Statistical Analysis of Random Reflectance Measurements of Biochar

Abstract

Random reflectance (Ro) measurements are widely used to characterize the structural order of biochar, yet their interpretation is often complicated by distributional heterogeneity arising from microstructural variability, measurement geometry, and sample preparation effects. We present a statistical framework for analysing Ro point clouds in which the cumulative distribution function (CDF) constitutes the principal parameter for describing reflectance heterogeneity, as it captures global deviations from a compact, approximately Gaussian distribution across the full data range. Complementary diagnostics are derived from moment-based descriptors, Q–Q analyses, and upper-envelope metrics based on the highest reflectance values after robust outlier cleaning, supporting interpretation of distribution shape, dispersion, and tail behaviour. These diagnostics are integrated into a descriptive heterogeneity score that is fundamentally anchored in CDF-based deviation, with higher-order moments and tail metrics serving as contextual refinements. The score is not intended as a persistence classifier or a formal normality test, but as an indicator of how representative a single mean Ro value is for a given dataset. A Python code is provided via GitHub that generates publication-ready summary statistics, plots, tables, and concise interpretations, facilitating consistent reporting and comparative evaluation of biochar reflectance data.

1. Introduction

Random reflectance (Ro) measurements are obtained microscopically on polished epoxy-mounted sections using reflected-light photometry, with measurement points selected on carbonaceous material while avoiding non-target phases (e.g., epoxy-filled voids, minerals) and measurement artefacts (e.g., points too close to particle edges or poorly focused surfaces). This point selection requires analyst judgement and training, and inter-analyst variability is a known contributor to scatter in reflectance datasets; reflectance standards therefore emphasize representative sampling, unbiased measurement procedures, and qualified analysts.

Because a small fraction of measurements can still represent clear artefacts rather than true sample carbon, a conservative outlier exclusion step was tested as a first data treatment prior to calculating the mean Ro, provided the outlier criterion and the number of excluded points are reported transparently. In addition, a trimmed mean based on the central 5th–95th percentile range can be reported as a robustness (sensitivity) statistic that reduces the influence of extreme tails. However, distributional analyses such as testing for mono- or multimodality, skewness, tail

behavior, or Q–Q deviations should be performed on the full dataset and compared against the outlier-cleaned dataset, since trimming the tails artificially suppresses distributional information.

In earlier Ro studies, sample homogeneity was commonly inferred from unimodal histograms and low standard deviations of point-wise reflectance values. Standard deviation is reported in absolute Ro units and typically remains below 0.5 %Ro for most biochar samples. However, low standard deviation alone does not preclude asymmetric tails or thin high-reflectance sub-populations, motivating the additional heterogeneity diagnostics introduced here.

In the following, we first compare three approaches to Mean Ro: (i) analysis on all measurements, (ii) analysis restricted to the 5th–95th percentile range, and (iii) analysis after conservative outlier exclusion, to quantify how these treatments influence mean Ro and distributional interpretation. We then outlie the main principles of Q-Q diagnostics and present the maximum reflection metric. In the last step we introduce the Heterogeneity class – a descriptive indicator of Ro distribution complexity

2. Mean Ro Approaches

2.1 Data Treatment

The following two statistical data treatments are not mutually exclusive but complementary:

1. Fixed trimming (5th–95th percentile), which removes a constant 10% of observations irrespective of whether excluded values represent artefacts or valid measurements.
2. Outlier exclusion, which removes only a small number of measurements identified as implausible (e.g., inconsistent with expected reflectance behavior or measurement plausibility criteria) and reports the criterion applied and the number of excluded points.

For Ro point clouds ($n \approx 500$) used to compute the mean Ro and to evaluate distributional properties (Gaussian reference overlay and Q–Q diagnostics), outlier exclusion is generally the more defensible default because it preserves the full distribution while removing likely artefacts. Percentile trimming is best treated as a robustness (sensitivity) statistic, as it suppresses tail information by construction and therefore weakens interpretation of skewness and tail behavior. However, for biochar persistence assessment, reporting a trimmed mean can be justified by convention as a robust estimator.

