# Ratio estimation: The first statistical analysis of a sample survey

## Unit 4 Lecture 2

Jonathan Auerbach
STAT 489 Pre-Cap Prof Development
jauerba@gmu.edu

November 2, 2021

# Learning Objectives

After this lecture, you will be able to:

1. Define the ratio estimator and describe how Laplace used it to estimate the population of France from the number of births.
2. Use the `ggrepel` package to visualize which administrative regions if any have atypical birth rates.
3. Derive the asymptotic variance of the ratio estimator by modeling the relationship between the population and the number of births.
4. Derive the asymptotic variance of the ratio estimator by modeling how Laplace selected his sample.

# These slides use the following R packages

Setup:

```
library("readxl")
library("knitr")
library("tidyverse")
library("ggplot2")
library("ggrepel")
theme_set(theme_bw(base_size = 20))
```

# Ratio estimation: The first sample survey analysis

- Laplace wanted to estimate the total population of France in 1802.
  - But he was only able to sample the population in a handful of administrative regions.

- Laplace used the number of births—known at all locations from government records—to estimate the total population in three steps:
  1. He first assumed the ratio of births to population (i.e. the birth rate) was the same for all locations.
  2. He then estimated this ratio using data from 30 regions in which both the number of births and the population were sampled.
  3. Finally, he divided the total number of births in France by the ratio, yielding the estimated total population of France.

- These steps as often referred to as "ratio estimation," and the estimated population as the "ratio estimator."

# Ratio estimation: The first sample survey analysis

- Laplace studied ratio estimation over a thirty-year period.
  - Laplace first examined the accuracy of ratio estimation in *On births, marriages, and deaths in Paris from 1771 to 1784* (1783).
  - He published the results of the 1802 population estimate in *Analytic Theory of Probabilities* (1812, Book 2 Chapter 6 Section 31).
  - Finally, Laplace further discussed the estimate in *A Philosophical Essay on Probabilities* (1814, Chapter 8).
- Laplace was not the first to estimate a population using ratios.
  - Graunt (1654) used church records to estimate the total population of London more than a hundred years earlier.
  - But Laplace was the first to derive the asymptotic distribution of the ratio estimator, characterizing its accuracy relative to the unobserved total population.
- Today, ratio estimation is commonly used to analyze survey data.

# Pierre-Simon Laplace, *Analytic Theory* (1812)

6. ON THE PROBABILITY OF CAUSES AND OF FUTURE EVENTS 161

If we make $\gamma z'' = t$, we will have

$$\frac{z''}{s'} = t\sqrt{\frac{2(p'-s')}{p's'}},$$

and the probability $P$ that the ratio of the error of the number $s'$ from the table, to this number itself, will be comprehended within the limits $\pm t\sqrt{\frac{2(p'-s')}{p's'}}$ is

$$P = 2\int \frac{dt}{\sqrt{\pi}}\, c^{-t^2},$$

the integral being taken from $t$ null. We see thus that the value of $t$, and consequently the probability $P$ remaining the same, this ratio increases when $s'$ diminishes; thus the numbers from the table are so much less certain, as they are more extended from the first $p'$. We see further that this ratio diminishes in measure as $p'$ increases, or in measure as we multiply the observations; in a manner that we are able by this multiplication, to diminish at the same time this ratio and to increase $t$; this ratio [391]
becoming null when $p'$ is infinite, and $P$ becoming then equal to unity.

§31. Let us apply the preceding analysis to the research on the population of a great empire. One of the simplest and most proper ways to determine this population, is the observation of the annual births of which we are obliged to take account in order to determine the civil state of the infants. But this way supposes that we know very nearly the ratio of the population to the annual births, a ratio that we obtain by making at many points of the empire, the exact denumeration of the inhabitants, and by comparing it to the corresponding births observed during some consecutive years: we conclude from it next, by a simple proportion, the population of all the empire. The government has well wished, at my prayer, to give orders to have with precision, these data. In thirty departments distributed over the area of France, in a manner to outweigh the effects of the variety of climates, we have made a choice of the townships of which the mayors, by their zeal and their intelligence, would be able to furnish the most precise information. The exact denumeration of the inhabitants of these townships, for 22 September 1802, is totaled to 2037615 individuals. The summary of the births, of the marriages and of the deaths, from 22 September 1799 to 22 September 1802, has given for these three years,

| *Births* | *Marriages* | *Deaths* |
|---|---|---|
| 110312 boys, | 46037, | 103659 males, |
| 105287 girls, | | 99443 females. |

