COMP229: Introduction to Data Science
Lecture 3: box plots of data samples

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Biased vs unbiased variance

Consider a sample $A = \{a_1, \ldots, a_n\}$ with a mean $\bar{a}$.

Sample standard deviation	Population standard deviation
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2}$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2}$

σ is used when a **sample is a whole population**.

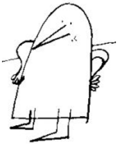
If a sample isn't a whole population, then the difference between the *population* average and the *sample* average requires correction by the factor $\frac{n}{n-1}$, which is called Bessel's correction.

Other common values for the denominator can correct other types of error. Four common values for the denominator are $n-1, n$ and (for normal distribution) $n+1, n-1.5$. The last three factors are beyond the scope of COMP229.

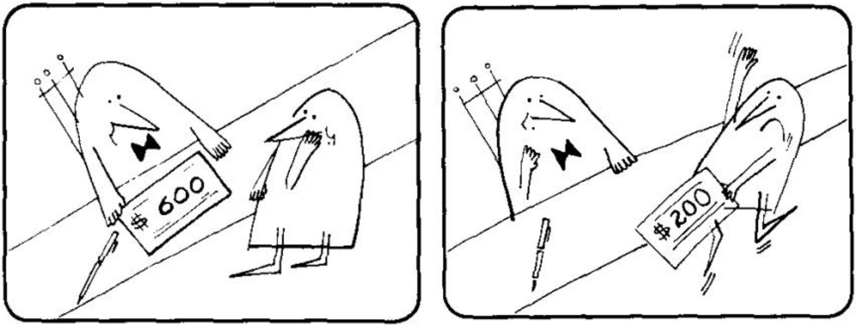
Similarly, the *sample variance* is $Var = s^2$ and the *population variance* is $Var = \sigma^2$.

In this module we will mostly use the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2}$.

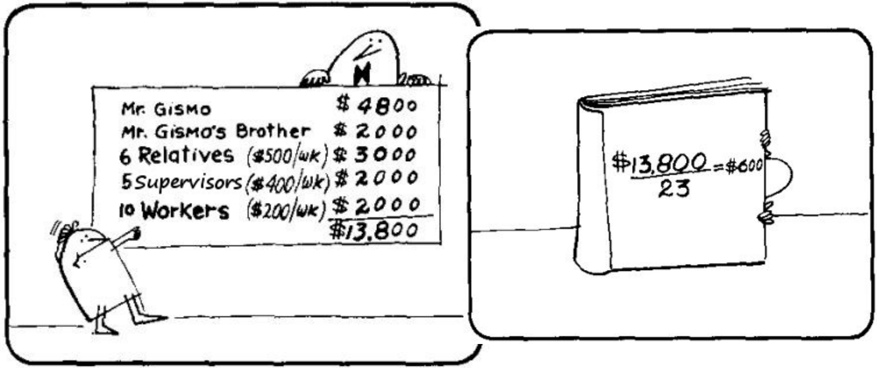
Alex's story



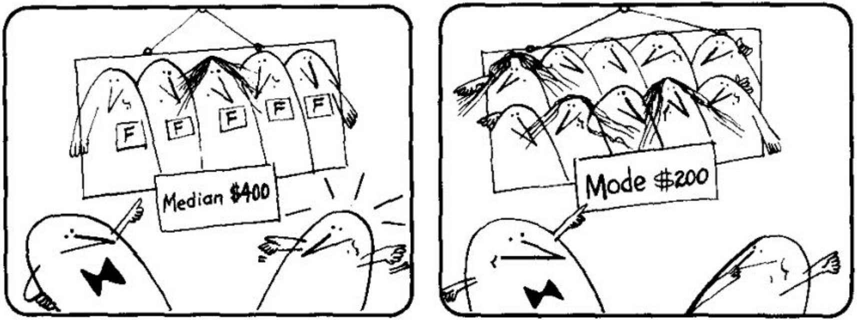
from Aha! Gotcha by Martin Gardener



Alex's story continued



Alex's story - happy ending



The median of a data sample

Definition 3.1. The *median* of a data sample $A = \{a_1, \ldots, a_n\}$ is the value separating the lower half of the ordered sample from the upper half. For $2k+1$ values $a_1 \leqslant \cdots \leqslant a_{k+1} \leqslant \cdots \leqslant a_{2k+1}$, the median is the middle value a_{k+1}. For $2k$ values $a_1 \leqslant \cdots \leqslant a_k \leqslant a_{k+1} \leqslant \cdots \leqslant a_{2k}$, the median is the average of two middle values $\dfrac{a_k + a_{k+1}}{2}$.

