Comments on: Cho, Lim, and Jang

John Londregan

Princeton

Asian Polmeth 2018

January 11, 2018

Attendance vs Crowd Size.

Attendance vs Crowd Size.

- The police

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers
  - Use attendance to demonstrate support for an agenda

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers
  - Use attendance to demonstrate support for an agenda
  - Care about the total number mobilized

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers
  - Use attendance to demonstrate support for an agenda
  - Care about the total number mobilized
- The police take the official count.

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers
  - Use attendance to demonstrate support for an agenda
  - Care about the total number mobilized
- The police take the official count.
- Organizers can't be everywhere.

Attendance vs Crowd Size.

- The police
  - Allocate resources for crowd control.
  - Care about peak crowd size.
- Protest organizers
  - Use attendance to demonstrate support for an agenda
  - Care about the total number mobilized
- The police take the official count.
- Organizers can't be everywhere.
- How to repurpose the police estimate?

Estimating the undetected population.

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
    - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
    - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\hat{N} \approx \frac{n_0^2}{n_1}$

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\hat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\widehat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by
  - the inverse probability of recapture: $\frac{n_0}{n_1}$

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\widehat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by
  - the inverse probability of recapture: $\frac{n_0}{n_1}$
- Assumptions:

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\widehat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by
  - the inverse probability of recapture: $\frac{n_0}{n_1}$
- Assumptions:
  - all birds are equally dumb

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\widehat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by
  - the inverse probability of recapture: $\frac{n_0}{n_1}$
- Assumptions:
  - all birds are equally dumb
  - captured birds grow neither more nor less fond of bait

Estimating the undetected population.

- What Rumsfeld might call “known unknowns"
- Estimating undiscovered petroleum reserves.
- Capture-Recapture methods.
  - $Pr\{Caught\ at\ t_0|Exist\} = Pr\{Recaptured\ at\ t_1|Caught\ at\ t_0\}$
  - $\frac{n_0}{N} \approx \frac{n_1}{n_0}$
  - $\widehat{N} \approx \frac{n_0^2}{n_1}$
  - Each of n_0 recaptured birds increases the expectation of N by
  - the inverse probability of recapture: $\frac{n_0}{n_1}$
- Assumptions:
  - all birds are equally dumb
  - captured birds grow neither more nor less fond of bait
  - :

An Inverse Probability Weight Estimator

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees
  - The time the count is taken: t_0

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees
  - The time the count is taken: t_0
  - $p_0 = Pr\{t_0 \in (T_{i,1}, T_{i,2})\}$

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees
  - The time the count is taken: t_0
  - $p_0 = Pr\{t_0 \in (T_{i,1}, T_{i,2})\}$
  - Crowd Size at time t_0: S_0

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees
  - The time the count is taken: t_0
  - $p_0 = Pr\{t_0 \in (T_{i,1}, T_{i,2})\}$
  - Crowd Size at time t_0: S_0
- CLJ advocate $\widehat{N} = \dfrac{S_0}{\widehat{p}_0}$

An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
- Ingredients:
  - The window during which i attends: $(T_{i,1}, T_{i,2})$
  - This is estimated based on a survey of a subset of attendees
  - The time the count is taken: t_0
  - $p_0 = Pr\{t_0 \in (T_{i,1}, T_{i,2})\}$
  - Crowd Size at time t_0: S_0
- CLJ advocate $\hat{N} = \dfrac{S_0}{\hat{p}_0}$
- They then formulate estimates of p_0

Estimating p_0

Estimating p_0

- Parametric version: attendee start times T_1 are normal.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.
- In either case, let $y_1 = T_1$, $y_2 = log(T_2 - T_1)$.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.
- In either case, let $y_1 = T_1$, $y_2 = log(T_2 - T_1)$.
- The fraction of the population present at a fixed time $\bar{t}$ is:

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.
- In either case, let $y_1 = T_1$, $y_2 = log(T_2 - T_1)$.
- The fraction of the population present at a fixed time $\bar{t}$ is:
- $$f_0 = \int_{\infty}^{\bar{t}} \int_{log(\bar{t}-y_1)}^{\infty} f(y_1, y_2) dy2dy1$$

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.
- In either case, let $y_1 = T_1$, $y_2 = log(T_2 - T_1)$.
- The fraction of the population present at a fixed time $\bar{t}$ is:
- $$f_0 = \int_{\infty}^{\bar{t}} \int_{log(\bar{t}-y_1)}^{\infty} f(y_1, y_2) dy2dy1$$
- They calculate this using Monte Carlo methods.

