Strategic Sample Selection

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

Workshop on Scientific Rationality & Strategic Interaction
Bocconi, November 8, 2019

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

Information Value of Sample Selection

Inference is often based on selected data

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

We often have control on extent of selection

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

We often have control on extent of selection:

- Bid solicitation

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

We often have control on extent of selection:

- Bid solicitation
- Test format design

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

We often have control on extent of selection:

- Bid solicitation
- Test format design
- Standards for research, etc.

Information Value of Sample Selection

Inference is often based on selected data:

- Winning bids in auctions
- Students' answers under examinee choice
- Researcher's data selection, etc.

We often have control on extent of selection:

- Bid solicitation
- Test format design
- Standards for research, etc.

So we want to understand:

How does sample selection affect quality of inference?

Environment: IDO Preferences

An evaluator faces a decision problem under uncertainty:

- States: $\Theta \subseteq \mathbb{R}$ (either finite or a possibly unbounded interval)
- Actions: $a_1 < \cdots < a_L$ (extended to continuous A)

Environment: IDO Preferences

An evaluator faces a decision problem under uncertainty:

- States: $\Theta \subseteq \mathbb{R}$ (either finite or a possibly unbounded interval)
- Actions: $a_1 < \cdots < a_L$ (extended to continuous A)
- Preferences: IDO, as in Quah-Strulovici (2009):

  for all $\theta' > \theta$ and $a'' > a'$,

  $$u(\theta, a'') \geq (>)\ u(\theta, a') \quad \Longrightarrow \quad u(\theta', a'') \geq (>)\ u(\theta', a')$$

  whenever $u(\theta, a'') \geq u(\theta, a)$ for all a with $a' \leq a \leq a''$

  This class of preferences includes

  - Milgrom-Shannon (1994) single-crossing preferences &
  - Karlin-Rubin (1956) monotone preferences

Environment: MLR Experiments

- Before acting, evaluator observes realization of MLR experiment:

$$X = (X_1, \ldots, X_n) \sim g(\cdot|\theta) \quad \text{such that}$$

$$x' \geqq x \quad \Rightarrow \quad g(x'|\theta)/g(x|\theta) \quad \text{increasing in } \theta$$

Environment: MLR Experiments

- Before acting, evaluator observes realization of MLR experiment:

$$X = (X_1, \ldots, X_n) \sim g(\cdot|\theta) \quad \text{such that}$$

$$x' \geqq x \quad \Rightarrow \quad g(x'|\theta)/g(x|\theta) \quad \text{increasing in } \theta$$

- IDO + MLR $\Rightarrow$ optimal action increasing in x

Environment: MLR Experiments

- Before acting, evaluator observes realization of MLR experiment:

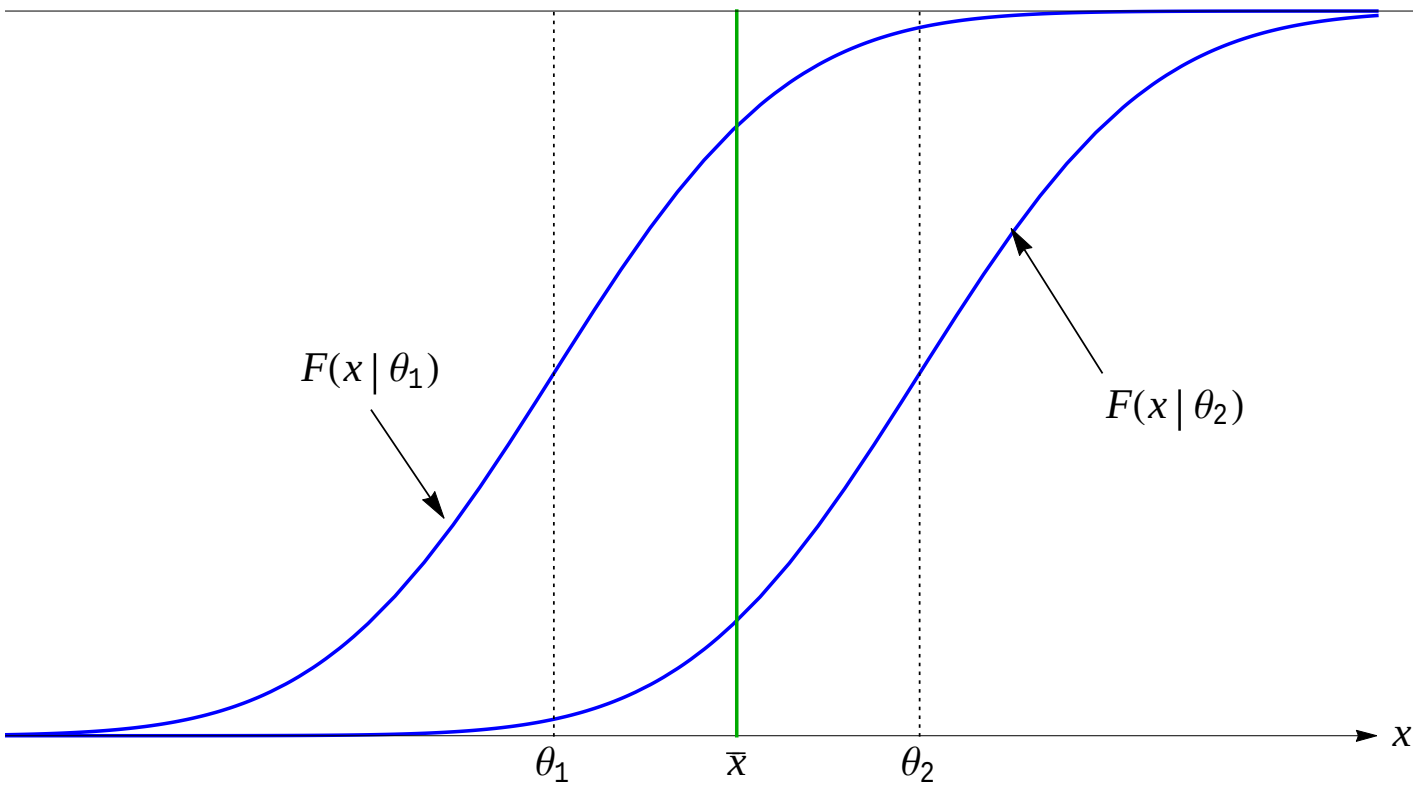
$$X = (X_1, \ldots, X_n) \sim g(\cdot|\theta) \quad \text{such that}$$

$$x' \geqq x \quad \Rightarrow \quad g(x'|\theta)/g(x|\theta) \quad \text{increasing in } \theta$$

