Life tables, EDA, Mortality quotients

2024-09-02

```r
stopifnot(
  require(patchwork),
  require(glue),
  require(here),
  require(tidyverse),
  require(plotly),
  require(DT),
  require(GGally),
  require(ggforce),
  require(ggfortify)
)
tidymodels::tidymodels_prefer(quiet = TRUE)

old_theme <-theme_set(theme_minimal(base_size=9, base_family = "Helvetica"))
```

- **M1 MIDS/MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

! **Objectives**

Data sources

Life tables have been downloaded from https://www.mortality.org.

We investigate life tables describing countries from Western Europe (France, Great Britain –actually England and Wales–, Italy, the Netherlands, Spain, and Sweden) and the United States.

Life tables used here have been doctored and merged so as to simplify discussion.

We will use the next lookup tables to recode some factors.

```
country_code <- list(fr_t='FRATNP',
                     fr_c='FRACNP',
                     be='BEL',
                     gb_t='GBRTENW',
                     gb_c='GBRCENW',
                     nl='NLD',
                     it='ITA',
                     swe='SWE',
                     sp='ESP',
                     us='USA')

countries <- c('fr_t', 'gb_t', 'nl', 'it', 'sp', 'swe', 'us')

country_names <- list(fr_t='France',     # total population
                     fr_c='France',      # civilian population
                     be='Belgium',
                     gb_t='England & Wales',    # total population
                     gb_c='England & Wales',    # civilian population
                     nl='Netherlands',
                     it='Italy',
                     swe='Sweden',
                     sp='Spain',
                     us='USA')

gender_names <- list('b'='Both',
                     'f'='Female',
                     'm'='Male')
```

```
datafile <- 'full_life_table.Rds'
fpath <- stringr::str_c("../DATA/", datafile) # here::here('DATA', datafile)   # check getwd()

if (! file.exists(fpath)) {
  download.file("https://stephane-v-boucheron.fr/data/full_life_table.Rds",
                fpath,
                mode="wb")
}

life_table <- readr::read_rds(fpath)
```

```
life_table <- life_table |>
  mutate(Country = as_factor(Country)) |>
  mutate(Country = fct_relevel(Country, "Spain", "Italy", "France", "England & Wales", "Nether
  mutate(Gender = as_factor(Gender))

life_table <- life_table |>
  mutate(Area = fct_collapse(Country,
                        SE = c("Spain", "Italy", "France"),
                        NE = c("England & Wales", "Netherlands", "Sweden"),
                        USA="USA"))
```

Document Tables de mortalité françaises pour les XIXe et XXe siècles et projections pour le XXIe siècle contains detailed information on the construction of Life Tables for France.

Two kinds of Life Tables can be distinguished: *Table du moment* which contain for each calendar year, the mortality risks at different ages for that very year; and *Tables de génération* which contain for a given birthyear, the mortality risks at which an individual born during that year has been exposed.

The life tables investigated in this lab are *Table du moment.* According to the document by Vallin and Meslé, building the life tables required decisions and doctoring.

Have a look at Lexis diagram.

Definitions can be obtained from www.lifeexpectancy.org. We translate it into mathematical (rather than demographic) language. Recall that the quantities define a probability distribution over $\mathbb{N}$. This probability distribution is a *construction* that reflects the health situation in a population at a given time (year). This probability distribution does not describe the sequence of sanitary situations experienced by a *cohort* (people born during a specific year).

