Digital Echoes:  
Understanding Patterns of Mass Violence  
with Data and Statistics

Patrick Ball  
Director of Research, HRDAG  
November, 2023
Post-truth?
That’s nothing new for human rights activists

People and institutions that commit mass violence nearly always lie about it. The lies are often grotesque and easily disproven.

Human rights campaigns are successful because we’re persistent, we speak with the moral authority of the victims, and because we’re committed to the truth.

In human rights, we speak truth to power — so it better be true.

Statistics should be a footnote -- but the statistics must be right.
In human rights data collection, we (usually) don’t know what we don’t know.
In human rights data collection, we (usually) don’t know what we don’t know. Which means that we don’t know if what we don’t know is systematically different from what we do know.
Imagine that you collect and combine three databases:

Which picture is accurate?
What is it that we don’t know?

Sendero Luminoso?

Peruvían Army?

The relationship between what is observed (the sample) and what is true (the population) can be very complicated. Only a mathematical, probability-based model can bridge the gap.
An introduction to MSE

\[ p(A) = \frac{A}{N} \]

\[ p(B) = \frac{B}{N} \]

\[ p(M) = \frac{M}{N} = \frac{A}{N} \cdot \frac{B}{N} \]

\[ MN = AB \]

\[ \hat{N} = \frac{AB}{M} \]

There are decades of development on this method: see Lum et al. (2013), *The American Statistician*, 67(4):191-200.
Police homicides in the US

- US Bureau of Justice Statistics used MSE to estimate police homicides.

- Two datasets: ARD (by the Bureau of Justice Statistics) and the SHR (by the FBI)

- Covering 2003-2009 and 2011 (what happened to 2010?)

- Estimate assumes that the part in the middle is independent of the parts on the sides: not true.
Police homicides in the US

- The BJS estimate was flawed by assuming list independence.
- What list dependence do we see in other countries?
Police homicides in the US

- If list dependence in the US is like list dependence in other countries, what is the total estimated number of police homicides?

- (Colombia is the context most like the US in terms of the lists available)

- The key question in the US is magnitude.
Police killings in the Philippines (2019)

- Project with faculty and students at the Stabile Center for Investigative Journalism at Columbia University
- Data from eight sources (two sources were themselves integrated from more underlying sources)
- Estimates for Manila, Quezon City, and Caloocan covering July 2016 through December 2017
- Main outcome was long-form article in *The Atlantic* magazine; widely discussed in the Philippines, including among legislators
- Nonetheless, violence by police remains popular
Data from Manila (police only)
Data from Manila (police only)

- density graph is 95% credible interval
- capture heterogeneity is apparent in graph showing estimates with omitted sources
- note: bias tends to be downward
Overall results

Table 2: Documented, Reported, and Estimated, Killings in Three Cities

<table>
<thead>
<tr>
<th>Region</th>
<th>Perpetrator</th>
<th>Documented</th>
<th>Reported by Police</th>
<th>Estimated</th>
<th>Est/Police</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manila</td>
<td>police</td>
<td>488</td>
<td>313</td>
<td>575</td>
<td>1.8</td>
</tr>
<tr>
<td>Manila</td>
<td>unidentified</td>
<td>245</td>
<td>158</td>
<td>414</td>
<td>2.6</td>
</tr>
<tr>
<td>Quezon City</td>
<td>police</td>
<td>337</td>
<td>245</td>
<td>348</td>
<td>1.4</td>
</tr>
<tr>
<td>Quezon City</td>
<td>unidentified</td>
<td>357</td>
<td>24</td>
<td>428</td>
<td>17.8</td>
</tr>
<tr>
<td>Caloocan</td>
<td>police</td>
<td>286</td>
<td>206</td>
<td>331</td>
<td>1.6</td>
</tr>
<tr>
<td>Caloocan</td>
<td>unidentified</td>
<td>599</td>
<td>19</td>
<td>745</td>
<td>39.2</td>
</tr>
</tbody>
</table>

- Murders of low-level drug users and sellers were encouraged by commanders and the government.
- violence by "unidentified perpetrators" tends to be unreported by police, sometimes entirely.
- "unidentified" may just mean violence by police that officers chose not to report to supervisors: with the delegation of violence, there is always a principal-agent problem!
Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

- [https://hrdag.org/2022/09/20/colombia-100-databases/](https://hrdag.org/2022/09/20/colombia-100-databases/)

