KOSTAT-UNFPA Summer Seminar on Population

Workshop 1. Demography in R

Day 8: Advanced processing and visualization

Instructor: Tim Riffe
tim.riffe@ehu.eus

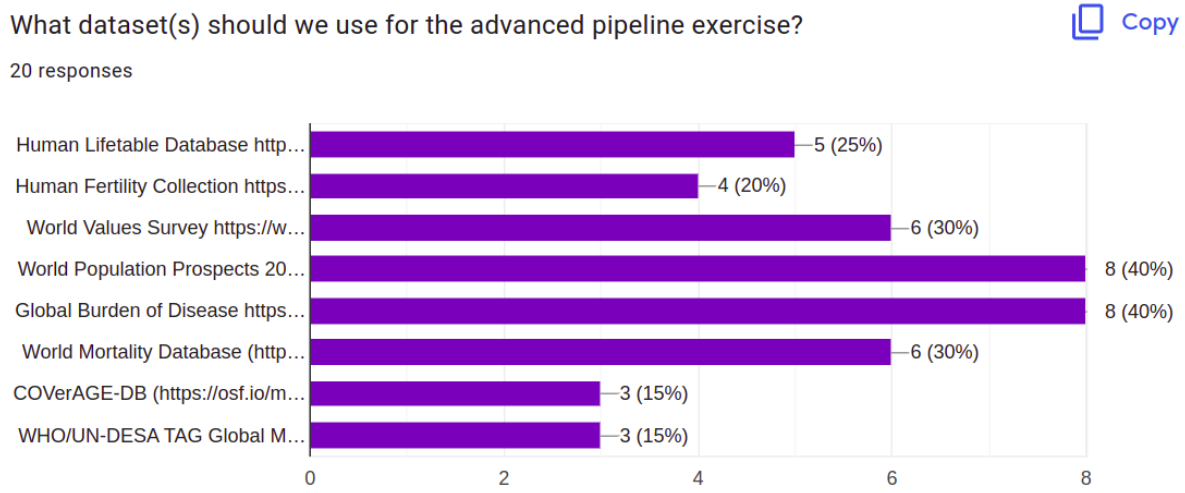
Assistants:
Jinyeon Jo: jyjo43043@gmail.com
Rustam Tursun-Zade: rustam.tursunzade@gmail.com

5 August 2022

Contents

1 Summary

In this lesson we try our hand at a spontaneously designed analytic exercise, making use of datasets chosen by participants in prior days. I suggested eight different *global* or *almost*-global datasets that we might tackle, and put it up to a vote. Here were the results:



Based on this, I decided to use both World Population Prospects 2022 (United Nations Population Division (2022)) data and an extract from the 2019 Global Burden of Disease (Global Burden of Disease Collaborative Network (2020)). For the WPP we download a csv file manually, and for the GBD I made a selection from the online results tool, which generated 60 zip files, which I pre-process some in a supplement to this handout called `08_data_prep.pdf`. I shared the results of that processing in a file called `gbd-share.csv.gz`, which we read in below.

The objective of this exercise we to combine lifetables from WPP with prevalence estimates of various impairments from GBD and to use these to calculate different kinds of impairment and impairment-free life expectancy using the so-called Sullivan Method Sullivan (1971) .

Most tools used have already been covered in the workshop, this is an agility exercise as much as anything, meant to drive home key concepts from the workshop. However, since this is a fresh exercise, I will reveal the bulk of the process, and indeed do most of the deriving as we go. With the clock ticking since we only have 1.5 hours. How did I choose to do things one way rather than another, and why do things in a particular order? This is important to convey because when you bring these tools to your own work, we sometimes get stuck, either because things don't work on the first try, or because we're unsure of process design aspects. Therefore, we will spend some time in this lesson charting out the steps in advance of actually writing them.

The code below is pasted from the session script (unlike the other handouts) due to its spontaneous nature.

