Bootstrap using R: A useful approach for handling chunky data

Even though measurement is the foundation of all scientific and industrial processes, experimental data, precisely because it is derived from measurements, is inherently 'imperfect' as it is subject to various types of errors: random, systematic, or merely trivial.

The measurement "is the process of assigning numbers to represent qualities" [1] and is accomplished using a "measurement system", *i.e.*, "a collection of instruments or gages, standards, operations, methods, fixtures, software, personnel, environment, and assumptions used to quantify a unit of measure or fix assessment to the feature characteristic being measured".[2]

Precisely from this definition it is evident that the measurement process and the measurement itself are influenced by many factors such as:

- environment: external conditions such as:
  - temperature,
  - humidity,
  - vibrations, and
  - heat radiation

  can affect the measurements. These environmental factors can introduce disturbances or distort the measured values.

- object to measure: the characteristics of the object to be measured can significantly influence the accuracy and precision of the measurement. Variations in:
  - shape,
  - texture or
  - composition

  can introduce inherent limitations to the measurement process.

- method: the methodology chosen for the measurement, including the techniques employed and the procedures followed, can introduce variability in the measurement process. Factors such as:
  - sampling techniques,
  - data collection methods or
  - experimental setups

  contribute to the overall measurement uncertainty.

- operator: the role of the operator is significant in measurements. Factors related to the operator such as:
  - training,
  - skill,
  - sense of appreciation for precision,

    - attitudes toward personal accuracy

  can introduce measurement variation and error. Different operators may interpret measurement protocols or instrument readings differently, leading to inconsistent results.

- uncertainty on the measurement: every measurement inherently carries some level of uncertainty. Uncertainty arises from various sources, such as:
    - instrument limitations,
    - sample variations, or
    - errors in the measurement process itself.

  Proper estimation and uncertainty reporting are crucial for accurate data interpretation.

- measuring instrument, and its calibration: the quality and calibration of the instruments used directly affect the accuracy and reliability of the measurements. Regular calibration procedures, adherence to instrument specifications, and proper maintenance help keep measurement errors to a minimum.

- *etc.*

The number of digits with which they are reported also contributes significantly to the imperfection of the data and in this case, we speak of a rounding or truncation error.

In this technical note, the focus is on an in-depth analysis of the so-called 'chunky data' and an exploration of methods for their practical utilization, particularly in the context of data series comparison.

Chunky data is a term coined by Dr. Wheeler [3] who has extensively studied this type of data and discussed it in numerous publications [3-8]. The term chunky data is used to describe "measurements are made using measurement increments which are too large for the job" [4] and for which, therefore, "the distance between the possible values becomes too large [4].

Data of this type are frequently encountered in pharmaceutical QA / QC and are often the result of rounding off the experimental measurements [5].

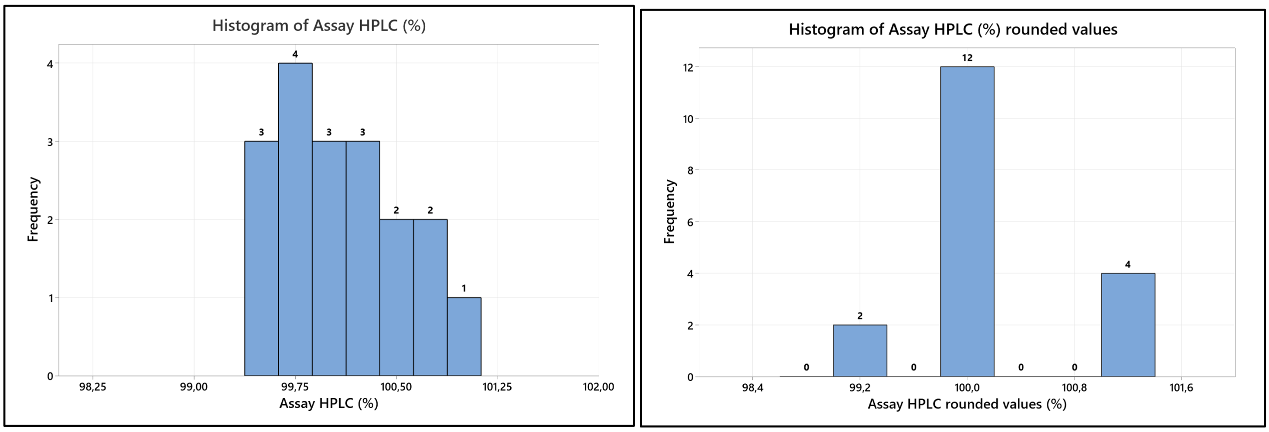
The impact this has in greatly reducing variability within a data set is evident. For example, consider the following two sets of HPLC assay measurements below. Each series consists of three groups of measures which, in one case, are reported with a decimal figure while in the other they are rounded to the corresponding integer.

