KM 38 (1991)
pg.5-16

AN IMPROVED FREQUENCY POLYGON

by

Pieter H.F.M. van Casteren

ABSTRACT

The frequency polygon, as it is commonly constructed, is not compatible with the frequency distribution from which it is derived. As a result it is too flat. Therefore a more accurate frequency polygon is constructed and the corresponding ogive is constructed as well.

KEY WORDS: frequency distribution, polygon, ogive.

1. Introduction

Suppose a number of observations on a certain variable is given. It may then be useful to make a diagram in order to get a visual impression of the distribution of these observations. Many types of diagrams may serve this purpose, such as the quantile plot, scatter plot, box plot, stem and leaf diagram, histogram, frequency polygon, density trace, ogive (see e.g. Chambers et al. 1983 Ch.2, Schmid 1983 Ch.4). This paper concentrates on the histogram and the polygon. Both are representations of a frequency distribution (e.g. table 1), hence it is not necessary to know the exact observed values.

The histogram is explained in most statistical textbooks (e.g. Harnett 1982) and it is incorporated in many statistical packages (see Ootjers 1987). An example is given in table 1 and figure 1. The histogram is discontinuous at the class bounds, so if a continuous representation is preferred, another kind

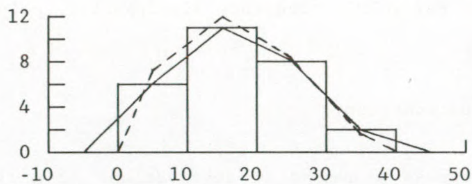
Drs. P.H.F.M. van Casteren
Vakgroep Kwantitatieve Methoden, Universiteit van Amsterdam
Jodenbreestraat 23, 1011 NH Amsterdam
Tel. 020 - 5254232

of graph is needed. This may especially be the case if the random variable is continuous, and if it is assumed to have a continuous density function. Continuous representations may also be preferable to histograms, when two or more distributions are to be compared on the same chart, because the overlaying of several histograms may hamper recognition of the individual histograms. The frequency polygon is such a continuous representation of a frequency distribution. It is commonly constructed by connecting the consecutive midpoints of the tops of the bars of the histogram by straight lines (see figure 1). In section 2 this *common* polygon, is reviewed and it is argued that it is inappropriate. Hence, section 3 provides an *alternative* polygon and section 4 presents the corresponding ogive. Section 5 comments briefly on the estimation of a probability density function.

Table 1. Frequency distribution, example 1.

class	frequency
0-<10	6
10-<20	11
20-<30	8
30-<40	2

Figure 1. Frequency histogram and common (——) and alternative (---) polygon, example 1.



2. The common frequency polygon

Anticipating the analysis to follow, an exact description of the construction of a histogram and a common polygon for the general case of nonequal class widths is given below. Suppose that for some random variable **x** a frequency distribution is given, with $c \geq 2$ the number of classes, u_i the upperbound of class i for **x** such that $u_i < u_{i+1}$ for all $i=1,...,c-1$, u_0 the lowerbound of class 1, and f_i the frequency of class i, with $f_1 > 0$ and $f_c > 0$, while $f_1 + ... + f_c$ is equal to n (the number of observations) or 1 (if the frequencies are relative). Then the class midpoints $m_i = ½(u_{i-1} + u_i)$ and the class widths $w_i = u_i - u_{i-1}$ can be calculated immediately. It is required to choose a standard class width, say w_s (often chosen equal to the predominant class width or equal to one), so that for each class i the average frequency density $d_i = w_s f_i / w_i$ (interpreted as the amount of frequency per standard class width) can be calculated as well. If all the class widths are equal, say $w_i = w$

for all $i=1,\dots,c$, and $w_s=w$ is chosen, then $d_i=f_i$.

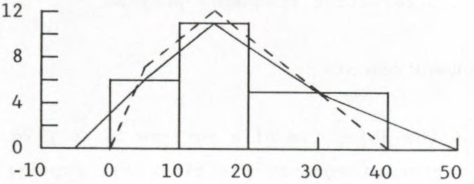
The histogram is now constructed by drawing a bar between u_{i-1} and u_i with height d_i for each class i. See table 2 and figure 2. The surfaces of the bars (not the heights) are associated with frequencies, see Velleman and Hoaglin (1981 p.258), Schmid (1983 p.71), Ootjers (1983 p.32). Thus, as the surface of the i-th bar $d_i w_i = f_i w_s$ is proportional to f_i, the histogram is *compatible* with the given frequencies. Note that w_s can be chosen as to achieve any preferred positive value for the total surface of the histogram $(f_1+\dots+f_c)w_s$. The histogram generates a frequency density function, which is obtained by taking it's height as a function of x.

