Published in final edited form as:

Stat Biopharm Res. 2014 January 1; 6(1): 30–40. doi:10.1080/19466315.2013.826597.

Empirical Likelihood Approaches to Two-Group Comparisons of Upper Quantiles Applied to Biomedical Data

Jihnhee Yu^{1,*}, Albert Vexler¹, Alan D. Hutson¹, and Heinz Baumann²

¹Department of Biostatistics, University at Buffalo, State University of New York, USA

²Department of Molecular and Cellular Biology, Roswell Park Cancer Institute, USA

Abstract

In many biomedical studies, a difference in *upper quantiles* is of specific interest since the upper quantile represents the upper range of biomarkers and/or is used as the cut-off value for a disease classification. In this article, we investigate two-group comparisons of an upper quantile based on the empirical likelihood methodology. Two approaches, the classical empirical likelihood and ‘plug-in’ empirical likelihood are used to construct the test statistics and their properties are theoretically investigated. Although the plug-in method is developed by the frame work of the empirical likelihood, the test statistic is not based on the maximization of the empirical likelihood, and is simplified by using indicator function in its construction, making it a unique test to investigate. Extensive simulation results demonstrate that the ‘plug-in’ empirical likelihood approach performs better to compare upper quantiles across various underlying distributions and sample sizes. For the actual application, we employ the developed methods to test the differences in upper quantiles in two different studies, the oral colonization of pneumonia pathogens for intensive care unit patients treated by two different oral treatments, and the biomarker expressions of normal and abnormal bronchial epithelial cells.

Keywords

Comparative effectiveness research; 0.9-quantile; 0.95-quantile; Quantile comparison; Reference range

1. Introduction

There are many advantages to using quantiles in biomedical research. Quantiles, particularly very high ones provide the relevant information regarding ranges of the outcomes. Investigation into outcome ranges is important in the context of the reference ranges of biomarkers, which provide criteria for classifying the normal and abnormal groups. Also, comparisons of upper quantiles are of great interest in the contexts of receiver operating characteristic curve analyses related to the sensitivity and specificity between different diagnostic tools or biomarkers (e.g., Zhou and Qin, 2005). The following examples from the available literature demonstrate the relevance of upper quantile research in medical practices.

1. Wener *et al.* (2000) investigated the upper reference limit of serum C-reactive protein (CRP) by different demographic groups. They showed that the sample 0.95-quantile value of CRP in the overall population was 0.95 mg/dl for males and 1.39

*Corresponding author: jinheeyu@buffalo.edu, phone: 1-716-829-6029, fax: 1-716-829-2200.

mg/dl for females and varied with age and race as well. They stated that the upper reference limits of CRP should consider demographic factors.

2. He *et al.* (2004) investigated the distribution of serum prostate-specific antigen in different ethnic groups. The 'normal' distribution of serum prostate-specific antigen (PSA) levels in healthy Chinese men was evaluated. They demonstrated a gradual increase of the sample median and 0.95-quantiles of serum PSA levels by age. He *et al.* used the 0.95-quantile for the comparison between groups since it was "appropriate upper limit" of the serum PSA concentration.

Oremek and Seiffert (1996) showed dramatic changes in PSA levels among men of various ages after 15 minutes of exercise. The investigators presented the sample 0.95-quantiles of PSA levels before and after exercise for practical usage. The reason of choosing 0.95-quantile to show PSA changes in different treatment groups was because of 0.95-quantile's "practical use in clinical management" (Oremek and Seiffert, 1996).

3. Boucai and Surks (2009) showed that reference limits for thyroid-stimulating hormone (TSH) differed between races and with age based on a cross-sectional study of an urban outpatient medical practice. Use of race- and age-specific reference limits was strongly recommended as such practices will decrease misclassification of patients with raised TSH.

Some other meaningful interpretations of upper quantiles are also found in the literature. For example, Castorina *et al.*, (2010) used the 0.95-quantile of metabolic levels since different 0.95-quantiles may indicate different levels of exposures potentially related to harmful substances (e.g., pesticides). In the view of showing group differences, we note that many researches in the medical-related areas consider the quantile differences as meaningful differences (e.g., Castorina *et al.*, 2010; Schull *et al.*, 2003) although the group classification may not be a result of randomization of a certain treatment.

The direct comparison of quantiles allows investigators to avoid the dichotomization of the observed values, i.e., grouping observations in terms of a cutoff point such as a quantile. Some loss of information related to dichotomization is well addressed in Fedorov *et al.* (2009).

In this article, we construct tests for two-group comparisons of upper quantiles based on the two empirical likelihood (EL) approaches, namely the classical EL approach (Owen, 1988; Owen, 2001) and 'plug-in' approach (e.g., Hjort *et al.*, 2009), in which the latter utilizes the separately obtained sample statistic for nuisance parameters to estimate the likelihood function. These EL approaches allow the incorporation of restrictions on parameters into the likelihood function estimation, without assuming an underlying distribution, and thus are intrinsically nonparametric methods.

The classical EL approach has demonstrated robust performances on inferences at various underlying distributions (e.g., Qin and Wong, 1996; Peng and Qi, 2006; Yu *et al.*, 2010; Vexler *et al.*, 2010). Some investigations regarding inference for quantiles based on the EL approach are found in the literature. Chen and Hall (1993) showed that smoothed EL confidence intervals show an excellent coverage rate. Chen and Chen (2000) investigated the properties of EL quantile estimation in large samples, Zhou and Jing (2003a) proposed an alternative smoothed EL approach in which the likelihood ratio has an explicit form based on the concept of M-estimators. Lopez *et al.* (2009) investigated testing general parameters that are determined by the expectation of non-smooth functions. Chen and Lazar (2010) investigated the quantile estimation for discrete distributions. Zhou and Jing (2003b)

investigated the confidence interval estimation of the quantile difference within one group e.g., interquartile range.

