# How do populations aggregate?

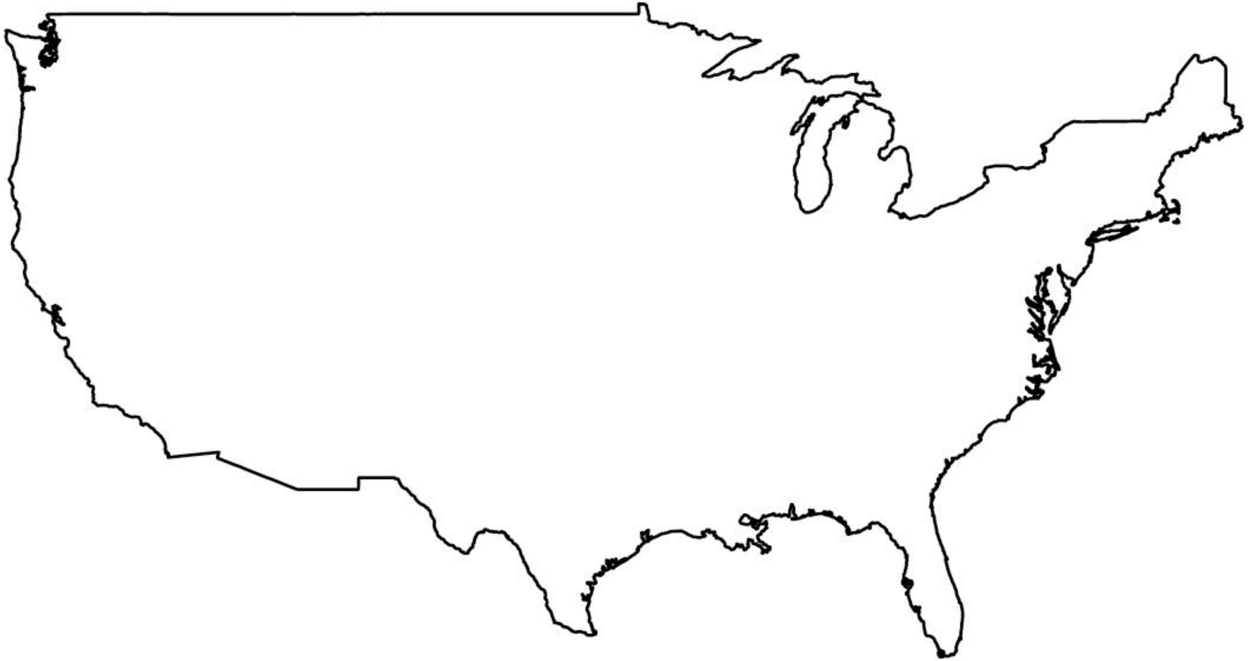
Dennis M. Feehan
UC Berkeley

Elizabeth Wrigley-Field
University of Minnesota

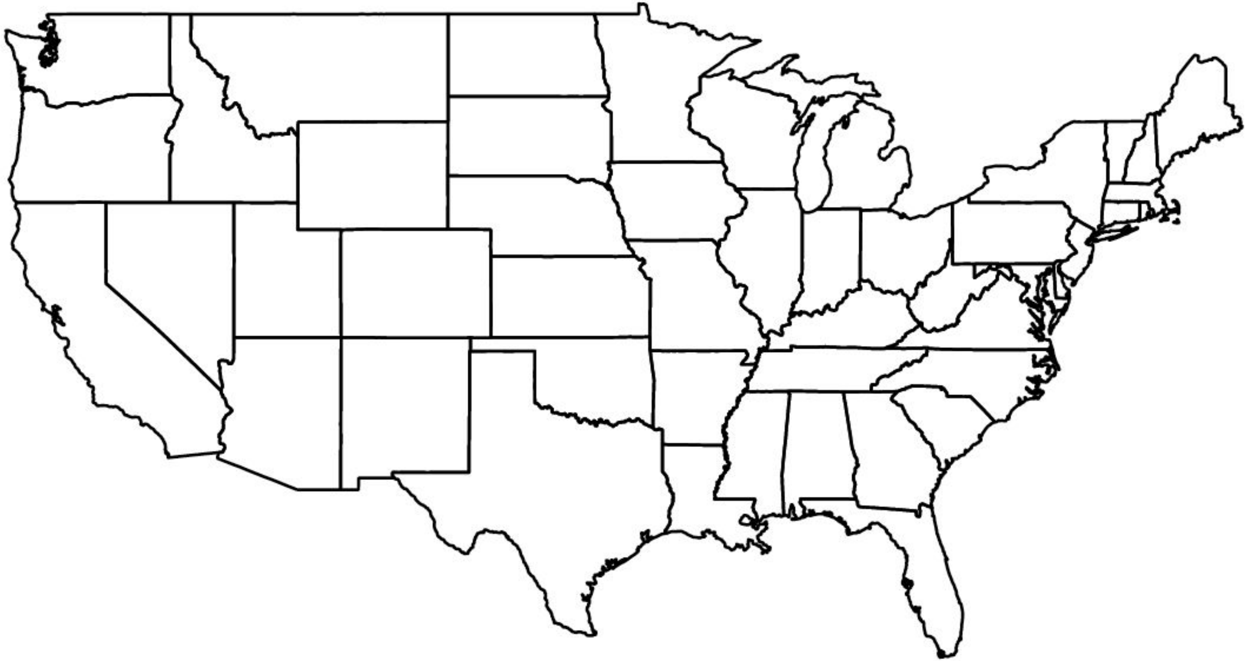
# Overview

- Aggregation and the harmonic mean
- Three applied examples
- Next steps

# Different types of aggregation



# Different types of aggregation



# Different types of aggregation: life table quantities

- Conditional probabilities

- Occurrence/exposure rates in a cohort

# Different types of aggregation: life table quantities

- Conditional probabilities

  => aggregate like the normal (arithmetic) mean, weighted by the number of survivors at the start of the age interval

- Occurrence/exposure rates in a cohort

# Different types of aggregation: life table quantities

- Conditional probabilities

  => aggregate like the normal (arithmetic) mean, weighted by the number of survivors at the start of the age interval

- Occurrence/exposure rates in a cohort

  => aggregate in two ways:
  (1) arithmetic mean of rates, weighted by exposure
  (2) harmonic mean of rates, weighted by events

# Different types of aggregation: life table quantities

- Conditional probabilities

  => aggregate like the normal (arithmetic) mean, weighted by the number of survivors at the start of the age interval

- Occurrence/exposure rates in a cohort

  => aggregate in two ways:
    (1) arithmetic mean of rates, weighted by exposure
    (2) harmonic mean of rates, weighted by events

# Harmonic mean decomposition

$$M = H[\vec{M}; \vec{D}] = \frac{\sum_i {}_nD_x^i}{\sum_i \frac{{}_nD_x^i}{{}_nM_x^i}},$$

Aggregate death rate

Harmonic mean of the subgroup death rates
(weighted by number of deaths in each subgroup)

# Background: generalized means

For a continuous, monotone function g, the **generalized mean** of $x_1$, $x_2$, ... , $x_K$ can be written

$$M_g[\vec{x}] = g^{-1}\left(\frac{1}{k}\sum_i g(x_i)\right)$$

[See, e.g., Carvalho (2016) for a review]

# Background: generalized means

$$M_g[\vec{x}] = g^{-1}\left(\frac{1}{k}\sum_i g(x_i)\right)$$

