MONASH University

Gender birth rate trend analysis in the United States

Arindam Baruah
abar0090@student.monash.edu (32779267)

Report for
ETC 5242 Task 1

8 August 2023

MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Contents

1 Introduction **2**

2 Research Questions **3**

3 Analysis **3**

3.1 Query 1 3

3.2 Query 2 5

3.3 Query 3 6

3.4 Query 4 9

3.5 Query 5 10

4 Conclusion **11**

5 Resources **12**

1 Introduction

The current study delineates a detailed analysis of the birth counts for boys and girls born in the country of The United States of America. The study is based on two separate datasets, the first of which contain gender wise birth data from the years of **1629 to 1710** and the second, containing the gender birth data from the years of **1940 to 2002** in the country of The United States. Utilisation of statistical techniques and subsequent feature engineering of the current and the historical data would further provide key insights into the changes of gender birth rate in the aforementioned country of interest. While the studies of Mathews, Hamilton, et al. (2005) and Ritchie and Roser (2019) have uncovered change of birth rates as a result of various socio-economic causes such as demographic changes, **the current study would however be limited to exploring the key changes in the number of births for girls and boys using statistical methods and feature engineering.**

It is critical to acknowledge that while the study by Pryzgoda and Chrisler (2000) defines the terminologies of “Sex ratio” and “Gender ratio” as two different statistics, however for the sake of simplicity, these terms may be used interchangeably in the context of the current study.

2 Research Questions

In this section of the study, the following key research questions pertaining to the new births in The United States will be formulated and further analysed in section 3 :

1. **How do the datasets belonging to the periods of 1629-1710 and 1940-2002 differ from one another** ?
2. **How has the birth rate for girls changed over the period of 1940-2002** ?
3. **Are there any similarities in the birth rate of girls between the time periods of 1629-1710 and 1940-2002** ?
4. **Would creating new statistical features allow us to gain better insights into the data** ?
5. **Are boys born in greater proportion to girls during the period of 1940-2002** ?

3 Analysis

The current section will provide a step by step analysis for each of the formulated research questions of section 2.

3.1 Query 1

A basic understanding on the size of the data and the number of entries for the datasets belonging to each of the two time periods will be explored in this section.

```
head(df_present) %>%
  kable(caption = "Preview of the data containing number of boys and girls born in The United States between 1940-2002.") %>%
  kable_styling(bootstrap_options = c("bordered","hover"),
                                  latex_options = "HOLD_position")
```

Table 1: *Preview of the data containing number of boys and girls born in The United States between 1940-2002.*

year	boys	girls
1940	1211684	1148715
1941	1289734	1223693
1942	1444365	1364631
1943	1508959	1427901
1944	1435301	1359499
1945	1404587	1330869

The above code-chunk and its output as observed through table 1 provides us with the glimpse of the dataset. The dataset contains three features. These features are explained as below:

- **Year** : The year pertaining to the count of new births in the United States.
- **boys** : The number of births classified as “boys” for the corresponding year.
- **girls** : The number of births classified as “girls” for the corresponding year.

Let us observe how do the values of the number of births between the periods of 1629-1710 and 1940-2002 vary from one another through tables 2 and 3 respectively.

```
summary(df_arb) %>% kable(caption = "Summary statistics for data between 1629-1710.") %>%
  kable_styling(bootstrap_options = c("bordered","hover"),
                                    latex_options = "HOLD_position")
```

Table 2: *Summary statistics for data between 1629-1710.*

	year	boys	girls
	Min. :1629	Min. :2890	Min. :2722
	1st Qu.:1649	1st Qu.:4759	1st Qu.:4457
	Median :1670	Median :6073	Median :5718
	Mean :1670	Mean :5907	Mean :5535
	3rd Qu.:1690	3rd Qu.:7576	3rd Qu.:7150
	Max. :1710	Max. :8426	Max. :7779

```
summary(df_present) %>% kable(caption = "Summary statistics for data between 1940-2002.") %>%
  kable_styling(bootstrap_options = c("bordered","hover"),
                                    latex_options = "HOLD_position")
```