The common polygon is drawn as a linear spline function, connecting consecutively the points (m_0,d_0), $(m_1,d_1),\dots,(m_c,d_c)$, (m_{c+1},d_{c+1}), where $m_0=u_0-\frac{1}{2}w_1$, $m_{c+1}=u_c+\frac{1}{2}w_c$ and $d_0=d_{c+1}=0$ (see figure 2). The first and the last point in this series are added in order to make the polygon a continuous function of x. Their positions on the x-axis can be considered the midpoints of two imaginary classes (i=0,c+1) with zero frequency and width w_1 and w_c respectively. Their position is rather disputable in the case of unequal class widths (an alternative would be to choose both imaginary class widths equal to the average class width).

Table 2. Frequency distribution, example 2 (choosing $w_s=10$).

class	f_i	d_i
0-<10	6	6
10-<20	11	11
20-<40	10	5

Figure 2. Frequency histogram and common (——) and alternative (---) polygon, example 2.



There is however one fundamental objection against the common polygon: it is not constructed such that the surface under the polygon within each class is proportional to the frequency of that class, which means it is not constructed to be compatible with the given frequency distribution. One can see this by comparing the surfaces under the polygon with those of the histogram. In particular, the surface under the polygon is too low within the peak classes (10-20 in figures 1 and 2) and too high in the imaginary classes directly below class 1 (<0) and above class c (>40). This means that the common polygon is too flat and therefore it is not a proper representation of the given frequency distribution. A related problem is that the common polygon may allocate positive surface to imaginary classes, consisting of impossible values of **x** (this happens in figures 1 and 2, if **x** is a nonnegative variable).

In fact a disaccord between the surface under the polygon and that of the histogram may arise within each class $i=0,\ldots,c+1$ (see also Yule and Kendall 1950, p.78-80). This can be explained as follows. The polygon attributes the function value d_i to $x=m_i$ for $i=0,\ldots,c+1$ and then interpolates between the successive class midpoints. If as a result the polygon is concave in class i (e.g. peak classes), the average height of the polygon in class i is smaller than d_i, so that the surface of class i is too small compared to the histogram. If the polygon is convex in class i (e.g. $i=0,c+1$), the average height in class i is larger than d_i, so that the surface of class i is too large. Hence, in order to get the surface of class i in accord with the histogram, one could draw a similar polygon, but with a larger value instead of d_i if the common polygon is concave in class i and with a smaller value instead of d_i if the common polygon is convex in class i. This basic idea is effectuated by the alternative polygon presented in the next section. Since Yule and Kendall (1950 p.80) and Schmid (1983 p.71) preferred using the histogram rather than the common polygon, because of the latter's incompatibility, this alternative might be a welcome new option.

3. An alternative frequency polygon

(a) Requirements

As the objective of a polygon is to give a proper continuous representation of a given frequency distribution, a polygon should be: (1) *continuous*; (2) *nonnegative*; (3) *compatible* with the given frequency distribution, that is the surface under the polygon within each class should be proportional to the frequency of that class. The compatibility requirement is included to avoid misrepresentation of the given frequencies. As this rules out the common polygon, the alternative is designed to satisfy all three requirements.

(b) The basic procedure

Suppose a frequency distribution, as specified in section 2, is given. In principle the alternative polygon is drawn as a linear spline function, connecting consecutively the points $(u_0,0)$, $(m_1,a_1),\ldots,(m_c,a_c)$, $(u_c,0)$, where the values of $a_1,\ldots,a_c$ are yet to be determined such that, choosing the standard class width w_s, the surface under the alternative polygon within each class i equals the corresponding surface under the histogram $f_i w_s$. Because each surface $f_i w_s$ is proportional to f_i, this will satisfy the compatibility requirement. Writing b_i for the function value of the polygon in $x=u_i$ for $i=0,\ldots,c$ (see figure 3), which means that $b_0=0$ and $b_c=0$ are known, while $b_i=(w_{i+1}a_i+w_i a_{i+1})/(w_{i+1}+w_i)$ for $i=1,\ldots,c-1$ depend on $a_1,\ldots,a_c$, the values of $a_1,\ldots,a_c$ are to be determined such that