- IDO + MLR $\Rightarrow$ optimal action increasing in x
- Evaluator's welfare: $U(X)$

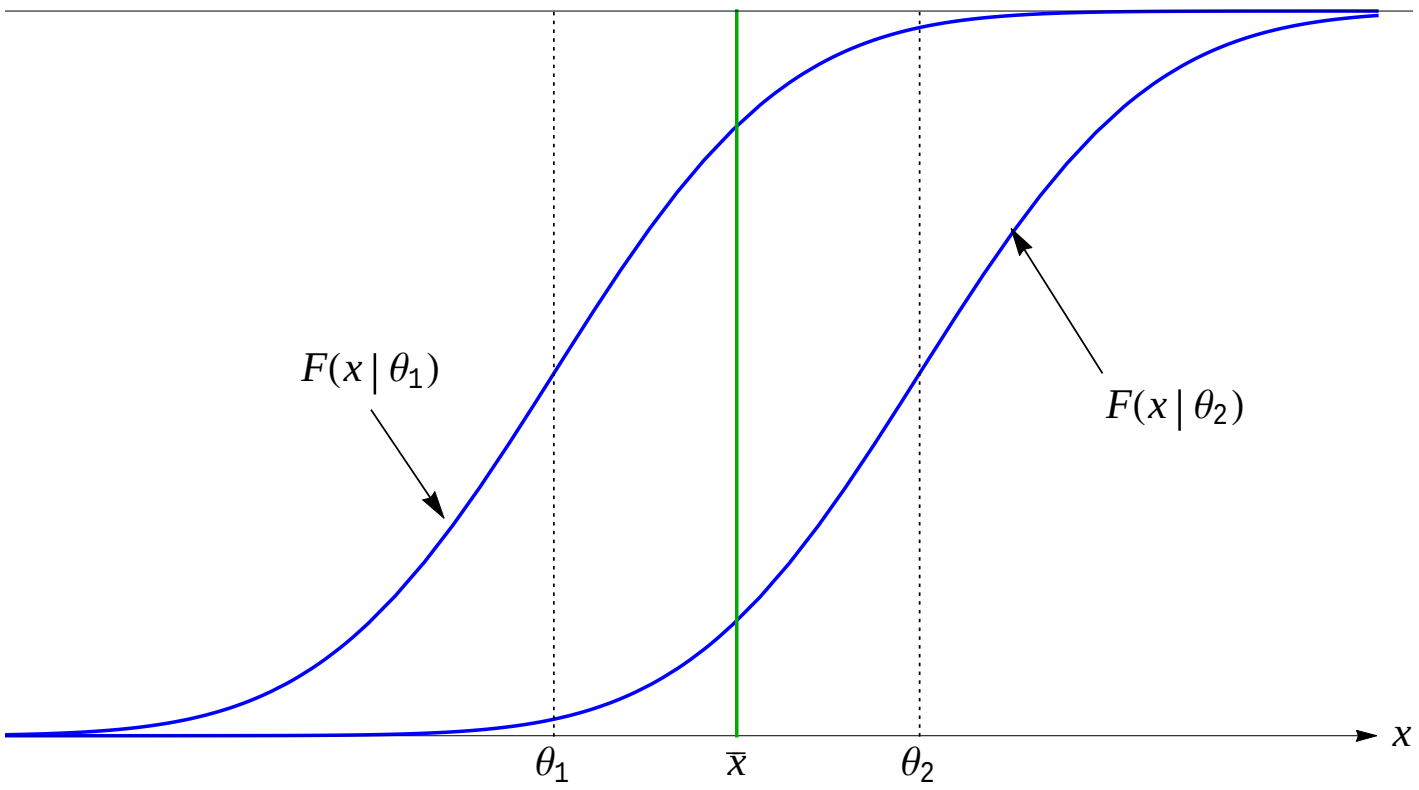
Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$

Experiment: $X = \theta + \varepsilon$ with $n = 1$ and $\varepsilon \sim$ Normal



Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$

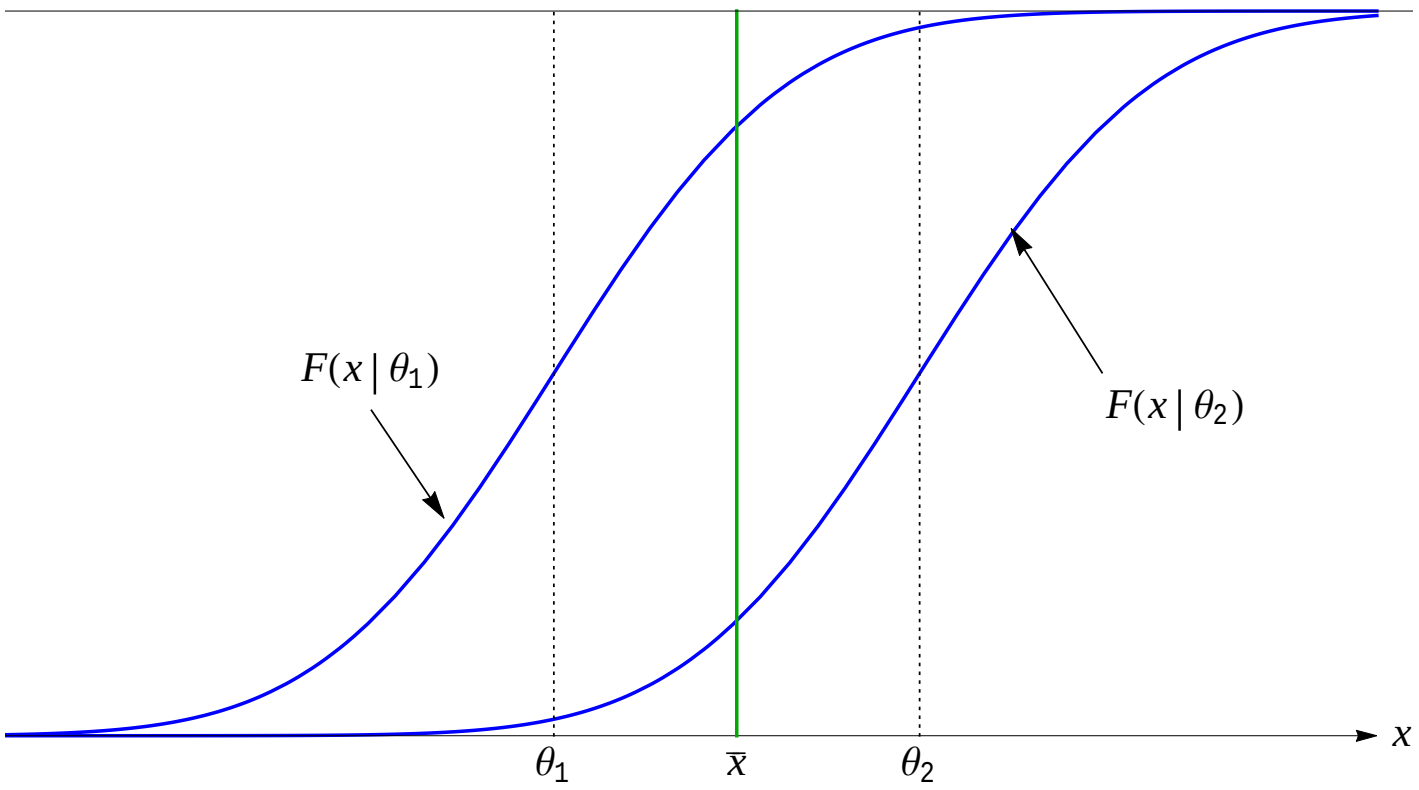
Experiment: $X = \theta + \varepsilon$ with $n = 1$ and $\varepsilon \sim$ Normal



