Estimation of Confidence Intervals for the Log-normal Means and for the Ratio and the Difference of Log-normal Means with Application of Breath Analysis

K. Cimermanová

Department of Theoretical Methods, Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 01 Bratislava, Slovakia
Email: katarina.cimermanova@gmail.com

Abstract. *Cancer is one of the leading causes of death in the world. Breath analysis is a possible diagnostic tool for early detection of lung and esophageal cancer (for more details consult e.g. the FP6 project of European Commission BAMOD – Breath-gas analysis for molecular oriented detection of minimal diseases). In this paper we deal with estimating generalized confidence intervals for the ratio and the difference of log-normal means, which are based on generalized pivotal quantities. We provide series of computer simulations and apply this method to the breath analysis data. We present the estimated 95% generalized confidence intervals for the ratio and the difference of mean concentrations of selected volatile organic compounds for patients with lung cancer and healthy volunteers. From results we would like to find potential bio-marker of lung cancer.*

Keywords: breath compounds, confidence intervals, generalized pivotal approach, generalized pivotal quantities

1. Introduction

The human breath gas contains a large number of compounds (300-3000 different compounds), see [1]. In particular volatile organic compounds (VOCs) which are produce endogenously by human body during metabolic processes, have a potential to participate in cancer detection as an effective bio-markers. Small concentrations of VOCs (ranging from a few ppb to a few hundred ppb – particles per billion in volume) are measured by non-invasive technologies. It is very attractive because it can be easily applied to sick patients, including children and elderly people. Technologies are gas chromatography with mass spectrometry (GC-MS), selected ion flow tube mass spectrometry (SIFT-MS), proton transfer reaction mass spectrometry (PTR-MS), ion mobility spectrometry (IMS) and laser optical spectroscopy, which are in details described e.g. in [1, 2].

In what follows, we consider the concentration measurements derived by PTR-MS. Quantification of the specific breath compounds by PTR-MS is performed by considering the kinetics of the chemical ionization. It is based on proton-transfer reactions with protonated water H_3O^+ as the primary reactant ion ($H_3O^+ + M \rightarrow MH^+ + H_2O$, where M denotes the selected specific VOC), and further, on the compound-specific rate coefficient k_M and the time t_{drift} which the swarm of ions take to traverse the drift chamber of the PTR-MS device. Identification of substances is not always possible with PTR-MS. A few compounds can usually be attributed to the detectable molecular masses, i.e. certain mass to charge ratios (m/z values) of product ions. Molecular masses detectable by PTR-MS range from m/z 21 to m/z 230. The measured quantities are transformed using the knowledge of chemistry kinetics and the reaction constants [5] to concentrations of VOCs in ppb levels.

The breath gas samples of 125 lung carcinoma patients ($N_1 = 125$) and 279 healthy test volunteers ($N_2 = 279$) have been collected in Tedlar bags (sampling bags) with parallel collection of the ambient air (room air). For each breath sample collected in Tedlar bag, there is a set of several measurements in the database. Medians of the bag's replicated measurements were used for statistical analysis. Further pre-processing of the raw measurements (including removing of outliers) was performed. The applied filter discarded all those concentrations of VOCs in the exhaled breath which were less than double the respective inspiratory (room-air) concentration.

Selection of the VOCs used for further analysis was based on previous research and knowledge [2, 3, 5]. The following 12 VOCs ($n = 12$) with detected molecular mass to charge ratio m/z 28 (tentatively identified as hydrogen cyanide), m/z 31 (formaldehyde), m/z 33 (methanol), m/z 42 (acetonitrile), m/z 59 (acetone), m/z 61 (acetic acid), m/z 63, m/z 79 (benzene), and further m/z 97, m/z 105, m/z 109 and m/z 123, presented also in [3, 5], were included in the study. Recent result suggests that breath concentration could be expected to be log-normally distributed and the logarithmic transformation of data could be profitable. Therefore we assume that the concentration of VOCs in breath gas, as positive random variables X_j (j = 1... n), follow log-normal distribution, for more details see [3].