$$\tfrac{1}{4}(b_{i-1}+2a_i+b_i)w_i=f_i w_s \quad \text{for } i=1,\ldots,c.$$

Substituting b_{i-1}, b_i and f_i this is equivalent to

$$\begin{cases} \left[2+\frac{w_2}{w_2+w_1}\right]a_1+\left[\frac{w_1}{w_2+w_1}\right]a_2 = 4d_1 \\ \left[\frac{w_i}{w_i+w_{i-1}}\right]a_{i-1}+\left[\frac{w_{i-1}}{w_i+w_{i-1}}+2+\frac{w_{i+1}}{w_{i+1}+w_i}\right]a_i+\left[\frac{w_i}{w_{i+1}+w_i}\right]a_{i+1} = 4d_i \quad \text{for } i=2,\ldots,c-1 \\ \left[\frac{w_c}{w_c+w_{c-1}}\right]a_{c-1}+\left[\frac{w_{c-1}}{w_c+w_{c-1}}+2\right]a_c = 4d_c, \end{cases}$$

which can be written in the matrixform

$$\begin{bmatrix} q_1 & r_1 & 0 & 0 & \ldots & & & \\ p_2 & q_2 & r_2 & 0 & \ldots & & & \\ & & & & \ldots & & & \\ & & & & \ldots & p_{c-1} & q_{c-1} & r_{c-1} \\ & & & & \ldots & 0 & p_c & q_c \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_{c-1} \\ a_c \end{bmatrix} = \begin{bmatrix} 4d_1 \\ 4d_2 \\ \ldots \\ 4d_{c-1} \\ 4d_c \end{bmatrix}$$

with obvious definitions of p_i, q_i and r_i. There is always a solution for $(a_1,\ldots,a_c)$, which can be computed by an algorithm which exploits the tridiagonal form of the matrix (see Press, et al. 1988, section 2.8). The algorithm is always successful, because $|q_i|>|p_i|+|r_i|$ for all $i=1,\ldots,c$. In the special case of equal class widths $w_i=w$ for $i=1,\ldots,c$ the system becomes

$$\begin{cases} 5a_1 + a_2 = 8d_1 \\ a_{i-1} + 6a_i + a_{i+1} = 8d_i \quad \text{for } i=2,\ldots,c-1 \\ a_{c-1} + 5a_c = 8d_c. \end{cases}$$

Examples of the alternative frequency polygon are given in figure 1 ($a_1=7.19$, $a_2=12.07$, $a_3=8.40$, $a_4=1.52$) and figure 2 ($a_1=7.15$, $a_2=12.23$, $a_3=5.08$).

Figure 3. The alternative polygon within class i.

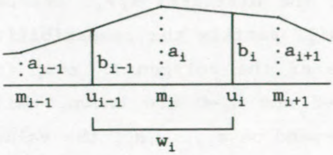
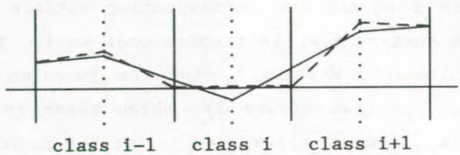


Figure 4. The alternative polygon before (——) and after (---) the first step of adjustment.



(c) Adjustments

The procedure described above yields a polygon that satisfies the continuity and compatibility requirement. However, the nonnegativity requirement is violated, if $a_i<0$ for some $i\in\{1,\ldots,c\}$. Roughly speaking this occurs if f_i is small relative to $f_{i-1} + f_{i+1}$. Such cases are presumably rare, but require an adjustment of the basic procedure given above. If possible, choosing a different set of class bounds (enlarging the classes with $a_i<0$) might solve the problem. Otherwise, a simple adjustment is to replace each negative value of a_i $(i=1,\ldots,c)$ by zero. This will obviously violate the compatibility requirement, but presuming that shifts of the alternative polygon due to this adjustment will be relatively rare and small in practice, it can be considered an acceptable practical solution. Nevertheless, the ideal solution would be an adjustment that respects the given class bounds and complies with all three requirements. Such an adjustment is described below.

This more complex adjustment starts after the values of $a_1,\ldots,a_c$ and also those of $b_0,\ldots,b_c$ have been calculated. Then, constructing the polygon as a linear spline function connecting the points (u_0,b_0), (m_1,a_1), (u_1,b_1), (m_2,a_2), $(u_2,b_2),\ldots,$ (m_c,a_c), (u_c,b_c), the adjustment consists of two steps in which some of these points may be replaced by some other points in order to achieve nonnegativity, while preserving continuity and compatibility.

