Introduction to Statistical Inference (QTM 100 Lab)

Lecture 1: Introduction to Data

Justin Eloriaga — Emory University

Fall 2024

Gameplan

Getting Started with Data

Working with Data

Plotting Data

Getting Started with Data

Dr. Arbuthnot's Baptism Records

- Dr. John Arbuthnot, 18th century all arounder.
- Gathered a dataset with baptism records for children born in London every year from 1629 to 1710

Downloading the Dataset

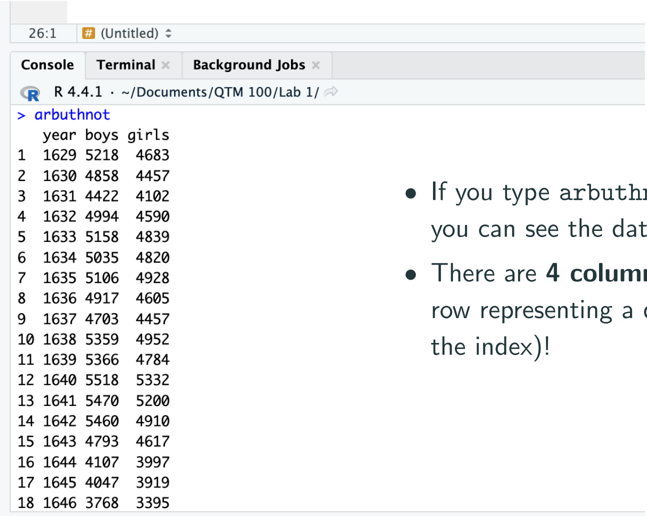
- You will need to download the dataset `arbuthnot.csv` from Canvas and load it
- To import this dataset, you need to use one of the methods discussed (point and click or working directory)

```
setwd("YourFilePath")

arbuthnot <- read.csv("arbuthnot.csv", header = TRUE)
```

- The variable's name (data frame) can be anything. It is not necessary to be `arbuthnot`.

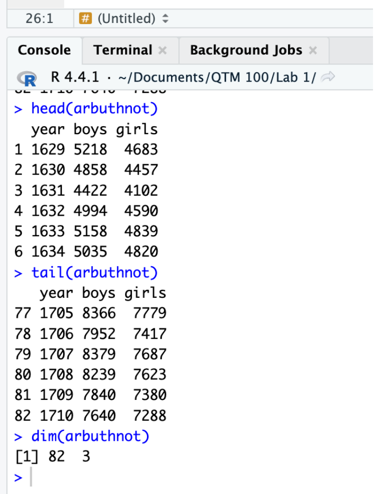
Looking at the data



```
> arbuthnot
   year boys girls
1  1629 5218  4683
2  1630 4858  4457
3  1631 4422  4102
4  1632 4994  4590
5  1633 5158  4839
6  1634 5035  4820
7  1635 5106  4928
8  1636 4917  4605
9  1637 4703  4457
10 1638 5359  4952
11 1639 5366  4784
12 1640 5518  5332
13 1641 5470  5200
14 1642 5460  4910
15 1643 4793  4617
16 1644 4107  3997
17 1645 4047  3919
18 1646 3768  3395
```

- If you type `arbuthnot` in the console, you can see the data.
- There are **4 columns** of numbers, each row representing a different year (+ the index)!

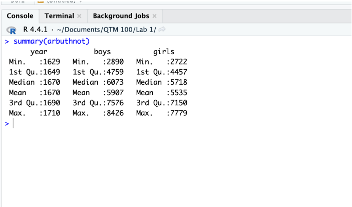
`head()`, `tail()`, and `dim()`



```
> head(arbuthnot)
  year boys girls
1 1629 5218  4683
2 1630 4858  4457
3 1631 4422  4102
4 1632 4994  4590
5 1633 5158  4839
6 1634 5035  4820
> tail(arbuthnot)
   year boys girls
77 1705 8366  7779
78 1706 7952  7417
79 1707 8379  7687
80 1708 8239  7623
81 1709 7840  7380
82 1710 7640  7288
> dim(arbuthnot)
[1] 82  3
>
```

- We use the `head()` function to display the **first** six entries.
- We use the `tail()` function to display the **last** six entries
- We use the `dim()` function to display the dimensions of the data. In this case, it has **82** rows and **3** columns.

The `summary()` function

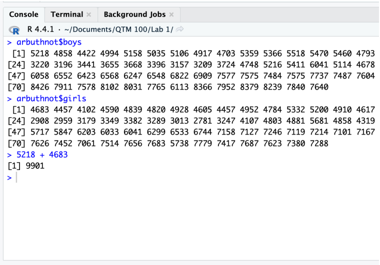


```
> summary(arbuthnot)
      year           boys           girls     
 Min.   :1629   Min.   :2890   Min.   :2722  
 1st Qu.:1649   1st Qu.:4759   1st Qu.:4457  
 Median :1670   Median :6073   Median :5718  
 Mean   :1670   Mean   :5907   Mean   :5535  
 3rd Qu.:1690   3rd Qu.:7576   3rd Qu.:7150 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

- We use the `summary()` function to get a quick overview of the dataset. Particularly, this gives us a notion of *range*!

Working with Data

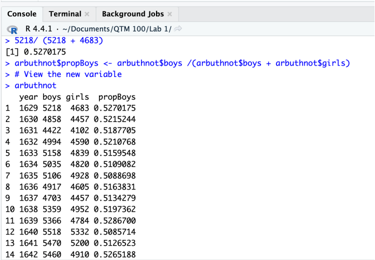
Selecting a column