> One works with a period, or current, life table (*table du moment*). This summarizes the mortality experience of persons across all ages in a short period, typically one year or three years. More precisely, the death probabilities $q(x)$ for every age x are computed for that short period, often using census information gathered at regular intervals. These $q(x)$'s are then applied to a hypothetical cohort of 100000 people over their life span to produce a life table.

```
life_table |>
  filter(Country=='France', Year== 2010, Gender=='Female', Age < 10 | Age > 80 & Age <90) |>
  knitr::kable()
```

Year	Age	mx	qx	ax	lx	dx	Lx	Tx	ex	Country	Gender	Area
2010	0	0.00325	0.00324	0.14	100000	324	99722	8465207	84.65	France	Female	SE
2010	1	0.00032	0.00032	0.50	99676	32	99660	8365484	83.93	France	Female	SE
2010	2	0.00015	0.00015	0.50	99645	15	99637	8265824	82.95	France	Female	SE
2010	3	0.00011	0.00011	0.50	99630	11	99624	8166187	81.97	France	Female	SE
2010	4	0.00008	0.00008	0.50	99619	8	99615	8066563	80.97	France	Female	SE
2010	5	0.00005	0.00005	0.50	99611	5	99608	7966948	79.98	France	Female	SE
2010	6	0.00008	0.00008	0.50	99606	8	99602	7867339	78.98	France	Female	SE
2010	7	0.00008	0.00008	0.50	99598	8	99594	7767737	77.99	France	Female	SE
2010	8	0.00008	0.00008	0.50	99590	8	99586	7668143	77.00	France	Female	SE
2010	9	0.00007	0.00007	0.50	99582	7	99578	7568557	76.00	France	Female	SE
2010	81	0.03516	0.03455	0.50	73367	2535	72099	727802	9.92	France	Female	SE
2010	82	0.04059	0.03978	0.50	70832	2818	69423	655702	9.26	France	Female	SE
2010	83	0.04754	0.04644	0.50	68014	3158	66435	586280	8.62	France	Female	SE
2010	84	0.05536	0.05386	0.50	64856	3493	63109	519845	8.02	France	Female	SE
2010	85	0.06295	0.06103	0.50	61362	3745	59490	456736	7.44	France	Female	SE

Year	Age	mx	qx	ax	lx	dx	Lx	Tx	ex	Country	Gender	Area
2010	86	0.07246	0.06993	0.50	57617	4029	55603	397246	6.89	France	Female	SE
2010	87	0.08256	0.07929	0.50	53588	4249	51464	341643	6.38	France	Female	SE
2010	88	0.09660	0.09215	0.50	49339	4547	47066	290180	5.88	France	Female	SE
2010	89	0.11088	0.10505	0.50	44792	4706	42440	243114	5.43	France	Female	SE

> Check on http://www.mortality.org the meaning of the different columns.

In the sequel, we denote by F_t the *cumulative distribution function* for year t. $F_t(x)$ represents the *probability* of dying at age not larger than x.

We agree on $\overline{F}_t = 1 - F_t$ and $F_t(-1) = 0$.

The life tables are highly redundant. Provided we get the right conventions we can derive almost all columns from column `qx`.

```r
life_table |>
  filter( Year>=1948) |>
  group_by(Country, Year, Gender) |>
  summarise(m1 =max(abs(lx -dx -lead(lx)), na.rm = T),
            m2 =max(abs(lx * qx -dx), na.rm=T),
            m3 =max(abs(Lx -lx * (1 + qx * (ax-1))), na.rm=T),
            m4 =max(abs(1-exp(-mx)-qx), na.rm=T)) |>
  select(Year, Country, Gender, m1, m2, m3, m4) |>
  ungroup() |>
  group_by(Country, Gender) |>
  slice_max(order_by = desc(m4), n = 1)
```

```
`summarise()` has grouped output by 'Country', 'Year'. You can override using
the `.groups` argument.
```

```
# A tibble: 21 x 7
# Groups:   Country, Gender [21]
    Year Country         Gender    m1    m2    m3      m4
   <int> <fct>           <fct>  <int> <dbl> <dbl>   <dbl>
 1  1948 Spain           Both       1 0.874 2.20  0.00838
 2  1948 Spain           Female     1 0.789 1.56  0.00816
 3  1952 Spain           Male       1 0.802 5.5   0.0119
 4  2004 Italy           Both       1 0.836 0.968 0.0150
 5  2004 Italy           Female     1 0.875 1.03  0.0149
 6  1984 Italy           Male       1 0.774 5.56  0.0146
 7  2007 France          Both       1 0.887 0.976 0.0152
 8  2007 France          Female     1 0.890 0.980 0.0151
 9  1979 France          Male       1 0.764 4.97  0.0161
10  1992 England & Wales Both       1 0.898 2.42  0.0135
# i 11 more rows
```