A few articles discussing plug-in approaches in the construction of EL are available in the literature (e.g., Qin and Jing, 2001; Wang and Jing, 2001; Li and Wang, 2003). In the context of one sample problem, Hjort *et al.* (2009) showed that the limit distribution of $-2 \log$ of the EL ratio (ELR) test is a sum of a weighted χ^2 distribution, where the weights are often intractable because of the complexity of the asymptotics. Recently, for median comparisons between two groups, Yu *et al.* (2011) obtained the analytic form of the asymptotic distribution of $-2 \log$ of the ELR test based on the plug-in EL approach, which has a weighted χ^2 distribution. They also showed that both classical and plug-in ELR tests outperform the other alternative tests (e.g., Wilcoxon rank sum test) in terms of the Type I error control when the underlying distributions of two groups are seemingly different, or put in other terms, violate the exchangeability assumption under the null hypothesis (Hutson, 2007).

We note that none of these available articles specifically discuss upper quantile comparisons. In this article, we consider that readers can be informed regarding the development of general quantile tests based on the EL approaches and their viability in testing upper quantiles such as the 0.9- or 0.95-quantiles. One of research questions in this paper is to compare plug-in and classical EL ratio statistics theoretically and experimentally, since this question is complicated in theoretical context. Commonly, there is no theoretical guarantee that the plug-in EL approach performs better asymptotically than the classical EL approach with the exception of some cases involving large number of estimating equations (Hjort *et al.*, 2009). The usage of plug-in methods may have some advantages over the classical EL approaches, e.g., the plug-in method may allow a more flexibility to construct the sample form of estimating equations comparing with the classical EL method (e.g., Qin and Zhou, 2006), and, in the calculation of the EL test statistic, the maximization of the EL to obtain the nuisance parameters can be avoided. The latter point may give rise to a more succinct form of the EL test statistic comparing with the classical EL test.

This article has the following structure. In Section 2 we develop tests for general quantile comparisons based on the classical EL and plug-in EL approaches. In Section 3 we investigate the performance of the proposed methods on testing differences in 0.9- and 0.95-quantiles via an extensive Monte-Carlo study. Also, the two developed methods are compared with a sample quantile-based test. In Section 4 we apply the proposed methods to analyze the number of pneumonia pathogens and biomarker expressions. Section 5 is devoted to concluding remarks.

2. Main results

For each group, $i = 1, 2$, there are n_i experimental units. Let X_{ij} denote the random variable of the j -th unit from group i with the continuous distribution function F_i , and satisfy $E|X_{ij}|^3 < \infty$, and F_i have at least two times differentiable in some neighborhood of the q -quantile ($0 < q < 1$). Also, let $f_i(x)$ and $f(x)$ indicate the density functions of group i and the pooled data, respectively, and F is the distribution function of the pooled data. We are interested in comparison of the q -quantile between the two treatment groups, where the q -quantile of group i is $F_i^{-1}(q) = \inf\{x: F_i(x) \geq q\}$. Let $F_i^{-1}(q) = \theta_i$. Under H_0 , we assume

$$H_0: \theta_1 = \theta_2 = \theta. \quad (1)$$

The ELR test is obtained as follows. The relevant likelihood function of X_{ij} is

$$\prod_{i=1}^2 \prod_{j=1}^{n_i} dF_i(x_{ij}). \quad (2)$$

Following the EL concept, (2) can be expressed in the nonparametric form of

$$\prod_{i=1}^2 \prod_{j=1}^{n_i} p_{ij}, \quad (3)$$

where p_{ij} represent the empirical probabilities replacing $dF_i(x_{ij})$ and satisfy the constraint

$$\sum_{j=1}^{n_1} p_{1j} = \sum_{j=1}^{n_2} p_{2j} = 1, 0 \leq p_{ij} \leq 1. \quad (4)$$

In addition to (4), under H_0 , the probabilities p_{ij} are obtained to maximize the EL function (3) subject to relevant constraints with respect to the hypothesis (1).

Using the definition of the quantile and the empirical probabilities, for each group i ($i = 1, 2$), we can establish the empirical equality as $\sum_{j=1}^{n_i} p_{ij} \Delta(\theta_i - X_{ij}) = q$ under H_0 , where $\Delta(x)$ is an appropriate function for obtaining the empirical distribution. One example of $\Delta(x)$ is an indicator function $I(x)$ defined as $I(x) = 0$ if $x < 0$ and $I(x)$ is 1 if $x \geq 0$. Following the classical EL approach, we have the empirical constraints consistent with (1) in a form of

$$\sum_{j=1}^{n_1} p_{1j} (\Delta(\tilde{\theta} - X_{1j}) - q) = 0, \text{ and } \sum_{j=1}^{n_2} p_{2j} (\Delta(\tilde{\theta} - X_{2j}) - q) = 0, \quad (5)$$

where, $\tilde{\theta}$ is the EL estimator of θ that maximizes (3). According to the constraints (4) and (5), the log of the likelihood function $\log \prod_{j=1}^{n_i} p_{ij}$ is maximized subject to

$$\sum_{j=1}^{n_i} p_{ij} \Delta(\tilde{\theta} - X_{ij}) = q, \sum_{j=1}^{n_i} p_{ij} = 1, 0 \leq p_{ij} \leq 1,$$

for each group i , $i = 1, 2$. Maximization can be achieved based on the Lagrange multiplier method through the function

$$\sum_{j=1}^{n_i} \log p_{ij} + \lambda_{i1} \left(1 - \sum_{j=1}^{n_i} p_{ij} \right) + \lambda_{i2} \left(q - \sum_{j=1}^{n_i} p_{ij} \Delta(\tilde{\theta} - X_{ij}) \right), \quad (6)$$

where λ_{ij} ($i, j = 1, 2$) are Lagrange multipliers. Let $\Delta(x) = I(x)$ in (6). Following the standard Lagrange multiplier solutions, we obtain

$$\lambda_{i1} = n_i - \lambda_{i2} q, p_{ij} = \left(n_i + \lambda_{i2} (\Delta(\tilde{\theta} - X_{ij}) - q) \right)^{-1}. \quad (7)$$

