MATHEMATICSCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport SP 30A

The plotting of observations on probability paper.[1]

by

A. Benard and E.C. Bos-Levenbach

1955

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

The cumulative-normal-distribution curve can be represented by a straight line by means of plotting on normal probability paper. In general, probability paper can be designed on which continuous distributionfunctions of the type $F(\alpha x+\beta)$, F being a known distributionfunction, appear as a straight line. This kind of paper can be used for several purposes, for instance
1) To get an indication whether the sample arises from a probability distribution of the type in question, or
2) To get a quick estimate of the parameters α and β, based on a given random sample $x_1, \ldots, x_n$.

In both cases the observations $x_1, \ldots, x_n$ are plotted on probability paper. The points are plotted according to increasing size along the horizontal axis and an estimate of F is used for the corresponding ordinate. This can be done in several ways. Some existing methods are described in this paper and compared with a new method, having the property that with a very good approximation the medians of the x_i are situated on a straight line. This method is especially useful for the first purpose, e.g. to find out whether the sample originates from a distribution with a specified, e.g. a normal, probabilitydistribution. A paper about the second purpose mentioned above has been published some time ago by CHERNOFF AND LIEBERMANN [1].

1. Some current methods.

The observations $x_1, \ldots, x_n$ are supposed to be arranged according to increasing size

(1) $$x_1 < x_2 < \cdots < x_n,$$

no equal values occurring among the x_i. The probability of ties is equal to 0 for continuous distributions and only these are considered. A remark on the treatment of ties, due to grouping, is given in section 4.

The various estimates for $F(x_i)$ in current use will be denoted by $\varphi_1(i)$, $\varphi_2(i)$, etc.; they depend only upon the serial number i of the observation in question. The points plotted on the probability paper are then $(x_1, \varphi(1)), \ldots, (x_n, \varphi(n))$.

---

1) Report SP 30A of the Statistical Department of the Mathematical Centre, Amsterdam. Head: PROF.DR.D. VAN DANTZIG.

First we consider the function

(2) $$\varphi_1(i) = i/n$$

Though this estimate is used very often, it has the disadvantage that in the case of normal probability paper for $i=n$ the ordinate $\varphi_1(n)=1$ cannot be plotted if the distribution under investigation has infinite range. The point $(x_n, \varphi_1(n))$ will then lie at infinity.

The same objection holds for $i=1$ for the ordinate

(3) $$\varphi_2(i) = (i-1)/n.$$

This difficulty can e.g. be avoided by using one of the following functions:

(4) $$\varphi_3(i) = (i-\tfrac{1}{2})/n$$

or

(5) $$\varphi_4(i) = i/(n+1)$$

The choice between the different ordinate-functions depends necessarily on their properties, which will therefore be investigated further. It may be remarked in advance that all methods considered in this paper are asymptotically, for $n\to\infty$, equivalent. That this is true for $\varphi_1, \cdots, \varphi_4$ follows at once from their definitions (2),..,(5).

2. The position of the point $(x_i, \varphi(i))$.

$\underline{x}_i$ [2] the ith order statistic in the sample, being a random variable the same holds for every function of $\underline{x}_i$, in particular for $F(\alpha\underline{x}_i+\beta)$. The graph $y=F(\alpha x+\beta)$, which is a straight line on the probability paper for F, represents the point $(x_i, F(\alpha x_i+\beta))$ for every value of x_i. The random variable $(\underline{x}_i, F(\alpha\underline{x}_i+\beta))$ therefore has a probability distribution on this line.[3] The line itself is, of course, unknown, α and β being unknown parameters.

---

2) The random character of a variable is denoted by underlining its symbol. The same symbol, not underlined, can then be used for values which may be assumed by this random variable.

3) This probability distribution is not the rectangular distribution, $\underline{x}_i$ being the ith order statistic; the distribution is thus different for different values of i.

The random variable

(6) $\underline{y}_i = F(\alpha \underline{x}_i + \beta) \qquad (i = 1, \ldots, n)$

has a probability distribution [4] with mean

(7) $\mathcal{E}\,\underline{y}_i = i/(n+1)$

and mode

(8) $\mathrm{Mod}\;\underline{y}_i = (i-1)/(n-1)$.

Furthermore the median y_i^* of $\underline{y}_i$ is approximately equal to:

(9) $y_i^* = \mathrm{Med}\;\underline{y}_i \approx (i-0.3)/(n+0.4)$.

