Babynames I

2024-09-02

```
require(patchwork)
require(httr)
require(glue)
require(ineq)
require(here)
require(skimr)
require(magrittr)
require(tidyverse)

old_theme <- theme_set(theme_minimal())
```

- **L3 MIASHS**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **!** **Objectives**

Naming babies

French data

The French data are built and made available by INSEE (French Governement Statistics Institute)

- https://www.insee.fr/fr/statistiques/fichier/2540004/nat2021_csv.zip

This dataset has been growing for a while. It has been considered by social scientists for decades. Given names are meant to give insights into a variety of phenomena, including religious observance.

A glimpse at that body of work can be found in *L'archipel français* by Jérome Fourquet, Le Seuil, 2019

Read the File documentation

```
path_data <- 'DATA'
fname <- 'nat2021_csv.zip'
fpath <- here(path_data, fname)

if (!file.exists(fpath)){
  url <- "https://www.insee.fr/fr/statistiques/fichier/2540004/nat2021_csv.zip"
  download.file(url, fpath, mode="wb")
}

df_fr <- readr::read_csv2(fpath)

# df_fr |> glimpse()
```

US data

US data may be gathered from

Baby Names USA from 1910 to 2021 (SSA)

See https://www.ssa.gov/oact/babynames/background.html

It can also be obtained by installing and loading the "babynames" package.

Full baby name data provided by the SSA. This includes all names with at least 5 uses.

```
if (!require("babynames")){
  install.packages("babynames")
  stopifnot(require("babynames"), "Couldn't install and load package 'babynames'")
}
```

```
?babynames
```

Tidy the French data

Rename columns according to the next lookup table:

```
lkp <- list(year="annais",
  sex="sexe",
  name="preusuel",
  n="nombre")
```

```
df_fr <- df_fr |>
  rename(!!!lkp) |>                                                    ①
  mutate(country='fr') |>
  mutate(sex=as_factor(sex)) |>
  mutate(sex=fct_recode(sex, "M"="1", "F"="2")) |>
  mutate(sex=fct_relevel(sex, "F", "M")) |>
  mutate(year=ifelse(year=="XXXX", NA, year)) |>
  mutate(year=as.integer(year))
```

```
df_fr  |>
  sample(5) |>
  glimpse()
```

① `!!!` (bang-bang-bang) is offered by `rlang` package. Here, we use it to perform *list unpacking* (with the same intent and purposes we use dictionary unpacking in Python)

```
Rows: 686,538
Columns: 5
$ name    <chr> "_PRENOMS_RARES", "_PRENOMS_RARES", "_PRENOMS_RARES", "_PRENOM~
$ country <chr> "fr", "fr", "fr", "fr", "fr", "fr", "fr", "fr", "fr", "fr", "f~
$ n       <dbl> 1249, 1342, 1330, 1286, 1430, 1472, 1451, 1514, 1509, 1526, 16~
$ sex     <fct> M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M,~
$ year    <int> 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 19~
```

Download 'Naissances totales par sexe' from URL `https://www.ined.fr/fichier/s_rubrique/168/t35.fr.xls` from INED.

```
births_fr_path <- here(path_data, 't35.fr.xls')
births_fr_url <- 'https://www.ined.fr/fichier/s_rubrique/168/t35.fr.xls'

if (!file.exists(births_fr_path)) {
  download.file(births_fr_url, births_fr_path)
}
```

```
births_fr <-  readxl::read_excel(births_fr_path, skip = 3)

births_fr <- births_fr[-1, ]


births_fr |>
  glimpse()
```

