Int. J. Metrol. Qual. Eng. **3**, 169–178 (2012)

DOI: 10.1051/ijmqe/2012029

Confidence intervals and other statistical intervals in metrology

R. Willink*

Received: 17 October 2012 / Accepted: 28 October 2012

Abstract. Typically, a measurement is regarded as being incomplete without a statement of uncertainty being provided with the result. Usually, the corresponding interval of measurement uncertainty will be an evaluated confidence interval, assuming that the classical, frequentist, approach to statistics is adopted. However, there are other types of interval that are potentially relevant, and which might wrongly be called a confidence interval. This paper describes different types of statistical interval and relates these intervals to the task of obtaining a figure of measurement uncertainty. Definitions and examples are given of probability intervals, confidence intervals, prediction intervals and tolerance intervals, all of which feature in classical statistical inference. A description is also given of credible intervals, which arise in Bayesian statistics, and of fiducial intervals. There is also a discussion of the term "coverage interval" that appears in the *International Vocabulary of Metrology* and in the supplements to the *Guide to the Expression of Uncertainty in Measurement*.

1 Introduction

The full analysis of experimental data recognises and accounts for variability. Usually a datum is seen as the outcome of a process with a random element, and a probability distribution, either known or unknown, is subsequently associated with this process. The datum, x, is regarded as the realization or outcome of a "random variable", X, possessing that probability distribution. The technical definition of a random variable is somewhat impenetrable. However, provided that we distinguish between the random variable and its outcome, i.e., the value that it takes, it is sufficient to consider a random variable to be "something about which a probability statement might be made". Thus, the basic, classical, view of a measurement process is that at the beginning potential results can be made the subjects of probability statements but that during the process this randomness is worked through to leave fixed and non-random outcomes, whether known or unknown.

Suppose that we are measuring a quantity with unknown value θ and that our procedure will incur an error drawn from a continuous symmetric distribution with mean and median 0. Then before we make the next measurement we can state that

$$\Pr(X > \theta) = 0.5 \tag{1}$$

where X is the random variable for the next measurement result. Suppose the measurement is made and we obtain the result 23.1 (in some units). Then the number 23.1 is the realization or outcome of the random variable X, and so we will often write $x = 23.1$. It is wise practice to differentiate in notation between a random variable and the value that it takes. As just exemplified, we will use a capital letter like X to indicate a random variable and the corresponding lower-case letter to indicate the realization of the random variable, be it a number or a dummy variable. In this way, if the distribution of X is normal with mean θ and variance σ^2 then we can write

$$\Pr(X \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(z-\theta)^2}{2\sigma^2}\right\} dz. \tag{2}$$

It is important to realise that (1) and (2) are probability statements about the next measurement result X, not about the quantity being measured θ. It is also important to realise that – without adopting the fiducial approach to statistics – we cannot rearrange (1) and replace X by x to obtain the statement

$$\text{"}\Pr(\theta < x) = 0.5\text{"}, \tag{3}$$

in which θ is treated as a random variable. If we wish to make a probability statement whose subject is θ, as in (3), then we must adopt the fiducial approach or a Bayesian approach, as described in Section 6. Until then, we will consider only the classical, frequentist, approach to statistics, which is the paradigm of greatest familiarity.

So the probability statements that are encountered in the classical statistical approach to measurement are about potential estimates of the actual quantity of interest, θ, and not about θ itself. Yet typically we want to obtain an interval in which, in some sense, we can have 95% assurance that θ lies, so that this interval can be taken as an interval of measurement uncertainty. Therefore there is potential for confusion: the subject of a legitimate probability statement is not the entity of interest, and not all of

* Correspondence: robin.willink@gmail.com

the "intervals" that might be constructed using the concept of probability will answer the correct question. In particular, there are at least four different types of classical statistical interval that can be distinguished – as described more fully by Hahn and Meeker [1]. Section 2 describes the first, and the simplest, which is the "probability interval". Sections 3–5 describe the "confidence interval", the "prediction interval" and the "tolerance interval" respectively. Section 6 takes us outside of the classical approach to statistics to describe intervals calculated according to the fiducial and Bayesian approaches. Finally, Section 7 examines the use of the term "coverage interval" by the *International Vocabulary of Metrology* [2] and the first two supplements [3, 4] to the *Guide to the Expression of Uncertainty in Measurement* [5]. All the definitions will be given using customary high probabilities like 0.95.

2 Probability interval

As we have implied, a continuous random variable X possesses a probability distribution function $\Pr(X \leq x)$. The first type of interval that we consider is straightforward.

Definition: A 95% *probability interval* for the random variable X is any interval with non-random limits such that the probability that X lies between these limits is 0.95.

Equivalently, if X is random, a and b are non-random and

$$\Pr(a \leq X \leq b) = 0.95 \tag{4}$$

then $[a, b]$ is a 95% probability interval for X. It is helpful to emphasise the subject of the probability statement, X, by placing it on the left of the mathematical sentence, as in the basic English sentence of "subject verb object". So we might instead write (4) as

$$\Pr(X \in [a, b]) = 0.95.$$

We see that the entity within a probability interval is a random variable. If this random variable represents a measurement result yet to be obtained then the probability interval provides statistical bounds on that measurement result. This interval has no direct connection with the actual value of the quantity measured, e.g. θ in (1) and so it is not an interval that describes measurement uncertainty.

Example 1

The temperature in some environment is intended to be kept at 20°. Every hour the temperature is automatically measured using a process whose statistical properties are well known from previous study. The measurement process is known to give an unbiased estimate of the actual temperature with an error that is drawn from a normal distribution with standard deviation 0.2°. An alarm sounds if the result of measurement lies outside the interval [19.5°, 20.5°]. Suppose the actual temperature is equal to the desired figure, 20°. Is there a large probability that the alarm will sound at the next measurement?

