

Bands for distributions

This chapter considers confidence bands for a distribution function and some related functions. Chapter 5.8 describes bands for kernel density estimates. For $X \in \mathbb{R}$, the cumulative distribution function is

$$F(x) = F((-\infty, x]) = \Pr(X \leq x)$$

taken as a function of x .

A confidence band for $F(x)$ is a pair of functions $L(x)$ and $H(x)$ for which

$$\Pr(L(x) \leq F(x) \leq H(x), \forall x \in \mathbb{R}) = 1 - \alpha \quad (7.1)$$

under independent sampling of $X_i \sim F$. The randomness in (7.1) arises from the fact that L and U depend on X_1, \dots, X_n , although this is suppressed from the notation. Some exact confidence bands are available, others are asymptotic.

If the inequalities in (7.1) were imposed only at B points x , the result could be described as a B -dimensional hyper-rectangular confidence region. Bands are essentially infinite dimensional hyper-rectangles. As such, they do not necessarily correspond to tests with the greatest power. Ellipsoids or other shapes are often better. Bands have the advantage that they can be easily plotted.

Bands are also of interest for some related functions. The quantile function $Q(u)$ is defined through

$$Q(u) = F^{-1}(u) \equiv \inf\{x \mid u \leq F(x)\}, \quad 0 < u < 1. \quad (7.2)$$

The definition (7.2) makes Q unique even when $F(x) = F(x') = u$ for $x \neq x'$.

For independent real-valued data $X_1, \dots, X_n \sim F$ and $Y_1, \dots, Y_m \sim G$, the QQ plot is formed by plotting an estimate of $QQ(x) = G^{-1}(F(x))$. If the sample QQ plot lies far from the 45° line $QQ(x) = x$, then the distributions F and G differ.

For three or more samples from distributions F_1, \dots, F_k , we can select one of the distributions, say F_1 , as a baseline, and define a $k - 1$ dimensional quantile-quantile function by $(F_2^{-1}(F_1(x)), \dots, F_k^{-1}(F_1(x)))$, over x .

The survival function is $S(t) = F((t, \infty)) = 1 - F((-\infty, t])$. It is widely used in medical applications, as is the cumulative hazard function

$$\Lambda(t) = \int_0^t \frac{dF(u)}{F((-\infty, u))}.$$

These are discussed in Chapter 6.5.

7.1 The ECDF

The empirical CDF is the value $\hat{F}(x) = \#\{X_i \leq x\}/n$, taken as a function of x . The 95% Kolmogorov-Smirnov bands for F are of the form $\hat{F}(x) \pm D_n^{0.95}$, where $D_n^{1-\alpha}$ is defined in terms of the random variable

$$D_n \equiv \sup_{-\infty < x < \infty} \left| \hat{F}(x) - F(x) \right|, \quad (7.3)$$

by $\Pr(D_n \leq D_n^{1-\alpha}) = 1 - \alpha$.

Such bands can have exact coverage levels for finite n , because the distribution of D_n for $X_i \sim F$ is the same for any continuous distribution F . If F is not continuous, then Kolmogorov-Smirnov bands have greater than the nominal coverage level. To see why the distribution of D_n does not depend on F , write the order statistics of the sample as $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, and introduce random variables $U_i = F(X_i)$. The U_i are independent observations from the $U(0, 1)$ distribution, and have order statistics $U_{(i)} = F(X_{(i)})$. For continuous F the supremum in (7.3) occurs either immediately to the left or right of an observation $X_{(i)}$, so

$$\begin{aligned} D_n &= \max_{1 \leq i \leq n} \max \left(\left| \frac{i-1}{n} - F(X_{(i)}) \right|, \left| \frac{i}{n} - F(X_{(i)}) \right| \right) \\ &= \max_{1 \leq i \leq n} \max \left(\left| \frac{i-1}{n} - U_{(i)} \right|, \left| \frac{i}{n} - U_{(i)} \right| \right). \end{aligned}$$

For any continuous F , D_n can be expressed in terms of the order statistics of a uniform sample, and so $D_n^{1-\alpha}$ can be calculated for one distribution, such as $F = U(0, 1)$, and then applied to any continuous distribution. The hypothesis that X_i have CDF F is rejected at level α when F is not contained within the bands at all t .

