

PHYSICS FOR ENGINEERING STUDENTS

Confidence Levels in Physical Measurements

Probability intervals, confidence intervals and Student's t distribution

Lecture 9 — Book-style student notes

University of Szeged · Institute of Physics

Confidence Levels in Physical Measurements

An uncertainty interval is meaningful only when its statistical basis is stated. This lecture connects repeated measurements to sampling distributions, probability intervals, confidence intervals, Student's t distribution and expanded uncertainty. The emphasis is on the correct interpretation of confidence and on choosing a coverage factor that reflects the available sample size.

Learning objectives

- Distinguish a parent population, a sample and a sampling distribution.
- Separate probability intervals from confidence intervals.
- Use z when σ is known and Student's t when σ is estimated.
- Relate confidence level, degrees of freedom and coverage factor.
- Report expanded uncertainty and evaluate unusual observations responsibly.

Central principle

Uncertainty gives the scale of an interval; the reference distribution and coverage factor determine its confidence. When σ is estimated from a small sample, Student's t distribution provides the finite-sample correction.

Further interpretation

A confidence level describes the long-run success rate of an interval procedure under its assumptions. It is not a guarantee for one experiment and not a probability assigned to a fixed true mean after the interval has been observed.

1. Parent populations and finite samples

Statistical inference begins by distinguishing the parent population from the finite sample actually observed. The parent population represents all results that could be produced under defined measurement conditions. Its true mean μ and true standard deviation σ are fixed parameters, but they are normally unknown. An experiment provides only N observations drawn from that population, so the sample contains incomplete information about both parameters.

$$X \sim N(\mu, \sigma^2)$$

The sample mean $\bar{x}$ is a point estimate of μ and the sample standard deviation s estimates σ . A different repetition of the entire experiment would produce a different sample and therefore different values of $\bar{x}$ and s . This sampling variability is the foundation of interval estimation. The result of one experiment should not be treated as if it described the complete population without uncertainty.

$$\bar{x} = (1/N)\sum_i x_i$$

Further interpretation

The population is defined by the measurement procedure and conditions, not by an abstract unlimited set of numbers. Changing the instrument, operator, environment or specimen population can change the parent distribution. Consequently, any confidence statement applies only to the population represented by the sampling design. Repeated readings under unchanged conditions may characterize repeatability while saying little about broader reproducibility.

Sampling language should be tied to an operational definition. If a length is measured six times without repositioning the specimen, the sample may represent electronic repeatability but not mounting variation. If six different specimens are used, it represents a population of objects instead. These two experiments can produce the same numerical mean and standard deviation while supporting different conclusions. State what was sampled, how observations were selected and what conditions were held fixed. This is essential because confidence intervals inherit their meaning from the repeated-sampling process that the actual experiment is intended to represent.

A representative sample is more important than a large convenient sample. Repeating a biased selection many times can yield a very narrow interval around the wrong population mean. Randomization, balanced selection and coverage of relevant operating conditions protect the interpretation of μ . Statistical formulas quantify sampling variation but cannot diagnose every selection mechanism from the observed numbers alone.

In practice, write one sentence defining the population represented by the observations. If that sentence is ambiguous, the target mean μ is ambiguous too, and no confidence calculation can repair the missing definition.

2. Sampling distribution of the mean

Imagine repeating the same experiment many times. Each experiment collects N measurements and produces one sample mean. The collection of those possible means forms the sampling distribution of $\bar{x}$. If the parent population is normal, the sampling distribution is exactly normal with mean μ and standard deviation $\sigma/\sqrt{N}$. More generally, the central limit theorem often makes the distribution approximately normal for sufficiently large N when observations are independent and have finite variance.

$$E[\bar{x}] = \mu$$

The standard deviation of the sampling distribution is called the standard error of the mean. It decreases as $1/\sqrt{N}$, so four times as many independent observations are required to halve it. This law describes random sampling uncertainty. It does not reduce systematic bias, common drift, or a modeling error that is repeated in every observation.

$$\sigma_{\bar{x}} = \sigma/\sqrt{N}$$

Further interpretation

The square-root law assumes that the N observations contribute independent information. Autocorrelation in a time series, spatial clustering, or repeated processing of the same raw signal reduces the effective sample size. Before using $\sqrt{N}$, inspect the acquisition sequence and sampling design. More data improve precision only when they add genuinely new information from a stable process.

