

keenonmaths

Lying with statistics

seventy years on

Part 2 · Samples, ranges and correlations

Science in the news

written by **Nicola Maglione** — keenonmaths

@keenonmaths

November 2026

Contents

Where we left off 2

1. From the sample to the population: the range 2

2. Correlation is not causation 7

3. Lots of data are not enough: quality matters more than quantity 9

Let's recap 11

Epilogue: Huff and smoking 11

Sources 12

Where we left off

In the first part we saw how data are described: mean, median and mode, variance and standard deviation, percentage points and charts. And we saw how each of these tools, used with a touch of malice, can turn true numbers into misleading headlines. The inspiration, seventy years ago, came from a short book by the American journalist Darrell Huff, *How to Lie with Statistics* [1, 2].

There is, however, something we took for granted. Almost all the numbers we used, average household spending, the unemployment rate, PISA scores, do not come from counting the *whole* population, but from a **sample**: a few tens of thousands of households or students, chosen at random, from which the value for everyone is estimated.

> **Definition — Population and sample** *(from the first part)*
> The **population** is the set of all the elements we are interested in. A **sample** is the part of the population on which measurements are actually made. If it is chosen **at random**, following precise rules, it makes it possible to **estimate** the characteristics of the population.

An estimate, however, is never exact. In this second part we will see how far off it can be, how that error is measured, and what tricks arise when it is forgotten. Then we will tackle perhaps the most widespread trick of all, correlation mistaken for causation; and finally, an error that no sample, however large, can correct.

1. From the sample to the population: the range

Every estimate has a margin of error

Back to the sauce from the first part. Two spoonfuls taken from the same pan never taste exactly the same. In the same way, if Istat repeated its household spending survey with another 32,500 randomly drawn households, it would get a slightly different mean. This uncertainty, which arises from observing only part of the population, is called **sampling error**, and it can be measured.

Imagine we could repeat the survey many times, each time with a new random sample of the same size. Each repetition would give a slightly different estimate, and all these estimates would in turn form a distribution, with its own mean and its own spread. If the sample is well chosen, the estimates cluster around the true value; how spread out they are tells us how far we can trust a single estimate.

> **Definition — Standard error (or sampling error)**
> The **standard error** of an estimate is the **standard deviation** of the estimates we would obtain by repeating the survey many times, with different random samples of the same size. It measures how far, typically, a single estimate lies from the true value.
>
> Divided by the estimate, it becomes the **relative error**, expressed as a percentage.

In reality a survey is carried out only once. The standard error is then **estimated** from the

data of that same sample, using formulas derived from probability theory. The two simplest, valid for a simple random sample of n elements, are these.

Formulas — Standard error of a mean and of a percentage

- For a **mean**:

$$\text{standard error} \approx \frac{s}{\sqrt{n}},$$

  where s is the sample standard deviation (the one with $n-1$ in the denominator, seen in the first part).

- For a **percentage** p (written as a number between 0 and 1):

$$\text{standard error} \approx \sqrt{\frac{p\,(1-p)}{n}}.$$

The formulas tell us two important things. First: the error grows when the data are very spread out (s large). Second: the error shrinks as the sample grows, but slowly, like $1/\sqrt{n}$. To halve the error, doubling the sample is not enough: it has to be made **four times** larger.

An example: in a random sample of 1,000 people, 50% answer "yes" to a question. The standard error is $\sqrt{0.5 \times 0.5/1{,}000} \approx 0.016$, that is about 1.6 percentage points. This is why many polls with about a thousand respondents state a margin of error of about ±3 points: it is twice the standard error, as we will see in a moment.

Istat's surveys use more elaborate sampling designs (first municipalities are drawn, then households), and the formulas are more complicated; but the idea is the same, and Istat publishes the relative errors of its estimates directly.

Definition — Confidence interval (the "range")

A **95% confidence interval** is obtained, for estimates like these, by adding to and subtracting from the estimate about **twice** (1.96 times, to be precise) its sampling error. The method used to build it "captures" the true value of the population in 95 cases out of 100.

