

Article

# Likelihood Confidence Intervals When Only Ranges Are Available

Szilárd Nemes 

Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, 405 30 Gothenburg, Sweden; nemessz@gmail.com

Received: 22 December 2018; Accepted: 3 February 2019; Published: 6 February 2019



**Abstract:** Research papers represent an important and rich source of comparative data. The change is to extract the information of interest. Herein, we look at the possibilities to construct confidence intervals for sample averages when only ranges are available with maximum likelihood estimation with order statistics (MLEOS). Using Monte Carlo simulation, we looked at the confidence interval coverage characteristics for likelihood ratio and Wald-type approximate 95% confidence intervals. We saw indication that the likelihood ratio interval had better coverage and narrower intervals. For single parameter distributions, MLEOS is directly applicable. For location-scale distribution is recommended that the variance (or combination of it) to be estimated using standard formulas and used as a plug-in.

**Keywords:** range; likelihood; order statistics; coverage

## 1. Introduction

One of the tasks statisticians face is extracting and possibly inferring biologically/clinically relevant information from published papers. This aspect of applied statistics is well developed, and one can choose to form many easy to use and performant algorithms that aid problem solving. Often, these algorithms aim to aid statisticians/practitioners to extract variability of different measures or biomarkers that is needed for power calculation and research design [1,2].

While these algorithms are efficient and easy to use, they mostly are not probabilistic in nature, thus they do not offer means for statistical inference. Yet another field of applied statistics that aims to help practitioners in extracting relevant information when only partial data is available propose a probabilistic approach with order statistics. This approach has a long history and special focus was/is paid for samples with censored observations [3] or extremes [4]. Arnold and collaborators [5] offer a comprehensive overview of order statistics including how and when minimum and maximum of a sample may become sufficient statistic and some exact formulas and closed solution for the distribution of some order statistics.

The likelihood, the joint density of the observed data, is at the core of most statistical estimation and/or inference. Whereas formulation of the likelihood in most cases is based on complete samples there are situations when we observe only parts of the data. This could be due to censoring of different kind, or perhaps we do not have access at the full data but only at minimum and maximum values. Combination of the likelihood theory and order statistics is straightforward [6].

Herein, we aim to investigate the performance of likelihood-based confidence intervals when only minimum–maximum and sample size is available. Using Monte Carlo simulations, we examine coverage and interval width for location parameters estimated from ranges. Additionally, we compare estimation and inference from ranges with estimation and inference based on full samples. Lastly, we examine if the effect of sample size on estimation and inference on ranges.

In the following, we give a brief background for likelihood, order statistics and likelihood for order statistics. Then outline the Monte Carlo simulation. Thereafter we list the results of the simulation, an illustrative application and conclude with a brief general discussion.

## 2. Likelihood and Order Statistics

We assume that each of the  $n$  iid random observations in the sample  $Y_1, \dots, Y_n$  have probability mass function  $f(y; \theta)$ . In all cases the likelihood is the joint density of the observed data and

$$L(\theta|Y) = \prod_{i=1}^n f(Y_i; \theta). \tag{1}$$

We aim to estimate the parameter  $\hat{\theta}_{MLE}$  that makes  $Y_1, \dots, Y_n$  most probable, or most likely under the assumed probability mass function. Based on the likelihood function for the simple null-hypothesis of  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$  three test statistics can be formulated, the Wald ( $T_W$ ), score ( $T_s$ ), and likelihood ratio ( $T_{LR}$ ) statistics. These three test statistics are asymptotically equivalent and as  $n \rightarrow \infty$ , the test statistics converges in distribution to  $\chi_r^2$ . Confidence regions for the parameter(s) of interest are given by  $P\{T(\theta) \leq \chi_r^2(1 - \alpha)\}$ , which is a random set that contains the nonrandom true value with nominal probability of  $1 - \alpha$ . Of the three alternative test statistics the Wald- statistics and the Wald (or approximate) confidence intervals are the most common. The Wald-statistics is estimated as  $T_W = (\hat{\theta}_{MLE} - \theta_0)^T \{I_T(\hat{\theta}_{MLE})\}(\hat{\theta}_{MLE} - \theta_0)$ . The Wald-interval is not invariant to the parametrization of the estimate of the interest.

