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)
```

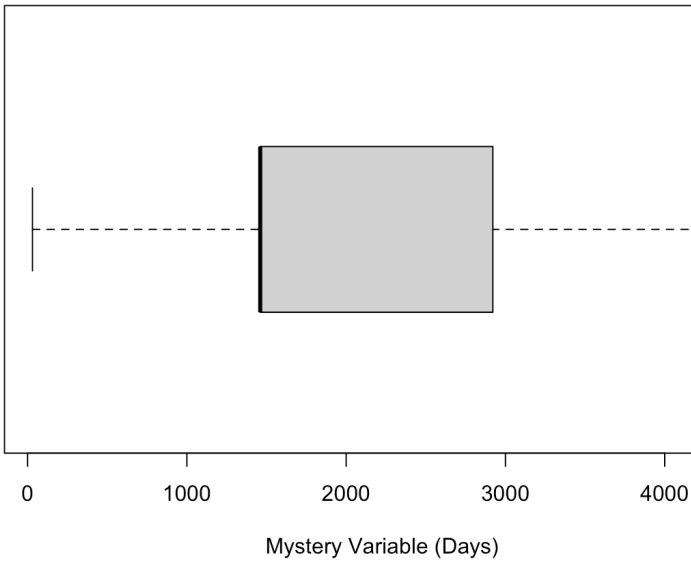
```
##  [1]   31  199  491  881  895  967  989 1036 1260 1418 1427 1460 1460 1460 1460
## [16] 1460 1460 1460 1460 1460 1460 1460 1460 1461 1461 1461 1503 1655 1886 2027
## [31] 2039 2727 2810 2864 2921 2921 2921 2921 2921 2921 2922 2922 2922 2922 2922
## [46] 4422
```

```
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

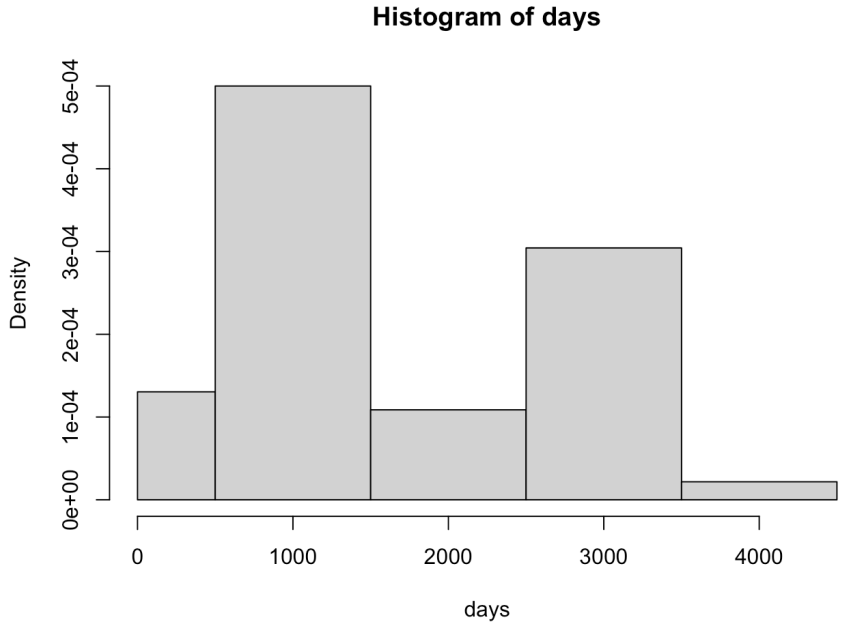


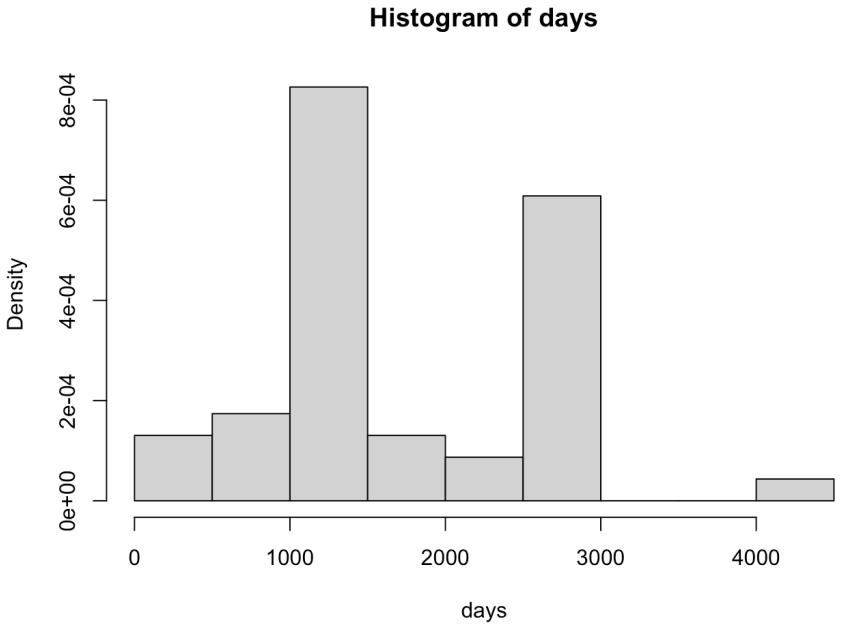
Create a boxplot of these data.

