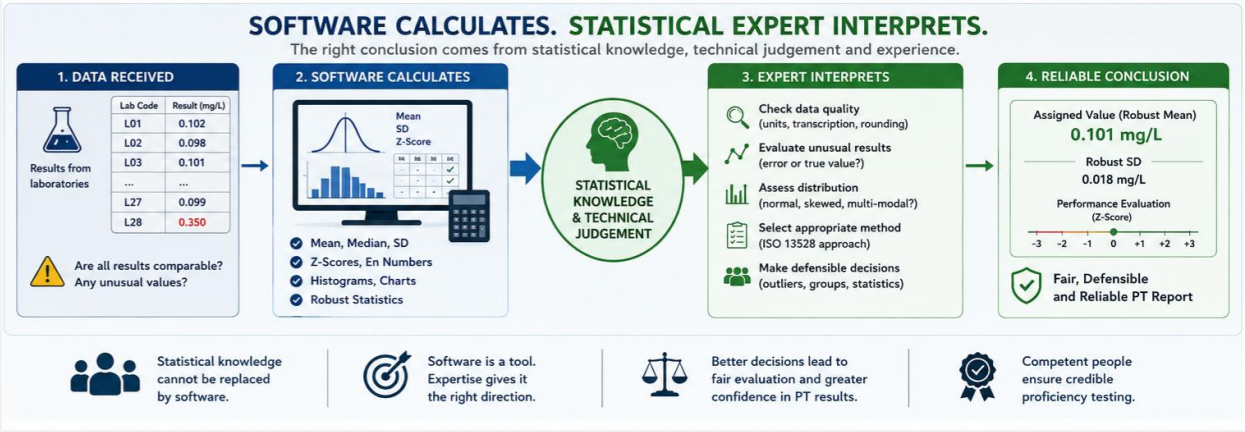
Statistical Software *Versus* Human Statistical Expertise *in* Proficiency Testing

Statistical software is a tool; statistical knowledge is the competence that determines whether the right tool, the right method, and the right interpretation are being used.



Why Statistical Knowledge Matters More Than Statistical Software in Proficiency Testing?

A statistical software package can calculate a number in seconds. But it cannot decide whether that number makes scientific sense.

In proficiency testing (PT), statistical analysis is not simply a matter of entering participant results into a software program and generating a mean, standard deviation, z-score or histogram. The quality of a PT statistical evaluation depends first on the **statistical understanding and technical judgement of the personnel performing the analysis**.

This distinction is particularly important because PT results are used to evaluate the performance of laboratories. A statistical decision that is technically inappropriate can therefore affect not only a PT report, but also the interpretation of a laboratory's analytical competence.

ISO 13528:2022 provides detailed statistical methods for PT providers to design proficiency testing schemes and analyse the resulting data. It also addresses the interpretation of PT data.

Software can calculate - but it cannot think statistically

Modern statistical software can perform sophisticated calculations very quickly. It can calculate averages, medians, standard deviations, robust statistics, confidence intervals, regression parameters, distributions and graphical outputs.

But the software normally does not understand the **technical story behind the data**.

For example, suppose a PT provider receives 25 participant results:

10.2, 10.4, 10.3, 10.5, 10.1, 10.4, 10.2, 10.3, 10.5, 10.4, 10.3, 10.2, 10.4, 10.3, 10.5, 10.2, 10.4, 10.3, 10.5, 10.4, 10.3, 10.2, 10.4, 10.3, 15.8

A software package can immediately calculate the mean and standard deviation.

But the important question is:

What is 15.8?

Is it:

- a genuine laboratory result?
- a transcription error?
- a unit conversion problem?
- a different analytical method?
- a result affected by sample preparation?
- a technically valid but unusual result?
- or an observation that requires investigation before statistical treatment?

The software can identify that the value is unusual. **It cannot establish the technical reason for that unusual result.**

That requires statistical knowledge combined with PT and measurement expertise.

The real challenge begins before the calculation

One of the most important misunderstandings in PT statistics is that statistical analysis starts when the data are entered into Excel or another software package.

In reality, statistical thinking begins much earlier.

A competent PT statistician should ask:

What type of data have we received?

Are the results comparable?

Are participants using different methods?

Is there evidence of more than one population or method group?

Is the distribution approximately symmetrical, skewed or multimodal?

Are there sufficient results for the proposed statistical approach?

Are some results reported below a detection or quantification limit?

Are the units consistent?

Have rounding or reporting conventions influenced the data?

These questions cannot be answered simply by clicking a statistical function.

This is why statistical knowledge of PT personnel **cannot be equated with knowledge of statistical software**.

Ground reality: PT data are rarely "perfect textbook data"

Real PT datasets are often considerably more complicated than examples found in statistical textbooks.

A PT provider may receive results from laboratories using different instruments, analytical principles, extraction procedures, calibration systems, reporting units and levels of measurement capability.

For example, a PT scheme may produce a distribution containing two or more clusters. A software program may calculate one overall mean and standard deviation without difficulty.

But statistically and technically, the important question may be:

Should these results actually be treated as one population?

