VE401 Recitation Class Note6

Sample and Data

Chen Siyi
siyi.chen_chicy@sjtu.edu.cn

Some reminders:

1. Statistic needs more work to understand.
2. Make sure you have at least a tool to do statistic.

   I recommend mma(Mathematica). And from next week on, sample usage of mma will be attached to the RC material.

1 Probability and Statistic

Probability:
The distribution of the random variable is fully known. We draw inferences from the know information.

Statistic:
The distribution of the random variable is not fully known. We attempt to gain information on type of distribution, value of parameters, and so on, through experiments

2 Sample

Random Sample:
A random sample of size n from the distribution of X is a collection of n independent random variables X_1, ... , X_n, each with the same distribution as X.
Which means X_1, ... , X_n are independent, identically distributed (i.i.d.) random variables as X.

Comment:
Notice sometimes i.i.d may be the idea case, and not exactly satisfied in our daily life. For example, consider the below samples.

Question: i.i.d Random Sample

(Explain X and X_1, ... , X_n)
In the following four cases, are the samples independent?
In the following four cases, are the samples independent to X?

1. There are a box containing 300 balls, red or black. You do not know how many red ones. You make an experiment by drawing out 1 ball out and put it back; totally 10 times.
2. There are a box containing 300 balls, red or black. You do not know how many red ones. You make an experiment by drawing out 10 balls out one by one without putting back.
3. There are a box containing 300 balls, red or black. Red balls are heavier. You close your eyes, and separate the most heavier 150 balls to a box B. Then make an experiment by drawing out 1 ball out from B and put it back to B; totally 10 times.
4. There are a box containing 300 balls, red or black. Red balls are heavier. You close your eyes, and separate the most heavier 150 balls to a box B. Then make an experiment by drawing out 10 balls out from B one by one without putting back.

1. Independent and identical
2. Not independent
3. Not identical as X
4. Not independent or identical as X

In general, we can make X_1, ... , X_n identical to X, by controlling the way we gather samples.

And many experiments in our daily life are actually of type 2 above. To make the samples more like "independent", we control the sample size. You can recall the discussion of Hypergeometric distribution in the previous RC.

Sample Size:

The required minimum size of n is absolute, independent of the population size.

The required maximum size of n is relative to the population size. Should always smaller than 5% of the population.

3 Data

3.1 Percentiles and Quartiles

Percentile:

The x_{th} percentile is defined as the value d_x of the data such that x% of the values of the data are less than or equal to d_x.

Quartiles:

1. The first quartile q_1: 25% of the data are no greater than q_1
2. The second quartile q_2: 50% of the data are no greater than q_2
3. The third quartile q_3: 75% of the data are no greater than q_2

Calculate Quartiles: Suppose n data: $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$. Then:

1. $$q_2 = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$
2. q_1:

   A: q_2 of the smallest n/2 elements if n is even.

   B: the average of q_2 of the smallest (n − 1)/2 elements and q_2 of the smallest (n + 1)/2 elements of the list if n is odd.
3. q_3 is similar as q_1.

4 Data Visualization

4.1 Histogram

Two values, three steps:

1. **The bin width: h**

   h should be rounded up to the precision of the data.

   If h is already at the precision of the data, one smallest decimal unit should be added to h.

2. **The smallest bin boundary:**

   the smallest datum subtract one-half of the smallest decimal of the data.

3. Successively add the bin width to obtain the bins, until all the data are contained in the histogram.

In some methods, h is determined after we decide the number of bins, k.

The Sturges Method to get h:

1. $k = \lceil \log_2(n) \rceil + 1$
2. $h = \frac{\max\{x_i\} - \min\{x_i\}}{k}$

The Freedman-Diaconis Method to get h:

