COMPARISON OF COUNT MODELING TECHNIQUES FOR ESTIMATING ENVIRONMENTAL MONITORING LIMITS IN CLEAN ROOMS

By Plinio De los Santos, Ji Young Kim, Pieta IJzerman-Boon, George Kariuki and Brandye Smith-Goettler

Center for Mathematical Sciences, Merck Manufacturing Division

June, 2019

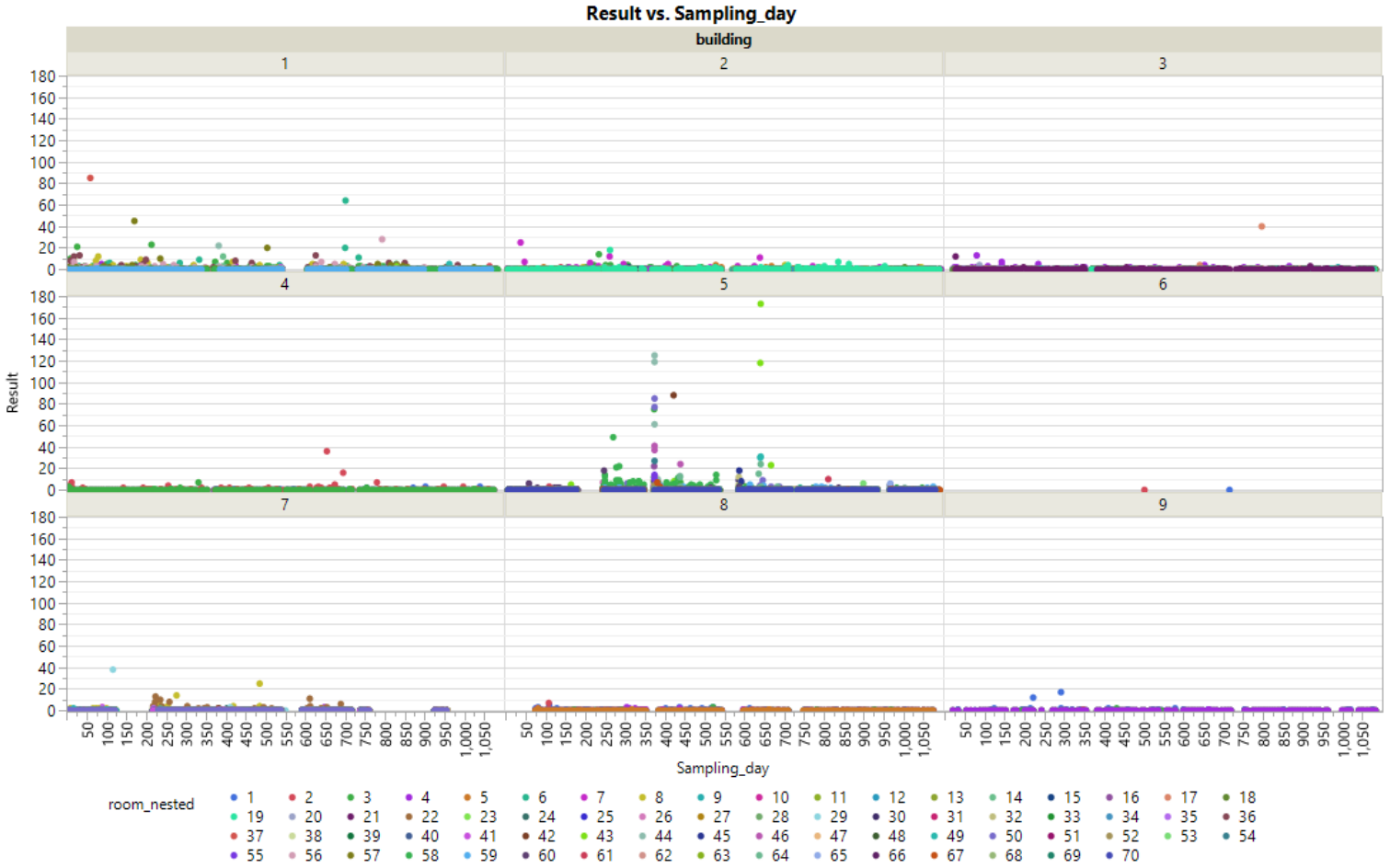
Objective

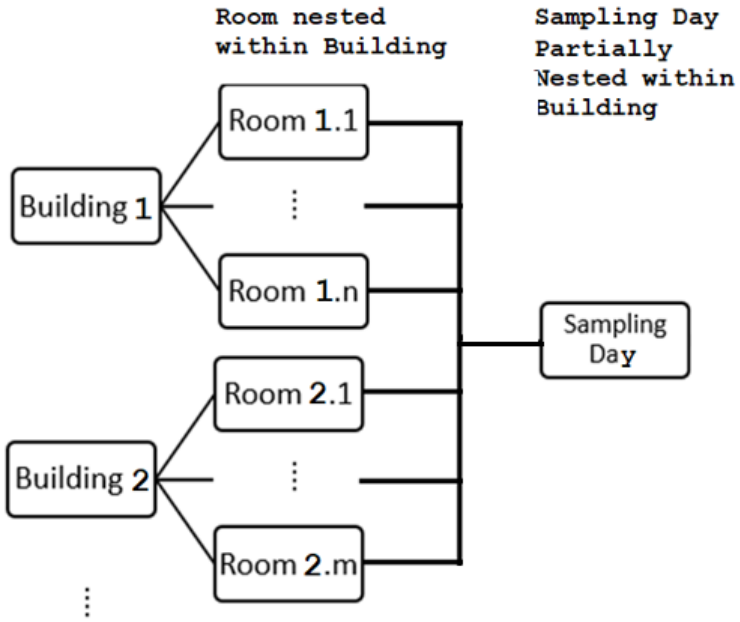
Pharmaceutical and biotechnology industries manufacture their products in clean rooms, which are designed to hold low levels of particulates (like microorganisms recovered from the air or from the clean room surfaces).

- Alert and action limits are employed to monitor and control the state of the room, keeping the level of particulates at appropriate levels.
- Particulate monitoring systems could generate count data with the following characteristics: are repeated counts subject to nested data structures, could be inflated at zero or at low counts, and on instances could exhibit long thin tails to the right with potential outliers.

During the presentation we will compare multiple statistical modeling techniques for setting alert and action limits using environmental monitoring data, to better understand the strengths and limitations of these techniques.

