INT. J. MATH. EDUC. SCI. TECHNOL., 1978, VOL. 9, NO. 2, 127–137

Estimation of probability densities by empirical density functions†

by M. S. WATERMAN and D. E. WHITEMAN
Los Alamos Scientific Laboratory, Los Alamos, New Mexico, U.S.A.

(*Received* 17 *March* 1977)

The empirical density function, a simple modification and improvement of the usual histogram, is defined and its properties are studied. An analysis is presented which enables the interval width to be chosen. The estimators are modified for the important practical case of bounded random variables. Finally, the problems of writing a programme to compute the functions are considered along with some Monte Carlo examples and a practical example from the National Uranium Resource Evaluation study conducted by the United States Energy Research and Development Administration. It is recommended that these techniques be introduced at all levels of statistical courses so that they will become more widely utilized.

1. Introduction

The first technique to be taught in an elementary statistics course is often the construction of histograms. On the surface, constructing histograms is easy, and students quickly understand why histograms should be constructed. For applications, displaying data in the form of histograms is a most important statistical activity. A well-constructed histogram often tells an experimenter things at a glance that would be almost impossible to ascertain were the data displayed in any other manner. Therefore, even though the elementary course quickly moves on to more 'interesting' topics, it is important to carefully consider histograms and the display of data.

Before constructing a histogram, three factors must be determined: (i) the number of classes, (ii) the width of each class, and (iii) the lower limit of the first class. None of these choices is easy, although certain rules of thumb have been developed. The question of the number of classes was considered by Sturges [1] who in 1926 proposed that $1+\log_2(n)$ classes should be used for n observations. His analysis was based on the normal distribution, and frequently gives too few classes. Recently, Doane [2] considered these problems and devised an algorithm for constructing histograms that can be implemented on a computer. It is worthy of note that such an algorithm was not developed until late 1976.

The object of this paper is to present and study empirical density functions, a histogram-like estimate of the underlying density function that is easy to understand and introduce. It is frequently observed that the convergence of empirical density functions to the density function is not good, but it is also most important to observe that they are superior to the usual histograms.

 This work was carried out under the auspices of the United States Energy Research and Development Administration under contract W-7405-ENG. 36.

Empirical density functions belong to a large class of non-parametric density estimators, a field surveyed by Wegman [3, 4]. We restrict ourselves to empirical density functions (' naïve estimators ') because (i) they are a simple modification of histograms and therefore easily introduced, and (ii) their convergence properties are equivalent to those of the larger class of estimators. We should point out that while our material could be naturally introduced after empirical distribution functions in any senior level course, it is rarely mentioned in such texts [5, 6, 7].

We introduce our estimators, derive some simple properties, and then present an analysis which enables an interval width to be chosen. Next, the estimators are modified for the important case of bounded random variables, and the expected sample moments of the empirical density function are calculated. Finally, the problems of writing a programme to compute the functions are considered along with some Monte Carlo examples and a practical example.

We hope this paper will also show that these estimators can be introduced and studied in both elementary and more advanced courses. We feel it is important that this be done.

2. Empirical density functions

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with density function $f(x)$ and distribution function $F(x)$. The empirical distribution function is defined by

$$F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n}$$

where $\#A$ denotes the number of elements in the set A. By an easy application of the binomial distribution [5], it is seen that

$$\lim_{n\to\infty} F_n(x) = F(x)$$

with probability 1. In fact, the Kolmogorov–Smirnov test [6] is based on the relationship between $F_n(x)$ and $F(x)$.

A natural estimator for $f(x)$ is suggested by the fact $dF(x)/dx = f(x)$. Therefore, we consider the approximate derivative of F_n,

$$g_n(x) = \frac{F_n(x+\lambda) - F_n(x-\lambda)}{2\lambda}$$

where $\lambda > 0$. This estimator is referred to as the naïve estimator in some of the literature [3], but here we follow Révész [8] and refer to $g_n(x)$ as the empirical density function.

