STAT0002 Introduction to Probability and Statistics Weekly Exercises

Ge Li (Sunny)

November 3 2023

Contents

Preface

This document is used to document my assignments and weekly exercises of STAT0002.

Week 1

Exercise 1

1.1 Question 1

Import a set of data "days" from STAT0002 library. Calculate the five-number summary of these data.

```
library(stat0002)
# To give us a better understanding of the data imported:
sort(days)
```

```
##  [1]   31  199  491  881  895  967  989 1036 1260 1418 1427 1460 1460 1460 1460
## [16] 1460 1460 1460 1460 1460 1460 1460 1460 1461 1461 1461 1503 1655 1886 2027
## [31] 2039 2727 2810 2864 2921 2921 2921 2921 2921 2921 2922 2922 2922 2922 2922
## [46] 4422
```

```
fivenum(days)
```

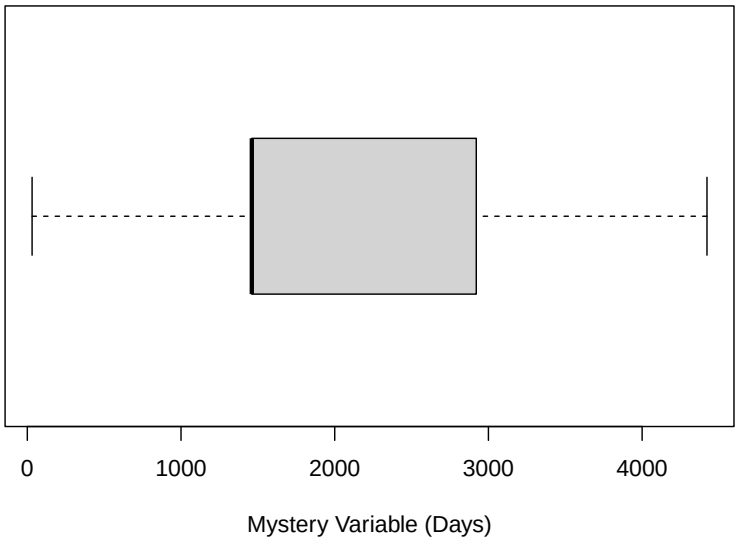
```
## [1]   31.0 1460.0 1460.5 2921.0 4422.0
```

Therefore, the five-number summary of the dataset is (31, 1460, 1460.5, 2921, 4422) days.

1.2 Question 2

Create a boxplot of these data.

```
boxplot(days, horizontal = T, xlab = "Mystery Variable (Days)")
```

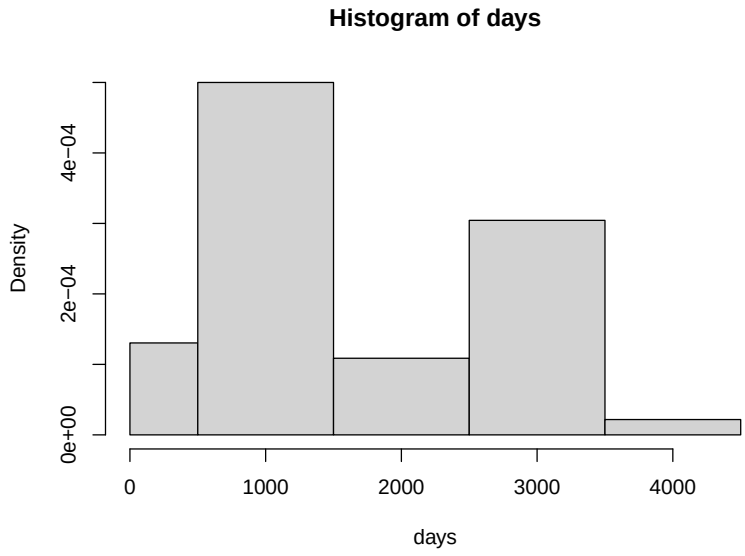


Combining the shape of this boxplot and the data summary we acquired above, the shape of the distribution of this dataset is slightly positively skewed. There are no outliers. All the data values are concluded in the boxplot. The boxplot visualizes the five-number summary as well as the location, shape, and spread of the distribution.

