UB Biostatistics Technical Report

2008:1 17

Empirical and parametrical likelihood interval estimations for populations containing many zero values with application to the "organochlorine pesticide exposure and risk of endometriosis" study

Le Kang[1,*,†], Albert Vexler[1], Lili Tian[1] and Germaine M. Buck Louis[2]

[1]*Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, U.S.A.*

[2]*Epidemiology Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, U.S.A.*

SUMMARY

Background: Endometriosis is a gynecologic disorder that occurs primarily in women of reproductive age. It is a disease exclusive to species that menstruate, such as humans and primates. The recent epidemiological literature suggests that hormonally active environmental chemicals such as Aldrin or beta-hexachlorocyclohexane (β-HCH) may be involved in the etiology of endometriosis, given that affected women are reported to have higher serum concentrations than unaffected women. However, many women will have no measurable concentration because they are either unexposed or because their measurements are below the limits of detection of laboratory equipments. These cases are recorded as

*Correspondence to: Le Kang, Department of Biostatistics, 249 Farber Hall, 3435 Main St. Bldg. 26, Buffalo, NY 14214-3000, U.S.A.

†E-mail: lekang@buffalo.edu

zero.

Methods: An observational study focusing on the relation between serum concentrations of polyhalogenated aromatic hydrocarbons such as PCBs and OCPs and diagnosis of endometriosis has been conducted. One of the goals in the study was to determine whether concentrations of Aldrin and β-HCH differed by disease states. The study comprised 84 patients of which 52 of them do not have disease. Among these 52 women without disease, 7 of them have detectable concentrations of Aldrin, and 8 of them have detectable concentrations of β-HCH. Among 32 women with disease, 5 have detectable concentrations of Aldrin, and 5 have detectable concentrations of β-HCH. To provide investigators with methodologies for assessing endometriosis risk when considerable zeros exist in a dataset, we extended the empirical likelihood techniques to construct confidence interval estimations for populations containing many zero values. Specifically, we proposed and examined the empirical likelihood interval estimations of the mean and mean difference via a broad Monte Carlo study that compares the proposed method with parametric techniques in the context of coverage probability.

Results: We conclude that the proposed method can be used as a very efficient statistical tool for confidence interval estimation. For the real data from "organochlorine pesticide exposure and risk of endometriosis" study, 95% empirical likelihood confidence intervals are $(0.000803, 0.004211)$ and $(0.000338, 0.003561)$ for the mean concentration of Aldrin for women without and with endometriosis respectively, and $(-0.001563, 0.003025)$ for mean Aldrin concentration difference; 95% empirical likelihood confidence intervals are $(0.000680, 0.003807)$ and $(0.000493, 0.005869)$ for the mean concentration of β-HCH for women without and with endometriosis respectively, and $(-0.003522, 0.002890)$ for mean β-HCH concentration difference.

Discussions: The empirical likelihood method for confidence interval estimations for populations containing many zero values is presented. We found no evidence of a clinically significant relation between Aldrin or β-HCH and endometriosis.

KEY WORDS: Aldrin, β-HCH; confidence interval estimation; coverage probability; endometriosis; empirical likelihood interval estimation; monte carlo; nonparametric interval estimation; parametric likelihood interval estimation

1. INTRODUCTION

Recent studies have suggested a possible etiologic role of persistent environmental chemicals, particularly hormonally active compounds such as dioxin, polychlorinated biphenyls, polychlorinated dibenzo-p-dioxins, and dibenzofurans [, , ,] in the development of endometriosis. Many of such compounds are reported to be higher among women with endometriosis than unaffected women. These findings have been corroborated in experimental rodent and primate models underscoring the importance of assessing this work in human populations [, ,].

Growing speculation suggests organochlorine pesticides (OCPs) also may be involved. Such pesticides are ubiquitous in the environment with most individuals having measurable concentrations (e.g., DDE). Other pesticides may be unique to select subgroups of the populations resulting in zero measurements for many individuals. We selected two such organochlorine pesticides Aldrin and β-HCH to develop methods for estimating the mean difference between women with and without endometriosis. A zero measurement of Aldrin or β-HCH could denote either the absence of exposure or the laboratory limits of detection for equipments such as gas chromatography with electron capture detection. A more complete description of this issue is presented elsewhere [,].

The objectives of this study is to extend the well-known empirical likelihood ratio techniques (e.g., Owen [ ]) to construct two-sample interval estimations for populations containing many zero values. We also apply parametric methods to analyze such data. Note that pesticide measurements need to be tested for parametric assumptions before applying any parametric

approach. To estimate mean concentration of Aldrin and β-HCH and their differences, two challenges need to be addressed: 1) a high proportion of women will have zero measurement and 2) nonzero concentrations are skewed to the right and may have log-normal distributions. This is a common scenario encountered by environmental epidemiologists and for which no well-developed methods for two-sample interval estimations currently exist. For this reason, investigators frequently ignore such chemical exposures despite their potential health effects. This critical data gap deserves further methodological investigations and serves as the impetus for study.

