Biometrika (2012), **99**, 1, *pp*. 1–12

Printed in Great Britain

Advance Access publication on 31 July 2012

Shape-constrained partial identification of a population mean under unknown probabilities of sample selection

BY L. W. MIRATRIX

Harvard Graduate School of Education, 13 Appian Way, Cambridge, Massachusetts 02138, U.S.A.

lmiratrix@stat.harvard.edu

S. WAGER

Stanford Graduate School of Business, 655 Knight Way, California 94035, U.S.A.

swager@stanford.edu

AND J. R. ZUBIZARRETA

Harvard Medical School, 180 Longwood Avenue, Boston 02115, U.S.A.

zubizarreta@hcp.med.harvard.edu

SUMMARY

Estimating a population mean from a sample obtained with unknown selection probabilities is an important problem in the biomedical and social sciences. Using a ratio estimator, Aronow & Lee (2013) proposed a method for partial identification of the mean by allowing the unknown selection probabilities to vary arbitrarily between two fixed values. In this paper, we show how to use auxiliary shape constraints on the population outcome distribution, such as symmetry or log-concavity, to obtain tighter bounds on the population mean. We use this method to estimate the performance of Aymara students, an ethnic minority in the north of Chile, in a national educational standardized test. We implement this method in the package `scbounds` for `R`.

Some key words: Partial identification; Sensitivity analysis; Survey sampling.

1. INTRODUCTION

A common challenge in the biomedical and social sciences is to estimate a population mean from a sample obtained with unknown probabilities of sample selection. This is often the case when drawing inferences about mobile populations, such as the homeless and hospital outpatients, as well as with hard-to-reach populations, such as injection drug users and some ethnic minorities. In general, this problem arises when a sampling frame is unavailable or unreliable, and when there is no or limited information about the sampling design.

The estimation problem can be formalized as follows. Let $\mathcal{P}$ denote a potentially infinite population, and let F denote the cumulative distribution function of our outcome of interest Y over $\mathcal{P}$. Our goal is to estimate the population mean $\mu = E_F(Y)$. To do so, we have access to a random sample $\mathcal{S} = \{Y_i\}$ of size n obtained via biased sampling. Concretely, we can imagine that $\mathcal{S}$ was generated using an accept-reject scheme as follows: until we have n observations, repeatedly draw independent and identically distributed pairs $(Y, \pi) \in \mathbb{R} \times (0, 1]$ where $Y \sim F$, and then

add Y to our sample $\mathcal{S}$ with probability π. Whenever the inverse sampling probabilities π_i^{-1} are correlated with Y_i, the sample mean will be an inconsistent estimator for the population mean.

If these sampling probabilities π_i for our n sampled observations were known, then we could use the following ratio estimator, which is consistent for μ under weak conditions (Hájek, 1958; Cochran, 1977),

$$\hat{\mu}^* = \sum_{i=1}^{n} \pi_i^{-1} Y_i \Big/ \sum_{i=1}^{n} \pi_i^{-1}. \tag{1}$$

Here, however, we suppose that the sampling weights π_i are unknown. Aronow & Lee (2013) showed that it is possible to obtain meaningful identification intervals for μ in the sense of Manski (2003), even if we only have bounds on the sampling weights π_i. Suppose that we know that $\max(\pi_i)/\min(\pi_i) \leq \gamma$ for some constant $\gamma < \infty$. This gives an asymptotically consistent identification interval $\mathcal{I}_{\text{AL}} = [\hat{\mu}^-_{\text{AL}}, \hat{\mu}^+_{\text{AL}}]$ for μ, where

$$\hat{\mu}^+_{\text{AL}} = \sup \left\{ \sum_{i=1}^{n} w_i Y_i : \sum_{i=1}^{n} w_i = 1,\ \frac{\max(w_i)}{\min(w_i)} \leq \gamma \right\} \tag{2}$$

and $\hat{\mu}^-_{\text{AL}}$ is the $\inf$ over the same set. Aronow & Lee (2013) also develop an efficient algorithm for computing these bounds.

While this approach that can yield identification intervals for μ under weak assumptions, the Aronow–Lee bounds can be unnecessarily pessimistic. To understand why, it is helpful to consider their method as first estimating the true population F as

$$\widehat{F}_w(y) = \sum_{i=1}^{n} w_i \, 1(Y_i \leq y),$$

where w_i are the maximizing or minimizing weights in (2), and then setting the limits of the interval as $E_{\widehat{F}_w}(Y)$ for the two resulting extreme sets of weights. The problem is that the population distributions implied by these extreme weights are often rather implausible in practice.

Specifically, the weights w_i induced by (2) correspond to a step function depending on whether or not Y_i falls below some threshold, and so the weighted empirical distribution functions $\widehat{F}^+_{\text{AL}}$ and $\widehat{F}^-_{\text{AL}}$have a sharp change in slope at that threshold, as illustrated in Figure 1(a) in Section 3. This threshold can also be interpreted as a substantial discontinuity in an associated density, provided we are willing to posit the existence of such a density; see Figure 1(b). Such sharp elbows in the estimated cumulative distribution function often contradict expert knowledge about the true population distribution F. For example, physical measurements in the biological and medical sciences often exhibit a bell-shaped distribution, as do stock returns and other indicators in finance, and mechanical error measurements in industry.

