Journal of Statistical Software

August 2010, Volume 36, Issue 5. *http://www.jstatsoft.org/*

tolerance: An R Package for Estimating Tolerance Intervals

Derek S. Young
The Pennsylvania State University

Abstract

The **tolerance** package for R provides a set of functions for estimating and plotting tolerance limits. This package provides a wide-range of functions for estimating discrete and continuous tolerance intervals as well as for estimating regression tolerance intervals. An additional tool of the **tolerance** package is the plotting capability for the univariate and regression settings as well as for the multivariate normal setting. The **tolerance** package's capabilities are illustrated using simulated data sets. Formulas used for the estimation procedures are also presented.

Keywords: acceptance limit, coverage, k factor, nonlinear regression, nonparametric regression, tolerance interval.

1. Introduction

Tolerance intervals (also referred to as *statistical design limits*) provide limits where at least a certain proportion of the population (P) falls for a given confidence level ($1 - \alpha$). They are an important tool often utilized in areas such as engineering, manufacturing, and quality control. Some applications of tolerance intervals include:

- Controlling the manufacturing process and prediction of parameter variation of circuit boards for new computers.
- Establishing (design) limits of fuel per inch of plate surface in a reactor core with respect to the nominal level (see Burrows 1963).
- Establishing an upper limit on the number of items allowed to be missing when constructing an audit sampling plan.

The first two applications are examples of two-sided limits, where it is necessary to establish a lower and upper limit for the tolerance interval. The last application is an example where only an upper limit is necessary for the sampling plan since you can have no fewer than 0 missing items.

Tolerance intervals have been developed extensively in the literature for many distributions and settings. Burrows (1963) provides a general introduction to tolerance intervals and serves as a good starting point to understand the utility of tolerance intervals. Patel (1986) provides a review (which was fairly comprehensive at the time of publication) of tolerance intervals for many distributions as well as a discussion of their relation with confidence intervals for percentiles and prediction intervals. One caveat with Patel (1986) is that there are some inconsistencies with the notation used, so it is best to refer back to the primary sources when studying the formulas. Many of the references cited within Patel (1986) were used in the development of the tolerance intervals discussed here. Finally, Krishnamoorthy and Mathew (2009) provide one of the more detailed texts concerning the theory and application of statistical tolerance regions.

The structure of the data and the assumptions made affect the calculations of the desired tolerance intervals. For example:

- The structure of the data may be univariate or multivariate.
- The data may be assumed to follow a certain distribution or no distribution may be assumed at all.
- The data may consist of a response variable (dependent variable) which is assumed to be modeled as a function of one or more predictor variables (independent variables), thus requiring a regression model.

In this paper, we discuss a variety of tolerance intervals for such settings. We also present the **tolerance** package (Young 2010) within the R statistical environment (R Development Core Team 2010), which provides estimation and plotting capabilities of the tolerance intervals discussed. The newest released version of **tolerance** (version 0.2.2, as of this writing) is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=tolerance`. Help with installing R packages can be found by typing

```
R> help("install.packages")
```

Upon successfully downloading the **tolerance** package, it can then be loaded by typing

```
R> library("tolerance")
```

The user may then type `help(package = "tolerance")` to see a list of available functions.

2. Technical details

Suppose a continuous random variable X has probability density function $f_X(\cdot;\boldsymbol{\theta})$ with cumulative distribution function $F_X(\cdot;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters characterizing the assumed distribution. Then

$$C_X(L,U;\boldsymbol{\theta}) = F_X(U;\boldsymbol{\theta}) - F_X(L;\boldsymbol{\theta})$$

is the coverage of the two-sided interval $[L, U]$, where L and U are statistics calculated from the data such that $L < U$.

Like other statistical intervals, one-sided and two-sided tolerance intervals can be constructed. For the following definitions, $\Pr[\cdot]$ denotes the probability set function.

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower tolerance interval $[L, +\infty)$ requires finding the lower tolerance limit L such that

  $$\Pr[1 - F_X(L; \boldsymbol{\theta}) \geq P] \geq 1 - \alpha. \tag{1}$$

  If an upper bound on the data is specified, then $+\infty$ is replaced with the known value and the right-hand side of the interval is closed with a brace.

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ upper tolerance interval $(-\infty, U]$ requires finding the upper tolerance limit U such that

  $$\Pr[F_X(U; \boldsymbol{\theta}) \geq P] \geq 1 - \alpha. \tag{2}$$

  If a lower bound on the data is specified, then $-\infty$ is replaced with the known value and the left-hand side of the interval is closed with a brace.

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided tolerance interval $[L, U]$ requires finding L and U such that

  $$\Pr[C_X(L, U; \boldsymbol{\theta}) \geq P] \geq 1 - \alpha. \tag{3}$$

Tolerance intervals can also be constructed when X is a discrete random variable, where $f_X(\cdot; \boldsymbol{\theta})$ is now a probability mass function with corresponding cumulative distribution function $F_X(\cdot; \boldsymbol{\theta})$. The coverage of the two-sided interval $[L, U]$ for discrete X is given by

$$C_X(L, U; m, \boldsymbol{\theta}_L, \boldsymbol{\theta}_U) = F_X(U; m, \boldsymbol{\theta}_U) - F_X(L-1; m, \boldsymbol{\theta}_L),$$

where $\boldsymbol{\theta}_L$ and $\boldsymbol{\theta}_U$ are $100 \times (1-\alpha)\%$ lower and upper confidence bounds, respectively, for the parameter $\boldsymbol{\theta}$. Note that the coverage also depends on m, which denotes some future quantity of interest (e.g., future sample size or future period of time). The formulas of tolerance limits for discrete distributions also reflect these differences.

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower tolerance interval $[L, +\infty)$ for a future quantity m, requires finding the largest integer L such that

  $$\Pr[1 - F_X(L-1; m, \boldsymbol{\theta}_L) \geq P] \geq 1 - \alpha. \tag{4}$$

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ upper tolerance interval $(-\infty, U]$ for a future quantity m, requires finding the smallest integer U such that

  $$\Pr[F_X(U; m, \boldsymbol{\theta}_U) \geq P] \geq 1 - \alpha. \tag{5}$$

- A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided tolerance interval $[L, U]$ for a future quantity m, requires finding the appropriate integers L and U such that

  $$\Pr[C_X(L, U; m, \boldsymbol{\theta}_L, \boldsymbol{\theta}_U) \geq P] \geq 1 - \alpha. \tag{6}$$

In most practical applications, $1-\alpha$ and P are typically chosen from the set of values $\{0.90, 0.95, 0.99\}$ (see Krishnamoorthy and Mathew 2009).

The basic form of tolerance intervals is similar to other types of statistical intervals. Let $x_1, \ldots, x_n$ denote an independent and identically distributed sample from some distribution. Suppose that the mean μ and the standard deviation σ of the population are unknown (which is often the case). Tolerance intervals are usually computed using sample estimates of the mean and standard deviation ($\bar{x}$ and s, respectively) and have the form

$$\bar{x} \pm ks,$$

where the factor k accounts for sampling errors in $\bar{x}$ and s, the confidence level $1-\alpha$, and the population proportion of interest P. However, exact k factors are often not available (with the one-sided normal setting being one exception). Thus, most of the intervals discussed in this paper are considered "approximate" $[100 \times (1-\alpha)\%]/[100 \times P\%]$ tolerance intervals. For the majority of these approximations, it can only be stated that with *at least* $100 \times (1-\alpha)\%$ confidence that *at least* $100 \times P\%$ of the population falls within the tolerance limits calculated.

As noted in Guenther (1972) and Hahn and Meeker (1991), one-sided tolerance limits can be used to obtain "approximate" two-sided tolerance intervals by applying Bonferroni's inequality. The Bonferroni approximation will be applied to control the central $100 \times P\%$ of the sampled population while controlling both tails to achieve at least $100 \times (1-\alpha)\%$ confidence. Thus, $[100 \times (1-\alpha/2)\%]/[100 \times (P+1)/2\%]$ one-sided lower and upper tolerance limits will be calculated and used to approximate a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided tolerance interval. The interval is conservative and will only be applied where two-sided tolerance interval procedures are currently unavailable in the literature.

The following three sections are devoted to tolerance intervals for discrete distributions, continuous distributions, and regression settings. Each section consists of subsections devoted to a particular tolerance interval. Furthermore, each subsection will briefly describe the form of the tolerance interval, outline the arguments of the corresponding function in the **tolerance** package, and provide analysis of a simulated data set.

3. Tolerance intervals for discrete distributions

While the **tolerance** package consists mainly of tolerance interval functions for continuous distributions, there are a few discrete distributions that the package handles. There are tolerance interval functions for the binomial and Poisson distributions as well as a function for calculating acceptance limits in sampling plans, which is based on the hypergeometric distribution.

3.1. Binomial tolerance intervals

A company which manufactures light bulbs selects a random sample for inspection and the number of defective light bulbs is counted. Suppose the company ships a certain number of units in a package and they wish to claim with confidence level $(1-\alpha)$, that a proportion P of such packages contain no more than a calculated quantity of defective units. A one-sided upper binomial tolerance interval can be used to provide such an answer for the company.

A random variable X is binomially distributed if it has cumulative distribution function

$$F_X(x; n, p) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i},$$

for $x = 0, \ldots, n$ and $0 \leq p \leq 1$. In the framework of binomial tolerance intervals, X is considered the number of defective (or acceptable) units in a random sample of size n with proportion p of such units in the population. The tolerance limits are calculated using confidence bounds for p. However, interval estimation for a binomial proportion is actually very complex and there are numerous methods that can be employed. Brown, Cai, and DasGupta (2001) discuss various estimation methods, some of which are provided as options in `bintol.int`, the function for calculating binomial tolerance intervals. Further details on the following discussion can be found in Hahn and Chandra (1981).

The proportion of defective items p is estimated using x/n. Let $100\times(1-\alpha)\%$ lower and upper confidence bounds for p (p_L and p_U, respectively) be found by using one of the methods in Brown *et al.* (2001). Suppose that tolerance limits on the number of defective units in future groups of size m is of interest. Then $[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper binomial tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided binomial tolerance interval are found according to equations (4), (5), and (6), respectively, such that $\boldsymbol{\theta}_L = p_L$ and $\boldsymbol{\theta}_U = p_U$.

Returning to the example, a random sample of $n = 1000$ light bulbs were selected for inspection, of which $x = 10$ of them failed (so an estimate of p would be given by $\hat{p} = 10/1000 = 0.01$). Suppose they ship packages of $m = 50$ units and the manufacturer wishes to claim, with 95% confidence, that 95% of such packages contain U or fewer defective units. Evaluation of a 95%/95% upper binomial tolerance limit is found by applying the `bintol.int` function:

```
R> bintol.int(x = 10, n = 1000, m = 50, alpha = 0.05, P = 0.95, side = 1,
+    method = "LS")
```

```
  alpha    P p.hat 1-sided.lower 1-sided.upper
1  0.05 0.95  0.01             0             2
```

The output for this example gives α, P, $\hat{p}$, and both one-sided tolerance limits (even though in this example only the upper limit is requested). So with 95% confidence, 95% of the future packages of size 50 will contain no more than 2 defective light bulbs.

As in many of the functions we discuss, `bintol.int` has the argument `side`, which can be set to `1` or `2` for calculating one-sided or two-sided tolerance limits, respectively. `bintol.int` also includes the argument `method`, which specifies the method to use for calculating the confidence bounds p_L and p_U. The available methods are

- `"LS"`: The large-sample method (the default method), which is appropriate when the sample size is large (e.g., $n \geq 50$) and $n\hat{p}$ and $n(1-\hat{p})$ are both ≥ 10. Otherwise, the coverage may not achieve the nominal level.

- `"WS"`: Wilson's method, which is appropriate even when the sample size is small (e.g., $n \leq 40$). The coverage probability fluctuates acceptably near the nominal level provided p does not approach 0 or 1.

- `"AC"`: The Agresti-Coull method, which is appropriate when the sample size is large (e.g., $n \geq 40$). This method is comparable to Wilson's method for large n.

- `"JF"`: Jeffreys' method, which is a Bayesian approach to the estimation. For this method, the prior distribution for p is assumed to be a beta distribution with parameters `a1` and `a2`, both of which must be specified by the user.

- `"CP"`: The Clopper-Pearson method (sometimes referred to as an "exact" procedure due to its derivation from the binomial distribution), which provides a more conservative interval and can be much larger than the nominal level as $n \to \infty$.

- `"AS"`: The arcsine method, which is appropriate when p is not too close to 0 or 1. Otherwise, the coverage approaches 0.

- `"LO"`: The logit method, which is appropriate when p is not too close to 0 or 1, but yields a more conservative interval.

Further details on these methods regarding interval estimation for a binomial proportion, including motivation and context, are described in Brown *et al.* (2001).

3.2. Poisson tolerance intervals

Suppose a company wishes to state with confidence level $(1-\alpha)$, the maximum number of unscheduled shutdowns that can be expected to occur in a proportion P of their systems over a period of time. Historical data over the past few months is available to determine the recent rate of unscheduled shutdowns for their systems. This historical data can be used to calculate a one-sided upper Poisson tolerance interval, which can provide such an answer for the company.