For an alternate derivation, consider a modification of the classical histogram [7] in which a rectangle of height 1 and width 2λ is centred at each data point X_i. At any real number x, consider the contribution from all X_i:

$$Y(x) = \#\{i : X_i - \lambda < x \leq X_i + \lambda\}$$

Clearly, since there are n rectangles, each of area 2λ,

$$\int_{-\infty}^{\infty} Y(x)\, dx = 2n\lambda$$

Thus, $Y(x)/2n\lambda$ will be a normalized function. Also, by the binomial distribution,

$$P(Y(x)=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

where $p=F(x+\lambda)-F(x-\lambda)$ is the probability of each X_i satisfying $X_i-\lambda<x\leq X_i+\lambda$ or $x-\lambda\leq X_i<x+\lambda$. The connection with $g_n(x)$ arises because

$$Y(x)=n(F_n(x+\lambda)-F_n(x-\lambda))$$

so that

$$g_n(x)=Y(x)/2n\lambda$$

As with the empirical distribution function,

$$\lim_{n\to\infty}g_n(x)=g(x)\equiv\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}$$

with probability one. From this result it is clear that as $n\to\infty$, we should take $\lambda=\lambda_n\to 0$. It is important to find the rate at which λ_n converges to 0, a point which is taken up in the next section.

3. Choice of λ

Rosenblatt [9] has shown that for a criterion of expected mean-square error, the best $\lambda=\lambda_n$ is a constant times $n^{-1/5}$, assuming, however, the existence of three derivatives of f at x. Parzen [10] has also treated the problem in a similar way. We employ the Kolmogorov–Smirnov statistic to obtain bounds using only properties of $f'(x)$.

Now,

$$\left|\frac{Y(x)}{2n\lambda}-f(x)\right|\leq\left|\frac{Y(x)}{2n\lambda}-\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}\right|+\left|\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}-f(x)\right|$$

Then, to bound the first quantity on the right-hand side,

$$\left|\frac{Y(x)}{2n\lambda}-\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}\right|\leq\left|\frac{F_n(x+\lambda)-F(x+\lambda)}{2\lambda}\right|+\left|\frac{F_n(x-\lambda)-F(x-\lambda)}{2\lambda}\right|$$

Let $D_n(\alpha)$ satisfy the equation $P\left(\max_x|F_n(x)-F(x)|>D_n(\alpha)\right)=\alpha$.

Then

$$\left|\frac{Y(x)}{2n\lambda}-\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}\right|\leq\frac{D_n(\alpha)}{\lambda}$$

with probability at least $1-\alpha$. To bound the second quantity on the right-hand side, note that

$$F(x\pm\lambda)=F(x)+F'(x)\,(\pm\lambda)+\frac{F''(x_i)(\pm\lambda^2)}{2}$$

Then

$$\left|\frac{F(x+\lambda)-F(x-\lambda)}{2\lambda}-f(x)\right| = \left|f(x)+\frac{\lambda}{4}(f'(x_1)-f'(x_2))-f(x)\right|$$

$$= \frac{\lambda}{4}|f'(x_1)-f'(x_2)|$$

If $|f'(x)| \leq C$ for all x, then

$$|g_n(x)-f(x)| \leq \frac{D_n(\alpha)}{\lambda}+\frac{C\lambda}{2} \equiv B_1(\lambda)$$

(Small λ makes the approximate derivative of F close to f but too small a value for λ makes $g_n(x)$ very rough.) The λ which minimizes $B_1(\lambda)$ is

$$\lambda^{(1)} = (2D_n(\alpha)/C)^{1/2}$$

and, if we let $D_n(\alpha) = K(\alpha)n^{-1/2}$, the asymptotic form given in [6], then

$$\lambda^{(1)} = (2K(\alpha)/C)^{1/2}n^{-1/4}$$

If $|f'(x)-f'(y)| \leq L|x-y|$, then

$$|g_n(x)-f(x)| \leq \frac{\lambda^2 L}{4}+\frac{D_n(\alpha)}{\lambda} = B_2(\lambda)$$

The λ which minimizes $B_2(\lambda)$ is

$$\lambda^{(2)} = (D_n(\alpha)/2L)^{1/3}$$

and the asymptotic form is

$$\lambda^{(2)} = (K(\alpha)/2L)^{1/3}n^{-1/6}$$

It should be emphasized that the results derived here hold uniformly (for all x) with probability at least $1-\alpha$. The approach of Rosenblatt [9] yields an optimal λ of $O(n^{-1/5})$ which is between our answers of $O(n^{-1/4})$ and $O(n^{-1/6})$.

4. Bounded random variables

For many applications, it is known that $a \leq X$ and/or $X \leq b$ for given constants a and b. However, the data will often be such that $g_n(x)$ is positive for $x < a$ or $b < x$ and it is therefore important to adjust $g_n(x)$ so that it is zero for $x < a$ or $b < x$. Later we illustrate how important this is in estimating $f(x)$ by a Monte Carlo example.

