Computing exact factors for two-sided tolerance limits in a normal distribution with unknown parameters in MATLAB

N M Borbatc, N Yu Chistokletov and T V Shkolina
Bryansk State Technical University, boulevard 50-let Oktyabrya 7, Bryansk, Russia

E-mail: shkolina.tv@yandex.ru

Abstract. MATLAB procedure is presented for calculating the exact factors used to find of two-sided tolerance limits in a normal distribution with unknown parameter values. The proposed procedure allows using arbitrary combinations of the sample size and the degrees of freedom of estimated variance. The procedure can be used in systems for processing the results of statistical quality control created on the basis of MATLAB.

1. Introduction
In the practice of quality management, valuable information regarding the distribution of a controlled quality characteristic is provided by the limits of the so-called tolerance intervals. The tolerance interval is an interval determined by the data of a random sample, relative to which it can be argued with a level of confidence \( \gamma = 1 - \alpha \) that it contains at least a given fraction \( p \) of the population of the random variable under study. The limits of the tolerance interval can be used in the statistical process control to assess their capabilities or in monitoring the quality of production batches to assess compliance with established requirements. So, based on the data of selective control, the values \( \xi \) and \( \bar{\xi} \) of the lower and upper limits of the tolerance interval are calculated, with respect to which, it can be argued with probability \( \gamma = 95\% \) that the units of production in a batch \( p = 99\% \) have quality characteristic values between them. Then, based on the results of comparison \( \xi \) and \( \bar{\xi} \) with the maximum permissible values (specification limits), you can decide whether to accept or reject the entire batch. Tolerance intervals also find application in solving many other practical problems [1 – 5].

In the case of the normal distribution of the controlled quality characteristic, special coefficients (tolerance factors) are used to calculate the limits \( \xi \) and \( \bar{\xi} \) of the tolerance interval. However, with unknown distribution parameters, the exact values of the factors in practice are determined by using look-up tables. Applying these tables is inconvenient in automated control systems for obvious reasons, and applying approximate formulae, under certain conditions, leads to unsatisfactory accuracy. In this regard, it remains an urgent task to develop algorithms for calculating the exact values of tolerant factors using various programming languages.

2. The problem of finding exact tolerance factors
Let \( X \) be a normally distributed random variable with parameters \( \mu \) and \( \sigma \) (i.e. \( X \sim N(\mu, \sigma) \)), and \( x_1, x_2, ..., x_n \) are sample volume data \( n \). With unknown parameter values \( \mu \) and \( \sigma \) their estimates
are found according to sample data which are sample mean \( \bar{x} \) and the sample standard deviation \( s \), respectively. At that \( s = \sqrt{s^2} \), where \( s^2 \) is an estimated variance with the degrees of freedom \( v \), but \( \bar{x} \) and \( s^2 \) are independent random variables. Then a two-sided tolerance interval containing at least \( p \)-n fraction of the distribution of the random variable \( X \), with a confidence level \( \gamma = 1 - \alpha \) is determined as:

\[
P_{\bar{x},s} \left[ P_X \left[ \frac{\bar{x} - k s}{\bar{x} + k s} \right] \geq p \right] = \gamma,
\]

where \( \bar{x} = \bar{x} - k s \) and \( \bar{x} = \bar{x} + k s \) are lower and upper limits of the tolerance interval, respectively.

The factor \( k \) used in the calculation \( \bar{x} \) and \( \bar{x} \), is defined as a solution to an integral equation of the form [6]:

\[
A(\gamma, k) = \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{+\infty} F(x, k) e^{-\frac{nx^2}{2}} dx - \gamma,
\]

where

\[
F(x, k) = P_{\chi^2_v} \left[ \frac{q^2}{2} > \frac{vr^2}{k^2} \right],
\]

where \( \chi^2_v \) is a random variable having a chi-square distribution with the degrees of freedom \( v \), and \( r \) is the root of the equation

\[
\Phi(x + r) - \Phi(x - r) = p,
\]

where \( \Phi(\cdot) \) is the cumulative function of the standard normal distribution.