A random variable X is Poisson distributed if it has cumulative distribution function

$$F_X(x; n, \lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

for $x = 0, 1, \ldots$ and where $\lambda \geq 0$ is a mean (or rate) parameter. In the framework of Poisson tolerance intervals, X is considered the number of randomly occurring events in n units of time with a mean occurrence of λ. The Poisson tolerance intervals provide bounds for the distribution of the number of occurrences in a specified future time period of length m. These are calculated using confidence bounds for λ, similar to the binomial setting.

The mean occurrence of events λ is estimated using x/n. Let λ_L and λ_U be $100 \times (1-\alpha)\%$ lower and upper confidence bounds for λ, respectively. Suppose the number of events that occur in a future time period of length m is of interest. Then $[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper Poisson tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided Poisson tolerance interval are found according to equations (4), (5), and (6), respectively, such that $\boldsymbol{\theta}_L = \lambda_L$ and $\boldsymbol{\theta}_U = \lambda_U$.

We present two methods for finding λ_L and λ_U as discussed in Hahn and Meeker (1991). The first method is the tabular method, which for one-sided limits has the form

$$\lambda_L = \frac{\chi^2_{2x;\alpha}}{2n}$$
$$\lambda_U = \frac{\chi^2_{2x+2;1-\alpha}}{2n},$$

where n is the sample size and $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$-th quantile of a χ^2 distribution with d degrees of freedom. The second method is based on large sample theory and has the form

$$\lambda_L = \hat{\lambda} - z_{1-\alpha}\sqrt{\frac{\hat{\lambda}}{n}}$$
$$\lambda_U = \hat{\lambda} + z_{1-\alpha}\sqrt{\frac{\hat{\lambda}}{n}},$$

which is quite good and usually preferred for $x > 20$ (see Hahn and Meeker 1991). For both methods in the two-sided setting, simply replace α by $\alpha/2$ and P by $(P+1)/2$ in the procedure outlined above. Further details on Poisson tolerance intervals can be found in Hahn and Chandra (1981).

Returning to the example, suppose the company needs to be 95% confident on the maximum number of unscheduled shutdowns that can be expected to occur in 90% of their systems over a period of $m = 3$ months. In the past $n = 9$ months, there were a total of $x = 45$ shutdowns. Evaluation of a 95%/90% upper Poisson tolerance limit is found by applying the `poistol.int` function:

```
R> poistol.int(x = 45, n = 9, m = 3, alpha = 0.05, P = 0.90, side = 1,
+    method = "LS")

  alpha   P lambda.hat 1-sided.lower 1-sided.upper
1  0.05 0.9          5             7            24
```

Notice that the `poistol.int` function also includes the argument `method`, which determines how to calculate the confidence bounds λ_L and λ_U. The user can specify either `"TAB"` for the tabular method or `"LS"` for the large-sample method.

The output for this example yields α, P, $\hat{\lambda}$, and both one-sided tolerance limits (even though in this example we are only interested in the upper limit). So, the company can state with 95% confidence that approximately 24 shutdowns will occur in 90% of their systems within the next 3 months.

3.3. Acceptance limits for sampling plans

Suppose an inventory checklist states that a factory should have a certain number of plastic shipping containers in their possession. An auditor can afford to sample only a select number of the containers in the warehouse. The auditor's guidelines state that an acceptance limit must be established so that with confidence level $(1-\alpha)$, a proportion P of the containers must be accounted for or else a census will be performed.

Once the acceptance limit is determined, then the auditor must also establish the producer's risk and the consumer's risk. The producer's risk is the probability of rejecting an audit of a good inventory (i.e., a type I error). The calculation of this probability depends on a specified acceptable quality level (AQL). If the sampling plan were to be repeated numerous times as a process, then the AQL is the proportion of missing items considered acceptable from the *process* on a whole. The consumer's risk is the probability of accepting an audit of a bad inventory (i.e., a type II error). The calculation of this probability depends on a specified rejectable quality level (RQL), which is the proportion of missing items in a *sample* the auditor is willing to tolerate. Note that $0 < \text{AQL} < \text{RQL} < 1$. Further details on acceptance sampling can be found in Montgomery (2005).

Acceptance limits for sampling plans can roughly be viewed as upper hypergeometric tolerance limits. Sampling plans are often used for performing inventories at a warehouse or facility, since it is often not desirable (or even feasible) to do a census on an entire inventory. A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ *acceptance limit* specifies with $100 \times (1-\alpha)\%$ confidence that at least $100 \times P\%$ of the entire inventory population can be located. This upper limit is determined when given a sample size n which needs to be drawn for inventory from a specified inventory of size N, where N is considered a known population quantity. This limit is found using the hypergeometric distribution.

A random variable X is hypergeometrically distributed if it has cumulative distribution function

$$F_X(x; n, D, N) = \sum_{i=0}^{x} \frac{\binom{D}{i}\binom{N-D}{n-i}}{\binom{N}{n}}.$$

A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ acceptance limit is given by

$$F_X(x; n, D, N) \leq \alpha.$$

The following quantities are used in the above formulas:

- $F_X(x; n, D, N)$ = the probability of accepting the inventory;
- D = the number of missing items in the inventory ($D = \lfloor N \times 100(1-P)/P \rfloor$) such that $\lfloor x \rfloor$ is the next lowest integer value;
- P = the proportion of items in the inventory which are accountable;
- x = the upper acceptance limit;
- i = the number of unaccountable items in the sample;
- $1-\alpha$ = the confidence level used when bounding $F_X(x; n, D, N)$.

Returning to the example, suppose an inventory checklist says that a factory should have $N = 960$ plastic shipping containers in their possession. An auditor can afford to sample $n = 450$ of the containers in the warehouse. The auditor's guidelines state that with 90% confidence that 90% of the containers must be accounted for. The guidelines also stipulate that an AQL of 0.07 and an RQL of 0.10 must be used. The acceptance limit for this situation is found by implementing the `acc.samp` function:

```
R> acc.samp(n = 450, N = 960, alpha = 0.10, P = 0.90, AQL = 0.07, RQL = 0.10)
```

```
acceptance.limit  38.0000
lot.size         960.0000
confidence         0.9000
P                  0.9000
AQL                0.0700
RQL                0.1000
sample.size      450.0000
prod.risk          0.0359
cons.risk          0.0802
```

The interpretation of the output for this example is that the auditor can be 90% confident that at least 90% of the containers are still accountable if no more than 38 containers from the sample are actually missing. In fact, the consumer's risk (which gives the actual confidence level of this sampling plan) shows that the auditor's confidence level is $100 \times (1 - 0.0802)\% = 91.98\%$. Further details on the quantities reported in the output can be found by typing `help("acc.samp")`.

4. Tolerance intervals for continuous distributions

Tolerance intervals for continuous data can be displayed using a control chart, histogram, or both, while tolerance regions for the multivariate normal setting can be overlayed on a scatterplot of the sample data. Plots for such settings can be constructed using the function called `plottol`. Details on the various plotting options can be obtained by typing `help("plottol")`. In the examples that follow, we highlight some of the arguments available with this function.

4.1. Cauchy tolerance intervals

Consider a physical situation involving resonant systems. A resonator (such as a pipe organ, laser rod, or quartz crystal) is a function of the measure of the linewidth of the resonance frequency and the driving force for the intensity of the oscillations. The resonance frequencies of the oscillator of the system can be adequately modeled using a Cauchy distribution. Suppose a physicist takes resonance frequency measurements on the system and wishes to claim with confidence level $(1 - \alpha)$, that a proportion P of the resonance frequencies are between two hertz values. Calculation of a two-sided Cauchy tolerance interval can provide such an answer for the physicist.

A random variable X is Cauchy distributed (also called the Lorentz distribution) if it has cumulative distribution function

$$F_X(x; \theta, \sigma) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x - \theta}{\sigma}\right),$$

where $-\infty < x < +\infty$, $-\infty < \theta < +\infty$ is a location parameter (not the mean of the distribution since the mean of a Cauchy distribution does not exist), and $\sigma > 0$ is a scale parameter. Cauchy tolerance intervals require estimates of both parameters (call them $\hat{\theta}$ and $\hat{\sigma}$, respectively) which will be estimated using the sample data in the code below. Patel (1986) and Bain and Engelhardt (1991) both discuss Cauchy tolerance intervals and the estimation procedure in greater detail.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper Cauchy tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided Cauchy tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\theta}, \hat{\sigma})^\top$. The formulas given in Patel (1986) for estimating L and U in the one-sided setting are:

$$L = \hat{\theta} - k_{\alpha,P}\hat{\sigma}$$
$$U = \hat{\theta} + k_{\alpha,P}\hat{\sigma},$$

where

$$k_{\alpha,P} = \frac{z_{1-\alpha}}{\sqrt{n}}\sqrt{2 + 2[F_X^{-1}(1-P; \theta = 0, \sigma = 1)]^2} - F_X^{-1}(1-P; \theta = 0, \sigma = 1),$$

n is the sample size, and $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of a standard normal distribution. Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

Returning to the example, suppose the physicist wishes to claim, with 95% confidence, that 90% of the resonance frequencies are between L and U Hz. The data were simulated using a Cauchy distribution with $\theta = 1 \times 10^5$ (i.e., 100 MHz) and $\sigma = 10$. Estimation of a 95%/90% two-sided Cauchy tolerance interval is found by implementing the `cautol.int` function:

```
R> set.seed(100)
R> x <- rcauchy(1e3, 1e5, 10)
R> out <- cautol.int(x = x, alpha = 0.05, P = 0.90, side = 2)
R> out

  alpha   P 2-sided.lower 2-sided.upper
1  0.05 0.9      99929.52      100069.4
```

The output for this example yields α, P, and the two-sided tolerance limits. So, the physicist can be 95% confident that at least 90% of the resonance frequencies will be between 99.930 MHz and 100.069 MHz. Figure 1 gives a control chart and histogram of this data with the tolerance limits shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "two",
+    x.lab = "Cauchy Data", lty = 0)
```

Notice how the tolerance limits appear to be narrow due to the high probability of getting extreme observations in either tail when data is Cauchy distributed.

4.2. Exponential tolerance intervals

Consider the lifetime of small electric motors. Suppose these motors can be adequately modeled according to an exponential distribution. Suppose further that the manufacturer of the motors wishes to claim with confidence level $(1-\alpha)$, that a proportion P of the motors have a service life below a certain number of hours. Calculation of a one-sided upper exponential tolerance interval can provide such an answer for the manufacturer.

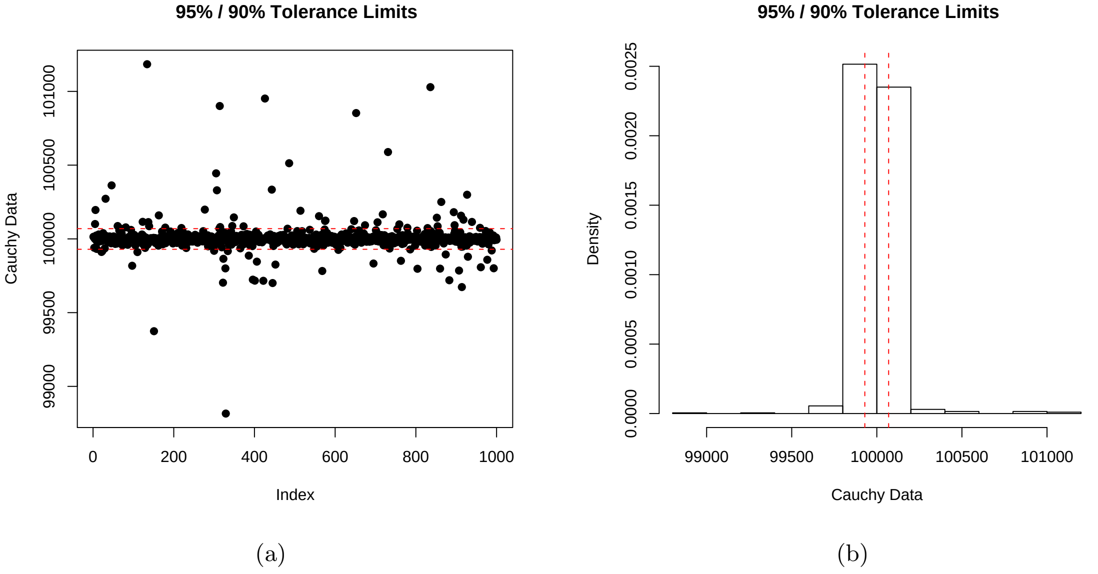


Figure 1: (a) Control chart and (b) histogram of the $n = 1000$ Cauchy distributed values. In both plots, the dashed red lines give the two-sided tolerance limits.

A random variable X is exponentially distributed if it has cumulative distribution function

$$F_X(x; \lambda) = 1 - e^{-\lambda x},$$

where $x > 0$ and $\lambda > 0$ is a scale parameter (sometimes called the rate). Exponential tolerance intervals require an estimate of this parameter ($\hat{\lambda}$) which can be obtained using maximum likelihood estimation. Blischke and Murthy (2000) discusses exponential tolerance intervals and the following estimation procedure in greater detail.

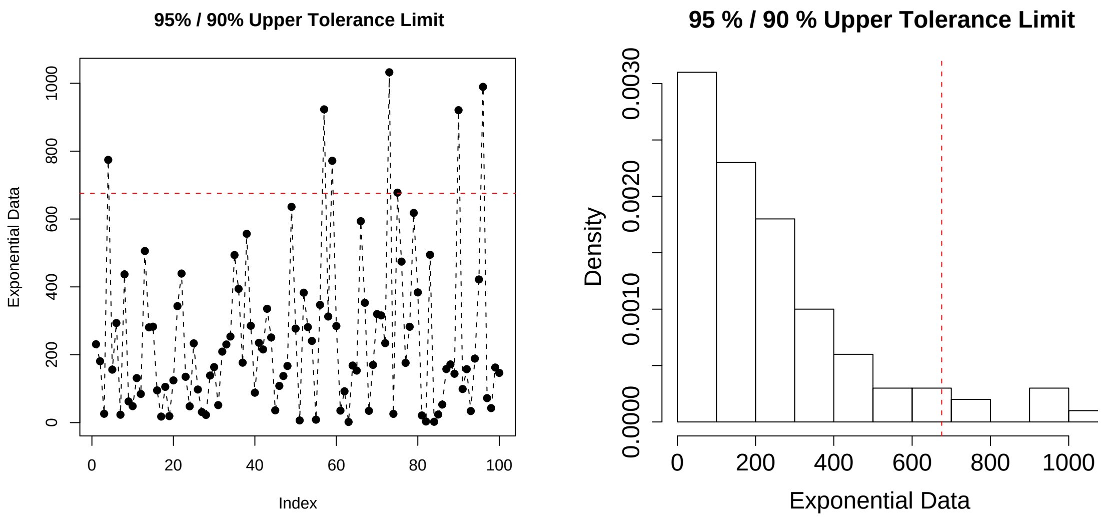
$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper exponential tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided exponential tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = \hat{\lambda}$. The formulas for estimating L and U in the one-sided setting are:

$$L = \frac{2n\hat{\lambda}\ln(1-P)}{\chi^2_{2n;1-\alpha}}$$

$$U = \frac{2n\hat{\lambda}\ln(P)}{\chi^2_{2n;1-\alpha}},$$

such that n is the sample size and $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$-th quantile of a χ^2 distribution with d degrees of freedom. Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

Returning to the example, suppose the manufacturer wishes to claim, with 95% confidence, that 90% of the motors have a service life below U hours. The data were simulated using an exponential distribution with $\lambda = 4 \times 10^{-3}$ (i.e., they have an average service life of 250 hours). Evaluation of a 95%/90% upper exponential tolerance interval is found by implementing the `exptol.int` function:



(a) (b)

Figure 2: (a) Control chart and (b) histogram of the $n = 100$ exponentially distributed values. In each plot, the dashed red line gives the one-sided upper tolerance limit.