$$g(x) = x$$
$$g(x) = \log(x)$$
$$g(x) = 1/x$$

# Background: generalized means

$$M_g[\vec{x}] = g^{-1}\left(\frac{1}{k}\sum_i g(x_i)\right)$$

$$g(x) = x \longrightarrow \text{arithmetic mean}$$

$$g(x) = \log(x)$$

$$g(x) = 1/x$$

# Background: generalized means

$$M_g[\vec{x}] = g^{-1}\left(\frac{1}{k}\sum_i g(x_i)\right)$$

$$g(x) = x$$ → arithmetic mean

$$g(x) = \log(x)$$ → geometric mean

$$g(x) = 1/x$$

# Background: generalized means

$$M_g[\vec{x}] = g^{-1}\left(\frac{1}{k}\sum_i g(x_i)\right)$$

$g(x) = x \longrightarrow$ arithmetic mean

$g(x) = \log(x) \longrightarrow$ geometric mean

$g(x) = 1/x \longrightarrow$ harmonic mean

# Background: the harmonic mean

The **harmonic mean** arises when g(x)=1/x; the harmonic mean of $x_1$, $x_2$, ... , $x_K$ is

$$H[\vec{x}] = \left( \frac{1}{K} \sum_i x_i^{-1} \right)^{-1} = \frac{K}{\frac{1}{x_1} + \cdots + \frac{1}{x_K}}$$

(for all $x_i > 0$)

# The harmonic mean of rates: a physical example

Imagine two cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second car travels at 40 miles per hour

What is the average speed at which the two cars travel?

# The harmonic mean of rates: a physical example

Imagine two cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second car travels at 40 miles per hour

What is the average speed at which the two cars travel?

$$\text{avg speed} = \frac{\text{total distance}}{\text{total time}}$$

# The harmonic mean of rates: a physical example

Imagine two cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second car travels at 40 miles per hour

What is the average speed at which the two cars travel?

$$\begin{aligned}\text{avg speed} &= \frac{\text{total distance}}{\text{total time}} \\ &= \frac{2d}{\frac{d}{20} + \frac{d}{40}}\end{aligned}$$

# The harmonic mean of rates: a physical example

Imagine two cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second car travels at 40 miles per hour

What is the average speed at which the two cars travel?