Using (7) and the constraint (5), we obtain

$$\lambda_{i2} = (r_i / q - n_i) / (1 - q),$$

where $r_i = n_i \hat{F}_i(\tilde{\theta}) = \sum_{j=1}^{n_i} \Delta(\tilde{\theta} - X_{ij})$. For the EL function under the alternative hypothesis $H_1: \theta_1 \neq \theta_2$, we impose the restriction of (4) only to maximize the EL function, resulting in $L = \prod_{i=1}^2 \prod_{j=1}^{n_i} 1/n_i = n_1^{-n_1} n_2^{-n_2}$. Consequently, we obtain $-2 \log$ of the ELR test,

$$-2\log R = -2 \left[\sum_{i=1}^2 n_i \log n_i - (n_i - r_i) \log \left(n_i - \frac{q}{1-q} (r_i/q - n_i) \right) - r_i \log (r_i/q) \right]. \quad (8)$$

Then, we have the following result.

PROPOSITION 1: Under H_0 and some regularity conditions, $-2 \log R$ in (8) converges in distribution to χ_1^2 distribution as $n_i \rightarrow \infty$, $i = 1, 2$.

Proposition 1 can be proven by applying the results from Lopez *et al.* (2009). In actual applications of the test statistic (8), the function $\Delta(x)$ can be a smoothed function using the kernel method (Nadaraya, 1964; Azzalini, 1981) to obtain r_i . Specifically, for

$\sum_{j=1}^{n_i} \Delta(\tilde{\theta} - X_{ij})$, we can use $\sum_{j=1}^{n_i} K((\tilde{\theta} - X_{ij})/h)$, where $K(u) = \int_{-\infty}^u k(u) du$, and k is a nonnegative, differentiable function satisfying $\int_{-\infty}^{\infty} k(u) du = 1$, $\int_{-\infty}^{\infty} |u| k(u) du < \infty$, and $\int_{-\infty}^{\infty} |k'(u)| du < \infty$ and h is a bandwidth. It has been shown that the performance of the smoothed version of the ELR test can be improved in terms of the Type I error and power comparing the ELR test based on the identity function (e.g., Zhou and Jing, 2003; Yu *et al.*, 2011). For the EL confidence interval, Chen and Hall (1993) showed that the smoothed EL confidence interval for quantiles has faster convergence in terms of the coverage error. Thus, in this article, we primarily use the smoothed function for $\Delta(x)$. The bandwidth h is commonly a function of the sample size and other parameters that are estimated based on the sample (e.g., Altman and Léger, 1995). An extensive Monte-Carlo study demonstrated that the proposed tests are robust to the choice of different bandwidths; however, in the context of the approach of Hyndman and Yao (2002), we chose a bandwidth of $h = 0.2n_i^{-1/6}$ for group i for actual applications and simulation studies, which showed empirically reasonable performances among many available methodologies.

We now propose a test statistic based on the plug-in EL approach. The plug-in EL approach simply replaces the EL estimator θ by the sample q -quantile estimator $\hat{\theta}$ based on the pooled sample, and subsequently we define

$$r_i = n_i \hat{F}_i(\hat{\theta}) = \sum_{j=1}^{n_i} \Delta(\hat{\theta} - X_{ij}). \quad (9)$$

This replacement gives rise to a change in the asymptotic distribution of the test statistic given in (8). Although the plug-in method is developed by the frame work of the empirical likelihood, the test statistic is not based on the maximization of the empirical likelihood, and is simplified by using indicator function in the development of (8), making it a unique test to investigate. We can show the following result.

PROPOSITION 2: When $n_i \rightarrow \infty$, $i = 1, 2$ and $n_1 / (n_1 + n_2) \rightarrow \eta$, under H_0 and given some regularity conditions, the statistic $-2 \log R / \nu$ based on (8) with r_i in (9) converges in

distribution to χ_1^2 distribution, where $\nu = \eta \frac{1}{f(\theta)} \left(f_1(\theta)^2 + f_2(\theta)^2 \left(\frac{1}{\eta} - 1 \right) \right)$.

We provide a sketch of the proof of Proposition 2 in Appendix. In the application of Proposition 2, the density functions need to be estimated using the kernel density estimation. In the next section, we show the performance of these developed test statistics based on an extensive Monte Carlo study.

