# Measuring Center in Quantitative Data

17.06.2025

## Statistics Intro: Mean, median, mode

* Statistics is the science of collecting, analyzing, interpreting, and presenting data. Descriptive statistics summarizes and describes the main features of a dataset; while Inferential statistics draws conclusions or makes predictions about a population based on sample data.

## Describing central tendency

* "What is the center/middle/typical value/average of this data?"
  * Arithmetic Mean: Sum of all values in a dataset, divided by the number of values. Perfect when the distribution is not skewed and has no outliers.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \quad \text{(For the sample (a subset) of the population)}$$

$$\mu = \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right) \quad \text{(For the population)}$$

  * Median: The middle value in an ordered dataset. Better when the data is skewed or has extreme outliers.

$$\text{Median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & n \text{ is even} \end{cases}$$

  * Mode: The value that appears the most frequently in a dataset. Mostly used only when the first two don't apply.

## Mean, median, mode example

* Find the mean, median, and mode of [23, 29, 20, 32, 23, 21, 33, 25] (8 values)

$$X = [20, 21, 23, 23, 25, 29, 32, 33]$$

$$\bar{x} = \frac{1}{8}\cdot\sum_{i=1}^{8} x_i = \frac{20+21+23+23+25+29+32+33}{8} = \frac{206}{8} = 25.75$$

$$\text{Median} = \frac{x_4 + x_5}{2} = \frac{23+25}{2} = 24 \qquad \text{Mode} = 23 \text{ (appears two-times)}$$

* Statistics is the science of collecting, analyzing, interpreting, and presenting data. Descriptive statistics summarizes and describes the main features of a dataset using measures of central tendency and spread, while inferential statistics draws conclusions or makes predictions about a population based on sample data using hypothesis testing, confidence intervals, and regression analysis.
* Measures of central tendency help us answer the question, "What is the center/middle/typical value/average of this data?". These measures are mean, median, and mode.

# More on Mean and Median

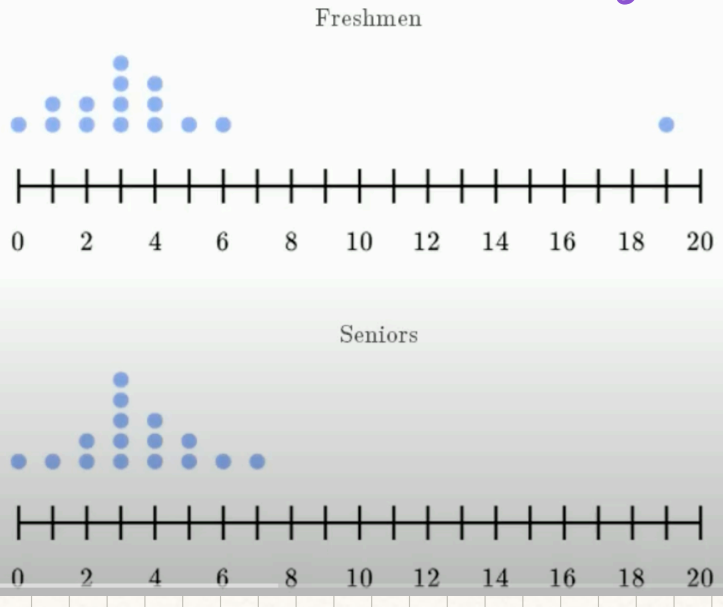
17.06.2025

## Comparing means of distributions

Kenny interviewed freshmen and seniors at his high school, asking them how many pieces of fruit they eat each day. The results are shown in the 2 plots below.

The mean number of fruits is greater for [ ]. Freshmen, b/c of the outlier

The mean is a good measure for the center of the distribution of [ ]. Seniors, b/c no outlier.




## Means and medians of different distributions

For his senior project, Richard is researching how much money a college graduate can expect to earn based on his or her major. He finds the following interesting facts:

- Basketball superstar Michael Jordan was a geology major at the University of North Carolina.
- There were only 3 civil engineering majors from the University of Montana. They all took the exact same job at the same company, earning the same salary.
- Of the 35 finance majors from Wesleyan University, 32 got high-paying consulting jobs, and the other 3 were unemployed.

For geology majors from UNC, the median income will likely be [ < ] the mean. (Right Tail)

For civil engineering majors from Montana, the median income will be [ = ] the mean. Symmetrical

For finance majors from Wesleyan, the median income will likely be [ > ] the mean.