In the first step all zero-frequency classes $(f_i=0)$ are considered subsequently. For such a class i the integral of the polygon is zero, so that either $[b_{i-1}=a_i=b_i=0]$ or $[b_{i-1}<0 \vee a_i<0 \vee b_i<0]$ holds. In the latter case the three prevailing values are replaced by $b^*_{i-1}=0$, $a^*_i=0$ and $b^*_i=0$, so that the requirements are satisfied within class i, and at the same time a_{i-1} and a_{i+1} are replaced by a^*_{i-1} and a^*_{i+1} which are chosen such that compatibility is preserved in classes i-1 and i+1 (see figure 4):

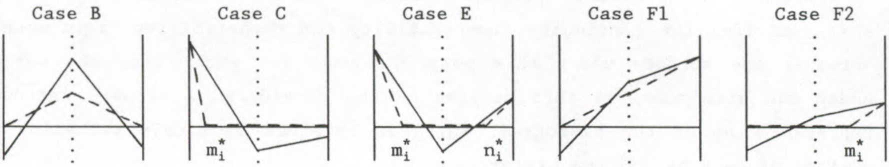
$$\tfrac{1}{4}(b_{i-2} + 2a^*_{i-1} + b^*_{i-1})w_{i-1} = f_{i-1}\, w_s \;\Leftrightarrow\; a^*_{i-1} = \tfrac{1}{2}(4d_{i-1} - b_{i-2}),$$
$$\tfrac{1}{4}(b^*_i + 2a^*_{i+1} + b_{i+1})w_{i+1} = f_{i+1}\, w_s \;\Leftrightarrow\; a^*_{i+1} = \tfrac{1}{2}(4d_{i+1} - b_{i+1}).$$

In the second step of adjustment all nonzero-frequency classes ($f_i>0$) are considered subsequently. Indicating all the values obtained after the first step without an asterisk, the new values can be indicated with an asterisk again. The polygon can be left unchanged within the domain of class i, if $b_{i-1} \geq 0$, $a_i \geq 0$ and $b_i \geq 0$. If this is not so, an adjustment is required. As a general principle any negative b_{i-1} or b_i will be replaced by zero, while positive values of b_{i-1} and b_i will remain unchanged. Any further adjustments depend on the signs of b_{i-1}, a_i and b_i. Table 3 gives the relevant cases and figure 5 illustrates the adjustments.

Table 3. Cases with negative values of b_{i-1}, a_i or b_i.

cases	A	B	C	D	E	F	G
b_{i-1}	–	–		–		–	
a_i	–		–	–	–		
b_i	–	–	–				–

Figure 5. The alternative polygon before (——) and after (- - -) the second step of adjustment.



Case A cannot occur, because $f_i>0$ and compatibility has been preserved. In case B the negative b_{i-1} and b_i are replaced by $b^*_{i-1}=0$ and $b^*_i=0$, while a_i is replaced by a^*_i in order to preserve compatibility:

$$\tfrac{1}{4}(b^*_{i-1} + 2a^*_i + b^*_i)w_i = f_i\, w_s \;\Leftrightarrow\; a^*_i = 2d_i \;\; (>0).$$

In case C the negative b_i is replaced by $b^*_i=0$ and the point (m_i, a_i) is replaced by the point $(m^*_i, 0)$ such that

$$\tfrac{1}{2}b_{i-1}(m^*_i - u_{i-1}) = f_i\, w_s \;\Leftrightarrow\; m^*_i = u_{i-1} + 2f_i\, w_s / b_{i-1} \;\; (>u_{i-1} \text{ and } <m_i).$$

Case D is similar to case C. In case E the point (m_i, a_i) is replaced by two other points, $(m_i^*, 0)$ and $(n_i^*, 0)$ respectively, satisfying

$$m_i^* - u_{i-1} = u_i - n_i^* \quad \text{and} \quad \tfrac{1}{2}b_{i-1}(m_i^* - u_{i-1}) + \tfrac{1}{2}b_i(u_i - n_i^*) = f_i w_s \quad \Leftrightarrow$$

$$\begin{cases} m_i^* = u_{i-1} + 2f_i w_s/(b_{i-1} + b_i) & (> u_{i-1} \text{ and } < m_i) \\ n_i^* = u_i - 2f_i w_s/(b_{i-1} + b_i) & (> m_i \text{ and } < u_i). \end{cases}$$

