Demographic Methods

Monica Alexander

Week 6: Wrap up

Methodological challenges in demography (and social science)

Methodological challenges in demography (and social science)

1. Inference from biased samples
2. Combining different data sources
3. Ethics of estimation and data collection

Inference from biased samples

Inference from biased samples

- Huge amount of data being produced (online, administrative records, unstructured data [text, images, videos])
- Not created for the purposes of social research but could be potentially used in this capacity
- Large data does not mean good data
- How do we use non-probability samples to make population-based inferences?

Inference from biased samples

Example: migration trends from Facebook data

- Can collect data that gives some information about migration in essentially real time
- Potentially useful to study short-term changes in migration where official data are lacking

But

- We know Facebook is not representative
- We know Facebook itself is a population that is changing over time (ageing, switching between platforms, etc)

Inference from biased samples

Example: studying kin and other support networks

- Studying changes in family is an important part of demography
  - fertility
  - care in old age
  - support networks in times of hardship
- Traditional data sources give us some information about immediate kin
  - members in household
  - info about spouse, siblings, children
- But extended kin structures are important to understand

Inference from biased samples

We can use demographic approximation methods to get estimates for extended kin:

1. Knowing fertility and mortality rates gives us a way of calculating the expected number of surviving kin of different types:

e.g. number of surviving daughters to a mother aged a is:

$$\int_{\alpha}^{a} l_{a-x} m_x dx$$

Here, m_x is the female-specific fertility rate. Grand-daughters:

$$\int_{\alpha}^{a} \left[\int_{\alpha}^{a-x} l_y m_y l_{a-x-y} dy \right] m_x dx$$

Estimating kin

Surviving mothers:

$$M_1(a) = \int_\alpha^\beta \frac{l_{x+a}}{l_x} W(x|t-a)dx$$

$W(x|t-a)$ is the age distribution (at time $t-a$) of women who gave birth to a daughter at time $t-a$.

Can keep going to get grandmothers, aunties, cousins, etc.

Issues:

- easiest to do for female population only
- easiest to do assuming stable populations

