PERCENTILE ESTIMATES RELATED TO EXPONENTIAL AND PARETO DISTRIBUTIONS

INTRODUCTION

The paper as posted to my website examined percentile statistics from a parent-offspring or Neyman-Scott spatial pattern. There are numerous applications of this pattern in meteorology, environmental science and epidemiology. [1,2] The paper identified that certain data errors in parent-offspring pairing can significantly increase the upper confidence limit (CL) and lead to overestimation of the P_{95} statistic. In a recent article, Ferrandino et al. refers to a Neyman-Scott pattern as a "mock epidemic" to differentiate the computer generated pattern from a real epidemic. [1] In their article, the mock epidemic was used to analyze search strategies for diseased plants.

This note expands on the "perfect information" case (Case 1) in the posted paper. For this case, a single parent responsible for infecting *n* offspring at distances distributed by the exponential distribution. The analytical solution for sampling distribution was identified in the paper. All other cases contain the potential for data error.

For the perfect information case with a given probability distribution and paramenters, the only variable affecting estimation of percentile values is sample size. From a sample as small as three numbers, theoretically we can generate every percentile value from a probablity of 1% to 99% (P_1 to P_{99}). The central question is not can we calculate percentiles with very limited data, but rather can we trust them. Or perhaps, the central question should be how much trust can be placed in these measures.

Percentile statistics are summary statistics, beneath which our raw data, including the flaws in record keeping, assumptions and computational methodology can hide. While there is no rigorous definition for "bad" in statistics, all signs will point in this direction when high end statistics (P_{90}+) are calculated with flaws in data and a lack of knowledge of the parent distribution.

Certainly, estimates of the center are more robust. It is out of necessity that the high range percentile estimates are computed. Applications such as encountered in disease control efforts, necessitate the calculation high end percentile estimates, in order to identify the longer traveled pathogens. Similar application occurs in environmental problems, where in order to control the spread of pollutants, the longer distances are the most relevant measures.

Environmental, mineral exploration and epidemiology share a common challenge of making statistical estimates in the high end of percentiles (P_{95}- P_{99}) where the data may be sparse with limited accuracy. Consider the spread of an oil spill. The far edges of the spill are the most important as may be related to the limit of potential environmental damage. Far removed from the foci of the spill, minute droplets of the most degraded oil in the spill must be used in the statistical analysis.

Theoretical results from identified stochastic pattern satisfies one element of technical validity- the results are easily reproduced. However, this does not ensure the inferences drawn from model statistics

will be correct to real world application. Stochastic analysis (forward analysis) and statistical analysis (inverse analysis) are both founded in the mathematics of probability. To use an analogy, the highway is proven theory. Forward analysis travels from a rigorous "given" to an outcome. Inverse analysis begins with outcomes, and travels the other direction towards the "givens."

EXTENSION OF PRIOR WORK

In this note, I have provided 1) Order statistics relationship to the beta distribution in a more generalized and rigorous manner, 2) Relationship of order statistics to percentile measures, and 3) Sampling distribution and simulation result for interpolated percentiles.

I have used the term percentile to be consistent with the paper, however all equations apply to quantiles. Further the Greek symbol ρ is used when defining probabily levels of confidence limits of statistics (See equations 8- 10, Table 1) and p for percentile levels. Confidence limits at a stated probability level, are denoted as $CL(\rho)$. Common high and low limits are CL(0.95) and CL(0.05), respectively.

SAMPLING DISTRIBUTION OF I.I.D. ORDER STATISTICS

For an independent and identically distributed (i.i.d) sample of size *n*, the probability density function (pdf) of the k[th] order statistics, ranked from lowest to highest with $1 \leq k \leq n$, is

$$g(x) = \frac{n!}{(k-1)!\,(n-k)!} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k} \tag{1}$$

where f(x) and F(x) are the pdf and cdf , respectively, of the parent distribution. [3]

Substituting $a = k$ and $b = n - k + 1$, into the above equation, results in

$$g(x) = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} f(x)[F(x)]^{a-1}[1-F(x)]^{b-1} \tag{2}$$

Since $\Gamma(n) = (n-1)!$ for every positive integer (proof given on page 296 of Reference 2), then

$$\frac{(a+b-1)!}{(b-1)!\,(a-1)!} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{3}$$

