Published on *IVT Network* (http://www.ivtnetwork.com)

Bayesian Tolerance Intervals for Zero-Inflated Data with Applications in Pharmaceutical Quality Control

By **Binbing Yu** ,

Harry Yang

Pin Ren **Jul 5, 2017 7:00 am PDT**

Declaration of Interest Statement: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper. The views of the authors are solely their own and do not necessarily reflect the views of MedImmune/AstraZeneca. ? Correspondence: Binbing Yu, Email: yub@medimmune.com.

Abstract

Quality control is an essential operation of the pharmaceutical industry. Tolerance intervals are commonly used in quality control to assure that a large proportion of critical quality attributes meet certain specification criteria with high confidence. Discrete measurements, such as microbial cell count, number of defective devices or discretized continuous responses due to limit of quantitation or rounding, are abundant in pharmaceutical manufacturing and quality control. Tolerance intervals for discrete variables are widely used in industrial appli- cations to set alert and action limits for critical quality attributes for the purpose of control and surveillance. A substantial amount of discrete data consist of excess number of zeros and often displays over-dispersion in variance. By far, there is no statistical method available

Keywords: Limit of quantitation, quality control, tolerance interval, zero-inflated distribution

1. Introduction

Quality control (QC) plays a critical role in pharmaceutical production, for both in-process and finished product testing [13]. The QC laboratories not only monitor and control the quality of incoming active pharmaceutical ingredients (APIs), and other supplies used in the manufacturing process; they are also instrumental in the batch release process. Regulatory authorities such as the FDA and EMA tightly control the manufacture of pharmaceutical and biopharmaceutical drugs [8].

Discrete quality attributes [1] are abundant in pharmaceutical manufacturing and quality control. Some discrete attributes are intrinsic count data. For example, microbial counts in raw materials, products, and water for pharmaceutical use (WPU), environmental monitoring (EM) microbial levels in production areas, number of complaints or number of noncon- forming items during the product quality monitoring. Other discrete attributes arise from the discretization of latent continuous variable due to rounding and limit of quantitation of assay results. For example, the antibody titer value in the enzyme-linked immunosorbent assay (ELISA), which is the reciprocal of the highest dilution where the readout occurs, is usually recorded as the fold of dilution 8, 16, 32, 64 etc. One common phenomenon of certain discrete quality attributes is the excess number of zeros. This type of quality attributes is usually called a zero-inflated variable. Alert and action limits in pharmaceutical manufacturing are a mandatory requirement for cGMP. These limits are used by the manufacturer to maintain effective control over the process. It is not necessary to take any action when assay results reach the alert limit, but the alert limit is an alarm that something might go wrong and certain preventive action is needed. The action limit is a trigger that mandatory action is needed to restore the process that might be out of control.

One common practice in the pharmaceutical industry involves using standard deviations to set the alert and action levels [2]. Tolerance intervals (TIs) are increasingly used in the development of quality control systems in the manufacturing and pharmaceutical industry [3, 17, 19, 22]. TI is an interval range to assure a certain proportion of the quality attribute covered within a pre-specified confidence [10]. The construction of TIs for continuous and discrete quality attributes have been discussed extensively [10, 23]. For example, Hahn and Chandra [7] investigated a general problem on controlling the number of unscheduled shut-downs of a complex system. Young [25] developed an R package tolerance for calculating

TIs for a wide range of distributions and commonly used regression models. However, the TIs for discrete data with excess number of zeros have not been well investigated. As zero- inflated discrete quality attributes are common in pharmaceutical manufacturing and quality control, it is useful to develop the TIs for discrete variables with zero-inflation. Although the non-parametric TIs can be used to set the alert and action limits, the reliability of the non-parametric approach depends on the amount of data [24]. In the presence of outlying data points caused by out-of-control conditions, the non-parametric method may produce extreme control limits with very wide ranges. Therefore, we propose using the Bayesian method to derive the tolerance interval of zero-inflated discrete data based on parametric zero-inflated distributions.