Estimating kin

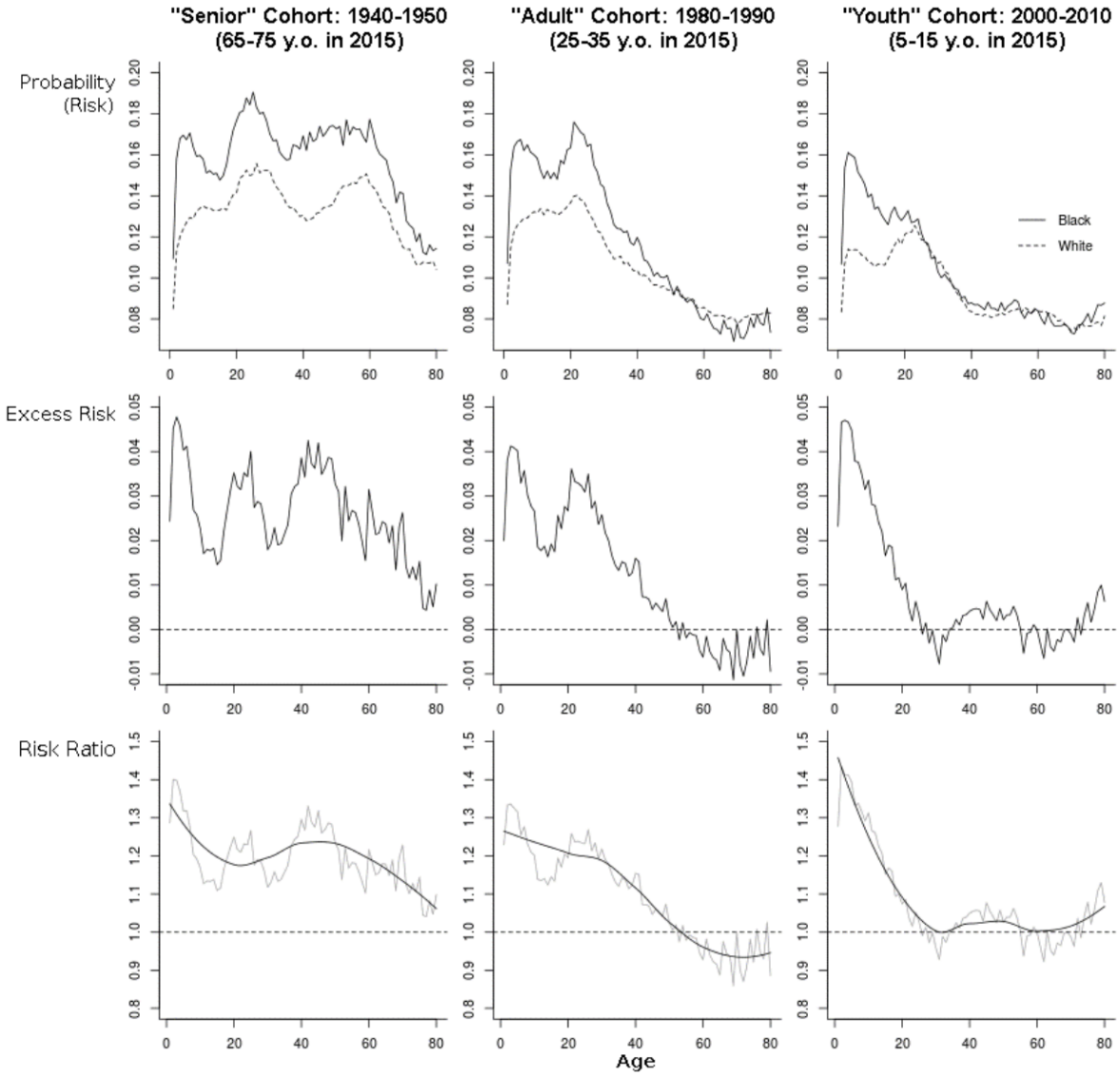
Another approach: simulation

- Know age-specific fertility rates and mortality rates in a population
- Simulate lifelines of a whole bunch of individuals to get estimates of kin structures.

E.g. Chung did this for black and white Americans

Estimating kin

Figure 7. Probability of experiencing the death of select kin (grandchild, child, sibling, parent, and grandparent) by age. The top row shows race-specific probability curves. The middle row shows the absolute difference (black-white) of these probability curves. The bottom row shows the ratio (black/white) of these probability curves: the jagged gray curve represents the raw ratios at each age, and the solid black curve is a LOESS smoother applied to the raw ratios.



Estimating kin

Social networks observed online have potential to complement traditional data sources in studying kin and other social networks.

- Very rich datasets of fine-grained interactions
- But usually only observing one 'type' of interaction
- How to use this data to make inferences about how support networks in general are changing over time?

Combining different data sources

Combining different data sources

- Usually have piecemeal information about different aspects of a demographic process
- How to combine in the one framework, taking into account different sorts of uncertainty and biases

Combining different data sources

Example: mortality estimation

- Have data from different sources (censuses, national surveys, vital registration systems)
- Trying to use all available information to get best estimates

In particular, how to use data from Health and Demographic Surveillance Systems (HDSS)

- detailed, high-quality data on a small population (e.g. town)
- capture all deaths and births for the whole population
- e.g. Agincourt in South Africa, Matlab in Bangladesh
- more viable in short term to set up more HDSS rather than go for a complete vital registration system
- how do we use this data?