Table 3: *Summary statistics for data between 1940-2002.*

	year	boys	girls
	Min. :1940	Min. :1211684	Min. :1148715
	1st Qu.:1956	1st Qu.:1799857	1st Qu.:1711404
	Median :1971	Median :1924868	Median :1831679
	Mean :1971	Mean :1885600	Mean :1793915
	3rd Qu.:1986	3rd Qu.:2058524	3rd Qu.:1965538
	Max. :2002	Max. :2186274	Max. :2082052

As we can clearly observe, the **magnitude of the boys and girls born during the period of 1940-2002 are much larger than that for the period of 1629-1710** as reported by the study of Arbuthnot (1710). This is expected as a result of the global rise of population owing to factors such as better infrastructure, better lifestyle, better socio-economic factors and improvement in medical sciences.

Let us try to visualise the same through figure 1.

```r
options(scipen = 999) # To remove scientific notation

df_arb_long <-
  pivot_longer(
    df_arb,
    names_to = "gender",
    values_to = "born",
    cols = c(boys, girls)
  )
pl1 <-
  ggplot(data = df_arb_long, aes(x = year, y = born, fill = gender)) +
  geom_area(color = 'black') + theme_classic() +
  ggtitle("Birth statistics between 1629-1710") +
  theme(plot.title = element_text(hjust = 0.5),aspect.ratio = 0.5) +
  labs(fill = "Gender", x = "Year", y = "Number of births")


df_present_long <-
  pivot_longer(
    df_present,
    names_to = "gender",
    values_to = "born",
    cols = c(boys, girls)
  )
pl2 <-
  ggplot(data = df_present_long, aes(x = year, y = born, fill = gender)) +
  geom_area(color = 'black') +
  theme_classic() +
  ggtitle("Birth statistics between 1940-2002 \n in The US") +
  theme(plot.title = element_text(hjust = 0.5),aspect.ratio = 0.5) +
  labs(fill = "Gender", x = "Year", y = "Number of births")

plot_grid(pl1, pl2, labels = "AUTO", ncol = 1)
```

As we can observe through figure 1, the magnitude of births are observed to be **significantly higher during the period between 1940-2002 in the United States** when compared to the period between 1620-1710 of Arbuthnot's data.

3.2 Query 2

In this section, the number of births classified as "girls" will be focused on to gain further insights through a scatter plot as visualised through figure 2.

```r
pl3 <- ggplot(data = df_present, aes(x = year, y = girls)) +
  geom_point(shape = 2, color = 'red') + theme_classic() +
  ggtitle("Number of girls born during 1940-2002 \n in the US") +
  theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 0.5) + labs(x = "Year",
                                                                           y = "Number of births")
pl4 <- ggplot(data = df_arb, aes(x = year, y = girls)) +
  geom_point(shape = 2, color = 'blue') + theme_classic() +
  ggtitle("Number of girls born during 1629-1710") +
  theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 0.5) + labs(x = "Year",
                                                                           y = "Number of births")
plot_grid(pl3,pl4,ncol =1)
```