Where the upper band goes above 1 it is replaced by 1, and similarly the lower band is replaced by 0 where it goes below 0. The Kolmogorov-Smirnov bands are widely used, but they are not particularly sensitive in the tails. To address this problem, weighted Kolmogorov-Smirnov bands, of the form

$$D_{n\psi} = \sup_{-\infty < x < \infty} \psi(F(x)) \left| \hat{F}(x) - F(x) \right|,$$

have been proposed. For example, the choice

$$\psi(z) = (z(1-z))^{-1/2} \quad (7.4)$$

weights each point x in inverse proportion to the standard deviation of $\hat{F}(x)$, and so puts more weight on the tail regions.

The random variable $n\hat{F}(x)$ has the binomial distribution with parameters n and $p = F(x)$. Kolmogorov-Smirnov bands are based on the most extreme discrepancy between the observed and expected binomial random variables. The weighted version with weights (7.4) takes account of the unequal variances of

those binomial random variables. Empirical likelihood bands may be constructed using the most extreme binomial likelihood at any x .

Empirical likelihood for $F(x)$ at a single point x was presented in Chapter 3.6. For $0 < p < 1$, and $-\infty < q < \infty$, define

$$\mathcal{R}(p, q) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=0}^{n+1} w_i Z_i(p, q) = 0, w_i \geq 0, \sum_{i=0}^{n+1} w_i = 1 \right\},$$

with $Z_i(p, q) = 1_{X_i \leq q} - p$, taking $X_0 = -\infty$ and $X_{n+1} = \infty$, so that $Z_0 = 1 - p$ and $Z_{n+1} = -p$. An asymptotic confidence interval for $F(x)$ is $\{p \mid -2 \log \mathcal{R}(p, x) \leq \chi_{(1)}^2\}$.

To get a confidence band for F , we consider the distribution of the most extreme pointwise likelihood, via

$$E_n = \sup_{-\infty < x < \infty} -\log \mathcal{R}(F(x), x).$$

Let $c_n^{1-\alpha}$ satisfy $\Pr(E_n \leq c_n^{1-\alpha}) = 1 - \alpha$. Then the band $(L(x), H(x))$ with

$$\begin{aligned} L(x) &= \min \{p \mid -\log \mathcal{R}(p, x) \leq c_n^{1-\alpha}\} \\ H(x) &= \max \{p \mid -\log \mathcal{R}(p, x) \leq c_n^{1-\alpha}\} \end{aligned}$$

is a $100(1 - \alpha)\%$ confidence band for $F(x)$. First we consider constructing L and H given $c_n^{1-\alpha}$, then we consider how to find $c_n^{1-\alpha}$.

7.2 Exact calibration of ECDF bands

It is computationally easy to obtain an exact calibration for empirical likelihood bands. The reason is that for any set of numbers a_1, \dots, a_n and b_1, \dots, b_n , there is a recursive algorithm to compute

$$\Pr(a_i \leq U_{(i)} \leq b_i, \quad i = 1, \dots, n).$$

See the discussion of Noé's recursion in Chapter 7.4. Noé's recursion also applies to weighted Kolmogorov-Smirnov confidence bands.

From equation (3.15) in Chapter 3.6,

$$-\frac{1}{n} \log \mathcal{R}(p, x) = \hat{p} \log(\hat{p}/p) + (1 - \hat{p}) \log((1 - \hat{p})/(1 - p)), \quad (7.5)$$

where $\hat{p} = \hat{p}(x) = \#\{X_i \leq x\}/n = F_n((-\infty, x])$, and $p = F(x)$. For fixed \hat{p} , equation (7.5) is a convex function of p with a minimum of 0 at $p = \hat{p}$. Thus $L(x)$ and $H(x)$ can be easily found by safeguarded searches, like those described in Chapter 2.9, starting in the intervals $(0, \hat{p})$ and $(\hat{p}, 1)$, respectively. Convexity in p of (7.5) implies that $-\log \mathcal{R}(p, x) \leq c_n^{1-\alpha}$ if and only if $L(x) \leq p \leq H(x)$. The bands $L(x)$ and $H(x)$ are piecewise constant functions, taking jumps at the n observed values $X_{(i)}$. Therefore, it is only necessary to compute them at $n + 1$ different points. Let L_i and H_i be the values of $L(x)$ and $H(x)$, respectively,

on the open interval $(X_{(i)}, X_{(i+1)})$, for $i = 0, \dots, n$, with $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$.