Table 1

Measurement	Assay HPLC (%)	Assay HPLC rounded values (%)
1	100,0	100
2	100,3	100
3	99,7	100
4	99,4	99
5	99,5	100
6	99,8	100
1	100,3	100
2	100,5	100
3	100,6	100
4	99,9	101
5	99,4	100
6	100,7	99
1	100,8	101
2	100,9	101
3	100,3	101
4	100,1	100
5	99,7	100
6	99,8	100

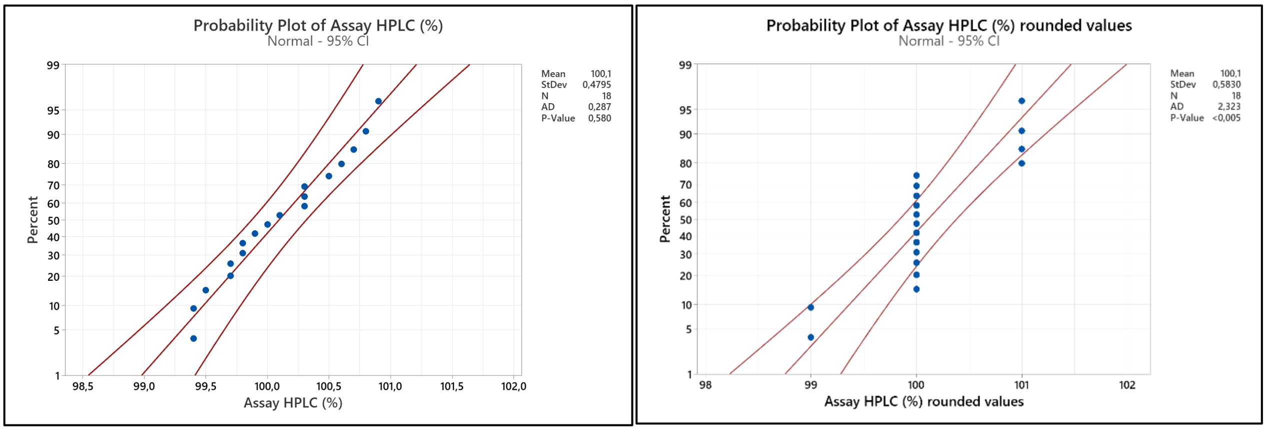
The comparison of the histograms relating to the two series of measures already shows at a glance the diversity existing between the two data sets (Figure 1). In the second case, in fact, three bins are sufficient to collect the eighteen values while in the first seven are needed.

Figure 1



Furthermore, while the first set of data is normally distributed, the second does not pass the Anderson-Darling test of normality. This is clearly visible from the comparison of the probability plots relating to the two distributions (Figure 2).

