Strategic Sample Selection

Alfredo Di Tillio[1] Marco Ottaviani[2] Peter N. Sørensen[3]

October 28, 2017

[1]Bocconi
[2]Bocconi
[3]Copenhagen

Impact of Sample Selection on Quality of Inference?

- Typically **observational data** are *non-randomly selected*:
  - Either self selection induced by choices made by subjects
  - Or selection from sample inclusion decisions made by analysts
- **Experimental data** can also suffer from selection problems challenging *internal validity*: **subversion of randomization** to treatment/control
  - Inadequate allocation concealment increases treatment effects by as much as 41%, according to Schulz et al. (1995)
  - Berger (2005) documents researchers' ability to subvert assignment of patients depending on expected outcomes toward end of block
- When treatment is given to healthiest rather than random patient:
  - Favorable outcomes are weaker evidence that treatment is effective
  - But how is accuracy of evaluator's inference affected?

Impact of Selection on Inference?

- When feeding a **consumer review** to potential buyers with limited attention:
  - Should an e-commerce platform post a random review or allow merchant to cherry-pick one?
- Similarly, **peremptory challenge** gives a defendant the right to strike down a number of jurors:
  - Given that the defendant selects the most favorable jurors, how is quality of final judgement affected?
- When **testing** a student in an **exam**:
  - Should teacher pick a question at random or allow student to select most preferred question out of a batch?

Outline

1. **Statistical model: Simple hypothesis testing under MLRP**
2. Global impact of selection on evaluator
    - Lehmann's dispersion for comparison of location experiments
    - Analysis of F^k, distribution of max of k iid variables, as k varies
3. Local impact of selection on evaluator
    - Local version of Lehmann's dispersion
    - Local effects of varying k
    - Extreme selection $k \to \infty$ and link to extreme value theory
4. Strategic selection
    - Equilibrium persuasion
    - Impact on researcher's payoff from selection
    - Impact of uncertain and unanticipated selection

Setup

- Evaluator interested in the true value of unknown state $\theta \in \{\theta_L, \theta_H\}$
  - Here, $\theta_H > \theta_L$, and prior $p = \Pr(\theta_H)$
- Data: Evaluator observes a signal $x = \theta + \epsilon$
- Noise ϵ independent from θ, with known c.d.f. F (experiment)
  - Assume **logconcave** density f

- Manipulation will shift the distribution of ε
- Specifically: F is shifted to F^k where $k > 1$
  - First-order stochastic higher ε and x
- As if ε is **selected**: best of k independent draws
- We will focus on a rational evaluator, aware of selection
  - For this evaluation, can proceed for now with some known F

Information and Optimal Decision

- Evaluator's reservation utility R
- Decision payoff for Evaluator:

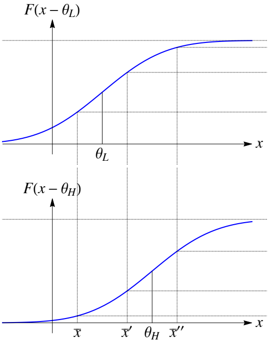
	state θ_L	state θ_H
reject	R	R
accept	θ_L	θ_H

  - Case of interest: $\theta_L < R < \theta_H$
- Evaluator accepts iff $\Pr(\theta_H|x)\theta_H + (1 - \Pr(\theta_H|x))\theta_L \geq R$
- Optimal strategy is a **cutoff rule**: accept iff

$$\underbrace{\ell_F(x) := \frac{f(x-\theta_H)}{f(x-\theta_L)}}_{\text{Likelihood Ratio}} \geq \underbrace{\bar{\ell} := \frac{1-p}{p}\frac{R-\theta_L}{\theta_H - R}}_{\text{Acceptance Hurdle}}$$

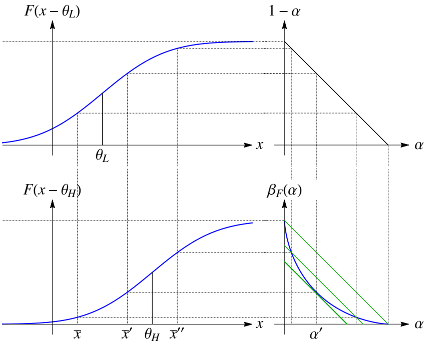
- Log-concavity of $f \Rightarrow$ **M**onotone **L**ikelihood **R**atio **P**roperty
  - $\ell_F(x)$ is increasing $\Rightarrow$ Optimal to accept iff $x \geq \bar{x}_F^*(\bar{\ell})$

False Positives v. False Negatives



- For every $-\infty \leq \bar{x}_F^*(\bar{\ell}) \leq \infty$, $\alpha = 1 - F(\bar{x} - \theta_L)$ and $\beta = F(\bar{x} - \theta_H)$
- Higher cutoff $\bar{x}_F^*(\bar{\ell})$ results in
  - decrease in type I errors (false positives) α
  - increase in type II errors (false negatives) β