The sampling distribution is conceptual; it usually cannot be observed directly because only one experiment is performed. Simulation can make it visible by drawing many samples from a known model, but the mathematical result is what supports inference in real work. The standard error describes the variation of sample means, not the spread of future individual readings. A prediction interval for a new observation is wider because it includes both uncertainty in the estimated mean and natural scatter of individuals. Confusing these intervals can lead to unrealistically tight expectations for future measurements.

The standard error can be estimated empirically by resampling methods when analytic assumptions are complicated, but the resampling scheme must mirror the dependence structure. Ordinary bootstrap resampling of individual points is inappropriate for strongly correlated time series. Blocks, clusters or complete experimental units may need to be resampled so that the simulated sampling distribution reflects how new data would actually be obtained.

In practice, explain why s describes individual scatter while $s/\sqrt{N}$ describes variation of sample means. Being able to state this distinction in words is as important as remembering the formula.

3. Standardization and the z statistic

When σ is known, the deviation of the sample mean from μ can be expressed in standard-error units. The standardized variable $z = (\bar{x} - \mu)/(\sigma/\sqrt{N})$ follows the standard normal distribution for a normal parent population. Standardization removes the original unit and maps many measurement problems to a common probability scale. Tables or software can then provide central probabilities and tail areas.

$$z = (\bar{x} - \mu)/(\sigma/\sqrt{N})$$

A z value of 2 means that the observed sample mean lies two known standard errors above the population mean. It is not automatically evidence of a mistake; its interpretation depends on the question, the tail convention, and the validity of the normal model. Standardization is a mathematical transformation, while scientific meaning still comes from the measurement context.

$$z \sim N(0,1)$$

Further interpretation

Known σ is uncommon in ordinary laboratory work. It may be defensible when a stable process has been characterized by extensive previous data and current conditions match that characterization. Substituting the small-sample estimate s into a z formula without changing the reference distribution understates tail uncertainty. That substitution leads to Student's t distribution, discussed below.

A z score must be paired with a probability convention. A two-sided question uses probability in both tails beyond $|z|$, while a one-sided question uses only the relevant tail. The same numerical z therefore produces different p-values depending on the pre-specified alternative. In confidence intervals the analogous choice determines whether α is divided by two. The physical problem should decide the convention: agreement around a target is commonly two-sided, whereas an upper contamination limit or minimum strength requirement may justify a one-sided bound.

Standardization also enables comparison across quantities with different units, but only after each deviation has been divided by its own valid standard error. Comparing raw deviations or percentages can be misleading when scales and precisions differ. A dimensionless statistic creates a common reference, while the underlying measurement units should remain visible in the reported estimates and intervals.

In practice, label whether a probability is one-sided or two-sided before consulting software or a table. Then sketch the shaded tail area. The sketch often prevents use of the wrong cumulative probability.

4. Probability intervals for a random sample mean

In the idealized case where μ and σ are known, a probability interval describes where a future random sample mean is likely to fall. For a chosen coverage factor k, the central interval $\mu - k\sigma_{\bar{x}} < \bar{x} < \mu + k\sigma_{\bar{x}}$ contains a specified fraction of the sampling distribution. Under the standard normal model, $k = 1$ gives 68.27%, $k = 1.96$ gives 95.00%, and $k = 2$ gives approximately 95.45%.

$$P(\mu - k\sigma_{\bar{x}} < \bar{x} < \mu + k\sigma_{\bar{x}}) = 1 - \alpha$$

The center μ is fixed and known in this imaginary setup; the random object is $\bar{x}$. Repeating the experiment produces different sample means, and the probability statement concerns their long-run frequency within the fixed interval. This predictive framing is useful for understanding the mathematics but differs from the usual experimental problem, where μ is unknown and the observed $\bar{x}$ is already fixed.

$$P(|z| < k) = 1 - \alpha$$

Further interpretation

Coverage grows as k increases because the interval includes more of the distribution. The chosen central area leaves $\alpha/2$ probability in each tail for a two-sided interval. One-sided limits allocate the tail probability differently and require different critical values. The interval form must therefore match the scientific question before a table value is selected.