Table 1: Comparison of fixed percentile trimming (5th–95th) and conservative outlier exclusion for the statistical treatment of Ro point clouds. In this study, trimming is used only as a robustness (sensitivity) statistic for the mean Ro, whereas distributional analyses (Gaussian reference overlay, Q–Q diagnostics, skewness and multimodality screening) are performed on the full dataset and evaluated in parallel on an outlier-cleaned dataset.

Aspect	Fixed trimming (5th–95th percentile)	Outlier exclusion (artefact cleaning)
Goal	Robust mean estimate by down-weighting tails	Remove likely artefacts while preserving distribution
Strength	Simple and fully reproducible	Maintains tails and shape; Q–Q and modality remain interpretable
Limitation	Truncates the measured distribution; forces "more normal" appearance	Requires a defensible, conservative criterion; may remove real subpopulations if too aggressive
Recommended use	Secondary robustness statistic for mean Ro	Default dataset for mean Ro and all distributional diagnostics

2.2 Outlier exclusion using a MAD-based criterion

Outliers are identified using a robust median absolute deviation (MAD) approach, which does not assume a Gaussian distribution and is therefore suitable for Ro point clouds that may show skewness or mixed populations. First, the median Ro is calculated as a robust location estimate. Absolute deviations from the median are computed for all measurements and the MAD is obtained as the median of these absolute deviations. Each observation is then converted to a modified z-score, defined as

$$z_i^{\backslash *} = 0.6745\ (x_i - \text{median})/\text{MAD}$$

and values with $|\ z_i^{\backslash *}\ | > 3.5$ are classified as outliers and excluded from the cleaned dataset. The number and fraction of excluded measurements are reported for each sample and the mean calculated from the cleaned dataset.

2.3 Recommend statistical analysis workflow

1. Primary analysis on the full dataset (no lower cutoff; no percentile trimming)
   – Report mean Ro, standard deviation, and key quantiles (5/25/50/75/95).
   – Use the Q–Q plot on the full dataset to assess deviations from a Gaussian reference across the entire range.
2. Conservative and transparent outlier handling
   – Apply a clearly defined, robust outlier criterion to exclude only measurements that are implausible or extreme and likely represent artefacts rather than sample carbon.
   – Report the outlier rule, the number of excluded points (n removed), and the final sample size used for the outlier-cleaned analysis.
3. Robustness reporting (recommended as a sensitivity check)
   – Report a 10% trimmed mean (5th–95th percentile) as an additional robust estimator of

central tendency.
– Do not use the trimmed dataset for Q–Q interpretation, skewness, kurtosis, or multimodality assessment, as trimming suppresses tail information by construction

3. Q–Q diagnostics for Ro point clouds (quantifying deviations from a Gaussian reference)

Normal Q–Q plots are a standard visual tool to compare an empirical Ro distribution against a Gaussian reference. In a Q–Q plot, the ordered Ro values are plotted against the corresponding theoretical quantiles of a normal distribution. If the distribution is close to Gaussian, the points fall approximately on a straight line. For Ro datasets with large point counts (typically $n \approx 500$), visual inspection alone can be subjective, and small but systematic tail features may be difficult to compare consistently across samples. Therefore, in addition to the plotted Q–Q diagram, we compute compact numerical diagnostics that quantify how closely the data follow the fitted Q–Q line and whether deviations are concentrated in the lower or upper tail.

The first diagnostic is the squared correlation coefficient of the Q–Q fit ($qq\text{-}r^2$). It is obtained from the linear fit of the ordered Ro values against the theoretical normal quantiles and provides a simple measure of overall linearity. High $qq\text{-}r^2$ indicates that most of the distribution follows an approximately linear relationship to the Gaussian reference, whereas lower $qq\text{-}r^2$ indicates systematic curvature or spread that reflects departures from a single compact shape. In this study, $qq\text{-}r^2$ is used as a descriptive indicator of conformity to a Gaussian reference and not as a formal normality test.