Information Constraint (a.k.a. ROC curve, qq plot)



- Define the **Info**rmation **C**onstraint of Experiment F as

$$\beta = \beta_F(\alpha) = F\big(F^{-1}(1-\alpha) + \theta_L - \theta_H\big),$$

decreasing and convex (by logconcavity/MLRP)

Problem Reformulation

- Reformulate evaluator problem in terms of α and β
- Disregarding constants, evaluator maximizes

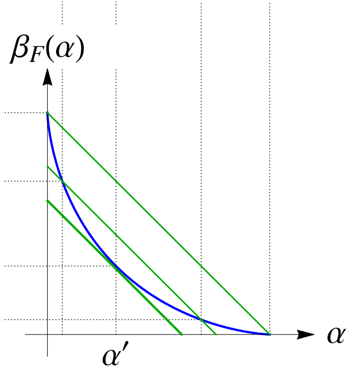
$$\underbrace{-(1-p)(R-\theta_L)}_{\text{MC False Pos.}}\alpha \overbrace{-p(\theta_H - R)}^{\text{MC False Neg.}}\beta$$

subject to the **InfoC**

$$\beta_F(\alpha) = F\left(F^{-1}(1-\alpha) + \theta_L - \theta_H\right)$$

- Substituting **InfoC** & $\bar{\ell} = \frac{1-p}{p}\frac{R-\theta_L}{\theta_H - R}$ into objective function, problem is

$$\min_{\alpha} \bar{\ell}\alpha + \beta_F(\alpha)$$



$$\begin{aligned} \min_{\alpha} \quad & \bar{\ell}\alpha + \beta \\ \text{s.t.} \quad & \beta = \beta_F(\alpha) = F\big(F^{-1}(1-\alpha) + \theta_L - \theta_H\big) \end{aligned}$$

Random v. Selected Experiment

- Compare two regimes:
  - Random data point, experiment F
  - Selected data point, experiment F^k
    - density $kF^{k-1}f$ still logconcave by Prekopa's theorem
- Threshold becomes

$$\ell_{F^k}(x) = \left[\frac{F(x-\theta_H)}{F(x-\theta_L)}\right]^{k-1} \ell_F(x) \geq \bar{\ell}.$$

- Is evaluator better off with F or with $G = F^k$?
- More generally, let's compare F and G

Outline

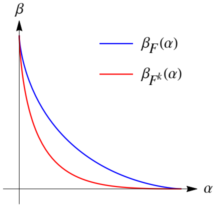
1. Statistical model: Simple hypothesis testing under MLRP
2. **Global impact of selection on evaluator**
   - **Lehmann's dispersion for comparison of location experiments**
   - **Analysis of F^k as k varies**
3. Local impact of selection on evaluator
   - Local version of Lehmann's dispersion
   - Local effects of varying k
   - Extreme selection $k \to \infty$ and link to extreme value theory
4. Strategic selection
   - Equilibrium persuasion
   - Impact on researcher's payoff from selection
   - Impact of uncertain and unanticipated selection

Comparison of Experiments

- G is preferred to F iff

$$\bar{\ell}\alpha_G^*(\bar{\ell}) + \beta_G(\alpha_G^*(\bar{\ell})) \leq \bar{\ell}\alpha_F^*(\bar{\ell}) + \beta_F(\alpha_F^*(\bar{\ell})),$$

- So, G is *globally* $(\forall\ R, q, \theta_H > \theta_L)$ preferred to F iff $\beta_G(\alpha) \leq \beta_F(\alpha)\ \forall\ \alpha$



- <u>Example</u>: $F = \mathcal{N}(0, 1)$ and $G = F^k$
- Laxer constraint: better power $1 - \alpha$ for any significance $1 - \beta$

Comparison of Experiments

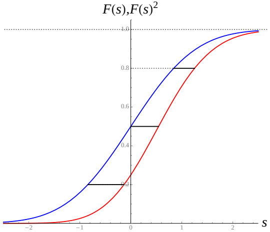
- Lehmann (1988) orders experiments without computing **InfoC**:

G is *globally* preferred to $F \Leftrightarrow G$ is *less dispersed* than F

- Definition G less dispersed than F: $G^{-1} - F^{-1}$ is decreasing, i.e.

$$G^{-1}(v) - F^{-1}(v) \leq G^{-1}(u) - F^{-1}(u) \qquad \text{for all } 0 < u < v < 1.$$

- Intuition: Constraint $F^{-1}(1-\alpha) - F^{-1}(\beta) = \theta_H - \theta_L$ relaxed with G



