Distribution-free inference on the extremum of conditional expectations via classification

Lihua Lei [*]

June 30, 2023

1 Problem setup

Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random vectors taking values in $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$. Let (X, Y) denote a generic random vector drawn from the same distribution and $f(x) = \mathbb{E}[Y \mid X = x]$ denote the conditional expectation and

$$f_{\min} = \inf_{x \in \mathcal{X}} f(x), \quad f_{\max} = \sup_{x \in \mathcal{X}} f(x)$$

denote the infimum and supremum of $f(x)$. By symmetry, we only focus on making inferential claims on $f_{\min}$. The goal of this note is to obtain an upper confidence bound $\hat{f}_{\min}$ on $f_{\min}$ such that

$$\mathbb{P}(f_{\min} \le \hat{f}_{\min}) \ge 1 - \alpha. \tag{1}$$

where α is the target Type-I error. In particular, we want the guarantee (1) to hold in finite samples without any assumption on $f(x)$, in which case no consistent estimate of $f(x)$ is guaranteed to exist. Moreever, we want the method to be able to wrap around any estimator of $f(x)$ so that one can apply flexible machine learning algorithms without worrying about potential failure modes. It is not hard to see that no nontrivial lower confidence bound on $f_{\min}$ exists without assumptions on f since a perturbation of $f(x)$ in a tiny region can change $f_{\min}$ substantially while has little effect on the observed values.

2 Preliminaries

2.1 Covariate standardization

As in Lei et al. [2021], we first split the data into two folds and compute an estimate of the conditional expectation $\hat{f}(\cdot)$ on the first fold of data using an arbitrary method. With a slight abuse of notation, we let the second fold of data be $(X_1, Y_1), \ldots, (X_n, Y_n)$. The following result shows that transforming X_i never reduces $f_{\min}$.

Proposition 1. *For any estimate $\hat{f}$ that is independent of $(X_i, Y_i)_{i=1}^n$,*

$$f_{\min} \le \mathbb{E}[Y \mid \hat{f}(X)], \quad \textit{almost surely.}$$

[*]The author would like to thank Jesse Shapiro for the discussion that motivates this note.

Let $Z_i = \hat{f}(X_i)$. Then we are left to find an upper confidence bound for $g_{\min} \triangleq \inf_{z\in\mathbb{R}} g(z)$ where

$$g(z) = \mathbb{E}[Y \mid Z = z].$$

Throughout the rest of the note, we will construct upper confidence bounds on $g_{\min}$.

2.2 Inverting hypothesis tests

When Y is binary, the classification-error (CE) O-value in Lei et al. [2021] only works for $\inf_x \min\{f(x), 1 - f(x)\}$ and does not directly apply to $\inf_x f(x)$. For the latter estimand, we will take a somewhat different strategy by exploiting the duality between confidence intervals and hypothesis testing. Specifically, for any $c \in \mathbb{R}$, consider the null hypothesis $H_0(c) : g_{\min} \geq c$. Suppose that, for each $c \in \mathbb{R}$, we find a test ϕ_c that maps the data to $\{0, 1\}$ such that

$$\mathbb{P}_{H_0(c)}(\phi_c = 1) \leq \alpha.$$

When ϕ_c is monotonic in the sense that $\phi_{c_1} \leq \phi_{c_2}$ almost surely for any $c_1 < c_2$ (i.e., $H_0(c_2)$ is rejected if $H_0(c_1)$ is so), an upper confidence bound can be obtained by simply inverting the test, i.e.,

$$\hat{f}_{\min} = \inf\{c \in \mathbb{R} : \phi_c = 1\}. \tag{2}$$

However, for the problem considered in this note, it is unclear how to construct a monotonic decision. When ϕ_c is not guaranteed to be monotonic, we can instead define

$$\hat{f}_{\min} = \inf\{c : \phi_{c'} = 1,\ \forall c' \geq c\}. \tag{3}$$

The following result shows that it is a valid upper confidence bound.