This paper studies how to use such auxiliary information about the shape of the population outcome distribution F to get shorter identification intervals for μ by ruling out implausible weightings in order to tighten the resulting identification bounds. We allow for various types of specifications for F, such as parametric assumptions, and shape constraints based on symmetry or log-concavity. In general, the more we are willing to assume about F, the shorter the resulting identification intervals. At one extreme, if we know that F is Gaussian, then we can substantially shorten identification intervals, while if we make weaker assumptions, e.g., that F has a log-concave density, then we get smaller but still noticeable improvements over $\mathcal{I}_{\text{AL}}$. We focus on the situation when F has real-valued support; when F has categorical support, it is less common to have access to plausible shape constraints. This paper relates to the literature on biased sampling,

empirical likelihood, and exponential tilting (Owen 1988, Efron & Tibshirani 1996, de Carvalho & Davison 2014, Fithian & Wager 2015, Qin 2017). We implement these methods in the `R` package `scbounds`.

2. TIGHTER IDENTIFICATION BOUNDS VIA SHAPE CONSTRAINTS

Our goal is to use existing information about the population distribution F, e.g., that F is symmetric or log-concave, to obtain tighter identification bounds for $\mu = E_F(Y)$. Operationally, we seek to encode such information about F into constraints that can be added to the optimization problem (2). Throughout the paper, we assume that μ is well-defined and finite.

Our analysis focuses on the weighted empirical distribution function $\widehat{F}^*$ induced by the oracle ratio estimator $\hat{\mu}^*$ (1) that has access to the true sampling probabilities π_i that give corresponding oracle weights w_i^*:

$$\widehat{F}^*(y) = \sum_{i=1}^n \pi_i^{-1} \, 1(Y_i \leq y) \Big/ \sum_{i=1}^n \pi_i^{-1} = \sum_{i=1}^n w_i^* \, 1(Y_i \leq y). \tag{3}$$

Any shape constraint on F that lets us control the behavior of $\widehat{F}^*$ induces an asymptotically consistent identification interval for the population mean μ. The intuition is that if we can construct sets of distribution functions that are as constrained as possible, while still containing our oracle $\hat{F}^*$ with high probability, then the optimization problem will contain the oracle estimate $\hat{\mu}^*$, giving short but still consistent bounds.

THEOREM 1. *Suppose that we have auxiliary information on F that lets us construct sets of distribution functions $\mathcal{C}_{\gamma,\,n}$ with the property that $\lim_{n\to\infty} \mathrm{pr}(\widehat{F}^* \in \mathcal{C}_{\gamma,\,n}) = 1$. Then, if we write*

$$\hat{\mu}^+ = \sup \left\{ \sum_{i=1}^n w_i Y_i : \sum_{i=1}^n w_i = 1, \; \frac{\max(w_i)}{\min(w_i)} \leq \gamma, \; \widehat{F}_w \in \mathcal{C}_{\gamma,\,n} \right\}, \tag{4}$$

and $\hat{\mu}^-$ as the infimum in the analogous optimization problem, the resulting identification interval $\mathcal{I} = [\hat{\mu}^-, \, \hat{\mu}^+]$ is asymptotically valid in the sense that $\Delta(\mu, \mathcal{I})$ converges in probability to 0, where $\Delta(\mu, \mathcal{I})$ is the distance between μ and the nearest point in the interval $\mathcal{I}$.

The intervals $\mathcal{I}$ can never be wider than those of Aronow & Lee (2013), because the optimization problem (4) has strictly more constraints than the original optimization problem (2).

Theorem 1 shows that if we have any auxiliary information about F, then the identification bounds of Aronow & Lee (2013) are needlessly long. However it cannot directly guide practical data analysis. First, it leaves open the problem of how to turn shape constraints on F into plausibility sets $\mathcal{C}_{\gamma,\,n}$ that contain $\widehat{F}^*$ with high probability. Second, Theorem 1 is not useful if we cannot solve the optimization problem (4) in practice. Our next concern is to address these issues given specific side-information about F.

2·1. *Identification bounds in parametric families*

Although our main goal is to provide inference under shape constraints on F, we begin by considering the parametric case $F = F_\theta$ for some $\theta \in \Theta$, which allows us to construct particularly simple plausibility sets $\mathcal{C}_{\gamma,\,n}$. Our approach is built around a Kolmogorov–Smirnov type concentration bound for ratio estimators, and relies on finding the worst-case weighted distribution with $\pi_{\max}/\pi_{\min} \leq \gamma$ in terms of the characterization of Marcus & Shepp (1972) for the tails of Gaussian processes; see the proof in Appendix.

LEMMA 1. *Suppose that we have a population and sampling scheme for which* $\pi_{\min} \leq \pi_i \leq \pi_{\max}$ *with* $\pi_{\max}/\pi_{\min} \leq \gamma$*; and, for* $\alpha > 0$*, define*

$$\rho_{\alpha,n} = \text{pr}\left[n^{1/2} \sup_{y\in\mathbb{R}} \left| \widehat{F}^*(y) - F(y) \right| \geq \left\{ \frac{\sigma_\gamma^2 \left(1+\gamma\right)\left(1+\gamma^{-1}\right)\log \alpha^{-1}}{2} \right\}^{1/2} \right], \quad (5)$$

for a constant $\sigma_\gamma^2 \leq 1$ *defined as the maximum of a concave function specified in the proof. Then the limiting probability* $\rho_{\alpha,n}$ *of large tail exceedances is bounded as follows:*

$$\limsup_{\alpha\to 0} \left(\limsup_{n\to\infty} \log \rho_{\alpha,n}^{-1} \Big/ \log \alpha^{-1} \right) \geq 1.$$