In this paper we are interested in comparison of the log-normal means of the selected VOC-concentrations based on generalized pivotal approach, introduced by Krishnamoorty and Mathew in [4], for construction of confidence intervals. The derived confidence intervals are based on statistical method with the generalized pivotal quantities, introduced by Weerahandi in [6]. The method is in detail described in the following section.

2. Generalized pivotal approach

Now, we define the generalized confidence interval for the ratio and the difference of two log-normal means as derived in [4].

Let X be a positive random variable whose distribution depends on parameters (θ, δ), where θ represents the scalar parameter of interest and δ represents a nuisance (possibly vector) parameter. Let x denote observed value of X. Consider that generalized pivot $T(X; x, \theta, \delta)$ depends on random variable X, its observed value x and parameters (θ, δ). This quantity satisfies the following conditions: for fixed x, its distribution does not depend on any unknown parameters and observed value of $T(X; x, \theta, \delta)$, i.e. $T(x; x, \theta, \delta)$ is free of the nuisance parameter δ.

Further, given the observed value x, let t_1 and t_2 be such values that $P(t_1 \leq T(X; x, \theta, \delta) \leq t_2) = 1- \alpha$ for chosen significant level $\alpha \in (0, 1)$, than the confidence interval for parameter θ defined by $\{\theta : t_1 \leq T(X; x, \theta, \delta) \leq t_2\}$ is a 100(1- α)% generalized confidence interval for θ.

We assume that X_i (i = 1, 2, X_i represents measured concentration of a volatile organics compound from population i) have a log-normal distribution with parameters μ_i and σ_i. Then the random variable $Y_i = \ln(X_i)$ follows normal distribution $N(\mu_i, \sigma_i)$. The log-normal mean m_i is

$$m_i = \exp\{\mu_i + \sigma_i^2 / 2\}. \tag{1}$$

The ratio of two lognormal means of two independent populations (population of lung cancer patients with log-normal mean m_1 and population of healthy test volunteers with log-normal mean m_2) is

$$m_1 / m_2 = \exp\{\mu_1 + \sigma_1^2 / 2 - \mu_2 - \sigma_2^2 / 2\}, \tag{2}$$

and the difference of two lognormal means is

$$m_1 - m_2 = \exp\{\mu_1 + \sigma_1^2 / 2\} - \exp\{\mu_2 + \sigma_2^2 / 2\}, \tag{3}$$

where m_i is the log-normal mean of the ith population (i = 1, 2) and μ_i and σ_i are the unknown parameters of log-normal distribution of the ith population.

The parameter μ_i of log-normal distribution could be estimated by the sample mean

$$\overline{Y_i} = \frac{1}{N_i} \sum_{k=1}^{N_i} Y_{ik}, \tag{4}$$

and the parameter σ_i^2 of log-normal distribution by the sample variance

$$S_i^2 = \frac{1}{(N_i - 1)} \sum_{k=1}^{N_i} (Y_{ik} - \overline{Y_i})^2 \tag{5}$$

based on random sample $Y_{i1} \ldots Y_{iNi}$ ($Y_{ik} = \ln(X_{ik})$, i = 1, 2; k = 1… N_i), where N_i is sample size of the ith population. We denote $\overline{y}_i$ and s_i^2, the observed values of $\overline{Y_i}$ and S_i^2.