Global Comparison Based on Dispersion

- **Double Logconvexity** Theorem

$$F^k \text{ is less (more) dispersed the greater is } k \geq 1$$

$$\Leftrightarrow$$

$$-\log\left(-\log F\right) \text{ is convex (concave)}$$

- Corollary
  The evaluator prefers $F^{k'}$ to F^k (resp. F^k to $F^{k'}$) for all $k' \geq k \geq 1$ and all parameter values (θ_L, θ_H, p, and R) if and only if $-\log(-\log F)$ is convex (resp. concave)

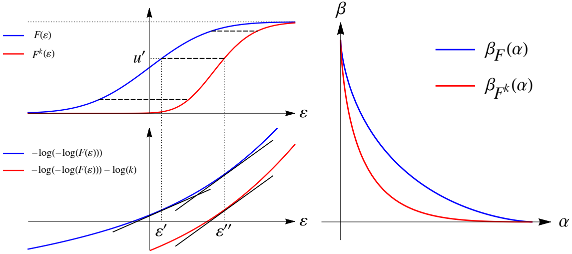
Double Logconvexity Theorem: Intuition

- Rewrite the condition of G less dispersed than F as:

$$f\big(F^{-1}(u)\big) \leq g\big(G^{-1}(u)\big) \qquad \text{for all } 0 < u < 1.$$

G at quantile $G^{-1}(u)$ is steeper than F at quantile $F^{-1}(u)$, $\forall u$

- Transform F and F^k by strictly increasing $u \mapsto -\log(-\log u)$
- Transformed functions are parallel shifts of each other: $-\log(-\log F^k) = -\log(-\log F) - \log k$



Special cases

- Gumbel's Extreme Value Distribution $F(\varepsilon) = \exp(-\exp(-\varepsilon))$
  - F is such that $-\log(-\log F)$ is linear—both convex and concave
  - For every k the experiment F^k is neither less nor more dispersed than F and the evaluator is therefore indifferent to selection
- Logistic distribution: $F(\varepsilon) = \frac{1}{1+e^{-\varepsilon}}$
  - Double logconvex, so selection benefits evaluator
- Exponential distribution: $F(\varepsilon) = 1 - e^{-\varepsilon}$, for $\varepsilon \geq 0$
  - Double logconcave, so selection harms evaluator

Analysis of Double Logconvexity

- $-\log(-\log F)$ is convex function if and only if

$$\frac{\overbrace{\frac{f(\varepsilon)}{F(\varepsilon)}}^{\text{reverse hazard rate}}}{\underbrace{\log F(\varepsilon)}_{\text{reverse hazard function}}} \text{ is decreasing}$$

  - The reverse hazard rate decreases less fast than the cumulative reverse hazard rate increases

- Equivalently, F has a quantile density function **less elastic** than Gumbel's

$$-\frac{\frac{f'(\varepsilon)}{f(\varepsilon)}}{\frac{f(\varepsilon)}{F(\varepsilon)}} < -\frac{1+\log F\left(\varepsilon\right)}{\log F\left(\varepsilon\right)} \qquad \text{for all } \varepsilon$$

Empirical Diagnostic Test

- We derive a practical diagnostic test in actual experimental studies
  - where it may be *unknown* whether selection occurred
- **Selection-invariance property** of double logconvexity/logconcavity:
  - $-\log(-\log F)$ and $-\log(-\log F^k)$ differ only by a constant $\Longrightarrow$

$$F \text{ double log-concave} \Longleftrightarrow F^k \text{ double log-concave}$$

- **Double log-concave data distributions should "raise a flag"**
  - *if selection does occur, analyst is bound to having less informative data*
- If data is double logconvex instead
  - selection actually results in a more informative experiment, if analyst properly adjusts for selection

Outline

1. Statistical model: Simple hypothesis testing under MLRP
2. Global impact of selection on evaluator
   - Lehmann's dispersion for comparison of location experiments
   - Analysis of F^k as k varies
3. **Local impact of selection on evaluator**
   - **Local version of Lehmann's dispersion**
   - **Local effects of varying** k
   - **Extreme selection** $k \to \infty$ **and link to extreme value theory**
4. Strategic selection
   - Equilibrium persuasion
   - Impact on researcher's payoff from selection
   - Impact of uncertain and unanticipated selection

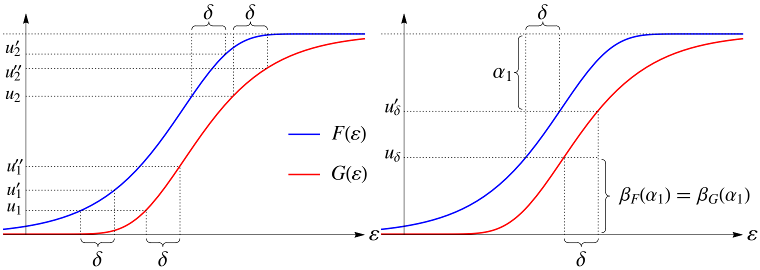
Local Dispersion

- First rewrite the condition of G less dispersed than F as:

$$F(F^{-1}(u)+\delta) \leq G(G^{-1}(u)+\delta) \qquad \text{for all } \delta > 0 \text{ and } 0 < u < 1.$$

