MAPAN-Journal of Metrology Society of India (September 2020) 35(3):457–470
https://doi.org/10.1007/s12647-020-00383-4

REVIEW PAPER

A Review on the Statistical Methods and Implementation to Homogeneity Assessment of Certified Reference Materials in Relation to Uncertainty

A. Kumar and D. K. Misra*

CSIR National Physical Laboratory, Dr. K. S. Krishnan Marg, New Delhi 110012, India

Received: 26 October 2019 / Accepted: 16 June 2020 / Published online: 30 July 2020

Abstract: An importance of data analysis, methods for homogeneity test and standard uncertainty evaluation associated in any measurement for exact quantification of certified value of any product is vital to be stressed in the scientific community. Herein, we have collectively summarized the detailed discussion on the basics of statistical parameters such as mean, median, mode, standard deviation, variance, range, normal distribution, and central limit theorem. Various statistical analysis methods such as z test, t test, Chi-squared test, and ANOVA including F test have also been discussed in great detail to test the homogeneity of samples for certification of the reference material. The ISO guide 35 (2006) and Guide to Uncertainty in Measurement (GUM) are primarily considered to describe the basic concept of evaluating the associated uncertainty in the light of GUM modelling approach to avoid the error in the measurement which normally occurs in many scientific reports.

Keywords: Homogeneity; Certified reference material; Associated uncertainty; Certification; Measurement

List of Symbols

$\overline{X}$	Sample mean
μ	Population mean
σ^2	Variance
n	Number of data in sample
N	Number of data in population
SS_{within}	Sum of squares within the samples
MS_{within}	Mean of the square within the samples
X_t	True value
u_s	Standard uncertainty
U_{exp}	Expanded uncertainty
u_{bb}	Between the bottle uncertainty
u_r	Uncertainty due to repeatability
k	Coverage factor
σ	Standard deviation
σ_x	Standard error
H_0	Null hypothesis
H_a	Alternative hypothesis
$\overline{\overline{X}}$	Grand mean
$SS_{between}$	Sum of squares between the samples
$MS_{between}$	Mean of the square between the samples
X_{meas}	Measured value
u_c	Combined uncertainty
u_{stab}	Uncertainty due to stability
u_{within}	Within the bottle uncertainty
u_{meas}	Measurement uncertainty

1. Introduction

Statistical methods play a crucial role in data analysis and interpretation of results and have found great importance in various fields of physical, chemical, biological, agricultural, environmental, and social sciences with more applications in engineering. The inadequate interpretations of statistical analysis may lead to erroneous conclusions of any scientific study. The studies resulting in a large volume of raw data require suitable reduction so that data could be understood in a facile manner and provide information without affecting the result. It is worth mentioning that statistical errors can be found in many procedures such as assumptions-based statistical methods, traditional analysis, and randomly observed experimental data. The International Standard Organization (ISO) has formulated a set of basic guidelines for statistical analysis and enforced to perform detailed statistical analysis enabling a reader to

*Corresponding author, E-mail: dakkmisra@gmail.com

verify the reported results with associated uncertainty in the measurement [1–5].

A standard statistical procedure usually involves the test of the relationship between two statistical data sets, or a data set and synthetic data drawn from the idealized model. Basically, there are two main statistical methods used in data analysis: descriptive statistics summarizes data from a sample using indexes such as the range, the average or true mean, median, mode, variance, and standard deviation, while the inferential statistics summarizes the data obtained by random variations of the variables (e.g. observational errors, sampling variation). Descriptive statistics are primarily associated with population distribution, namely normal distribution, which is characterized by central limit theorem and secondly with dispersion that characterizes how random variables of the distribution depart from its centre.

The two statistical methods are normally compared by defining the hypothesis: null hypothesis and alternative hypothesis. When our observed sample has similar results as of the hypothesized results, it is called null hypothesis which means the results of sample and population are equally good, but when our observed results are not in support of hypothesized results then we conclude it as alternative hypothesis which contradicts the null hypothesis and we conclude one method is superior to other. The null hypothesis is performed using the statistical tests that give the acceptance or rejection of the null hypothesis. Type I errors (wrong rejection of null hypothesis) and Type II errors (wrong acceptance of the null hypothesis) are recognized in the null hypothesis. Several tests such as z test, t test, Chi-squared (χ^2) test, and analysis of variances (ANOVA) can be performed in the light of null hypothesis to verify the measurement errors and homogeneity assessment. Among all these tests, the analysis of variances (ANOVA) is commonly used in determining homogeneity using F test. It is worth mentioning that the certification of any reference materials requires performing homogeneity test and its validity to the certified values with associated uncertainty. Different types of reference materials on the basis of International Laboratory Accreditation Cooperation (ILAC) and ISO guide 35 have been discussed briefly with the principal steps of production, which is followed by the certification procedure in simple way for the better understanding of the reader. This will be discussed later in the light of ISO guide 35.

Uncertainty is another important parameter which is a quantitative measure of the quality of the value and needs to be emphasized in detail. Uncertainty can be derived by the values expressed as a standard deviation so-called standard uncertainty. Further, if transformed to the coverage interval, it gives expanded uncertainty with a stated coverage probability. The "Guide to the expression of uncertainty in measurement" (GUM) is considered as the primary document regarding the evaluation of uncertainty in measurement. An example of detection of creatinine and albumin detection in human urine for early stage detection of Kenny disease is explained to better understand the approach. Thus, to maintain the quality of measurement, calibration of instruments, traceability to SI unit of standards, and establishment of national and international measurement standards comparison in support of research and development and extensive knowledge of mathematical, statistical and metrological are primarily required.

Thus, from the above point of view, an overview of statistical parameters and statistical methods, including methods for homogeneity test and determination of standard uncertainty associated in any measurement, and how to correctly present and interpret the results is summarized in this review article. We have also discussed the most appropriate analysis methods to be implemented for different kinds of problems to avoid the error.

2. Basics of Statistics

Statistics is a branch of mathematics that deals with data collection, organization, analysis, and interpretation of the result.

2.1. Population and Sample

The population is the collection of all type of data or items involved in the analysis under investigation, while the sample is the subset of the population from which information is collected and results are provided [6].

2.2. Type of Statistics

There are two types of statistics, which are:

1. *Inferential statistics* consist of methods in which we draw the conclusion and measuring its reliability about population and giving the predictions based on information obtained from a randomly selected sample of the population.
2. *Descriptive statistics* consist of methods for data organization and summarizing the information from the same. It helps us constructing graphs, charts, and tables, and the calculation of various parameters such as averages, mode, median, range, standard deviation, variance, central tendency, and percentiles [7].

2.3. Mode

It obtains the frequency of the given variable and provides the variable which has the greatest occurrence in the given data. It is worth mentioning here that the variable has no mode if the greatest frequency of variable is one. Further, any value that occurs maximum times is called mode of a sample if the frequency of occurrence is 2 or greater.

2.4. Median

The sample median of the given variable is obtained by arranging observed values in a data set in increasing order, and then the middle value of the ordered list will be median. Two middle values exist if the number of observations is even; hence, the sample median is half of the sum of those two numbers.

2.5. Mean

It is the parameter that is generally used to calculate the centre for a given variable. It is also called a arithmetic mean or average in general understanding. It is the sum of all the observed values of given variable in data divided by the number of observations.