The likelihood ratio statistic does not depend on model parametrization and the not necessary symmetric around the point estimate. The likelihood ratio statistic is given by  $T_{LR} = \left\{ \frac{\sup_{\theta \in H_0} L(\theta|Y)}{\sup_{\theta \in \Theta} L(\theta|Y)} \right\}$ .

For the confidence intervals based on the Wald statistics (or Score statistics in most cases) there are closed form solutions. The intervals derived from  $T_{LR}$  require numerical estimation. This is done on log scale  $T_{LR} = -2\{l(\theta_0) - l(\hat{\theta}_{MLE})\}$ , where  $l$  is the log likelihood function

$$l(\theta|Y) = \sum_{i=1}^n \log f(Y_i; \theta). \tag{2}$$

If we do not know the full  $Y_1, \dots, Y_n$ , but we only know  $Y_{(1)} = \min\{Y_1, \dots, Y_n\}$  and  $Y_{(n)} = \max\{Y_1, \dots, Y_n\}$ , the likelihood function and the test statistics cannot be calculated as described above. For iid continuous variables the distribution of the range is

$$G(t) = n \int_{-\infty}^{\infty} f(y) \{F(y+t) - F(y)\}^{n-1} dy \tag{3}$$

where  $F(y)$  is the cdf and  $f(y)$  is the pdf of  $y$ . For simpler cases it is possible to obtain closed form solutions [7]. However, this formulation can be rather unpractical with difficult optimization.

Glen [8] offered a simpler solution, MLEOS, maximum likelihood estimation with order statistics defined as  $L_K(\theta|Y) = \prod_{i \in K} g_y(y_{(i)}, \theta)$ , where  $K$  is the set of order statistic indices (in or case 1 and  $n$ ) and  $g_{y_{(i)}}$  is the pdf for the order statistics. The pdf for the  $r^{\text{th}}$  order statistics is

$$g_r(y) = \frac{n!}{(r-1)!(n-r)!} [F(y)]^{n-r} [1 - F(y)]^{r-1} f(y). \tag{4}$$

We can obtain likelihood ratio statistics and confidence intervals by numerical optimization of MELOS. When full data is available, we can use closed form solutions for the Wald-statistics and approximate confidence intervals. It is possible to obtain exact formulas of the Information Matrix for MLEOS, however with considerable difficulty [9] and this is not always needed as optimization routines return the value of the Hessian matrix at  $\hat{\theta}_{MLE}$ .

### 3. Simulation Settings

We simulated a random variable  $y$  with sample sizes of 25, 50, 100, 500, and 1000. We assumed that the simulated  $y$  is iid following exponential and normal distribution. After simulating  $y_i$  with  $i = 1, \dots, n$  we used maximum likelihood estimation to obtain likelihood ratio and Wald-type approximate confidence intervals. Thereafter we extracted  $y_{(1)}$  and  $y_{(n)}$  and used MELOS to obtain likelihood ratio and Wald-type approximate confidence intervals. Thereafter, we repeated the procedure 1000 times. For each iteration, we noted if the confidence interval covers the true parameter value and the width of the confidence interval. Assuming binomial distribution for the confidence interval coverage, we expect that the coverage should be within 1.96 standard errors of the nominal coverage probability. As we run 1000 simulations, we expect that the coverage should be between tolerance limits of 0.936 and 0.963 [10].

For the normal distribution, we employed two types of estimation. First, we used MLEOS to estimate both mean and variance from the ranges. Second, we used the Wan-estimator to estimate the sample standard deviation from the ranges as

$$sd = \frac{y_{(n)} - y_{(1)}}{2\Phi^{-1}\left(\frac{n-0.375}{n+0.250}\right)} \quad (5)$$

where  $\Phi^{-1}(z)$  is the inverse function of standard normal distribution or equivalently, the upper  $z^{\text{th}}$  percentile of the standard normal distribution [2]. All analyses were conducted in R 3.5.1 [11].