The beta function, $\beta(x)$, as a function of F(x) is

$$\beta(F(x)|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} F(x)^{a-1}[1-F(x)]^{b-1} \tag{4}$$

Function g(x) expressed in (2) is identical to the product of (4) and *f(x)* or

$$g(x) = \beta(F(x)|a,b) \cdot f(x) \tag{5}$$

Since an equivalent g(x) expression is

$$g(x) = \frac{dB(F(x))}{dx} \cdot \frac{dF(x)}{dx} \tag{6}$$

then

$$G(x) = B(F(x)|a,b). \tag{7}$$

Setting $G(x) = \rho,$ so $G^{-1}(\rho) = x$ and since $(B \circ F)^{-1} = F^{-1} \circ B^{-1}$, then

$$G^{-1}(\rho) = F^{-1}[B^{-1}(\rho|a,b)]. \tag{8}$$

Equation (8) is used to calculate confidence limits. For example , a high confidence limit, CL(0.95) of the k[th] ranked order statistic $CL(0.95)$ equals $G^{-1}(0.95)$. For the following distributions, the cumulative distribution and inverse (quantile function) are presented.

Table 1: Cumulative and Inverse Distributions

Distribution	$\boldsymbol{F(x)}$	$\boldsymbol{F^{-1}(\rho)}$
Exponential	$1 - \frac{1}{\theta} e^{-\frac{x}{\theta}}$	$-\theta\, ln(1-\rho)$
Pareto -I	$1 - \left(\frac{\beta}{x}\right)^{\alpha}$	$\beta(1-\rho)^{-1/\alpha}$

In all cases, $x > 0,$ and $0 < \rho < 1.$ All parameters of these distributions are positive values. The Pareto distribution is applicable for $x > \beta$. It is noted that the inverse expressions, as given above, would be the identical in the random number generator where $u(0,1)$, a uniform random deviate from 0 to 1, is set equal to *1 – p*.

Substituting the inverse functions results into equation 8 results in:

Exponential Distribution: $G^{-1}(\rho) = -\theta \ln(1 - B^{-1}(\rho|a,b))$ (9)

Pareto-I: $$G^{-1}(\rho) = \beta[1 - B^{-1}(\rho|a,b)]^{-1/\alpha} \quad (10)$$

The means of g(x) for the exponential and Pareto distributions are shown as equations 11 and 12, from reference 3.

Mean of Sampling Distribution for Exponential $$\mu_{n,k} = \theta \cdot \sum_{i=n+k+1}^{n} 1/i \quad (11)$$

Mean of Sampling Distribution for Pareto $$\mu_{n,k} = \beta \frac{n!}{(n-k)!} \frac{\Gamma(n-k+1-\frac{1}{\alpha})}{\Gamma(\mathrm{n}+1-\frac{1}{\alpha})} \quad (12)$$

RELATIONSHIP OF ORDER STATISTICS TO PERCENTILES

Various methods have been developed for estimating percentiles. Consider an order set of data, with x_1 as the lowest and x_n as the highest. We define V_p as the value of the p^{th} percentile, where p is in the range of $\{ 0 < p < 1\}$.

- METHOD 1 (NO INTERPOLATION)

No interpolation is used. The order statistic is the ceiing function of np. For example, if $np = 4.1$, then $V_p = x_5$. For $p = 0$ and $p = 1$, the percentile value is equal to the minimum and the maximum of the sample, respectively. Vp for small sets of data, will increase in a step-wise manner, and can not be inverted, as shown in figure 1.

$k = \lceil np \rceil$ $V_p = x_k$	**(13)**

- METHOD 2

(linear interpolation when k is greater than 1 or less than n)

For $p = 0$ and $p = 1$, the percentile value is equal to the minimum and the maximum of the sample, respectively. In all other percentile values are calculated based on a linear interpolation of the k^{th} order statistic as indicated by Method 1 and k+1 order statistic:

$$k = \lceil np \rceil$$ $$d = k - np + \frac{1}{2}$$ $$V_p = d \cdot x_k + (1 - d) \cdot x_{k+1}$$	(14)

- METHOD 3: (INVERTIBLE Vp AND LINEAR INTERPOLATION)