If we assume that $a \leq X$ and use $\lambda = \lambda_n > 0$, the x values of interest are $a \leq x < a+\lambda$. For these values of x,

$$P(Y(x)=k) = F(x+\lambda)-F(a) = F(x+\lambda)$$

Rather than estimate $f(x)$ by

$$\frac{F(x+\lambda)}{2\lambda} = \frac{F(x+\lambda)-F(a)}{2\lambda}$$

we use

$$\frac{F(x+\lambda)}{x+\lambda-a} = \frac{F(x+\lambda)-F(a)}{x+\lambda-a}$$

The corresponding (modified) empirical density function is then

$$g_n^*(x)=\frac{F_n(x+\lambda)-F_n(a)}{x+\lambda-a}=\frac{Y(x)}{n(x+\lambda-a)}$$

For the general case of $a\leq X$ and/or $X\leq b$, the function $g_n^*(x)$ is defined by

$$g_n^*(x)=\begin{cases}\dfrac{Y(x)}{n(x+\lambda-a)}, & \text{if } a\leq x<a+\lambda\\[2ex] \dfrac{Y(x)}{2n\lambda}, & \text{if } a+\lambda\leq x\leq b-\lambda\\[2ex] \dfrac{Y(x)}{n(b+\lambda-x)}, & \text{if } b-\lambda<x\leq b\\[2ex] 0, & \text{otherwise}\end{cases}$$

Notice that we still have the property that $g_n^*(x)\to f(x)$ for all x, with probability one. This estimator does not not seem to have been given elsewhere, probably due to the emphasis on theoretical results.

5. Expected moments

Since an experimenter obtains information about the distribution by examining $g_n(x)$, it is of some interest to calculate the expected moments of $g_n(x)$. Let

$$m_n'=\int_{-\infty}^{\infty}x^n g_n(x)\,dx$$

and

$$\mu_n'=E(X^n)$$

where $n\geq 0$. Now, if $A=(x-\lambda,\ x+\lambda)$,

$$E(m_n')=E\left\{\int_{-\infty}^{\infty}x^n\frac{Y(x)}{2n\lambda}\,dx\right\}$$

$$=\frac{1}{2\lambda}\int_{-\infty}^{\infty}x^n(F(x+\lambda)-F(x-\lambda))\,dx$$

$$=\frac{1}{2\lambda}\int_{-\infty}^{\infty}x^n\int_{-\infty}^{\infty}I_A(t)f(t)\,dt\,dx$$

$$=\frac{1}{2\lambda}\int_{-\infty}^{\infty}f(t)\int_{-\infty}^{\infty}x^n I_A(t)\,dx\,dt$$

$$=\frac{1}{2\lambda(n+1)}\int_{-\infty}^{\infty}f(t)((t+\lambda)^{n+1}-(t-\lambda)^{n+1})\,dt$$

Using the binomial theorem,

$$(t+\lambda)^{n+1}-(t-\lambda)^{n+1}=\sum_{k=0}^{n+1}\binom{n+1}{k}t^k\lambda^{n+1-k}(1-(-1)^{n+1-k})$$

$$=2\sum_{\substack{0\leqslant k\leqslant n+1\\ n-k\text{ even}}}\binom{n+1}{k}t^k\lambda^{n+1-k}$$

Therefore

$$E(m_n')=\frac{1}{n+1}\sum_{n-k\text{ even}}\binom{n+1}{k}\mu_k'^{n-k}$$

For $n=1, 2$,

$$E(m_1')=\tfrac{1}{2}\binom{2}{1}\mu_1'=\mu_1'$$

and

$$E(m_2')=\tfrac{1}{3}[\mu_0'\lambda^2+3\mu_2']=\frac{\lambda^2}{3}+\mu_2'$$

The expected variance, then, of $f_n(x)$ is

$$E(m_2'-(m_1')^2)=\frac{\lambda^2}{3}+\sigma^2$$

where σ^2 is the variance of the underlying distribution.

6. Computing empirical density functions

As is frequently the case, special problems arise when empirical density functions are actually being computed for real data. In this section we consider these problems, and illustrate the technique by two Monte Carlo examples and one practical example.