At given values of \( n, v, p, \) and \( \gamma \) there is such value \( k \), which exactly satisfies equation (1), i.e. for a normal distribution with unknown parameters \( \mu \) and \( \sigma \), the problem of finding the limits of a two-sided tolerance interval has an exact solution. However, due to the nature of equation (1) its analytical solution regarding \( k \) is impossible, and therefore in practice, the approximation of the form is often used [6, 7]:

\[
k \equiv \left( \frac{v \chi^2_{1-v} / (1/n)}{\chi^2_{1-v} / 2} \right)^{1/2},
\]

where \( \chi^2_{\alpha, v} \) is a quantile of \( \alpha \) the chi-square of the distribution with the degrees of freedom \( v \), and \( \chi^2_{p, v}(\delta) \) is a quantile of \( p \) a noncentral chi-square distribution with the degrees of freedom \( v \) and noncentrality parameter \( \delta \).

However, as was shown [8], the accuracy of approximation by relation (3) is satisfactory only for the case when the degrees of freedom \( v \) of the sample variance \( s^2 \) equal \( n - 1 \). At the same time, in practice, situations arise in which \( v \neq n - 1 \), in particular when carrying out regression analysis or analysis of variance. For example, let a one-way analysis of variance be performed using a fixed-effect model to compare \( a \) machines performance. The output from each machine is described by the normal distribution with the same variance \( \sigma^2 \) and possibly different mean values. If for each machine \( n \) observations were carried out, then if it is necessary to build a two-sided tolerance interval for the output of one of the machines it turns out, that \( v = a(n - 1) \) (the degrees of freedom of the mean squared error). In this case, the greater the value \( v \) in comparison with \( n \), the greater the error is when
using approximation (3), therefore, when \( \nu \neq n - 1 \) it is necessary to use special approximation methods to find the value of the tolerance factor.

The exact values of the factor \( k \) satisfying equation (1) were first obtained using numerical methods and were summarized in look-up tables. Similar tables are given, for example, in ISO 16269-6:2014 “Statistical interpretation of data – Part 6: Determination of statistical tolerance intervals”. Moreover, these table contain values of tolerance factors only for a limited number of values of the degrees of freedom for estimated variance \( \nu \), the proportion of the distribution \( p \), and the level of confidence \( \gamma \). However, even more detailed look-up tables [9, 10] turn out to be inconvenient for using in various systems of automated processing of control results. In this regard, the urgent task is to develop a procedure for calculating the exact values of tolerant factors for any combination of values \( n \), \( \nu \), \( p \), and \( \gamma \) in a suitable programming language.

3. The procedure for calculating exact tolerance factors in MATLAB

The procedures for calculating the exact values of the factors \( k \) to find the limits of two-sided tolerance intervals were developed in the programming languages FORTRAN [11] and BASIC [12]. Recently, MATLAB has become increasingly popular. MATLAB is a package of application programs and the similarly-named programming language used for various kinds of technical calculations in different fields, including statistical quality control. However, despite having all the variety of statistical tools, MATLAB it does not provide for the possibility of finding the limits of tolerant intervals. In this regard, the authors developed a procedure (user-defined M file function) to find the exact values of the factors for two-sided tolerance limits with unknown normal distribution parameters in MATLAB (version R2017b), a full listing of which is shown in Figure 1.

The procedure is called with three required parameters: sample size \( n \); share of distribution \( p \) within the tolerance interval; level of confidence \( \gamma \). As a fourth, optional argument, the degrees of freedom \( \nu \) for the estimated variance can be specified. If the fourth input argument is omitted, then the default is \( \nu = n - 1 \). All input arguments must be scalar values and checked against their valid values: \( n \geq 2 \), \( 0 < p < 1 \), \( 0 < \gamma < 1 \), and \( \nu \geq 1 \). For given input arguments, the procedure returns the exact value of the two-sided tolerance factor \( k \).

A procedure is a hierarchy of nested functions, each of which solves its own problems. Inside the main function, the search for the value of the tolerance factor \( k \), which is an exact solution of equation (1), is performed using the built-in function MATLAB fzero. As initial value \( k_0 \), required to call the function fzero, the approximate value of the tolerance factor found from relation (3) is used. At the same time, built-in function nncx2inv is used to calculate the quantile of the noncentral chi-square distribution and built-in function chi2inv is used to calculate the quantile of the chi-square distribution. The very solvable equation (1) is passed to the fzero function using a handle to the first nested function FINTOLMUL.