The rest of the paper is organized as follows. In Section 2, we present some zero-inflated discrete distributions used in pharmaceutical quality control, describe the Bayesian method for constructing TI, and later discuss some computational issues. A limited simulation study is presented in Section 3. The proposed method is applied to two applications from environ- mental monitoring and cleaning validation in Section 4. The paper ends with discussions and future directions in Section 5.

2. STATISTICAL METHOD

2.1 Zero-inflated discrete distributions

Count data is in abundance in clinical trials, pharmaceutical development, and manufacturing. In preclinical or clinical trials, counts like number of asthma exacerbations per week, or number of migraine headaches per day are collected for measuring the severity of symptoms. These types of data are intrinsic integers ranging from 0 to infinity. Another type of discrete data come from discretization of continuous responses. The reportable values from many analytical methods are integers multiplied by the smallest usable unit [15]. For example, relative potency is often rounded to the nearest integer and maximum daily drug doses may be rounded according to certain rules [9]. In addition, analytical methods may be subject to limits of quantitation (LOQ), which is the smallest concentration or signals that can be reliably measured. The assay result under the LOQ is often reported as <LOQ or simply coded as zero. This type of zero is actually an interval-censored observation with range [0, LOQ).

Let X denote the discrete random variable. Poisson and negative binomial distributions are two popular choices for modeling discrete data. The probability function for a Poisson distribution with a mean ? is given by

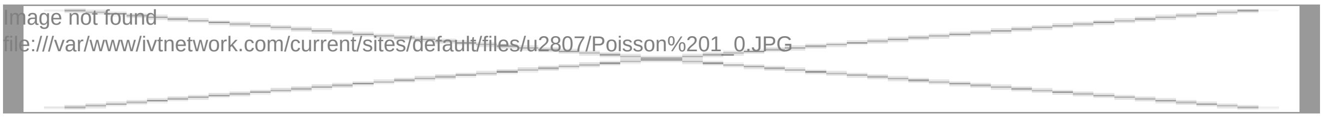
Image not found
file:///var/www/ivtnetwork.com/current/sites/default/files/u2807/Poisson%201_0.JPG

The probability function of a negative binomial distribution is

Image not found
file:///var/www/ivtnetwork.com/current/sites/default/files/u2807/Poisson%202_1.JPG

The mean and the variance of the negative binomial distribution are ? (1 ? ?)/? and ? (1 ??)/?2. A typical scenario of using the negative binomial distribution is during the inspection of medical devices with defect probability ?, the number of total devices, which is equal to ? + x, to be inspected in order to detect ? defected devices follows the negative binomial distribution.

If there is excess number of zeros in the responses, zero-inflated Poisson (ZIP) or zero- inflated negative binomial (ZINB) distributions are often used [11, 18]. The ZIP and ZINB distributions have the following probability function:

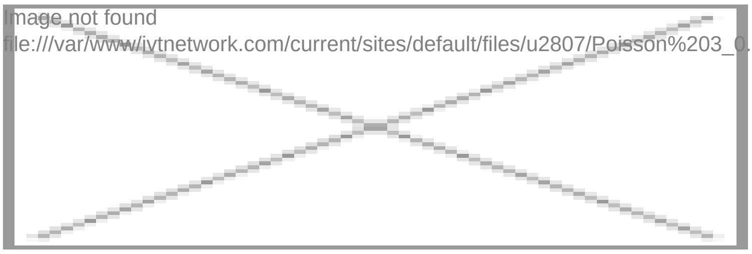
Image not found
file:///var/www/ivtnetwork.com/current/sites/default/files/u2807/Poisson%203_0.JPG

where p is the probability of excess number of zeros, f (y|?) is the regular density function with parameter ? for discrete variables, e.g., the Poisson or negative binomial distributions.

2.2 Bayesian Tolerance Interval

In this section, we briefly review the definition of TI and describe the Bayesian approach for obtaining the TI for discrete response variables with zero inflation. Let F denote the cumulative distribution for the random variable X with parameter ?. A ?-content, 1 ? ? confidence TI is defined as the interval (LX , UX) such that

P r? {[F (UX) ? F (LX)] ? ?} = 1 ? ?. (3)

By definition, with a specified degree of confidence, 1 ? ?, that a specified proportion ? or more of the quality attributes lie between LX and UX .