## Impact on median & mean: Removing an outlier

* Removing an outlier significantly affects the mean, pulling it toward the remaining data, while the median remains relatively stable since it depends only on the middle values.

* Mean is the sum of all values in a dataset, divided by the number of values. Median is the middle value in an ordered dataset. Mode is the most frequent value in a dataset.

* Mean (for the sample) $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (for the population): $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

* Median $= \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ is even} \end{cases}$

**Impact on median & mean: Increasing an outlier**

* Increasing an outlier pulls the mean strongly in its direction but has little to no effect on the median, which depends only on the middle-ranked data points.

**Mean as the balancing point**

* We can think of the mean as the balancing point, because the total distance from the mean to the data points below the mean is equal to the total distance from the mean to the data points above the mean.

**Median & range puzzles**

* 11 sales people. Median cars sold = 6. Range cars sold = 4.

$$x = [\frac{x}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{6}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{\_}{6}, \frac{x+4}{10}]$$

* At least one sales person sold more than 10 cars:

True / False / Not Know

If $x+4 = 11$

$x = 7$ which means median can't be 6.

* Removing an outlier significantly affects the mean, pulling it toward the remaining data, while the median remains relatively stable since it depends only on the middle values. Increasing an outlier pulls the mean strongly in its direction but has little to no effect on the median. Bic of these reasons, mean is meaningful only if the data is not-skewed and there are not extreme outliers. If one of these conditions exist, it's better to use the median. If neither are applicable (e.g. data is categorical), the mode is used.

# Interquartile Range (IQR)

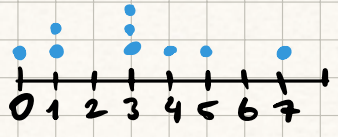
17.06.2025

* IQR is the amount of spread in the middle 50% of a dataset, which equals to the distance between the first quartile ($Q_1$) and the third quartile ($Q_3$), where $Q_1$ is the median of the data points to the left of the median (in the ordered list) and $Q_3$ is median of the data points to the right of the median.

* $X = [7, 3, 6, 4, 1, 5, 4, 5, 9, 1, 5] \Rightarrow IQR = ?$

$X_{sorted} = [1, 1, \boxed{3}, 4, 4, \boxed{5}, 5, 5, \boxed{6}, 7, 9] \Rightarrow IQR = 6 - 3 = 3$

($Q_1$ $Q_2$ $Q_3$)

* 

$\Rightarrow IQR = ?$

$X = [0, 1, 1, 3, \boxed{3}, 3, 4, 5, 7] \Rightarrow IQR = Q_3 - Q_1 = 4.5 - 1 = 3.5$

($Q_1 = 1$, $Q_2$, $Q_3 = 4.5$)

* $X = [81, 82, 83, 83, 84, 84, 84, 85] \Rightarrow IQR = ?$

($Q_1$ 82.5, $Q_2 = 83.5$, $Q_3$ 84)

$= 84 - 82.5 = 1.5$

**Range vs IQR**

* Range is simple but misleading for skewed data, while IQR focuses on central data, ignoring outliers.

---

* We can think of the mean as the balancing point, b/c the total distance from it to the data points below is equal to the total distance from it to the data points above it.
* Interquartile Range (IQR) is the amount of spread in the middle 50% of a dataset, which equals to the distance between the first quartile ($Q_1$) and the third ($Q_3$), where $Q_1$ is the median of the data points to the left of the median (in the ordered list) and $Q_3$ is the median of the data points to the right of the median.
* Range (max−min) is simple but misleading for skewed data, while IQR focuses on central data, ignoring outliers.

# Variance and Standard Deviation of a Population

18.06.2025

**Measures of Spread: Range, Variance, and Standard deviation**

* Measures of spread help us to answer the question. "How spread apart is the data?" These measures are range, variance, and standard deviation.

* $X_1 = [-10, 0, 10, 20, 30]$ $\qquad$ $X_2 = [8, 9, 10, 11, 12]$

$\mu_{X_1} = \frac{1}{5}(50) = 10$ $\qquad = \qquad$ $\mu_{X_2} = \frac{1}{5}(50) = 10$

* Their center is the same, so are they similar? NO! B/C: $X_1$ is more disperse (data points are farther from the mean).