Probability intervals are also useful for quality control when process parameters are treated as known from a long stable history. They predict the expected variation of future sample means of a specified size. If a new mean falls outside the interval, possible explanations include a rare fluctuation, a changed process, underestimated σ or dependence among observations. The event is evidence to investigate, not proof of a specific cause. Because the interval width depends on N, control limits for individual readings cannot be applied unchanged to averages of several readings.

The central normal interval is symmetric because the standard normal distribution is symmetric. If the sampling distribution is skewed or bounded, equal-tail and shortest intervals can differ and neither may be

symmetric around the estimate. In those cases a transformation, exact distribution or simulation-based interval may provide more appropriate coverage than applying a normal factor by habit.

In practice, identify the fixed center and the random quantity in a probability interval. This single labeling step separates prediction of $\bar{x}$ from estimation of μ .

5. Confidence intervals when σ is known

A real experiment reverses the roles. After observation, $\bar{x}$ is known, while μ remains fixed but unknown. Algebraically rearranging the probability statement yields a random interval procedure: $\bar{x} - k\sigma/\sqrt{N} < \mu < \bar{x} + k\sigma/\sqrt{N}$. Before data are collected, the interval endpoints are random because they depend on the future sample. After calculation, one particular interval has been obtained.

$$CI\mu: \bar{x} \pm k\sigma/\sqrt{N}$$

A $100(1-\alpha)\%$ confidence interval is centered on the observed sample mean and has half-width $k\sigma/\sqrt{N}$. In repeated applications of the same procedure, the stated proportion of intervals would contain μ . Larger N narrows the interval, a larger σ widens it, and a higher confidence level requires a larger k . These tradeoffs connect experimental design directly to the desired precision and confidence.

$$\text{width} = 2k\sigma/\sqrt{N}$$

Further interpretation

Probability and confidence intervals may look algebraically similar but answer different questions. A probability interval predicts a random statistic around a known parameter; a confidence interval estimates an unknown parameter using a random interval procedure. Keeping the random object explicit prevents incorrect verbal interpretations and helps distinguish interval estimation from prediction of a future individual observation.

The confidence interval formula shows how to plan sample size when σ is available from a pilot study or established process. To target a normal-theory half-width E at confidence factor k , solve $N \approx (k\sigma/E)^2$ and round upward. For small planned samples, the t factor depends on the unknown final degrees of freedom, so planning is iterative or conservative. The calculation also assumes the future σ resembles the planning value. Allowance for unusable observations, clustering and practical constraints should be included before committing to a measurement campaign.

Sample-size planning should also account for the distinction between statistical and practical precision. An extremely large N can make a trivial difference statistically detectable without making it scientifically important. Define the required half-width or decision-relevant effect in physical units first, then determine whether the planned measurement resolution and control of systematic effects can support that target.

In practice, identify which parts of the confidence interval would change in a repeated experiment. The endpoints and $\bar{x}$ change; μ remains fixed. This is the core of the frequentist interpretation.

6. The frequentist meaning of confidence

A 95% confidence level describes the long-run performance of the interval construction method. If the experiment and interval calculation were repeated under the same conditions many times, about 95% of the resulting intervals would contain the fixed true mean and about 5% would miss it. The statement is about the procedure, not a probability distribution assigned to μ .

$$P_{\mu}(L(X) \leq \mu \leq U(X)) = 1 - \alpha$$

Once a particular interval has been computed, μ is either inside it or outside it. In the classical frequentist framework it is not correct to say that this fixed μ has a 95% probability of lying in the observed interval. The available confidence level still communicates how reliably the method captures μ across repeated sampling and is the standard interpretation expected in measurement reporting.

Further interpretation

Clear language matters because the same numerical interval can be misunderstood. Prefer: “The result is $\bar{x} \pm U$ at 95% confidence, using the stated procedure and assumptions.” Avoid suggesting certainty, guaranteed containment, or a posterior probability for μ . Confidence also does not measure model validity; violations of independence, normality or stable conditions can make the nominal 95% coverage incorrect.

Frequentist confidence is sometimes explained with a net that captures a fixed fish in a known fraction of repeated throws. The net changes from sample to sample; the parameter does not. This analogy emphasizes why a confidence procedure can be calibrated even though the truth is unknown. It also highlights a limitation: the advertised capture rate holds only under the model used to design the net. Selection bias, unmodeled drift or a different parent distribution can change actual coverage without changing the printed “95%” label.