Suppose that the sample is $X_1, \ldots, X_n$. Since each X_i has a rectangle centred at X_i which contributes to $g_n(x)$, then there are (no more than) $2n$ discontinuity points in $g_n(x)$. The $2n$ points are then ordered as $X_1-\lambda$, $X_1+\lambda$, $X_2-\lambda, \ldots, X_n+\lambda$. Then, between any two adjacent points, the height of $g_n(x)$ is constant and proportional to the number of X_i's within a distance λ from x. This property makes $g_n(x)$ easy to plot.

The main problem, then, is to find a good first guess for $\lambda=\lambda_n$. In § 3 we found that

$$\lambda^{(1)}=(2K(\alpha)/C)^{1/2}n^{-1/4}$$

where $K(\alpha)$ is a constant from the Kolmogorov–Smirnov statistic and C is an upper bound for $|f'(x)|$. For a value of C, we consider the normal density

$$f(x)=\frac{1}{\sqrt{(2\pi)}\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

and by elementary calculus we obtain

$$|f(x)|\leq(\sqrt{(2\pi)}e\sigma^2)^{-1}$$

If $\alpha=0{\cdot}05$, then $K(\alpha)=1{\cdot}36$ and

$$\lambda^{(1)}=3{\cdot}35\sigma n^{-1/4}$$

Of course, in practice one can let

$$\lambda^{(1)}=3{\cdot}35sn^{-1/4}$$

where $s=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the usual estimate of σ. However, using s to estimate σ is very sensitive to outliers. We therefore use the quartile deviation

$$\text{qd}=(\xi_{0{\cdot}75}-\xi_{0{\cdot}25})/2$$

where ξ_α is a number such that $100\alpha\%$ of the data lie below. An elementary calculation with the normal density used above shows the average value of qd to be $0{\cdot}676\sigma$. Therefore we modify $\lambda^{(1)}$ to

$$\lambda^{(1)}=3{\cdot}35\,\frac{\text{qd}}{0{\cdot}676}\,n^{-1/4}$$

$$=4{\cdot}96\ \text{qd}\ n^{-1/4}$$

In figure 1, we illustrate our techniques with a sample of 15 normal random variables with mean 0 and variance 1. These were generated by a standard random variable generator. Figure 1 (*a*) has λ smaller than $\lambda^{(1)}$ and hence is a very uneven plot. Notice, however, that a small value of λ allows one to look at the structure of the data itself in some detail. Figure 1 (*b*) uses $\lambda=\lambda^{(1)}$ and is a good estimate of the underlying density. Finally, figure 1 (*c*) uses a large value of λ and consequently has a very flat plot.

In figure 2, we consider a Monte Carlo sample of 50 exponential random variables with density

$$f(x)=\begin{cases}\exp(-x), & x\geq 0\\ 0, & x<0\end{cases}$$

Figure 2 (*a*) uses $\lambda^{(1)}$ without the information that $P(X\geq 0)=1$. Note that the estimate suffers from the fact that $g_{50}(x)>0$ for $x<0$. By using the estimator $g_{50}{}^{*}(x)$ in § 4, figure 2 (*b*) is produced and, as can be seen, gives a much better estimate of $f(x)$.

Part of the National Uranium Resource Evaluation programme [11], is an aerial radiometric reconnaissance in which gamma radiation intensities are measured by instruments on a low flying aircraft. The aircraft flies transects or map lines which are 50 to 100 miles long and approximately five miles apart. The problem arose of producing a graphical representation of the data for 100 consecutive map lines in a region near the Texas gulf coast.

An empirical density function was computed for each map line according to the above techniques and, to present the data in one picture, map lines are given on one axis while counts per second are given on the other. These 100 empirical density functions appear in figure 3 as a three-dimensional plot.

Since many properties are obscured by higher densities on earlier map lines, the information contained in figure 3 (*a*) is represented in figure 3 (*b*), as a lightness–darkness plot. Darker points correspond to points of higher density, and hence the representation is two-dimensional.

