keenonmaths

Lying with statistics

seventy years on

Part 1 · Averages, spread and charts

Science in the news

written by **Nicola Maglione** — keenonmaths

@keenonmaths

November 2026

Contents

Introduction 2

1. Before counting: who are we looking at? 3

2. Mean, median and mode: three ways of saying “usually” 4

3. How spread out are the data: variance and standard deviation 8

4. Percentage points and charts with a truncated axis 10

Let’s recap 11

In the second part 12

Appendix — For those who want the maths 12

Sources 14

Introduction

On 29 October 2026 Istat, Italy's national statistics institute, published the September labour market figures [1]. With those numbers you can write two opposite headlines without telling a single lie:

Headline trap *(made up)*

Unemployed up 7.8% in a year: jobs alarm

Headline trap *(made up)*

Jobs: 198,000 more people in work than a year ago

Both are true. But how can it be that, in the same year, *both* the number of people in work *and* the number of people looking for work without finding it went up?

The key is a group of people who almost never appear in headlines. Istat divides working-age people into three groups, not two:

Definition — Employed, unemployed, inactive

- **Employed**: people who have a job.
- **Unemployed**: people who do not have a job, are **actively looking** for one and are ready to start straight away.
- **Inactive**: people who do not have a job and are **not looking** for one: students, retired people, people looking after the home or the family, people who have become discouraged and stopped looking.

Between September 2025 and September 2026 the number of inactive people aged 15 to 64 fell by almost 300,000 [1]. Many of them are people who were not looking for work before and started looking again: some found a job and became employed, others are still looking and became unemployed. This is why the two numbers can rise together. And the rise in unemployment, which the first headline presents as an alarm, is also a sign of people getting back into the job market.

Then there is a second, subtler trick: the choice between percentages and absolute numbers. There are about 1.6 million unemployed people in Italy: on such a small base, the 117,000 extra unemployed become an impressive +7.8%. There are more than 24 million employed people: 198,000 more, as a percentage, would be a mere +0.8%. The alarmist headline uses the percentage, which sounds huge for the unemployed; the optimistic one uses the absolute number, which sounds big for the employed. Swap the choices, "employment +0.8%" and "117,000 more unemployed", and the effect is reversed.

Neither headline contains a wrong number. Yet someone who reads only the first and

someone who reads only the second come away with two opposite pictures of the same country. You do not need to be an expert to notice: a few tools are enough, the ones taught at school and often seen as a mere numerical exercise, an end in itself.

At school, in fact, statistics is often treated as a secondary topic, a chapter to rush through, time permitting, after the "important" ones. Mean, median, mode, variance: they look like test material, far from real life. The opposite is true: they are among the most useful tools for reading the world, and also among the easiest to misuse. If you know the rules of the game, you notice when someone cheats.

In 1954 the American journalist Darrell Huff published a short book with a provocative title: *How to Lie with Statistics* [2, 3]. It was not a manual for cheats, quite the opposite: it showed the most common tricks so that readers could learn to recognise them. It became the most widely read popular statistics book of all time, and it is a book I quote often whenever statistics comes up. This article was born from it. More than seventy years have passed, charts are now made on computers and headlines travel on social media, but the tricks Huff described are more relevant than ever.

In two articles we will go looking for them using only recent official data: Istat's surveys of household spending and of the labour market, the Bank of Italy's data on wealth, the national school tests run by INVALSI (Italy's institute for the evaluation of the school system) and the OECD's international PISA survey (Programme for International Student Assessment). For each statistical tool we will first see what it measures, then the "trick" that can bend it. In this first part we deal with how data are **described**: means, medians, spread, percentages and charts. In the second part we will move on to how they are **estimated**: samples, margins of error and correlations. The headline traps we will meet are made up by me, but the numbers they are built on are all real.

1. Before counting: who are we looking at?

Nobody has interviewed every Italian household to find out how much they spend. Istat interviewed about 32,500 of them in a year, in 542 municipalities, picked at random from the population register [4]. Starting from those households, it **estimates** the figures for all households living in Italy.

