Introduction to Bayesian Demography

Advanced Demographic Methods Guest Lecture
April 21 2025

Monica Alexander, University of Toronto

Roadmap

Materials: https://mjalexander.github.io/bayesian-demography-lecture/

Lecture:

- What is Bayesian probability
- A brief history
- But what does this mean for my regression
- This seems unnecessary why would I do this

Lab:

- How do I fit these models
- I’m drowning in samples please help
- Okay but also how do I plot
- Examples for mortality and migration (may not cover everything)

Bayesian probability

Different probability interpretations

- Just like there are different types or interpretations of reasoning (inductive, deductive), there are different interpretations of probability
- Two major interpretations: **Frequentist** and **Bayesian**

Frequentist probability

- The probability of an event is defined as the long run relative frequency of that event in (infinitely many) trials
- If we had infinite time and were able to repeat the process that we're interested in over and over, we would calculate the probability as the number of times the event happened divided by the total number of trials
- E.g. the probability of heads on a coin, the probability of scoring 20 in crib, the probability of a toddler drinking a cup of water without spilling any
- Frequentist probability is devoid of opinion, the interpretation is literally just counting events

Frequentist probability

- We are interested in some parameter, call it θ. E.g. θ is the probability of a head, the probability of a hand of 20, the probability of no water spilled
- In frequentist probability, we treat θ as fixed i.e. there is only one true value of θ, and we want to estimate it
- We collect data y (tossing a coin, playing crib, giving the kid water). Then to estimate θ we are fundamentally interested in

$$P(y \mid \theta)$$

i.e. given a certain value of our parameter of interest, how likely is it that we observed that data?

Bayesian probability

- Here, probability is interpreted as a state of knowledge about the world, where this could be a reasonable expectation, or a personal belief
- Before seeing any data, we can have an opinion about the probability of an event happening
- We can then update that belief based on any data we do see
- Probability is a measure of strength of belief

Bayesian probability

- We are still interested in parameter θ as before
- But now θ is not treated as fixed, but random, with its own set of likely values based on our knowledge / the data that we see
- In the Bayesian framework we can say something about θ even before seeing any data. We can then update our beliefs about θ once we've seen data
- To estimate θ we are interested in

$$P(\theta \mid y)$$

i.e. conditional on seeing a set of observations (data), what is the probability of θ being certain values?

Bayes rule

$$P(\theta \mid y)$$

- This is call the **posterior probability** distribution
- Before we collect any data, our beliefs about θ are encoded in the **prior probability** $P(\theta)$
- Information gained from data collection is encoded in the **likelihood** $P(y \mid \theta)$
- Information from the prior and likelihood are combined to obtain the posterior through Bayes' rule:

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

A simple example

- Breast cancer screening (using mammograms) in Germany. Imagine we know
    - The probability an asymptomatic woman has breast cancer is 0.8%.
    - If she has breast cancer, the probability is 90% that she has a positive mammogram
    - If she does not have breast cancer, the probability is 7% that she still has a positive mammogram
- Suppose a woman has a positive mammogram: What is the probability she actually has breast cancer?

Bayes rule for events

$$P(C = 1 \mid M = 1) = \frac{P(M = 1 \mid C = 1)P(C = 1)}{P(M = 1)}$$

A brief history

Thomas Bayes

- Presbyterian minister (1701-1761)
- Studied logic and theology, interested in probability
- "An Essay towards solving a Problem in the Doctrine of Chances" published in 1763 by his friend Richard Price

Pierre-Simon Laplace

- Did a bunch of stuff (1749-1827)
- Largely responsible for development of Bayesian interpretation of probability ("law of inverse probability")
- **The OG Bayesian demographer**
  - Studying the sex ratio at birth in Paris (1781)
  - Over 1745-1770, observed 251,527 boys and 241,945 girls
  - Denote $x =$ the probability that a given birth was male
  - Laplace was interested in estimating $P(x \leq 0.5)$
  - Using inverse probability he computed $P(x \leq 0.5 \,|\, p = 251{,}527, q = 241{,}945) = 1.152 \times 10^{-42}$

> As it is exceedingly small, we can assert, with the same certainty as any other moral truth, that the difference observed in Paris between births of boys and those of girls is due to a greater likelihood for births of boys (Laplace 1781).