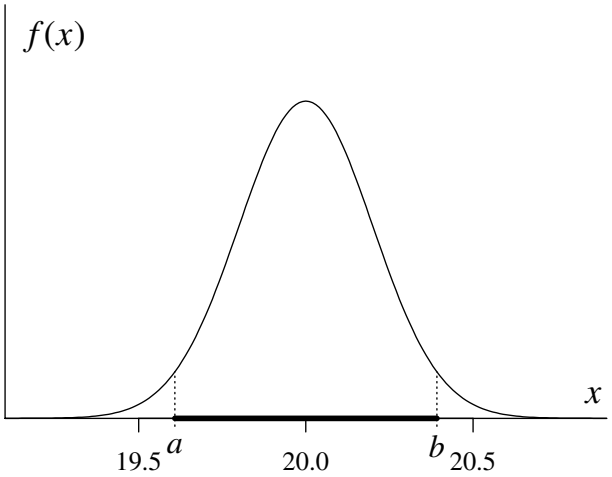

Fig. 1. A 95% probability interval $[a, b]$ for the random variable X with the normal distribution with mean 20 and standard deviation 0.2. $a = 19.61,\ b = 20.39$.

Let X be the random variable for the next measurement result. So, in degrees, X has the normal distribution with mean 20 and standard deviation 0.2, i.e.

$$X \sim \mathrm{N}(0, 0.2^2).$$

The figure 1.96 is the 0.975 quantile of the standard normal distribution, so a 95% probability interval for X is $[20-1.96\times0.2,\ 20+1.96\times0.2] = [19.61,\ 20.39]$. This situation is depicted in Figure 1. The alarm will not sound for any measurement result within this interval, so the probability that an alarm will sound after the next measurement does not exceed $1 - 0.95 = 0.05$.

Example 2

Let us now consider a situation involving many measurements that is represented by the repeated realization of a random variable. Concrete is manufactured in a procedure known to produce blocks with masses following a normal distribution with mean 3200 g and standard deviation 30 g. The mass of each block is measured automatically at the end of the production line using a method with negligible error. A large order is received for blocks with masses no less than 3100 g and no more than 3250 g. What proportion of blocks manufactured will need to be removed to fulfil the order?

In this case, the random variable X corresponds to the mass of a general block yet to be manufactured. Here $X \sim N(3200, 30^2)$, so

$$\Pr(3100 \leq X \leq 3250) = 0.952$$

which means that the interval $[3100, 3250]$ is a 95.2% probability interval for the mass of a manufactured block. Therefore, 4.8% of the blocks will need to be removed to fulfil the order.

In this example, the randomness resides in the generation of the actual values of the quantities being measured, not in the measurement process. So while this kind of situation might be relevant to many industrial practices, it does not correspond to the concept of scientific measurement emphasised in this paper, where in any well-defined measurement the measurand has a unique value to be estimated. The rest of the paper will involve the idea of a fixed unique value of θ.

Comments

We have discussed a probability interval first because it is the type of interval that immediately arises from the notion of a continuous probability distribution. Perhaps because of this immediacy, someone might think of a probability interval when the term "confidence interval" is encountered. However, the idea behind a confidence interval is quite different, as shall be seen in the next section. Similarly, the ideas behind prediction intervals and tolerance intervals also differ substantially from the idea of a probability interval. In particular, confidence intervals, prediction intervals and tolerance intervals are intervals with random limits, whereas a probability interval has fixed limits, e.g. a and b or 19.61 and 20.39 in Example 1.

In short, we may describe a probability interval as a fixed interval with a random subject. Its role is to make statistical inference about the future outcome of this random variable.

3 Confidence interval

Usually, measurement can be understood as a process of estimating or approximating an unknown actual value, whether it be called a "true value" or "target value". This is reflected in the *Guide to the Expression of Uncertainty in Measurement* which states that "The measurand should be defined... so that for all practical purposes associated with the measurement its value is unique." [5, Sect. 3.1.3]. The relevant field of statistics is that of parameter estimation, where a data-generating process is deemed to be governed by one or more unknown fixed quantities, called parameters, and where our attention is on estimating (the value of) one of these parameters. In a measurement situation, the quantity that is being measured, θ, affects the distribution of potential measurement results, so estimating this quantity means estimating a parameter of that distribution.

A *point estimate* of θ is a single number, say the mean of n measurement results. An *interval estimate* of θ is an interval, say $[x_{\mathrm{L}}, x_{\mathrm{H}}]$, about which we have a high level of assurance that it contains θ. The idea of a "confidence interval" relates to the calculation of an interval estimate of θ. Because of unpredictable influences in the measurement process, the limits of an interval estimate of θ would be different if the experiment were carried out a second time. The limits x_{L} and x_{H} are therefore the outcomes of random variables, which we can call X_{L} and X_{H}. The probability statement underlying the idea of the confidence interval involves these random variables X_{L} and X_{H}, not their outcomes x_{L} and x_{H}. If these random variables X_{L} and X_{H} are distributed such that

$$\Pr(X_{\mathrm{L}} < \theta \text{ and } X_{\mathrm{H}} > \theta) = 0.95,$$

which is

$$\Pr([X_{\mathrm{L}},\ X_{\mathrm{H}}] \ni \theta) = 0.95,$$

then the interval with random limits X_{L} and X_{H} is called a 95% *confidence interval* for θ. The interval $[x_{\mathrm{L}}, x_{\mathrm{H}}]$, which is formed from the experimental observations, is to be seen as the realization of this confidence interval, not the confidence interval itself. We thus can make the following definition.

Definition: A 95% *confidence interval* for an unknown constant θ is a random interval $[X_{\mathrm{L}}, X_{\mathrm{H}}]$ with probability 0.95 of covering θ.

