Computing Statistical Tolerance Regions Using the **R** Package `tolerance`

Derek S. Young
Dr. Bing Zhang Department of Statistics
University of Kentucky

DATAWorks 2022
April 27th, 2022

[*] *Joint work with Kedai Cheng, University of North Carolina - Asheville*

College of Arts
and Sciences
Dr. Bing Zhang Department of Statistics

Funding Acknowledgment

This work is supported by the Chan Zuckerberg Initiative: Essential Open Source Software for Science (Grant Number 2020-225193).

Outline of Topics

Preliminaries

`tolerance`

- Normal Tolerance Intervals
- Other Univariate Tolerance Intervals
- Regression Tolerance Intervals
- Multivariate Tolerance Regions
- `Shiny` App

Resources

Preliminaries

Materials: https://github.com/dsy109/Supplemental/tree/main/DATAWorks 202022

Why Tolerance Regions?

- Based on a random sample, three primary statistical regions can be calculated:
  - **Confidence regions** → provide regions for an unknown population parameter (e.g., mean vector, variance-covariance matrix)
  - **Prediction regions** → provide regions for one or more future observations from the sampled population
  - **Tolerance regions** → provide regions that are expected to contain at least a specified proportion of the sampled population
- Typical applications of tolerance regions (or **tolerance intervals** for the univariate setting) include clinical and industrial studies, statistical quality control, environmental monitoring, and setting statistically-based engineering design limits
- Tolerance intervals are used in regulations published by the EPA (Environmental Protection Agency, 2006), the IAEA (International Atomic Energy Agency, 2008), and standard 16269-6 of the ISO (International Organization for Standardization, 2014)
- Krishnamoorthy and Mathew (2009) is a good reference on tolerance regions

Statistical Tolerance Intervals

Definition

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample of continuous random variables that have cumulative distribution function F_X, which is parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Let $X \sim F_X$, independently of $\mathbf{X}$. A (P, γ) **one-sided upper tolerance limit** $(U_1(\mathbf{X}))$ and (P, γ) **one-sided lower tolerance limit** $(L_1(\mathbf{X}))$ satisfy the expressions

$$\Pr_{\mathbf{X}} \left(\Pr_X \left[X \leq U_1(\mathbf{X}) | \mathbf{X} \right] \geq P \right) = \gamma \tag{1}$$

and

$$\Pr_{\mathbf{X}} \left(\Pr_X \left[L_1(\mathbf{X}) \leq X | \mathbf{X} \right] \geq P \right) = \gamma, \tag{2}$$

respectively. Similarly, a (P, γ) **two-sided tolerance interval**, $(L_2(\mathbf{X}),\ U_2(\mathbf{X}))$, satisfies

$$\Pr_{\mathbf{X}} \left(\Pr_X \left[L_2(\mathbf{X}) \leq X \leq U_2(\mathbf{X}) | \mathbf{X} \right] \geq P \right) = \gamma. \tag{3}$$

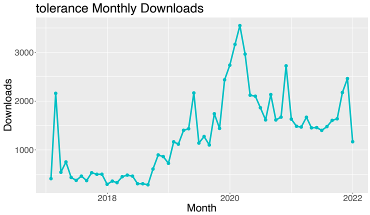
Sometimes, controlling the proportion in the tails is required, in which case we have a (P, γ) **equal-tailed tolerance interval**, $(L_e(\mathbf{X}),\ U_e(\mathbf{X}))$, which satisfies

$$\Pr_{\mathbf{X}} \left(\left\{ \Pr_X \left[L_e(\mathbf{X}) \leq X | \mathbf{X} \right] \leq (1-P)/2 \right\} \cap \left\{ \Pr_X \left[U_e(\mathbf{X}) \geq X | \mathbf{X} \right] \leq (1-P)/2 \right\} \right) = \gamma. \tag{4}$$

tolerance

tolerance (Young, 2010)

