Math 170S

Lecture Notes Section 6.2 [*][†]

Exploratory Data Analysis

Instructor: Swee Hong Chan

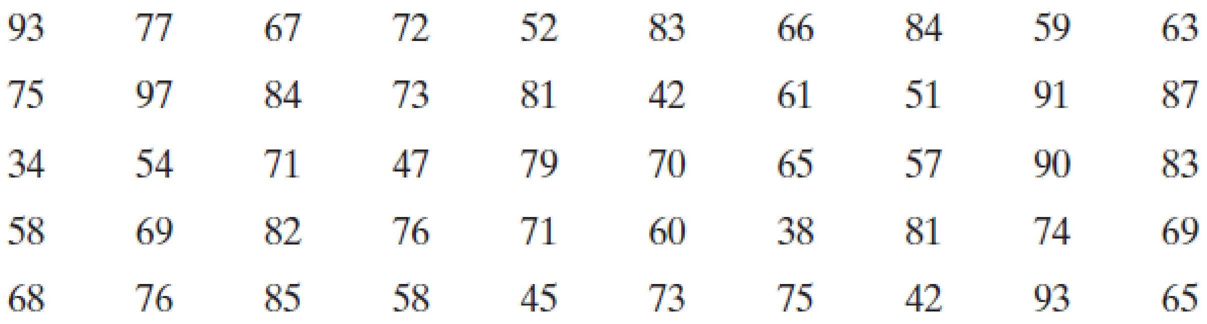
NOTE: The notes is a summary for materials discussed in the class and is not supposed to substitute the textbook. Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

[*] Version date: Sunday 4th October, 2020, 16:17.

[†] This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Stem-and-leaf display

You are asked to help the friendly instructor to analyze the scores of the final exam of a statistic class of 50 people.



93	77	67	72	52	83	66	84	59	63
75	97	84	73	81	42	61	51	91	87
34	54	71	47	79	70	65	57	90	83
58	69	82	76	71	60	38	81	74	69
68	76	85	58	45	73	75	42	93	65

Figure 1: The scores of the final exam of a class from a parallel universe, taken from the textbook.

It is hard to analyze raw data with naked eyes, so we make the **ordered stem-and-leaf display**.