## 4. Results

### 4.1. Exponential Distribution

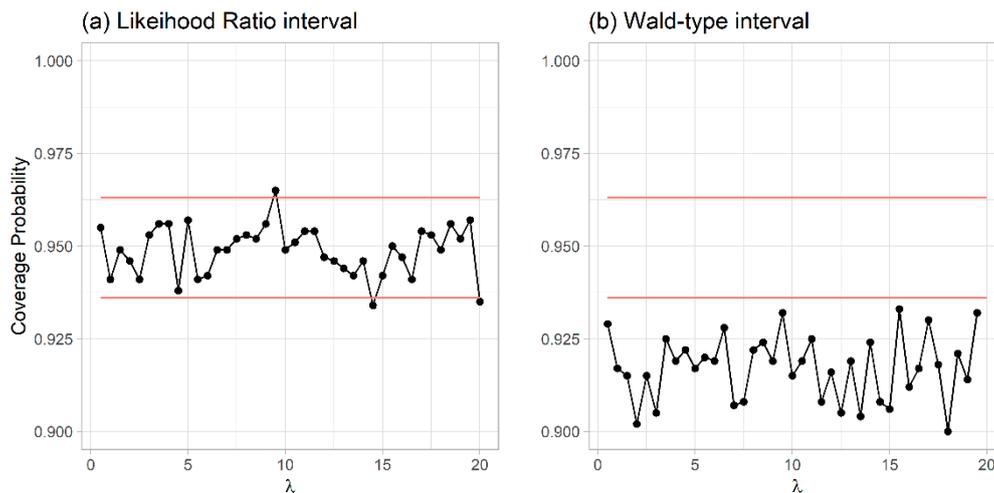
Table 1 presents the results for the exponential distribution with an intensity of 2. Likelihood ratio confidence intervals had coverage values within the tolerance limits Wald-type confidence intervals had coverage values under the lower tolerance limit of 0.936 when estimated either on full samples or ranges (Table 1).

**Table 1.** Confidence interval coverage and confidence interval width (in parenthesis) for likelihood ratio (LR) and Wald-type confidence intervals constructed on the full sample and on ranges.

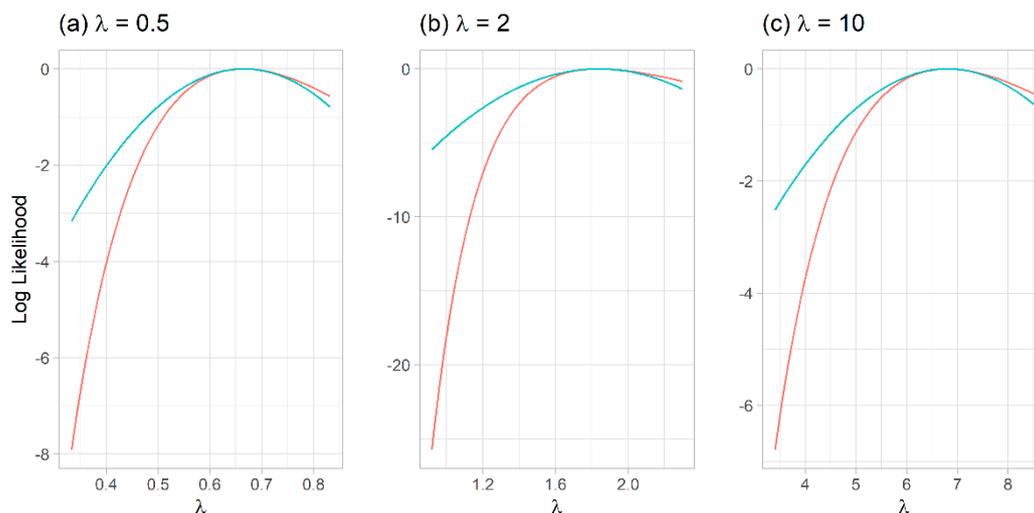
	N = 25	N = 50	N = 100	N = 500	N = 1000
<b>Full Data</b>					
LR-interval	0.948(1.62)	0.945 (1.13)	0.947 (0.79)	0.949 (0.35)	0.940 (0.24)
Wald-interval	0.845(1.15)	0.822 (0.798)	0.835 (0.56)	0.852 (0.24)	0.872 (0.17)
<b>Range data</b>					
LR-interval	0.945 (2.65)	0.942 (2.22)	0.957 (1.88)	0.945 (1.41)	0.946 (1.28)
Wald-interval	0.938 (2.42)	0.92 (2.02)	0.919 (1.174)	0.900 (1.27)	0.913 (1.15)

This pattern was consistent for different intensities (Figure 1).

Quadratic (Taylor-series) approximation of the order statistics likelihood showed poor normal approximation, explaining the subpar performance of the Wald-type intervals (Figure 2).