This median as well as its approximate value are situated between the mode and the mean,[5] Now consider values of $i > \frac{1}{2}(n+1)$ (For values of $i < \frac{1}{2}(n+1)$ analogous results follow from analogous arguments) The following inequality holds for $i > \frac{1}{2}(n+1)$

(10) $\mathcal{E}\,\underline{y}_i < \mathrm{Med}\;\underline{y}_i < \mathrm{Mod}\;\underline{y}_i$

and this implies that the value of $\underline{y}_i$ corresponding to $\underline{x}_i$ (cf.(6)) is more often situated above than below its mean $\mathcal{E}\,\underline{y}_i$, while the reverse is true for the mode. As the point $(\underline{x}_i, \underline{y}_i)$ is always situated on the unknown line $y = F(\alpha x + \beta)$, the point $(\underline{x}_i, i/(n+1))$ will more often be situated below the line than above, and the point $(\underline{x}_i, (i-1)/(n-1))$ more often above than below the line Consequently the use of φ_4 (cf. (5)) which (cf. (7)) is equal to $\mathcal{E}\,\underline{y}_i$, has the disadvantage that for $i > \frac{1}{2}(n+1)$ the points plotted on probability paper are mostly situated below the line representing the unknown probability distribution and vice versa in the case of $i < \frac{1}{2}(n+1)$.

---

4) Proofs and references are given in an appendix.

5) It is easy to see from (7),(8) and (9) that this is true for the approximate value. The proof for the median itself will not be given in this paper. It follows readily from C.G. LEKKERKERKER [4] .

Therefore this way of plotting the points leads to an estimate of the unknown line which will more often than not underestimate its slope thus overestimating the variance of the original distribution.

Furthermore this effect is especially pronounced for small and large values of i; consequently the points thus plotted will have a tendency to result in an S shaped graph on probability paper.

A similar objection holds for φ_3 (cf (4)), as, for $i > \frac{1}{2}(n+1)$

$$(i-\tfrac{1}{2})/n > (i-0.3)/(n+0.4)$$

and the other way around for $i < \frac{1}{2}(n+1)$ Thus this method involves frequent underestimation of σ and has the tendency to produce a reversed S shaped curve on probability paper

3. A new method

These objections may be met by using

$$(11) \qquad \varphi_5(i) = (i-0.3)/(n+0.4)$$

because the points $(x_i, (i-0.3)/(n+0.4))$ will for every i be situated about as often below as above the unknown line under investigation. The situation is summarized in figure 1.