Commonly, the confidence interval estimation of the difference of means between two populations is built upon the Student t statistics that have asymptotic normal distribution. The actual coverage probability of the Student t interval might differ greatly from an expected coverage level, if biomarkers' measurements are not normally distributed. Given parametrical forms of biomarkers' measurement distributions, the parametric likelihood intervals are estimators commonly accurate. For the case with nonzero observations following a log-normal distribution, Zhou *et al.* [.] proposed a parametric likelihood ratio type confidence interval for the difference of means of independent populations that are assumed to satisfy a mixture model. Tian *et al.* [.,] proposed a novel approach using the concept of generalized variable for inferences on the mean of zero-inflated log-normal data. This method provides satisfactory solutions that are applicable to small samples. However, we must assume that there is prior knowledge related to parametric distributions of the nonzero part of data. Note that, testing for goodness-of-fit in these situations is a relevant problem. Furthermore, parameter estimation might be awkward if the number of nonzero observations is too small. Without the parametric assumptions, completely nonparametric statistical methods (e.g., the Wilcoxon-signed-rank

(WSR) interval) can be applied. However, we might lose the accuracy of estimation, since neither the model structure nor the principle of the likelihood methodology is utilized.

In this article, we assume that biomarkers Aldrin and β-HCH follow nonparametric mixture models. We employ the empirical likelihood methodology (e.g., Owen [, ,]) to develop a nonparametric likelihood approach for two-sample interval estimation which is proposed as a special case following the existing empirical likelihood ratio techniques based on one-sample populations (e.g., [,]). Note that, Liu *et al.* [ ] proposed an extension of the empirical likelihood methodology to a simple two-sample mean problem. The empirical likelihood intervals for the mean and mean difference of two independent populations with mass at zero are calculated and examined. We also apply the parametric likelihood interval estimation to compare with the proposed interval estimation as well as to analyze the data from the "organochlorine pesticides exposure and risk of endometriosis" study. We demonstrate results of a Monte Carlo study to confirm analytical propositions of this article, and to examine coverage probability related to the analyzed intervals.

2. METHODS

2.1. One-sample nonparametric interval estimation

Suppose we observe a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$. One can present the empirical likelihood ratio statistic in the form of

$$ELR(\mu_0) = \frac{\max_{p_i}\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1\}}{\max_{p_i}\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(x_i) = \mu_0\}} = \prod_{i=1}^{n} \frac{1}{np_i},$$

where $p_i = P(X = x_i)$, $i = 1, 2, \ldots, n$, satisfy the empirical constrain that corresponds to $Eg(\mathbf{X}) = \mu_0$ and maximize $\prod_{i=1}^{n} p_i$. g is a fixed function (e.g., Owen [ ]). Thus, the Lagrange

multipliers method yields $p_i = [n + \lambda (g(x_i) - \mu_0)]^{-1}$, where λ is a root of the equation

$$\sum_{i=1}^{n} \frac{g(x_i) - \mu_0}{n + \lambda(g(x_i) - \mu_0)} = 0.$$

Owen[ ] showed that $2 \log ELR(\mu_0)$ has asymptotic $\chi^2_{(1)}$ distribution as $n \to \infty$.

The limiting $\chi^2_{(1)}$ distribution implies that the $(1-\alpha)\,100\%$ empirical likelihood intervals for the true parameter μ_0 is $\left\{\mu : 2 \log ELR(\mu) < \chi^2_{(1),1-\alpha}\right\}$, where $\chi^2_{(1),1-\alpha}$ is the α-upper quantile of $\chi^2_{(1)}$ distribution.

We could easily extend this result to one-sample case with many zero values. Suppose we have a sample $\mathbf{U} = (U_1, U_2, \ldots, U_m)'$, which contains positive and zero value observations. Let $\delta_1 = P(g(U_i) = 0) > 0$ be the probability of taking zero values and $\mu_1 = Eg(\mathbf{U}) = (1 - \delta_1)E(g(\mathbf{U})|g(\mathbf{U}) > 0)$ be the population mean as well as (m_0, m_1) be the number of zero and nonzero observations in the sample. Also, we denote the nonzero observations as $g(x_1), ..., g(x_{m_1})$. In this case, one can present the empirical likelihood ratio statistic in the form of

$$ELR(\mu_1) = \frac{\max_{p_i^X, \delta_1} \left\{\delta_1^{m_0}(1-\delta_1)^{m_1} \prod_{i=1}^{m_1} p_i^X : \sum_{i=1}^{m_1} p_i^X = 1\right\}}{\max_{p_i^X, \delta_1} \left\{\delta_1^{m_0}(1-\delta_1)^{m_1} \prod_{i=1}^{m_1} p_i^X : \sum_{i=1}^{m_1} p_i^X = 1, (1-\delta_1)\sum_{i=1}^{m_1} p_i^X g(x_i) = \mu_1\right\}},$$

where p_i^X, $i = 1, 2, \ldots, m_1$, satisfy the empirical constrains corresponding to $Eg(\mathbf{U}) = \mu_1$ and maximize $\delta_1^{m_0}(1-\delta_1)^{m_1} \prod_{i=1}^{m_1} p_i^X$, again, g is a fixed function.

Chen and Qin [ ] proposed to ML-estimate δ_1 as $\hat{\delta}_1 = m_0/m$.