Let the sample size of the variable x_i be n, and then the mean is expressed operationally

$$\overline{X} = \frac{\left(\sum_{i=1}^{n} x_i\right)}{n} \tag{1}$$

2.6. Range

The sample range of a sample is the difference between its highest and lowest values of the variable in a given data set, i.e.

$$\text{Range} \;=\; V_{\max} - V_{\min} \tag{2}$$

2.7. Standard Deviation

The standard deviation of a sample is the most commonly used parameter that measures the variability, although it is different from range. It can be understood as an average of the difference of mean value from the observed value of the variable in the given data which is also known as the deviation from means position and is presented by SD

$$\sigma_s = \sqrt{\frac{\left\{\sum_{i=1}^{n}(x_i - x)^2\right\}}{n-1}}, \quad \sigma_p = \sqrt{\frac{\left\{\sum_{i=1}^{n}(x_i - x)^2\right\}}{N}} \tag{3}$$

While dealing with population in place of sample, the "n − 1" term is replaced by "n".

2.7.1. Standard Error of the Mean

The standard error of the mean is generally evaluated by dividing the standard deviation by the square root of the sample size and can be expressed as:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \tag{4}$$

where "s" and "n" are the usual notation of standard deviation and sample size, respectively.

2.8. Variance

It is the square of standard deviation, which can be estimated after evaluating the standard deviation. Population variance (σ^2) is the sum of squared deviation or variable from mean divided by n and may be expressed as

$$\sigma^2 = \left[\frac{\left\{\sum_{i=1}^{n}(x_i - x)^2\right\}}{N}\right] \tag{5}$$

On the other hand, the sample variance (σ^2) is the sum of squared deviation or variable from mean divided by ($n - 1$) and can be expressed as

$$\sigma^2 = \left[\frac{\left\{\sum_{i=1}^{n}(x_i - x)^2\right\}}{n-1}\right] \tag{6}$$

where n is the number of observations taken into account; one should note that when we deal with population, the denominator is "N" but in case of sample it changes to "n − 1" (for both standard deviation and variance) which is called degree of freedom.

2.9. Normal Distribution

The normal distribution is the most commonly used statistics based on the assumption that the probability of random distribution of measurement is normal in its distribution. It is also known as Gaussian distribution. A random variable X is said to have a normal distribution if its density curve is symmetric and continuous similar to Gaussian curve. The normal distribution is characterized by its mean μ and standard deviation σ and is denoted by, $N(\mu, \sigma)$ [8]. The normal distribution curve is generated with a predefined special function, which represents the density of any random variable x and is expressed as

$$F(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} \tag{7}$$

From the above equation, it can be seen that the normal distribution function is completely determined by μ and σ (Fig. 1).

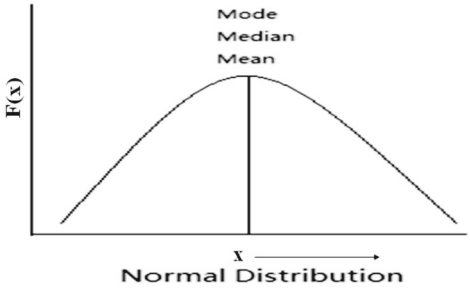

Fig. 1 Normal distribution representation with mean, median, and mode

In a normal distribution, the mean, median, modes all lie on the same line. So basically a distribution will be called a normal distribution if it's mean mode and median has the same value. Where "x" is independent parameter and "F(x)" is density of variable x at every possible x value.

It is important to note that if the distribution is not normal, it will be skewed to either the left or right side [9]. Here, the skewness is created which provides us a different value for the same parameters, as shown in Fig. 2.

The standard normal distribution (Fig. 3) graph can be generated by using normal distribution function $F(x)$ by taking a mean of zero and a standard deviation of 1 which is represented below.

In a standard normal distribution, the coverage factor is an important parameter which integrates the standard deviation for the estimation of expanded uncertainty. Thus, a number which is multiplied with the standard deviation is called its coverage factor. This number represents how many standard deviations have occurred in an observation.

As shown in Fig. 3, for every number k, the probability density within the interval $[\mu - k\sigma, \mu + k\sigma]$ is similar for all the normal distributions, where k is known as the coverage factor of the distribution. Particularly, the probabilities of distribution of observation are as follows for $k = 1$, 2 & 3, respectively.

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683 \quad (8)$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954 \quad (9)$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997 \quad (10)$$

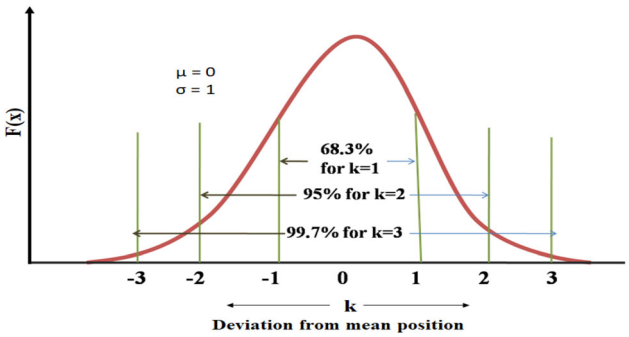

Fig. 3 Normal distribution with mean ($\mu = 0$) with a standard deviation ($\sigma = 1$), where k represents the coverage factor (either + ve and − ve) and $F(x)$ is density of x

2.10. Central Limit Theorem

It states that the sampling distribution of the mean of randomly taken data will approximately tend to a normal distribution (Fig. 4) as the number of observations increases [10]. So, the distribution of means of random samples having mean μ and finite variance σ^2 of a population with n (size of sample) will become the normal distribution with the same mean and variance as

$$\sigma_x = \sqrt{\frac{\sigma^2}{N}}$$

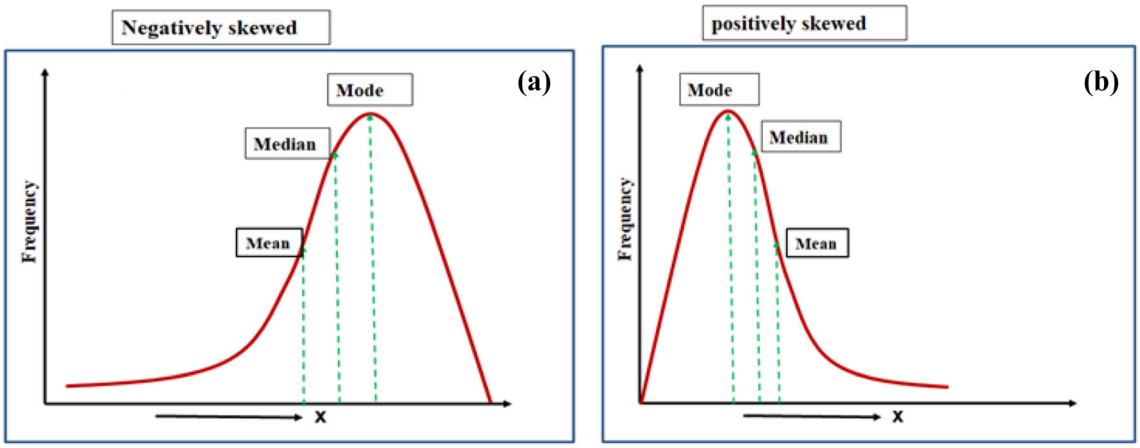

Fig. 2 Variation of mean median mode with a variation in distribution. **a** Negatively skewed distribution, **b** positively skewed distribution

which can be seen below in the diagram.

