Data tidying

Gabriel Montes

2025-07-17

1. For each of the sample tables, describe what each observation and each column represents.

in table 1 each row represents a unique combination of country and year each column is a variable, country, year, cases, population

in table 2 each row represents a single measurement each column also are features of the measurement, country, year, type count

in table3 each row represents a country and year pair with the rate of TB cases each colum are the features, country year rate

2. Sketch out the process you'd use to calculate the rate for table2 and table3. You will need to perform four operations:

for table 2

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
table2_wide <- table2 |>
  pivot_wider(names_from = type, values_from = count)

table2_rate <- table2_wide |>
  mutate(rate = (cases / population) * 10000)
```

for table 3

```r
table3_separated <- table3 |>
  separate(rate, into = c("cases", "population"), sep = "/") |>
  mutate(
    cases = as.numeric(cases),
    population = as.numeric(population)
  )

table3_rate <- table3_separated |>
  mutate(rate = (cases / population) * 10000)
```