**Figure 1.** Coverage probabilities for the (a) likelihood ratio confidence interval and (b) Wald-type confidence interval for exponential distribution with different intensities ( $\lambda$ ). Vertical lines represent the tolerance limits for 1000 simulation and 95% nominal coverage probability.



**Figure 2.** Order statistics likelihood (blue line) and quadratic approximation (red line) for an exponential random variable with intensity (a) 0.5; (b) 0.2; and (c) 10.

#### 4.2. Normal Distribution

Table 2 summarizes the coverage probabilities for confidence intervals for the mean of a normally distributed variable with  $\mu = 0.5$  and  $\sigma = 2$ . Apart from  $n = 25$  the coverage of both likelihood ratio and Wald-type intervals had coverage values within the tolerance limits.

Concomitant estimation of both means and associated standard deviations leads to a substantial under-coverage of likelihood ratio intervals based on ranges. The Wald-type confidence interval had coverage values within the tolerance limits. Using the Wan-estimator (Equation (5)) as a plug-in kept the coverage properties of the Wald-type confidence interval. The coverage of the likelihood ratio interval improved and attained coverage value within the tolerance limits and close to the nominal coverage. Additionally, likelihood ratio intervals were slightly narrower than Wald-type intervals, suggesting improvement in statistical power.

**Table 2.** Confidence interval coverage and confidence interval width (in parenthesis) for likelihood ratio (LR) and Wald-type confidence intervals constructed on the full sample and on ranges.

	N = 25	N = 50	N = 100	N = 500	N = 1000
<b>SD estimated with Maximum Likelihood</b>					
<b>Full Data</b>					
LR-interval	0.926 (1.51)	0.940 (1.09)	0.956 (0.78)	0.949 (0.35)	0.942 (0.24)
Wald-interval	0.929 (1.55)	0.942 (1.10)	0.957 (0.78)	0.949 (0.35)	0.942 (0.24)
<b>Range data</b>					
LR-interval	0.835 (2.00)	0.839 (1.80)	0.831 (1.64)	0.829 (1.37)	0.837 (1.28)
Wald-interval	0.949 (2.95)	0.951 (2.55)	0.951 (2.32)	0.939 (1.94)	0.952 (1.81)
<b>Wan-type plug-in SD estimator</b>					
<b>Full Data</b>					
LR-interval	0.938 (1.52)	0.938 (1.09)	0.945 (0.77)	0.956 (0.35)	0.947(0.24)
Wald-interval	0.943 (1.55)	0.940 (1.10)	0.946 (0.78)	0.956 (0.35)	0.947 (0.24)
<b>Range data</b>					
LR-interval	0.943 (2.74)	0.947 (2.51)	0.956 (2.28)	0.944 (1.96)	0.945 (1.85)
Wald-interval	0.951 (2.80)	0.958 (2.58)	0.962 (2.36)	0.953 (2.07)	0.953 (1.96)

## 5. Illustrative Application

### 5.1. Exponential Distribution: Survival Data

In a recent abstract Okiror and collaborators [12] reported survival statistics for patients undergoing pulmonary metastasectomy for sarcoma. The authors report that for the 66 patients with metastatic sarcoma that they followed up the median disease-free interval was 25 months, ranging between 0 and 156. We could assume that the disease-free survival is exponentially distributed with intensity  $\lambda$ . Of interest is to estimate  $\lambda$  and construct a 95 % confidence interval around it. In this case we have access to median and range so there are several ways to proceed. The range is the least robust statistics, as they are maximally sensitive to outliers. Thus, using the median to estimate  $\lambda$ , is the better option. The median of an exponentially distributed variable is given by  $\lambda^{-1}\ln(2)$  which in our case gives an intensity estimate of 0.0277. Additionally, we could use the ranges. The range reported can be used in two ways. First, knowing that the variance of an exponentially distributed variable is  $\lambda^{-2}$  and using the Wan [2] equation for the standard deviation we can get an intensity estimate of 0.0301. Neither, approach offers straightforward way of inference, however Monte Carlo methods could be considered. Lastly, we could use MLEOS for point and interval estimation. This resulted in an intensity estimate of 0.0303 and associated 95 % confidence interval of 0.0192 to 0.0535. It is worth to note that this intensity underestimates the median, by two months (23 instead of 25). This can be due to multiple reasons. Partly the sensibility of ranges to outliers and equally importantly the possible deviation from the assumed distribution. However, a reparameterization to  $\ln(2)/\text{Median}$  and optimization with MLEOS gave a 95 % confidence interval for the median of 13 to 36 months. Thus, we cannot conclude without reasonable doubt that the two months' deviation is a genuine one.

