Empirical Likelihood Ratio Test on Quantiles under a Density Ratio Model

Archer Gong Zhang

Department of Statistics
University of British Columbia

Joint work with Dr. Jiahua Chen and Dr. Guangyu Zhu

SSC 2021 Annual Meeting

Outline

Outline

Motivation

In many disciplines, data are collected as multiple samples from similar and connected populations.
For example,

- in socio-economic studies, researchers collect survey data on household characteristics from year to year;
- in network studies, people's activities on social networks in different periods of time are collected as multiple samples;
- etc...

Example: How to analyze data look like these?

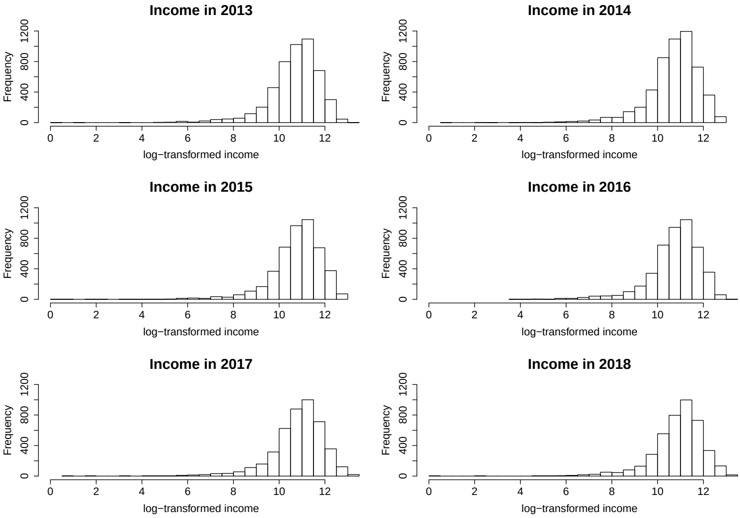

Figure: Histograms of log annual household incomes from 2013 to 2018. Data source: US Consumer Expenditure Surveys `https://www.bls.gov/cex/pumd.htm`.

Outline

Research problem

- We study hypothesis test on quantiles of multiple populations.
- Given $m+1$ independent samples from multiple populations:

$$
\begin{aligned}
x_{0,1}, x_{0,2}, \ldots, x_{0,n_0} &\overset{i.i.d.}{\sim} G_0(x) \\
x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} &\overset{i.i.d.}{\sim} G_1(x) \\
&\vdots \\
x_{m,1}, x_{m,2}, \ldots, x_{m,n_m} &\overset{i.i.d.}{\sim} G_m(x).
\end{aligned}
$$

- Consider $G_0, G_1, \ldots, G_m$ share some common features.
- Let ξ_r be the τ_r-th quantile of the r-th population.
- Hypothesis test:

$$H_0 : \boldsymbol{\xi} := (\xi_0, \xi_1, \ldots, \xi_m) = \boldsymbol{\xi}^* \quad \text{versus } H_1 : \boldsymbol{\xi} \neq \boldsymbol{\xi}^*,$$

for some given vector $\boldsymbol{\xi}^*$.

Different approaches to statistical analysis

- A fully parametric approach:
  - assumes a suitable parametric model for each population
  - there is a risk of model misspecification
- A fully non-parametric approach:
  - does not place distributional assumptions on the populations
  - free from the risk of model misspecification, but usually leads to low statistical efficiency
- ✓ a semi-parametric approach: density ratio model [Anderson, 1979]:
  - does not place parametric assumptions on each population
  - models the connection between the multiple populations to account for the latent structure they share
  - a flexible but efficient compromise between the parametric and non-parametric approaches

Density ratio model (DRM)

- The DRM models the relationship between $\{G_k\}_{k=0}^m$ by assuming the ratios of their densities $\{g_k\}_{k=0}^m$ have certain forms:

$$\frac{g_k(x)}{g_0(x)} = \exp\left\{\boldsymbol{\theta}_k^\top \mathbf{q}(x)\right\}, \qquad k = 0, 1, \ldots, m.$$

- $\mathbf{q}(x)$ is some given vector-valued function, called the basis function; we require the first component of $\mathbf{q}(x)$ to be 1.
- $\boldsymbol{\theta}_k$ is some unknown vector-valued parameters to be estimated; the first component of $\boldsymbol{\theta}_k$ is a normalizing constant.

