Week 2 Methods: Mortality

ECON 125: The Science of Population

Setup

Today, we analyze mortality across countries. Our data are:

- Deaths and mid-year population in 2023 (in 1000s)
- One row for each single year of age for every country
- From United Nations World Population Prospects

Start by setting up R and loading the dataset

```
# Load tidyverse and clear the R environment
library(tidyverse)
rm(list=ls())

# Load dataset and assign it the name country_year_df
country_age_df <- read_csv(url("https://github.com/tomvogl/econ125/raw/
```

Variables

Let's look at the first few rows of the dataset

```
head(country_age_df, 3)
```

```
## # A tibble: 3 x 4
##   country   age deaths   pop
##   <chr>   <dbl>  <dbl> <dbl>
## 1 Burundi     0  16.3   443.
## 2 Burundi     1   1.38  437.
## 3 Burundi     2   1.47  432.
```

Demographers use x to denote age, so let's rename age as `x` in the data

```
country_age_df <- country_age_df |> rename(x = age)
```

Our building block for today is the age-specific mortality rate at age x

$$m_x = 1000 * \frac{d_x}{p_x}$$

where d_x = deaths at age x and p_x = midyear population of age x

```
country_age_df <- country_age_df |> mutate(m = 1000*deaths/pop)
```

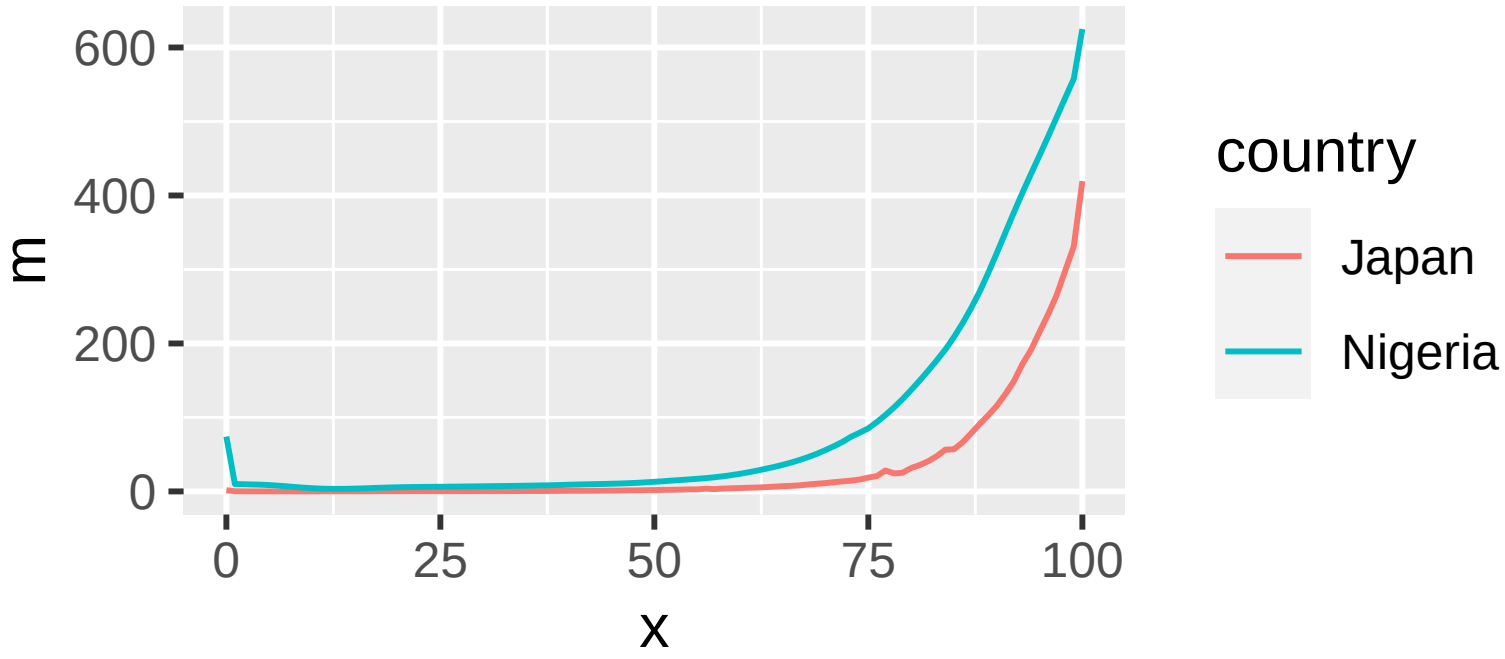
Running example: Japan vs Nigeria

Let's split off a data frame with just Japan and Nigeria

```
j_n_df <- country_age_df |> filter(country=="Japan"|country=="Nigeria")
```

Now let's plot the age pattern of mortality by country

```
ggplot(j_n_df, aes(x = x, y = m, color = country)) +
  geom_line()
```



