One-dimensional data analysis

I. Key mathematical terms

Terms	Symbol	Chinese translation
mode		
arithmetic mean		
geometric mean		
median		
quartile		
percentile		

II. Statistic and statistical charts

There are several commonly used statistics that we learned in junior high school. Let's review the definitions and formulas of these statistics.

(1) Mode 眾數 (　　)

The value that appears most frequently in a data set.

(2) Arithmetic mean/Arithmetic average 算術平均數 (　　)

The sum of all values in a dataset divided by the numbers of the values.

For example:

A data set consisting of values $x_1, x_2, x_3, ..., x_n$, the arithmetic mean is defined by the formula: $\mu = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$

(3) Median 中位數 (　　)

The vale separating the lowest 50% from the highest 50% of the data.

To determine the median, you need to arrange the data points in either ascending or descending order. If there is an odd number of observations, the median is the middle number. If there is an even number of observations, the median is the average of the two middle numbers.

(4) Quartile 四分位數（　　）

Quartiles are values that divide a dataset into four equal parts. To find quartiles, arrange the data in ascending order and then divide them into four segments.

The quartiles are defined as follows:

Q1 (lower quartile):

The value separating the lowest 25% of the data from the highest 75%.

Q2 (median):

The value separating the lowest 50% from the highest 50% of the data.

Q3 (upper quartile):

The value separating the lowest 75% of the data from the highest 25%.

If the division between two numbers fall exactly between data points, the quartile is the average of those two numbers.

After reviewing the definition and calculation about the statistics we learned in junior high, let's use an example to let you try it out.

Example 1

The following table shows the record that how many times 20 students are late in a school year

1	5	3	4	2	0	0	4	5	5
3	4	4	5	1	3	3	4	5	5

(1) Write down the mode of the data

(2) Work out the mean number of late arrivals

(3) Work out the median number of late arrivals

(4) Work out the quartile numbers Q1 and Q3 of late arrivals

<key>

You have to rearrange the data points in ascending form before you find the median and quartiles.

<sol>

Absolute frequency distribution table

The absolute frequency distribution table displays the frequency of each data set in an organized way. It helps us to find the patterns in the data. To complete the distribution table, you need to collect data and organize it in the form of a frequency distribution table. Please take a look at the following example:

Example 2

The school organize a blood donation event. A group of 30 students' blood group are recorded as follows:

A, B, O, O, AB, B, B, B, A, O, O, O, O, AB, O, A, A, B, O, O, AB, O, A, A, B, B, B, O, O, O

Represent this data in the form of a frequency distribution table.

Blood group	Number of students
A	
B	
AB	
O	
Total	

Absolute cumulative frequency

There are two types of absolute cumulative frequency, the 'lesser than cumulative frequency' and 'greater than cumulative frequency'. For the lesser than cumulative frequency, the cumulate starts from the lowest to the highest and the greater than cumulative frequency, the cumulate starts from the highest to the lowest.

Relative frequency and relative cumulative frequency

The relative frequency is expressed in percentages. To find the relative frequency, calculate the proportion or percentage of times a particular value is observed in the interval. For more details and examples, please read the paragraph in the following link:

https://edu.gcfglobal.org/en/statistics-basic-concepts/frequency-tables/1/

Geometric mean

The geometric mean (GM) is the average value of n numbers where we multiply the numbers together and then take the n-th root.

A set of positive numbers: $x_1, x_2, x_3, \ldots, x_n$, the geometric mean is defined as:

$$G.M. = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

Example 3

The revenue growth rates of a company from 2019 to 2023 were as follows:

50%, 20%, -10%, 8%

Find the average growth rate of the revenue of this company.

III. Percentile

The percentile is a number that tells the percentage of scores that fall below the given number. It is frequently used for grading test scores and biometric measurements. You can calculate the percentile by the following steps:

$P_k =$ the k^{th} percentile. (It may or may not be part of the data)

$i =$ the index helping us finding the percentile.

$n =$ the total number of the data.

Step1: Arrange the data points in ascending order.

Step2: Calculate the index $i = n \times \frac{k}{100}$.

Step3-1: If i is an integer. Then $P_k = \frac{1}{2}(x_i + x_{i+1})$.

Step3-2: If i is **not** an integer. Then $P_k = x_m$. (m is the least integer greater than i.)