Another reason to dislike Fisher?

- Until early 1920s, the inverse probability method, which is based on what is now called Bayes's Theorem, was pretty much the predominant point of view of statistics.
- R. A. Fisher and Jerzy Neyman, criticized Bayesian inference for the use of subjective elements in an objective discipline. In Fisher's words "The theory of inverse probability is founded upon an error, and must be wholly rejected"
- Frequentist methods became the norm; hypothesis testing developed by Fisher, Pearson, Neyman in early 20th century

Probability does not exist

- De Finetti (1974)
- Exchangeability
- Motivates Bayesian thinking and hierarchical models

Probabilistic reasoning –always to be understood as subjective– merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforeseeable future, or to an unnoticed past, or to a past doubtfully reported or forgotten. . . The only relevant thing is uncertainty - the extent of our own knowledge and ignorance.

de Finetti, 1974

Bayesian bubblings

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

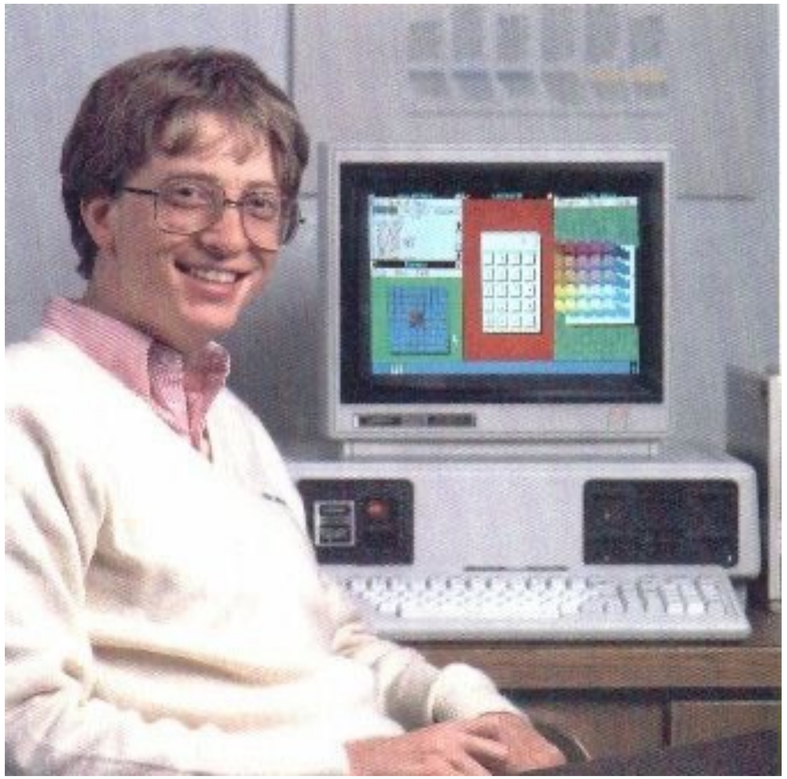
- Beyond the agenda of a few influential statisticians, a big problem with using Bayesian probability for statistical inferences was the lack of computing power
- Bayesian inference often ends up in complicated expressions that cannot be solved on paper
- Need computational algorithms to solve a lot of problems
- Work by physicists in the 1950s and 1960s laid the foundations of algorithms used today (Ulam, Metropolis (Arianna Rosenbluth), Hastings)

Hard to work out!!!!

https://www.nytimes.com/2021/02/09/science/arianna-wright-dead.html

The golden age

- Bayesian methods reappeared in the 1980s/1990s (important paper by Geman and Geman in 1980)
- Rise and rise of Markov Chain Monte Carlo (MCMC) algorithms, which allow for complex Bayesian models to be estimated
- Coupled with rise of more complex data structures and data problems



Bayesian inference in the context of regression

The regression context

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the frequentist set-up:

- estimate β_0, β_1 using OLS
- Make assumptions about normality
- This allows us to write down sampling distribution for, say $\hat{\beta}_1$
- This allows us to perform hypothesis testing about likely value of true β_1

The regression context

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the Bayesian set-up:

- β_0 and β_1 are random variables
- We have some prior knowledge about their values, which we can encode in a prior probability distribution (if we don't know anything, then use non-informative priors)
- After seeing data, use Bayes rule to estimate posterior probability distribution of $\hat{\beta}_0, \hat{\beta}_1$
- Use this distribution to get values of expected value of β_1, variance of β_1, etc

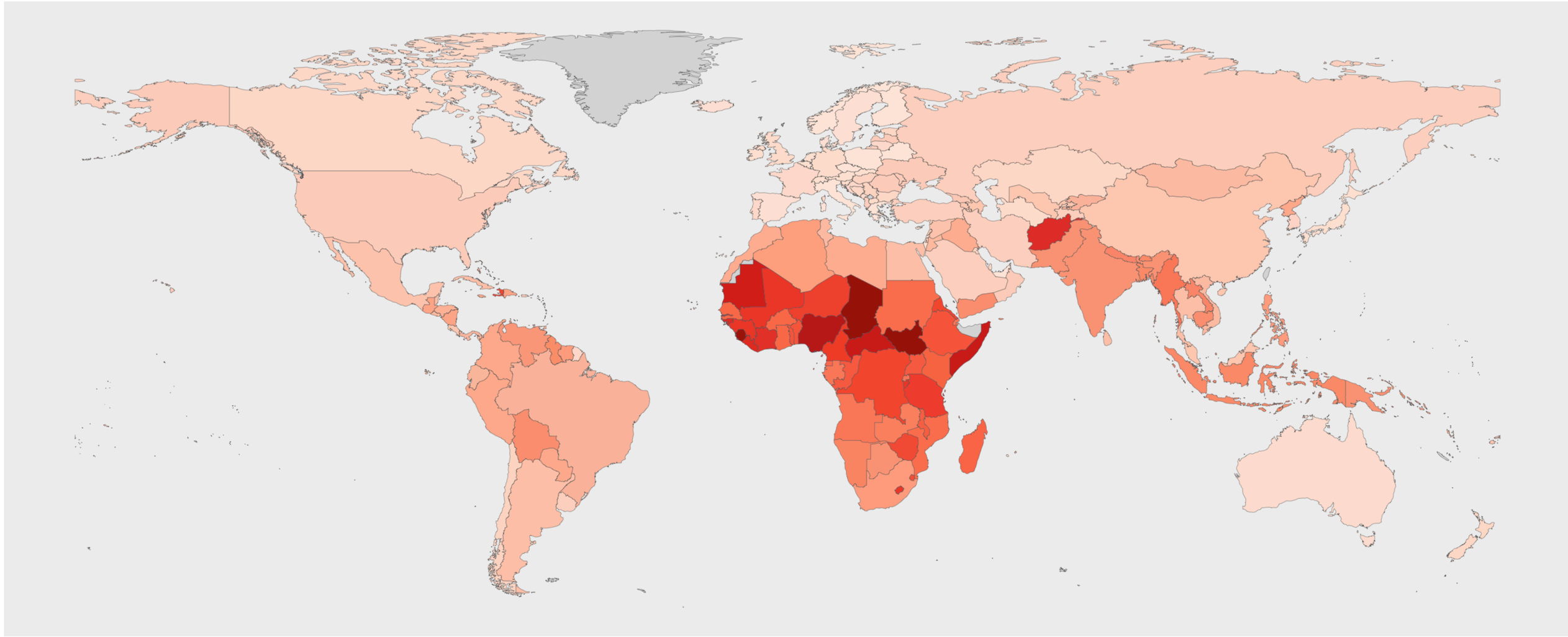
Okay but why

Some common estimation issues

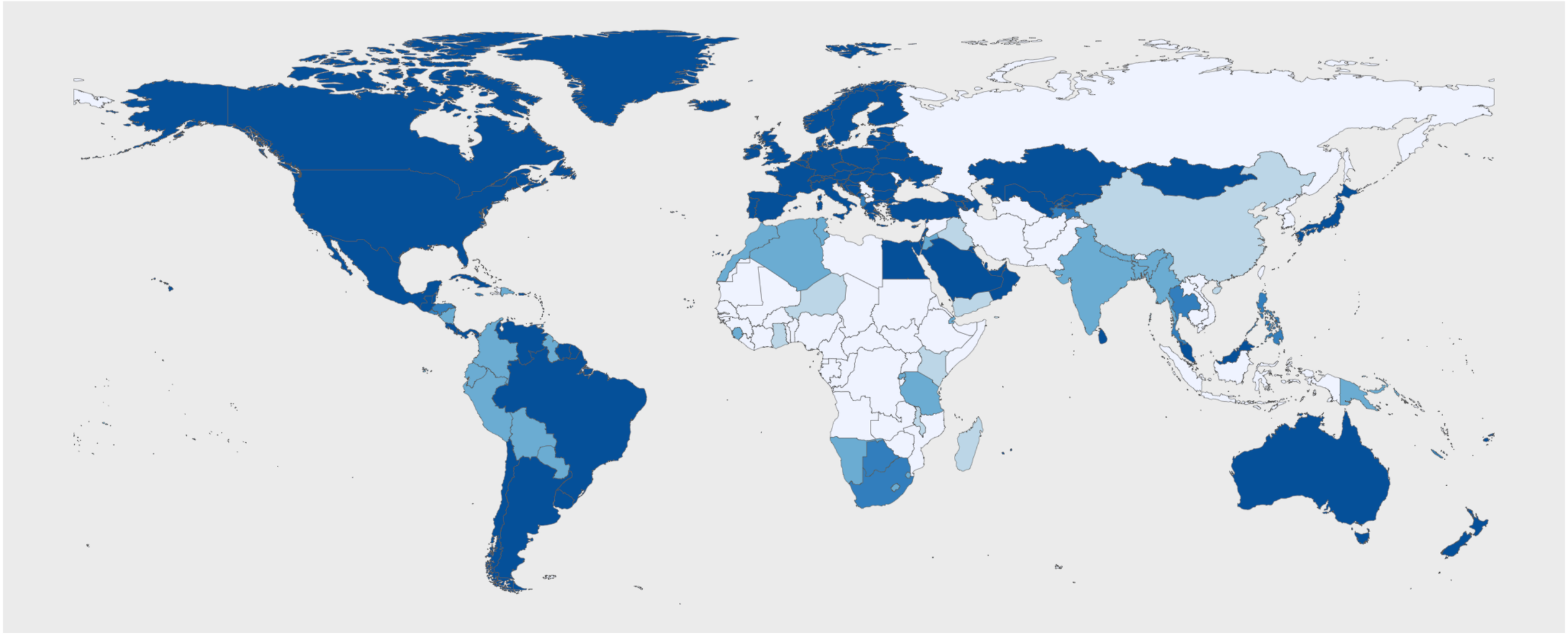
Missing data:

- Even the most fundamental vital statistics (births, deaths) are not recorded for many populations
- Usually data availability is the worst for the most disadvantaged populations