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.





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

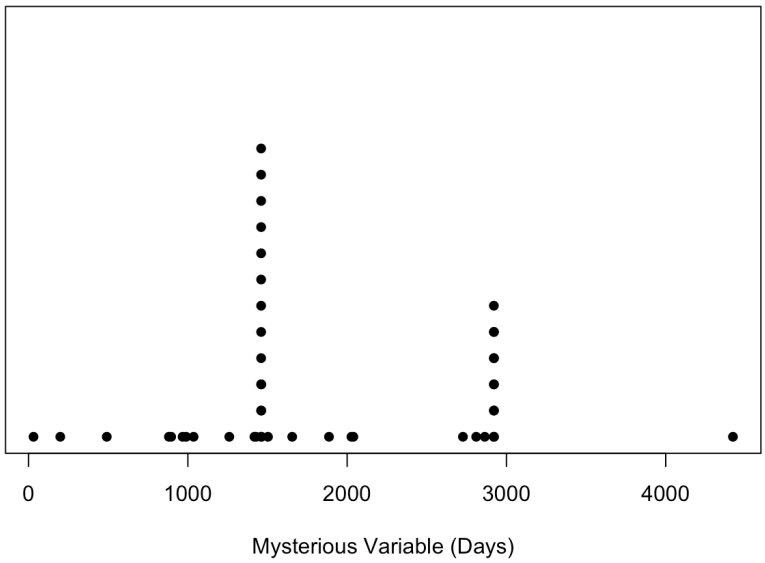
Create a stem-and-leaf plot of the data.

```
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.

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.

2.2 Question 2 - Spread

(a) Standard Deviation

$$
\begin{aligned}
S_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
S_y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2} \\
S_y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(ax_i+b-(a\bar{x}+b))^2} \\
&= \sqrt{\frac{a^2}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= a\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
\therefore S_y &= aS_x
\end{aligned}
$$

(b) Sample Interquartile Range

$$
\begin{aligned}
q_U^x - q_L^x &= x_{\frac{3}{4}(n+1)} - x_{\frac{1}{4}(n+1)} \\
q_U^y - q_L^y &= y_{\frac{3}{4}(n+1)} - y_{\frac{1}{4}(n+1)} \\
q_U^y - q_L^y &= ax_{\frac{3}{4}(n+1)} + b - (ax_{\frac{1}{4}(n+1)} + b) \\
q_U^y - q_L^y &= a(x_{\frac{3}{4}(n+1)} - x_{\frac{1}{4}(n+1)}) \\
\therefore q_U^y - q_L^y &= a(q_U^x - q_L^x)
\end{aligned}
$$

(c) Summary: The above results about sample interquartile range and standard deviation make sense based on mathematical proof.