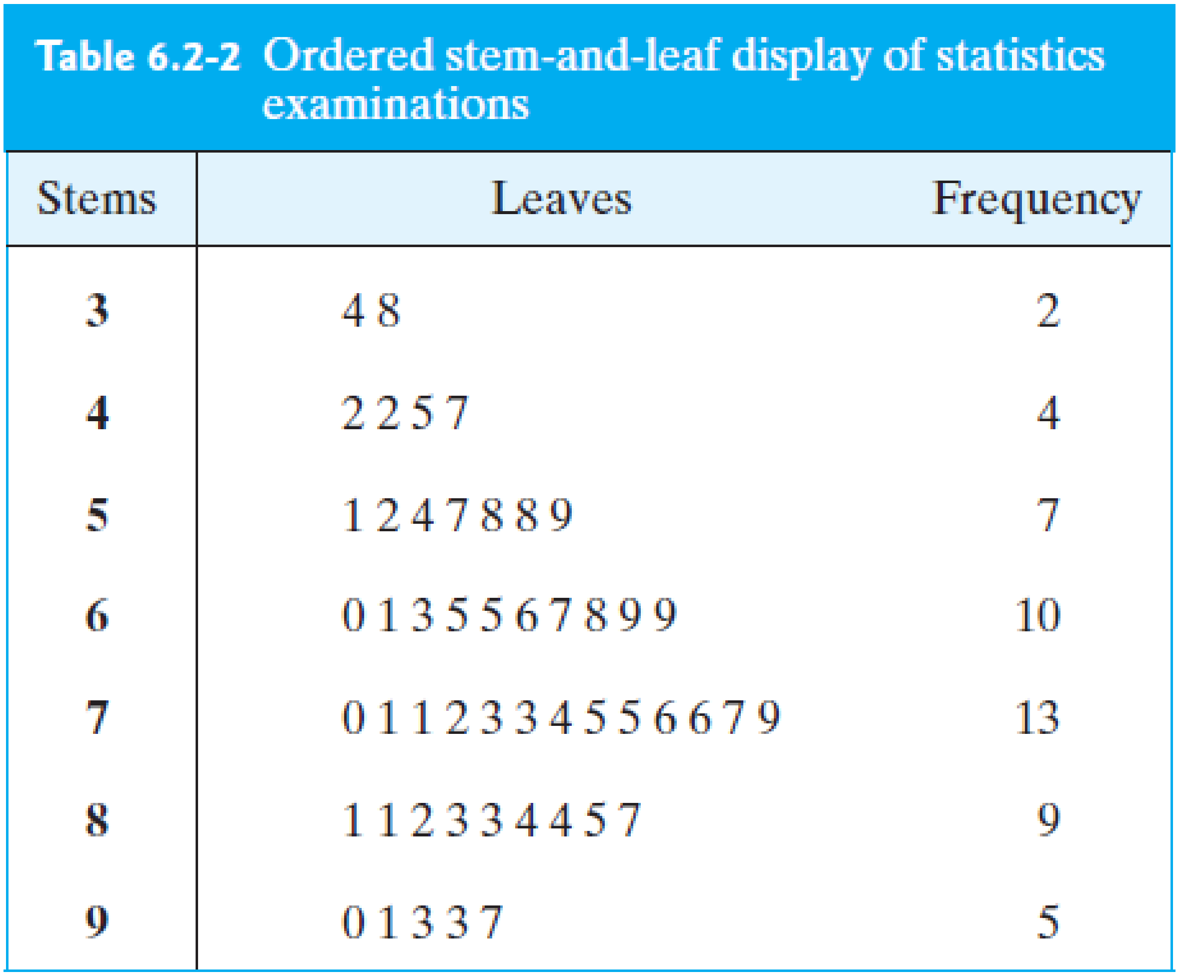


Table 6.2-2 Ordered stem-and-leaf display of statistics examinations

Stems	Leaves	Frequency
3	4 8	2
4	2 2 5 7	4
5	1 2 4 7 8 8 9	7
6	0 1 3 5 5 6 7 8 9 9	10
7	0 1 1 2 3 3 4 5 5 6 6 7 9	13
8	1 1 2 3 3 4 4 5 7	9
9	0 1 3 3 7	5

Figure 2: The ordered stem-and-leaf display for the exam scores presented in Figure 1, taken from the textbook.

Remark 1. Can you find a bell curve hidden somewhere in the plain sight in the table?

The rule for creating this particular display is

- The first digit of the sample value will be the ‘stem’, while the second digit will be the ‘leaf’ of the sample value.

  For example, a sample value 93 will have a stem of 9 and a leaf of 3.

- We will order the stems vertically in increasing order (from the smallest to largest), and order the leaves horizontally in increasing order.

- We will also record the frequency of each stem, which is equal to the number of its leaves.

2 Order statistics

Let $x_1, x_2, \ldots, x_n$ be a given sample values. We order them from the smallest to the largest:

- y_1 is the smallest of $x_1, \ldots, x_n$;
- y_2 is the second smallest of $x_1, \ldots, x_n$;
- $\ldots$
- y_n is the largest of $x_1, \ldots, x_n$.

The number y_k is called the **k-th order statistics of the sample**.

Example 2. Suppose that the samples are given by

$$x_1 = 3; \quad x_2 = 8; \quad x_3 = 5; \quad x_4 = 1.$$

The y_k's are then given by

$$y_1 = 1; \quad y_2 = 3; \quad y_3 = 5; \quad y_4 = 8.$$

3 Median

Definition 3. The **median** of $x_1, x_2 \ldots, x_n$ is the value m so that half of the sample values are less than m, and the other half are greater than m.

- If $n = 2h+1$ is an odd number, the median is y_{h+1};
- If $n = 2h$ is an even number, the median is $\frac{y_h+y_{h+1}}{2}$.

Example 4. Suppose that the samples are

$$1 \quad 3 \quad 5 \quad 9 \quad 13.$$

Then median is 5, the number in the middle.

Suppose that the samples are

$$1 \quad 4 \quad 11 \quad 12;$$

The median is $\frac{4+11}{2} = 7.5$, the average of the two numbers in the middle.

Remark 5. The median should NOT be confused with the mean of the sample. Indeed, the samples

$$1 \quad 4 \quad 13$$

has median 4 but mean $\frac{1+4+13}{3} = 6$.

4 Sample percentile

Fix $p \in (0,1)$. The **(100p)th sample percentile** (or **sample percentile of order p**) is the value $\widetilde{\pi}_p$ such that np of the sample values are less than, and the rest $n(1-p)$ of them are larger than $\widetilde{\pi}_p$.

- If $(n+1)p$ is an integer, then $\widetilde{\pi}_p$ is equal to $y_{(n+1)p}$.