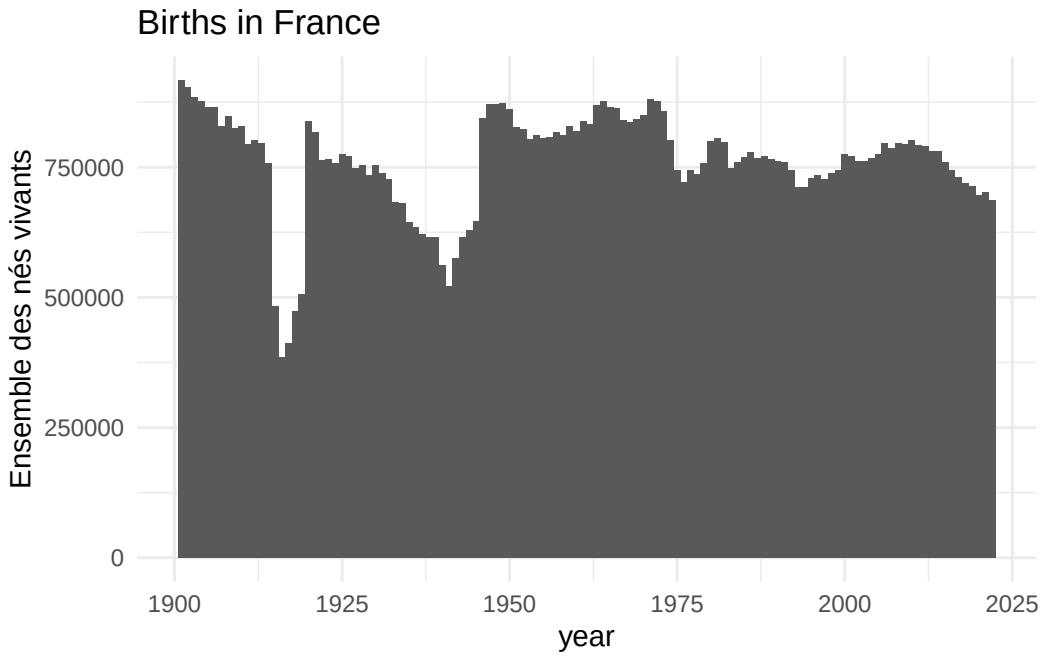
```
Rows: 130
Columns: 10
$ `Répartition par sexe et vie`                <chr> "1901", "1902", "1903", "~
$ `Ensemble des nés vivants`                   <dbl> 917075, 904434, 884498, 8~
$ `Nés vivants - Garçons`                      <dbl> 468125, 462097, 451510, 4~
$ `Nés vivants - Filles`                       <dbl> 448950, 442337, 432988, 4~
$ `Ensemble des enfants sans vie`              <dbl> 32410, 32000, 31076, 3067~
$ `Enfants sans vie - Garçons`                 <chr> "18522", "18172", "17875"~
$ `Enfants sans vie - Filles`                  <chr> "13888", "13828", "13201"~
$ `Garçons vivants pour 100 nés\nvivants`      <dbl> 51.0, 51.1, 51.0, 51.0, 5~
$ `Garçons vivants pour 100\nfilles vivantes`  <dbl> 104.3, 104.5, 104.3, 104.~
$ `Garçons sans vie pour 100\nfilles sans vie` <chr> "133.40000000000001", "13~
```

If you have problems with the excel reader, feel free to download an equivalent `csv` file from url

```r
names(births_fr)[1] <- "year"
```

```r
births_fr <- births_fr |>
  mutate(year=as.integer(year)) |>
  drop_na()
```

```r
births_fr |>
  ggplot() +
  aes(x=year, y=`Ensemble des nés vivants`) +
  geom_col() +
  labs(title="Births in France")
```



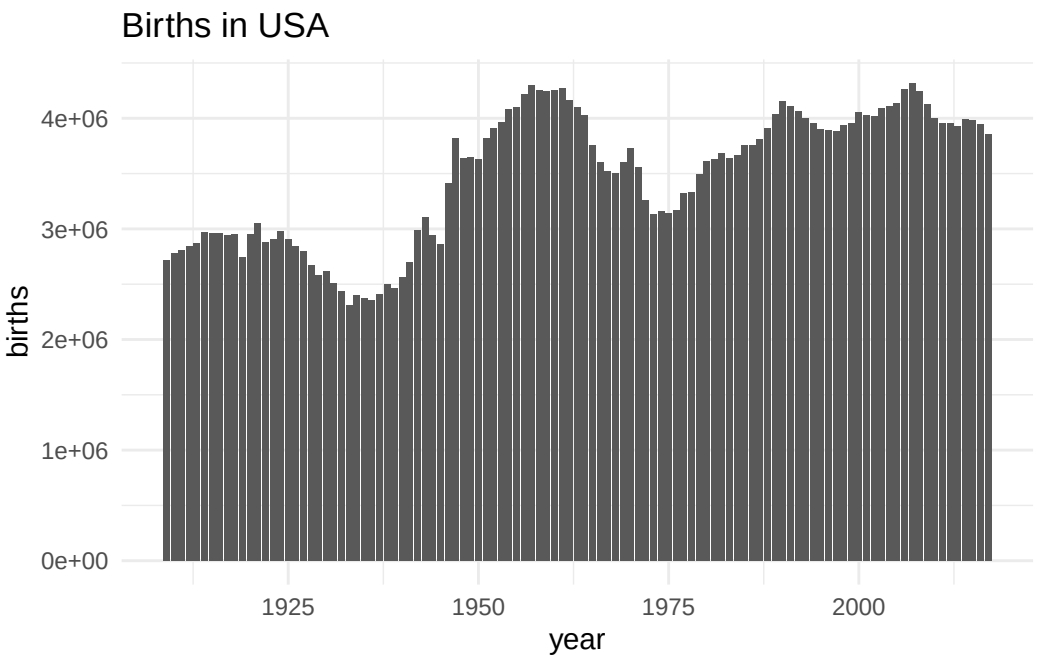
Tidy the American data

```r
babynames <- babynames |>
  mutate(country='us') |>
  mutate(sex=as_factor(sex))

babynames |>
  glimpse()
```

```
Rows: 1,924,665
Columns: 6
$ year    <dbl> 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 18~
$ sex     <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F,~
$ name    <chr> "Mary", "Anna", "Emma", "Elizabeth", "Minnie", "Margaret", "Id~
$ n       <int> 7065, 2604, 2003, 1939, 1746, 1578, 1472, 1414, 1320, 1288, 12~
$ prop    <dbl> 0.07238359, 0.02667896, 0.02052149, 0.01986579, 0.01788843, 0.~
```

```
$ country <chr> "us", "us", "us", "us", "us", "us", "us", "us", "us", "us", "u~
```

```r
births_us <- births

births_us  |>
  ggplot() +
  aes(x=year, y=births) +
  geom_col() +
  labs(title="Births in USA")
```