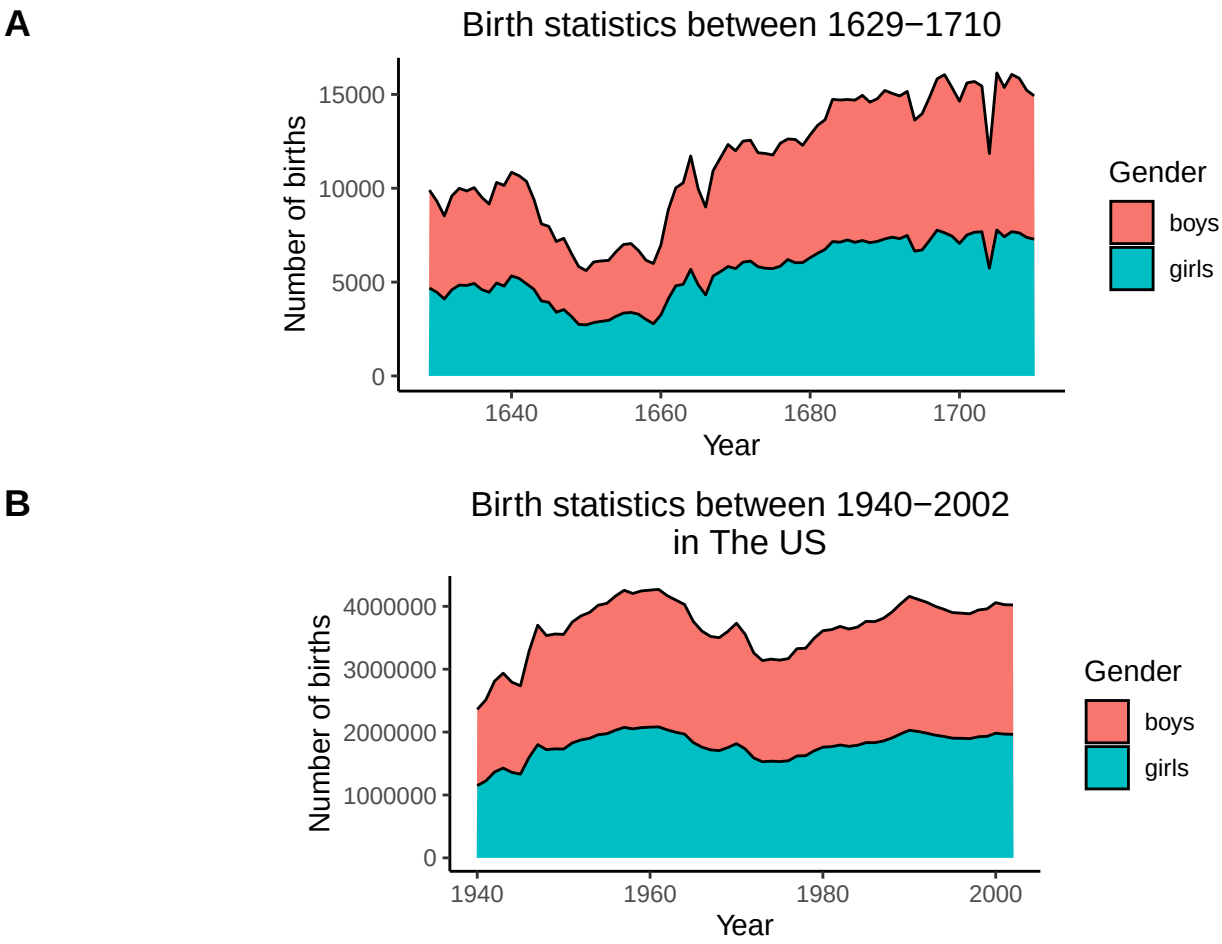

Figure 1: *Total number of births in two time periods*

3.3 Query 3

Based on the analysis of figures 2 and 3, we can observe the following key points:

1. The magnitude of girls born in the period between 1940-2002 is **significantly higher** than in the period between 1629-1710.
2. Although the magnitude of girls born for the two time periods vary significantly, however a few similarities in the trends may be observed between the two time periods. These include the sudden dip in births followed by a rise and eventual stagnation of the numbers.
3. During the period between 1940-1960, there was **a consistent increase in the number of girls born in the country**.
4. The period between 1960-1975 observed a **sudden drop in the number of girls born**.
5. However, the number of girls born were **back on the rise** after the year 1975.
6. Between the period of 1990-2000, the **number of girls born were observed to stagnate at approximately 200,0000 births a year**.

