable. The program implementing the algorithm presented in subsection 4.1.2 is as follows:

```
dx=c(1,2,4,10,100,4,150,3,5,6,9)
dx=as.matrix(dx)
inkl1=function(n,dx)

n1=n
dx1=dx
N=nrow(dx)
pi1=matrix(0,ncol=1,nrow=N)
sp=sum(dx)
pi1=dx*(n/sp)
pi1
km=1
m=pi1[1]
for (k in 2:N) if (pi1[k]>m) {m=pi1[k]; km=k}
if (m>1) {pi1[km]=1;dx1[km]=0}
while (m>1)
{
n1=n1-1
sp=0
for (k in 1:N) sp=sp+dx1[k]
for (k in 1:N) if (dx1[k]>0) pi1[k]=dx1[k]*(n1/sp)
m=pi1[1]
for (k in 2:N) if (pi1[k]>m) {m=pi1[k]; km=k}
if (m>1) {pi1[km]=1;dx1[km]=0}
}
pi1
}
a=inkl1(3,dx)
a
```

6.2.7 Hartley-Rao sampling scheme

The program implementing the Hartley-Rao sampling scheme (explained in subsection 4.1.4) is as follows.

```
losHartleyRao=function(n,N,pi1)
{skum=matrix(0,ncol=1,nrow=N+1)
```

```
skum[2]=pi1[1]
for (k in 2:N) skum[k+1]=skum[k]+pi1[k]
u=runif(1)
snmatrix(0,ncol=1,nrow=n)
z=1
while (z≤n)
{
k=1;q=0
while (q==0)
{
if ((skum[k]<u+z-1)&(u+z-1≤skum[k+1])) {q=1; sn[z]=k}
k=k+1
}
z=z+1
};sn}
```

6.2.8 *Evaluation of inclusion probabilities of the sampling design proportionate to the function of one quantile*

The program below let us calculate the first- and second-order inclusion probabilities. The inclusion probabilities characterize the conditional sampling design proportionate to the r-th order statistic of a positively valued auxiliary variable. These inclusion probabilities are defined in subsection 4.1.8 in theorems 4.4 and 4.5.

```
inkluzje1i2=function(n,N,r,u,v,x) {
#Inclusion probabilities of degree 1 and 2 for the sampling design proportional
#to one order statistics of the auxilairy variable x
#inpute: n - sample size
#N - population size
#r - degrees of the order statistic
#u, v - constrains: u ≥ r and u ≤ v; v ≤ N − n + r and v ≥ u
#x -vector of the positively valued observations of the auxiliary variable
# in the population
#output: the matrix (of degree N) of inclussion probabilities
#the diagonal element of the matrix are equal to inclusion probabailities
# of the first degree
Newton=function(N,n) {if ((N==n)—(n==0))    b=1    else
{b=1;for (i in 1:n) b=(b/(n-i+1))*(N-i+1)};b}
z=0; for (i in u:v) z=z+x[i]*Newton(i-1,r-1)*Newton(N-i,n-r)
pi=array(0,dim=c(N,N))
# inkl. degree 1
if ((r-1>0)&(v-1>0)&&(u-1>0)) {a=0; for (i in u:v)
a=a+Newton(i-2,r-2)*Newton(N-i,n-r)*x[i];
```

```
for (i in 1:(u-1)) pi[i,i]=a};
for (k in u:v)
{pi[k,k]=Newton(k-1,r-1)*Newton(N-k,n-r)*x[k];
if ((n-r>0)&&(k-u>0)&&(k-1>0)) for (i in u:(k-1))
pi[k,k]=pi[k,k]+Newton(i-1,r-1)*Newton(N-i-1,n-r-1)*x[i];
if ((r-1>0)&&(v-k>0)) for (i in (k+1):v)
pi[k,k]=pi[k,k]+Newton(i-2,r-2)*Newton(N-i,n-r)*x[i]};
for (k in u:v) pi[k,k]=pi[k,k];
if ((n-r>0)&&(N-v>0)) {a=0;
for (i in u:v) a=a+Newton(i-1,r-1)*Newton(N-i-1,n-r-1)*x[i];
for (i in (v+1):N) pi[i,i]=a};
```

6.2.9 Sampling scheme of the sampling design proportionate to the function of one quantile

The program below implements the conditional sampling design proportionate to the r-th order statistic of a positively valued auxiliary variable. The sampling scheme is defined in subchapter 4.1.8.

```
losq=function(n,N,r,u,v,x) {
#Drawing the sample according to thesampling design proportional
#to one order statistics of the auxilairy variable x
#inpute: n - sample size
#N - population size
#r - degrees of the order statistic
#u, v - constrains: u ≥ r and u ≤ v; v ≤ N − n + r and v ≥ u,
#x -vector of the positively valued observations of the auxiliary variable
# in the population
#output: the vector with the selected (and sorted) number of the population elements
Newton=function(N,n) {if ((N == n)|(n == 0)) b=1 else {b=1;
for (i in 1:n) b=(b/(n-i+1))*(N-i+1)};b}
for (i in u:v) x[i]=x[i]*Newton(i-1,r-1)*Newton(N-i,n-r);
a=0;for (i in u:v) a=a+x[i];p=1:(v-u+1);for (i in u:v) p[i-u+1]=x[i]/a;x1=u:v;
nx=matrix(0,1,n);if (u==v) nx[r]=u else nx[r]=sample(x1,1,replace=TRUE,p)
x1=1:(nx[r]-1);
if (r − 1 > 0) {s1=sample(x1,r-1);for (i in 1:(r-1)) nx[i]=s1[i]};
x1=(nx[r]+1):N;if (n − r > 0) {s1=sample(x1,n-r);
for (i in (r+1):n) nx[i]=s1[i-r]};nx=sort(nx);nx}
```