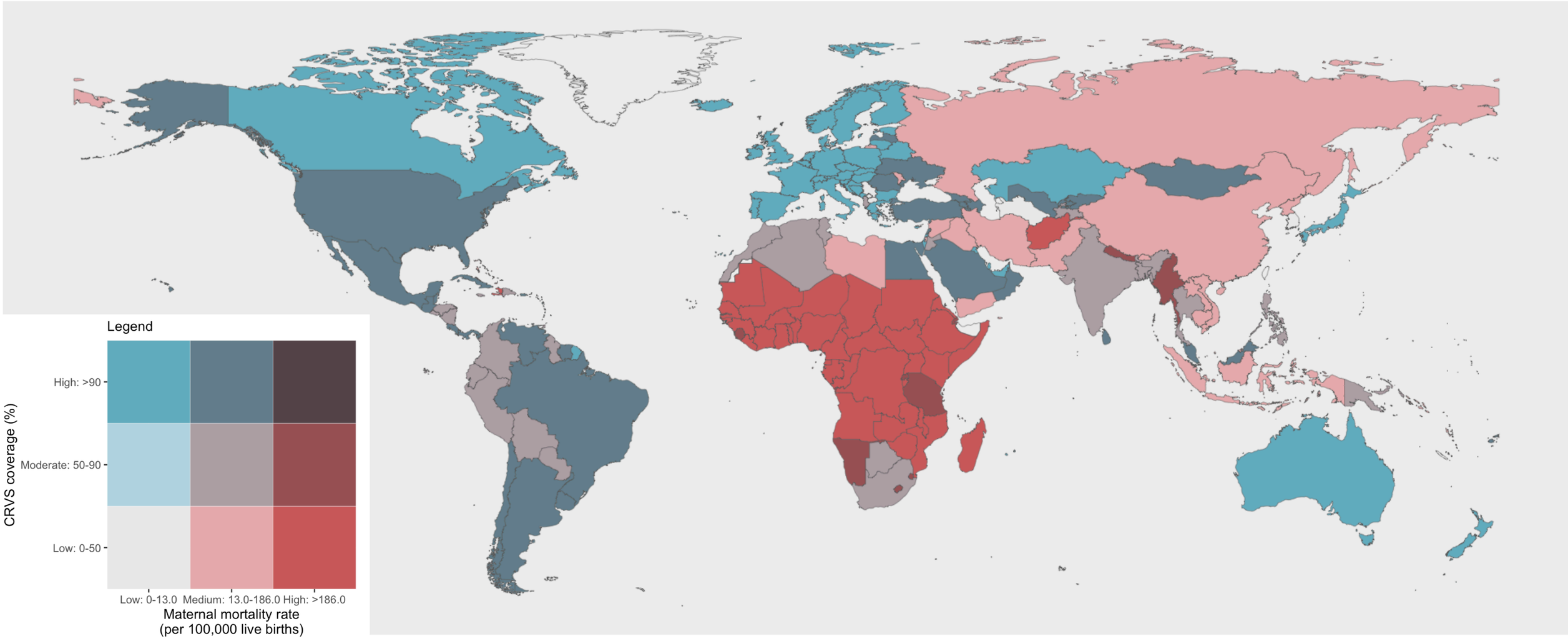
Substantial variation in maternal mortality



