Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 6

11-1-2009

Least Error Sample Distribution Function

Vassili F. Pastushenko
Johannes Kepler University of Linz, Austria, vassili.pastushenko@jku.at

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 396-408

1538 – 9472/09/$95.00

Least Error Sample Distribution Function

Vassili F. Pastushenko
Johannes Kepler University of Linz, Austria

The empirical distribution function (ecdf) is unbiased in the usual sense, but shows certain order bias. Pyke suggested discrete ecdf using expectations of order statistics. Piecewise constant optimal ecdf saves 200%/N of sample size N. Results are compared with linear interpolation for U(0, 1), which require up to sixfold shorter samples at the same accuracy.

Key words: Unbiased, order statistics, approximation, optimal.

Introduction

Natural sciences search for regularities in the chaos of real world events at different levels of complexity. As a rule, the regularities become apparent after statistical analysis of noisy data. This defines the fundamental role of statistical science, which collects causally connected facts for subsequent quantitative analysis. There are two kinds of probabilistic interface between statistical analysis and empirical observations. In differential form this corresponds to histograms, and in integral form to the so-called sample distribution function or empirical distribution function (edf or ecdf in Matlab notation), c.f. Pugachev (1984), Feller (1971), Press, et al. (1992), Abramowitz & Stegun (1970), Cramér (1971), Gibbons & Chakraborti (2003). If histogram bins contain sufficiently big numbers of points, the usual concept of ecdf is more or less satisfactory. The focus of this paper is on short samples, where a histogram approach is not possible and an optimal integral approach is welcome. Consider i.i.d. sample X with N elements, numbered according to their appearance on the x-axis

$$X= [X_1, X_2,\ldots,X_N], \tag{1}$$

$$X_1 \le X_2 \le\ldots\le X_N. \tag{2}$$

Sorted X-values are sometimes denoted $X_{(n)}$, but here parentheses are omitted. Parent d.f. F(x) is connected with corresponding p.d.f. $f(x)$

$$F(x) = \int_{-\infty}^{x} f(x)dx \tag{3}$$

F(x) is defined for the whole range of possible sample values between extreme x-values X_0 and X_{N+1} (denoted similarly to X for formal convenience):

$$X_0 =\inf(x), X_{N+1}= \sup(x) \tag{4}$$

Due to the fact that f(x) ≥ 0, F(x) is non-decreasing. Therefore the exact random values of F(X), usually unknown in practical ecdf applications, are ordered according to positions of X elements at x-axis,

$$F_1 \le F_2 \le\ldots\le F_N. \tag{5}$$

where $F_1 = F(X_1)$, $F_2 = F(X_2)$, …, $F_N = F(X_N)$. For this reason values (5) are called order statistics, Gibbons & Chakraborti (2003). In literature ecdf is frequently denoted as $F_n(x)$ meaning that a sample consists of n elements. Here the notations are different. As defined in (5), $F_n = F(X_n)$, n = 1:N (colon is a convenient notation of MathWorks, meaning arithmetic progression between delimited expressions, here with an increment 1, more generally start : increment : finish). Usually ecdf is denoted $F_*(x, X)$, where x is the independent variable,

Email Vassili F. Pastushenko at vassili.pastushenko@jku.at.

sometimes called parameter, taking any value of the principally possible X-values

$$F_*(x, X) = \frac{1}{N}\sum_{n=1}^{N} H(x - X_n) \qquad (6)$$

H(t) is Heaviside unit step function, H = 1 for t≥0, otherwise H=0. Function (6) as a piecewise constant approximation of F(x) takes N+1 values (levels) equal to (0:N)/N in N+1 x-intervals between and outside of N sample values. F_* is continuous from the right, although Pugachev (1984) suggested that the continuity from the left would be more reasonable. A centrally symmetrical version could be a compromise (H = 0.5 at t = 0). Middle points between adjacent F_* levels are

$$m = (n-0.5)/N, \quad n = 1:N \qquad (7)$$

For convenience, an example of F_* is shown for N = 3, Figure 1 A. Expected F_n-values (E_n, c.f. next section), shown by circles, are different from m.

Eq. (6) is constructed as an arithmetic mean of N ecdf, each corresponding to a 1-point sample,

$$F_*(x, X) = \frac{1}{N}\sum_{n=1}^{N} F_*(x, X_n) \qquad (8)$$

This shows that E[F_*(x, X)] = F(x) for any N, where E[…] denotes the mathematical expectation, because this expectation corresponds to F_* for an infinitely long sample. In other words, for any fixed-in-advance x-value F_*(x, X) represents an unbiased estimation of F(x). The name empirical reflects the similarity between F_*(x, X), which gives the proportion of sample elements r satisfying r ≤ x, and F(x) = Prob(r≤x), r being a single random number. However, this similarity contains an arbitrary assumption. Indeed, differentiation of (8) with respect to x gives empirical p.d. f. $f_*(x, X)$, Feller (1971)

$$f_*(x) = \frac{1}{N}\sum_{n=1}^{N} \delta(x - X_n), \qquad (9)$$

δ(*x*) being the Dirac delta. As can be seen from this expression, ecdf (8) attributes probability measure 1/N to each sample element. As a result, any sample is represented as a set of measure 1, whereas in reality it represents a set of measure zero, which is obvious for the main class of continuous distributions, and discontinuous distributions can be considered as a limit of continuous ones. This contradiction is especially strongly expressed in eq.(6), where measure 1 is attributed to every single-point sample on the right-hand side, which should mean that every sample element is a deterministic, not a stochastic item.