The central limit theorem allows us to make inferences about population parameters without knowing about frequency and variability.

3. Hypothesis

The hypothesis is a statement about some characteristics of a variable or a collection of variables [11] which is generally used to obtain inferences by comparing the data to the hypothesized values. There are two types of hypothesis, namely null hypothesis ($\boldsymbol{H_0}$) and alternative hypothesis ($\boldsymbol{H_a}$), which either supports the superiority or inferiority of the methodology. For example, if a country (population) has average 65 (hypothesized mean) publications per university and the observer considers 5 states (sample) for his evaluation and if he finds 65 as the average publication in these 5 states, then it will be considered as null hypothesis which supports the hypothesis prediction, whereas if the publications are more or less, then the hypothesized value is contradicted; hence, it will be called alternative hypothesis that means one method of evaluation is superior to others.

$$\mu = \mu(H_0 = 65) \text{ [null hypothesis]} \tag{11}$$

$$\mu \neq \mu(H_0 = 65) \text{ [alternative hypothesis]} \tag{12}$$

The level of significance (α) and level of confidence are a very important parameter in the testing of hypothesis. The significance level represents the probability of occurring our observation in the critical region, and the **level of confidence** witnesses how confident one can be with the observation that falls in a specific region of our interest, so-called confidence region [12].

The significance level may be one-tailed or two-tailed depending upon the condition of a range of random variables for a given population. A two-tailed level of significance with a level of confidence 95% is shown in Fig. 5.

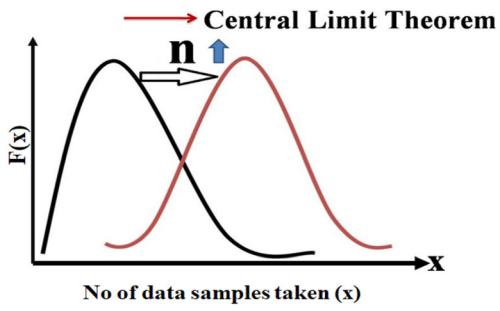

Fig. 4 Approaching to centre limit theorem with an increasing number of observations. As sampling size (n) increases, any distribution approaches to normal distribution

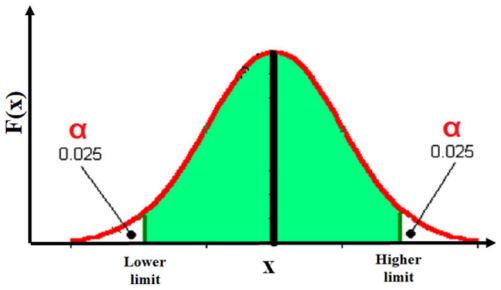

Fig. 5 Distribution of random data analysis exhibiting the level of confidence of 95%, where x is the independent variable with its density function $f(x)$

Further, on the way of testing of hypotheses, generally one can encounter two types of errors, namely Type I and Type II errors [11]. For example, one may reject a true hypothesis which is true H_0 and may accept the hypothesis which is false. The former is Type I error, and the later one is Type II error. Table 1 illustrates the errors for better understanding.

With a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. It is worth mentioning that both types of errors cannot be reduced simultaneously.

4. Methods for Hypothesis Testing

On the basis of a sample data available, there are various methods for hypothesis testing which help to decide about the trueness of a hypothesis. Many statistical tests have been developed by statisticians for testing a given hypothesis. These are as follows

(1) *z-test* It is commonly used for (a) the comparison of a sample mean to some hypothesized mean and (b) identifying the significance of the difference between means of two independent samples for large sample (≥ 30) with a known population variance [10].

There are **two cases** for z test, which are given here: *Case-I* When the population is normal and *infinite* and large sample size with the known variance σ_p^2 of the population, the z test can be defined as

Table 1 The type of errors

	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error
H_0 (false)	Type II error	Correct decision

$$z = \frac{X - \mu(H_0)}{\sigma_p^2/\sqrt{n}} \tag{13}$$

Case-II When the population is normal and *finite* and large sample size with the known variance of the population, the z test can be defined as

$$z = \frac{X - \mu(H_0)}{\sigma_p^2/\sqrt{n}\sqrt{\frac{N-n}{N-1}}} \tag{14}$$

where "N" is population size (total number of observations), "n" is sample size (number of observations in sample), "X" is sample mean and $\mu(H_0)$ is hypothesized population mean, and σ^2 is variance.

In such a test, it is important to determine the confidence interval for μ when σ has known to assume that the population variable Z is normally distributed. The normal z table [13] shows that within 1.96 of standard deviations from its mean, a probability of 95% for a random variable occurs [14], i.e.

$$P\left(\mu - 1.96\frac{\sigma_p^2}{\sqrt{n}} < Z < \mu + 1.96\frac{\sigma_p^2}{\sqrt{n}}\right) = 0.95 \tag{15}$$

and the random interval is

$$\left(Z - 1.96\frac{\sigma_p^2}{\sqrt{n}},\ Z + 1.96\frac{\sigma_p^2}{\sqrt{n}}\right) \tag{16}$$

(2) *t test* It is commonly used in the case when the sample size is small (≤ 30) with an unknown population variance. The only difference between the z test and t test is the size of the sample based on which one of the two tests is used.

There are two cases for t test, which are given here:

Case-1 When population is normal and *infinite*, with a small sample size together with unknown population variance, the t test can be estimated as

$$t = \frac{X - \mu(H_0)}{\sigma_s/\sqrt{n}}, \quad \sigma_s = \sqrt{\frac{\sum(X_i - X)^2}{n-1}} \tag{17}$$

where $(n-1)$ is degrees of freedom (d.f), X_i is observed ith value, X is the sample mean; here as we do not know the population variance, we will use sample variance as population variance.

Case-2 When Population is normal and *finite*, with a small sample size along with unknown population variance, the t test can be estimated as

$$z = \frac{X - \mu(H_0)}{\sigma_s^2/\sqrt{n}\sqrt{\frac{N-n}{N-1}}} \tag{18}$$

Similar to z test, the confidence interval of 95% for μ when σ is known assuming that the population variable X is normally distributed which can be written as

$$P\left(\mu - 1.96\frac{\sigma_s^2}{\sqrt{n}} < X < \mu + 1.96\frac{\sigma_s^2}{\sqrt{n}}\right) = 0.95 \tag{19}$$

and the random interval is:

$$\left(X - 1.96\frac{\sigma_s^2}{\sqrt{n}},\ X + 1.96\frac{\sigma_s^2}{\sqrt{n}}\right) \tag{20}$$

5. Chi-squared test

Chi-squared (χ^2) test is employed to have information about the sampling analysis for comparing a variance with the expected theoretical variance. It is generally used when one sample is given and we have to compare its variance with the population. It is worth mentioning that this technique helps the researcher to

(1) Identify the fitness of data [15];
(2) Test the significance of the connection between two parameters and
(3) Homogeneity test

This test is helpful in identifying whether a random sample belongs to the normal population with a mean (μ) and with a specified variance (σ).