### 5.2. (Log) Normal Distribution: Exhaled Nitric Oxide Test

Early phase clinical trials for new asthma medicines often take advantage of allergen challenges. In these challenges healthy volunteers are subjected to allergens that cause adverse airway responses and different biomarkers are measured and compared between the placebo and active arms. Research planning often extracts data from published articles either for setting reasonable target values or for variance estimation that needed for power calculations. Barchuk and collaborators [13] present such an allergen challenge study where among others they show data for  $F_{E}NO$ .  $F_{E}NO$  test (exhaled nitric oxide test) is a way to determine how much lung inflammation is present and how well inhaled steroids are suppressing this inflammation in allergic or eosinophilic asthma patients. Barchuk and

collaborators [13] present the geometric mean and range for FENO readings during a bronchial allergen challenge (BAC) (Table 3).

**Table 3.** MELOS estimates for the geometric mean (GM) and standard deviation for F<sub>E</sub>NO test data for in allergen challenge setting.

Period	GM	Range	GM <sub>MELOS</sub>	95 % ci	$\sigma_{Wan}$	$\sigma_{MLEOS}$	95 % ci
Pre BAC	44	12; 177	46	27; 77	0.73	0.72	0.50; 1.09
11 h post BAC	45	13; 156	45	27; 73	0.67	0.67	0.47; 1.01
24 h post BAC	65	17; 192	58	35; 92	0.65	0.65	0.45; 0.99
Predose	42	13; 151	44	27; 71	0.66	0.66	0.46; 1.00
24 h pre BAC	41	11; 178	44	25; 76	0.75	0.75	0.52; 1.14

We applied MLEOS in two different setting to this data. First, we used the range data to estimate standard deviations using the Wan-estimator and then assuming log-normally distributed data we estimated the geometric means and associated 95% confidence intervals. With one notable exception the estimation was acceptable. Second, using log-scale (normal distribution) we estimated standard deviations both with the Wan-estimator and MLEOS. F<sub>E</sub>NO is modeled on log scale, thus we modeled  $\sigma$ , the standard deviation for the normal distribution and not the standard deviation of the log-normal distribution. Here, MLEOS provided standard deviation estimates that deviated only on third or fourth decimal from the Wan-estimator. As the later estimator is validated it can be used as a golden standard. Thus, we concluded that MLEOS is a feasible and easy way to obtain standard deviation estimates from ranges. In addition to the Wan-estimator MLEOS provides future inference that can be extremely valuable for a research planning.

## 6. Discussion

In this note, we showed that it is possible to construct likelihood-based confidence intervals for means when the only available data is the minimum and maximum value of a sample. The range carries more information than the values of the two measurements, it also indicates that the rest of the values are within these values. This combined with an assumed distribution allowed construction of the confidence intervals. Of course, confidence intervals have meaning only if the parameter estimates are unbiased or if the bias it relatively low compared to the variance of the estimate. MLEOS had been proved to provide unbiased point estimates [8], and this was confirmed by or simulation (data not presented). Glen [8] focused to show the value of MLEOS for various censorship scenarios. Building on his work, we took a step forward and assessed the feasibility of MLEOS not only for point estimates but for inference. For the one parameter distribution, like the exponential distribution where the intensity characterizes both the expected value and variance estimation is straightforward. Order statistics likelihood estimation for the normal distribution assumes estimation of two model parameters the mean and variance. If full data are available, this estimation is straightforward; however, if only minimum and maximum values are available, then the number of parameters to be estimated matches the number of data points. Additionally, the maximum likelihood estimate of the standard deviation is biased. Here it is recommended to use plug-in estimator for the variance, such as the Wan-estimator [2]. Glen [8] observed that the censoring pattern greatly influences the recorded bias. Unfortunately, he did not considered ranges. We expected that the standard deviation estimated by MLEOS would be higher than for the Wan-estimator, however this was not the case. While both estimators explicitly consider the sample sizes, their aim is somewhat differ. The Wan-estimator aims to give an estimate the sample standard deviation, MLEOS aims to estimate parameters that makes toe observed ranges most likely.

