Tutorial of the new denominator

Hsin-wen Chang*

December 23, 2021

Abstract

1 Method

1.1 Biased sampling setup

The observed samples $\{X_{ij}, i = 1, \ldots, n_j\}$ $(j = 1, 2)$ are independent from

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

where $w_j(x) > 0$ are known biasing functions 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 nonparametric maximum likelihood estimate (NPMLE) based on the full likelihood for F_j (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 $\hat{W}_j = n_j / \sum_{i=1}^{n_j}(1/w_j(X_{ij}))$.

To see if $\hat{F}_1$ and $\hat{F}_2$ cross, we can check the cross product between $\hat{F}_1(x) - \hat{F}_2(x)$ at different x values. If $\{\hat{F}_1(x) - \hat{F}_2(x)\}\{\hat{F}_1(y) - \hat{F}_2(y)\}$ are non-negaive for all (x, y) pairs, then there is no crossing, and vice versa. From this, we conclude that a test statistic for H_0 vs H_1 would be sensative to crossing if it takes into account the aforementioned cross products.

1.2 AD6

An Anderson–Darling type statistic for testing H_0 can be obtained as

$$AD6 = AD_n = n \int_{t_1}^{t_2} \frac{\left\{\hat{F}_1(t) - \hat{F}_2(t)\right\}^2}{\hat{\theta}(t, t)} d\hat{H}(t)$$

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

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, $\hat{\theta}(s,t) = \sum_{j=1}^{2}\sum_{i=1}^{n_j}\hat{W}_j^2\{I_{X_{ij}\le s}-\hat{H}(s)\}\{I_{X_{ij}\le t}-\hat{H}(t)\}/(n\kappa_j^2 w_{ij}^2)$, $\kappa_j = n_j/n$, $w_{ij} = w_j(X_{ij})$, and we adopt the convention that $0/0 = 0$. However, in the presence of crossing CDFs, this statistic may not perform well, as it does not incorporate the cross products mentioned in the previous section.

To calibrate the above test, we utilize a similar multiplier bootstrap approach as the one proposed in Chang et al. (2016) for biased sampling data. Specifically, it suffices to bootstrap the numerator of the integrand, which can be done by replacing $\hat{F}_1(t) - \hat{F}_2(t) = \hat{F}_1(t) - F_0(t) - \{\hat{F}_2(t) - F_0(t)\} \equiv D_1(t) - D_2(t)$ by $D_1^*(t) - D_2^*(t)$, 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 i.i.d. standard normal random variables independent of the data. Denote the resulting bootstrap AD_n^*. To calibrate the test we simulate AD_n^* by repeatedly generating samples of Gaussian random multipliers $\{\xi_{ij}\}$. We then compare the empirical quantiles of these bootstrapped values AD_n^* with our test statistic AD_n.

1.3 AD7 and AD7'

To calibrate the new two-sample tests AD7 and AD7', we use bootstrap. For comparability we use bootstrap in the two-sample version of AD6, too.

AD7

$$AD_7 = n \iint_{\substack{s,t\in[t_1,t_2]\\ s<t}} \mathcal{I}(s,t)\; d\hat{H}(s)d\hat{H}(t),$$

where

$$\mathcal{A}_2(s,t) = \begin{pmatrix} \hat{F}_1(s) - \hat{F}_2(s) \\ \hat{F}_1(t) - \hat{F}_2(t) \end{pmatrix}^T \begin{pmatrix} \hat{\theta}(s,s) & \hat{\theta}(s,t) \\ \hat{\theta}(t,s) & \hat{\theta}(t,t) \end{pmatrix}^{-1} \begin{pmatrix} \hat{F}_1(s) - \hat{F}_2(s) \\ \hat{F}_1(t) - \hat{F}_2(t) \end{pmatrix}$$

AD7' = AD6 + AD7

Quadratic form has been utilized to combine information across different tests (see, e.g., Murakami, 2016), and in combining information across different time points in a cdf (see eg qudratic distance), but to our knowledge no two-sampel test under size bias has been developed.

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.

Murakami, H. (2016). A moment generating function of a combination of linear rank tests and its asymptotic efficiency. *TEST*, 25(4):674–691.

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