The calculation is not a textbook exercise: Istat publishes it itself. Average household spending in 2025 is estimated at 2,782 euros, with a relative error of 0.6%, that is about 17 euros. The true value, the one we would get by interviewing every household, therefore lies, with 95% confidence, **between about 2,749 and 2,816 euros** [3].

The case of Umbria and the case of Marche

The smaller the sample, the wider the range. At the level of a single region far fewer households are interviewed, and some estimates become uncertain.

Take Istat's estimates of average monthly spending in Umbria, a small region in central Italy: 2,640 euros in 2024, 3,257 euros in 2025 [3]. Let's do the arithmetic:

$$\frac{3{,}257}{2{,}640} \approx 1.234,$$

that is, a rise of about **23%** in a single year. And at 3,257 euros, Umbria comes out as Italy's second-highest-spending region, behind only Trentino-Alto Adige (3,444 euros) [3]. The headline writes itself.

Headline trap *(made up)*

Umbria's spending boom: up 23% in a year

Now Italy's second-highest-spending region

All the numbers are in the Istat report, and the arithmetic is right. Let's do the same calculation for Marche, the neighbouring region: from 2,463 to 3,084 euros, a rise of about **25%** [3]. Two almost identical increases. But let's look at the ranges.

Region	Year	Estimated average spending	Relative error	95% range
Umbria	2024	€2,640	2.0%	about €2,540 – 2,740
Umbria	2025	€3,257	**12.1%**	about €2,490 – 4,030
Marche	2024	€2,463	3.0%	about €2,320 – 2,610
Marche	2025	€3,084	2.9%	about €2,910 – 3,260

The ranges are obtained as shown above: estimate $\pm 1.96\times$ relative error $\times$ estimate. For Umbria in 2025, for example, $1.96 \times 0.121 \times 3{,}257 \approx 773$ euros, to be subtracted from and added to the estimate.

Why Umbria, and why 2025? That 12.1% error stands out, and might suggest a wrong figure. It is not: Istat itself publishes it and flags it in a note [3]. To understand where it comes from, recall the formula for the standard error of a mean, $s/\sqrt{n}$. The error is large when few households are interviewed (n small) or when their spending varies a lot from one to another (s large).

- **Few households.** In small regions the sample contains few households, and indeed all of them have errors higher than the national figure (0.6%): in 2025, 5.4% in Valle d'Aosta, 4.9% in Basilicata, 3.2% in Molise, against 1.8% in Lombardy [3].
- **Very spread-out spending.** This alone is not enough to explain the 12.1%, which is more than twice that of any other region (the second highest is Valle d'Aosta's 5.4%), and which for Umbria in 2024 was only 2.0%. The most likely explanation, although the report does not go into detail, is that the 2025 Umbrian sample happened to include a few households with exceptionally high spending, for example on home renovation: an item so irregular that even at national level its relative error is 11.4% [3]. A few households very "far" from the others raise the mean and, even more, the standard deviation: it is the "millionaire in the classroom" of the first part.