If the groups arise from genuinely different measurement principles, combining them without appropriate consideration may produce a misleading estimate of the consensus value or variability.

Similarly, a highly dispersed dataset does not automatically mean that the PT item is unsuitable. The dispersion may reflect genuine differences between analytical methods, laboratories or measurement processes.

The person evaluating the data therefore needs to understand **why the data look the way they do**, not merely what the software reports.

Outliers are another practical example

Outlier treatment is perhaps one of the clearest examples of why statistical knowledge is essential.

A common misconception is:

“The software identified an outlier, so remove it.”

That is not a statistically sound principle.

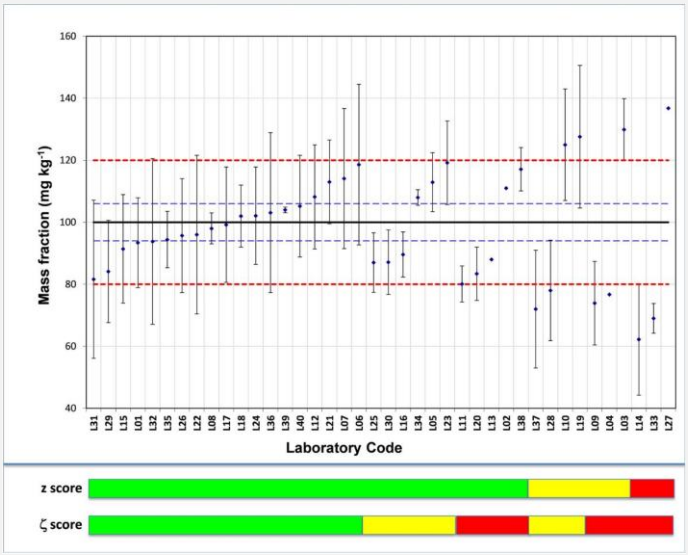
An extreme observation is not automatically an invalid observation.

Removing an observation can change the assigned value, standard deviation, uncertainty and ultimately the performance evaluation of participating laboratories.

Therefore, an experienced PT statistical evaluator should understand:

- how unusual observations are identified;
- the difference between an extreme result and an erroneous result;
- when robust statistical approaches are appropriate;
- how outlier treatment affects the final statistical estimates;
- whether the underlying PT statistical model remains appropriate.

ISO 13528 provides statistical approaches specifically for PT data, rather than simply requiring providers to apply generic statistical functions.



A spreadsheet error can become a statistical error

There is also a very practical reason why human statistical competence matters.

Imagine that a PT provider receives a participant result reported as:

0.25 mg/kg

while another laboratory reports:

250 µg/kg

Numerically, these represent the same concentration.

However, if the data are entered or converted incorrectly, the statistical software will faithfully analyse the wrong value.

The software has done exactly what it was instructed to do.

The problem is not the calculation.

The problem is the statistical and technical judgement before the calculation.

The same issue can occur with decimal separators, duplicated results, incorrect sample identification, wrong units, transcription errors, method-group classification and incorrectly entered uncertainty values.

Statistical software generally has no understanding of the laboratory's test method or the physical meaning of the result.

Software should support statistical competence - not replace it

This does not mean that PT providers should avoid statistical software.

Quite the opposite. Good statistical software can provide enormous benefits:

- rapid and reproducible calculations;
- reduced manual arithmetic errors;
- graphical examination of datasets;
- automated statistical procedures;
- traceable calculations;
- efficient handling of large datasets;
- repeatability of analysis;
- and easier generation of statistical reports.

The correct approach, therefore, is not:

Human vs Software

but:

Statistical competence + Appropriate software + Technical judgement

Software should make a competent statistical professional **faster and more reliable**. It should not be expected to make an untrained person statistically competent.

The PT provider's responsibility

For a PT provider, statistical analysis is not simply a back-office calculation.

The statistical evaluation influences:

participant performance → assigned value → performance criteria → z-score / other performance statistic → interpretation → PT report

An error at the statistical stage can therefore propagate through the entire PT process.

This is why competent PT personnel should understand the principles behind the statistical methods they apply. They should be capable of explaining not only **"what result did the software produce?"**, but also:

“Why was this statistical method selected, what assumptions or considerations apply, and why is the resulting conclusion technically justified?”

That is the real difference between **using statistics** and **understanding statistics**.

The final perspective

A PT provider does not demonstrate statistical competence merely by possessing an advanced statistical software package.

- A computer can calculate a mean.
- A computer can calculate a robust estimate.
- A computer can generate a histogram.
- A computer can calculate a z-score.

But **the responsibility for deciding whether those calculations are appropriate remains with competent personnel**.

For this reason, statistical knowledge should be regarded as a fundamental competence of PT personnel, while statistical software should be regarded as an important supporting tool.

- **The software performs the calculation.**
- **The statistician decides what should be calculated, why it should be calculated, whether the result is meaningful, and how it should be interpreted.**

In proficiency testing, that distinction is not merely academic.

It is part of ensuring that statistical conclusions are scientifically defensible, technically meaningful and fair to participating laboratories.

In one sentence:

A powerful statistical software package in the hands of someone without statistical understanding can produce a very precise answer to the wrong question.