The lower limit of the confidence interval might be $-\infty$ or the upper limit might be $+\infty$, but often both limits of the interval will be random variables, so that the interval can be represented as $[X_{\mathrm{L}}, X_{\mathrm{H}}]$, as in our basic definition. The experiment is carried out and the observations are made. The random limits X_{L} and X_{H} take their realized values x_{L} and x_{H} and we form the known numerical interval $[x_{\mathrm{L}}, x_{\mathrm{H}}]$. This known interval is the outcome or realization of the confidence interval.

Regrettably, authoritative sources give two different definitions of a confidence interval. The *International Dictionary of Statistics* [6] considers the confidence interval to be the random interval $[X_{\mathrm{L}}, X_{\mathrm{H}}]$, as above, but the *Encyclopedia of Statistical Sciences* [7] takes the confidence interval to be the numerical interval $[x_{\mathrm{L}}, x_{\mathrm{H}}]$. In the same way, some statistical books take a confidence interval to be random, e.g. [8, 9], and others take it to be numerical, e.g. [10, 11]. There is some merit in each of these definitions, but no merit whatsoever in the existence of two different definitions! Perhaps much of the misunderstanding in applied science about the idea of confidence interval is related to this ambiguity. We prefer the first definition, where the confidence interval is the random interval, not the numerical interval. This preserves and emphasises the important concept of a random interval with a specified probability of enclosing a fixed target point. This also means that a participle such as "realized", "calculated" or "evaluated" is required when referring to the numerical interval.

So a 95% confidence interval is a random interval that has probability 0.95 of enclosing a constant. The merit of this idea lies in the fact that if a 95% confidence interval is calculated in every measurement problem then, in the long-run, 95% of the intervals obtained will contain the actual values of the measurands. Thus, unless there is relevant information that we have failed to take into account in our particular situation, such as a physical bound on θ, we can be 95% assured that θ lies in the numerical interval obtained.

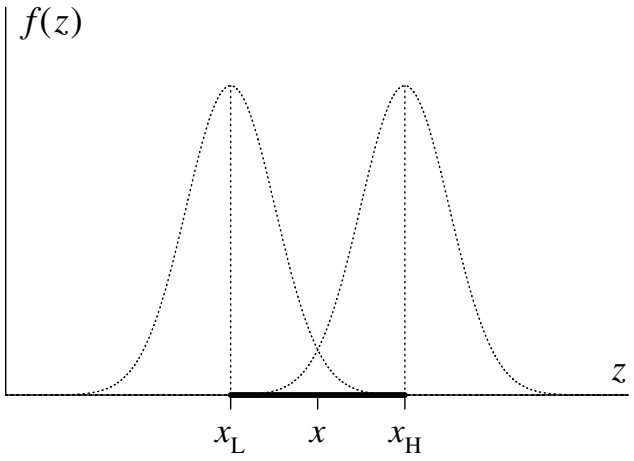

Fig. 2. An evaluated 95% confidence interval $[x_{\mathrm{L}}, x_{\mathrm{H}}] = [x - 1.96\sigma,\ x + 1.96\sigma]$ for θ when the measurement result x is drawn from a normal distribution with mean θ and known standard deviation σ.

Example 3: Known error variance

Our first example of a confidence interval is given for the simple situation where a quantity θ is measured once using an unbiased method that incurs a normally distributed error with known variance σ^2. The random variable for the measurement result X has the normal distribution with mean θ and variance σ^2, so

$$\Pr(X > \theta - 1.96\sigma \ \text{ and } \ X < \theta + 1.96\sigma) = 0.95.$$

Thus

$$\Pr(X + 1.96\sigma > \theta \ \text{ and } \ \theta > X - 1.96\sigma) = 0.95,$$

which means that

$$\Pr([X - 1.96\sigma,\ X + 1.96\sigma] \ni \theta) = 0.95.$$

So the random interval $[X - 1.96\sigma,\ X + 1.96\sigma]$ is a 95% confidence interval for θ. If x is the numerical measurement result then the interval

$$[x - 1.96\sigma,\ x + 1.96\sigma] \tag{5}$$

is the evaluated 95% confidence interval for θ.

The simplicity of this example enables us to illustrate the calculation of the numerical interval in a different way. Figure 2 shows this interval $[x - 1.96\sigma,\ x + 1.96\sigma]$ and shows the distributions of X that would be applicable if θ were equal to these limits. The area under the left-hand distribution to the right of x is 0.025 and the area under the right-hand distribution to the left of x is 0.025. The confidence interval procedure is thus obtaining potential values for θ beyond which the observation x is deemed too unlikely to have occurred. So the numerical interval $[x_{\mathrm{L}}, x_{\mathrm{H}}]$ is seen to connect the parameters of two different distributions, in contrast to the idea that the probability interval connects two different quantiles of the same distribution.

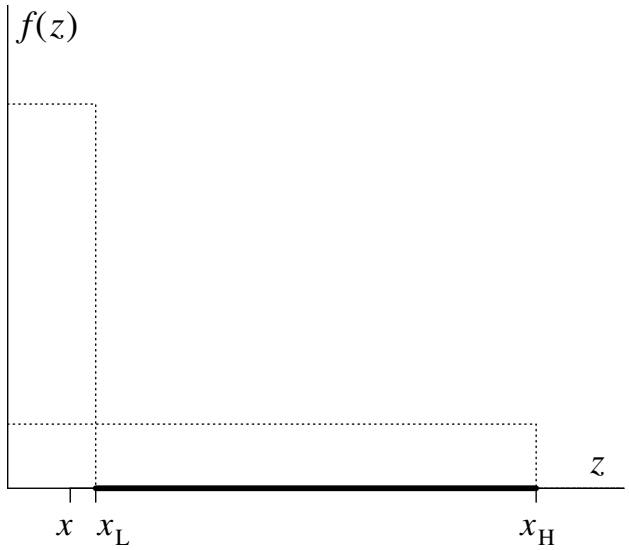

Fig. 3. An evaluated 95% confidence interval $[x_{\mathrm{L}},\ x_{\mathrm{H}}]$ for θ when the measurement result x is drawn from a uniform distribution with lower limit 0 and upper limit θ. (The figure is not drawn to scale.)

