Statistics Basics

Mean, median, mode and range are the four numbers that summarise a whole column. What each one means, why the average alone can lie, and one simple distribution example.

Introduction

A class has thirty students. Nobody wants to read thirty marks to understand the class — they want one or two numbers that summarise them. **Statistics** is the set of tools that turns a whole column of numbers into a few honest summary numbers, and tells you how far you can trust them.

The four summaries every beginner needs are the mean (the average), the median (the middle value), the mode (the most common value), and the range (highest minus lowest). They sound similar, but they answer different questions, and on messy data they can disagree wildly. Knowing which one to use — and when the average is lying to you — is the difference between a useful summary and a misleading one.

A quick rule of thumb will save you many wrong summaries. If the data is fairly even — no extreme values — the mean is fine and uses every value. If the data is lopsided — one huge sale, one top scorer far ahead — report the median beside the mean, because the median describes the typical value better. And always pair a middle number with a spread number: an average of 70 means something very different when marks run from 68 to 72 than when they run from 20 to 100. Middle without spread is half a summary.

(Roman Urdu: Tees numbers parhna mushkil hai, is liye hum chand numbers se poori class ka haal batate hain — average, darmiyani number, sab se aam number, aur sab se bara aur chhota ka farq. Har ek ka apna kaam hai.)

Key points

- **Mean = add everything, divide by how many.** It uses every value, so one extreme value pulls it hard.
- **Median = the middle value when sorted.** Half the values sit below it, half above. Extreme values barely move it.
- **Mode = the most common value.** Useful for categories too: the most common subject choice, the most common shoe size.
- **Range = highest minus lowest.** It tells you the spread in one number, but it only looks at the two extremes.
- **The average can lie when data is uneven.** If one student scores 100 and nine score 40, the mean is 46 — but "average 46" hides that almost everyone scored 40.
- **When mean and median disagree, look closer.** A big gap between them is a warning: the data is lopsided, and you should look at the full spread, not one number.
- **A distribution is just the shape of the data:** are values bunched in the middle, spread flat, or piled at one end? A histogram from lesson DS.5 shows this shape at a glance.

- **Even counts need a two-step median.** With an even number of values, there is no single middle value — take the average of the two middle values instead (Example 4 shows this).
- **Standard deviation, in plain words, is the typical distance from the average.** A small standard deviation means values hug the average; a big one means they scatter far from it. It is the grown-up version of the range, using every value instead of just the two extremes.

Examples

Example 1 — Mean, median, mode and range by hand

```
Marks, sorted: 40, 45, 50, 55, 60, 60, 70, 80, 95

Mean:   (40 + 45 + 50 + 55 + 60 + 60 + 70 + 80 + 95) / 9
      = 555 / 9 = 61.7
Median: the middle value (5th of 9) = 60
Mode:   the most common value = 60 (it appears twice)
Range:  95 - 40 = 55
```

What this means: here the mean (61.7) and the median (60) agree closely, so the average is a fair summary. The range of 55 warns you that the class is still spread wide — the average alone would hide that spread.

Example 2 — When the average lies

Plain Python — no installs needed:

```
marks = [40, 40, 40, 40, 40, 40, 40, 40, 40, 100]

mean = sum(marks) / len(marks)
print("Mean:", mean)  # 46.0

sorted_marks = sorted(marks)
middle = sorted_marks[len(sorted_marks) // 2]
print("Median:", middle)  # 40
```

What this means: the mean says 46, but nine out of ten students scored 40. The single 100 pulled the average up by 6 marks. The median says 40, which describes the typical student far better. Lesson: on lopsided data, report the median beside the mean, and say why they differ.

Example 3 — The same summaries with NumPy

This example needs `pip install numpy` first:

```
import numpy as np

marks = np.array([40, 45, 50, 55, 60, 60, 70, 80, 95])

print("Mean:", np.mean(marks))      # 61.7
print("Median:", np.median(marks))  # 60.0
print("Range:", np.max(marks) - np.min(marks))  # 55
```

What this means: NumPy has `mean` and `median` built in, and the range is just max minus min. There is no built-in mode in NumPy — for the mode, count values by hand on small data, or use a pandas column's `value_counts()` from lesson DS.3, which counts every value and shows the most common one at the top.