3. Simulation study

We investigated the performance of each developed method to compare upper quantiles (0.9 and 0.95). Various combinations of underlying distributions which reflect a violation of the exchangeability assumption under H_0 were used for the simulation. Figure 1 shows some examples that we investigated for 0.95-quantile comparisons, where each plot describes two groups that satisfy the null hypothesis (i.e., the same 0.95-quantiles), but have vastly different underlying distributions. In this simulation, we compared the proposed methods to the test statistic using the sample quantile constructed in the following way. Let $\hat{\theta}_i$ and $\hat{\theta}$ indicate sample q -quantiles for the group i and the pooled sample, $\hat{f}(x)$ indicate the estimated density at x with the pooled sample. Based on the asymptotic distribution of $\hat{\theta}_i$ (Serfling, 1980),

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2)}{\sqrt{q(1-q)} \sqrt{\left(n_1(\hat{f}(\hat{\theta}))^2\right)^{-1} + \left(n_2(\hat{f}(\hat{\theta}))^2\right)^{-1}}}, \quad (10)$$

which asymptotically has the standard normal distribution under $H_0 : \theta_1 = \theta_2$ (henceforth, referred as to the sample quantile test). Note that we can construct a few other similar test statistics based on the asymptotic concept of the sample quantiles; we choose to use (10) since, through extensive simulations, the test statistic (10) performs best among the potential test statistics based on the sample quantiles. We also note that we examined the permutation test based on the difference of the sample quantiles, which gave uncontrolled Type I errors as expected (e.g., for Normal(0,1) vs. Normal(-3.2871,9) with $n_1 = n_2 = 200$, simulated Type I error of 0.2038).

The simulated Type I errors for the proposed methods to compare 0.95- and 0.9-quantiles are shown in Tables 1 and 2, respectively (significance level = 0.05). Each scenario presented in tables is based on 5,000 simulations. In the tables, parameters and distributions are chosen to demonstrate the performance of the proposed tests in the comparison of two considerably different distributions. For example, in Tables 1 and 2, we compare the case of the normal distribution with variance 1 to that with variance 9. In order to have matching 0.95-quantiles, the mean of -3.2871 is chosen for the normal distribution with variance 9 (Table 1). Likewise, the mean of -2.5631 for the normal distribution with the variance 9 is chosen to have matching 0.9-quantile (Table 2). Based on these parameters of null distributions, some differences are added to the parameters to investigate the power properties in Tables 3 and 4. We also remark that we choose parameters that are relatively simple to present but also produce sufficiently different distributions. For example, in Table 1, one lognormal distribution has the location parameter 1 while the other lognormal distribution has the scale parameter 1. The rest of the parameters are chosen in an effort to make two distributions have matching quantiles. Note that Lognormal (0, 2.25) has the variance of 80.53 while Lognormal (0.8224, 1) has the variance of 24.19, showing sizable variability differences.

With the relatively small sample sizes (i.e., $n_1 = n_2 = 15$ or 30), the sample quantile test was not reliable as many simulated Type I errors are much higher than 0.05. On the contrary, although it is not consistent, the classical EL has overall much lower simulated Type I errors than 0.05 as too few were rejected in some scenarios. This tendency becomes more severe with the higher quantile (i.e., 0.95-quantile) comparisons. Note that both the sample quantile test and the classical ELR test improve their performances when the sample size increases. On the other hand, the plug-in ELR test shows viable Type I error control even with the small sample sizes. In fact, it has stable Type I error controls for various underlying

distributions and sample sizes, indicating that the plug-in ELR test can be used reliably to test upper quantiles with unknown underlying distributions.

The Monte Carlo study to investigate the power is presented in Tables 3 and 4 for 0.95- and 0.9-quantiles, respectively. The plug-in ELR test shows the best simulated power with the relatively small sample sizes. The differences in the power between the two ELR methods decrease when the sample sizes increase, showing that, with a large sample size, the two methods are comparable. The sample quantile test does not have a consistently lower or higher power than the ELR tests, but caution is required when using the sample quantile test with relatively small sample sizes since the Type I error is often not well controlled as shown in Tables 1 and 2.

In large sample cases (e.g., $n_1 = n_2 = 200$), we also note that the EL approaches may tend to provide better powers than the sample quantile test if extreme quantiles are of interest (e.g., 0.95-quantile). In order to demonstrate this point better, we carried out the additional comparisons of 0.925-quantile and 0.8-quantile with the lognormal case in Table 3 (i.e., Lognormal (0, 2.25) vs. Lognormal (1.3224, 1)). When compared 0.8-quantile, the simulation results showed that the sample quantile test had the simulated power, 0.1996 while the classical EL and plug-in EL methods showed the power of 0.1544 and 0.1546, respectively. However, when compared 0.925-quantile, the simulated power of the sample quantile test was 0.7748 while the classical EL and plug-in methods showed the power of 0.9166 and 0.9148, respectively.

It is noteworthy that the classical ELR test often lacks a power to detect the difference with the relatively small sample sizes, although this lack of power improves when the lower quantile (i.e., 0.9-quantile) is compared.

Overall, when comparing upper quantiles, the simulation results show that the plug-in ELR performs very well with various underlying distributions and sample sizes.

We remark that, for the density estimation in application of Proposition 2, R function “density” is used with Silverman’s rule of thumb as the choice of the bandwidth method, a default of the R’s density estimation function. R’s density estimation function also provides other bandwidth choices such as the method proposed by Sheather and Jones (1991), which shows overall good performances and generally recommended for bandwidth choices (Sheather, 2004; R development Core Team, 2011). We note that, when we used Sheather-Jones plug-in bandwidth, we observed an improvement of the Type I errors for plug-in EL methods in the case of Normal(0, 1) vs. Normal(−3.2871, 9) (e.g., the simulated Type I error was 0.0510 for the sample sizes of $n_1 = n_2 = 200$). In light of this observation and the general recommendation for bandwidth choices, we remark that the density estimation based on Sheather-Jones plug-in bandwidth may provide more robust Type I errors for the testing procedure using the plug-in EL method comparing with the default bandwidth choice in R’s density function (i.e., Silverman’s rule of thumb approach). In the other cases presented in the tables, the accuracies of the Type I errors and the power properties were consistent whether we used the method of Sheather and Jones or the default bandwidth choice.