Evaluator chooses a_1 if $x < \bar{x}$ and a_2 if $x \geq \bar{x}$.

Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$

Experiment: $X = \theta + \varepsilon$ with $n = 1$ and $\varepsilon \sim$ Normal



Evaluator chooses a_1 if $x < \bar{x}$ and a_2 if $x \geq \bar{x}$. FP and FN.

Main Question: Comparing Selected Experiments

Family of distributions: $F(\cdot|\theta)$ for $\theta \in \Theta$

Selected experiment: observe n highest of $k \geq n$ iid$|\theta$ draws

$$g(x|\theta) = \frac{k!}{(k-n)!} F^{k-n}(x_n|\theta) f(x_1|\theta) \cdots f(x_n|\theta)$$

(for $x_1 \geq \cdots \geq x_n$)

Main Question: Comparing Selected Experiments

Family of distributions: $F(\cdot|\theta)$ for $\theta \in \Theta$

Selected experiment: observe n highest of $k \geq n$ iid$|\theta$ draws

$$g(x|\theta) = \frac{k!}{(k-n)!} F^{k-n}(x_n|\theta) f(x_1|\theta) \cdots f(x_n|\theta)$$

(for $x_1 \geq \cdots \geq x_n$)

n sample size, $k \geq n$ presample size (no selection: $k = n$)

More selection: $k \uparrow$

Main Question: Comparing Selected Experiments

Family of distributions: $F(\cdot|\theta)$ for $\theta \in \Theta$

Selected experiment: observe n highest of $k \geq n$ iid$|\theta$ draws

$$g(x|\theta) = \frac{k!}{(k-n)!} F^{k-n}(x_n|\theta) f(x_1|\theta) \cdots f(x_n|\theta)$$

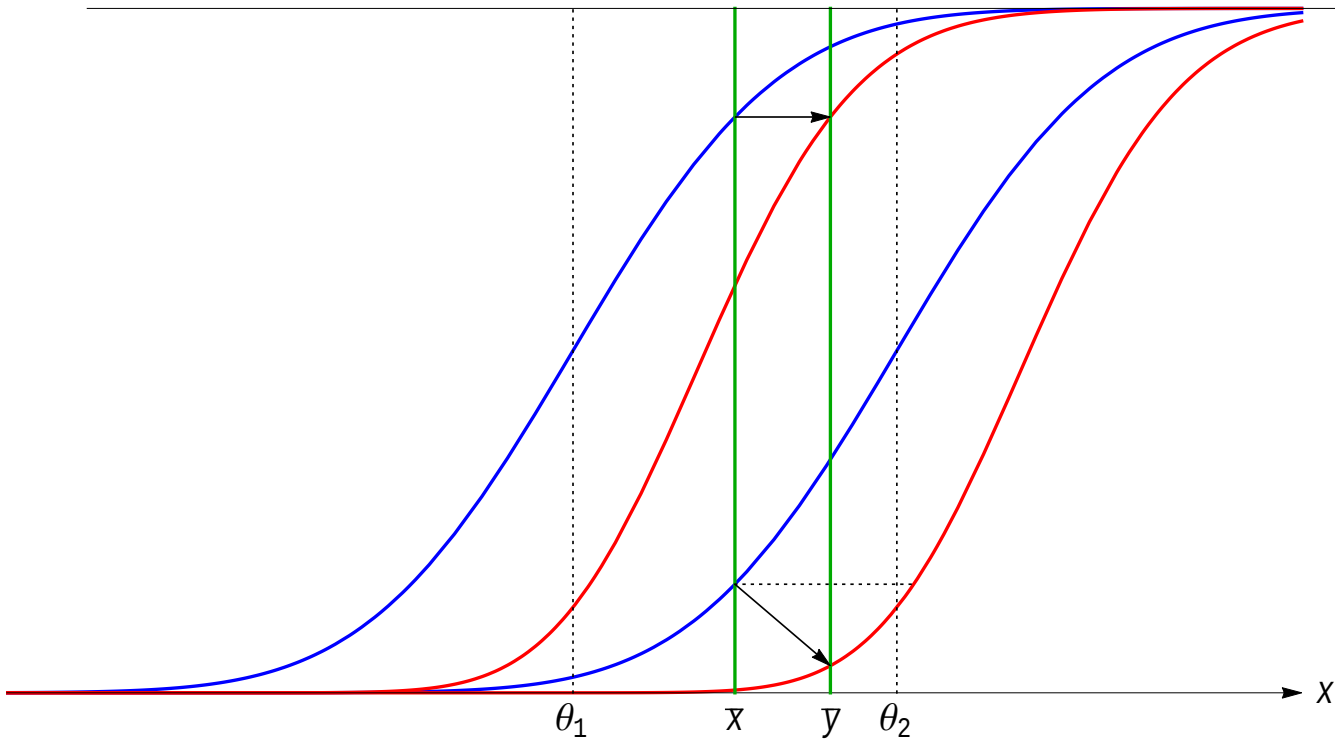
(for $x_1 \geq \cdots \geq x_n$)

n sample size, $k \geq n$ presample size (no selection: $k = n$)

More selection: $k \uparrow$

When does more selection benefit/hurt evaluator?

Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$
Normal noise, $k = n = 1$ vs $k > n = 1$



Selection benefits: as selection increases, adjusting cutoff from $\bar{x}$ to $\bar{y}$ matches FP; matching FN would require larger adjustment.

Accuracy and Welfare

To answer question, we generalize Lehmann's notion of accuracy.

- Consider family of experiments $X(t)$ parametrized by $t \in [0,1]$
- For every θ and all s, t in $[0,1]$, define quantile transformation $x \mapsto \varphi_{s,t}(x|\theta) = (z_1, \ldots, z_n)$, where $z_1, \ldots, z_n$ are defined by

$$G_1(t, z_1|\theta)) = G_1(s, x_1|\theta) \quad \ldots$$
$$G_i(t, z_i|\theta, z_{<i}) = G_i(s, x_i|\theta, x_{<i}) \quad \ldots$$

- $X(t)$ ordered by accuracy if for all x and $t > s$,
  $\varphi_{s,t}(x|\theta)$ is increasing in θ.
  Intuition: matching FN requires larger cutoff adjustment.

THEOREM 0
If $X(t)$ is ordered by accuracy, then $U(X(t))$ increases in t.

Welfare Impact of Selection: Location, $n = 1$

THEOREM 1

With sample size $n = 1$ *from location experiment* $F(x|\theta) = F(x - \theta)$, *higher presample size* k *increases (decreases) payoff in all IDO problems iff reverse hazard function* $-\log F$ *is logconcave (logconvex).*

Welfare Impact of Selection: Location, $n = 1$

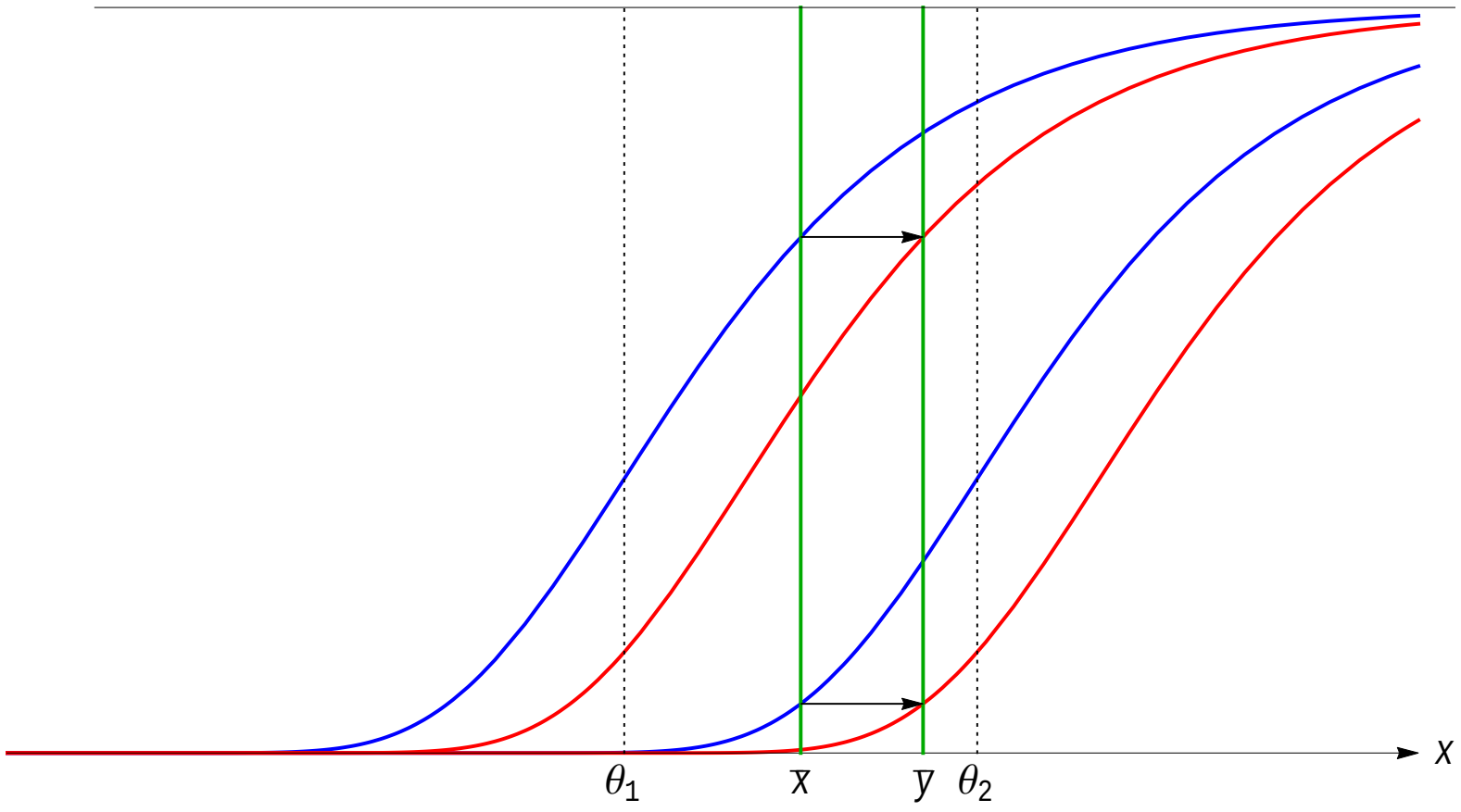
THEOREM 1

With sample size $n = 1$ from location experiment $F(x|\theta) = F(x - \theta)$, higher presample size k increases (decreases) payoff in all IDO problems iff reverse hazard function $-\log F$ is logconcave (logconvex).