Here, we assumed that we know the distribution of the observations. However, this is not always the case. Confidence interval construction for ranked set samples [14–16] including non-parametric interval received considerable attention. Confidence intervals for medians can be constructed based on adjacent order statistics with nonlinear interpolation [17,18]. Moreover, there are available routines for

likelihood estimation for miss-specified or partially miss-specified models [19]. Thus, it is of interest to assess in a future work the effect of model miss-specification and possible remedies.

**Funding:** This research received no external funding.

**Acknowledgments:** I would like to thank the reviewers to comments and improvement suggestions. This paper was written during a visit to Jakarta, Indonesia.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Hozo, S.P.; Djulbegovic, B.; Hozo, I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med. Res. Methodol.* **2005**, *5*, 13. [[CrossRef](#)] [[PubMed](#)]
2. Wan, X.; Wang, W.; Liu, J.; Tong, T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med. Res. Methodol.* **2014**, *14*, 135. [[CrossRef](#)] [[PubMed](#)]
3. Harter, H.L.; Moore, A.H. Maximum-Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and from Censored Samples. *Technometrics* **1965**, *7*, 639–643. [[CrossRef](#)]
4. Gnanadesikan, R.; Pinkham, R.S.; Hughes, L.P. Maximum Likelihood Estimation of the Parameters of the Beta Distribution from Smallest Order Statistics. *Technometrics* **1967**, *9*, 607–620. [[CrossRef](#)]
5. Arnold, B.C.; Balakrishnan, N.; Nagaraja, H.N. *A First Course in Order Statistics*; SIAM: Philadelphia, PA, USA, 1992; Volume 54.
6. Balakrishnan, N.; Cohen, A.C. *Order Statistics & Inference: Estimation Methods*; Elsevier: Amsterdam, The Netherlands, 2014.
7. Shahbaz, M.Q.; Ahsanullah, M.; Shahbaz, S.H.; Al-Zahrani, B.M. *Ordered Random Variables: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
8. Glen, A.G. Maximum likelihood estimation using probability density functions of order statistics. *Comput. Ind. Eng.* **2010**, *58*, 658–662. [[CrossRef](#)]
9. Park, S.; Kim, C.E. A Note on the Fisher Information in Exponential Distribution. *Commun. Stat. Theory Methods* **2006**, *35*, 13–19. [[CrossRef](#)]
10. Burton, A.; Altman, D.G.; Royston, P.; Holder, R.L. The design of simulation studies in medical statistics. *Stat. Med.* **2006**, *25*, 4279–4292. [[CrossRef](#)] [[PubMed](#)]
11. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
12. Okiror, L.; Peleki, A.; Moffat, D.; Bille, A.; Bishay, E.; Rajesh, P.; Steyn, R.; Naidu, B.; Grimer, R.; Kalkat, M. Survival following Pulmonary Metastasectomy for Sarcoma. *Thorac. Cardiovasc. Surg.* **2016**, *64*, 146–149. [[CrossRef](#)] [[PubMed](#)]
13. Barchuk, W.; Lambert, J.; Fuhr, R.; Jiang, J.Z.; Bertelsen, K.; Fourie, A.; Liu, X.; Silkoff, P.E.; Barnathan, E.S.; Thurmond, R. Effects of JNJ-40929837, a leukotriene A4 hydrolase inhibitor, in a bronchial allergen challenge model of asthma. *Pulm. Pharmacol. Ther.* **2014**, *29*, 15–23. [[CrossRef](#)] [[PubMed](#)]
14. Balakrishnan, N.; Li, T. Confidence intervals for quantiles and tolerance intervals based on ordered ranked set samples. *Ann. Inst. Stat. Math.* **2006**, *58*, 757–777. [[CrossRef](#)]
15. Chen, Z. Density estimation using ranked-set sampling data. *Environ. Ecol. Stat.* **1999**, *6*, 135–146. [[CrossRef](#)]
16. Frey, J. Distribution-free statistical intervals via ranked-set sampling. *Can. J. Stat.* **2007**, *35*, 585–596. [[CrossRef](#)]
17. Hettmansperger, T.P.; Sheather, S.J. Confidence intervals based on interpolated order statistics. *Stat. Probab. Lett.* **1986**, *4*, 75–79. [[CrossRef](#)]
18. Hutson, A.D. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *J. Appl. Stat.* **1999**, *26*, 343–353. [[CrossRef](#)]
19. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference: Theory and Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 120.