Sex ratios

> **Question**
>
> In dataset `df_fr` compute the total number of reported male and female births per year. Compute and plot the sex ratio.

> **Question**
>
> Compare with sex ratio as given in dataset from INED

> **Question**
>
> Consider the fluctuations of the sex ratio through the years.
> Are they consistent with the hypothesis: the sex of newborns are independently. identically distributed with the probability of getting a girl equal to .48?

Question

Consider again the fluctuations of the sex ratio through the years.
Assume that for each year the sex of newborns are independently. identically distributed with the probability of getting a girl depending on the year.
Are the data consistent with the hypothesis: the probability of getting a girl remains constant thoughout the years?

Picturing concentration of babynames distributions

Every year, in each country, for both sex, the name counts define a discrete probability distribution over the set of names (the universe).

This distribution, just as an income or wealth distribution, is (usually) far from being uniform. We want to assess how uneven it is.

We use the tools developed in econometrics.

Without loss of generality, we assume that we handle a distribution over positive integers $1, \dots, n$ where n is the number of distinct names given during a year.

We assume that frequencies $p_1, p_2, \dots, p_n$ are given in ascending order, ties are broken arbitrarily.

The `Lorenz function` (Lorenz not `Lorentz`) maps $[0,1] \to [0,1]$.

$$L(x) = \sum_{i=1}^{\lfloor nx \rfloor} p_i.$$

Note that this is a piecewise constant function.

Question

Compute and plot the Lorenz function for a given `sex`, `year` and `country`

Question

Design an animated plot that shows the evolution of the Lorenz curve of babynames distribution through the years for a given sex and country.

Inequality indices

The Lorenz curve summarizes how far a discrete probability distribution is from the uniform distribution. This is a very rich summary and it is difficult to communicate this message to a wide audience. People tend to favor numerical indices (they don't really understand, but they get used to it): Gini, Atkinson, Theil, …

The Gini index is twice the surface of the area comprised between curves $y = x$ and $y = L(x)$.

$$G = 2 \times \int_0^1 (x - L(x))\mathrm{d}x$$

The next formula allows us to compute it efficiently.

$$G = \frac{2\sum_{i=1}^{n} ip_i}{n\sum_{i=1}^{n} p_i} - \frac{n+1}{n}.$$

> **Question**
>
> Compute and plot Gini index of names distribution over time for sex and countries

`PRENOMS RARES` in France

> **Question**
>
> For each sex, Plot the proportion of births given `_PRENOMS_RARES` as a function of year.

> **Look for `Mary` in US Data**

Marie, Jeanne and France in France

> **Question**
>
> Plot the proportion of female births given name 'MARIE' or 'MARIE-...' as a function of year. Proceed in such a way that the reader can see the share of compounded names. We are expecting an *area plot*

Have a look at r-graph-gallery: stacked area and at ggplot documentation. Pay attention on the way you stack the area corresponding to names matching pattern 'MARIE-...' over or under the are corresponding to babies named 'MARIE'

> **Question**
>
> Answer the same question for JEANNE and FRANCE

Patterns of popularity

Question

Plot the popularities of `KEVIN`, `ENZO`, `STÉPHANE` as a function of `year`.

Question

Plot the popularities of “JEAN”, “LUC”, “MATHIEU”, “MARC”, “PAUL”, “PIERRE”, “JOSEPH”, “FRANÇOIS” as a function of `year`. Use stacked area style plot.

Question

Plot the popularities of “JEAN”, “LUC”, “MATHIEU”, “MARC”, “PAUL”, “PIERRE”, “JOSEPH”, “FRANÇOIS” as a function of `year`. Use line plot.

Question

Look for the translation of these names in US Data

Grouping names by patterns of popularity

Patterns of popularity

Fitting a Zipf distribution

Choosing scales

Animation

Classifying names according to their pattern of popularity

Now, we focus on names that made it to the top 300 at least once since year 1948. We attempt to classify them according to their pattern of popularity,