STATCON
Statistik: Consulting • Training • Software

# Empirical Likelihood

## 1 Introduction

In industry production, tolerance intervals are widely used to determine the quality of a process.

The commonly used tolerance Intervals however assume normal distribution of the data, which is problematic for many processes. Nonparametric approaches on the other hand need a large sample size to secure sufficient coverage. This is especially problematic for processes, where evaluating products is costly and/or destructive.

Following a request of a client we came up with a possibility to calculate nonparametric tolerance intervals by calculating confidence Intervals for quantiles using the empirical likelihood approach implemented in JMP. As the desired sample sizes become very small, the traditional nonparametric confidence intervals tend to return unstable results.

In this paper we discuss the performance of the existing smoothed empirical likelihood approaches and compare different smoothing methods (Section 2). Additionally, we developed an extension to the existing methods in order to make the approach more stable for small sample sizes in combination with extreme quantiles. E.g. A 1% Quantile is already extreme in the situation of small samples (Section 3). Furthermore, the performance of the extension is evaluated in a simulation study (Section 4). Lastly, the Implementation of the method using JSL is discussed and demonstrated in an example.

## 2 Existing Methods

The subject of empirical likelihood was first introduced by Owen (1988), who also showed that the Wilks theorem is applicable to empirical likelihood, so that asymptotic Confidence intervals can be calculated. As the empirical likelihood is represented by a step function, depending on the empirical distribution function of the observed sample, several methods of smoothing have been proposed.

Chen and Hall (1993) proposed a method of smoothing using a kernel function. JMP uses this smoothing approach for its calculations. Furthermore Adimari (1998) proposed a linear smoothing of the empirical distribution function.

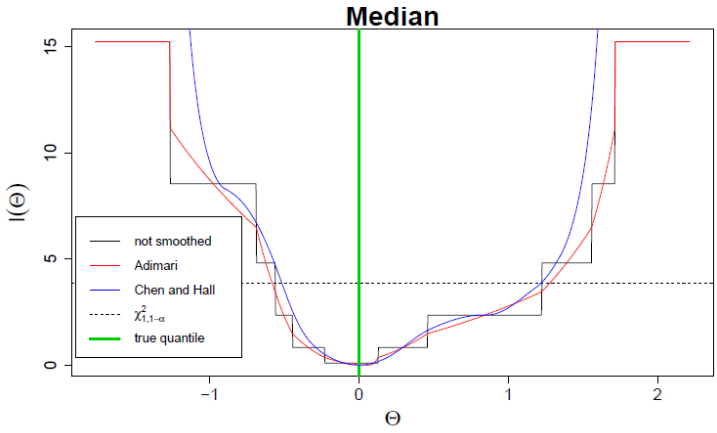


**Figure 1: Empirical likelihood functions for the median (n=11, α=0.05)**

Figure 1 shows the empirical likelihood functions for the Median of 11 observations from a standard normal distribution. It is visible, that all three approaches cover the true median in their respective confidence Interval, all in all, the confidence intervals do not differ substantially from another.

However, if we look at the 1% quantile of the same sample, the problems of the different methods become apparent.

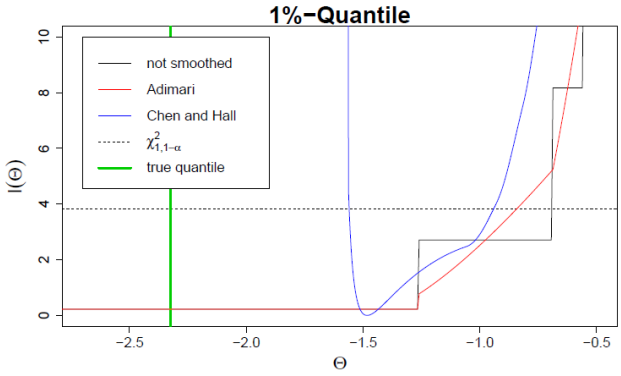


**Figure 2: Empirical Likelihood functions for the 1% - Quantile (n=11, α=0.05)**

Figure 2 shows the problems of empirical likelihood confidence intervals when it comes to extreme quantiles in combination with small sample sizes. The kernel approach goes off to infinity way to early, thus missing the true 1% quantile. The linear smoothing approach, as well as the original empirical likelihood function, however, remain constant for values smaller than the minimal observation. This results in infinite and thus unusable confidence intervals.