To make tail behavior explicit, we further quantify deviations in the lower and upper tails using the mean absolute error (MAE) relative to the fitted Q–Q line. After fitting the Q–Q line, residuals are computed as the difference between each ordered Ro value and its predicted value on the fitted line. The MAE of these residuals is then calculated separately for the lowest 5% and highest 5% of points. These tail MAE values are expressed in Ro units and represent the average absolute departure from the Gaussian reference line within each tail. Comparing the lower-tail and upper-tail MAE provides a concise description of whether deviations are more pronounced in the low-reflectance fraction (e.g., contributions from less carbonised domains or reactive material) or in the high-reflectance fraction (e.g., inclusions, strong heterogeneity, or the upper reflectance envelope captured by the maximum reflection metric).

Together, $qq\text{-}r^2$ and the tail MAE diagnostics provide a transparent and reproducible summary of Q–Q behaviour. They enable objective comparison across samples and across data treatments (full dataset versus outlier-cleaned), while keeping the Q–Q plot itself as the primary visualization. These diagnostics are therefore used to describe Q–Q linearity and to localize deviations to the lower or upper tail, without implying that the data are truly normal or that deviations are necessarily artefactual.

4. Maximum reflection metric

In reflected-light Ro measurements, the highest reflectance values within a point cloud can carry specific methodological information. A subset of measurement points may correspond to locations where the carbon surface is optimally prepared and measured – i.e., a well-polished, flat carbon area with ideal focus and measurement geometry, and minimal influence from adjacent phases or boundary effects. Under such conditions, measured reflectance approaches the intrinsic reflectance of the carbon domain more closely than points affected by edges, surface relief, epoxy proximity, or imperfect focusing.

To capture this "best-case" measurement subset in a reproducible way, we define the maximum reflection as the mean Ro of the 20 highest measurements after conservative outlier exclusion (MAD-based cleaning). This estimator is intended as a descriptive metric of the upper reflectance envelope under high-quality measurement conditions, rather than as a distributional tail diagnostic. In addition, the ratio between maximum reflection and the mean Ro of the cleaned dataset provides a dimensionless indicator of how strongly the upper-envelope reflectance exceeds the typical reflectance of the sample. Both metrics are reported alongside the full-data and cleaned means to support inter-sample comparison and to document measurement quality and heterogeneity without relying on a single central-tendency statistic.

5. Moment-based descriptors (skewness and excess kurtosis)

To summarize asymmetry and tail weight of the Ro distribution in compact form, we compute the third and fourth standardized moments: skewness and excess kurtosis. Skewness measures the direction and magnitude of asymmetry around the mean. Positive skewness indicates a longer right tail (more high-Ro values relative to the centre), whereas negative skewness indicates a longer left tail (more low-Ro values). Values close to zero indicate an approximately symmetric distribution.

Excess kurtosis measures tail weight relative to a Gaussian reference, with excess kurtosis of zero corresponding to a normal distribution. Positive excess kurtosis indicates heavier-than-Gaussian tails, while negative excess kurtosis indicates lighter tails. In this study, skewness and excess kurtosis are treated as descriptive indicators of distribution shape and are used as part of the heterogeneity scoring system. They are not interpreted as formal tests of normality and are best considered alongside quantiles, Q–Q diagnostics, and the outlier-cleaned summary statistics.

6. Local peak count from smoothed histogram (shape complexity indicator)

To provide a simple indicator of distributional shape complexity, we quantify the number of local maxima ("peaks") in the Ro histogram after mild smoothing. The Ro values are first binned into a fixed number of histogram bins (here: 30 bins). The resulting density histogram is then smoothed using a short moving average filter (here: a 3-bin window) to reduce noise-driven fluctuations. Local maxima are identified as bins whose smoothed density is higher than the adjacent bins, using a standard peak-finding algorithm. The resulting peak count is reported as the number of detected local maxima in the smoothed histogram.