In case F the negative b_{i-1} is replaced by $b_{i-1}^* = 0$; if $\tfrac{1}{4}b_i w_i \leq f_i w_s$ (case F1) then a_i is also replaced by a_i^*, satisfying

$$\tfrac{1}{4}(b_{i-1}^* + 2a_i^* + b_i)w_i = f_i w_s \quad \Leftrightarrow \quad a_i^* = 2d_i - \tfrac{1}{2}b_i \ (\geq 0),$$

but if $\tfrac{1}{4}b_i w_i > f_i w_s$ (case F2) the point (m_i, a_i) is replaced by $(m_i^*, 0)$ with

$$\tfrac{1}{2}b_i(u_i - m_i^*) = f_i w_s \quad \Leftrightarrow \quad m_i^* = u_i - 2f_i w_s/b_i \ (> m_i \text{ and } < u_i).$$

Case G, finally, is similar to case F.

(d) Corollary

Obviously the procedure presented above leads to an alternative polygon that satisfies the continuity, nonnegativity and compatibility requirements. Moreover the surface under this polygon equals for each class the surface under the histogram, so this polygon can be considered a proper continuous representation of the histogram, which is obtained by simply reshaping the surface of each bar of the histogram.

(e) Modifications

It is worth noting that the alternative procedure can easily be modified by choosing positive values for b_0 and b_c rather than $b_0=0$ and $b_c=0$. One can also endogenize b_0 or b_c or both, for instance by taking $b_0=0$ and $b_c=\tfrac{1}{2}a_c$ or $b_0=\tfrac{1}{2}a_1$ and $b_c=\tfrac{1}{2}a_c$. In such cases one would obtain a compatible, but discontinuous polygon, which could be made continuous by extending it with one straight line from $(u_0-\tfrac{1}{2}w_1, 0)$ to (u_0, b_0) and another from (u_c, b_c) to $(u_c+\tfrac{1}{2}w_c, 0)$. For $b_0=\tfrac{1}{2}a_1$ and $b_c=\tfrac{1}{2}a_c$ such an extended polygon would be similar to the common polygon, yet compatible over the classes 1 to c.

(f) Choice of class bounds

When one looks at a histogram or either type of polygon, one should be aware that the shape of the figure depends heavily on the position of the class bounds $u_0,\ldots,u_c$. Large class widths make the figure smooth, while small class widths lead to a figure with more detail (see Chambers et al. 1983 Ch.2). If the exact observations are available, one is in a position to choose a set of preferred class bounds. This choice requires judgement and it is not possible to give precise rules for it (Schmid 1983 p.68), yet some helpful guidelines are given by Ootjers (1987) and Doane (1976).

(g) Interpretation of the function values

The values indicated on the vertical axis give the frequency density, i.e. the frequency per standard class width, so they can be interpreted by multiplying with a certain number of standard class widths.

With the common polygon the average density of each class d_i can be read directly from the figure. This advantage is lost when the alternative polygon is used. On the other hand the height of the polygon becomes more meaningful with the alternative polygon, because the height (density) derives its meaning from its integral (frequency) and the integral over each class is correct with the alternative polygon, but not with the common polygon.

(h) Conclusion

The alternative procedure leads to a continuous representation of a given frequency distribution, avoiding misrepresentations as produced by the common procedure. The price is greater complexity and more computations, but the basic idea of reshaping the bars of the histogram into a polygonal form is a trivial matter and in this computer era the computational burden should not be an impediment for implementation. Hence, statisticians might consider to incorporate the alternative polygon as a continuous version of the histogram in statistical textbooks and software, optionally with the simple or more complex adjustment procedure.

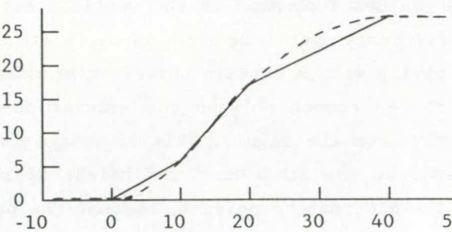
4. The common and alternative ogive

The ogive is a graphical representation of the cumulative frequency as a function of x. In the terminology of a given frequency distribution of section 2 the ogive is commonly constructed as a linear spline function, connecting consecutively the points $(u_0,0)$, (u_1,f_1), (u_2,f_1+f_2), ..., $(u_c,f_1+f_2+\ldots+f_c)$. See table 4 and figure 6. This *common* ogive is also called the cumulative frequency polygon. Obviously it is compatible with the given frequencies. The common ogive is in fact the integral of the height of the frequency histogram as a function of x, yet after dividing each height d_i by w_s, so that the height-measure becomes frequency per unit of **x**.