- Normal distribution has logconcave $-\log F$
- Gumbel extreme value distribution has loglinear $-\log F$
- Exponential distribution has logconvex $-\log F$

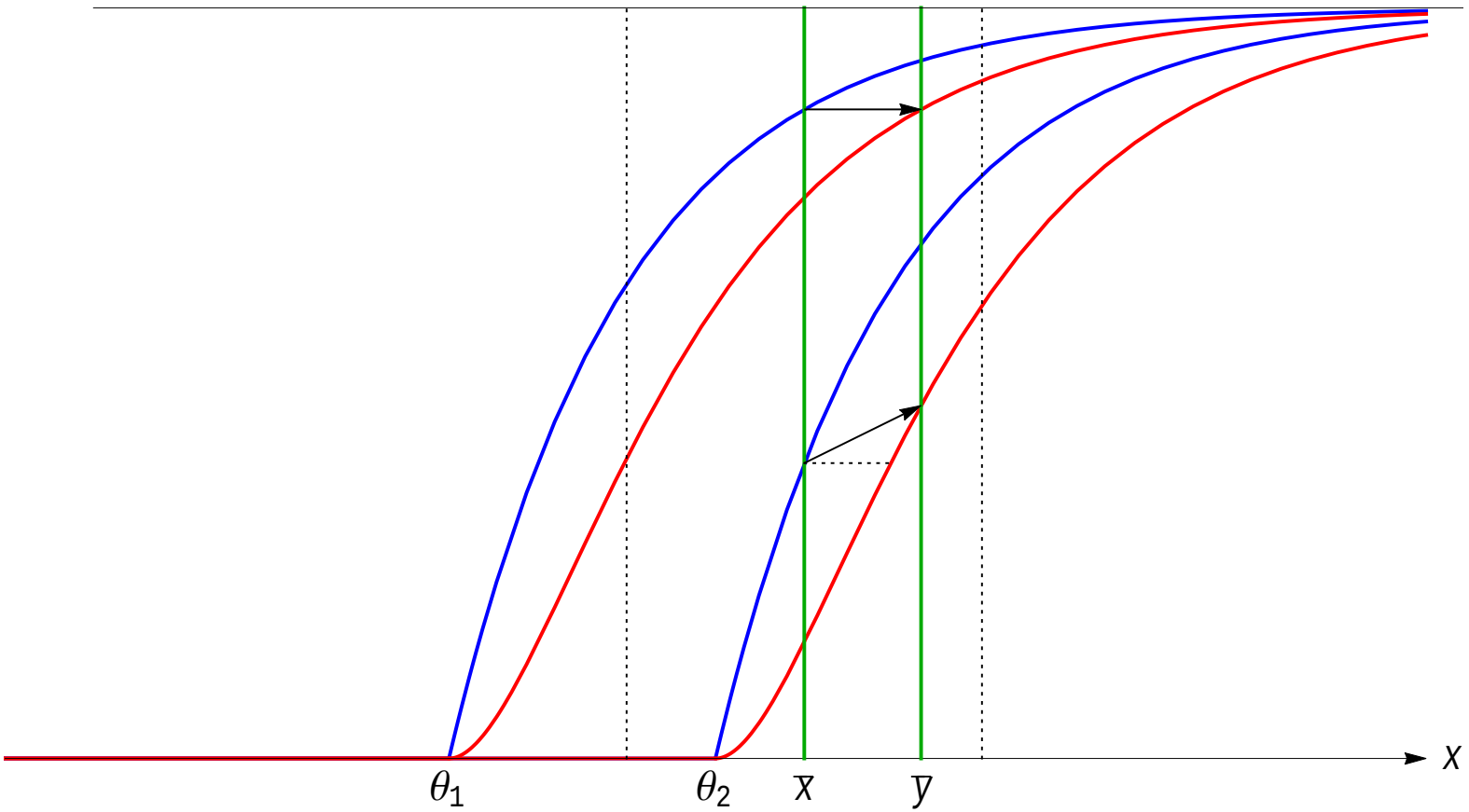
Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$

Gumbel noise, $k = n = 1$ vs $k > n = 1$



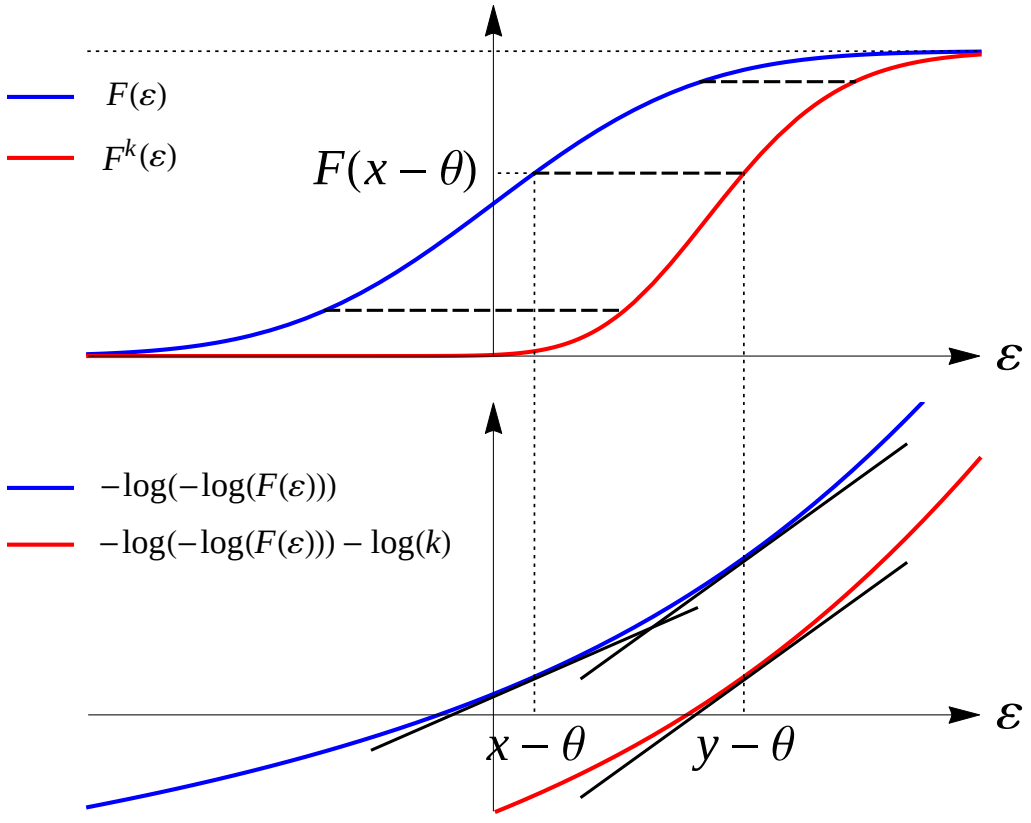
Example: actions $a_1 < a_2$ and states $\theta_1 < \theta_2$