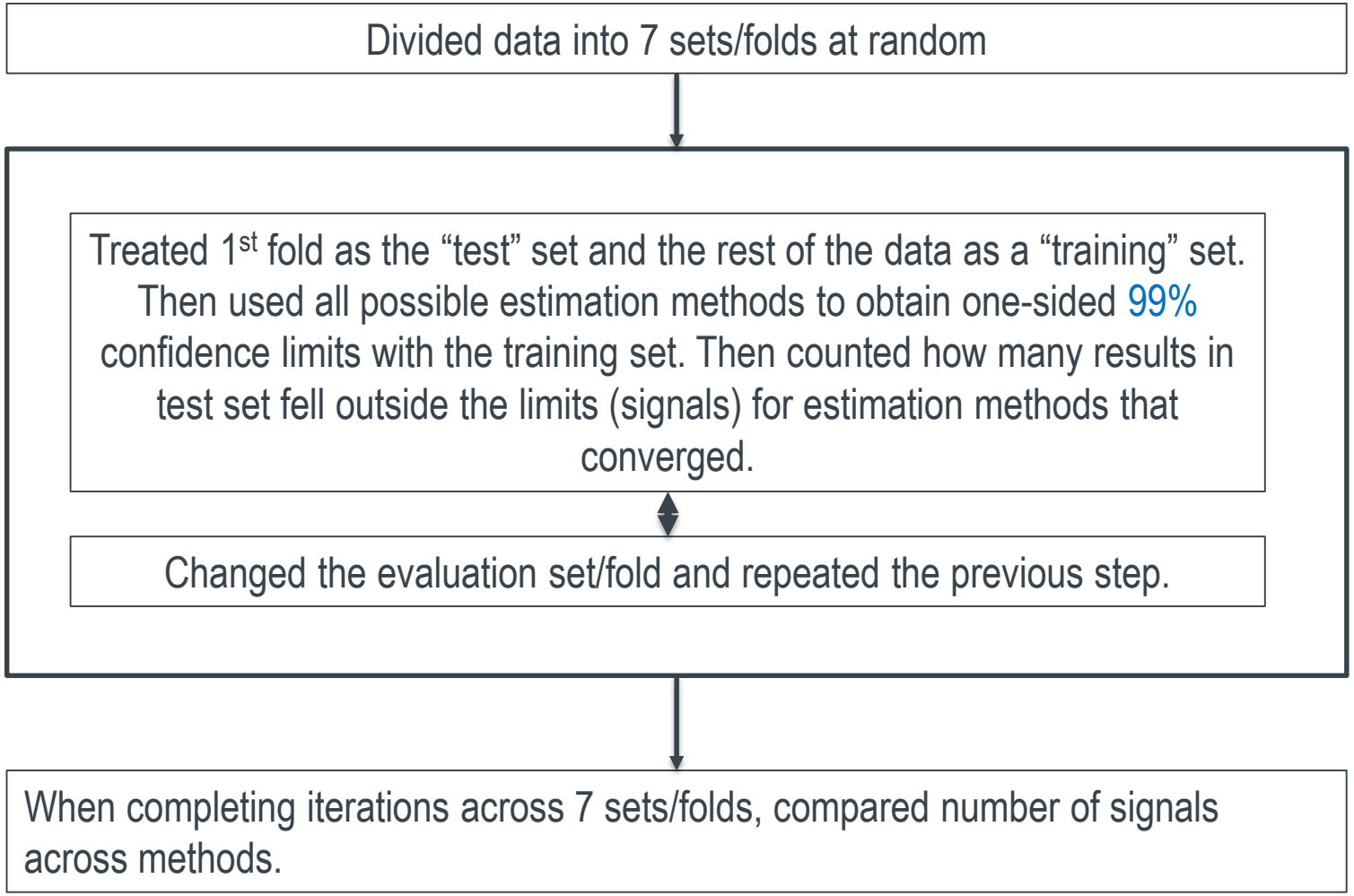
Evaluated Dataset with 57,175 Data Rows…





MERCK
INVENTING FOR LIFE

