A test for stochastic ordering under biased sampling

Hsin-wen Chang

Joint work with Dr. Hammou El Barmi and Dr. Ian W. McKeague
Institute of Statistical Science, Academia Sinica

June 21, 2016

Outline

- Introduction
  - biased sampling
  - motivation
  - stochastic ordering and empirical likelihood (EL)
- Method: EL test for stochastic ordering between two distributions under biased sampling
- Simulation study
- Discussion
  - back to motivating example
  - summary and future directions

Sampling bias

- The probability of a datum being selected into a sample depends on the datum's magnitude
  - a.k.a. size bias, selection bias, ascertainment bias (in genetics), visibility bias (in animal studies)
- $P(X^* \text{is selected} | X^* = x) \propto w(x)$
  - $w(x) = x$: length bias, e.g. family size; time
  - $w(x) = x^3$: 'volume' bias, e.g. factories sampling 3-D objects

Two-sample framework

- Due to sampling bias, instead of observing samples from $F_j = 1 - S_j$ directly ($j = 1, 2$), we observe samples from a biased version of F_j:

$$G_j(x) = \int_0^x \frac{w_j(u)}{W_j} dF_j(u)$$

  according to some biasing or weight function $w_j(\cdot) > 0$, where $W_j = \int_0^\infty w_j(u) dF_j(u) < \infty$ is the normalizing constant
- F_j: unbiased distribution function; G_j: biased distribution function
- Want to compare F_1 and F_2

Literature review

- For groups of size-biased data, NPMLE for the unbiased distribution function and its weak convergence have been established [Vardi, 1982, Vardi, 1985, Gill et al., 1988]
  - A two-sample test based on the NPMLEs from each sample: only point-wise comparison feasible
- EL has been applied to biased sampling problems [Qin, 1993, El Barmi and Rothmann, 1998, Davidov et al., 2010]
  - However, simultaneous confidence bands and hypothesis testing have not been considered
- We develop an EL test that compares the underlying distribution functions uniformly

Motivating example

- Compare blood alcohol concentration of young and old drivers
  - drivers with higher alcohol levels are more likely to be sampled
  - 125 drunken drivers, 67 young and 58 old (cutoff age: 30)

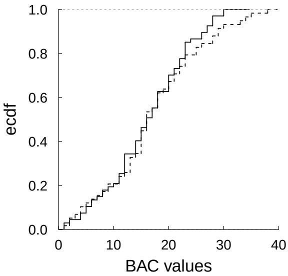


Figure: The empirical cdf of observed BAC values for drivers of age less than 30 (solid) and at least 30 (dashed).

Motivating example (cont.)

- Bias differs btwn young & old [Ramírez and Vidakovic, 2010]
- Consider $w_o(x) = x$, $w_y(x) = x^r$ $(r \in (0, 1))$
  - to upweight sampling at lower levels of BAC in the younger group
  - $r = 1/2$ in [Ramírez and Vidakovic, 2010]

Motivating example (cont.)

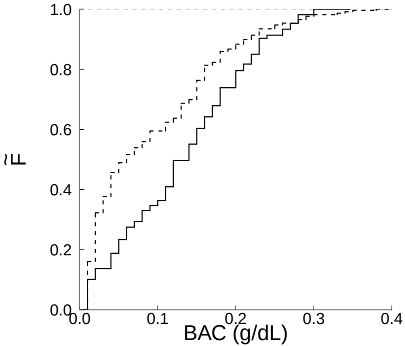


Figure: The NPMLE for the underlying distribution function of BAC values for drivers of age less than 30 (solid) and at least 30 (dashed); the weight functions for the NPMLEs are taken to be $w_y(x) = \sqrt{x}$ and $w_o(x) = x$, respectively.

Stochastic ordering

- Goal: to detect whether the survival function is uniformly higher in one group than the other
- Framed in terms of the classical notion of stochastic ordering:
  - a survival function S_1 is said to be *stochastically larger* than another survival function S_2 if $S_1(t) \geq S_2(t)$ for all $t \geq 0$
  - $\succ$: $\geq$ for all t and $>$ for some t
- We will be testing

$$H_0 : S_1 = S_2 \text{ versus } H_1 : S_1 \succ S_2$$

based on size-biased random samples from each population

Empirical likelihood (EL)

- EL involves forming a ratio of two nonparametric likelihoods subject to constraints on the parameters of interest
- Two early papers: [Thomas and Grunkemeier, 1975], [Owen, 1988]
- Produces highly accurate confidence regions [Owen, 2001] and tests with optimal power [Kitamura et al., 2012]