Typical rich vs poor comparison: Japanese mortality is lower at every age

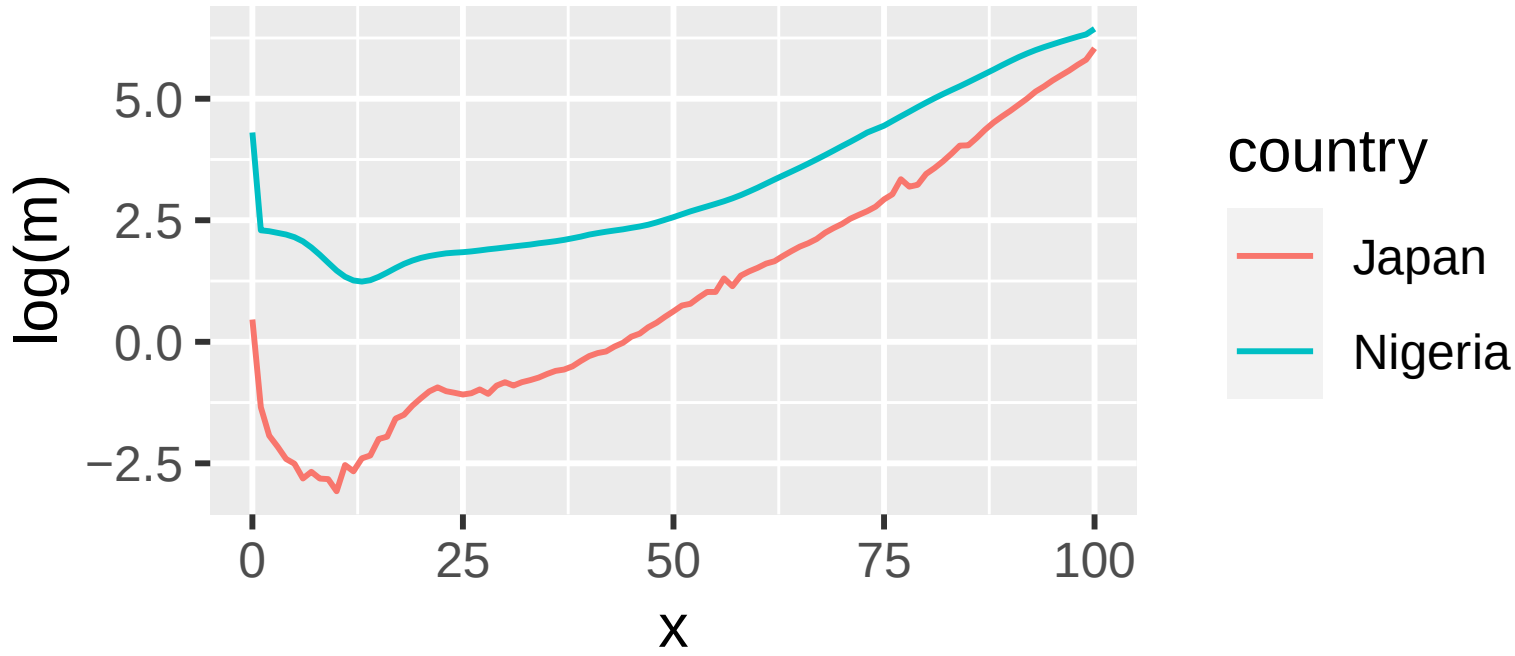
Mortality higher in infancy, lower in older childhood, rising in adulthood

Rescaling y-axis: Japan vs Nigeria

Common to take natural log of mortality rate to see more detail at younger ages

Just write `y = log(m)` instead of `y = m` (equivalent to using a log scale)

```
ggplot(j_n_df, aes(x = x, y = log(m), color = country)) +
  geom_line()
```



Many call this pattern the mortality 'swoosh'

Since $\ln(\frac{a}{b}) = \ln(a) - \ln(b)$, gap in $\ln(m)$ informative about ratio of m

Crude mortality rate: definition

Often, we want a single measure to describe the burden of mortality

The simplest measure is the crude mortality rate

- *CMR* equals total deaths divided by total population
- Low information requirement: only deaths and people, no age
- Commonly applied to settings without death registration systems
- Same as average of age-specific mortality rates, weighted by age distribution

$$CMR = \sum_x \text{share}_x * m_x$$

share_x is the share of people aged x, m_x is the mortality rate at age x

Crude mortality rate: calculation

Let's try it out for Japan and Nigeria

```
j_n_df |>
  group_by(country) |>
  summarise(cmr = sum(pop/sum(pop) * m))
```

```
## # A tibble: 2 x 2
##   country   cmr
##   <chr>   <dbl>
## 1 Japan    12.2
## 2 Nigeria  11.9
```

In both countries, roughly 12 deaths per 1000 people

Japanese mortality is lower than Nigerian at every age, but *CMR* is **higher**

What is going on?

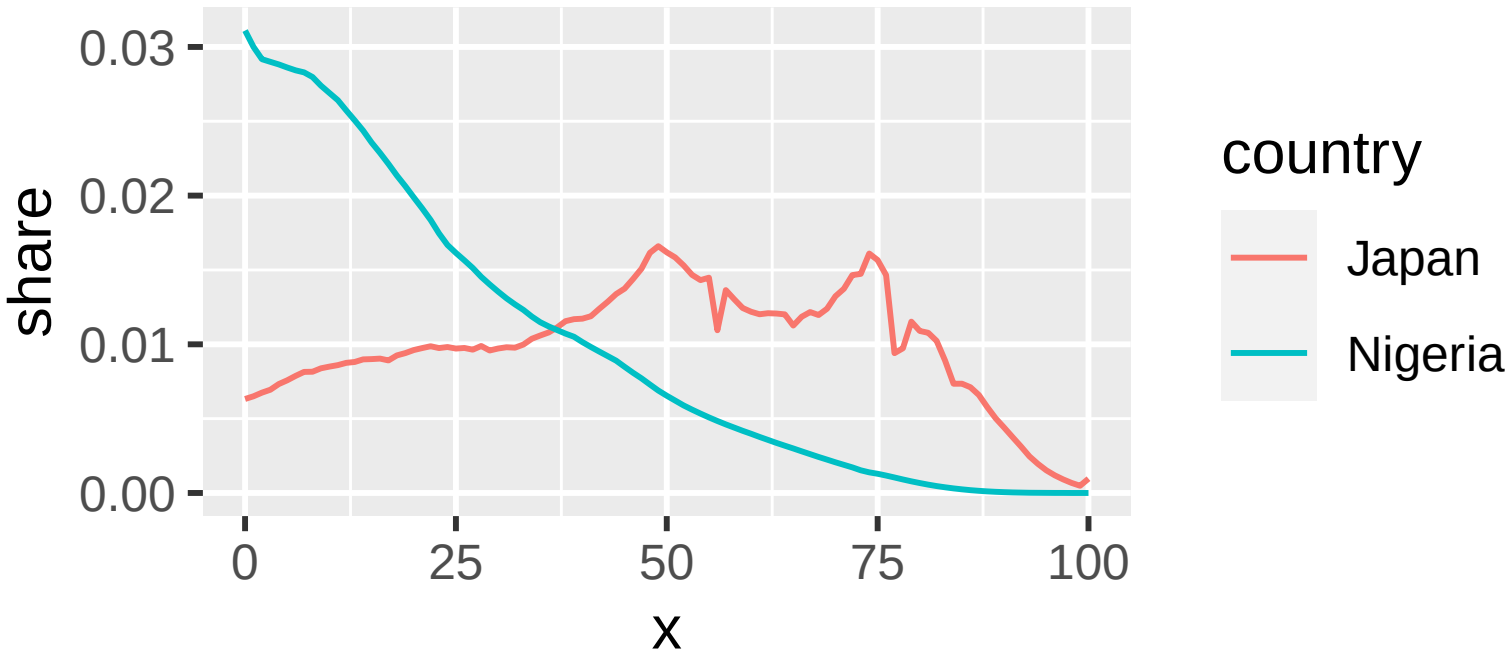
Problem: Japan is much older than Nigeria