One can estimate $\boldsymbol{\chi^2}$-distribution by multiplying degrees of freedom $(n-1)$ with the ratio of sample variances to the known population variance [16]. Thus, $\boldsymbol{\chi^2}$-distribution is expressed as

$$\chi^2 = \frac{\sigma_s^2(n-1)}{\sigma_p^2} \tag{21}$$

The $\boldsymbol{\chi^2}$-distribution is not symmetrical until we have large degrees of freedom (≥ 50). The smaller sample size gives low degrees of freedom, resulting in skewed distribution which can be better understood in Fig. 6.

6. Analysis of Variance (ANOVA)

ANOVA is used when multiple groups of sampling are to be analysed. The ANOVA technique enables us to compare the variance of more than two populations, such as in comparing the dose–effect from various categories of medicines on patients. Its principle is to test for estimation of variation of means of the populations by evaluating the variation of means within each of the samples, relative to the variation of means between the samples available. It is considered an ANOVA that each of the samples has to be drawn from a normal population and that each of these populations has the same variance. In other words, one can concern only those factor(s) which has to be studied by

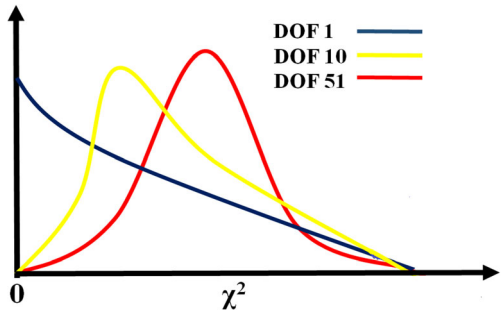

Fig. 6 Fitness of data can be obtained by comparing the calculated χ^2-value with the χ^2-tabular value [17]

neglecting others which may affect the results. In brief, one has to make two estimates of population variance: the first one is on the basis of between-sample variance and the other on the basis of within-sample variance [18]. Thus, the above-said two estimations are compared as follows

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{An estimate of population variance based on within samples variance}} \quad (22)$$

Technically speaking, this F value must be compared to the F-limit for given degrees of freedom (DOFs). If estimated Value exceeds the F-limit value, one may infer that there are significant differences between the sample means.

One-way ANOVA and two-way ANOVA are used to test the differences in means for the homogeneity test between the samples and within the sample. In one-way ANOVA, only one factor is considered, while in two-way ANOVA two factors are considered for testing the homogeneity. Generally, one-way ANOVA is frequently used for homogeneity test and to estimate the uncertainty aroused due to inhomogeneity in the materials which has been described in detail below.

6.1. Implementation of One-Way ANOVA Technique

In one-way ANOVA, we consider only one factor and one have to be ensured that several possible types of samples can occur within that factor [19]. Let us consider an example where some student measures three liquid samples of same volume in microlitres to understand the procedure for ANOVA.

Step 1 Estimation of the mean of each sample for a set of k samples, i.e.

$$\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots \overline{X}_k \quad (23)$$

Step 2 Estimation of the mean of the sample means as follows

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \cdots + \overline{X}_k}{\text{No. of samples } (k)} \quad (24)$$

Step 3 Determination of sum of squares between the samples (SS between).

The sum of squared deviations of the sample means from the grand mean when multiplied by the number of items in the corresponding sample results in the sum of squares between the samples ($SS_{between}$). Simply, it is variability of all samples from same population, as every sample can have a different mean than the grand mean as shown in Table 2. The number of items which is 5 as obtained from 5 measurements in each sample is multiplied with its square of difference to evaluate the variability of each sample from the grand mean which is known as squared deviation between the samples. Their sum is called sum of squared deviation and it can be expressed as

$$SS_{between} = n_1\left(\overline{X}_1 - \overline{\overline{X}}\right)^2 + n_2\left(\overline{X}_2 - \overline{\overline{X}}\right)^2 + \cdots + n_k\left(\overline{X}_k - \overline{\overline{X}}\right)^2$$

where, $n_1\left(\overline{X}_1 - \overline{\overline{X}}\right)^2$, $n_2\left(\overline{X}_2 - \overline{\overline{X}}\right)^2$ are the deviation of each sample mean from grand mean.

$$SS_{between} = 5(1) + 5(0) + 5(1) = 10 \quad (25)$$

Step 4 Estimation of the mean of the square between the samples: the squared deviation is obtained between the samples as above rather than standard deviation, and its mean is taken to standardize which can be obtained by dividing $SS_{between}$ by the DOF between the samples. Symbolically, it can be written as

$$MS_{between} = \frac{SS_{between}}{(k-1)} \quad (26)$$

where k specifies the number of samples and $(k-1)$ is DOF between the samples.

$$MS_{between} = 10/2 = 5$$

Step 5 Determination of sum of squares within the samples: it is the sum of squared deviations of the samples from the sample mean within the samples (SS_{within}). Basically, it expresses the variability occurring in measurement of same sample, as each student measures the different volume of same sample and hence can be expressed as

$$SS_{within} = \sum\left(X_{1i} - \overline{X}_1\right)^2 + \sum\left(X_{2i} - \overline{X}_2\right)^2 + \cdots + \sum\left(X_{ki} - \overline{X}_k\right)^2 \quad (27)$$
$$i = 1, 2, 3, \ldots$$

where $\sum\left(X_{1i} - \overline{X}_1\right)^2$, $\sum\left(X_{2i} - \overline{X}_2\right)^2$ are the deviation of each sample value from same sample mean.

Table 2 Analysis of Variance (ANOVA) example of 5 students measured volume of three samples

	Sample 1	Sample 2	Sample 3
Student 1	14	15	18
Student 2	15	14	15
Student 3	16	15	16
Student 4	15	13	16
Student 5	15	13	15
Mean	$\overline{X}_1 = 15$	$\overline{X}_2 = 14$	$\overline{X}_3 = 16$
Grand mean $\left(\overline{\overline{X}}:\right) = 15$			

$$\mathrm{SS}_{\mathrm{within}} = 2 + 4 + 6 = 12$$

Step 6 Estimation of the mean of the square within the samples: this is obtained by standardizing the deviation obtained within the samples which is obtained by dividing $\mathrm{SS}_{\mathrm{within}}$ by the degrees of freedom within the samples. Symbolically, it can be written as

$$\mathrm{MS}_{\mathrm{within}} = \frac{\mathrm{SS}_{\mathrm{within}}}{(n-k)} \tag{28}$$

where $(n - k)$ is DOF within the samples, n = total number of items in all the sample, and k = number of samples.

$$\mathrm{MS}_{\mathrm{within}} = \frac{12}{(5-3)} = 6$$

Step 7 Determination of the total variance can also be obtained by adding the squares of deviations of individual items from the grand mean and can be written as:

$$\begin{aligned} &\text{SS for total variance} = \sum \left(X_{ij} - \overline{\overline{X}}\right)^2 \quad i = 1, 2, 3, \ldots \\ &\mathrm{SS}_{\mathrm{total}} = 22 \qquad\qquad\qquad\qquad\qquad\quad j = 1, 2, 3, \ldots \end{aligned} \tag{29}$$

This total variance should be equal to the sum of the result of step 3 and step 5 explained above

$$\text{SS for total variance} = \mathrm{SS}_{\mathrm{between}} + \mathrm{SS}_{\mathrm{within}} \tag{30}$$

Here, the degrees of freedom $(n - 1)$ for total variance will be equal to the sum of degrees of freedom for between $(k - 1)$ the samples and degrees of freedom within the sample $(n - k)$

$$(n-1) = (k-1) + (n-k) = 4 \tag{31}$$

Step 8 Finally, F-ratio can be estimated as

$$F \text{ - ratio} = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} = > \frac{5}{6} = 0.8333 \tag{32}$$

The F-ratio helps identify whether there is an epochal difference between the sample means or is just fluctuations due to some human or machine error. To identify it, one can look at F table available [20] for different DOFs at various levels of significance, and accordingly, depending upon the value of F test, the relevant hypothesis is checked to infer the accurate results. For instance, if F test estimated values are less than the table value of F, the difference is negligible, i.e. the null hypothesis of no difference between the sample means becomes effective and considered to be truly accepted. On the other hand, if the value of F is found to be more than its table value, the difference is epochal as and larger than the critical value and consequently null hypothesis cannot be accepted (Table 3).