4. Applications

4.1 Oral health data

The data is obtained from Oral Health and Ventilator-Associated Pneumonia study (OHP study) that was performed during the period of 12/1/2004 – 11/31/2007. The primary objective of the OHP study was to determine the effects of oral decontamination with chlorhexidine (CHX) on reducing the colonization of potential respiratory pathogens in the

oral cavity. The target group for this study was patients admitted to an intensive care unit (ICU) who were mechanically ventilated. A total of 175 patients met the eligibility requirements and consented to participate in the study. These 175 patients were randomized into three treatment groups (control: 59, CHX once: 58, CHX twice: 58). Primary analysis based on mean comparisons (ANOVA) was completed and the total numbers of colony forming units per ml for the potential pathogens were not reduced by CHX at any time point (Scannapieco *et al.* 2009) except *S. aureus*.

Here, we revisit the problem, and investigate the early colonization (at day 6) of potential respiratory pathogens among the ICU patients. The aggregated and log-transformed values of the potential pathogens (*S. aureus*, *P. aeruginosa*, *Acinetobacter* spp., and enteric organisms) are obtained as the study originally intended (Scannapieco *et al.* 2009). We carry out the 0.95-quantile comparisons between the control group and CHX treated group. The comparison of the upper quantile is meaningful since a highly concentrated colonization of specific pathogens is evidence of lung infection, which may not be properly tested based on the center of the distributions such as the mean or median. The final sample includes 23 controls and 51 CHX-treated patients. Visual inspection of the histogram and box plots for the two groups shows that the distributions between the two groups are different (Figure 2). Also note that the standard deviations for the control and CHX groups were 4.92 and 5.67, respectively and the inter-quartile ranges for the control and CHX groups were 6.28 and 12.22, respectively, indicating the differences in the distributions. The sample 0.95-quantiles for control and CHX groups are 14.70 and 15.35, suggesting not much difference in those quantiles. The classical ELR test statistic was 1.097 (p-value = 0.478) and the plug-in ELR test was 0.503 (p-value = 0.295). The both test results fail to reject the null hypothesis, indicating that there is not enough evidence of different 0.95-quantiles between the two treatment groups.

4.2 Biomarker expressions in a lung cancer study

Cells with constant exposure to irritants and pathogens may be damaged and have altered inhibitory mechanisms to suppress cell proliferation. Loewen *et al.* (2005) investigated that the early stage premalignant lesions from bronchial epithelial brushings might have altered cell proliferation resulting in the promotion of tumorigenesis. To test this, a cytokine called oncostatin M (OSM) was applied to both normal and abnormal (metaplastic or dysplastic) lung cell cultures and phosphorylation of various biomarkers (ERK, STAT1 and STAT3) was measured using Western blot analysis. The boxplots of quantified responses of biomarkers after the cytokine treatment are shown in Figure 3. The values of each biomarker to construct the boxplot are relative standing in phosphorylation in the abnormal or normal cells compared to its matched normal cells' OSM receptor expressions from the same subject. The figure demonstrates that the underlying distributions of normal and abnormal cells for each biomarker are different. We note that not all analyses are based on the same subject numbers due to incomplete biomarker evaluations in some cells. The subject numbers for constructing the boxplots and the data analysis are shown in Figure 3. It is suspected that an increased sensitivity of cytokines which demonstrates the growth-stimulatory activity would be observed in upper quantiles rather than center of the distribution. Thus, we compare 0.9-quantiles of normal and abnormal cells. Using the classical ELR test, the p-values for ERK, STAT1 and STAT3 were 0.0511, 0.5138 and 0.7539. Using the plug-in ELR test, the p-values for ERK, STAT1 and STAT3 were 0.0451, 0.3954 and 0.3821. The both ELR tests present some degree of evidence that there is a significant 0.9-quantile difference with the response of ERK. In particular, the plug-in ELR test result was significant at the level of 0.05. Note that, as discussed in Simulation Study section, with the small sample sizes such as this study, the plug-in ER ratio test provides

more reliable results. This result does indeed support the conclusion that abnormal bronchial epithelial cells have a higher probability for increase signaling toward the ERK pathway.

5. Conclusion

We proposed the two ELR tests to compare the general quantiles, and investigated their performance in the comparisons of upper quantiles. Although the plug-in approach uses the same EL principle as the classical ELR test to maximize the nonparametric likelihood function, we showed that the resulting distribution of the plug-in approach was different from that of the classical ELR test. Also, the plug-in ELR test shows well controlled Type I errors and better (or comparable) powers even with relatively small sample sizes, where the other tests were unstable. Thus, we recommend the plug-in ELR test to compare upper quantiles between groups.

Acknowledgments

This work was financially supported by a grant from the National Institute of Health, National Institute of Dental and Craniofacial Research in USA (Grant No. 1R03DE020851 - 01A1). The authors are grateful to the Editor and the referees for suggestions that led to a substantial improvement in this paper.

