Korea Advanced Institute of Science and Technology

Probability and Statistics

Homework 6

Professor:
Sung-Ho Kim

Course ID:
CC511

Student:
Federico Berto

ID number:
20204817

Exercise 5.4.3

Consider a sequence of random variables X i that are independently identic- ally distributed with a positive state space. Explain why the central limit theorem implies that the random variable

$$X = X_1 \times \ldots \times X_n \tag{1}$$

has approximately a lognormal distribution for large values of n.

Solution:

Let's consider the random variables Y_i, since the X_i are independently identically distributed:

$$Y_i = ln(X_i) \tag{2}$$

which are also identically distributed with the parameters μ and σ^2. For the central limit theorem we have that:

$$Y = Y_1 + Y_2 + \cdots + Y_n \simeq N(n\mu, n\sigma^2) \tag{3}$$

This yields

$$ln(X_1) + ln(X_2) + \cdots + ln(X_n) = ln(X_1 \times X_2 \times \ldots \times X_n) \tag{4}$$

thus,

$$Y = ln(X_1 \times X_2 \times \ldots \times X_n) \simeq N(n\mu, n\sigma^2) \tag{5}$$

Exercise 5.4.8

(a) There is a probability of 0.90 that a t random variable with 23 degrees of freedom lies between −x and x. Find the value of x.

Solution:

Since we know that:

$$1 - \alpha = P\left(\mid X \mid \leq t_{\frac{\alpha}{2},\nu}\right) = P\left(-t_{\frac{\alpha}{2},\nu} \leq X \leq t_{\frac{\alpha}{2},\nu}\right) \tag{6}$$

then, given $\alpha = 0.10$, then $x = t_{0.05,23} = 1.714$.

(b) There is a probability of 0.975 that a t random variable with 60 degrees of freedom is larger than y. Find the value of y.

Solution:

We have that

$$\alpha = P(X \geq t_{\alpha,\nu}) \tag{7}$$

Using the distribution's symmetry $t_{\alpha,\nu} = -t_{1-\alpha,\nu}$, we get the probability $y = -t_{0.025,60} = -2.000$.

(c) What is the probability that a chi-square random variable with 29 degrees of freedom takes a value between 19.768 and 42.557?

Solution:

The probability will be the following:

$$P(Y \leq 42.557) - P(Y \leq 19.768) \tag{8}$$

which yields $P = 0.84999269$.
Source code:
The source code for the calculations is the following:

```
from scipy.stats import t, chi2
print("Result (a): ", t.ppf(1-0.05, 23))
print("Result (b): ", t.ppf(1-0.975, 60))
print("Result (c): ", chi2.cdf(42.557, 29) - chi2.cdf(19.768, 29))
```

Exercise 5.4.12

Use your computer package to find the following critical points:

- (a) $\chi^2_{20.12,8}$
- (b) $\chi^2_{20.54,19}$
- (c) $\chi^2_{20.023,32}$

If the random variable X has a chi-square distribution with 12 degrees of freedom, use your computer package to find:

- (d) $P(X \leq 13.3)$
- (e) $P(9.6 \leq X \leq 15.3)$

Solution:

- (a) 12.77032873743455

- (b) 17.737971049053566
- (c) 49.859048172797955
- (d) 0.6523822453579444
- (e) 0.42556755784723643

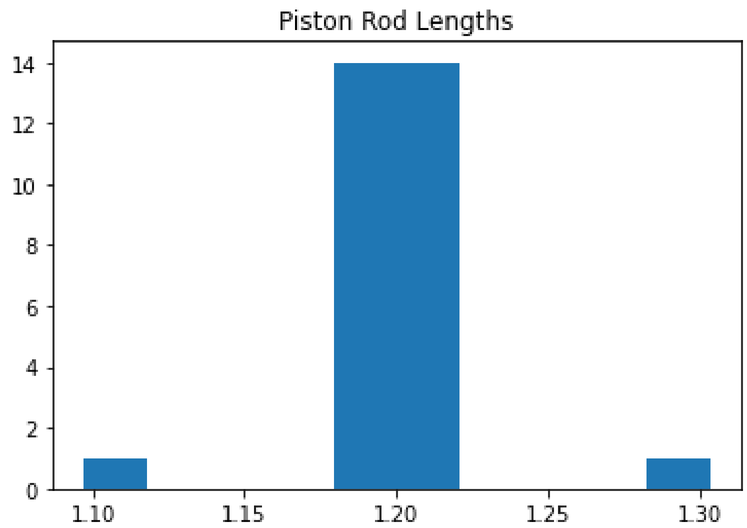
Source code:

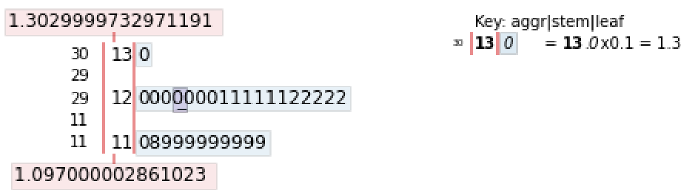
```
from scipy.stats import t, chi2
a = chi2.ppf(1-0.12,8)
b = chi2.ppf(1-0.54,19)
c = chi2.ppf(1-0.023,32)
d = chi2.cdf(13.3, 12)
e = chi2.cdf(15.3, 12) - chi2.cdf(9.6, 12)
```

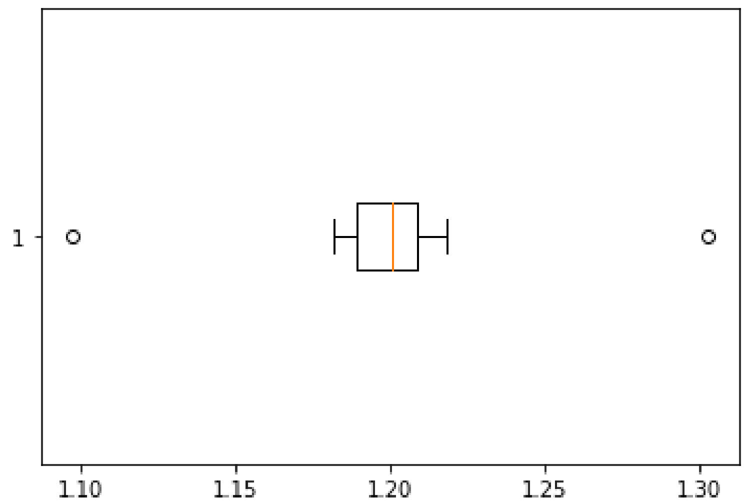
Exercise 6.2.3 Piston Rod Lengths

DS 6.2.3 shows the lengths of 30 piston rods. Construct a histogram of the data set with appropriate band widths. Do you think that there are any outliers in the data set?
Solution:

(Source code is after the exercises)
We plot down here are the histogram, box plot and steam leaf plots in this order:







The main outliers are, as we can also see from the box plot, the following: 1.097, 1.303.

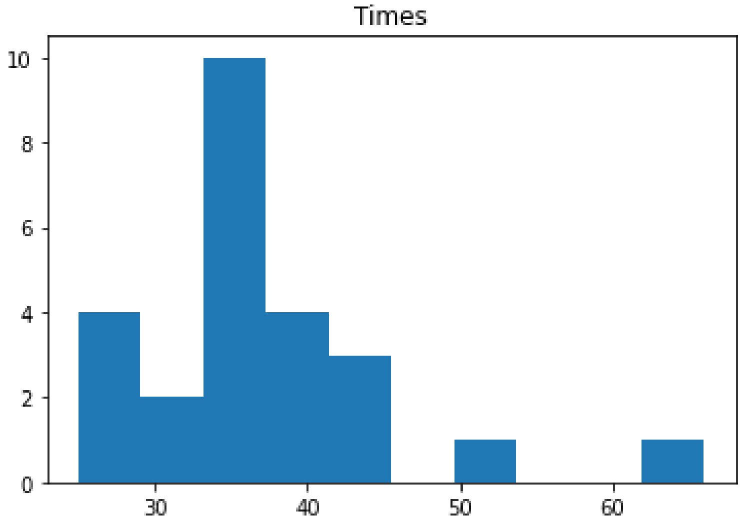
Exercise 6.2.4 Physical Training Course Completion Times

DS 6.2.4 shows the times taken by 25 students to finish a physical training course. Construct a histogram of the data set with appropriate band widths. Do you think that there are any outliers in the data set?

