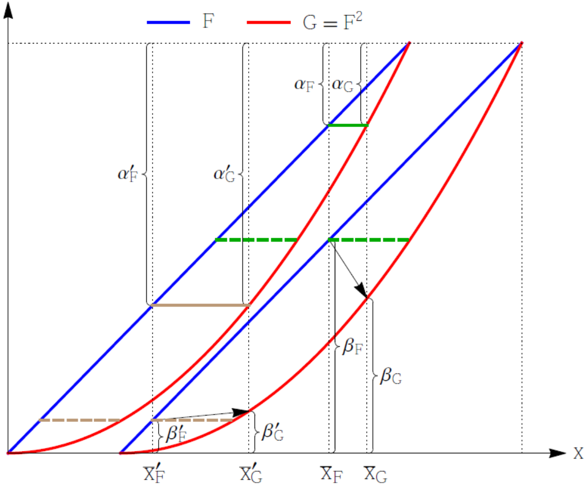
Discussion of "Strategic Sample Selection"

Di Tillio, Ottaviani, and Sørensen

Montreal, October 2017

Strategic Sample Selection



The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.
- The outcome is $x(\theta, \varepsilon)$. Is the evaluator better off with a *random experiment* or a *selected experiment*? Good question!

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.
- The outcome is $x(\theta, \varepsilon)$. Is the evaluator better off with a *random experiment* or a *selected experiment*? Good question!
- You can think of the evaluator as minimizing a weighed average of Type I and Type II errors.

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.
- The outcome is $x(\theta, \varepsilon)$. Is the evaluator better off with a *random experiment* or a *selected experiment*? Good question!
- You can think of the evaluator as minimizing a weighed average of Type I and Type II errors.
- The experiment is a *location experiment*: $x(\theta, \varepsilon) = \theta + \varepsilon$.

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.
- The outcome is $x(\theta, \varepsilon)$. Is the evaluator better off with a *random experiment* or a *selected experiment*? Good question!
- You can think of the evaluator as minimizing a weighed average of Type I and Type II errors.
- The experiment is a *location experiment*: $x(\theta, \varepsilon) = \theta + \varepsilon$.
  - The term "location experiment" is used only in the abstract.

The setting

- An evaluator makes a decision (accept/reject) based on an *experiment* conducted by a researcher.
- Two possible states, θ_H and θ_L. Noise ε, with CDF F.
- The outcome is $x(\theta, \varepsilon)$. Is the evaluator better off with a *random experiment* or a *selected experiment*? Good question!
- You can think of the evaluator as minimizing a weighed average of Type I and Type II errors.
- The experiment is a *location experiment*: $x(\theta, \varepsilon) = \theta + \varepsilon$.
  - The term "location experiment" is used only in the abstract.
  - It is left to the reader to decide how important or realistic this assumption is.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0,1]$.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0, 1]$.

- With a location experiment $(x = \theta + \varepsilon)$ there is no "global" ranking between the random and the selected experiment; the evaluator's preferences matter.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0, 1]$.
- With a location experiment ($x = \theta + \varepsilon$) there is no "global" ranking between the random and the selected experiment; the evaluator's preferences matter.
- But with a scale experiment ($x = \theta\varepsilon$), it can be shown that the selected experiment is globally better than the random experiment.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0,1]$.
- With a location experiment ($x = \theta + \varepsilon$) there is no "global" ranking between the random and the selected experiment; the evaluator's preferences matter.
- But with a scale experiment ($x = \theta\varepsilon$), it can be shown that the selected experiment is globally better than the random experiment.
- Conclusion: The nature of the experiment matters.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0,1]$.
- With a location experiment ($x = \theta + \varepsilon$) there is no "global" ranking between the random and the selected experiment; the evaluator's preferences matter.
- But with a scale experiment ($x = \theta\varepsilon$), it can be shown that the selected experiment is globally better than the random experiment.
- Conclusion: The nature of the experiment matters.
  - If the empirical relevance of $-\log(-\log F)$ is to be a selling point, then it is important to argue that location experiments are realistic.

Different types of experiment

- Example: Assume ε is uniformly distributed on $[0,1]$.
- With a location experiment $(x = \theta + \varepsilon)$ there is no "global" ranking between the random and the selected experiment; the evaluator's preferences matter.
- But with a scale experiment $(x = \theta\varepsilon)$, it can be shown that the selected experiment is globally better than the random experiment.
- Conclusion: The nature of the experiment matters.
  - If the empirical relevance of $-\log(-\log F)$ is to be a selling point, then it is important to argue that location experiments are realistic.
- NB: In the scale experiment, $\log x = \log\theta + \log\varepsilon$, where the distribution of $\log\varepsilon$ satisfies $-\log(-\log F)$ convex.