As indicated by Pyke (1959), a more reasonable approach should consider a sample as a set of measure zero, which delimits N+1 nonzero-measure intervals on the x-axis. This is consistent with the point of view that the sampling procedure represents mapping of N random values of parent F(x) to the x-axis. A single random F-value is uniformly distributed in (0, 1), i.e., $F \in U(0, 1)$. Each of the F-values mapped into the sample values is selected independently. However, these values finally appear on the F-axis as an ordered sequence, so that the neighbouring elements of the sequence are no longer independent. Order statistics F_1, …, F_N have their own distributions. Therefore, an optimal ecdf must use this information. Probability densities for random u ∈ U(0,1), u = F_n, are c.f. Gibbons & Chakraborti (2003), Durbin (1973), Pyke (1959):

$$f_{N,n}(u) = u^{n-1}(1-u)^{N-n}\,\frac{N!}{(n-1)!\,(N-n)!}, \quad n = 1:N. \qquad (10)$$

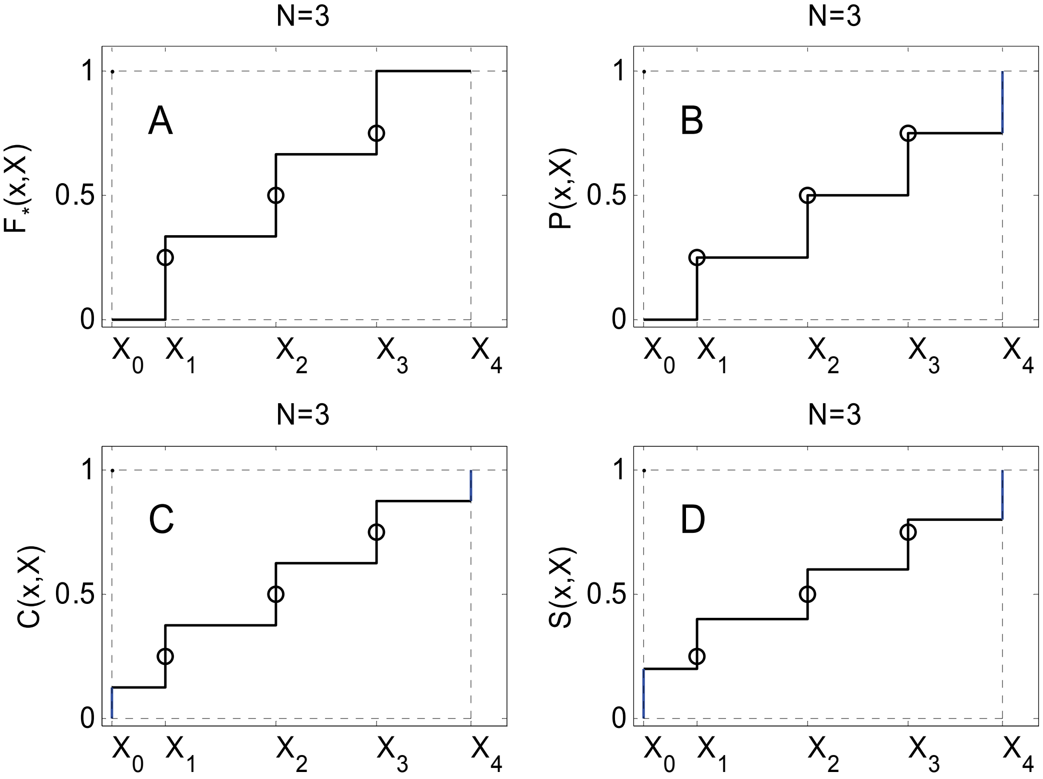
The first two moments of these distributions, or expected values and variances of F_n, denoted E_n and V_n respectively, are (c.f. Gibbons & Chakraborti (2003)):

$$E_n = E[F_n] = \int_0^1 x f_{N,n}(x)dx = \frac{n}{N+1}, \ n = 1:N \qquad (11)$$

and

$$V_n = \int_0^1 (x - E_n)^2 f_{N,n}(x)dx = \frac{n(N+1-n)}{(N+1)^2(N+2)} = \frac{E_n(1-E_n)}{N+2}, \ n = 1:N \qquad (12)$$

Figure 1: Different Ecdf-Versions for a Sample with 3 Elements
Expectations of order statistics are shown by circles. A: F*(x, X); B: P(x, X); C: C(x, X); D: S(x, X). Note that the numbers of jumps are different, A: N; B: N+1; C: N+2; D: N+2.



As a research tool, F* is expected to optimally reproduce parent F(x). However, there are some discrepancies with predictions of order statistics (11-12). It follows from (6) that E[F*(X_n, X)] = E[n/N] = n/N, n=1:N, whereas the correct expectation (11) is different , Pyke (1959). This discrepancy means a certain order bias.

Pyke (1959) considered order statistics F_n as zero-measure delimiters of probability intervals, created by the sample. He also considered statistic $C_N^+ = E_n - F_n$ instead of the usual statistic, $D_N^+ = n/N - F_n$, n = 1:N. This was interpreted by Brunk (1962), Durbin (1968) and Durbin and Knott (1972) as a discrete modification of ecdf. In particular, Brunk mentions Pyke's (1959) suggestion that the plotted points (F_n, n/(N+1)) in the Cartesian plane replace the empirical distribution function. In fact, as Hyndman and Fan (1996) mentioned, similar suggestions were made much earlier by Weibull (1939) and Gumbel (1939). These suggestions were partly considered for applications using ecdf values only at x = X, such as two-sided Kolmogorov-Smirnov test, Durbin (1968). However, any generalization for arbitrary x-values was not presented, although Hyndman and Fan (1996) discuss similar ideas concerning distribution quantiles.

To find an alternative for ecdf, an optimality criterion must be selected. Because the aim of ecdf is to approximate F(x), the main criterion is the approximation accuracy. Additionally convenience or simplicity may be discussed, but these aspects are almost the same within the class of piecewise constant approximations which are considered here.

The approximation accuracy needs a definition of distance between two compared distribution functions. The distance between two distributions, e.g. between exact F(x) and its empirical approximation, is frequently characterized by the biggest (supremum) absolute value of their difference. Another possible measure could be an average absolute value of the difference. A more commonly used statistical measure is mean squared deviation, calculated either as a sum in discrete approaches or as an integral in continual approaches. For discrete applications, which only need ecdf values for the sample elements, Pyke's approach already gives an optimal solution in the sense of minimal rms deviation. Indeed, if it is desirable to replace a random F_n by a single number with minimal rms error, this number is E_n.