Figure 2

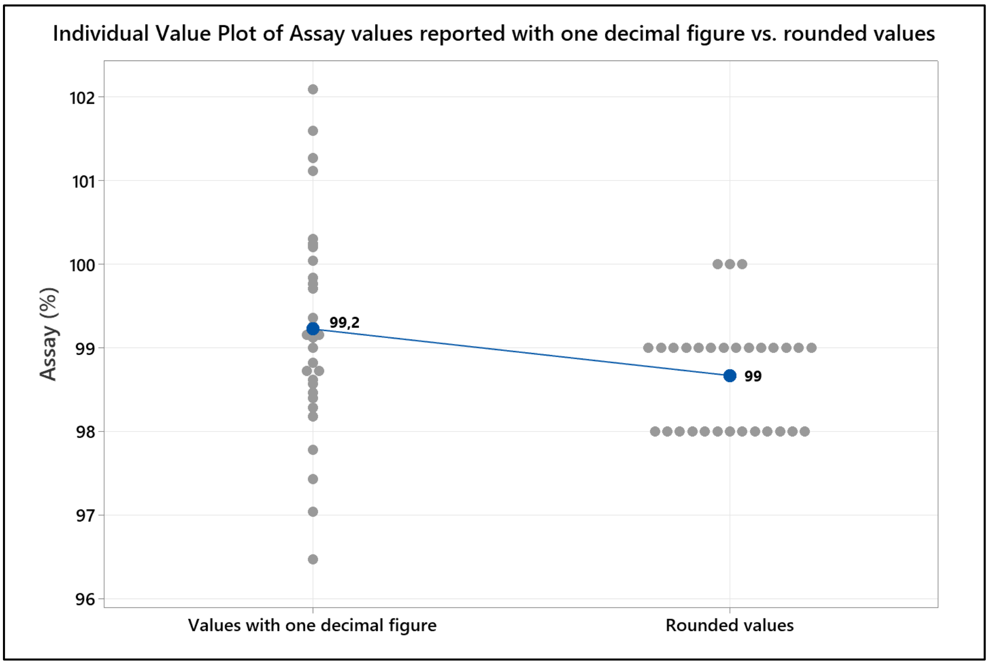


The presence of *chunky data* is particularly evident in probability plots. In fact, if present, they appear as groups of equal values stacked on top of each other as shown in the second probability plot. In the first case, however, thanks to the presence of the decimal figure, the variability of the data is maintained, and the values are distributed harmoniously along the straight line.

The normality tests are based on the hypothesis that the data are taken from a continuous distribution, and it is therefore possible that they detect as "non-normal" distributions with clustered values such as those resulting from the presence of chunky data.

Again, from a graphical point of view, the presence of chunky data is particularly evident also from the so-called *individual value plots*. The one shown in Figure 3 clearly shows the structure of the data in the two cases.

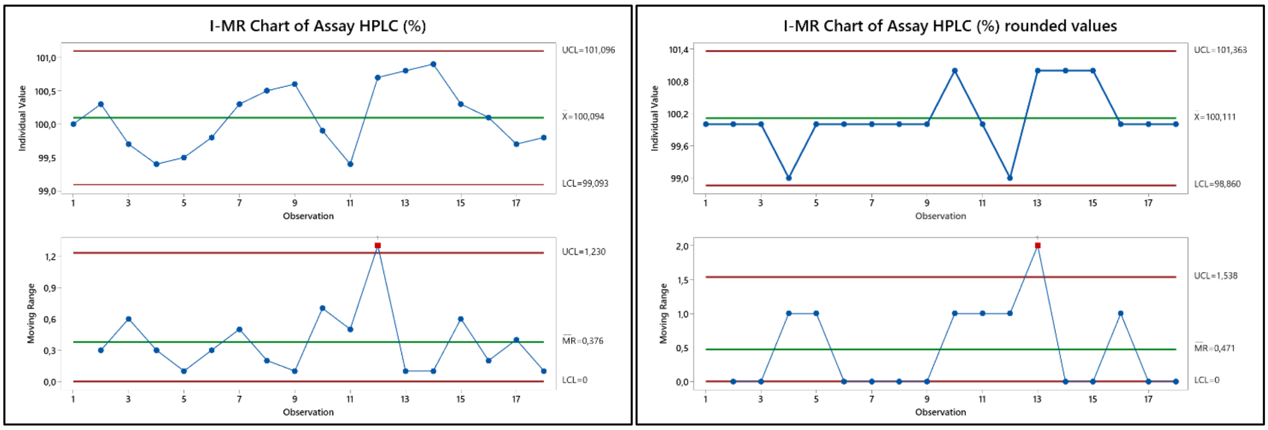
Figure 3



The effective difference between the average values is greater than what can be obtained from the comparison of the two numerical values shown in Figure 3 as the whole value is rounded up.

The presence of *chunky data* is also evident when using control charts. In fact, the following graphs illustrate how the *I-MR* charts relating to the two series of values shown in Table 1 appear. As described in the literature [4] the presence of *chunky data* could be the cause of numerous false alarms if the values are close to the control limits.

