Sample size determination strategies for normal tolerance intervals using historical data

Derek S. Young, Charles M. Gordon, Shihong Zhu & Bryan D. Olin

To cite this article: Derek S. Young, Charles M. Gordon, Shihong Zhu & Bryan D. Olin (2016) Sample size determination strategies for normal tolerance intervals using historical data, Quality Engineering, 28:3, 337-351, DOI: 10.1080/08982112.2015.1124279

To link to this article: http://dx.doi.org/10.1080/08982112.2015.1124279

Published online: 06 Apr 2016.
Sample size determination strategies for normal tolerance intervals using historical data

Derek S. Young, Charles M. Gordon, Shihong Zhu, and Bryan D. Olin

Department of Statistics, University of Kentucky, Lexington, Kentucky; Cyberonics, Houston, Texas; Fifth Third Bank, Cincinnati, Ohio

ABSTRACT

Statistical tolerance intervals are often used during design verification or process validation in diverse applications, such as the manufacturing of medical devices, the construction of nuclear reactors, and the development of protective armor for the military. Like other statistical problems, the determination of a minimum required sample size when using tolerance intervals commonly arises. Under the Faulkenberry-Weeks approach for sample size determination of parametric tolerance intervals, the user must specify two quantities—typically set to rule-of-thumb values—that characterize the desired precision of the tolerance interval. Practical applications of sample size determination for tolerance intervals often have historical data that one expects to closely follow the distribution of the future data to be collected. Moreover, such data are typically required to meet specification limits. We provide a strategy for specifying the precision quantities in the Faulkenberry-Weeks approach that utilizes both historical data and the required specification limits. Our strategy is motivated by a sampling plan problem for the manufacturing of a certain medical device that requires calculation of normal tolerance intervals. Both classical and Bayesian normal tolerance intervals are considered. Additional numerical studies are provided to demonstrate the general applicability of our strategy for setting the precision quantities.

Introduction

A statistical tolerance interval is constructed based on a random sample so that it includes at least a specified proportion $P$ of the sampled population with a given confidence level $(1 - \alpha)$. Statistical tolerance intervals were developed as a means to assess whether or not products meet specification (spec) limits. Many examples of tolerance intervals applied to real engineering applications are discussed in the texts by Hahn and Meeker (1991) and Krishnamoorthy and Mathew (2009). Other applications of tolerance intervals can be found in clinical studies (Katki et al. 2005; Cummings et al. 2011; Zaslavsky 2007), bioequivalence studies (Westlake 1979; Esinhart and Chinchilli 1994; Brown et al. 1994), and environmental monitoring (Allen and Jones 1998; Smith 2002; Krishnamoorthy et al. 2008). More generally, statistical tolerance intervals can be used for process control or to demonstrate conformance to limits established by federal U.S. agencies, like the Food and Drug Administration (FDA), National Institute of Standards and Technology (NIST), and Environmental Protection Agency (EPA).

Bayesian tolerance intervals are, perhaps, less employed in practice, but have been developed in the literature for various settings. Bayesian normal tolerance intervals were first derived in Aitchison (1964). Bayesian tolerance intervals for balanced random effects models were developed in Wolfinger (1998) and Van der Merwe and Hugo (2007). Choudhary (2007) developed Bayesian semiparametric regression tolerance bands for assessing agreement between two methods that measure a continuous variable. Asymptotic results have also been derived for two-sided Bayesian tolerance intervals—including the characterization of probability-matching priors—in Pathmanathan et al. (2014). A general treatment of Bayesian tolerance intervals is provided in Chapters 7–9 of Guttman (1970) and Chapter 11 of Krishnamoorthy and Mathew (2009).

CONTACT

Derek S. Young derek.young@uky.edu Department of Statistics, University of Kentucky, 323 Multidisciplinary Science Building, 725 Rose Street, Lexington, KY 40536.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.

© 2016 Taylor & Francis
To motivate the need for methods to identify adequate sample sizes, we focus on the application of tolerance intervals during design verification or process validation in the medical device industry. Prior to regulatory approval, medical devices must satisfactorily meet numerous technical specifications and user requirements. Conformance to specifications can be assessed by showing that some minimum proportion conforms at a pre-determined confidence level; consequently, tolerance intervals are often relied upon. Determining the appropriate sample size involves a trade-off between the risk associated with the product or characteristic under test as well as economic and resource constraints, such as sample cost and testing time. Furthermore, the FDA and international standards mandate any sampling plans be based on a valid statistical rationale.

The application central to our discussion concerns the design verification stage of the product development process at Cyberonics, a manufacturer of neuromodulation devices. The company developed and markets the Vagus Nerve Stimulation (VNS) Therapy® system, which is FDA-approved for the treatment of refractory epilepsy and treatment-resistant depression. The VNS Therapy system uses a surgically implanted medical device that delivers pulsed electrical signals to the vagus nerve. It is common for such medical devices to have existing data to study from either predicate devices or characterization work conducted prior to finalizing (or freezing) the device design. These data characterize process stability and capability. From a capability standpoint, the historical data can be compared to the specification limits to inform the minimum sample size. If these historical data indicate that the device meets specification by a wide margin, then fewer samples should be needed to demonstrate conformance for a given confidence level.

There is limited literature on how to determine sample sizes for tolerance intervals and none of that literature provides guidance on how to account for the existence of prior data, which is crucial for the economic planning of scientific studies. We consider the traditional sample size determination methodology of Faulkenberry and Weeks (1968)—henceforth referred to as the Faulkenberry-Weeks approach—in order to address the statistical problem of the medical device data. In the Faulkenberry-Weeks approach, one needs to specify two precision quantities, \( \delta \) and \( P' \), in addition to the levels that characterize the desired tolerance interval. Since rule-of-thumb values for \( \delta \) and \( P' \) are always used in the literature, we propose an informed strategy that uses parameter estimates from representative historical data as well as specification limits for setting the levels of \( \delta \) and \( P' \). Our proposed strategy is adaptable to both classical and Bayesian normal tolerance intervals. We also provide readily-available \( \texttt{R} \) (R Development Core Team 2014) code for performing the proposed sample size determination procedure that leverages the efficient and accurate routines for computing noncentral distributions and performing numerical integration. Such calculations as they relate to tolerance intervals are found in the \( \texttt{R} \) package \texttt{tolerance} (Young 2010, 2014).

The rest of this article is organized as follows. We first present the theory for tolerance intervals as well as the formulas for normal tolerance intervals—both classical and Bayesian. We then discuss sample size determination for tolerance intervals as well as our strategy for selecting \( \delta \) and \( P \) in the Faulkenberry-Weeks approach. We then summarize some numerical results, analyze the medical device data, and determine an appropriate sample size for future manufactured products while considering design specification limits. Finally, we provide a discussion that highlights the feasibility with extending our strategy to other parametric distributions.