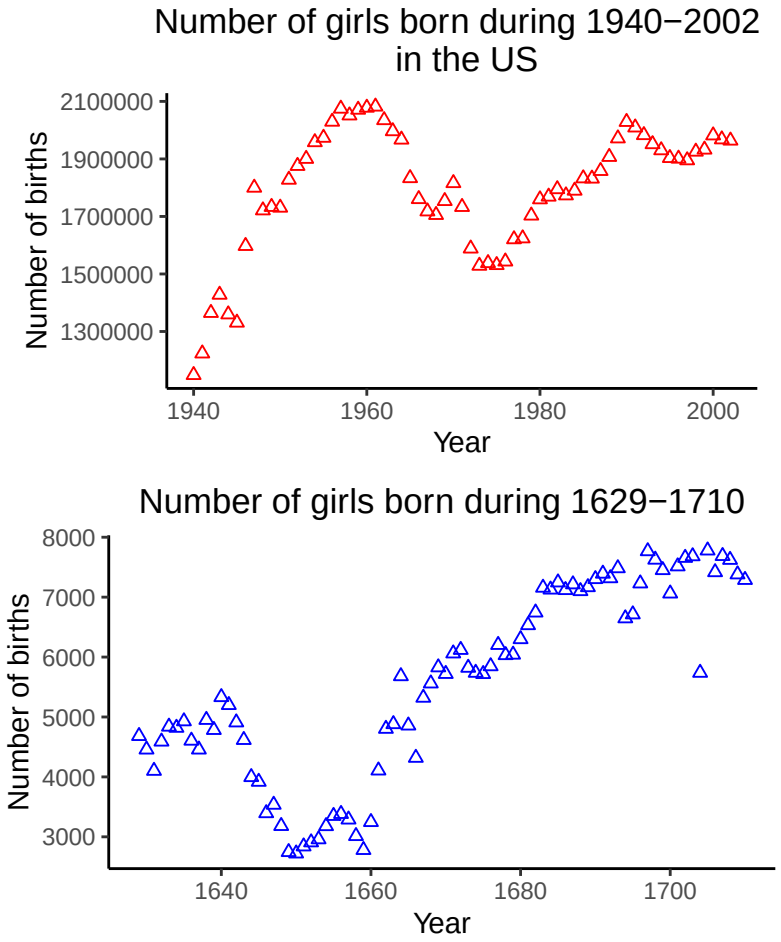

Figure 2: *Girls born during each periods*

```
pl5 <- ggplot(data = df_present, aes(x = year, y = girls)) +
  geom_point(shape = 2, color = 'red') + theme_classic() +
  ggtitle("Number of girls born during 1940-2002 \n in the US") +
   labs(x = "Year", y = "Number of births") + theme(plot.title = element_text(hjust = 0.5),
                                                    aspect.ratio = 0.7) +
  annotate(
    "segment",
    x = 1945,
    y = 1900000,
    xend = 1950  ,
    yend = 1800000,
    arrow = arrow(type = "closed",
                  length = unit(0.02, "npc"))
  ) +
  annotate(
    "text",
    x = 1946,
    y = 1980000,
    colour = "black",
    label = 'Consistent rise in number of girls \n born between 1940-1960',
    size = unit(2.5, "pt")
  ) +
  annotate(
    "segment",
    x = 1965,
    y = 1300000,
    xend = 1973  ,
    yend = 1500000,
    arrow = arrow(type = "closed",
```

```
                  length = unit(0.02, "npc"))
  ) +
  annotate(
    "text",
    x = 1965,
    y = 1250000,
    colour = "blue",
    label = 'Dip in number of girls born \n between 1960-1975',
    size = unit(2.5, "pt")
  ) +
  annotate(
    "segment",
    x = 1983,
    y = 1500000,
    xend = 1985  ,
    yend = 1700000,
    arrow = arrow(type = "closed",
                  length = unit(0.02, "npc"))
  ) +
  annotate(
    "text",
    x = 1985,
    y = 1450000,
    colour = "darkgreen",
    label = 'Girl child births start \n increasing again from 1975',
    size = unit(2.5, "pt")
  ) +
    annotate(
    "segment",
    x = 1995,
    y = 1700000,
    xend = 1995  ,
    yend = 1850000,
    arrow = arrow(type = "closed",
                  length = unit(0.02, "npc"))
  ) +
  annotate(
    "text",
    x = 1995,
    y = 1600000,
    colour = "darkred",
    label = 'Stagnation in the number \n of girls born after 1990',
    size = unit(2.5, "pt")
  )
pl5
```