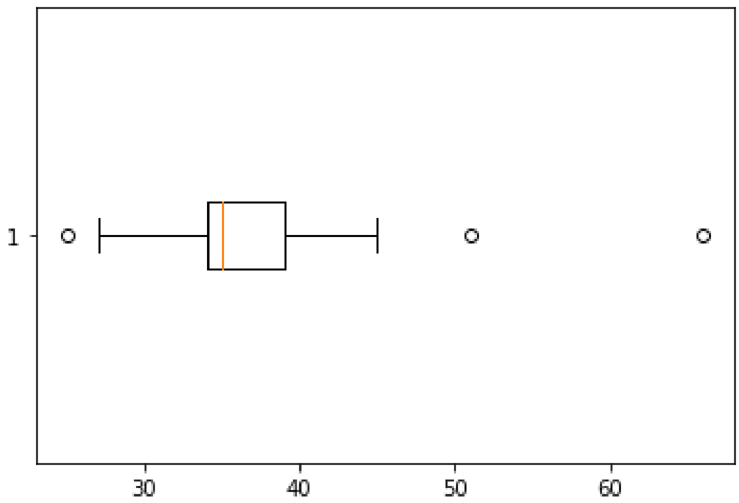
Solution:

(Source code is after the exercises)

We plot down here are the histogram, box plot and steam leaf plots in this order:


Times
0
2
4
6
8
10
30
40
50
60


Key: aggr|stem|leaf
6 | 6 = 6.6x10 = 66.0
25 | | 6
24 | 6 |
24 | |
24 | 5 | 1
23 | | 5
22 | 4 | 122
19 | | 55677899
11 | 3 | 2344444
4 | |
4 | 2 | 5789


1
30
40
50
60

The main outliers are, as we can also see from the box plot, the following: 51, 66.

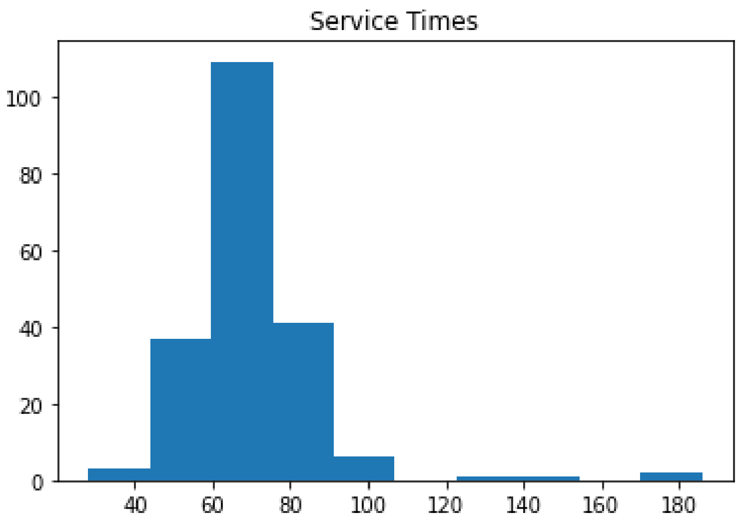
Exercise 6.2.8 Restaurant Service Times

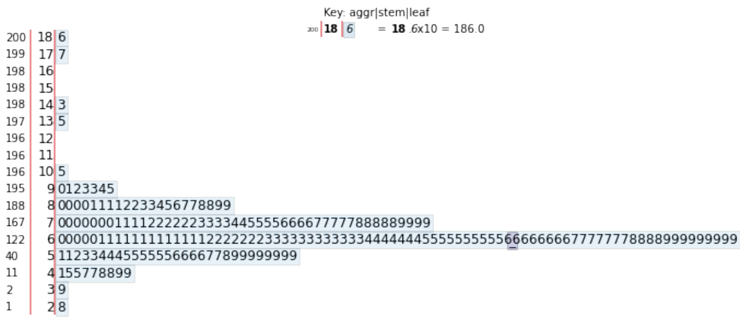
The data set of service times given in DS 6.1.4. Use a statistical software package to obtain appropriate graphical presentations of the data set. Obtain more than one graphical presentation where appropriate. Indicate any data observations that might be considered to be outliers. What do your pictures tell you about the data sets?

