BSRP Problem Set 1

your-name-here

Introduction

The main goal of this problem set is to briefly look at the metadata from the Cancer Cell Line Encyclopedia (CCLE). Please read the press release of CCLE before starting this problem set!

Loading the metadata in

Place the metadata file in the same directory as this document. Load it in with the following code chunk below. Notice that we access this file from a relative path, not an absolute path (don't worry if you don't know what that means).

Problem 1

How many cell lines are in this dataset?

```
## [1] 1804   24
```

There are 1804 cell lines.

Look at the column `primary_disease`, and create a frequency table using `table` function. What is the most common cancer type profiled, and what is the least common cancer type profiled?

```
##
##             Adrenal Cancer           Bile Duct Cancer
##                          1                         36
##             Bladder Cancer               Bone Cancer
##                         39                         75
##               Brain Cancer             Breast Cancer
##                        107                         82
##            Cervical Cancer    Colon/Colorectal Cancer
##                         22                         83
##           Embryonal Cancer Endometrial/Uterine Cancer
##                          3                         39
##                 Engineered          Esophageal Cancer
##                         14                         38
##                 Eye Cancer                Fibroblast
##                          9                         43
##         Gallbladder Cancer             Gastric Cancer
##                          7                         49
##       Head and Neck Cancer              Kidney Cancer
```

```
##                      76                      56
##                Leukemia              Liposarcoma
##                     132                      11
##            Liver Cancer              Lung Cancer
##                      27                     273
##                Lymphoma                 Myeloma
##                     109                      34
##          Neuroblastoma             Non-Cancerous
##                      46                       5
##          Ovarian Cancer         Pancreatic Cancer
##                      74                      59
##         Prostate Cancer                Rhabdoid
##                      13                      20
##                Sarcoma              Skin Cancer
##                      42                     113
##                Teratoma            Thyroid Cancer
##                       1                      21
##                 Unknown
##                      45
```

Lung cancer. Adrenal cancer and teratoma.

Problem 2

What is the mean age of the patient whose original sample was used for establish the cell line? Add the argument `na.rm = T` to the `mean()` function so that it removes any missing values `NA` before making the calculation.

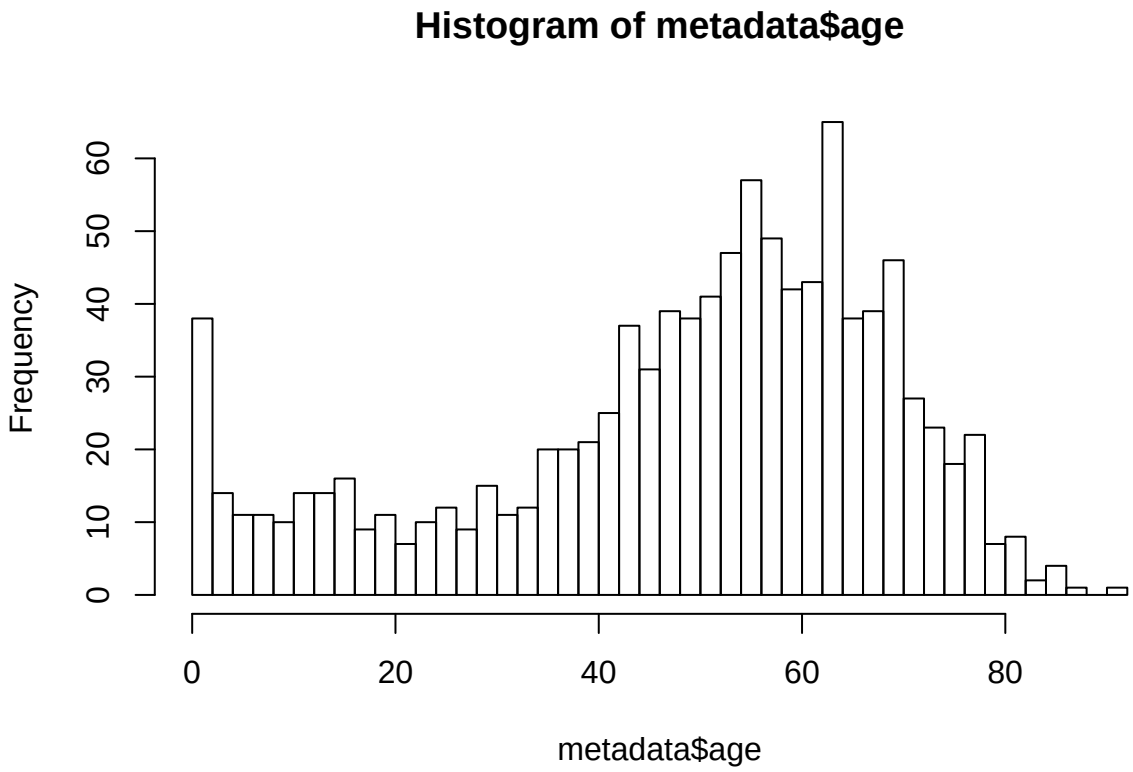
```
## [1] 48.582
```

Compute the median age also. What do you notice between the mean and median age? What does the difference say about the variability of age?

```
## [1] 53
```

Solution: The median is a bit larger than the mean. This suggests that there are outliers towards the younger population, shifting the mean to be smaller than the median.

Use the `hist()` function to create a histogram of age. You can toggle with the bin size by adding a second argument on the numbers of bins used, such as `hist(x, 50)`. Does this visualization help explain your previous answer?


Histogram of metadata$age
Frequency
0
10
20
30
40
50
60
0
20
40
60
80
metadata$age