> **Definition — Population and sample**
>
> The **population** is the set of all the elements we are interested in: all households living in Italy, all students in the last year of upper secondary school, all people aged 15 to 74.
>
> A **sample** is the part of the population on which measurements are actually made. If it is chosen **at random**, following precise rules, the sample "resembles" the population and makes it possible to estimate its characteristics.

Tasting a spoonful of sauce is enough to tell whether it is salty, provided you have stirred it well first. A random sample is a spoonful taken from well-stirred sauce; a badly chosen sample (only the people who answer an online poll, only your friends, only the people who live in the town centre) is a spoonful taken from the surface of a sauce that was never stirred.

2. Mean, median and mode: three ways of saying “usually”

Three definitions

Definition — Arithmetic mean

Given n values $x_1, x_2, \ldots, x_n$, their **mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

It is the value each element would have if the total were divided into equal parts.

Definition — Median

Sort the values from smallest to largest. The **median** is the value in the middle: half of the data are less than or equal to it, half are greater than or equal to it. If there is an even number of values, take the mean of the two middle ones.

Definition — Mode

The **mode** is the value that occurs most often. When data are grouped into intervals (for example spending bands), we speak of the **modal class**: the interval containing the most data.

A small example

Nine friends receive the following monthly pocket money, in euros, already sorted from smallest to largest:

$$15, \quad 20, \quad 20, \quad 20, \quad 25, \quad 30, \quad 30, \quad 35, \quad 165.$$

- The **mean** is

$$\bar{x} = \frac{15 + 20 + 20 + 20 + 25 + 30 + 30 + 35 + 165}{9} = \frac{360}{9} = 40 \text{ euros}.$$

- The **median** is the fifth value, the one with four values before it and four after it: **25 euros**.
- The **mode** is **20 euros**, the value that occurs most often (three times).

Which number best describes the “typical pocket money” of the group? Eight friends out of nine get 35 euros or less, yet the mean says 40, more than anyone except one receives. A single value much larger than the others is enough to drag it upwards. The median, on the other hand, barely notices: if the lucky friend received 1,000 euros instead of 165, the mean would rise to about 133 euros, while the median would stay at 25.

Remark — The mean is sensitive to extreme values, the median is not

The mean uses the value of every data point, so one data point far from the others moves it a lot. The median uses only position: what counts is who is in the middle, not how large the values at the top are.

Italian household spending in 2025

Now the same tools on real data. According to Istat, in 2025 a household living in Italy spent on average **2,782 euros a month** on consumption; median spending was **2,338 euros** [4]. The mean exceeds the median by about 445 euros, almost 20%.

The reason is the same as with the pocket money. Household spending is **skewed**: most households are concentrated at low-to-middle levels, while a smaller number of households spend a lot, stretch the distribution to the right and pull the mean up. This is why, as Istat itself points out, the majority of households spend less than the mean [4].

Definition — Skewed distribution

A distribution is **symmetric** if the data are arranged in the same way to the left and to the right of the central value (the median), like the two halves of a mirror image; in that case the mean coincides with the median. It is **right-skewed** (or positively skewed) if it has a long "tail" towards large values: then, as a rule, mode $<$ median $<$ mean.

To show the shape of the distribution, Istat publishes the **deciles** of spending.

Definition — Quantiles: quartiles, quintiles, deciles

Once the data are sorted, **quantiles** are the values that split them into groups of equal size. **Deciles** split them into 10 groups, **quintiles** into 5, **quartiles** into 4. The median is the fifth decile, and also the second quartile. Mind the names: quantiles are the boundary values; the groups that form between one boundary and the next are called **tenths**, **fifths**, **quarters**.

There is a detail, though. In the Istat survey we are using, the deciles do not refer to households' actual spending, but to a slightly different quantity, **equivalised spending**. The reason is simple: a family of four spends more than a person living alone, but that does not mean it lives better. To compare households of different sizes, spending has to be "corrected".