Methods 1 and 2 do not create a unique V_p for every value of p. So, while a functional relationship is developed, for p as the independent variable, there is no functional relationship for V_p as the independent variable. Method 3 will produce an invertible series. For $p = 0$ and $p = 1$, the percentile value is equal to the minimum and the maximum of the sample, respectively. In all other percentile values are calculated based on a linear interpolation of the k^{th} order statistic as indicated by Method 1 and k+1 order statistic:

$$k = \lceil (n-1)p \rceil$$ $$d = k - (n-1)p$$ $$V_p = d \cdot x_k + (1 - d) \cdot x_{k+1}$$	(15)

- EXAMPLE OF METHODS

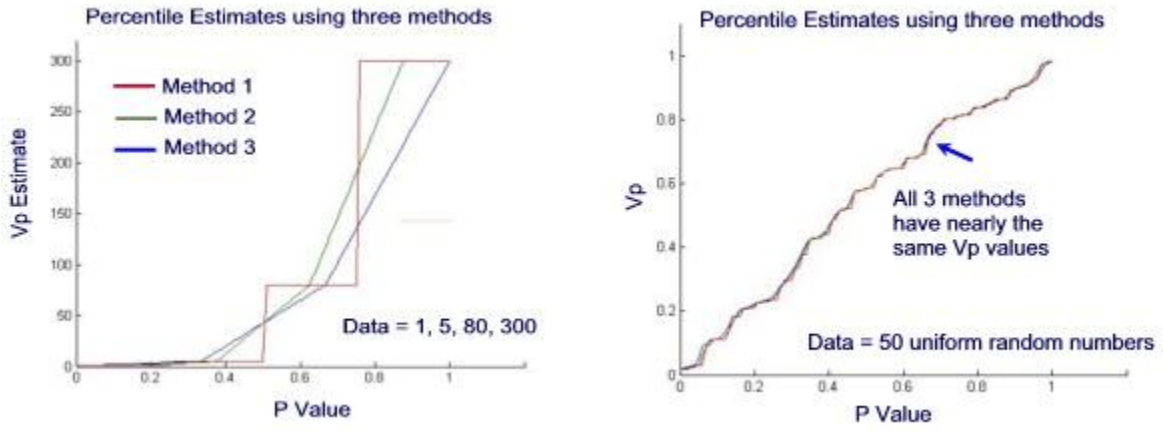
To demonstrate the differences among these methods, the p^{th} percentile is calculated below for a very small data sample x = [1, 5, 80, 300]. Generally, as the sample increases, the difference among these methods decreases. It is noted that Matlab program uses Method 2 and Microsoft Excel uses Method 3.

Vp Estimates

p	Method 1	Method 2	Method 3
0	1	1	1
0.50	5	42.5	42.5
0.51	80	45	44.75
0.75	80	190	135
0.95	300	300	267
1.00	300	300	300

Methods 2 and 3 have the identical median, consistent with generally accepted definition of the median as shown below. Differences are less with more data and less dispersion.

Figure 1: Comparison of Percentile Estimates using the three methods



EXAMPLES OF CONFIDENCE INTERVALS BASED ON SIMULATION

Exponential distribution cases were run with $\theta = 10$ which corresponds to $F^{-1}(0.95) = 30$ and $\sigma = 100$.

For method 1, the analytical solution agreed within 2 decimal places for the cases shown below when the number of simulation runs was one million. It is noted that due to vectorized coding, these runs took less than 5 seconds on a PC.