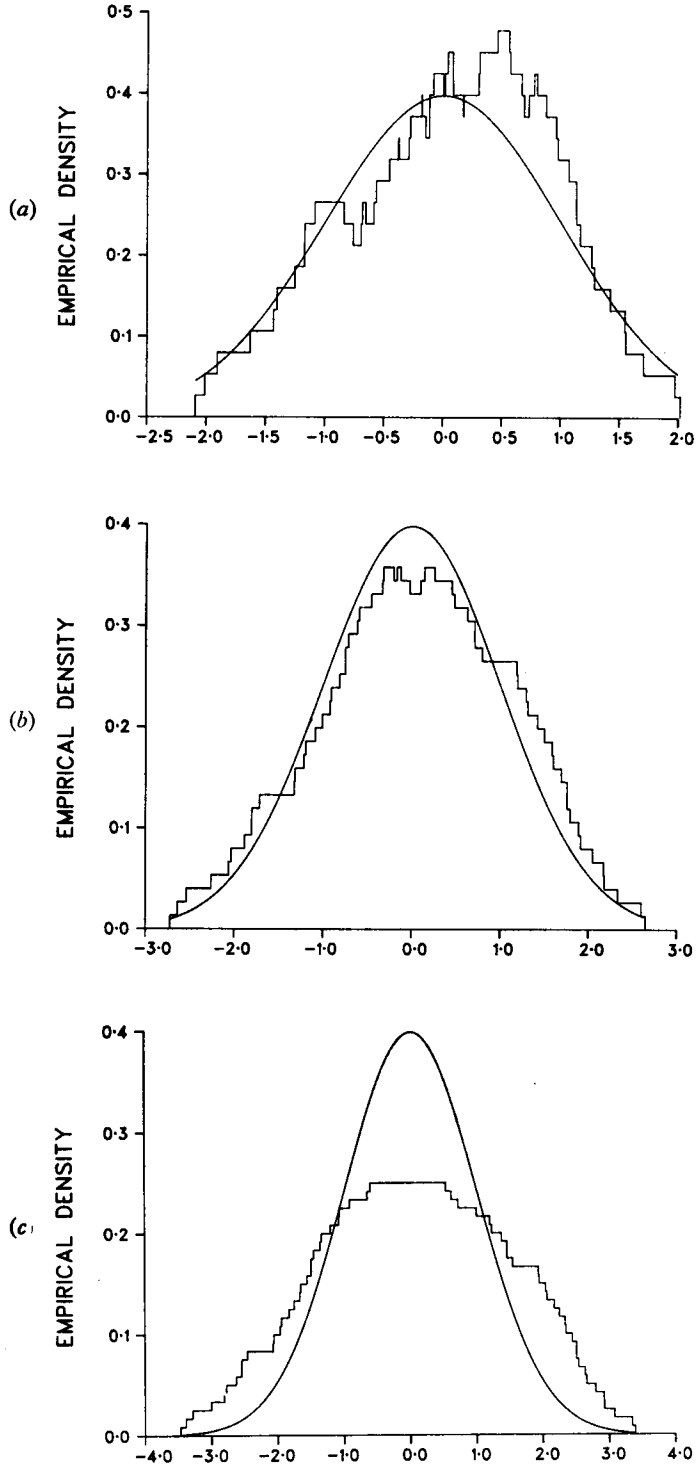


Figure 1. Empirical density functions for a sample of size 30 from a standard normal population. (*a*) $\lambda = 0{\cdot}625 < \lambda^{(1)}$. (*b*) $\lambda^{(1)} = 1{\cdot}26$. (*c*) $\lambda = 2{\cdot}0 > \lambda^{(1)}$.

