Tutorial of the new endpoints, bootstrap and k-sample statistic

Hsin-wen Chang*

April 13, 2022

1 Testing under biased sampling and crossing CDFs

1.1 Biased sampling and crossing CDFs

Instead of observing directly from F_j, the observed samples $\{X_{ij}, i = 1, \ldots, n_j\}$ ($j = 1, 2$) are size-biased and i.i.d. from

$$G_j(t) = \int_{-\infty}^{t} \frac{w_j(u)}{W_j} dF_j(u),$$

where $w_j(t) > 0$ are known biasing functions depending on the size t of the datum, and $W_j = \int_{-\infty}^{\infty} w_j(u)\, dF_j(u) < \infty$ are the normalizing constants. Note that a constant biasing function yields the special case of no biased sampling. The NPMLE $F_j(t)$ (see, e.g., Owen, 2001, Ch. 6.1) is given by $\hat{F}_j(t) \equiv \sum_{i=1}^{n_j} \hat{p}_{ij} I_{X_{ij} \le t}$, where $\hat{p}_{ij} = \hat{W}_j / \{n_j w_j(X_{ij})\}$ and $\hat{W}_j = n_j / \sum_{i=1}^{n_j} (1/w_j(X_{ij}))$. Let $\hat{D}(t) = \hat{F}_1(t) - \hat{F}_2(t)$ denote the estimate for the quantity we are interested in: $F_1(t) - F_2(t)$ for $t \in [\tau_1, \tau_2]$.

To test for H_0, we start with an AD-type statistic:

$$B_n = n \int_{\tau_1}^{\tau_2} \frac{\hat{D}^2(t)}{\hat{\theta}(t,t)} d\hat{H}(t),$$

where $n = n_1 + n_2$, $\hat{H}(t) = \{n_1 \hat{F}_1(t) + n_2 \hat{F}_2(t)\}/n$ is an estimate for the common CDF F_0 under H_0, $\hat{\theta}(s,t) = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \hat{W}_j^2 \{I_{X_{ij} \le s} - \hat{H}(s)\}\{I_{X_{ij} \le t} - \hat{H}(t)\} / \{n \kappa_j^2 w_j^2(X_{ij})\}$, $\kappa_j = n_j/n$, and we adopt the convention that $0/0 = 0$. Note that the denominator of the integrand of B_n is an estimate of the asymptotic variance of $\sqrt{n}\hat{D}(t)$; this formulation takes into account the fact that the integrand of a usual AD statistic is studentized. An alternative is to simply replacing the elements in a usual two-sample AD statistic

*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

(Pettitt, 1976) by their counterparts in biased-sampling settings, leading to

$$A_n = \frac{n_1 n_2}{n} \int_{\tau_1}^{\tau_2} \frac{\hat{D}^2(t)}{\hat{H}(t)\left\{1 - \hat{H}(t)\right\}} d\hat{H}(t).$$

Unfortunately, this statistic did not perform well in our simulation study; this shows the importance of studentizing the integrand properly in an integral-type statistic.

Although both B_n and A_n incorporate variance of the estimated quantity of interest $\hat{D}(t)$, they do not consider the correlations between $\hat{D}(s)$ and $\hat{D}(t)$ for different s and t. To incorporate the correlations, consider replacing the integrand of B_n by an estimate of the Mahalanobis distance between $\sqrt{n}[\hat{F}_1(s), \hat{F}_1(t)]^T$ and $\sqrt{n}[\hat{F}_2(s), \hat{F}_2(t)]^T$: $M_n^B(s,t) \equiv n[\hat{D}(s), \hat{D}(t)]\hat{\Theta}^{-1}(s,t)[\hat{D}(s), \hat{D}(t)]^T$ and changing the single integral into double integrals, where

$$\hat{\Theta}(s,t) = \begin{bmatrix} \hat{\theta}(s,s) & \hat{\theta}(s,t) \\ \hat{\theta}(t,s) & \hat{\theta}(t,t) \end{bmatrix}.$$