`qx` (age-specific) risk of death at age x, or *mortality quotient* at given age x for given year t: $q_{t,x} = \frac{\overline{F}_t(x) - \overline{F}_t(x+1)}{\overline{F}_t(x)}$.
For each year, each age, $q_{t,x}$ is determined by data from that year.

We also have

$$\overline{F}_t(x+1) = \overline{F}_t(x) \times (1 - q_{t,x+1}) \, .$$

`mx` *central death rate* at age x during year t. This is connected with $q_{t,x}$ by

$$m_{t,x} = -\log(1 - q_{t,x}) \, ,$$

or equivalently $q_{t,x} = 1 - \exp(-m_{t,x})$.

`lx` the so-called *survival function*: the scaled proportion of persons alive at age x. These values are computed recursively from the $q_{t,x}$ values using the formula

$$l_t(x+1) = l_t(x) \times (1 - q_{t,x}) \, ,$$

with $l_{t,0}$, the "radix" of the table, arbitrarily set to 100000. Function $l_{t,\cdot}$ and $\overline{F}_t$ are connected by

$$l_{t,x+1} = l_{t,0} \times \overline{F}_t(x) \, .$$

Note that in Probability theory, $\overline{F}$ is also called the survival or tail function.

`dx` $d_{t,x} = q_{t,x} \times l_{t,x}$

`Tx` Total number of person-years lived by the cohort from age x to $x+1$. This is the sum of the years lived by the $l_{t,x+1}$ persons who survive the interval, and the $d_{t,x}$ persons who die during the interval. The former contribute exactly 1 year each, while the latter contribute, on average, approximately half a year, so that $L_{t,x} = l_{t,x+1} + 0.5 \times d_{t,x}$. This approximation assumes that deaths occur, on average, half way in the age interval x to x+1. Such is satisfactory except at age 0 and the oldest age, where other approximations are often used; *We will stick to a simplified vision $L_{t,x} = l_{t,x+1}$*

`ex:` Residual Life Expectancy at age x and year t

See: *Demography: Measuring and Modeling Population Processes* by SH Preston, P Heuveline, and M Guillot. Blackwell. Oxford. 2001.

- Chapter 2, on *Age-specific rates and Probabilities.* The comparison between Sweden and Kazakhstan illustrates the distinction between *crude death rates* and *age-specific death rates*, as well as the dependence of *crude death rates* on the age structure/distribution of the population. Moreover the Sweeden/Kazakhstan comparison offers a clear-cut example of the Yule-Simpson paradox.

- Chapter contains an important discussion of *age standardization* for cross country comparisons, why it matters, why it is difficult and remains a matter of taste. The definitions of *Life Expectancy at Birth*, or *Residual Life Expectancies* are example of age standardizations.