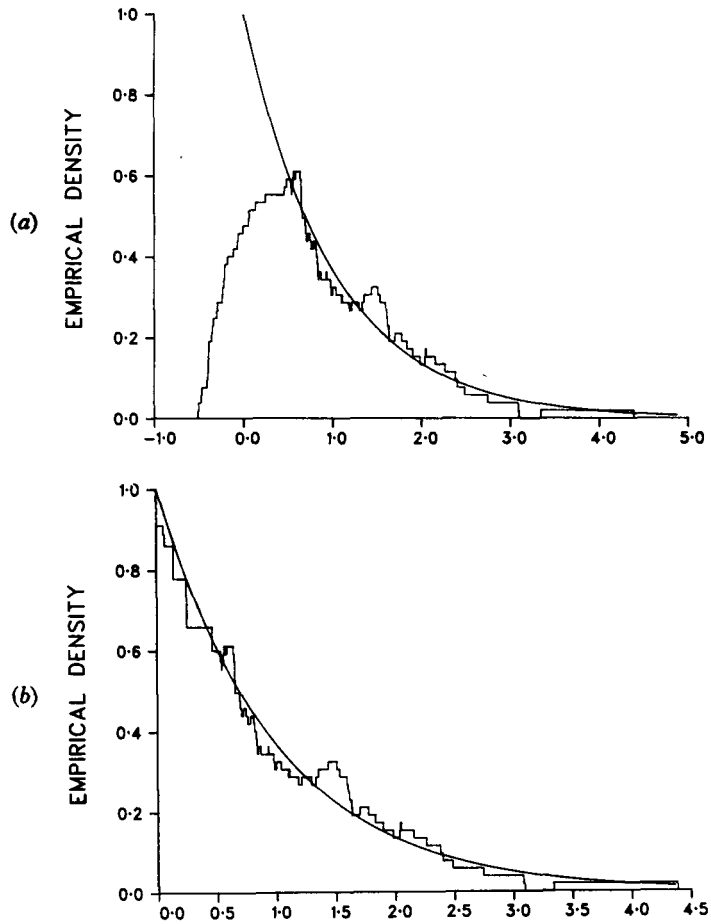


Figure 2. Empirical density functions for a sample of size 50 from an exponential population with mean 1. (*a*) No use of $X \geqslant 0$. (*b*) Use of $X \geqslant 0$. (See § 4.)

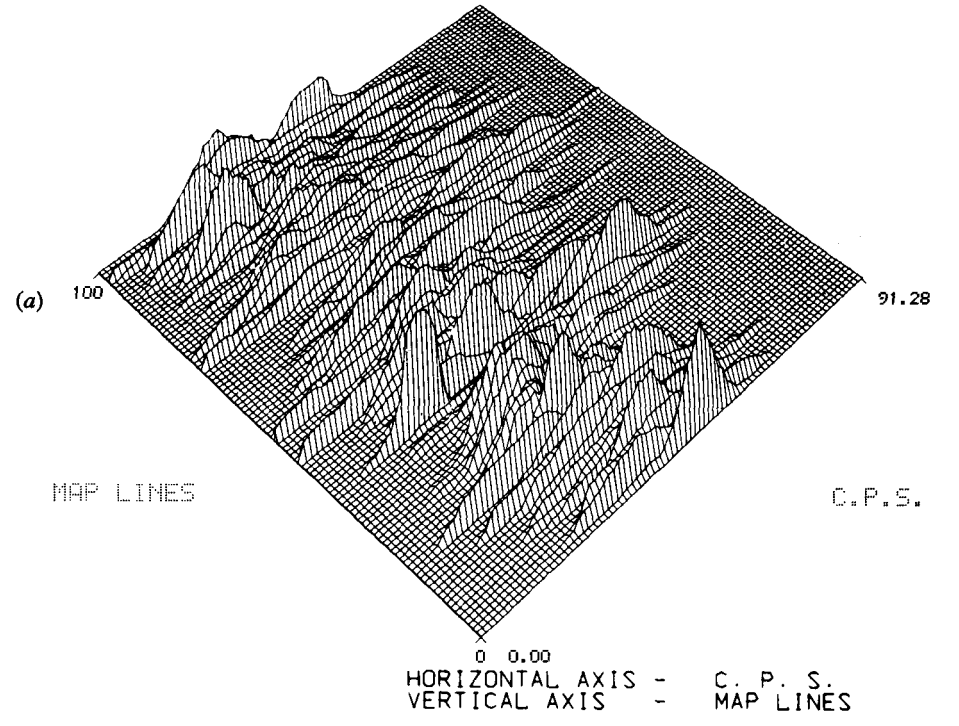
7. Conclusion

We have demonstrated the usefulness of density estimation, both at various levels of statistical courses and in practical situations. The utilization of density estimation in practice will not become widespread until it is commonly taught in courses.

Of course, empirical density functions are only the simplest of a large class of density estimators. The general estimator [12] has the form

$$g_n(x) = \{n\lambda_n\}^{-1} \sum_{i=1}^{n} w\left(\frac{x - X_i}{\lambda_n}\right)$$

where $w(\cdot)$ is called the kernel. It is clearly possible to use smooth kernels such as normal densities, which allows very smooth estimates $g_n(\cdot)$. However, the convergence properties do not seem to depend on $w(\cdot)$ [3], and we chose to work with the empirical density functions for this reason and because of their probabilistic simplicity.