Let's calculate age shares and add them to `j_n_df`

```
j_n_df <-
  j_n_df |>
  group_by(country) |>
  mutate(share = pop/sum(pop))
```

Plot them by country, and see the Nigeria has lots more young people

```
ggplot(j_n_df, aes(x = x, y = share, color = country)) +
  geom_line()
```



Solution: age-standardized mortality rate

The age-standardized mortality rate (*ASMR*) uses same shares for both countries

Compute *ASMR* using Nigeria's shares → Nigeria *ASMR* 11x Japan *ASMR*

```
# Step 1: compute Nigeria's shares, save them to a new data frame
nigeria_df <-
  country_age_df |>
  filter(country == "Nigeria") |>
  mutate(nigeria_share = pop / sum(pop)) |>
  select(x, nigeria_share) # keep these variables

# Step 2: apply Nigeria's shares to both countries, compute ASMR
j_n_df |>
  left_join(nigeria_df, by = "x") |>
  group_by(country) |>
  summarise(asmr = sum(nigeria_share * m))
```

```
## # A tibble: 2 x 2
##   country  asmr
##   <chr>   <dbl>
## 1 Japan    1.08
## 2 Nigeria 11.9
```

Problem: choice of age structure is arbitrary

Very different answers using Japan's vs Nigeria's age structure: 4x vs 11x

```
# Step 1: compute Japan's shares, save them to a new data frame
japan_df <-
  country_age_df |>
  filter(country == "Japan") |>
  mutate(japan_share = pop / sum(pop)) |>
  select(x, japan_share)

# Step 2: apply Japan's shares to both countries, compute ASMR
j_n_df |>
  left_join(japan_df, by = "x") |>
  group_by(country) |>
  summarise(asmr = sum(japan_share * m))
```

```
## # A tibble: 2 x 2
##   country  asmr
##   <chr>   <dbl>
## 1 Japan    12.2
## 2 Nigeria  46.3
```

Solution: life expectancy

In terms of a **hypothetical person**, period life expectancy at age x asks...

How many more years would a person aged x expect to live if she experienced current age-specific mortality rates for the rest of her life?

Solution: life expectancy

In terms of a **hypothetical person**, period life expectancy at age x asks...

How many more years would a person aged x expect to live if she experienced current age-specific mortality rates for the rest of her life?

Can equivalently ask in terms of a **hypothetical group of people**...

How many more years would a group of people aged x live on average if they experienced current age-specific mortality rates for the rest of their lives?

Functions of the life table

Suppose...

- l_x people are celebrating their x^{th} birthday
- q_x is the probability of dying before the $(x+1)^{th}$ birthday

Functions of the life table

Suppose...

- l_x people are celebrating their x^{th} birthday
- q_x is the probability of dying before the $(x+1)^{th}$ birthday

Then...

- $d_x = l_x \times q_x$ people die in their x^{th} year
- $l_{x+1} = l_x - d_x = l_x \times (1 - q_x)$ survive to the $(x+1)^{th}$ birthday

Functions of the life table

Suppose...

- l_x people are celebrating their x^{th} birthday
- q_x is the probability of dying before the $(x+1)^{th}$ birthday

Then...

- $d_x = l_x \times q_x$ people die in their x^{th} year
- $l_{x+1} = l_x - d_x = l_x \times (1 - q_x)$ survive to the $(x+1)^{th}$ birthday

Assume that decedents die halfway through the year on average, so...

- $L_x = 0.5 \times d_x + l_{x+1}$ person-years are lived in the x^{th} year

Functions of the life table

Suppose...

- l_x people are celebrating their x^{th} birthday
- q_x is the probability of dying before the $(x+1)^{th}$ birthday

Then...

- $d_x = l_x \times q_x$ people die in their x^{th} year
- $l_{x+1} = l_x - d_x = l_x \times (1 - q_x)$ survive to the $(x+1)^{th}$ birthday

Assume that decedents die halfway through the year on average, so...

- $L_x = 0.5 \times d_x + l_{x+1}$ person-years are lived in the x^{th} year

If we repeat these calculations for ages $x+1$ to x_{max}, then...

- $T_x = L_x + L_{x+1} + \cdots + L_{x_{max}}$ person-years are lived above age x

And the average person lives...

- $e_x^o = \frac{T_x}{l_x}$ more years after the x^{th} birthday

Life table

Putting these all together, we obtain a life table

- The initial value of l_x (called the *radix*) is made up and does not affect e_x^o
- The death probabilities q_x come from the data
- All other elements are derived from q_x and the initial value of l_x

Table 1. Life table for the total population: United States, 2021

Spreadsheet version available from: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/NVSR/72-12/Table01.xlsx.