Western countries in 1948

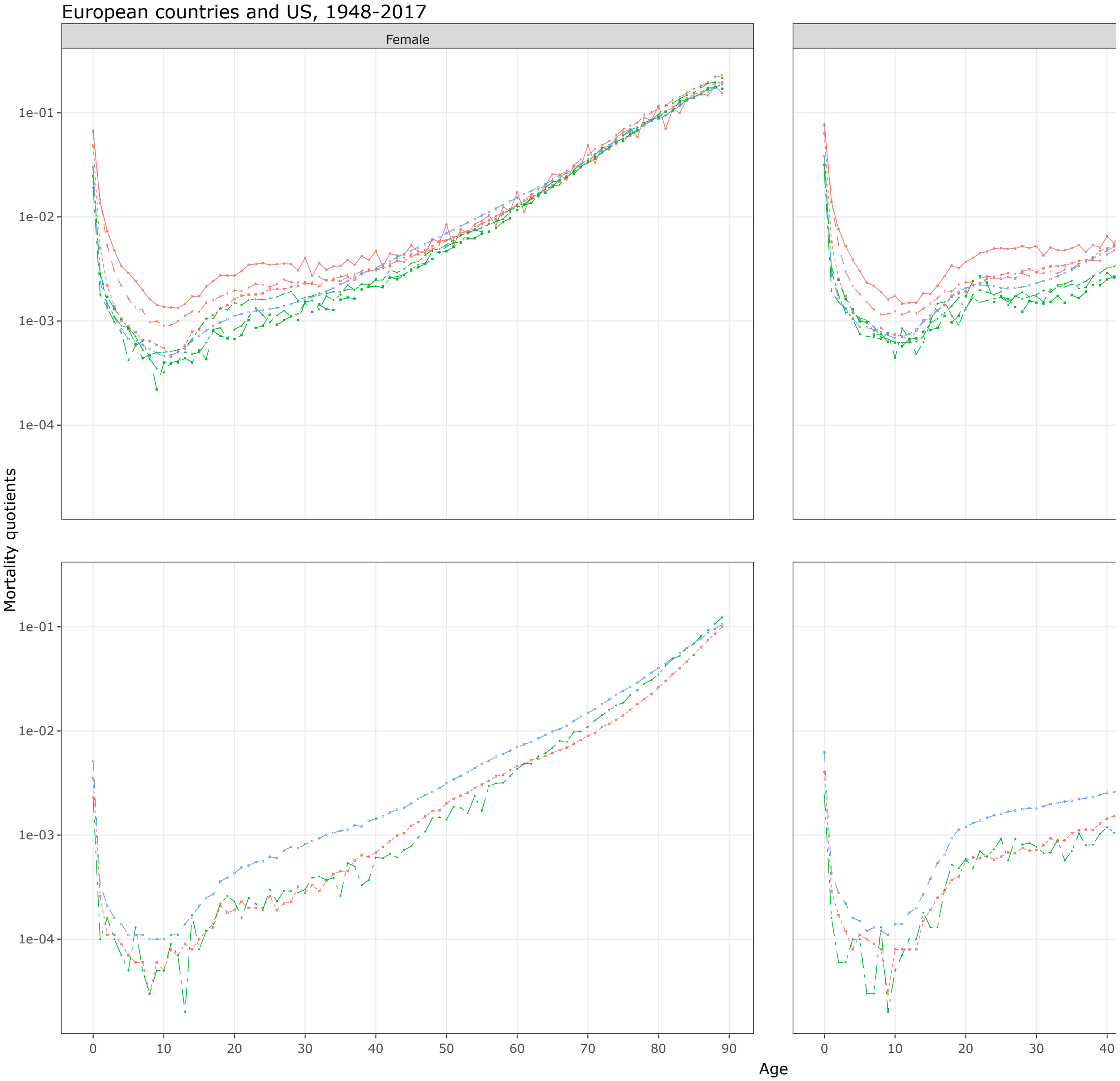
Several pictures share a common canvas: we plot mortality quotientss against ages using a logarithmic scale on the y axis. Countries are identified by aesthetics (shape, color, linetypes). Abiding to the DRY principle, we define a prototype `ggplot` (alternatively `plotly`) object. The prototype will be fed with different datasets and decorated and arranged for the different figures.

```
dummy_data <- dplyr::filter(life_table, FALSE)

proto_plot <- ggplot(dummy_data,
                     aes(x=Age,
                         y=qx,
                         col=Area,
                         linetype=Country,
                         shape=Country)) +
              scale_y_log10() +
              scale_x_continuous(breaks = c(seq(0, 100, 10), 109)) +
              ylab("Mortality quotients") +
              labs(linetype="Country") +
              theme_bw()
```