Example 4: Uniform distribution with one limit known

Our second example of a confidence interval is rather artificial, but it does serve to broaden understanding of the essential concept. Consider the measurement of a quantity θ using a technique that returns a value drawn randomly from the interval between 0 and θ. A single measurement is made, the result is x, and we wish to construct an interval estimate of θ with "95% reliability".

The result x is seen as the outcome of a random variable X distributed uniformly on the interval $[0, \theta]$. The random variable X has probability 0.95 of lying in the interval $[0.025\,\theta, 0.975\,\theta]$, i.e.

$$\Pr(0.025\,\theta < X \ \text{ and } \ X < 0.975\,\theta) = 0.95.$$

Thus

$$\Pr(X/0.025 > \theta \ \text{ and } \ \theta > X/0.975) = 0.95,$$

which means that

$$\Pr([40X/39,\ 40X] \ni \theta) = 0.95.$$

So the random interval $[40X/39,\ 40X]$ is a 95% confidence interval for θ. The evaluated 95% confidence interval for θ is $[40x/39,\ 40x]$.

Figure 3 shows the datum x, the uniform distribution for the smallest potential value of θ that would admit the statement $\Pr(X > \theta) = 0.025$, which is x_{L}, and the uniform distribution for the largest potential value of θ that would admit the statement $\Pr(X < \theta) = 0.025$, which is x_{H}. These extreme values for θ are the limits of the evaluated confidence interval. As in Figure 2, the interval is seen to lie between corresponding parameter values of two different probability distributions.

Example 5: Unknown error variance

The confidence interval that is most relevant in metrology is the confidence interval for a quantity constructed from a random sample of a fixed size n. This the archetypal situation in "Type A evaluation" of measurement uncertainty, i.e. evaluation by statistical means [5]. The data represent measurement results of n repeated measurements of θ that are assumed to incur independent errors from a normal distribution with mean zero but unknown variance. So the measurement results $x_1, \ldots, x_n$ are assumed to be a sample drawn randomly from a normal distribution with unknown mean θ and unknown variance σ^2, and we wish to obtain an interval estimate of θ that we can see as being 95% reliable in containing θ.

The number x_i is seen as the realization of a random variable X_i having the distribution $\mathrm{N}(\theta, \sigma^2)$. Let us define the familiar random variables

$$\bar{X} \equiv \frac{\sum_{i=1}^{n} X_i}{n}$$

and

$$S^2 \equiv \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}.$$

It follows from standard statistical theory that the random variable

$$T \equiv \frac{\bar{X} - \theta}{S/\sqrt{n}}$$

has Student's t-distribution with $n-1$ degrees of freedom, and so that

$$\Pr\left(-t_{n-1,0.975} < T < t_{n-1,0.975}\right) = 0.95$$

where $t_{n-1,0.975}$ is the 0.975 quantile of that distribution. (Thus, $[-t_{n-1,0.975}, t_{n-1,0.975}]$ is a 95% probability interval for T.) For simplicity, let us simply write t for $t_{n-1,0.975}$. Then

$$\Pr\left(-t < T \quad \text{and} \quad T < t\right) = 0.95$$

which means that

$$\Pr\left(\bar{X} + tS/\sqrt{n} > \theta \ \text{ and } \ \bar{X} - tS/\sqrt{n} < \theta\right) = 0.95.$$

That is,

$$\Pr\left([\bar{X} - tS/\sqrt{n},\ \bar{X} + tS/\sqrt{n}] \ni \theta\right) = 0.95,$$

and the random interval $[\bar{X} - tS/\sqrt{n},\ \bar{X} + tS/\sqrt{n}]$ is seen to be a 95% confidence interval for θ. The figures

$$\bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$s^2 \equiv \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

are the realizations of $\bar{X}$ and S^2, and the interval $[\bar{x} - ts/\sqrt{n},\ \bar{x} + ts/\sqrt{n}]$ is the realization of this confidence interval.

Example 6: Linear regression analysis

A common situation in calibration is where an unknown relationship $\tilde{y} = f(x)$ exists between a stimulus x and a response $\tilde{y}$, and where this function is known to be approximately linear. Values $x_1, \ldots, x_n$ are chosen for the stimulus and we prepare to measure the underlying response of the system $\tilde{y}_1, \ldots, \tilde{y}_n$. Because of the presence of error, we will not observe the figure $\tilde{y}_i$ but will instead obtain the figure

$$y_i = \tilde{y}_i + e_i$$

where each e_i is an error regarded as being independently drawn from a normal distribution with mean 0 and some unknown variance σ^2. So we think of y_i as being the outcome of a random variable Y_i having the normal distribution with mean $\tilde{y}_i$ and variance σ^2. Suppose that one of the purposes of carrying out this analysis is to estimate the value of the function $f(x)$ at another x-value, say x_0. Thus, the quantity of interest is the constant $\theta = f(x_0)$.

Let us define the constant $\bar{x} \equiv \sum x_i/n$ and the random variables

$$\bar{Y} \equiv \frac{\sum Y_i}{n}$$

$$B \equiv \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2}$$

$$A \equiv \bar{Y} - B\bar{x}$$

$$S^2 \equiv \frac{\sum (Y_i - A - Bx_i)^2}{n-2},$$

where summation is from $i = 1$ to $i = n$. An approximate 95% confidence interval for $f(x_0)$ is the interval with limits [12, p. 176]

$$A + Bx_0 \pm t_{n-2,0.975} \frac{S}{\sqrt{n}} \sqrt{1 + \frac{n(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \qquad (6)$$

Evaluating this would give intervals that contain $f(x_0)$ on approximately 95% of occasions, hence the name "approximate 95% confidence interval". If the unknown function were truly linear and the distribution of errors truly normal then the interval would be exact.