Combining different data sources

Example: lifecycle of opioid (/drug) use

- We observe that opioid-related mortality is rising substantially
- There's a whole series of decisions / events that may occur before eventual death
- Important to understand these steps to understand epidemic

1. Starting to use
2. Starting to abuse
3. Non-fatal complications (e.g. non-fatal overdoses, other health complications)
4. Death

Lifecycle of opioid use

1. Starting to use
   - Demand side: legitimate pain, change in life circumstances (e.g. PTSD)
   - Supply side: increase in supply, ease of access to prescriptions, substitution for alternatives
2. Starting to abuse
   - Demand side: increased tolerance, change in life circumstances
   - Supply side: restricted access to legitimate opioids, increase of illicit substances
3. Non-fatal complications
   - Non-fatal overdoses (Naloxone administered at home or hospital or community center)
   - Seeking treatment (for addiction or related issues)
4. Death

What do we have data on?

Lifecycle of opioid use

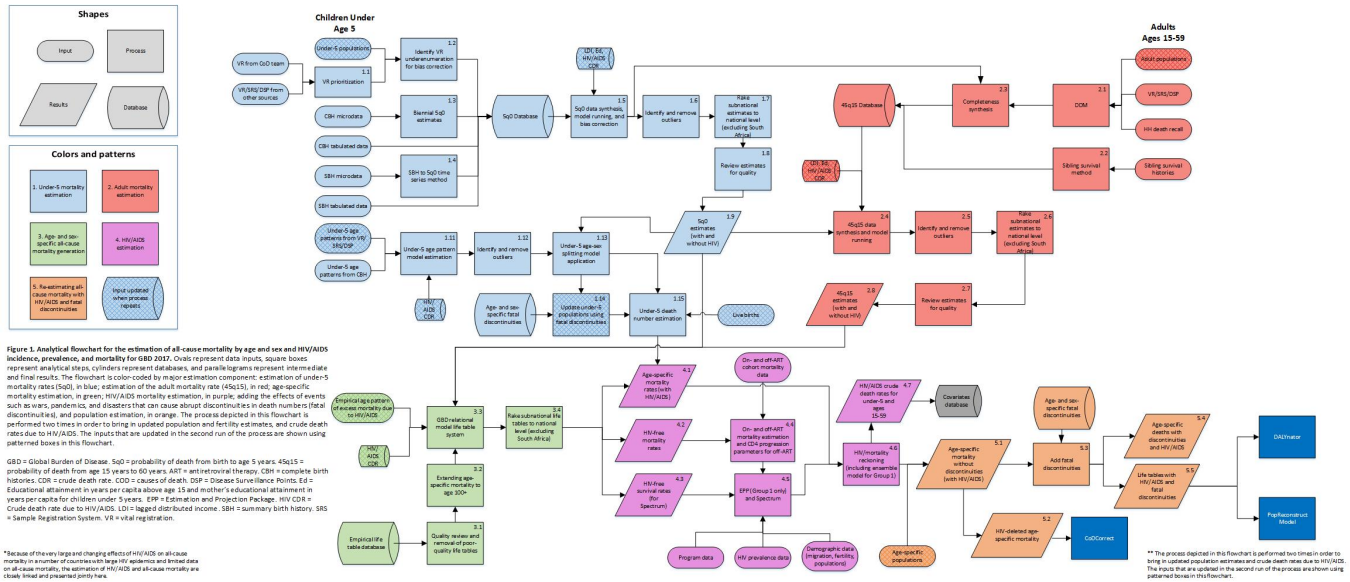
What do we have data on?

- Deaths
- Reasonably good data on prescriptions administered
- Illicit drug supply data are patchy:
  - police seizures
  - small-scale surveys
  - Reddit???
  - Note: not just amount but type and mixtures, especially wrt fentanyl
- Reasonably good data on interactions with healthcare system (in Canada anyway)
  - but missing important piece: non-fatal overdoses outside of healthcare system

Ethics of estimation and data collection

Ethics of estimation

- Modeled estimates versus data
  - we can report uncertainty, but often people don't understand or use it
  - what happens when a country has no data?
- Transparency in modeling process



Ethics of data collection

- online data
  - using data for something other that what it was collected for
  - extreme case: Cambridge Analytica
  - more borderline case: OKCupid paper
- linking of datasets
  - even if identifiers are removed may be identifiable through linkage

Presentations