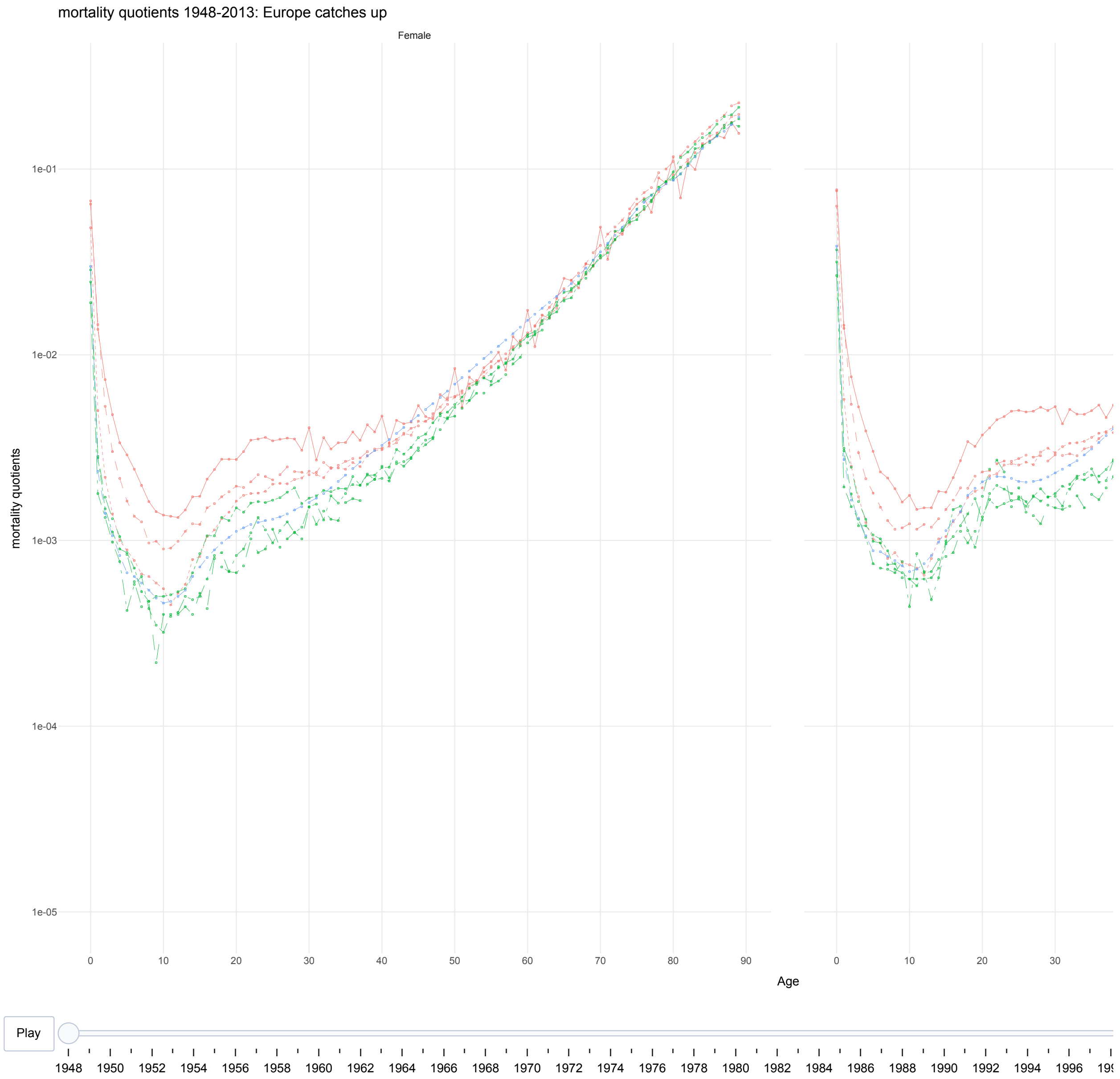
Question

Plot `qx` of all Countries at all ages for years 1948 and 2013.



```r
proto_plt2 <-
  ggplot() +
  aes(x=Age, y=qx, colour=Area, frame=Year, linetype=Country) +
  geom_point(size=.1) +
  geom_line(size=.1) +
  scale_y_log10() +
  labs(linetype=c("Country")) +
  scale_x_continuous(breaks = c(seq(0, 100, 10), 109)) +
  xlab("Age") +
  ylab("mortality quotients") +
  facet_grid(cols=vars(Gender))
```

```r
with(params,
(proto_plt2 %+%
  (life_table |> filter(between(Year, year_p, year_e), Gender != 'Both', Age < 90))  +
  ggtitle("mortality quotients 1948-2013: Europe catches up"))) |>
  plotly::ggplotly()
```



ℹ The animated plot allows to spot more details. It is useful to use color so as to distinguish threee areas: USA; Northern Europe (NE) comprising England and Wales, the Netherlands, and Sweden; Southern Europe (SE) comprising Spain, Italy, and France. In 1948, NE and the USA exhibit comparable central death reates at all ages for the two genders, the USA looking like a more dangerous place for young adults. Spain lags behind, Italy and Frane showing up at intermediate positions.
By year 1962, SE has almost caught up the USA. Italy and Spain still have higher infant mortality while mortality quotients in the USA and France are almost identical at all ages for both genders. mortality quotients attain a minimum around 10-12 for both genders. In Spain the minium central death rate has been divided by almost ten between 1948 and 1962.
If we dig further we observe that the shape of the male mortality quotients curve changes over time. In 1962, in the USA and France, mortality quotients exhibit a sharp increase between years 12 and 18, then remain almost constant between 20 and 30 and afterwards increase again. This pattern shows up in other countries but in a less spectacular way.
Twenty years afterwards, during years 1980-1985, death rates at age 0 have decreased at around 1% in all countries while it was 7% in Spain in 1948. The male central death curve exhibits a plateau between ages 20 and 30. mortality quotients at this age look higher in France and the USA.