  For example, suppose that $n = 99$ with samples

  $$1 \quad 2 \quad 3 \quad \ldots \quad 98 \quad 99.$$

  Let $p = 0.42$. Then

  $$(n+1)p = (99+1)(0.42) = 42,$$

  which is an integer. Then $\widetilde{\pi}_p$ is equal to

  $$\widetilde{\pi}_p \; = 42.$$

- $(n+1)p$ is not an integer, but is equal to

  $$(n+1)p = r + \delta,$$

  where r is an integer, and δ is a real number such that $0 \leq \delta < 1$ (such r and δ always exist, and is unique). Then $\widetilde{\pi}_p$ is equal to

  $$\widetilde{\pi}_p := (1-\delta)y_r + \delta y_{r+1} \; = \; y_r + \delta(y_{r+1} - y_r).$$

  That is to say, $\widetilde{\pi}_p$ is a linear interpolation between y_r and y_{r+1}.

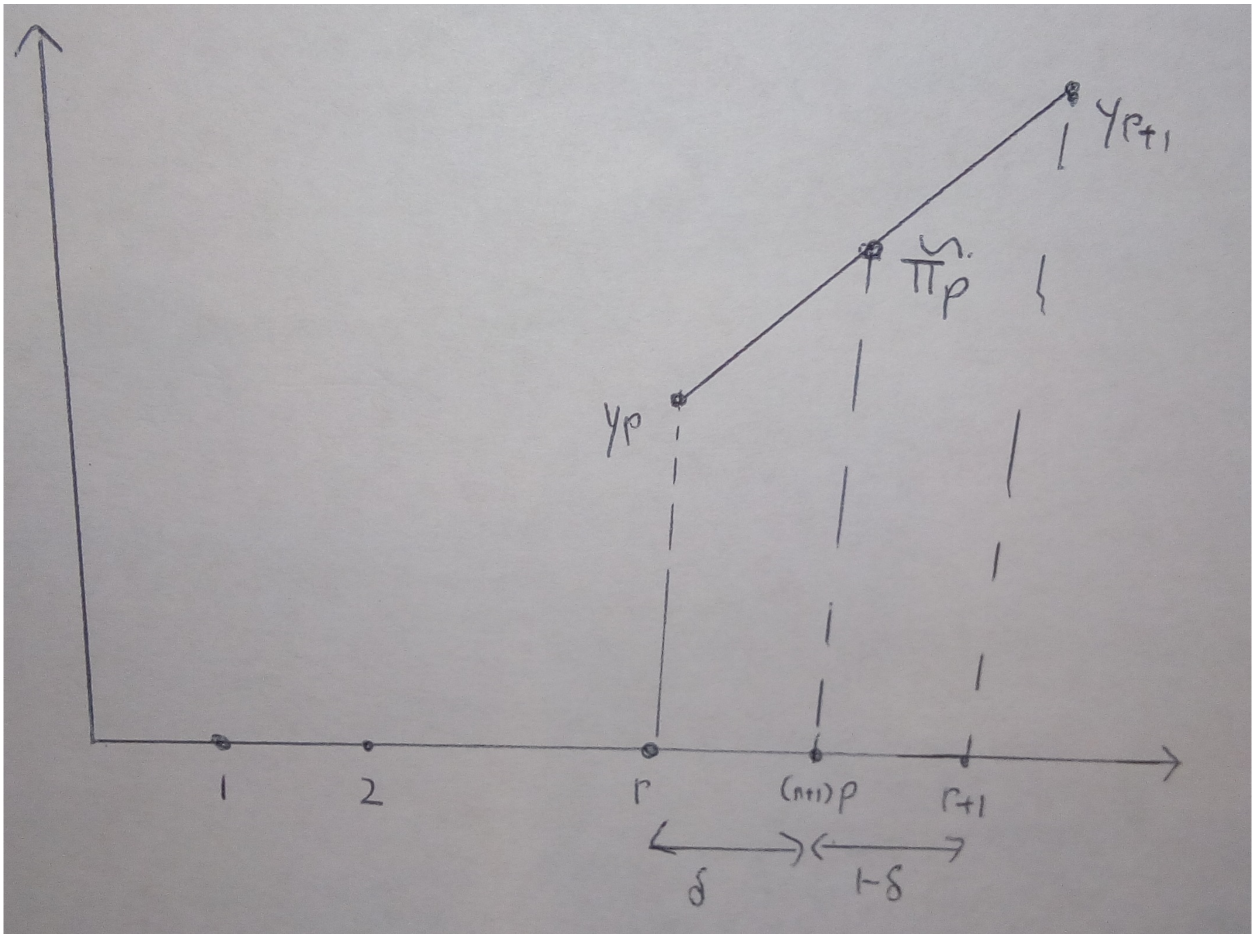


Figure 3: The location of the 100p-th percentile with respect to y_r and y_{r+1}.