Possible DRM-based approaches

Some possible approaches under the DRM:

- Wald-type methods [Chen and Liu, 2013];
- Likelihood ratio test (our approach).

Wald method

- Chen and Liu [2013] propose a quantile estimator $\hat{\xi}$ that is asymptotically normal with covariance Σ.
- The Wald method is used for $H_0 : \xi = \xi^*$, with the test statistic

$$n(\hat{\xi} - \xi^*)^\top \Sigma^{-1} (\hat{\xi} - \xi^*).$$

- A consistent and stable estimate of Σ must be provided.
- [Chen et al., 2016] suggest a resampling scheme for an estimate of Σ.

Our approach: likelihood ratio test

- We investigate the use of the likelihood ratio test (LRT).
- The LRT is generally believed to be more powerful, suggested by the Neyman–Pearson lemma.
- The LRT confidence regions have data-driven shapes, while those by the Wald method are oval-shaped.
- In fact, the LRT approach is the core of the foundational work of the empirical likelihood by Owen [1988].

Outline

Empirical Likelihood

- We use a non-parametric inference method: the empirical likelihood (EL).
- Owen [2001]: "EL keeps the effectiveness of likelihood methods and does not impose a known family distribution on the data."
- There have been many works on the EL approach under the DRM [e.g., Qin, 1993; Qin and Zhang, 1997; Fokianos et al., 2001; Qin, 1998; Chen and Liu, 2013; Cai et al., 2017].

EL under DRM

- Let x_{kj} be the j-th observation from the k-th population, and let $p_{kj} = \mathrm{d}G_0(x_{kj}) = P(X = x_{kj}; G_0)$.
- The principle of EL leads to the EL under the DRM:

$$L_n(G_0, \ldots, G_m) = \prod_{k,j} \mathrm{d}G_k(x_{kj}) = \Big\{ \prod_{k,j} p_{kj} \Big\} \times \exp \Big\{ \sum_{k,j} \boldsymbol{\theta}_k^\top \mathbf{q}(x_{kj}) \Big\}.$$

- The log-EL regarded as a function of $\boldsymbol{\theta}$ and G_0:

$$\ell_n(\boldsymbol{\theta}, G_0) = \log L_n(G_0, \ldots, G_m) = \sum_{k,j} \log p_{kj} + \sum_{k,j} \boldsymbol{\theta}_k^\top \mathbf{q}(x_{kj}).$$

An empirical likelihood ratio test (ELRT) approach

- Recall: $H_0 : \xi = \xi^*$ versus $H_1 : \xi \neq \xi^*$.
- The test statistic R_n is twice the difference between the two largest possible values of the log-EL $\ell_n(\boldsymbol{\theta}, G_0)$:
  - one is attained within the space of all DRM distributions $G_0, \ldots, G_m$: $H_0 \cup H_1$;
  - one is attained in the subset where their quantiles are ξ^*: H_0.
- Reject H_0 if this difference is too large by some standard formed by the distribution of R_n under H_0.