Thus, the Lagrange multipliers method yields $p_i^X = \left[m_1 + \lambda\left(g(x_i) - \frac{\mu_1}{1-\hat{\delta}_1}\right)\right]^{-1}$, where λ is a root of the equation

$$\sum_{i=1}^{m_1} \frac{g(x_i) - \frac{\mu_1}{1-\hat{\delta}_1}}{m_1 + \lambda\left(g(x_i) - \frac{\mu_1}{1-\hat{\delta}_1}\right)} = 0\,.$$

One can show that $2 \log ELR(\mu_1)$ has asymptotic $\chi^2_{(1)}$ distribution as $m \to \infty$, $m_1/m < \infty$. (See Appendix for details). This result implies that the $(1-\alpha)100\%$ empirical likelihood

intervals for the true parameter μ_1 is $\left\{\mu : 2\log ELR(\mu) < \chi^2_{(1),1-\alpha}\right\}$, where $\chi^2_{(1),1-\alpha}$ is the α-upper quantile of $\chi^2_{(1)}$ distribution.

When $g(\mathbf{U}) = \mathbf{U}$, the parameter μ_1 has a simple meaning of the expectation.

2.2. Two-sample nonparametric interval estimation

Suppose we have second sample $\mathbf{V} = (V_1, V_2, \ldots, V_n)'$. The general empirical constraint concerning $Eg(\mathbf{U}, \mathbf{V}) = \theta_0$ is $\sum_{i=1}^{m}\sum_{j=1}^{n} p_i p_j g(u_i, v_j) = \theta_0$, where $\sum_{i=1}^{m} p_i = 1$, $\sum_{j=1}^{n} p_j = 1$. Without loss of generality, consider a special case where $g(\mathbf{U}, \mathbf{V}) = \mathbf{U} - \mathbf{V}$, $Eg(\mathbf{U}, \mathbf{V}) = E\mathbf{U} - E\mathbf{V}$. In this case, we are interested in the mean difference between two populations.

The constraint $\sum_{i=1}^{m}\sum_{j=1}^{n} p_i p_j (u_i - v_j) = \sum_{i=1}^{m} p_i (u_i - \sum_{j=1}^{n} p_j v_j) = \sum_{i=1}^{m} p_i u_i - \sum_{j=1}^{n} p_j v_j = \theta_0$ is about the empirical mean difference between two populations.

In the similar manner to section 2.1, define $\delta_2 = P(V_j = 0) > 0$ to be the probability of taking zero values and $\mu_2 = E(\mathbf{V}) = (1 - \delta_2)E(\mathbf{V}|\mathbf{V} > 0)$ to be the population mean. Let (n_0, n_1) be the number of zero and nonzero observations in the second sample, $y_1, \ldots, y_{n_1}$ be the nonzero observations. The interval estimation of interest is

$$Eg(\mathbf{U}, \mathbf{V}) = \theta_0 = \mu_1 - \mu_2. \tag{1}$$

The two-sample empirical likelihood function based on the binomial likelihood for δ's and the empirical likelihood for μ's has the form of

$$EL(\theta) = \delta_1^{n_0}(1-\delta_1)^{n_1}\prod_{i=1}^{n_1} p_i^X \times \delta_2^{n_0}(1-\delta_2)^{n_1}\prod_{j=1}^{n_1} p_j^Y. \tag{2}$$

Thus the empirical likelihood ratio statistic is

$$ELR(\theta_0) = \frac{\max_{p_i^X, p_j^Y, \delta_1, \delta_2}\left\{EL(\theta) : \sum_{i=1}^{m_1} p_i^X = 1, \sum_{i=1}^{n_1} p_i^Y = 1\right\}}{\max_{p_i^X, p_j^Y, \delta_1, \delta 2}\left\{EL(\theta) : \sum_{i=1}^{m_1} p_i^X = 1, \sum_{j=1}^{n_1} p_j^Y = 1, (1-\delta_1)\sum_{i=1}^{m_1} p_i^X x_i - (1-\delta_2)\sum_{j=1}^{n_1} p_j^Y y_j = \theta_0\right\}}$$

One can show the following proposition.

Proposition 1 *The empirical likelihood ratio statistic* $2\log ELR(\theta_0)$ *converges in distribution to* $\chi^2_{(1)}$ *as* $m, n \to \infty$ *and* $m_1/m, n_1/n, m_1/n_1 < \infty$.

Proof is in Appendix.

This implies the $(1-\alpha)\,100\%$ empirical likelihood interval for true mean difference θ_0 in (.) is $\left\{\theta : 2\log ELR(\theta) < \chi^2_{(1),1-\alpha}\right\}$, where $\chi^2_{(1),1-\alpha}$ is the α-upper quantile of $\chi^2_{(1)}$ distribution.

3. THE DATASET RESULTS AND RELEVANT MONTE CARLO STUDY

Buck Louis *et al.* [ ] have conducted an observational study focusing on the relation between serum concentrations of polyhalogenated aromatic hydrocarbons such as PCBs and OCPs and diagnosis of endometriosis. One of the goals in the study was to determine whether concentrations of Aldrin and β-HCH differed by disease states.