The peak count is interpreted strictly as a heuristic descriptor of histogram irregularity, not as evidence of distinct modes. A single dominant peak is consistent with a compact, approximately unimodal shape. Two peaks can indicate a pronounced secondary structure. More than two peaks typically reflects increased scattering or fine-scale irregularity rather than a clearly separable mixture. Because peak counts depend on binning, smoothing, and sample size, they are used conservatively and only as one component of the overall heterogeneity score.

7. Cumulative distribution functions as a primary representation of Ro data

Ro measurements form point clouds that describe the optical response of a carbonaceous material at the scale of individual particles or domains. Any statistical description of such data must therefore address not only a single central value, but how measurements are distributed across the full reflectance range.

A cumulative distribution function (CDF) provides a direct and non-parametric representation of the data. This means that the measurements are not fitted to a predefined functional form or shape (such as a Gaussian bell curve), but are represented directly in cumulative form. For a given reflectance value Ro_i, the CDF counts how many of the other Ro measurements are less than or equal to Ro_i. This count is then divided by the total number of measurements, yielding a cumulative fraction between 0 and 1 (or equivalently 0–100%).

Formally, for a dataset of n measurements $\{Ro_1, Ro_2, \dots, Ro_n\}$, the empirical cumulative distribution function is defined as

$$F(Ro) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(R\,o_i \leq Ro),$$

where $\mathbf{1}(\cdot)$equals 1 if the condition is true and 0 otherwise. Thus, $F(Ro)$ represents the fraction of measurements with reflectance less than or equal to the value Ro.

When plotted, the empirical CDF (c.f., black curve in Figure 1 below) is obtained by sorting the Ro values in ascending order and assigning each ordered value its corresponding cumulative fraction between 0 and 1. The cumulative fractions themselves depend only on the sample size, but the reflectance values at which these fractions are reached are determined entirely by the data. It is precisely this mapping between cumulative fraction and Ro value that encodes the distributional structure of the reflectance dataset.

In contrast to histograms or other binned summaries, the CDF uses all measurements directly and does not depend on bin widths, smoothing parameters, or other subjective analysis choices. It therefore provides a stable and complete view of the reflectance distribution, capturing both the compactness of the central region and the behavior of the lower and upper tails within a single representation.

7.1 CDF-based comparison to a Gaussian reference

To interpret the empirical CDF, it is useful to compare it to a compact reference distribution. In this framework, a Gaussian reference CDF is constructed using the sample mean and standard deviation as descriptive parameters. This reference does not imply that the Ro data are expected to be normally distributed, nor is it used as a hypothesis test of normality. Instead, it serves as a standardized baseline representing a compact, unimodal distribution with the same central location and overall dispersion as the observed data.

Overlaying the empirical CDF with this Gaussian reference (cf. the red curve in Figure 1) makes cumulative deviations visible across the entire Ro range. Deviations that are confined to the lower or upper tails can be distinguished from deviations that are distributed across the central portion of the distribution. As shown in the next subchapter, this CDF-based comparison provides a direct visual and calculable measure of how strongly the observed reflectance point cloud departs from a simple, compact reference shape.

