Babynames II: patterns of popularity

2024-09-02

```r
stopifnot(
  require(patchwork),
  require(httr),
  require(glue),
  require(ineq),
  require(here),
  require(slider),
  require(tidyverse),
  require(gtools)
)

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

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

- Moodle

> ❗ **Objectives**

Setup

```r
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)
```

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

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

```r
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)
}
```

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

births_fr <- births_fr[-1, ]
names(births_fr)[1] <- "year"

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

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

births_us <- births
```

```r
df <- bind_rows(babynames, df_fr)
```

```r
df <- df |>
  filter(year > 1947) |>
  drop_na() |>
  filter(name!='_PRENOMS_RARES')
```

```r
df <- df |>
  group_by(year, sex, country) |>
  arrange(desc(n), .by_group=T) |>
  mutate(rnk=row_number(),
         rrnk=rnk/n(),
         cprop=cumsum(prop)) |>
  ungroup()
```

```r
min_maj <- function(cprop, rrnk){
  1- rrnk[findInterval(.5, cprop)]
}
```

```r
last_dec <- function(cprop, rrnk) {
  cprop[findInterval(.1, rrnk)]
}
```

```r
ineq_idx_fns <- list(
  gini=Gini,
  atkinson=Atkinson,
  ent=entropy,
  theil=Theil)

ineq_idxes <- df |>
  summarize(
    across(n, .fns=ineq_idx_fns),
    n_alker=min_maj(cprop, rrnk),
    n_last_dec=last_dec(cprop, rrnk),
    .by= c(year, sex, country),
    ) |>
  pivot_longer(
    cols=starts_with("n"),
    names_to="index_name",
    values_to="index")
```

```r
ineq_idxes |>
  ggplot() +
  aes(x=year, y=index, color=sex) +
  geom_line() +
  facet_grid(rows=vars(index_name), cols=vars(country), scales="free_y")
```
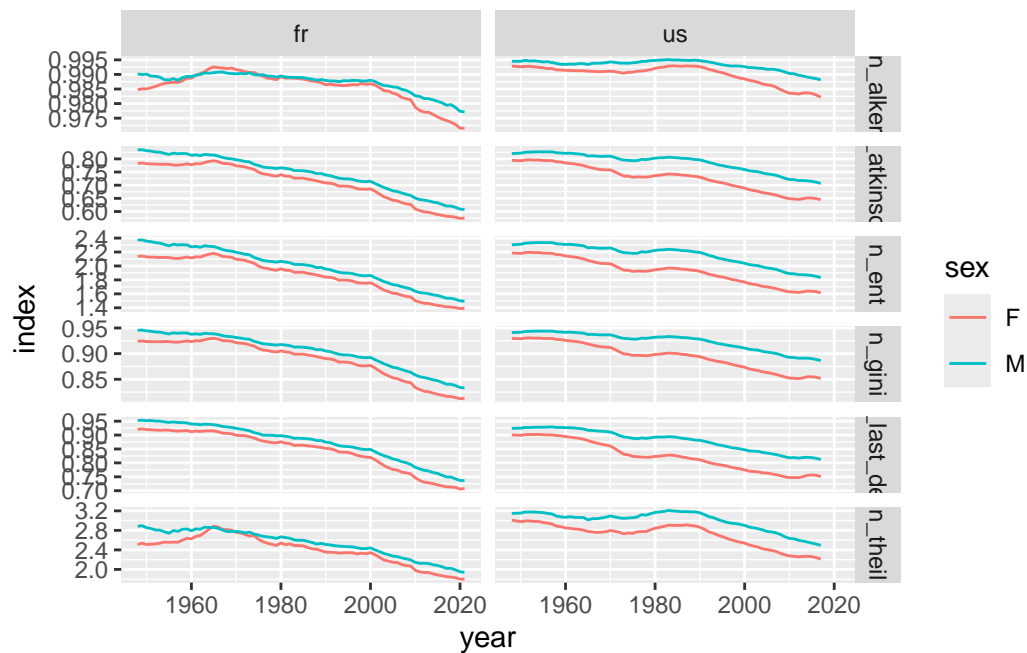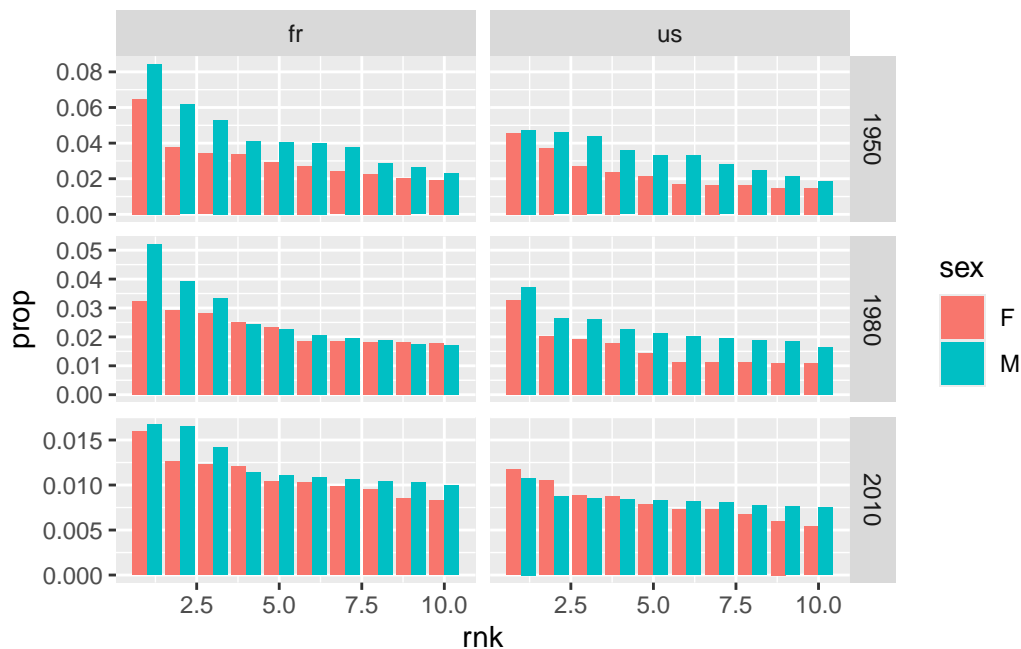
```r
df <- df |>
  group_by(country, sex, name) |>
  mutate(best_rnk=min(rnk)) |>
  ungroup()
```

```r
(
  df |>
  filter(rrnk<.9, round(10000*rrnk)%%10==1) |>
  ggplot() +
    aes(x=1-rrnk, y=1-cprop, color=sex, frame=year) +
    geom_point(size=.2)  +
    coord_fixed() +
    facet_wrap(~ country)
) |>
  plotly::ggplotly()
```
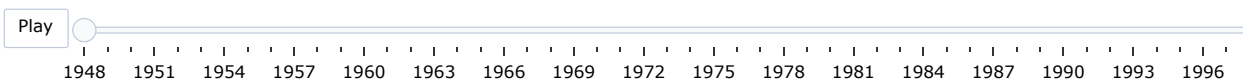
```
# |>  plotly::ggplotly()
```

```
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)) |>
  group_by(year,sex) |>
  mutate(prop=n/sum(n)) |>
  ungroup() |>
  select(year, sex, name, n, prop, country)
```