Confidence intervals from different methods need not have the same coverage accuracy for small samples. A method can be exact under normality, conservative, or approximate. State the assumptions and procedure rather than treating “95%” as a universal property of two endpoints. This becomes especially important for ratios, bounded parameters and nonlinear transformations of an estimated mean.

In practice, replace the phrase “probability that μ is inside” with a statement about the percentage of repeated intervals that contain μ . The revised wording should retain the stated confidence level.

7. Unknown σ and the estimated standard error

In practice, σ is usually unknown and is estimated from the same N measurements. The sample standard deviation uses $N-1$ in the denominator because the deviations from $\bar{x}$ must sum to zero; one degree of freedom has been consumed by estimating the mean. The estimated standard error of the mean is $s_{\bar{x}} = s/\sqrt{N}$. Since s varies from sample to sample, this denominator introduces additional uncertainty into the standardized statistic.

$$s = \sqrt{\{\sum_i (x_i - \bar{x})^2 / (N - 1)\}}$$

The sample standard deviation characterizes scatter of individual observations, while $s/\sqrt{N}$ characterizes uncertainty of the estimated mean under the sampling model. They answer different questions and should not be interchanged. Reporting only s can make the mean appear less precise than it is; reporting only $s/\sqrt{N}$ hides the variability of individual observations and the sample size.

$$s_{\bar{x}} = s/\sqrt{N}$$

Further interpretation

The estimated standard error inherits all limitations of the sample. With very small N , s is unstable and may seriously underestimate or overestimate σ . A time trend, outlier, rounding effect or multimodal population can make it misleading. Graphical inspection and procedural knowledge should accompany the calculation, particularly when the interval will support a high-stakes comparison.

The Bessel correction makes the usual sample variance unbiased under independent sampling, but unbiasedness alone does not guarantee a precise estimate. The distribution of s is broad at small N . Confidence intervals for σ or variance can be built from a chi-square distribution under normality, although that extension is beyond the central calculation here. The key practical lesson is that uncertainty in scatter must not be ignored. Student's t incorporates it into inference for μ and produces appropriately heavier tails and wider intervals.

Independence is again central to the $\sqrt{N}$ reduction. If successive readings share drift, the effective number of observations may be far below N . Plotting residuals against acquisition order and calculating autocorrelation can reveal this problem. A better design may alternate conditions, randomize order or allow stabilization time instead of collecting many closely spaced readings that repeat the same disturbance.

In practice, calculate all deviations from $\bar{x}$ and verify that their sum is zero apart from rounding. This confirms the constraint responsible for losing one degree of freedom.

8. Student's t statistic and distribution

Replacing σ by the random estimate s changes the standardized variable. For independent observations from a normal population, $t = (\bar{x} - \mu)/(s/\sqrt{N})$ follows Student's t distribution with $\nu = N - 1$ degrees of freedom. The t distribution is symmetric and centered at zero but has heavier tails than the standard normal distribution. The heavier tails account for uncertainty in estimating σ from limited data.

$$t = (\bar{x} - \mu)/(s/\sqrt{N}), \quad \nu = N - 1$$

Student's t is a family of distributions. Small ν produces broad tails and large critical factors; as ν increases, s becomes a more stable estimate of σ and the t distribution approaches $N(0,1)$. At $\nu = 1$ it is the Cauchy distribution, whose mean and variance are not defined. For $\nu > 2$ its variance is $\nu/(\nu-2)$, which approaches one as ν grows.

$$t_{\nu} \rightarrow N(0,1) \text{ as } \nu \rightarrow \infty$$

Further interpretation

The t result is exact for a normal parent population. It is often robust to moderate deviations from normality when data are independent and not strongly skewed, but small samples provide little protection against outliers or heavy tails. The distribution should not be invoked mechanically: examine how the observations were generated and whether a mean-based interval is scientifically appropriate.

The name Student came from William Sealy Gosset's pseudonym, adopted when publishing work motivated by small-sample industrial experiments. The historical context matches the mathematical purpose: reliable conclusions were needed when collecting large datasets was expensive. The distribution is not restricted to classroom examples. It remains fundamental in calibration, materials testing, chemical analysis and any repeated-measurement problem where the population standard deviation is estimated from limited observations.