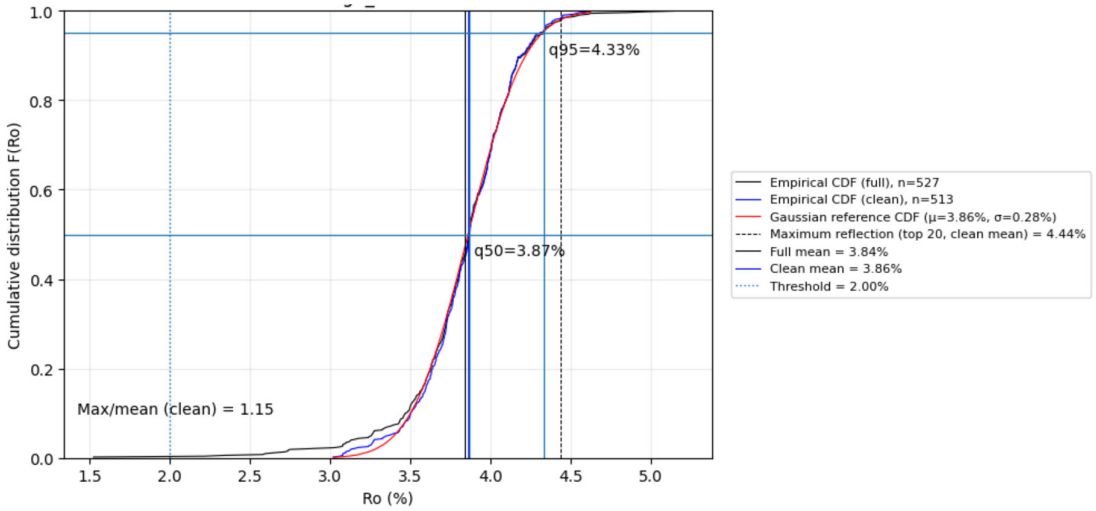

Figure 1: Empirical cumulative distribution functions (CDFs) of random reflectance (Ro) for a representative biochar sample, shown for the full dataset (black line, n = 527) and the outlier-cleaned dataset (blue line, n = 513). The CDF expresses, for each Ro value on the x-axis, the fraction of measurements with reflectance less than or equal to that value.

The close overlap of the full and cleaned CDFs across most of the reflectance range indicates that outlier exclusion has little influence on the central structure of the distribution, with minor differences confined to the extreme tails. A Gaussian reference CDF (red dashed line), constructed using the sample mean (μ = 3.86%) and standard deviation (σ = 0.28%) as descriptive parameters, is overlaid as a compact reference for visual comparison. This reference does not imply normality, but serves to highlight cumulative deviations of the empirical distribution from a simple unimodal shape across the full Ro range.

Horizontal reference lines indicate selected cumulative levels, including the median (q50 = 3.87%) and the upper quantile (q95 = 4.33%), which are annotated directly on the plot. Vertical lines mark the full-sample mean, the cleaned mean, the maximum-reflection metric defined as the mean of the highest 20 cleaned values, and the reactive-carbon threshold at 2.00% Ro. Together, these annotations illustrate how central tendency, upper-envelope behavior, and threshold-based metrics are embedded within the cumulative structure of the reflectance distribution.

8. Cumulative distribution–based diagnostics using the Cramér–von Mises statistic

While visual comparison of empirical and reference CDFs is informative, a quantitative measure is required to summarize cumulative deviations in a single, comparable statistic. For this purpose, we employ a cumulative distribution–based measure derived from the Cramér–von Mises (CvM) statistic that complements histogram- and Q–Q–based shape diagnostics

The CvM statistic quantifies the integrated squared deviation between the empirical CDF of the Ro dataset and a reference CDF, here chosen as the Gaussian reference constructed from the same sample mean and standard deviation. In contrast to pointwise diagnostics, which focus on the largest local deviation, the CvM statistic accumulates deviations across the entire reflectance range. It therefore provides a global measure of how strongly the observed reflectance distribution departs from a compact reference shape.

In the context of Ro point clouds, this property is particularly advantageous. Reflectance datasets often exhibit minor local irregularities arising from polishing artefacts, operator effects, or rare inclusions. Metrics that emphasize extreme deviations, such as Kolmogorov-type distances, can be dominated by a small number of points and may therefore overstate heterogeneity. By contrast, the CvM statistic responds primarily to systematic deviations that are distributed across the bulk or tails of the distribution, and is comparatively insensitive to isolated outliers.