Definition — Equivalised spending

A household's **equivalised spending** is its spending divided by a coefficient that depends on the number of household members. Istat uses the **Carbonaro** equivalence scale, which takes a two-person household as its reference:

Members	1	2	3	4	5	6	7 or more
Coefficient	0.60	1.00	1.33	1.63	1.90	2.16	2.40

The coefficients grow less than proportionally with the number of people, because many expenses are shared: the gas bill of a family of four is not four times that of a person living alone [4].

An example. In 2025 a person living alone spent on average 2,070 euros a month [4]; divided by 0.60, that is about 3,450 equivalised euros: what a couple would have to spend to enjoy the same standard of living. There are many people living alone in Italy, which is why the mean and median of equivalised spending are higher than those of actual spending: the equivalised median is 2,467 euros, against 2,338 for actual spending. It is a good reminder of a rule that will come up often: before comparing two numbers, make sure they measure **the same thing**.

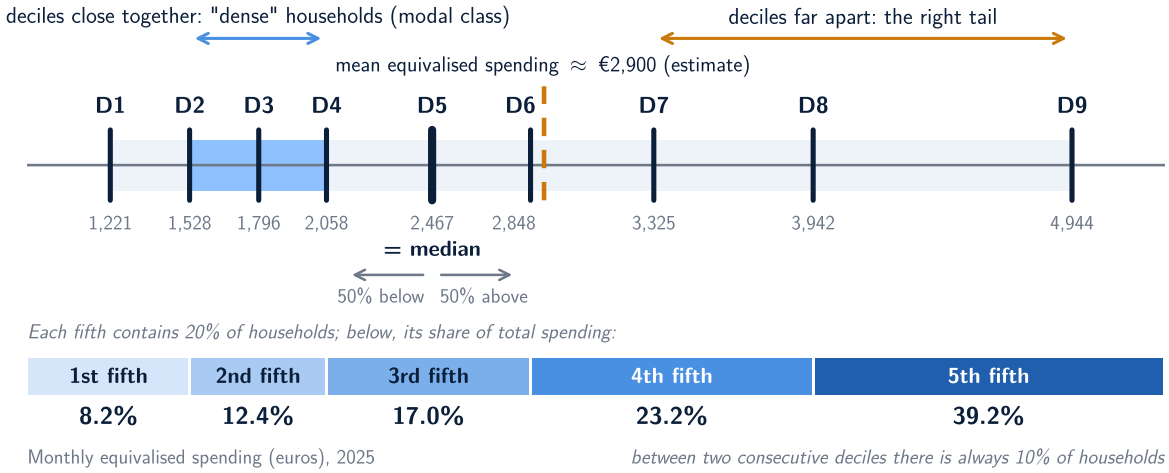

Figure 1. The nine deciles of monthly equivalised spending of Italian households in 2025, marked on a line. Between one decile and the next there is always 10% of households. Where the marks are close together, households are "dense"; where they move apart, households are sparse: the long tail is on the right. The fifth decile, D5, is the median: half of households spend less, half spend more. The dashed line marks the mean of equivalised spending, about 2,900 euros: the Istat report does not publish it, and we estimated it from the spending shares of the five household quintiles. Mean to the right of the median: the signature of a right-skewed distribution. In the band at the bottom, households are split into five fifths, bounded by the deciles D2, D4, D6 and D8: each fifth contains 20% of households, and below each one is the share of total spending it accounts for. The width of the blocks measures the range of spending, not the number of households; the first and last fifths have no outer boundary and continue beyond the edges of the figure. The mean and median in the figure differ from those of actual spending quoted in the text (2,782 and 2,338 euros) because they refer to equivalised spending. Source: Istat, *Le spese per i consumi delle famiglie – Anno 2025* (Household consumption expenditure – Year 2025), glossary.

The deciles closest together are those between about 1,530 and 2,060 euros: that is where the modal class lies, the "most common level of spending". Moving to the right, the intervals grow wider and wider: between the eighth and the ninth decile there are almost a thousand euros [4].