The evaluated confidence interval for $f(x_0)$ is the interval with limits

$$a + bx_0 \pm t_{n-2,0.975} \frac{s}{\sqrt{n}} \sqrt{1 + \frac{n(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}},$$

where a, b and s are defined in the same way as A, B and S but using the y_i values. Figure 4 shows this interval for a certain set of data with $n = 7$ and for a certain choice of x_0.

Comments

In summary, we may say that a confidence interval is a random interval used to put statistical bounds on a nonrandom quantity. Accordingly, a 95% confidence interval

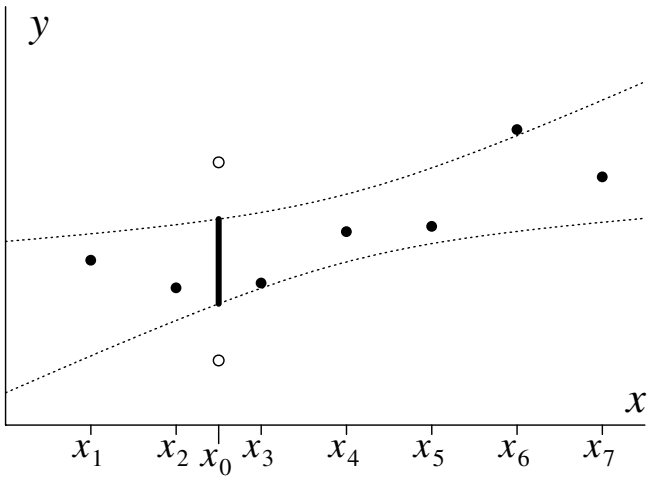

Fig. 4. A set of data $\{x_i, y_i\}$ for $i = 1, \ldots, 7$ obtained when studying the function $f(x)$, and an evaluated 95% confidence interval for $f(x_0)$. The interval lies between dotted curves that show the functions $a + bx \pm t_{n-2,0.975}(s/\sqrt{n}) \times \sqrt{1 + n(x - \bar{x})^2 / \sum(x_i - \bar{x})^2}$. (The hollow markers give the limits of the evaluated prediction interval described in Example 8.)

for θ is an interval with one or two random limits such that there is probability 0.95 that the interval encloses θ. This idea of a random interval and a fixed target contrasts with the idea of a fixed interval and a random target, which is the idea of the probability interval.

Some authors define a 95% confidence interval as a random interval with probability *at least* 0.95 of covering the true value, so that they would write "0.95 or more" in our definition above, e.g. [9]. This is entirely sensible, because 0.95 represents a high figure used as a threshold in the process of decision making, and the same decision would also be made if the actual probability was greater than 0.95.

4 Prediction interval

We now consider a little-known type of interval called a prediction interval. The term "prediction" often carries with it the connotation of the future, so – like the probability interval – this interval is about predicting the outcome of some random variable. In particular, this interval is about examining random variables that are relevant now in order to predict the outcome of a random variable that will be relevant later. And – like the confidence interval – it is a random interval.

Definition: A 95% *prediction interval* for a random variable X is a random interval $[X_L, X_H]$ with probability 0.95 of covering the value that will be taken by X.

Equivalently, with regard to the joint distribution of all three random variables X_L, X_H and X, the random interval $[X_L, X_H]$ is a 95% prediction interval for X if

$$\Pr(X_L < X < X_H) = 0.95.$$

In its simplest form, a prediction interval arises when something is to be measured n times and the results are to be used to place predictive bounds on the result of a further measurement. Our next example is of this form.

Example 7: Predicting a future sample element

The random variables $X_1, \ldots, X_n$ will be observed and the value to be taken by X_{n+1} is to be predicted. Consider the elementary case where $n = 1$. It is not difficult to see that $\Pr(X_1 < X_2 < \infty) = 0.5$ so we can say that the random interval $[X_1, \infty)$ is a 50% prediction interval for X_2. The first measurement is made and the result is x_1. The numerical interval $[x_1, \infty)$ is therefore the realization of a 50% prediction interval for X_2.

More generally, from a consideration of symmetry, it is apparent that

$$\Pr\left(X_{n+1} < \min\{X_1, \ldots, X_n\}\right) = 1/(n+1)$$

and

$$\Pr\left(X_{n+1} > \max\{X_1, \ldots, X_n\}\right) = 1/(n+1).$$

So if $X_{\max}$ and $X_{\min}$ denote the random variables for the maximum and minimum in 39 measurement results then

$$\Pr\left(X_{\min} \le X_{40} \le X_{\max}\right) = 0.95.$$

Therefore, the random interval $[X_{\min}, X_{\max}]$ is a 95% prediction interval for X_{40} and, by implication, for any particular future measurement result.

The prediction intervals just described may also be called "distribution-free" or "non-parametric" because they are constructed without making any assumptions about the parent probability distribution of the data (except that it is continuous). In contrast, the next example describes a prediction interval that involves an assumption of distributional form. The assumption is that the measurement errors are drawn from a single normal distribution.