The usual EL without sampling bias

Given $X_1, \ldots, X_n$ i.i.d. from some unknown cdf F_0 and let $\mathcal{F}_X$ be the space of all distribution functions supported on $\{X_1, \ldots, X_n\}$:

- The nonparametric likelihood ratio for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, where $\theta = \theta(F_0)$:

$$\mathcal{R}(\theta) = \frac{\sup \{L(F) : \theta(F) = \theta_0, F \in \mathcal{F}\}}{\sup \{L(F) : F \in \mathcal{F}\}}$$

- For example, for the mean $\mu \equiv E(X_1)$, the (empirical) likelihood ratio for $H_0 : \mu = \mu_0$:

$$\mathcal{R}(\mu_0) = \frac{\sup \{\prod_{i=1}^n p_i | \sum_{i=1}^n p_i X_i = \mu_0, p_i \geq 0, \sum_{i=1}^n p_i = 1\}}{\sup \{\prod_{i=1}^n p_i | p_i \geq 0, \sum_{i=1}^n p_i = 1\}}$$

The usual EL without sampling bias (cont.)

- If $0 < \mathrm{Var}(X_i) < \infty$, then

$$-2\log\left(\mathcal{R}(\mu_0)\right) \xrightarrow{d} \chi^2_{(1)}$$

as $n \to \infty$

- Hypothesis testing can be conducted and the level $1-\alpha$ confidence interval for μ is

$$\left\{\mu_0 : -2\log\left(\mathcal{R}(\mu_0)\right) \leq \chi^{2,1-\alpha}_{(1)}\right\}$$

Idea behind the procedure

- First construct the EL test statistic for testing the "local" hypotheses $H_0^t : S_1(t) = S_2(t)$ versus $H_1^t : S_1(t) > S_2(t)$ for a given t
- The local EL ratio at t is

$$\mathcal{R}(t) = \frac{\sup\{L(S_1, S_2) : S_1(t) = S_2(t)\}}{\sup\{L(S_1, S_2) : S_1(t) \geq S_2(t)\}}$$

- Then use the maximally selected localized statistic for the general hypothesis

Nonparametric likelihood for size-biased data

- Nonparametric likelihood can be written as

$$\prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{w_{ij} p_{ij}}{W_j},$$

where $w_{ij} \equiv w_j(X_{ij})$ and $p_{ij} \equiv dF_j(X_{ij})$

- The NPMLE (the unconstrained maximizer of $L(S_1, S_2)$) is given by $\tilde{S}_j(t) \equiv 1 - \sum_{i=1}^{n_j} \tilde{p}_{ij} I_{X_{ij} \leq t}$, where $\tilde{p}_{ij} = \tilde{W}_j/(n_j w_{ij})$ and $\tilde{W}_j = n_j / \sum_{i=1}^{n_j} (1/w_{ij})$

Two-sample framework

- Due to sampling bias, instead of observing samples from $F_j = 1 - S_j$ directly ($j = 1, 2$), we observe samples from a biased version of F_j:

$$G_j(x) = \int_0^x \frac{w_j(u)}{W_j} dF_j(u)$$

  according to some biasing or weight function $w_j(\cdot) > 0$, where $W_j = \int_0^\infty w_j(u) dF_j(u) < \infty$ is the normalizing constant
- F_j: unbiased distribution function; G_j: biased distribution function
- Want to compare F_1 and F_2

Deriving $\mathcal{R}(t)$

- When $\tilde{S}_1(t) \geq \tilde{S}_2(t)$:
  - the denominator of $\mathcal{R}(t)$ is the unconstrained maximum given by $\prod_{i=1}^{n_j}(w_{ij}\tilde{p}_{ij})/\tilde{W}_j = \prod_{i=1}^{n_j}(1/n_j)$
  - the numerator can be obtained by the method of Lagrange multipliers
- When $\tilde{S}_1(t) < \tilde{S}_2(t)$:
  - the constrained maximum in the denominator is attained on the boundary of the constraint set, and then $\mathcal{R}(t) = 1$

Deriving $\mathcal{R}(t)$ (cont.)