- First release: 2009
- The package includes: tolerance interval procedures for numerous parametric distributions, nonparametric settings, regression models, and some multivariate settings, as well as visualizations
- Used by NASA, 3M, EcoLab, PepsiCo, and NIST, among others
- The figure below also shows a mostly increasing trend over the years as the average number of monthly downloads per year is approximately 612, 506, 1597, 2263, and 1622, in 2017, 2018, 2019, 2020, and 2021, respectively



General Capabilities of `tolerance`

Parametric Procedures for Univariate Data

- Continuous distributions
  - Normal (log-normal) [`normtol.int`, `bayesnormtol.int`, `simnormtol.int`]
  - Gamma (log-gamma) [`gamtol.int`]
  - Laplace [`laptol.int`]
- Discrete distributions
  - Binomial [`bintol.int`]
  - (Negative) hypergeometric [`hypertol.int`, `neghypertol.int`]
  - Poisson [`poistol.int`]

Nonparametric Procedures for Univariate Data

- (P, γ) nonparametric tolerance intervals [`nptol.int`]
- β-expectation nonparametric tolerance intervals [`npbetol.int`]
- Sample size determination for nonparametric tolerance intervals [`np.order`]

General Capabilities of `tolerance` (ctd.)

Regression Tolerance Intervals

- Linear regression tolerance intervals [`regtol.int`]
- Nonlinear regression tolerance intervals [`nlregtol.int`]
- Nonparametric regression tolerance intervals [`npregtol.int`]

Multivariate Tolerance Regions

- Multivariate normal tolerance regions [`mvtol.region`]
- Multivariate regression tolerance regions [`mvregtol.region`]
- Nonparametric multivariate tolerance regions [`npmvtol.region`]

Visualizations

- Plotting control charts, histograms, and scatterplots with tolerance limits/regions [`plottol`]
- Operating characteristic curves for k-factors [`norm.oc`]

Format of Typical `tolerance` Function

```
disttol.int(data, alpha, P, side, method)
```

Function Inputs

- `data`: Often a vector, matrix, or data frame
- `alpha`: The significance level, such that 1-`alpha` equals γ
- `P`: The content level
- `side`: A way to specify either one-sided or two-sided tolerance intervals
- `method`: Different approximations for calculating the corresponding tolerance intervals/regions

Function Outputs

- Parameter estimates (for parametric tolerance intervals)
- The requested tolerance intervals/regions

Normal Tolerance Intervals

Normal Tolerance Intervals

- Let $X_1, \ldots, X_n$ be *iid* $\mathcal{N}(\mu, \sigma^2)$; i.e. a normal distribution with unknown mean μ and unknown variance σ^2
- Let $\bar{X}$ and S^2 denote the sample mean and sample variance, respectively
- The formulas for (P, γ) lower and upper normal tolerance limits are

$$L_h(\mathbf{X}) = \bar{X} - k_h(n, \gamma, P)S \quad \text{and} \quad U_h(\mathbf{X}) = \bar{X} + k_h(n, \gamma, P)S, \tag{5}$$

  respectively, where $h \in \{1, 2, e\}$

  - h is an index specifying whether we want one-sided tolerance limits, two-sided tolerance intervals, or equal-tailed tolerance intervals
- $k_1(n, \gamma, P)$, $k_2(n, \gamma, P)$, and $k_e(n, \gamma, P)$ are the k-**factors** for these settings
  - The k-factor ensures that we capture at least a proportion P of the sampled population with confidence level γ

k-Factors in `tolerance`

We have four functions regarding k-factors in `tolerance`

1. `K.factor`: Computes any of the three k-factors we presented
   - $k_1(n, \gamma, P)$ is computed by setting `side = 1`
   - $k_2(n, \gamma, P)$ is computed by setting `side = 2`; various (approximate) methods are available, however, `method = "EXACT"` is the most accurate as it solves an integral equation to find the exact k-factor
   - $k_e(n, \gamma, P)$ is computed by setting `side = 2` and `method = "OCT"`
2. `K.table`: Tabulates k-factors under numerous user-provided conditions
3. `K.factor.sim`: Estimates k-factors for simultaneous tolerance intervals based on normality
4. `norm.OC`: Provides operating characteristic curves for either the k-factor, n, γ, or P when given levels of the other three quantities (Young, 2016)

