COMP229: Introduction to Data Science
Lecture 13: descriptive statistics

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

Descriptive vs inferential statistics

Descriptive statistics quantitatively summarises features of *sample* data by numbers, diagrams.

This approach differs from *inferential statistics* that aims to learn about the whole *population* (all objects in question) from a smaller sample.

This class can be viewed as a sample of all students studying in the UK. If we only find the average age in this class, it's a simple description. If we infer (make a conclusion about) ages of all UK students, this is an *inference* problem.

The range of a data sample

The first good question to ask about data:
what is the range of a given sample?

Definition 13.1. The *range* of scalar data is the interval from the minimum to the maximum value.

Example. What's the range of this data sample?
$9, 4, 2, 4, 6, 4, 7, 4, 6, 4$.

It might help to put the values in the increasing order from the minimum to the maximum value:
$2, 4, 4, 4, 4, 4, 6, 6, 7, 9$.

3 descriptors of a data sample

The range of $9, 4, 2, 4, 6, 4, 7, 4, 6, 4$ is $[2, 9]$.

Definition 13.2. The sample *mean* (or the arithmetic average) of n values $a_1, \ldots, a_n$ is

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

The data sample S above has the mean $\bar{a} = \frac{1}{10}(9 + 4 + 2 + 4 + 6 + 4 + 7 + 4 + 6 + 4) = 5.$

Note: $\frac{1}{2}\left(\min_{i=1,\ldots,n} a_i + \max_{i=1,\ldots,n} a_i\right) = 5.5 \neq \bar{a} = 5.$

2 more descriptors of a sample

Definition 13.3. The *mode* of a sample is the most frequent value (not always unique).

The *median* of a sample is the value separating the lower half of the sample from the upper half.

For $2k+1$ values $a_1 \leq \cdots \leq a_{k+1} \leq \cdots \leq a_{2k+1}$, the median is the middle value a_{k+1}. For $2k$ values $a_1 \leq \cdots \leq a_k \leq \cdots \leq a_{2k}$, the median is the average of two middle values $\dfrac{a_k + a_{k+1}}{2}$.

The sample $2, 4, 4, 4, 4, 4, 6, 6, 7, 9$ has mode 4 (appearing 5 times) and median $0.5(4+4) = 4$.

The 1st and 3rd quartiles Q_1 and Q_3

Definition 13.4. The *1st quartile* Q_1 separates the lowest 25% of the data from the highest 75%.

The *3rd quartile* Q_3 separates the lowest 75% of the sample from the highest 25% of the sample.

If a sample has $2k$ values, Q_1 is the median of the lowest half, Q_3 is the median of the highest half.

The median from Def 13.3 is the 2nd quartile Q_2.

Find the 1st/3rd quartiles of $9, 4, 2, 4, 6, 4, 7, 4, 6, 4$.

Quartiles for an odd number of values

After ordering, the sample $2, 4, 4, 4, 4, 4, 6, 6, 7, 9$ of 10 values has $Q_1 = 4$ and $Q_3 = 6$.

For $4k + 1$ ordered values $a_1 \leq \cdots \leq a_{4k+1}$, set $Q_1 = 0.75a_k + 0.25a_{k+1}, Q_3 = 0.25a_{3k+1} + 0.75a_{3k+2}$.
For $4k - 1$ ordered values $s_1 \leq \cdots \leq s_{4k-1}$, set $Q_1 = 0.75a_k + 0.25a_{k+1}$, $Q_3 = 0.25a_{3k-1} + 0.75a_{3k}$.

Find the 1st and 3rd quartiles of the samples $S_1 = \{6, 5, -2, 7, 12, 5, 6, 3, 4\}$ and $S_2 = \{6, 4, 8, 6, 12, 9, 5, 7, 6, 1, 6\}$.

The 5-number summary of a sample

$\{-2, 3, 4, 5, 5, 6, 6, 7, 12\}$, $Q_1 = 3.25$, $Q_3 = 6.75$

$\{1, 4, 5, 6, 6, 6, 6, 7, 8, 9, 12\}$, $Q_1 = 5.25$, $Q_3 = 7.75$.

Definition 13.5. The 5-number summary of a sample consists of the minimum, 3 quartiles and maximum: $\min_{i=1,\ldots,n} a_i \leq Q_1 \leq Q_2 \leq Q_3 \leq \max_{i=1,\ldots,n} a_i$.

The *Interquartile Range* $IQR = Q_3 - Q_1$.

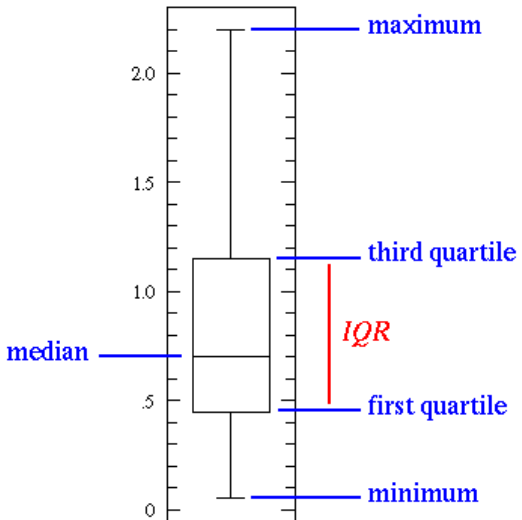
The $1.5 \times IQR$ rule says that a value outside $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ is an *outlier*.

Find outliers in the samples S_1 and S_2 above.

The box plot of a data sample

The sample $S_1 = \{-2, 3, 4, 5, 5, 6, 6, 7, 12\}$ has $IQR = 6.75 - 3.25 = 3.5$ and is contained in $[3.25 - 5.25, 6.75 + 5.25]$, hence has no outliers.

The sample $S_2 = \{1, 4, 5, 6, 6, 6, 6, 7, 8, 9, 12\}$ has $IQR = 7.75 - 5.25 = 2.5$. The values $1, 12$ are outside $[5.25 - 3.75, 7.75 + 3.75]$, hence outliers.



The box plot on the left visualises the 5-number summary of a data sample.

The sample standard deviation

Definition 13.6. A sample of n values $a_1, \ldots, a_n$ with a mean $\bar{a}$ has the (unbiased) *sample standard deviation* $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2}$.

The sum is divided by $n-1$, not by n, because the differences $a_i - \bar{a}$ have sum 0, hence $n-1$ "degrees of freedom" (independent variables).

This deviation s and IQR show how widely values are distributed around the mean $\bar{a}$ or median.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Find the sample standard deviation s of the data sample $S = \{9, 4, 2, 4, 6, 4, 7, 4, 6, 4\}$

Answer to the quiz and summary

Answer. $\{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$ has $\bar{a} = 5$, $s =$

$$\sqrt{\frac{(2-5)^2 + 5(4-5)^2 + 2(6-5)^2 + (7-5)^2 + (9-5)^2}{9}}$$

$$= \sqrt{\frac{1}{9}(9 + 5 + 2 + 4 + 16)} = \sqrt{36/9} = \sqrt{4} = 2.$$

- A data sample $a_1, \ldots, a_n$ of n values has the sample mean $\bar{a} = \dfrac{1}{n}\sum_{i=1}^{n} a_i$ and the sample standard deviation $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2}$.