Bayesian Tolerance Region

Xiang-Yi (Gregory) Chen

Center for Mathematical Sciences,
Merck / MSD

Outline

- Tolerance interval is not a straightforward concept, even for statisticians
- However, it is of much more merits than 3-σ limits
- It is hard to implement it for a linear mixed model in a frequentist formulation
- Bayesian approach can provide a well-perform and easy-to-generalize solution

Outline

- Tolerance interval is not a straightforward concept, even for statisticians
- However, it is of much more merits than 3-σ limits
- It is hard to implement it for a linear mixed model
- Bayesian approach can provide a well-perform and easy-to-generalize solution

Notations

Y :	Observed data, say, $Y \in \mathbb{R}^n$
P_θ :	Probability distribution for Y, depending on θ
Z :	A future observation, say, $Z \in \mathbb{R}$
Q_θ:	Probability distribution for Z, depending on θ
Θ :	Parameter space of θ
$\pi(\cdot)$:	Prior on Θ
$\Pi(\cdot\,\|Y)$:	Posterior given observed data Y, on Θ

Tolerance Interval: Frequentist Definition

Treat data as random, formulated via a parametric family $\{P_\theta\colon \theta \in \Theta\}$ on $\mathbb{R}^n$.

$\mathcal{R}(Y)$ is a ***(c,δ)-tolerance region*** if

$$P_\theta\big(Y\colon Q_\theta\big(\mathcal{R}(Y)\big) \geq c\big) = \delta\,, \qquad \forall\theta.$$

("region" is usually replaced by "interval" when $Z \in \mathbb{R}$)

Difficulties:

- Complexity of the outer probability statement to non-statisticians
- Be aware that the lefthand side need to be independent of θ, formulate two-sided $\mathcal{R}(Y)$

* $Q_\theta\big(\mathcal{R}(Y)\big)$ denotes the probability of $\mathcal{R}(Y)$ under Q_θ

Tolerance Interval: Frequentist Definition (cont'd)

One-sided (c,δ)-tolerance region, where $\mathcal{R}(Y) = (-\infty, U(Y)]$ or $[L(Y), +\infty)$

If Z is real-valued, so that Q_θ is a distribution on $\mathbb{R}$,

- $(-\infty, U(Y)]$ *is a δ confidence interval for the induced parameter $Q_\theta^{-1}(c)$ *;*
- $[L(Y), +\infty)$ *is a δ confidence interval for $Q_\theta^{-1}(1-c)$*

Two-sided (c,δ)-tolerance region, where $\mathcal{R}(Y) = [L(Y), U(Y)]$

Let $v\big(\mathcal{R}(Y)\big)$ be the an criterion on the tolerance interval desirably to be optizimied, e.g. $v\big(\mathcal{R}(Y)\big)$ can be the size/length/volume of the tolerance interval

$$\arg\min_{\mathcal{R}(Y)} v\big(\mathcal{R}(Y)\big) \qquad s.t.\quad P_\theta(Y\colon Q_\theta(\mathcal{R}(Y)) \geq c) = \delta,\ \ \forall\theta$$

*Under normality** assumption and mutual independency among elements of Y, $\mathcal{R}(Y)$ is chosen to be of a form $\hat{\mu} \pm E$, where E is the uncertainy margin*

* $Q_\theta^{-1}(\cdot)$ *is the quantile function of* Q_θ

** *It can be generalized to any unimodal and symmetric distribution*

Tolerance Interval: Bayesian Definition

Treat θ as random, formulated via the posterior $\Pi(\cdot\,|Y)$ on Θ.
Probability is regarded as "degrees-of-believe" instead of "frequency".

$\mathcal{R}(Y)$ is a ***(c,δ)-tolerance region*** if

$$\Pi\left(\theta:\ Q_\theta\left(\mathcal{R}(Y)\right) \geq c \mid Y\right) = \delta,$$

assuming Y and Z are independent given θ.

- Easy to interpret, maybe more natural to engineers' intuition
- In case of two sided $\mathcal{R}(Y)$, it is easier than its frequentist counterpart to solve

  $$\arg min_{\mathcal{R}(Y)}\, v\left(\mathcal{R}(Y)\right) \qquad s.t.\quad \Pi(\theta: Q_\theta(\mathcal{R}(Y)) \geq c|Y) = \delta$$

  since it is coniditonal on observed data, but depends on the prior.

Outline

- Tolerance interval is not a straightforward concept, even for statisticians
- However, it is of much more merits than 3-σ limits
- It is hard to implement it for a linear mixed model
- Bayesian approach can provide a well-perform and easy-to-generalize solution

3-σ Limits

Under normality assumption, 3-σ limits ($\bar{Y}_n \pm 3s_n$) ***approximate*** the two-sided 99.73% quantiles. However, the accuracy of related probablity statement depends on the data size n, and it is complex to picture the convergence rate when Y is not a collection of mutually-independent results.

Trivial example:
Considering a data set Y of n independent repeats of certein experiment, let $\mathcal{R}(Y)$ take the form of 3-σ limits and work out the corresponding c or δ one at a time by fixing the other at certain value.

