STT 5811 Lab01: Data Basics and Study Design

SOLUTIONS

Tuesday, January 27, 2026 @ 10:54 PM

Contents

Instructions

This data lab is adapted from the **Intro to R - Birth Rates** lab in **Section 3.4** of the IMS2E textbook. Read through the textbook's version of the lab (you can skip the videos), then complete your analysis here. Use complete sentences when answering questions.

Packages Used

Load the `tidyverse` and `openintro` packages.

```
library(tidyverse)
library(openintro)
```

Arbuthnot Dataset

Dataset Description

The `arbuthnot` dataset in the `openintro` package comes from the work of Dr. John Arbuthnot (1667-1735), an 18th century physician, writer, and mathematician. He investigated the ratio of newborn boys to girls, because he believed boys and girls were not born in equal numbers. If that was shown to be true beyond what was explainable by random chance, he asserted it would demonstrate "divine providence." This study is one of the earliest extant examples of a researcher using real-world data and probability to test a hypothesis. For data, Arbuthnot gathered baptismal records for babies born in London between 1629 and 1710.

Arbuthnot, John. (1710). An argument for Devine Providence, taken from the constant Regularity observ'd in the Births of both Sexes. *Philosophical Transactions*, 27, 186-190. **https://doi.org/10.1098/rstl.1710.0011**

Explore the Dataset

Assign the `arbuthnot` dataset to an object named `arbuthnot` in your Environment.

```
arbuthnot <- arbuthnot
```

Use `glimpse()` to explore the dataset.

```
glimpse(arbuthnot)
```

```
Rows: 82
Columns: 3
$ year  <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1639~
$ boys  <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, 5366~
$ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, 4784~
```

Use `head()` to print out the first five rows of the dataset.

```
head(arbuthnot, n = 5)
```

```
# A tibble: 5 x 3
   year  boys girls
  <int> <int> <int>
1  1629  5218  4683
2  1630  4858  4457
3  1631  4422  4102
4  1632  4994  4590
5  1633  5158  4839
```

Extract the column representing boys from the dataset using `$` and `sum()` the values.

```
sum(arbuthnot$boys)
```

```
[1] 484382
```

Extract the column representing girls from the dataset using `$` and `sum()` the values.

```
sum(arbuthnot$girls)
```

```
[1] 453841
```

QUESTIONS

1. How many observational units are there? What does each unit (row) represent?

   There are 82 observational units. Each row represents one year between 1629 and 1710 (inclusive).

2. How many variables are there? What variable type is each? What vector type?

   There are 3 variables. All of them are discrete quantitative variables and are stored as integers.

3. What piece of information does each variable tell you for each observation?

   The dataset contains annual counts of chidren who were baptized over a period of 82 years in Anglican churches in London. The variable `year` indicates what year the baptism counts come from; `boys` is the number of boys baptised that year and `girls` is the number of girls baptized.

4. What are the total numbers of boys and girls? Which total is larger?

   Overall, there are 484,382 boys and 453,841 girls. The total count of boys is larger by 30,541.

Create New Variables

NOTE: Your choice of variable names may differ from those below. In general, you should choose names that are at least somewhat descriptive and make sense in context. Use of either pipe operator is acceptable.

Create a new variable to represent total births. Reassign the dataset to save the change.

```
arbuthnot <- arbuthnot |>
  mutate(total = boys + girls)
```

Create a new variable to represent the ratio of boys to girls. Be sure to save the change.

```
arbuthnot <- arbuthnot |>
  mutate(ratiobg = boys/girls)
```

Create a new variable to represent the proportion of boys. Be sure to save the change.

```
arbuthnot <- arbuthnot |>
  mutate(propboys = boys/total)
```

Create a new variable to represent whether boys > girls. Be sure to save the change.

```
arbuthnot <- arbuthnot |>
  mutate(moreboys = boys > girls)
```

Use `glimpse()` to verify the new structure of your dataset with the added variables.

```
glimpse(arbuthnot)
```

```
Rows: 82
Columns: 7
$ year     <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1~
$ boys     <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917, 4703, 5359, 5~
$ girls    <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605, 4457, 4952, 4~
$ total    <int> 9901, 9315, 8524, 9584, 9997, 9855, 10034, 9522, 9160, 10311,~
$ ratiobg  <dbl> 1.114243, 1.089971, 1.078011, 1.088017, 1.065923, 1.044606, 1~
$ propboys <dbl> 0.5270175, 0.5215244, 0.5187705, 0.5210768, 0.5159548, 0.5109~
$ moreboys <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T~
```

QUESTIONS

1. How many observational units are there? Why does this make sense?

   There are still 82 observational units. This makes sense because we did not add any observations (years) to the dataset.