Table 4. Cumulative frequency distribution, example 2.

u_i	$f_1+\ldots+f_i$
0	0
10	6
20	17
40	27

Figure 6. Common (——) and alternative (---) ogive, example 2.



Similar, an *alternative* ogive can be derived by integrating the alternative polygon of section 3, after dividing the polygon function by w_s. For convenience the alternative polygon is described as a linear spline function, connecting p points (x_1,y_1), ..., (x_p,y_p), where $x_1<x_2<\ldots<x_p$ and $y_1,\ldots,y_p\geq 0$, while in particular $(x_1,y_1) = (u_0,0)$ and $(x_p,y_p) = (u_c,0)$. This polygon can be formulated as

$$p(x) = \begin{cases} \sum_{i=1}^{p-1} \alpha_i (x-x_i)_+ & \text{for } x_1<x<x_p \\ 0 & \text{elsewhere,} \end{cases}$$

where

$$(x-x_i)_+ = \begin{cases} 0 & \text{if } x\leq x_i \\ x-x_i & \text{if } x>x_i , \end{cases}$$

$$\alpha_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \sum_{j=1}^{i-1} \alpha_j \qquad \text{for } i=1,\dots,p-1.$$

The alternative ogive can now be derived as follows:

$$q(x) = [\int_{x_1}^{x} p(x)dx\ ]/w_s = \begin{cases} 0 & \text{for } x \le x_1 \\ [\sum_{i=1}^{p-1} \frac{1}{2}\alpha_i (x-x_i)_+^2\ ]/w_s & \text{for } x_1 < x < x_p \\ [\sum_{i=1}^{p-1} \frac{1}{2}\alpha_i (x_p-x_i)_+^2\ ]/w_s & \text{for } x \ge x_p . \end{cases}$$

An example is given in figure 6, where $p=5$, $(x_1,\dots,x_5) = (0,5,15,30,40)$ and $(\alpha_1,\dots,\alpha_4) = (1.43,-0.922,-0.9847,-0.0313)$.

The alternative ogive is a spline function of the second degree. It is: (1) *continuous* and *differentiable*; (2) *nondecreasing*; (3) *compatible* with the given frequency distribution. It is smoother than the common ogive, which is not differentiable. If the random variable $\mathbf{x}$ is assumed to have a continuous density function, then the cumulative density function is differentiable, so that in such cases the alternative ogive might be preferred. Therefore it might be incorporated in statistical textbooks and software along with the alternative polygon as two appropriate and mutually consistent representations of frequency distributions of random variables with a continuous density function.

5. Some remarks on density estimation

In the above the histogram and both types of polygons were discussed as tools of descriptive statistics. Nevertheless the resulting frequency density functions could also be used as estimators of the unknown probability density function of $\mathbf{x}$. For this purpose the total surface under the density estimator should be equal to one, so one should choose $w_s=1/(f_1+\dots+f_c)$, and with the common polygon all class widths should be equal (see figures 1 and 2). Scott (1985) gave results on the efficiency and the optimal class widths of the histogram and the common polygon. Similar results on the alternative polygon would be required in order to justify it's use as a density estimator. Hence, research on this subject is welcome.

References

Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey (1983): *Graphical Methods for Data Analysis*, Wadsworth, Belmont (California).

Doane, D.P. (1976): "Aesthetic Frequency Classifications", *The American Statistician*, 30, 181-183.

Harnett, D.L. (1982): *Statistical Methods*, 3rd ed., Addison-Wesley, Reading.

Ootjers, M.G. (1987): "Staafdiagrammen en Histogrammen in Statistische Programmatuur, in: *Symposium Statistische Software 1987*, Rekencentrum Rijksuniversiteit Groningen, pp. 27-43.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1988): *Numerical Recipes in C*, Cambridge University Press, Cambridge.

Schmid, C.F. (1983): *Statistical Graphics*, Wiley, New York.

Scott, D.W. (1985): "Frequency Polygons: Theory and Application", *Journal of the American Statistical Association*, 80, 348-354.

Velleman, P.F. and D.C. Hoaglin (1981): *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

Yule, G.U. and M.G. Kendall (1950): *An Introduction to the Theory of Statistics*, 14th ed., Griffin, London.

ontvangen 20-11-1990
geaccepteerd 3-07-1991