The same imbalance can be seen in another way. Line up households, as in Figure 1, from the lowest equivalised spending to the highest, and split them into five groups with the same number of households, which Istat calls **fifths** (they are the band at the bottom of the figure). Then add up the spending of each group and see what share it represents of the spending of all households together. If everyone spent the same, each group would account for 20%. In 2025, however, the fifth that spends least accounts for 8.2% of total spending and the one that spends most for 39.2% [4]: the top fifth spends about 4.8 times as much as the bottom one.

Headline trap *(made up)*

Italian households spend 2,782 euros a month on average

Mathematically it is correct. But readers tend to picture a "normal" household spending roughly that amount, and that is not the typical household at all: most households spend less. The household in the middle, the median one, spends about 2,338 euros, and the most frequent level of spending is lower still. To describe the "typical" household, the right numbers are the median or the mode, not the mean.

An extreme case: wealth

With household wealth (homes, savings, investments, net of debts) the skew becomes enormous. According to the Bank of Italy, at the end of 2025 **mean** net wealth was 453,000 euros per household; but the richest 10% of households owned 60.6% of it, while the less wealthy half owned just 7.2% in total [5].

"The average Italian household owns 453,000 euros" is a true sentence. Said on its own, though, it suggests a widespread wealth that, for half of households, does not exist.

Key figures — Mean and median in official data

- Consumption spending 2025: mean €2,782, median €2,338 a month [4].
- North-East: mean €3,087, median €2,587. South: mean €2,215, median €1,913 [4].
- Net wealth, end of 2025: mean €453,000 per household; the richest 10% own 60.6% of it, the less wealthy half 7.2% [5].

3. How spread out are the data: variance and standard deviation

Mean and median tell us where the "centre" is. They do not tell us how close to or far from that centre the data are.

Two classes take the same test. In the first, the marks are $6, 6, 6, 6, 6$; in the second, $4, 5, 6, 7, 8$. In both, the mean is 6 and the median is 6. Yet they are very different classes: in the first, everyone is at the same level; in the second, some students are struggling and some are doing very well. We need a number that measures **spread**.

Definition — Variance and standard deviation

For each data point, compute its **deviation** from the mean, $x_i - \bar{x}$. The **variance** is the mean of the squared deviations:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}.$$

The **standard deviation** is its square root, $\sigma = \sqrt{\sigma^2}$. It has the same unit of measurement as the data and indicates, roughly, how far a data point "typically" lies from the mean.

Why the squares? The natural idea would be simply to take the mean of the deviations. But positive and negative deviations cancel out exactly, for any set of data: **their sum is always zero** (the two-line proof is in the Appendix, A.1). In words: the mean is the "balance point" of the data. If the data were equal weights placed on a rod, the mean would be the point where you would have to support the rod to keep it balanced.

The mean of the deviations, therefore, is always zero and says nothing about spread. Squaring makes all the deviations positive, and gives large ones more weight. (One could also use the absolute value of the deviations, and that measure does exist; squares, however, have much more convenient mathematical properties.)

In the second class the deviations are $-2, -1, 0, 1, 2$ and their squares $4, 1, 0, 1, 4$: the variance is $10/5 = 2$ and the standard deviation is $\sqrt{2} \approx 1.4$ marks. In the first class all deviations are zero: variance and standard deviation are both 0.

Why do we sometimes divide by $n - 1$?

Scientific calculators and spreadsheets often offer two different standard deviations: one with n in the denominator and one with $n - 1$. The one with $n - 1$ is used when the data are a **sample** and we want to estimate the variance of the **population** they come from:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.$$

Where does that -1 come from?