The FINTOLMUL nested function is equation (1) solved with respect to \( k \), which, for convenience, is expressed in the form:

\[
A(\gamma, k) = 2 \int_{0}^{+\infty} F\left(\frac{z}{\sqrt{n}}, k\right) \varphi(z) dz - \gamma, \quad (4)
\]

where \( \varphi(z) \) is density function of standard normal distribution. Relation (4) is obtained from formula (1) using the substitution \( z = x\sqrt{n} \) and taking into account the parity property of the integrand [11].

The integral on the right-hand side of expression (4) is calculated with each call to FINTOLMUL using the built-in function of numerical integration MATLAB integral. The integrand is passed to the integral function by a handle to the second nested INTEGRFUN function. Inside this function to calculate \( F\left(z/\sqrt{n}, k\right) \) built-in function chi2cdf is used, and to calculate \( \varphi(z) \) built-in function normpdf is
used. At the same time, each call to the \textit{INTEGRFUN} function to calculate \( F\left(z/\sqrt{n}, k\right) \) using the \textit{fzero} function, value \( R \) is calculated, which is a solution to the equation:

\[
\Phi\left(z/\sqrt{n} + r\right) - \Phi\left(z/\sqrt{n} - r\right) = p \tag{5}
\]

which is obtained from expression (2) when replacing by a variable \( z = x\sqrt{n} \).

function \( k = \text{tolfact}(n,p,\text{gamma},\text{nu}) \)
%Calculates the exact value of the tolerance factor for finding the limits
%of the two-sided tolerance interval with unknown normal distribution
%parameters.
% Input arguments:
% \( n \) - data volume;
% \( p \) - fraction of distribution within the interval;
% \text{gamma} - confidence level;
% \text{nu} - degrees of freedom of variance estimation (optional).
% Output argument:
% \( k \) - value of tolerance factor.
%
%---------------------------------------------------------------------
% Designed by: Bryansk State Technical University
% The authors: Borbat N.M., Shkolina T.V.
% E-mail: borbact@mail.ru; shkolina.tv@yandex.ru
%
%---------------------------------------------------------------------
% checking input arguments
if nargin < 4 || isempty(nu)
    nu = n-1; \% default degrees of freedom of variance estimation
end
if ~isscalar(n) || ~isscalar(p) || ~isscalar(gamma) || ~isscalar(nu)
    error('All input arguments must be scalars.')
end
if n < 2 || (p <= 0 || p >= 1) || (gamma <= 0 || gamma >= 1) || nu < 1
    error('Invalid input argument values.')
end
% the approximate value is used as the starting point of the root search
k0 = sqrt(nu*ncx2inv(p,1.1/n)/chi2inv(1-gamma,nu));
% finding the root of the equation
k = fzero(@FINTOLMUL,k0);
% integral function
function I = FINTOLMUL(n)
    I = 2*integral(@INTEGRFUN,0,Inf,'ArrayValue',true)-gamma;
% integrand function
    function t = INTEGRFUN(z)
        R0 = sqrt(ncx2inv(p,1.1/sqrt(n)));
        R = fzero(@(FindR,R0);
        t = chi2cdf(nu.*R.^2,h.^2,nu,'upper')*normpdf(z);
        function y = FindR(r)
            y = normcdf(z/sqrt(n)+r)-normcdf(z/sqrt(n)-r)-p;
        end
    end
end

\textbf{Figure 1.} Listing procedure for finding the exact two-sided tolerance factors in MATLAB.

In turn, to find the root of function (5), the value \( R_0 \) is passed to the \textit{fzero} function as the starting point of the search. The value \( R_0 \) is calculated by the relation:

\[
R_0 = \sqrt{\chi^2_{p,n}(\delta)}
\]

where \( \chi^2_{p,n}(\delta) \) is the quantile of the level \( p \) of noncentral chi-square distribution with the degrees of freedom \( n \) and the noncentrality parameter \( \delta \). The very solvable equation (5) is passed using a
pointer to the third nested function FindR, inside which calculating cumulative function of the standard normal distribution $\Phi(\cdot)$ is carried out using the built-in function normcdf.