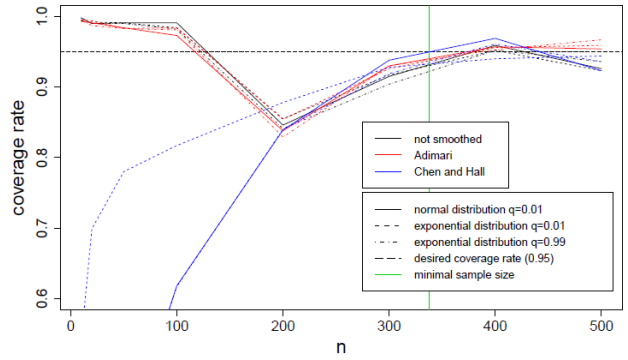


**Figure 3: Coverage rates based on 1000 samples for the different methods**

Figure 3 shows the problems observed in Figure 2 in a more general form. The kernel approach by Chen and Hall achieves coverage rates way below the desired coverage. The linear smoothing approach however returns infinite confidence Intervals for small sample sizes, which result in coverage rates above the desired value. Regardless of the approach, sample sizes of about n=330 observations are required to receive finite confidence intervals with sufficient coverage for a 1% quantile of a population. This required sample size has no relevant advantage compared to traditional nonparametric tolerance intervals.

## 3 Extension of the existing methods

Because of the problems regarding the existing methods, discussed in Section2, an extension was developed.

By extending the existing methods, the goal is to obtain finite confidence intervals which grant sufficient coverage rates, even for small sample sizes. To achieve this, the linear smoothing approach was chosen. The linear approach uses a smoothed variation of the empirical distribution function:

**Equation 1: Smoothed empirical distribution function (Adimari, 1998)**

$$F_n^*(\Theta) = \begin{cases} 0 & \text{if } \Theta < x_{(1)} \\ H(\Theta) & \text{if } \Theta \in [x_{(1)}, x_{(n)}) \\ 1 & \text{if } \Theta \geq x_{(n)} \end{cases}$$

Where

$$H(\Theta) = \begin{cases} \frac{2i-1}{2n} & \text{if } \Theta = x_{(i)}; i \in \{1,..,n-1\} \\ (1-\lambda)\frac{2i-1}{2n} + \lambda\frac{2i+1}{2n} & \text{if } \Theta \in (x_{(i)}, x_{(i+1)}); \lambda = \frac{\Theta - x_{(i)}}{x_{(i+1)} - x_{(i)}}; i \in \{1,..,n-1\} \end{cases}$$

We extended this function by keeping the linear smoothing even for values outside of the sample, so that every value between 0 and 1 is realized.

**Equation 2: Extended empirical distribution function**

$$F_{ext}(\Theta) = \begin{cases} 0 & \text{if } \Theta \le x_{(1)} - d_1 c \\ \frac{1}{2n} - \frac{1}{2n * d_1 * c}(x_{(1)} - \Theta) & \text{if } x_{(1)} - d_1 c < \Theta < x_{(1)} \\ H(\Theta) & \text{if } x_{(1)} \le \Theta \le x_{(n)} \\ \frac{2n-1}{2n} + \frac{1}{2n * d_2 * c}(\Theta - x_{(n)}) & \text{if } x_{(n)} < \Theta < x_{(n)} + d_2 c \\ 1 & \text{if } \Theta \ge x_{(n)} + d_2 c \end{cases}$$

Where $c \ge 1$; $d_1 = \frac{1}{10}\sum_{i=1}^{5}(x_{(i+1)} - x_{(i)})$ and $d_2 = \frac{1}{10}\sum_{i=1}^{5}(x_{(n-i+1)} - x_{(n-i)})$

Equation 2 shows the extended empirical distribution function, two linear segments were added, to realize all values between 0 and 1, the slope of the two linear segments is defined by the scaling parameters $d_1$ and $d_2$, which are meant to make the method independent from the scale of the measurements and the extension parameter c. Through a sensible setting of this extension parameter, the desired coverage can be achieved.

As the likelihood function resulting from this extension still has constant values for arguments where $F_{ext}(\theta)$ is 0 or 1 respectively, the likelihood function will be extended linearly for those values.