This exercise helps to draw some analogy between the 3-σ limits and a two-sided tolerance interval. One can extend the exercise to a mixed model.

Done by 100,000 simulations with varying n and $N(0{,}100)$ for the individual experiment. Since the form of 3-σ limits is independent of the true parameter θ_0 and the mutual independence assumption, the simulation results should hold $\forall\theta$.

3-σ Limits (cont'd)

Sample Size (n)	Avg. Coverage	$c = 99.73\%$ Approximated δ	$\delta = 99\%$ Approximated c
5	0.9483	0.343	0.531
6	0.9613	0.350	0.633
7	0.9688	0.356	0.703
8	0.9746	0.366	0.757
9	0.9781	0.372	0.795
10	0.9812	0.376	0.826
20	0.9913	0.406	0.935
30	0.9938	0.422	0.961
40	0.9949	0.431	0.972
50	0.9954	0.438	0.977
60	0.9958	0.443	0.981
70	0.9960	0.445	0.983
80	0.9962	0.450	0.985
90	0.9963	0.452	0.986

Outline

- Tolerance interval is not a straightforward concept, even for statisticians
- However, it is of much more merits than 3-σ limits
- It is hard to implement it for a linear mixed model
- Bayesian approach can provide a well-perform and easy-to-generalize solution

Frequentist Tolerance Interval: Mixed Model

Reference: *Gaurav Sharma and Thomas Mathew (JASA, 2012)*

It gave very good summary of existing works taking frequentist's perspective.

- Data : $Y \sim N\left(X\beta, \sum_{i=1}^{q-1} \sigma_i^2 K_i K_i' + \sigma_e^2 I_n\right), \quad Y \in \mathbb{R}^n$
- Future obs : $Z \sim N(x_0'\beta, z_0' D z_0 + \sigma_e^2)$

To derive the tolerance limits, introduce a notation for interested parameter

$$\varphi = x_0'\beta \pm z_q \sqrt{z_0' D z_0 + \sigma_e^2}, \text{ in one-sided scenario}$$

$$\varphi = x_0'\beta + E, \text{ in two-sided scenario.}$$

Rely on *modified signed log-likelihood ratio test statistic* for estimating $\hat{\varphi}$

Frequentist Tolerance Interval: Mixed Model

Reference: *Gaurav Sharma and Thomas Mathew (JASA, 2012), [cont'd]*

Parameterize $\theta = (\varphi, \lambda')'$, φ is a scalar parameter of interest, λ' is nuisance. Let $\hat{\lambda}_\varphi$ denote the constrained MLE of λ for a fixed φ, and $\hat{\theta}_\varphi = (\varphi, \lambda'_\varphi)'$.

Signed log-likelihood ratio test statistic

$$r(\varphi) = sign(\hat{\varphi} - \varphi)\left[2\left(\ell(\hat{\theta}) - \ell(\hat{\theta}_\varphi)\right)\right]^{1/2}$$

with first-order accuracy, i.e. error $\mathcal{O}\left(n^{-1/2}\right)$.

Modification made to this statistic for high-order accuracy, e.g. ($\mathcal{O}\left(n^{-3/2}\right)$ or $\mathcal{O}(n^{-1})$) ,

$$r(\varphi) + \frac{1}{r(\varphi)}\ln\left(\frac{q(\varphi)}{r(\varphi)}\right).$$

Different formulations of $q(\varphi)$ are needed for balanced and unbalanced data

Frequentist Tolerance Interval: Mixed Model

Reference: *Gaurav Sharma and Thomas Mathew (JASA, 2012), [cont'd]*

Pros:

- *It might be optimal regarding accuracy and convergence rate*
- *The modeling can be applied to virtually any linear mixed model, under regular assumptions*

Cons:

- *Very complicated in computational aspects and in turn hard to have a generalizable macro*
- *Not flexible when excusing the normality assumption*

Outline

- Tolerance interval is not a straightforward concept, even for statisticians
- However, it is of much more merits than 3-σ limits
- It is hard to implement it for a linear mixed model in a frequentist formulation
- Bayesian approach can provide a well-perform and easy-to-generalize solution

Bayesian Tolerance Interval: Mixed Model

Consider the same linear mixed model

- Data : $Y \sim N\left(X\beta, \sum_{i=1}^{q-1} \sigma_i^2 K_i K_i' + \sigma_e^2 I_n\right), \quad Y \in \mathbb{R}^n$
- Future obs : $Z \sim N(x_0'\beta, z_0' D z_0 + \sigma_e^2)$

The Bayesian formulation is very clear for the one-sided tolerance interval. Focus on the two-sided scenario, we shall try the same form as Sharma, i.e. Let $\mathcal{R}(Y)$ be of a form $x_0'\hat{\beta} + E$, where $\hat{\beta}$ is the estimate at its posterior mean and E is the uncertainy margin.