Age (years)	Probability of dying between ages x and $x+1$ q_x	Number surviving to age x l_x	Number dying between ages x and $x+1$ d_x	Person-years lived between ages x and $x+1$ L_x	Total number of person-years lived above age x T_x	Expectation of life at age x e_x
0–1	0.005446	100,000	545	99,522	7,637,023	76.4
1–2	0.000403	99,455	40	99,435	7,537,501	75.8
2–3	0.000254	99,415	25	99,403	7,438,065	74.8
3–4	0.000192	99,390	19	99,381	7,338,663	73.8
4–5	0.000161	99,371	16	99,363	7,239,282	72.9
5–6	0.000143	99,355	14	99,348	7,139,919	71.9
6–7	0.000130	99,341	13	99,334	7,040,571	70.9
7–8	0.000119	99,328	12	99,322	6,941,237	69.9
8–9	0.000107	99,316	11	99,311	6,841,915	68.9
9–10	0.000095	99,305	9	99,301	6,742,604	67.9
. . .						
95–96	0.239623	7,260	1,740	6,390	21,928	3.0
96–97	0.259772	5,520	1,434	4,803	15,538	2.8
97–98	0.280504	4,086	1,146	3,513	10,735	2.6
98–99	0.301662	2,940	887	2,497	7,222	2.5
99–100	0.323082	2,053	663	1,721	4,725	2.3
100 and older	1.000000	1,390	1,390	3,004	3,004	2.2

SOURCE: National Center for Health Statistics, National Vital Statistics System, mortality data file.

Deriving q_x

The death probability q_x is a little different from the mortality rate m_x

$$q_x = \frac{\text{deaths}}{\text{starting population}} \qquad m_x = \frac{\text{deaths}}{\text{midyear population}}$$

Special case of a general demographic principle

$$\text{probability} = \frac{\text{number of occurrences}}{\text{number of trials}} \qquad \text{rate} = \frac{\text{number of occurrences}}{\text{number of person-years lived}}$$

Deriving q_x

The death probability q_x is a little different from the mortality rate m_x

$$q_x = \frac{\text{deaths}}{\text{starting population}} \qquad m_x = \frac{\text{deaths}}{\text{midyear population}}$$

Special case of a general demographic principle

$$\text{probability} = \frac{\text{number of occurrences}}{\text{number of trials}} \qquad \text{rate} = \frac{\text{number of occurrences}}{\text{number of person-years lived}}$$

These definitions are consistent since we assumed deaths halfway through year

- Assumption works well for most ages, though not the very young or very old
- To keep things simple, we will use the halfway assumption for every age

Under the halfway assumption...

$$q_x = \frac{m_x}{1 + 0.5 \times m_x}$$

which adds people who died in the first half if the year back into the denominator

Calculating q_x in the data

First convert the mortality rate from a 1000 scale to a 1 scale

```
country_age_df <- country_age_df |> mutate(m = m/1000)
```

Now calculate the probability of dying q_x

```
country_age_df <- country_age_df |> mutate(q = m / (1 + 0.5 * m))
```

But revise so that we assume everyone at the oldest age (100) dies

```
country_age_df <-
  country_age_df |>
  mutate(q = if_else(x==100, 1, q))
```

The function `if_else(A, B, C)` returns `B` if condition `A` is true and `C` otherwise

Here, it replaces `q` with the value 1 if `x` is 100 but leaves it unchanged otherwise