- <u>Definition</u> (Local Dispersion)
  Experiment G is *locally less δ-dispersed* than experiment F on $[u_1, u_2] \subseteq [0,1]$ if

$$F(F^{-1}(u)+\delta) \leq G(G^{-1}(u)+\delta) \qquad \text{for all } u_1 \leq u \leq u_2$$



Local Dispersion Theorem

- Equivalence between:
  - G less dispersed than F for a specific δ and for all u in some interval
  - G preferred to F in a corresponding interval
- **Local Dispersion** Theorem: Let $\delta = \theta_H - \theta_L$. For all $N \geq 1$, the following conditions are equivalent:

(L) There exist $0 = \ell_1 \leq \cdots \leq \ell_{2N+1} = \infty$: for all $n = 1, \ldots, N$, the evaluator prefers F to G for $\bar{\ell} \in [\ell_{2n-1}, \ell_{2n}]$ and G to F for $\bar{\ell} \in [\ell_{2n}, \ell_{2n+1}]$.

(A) There exist $1 = \alpha_1 \geq \cdots \geq \alpha_{2N+1} = 0$: $\forall n = 1, \ldots, N$, $\beta_F(\alpha) \leq \beta_G(\alpha)$ for all $\alpha \in [\alpha_{2n}, \alpha_{2n-1}]$ and $\beta_F(\alpha) \geq \beta_G(\alpha)$ for all $\alpha \in [\alpha_{2n+1}, \alpha_{2n}]$.

(D) $\exists 0 = u_1 \leq \cdots \leq u_{2N+1} = 1$: $\forall n = 1, \ldots, N$, F is locally less δ-dispersed than G on $[u_{2n-1}, u_{2n}]$ and more δ-dispersed than G on $[u_{2n}, u_{2n+1}]$.

Local Dispersion: Idea

- Fix $\delta = \theta_H - \theta_L > 0$
- Consider any given β
- Under F, we obtain α_F on the information constraint curve,

$$\delta = F^{-1}(1 - \alpha_F) - F^{-1}(\beta)$$

- G does better with this β

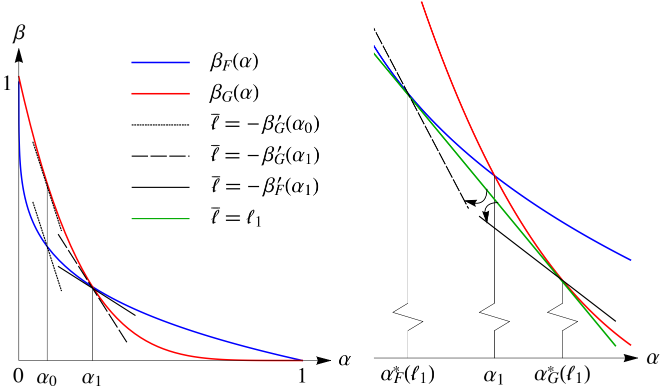
$$G^{-1}(1 - \alpha_F) - G^{-1}(\beta) < \delta = F^{-1}(1 - \alpha_F) - F^{-1}(\beta)$$

  i.e.,

$$G(G^{-1}(\beta) + \delta) < F(F^{-1}(\beta) + \delta)$$

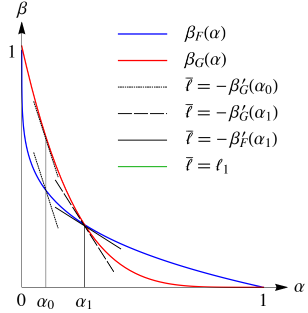
- In particular, if G^{-1} is flatter than F^{-1} at β, this is true when δ is small

Info Constraint Crossing Really Matters



Bayesian vs. Frequentist Evaluator

- Frequentist Evaluator fixes $\tilde{\alpha}$ and prefers the experiment with higher $\beta(\tilde{\alpha})$
- Bayesian Evaluator reoptimizes $\tilde{\alpha}$ for every experiment



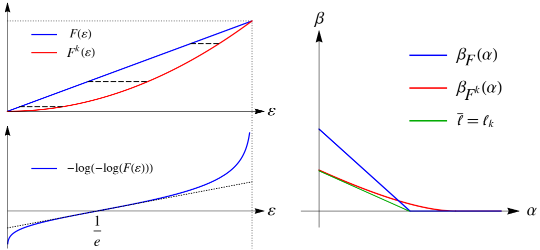
- Bayesian and Frequentist Evaluator agree iff $\beta_G(\alpha) \leq \beta_F(\alpha)\ \forall \alpha$