7. Uncertainty

Measurement of a physical quantity is always associated with the uncertainty, and hence, a quantitative determination is required to check the quality of the products that one can reasonably high level of confidence with the product. With such information, one can compare with the available results associated with the standard reference data. Thus, it required its ready implementation and easily understandable and commonly accepted procedure for characterizing the physical parameters with associated uncertainty [21].

Though analysis of error in measurement has been the common practice in metrology, accurate determination of uncertainty is relatively new in the history of measurement, which is now widely recognized. It is worth mentioning that even after the appropriate corrections of all the known or suspected components of error, uncertainty is persisted.

One may notice that as expressed in GUM [22] (Guide to the expression of uncertainty in measurement), the most accurate method for evaluation and expression of measurement uncertainty of a result must be:

(1) *Universal* the method that can be applied to every type of measurements and its input parameters of measurement.
(2) The uncertainty quantifying quantity must be:
    (1) Internally consistent.
    (2) Transferable [22].

In order to understand the concept of uncertainty in measurement, it is important to know initially the measurand which is the quantity that we intend to measure. Primarily, the aim of a measurement is to obtain the **true value** of the measurand [23]. Many efforts are normally made to optimize the **measurement procedure** in such a way that the measured value can be close to the true value as much possible. However, our measurement results are just an estimate of the true value and the actual true value, which is always unknown to us. Therefore, we cannot

Table 3 The format of the ANOVA table

Source of variation	SS	DF	MS	*F*-ratio
Between sample	$SS_{between}$ (10)	$k - 1$ (2)	$MS_{between}$ (5)	$F = \frac{MS_{between}}{MS_{within}}$ (0.8333)
Within sample	SS_{within} (12)	$n - k$ (2)	MS_{within} (6)	
Total	SS_{total} (22)	$n - 1$ (4)		

reveal exactly how close our measured value is to the true value giving rise an estimate of some associated uncertainty.

Further, it is important to know about the error which is the difference between the measured value and the true value. It can be positive or negative and can be regarded as being composed of two parts, namely random error and systematic error, which will be discussed in more detail in the subsequent part of the article.

The associated **uncertainty** in the measurement of a quantity accounts to the accuracy of the measurement. So, uncertainty can be defined as the dispersion region of the measured value ($C_{MEASURED}$) is the region in which somewhere true value (C_{TRUE}) lies. Half of that dispersion is known as measurement uncertainty U (positive always) [24]. The same can be better visualized as in the schematic diagram (Fig. 7).

The Guide to Uncertainty in Measurement (GUM), published in 1992 by ISO/BIPM, has been the origin of determining the uncertainty in measurement. Primarily, the GUM introduced two categories of the elemental uncertainties, namely Type A and Type B [25]. The GUM imposed that Type A uncertainties can be evaluated by statistical methods while Type B uncertainties can be obtained by some other means, e.g. specifications or manuals of instruments, estimates based on long-term experience, and certificates of reference materials.

The main uncertainty sources of measurement are generally occurred due to repeatability, calibration, and parametric as well as conditional effects. Uncertainty due to the non-ideal repeatability of measurement, often called repeatability uncertainty, is originated due to randomness in the measurement. Repeatability is a typical random

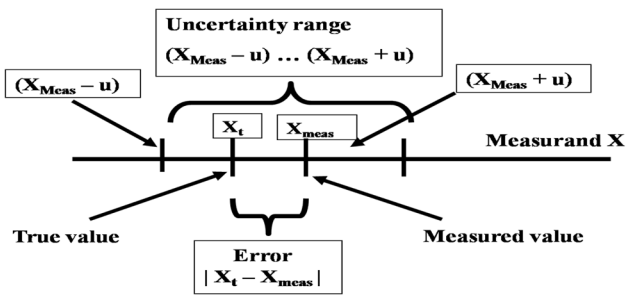

Fig. 7 Brief illustration of interrelations of true value, measured value, error, and associated uncertainty

effect that contributes to uncertainty which further can be decreased by making repeated measurements but cannot be entirely eliminated. This type of uncertainty is considered as type A uncertainty. The uncertainty aroused due to calibration of the equipment, often called calibration uncertainty which is not statistics dependent, can be considered as type B uncertainty. Repeated calibration can reduce the level of Type B uncertainty, but there will always be a residual amount that cannot be determined. Determination of standard uncertainty can be performed as follows

(a) *Type A evaluation of standard uncertainty* As referred in GUM [25], the best available estimate of the expectation or expected value μ of a quantity x that varies randomly for n number of observations taken in the same condition is determined by arithmetic mean or average x of the n observations:

$$x = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{33}$$

Further, the estimate of variance of the n observations σ^2 is given by

$$\sigma^2 = \left[\frac{\left\{\sum_{i=1}^{n}(x_i - x)^2\right\}}{n-1}\right] \tag{34}$$

and the standard uncertainty is

$$u_s = \frac{\sigma}{\sqrt{n}} \tag{35}$$

(b) *Type B evaluation of standard uncertainty* Type B evaluation of standard uncertainty reveals the experience and related skill that can biological through practice [26].

If the estimation x_i that has not been obtained from repeated observations but depends on an independent quantity which is available from the system (X_i), the estimation of the standard uncertainty u (x_i) or variance $\boldsymbol{\sigma}^2(x_i)$ can be done by ones scientific ability and judgment that is based on the availability of information and the variation of X_i. The information may be in the form of already measured data, working experience or general knowledge of the

characteristics and behaviour of given materials along with the properties of instruments, specifications that are provided by the manufacturer, and standard method of calibration.

7.1. The ISO GUM Modelling Approach

The model equation enables calculating the output quantity value (result value) from the input quantity values. Input quantities are directly measured or known quantities. Thus, calculating the output requires designing a model for calculating the measurand (Y) from the directly measured input quantities (X_i) as

$$Y = f(X_1, X_2, X_3 \ldots X_i) \tag{36}$$

The possible input sources for uncertainty can be due to non-representativeness of sampling, sample preparation, weighing, calibration of instrument, and measurement of the sample [27] (Fig. 8).