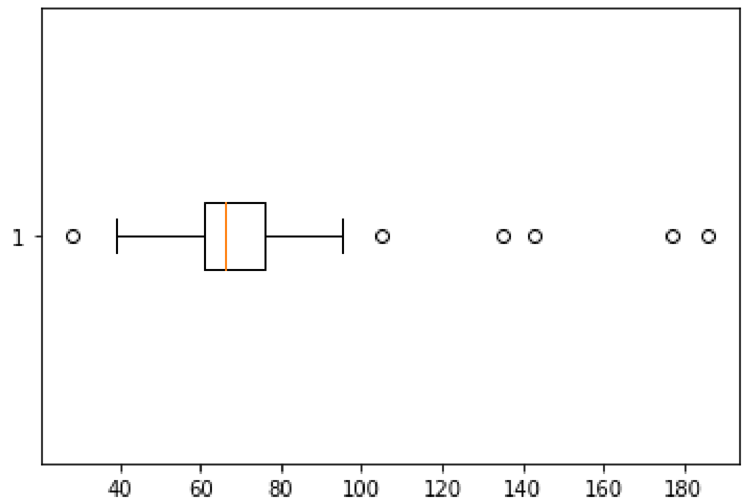
Solution:

(Source code is after the exercises)

We plot down here are the histogram, box plot and steam leaf plots in this order:







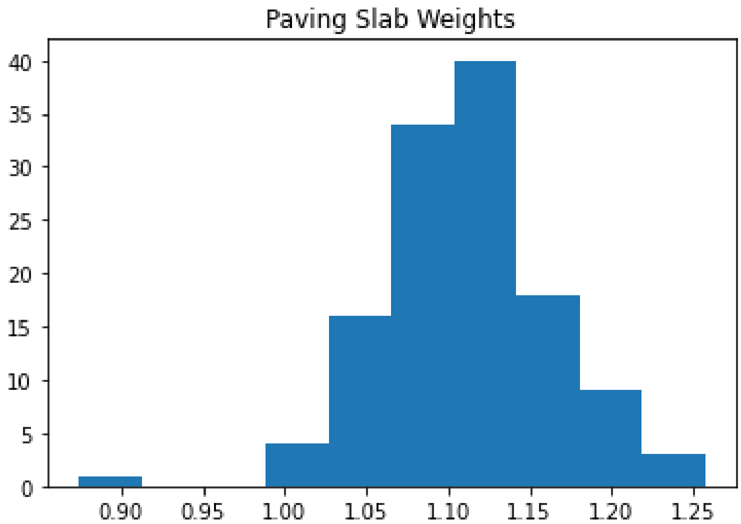
The main outliers are, as we can also see from the box plot, the following: 135, 143, 177, 186.
The distribution also be summarized by the following table, presenting its main characteristics:

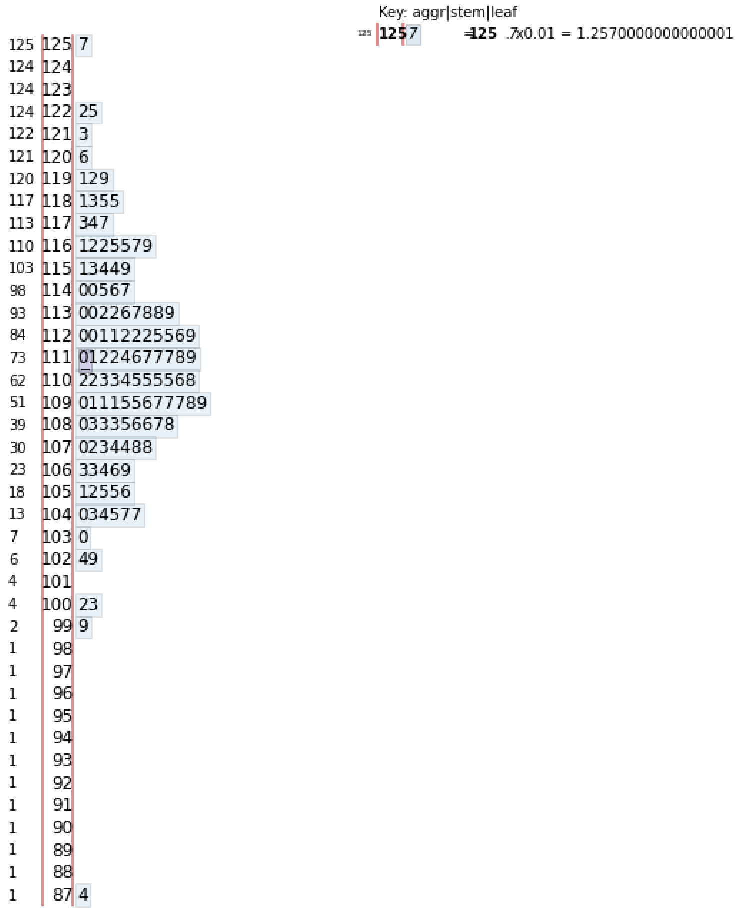
Mean	69.345
Median	66
Trimmed mean	67.8833
Standard deviation	17.5872
Upper quantile	76
Lower quantile	61