Locally Variable Impact of Selection

- Back to comparison of F and F^k
- Focus on F^k first more & then less locally dispersed than F
- Proposition: Let F be an experiment such that $-\log(-\log(F))$ is first concave (resp. convex) and then convex (resp. concave). Then for every $k \geq 1$ there exists ℓ_k such that the evaluator prefers F to F^k (resp. F^k to F) for $\bar{\ell} \leq \ell_k$ and F^k to F (resp. F to F^k) for $\bar{\ell} \geq \ell_k$
- If F is first double log-concave and then double log-convex
  - quantile difference $(F^k)^{-1}(u) - F^{-1}(u)$ is first increasing and then decreasing in u
- Selection hurts evaluator less concerned about type I errors: low $\bar{\ell}$
  - benefits for high acceptance hurdle $\bar{\ell}$

Uniform Example

- Uniform distribution, $F(\varepsilon) = \epsilon$ for $\epsilon \in [0, 1]$
- Double-log transformation of F is $-\log(-\log(\varepsilon))$
- Concave for $\varepsilon \leq 1/e$ & convex for $\varepsilon \geq 1/e$
- **Bell-shaped quantile difference**



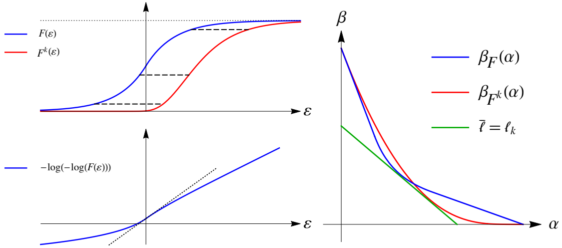
- Evaluator is hurt by selection when concerned about type II errors (low $\bar{\ell}$)
    - benefits from selection when more concerned about type I errors (high $\bar{\ell}$)

Laplace Example

- Laplace distribution

$$F(\varepsilon) = \begin{cases} \frac{e^{\varepsilon}}{2} & \text{for } \varepsilon < 0 \\ 1 - \frac{e^{-\varepsilon}}{2} & \text{for } \varepsilon \geq 0 \end{cases}$$

- Double-log transformation of F is convex for $\varepsilon < 0$ and concave for $\varepsilon > 0$
- **U-shaped quantile difference**



- Evaluator benefits from selection for low $\bar{\ell}$ but is hurt for high $\bar{\ell}$

Extreme Selection

- What happens when presample size $k \to \infty$?
- Suppose that, for some nondegenerate distribution $\bar{F}$ and for some location and scale **normalization sequences** b_k and $a_k > 0$

$$F^k(b_k + a_k\varepsilon) \to \bar{F}(\varepsilon)$$

  for every continuity point ε of $\bar{F}$
- By the **Fundamental Theorem of Extreme Value Theory**
  - $\bar{F}$ is Gumbel, Extreme Weibull or Frechet
  - For logconcave F, either Gumbel or Extreme Weibull

Extreme Selection: Results

- Distribution of noise term is systematically shifted upwards as k increases
- Location *normalization sequence* b_k is growing
  - but evaluator can adjust for any translation without impact on payoff
- **Limit impact of selection thus hinges on**
  - **whether the scale normalization sequence a_k shrinks to zero or not**

1. If $a_k \to 0$, noise distribution is less and less dispersed as k grows
   - evaluator gets arbitrarily precise information about the state
2. If instead we can choose a constant sequence a_k
   - extreme selection based on experiment F amounts to a **random experiment based on $\bar{F}$**

Extreme Selection - Exponential Power Family

- Proposition: Let F be an **exponential power distribution**

$$f(\varepsilon) = \frac{s}{\Gamma(1/s)} e^{-|\varepsilon|^s}$$

of shape $s > 1$. As $k \to \infty$, the limiting distribution has Gumbel shape, and there is arbitrarily precise information about the state

  - But the limit result is very different when $s = 1$, Laplace
  - Laplace (like exponential) converges to Gumbel with $a_k = 1$ for each k

Outline

1. Statistical model: Simple hypothesis testing under MLRP
2. Global impact of selection on evaluator
   - Lehmann's dispersion for comparison of location experiments
   - Analysis of F^k as k varies
3. Local impact of selection on evaluator
   - Local version of Lehmann's dispersion
   - Local effects of varying k
   - Extreme selection $k \to \infty$ and link to extreme value theory
4. **Strategic selection**
   - **Equilibrium persuasion**
   - **Impact on researcher's payoff from selection**
   - **Impact of uncertain and unanticipated selection**

Strategic Selection

- So far we assumed that the researcher is willing to show a selected experiment to the evaluator
- We now verify this posited behavior is an **equilibrium in natural game**
- Assume **researcher is fully biased toward acceptance**
  i.e. bears no losses due to type I errors