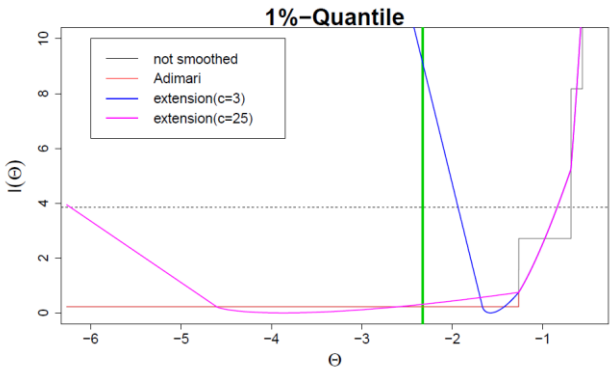


**Figure 4: Extended likelihood function for the sample in Figure 2**

Figure 4 shows how the extension parameter c influences the likelihood function and thus the width and coverage of the confidence Interval.

For valid confidence intervals, the right selection of a value for the extension parameter is essential. Here, the smallest possible value for c, which grants the desired coverage rates is wanted. To find those values for different situations a simulation study was carried out. For different combinations of quantile (q), sample size (n) and significance level (α), the coverage of the confidence intervals was evaluated using 5000 samples. The smallest natural number for c which grants the desired coverage rate was used to model a general selection of c based on the 3 mentioned parameters.

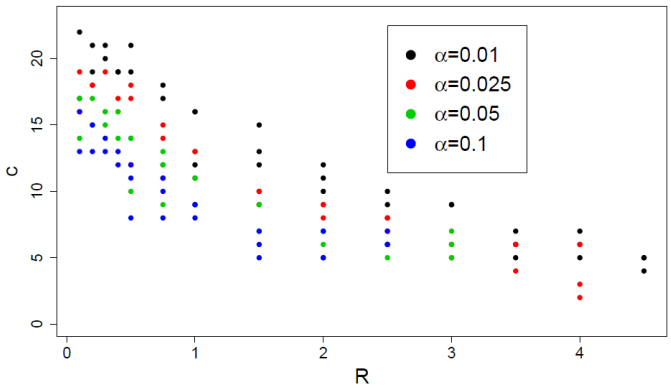


**Figure 5: Training data to model the required value of the extension parameter**

Figure 5 shows the simulated values for the extension parameter in relation to R, the product of quantile and sample size:

$$R := \begin{cases} q * n & \text{if } q \le 0.5 \\ (1-q) * n & \text{if } q > 0.5 \end{cases}$$

A linear model was chosen to predict the required value for c was chosen: $\hat{c} =$ 12.344 - 7.082$\sqrt{R}$ - 2.454log($\alpha$) − 75.125q - 0.004n The model achieves a relatively high goodness of Fit $(R^2_{adj.} = 0.933)$ reflecting the restriction to natural numbers in the training data.

## 4 Performance

With the modelling approach discussed in Section 3, a rule for the selection of the extension parameter c was developed. Using this approach, the performance of the extension is evaluated by calculating coverage rates based on 1000 samples for different situations.

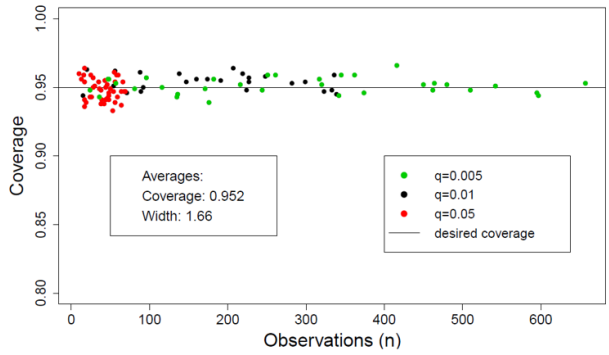


**Figure 6: Coverage rates based on 1000 samples from a normal distribution**

Figure 6 shows a drastic improvement compared to the existing methods when considering coverage rates. Even for samples as small as 10 observations, sufficient coverage is achieved, even for extreme quantiles. As the training dataset was derived from gaussian samples, the extension has a few problems when the data heavily differs from a normal distribution.