Figure 4



The easiest way to eliminate the chunky data problem would be to repeat measurements using smaller measurement increments [4]. In many cases, however, this is impossible. Consider, for example, the comparison between two sets of data that must be used as they are. A typical example is that represented by the comparative comparison between supplier data and the corresponding ones measured in-house on a given incoming product. In this case, the non-normality of one of the two series can hinder the correct application of tests such as the F-Test for Equality of Two Variances or the Two-Sample t-test which, in fact, require normality or quasi- normality of both datasets.

When the data do not satisfy the assumption of normality, several alternatives exist, including the use of nonparametric tests, such as the Mann-Whitney test to compare the means of the two groups, and the Levene test to verify homoskedasticity (equality of variances).

An important and useful alternative to nonparametric methods is represented by the *bootstrap* technique.

Bootstrap methods were introduced in 1979 by Efron [8] and are a class of nonparametric Monte Carlo methods that estimate the distribution of a population by resampling.

In practice bootstrapping is a statistical procedure that does not make any assumptions about data distribution and resamples, over and over with replacement, a single dataset to create many simulated samples. The central assumption for bootstrapping is that the original sample is representative of the actual population. It is therefore obvious that the larger the sample size, the better the result.

Let us consider, for example, precisely the case in which we want to evaluate whether there are significant differences between the average assay values and their dispersions, measured by the Supplier and obtained in-house by analyzing the incoming product.

As we said initially, the simplest approach in this case would be to compare the two data series using the *Two-Sample t-test* and the *F-Test for Equality of Two Variances*.

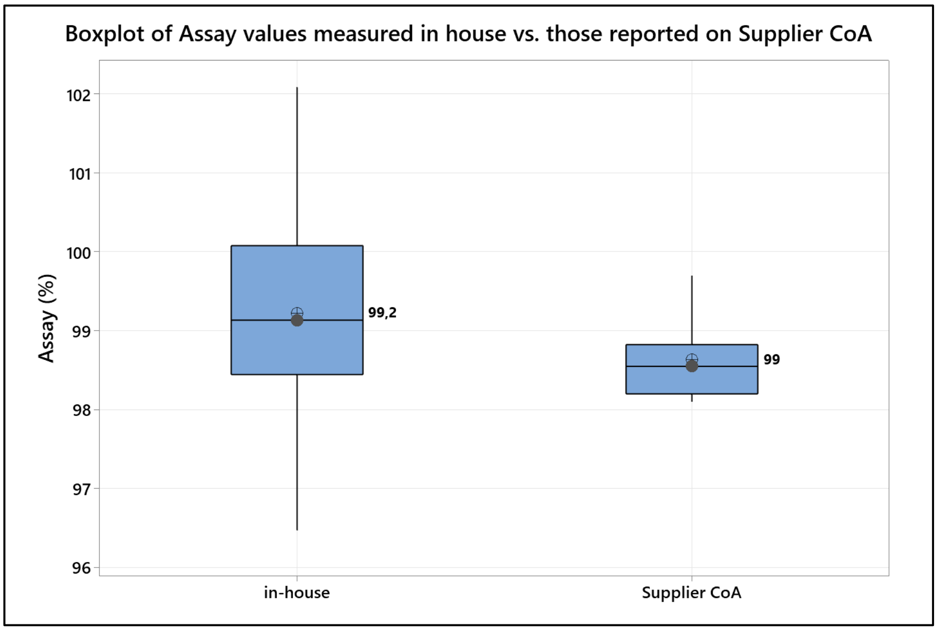
For example, consider the case of the thirty values shown in Table 2 and which refer to the HPLC assay values of as many lots of a given raw material measured on the incoming product and reported on the accompanying Supplier's certificate.

Table 2

Lot No.	Assay in-house	Assay Supplier CoA
1	99,2	99
2	100,2	98
3	99,7	98
4	99,3	98
5	99,2	98
6	99,4	98
7	98,8	98
8	100,3	98
9	102,1	99
10	101,6	99
11	101,3	99
12	101,1	99
13	97,8	100
14	99,1	99
15	100,0	99
16	100,2	98
17	99,8	99
18	98,6	99
19	98,3	99
20	97,0	98
21	97,4	98
22	98,6	98
23	98,2	98
24	99,8	99
25	98,7	99
26	98,7	100
27	98,5	99
28	98,4	99
29	99,0	98
30	96,5	100