6.2.10 Evaluation of first-order inclusion probabilities for a conditional sampling design dependent on two order statistics

The program below lets us calculate first-order inclusion probabilities. The inclusion probabilities characterize the conditional sampling design proportionate to the positive function of the r-th and u-th order statistics of a positively valued auxiliary variable. The inclusion probabilities are expressed in subsection 4.1.9 by means of theorem 4.6.

```
inkluzje1fsum2kw=function(n,N,r,u,x,d12) {
#Inclusion probabilities of degree 1 for the conditional sampling design proportional
#to the positive function of two order statistics of the auxilairy variable x
#inpute: n - sample size
#N - population size
#r < u- degrees of the order statistic
#x -vector of the positively valued observations of the auxiliary variable in the population
#d12 - value defining the condition
#output: the vector (of size N) of inclussion probabilities of the first degree
h=function(x[i],x[j]){
#definition of the function; for instance:
d=x[j]+x[i];d}
warunek=function(mx,d,d12){
#the function testing the condition
a=(d>d12);a}
#for instance another function testing condition:
warunek=function(x,i,j,d1,d2){a=(x[i]≤d1)&(x[j]>d2);a}
z=0;mx=mean(x)
for (i in r:(N-n+r)) for (j in (i+u-r):(N-n+u))
{d=h(x[i],x[j]);
if (warunek(mx,d,d12))
z=z+Newton(i-1,r-1)*Newton(j-i-1,u-r-1)*Newton(N-j,n-u)*d}; pi=matrix(0,1,N+1)
for (k in 1:N) {
if ((k<r)&(r>1)) for (i in r:(N-n+r)) for (j in (i+u-r):(N-n+u))
{d=h(x[j],x[i]); if (warunek(mx,d,d12))
pi[k]=pi[k]+Newton(i-2,r-2)*Newton(j-i-1,u-r-1)*Newton(N-j,n-u)*d};
if ((k>r-1)&(k<N-n+u+1)) {
if ((k>u)&(n>u)) for (i in r:(k-u+r-1)) for (j in (i+u-r):(k-1)) {d=h(x[j],x[i]);
if (warunek(mx,d,d12))
pi[k]=pi[k]+Newton(i-1,r-1)*Newton(j-i-1,u-r-1)*Newton(N-j-1,n-u-1)*d};
if(k>u-1) {a=0; for (i in r:(k-u+r)) {d=h(x[k],x[i]);
if (warunek(mx,d,d12)) a=a+Newton(i-1,r-1)*Newton(k-i-1,u-r-1)*d};
pi[k]=pi[k]+a*Newton(N-k,n-u)};
```

```
if ((k>r)&(u>r+1)&(N-n+u>k)) for (i in r:(k-1)) for (j in (k+1):(N-n+u))
{d=h(x[j],x[i]); if (warunek(mx,d,d12))
pi[k]=pi[k]+Newton(i-1,r-1)*Newton(j-i-2,u-r-2)*Newton(N-j,n-u)*d}
if (k<N-n+r+1) {a=0; for (j in (k+u-r):(N-n+u)) {d=h(x[j],x[k]); if (warunek(mx,d,d12))
a=a+Newton(j-k-1,u-r-1)*Newton(N-j,n-u)*d};pi[k]=pi[k]+a*Newton(k-1,r-1)};
if ((r>1)&(k<N-n+r)) for (i in (k+1):(N-n+r)) for (j in (i+u-r):(N-n+u))
{d=h(x[j],x[i]); if (warunek(mx,d,d12))
pi[k]=pi[k]+Newton(i-2,r-2)*Newton(j-i-1,u-r-1)*Newton(N-j,n-u)*d}};
if ((k>N-n+u)&(n>u)) for (i in r:(N-n+r)) for (j in (i+u-r):(N-n+u))
{d=h(x[j],x[i]); if (warunek(mx,d,d12))
pi[k]=pi[k]+Newton(i-1,r-1)*Newton(j-i-1,u-r-1)*Newton(N-j-1,n-u-1)*d} }
for (i in 1:N) pi[i]=pi[i]/z; pi }
```

6.2.11 *Implementing a sampling scheme of a conditional sampling design proportionate to the function of two order statistics*

The program below implements a conditional sampling design proportionate to two order statistics of a positively valued auxiliary variable. The sampling scheme is explained in subchapter 4.1.9.

```
sum2kwschemlos=function(n,N,r1,r2,x,d12) {
# Random selecting the sample according to the condit. sampling design
# proportional to the positive function of the two order statistics
# of the auxilairy variable x
# inpute: n - sample size
# N - population size
# r1 < r2- degrees of the order statistic
# x -vector of the positively valued observations of the auxiliary variable
# in the population
# d12 - value defining the condition
# output: the vector with the selected number of the population elements
h=function(x[i],x[j]){
# definition of the function; for instance: d=x[j]+x[i];d}
warunek=function(mx,d,d12){ #the function testing the condition a=(d>d12);a}
p1=array(0,dim=c(N,1));
p2=array(0,dim=c(N,1));
for (i in r1:(N-n+r1)) for (j in (i+r2-r1):(N-n+r2)) {d=h(x[i],x[j]);
if (warunek(mx,d,d12))
p1[i]=p1[i]+Newton(i-1,r1-1)*Newton(j-i-1,r2-r1-1)*Newton(N-j,n-r2)*d};
z=0;for (i in r1:(N-n+r1)) z=z+p1[i]; p1=p1/z; ns=array(0,dim=c(n,1));
ns[r1]=sample(N,1,replace=FALSE,p1);
for (j in (ns[r1]+1):(N-n+r2)) {d=h(x[ns[r1]],x[j]) if (warunek(mx,d,d12))
p2[j]=Newton(ns[r1]-1,r1-1)*Newton(j-ns[r1]-1,r2-r1-1)*Newton(N-j,n-r2)*d};
```

```
p2=p2/p1[ns[r1]]; ns[r2]=sample(N,1,replace=FALSE,p2);
if (r1 > 1) {s1=sample(1:(ns[r1]-1),r1-1); for (i in 1:(r1-1)) ns[i]=s1[i]}
if (r2 > r1 + 1) {s2=sample((ns[r1]+1):(ns[r2]-1),r2-r1-1);
for (i in 1:(r2-r1-1)) ns[r1+i]=s2[i]} if (n > r2) {s3=sample((ns[r2]+1):N,n-r2);
for (i in 1:(n-r2)) ns[r2+i]=s3[i]} ns=sort(ns);ns }
```