Exponential noise, $k = n = 1$ vs $k > n = 1$



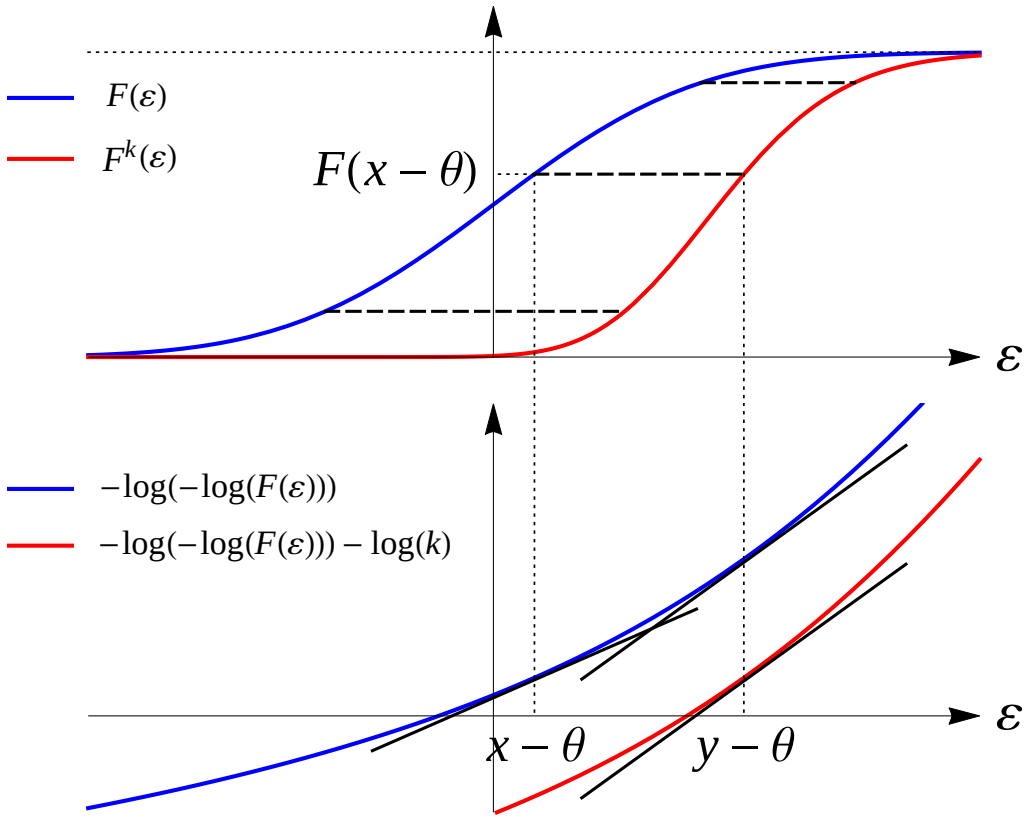
Proof of Thm 1: Dispersion and Selection

More selection benefits when it reduces noise dispersion, or F^k steeper than F at corresponding quantiles:



Proof of Thm 1: Dispersion and Selection

More selection benefits when it reduces noise dispersion, or F^k steeper than F at corresponding quantiles:



Transform both F and F^k by $u \mapsto -\log(-\log u)$,
we see more selection benefits/hurts if $-\log F$ is logcav (logvex).

Impact of Selection: General Selected Experiments

THEOREM 2

For fixed sample size $n \geq 1$, higher pre-sample size k increases (decreases) payoff in all IDO problems if reverse hazard rate $f(x|\theta)/F(x|\theta)$ is log-supermodular (log-submodular, with support unbounded above).

That is, if for all $\theta' > \theta$

$$\frac{f(x|\theta')/F(x|\theta')}{f(x|\theta)/F(x|\theta)} \quad \textit{increasing (decreasing) in } x$$

Proof of Thm 2: Log-spm of RHR and FN gain

Selected experiment $X(t)$ with t parametrizing extent of selection:
($t = 0$ less selection, $t = 1$ more selection)

Recall $x \mapsto \varphi_{s,t}(x|\theta) = z$.

$X(t)$ ordered by accuracy if for $\theta' > \theta$ we have

$$\frac{G_1(t, z_1|\theta'))}{G_1(s, x_1|\theta')} \leq 1 \quad \ldots \quad \frac{G_i(t, z_i|\theta', z_{<i})}{G_i(s, x_i|\theta', x_{<i})} \leq 1 \quad \ldots$$

FN gain in limit: ratios are ≤ 1 as each $x_i \to \infty$.

Log-spm of RHR ensures ratios are increasing:
FN gain relatively even larger for $x_i < \infty$

Method of proof applicable to any family of experiments.

Extreme Selection

What happens when presample size $k \to \infty$?

Extreme Selection

What happens when presample size $k \to \infty$?
Focus on location experiments, hold sample size $n = 1$ fixed.

Extreme Selection

What happens when presample size $k \to \infty$?
Focus on location experiments, hold sample size $n = 1$ fixed.

Fundamental Theorem of Extreme Value Theory.
Take noise distribution F. Suppose that, for some nondegenerate distribution $\bar{F}$ and normalization sequences b_k and $a_k > 0$,

$$\Pr\left[\frac{\max\{\varepsilon_1, \ldots, \varepsilon_k\} - b_k}{a_k}\right] \equiv F^k\left(b_k + a_k\varepsilon\right) \to \bar{F}\left(\varepsilon\right) \quad \text{as } k \to \infty$$

for every continuity point ε of $\bar{F}$. Then $\bar{F}$ is Gumbel, Extreme Weibull or Frechet.

