# AN APPROXIMATION TO THE DENSITY FUNCTION

By HIROTUGU AKAIKE

(Received Aug. 30, 1954)

## 1. Introduction

In many practical analyses of statistical data from continuous distributions, we use histograms. From the histogram we can get approximate values of probabilities that the data we observe fall in some preassigned class-intervals. We usually treat histograms from this point of view. This point of view is theoretically reasonable, and the confusion between the histogram and the density function will never take place.

There are, however, problems in which we want to use directly density functions like that of determining some domain by the likelihood-ratio criterion. In such cases we often use the histogram as an approximate image of the density function, or we first obtain informations about the density function from the histogram. Concerning this problem consider the next one. Let the data we observe be from the two-dimensional population $\pi_1$ or $\pi_2$ with density functions $f_1(x, y)$, $f_2(x, y)$, respectively, and let the a priori probabilities with which the data are taken from $\pi_1$ or $\pi_2$ be $P_1$ and $P_2$, respectively. Then, if we want to classify the observed data as from $\pi_1$ or $\pi_2$, we consider the data falling into

$$S_1=\{(x, y);\ P_1\cdot f_1(x, y)\geq P_2\cdot f_2(x, y)\}$$

to be from $\pi_1$ and those falling into

$$S_2=\{(x, y);\ P_1\cdot f_1(x, y)< P_2\cdot f_2(x, y)\}$$

to be from $\pi_2$.

As is well known, this procedure assures us the maximum rate of success which is given by

$$P_1\cdot\iint_{S_1} f_1(x, y)\,dx\,dy+P_2\cdot\iint_{S_2} f_2(x, y)\,dx\,dy.$$

Now, if $f_1(x, y)$ and $f_2(x, y)$ are unknown but $P_1$ and $P_2$ are known, and if we want to apply, formally, the above idea of obtaining the maximum rate of success to this case, we may take

$$S_1^* = \{(x_i^{(1)}, y_i^{(1)});\ i=1.2, \cdots n\}$$

and

$$S_2^* = \{(x_i^{(2)}, y_i^{(2)});\ i=1.2, \cdots m\}$$

as approximations to $S_1$ and $S_2$, where $(x_i^{(k)} y_i^{(k)})$s are the formerly obtained data from $\pi_k (k=1, 2)$. Then, we almost always have $S_1^* \cap S_2^* = \phi$ and

$$P_1 \times \frac{n}{n} + P_2 \times \frac{m}{m} = 1 \quad \text{(visional rate of success)}.$$

But, as is obvious,

$$P_1 \iint_{S_1^*} f_1(x, y)\, dx\, dy + P_2 \iint_{S_2^*} f_2(x, y)\, dx\, dy = 0$$

which shows that $S_1^*$ and $S_2^*$ are of no use for our purpose. To avoid such circumstances, we usually divide the whole space into cells or class-intervals and arrive at the idea of histogram.

In considering the histogram, however, arises the problem to determine the sizes of these cells. When we make the sizes of cells larger, we get the more reliable but dull results. When we make the sizes of cells smaller, we get the more accurate but unreliable results.

In this paper, we shall give an approximation to the density function not by histogram but directly. Through it the problem of optimum sizes of the cells in making histogram will be considered.

## 2. ε-Approximation to the density function

We shall restrict our consideration to one-dimensional space $R$ with Lebesgue measure $m$ on it, but the results will easily be extended to a space of any dimension.

Given a random sample $(x_1, x_2, \cdots, x_N)$ of size $N$ from the population with the bounded density function $f(x)$ in respect to $m$, we define an ε-approximate density function $\hat{f}_{N,\varepsilon}(x)$ by the following formula. Put

$$U_\varepsilon(x) \equiv [x-\varepsilon, x+\varepsilon) = \{x';\ x-\varepsilon \leq x' < x+\varepsilon\}$$

$$d_{N,\varepsilon}(x) \equiv \text{number of } x_i s \text{ such that } x_i \in U_\varepsilon(x).$$

Then the ε-approximate density function $\hat{f}_{N,\varepsilon}(x)$ is given by

$$\hat{f}_{N,\varepsilon}(x) = \frac{d_{N,\varepsilon}(x)}{N \cdot m(U_\varepsilon)}$$

where $m(U_\varepsilon) \equiv 2\varepsilon$. Our $\hat{f}_{N,\varepsilon}(x)$ is defined for all $x$ in $R$. The height of the ordinary histogram at the center of every class-interval with width $2\varepsilon$

just coincides with the value of $N\cdot m(U_\varepsilon)\cdot f_{N\cdot\varepsilon}(x)$ at that point. Therefore, the ordinary histogram can be considered as an incomplete graph of $\hat{f}_{N\cdot\varepsilon}(x)$.