Student's t interval is invariant to changes of measurement unit: multiplying every observation by a constant multiplies $\bar{x}$, s and the interval endpoints by that constant while leaving t unchanged. This provides a useful check. Converting millimetres to metres should change only the numerical scale of the reported estimate and uncertainty, never the confidence level or degrees of freedom.

In practice, compare a t critical factor with the corresponding normal factor. The t value should be larger for finite ν and should approach the normal value as ν increases.

9. Degrees of freedom and finite-sample penalty

Degrees of freedom count the independent pieces of information remaining after constraints are imposed. Once $\bar{x}$ is computed, the N deviations $x_i - \bar{x}$ must sum to zero. If $N-1$ deviations are known, the last is determined, so only $\nu = N-1$ can vary independently in estimating variance. This loss of one degree of freedom is reflected in both the Bessel correction and the t distribution.

$$\nu = N - 1$$

For a fixed factor k , central coverage under t is smaller than under the normal distribution because more probability lies in the tails. Conversely, to obtain the same target confidence with small ν , the t critical factor must be larger. For a two-sided 95% interval, $t_{0.975,1} = 12.706$, $t_{0.975,4} = 2.776$, $t_{0.975,9} = 2.262$, and the large-sample limit is 1.960.

$$k = t_{1-\alpha/2, \nu}$$

Further interpretation

The finite-sample penalty can be substantial. A small dataset may have a modest s yet still require a wide confidence interval because s is poorly known. Adding measurements narrows the interval in two ways: it reduces $s/\sqrt{N}$ and increases ν , which lowers the t factor. Planning should use both effects rather than the square-root rule alone.

Degrees of freedom need not always equal $N-1$. That value applies to the one-sample mean and variance problem. Regression, weighted combinations, and uncertainty budgets can have different or effective degrees of freedom because several parameters are estimated or components have different information bases. The Welch–Satterthwaite approximation is often used to obtain an effective ν for a combined standard uncertainty. Whatever method is used, report it explicitly; inserting $N-1$ automatically into every t calculation is not generally valid.

In regression, residual degrees of freedom equal the number of observations minus the number of fitted parameters under the standard linear model. This generalizes the one-sample loss of one degree of freedom. It reminds us that information spent estimating nuisance parameters is unavailable for estimating scatter, and small residual degrees of freedom require wider t -based intervals for fitted coefficients.

In practice, for $N=2, 5$ and 30 , locate ν and the 95% critical factor. The sequence makes the finite-sample penalty visually clear and builds confidence in reading tables.

10. Choosing confidence level and critical value

For a two-sided central confidence interval with level $1-\alpha$, the total excluded probability α is divided equally between the two tails. The upper cutoff is therefore the $1-\alpha/2$ quantile. At 95% confidence, $\alpha = 0.05$ and the required quantile is 0.975. The resulting interval is $\bar{x} \pm t_{1-\alpha/2, \nu} s/\sqrt{N}$. Tables and software commonly list this upper cumulative probability.

$$CI_{\mu} = \bar{x} \pm t_{1-\alpha/2, \nu} s/\sqrt{N}$$

The confidence level should be selected before inspecting the data whenever possible. Higher confidence produces a wider interval; lower confidence produces a narrower one but a larger miss rate in repeated use. The choice depends on the consequences of error, disciplinary convention and intended decision. Reporting several levels after seeing the result can invite selective interpretation.

$$\alpha/2 + (1 - \alpha) + \alpha/2 = 1$$

Further interpretation

One-sided confidence bounds use the $1-\alpha$ quantile rather than $1-\alpha/2$ because all excluded probability lies in one tail. This distinction is important for detection limits, upper contamination bounds and safety thresholds. Always state whether the interval is one- or two-sided, the confidence level, degrees of freedom and the distribution used to obtain the critical factor.

Critical-value tables are indexed by both probability and degrees of freedom, so the correct row and column must be identified carefully. A two-sided 95% interval uses cumulative probability 0.975, not 0.95. Software may request the central confidence level, upper-tail probability, or cumulative quantile, and these interfaces are not interchangeable. Verify one familiar case, such as 1.960 for the normal 95% two-sided limit, before trusting the calculation. This simple check prevents a common and consequential tail-selection error.

