Estimation of the quantile function using Bernstein-Durrmeyer polynomials

Andrey Pepelyshev
Ansgar Steland
Ewaryst Rafajłowic

Institut für Statistik & Wirtschaftsmathematik RWTH AACHEN

The work is supported by the German Federal Ministry of the Environment, Nature and Nuclear Safety, grant 0325226.

Hamburg
February 20, 2013

Contents

Bernstein-Durrmeyer operator

The BD operator $D_N(u(x))$ has the kernel form

$$D_N(u(x)) = \int_0^1 K_N(x,z)u(z)dz$$

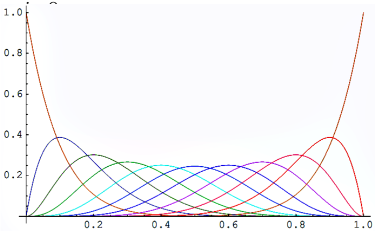
where $K_N(x,z) = (N+1)\sum_{i=0}^{N} B_i^{(N)}(x)B_i^{(N)}(z)$, or

$$D_N(u(x)) = (N+1)\sum_{i=0}^{N} a_i B_i^{(N)}(x)$$

where

$$a_i = \int_0^1 u(x)B_i^{(N)}(x)dx,$$

$$B_i^{(N)}(x) = \frac{N!}{i!(N-i)!}x^i(1-x)^{N-i}.$$



BD operator for deterministic functions

Derriennic (1981) shown that $D_N(u)$ converges uniformly,

$$\sup_{x\in[0,1]} |D_N(u(x)) - u(x)| \leq 2\omega_u(N^{-1/2}),$$

where ω_u denotes the modulus of continuity of u.
Chen and Ditzian (1991) established the fact that

$$\|u(x) - D_N(u(x))\|_p = O(N^{-\delta/2}),$$

if and only if

$$\epsilon(u) = \inf_{a_0,\ldots,a_N} \left\{ \|u(x) - a_0 - a_1 x - \cdots - a_N x^N\|_p \right\} = O(N^{-\delta}),$$

as $N \to \infty$ for some $\delta \in (0, 2)$ and $p \in \{2, \infty\}$, where
$\|f\|_p = \left(\int_I |f(x)|^p \, dx\right)^{1/p}$.

BD operator for the distribution function

Let $F_m(x)$ be the empirical d.f. for a sample $X_1, \dots, X_m \sim F_{[0,1]}$ with density f and

$$\widetilde{f}_{m,N}(x) = \frac{1}{m} \sum_{j=1}^{m} K_N(X_j, x), \quad \widetilde{F}_{m,N}(x) := \int_0^x \widetilde{f}_{m,N}(t) dt$$

Theorem. (Ciesielski, 1988) If $F \in C[0,1]$, then

$$\Pr(\|\widetilde{F}_{m,N}(x) - F\|_\infty \to 0 \text{ as } m, N \to \infty) = 1.$$

If $f \in L_p[0,1]$ and $m = \lfloor N^\beta \rfloor$ for some $\beta \in (0, 0.5)$, then

$$\Pr(\|\widetilde{f}_{m,N}(x) - f\|_p = o(1) \text{ as } N \to \infty) = 1.$$

BD operator for the quantile function

Let $Q_m(x)$ be the empirical q.f. for a sample $X_1, \ldots, X_m$.
Version 1:

$$\widetilde{Q}_{m,N}(x) := D_N(Q_m(x)) = (N+1)\sum_{i=0}^{N} \widetilde{a}_i B_i^{(N)}(x),$$

$$\widetilde{a}_i = \int_0^1 Q_m(x) B_i^{(N)}(x)dx, \qquad i = 1, \ldots, m,$$

Version 2:

$$\widehat{Q}_{m,N}(x) = (N+1)\sum_{i=0}^{N} \widehat{a}_i B_i^{(N)}(x),$$

$$\widehat{a}_i = \frac{1}{m}\sum_{j=1}^{m} X_{(j)} B_i^{(N)}(\frac{j-1}{m-1})$$

Note that $\widetilde{a}_i$ are asymptotically equivalent to $\widetilde{a}_i$ in mean square (Yang, 1985).