However, some applications need an extension of Pyke's discrete function to the whole range of possible x-values. Interpolation based on known knots (X_n, E_n) is one option. Linear interpolation, which was probably meant by Brunk's suggestion to plot (X_n, E_n), may work well numerically in some cases, such as uniform parent distribution, however, it is badly suited for unlimited distributions and it is difficult to obtain general, distribution-independent results. This article focuses on nearest interpolation, for which two versions are possible depending on the choice of an independent variable. A more attractive version corresponds to independent F. In this way, interpolating from known knots (E_n, X_n) to arbitrary (C, x), ecdf C(x, X) (Figure 1C) with expected E_n (circles) in the centres of corresponding probability intervals may be obtained. Table 1 lists the various notations used throughout this article.

Methodology

A family of sample distribution functions. An ecdf-like function P(x, X) can be constructed using (X_n, E_n):

$$P(x,X)=\frac{1}{N+1}\sum_{n=1}^{N}H(x-X_n)\,. \qquad (13)$$

This function exactly corresponds to Pyke's suggestion at x = X. Nevertheless, P(x, X) is not very useful for arbitrary x-values, because it corresponds to a one-directional near interpolation, extending the optimal values only to the right. This is illustrated by Figure 1, B (E_n are shown by circles).

Vectors E = [E_1, …, E_n], and X = [X_1 , …, X_n], n=1:N, can be complimented by extremal values of F_n and x in order to enable interpolation in the entire range of x- and F-values. This leads to extended vectors **E** and **X**, each of size N+2:

$$\mathbf{E} = [0, E, 1] \qquad (14)$$

$$\mathbf{X}=[X_0, X, X_{N+1}] \qquad (15)$$

Two versions of the nearest interpolation are possible. In MathWorks syntax:

$$C= \text{interp}(\mathbf{X} , \mathbf{E}, x,\text{'nearest'}),\ X_0 \le x \ \le X_{N+1} \qquad (16)$$

and

$$x = \text{interp}(\mathbf{E}, \mathbf{X}, C,\text{'nearest'}),\ 0 \le C \le 1. \qquad (17)$$

Version (17) is more attractive for two reasons. First, **E** has known boundaries 0 and 1, whereas X_0 and/or X_{N+1} can be either unknown or infinite. Second, eq. (16) is less convenient for analysis because it involves middle points $m_{xn}=(\mathbf{X}_n+\mathbf{X}_{n-1})/2$, n=1:N+1, where any exact calculation of $E[F(m_{xn})]$ and $E[F(m_{xn})^2]$ for an unknown F(x) is not possible. As follows from (17),

$$C(x,X) = \quad \frac{1}{N+1}\sum_{n=0}^{N+1}w_n H(x-\mathbf{X}_n) \qquad (18)$$

Weight coefficients w_n are equal to 1 except for n = 0 and n = N + 1, where w_n = 0.5. Thus eq. (18) attributes probability measure of 0.5/(N+1) to x-values below X_1 and above X_N respectively, formally to extremal x-values X_0 and X_{N+1}, and measure of 1/(N+1) to every sample element. Altogether, measure of N/(N+1) < 1 is now attributed to the very sample. Incomplete measure does not lead to any difficulty, because sample estimations based on (18) should be considered as conditional ones, and therefore the result should be normalized by condition probability N/(N+1). Thus, estimation of expected value of some function t(x) results in a traditional answer, mean(t(X)):

$$\frac{N+1}{N}\int_{X_1-0}^{X_N+0} t(x)\frac{d\,C(x,X)}{dx}\,dx = \frac{1}{N}\sum_{n=1}^{N} t(X_n) \qquad (19)$$

Because extremal x-values acquire a probability measure, the first and last summands can be simplified in (18), which results in an equivalent of C in the entire x-range:

$$\mathrm{C(x,X)} = \frac{1}{N+1}\left(\frac{1}{2}+\sum_{n=1}^{N} H(x-X_n)\right),$$

$$X_0 < x < X_{N+1}. \qquad (20)$$

Table 1: Notations

[A, B, …]	Concatenation of A, B, …, a set consisting of A, B, …
C(x, X)	Centred ecdf, E-values are in the centres of corresponding probability intervals
d	Defect of levels, sum of their squared deviations from the optimal (natural) levels
D(α)	Total expected squared Deviation of s(x, X, α) from F(x)
D_*, D_C, D_S, …	Total expected squared deviations for F*(x, X), C(x, X), S(x, X), …
E = n/(N+1)	n = 1:N vector of expected order statistics.
E_n	n^{th} element of E
E = [0, E_1, …, E_N, 1]	Vector E extended by extremal E-values
E[<abc>]	Mathematical expectation of an expression <abc>
F(x)	Parent d.f.
f(x)	p.d.f., f = dF/dx
F_*(x, X)	Presently accepted ecdf
f_*(x)	Empirical p.d.f., f_* = dF_*(x)/dx
$f_{N,n}(u)$	p.d.f. of n-th order statistic u ∈ U(0,1), n = 1:N
g_z	Gain, relative total squared deviation (in units of total deviation for F_*), $g_z = D_z/D_*$, z = C,S,…
H(t)	Heaviside unit step, H=1 if t ≥ 0, otherwise H = 0. In Matlab: H = t >= 0
M = mean(X)	Average of sample elements
N	Sample size (length of i.i.d. sample)
P(x, X) NF_*(x, X)/(N+1)	Pyke function
s(x, X, α)	Family of ecdf with parameter α, 0 ≤ α < 0.5
s_n	Levels of s(x, X, α), n = 1:N+1
S(x, X)	Optimal member of s-family, minimizing D(α)
u	Uniform random variable, 0 ≤ u ≤ 1
U(0, 1)	Standard uniform distribution, F(u) = u, 0 ≤ u ≤ 1
X = [X_1, X_2, …, X_N]	i.i.d. sample with parent d.f. F(x)
X = [X_0, X_1, X_2, …, X_N, X_{N+1}]	Extended sample X by adding extremal x-values, size(**X**)= N+2
x	A number ∈ (set of possible X-values)
α, β	Parameters of ecdf family s(x, X, α, β)
δ(x)	Dirac delta
δ_{xX}	Kronecker symbol. In Matlab: δ_{xX} = any(x == X)
Δ	The deviation of an ecdf from the parent d.f.
Φ(x, X)	Hybrid of S and P for both continual and discrete applications

An example of C(x, X) is shown in Figure 1,C. Note that eq. (20) follows from eq. (13) by adding 0.5/(N+1), and C(x, X) has expected values E_n in the middle of the corresponding probability intervals. Therefore if the centrally symmetric unit step H(t) is accepted, C(X_n, X) automatically gives expected value E_n.