Inserting $\alpha = n^{-1/2}$ into equation (5) and considering the union of all possible population distributions $F(\cdot) = F_\theta(\cdot)$ with $\theta \in \Theta$ suggests using the plausibility set given by

$$\begin{aligned} \mathcal{C}_{\gamma,n}(\Theta) &= \cup_{\theta\in\Theta} \left\{ H : \sup_{y\in\mathbb{R}} |H(y) - F_\theta(y)| \leq \delta_{\gamma,n} \right\}, \\ \delta_{\gamma,n} &= \left\{ \frac{\sigma_\gamma^2 \left(1+\gamma\right)\left(1+\gamma^{-1}\right)\log n}{4n} \right\}^{1/2}. \end{aligned} \quad (6)$$

Regardless of the true parameter value $\theta \in \Theta$, we have $\hat{F}^* \in \mathcal{C}_{\gamma,n}(\Theta)$ with probability tending to unity, so we immediately have:

COROLLARY 1. *Suppose that, under the conditions of Lemma* 1*, we know that* $F = F_\theta$ *for some* $\theta \in \Theta$*, and set* $\mathcal{C}_{\gamma,n}(\Theta)$ *as in* (6)*. Then,* (4) *provides an asymptotically valid identification interval for* μ *as* $n \to \infty$*.*

Furthermore, we can also check that the resulting intervals are asymptotically sharp.

We implement this procedure by first solving the optimization problem

$$\hat{\mu}_\theta^+ = \sup\left\{ \sum_{i=1}^n w_i Y_i : \sum_{i=1}^n w_i = 1,\ \frac{\max(w_i)}{\min(w_i)} \leq \gamma,\ \sup_{y\in\mathbb{R}} \left| \widehat{F}_w(y) - F_\theta(y) \right| \leq \delta_{\gamma,n} \right\}, \quad (7)$$

over a grid of candidate values $\theta \in \Theta$, and then set $\hat{\mu}^+ = \sup_\theta \hat{\mu}_\theta^+$. The fractional programming problem (7) can be solved as a linear program using standard optimization methods; see, e.g., section 4.3 of Boyd & Vandenberghe (2004).

2·2. *Relaxation of sharp shape constraints*

Consider the parametric family case, above. Lemma 1 implies that $\widehat{F}^*(\cdot)$ will grow arbitrarily close to some F_θ with probability unity. This can be a very strong assumption: under a Gaussian family, for example, this implies that, with increasing n, not only are our bounds sharp, but they will collapse to a single point as $\mathcal{C}_{\gamma,n}(\Theta)$ shrinks due to the tightening of $\delta_{\gamma,n}$. If we instead impose a shape constraint of $\max_y |F(y) - F_\delta(y)| \leq \delta^*$, for some δ^* and unknown $\theta \in \Theta$, we can expand our set $\mathcal{C}_{\gamma,n}(\Theta)$ to

$$\mathcal{C}_{\gamma,n}(\Theta) = \cup_{\theta\in\Theta} \left\{ H : \sup_{y\in\mathbb{R}} |H(y) - F_\theta(y)| \leq \delta_{\gamma,n} + \delta^* \right\}.$$

Due to the triangle inequality on the Kolmogorov–Smirnov distances, we still have, in the limit, $\widehat{F}^*$ in our set with probability 1, and therefore valid bounds on our mean. This relaxation allows

for restricting the shape of our unknown distribution to be near a given parametric family without imposing a strong parametric assumption. The δ^* is a sensitivity parameter, and practitioners could examine how the bounds respond to different choices. We, however, instead investigate methods that make general shape constraint assumptions instead of these parametric ones.

2·3. *Identification bounds with symmetry*

We now move back to our main topic of interest, i.e., how to leverage shape constraints on F to obtain improved identification bounds for μ. The difference between parametric versus shape-constrained side information about F is that grid search is not usually practical over all distributions in some shape-constrained class, so the algorithm based on (7) does not generalize. Rather, for each candidate shape-constrained class, we may need to find an ad-hoc way to avoid a full grid search.

First, we consider the case where F is symmetric, i.e., there is some value $m \in \mathbb{R}$ such that $F(m+y) = 1 - F(m-y)$ for all $y \in \mathbb{R}$. Such symmetry constraints mesh particularly nicely with our approach. In order to make use of them, we establish the following analogue to Lemma 1. Unlike Lemma 1, this result holds for any value of $\alpha > 0$ and not only for small $\alpha \to 0$; from a technical perspective, this result follows directly from Donsker arguments used to prove the classical Kolmogorov–Smirnov theorem, and does not require the additional machinery of Marcus & Shepp (1972).

LEMMA 2. *Suppose that we have a population and sampling scheme for which* $\pi_{\min} \leq \pi_i \leq \pi_{\max}$ *for some* $\pi_{\max}/\pi_{\min} \leq \gamma$. *Then, for any* $\alpha > 0$,

$$\lim_{n\to\infty} \text{pr}\left[n^{1/2} \sup_{q\in[0,\,1]} \left|\widehat{F}^*\{F^{-1}(q)\} + \widehat{F}^*\{F^{-1}(1-q)\} - 1\right| \geq \zeta_{\gamma,\,\alpha}\right] \leq \alpha, \tag{8}$$

with

$$\zeta_{\gamma,\,\alpha} = \Phi^{-1}\left(1 - \frac{\alpha}{4}\right)\left\{\frac{(1+\gamma)\left(1+\gamma^{-1}\right)}{4}\right\}^{1/2}. \tag{9}$$

To draw a connection to Lemma 1, we note that $\Phi^{-1}(1-\alpha/4) \asymp (2\log\alpha^{-1})^{1/2}$ for small values of α. Lemma 2 holds regardless of symmetry.