```
R> set.seed(100)
R> x <- rexp(100, 0.004)
R> out <- exptol.int(x = x, alpha = 0.05, P = 0.90, side = 1, type.2 = FALSE)
R> out
```

```
  alpha   P lambda.hat 1-sided.lower 1-sided.upper
1  0.05 0.9    246.869      22.23152      675.5904
```

Notice that the `exptol.int` function also includes the argument `type.2`. This logical argument is used to indicate whether or not the data has type II censoring, which occurs when a measurement is taken on an object still in operation (or surviving) past the conclusion of the study. The tolerance interval formulas change slightly to accommodate type II censored data, in which case the user would specify `type.2 = TRUE`. See Patel (1986) for details.

The output for this example yields α, P, $\hat{\lambda}$, and both one-sided tolerance limits. So, the manufacturer can be 95% confident that at least 90% of the electric motors have a lifetime less than 675.6 hours. Figure 2 gives a control chart and histogram of this data with the upper tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "upper",
+    x.lab = "Exponential Data", lty = "dashed")
```

4.3. (2-Parameter) exponential tolerance intervals

Consider a situation involving the age of individuals who retire from a company. Suppose the retirees have an average retirement age past a certain threshold, which is the youngest age of a retiree assuming that the company does not have a required minimum age for when an individual may retire. The ages can be adequately modeled according to a 2-parameter exponential distribution. Suppose further that the company's human resources department wishes to state with confidence level $(1-\alpha)$, that a proportion P of the retirees will be below a certain age. Calculation of a one-sided upper 2-parameter exponential tolerance interval can provide such an answer for the human resources department.

A random variable X is distributed according to the 2-parameter exponential distribution if it has cumulative distribution function

$$F_X(x;\lambda,\mu) = 1 - e^{-\lambda(x-\mu)},$$

where $x > \mu$ such that μ is a location parameter and $\lambda > 0$ is a scale parameter (or rate). 2-parameter exponential tolerance intervals require estimates of both of these parameters ($\hat{\mu}$ and $\hat{\lambda}$, respectively) which can be obtained using maximum likelihood estimation. Notice that this is just a shifted version of the exponential distribution. Dunsmore (1978) and Guenther, Patil, and Uppuluri (1976) both discuss 2-parameter exponential tolerance intervals and the estimation procedure in greater detail. Engelhardt and Bain (1978) discusses how to modify the formulas when dealing with type II censored data.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper 2-parameter exponential tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided 2-parameter exponential tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\lambda}, \hat{\mu})^\top$. The formulas for estimating L and U in the one-sided setting (see Guenther *et al.* 1976) are:

$$L = \hat{\mu} + \frac{k_1}{\hat{\lambda}}$$
$$U = \hat{\mu} + \frac{k_2}{\hat{\lambda}},$$

where

$$k_1 \approx 1 - \left(\frac{P^n}{\alpha}\right)^{1/(n-1)},$$
$$k_2 \approx n\left(\frac{\chi^2_{2;P}}{\chi^2_{2n-2;P}}\right),$$

n is the sample size, and $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$-th quantile of a χ^2 distribution with d degrees of freedom. However, Dunsmore (1978) suggests that an improvement can be made on the upper bound by penalizing k_2 such that

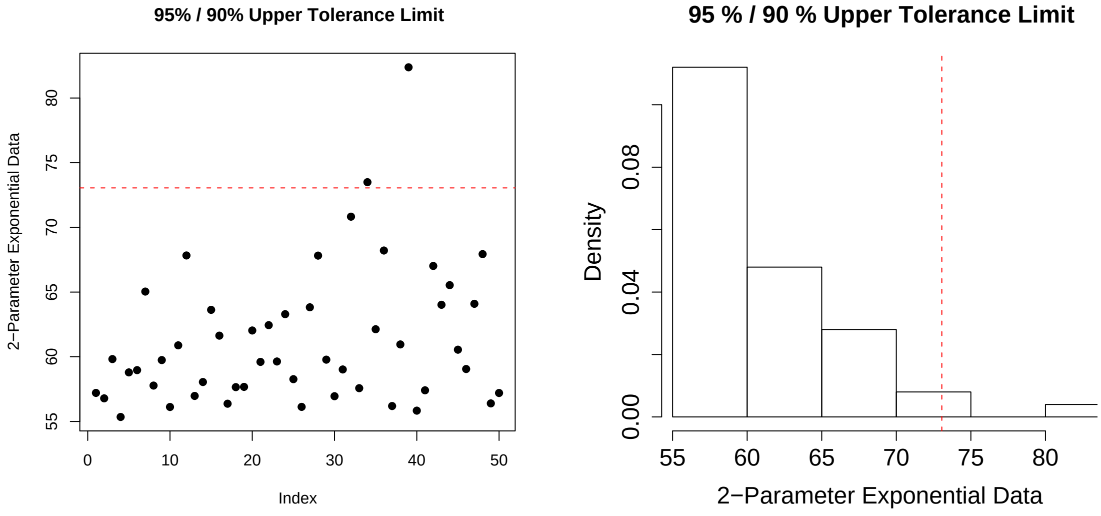
$$k_2^* \approx k_2 - n(\gamma/n)^{1.63+0.39\gamma},$$

where

$$\gamma \approx 1.71 + 1.57\ln(\ln(\alpha^{-1}))$$

Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

Returning to the example, suppose the human resources department wishes to state with 95% confidence that 90% of the retirees will be below a certain age. Data were simulated using a 2-parameter exponential distribution with $\lambda = 6$ and $\mu = 55$. Evaluation of a 95%/90% upper 2-parameter exponential tolerance interval is found by implementing the `exp2tol.int` function:



(a) (b)

Figure 3: (a) Control chart and (b) histogram of the $n = 50$ 2-parameter exponentially distributed values. In each plot, the dashed red line gives the one-sided upper tolerance limit.

```
R> set.seed(100)
R> x <- r2exp(50, rate = 6, shift = 55)
R> out <- exp2tol.int(x = x, alpha = 0.05, P = 0.90, side = 1,
+    method = "DUN", type.2 = FALSE)
R> out

  alpha   P 1-sided.lower 1-sided.upper
1  0.05 0.9      55.61514      73.05272
```

Note that the `exp2tol.int` function contains the argument `method`, which allows the user to specify which way to estimate k_2. The two options are `"GPU"` (see Guenther *et al.* 1976) and `"DUN"` (see Dunsmore 1978), where the latter has been shown to yield slightly improved estimates for $n \geq 8$. Also note in the code above that the data was simulated using the function `r2exp`, which performs random generation from the 2-parameter exponential distribution. Other functions are available in the **tolerance** package for the density, distribution function, and quantile function (type `help("TwoParExponential")` for more information).

The output for this example yields α, P, and both one-sided tolerance limits. So, the human resources department can be 95% confident that at least 90% of the retirees will be below the age of 73.1 years. Figure 3 gives a control chart and histogram of this data with the upper tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "upper",
+    x.lab = "2-Parameter Exponential Data", lty = 0)
```

4.4. Gamma tolerance intervals

An environmental engineer measures air quality using an index which considers various measurements from air monitoring data. Anything above an established value on this fictional scale indicates poor air quality. The daily measurements can be modeled appropriately with a gamma distribution. The engineer wishes to state with confidence level $(1-\alpha)$, that a certain proportion P of the days will be below a particular index measurement. Calculation of a one-sided upper gamma tolerance interval can provide such an answer for the engineer.

A random variable X is gamma distributed if it has cumulative distribution function

$$F_X(x;\theta,\beta) = \int_{t=0}^{x} \frac{t^{\theta-1}e^{-t/\beta}}{\beta^{\theta}\Gamma(\theta)}dt,$$

where $x > 0$, $\theta > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, and $\Gamma(\cdot)$ is the gamma function. Gamma tolerance intervals require estimates of both parameters ($\hat{\theta}$ and $\hat{\beta}$) which can be obtained using a nonlinear optimization routine to find the maximum likelihood estimates. Krishnamoorthy, Mathew, and Mukherjee (2008) estimate gamma tolerance intervals using a normal approximation, which is the method adopted here. Their simulation results show that this approximate procedure is very simple and provides accurate estimates.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper gamma tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided gamma tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\theta}, \hat{\beta})^{\top}$. Gamma tolerance intervals can be estimated using a normal approximation to the gamma distribution. As discussed in Krishnamoorthy *et al.* (2008), if X is a gamma random variable with shape parameter θ and scale parameter β, then $X^{1/3}$ is (approximately) a normal random variable with mean μ and variance σ^2 such that

$$\begin{aligned}\mu &= \frac{\beta^{1/3}\Gamma(\theta+1/3)}{\Gamma(\theta)}\\ \sigma^2 &= \frac{\beta^{2/3}\Gamma(\theta+2/3)}{\Gamma(\theta)} - \mu^2.\end{aligned}$$

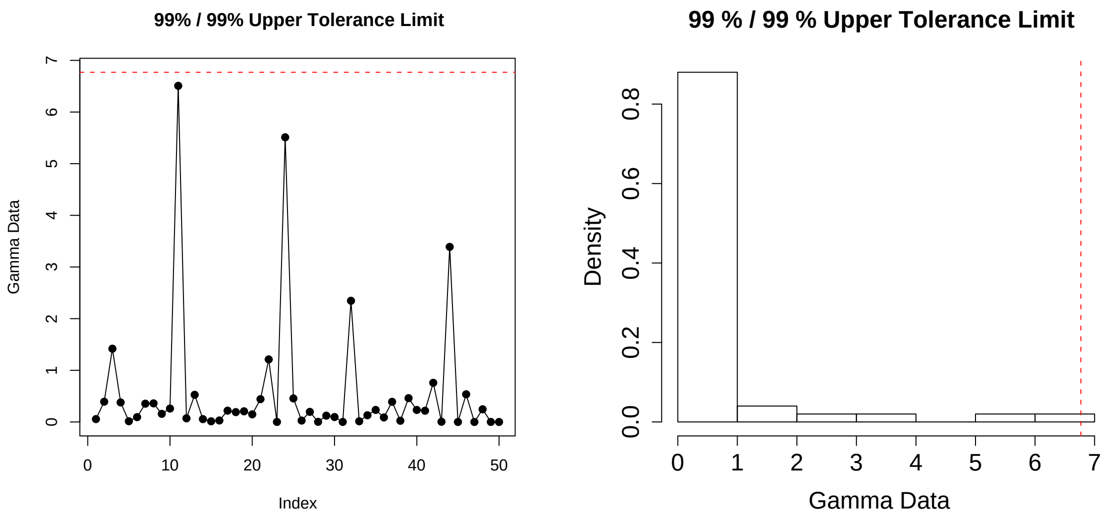
Estimates of the above parameters ($\hat{\mu}$ and $\hat{\sigma}^2$), are found by plugging in the estimates $\hat{\theta}$ and $\hat{\beta}$. The formulas for estimating L and U in the setting for normally distributed data are:

$$\begin{aligned}L_N &= \hat{\mu} - k\hat{\sigma}\\ U_N &= \hat{\mu} + k\hat{\sigma},\end{aligned}$$

where the functional form of k, for both the one-sided and two-sided settings, is discussed later in the normal tolerance intervals section. Once the above are calculated, it is then necessary to transform back to obtain the estimates of L and U for gamma tolerance intervals:

$$\begin{aligned}L &= L_N^3\\ U &= U_N^3.\end{aligned}$$

If X^* is a log-gamma distributed random variable, then $\ln(X^*) = X$ is a gamma distributed random variable. Thus, tolerance intervals can also be found for log-gamma data. This is



(a) (b)

Figure 4: (a) Control chart and (b) histogram of the $n = 50$ gamma distributed values. In each plot, the dashed red line gives the one-sided upper tolerance limit.

accomplished by the procedure outlined above, with the additional step of transforming back to the log-gamma scale by taking e^L and e^U for the lower and upper limits, respectively.

Returning to the example, suppose the engineer wishes to state with 99% confidence that 99% of the days will be below a certain index measurement. Data were simulated using a gamma distribution with $\theta = 0.30$ and $\beta = 2$. Evaluation of a 99%/99% upper gamma tolerance interval is found by implementing the `gamtol.int` function:

```
R> set.seed(100)
R> x <- rgamma(50, shape = 0.30, scale = 2)
R> out <- gamtol.int(x = x, alpha = 0.01, P = 0.99, side = 1,
+    method = "HE", log.gamma = FALSE)
R> out
```

```
  alpha    P 1-sided.lower 1-sided.upper
1  0.01 0.99             0      6.769559
```

Note that the `gamtol.int` function has the argument `method`, which allows the user to specify which way to estimate k. The two options are `"HE"` (see Howe 1969) and `"WBE"` (see Weissberg and Beatty 1969), both of which are discussed in greater detail in the section on normal tolerance intervals.

The output for this example yields α, P, and both one-sided tolerance limits. So, the engineer can be 99% confident that at least 99% of the days will have an air quality index measure below 6.8. Figure 4 gives a control chart and histogram of this data with the upper tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "upper",
+    x.lab = "Gamma Data")
```

4.5. Laplace tolerance intervals

Suppose the performance of an assembly line worker can be measured on a (continuous) scale from 0 to 100, where any performance score above 60 is considered acceptable by the company's standards. The performance scores of the workers are adequately modeled by a Laplace distribution. The quality engineer performing the study wishes to state with confidence level $(1-\alpha)$, that a proportion P of the workers will be above a certain score. Calculation of a one-sided lower Laplace tolerance interval can provide such an answer for the engineer.

A random variable X is Laplacian distributed if it has cumulative distribution function

$$F_X(x;\theta,\sigma) = \int_{t=-\infty}^{x} \frac{1}{2\sigma} e^{-\frac{|t-\theta|}{\sigma}}\, dt,$$

where $-\infty < x < +\infty$, $-\infty < \theta < +\infty$ is a location parameter, and $\sigma > 0$ is a scale parameter. Laplace tolerance intervals require estimates of both parameters ($\hat{\theta}$ and $\hat{\sigma}$) which can be obtained using maximum likelihood estimation. The Laplace distribution is also referred to as the double exponential distribution. Bain and Engelhardt (1973) discusses Laplace tolerance intervals and the following estimation procedure in greater detail.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper Laplace tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\theta}, \hat{\sigma})^\top$. The formulas for estimating L and U for Laplacian distributed data in the one-sided setting are:

$$L = \hat{\theta} - k\hat{\sigma}$$
$$U = \hat{\theta} + k\hat{\sigma},$$

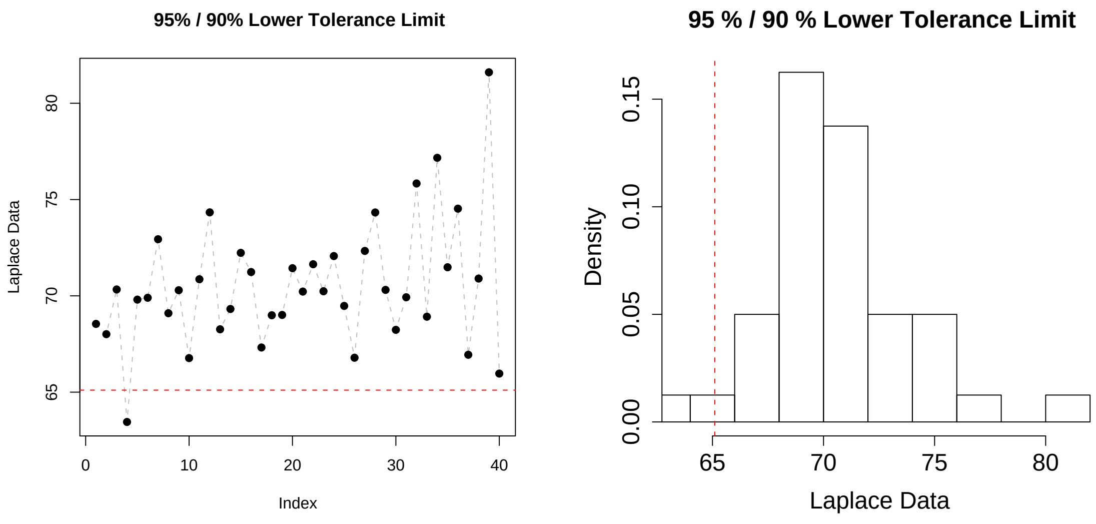
where

$$k \approx -nk_P + \frac{z_{1-\alpha}}{n - z_{1-\alpha}^2}\sqrt{n(1+k_P^2) - z_{1-\alpha}^2},$$

such that n is the sample size, $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of a standard normal distribution, and $k_P = \ln[2(1-P)]$. Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

Returning to the example, suppose the quality engineer performing the study wishes to state with 95% confidence that 90% of the workers will be above a certain score and utilizes the worker's previous year scores to construct a tolerance interval. Data were simulated using a Laplace distribution with $\theta = 70$ and $\sigma = 3$. Evaluation of a 95%/90% lower Laplace tolerance interval is found by implementing the `laptol.int` function:

```
R> set.seed(100)
R> tmp <- runif(40)
R> x <- rep(70, 40) - sign(tmp - 0.5)*rep(3, 40)*
+    log(2*ifelse(tmp < 0.5, tmp, 1 - tmp))
R> out <- laptol.int(x = x, alpha = 0.05, P = 0.90)
R> out
```



(a) (b)

Figure 5: (a) Control chart and (b) histogram of the $n = 40$ Laplacian distributed values. In each plot, the dashed red line gives the one-sided lower tolerance limit.

```
  alpha   P 1-sided.lower 1-sided.upper
1  0.05 0.9      65.10457      75.35777
```

The output for this example yields α, P, and both one-sided tolerance limits. So, the engineer can be 95% confident that at least 90% of the assembly line workers will have a performance score above 65.1. Figure 5 gives a control chart and histogram of this data with the lower tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "lower",
+    x.lab = "Laplace Data", lty = "dashed", col = "gray")
```

4.6. Logistic tolerance intervals

Suppose a researcher is studying the effects of a new drug on the sleeping habits of individuals with insomnia. The number of hours of sleep are adequately modeled by a logistic distribution. The researcher wishes to claim with confidence level $(1-\alpha)$, that a proportion P of the subjects get sleep above a certain number of hours. Calculation of a one-sided lower logistic tolerance interval can provide such an answer for the researcher.

A random variable X follows a logistic distribution if it has cumulative distribution function

$$F_X(x;\theta,\sigma) = \int_{t=-\infty}^{x} \left[1 + e^{-\frac{\pi(t-\theta)}{\sqrt{3\sigma^2}}}\right]^{-1} dt$$

where $-\infty < x < +\infty$, $-\infty < \theta < +\infty$ is a location parameter, and $\sigma > 0$ is a scale parameter. Logistic tolerance intervals require estimates of both parameters ($\hat{\theta}$ and $\hat{\sigma}$) which can be obtained using a nonlinear optimization routine to find the maximum likelihood estimates. Hall (1975) discusses one-sided logistic tolerance intervals and their properties in greater detail.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper logistic tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\theta}, \hat{\sigma})^\top$. The formulas for estimating L and U for logistically distributed data are:

$$
\begin{aligned}
L &= \hat{\theta} - k_1\hat{\sigma} \\
U &= \hat{\theta} + k_2\hat{\sigma},
\end{aligned}
$$

which are based on a normal approximation method (see Balakrishnan 1991). The estimates for L and U require additional calculations which require further optimization. In particular, both estimates rely on the variances of $\hat{\theta}$ and $\hat{\sigma}$ ($\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, respectively) and their covariance ($\hat{\sigma}_{1,2}$). These can be estimated using the Hessian matrix of the optimization algorithm to find $\hat{\theta}$ and $\hat{\sigma}$. The formulas for the estimates of k_1 and k_2 are

$$
\begin{aligned}
k_1 &\approx \frac{t_1 + \sqrt{t_1^2 - uv}}{v} \\
k_2 &\approx \frac{t_2 + \sqrt{t_2^2 - uv}}{v},
\end{aligned}
$$

such that

$$
\begin{aligned}
t_1 &= F_X^{-1}(P; \theta = 0, \sigma = 1) - \hat{\sigma}_{1,2} z_{1-\alpha}^2 \\
t_2 &= F_X^{-1}(P; \theta = 0, \sigma = 1) + \hat{\sigma}_{1,2} z_{1-\alpha}^2 \\
u &= [F_X^{-1}(P; \theta = 0, \sigma = 1)]^2 - \hat{\sigma}_1^2 z_{1-\alpha}^2 \\
v &= 1 - \hat{\sigma}_2^2 z_{1-\alpha}^2.
\end{aligned}
$$

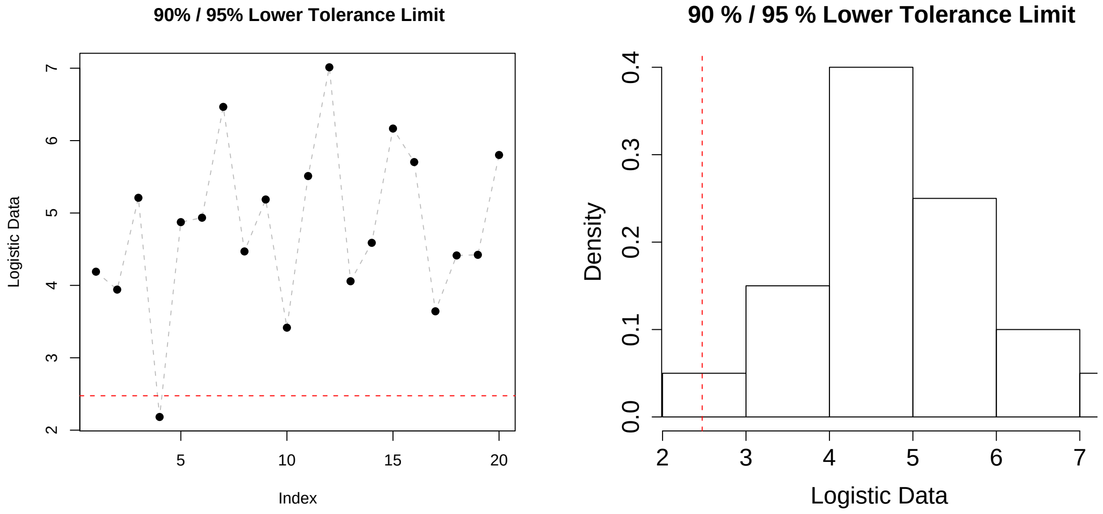
Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

It should be noted that Balakrishnan (1991) pointed out that the equation for k_2 in Hall (1975) was reported incorrectly with a positive sign under the radical. Further details on this method, as well as a discussion on two-sided logistic tolerance intervals not using the Bonferroni approximation, can be found in Balakrishnan (1991).

If X^* is a log-logistic random variable, then $\ln(X^*) = X$ is a logistic random variable. Thus, tolerance intervals can also be found for log-logistic data. This is accomplished by the procedure outlined above, with the addition of transforming back to the log-logistic scale by taking e^L and e^U for the lower and upper limits, respectively.

Returning to the example, suppose the researcher wishes to claim with 90% confidence that 95% of the subjects get sleep above a certain number of hours. Data were simulated using a logistic distribution with $\theta = 5$ and $\sigma = 1$. Evaluation of a 90%/95% lower logistic tolerance interval is found by implementing the `logistol.int` function:

```
R> set.seed(100)
R> x <- rlogis(20, 5, 1)
R> out <- logistol.int(x = x, alpha = 0.10, P = 0.95, log.log = FALSE)
R> out
```



(a) (b)

Figure 6: (a) Control chart and (b) histogram of the $n = 20$ logistically distributed values. In each plot, the dashed red line gives the one-sided lower tolerance limit.

```
  alpha    P 1-sided.lower 1-sided.upper
1   0.1 0.95      2.475273      7.143847
```

The output for this example yields α, P, and both one-sided tolerance limits. So, the researcher can be 90% confident that at least 95% of the subjects will get at least 4.1 hours of sleep per night. Figure 6 gives a control chart and histogram of this data with the lower tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "lower",
+    x.lab = "Logistic Data", lty = "dashed", col = "gray")
```

4.7. Nonparametric tolerance intervals

Consider the logistic tolerance interval example from earlier. The researcher still wishes to claim with confidence level $(1 - \alpha)$, that a proportion P of the subjects get sleep above a certain number of hours; however, now no distributional assumptions are made about the data. Calculation of a one-sided lower nonparametric tolerance interval can provide an answer for the researcher.