This can be seen as a bivariate generalization of B_n. A similar generalization of A_n can be achieved by using the integrand $M_n^A(s,t) \equiv n[\hat{D}(s), \hat{D}(t)]\hat{\Psi}^{-1}(s,t)[\hat{D}(s), \hat{D}(t)]^T$, where

$$\hat{\Psi}(s,t) = \begin{bmatrix} \hat{H}(s)\left\{1 - \hat{H}(s)\right\} & \hat{H}(s)\left\{1 - \hat{H}(t)\right\} \\ \hat{H}(s)\left\{1 - \hat{H}(t)\right\} & \hat{H}(t)\left\{1 - \hat{H}(t)\right\} \end{bmatrix}.$$

The resulting statistics

$$BA_n = \kappa_1 \kappa_2 \iint\limits_{\substack{s,t \in [\tau_1, \tau_2] \\ s<t}} M_n^A(s,t)\ d\hat{H}(s) d\hat{H}(t) \text{ and}$$

$$BB_n = \iint\limits_{\substack{s,t \in [\tau_1, \tau_2] \\ s<t}} M_n^B(s,t)\ d\hat{H}(s) d\hat{H}(t)$$

can be viewed as testing the bivariate hypotheses H_0' versus H_1'. Although this testing problem is equivalent to testing the univariate H_0 versus H_1, the statistics for the bivariate hypotheses incorporate the (pairwise) interactions among $\hat{D}(t)$ at different t values, and hence may have better power in detecting difference between the underlying CDFs.

Remark 2. The condition $0 < F_0(t_1) < F_0(t_2) < 1$ is needed for $H(s,t)$ and $\Theta(s,t)$ to be invertable. This suggests a data-driven rule for τ_1 and τ_2: $\tau_1 = \inf\{t : \hat{F}_1(t) > 0 \text{ and } \hat{F}_2(t) > 0\}$ and $\tau_2 = \sup\{t : \hat{F}_1(t) < 1 \text{ and } \hat{F}_2(t) < 1\}$. This is what we use in the later simulation runs and data analysis.

The limiting distributions need to be estimated because they are not distribution-free. This can be done using a similar multiplier bootstrap approach as the one proposed in Chang et al. (2016) for biased sampling data. Specifically, it suffices to bootstrap the key component $\hat{D}(t)$ in the statistics by $D^*(t) = D_1^*(t) - D_2^*(t)$, where $D_j^*(t) = \sum_{i=1}^{n_j} \xi_{ij}\hat{p}_{ij}\left\{I_{X_{ij}\le t} - \hat{H}(t)\right\}$ and ξ_{ij}s are (t) standard normal random variables independent of the data. The bootstrap procedure is given as follows.

(a) Given each draw of the multipliers $\{\xi_{ij}, i = 1, \dots, n_j, j = 1, 2\}$, compute a value for each of

$$A_n^* = \frac{n_1 n_2}{n}\int_{\tau_1}^{\tau_2} \frac{D^{*2}(t)}{\hat{H}(t)\left\{1-\hat{H}(t)\right\}} d\hat{H}(t),\ B_n^* = n\int_{\tau_1}^{\tau_2} \frac{D^{*2}(t)}{\hat{\theta}(t,t)} d\hat{H}(t),$$

$$BA_n^* = \kappa_1\kappa_2 \iint\limits_{\substack{s,t\in[\tau_1,\tau_2]\\ s<t}} n[D^*(s), D^*(t)]\hat{\Psi}^{-1}(s,t)[D^*(s), D^*(t)]^T\ d\hat{H}(s)d\hat{H}(t),\ \text{and}$$

$$BB_n^* = \iint\limits_{\substack{s,t\in[\tau_1,\tau_2]\\ s<t}} n[D^*(s), D^*(t)]\hat{\Theta}^{-1}(s,t)[D^*(s), D^*(t)]^T\ d\hat{H}(s)d\hat{H}(t).$$

Alternatively, change $\hat{\theta}(s,t)$ in B_n^* and BB_n^* above to $\hat{\theta}^*(s,t) = \sum_{j=1}^{2}\hat{\theta}_j^*(s,t)$, where $\hat{\theta}_j^*(s,t)$ is the sample covariance of $V_{ij}^*(t) = \xi_{ij}\hat{W}_j[I_{X_{ij}\le t} - \hat{H}(t)]/\{\sqrt{\kappa_j}w_j(X_{ij})\}$.

