KOSTAT-UNFPA Summer Seminar on Population

Workshop 1. Introduction to Demography

Day 3: Population Structure

Instructor:
Tim Riffe `tim.riffe@ehu.eus`

Assistant:
Inchan Hwang `inchanhwang@utexas.edu`

25 June 2025

Contents

1 Population structure

Population structure refers to population definitions (e.g. age, sex, location, status) that may define universe strata within a dataset with many such strata, or differentiate risk categories within populations considered together, such as age, health, or separate risk strata being combined to represent a whole population (e.g. life expectancy controlling for educational composition). Often, we wish to control for the effects of within-population structure when comparing metrics. The lifetable is intended to do this, as are measures such as TFR, which we saw yesterday.

Some aspects of population structure are not observed, or are only observed in retrospect. For example, the frailty of an individual if inferred using proxy variables, such as questions on disability. Other important strata might be an individual's social position, which might confer advantages in life. This is most often approximated with observable variables such as employment categories, educational attainment, wealth, or income. Trying to define and identify such drivers is a key subfield of contemporary demography. How much do such unobservable aspects of population structure really determine the demography of a population?

1.1 Observed structure

Demographic structure is often used to infer the economic or demographic potential of a population, such as the so-called demographic dividend. Often, this is approximated by the share of population in working ages, with the potential support ratio PSR:

$$PSR = \frac{P(\text{age } 65+)}{P(\text{ages } 15-64)}$$

That is the ratio of the population size above age 65 to that in working ages, where the threshold ages may move up or down depending on what is common in different populations. Does this measure remind you of the GFR? This measure gives some information, but it says nothing of the value of work versus the actual costs of support, and ignores age structure between ages 15 and 64, that some people of working age do not work, and many above age 65 require no support. These and other biases can push the indicator either up or down, and so when possible we prefer informative indicators, such as those based on more granular demographic data. For example, we may structure a population on health, or social capital.

1.2 Social structure

For the case of social capital, let's have a look at how social capital can vary over time using data from the Human Capital Data Explorer Demography and Capital (2018) . This data extract was prepared in advance by Rustam Tursun-zade (`rustam.tursunzade at gmail.com`) using the`wcde` Abel (2023) R package:

Compare

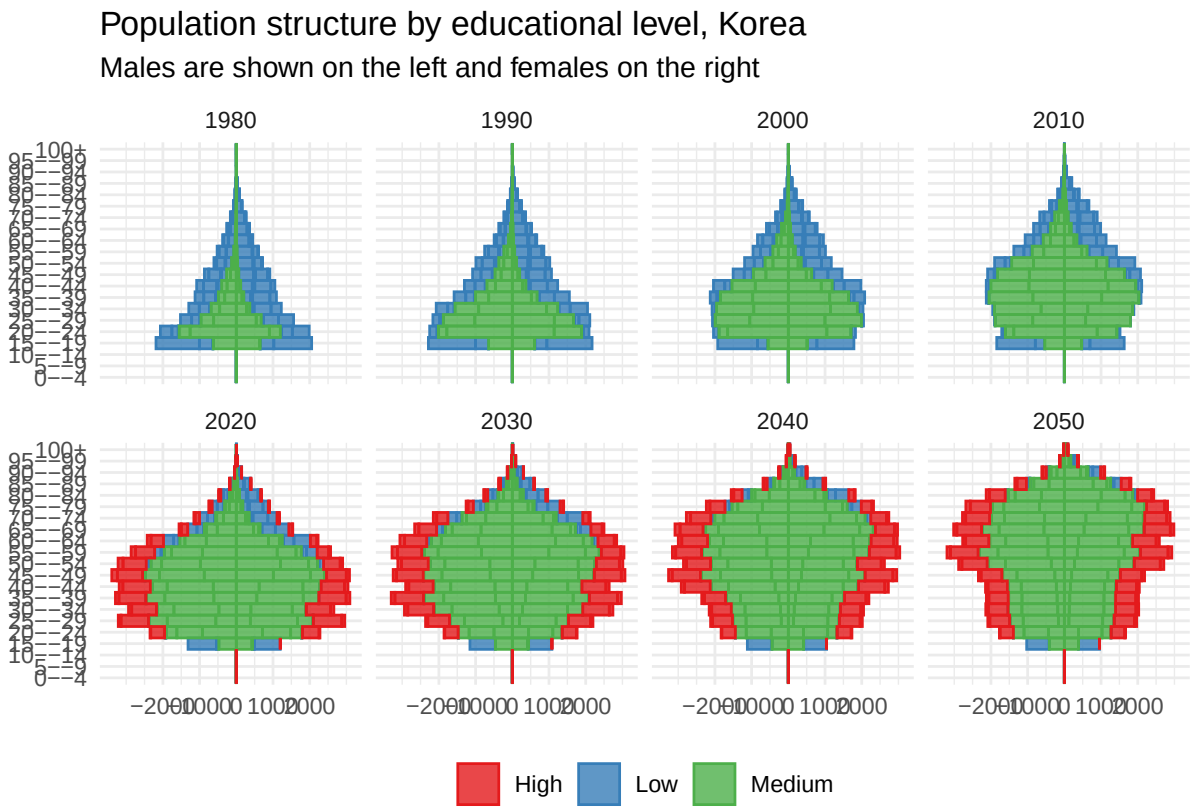
```
library(tidyverse)
# read the data. Note we have to skip first 8 rows with metainformation
wcde <-  read_csv("Data/wcde_data.csv",
                  skip = 8,
                  show_col_types = FALSE) |>
  # transform names to lowercase
  rename_with(tolower)
```

```r
# let us prepare our data
wcde <- wcde |>
  # remove excess educational cathegories
  filter(!education %in% c("Total", "Under 15"),
         # remove info on all ages
         age != "All") |>
  # transfrom educational categories into 3 groups
  mutate(
    education = case_when(
      education %in% c(
        "No Education",
        "Incomplete Primary",
        "Primary",
        "Lower Secondary"
      ) ~ "Low",
      education %in% c("Post Secondary", "Short Post Secondary",
                       "Upper Secondary") ~ "Medium",
      education %in% c("Bachelor", "Master and higher") ~ "High"
    )
  ) |>
  # choose country and years
  filter(area == "Republic of Korea",
         year %in% seq(1980, 2050, 10)) |>
  # note we have to relevel age cathegories for correct visual representation
  mutate(age = fct_relevel(age, "5--9", after = 1)) |>
  mutate(age = fct_relevel(age, "100+", after = Inf)) |>
  mutate(population = ifelse(sex == "Male", -population, population))
```

How has the human capital structure of Korea changed in the past, and how has it been projected into the future?

```r
wcde |>
  ggplot(aes(
    x = age,
    y = population,
    fill = education,
    color = education
  )) +
  geom_bar(stat = "identity",
           width = 1,
           alpha = 0.8) +
  facet_wrap(~ year, ncol = 4) +
  coord_flip() +
  scale_color_brewer(palette = "Set1") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal() +
  scale_y_continuous(breaks = seq(-2000, 2000, 1000)) +
  theme(
    legend.position = "bottom",
    axis.title = element_blank(),
    legend.title = element_blank()
```

```
  ) +
  # add title and subtitle
  ggtitle("Population structure by educational level, Korea",
          subtitle = "Males are shown on the left and females on the right")
```



1.2.1 Social structure and mortality

Certainly behavioral outcomes vary by social structure. Unfortunately so does mortality. Let's look at mortality data by low and high educational attainment. These data come from Bramajo, Permanyer, and Blanes (2023) (many thanks to the authors for sharing the data!), and they may give a sense of how much mortality levels can vary along social strata. To calculate lifetable indicators, I've copied our lifetable functions from yesterday's lesson into an `R` script. We can load these into today's workspace, by `source()`ing the file:

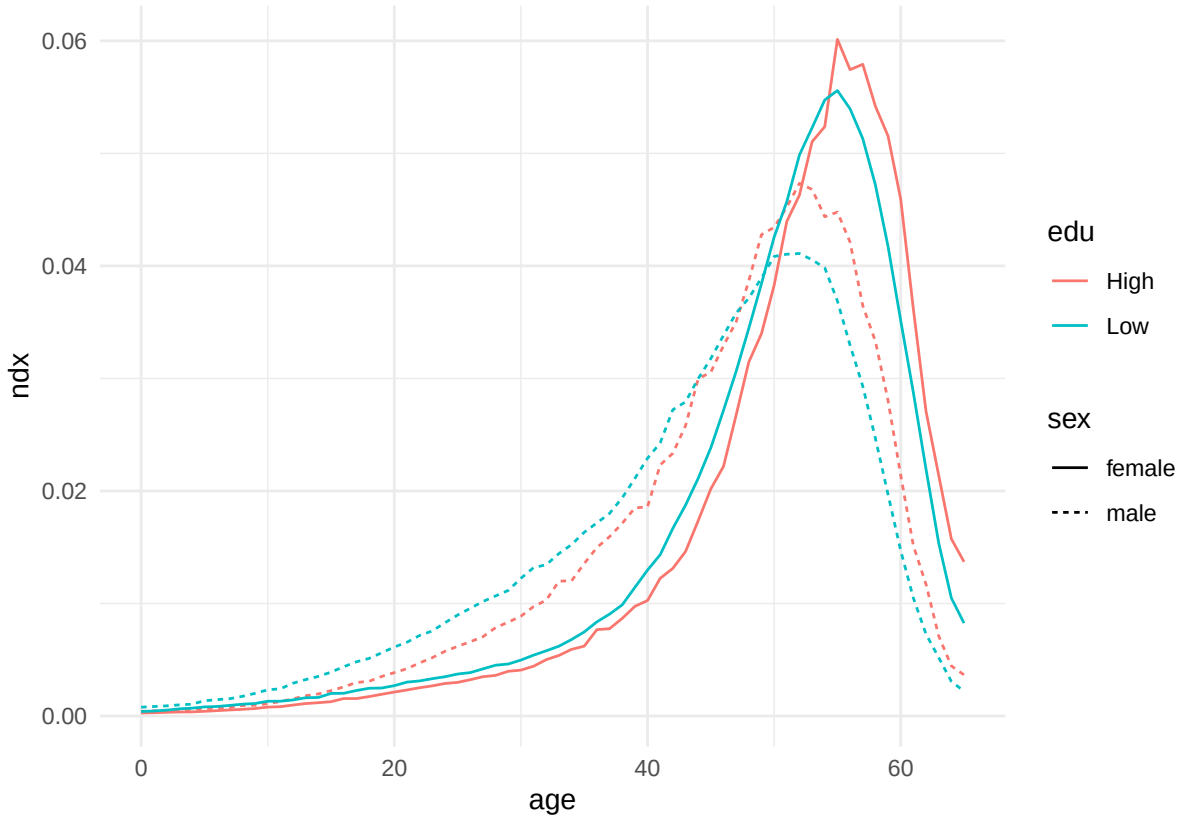
```
source("lifetable_functions.R")

edu_lt <-
  read_delim("Data/DeathsSpain.csv",
                 delim = ";",
               show_col_types = FALSE) |>
  filter(!is.na(ccaa)) |>
  pivot_longer(-(1:3),
               names_to = c("measure","edu"),
               values_to ="value",
               names_sep = "_") |>
  pivot_wider(names_from = "measure", values_from = "value") |>
```

```
  rename(sex = sexo,age=edad) |>
  group_by(sex,age,edu) |>
  summarize(deaths = sum(Deaths),
            pop = sum(Pop), .groups = "drop") |>
  mutate(sex = if_else(sex == 1, "male","female"),
         nMx = deaths / pop,
         nAx = .5,
         n = 1) |>
  group_by(sex,edu) |>
  group_modify(~calc_LT_tidy(data = .x, radix = 1e5)) |>
  ungroup()
```

Let's compare death distributions:

```
edu_lt |>
  ggplot(aes(x = age,
             y = ndx,
             color = edu,
             linetype = sex)) +
  geom_line() +
  theme_minimal()
```



Just eyeballing premature mortality, it seems that (in the long run) it's worse to be male than to have low education in this particular data.

Note the large heterogeneity in ages at death even within education-sex-specific strata. This is a key observation: We could imagine data with more and finer categorization of risk groups, and so stratify the lifetables even more, but there will be diminishing returns to the variance explained

by these groups. Most mortality variation is *within* groups rather than *between* groups. This turns out to be the case even after accounting for plausible levels of unobserved heterogeneity. Age on the other hand, specifically the orders-of-magnitude change in mortality risk over age, explains a lot. It would be hard to find an element of social strata as powerful as age in explaining differences between groups. This is a recent argument of Hal Caswell (Caswell (2023)). This is why we always deal explicitly with age structure when calculating anything in demography.