For a continuous distribution, there may exist a TI (LX , UX) that satisfies (3) for all parameters ?. But for a discrete distribution, the definition of TI can be modified as

P r? {[F (UX) ? F (LX)] ? ?} ? 1 ? ?. (4)

and there exists a ? such that the equality holds. As the higher counts typically indicates a higher risk of failure in pharmaceutical quality control, we focus on the one-sided upper tolerance limit UX , where

P r? {F (UX) ? ?} ? 1 ? ?. (5)

Given the observed values x = (x1 , ..., xn) of random variable X , the (?, 1 ? ?) upper tolerance limit can be obtained using the following two-steps:

- Construct a one-sided (1??) upper confidence interval (0, u) for ?, where u is a function of x.
- Find a minimum value Ux such that P ru(X ? Ux) ? ?.

Although this provides a simple recipe for obtaining upper tolerance limit for common dis- tributions like Poisson, binomial, the derivation of tolerance limits for zero-inflated discrete data is not straightforward. Krishnamoorthy and Mathew [10] provided a simple solution for obtaining the Bayesian tolerance intervals. The Bayesian one-sided upper tolerance intervals can be constructed as follows. Let q? (?) denote the ? quantile of X . Under the Bayesian framework, the 1 ? ? upper confidence limit for q? (?) can be obtained based on the posterior distribution q? (?) given the observed data x. The upper (?, 1??) tolerance limit UX satisfies that

P r[q? (?) ? UX] = 1 ? ?, (6)

where the probability is computed with respect to the posterior distribution f1(q? (?)|x). When there is no closed analytic form for the posterior distribution f1(q? (?)|x), Bayesian simulation method can be used. Let ?1, ..., ?S denote the simulated samples of parameter ? from the posterior distribution f1(?|x). The corresponding ? quantile of X can be computed as q? (?1), q? (?2), ..., q? (?S). The upper tolerance limits can be obtained as 100(1 ? ?)th percentile of the q? (?i), i = 1, ..., S.

2.3 Computational issue

Under the Bayesian formulation, the posterior distribution f1 (?|x) is proportional to the product of likelihood function L(x|?) and the prior distribution f0 (?):

f1 (?|x) ? L(x|?)f0(?). (7)

The posterior distribution can be obtained using the Bayesian Markov chain Monte Carlo (MCMC) method [4], which has been implemented in public available software OpenBUGS (openbugs.net) [12].

To simulate the random variable from the ZIP distribution, one can use the mixed Bernoulli-Poisson representation [5], where

X |D ? Poisson(?(1 ? D)) with D ? Bernoulli(p). (8)

By a reparametrization ? = ? (1 ? ?)/?, the zero-inflated negative binomial model can be derived by a mixed representation as well

X |D ? PN B (x|?, ?) with ? = ? /(? + ?(1 ? D)), D ? Bernoulli(p). (9)

With the mixed representation, the sampling of the ZIP and ZINB random variable can be easily implemented in OpenBUGS. The interval-censoring of continuous random variable can be implemented using the I (a, b) function in WinBUGS or C (a, b) in OpenBUGS. For example, a negative binomial random variable under the LOQ can be generated as x?dnegbin(?, ?)I(0,LOQ).

To complete specification of a Bayesian model it is necessary to choose prior distributions.

For the probability p of excess number of zeros, we specify normal prior for the logit of p, i.e., logit(p) ? N (0, 1000). For the ZIP distribution, the prior for the mean parameter is ? ? Gamma(0.01, 0.01). For the ZINB distribution, we specify the priors ? ? Gamma(0.01, 0.01) and ? ? Gamma(0.01, 0.01), where Gamma(a, b) is the Gamma distribution with shape parameter a and scale parameter b. The mean and variance of the Gamma(a, b) distribution is a/b and a/b2. The sample OpenBUGS code for fitting the ZINB distribution is shown in the Appendix.