The study comprised 84 patients of which 52 of them do not have disease. Among these 52 women without disease, 7 of them have detectable concentrations of Aldrin, and 8 of them have detectable concentrations of β-HCH. Among 32 women with disease, 5 have detectable concentrations of Aldrin, and 5 have detectable concentrations of β-HCH.

The sample means and sample variances for Aldrin are 0.001970 and 3.452579e-5 for women without disease, 0.001239 and 1.797720e-5 for women with disease; the sample means and sample variances for β-HCH are 0.001793 and 3.130589e-5 for women without disease, 0.002109 and 5.480369e-5 for women with disease. Using only the positive observations, we obtain sample means and sample variances for Aldrin are 0.014073 and 8.325109e-5 for women without disease, 0.007434 and 7.277152e-5 for women with disease; the sample means and sample variances for β-HCH are 0.011208 and 9.854232e-5 for women without disease, 0.012655

and 2.304925e-4 for women with disease. The estimated 95% empirical likelihood intervals for the mean concentration of Aldrin are $(0.000803, 0.004211)$ and $(0.000338, 0.003561)$ for women without and with endometriosis respectively, and $(-0.001563, 0.003025)$ for Aldrin mean concentration difference. Also, the estimated 95% empirical likelihood intervals for the mean concentration of β-HCH are $(0.000680, 0.003807)$ and $(0.000493, 0.005869)$ for women without and with endometriosis respectively, and $(-0.003522, 0.002890)$ for β-HCH mean concentration difference. Thus, we conclude that the average concentrations of Aldrin and β-HCH between women with and without disease are not statistically different.

We also present the results of Monte Carlo simulations conducted to compare the coverage probabilities of the proposed empirical likelihood intervals with those of the corresponding parametric likelihood intervals, classical Student-t intervals, and the nonparametric WSR intervals about the mean difference. Throughout the simulation study, a nominal coverage level is $1 - \alpha = 0.95$ and the number of Monte Carlo repetitions is 10000.

<Tables 1-2 Go Here>

The parameter settings of designs 1-5 in Table 1 were made to represent different situations of the null constraint of the proposed interval estimation. Design 1 corresponds to the situation where all samples values are strictly positive and follow normal distributions; Design 2-5 correspond to the situation where sample data contain both normal and zero values; Design 5 represents the parameter setting similar to our real example. The parameter settings of designs 6-10 in Table 2 were similar except that the nonzero sample values follow log-normal distributions. Generally speaking, the proposed empirical likelihood intervals have relatively satisfactory coverage probabilities.

4. DISCUSSIONS

In this article, we derived the nonparametric empirical likelihood interval estimation for populations containing many zero values. For the scenarios we considered, the proposed interval estimation is accurate even in unbalanced cases. Note that the proposed empirical likelihood interval estimation does not require parametric assumption, which is not easy to assume or test. However, under certain parameter settings, coverage probability tends to be liberal. Further study on how different parameter settings affecting the coverage probability is needed.

ACKNOWLEDGEMENTS

This work was supported in part with funding from the American Chemistry Council (ACC) and the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD). The opinions expressed are those of the authors and not necessarily of the National Institutes of Health.

REFERENCES

1. Mayani A, Barel S, Soback S, Almagor M. Dioxin concentrations in women with endometriosis. *Human Reproduction* 1997; **12**:373–375.
2. Heilier JF, Verougstraete V, Nackers F, Tonglet R, Donnez J, Lison D. Assessment of cadmium impregnation in women suffering from endometriosis: a preliminary study. *Toxicology Letters* 2004; **154**:89–93.
3. Buck Louis GM, Weiner JM, Whitcomb BW, Sperrazza R, Schisterman EF, Lobdell DT, Crickard K, Greizerstein H, Kostyniak PJ. Environmental PCB exposure and risk of endometriosis. *Human Reproduction* 2005; **20**:279–285.
4. Porpora MG, Ingelido AM, Domenico AD, Ferro A, Crobu M, Pallante D, Cardelli M, Cosmi EV, Felip ED. Increased levels of polychlorobiphenyls in Italian women with endometriosis. *Chemosphere* 2006; **63**:1361–1367.

5. Rier SE, Martin DC, Bowman RE, Dmowski WP, Becker JL. Endometriosis in rhesus monkeys (Macaca mulatta) following chronic exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin. *Fundamental and Applied Toxicology* 1993; **21**:433–441.

6. Cummings AM, Metcalf JL, Birnbaum L. Promotion of endometriosis by 2,3,7,8-tetrachlorodibenzo-p-dioxin in rats and mice: Time and dose dependence and species comparison. *Toxicology and Applied Pharmacology* 1996; **138**:131–139.

7. Birnbaum LS, Tuomisto J. Non-carcinogenic effects of TCDD in animals. *Food Additives and Contaminants* 2000; **17**:275–288.

8. Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* 2006; **163**:374–383.

9. Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* 2008; **64**:895–903.

10. Vexler A, Liu A, Schisterman EF. Efficient design and analysis of biospecimen with measurements subject to detection limit. *Biometrical Journal* 2006; **48**:780–791.

11. Owen A. *Empirical likelihood*. Chapman & Hall: London, 2001.

12. Zhou XH, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* 1999; **55**:645–651.

13. Zhou XH, Gao S, Hui SL. Methods for comparing the means of two independent log-normal samples. *Biometrics* 1997; **53**:1129–1135.

