The plotting of observations on probability paper. [1)]

By A. Benard and E.C. Bos-Levenbach
Translated by Ronald Schop, Sr. Reliability Engineer, DAF Trucks N.V.

SUMMARY

The plotting of observations on probability paper.

The mathematical foundation of probability paper for a variate with a cumulative distribution function F ($\alpha x + \beta$) is explained as well as its purpose.

To plot the observations it is necessary to use an estimate of F (x_i), x_i being the i^{th} order statistic in the sample.

Several methods are described and compared, and a new one is developed, having the property that with a very good approximation the medians of x_i are situated on a straight line.

The derivations are given separately in an appendix.

1- Introduction.

When $\underline{x}$ is a normal distributed stochastic quantity [2)] with an average of μ and a spread σ with the probability function:

$$F(x) = P[\underline{x} \leqslant x] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du \qquad (1)$$

then plotting the mathematical positions of the points *(x,F(x))* on normal millimetre paper shows an appearance as in figure 1.

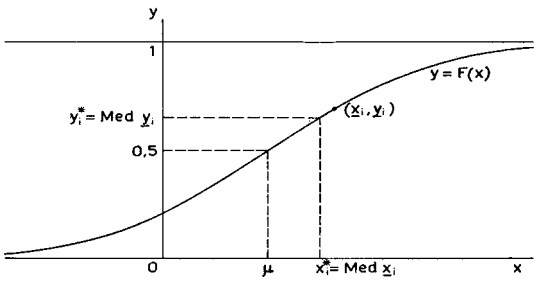

Fig. 1. Normale verdelingsfunctie op gewoon grafiekenpapier.

Fig. 1: Normal distribution on average graphic paper.

By transforming the vertical scale, this curved line (for any μ and σ) can be transformed in a straight line. This transformation consists of creating, on the vertical scale, at the point with co-ordinate *y*, a new co-ordinate y', the number $\Phi(y)$, in what Φ represents the distribution-function of the standardized normal distribution:

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{u^2}{2}} du \qquad (2)$$

[1)] Report SP 30 of the Statistical Department of the Mathematics Centrum, Amsterdam. This department is under supervision of Prof. Dr. D. van Dantzig.

[2)] Underlining shows that this quantity is stochastic, meaning it belongs to a probability distribution.

As:

$$F(x) = P[\underset{\sim}{x} \leqslant x] = P\left[\frac{\underset{\sim}{x} - \mu}{\sigma} \leqslant \frac{x - \mu}{\sigma}\right] = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

where *y' = F(x)* belonging value of the *y* equals *(x-μ)/σ*. The geometrical position of the points *(x, y')* with *y' = F(x)* becomes on the original linear *y*-scale the geometrical position of the points *(x, (x-μ)/σ)*, being a straight line. In this, *μ* is the x co-ordinate (abscissa) of the point with co-ordinate *y'= ½*, while *1/σ* is the direction-coefficient of the line.

Except for this normal probability paper, on analogous way, starting with an arbitrarily distribution-function *F(x),* probability paper can be created on which distribution functions of the type *F(αx + β),* with α and β as parameters, are shown in a straight line. It is also possible e.g. to create probability paper belonging to distributions of the type *F($e^{\alpha x + \beta}$)* . This by transforming one scale logarithmic and the other scale according to *F(x).* The same applies for distributions of the type *F(φ(αx + ß)),* in which φ is an arbitrarily monotone function. The one scale will be transformed to φ, the other to *F*.

The considerations in the next paragraphs apply to all these kinds of probability paper.

2 . The purpose of the probability paper.

Probability paper can, among others, be used to obtain a quick estimation of both the parameters μ and σ, based on a given sample-check $x_1, \ldots, x_n$, if the shape of the probability distribution of $\underset{\sim}{x}$ is known; or to obtain a visual impression for answering the question if the quantity owns a distribution of the considered type.