The variance of the population is computed from the deviations from the *true* mean of the population, which we call μ. But we do not know μ: we only have the sample mean $\bar{x}$, and we use that instead. The trouble is that $\bar{x}$ is computed from those very data, so it "sits closer to them" than μ does: the deviations from $\bar{x}$ are, in total, systematically smaller than the deviations from μ. Dividing by n, the variance of the population would be **underestimated**: on average we would get $\frac{n-1}{n}\,\sigma^2$ instead of σ^2. Dividing by $n - 1$ corrects exactly this flaw. The

full calculation, which uses only high-school algebra plus one result from probability, is in the Appendix (A.2 and A.3).

An example you can check by hand. A population contains only two values, 4 and 8. Its mean is $\mu = 6$ and its variance is $\sigma^2 = \frac{(4-6)^2+(8-6)^2}{2} = 4$. We draw two values at random, putting the drawn value back into the urn each time. There are four possible samples, all equally likely:

Sample	**Mean $\bar{x}$**	**Sum of squared deviations**	**Divided by $n = 2$**	**Divided by $n - 1 = 1$**
4, 4	4	0	0	0
4, 8	6	8	4	8
8, 4	6	8	4	8
8, 8	8	0	0	0
Mean over the four samples			**2**	**4**

Dividing by n gives 2 on average: half the true variance, exactly as predicted: $\frac{n-1}{n}\,\sigma^2 = \frac{1}{2} \cdot 4 = 2$. Dividing by $n - 1$ gives exactly 4 on average. With only two data points the error is huge; with 1,000 data points, dividing by 1,000 or by 999 makes very little difference.

Remark — Degrees of freedom

There is also a more intuitive way of seeing that -1. The n deviations from the mean are not free: their sum must be zero. If you know $n - 1$ of them, the last one is already determined. The independent pieces of information about spread are therefore $n - 1$, and statisticians say that the estimate has $n - 1$ **degrees of freedom**.

PISA: what is a point worth?

The PISA survey of the OECD, the Organisation for Economic Co-operation and Development, measures the skills of 15-year-olds in dozens of countries in reading, mathematics and science.

A PISA score, however, is not a natural quantity like a length in centimetres. Students answer questions of varying difficulty, and a statistical model turns their answers into a number; but the **unit of measurement** of that number has to be chosen, just as for temperatures someone had to decide where to put the 0 and the 100 of the Celsius scale. When the scale was built (for reading in 2000, for mathematics and science in later cycles), the OECD calibrated it so that, among students in OECD countries, the mean was about **500 points** and the standard deviation about **100 points** [6].

Those two values, then, are a convention fixed once and for all, not the result of each survey. In later cycles means and standard deviations are **measured** on that same scale, and they can change: in 2025 the OECD average in science is 482 points, no longer 500. The standard deviation, however, remains of the order of 100 points: students "typically" lie about 100 points away from the mean.

In the 2025 cycle, published on 8 September 2026, Italian students scored 483 points in science, against an OECD average of 482 [7, 8].

Headline trap *(made up)*

PISA: Italian students beat the OECD average in science

True, by one point. On a scale where students typically differ from each other by 100 points, one point is a hundredth of a standard deviation. To get an order of magnitude, the OECD estimates that a year of schooling is worth, very roughly, 25–30 points [6]: one point corresponds to just over a week of lessons.

Rankings turn any difference, however tiny, into an overtaking. The standard deviation is the yardstick for telling whether a difference is large or small. And that is not enough: we also have to ask whether it is *real*, or just an effect of the sample. That is the question the second part will start from.

4. Percentage points and charts with a truncated axis

Two ways of saying "up"

Back to jobs. According to Istat, the unemployment rate was 5.9% in September 2025 and is 6.3% in September 2026 [1].

Definition — Percentage points and percentage change

When a quantity is already a percentage, its change can be expressed in two ways.

- **Difference in percentage points** (pp): subtract the two percentages. From 5.9% to 6.3% is +0.4 points.
- **Percentage (or relative) change**: divide the difference by the starting value. From 5.9% to 6.3% the change is $0.4/5.9 \approx +6.8\%$.