- Numerator of $\mathcal{R}(t)$:
  - first maximize

$$L(S_1, S_2) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{w_{ij} p_{ij}}{\sum_{i=1}^{n_j} w_{ij} p_{ij}}$$

    subject to

$$\sum_{i=1}^{n_j} p_{ij} = 1, \sum_{i=1}^{n_j} p_{ij} \left(I_{X_{ij} \leq t} - F_0(t) \right) = 0, \text{ and } \sum_{i=1}^{n_j} p_{ij} \left(w_{ij} - W_j \right) = 0,$$

    for fixed W_j and $F_0(t)$, $j = 1, 2$.
  - then plugging the resulting $p_{ij}(W_j, F_0(t))$ to get a profile log-likelihood
  - maximize the profile log-likelihood over $(W_1, W_2, F_0(t))$

$$\mathcal{R}(t) = \begin{cases} 1 & \text{if } \tilde{S}_1(t) < \tilde{S}_2(t), \\ \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{n_j w_{ij} \hat{p}_{ij}}{\hat{W}_j} & \text{if } \tilde{S}_1(t) \geq \tilde{S}_2(t), \end{cases}$$

where $\hat{p}_{ij}, \hat{W}_j, \hat{\lambda}$, and $\hat{F}_0(t)$ satisfy the system of equations

$$\hat{p}_{ij} = \frac{1}{n} \frac{1}{(\kappa_j w_{ij})/\hat{W}_j + \hat{\lambda}(-1)^{j-1}(I_{X_{ij} \leq t} - \hat{F}_0(t))},$$

$$\sum_{i=1}^{n_j} \hat{p}_{ij} \left(w_{ij} - \hat{W}_j \right) = 0, \quad \sum_{i=1}^{n_j} \hat{p}_{ij} \left(I_{X_{ij} \leq t} - \hat{F}_0(t) \right) = 0.$$

Under H_0^t, $\hat{F}_0(t)$ is the maximum EL estimate of the common distribution function at t.

Large sample properties of $-2\log\mathcal{R}(t)$ under H_0

- We can show

$$-2\log\mathcal{R}(t) = U_n^2(t) I_{U_n(t)\geq 0} + o_p(1),$$

where $U_n(t) = \hat{\sigma}^{-\frac{1}{2}}(t,t)\left[V_2(t) - V_1(t)\right]$,

$$V_j(t) = \frac{W_j}{\sqrt{n_j}\sqrt{\kappa_j}} \sum_{i=1}^{n_j} \frac{I_{X_{ij}\leq t} - F_0(t)}{w_{ij}}$$

and the o_p term holds uniformly in t over $[t_1, t_2]$, for t_1 and t_2 satisfying $0 < F_0(t_l) < 1$ $(l = 1, 2)$

  - $\hat{\sigma}(t,t) = \sum_{j=1}^{2}(\hat{W}_j^2/\kappa_j)\sum_{i=1}^{n_j}[(I_{X_{ij}\leq t} - \hat{F}_0(t))/w_{ij}]^2/n_j$

Large sample properties of $-2\log\mathcal{R}(t)$ under H_0 (cont.)

- Can show

$$U_n(t) \xrightarrow{d} U(t)$$

in $l^\infty[t_1, t_2]$, where $U(t)$ is a mean 0 Gaussian process with covariance $\mathrm{cov}(U(s), U(t)) = \sigma(s,t)/\sqrt{\sigma(s,s)\sigma(t,t)}$

- By continuous mapping theorem, we have

$$-2\log\mathcal{R}(t) \xrightarrow{d} U_+^2(t)$$

in $l^\infty[t_1, t_2]$, where $U_+ = \max(U, 0)$

For the general hypotheses

- To test for the alternative of stochastic ordering, consider the maximally selected EL statistic $M_n \equiv \sup_{t \in [t_1, t_2]} [-2 \log \mathcal{R}(t)]$
- Connections to the one-sided two-sample Kolmogorov–Smirnov statistic $\sup_{t \in [t_1, t_2]} [F_{n_2 2}(t) - F_{n_1 1}(t)]_+$:
    - because $U_n(t)$ is asymptotically equivalent to $\hat{\sigma}^{-\frac{1}{2}}(t,t)\sqrt{n}\left[\tilde{F}_2(t) - \tilde{F}_1(t)\right]$,
    - $\tilde{F}_j(t)$ reduces to $F_{n_j j}(t)$ when there is no size bias (i.e., $w_j(\cdot) \equiv 1$)