Selective Sampling Game - Setting

- **Timeline**
  1. Researcher privately observes $\varepsilon_1, \ldots, \varepsilon_k$
  2. Researcher chooses $i \in \{1, \ldots, k\}$
  3. Evaluator observes $x_i = \theta + \varepsilon_i$
  4. Evaluator chooses whether to accept or reject

- **Payoffs**
  - Evaluator: Same as before
  - Researcher:
    - 0 if the evaluator rejects
    - 1 if the evaluator accepts

Selective Sampling Game - Equilibrium

- Proposition: There exists a Bayes Nash equilibrium where the researcher chooses maximal selection, $i \in \arg\max_{1 \leq j \leq k} \varepsilon_j$, and the evaluator accepts for signals x satisfying

$$\frac{F^{k-1}(x-\theta_H)f(x-\theta_H)}{F^{k-1}(x-\theta_L)f(x-\theta_L)} \geq \bar{\ell}$$

- The researcher's strategy is a best response because the evaluator will observe a higher signal and will be more likely to accept

Equilibrium Impact of Selection on Researcher's Welfare

- Impact of selection on researcher's welfare
  - depends on direction of change in pair (α, β) chosen by evaluator
- For any pair (α, β), the researcher's payoff is

$$p(1-\beta) + (1-p)\alpha.$$

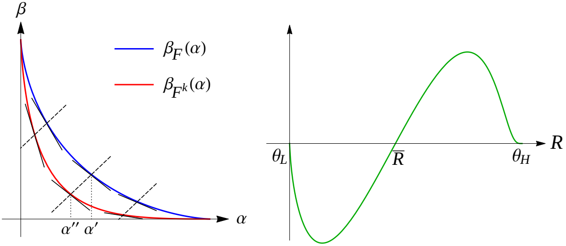
- Thus, a generic indifference curve of the researcher is a line of the form

$$\beta = \left(1 - \frac{u}{p}\right) + \frac{1-p}{p}\alpha,$$

  where $0 \leq u \leq 1$ is researcher's payoff
- Researcher benefits from selection $\Leftrightarrow$ Evaluator reacts to selection (experiment F^k) by choosing a new pair (α', β') **below and to right** of indifference line going through optimal pair in experiment F

Equilibrium Impact of Selection on Researcher's Welfare

- Intuitively:
  - If R is high, informative selection increases the acceptance chance
  - but info-reducing selection reduces acceptance
  - Conversely, when R is low
- To illustrate consider normal noise, with $\beta_{F^k}(\alpha) \geq \beta_F(\alpha)$



Impact of Selection on Researcher's Welfare: Examples

- **Gumbel example–pure rat race**
  - Selection is welfare neutral for evaluator & researcher

- Laplace example:
  - Evaluator is worse off with F^k than with F for large values of R
  - Researcher is hurt by selection for small or large values of R, but benefits for intermediate values
  - **Credibility Crisis** at high R — both parties lose from selection

- Uniform example:
  - Evaluator is better off with F^k than with F for large values of R
  - Researcher benefits for small or large values of R, but hurt for intermediate values

Data Production

- At $t = 0$ researcher privately sets presample size k
  - at increasing & convex cost $C(k)$
- Evaluator correctly anticipates k optimally chosen by the researcher:
  - best responds with acceptance at $\bar{x}$
- Researcher correctly anticipates acceptance threshold $\bar{x}$ and

$$\max_k p\left(1 - F^k(\bar{x} - \theta_H)\right) + (1 - p)\left(1 - F^k(\bar{x} - \theta_L)\right) - C(k),$$

concave problem

Equilibrium with Data Production

- Proposition: Equilibrium is characterized as the solution $(\bar{x}, k)$ to

$$\frac{F^{k-1}(\bar{x}-\theta_H)f(\bar{x}-\theta_H)}{F^{k-1}(\bar{x}-\theta_L)f(\bar{x}-\theta_L)} = \bar{\ell}$$

  and

$$-p\log\left(F\left(\bar{x}-\theta_H\right)\right)F^k\left(\bar{x}-\theta_H\right)-(1-p)\log\left(F\left(\bar{x}-\theta_L\right)\right)F^k\left(\bar{x}-\theta_L\right) = C'\left(k\right)$$