With the same data, then, one can write "unemployment up 0.4 points" or "unemployment up 6.8%". Or again, counting people instead of the rate, "unemployed up 7.8%" [1]. All three numbers are correct. The first sounds like a small wobble, the other two like a jump. An honest article says which of the three it is using; a sensational headline picks the one that impresses most.

The chart that says what we want

The same trick can be played with a picture.

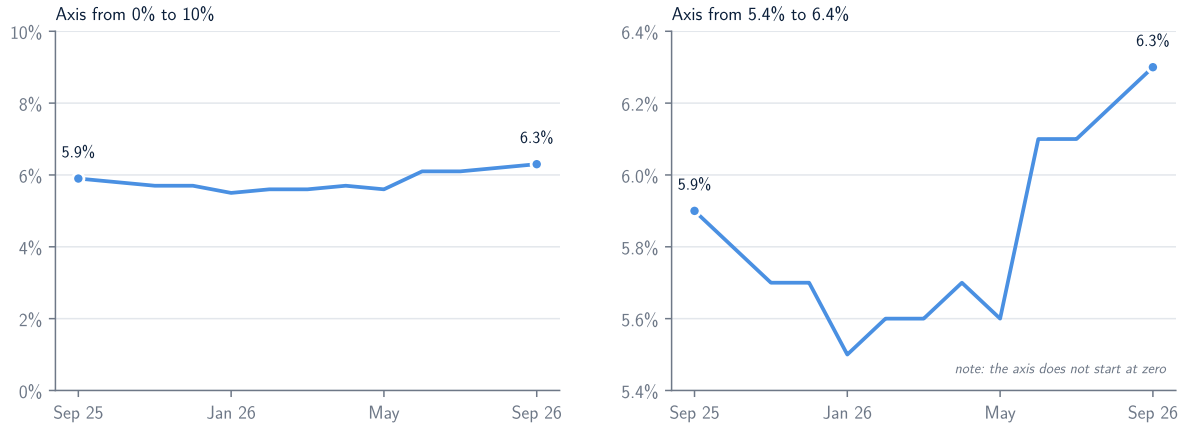

Figure 2. Monthly unemployment rate in Italy, from September 2025 to September 2026. The same thirteen numbers are drawn twice: on the left the vertical axis runs from 0% to 10%, on the right from 5.4% to 6.4%. Source: Istat, *Occupati e disoccupati – Settembre 2026* (Employment and unemployment – September 2026), Table 7.

On the left the line is almost flat; on the right it looks like a surge. Neither chart is false: the one on the right "zooms in" on part of the vertical axis. Zooming in is sometimes useful, to see details that would otherwise be lost; but if the axis does not start at zero this must be said clearly, and readers should always check where it starts.

Remark — Three questions to ask about a chart

Where does the vertical axis start? What period is shown, and why that one? Are the units points, percentages or absolute numbers?

Let's recap — The first four questions to ask about a number

1. **Mean or median?** If the distribution is skewed, as with incomes, spending and wealth, the mean tells a different story from that of the "typical" household.
2. **How spread out are the data?** A difference of a few points can be tiny compared with the standard deviation.
3. **Points or percentage? Rate or people?** The same change can become +0.4, +6.8% or +7.8%.
4. **Where does the axis start, and from which month?** Charts and periods can be chosen to make something small look big.

In the second part

So far we have treated Istat's numbers as if they were exact. They are not: almost all of them come from a sample, and every estimate from a sample carries a margin of error. In the second part we will see how that "range" is calculated and why a 23% rise in household spending may mean nothing at all; we will find out that a strong correlation between household spending and school results almost vanishes as soon as we look at the data more closely; we will tell the story of the largest poll in history, and the most wrong. And we will find out what became of Darrell Huff.

Read the second part →

Appendix — For those who want the maths

This appendix collects the proofs of the results used in section 3. High-school algebra is enough, plus, at a single point, a probability result that is proved at university. These are classic results, found in any good textbook: for example [9] for measures of location and spread (A.1 and A.2) and [10] for probability (A.3).