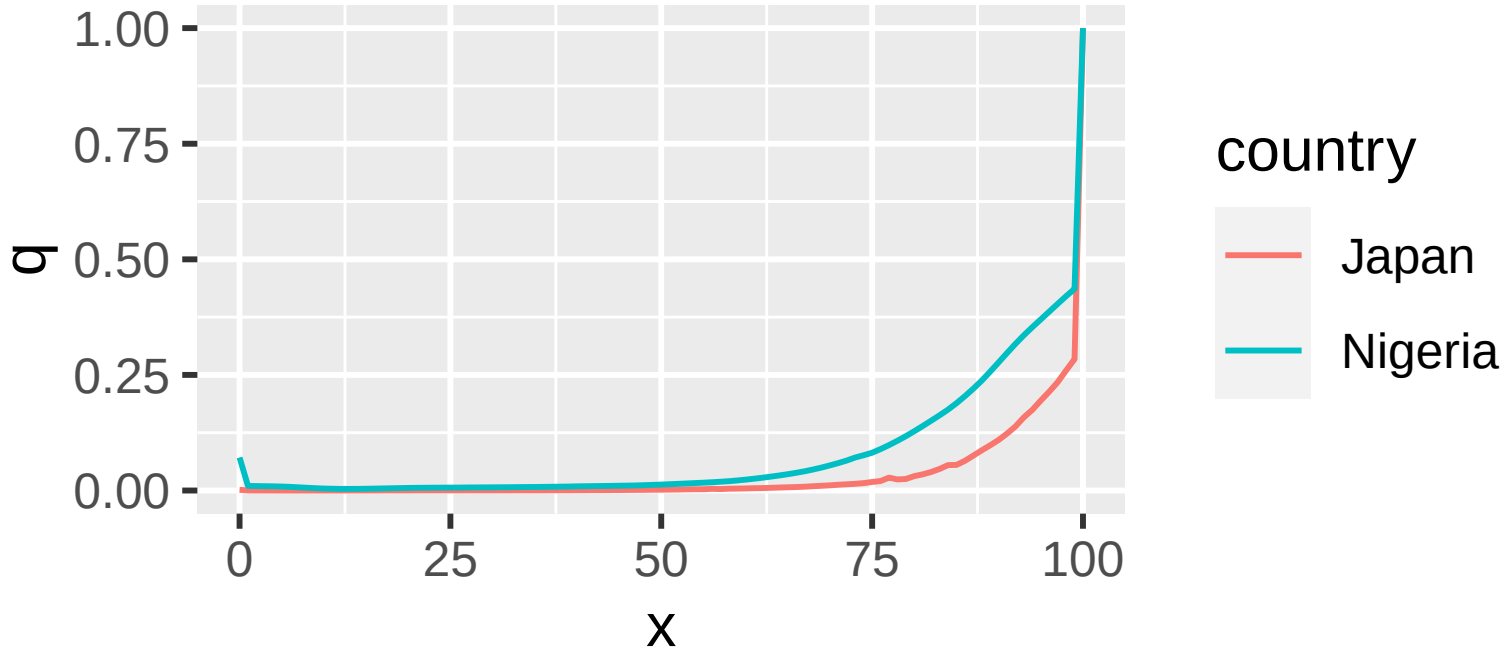
Plotting q_x: Japan vs Nigeria

Let's split off Japan and Nigeria again

```
j_n_df <- country_age_df |> filter(country=="Japan"|country=="Nigeria")
```

Plot q_x over age by country

```
ggplot(j_n_df, aes(x = x, y = q, color = country)) +
  geom_line()
```



Very similar to m_x except at the oldest age

Calculating l_x and d_x in the data

Within each country, let's start with a radix (l_0) of 100,000 and compute all subsequent l_x and d_x

- `arrange()` sorts the data by age
- `lag()` takes the value of q from the last age
- `default = 1` sets the lagged survival probability to 1 in the first row
- `cumprod()` multiplies together all of the past survival probabilities

```
country_age_df <-
  country_age_df |>
  group_by(country) |>
  arrange(x) |>
  mutate(l = 100000 * cumprod(lag(1-q, default=1)),
         d = l*q)
```

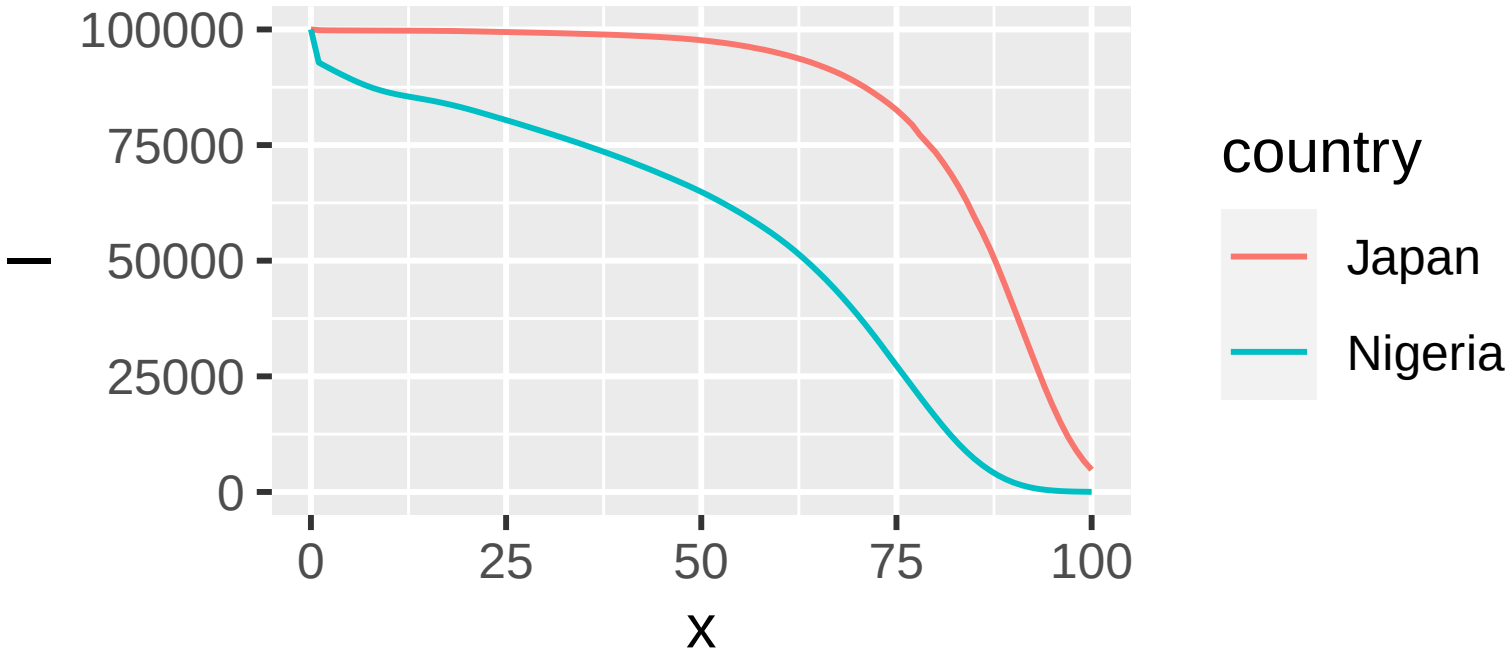
Plotting l_x: Japan vs Nigeria

Split off Japan and Nigeria again

```
j_n_df <- country_age_df |> filter(country=="Japan"|country=="Nigeria")
```

Plot the number of survivors by age, l_x

```
ggplot(j_n_df, aes(x = x, y = l, color = country)) +
  geom_line()
```