```
> arbuthnot$boys
 [1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460 4793
[24] 3220 3196 3441 3655 3668 3396 3157 3209 3724 4748 5216 5411 6041 5114 4678
[47] 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575 7484 7575 7737 7487 7604
[70] 8426 7911 7578 8102 8031 7765 6113 8366 7952 8379 8239 7840 7640
> arbuthnot$girls
 [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910 4617
[24] 2908 2959 3179 3349 3382 3289 3013 2781 3247 4107 4803 4881 5681 4858 4319
[47] 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127 7246 7119 7214 7101 7167
[70] 7626 7452 7061 7514 7656 7683 5738 7779 7417 7687 7623 7380 7288
> 5218 + 4683
[1] 9901
> 
```

- We use the \$ to select a specific column from a data frame. It allows us to access columns from a data frame.
- `arbuthnot$boys` and `arbuthnot$girls` selects the column for boys and girls respectively
- What is the first number? A: The births in 1629 (the first year in the dataset)
- $5218 + 4683 = 9901$ → total births in 1629!

Working with Proportions and Variable Assignment

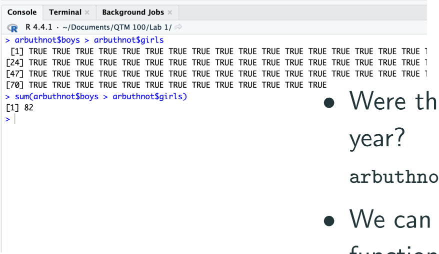


```
> 5218/ (5218 + 4683)
[1] 0.5270175
> arbuthnot$propBoys <- arbuthnot$boys /(arbuthnot$boys + arbuthnot$girls)
> # View the new variable
> arbuthnot
   year boys girls  propBoys
1  1629 5218  4683 0.5270175
2  1630 4858  4457 0.5215244
3  1631 4422  4102 0.5187705
4  1632 4994  4590 0.5210768
5  1633 5158  4839 0.5159548
6  1634 5035  4820 0.5109082
7  1635 5106  4928 0.5088698
8  1636 4917  4605 0.5163831
9  1637 4703  4457 0.5134279
10 1638 5359  4952 0.5197362
11 1639 5366  4784 0.5286700
12 1640 5518  5332 0.5085714
13 1641 5470  5200 0.5126523
14 1642 5460  4910 0.5265188
```

- What is the proportion of boys to the total number of births in 1629? A: 0.5270175
- We can go further! We can create a new column which is the proportion of boys to the total of births for each year. Use the command below!
- **Keep in mind the order of operations!!**

```
arbuthnot$propBoys <- arbuthnot$boys /(arbuthnot$boys + arbuthnot$girls)
```

Making Comparisons



```
> arbuthnot$boys > arbuthnot$girls
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE T
[24] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE T
[47] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE T
[70] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> sum(arbuthnot$boys > arbuthnot$girls)
[1] 82
> 
```

- We can compare two variables using some logical operators

$$>, <, ==, <=, >=$$

- Were there more boys or girls born in each year?

  `arbuthnot$boys > arbuthnot$girls`

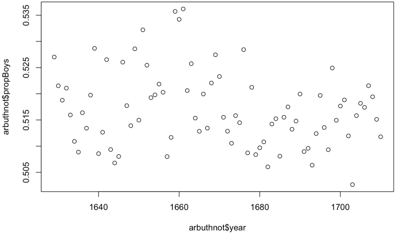
- We can view this another way using the `sum()` function

  `sum(arbuthnot$boys > arbuthnot$girls)`

- What do you observe?

Plotting Data

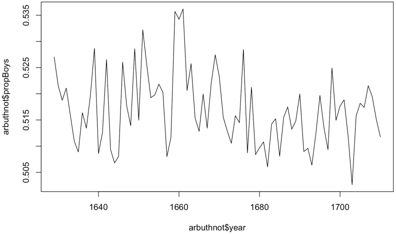
The `plot()` function



- What is the proportion of male baptisms, and does it vary by year?
- We can answer this question using a scatterplot!
- By default `R`, using the `plot` command, will generate a scatter plot

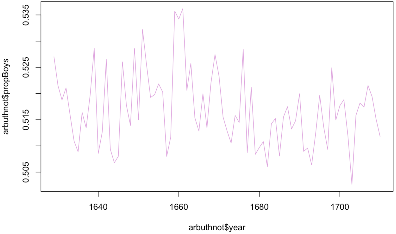
  ```
  plot(x = arbuthnot$year, y =
  arbuthnot$propBoys
  ```
- Do you notice anything?

Connecting with lines



- If we modify the command to `plot(x = arbuthnot$year, y = arbuthnot$propBoys, type ="l")`, we can connect the points with lines!
- When do you know whether you can add another argument like `type`? You can use the `?plot` command (or the question mark to be specific!)

Changing colors



- We can modify the code to change many things, one of which is the color! There are many ways to do that though...

```
plot(x = arbuthnot$year, y =
arbuthnot$propBoys, type
="l",col ="2" )
```

- You can even be more specific

```
plot(x = arbuthnot$year, y =
arbuthnot$propBoys, type
="l",col ="plum" )
```