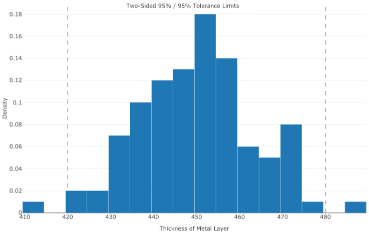
Example: Thickness of Silicon Wafers

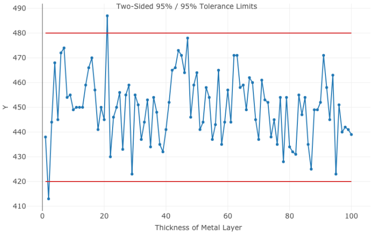
- Data were collected on the thickness of a metal layer on $n = 100$ silicon wafers resulting from a chemical vapor deposition (CVD) process in a semiconductor plant
- The thickness is measured in angstroms (Å)
- The data appear (approximately) normally distributed
- We we will construct a $(0.95, 0.95)$ two-sided normal tolerance interval to quantify where we expect 95% of such measurements from this process should fall with 95% confidence
- Source: D. C. Montgomery (2009), *Introduction to Statistical Quality Control, 6^{th} ed.* Wiley: Hoboken, NJ.

Example: Thickness of Silicon Wafers (ctd.)

```
normtol.int(x = wafer, alpha = 0.05, P = 0.95, side = 2, method = "EXACT")

  alpha    P  x.bar 2-sided.lower 2-sided.upper
1  0.05 0.95 450.01       420.015       480.005
```





Other Univariate Tolerance Intervals

Non-Normal and Nonparametric Tolerance Intervals

- Although normal tolerance intervals are, perhaps, the most frequently calculated in practice, tolerance intervals for non-normal distributions are also quite common
- The `tolerance` package handles numerous discrete and continuous non-normal distributions
- Tolerance intervals for discrete distributions mostly use the method of Hahn and Chandra (1981), which requires the discrete distribution to be stochastically increasing in the main parameter (e.g., the rate parameter or location parameter), and then the confidence interval on that parameter is used to construct the tolerance interval
- For continuous univariate populations, but where a distribution assumption is not made, we can construct nonparametric (P, γ) tolerance limits based on order statistics
- We will briefly illustrate tolerance intervals for the binomial distribution and the Weibull distribution as well as nonparametric tolerance intervals

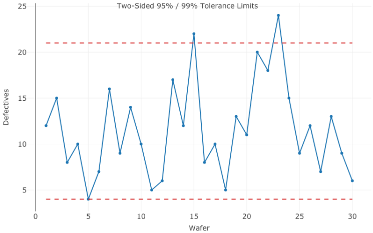
Example: Defective Chips

- For quality control purposes, the location of chips on a wafer is measured on 30 wafers
- On each wafer, 50 chips are measured
- A chip is recorded as defective whenever a misregistration, in terms of horizontal and/or vertical distances from the center, occurs
- Thus, we have $n = 50 * 30 = 1500$ samples of chips for which we can calculate the proportion of defects
- We we will construct a $(0.99, 0.95)$ two-sided binomial tolerance interval for a future sample of size $m = 50$ chips
- Source: NIST (2013). *NIST/SEMATECH e-Handbook of Statistical Methods*, https://www.itl.nist.gov/div898/handbook/pmc/section3/pmc332.htm, accessed: April 7th, 2022.

Example: Defective Chips (ctd.)

```
bintol.int(x = defects, n = 50*30, m = 50, alpha = 0.05, P = 0.99, side = 2,
           method = "CP")
```

```
  alpha    P  p.hat 2-sided.lower 2-sided.upper
1  0.05 0.99 0.2313             4            21
```



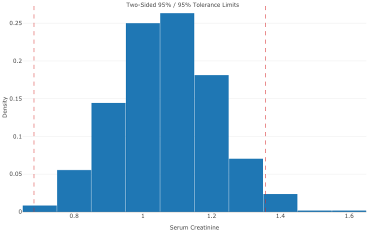
Example: Serum Creatinine Levels

- Serum creatinine is a blood test to measure the creatinine levels (mg/dL) in the blood
- Reference ranges for creatinine levels are typically $0.7 - 1.3$ mg/dL for males and $0.6 - 1.1$ mg/dL for females
- We will construct a $(0.95, 0.95)$ two-sided Weibull tolerance interval for serum creatinine from a sample of $n = 596$ healthy individuals (sex is not distinguished)
- Source: E. Harris and J. C. Boyd (1995), *Statistical Bases of Reference Values in Laboratory Medicine.* Marcel-Dekker, NY.