Concerning $\hat{f}_{N\cdot\varepsilon}(x)$, we have

$$\int_R \hat{f}_{N\cdot\varepsilon}(x)\,dm=1\,,$$

$$E\hat{f}_{N\cdot\varepsilon}(x)=\frac{P_r\{U_\varepsilon(x)\}}{m(U_\varepsilon)} \quad \text{where } P_r\{U_\varepsilon(x)\}=\int_{U_\varepsilon(x)} f(t)\,dm,$$

$$E(\hat{f}_{N\cdot\varepsilon}(x)-f(x))^2=\frac{P_r\{U_\varepsilon(x)\}\cdot[1-P_r\{U_\varepsilon(x)\}]}{N\cdot\{m(U_\varepsilon)\}^2}+\frac{[f(x)\cdot m(U_\varepsilon)-P_r\{U_\varepsilon(x)\}]^2}{\{m(U_\varepsilon)\}^2},$$

for it holds that

$$\int_R \hat{f}_{N\cdot\varepsilon}(x)\,dm=\int_{U_\varepsilon(0)}\sum_{k=-\infty}^{+\infty}\hat{f}_{N\cdot\varepsilon}(x+2k\varepsilon)\cdot dm$$

$$=\int_{U_\varepsilon(0)}\frac{1}{m(U_\varepsilon)}\cdot dm=\frac{2\varepsilon}{2\varepsilon}=1$$

and $d_{N\cdot\varepsilon}(x)$ has the binomial distribution such that

$$P_r\{d_{N\cdot\varepsilon}(x)=k\}={}_NC_k\cdot[P_r\{U_\varepsilon(x)\}]^k\cdot[1-P_r\{U_\varepsilon(x)\}]^{N-k}$$

(=probability that $x$ is covered by just $k$ $U_\varepsilon(x_i)$s).

Then, it is seen that for almost all $x$

$$\lim_{\varepsilon\to 0} E\hat{f}_{N\cdot\varepsilon}(x)=f(x)\,,$$

$$\lim_{\varepsilon\to 0} E(\hat{f}_{N\cdot\varepsilon}(x)-f(x))^2=+\infty \quad \text{for } x \text{ where } f(x)>0\,,$$

$$\lim_{N\to+\infty} E(\hat{f}_{N\cdot\varepsilon}(x)-f(x))^2=\frac{[f(x)\cdot m(U_\varepsilon)-P_r\{U_\varepsilon(x)\}]^2}{\{m(U_\varepsilon)\}^2}\,.$$

Now, the problem of determining the best $\varepsilon$ may be treated as follows.

Taking some appropriate weight function $w(x)$, we calculate

$$D_N(\varepsilon)=\int_R E(\hat{f}_{N\cdot\varepsilon}(x)-f(x))^2\cdot w(x)\cdot dm$$

and define $\varepsilon$ which minimizes $D_N(\varepsilon)$ as the best for the $N$. As the weight function, $w(x)\equiv 1$ or $w(x)\equiv f(x)$ may be considered. If we can evaluate

$$\rho_N(x)=\int_R E\sqrt{\hat{f}_{N\cdot\varepsilon}(x)}\cdot\sqrt{f(x)}\cdot dm^{*)}\,,$$

*) Concerning the quantity $\rho_N(\varepsilon)$, see K. Matusita, On the theory of statistical decision functions, *Ann. Inst. Stat. Math.* Vol. III, 1951, K. Matusita and H. Akaike, Note on the decision problem, *Ann. Inst. Stat. Math.* Vol. IV, 1951.