6.2.12 Evaluation of first-order inclusion probabilities for a conditional sampling design dependent on three order statistics

The program below lets us calculate first-order inclusion probabilities. The inclusion probabilities characterize the conditional sampling design proportionate to the positive function of three order statistics of a positively valued auxiliary variable. The program is prepared on the basis of expressions presented in the book by Wywiał (2015).

```
inkluzje1YK=function(n,N,r1,r2,r3,x,d1,d2,d21) {
# Inclusion probabilities of degree 1 for the conditional sampling
# design proportional
# to the positive function of three order statistics of the auxilairy
# variable x
# inpute: n - sample size
# N - population size
# r1¡r2¡r3- degrees of the order statistic
# x -vector of the positively valued observations of the auxiliary
# variable in the population
# d1,d2,d12 - value defining the condition
# output: the vector (of size N) of inclusion probabilities
# of the first degree

Newton=function(N,n) {b=0; if ((N>=n)&(n>=0)&(N>=0)) {b=1;
if ((N>n)&(n>0)) for (i in 1:n) b=(b/(n-i+1))*(N-i+1)};b}

funkcja=function(i1,i2,i3,x){
d=(x[i1]+x[i2]+x[i3])/3
# d=var(c(x[i1],x[i2],x[i3])); d}

warun=function(x,i1,i2,i3,d1,d2){a=((x[i1]<=d1)&(x[i3]>d2));a}
# for example:
# warun=function(x,mx,d,d12,a1,i1,a2,b1,i2,b2,c1,i3,c2){
# a=(abs(d-mx)<=d12);
# b=((x[i1]>=a1)&(x[i1]<=a2)&(x[i2]>b1)&
# (x[i2]<=b2)&(x[i3]>c1)&(x[i3]<=c2));a&b}
```

```
z=0;for (i1 in r1:(N-n+r1)) for (i2 in (i1+r2-r1):(N-n+r2)) for (i3 in (i2+r3-r2):(N-
n+r3))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x);
if (d>=d21) z=z+Newton(i1-1,r1-1)*
Newton(i2-i1-1,r2-r1-1)*Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d};
pi=matrix(0,N,1); mx=mean(x)
for (k in 1:N)
{
if ((k<r1)&(r1>1)) for (i1 in r1:(N-n+r1)) for (i2 in (i1+r2-r1):(N-n+r2))
for (i3 in (i2+r3-r2):(N-n+r3)) if (warun(x,i1,i2,i3,d1,d2))
{d=funkcja(i1,i2,i3,x); if (d>=d21)
pi[k]=pi[k]+Newton(i1-2,r1-2)*Newton(i2-i1-1,r2-r1-1)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d}
if ((k>r1-1)&(N-n+r1>k)&(r1>1)) for (i1 in (k+1):(N-n+r1))
for (i2 in (i1+r2-r1):(N-n+r2)) for (i3 in (i2+r3-r2):(N-n+r3))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x); if (d>=d21)
pi[k]=pi[k]+Newton(i1-2,r1-2)*Newton(i2-i1-1,r2-r1-1)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d}

if ((k>r1)&(r2-r1>1)&(N-n+r2>k)) for (i1 in r1:(k-1))
for (i2 in (k+1):(N-n+r2)) for (i3 in (i2+r3-r2):(N-n+r3))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x);
if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*Newton(i2-i1-2,r2-r1-2)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d}

if ((k>r2)&(r3-r2>1)&(N-n+r3>k)) for (i1 in r1:(k-r2+r1-1))
for (i2 in (i1+r2-r1):(k-1)) for (i3 in (k+1):(N-n+r3))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x);
if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*Newton(i2-i1-1,r2-r1-1)*
Newton(i3-i2-2,r3-r2-2)*Newton(N-i3,n-r3)*d}

if ((n>r3)&(N-n+r3+1>k)&(k>r3)) for (i1 in r1:(k+r1-r3-1))
for (i2 in (i1+r2-r1):(k+r2-r3-1)) for (i3 in (i2+r3-r2):(k-1))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x);
if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*Newton(i2-i1-1,r2-r1-1)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3-1,n-r3-1)*d}

if ((n>r3)&(k>N-n+r3)) for (i1 in r1:(N-n+r1))
for (i2 in (i1+r2-r1):(N-n+r2)) for (i3 in (i2+r3-r2):(N-n+r3))
if (warun(x,i1,i2,i3,d1,d2)) {d=funkcja(i1,i2,i3,x);
if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*Newton(i2-i1-1,r2-r1-1)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3-1,n-r3-1)*d}

if ((N-n+r1+1>k)&(k>r1-1)) for (i2 in (k+r2-r1):(N-n+r2))
for (i3 in (i2+r3-r2):(N-n+r3)) if (warun(x,k,i2,i3,d1,d2))
```

```
{d=funkcja(k,i2,i3,x); if (d>=d21) pi[k]=pi[k]+Newton(k-1,r1-1)*
Newton(i2-k-1,r2-r1-1)*Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d}

if ((N-n+r2+1>k)&(k>r2-1)) for (i1 in r1:(k-1))
for (i3 in (k+r3-r2):(N-n+r3)) if (warun(x,i1,k,i3,d1,d2))
{d=funkcja(i1,k,i3,x);
if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*Newton(k-i1-1,r2-r1-1)*
Newton(i3-k-1,r3-r2-1)*Newton(N-i3,n-r3)*d}

if ((N-n+r3+1>k)&(k>r3-1)) for (i1 in r1:(N-n+r1))
for (i2 in (i1+r2-r1):(N-n+r2)) if (warun(x,i1,i2,k,d1,d2))
{d=funkcja(i1,i2,k,x); if (d>=d21) pi[k]=pi[k]+Newton(i1-1,r1-1)*
Newton(i2-i1-1,r2-r1-1)*Newton(k-i2-1,r3-r2-1)*Newton(N-k,n-r3)*d}}
pi=pi/z; pi}
```