2.3 Question 3 - Shape

(a) Standardized Sample Skewnewss

$$
\begin{aligned}
Skewness_x &= \frac{1}{n}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{S_x^3} \\
Skewness_y &= \frac{1}{n}\cdot\frac{\sum_{i=1}^{n}(y_i-\bar{y})^3}{S_y^3} \\
Skewness_y &= \frac{1}{n}\cdot\frac{a^3\cdot\sum_{i=1}^{n}(x_i-\bar{x})^3}{a^3\cdot S_x^3} \\
\therefore Skewness_y &= Skewness_x
\end{aligned}
$$

(b) Sample Quartile Skewness

$$Skewness_x = \frac{(q_U^x - m_x) - (m_x - q_L^x)}{q_U^x - q_L^x}$$
$$Skewness_y = \frac{(q_U^y - m_y) - (m_y - q_L^y)}{q_U^y - q_L^y}$$
$$Skewness_y = \frac{a((q_U^x - m_x) - (m_x - q_L^x))}{a(q_U^x - q_L^x)}$$
$$\therefore Skewness_y = Skewness_x$$

(c) Summary: The above results about standardized sample skewness and sample quartile skewness are correct based on mathematical proof.

2.4 Question 4 - About Sample Size

Explain the reason why these summary statstics require $n \geq 3$.
The precondition is $n \geq 3$ because when the sample size is too small, it's meaningless to investigate summary statistics. There's nothing to investigate about a set of variables with only one or two data. A smaller sample size also leads to increasing percentage uncertainty while calculating the data. Finally, $n \neq 1$ also because that the process of calculating standard deviation requires we use $(n-1)$ as a denominator, and 0 can't be a denominator.

Week 3

Exercise 3

Two fair dice, one red and one blue, are thrown. For each die: the possible scores are 1,2,3,4,5,6. Each score is equally likely, so each score has probability $\frac{1}{6}$. You may assume that the score on red die is independent of the score on the blue die.

We define the following events:
A: sum of the scores on the dice is 7.
B: the score on the red die is different from the score on the blue die.
C: the sum of the scores on the dice is 11.
D: the score on the red die is 6.

3.1 Question 1

(a) Calculate $P(A), P(B), P(C)$, and $P(D)$.

$$P(A) = (\frac{1}{6} \cdot \frac{1}{6}) \cdot 6 = \frac{1}{6} P(B) = \frac{5}{6} P(C) = (\frac{1}{6} \cdot \frac{1}{6}) \cdot 2 = \frac{1}{18} P(D) = \frac{1}{6}$$

(b) Calculate $P(A, B), P(A, C)$, and $P(A, D)$.

$$P(A, B) = P(A) = \frac{1}{6} P(A, C) = 0 P(A, D) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

(c) Calculate $P(AorB), P(AorC)$, and $P(AorD)$.

$$P(A \text{ or } B) = P(B) = \frac{5}{6} P(A \text{ or } C) = P(A) + P(C) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9} P(A \text{ or } D) = \frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} = \frac{11}{36}$$

(d) Calculate $P(B \mid A), P(C \mid A)$, and $P(D \mid A)$.

$$P(B \mid A) = \frac{P(A, B)}{P(A)} = 1 P(C \mid A) = \frac{P(A, C)}{P(A)} = 0 P(D \mid A) = \frac{P(A, D)}{P(A)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

3.2 Question 2

Which of the following pairs of events are independent and why?
(a) A and B.
$P(A, B) = \frac{1}{6} \neq P(A) \cdot P(B)$, $\therefore A$ is not independent of B.
Events A and B are not independent if A happens (the sum of the scores on 2 dice equal 7), B must happen since the score on the red die must be different from the score on the blue one.

(b) A and C.
$P(A, C) = 0$, $\therefore A$ is not independent of C.
Events A and C are not independent, they are mutually exclusive since these 2 events cannot happen at the same time.

(c) A and D.
$P(A, D) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$, $\therefore A \perp B$.
Events A and D are independent.

3.3 Question 3

Now let E be the event that at least one 6 is thrown and F the event that two 6s are thrown. Calculate:
(a) $P(F \mid D)$.

$$P(F \mid D) = \frac{P(F, D)}{P(D)} = \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6}} = \frac{1}{6}$$

(b) $P(F \mid E)$.