$\varepsilon$ maximizing $\rho_N(\varepsilon)$ is desirable, but for the present $\rho_N(\varepsilon)$ seems not easy to calculate.

## 3. Examples

Applying the above minimum-$D_N(\varepsilon)$ method to some types of $f(x)$ and $w(x)$, we get the following relations between $N$ and the best $\varepsilon$.

I. In case $f(x)=e^{-x}$ for $x \geq 0$, $f(x)=0$ for $x<0$. and $w(x)=1$, we have

$$N_{\mathrm{I}}=\frac{-1+e^{-2\varepsilon}+\varepsilon e^{-3\varepsilon}}{-1+e^{-2\varepsilon}+\varepsilon(-1+2e^{-\varepsilon}+e^{-2\varepsilon})+2\varepsilon^2\cdot e^{-\varepsilon}} .$$

II. In case $f(x)=e^{-x}$ for $x \geq 0$, $f(x)=0$ for $x<0$, and $w(x)=f(x)$, we have

$$N_{\mathrm{II}}=\frac{-4e^{-\varepsilon}+2e^{-2\varepsilon}+2e^{-3\varepsilon}+\varepsilon(-2e^{-\varepsilon}+2e^{-2\varepsilon}+3e^{-3\varepsilon})}{6-10e^{-\varepsilon}+2e^{-2\varepsilon}+2e^{-3\varepsilon}+\varepsilon(-6-e^{-\varepsilon}+4e^{-2\varepsilon}+3e^{-3\varepsilon})+\varepsilon^2(4e^{-\varepsilon}+4e^{-2\varepsilon})} .$$

III. In case $f(x)=\frac{1}{2}e^{-|x|}$ and $w(x)=1$, we have

$$N_{\mathrm{III}}=\frac{-3+3e^{-2\varepsilon}+4\varepsilon e^{-2\varepsilon}+2\varepsilon^2 e^{-2\varepsilon}}{-3+3e^{-2\varepsilon}+\varepsilon(-2+4e^{-\varepsilon}+4e^{-2\varepsilon})+\varepsilon^2(4e^{-\varepsilon}+2e^{-2\varepsilon})+2\varepsilon^3 e^{-\varepsilon}} .$$

IV. In case $f(x)=\frac{1}{2}e^{-|x|}$ and $w(x)=f(x)$, we have

$$N_{\mathrm{IV}}=\frac{-2e^{-\varepsilon}+4e^{-2\varepsilon}-2e^{-3\varepsilon}+\varepsilon(-4e^{-\varepsilon}+4e^{-2\varepsilon}-3e^{-3\varepsilon})-3\varepsilon^2 e^{-\varepsilon}}{-12+6\varepsilon+14e^{-\varepsilon}-4e^{-2\varepsilon}+2e^{-3\varepsilon}+\varepsilon(5e^{-\varepsilon}-2e^{-2\varepsilon}+3^{-3\varepsilon})+\varepsilon^2(-2e^{-\varepsilon}+4e^{-2\varepsilon})} .$$

V. In case $f(x)=\frac{1}{a}$ for $x \in [0, a]$, $f(x)=0$ for $x \notin [0, a]$, and $w(x)=1$, we have

$$N_{\mathrm{V}}=\frac{3}{2}\cdot\frac{a^2}{\varepsilon^2}-2 .$$

VI. In case $f(x)=\frac{1}{a}$ for $x \in [0, a]$, $f(x)=0$ for $x \notin [0, a]$, and $w(x)=f(x)$, we have

$$N_{\mathrm{VI}}=3\cdot\frac{a^2}{\varepsilon^2}-5 .$$

Numerical values of $\varepsilon$ and $N_{\mathrm{I}}, N_{\mathrm{II}}, \cdots, N_{\mathrm{VI}}$ are given below

| $\varepsilon$ | $\frac{2\varepsilon}{L}$ * | $N_{\mathrm{I}}$ | $N_{\mathrm{II}}$ |
|---|---|---|---|
| 0.05 | 0.043 | 1186 | 1173 |
| 0.07 | 0.061 | 610 | 569 |
| 0.10 | 0.087 | 300 | 275 |
| 0.15 | 0.130 | 133 | 117 |
| 0.20 | 0.174 | 75 | 63 |
| 0.30 | 0.261 | 33 | 26 |

