Zastosowanie mieszaniny dwóch rozkładów gamma do audytu finansowego
Application of two gamma distributions mixture to financial auditing

Janusz L. Wywiał

Department of Statistics, Econometrics and Mathematics
Management Faculty, University of Economics in Katowice
janusz.wywial@ue.katowice.pl
Result of a grant supported by the *National Science Centre, Poland*,
no. DEC-2012/07/B/HS4/03073

2nd Congress of Polish Statistics, Warsaw, 10-12 June 2018

Contents

- Model of accounting observations.
- Mixture of gamma distributions.
- Moment method of estimation.
- Likelihood ratio test.
- Monte Carlo test.
- Conclusion.
- Reference.

- U-population of size N, s-sample of size $n \leq N$;
- vector of book values (auxiliary variable): $\boldsymbol{x} = [x_1 ... x_N]$, $\boldsymbol{x} \in R_+^N$ is outcome of $\boldsymbol{X}^T = [X_1 ... X_N]$;
- true (without errors) accounting amounts: $\mathbf{y} = [y_1 ... y_N]$, $\boldsymbol{y} \in R_+^N$ is outcome of $\boldsymbol{Y}^T = [Y_1 ... Y_N]$;
- accounting amounts contaminated by errors: $\mathbf{w} = [w_1 ... w_N]$, $\boldsymbol{w} \in R_+^N$ is outcome of $\boldsymbol{W}^T = [W_1 ... W_N]$;
- $\mathbf{Z}^T = [Z_1 ... Z_N]$, $Z_i = 0$ $(Z_i = 1) \Leftrightarrow X_i = Y_i$ $(X_i = W_i)$, $i \in U$;

- Rows: $[X_i\ Y_i\ W_i\ Z_i]$ of matrix $[\mathbf{X}\ \mathbf{Y}\ \mathbf{W}\ \mathbf{Z}]$ are independent and identically distributed as random vector $[X\ Y\ W\ Z]$;

$$X = (1 - Z)Y + ZW \qquad \text{or} \qquad X = Y + ZR \tag{1}$$

  $R = W - Y$ is the auditing error, we assume that $R \geq 0$;

- The probability distribution of the matrix $[\mathbf{X}\ \mathbf{Y}\ \mathbf{W}\ \mathbf{Z}]$ is the population model.
- Distribution of X is the following mixture of distributions:

$$F(x|\boldsymbol{\theta}) = (1 - p)F_0(x|\boldsymbol{\theta}_0) + pF_1(x|\boldsymbol{\theta}_1), \tag{2}$$

  $P(Z = 1) = p$, $P(Z = 0) = 1 - p$;
  $F_0(x|\boldsymbol{\theta}_0) = F(x|Z = 0) = F_0(y|\boldsymbol{\theta}_0)$,
  $F_1(x|\boldsymbol{\theta}_1) = F(x|Z = 1) = F_1(w|\boldsymbol{\theta}_1)$,
  $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \cup \boldsymbol{\theta}_1$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$,

$$f(x|\boldsymbol{\theta}) = (1 - p)f_0(x|\boldsymbol{\theta}_0) + pf_1(x|\boldsymbol{\theta}_1). \tag{3}$$

- Expected mean accounting error:

  $\tau = E(\bar{X} - \bar{Y}) = p(E(W|\theta_1) - E(Y|\theta_0))$

  or expected total account. error:

  $N\tau = E(\sum_{i\in U} X_i - \sum_{i\in U} Y_i)$;

- Hypotheses: $H_0 : \quad \tau = \tau_0, \qquad H_1 : \quad \tau = \tau_1 > \tau_0$

  τ_0, (τ_1): admissible, (un-admissible) exp. account. error.

- Let α-significance level (risk of incorrect rejection of H_0),

  $(1 - \beta)$-probability of II kind error (risk of incorrect acceptance of H_0) where β-power of the test.

- Before auditing process the following data are observed:

$$\boldsymbol{X} = (X_i : i \in U) = (\boldsymbol{X}_s, \boldsymbol{X}_{U-s})$$

where

$$\boldsymbol{X}_s = (X_i : i \in s), \quad \boldsymbol{X}_{U-s} = (X_i : i \in U - s)$$

- After the auditing process the following data are observed:

$$\mathcal{D} = (\mathcal{D}_s, \boldsymbol{X}_{U-s}), \quad \mathcal{D}_s = ((X_i, Z_i) : i \in s) = (\boldsymbol{Y}_{s_0}, \boldsymbol{W}_{s_1}).$$

$\boldsymbol{d}$, $\boldsymbol{d}_s$, $\boldsymbol{x}$, $\boldsymbol{x}_s$, $\boldsymbol{x}_{U-s}$, $\boldsymbol{y}_{s_0}$ and $\boldsymbol{w}_{s_1}$ are outcomes of $\mathcal{D}$, $\mathcal{D}_s$, $\boldsymbol{X}$, $\boldsymbol{X}_s$, $\boldsymbol{X}_{U-s}$, $\boldsymbol{Y}_{s_0}$ and $\boldsymbol{W}_{s_1}$, respectively.