$$\begin{aligned} \text{avg speed} &= \frac{\text{total distance}}{\text{total time}} \\ &= \frac{2d}{\frac{d}{20} + \boxed{\frac{d}{40}}} \end{aligned}$$

distance / rate gives us the time it took at each rate

# The harmonic mean of rates: a physical example

Imagine two cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second car travels at 40 miles per hour

What is the average speed at which the two cars travel?

$$
\begin{aligned}
\text{avg speed} &= \frac{\text{total distance}}{\text{total time}} \\
&= \frac{2d}{\frac{d}{20} + \frac{d}{40}} \\
&= \frac{2}{\frac{1}{20} + \frac{1}{40}}
\end{aligned}
$$

The harmonic mean of 20mph and 40mph

(which equals 26.66mph)

# The harmonic mean of rates: a physical example

Now imagine **three** cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second and third cars travels at 40 miles per hour

What is the average speed at which the three cars travel?

$$\begin{aligned}
\text{avg speed} &= \frac{\text{total distance}}{\text{total time}} \\
&= \frac{3d}{\frac{d}{20} + \frac{d}{40} + \frac{d}{40}} \\
&= \frac{3}{\frac{1}{20} + \frac{1}{40} + \frac{1}{40}} = \frac{3}{\frac{1}{20} + \frac{2}{40}}
\end{aligned}$$

# The harmonic mean of rates: a physical example

Now imagine **three** cars traveling from city A to city B, a distance of *d* miles.

The first car travels at 20 miles per hour
The second and third cars travels at 40 miles per hour

What is the average speed at which the three cars travel?

$$\begin{aligned}\text{avg speed} &= \frac{\text{total distance}}{\text{total time}} \\ &= \frac{3d}{\frac{d}{20} + \frac{d}{40} + \frac{d}{40}} \\ &= \frac{3}{\frac{1}{20} + \frac{1}{40} + \frac{1}{40}} = \frac{3}{\frac{1}{20} + \frac{2}{40}}\end{aligned}$$

We can think of this as a **weighted** harmonic mean of the two speeds, weighted by the number of cars traveling at each speed

The **weighted harmonic mean** of $x_1$, $x_2$, ... , $x_K$ with weights given by $w_1$, $w_2$, …, $w_K$ can be written

$$H[\vec{x};\vec{w}] = \frac{w_1 + \ldots w_K}{\frac{w_k}{x_1} + \cdots + \frac{w_K}{x_K}} = \frac{\sum_i w_i}{\sum_i \frac{w_i}{x_i}}$$

The **weighted harmonic mean** of $x_1$, $x_2$, ... , $x_K$ with weights given by $w_1$, $w_2$, …, $w_K$ can be written

$$H[\vec{x};\vec{w}] = \frac{w_1 + \ldots w_K}{\frac{w_k}{x_1} + \cdots + \frac{w_K}{x_K}} = \frac{\sum_i w_i}{\sum_i \frac{w_i}{x_i}}$$

… **continuous version**: $$H[\vec{x};\vec{w}] = \frac{\int_0^\infty w(x)dx}{\int_0^\infty \frac{w(x)}{f(x)}dx}$$

# Back to deaths

Suppose we have two subpopulations with death rates $m_1=d_1/L_1$ and $m_2=d_2/L_2$

Say $m_1 < m_2$

And suppose we know that an equal number of deaths $d_0$ took place in each subgroup

Then our relationship says that the aggregate death rate is given by $$\frac{2d_0}{\frac{d_0}{m_1} + \frac{d_0}{m_2}}$$

What is going on here?

In order to see the same number of deaths from groups that have different death rates, we must have had different amounts of exposure;
since $m_1 < m_2$, and $d_1=d_2=d_0$, it must be the case that $L_2 < L_1$.

In other words, the terms in the denominator are estimating the different exposures

# Connection to length-biased sampling

Harmonic means often arise in situations where something is being sampled with probability proportional to its value -- **length-biased sampling**

$$f^\star(x_{lb}) \propto x f(x)$$

for x > 0

Prob density fn for x under length-biased sampling

Prob density fn for x in the population

# Aggregating rates using the harmonic mean

Harmonic means often arise in situations where something is being sampled with probability proportional to its value -- **length-biased sampling**

$$f^{\star}(x_{lb}) \propto x f(x)$$

for x > 0

Prob density fn for x under length-biased sampling

Prob density fn for x in the population

$$f^{\star}(x_{lb}) = \frac{x f(x)}{\int x f(x) dx} = \frac{x f(x)}{\mu_x}$$

So, under a **length-biased sample**

$$f^\star(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\iff E^\star\left[\frac{1}{x_{lb}}\right] = \int_0^\infty \frac{1}{x}\frac{xf(x)}{\mu_x} = \frac{1}{\mu_x}$$

[See, e.g., Carvalho (2016)]

The expected value of the reciprocal of a draw from a length-biased sample is equal to the reciprocal of the population mean (i.e., the mean of a sample that is not length-biased)