1.1 load dependencies

```
library(tidyverse)
library(vroom)
library(janitor)
library(countrycode)
```

1.2 Read in WPP 2022 abridged lifetables

This file contains both countries, territories, and aggregates thereof, as well as different kinds of geocodes. We have abridged-age lifetables for males, females and the total population. The

code below filters the file down to just those countries whose UN location ID is included the the `countrycode` package lookup table. This is just a cheap way to remove the regional aggregates. In the end, we'll use the ISO3 code for matching to GBD data anyway. Note we also divide out the lifetable radix from the `Lx` column, such that $e_0 = \sum L_x$. That just simplifies formulas for impairment-free life expectancy.

```
wpp <- read_csv("Data/wpplt.zip",
                show_col_types = FALSE)
wpp <-
  wpp %>%
  mutate(location = countrycode(LocID,
                                origin = "un",
                                destination = "country.name.en",
                                warn = FALSE)) %>%
  filter(!is.na(location)) %>%
  select(iso3 = ISO3_code,
         sex = Sex,
         year = Time,
         age = AgeGrpStart,
         mx, Lx, ex) %>%
  mutate(Lx = Lx / 100000)
```

2 Read in GBD

This was pre-processed in `06_data_prep.pdf`, have a look there if interested. Registration is free and easy for the GBD results tool.

```
gbd <- read_csv("Data/gbd_share.csv.gz",
                show_col_types = FALSE)

gbd <-
  gbd %>%
  mutate(iso3 = countrycode(location,
                            origin = "country.name.en",
                            destination = "iso3c")) %>%
  arrange(location, measure, sex, year, age)

object.size(gbd) %>% print(units = "Mb")
```

```
## 503.3 Mb
```

2.1 What measures do we have?

A few checks to decide what and how we calculate things. I see we have different severity breakdowns of a reduced set of potential impairments or deficiencies.

```
gbd %>% pull(measure) %>% unique()
```

```
##  [1] "Anemia"
##  [2] "Blindness"
##  [3] "Blindness and vision loss"
##  [4] "Borderline intellectual disability"
##  [5] "Complete hearing loss"
##  [6] "Developmental intellectual disability"
```

```
##  [7] "Epilepsy"
##  [8] "Guillain-Barré syndrome"
##  [9] "Hearing loss"
## [10] "Heart failure"
## [11] "Infertility"
## [12] "Mild anemia"
## [13] "Mild hearing loss"
## [14] "Mild heart failure"
## [15] "Mild intellectual disability"
## [16] "Moderate anemia"
## [17] "Moderate epilepsy"
## [18] "Moderate hearing loss"
## [19] "Moderate heart failure"
## [20] "Moderate intellectual disability"
## [21] "Moderate pelvic inflammatory disease"
## [22] "Moderate vision loss"
## [23] "Moderately severe hearing loss"
## [24] "Pelvic inflammatory disease"
## [25] "Presbyopia"
## [26] "Primary infertility"
## [27] "Profound hearing loss"
## [28] "Profound intellectual disability"
## [29] "Secondary infertility"
## [30] "Severe anemia"
## [31] "Severe epilepsy"
## [32] "Severe hearing loss"
## [33] "Severe heart failure"
## [34] "Severe intellectual disability"
## [35] "Severe pelvic inflammatory disease"
## [36] "Severe vision loss"
## [37] "Treated epilepsy"
## [38] "Treated heart failure"
```

2.2 Check sums

Is Anemia equal to the sum of mild and severe anemia? Let's do a check and see how this works so that we can do more intelligent calculations later

```
gbd %>%
  filter(iso3 == "AFG",
         year == 2019,
         sex == "Female",
         measure %in% c("Anemia", "Mild anemia", "Moderate anemia", "Severe anemia")) %>%
  pivot_wider(names_from = measure, values_from = prev) %>%
  mutate(check_sum = `Mild anemia` + `Moderate anemia` + `Severe anemia`) %>%
  select(age, Anemia, check_sum)
```