6.2.13 Implementing a sampling scheme of a conditional sampling design proportionate to the function of three order statistics

The program below implements a conditional sampling design proportionate to the positive function of three order statistics of an auxiliary variable. The program is prepared on the basis of the algorithm presented in the book by Wywiał (2015).

```
schemlosYK=function(n,N,r1,r2,r3,x,d1,d2,d21,p,z) {
# Random selecting the sample according to the conditional
# sampling design proportional to the positive function of three
# order statistics of the auxilairy variable x
# inpute: n - sample size
# N - population size
# r1¡r2- degrees of the order statistic
# x -vector of the positively valued observations of the auxiliary variable
# in the population
# d12 - value defining the condition
# output: the vector with the selected number of the population
# elements

# Newton=function(N,n) {b=0; if ((N>=n)&(n>=0)&(N>=0))
# {b=1; if ((N>n)&(n>0)) for (i in 1:n) b=(b/(n-i+1))*(N-i+1)};b}

funkcja=function(i1,i2,i3,x){(x[i1]+x[i2]+x[i3])/3}

warun=function(x,i1,i2,i3,d1,d2){a=((x[i1]<=d1)&(x[i3]>d2));a}
#For instance:
# warun=function(x,mx,d,d12,a1,i1,a2,b1,i2,b2,c1,i3,c2)
```

```
#{a=(abs(d-mx)<=d12); b=((x[i1]>=a1)&(x[i1]<=a2)
#&(x[i2]>b1)&(x[i2]<=b2)&(x[i3]>c1)&(x[i3]<=c2));a&b}

s=matrix(0,n,1)
p=matrix(0,N,1)
for (i1 in r1:(N-n+r1)) for (i2 in (i1+r2-r1):(N-n+r2))
for (i3 in (i2+r3-r2):(N-n+r3)) if (warun(x,i1,i2,i3,d1,d2))
{d=funkcja(i1,i2,i3,x);
if (d>=d21) p[i1]=p[i1]+Newton(i1-1,r1-1)*Newton(i2-i1-1,r2-r1-1)
*Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d};
z=0;for (i1 in r1:(N-n+r1)) z=z+p[i1]; p=p/z
t1=sample(1:N,1,replace=FALSE,p)
a1=p[t1]*z
p=matrix(0,N,1)
for (i2 in (t1+r2-r1):(N-n+r2)) for (i3 in (i2+r3-r2):(N-n+r3))
if (warun(x,t1,i2,i3,d1,d2)) {d=funkcja(t1,i2,i3,x);
if (d>=d21) p[i2]=p[i2]+Newton(t1-1,r1-1)*Newton(i2-t1-1,r2-r1-1)*
Newton(i3-i2-1,r3-r2-1)*Newton(N-i3,n-r3)*d};
p=p/a1
t2=sample(1:N,1,replace=TRUE,p)
a2=p[t2]*a1
p=matrix(0,N,1)
for (i3 in (t2+1):(N-n+r3)) if (warun(x,t1,t2,i3,d1,d2))

{d=funkcja(t1,t2,i3,x); if (d>=d21)
p[i3]=p[i3]+Newton(t1-1,r1-1)*Newton(t2-t1-1,r2-r1-1)*
Newton(i3-t2-1,r3-r2-1)*Newton(N-i3,n-r3)*d};
p=p/a2
t3=sample(1:N,1,replace=TRUE,p)
s[r1]=t1;s[r2]=t2;s[r3]=t3;
if (r1>1) s[1:(r1-1)]=sample(1:(t1-1),r1-1,replace=FALSE)
if (r2-r1>1) if(t1+1==t2-1) s[r1+1]=t1+1 else
s[(r1+1):(r2-1)]=sample((t1+1):(t2-1),r2-r1-1,replace=FALSE)
if (r3-r2>1) if((t2+1)==(t3-1)) s[r2+1]=t2+1 else
s[(r2+1):(r3-1)]=sample((t2+1):(t3-1),r3-r2-1,replace=FALSE)
if (r3<n) if(t3+1==N) s[n]=N else
s[(r3+1):n]=sample((t3+1):N,n-r3,replace=FALSE); s }
```

Chapter 7
Bibliography

Agresti A, Coull B. A. (1998). Approximate is beteer than "exact" for interval estimation of binomial proportion. *American Statistician*, vol. 54, 280-288.

Antal E., Tillé Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association* vol. 106, 534-543.

Ardilly P., Tillé Y. (2005). *Sampling Methods: Exercises and Solutions*. Springer.

Arens A.A., Loebbecke J.K. (1981). *Application of Statistical Sampling to Auditings*. Prentice-Hall, Englewood Cliff.

Arnold B.C., Balakrishnan N., Nagaraja H.N. (2008). *A First Course in order Statistics*. SIAM, Philadelphia.

Artifcial Intelligence in Accounting and Auditing. M.A. Vasarhelyi (Ed.). (1995) vol.3, Markus Wiener Publishers, Princeton.