Formally, the CvM statistic is computed from the empirical CDF $F(Ro)$ and the Gaussian reference CDF $F_{\mathrm{ref}}(Ro)$ as an integrated squared difference over all ordered observations. In practice, this integral is evaluated numerically using the ordered Ro values, yielding a single non-negative scalar that increases with increasing global distributional scatter.

For compact Ro datasets, CvM values are typically small, often on the order of $10^{-3}$to 10^{-2}. Such values should not be interpreted as numerical insignificance or absence of heterogeneity. Rather, they indicate that deviations from the Gaussian reference are weak and distributed, rather than concentrated in specific regions of the CDF. To improve readability and facilitate comparison across samples, CvM values are therefore reported throughout this work as $1000 \times \mathrm{CvM}$.

The purpose of CvM in this framework is not to reject a distributional model, but to provide a stable, scale-free indicator of global distributional complexity that can be compared consistently across samples and data treatments. The CvM statistic provides a compact numerical summary of cumulative distributional deviation. It complements tail-focused metrics and envelope-based descriptors by capturing the overall degree of reflectance scattering in a single, reproducible quantity, and forms a central component of the heterogeneity assessment introduced in the following section.

9. The Heterogeneity Score – a descriptive indicator of Ro distribution complexity

Random reflectance (Ro) measurements of biochar can exhibit very different distributional structures, even at similar mean values. Some samples are highly uniform and yield a narrow, near-symmetric reflectance distribution. Others show broader cumulative spread, pronounced upper envelopes, or clear departures from compact distributional shapes, reflecting internal heterogeneity of the carbon matrix, variable conversion conditions, mineral inclusions, or measurement-level variability. These differences matter for interpretation and quality control, because a single mean Ro value may conceal whether a dataset represents a compositionally uniform material or is instead influenced by a small fraction of highly reflective or weakly reflective domains.

For this reason, we introduce a heterogeneity score as a purely descriptive indicator that summarizes the complexity of the Ro distribution. The heterogeneity score has no direct impact on persistence classification and does not modify the mean Ro threshold approach. Instead, it provides complementary context for interpreting Ro datasets, particularly for borderline cases close to persistence thresholds and for deciding whether closer inspection or additional replication may be warranted.

9.1 Definition of heterogeneity in the context of Ro analysis

The heterogeneity score represents how strongly the observed Ro distribution deviates from a simple, compact shape that would be expected from a compositionally uniform material. Low heterogeneity corresponds to distributions with weak global cumulative scatter, limited upper-envelope contrast, modest asymmetry and tail weight, and a Q–Q relationship that remains close to linear across most of the reflectance range. Higher heterogeneity corresponds to distributions with broader cumulative structure, pronounced upper-envelope contributions, and clear departures from Q–Q linearity.

The heterogeneity score is a descriptive measure of distributional complexity and interpretability, indicating how well a single mean Ro value represents the underlying reflectance point cloud.

9.2 Inputs used for heterogeneity scoring

The heterogeneity score is computed from four parameter groups derived from the same Ro dataset:

1. **Global cumulative scatter indicator** based on the Cramér–von Mises (CvM) statistic, which quantifies the integrated squared deviation between the empirical Ro cumulative distribution function and a Gaussian reference constructed from the same mean and standard deviation (Chapter 8). For ease of reporting and scoring, the CvM statistic is expressed as a scaled quantity (1000 × CvM).
2. **Q–Q diagnostics** quantify how closely the ordered Ro values follow a Gaussian reference line. This is represented by qq-r^2, the squared correlation coefficient of the Q–

Q fit, and by tail mean absolute errors describing deviations in the lower and upper tails (Chapter 3).

3. **Moment-based descriptors**, namely skewness and excess kurtosis, which summarize global asymmetry and tail weight (Chapter 5).

4. **Upper-envelope indicator** based on the maximum-reflection metric, defined as the mean Ro of the highest 20 values remaining after robust outlier cleaning, expressed as the ratio of this maximum-reflection mean to the cleaned mean Ro (Chapter 4).