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

- Location experiments are "special":

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

- Location experiments are "special":

    - The distribution of x in state θ is $H(x|\theta) = F(x - \theta)$.

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

- Location experiments are "special":

    - The distribution of x in state θ is $H(x|\theta) = F(x - \theta)$.

    - Knife-edge: $H(x|\theta)$ is "equally disperse" regardless of θ *if and only if* the experiment is a location experiment (the CDF shifts horizontally).

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

- Location experiments are "special":
  - The distribution of x in state θ is $H(x|\theta) = F(x - \theta)$.
  - Knife-edge: $H(x|\theta)$ is "equally disperse" regardless of θ *if and only if* the experiment is a location experiment (the CDF shifts horizontally).
  - In other words dispersion is, conveniently, "held constant" across states. Note that $-\log(-\log H(x|\theta))$ has the same curvature for all θ.

Location experiments and dispersion

- A main insight of the paper: To rank the random and the selected experiments, it is important to compare the *dispersion* of the corresponding distribution functions, F and $G = F^k$.

- Location experiments are "special":

  - The distribution of x in state θ is $H(x|\theta) = F(x - \theta)$.

  - Knife-edge: $H(x|\theta)$ is "equally disperse" regardless of θ *if and only if* the experiment is a location experiment (the CDF shifts horizontally).

  - In other words dispersion is, conveniently, "held constant" across states. Note that $-\log(-\log H(x|\theta))$ has the same curvature for all θ.

  - Thus, we only have to worry about relative dispersion of F and G.

Location experiments and dispersion

- Proposition 2: The selected experiment is globally preferred to the random experiment if and only if $-\log(-\log F)$ is convex.

Location experiments and dispersion

- Proposition 2: The selected experiment is globally preferred to the random experiment if and only if $-\log(-\log F)$ is convex.
- Can a weaker version be obtained, e.g.:

Location experiments and dispersion

- Proposition 2: The selected experiment is globally preferred to the random experiment if and only if $-\log(-\log F)$ is convex.

- Can a weaker version be obtained, e.g.:

  - The selected experiment is globally preferred if (i) $-\log(-\log H(x|\theta))$ is convex for all θ and (ii) $H(x|\theta_H)$ is more disperse than $H(x|\theta_L)$??

Location experiments and dispersion

- Proposition 2: The selected experiment is globally preferred to the random experiment if and only if $-\log(-\log F)$ is convex.

- Can a weaker version be obtained, e.g.:

    - The selected experiment is globally preferred if (i) $-\log(-\log H(x|\theta))$ is convex for all θ and (ii) $H(x|\theta_H)$ is more disperse than $H(x|\theta_L)$??

    - Or (equivalently?) if (i) $-\log(-\log F)$ is convex and (ii) the transformation from $x(\theta_L, \varepsilon)$ to $x(\theta_H, \varepsilon)$ satisfies some condition.

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

- Recall that F and G have the same support.

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

- Recall that F and G have the same support.
  - However, two CDFs with the same **bounded** support can **never** be ranked in terms of the dispersive order.

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

- Recall that F and G have the same support.
  - However, two CDFs with the same **bounded** support can **never** be ranked in terms of the dispersive order.
  - Direct implication: There can be no global ranking of the random and selected experiments if the support of ε is bounded.

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.
- Recall that F and G have the same support.
  - However, two CDFs with the same **bounded** support can **never** be ranked in terms of the dispersive order.
  - Direct implication: There can be no global ranking of the random and selected experiments if the support of ε is bounded.
- I assume $-\log(-\log F)$ can be neither convex nor concave when the support is bounded?

Peculiarities of the dispersive order

- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

- Recall that F and G have the same support.
  - However, two CDFs with the same **bounded** support can **never** be ranked in terms of the dispersive order.
  - Direct implication: There can be no global ranking of the random and selected experiments if the support of ε is bounded.

- I assume $-\log(-\log F)$ can be neither convex nor concave when the support is bounded?
  - Hence, the relevance of $-\log(-\log F)$ convex relies on (i) location experiments and (ii) unbounded support.