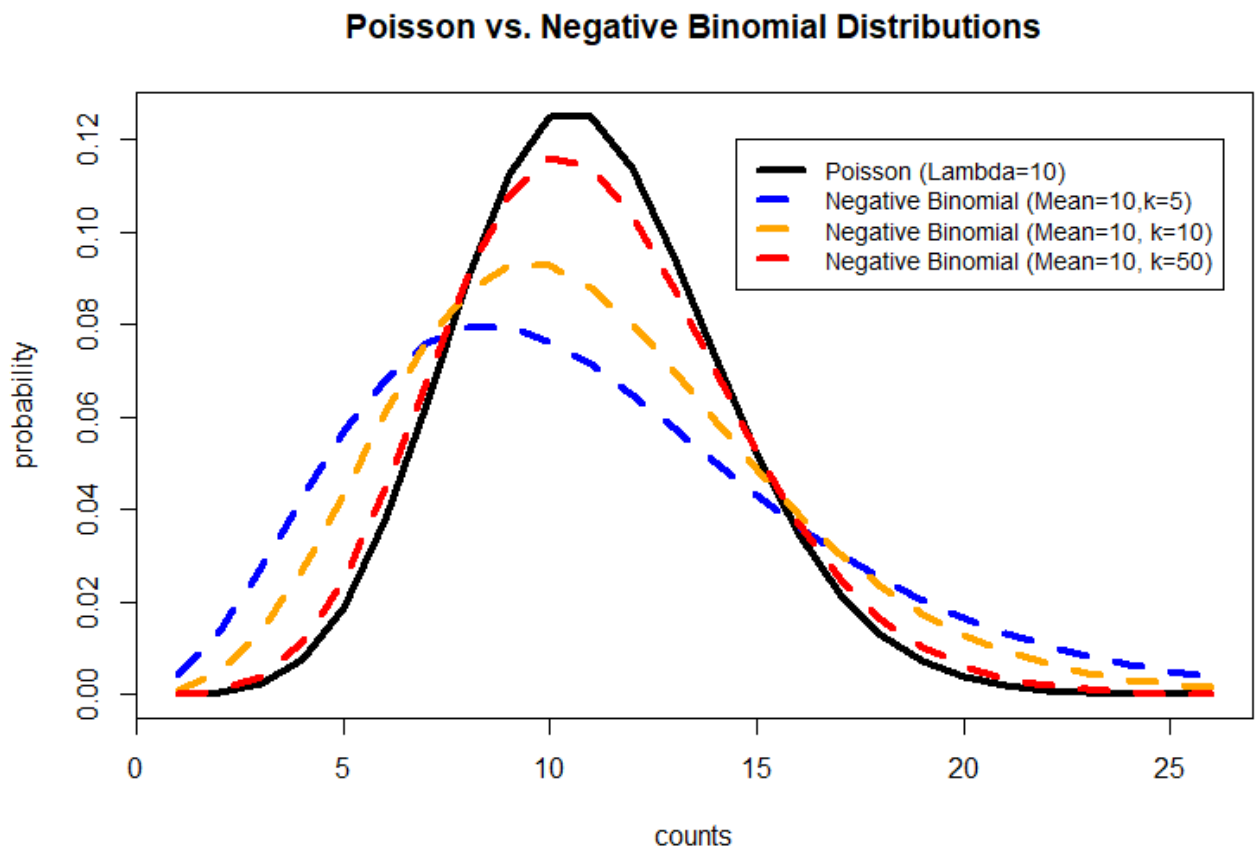
7-Fold Cross Validation Schema Employed During the Estimation