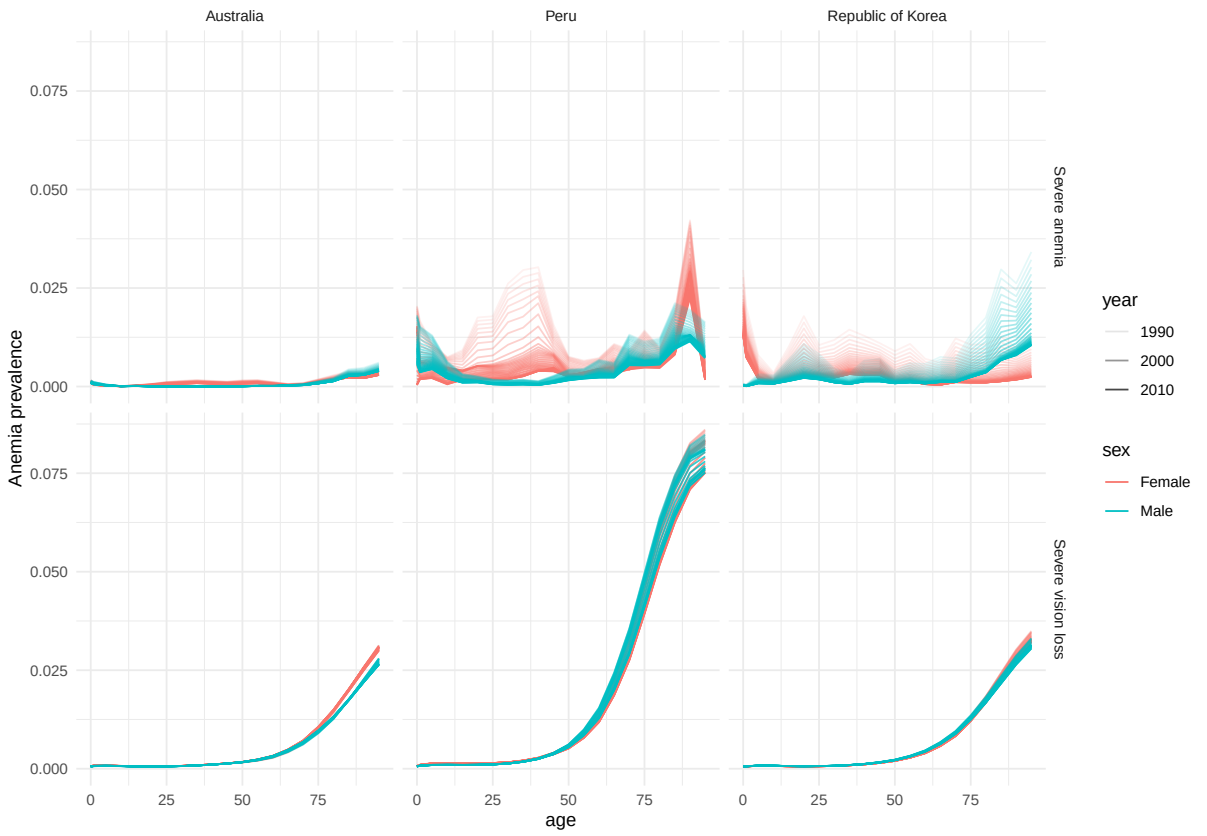
```
## # A tibble: 21 x 3
##      age Anemia check_sum
##    <dbl>  <dbl>     <dbl>
##  1     0  0.529     0.529
##  2     1  0.374     0.374
##  3     5  0.261     0.261
```

```
##  4    10  0.175     0.175
##  5    15  0.194     0.194
##  6    20  0.194     0.194
##  7    25  0.199     0.199
##  8    30  0.207     0.207
##  9    35  0.227     0.227
## 10    40  0.234     0.234
## # ... with 11 more rows
## # i Use `print(n = ...)` to see more rows
```

Yes, the basic illness indicators are the sum of the severity breakdowns.