Properties of the BD estimator

Lemma. For $\widetilde{a}_i$ it is hold

$$\mathrm{Var}(\widetilde{a}_i) \leq \frac{1}{(N+1)^2} \max_{j=1,\dots,m} \mathrm{Var}(X_{(j)})$$

Proof. Set

$$\widetilde{w}_j = \frac{\int_{(j-1)/m}^{j/m} B_i^{(N)}(t)\,dt}{\sum_{l=1}^m \int_{(l-1)/m}^{l/m} B_i^{(N)}(t)\,dt},$$

$j = 1, \dots, m$. Then we have that

$$\mathrm{Var}(\widetilde{a}_i) = \left(\sum_{l=1}^m \int_{(l-1)/m}^{l/m} B_i^{(N)}(t)\,dt \right)^2 \mathbf{E} \sum_{j=1}^m (X_{(j)} - \mathbf{E}X_{(j)})^2 \widetilde{w}_j.$$

Apply the Jensen inequality $(\mathbf{E}_\zeta(a_\zeta - \mathbf{E}a_\zeta))^2 \leq \mathbf{E}_\zeta(a_\zeta - \mathbf{E}a_\zeta)^2$, where ζ is a random variable such that $\mathbb{P}(\zeta = j) = w_j$.

Properties of the BD estimator

Lemma. For the BD estimator it is hold

$$\mathrm{Var}(\widetilde{Q}_{m,N}(x)) \leq C_m := \max_{j=1,\dots,m} \mathrm{Var}(X_{(j)})$$

and

$$\int_0^1 \mathrm{Var}(\widetilde{Q}_{m,N}(x))dx \leq C_m.$$

If the derivative of $Q(x)$ is continuous, then $C_m = O(1/m)$.

Proof. If $X_1, \dots, X_m \sim U(0,1)$ then $X_{(j)} \sim \beta(j, m-j+1)$. Note that the variance of the beta distribution $\beta(a,b)$ is $\frac{ab}{(a+b)^2(a+b+1)}$. Therefore, it is easy to see that $C_m = O(1/m)$.

Maximum variance of order statistics

Since $Q'(x)$ is continuous, then we can write $X_{(j)} = Q(U_{(j)})$. Applying the Lagrange–Taylor formula for $Q(U_{(j)})$ at the point $\mathbf{E}U_{(j)} = j/(m+1) = p_j$, we obtain

$$X_{(j)} = Q(p_j) + (U_{(j)} - p_j)Q'(\zeta_j)$$

for some $\zeta_j \in [\min\{U_{(j)}, p_j\}, \max\{U_{(j)}, p_j\}]$. Then we have

$$\begin{aligned} \mathrm{Var}(X_{(j)}) &= \mathrm{Var}((U_{(j)} - p_j)Q'(\zeta_j)) \\ &\leq \max_x Q'^2(x)\mathbf{E}(U_{(j)} - p_j)^2 \\ &= \frac{p_j(1-p_j)}{m+2} \max_x Q'^2(x). \end{aligned}$$

MSE and MISE consistency

Let $X_1, \ldots, X_m$ be m random variables with common quantile function $Q(x)$ with a continuous derivative on $[a, b] \subset [0, 1]$ that satisfies the Chen-Ditzian condition with $p = \infty$. Then the Bernstein-Durrmeyer estimator with integral weights, $\widetilde{Q}_{m,N}(x)$, satisfies

$$\max_{x \in [a,b]} \mathbf{E}(\widetilde{Q}_{m,N}(x) - Q(x))^2 = O(1/m + N^{-\delta})$$

and

$$\mathbf{E} \int_a^b (\widetilde{Q}_{m,N}(x) - Q(x))^2 dx = O(1/m + N^{-\delta})$$

as $m \to \infty$ and $N \to \infty$.

The BD estimator with correction term

Let $\widetilde{e}_l(x) = D_l\Big(Q_m(x) - \widetilde{Q}_{m,N}(x)\Big)$.
Define the *error-corrected Bernstein-Durrmeyer estimator*

$$\widetilde{Q}_{m,N,l}(x) = \widetilde{Q}_{m,N}(x) + \widetilde{e}_l(x).$$

Theorem. Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with common quantile function $Q(x)$ and distribution function $F(x)$ satisfying $\int |x|^p dF(x) < \infty$ for some $1 \leq p \leq \infty$. Then the error-corrected estimator $\widetilde{Q}_{m,N,l}(x)$ is consistent in the pth mean for the true quantile function $Q(x)$,

$$\|\widetilde{Q}_{m,N,l} - Q\|_p \to 0,$$

as $m \to \infty$ and $N, l \to \infty$, with probability 1, and in the mean, i.e.

$$E\|\widetilde{Q}_{m,N,l} - Q\|_p \to 0,$$

as $m \to \infty$ and $N, l \to \infty$.