- Disponible en la página web de la CEV. Licencia CC BY-SA 4.0.
Nuestros insumos iniciales

Resulta en 125 archivos, 26M de registros
(no todos son víctimas)
Sobre el proyecto

3 años, 25 programadores, 11542 commits en git

<table>
<thead>
<tr>
<th>Language</th>
<th>archivos</th>
<th>comentario</th>
<th>código</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAML</td>
<td>198</td>
<td>244</td>
<td>1347880</td>
</tr>
<tr>
<td>R</td>
<td>1695</td>
<td>23480</td>
<td>148694</td>
</tr>
<tr>
<td>make</td>
<td>1013</td>
<td>7950</td>
<td>29816</td>
</tr>
<tr>
<td>Rmd</td>
<td>85</td>
<td>8138</td>
<td>4489</td>
</tr>
<tr>
<td>TeX</td>
<td>32</td>
<td>14</td>
<td>4019</td>
</tr>
<tr>
<td>XML</td>
<td>2</td>
<td>0</td>
<td>3394</td>
</tr>
<tr>
<td>Python</td>
<td>29</td>
<td>435</td>
<td>2501</td>
</tr>
<tr>
<td>Jupyter</td>
<td>21</td>
<td>5501</td>
<td>1309</td>
</tr>
<tr>
<td>Julia</td>
<td>7</td>
<td>88</td>
<td>940</td>
</tr>
<tr>
<td>bash</td>
<td>9</td>
<td>89</td>
<td>274</td>
</tr>
<tr>
<td>Markdown</td>
<td>3</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Lua</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Objetivos principales

- Estimar campos faltantes
- Estimar víctimas no observadas
- Propagar la incertidumbre a una única estimación de varianza
Número de víctimas

Observado  Promedio de imputación  Rango de estimación
Más detalles en el reporte

6. Estimación por sistemas múltiples 45
   6.1. Detalles técnicos ................................................................. 45
   6.2. Implementación con datos imputados .................................. 47

7. Sobre las estimaciones 48
   7.1. Sesgo ................................................................................. 49
       7.1.1. Detalles sobre sesgo por sobreestatificación .................. 50
   7.2. Varianza ............................................................................. 51

8. Limitaciones y trabajo futuro 52
   8.1. Datos .................................................................................. 53
       8.1.1. Sobre los datos de la Fiscalía General de la Nación ........ 54
   8.2. Vinculación de registros .................................................... 54
       8.2.1. Campos mínimos ........................................................... 54
       8.2.2. Falsos positivos vs. falsos negativos ......................... 54
       8.2.3. Casos de ambigüedad .................................................. 55
   8.3. Imputación estadística ........................................................ 55
       8.3.1. Bases especializadas vs. bases no especializadas .......... 56
       8.3.2. Extensiones de la imputación ........................................ 56
       8.3.3. Imputación guerrilla sin especificar ............................... 56
       8.3.4. Modelo de clasificación ............................................... 58
   8.4. Estimación por sistemas múltiples ...................................... 59
       8.4.1. Método ......................................................................... 59
       8.4.2. Estratificación ............................................................... 59
       8.4.3. Método para estimar la varianza ................................. 59
   8.5. Trabajo futuro ................................................................. 60
Three ways to have rigorous statistics (and machine learning predictions, btw)

• **A perfect census:** if you have all the possible data, you can do anything you like. This is what “big data” should mean. There are a very small number of projects in human rights in which we have all the data, but even there, the only proof of complete data is modeling.

• **A random sample** of the population (or a probability sample of some kind). Very hard to do, and many challenging technical issues.

• **Posterior modeling** of the sampling process (e.g., capture-recapture, raking). Requires exactly the right data, plus a lot of math and computing capacity.
Questions about human rights statistics

- Do we have all the data? What's missing? Who is excluded, silenced, marginalized? Or by contrast, what groups are overpoliced (i.e., minor offenses treated as serious crimes)?

- If we don't have all the data (and we almost certainly do not), how do we manage the missingness? Can we quantify our uncertainty? (if not, what do we really know?)

- What is the point of a statistic, analysis, or model? Is to to show "how much?" Or are we interested in a pattern? Are all the elements of the pattern equally represented (or estimated) in the analysis, i.e., has the model corrected for -- or amplified and made worse -- selection bias in the data.

Because we must be right.
If we get it right:
The accused were sentenced to 40 years in prison.