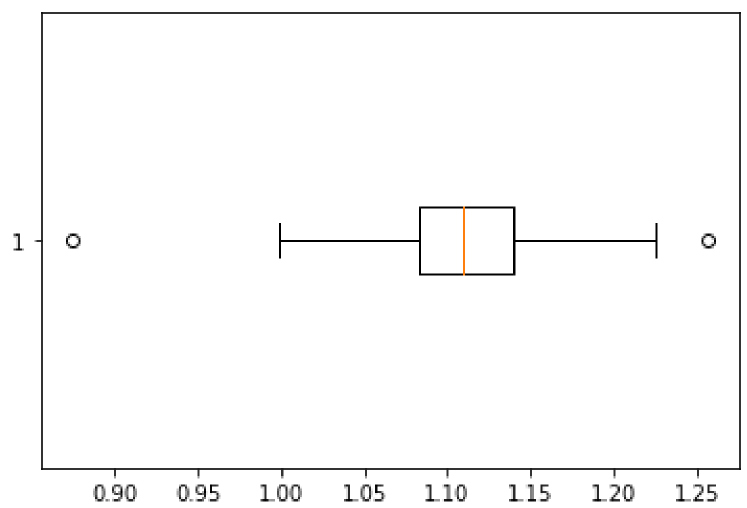
Exercise 6.3.8 Paving Slab Weights

The data set of paving slab weights given in DS 6.1.7. Use a statistical software package to obtain sample statistics and boxplots for the data set. What do the sample statistics and boxplots tell you about the data set?
Solution:

(Source code is after the exercises)
We plot down here are the histogram, box plot and steam leaf plots in this order:


Paving Slab Weights
40
35
30
25
20
15
10
5
0
0.90
0.95
1.00
1.05
1.10
1.15
1.20
1.25


Key: aggr|stem|leaf
125 |125|7 =125.7x0.01 = 1.2570000000000001
125 125 7
124 124
124 123
124 122 25
122 121 3
121 120 6
120 119 129
117 118 1355
113 117 347
110 116 1225579
103 115 13449
98 114 00567
93 113 002267889
84 112 00112225569
73 111 01224677789
62 110 22334555568
51 109 011155677789
39 108 033356678
30 107 0234488
23 106 33469
18 105 12556
13 104 034577
7 103 0
6 102 49
4 101
4 100 23
2 99 9
1 98
1 97
1 96
1 95
1 94
1 93
1 92
1 91
1 90
1 89
1 88
1 87 4


1
0.90
0.95
1.00
1.05
1.10
1.15
1.20
1.25

The main outliers are, as we can also see from the box plot, the following: 0.874, 1.257. The distribution also be summarized by the following table, presenting its main characteristics:

Mean	1.1105
Median	1.1100
Trimmed mean	1.1112
Standard deviation	0.0530
Upper quantile	1.1340
Lower quantile	1.0821

Source code for graph plotting, outlier detection and distribution characteristics

This is the source code for the previous four exercises. Notice that by changing the file name and plot names accordingly, we can plot all of the above graphs.

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
from scipy.ndimage import mean, median
from scipy.mstats import mquantiles
from stemgraphic import stem_graphic as sg
df = pd.read_excel('DS 6.2.3.xls')
# Plotting
auto.hist(['Piston Rod Lengths'], grid=False)
plt.show()
sg(auto['Piston Rod Lengths'])
plt.show()
plt.boxplot(auto['Piston Rod Lengths'], vert=False)
# Outliers detection
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
mask = ((df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR)))
filtered_data = (mask*auto).to_numpy()
print("Outliers:")
for datum in filtered_data:
    if(datum != 0):
        print(datum)
```

```python
# Distribution characteristics
print("Sample mean: ", mean(df))
print("Sample median: ", median(df))
print("Sample trimmed mean: ", stats.trim_mean(df, 0.05))
print("Sample standard deviation: ", stats.tstd(df))
print("Upper sample quartile: ", mquantiles(df, prob=[0.75]))
print("Lower sample quartile: ", mquantiles(df, prob=[0.25]))
```

Exercise 6.3.14

If a histogram is skewed with a long left tail, which of the following must be correct?

Solution:

The correct answer is B: *"The sample mean is smaller than the sample median"*.