For reproducibility, record the exact critical value rather than only writing “from a table.” Tables can differ in whether they show one-tail probabilities, two-tail probabilities, or central areas. Software output should be labeled with distribution, cumulative probability, and v . A short notation such as $t_{0.975,5}=2.571$ removes ambiguity and lets the reader verify the calculation directly.

In practice, for a 95% two-sided interval, write $\alpha=0.05$, $\alpha/2=0.025$ and $1-\alpha/2=0.975$ before selecting the quantile. This prevents the common 0.95-versus-0.975 error.

11. Standard and expanded uncertainty

The standard uncertainty of the mean is $u(\bar{x}) = s/\sqrt{N}$ when evaluated from repeated observations. Expanded uncertainty multiplies a standard uncertainty by a coverage factor k : $U = k u(\bar{x})$. When σ is unknown and the confidence statement is based only on N normal observations, k should be the appropriate Student t critical factor rather than an automatic value of two. The factor carries the chosen confidence level and degrees of freedom.

$$u(\bar{x}) = s/\sqrt{N}$$

A measurement result may be reported as $X = \bar{x} \pm U$ with unit and stated confidence. This compact form must be accompanied by enough information to interpret U : the coverage factor or confidence level, degrees of freedom or sample size, evaluation method, and important conditions. Without these, two identical plus-or-minus values can represent different statistical scales and cannot be compared safely.

$$U = k u(\bar{x})$$

Further interpretation

The widely used $k \approx 2$ convention corresponds closely to 95% only under normal or large-degree-of-freedom conditions. For small samples it can under-cover substantially. If several uncertainty components are combined and an effective degrees of freedom is used, document that calculation. Expanded uncertainty is not an additional physical error; it is a scaled interval statement based on the combined standard uncertainty.

Expanded uncertainty should not be presented with more digits than the data support. A conventional approach rounds U to one or two significant digits and rounds $\bar{x}$ to the same decimal place. The unrounded values should be retained internally so that premature rounding does not shift the final interval. The statement should distinguish standard deviation of observations s , standard uncertainty of the mean $u(\bar{x})$, and expanded uncertainty U . These three quantities can differ substantially and serve different interpretive purposes.

Coverage factor and confidence level are linked through a distribution, not interchangeable labels. The statement $k=2$ is incomplete unless the assumed distribution and degrees of freedom are known. Conversely, a stated 95% confidence level is incomplete unless the method used to obtain the interval is clear. Complete reporting prevents a conventional shorthand from being mistaken for a universally exact probability.

In practice, state whether each symbol is s , $u(\bar{x})$ or U . If two are assigned the same numerical value without explanation, the statistical scale has probably been confused.

12. Worked 95% interval for repeated length

Suppose $N = 6$ repeated length measurements give $\bar{x} = 25.42$ mm and $s = 0.18$ mm. The estimated standard error is $0.18/\sqrt{6} = 0.0735$ mm. Degrees of freedom are $\nu = 5$. For a two-sided 95% interval the Student factor is $t_{0.975,5} = 2.571$, so the half-width is $U = 2.571 \times 0.0735 = 0.189$ mm.

$$s_{\bar{x}} = 0.18/\sqrt{6} = 0.0735 \text{ mm}$$

The result is therefore $\bar{x} = (25.42 \pm 0.19)$ mm at 95% confidence with $\nu = 5$. If the normal factor 1.96 were used instead, the half-width would be only 0.144 mm. That narrower interval would ignore the finite-sample uncertainty in s and would not achieve the intended coverage under the model.

$$U = 2.571 \times 0.0735 = 0.189 \text{ mm}$$

Further interpretation

The calculation is short, but its interpretation depends on design. The six readings must be independent observations from a stable population, and the normal model must be reasonable. If the measurements are rounded coarsely or collected during drift, a t interval around the mean may not describe the actual uncertainty. Numerical correctness cannot compensate for a sampling procedure that fails to represent the measurand.

The worked length example can be audited by reversing the calculation. Dividing $U = 0.189$ mm by $t = 2.571$ recovers the standard error 0.0735 mm, and multiplying by $\sqrt{6}$ recovers $s \approx 0.18$ mm. This consistency check catches unit or transcription errors. The interval endpoints, 25.231 mm and 25.609 mm before reporting-

rounding, are not additional measured values; they are the limits produced by the confidence procedure. Their apparent precision should not exceed that of the rounded result.