A.1 The sum of the deviations from the mean is zero

Property

For any set of data $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0.$$

Proof. Add up all the deviations and separate the terms:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = (x_1 + x_2 + \cdots + x_n) - n\,\bar{x}.$$

By the definition of the mean, $x_1 + x_2 + \cdots + x_n = n\,\bar{x}$. So the sum equals $n\,\bar{x} - n\,\bar{x} = 0$. ■

A.2 The mean is the value closest to the data

Notation — The summation sign

To write a sum of many terms concisely, we use the Greek letter Σ (capital sigma):

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

It is read "the sum for i from 1 to n of a_i". Two rules we will use: a factor that does not depend on i can be taken outside the sum, $\sum_{i=1}^{n} k\,a_i = k\sum_{i=1}^{n} a_i$; and adding the same number k n times gives $\sum_{i=1}^{n} k = n\,k$.

Property — The mean is the value closest to the data

For any number c,

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\,(\bar{x} - c)^2. \tag{1}$$

Proof. *Step 1.* Add and subtract $\bar{x}$ inside each bracket:

$$x_i - c = (x_i - \bar{x}) + (\bar{x} - c).$$

Step 2. Square, using the rule for the square of a binomial, $(A + B)^2 = A^2 + 2AB + B^2$, with $A = x_i - \bar{x}$ and $B = \bar{x} - c$:

$$(x_i - c)^2 = (x_i - \bar{x})^2 + 2\,(x_i - \bar{x})(\bar{x} - c) + (\bar{x} - c)^2.$$

Step 3. Sum over i from 1 to n, term by term:

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}2\,(x_i - \bar{x})(\bar{x} - c) + \sum_{i=1}^{n}(\bar{x} - c)^2.$$

Step 4. In the second term, the factor $2(\bar{x} - c)$ is the same for every i, so it comes out of the sum; what is left is the sum of the deviations from the mean, which is zero:

$$\sum_{i=1}^{n}2\,(x_i - \bar{x})(\bar{x} - c) = 2\,(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) = 2\,(\bar{x} - c)\cdot 0 = 0.$$

Step 5. In the third term, the same number $(\bar{x} - c)^2$ is added n times:

$$\sum_{i=1}^{n}(\bar{x} - c)^2 = n\,(\bar{x} - c)^2.$$

Putting steps 3, 4 and 5 together gives (1). ■

In (1), the first term on the right does not depend on c, while the second, $n(\bar{x} - c)^2$, is never negative and is zero only for $c = \bar{x}$. So the sum of the squared distances of the data from a number c is **smallest when c is the mean**.

A.3 Why we divide by $n - 1$

The data $x_1, \ldots, x_n$ are now a random sample drawn from a population with mean μ and variance σ^2. In (1), choose $c = \mu$:

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\,(\bar{x} - \mu)^2,$$

and therefore

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \mu)^2 - n\,(\bar{x} - \mu)^2. \tag{2}$$

The deviations from the sample mean are therefore, in total, systematically smaller than the deviations from the true mean: if we divided by n, the population variance would be **underestimated**. By how much?

To answer, we need to reason about the **mean over all possible samples**, which we write $\mathbb{E}[\,\cdot\,]$ (in probability it is called the *expected value*). We need three facts.

- **(a)** The mean of a sum is the sum of the means, and a constant factor comes out of the mean: $\mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$ and $\mathbb{E}[k\,A] = k\,\mathbb{E}[A]$.
- **(b)** Each data point x_i is drawn at random from the population, so its squared distance from μ is, on average, exactly the population variance: $\mathbb{E}[(x_i-\mu)^2] = \sigma^2$. This is the very definition of variance.
- **(c)** The sample mean $\bar{x}$ fluctuates around μ, and its variance is $\mathbb{E}[(\bar{x}-\mu)^2] = \sigma^2/n$. This is the only step we do not prove: it is done in university probability courses (see for example [10]). It is the same fact that gives the standard error $\sigma/\sqrt{n}$, which we will discuss in the second part.