Comparison Methods

Method	Process	Distributions
Traditional	Fit a distribution to the data, and get an upper limit.	Poisson, Negative Binomial, Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB)
Parametric Bootstrap	Fit a distribution to the data.	Poisson, Negative Binomial, Zero-Inflated Poisson (ZIP)
	Resample the assumed distribution B times using a sample size "n" and get each time a percentile based limit. ("n" equals the observed # of samples in the room evaluated.)	
	Set the environmental monitoring limit equal to the median of the B percentile based limits.	
Bayesian	Iteratively fit the distribution to the data. (Employed non-informative prior)	Zero - and One - Inflated Poisson (ZOIP): Counts show a noticeable inflation at zero, and a relatively minor inflation at one as well.
	At each iteration, obtain distribution parameter estimates and obtain a percentile based limit.	
	Set the environmental monitoring limit equal to the median of the percentile based limits.	

Comparison of Poisson and Negative Binomial Mass Functions



Poisson Distribution (λ):

$$V[X] = E[X] = \mu \ \ (a.k.a. "\lambda")$$

Negative Binomial (k, p):

$$E[X] = \mu$$

$$V[X] = \mu\left(1 + \frac{\mu}{k}\right)$$

$$\hat{p} = \frac{k}{\mu + k}$$

Zero Inflation

A zero-inflated probability distribution is one that allows for recurrent zero values. This is:

$$g(x) = P(X = x) = \begin{cases} \theta_0 + (1 - \theta_0) f(x), & for\ x = 0 \\ (1 - \theta_0) f(x), & for\ x > 0 \end{cases}$$

When the particular probability mass function "$f(x)$" is:

- Poisson distributed, we get a zero-inflated Poisson (or ZIP) distribution.
- Negative Binomial distributed, we get a zero-inflated Negative Binomial (or ZINB) distribution.

Parameter Estimation Using Method of Moments (MOM) for Parametric Bootstrap Method

Distribution	Parameter Estimator by MOM
Poisson (λ)	$\hat{\lambda} = \bar{x}$, where: "λ" is the expected Poisson count.
Negative Binomial (k, μ)	$\hat{k} = \frac{\bar{x}^2}{S^2 - \bar{x}}$, $\hat{\mu} = \bar{x}$, where: "k" is the dispersion parameter, and "μ" is the mean observed counts. A parameter "p" can be estimated as a function of the previous two parameters: $\hat{p} = \frac{\hat{k}}{\bar{x} + \hat{k}}$
Zero Inflated Poisson "ZIP" (λ, θ_0)	$\hat{\lambda} = \frac{S^2 + \bar{x}^2}{\bar{x}} - 1, \hat{\theta}_0 = \frac{S^2 - \bar{x}}{S^2 + \bar{x}^2 - \bar{x}}$, where: "λ" is the expected Poisson count and "θ_0" is the probability of extra zeros.