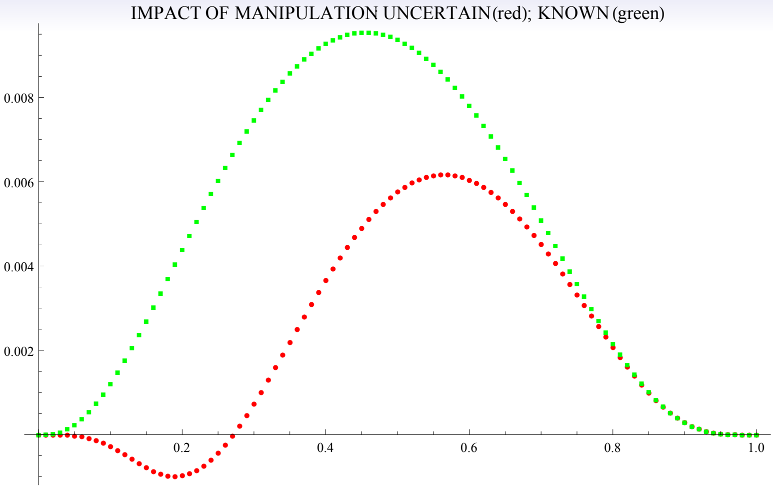
- Rat race effect:
  - Evaluator correctly anticipates degree k of selection
  - $\Rightarrow$ manipulation cost $C\left(k\right)$ largely wasted
- Gumbel example:
  - Apart from $C(k)$, payoffs independent of k
  - Researcher would gain from making k observable

Evaluator's Value of Commitment

- Slope of researcher's best response $k(\bar{x})$ depends on parameters:
  - When prior strongly favors rejection, $F(\bar{x} - \theta_L)$ is sufficiently small
    - best response k is an increasing function of $\bar{x}$
  - When the prior strongly favors acceptance
    - best response k is a decreasing function of $\bar{x}$
- Under double logconvexity, evaluator wants to induce greater k
  - **Commit** to a weaker standard for high R
- Conversely, when evaluator loses from greater k

Uncertainty of Manipulation: Negative Impact

- Under uncertain selection, evaluator does not know whether researcher manipulates — the number k is random
  - In location experiment, difficult to adjust estimate correctly
  - Logconcavity could fail, so monotonicity could fail: some experimental results may be "too good to be true"
- Consider the Gumbel case
  - If the evaluator knew realized k, since F^k is as effective as F, the randomness made no difference
  - Not knowing k is then Blackwell worse
- More general force: Uncertainty in selection harms evaluator



- Evaluator's payoff gain at $k = 2$ over benchmark, Normal example
  - Red curve has equal chance of $k = 1, 2$

Impact on Unwary Evaluator

- We have assumed that the evaluator correctly anticipates k
- If not, the threshold s^* does not adjust to k
  - No doubt that the researcher gains from raising k (gross)
- The impact on the evaluator turns out to be ambiguous
  - Wrong threshold: bad
  - More informative experiment: good
- Under symmetry and equipoise, indifference to $k = 1, 2$
  - Equipoise: $\bar{\ell} = 1$. Symmetry, $F(1 - \varepsilon) = 1 - F(\varepsilon)$

Literature

- **Selection bias**: Heckman (1979)
  - Methods to estimate and test, correcting for bias
- **Subversion of randomization**: Blackwell and Hodges (1957) *assume* Evaluator loses from manipulation
  - Characterize optimal randomization mechanism to be unpredictable
- **Disclosure**: Grossman (1980), Milgrom (1981), ..., Henry (2009)
- **Information provision & persuasion**: Johnson and Myatt (2005) & Kamenica and Gentzkow (2011)
  - The researcher freely chooses an experiment
- **Selective disclosure**: Fishman and Hagerty (1990), Glazer and Rubinstein (2004), Hoffmann, Inderst & Ottaviani (2014)
  - Here, systematic study of selection, based on statistical properties
- **Signal jamming**: Holmström (1999)
- **Selective trials**: Chassang, Padró and Snowberg (2012)

Summary

- We develop tractable model of challenges to **internal validity:**
  1. Dispersion of $G = F^k$ decreases in $k \Leftrightarrow -\log(-\log F(\varepsilon))$ is convex
     - Then, selection has **global and monotonic impact** on evaluator
  2. To provide general characterization of impact of selection, we compare **any** two experiments F, G based on **local dispersion**, for a **subset** of parameters
     - We compare experiments when $G^{-1}(p) - F^{-1}(p)$ is not monotonic
- Evaluator **benefits** from **known** sample selection *unless*
  - Data has sufficiently thin tails & prior strongly favors acceptance
  - Data has sufficiently thick tails & prior strongly favors rejection
- **Uncertain** manipulation tends to **harm** evaluator

Open Questions

- In companion paper we developed toy (all-binary) model of sample selection challenging *external validity*
  - Alcott (2015) documents hard-to-control-for site selection: study sample **not representative** of population of interest
  - Initial trials are implemented in high impact sites, then impact declines, $\Rightarrow$ no reliable inference of ATE even after sample of 8 million Americans!