Appendix

Proof of Proposition 2: The plug-in ELR statistic can be obtained similarly to (8) using the indicator function, and has the form of

$$-2\log R = -2 \left[\sum_{i=1}^2 n_i \left\{ (1 - \hat{F}_i(\hat{\theta})) \log(1 - q) + \hat{F}_i(\hat{\theta}) \log(q) - \hat{F}_i(\hat{\theta}) \log(\hat{F}_i(\hat{\theta})) - (1 - \hat{F}_i(\hat{\theta})) \log(1 - \hat{F}_i(\hat{\theta})) \right\} \right]. \quad (\text{A-1})$$

Since

$$\hat{F}_i(\hat{\theta}) - q = \left\{ \hat{F}_i(\hat{\theta}) - \hat{F}_i(\theta) - (F_i(\hat{\theta}) - q) \right\} - 2q + \hat{F}_i(\theta) + F_i(\hat{\theta}), \quad (\text{A-2})$$

where the first curly bracket has the order of $o(n_2^{-1/2+\alpha})$ for some $\alpha > 0$, and

$\hat{F}_i(\theta) = F(\theta) + o(n_i^{-1/2+\alpha})$ (Winter, 1979) and $F_i(\hat{\theta}) = F(\theta) + o(n_i^{-1/2+\alpha})$ (Serfling, 1980) under H_0 , we have

$$\hat{F}_1(\hat{\theta}) = \left((n_1 + n_2)q - n_2 \hat{F}_2(\hat{\theta}) \right) / n_1 + o(n_*^{-1/2+\alpha}), \quad (\text{A-3})$$

where n_* is the minimum of n_1 and n_2 . Using (A-1), (A-3) and the Taylor's expansion up to the order $(\hat{F}_1(\hat{\theta}) - q)^2$, we have

$$-2\log LR = (n_1 + n_2)(n_1 q(1 - q))^{-1} n_2 \left(\hat{F}_2(\hat{\theta}) - q \right)^2 + o(n_*^{-1/2+\alpha}). \quad (\text{A-4})$$

Using the relationship (A-2) and approximating $F_2(\hat{\theta}) - q$ by $(\hat{\theta} - \theta)f_2(\theta)$, we have

$$\hat{F}_2(\hat{\theta}) - q = \hat{F}_2(\theta) - q + f_2(\theta)(\hat{\theta} - \theta) + o(n_2^{-1/2+\alpha}). \quad (\text{A-5})$$

Based on the result in Azzalini (1981), we substitute $(\hat{\theta} - \theta)$ with $(q(n_1 + n_2) - n_1 \hat{F}_1(\hat{\theta}) - n_2 \hat{F}_2(\hat{\theta})) / ((n_1 + n_2)f(\theta))$, then (A-5) can be expressed as

$$\hat{F}_2(\hat{\theta}) - q = n_2^{-1/2} \left[\left(1 - (1 - \eta) \frac{f_2(\theta)}{f(\theta)} \right) (\hat{F}_2(\theta) - q) - \left(\frac{f_2(\theta)}{f(\theta)} \eta (\hat{F}_1(\theta) - q) \right) \right] + o(n_2^{-1/2+\alpha}). \quad (\text{A-6})$$

Based on equations (A-4) and (A-6), the variance of $F_i(\hat{\theta})$ in Azzalini (1981), and applying the central limit theorem, the result follows.

References

- Altman N, Legér C. Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*. 1995; 46:195–214.
- Azzalini A. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*. 1981; 68:326–328.
- Boucai L, Surks MI. Reference limits of serum TSH and free T4 are significantly influenced by race and age in an urban outpatient medical practice. *Clinical Endocrinology*. 2009; 70:788–793. [PubMed: 18727705]
- Castorina R, Bradman A, Fenster L, Barr DB, Bravo R, Vedar MG, Harnly ME, McKone TE, Eisen EA, Eskenazi B. Comparison of current-use pesticide and other toxicant urinary metabolite levels among pregnant women in the CHAMACOS cohort and NHANES. *Environmental Health Perspectives*. 2010; 118:856–863. [PubMed: 20129873]
- Chen H, Chen J. Bahadur representations of the empirical likelihood quantile processes. *Nonparametric Statistics*. 2000; 12:645–660.
- Chen J, Lazar NA. Quantile estimation for discrete data via empirical likelihood. *Journal of Nonparametric Statistics*. 2010; 22:237–255.
- Chen SX, Hall P. Smoothed Empirical Likelihood Confidence Intervals for Quantiles. *The Annals of Statistics*. 1993; 21:1166–1181.
- Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharmaceutical Statistics*. 2009; 8:50–61. [PubMed: 18389492]
- He D, Wang M, Chen X, Gao Z, He H, Zhau HE, Wang W, Chung LW, Nan X. Ethnic differences in distribution of serum prostate-specific antigen: a study in a healthy Chinese male population. *Urology*. 2004; 63:722–726. [PubMed: 15072888]
- Hjort NL, McKeague IW, Van Keilegom I. Extending the scope of empirical likelihood. *The Annals of Statistics*. 2009; 37:1079–1111.
- Hutson AD. An “exact” two-group median test with an extension to censored data. *Journal of Nonparametric Statistics*. 2007; 19:103–112.
- Hyndman RJ, Yao Q. Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics*. 2002; 14:259–278.
- Lopez EMM, Van Keilegom I, Veraverbeke N. Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics*. 2009; 36:413–432.
- Li G, Wang QH. Empirical likelihood regression analysis for right censored data. *Statistica Sinica*. 2003; 13:51–68.
- Nadaraya EA. Some new estimates for distribution function. *Theory of Probability and its Applications*. 1964; 15:497–500.
- Oremek GM, Seiffert UB. Physical activity releases prostate-specific antigen (PSA) from the prostate gland into blood and increases serum PSA concentrations. *Clinical Chemistry*. 1996; 42:691–695. [PubMed: 8653893]
- Owen A. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988; 75:237–249.
- Owen, A. *Empirical Likelihood*. New York: Chapman & Hall; 2001.
- Peng L, Qi Y. Confidence regions for high quantiles of a heavy tailed distribution. *The Annals of Statistics*. 2006; 34:1964–1986.
- Qin G, Jing BY. Empirical likelihood for censored linear regression. *Scandinavian Journal of Statistics*. 2001; 28:661–673.