Zero- and One-inflated Poisson (ZOIP) Distribution Employed for Bayesian Implementation

In this method, the dispersion of the distribution is modeled by a Poisson distribution, with inflation parameters at zero and at one:

X ~ Zero- and One- inflated Poisson $(\lambda, \theta_0, \theta_1)$

θ_0: Zero-inflation probability (0, 1)

θ_1: One probability (0, 1)

f(x): probability mass function for Poisson(λ), i.e. $\mathrm{f(x)} = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, \cdots$

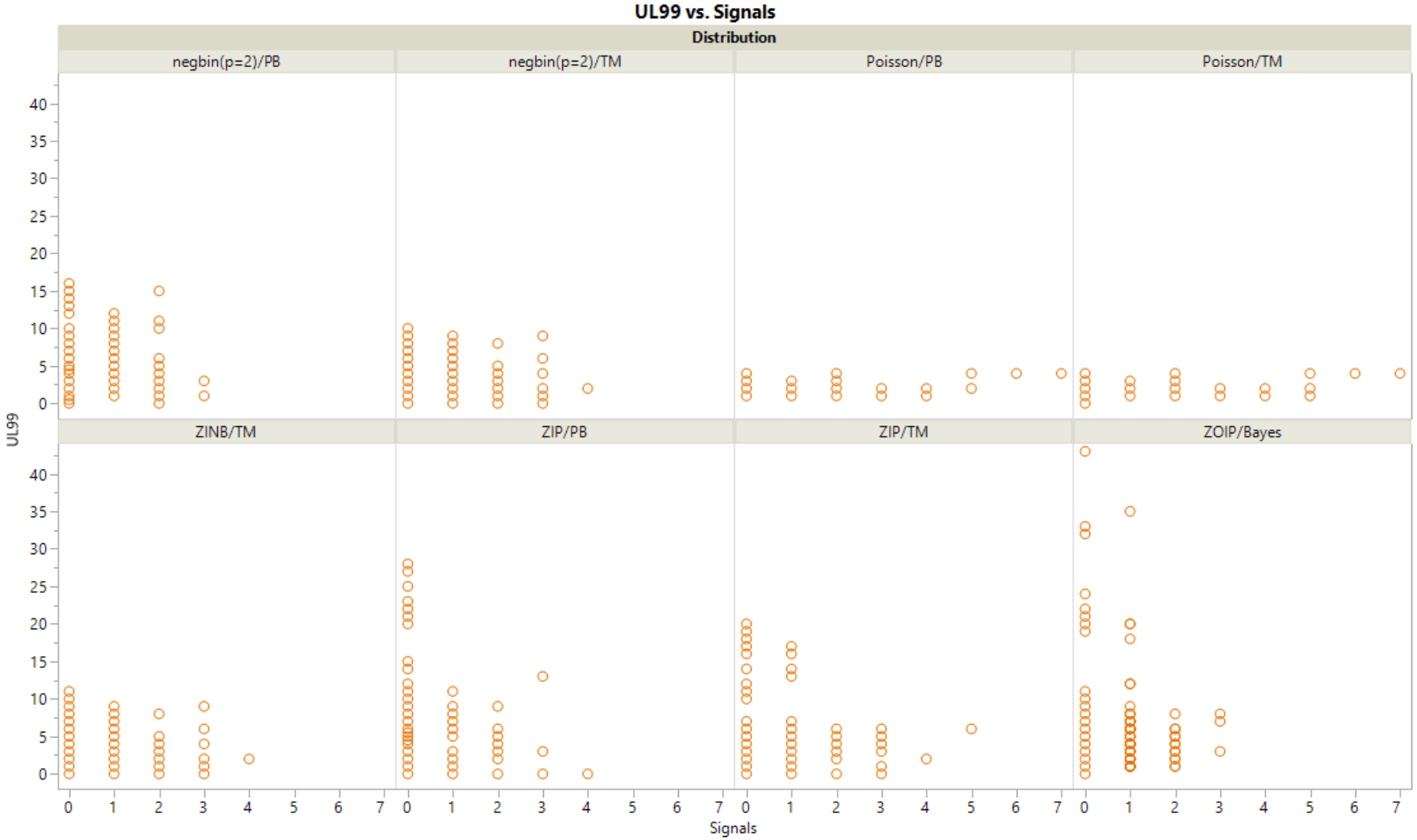
$$g(x) = P(X = x) = \begin{cases} \theta_0 + (1 - \theta_0 - \theta_1) f(0) = \theta_0 + (1 - \theta_0 - \theta_1) e^{-\lambda}, & x = 0 \\ \theta_1 + (1 - \theta_0 - \theta_1) f(1) = \theta_1 + (1 - \theta_0 - \theta_1) \lambda e^{-\lambda}, & x = 1 \\ (1 - \theta_0 - \theta_1) f(x) = (1 - \theta_0 - \theta_1) \frac{\lambda^x e^{-\lambda}}{x!}, & x > 1 \end{cases}$$