2.3 examine some subsets

```
gbd %>%
  filter(iso3 %in% c("AUS","KOR","PER"),
         measure %in% c("Severe anemia", "Severe vision loss")) %>%
  ggplot(mapping = aes(x = age,
                       y = prev,
                       color = sex,
                       alpha = year,
                       group = interaction(year, sex))) +
  geom_line() +
  # new! gridded facetting, note prev measures are in rows
  # and locations in columns
  facet_grid(vars(measure),
             vars(location)) +
  labs(y = "Anemia prevalence") +
  theme_minimal()
```



2.4 Test merge

```
wpp_test <- wpp %>%
  filter(iso3 == "KOR")

gbd_test <- gbd %>%
  filter(iso3 == "KOR")

# success
left_join(gbd_test,
          wpp_test,
          by = c("iso3","sex","year","age"))
```

```
## # A tibble: 45,990 x 10
##    location          measure sex     year   age  prev iso3        mx    Lx    ex
##    <chr>             <chr>   <chr>  <dbl> <dbl> <dbl> <chr>    <dbl> <dbl> <dbl>
##  1 Republic of Korea Anemia  Female  1990     0 0.509 KOR   0.0124    0.989  76.4
##  2 Republic of Korea Anemia  Female  1990     1 0.335 KOR   0.000533 3.95   76.4
##  3 Republic of Korea Anemia  Female  1990     5 0.201 KOR   0.000580 4.92   72.5
##  4 Republic of Korea Anemia  Female  1990    10 0.125 KOR   0.000444 4.91   67.7
##  5 Republic of Korea Anemia  Female  1990    15 0.285 KOR   0.000616 4.90   62.9
##  6 Republic of Korea Anemia  Female  1990    20 0.342 KOR   0.000849 4.88   58.0
##  7 Republic of Korea Anemia  Female  1990    25 0.294 KOR   0.000879 4.86   53.3
##  8 Republic of Korea Anemia  Female  1990    30 0.321 KOR   0.00104  4.83   48.5
##  9 Republic of Korea Anemia  Female  1990    35 0.342 KOR   0.00161  4.80   43.7
## 10 Republic of Korea Anemia  Female  1990    40 0.333 KOR   0.00241  4.75   39.1
## # ... with 45,980 more rows
```

```
## # i Use `print(n = ...)` to see more rows
```

2.5 a little bit of lifetable rigor

Note, the lifetables close out at 100+, but the prevalence closes out at 95+, so we should do a little bit of adjustment to the lifetables to make these match.

```
wpp <-
  wpp %>%
  mutate(age = ifelse(age > 95, 95, age)) %>%
  group_by(iso3, sex, year, age) %>%
  summarize(mx = mx[1],
            Lx = sum(Lx),
            ex = ex[1],
            .groups = "drop")
```