Babu G. J., Singh K. (1984). On one-therm Edgeworth corection by Efron's bootstrap. *Sankhya A*, vol. 46, 219-232.

Babu G. J., Singh K. (1985). Edgeworth expansions for sampling without replacement from finite population. *Journal of Multivariate Analysis* vol. 17, 261-278.

Barbiero A., Mecatti F. (2010). Bootstrap algorithms for variance estimation in PS sampling. In: *Complex Data Modeling and Computationally Intensive Statistical Methods*. Edited by Mantovan P. and Secchi P. Springer-Verlag Italia, pp. 57-70.

Beck P. J., Solomon I. (1985). Sampling risks and audit consequences under alternative testing approaches. *The Accounting Review*, vol. LX, 4.

Bellhouse D. R. (2001). The central limit theorem under simple random sampling. *The American Statistician* vol. 55, 352-357.

Berger, Y.G. (1998). Rate of convergence to normal distribution for Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* 67, 209-226.

Berger Y. G., Skinner C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Jouranl of the Royal Statistical Society* series B, vol. 67, 79-89.

Brewer K.R.W.,Hanif M. (1983). *Sampling with Unequal Probabilities.* Springer Verlag, New York-Heidelberg-Berlin.

Brown L. D., Cai T. T., Das Gupta A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics*, vol. 30, 160-201.

Bühler W., Deutler T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Campbell C. (1980). A different view of finite population estimation *Proceedings of Survey Methods Section of American Statistical Association*, 319-324.

Cassel C.M., Särndal C.E., Wretman J.W. (1977). *Foundation of Inference in Survey Sampling*. John Wiley & Sons, New York, London, Sydney, Toronto.

Chao M. T., Lo A. Y. (1985). A bootstrap method for finite population. *Sankhya* vol. A47, 399-405.

Chaudhuri A., Stenger H. (2005). *Survey Sampling. Theory and Methods*, Second Edition, Chapman & Hall CRC, Boca Raton-London-New York-Singapore.

Chaudhuri A., Vos J.W.E. (1988). *Unified Theory of Survey Sampling.* North Holland, Amster-dam-New York-Oxford-Tokyo.

Chauvet G. (2007). Méthodes de bootsrap en population finie. PhD Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Université de Rennes 2. Available at http://tel.archives-ouvertes.fr/docs/00/26/76/89/PDF/thesechauvet.pdf

Chen H. (1990). The Accuracy of approximate intervals for binomial parameter. *Journal of the American Statistical Association*, vol. 85, 514-518.

Chen J., Chen S. Y., Rao J. N. K. (1998). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics* vol. 31, no. 1, pp. 53-68.

Chernick M. R., Liu C. Y. (2002). The saw-toothed behavior of the power versus sample and software solutions: Single binomial proportion using exact methods. *The American Statistician* vol. 56, pp. 149-155.

Colopper C. J., and Pearson E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* vol. 26, 404-413.

Cochran W.G. (1952). The chi-square test of goodness of fit. *Annals of Mathematical Statistics* vol. 23, 315-345.

Cochran W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, vol. 38, no. 2, Tokyo, 345-358.

Cochran W.G. (1977). *Sampling techniques.* John Wiley & Sons, New York.

Copas J. B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika* 59, 349-360.

Cordy C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters* 18, 353-362.

Cox D. R., Snell E. J. (1979). On sampling and estimation of rare errors. *Biometrika* 69, 1, 125-132.

Cramér H. (1946). *Mathematical Methods of Statistics*. Princenton University Press.

Dalenius T. (1950). The problem of optimum stratification. *Scandinavisk Aktuarietidskrift* vol. 3-4, 203-213.

Dalenius T. (1957). *Sampling in Sweden. Contribution to the Methods and Theories of Sample Survey Practice.* Almqvist and Wicksell, Stockholm.

Dalenius T., Hodges J. L. (1959). Minimum variance stratification. *Journal of the American Statistical Association* vol. 54, no. 285, 88-101.

Deville J. C. (1993). Estimation de la variance pur les enquêtes en deux phases. Manuscript INSEE.

Deville J. C., Tillé Y. (2005). Variance approximation under balanced sampling *Journal of Statistical Planning and Inference* vol. 128, 411-425.

Domański Cz., Pruska K. (2001). *Metody statystyki małych obszarów*. Wydawnictwo Uniwersytetu Łódzkiego.

Domański Cz., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014). Testy statystyczne w procesie podejmowania decyzji. Wydawnictowo Uniwersytetu Łódzkiego.

Edgeworth F. Y. (1907). On the representation of a statistical frequency by a series. *Journal of the Royal Statistical Society* vol. A 70, 102-106.

Efron B. (1979). Bootstrap methods. Another look at the jackknife. *Annals of Statistics* vol. 1, 1-26.

Erdös P., Rényi A. (1959). On the central limit theorem from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, Trudy Publications, vol. 4, 49-57.

Fisz M. (1967). *Rachunek prawdopodobieństwa i statystyka matematyczna* (in Polish). PWN, Warszawa.

Fienberg S.E., Neter J., Leith R. A. (1997). Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, vol. 72, no. 358, 295-302.

Fuller W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Gerstenkorn T., Śródka T. (1974). *Kombinatoryka i rachunek prawdopodobieństwa* (in Polish). PWN Warszawa.

Gamrot W. (2014). *Estymacja wartości przeciętnej uwzględniająca koszt pozyskania danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

Ghosh M., Meeden G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall.

Ghosh M., Rao J. N. K. (1994). Small area estimation: An appraisal. Statistical Science, vpl. 9, 55-93.

Griffiths B., Krutchokoff R. (1971). Optimal linear estimates: An empirical Bayes version with application to the binomial distribution. *Biometrika* 58, 195-201.