Example 4 — Median with an even count of values

```
Marks, sorted: 50, 60, 70, 80   (4 values, even count)

Two middle values: 60 and 70
Median: (60 + 70) / 2 = 65
Mean:   (50 + 60 + 70 + 80) / 4 = 65
Range:  80 - 50 = 30
```

What this means: with an even count you cannot just point at the middle — you average the two values straddling it. Here mean and median agree at 65, which again tells you the data is fairly even. If you had picked just 60 or just 70 as "the median", you would have been wrong by 5 marks either way.

Example 5 — The mode for categories, with pandas

This example needs `pip install pandas` first:

```
import pandas as pd

choices = pd.Series(["Maths", "Maths", "English", "Maths",
                     "Urdu", "English", "Maths"])
print(choices.value_counts())
print("Mode:", choices.mode()[0])
```

Output:

```
Maths      4
English    2
Urdu       1
Mode: Maths
```

What this means: the mode is not only for numbers — for categories it is simply the most popular choice. Four students picked Maths, so Maths is the mode. A school choosing which subject needs an extra teacher would

start exactly here: the most common choice is where the demand sits.

Example 6 — Real-world case: a shop's daily sales, and one festival day

A small shop records daily sales for ten days: nine ordinary days around 5,000 rupees, and one festival day of 40,000. The mean is (9 x 5,000 + 40,000) / 10 = 8,500 — but no ordinary day ever reached 8,500, so "average day 8,500" misleads anyone planning stock. The median is 5,000, which describes a normal day honestly. The same trap appears in salaries: a company where the boss earns ten times the staff can truthfully advertise a high "average salary" while most workers earn far less. Whenever a mean sounds too good, ask for the median — and for how many values sit behind it.

(Roman Urdu: Festival wale din ki sale average ko upar khainch leti hai, lekin aam din ki asal tasveer median batata hai.)

Quick recap

- Mean: total divided by count; sensitive to extreme values.
- Median: middle value when sorted; steady even with extremes.
- Mode: most common value; works for numbers and categories.
- Range: highest minus lowest; shows spread, but only from the extremes.
- If mean and median disagree a lot, the data is lopsided — report both.

Common mistakes

1. **Reporting only the mean on lopsided data** — one rich shop, one top scorer, or one huge sale drags the average away from the typical value. Fix: report the median beside it, and check the range.
2. **Thinking the median is "the average" in every case** — the median ignores how far values sit from the middle. Fix: use mean and median together; they describe different things.
3. **Forgetting to sort before finding the median** — the middle of an unsorted list is just a random value. Fix: sort first, then take the middle; on even counts, take the average of the two middle values.
4. **Reporting the range as if it described the whole spread** — the range uses only the highest and lowest values, so one extreme value makes the spread look huge even when everyone else is bunched together. Fix: report the median and the middle of the data too, or the standard deviation, not the range alone.
5. **Comparing averages from groups of very different sizes** — an average from 3 students and an average from 300 students are not equally solid. Fix: always give the count beside the average: "average 82 (30 students)" versus "average 88 (3 students)".

Practice

1. Take these marks: 55, 70, 70, 80, 95. Find the mean, median, mode and range by hand, then check with Python.
2. Add one extreme value — 20 — to that list and recompute. Which summary moves most, and which barely moves?

3. Pick a real set of numbers from your life (daily steps for a week, or prices of five notebooks) and summarise it with all four numbers in two sentences.

1. Take these marks: 60, 70, 80, 90. Find the median step by step (even count), then check with `np.median`. Why is the median not simply 70 or 80?
2. Make a list of 8 values where the mean and the median differ by more than 10. Write one sentence explaining, in plain words, why they differ.
3. Ask 5 friends or family members for a simple category (favourite subject, favourite fruit). Count the answers and state the mode. What decision could that mode help someone make?

Practice quiz

1. How do you find the mean?

- Pick the middle value
- Add all values and divide by how many there are
- Pick the most common value
- Subtract smallest from largest

Answer: Add all values and divide by how many there are

2. What is the median?

- The average
- The middle value when values are sorted
- The most common value
- The largest value

Answer: The middle value when values are sorted

3. Why can an average lie?

- Averages are always wrong
- A few very large values pull the mean up, so it no longer describes a typical value
- Averages ignore small values
- Means cannot be trusted at all

Answer: A few very large values pull the mean up, so it no longer describes a typical value

---

Watch the video lesson: https://youtu.be/pHi3x-4sHyo

Read online: https://scopetreasure.com/notes/ds-statistics-basics · More free tools and notes at https://scopetreasure.com