Example: Serum Creatinine Levels (ctd.)

```
exttol.int(x = kidney$SCR, alpha = 0.05, P = 0.95, side = 2, dist = "Weibull")

  alpha    P shape.1 shape.2 2-sided.lower 2-sided.upper
1  0.05 0.95    7.78    1.13        0.6832         1.357
```



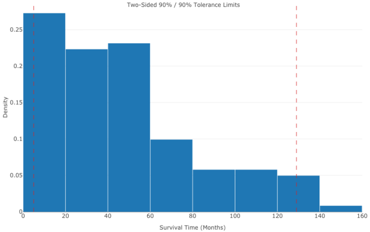
Example: Breast Cancer Survival Times

- We analyze a subset of data from a larger study on the survival times in months for $n = 121$ breast cancer patients
- We will construct a nonparametric $(0.90, 0.90)$ two-sided tolerance interval for the remission times
- Such a tolerance interval could provide reasonable expectations for the duration of long-term care of individuals
- Source: M. W. A. Ramos et al. (2013), "The Zografos-Balakrishnan Log-Logistic Distribution: Properties and Applications." *Journal of Statistical Theory and Applications*; 13(1):65–82.

Example: Breast Cancer Survival Times (ctd.)

```
nptol.int(x = bcancer, alpha = 0.10, P = 0.90, side = 2, method="WILKS")

    alpha   P 2-sided.lower 2-sided.upper
V14   0.1 0.9             5           129
```



Regression Tolerance Intervals

Including Covariates

- All procedures discussed thus far are for univariate settings
- There are many methods for constructing tolerance regions in regression settings, and the `tolerance` package has procedures for linear regression, nonlinear regression, nonparametric regression, and multivariate linear regression
- Wallis (1951) first developed pointwise tolerance intervals for linear regression models
- Young (2013) also discussed pointwise tolerance intervals for linear regression models, but proposed procedures for the nonlinear regression and nonparametric regression settings
- Various approaches have been proposed for simultaneous tolerance intervals in regression settings, however, those are not currently implemented in `tolerance`

Example: Hospital Infections

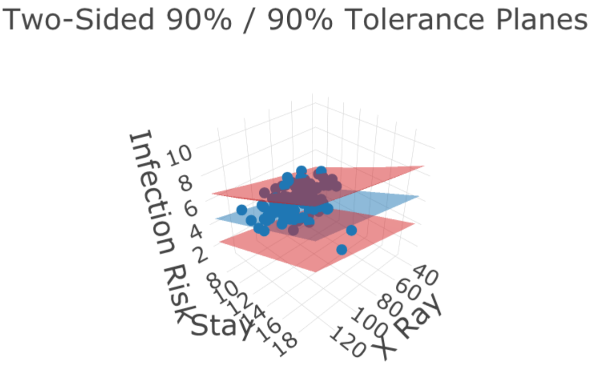
- We analyze data from the Study on the Efficacy of Nosocomial Infection Control (SENIC Project)
- This study's primary aim was to determine whether infection surveillance and control programs helped to reduce the rates of nosocomial (hospital-acquired) infection in US hospitals
- This dataset is a random sample of $n = 113$ hospitals from the original 338 hospitals surveyed
- The response variable (Y) of infection risk will be regressed on the average length of stay of all hospital patients (X_1) and the ratio of number of chest X-rays performed to number of patients without signs or symptoms of pneumonia, times 100 (X_2)
- We will construct pointwise $(0.90, 0.90)$ two-sided linear regression tolerance intervals, including for a new observation that has an average length of stay of 12 days and a chest X-ray ratio (times 100) of 50
- Source: J. Neter et al. (1996), *Applied Linear Statistical Models*, 4^{th} *ed.* McGraw-Hill, Boston.