Functions P(x, X) and C(x, X) represent linear transformations of F_*, therefore F_*, P and C could be considered as members of two-parametric ecdf family *s*:

$$s(x,X,\alpha)=\alpha+\beta\,F_*(x,X) \qquad (21)$$

Thus, $\alpha = 0, \beta = 1$ leads to s = F_*(x, X); $\alpha = 0$, β = N/(N+1) gives s = P(x ,X) and α = 0.5/(N+1), β = N/(N+1) gives s = C(x, X). Levels of P(x, X) are not symmetrical with respect to probability centre 0.5, i.e. not invariant in transformation levels → 1-levels. Therefore, although P(x, X) has expected values at x = X, it cannot be considered as a real alternative to F_*. Excluding P(x, X) from the s-family, the number of parameters may be reduced by setting $\beta = 1-2\alpha$, which enables the automorphism levels → 1-levels. This leads to one-parametric s-family

$$s(x,X,\alpha)=\alpha+(1-2\alpha)F_*(x,X), \quad 0 \le \alpha < 0.5. \qquad (22)$$

Levels s_n of s(x, X, α) follow from the levels of F_*:

$$s_n=\alpha+(1-2\alpha)(n-1)/N \text{ , n=1:N+1,} \qquad (23)$$

where α = 0 corresponds to F_*(x, X), and α = 0.5/(N+1) to C(x, X). Consider the properties of s(x, X, α) in terms of order statistics and squared deviation of s(x, X, α) from F(x).

Mean Values of F(x) Between Adjacent X_n and Natural Levels

As noted above, the mapping of F(x) to sample X leads to certain order statistics predictions (11-12), therefore,

$$E[F_n^2]=E_n^2+V_n=\frac{(n+1)E_n}{N+2}\text{; n=1:N} \qquad (24)$$

In order to see how the levels s_n (23) agree with these predictions, different ecdf versions must be compared with F(x) within intervals ($\mathbf{X}_{n-1}$, $\mathbf{X}_n$) numbered by n=1:N+1. Consider the integrals:

$$I_{F,n}=\int_{X_{n-1}}^{X_n}\mathrm{F}(x)\mathrm{f}(x)dx=\int_{F_{n-1}}^{F_n}\mathrm{F}dF=\frac{F_n^2-F_{n-1}^2}{2}; \quad \text{n=1:N+1} \qquad (25)$$

and

$$I_{s,n}=\int_{X_{n-1}}^{X_n}\mathrm{s}(x,X,)\mathrm{f}(x)dx=s_n(F_n-F_{n-1}); \quad \text{n=1:N+1} \qquad (26)$$

Integrals (25-26) represent another kind of order statistics. Natural levels S_n can be found from a comparison of their mathematical expectations, that is, from E[$I_{s,n}$] = E[$I_{F,n}$], where

$$E[I_{F,n}]=\frac{E_n}{N+2}\text{; n=1:N+1} \qquad (27)$$

and

$$E[I_{s,n}]=\frac{s_n}{N+1}\text{; n= 1:N+1.} \qquad (28)$$

Equality of (27) and (28) leads to natural levels:

$$S_n=\frac{n}{N+2}\text{; n = 1:N+1.} \qquad (29)$$

The levels follow if the right hand sides of (25 and 26) are equated and divided by F_n-F_{n-1}. The mathematical expectations found lead to levels of C(x, X):

$$C_n=E[\frac{F_n^2-F_{n-1}^2}{2(F_n-F_{n-1})}]=E[\frac{F_n+F_{n-1}}{2}]=\frac{2n-1}{2(N+1)}; \quad \text{n = 1:N+1} \qquad (30)$$

Comparing the levels of F_*, given by (n-1)/N, and C_n (30) with natural levels S_n (29), n = 1:N+1, both are smaller than S_n below 0.5 and bigger than S_n above 0.5. If the ratio of differences between these levels is constructed and the natural ones, this ratio (for nonzero values) appears to be greater than 2 at any N (zeros happen at the median level 0.5 for even N):

$$\frac{(n-1)/N - n/(N+2)}{(n-0.5)/(N+1) - n/(N+2)} = 2 + \frac{2}{N} \quad (31)$$

Thus, the detailed comparison leads to a conclusion: both levels of F∗(x, X) and of C(x, X) show certain order bias, because in average these levels do not match the expected behaviour of integrals of F between order statistics F_n. They are insufficiently big below the sample median, and too big above it.

The defect d of s(x, X, α) is introduced as a sum of squared deviations of s_n from natural levels (29),

$$d = \sum_{1}^{N+1} (s_n - S_n)^2. \quad (32)$$

The defect of F∗ is

$$d_* = \sum_{n=1}^{N+1} \left(\frac{n-1}{N} - \frac{n}{N+2}\right)^2 = \frac{N+1}{3N(N+2)}, \quad (33)$$

and the defect of C is

$$d_C = \sum_{n=1}^{N+1} \left(\frac{n-0.5}{N+1} - \frac{n}{N+2}\right)^2 = \frac{N}{12(N+1)(N+2)}. \quad (34)$$

In agreement with eq. (31), the ratio of these defects is:

$$\frac{d_*}{d_C} = 4\left(1 + \frac{1}{N}\right)^2 \quad (35)$$

Two conclusions can be made. First, although near interpolation seems to be attractive in the sense that in puts expected values E_n exactly in the middle between C- levels, it is still not yet optimal S(x, X), based on natural levels (29):

$$S(x, X) = s(x, X, 1/(N+2)). \quad (36)$$

Thus, the optimum should occur at:

$$\alpha = \frac{1}{N+2}. \quad (37)$$

Ecdf S(x, X) formally ascribes to every element of the extended sample **X** probability measure of 1/(N+2):

$$\mathrm{S}(\mathrm{x}, \mathrm{X}) = \frac{1}{N+2} \sum_{n=0}^{N+1} H(x - \mathbf{X}_\mathrm{n}). \quad (38)$$

Ecdf S(x, X) has zero defect d by definition. Similar to C(x, X), the expression for S may be simplified as:

$$\mathrm{S}(\mathrm{x}, \mathrm{X}) = \frac{1}{N+2}\left(1 + \sum_{n=1}^{N} H(x - X_n)\right), \quad X_0 < x < X_{N+1}. \quad (39)$$

An illustration of S(x, X) for N = 3 is given by Figure 1, D.