Estimating p_0

- Parametric version: attendee start times T_1 are normal.
- the duration of attendance, $T_2 - T_1$ is log normal.
- use an attendee sample to estimate the joint distribution.
- Nonparametric version: use a kernel estimator based on the same attendee sample.
- In either case, let $y_1 = T_1$, $y_2 = log(T_2 - T_1)$.
- The fraction of the population present at a fixed time $\bar{t}$ is:
- $$f_0 = \int_{\infty}^{\bar{t}} \int_{log(\bar{t}-y_1)}^{\infty} f(y_1, y_2) dy2dy1$$
- They calculate this using Monte Carlo methods.
- Then they bootstrap to calibrate the precision of their $\hat{p}(0)$.

Trouble with the Police!

Trouble with the Police!

- Problem: the police report the modal crowd size...

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...or at least they try to.

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...or at least they pick a moment to measure the crowd based on the density of attendance.

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...so the time of peak attendance t_0 is a random variable dependent on (y_1, y_2)!

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...so the time of peak attendance t_0 is a random variable dependent on (y_1, y_2)!
- Formally it is straightforward to amend their framework to:

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...so the time of peak attendance t_0 is a random variable dependent on (y_1, y_2)!
- Formally it is straightforward to amend their framework to:
- $$p_0 = \max_t \int_{\infty}^{t} \int_{log(t-y_1)}^{\infty} f(y_1, y_2) dy2dy1$$

Trouble with the Police!

- Problem: the police report the modal crowd size...
- ...so the time of peak attendance t_0 is a random variable dependent on (y_1, y_2)!
- Formally it is straightforward to amend their framework to:
- $$p_0 = \max_t \int_{\infty}^{t} \int_{log(t-y_1)}^{\infty} f(y_1, y_2) dy2dy1$$
- To do: derive sampling properties for $\hat{N}$

The Bootstrap

The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.

The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.
- Amundsen made full use of his dogs

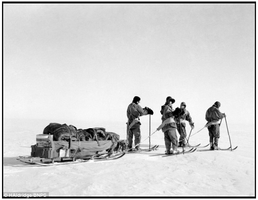
The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.
- Amundsen made full use of his dogs
- 

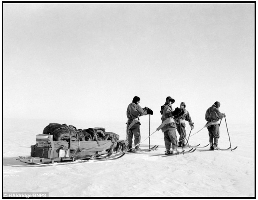
The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.
- Amundsen made full use of his dogs
- 
- Scott didn't

The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.
- Amundsen made full use of his dogs
  - 
- Scott didn't
  - 

The Bootstrap

- The non-parametric bootstrap makes full use of Efron's insight.
- Amundsen made full use of his dogs
  - 
- Scott didn't
  - 
- Moral: Make full use of your options.

The Bootstrap

The Bootstrap

- So use the non-parametric bootstrap!

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample
- repeat

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample
- repeat
- this gives us the variance of $\hat{p}$ directly

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample
- repeat
- this gives us the variance of $\hat{p}$ directly
- as an extra we get an estimate of **bias**

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample
- repeat
- this gives us the variance of $\hat{p}$ directly
- as an extra we get an estimate of **bias**
- measured as the gap between $p(0)$ and the mean of our bootstrap repliquees

The Bootstrap

- So use the non-parametric bootstrap!
- Draw $\{(T_{1i}, T_{2i})\}_{i=1}^{n}$ from the sample population
- calculate the peak attendance fraction for the drawn pseudosample
- repeat
- this gives us the variance of $\hat{p}$ directly
- as an extra we get an estimate of **bias**
- measured as the gap between $p(0)$ and the mean of our bootstrap repliquees
- a useful (if not always welcome) reality check.

Conclusion

Conclusion

- CLJ identify an important aspect of the crowd measurement problem

Conclusion

- CLJ identify an important aspect of the crowd measurement problem
- The offer an interesting fix

Conclusion

- CLJ identify an important aspect of the crowd measurement problem
- The offer an interesting fix
- **Nice project!**