For example, suppose $n = 99$ with samples

$$1 \quad 2 \quad 3 \quad \ldots \quad 98 \quad 99.$$

Let $p = 0.347$. Then

$$(n+1)p = 34.7 = 34 + 0.7 = r + \delta,$$

so $r = 34$ and $\delta = 0.7$. Then $\widetilde{\pi}_p$ is equal to

$$\begin{aligned} \widetilde{\pi}_p &= y_r + \delta(y_{r+1} - y_r) \\ &= (34) + (0.7)[(35) - (34)] = 34.7\,. \end{aligned}$$

Remark 6. Percentiles is NOT defined when $p < \frac{1}{n+1}$ or when $p > \frac{n}{n+1}$.

- In the former case, we have $(n+1)p = r + \delta$ with $r = 0$, so $\widetilde{\pi}_p$ is a linear interpolation between y_0 and y_1. However, y_0 does NOT exist.
- In the latter case, we have $(n+1)p = r + \delta$ with $r = n$, so $\widetilde{\pi}_p$ is a linear interpolation between y_n and y_{n+1}. However, y_{n+1} does NOT exist.

Remark 7. The reason why in the definition of percentile we use the $(n+1)p$ rather than np is so that median is equal to the 50th percentile.

5 Quartiles

Special names are given to certain sample percentiles. The 25th, 50th, and 75th percentiles are called the **first, second, and third quartiles**, respectively. We also use special notations for them:

$$\widetilde{\pi}_1 = \widetilde{\pi}_{0.25}, \qquad \widetilde{\pi}_2 = \widetilde{\pi}_{0.5}, \qquad \widetilde{\pi}_3 = \widetilde{\pi}_{0.75}\,.$$

Remark 8. There is no universal agreement on selecting quartile values. For this course, we will always use the definition presented here.

6 Box plot: Example

When one reads a financial report, one often sees data presented in the form of box plots. We discuss how to interpret these pictures here.