**Tolerance intervals**

Let \( \mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\} \) be a set of parametric univariate distributions. Suppose that \( X = (X_1, X_2, \ldots, X_n) \) is a random sample of size \( n \) from \( F_\theta \in \mathcal{F} \) and that \( X \) is a random variable that also follows \( F_\theta \), independently of \( X \). Let \( T(X) \) be a random subset of \( \mathbb{R} \) and define \( C_{\theta_0}(T(X)) \) to be the probability content under \( F_\theta \) of the set \( T(X) \). If
\[
\inf_{\theta \in \mathcal{F}} P(C_\theta(T(X)) \geq P) = 1 - \alpha, \tag{1}
\]
then \( T(X) \) is called a \((P, 1 - \alpha)\) tolerance set of the underlying distribution. \( P \in (0, 1) \) and \((1 - \alpha) \in (0, 1)\) are the desired (probability) content level and confidence level, respectively, for the tolerance set.

---

Using (1), a \((P, 1 - \alpha)\) one-sided upper tolerance limit \((U_1(X))\) and one-sided lower tolerance limit \((L_1(X))\), both of which depend on the random sample \(X\), satisfy the expressions

\[
P_X (P_X [X \leq U_1(X)|X] \geq P) = 1 - \alpha \quad [2]
\]
and

\[
P_X (P_X [L_1(X) \leq X|X] \geq P) = 1 - \alpha, \quad [3]
\]
respectively. Similarly, a \((P, 1 - \alpha)\) two-sided tolerance interval, \((L_2(X), U_2(X))\), satisfies

\[
P_X (P_X [L_2(X) \leq X \leq U_2(X)|X] \geq P) = 1 - \alpha. \quad [4]
\]

For the Bayesian set-up, let \(\mathbf{x}\) be a vector of realizations of \(X\), \(L(\theta|x)\) be the likelihood function, and \(\pi(\theta)\) denote a prior distribution for \(\theta\). The posterior distribution of \(\theta\) is then given by

\[
p(\theta|x) = \frac{L(\theta|x)\pi(\theta)}{\int_\theta L(\theta|x)\pi(\theta)\,d\theta}.
\quad [5]
\]
Bayesian one-sided upper and lower tolerance limits satisfy

\[
P_\theta (P_X [X \leq U_1(\theta)|\theta] \geq P|X) = 1 - \alpha \quad [6]
\]
and

\[
P_\theta (P_X [L_1(\theta) \leq X|\theta] \geq P|X) = 1 - \alpha, \quad [7]
\]
respectively, while a \((P, 1 - \alpha)\) Bayesian two-sided tolerance interval satisfies

\[
P_\theta (P_X [L_2(\theta) \leq X \leq U_2(\theta)|\theta] \geq P|X) = 1 - \alpha. \quad [8]
\]
Notice that the probability content calculated for Eqs. [6]–[8] are each defined as functions of the parameter vector \(\theta\), which we then calculate with respect to the posterior probability measure \(P_\theta\) of being greater than or equal to \(P\). Moreover, the value of \(\theta\) used to calculate the limits in Eqs. [6]–[8] is the maximum likelihood estimator \(\hat{\theta}\).

**Classical normal tolerance intervals**

Tolerance intervals for normal distributions have been extensively studied in the literature, with some of the earliest works being Wilks (1941, 1942), Wald (1943), and Wald and Wolfowitz (1946). Let \(X_1, \ldots, X_n\) be independent and identically distributed random normal variables with mean \(\mu\) and variance \(\sigma^2\), both of which are unknown. Let \(\bar{X}\) denote the sample mean and \(S^2\) denote the sample variance. The formulas for \((P, 1 - \alpha)\) lower and upper normal tolerance limits are

\[
L_j(X) = \bar{X} - k_j(n, \alpha, P)S \quad \text{and} \quad U_j(X) = \bar{X} + k_j(n, \alpha, P)S, \quad [9]
\]
respectively, where \(j = 1, 2\). \(k_1(n, \alpha, P)\) and \(k_2(n, \alpha, P)\) are the \(k\)-factors for the one-sided and two-sided tolerance limits, respectively. \(k_1(n, \alpha, P)\) is calculated by

\[
k_1(n, \alpha, P) = \frac{1}{\sqrt{n}}t_{n-1; 1 - \alpha} (\sqrt{n}z_p), \quad [10]
\]
where \(t_{d,q}(\xi)\) is the \(q\)th quantile of a noncentral \(t\)-distribution with \(d\) degrees of freedom and noncentrality parameter \(\xi\) and \(z_p\) is the \(p\)th quantile of the standard normal distribution. \(k_2(n, \alpha, P)\) is the solution to the integral equation

\[
\sqrt{\frac{2n}{\pi}} \int_0^\infty P(X^2 > \frac{(n - 1)\chi_{f; p}^2(z^2)}{k_2(n, \alpha, P)^2}) e^{-\frac{1}{2}z^2} \, dz = 1 - \alpha, \quad [11]
\]
where \(\chi_{f; p}^2\) is the chi-square random variable with \(f\) degrees of freedom and \(\chi_{f; q}(\xi)\) is the \(q\)th quantile of the noncentral chi-squared distribution with \(f\) degrees of freedom and noncentrality parameter \(\xi\). The formulas for \(k_1(n, \alpha, P)\) and \(k_2(n, \alpha, P)\) are due to Lemma A.1 given in the Appendix.

Normal tolerance intervals constructed using the two-sided tolerance limits above are, perhaps, the most commonly used. However, one can also construct two-sided tolerance intervals that contain at least a proportion \(P\) of the center of the normal population with confidence level \((1 - \alpha)\). Such equal-tailed tolerance intervals still take the form of [9], but where the tolerance factor \(k_0(n, \alpha, P)\) is found as the solution to the integral equation

\[
\left(\frac{2n}{\pi} \right)^{1/2} \Gamma \left(\frac{n - 1}{2}\right)^{-1} \int_{-\delta}^{\delta} \left(2\Phi \left(-\delta + \frac{k_0\sqrt{n}z_p}{\sqrt{n - 1}}\right) - 1\right) e^{-z^2/2} \, dz = 1 - \alpha, \quad [12]
\]
where \(\delta = \sqrt{n}z_{1 - \alpha}\) and \(\Phi(\cdot)\) denotes the standard normal cumulative distribution function.