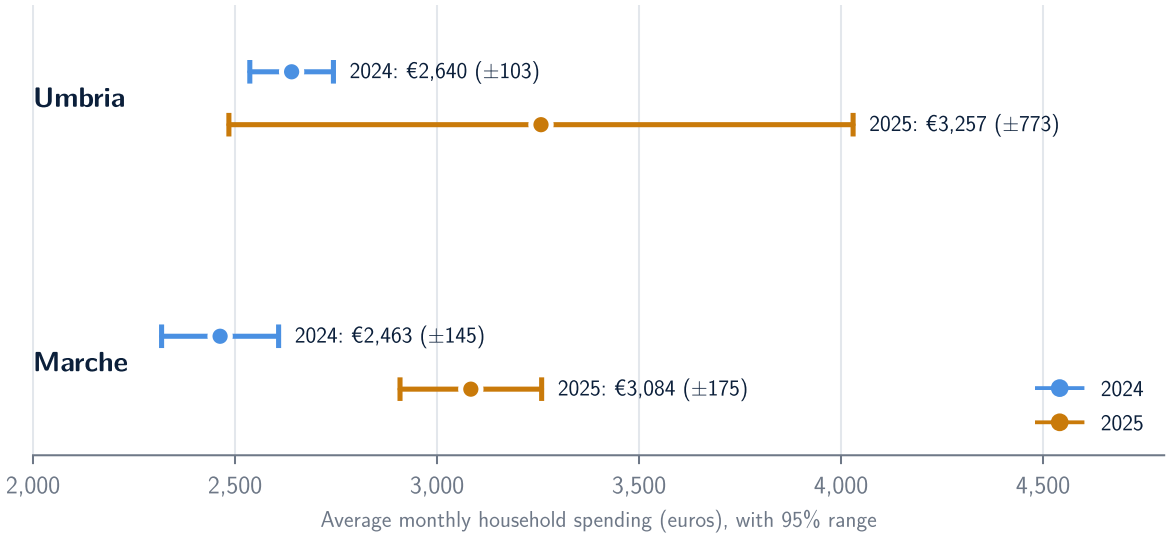

Figure 1. Average monthly household spending in Umbria and Marche, 2024 and 2025, with 95% ranges. The ranges are computed from the estimates and relative errors published by Istat. Source: Istat, *Le spese per i consumi delle famiglie – Anno 2025* (Household consumption expenditure – Year 2025), Table B.

For Marche, the 2024 and 2025 ranges are well apart: the increase is **statistically significant**, and Istat marks it with an asterisk. For Umbria the 2025 range is very wide, overlaps the 2024 range and even contains the 2024 estimate: the increase might be real, but it might also be an effect of the sample. Istat itself warns that the Umbrian estimate is unreliable [3].

Watch out — Comparing ranges is a rule of thumb

Checking whether two ranges overlap is a good first test, but it is not the exact calculation: to decide whether a change is significant, statisticians build the range of the *difference* between the two estimates. Two ranges that barely touch can still correspond to a significant difference; two ranges well apart, like those for Marche, always indicate a significant difference.

Definition — Statistically significant difference

A difference between two estimates is **statistically significant** when it is too large to be explained, with reasonable confidence, by sampling error alone. "Statistically significant" does not mean "important": it means "probably not due to chance". A difference can be significant and tiny, or large but not significant, if the estimates are too uncertain.

The headline about Umbria, then, rests on precisely the number that Istat itself describes as unreliable. The one about Marche, which nobody would write because the region does not make the podium, would instead describe a well-documented increase.

The range of the unemployment rate

The same applies to the monthly labour market figures, which come from a survey of about 18,000 households a month [4]. In the methodological note to its September 2026 release, Istat reports that the 95% range for the unemployment rate is about **±0.3 points**, and that for the number of employed people about **±140,000** [4].

Month-on-month changes are often 0.1 or 0.2 points: they fall within the range. A headline such as "unemployment down 0.1 points" or "20,000 more people in work in a month" reports a change that the sample cannot distinguish from zero. This is why it is better to look at comparisons over longer periods, such as quarterly averages or year-on-year comparisons.

There is another consequence. With data that go up and down slightly every month, anyone who wants to prove a point can always pick the starting month that suits them. A rise from May to September and stability from September to September can come out of the same series of numbers.

Remark — Data change too

Monthly labour market data are *provisional*: Istat revises them as new information arrives. A headline written on the day of release may refer to a number that is different a few months later.

Counting who is there

One last example from the labour market. In a year, the number of employed people aged 50 and over rose by 339,000 [4]. Here is the easy headline:

Headline trap *(made up)*

Jobs boom for the over-50s

339,000 more people in work in a year

But in Italy there are more and more people aged 50 to 64, because the population is ageing: even if each of them had the same chance of working as a year earlier, the number of employed people in that age group would grow anyway. Istat does the calculation: employment among 50-to-64-year-olds grows by 2.3%, but once the effect of the larger population is removed, growth falls to 1.7% [4]. Part of the "boom" is just demography.