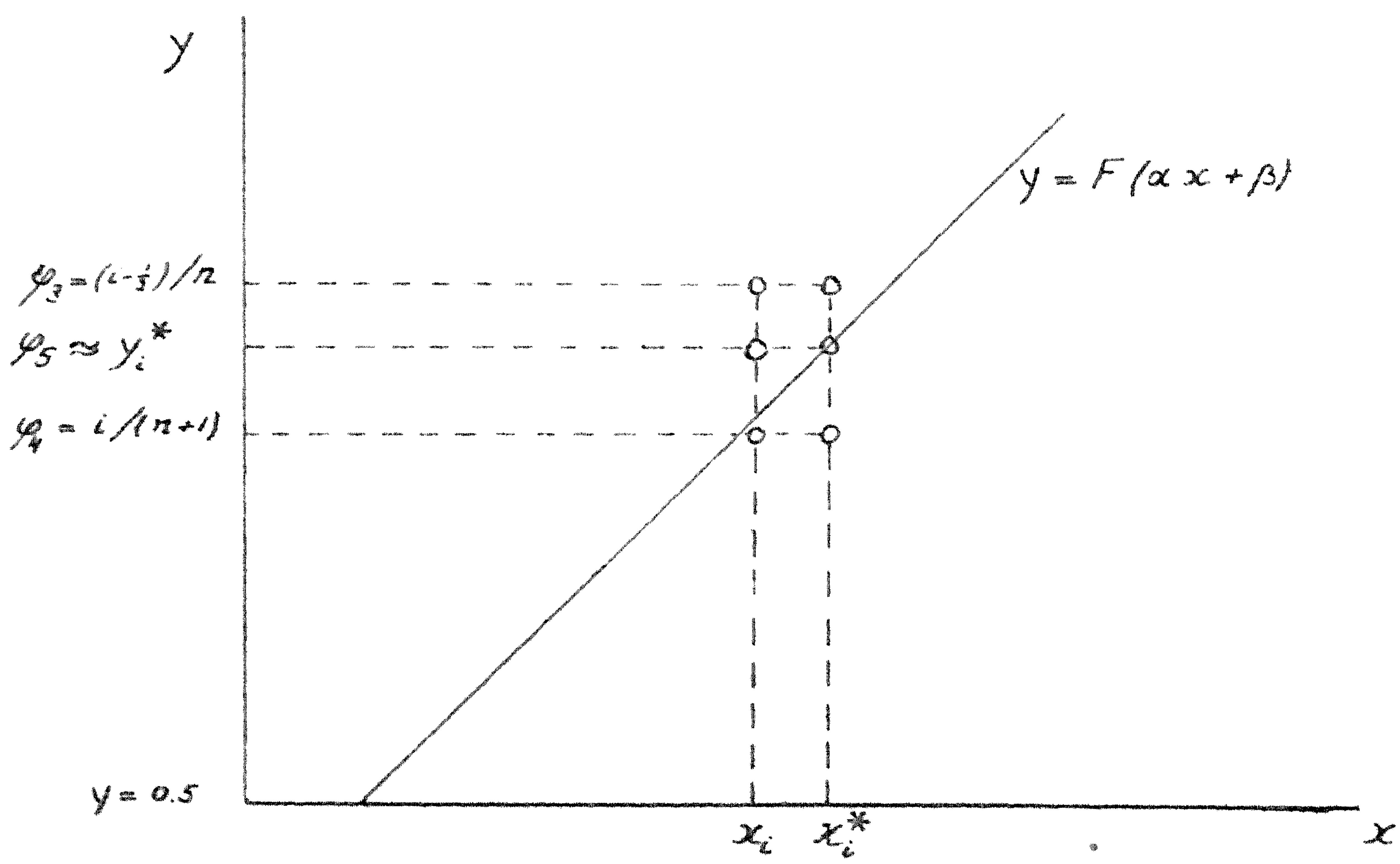

Figure 1: Three methods for plotting observations on probability paper.

The straight line represents the (unknown) distributionfunction $y = F(\alpha x + \beta)$. x_i^* represents the median of $\underline{x}_i$ and y_i^* the median of y_i . The point (x_i^*, y_i^*) is situated on the line $y = F(\alpha x + \beta)$ and the point $(\underline{x}_i, y_i)$ has a probability distribution on the line $y = F(\alpha x + \beta)$ with equal probabilities to lie to the left and to the right of the point (x_i^*, y_i^*) The figure shows that the point $(\underline{x}_i, \varphi_3(i))$ will be situated as often to the left as to the right of $(x_i^*, \varphi_3(i))$, and therefore more often above than under $y = F(\alpha x + \beta)$, while the reverse holds for the point $(\underline{x}_i, \varphi_4(i))$. The point x_i in figure 1 lies to the left of x_i^*, while $(x_i, \varphi_4(i))$ is nevertheless still situated below instead of above the line.

4. Remarks.

1) A sixth possibility

(12) $$\varphi_6(i) = (i-1)/(n-1),$$

of which the right side represents the mode of $\underline{y}_i$, combines the disadvantages of φ_1, φ_2 and φ_3.

2) As y_i is a monotonous function of x_i $(y_i = F(\alpha x_i + \beta))$, an approximation of the median of $\alpha x_i + \beta$ is found by substituting $y_i = (i-0{,}3)/(n+0{,}4)$ in the inverse function $F^{-1}(y_i)$ of F; i.e. $Med(\alpha \underline{x}_i + \beta) \simeq F^{-1}(i-0{,}3)/(n+0{,}4)$. If α and β are known this leads to an approximation of $Med\ \underline{x}_i$; if α and β are unknown the abscissa corresponding to the ordinate $(i-0{,}3)/(n+0{,}4)$ estimated bymeans of the estimated straight line, is an estimate of $Med\ \underline{x}_i$

3) If tied observations are present, due to grouping or rounding off, it seems best to use the method of average ranks, i.e to assign the average of the ranks of a tie to all of its numbers.

4) There are still other methods besides those described here. E.J. GUMBEL [2] e.g suggests one for a special type of probability paper of his own design for extreme values (the double exponential distribution) which for $i = 1$ nearly coincides with the use of φ_1, and for $i = n$ with that of φ_2 , with lineair interpolation for the ordinates of the other points. The method is especially adepted to the distribution considered by GUMBEL and is based on the use of the modes of $\underline{x}_1$ and $\underline{x}_n$ for this distribution.