14. Tian L. Inferences on the mean of zero-inflated lognormal data: The generalized variable method. *Statistics in Medicine* 2005; **24**:3223–3232.

15. Tian L, Wu J. Confidence intervals of the mean for lognormal data containing zeros. *Biometrical Journal* 2006; **48**:149–156.

16. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988; **75**:237–249.

17. Owen A. Empirical likelihood for linear models. *The Annals of Statistics* 1991; **19**:1725–1747.

18. Chen J, Chen SY, Rao JNK. Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics* 2003; **31**:53–68.

19. Chen SX, Qin J. Empirical likelihood-based confidence intervals for data with possible zero observations. *Statistics and Probability Letters* 2003; **65**:29–37.

20. Liu Y, Zou C, Zhang R. Empirical likelihood for the two-sample mean problem. *Statistics and Probability*

Letters 2008; **78**:548–556.

Appendix

1. A Proof for One-sample Chi-square Asymptotic Distribution

$$\begin{aligned} ELR(\mu_1) &= \frac{\max_{p_i^X,\delta_1}\left\{\delta_1^{m_0}(1-\delta_1)^{m_1}\prod_{i=1}^{m_1}p_i^X : \sum_{i=1}^{m_1}p_i^X = 1\right\}}{\max_{p_i^X,\delta_1}\left\{\delta_1^{m_0}(1-\delta_1)^{m_1}\prod_{i=1}^{m_1}p_i^X : \sum_{i=1}^{m_1}p_i^X = 1, (1-\delta_1)\sum_{i=1}^{m_1}p_i^X g(x_i) = \mu_1\right\}} \\ &= \frac{\max_{p_i^X}\left\{\prod_{i=1}^{m_1}p_i^X : \sum_{i=1}^{m_1}p_i^X = 1\right\}}{\max_{p_i^X}\left\{\prod_{i=1}^{m_1}p_i^X : \sum_{i=1}^{m_1}p_i^X = 1, \frac{m_1}{m}\sum_{i=1}^{m_1}p_i^X g(x_i) = \mu_1\right\}} \\ &= \prod_{i=1}^{m_1}\frac{1}{m_1 p_i^X}. \end{aligned}$$

where $p_i^X, i = 1, 2, \ldots, m_1$, satisfy the empirical constrains and maximize $\prod_{i=1}^{m_1}p_i^X$. The Lagrange multipliers method yields $p_i^X = \frac{1}{m_1 + \lambda(g(x_i) - \frac{m}{m_1}\mu_1)}$, where λ is a root of the equation $\sum_{i=1}^{m_1}\frac{g(x_i) - \frac{m}{m_1}\mu_1}{m_1 + \lambda(g(x_i) - \frac{m}{m_1}\mu_1)} = 0$.

By Taylor expansion,

$$0 = \sum_{i=1}^{m_1}\frac{g(x_i) - \frac{m}{m_1}\mu_1}{m_1 + \lambda(g(x_i) - \frac{m}{m_1}\mu_1)} = \sum_{i=1}^{m_1}\frac{g(x_i) - \frac{m}{m_1}\mu_1}{1 + \frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1)} \approx \sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)\left(1 - \frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1)\right)$$

Thus, we could have $\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right) \approx \frac{\lambda}{m_1}\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)^2$ and $\lambda \approx \frac{m_1\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)}{\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)^2}$

Plug λ into p_i^X and further $ELR(\mu_1)$, we have

$$\begin{aligned} -2\log(ELR(\mu_1)) &= -2\sum_{i=1}^{m_1}\log\left(m_1 p_i^X\right) \\ &= -2\sum_{i=1}^{m_1}\log\frac{1}{1 + \frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1)} \\ &= 2\sum_{i=1}^{m_1}\log\left(1 + \frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1)\right) \\ &\approx 2\sum_{i=1}^{m_1}\left(\frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1) - \frac{1}{2}\left\{\frac{\lambda}{m_1}(g(x_i) - \frac{m}{m_1}\mu_1)\right\}^2\right) \quad \text{(Taylor Expansion)} \\ &\approx \frac{\left\{\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)\right\}^2}{\sum_{i=1}^{m_1}\left(g(x_i) - \frac{m}{m_1}\mu_1\right)^2} \longrightarrow \chi^2_{(1)} \end{aligned}$$

2. A Proof for Two-sample Chi-square Asymptotic Distribution

Introducing Lagrange multiplier λ's, the log-likelihood function without null constraint is

$$l = m_0\log\frac{m_0}{m} + m_1\log\frac{m_1}{m} + n_0\log\frac{n_0}{n} + n_1\log\frac{n_1}{n} + \sum_{i=1}^{m_1}\log p_i^X + \sum_{j=1}^{n_1}\log p_j^Y$$

$$- \lambda_1 \left(\sum_{i=1}^{m_1} p_i^X - 1 \right) - \lambda_2 \left(\sum_{j=1}^{n_1} p_j^Y - 1 \right)$$

Differentiating l with respect to p_i^X, we have $\lambda_1 = m_1$, $p_i^X = \frac{1}{m_1}$