The ratio of the births of boys to those of girls, that this summary presents, is the one of 22 to 21; and the marriages are to the births, as 3 to 14; the ratio of the population to the annual births is 28,352845. In supposing therefore the number of annual births in France, equal to one million, that which deviates little from the truth; we will have, by multiplying by the preceding ratio, this last number, the population of France equal to 28352845 individuals. Let us see the error that we are able to fear in this evaluation.

Source: *https://www.newworldencyclopedia.org/entry/Pierre-Simon_Laplace*
*https://gdz.sub.uni-goettingen.de/id/PPN129323640_0018*

# Laplace first to formally study ratio estimation

- Let $Y$ denote the total population of France in 1802 and $X$ the total number of births.
  - Laplace wanted to estimate $Y$ from $X$.
  - Graunt had previously observed that $Y$ could be determined by multiplying $X$ by $\beta$, the population per birth—or equivalently, divide by $p = 1/\beta$, the number of births per person.
- Laplace estimated $\beta$ and $p$ using a somewhat systematic sample of regions. Let $y$ denote the population in Laplace's sample regions and $x$ the corresponding number of births.
  - Laplace calculated the ratio estimator $\hat{Y} = \frac{y}{x} X = \hat{\beta} X = X/\hat{p}$.
- n.b. the ratio estimator is also obtained if the sample ratio is used to estimate the population not sampled, $X - x$. i.e.

$$\tilde{Y} := y + \frac{y}{x}(X - x) = y + \frac{y}{x}X - \frac{y}{x}x = \frac{y}{x}X = \hat{Y}$$

# Laplace sampled the population in 30 regions and calculated number of births in the preceding 3 years

```
sample_fr <- tibble(                    # data from Bru (1988)
 region = c("Alpes basses", "Ardennes", "Aube",
   "Bouches-du-Rhone", "Charente", "Doubs", "Dyle", "Gard",
   "Herault", "Ille et Villaine","Jura", "Liamone",
   "Loire inferieure", "Lozere", "Meuse",
   "Meuse inferieure", "Mont Blanc", "Mont Tonnerre",
   "Nord", "Puy-de-Dome", "Rhin bas", "Sarre", "Seine",
   "Seine inferieure", "Seine-et-Oise", "Sesia",
   "Deux-Sevres", "Stura", "Var", "Vienne"),
 population = c(51678, 50900, 51717, 49996, 58229, 50170,
   109568, 65526, 107227, 106157, 58514, 14509, 97778,
   50867, 72419, 45998, 50056, 50507, 51796, 48265, 49999,
   55002, 52585, 135497, 55334, 209510, 49993, 86315,
   49957, 51546),
 births = c(6094, 5210, 6071, 5471, 5961, 5393, 12010,
   7352, 12247, 12246, 5780, 1422, 9644, 4075, 7772, 3927,
   5215, 6070, 5876, 5050, 5758, 6174, 5499, 13584, 4846,
   22382, 5058, 9446, 5325, 4641) / 3)
```

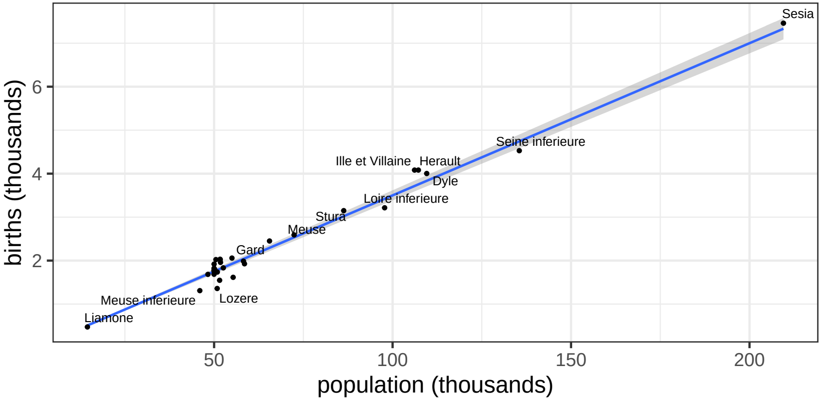
# Laplace sampled the population in 30 regions and calculated number of births in the preceding 3 years

```
sample_fr %>%
  head(10) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| region | population | births |
|---|---:|---:|
| Alpes basses | 51,678 | 2,031.33 |
| Ardennes | 50,900 | 1,736.67 |
| Aube | 51,717 | 2,023.67 |
| Bouches-du-Rhone | 49,996 | 1,823.67 |
| Charente | 58,229 | 1,987.00 |
| Doubs | 50,170 | 1,797.67 |
| Dyle | 109,568 | 4,003.33 |
| Gard | 65,526 | 2,450.67 |
| Herault | 107,227 | 4,082.33 |
| Ille et Villaine | 106,157 | 4,082.00 |

# Nearly all regions had 28.352845 people per birth

