Comments on: Cho, Lim, and Jang

John Londregan
Princeton
Asian Polmeth 2018
January 11, 2018
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?
Attendance vs Crowd Size.

- The police

- 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?
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.

- 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?
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.

What Rumsfeld might call "known unknowns" are undiscovered petroleum reserves.

Capture-Recapture methods.

\[\text{Pr}\left\{\text{Caught at } t_0 | \text{Exist}\right\} = \text{Pr}\left\{\text{Recaptured at } t_1 | \text{Caught at } t_0\right\}\]

\[n_0 N \approx n_1\]

Each of \(n_0\) recaptured birds increases the expectation of \(N\) by the inverse probability of recapture:

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 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.
  - \( \Pr\{\text{Caught at } t_0 | \text{Exist}\} = \Pr\{\text{Recaptured at } t_1 | \text{Caught at } t_0\} \)

Assumptions:
- All birds are equally dumb
- Captured birds grow neither more nor less fond of bait

Londregan

Comments: Candlelight

Asian Polmeth
Estimating the undetected population.

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

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.
  
  $$Pr\{Caught \ at \ t_0|Exist\} = Pr\{Recaptured \ at \ t_1|Caught \ at \ t_0\}$$

  $$\frac{n_0}{\hat{N}} \approx \frac{n_1}{n_0}$$

  $$\hat{N} \approx \frac{n_0^2}{n_1}$$

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.
  - \( Pr\{\text{Caught at } t_0|\text{Exist}\} = Pr\{\text{Recaptured at } t_1|\text{Caught at } t_0\} \)
  - \( \frac{n_0}{\hat{N}} \approx \frac{n_1}{n_0} \)
  - \( \hat{N} \approx \frac{n_0^2}{n_1} \)
  - 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.

\[ \Pr\{\text{Caught at } t_0|\text{Exist}\} = \Pr\{\text{Recaptured at } t_1|\text{Caught at } t_0\} \]
\[ \frac{n_0}{\hat{N}} \approx \frac{n_1}{n_0} \]
\[ \hat{N} \approx \frac{n_0^2}{n_1} \]
- Each of \( n_0 \) recaptured birds increases the expectation of \( N \) by
- 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.
  - \( \Pr\{\text{Caught at } t_0 | \text{Exist}\} = \Pr\{\text{Recaptured at } t_1 | \text{Caught at } t_0\} \)
  - \( \frac{n_0}{\hat{N}} \approx \frac{n_1}{n_0} \)
  - \( \hat{N} \approx \frac{n_0^2}{n_1} \)
  - Each of \( n_0 \) recaptured birds increases the expectation of \( N \) by \( \frac{n_0}{n_1} \)

- Assumptions:
Estimating the undetected population.

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

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 \( ^\wedge N = S_0 ^p_0 \)

They then formulate estimates of \( p_0 \)
An Inverse Probability Weight Estimator

- Cho, Lim, and Jang (hereafter CLJ) exploit and generalize the insight of recapture methods.
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 $\hat{N} = S_0 p_0$
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 $^\hat{N} = S_0^p_0$ They then formulate estimates of $p_0$
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

$P_0 = \Pr\{t_0 \in (T_{i,1}, T_{i,2})\}$

CLJ advocate $\hat{N} = S_0^P$
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 \( t_0 \):

\[ S_0 = S_0^p = \ldots \]
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: \( t_0 \)
- \( p_0 = \Pr\{t_0 \in (T_{i,1}, T_{i,2})\} \)
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 \( \hat{N} = \frac{S_0}{\hat{p}_0} \)
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 \( \hat{N} = \frac{S_0}{\hat{p}_0} \)

They then formulate estimates of \( p_0 \)
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:

$$f_0 = \frac{\bar{t}}{\int \int \log(\bar{t} - y_1) f(y_1, y_2) \, dy_2 \, dy_1}$$

They calculate this using Monte Carlo methods. Then they bootstrap to calibrate the precision of their $\hat{p}_0$. 

Comments: Candlelight
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:

$$f_0 = \int_{\infty}^{\bar{t}} \int_{\infty}^{\infty} f(y_1, y_2) dy_2 dy_1$$

$$= \int_{\infty}^{\bar{t}} \int_{\infty}^{\infty} f(y_1, y_2) dy_2 dy_1$$
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:

$$f_0 = \int_{\infty}^{\bar{t}} \int_{\infty}^{\infty} f(y_1, y_2) dy_2 dy_1$$

- 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:

\[
    f_0 = \int_{\infty}^{\bar{t}} \int_{\infty}^{\infty} f(y_1, y_2) \, dy_2 \, dy_1
    \log(\bar{t} - y_1)
\]

- They calculate this using Monte Carlo methods.
- Then they bootstrap to calibrate the precision of their $\hat{p}(0)$. 
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...
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.
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:

$$p_0 = \max_t \int \int \log(t - y_1) f(y_1, y_2) \, dy_2 \, dy_1$$
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:

$$p_0 = \max_t \int_t^\infty \int_\infty^{\log(t-y_1)} f(y_1, y_2) \, dy_2 \, dy_1$$
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:

$$p_0 = \max_t \int_t^\infty \int_\infty^{\log(t-y_1)} f(y_1, y_2) \, dy_2 \, dy_1$$

- To do: derive sampling properties for $\hat{N}$
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.

Londregan

Comments: Candlelight

Asian Polmeth 9 / 11
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

So use the non-parametric bootstrap!

\[
\{ (T_1^i, T_2^i) \}_{i=1}^n
\]

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.
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

\[ n_i = 1 \]

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 replicates. A useful (if not always welcome) reality check.

Londregan

Comments: Candlelight

Asian Polmeth 10 / 11
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
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 $p$ directly as an extra we get an estimate of bias measured as the gap between $p(0)$ and the mean of our bootstrap replicates, a useful (if not always welcome) reality check.
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

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

Nice project!
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!