If the distribution $F(\cdot)$ is symmetric around m, then any pair $\left\{F^{-1}(q),\, F^{-1}(1-q)\right\}$ with $q \in [0,\,1]$ can be written as $(m-y,\, m+y)$ for some $y \in \mathbb{R}$. Thus, in the case of symmetric distributions, (8) provides a tail bound on the supremum of $\widehat{F}^*(m+y) + \widehat{F}^*(m-y) - 1$ over $y \in \mathbb{R}$, and suggests using the estimator given by

$$\begin{aligned}
\hat{\mu}_m^+ &= \sup\left\{\sum_{i=1}^n w_i Y_i : \sum_{i=1}^n w_i = 1,\ \frac{\max(w_i)}{\min(w_i)} \leq \gamma,\ \widehat{F}_w \in \mathcal{C}_{\gamma,\,n}^{\text{SYM}}\right\}, \\
\mathcal{C}_{\gamma,\,n}^{\text{SYM}} &= \bigcup_{m\in\mathbb{R}}\left\{H : \sup_y |H(m+y) + H(m-y) - 1| \leq \zeta_{\gamma,\,n^{-1/2}}\right\}.
\end{aligned} \tag{10}$$

The lower bound $\hat{\mu}^-$ is computed analogously. This algorithm thus enables us to use symmetry constraints while only performing a grid search over a single parameter m, i.e., the center of symmetry. The identification intervals defined by (10) are again asymptotically valid and sharp.

2·4. *Identification bounds with log-concavity*

Finally, we consider the case where F is known to have a log-concave density. Imposing log-concavity constraints appears to be a promising method for encoding side information about F:

the class of log-concave distributions is quite flexible, including most widely-used parametric distributions with continuous support, while enforcing regularity properties such as unimodality and exponentially decaying tails (Walther, 2009).

Unlike in the case of symmetry, there does not appear to be a simple way to turn log-concavity constraints into asymptotically sharp identification bounds for μ using only linear programming and a low-dimensional grid search. However, we can still use log-concavity to provide computationally tractable improvements on the bounds of Aronow & Lee (2013). Below, we detail our procedure for obtaining $\hat{\mu}^+$; to obtain $\hat{\mu}^-$ we can apply the same procedure to $-Y_i$. We posit the existence of a known upper bound $Y_i \leq y_{\max}$ for the outcomes Y_i to ensure integrability.

1. Let $\widehat{S}(y) = n^{-1}\sum_i 1(Y_i \leq y)$ and $\widehat{S}_{\mathrm{KS}}(y) = \max\{\widehat{S}(y) - n^{-1/2}D_{\mathrm{KS}}^{(-1)}(1-n^{-1/2}),\, 0\}$, where $D_{\mathrm{KS}}^{(-1)}(\cdot)$ denotes the inverse Kolmogorov–Smirnov cumulative distribution function. By the Kolmogorov–Smirnov theorem, $S(y) \geq \widehat{S}_{\mathrm{KS}}(y)$ for all $y \in \mathbb{R}$ with probability tending to 1, where $S(y)$ is the distribution function of the observed, biased, sample.
2. Next, in the proof, we show that for some $m \in \mathbb{R}$,

$$\widehat{U}_m(y) = \left[\widehat{S}_{\mathrm{KS}}(y) + (\gamma-1)\left\{\widehat{S}_{\mathrm{KS}}(y) - \widehat{S}_{\mathrm{KS}}(m)\right\}_+\right] \Big/ \left[\widehat{S}_{\mathrm{KS}}(m) + \gamma\left\{1 - \widehat{S}_{\mathrm{KS}}(m)\right\}\right],$$

   is a lower bound for the population distribution of interest, $F(y)$, with probability tending to 1. We also set $\widehat{U}_m(y_{\max}) = 1$, to ensure that $\widehat{U}_m(\cdot)$ is a cumulative distribution function.
3. We now use the fact that our distribution is log-concave. It is well known that if F has a log-concave density, then $\log F(y)$ must be concave (Prékopa, 1973). Thanks to this fact, we know that if $F(y) \geq \widehat{U}_m(y)$, then also $F(y) \geq \widehat{L}_m(y)$ where

$$\widehat{L}_m(\cdot) = \arg\min_{L(\cdot)}\left\{\int_{-\infty}^{y_{\max}} L(y)\,dy : \log L(y) \text{ is concave, and } L(y) \geq \widehat{U}_m(y) \text{ for all } y \in \mathbb{R}\right\}.$$

   $\widehat{L}_m(\cdot)$ is, for m, the lowest function for which $\log L(y)$ is concave that still lies above our overall lower bound $\widehat{U}_m(y)$.
4. Finally, we define $C_{\gamma,\,n}$ as the set of distributions satisfying at least one of these lower bounds:

$$C_{\gamma,\,n}^{\mathrm{LC}+} = \bigcup_{m\in\mathbb{R}}\left\{H : H(y) \geq \widehat{L}_m(y),\ \ y \in \mathbb{R}\right\}.$$

Given this construction, we can obtain an upper endpoint for our identification interval as usual,

$$\hat{\mu}^+ = \sup_w\left\{\sum_{i=1}^n w_i Y_i : \sum_{i=1}^n w_i = 1,\ \frac{\max(w_i)}{\min(w_i)} \leq \gamma,\ \widehat{F}_w(y) \in C_{\gamma,\,n}^{\mathrm{LC}+}\right\}.$$

Lemma 3 shows that $C_{\gamma,\,n}^{\mathrm{LC}+}$ contains the population sampling distribution with probability tending to 1, and so Theorem 1 establishes the validity of our identification intervals. Unlike in the parametric or symmetric cases, our log-concave identification bounds are not asymptotically sharp, i.e., they may not converge to the shortest possible identification interval given our assumptions about log-concavity and bounded sampling ratios; however, they still provide tighter intervals than do Aronow & Lee (2013).