Asymptotic null distribution of our test statistic

Theorem 1

Suppose $0 < F_0(t_1) < F_0(t_2) < 1$ *and* $\int_0^\infty w_j(u)^{-1}\, dF_0(u) < \infty$. *Then, under* H_0

$$M_n \xrightarrow{d} \sup_{t \in [t_1, t_2]} \left[U_+^2(t) \right].$$

Equivalent form of $U(t)$

$$U(t) \stackrel{d}{=} \frac{\sqrt{c}}{\sigma(t,t)} \left\{ B(x) + \left[x - F_0(H^{-1}(x)) \right] Z \right\},$$

where B is a standard Brownian bridge on $[0,1]$, $Z \sim N(0,1)$, $x = H(t)$,

$$H(t) = \sum_{j=1}^{2} \frac{W_j^2}{c\kappa_j} E_{G_j} \left(\frac{I_{X_{ij} \le t}}{w_{ij}^2} \right)$$

and $c = \sum_{j=1}^{2} W_j^2 / \kappa_j \times E_{G_j}(1/w_{ij}^2)$ as the sum of normalizing constants

Calibration: a Gaussian multiplier bootstrap approach

- Define a Gaussian multiplier bootstrap for M_n by $M_n^* \equiv \sup_{t\in[t_1,t_2]}\left[U_n^{*2}(t) I_{U_n^*(t)\geq 0}\right]$, where $U_n^*(t) = \hat{\sigma}^{-\frac{1}{2}}(t,t)\left[V_2^*(t) - V_1^*(t)\right]$,

$$V_j^*(t) = \frac{\hat{W}_j}{\sqrt{n_j}\sqrt{\kappa_j}} \sum_{i=1}^{n_j} \xi_{ij} \frac{I_{X_{ij}\leq t} - \hat{F}_0(t)}{w_{ij}},$$

  ξ_{ij} $(i = 1, \ldots, n_j,\ j = 1, 2)$ are i.i.d. $N(0,1)$ RVs $\perp\!\!\!\perp$ $\{X_{ij}\}$
- To calibrate the test:
  - compare the empirical quantiles of these bootstrap values M_n^* with our test statistic M_n

Large sample properties of $-2\log\mathcal{R}(t)$ under H_0

- We can show

$$-2\log\mathcal{R}(t) = U_n^2(t) I_{U_n(t)\geq 0} + o_p(1),$$

where $U_n(t) = \hat{\sigma}^{-\frac{1}{2}}(t,t)\left[V_2(t) - V_1(t)\right],$

$$V_j(t) = \frac{W_j}{\sqrt{n_j}\sqrt{\kappa_j}} \sum_{i=1}^{n_j} \frac{I_{X_{ij}\leq t} - F_0(t)}{w_{ij}}$$

and the o_p term holds uniformly in t over $[t_1, t_2]$, for t_1 and t_2 satisfying $0 < F_0(t_l) < 1$ $(l = 1, 2)$

  - $\hat{\sigma}(t,t) = \sum_{j=1}^{2}(\hat{W}_j^2/\kappa_j)\sum_{i=1}^{n_j}[(I_{X_{ij}\leq t} - \hat{F}_0(t))/w_{ij}]^2/n_j$

Bootstrap consistency theorem

Theorem 2

Assume the conditions of Theorem 1. Then conditionally on $X_{11}, X_{21}, \ldots, X_{12}, X_{22}, \ldots,$

$$M_n^* \xrightarrow{d} \sup_{t \in [t_1, t_2]} \left[U_+^2(t) \right]$$

a.s.

Simulation study

- Tests for comparison
  1. M_n^{ign}: counterpart of M_n when size bias is ignored (i.e. mistaking G_j as F_j)
  2. Wald: $\sup_{t\in[t_1,t_2]}[U_n^2(t)I_{U_n(t)\geq 0}]$, with W_j and $F_0(t)$ replaced by their consistent estimate $\hat{W}_j$ and $\hat{F}_0(t)$, respectively
- Power comparisons:
  - Underlying distributions:
    - Model A: smaller difference
    - Model B: larger difference
  - Biasing functions: $w_1(x) = \sqrt{x}$ and $w_2(x) = x$
    - The weight functions make the difference between G_1 and G_2 smaller than the difference between F_1 and F_2
    - M_n^{ign} is expected have lower power