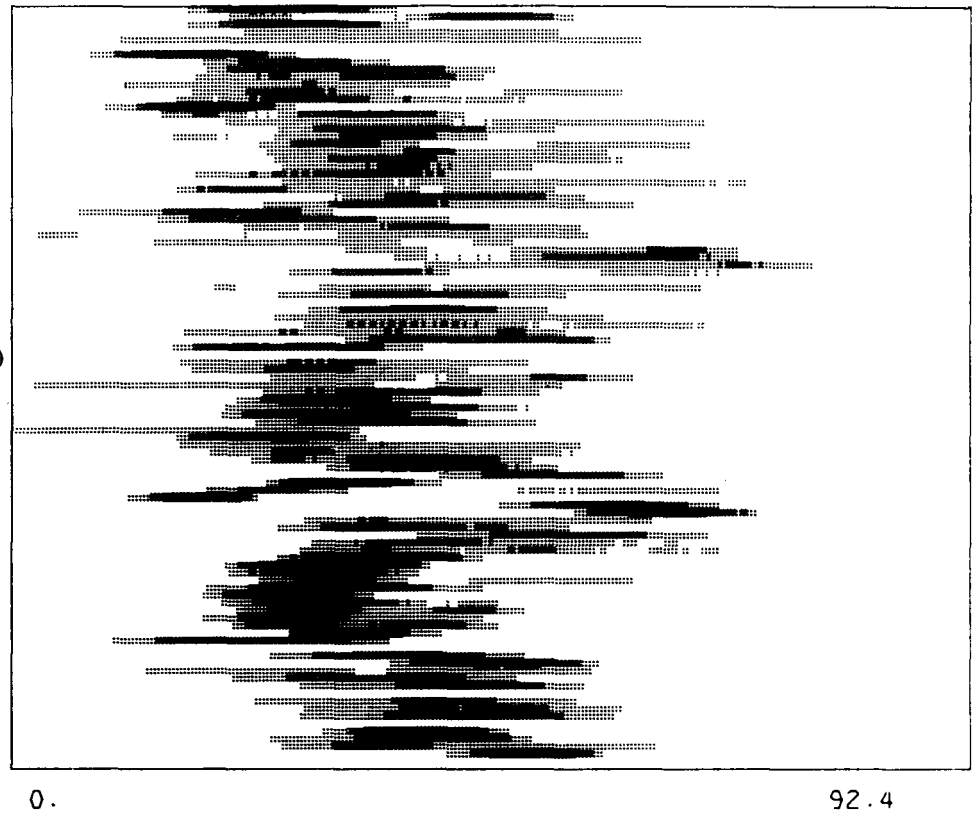


Figure 3. Empirical density functions for gamma counts per second for 100 adjacent map lines. (*a*) Three-dimensional representation. (*b*) Lightness–darkness representation.

Acknowledgments

The first author expresses his thanks to Professor T. F. Smith of Marquette, Michigan, U.S.A. who pointed out a form of density estimation [13] and initiated our interest in these topics. Thanks are also due to James Beach of Pocatello, Idaho, U.S.A. who provided assistance in the early stages of this study.

References

[1] STURGES, H. A., 1926, *J. Am. statist. Ass.*, **21**, 65.
[2] DOANE, D. P., 1976, *Am. Statistn.*, **30**, 181.
[3] WEGMAN, E. J., 1972, *Technometrics*, **14**, 533.
[4] WEGMAN, E. J., 1972, *J. Statist. Comput. Simul.*, **1**, 225.
[5] HOGG, R. V., and CRAIG, A. T., 1970, *Introduction to Mathematical Statistics*, third edition (London: Macmillan).
[6] LINDGREN, B. W., 1968, *Statistical Theory*, second edition (London: Macmillan).
[7] MORAN, P. A. P., 1968, *Introduction to Probability Theory* (Oxford: Clarendon Press).
[8] RÉVÉSZ, P., 1972, *Period. Math. Hung.*, **2**, 85.
[9] ROSENBLATT, M., 1956, *Ann. math. Statist.*, **27**, 832.
[10] PARZEN, E., 1962, *Ann. math. Statist.*, **33**, 1065.
[11] ZEIGLER, R. K., WHITEMAN, D. E., WATERMAN, M. S., and BEMENT, T. R., 1976, *Proceedings of the Second ERDA Statistical Symposium.* (Los Alamos Scientific Laboratory Publication, LA–6758–C), p. 212.
[12] TARTER, M. E., and KRONMAL, R. A., 1976, *Am. Statistn*, **30**, 105.
[13] SMITH, T. F., and SADLER, J. R., 1971, *J. molec. Biol.*, **57**, 273.