It turns out that, under length-biased sampling

$$E^\star\left[\frac{1}{x_{lb}}\right] = \frac{1}{\mu_x}$$

So, if we want to estimate the population mean from a LB sample, we take one over the mean of the reciprocals - i.e., the harmonic mean

It turns out that, under length-biased sampling

$$E^{\star}\left[\frac{1}{x_{lb}}\right] = \frac{1}{\mu_x} \longrightarrow$$

So, if we want to estimate the population mean from a LB sample, we take one over the mean of the reciprocals - i.e., the harmonic mean

So in our decomposition of aggregate death rates, we can interpret the harmonic mean as telling us that the death rates aggregate like a size-biased sample:
we see more deaths from higher-mortality subpopulations

$$\frac{2d_0}{\frac{d_0}{m_1} + \frac{d_0}{m_2}}$$

# Arithmetic mean of length-biased sample

$$f^{\star}(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\Longleftrightarrow E^{\star}[x_{lb}] = \int_0^{\infty} x \, \frac{xf(x)}{\mu_x} dx$$

What do we get if we calculate the arithmetic mean of length-biased samples?

# Arithmetic mean of length-biased sample

$$f^\star(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\iff E^\star[x_{lb}] = \int_0^\infty x\,\frac{xf(x)}{\mu_x}dx$$

$$= \frac{1}{\mu}\int_0^\infty x^2 f(x)dx$$

What do we get if we calculate the arithmetic mean of length-biased samples?

# Arithmetic mean of length-biased sample

$$f^{\star}(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\Longleftrightarrow E^{\star}[x_{lb}] = \int_0^{\infty} x \, \frac{xf(x)}{\mu_x} dx$$

$$= \frac{1}{\mu} \boxed{\int_0^{\infty} x^2 f(x) dx}$$

What do we get if we calculate the arithmetic mean of length-biased samples?

This quantity is the expected value of x squared, $E[x^2]$

# Aside: $E[x^2]$

By definition,

$$\mathrm{Var}[x] = E[x^2] - E[x]^2$$

# Aside: $E[x^2]$

By definition,

$$\begin{aligned} \mathrm{Var}[x] &= E[x^2] - E[x]^2 \\ \iff E[x^2] &= \mathrm{Var}[x] + E[x]^2 \end{aligned}$$

# Aside: $E[x^2]$

By definition,

$$
\begin{aligned}
\mathrm{Var}[x] &= E[x^2] - E[x]^2 \\
\iff E[x^2] &= \mathrm{Var}[x] + E[x]^2 \\
\iff E[x^2] &= E[x]^2 \left[1 + \frac{\mathrm{Var}[x]}{E[x]^2}\right]
\end{aligned}
$$

(assuming E[x] > 0)

# Aside: $E[x^2]$

By definition,

$$\begin{aligned}
\mathrm{Var}[x] &= E[x^2] - E[x]^2 \\
\iff E[x^2] &= \mathrm{Var}[x] + E[x]^2 \\
\iff E[x^2] &= E[x]^2 \left[1 + \frac{\mathrm{Var}[x]}{E[x]^2}\right] \\
\iff E[x^2] &= E[x]^2 \left[1 + \mathrm{cv}^2(x)\right]
\end{aligned}$$

(assuming E[x] > 0)

# Arithmetic mean of length-biased sample

$$f^\star(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\Longleftrightarrow E^\star[x_{lb}] = \int_0^\infty x \frac{xf(x)}{\mu_x} dx$$

$$= \frac{1}{\mu} \boxed{\int_0^\infty x^2 f(x) dx}$$

What do we get if we calculate the arithmetic mean of length-biased samples?

This quantity is the expected value of x squared,

$$E[x^2]$$

# Arithmetic mean of length-biased sample

$$f^{\star}(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\Longleftrightarrow E^{\star}[x_{lb}] = \int_0^{\infty} x \, \frac{xf(x)}{\mu_x} dx$$

$$= \frac{1}{\mu} \boxed{\int_0^{\infty} x^2 f(x) dx}$$

What do we get if we calculate the arithmetic mean of length-biased samples?

This quantity is the expected value of x squared,

$$E[x^2] = \mu^2 \left[1 + \mathrm{cv}^2(x)\right]$$

# Arithmetic mean of length-biased sample

$$f^\star(x_{lb}) = \frac{xf(x)}{\mu_x}$$

$$\begin{aligned}\iff E^\star\left[x_{lb}\right] &= \int_0^\infty x\ \frac{xf(x)}{\mu_x}dx \\ &= \frac{1}{\mu}\int_0^\infty x^2 f(x)dx \\ &= \mu\left[1 + \mathrm{cv}^2(x_{lb})\right]\end{aligned}$$

What do we get if we calculate the arithmetic mean of length-biased samples?