Proposition 2. *If $\mathbb{P}_{H_0(c)}(\phi_c = 1) \leq \alpha$ for any $c \in \mathbb{R}$,*

$$\mathbb{P}\left(f_{\min} \leq \hat{f}_{\min}\right) \geq 1 - \alpha.$$

Proof. By definition, if $f_{\min} > \hat{f}_{\min}$, $\phi_{f_{\min}} = 1$. Since $H_0(f_{\min})$ is true,

$$\mathbb{P}\left(f_{\min} > \hat{f}_{\min}\right) \leq \mathbb{P}\left(\phi_{f_{\min}} = 1\right) = \mathbb{P}_{H_0(f_{\min})}\left(\phi_{f_{\min}} = 1\right) \leq \alpha.$$

□

In some cases, (3) is hard to compute because it requires the entire path on the right of c. Instead, we can start by discretizing c into a grid $0 = c_0 < c_1 < \ldots < c_N < c_{N+1} = 1$ and then define

$$\hat{f}_{\min} = c_{\hat{j}}, \quad \hat{j} = \min\left\{j : \phi_{c_{j'}} = 1,\ j' \geq j\right\}. \tag{4}$$

This is equivalent to apply the fixed sequence test that has over 35 years of history in medical statistics [Sonnemann et al., 1986, Bauer, 1991]. The benefit is that it involves absolutely no multiple testing adjustment and the test ϕ_c is just required to be pointwise valid for $H_0(c)$. The number of grid points is entirely driven by the computation budget.

Here we provide a self-contained proof without resorting to the general argument.

Proposition 3. *Proposition 2 holds for the upper confidence bound defined in* (4).

Proof. Let $j_0 = \min\{j : c_j \geq f_{\min}\}$. Then $H_0(c_{j_0})$ holds and

$$\mathbb{P}(c_{\hat{j}} < f_{\min}) = \mathbb{P}(\hat{j} < j_0) \leq \mathbb{P}(\phi_{j_0} = 1) = \mathbb{P}_{H_0(c_{j_0})}(\phi_{j_0} = 1) \leq \alpha.$$

□

In the following sections, we will construct valid tests for $H_0(c)$ with a fixed $c \in \mathbb{R}$.

3 Method

3.1 Binary outcomes

In this subsection we assume Y_i is binary. Let $Y_{(i)}$ be the outcome corresponding to the i-th largest Z's, i.e., $Y_{(i)} = Y_{R_i}$ where $Z_{R_1} \leq Z_{R_2} \leq \ldots \leq Z_{R_n}$. Conditional on $\{Z_1, \ldots, Z_n\}$, $Y_{(1)}, \ldots, Y_{(n)}$ are independent Bernoulli variables. Under $H_0(c)$, for any entrywise increasing function $u : [0,1]^n \mapsto \mathbb{R}$

$$u(Y_{(1)}, \ldots, Y_{(n)}) \succeq u(B_1(c), \ldots, B_n(c)) \tag{5}$$

where $B_i(c) \overset{i.i.d.}{\sim} \mathrm{Ber}(c)$ and $\succeq$ denotes stochastic dominance. Let

$$q_n(\alpha, c; u) = \sup\left\{x : \mathbb{P}(u(B_1(c), \ldots, B_n(c)) \leq x) < \alpha\right\}.$$

For any given u, $q_n(\alpha, c; u)$ can be computed to any acculation by the Monte-Carlo method. Then we can define a valid test for $H_0(c)$ as

$$\phi_c = I\left\{u(Y_{(1)}, \ldots, Y_{(n)}) \leq q_n(\alpha, c; u)\right\}.$$

One reasonable option for u is

$$u(y_1, \ldots, y_n) = \min_{k \in \{1, \ldots, n\}} \frac{y_1 + \ldots + y_k}{k}. \tag{6}$$

A shortcoming of this test statistic is that it could be dominated by the first few observations (e.g., $y_1 = 0$). Another option is

$$u_f(y_1, \ldots, y_n) = \max_{n/f(n) \leq k \leq n} \frac{S_k}{\sqrt{kc(1-c)}}, \quad \text{where } S_k = \sum_{i=1}^{k}(y_i - c) \tag{7}$$

for some $f(n) \in [1, n]$. In practice, one can simply choose $f(n) = n$. Unlike (6), the maximizer of (7) diverges, thereby allowing the statistic to account for a majority of observations. Furthermore, the critical value can be approximated by a version of Darling-Erdös theorem.