- Let: Let $Y \sim G(a, c)$ and $R \sim G(b, c)$ be independent, then variable $W = Y + R \sim G(a + b, c)$;
- the mixture:

$$f(x|a, b, c, p) = pf_1(x|a, b, c) + (1 - p)f_0(x|a, c) \qquad (4)$$

where

$$f_1(x|a, b, c) = \frac{c^{a+b}}{\Gamma(a+b)} x^{a+b-1} e^{-cx}, \quad f_0(x|a, c) = \frac{c^a}{\Gamma(a)} x^{a-1} e^{-cx},$$

$$\tau = p(E(X|a, b, c, p) - E(Y|a, c)) = \frac{pb}{c}. \qquad (5)$$

- The solution $\{p_U(x),\ a_U(x),\ b_U(x),\ c_U(x)\}$ of the equation system, Wywiał(2016, 2018):

$$E(X^e) = m_e(x), \quad e = 1, 2, 3, 4, \quad N > 4,$$

$m_e(x) = \frac{1}{N}\sum_{i \in U} x_i^e$, is the estimator of $\{p,\ a,\ b,\ c\}$.

- Test statistic:

$$\hat{G}_1 = \frac{\hat{\tau}_1 - \tau_0}{\sqrt{Q_U(\mathcal{D})}}, \qquad \hat{\tau}_1 = \frac{p_U b_U}{c_U},$$

$Q_U(\mathcal{D})$ - e.g. bootstrap type estimator.

- p-value could be evaluated based on limit distribution of $\hat{G}_1$ or Monte-Carlo procedures.

Moment method of estimation

The sample s is not selected, $s = s_0 \cup s_1$, $s_0 \neq \varnothing$, $s_1 \neq \varnothing$

- Test statistic, Wywiał(2018):

$$\hat{G}_2 = \frac{\hat{\tau}_2 - \tau_0}{\sqrt{\frac{V_{U-s}(X)}{N-n} + \frac{V_{s_0}(Y)}{n_0}}}, \qquad \hat{\tau}_2 = \bar{X}_{U-s} - \bar{Y}_{s_0}. \qquad (6)$$

- where: $\{P_U, A_{s_0}, B_s, C_{s_0}\}$ are estimators of $\{p, a, b, c\}$:

$$\begin{cases} P_U = \frac{\bar{X}_{U-s} - \bar{Y}_{s_0}}{\bar{R}_{s_1}}, \quad A_{s_0} = \frac{\bar{Y}^2_{s_0}}{V_{s_0}(Y)}, \\ B_s = \frac{\bar{Y}_{s_0}\bar{R}_{s_1}}{V_{s_0}(Y)}, \quad C_{s_0} = \frac{\bar{Y}_{s_0}}{V_{s_0}(Y)} \end{cases} \qquad (7)$$

  provided denominators of the above ratios are positive.

- p-value could be evaluated based on limit distribution of $\hat{G}_2$ or Monte-Carlo methods.
- Cases: $(s_0 \neq \varnothing,\ s_1 = \varnothing)$ and $(s_0 = \varnothing,\ s_1 \neq \varnothing)$ are considered by Wywiał (2016)

- Log-likelihood function:

$$l\left(\boldsymbol{d}|\theta\right) = ln\left(L\left(\boldsymbol{d}|\theta\right)\right) = k ln(p) + (n-k) ln(1-p) +$$
$$+ \sum_{i \in s_1} ln(f_1(x_i|\boldsymbol{\theta}_1)) + \sum_{i \in s_0} ln(f_0(x_i|\boldsymbol{\theta}_0)) + \sum_{i \in U-s} ln(f(x_i|\boldsymbol{\theta})).$$

- Log-likelihood function in the case of gamma-mixture distribution:

$$l(\boldsymbol{d}, a, b, c, p) = k\ ln(p) + (n-k) ln(1-p) + Na\ ln(c) + kb\ ln(c) +$$
$$-k\ ln(\Gamma(a+b)) - (n-k) ln(\Gamma(a)) + (a-1) \sum_{j \in U} ln(x_j) + b \sum_{j \in s_1} ln(x_j) +$$
$$- c \sum_{j \in U} x_j + \sum_{j \in U-s} ln\left(\frac{1-p}{\Gamma(a)} + \frac{p(cx_j)^b}{\Gamma(a+b)}\right). \quad (8)$$

- Likelihood ratio statistic:

$$\lambda = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\tau(\theta)=\tau_0} L(\mathcal{D}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\mathcal{D}|\boldsymbol{\theta})}. \tag{9}$$

- Distribution of $ln(\lambda)$ is approximated by chi-square distribution with 1 degree of freedom under the sufficiently large size of sample.