So it turns out that, under length-biased sampling, it's also the case that

$$E^\star\left[x_{lb}\right] = \mu_x[1 + \mathrm{cv}^2(x)]$$

The arithmetic mean of length-biased samples

Population mean

Squared coefficient of variation (mean over sd)

# Does it make a difference?

- Data: life tables for males and females in US states in 2010 from the US Mortality database
- Based on these life tables, simulate an aggregate population with 51 subnational units of equal size
- Then compare three aggregation strategies:
  - The correct one (equivalent to harmonic mean of deaths, or arithmetic mean of exposure)
  - The arithmetic mean of the rates, weighted by number of deaths
  - The arithmetic mean of the rates, unweighted
- Results suggest that, yes, this can make an appreciable difference - up to 5 or 10% relative error, in some cases

# Example: developing sensitivity frameworks

# Example: developing sensitivity frameworks

- Many demographic methods have been developed to help solve challenging estimation problems
- These approaches often require assumptions to be made about one group being the same as another group
- Formally understanding aggregation can be helpful for understanding how sensitive these methods are to the assumptions they rely upon
- Example: sibling survival
  (this example comes from work in progress with Gabriel Borges at IBGE)

# Example: developing sensitivity frameworks

- Example: sibling survival
  - Goal: estimate death rates in settings without gold-standard death certificate data
  - Approach: conduct a sample survey and ask respondents to report about deaths and exposure among their siblings
  - Problem: some people have no siblings who are eligible to respond to the survey - they are invisible. We can only estimate death rates for the group that is visible to the sibling histories
- So, it would be useful to understand how important this assumption is -- i.e., we care about the aggregate death rate across visible and invisible people.
- How misleading is it to use the visible death rate to estimate this aggregate?

# Example: developing sensitivity frameworks

How misleading is it to use the visible death rate to estimate this aggregate?

| | |
|---|---|
| $M^I$ | Death rate among people invisible to sibling histories |
| $M^V$ | Death rate among people visible to sibling histories |
| $M$ | Aggregate death rate (which we interested in) |

# Example: developing sensitivity frameworks

How misleading is it to use the visible death rate to estimate this aggregate?

$M^I$ Death rate among people invisible to sibling histories

$M^V$ Death rate among people visible to sibling histories

$M$ Aggregate death rate (which we interested in)

$K > 0$ Multiplicative factor by which invisible and visible death rate differ

$$\Downarrow$$

$$M^I = KM^V$$

$p = \dfrac{D^I}{D^I + D^V}$ Fraction of deaths that is invisible

# Example: developing sensitivity frameworks

$$M = \frac{D^V + D^I}{\frac{D^V}{M^V} + \frac{D^I}{M^I}}$$

Decompose the aggregate death rate using the harmonic mean relationship

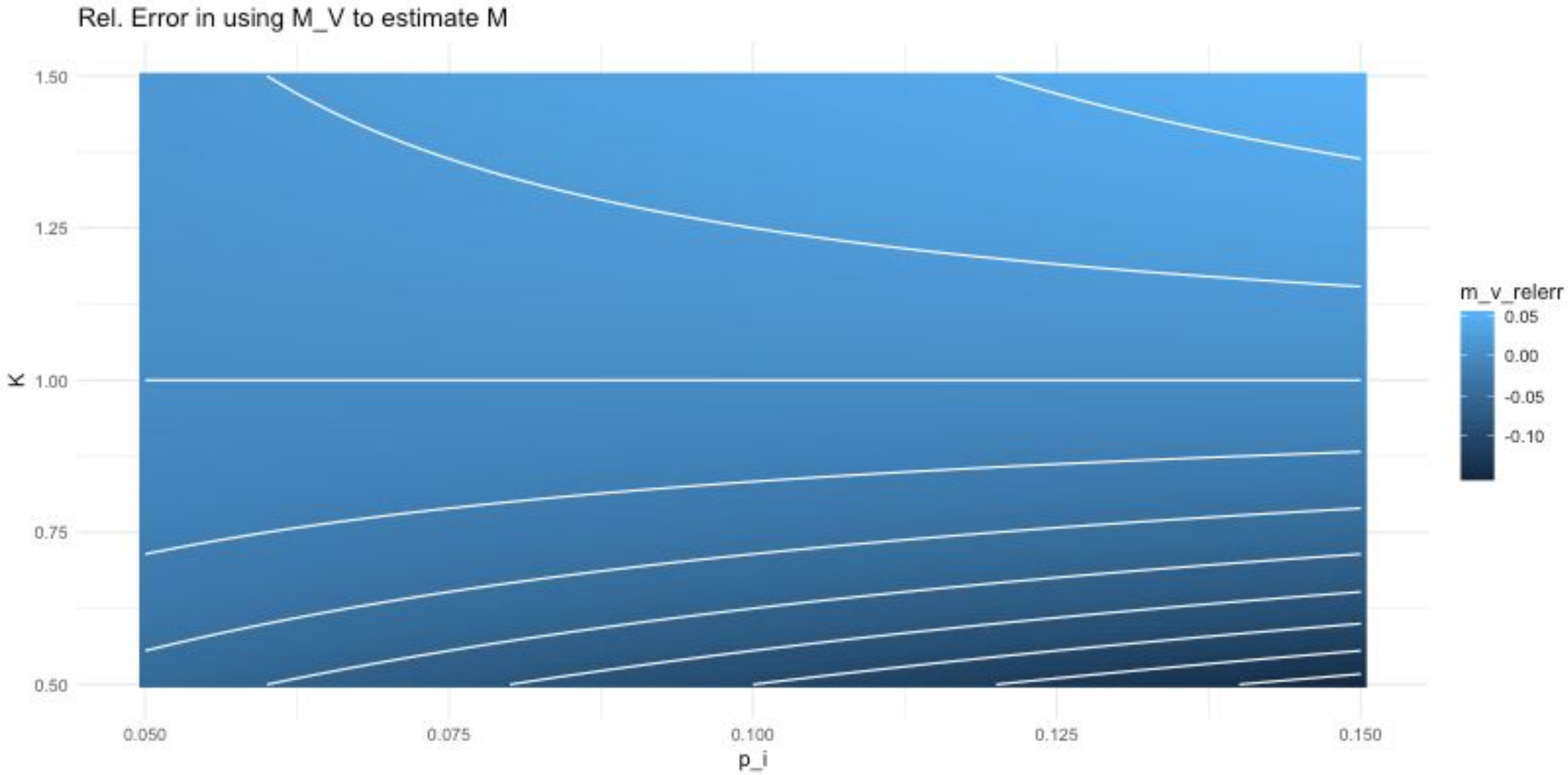
# Example: developing sensitivity frameworks