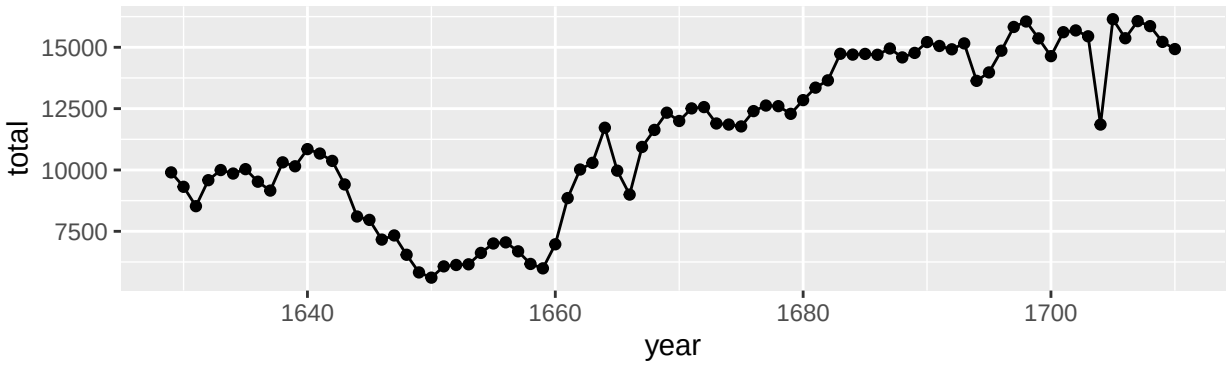
2. What are the variable and vector types for your new variables?

   Like the original three variables, the variable that represents the total number of children is a discrete numerical variable stored as an integer. This makes sense, since it is a sum of two integers. The variables that represent the ratio of boys to girls and the proportion of boys are both continuous numerical variables stored as numbers (double-precision floating-point number). This makes sense, because they both arise from the division of two numbers, the results of which may not be an integer. The variable that tells us whether there are more boys than girls is categorical (`TRUE` or `FALSE`) and is stored as a logic vector, because it results from a logical test.

Total Baptisms Over Time

Insert your total variable name to plot total number of children baptized over time.

```
ggplot(data = arbuthnot, aes(x = year, y = total)) +
  geom_line() +
  geom_point()
```



Use code to find and display the 10 years with the lowest total numbers of children.

```
arbuthnot |>
  arrange(total) |>
  head(n = 10)      # you can also use slice_head() here
```

```
# A tibble: 10 x 7
    year  boys girls total ratiobg propboys moreboys
   <int> <int> <int> <int>   <dbl>    <dbl> <lgl>
 1  1650  2890  2722  5612    1.06    0.515 TRUE
 2  1649  3079  2746  5825    1.12    0.529 TRUE
 3  1659  3209  2781  5990    1.15    0.536 TRUE
 4  1651  3231  2840  6071    1.14    0.532 TRUE
 5  1652  3220  2908  6128    1.11    0.525 TRUE
 6  1653  3196  2959  6155    1.08    0.519 TRUE
 7  1658  3157  3013  6170    1.05    0.512 TRUE
 8  1648  3363  3181  6544    1.06    0.514 TRUE
 9  1654  3441  3179  6620    1.08    0.520 TRUE
10  1657  3396  3289  6685    1.03    0.508 TRUE
```

QUESTIONS

1. Is there an apparent trend (or trends) over the years in the number of children that were baptized? Describe what you see in words. Are there any notable intervals of time or deviations?

   The annual number of baptized children generally rose between 1629 and 1710, from about 10,000 to about 15,000, but it was not a continual increase. Beyond the random variations we would expect in real-life data, there is a more sizable dip starting around 1640; the total roughly halved between 1640 and 1650. Baptisms did not return to a value near 1640 levels until after 1660.

2. What was happening in England between 1640 and 1660 that might explain what the graph shows?

   Civil wars, the execution of King Charles I, and a governmental shift to a Puritan Commonwealth (later dictatorship) under Oliver Cromwell destabilized English society and the Anglican church. There were likely fewer babies born, and when they were, some parents probably delayed baptism. The monarchy and the Anglican church returned when Charles II was crowned in 1660.