By year 2000, France is back amongst European countries (at least with respect to mortality quotients). Young adult mortality rates are higher in the USA than in Europe. This phenomenon became more pregnant during the last decade.

ℹ **Question**

Plot ratios between mortality quotients (`qx`) in European countries and mortality quotients in the USA in 1948.

```
simplified_life_table <- with(params,
                              life_table |>
  dplyr::filter(between(Year, year_p, year_e), Age<90, Gender!="Both") |>
  dplyr::select(Age, Year, Country, qx, Gender, Area))

eur_table <- simplified_life_table |>
  dplyr::filter(Country!='USA')

us_table <- simplified_life_table |>
  dplyr::filter(Country=='USA') |>
  dplyr::select(-Area, -Country)

eur_us_table <-  eur_table |>
  dplyr::inner_join(us_table, by=c('Age', 'Year', 'Gender')) |>
  dplyr::mutate(Ratio=qx.x/qx.y)
```

```
with(params,
(eur_us_table  |>
```

```r
  ggplot(aes(x=Age,
             y=Ratio,
             col=Area,
             frame=Year,
             linetype=Country)) +
  scale_y_log10() +
  scale_x_continuous(breaks = c(seq(0, 100, 10), 109)) +
  geom_point(size=.1) +
  geom_smooth(method="loess", se=FALSE, span=.1, size=.1) +
  ylab("Ratio of mortality quotients with respect to US") +
  labs(linetype="Country", color="Area") +
  ggtitle(label = stringr::str_c("European countries with respect to US,", year_p,'-', year_e,
  facet_grid(rows = vars(Gender))) |>
  ggplotly()
)
```

This animation reveals less than the preceding one since we just have ratios with respect to the USA. But the patterns followed by European societies emerge in a more transparent way. The divide between northern and southern Europe at the onset of the period is even more visible. The ratios are important across the continent: there is a factor of 10 between spanish and swedish infant mortality rates. But the ratios at ages 50 and above tend to be similar. By the early 60s, the gap between southern and northern Europe has shrinked. By now, the ratios between mortality quotients tend to be within a factor of 2 across all ages, and even less at ages 50 and above.