1.2 k-sample statistic

$$B_n = \int_{\tau_1}^{\tau_2} \widehat{\mathrm{SSB}}(t)d\hat{H}(t),$$

where

$$\widehat{\mathrm{SSB}}(t) = \sum_{j=1}^{k}\hat{w}_j(t)\left\{\frac{\hat{\Psi}_j(t)}{\sqrt{\hat{w}_j(t)}} - \check{\Psi}(t)\right\}^2, \tag{1}$$

$\check{\Psi}(t) = \sum_{j=1}^{k}\sqrt{\hat{w}_j(t)}\hat{\Psi}_j(t)$, $\hat{\Psi}_j(t) = \sqrt{n}\{\hat{F}_j - \hat{H}\}(t)/\sqrt{\hat{\theta}_j(t,t)}$, $\hat{\theta}_j(s,t) = \sum_{i=1}^{n_j}\hat{W}_j^2\{I_{X_{ij}\le s} - \hat{H}(s)\}\{I_{X_{ij}\le t} - \hat{H}(t)\}/\{n\kappa_j^2 w_j^2(X_{ij})\}$, and the time-varying group-specific weights $\hat{w}_j(t) \propto 1/\hat{\theta}_j(t,t)$, $j = 1, \dots, k$. The weights $w_j(t)$ are normalized to sum to 1. That is, $w_j(t) = \prod_{g\in E_j}\hat{\theta}_g(t,t)/\hat{\phi}(t) > 0$, where $E_j = \{1, \dots, k\}\setminus\{j\}$ and $\hat{\phi}(t) = \sum_{l=1}^{k}\prod_{g\in E_l}\hat{\theta}_g(t,t)$. This k-sample statistic reduces to the B_n in Section 1.1 when $k = 2$. A data-driven rule for τ_1 and τ_2: $\tau_1 = \inf\{t : \hat{F}_j(t) > 0 \text{ for all } j = 1, \dots, k\}$ and $\tau_2 = \sup\{t : \hat{F}_j(t) < 1 \text{ for all } j = 1, \dots, k\}$.

For calibration, again we use a similar multiplier bootstrap method as in Section 1.1, based on sampling i.i.d. standard normal random variables (multiliers) ξ_{ij}s independent of the data. Specifically, it suffices to bootstrap the key component $\widehat{\mathrm{SSB}}(t)$ in B_n by $\widehat{\mathrm{SSB}}^*(t)$, where

$$\widehat{\mathrm{SSB}}^*(t) = \sum_{j=1}^{k}\hat{w}_j(t)\left\{\frac{\hat{\Psi}_j^*(t)}{\sqrt{\hat{w}_j(t)}} - \check{\Psi}^*(t)\right\}^2,$$

$\hat{\Psi}^*_j(t) = \sqrt{n}D^*_j(t)/\sqrt{\hat{\theta}^*_j(t,t)}$, $\hat{\theta}^*_j(t,t)$ is the sample variance of $V^*_{ij} = \xi_{ij}\hat{W}_j[I_{X_{ij}\leq t} - \hat{H}(t)]/\{\sqrt{\kappa_j}w_j(X_{ij})\}$, and $\check{\Psi}^*(t) = \sum_{j=1}^k \sqrt{\hat{w}_j(t)}\hat{\Psi}^*_j(t)$. The resulting bootstrap for B_n is $B^*_n = \sup_{a\in[\alpha_1,\alpha_2]} \widehat{\mathrm{SSB}}^*(t)$. To calibrate the test, we compare the upper α-quantile of the K^*_n values obtained from $B = 1000$ bootstrap samples (as in Section 1.1) with our test statistic B_n.

References

Chang, H., El Barmi, H., and McKeague, I. W. (2016). Tests for stochastic ordering under biased sampling. *Journal of Nonparametric Statistics*, 28(4):659–682.

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

Pettitt, A. N. (1976). A two-sample anderson-darling rank statistic. *Biometrika*, 63:161–168.