In the past, challenges with computing noncentral distributions necessitated the use of approximations for \(k_1(n, \alpha, P)\), \(k_2(n, \alpha, P)\), and \(k_0(n, \alpha, P)\). Moreover, the same challenges were faced with solving the integral equations in [11] and [12], which resulted in tables being published for certain values of \(n, P,\) and \(\alpha\); see,
Bayesian normal tolerance intervals

For our discussion of Bayesian normal tolerance intervals, we consider the setting where both the parameters $\mu$ and $\sigma^2$ are unknown. We use the conjugate priors $\pi(\mu|\sigma^2)$ and $\pi(\sigma^2)$, which are

$$
\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/n_0)
$$

and

$$
\sigma^2 \sim \text{Scale-inv-}\chi^2(m_0, \sigma_0^2),
$$

respectively, where $\mathcal{N}(\gamma, \xi^2)$ denotes the normal distribution with mean $\gamma$ and variance $\xi^2$ and $\text{Scale-inv-}\chi^2(\nu, \tau^2)$ denotes the scaled inverse chi-squared distribution with $\nu$ degrees of freedom and scaling parameter $\tau^2$. Note that when a scaled inverse chi-squared random variable is divided by $\nu\tau^2$, it results in an inverse chi-squared random variable with $\nu$ degrees of freedom. The joint prior density of $(\mu, \sigma^2)$ is then given by

$$
\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2).
$$

The prior structure has four hyperparameters: $\mu_0 \in \mathbb{R}$, $\sigma_0^2 > 0$, $m_0 > 0$, $n_0 > 0$, $m_0$ and $n_0$ are not prior sample size quantities, but are intended to reflect the prior precision relative to the sample size. They are tunable by the researcher to reflect their prior confidence and, thus, are chosen subjectively. We briefly discuss our choices for these quantities for our numerical study and real data analysis presented later. The joint posterior distribution is then determined by

$$
p(\mu, \sigma^2|x) = p(\mu|\sigma^2)p(\sigma^2),
$$

where $p(\mu|\sigma^2)$ and $p(\sigma^2)$ are the distributions

$$
\mu|\sigma^2 \sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n_0+n}\right)
$$

and

$$
\sigma^2 \sim \text{Scale-inv-}\chi^2(m_0+n-1, q^2),
$$

respectively, such that

$$
\bar{x} = \frac{n_0\mu_0+n\bar{x}}{n_0+n}
$$

and

$$
q^2 = (m_0+n-1)^{-1}
$$

$$
\times \left[ m_0\sigma_0^2 + (n-1)s^2 + \frac{n_0n}{n_0+n}(\bar{x} - \mu_0)^2 \right].
$$

See Robert (2007) or Gelman et al. (2013) for details on the above derivations. Note that in the Bayesian set-up, we condition on realizations of the observed data, i.e., $X = x$. Therefore, the formulas in the Bayesian set-up are written in terms of the observed sample $x$, where the sample estimates of the mean and variance are $\bar{x}$ and $s^2$, respectively.

The formulas for one-sided Bayesian normal tolerance limits and two-sided Bayesian normal tolerance intervals are similar to the classical setting. ($P, 1 - \alpha$) lower and upper Bayesian normal tolerance limits are

$$
L_j(\bar{x}, s^2) = \bar{x} - k_j(n, n_0, m_0, \alpha, P)q
$$

and

$$
U_j(\bar{x}, s^2) = \bar{x} + k_j(n, n_0, m_0, \alpha, P)q,
$$

respectively, where $j = 1, 2$. Note that these limits are expressed in terms the maximum likelihood estimates of $\mu$ and $\sigma^2$, which occur through how $\bar{x}$ and $q$ are defined. The one-sided $k$-factor is calculated by

$$
k_j(n, n_0, m_0, \alpha, P) = \frac{1}{\sqrt{n + n_0}}_k \frac{tm_m - tn + \frac{t}{|\alpha - 1|}}{(\sqrt{n + n_0}z_\alpha)},
$$

which follows from part (a) of Lemma A.1 in the Appendix. The two-sided $k$-factor is calculated by finding the solution to the integral equation

$$
\sqrt{\frac{2(n + n_0)}{\pi}} \int_0^\infty q^n_p \left( \frac{\chi_{m_0+n}^2}{k_j(n, n_0, m_0, \alpha, P)} \right) e^{-\frac{q(n+n_0)}{2}} dz = 1 - \alpha.
$$

Furthermore, if one considers the non-informative prior distribution

$$
\pi(\mu, \sigma^2) \propto \sigma^{-2},
$$

it can be shown that the solutions for the one-sided Bayesian normal tolerance limits and two-sided Bayesian normal tolerance intervals are the same as for the classical setting given in Eqs. [9]–[11]. Details on the derivation for the one-sided limits—under the conjugate prior and non-informative prior—can be found in Chapter 11 of Krishnamoorthy and Mathew (2009). The derivations for the two-sided limits under both prior structures are given in the Appendix.

Sample size determination strategies

Direct optimization when given specification limits

For planning purposes, suppose that we assume $\mu$ and $\sigma$ are known. We first consider the problem of designing a study to demonstrate that a process or product exceeds a lower specification limit, $S_L$, and/or falls below an upper specification limit, $S_U$. We are interested in the minimum sample size necessary such that...
It is that they had to rely on inefficient approximations for the $k$-factor. However, greater computing power was available for Odeh et al. (1987), which allowed for highly accurate $k$-factor calculations.

Since the calculation of exact $k$-factors for normal tolerance intervals is no longer a practical challenge, we will be able to easily implement the Faulkenberry-Weeks approach for sample size determination.

Recall that $X$ is a sample of size $n$ from a $\mathcal{N}(\mu, \sigma^2)$ distribution, $\bar{X}$ and $S^2$ are the sample mean and variance, respectively, and $X \sim \mathcal{N}(\mu, \sigma^2)$ independently of $X$. Let $F_{\mu, \sigma}$ denote the $\mathcal{N}(\mu, \sigma^2)$ distribution. We write the content of the $(P, 1 - \alpha)$ two-sided tolerance interval, respectively, as

$$C_{F_{\mu, \sigma}}(X) = P_X(\bar{X} - k_1S \leq X | X)$$

$$C_{F_{\mu, \sigma}}(X) = P_X(X \leq \bar{X} + k_1S | X)$$

$$C_{F_{\mu, \sigma}}(X) = P_X(\bar{X} - k_2S \leq X \leq \bar{X} + k_2S | X).$$

Since the Faulkenberry-Weeks approach can be formulated similarly for each setting above, we will generically denote the content as $C_{F_{\mu, \sigma}}(X)$. Therefore, the $(P, 1 - \alpha)$ tolerance interval of interest is determined such that

$$P_X(C_{F_{\mu, \sigma}}(X) \geq P) = 1 - \alpha. \quad [25]$$

The Faulkenberry-Weeks approach is designed to ensure that the calculated tolerance limits are “close” to the quantiles that result in a content level greater than or equal to $P$. So to ensure the “goodness” of the tolerance limits, one must choose some arbitrary $P' > P$ and small $\delta > 0$ to determine a sample size $n^*$ such that

$$P_X(C_{F_{\mu, \sigma}}(X) \geq P') \leq \delta. \quad [26]$$

Criterion [26] implies that there is only a small probability $\delta$ that the $(P, 1 - \alpha)$ tolerance interval of interest will have a content that exceeds $P$ by a margin of $(P' - P)$. Note that [25] and [26] are the traditional tolerance interval calculations, which are used to solve for an unknown $n^*$. The complexity of the approach is due to the messy nonlinear form of the $k$-factor.