5 Appendix.

Let $\mathcal{G}_i$ be the cumulative distributionfunction of $\underline{y}_i$, then

$$\mathcal{G}_i(y) = \sum_{k=i}^{n} \binom{n}{k} y^k (1-y)^{n-k} \qquad (0 \leq y \leq 1)$$

This can also be written in the following form

$$\mathcal{G}_i(y) = \frac{n!}{(i-1)!(n-i)!} \int_0^y u^{i-1}(1-u)^{n-i}\,du \tag{13}$$

(cf e.g. M.G. KENDALL [3] , p. 120). The mode of $\underline{y}_i$, cf. (8), can be deducted from the equation

$$\frac{\partial^2 \mathcal{G}_i(y)}{\partial y^2} = 0$$

while the expectation of $\underline{y}_i$ (cf. (7)) can be inferred from (13). The numerical values of y_i^*, the median of $\underline{y}_i$, can be found by means of the tables of the incomplete bêta-function (C.M. THOMPSON [6]).

The approximation (9) can be deducted as follows:

For every value of the cumulative distribution function $\mathcal{G}_i(y)$ the following relation holds:

$$\mathcal{G}_i(y) = \sum_{k=i}^{n} \binom{n}{k} y^k (1-y)^{n-k} = 1 - \sum_{k=0}^{i-1} \binom{n}{k} y^k (1-y)^{n-k} =$$

$$= 1 - \sum_{k=n-i+1}^{n} \binom{n}{k} y^{n-k}(1-y)^k = 1 - \mathcal{G}_{n+1-i}(1-y).$$

From this it follows that, if the observations are arranged according to decreasing instead of increasing size, i is replaced by $n+1-i$, y by $1-y$ and $\mathcal{G}$ by $1-\mathcal{G}$. It is obvious that φ should also satisfy this symmetrical relation, i.e.

$$\varphi(i) = 1 - \varphi(n+1-i)$$

$\varphi_3, \dots, \varphi_6$ meet this requirement, φ_1 and φ_2 do not.

Let y_i^* be the median of $\underline{y}_i$, then

$$\left\{ \begin{array}{l} \mathcal{G}_i(y_i^*) = \frac{1}{2} \\ 1 - \mathcal{G}_{n+1-i}(1-y^*) = \frac{1}{2} \end{array} \right. \tag{14}$$

and $1 - y_i^*$ is the median of $\underline{y}_{n+1-i}$.

Now, writing for y_i^*

(15) $y_i^* = (i-a)/(n+b)$,

where a and b are functions of i and n to be determined below, (14) and (15) give

$$1-(i-a)/(n+b) = (n+1-i-a)/(n+b),$$

or

(16) $b = 1-2a$.

Formula (15) now becomes

(17) $y_i^* = (i-a)/(n+1-2a)$.

Furthermore according to (14) a should satisfy the relation

(18) $\sum_{k=i}^{n} \binom{n}{k} \left((i-a)/(n+1-2a)\right)^k \left(1-(i-a)/(n+1-2a)\right)^{n-k} = 1/2$.

This can also be written as follows

(19) $\sum_{k=0}^{i-1} \binom{n}{k} \left((i-a)/(n+1-2a)\right)^k \left(1-(i-a)/(n+1-2a)\right)^{n-k} = 1/2$.

The value of a which satisfies this relation may be found by means of the tables of the incomplete bêta-function, and depends on i and n.

In table I the values of a are given for $n=1,2,\ldots,8$ and $i=1,\ldots,[n/2]$. Using the formula $a(i,n) = a(n+1-i,n)$, a can be computed for all other values of i, except $i=\frac{1}{2}n+\frac{1}{2}$ when n = odd.

Table I

Values of $a \times 10^3$ for some small values of i and n.

i \ n	2	3	4	5	6	7	8
1	293	298	300	301	302	303	303
2			312	315	316	317	318
3					319	320	321
4							322

Table I shows, that for small values of n and for $i = 1$ the value $a = 0,3$ is a good enough approximation of the true value of a, which is itself not constant. One might, of course, use the exact values of a and substitute these in (17), plotting the points (x_i, y_i^*), but this is a rather cumbersome method. On the other hand small changes in a do not have much influence on y_i^* in (17) if i and n are not very small. This seems sufficient reason to use the value $a = 0,3$ for every i and n, thus leading to φ_5 (cf. (11)).