$\text{Range}_{X_1} = 30 - (-10) = 40$ $\qquad$ $\text{Range}_{X_2} = 12 - 8 = 4$

* Variance (for population) $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$

Variance (for sample) $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

* $\sigma^2_{X_1} = \frac{1}{5}\left((-10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2 + (30-10)^2\right)$

$= \frac{1}{5}[400 + 100 + 0 + 100 + 400] = 200$

$\sigma^2_{X_2} = \frac{1}{5}[(8-4)^2 + (9-4)^2 + (10-4)^2 + (11-4)^2 + (12-4)^2]$

$= \frac{1}{5}[16 + 25 + 36 + 49 + 64] = 38$

* $200 > 38 \Leftrightarrow X_1$ is more disperse than $X_2$.

* Standard deviation is the square root of variance. Because it shares the same unit of measurement as the original dataset, it provides a more intuitive and practical measure of spread than variance.

---

* "How spread apart is the data?": Measures of spread (range, variance, standard deviation)
* Range $= X_{max} - X_{min}$
* Variance (for population) $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$ * Variance (for sample) $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
* Standard deviation (for population) $= \sigma = \sqrt{\sigma^2}$ * Standard deviation (for sample) $= s = \sqrt{s^2}$

## Variance of population

* Years of experience = [1, 3, 5, 7, 14] ⇒ $\mu_x = ?$, $\sigma = ?$
② Khan Academy

$$\mu = \frac{1}{5}[1+3+5+7+14] = 6 \text{ years}$$

$$\sigma^2 = \frac{1}{5}\left[(1-6)^2+(3-6)^2+(5-6)^2+(7-6)^2+(14-6)^2\right]$$

$$= \frac{1}{5}[25+9+1+1+64] = 20 \text{ squared years}$$

$$\Rightarrow \sigma = \sqrt{20} \approx 4.47 \text{ years}$$

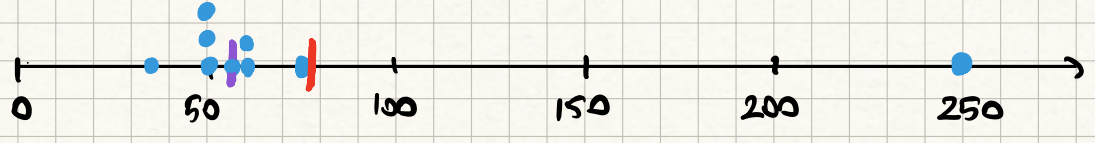
## Mean and standard deviation vs. median and IQR

* $X = [35, 50, 50, 50, 56, 60, 60, 75, 250]$

$\mu \approx 76.2$ Median = 56

$\sigma \approx 62.3$ IQR = 17.5

⇒ What is the central tendency and spread of the dataset?




* We have a skewed dataset, so median is better for central tendency. Since the standard deviation is based on the mean, it's not the better measure for this dataset.

* Standard deviation shares the same unit of measurement as the original dataset, so it provides a more intuitive and practical measure of spread than variance.

* If the data is skewed (and/or has extreme outliers), the standard deviation (or variance) is not the best measure to use for spread, b/c it's based on the mean, which itself is not the best measure to use for the central tendency in these conditions. For these type of data, the median for central tendency and the IQR for spread are better choices.

## Alternate variance formulas

* $\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$

$= \frac{1}{N} \left( \sum_{i=1}^{N} (x_i^2 - 2x_i\mu + \mu^2) \right)$

$= \frac{1}{N} \left( \sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \sum_{i=1}^{N} 1 \right)$

* means we add 1 to itself N times, which is N times 1 = N

$= \frac{1}{N} \left( \sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \cdot N \right)$

* Adding all data points and dividing by N, is same as finding the mean.

$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu\mu + \mu^2$

$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{\left[\sum_{i=1}^{N} x_i\right]^2}{N^2}$

* $x = [8, 9, 10, 11, 12] \Rightarrow \sigma^2 = ?$

$= \frac{(64+81+100+121+144)}{5} - \frac{(8+9+10+11+12)^2}{25}$

$= \frac{510}{5} - \frac{50 \cdot 50}{25} = 102 - 100 = 2$

# Variance and Standard Deviation of a Sample

18.06.2025

## Sample Variance

* $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$

* What is the average TV watch time for 300 million Americans? And what is the variance? We can't observe all of them so we observe a sample: [1.5, 2.5, 4, 2, 1, 1] (6 samples not enough ofc but for simplicity)

$$\bar{x} = \frac{1}{6}(1.5 + 2.5 + 4 + 2 + 1 + 1) = 2 \text{ hours}$$

↳ sample mean

$$s^2 = \frac{1}{6-1}\left((0.5)^2 + (0.5)^2 + (2)^2 + (0)^2 + (1)^2 + (1)^2\right)$$

$$= \frac{1}{5}\left(\frac{1}{4} + \frac{1}{4} + 4 + 0 + 1 + 1\right)$$

$$s^2 = \frac{1}{5}(6.5) = 1.5 \text{ squared hours}$$

$s^2$ ← sample variance

$$s \approx 1.22 \text{ hours}$$

$s$ ← sample std. deviat.