If the length data were collected with a resolution of 0.1 mm, reporting a standard error to four decimal places would imply unsupported detail. Internal calculations may retain digits, but the final uncertainty and estimate should respect both statistical rounding and instrumental resolution. A well-rounded result is easier to read and less likely to suggest precision that the experiment cannot deliver.

In practice, reproduce the half-width using both $U = t \cdot s / \sqrt{N}$ and $U = t \cdot s_{\bar{x}}$. Both routes must agree before rounding and carry the same measurement unit.

13. Compatibility, unusual observations and Chauvenet's criterion

Two results can be screened for compatibility by comparing their difference with the standard uncertainty of that difference. For independent estimates $x_1 \pm u_1$ and $x_2 \pm u_2$, the normalized discrepancy is $|x_1 - x_2| / \sqrt{u_1^2 + u_2^2}$. A value near one is unsurprising under a common-value model, while a large value motivates investigation. Correlation and differing measurand definitions must be resolved before interpreting the statistic.

$$t_{\text{comp}} = |x_1 - x_2| / \sqrt{u_1^2 + u_2^2}$$

An unusual observation may be a rare random fluctuation, an unidentified systematic error, or a genuine physical effect. Visual surprise alone is not grounds for deletion. Chauvenet's criterion calculates the probability P_{worse} of an equal or more extreme deviation and considers rejection only when $NP_{\text{worse}} < 0.5$. In the lecture example, six values including 1.5 give approximately $P_{\text{worse}} = 0.10$, so $NP_{\text{worse}} = 0.6$ and rejection is not justified.

$$\text{Chauvenet screen: } NP_{\text{worse}} < 0.5$$

Further interpretation

Any rejection rule must be specified and documented. Recomputing the mean and standard deviation after removing a point changes the test and can bias the result if repeated opportunistically. First inspect records, instrument status and physical conditions. If a demonstrable procedural error exists, exclusion may be justified independently of a statistical screen; otherwise retain the observation or present a sensitivity analysis with transparent reasoning.

Compatibility and outlier tests answer narrower questions than scientific agreement. A small normalized difference indicates that the observed discrepancy is plausible given the quoted uncertainties, but it does not prove absence of common bias. A large difference signals tension but does not identify which result is wrong. Similarly, Chauvenet's criterion depends on a normal model and on estimates influenced by the suspected point. Robust summaries, replicate checks and physical diagnostics can supplement the screen. Transparency is more valuable than forcing every dataset into a tidy accepted/rejected classification.

For compatibility, shared calibration uncertainty can cancel in a difference. If two measurements use the same reference, treating their full uncertainties as independent may overstate the uncertainty of their difference and

hide a real discrepancy. Separate common and independent components before calculating t_{comp} . This is the same covariance principle used in uncertainty propagation and illustrates how the course topics connect.

In practice, calculate a normalized difference only after deciding whether the two uncertainties are independent. List shared calibration or reference components explicitly before forming the denominator.

14. Integrated confidence workflow and reporting

Begin with a defined population and a sampling plan. Collect independent observations, retain raw data and conditions, and inspect the sequence for drift or anomalies. Calculate $\bar{x}$ and s , then compute $u(\bar{x}) = s/\sqrt{N}$. Decide whether σ is genuinely known; if not, use Student's t with $\nu = N-1$. Choose the confidence level and whether the interval is one- or two-sided before selecting the critical factor.

$$\text{data} \rightarrow \bar{x}, s \rightarrow u(\bar{x}) \rightarrow z \text{ or } t \rightarrow k \rightarrow U \rightarrow \text{report}$$

Construct the expanded interval $U = ku(\bar{x})$ and report $\bar{x} \pm U$ with unit, confidence level, k and degrees of freedom. Check that the parent model and independence assumptions are adequate. When comparing results, propagate the uncertainty of the difference and account for correlation. Treat outliers as questions requiring physical and statistical evidence rather than inconvenient values to be removed.

$$X = (\bar{x} \pm U) \text{ unit, at stated confidence}$$

The logical chain is repeated measurement, summary statistics, standard uncertainty of the mean, selection of the appropriate reference distribution, critical factor, expanded uncertainty and complete reporting. Confidence communicates the long-run reliability of this interval procedure. It does not turn an incomplete model into a valid one, and it does not guarantee that the particular interval contains the true value.