- Qin J, Wong A. Empirical likelihood in a semi-parametric model. *Scandinavian Journal of Statistics*. 1996; 23:209–219.
- Qin GS, Zhou XH. Empirical likelihood inference for the area under the ROC curve. *Biometrics*. 2006; 62:613–622. [PubMed: 16918927]
- R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Scannapieco FA, Yu J, Raghavendran K, Vacanti A, Owens SI, Wood K, Mylotte JM. A randomized trial of chlorhexidine gluconate on oral bacterial pathogens in mechanically ventilated patients. *Critical Care*. 2009; 13(4):R117. [PubMed: 19765321]
- Schull MJ, Morrison LJ, Vermeulen M, Redelmeier DA. Emergency department overcrowding and ambulance transport delays for patients with chest pain. *Canadian Medical Association Journal*. 2003; 168:277–283. [PubMed: 12566332]
- Serfling, R. Approximation theorems of mathematical statistics. New York: John Wiley and. Sons; 1980.
- Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*. 1991; 53:683–690.
- Silverman, BW. Density Estimation for Statistics and Data Analysis. London: Chapman & Hall; 1986.
- Vexler A, Yu J, Tian L, Liu S. Two-sample nonparametric likelihood inference based on incomplete data with an application to a pneumonia study. *Biometrical Journal*. 2010; 52:348–361. [PubMed: 20533413]
- Wang QH, Jing BY. Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*. 2001; 53:517–527.
- Wener MH, Daum PR, McQuillan GM. The influence of age, sex, and race on the upper reference limit of serum C-reactive protein concentration. *The Journal of Rheumatology*. 2000; 27:2351–2359. [PubMed: 11036829]
- Winter BB. Convergence rate of perturbed empirical distribution functions. *The Journal of Applied Probability*. 1979; 16:163–173.
- Yu J, Vexler A, Kim SE, Hutson AD. Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluations. *The Canadian Journal of Statistics*. 2011; 39:671–689.
- Yu J, Vexler A, Tian L. Analyzing incomplete data subject to a threshold using empirical likelihood methods: an application to a pneumonia risk study in an ICU setting. *Biometrics*. 2010; 66:123–130. [PubMed: 19432776]
- Zhou W, Jing BY. Adjusted empirical likelihood method for quantiles. *Annals of the Institute of Statistical Mathematics*. 2003a; 55:689–703.
- Zhou W, Jing BY. Smoothed empirical likelihood confidence intervals for the difference of quantiles. *Statistica Sinica*. 2003b; 13:83–95.
- Zhou XH, Qin G. Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in Medicine*. 2005; 24:465–477. [PubMed: 15635678]

Figure 1.

Descriptions of the same 0.95-quantiles with different underlying distributions. The vertical dotted lines indicate the 0.95-quantiles.

Figure 2. The histograms and boxplots for the distributions of pathogens of control (group 1, top) and CHX treated groups (group 2, bottom) in the OHP study.

Figure 3.

The boxplots of biomarker expressions to compare normal and abnormal groups (left and right plots for each biomarker) for ERK, STAT1 and STAT3. Next to each biomarker name, the numbers in the parenthesis indicate the sample sizes for normal and abnormal groups used to construct the boxplots and data analysis.

The Monte Carlo Type I errors to compare *0.95-quantiles* based on various underlying distributions. The significance level is 0.05. Note that the distributions in each scenario have a matching 0.95-quantile. Note that the second parameters in the normal and lognormal distributions are the variance or the variance of the logarithmic value. For the method column, sample, classic and plug-in indicate the sample quantile test, classical ELR test and the plug-in ELR test.

Table 1

Distributions	Method	n_1, n_2				
		15,15	30,30	50,100	200,200	
Normal(0,1) vs. Normal(0,1)	Sample	0.0546	0.0282	0.0392	0.0474	
	Classic	0.0000	0.0000	0.0362	0.0338	
	Plug-in	0.0100	0.0506	0.0410	0.0378	
Normal(0,1) vs. Normal(-3.2871,9)	Sample	0.0222	0.0122	0.0398	0.0398	
	Classic	0.0000	0.0016	0.0376	0.0350	
	Plug-in	0.0488	0.0698	0.0668	0.0728	
Lognormal(0, 2.25) vs. Lognormal(0.8224,1)	Sample	0.1904	0.1062	0.0876	0.0538	
	Classic	0.0000	0.0868	0.049	0.0322	
	Plug-in	0.0582	0.0632	0.0572	0.0402	
Normal(4,9) vs. Lognormal(0.5451, 1)	Sample	0.1004	0.0494	0.0476	0.0430	
	Classic	0.0000	0.0666	0.0478	0.0458	
	Plug-in	0.0132	0.0544	0.0560	0.0468	
Exponential(1) vs. Lognormal(-1.3701,2.25)	Sample	0.1760	0.0914	0.0792	0.0582	
	Classic	0.0000	0.0222	0.0370	0.0366	
	Plug-in	0.0386	0.0504	0.0458	0.0480	
Lognormal(0, 1) vs. Gamma(3,1.215)	Sample	0.1218	0.0596	0.0596	0.0394	
	Classic	0.0000	0.0078	0.0546	0.0350	
	Plug-in	0.0192	0.0366	0.0558	0.0428	

The Monte Carlo Type I errors to compare *0.9-quantiles* based on various underlying distributions. Note that the distributions in each scenario have a matching 0.9-quantile. The significance level is 0.05. Note that the second parameters in the normal and lognormal distributions are the variance or the variance of the logarithmic value. For the method column, sample, classic and plug-in indicate the sample quantile test, classical ELR test and the plug-in ELR test.