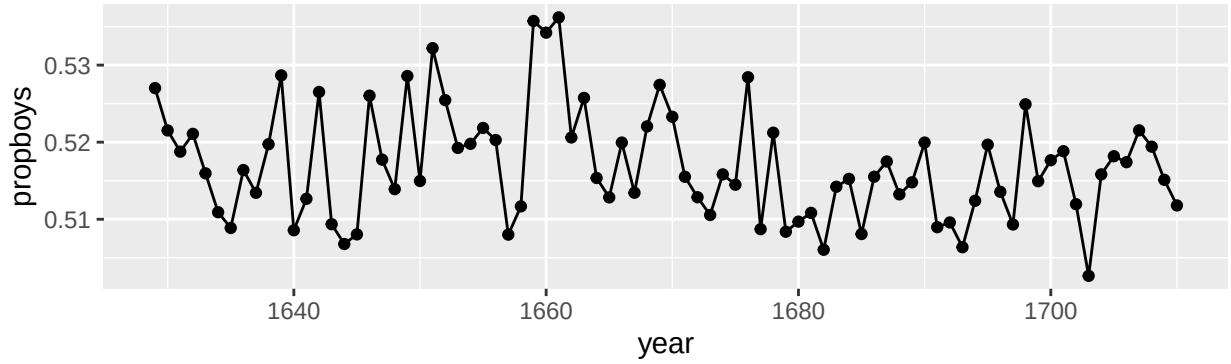
3. What major event happened in England in 1703 that might explain the data point for 1704?

   In November of 1703, an extratropical cyclone known as the "Great Storm" struck England. It was the most destructive storm in British history, killing thousands of people and destroying countless homes, as well as hundreds of civilian and naval ships. London was particularly affected by widespread flooding, collapsed chimneys, and torn-off roofs. This undoubtedly impacted both the number of children born and baptized in London during the subsequent year 1704.

Comparing Boys and Girls

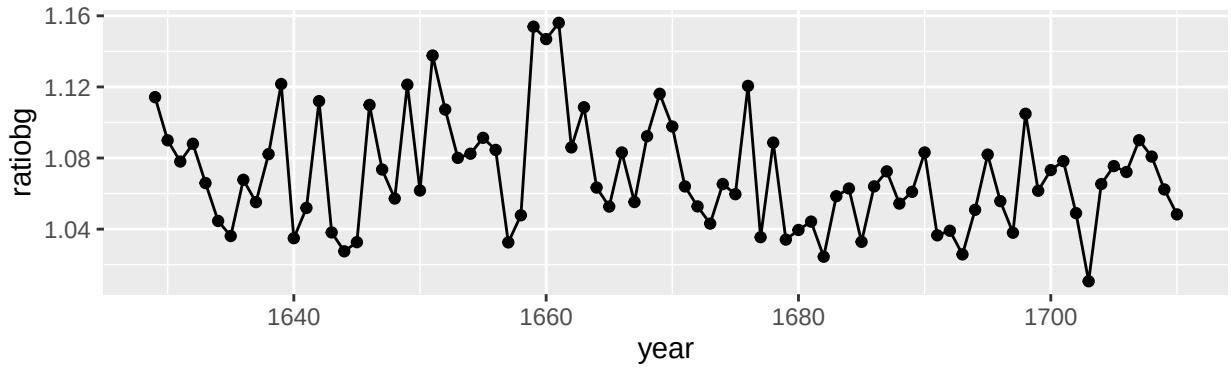
Create a plot of the proportion of boys over time using both points and a line.

```
ggplot(data = arbuthnot, aes(x = year, y = propboys)) +
  geom_line() +
  geom_point()
```



Create a plot of the ratio of boys to girls over time using both points and a line.

```
ggplot(data = arbuthnot, aes(x = year, y = ratiobg)) +
  geom_line() +
  geom_point()
```