## Sample standard deviation and bias

* When the sample mean ($\bar{x}$) is calculated, $(x_i - \bar{x})$ are smaller than true deviations $(x_i - \mu)$, which causes the sum of squared deviations to be smaller than it should be, leading to underestimation of the variance. To adjust this, we divide by $(n-1)$ instead of $n$.

* When the sample mean ($\bar{x}$) is calculated, $(x_i - \bar{x})$ are smaller than true deviations $(x_i - \mu)$, which causes the sum of squared deviations to be smaller than it should be, leading to underestimation of the variance. To adjust this, we divide by $(n-1)$ instead of $n$.

# More on Standard Deviation

19.06.2025

## Why we divide by (n-1) in variance

* The best way to understand why using (n) instead of (n-1) for samples underestimates the variance is through simulation. You can check:

https://www.khanacademy.org/cs/unbiased-variance-visualization/1167453164

* Why (n-1)? Not any other number?

* (n-1) arises because one degree of freedom is lost. If we were estimating more parameters (e.g. in linear regression) we might divide by (n-p) where p is the number of estimated parameters.

* (n-1) arises because one degree of freedom is lost. If we were estimating more parameters (e.g. in linear regression) we might divide by (n-p) where p is the number of estimated parameters.

# Box and Whisker Plots

19.06.2025

## Constructing a box plot

* A box-and-whisker plot displays the five-number summary of a dataset:

① The minimum (lower whisker)

② First quartile, $Q_1$ (box edge)

③ Median, $Q_2$ (line inside the box)

④ Third quartile, $Q_3$ (box edge)

⑤ The maximum (upper whisker)

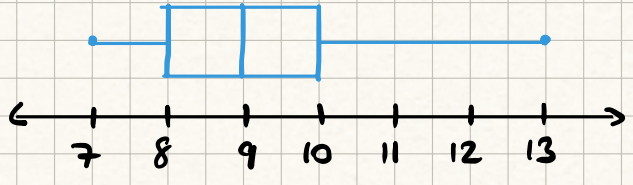
with potential outliers marked individually.

## Example with odd number of data points

* $X = [13, 9, 8, 10, 7, 10, 9, 8, 9, 8, 9]$ as a box-and-whisker plot?

$X_{sorted} = [7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 13]$

min $Q_1$ $Q_2$ $Q_3$ max

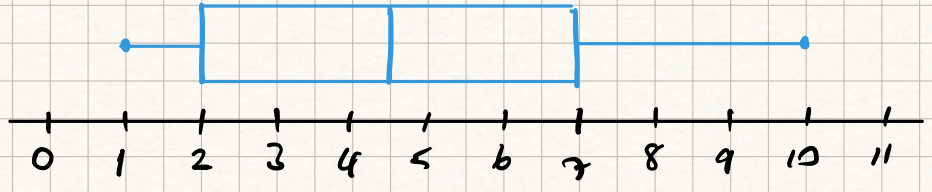



## Example with even number of data points

* $X = [5, 2, 10, 2, 6, 8, 8, 1, 2, 5, 7, 3, 3, 4]$

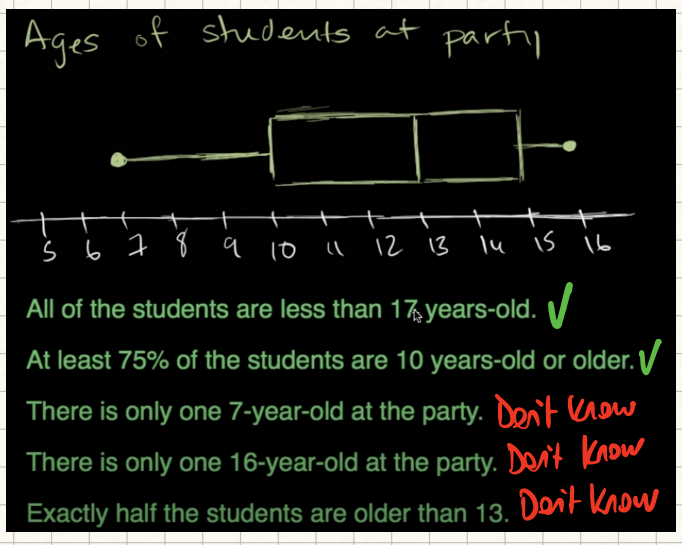
$X_{sorted} = [1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 10]$