3. SIMULATION

In this section, we conduct a simulation study to evaluate the performance of the proposed Bayesian tolerance limits. The underlying distribution is ZINB with parameters ? = (p, ?, ?). The parameters are chosen to mimic the distributions in the two case studies. The zero- inflation probability p = 0.1 or 0.2. The parameter ? = 1.5 or 3 and the parameter ? = 0.1, so the probability of observing an event is low (? 6.25%). For each parameter combination,

Table 1: Coverage rate of the (?, 1 ? ?) tolerance limits of containing the ? quantile q?

			Content ? =0.95				Content ? =0.99			
p	?	n	? =0.80	? =0.90	? =0.95	? =0.99	? =0.80	? =0.90	? =0.95	? =0.99
0.1	1.5	250	81.2	91.3	96.2	99.2	79.4	90.2	94.4	98.8
		500	81.4	91.8	95.2	99.6	81.6	91.4	95.6	98.6
	1.5	250	81.0	91.2	95.8	99.2	79.0	91.6	95.0	99.6
		500	82.2	92.4	96.0	99.4	79.6	90.6	95.4	99.2
0.2	3	250	81.6	91.8	96.6	99.8	80.0	92.0	96.0	99.4
		500	81.4	91.4	96.2	99.2	82.2	90.8	95.2	99.2
	3	250	80.2	90.5	96.4	99.4	78.4	89.2	94.6	99.4
		500	81.4	90.4	96.6	99.8	79.6	90.4	95.2	99.1

two data sets are generated, one has sample size n = 250 and the other has n = 500. The ? content and (1 ? ?) upper tolerance limits are obtained for each data sets.

The simulation is replicated 1000 times to calculate the rate that the upper (?, 1 ? ?) tolerance limits cover the ? quantile. Table 1 shows the actual coverage rates of (?, 1 ? ?) upper tolerance limits with respect to different values of p, ?, n. The actual coverage rates are very close to, but slightly above, the nominal levels for the values considered in the simulation. This indicates that the Bayesian TIs for the ZINB distribution tend to be slightly conservative.

4. Applications

4.1 Environmental Monitoring

The primary purpose of an environmental monitoring program is to provide oversight for microbiological cleanliness of manufacturing operation and document the state of control of the facility. A key to the success of the program is the establishment of alert and action limits. In practice, several statistical methods including normal, Poisson, and negative binomial modeling which have been routinely used to set these limits. However, data collected from clean

Image not found
file:///var/www/ivtnetwork.com/current/sites/default/files/u2807/binbing_Figure%201_0.JPG