Adaptive choice of N

We modify an algorithm from (Golyandina, Pepelyshev, Steland, 2012) whose idea is based on a balance of

- the closeness of the BD estimator to the empirical q.f. in terms of the law of iterated logarithms and
- the stability of the number of modes of the corresponding density estimator.

Note that the Bernstein polynomial $B_i^{(N)}(x)$ can be approximated by the density of the normal distribution $\phi((i/N-x)/s)$, where $s=\sqrt{x(1-x)/N}$ and $\phi(x)=(2\pi)^{-0.5}e^{-x^2/2}$. Thus, we can introduce the quantity $h=1/\sqrt{N}$ which plays the role of the bandwidth.

Algorithm of adaptive choice of N

1] Compute

$$\bar{h} = \max\Big\{ \; h \in (0,1] : \max_q \big| F_m(\widehat{Q}_{m,\lceil 1/t^2 \rceil}(q)) - q \big|_\infty \le 1/R_m \;\; \forall\, t \in (0,h) \Big\}$$

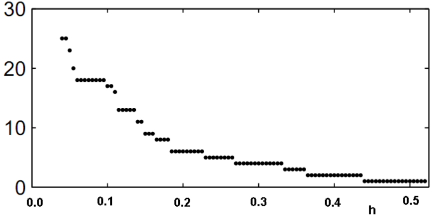
where $F_m(x)$ is the empirical distribution function, $\lceil z \rceil$ is the smallest integer that is larger or equal to z, and

$$R_m = 2\sqrt{m}/\sqrt{2\log\log m}.$$

Algorithm of adaptive choice of N

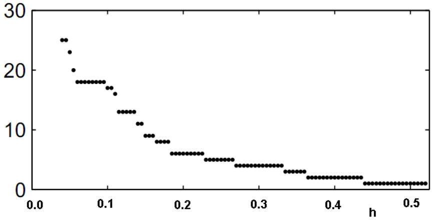
2] Define a set $\{h_1, \ldots, h_n\} \in (0, \bar{h})$ such that $\max_{h \in (0,\bar{h})} \min_{j=1,\ldots,n} |h - h_j|$ is small and compute the sequence $M_1, \ldots, M_n$, where M_j is the number of local minimums of the Bernstein-Durrmeyer estimator $\widehat{Q}_{m,N}(x)$ with $N = \lceil 1/h_j^2 \rceil$.

3] Compute $\check{M}_j = \min\{M_1, M_2, \ldots, M_j\}$, $j = 1, \ldots, n$.



Algorithm of adaptive choice of N

4] Divide the set $\{h_1, \ldots, h_n\}$ into groups as follows. Define a_i and b_i such that $a_1 \le b_1 < a_2 \le b_2 < \ldots < a_k \le b_k$ and $\check{M}_i = \check{M}_j$ for all $h_i, h_j \in [a_l, b_l]$ for $l \in \{1, \ldots, k\}$.



5] Compute $\widehat{h} = \sum_{i=1}^{k} a_i w_i$, where $w_i = (b_i - a_i) / \sum_{j=1}^{k} (b_j - a_j)$, and then set $\widehat{N} = \lceil 1/\widehat{h}^2 \rceil$.

Consistency of the BD estimator with $\hat{N}$

Let $X_1, \ldots, X_m$ be m random variables with the continuously differentiable quantile function Q, $R_m = \frac{2\sqrt{m}}{\sqrt{2 \log\log m}}$.
Then the adaptive Bernstein-Durrmeyer estimator $\widehat{Q}_{m,\widehat{N}}(q) = D_{\widehat{N}}(Q_m(q))$ where $\widehat{N}$ is selected by the above algorithm is consistent as $m \to \infty$. Moreover, we have the a.s. uniform error bound

$$\sup_{-\infty<x<\infty} |\widehat{Q}^{-1}_{m,\widehat{N}}(x) - F(x)| \leq \sqrt{2 \log\log m}/\sqrt{m}$$

for the estimator $\widehat{Q}^{-1}_{m,\widehat{N}}(x)$ of the distribution function $F(x)$, and the a.s. uniform error bound

$$\sup_{q} |\widehat{Q}_{m,\widehat{N}}(q) - Q(q)| \leq 2\sqrt{2 \log\log m}/\sqrt{m}$$

for the quantile estimator $\widehat{Q}_{m,\widehat{N}}(x)$.