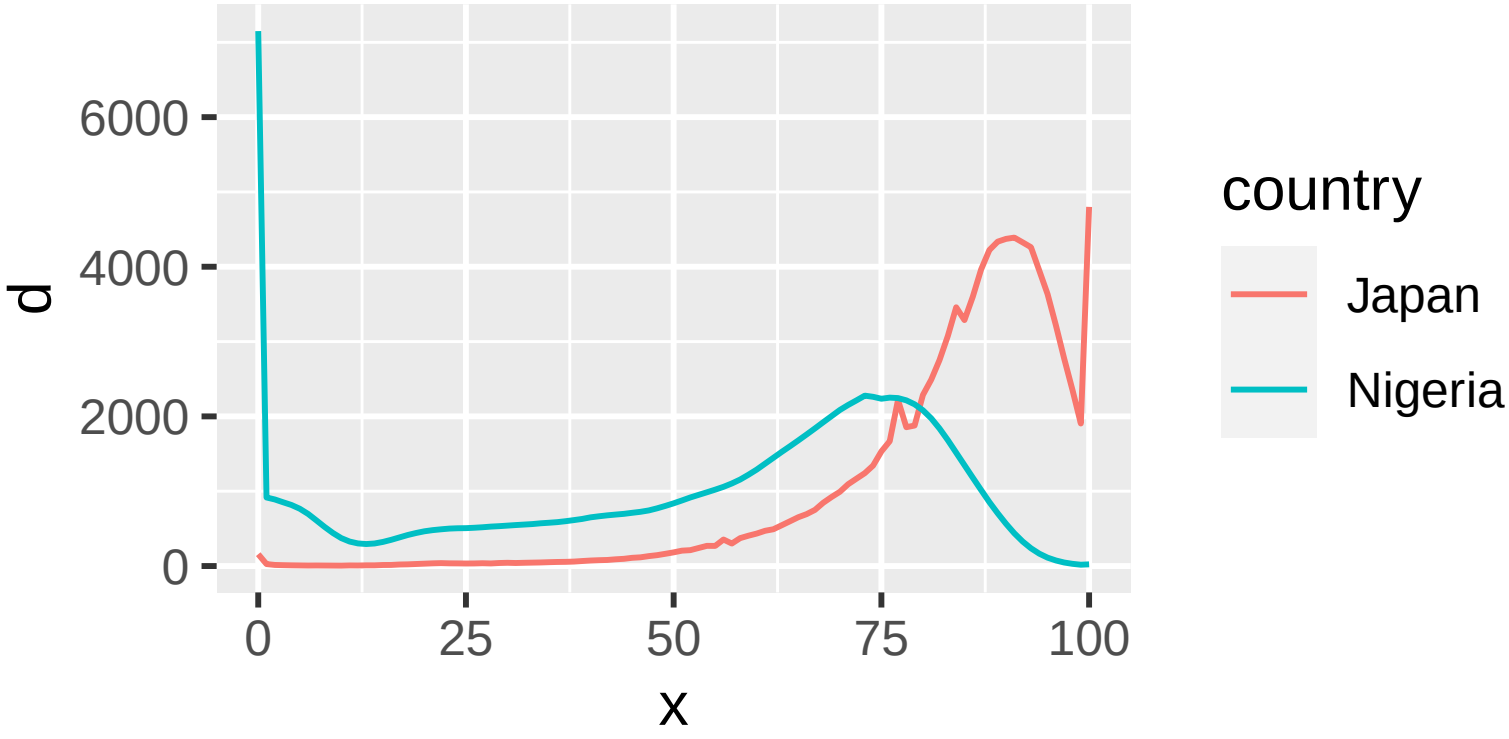


Also known as the survival function, common in biomedical research

Plotting d_x: Japan vs Nigeria

Plot the number of deaths by age, d_x

```
ggplot(j_n_df, aes(x = x, y = d, color = country)) +
  geom_line()
```



Lots of elderly deaths in Japan, childhood deaths in Nigeria

This result is due to the shape of m_x, not the age structure of the population

Calculating L_x, T_x, e_x^o in the data

Set person-years lived at x, L_x, to be

- 0.5 for those who die (d)
- 1 for those who survive the age ($l - d$)

```
country_age_df <- country_age_df |> mutate(L = 0.5*d + (l-d))
```

Calculating L_x, T_x, e_x^o in the data

Set person-years lived at x, L_x, to be

- 0.5 for those who die (d)
- 1 for those who survive the age ($l - d$)

```
country_age_df <- country_age_df |> mutate(L = 0.5*d + (l-d))
```

To obtain person-years lived after x, T_x, for each country...

- Sum person-years lived at all ages using `sum()`
- Subtract person-years lived up through age x using `cumsum()`
- Add back in person-years lived at age x

```
country_age_df <-
  country_age_df |>
  group_by(country) |>
  arrange(x) |>
  mutate(T = sum(L) - cumsum(L) + L)
```

Calculating L_x, T_x, e_x^o in the data

Set person-years lived at x, L_x, to be

- 0.5 for those who die (d)
- 1 for those who survive the age ($l - d$)

```
country_age_df <- country_age_df |> mutate(L = 0.5*d + (l-d))
```

To obtain person-years lived after x, T_x, for each country...

- Sum person-years lived at all ages using `sum()`
- Subtract person-years lived up through age x using `cumsum()`
- Add back in person-years lived at age x

```
country_age_df <-
  country_age_df |>
  group_by(country) |>
  arrange(x) |>
  mutate(T = sum(L) - cumsum(L) + L)
```

To obtain life expectancy e_x^o, divide T_x by l_x

```
country_age_df <- country_age_df |> mutate(e = T/l)
```