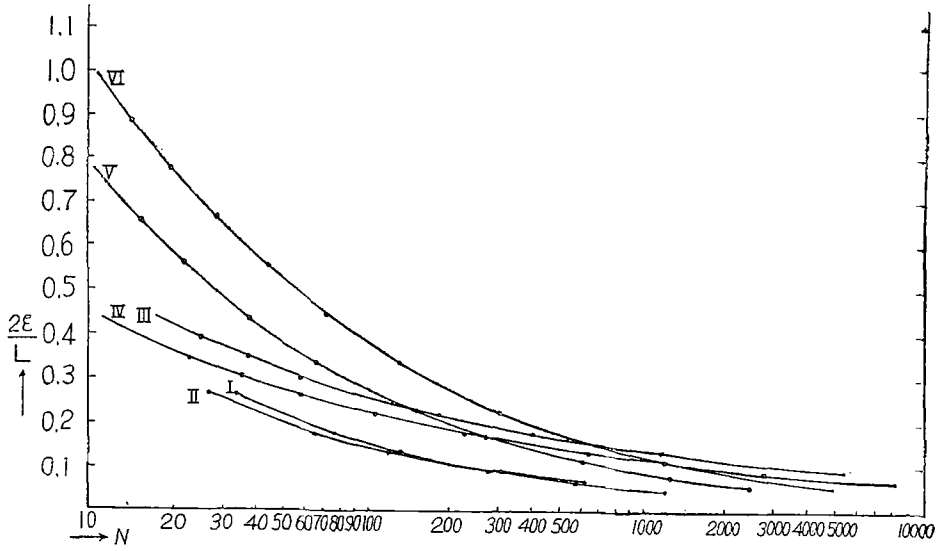
| $\varepsilon$ | $\frac{2\varepsilon}{L}$ * | $N_{\mathrm{III}}$ | $N_{\mathrm{IV}}$ |
|---|---|---|---|
| 0.15 | 0.065 | 15000** | 8050** |
| 0.20 | 0.087 | 5200** | 2720** |
| 0.30 | 0.130 | 1139 | 630 |
| 0.40 | 0.174 | 398 | 226 |
| 0.50 | 0.217 | 180 | 104 |
| 0.60 | 0.261 | 95 | 56 |
| 0.70 | 0.304 | 56 | 34 |
| 0.80 | 0.347 | 36 | 22 |
| 0.90 | 0.391 | 24 | 15 |
| 1.00 | 0.434 | 17 | 11 |

**; significant only for two figures

| $\frac{2\varepsilon}{a}$ | $\frac{2\varepsilon}{L}$ * | $N_{\mathrm{V}}$ | $N_{\mathrm{VI}}$ |
|---|---|---|---|
| 0.05 | 0.056 | 2398 | 4795 |
| 0.07 | 0.078 | 1222 | 2444 |
| 0.10 | 0.111 | 598 | 1195 |
| 0.15 | 0.167 | 265 | 528 |
| 0.20 | 0.222 | 148 | 295 |
| 0.30 | 0.333 | 65 | 128 |
| 0.40 | 0.444 | 36 | 70 |
| 0.50 | 0.556 | 22 | 43 |
| 0.60 | 0.667 | 15 | 28 |
| 0.70 | 0.778 | 10 | 19 |
| 0.80 | 0.889 | 7 | 14 |
| 0.90 | 1.000 | 5 | 10 |
| 1.00 | 1.111 | 4 | 7 |

*; $L=\begin{cases} e^{-L}=0.1 \text{ or } L=2.303 & \text{for I II} \\ 2\times 2.303 & \text{for III IV} \\ 0.9a & \text{for V VI} \end{cases}$

These relations between $N$ and $\varepsilon$ are graphically represented below. From this graph, it can be seen that the usual procedure of taking about from 10 to 20 class-intervals for histograms based on samples of about 500 or more may be considered reasonable. Moreover, taking into account the fact that the sensibility of the best $\varepsilon$ to $w(x)$ is rather low, it can also be seen that our approach to the density function has thrown some light on practical procedures in statistical analyses.




THE INSTITUTE OF STATISTICAL MATHEMATICS