In order to better understanding, let us take an example of detection of *albumin and creatinine* ***(measurand)*** in human urine for the detection of *Kenny diseases*. As we have seen from the equation that the measurand quantity Y is dependent on input quantities (X_1, X_2, $X_3 \ldots$), the uncertainty will also be adjoint to these parameters together with uncertainty of instruments used for measurement. The following steps can be followed up which better explain the ISO Gum modelling approach

Step 1: Procedure It is utmost important to standardize a procedure for the analysis, which includes constructing the generalized graph from the population. Construction of parametric standards can be based on the fact that whether if the urine will be diluted or not, what will be the size of antibody for detection of creatinine, which electrodes (either platinum or silver) can be used as reference electrode and what will be the process and materials to be used and finally which instrumental technique can be used for evaluation, e.g. cyclic voltammetry or linear sweep voltammetry (Fig. 9).

Step 2: Quantifications and formulation Formulation of analytical formulae and the uncertainty sources occurring in the procedure must be taken into consideration for evaluating the concentration of creatinine and albumin as follows:

$$C_{c,a}\left(\frac{X_i - \text{slope of graph}}{\text{intercepts of graph on measurand axis}} \times F_{\text{d}}\right) + \Delta U_{a,e} \tag{37}$$

where X_i is sample value, F_{d} is urine dilution factor and $\Delta U_{a,e}$ is all the uncertainty factor that comes into consideration by electrode kinetics and mass transport along with the techniques and measurement involving.

Step 3: Uncertainty identification It involves to identify the sources of uncertainty responsible for the uncertainty in the output (Fig. 10).

Step 4: Calibration This includes using of computer software for the calibration of the data and constructing the graphs to get some parametric values for the parameters in standard units of the quantities used as in Eq. 37.

Step 5: Standard uncertainty This quantifies the evaluation of the standard uncertainty in measurement from the all input sources, which already are discussed in Eqs. 34 and 35.

Step 6: Combined standard uncertainty The GUM provides us the method for combined uncertainty measurements of all type of uncertainty included in the measurement and can be determined by combining the standard uncertainties from all the input sources (x_1, x_2,…, x_N) [29]. The combined standard uncertainty $u_{\text{c}}(y)$ is the positive square root of the combined variance $u_{\text{c}}^2(y)$, which is given by

$$u_{\text{c}}(y) = \sqrt{u(x_1)^2 + u(x_2)^2 + u(x_3)^2 + u(x_4)^2} \tag{38}$$

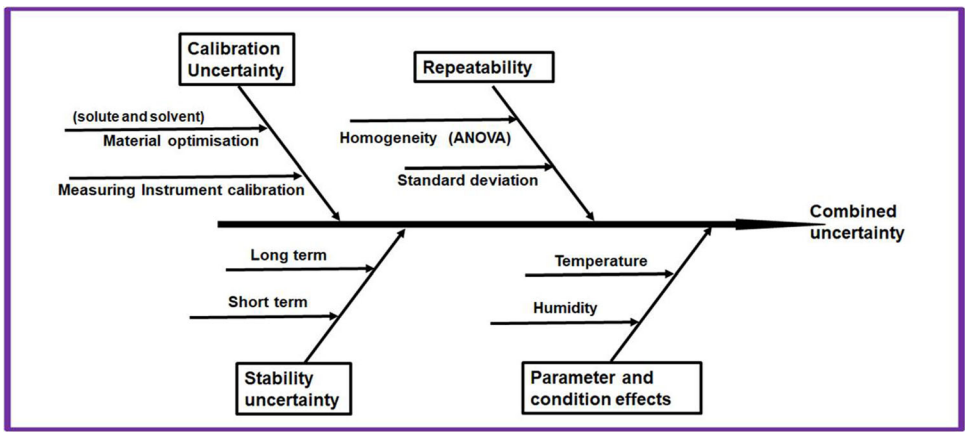

Fig. 8 Schematic diagram showing causes and their effects depicting uncertainty components which is also known as "fishbone diagram" [28]

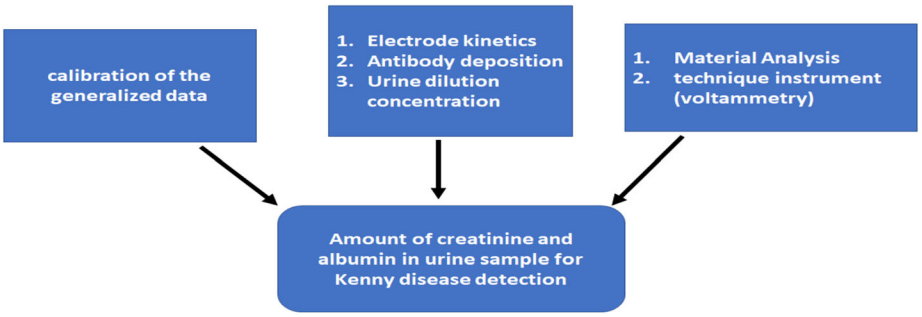

Fig. 9 Step 1 defining the procedure of the model

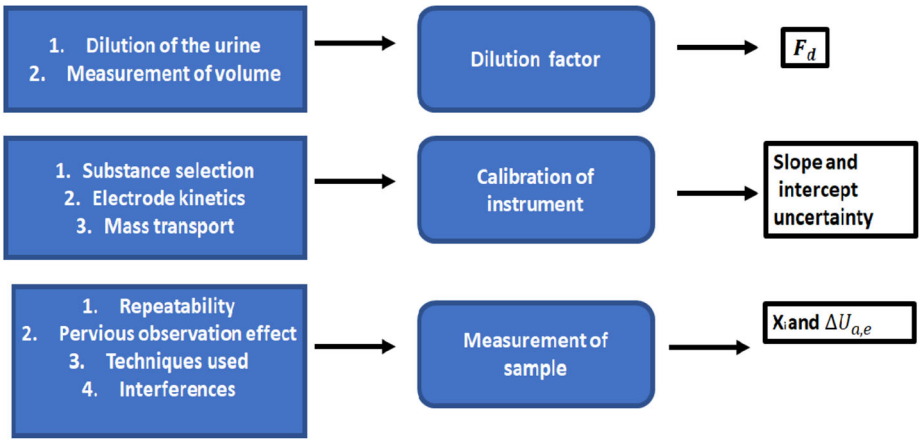

Fig. 10 Identification of the uncertainty sources factors

where $u(x_1)$ = uncertainty due to repeatability, $u(x_2)$ = uncertainty due to calibration, $u(x_3)$ = uncertainty due to input parameters, and $u(x_4)$ = uncertainty due to stability, which will be discussed in the next section.

Step 7: Expanded uncertainty The measurement uncertainty is generally presented to customers as expanded uncertainty, U. It is estimated by multiplying the combined uncertainty with a coverage factor, **k**. The expanded uncertainty can be found at two different levels of sophistication. The simpler approach uses simply a preset coverage factor defining the probability of occurring the results. Very often $k = 2$ which is a case of the normal distribution corresponds to 95% probability.

Thus,

$$\text{Expanded uncertainty}\left(U_{\text{exp}}\right) = k * u_{\text{c}}(y) \tag{39}$$

Further, researchers can construct a pie chart (Fig. 11) for a better representation of the uncertainty sources which will be helpful for the evaluation of the product to the end user.