LEMMA 3. *Suppose that we have a population for which $\pi_{\min} \leq \pi_i \leq \pi_{\max}$ for some $\pi_{\max}/\pi_{\min} \leq \gamma$, and that F has a log-concave density. Then,* $\lim_{n\to\infty}\mathrm{pr}(\widehat{F}^* \in C_{\gamma,\,n}^{\mathrm{LC}+}) = 1$.

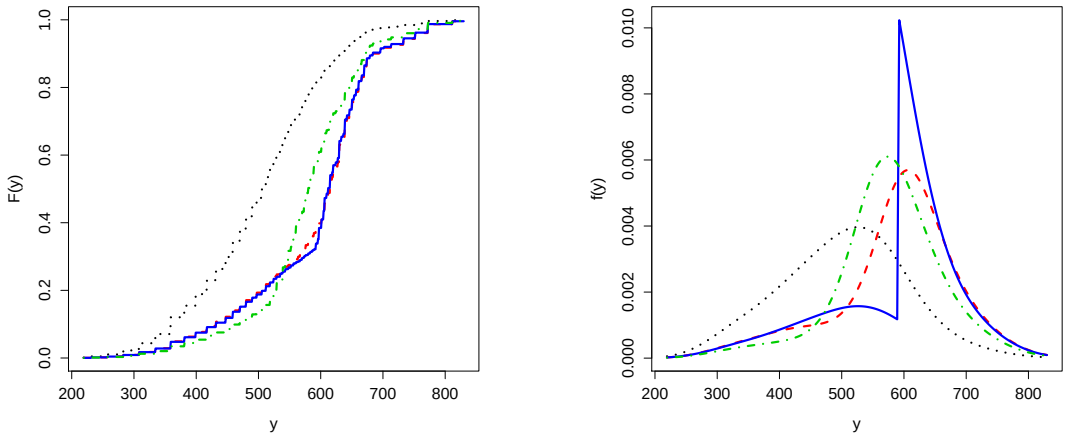
3. APPLICATION: SAMPLING ETHNIC MINORITIES

The Aymara are an indigenous population of the Andean plateau of South America, who live predominantly in Bolivia and Peru; only a small proportion of them live in the north of Argentina and Chile. In Chile, they constitute a minority of nearly 50,000 in a country of approximately 18 million. Across the world, it is of great importance to understand how ethnic minorities fare in order to design effective affirmative action policies. Here, we use the proposed method to bound the average performance of the Aymara students in the national standardized test held in Chile for admission to higher education. This test is called Prueba de Selección Universitaria and nearly 90% of enrolled high school students take it every year; however, this figure is known to be lower in vulnerable populations such as the Aymara in northern Chile.

Using the sample of 847 Aymara students that took the test in mathematics in 2008, we seek identification intervals for the population mean counterfactual test score had everyone taken the test. We assume that the sampling ratio is bounded by $\max(\pi_i) \,/\, \min(\pi_i) \leq \gamma = 9$, and consider inferences under the assumptions that the population test score distribution is symmetric, and that it is log-concave. We also consider the approach of Aronow & Lee (2013) without shape constraints.

Here, the observed test scores have a mean of 502 with a sample standard deviation of 104. Given $\gamma = 9$, we obtain population identification intervals of $(426,\ 578)$ assuming symmetry, $(414,\ 589)$ assuming log-concavity, and $(410,\ 591)$ without any constraints. Thus, in this example, assuming symmetry buys us shorter identification intervals than assuming log-concavity.

Figure 1 depicts the weighted distribution functions $\widehat{F}_w(\cdot)$ underlying the upper endpoints of all 3 identification intervals. The sharp threshold of the weights w_i resulting from the unconstrained method of Aronow & Lee (2013) is readily apparent. Assuming either symmetry or log-concavity of the population sampling distribution yields more regular-looking distributions.



(a) Weighted cumulative distribution functions

(b) Weighted density functions

Fig. 1: Illustration of the weighted distribution functions $\widehat{F}_w(\cdot)$ used to obtain upper endpoints $\hat{\mu}^+$ for our identification intervals. Panel (a) shows the raw cumulative distribution functions used to compute $\hat{\mu}^+$, while panel (b) uses smoothing splines to help visualize the underlying estimated densities. The dotted curve shows the observed empirical cumulative distribution, while the dashed, dot-dashed, and solid lines denote $\widehat{F}_w(\cdot)$ obtained with log-concavity, symmetry, or no constraints on the population distribution.

REFERENCES

ARONOW, P. M. & LEE, D. K. (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika* **100**, 235–240.

BOYD, S. & VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

COCHRAN, W. G. (1977). *Sampling Techniques*. New York: Wiley.

DE CARVALHO, M. & DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association* **109**, 764–776.

EFRON, B. & TIBSHIRANI, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics* **24**, 2431–2461.

FITHIAN, W. & WAGER, S. (2015). Semiparametric exponential families for heavy-tailed data. *Biometrika* **102**, 486–493.

HÁJEK, J. (1958). On the theory of ratio estimates. *Aplikace Matematiky* **3**, 384–398.

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

MARCUS, M. B. & SHEPP, L. A. (1972). Sample behavior of Gaussian processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. University of California Press.

MÖRTERS, P. & PERES, Y. (2010). *Brownian Motion*. Cambridge: Cambridge University Press.