Substantial variation in data availability



The dual burden at the global scale

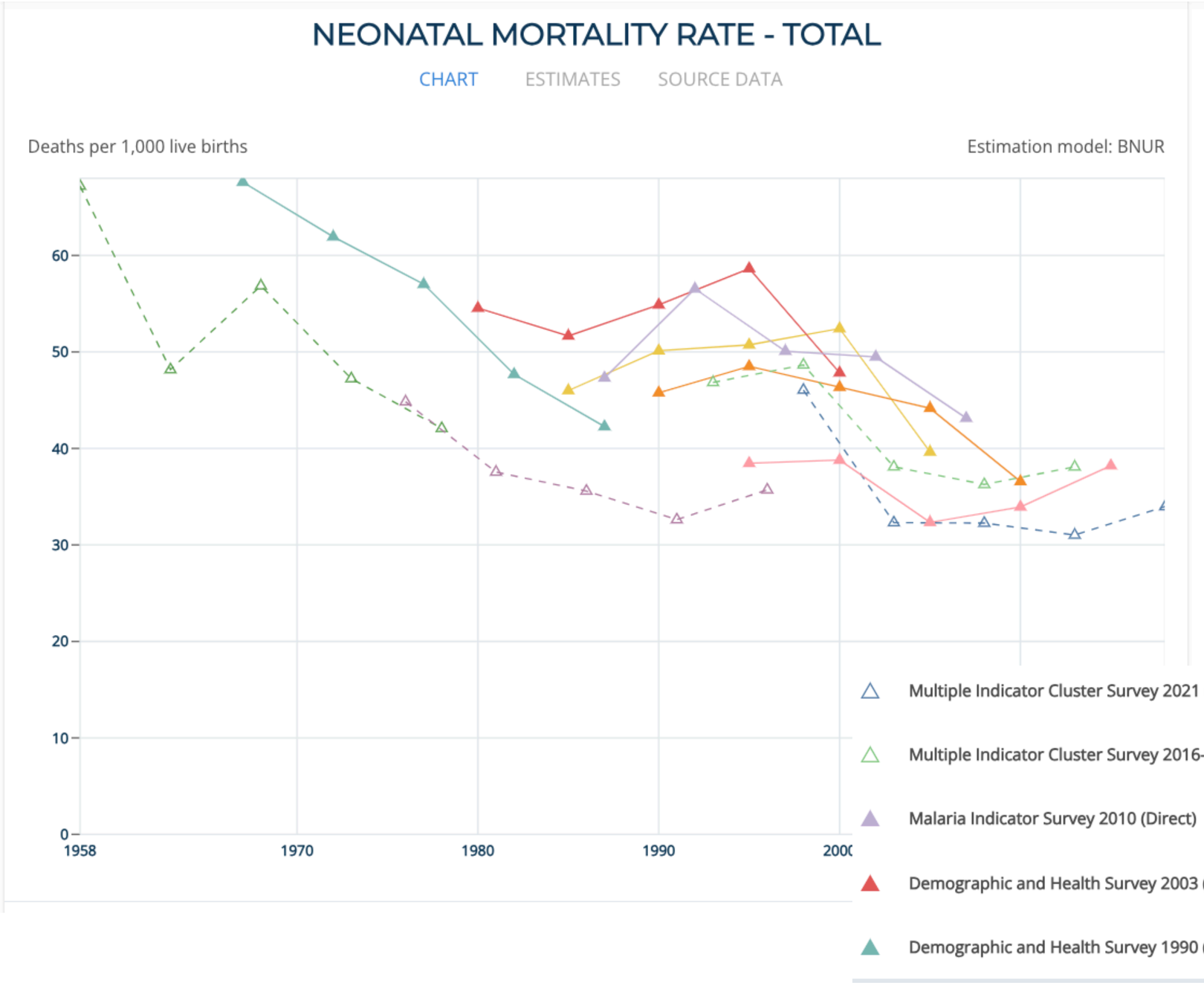


Some common estimation issues

Multiple data sources:

- Often demographic estimation involves reconciling multiple data sources that may have different sampling schemes, coverage, measurement error, etc
- For example:
  - Numerator / denominator problem
  - Multiple surveys on infant mortality
  - Migration flows reported from the origin and destination countries