8. Homogeneity Test and Uncertainty Determination

In order to demonstrate the validity of the certified values and their uncertainties in the analysis of individual units, the homogeneity analysis is of the prime importance for reference materials certification [30]. Inhomogeneity usually occurs within a few exceptions and inherent in the materials itself. Therefore, prior to certification, reference materials are usually subjected to homogeneity testing. The homogeneity test is employed to verify a random number of units provided they are identical while performing testing of between-sample homogeneity. The homogeneity test can be performed by employing ANOVA test described above in Sect. 4 wherein the homogeneity can be envisaged from the comparison of estimated F and critical F value obtained from the table for given degrees of freedom. Prior to further discussion on homogeneity test, one must know briefly about the reference materials.

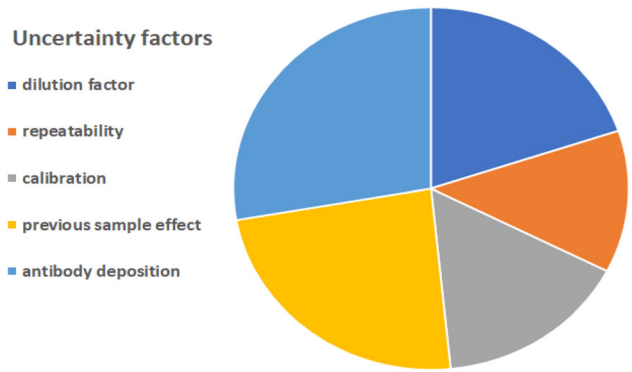

Fig. 11 Uncertainty sources responsible for the overall product uncertainty

Different types of reference materials have been discussed on the basis of International Laboratory Accreditation Cooperation (ILAC) which are (a) *Objective reference materials* These are materials used specifically for detection of given property, e.g. as litmus paper used for testing acidity and basicity (b) *Pure material* Materials used for tracing the purity or impurity of material (c) *Physical–chemical reference materials* These RMs are incorporated with properties such as density, viscosity, electrochemical activities of materials, boiling melting points (d) *Standard solutions and gas mixtures* Such RMs are used for calibration of solution and gas, which requires gravimetrical preparation from pure material (e) *Matrix reference materials* These RMs characterized for the composition of unique major, minor, or trace chemical constituents. As per requirements, the materials can be prepared artificially by making grid (matrices) for the components of our interest.

It is worth mentioning here that the prime aim of all reference materials for analytic purpose is to know the sensitivity and specificity of the analysis for the materials required for particular purpose without worrying about the undesired and unwanted factors which may affect the outcomes of laboratory. All analytical techniques suffer from some shape of influence depending on the awareness of other elements in the sample matrix. The matrix could have a sizeable effect in a manner the testing is performed and better outcomes can be achieved; the effects are known as matrix effects. One can understand it in such a way that as we change the concentration, the antibody performance of a reference electrode changes. To avoid such effects, researchers are usually advised to draw the calibration curve which further increase the importance of matrix reference materials. However, as per categorized by ISO guide 170 30 (1992), A **reference material (RM)** is a material which is necessarily stable and homogeneous for one or more properties throughout the evaluation, so that it can be used as a reference for measurement of required features, while a **certified reference material (CRM)** is categorized by specified metrological procedure (ISO guide 17034) which are reasonable and logical for one or more properties escorted by the certificates (ISO guide 17031) that proves the legality of specified property with associated uncertainty and metrological traceability. One should also note that the properties of reference materials can be anything chemical, physical, and biological. The deficiency of quality-controlled measurement performed by laboratories in evaluations may encounter severe economic losses due to corrupted faulty measurements/records. With reference to manufactured products, inaccurate measurements may cause a wrong evaluation of the high-quality product. Quality of the products has turn out to be a primary function since early stage of twentieth century, and diverse movements are currently undertaken to make sure that the facts produced whether ensures better quality of life for the consumer. The principal steps for the production of the reference material (ISO guide 17035 and ISO Guide 17034 [2006]) are as follows:

- Data collection and material synthesis.
- Sample preparation (that includes between the bottle and within the bottle preparation).
- Testing the material homogeneity and heterogeneity.
- Testing the stability of material.
- Traceability testing.
- Assigning the value with associated uncertainty.

Certified reference materials (CRMs) represent a key device for laboratories to confirm the qualitative accuracy in their measurements within the framework of their internal quality control process. These include QA standards (ISO 9000), accreditation, and participation in proficiency testing. A procedure for certification of reference material is shown below for the better understanding of discussion (Fig. 12).

Further, it is important to know about material homogeneity, a material is recognized as perfectly homogenous in aspect of a given property, if there's no change in assets in belongings from one component to another, while a homogeneous material has no significance change in the specified property of the material. Homogeneity testing gives an idea that there is no significant difference between the samples as compared to the certified uncertainty interval for which the best possible way is to estimate the relative value of each unit. The level of homogeneity can be assessed in terms of the uncertainty range within the given confidence level or coverage interval [31]. It is worth stating that the measured uncertainty (u_{exp}) is commonly overestimated the value of between-sample variations (u_{bb}), which is a combination of analytical measurement (mean_{s}) and the between-sample variation,

$$U_{\text{exp}}^2 = u_{\text{meas}}^2 + u_{\text{bb}}^2 \tag{40}$$

where

$$u_{\text{bb}} = (\text{MS}_{\text{within}}/n)^{1/2} \cdot \left(\frac{2}{\vartheta}\right)^{1/4} \tag{41}$$

ϑ is the DOF of mean square of within the samples and n number of replicates performed on each sample.

It is worth mentioning that if the between-sample variability u_{bb} is insignificant when compared with the certified uncertainty interval and hence can be neglected. Therefore, the uncertainty estimated from the measurement is used to describe CRM uncertainty. On the other hand, when the between-sample variability u_{bb} is significant in the total uncertainty, then in such cases, the batch may either be rejected or regenerated, or each sample must be

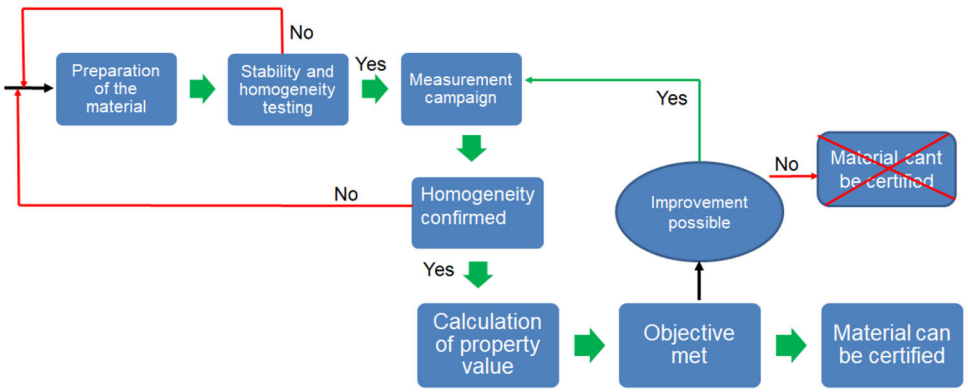

Fig. 12 Procedure for certification of reference material

individually certified. In case of within subsampling scheme, the same general principles can also be applied as it was applied to the between-sample homogeneity testing [32]. Thus, the experimental within-unit uncertainty (u_{exp}) is commonly its overestimated value, which is a combination of analytical measurement (u_{meas}) and the material inhomogeneity (u_{within})

$$U_{\text{exp}}^2 = u_{\text{meas}}^2 + u_{\text{within}}^2 \tag{42}$$

where

$$u_{\text{within}} = (\text{MS}_{\text{within}})^{1/2} \cdot \left(\frac{2}{\vartheta}\right)^{1/4} \tag{43}$$

ϑ is the DOF of mean square of within the samples.