$$
\begin{aligned}
M &= \frac{D^V + D^I}{\frac{D^V}{M^V} + \frac{D^I}{M^I}} \\
&= \frac{D\left[(1-p) + p\right]}{D\left[\frac{1-p}{M^V} + \frac{p}{M^I}\right]} \\
&= \left(\frac{1-p}{M^V} + \frac{p}{KM^V}\right)^{-1} \\
&= \left(\frac{K(1-p) + p}{KM^V}\right)^{-1} \\
&= M^V\left(\frac{K}{p + K(1-p)}\right).
\end{aligned}
$$

# Example: developing sensitivity frameworks

$$
\begin{aligned}
M &= \frac{D^V + D^I}{\frac{D^V}{M^V} + \frac{D^I}{M^I}} \\
&= \frac{D\left[(1-p)+p\right]}{D\left[\frac{1-p}{M^V} + \frac{p}{M^I}\right]} \\
&= \left(\frac{1-p}{M^V} + \frac{p}{KM^V}\right)^{-1} \\
&= \left(\frac{K(1-p)+p}{KM^V}\right)^{-1} \\
&= M^V\left(\frac{K}{p+K(1-p)}\right).
\end{aligned}
$$

An expression that relates the aggregate death rate to the visible death rate (which can be estimated), in terms of
K - difference between $M^I$ and $M^V$
p - proportion of deaths that is invisible


Rel. Error in using M_V to estimate M
K
1.50
1.25
1.00
0.75
0.50
p_i
0.050
0.075
0.100
0.125
0.150
m_v_relerr
0.05
0.00
-0.05
-0.10

# Example: developing sensitivity frameworks

- So understanding aggregation can be helpful for assessing how sensitive some demographic estimation procedures are to important assumptions

- Note that the arithmetic mean can be used here, too

- And that there are other techniques apart from the sibling method for which this could potentially be useful (eg: Gabriel Borges has applied this to some fertility estimation techniques in his work in Brazil)

# Example: expected lifespan of the living in a stationary population

In a stationary population, the average lifespan of a cohort is e(0), life expectancy at birth.

# Example: expected lifespan of the living in a stationary population

In a stationary population, the average lifespan of a cohort is e(0), life expectancy at birth.

Suppose we take a snapshot of (living) members of a stationary population at a point in time, then follow those people until they die. Call the total lifespan

$$t(x) = x + e(x)$$

Total lifespan of people observed alive at age x in our snapshot

Life lived so far at age x

Average years of life remaining at age x

# Example: expected lifespan of the living in a stationary population

In a stationary population, the average lifespan of a cohort is e(0), life expectancy at birth.

Suppose we take a snapshot of (living) members of a stationary population at a point in time, then follow those people until they die. Call the total lifespan

$$t(x) = x + e(x)$$

What is the relationship between the average total lifespan of this snapshot and life expectancy at birth?

# Example: expected lifespan of the living in a stationary population

Average total lifespan will be

$$\bar{t} = \frac{\int_0^\infty t(x)l(x)dx}{\int_0^\infty l(x)dx}$$

A weighted (arithmetic) average, with weights given by l(x).