Death rates evolution since WW II

Question

Plot mortality quotients (column `qx`) for both genders as a function of `Age` for years `1946, 1956, ...` up to `2016` .

```r
post_ww_II <- with(params, seq(year_p, year_e, 10))

p <- life_table |>
  filter(FALSE) |>
  ggplot(aes(x=Age,
             y=qx,
             col=forcats::as_factor(Year),
             linetype=forcats::as_factor(Year))) +
  geom_smooth(se=FALSE, method="loess", span= .1, size=.2) +
  labs(colour="Year", linetype="Year")   +
  scale_y_log10() +
  facet_grid(rows=vars(Country), cols=vars(Gender))
```

```
(p  %+%
  (life_table |> dplyr::filter(Year %in% post_ww_II, Gender!="Both",
                                Age < 90,
                                Country %in% c('Spain', 'USA'))) +
  ggtitle(stringr::str_c("Mortality quotient per Age", sep=", "),
          subtitle = "Post WW II")) |>
  ggplotly()
```

Question

Write a function `ratio_mortality_rates` with signature `function(df, reference_year=1946, target_years=seq(1946, 2016, 10))` that takes as input:

- a dataframe with the same schema as `life_table`,
- a reference year `ref_year` and
- a sequence of years `target_years`

and that returns a dataframe with schema:

Column Name	Column Type
Year	integer
Age	integer
qx	double
qx.ref_year	double
Country	factor
Gender	factor

where `(Country, Year, Age, Gender)` serves as a *primary key*, `mx` denotes the central death rate at `Age` for `Year` and `Gender` in `Country` whereas `qx_ref_year` denotes mortality quatient at `Age` for argument `reference_year` in `Country` for `Gender`.

```
ratio_mortality_rates <- function(df,
                                  reference_year=1946,
                                  target_years=seq(1946, 2016, 10)){
  dplyr::filter(df, Year %in% target_years, Age <90) |>
  dplyr::select("Age", "Area", "Gender", "Country", "qx", "Year") |>
  dplyr::inner_join(y=df[df$Year==reference_year,
                         c("Age", "Gender", "Country", "qx")],
                    by=c("Age", "Gender", "Country"))
}
```

Question

Draw plots displaying the ratio $m_{x,t}/m_{x,1946}$ for ages $x \in 1, \ldots, 90$ and year t for $t \in 1946, \ldots, 2016$ where $m_{x,t}$ is the central death rate at age x during year t.
Handle both genders and countries `Spain`, `Italy`, `France`, `England & Wales`, `USA`, `Sweden`, `Netherlands`.

One properly facetted plot is enough.

```
df_ratios <- ratio_mortality_rates(filter(life_table,
                                          Gender!="Both"),
                                   reference_year=1948,
                                   target_years=seq(1948, 2013, 1))
```

```
q <- df_ratios |>
  dplyr::filter(FALSE) |>
  ggplot(aes(x=Age,
             y=qx.x/qx.y,
             linetype=forcats::as_factor(Year),
             col=forcats::as_factor(Year))) +
  geom_smooth(method="loess",
              se= FALSE,
              size =.2,
              span= .1) +
  scale_y_log10() +
  ylab("Ratio of mortality rates, reference Year 1946") +
  labs(linetype="Year", col="Year") +
  scale_colour_brewer()
```

```
qf <- df_ratios |>
#  dplyr::filter(FALSE) |>
  ggplot(aes(x=Age,
             y=qx.x/qx.y,
             linetype=Country,
             frame=Year,
             col=Area)) +
  geom_smooth(method="loess",
              se= FALSE,
              size =.2,
              span= .1) +
  scale_y_log10() +
  scale_x_continuous(breaks = c(seq(0, 100, 10), 109)) +
  ylab("Ratio of mortality rates, reference Year 1946") +
  labs(linetype="Country") +
  facet_grid(rows=vars(Gender))
```

```
qf |> ggplotly()
```