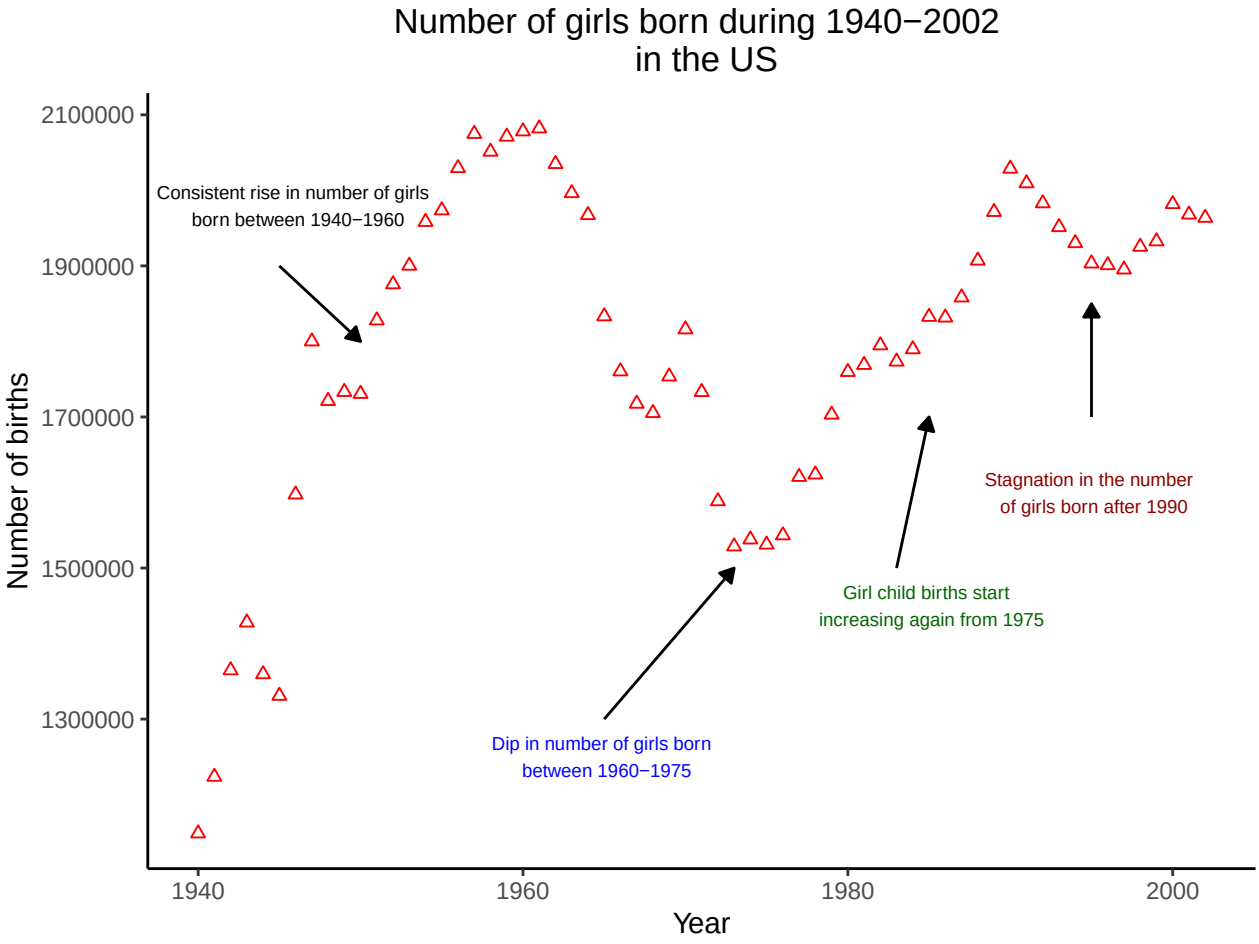

Figure 3: *Trend analysis of girls child births in The United States during 1940-2002*