8.1. Stability of Reference Material

Once the homogeneity of a material is confirmed, the next property which is equally important for certification of the reference material is its stability. There are two types of stability: long-term stability (which includes the life of reference material/certified reference material) and short-term stability (which includes transport from one place to other). To test the material stability, one requires to measure the property value and its associated uncertainty which must be always lower than the certified uncertainty of the reference material. Hence, the condition for stability testing of reference material can be expressed as

$$|x_{\text{CRM}} - x_{\text{meas}}| \leq k\left(u_{\text{CRM}}^2 - u_{\text{meas}}^2\right)^{1/2} \tag{44}$$

This signifies that the material is sufficiently stable and the evaluation process of the material is unbiased. Further, there are some uncertainty components which must be taken into consideration for the stability analysis of material (ISO guide 25 and ISO guide 17034, 2006):

- Between the bottle homogeneity.
- Repeatability of measurement.
- Instability of the material.
- Reproducibility.
- Instability of the measurement.

Further, it is important to mention here that in the study of a cyclical stability, the parameters of uncertainty are reduced and hence the associated uncertainty decreases by some amount, and therefore, the total uncertainty of the reference material can only be estimated as

$$U = \left(u_{\text{stab}}^2 + u_r^2 + u_{\text{bb}}^2\right)^{1/2} \tag{45}$$

9. Summary

Various relevant statistical parameters have been discussed in this review article for better readership. We have also stressed the homogeneity testing through various statistically adopted methods and have summarized a very simplified example of gravimetrical measurement of volume for better understanding of homogeneity. In addition to this, a prime focus has also been made on the importance of the ISO guide and GUM modelling approach. Further, for a detailed understanding of researchers we have elaborated the procedure steps of ISO GUM modelling approach in very accurate manner by taking the example of element detection in urine for early detection of Kenny disease. Further, for homogeneity discussion and testing of material, ISO guide (17030& 17035) has been taken into consideration for describing, differentiating, and defining the reference material. The standard procedure for certification of reference material has been also been explained in lucid manner with the help of procedural flow charts along with their stability study and analysis for better readership. This review article primarily emphasizes the relation between basic statistics, homogeneity, and uncertainty in the measurements collectively. We strongly believe that this article will be relevant for accurate and precise measurement for the betterment of the value of products, which in turn may improve the quality of life.

Acknowledgements This review article is evolved under the activity of Bharatiya Nirdeshak Dravya-Indian Reference Material at CSIR-

NPL. The authors are grateful to our head, Dr. R. P. Pant and Director, NPL, for his constant mentoring and support for this project. The authors also acknowledge Dr. S. P. Singh and Sahiba Shaikh from CSIR-NPL and Dr. Chandrashekhar from Amity University Haryana for their useful discussion.

References

[1] S.M. Gore, I.G. Jones and E.C. Rytter, Misuse of statistical methods: a critical assessment of articles in BMJ from January to March 1976, Br. Med. J., **1** (1977) 85–87.
[2] S. Schor and I. Karten, Statistical evaluation of medical journal manuscripts, JAMA, **195** (1966) 1123–1128.
[3] J.S. Kim, D.K. Kim and S.J. Hong, Assessment of errors and misused statistics in dental research, Int. Dent. J., **61** (2011) 163–167.
[4] R.D. MacArthur and G.G. Jackson, An evaluation of the use of statistical methodology in the Journal of Infectious Diseases, J. Infect. Dis., **149** (1984) 349–354.
[5] International Committee of Medical Journal Editors, Uniform requirements for manuscripts submitted to biomedical journals, N. Engl. J. Med., **336** (1997) 309–315.
[6] N.A. Weiss, Introductory statistics, Addison Wesley, Boston (1999).
[7] G.M. Clarke and D. Cooke, A basic course in statistics, 5ed; Arnold, London (1998).
[8] R. Williams, Normal distribution, 1ed; (2004) (www3.nd.edu/~rwilliam/stats1/x21.pdf).
[9] C.R. Kothari, Research methodology methods and techniques, 2ed; New Age International Publishers, New Delhi, (2004) pp. 256–260.
[10] C.R. Kothari, Research methodology methods and techniques, 2ed; New Age International Publishers, New Delhi, (2004) p. 157.
[11] R. Kumar, Research methodology, 3ed; SAGE Publications Ltd, New Delhi (2011).
[12] M.J. de Smith, Statistical analysis handbook, 1ed; Troubador Publishing Ltd., Leicester (2018) p. 132.
[13] http://www.cs.bilkent.edu.tr/~korpe/courses/cs515-fall2002/z-table.jpeg.
[14] J. Isotalo, Basics of statistics, University of Tampere, Finland, (2001) pp. 53–54.
[15] M.J. de Smith, Statistical analysis handbook, 1ed; Troubador Publishing Ltd., Leicester, (2018) p. 423.
[16] M.J. de Smith, Statistical analysis handbook, 1ed; Troubador Publishing Ltd., Leicester, (2018) pp. 443–445.
[17] https://math.stackexchange.com/questions/2009660/basic-question-about-using-the-chi-square-table.
[18] https://www.calvin.edu/~scofield/courses/m143/materials/handouts/anova1And2.pdf.
[19] C.R. Kothari, Research methodology methods and techniques, 2ed.
[20] http://ssl.safaribooksonline.com/getfile?item=ODE3ZzRkLzR0c2NpZ3AvbS85ZTAxOXI3NzU0YWFyYWxfcGlzaGJ0NGMvMDkxYWcuZnQ2aQ.
[21] BIPM JCGM 100:2008, Evaluation of measurement data—guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections), p. 71.
[22] Guide to the expression of uncertainty in measurement—JCGM 100:2008 (GUM 1995 with minor corrections—evaluation of measurement data).
[23] Guide to the expression of uncertainty in measurement, (1995), ISO, ISBN 92-67-10188-9.
[24] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, Evaluation of measurement data—concepts and basic principles, Joint Committee for Guides in Metrology, JCGM 105, (in preparation), p. 2.
[25] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, Evaluation of measurement data—supplement 3 to the "guide to the expression of uncertainty in measurement"—modelling, Joint Committee for Guides in Metrology, JCGM 103, (in preparation) pp. 10–11.
[26] C. Elster, Calculation of uncertainty in the presence of prior knowledge. Metrologia, **44** (2007) p. 111.
[27] Introducing the concept of uncertainty of measurement in testing in association with the application of the standard ISO/IEC 17025 ILAC-G17, p. 5.
[28] EURACHEM/CITAC, Quantifying uncertainty in analytical measurement, 2ed; Technical report guide CG4, EURACHEM/CITEC,[EURACHEM/CITAC Guide], (2000) pp. 101–103.
[29] National Institute of Standards and Technology (NIST), Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST technical note 1297, NIST, Gaithersburg, MD, (1994) pp. 3–6.
[30] Certification of reference materials—general and statistical principles, (1989) ISO Guide 35.
[31] Quality system guidelines for the production of reference materials, (1996) ISO Guide 34.
[32] Certification of reference materials—general and statistical principles (1989) ISO Guide 35, pp. 18–19.