Results

Function S(x, X) Minimizes the Expected Total Error of F(x) Approximation.

It can be shown that S(x, X) minimizes the error of F(x) approximation by calculating total squared deviation D of s(x, X, α) from F(x) and finding an optimal α as argmin(D(α)), getting in this way again α = 1/(N+2) as the optimal value. Total expected approximation error, or expected squared deviation is

$$D(\alpha) = E\left[\int_{X_0}^{X_{N+1}} (\mathrm{F}(x) - \mathrm{s}(\mathrm{x}, \mathrm{X}, \alpha))^2 \mathrm{f}(\mathrm{x}) dx\right] \quad (40)$$

The optimality of S is confirmed by following theorem and proof.

Theorem

S(x, X) represents least error approximation of F(x) at the family s(x, X, α), because it minimizes the total squared approximation error (40).

Proof

Consider deviation Δ,

$$\Delta = \mathrm{F}(x) - \mathrm{s}(x, X, \alpha) \quad (41)$$

as a random quantity at every fixed x due to randomness of X. Mathematical expectation of

Δ, taking into account eq. (22) and $E[F_*(x, X)] = F(x)$, is:

$$E[\Delta] = F(x) - (\alpha+(1-2\alpha)F(x)) = \alpha\,(2F(x)-1). \tag{42}$$

The goal is to find $D = E[\Delta^2]$, therefore the variance, var(Δ), is needed. This can be found using the variance of $F_*(x, X)$, expressed as $F(x)(1-F(x))/N$, Gibbons and Chakraborti (2003). Because in (41) F(x) is a deterministic function, only the second term in (41) contributes to var(Δ):

$$V_\Delta = \mathrm{var}(\Delta) = (1-2\alpha)^2 F(x)(1-F(x))/N. \tag{43}$$

Therefore, the expected squared deviation is:

$$\begin{aligned} E[\Delta^2 &= V_\Delta + E[\Delta]^2 \\ &= (1-2\alpha)^2 F(x)(1-F(x))/N + \alpha^2 (2F(x)-1)^2 \end{aligned} \tag{44}$$

Substituting (44) into (40) leads to total expected squared deviation D

$$\begin{aligned} D(\alpha) &= \int_{X_0}^{X_{N+1}} E[(\mathrm{F}(x)\text{-}\mathrm{s}(\mathrm{x,X},\alpha))^2]\mathrm{f(x)}dx \\ &= \int_{X_0}^{X_{N+1}} \begin{bmatrix} (1\text{-}2\alpha)^2 \dfrac{F(x)(1-F(x))}{N} \\ +\alpha^2(2F(x)-1)^2 \end{bmatrix} \mathrm{f(x)}dx \\ &= \int_0^1 \left[(1\text{-}2\alpha)^2 \frac{F(1-F)}{N} + \alpha^2(2F-1)^2 \right] dF \\ &= \frac{2(N+2)\alpha^2 - 4\alpha + 1}{6N}. \end{aligned} \tag{45}$$

Thus, $D(\alpha)$ is quadratic in α with minimum at α defined by (37), which proves the theorem.

Now consider expected squared deviations for three different α-values leading to F_*, C and S. For $\alpha = 0$ eq. (45) yields known result for F_*,

$$D_* = D(0) = \int_0^1 \frac{F(1-F)}{N} dF = \frac{1}{6N}. \tag{46}$$

For C(x, X), i.e. for α=0.5/(N+1):

$$D_C = D(\frac{0.5}{N+1}) = \frac{2N+1}{12(N+1)^2}, \tag{47}$$

and correspondingly, for S(x, X),

$$D_S = D(\frac{1}{N+2}) = \frac{1}{6(N+2)}. \tag{48}$$

Parabolic dependency of $D(\alpha)$, eq. (45) is illustrated in Figure 2 for several N-values. The values of D for three ecdf versions, F_*, C and S (46- 48), are indicated by special markers.

Linear Interpolation for Uniformly Distributed Data

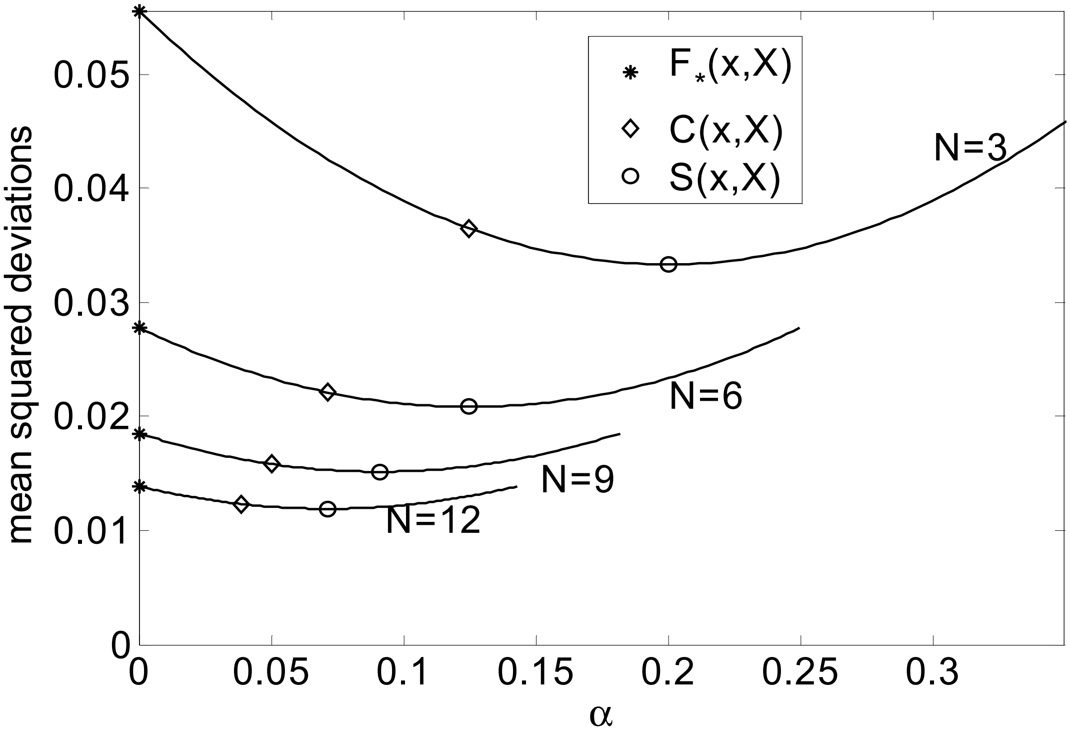
Compare the piecewise constant approximation in versions presented above with possibilities of different linear interpolations. In the case of a general parent d.f. F(x), it is difficult to get any analytical results. Therefore, F(x) is taken as standard uniform distribution, U(0, 1). However, this is more than a mere numerical example. Any known F(x) can be transformed to U(0, 1) by probability integral transformation u = F(x). Although in practice F(x) is mostly unknown, sometimes the transformation is possible, e.g. in fitting distribution parameters to X. Another meaningful aspect is - assuming that F(x) is known and transformed to standard uniform - the potentials of the linear interpolation become apparent.