LITERATURE on Stochastic Orders of Order Statistics

- No existing results for strictly logconcave distributions
- Khaledi and Kochar's (2000) Thm 2.1: if X_i's are i.i.d. according to F with Decreasing Hazard Rate (DHR), $X_{i:n}$ is less dispersed than $X_{j:m}$ whenever $i \leq j$ and $n - i \geq m - j$. Thus, for $i = n = 1$ and $j = m = k$: **If F has DHR, F^k is more dispersed than F**
- By Prekopa's Thm: **Logconcavity => IHR**
- Thus exponential (loglinear, with constant HR) is the only logconcave distribution to which Khaledi and Kochar applies
- Converse of Khaledi and Kochar'sThm 2.1 *not* valid for IHR distribution
- Our characterization applies to logconcave distributions

Testing for Double Logconvexity: Approach

- Suppose researcher obtains data $(x_1, ..., x_N)$ and estimates $\hat{\theta}$
- Residuals $\varepsilon_n = x_n - \hat{\theta}$ are independent draws from F^k
- Under assumption of homogeneous treatment effect, test can be performed on ε_n or $(x_1, ..., x_N)$
  - Use Kolmogorov-Smirnov 2-sample test to evaluate homogeneity in treatment effect, comparing treatment and control distribution
- Double logconvexity of F is equivalent to concavity of $\log(-\log F)$
  - **IDEA: test for logconcavity of** $-\log F$
- Similarly, to test double logconcavity of F $\Leftrightarrow$logconcavity of $\frac{1}{-\log F}$

Testing for Double Logconvexity/Logconcavity: Procedure

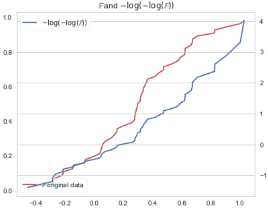
- We extend Hazelton's (2011) test for logconcavity
  - start from empirical CDF F of an outcome variable
  - compute the non-negative transformation $-\log F$
  - rescale it to integrate to one over the original support
- The test requires as input a sample generated by the density whose logconcavity we want to test, so we cannot just use original sample,but
  - we can treat this transformation as a PMF and
  - draw an independent random sample from it
- Run the test for logconcavity on the simulated sample:

$$\begin{cases} H_0 : \text{transformed density is logconcave (=dlogcx)} \\ H_1 : \text{transformed density is not logconcave} \end{cases}$$

- Replacing $-\log F$ with $\frac{1}{-\log F}$ we have H_0 =dlogcv

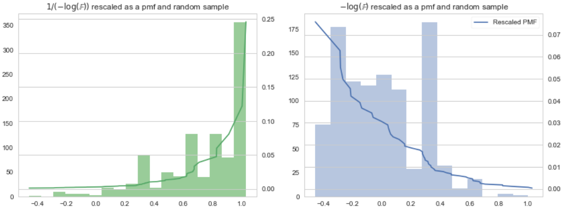
Application to Andrabi, Das, and Khwaja (2017) AER

- Field experiment on
  - impact of providing test scores on educational markets
- Considered outcome variable: scores in treated villages
  - K-S test for homogeneity in distributions returns p-value$>$0.3
  - Test for logconcavity of original sample: p-value$>$0.77
- Left: Distribution of original outcome variable
- Right: Computed empirical F (red) and $-\log(-\log F)$ (blue)



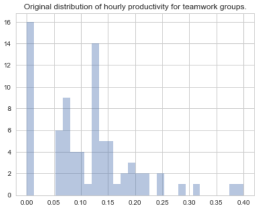
Application to Andrabi, Das, and Khwaja (2017), Cont.

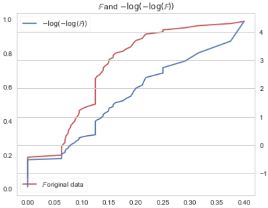
- Rescaled $-\log F$ to fit a PMF & sample of 1,000 iid obs from it (right panel)
- Rescaled $\frac{1}{-\log F}$ to fit a PMF & sample of 1,000 iid obs from it (left panel)
- Test p-value: 0.9 for H_0: transformed density $-\log F$ is logconcave;
  - evidence in favor of F double logconvex
- Test p-value $= 0$ for H_0: transformed density $\frac{1}{-\log F}$ is logconcave



Application to Lyons (2017) AEJ Applied Econ

- Field experiment on
  - impact of teamwork on productivity
- Outcome variable: productivity for groups allowed to work in teams
  - K-S test for homogeneity in distributions returns p-value>0.97
  - Test for logconcavity of original sample: p-value>0
- Left: Distribution of original outcome variable
- Right: Computed empirical F (red) and $-\log(-\log F)$ (blue)





Application to Lyons (2017) AEJ Applied Econ – cont.

- Rescaled $-\log F$ to fit a PMF & sample of 1,000 iid obs from it (right panel)
- Rescaled $\frac{1}{-\log F}$ to fit a PMF & sample of 1,000 iid obs from it (left panel)
- Test p-value $= 0.99$ for H_0: transformed density $\frac{1}{-\log F}$ is logconcave
- Test p-value: 0 for H_0: transformed density $-\log F$ is logconcave;
  - evidence in favor of F double logconcavity