Gross S. (1980). Median estimation in sample surveys. *Proceedings of Section on Survey Research Methods* American Statistical Association, Washington, 181-184.

Guilford J. P. (1971). *Fundamental Statistics in Psychology and Education*. McGraw-Hill, New York.

Guy, D. M., & Carmichael, D. R. (1986). *Audit sampling: An introduction to statistical sampling in auditing*. New York: John Wiley & Sons Inc.

Hájek J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, Trudy Publications, vol. 5, 361-374.

Hájek J. (1964). Asymptotic theory of of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* vol. 35, 1491-1523.

Hájek J. (1981). *Sampling from a Finite Population*. Edit. by V. Dupač. Marcel Dekker, Inc. New York and Basel.

Hall P. (1992). *The Bootstrap and Edgewrth Expansion*. Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

Hartley H.O., Rao J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics* 33, 350-374.

Helmers R. (2000). Inference on rare errors using asymptotic expansions and bootstrap calibration. *Biometrika* 87, 689-694.

Hess I., Sethi V. K., Balakrishnan T. R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association* vol. 61, no. 313, 74-90.

Holmberg, A. (1998). A bootstrap approach to probability proportional to size sampling. In *Proceedings of Section on Survey Research Methods*, American Statistical Association, 378-383.

Hołda A., Pociecha J. (2004). *Rewizja finansowa* (in Polish). Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

Hołda A., Pociecha J. (2009). *Probabilistyczne metody badania sprawozdań finansowych* (in Polish). Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Horvitz, D., G., Thompson, D., J. (1952). A generalization of the sampling without replacement from finite universe. *Journal of the American Statistical Association* vol. 47, 663-685.

Ijiri Y., and Kaplan R. S. (1971). A model for integrating sampling objectives in auditing. *Journal of Accounting Research* vol. 9, no, 1, 73-87.

International Encyclopedia of Statistical Science. Edt. M. Lovric. (2011). Springer.

Jeffreys H. (1961). *Theory of Probability*. Oxford University Press, Oxford, UK.

Johnson N. L., Kotz S., Kemp A. W. (1992). *Univariate Discrete Distributions*. John Wiley, New York.

Jowett D. H. (1963). The relationship between the binomial and the F distributions. *The American Statistician* vol. 31(1), 55-57.

Kaplan R. S. (1973). A stochastic model fo auditing. *Journal of Accounting Research*, No 11, 38-46.

Karliński W. (2005). *Dobór próby w audycie* (in Polish). Instytut Rachunkowości i Podatków Warszawa.

Kass R. E., Raftery A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, vol. 90, 773-795.

Khan M. G. M., Nand N., Ahmad N. (2008). Determining the optimum strata boundary points using dynamic programming. Survey Methodology 34 (2), 205-214.

Kish L. (1965). *Survey sampling.* John Wiley & Sons, Inc. New York- London-Sydney.

Klima D. (2005). *Statystyka dla audytorów*. InfoAudit Warszawa.

Konijn H.S. (1973). *Statistical theory of sample survey and analysis.* North-Holland Publishing Company, Inc., Amsterdam-London, American Elsevier Publishing Company, Inc., New York.

Kozak M. (2004). Modyfikacja metody Daleniusa i Hodgesa wyznaczania przybliżonych granic warstw. *Wiadomości Statystyczne*, 10, 10-16.

Kozak M. (2004a). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806.

Kozak M. (2011). Comparison of efficiency of geometric stratification and k-means algorithm in univariate stratification of skewed population. *International Journal of Agricultural and Statistical Sciences* **7**(1), 341-344.

Krzyśko M. (2000). *Wykłady z teorii prawdopodobieństwa*. WNT, Warszawa.

Krzyśko M. (2004). *Statystyka matematyczna*. Wydawnictwo Naukowe UAM, Poznań.

Kuk A. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size. *Journal of Statistical Computation and Simulation* 31, 73-82.

Kvanli A. H., Shen Y. K., Deng L. Y. (1998). Construction of confidence intervals for the mean of a population containing many zero values. *Journal of Business and Economic Statistics* vol. 16, pp. 362-368.

Lahiri D.B. (1951). A method of sample selection providing unbiased ratio estimator. *Bulletin of the International Statistical Institute*, vol. **33**, 2, pp. 133-140.

Lalu N. M., Krishnan P. (1978). Sequential procedure for estimating the sample size needed for normal approximation in finite population sampling. *Proceedings of the Survey Research Method* American Statistical Association, 621-623.

Lavallée P., Hidiroglou M. (1988). On the stratification of skewed population. *Survey Methodology*, 14, 33-43.

Lednicki B., Wieczorkowski R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

Leslie D. A., Teitlebaum A. D., Anderson R. J. (1979). *Dollar Unit Sampling: a Practical Guide for Auditors*. Copp Clark Pitman.

Lodewyckx T., Kim W., Lee M. D., Tuerlinckx F., Kuppens P., Wagenmakers E. J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, vol. 55, 331-347.

Madow W. G. (1948). On the limiting distribution of estimates based on samples from finite universes. *Annals of Mathematical Statistics* vol. 19, 535-545.

McCray J. H. (1984). A quasi-Bayesian audit risk model for dollar unit sampling. *The Accounting Review* vol. 59, No. 1, 35-51.

Meeden G. (2003). A Bayesian solution for statistical auditing problem. *Journal of the American Statistical Association* vol. 98, 735-740.

Midzuno H. (1952). On the sampling system with probability proportional to sum of sizes *Annals of the Institute of Matematics and Statistisc* 3, 99-107.

Neter J., and Loebbecke J. K. (1975). *Behavior of Major Statistical Estimators in Sampling Accounting Populations: An Empirical Study*. New York: American Institute of Certified Public Accountants.