(For logconcave f, $\bar{F}$ is always either Gumbel or Weibull.)

Extreme Selection

Selection shifts noise distribution upward: location normalization sequence b_k grows. But location (where noise is centered) is irrelevant in location experiments.

Limit impact hinges on scale normalization sequence a_k.

Does a_k shrink to zero as k increases?

Extreme Selection

Selection shifts noise distribution upward: location normalization sequence b_k grows. But location (where noise is centered) is irrelevant in location experiments.

Limit impact hinges on scale normalization sequence a_k.

Does a_k shrink to zero as k increases?

THEOREM 3

With sample size $n = 1$ from location experiment $F(x|\theta) = F(x-\theta)$, as presample size $k \to \infty$ welfare converges to full information payoff if and only if at least one of the following holds: (i) the support of F is bounded above; (ii) hazard rate $f/(1-F)$ is unbounded.

If neither holds, limit welfare is welfare from unidimensional experiment with Gumbel noise $\bar{F}(\cdot/\alpha)$, where $\alpha = \lim_{\varepsilon\to\infty}[1-F(\varepsilon)]/f(\varepsilon)$.

Application 1: Information Aggregation in Auctions

Consider equilibrium in Milgrom-Weber general symmetric model of discriminatory auction:

- n objects, $k > n$ symmetric bidders
- one-dimensional state θ
- bidder i's signal: $X_i = \theta + \varepsilon_i$
- bidder i's actual value for an object: $v_i(\theta, x_1, \ldots, x_k)$
- equilibrium strategy $b(\cdot)$ strictly increasing

Application 1: Information Aggregation in Auctions

Consider equilibrium in Milgrom-Weber general symmetric model of discriminatory auction:

- n objects, $k > n$ symmetric bidders
- one-dimensional state θ
- bidder i's signal: $X_i = \theta + \varepsilon_i$
- bidder i's actual value for an object: $v_i(\theta, x_1, \ldots, x_k)$
- equilibrium strategy $b(\cdot)$ strictly increasing

Now consider external observer (evaluator) who sees winning bids. These have same information value as highest signals!

$\big(b(X_{k:k}), \ldots, b(X_{k-n+1:k})\big)$ as accurate as $\big(X_{k:k}, \ldots, X_{k-n+1:k}\big)$

Application 1: Information Aggregation in Auctions

PROPOSITION 1

1.A. If reverse hazard rate f/F is logconcave, information increases monotonically with competition k. If support of f is bounded or hazard rate $f/(1-F)$ is unbounded, information is full for $k \to \infty$ (e.g. normal). If hazard rate $f/(1-F)$ is bounded, information is not full for $k \to \infty$ (e.g. logistic).

1.B. If reverse hazard rate f/F is logconvex (with hazard rate necessarily bounded), information decreases monotonically with competition k (e.g. exponential).

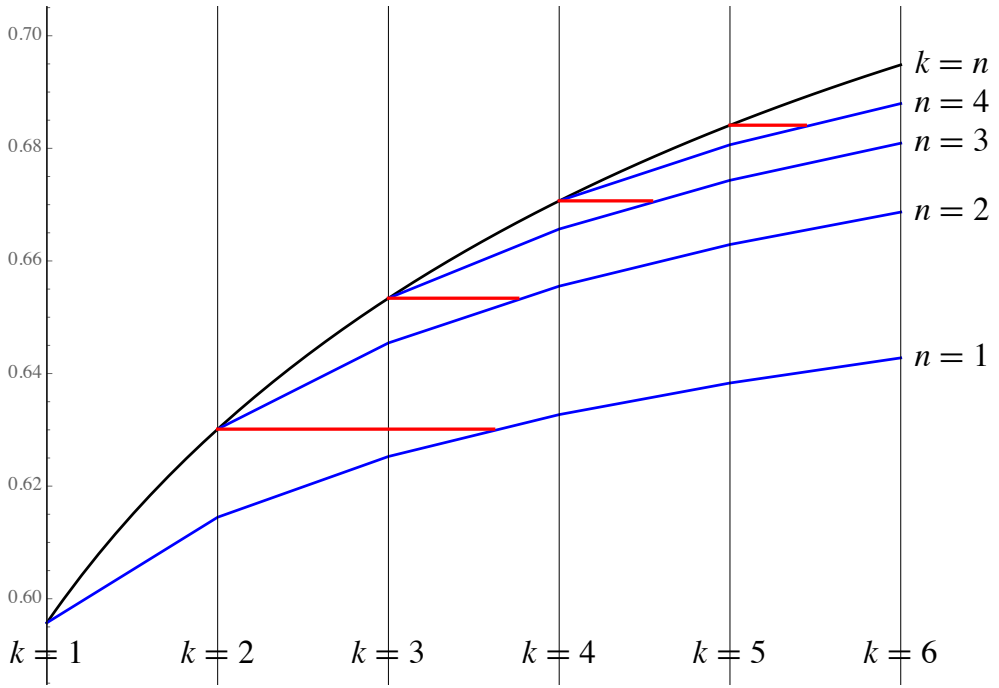
Using variations of Thm 3, more limit results can be given for
- objects/bidders ratio n/k fixed as $k \to \infty$, or
- both n and $k-n$ growing without bound
- uniform-price auctions

Application 2: Examinee Choice

- Evaluator tests examinee's ability θ through n questions.
- Examinee's answer in any given question: $\theta + \varepsilon$, where $\varepsilon \sim F$.
- Should evaluator allow examinee to choose n out $k > n$ questions?
- Evaluator bears question preparation cost c_P, grading cost c_S
  - Total cost: $kc_P + nc_S$
- Optimal test format (n, k)?