Example: Hospital Infections (ctd.)

```
out <- lm(InfctRsk ~ Xray + Stay, data = hospitals)
regtol.int2(out, new.x = data.frame(Xray = 50, Stay = 12), alpha = 0.10,
            P = 0.90, side = 2, new = TRUE)

      y   y.hat 2-sided.lower 2-sided.upper
110 5.8 4.09399       2.12606       6.06192
111 4.4 3.50079       1.52566       5.47592
112 5.9 7.01366       4.87898       9.14833
113 3.1 4.48808       2.51959       6.45657
114  NA 4.41087       2.38824       6.43349
```

Example: Hospital Infections (ctd.)



Two-Sided 90% / 90% Tolerance Planes

Multivariate Tolerance Regions

Multivariate Data

- When an observation has many variables measured, it might not be tenable to build a regression relationship to characterize dependencies between the variables
- In the absence of assuming a dependent variable – either through a designed experiment or observational study – we can still look at the relationship between all of the measured variables through multivariate analysis
- Just as in the previous settings, multivariate tolerance regions serve the same purpose and have the same interpretation, granted in a multivariate context
- There are far fewer procedures for multivariate tolerance regions, and the `tolerance` package has procedures for multivariate normal tolerance regions, multivariate linear regression tolerance regions, and nonparametric hyperrectangular tolerance regions
  - Young and Mathew (2020) developed nonparametric (hyper)rectangular tolerance regions for constructing (hyper)rectangular reference regions by using data depth to provide a center-outward ranking of the data, followed by bounding the region of a depth-trimmed version of the data

Example: Adolescent Kidney Function Reference Regions

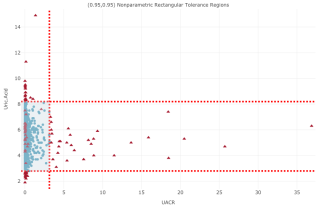
- Kidney function laboratory tests include a urinalysis to screen for the presence of protein and blood in the urine, a blood urea nitrogen (BUN) test to check for waste product in the urine, and a test to obtain the estimated glomerular filtration rate (eGFR), which is used to detect the presence and cause of kidney disease
- Little information is available regarding normal reference values for kidney function in adolescents, which impacts how physicians diagnose and manage diabetes in this population
- The reference population studied is healthy US adolescents between 12 and 17 years of age, with a number of criteria used to determine "healthy"
- We will construct nonparametric $(0.95, 0.95)$ semi-space rectangular tolerance regions to represent the reference regions of normal adolescent kidney function; we will look at the males in this sample, yielding $n = 2529$ subjects in the reference sample
- Source: D. S. Young and T. Mathew (2020), "Nonparametric Hyperrectangular Tolerance and Prediction Regions for Setting Multivariate Reference Regions in Laboratory Medicine." *Statistical Methods in Medical Research*; 29(12):3569–3585.

Example: Adolescent Kidney Function Reference Regions (ctd.)

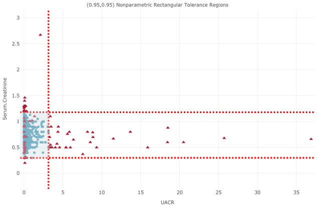
```
npmvtol.region(x = as.matrix(ref.males), alpha = 0.05, P = 0.95,
               depth.fn = Elliptical, type = "semispace", adjust = "ceiling",
               semi.order = list(lower = NULL, center = 2:3, upper = 1))
```

```
                 Lower  Upper
UACR              -Inf 3.1486
Uric.Acid          2.8 8.2000
Serum.Creatinine   0.3 1.1800
```

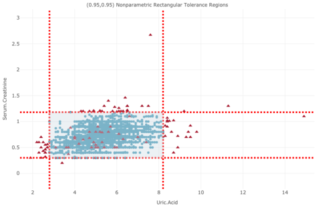
Example: Adolescent Kidney Function Reference Regions (ctd.)