Example 8: Linear regression analysis (continued)

Consider again the linear regression situation of Example 6, and suppose that the underlying function is exactly linear. Suppose also that, instead of estimating $f(x_0)$, we wish to predict the result in a measurement when the stimulus is x_0. So we now wish to predict the value that will be taken by a random variable Y_0 having a normal distribution with mean $f(x_0)$ and variance σ^2. It can be shown that there is probability 0.95 that the random interval with limits [13, p. 36] [14, p. 455]

$$A + Bx_0 \pm t_{n-2,0.975} \frac{S}{\sqrt{n}} \sqrt{1 + n + \frac{n(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

will cover the value taken by this random variable. So this random interval is a 95% prediction interval for Y_0,

a potential measurement result at $x = x_0$. The numerical interval with limits

$$a + bx_0 \ \pm \ t_{n-2,0.975}\frac{s}{\sqrt{n}}\sqrt{1+n+\frac{n(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

is the realized or evaluated prediction interval.

Comments

In short, the prediction interval is a random interval with a random subject. Like a confidence interval, it is a random interval; one or both of its limits is a random variable. As with a confidence interval, confusion can arise from using the unqualified term "prediction interval" to refer to the numerical interval instead of the random interval. This can be avoided by using an adjective like "realized".

The difference between a prediction interval and a confidence interval is the nature of the subject. Like a probability interval, a prediction interval has a random subject; it is a tool for making inference about the outcome of random variable. In contrast, the subject of a confidence interval is a non-random quantity. Pfanzagl [6] writes helpfully "hence prediction intervals are subsets of the sample space whereas confidence intervals are subsets of the parameter space."

5 Tolerance interval

The last of the classical intervals that we consider is the "tolerance interval" or "statistical tolerance interval", as it might be known in engineering contexts. Like the confidence interval and prediction interval, the tolerance interval is a random interval. Consequently, the outcome of this interval should be called a realized tolerance interval or evaluated tolerance interval.

The idea of a tolerance interval can be introduced by placing it alongside a prediction interval. The prediction interval takes as its subject the potential outcome of a random variable, X. In contrast, the tolerance interval takes as it subject the distribution of potential outcomes of the random variable, which is represented by the distribution function $F(x) \equiv \Pr(X \le x)$. So with a tolerance interval, the relevant probability statement is a statement about $F(x)$, not directly about X.

Definition: A 95%-content tolerance interval for a random variable X with confidence coefficient 0.99 is a random interval $[X_\mathrm{L}, X_\mathrm{H}]$ that has probability 0.99 of covering at least 95% of (the probability content of) the distribution of X.

Equivalently, if X_L and X_H are random variables distributed such that

$$\Pr\left(\int_{X_\mathrm{L}}^{X_\mathrm{H}} f(x)\,dx \ge 0.95\right) = 0.99$$

where $f(x)$ is the density function of a random variable X then the interval with random limits X_L and X_H is a 95%-content tolerance interval for X with confidence level 0.99 [6]. This relationship can also be written as

$$\Pr\{F(X_\mathrm{H}) - F(X_\mathrm{L}) \ge 0.95\} = 0.99,$$

where $F(x) = \int_{-\infty}^{x} f(z)\,dz$, which shows that the probability statement is a statement about the distribution function $F(x)$. In the absence of external information, we can be 99% sure that at least 95% of potential measurement results will lie in the realized interval $[x_\mathrm{L}, x_\mathrm{H}]$.

Example 9: A normal distribution

Suppose we wish to study the distribution of the potential results of a measurement and that this distribution can be assumed to be normal. Let the sample size n be predetermined and let the random variables $\bar{X}$ and S^2 be defined as in Example 5. Set

$$k = 1.96 \times \sqrt{\frac{(n^2-1)/n}{\chi^2_{0.01,\ n-1}}}$$

where $\chi^2_{0.01,\ n-1}$ indicates the first percentile of the chi-square distribution with $n-1$ degrees of freedom. Then the random interval with limits $X_\mathrm{L} = \bar{X} - kS$ and $X_\mathrm{H} = \bar{X} + kS$ has probability approximately 0.99 of covering 95% of the unknown normal distribution [15, 16]. So the random interval $[\bar{X} - kS,\ \bar{X} + kS]$ is a 95% tolerance interval for a future measurement result with level of confidence approximately 99%.

The measurements are made and $\bar{x}$ and s^2 are the observed values of $\bar{X}$ and S^2. So $[\bar{x} - ks,\ \bar{x} + ks]$ is the realization of a 95% tolerance interval for a future measurement result with level of confidence approximately 99%. Unless there is additional information that casts doubt on the suitability of this specific numerical interval, we can be approximately 99% sure that it covers 95% of potential measurement results.

Example 10: Uniform distribution

Let X_max and X_min denote the random variables for the maximum and minimum observations when n elements are drawn independently from a continuous uniform distribution with unknown limits. The probability distribution of $F(X_\mathrm{max}) - F(X_\mathrm{min})$ is the beta distribution with parameters $n-1$ and 2 [17, Eq. (2.3.4)], from which we can show that if $n = 50$ there is probability 0.99 that the interval $[X_\mathrm{min},\ X_\mathrm{max}]$ covers at least 87.4% of the uniform distribution. Therefore, if $n = 50$ the random interval $[X_\mathrm{min},\ X_\mathrm{max}]$ is a 0.874-content tolerance interval with confidence coefficient 0.99. That is, if there is a long series of experiments of this type each involving the calculation of an interval $[x_\mathrm{min},\ x_\mathrm{max}]$, those intervals will contain at least 87.4% of the uniform distribution on 99% of occasions.

Comments

Like a confidence interval and a prediction interval, a tolerance interval is a random interval. Consequently, a tolerance interval should not be confused with the realization of a tolerance interval, which is a numerical interval. The subject of a tolerance interval is the distribution fonction of a random variable, not the outcome of a random variable. In this way, it is similar to a confidence interval, which has an unchanging subject.