Proposition 4. *[Berkes and Weber [2006]] Assume that* $B_1(c), \ldots, B_n(c) \overset{i.i.d.}{\sim} \mathrm{Ber}(c)$. *As* $n \to \infty$,

$$a_{n,f}(u_f(B_1(c), \ldots, B_n(c)) - b_{n,f}) \overset{d}{\to} H,$$

where

$$a_{n,f} = \sqrt{2 \log\log f(n)}, \quad b_{n,f} = a_n + \frac{\log\log\log f(n) - \log 4\pi}{2a_n},$$

and H *is the distribution with CDF* $\exp\{-\exp\{-x\}\}$. *In particular,*

$$\lim_{n\to\infty} \mathbb{P}\left(u_f(B_1(c), \ldots, B_n(c)) \leq b_{n,f} - a_{n,f}^{-1} \log\log\left(\frac{1}{\alpha}\right)\right) = \alpha.$$

3.2 Bounded outcomes

In this subsection, we consider the case of bounded outcomes. Without loss of generality, we assume $Y_i \in [0,1]$. Let

$$\mathcal{E}(\eta) = \mathbb{E}[(1-Y)I(Z>\eta)] + \mathbb{E}[YI(Z\le\eta)]\cdot\frac{1}{1+c} + \mathbb{E}[YI(Z>\eta)]\cdot\frac{c}{1+c}.$$

Then we can rewrite $\mathcal{E}(\eta)$ as

$$\begin{aligned}\mathcal{E}(\eta) &= \mathbb{E}\left[\left(1-Y+\frac{c}{1+c}Y\right)I(Z>\eta)+\frac{1}{1+c}YI(Z\le\eta)\right]\\ &= \mathbb{E}\left[\mathbb{E}[\left(1-\frac{1}{1+c}Y\right)I(Z>\eta)+\frac{1}{1+c}YI(Z\le\eta)\mid Z]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{1}{1+c}g(Z)\right)I(Z>\eta)+\frac{1}{1+c}g(Z)I(Z\le\eta)\mid Z\right]\right].\end{aligned}$$

Since $I(Z>\eta)+I(Z\le\eta)=1$,

$$\mathcal{E}(\eta) \ge \mathbb{E}\left[\min\left\{1-\frac{1}{1+c}g(Z),\frac{1}{1+c}g(Z)\right\}\right].$$

Under $H_0(c)$, it is clear that

$$\mathcal{E}(\eta) \ge \frac{c}{1+c}.$$

This yields a testable implication. We estimate $\mathcal{E}(\eta)$ by the empirical analogue:

$$\hat{\mathcal{E}}(\eta) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left\{I(Y_i=0,Z_i\ge\eta)+I(Y_i=1,Z_i<\eta)\cdot\frac{1}{1+c}+I(Y_i=1,Z_i\ge\eta)\cdot\frac{c}{1+c}\right\}. \tag{8}$$

Note that

$$I(Y_i=0,Z_i\ge\eta)+I(Y_i=1,Z_i<\eta)\cdot\frac{1}{1+c}+I(Y_i=1,Z_i\ge\eta)\cdot\frac{c}{1+c}\in[0,1].$$

By Theorem 1 in Appendix A, we can compute $t_n(\alpha,\xi)$ by inverting the tail probability bound such that with probability $1-\alpha$,

$$\sup_{\eta\in[0,1]}\frac{\mathcal{E}(\eta)-\hat{\mathcal{E}}(\eta)}{\sqrt{\mathcal{E}(\eta)+\xi}} \le t_n(\alpha,\xi).$$

In particular, we can choose ξ based on the strategy discussed in Appendix F.3 of Angelopoulos et al. [2021]. As a result, with probability $1-\alpha$,

$$\mathcal{E}(\eta) \le \hat{\mathcal{E}}(\eta)+\frac{t_n^2(\alpha,\xi)}{2}+t_n(\alpha,\xi)\sqrt{\hat{\mathcal{E}}(\eta)+\xi+\frac{t_n^2(\alpha,\xi)}{4}},\quad \forall\eta\in[0,1].$$

This implies a valid test for $H_0(c)$:

$$\phi_c = I\left(\inf_{\eta\in[0,1]}\hat{\mathcal{E}}(\eta)+\frac{t_n^2(\alpha,\xi)}{2}+t_n(\alpha,\xi)\sqrt{\hat{\mathcal{E}}(\eta)+\xi+\frac{t_n^2(\alpha,\xi)}{4}}<\frac{c}{1+c}\right).$$

Note that ϕ_c is non-monotonic in c. In addition, ϕ_c is not easy to compute when ξ is chosen based on the technique described in Appendix F.3 of Angelopoulos et al. [2021] which depends on c intricately. Nonetheless, we can compute an upper bound by (4).