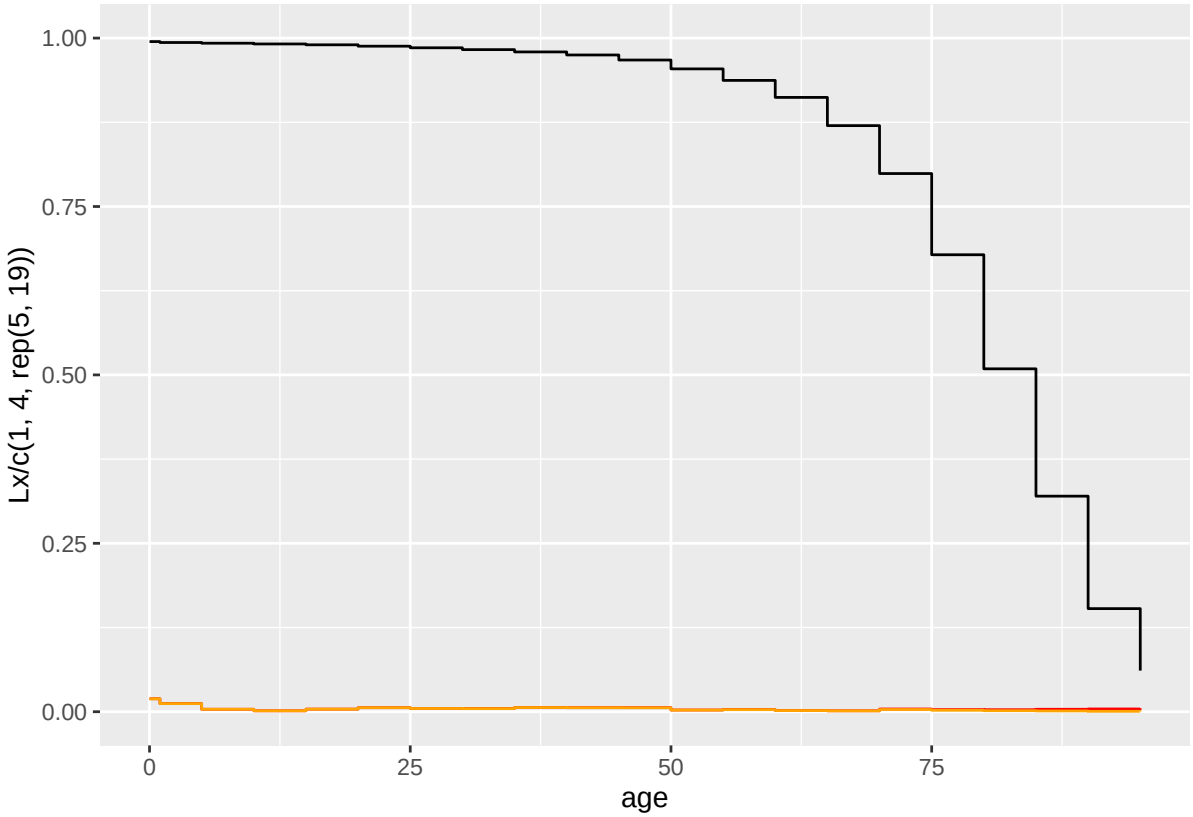
Let's redo an example join to demonstrate the Sullivan method.

```
wpp_test <-
  wpp %>%
  filter(year == 2000,
         sex == "Female",
         iso3 == "KOR")

gbd_test <-
  gbd %>%
  filter(year == 2000,
         sex == "Female",
         iso3 == "KOR",
         measure == "Severe anemia")
join_test <-
left_join(gbd_test,
          wpp_test,
          by = c("iso3", "sex", "year", "age"))
```

2.6 Demonstrate Sullivan

We have a survival step function, a red line for prevalence (conditional on survival), and and orange line showing the *lifetable burden* of the given condition. See how we divide out the interval width for L_x? That's just for the plot! We don't do it in the actual formula!

```
join_test %>%
  ggplot(mapping = aes(x = age,
                       y = Lx / c(1,4,rep(5,19)))) +
  geom_step() +
  ylim(0,1) +
  geom_step(mapping = aes(x = age,
                          y = prev),
            color = "red") +
  geom_step(mapping = aes(x = age,
                          y = Lx / c(1,4,rep(5,19)) * prev),
            color = "orange")
```



Actual HLE calculation is straightforward like so:

```
join_test %>%
  summarize(Anemia_LE = sum(Lx * prev),
            Anemia_free_LE = ex[age == 0] - Anemia_LE,
            LE = ex[age == 0])
```

```
## # A tibble: 1 x 3
##   Anemia_LE Anemia_free_LE    LE
##       <dbl>          <dbl> <dbl>
## 1     0.366           79.8  80.2
```

2.7 Hearing loss

Finally, we select some countries just for the sake of comparison, then calculate the various hearing-loss related expectancies.

```
countrycode("Somalia", origin = "country.name.en", destination = "iso3c")
```

```
## [1] "SOM"
```

