EFFICIENCY OF THE KERNEL METHOD FOR ESTIMATING A DISTRIBUTION FUNCTION AND PERCENTAGE POINTS

by A. Azzalini
Department of Mathematics, Imperial College, London

1979

SUMMARY

A distribution function is estimated by integrating a kernel estimator of the density. Properties of the resulting estimator are discussed and a rule for choosing bandwidth is given. The associated estimator of a percentage point is also considered; approximations to its expectation and mean square error are given.

Keywords: non-parametric estimation, distribution function, percentage points, kernel method, simulation.

1. INTRODUCTION

The usual estimate of a distribution function F, say, is the empirical distribution function (EDF): the present paper compares the EDF with another estimator. The problem is considered in the framework of estimation of the distribution function of some statistic via simulation, but a good deal of what follows can be applied for other purposes.

Generally, we deal with distribution functions which admit a density with respect to Lebesgue measure and, fairly often, there are reasons for believing that this density is "smooth", at least locally. This suggests estimators of F which use this information, for instance by integrating a nonparametric estimate of the density (Nadaraya, 1964).

Comparison of this alternative estimate $\hat{F}_n(x)$, say, with EDF is made by the mean square error (MSE) at a particular value of x. In section 2 the bias and the MSE of $\hat{F}_n(x)$ are evaluated and a rule for choosing the bandwidth is derived.

In section 3 the inverse problem is considered. Let us define the p-th quantile of F as the root ξ_p of

$$p = F(\xi_p) \qquad (0<p<1)$$

assumed to be unique. We estimate ξ_p by x_p which is the root of

$$p = \hat{F}_n(x_p) \; ; \qquad (1)$$

approximate values of the mean and the MSE of x_p are given.

The present paper is a shorter version of the author's M.Sc. report at the Imperial College, where more extensive results are given. In particular, estimators defined by Kronmal and Tarter (1968) are considered; these results are not reported here.

2. ESTIMATING THE DISTRIBUTION FUNCTION

Let X be a random variable with absolutely continuous distribution function F and probability density function f. Let $X_1,\ldots,X_n$ be independently and identically distributed random variables with distribution function F. Let us consider estimators of f(x), where x is some given real number, of the form

$$\hat{f}_n(x) = \frac{1}{n\,b} \sum_{i=1}^{n} w\left(\frac{x-X_i}{b}\right) \qquad (2)$$

where w is a bounded density function called the kernel and b is a positive number called the bandwidth.

The "obvious" estimator of $F(x)$ is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} W\left(\frac{x-X_i}{b}\right) \tag{3}$$

where

$$W(t) = \int_{-\infty}^{t} w(u)\, du \ .$$

Nadaraya (1964) has proved under mild conditions that $\hat{F}_n(x)$ is an asymptotically unbiased and consistent estimator for $F(x)$, and

$$\lim_{n\to\infty} n \operatorname{var} \hat{F}_n(x) = F(x)(1-F(x)) \ .$$

It has been pointed out by several authors that the choice of w is not critical, at least estimating a density function; see e.g. Wertz (1978, p.45). However, there are reasons for choosing w with finite range, i.e.

$$w(t) = 0 \qquad \text{if } t \notin (-h, h)$$

where h is some positive number depending on w. In fact, (i) this allows numerical efficiency; (ii) any regularity condition on F has to be satisfied only locally, namely in (x_1, x_2), where

$$x_1 = x-hb \ , \qquad x_2 = x+hb \ .$$

If we consider kernels with finite range, then

$$E\, \hat{F}_n(x) = \alpha_{10} + \vartheta \tag{4}$$

$$\operatorname{var} \hat{F}_n(x) = (\alpha_{20} + \vartheta - (\alpha_{10} + \vartheta)^2)/n \tag{5}$$

where

$$\vartheta = \int_{-\infty}^{x_1} f(t)\, dt$$

$$\alpha_{rs} = \int_{x_2}^{x_1} \left\{ W\left(\frac{x-t}{b}\right) \right\}^r \left\{ w\left(\frac{x-t}{b}\right) \right\}^s f(t)\, dt, \quad (r,s=0,1,\ldots)$$