The sample size $n^*$ determined using the Faulkenberry-Weeks approach depends on how large (or small) $P'$ and $\delta$ are chosen. There is no direct guidance in the literature on how to choose $P'$ other than it must be greater than $P$, while $\delta$ is usually chosen from the set of small values $\{0.01, 0.05, 0.10\}$. For a given application, the choice of appropriate levels might not be clear. A lack of such clarity could cause a practitioner to select levels that yield sample sizes larger than what may be required and, thus, incur unnecessary production costs. Or the practitioner...
could select levels that yield smaller sample sizes and run the risk of obtaining wider than expected tolerance intervals, possibly triggering unnecessary adjustments to the process or additional testing. For our medical device data, we developed a strategy that helps remove some of the subjectivity when choosing \( P' \) and \( \delta \). Such a strategy is general enough that it is portable to other applications that require sample size determination for normal tolerance intervals. By having historical data that has been shown to be in statistical control as well as within specification limits, we can leverage these features when applying the Faulkenberry-Weeks approach to remove some of the subjectivity in choosing \( P' \) and \( \delta \). The strategy can be used with classical or Bayesian normal tolerance intervals.

**Modified approach**

When designing a study for a process or product, there is often data from feasibility/pilot studies or historical data that can be used as a guide to intelligently select safety margins. This data is crucial for specifying conditions in experimental designs, but leveraging such data in the context of sample size determination for tolerance intervals has not been discussed. We address this problem using a modification to the Faulkenberry-Weeks approach. In our approach, we assume that the historical data are representative. We consider data to be representative when a subject-matter expert has assured that the historical (or sometimes pilot) data used has met some practical requirements, such as the following.

1. The design of the historical and current products are similar with respect to the characteristic being tested.
2. The manufacturing process used to make the historical and current products are similar, e.g., people, processes, and procedures do not vary dramatically in the view of the subject-matter expert.
3. The process generating the historical data is stable.

Taken together, requirements like the above provide a practical approach for engineers when assessing the appropriateness of the assumption that data are independent and identically distributed from the same distribution, which should be tested for normality in our application.

Let \( \bar{\mu} \) and \( \bar{\sigma} \) be posited values for the mean and standard deviation, which are calculated from our historical data, \( \bar{X} \). For the Bayesian tolerance intervals, these will be hyperparameters for the conjugate priors in [13], but the user will also need to specify values for \( m_0 \) and \( m_0 \). Let \( X^* \) be our most recent data, independent of \( \bar{X} \), that we have observed and \( C_{\bar{F}_\mu, \bar{\sigma}}(X^*) \) be the probability content of the calculated \((P, 1 - \alpha)\) tolerance interval (one-sided or two-sided) under a \( \mathcal{N}(\bar{\mu}, \bar{\sigma}^2) \) distribution. It is important to emphasize that \( C_{\bar{F}_\mu, \bar{\sigma}}(X^*) \) is reflecting uncertainty in the \((P, 1 - \alpha)\) tolerance interval calculated using \( X^* \) by computing the content of that tolerance interval using the posited values of \( \bar{\mu} \) and \( \bar{\sigma} \) determined from the historical data \( \bar{X} \). Since \( \delta \) in the Faulkenberry-Weeks approach is intended to be a small probability such that the tolerance interval will exceed \( P \) by a certain margin, we propose using a relative error measure with respect to the specified content level \( P \),

\[
\delta = \frac{|C_{\bar{F}_\mu, \bar{\sigma}}(X^*) - P|}{P},
\]

which we refer to as a *relative content error* (or RACE) measure. Thus, if \( X^* \) follows a normal distribution with mean and variance close to the posited (or historical) values \( \bar{\mu} \) and \( \bar{\sigma}^2 \), then we would expect a tighter precision value based on the RACE measure. Note that [27] is applicable to sample size determination for either a one-sided or two-sided \((P, 1 - \alpha)\) normal tolerance interval.

\( P' \) reflects the proportion of the sampled population that we are willing to let our \((P, 1 - \alpha)\) tolerance interval actually cover. Since the tolerance interval based on the collected data must be within the specification limits \((S_L, S_U)\), a useful value for \( P' \) would be one that reflects the probability content relative to these limits. Thus, we propose using the estimated probability of conformance to the specifications to estimate \( P' \) given the historical data as follows:

\[
P' = \begin{cases} 
\Phi \left( \frac{S_U - \bar{\mu}}{\bar{\sigma}} \right) - \Phi \left( \frac{S_L - \bar{\mu}}{\bar{\sigma}} \right) & \text{for two-sided tolerance intervals;} \\
1 - \Phi \left( \frac{S_L - \bar{\mu}}{\bar{\sigma}} \right) & \text{for one-sided lower tolerance intervals;} \\
\Phi \left( \frac{S_U - \bar{\mu}}{\bar{\sigma}} \right) & \text{for one-sided upper tolerance intervals.}
\end{cases}
\]
Note that $P' \leq P$ will occur if the posited values $\tilde{\mu}$ and $\tilde{\sigma}^2$ result in a wide $(P, 1 - \alpha)$ tolerance interval relative to the specification limits. In such a case, we recommend that the user examine a histogram of the data to determine if this is caused by the central tendency of the data being near one of the specification limits and/or high variability. Depending on the cause, this should motivate consideration for either redesign of the product, process, or test methods, or an increased sample size to demonstrate that the process or product meets the specification limits. Finally, if no specification limits are established and the user does not propose a value for $P'$, then we simply recommend using $P' = (1 + P)/2$. Such a rule-of-thumb will be more pragmatic since it does not result in too large or too small of a sample size as $P'$ approaches $P$ and 1, respectively.

**Numerical results**

In this section, we provide various numerical studies regarding the approaches discussed above.

We begin by briefly illustrating the direct method, where we assume that the population mean and standard deviation are known. We then compute the minimum sample sizes for $(P, 1 - \alpha)$ tolerance intervals to ensure that they are within the given specification limits. Without loss of generality, we assume that the specification limits are symmetric and that the underlying distribution is standard normal. We considered levels of $(1 - \alpha)$ and $P$ from the set {0.90, 0.95, 0.99}—both for the one-sided and two-sided settings—and considered values of the specification limits between 2.75 and 6 standard deviations from the mean of 0.

Figure 1 summarizes these results, where the black
lines and gray lines correspond to the one-sided and two-sided settings, respectively. As to be expected, the sample size increases when \((1 - \alpha)\) and \(P\) increase. Moreover, the sample size increases as the specification limits become narrower relative to the inherent variability in the data. By standardizing the data and specification limits, Figure 1 assists in identifying the general range of \(n^*\) values needed for a particular data problem. Figure 1 also shows a setting in which the historical data values needed for a particular data problem. Figure 1 assists in identifying the general range of \(n^*\) values needed for a particular data problem. Figure 1 also shows a setting in which the historical data indicate that \(n^*\) can reach small values. For example, in the \((0.90, 0.90)\) setting, \(n^* \leq 5\) when the mean is a little more than four standard deviations from the specification limit. However, because the purpose of a tolerance interval is to describe a certain proportion of the sampled population, it is not recommended to choose such a small value. While there is no formal criterion of an absolute minimum sample size, we recommend that \(n^* \geq 7\).