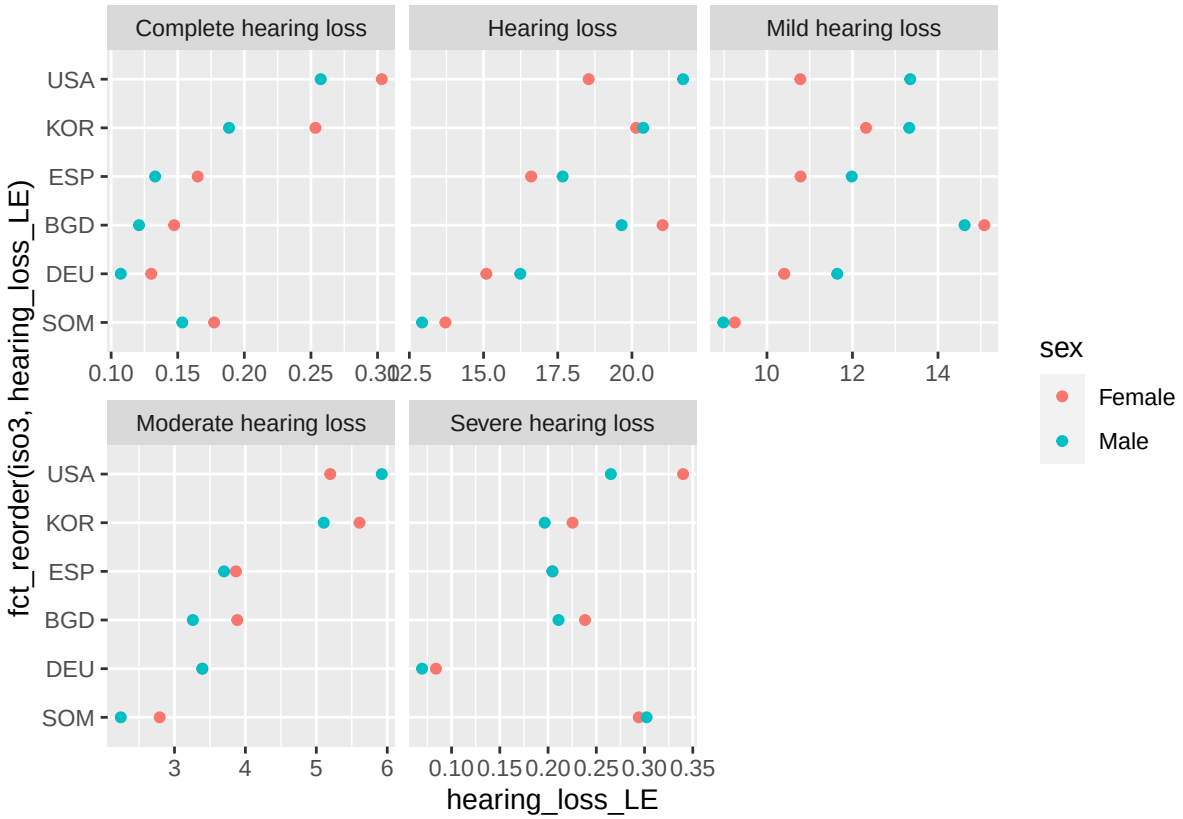
```
codes <- c("ESP","USA","DEU","KOR","BGD","SOM")
measures <- c("Hearing loss","Complete hearing loss","Mild hearing loss","Moderate hearing l

# all(measures %in% gbd$measure)

hearing <-
gbd %>%
  filter(year == 2019,
         iso3 %in% codes,
```

```
        measure %in% measures) %>%
  left_join(wpp, by = c("iso3","year","sex","age")) %>%
  group_by(iso3, sex, measure) %>%
  summarize(hearing_loss_LE = sum(prev * Lx),
            better_hearing_LE = sum((1 - prev) * Lx),
            LE = sum(Lx),
            .groups = "drop")
```

2.8 visualize

And finally we visualize this as a dot-plot, including some value-sorting of the country-rows using the handy `fct_reorder()` function (loads with `tidyverse`).

```
hearing %>%
  ggplot(mapping = aes(x = hearing_loss_LE,
                       y = fct_reorder(iso3, hearing_loss_LE),
                       color = sex)) +
  facet_wrap(~measure, scales = "free_x") +
  geom_point()
```



Had there been more time we would have also sorted the panels (currently alphabetical!) according to severity.

References

Global Burden of Disease Collaborative Network. 2020. "Global Burden of Disease Study 2019 (GBD 2019) Results."

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

United Nations Population Division. 2022. *World Population Prospects 2022*. http://population.un.org/wpp.