4. Comparison calculation results with the known values

Table 1 compares the exact values of the factors $k$, calculated by the proposed procedure, with the exact values $k^*$, given in [11] for various values of $n$ and $v$ at $p = \gamma = 0.99$, and table 2 compares the exact values of the factors $k$, calculated by the proposed procedure, with the exact values $k^{**}$, taken from tables in ISO 16269–6:2014 for various combinations of $p$, $\gamma$, $n$, and $v$.

| Factor | $n = 25$ | $n = 10$ |
|--------|---------|---------|
| $k$    | 26.1698 | 5.1956  |
| $k^*$  | 26.17   | 5.196   |

| Factor | $n = 2$ | $n = 3$ | $n = 5$ | $n = 10$ |
|--------|---------|---------|---------|---------|
| $k$    | 4.3174  | 3.9874  | 3.6633  | 3.361   |
| $k^*$  | 4.317   | 3.987   | 3.663   | 3.361   |

Table 2. Comparing calculation results for the proposed procedure with tabular values

| Factor | $p = 0.90; \gamma = 0.90$ | $p = 0.90; \gamma = 0.95$ | $p = 0.95; \gamma = 0.99$ |
|--------|--------------------------|--------------------------|--------------------------|
| $n = 5$ | 3.4993                   | 2.2383                   | 1.8757                   |
| $n = 10$ | 3.3690                   | 2.1724                   | 1.8000                   |
| $n = 20$ | 3.3000                   | 2.7924                   | 1.7200                   |
| $n = 5$ | 3.3300                   | 2.7924                   | 1.7200                   |
| $n = 100$ | 182.7201                 | 4.6162                   | 2.5791                   |

The tables show that the values of the two-sided tolerance factors calculated by the proposed procedure with an accuracy of three decimal places coincide with the known values.

5. Conclusion

The procedure proposed by the authors allows one to find the exact values of the factors, on the basis of which the limits of the two-sided tolerance intervals are calculated under a normal distribution with unknown parameter values. This procedure can be used as a composite module of the statistical processing system of quality control results. This system is based on MATLAB and comparison of the tolerance interval limits with the specification limits on the controlled quality characteristic.

References

[1] Meeker W Q, Hahn G J and Escobar L A 2017 Statistical Intervals. A Guide for Practitioners and Researchers. Second Edition (Hoboken: Wiley)
[2] Pusztai É and Kemény S 2020 Process capability indices when two sources of variability present, a tolerance interval approach Journal of Chemometrics e3213
[3] Franço B G, Lin D and Hoyer W 2019 Confidence, prediction, and tolerance in linear mixed models Statistics in Medicine 38(30) 5603-22
[4] Flouri M, Zhai S, Mathew T and Bebu I 2017 Tolerance limits and tolerance intervals for ratios
of normal random variables using a bootstrap calibration *Biometrical Journal* 59(3) 550-66

[5] Tsong Y, Wang T and Hu X 2019 Statistical Considerations in Setting Quality Specification Limits Using Quality Data *Springer Proceedings in Mathematics and Statistics* 218 3-12

[6] Wald A and Wolfowitz J 1946 Tolerance Limits for a Normal Distribution *Annals of Mathematical Statistics* 17(2) 208-15

[7] Krishnamoorthy K and Mathew T 2009 *Statistical Tolerance Region: Theory, Applications, and Computation* (Hoboken: Wiley)

[8] Ellison B E 1964 On Two-Sided Tolerance Intervals for a Normal Distribution *Annals of Mathematical Statistics* 35(2) 762-72

[9] Garaj I and Janiga I 2002 Two-sided Tolerance Limits of Normal Distribution for Unknown Mean and Variability (Bratislava: Vydavatel’stvo STU)

[10] Garaj I and Janiga I 2004 Two-sided Tolerance Limits of Normal Distribution with Unknown Mean and Unknown Common Variability. (Bratislava: Vydavatel’stvo STU)

[11] Eberhardt K R, Mee R W and Reeve C P 1989 Computing factors for exact two-sided tolerance limits for a normal distribution *Communications in Statistics – Simulation and Computation* 18(1) 397-413

[12] Fujino Y 1989 Exact Two-sided Tolerance Limits for a Normal Distribution *Japanese Journal of Applied Statistics* 18(1) 29-36