2. Correlation is not causation

A headline built on true data

Let's put two official surveys together. For each region we take:

- **average monthly household spending** in 2025 (Istat [3]);
- the **percentage of students in the last year of upper secondary school who reach the expected standard** in the 2026 INVALSI mathematics tests (INVALSI [5]). INVALSI is Italy's national institute for the evaluation of the education and training system, which every year gives standardised tests to students across the country.

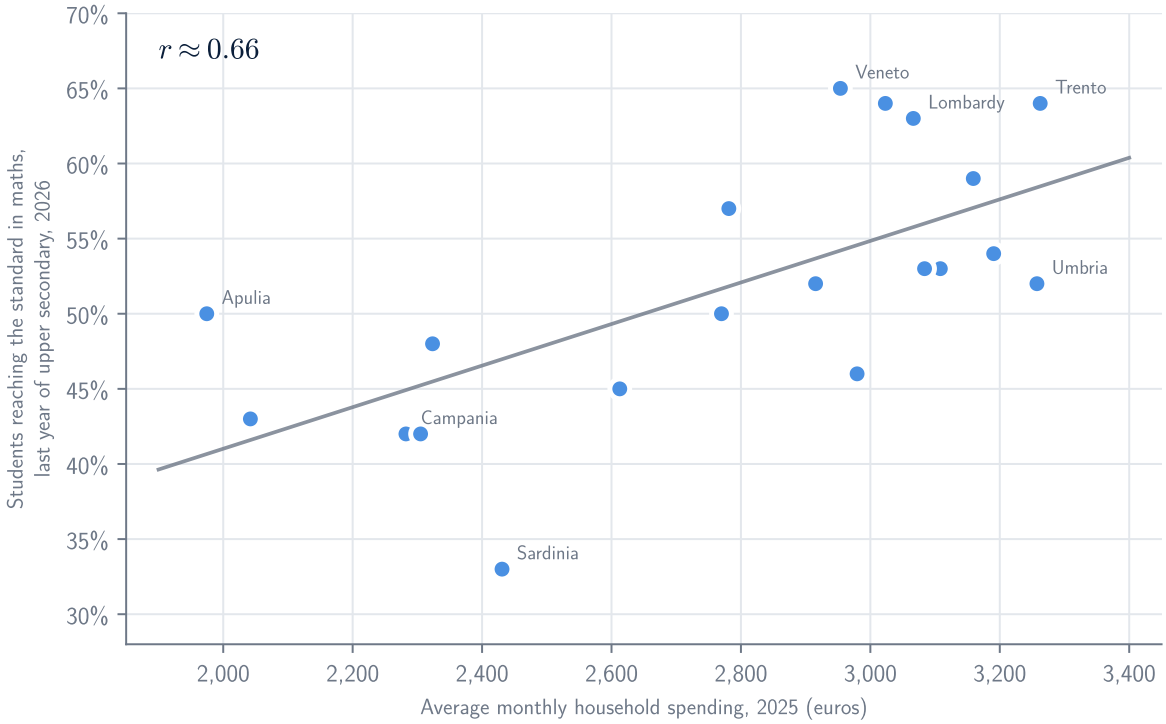

Figure 2. Each point is a region (Trento separate from Bolzano). Horizontally, average monthly household spending in 2025; vertically, the share of students in the last year of upper secondary school reaching the expected standard in mathematics in 2026. The Province of Bolzano is excluded, because its INVALSI data refer separately to Italian-language and German-language schools, while spending refers to the whole province. Sources: Istat, *Le spese per i consumi delle famiglie – Anno 2025*, Table B; INVALSI, *Rapporto nazionale 2026* (National report 2026), Figure 6.2.2.3. Calculations by keenonmaths.

The points rise from left to right: where households spend more, more students reach the expected standard. The link can be measured with a number.