In exercises, we will calculate within-between metrics of mortality inequality for different populations to back this up, and I will give an example of a future scenario of social differences in mortality that runs counter to this lesson.

2 Population aging

We have established that age is an important determinant of demographic phenomena, but now I will argue that an age-centric view of health can miss key aspects of the demographic present and the future we foresee in countries such as South Korea. Measures such as TFR, classic support ratios, or other summary metrics, such as disability expectancy might miss important perspectives on aging and steer us into thinking that there is a tradeoff between quality and quantity of years lived. Notably, sustained improvements in mortality are probably driven by sustained improvements in health, right? So, maybe longer lives are not so bad in terms of disability or other forms of burden: maybe our standard health measures and assumptions are missing something?

2.1 Flexible threshold ages

Many have argued in the literature for the use of different demographic perspectives when it comes to aging. For example, Sanderson and Scherbov (2007) argues for the use of age thresholds that are rescaled to ages at which demographic thresholds are met rather than subjectively chosen ages. For example, a prospective aging indicator might define a support ratio based on the age at which remaining life expectancy drops to 10 or 15 years, or the age at which a given fraction of the population remains alive (see Alvarez and Vaupel 2023 for further development of this idea). Let's develop each of these ideas.

This time we'll calculate for the entire HMD. This will take no extra effort, since we have our lifetable functions already written!

```
hmd <- read_csv("Data/hmd.csv.gz",
                show_col_types = FALSE) |>
  filter(!is.na(mx)) |>
  mutate(n = 1) |>
  rename(population = country,
         nMx = mx,
         nAx = ax) |>
  group_by(population, sex, year) |>
  group_modify(~calc_LT_tidy(data = .x, radix = 1), .groups = "keep") |>
  filter(sum(is.nan(ex)) < 5) |>
  ungroup() |>
  mutate(ex = if_else(is.nan(ex),0,ex))
```

Now let's write a function that tells us the threshold age for a given level of $l(x)$ or $e(x)$. Note, we'll want a continuous representation of these two functions. While $l(x)$ is guaranteed to be monotonically decreasing, $e(x)$ is not. However, $x+e(x)$ is guaranteed to monotonically decrease:

```r
lx <- hmd |>
  filter(population == "Sweden", year == 2000, sex == "f") |>
  pull(lx)
ex <- hmd |>
  filter(population == "Sweden", year == 2000, sex == "f") |>
  pull(ex)

calc_threshold_age <- function(y, x, threshold, closeout = 115, measure = "lx"){
  x <- c(x,115)
  y <- c(y,0)
  if (measure == "ex"){
    x <- x[-(1:10)]
    y <- y[-(1:10)]
  }
  splinefun(x~y,
            ties = max,
            method = "monoH.FC")(threshold)
}
calc_threshold_age(y = lx, x = 0:110, threshold = .2, measure = "lx")
```