Monte Carlo test 1

Parameter of gamma distribution estimated by method of moments

- $f(x|a, b, c, p)$ is transformed by means $c = \frac{pb}{\tau}$ into:

$$f(x|a, b, \tau, p) = pf_1(x|a, b, \tau) + (1-p)f(x|a, \tau).$$

- Parameters a, b are replaced with estimators given by (7), see Dufour (2006).
- Data $\boldsymbol{d}^{(0,i)} = \left(\boldsymbol{y}_{s_0}^{(0,i)}, \boldsymbol{w}_{s_1}^{(0,i)}, \boldsymbol{x}_{U-s}^{(0,i)}\right)$, $i = 1, ..., m$, are generated according to $f_0(y|a_{s_0}, \tau_0)$ with probability $(1-p)$ and $f_1(w|a_{s_0}, b_s, \tau_0)$ with prob. p;
- Data $\boldsymbol{d}^{(1,i)} = \left(\boldsymbol{y}_{s_0}^{(1,i)}, \boldsymbol{w}_{s_1}^{(1,i)}, \boldsymbol{x}_{U-s}^{(1,i)}\right)$ are generated m-times according to $f_0(y|a_{s_0}, \tau_1)$ with probab. $(1-p)$ and $f_1(w|a_{s_0}, b_s, \tau_1)$ with prob. p;

- Test statistic, Wywiał(2018):

$$\hat{g}_2^{(e,i)} = \frac{\tau^{(e,i)} - \tau_0}{\sqrt{\frac{v_{U-s}\left(\boldsymbol{x}^{(e,i)}\right)}{N-n} + \frac{v_{s_0}\left(\boldsymbol{y}^{(e,i)}\right)}{n_0}}}, \quad \tau^{(e,i)} = \bar{X}_{U-s}^{(e,i)} - \bar{Y}_{s_0}^{(e,i)}, \; e = 0, 1.$$

- Data $\{\hat{g}_2^{(e,i)}, i = 1, ..., m\}$ approximates the distrib. of $\hat{G}_2$ when hypothesis H_e is true, $e = 0, 1$;
- Let (see: Dufour and Khalaf (2001)):

$$\eta_e = \frac{m\omega_e}{m+1}, \quad \omega_e = \frac{1}{m}\sum_{i=1}^{m} I(\hat{g}_2^{(e,i)}), \quad I(\hat{g}_2^{(e,i)}) = \begin{cases} 1, & \text{if } g \geq \hat{g}_2 \\ 0, & \text{if } g < \hat{g}_2 \end{cases},$$

ω_e is equal to the frequency of appearing inequalities $\hat{g}_2^{(e,i)} \geq \hat{g}_2$, $i = 1, ..., m$, $e = 0, 1$.

Monte Carlo test 3

Simulated distribution of the test statistic. Decisions

- *p*-value the power of the test is assessed by $\hat{\alpha} = \eta_0$, and $\hat{\beta} = \eta_1$, respectively.
- If $\hat{\alpha} \leq \alpha$, H_0 is rejected, α is the risk of incorrect rejection;
- If $\hat{\alpha} > \alpha$, H_0 is accepted, $(1 - \hat{\beta})$ is the risk of incorrect acceptance.

- After substituting c for $\frac{pb}{\tau}$ in $l(\boldsymbol{d}, a, b, c, p)$, (see expr. (8)):

$$l(\boldsymbol{d}, a, b, \tau, p) = k\ ln(p) + (n-k) ln(1-p) + Na\ ln(c) + kb\ ln(c) +$$
$$-k\ ln(\Gamma(a+b)) - (n-k) ln(\Gamma(a)) + (a-1) \sum_{j \in U} ln(x_j) + b \sum_{j \in s_1} ln(x_j) +$$
$$-\frac{pb}{\tau} \sum_{j \in U} x_j + \sum_{j \in U-s} ln\left(\varphi(a, b, \tau, p, x_j)\right) \quad (10)$$

where

$$\varphi(a, b, \tau, p, x_j) = \frac{1-p}{\Gamma(a)} + \frac{p^{b+1} (b x_j)^b}{\tau^b \Gamma(a+b)}.$$

- Now: $E(X) = \tau$.