Now take the mean over all samples of both sides of (2). By fact (a), we can do it term by term.

First term, using (a) and (b):

$$\mathbb{E}\Big[\sum_{i=1}^{n}(x_i-\mu)^2\Big] = \sum_{i=1}^{n}\mathbb{E}\big[(x_i-\mu)^2\big] = \sum_{i=1}^{n}\sigma^2 = n\,\sigma^2.$$

Second term, using (a) and (c):

$$\mathbb{E}\big[n\,(\bar{x}-\mu)^2\big] = n\,\mathbb{E}\big[(\bar{x}-\mu)^2\big] = n\cdot\frac{\sigma^2}{n} = \sigma^2.$$

Substituting into (2):

$$\mathbb{E}\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\Big] = n\,\sigma^2 - \sigma^2 = (n-1)\,\sigma^2.$$

Here is the result. If we divide the sum of the squared deviations by n, on average we get $\frac{n-1}{n}\,\sigma^2$, a little less than the true variance. If we divide it by $n-1$, on average we get exactly σ^2:

$$\mathbb{E}[s^2] = \frac{1}{n-1}\,\mathbb{E}\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\Big] = \frac{(n-1)\,\sigma^2}{n-1} = \sigma^2.$$

Sources

1. Istat, *Occupati e disoccupati (dati provvisori) – Settembre 2026* (Employment and unemployment, provisional data – September 2026), 29 October 2026 (Tables 4, 6, 7, A and B). Press release · PDF
2. D. Huff, *How to Lie with Statistics*, illustrated by I. Geis, W. W. Norton & Company, New York, 1954 (paperback edition: ISBN 978-0-393-31072-6). Wikipedia entry

3. D. Huff, *Mentire con le statistiche* (Italian edition of *How to Lie with Statistics*), illustrated by I. Geis, edited by G. Livraghi and R. Puglisi, Monti & Ambrosini, Pescara, 2007, «Diogene» series, ISBN 978-88-89479-09-4. Istat digital library

4. Istat, *Le spese per i consumi delle famiglie – Anno 2025* (Household consumption expenditure – Year 2025), Statistica Report, 8 October 2026 (methodological note, Tables A and B, glossary). istat.it

5. Bank of Italy, *Conti distributivi sulla ricchezza delle famiglie italiane: statistiche aggiornate al quarto trimestre del 2025* (Distributional wealth accounts of Italian households, updated to Q4 2025), 3 June 2026. bancaditalia.it

6. OECD, *How PISA results are reported: What is a PISA score?* oecd-ilibrary.org

7. OECD, *PISA 2025 Results (Volume I): Italy*, 8 September 2026. oecd.org

8. INVALSI, *PISA 2025. I risultati degli studenti italiani a colpo d'occhio* (PISA 2025: Italian students' results at a glance), September 2026. invalsi.it

9. D. Piccolo, *Statistica*, Il Mulino, Bologna (in Italian).

10. A. Papoulis, S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill.

Further reading

- J. M. Steele, "Darrell Huff and Fifty Years of *How to Lie with Statistics*", *Statistical Science*, 20(3), 2005. projecteuclid.org

About the author. Nicola Maglione is an electronic engineer (degree in Electronic Engineering, University of Naples, 2001), with over twenty years of experience in the semiconductor industry as a design and program manager, specialized in NAND Flash memory and high-speed interfaces, and holder of several patents in memory architectures. Alongside his professional work, he has always cultivated a passion for mathematics – particularly Analysis, Abstract Algebra, and Number Theory – and for physics, which he shares through keenonscience, his science-outreach page, and on the website keenonmaths.org.

© 2026 Nicola Maglione. Distributed under a Creative Commons Attribution – NonCommercial – NoDerivatives 4.0 International license (CC BY-NC-ND 4.0): sharing is permitted with attribution; commercial use and derivative works are not permitted.
creativecommons.org/licenses/by-nc-nd/4.0