Both versions of interpolation, eq. (16) and (17) are now considered linear instead of nearest. Let $E_{\mathrm{lin}}(x, X)$ be ecdf, defined as interpolation between Pyke points (X_n, E_n) according to (16)

$$\begin{aligned} &\mathrm{E_{lin}}(\mathrm{x,X}) = E_{n-1} + \frac{x - X_{n-1}}{X_n - X_{n-1}}(E_n - E_{n-1}); \\ &X_{n-1} \le x \le X_n, \\ &\qquad n = 1{:}N{+}1,\ \mathrm{X}_{(1:N)} \in \mathrm{U}(0, 1). \end{aligned} \tag{49}$$

Here X in the left hand side is usual sample, and *X* in the right hand side is the extended sample, $X_0 = E_0 = 0$, $X_{N+1} = E_{N+1} = 1$.

Figure 2: Total Squared Expected Error D(α) of the Family s(x, X, α) for Several N-values
The cases of F*(x, X), C(x, X) and S(x, X) as members of s-family are shown by special symbols; note that min(D(α))-values (circles) linearly depend on optimal α-values.



Eq. (49) is nonlinear with respect to random numbers X. Correspondingly, the expectation $E[E_{lin}(x, X)]$ deviates from x. Expected squared deviations $E[(E_{lin}(x, X)-x)^2]$ were estimated numerically as an average over 10^5 X-samples at N = 5. Figure 3 compares the result with $E(\Delta^2_*)$, $E(\Delta^2_C)$ and $E(\Delta^2_S)$. The left figure shows these expectations for all four compared versions, and the right figure shows their integrals in (0, x), which give at x = 1 corresponding total errors D. The gains, shown on the top of the right figure, represent the relative total errors, i.e. D_C/D_*, D_S/D_* and D_{lin}/D_* respectively.

The total approximation error is notably smaller for linear interpolation, as reflected by g_C (1.31), g_S (1.4) and g_{lin} (1.68). As illustrated in Figure 3 (left), the total squared error is smaller for E_{lin} than for C at any x, and it is smaller than that for F_* almost everywhere, with exception of narrow intervals near x = 0 and x = 1. In addition, E_{lin} loses to S around x = 0.5, but wins in wide intervals near x = 0 and x = 1.

More interesting results follow if linear interpolation is made according to eq. (17). Now the interpolation target is x, i.e. ecdf-values are selected as an independent variable *e*. In this case the implicitly defined ecdf *e*(x, X) is given by:

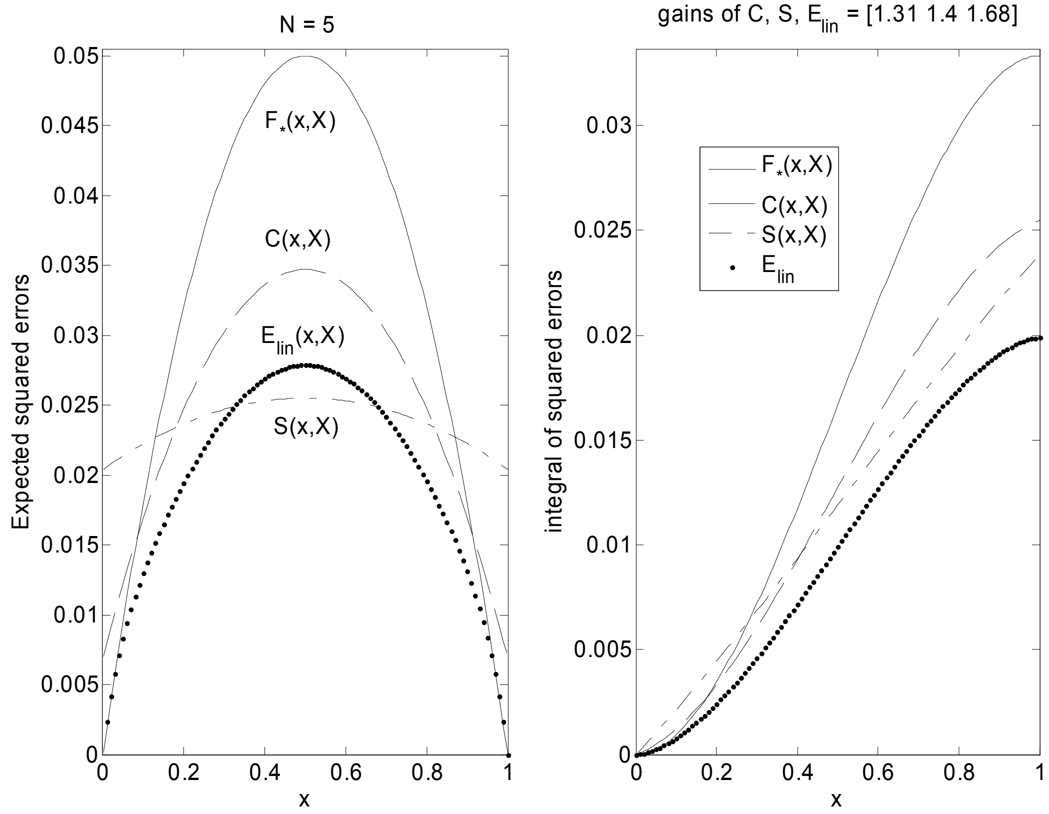
$$x(e, X) = \mathbf{X}_{n-1}(1-\lambda)+ \mathbf{X}_n\lambda;\ n = 1:N+1. \quad (50)$$