- The likelihood ratio test statistic, Wywiał(2018):

$$t = 2\left(l(\boldsymbol{d}, \hat{a}, \hat{b}, \hat{\tau}, \hat{p}) - l(\boldsymbol{d}, \tilde{a}, \tilde{b}, \tau_e, \tilde{p}) \right)$$

$(\hat{a}, \hat{b}, \hat{\tau}, \hat{p})$ maximizes $l(\boldsymbol{d}, a, b, \tau, p)$ (see: (10)),
$(\tilde{a}, \tilde{b}, \tilde{p})$ maximizes $l(\boldsymbol{d}, a, b, \tau_0, p)$.

- $\boldsymbol{d}^{(e,i)}$, $e = 0, 1$, is generated according to $f(x|\tilde{a}, \tilde{b}, \tau_e, \tilde{p})$.
- Simulated distribution of test statistic:

$$t_i^{(e)} = 2\left(l(\boldsymbol{d}^{(e,i)}, \hat{a}^{(i)}, \hat{b}^{(i)}, \hat{\tau}^{(i)}, \hat{p}) - l(\boldsymbol{d}^{(e,i)}, \tilde{a}^{(i)}, \tilde{b}^{(i)}, \tau_e, \tilde{p}^{(i)}) \right),$$

$(\hat{a}^{(i)}, \hat{b}^{(i)}, \hat{\tau}^{(i)}, \hat{p}^{(i)})$ maximizes $l(\boldsymbol{d}^{(e,i)}, a, b, \tau, p)$,
$(\tilde{a}^{(i)}, \tilde{b}^{(i)}, \tilde{p}^{(i)})$ maximizes $l(\boldsymbol{d}^{(e,i)}, a, b, \tau_e, p)$.

Monte Carlo test 6

Simulated distribution of the likelihood ratio test. Decisions

- t_α is the critical value of the test defined as the sample quantile of order $(1-\alpha)$ of $\{t_i^{(0)}, i = 1, ..., m\}$;
- $\hat{\beta}$, is the power evaluated as the frequency of appearing inequalities $t_i^{(1)} \geq t_\alpha$, $i = 1, ..., m$;
- if $t \geq t_\alpha$, H_0 is rejected, α is the risk of incorrect rejection;
- if $t < t_\alpha$, H_0 is accepted, $(1-\hat{\beta})$ is the risk of incorrect acceptance.

Conclusions

- Model of accounting data is defined as mixture of two distributions.
- In particular, the mixture of two gamma distribution is considered.
- Hypothesis on the mean accounting error is tested by means of studentized estimator of the mean or likelihood ratio test.
- Monte-Carlo methods of approximation distributions of test statistics is proposed.
- Specification of the alternative distribution let us control the error of II kind (risk of incorrect acceptance).
- It is possible to test the hypothesis without auditing process of data provided the mixture model is true.
- Mixtures of other distributions can be considered.

Reference

- Cox D. R., Snell E. J. (1979). On sampling and the estimation of rare errors. *Biometrika*, 66, No. 1, pp. 125-132. Errata: *Biometrika*, 1982, 69, No. 2, p. 491.
- Dufour J. M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics* vol. 133, pp. 443-477.
- Dufour J. M., Khalaf L. (2001). Monte Carlo test methods in econometrics. In: *Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford, U.K., pp. 494-519.
- Fienberg S. E., Nether J., Leitch R. A. (1977). Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, vol. 72, pp. 295-302.
- Frost P. A., Tamura H. (1986). Accuracy of auxiliary information interval estimation in statistical auditing. *Journal of Accounting Research* vol. 24, pp. 57-75.

Reference

- Hall P. (1992). *The Bootstrap and Edgewrth Expansion.* Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.
- Hope C.A. (1968). A simplified Monte-Carlo significance test procedure. *Journal of the Royal Statistical Society* B30, no. 3, 582-598.
- MacKinnon J. (2007): Bootstrap hypothesis testing. Queen's Economics Department, *Working Paper* no. 1127.
- Marazzi A., Tillé Y. (2016). Using past experience to optimize audit sampling design. *Review of Quantive Finance Accounting* p. 1-28, doi:10.1007/s11156-016-0596-7.
- McLachlan G., Peel D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc. New York Chichester Weinheim Brisbane Singapore Toronto.

Reference

- Rao C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New-York - London - Sydney - Toronto.
- Silvey S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics* vol. 30, no. 2, pp. 389-407.
- Statistical models and analysis in auditing. Panel on Nonstandard Mixtures of Distributions. (1989). *Statistical Science*, vol.4, nr 1, 2-33.
- Wywiał J. L. (2016). *Contributions to Testing Statistical Hypotheses in Auditing*. PWN, Warsaw.
- Wywiał, J. L. (2018). ***Application of two gamma distributions mixture to financial auditing***. ***Sankhya B***, vol. 80, issue 1, 1-18.

***Thank you very much for attention*.**