(0.95,0.95) Nonparametric Rectangular Tolerance Regions



(0.95,0.95) Nonparametric Rectangular Tolerance Regions

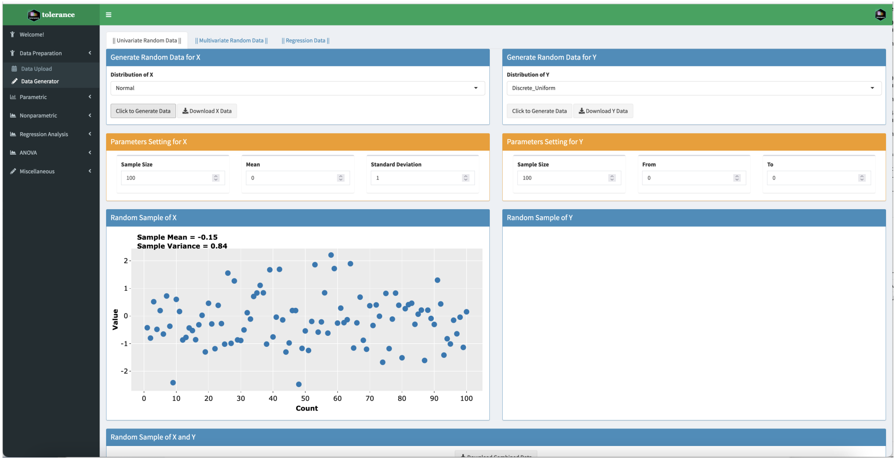


(0.95,0.95) Nonparametric Rectangular Tolerance Regions

Shiny App

Link: https://tolerance.as.uky.edu/

Shiny App Demo



Distribution of X: Normal

Sample Size: 100

Mean: 0

Standard Deviation: 1

Distribution of Y: Discrete_Uniform

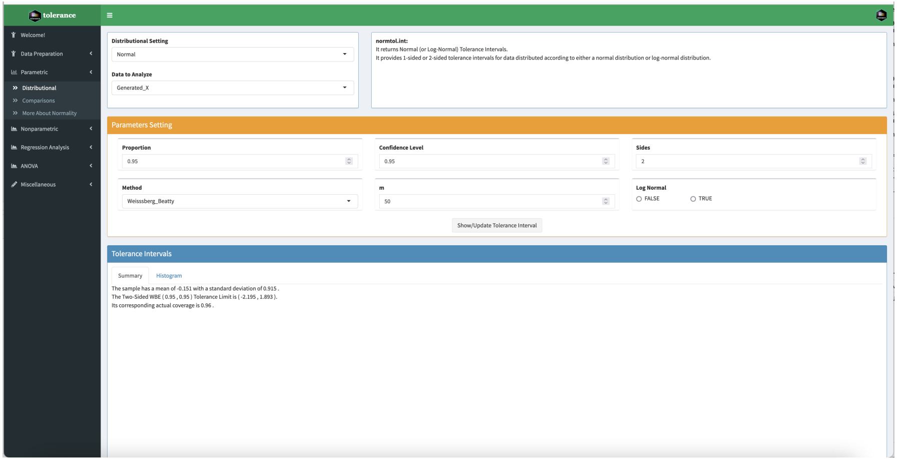
Sample Size: 100

From: 0

To: 0

Sample Mean = -0.15
Sample Variance = 0.84

Shiny App Demo (ctd.)



Distributional Setting
Normal

Data to Analyze
Generated_X

normtol.int:
It returns Normal (or Log-Normal) Tolerance Intervals.
It provides 1-sided or 2-sided tolerance intervals for data distributed according to either a normal distribution or log-normal distribution.