Here, λ is the interpolation variable,

$$\lambda=(e-\mathbf{E}_{n-1})(N+1),\ 0 \le \lambda \le 1\ (\mathbf{E}_{n-1} \le e \le \mathbf{E}_n). \quad (51)$$

Note that an equation similar to (50) was used by Hyndman & Fan (1996), eq. (1), who applied linear interpolation for calculating distribution quantiles. Due to the linearity of eq. (50) with respect to random X-values, this equation represents an unbiased empirical estimation of parent U(0, 1), that is, $E[x(e, X)] = e$, which

Figure 3: Expected Squared Errors of Different Versions of ecdf for Samples from U(0, 1), N=5
Left: $E[\Delta^2]$; right: integrals of left curves in (0, x), which define at x = 1 the total expected errors.



immediately follows from E[**X**] = **E**. This is interesting, because it shows that F*(x, X) is not the only possible unbiased estimation of F(x). The squared error of x_{lin} defined by (50) is:

$$E[\Delta_{lin}{}^2] = E[(x(e,\mathbf{X})-e)^2]$$
$$= E[((\mathbf{X}_{n-1} - \mathbf{E}_{n-1})(1-\lambda) + (\mathbf{X}_n - \mathbf{E}_n)\lambda)^2]$$
$$= \mathbf{V}_{n-1}(1-\lambda)^2 + \mathbf{V}_n\lambda^2 + 2c(n,n+1)\lambda(1-\lambda),\ n = 1{:}N+1. \quad (52)$$

Here c = cov(**X**), a covariance matrix of extended sample **X**, and **V**=[0 V 0] is the variance (12), extended by the values $V_0 = V_{N+1} = 0$. As can be seen from eq. (52), expected squared approximation error in every interval $\mathbf{E}_{n-1} \le e \le \mathbf{E}_n$ is given by parabola, connecting adjacent points ($\mathbf{E}_n$, $\mathbf{V}_n$). This is illustrated in Figure 4. The integral of (52) in (0, e) is now represented by piecewise cubic parabolas.

The gain of linear interpolation is now the same as in Figure 3, that is, the linear gain is invariant with respect to the interpolation target. The value of the linear gain for N > 1 is well approximated by g = 1 + 6/(2N-1), which means about 300%/N savings on sample size in comparison with F*. This raises the question about how such gain correlates with the quality of predictions based on linear interpolation.

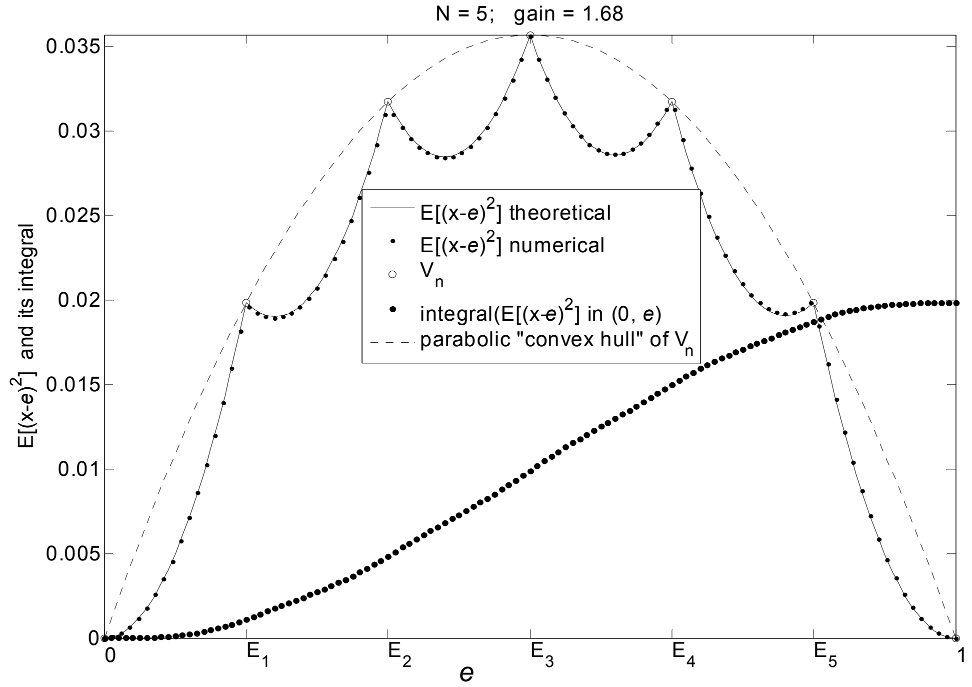
Eq. (8) can be directly applied to linear interpolation, which gives unbiased estimation and therefore eq. (8) should be valid. Given M = mean(X), eq. (8) suggests to represent x(*e*, X) as x(*e*, M(X)):

$$x = 2eM,\ e \le 0.5$$

and

$$x = 2(1-e)\,M + 2e-1,\ e > 0.5. \quad (53)$$

Figure 4: Expected Squared Error of the Linear Approximation (50), $E[(x-e)^2]$ and its Integral in (0, *e*)



Because E[M] = 0.5 (uniform data), (53) is indeed an unbiased estimation of x = *e*, the expected squared deviation of x from *e* is given by

$$\mathrm{var}(x) = 4e^2 V_M,\ e \le 0.5$$

and

$$\mathrm{var}(x) = 4(1-e)^2\ V_M,\ \ e > 0.5. \qquad (54)$$

Where

$$V_M = 1/(12N) \qquad (55)$$

V_M is the variance of M. Integrating (54) over *e* in (0, 1), the total mean squared deviation D_M is obtained as

$$D_M = 1/(36N). \qquad (56)$$

This result seems to be extraordinary, because it means a gain of ecdf (53) equal to 6, that is, 6 times shorter samples in comparison with F_* at the same approximation error, and this happens at any N value! Is it possible to get some practical advantages out of such a precise approximation?