The 0.95-content tolerance interval with confidence coefficient 0.99 that we have described here can be distinguished from a "0.95-expectation tolerance interval" $[X_{\mathrm{L}}^*, X_{\mathrm{H}}^*]$, which is an interval satisfying

$$\mathcal{E}\left(\int_{X_{\mathrm{L}}^*}^{X_{\mathrm{H}}^*} f(x)\,dx\right) = 0.95$$

where $\mathcal{E}(\cdot)$ denotes the expected value [18].

This brings us to the end of our presentation of intervals in classical statistics. We see that the only type of interval that has the measurand θ as its subject is the confidence interval, so that the confidence interval is the type of interval that is directly relevant to the statement of uncertainty in an individual measurement. In contrast, the probability interval, predictive interval and tolerance interval are focused on the spread of measurement results, and so these intervals are more associated with the characterisation of a measurement process or technique.

6 Non-classical intervals

The preceding material has described four types of statistical interval that arise under the classical, i.e. frequentist, view of statistics. We now turn our attention to intervals arising in two other approaches to statistics, namely the fiducial approach and the Bayesian approach; see e.g. [19].

The essential idea shared by the fiducial and Bayesian approaches to statistics is the idea that direct statements of probability are made about an unknown constant being studied, such as the quantity of interest in a measurement, θ. Thus, a statement of the form "$\Pr(\theta > 10) = 0.54$" can be deemed meaningful in fiducial or Bayesian inference. In fiducial inference this would be a statement of "fiducial probability" and in Bayesian inference this would be a statement of "strength of belief".

The idea that a constant such as θ can be the subject of a probability statement like (3) means that it can also be considered to have a probability distribution. So now θ is regarded as a random variable. The idea that θ has a probability distribution naturally leads to the idea of a interval in which θ is said to lie with 0.95 probability. In the fiducial case this is a "95% fiducial interval for θ" while in the Bayesian case this is called a "95% credible interval for θ". (Occasionally, a credible interval might be called a Bayesian interval or a Bayesian confidence interval.) It can be seen that fiducial intervals and credible intervals are both probability intervals in their own contexts.

The fiducial approach has only been developed for a subset of problems [20]. It has been controversial and, currently, it is little used. Put simply, the fiducial argument allows a probability distribution for θ to be constructed using only the observation x and the probability distribution of the corresponding random variable X. The relationship (3) becomes a consequence of (1) provided that the probability in (3) is understood to be "fiducial". Similarly, the probability distribution formed for θ is known as a fiducial distribution. For example, if a measurement result x is taken to be the outcome of a normal random variable X with mean equal to the unknown value of the measurand θ and with known standard deviation σ then the fiducial distribution for θ becomes the normal distribution with mean x and standard deviation σ, and so a 95% fiducial interval for θ is the interval $[x - 1.96\sigma, x + 1.96\sigma]$. The equivalence of this interval with the realized confidence interval for θ given by (5) hides the fact that a controversial and unaccepted idea is behind this claim.

Bayesian statistics offers an alternative paradigm that, theoretically, is complete in scope. This approach, and in particular the "objective Bayesian" approach, is also controversial. For the Bayesian statistician, all unknown fixed or unrealized quantities are attributed probability distributions that describe someone's belief about them [21]. These distributions are updated on receipt of new data, so that a prior distribution for a quantity, say θ, becomes a posterior distribution after the measurement results are processed. At all times, the Bayesian statistician claims to be able to construct a meaningful probability distribution for θ, and therefore an interval within which θ is said to lie with 95% probability. Perhaps because the idea of belief is at the heart of the Bayesian understanding of probability, such an interval is called a 95% "credible interval" for θ.

The concept of a credible interval does not only apply to parameters like θ that a frequentist would estimate using a confidence interval. To the Bayesian statistician, there are only two sorts of entity, those that are known and those that are unknown [21], and Bayesian inference involves forming probability distributions for all unknowns that are relevant. So the term "credible interval" is equally applicable if the subject is the potential result of the next measurement instead of the existing constant θ.

7 On a "coverage interval"

The first and second supplements to the *Guide to Expression of Uncertainty in Measurement* [3,4] describe an approach to the evaluation of measurement uncertainty that is broadly consistent with a Bayesian analysis. They advocate that the resulting interval of uncertainty be called a "coverage interval" and that the probability attributed to the idea that the measurand of lies within that interval be called a "coverage probability". The first supplement also notes that "a coverage interval is sometimes known as a credible interval or a Bayesian interval" [3, 3.12].

The term "coverage probability" is, however, also found in frequentist statistics, where it is often used to

describe the actual probability that a confidence interval covers the target value θ. For example, consider the measurement of the quantity $\theta = c_1\theta_1 + c_2\theta_2$ where θ_1 and θ_2 are two quantities that are themselves estimated by averaging n_1 and n_2 measurement results respectively, with these results being regarded as drawn from normal distributions with means θ_1 and θ_2 and unknown variances. By the Welch-Satterthwaite approximation [5], the random variable

$$\frac{c_1\bar{X}_1 + c_2\bar{X}_2 - \theta}{\sqrt{\left(S_1^2/n_1 + S_2^2/n_2\right)}}$$

has approximately Student's t-distribution with

$$M = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{S_1^4/\left\{n_1^2\left(n_1-1\right)\right\} + S_2^4/\left\{n_2^2\left(n_2-1\right)\right\}}$$

degrees of freedom (with M seen to be a random variable). Therefore an approximate 95% confidence interval for θ is the random interval with limits

$$c_1\bar{X}_1 + c_2\bar{X}_2 \pm t_{M,0.975}\sqrt{\left(S_1^2/n_1 + S_2^2/n_2\right)}\,.$$

Experimentation shows that this interval encloses θ with probability approximately 0.95 over the bulk of the parameter space. For example, this probability is approximately 0.952 when the two unknown variances are equal and $n_1 = n_2 = 8$ [22, table 3]. Many frequentist statisticians would then say "the coverage probability of this interval in that situation is 0.952". (An example of this usage of this term can be found in an article of Dawid [23, p. 233], who discusses a type of prior distribution in Bayesian statistics. This suggests that the Bayesian community might also have that understanding of the term.)

