Confidence Intervals for the Empirical CDF (ECDF)

Conversation: user's question and detailed explanation

Generated for local compilation

User's Original Question

You wrote:

> I am interested in gaining a better understanding of the statistical confidence intervals around an empirical cumulative distribution (ECDF). The ECDF of a sorted data sample
>
> $$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(i-1)} \leq x_{(i)} \leq x_{(i+1)} \leq \cdots \leq x_{(N)}$$
>
> is $F(x_{(i)}) = i/(N+1)$ according to the 1958 text by Emil Gumbel called *Extreme Value Statistics*. The variance of $F(x_{(i)})$ is
>
> $$V(F) = \frac{F(1-F)}{N+2}.$$
>
> So the variance tapers to zero at $F = 1$ and $F = 0$. The confidence interval would be
>
> $$F(x) \pm z_{\alpha/2}\sqrt{V(x)}.$$
>
> Three conundrums about this ...
>
> 1. The expression F(x) +/- z(alpha) * sqrt(V(x)) presumes that F(x) is normally distributed for any value of x, but F(x) must lie between 0 and 1, and so the confidence interval should be symmetric about F(x) only at $F = 1/2$. So the normal distribution does not seem appropriate for this case.
> 2. related to conundrum (1). The confidence $F(x) \pm z(\alpha)\sqrt{V}$ must be bounded within $[0, 1]$. But for large values of z, it can exceed this range. Note here thate the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality bounds can also exceed the $[0, 1]$ interval. This issue ties to the variance of the ECDF V(F) which certainly goes to zero as F goes to zero or one.
> 3. Why would the confidence interval on the ECDF not be larger at extreme values, where data is more sparse? At extreme values the Gumble bound is tighter and the DKW bound is uniform.
>
>   It seems that the standard normal variate is not approriate. It seems that an appropriate distribution would be bounded within the domain $[0 < F < 1]$ and is

symmetric at ($F = 1/2$) and has a two variables ... a mean (or median) F and a variance V.

Question: What is the appropriate distribution function for confidence intervals on the ECDF?

Detailed Answer

Short Answer

For finite-sample ECDF confidence intervals, the correct distributions are:

- **Binomial** for fixed x (you observe k of N samples $\leq x$): the ECDF value $F^N(x) = k/N$ has an exact Binomial distribution: $k \sim Binomial(N, F(x))$. Confidence intervals should be formed from the binomial distribution (exact Clopper-Pearson) or good approximations (Wilson, Jeffreys/Beta) These are automatically bounded in [0,1] and generally asymmetric near the edges.
- **Beta** for order statistics: $U_i \sim \text{Beta}(i, N + 1 - i)$

  The random variable $U_i = F(X(i))$ (the population CDF evaluated at the i-th order statistic) has the Beta distribution

  $$U_i \sim \text{Beta}(i, N + 1 - i),$$

  with mean $i/(N + 1)$ and variance $i(N + 1 - i)/((N + 1)^2(N + 2))$. Use Beta quantiles for exact, bounded, asymmetric intervals around the plotting positions.

  These produce bounded, asymmetric, exact confidence intervals.
- **DKW and normal approximations** are useful for quick, rough, or uniform (simultaneous) bands but they are conservative and/or inappropriate at the extremes. They do not contradict the binomial/Beta facts — they are simply different (and usually looser) guarantees.

Full Explanation

1. Why the normal approximation fails

The ECDF lies in $[0, 1]$, is discrete for finite N, and is skewed near 0 and 1. A normal CI can exceed the unit interval.

2. Correct distribution at fixed x

If k points fall $\leq x$, then $k \sim \text{Binomial}(N, F(x))$.

Exact Clopper–Pearson CI:

$$p_L = \text{Beta}^{-1}(\alpha/2; k, N - k + 1), \quad p_U = \text{Beta}^{-1}(1 - \alpha/2; k + 1, N - k).$$

3. Correct distribution for order statistics

$$U_i = F(X_{(i)}) \sim \text{Beta}(i, N+1-i),$$

with variances matching Gumbel's formula.

The CI is:

$$\left(\text{Beta}^{-1}\left(\frac{\alpha}{2}; i, N+1-i\right), \text{Beta}^{-1}\left(1-\frac{\alpha}{2}; i, N+1-i\right)\right).$$

4. Uniform (DKW) bands

DKW gives a distribution-free bound:

$$\sup_x |\hat{F}_N(x) - F(x)| \leq \sqrt{\frac{\ln(2/\alpha)}{2N}}.$$

Useful for full-range coverage but conservative.

Practical Recommendations

- Use Clopper–Pearson, Wilson, or Jeffreys intervals for fixed x
- Use Beta quantiles for plotting positions
- Use DKW only for uniform bands

Final Conclusion

The appropriate distributions for ECDF confidence intervals are:

Binomial (fixed x) and Beta (order statistics).

These provide proper bounded, asymmetric, exact intervals.