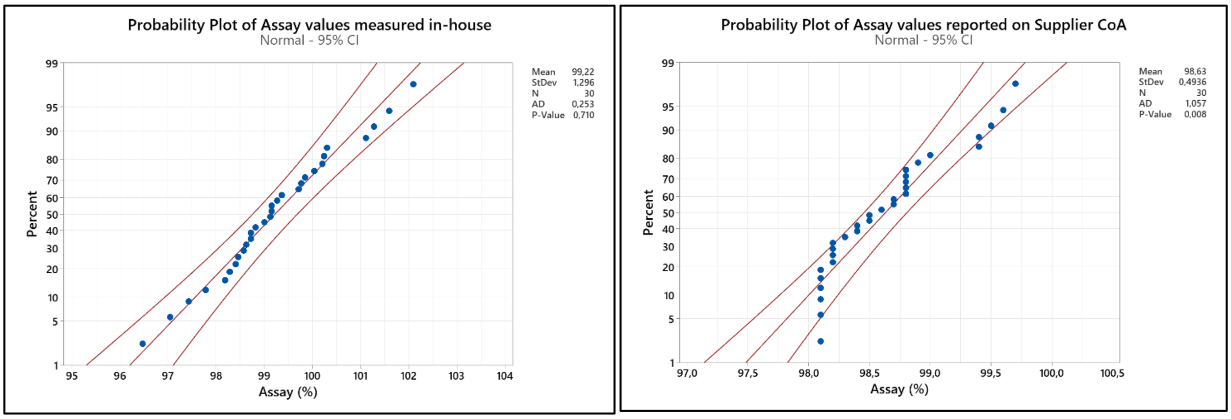
Even a simple data visualization, such as the one provided by the boxplots shown below, immediately returns the image of two quite different data distributions, the first more symmetrical and dispersed, while the second is more asymmetrical and compact.

Figure 5



Examination of the probability plots shows that the values measured in-house (and characterized by a decimal digit) appear normally distributed while those reported on the Supplier's certificates (and without decimals) are not normally distributed and this due to *chunky data*.

Figure 6



Unfortunately, the requirement of normality, or at least near-normality, is essential for the correct application of inferential tests such as:

- *2-Sample t-test*: which allows you to check if the means of the two distributions differ significantly from each other or not,
- F-Test for Equality of Two Variances: test for the null hypothesis that two normal populations have the same variance.

Now, in cases like these, the application of the bootstrap technique proves to be particularly useful because it is a very powerful method that can be used to estimate the distribution of the data sample and allows inferences to be made without having to make hard assumptions about the distribution of the initial data.

The most direct approach to comparing the two groups of data is certainly the one that uses the *bootstrap* to obtain confidence intervals for the difference between the means of the two groups, or for the difference between the variances, and then see if these intervals contain zero.

In practice the process would be as follows:

- For each of the two initial datasets, many bootstrap samples (for example, 1000) are generated and the mean or variance is calculated for each of them. In this way, for each group, there is a distribution of means or variances.
- For each pair of bootstrap means (or variances), the difference is calculated. This difference forms a new bootstrap distribution, which is the distribution of the difference between the bootstrap means (or variances) of the two groups.
- Now we calculate a confidence interval for this distribution of differences. This confidence interval is an estimate of where the "true" difference between the means (or variances) of the two groups is.
- If the confidence interval contains zero, one cannot reject the null hypothesis that the means (or variances) of the two groups are equal.

Thus, the confidence interval that you compute is a confidence interval for the difference between the means (or variances) of the bootstrap samples and serves as an estimate for the difference between the means (or variances) of the original groups.