Occasionally, a researcher may not wish to make any distributional assumptions regarding their data. All that is assumed about the random sample $X_1 = x_1, \ldots, X_n = x_n$ is that the underlying distribution function F_X is a continuous, non-decreasing, probability distribution. The basic form of $[100 \times (1-\alpha)\%]/[100 \times P\%]$ upper and lower nonparametric tolerance limits is

$$L = x_{(r)}$$
$$U = x_{(s)}.$$

Here, $x_{(j)}$ corresponds to the j-th value from the ordered sequence of the original $x_1, \ldots, x_n$ values. Computing nonparametric tolerance intervals involves finding the appropriate r and s values, which is typically done using the beta distribution or the binomial distribution.

There are three methods the `nptol.int` function offers for computing r and s:

1. `"HM"`: The Hahn-Meeker method (see Hahn and Meeker 1991), which uses an estimate based on the binomial distribution; however, two intervals may be reported if an odd number of observations must be trimmed from both sides.
2. `"WILKS"`: The Wilks method (see Wilks 1941), which uses an estimate based on the beta distribution to omit a certain number of observations from either side. For the two-sided intervals, the tolerance intervals are symmetric about the center of the observed data.
3. `"WALD"`: The Wald method (see Wald 1943), which is the same as the Wilks method for the one-sided setting. For the two-sided setting, symmetry about the center of the observed data is not assumed and this method finds all possible tolerance intervals, each having at least the specified confidence level.

Continuing with the earlier example and the data generated earlier, the researcher still wishes to claim with 90% confidence that 95% of the subjects get sleep above a certain number of hours. Evaluation of a 90%/95% lower nonparametric tolerance interval is found by implementing the `nptol.int` function:

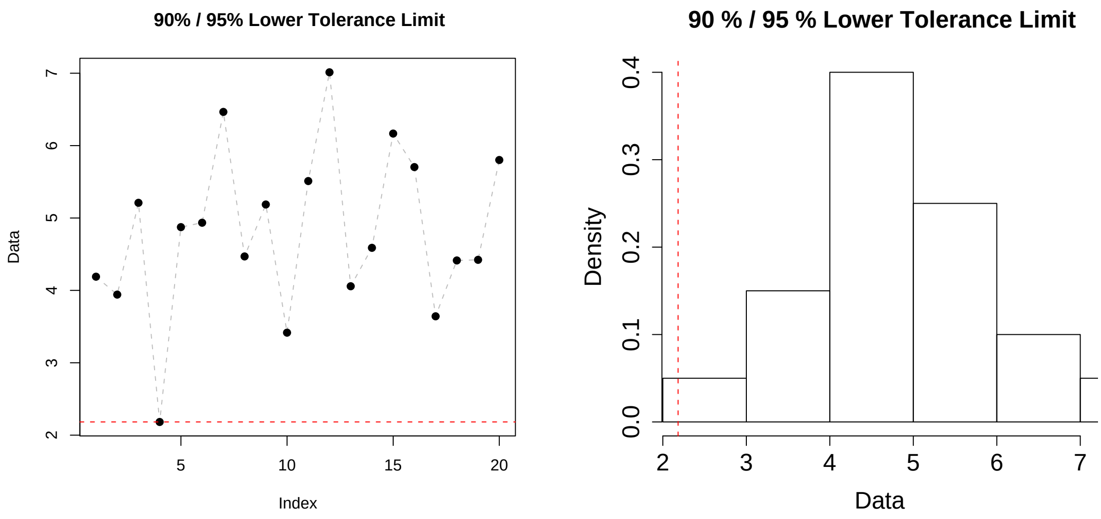
```
R> set.seed(100)
R> x <- rlogis(20, 5, 1)
R> out <- nptol.int(x = x, alpha = 0.10, P = 0.95, side = 1,
+    method = "WILKS", upper = NULL, lower = NULL)
R> out
```

```
  alpha    P 1-sided.lower 1-sided.upper
1   0.1 0.95       2.18245      7.013099
```

Notice that the `nptol.int` function also has the arguments `upper` and `lower`, which can (optionally) be used to specify known upper and lower limits for the data, respectively. Otherwise, these are taken to be the maximum and minimum of the data set.

The output for this example yields α, P, and both one-sided tolerance limits. So, the researcher can be 90% confident that at least 95% of the subjects will get at least 2.2 hours of sleep per night, which is a more conservative estimate than that obtained when assuming the data came from a logistic distribution. Figure 7 gives a control chart and histogram of this data with the lower tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "lower",
+    x.lab = "Data", lty = "dashed", col = "gray")
```



(a) (b)

Figure 7: (a) Control chart and (b) histogram of the $n = 20$ logistically distributed values. In each plot, the dashed red line gives the one-sided lower nonparametric tolerance limit.

Notice in the control chart how the red line for the tolerance limit passes through the actual data point.

4.8. (Univariate) normal tolerance intervals

Suppose a company manufactures metal rings which are to be 5 inches thick. A quality engineer wishes to assess how much the measurements of each ring differ from the nominal thickness. The difference in thickness values appear to be normally distributed. The company wishes to state with confidence level $(1-\alpha)$, that a proportion P of the metal rings are within a certain range of thicknesses. Calculation of a two-sided normal tolerance interval can provide such an answer for the company.

A random variable X follows a normal distribution if it has cumulative distribution function

$$F_X(x;\mu,\sigma) = \int_{t=-\infty}^{x} (2\pi\sigma^2)^{-1/2} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt,$$

where $-\infty < x < +\infty$, $-\infty < \mu < +\infty$ is the mean, and $\sigma > 0$ is the standard deviation. Normal tolerance intervals require estimates of both parameters ($\hat{\mu}$ and $\hat{\sigma}$) which can be obtained by using maximum likelihood estimation. There is a great deal of literature on normal tolerance intervals, but some of the more often cited references include Howe (1969) and Weissberg and Beatty (1969).

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper normal tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided normal tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\mu}, \hat{\sigma})^\top$. The formulas for estimating L and U for normally distributed data are:

$$L = \hat{\mu} - k\hat{\sigma}$$
$$U = \hat{\mu} + k\hat{\sigma}.$$

The estimate for k requires additional calculations regarding which type of normal tolerance interval will be calculated. In the one-sided setting, there is an exact solution to k. Unfortunately, this is not the case for the two-sided setting. However, there are two methods commonly employed for estimating k in the two-sided setting: the Weissberg-Beatty method (Weissberg and Beatty 1969) and the Howe method (Howe 1969). The Howe method is an improved approximation over the Weissberg-Beatty method, but both methods are available options in this function for comparative purposes.

For the one-sided tolerance intervals,

$$k = \frac{1}{\sqrt{n}} t^*_{n-1;1-\alpha}(\sqrt{n} z_P),$$

such that n is the sample size, $t^*_{d;1-\alpha}(\delta)$ is the $(1-\alpha)$-th quantile of a non-central t distribution with d degrees of freedom and non-centrality parameter δ, and z_P is the P-th percentile of the standard normal distribution.

For the two-sided tolerance interval, the Howe method finds

$$k = uvw,$$

where

$$u = z_{\frac{1+P}{2}} \sqrt{1 + n^{-1}}$$
$$v = \sqrt{\frac{n-1}{\chi^2_{n-1;\alpha}}}$$
$$w = \sqrt{1 + \frac{n - 3 - \chi^2_{n-1;\alpha}}{2(n+1)^2}},$$

such that n is the sample size and $\chi^2_{d;\alpha}$ is the α-th quantile of a χ^2 distribution with d degrees of freedom. The Weissberg-Beatty method finds

$$k = rv,$$

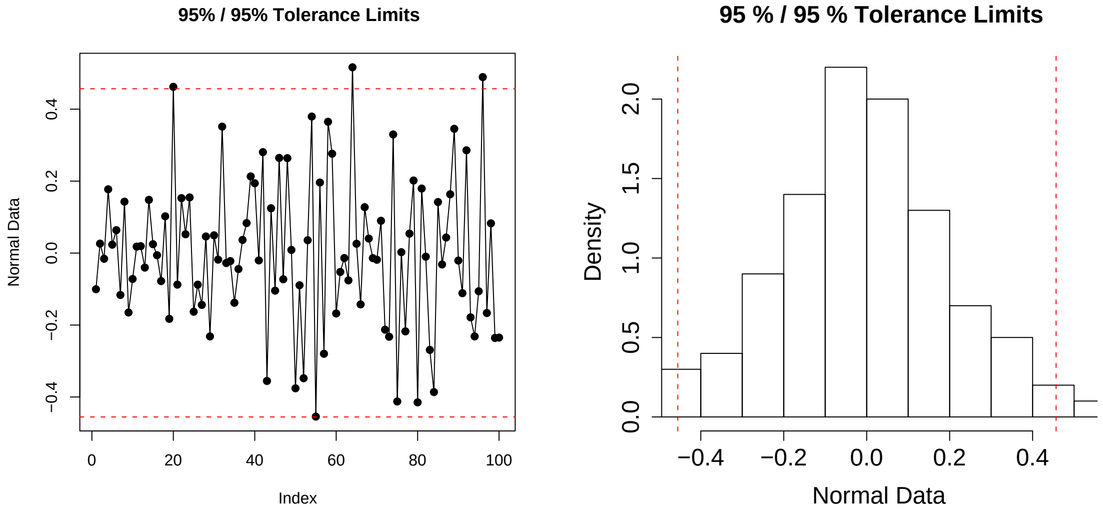
where

$$\Phi(r + \sqrt{n^{-1}}) - \Phi(r - \sqrt{n^{-1}}) = P$$

is solved for r using numerical methods (e.g., Newton's method). In the above, $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Tolerance intervals can also be found for log-normal data. If X^* is a log-normal random variable, then $\ln(X^*) = X$ is a normal random variable. The procedure outlined above can be performed, except at the end a transformation back to the log-normal scale is performed by taking e^L and e^U for the lower and upper limits, respectively.

Returning to the example, suppose the researcher has a batch of $n = 100$ rings and he wishes to state with 95% confidence that 95% of the ring measurements fall within a certain



(a) (b)

Figure 8: (a) Control chart and (b) histogram of the $n = 100$ normally distributed values. In each plot, the dashed red lines give the two-sided tolerance limits.

interval. Data were simulated using a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.2$. Evaluation of a 95%/95% two-sided normal tolerance interval is found by implementing the `normtol.int` function:

```
R> set.seed(100)
R> x <- rnorm(100, 0, 0.2)
R> out <- normtol.int(x = x, alpha = 0.05, P = 0.95, side = 2,
+    method = "HE", log.norm = FALSE)
R> out
```

```
  alpha    P        x.bar 2-sided.lower 2-sided.upper
1  0.05 0.95 0.0005825125    -0.4554493     0.4566144
```

Note that the `normtol.int` function contains the argument `method` for specifying whether to use the Howe method (`"HE"`) or the Weissberg-Beatty method (`"WBE"`).

The output for this example yields α, P, the sample mean (which is an estimate of μ) and the two-sided tolerance limits. So, the engineer can be 95% confident that at least 95% of the rings will have measurements between about 4.5446 inches and 5.4566 inches. Figure 8 gives a control chart and histogram of this data with the two-sided tolerance limits shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "two",
+    x.lab = "Normal Data")
```

4.9. (Multivariate) normal tolerance regions

Suppose the company which manufactures the metal rings from the univariate normal tolerance interval example also specifies that the diameter must be 15 inches. A quality engineer wishes to assess how much the thickness and diameter measurements of each ring differ from the specified values. These differences are assumed to follow a bivariate normal distribution. Calculation of a multivariate normal tolerance region can provide such an answer for the company.

A random vector $\underline{X} = (X_1, \ldots, X_p)^\top$ follows a multivariate normal distribution if it has the density function

$$f_{\underline{X}}(\underline{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(1/2)\|\Sigma^{-1/2}(\underline{x}-\boldsymbol{\mu})\|^2},$$

where $\underline{x} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$, Σ is a $p \times p$ symmetric, positive-definite matrix, and $\|\mathbf{A}\|^2 = \mathbf{A}^\top\mathbf{A}$ for any matrix $\mathbf{A}$. Unlike the univariate setting, less literature on multivariate normal tolerance regions is available. The method used here was developed in Krishnamoorthy and Mondal (2006).

Normal tolerance regions require estimates of both parameters $\boldsymbol{\mu}$ and Σ. The usual estimates are given by

$$\underline{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})^\top,$$

respectively. A $[100 \times (1-\alpha)\%]/[100 \times P\%]$ normal tolerance region gives, with $100 \times (1-\alpha)\%$ confidence, a region where at least $100 \times P\%$ of the population will be enclosed. The tolerance region is given by

$$\{\underline{x} : \|S^{-1/2}(\underline{x}_i - \underline{\bar{x}})\|^2 \leq c\},$$

where c is the tolerance factor to be determined such that

$$\Pr[\Pr_{\underline{X}}\{\|S^{-1/2}(\underline{x}_i - \underline{\bar{x}})\|^2 \leq c; \underline{x}, S\} \geq P] = 1 - \alpha.$$

The method for computing c requires a more detailed discussion which can be found in Krishnamoorthy and Mondal (2006). A basic sketch of the Monte Carlo algorithm is as follows:

1. Generate independent χ^2 random variables and Wishart random matrices.
2. Compute the eigenvalues of the randomly generated Wishart matrices.
3. Iterate the above B times to generate the sample values $T_1, \ldots, T_B$, where T_i is a function of the randomly generated quantities in steps 1 and 2. Then the $100 \times (1-\alpha)$-th percentile of the T_i's is an approximate tolerance factor c.