U+0128

```
extract_pattern <- \(x)
  str_c((as.character(lkp[as.character(x[x!=0 & !is.na(x)])])), collapse="")
```

```
df <- df |>
  group_by(country,sex, name) |>
  arrange(year) |>
  mutate(sprop=slide_dbl(pmax(prop, 1e-4), mean, .before=2, .after =2)) |>
  ungroup()
```

```
df_patterns <- df |>
  group_by(country,sex, name) |>
  arrange(year) |>
  mutate(change=log(sprop)) |>
```

```
    mutate(change=sign(change-lag(change, default = change[1]))) |>
    summarise(change_pattern=extract_pattern(change), .groups = "drop") |>
    arrange(country,sex, change_pattern)
```

```
df_patterns |>
  filter(name %in% c('JULES', 'KEVIN', 'STÉPHANE', 'ARTHUR', 'MICHEL', 'EMILE'), sex=='M')
```

```
# A tibble: 6 x 4
  country sex   name        change_pattern
  <chr>   <fct> <chr>       <chr>
1 fr      M     KEVIN       NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
2 fr      M     ARTHUR      NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
3 fr      M     JULES       NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
4 fr      M     EMILE       NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
5 fr      M     MICHEL      NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
6 fr      M     STÉPHANE    NULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNULLNU~
```

```
df  |>
  filter(name %in% c('STÉPHANE', 'KEVIN', 'ENZO'), sex=='M') |>
  ggplot() +
  aes(x=year) +
  geom_point(aes(y=prop, shape=name), color="blue", alpha=.5, size=.2) +
  geom_line(aes(y=sprop, linetype=name), color="red", linewidth=.2) +
  scale_y_log10()
```