An R script implementing this approach could, for example, use the sample() function in R to generate the bootstrap samples, and then the mean() and var() functions to compute the means and variances. The R code here below can be downloaded from my repository on GitHub at https://github.com/rbonfichi/bootstrap :

```
# Read data:

df <- read.csv2("C:/Users/Utente/Desktop/assay.csv")

data1 <- df$assay_1
data2 <- df$assay_2

# Number of bootstraps
n_bootstraps <- 1000

# Length original data
n <- nrow(df)
```

```
# Initialize vectors to store bootstrap means and variances
bootstrap_means_data1 <- numeric(n_bootstraps)
bootstrap_means_data2 <- numeric(n_bootstraps)
bootstrap_vars_data1 <- numeric(n_bootstraps)
bootstrap_vars_data2 <- numeric(n_bootstraps)

# Generate bootstrap samples and calculate means and variances
for (i in 1:n_bootstraps) {
  bootstrap_sample_data1 <- sample(data1, n, replace = TRUE)
  bootstrap_sample_data2 <- sample(data2, n, replace = TRUE)

  bootstrap_means_data1[i] <- mean(bootstrap_sample_data1)
  bootstrap_means_data2[i] <- mean(bootstrap_sample_data2)

  bootstrap_vars_data1[i] <- var(bootstrap_sample_data1)
  bootstrap_vars_data2[i] <- var(bootstrap_sample_data2)
}

# Calculates the differences between the bootstrap means and variances
mean_differences <- bootstrap_means_data1 - bootstrap_means_data2
var_differences <- bootstrap_vars_data1 - bootstrap_vars_data2

# Calculate 95% confidence intervals for differences
mean_difference_ci <- quantile(mean_differences, c(0.025, 0.975))
var_difference_ci <- quantile(var_differences, c(0.025, 0.975))

# Print confidence intervals
print(round(mean_difference_ci, digits =4))
print(round(var_difference_ci, digits = 4))
```

Basically, the above code:

1. Reads a CSV file from the local disk containing the experimental data you want to compare, which is stored in columns "assay_1" and "assay_2".
2. Set the number of bootstraps to 1000.
3. Initializes four vectors to store the bootstrap sample means and variances for each of the two datasets.
4. Generates bootstrap samples for each dataset and calculates means and variances for each sample.
5. Calculates the difference between the bootstrap sample means and variances for each group.
6. Calculates 95% confidence intervals for the differences between the bootstrap sample means and variances.
7. Prints the confidence intervals for the differences between the means and variances

Running this R script on the data listed in Table 2, the following output is generated:

```
> print(round(mean_difference_ci, digits =4))
  2.5%      97.5%
0.0232      1.0401
> print(round(var_difference_ci, digits = 4))
  2.5%      97.5%
0.4159      2.0665
```

The results obtained indicate that the 95% confidence interval for the difference between the means of the two data sets ranges from 0.043 to 1.050. Since this interval does not contain zero, one can reject the null hypothesis that the means of the two groups are equal, suggesting that there is a statistically significant difference between the means of the two data groups.

In addition, the 95% confidence interval for the difference between the variances of the two data sets ranges from 0.397 to 2.051. This interval also does not contain zero, which means that the null hypothesis that the variances of the two groups are equal can be rejected. This suggests that there is a statistically significant difference between the variances of the two data sets.
In summary, based on the results of the bootstrap analysis, there appears to be a statistically significant difference between both the means and the variances of the two data groups.

It is interesting to observe how, applying anyway the *F-Test for Equality of Two Variances* and the *Two-Sample t-test* to the initial data of Table 2, while the first unequivocally confirms the difference between the dispersions between the two data series, the *Two -Sample t-test* returns a result at least formally borderline. Indeed:

Test and CI for Two Variances: in-house vs. supplierCoA

Method

σ_1: standard deviation of in-house
σ_2: standard deviation of supplierCoA
Ratio: σ_1/σ_2
The Bonett and Levene's methods are valid for any continuous distribution.

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
in-house	30	1,296	1,680	(1,015; 1,770)
supplierCoA	30	0,661	0,437	(0,530; 0,882)