Returning to the example, suppose the researcher uses the same batch of $n = 100$ metal rings. He wishes to state with 95% confidence that 95% of the ring measurements fall within a certain region. Data were simulated using a bivariate normal distribution with

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.5 \end{pmatrix}.$$

Computing the tolerance factor for a 95%/95% multivariate normal tolerance region is found by implementing the `mvtol.region` function:

```
R> set.seed(100)
R> x1 <- rnorm(100, 0, 0.2)
R> x2 <- rnorm(100, 0, 0.5)
R> x <- cbind(x1, x2)
R> out <- mvtol.region(x = x, alpha = 0.05, P = 0.95, B = 10000)
R> out

         0.05
0.95 7.456825
```

Notice that the structure of the multivariate data assumes that each row of the matrix is an observation. The `mvtol.region` takes the argument `B`, which is used to specify the size of the Monte Carlo sample. Also, the arguments `alpha` and `P` can actually take a vector of values. If specifying a vector of values, then the output will be a matrix of tolerance factors with number of rows equal to the length of `P` and number of columns equal to the length of `alpha`. The output for this example yields the tolerance factor for the given combination of α and P. So, the engineer can be 95% confident that at least 95% of the rings will have measurements within the region shown in Figure 9(a). The scatterplot for this bivariate normal tolerance region is obtained by typing

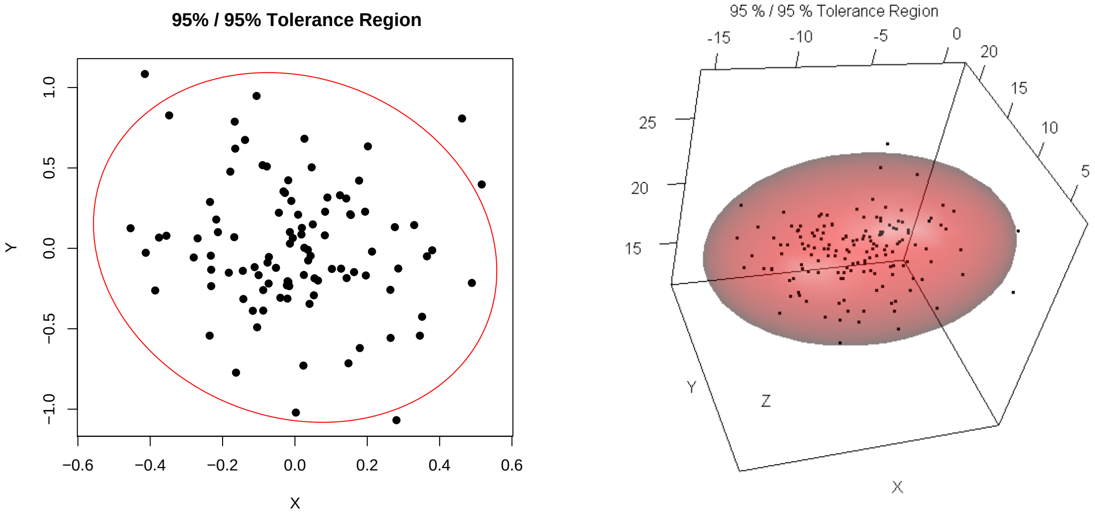
```
R> plottol(out, x)
```

For an example with trivariate data (i.e., $p = 3$), suppose a researcher wishes to obtain the 95%/95% tolerance factor for some population when a sample of size $n = 150$ is collected. Data were simulated using a trivariate normal distribution with

$$\boldsymbol{\mu} = \begin{pmatrix} -7 \\ 13 \\ 20 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 8.7 & -3.2 & 1.3 \\ -3.2 & 9.6 & -3.1 \\ 1.3 & -3.1 & 6.9 \end{pmatrix}.$$

Generation of this data and computation of the tolerance factor for a 95%/95% multivariate normal tolerance region is found by typing the following code:

```
R> set.seed(100)
R> x1 <- rnorm(150, 0, 1)
R> x2 <- rnorm(150, 0, 1)
R> x3 <- rnorm(150, 0, 1)
R> Sig <- matrix(c(8.7, -3.2, 1.3, -3.2, 9.6, -3.1, 1.3, -3.1, 6.9),
+    3, 3)
R> tmp <- eigen(Sig)
R> Sig.5 <- cbind(tmp$vectors) %*% diag(sqrt(tmp$values)) %*%
+    solve(cbind(tmp$vectors))
R> x <- t(Sig.5 %*% rbind(x1, x2, x3) + c(-7, 13, 20))
R> out1 <- mvtol.region(x = x, alpha = 0.05, P = 0.95, B = 10000)
R> out1
```



(a) (b)

Figure 9: (a) Scatterplot of the bivariate normal data with a 95%/95% tolerance region. (b) 3D scatterplot of the trivariate normal data with a 95%/95% tolerance region.

```
         0.05
0.95 9.130484
```

Figure 9(b) displays a 95%/95% trivariate normal tolerance region for this simulated data (which has an approximate tolerance factor $c = 9.1305$ from the above output). The 3D scatterplot for this trivariate normal tolerance region is obtained by typing

```
R> plottol(out1, x)
```

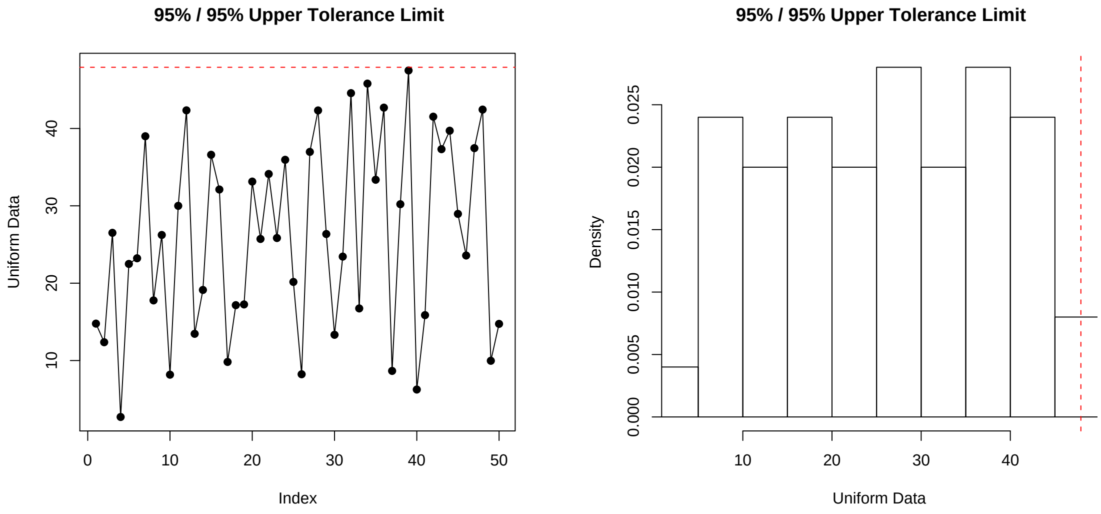
The scatterplot in Figure 9(b) is generated using the `plot3d` function in the **rgl** package (Adler and Murdoch 2010). The advantage of using this package is that it is a 3D real-time rendering program. So one can use their mouse cursor on scatterplots like Figure 9(b) and rotate it in real-time.

4.10. Uniform tolerance intervals

Suppose the time (in hours) between successive shutdowns of an unreliable machine is uniformly distributed. The company which owns this machine wants to state with confidence level $(1 - \alpha)$, an upper limit on the amount of time when a proportion P of all unscheduled shutdowns will occur. Calculation of a one-sided upper uniform tolerance interval can provide such an answer for the company.

A random variable X is uniformly distributed if it has cumulative distribution function

$$F_X(x; \theta_1, \theta_2) = \frac{x - \theta_1}{\theta_2 - \theta_1},$$



(a) (b)

Figure 10: (a) Control chart and (b) histogram of the $n = 50$ uniformly distributed values. In each plot, the dashed red line gives the one-sided upper tolerance limit.

where $\theta_1 < x < \theta_2$. The maximum likelihood estimates of θ_1 and θ_2 are $x_{(1)}$ and $x_{(n)}$ (i.e., the minimum and maximum values from a sample of size n), respectively.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ lower and upper uniform tolerance limits and a $[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided uniform tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^\top$. The formulas for estimating uniform tolerance intervals (see Faulkenberry and Weeks 1968) are:

$$L = \frac{(x_{(n)} - x_{(1)})(1-P)}{(1-\alpha)^{1/n}} + x_{(1)}$$

$$U = \frac{(x_{(n)} - x_{(1)})P}{\alpha^{1/n}} + x_{(1)}.$$

Approximate tolerance limits for the two-sided setting are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

Returning to the example, the company needs to be 95% confident of the upper time limit when 95% of all unscheduled shutdowns will occur. The data were generated using a uniform distribution with $\theta_1 = 0$ and $\theta_2 = 48$. Evaluation of a 95%/95% upper uniform tolerance interval is found by implementing the `uniftol.int` function:

```
R> set.seed(100)
R> x <- runif(50, 0, 48)
R> out <- uniftol.int(x = x, alpha = 0.05, P = 0.95, lower = 0, side = 1)
R> out

  alpha    P 1-sided.lower 1-sided.upper
1  0.05 0.95      2.377392      47.91036
```

The output for this example yields α, P, and both one-sided tolerance limits (even though in this example we are only interested in the upper limit). So the company can state with 95% confidence that 95% of the shutdowns of this machine will occur before 47.9 hours since the previous shutdown. Notice that `lower = 0` since we know that the distribution of these times is bounded below by 0.

Finally, Figure 10 gives a control chart and histogram of this data with the upper tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "upper",
+    x.lab = "Uniform Data")
```

4.11. Weibull and extreme-value tolerance intervals

Suppose an engineer is testing the endurance (in millions of revolutions) of ball bearings, which are adequately modeled using a Weibull distribution. The engineer wishes to state with confidence level $(1-\alpha)$, that a proportion P of all bearings will have at least a certain number of revolutions. Calculation of a one-sided lower Weibull tolerance interval can provide such an answer for the engineer.

A random variable X has a Weibull distribution if it has cumulative distribution function

$$F_X(x;\theta,\beta) = 1 - e^{-(x/\theta)^{\beta}},$$

where $x > 0$, with shape parameter $\beta > 0$ and scale parameter $\theta > 0$. Let a random variable Y be such that $Y = \ln(X)$. Then Y has an extreme-value distribution (also called the Gumbel distribution for the minimum) if it has cumulative distribution function

$$F_Y(y;\xi,\delta) = 1 - \exp\left\{-e^{\frac{y-\xi}{\delta}}\right\},$$

where $-\infty < y < +\infty$, $\xi = \ln(\theta)$ (so $-\infty < \xi < +\infty$), and $\delta = \beta^{-1}$ (so $\delta > 0$). The maximum likelihood estimates of the parameters ($\hat{\delta}$ and $\hat{\xi}$) can be found by using a Newton-Raphson algorithm initialized with the method of moments estimates. The maximum likelihood estimates of the parameters ($\hat{\beta}$ and $\hat{\theta}$) can be found by taking a log transformation on the data, finding the maximum likelihood estimates for δ and ξ, and then transforming those estimates back to the Weibull scale.

$[100\times(1-\alpha)\%]/[100\times P\%]$ lower and upper extreme-value tolerance limits and a $[100\times(1-\alpha)\%]/[100\times P\%]$ two-sided extreme-value tolerance interval are found according to equations (1), (2), and (3), respectively, such that $\boldsymbol{\theta} = (\hat{\xi},\hat{\delta})^{\top}$. Then, letting $\lambda_w = \ln(-\ln(w))$, n be the sample size, and $t_{d;1-\alpha^*}(\gamma)$ be the $(1-\alpha)$-th quantile of a non-central t distribution with d degrees of freedom and non-centrality parameter γ, the formulas for estimating the one-sided extreme-value tolerance limits (based on those in Bain and Engelhardt 1981) are:

$$\begin{aligned} L &= \hat{\xi} - \frac{\hat{\delta}t^*_{n-1;\alpha}(-\sqrt{n}\lambda_P)}{\sqrt{n-1}} \\ U &= \hat{\xi} - \frac{\hat{\delta}t^*_{n-1;1-\alpha}(-\sqrt{n}\lambda_{1-P})}{\sqrt{n-1}}. \end{aligned}$$

Furthermore, the formulas for estimating the one-sided Weibull tolerance limits are:

$$\begin{aligned} L_W &= e^L \\ U_W &= e^U. \end{aligned}$$

Finally, a random variable Z follows a Gumbel distribution for the maximum if it has cumulative distribution function

$$F_Z(z;\xi,\delta) = 1 - \exp\left\{-e^{-(\frac{z-\xi}{\delta})}\right\},$$

where $-\infty < z < +\infty$, $-\infty < \xi < +\infty$, and $\delta > 0$. The maximum likelihood estimates of the parameters ($\hat{\delta}$ and $\hat{\xi}$) can also be found by using a Newton-Raphson algorithm. Then, the formulas for estimating the one-sided tolerance limits for the Gumbel distribution for the maximum are:

$$\begin{aligned} L &= \hat{\xi} + \frac{\hat{\delta} t^*_{n-1;\alpha}(-\sqrt{n}\lambda_{1-P})}{\sqrt{n-1}} \\ U &= \hat{\xi} + \frac{\hat{\delta} t^*_{n-1;1-\alpha}(-\sqrt{n}\lambda_{P})}{\sqrt{n-1}}. \end{aligned}$$

Approximate tolerance intervals for the two-sided settings of all three distributions discussed above are found by replacing α by $\alpha/2$ and P by $(P+1)/2$ in the above formulas.

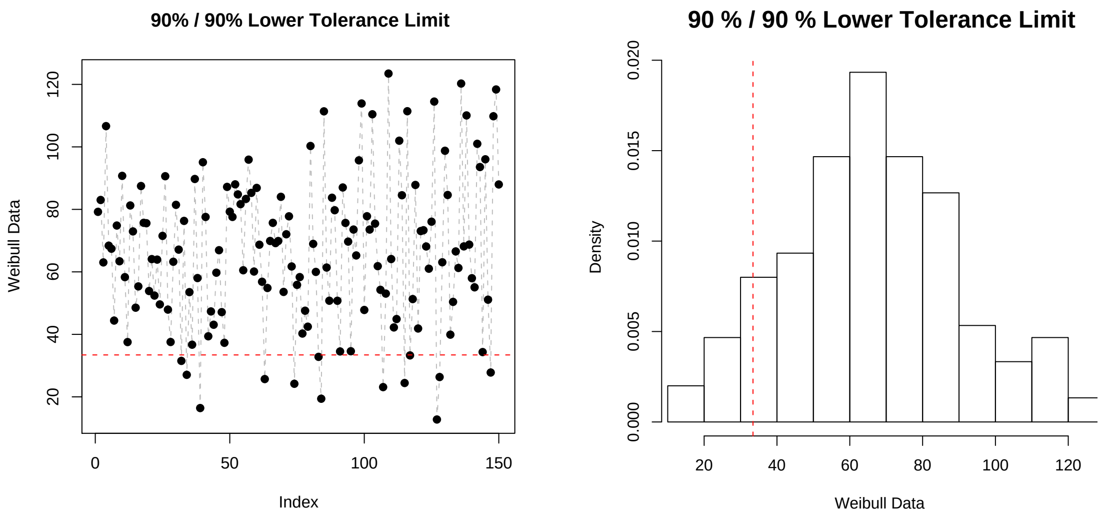
Returning to the example, the engineer wishes to be 90% confident that 90% of all bearings will have at least a certain number of revolutions. The data were generated using a Weibull distribution with $\beta = 3$ and $\theta = 75$. Evaluation of a 90%/90% lower Weibull tolerance interval is found by implementing the `exttol.int` function:

```
R> set.seed(100)
R> x <- rweibull(150, 3, 75)
R> out <- exttol.int(x = x, alpha = 0.10, P = 0.90,
+    dist = "Weibull", NR.delta = 1e-8)
R> out
```

```
  alpha   P  shape.1  shape.2 1-sided.lower 1-sided.upper
1   0.1 0.9 3.084836 74.48706      33.41646      101.8583
```

Notice that the `exttol.int` function has the argument `dist` for specifying whether the data is assumed to come from a Weibull distribution (i.e., `dist = "Weibull"`) or from an extreme-value distribution (i.e., `dist = "Gumbel"`). If the data follows one of the Gumbel distributions, then the argument `ext` is used for specifying whether the data is modeling the distribution of maximum values (i.e., `ext = "max"`) or minimum values (i.e., `ext = "min"`). Also, the option `NR.delta` specifies the stopping criterion used for the Newton-Raphson algorithm for the maximum likelihood estimates.

The output for this example yields α, P, estimates of the two distributional parameters using Newton-Raphson (i.e., shape and scale parameters for the Weibull distribution or location and scale parameters for the extreme-value distribution), and both one-sided tolerance limits



(a) (b)

Figure 11: (a) Control chart and (b) histogram of the $n = 150$ Weibull distributed values. In each plot, the dashed red line gives the one-sided lower tolerance limit.

(even though in this example we are only interested in the lower limit). So, the engineer can state with 90% confidence that 90% of the ball bearings should have at least 33.42 million revolutions. Figure 11 gives a control chart and histogram of this data with the lower tolerance limit shown. These plots are obtained by typing

```
R> plottol(out, x = x, plot.type = "both", side = "lower",
+    x.lab = "Weibull Data", lty = "dashed", col = "gray")
```

5. Regression tolerance intervals

As with tolerance regions for multivariate normal data, the calculated regression tolerance intervals can be overlayed on a scatterplot of the sample data using the `plottol` function. All three regression settings we discuss can be plotted in a similar manner, provided that the regression model is a function of only one predictor.

5.1. Linear regression tolerance intervals

Suppose a quality engineer wishes to model the quality scores of small businesses as a function of the amount spent on program funding (in thousands of dollars). The engineer wishes to claim with confidence level $(1 - \alpha)$, that a proportion P of all such businesses that spend a given amount are within certain limits on their quality scores. Calculation of two-sided linear regression tolerance intervals can provide such a quantification for the engineer.

A multiple linear regression model is used to model the linear relationship between a response variable Y with a given set of predictor variables $X_1, \ldots, X_{p-1}$. The multiple regression model is defined as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \epsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the p regression parameters and ϵ is a normally distributed error term with mean 0 and variance σ^2. For a data set of size n, the estimated regression model (estimated using ordinary least squares) is given by

$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_{p-1} x_{i,p-1} + e_i \\ &= \hat{y}_i + e_i, \end{aligned}$$

where the e_i's are the residuals and the $\hat{y}_i$'s are the fitted values for this regression equation. Tolerance limits can then be constructed using these fitted values.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ one-sided regression tolerance limits for each observation i are given by

$$\begin{aligned} L &= \hat{y}_i - \hat{\sigma} k_{1,i} \\ U &= \hat{y}_i + \hat{\sigma} k_{1,i}, \end{aligned}$$

respectively. $\hat{\sigma}$ is estimated by the root mean square error and

$$k_{1,i} = \frac{t^*_{n-p;1-\alpha}(\sqrt{n^*_i} z^*_P)}{\sqrt{n^*_i}},$$

where $t^*_{d;1-\alpha}(\gamma)$ is the $(1-\alpha)$-th quantile of a non-central t distribution with d degrees of freedom and non-centrality parameter γ, z^*_P is the P-th quantile of a standard normal distribution, and

$$n^*_i = \frac{\hat{\sigma}^2}{\text{s.e.}(\hat{y}_i)^2}$$

such that s.e.$(\hat{y}_i)$ is the standard error of $\hat{y}_i$. The value n^*_i is called the "effective number of observations" by Wallis (1946), which means when n^*_i is divided into the variance of an observation, then the result is the variance of the statistic.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ two-sided regression tolerance limits for each observation i are given by

$$\begin{aligned} L &= \hat{y}_i - \hat{\sigma} k_{2,i} \\ U &= \hat{y}_i + \hat{\sigma} k_{2,i}, \end{aligned}$$

where $k_{2,i}$ is estimated according to the formula in Krishnamoorthy and Mathew (2009). Let $f = n - p$. Then

$$k_{2,i} = \sqrt{\frac{f \chi^2_{1;P}(1/n^*_i)}{\chi^2_{f;\alpha}}}$$

where $\chi^2_{d;\alpha}(\delta)$ is the α-th quantile of a non-central χ^2 distribution with d degrees of freedom and non-centrality parameter δ.

Returning to the example, suppose the quality engineer has data from $n = 100$ small businesses. The engineer wishes to be 95% confident that 95% of all such businesses that spend a given amount are within certain limits on their quality scores. The data were generated assuming $(\beta_0, \beta_1) = (20, 5)$ and that the random error follows a normal distribution with mean 0 and standard deviation $\sigma = 3$. Evaluation of 95%/95% two-sided regression tolerance limits is found by implementing the `regtol.int` function:

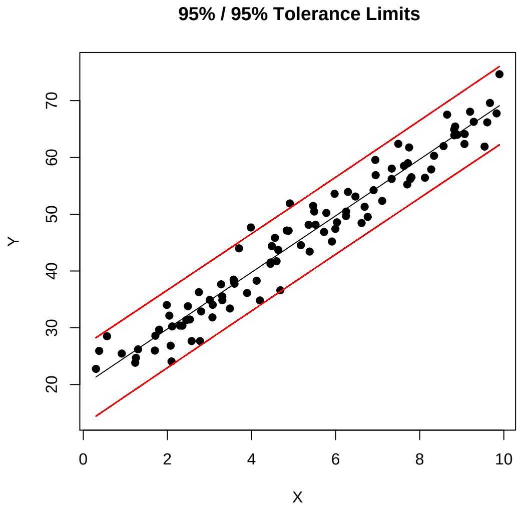


Figure 12: Scatterplot of the simulated linear regression data with the ordinary least squares line in black and 95%/95% linear regression tolerance limits in red.

```
R> set.seed(100)
R> x <- runif(100, 0, 10)
R> y <- 20 + 5 * x + rnorm(100, 0, 3)
R> out <- regtol.int(reg = lm(y ~ x), new.x = NULL, alpha = 0.05, P = 0.95,
+    side = 2)
R> out[1:5, ]
```

```
     alpha    P        y    y.hat 2-sided.lower 2-sided.upper
[1,]  0.05 0.95 22.76428 21.33912      14.43466      28.24357
[2,]  0.05 0.95 25.93468 21.72047      14.81959      28.62136
[3,]  0.05 0.95 28.51155 22.64589      15.75344      29.53834
[4,]  0.05 0.95 25.46921 24.39543      17.51807      31.27280
[5,]  0.05 0.95 23.84511 25.98808      19.12346      32.85270
```

The `regtol.int` function takes an object of class `lm` and uses the appropriate estimates from that fitted object to estimate the linear regression tolerance limits. This function also has the option `new.x` which can be used to specify new levels of the predictor for which the user wishes to construct tolerance limits.

The output for this example yields α, P, y_i, $\hat{y}_i$, and the two-sided tolerance limits. The output will have one row for each observation as well as any new levels of predictor(s) which the user has specified through the `new.x` argument. The output for this example has been truncated to show only the first five observations. Figure 12 gives a scatterplot of this data along with the least squares regression line and the tolerance limits. This plot is obtained by typing

```
R> plottol(out, x = cbind(1, x), y = y, side = "two", x.lab = "X",
+    y.lab = "Y")
```

5.2. Nonlinear regression tolerance intervals

Suppose an engineer is dealing with a physical process which is known to have a specified nonlinear relationship. The engineer wishes to claim with confidence level $(1-\alpha)$, that a proportion P of all responses for a given level of the predictor are within certain limits. Calculation of two-sided nonlinear regression tolerance intervals can provide such a quantification for the engineer.

A nonlinear regression model is used to model the nonlinear relationship between a response variable Y with a given set of predictor variables $X_1, \ldots, X_p$. The nonlinear regression model is defined as

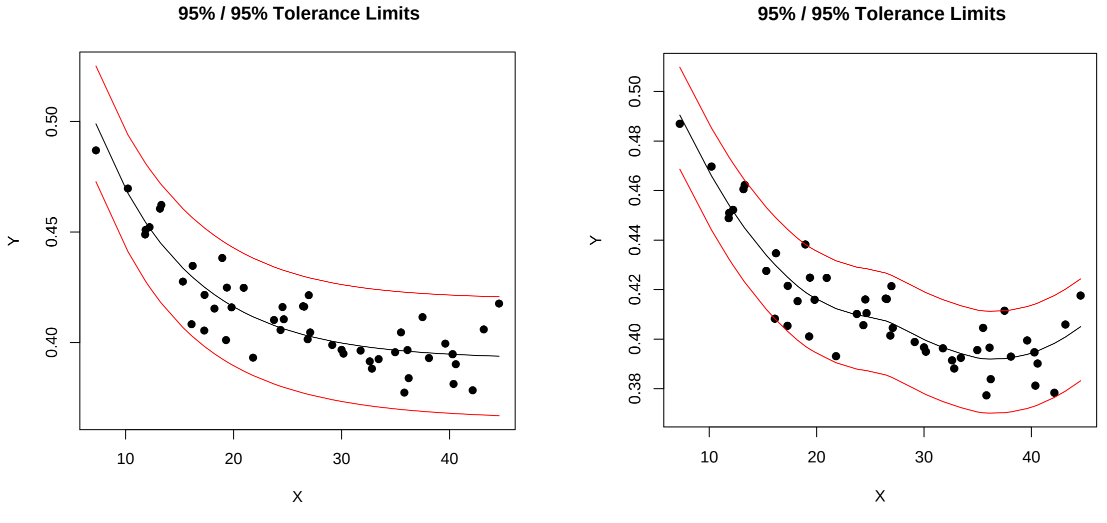
$$Y = f(\boldsymbol{\beta}, X_1, \ldots, X_p) + \epsilon,$$

where $\boldsymbol{\beta}$ is a vector of regression parameters and ϵ is an error term following a specified distribution which is not necessarily normal. For a data set of size n, the estimated nonlinear regression model is given by

$$\begin{aligned} y_i &= f(\hat{\boldsymbol{\beta}}, x_{i,1}, \ldots, x_{i,p}) + e_i \\ &= \hat{y}_i + e_i, \end{aligned}$$

where the e_i's are the residuals and the $\hat{y}_i$'s are the fitted values for this regression equation. Tolerance limits are again constructed based on these fitted values. The estimation done in the `nlregtol.int` function is through the nonlinear least squares routine `nls`.

$[100 \times (1-\alpha)\%]/[100 \times P\%]$ nonlinear regression tolerance limits are constructed in a similar manner as for the linear regression case. The only difference is how the effective sample



(a) (b)

Figure 13: Scatterplot of the simulated nonlinear regression data with (a) a nonlinear least squares curve in black and 95%/95% nonlinear regression tolerance limits in red and (b) LOESS curve in black and 95%/95% nonparametric regression tolerance limits in red.

size n_i^* is calculated. For the nonlinear setting, n_i^* is a function of the partial derivatives of $f(\boldsymbol{\beta}, x_{i,1}, \ldots, x_{i,p})$ with respect to each of the regression parameters (i.e., the gradient of $f(\cdot)$). The remaining formulas are the same. Further details can be found in Wallis (1946).

Returning to the example, suppose the physical process has the nonlinear relationship

$$Y = \beta_1 + (0.49 - \beta_2)e^{-\beta_2(X-8)},$$

where Y is the response and X is some predictor. The engineer wishes to be 95% confident that 95% of all observed responses are within certain limits for a given level of the predictor. The data were generated assuming $(\beta_1, \beta_2) = (0.39, 0.11)$ and that the random error follows a normal distribution with mean 0 and standard deviation 0.01. Evaluation of 95%/95% two-sided nonlinear regression tolerance limits is found by implementing the `nlregtol.int` function:

```
R> set.seed(100)
R> x <- runif(50, 5, 45)
R> f1 <- function(x, b1, b2) b1 + (0.49 - b1) * exp(-b2 * (x - 8)) +
+    rnorm(50, 0, 0.01)
R> y <- f1(x, 0.39, 0.11)
R> formula <- as.formula(y ~ b1 + (0.49 - b1) * exp(-b2 * (x - 8)))
R> out <- nlregtol.int(formula = formula, xy.data = data.frame(cbind(y, x)),
+    x.new = NULL, side = 2, alpha = 0.05, P = 0.95)
R> out[1:5, ]

     alpha    P     y.hat         y 2-sided.lower 2-sided.upper
[1,]  0.05 0.95 0.3961376 0.3773059     0.3695132     0.4227620
[2,]  0.05 0.95 0.3942112 0.3783508     0.3673801     0.4210423
[3,]  0.05 0.95 0.3946146 0.3812180     0.3678379     0.4213912
[4,]  0.05 0.95 0.3959679 0.3838453     0.3693298     0.4226059
[5,]  0.05 0.95 0.3976938 0.3881386     0.3711657     0.4242218
```

The `nlregtol.int` function requires the user to specify `formula`, which is a nonlinear model formula consistent with that used in `nls` (type `help("nls")` for more details). Also, the data must now be entered as a `data.frame` for the argument `xy.data`. Note that the response must appear in the first column of the specified `data.frame`.

The output for this example yields α, P, $\hat{y}_i$, y_i, and the two-sided tolerance limits. The output will have one row for each observation as well as any new levels of predictor(s) which the user has specified. The output for this example has again been truncated after the first five observations. Figure 13(a) gives a scatterplot of this data along with the nonlinear least squares regression fit and the tolerance limits. This plot is obtained by typing

```
R> plottol(out, x = x, y = y, side = "two", x.lab = "X", y.lab = "Y")
```

5.3. Nonparametric regression tolerance intervals

Consider the nonlinear regression tolerance bounds example from earlier. Suppose now that the engineer does not know the functional form of the relationship and decides to fit a non-parametric curve to the data. The engineer still wishes to claim with confidence level $(1 - \alpha)$, that a proportion P of all responses for a given level of the predictor are within certain limits. Calculation of two-sided nonparametric regression tolerance bounds can provide such a quantification for the engineer.

A nonparametric regression model is used to model the nonlinear relationship between a response variable Y with a given set of predictor variables $X_1, \ldots, X_p$, but with no parameters. The nonparametric regression model is defined as

$$Y = f(X_1, \ldots, X_p) + \epsilon,$$

which is free of any parameters and ϵ is a random error term which is only assumed to have mean 0. For a data set of size n, the estimated regression model (estimated using a nonparametric smoothing technique) is given by

$$\begin{aligned} y_i &= \hat{f}(x_{i,1}, \ldots, x_{i,p}) + e_i \\ &= \hat{y}_i + e_i, \end{aligned}$$

where the e_i's are the residuals and the $\hat{y}_i$'s are the fitted values for the chosen nonparametric regression routine. Tolerance limits for nonparametric regression can be constructed based on the residuals. First it is necessary to find $e_{(r)}$ and $e_{(s)}$ in the same manner as done for nonparametric tolerance intervals (where $e_{(j)}$ corresponds to the j-th value from the ordered sequence of the residuals). Then, $[100 \times (1-\alpha)\%]/[100 \times P\%]$ upper and lower nonparametric regression tolerance limits are given by

$$\begin{aligned} L &= \hat{y}_i + e_{(r)} \\ U &= \hat{y}_i + e_{(s)}. \end{aligned}$$

The methods for determining r and s are given in the section on nonparametric tolerance intervals.

Returning to the example and the data generated earlier, suppose the engineer fits a LOESS curve (Cleveland, Devlin, and Grosse 1988) to the data. The engineer wishes to be 95% confident that 95% of all observations are within certain limits. Evaluation of 95%/95% two-sided nonparametric regression tolerance limits is found by implementing the `npregtol.int` function:

```
R> set.seed(100)
R> x <- runif(50, 5, 45)
R> f1 <- function(x, b1, b2) b1 + (0.49 - b1)*exp(-b2*(x - 8)) +
+    rnorm(50, 0, 0.01)
R> y <- f1(x, 0.39, 0.11)
R> y.hat <- fitted(loess(y ~ x))
R> out <- npregtol.int(x = x, y = y, y.hat = y.hat, alpha = 0.05, P = 0.95,
+    side = 2, method = "WILKS", upper = NULL, lower = NULL)
R> out[1:5, ]
```

```
     alpha    P        x         y     y.hat 2-sided.lower 2-sided.upper
[1,]  0.05 0.95 36.21434 0.3838453 0.3919522     0.3701138     0.4112673
[2,]  0.05 0.95 36.10338 0.3965609 0.3919601     0.3701217     0.4112752
[3,]  0.05 0.95 35.81206 0.3773059 0.3919991     0.3701607     0.4113142
[4,]  0.05 0.95 35.50204 0.4045567 0.3920930     0.3702547     0.4114081
[5,]  0.05 0.95 37.49610 0.4114723 0.3921572     0.3703188     0.4114723
```

Notice that the `npregtol.int` function also has a `method` argument which specifies the estimation method to use for determining the indices of the ordered residuals to use for the tolerance limits. As with the `nptol.int` function, the possible methods include the Wilks method (`"WILKS"`), the Wald method (`"WALD"`), and the Hahn-Meeker method (`"HM"`).

The output for this example yields α, P, x_i, y_i, $\hat{y}_i$, and the two-sided tolerance limits. The output will have one row for each observation. The output for this example has also been truncated to show only the first five observations. Figure 13(b) gives a scatterplot of this data along with the fitted regression line obtained using the LOESS procedure and the two-sided tolerance limits. This plot is obtained by typing

```
R> plottol(out, x = x, y = y, y.hat = y.hat, side = "two", x.lab = "X",
+    y.lab = "Y")
```

Details on the LOESS procedure in R can be found by typing `help("loess")`.

6. Discussion

Tolerance limits enjoy a fairly rich history in the literature and have a very important role in engineering and manufacturing applications. The **tolerance** package was developed to provide a central collection of functions to estimate tolerance limits for some of the more frequent settings found in practice, including certain discrete and continuous univariate distributions, the multivariate normal distribution, and common regression settings. This package by no means provides an exhaustive collection of functions for the many tolerance intervals available; however, it will be maintained to supplement its current capabilities. Functions under consideration include tolerance limits for survival regressions, semi-parametric regression models, and non-normal multivariate distributions. Additional plotting capabilities for the multiple regression setting will also be explored.

Acknowledgments

The author is grateful to two anonymous referees for numerous helpful comments regarding both the **tolerance** package and the preparation of this article.

References

Adler D, Murdoch D (2010). ***rgl**: 3D Visualization Device System (OpenGL).* R package version 0.91, URL http://CRAN.R-project.org/package=rgl.

Bain LJ, Engelhardt M (1973). “Interval Estimation for the Two-Parameter Double Exponential Distribution.” *Technometrics*, **15**, 875–887.

Bain LJ, Engelhardt M (1981). “Simple Approximate Distributional Results for Confidence and Tolerance Limits for the Weibull Distribution Based on Maximum Likelihood.” *Technometrics*, **23**, 15–20.

Bain LJ, Engelhardt M (1991). *Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods.* 2nd edition. Marcel Dekker, Inc., New York.

Balakrishnan N (1991). *Handbook of the Logistic Distribution.* CRC Press, New York.

Blischke WR, Murthy DNP (2000). *Reliability: Modeling, Prediction, and Optimization.* John Wiley & Sons, New York.

Brown LD, Cai TT, DasGupta A (2001). "Interval Estimation for a Binomial Proportion." *Statistical Science*, **16**, 101–133.

Burrows GL (1963). "Statistical Tolerance Limits – What Are They?" *Applied Statistics*, **12**, 133–144.

Cleveland WS, Devlin SJ, Grosse E (1988). "Regression by Local Fitting." *Journal of Econometrics*, **37**, 87–114.

Dunsmore IR (1978). "Some Approximations for Tolerance Factors for the Two Parameter Exponential Distribution." *Technometrics*, **20**, 317–318.

Engelhardt M, Bain LJ (1978). "Tolerance Limits and Confidence Limits on Reliability for the Two-Parameter Exponential Distribution." *Technometrics*, **20**, 37–39.

Faulkenberry GD, Weeks DL (1968). "Sample Size Determination for Tolerance Limits." *Technometrics*, **10**, 343–348.

Guenther WC (1972). "Tolerance Intervals for Univariate Distributions." *Naval Research Logistics Quarterly*, **19**, 309–333.

Guenther WC, Patil SA, Uppuluri VRR (1976). "One-Sided β-Content Tolerance Factors for the Two Parameter Exponential Distribution." *Technometrics*, **18**, 333–340.

Hahn GJ, Chandra R (1981). "Tolerance Intervals for Poisson and Binomial Random Variables." *Journal of Quality Technology*, **13**, 100–110.

Hahn GJ, Meeker WQ (1991). *Statistical Intervals: A Guide for Practitioners.* John Wiley & Sons, New York.

Hall IJ (1975). "One-Sided Tolerance Limits for a Logistic Distribution Based on Censored Samples." *Biometrics*, **31**, 873–880.

Howe WG (1969). "Two-Sided Tolerance Limits for Normal Populations – Some Improvements." *Journal of the American Statistical Association*, **64**, 610–620.

Krishnamoorthy K, Mathew T (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation.* Wiley, Hoboken.

Krishnamoorthy K, Mathew T, Mukherjee S (2008). "Normal-Based Methods for a Gamma Distribution: Prediction and Tolerance Intervals and Stress-Strength Reliability." *Technometrics*, **50**, 69–78.

Krishnamoorthy K, Mondal S (2006). "Improved Tolerance Factors for Multivariate Normal Distributions." *Communications in Statistics: Simulation and Computation*, **35**, 461–478.

Montgomery DC (2005). *Introduction to Statistical Quality Control*. 5th edition. John Wiley & Sons, Hoboken, NJ.

Patel JK (1986). "Tolerance Limits – A Review." *Communications in Statistics: Theory and Methodology*, **15**, 2719–2762.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Wald A (1943). "An Extension of Wilks' Method for Setting Tolerance Limits." *Annals of Mathematical Statistics*, **14**, 44–55.

Wallis WA (1946). "Tolerance Intervals for Linear Regression." In J Neyman (ed.), *Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 43–51. University of California Press, Berkely, CA.

Weissberg A, Beatty G (1969). "Tables of Tolerance Limits Associated with Engineering Models." *Technometrics*, **2**, 483–500.

Wilks SS (1941). "Determination of Sample Sizes for Setting Tolerance Limits." *Annals of Mathematical Statistics*, **12**, 91–96.

Young DS (2010). ***tolerance**: Functions for Calculating Tolerance Intervals*. R package version 0.2.2, URL `http://CRAN.R-project.org/package=tolerance`.

Affiliation:

Derek S. Young
Department of Statistics
326 Thomas Building
Pennsylvania State University
University Park, PA 16802, United States of America
Telephone: +1/814-865-1348
Fax: +1/814-863-7114
E-mail: `dsy109@stat.psu.edu`
URL: `http://www.stat.psu.edu/~dsy109/`

Journal of Statistical Software `http://www.jstatsoft.org/`
published by the American Statistical Association `http://www.amstat.org/`
Volume 36, Issue 5 *Submitted:* 2009-07-01
August 2010 *Accepted:* 2010-06-24