Having found either the L_i or the H_i , the other ones can be found by symmetry through

$$L_i = 1 - H_{n-i}.$$

Note that $L(X_{(i)}) = \min(L_{i-1}, L_i) = L_{i-1}$ and $H(X_{(i)}) = \max(H_{i-1}, H_i)$, for $1 \leq i \leq n$. Therefore, $H(x)$ is continuous from the right and $L(x)$ is continuous from the left.

To calibrate the curves we need to find $c_n^{1-\alpha}$. The extreme value of E_n must take place at or just to the left of an order statistic $X_{(i)}$. Thus

$$E_n = \max_{1 \leq i \leq n} \max \left(-\log \mathcal{R} \left(F(X_{(i)}), X_{(i)-} \right), -\log \mathcal{R} \left(F(X_{(i)}), X_{(i)} \right) \right).$$

Suppose that F is continuous. Then $\mathcal{R}(p, q)$ with $X_i \sim F$ is the same as $\mathcal{R}(p, F(q))$ on data $U_i = F(X_i)$. Thus we may write

$$\begin{aligned} E_n &= \max_{1 \leq i \leq n} \max \left(-\log \mathcal{R} \left(U_{(i)}, \frac{i}{n} - \right), -\log \mathcal{R} \left(U_{(i)}, \frac{i}{n} \right) \right) \\ &= \max_{1 \leq i \leq n} \max \left(-\log \mathcal{R} \left(U_{(i)}, \frac{i-1}{n} \right), -\log \mathcal{R} \left(U_{(i)}, \frac{i}{n} \right) \right). \end{aligned}$$

Now $E_n \leq c_n^{1-\alpha}$ is equivalent to

$$a_i \equiv L_{i-1} \leq U_{(i)} \leq H_{(i)} \equiv b_i, \quad i = 1, \dots, n.$$

It follows that Noé's algorithm can be employed to find the coverage probability for any value of $c_n^{1-\alpha}$. A one-dimensional numerical search can then be employed to find the value of $c_n^{1-\alpha}$.

Critical values $c_n^{1-\alpha}$ can be precomputed and tabulated. It may be more convenient to store them as a function of n . The function values in [Table 7.1](#) give very accurate coverage for the standard coverage levels 0.95 and 0.99, for sample sizes up to 1000.

7.3 Asymptotics of bands

The confidence bands of the previous section were constructed without employing any asymptotics. This was made possible by Noé's recursion. These bands have good power properties. Suppose that X_i have a continuous distribution F . Then the empirical likelihood confidence band of level $1 - \alpha$ has better asymptotic power for rejecting an alternative $\tilde{F} \neq F$ than a weighted Kolmogorov-Smirnov band of level $1 - \alpha$. This holds simultaneously for all weighted Kolmogorov-Smirnov bands and all alternatives $\tilde{F} \neq F$. Such universal optimality is surprising because \tilde{F} might only differ from F in a narrow interval, and a weighted Kolmogorov-Smirnov band might be constructed to be particularly sensitive to departures from F in just that one interval. See Chapter 7.4. The power consid-

Coverage 95% to 95.01%Sample size $n = 1$:

$$2.9957$$

Sample sizes $1 < n \leq 100$:

$$3.0123 + 0.4835 \log(n) - 0.00957 \log(n)^2 - 0.001488 \log(n)^3$$

Sample sizes $100 < n \leq 1000$:

$$3.0806 + 0.4894 \log(n) - 0.02086 \log(n)^2$$

Coverage 99% to 99.01%Sample size $n = 1$:

$$4.60517$$

Sample sizes $1 < n \leq 100$:

$$4.626 + 0.541 \log(n) - 0.0242 \log(n)^2$$

Sample sizes $100 < n \leq 1000$:

$$4.71 + 0.512 \log(n) - 0.219 \log(n)^2$$

Table 7.1 Shown are approximate critical values $c_n^{1-\alpha}$, for empirical likelihood confidence bands for the CDF from Owen (1995). The nominal coverage level is $1 - \alpha$, either 0.95 or 0.99. The actual coverage level is between the nominal level, and the nominal plus 0.0001. The sample sizes are from $n = 1$ to $n = 1000$.

ered is of large deviations type. Further large deviations results are described in Chapter 13.5.