The generalized pivotal quantity for parameter $\theta_i = \ln(m_i)$ (logarithm of log-normal mean) is

$$T_i = \bar{y}_i - \frac{\bar{Y}_i - (\theta_i - \sigma_i / 2)}{S_i / \sqrt{N_i}} \frac{s_i}{\sqrt{N_i}} + \frac{1}{2} \frac{\sigma_i^2}{S_i^2} s_i^2 . \tag{6}$$

This quantity is a function of the random variables $\bar{Y}_i$ and S_i^2, their observed values $\bar{y}_i$ and s_i^2 and it further depends on the unknown parameters μ_i and σ_i^2 (in fact the only parameter of interest is $\theta_i = \ln(m_i) = (\mu_i + \sigma_i^2/2)$ and σ_i^2 is considered to be a nuisance parameter). After substitution of $\bar{y}_i$ and s_i^2 for the random variable $\bar{Y}_i$ and S_i^2 we get the observed value $T_i = \mu_i + \sigma_i^2/2 = \theta_i$ which is free of the nuisance parameter. Note that T_i can be rewritten as follows

$$T_i = \bar{y}_i - \frac{Z_i}{Q_i / \sqrt{N_i - 1}} \frac{s_i}{\sqrt{N_i}} + \frac{1}{2} \frac{s_i^2}{Q_i / \sqrt{N_i - 1}} , \tag{7}$$

where $i = 1, 2,$ $Z_i = \sqrt{N_i}(\bar{Y}_i - \mu_i) / \sigma_i \sim N(0, 1)$ and $Q_i^2 = (N_i - 1) S_i^2 / \sigma_i \sim \chi^2_{N_j - 1}$. Z_i and Q_i are independent and are generated as random values of given distributions. From that it is clear that the distribution of T_i does not depend on the unknown parameters (it directly depends on the realized values $\bar{y}_i$ and s_i^2 and the sample size N_i, only).

We get 100(1-α) % confidence interval for $\ln(m_i)$ as α / 2 and (1 - α / 2) quantiles (l – lower bound, u – upper bound) obtained from k generated values of T_i. The generalized confidence interval (GCI) for log-normal mean is

$$[\exp\{T_{i(l)}\}, \exp\{T_{i(u)}\}] . \tag{8}$$

Generalized pivotal quantity for logarithm of the ratio of two log-normal means $\ln(m_1 / m_2)$ is

$$T_R = T_1 - T_2, \tag{9}$$

where the quantities T_1 and T_2 from Eq. 6 are for the same VOC. Generalized pivotal quantity for the difference of log-normal means (m_1 - m_2) is

$$T_D = \exp\{T_1\} - \exp\{T_2\} . \tag{10}$$

For each volatile organic compound (n = 12), a GCI is constructed by generating k random values T_i from Eq. 7 for the ith population (i = 1, 2; population of lung carcinoma patients and population of healthy test volunteers). Then we sort T_R (T_D) from Eq. 9 (from Eq. 10) in increasing order and find α / 2 and (1 - α / 2) quantiles (l – lower bound, u – upper bound). Then the confidence bounds for the ratio of two log-normal means are

$$[\exp\{T_{Ri(l)}\}, \exp\{T_{Ri(u)}\}] \tag{11}$$

and for the difference of two log-normal means the GCI is

$$[T_{Di(l)}, T_{Di(u)}] . \tag{12}$$

3. Statistical properties of the generalized confidence intervals - a simulation study

It is known that generalized confidence intervals do not always have exact frequentist coverage. In order to check statistical properties we conducted a large simulation study. Results from this study are in tables 6, 7, 8, 9 and 10 in [3]. We simulated the 95 % GCIs for log-normal mean m with the values of true parameters μ = [-5, -2, 0, 2, 5] and σ^2 = [0.1, 0.5, 1, 2, 5, 20] with sample size N = [4, 8, 12, 30, 300]. Parameters represent typical situations observed during concentration measurements of