Ratio of Standard Deviations

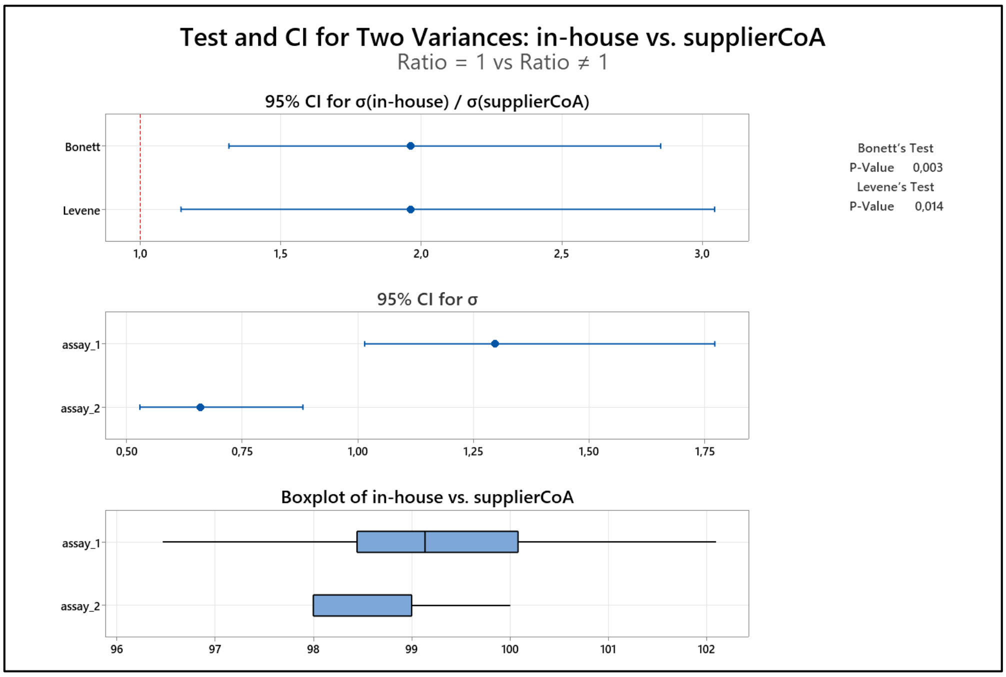
Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
1,96127	(1,317; 2,850)	(1,146; 3,043)

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0{,}05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	8,70	1		0,003
Levene	6,37	1	58	0,014

Figure 9



Two-Sample T-Test and CI: in-house vs. supplierCoA

Method

μ_1: population mean of in-house
μ_2: population mean of supplierCoA
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
in-house	30	99,22	1,30	0,24
supplierCoA	30	98,667	0,661	0,12

Estimation for Difference

Difference	95% CI for Difference
0,556	(0,020; 1,092)

Test

Null hypothesis	H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis	H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2,09	43	0,042

The study of chunky data and their peculiarities has offered the opportunity to apply the bootstrapping technique which has proved to be very useful for managing a situation of comparison between two series of data, one of which was normally distributed and the other was not.

Apart from the fact that, in the specific case, this difference was clearly due to the presence of chunky data, it occurs quite frequently in the practice of comparing experimental data series in the Quality Control / Quality Assurance pharmaceutical field. In fact, it is enough to think of the comparison between series of "naturally limited" data, higher or lower (*e.g.*, content of related substances), which are typically non-normal. Future studies could therefore focus on the application of bootstrap techniques to such experimental datasets. As the field of data science continues to evolve, it is critical to continually investigate and develop new methods for handling different types of data.

The case study addressed then highlighted the importance of defining reporting criteria for experimental data before they are generated. In fact, operator training in this regard can lead to more accurate data collection and thus improve the comparison and interpretation of experimental results. Often, in fact, it is not possible to repeat the measurements once we have noticed the problem.

In conclusion, the domain of chunky data analysis, as already highlighted by Drs. Wheeler [3-7] and Sleeper [9] several years ago, has proved to be full of opportunities for innovation and progress. Through continuous research and exploration, we can develop more effective strategies to handle complex data and leverage their unique characteristics to gain deeper insights into our data.

Bibliography:

1. Norman R. Campbell, *Foundations of Science*, Dover Publications, New York (1957) page 267
2. *Measurement Systems Analysis Reference Manual*, 4th Ed., AIAG, USA (2010)
3. D. Wheeler, *Don't be deceived by chunky data*, Quality Magazine (1999)
4. D. J. Wheeler, *What Is Chunky Data?*, Quality Digest (2011)
5. D. J. Wheeler, J. Beagle III, *Is That Last Digit Really Significant?*, Quality Digest (2018)
6. D. J. Wheeler, *EMP III (Evaluating the Measurement Process) Using Imperfect Data*, SPC Press Knoxville, Tennessee (2006)
7. D. J. Wheeler, *The Problem of Chunky Data*, Quality Digest (2023)
8. B. Efron, *Bootstrap methods: another look at the jackknife*, Annals of Statistics, 7:1–26, 1979
9. A. Sleeper, *Six Sigma Distribution Modeling*, McGraw Hill (2007)