```r
sample_fr %>% ggplot(aes(population/1000, births/1000)) +
  geom_smooth(method = "lm", formula = y ~ x + 0,
              aes(weight = 1000/births)) +
  geom_point() + geom_text_repel(aes(label = region)) +
  labs(x = "population (thousands)",
       y = "births (thousands)")
```



# Laplace concluded France had 28,352,845 people

- Laplace assumed there were one million births in France in 1802.
  - Multiplying one million by the sample number of people per birth produced an estimated total population of 28,352,845 people.

```
sample_total_fr <-
  sample_fr %>%
  summarize(x = sum(births),
            y = sum(population),
            y/x,
            X = 1e6,
            `X(y/x)` = X * y / x)

sample_total_fr %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| x | y | y/x | X | X(y/x) |
|---:|---:|---:|---:|---:|
| 71,866.33 | 2,037,615 | 28.35 | 1e+06 | 28,352,845 |

# How accurate is the ratio estimator?

- Laplace initially assumed $x \sim \text{Binomial}(p, y)$. He then derived the asymptotic (posterior) distribution of $X/\hat{p} = \frac{y}{x}X$.
  - Cochran (1978) presents a similar but modern argument. Observe that $\sqrt{y}(\hat{p} - p) \to \mathcal{N}(0,\ p(1-p))$

- To determine the distribution of $X/\hat{p}$, recall the Delta Method:

if $\sqrt{n}(\hat{\mu}_n - \mu) \to \mathcal{N}(0,\ \sigma^2),$ then $\sqrt{n}\,(f(\hat{\mu}_n) - f(\mu)) \to \mathcal{N}(0,\ f'(\mu)^2\sigma^2)$

  - Substituting $y$ for $n$ and $p$ for $\mu$; setting $f(\hat{p}) = X/\hat{p}$ and $f(p) = X/p$; and noting $f'(p)^2 = X^2/p^4$, it follows that

$$X/\hat{p} \mathrel{\dot\sim} \mathcal{N}\left(X/p,\ \frac{p(1-p)}{yp^4}X^2\right)$$

- The variance can be approximated by substituting $\hat{p}$ for $p$,

$$\text{Var}(X/\hat{p}) \approx \frac{\frac{x}{y}(1 - \frac{x}{y})}{y(\frac{x}{y})^4}X^2 = \frac{(y-x)y}{x^3}X^2$$