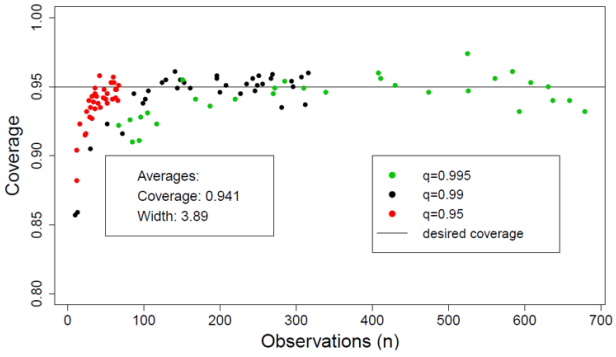


**Figure 7: Coverage rates based on 1000 samples from an exponential distribution**

For the high end of an exponential distribution, a very heavy tailed distribution, the desired coverage is not quite reached for small samples. However, the method grates good coverage for sample sizes of n>100 for the 99%-Quantile which is still substantially less than the existing methods.

On top of that, a semi parametric approach to the method is feasible, where a model similar to Section 3 is developed by using samples from an arbitrary, assumed distribution. The semi parametric approach was tested by developing a model for exponentially distributed data, which creates comparable results to the normal distribution (Figure 6).

All in all, the developed extension of the smoothed empirical likelihood approach is a clear improvement of the existing methods and offers the possibility of calculating valid confidence intervals, even when considering small sample sizes and extreme Quantiles.

## 5 Implementation

To put the developed method to practice, a JMP application was developed, which can calculate confidence Intervals for quantiles using the developed extension in combination with the Models for the parameter “c”.

To motivate the practical application of the method, consider the following example:

A sample of 100 observations of quality measurements from a production line is observed. The quality measurement has an upper specification limit of 85. The customer is willing to accept the batch of produced units if they’re assured, that at least 99% of the parts fulfill the quality requirement.

For this hypothetical sample, we compare different methods you could use for such a

quality assurance problem. In this case, the data is generated from a mixture of 2 normal distributions. This is quite often the case, the two normal distributions could be generated by different shifts, machines or suppliers of raw materials. The true 99%-quantile of the population lies at 79.6, so indeed, more than 99% of the population meet the specifications. However, in practice it is not known that more than one distribution is in place and the assumption of two mixed normal distributions is quite specific and very rarely used.

Nevertheless, we want to assure the customer that at least 99% of the products meet the specifications. Multiple approaches are feasible to generate such a statement.

The first approach would be a one-sided tolerance interval for 99% of the population with a significance level of $\alpha$=0.05. If normality is assumed, the upper limit calculated by JMP is 90.48, so that one would not be able to assure sufficient quality. The nonparametric tolerance interval with the same settings cannot be calculated due to lack of sample size.

The third approach would be to calculate a smoothed empirical likelihood confidence interval for the 99% -quantile of the population. For this approach JMP returns an upper limit of 78.58, this upper limit lies within the specification limits, so that one would feel confident to assure sufficient quality. However, the true 99%-quantile lies outside of the confidence interval due to the low coverage rates observed in Section 2. So, the implemented empirical likelihood method can not be trusted for this combination of quantile and sample size. On the other hand, the introduced extension of the empirical likelihood method returns an upper limit of 80.33, so both requirements are met: The confidence interval includes the true value and the interval is narrow enough to be confident in assuring quality for the customer.

This example represents a realistic application of the developed method and shows that the method outperforms the implemented alternatives in JMP under these circumstances.

## 6 Conclusion

In industry production, it is often of interest to give an upper limit for the fraction of low-quality products. When normal distribution cannot be assumed and only small samples are available, the existing methods perform very poorly. Therefore, we developed a method that extends the empirical likelihood method and can generate stable, nonparametric confidence intervals for quantiles at the tail end of the population even with small samples.

We showed that the developed method outperforms the existing methods when it comes to extreme quantiles in combination with small sample sizes. This way, using JMP in combination with the implemented method, we are able to assure quality of processes where measuring quality is very costly and/or time consuming.

## References

Adimari, G. (1998). An empirical likelihood statistic for quantiles. Journal
of Statistical Computation and Simulation, 60(1) pages 85-95.

Chen, S. X., Hall, P. (1993). Smoothed empirical likelihood con_dence
intervals for quantiles. The Annals of Statistics, pages 1166-1181.

Owen, A. B. (1988). Empirical Likelihood Ratio Con_dence IIntervals for a
Single Functional. Biometrika, Vol. 75, No. 2(Jun. 1988), pages
237-249.