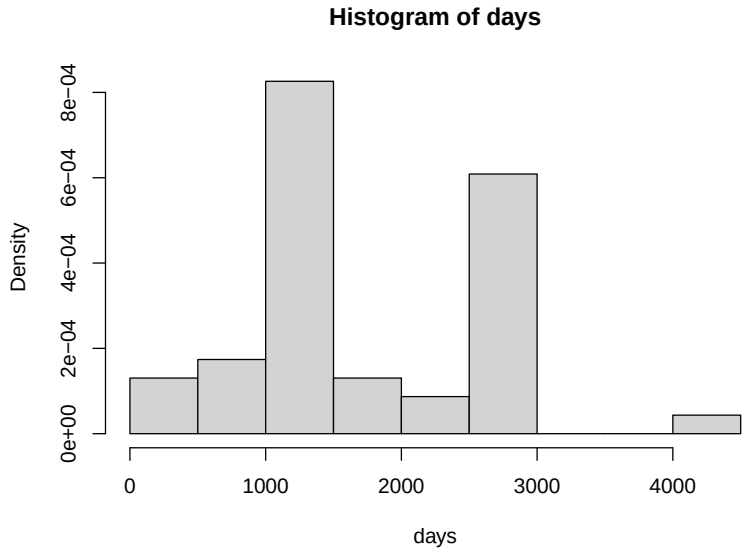
1.3 Question 3

Create histograms of these data with with different breaks.

```
hist(days, breaks = c(0, 500, 1500, 2500, 3500, 4500), freq = F)
```




```
hist(days, breaks = c(0, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500), freq = F)
```



While the first histogram shows us a rough shape and distribution, the second histogram is more detailed with more breaks. We can easily identify the modes of distribution using histograms and acquire a basic shape of the distribution as well.

1.4 Question 4

Create a stem-and-leaf plot of the data.

```
stem(days)
```

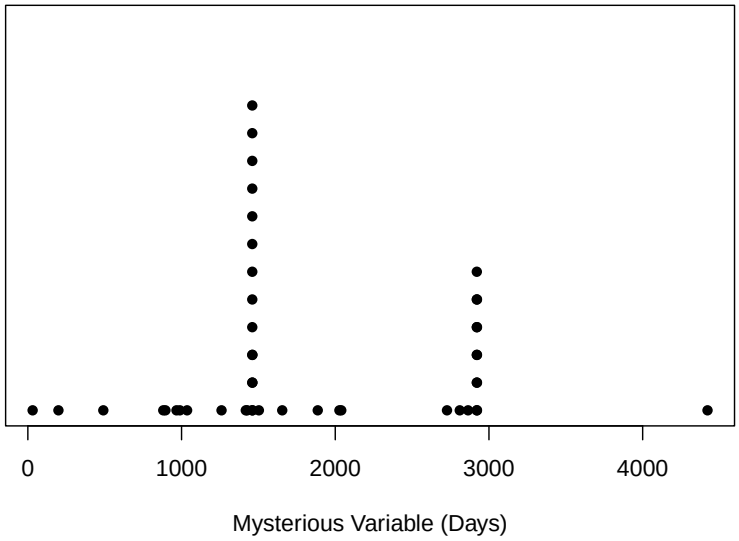
```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##   0 | 02
##   0 | 599
##   1 | 000344
##   1 | 555555555555555579
##   2 | 00
##   2 | 78999999999999
##   3 |
##   3 |
##   4 | 4
```

We could conclude from this plot that the shape of this dataset follows a bimodal distribution, where a lot of values in the data fall around 1500 and 2900.

1.5 Question 5

Create a dot plot of the data.

```
stripchart(days, method = "stack", at = 0, offset = 1, pch = 16, xlab = "Mysterious Va
```



The dot plot confirms our conclusion from the stem-and-leaf plot stated above.

1.6 Question 6

Find if there are any outliers in this dataset.

```
IQR <- IQR(days)
UpperBound <- 2921 + 1.5 * IQR
LowerBound <- 1460 - 1.5 * IQR
UpperBound
```

```
## [1] 5091.125
```

```
LowerBound
```

```
## [1] -710.125
```

All the values from the data fall in the range $\in [-710.125, 5091.125]$. Therefore, there aren't any outliers. Since 2922 days are approximately equivalent to 8 years, these data might represent the time it takes for people to get a Ph.D., the 8-year cycle of Venus, or the length of the presidency of the United States.

Week 2

Exercise 2

2.1 Question 1 - Location

(a) Sample Mean

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (ax_i + b) \\
&= \frac{1}{n}(a\frac{1}{n}\sum_{i=1}^{n} x_i + bn) \\
&= \frac{a}{n}\sum_{i=1}^{n} x_i + b \\
\therefore \bar{y} &= a\bar{x} + b
\end{aligned}
$$

(b) Sample Median: The order of the values in the dataset sorted will remain the same after the transformation. As a result, $y_{1/2} = ax_{1/2} + b$.
(c) Summary: The above results about sample mean and sample median are reasonable based on mathematical proof.