Proportion
0.95

Confidence Level
0.95

Sides
2

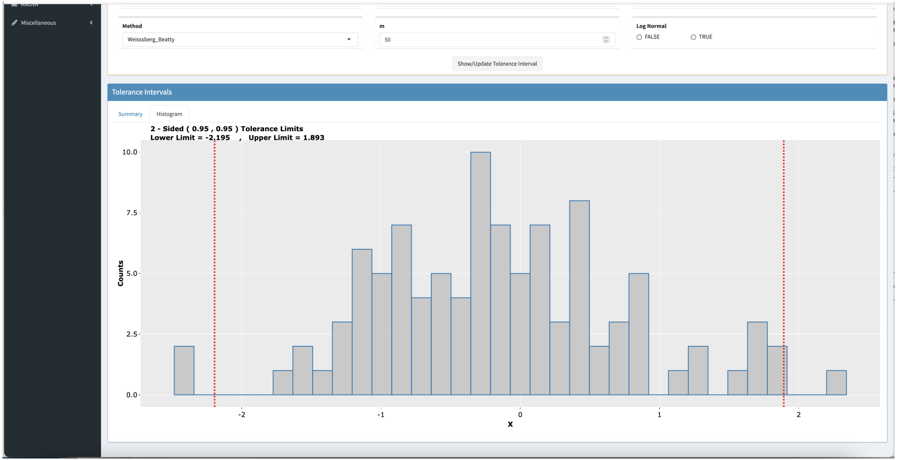
Method
Weisssberg_Beatty

m
50

Log Normal
○ FALSE ○ TRUE

The sample has a mean of -0.151 with a standard deviation of 0.915 .
The Two-Sided WBE (0.95 , 0.95) Tolerance Limit is (-2.195 , 1.893).
Its corresponding actual coverage is 0.96 .

Shiny App Demo (ctd.)



Method: Weisssberg_Beatty

m: 50

Log Normal: ○ FALSE ○ TRUE

2 - Sided (0.95 , 0.95) Tolerance Limits
Lower Limit = -2.195 , Upper Limit = 1.893

Resources

Links to Additional Resources for `tolerance`

Original Paper by Young (2010):

- https://www.jstatsoft.org/article/view/v036i05

`GitHub` Repo

- https://github.com/dsy109/tolerance

`Shiny` App

- https://tolerance.as.uky.edu/

References I

Environmental Protection Agency. (2006). Data Quality Assessment: Statistical Methods for Practitioners [Computer software manual]. Washington, DC, USA. Retrieved from http://www.epa.gov/sites/production/files/2015-08/documents/g9s-final.pdf

Hahn, G. J., & Chandra, R. (1981). Tolerance Intervals for Poisson and Binomial Random Variables. *Journal of Quality Technology*, *13*(2), 100–110.

International Atomic Energy Agency. (2008). Safety Report Series No. 52: Best Estimate Safety Analysis for Nuclear Plants: Uncertainty Evaluation [Computer software manual]. Vienna, Austria. Retrieved from http://www-pub.iaea.org/MTCD/publications/PDF/Pub1306_web.pdf

International Organization for Standardization. (2014). ISO 16269-6: Statistical Interpretation of Data – Part 6: Determination of Statistical Tolerance Intervals [Computer software manual]. Geneva, Switzerland. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=57191

References II

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

Wallis, W. A. (1951). Tolerance Intervals for Linear Regression. In J. Neyman (Ed.), *Second berkeley symposium on mathematical statistics and probability* (pp. 43–51). Berkeley, CA: University of California Press.

Young, D. S. (2010). tolerance: An R Package for Estimating Tolerance Intervals. *Journal of Statistical Software*, *36*(1), 1–39. Retrieved from http://www.jstatsoft.org/v36/i05/

Young, D. S. (2013). Regression Tolerance Intervals. *Communications in Statistics - Simulation and Computation*, *42*(9), 2040–2055.

Young, D. S. (2016). Normal Tolerance Interval Procedures in the tolerance Package. *The R Journal*, *8*(2), 200–212.

References III

Young, D. S., & Mathew, T. (2020). Nonparametric Hyperrectangular Tolerance and Prediction Regions for Setting Multivariate Reference Regions in Laboratory Medicine. *Statistical Methods in Medical Research*, *29*(12), 3569–3585.

Contact Information

✉ derek.young@uky.edu

🌐 http://young.as.uky.edu

https://github.com/dsy109