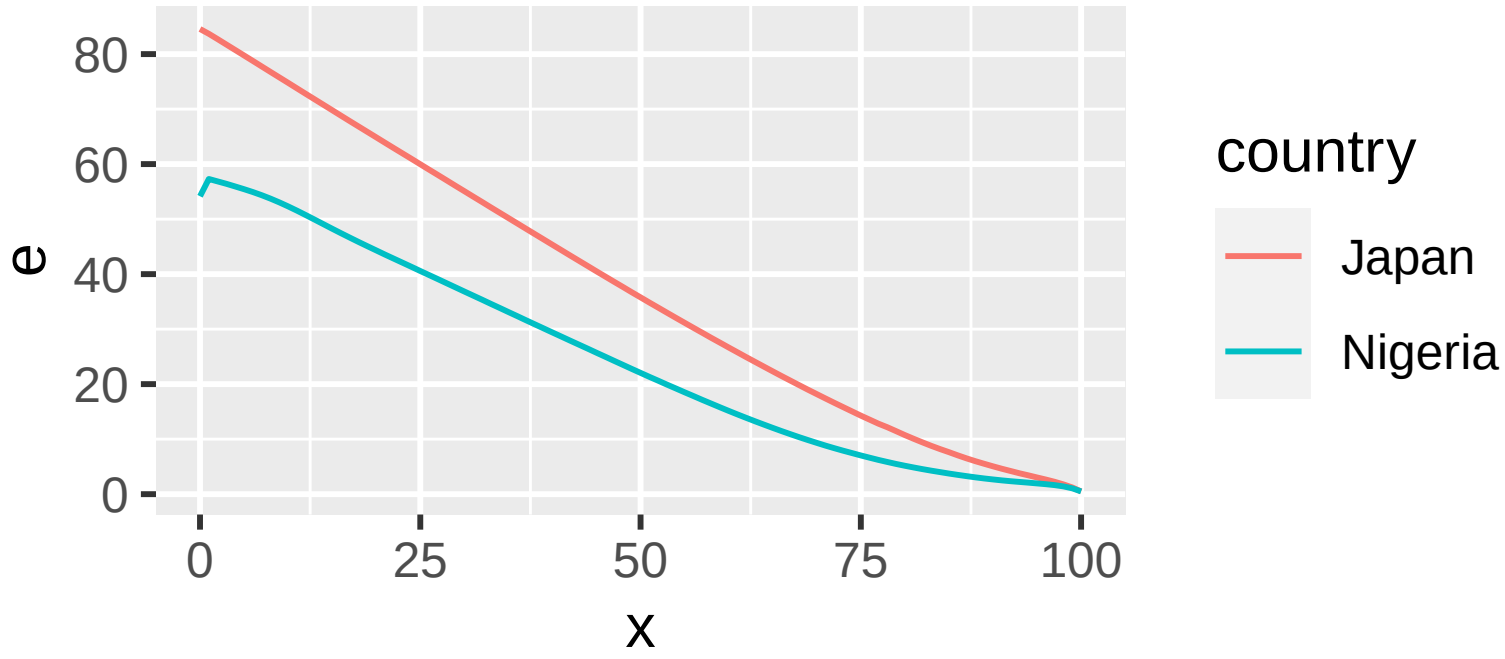
Plotting e_x^o: Japan vs Nigeria

Split off Japan and Nigeria again

```
j_n_df <- country_age_df |> filter(country=="Japan"|country=="Nigeria")
```

Plot life expectancy at each age, e_x^o

```
ggplot(j_n_df, aes(x = x, y = e, color = country)) +
  geom_line()
```



Can easily see the 30-year gap in life expectancy at birth

In Nigeria, $e_1^o > e_0^o$ due to high infant mortality

Global distribution of e_0^o

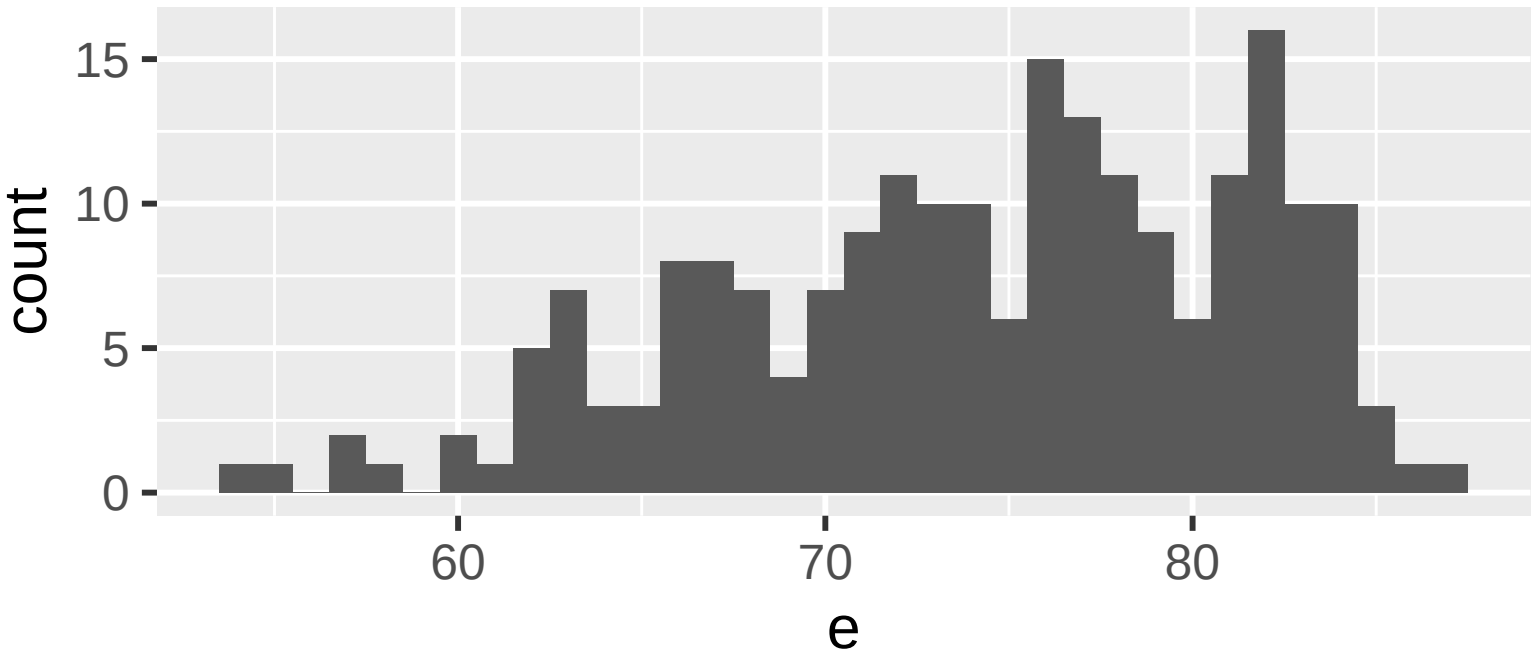
Let's plot a histogram of life expectancy **at birth** across countries in 2023

