How sure are we? A plain-English guide

Seven ideas that explain what our numbers can — and can't — claim. No statistics background needed.

Bluefruit Software — internal reference

Every number needs three friends

A measurement on its own means very little. Always ask for its three companions: **how many times we ran it**, **how much it wobbled** (the ± bit), and **the conditions it was measured under**.

Measuring the wobble: repeat the measurement a few times and take the standard deviation of the repeats — one spreadsheet formula, **=STDEV.S(your runs)**. That number is the ±.

"Resistance was 3.2 kΩ" tells you almost nothing. "3.2 ± 0.3 kΩ across five runs, same rig, same day" you can rely on.

Consistent isn't the same as correct

If repeats agree with each other, the measurement is **consistent** — but it could still be consistently *wrong*. A stopped clock is very consistent.

EIS example: ten sweeps without touching the rig will agree beautifully — that only proves the electronics are quiet. Remount the probe between runs and the spread jumps. **Repeat the step your claim depends on** — remount, re-prep — or your ± flatters you.

And only a trusted reference reveals a built-in offset — **more runs never fix bias**. Wobble averages away; being systematically off doesn't.

Three runs is a hint, not proof

With only a handful of runs, the honest uncertainty is far wider than people expect. Three repeats give an **indication**, not a claim.

Honest range = average ± t × wobble ÷ √n. The catch is t: at n=3 it's 4.3; at n=5, 2.8; by n=10 it settles near 2. Small n makes the range balloon.

Same story for percentages: 19/20 right sounds like "95% accurate", but with that few tests the truth could be anywhere from **75% to 99%**.

Is the difference real, or just noise?

A 60-second check: average each condition; take the gap between the two averages; measure each condition's wobble (=STDEV.S); then divide the gap by the typical wobble.

That ratio is "how many noises wide" the difference is. **Around 1** — could easily be luck. **Around 2** — promising, worth the proper sums. **5 or more** — obvious to anyone.

Decide how many runs before you start

The subtler the effect you're looking for, the more runs you need — and the answer should be agreed **before** the experiment, not after.

Huge effect	Modest effect	Subtle effect
~4 runs each	~16 runs each	~64 runs each

Too few runs and "inconclusive" just means **we didn't look hard enough** — not "there's no effect".

You can't measure better than your ruler

Every claim is checked against a reference method, and it can never be more precise than that reference — **however good the instrument is**.

If two people counting cells under a microscope disagree by ±10%, then no viability claim built on that counting can be tighter than ±10%.

Total error comes from several sources — fixture, sample prep, drift, electronics. Find the biggest one and fix *that*; ignore the rest.

Rules of thumb

Under 5 runs? Call it a hint, not a result.

Halving the uncertainty takes four times the runs.

Gap twice the scatter — promising. Five times — obvious to anyone.

Quoting an accuracy %? Expect to need 70+ blind tests first.

Error bars apart — probably real. Touching — ask for the sums.

Decimal places: quote one fewer than you're tempted to.

Found it by hunting everywhere? That's a new question, not an answer.

A claim is only as good as the reference it was checked against.

Play it straight

Say how many runs sit behind every figure. Don't quote more decimal places than the uncertainty justifies. Don't test lots of things and report only the best one — decide what you're looking for **before** collecting the data. And judge results **blind** — if you knew the answer while measuring, the claim only counts as a demo.