Differentiating l with respect to p_j^Y, we have $\lambda_2 = n_1$, $p_j^Y = \frac{1}{n_1}$

With null constraint: $\left(1 - \hat{\delta}_1\right) \sum_{i=1}^{m_1} p_i^X X_i - \left(1 - \hat{\delta}_2\right) \sum_{j=1}^{n_1} p_j^Y Y_j = \theta_0 = 0$

Therefore, the log-likelihood function with null constraint is

$$l = m_0 \log \frac{m_0}{m} + m_1 \log \frac{m_1}{m} + n_0 \log \frac{n_0}{n} + n_1 \log \frac{n_1}{n} + \sum_{i=1}^{m_1} \log p_i^X + \sum_{j=1}^{n_1} \log p_j^Y$$
$$- \lambda_1 \left(\sum_{i=1}^{m_1} p_i^X \quad 1 \right) - \lambda_2 \left(\sum_{j=1}^{n_1} p_j^Y - 1 \right) \quad \lambda_3 \left(\frac{m_1}{m} \sum_{i=1}^{m_1} p_i^X (X_i - \mu_x) - \frac{n_1}{n} \sum_{j=1}^{n_1} p_j^Y (Y_j - \mu_y) \right)$$

where $\mu_x = \frac{m}{m_1}\mu$, $\mu_y = \frac{n}{n_1}\mu$, while μ is the common mean.

Still, differentiating l with respect to p_i^X, p_j^Y, we could have

$$\begin{cases} \dfrac{\partial l}{\partial p_i^X} = \dfrac{1}{p_i^X} - \lambda_1 - \lambda_3 \dfrac{m_1}{m}(X_i - \mu_x) \equiv 0 \text{ and } p_i^X = \dfrac{1}{\lambda_1 + \lambda_3 \frac{m_1}{m}(X_i \quad \mu_x)} \\ \dfrac{\partial l}{\partial p_j^Y} = \dfrac{1}{p_j^Y} - \lambda_2 + \lambda_3 \dfrac{n_1}{n}(Y_j - \mu_y) \equiv 0 \text{ and } p_j^Y = \dfrac{1}{\lambda_2 - \lambda_3 \frac{n_1}{n}(Y_j - \mu_y)} \end{cases}$$

Through some mathematical manipulation, we could get

$$\begin{cases} m_1 = \lambda_1 + \lambda_3 \dfrac{m_1}{m} \sum_{i=1}^{m_1} p_i^X (X_i - \mu_x) \\ n_1 = \lambda_2 - \lambda_3 \dfrac{n_1}{n} \sum_{j=1}^{n_1} p_j^Y (Y_j - \mu_y) \end{cases}$$, and it is easy to see that $m_1 + n_1 = \lambda_1 + \lambda_2$

Using the fact $\sum_{i=1}^{m_1} \frac{1}{\lambda_1 + \lambda_3 \frac{m_1}{m}(X_i - \mu_x)} = 1$, we obtain that $\sum_{i=1}^{m_1} \frac{m_1 \quad \lambda_1 - \lambda_3 \frac{m_1}{m}(X_i - \mu_x)}{\lambda_1 + \lambda_1 \frac{m_1}{m}(X_i \quad \mu_x)} = 0$,

which is equivalent to $\sum_{i=1}^{m_1} \frac{m_1 - \lambda_1 - \lambda_3 \frac{m_1}{m}(X_i - \mu_x)}{1 + \frac{\lambda_3}{\lambda_1} \frac{m_1}{m}(X_i - \mu_x)} = 0$

Apply the Taylor expansion about the denominator, we obtain

$$\sum_{i=1}^{m_1} \left[m_1 - \lambda_1 - \lambda_3 \frac{m_1}{m}(X_i \quad \mu_x) \right] \left[1 - \frac{\lambda_3}{\lambda_1} \frac{m_1}{m}(X_i \quad \mu_x) \right] = 0$$

Simplify the expression above will give us

$$m_1 - \lambda_1 - \frac{\lambda_3}{\lambda_1} \frac{m_1}{m} \sum_{i=1}^{m_1} (X_i \quad \mu_x) + \frac{\lambda_3^2}{\lambda_1} \frac{m_1}{m^2} \sum_{i=1}^{m_1} (X_i - \mu_x)^2 = 0 \tag{3}$$

In the same way, we could obtain that

$$n_1 - \lambda_2 + \frac{\lambda_3}{\lambda_2} \frac{n_1}{n} \sum_{j=1}^{n_1} (Y_j - \mu_y) + \frac{\lambda_3^2}{\lambda_2} \frac{n_1}{n^2} \sum_{j=1}^{n_1} (Y_j - \mu_y)^2 = 0 \tag{4}$$

From (3), we could get

$$\frac{\lambda_1}{m_1} - \frac{\lambda_1^2}{m_1^2} - \frac{\lambda_3}{m_1 m} \sum_{i=1}^{m_1} (X_i - \mu_x) + \frac{\lambda_3^2}{m_1 m^2} \sum_{i=1}^{m_1} (X_i - \mu_x)^2 = 0$$

Also, we may ignore the last two terms when $m_1 \to \infty$ and obtain $\lambda_1 \to m_1$