1. $h = \frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}}$

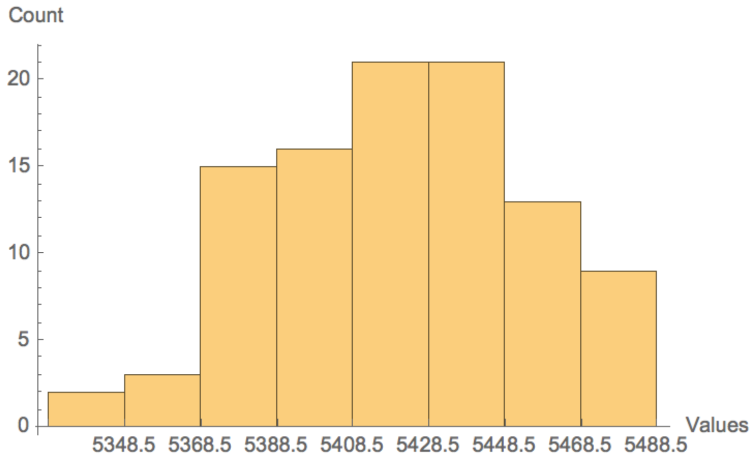
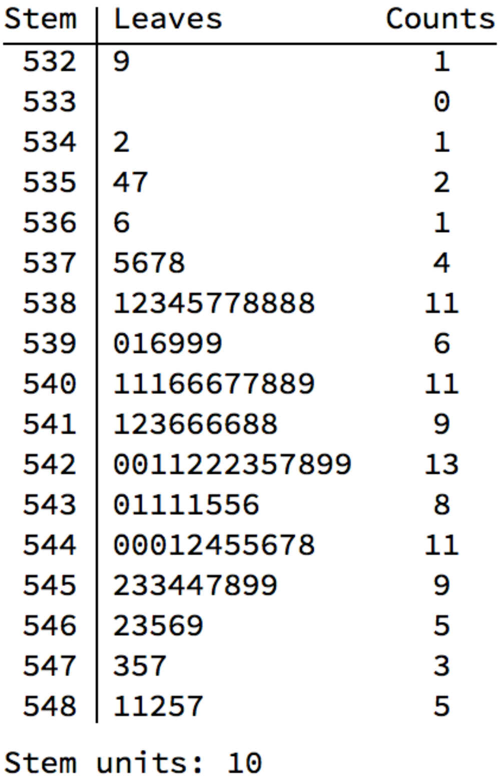


Figure 1: The Histogram

4.2 Stem-and-Leaf Diagram

Four steps:

1. Choose a convenient number of leading decimal digits to serve as stems
2. label the rows using the stems
3. for each datum of the random sample, note down the digit following the stem in the corresponding row
4. turn the graph on its side to get an impression of its distribution



```
Stem | Leaves          Counts
-----------------------------
 532 | 9                   1
 533 |                     0
 534 | 2                   1
 535 | 47                  2
 536 | 6                   1
 537 | 5678                4
 538 | 12345778888        11
 539 | 016999              6
 540 | 11166677889        11
 541 | 123666688           9
 542 | 0011222357899      13
 543 | 01111556            8
 544 | 00012455678        11
 545 | 233447899           9
 546 | 23569               5
 547 | 357                 3
 548 | 11257               5

Stem units: 10
```

Figure 2: The Stem-and-Leaf Diagram

4.3 Box-and-Whisker Plot(Boxplot)

Four steps:

1. Calculate 3 quartiles as the box, also obtain IQR.
2. Calculate adjacent values as whiskers:

$$a_1 = \min \{x_k : x_k \geq f_1\}, \quad a_3 = \max \{x_k : x_k \leq f_3\}$$

3. Calculate inner fences and draw:

$$f_1 = q_1 - \frac{3}{2}\text{IQR}, \quad f_3 = q_3 + \frac{3}{2}\text{IQR}$$

Calculate outer fences and draw:

$$F_1 = q_1 - 3\text{IQR}, \quad F_3 = q_3 + 3\text{IQR}$$

4. Find near outliers that in (F_1,f_1) or (f_3,F_3),

Find far outliers that outside (F_1,F_3), draw.

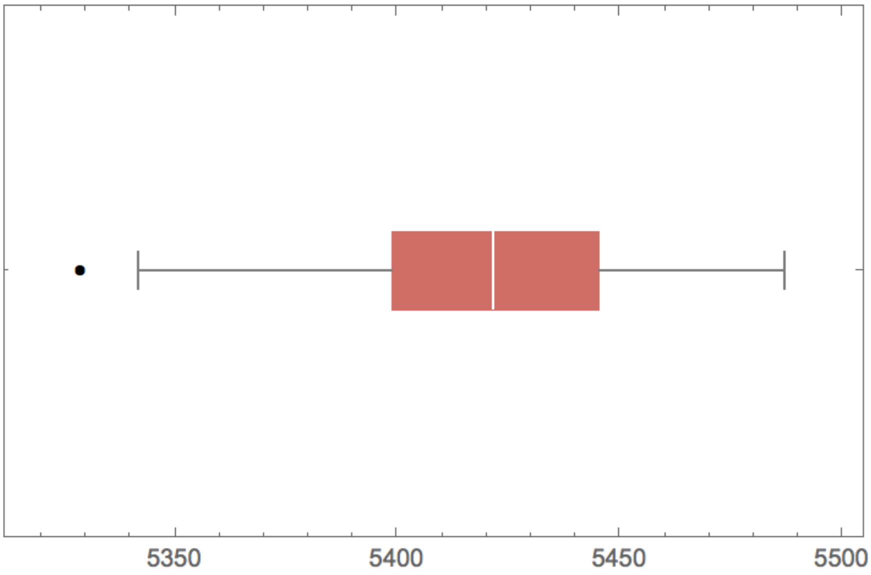


Figure 3: The Box-and-Whisker Plot

*Attached mathematica codes in demo1. Please understand before using it to help check.