Peculiarities of the dispersive order

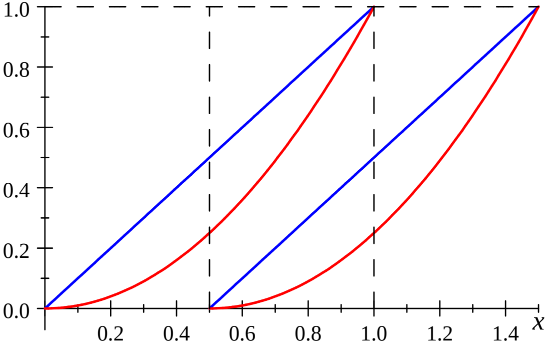
- It matters greatly if the support is bounded or not. An explicit discussion of this fact would be welcome.

- Recall that F and G have the same support.
  - However, two CDFs with the same **bounded** support can **never** be ranked in terms of the dispersive order.
  - Direct implication: There can be no global ranking of the random and selected experiments if the support of ε is bounded.

- I assume $-\log(-\log F)$ can be neither convex nor concave when the support is bounded?
  - Hence, the relevance of $-\log(-\log F)$ convex relies on (i) location experiments and (ii) unbounded support.
  - Explicitly state and justify/discuss both assumptions.

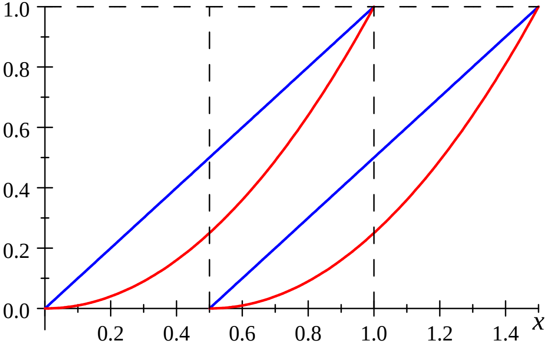
Peculiarities of the dispersive order

- Another way to think about this: With bounded support, it is possible to *eliminate* Type I or Type II errors; corner solutions.



Peculiarities of the dispersive order

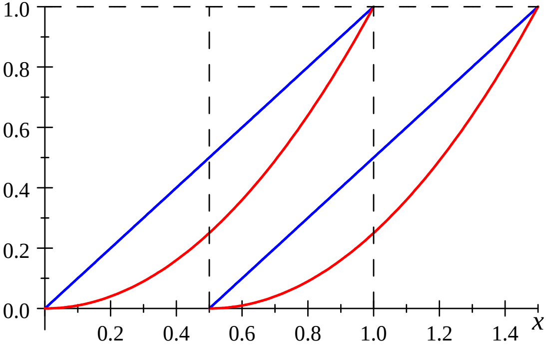
- Another way to think about this: With bounded support, it is possible to *eliminate* Type I or Type II errors; corner solutions.



- $x = 0.5$: Zero Type II errors ($\beta = 0$). By FOSD blue has fewer Type I errors than red.

Peculiarities of the dispersive order

- Another way to think about this: With bounded support, it is possible to *eliminate* Type I or Type II errors; corner solutions.



- $x = 0.5$: Zero Type II errors ($\beta = 0$). By FOSD blue has fewer Type I errors than red.
- $x = 1.5$: Zero Type 1 errors ($\alpha = 0$). By FOSD, red has fewer Type II errors than blue.

Peculiarities of the dispersive order

- Corner solutions are perhaps implicitly ignored at times?

Peculiarities of the dispersive order

- Corner solutions are perhaps implicitly ignored at times?
  - When moving from random to selected experiments, with F symmetric: "by symmetry the loss associated to the increase in false positives (higher α) is exactly offset by the benefit associated to the reduction in false negatives (lower β)."

Peculiarities of the dispersive order

- Corner solutions are perhaps implicitly ignored at times?
  - When moving from random to selected experiments, with F symmetric: "by symmetry the loss associated to the increase in false positives (higher α) is exactly offset by the benefit associated to the reduction in false negatives (lower β)."
  - This assumes an interior solution because α, β are the same for the random and the selected experiments at the corners.

Peculiarities of the dispersive order

- Corner solutions are perhaps implicitly ignored at times?
  - When moving from random to selected experiments, with F symmetric: "by symmetry the loss associated to the increase in false positives (higher α) is exactly offset by the benefit associated to the reduction in false negatives (lower β)."
  - This assumes an interior solution because α, β are the same for the random and the selected experiments at the corners.
- Conclusion: The theory is "neater" when ε is unbounded both below and above (ensures α, β interior).