3.4 Query 4

While the raw dataset provides an overall understanding of the gender birth number in the United States, this section will focus on the creation of new statistical features which will help provide deeper insights. Based on the available data, a total of 4 new features will be created. These features may be described as follows:

1. **Total** : Total number of child births for the corresponding year in the United States.
2. **Ratio** : This feature maybe defined as the ratio of the number of girls born to the number of boys born for a particular year.
3. **Prop_girls**: This feature maybe defined as the ratio of the number of girls born to the total number of children born for a particular year.
4. **Prop_boys**: This feature maybe defined as the ratio of the number of boys born to the total number of children born for a particular year.

The above new features will be calculated through the following code chunk below and the data has been tabulated in table 4.

```
df_present <- df_present %>% mutate(Total = boys + girls) %>%
  mutate(Ratio = girls/boys) %>%
  mutate(Prop_girls = girls/Total) %>%
  mutate(Prop_boys = boys/Total)

head(df_present) %>% kable(caption = "Addition of new features in the data for births between 1940-2002.") %>%
  kable_styling(bootstrap_options = c("bordered","hover"),
                                  latex_options = "HOLD_position")
```

Table 4: *Addition of new features in the data for births between 1940-2002.*

year	boys	girls	Total	Ratio	Prop_girls	Prop_boys
1940	1211684	1148715	2360399	0.9480318	0.4866614	0.5133386
1941	1289734	1223693	2513427	0.9487949	0.4868624	0.5131376
1942	1444365	1364631	2808996	0.9447965	0.4858074	0.5141926
1943	1508959	1427901	2936860	0.9462822	0.4861999	0.5138001
1944	1435301	1359499	2794800	0.9471874	0.4864387	0.5135613
1945	1404587	1330869	2735456	0.9475162	0.4865255	0.5134745

3.5 Query 5

This section shall aim to understand the difference between the number of boys and girls born during the period between 1940-2002 in The United States. For gaining in-depth insights, the proportionality of male births will be visualised. The code chunk below provides the visualisation technique to be used for understanding the change in child sex-ratio during the period of interest.

```
pl6 <- ggplot(data = df_present,aes(x = year, y = Prop_boys)) +
  geom_line(size=1.5,alpha=0.6) + ylim(0.475,0.53) + theme_classic() + geom_point(color = "darkred",size=1) +
  geom_hline(yintercept = 0.5,
             color = 'blue',
             linetype = "twodash",
             size = 2,alpha=0.6) + labs(x = "Year", y= "Proportion of boys") +
  ggtitle("Proportion of boys in the US \n between 1940-2002") +
  theme(plot.title = element_text(hjust = 0.5)) +

  annotate(
    "text",
    x = 1980,
    y = 0.52,
    colour = "darkred",
    label = 'Higher proportion of boys born \n between 1940-2002',
    size = unit(2.5, "pt")
  ) + theme(aspect.ratio = 1)
pl6
```

Figure 4 suggests that the proportion of boys and consequently, the number of boys born in the United States during the period between 1940-2002 is **indeed greater than the proportion of girls born in**