Let us assume that w is an even density. The following relationships hold

$$\int_{-h}^{h} W(t)\, dt = h\,, \qquad \int_{-h}^{h} W(t)\, w(t)\, dt = 1/2\,,$$

$$\int_{-h}^{h} t\, W(t)\, dt = \left\{ h^2 - \int_{-h}^{h} t^2\, w(t)\, dt \right\}/2 \quad .$$

Then,

$$E\,\hat{F}_n(x) - F(x) = \int_{x_1}^{x_2} W\left(\frac{x-t}{b}\right) f(t)dt - \int_{x_1}^{x} f(t)\, dt$$

$$= \frac{1}{2}\, b^2\, f'(x) \int_{-h}^{h} t^2\, w(t)\, dt + o(b^2) \qquad (6)$$

assuming that $f'(x)$ exists. The assumption of continuity of $f(t)$, $t\epsilon(x_1,x_2)$, is sufficient to say that the bias of $\hat{F}_n(x)$ is $O(b^2)$ and

$$\text{var}\,\hat{F}_n(x) = F(x)(1-F(x))/n +$$

$$- f(x)b\left\{ h - \int_{-h}^{h} W^2(t)\, dt \right\}/n + O(b^2/n),$$

where the term in braces is positive. Therefore

$$E(\hat{F}_n(x)-F(x))^2 \sim F(x)(1-F(x))/n - ub/n + vb^4 \qquad (7)$$

where

$$u = f(x)\left\{ h - \int_{-h}^{h} W^2(t)\, dt \right\}, \qquad v = \left\{ \frac{1}{2}\, f'(x) \int_{-h}^{h} t^2\, w(t)\, dt \right\}^2.$$

We get the minimum of of the right hand side of (7) at

$$b = \left(\frac{u}{4\, v\, n}\right)^{1/3} \qquad (8)$$

for which

$$E(\hat{F}_n(x)-F(x))^2 \sim \frac{F(x)(1-F(x))}{n} - \frac{3}{v^{1/3}}\left(\frac{u}{4\,n}\right)^{4/3}.$$

Therefore, if we choose b of the form

$$b = K\,n^{-\delta} \qquad (K,\delta > 0)$$

then $\delta=1/3$. We note that this value is different from the optimum value for density estimation ($\delta=1/5$). Clearly, in practical situations, we must choose K ignoring u and v. The literature on density estimation is not very helpful about this point. Silverman (1978) has given a rule for choosing b, but it is not suitable for computer simulation because it involves the computation of the second derivative of the estimate for several values of b and a decision "by eye".

An heuristic argument suggests putting $K=\sigma$, where σ is the standard deviation of X. We can improve this slightly in the following way. Let us consider a special form of w, e.g.