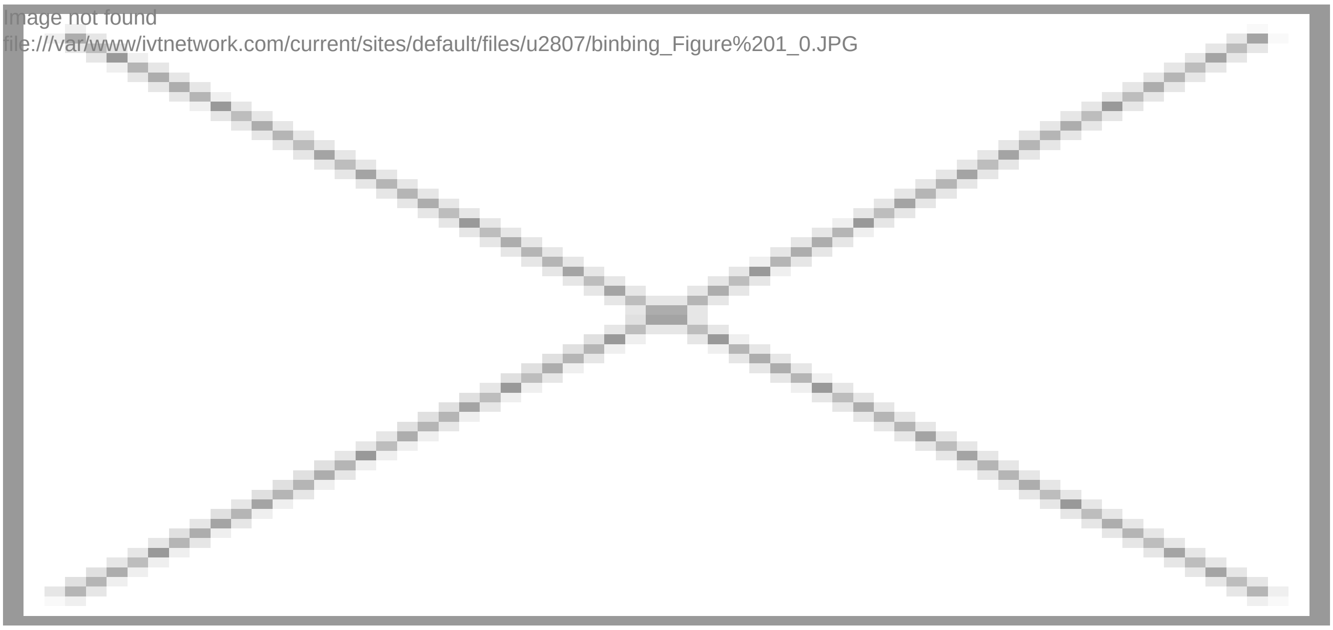

Figure 1: Histogram and estimated probability density for microbial cell counts

rooms or controlled locations often display an excess of zeros and overdispersion, caused by sampling population heterogeneity. Therefore, it is not appropriate to use the traditional methods to set alert and action levels. The ZINB method showed a clear improvement in terms of model fitting and parameter estimation. Therefore, Yang et al. [24] proposed to use the 95% and 99% percentile from a ZINB distribution model to set the alert and action limits for microbial counts of samples taken in a monitoring cycle. A more meaningful approach is to use the upper tolerance limits by controlling both the proportion of microbial counts and the confidence level.

We then analyzed the microbial cell counts from a typical environment study. The proportion of zero cell count is above 0.6, indicating a zero-inflated distribution. A ZINB distribution was fitted to the observed cell count data. Figure 1 shows the histogram of the observed microbial cell counts (red bars) and the estimated probability density from the ZINB distribution (blue bars). We can see that the estimated probabilities are very close to the observed cell count frequency, indicating a good fit. The (95%, 95%) and (95%, 99%) upper tolerance limits for cell counts are 8 and 9, respectively.

4.2 Cleaning Validation

Cleaning, sanitization and maintenance (Code of Federal Regulations Title 21 Part 211.67) are among the 10 most cited observations for drug inspection [6]. A series of articles have been published to discuss the methodology and practices in pharmaceutical cleaning validation [2, 20, 21]. Mott et al. [14] proposed several approaches for setting control limits of process residual in cleaning validation on the premises that Active Pharmaceutical Ingredient (API) inactivation by the cleaning process has been demonstrated. Performance Control Limits may be used if the cleaning validation studies have been completed and routine cleaning consistently demonstrates the equipment cleaning process removes process residue below the acceptance limits, especially if the residual is considerably lower than the acceptance limit [14]. The Performance Control Limit, sometimes referred to as an alert limit, enables detection of a change in the performance of the cleaning process and allows for a proactive investigation into a potential cleaning process issue. The alert limit is calculated from the total organic carbon (TOC) data collected from routine cleaning studies.

The current statistical methods for deriving performance control limits are based on the assumption of normality and independence of the TOC results. However, data generated from effective cleaning processes are usually not normally distributed, as shown in the example of cleaning verification data in Mott et al. [14]. Mott et al. suggest some transformations like Box-Cox transformation to normalize data. The Box-Cox transformation can only applied to non-zero values. In fact, a substantial number of TOC results are set to be zero because of the limit of quantitation. Therefore, a non-normal distribution with zero-inflation is more appropriate.

We analyze the TOC data from a cleaning validation study for a bioreactor. The TOC data are obtained by swabbing multiple locations from a bioreactor tank. The LOQ is 6μg per swab. Majority (93.7%) of the TOC results are below the

LOQ. The ZINB model is used to fit the observed TOC data and the results below LOQ are treated as interval censored [0, 6). Figure 2 compares the normal Q-Q plot and the ZINB Q-Q plot for the