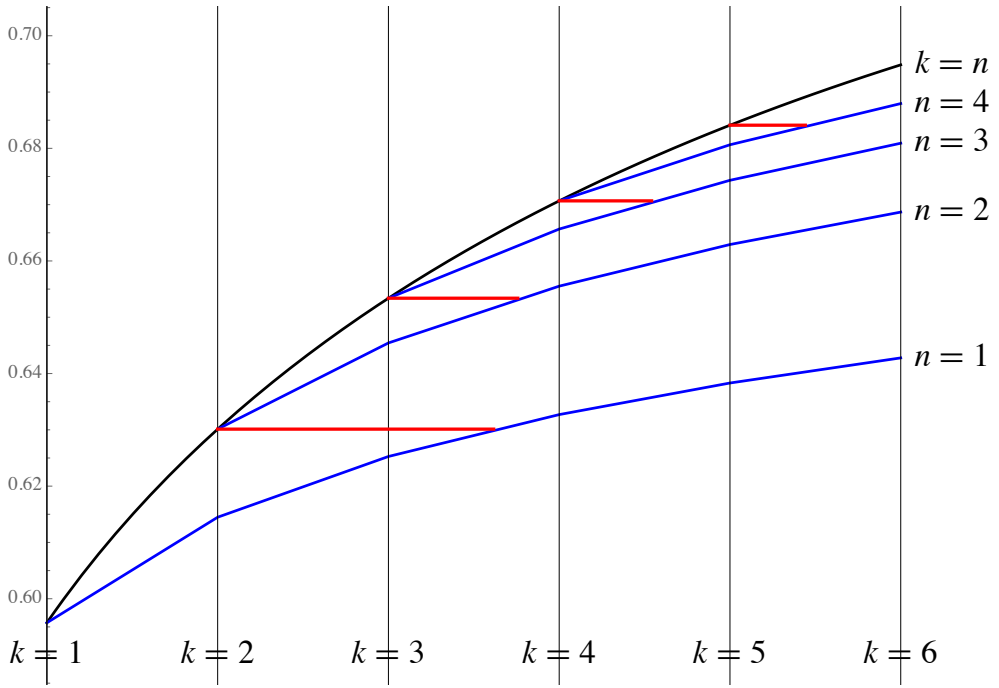
Application 2: Examinee Choice

- Under log-cav f/F, value increases in n and k: two goods
- For indifference, sample size $n \downarrow$ requires presample size $k \uparrow$. With normal F, additional presample needed <u>decreases</u> in n.



Application 2: Examinee Choice

- Under log-cav f/F, value increases in n and k: two goods
- For indifference, sample size $n \downarrow$ requires presample size $k \uparrow$. With normal F, additional presample needed decreases in n.



- Optimal (n, k) depends on relative price c_P/c_S
- But, with M students, $k = n$ never optimal if $2c_P < Mc_S$!

Optimal Test Design with Small Presampling Costs

- Evaluator exploits selection to economize on sample size
- By Thm 3, sample selection is always beneficial for small c_P

PROPOSITION 2 (OPTIMAL PRESAMPLING)
In a location experiment design problem with optimal presampling and noise distribution having either (i) support bounded above or (ii) unbounded hazard rate, for every $c_S > 0$ there exists $\bar{c}_P > 0$ such that if $c_P \leq \bar{c}_P$ then optimal experiment format (k, n) is such that $k > n$.

Application 3: Researcher Bias

Game of experimental design:

1. Evaluator sets sample size n & allows or not sample selection
2. Sender privately chooses $k \geq n$ or opts out

Application 3: Researcher Bias

Game of experimental design:

1. Evaluator sets sample size n & allows or not sample selection
2. Sender privately chooses $k \geq n$ or opts out

Sender bears presample cost kc_P.
Sample cost nc_S split between sender and evaluator.

Application 3: Researcher Bias

Game of experimental design:

1. Evaluator sets sample size n & allows or not sample selection
2. Sender privately chooses $k \geq n$ or opts out

Sender bears presample cost kc_P.
Sample cost nc_S split between sender and evaluator.

PROPOSITION 3 (STRATEGIC PRESAMPLING)
In a location experiment design problem with strategic presampling and noise distribution with either (i) support bounded above or (ii) unbounded hazard rate, for every $c_S > 0$ there exists $\bar{c}_P > 0$ such that if $c_P \leq \bar{c}_P$ then it is optimal for evaluator to allow sample selection.

Researcher Bias: RCT setup

- Potential outcomes of individual i:

$$X_{1,i} := \varepsilon_i \qquad \text{(control)}$$
$$X_{2,i} := \varepsilon_i + \theta \qquad \text{(treatment)}$$

- Outcome distribution of control assumed known: $\varepsilon_i \sim F$
- Treatment effect θ of new drug constant but unknown
- Evaluator would like to approve a_2 if $\theta \geq \bar{\theta}$ & reject a_1 otherwise

Researcher Bias: RCT setup

Researcher runs RCT treating $i_1, \ldots, i_n$ & untreating $i_{n+1}, \ldots, i_{2n}$.

Evaluator observes:

Treatment Group	Control Group
$X_{2,i_1} = \varepsilon_{i_1} + \theta$	$X_{1,i_{n+1}} = \varepsilon_{i_{n+1}}$
$\vdots$	$\vdots$
$X_{2,i_n} = \varepsilon_{i_n} + \theta$	$X_{1,i_{2n}} = \varepsilon_{i_{2n}}$

Compare two scenarios:

Random: Researcher selects $2n$ individuals at random, and randomly assigns n to treatment and n to control.

Strategic: Researcher has information on baseline outcome for $k > 2n$ individuals, and on this basis
(i) selects $2n$ individuals for RCT and
(ii) assigns n to treatment and n to control

Random Selection and Assignment to Treatment

- Random scenario
  - distribution F of baseline outcome is known
  - control group adds no information
- Experiment boils down to observation of treatment group only, i.e.

$$\begin{aligned} &\varepsilon_{i_1} + \theta \\ &\vdots \\ &\varepsilon_{i_n} + \theta \end{aligned}$$

where $\varepsilon_{i_1}, \ldots, \varepsilon_{i_n}$ are iid draws from F

Strategic Selection and Assignment to Treatment

- Out of k pre-sampled individuals, Researcher assigns
  - n with highest X_1 to treatment and
  - n with lowest X_1 to control
- Let $(i{:}k)$ denote individual with ith highest X_1
- Evaluator then observes

$$\begin{matrix} \varepsilon_{(k:k)} + \theta & \varepsilon_{(n:k)} \\ \vdots & \vdots \\ \varepsilon_{(k-n+1:k)} + \theta & \varepsilon_{(1:k)} \end{matrix}$$

- Control group can only add information (and here it does, because order statistics are correlated). Thus experiment gives at least as much information as vector on LHS
- By Theorem 1, if f/F is log-concave
  - strategic selection and assignment benefit evaluator!