FIGURE A

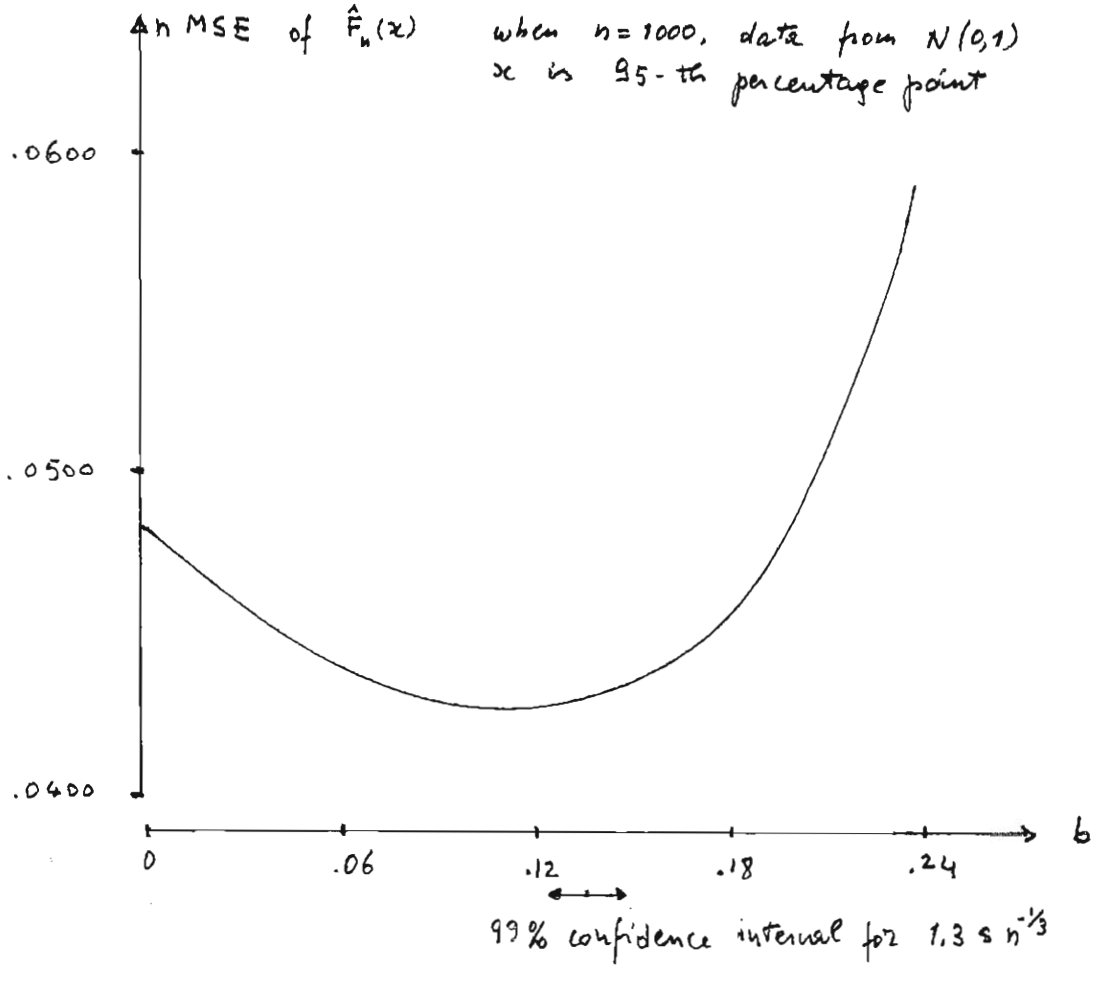

FIGURE B

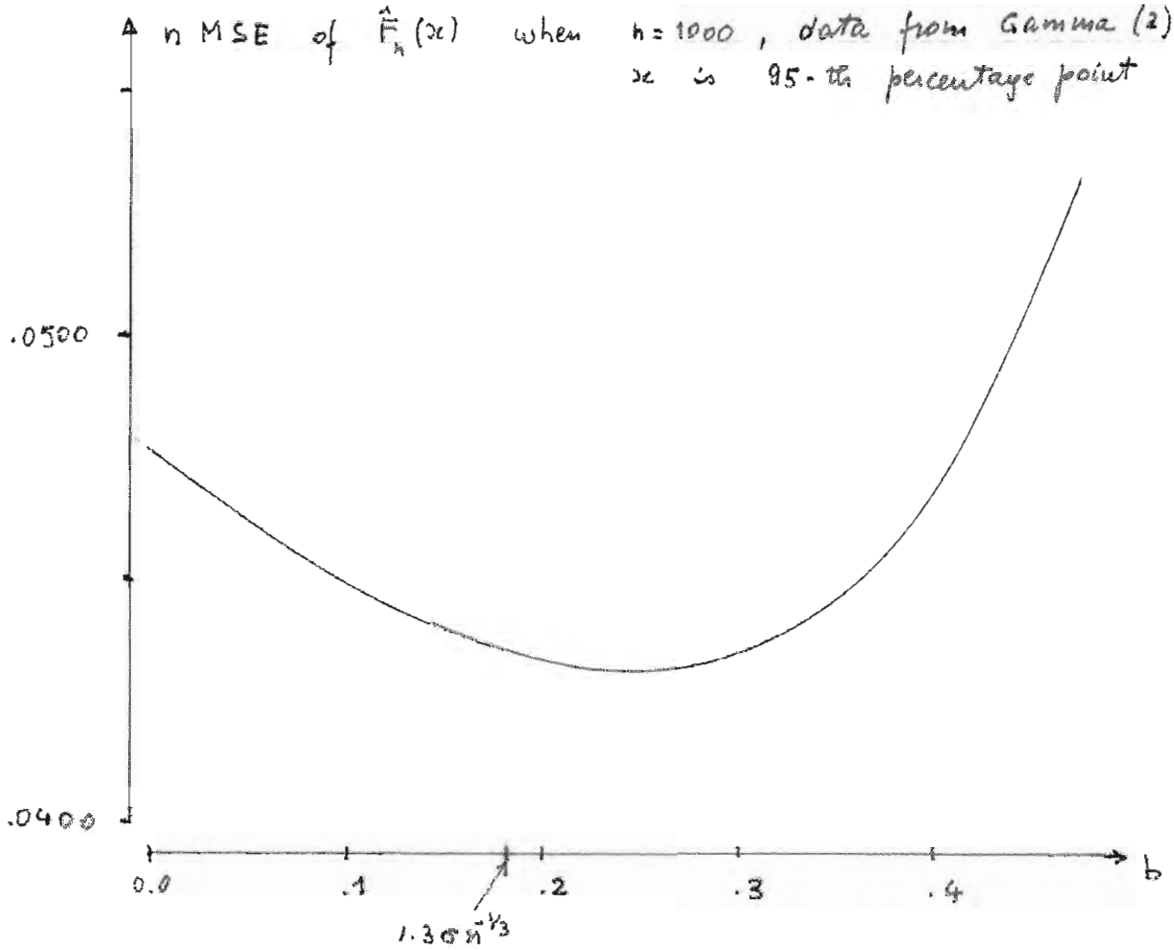

FIGURE C

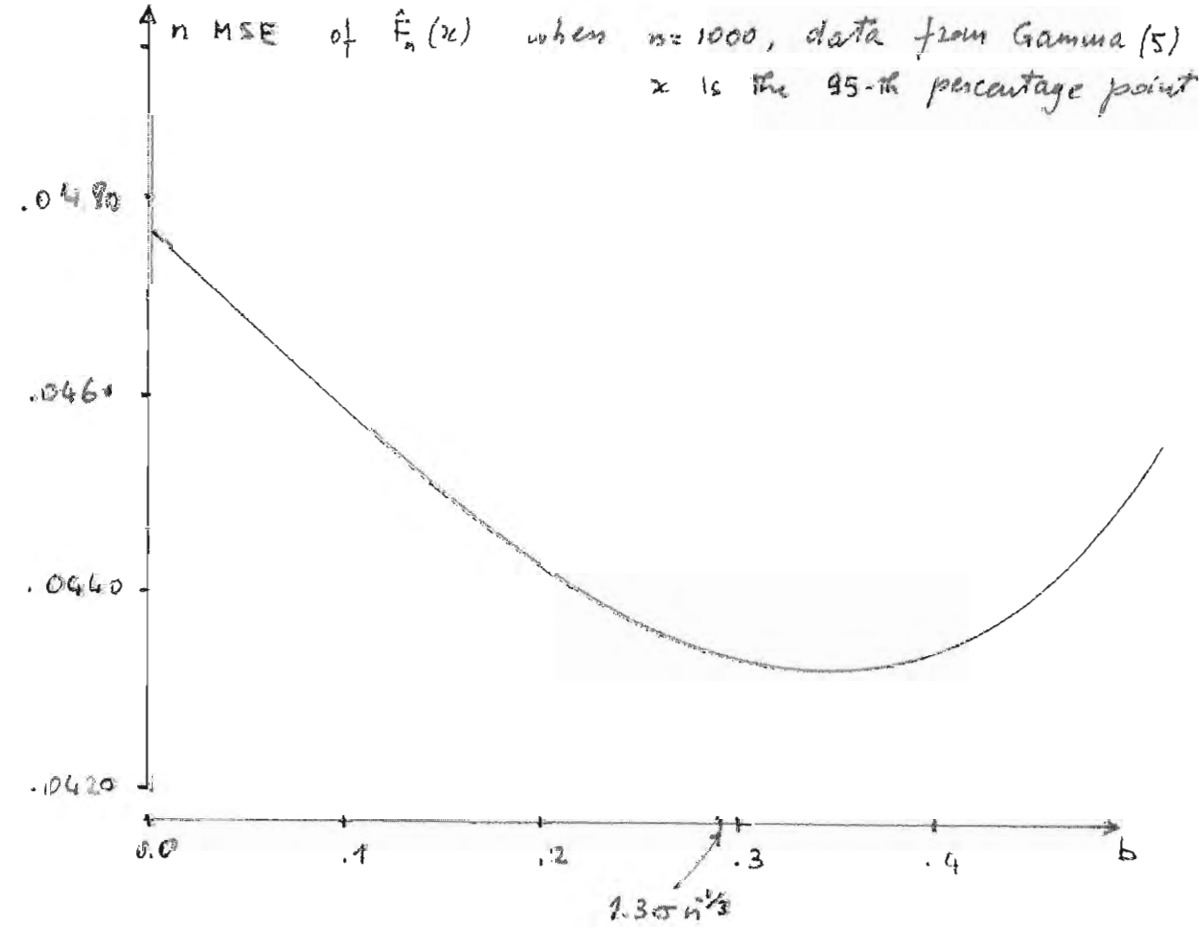

TABLE 1

Optimum value of K and minimum n MSE when n=1000 and x is the p-th percentage point

Variable	K	σ	n.MSE	K/σ
p=0.95, Beta (1,1)	.3573	.2687	.02960	1.237
Beta (1,2)	.3634	.2357	.04109	1.542
Beta (2,5)	.2342	.1598	.04261	1.465
Beta (1,5)	.2587	.1409	.04290	1.836
Beta (5,5)	.1690	.1508	.04228	1.121
Gamma (1/2)	1.1619	.7071	.04375	1.643
Gamma (1)	2.0112	1.	.04359	2.011
Gamma (2)	2.5164	1.4142	.04346	1.780
Gamma (5)	3.4781	2.2361	.04331	1.555
Gamma (10)	4.5465	3.1623	.04322	1.438
Gamma (20)	6.0512	4.4721	.04314	1.353
Normal	1.1459	1.	.04289	1.146
p=0.05, Beta (1,2)	.1887	.2357	.02901	.800
Beta (2,5)	.1292	.1597	.04005	.809
Beta (1,5)	.0780	.1409	.02868	.554
Gamma (1)	.3988	1.	.02846	.399
Gamma (2)	.8003	1.4142	.03957	.566
Gamma (5)	1.6657	2.2361	.04178	.745
Gamma (10)	2.7094	3.1623	.04226	.856
Gamma (20)	4.2026	4.4721	.04250	.940