- EXPONENTIAL CASES

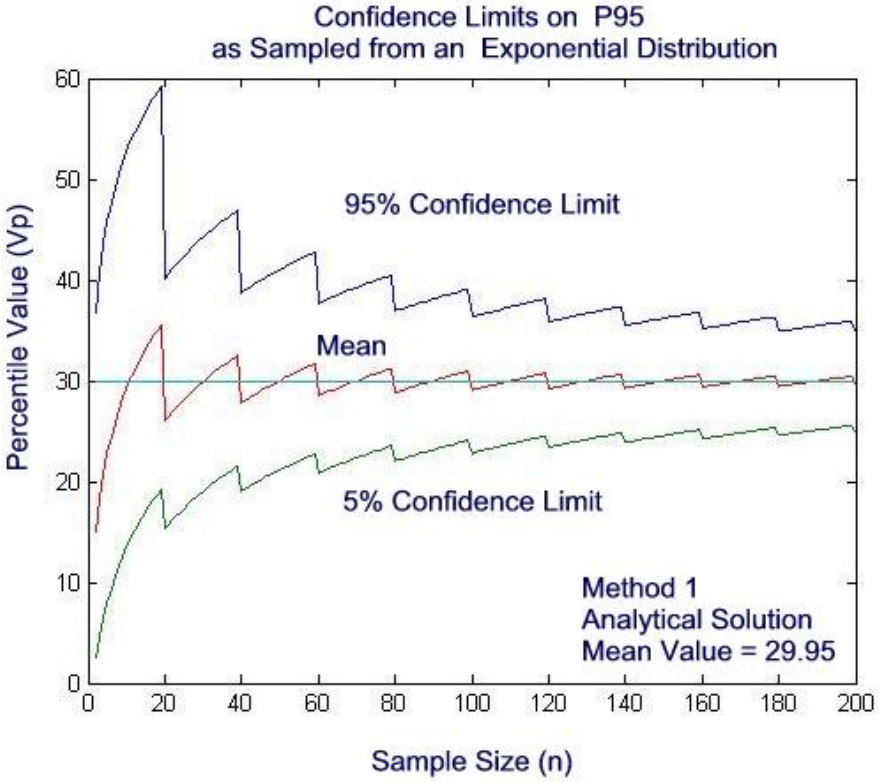
Table 2: P95 – Simulated P_{95} Mean Values (true value = 30)

Sample Size	Method 1	Method 2	Method 3
10	29.3	29.3	24.8
20	26.0	31.0	26.5
30	29.9	29.9	27.7

Lower (5%) and Upper (95%) Confidence Levels of P_{95}

Sample Size	Method 1	Method 2	Method 3
10	14 to 53	14 to 53	12 to 42
20	15 to 40	18 to 48	16 to 41
30	19 to 44	18 to 44	18 to 40

Using method 1 for calculating confidence intervals, there is a saw tooth pattern, due to the discrete nature of the ceiling function. These discrete changes occurs between sample sizes 19 and 20, 39 and 40, 59 and 60, etc.


Confidence Limits on P95
as Sampled from an Exponential Distribution
95% Confidence Limit
Mean
5% Confidence Limit
Method 1
Analytical Solution
Mean Value = 29.95
Percentile Value (Vp)
Sample Size (n)
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140
160
180
200

- PARETO I CASES

The Pareto I probability density distribution is given as:

$$f(x) = \frac{\alpha \beta^{\alpha}}{x^{\alpha+1}} \tag{16}$$

The distribution parameters used for the Pareto I cases were $\alpha = 2.9$ and $\beta = 10.7$. These parameters were chosen to match the exponential case values of $F^{-1}(0.95) = 30$ and $\sigma^2 = 100$.

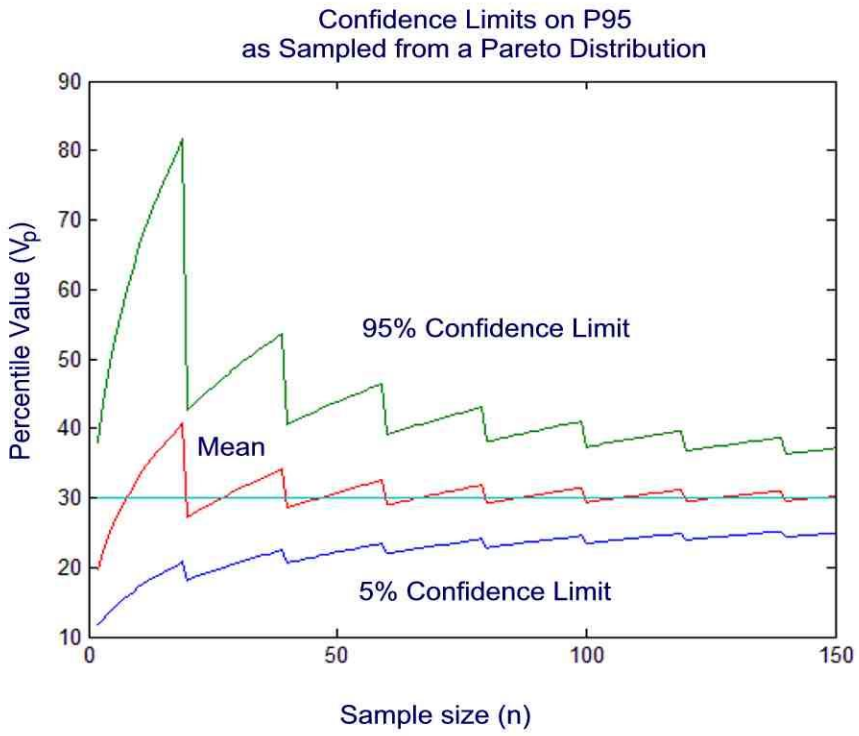
Table 3: P95 – Simulated P_{95} Values (average of one million statistics)