To make it easier for you to understand percentiles, let's look at the example below.

Example 4

We have test score for 10 students: 55, 60, 65, 70, 75, 80, 85, 90, 95, 100.

Find the 40th percentile.

<illustration>

Step1: Sort the scores. (Already sorted in ascending order.)

Step2: Calculate the index $i = 10 \times \frac{40}{100} = 4$ (It's an integer.)

Step3-1: $P_{40} = \frac{x_4 + x_5}{2} = \frac{70 + 75}{2} = 72.5$.

Now, please try to solve the following example on your own.

Example 5

The following list are 25 ages for Academy Award-winning best actors in order from smallest to largest: 18, 22, 25, 26, 27, 29, 30, 31, 36, 37, 41, 42, 47, 52, 57, 58, 62, 64, 67, 69, 71, 72, 74, 76, 77.

(1) Find the percentile for 20

(2) Find the percentile for 71

IV. Measures of data dispersion

Besides the statistical measures introduced earlier, there are some that are specifically used to analyze the dispersion of data. Let's take a look at the ones we will encounter in high school.

(1) Range 全距 (　　)

The difference between the smallest and the largest value.

(2) Interquartile range 四分位距 (　　)

The difference between the 75th and 25th percentiles. (You can use the method mentioned above to find the 75th and 25th percentiles and find the difference.)

Example 6

Use the data from the previous question, calculate the range and interquartile range.

(3) Variance 變異數 (　　)

Variance is the statistical measure that quantifies the spread or dispersion of set of data points around the mean. It provides an indication of how much the values in a dataset vary from the mean value.

(4) Standard deviation 標準差 (　　)

Standard deviation is the square root of the variance, providing a more interpretable measure of spread. (Since it is in the same units as the data.)

Next, let's take a look at the formulas for calculating variance and standard deviation.

Variance and Standard deviation

A set of data points: $x_1, x_2, x_3, ..., x_n$. $\mu = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$

Variance: $\sigma^2 = \frac{1}{n}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2\right]$

Standard deviation: $\sigma = \sqrt{\frac{1}{n}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2\right]}$

In addition to the formula above, there is another way to calculate variance and standard deviation. This formula can be derived by expanding each square term and simplifying. For this proof, we'll left it as an after-class exercise.

A set of data points: $x_1, x_2, x_3, ..., x_n$.

Variance: $\sigma^2 = \frac{1}{n}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2\right] = \frac{1}{n}({x_1}^2 + {x_2}^2 + ... + {x_n}^2) - \mu^2$

Standard deviation: $\sigma = \sqrt{\frac{1}{n}({x_1}^2 + {x_2}^2 + ... + {x_n}^2) - \mu^2}$

Let's try to use the formulas above to answer the question together.

Example 7

Consider the number of gold coins 10 pirates have: 2,2,3,4,5,6,7,7,8,8. Find the variance and standard deviation. (You can use the calculator to help you.)

Example 8

There are 10 students in the club. They took a general knowledge quiz. The average score is 62, the standard deviation is 8. It knows that the scores of 8 out of these 10 students are as follows: 50, 54, 56, 58, 62, 64, 70, 72. Find the scores of the other two students.

<資料來源>

1. Statistic charts

Quartiles: https://www.mathsisfun.com/data/quartiles.html

Frequency distribution table:

https://www.cuemath.com/data/frequency-distribution-table/

Cumulative Frequency and relative frequency:

https://www.cuemath.com/data/cumulative-frequency/

https://edu.gcfglobal.org/en/statistics-basic-concepts/frequency-tables/1/

https://stats.libretexts.org/Bookshelves/Introductory_Statistics

https://edu.gcfglobal.org/en/statistics-basic-concepts/frequency-tables/1/

2. Geometric mean

https://www.mathsisfun.com/numbers/geometric-mean.html

3. Percentile

https://www.cuemath.com/percentile-formula/

https://www.mathsisfun.com/data/percentiles.html

https://www.mathsisfun.com/definitions/percentile-rank.html

https://www.texasgateway.org/resource/23-measures-location-data

4. Standard deviation

https://byjus.com/maths/standard-deviation/

5. 南一書局數學（二）

製作者：國立臺灣師範大學附屬高級中學 蕭煜修