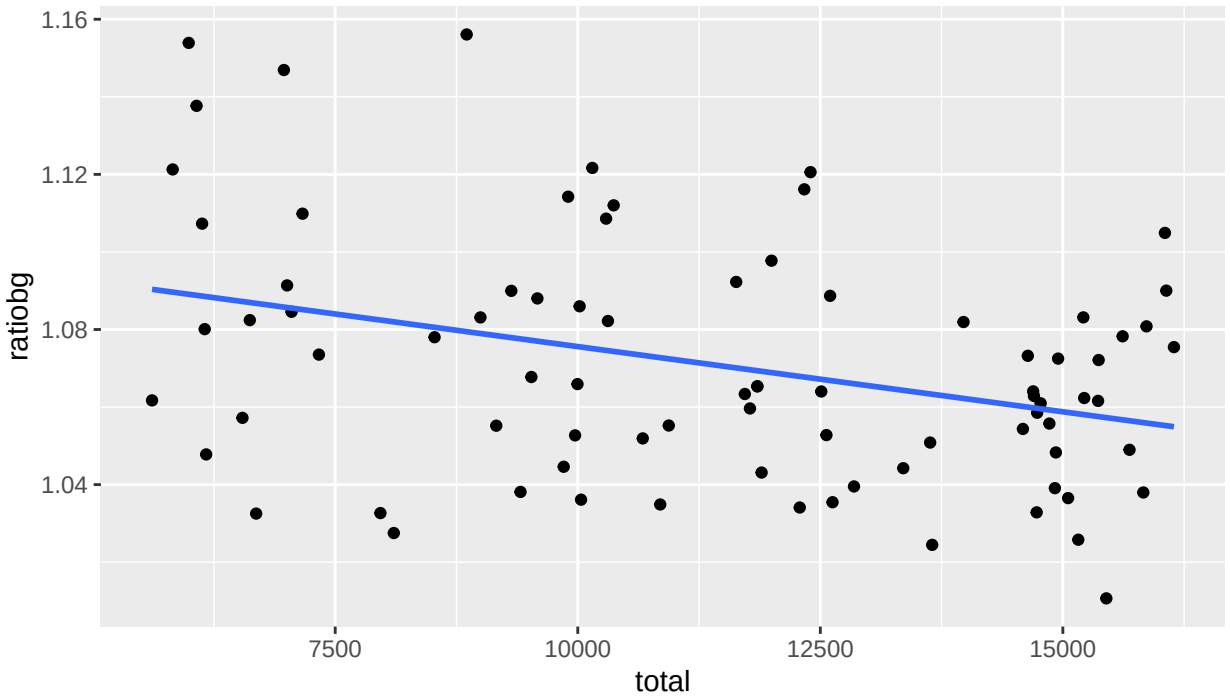
Use `count()` to find out how many observed years had a larger number of boys than girls.

```
arbuthnot |>
  count(moreboys)
```

```
# A tibble: 1 x 2
  moreboys     n
  <lgl>    <int>
1 TRUE        82
```

Insert variables names to make a scatterplot of the ratio of boys as a function of total.

```
ggplot(data = arbuthnot, aes(x = total, y = ratiobg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



NOTE: The following two bits of code were not required, but I used the values in my discussion of the plots. You can also estimate values of data points from the plots themselves.

```
# minimum proportion of boys for all years
arbuthnot |> arrange(propboys) |> head(n = 1)
```

```
# A tibble: 1 x 7
   year  boys girls total ratiobg propboys moreboys
  <int> <int> <int> <int>   <dbl>    <dbl> <lgl>
1  1703  7765  7683 15448    1.01    0.503 TRUE
```

```
# maximum proportion of boys for all years
arbuthnot |> arrange(desc(propboys)) |> head(n = 1)
```