Nigeria

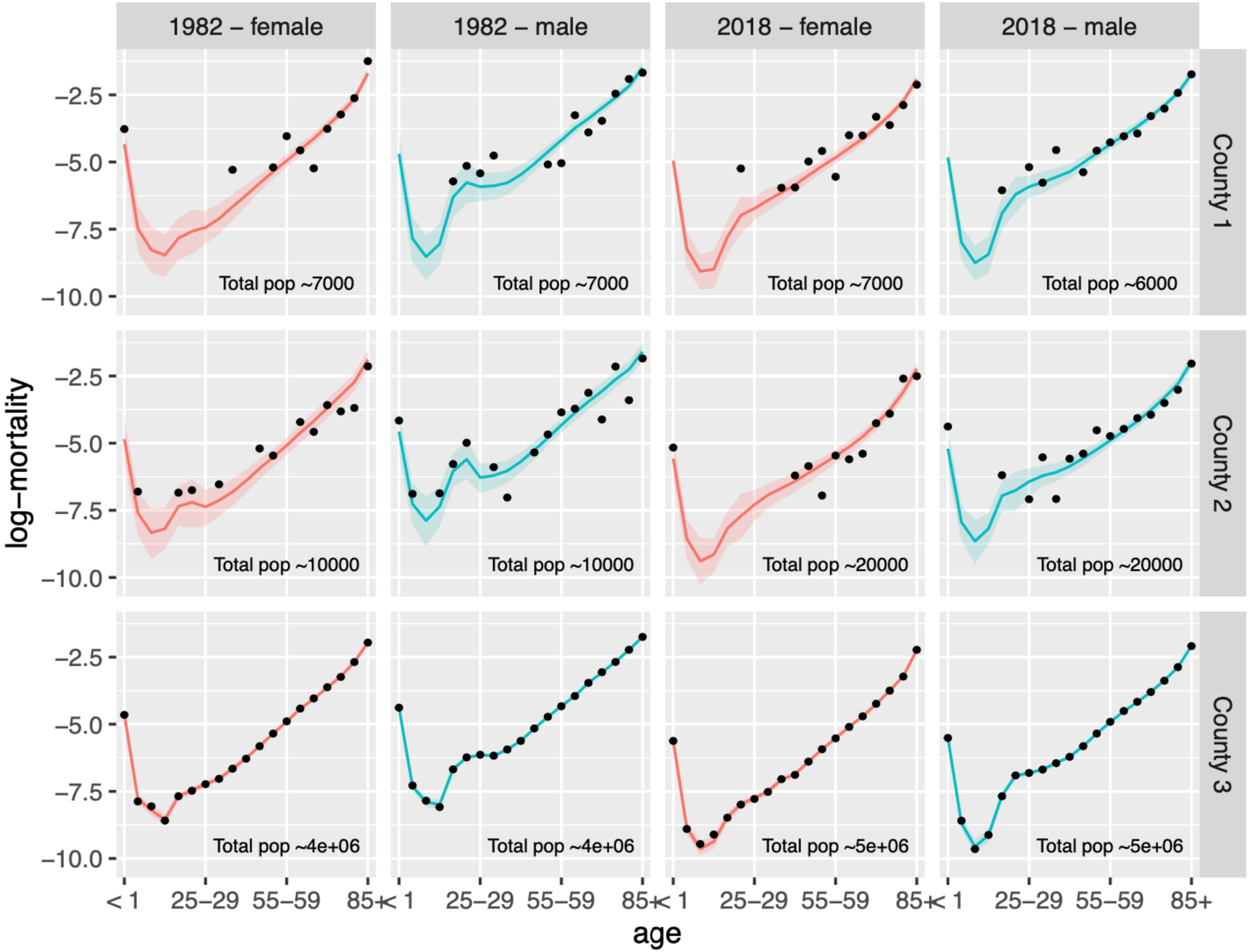


Source: https://childmortality.org/all-cause-mortality/data?indicator=MRM0&refArea=NGA

Some common estimation issues

Small populations / low risk of events:

- Even in populations where the available data is good quality, there still may be estimation issues with small populations
- Increasing demand for estimates at smaller and smaller geographic levels
- Some years / age groups may have zero event counts (e.g. no deaths in a particular year)
- But we still want an estimate of the underlying latent risk of the event



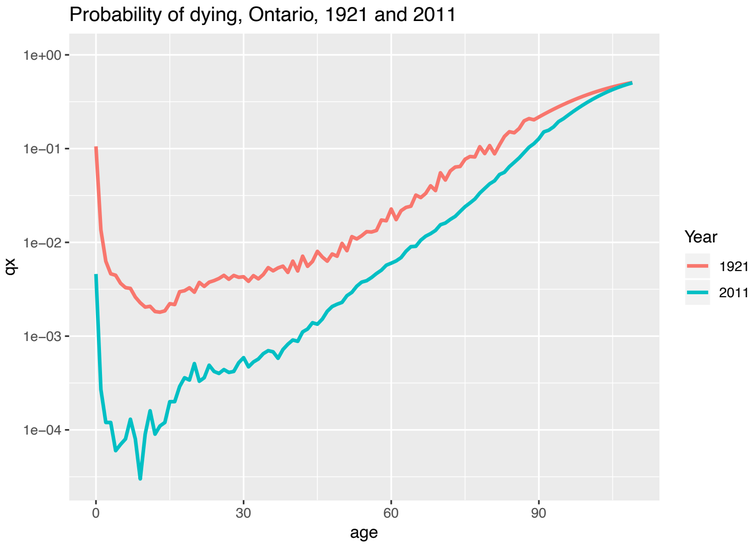
From: Jointly Estimating Subnational Mortality for Multiple Populations" (Dharamshi, Alexander, Winant, and Barbieri, 2024)

Why are Bayesian methods useful?