The empirical likelihood confidence bands are based on the distribution of the most extreme of $2n$ binomial p -values, arising from an upper and a lower bound at each of n points. These p -values are strongly correlated with each other because they are based on the same data. It is interesting to compare the critical value of the likelihood used in setting bands with the finite degrees of freedom case. Figure 7.1 plots $c_n^{0.95}$ versus n for $1 \leq n \leq 1000$. The effective degrees of freedom corresponding to c_n are defined to be d such that $\Pr(\chi_{(d)}^2 \leq 2c_n) = 0.95$. The factor of 2 enters because in parametric settings the test statistic is minus twice a log likelihood where $c_n^{0.95}$ was developed for a negative log likelihood. Chisquareds on fractional degrees of freedom are Gamma distributions.

For $n = 1$, the effective degrees of freedom are $d = 2$. The effective degrees of freedom increase very slowly with n , to $d = 3$ at $n = 7$, to $d = 4$ at $n = 62$, and to $d = 5$ at some $n > 1000$. The effective degrees of freedom would be slightly different at a confidence level other than 0.95. The effective degrees of freedom are very nearly linear in c_n .

The case $n = 1$ is interesting. It involves just one quantile. As $n \rightarrow \infty$ for one quantile a $\chi_{(1)}^2$ limit is appropriate. The effect of $n = 1$ instead of $n = \infty$ is to change the degrees of freedom from 1 to 2.

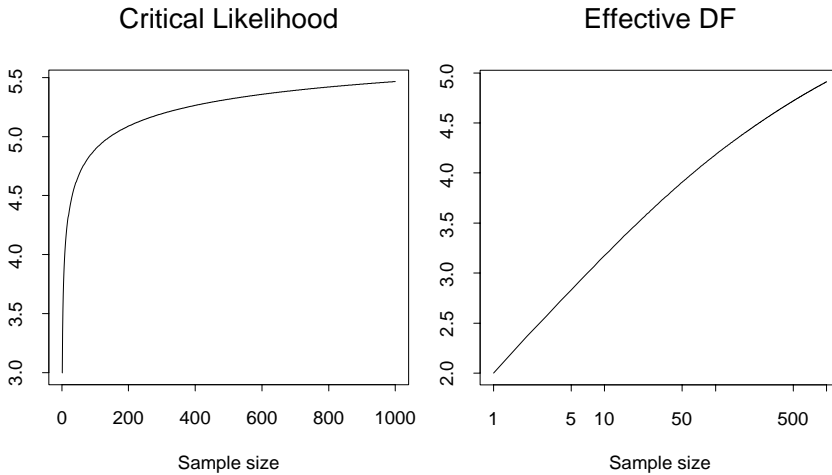


Figure 7.1 The left plot shows the critical likelihood threshold for exact 95% empirical likelihood confidence bands for the distribution function. The sample sizes range from 1 to n . A critical likelihood of c corresponds to an effective degrees of freedom of d where $\Pr(\chi_{(d)}^2 \leq 2c) = 0.95$. The right plot shows effective degrees of freedom versus sample size. The two quantities have nearly the same dependence on sample size. This is nearly linear on a log scale as shown in the right plot.

7.4 Bibliographic notes

Exact confidence bands for the CDF based on empirical likelihood were published by Owen (1995). Hollander, McKeague & Yang (1997) find asymptotic confidence bands for the survival function, $1 - F$, from right-censored data.

The weights (7.4) were proposed by Anderson & Darling (1952). The better known Anderson-Darling statistic is based on an integral over x , not an extreme as presented here. It corresponds to an infinite dimensional ellipsoidal region instead of an infinite dimensional hyper-rectangle.

The recursive algorithm for finding the probability that the ECDF from a $U(0, 1)$ sample stays within a given band is due to Noé (1972). It takes $O(n)$ space, and appears to be numerically stable for $n \leq 1000$. Noé's algorithm is given in Shorack & Wellner (1986). The fact that the bands described here give a test with better asymptotic power than any weighted Kolmogorov-Smirnov test at any alternative to $U(0, 1)$ was proved by Berk & Jones (1979) using the notion of relative optimality discussed in Berk & Jones (1978).

Qin & Lawless (1994) show that the error in estimating a distribution function is smaller if side information is used. Zhang (1996a) and Zhang (1999) describe confidence bands for the distribution function, given some side information expressed through estimating equations.

Switzer (1976) computes a confidence band for the QQ function by inverting Smirnov's two sample rank test. Confidence bands for the quantile function are given by Zhang (1997), by resampling from the NPMLE. Li, Hollander, McKeague & Yang (1996) present confidence bands for the quantile function from censored data. Einmahl & McKeague (1999) create empirical likelihood-based confidence tubes for QQ plot relating samples from two or more populations.