Invariance principle

Theorem. If R_m is chosen such that $R_m^{-1} = o(m^{-1/2})$, then for the Bernstein-Durrmeyer polynomial estimator with data-adaptive selection of the degree N we have

$$\{\sqrt{m}[\widehat{Q}_{m,\widehat{N}}^{-1}(x) - F(x)] : \infty < x < \infty\} \Rightarrow \{B^0(F(x)) : -\infty < x < \infty\},$$

as $m \to \infty$, in the Skorohod space $D(\mathbb{R};\mathbb{R})$, where $B^0(t) = W_t - tW_1,\ t \in [0,1]$, is a Brownian bridge process.

Application in photovoltaics

Parameters of acceptance sampling for the quality control

- AQL, the acceptable quality level
- RQL, the rejectable quality level
- α, the producer risk
- β, the consumer risk

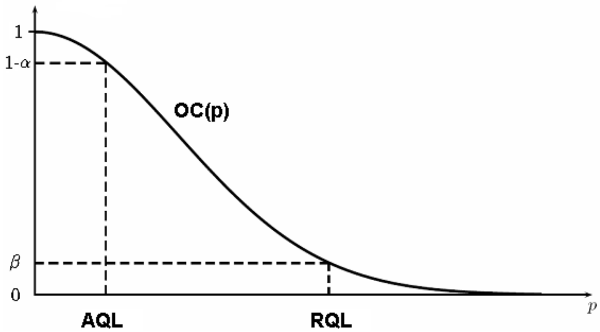
The sampling plan

- n, the number of items to be checked from a lot
- c, the critical value for the T-statistic

The meaning of parameters

The lot is accepted if and only if $T_n > c$.

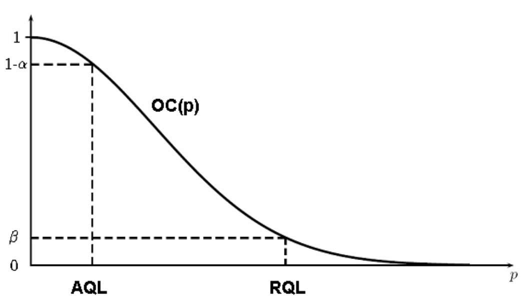
The operational characteristic $\mathrm{OC}_{n,c}(p) = \mathbb{P}(T_n > c)$ is the probability of the lot acceptance for given p, the true proportion of non-conforming items.



Requirements for the sampling plan

The sampling plan is a minimal solution satisfying the following conditions

$$\begin{cases} \mathrm{OC}_{n,c}(p) \geq 1 - \alpha \text{ for } p \leq \mathrm{AQL}, \\ \mathrm{OC}_{n,c}(p) \leq \beta \text{ for } p \geq \mathrm{RQL} \end{cases}$$



Sampling plans for quality control

If measurements are distributed according to a distribution $G(x)$ with mean a and variance σ^2, then

$$\mathrm{OC}_{n,c}(p) \cong 1 - \Phi\big(c + \sqrt{n} G^{-1}(1-p)\big).$$

The asymptotically optimal sampling plan (n, c) is

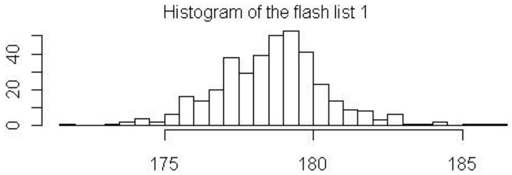
$$\begin{aligned} n &= \left\lceil \frac{\big(\Phi^{-1}(\alpha) - \Phi^{-1}(1-\beta)\big)^2}{\big(F^{-1}(\mathrm{AQL}) - F^{-1}(\mathrm{RQL})\big)^2} \right\rceil, \\ c &= -\frac{\sqrt{n}}{2} \left(F^{-1}(\mathrm{AQL}) + F^{-1}(\mathrm{RQL})\right), \end{aligned}$$

where $F(x) = G((x-a)/\sigma)$ and $\Phi(x)$ is the standard normal distribution.

Numerical results

Characteristics of estimation of the sampling plan (n, c) for two quantile estimators,
$\alpha = \beta = 5\%$, $\mathrm{AQL} = 2\%$ and $\mathrm{RQL} = 5\%$.

	Bernstein-Durrmeyer estimator				Empirical quantile estimator			
	n		c		n		c	
m	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.
200	57.9	24.8	13.3	2.3	84.0	178.3	14.7	7.7
400	57.9	19.9	13.5	1.9	59.3	40.0	13.5	3.8
800	55.7	14.0	13.4	1.5	54.1	23.4	13.2	2.6
1600	54.3	9.9	13.4	1.0	53.9	18.2	13.3	1.9



Thank you for your attention!