```
# A tibble: 1 x 7
   year  boys girls total ratiobg propboys moreboys
  <int> <int> <int> <int>   <dbl>    <dbl> <lgl>
1  1661  4748  4107  8855    1.16    0.536 TRUE
```

QUESTIONS

1. How do the analyses support the idea that more boys than girls were born between 1629 and 1710? Explain, using information from your plots and counts to support your answer.

   There were 30,541 more boys overall. For every year in the dataset, more boys than girls were baptized, though the margin is sometimes slim. The ratio of boys to girls varies from 1.01 to 1.16. The proportion of baptisms that were boys varies from 0.503 (50.3%) to 0.536 (53.6%). However, Arbuthnot assumed that baptisms were a good proxy for births, which may not be true.

2. Does the scatterplot suggest that boys-to-girls ratio might be related to the total number of babies? Explain, using information from the plot to support your answer.

   The trend line has a negative slope, which suggests that when there is a larger number of baptisms, the ratio of boys to girls is smaller. However, there is also a fairly wide scattering (variability) of points around the trend line in the Y direction, so the relationship is not strong.

Think and Reflect on the Data

QUESTIONS

1. Arbuthnot used baptismal records rather than birth records, which would not have existed at the time, and only from Anglican churches in London. How might this impact his data, with respect to what he was trying to measure and investigate (i.e., is his sample biased)?

   There are several possible sources of bias. Arbuthnot was interested in births for people in general, but he used a convenience sample that was restricted to only London and to children baptized in Anglican churches (though Anglicans tended to be a majority, especially post-1660). He would have collected data by hand, so he may have missed some records or recorded some information incorrectly. The church records themselves may also have had errors. Finally, he assumed that baptisms accurately reflected births, though some babies would have died before being baptized.

   NOTE: Historically, infant mortality is higher among boys, and there is usually a delay between birth and baptism, so these data might *underestimate* male births. While this sample is sub-optimal in many ways, it was probably the only feasible method in Arbuthnot's time.

2. In his paper, Arbuthnot described his methodology: "Let there be a Die of Two sides, M and F... to find all the Chances of any determinate Number of such Dice, let the Binome M + F be raised to the Power, whose Exponent is the Number of Dice given... For Example, in Two Dice of Two sides M + F, the Chances are $M^2 + 2MF + F^2$, that is, One Chance for M double, One for F double, and Two for M single and F single..." M is defined as observing more boys than girls in a given year, while F is more girls than boys. What probability model does this sound like? If boys and girls are truly equally likely, how could you use this model to calculate the probability of the ratios you saw in Arbuthnot's data? You do not have to do the calculation, just describe how you would.

   This sounds like a binomial distribution with parameters $n = 82$, because each year would be one trial, and $p = 0.5$, because Arbuthnot used a "two-sided die" (coin flip) to model equal chances of boys vs. girls. We would have to assume that all the years are independent. The random variable X is the number of years in which boys outnumber girls, and Arbuthnot's outcome is $x = 82$ successes. We can solve for the probability using the binomial pdf.

   NOTE: You did not have to write out the pdf or solve for the answer, but the solution is below.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \rightarrow P(X = 82) = \binom{82}{82} 0.5^{82}(1-0.5)^{82-82} = 0.5^{82} = 2.1 \times 10^{-25}$$

```
# dbinom(x, n, p)
dbinom(82, 82, 0.5)
```

```
[1] 2.067952e-25
```

Session Information

Reproducibility in statistics means being able to get the same results using the same data, code, and methods as the original researchers. It is a way ensure scientific transparency and verifiability. Session information is a snapshot of the `R` working environment you used. This helps with reproducibility and code debugging, since software evolves. Even with the same data and methods, code and analyses implemented using older or newer versions of `R` and its packages may behave differently and give different results.

```
sessionInfo()
```

```
R version 4.5.2 (2025-10-31)
Platform: x86_64-redhat-linux-gnu
Running under: Red Hat Enterprise Linux 9.7 (Plow)
```

```
Matrix products: default
BLAS/LAPACK: FlexiBLAS OPENBLAS-OPENMP;  LAPACK version 3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] openintro_2.5.0     usdata_0.3.1        cherryblossom_0.1.0
 [4] airports_0.1.0      lubridate_1.9.4     forcats_1.0.1
 [7] stringr_1.5.2       dplyr_1.1.4         purrr_1.1.0
[10] readr_2.1.5         tidyr_1.3.1         tibble_3.3.0
[13] ggplot2_4.0.0       tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] generics_0.1.4     stringi_1.8.7      lattice_0.22-7     hms_1.1.4
 [5] digest_0.6.37      magrittr_2.0.4     evaluate_1.0.5     grid_4.5.2
 [9] timechange_0.3.0   RColorBrewer_1.1-3 fastmap_1.2.0      Matrix_1.7-4
[13] tinytex_0.57       mgcv_1.9-3         scales_1.4.0       cli_3.6.5
[17] rlang_1.1.6        splines_4.5.2      withr_3.0.2        yaml_2.3.10
[21] tools_4.5.2        tzdb_0.5.0         vctrs_0.6.5        R6_2.6.1
[25] lifecycle_1.0.4    pkgconfig_2.0.3    pillar_1.11.1      gtable_0.3.6
[29] glue_1.8.0         xfun_0.53          tidyselect_1.2.1   rstudioapi_0.17.1
[33] knitr_1.50         dichromat_2.0-0.1  farver_2.1.2       htmltools_0.5.8.1
[37] nlme_3.1-168       rmarkdown_2.30     labeling_0.4.3     compiler_4.5.2
[41] S7_0.2.0
```