Table 2

Distributions	Method	n_1, n_2				
		15,15	30,30	50,100	200,200	
Normal(0,1) vs. Normal(0,1)	Sample	0.0156	0.0236	0.0262	0.0434	
	Classic	0.0030	0.0252	0.0304	0.0348	
	Plug-in	0.0584	0.0340	0.0410	0.0396	
Normal(0,1) vs. Normal(-2.5631,9)	Sample	0.0068	0.0176	0.0266	0.0392	
	Classic	0.0008	0.0402	0.0338	0.0476	
	Plug-in	0.0464	0.0516	0.0496	0.0516	
Lognormal(0, 6.25) vs. Lognormal(1.9223,1)	Sample	0.1150	0.1206	0.1100	0.0784	
	Classic	0.1764	0.0328	0.0504	0.0570	
	Plug-in	0.0778	0.0416	0.0388	0.0480	
Normal(9,9) vs. Lognormal(-0.0102, 4)	Sample	0.0866	0.0866	0.0658	0.0642	
	Classic	0.0888	0.0642	0.0706	0.0516	
	Plug-in	0.0692	0.0794	0.0634	0.0578	
Exponential(1) vs. Lognormal(-1.088,2.25)	Sample	0.0886	0.0660	0.0456	0.0406	
	Classic	0.0150	0.0310	0.0258	0.0368	
	Plug-in	0.0684	0.0486	0.0400	0.0428	
Lognormal(0, 1) vs. Gamma(3,1.478)	Sample	0.0606	0.0370	0.0456	0.0382	
	Classic	0.0056	0.0332	0.0366	0.0450	
	Plug-in	0.0660	0.0466	0.0498	0.0422	

Table 3

The Monte Carlo powers to compare *0.95-quantiles* based on various underlying distributions. The significance level is 0.05. Note that the second parameters in the normal and lognormal distributions are the variance or the variance of the logarithmic value. For the method column, sample, classic and plug-in indicate the sample quantile test, classical ELR test and the plug-in ELR test.

Distributions	Method	n_1, n_2			
		15,15	30,30	50,100	200,200
Normal(0,1) vs. Normal(0.5,1)	Sample	0.0698	0.0446	0.2866	0.4934
	Classic	0.0000	0.0066	0.7442	0.9924
	Plug-in	0.0118	0.0860	0.6948	0.9392
Normal(0,1) vs. Normal(-2.7871,9)	Sample	0.0366	0.0284	0.0670	0.2276
	Classic	0.0000	0.0024	0.0940	0.1766
	Plug-in	0.0812	0.1218	0.1722	0.2948
Lognormal(0, 2.25) vs. Lognormal(1.3224,1)	Sample	0.216	0.1334	0.273	0.3584
	Classic	0.0000	0.1644	0.1478	0.4100
	Plug-in	0.0930	0.1240	0.1634	0.4192
Normal(4,9) vs. Lognormal(0.7451, 1)	Sample	0.1412	0.0800	0.0934	0.3136
	Classic	0.0000	0.1066	0.1502	0.2446
	Plug-in	0.0178	0.0746	0.1834	0.2838
Exponential(1) vs. Lognormal(-1.0701,2.25)	Sample	0.1958	0.1168	0.1068	0.2538
	Classic	0.0000	0.0472	0.1008	0.1986
	Plug-in	0.0448	0.0666	0.1410	0.2356
Lognormal(0, 1) vs. Gamma(3, 3.215)	Sample	0.2168	0.1228	0.5966	0.1588
	Classic	0.0000	0.0932	0.9574	1.0000
	Plug-in	0.1624	0.4228	0.9194	0.9998

The Monte Carlo powers to compare 0.9 -quantiles based on various underlying distributions. The significance level is 0.05 . Note that the second parameters in the normal and lognormal distributions are the variance or the variance of the logarithmic value. For the method column, sample, classic and plug-in indicate the sample quantile test, classical ELR test and the plug-in ELR test.

Table 4

Distributions	Method	n_1, n_2			
		15,15	30,30	50,100	200,200
Normal(0,1) vs. Normal(0.5,1)	Sample	0.0228	0.0610	0.3906	0.8010
	Classic	0.0404	0.1406	0.3672	0.8100
	Plug-in	0.1090	0.1460	0.3816	0.8154
Normal(0,1) vs. Normal(-2.0631,9)	Sample	0.0180	0.0480	0.0622	0.2902
	Classic	0.0026	0.0574	0.1156	0.2554
	Plug-in	0.0916	0.1098	0.1962	0.2982
Lognormal(0.7, 6.25) vs. Lognormal(1.9223,1)	Sample	0.1470	0.1640	0.2424	0.4094
	Classic	0.2148	0.0694	0.1398	0.3750
	Plug-in	0.1262	0.0924	0.1386	0.3700
Normal(9,9) vs. Lognormal(0.1898, 4)	Sample	0.1024	0.1144	0.1126	0.1980
	Classic	0.1188	0.0768	0.1300	0.1484
	Plug-in	0.0976	0.1236	0.1824	0.1868
Exponential(1) vs. Lognormal(-0.788,2.25)	Sample	0.1038	0.1030	0.0834	0.3684
	Classic	0.0346	0.0480	0.1112	0.3026
	Plug-in	0.0964	0.0906	0.1548	0.3400
Lognormal(0, 1) vs. Gamma(3, 5.215)	Sample	0.1908	0.2984	0.9250	0.7290
	Classic	0.0506	0.7016	0.9688	1.0000
	Plug-in	0.3780	0.6782	0.9644	1.0000