Comment. During the last seventy years, death rates decreased at all ages in all seven countries. This progress has not been uniform across ages, genders and countries. Across most countries, infant mortality dramatically improved during the first post-war decade while death rates at age 50 and above remained stable until the mid seventies.

Trends

We noticed that mortality quotients did not evolve in the same way across all ages: first, the decay has been much more significant at low ages; second, the decay of mortality quotients at old ages (above 60) mostly took place during the last four decades. It is worth digging separately at what happened for different parts of life.

> **Question**
>
> Plot mortality quotients at ages $0, 1, 5$ as a function of time. Facet by Gender and Country

```r
ages <- c(0, 1, 5)

p_children <- filter(life_table, FALSE) |>
  ggplot(mapping=aes(x=Year, y=qx,
                     linetype=forcats::as_factor(Age),
                     shape=forcats::as_factor(Age),
                     col=forcats::as_factor(Age))) +
  geom_line(size=.2) +
  labs(linetype="Age", col="Age", shape="Age") +
  scale_y_log10() +
  scale_x_continuous(breaks=seq(1935,2010,5)) +
  facet_grid(cols=vars(Gender), rows=vars(Country))

p_children %+%
  filter(life_table,
          Age %in% ages,
          Gender != "Both",
          Year %in% 1933:2013) +
  ggtitle("Infant and child, mortality rate",
          subtitle = "Hygiene, Vaccination, Antibiotics")
```

All European countries achieved the same infant mortality rates after year 2000. The USA now lag behind.

During years 1940-1945, in the Netherlands and France, gains obtained before 1940 were reversed. Year 1945 was particularly difficult in the Netherlands.

> **Question**
>
> Plot mortality quotients at ages $15, 20, 40, 60$ as a function of time. Facet by `Gender` and `Country`

Life expectancy

Write a function that takes as input a vector of mortality quotients, as well as an age, and returns the residual life expectancy corresponding to the vector and the given age.

- Write a function that takes as input a dataframe with the same schema as `life_table` and returns a data frame with columns `Country`, `Gender`, `Year`, `Age` defining a primary key and a column `res_lex` containing *residual life expectancy* corresponding to the pimary key.

The next window function suffices to compute life expectancy at birth. It computes the logarithm of survival probabilities for each `Country`, `Year`, `Gender` (partition) at each `Age`. Note that the expression mentions an aggregation function `sum` and that the correction of the result is ensured by a correct design of the `frame` argument.

In order to compute Residual Life Expectancies at all ages, instead of performing aggregation, we compute a second window function. For each `Year`, `Country`, `Gender` (defining the partition), for each `Age`, the `Residual Life Expectancy` is the sum of survival probabilities over the `frame` defined by the current `Age` and all ages above.

Departing from the official method for computing residual life expectancy, we use the simplified recursion

$$e_{t,x} = (1 - q_{t,x}) \times (1 + e_{t,x+1}) \, .$$

That is, we assume that people dying between age x (included) and $x+1$ (non-included) die exactly on their x^{th} birthday. The official calculation assume that except at age 0 and great age, people die uniformly at random between age x and $x+1$:

$$e_{t,x} = (1 - q_{t,x}) \times (1 + e_{t,x+1}) + \frac{1}{2} \times q_{t,x}$$

Question

This recursion suggests a more efficient to compute *residual life expectancies* at all ages.

Question

Plot residual life expectancy as a function of `Year` at ages 60 and 65, facet by `Gender` and `Country`.

R4Data Science Tidy