ELRT statistic

- The space of $H_0 \cup H_1$ correspond to $\{\boldsymbol{\theta}, G_0\}$ satisfying

$$\sum_{k,j} p_{kj} \exp\{\boldsymbol{\theta}_r^\top \mathbf{q}(x_{kj})\} = 1. \tag{1}$$

- The space of H_0 correspond to $\{\boldsymbol{\theta}, G_0\}$ satisfying (1) and

$$\sum_{k,j} p_{kj} \exp\{\boldsymbol{\theta}_r^\top \mathbf{q}(x_{kj})\} 1(x_{kj} \le \xi_r^*) = \tau_r. \tag{2}$$

- Our ELRT statistic is defined as

$$R_n = 2\Big\{ \underbrace{\sup_{\boldsymbol{\theta}, G_0} \{\ell_n(\boldsymbol{\theta}, G_0) | (1)\}}_{H_0 \cup H_1} - \underbrace{\sup_{\boldsymbol{\theta}, G_0} \{\ell_n(\boldsymbol{\theta}, G_0) | (1), (2)\}}_{H_0} \Big\}.$$

Asymptotic chi-squaredness of the ELRT statistic

We have Wilks' Theorem as in the classical likelihood theory:

Theorem

Under some conditions and H_0, the ELRT statistic $R_n \xrightarrow{d} \chi^2_{m+1}$ as the total sample size $n = n_0 + \cdots + n_m \to \infty$.

- This result allows us to determine an approximate rejection region for the test.
- Reject H_0 at significance level α when $R_n \geq \chi^2_{1-\alpha, m+1}$.

Outline

US consumer expenditure surveys data

- We consider a survey data from the US consumer expenditure surveys, from 2013-2018, where ≈ 5000 households are contacted each year.
- Data available on `https://www.bls.gov/cex/pumd.htm`.
- The variable of interest is the annual sum of the income received by all household members.
- We log-transformed the income values to make the scale more suitable for numerical computation.

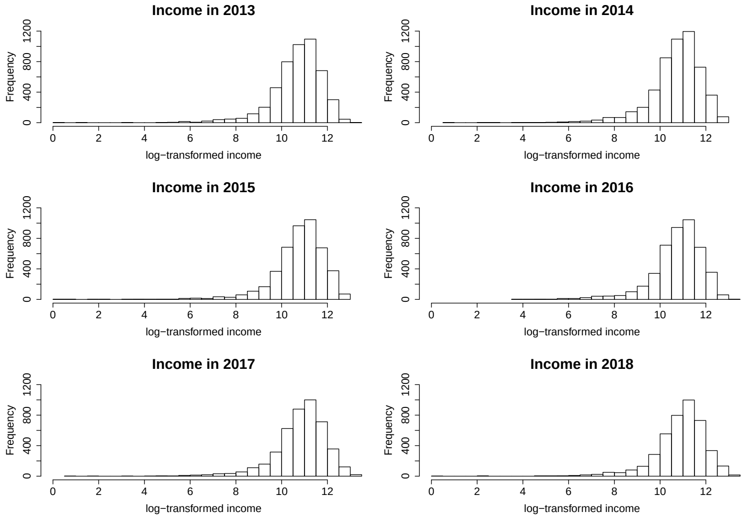

Figure: Histograms of log annual household incomes from 2013 to 2018.

Real-data based simulations

- Apparently, these distributions are connected.
- It is difficult to prescribe a suitable parametric model for these data sets, but a DRM may work well enough.
- We use real-data based simulations by sampling (with replacement) repeatedly from the 6 populations formed by the yearly 2013-2018 data sets to:
  1. check whether chi-square is a good approximation of the distribution of R_n under H_0
  2. study the confidence region based on our ELRT approach

Is chi-square is a good approximation?