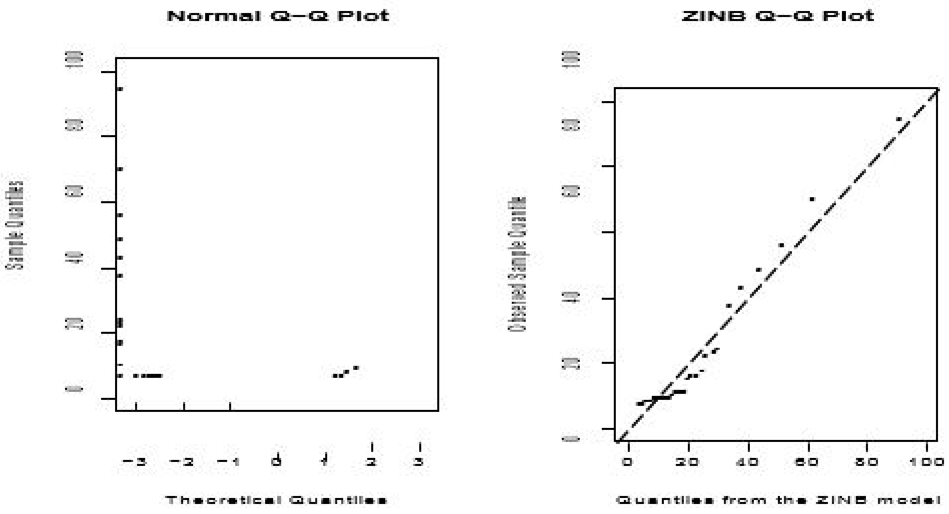

Figure 2: Normal and ZINB Q-Q plots for the observed TOC data observed TOC data

The observed data are significantly deviated from normal, while the ZINB distribution provides a much better fit. The estimated proportion of below LOQ is 92.8%, which is very close the observed proportion of below LOQ. The (95%, 95%) and (95%, 99%) upper tolerance limits for the TOC value are 11µg and 13µg per swab, respectively. Therefore, we set 11 and 13 as the alert and action limits for bioreactor TOC cleaning validation.

5. Discussion

In this article, we propose a Bayesian approach for obtaining the TIs for discrete responses with zero-inflation. The upper tolerance limits can be used to set the alert and action limits for pharmaceutical quality control. The proposed method is easy to implemented and tends to be slightly conservative. The Bayesian tolerance limits may by used by QC scientists for more vigilant monitoring of manufacturing processes. In the analysis of trend, the count data usually exhibits serial independence [16]. The Bayesian tolerance limits can be constructed for correlated zero-inflated count data by considering the autocorrelation structure.

APPENDIX A. SAMPLE OPENBUGS CODE FOR FITTING ZINB DISTRIBUTION

```
model{

# probability of zero-inflation logit(p)<-logit.p


# X(1:N0) are below LOQ

for(i in 1:N0){ X[i]~dnegbin(P.NB[i],tau)I(0,LOQ)


}
# X((N0+1):N) are positive TOC result for (i in (N0+1):N) {

X[i]~dnegbin(P.NB[i],tau)


}


# Bernoulli random variables d[i] are used to create zero-inflation for (i in 1:N) {

P.NB[i]<-tau/(tau+lambda*(1-d[i]))


d[i]~dbern(p)

}

# priors for the parameters logit.p~dnorm(0,0.01) tau~dgamma(0.1,0.1) lambda~dgamma(0.1,0.1)

}
```

REFERENCES

[1] Cholayudth, P. [2007], "Application of Poisson distribution in establishing control limits for discrete quality attributes," Journal of Validation Technology, 13(3), 196–205.

[2] Cholayudth, P. [2013], "Establishing a complete set of target, alert, and action limits for microbial counts in purified water," Journal of Validation Technology, 19, 44–53.

[3] Dong, X., Tsong, Y., Shen, M., and Zhong, J. [2015], "Using tolerance interval- s for assessment of pharmaceutical quality," Journal of Biopharmaceutical Statistics, 25(2), 317–327.

[4] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. [2013], Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC Texts in Statistical Science Taylor & Francis.

[5] Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. J. [2006], "Bayesian analysis of zero- inflated regression models," Journal of Statistical planning and Inference, 136(4), 1360–1375.

[6] Gietl, M. [2013], "Cleaning validation residue limits – how clean is clean?," Journal ofValidation Technology, 19(1).