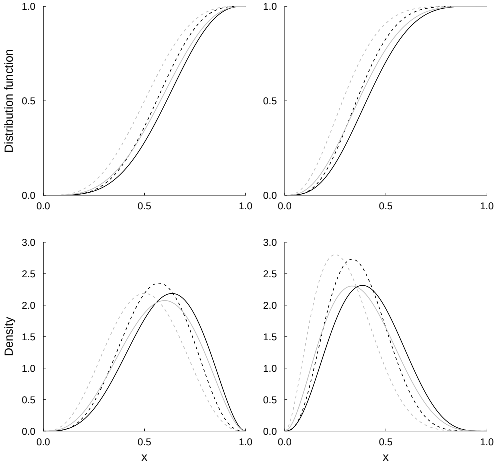


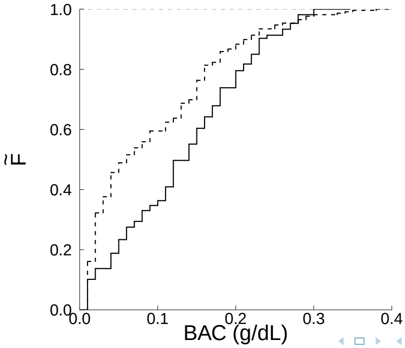
Figure: For power comparisons, the underlying (gray) and weighted (black) distribution (top row) and density (bottom row) functions in Scenario A (first column) and Scenario B (second column): F_1 and G_1

Scenario	group size	$\alpha = 0.05$			$\alpha = 0.01$		
		M_n	M_n^{ign}	Wald	M_n	M_n^{ign}	Wald
A	50	0.600	0.345	0.524	0.329	0.132	0.242
	80	0.791	0.484	0.736	0.530	0.229	0.440
B	50	0.757	0.405	0.674	0.494	0.176	0.365
	80	0.906	0.561	0.858	0.722	0.290	0.619

Table: Power simulation results based on 10,000 replications, each with 1000 bootstrap samples. **Scenario A**: smaller difference. **Scenario B**: larger difference.

Applying the proposed EL test

- Testing H_0: young= old vs H_1: young $\succ$ old:
  - $w_y(x) = \sqrt{x}$ and $w_o(x) = x$ [Ramírez and Vidakovic, 2010]
  - $M_n = 4.46$ ($p = 0.109$)
  - M_n^{ign} ($p = 0.841$) and Wald ($p = 0.168$)



Summary

- We develop an EL-based test for stochastic ordering in biased sampling models
- A simulation study shows that our test can be more powerful than the Wald test, and that considering size bias can result in a much more powerful inference than ignoring it
- We apply our test to blood alcohol measurements of drivers involved in car accidents and found a more significant result than the Wald test and test ignoring sampling bias

Future directions

- Explore the use of EL for size-biased data in other types of ordering between two distributions:
    - increasing convex ordering
    - uniform stochastic ordering (or hazard rate ordering)
- Develop a test for stochastic ordering in the k-sample case

Thank you!

Davidov, O., Fokianos, K., and Iliopoulos, G. (2010).
Order-restricted semiparametric inference for the power bias model.
Biometrics, 66(2):549–557.

El Barmi, H. and Rothmann, M. (1998).
Nonparametric estimation in selection biased models in the presence of estimating equations.
Journal of Nonparametric Statistics, 9(4):381–399.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988).
Large sample theory of empirical distributions in biased sampling models.
The Annals of Statistics, 16(3):1069–1112.

Kitamura, Y., Santos, A., and Shaikh, A. M. (2012).

On the asymptotic optimality of empirical likelihood for testing moment restrictions.
Econometrica, 80(1):413–423.

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

Owen, A. B. (2001).
Empirical Likelihood.
Chapman & Hall/CRC, Boca Raton.

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

Ramírez, P. and Vidakovic, B. (2010).

Wavelet density estimation for stratified size-biased sample.
Journal of Statistical Planning and Inference, 140(2):419–432.

Thomas, D. R. and Grunkemeier, G. L. (1975).
Confidence interval estimation of survival probabilities for censored data.
Journal of the American Statistical Association, 70:865–871.

Vardi, Y. (1982).
Nonparametric estimation in the presence of length bias.
The Annals of Statistics, 10(2):616–620.

Vardi, Y. (1985).
Empirical distributions in selection bias models.
The Annals of Statistics, 13(1):178–203.

Size simulation results

Table: Empirical significance levels based on 10,000 replications, each with 1000 bootstrap samples. **Scenario C**: $w_1(x) = x$ and $w_2(x) = \sqrt{x}$. **Scenario D**: $w_1(x) = \sqrt{x}$ and $w_2(x) = x$.

Scenario	group	$\alpha = 0.05$			$\alpha = 0.01$		
	size	M_n	M_n^{ign}	Wald	M_n	M_n^{ign}	Wald
C	50	0.053	0.153	0.053	0.012	0.044	0.012
	80	0.052	0.192	0.055	0.010	0.059	0.010
D	50	0.054	0.012	0.032	0.011	0.002	0.005
	80	0.055	0.010	0.032	0.011	0.001	0.005