Definition 3.2. A *mode* is a most frequent value (not always unique), i.e. a value that appears in the data sample a highest number of times.

Computing the median of a sample

Problem 3.3. Find the median of the data sample $A = \{9, 4, 2, 4, 6, 4, 7, 4, 6, 4\}$.

Solution 3.3. Order the sample:
$A = \{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$
The sample A has 10 values. The middle values (at places 5 and 6) are 4 and 4, hence the median is their arithmetic average $(4 + 4)/2 = 4$.

Problem 3.4. Is the median of any data sample always equal to the sample mean?

Median vs mean

Solution 3.4. $A = \{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$ has the mean $\bar{a} = 5$ not equal to the median 4.

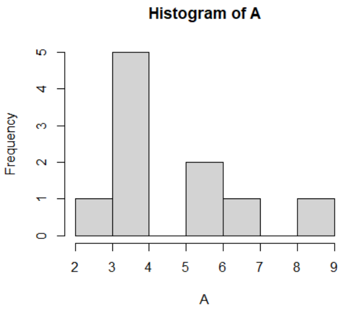
Problem 3.5. When are median and mean equal?

Solution 3.5. If a sample A is *symmetric* around its median a, i.e. splits into pairs $a \pm x$ for some x (possibly including a as an extra value) then median $=$ mean.

The data sample $B = \{2, 3, 3, 3, 4, 6, 6, 6, 8, 9\}$ has the mean $\bar{b} = 5$ equal to median 5, but isn't symmetric around 5.

Histogram

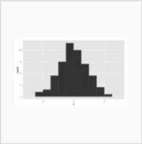
A **histogram** is a plot that depicts the number of observations that fall into each of the disjoint categories (known as **bins**):

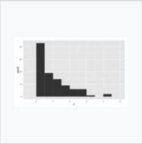


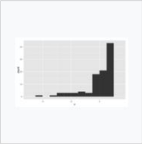
histogram of A $=$ $\{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$

Histogram shapes

It is handy to see the symmetries and other properties, especially for big data:


Symmetric, unimodal

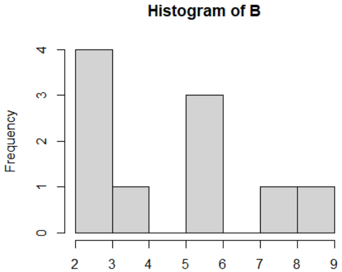

Skewed right


Skewed left


Bimodal


Multimodal

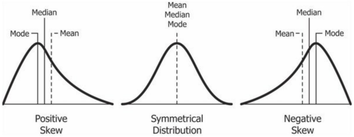

Symmetric



histogram of B $=$ $\{2, 3, 3, 3, 4, 6, 6, 6, 8, 9\}$, no symmetry.

Skewness and central tendencies

Usually for skewed data the mean lies toward the direction of skew (the longer tail) relative to the median,



but this is not a general rule and is very frequently violated for real life data!

In cases where one tail is *long* but the other tail is *fat*, skewness does not obey this simple rule.

Try, for example, $C = \{1, 1, 2, 2, 3\}$.

A mode of a data sample

Problem 3.6. Find modes for
$A = \{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$ and
$B = \{2, 3, 3, 3, 4, 6, 6, 6, 8, 9\}$.
Solution 3.6. A has the mode 4.

B is *bimodal* with two modes 3, 6.

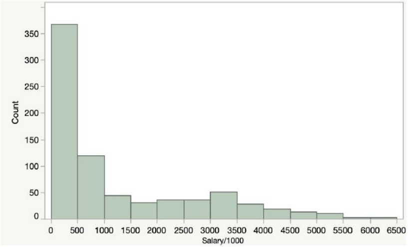
Actual instances of mean, median, mode

- Does the mean always belong to the sample? No. For example, the mean of integer numbers can be non-integer.
- Does the median always belong to the sample? Yes unless the sample size $n = 2k$ and $a_k \neq a_{k+1}$.
- Does a mode always belong to the sample? Yes, it does by the definition.