Why are Bayesian methods useful in demography?

Allow us to incorporate prior information

- In a lot of cases in demographic research, we may have reasonably strong prior information based on historical, biological or social processes.
    - We know that in human populations, age-specific mortality rates have a ‘J’ shape
    - Peak migration is likely to occur in working age groups
    - Total fertility rates are very unlikely to go above the levels historically seen in Hutterite populations



Why are Bayesian methods useful in demography?

Better propagation of uncertainty

- Treating the parameters as random variables allows for better propagation of different types of associated uncertainty
- An added bonus: it's easy to calculate estimates of uncertainty around functions of those parameters
- For demographers, if we're estimating mortality rates (with uncertainty), we can easily convert these to life expectancy estimates (with uncertainty), even though life expectancy is a non-linear function of mortality rates

Why are Bayesian methods useful in demography?

Facilitates combining different data sources, different errors

- May have multiple data sources measuring the same outcome (but with different errors)
- May have to combine different data sets to get measurements of different outcomes
- We can write down a data model / likelihood:

observed outcome = latent outcome + error

- We can then model the source of the **error** differently based on the data source and what we know about it:
  - sampling and non-sampling error
  - biases

Why are Bayesian methods useful in demography?

Natural framework for hierarchical models:

- Hierarchical models deal with natural structure of individuals within populations
- Very useful in data-limited contexts, where assumptions of exchangeability across, for example, geographic areas are reasonable
- Allow for estimates in areas / populations with very little information to be partially informed by similar areas that have more available information

Bayesian demography timeline

- 1700s: Laplace was the original bayesian demographer (e.g. sex ratios at birth in Paris)
- ????? For a long time (demography as ‘book-keeping’)
- 1980-1990s: Increasing interest in probabilistic population projections
- 2000s: A few papers advocating for ‘subjective Bayesianism’
- 2014: UN changed population projections to be probabilistic and Bayesian
- Now: Flood gates have opened, paradigm shift in the field, demand for estimates has changed

As it is exceedingly small, we can assert, with the same certainty as any other moral truth, that the difference observed in Paris between births of boys and those of girls is due to a greater likelihood for births of boys (Laplace 1781).

Rather than dismiss what is known about populations and the evaluation of data sources, Bayesian demography instead forces the demographer to confront uncertainty in the demographic phenomena and baseline data. The projec-

Daponte et al (1997)

Where to read more

A few different Bayesian demography groups:

- Leontine Alkema (UMass Amherst) (child mortality, family planning indicators)
- Jakub Bijak (Southampton) (migration, UK population)
- Doug Leasure (Oxford) (crisis migration and populations)
- Adrian Raftery (UW) (population forecasting, subnational forecasts)
- Carl Schmertmann (Florida) (estimation with not great data)
- Jon Wakefield (UW) (child mortality, small area estimation, spatial models)
- Emilio Zagheni (Max Planck) (migration)
- Me (Toronto) (mortality, migration, social media data)

Lab

Estimating posterior distributions

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

- It turns out posteriors are usually hard to write down in closed form for even relatively straightforward models
- Markov Chain Monte Carlo (MCMC) algorithms are a class of algorithms that are commonly used to obtain samples from a posterior distribution of interest
- Once we have a set of samples from the posterior of interest, we can use these to make inferences about the parameters of interest (using Monte Carlo approximation)

MCMC in R

- Different options, we will be using Stan
- Probabilistic programming language to implement HMC (a version of MCMC)
- Run through R using rstan package
- Very flexible, write models in a separate text file in ‘Stan’ language, but call and manipulate in R

Now let’s

and probably