Solve the following for an estimate $\hat{E}$ so that the tolerance interval is $x_0'\hat{\beta} + \hat{E}$:

$$\Pi\left(Q_\theta^{-1}\left(x_0'\hat{\beta} + E\right) - Q_\theta^{-1}\left(x_0'\hat{\beta} - E\right) \geq c \mid Y\right) = \delta.$$

Wolfinger (1998) proposed a computational solution for two-sided Bayesian tolerance interval under similar formulation, and Krishnamoorthy & Mathew (2009) made a correction.

Bayesian Tolerance Interval: Mixed Model

Consider the same linear mixed model

- Data : $Y \sim N\left(X\beta, \sum_{i=1}^{q-1} \sigma_i^2 K_i K_i' + \sigma_e^2 I_n\right), \quad Y \in \mathbb{R}^n$
- Future obs : $Z \sim N(x_0'\beta, z_0' D z_0 + \sigma_e^2)$

The Bayesian formulation is very clear for the one-sided tolerance interval. Focus on the two-sided scenario, we shall try the same form as Sharma, i.e. Let $\mathcal{R}(Y)$ be of a form $x_0'\hat{\beta} + E$, where $\hat{\beta}$ is the estimate at its posterior mean and E is the uncertainty margin.

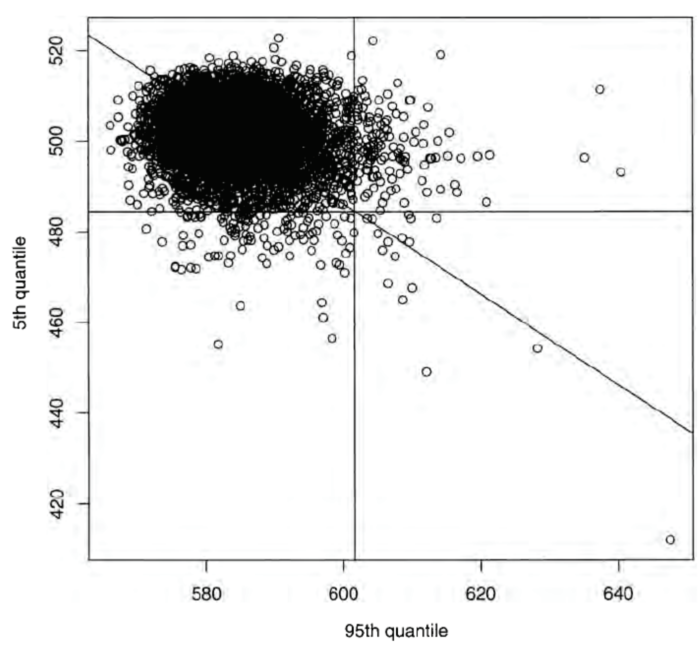
Solve the following for an estimate $\hat{E}$ so that the tolerance interval is $x_0'\hat{\beta} + \hat{E}$:

$$\Pi\left(Q_\theta^{-1}\left(x_0'\hat{\beta} + E\right) - Q_\theta^{-1}\left(x_0'\hat{\beta} - E\right) \geq c \mid Y\right) = \delta.$$

Wolfinger (1998) proposed a computational solution for two-sided Bayesian tolerance interval under similar formulation, and Krishnamoorthy & Mathew (2009) made a correction.

Bayesian Tolerance Interval: Mixed Model

Solution of Wolfinger (1998) and Krishnamoorthy & Mathew (2009)



Bayesian Tolerance Interval: Mixed Model

We propose another strategy for calculating

$$\Pi\left(Q_\theta^{-1}\left(x_0'\hat{\beta} + E\right) - Q_\theta^{-1}\left(x_0'\hat{\beta} - E\right) \geq c \mid Y\right) = \delta.$$

Simulation-based evaluation of the performances multiple estimators.

We take similar simulation scheme as Sharma et al. (2013, Table 1), so that we can compare the Bayesian and Frequentist estimators.

Simulation: One-way random effect model
6 runs, repeats per run = (2,3,4,2,3,4), $\beta_0 = 0$, $\sigma_{run}^2 = 1$,
Due to limitation of time, 1000 simulations for Bayesian approaches.

Bayesian Tolerance Interval: Mixed Model

($c = 0.90, \delta = 0.95$)- tolerance interval

	Bayesian (1000 simulations)			Frequentist (Sharma, 10000 simulations)	
Intra-Run Correlation	Wolfinger	K&M	OURS	$r(\varphi)$	$\tilde{r}(\varphi)$
0.10	0.866	0.994	***0.958***	0.9257	***0.9464***
0.30	0.824	0.981	***0.943***	0.9240	***0.9506***
0.50	0.820	0.976	***0.950***	0.9181	***0.9488***
0.70	0.859	0.975	***0.939***	0.9144	***0.9484***
0.90	0.875	0.970	***0.946***	0.9207	***0.9519***

Bayesian Tolerance Interval: Mixed Model

Pros:

- Easy to implement and fast calculation, for any linear mixed model
- Very comparable to the optimal frequentist counterpart
- It can be easily generalized to applications withou normality assumption

Cons:

- Dependent on the prior

Thank You !

Questions to gregory.chen@merck.com