$$P(F \mid E) = \frac{P(F, E)}{P(E)} = \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} \cdot 2} = \frac{1}{11}$$

Week 4

Exercise 4

4.1 Question 1

B_1 and B_2 are mutually exclusive, exhaustive events such that $P(B_1) = 0.05$. A is an event such that $P(A \mid B_1) = 0.75$ and $P(A \mid B_2) = 0.50$.
(a) Use the law of total probability to calculate $P(A)$.

$P(A) = P(A \mid B_1)P(B_1)+P(A \mid B_2)P(B_2) = 0.75*0.05.+0.50*(1-0.05) = 0.5125$

(b) Use Bayes' Theorem to calculate $P(B_1 \mid A)$.

$$P(B_1 \mid A) = \frac{P(A \mid B_1) \cdot P(B_1)}{P(A)} = \frac{0.75 \cdot 0.05}{0.5125} = 0.07317$$

4.2 Question 2

In a population, 5% of people have high blood pressure. Of people with high blood pressure, 75% drink alcohol whereas only 50% of people without high blood pressure drink alcohol. What percentage of drinkers of alcohol have high blood pressure?

Let event B_1 represent people having high blood pressure in the population, whereas event B_2 represents people without high blood pressure in this population.

$$\therefore P(B_1) = 0.05,\ P(B_2) = 0.95.$$

Let event A represent people who drink alcohol. Of people with high blood pressure, 75% drink alcohol while only 50% of people without high blood pressure drink alcohol.

$$\therefore P(A \mid B_1) = 0.75,\ P(A \mid B_2) = 0.50.$$

Therefore, the percentage of alcohol drinkers having high blood pressure can be interpreted as $PA(B_1 \mid A)$.

$$P(B_1 \mid A) = \frac{P(A \mid B_1) \cdot P(B_1))}{P(A)} = \frac{0.75 \cdot 0.05}{0.5125} = 0.07317$$

4.3 Question 3

Suppose that A and B are two events with $P(A) > 0$ and $P(B) > 0$. Suppose, in addition, that $P(A \mid B) < P(A)$. Using Bayes' Theorem, or otherwise, show that $P(B \mid A) < P(B$. Interpret this result using words.

$$P(A \mid B) < P(A)$$

$$\frac{P(B \mid A) \cdot P(A)}{P(B)} < P(B)$$

$$\because P(A) > 0, P(B) > 0$$

$$\therefore \frac{P(B \mid A) \cdot P(B) \cdot P(A)}{P(B) \cdot P(A)} < \frac{P(A) \cdot P(B)}{P(A)}$$

$$\therefore P(B \mid A) < P(B)$$

Interpretation: the conditional probability will always be less than or at least equal to the total proabability of an event.

Week 5

Exercise 5

5.1 Question 1

Let X be a random variable with mean μ and (positive) variance σ^2. Find the mean and variance of the random variable $Y = \frac{X-\mu}{\sigma}$.
Mean:

$$\because Y = \frac{X - \mu}{\sigma}$$

$$\therefore X = \sigma \cdot Y + \mu$$

$$\therefore \mathbb{E}(X) = \sigma \cdot \mathbb{E}(Y) + \mu$$

$$\therefore \mathbb{E}(Y) = \frac{\mathbb{E}(X) - \mu}{\sigma}$$

$$\because \mathbb{E}(X) = \mu$$

$$\therefore \mathbb{E}(Y) = 0$$

Variance:

$$\because X = \sigma \cdot Y + \mu$$

$$\therefore \mathbb{V}\text{ar}(X) = \sigma^2 \cdot \mathbb{V}\text{ar}(Y)$$

$$\because \mathbb{V}\text{ar}(X) = \sigma^2$$

$$\therefore \mathbb{V}\text{ar}(Y) = 1$$

Therefore, the mean and variance of the random variable Y is 0 and 1.

5.2 Question 2

Let X be the number of heads in 2 independent tosses of a fair coin.
(a) Find $\mathbb{E}(\frac{1}{1+X})$.