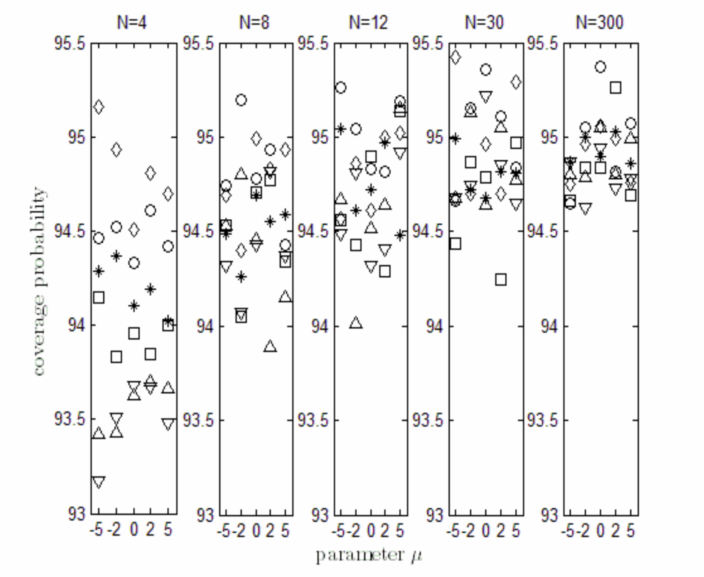

Fig. 1. Empirical coverage probability of generalized confidence intervals for lognormal mean (- σ²=0.1, ○ - σ²= 0.5, Δ - σ²=1,▼ - σ²=2, *- σ²=5, ◊ - σ²=20).

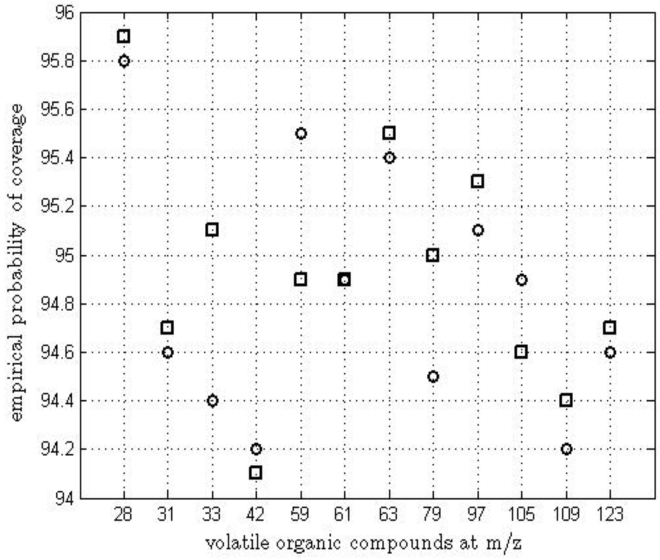

Fig. 2. Coverage probability of generalized confidence intervals for the ratio (○) and the difference () of two lognormal means.

the VOCs from the database. For each combination of them we generated 10 000 random samples $(Y_1 \ldots Y_N)$ from normal distribution $N(\mu,\sigma^2)$ together with realizations of $Z \sim N(0, 1)$ and $Q^2 \sim \chi^2_{N_j-1}$ which were used for evaluation of the generalized pivotal quantity T given by Eq. 7. Finally, the empirical quantiles T_l and T_u were used for evaluation of the 95% GCIs from Eq. 8 for the log-normal mean m. Results are showed in Fig. 1.

From the plotted results we see that the empirical coverage probability of the 95% GCIs for the true value of the log-normal mean m converges asymptotically (for large N) to the stated confidence level $(1 - \alpha) = 0.95$. Moreover, the empirical coverage probabilities are satisfactory also for small sample sizes.

Figure 2 shows the empirical coverage probabilities of the 95% GCIs for the ratio (m_1 / m_2) and the difference $(m_1 - m_2)$ of two log-normal means m_1 and m_2, obtained for parameters representing the selected VOC concentrations from our dataset. Similarly as before, for each selected VOC we generated 10 000 realizations of the GCIs from Eq. 11 and 12, respectively. From these results we see, that the empirical coverage probabilities of the 95% GCIs for the true values of (m_1 / m_2) and $(m_1 - m_2)$ are close to the stated 95% confidence level.