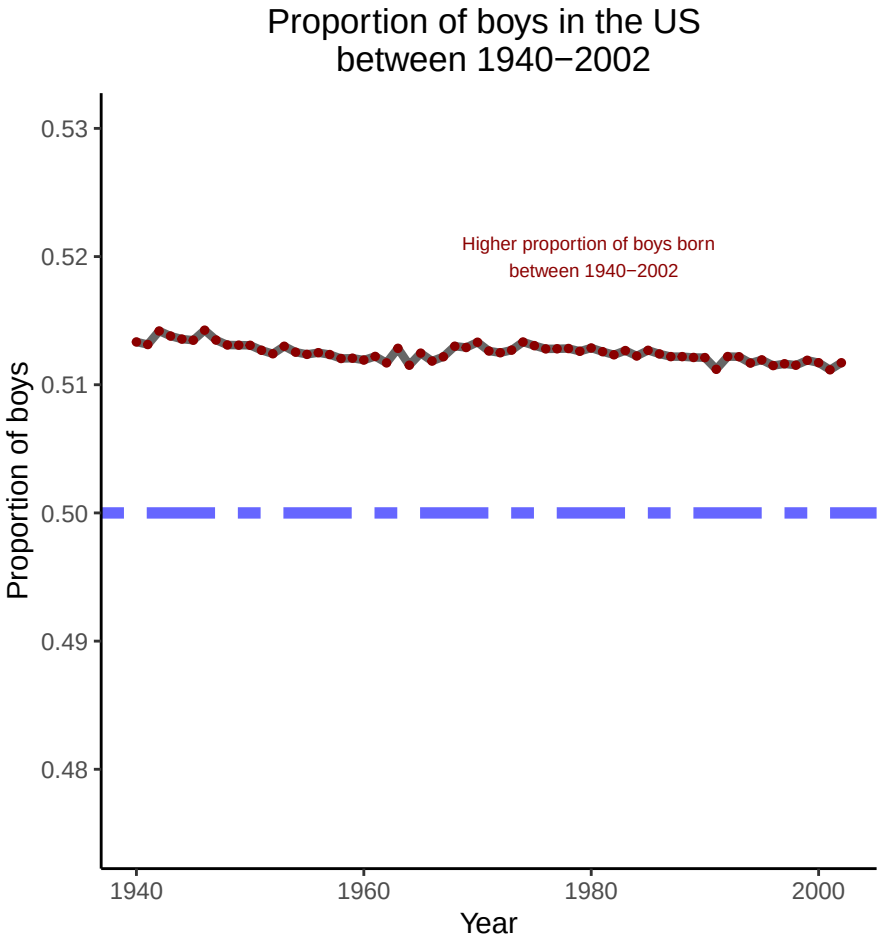

Figure 4: *Proportion of boys born between 1940-2002*

the same time period. As a result, the observations pertaining to the birth statistics of the United States between the period of 1940-2002 **closely corroborates** with the findings of Arbuthnot (1710) during the period of 1629-1710.

4 Conclusion

The key takeaways from the above analysis are as follows:

1. The magnitude of children born are significantly greater in the period of 1940-2002 when compared to the period of 1629-1710.
2. Statistics pertaining to female child birth were observed to undergo a sudden dip and then subsequently rise and stagnate for both the time periods. Hence, the two time periods were observed to report similarities in its trend.
3. Proportion of male child birth continue to be greater than female child birth in The United States during the period of 1940-2002. A similar observation was first reported by Arbuthnot (1710) for the period of 1629-1710.

5 Resources

The above analysis was undertaken with the help of the following software and packages:

1. **RStudio**: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
2. **ggplot2**: H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
3. **tidyverse**: Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.
4. **here**: Müller K (2020). *here: A Simpler Way to Find Your Files*. R package version 1.0.1, https://CRAN.R-project.org/package=here.
5. **knitr**: Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595.
6. **rmarkdown**: Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2023). *rmarkdown: Dynamic Documents for R*. R package version 2.23, https://github.com/rstudio/rmarkdown.
7. **cowplot**: Wilke C (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.1, https://CRAN.R-project.org/package=cowplot.
8. **kableExtra**: Zhu H (2023). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra

References

Arbuthnot, J (1710). II. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. By Dr. John Arbuthnott, Physitian in Ordinary to Her Majesty, and Fellow of the College of Physitians and the Royal Society. *Philosophical Transactions of the Royal Society of London* **27**(328), 186–190.

Mathews, T, BE Hamilton, et al. (2005). Trend analysis of the sex ratio at birth in the United States. *National vital statistics reports* **53**(20), 1–17.

Pryzgoda, J and JC Chrisler (2000). Definitions of gender and sex: The subtleties of meaning. *Sex roles* **43**, 553–569.

Ritchie, H and M Roser (2019). Gender Ratio. *Our World in Data*. https://ourworldindata.org/gender-ratio.