We next ran a numerical study to demonstrate the efficacy of our modified approach when varying significance levels, content levels, and specification limits. Without loss of generality, we consider a process that follows a standard normal distribution. For each combination of

- \(n \in \{20, 50, 100\},\)
- \((P, 1 - \alpha) \in \{(0.95, 0.90), (0.90, 0.95), (0.95, 0.95)\}); and
- specification limits in \([±3, ±4]\),

we compute both one-sided and two-sided tolerance intervals for the following three methods:

1. classical tolerance interval where \(\tilde{\mu} = 0\) and \(\tilde{\sigma} = 1;\)
2. Bayesian tolerance interval using the conjugate prior where \(\mu_0 = 0, \sigma_0 = 1,\) and \(m_0 = n_0 = n/10;\) and
3. Bayesian tolerance interval using the conjugate prior where \(\mu_0 = 0, \sigma_0 = 1,\) and \(m_0 = n_0 = n.\)

We simulate a sample of size \(n\) from a standard normal distribution and then calculate the values of \(\delta, P',\) and \(n^*\) for each of the three methods given above by applying [27] and [28] to the Faulkenberry-Weeks approach. Letting \(n^*_C, n^*_B1,\) and \(n^*_B2\) denote the sample sizes for the three methods, we simulate three sets of \(B = 1000\) samples using these sample sizes from a standard normal distribution. We then compute the \((P, 1 - \alpha)\) tolerance interval under the respective setting and calculate the proportion of the tolerance intervals within the specification limits, which we call \(p_C, p_{B1},\) and \(p_{B2}.\) We repeat this \(M = 1000\) times and retain the sample sizes \(n^*_C, n^*_B1,\) and \(n^*_B2\) and the proportions \(p_C, p_{B1},\) and \(p_{B2}.\) Note that we consider only the one-sided upper tolerance interval setting since results for the one-sided lower tolerance interval setting are similar.

The numerical results for the one-sided setting are presented in Table 1. As expected, for given levels of \(P, \alpha\), and \(n,\) the average values of the minimum sample sizes decrease when increasing the specification limits from \(±3\) to \(±4.\) Overall, the average proportion of times the tolerance intervals are within the

<table>
<thead>
<tr>
<th>((P, 1 - \alpha))</th>
<th>(90/95)</th>
<th>(95/90)</th>
<th>(95/95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specification Limits</td>
<td>(±3)</td>
<td>(±4)</td>
<td>(±3)</td>
</tr>
<tr>
<td>Classical TIs</td>
<td>(n = 20)</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9628)</td>
<td>(0.9564)</td>
</tr>
<tr>
<td></td>
<td>(n = 50)</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9716)</td>
<td>(0.9661)</td>
</tr>
<tr>
<td></td>
<td>(n = 100)</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9794)</td>
<td>(0.9744)</td>
</tr>
<tr>
<td>Bayesian TIs (#1)</td>
<td>(n = 20)</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9801)</td>
<td>(0.9904)</td>
</tr>
<tr>
<td></td>
<td>(n = 50)</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9945)</td>
<td>(0.9995)</td>
</tr>
<tr>
<td></td>
<td>(n = 100)</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9994)</td>
<td>(1.0000)</td>
</tr>
<tr>
<td>Bayesian TIs (#2)</td>
<td>(n = 20)</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.0000)</td>
<td>(1.0000)</td>
</tr>
<tr>
<td></td>
<td>(n = 50)</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.0000)</td>
<td>(1.0000)</td>
</tr>
<tr>
<td></td>
<td>(n = 100)</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.0000)</td>
<td>(1.0000)</td>
</tr>
</tbody>
</table>
specification limits are all very good for the conditions considered. The lowest proportions occur for the classical tolerance interval setting when \( n = 20 \), but as \( n \) increases, so does the average of the proportions. For the two Bayesian tolerance interval settings, using \( m_0 = n_0 = n \) yields overall better average proportions. This makes sense since we are using prior hyperparameters that reflect how we generated the underlying data, i.e., from a standard normal. Thus, the larger prior precision of \( n \) compared to \( n/10 \) yields better performance in terms of the average proportions.

The numerical results for the two-sided setting are presented in Table 2. The overall results are similar to those in the one-sided setting. However, the average minimum sample size in the two-sided setting is always higher than in the corresponding one-sided setting. This, too, is expected as typically a larger sample size is required when going from a one-sided testing scenario to a two-sided testing scenario. For example, the minimum sample size to achieve a target power in a two-sided \( t \)-test is larger than the minimum sample size for the corresponding one-sided \( t \)-test under identical conditions. The results we present here illustrate the accuracy one can expect for a given current sample size of \( n \).

Overall, the numerical results demonstrate the robustness of our procedure when specifying \( \delta \) and \( P' \). Whether calculating classical or Bayesian tolerance intervals, our strategy provides an appropriate sample size \( n^* \) such that the tolerance intervals constructed using that future sample will almost always be within the specification limits of the process. Thus, we have provided a more informed strategy for specifying \( \delta \) and \( P' \) over traditional rule-of-thumb values by leveraging parameter estimates from representative data as well as the likely spread of our data relative to established specification limits.

Analysis of medical device electrical performance data

The hermetically sealed electrical circuitry of the VNS Therapy system’s implantable pulse generator is comprised of many electrical components, such as resistors and capacitors, that cause small variations in electrical performance from device to device. Many aspects of the electrical performance of the generator must be verified by Cyberonics prior to submission to the FDA.

### Table 2

Results of the numerical study for the two-sided tolerance interval setting. Each cell gives the average sample size (\( n^*_{C}, n^*_{B1}, \) and \( n^*_{B2} \) respectively) rounded to the nearest integer. The average proportion of times the tolerance intervals were within the given specification limits is in parentheses.