References

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Peter Bauer. Multiple testing in clinical trials. *Statistics in medicine*, 10(6):871–890, 1991.

István Berkes and Michel Weber. Almost sure versions of the darling–erdős theorem. *Statistics & probability letters*, 76(3):280–290, 2006.

Lihua Lei, Alexander D'Amour, Peng Ding, Avi Feller, and Jasjeet Sekhon. Distribution-free assessment of population overlap in observational studies. Technical report, Working paper, Stanford University, 2021.

E. Sonnemann, H. Finner, and T. J. Kunert. Analyse von verlaufskurven. *Biometriekurs*, 1986.

Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 1995.

A A computable concentration inequality for self-normalized empirical processes

This section reviews the concentration inequality derived in Appendix G of Angelopoulos et al. [2021]. It is typically tighter than all other computable concentration inequalities (that is, the ones with explicit constants) in the past 50 years.

Let $W_1, \ldots, W_n$ be i.i.d. random variables and $S(\lambda; w)$ be a function of w indexed by a (potentially multivariate) parameter $\lambda \in \Lambda$ that takes value in $[0, 1]$ for any λ and w. In our context, $\lambda = \eta, W_i = (Y_i, Z_i)$ and

$$S(\lambda; W_i) = I(Y_i = 0, Z_i \geq \xi) + I(Y_i = 1, Z_i < \xi) \cdot \frac{1}{1+c} + I(Y_i = 1, Z_i \geq \xi) \cdot \frac{c}{1+c}.$$

Further let

$$\hat{s}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} S(\lambda; W_i), \quad s(\lambda) = \mathbb{E}[S(\lambda; W_i)]. \tag{9}$$

Furthermore, we define $\Delta(n)$ as the

$$\Delta(n) = \sup_{z_1, \ldots, z_n} \left| \left\{ \{ S(\lambda; z_1), \ldots, S(\lambda; z_n) \} : \lambda \in \Lambda \right\} \right|. \tag{10}$$

In the literature, $\log \Delta(n)$ is often referred to as the growth function ([Vapnik, 1995, Section 2]).

Theorem 1. *For any* $\xi \geq 0$,

$$\mathbb{P}\left(\sup_{\lambda\in\Lambda} \frac{s(\lambda)-\hat{s}_n(\lambda)}{\sqrt{s(\lambda)+\xi}} \geq t\right)$$

$$\leq \min\left\{ \inf_{\gamma\in(0,1),n'\in\mathbb{Z}^+} \frac{\Delta(n+n')\exp\{-g_2(t;n,n',\gamma,\kappa^-)\}}{1-\exp\{-g_1(t;n',\gamma,\xi)\}}, \inf_{\gamma\in(0,1)} \frac{\Delta(2n)\tilde{g}\left(\sqrt{\frac{n(1+\xi)}{2}}(1-\gamma)t\right)}{1-\exp\{-g_1(t;n,\gamma,\xi)\}} \right\},$$

where

$$g_1(t;n',\gamma,\kappa) = \max\left\{ \frac{n't^2}{2}\frac{\gamma^2}{1+\gamma^2t^2/36\kappa},\ \log\left(\frac{n't^2\gamma^2}{(\sqrt{1+\kappa}-\sqrt{\kappa})^2}\right)\right\},$$

$$g_2(t;n,n',\gamma,\kappa) = \frac{nt^2}{2}\left(\frac{n'}{n+n'}\right)^2 \frac{(1-\gamma)^2}{1+(1-\gamma)^2t^2/36\kappa},$$

$$\kappa^+ = \xi + \frac{t^2}{2} + t\sqrt{\frac{t^2}{4}+\xi}, \quad \kappa^- = \xi + \frac{n+\gamma n'}{n+n'}\sqrt{\kappa^+},$$

and $\tilde{g}(x) = \min\{\tilde{g}_1(x), \tilde{g}_2(x), \tilde{g}_3(x)\}$ *with*

$$\tilde{g}_1(x) = c_1(1-\Phi(x)), \quad c_1 = 1/4(1-\Phi(\sqrt{2})) \approx 3.178,$$

$$\tilde{g}_2(x) = 1-\Phi(x) + \frac{c_2}{9+x^2}\exp\left\{-\frac{x^2}{2}\right\}, \quad c_2 = 5\sqrt{e}(2\Phi(1)-1) \approx 5.628,$$

$$\tilde{g}_3(x) = \exp\left\{-\frac{x^2}{2}\right\}.$$