Number of Cases that Converged Across Methods

Traditional Methods ("TM")	Parametric Bootstrap ("PB")	Bayesian ("Bayes")
Poisson = 1,297 cases	Poisson = 1,151 cases	ZOIP = 601 cases
Negative Binomial (e.g., negbin(2)) = 815 cases	Negative Binomial (e.g., negbin(2)) = 671 cases	
ZIP = 1,278 cases	ZIP = 1,297 cases	
ZINB = 906 cases		

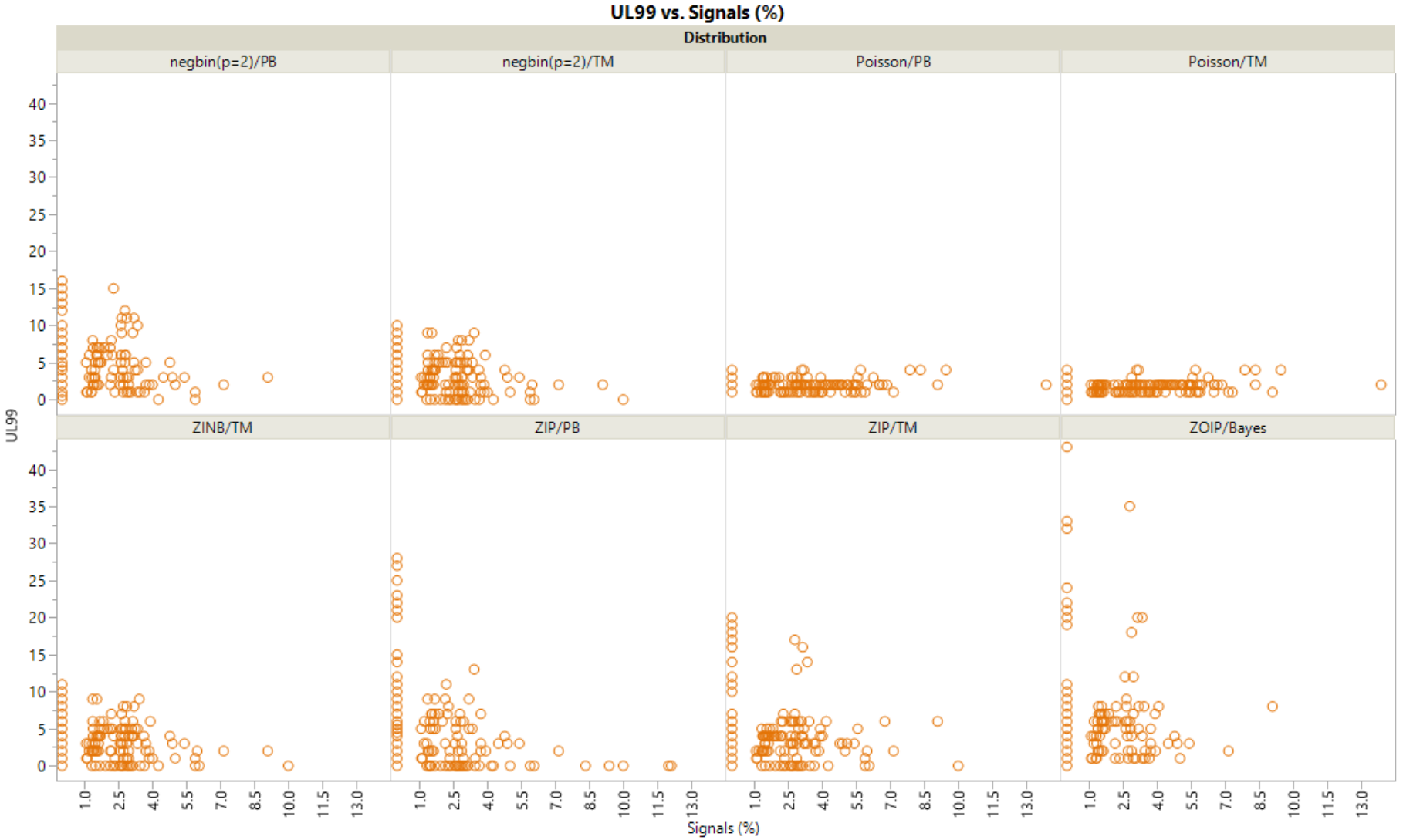
Note: since only 360 cases converged across all the methods, only these cases were used for the comparison.