Left: H_0 regarding 50% quantile in 2013;
Right: H_0 regarding 50% quantile in 2014;

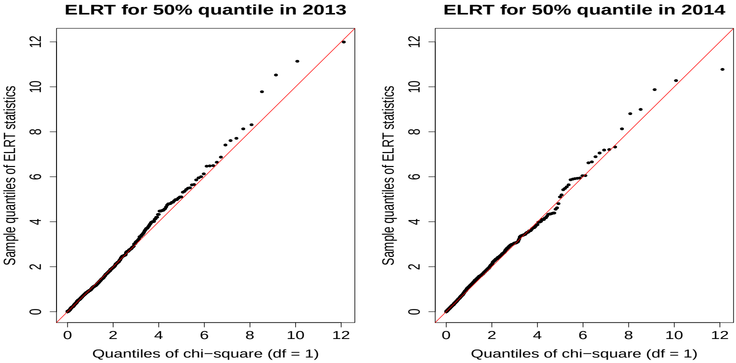

Figure: Q-Q plots of R_n values against χ_1^2, based on 1000 simulated real data sets of an equal sample size $n_r = 100$. We use $\mathbf{q}(x) = (1, x, x^2)^\top$.

Confidence region

H_0 regarding 20% quantiles in 2013 and 2018 simultaneously

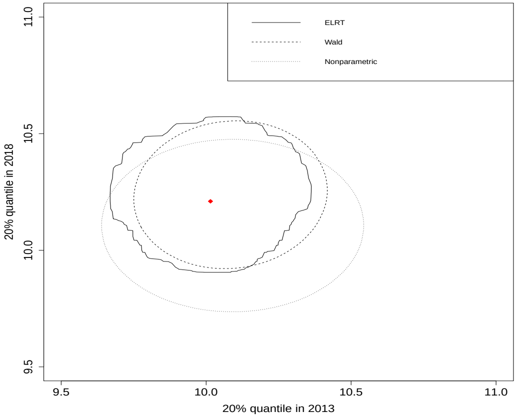

Figure: 95% confidence regions by three methods, based on one simulated real data set of an equal sample size $n_r = 100$. Diamond: location of the true quantiles. We use $\mathbf{q}(x) = (1, x, x^2)^\top$.

Numerical results

Table: Empirical coverage probabilities and average areas for 20% quantiles in 2013 and 2018 simultaneously, based on 1000 simulated real data sets of an equal sample size n_r.

Method	Nominal level: 90%		Nominal level: 95%	
	Coverage probability	Area	Coverage probability	Area
		$n_r = 100$		
ELRT	89.00%	0.284	94.20%	0.379
Wald	86.30%	0.245	91.80%	0.319
Nonparametric	87.20%	0.358	91.60%	0.466
		$n_r = 200$		
ELRT	88.20%	0.130	93.40%	0.171
Wald	86.10%	0.120	92.30%	0.156
Nonparametric	88.80%	0.183	93.80%	0.238

Summary on real-data analysis

- The points of R_n in the Q-Q plots are close to the 45-degree line: the chi-square approximation is satisfactory.
- The ELRT produces very satisfactory confidence regions that have data-driven shapes.
- The ELRT confidence regions improve the Wald confidence regions by rightfully increased area to achieve more accurate coverage probabilities. They are much more efficient than the nonparametric confidence regions.

Outline

Future work

ELRT for a composite hypothesis regarding function of quantiles

$$H_0 : \boldsymbol{g}(\xi^*) = \mathbf{0} \quad \text{against } H_1 : \boldsymbol{g}(\xi^*) \neq \mathbf{0}.$$

- Application: have the 5-th percentiles changed across years?

References I

J. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

S. Cai, J. Chen, and J. V. Zidek. Hypothesis testing in the presence of multiple samples under density ratio models. *Statistica Sinica*, 27:761–783, 2017.

J. Chen and Y. Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669–1692, 2013.

J. Chen, P. Li, Y. Liu, and J. V. Zidek. Monitoring test under nonparametric random effects model. *arXiv preprint arXiv:1610.05809*, 2016.

K. Fokianos, B. Kedem, J. Qin, and D. A. Short. A semiparametric approach to the one-way layout. *Technometrics*, 43(1):56–65, 2001.

A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

A. B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, New York, 2001.

J. Qin. Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21(3):1182–1196, 1993.

J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

J. Qin and B. Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

Thank you!