Definition — Correlation coefficient

The (Pearson) **correlation coefficient** r measures how much two quantities tend to rise or fall together *linearly*. It always lies between -1 and $+1$:

- r close to $+1$: when one rises, the other tends to rise;
- r close to -1: when one rises, the other tends to fall;
- r close to 0: no linear link.

For the regions in Figure 2 we get $r \approx 0.66$ in mathematics and $r \approx 0.71$ in Italian: a decidedly strong correlation. The headline trap follows naturally.

Headline trap *(made up)*

Where families spend more, students do better at maths

Istat and INVALSI data confirm it

And the next step is even more tempting: "to improve schools, families must be made to spend more".

The hidden variable

Now let's colour the points: in blue the regions of the Centre-North, in orange those of the South (the Mezzogiorno).

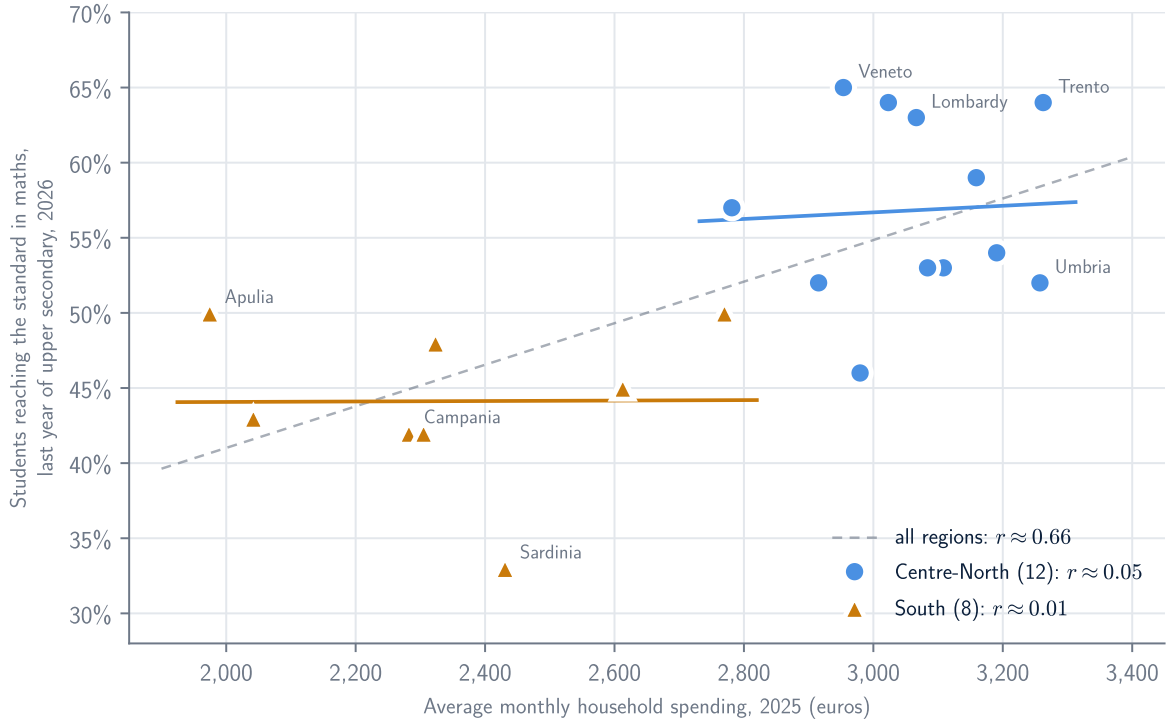

Figure 3. The same data as in Figure 2, with the regions of the Centre-North (12) in blue and those of the South (8) in orange. The lines show the trend within each group. Sources and calculations as in Figure 2.

The regions form two separate "clouds". The Centre-North spends more and has higher results; the South spends less and has lower results. But **within each group** the correlation almost disappears: $r \approx 0.05$ among the regions of the Centre-North and $r \approx 0.01$ among those of the South. Between Lombardy and Emilia-Romagna, or between Apulia and Calabria, spending more does not go with better results.