<table>
<thead>
<tr>
<th>(( P, 1 - \alpha ))</th>
<th>Specification Limits</th>
<th>90/95</th>
<th>95/90</th>
<th>95/95</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \pm 3 )</td>
<td>( \pm 4 )</td>
<td>( \pm 3 )</td>
<td>( \pm 4 )</td>
</tr>
<tr>
<td>Classical TIs</td>
<td>( n = 20 )</td>
<td>29</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.8013)</td>
<td>(0.9797)</td>
<td>(0.8846)</td>
</tr>
<tr>
<td></td>
<td>( n = 50 )</td>
<td>32</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9654)</td>
<td>(0.9973)</td>
<td>(0.9853)</td>
</tr>
<tr>
<td></td>
<td>( n = 100 )</td>
<td>34</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9984)</td>
<td>(0.9993)</td>
<td>(0.9961)</td>
</tr>
<tr>
<td>Bayesian TIs (#1)</td>
<td>( n = 20 )</td>
<td>30</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.8532)</td>
<td>(0.9932)</td>
<td>(0.9303)</td>
</tr>
<tr>
<td></td>
<td>( n = 50 )</td>
<td>32</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9856)</td>
<td>(0.9999)</td>
<td>(0.9974)</td>
</tr>
<tr>
<td></td>
<td>( n = 100 )</td>
<td>34</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9988)</td>
<td>(1.0000)</td>
<td>(0.9999)</td>
</tr>
<tr>
<td>Bayesian TIs (#2)</td>
<td>( n = 20 )</td>
<td>30</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.9959)</td>
<td>(1.0000)</td>
<td>(0.9999)</td>
</tr>
<tr>
<td></td>
<td>( n = 50 )</td>
<td>33</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.0000)</td>
<td>(1.0000)</td>
<td>(1.0000)</td>
</tr>
<tr>
<td></td>
<td>( n = 100 )</td>
<td>35</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.0000)</td>
<td>(1.0000)</td>
<td>(1.0000)</td>
</tr>
</tbody>
</table>
and subsequent FDA regulatory approval of the device. We focus on one critical electrical measurement that has a nominal value of 3 units and an acceptable tolerance of ±10%, i.e., the lower and upper specification limits are 2.7 and 3.3, respectively. Note that the exact type of units are not reported due to confidentiality issues.

Considerable historical data is available for this attribute under test from previously approved predicate pulse generator models manufactured by Cyberonics, as well as from characterization data on the design under development. The required sample size, analysis method, and acceptance criteria must be established in a protocol prior to conducting the final design verification study in order to meet FDA design control requirements, follow internal procedures, and maintain statistical hypothesis requirements, follow internal procedures, and maintain statistical hypothesis α-levels. Tolerance intervals are a suitable method to analyze the device verification data to demonstrate that the product design conforms to its specification at an acceptable rate at a given confidence level. The required sample size for constructing a future \( (P, 1 - \alpha) \) tolerance interval can be determined using the specification limits and the parameter estimates from our predicate data.

For our example, we have data on the electrical measurement for \( n = 59 \) previously manufactured components on a previous device design. This comprises our historical data, which due to the device design similarity is expected to be representative of the new device design under test. We also have an additional set of \( n = 30 \) units of the new device design that were developed as part of a pilot study to characterize the device performance prior to conducting a formal design verification test required for regulatory submission.

Normality is an appropriate assumption for each of the historical and characterization datasets since the Shapiro-Wilk test \( p \)-values are 0.3861 and 0.3562, respectively. Common variance across the two datasets is an appropriate assumption as Levene’s test yields a \( p \)-value of 0.3657. Finally, a two-sample \( t \)-test (assuming equal variances) using these datasets shows that their means are not significantly different, with a \( p \)-value of 0.5084 and 95% confidence interval of \((-0.02, 0.03)\).

Using the historical dataset of \( n = 59 \) previously manufactured components, the sample mean is 2.9978 and the sample standard deviation is 0.0523, which we use as our values for \( \bar{\mu} \) and \( \sigma \), respectively. We consider both the classical and Bayesian settings. For the Bayesian setting, \( \bar{\mu} \) and \( \sigma \) are the hyperparameters for the mean and standard deviation, respectively, and we set \( m_0 = n_0 = 30 \), where 30 is the sample size of the characterization dataset. We are interested in determining the sample size for \( (P, 1 - \alpha) \) normal tolerance intervals where \( \alpha = 0.05 \) and \( P \in \{0.90, 0.95, 0.99\} \). Using our strategy, the value of \( P' \) is

\[
P' = \Phi\left(\frac{3.3 - 2.9978}{0.0523}\right) - \Phi\left(\frac{2.7 - 2.9978}{0.0523}\right) 
\approx 1 - (9.82e^{-9}), \tag{29}
\]

which is very close to 1 since our data are well within the specification limits. For the different combinations of \( P \) and \( \alpha \), we use the value of \( P' \) in \([29]\) and calculate the RACE value of \( \delta \) using \((27)\) to determine the necessary sample size \( n^* \).

Table 3 summarizes the values of \( n^* \) for constructing \( (P, 1 - \alpha) \) tolerance intervals from a future sample. This table is constructed for the setting of \( \alpha = 0.05 \) and the common values of \( P \in \{0.90, 0.95, 0.99\} \). For each value of \( P \), there are three separate columns. The first column gives values of \( \delta \), of which the first two rows are those calculated using the RACE criterion under the classical and Bayesian settings, and the last three rows are the common rule-of-thumb values. The second column (\( P'_1 \)) gives the value of \( n^* \) for

<table>
<thead>
<tr>
<th>Method</th>
<th>( P = 0.90 )</th>
<th>( P = 0.95 )</th>
<th>( P = 0.99 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta )</td>
<td>( P'_1 )</td>
<td>( P'_2 )</td>
<td>( P'_1 )</td>
</tr>
<tr>
<td>Classical</td>
<td>0.0905</td>
<td>6</td>
<td>150</td>
</tr>
<tr>
<td>Bayesian</td>
<td>0.0682</td>
<td>6</td>
<td>164</td>
</tr>
<tr>
<td>Rule-of-Thumb</td>
<td>0.01</td>
<td>7</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>6</td>
<td>145</td>
</tr>
</tbody>
</table>
Figure 2. (a) Plot of the minimum sample size ($n^*$) vs $P'$ for the values of $P \in \{0.90, 0.95, 0.99\}$ and $\alpha = 0.05$. The three sets of curves are enhanced for a smaller subset of $P'$ values for the (b) 90/95, (c) 95/95, and (d) 99/99 settings. In each of the plots, the solid curves from top to bottom are for $\delta = 0.01, 0.05$, and 0.10, respectively. In the last three plots, the dashed curves are for the RACE values of $\delta$ in the Bayesian (top) and classical (bottom) settings, while the vertical line is the calculated value of $P'$. 

$P'$ calculated in [29]. The third column ($P_2'$) gives the value of $n^*$ for $P' = (1 + P)/2$. Clearly, the value of $P'$ calculated for our example yields a much smaller sample size for a future tolerance interval than using a criterion like $(1 + P)/2$. The smaller value from this more conservative criterion would require measurements from many more manufactured components to obtain a level of precision that is, perhaps, not necessary or cost-effective. We also note that the sample sizes for our value of $P'$ are similar for the five values of $\delta$. When using the value of $P' = (1 + P)/2$, there are greater differences between the values of $n^*$ across the five values of $\delta$. This further emphasizes the importance when choosing $P'$.

Figure 2a is a plot of the $n^*$ vs $P'$ for the values of $P \in \{0.90, 0.95, 0.99\}$ and $\alpha = 0.05$. Since $P' > P$, there is a separate group of curves starting just after each of the three levels of $P$. Clearly, this figure shows that as $\delta$ increases (i.e., less precision), then the minimum sample size decreases. Moreover, as $P$ increases, one will obviously need larger sample sizes since a larger proportion of the sampled population is to be captured.