Neyman J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, vol. 97, 558-606.

Niemiro W. (1999). Konstrukcja optymalnej stratyfikacji metodą poszukiwań losowych. *Wiadomości Statystyczne*, 10, 1-9.

Patel P. A., Patel J. S. (2010). A Monte Carlo comparison of some variance estimators of the HorvitzThompson estimator. *Journal of Statistical Computation and Simulation*. vol. 80, no. 5, pp. 489-502.

Pathak K. (1976). Unbiased estimation in fixed-cost sequential sampling schemes. *Annals of Statistics* vol. 4, no. 5, pp. 1012-1017.

Pekasiewicz D. (2010). Sequential method for estimating the sample size required for testing hypotheses on the population mean. *Acta Universitatis Lodziensis. Folia Oeconomica* No. 235. Wydawnictwo Uniwersytetu Łódzkiego, 99-108.

Pfefferman D. (2001). Small area estimation - New developments and directions. *International Statistical Review*, vol. 79, 125-143.

Pfefferman D. (2013). New important developments in small area estimation. *Statistical Science*, vol. 28 (1), 40-68.

Prášková Z. (1982). Rate of convergence for simple estimate in the rejective sampling. *Probability and Statistical Inference*, pp. 307-317.

Prášková Z. (1985). On the convergence to the Poisson distribution in rejective sampling from a finite population. In: *Probability and Mathematical Statistics with Applications* (Visegrad 1985, W. Grossmann et al. eds.), Reidel, Dordrecht 285-294.

Przybycin Z., Rojek P. (1966). Metody reprezentacyjne w badaniu sprawozdań finansowych. Stowarzyszenie Ksiegowych w Polsce Warszawa.

Raftery A. E. (1995). Bayesian model selection in social research. In P.V. Marsden (Ed.), *Sociological Methodology*, Blackwells, Cambridge, pp. 111-196.

Rao C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New-York - London - Sydney - Toronto.

Rao, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling Without Replacement. *Journal of the Indian Statistical Association* 3, 173-180.

Rao J.N.K. (2003). Small Area Estiamtion. John Wiley and Sons, New York.

Rao J.N.K., Hartley H.O., Cochran W.G. (1962). On a simple procedure of unequal probability sampling without replacement *Journal of the Royal Statistical Association* B 24, 2, 482-491.

Rao J.N.K., Wu C.F,J., Yue K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, vol. 18, no. 2, 209-217.

Rao T.J. (1977). Estimating the variance of the ratio estimator for the Midzuno-Sen sampling scheme. *Metrika*, vol. 24, pp. 203-208.

Rao T.J. (2004). Five decades of the Horvitz·Thompson estimator and furthermore. *Journal of the Indian Society of Agricultural Statistics* vol. 58, 177-189.

Rao T.V.H. (1962). An existence theorem in sampling theory. *Sankhya*, vol. A 24, pp. 327-330.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert C. (2007). *The Bayesian Choice*. Springer, New York.
Robert D. M. (1978). *Statistical Auditing*. AICPA, New York.
Rosén B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I, II. *Annals of Mathematical Statistics* vol. 43, 373-397, 748-776.
Ryan T. P. (2013). *Sample Size Determination and Power*. John Wiley & Sons, Hoboken, New Jersey.
Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* 54, 499-513.
Santner T.J, Duffy D. E. (1989). *The Statistical Analysis of Discret Data*. Springer-Verlag New York.
Särndal C. E., B. Swensson, J. Wretman (1992). *Model Assisted Survey Sampling.* Springer Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong Kong-Barcelona-Budapest.
Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* vol. 23, (49), 807-812.
Seber G. A. F. (2013). *Statistical Models for Proportions and Probabilities*. Springer Briefs in Statistics, Heidelberg New York Dordrecht London.
Sen A. R. (1953). On the estimate of variance in sampling with varying probabilities *Journal of the Indian Society of Agicultural Statistics* 5, 2, pp. 119-127.
Sen P. K. (1995). The Hájek asymptotics for finite population sampling and their ramification. *Kybernetika* vol. 31, no. 3, 251-268.
Serfling R. J. (1968). Approximately optimal stratification. *Journal of the American Statistical Association* vol. 63, no. 324, 1298-1309.
Sethi V. K. (1963). Note on optimum stratification of population for estimating the population means. *The Australian Journal of Statistics* vol. 5, 20-33.
Silvey S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics* vol. 30, no. 2, pp. 389-407.
Sorensen J. E. (1969). Bayesian analysis in auditing. *The Accounting Review* vol. 45, no. 3, 555-561.
Statistical models and analysis in auditing: Panel on Nonstandard Mixtures of Distributions. (1989). *Statistical Science*, vol.4, nr 1, 2-33.
Stringer K. W. (1963). Practical aspects of statistical sampling in auditing. In *Asa Proceeding of the Business and and Economic Statistics Section* American Statistical Association, 405-411.
Sunter A. B. (1977a). Responce burden, sample rotation, and classification renoval in economic surveys. *International Statistical Review*, vol. 45, 209-222.
Sunter A. B. (1977b). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics* vol. 26, 261-268.
Talens E. (2005). *Statistical auditing and the AOQL-method.* Thesis. University of Groningen. Labyrint Publications Offsetdrukkerij Ridderprint B.V., Ridderkerk. https://www.rug.nl/research/portal/publications/pub(29b99352-9f8a-4e5d-9ef2-cfa7fe902ce2).html

Tillé Y. (1998). Estimation in surveys using conditional inlusion probabilities: simple random sampling. *International Statistical Review*, vol. 66, 3, pp. 303-322.

Tillé Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Paris: Dunod.

Tillé Y. (2006). *Sampling algorithms*. Springer New York.