Use `filter()` to create a new data frame that only has rows with $x = 0$

```
country_e0 <- country_age_df |> filter(x==0)
```

- Use `geom_histogram()` from `ggplot()` to plot the histogram

```
ggplot(country_e0, aes(x = e)) +
  geom_histogram(binwidth=1)
```



Importance of mortality at younger ages

Life expectancy puts a lot of weight on mortality at younger ages

This is a feature of the measurement tool, not necessarily desirable

To see this point, let's simulate Japanese e_x^o under Nigerian infant mortality

Let's look at actual e_0^o and q_0 in Japan and Nigeria...

```
j_n_df |>
  filter(x==0) |>
  select(country, e,q)
```

```
## # A tibble: 2 x 3
##   country     e       q
##   <chr>   <dbl>   <dbl>
## 1 Nigeria  54.2 0.0715
## 2 Japan    84.6 0.00158
```

Simulating Japanese life expectancy with Nigerian infant mortality

Let's set Japan's probability of dying in the first year to Nigeria's level

```
j_n_df <-
  j_n_df |>
  mutate(q = if_else(x==0 & country=="Japan", 0.0715, q))
```

Now let's compute all the other life table functions as before

```
j_n_df <-
  j_n_df |>
  group_by(country) |>
  arrange(x) |>
  mutate(l = 100000 * cumprod(lag(1-q, default=1)),
         d = l*q,
         L = 0.5*d + (l-d),
         T = sum(L) - cumsum(L) + L,
         e = T/l)
```

Simulation results

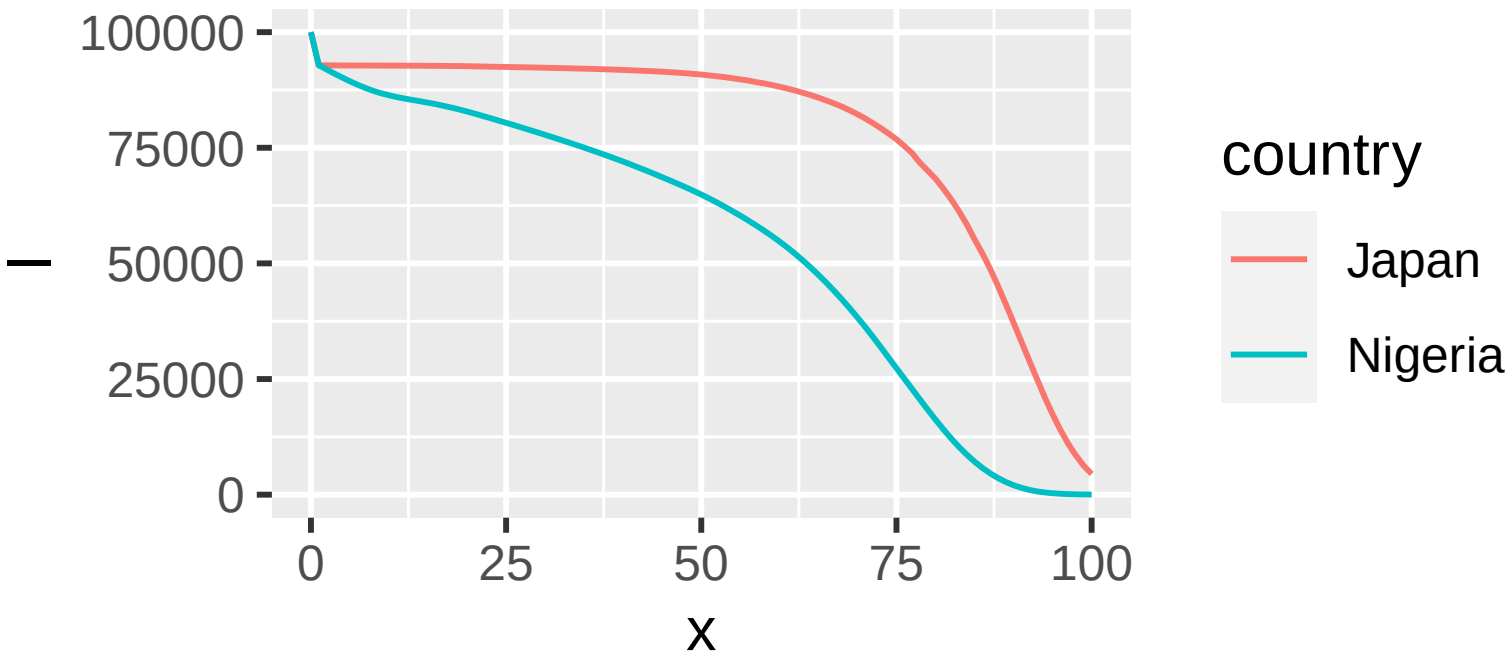
The simulation removed 6 years from Japanese life expectancy at birth!

```
j_n_df |> filter(x==0 & country=="Japan") |>  select(e)
```

```
## # A tibble: 1 x 1
##       e
##   <dbl>
## 1  78.7
```

The reason is that high infant mortality removes survivors from every age > 0

```
ggplot(j_n_df, aes(x = x, y = l, color = country)) +
  geom_line()
```



Period vs cohort

Remember that the preceding material all relates to **period** life expectancy

- Asks about a hypothetical group, a.k.a. *synthetic cohort*
- Nobody will ever experience current rates over their lives
- To calculate, only need a snapshot of age-specific mortality in a single year

Distinct from **cohort** life expectancy

- Asks about an actual group, an actual cohort
- To calculate, need to wait till the whole cohort dies
- Can use same formulas as above or directly compute average age of death