# How accurate is the ratio estimator?

```
sample_total_fr %>%
  transmute(Y_hat = y / x * X,
            se = sqrt((y - x) * y / x^3 * X^2),
            lower = Y_hat - 2 * se,
            upper = Y_hat + 2 * se) %>%
  rename(`$\\hat \\text Y$` = Y_hat) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| $\hat{\text{Y}}$ | se | lower | upper |
|---:|---:|---:|---:|
| 28,352,845 | 103,881.2 | 28,145,083 | 28,560,607 |

- Laplace assumed sample population $y$ was measured without error, and number of births $x$ varied due to binomial sampling variation.
  - Since $y$ is large, the ratio estimator is found to be very accurate.
  - But was the population measured with error? Does the birth rate vary, and the sample locations have unusual birth rates by chance?

# The population-error model

- Let $y_i$ denote the population at sampled region $i$ and $x_i$ the number of births. Suppose $y_i$ proportional to $x_i$ plus measurement error. i.e.

$$y_i = \beta x_i + \epsilon_i \text{ where } \epsilon_i \sim \mathcal{N}(0,\ \sigma^2 x_i)$$

  - Let $y = \sum y_i$ and $x = \sum x_i$ denote the population and births in the sampled regions. Define estimates $\hat{\beta} = 1/\hat{p} = \frac{y}{x}$ and $\hat{\epsilon}_i = y_i - \hat{\beta} x_i$.

- To find the distribution of $\hat{\beta} X$, note that $\epsilon = y - \beta x \sim \mathcal{N}(0,\ \sigma^2 x)$.
  - Multiplying by $X/x$,

$$\epsilon X/x = (y - \beta x)X/x = (\hat{\beta} X - \beta X) \sim \mathcal{N}(0,\ \sigma^2 X^2/x)$$

  - It follows that $\hat{\beta} X \sim \mathcal{N}(\beta X,\ \sigma^2 X^2/x)$

- We approximate $\sigma^2$ with sample variance of $\hat{\epsilon}_i/\sqrt{x_i(1 - x_i/x)}$ since

$$\begin{aligned}\mathrm{Var}(\hat{\epsilon}_i) &= \mathrm{Var}(y_i - \hat{\beta} x_i) = \mathrm{Var}(y_i - (y/x)x_i) = \\ &= \mathrm{Var}(y_i) + \mathrm{Var}(y)(x_i/x)^2 - 2\mathrm{Cov}(y_i, y)(x_i/x) \\ &= \sigma^2 x_i + \sigma^2 x (x_i/x)^2 - 2\sigma^2 x_i^2/x = \sigma^2 x_i(1 - x_i/x)\end{aligned}$$

# The population-error model

```
beta_hat <- sample_total_fr$y / sample_total_fr$x

e <-
  sample_fr %>%
  mutate(e = (population - beta_hat * births) /
              (sqrt(births * (1 - births / sum(births))))) %>%
  pull(e)

sample_total_fr %>%
  transmute(Y_hat = y / x * X,
            se = sd(e) * X / sqrt(x),
            lower = Y_hat - 2 * se,
            upper = Y_hat + 2 * se) %>%
  rename(`$\\hat \\text Y$` = Y_hat) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| $\hat{\text{Y}}$ | se | lower | upper |
|---:|---:|---:|---:|
| 28,352,845 | 466,905.5 | 27,419,034 | 29,286,656 |

# The sampling design-based model

- Let $y_i$ denote the population at region $i = 1, \ldots, N$ (including regions not sampled) and $x_i$ the number of births. Let $z_i$ denote if region $i$ sampled (i.e. $z_i = 1$ if region $i$ sampled and $z_i = 0$ if not).
  - Let $y = \sum_{i=1}^N y_i z_i$ and $x = \sum_{i=1}^N x_i z_i$ denote population and births at sample regions and $Y = \sum_{i=1}^N y_i$ and $X = \sum_{i=1}^N x_i$ at all regions.
  - Estimate $\beta = Y/X$ with $\hat{\beta} = 1/\hat{p} = \frac{y}{x}$.
- To determine distribution of $\hat{\beta}X$, assume $z_i \sim$ Bernoulli$(q)$ so that

$$\sqrt{N}\epsilon = \sqrt{N}(y - \beta x) = \sqrt{N}\sum_{i=1}^{N}(y_i - \beta x_i)z_i \to \mathcal{N}(0,\ q(1-q)\sigma^2)$$

  - Multiplying by $X/x$ (and ignoring the randomness of $x$),

$$\sqrt{N}\epsilon X/x = \sqrt{N}(y - \beta x)X/x = \sqrt{N}(\hat{\beta}X - \beta x) \to \mathcal{N}(0,\ q(1-q)\sigma^2 X^2/x^2)$$

  - If $x \approx Xq$, then $\hat{\beta}X \stackrel{.}{\sim} \mathcal{N}\Big(\beta X,\ \frac{1-q}{q}\sigma^2\Big)$
- We approximate $\sigma^2$ with the sample variance of $\hat{\epsilon}_i = y_i - \hat{\beta}x_i$.