From (4), we could get the similar conclusion that when $n_1 \to \infty$, $\lambda_2 \to n_1$

Combine (3) and (4), we have

$$-\frac{\lambda_3}{\lambda_1}\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)+\frac{\lambda_3^2}{\lambda_1}\frac{m_1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\frac{\lambda_3}{\lambda_2}\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)+\frac{\lambda_3^2}{\lambda_2}\frac{n_1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2=0$$

Simplify the expression above, we could get

$$\left[-\frac{1}{\lambda_1}\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)+\frac{1}{\lambda_2}\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right]+\lambda_3\left[\frac{m_1}{\lambda_1 m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\frac{n_1}{\lambda_2 n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2\right]=0$$

which means

$$\lambda_3=\frac{\frac{1}{\lambda_1}\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)-\frac{1}{\lambda_2}\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)}{\frac{m_1}{\lambda_1 m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\frac{n_1}{\lambda_2 n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}-\frac{\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)-\psi\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)}{\frac{m_1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\psi\frac{n_1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}=\frac{\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)-\psi\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)}{\frac{m_1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\psi\frac{n_1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2},$$

where $\psi=\frac{\lambda_1}{\lambda_2}\longrightarrow\frac{m_1}{n_1}$

Now let us look back to empirical likelihood ratio statistic.

$$2\log\frac{\max_\theta EL(p_i^X,p_j^Y,\delta_1,\delta_2)}{\max_{\theta_0}EL(p_i^X,p_j^Y,\delta_1,\delta_2)}=2\log\left[\left(\frac{1}{m_1}\right)^{m_1}\left(\frac{1}{n_1}\right)^{n_1}\Big/\prod_{i=1}^{m_1}\frac{1}{\lambda_1+\lambda_3\frac{m_1}{m}(X_i-\mu_x)}\prod_{j=1}^{n_1}\frac{1}{\lambda_2-\lambda_3\frac{n_1}{n}(Y_j-\mu_y)}\right]$$

$$-2\log\left[\left(\frac{1}{m_1}\right)^{m_1}\left(\frac{1}{n_1}\right)^{n_1}\Big/\left(\frac{1}{\lambda_1}\right)^{m_1}\left(\frac{1}{\lambda_2}\right)^{n_1}\prod_{i=1}^{m_1}\frac{1}{1+\frac{\lambda_3}{\lambda_1}\frac{m_1}{m}(X_i-\mu_x)}\prod_{j=1}^{n_1}\frac{1}{1-\frac{\lambda_3}{\lambda_2}\frac{n_1}{n}(Y_j-\mu_y)}\right]$$

$$=2\sum_{i=1}^{m_1}\log\left(1+\frac{\lambda_3}{\lambda_1}\frac{m_1}{m}(X_i-\mu_x)\right)+2\sum_{j=1}^{n_1}\log\left(1-\frac{\lambda_3}{\lambda_2}\frac{n_1}{n}(Y_j-\mu_y)\right)$$

$$=2\sum_{i=1}^{m_1}\left(\frac{\lambda_3}{\lambda_1}\frac{m_1}{m}(X_i-\mu_x)-\frac{1}{2}\frac{\lambda_3^2}{\lambda_1^2}\frac{m_1^2}{m^2}(X_i-\mu_x)^2\right)-2\sum_{j=1}^{n_1}\left(\frac{\lambda_3}{\lambda_2}\frac{n_1}{n}(Y_j-\mu_y)+\frac{1}{2}\frac{\lambda_3^2}{\lambda_2^2}\frac{n_1^2}{n^2}(Y_j-\mu_y)^2\right)$$

$$+2\lambda_3\left(\frac{\sum_{i=1}^{m_1}(X_i-\mu_x)}{m}-\frac{\sum_{j=1}^{n_1}(Y_j-\mu_y)}{n}\right)-\lambda_3^2\left(\frac{\sum_{i=1}^{m_1}(X_i-\mu_x)^2}{m^2}+\frac{\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}{n^2}\right)$$

Plug λ_3 into the first term.

$$2\left(\frac{\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)-\psi\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)}{\frac{m_1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\psi\frac{n_1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}\right)\frac{1}{m}\left(\sum_{i=1}^{m_1}(X_i-\mu_x)-\frac{m}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right)\to 2\left\{\frac{\left(\sum_{i=1}^{m_1}(X_i-\mu_x)-\frac{m}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right)^2}{\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\left(\frac{m}{n}\right)^2\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}\right\}$$

Plug λ_3 into the second term.

$$\frac{\left(\frac{m_1}{m}\sum_{i=1}^{m_1}(X_i-\mu_x)-\psi\frac{n_1}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right)^2}{\left(\frac{m_1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\psi\frac{n_1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2\right)^2}\left(\frac{\sum_{i=1}^{m_1}(X_i-\mu_x)^2}{m^2}+\frac{\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}{n^2}\right)\to-\frac{\left(\sum_{i=1}^{m_1}(X_i-\mu_x)-\frac{m}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right)^2}{m^2\left(\frac{1}{m^2}\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\frac{1}{n^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2\right)}$$