Almost all of the correlation, then, comes from the gap between two large areas of the country, which differ in a great many things at once: incomes, jobs, services, the types of school attended, history. Household spending and school results both depend on this set of conditions; that does not mean that one depends on the other.

Definition — Hidden (or confounding) variable

A **hidden variable** is a third quantity that influences both of the quantities observed and creates a correlation between them, even when neither causes the other.

Indeed, when INVALSI tries to estimate what really affects results, with all other conditions equal, it finds very specific factors: the type of school (students at a *liceo scientifico*, the science-oriented academic high school, score on average more than 22 points higher in mathematics than students at a technical institute), a regular school career (being held back a year costs about 10 points) and the socio-economic background, especially that of the school [5]. A region's average spending sums up all these differences in a single number, and very crudely.

Remark — Few points, a lot of caution

There are few regions: 12 in one group and 8 in the other. With so few data, the correlation coefficient itself is uncertain, and should be read as an indication, not as a precise measurement.

Remark — Growing together is not enough

Two quantities that both increase over time always turn out to be correlated, whatever they are: the number of smartphones sold and the average age of the population, for example. A correlation between two time series, on its own, proves almost nothing.

3. Lots of data are not enough: quality matters more than quantity

Two kinds of error

So far we have talked about only one kind of error, the one due to chance: the sample is part of the population, and another part would give a slightly different result. This error has a reassuring property: **it shrinks as the amount of data grows**. But there is a second kind of error, much more insidious, which does not shrink at all as the data grow.

Definition — Random error and systematic error

Random error moves measurements sometimes one way, sometimes the other, with no preference: when many measurements are averaged, it tends to cancel out.

Systematic error (or **bias**) always moves measurements in the same direction. Averaging many measurements does not cancel it: it all remains.

A domestic example. A set of bathroom scales always reads 2 kg more than the true weight. Weigh yourself once and you are 2 kg off; weigh yourself a thousand times and take the average, and you get an extremely precise number, stable to the gram, and 2 kg wrong. A thousand measurements make the result **precise**, not **accurate**. The same goes for statistical data: if the way they are collected is biased, multiplying them only produces a more convincing error.

The largest poll in history, and the most wrong

In 1936, in the United States, the magazine *Literary Digest* set out to predict the result of the presidential election between Franklin D. Roosevelt and Alf Landon. It mailed about **10 million** postcards and got almost **2.4 million** back: a gigantic sample. The result gave Landon 54% and Roosevelt 41%. At the election, Roosevelt won about **61%** of the vote and carried 46 states out of 48 [6, 7].

Other pollsters, including George Gallup, with much smaller but more carefully chosen samples, correctly predicted Roosevelt's victory [6].

What went wrong? The addresses came largely from telephone directories and car registration lists, and for decades this was blamed: in 1936 telephones and cars were more common among the better-off. More recent studies, however, suggest that another factor weighed most: **who chose to reply**. Fewer than one in four of the ten million postcards came back, and Landon's voters, more motivated to make themselves heard, replied much more often [6]. Either way, the error is systematic: 2.4 million replies were not enough to correct it.

Definition — Self-selection

Self-selection occurs when people themselves decide whether to enter the sample. Those who choose to reply are usually different from those who do not: more interested, angrier, more satisfied. The sample then no longer resembles the population, however large it is.

Today's examples are everywhere: online reviews (written mostly by people who are delighted or furious), social media polls (answered by the followers of that page), viewer votes on television programmes. They can collect hundreds of thousands of answers and still be unreliable.

Headline trap *(made up)*

Record poll: 70% of students reject the new reform

100,000 replied

A hundred thousand replies is a lot. But who gave them? If the poll was on a social media page followed mostly by opponents of the reform, the number of replies says nothing about what all students think.