Studies in [3] compare generalized pivotal approach to construction of confidence interval for logarithm of log-normal mean with 4 other approaches (naive method, Cox's method, Angus's conservative method and Angus's bootstrap method). These simulation studies also compare generalized pivotal approach to construction of confidence intervals for the ratio and the difference of two log-normal means with 2 other approaches (maximum likelihood method, bootstrap method). GCIs are the best with respect to the empirical coverage probability (which is close to the stated confidence level 1 − α) and also with respect to the expected length of the considered confidence intervals (GCIs tend to be the shortest). The procedure (GCI) is applicable and satisfactory also for small sample sizes. Moreover, it is easy to compute and implement.

4. Results from breath analysis

Primarily, we are interested in compounds which have significant differences between population of patients with lung cancer and healthy volunteers. For that we have made the following statistical analysis.

We have compared the log-means of the selected VOC-concentrations (i.e. the means of the associated log-normal distributions) by generalized pivotal approach described in section 2. We have estimated the 95% generalized confidence intervals for the ratio and the difference of two log-normal means of exhaled-breath concentrations in patients vs. healthy volunteers.

In Figure 3 and Table 1 there are 95% GCIs for the ratio (GCI_R) and the difference (GCI_D) of log-normal means for each VOC, constructed by generalized pivotal approach. If the GCI_R contains value 1 we can say that two populations are similar and there are not significant differences. The same is valid if the GCI_D contains 0.

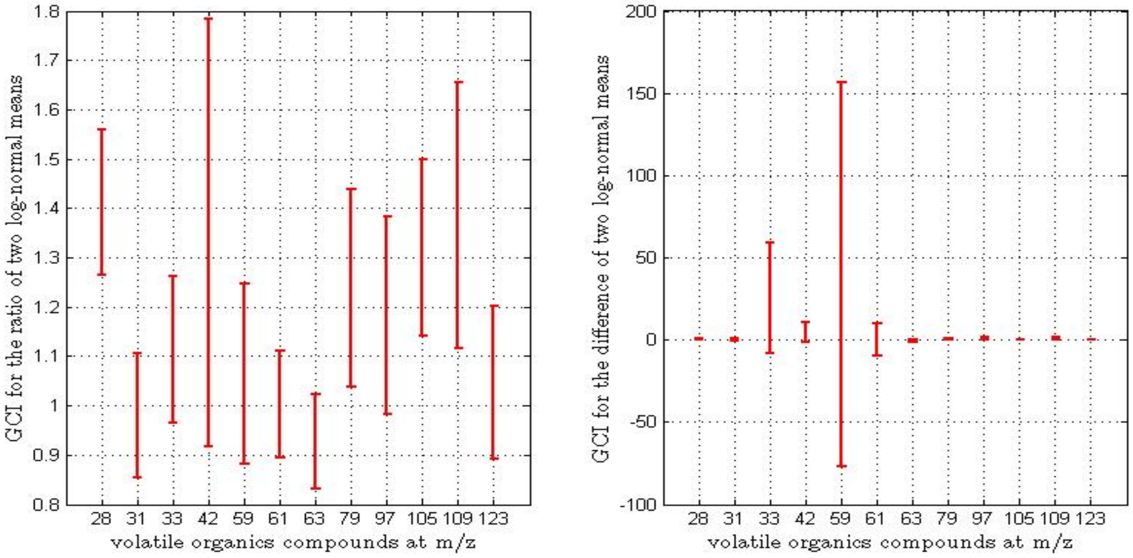

Fig. 3. Generalized confidence intervals for the ratio (left panel) and the difference (right panel) of two log-normal means of population of lung carcinoma patients (m_1) and of population of healthy test volunteers (m_2) for 12 volatile organic compounds.

Table 1. Generalized confidence intervals of the ratio (GCI_R) and the difference (GCI_D) of two log-normal means of population of lung carcinoma patients (m_1) and of population of healthy test volunteers (m_2) for 12 volatile organic compounds.