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

PRÉKOPA, A. (1973). Logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum* **34**, 334–343.

QIN, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. Singapore: Springer.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

WALTHER, G. (2009). Inference and modeling with log-concave distributions. *Statistical Science* **24**, 319–327.

A. APPENDIX: PROOFS

Proof of Theorem 1

First, mirroring the argument of Aronow & Lee (2013), we note that $\hat{F}^*$ is itself an estimator of the form $\hat{F}^*(y) = \sum_i w_i 1\left(Y_i \leq y\right)$ with $\sum_i w_i = 1$ and $\max w_i / \min w_i \leq \gamma$, for some vector of weights w. Further, if $\hat{F}^* \in \mathcal{C}_{\gamma,\, n}$ then $\mu^* \in \mathcal{I}$. Now, if $\hat{F}^* \in \mathcal{C}_{\gamma,\, n}$ then $\Delta(\mu,\, \mathcal{I}) \leq |\hat{\mu}^* - \mu|$, so $\Delta(\mu,\, \mathcal{I}) \leq |\hat{\mu}^* - \mu|$ with probability tending to one. Finally, as $|\hat{\mu}^* - \mu|$ converges to zero by the weak law of large numbers, we have $\Delta(\mu,\, \mathcal{I})$ also converging to zero, in probability.

Proof of Lemma 1

In order to derive this type of uniform tail bound, we proceed in two steps. First, we verify that $n^{1/2}\left\{\widehat{F}^*(\cdot) - F(\cdot)\right\}$ converges in distribution to a tight Gaussian process $G(\cdot)$; then, we can bound the tail probabilities of $\sup_{y\in\mathbb{R}} |G(y)|$ directly. The first step is a routine application of Donsker's theorem as presented, e.g., in Chapter 19 of van der Vaart (1998); here, we take the existence of a limiting process $G(\cdot)$ as given. Because the supremum of $n^{1/2}\left\{\widehat{F}^*(\cdot) - F(\cdot)\right\}$ is invariant to monotone transformations of Y, and is stochastically maximized when Y has a continuous density without atoms, we can without loss of generality assume that $Y \in [0,\, 1]$.

Our sampling framework can be modeled as first drawing triplets $W_i = (Y_i, \pi_i, u_i) \sim F$, with u_i marginally distributed as a uniform $[0, 1]$, and then thinning our sample if $u_i > \pi_i$. We then make the assumption that the pre-thinned sample size is Poisson which implies the actual sample size we observe is also Poisson, $n \sim \text{Poisson}(N)$, with $N \to \infty$; this does not affect our conclusions, but makes the derivation more direct.

Given the Poisson assumption, we can, again without loss of generality, further assume that y is scaled on the $[0,\, 1]$ interval such that there exists a $\omega > 0$ for which

$$N\text{var}\left\{\hat{H}(y)\right\} = \omega^2 y, \quad \hat{H}(y) = \frac{1}{N}\sum_{(i:Y_i\leq y)} \frac{1}{\pi_i} \Big/ E_{\text{obs}}\left(\frac{1}{\pi_i}\right), \tag{A1}$$

where $E_{\text{obs}}\left\{f(W_i)\right\} = E\left\{f(W_i)|u_i \leq \pi_i\right\}$ denotes expectations with respect to the observed, i.e. thinned and therefore biased, sampling distribution. Equation (A1) is simply rescaling y so the variance of $\hat{H}(y)$, the total mass sampled below y, linearly increases with Y, allowing conception of this object as, effectively, a random walk. We note that $\hat{H}(y)$ is an oracle unnormalized estimator of $F(\cdot)$, akin to a Horvitz–Thompson estimator where we normalize by an expected weight rather than actual weights. It is unbiased for $F(y)$ in that, for any y, $E_{\text{obs}}\{\hat{H}(y)\} = F(y)$.

Now, $\hat{H}(y)$ is a compound Poisson process with variable size increments proportional to $1/\pi_i$ and a rate controlled by $F(y)$ and N. Using standard results on such compound Poisson processes, we have

$$N\text{var}\left\{\hat{H}(y)\right\} = E_{\text{obs}}\left\{\frac{1(Y_i \leq y)}{\pi_i^2}\right\} \Big/ E_{\text{obs}}\left(\frac{1}{\pi_i}\right)^2. \tag{A2}$$

In connecting (A1) and (A2), note the additional variability due to the random sample size n plays a critical role; the variable sample size n allows $\text{var}\{\hat{H}(y)\}$ to increase as there is no normalization constraint that sets $\hat{H}(1)$ to 1.

Because we are not yet normalizing $\hat{H}(y)$ we can easily obtain the pairwise covariances of $\hat{H}(y_1)$ and $\hat{H}(y_2)$ for any $y_1 < y_2$,

$$\text{cov}\left\{\hat{H}(y_1), \hat{H}(y_2)\right\} = \text{var}\left\{\hat{H}(y_1)\right\}. \tag{A3}$$

These pairwise covariances coupled with the linearly increasing variance implies that, if Gaussian limits exist, and we know that they do by Donsker's theorem, we must have that $N^{1/2}\left\{\hat{H}(y) - F(y)\right\}$ converges in distribution to $\omega W(y)$, where $W(y)$ is a standard Wiener process on $[0,\, 1]$. Then, noting that $\widehat{F}^*(y) = \hat{H}(y)/\hat{H}(1)$, we have, using the delta method, the convergence in distribution

$$n^{1/2}\left\{\widehat{F}^*(\cdot) - F(\cdot)\right\} \to G(y) = \omega\left\{W(y) - F(y)W(1)\right\},\ n \to \infty. \tag{A4}$$

We can use $n^{1/2}$ instead of $N^{1/2}$ thanks to Slutsky's theorem, since $n^{1/2} = N^{1/2}\{1 + o_P(1)\}$.