Combine the above results together, we have

$$2\log\frac{\max_\theta EL(p_i^X,p_j^Y,\delta_1,\delta_2)}{\max_{\theta_0}EL(p_i^X,p_j^Y,\delta_1,\delta_2)}\to\frac{\left(\sum_{i=1}^{m_1}(X_i-\mu_x)-\frac{m}{n}\sum_{j=1}^{n_1}(Y_j-\mu_y)\right)^2}{\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\left(\frac{m}{n}\right)^2\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}=\frac{m_1^2\left(\bar{X}-\frac{m}{n}\frac{n_1}{m_1}\bar{Y}\right)^2}{\sum_{i=1}^{m_1}(X_i-\mu_x)^2+\left(\frac{m}{n}\right)^2\sum_{j=1}^{n_1}(Y_j-\mu_y)^2}$$

$$=\frac{m_1^2\left(\bar{X}-\frac{m}{n}\frac{n_1}{m_1}\bar{Y}\right)^2}{m_1^2\sum_{i=1}^{m_1}(X_i-\mu_x)^2/m_1^2+\left(\frac{m}{n}\right)^2\frac{n_1^2}{m_1^2}\sum_{j=1}^{n_1}(Y_j-\mu_y)^2/n_1^2}=\frac{\left(\bar{X}-\frac{m}{n}\frac{n_1}{m_1}\bar{Y}\right)^2}{\sigma_X^2/m_1+\left(\frac{m}{n}\right)^2\left(\frac{n_1}{m_1}\right)^2\sigma_Y^2/n_1}\to\chi^2_{(1)}\ \blacksquare$$

Table 1. Monte carlo coverage probability based on different designs (*Normal*)

Design	m	n	δ_1	μ_1	σ_1^2	δ_2	μ_2	σ_2^2	Monte Carlo Coverage Probability			
									EL	PL	Student-t	WSR
1	10	10	0.0	0.15	1.1	0.0	0.15	1.0	0.9319	0.9144	0.9501	0.9512
	12	12							0.9557	0.9448	0.9522	0.9549
	24	24							0.9532	0.9488	0.9472	0.9483
	48	48							0.9513	0.9515	0.9512	0.9501
2	8	8	0.1	0.1778	2.0	0.2	0.2	1.5	0.9338	0.9388	0.9516	0.9533
	10	10							0.9378	0.9392	0.9505	0.9546
	15	15							0.9400	0.9455	0.9499	0.9569
	20	20							0.9425	0.9498	0.9553	0.9578
3	9	12	0.3	0.4571	1.5	0.2	0.4	2.5	0.9079	0.9313	0.9597	0.9548
	9	24							0.9091	0.9379	0.9574	0.9552
	18	24							0.9454	0.9487	0.9535	0.9560
	18	36							0.9567	0.9505	0.9550	0.9555
4	6	8	0.2	0.3	1.0	0.25	0.32	1.2	0.9299	0.9303	0.9554	0.9626
	9	10							0.9407	0.9389	0.9573	0.9614
	12	12							0.9489	0.9470	0.9531	0.9603
	15	14							0.9511	0.9467	0.9516	0.9578
5	10	10	0.8	0.01	8e-5	0.8	0.01	8e-5	>0.9999	0.0091	>0.9999	0.9998
	20	20							>0.9999	0.0021	0.9982	0.9997
	40	40							0.9999	<0.0001	0.9965	0.9994
	80	80							0.9993	<0.0001	0.9952	0.9989

Table II. Monte carlo coverage probability based on different designs (*Log-Normal*)

Design	m	n	δ_1	μ_1	σ_1^2	δ_2	μ_2	σ_2^2	Monte Carlo Coverage Probability			
									EL	PL	Student-t	WSR
6	10	10	0.0	0.45	1.1	0.0	0.35	1.0	0.9307	0.9331	0.9674	0.9361
	12	12							0.9394	0.9363	0.9672	0.9295
	25	25							0.9449	0.9472	0.9623	0.8780
	50	50							0.9495	0.9493	0.9548	0.8549
7	8	8	0.1	0.2322	1.2	0.2	0.45	1.0	0.9183	0.9231	0.9858	0.9731
	10	10							0.9376	0.9299	0.9857	0.9711
	12	12							0.9422	0.9309	0.9863	0.9713
	15	15							0.9415	0.9335	0.9826	0.9695
8	9	12	0.2	0.7665	1.5	0.3	0.4	2.5	0.9246	0.9293	0.9851	0.9491
	12	18							0.9321	0.9390	0.9853	0.9347
	15	24							0.9430	0.9477	0.9817	0.9449
	18	30							0.9428	0.9495	0.9774	0.8965
9	6	8	0.3	0.4690	1.0	0.25	0.3	1.2	0.8366	0.8591	0.9979	0.9998
	9	12							0.8755	0.9088	0.9943	0.9996
	12	16							0.9112	0.9283	0.9905	0.9992
	15	20							0.9309	0.9333	0.9856	0.9991
10	10	10	0.8	0.01	8e-5	0.8	0.01	8e-5	>0.9999	0.0005	>0.9999	0.9998
	20	20							0.9997	0.0006	>0.9999	0.9998
	40	40							0.9989	0.0006	>0.9999	0.9996
	80	80							0.9967	<0.0001	>0.9999	0.9995