# Example: expected lifespan of the living in a stationary population

(Assuming throughout that radix l(0) = 1)

Average total lifespan will be

$$\bar{t} = \frac{\int_0^\infty t(x)l(x)dx}{\int_0^\infty l(x)dx}$$

$$= \frac{1}{e(0)}\int_0^\infty [x + e(x)]\, l(x)dx$$

$$= \frac{1}{e(0)}\int_0^\infty xl(x)dx + \frac{1}{e(0)}\int_0^\infty e(x)l(x)dx$$

# Example: expected lifespan of the living in a stationary population

Average total lifespan will be

$$\bar{t} = \frac{\int_0^\infty t(x) l(x) dx}{\int_0^\infty l(x) dx}$$

$$= \frac{1}{e(0)} \int_0^\infty \left[ x + e(x) \right] l(x) dx$$

$$= \frac{1}{e(0)} \int_0^\infty x l(x) dx + \frac{1}{e(0)} \int_0^\infty e(x) l(x) dx$$

Goldstein (2009) showed that these are equal, and are equal to the average age in the stationary population

# Example: expected lifespan of the living in a stationary population

Average total lifespan will be

$$\bar{t} = \frac{\int_0^\infty t(x)l(x)dx}{\int_0^\infty l(x)dx}$$

$$= \frac{1}{e(0)} \int_0^\infty [x + e(x)]\, l(x)dx$$

$$= \frac{1}{e(0)} \int_0^\infty xl(x)dx + \frac{1}{e(0)} \int_0^\infty e(x)l(x)dx = 2\ \bar{A}_p$$

# Example: expected lifespan of the living in a stationary population

So we have that the average total lifespan of the living is twice the average age:

$$\bar{t} = 2\ \bar{A}_p$$

Seems reasonable: on average, people sampled are halfway through their lives.

# Example: expected lifespan of the living in a stationary population

So we have that the average total lifespan of the living is twice the average age:

$$\bar{t} = 2\ \bar{A}_p$$

Seems reasonable: on average, people sampled are halfway through their lives.

But! We know that, in general,

$$2\ \bar{A}_p \neq e(0)$$

So, how should we think about the relationship between these two quantities?

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx$$

$$\Rightarrow IBP : dv = x dx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x) dx$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx \qquad \Rightarrow IBP : dv = x dx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x)dx$$

$$= \frac{2}{e(0)} \left[ \frac{x^2}{2}\ l(x) \right]_0^\infty - \frac{2}{e(0)} \int_0^\infty -\frac{x^2}{2} d(x) dx$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx \qquad \Rightarrow IBP : dv = xdx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x)dx$$

$$= \frac{2}{e(0)} \left[\frac{x^2}{2} l(x)\right]_0^\infty - \frac{2}{e(0)} \int_0^\infty -\frac{x^2}{2} d(x) dx$$

$$= 0 + \frac{1}{e(0)} \int_0^\infty x^2 d(x) dx$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)}\int_0^{\infty} x l(x) dx \qquad \Rightarrow IBP : dv = xdx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x)dx$$

$$= \frac{2}{e(0)} \left[\frac{x^2}{2}\, l(x)\right]_0^{\infty} - \frac{2}{e(0)}\int_0^{\infty} -\frac{x^2}{2} d(x)dx$$

$$= 0 + \frac{1}{e(0)}\boxed{\int_0^{\infty} x^2 d(x)dx}$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)}\int_0^\infty x l(x) dx$$

$$\Rightarrow IBP : dv = x dx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x)dx$$

$$= \frac{2}{e(0)}\left[\frac{x^2}{2}\ l(x)\right]_0^\infty - \frac{2}{e(0)}\int_0^\infty -\frac{x^2}{2} d(x)dx$$

$$= 0 + \frac{1}{e(0)}\boxed{\int_0^\infty x^2 d(x)dx}$$

d(x) is the density of deaths (since we've taken l(0) = 1), and so we saw earlier that this is

$$E[x^2] = \mu\left[1 + \mathrm{cv}^2(x)\right]$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx$$

$$\Rightarrow IBP : dv = x dx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x) dx$$

$$= \frac{2}{e(0)} \left[ \frac{x^2}{2} l(x) \right]_0^\infty - \frac{2}{e(0)} \int_0^\infty -\frac{x^2}{2} d(x) dx$$

$$= 0 + \frac{1}{e(0)} \boxed{\int_0^\infty x^2 d(x) dx}$$

d(x) is the density of deaths (since we've taken l(0) = 1), and so we saw earlier that this is

$$E[x^2] = \mu \left[1 + \mathrm{cv}^2(x)\right]$$

We also have

$$E[x] = \int_0^\infty x d(x) dx = e(0)$$

# Example: expected lifespan of the living in a stationary population

$$2\ \bar{A}_p = \frac{2}{e(0)} \int_0^\infty x l(x) dx \qquad \Rightarrow IBP : dv = x dx \Leftrightarrow v(x) = x^2/2$$