Each parameter group captures a distinct aspect of heterogeneity. The CvM statistic provides a global, bin-free measure of cumulative distributional scatter and forms the primary component of the score. Q–Q diagnostics complement this by emphasizing localized and tail-specific deviations. The maximum-reflection ratio characterizes the upper reflectance envelope and provides a physically interpretable indicator of the highest degree of carbon ordering achieved. Skewness and excess kurtosis contribute weakly as secondary descriptors of asymmetry and tail mass.

9.3 Calculation procedure and scoring scheme

The heterogeneity score is computed by converting each parameter group into a small sub-score and summing them. This approach is intentionally simple and robust, avoids reliance on a single sensitive statistic, and yields stable results across datasets of different size.

For each Ro dataset, four sub-scores are computed:

1. Global scatter score (S_2) from the scaled Cramér–von Mises statistic (1000 × CvM):

   Assign:
   – $S_1 = 0$ if $(1000 \times CvM) \leq 2$
   – $S_1 = 1$ if $2 < (1000 \times CvM) \leq 4$
   – $S_1 = 2$ if $4 < (1000 \times CvM) \leq 6$
   – $S_1 = 3$ if $6 < (1000 \times CvM) \leq 8$
   – $S_1 = 5$ if $(1000 \times CvM) > 8$

   Maximum contribution: 5 points

2. Q–Q deviation score (S_4) from qq-r^2, where qq-r^2 is the squared correlation coefficient of the Q–Q fit.

   Assign:
   – $S_2 = 0$ if qq-$r^2 \geq 0.95$
   – $S_2 = 1$ if $0.90 \leq$ qq-$r^2 < 0.95$
   – $S_2 = 2$ if $0.85 \leq$ qq-$r^2 < 0.90$
   – $S_2 = 3$ if qq-$r^2 < 0.85$

   Maximum contribution: 3 points

3. Moment score (S_3) from skewness and excess kurtosis. Each condition adds 0.5 points:

   Assign:
   – +0.5 if skewness ≤ -2.0 or ≥ 2.0
   – +0.5 if excess kurtosis ≥ 8.0
   Thus, $S_3 \in \{0, 0.5, 1.0\}$.
   Maximum contribution: 1 point

4. Upper-envelope score (S_1) from the maximum-reflection ratio R, where
   R = (mean Ro of top 20 cleaned values) / (cleaned mean Ro)

   Assign:
   – $S_1 = 0$ if $R \leq 1.12$
   – $S_1 = 1$ if $1.12 < R \leq 1.20$
   – $S_1 = 2$ if $1.20 < R \leq 1.35$
   – $S_1 = 3$ if $R > 1.35$

   Maximum contribution: 2 points

The total heterogeneity score is defined as:

$$\text{H-score} = S_1 + S_2 + S_3 + S_4,$$

with a maximum possible value of 10. The score is reported only if all required diagnostics are available; otherwise, it is reported as not available.

9.4 Interpretation and practical use

The heterogeneity score is a descriptive indicator of Ro distribution complexity and interpretability. It does not modify persistence class assignment and is not a formal test of normality. Instead, it supports quality control and indicates how representative a single mean Ro value is for the underlying dataset.

For practical use, the score is interpreted in four reporting bands:

- H-score ≤ 4: the dataset appears robust and suitable for reporting.
- 4 < H-score ≤ 6: the dataset shows some irregularities but still appears suitable for reporting.
- 6 < H-score ≤ 8: attention is recommended; targeted review of the CDF, Q–Q plot, and outlier handling is advised, and repeat imaging or additional measurements may be considered.
- H-score > 8: investigation of production conditions and/or repeat measurement is advisable.

High scores do not automatically indicate poor measurement quality. They may reflect genuine structural non-uniformity or mixture-like reflectance behavior at the scale of observation. However, they indicate that mean Ro values should be interpreted with increased caution, particularly near persistence thresholds.