# The sampling design-based model

```r
beta_hat <- sample_total_fr$y / sample_total_fr$x

e <- sample_fr$population - beta_hat * sample_fr$births

q <- sample_total_fr$x / sample_total_fr$X

N <- 30 / q

sample_total_fr %>%
  transmute(Y_hat = y / x * X,
            se = sqrt((1 - q) / q * N) * sd(e),
            lower = Y_hat - 2 * se,
            upper = Y_hat + 2 * se) %>%
  rename(`$\\hat \\text Y$` = Y_hat) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| $\hat{\text{Y}}$ | se | lower | upper |
|---:|---:|---:|---:|
| 28,352,845 | 403,708.7 | 27,545,427 | 29,160,262 |

# Which model is the right model?

- The intervals derived from the population-error and design-based models are 4x larger than the interval from the binomial model.
  - The design-based interval assumes regions are randomly selected.
  - In reality, Laplace picked regions evenly distributed across France. Assistants picked subregions until roughly 50,000 people per region.
- The binomial model produces intervals that appear to be too narrow, but since there was no 1802 census we cannot know for sure.
  - One way to compare the models is to use the 1801 census, which attempted to enumerate the entire population of France. (The 1801 census was not available when the 1802 sample was collected.)
  - Such a comparison suggests the binomial model intervals are too narrow while the other two intervals are appropriate. (See Appendix.) Note that the 1801 census was thought to be inaccurate (Bru 1988).

# Ratio estimation and its variants are popular today

- Population estimation has improved substantially since 1802.
  - But even modern surveys have errors due to small samples or respondants unable or unwilling to participate.
  - Auxiliary information, such as administrative records, are still used to adjust surveys to more accurately reflect the population.

- A common use of ratio estimation is in post-stratification:
  - For each strata (group) $s$, the survey taker obtains sums $y_s$ and $x_s$, denoting the outcome of interest and an auxiliary outcome for a sample of respondants.
  - To estimate $Y_s$, the sum of the auxiliary outcome for all individuals in the strata, $X_s$, is used to construct a ratio estimator for each strata.
  - Summing across strata produces an estimate of $Y$,

$$\hat{Y} = \sum_{s=1}^{S} \hat{Y}_s = \sum_{s=1}^{S} X_s \frac{y_s}{x_s}$$

# References

1. Bru, Bernard. "The estimates of Laplace. An example: Research concerning the population of a large empire, 1785-1812." Journal de la Société de statistique de Paris 129.1-2 (1988): 6-45.

2. Cochran, William G. Laplace's ratio estimator. Contributions to survey sampling and applied statistics. Academic Press, 1978.

3. Hald, Anders. A history of mathematical statistics from 1750 to 1930. John Wiley & Sons, 1998.

4. Historical data from the General Statistics of France. https://www.insee.fr/fr/statistiques/2591293?sommaire=2591397

5. Laplace, Pierre Simon. Analytic theory of probabilities. 1810.

6. Laplace, Pierre Simon. A philosophical essay on probabilities, book 2. 1814.

# Appendix: Download data from 1801 census

```r
url <-
"www.insee.fr/fr/statistiques/fichier/2591293/TERR_T86.xls"

download.file(url, destfile = "TERR_T86.xls")

france <-
 read_xls("TERR_T86.xls", skip = 7) %>%
 transmute(
  region =
    str_replace(`Nom de l'unité d'analyse`, "\\(", " \\("),
  region = str_replace_all(region, "-\\)", "\\)"),
  births =
    `Naissances légitimes et naturels: total, 1800 à 1801`,
  population = `Nombre d'habitants, 1801`) %>%
  filter(region != "TARN-ET-GARONNE")
```

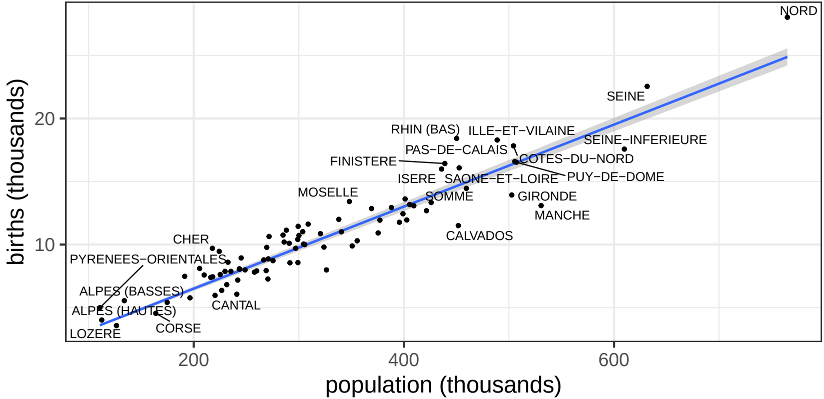
# Appendix: Table of first 10 of 85 census regions