$$u(x) = l(x) \Leftrightarrow du = -d(x) dx$$

$$= \frac{2}{e(0)} \left[ \frac{x^2}{2} l(x) \right]_0^\infty - \frac{2}{e(0)} \int_0^\infty -\frac{x^2}{2} d(x) dx$$

$$= 0 + \frac{1}{e(0)} \int_0^\infty x^2 d(x) dx$$

$$= e(0) \left[ 1 + \mathrm{cv}^2(x) \right]$$

# Example: expected lifespan of the living in a stationary population

So we have:

$$e(0)$$

Average years of life lived in a cohort

# Example: expected lifespan of the living in a stationary population

So we have:

$$e(0) \qquad 2\bar{A}_p$$

Average years of life lived in a cohort

Average years of life lived among people in our cross-sectional snapshot

# Example: expected lifespan of the living in a stationary population

So we have:

$$e(0) \qquad 2\bar{A}_p = e(0)\left[1 + \mathrm{cv}^2(x)\right]$$

Average years of life lived in a cohort

Average years of life lived among people in our cross-sectional snapshot

# Example: expected lifespan of the living in a stationary population

So we have:

$$e(0) \leq 2\bar{A}_p = e(0)\left[1 + \mathrm{cv}^2(x)\right]$$

Average years of life lived in a cohort

Average years of life lived among people in our cross-sectional snapshot

# Example: expected lifespan of the living in a stationary population

So we have:

$$e(0) \leq 2\bar{A}_p = e(0)\left[1 + \mathrm{cv}^2(x)\right]$$

Average years of life lived in a cohort

Average years of life lived among people in our cross-sectional snapshot

Idea: the people we see in a cross-section are a biased sample of members of the cohorts in the stationary population

The bias comes from the fact that, in the cross section, we see people from each cohort in proportion to their lifespans

# Example: models of heterogeneity in mortality

# Example: models of heterogeneity in mortality

- Understanding aggregation can potentially help provide an alternate way of thinking about theoretical issues
- Example: what Vaupel and Missov (2014) call the 'relative risks and fixed frailty' model
  - Every individual i in a cohort has a fixed frailty parameter $z_i > 0$
  - The hazard individual i faces at age x is given by

$$\mu(x, z) = z\mu(x)$$

where $\mu(x)$ is a baseline hazard for frailty at z=1

# Example: models of heterogeneity in mortality

Under this model, we can think of these deaths as being a size-biased sample of survival cohort members, where the ‘size’ is the frailty parameter z.

# Example: models of heterogeneity in mortality

Under this model, we can think of these deaths as being a size-biased sample of survival cohort members, where the 'size' is the frailty parameter z.

For example, Vaupel, Manton and Stallard (1979) showed that the population, or aggregate, frailty at age x under this model is

$$z^{\dagger}(x) = \bar{z}(x)\left[1 + \mathbf{CV}_z^2(x)\right]$$

Avg. frailty of the dead at x

Avg. population frailty at x

Squared coef. variation in frailty among those alive at x

[Equation from Vaupel and Missov (2014)]

# Example: models of heterogeneity in mortality

Under this model, we can think of these deaths as being a size-biased sample of survival cohort members, where the ‘size’ is the frailty parameter z.

For example, Vaupel, Manton and Stallard (1979) showed that the population, or aggregate, frailty at age x under this model is

$$z^{\dagger}(x) = \bar{z}(x)\left[1 + \mathbf{CV}_z^2(x)\right]$$

This is formula is what we saw earlier - the arithmetic mean from a size-biased sample

Avg. frailty of the dead at x

Avg. population frailty at x

Squared coef. variation in frailty among those alive at x

[Equation from Vaupel and Missov (2014)]

# Conclusion

- Formally understanding aggregation can matter in practical applications
- It can help better understand existing methods
- And it can potentially help conceptualize existing models in a different way

# Thanks!

- Feedback welcome, this idea is still being developed

References


- Borges, Gabriel. *Consistent Population Estimates: an Application to Brazil*. PhD Thesis, UC Berkeley, 2019.
- de Carvalho, Miguel. "Mean, What do You Mean?." *The American Statistician* 70.3 (2016): 270-274.
- Feehan, Dennis M. and Gabriel Borges. “Estimating adult death rates from sibling histories: A network approach.” Working paper (PAA draft).
- Goldstein, Joshua. "Life lived equals life left in stationary populations." *Demographic Research* 20 (2009): 3-6.
- Keyfitz, Nathan, and Gary Littman. "Mortality in a heterogeneous population." *Population studies* 33.2 (1979): 333-342.
- Vaupel, James W., Kenneth G. Manton, and Eric Stallard. "The impact of heterogeneity in individual frailty on the dynamics of mortality." *Demography* 16.3 (1979): 439-454.
- Vaupel, James W., and Trifon I. Missov. "Unobserved population heterogeneity: A review of formal relationships." *Demographic Research* 31 (2014): 659-686.
- United States Mortality Database, https://usa.mortality.org/