Why INVALSI uses a sample even though it has everyone's data

Now we can understand a choice that at first sight seems strange. The INVALSI tests are taken by **all** students in the classes concerned: in 2026, about 11,400 schools [5]. Yet, alongside the testing of everyone, INVALSI randomly selects a **sample of classes**, usually two in each selected school. In those classes an external observer checks that the test is carried out according to the protocol, and the report explains that the sample serves "to guarantee the reliability of the data underlying any comparison" [5]. In the second year of primary school, for example, in 2026 the students in sample classes numbered about 15,700 out of more than 380,000 participants [5].

Why trust 15,700 students more than 380,000? Because in a class without an observer systematic errors can creep in: help during the test, time limits not respected, conditions that differ from school to school. These errors all push in the same direction, and no amount of data can make them disappear. The supervised sample, on the other hand, is small but clean: it has a somewhat larger random error, which can however be calculated, and a much smaller systematic error.

Remark — Which error to fear

Random error is like fog: it makes the view less sharp, but you know by how much, and the range tells you. Systematic error is like a road sign turned the wrong way: you can see it perfectly, and that is exactly why it deceives you. No range will flag it.

Let's recap — Three more questions to ask about a number

(to add to the four from the first part)

5. **How wide is the range?** Every estimate from a sample has a margin of error; changes that fall within it are not news.
6. **Is there a third variable?** A correlation, however strong, does not tell you what causes what.
7. **Who chose the data?** Lots of badly collected data are still bad data: a small but random sample beats a huge sample that selected itself.

Epilogue: Huff and smoking

There is a little-known sequel to Huff's story. In 1964 the report of the *Surgeon General*, the highest public health authority in the United States, concluded that cigarette smoking causes lung cancer. The tobacco industry responded by looking for experts who would cast doubt on those conclusions, and among them it hired Huff himself. In 1965 Huff testified before a US Senate committee against the report's statistics; then, paid by the industry, he wrote a sequel to his most famous book: *How to Lie with Smoking Statistics* [8].

The book was never published. The manuscript resurfaced decades later, among the tobacco industry's internal documents made public as a result of the lawsuits of the 1990s [8]. And its arguments are exactly the ones we have met in these two parts: samples that are too small or biased, and above all "correlation is not causation". To amuse the reader, Huff recalled that in some areas the number of storks is correlated with the number of births: so why believe that smoking causes cancer just because smoking and cancer go together? [8]

The reasoning is correct in the abstract and wrong in the concrete case. The link between

smoking and cancer did not rest on a single correlation, but on a great many different observations all pointing in the same direction: the more people smoke, the higher the risk; quitting reduces it; the effects are also seen in the laboratory. No hidden variable capable of explaining all of this together has ever been found.

The moral is uncomfortable but important: the tools of statistics do not belong to one side only. Doubt, too, can be used to deceive: you can take apart an exaggerated headline, but you can also manufacture artificial doubt about a solid result. This is why knowing the tricks is not enough; they must be applied honestly, asking each time whether an objection is a legitimate question or a way of refusing to accept an answer.

Statistics is not there to prove that "you can make numbers say anything". Quite the opposite: it is there to understand what a number really says, and to recognise those who make it say more, or less. And the first step is often the simplest: whenever possible, go back to the source. If a headline quotes a figure, the Istat release, the INVALSI report or the study it comes from are almost always a few clicks away, and they say how that number was measured, on what sample and with what margin of error. That is often where an exaggerated headline deflates.

Sources

1. 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
2. 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
3. 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
4. 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
5. INVALSI, *Rapporto nazionale Prove INVALSI 2026* (INVALSI national report 2026), 16 July 2026 (section 1.2; Figures 6.1.2.3 and 6.2.2.3; sections 6.1.3 and 6.2.3). invalsi.it
6. D. Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?", *Social Science History*, 36(1), 2012. doi.org/10.1215/01455532-1461650
7. The American Presidency Project, University of California, Santa Barbara, *1936 Presidential Election*. presidency.ucsb.edu
8. A. Reinhart, "Huff and puff", *Significance*, 11(4), 2014, pp. 28–33. doi.org/10.1111/j.1740-9713.2014.00765.x · freely readable version

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