Since the curves in Figure 2a are a bit difficult to discern for larger values of $P'$, we enhanced the far
right-hand portion of the plot for the three different levels of \( P \). Figure 2b is a plot of the required sample sizes for the 90/95 tolerance interval when \( P' > 0.990 \). In this setting, the RACE values of \( \delta \) for both the classical and Bayesian settings are between the rule-of-thumb values of 0.05 and 0.10. Hence, the curves are bound between the profiles for these two \( \delta \) values. Figure 2c is a plot of the required sample sizes for the 95/95 tolerance interval when \( P' > 0.9975 \). In this setting, the RACE values of \( \delta \) for both the classical and Bayesian settings are between the rule-of-thumb values of 0.01 and 0.05, with the Bayesian method being very close to 0.05. Finally, Figure 2d is a plot of the required sample sizes for the 99/95 tolerance interval when \( P' > 0.9975 \). In this setting, the RACE values of \( \delta \) are both very close to 0.05. Hence, the strong overlap for the plotted curves. All of these figures can be used to determine a future sample size for the next set of manufactured components to be measured, with clear tradeoffs in the levels of \( P' \) and \( \delta \) being illustrated. We note that in this particular application the RACE values of \( \delta \) under both the classical and Bayesian settings are somewhat similar. Moreover, \( \delta \) calculated under the Bayesian setting is always smaller than the \( \delta \) calculated under the classical setting. However, this is not a strict relationship that holds within the procedure. For example, the RACE values of \( \delta \) can differ substantially depending on the prior hyperparameters selected in the Bayesian setting.

For demonstrating the efficacy of our approach, we also have a verification dataset of size \( n = 30 \), which consists of the most recent set of components that Cyberonics has manufactured and measured. The data is ordered according to when they were manufactured. Using the RACE values of \( \delta \) and \( P' \) from [29], we computed the corresponding \((P, 1 - \alpha)\) tolerance intervals using the first \( n^* \) values of the verification data. These tolerance intervals are reported in Table 4. As we can see, all of the tolerance intervals are within the specification limits.

We note that the value of \( P' \) for our analysis is very close to 1. This is not detrimental to the overall procedure, but rather reflects that the data are very close to the specification limits since the sample mean is over five standard deviations from the limits. This highlights that if the user has considerable room with the data to meet the specification limits, then they can use a smaller sample size. If faced with a large \( P' \), they could modify the criterion to further guarantee that the future data to be collected yields \((P, 1 - \alpha)\) tolerance intervals within the specification limits. They could use a more conservative criterion, like \( P' = (1 + P)/2 \), or compute \( P' \) with respect to specification limits that are tighter than those required. For example, if the specification limits are \( \pm 10\% \) of the nominal level, then consider \( \pm 7.5\% \) of the nominal level.

### Discussion

In this article, we have proposed a strategy for informing precision levels used in the Faulkenberry-Weeks approach for sample size determination of normal \((P, 1 - \alpha)\) tolerance intervals. Typically, rule-of-thumb values are used for these precision levels. Our work appears to be the first attempt at developing an informed strategy for setting these levels. Our strategy uses both specification limits and parameter estimates based on representative historical data to choose \( P' \) and \( \delta \). Moreover, using historical data can lead to an economical sample size, allowing for a better use of resources. We demonstrated the efficacy of this approach when considering either classical or Bayesian tolerance intervals. The general approach was motivated by the need to apply a statistically valid rationale to sample size determination for verifying the electrical performance characteristics of a neuromodulation device. This article is intended to provide the basis for practical methods that practitioners can use to formulate statistically valid sampling plans that can

---

**Table 4.** This table gives the \((P, 1 - \alpha)\) tolerance intervals (for \( \alpha = 0.05 \)) constructed using the first \( n^* \) values of the validation dataset. For convenience, the value of \( n^* \) that was determined in Table 3 is reported again beneath each tolerance interval.

<table>
<thead>
<tr>
<th>( P = )</th>
<th>0.90</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
</table>
| Classical | \[
| &nbsp; | \( \text{2.8672, 3.2550} \) | \( \text{2.8694, 3.2522} \) | \( \text{2.8415, 3.2631} \) |
| &nbsp; | \( n^* = 6 \) | \( n^* = 7 \) | \( n^* = 14 \) |
| Bayesian  | \[
| &nbsp; | \( \text{2.8887, 3.1281} \) | \( \text{2.8700, 3.1486} \) | \( \text{2.8295, 3.2008} \) |
| &nbsp; | \( n^* = 6 \) | \( n^* = 8 \) | \( n^* = 14 \) |
be described in test protocols used for design verification and validation.

We note that the calculation of tolerance limits can be viewed in the context of process capability. As stated in Chapter 7 of Montgomery (2013), “specifications are not necessary to process capability analysis.” Our approach does not run contrary to this statement. For our application, we want the tolerance limits to be within the specification limits, but demonstrating that they are even closer to a nominal content level $P$ does not provide additional value. In the absence of any guidance on the choices of $P$ and $\delta$ for the Falukenberry-Weeks approach, we leverage what information we have available. This includes our historical data and specification limits. Again, we emphasize that our approach is suggested for those settings where there is no guidance (e.g., by internal protocols) on the levels to use for $P$ and $\delta$.

A limitation of this approach is the degree to which the historical data is representative of the population. If the historical data is truly pilot data, then there is not enough data to support statements about the stability of the process or product being manufactured. One-off type of applications, such as the construction of nuclear cores, do not fit well in the paradigm we presented as any assumptions will have to rely on previous data from another similar, but not identical, product. In such cases, the users may want to either consider additional conservatism in their assumptions or a larger sample size than what is determined using our approach. Regardless, subject-matter experts should be involved throughout such a decision-making process.

If the historical data are biased, then the impact on our procedure will depend on how biased the data are with respect to the specification limit. Data biased away from a specification limit will require a lower sample size because the calculation assumes there is enough room for variability before exceeding the specification limit. Data biased towards a specification limit will require a larger sample size because there needs to be greater precision to ensure the tolerance interval does not exceed the specification limit. However, the premise of our article is about having representative historical data. If the historical data are biased, then the process is not in statistical control and one should not be haphazardly applying statistical procedures like those in this article.

We note that the application that motivated this work involves tolerance intervals for a single univariate normal population. We did not cover sample size determination for simultaneous tolerance limits (intervals) for several normal populations. One possibility is to extend the modified approach that we presented while using a simultaneous tolerance limits (intervals) procedure like that developed in Mee (1990).

Although probably the most common scenario that arises in practice, sample size determination for tolerance intervals is not relegated to the normal case. For example, some engineering tests involve collecting data on failure times or maximal events—such as for durability of pacemaker leads, cardiac stents, or heart valves—in which case tolerance limits (intervals) for generalized extreme value distributions are appropriate; see Lawless (1975) and Bain and Engelhardt (1981) for their development and relevant applications. We note that our approach can be extended to any such univariate parametric setting. Frequentist tolerance intervals are available for many distributions (see the tolerance package (Young 2010) for some common distributions) and the approach for Bayesian tolerance intervals can be applied to any parametric distribution (see Chapter 11 of Krishnamoorthy and Mathew 2009). Also, the Faulkenberry-Weeks approach can be applied to any parametric distribution. Thus, our work can be used as a general recipe for sample size determination for tolerance intervals under any parametric distribution.