The next step is to bound ω. We do this by expressing ω in terms of population moments on the π_i, and then optimizing over all possible distributions of π_i to maximize this expression. First plug $y = 1$ into (A2) and (A1) and set them equal to each other to obtain

$$\omega^2 = E_{\text{obs}}\left(\pi_i^{-2}\right) / E_{\text{obs}}\left(\pi_i^{-1}\right)^2. \tag{A5}$$

Next we, defining E_{pop} as the expectation without the thinning step, i.e., with respect to the underlying, unbiased, population, derive the following equalities:

$$E_{\text{obs}}\left({\pi_i}^{-1}\right) = E_{\text{pop}}\left(\pi_i\right)^{-1}, \;\; E_{\text{obs}}\left(\pi_i^{-2}\right) = E_{\text{pop}}\left({\pi_i}^{-1}\right) / E_{\text{pop}}\left(\pi_i\right).$$

These are derived by simply writing out the expectation integrals of the thinned sample. For example, letting Π be the population distribution of the π_i,

$$E_{\text{obs}}\left(\pi_i^{-1}\right) = \frac{1}{\int \pi_i d\Pi} \int \frac{1}{\pi_i} \pi_i d\Pi = \frac{1}{E_{\text{pop}}\left(\pi_i\right)}.$$

These relations inserted into (A5) imply that

$$\omega^2 = E_{\text{pop}}\left({\pi_i}^{-1}\right) E_{\text{pop}}\left(\pi_i\right) \leq \sup_{a \in [0,\,1]} \left\{1 + a\left(\gamma - 1\right)\right\} \left\{1 + a\left(\gamma^{-1} - 1\right)\right\}, \tag{A6}$$

where the last inequality follows from the fact that, because $1/x$ is a convex function of x, our expression of interest is maximized for some distribution of π_i with all weight on two discrete values representing the endpoints γ and 1 of the allowed range (up to a scaling constant). The supremum over a is then the supremum over all distributions of this form, with a being the probability of $\pi_i = \gamma$. We can check by calculus that bound (A6) is maximized at $a^* = 1/2$, resulting in

$$\omega^2 \leq (1 + \gamma)(1 + \gamma^{-1})/4. \tag{A7}$$

It remains to bound suprema of $X(y) = W(y) - F(y)W(1)$. If we had $F(y) = y$, then $X(y)$ would be a standard Brownian bridge and the next steps would be straightforward. In our case, however, $F(y)$ may belong to a larger class of functions, so we instead make use of the following technical Lemma, proved at the end of this Section.

LEMMA A4. *Suppose that $W(y)$ is a standard Wiener process, and that $F(y)$ is a cumulative distribution function on $[0,\,1]$ with $F(0) = 0$, $F(1) = 1$ and, for some constant $\gamma > 0$, $F(A)/F(B) \leq \gamma$ for any intervals A and B with $\lambda(A) = \lambda(B)$, where $\lambda(\cdot)$ is the Lebesgue measure. Then, there exists a constant σ_γ, bounded by unity, for which the stochastic process $X(y) = W(y) - F(y)W(1)$ satisfies the bound*

$$\lim_{u \to \infty} \frac{1}{u^2} \log \text{pr}\left\{\sup_{y \in [0,\,1]} |X(y)| > u\right\} \leq \frac{-1}{2\sigma_\gamma^2}. \tag{A8}$$

The γ condition in the above will correspond to our control on the weights π_i due to the rescaled y. To continue, letting $X(y) = G(y)/\omega$ and plugging this and $u = \sigma_\gamma(2\log\alpha^{-1})^{1/2}$ into (A8), gives

$$\limsup_{\alpha \to 0} \;\; \log \text{pr}\left\{\sup_{y \in [0,\,1]} |G(y)| > \omega\sigma_\gamma(2\log\alpha^{-1})^{1/2}\right\}^{-1} \Big/ \log\alpha^{-1} \geq 1.$$

The final form of the lemma follows from plugging in (A7) for ω.

Proof of Corollary 1

By Lemma 1, we know that $\widehat{F}^* \in \mathcal{C}_{\gamma,\,n}$ with probability tending to 1, so the result follows immediately from Theorem 1.

Proof of Lemma 2

Lemma 2 and Lemma 1 begin the same way to obtain the tight Gaussian process, but then diverge in how we bound the tail behavior of this process. We start Lemma 2 with the representation in (A4). Given a $q \in [0, 0.5]$, use (A4) twice with $y = F^{-1}(q)$ and $y = F^{-1}(1-q)$, take the sum, and simplify to obtain the following:

$$\begin{aligned}\omega^{-1} n^{1/2} \left[\widehat{F}^* \left\{F^{-1}(q)\right\} + \widehat{F}^* \left\{F^{-1}(1-q)\right\} - 1\right] &\to W\left\{F^{-1}(q)\right\} + W\left\{F^{-1}(1-q)\right\} - W(1) \\ &= W\left\{F^{-1}(q)\right\} - \left[W(1) - W\left\{F^{-1}(1-q)\right\}\right] \\ &= W\left\{F^{-1}(q) + 1 - F^{-1}(1-q)\right\},\end{aligned}$$