Sample Size, n	Method 1		Method 2		Method 3	
	Exponential	Pareto I	Exponential	Pareto I	Exponential	Pareto I
10	29.3	32.8	29.3	32.8	24.8	31.2
20	26.0	27.2	31.0	32.8	26.5	31.2
30	29.9	31.2	29.9	27.7	27.7	28.8

P95 – Lower (5%) and Upper (95%) Confidence Levels

Sample Size, n	**Method 1**		**Method 2**		**Method 3**	
	Exponential	Pareto I	Exponential	Pareto I	Exponential	Pareto I
10	14 to 53	17 to 66	14 to 53	17 to 66	12 to 42	16 to 49
20	15 to 40	18 to 44	18 to 48	20 to 60	16 to 41	19 to 44
30	19 to 44	21 to 49	18 to 44	21 to 49	18 to 40	20 to 43

The above tables show very comparable results. The graph of confidence limits provided on the following page, also shows similar confidence limits. Analytically calculated values agreed within two decimal points to the simulated results.



SUMMARY

1) Sampling distribution provide the uncertainty of statistics taken from limited sample. The derivation of sampling distributions and their inverse as they related to order statistics were shown in a general form.

2) The distribution and related confidence intervals of order statistics from samples taken from the exponential and Pareto distribution were identified. An analytical solution of the mean value of order statistics as published in the literature was presented.

3) Three common methods to calculate percentiles was presented. Method 1 (no interpolation) directly relates order statistics with percentiles, so the analytical distribution could be used to calculate mean and confidence limits to the sampling distribution. For Methods 2 and 3, Monte-Carlo simulation was used.

4) For Method 1, confidence intervals as calculated by the analytical solution and simulation were in close agreement for the exponential and Pareto cases.

ADDITIONAL WORK

The paper posted to my website explored the impact of data errors for a mock epidemic. This note extends the theoretical work in the most idealized case of the paper (Case 1: Perfect Information). As stated in the introduction, we have no rigorous means of separating "good" and "bad" statisitics. However, we can make imperfect assessments of the varying degrees of robustness or fragility of results.

Additional work is ongoing with the log normal distribution. I may explore the more realistic situation where a set of data is randomly drawn, and the distribution that seems to fit the best is used in the high-end estimates. This erroneous selection would introduce additional error into the estimation process.

The Pareto and exponential cases were expected to be similar, as the parent distributions had identical F(0.95) and variance values. Additional cases may be developed in the future, for parent distribution with similar means and variances and alternative means of calculated high end percentile values.

NOTATION

$\beta(x)$ Beta probability density distribution function (pdf)

B(x) Beta cumulative distribution function

$\Gamma(x)$ Gamma function

$\lceil\ \rceil$ Ceiling function, argument value is rounded to the next highest integer

REFERENCES:

1) Ferrandino, F.J., (2004), Measuring Spatial Aggregation in Binary Epidemics, Correlative Analysis and the Advantage of Fractal Based Sampling, Phytopathology, 94: 1215-1227.

2) Diggle, P.J., Spatial Analysis of Point Patterns, 2003, Oxford Press, New York.

3) David, H. A. and H.N. Nagarian, Order Statistics, 2003, John Wiley and Associates, New Jersey. Page 52, Exercises 3.2.1 and 3.2.3 presents equation for the moments of exponential and Pareto order statistics.

4) DeGroot, M.H. and Schervish, M.J., Probability and Statistics, 2002, Addison Wesley, New York.