Comparison of mean, median, mode

Mean	Median	Mode
+ Easy to calculate + Good for symmetrical distributions + Good for ordered data	+ Good for skewed distributions +Good for ordered data	+ Always belong to the sample + Good for qualitative data
- Influenced by outliers - Does not always belong to the sample	- No easy maths formula - Does not always belong to the sample	- No easy maths formula - Not always unique - Can be away from the centre

The more the merrier!



Baseball salaries (1994) have mean $= \$1{,}183{,}000$, mode $= \$250{,}000$ and median $= \$500{,}000$.

Each of those central tendency measurements provides useful information.

Are means useful?

Chicago Jury Project (1960): the same outcome was reached by judges and lay decision-makers (juries) in 89% of all cases.

Besides, mean is not always an *arithmetic* mean: there are lots of mean creatures!

Geometric mean

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \cdots a_n}$$

is always *smaller than arithmetic mean*.

The quartiles Q_1, Q_2, Q_3 of a sample

Definition 3.7. The median is the *2nd quartile* Q_2. The idea of the *1st quartile* Q_1 is to separate the lowest 25% data values from the highest 75%. Similarly, the *3rd quartile* Q_3 separates the lowest 75% data values from the highest 25% values.

Quartiles in even cases

If a data sample consists of $2k$ ordered values $a_1 \leqslant \cdots \leqslant a_k \leqslant a_{k+1} \leqslant \cdots \leqslant a_{2k}$, then

Q_1 is the median of the lowest half $a_1, \ldots, a_k$;

Q_3 is the median of the highest half $a_{k+1}, \ldots, a_{2k}$.

Problem 3.7. Find the quartiles of the samples
$A = \{9, 4, 2, 4, 6, 4, 7, 4, 6, 4\}$,
$B = \{9, 6, 2, 3, 6, 8, 3, 4, 6, 3\}$.

Solution 3.7. $A = \{2, 4, 4, 4, 4, 4, 6, 6, 7, 9\}$ has $Q_1 = 4$, $Q_3 = 6$. $B = \{2, 3, 3, 3, 4, 6, 6, 6, 8, 9\}$ has $Q_1 = 3$, $Q_3 = 6$.

Quartiles in odd cases

If a data sample has $2k+1$ ordered values $a_1 \leqslant \cdots \leqslant a_k \leqslant a_{k+1} \leqslant \cdots \leqslant a_{2k+1}$, we will remove one median value a_{k+1}, then split the data into equal parts containing k values each.

Q_1 is the median of the lowest half $a_1, \ldots, a_k$;

Q_3 is the median of the highest half $a_{k+2}, \ldots, a_{2k+1}$.

Computing quartiles Q_1, Q_2, Q_3

Problem 3.8. Find the median and quartiles of the data samples $C = \{6, 5, -2, 7, 12, 5, 6, 3, 4\}$ and $D = \{6, 4, 8, 6, 12, 9, 5, 7, 6, 1, 6\}$.

Solution 3.8. The sample $C = \{-2, 3, 4, 5, 5, 6, 6, 7, 12\}$ has the median $Q_2 = 5$ and quartiles $Q_1 = 3.5,\ Q_3 = 6.5$.

The sample $D = \{1, 4, 5, 6, 6, 6, 6, 7, 8, 9, 12\}$ has the median $Q_2 = 6$ and quartiles $Q_1 = 5,\ Q_3 = 8$.

Can all descriptors be ordered for any sample?

Bounds for the sample mean

Claim 3.9. The mean of a data sample $a_1, \ldots, a_n$ is within the range (non-strictly between the minimum and maximum values).

Start from writing the claim in terms of $a_1, \ldots, a_n$.

Proof. Taking the sum of $a_i \geqslant \min\limits_{i=1,\ldots,n} a_i$ over all i, dividing by $n > 0$, we get $\bar{a} = \dfrac{1}{n}\sum\limits_{i=1}^{n} a_i \geqslant \min\limits_{i=1,\ldots,n} a_i$.
Similarly, taking the sum of $a_i \leqslant \max\limits_{i=1,\ldots,n} a_i$ over all i, dividing by $n > 0$, we get $\bar{a} = \dfrac{1}{n}\sum\limits_{i=1}^{n} \leqslant \max\limits_{i=1,\ldots,n} a_i$. $\square$

Bounds for the median

Claim 3.10. The median is always within $[Q_1, Q_3]$.