Limits vs. Signals for 360 Evaluation Cases



MERCK
INVENTING FOR LIFE

Limits vs. Signals (as % of Results/ Room) for 360 Evaluation Cases



Findings from 360 Cases Based on 99% Limits

In terms of the number of signals:

- The Poisson distribution tended to generate more signals due to tighter limits.
- The Bayesian's ZOIP tended to generate less signals due to higher limits.
- The Negative Binomial methods tended to represent a mid-point position with respect to signal generation.

In terms of distribution fitting convergence, the Poisson distribution tended to converge more frequently (as compared to other distributions). That makes sense, because parameter-wise this is a relatively simpler distribution to estimate.

Next Steps

Plan to assess the parameter estimation techniques and implementation algorithms to better understand difference in convergence and goodness fit.

Will develop comparison metrics to more effectively assess the benefits and drawbacks of the estimation methods.

Plan to use an alternative dataset and (if feasible) synthetic data to develop recommendations for the estimation methods and the assumed distributions.

References

S. Banik, B. M. Golam Kibria, "On Some Discrete Distributions and their Applications with Real Life Data “, Journal of Modern Applied Statistical Methods, November 2009, V8(2), pp. 423-447.

D. Erdman, L. Jackson and A. Sinko, "Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure“, SAS Global Forum 2008, conference proceeding paper 322-2008.

A. Kulesa, M. Krzywinski, P. Blainey and N. Altman “Sampling Distributions and the Bootstrap”, Nature Methods, June 2015, V12(6), pp. 477–478.

S. Beckett, J. Jee, T. Ncube, S. Pompilus, Q. Washington, A. Singh and N. Pal, "Zero-inflated Poisson (ZIP) Distribution: Parameter Estimation and Applications to Model Data from Natural Calamities,“ Involve,2014, V7 (6), pp. 751–767.

F.C. Xie, J.G. Lin, B.C. Wei, “Zero Inflated Generalized Poisson Regression Model: Estimation and Case Influence Diagnostics,” Journal of Applied Statistics, 2014, V 41(6), pp. 1383-1392.

V. Savani and A. Zhigljavsky, "Efficient estimation of parameters of the negative binomial distribution", Communications in Statistics - Theory and Methods, 2006, V35(5), pp.1–17.

S. Yang and G. Berdine, "The Southwest Respiratory and Critical Care Chronicles", 2015, V3(10), pp. 50-53.

Acknowledgements

The team wants to recognize the support of Asabere Owusu and Perceval Sondag during the development of this presentation.