min $Q_1$ $Q_2 = 4.5$ $Q_3$ max




* A box-and-whisker plot displays the five-number summary of a dataset: ① The minimum ② First quartile (box edge) ③ Median (line inside the box) ④ Third quartile (box edge) ⑤ The maximum, with potential outliers marked individually.

## Interpreting box plots




12 14 ---- TRUE

13 ---- FALSE

## Judging outliers in a dataset

Outlier = A data point that lies an abnormal distance from other

Lower outlier threshold = $Q_1 - 1.5 \cdot IQR$

Higher Outlier threshold = $Q_3 + 1.5\ IQR$

$X = [1, 1, 6, 13, 13, 14, 14, 14, 15, 15, 16, 18, 18, 18, 19]$

$\Rightarrow Q_1 = 13$

$Q_2 = 14$

$Q_3 = 18$

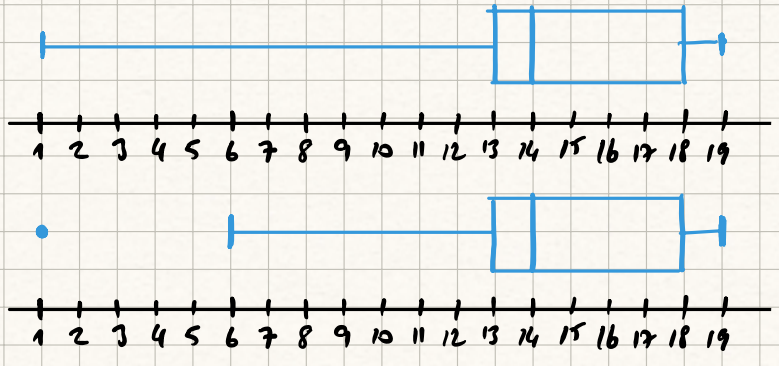
$IQR = 18 - 13 = 5$ Threshold $= 5(1.5) = 7.5$

Outliers → $x < 13 - 7.5 = x < 5.5$

→ $x > 18 + 7.5 = x > 25.5$

Outliers = $\{1, 1\}$

## Box plots excluding outliers




An outlier is a data point that lies on abnormal distance from the others. The standard threshold for outliers are $Q_1 - 1.5\ IQR$ (for lower) and $Q_3 + 1.5\ IQR$ (for upper).

# Other Measures of Spread

19.06.2025

## Range and mid-range

* Range $= x_{max} - x_{min}$
* Mid-Range $= \frac{x_{max} - x_{min}}{2}$
* Very sensitive to outliers!

## Mean absolute deviation

* $MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$ . MAD is less sensitive to extreme values than variance/standard deviation b/c it uses absolute differences instead of squares. However it lacks mathematical properties exploited in advanced statistics (e.g. differentiability). Mainly used where outlier distortion must be minimized (e.g. finances, engineering)

* $x_1 = [2,2,4,4]$

$\bar{x}_1 = \frac{2+2+4+4}{3} = 3$

$MAD_{x_1} = \frac{1}{4}(|2-3| + |2-3| + |4-3| + |4-3|)$

$= \frac{1}{4}(1+1+1+1) = 1$

$x_2 = [1,1,6,4]$

$\bar{x}_2 = \frac{1+1+6+4}{4} = 3$

$MAD_{x_2} = \frac{1}{4}(|1-3| + |1-3| + |6-3| + |4-3|)$

$= \frac{1}{4}(2+2+3+1)$

$= 2$

* "$x_2$ is more spread out than $x_1$"

---

* Range $(x_{max} - x_{min})$ and mid-range (range/2) are also measures of spread, but they're very sensitive to outliers.
* Mean Absolute Deviation = $MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$ . MAD is less sensitive to extreme values than variance/std. dev. b/c it uses abs. differences instead of squares. However, it lacks mathematical properties exploited in advanced statistics, such as differentiability.