Start from writing the claim in terms of $a_1, \ldots, a_n$.

Proof. The median Q_2 is within the range $\left[\min_{i=1,\ldots,n} a_i, \max_{i=1,\ldots,n} a_i\right]$. Also by definition, Q_1 is the median of a 'half-sample' within $\left[\min_{i=1,\ldots,n} a_i, Q_2\right]$.

Similarly, Q_3 by definition is the median of the upper 'half-sample' within $\left[Q_2, \max_{i=1,\ldots,n} a_i\right]$. □

The 5-number summary of a sample

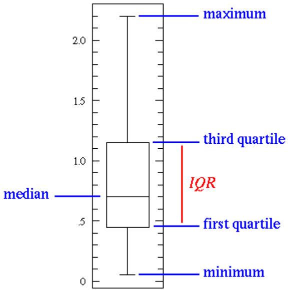
Definition 3.11. The 5-number summary of a data sample consists of the minimum, 3 quartiles and maximum: $\min\limits_{i=1,\ldots,n} a_i \leqslant Q_1 \leqslant Q_2 \leqslant Q_3 \leqslant \max\limits_{i=1,\ldots,n} a_i$.
The *Interquartile Range* $IQR = Q_3 - Q_1$.

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

Problem 3.12. Find outliers by the $1.5 \times IQR$ rule in the data samples $C = \{6, 5, 2, 7, 12, 5, 6, 3, 4\}$ and $D = \{6, 4, 8, 6, 12, 9, 5, 7, 6, 1, 6\}$.

Finding the Interquartile Range

Solution 3.12. $\{2, 3, 4, 5, 5, 6, 6, 7, 12\}$ has $IQR = Q_3 - Q_1 = 6.5 - 3.5 = 3$, so $12 \notin [3.5 - 1.5 \times 3, 6.5 + 1.5 \times 3]$ is an outlier.
$\{1, 4, 5, 6, 6, 6, 6, 7, 8, 9, 12\}$ has $IQR = 8 - 5 = 3$.
all values in $[5 - 1.5 \times 3, 8 + 1.5 \times 3]$, no outliers.



Definition 3.10. The *box plot* on the left represents any sample of scalar values by 2 intervals $[\min, \max]$ and $[Q_1, Q_3]$ using 5 descriptors.

Time to revise and ask questions

To benefit from the lecture, now you could

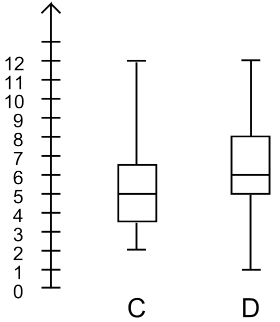
- ask or submit your questions
  on CANVAS or after the lecture;
- write down your summary in 2-3 phrases,
  e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Problem 3.13. Draw the box plots of the data samples $C = \{6, 5, 2, 7, 12, 5, 6, 3, 4\}$ and $D = \{6, 4, 8, 6, 12, 9, 5, 7, 6, 1, 6\}$.

Final solution

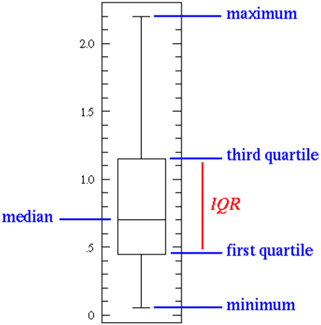
Solution 3.13. $C = \{2, 3, 4, 5, 5, 6, 6, 7, 12\}$ has 5 descriptors $2 < 3.5 < 5 < 6.5 < 12$.

The data sample $D = \{1, 4, 5, 6, 6, 6, 6, 7, 8, 9, 12\}$ has 5 descriptors $1 < 5 < 6 < 8 < 12$.



Summary

1) The mean, mode, median do not always coincide, each has its own uses and limitations.



2) The box plot represents a sample of scalar values by intervals $[\min, \max]$, $[Q_1, Q_3]$ and 5 descriptors $\min_{i=1,\ldots,n} a_i \leqslant Q_1 \leqslant Q_2 \leqslant Q_3 \leqslant \max_{i=1,\ldots,n} a_i$.