One such possibility is suggested by distribution parameter fitting. Thus, unknown parameter(s) q can be found as argmin((mean(F(X, X, q))-0.5)^2). This method works indeed, and it should be compared with others. However, fitting parameters is a special topic, which should be discussed separately.

Optimal ecdf S is constructed to minimize expected total error for continual applications. Discrete applications only need ecdf values at x = X, and then P(x, X) should be used. How is it possible to combine P(x, X) and S(x, X) into a universal ecdf, valid both for continual and discrete applications? This can be done by redefining S at x = X, e.g. by introducing function $\Phi(x, X) = S(x)$, if $x \neq X_n$, otherwise $\Phi(X_n, X) = E_n$, n = 1:N. Such switching between P(X, X) and S(x, X) can be expressed as a formal mixture of both functions, using Kronecker symbol δ_{xX}:

$$\Phi(x,X) = \delta_{xX}\, P(x,X) + (1-\delta_{xX})\, S(x,X),\ \delta_{xX} = 1,$$
$$\text{if any}(x{=}{=}X),\ \text{otherwise}\ \delta_{xX} = 0. \qquad (57)$$

Function $\Phi(x, X)$ is discontinuous at x = X both from left and right, which is physically and functionally more reasonable, than in the case of $F_*(x, X)$, continuous from the right only.

Conclusion

The least error piecewise constant approximation of F(x) was presented. The starting point was that ecdf ascribes total probability of 1 to the sample, whereas any finite sample represents a set of measure zero. An optimal approach should ascribe zero measure associated with the sample. However, due to its convenience, a piecewise constant formalism has been selected. As a result, a part of total probability, equal to N/(N+2), is still associated with the sample. However, the aim was roughly achieved, because this measure is now smaller than 1, and this enabled a higher accuracy.

Optimal ecdf S(x, X) was built as a result of eliminating order bias of levels in ecdf $F_*(x, X)$, which is an unbiased estimation of F(x) for any fixed-in-advance x-value. Are ecdf versions C and S also unbiased? If it is forgotten for a moment that C and S are not designed for straightforward averaging over different samples, E[s(x, X, α)] could be calculated. As follows from (22), the s-family is biased at $\alpha > 0$, i.e. C and S are biased. This bias asymptotically disappears as $N \rightarrow \infty$. Is this bias important or not? What is more important for practical applications, improved accuracy of F(x) approximation, or formal bias which is in fact artificially created?

This bias has no practical meaning. Versions C and S use all available sample elements by definition, and the way this is done is not reducible to simple averaging. In fact, the bias is created by violation of the procedures behind C and S. The correct comparison is not reduced to an averaging over several samples. Instead, all available samples should be fused into one long sample before C or S functions are found. As eq. (8) shows, in the case of F_* the averaging over many samples gives the same result, as one combined sample. This enables formal ubiasedness, but the consequence thereof is increased approximation error.

A correct comparison of D_{F*}, D_C and D_S should always be done using the same sample or set of samples. If $N \rightarrow \infty$, then F_*, C and S all converge to the same F(x). The only difference is that D_S is the smallest of the three at any N. For this reason, if N is not very large, S(x, X) should always be preferred in practice as the best piece-wise constant approximation of F(x).

The smallest possible error of empirical estimation of F(x) is desirable, regardless of whether the error is due to the variance or due to inexact mathematical expectation. An optimal method should minimize the total error, and exactly this is done by Φ(x, X) both for discrete and continual applications. Physically, S has a smaller approximation error, because it takes into account additional information, contained in the order statistics F(X), whereas F_* neglects this information. As a result, ecdf F_* has order bias and an unnecessarily big approximation error.

The optimal ecdf Φ(x, X), presented here, is based on the most popular optimality criterion in statistics, i.e. least squared deviation. Final decision about its superiority depends on the quality of statistical predictions produced by different ecdf versions.

Acknowledgements

It is a pleasure to acknowledge Dr. L. Bogachev from the University of Leeds, UK, for his quite detailed discussions, Dr. J. Hilbe for several useful remarks, Dr. T. Lane from MathWorks, Inc. for discussions, Cleve Moler and the MathWorks team for the efficient software used in numerical simulations and Q. Beatty for improving my English.

References

Abramowitz, M., & Stegun, I. A. (1962). *Handbook of mathematical functions*. NY: Dover.

Brunk, H. D. (1962). On the range of the differences between hypothetical distribution function and Pyke's modified empirical distribution function. *The Annals of Mathematical Statistics*, 33(2), 525-532.

Cramér, H. (1971). *Mathematical methods of statistics*, (*12th Ed.*). Princeton, NJ: Princeton University Press.

Durbin, J. (1968). The probability that the sample distribution function lies between two parallel straight lines. *The Annals of Mathematical Statistics*, *39*(2), 398-411.

Durbin, J. (1973). Distribution theory for tests based on the sample distribution theory. *Regional Conference Series in Applied Mathematics*, *9*. UK: London School of Economics and Political Science, University of London.

Durbin, J., & Knott, M. (1972). Components of Cramér-von Mises statistics I. *Journal of the Royal Statistical Society*, *B*(*34*), 290-307.

Feller, W. (1971). *An introduction to probability theory and its applications*, (*2nd Ed.*). NY: John Wiley and Sons, Inc.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference*, (*3rd Ed.*). NY: M.Dekker.

Gumbel, E. J. (1939). La probabilité des Hypothèses. *Comptus Rendus de l'Academie des Sciences*, *209*. 645-647.

Hyndman, R. J., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*(4), 361-365.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C*, (*2nd Ed.*). Cambridge, MA: Cambridge University Press.

Pugachev, V. S. (1984). *Probability theory and mathematical statistics for engineers*. Elsevier Science

Ltd. In Russian: В.С. Пугачев. *Теория вероятностей и математическая статистика*.

Москва, Наука, Главная редакция физико-математической литературы, 1979, стр. 303.

Pyke, R. (1959). The supremum and infimum of the Poisson process. *Annals of Mathematical Statistics*, *30*, 568-576.

Weibull, W. (1939). The phenomenon of rupture in solids. *Ingeniörs Vetenskaps Akademien Handlinger*, *153*, 17.