Epanechnikov's kernel, of which the integral is

$$W_o(t) = 0 \qquad \text{if } t \leqslant -\sqrt{5},$$

$$= \frac{1}{2} + \frac{3}{4}\frac{t}{\sqrt{5}} - \frac{t^3}{20\sqrt{5}} \qquad \text{if } -\sqrt{5} < t < \sqrt{5},$$

$$= 1 \qquad \text{if } t \geqslant \sqrt{5},$$

which is known to have optimal properties estimating the density function; for details see, e.g., Rosenblatt (1971). Let us fix a value of n, say n=1000; let us finally choose a few density functions and find the optimum value of b (and of K) in each case estimating the distribution function at the 5-th and 95-th percentage point. Interpolating these results the two following empirical approximations have been obtained

$$K = 1.3\,\sigma, \qquad K = 0.5\,\sigma.$$

The first one is devised to give best results in the long tail of the distribution (which is the most common case in practice), while the second one can be used over the whole range. See Figure A, B & C, and Table 1 for details.

We note that it is very frequently the case that the first two moments of the statistic we are studying are known. If σ is not known we estimate it from the sample. Since in simulation studies we use large samples, we expect to have a very precise estimate of σ. However, relatively large estimation errors do not affect drastically the results; Figure A, B & C illustrate this point.

Table 2 gives n MSE of $\hat{F}_n(t)$ using W_o, $\delta = 1/3$, n=100 & 1000, x being the p-th percentage point. The MSE has been evaluated using (4) and (5) with numerical computation of α_{10} and α_{20}. Accuracy is full in all cases except when estimating at the 5-th percentage point of Gamma(0.5), where only three digits can be relied on. We remind ourselves that, if we use the EDF (which corresponds to b=0), n MSE = p(1-p) for each entry.

Results show that the kernel estimator is nearly always

superior to the EDF. It happens otherwise with Gamma(0.5) estimating at the 5-th percentage point, because the density is unbounded in (x_1, x_2).

3. ESTIMATING THE PERCENTAGE POINTS

Let us consider x_p which is defined by (1). Nadaraya (1964) has proved that, under certain regularity conditions, x_p is asymptotically normal with mean ζ_p and variance

$$\frac{p(1-p)}{n\, f^2(\zeta_p)}$$

which is the same asymptotic distribution of the estimator using order statistics.

We attempt here to give closer approximations to the mean and the MSE of x_p. Let us start from the identity

$$Z + \hat{F}_n(x_p) - \hat{F}_n(\zeta_p) = 0$$

where

$$Z = \hat{F}_n(\zeta_p) - p \ .$$

If we assume that $\hat{F}_n$ has two derivatives, then

$$Z + (x_p-\zeta_p)\hat{f} + \frac{1}{2}(x_p-\zeta_p)^2\, \hat{f}' \simeq 0 \qquad (9)$$

where $\hat{f}=\hat{f}_n(\zeta_p)$, $\hat{f}'=\hat{f}_n'(\zeta_p)$ and the approximation sign is due to a random variable which is $o_p(1/n)$ if $b^4=o(1/n)$. Taking the approximation sign in (9) as an equality and expanding the solution by the binomial theorem we get

$$E(x_p) = \zeta_p - \frac{E(Z)}{f} + \frac{E(ZU)}{f^2} - \frac{f'\, E(Z^2)}{2\, f^3} + o(1/n) \qquad (10)$$

$$E(x_p-\zeta_p)^2 = \frac{E(Z^2)}{f^2} - \frac{2\,E(Z^2U)}{f^3} + \frac{E(Z^2U^2)}{f^4} + O(n^{-3/2}) \qquad (11)$$

where the argument of f anf f' is ζ_p and $U=\hat{f}-f$. For evaluating (10) and (11), E(Z) and $E(Z^2)$ are already known from section 2; E(ZU), $E(Z^2U)$ and $E(Z^2U^2)$ can be computed using

$$E(\hat{f}) = \alpha_{01}/b$$

$$\mathrm{var}(\hat{f}) = (\alpha_{02}-\alpha_{01}^2)/(b^2n)$$

$$\mathrm{cov}(\hat{f},\hat{F}) = (\alpha_{11}-\nu\alpha_{01})/(bn)$$

$$E(\hat{F}^2\hat{f}) = \{\alpha_{21}+(n-1)(\alpha_{20}+\vartheta)\alpha_{01}+2(n-1)\nu\alpha_{11}+ +(n-1)(n-2)\nu^2\alpha_{01}\}/(bn^2)$$