**About the authors**

Derek S. Young is an Assistant Professor of Statistics in the Department of Statistics at the University of Kentucky (UK). Before joining UK, he had seven years of experience working as a research mathematical statistician for the Naval Nuclear Propulsion Program and the United States Census Bureau. He has published numerous articles, mostly in the areas of statistical tolerance regions and mixture models. He is also the developer and maintainer for popular R packages in both of these areas. He is an Accredited Professional Statistician™ of the American Statistical Association. He received a B.S. degree in mathematics from the University of Michigan and M.S. and Ph.D. degrees in statistics from the Pennsylvania State University.

Charles M. Gordon is the Manager of Biostatistics at Cyberonics, Inc. In this and prior roles he has been responsible for writing and implementing procedures related to sample size determination including design verification, process validation, and test method qualification. Other interests include clinical study design and survival analysis. He received a B.S. degree in biomedical engineering and an M.S. degree in statistics, both from Texas A&M University.
Shihong Zhu recently graduated from the University of Kentucky with a Ph.D. in statistics. Before that, he obtained his master’s degree in mathematics from the Chinese Academy of Sciences. He is currently a Quantitative Analyst at Fifth Third Bank in Cincinnati, Ohio.

Bryan D. Olin, Ph.D. joined Cyberonics as Vice President, Quality in May 2009. In August 2009, Dr. Olin assumed interim responsibility for Clinical Affairs, and in November 2009, he was named Vice President, Clinical Affairs & Quality. He joined Cyberonics from Zeltiq Aesthetics, Inc., a privately held medical technology company in the San Francisco Bay Area, where he served as Sr. Director, Quality Assurance beginning in October 2007. Prior to Zeltiq Aesthetics, Dr. Olin was employed at the LifeScan and Cordis franchises of Johnson & Johnson from 1999–2007, holding several positions of increasing responsibility in quality assurance, statistics and clinical data management. Dr. Olin began his career with Procter & Gamble in 1993 after obtaining his Ph.D. in statistics from Iowa State University. Dr. Olin currently serves on the Board of Directors of both the Medical Device Innovation Consortium (www.mdic.org) and cerbomed GmbH (www.cerbomed.com).

Acknowledgments

We are thankful to two anonymous referees for their helpful comments, which greatly improved the quality of this article.

References


Appendix A: Two-sided Bayesian normal tolerance intervals

In order to derive the two-sided Bayesian normal tolerance intervals, we need the following lemma, which we only state. The proof of the lemma is part of Result 1.2.1 in Krishnamoorthy and Mathew (2009).

**Lemma A.1.** Let \( V \sim N(0, \tau) \) independently of \( W \sim \chi^2_m \) and let \( 0 < \alpha, P < 1 \).

(a) The factor \( k_1 \) that satisfies
\[
P_{V,W} \left( \Phi \left( V + k_1 \sqrt{W} \right) \geq P \right) = 1 - \alpha
\]
is given by
\[
k_1 = \sqrt{\tau} t_{m,1-\alpha} \left( z_\alpha / \sqrt{\tau} \right),
\]
where \( z_\alpha \) denotes the \( \alpha \)th quantile of a standard normal distribution and \( t_{m,\eta}(\delta) \) denotes the \( \eta \)th quantile of a noncentral \( t \) distribution with degrees of freedom \( m \) and noncentrality parameter \( \delta \).

(b) The factor \( k_2 \) that satisfies
\[
P_{V,W} \left( \Phi \left( V + k_2 \sqrt{W} \right) - \Phi \left( V - k_2 \sqrt{W} \right) \geq P \right) = 1 - \alpha
\]
is the solution to the integral equation
\[
\sqrt{\frac{2}{\pi \tau}} \int_0^\infty P_W \left( W \geq \chi^2_v(p(\delta)) \right) e^{-x^2/2} \, dx = 1 - \alpha,
\]
where \( \chi^2_v(\delta) \) denotes the \( \alpha \)th quantile of a noncentral \( \chi^2 \) distribution with degrees of freedom \( v \) and noncentrality parameter \( \delta \).

The limits \( b_L(x) < b_U(x) \) for a two-sided \((P, 1 - \alpha)\) Bayesian normal tolerance interval are chosen such that
\[
P_{\mu, \sigma^2} \left[ \Phi \left( \frac{\bar{x} - \mu}{\sigma} + k_2 \frac{q}{\sigma} \right) - \Phi \left( \frac{\bar{x} - \mu}{\sigma} - k_2 \frac{q}{\sigma} \right) \geq P \right] = 1 - \alpha,
\]
where the probability above is calculated with respect to the posterior distribution of \((\mu, \sigma^2)\) given in [14]. Using \( x \) and \( q^2 \) as defined in (16), we can set \( b_L(x) = \bar{x} - k_2 q \) and \( b_U(x) = \bar{x} + k_2 q \) for some \( k_2 > 0 \). This implies that
\[
P_{\mu, \sigma^2} \left[ \Phi \left( \frac{\bar{x} - \mu}{\sigma} + k_2 \frac{q}{\sigma} \right) - \Phi \left( \frac{\bar{x} - \mu}{\sigma} - k_2 \frac{q}{\sigma} \right) \right] = 1 - \alpha.
\]

Using the Bayesian normal tolerance interval set-up presented in the main text, we note that the posterior distributions
\[
\frac{\mu - \bar{x}}{\sigma} \sim N(0, (n_0 + n)^{-1}) \quad \text{and} \quad \frac{q^2}{\sigma^2} \sim \frac{\chi^2_{m_0+n-1}}{m_0 + n - 1}
\]
are independent of one another. By part (b) of the lemma, it follows that the two-sided \( k \)-factor for [31] is the solution to the integral equation given in [19].

Under the non-informative prior in [20], we have a joint posterior distribution for \((\mu, \sigma^2)\) similar to [14], but now where \( p(\mu|\sigma^2) \) and \( p(\sigma^2) \) are the distributions
\[
\mu|\sigma^2 \sim N(\bar{x}, \frac{\sigma^2}{n}) \quad \text{and} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1),
\]
respectively. The limits for a two-sided \((P, 1 - \alpha)\) Bayesian normal tolerance interval are found similarly to the conjugate prior setting. Specifically, we now set \( b_L(x) = \bar{x} - k_2 s \) and \( b_U(x) = \bar{x} + k_2 s \) for some \( k_2 > 0 \) and note that from (33), the posterior distributions
\[
\frac{\mu - \bar{x}}{\sigma} \sim N(0, n^{-1}) \quad \text{and} \quad \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n - 1}
\]
are independent of one another. Therefore, by part (b) of the lemma, the \( k \)-factor under the non-informative prior setting is the solution to the integral equation given in [11].
学霸图书馆

www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具