This argument may be supported by an additional investigation of the asymptotic behaviour of a for $n \to \infty$. Consider the limit for $n \to \infty$ of the left hand member of (19):

$$\lim_{n\to\infty} \sum_{k=0}^{i-1} \binom{n}{k} \left((i-a)/(n+1-2a)\right)^k \left(1-(i-a)/(n+1-2a)\right)^{n-k} =$$

$$\lim_{n\to\infty} \sum_{k=0}^{i-1} \left\{n(n+1)\cdots(n-k+1)/(n+1-2a)\right\} \times \left\{(i-a)^k/k!\right\} \times \left\{\left(1-(i-a)/(n+1-2a)\right)\right\}^{n-k} =$$

$$e^{-(i-a)} \sum_{k=0}^{i-1} (i-a)^k/k! \, .$$

Equalizing this limit to $\frac{1}{2}$, the problem is to find Poisson distributions with mean $i-a$ and median i.

A table of Poisson distributions (e.g. E.C. MOLINA [5]) gives for every i the corresponding value of a. The results are given in Table II.

Table II

Values of a for different i's, $n \to \infty$.

i	a
1	0,307
2	0,321
5	0,329
10	0,331
50	0,332
100	0,333

The exact value of a for $i = 1$ is obtained from

$$1/2 = e^{-(i-a)} \sum_{k=0}^{i} (i-a)^k/k! \to e^{-(1-a)} = e^{-\ln 2}$$

or $a = 1 - \ln 2$.[6)]

Besides the value of a for $i = 1$ and $n \to \infty$ one can also compute the value of a for i/n = constant and $n \to \infty$. Substituting $i/n = \alpha + \varepsilon_n/n$ and usin an approximative formula given by USPENSKY [7] p. 129 one finds $a = 1/3$.[7)]

Finally, to check the degree of approximation, for $i = 1$ and a few small values of n, the exact value of $G_1((i-a)/(n+1-2a))$, $G_1(0.7/(n+0.4))$ was computed. G_1 is equal to the probability that the first point plotted falls below the true line given by (6) and should thus be approximately equal to $\frac{1}{2}$.

The results, given in Table III, are favorable for the value of a chosen.

Table III

Exact values of $G_1(0.7/n+0.4)$

n	G
2	0,4983
3	0,4992
4	0,5000
5	0,5005

The approximation is very satisfying and has, moreover, been checked for $n = 10$ and 15 by computing the exact values of $\varphi_5(i)$ for $i = 1, \ldots, 10$ and $1, \ldots, 15$ respectively, and comparing these with the corresponding values of $(i-0,3)/(n+0,4)$. The difference proved to be smaller than 1% for every value investigated.

Acknowledgement.

The authors want to thank Prof J. HEMELRIJK for many helpfull suggestions.

----

6) This can also be deducted directly from formula (19) for $i = 1$, giving $(1-(1-a)/(n+1-2a))^n = 1/2$ expended $a = 1 - \ln 2 - (1{,}5\ln 2 - 1)\ln 2/n - (2\tfrac{1}{6}\ln 2 - 1{,}5)\ln^2 2/n^2 \ldots$

7) This computation, suggested by PROF.DR. D. VAN DANTZIG, was executed by H. KESTEN and TH.J. RUNNENBURG, assistants of the Statistical Department of the Mathematical Centre, Amsterdam.

Literature

[1] H. Chernoff and G.J. Liebermann, Use of normal probability paper, Journal of the American Statistical Association, (1954) 49, p. 778-785.

[2] E.J. Gumbel, The return period of flood flows, Annals of Mathematical Statistics 12 (1941), p. 163-190.

Symplified plotting of statistical observations. Trans-American Geophysical Union 26 (1945) p. 69-82.

[3] M.G. Kendall, The advanced theory of statistics, Vol. I, London 1947.

[4] C.G. Lekkerkerker, Rapport Z.W. 1953-016, Afd. Zuivere Wiskunde, Mathematisch Centrum, Amsterdam.

[5] E.C. Molina, Poisson's exponential binomial limit, D. van Nostrand, Comp., Inc., N.Y. 1945.

[6] C.M. Thompson, Tables of percentage points of the incomplete bêta-function, Biometrika 32 (1941) p. 168-181.

[7] J.V. Uspensky, Introduction to mathematical probability (6th impr. 1937). Mc Graw-Hill Book Cy., New York and London.