VOCs at m/z	28	31	33	42	59	61	63	79	97	105	107	123
GCI_R	1.26	0.85	0.96	0.92	0.88	0.90	0.83	1.04	0.98	1.14	1.12	0.89
	1.56	1.11	1.26	1.78	1.25	1.11	1.02	1.44	1.38	1.50	1.66	1.20
GCI_D	0.32	-0.86	-8.44	-1.19	-76.8	-9.60	-1.46	0.06	-0.06	0.06	0.30	-0.12
	0.66	0.57	59.38	10.87	156.3	9.62	0.17	0.65	1.35	0.20	1.50	0.20

Based on the generalized pivotal approach, the following volatile organics compounds have been identified as the compounds with significantly different log-normal means for the population of lung carcinoma patients and population of healthy test volunteers: m/z 28 (tentatively identified as hydrogen cyanide), m/z 79 (benzene), m/z 105 and m/z 109.

5. Discussion and conclusions

In this paper we have tried to compare the mean concentrations of the selected volatile organic compounds (VOCs) in exhaled breath of healthy volunteers and patients with lung cancer. For that, the small sample statistical properties (coverage probabilities) of the generalized confidence intervals proposed by Krishnamoorty and Mathew [4] have been studied and the generalized confidence intervals have been estimated for the ratio and the difference of breath-gas concentrations in two populations (sick patients vs. healthy volunteers).

Based on simulations, for construction of 95% GCIs for log-normal means and for construction of 95% GCIs for the ratio and the difference of two log-normal means we can recommend generalized pivotal approach. The GCIs have good statistical properties even for small samples. Moreover, the GCIs are easy to compute and implement. Coverage probability of confidence intervals constructed by generalized pivotal approach is stable for sample size, increasing variance and for different means. Coverage probability is between 93% and 95.5%.

The resulted confidence intervals are plotted in Fig. 3. The biggest differences between patients with lung cancer and healthy volunteers have been found at the compound m/z 28, tentatively identified as hydrogene cyanide). The mean concentration of this VOC is 1.26 to 1.56 times bigger for patients with lung cancer than the corresponding value for healthy volunteers.

Generalized pivotal approach can be helpful in finding relevant markers of lung cancer. A volatile organic compound for which the ratio and the difference of log-normal means are significant can be a substance whose detection indicates a particular disease state. However, to mark a substance as bio-marker of lung cancer more examination is necessary.

Acknowledgements

The paper was supported by grants VEGA1/3016/2006, VEGA2/7087/27, (APVV) RPEU-0008-06 and FP6 Project BAMOD (LSHC-CT-2005-019031 STREP).

References

[1] Amann, A., Smith, D., Breath Analysis for Clinical Diagnosis and Therapeutic Monitoring, *World Scientific, Singapure*, 2005

[2] Amann, A., Španel, P., Smith, D., Breath Analysis; the Approach Towards Clinical Applications, *MiniReviews in Medicinal Chemistry*, 7(2), 2007 , pp. 115-129

[3] Cimermanová, K., Statistical Methods and Algorithms to Research of Molecular Oriented Detection of Lung Cancer, Disertation Proposal (in Slovak), *Institute of Measurement Science, Slovak Academy of Sience, Bratislava*, 2007, http://www.um.sav.sk/sk/images/stories/dep03/doc/minimovka.pdf

[4] Krishnamoorty, K., Mathew, T., Inference on the Means of Lognormal Distributions Using Generalized P-values and Generalized Confidence Intervals, *Journal of Statistical Planning and Inference*, 115, 2003, pp. 103 – 120

[5] Kushch, I., et al., Compounds Enhanced in Exhaled Breath of Smokers as Determined in a Pilot Study Using PTR-MS, to appear

[6] Weerahandi, S., Exact Statistical Methods for Data Analysis., *Springer-Verlag*, *New York*, 1995, p. 146