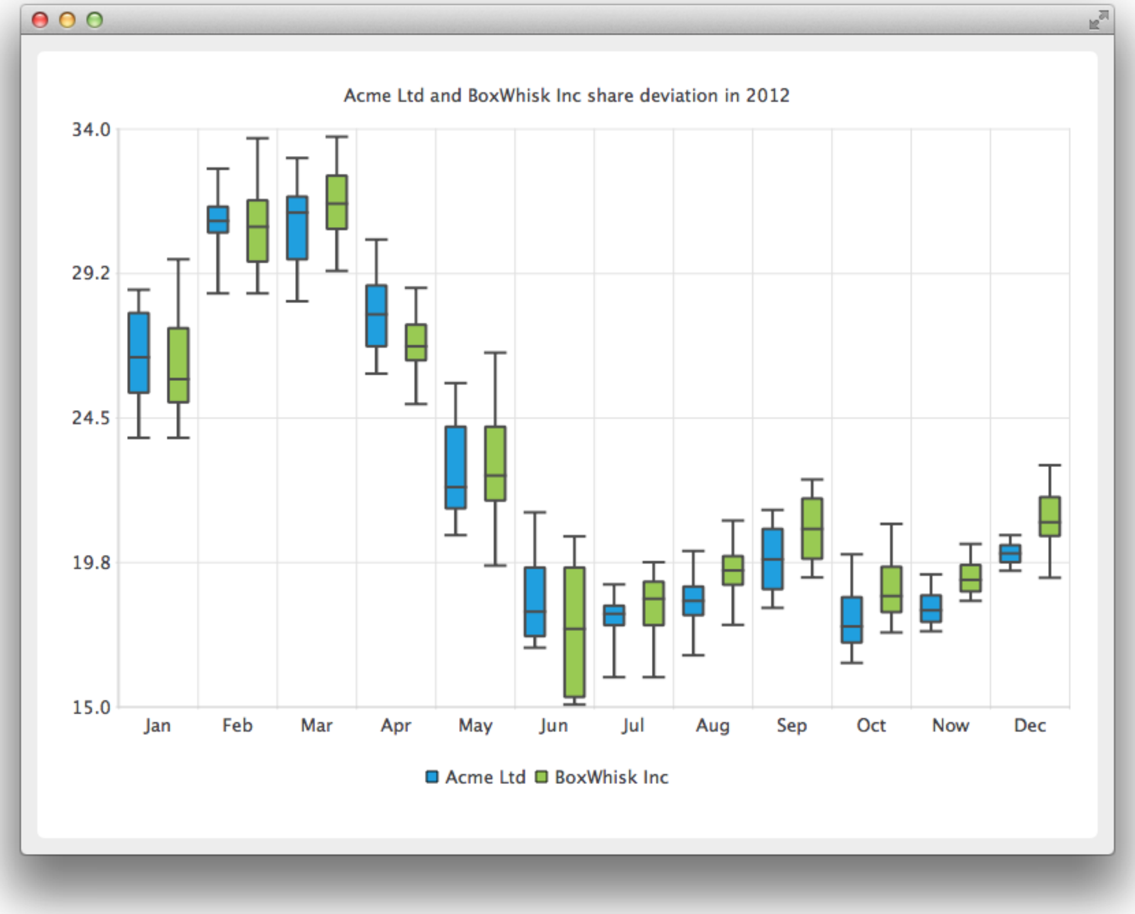


Figure 4: The box plot of the stock's value of a fictional company, taken from `www.qt.io`.

7 Five number summary

The **five-number summary** are these five numbers:

- First quartile $\widetilde{\pi}_1$;
- Second quartile $\widetilde{\pi}_2$;
- Third quartile $\widetilde{\pi}_3$;
- The minimum value y_1; and
- The maximum value y_n.

The **interquartile range (IQR)** is the difference between the first and third quartile,

$$IQR := \widetilde{\pi}_3 - \widetilde{\pi}_1 \,.$$

8 Box plot

The **box plot** or **box-and-whisker diagram** is a diagram that records the five-number summary as follows:

- The vertical line at the middle is the second quartile;
- The vertical line at the left side is the first quartile;
- The vertical line at the right side is the third quartile;
- the left end of the horizontal line is the minimum;
- the right end of the horizontal line is the maximum.

Note that the interquartile range IQR is equal to the length of the box in the box plot.

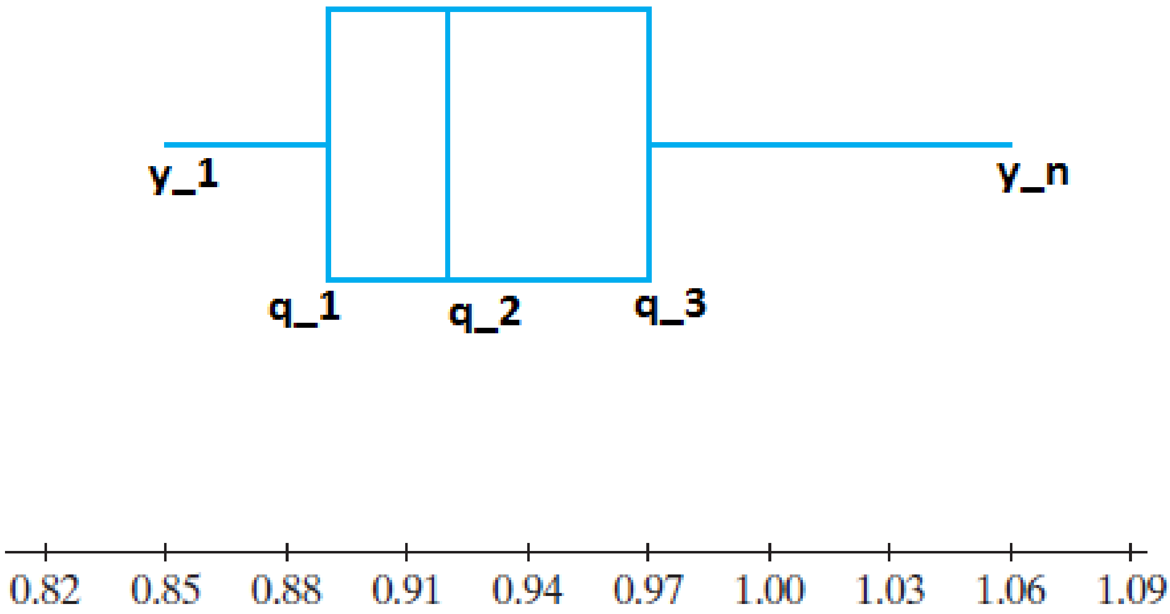


Figure 6.2-1 Box plot of fluoride concentrations

Figure 5: Box plot of the five point summary, taken from textbook

From the box plot we learn that

$$y_1 = 0.85,\ \ \widetilde{\pi}_1 = 0.89,\ \ \widetilde{\pi}_2 = 0.92,\ \ \widetilde{\pi}_3 = 0.97,\ \ y_n = 1.06.$$