$$E(\hat{F}\hat{f}^2) = \{\alpha_{12}+(n-1)\nu\alpha_{02}+2(n-1)\alpha_{01}\alpha_{11}+ +(n-1)(n-2)\alpha_{01}^2\nu\}/(b^2n^2)$$

$$E(\hat{F}^2\hat{f}^2) = \{\alpha_{22}+(n-1)(\alpha_{20}+\vartheta)\alpha_{02}+2(n-1)\nu\alpha_{12}+ +(n-1)(n-2)\nu^2\alpha_{02}+2(n-1)\alpha_{21}\alpha_{01}+ +(n-1)(n-2)(\alpha_{20}+\vartheta)\alpha_{01}^2+2(n-1)\alpha_{11}^2+ +4(n-1)(n-2)\alpha_{11}\nu\alpha_{01}+ +(n-1)(n-2)(n-3)\nu^2\alpha_{01}^2\}/(b^2n^3)$$

where $\nu=\alpha_{10}+\vartheta$ and $\hat{F}=\hat{F}_n(\zeta_p)$. It can be proved that $E(Z^2U)=O(\max\{1/n^2, b^2/n, b^6\})$. Therefore (8) is the asymptotically optimum bandwidth also for minimizing MSE of x_p. Correspondingly,

$$E\, x_p \sim \zeta_p - \frac{\mathrm{sign}(f')}{f\, v^{1/6}} \left(\frac{u}{4\, n}\right)^{2/3},$$

$$E(x_p - \zeta_p)^2 \sim \frac{p(1-p)}{n\, f^2} - \frac{3}{f^2\, v^{1/3}} \left(\frac{u}{4\, n}\right)^{4/3}.$$

If Epanechnikov's kernel is used the derivative of $\hat{f}_n(x)$ does not exist everywhere, but this happens to occur at ζ_p with probability 0. Moreover, for moderate n, $\hat{f}_n$ is so smooth that we can deal with it as if it were differentiable everywhere.

The mean and the MSE of x_p have been evaluated using (10) & (11) when X is a Gamma(1) or a standard Logistic random variable, p=.90, $b=1.3\,\sigma\, n^{-1/3}$, σ being 1 or $\pi/\sqrt{3}$, respectively. In these cases the mean and the variance of the s-th order statistic, where s=np is supposed to be an integer, are known exactly. See Johnson & Kotz (1970, vol. 1, p. 216) for the Gamma(1) and Johnson & Kotz (1970, vol. 2, p. 8) for the Logistic distribution. Results are shown in Table 3. Also in this case the α's integrals have been evaluated numerically; accuracy is full in all cases. For all entries x_p is rather better than the sample quantile, although not by a great deal.

In applications, for the actual computation of x_p, the Newton-Raphson search can be used, i.e. if x'_p is an approximation to x_p then

$$x''_p = x'_p + \frac{p - \hat{F}_n(x'_p)}{\hat{f}_n(x'_p)}$$

is the next approximation. It can be initialized with the sample quantile. A few simulations have been carried out with n=1000; in all cases two iterations were sufficient for 10^{-4} precision.

Acknowledgments

I would like to express my gratitude to Dr.A.C.Atkinson and to Professor D.R.Cox for several comments and suggestions. My work was supported by the Italian Ministry of Education.

REFERENCES

Azzalini, A. (1979). Comparison of methods for estimating the distribution function. M.Sc. report, Dept. of Mathematics, Imperial College, London.

Johnson, N.L. and Kotz, S. (1970). Distributions in statistics, continuous univariate distributions. Houghton Mifflin, Boston.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series method. J.Amer.Statist.Assoc., 63,925-952.