[7] Hahn, G. J., and Chandra, R. [1981], "Tolerance intervals for Poisson and binomial variables," Journal of Quality Technology, 13(2), 100–110.

[8] Haleem, R. M., Salem, M. Y., Fatahallah, F. A., and Abdelfattah, L. E. [2015], "Quality in the pharmaceutical industry–a literature review," Saudi Pharmaceutical Journal, 23(5), 463–469.

[9] ICH [2006], Q3B(R2) Impurities in new drug products.

[10] Krishnamoorthy, K., and Mathew, T. [2009], Statistical Tolerance Regions: Theory, Applications, and Computation John Wiley & Sons.

[11] Lambert, D. [1992], "Zero-inflated Poisson regression, with an application to defects in manufacturing," Technometrics, 34(1), 1–14.

[12] Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. [2009], "The BUGS project: Evolution, critique and future directions," Statistics in Medicine, 28(25), 3049–3067.

[13] May, M. [2014], “Leaning the quality control laboratory,” Pharmaceutical Manufacturing, .

[14] Mott, A., Henry, B., Wyman, E., Bellorado, K., Blu¨mel, M., Parks, M., Hayes, R., Runkle, S., and Luo, W. [2013], “Methodology for assessing product inactivation during cleaning Part II: setting acceptance limits of biopharmaceutical product carryover for equipment cleaning,” Journal of Validation Technology, 19(4), 20–26.

[15] Porter, W. R. [2014], “Measurement Uncertainty and Specifications for Kinetic Stud- ies.,” Journal of Validation Technology, 20(2).

[16] Rakitzis, A. C., Weiß, C. H., and Castagliola, P. [2017], “Control Charts for Monitoring Correlated Poisson Counts with an Excessive Number of Zeros,” Quality and Reliability Engineering International, 33(2), 413–430.

[17] Rebafka, T., Cl´emen¸con, S., and Feinberg, M. [2007], “Bootstrap-based tolerance in- tervals for application to method validation,” Chemometrics and Intelligent Laboratory Systems, 89(2), 69–81.

[18] Ridout, M., Dem´etrio, C. G., and Hinde, J. [1998], Models for count data with many zeros,, in Proceedings of the XIXth international biometric conference, Vol. 19, pp. 179–192.

[19] Rozet, E., Lebrun, P., Michiels, J.-F., Sondag, P., Scherder, T., and Boulanger, B. [2015], “Analytical procedure validation and the quality by design paradigm,” Journal of biopharmaceutical statistics, 25(2), 260–268.

[20] Sharnez, R., Spencer, A., Bussiere, J., Mytych, D., To, A., and Tholudur, A. [2013], “Biopharmaceutical Cleaning Validation: Acceptance Limits for Inactivated Product Based on Gelatin as a Reference Impurity.,” Journal of Validation Technology, 19(1).

[21] Spencer, A., Romero, J., Runkle, S., Carolan, C., Mott, A., Clark, M. E., Wyman, E., Rasmi, M., Donat, S., and Bellorado, K. [2012], “Methodology for assessing product inactivation during cleaning Part I: experimental approach and analytical methods,” Journal of Validation Technology, 18(4), 16–19.

[22] Von Collani, E., and Baur, K. [2002], “Prediction Intervals, Tolerance Intervals and Standards in Quality Control-Part 1,” Economic Quality Control, 17(1), 81–98.

[23] Wang, H., and Tsung, F. [2009], “Tolerance intervals with improved coverage probabilities for binomial and poisson variables,” Technometrics, 51(1), 25–33.

[24] Yang, H., Zhao, W., O’day, T., and Fleming, W. [2013], “Environmental monitoring: Setting alert and action limits based on a zero-inflated model,” PDA Journal of Pharmaceutical Science and Technology, 67(1), 2–8.

[25] Young, D. [2010], “Tolerance: An R package for estimating tolerance intervals,” Journal of Statistical Software, 36(5), 1–39.

Source URL: http://www.ivtnetwork.com/article/bayesian-tolerance-intervals-zero-inflated-data-applications-pharmaceutical-quality-control