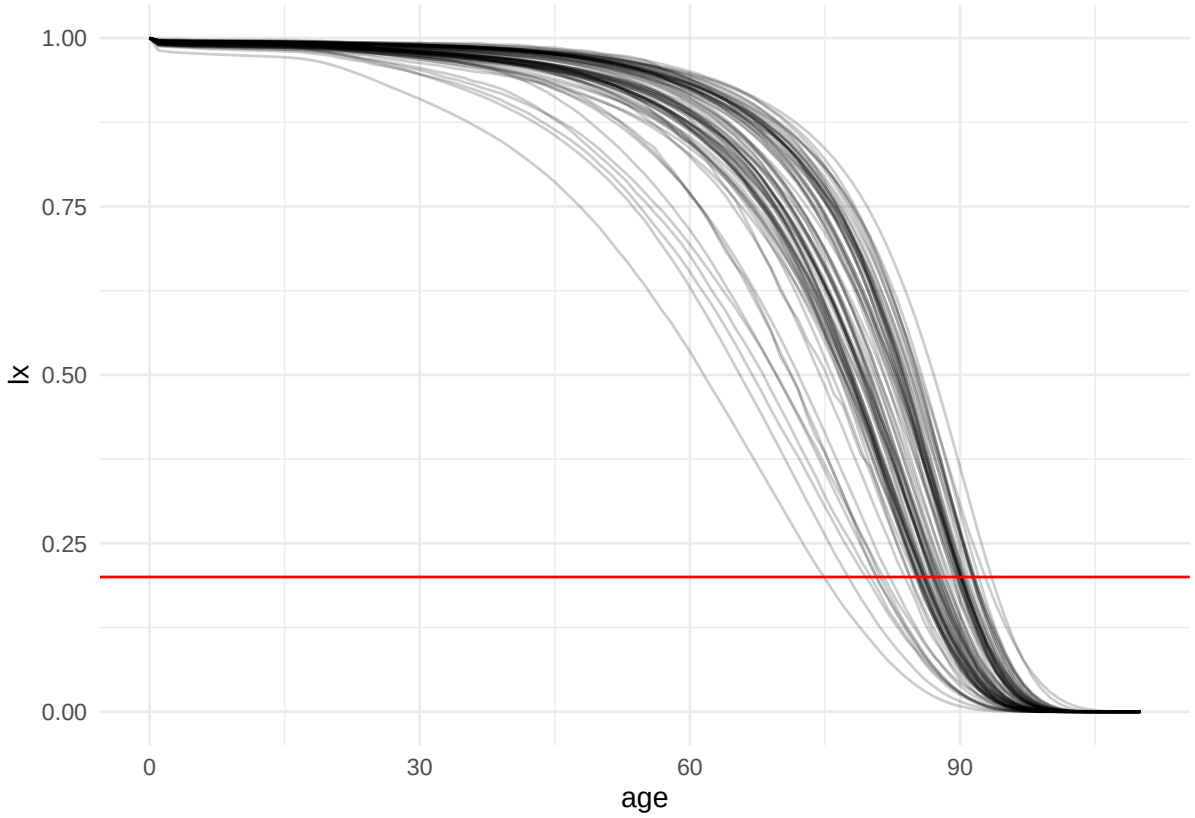
```
## [1] 90.89735
```

```r
calc_threshold_age(y = ex, x = 0:110, threshold = 15, measure = "ex")
```

```
## [1] 70.57082
```

Here's an illustration of how the lx-based threshold age is determined. For each different survival curve drawn, we want to find at which age the curve passes the value of 0.2 (for example):

```r
hmd |>
  filter(year == 2000) |>
  ggplot(aes(x = age, y = lx, group = interaction(population,sex,year))) +
  geom_line(alpha = .2) +
  geom_hline(yintercept = .2, color = "red") +
  theme_minimal()
```