Nadaraya, E.A. (1964). Some new estimates for distribution functions. Theory Prob.Appl., 15, 497-500.

Rosenblatt, M. (1971). Curve estimates. Ann.Math.Statist., 42, 1815-1842.

Silverman, B.W. (1978). Choosing the window width when estimating a density. Biometrika, 65, 1-11.

Wertz, W. (1978). Statistical density estimation: a survey. Vandenhoeck & Ruprecht, Göttingen.

Table 2. n MSE estimating the distribution function at the p-th quantile

Distribution	p=.05 p(1-p)=.0475		p=.90 p(1-p)=.09		p=.95 p(1-p)=.0475	
	K=1.3σ	K=0.5σ	K=1.3σ	K=0.5σ	K=1.3σ	K=0.5σ
(n=100)						
Normal	.04047	.04205	.07562	.08031	.04047	.04205
Gamma(1/2)	2.36340	.71606	.07923	.08491	.04212	.04508
Gamma(1)	.54426	.06218	.07821	.08428	.04141	.04466
Gamma(2)	.13431	.03410	.07732	.08360	.04080	.04422
Gamma(5)	.05758	.04033	.07647	.08275	.04024	.04367
Beta(5,5)	.03974	.04120	.07307	.07960	.03974	.04120
Beta(2,5)	.05642	.03618	.07437	.08199	.03894	.04280
Beta(2,10)	.07982	.03490	.07597	.08290	.03988	.04359
Beta(1,1)	.04452	.03030	.05095	.07212	.04452	.03030
Beta(1,4)	.23297	.03199	.07579	.08330	.03949	.04360
(n=1000)						
Normal	.04305	.04477	.08193	.08526	.04305	.04477
Gamma(1/2)	8.91245	1.97219	.08445	.08755	.04472	.04633
Gamma(1)	.71826	.03334	.08388	.08724	.04430	.04612
Gamma(2)	.09075	.03976	.08335	.08690	.04391	.04590
Gamma(5)	.05030	.04269	.08279	.08648	.04349	.04562
Beta(5,5)	.04254	.04434	.08070	.08583	.04254	.04434
Beta(2,5)	.04716	.04162	.08179	.08613	.04272	.04519
Beta(2,10)	.05782	.04074	.08264	.08657	.04334	.04559
Beta(1,1)	.02976	.03920	.06842	.08170	.02976	.03920
Beta(1,4)	.19292	.02995	.08271	.08678	.04323	.04560

Table 3. Estimating the 90-th quantile

Gamma(1): $\xi_{.90}=2.30259$

n	x_p mean	x_p n MSE	sample quantile mean	sample quantile n MSE
100	2.3368	7.4306	2.2584	8.7168
250	2.3237	7.8526	2.2847	8.8845
500	2.3163	8.0981	2.2936	8.9419
1000	2.3113	8.2911	2.2981	8.9708
2500	2.3073	8.4836	2.3008	8.9883
5000	2.5056	8.5930	2.3017	8.9942
10000	2.3045	8.6789	2.3021	8.9971
25000	2.3036	8.7648	2.3024	8.9988

using formulae of page 3 ($b = 1.7815\, n^{-1/3}$ from (8))

2.3770 7.3355

2.3186 8.2274

2.3060 8.6414 ~~8.4732~~

2.3044 8.7358

Logistic: $\xi_{.90}=2.19722$

n	x_p mean	x_p n MSE	sample quantile mean	sample quantile n MSE
100	2.2842	7.9059	2.1425	10.5141
250	2.2494	8.8583	2.1751	10.8509
500	2.2312	9.3976	2.1862	10.9770
1000	2.2190	9.8080	2.1917	11.0430
2500	2.2092	10.1996	2.1950	11.0836
5000	2.2048	10.4121	2.1961	11.0973
10000	2.2020	10.5726	2.1967	11.1042
25000	2.1998	10.7270	2.1970	11.1084