Tschuprow A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, vol. 2, 461-193, 646-683.

Víšek J. Á. (1979). Asymptotic distribution of sample estimate for rejective, Sampford and successive sampling. In: *Contributions to Statistics: Jaroslav Hájek Memorial* Volume (J. Jurečková, ed.), Academia, Prague and Reidel, Dordrecht, 363-376.

Walter G. G., Hamdani G. G. (1987). Empiric Bayes estimation of binomial probability. *Communication in Statistics - Theory and Methods* 16, 559-577.

Wendell J. P., and Schmee J. (1996). Exact inference for proportions from a stratified finite population. *Journal of the American Statistical Association* vol. 91, 825-830.

Wilcox R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* Elsevier Inc. Amsterdam, Boston, Heidelrberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo.

Wilks S.S. (1962). *Mathematical Statistics*. John Wiley and Sons, Inc. New York, London.

Wywiał J. L. (1981). O pewnych unormowanych współczynnikach asymetrii i spłaszczenia. *Przegląd Statystyczny* vol. 28, 263-269.

Wywiał J. L. (1982). O mierzeniu i testowani odchyleń od normalności rozkładu prawdopodobieństwa jednowymiarowej zmiennej losowej (in Polish). *Przegląd Statystyczny*, vol. 29, 1982, 415-425.

Wywiał J. L. (1992). *Statystyczna metoda reprezentacyjna w badaniach ekonomicznych*. Akademia Ekonomiczna w Katowicach, Katowice.

Wywiał J. L. (1995). *Wielowymiarowe aspekty metody reprezentacyjnej*. Ossolineum, Wrocław Warszawa Kraków.

Wywiał J. L. (2003). *Some Contributions to Multivariate Methods in Survey Sampling.* Katowice University of Economics, Katowice.
http://www.ue.katowice.pl/jednostki/wydawnictwo/darmowy-e-book.html

Wywiał J. L. (2003a). On conditional sampling strategies. *Statistical Papers* vol. 44, 3, pp. 397-419.

Wywiał, J. L. (2007). Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design proportionate to the order statistic of an auxiliary variable. *Statistics in Transition-new series* vol. 8 (1), pp. 125-137.

Wywiał, J. L. (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers* vol. 49 (2), pp. 277-289.

Wywiał, J. L. (2009). Performing quantiles in regression sampling strategy. *Model Assisted Statistics and Applications* vol. 4, No. 2, pp. 131-142.

Wywiał J. L. (2009a). Sampling design proportional to positive function of order statistics of auxiliary variable. *Studia Ekonomiczne-Zeszyty Naukowe*, vol. 53, pp. 35-60.

Wywiał J. L. (2010). Wprowadzenie do metody reprezentacyjnej (In Polish). Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice.

Wywiał J. L. (2011). Sampling designs proportionate to sum of two order statistics of auxiliary variables. *Statistics in Transition - new series* vol. 12, No. 2, pp. 231-248.

Wywiał J. L. (2012). On limit distribution of Horvitz-Thompson statistic under the rejective sampling. *Studia Ekonomiczne - Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach* no. 120, 84-96. http://www.ue.katowice.pl/uploads/media/8_J.L.Wywial_On_Limit_Distribution....pdf.

Wywiał J. L. (2012a). Application of order statistics of auxiliary variable to estimation the population mean. *Statistics in Transition - new series* vol. 13, No. 2, pp. 279-286.

Wywiał J. L. (2013). On limit distribution of Horvitz-Thompson statistic under Poisson sampling design. *Studia Ekonomiczne - Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach* no. 133, 61-70. http://www.ue.katowice.pl/uploads/media/4_J.L.Wywial_On_Limit_Distribution....pdf

Wywiał J. L. (2013a). O optymalizacji rozmiaru próby warstwowej w badaniu wiarygodności sprawozdań. *Zeszyty Teoretyczne Rachunkowości*, tom 70 (126), SKwP, Warszawa, s. 129-139.

Wywiał J. L. (2013b). Sampling designs proportionate to sum of two order statistics of auxiliary variable. *Statistics in Transition new series* vol. 14, No. 2, 231-248. http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/

Wywiał J. L. (2013c). Sampling design proportionate to non-negative functions of two quantiles of auxiliary variable. *Studia Ekonomiczne - Zeszyty naukowe Uniwersytetu Ekonomicznego w Katowicach*, nr 152, pp. 174-190. http://www.ue.katowice.pl/uploads/media/SE_152.pdf

Wywiał J. L. (2014). On Bayesian tests in auditing. In: *Proceedings of 17-th Conference Applications of Mathematics and Statistics in Economics* Editors: Z. Rusnak and B. Zmyślona. Wrocław University of Economics 2014, pp. 284-293, http://www.amse.ue.wroc.pl/proceedings.html.

Wywiał J. L. (2014a). *Próby losowe w audycie finansowym* (In Polish). Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

Wywiał J. L. (2014b). On conditional simple random sample. *Statistics in Transition new series* vol. 15, No. 4, 525-534. http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/

Wywiał J. L. (2015). *Sampling designs dependent on sample parameters of auxiliary variables*. Springer Briefs, Heidelberg, New York, Dordrecht, London.

Yates F., Grundy P.M. (1953). Selection without replacement from within strata with probability proporcional to size. *Journal of the Royal Statistical Society* vol. B15, pp. 235-261.

Zieliński W. (2010). *Estymacja wskaźnika struktury*. Wydawnictwo SGGW, Warszawa.

Żądło T. (2008). *Elementy statystyki małych obszarów z programem R*. Akademia Ekonomiczna w Katowicach, Katowice.

Żądło T. (2015). *Statystyka małych obszarów w badaniach ekonomicznych. Podejście modelowe i mieszane*. Uniwersytet Ekonomiczny w Katowicach, Katowice.