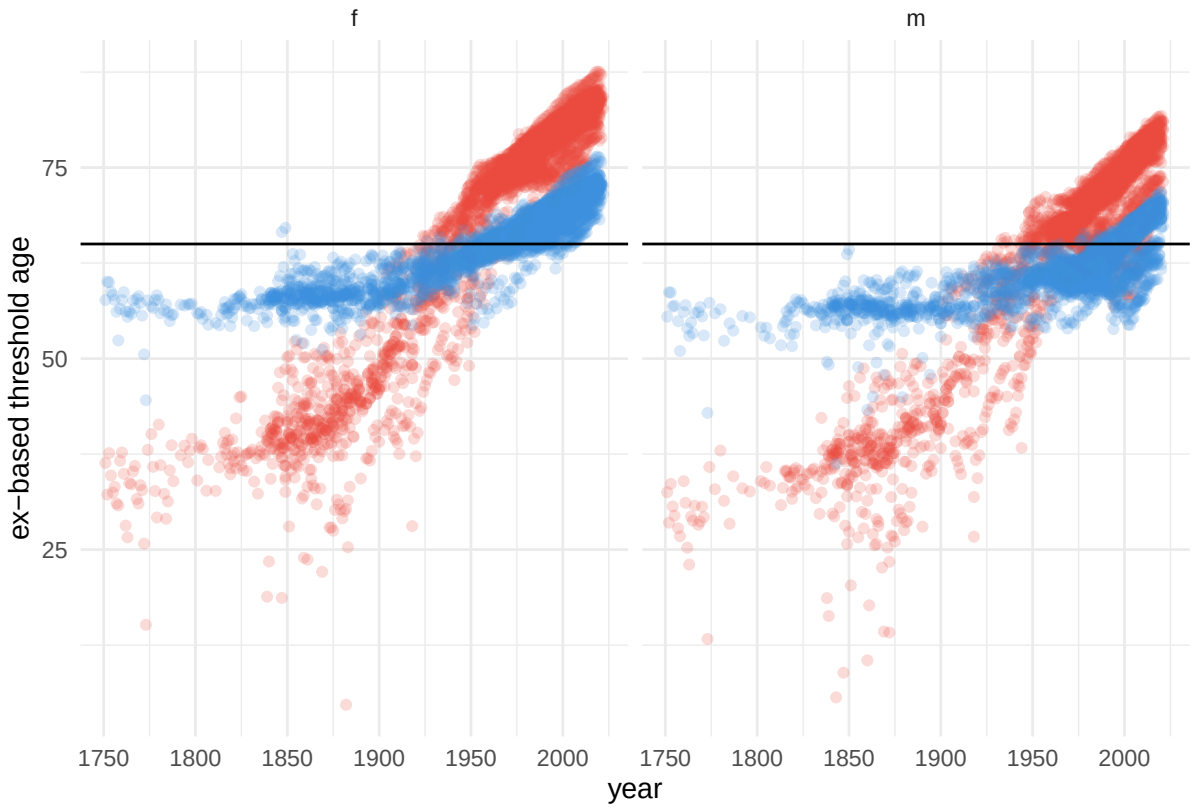


Now we can apply these in bulk. First let's compare how $e(0)$ compares to the two threshold ages:

```
t_ages <-
  hmd |>
  group_by(population,sex,year) |>
  summarize(e0 = ex[1],
            lx_age = calc_threshold_age(x = age, y = lx, .2, "lx"),
            ex_age = calc_threshold_age(x = age, y = ex, 15, "ex"),
            .groups = "drop")
```

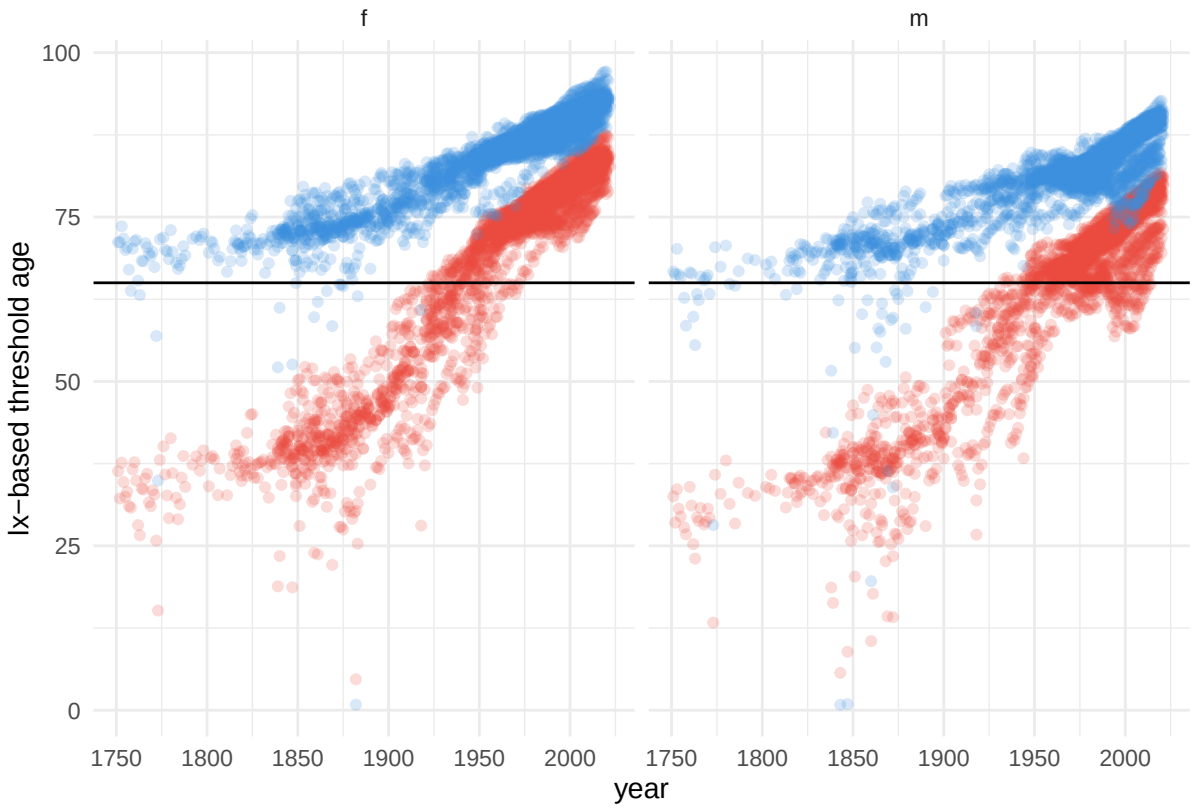
Here's how our results compare to life expectancy (red) at birth, and the age at which $e(x)$ crosses the threshold of 15 average remaining years of life (blue). Really, we're meant to compare with the horizontal line at age 65:

```
t_ages |>
  ggplot(aes(x = year, y = e0)) +
  geom_point(color = "#eb4b3f", alpha = .2) +
  geom_point(aes(x = year, y = ex_age), color = "#3c90de", alpha = .2) +
  facet_wrap(~sex) +
  theme_minimal() +
  geom_hline(yintercept = 65) +
  labs(y = "ex-based threshold age")
```