```r
france %>%
  head(10) %>%
  kable(format.args = list(big.mark = ","))
```

| region | births | population |
|---|---:|---:|
| FRANCE | 903,688 | 27,349,003 |
| AIN | 9,703 | 297,071 |
| AISNE | 13,335 | 425,981 |
| ALLIER | 7,991 | 248,864 |
| ALPES (BASSES) | 5,552 | 133,966 |
| ALPES (HAUTES) | 4,015 | 112,500 |
| ARDECHE | 8,784 | 266,656 |
| ARDENNES | 7,911 | 259,925 |
| ARIEGE | 5,773 | 196,454 |
| AUBE | 6,823 | 231,455 |

# Appendix: Visualization of 85 census regions

```r
france %>% filter(region != "FRANCE") %>%
  ggplot(aes(population/1000, births/1000)) +
  geom_smooth(method = "lm", formula = y ~ x + 0,
              aes(weight = 1000/births)) +
  geom_point() + geom_text_repel(aes(label = region)) +
  labs(x = "population (thousands)",
      y = "births (thousands)")
```



# Appendix: Table of regions randomly sampled

```
set.seed(1)

france_sample <- france %>%
  filter(region != "FRANCE")%>%
  filter(rbinom(length(region), 1, 1/8) == 1)

france_sample %>% kable(format.args = list(big.mark = ","))
```

| region | births | population |
|---|---:|---:|
| ALPES (BASSES) | 5,552 | 133,966 |
| ARDECHE | 8,784 | 266,656 |
| ARDENNES | 7,911 | 259,925 |
| CORREZE | 8,041 | 243,654 |
| COTES-DU-NORD | 17,829 | 504,303 |
| PUY-DE-DOME | 16,530 | 507,128 |
| SARTHE | 12,938 | 388,143 |
| SOMME | 14,458 | 459,453 |
| VENDEE | 8,080 | 243,426 |

# Appendix: Ratio estimation using sample

```
france_sample_total <- france_sample %>%
  summarize(x = sum(births), y = sum(population))

france_sample_total %>% mutate(y/x) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| x | y | y/x |
|---:|---:|---:|
| 100,123 | 3,006,654 | 30.03 |

```
france_total <- france %>% filter(region == "FRANCE") %>%
  select(X = births, Y = population)

france_total %>% bind_cols(france_sample_total) %>%
  transmute(X, Y, Y/X, `X(y/x)` = X * y / x) %>%
  kable(digits = 2, format.args = list(big.mark = ","))
```

| X | Y | Y/X | X(y/x) |
|---:|---:|---:|---:|
| 903,688 | 27,349,003 | 30.26 | 27,137,392 |

# Appendix: Comparison of accuracy estimates

```r
beta_hat <- france_sample_total$y / france_sample_total$x
e_1 <- (france_sample$population -
        beta_hat * france_sample$births)
e_2  <-
  e_1 / sqrt(france_sample$births *
   (1 - france_sample$births / sum(france_sample$births)))

france_sample_total %>% bind_cols(france_total) %>%
  transmute(Y, Y_hat = y / x * X,
  `se binomial` = sqrt((y-x)* y / x^3 * X^2),
  `se pop error` = sd(e_2) * X / sqrt(x),
  `se design` = sqrt((1 - 1/8) / (1/8) * 85) * sd(e_1)) %>%
  rename(`$\\hat \\text Y$` = Y_hat) %>%
  kable(digits = 0, format.args = list(big.mark = ","))
```

| Y | $\hat{\text{Y}}$ | se binomial | se pop error | se design |
|---:|---:|---:|---:|---:|
| 27,349,003 | 27,137,392 | 84,323 | 639,070 | 496,051 |