Therefore, the use of the term "coverage probability" in an analysis that uses and promotes the Bayesian view of statistics is a potential source of confusion. For this reason, if an analysis of measurement uncertainty is carried out using a Bayesian approach the term "credible interval" should be preferred to the term "coverage interval".

8 Conclusion

This paper has distinguished four types of interval that are found in classical statistics and also briefly described intervals that feature in fiducial and Bayesian statistics. Our concluding comments about these intervals are given by way of a summary.

Let a and b be known constants and θ be an unknown constant of interest. Also let X, X_1 and X_2 be random variables and let x_1 and x_2 be the values taken by X_1 and X_2, i.e. the observations of those random variables. Let $f(x)$ be the probability density function of X. Suppose we take the classical approach to statistical inference, where θ is not seen as a random variable.

- If $\Pr(a \le X \le b) = 0.95$ then $[a, b]$ is called a 95% probability interval for X.
- If $\Pr(X_1 \le \theta \le X_2) = 0.95$ then $[X_1, X_2]$ is called a 95% confidence interval for θ.
- If $\Pr(X_1 \le X \le X_2) = 0.95$ then $[X_1, X_2]$ is called a 95% prediction interval for X.
- If $\Pr\left(\int_{X_1}^{X_2} f(x)\,dx \ge 0.95\right) = 0.99$ then $[X_1, X_2]$ is called a 95%-content tolerance interval for X with confidence coefficient 0.99.

The confidence interval is the type of interval that is most relevant in the evaluation of measurement uncertainty because it focuses on the unknown value of the measurand. The other intervals are of greater relevance when the task is instead to characterise the measurement technique.

In the fiducial and Bayesian approaches to inference a probability distribution for θ is obtained, so that θ is treated as a random variable with some probability distribution $F_\theta(x) = \Pr(\theta \le x)$.

- If the analysis is fiducial and if $\Pr(a \le \theta \le b) = 0.95$ then $[a, b]$ is called a 95% fiducial interval for θ.
- If the analysis is Bayesian and if $\Pr(a \le \theta \le b) = 0.95$ then $[a, b]$ is called a 95% credible interval for θ.
- If the analysis is Bayesian and if $\Pr(a \le X \le b) = 0.95$ then $[a, b]$ is called a 95% credible interval for X.

The term "coverage probability" features in classical statistics, yet this term and the accompanying term "coverage interval" are being promoted for use in metrology in a context where the analysis is explicitly Bayesian. This seems likely to lead to some confusion.

References

1. G.J. Hahn, W.Q. Meeker, *Statistical Intervals: A Guide for Practitioners* (Wiley, 1991)
2. JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM) (2012), http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf
3. Joint Committee for Guides in Metrology, Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method (2006)
4. Joint Committee for Guides in Metrology, Evaluation of measurement data – Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Extension to any number of output quantities (2006)
5. Guide to the Expression of Uncertainty in Measurement (International Organization for Standardization, Geneva, 1995)
6. J. Pfanzagl, Estimation: Confidence Intervals and Regions, in *International Encyclopedia of Statistics*, edited by W.H. Kruskal, J.M. Tanur (The Free Press, Macmillan, 1978), pp. 259–267
7. G.K. Robinson, Confidence intervals and regions, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz, N.L. Johnson, C.B. Read (Wiley, 1982), Vol. 2, pp. 120–127
8. S.S. Wilks, *Mathematical Statistics* (Wiley, 1962)
9. H.J. Larson, *Introduction to Probability Theory and Statistical Inference*, 3rd edn. (Wiley, 1982)

10. A.M. Mood, F.A. Graybill, *Introduction to the Theory of Statistics*, 2nd edn. (McGraw-Hill, 1963)
11. R.E. Walpole, R.H. Myers, *Probability and Statistics for Engineers and Scientists*, 2nd edn. (Macmillan, 1978)
12. R.G. Miller Jr., *Simultaneous Statistical Inference*, 2nd edn. (Springer-Verlag, 1980)
13. F.S. Acton, *Analysis of Straight-Line Data* (Wiley, 1959)
14. B.W. Lindgren, *Statistical Theory* (Macmillan, 1968)
15. W.G. Howe, Two-sided tolerance limits for normal populations – some improvements, J. Am. Stat. Assoc. **64**, 610–620 (1969)
16. *NIST/SEMATECH e-Handbook of Statistical Methods* (2012), http://www.itl.nist.gov/div898/handbook/
17. H.A. David, *Order Statistics*, 2nd edn. (Wiley, 1981)
18. I. Guttman, Tolerance regions, statistical, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz, N.L. Johnson, C.B. Read (Wiley, 1988), Vol. 9, pp. 272–287
19. W.F. Guthrie, H. Liu, A.L. Rukhin, B. Toman, J.C.M. Wang, N. Zhang, Three Statistical Paradigms for the Assessment and Interpretation of Measurement Uncertainty, in *Data Modeling for Metrology and Testing in Measurement Science*, edited by F. Pavese, A.B. Forbes (Birkhäuser, 2009), pp. 71–115
20. A.W.F. Edwards, Fiducial probability, The Statistician **25**, 15–35 (1976)
21. D.V. Lindley, Bayesian inference, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz, N.L. Johnson, C.B. Read (Wiley, 1982), Vol. 1, pp. 197–204
22. R. Willink, B.D. Hall, A classical method for uncertainty analysis with multidimensional data, Metrologia **39**, 361–369 (2002)
23. A.P. Dawid, Invariant prior distributions, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz, N.L. Johnson, C.B. Read (Wiley, 1983), Vol. 4, pp. 228–236