And here's the same comparison for the age at which only 20% of the synthetic cohort remains alive:

```
t_ages |>
  ggplot(aes(x = year, y = e0)) +
  geom_point(color = "#eb4b3f", alpha = .2) +
  geom_point(aes(x = year, y = lx_age), color = "#3c90de", alpha = .2) +
  facet_wrap(~sex) +
  theme_minimal() +
  geom_hline(yintercept = 65) +
  labs(y = "lx-based threshold age")
```



From this, the lesson is that while the common definition of age is fixed, the experience of age can change, in this case in response to background mortality, which itself is responding to background health. Notably, any dependency ratio statistic will vary greatly depending on which threshold age is chosen, and if said age is a dynamic response to mortality itself, both trends and levels can turn out far more optimistic. In exercises, we will try to calculate trends in different old age indicators.

2.2 Different measures imply different burdens

Rustam Tursun-zade has kindly prepared a data extract consisting in prevalence data for a very-lethal (Stomach cancer) and a not-lethal (Seborrhoeic dermatitis) condition, each for Kenya and South Korea, each by sex and for a time series from 1990-2019. We will use these data to calculate health expectancy and related trends.

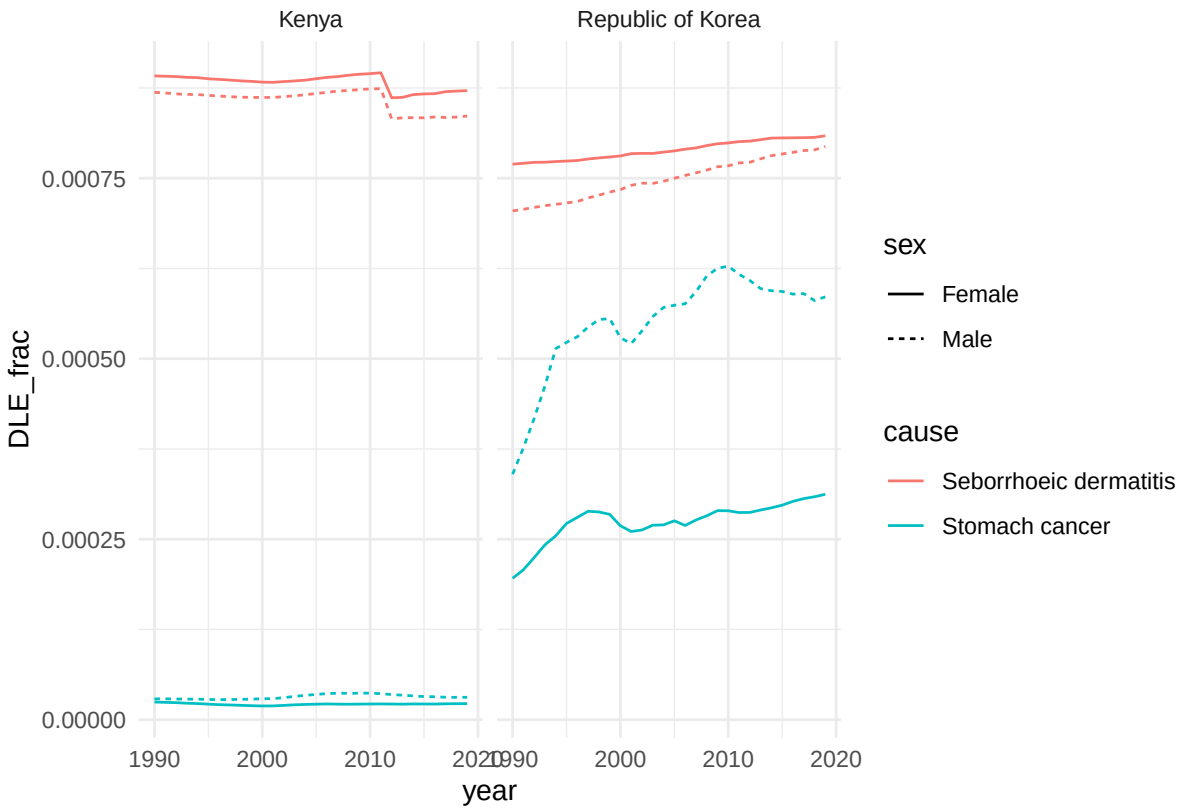
```
sull <- read_csv("Data/prev_mx.csv",
         show_col_types = FALSE) |>
  arrange(location, sex, year, cause, age) |>
  rename(nMx = mx) |>
  mutate(nAx = if_else(age == 0, .1, .5),
         n = 1) |>
  group_by(location, year, sex, cause) |>
  mutate(nqx = calc_nqx(nMx,nAx,n),
         lx = calc_lx(nqx,radix = 1),
         ndx = nqx * lx,
         nLx = calc_nLx(lx,ndx,nAx,n))
```

Sullivan's Sullivan (1971) method of calculating health expectancy takes prevalence and a lifetable as its basic inputs.

$$HLE(x) = \sum_{i=x}^{\omega} {}_nL_i/l_x \cdot {}_n\pi_i$$

Where π is prevalence, ranging from 0-1, and $L(x)$ comes from the lifetable. In R, it looks like so for this data (note we are calculating for age 0 with a radix of 1). We can plot the fraction of life expectancy taken up by the given condition:

```
sull |>
  group_by(location, sex, cause, year) |>
  summarize(DLE = sum(nLx * prev),
            HLE = sum(nLx * (1 - prev)),
            .groups= "drop") |>
  mutate(DLE_frac = DLE / (DLE + HLE)) |>
  ggplot(aes(x=year,y=DLE_frac, color = cause, linetype = sex)) +
  geom_line() +
  facet_wrap(~location) +
  theme_minimal()
```



In short, for consequential conditions (very lethal conditions), longer lives make little difference, whereas for non-lethal conditions (imagine conditions like minor vision impairment, minor hearing loss, mild disability, back pain) longer lives probably imply higher burden, all else equal.

References

Abel, Guy J. 2023. *Wcde: Download Data from the Wittgenstein Centre Human Capital Data Explorer*. https://guyabel.github.io/wcde/.

Alvarez, Jesús-Adrián, and James W Vaupel. 2023. "Mortality as a Function of Survival." *Demography* 60 (1): 327–42.

Bramajo, Octavio, Iñaki Permanyer, and Amand Blanes. 2023. "Regional Inequalities in Life Expectancy and Lifespan Variation by Educational Attainment in Spain, 2014–2018." *Population, Space and Place* 29 (3): e2628.

Caswell, Hal. 2023. "The Contributions of Stochastic Demography and Social Inequality to Lifespan Variability." *Demographic Research* 49: 309–54.

Demography, Wittgenstein Centre for, and Global Human Capital. 2018. *Wittgenstein Centre Human Capital Data Explorer*. http://dataexplorer.wittgensteincentre.org/wcde-v2/.

Sanderson, Warren C, and Sergei Scherbov. 2007. "A New Perspective on Population Aging." *Demographic Research* 16: 27–58.

Sullivan, Daniel F. 1971. "A Single Index of Mortality and Morbidity." *HSMHA Health Reports* 86 (4): 347.