where the last equality is in distribution. For the last step note that $W\left\{F^{-1}(q)\right\}$ and $W(1) - W\left\{F^{-1}(1-q)\right\}$ are two independent Gaussian processes over $q \in [0,\, 0.5]$; $W(1) - W(y_2)$ is independent of $W(y_1)$ if $y_1 < y_2$ as we are effectively looking at two non-overlapping, and therefore independent, sums of our sample. Then the sum of two independent random walks is a random walk of the total distance. Finally, we note that, for $q \in [0,\, 0.5]$, $F^{-1}(q) + 1 - F^{-1}(1-q)$ takes all values in $[0,\, 1]$, and so we find that for any threshold t,

$$\begin{aligned}\lim_{n\to\infty} \mathrm{pr}\left(n^{1/2} \sup_{q\in[0,\,0.5]} \left[\widehat{F}^* \left\{F^{-1}(q)\right\} + \widehat{F}^* \left\{F^{-1}(1-q)\right\} - 1\right] \geq \omega t\right) &= \mathrm{pr}\left\{\sup_{y\in[0,\,1]} W(y) \geq t\right\} \\ &= 2\Phi(-t),\end{aligned}$$

where the last equality is a consequence of the reflection principle for Brownian motion (e.g., Mörters & Peres 2010, page 44). The desired conclusion then follows by noting our bound (A7) on ω and applying the above bound to both tails. This lemma does not require any symmetry in $F(\cdot)$; the subsequent use of the lemma is what takes advantage of that structure.

Proof of Lemma 3

It suffices to verify all claims made in the four steps leading to our construction of $\mathcal{C}^{\mathrm{LC}+}_{\gamma,\,n}$. Step 1 is a direct consequence of the Kolmogorov–Smirnov theorem. Steps 3 and 4 are also immediate given the result of Prékopa (1973). The optimization problem used to define $\widehat{L}_m(y)$ is computationally tractable: Finding a concave upper-bound for a function $l(y)$ is equivalent to taking the convex hull of the curve $(y,\, l(y))$.

It remains to check Step 2, which is a direct analogue to the argument of Aronow & Lee (2013) applied to the sampling distribution $S(\cdot)$. Consider $r = \min(\pi_i)/E(\pi_i)$. If r were known (and recalling that $\max \pi_i / \min \pi_i \leq \gamma$), we can follow the argument of Aronow & Lee (2013) to verify that F is stochastically dominated by U_m, i.e., $F(y) \geq U_m(y)$ for all $y \in \mathbb{R}$, where $U_m(\cdot)$ is a reweighting of the observed cumulative sampling distribution $S(\cdot)$ given by

$$U_m(y) = \left[S(y) + (\gamma - 1)\left\{S(y) - S(m)\right\}_+\right] \,/\, \left[S(m) + \gamma\left\{1 - S(m)\right\}\right],$$

for a threshold m characterized by

$$S(m) \,/\, \left[S(m) + \gamma\left\{1 - S(m)\right\}\right] = r.$$

In other words, $U_m(y)$ corresponds to a re-weighting of $S(y)$ with sampling weights π_i jumping from a low value a to a high value γa at threshold m. In practice, of course, we do not know r, but we still know that F is stochastically dominated by U_m for some $m \in \mathbb{R}$. The claim made in Step 2 then follows by noting that $S(y) \geq \widehat{S}_{\mathrm{KS}}(y)$, implying $U_m(y) \geq \widehat{U}_m(y)$ for all $m,\, y \in \mathbb{R}$ with probability tending to 1 thanks to Step 1.

*Proof of Lemma A*4

We first recall the classic result of Marcus & Shepp (1972) which, in our situation, implies that the Gaussian process $X(y)$ satisfies

$$\lim_{u\to\infty} \frac{1}{u^2} \log \mathrm{pr}\left(\sup_{y\in[0,\,1]} |X(y)| > u \right) = \frac{-1}{2 \sup_{y\in[0,\,1]} \mathrm{var}\left\{X(y)\right\}};$$

it thus remains to use the shape restrictions on $F(y)$ to lower-bound the right-hand expression. To do so, it is helpful to decompose the stochastic process $X(y)$ into two independent parts:

$$X(y) = B(y) + W(1)\left\{y - F(y)\right\},$$

where $B(y) = W(y) - yW(1)$ is a standard Brownian bridge. Moreover, given our assumptions on $F(y)$,

$$|F(y) - y| \le A_\gamma(y), \;\; A_\gamma(y) = \frac{\gamma y}{1 - y + \gamma y} - y,$$

for $0 \le y \le 0.5$ ($A_\gamma(y)$ corresponds to how much more weight below y, as compared to a uniform distribution, is possible given the constraints of γ). A similar derivation shows $A_\gamma(y) = A_\gamma(1-y)$ for all $y \in [0.5,\, 1]$. Thus, noting that $\mathrm{var}\left\{B(y)\right\} = y(1-y)$ and $\mathrm{var}\left\{W(1)\right\} = 1$, we see that, using $A_\gamma(y)$ as a bound on the $y - F(y)$ scaling of the $W(1)$ term,

$$\mathrm{var}\left\{X(y)\right\} \le y(1-y) + A_\gamma(y), \;\; 0 \le y \le 1.$$

The above function is concave on $[0,\, 0.5]$, in particular, $A''_\gamma(y) = -2\gamma(\gamma-1) \,/\, (1 - y + \gamma y)^3$, and so the above expression has a unique maximizer y^* that can be derived numerically. The desired conclusion then holds for $\sigma^2_\gamma := y^*_\gamma(1 - y^*_\gamma) + A_\gamma(y^*_\gamma)$; the fact that $\sigma^2_\gamma \le 1$ is immediate by inspection.

[*Received April* 2012. *Revised September* 2012]