Let random variable $Y = \frac{1}{1+X}$,
Possible outcomes:

x

0

1

2

y

1

$\frac{1}{2}$

$\frac{1}{3}$

y^2

1

$\frac{1}{4}$

$\frac{1}{9}$

$P(X = x)$

$\frac{1}{4}$

$\frac{1}{2}$

$\frac{1}{4}$

$P(Y = y)$

$\frac{1}{4}$

$\frac{1}{2}$

$\frac{1}{4}$

$$\therefore \mathbb{E}(\frac{1}{1+X}) = \sum(\frac{1}{1+x_i})P(X = x_i) = \mathbb{E}(Y) = 1 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{4} = \frac{7}{12}$$

(b) Show that $\mathbb{E}(\frac{1}{1+X}) \neq \frac{1}{1+\mathbb{E}(X)}$.

$$\mathbb{E}(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$\therefore \frac{1}{1+\mathbb{E}(X)} = \frac{1}{2}$$
$$\because \mathbb{E}(\frac{1}{1+X}) = \frac{7}{12}$$
$$\because \frac{7}{12} \neq \frac{1}{2}$$
$$\therefore \mathbb{E}(\frac{1}{1+X}) \neq \frac{1}{1+\mathbb{E}(X)}$$

(c) Find $\mathbb{V}\text{ar}(\frac{1}{1+X})$.

$$\mathbb{E}((\frac{1}{1+X})^2) = 1 \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{9} \cdot \frac{1}{4} = \frac{29}{72}$$
$$\because \mathbb{V}\text{ar}(\frac{1}{1+X}) = \mathbb{E}((\frac{1}{1+X})^2) - (\mathbb{E}(\frac{1}{1+X}))^2 = \frac{29}{72} - (\frac{7}{12})^2 = \frac{1}{16}$$

5.3 Question 3

Let X_H be be the number of heads and X_T be the number of tails obtained in 4 independent tosses of a fair coin. You are given that $\mathbb{E}(X_H) = 2$ and $\mathbb{V}\text{ar}(X_H) = 1$.
(a) Find $\mathbb{E}(X_T)$.

$$X_H + X_T = 4$$
$$X_T = 4 - E_H$$
$$\begin{aligned}\mathbb{E}(X_T) &= \mathbb{E}(4 - X_H) \\ &= 4 - \mathbb{E}(X_H) \\ &= 4 - 2\end{aligned}$$
$$\therefore \mathbb{E}(X_T) = 2$$

(b) Show that $X_H - X_T = 2 \cdot X_H - 4$.

$$\because X_H + X_T = 4$$
$$\therefore X_T = 4 - X_H$$
$$\therefore X_H - X_T = X_H - (4 - X_H)$$
$$\therefore X_H - X_T = 2 \cdot X_H - 4$$

(c) Hence, or otherwise, find $\mathbb{E}(X_H - X_T)$ and $\mathbb{V}\text{ar}(X_H - X_T)$.

$$\mathbb{E}(X_H - X_T) = \mathbb{E}(2 \cdot X_H - 4) = 2 \cdot \mathbb{E}(X_H) - 4 = 0$$
$$\mathbb{V}\text{ar}(X_H - X_T) = \mathbb{V}\text{ar}(2 \cdot X_H - 4) = 2^2 \cdot \mathbb{V}\text{ar}(X_H) = 4$$

Week 6

Exercise 6

6.1 Question 1

A machine makes components whose lengths must be between the specification limits of 6.45cm and 6.55cm. Let L denote the length of a randomly chosen componet made by the machine. Suppose that L has a normal distribution with mean μ and variance σ^2.

Based on measurements of a very large number of components it is found that 5% of components are longer than 6.55cm and 5% of components are shorter than 6.45cm.

(a) Find μ and σ^2

$$P(X > 6.55) = 1 - \Phi(\frac{6.55 - \mu}{\sigma}) = 0.05$$

$$P(X < 6.45) = \Phi(\frac{6.45 - \mu}{\sigma}) = 0.05$$

By inspecting the Normal Distribution Function table,

$$z = \frac{6.55 - \mu}{\sigma} = 1.6449$$

$$\therefore \mu = 6.55 - 1.6449\sigma$$

Substitute this relationship into the second equation,

$$\frac{6.45 - (6.55 - 1.6449\sigma)}{\sigma} = -1.6449$$

$$2 * 1.6449\sigma = 0.1$$

$$\sigma = 0.0304$$