Further interpretation

Good statistical reporting preserves the link between the data-generating process and the final interval. More measurements help only when they add independent information; a larger coverage factor widens the interval but does not correct bias; and an outlier rule cannot replace investigation. The goal is not merely to quote 95%, but to build a procedure whose assumptions justify approximately 95% coverage in the intended measurement population.

A defensible final workflow includes an audit trail. Preserve the original observations, the rule used to calculate s , the selected confidence level, the source of the critical factor and any exclusions or corrections. Repeating the analysis with and without questionable modeling choices can show whether the conclusion is stable. When results inform a decision, state that decision threshold separately from the statistical interval. Confidence intervals quantify sampling uncertainty under assumptions; practical acceptance limits, scientific relevance and systematic uncertainty require their own justification. Keeping these roles separate makes the final note useful for both calculation and critical evaluation.

Before finalizing, ask whether the interval addresses the intended quantity: a population mean, a future observation, a fitted parameter or a derived measurand. These require different constructions. Then audit the numerical chain, assumptions, units and wording. A correct confidence statement is simultaneously

mathematical, experimental and linguistic; failure in any one of these layers can change what the reported interval actually means.

In practice, end the analysis with both a numerical result and a sentence describing its confidence procedure. The result should be understandable without relying on a table heading or unstated classroom convention.

Consider a second compact application. A temperature is measured $N=10$ times, giving $\bar{x}=21.36$ °C and $s=0.24$ °C. The standard uncertainty of the mean is $0.24/\sqrt{10}=0.0759$ °C and $\nu=9$. For a two-sided 95% interval, $t_{0.975,9}=2.262$, so $U=0.172$ °C. The result may be written $T=(21.36\pm0.17)$ °C at 95% confidence, $\nu=9$. A normal factor of 1.96 would give 0.149 °C and would again produce an interval that is too narrow for the stated finite-sample model. This example reinforces the same sequence while showing that the numerical Student correction remains relevant even near ten observations.

Now distinguish that interval from prediction. The 95% confidence interval estimates the population mean temperature, not the range in which a future individual reading will fall. A future observation contains the full population scatter as well as uncertainty in the mean, so its normal-theory prediction half-width is based on $s\sqrt{1 + 1/N}$ multiplied by a t factor. With the example data, this is much wider than 0.17 °C. The distinction is operationally important: use a confidence interval when estimating a stable mean, and a prediction interval when setting expectations for a new reading. Stating the target of the interval prevents the small standard error of a well-estimated mean from being mistaken for the much larger variability of individual measurements.

Finally, confidence analysis should be integrated with the broader uncertainty budget. Repeated-measurement statistics may quantify one Type A component, while calibration, resolution, correction models and environmental influences contribute additional standard uncertainties. These components are first combined through the measurement model. A coverage factor is then selected for the combined result using an appropriate distribution or effective degrees of freedom. Applying a Student factor only to $s/\sqrt{N}$ while ignoring other contributions can understate the reported interval; applying it independently to every component can overstate it. The exact method depends on the information basis, but the reporting principle is constant: identify the combined standard uncertainty, justify the coverage factor, and state the resulting expanded uncertainty without implying more confidence than the model supports.

A final planning example makes the confidence tradeoff concrete. If prior information suggests $\sigma\approx0.50$ unit and the desired 95% normal-theory half-width is 0.10 unit, the starting estimate is $N\approx(1.96\times0.50/0.10)^2=96.04$, so at least 97 independent observations are required. If only ten observations are feasible, the expected half-width is much larger and must be accepted or the measurement process improved. The calculation separates limitations due to sample size from limitations due to intrinsic scatter.

Before reporting such a plan, test whether independence is realistic. If measurements arrive in correlated groups, 97 recorded values may provide far fewer than 97 independent pieces of information. Sampling across time, operators, or specimens may improve representativeness but can also increase observed variability by including reproducibility effects. The population definition decides whether that broader variation belongs in σ . A transparent plan states which variation is intentionally sampled and which conditions are controlled.

Confidence analysis is most useful when it changes practice. It guides how many observations to collect, warns when small-sample factors matter, separates uncertainty of a mean from scatter of individuals, and provides disciplined language for comparison. Used together with the propagation framework, it connects raw observations to a final expanded uncertainty without erasing the assumptions made along the way.