Beautiful Ideas & AUMPUT for Monotone Selection Bias Study

Presented by Zixiang Xu
Joint work with Jiayang Sun, Michael Woodroofe, Mary Meyer

Introduction

As demonstrated in today's early presentation by Professor Jiayang Sun, "*testing for potential selection is crucial for effective decision-making*", data from observational studies, including crowdsourcing, may come with selection biases. This poster

- presents ideas behind ***the approaches*** in [2] and [3], coauthored by Woodroofe, which continue to influence the development of tests for assessing selection biases; and
- highlights my enjoyment in deriving an ***AUMPUT test*** that will serve as a benchmark for evaluating new selection bias tests.

Beautiful Ideas

Setting:

In paper [3], "Testing Uniformity Versus A Monotone Density" published in 1999, the authors (Woodroofe & Sun) focused on testing

$H_0: f = 1$ vs. $H_1: f$ is monotone decreasing

where f is the selection bias function.

One test statistic immediately coming to our head is the log-likelihood function:

$$l_{\alpha,\beta}(f) = \sum_{i=1}^{n} \log[f(x_i)]$$

It can be proved that the above test statistic is maximized when f is a step function with jumps only at the observation points. So, our target would be estimating f_k instead of f, where f_k is the value of f at each step.

According to theorem in book [1], solutions of f_k are

$$f_k = \min_{i \le k} \max_{j \ge k} \frac{c_i + \cdots + c_j}{w_i + \cdots + w_j} \quad (1)$$

where $c_i = 1/n$ and $w_i = \gamma(x_i - x_{i-1})$

Reason 1: Spiking Problem
Since the max-min formula (1) can output extreme values at the two ends. There will be a spiking problem when estimating the sequence f_k, especially when computing f_n.

The penalty terms would add a constant to the numerator when the numerator is too small and add a constant to the denominator when the denominator is too small. More details are illustrated in Reason 2 below.

Reason 2: Easy Calculation
This is not the first paper dealing with the spiling problem mentioned in Reason 1. The problem was even worse under the setting of paper [2], where the authors had to add a strict constraint for each f_k. However, that made the estimation algorithm more complicated and slower. The strict boundary proposed in [2] led to the situation that there's no analytical solution to f_k. The solution can sometimes only be found by running bisection method on non-convex function.

For the proposed method in this paper [3], after taking derivative, the added penalty terms could be absorbed into equation (1) and therefore would not change the format of solution. In fact, we only have to add β to c_n and α to w_1.

Method:

Rather than using the log-likelihood function as the test statistic, the authors introduced the penalized log-likelihood function

$$l_{\alpha,\beta}(f) = \sum_{i=1}^{n} \log[f(x_i)] - n\alpha f(0+) + n\beta \log[f(x_n)]$$

Reason 3: Identifiability Issue
Although there is no identifiability issue in [3], it is a big problem discussed in paper [2]. The authors in [2] had to add a penalty term like $n\alpha$ in certain case.

Hence, adding the two penalty terms can also address any identifiability issue in any similar studies in the future using the same idea of derivation.

Ideas behind the method:

The test statistic is actually the log-likelihood function plus some penalty terms, but why did the authors add those terms?

AUMPUT

Semiparametric test is useful when selection bias is an unknown monotone function. To see its power, it is important to compare it with a good baseline test that is optimal under a parametric model.

An UMP test would be the best for us. However, in our case we have more than one parameter. Therefore, an UMPUT would be more realistic since for exponential family distribution, conditional on one of the complete and sufficient statistic, the distribution would still be from the exponential family.

Theorem:

Assume X_i's are sample from $\text{Exp}(\theta)$ with bias $w(x) = x^\beta$ $(\beta > 0)$. That is,

$$\begin{cases} & F_0 \sim \text{Exp}(\theta), \\ X_i \overset{iid}{\sim} & F_X \sim \text{Exp}(\theta) \cdot w \sim \text{Gamma}(\beta + 1, \theta). \end{cases}$$

The hypothesis is therefore:

$$H_0 : \beta = 0 \quad \text{v.s.} \quad H_1 : \beta > 0$$

We found the test

$$\phi(\vec{X}) = \begin{cases} 1 & \text{if } \overline{\ln X} - \ln \overline{X} > c, \\ 0 & \text{o.w.} \end{cases}$$

to be UMPUT at least asymptotically, or, AUMPUT.

Proof:

The complete & sufficient statistic is $(\overline{X}, \overline{\ln X})$

Conditional on $T = \overline{X}, U = \overline{\ln X}$ is still exponential distribution with density:
$f_{U|T=t}(\vec{x}) = C_t(\beta) e^{\beta u} I_{\{x_{(1)}>0\}}$

Neyman-Pearson Lemma

The MP test for any simple test like above is:

$$\phi_1(\vec{X}) = \phi_1(u,t) = \begin{cases} 1 & \text{if } u > c(t), \\ 0 & \text{o.w.} \end{cases}$$

with condition:
$P_{\beta=0}(U > c(t)) = P_{\beta=0}(U > c(T)|T = t) = \alpha = E_{\beta=0}[\phi_1(U,T)|T = t]$

The test does not depend on the choice of β

The test is the UMP test for testing:
$H_0 : \beta = 0$ v.s. $H_1 : \beta > 0$
with condition:
$E_{\beta=0}[\phi_1(U,T)|T = t] = \alpha \ (\forall t)$

The condition is exactly the Neyman structure condition

The test is the UMP test with NS condition

T is boundedly complete

The test is the UMP similar test

Power function is continuous

The test below is the UMPUT

$$\phi_1(\vec{X}) = \begin{cases} 1 & \text{if } \overline{\ln X} > c(\overline{X}) \\ 0 & \text{o.w.} \end{cases}$$

Inspired by the format of LRT:
$\phi_2(\vec{X}) = \begin{cases} 1 & \text{if } \overline{\ln X} - \ln \overline{X} > c \\ 0 & \text{o.w.} \end{cases}$

Want to show $\ln\overline{X} - \overline{\ln X}$ and $\overline{\ln X}$ are independent

Unable to prove directly

Want to show $\ln\overline{X} - \overline{\ln X}$ and $\overline{\ln X}$ are independent asymptotically

CLT plus Delta's method

$$\sqrt{n}\left[\begin{pmatrix} \overline{\ln X} - \ln \overline{X} \\ \ln \overline{X} \end{pmatrix} - \begin{pmatrix} -\gamma + \ln\theta \\ -\ln\theta \end{pmatrix}\right] \xrightarrow{d} N(\vec{0}, \begin{pmatrix} \frac{\pi^2}{6} + \ln^2\theta + 2\gamma\ln\theta - 1 & 0 \\ 0 & 1 \end{pmatrix})$$

where γ is the Euler's constant

Theorem proved!

REFERENCES:

[1] Robertson, T., F. Wright, and R. Dykstra (1988). Order Restricted Statistical Inference. Probability and Statistics Series. Wiley.
[2] Sun, J. and M. Woodroofe (1996, 09). Semi-parametric estimates under biased sampling. Statisticaa Sinica 7.
[3] Sun, J. and M. Woodroofe (1999). Testing uniformity versus a monotone density. The Annals of Statistics 27 (1), 338–360.