In both cases the observations $x_1, \ldots, x_n$ have to be plotted on probability paper. This can be done on several different ways. The difficulty is that for each point $(x_i, F(x_i))$, plotted on probability paper resulting in exactly a straight line, only the co-ordinate x_i is known, while the $F(x_i)$ is unknown as the *μ* and the *σ* are unknown. Therefore an estimation of $F(x_i)$ is used instead of $F(x_i)$ its self. For this estimation several functions can be used.

3 . Some common methods.

In the next part we assume that the observations $x_1, \ldots, x_n$ are numbered in rising magnitude, while there are no equal observations (in a continuous distribution of x the probability for equal observations is equal to zero), so we have:

$$x_1 < x_2 < \ldots < x_n \tag{3}$$

The different estimations, used for $F(x_i)$ are given the names $\varphi_1(i)$, $\varphi_2(i)$, etc, as they all only depend on the rank-number *i* of the observed observations.

First we consider the usually applied function:

$$\varphi_1(i) = \frac{i}{n}. \tag{4}$$

This has the disadvantage that $\varphi_1(n) = 1$, so that the point $(x_n, \varphi_1(n))$ falls outside the probability paper. After all, *y' = 1* stands on the vertical scale at *y = ∞*, as *Φ(∞) = 1* and this point is not on the paper. The same objection, but now for *i=1* applies when one uses;

$$\varphi_2(i) = \frac{i - 1}{n} \tag{5}$$

This objection can be overcome in several ways, for example by using the following functions;

$$\varphi_3(i) = \frac{i - \frac{1}{2}}{n} \tag{6}$$

or

$$\varphi_4(i) = \frac{i}{n+1}. \tag{7}$$

Both functions are used and there are numerous other possibilities. In order to make a choice of the different possibilities the properties of the different methods have to be investigated.
In all cases the points

$$(x_1, \varphi(1)), \ldots, (x_n, \varphi(n))$$

are plotted on the probability paper, and as can be seen from the equations (4), . .,(7), all 4 above described methods are asymptotic equivalent for $n \to \infty$. This also is applicable for the later on introduced functions φ_5 and φ_6.

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

As $\underline{x}_i$, the i^{th} observation in order of magnitude, is a stochastic quantity, the same applies for any function of $\underline{x}_i$, so especially for $F(x_i)$. The point $(x_i, F(x_i))$ is for every value of x_i positioned on the curve $y = F(x)$ as in figure 1, as this curve is the mathematical position of these points. The stochastic point $(\underline{x}_i, F(x_i))$ thus has a probability distribution over this curve. [3] The stochastic quantity

$$\underline{y}_i = F(\underline{x}_i) \qquad (i = 1, \ldots, n) \tag{8}$$

owns a probability distribution with an average [4] of

$$\mathcal{E}\,\underline{y}_i = \frac{i}{n+1} \tag{9}$$

and a modus of

$$Mod\,\underline{y}_i = \frac{i-1}{n-1}. \tag{10}$$

For the median y_i^* of $\underline{y}_i$ is valid, by approximation:

$$y_i^* = Med\,\underline{y}_i \approx \frac{i - 0{,}3}{n + 0{,}4} \tag{11}$$

and we see that this latest number is always situated between both other numbers. It can be demonstrated that this is not only valid for the approximation of the median, but also for the median itself. [5]

We now observe the values of $i > (n+1)/2$ (for the values $< (n+1)/2$ similar conclusions are applicable). For an i like this is valid.

$$\mathcal{E}\,y_i < Med\,\underline{y}_i < Mod\,\underline{y}_i \qquad \left(i > \frac{n+1}{2}\right). \tag{12}$$

[3] This probability distribution is not the same as the original probability distribution of $\underline{x}$, as x_i is the i^{th} observation when ranked at magnitude.

[4] For justification and references see appendix of this article.

[5] The proof of this is not given in this article. It follows on a simply way form C. G. Lekkerkerker [4] (see literature list).

This means that the at $\underline{x}_j$ belonging value $\underline{y}_i$ (see (8)) is more often positioned above than below its average $\varepsilon\underline{y}_j$, while the opposite is valid for the modus. As the point $(\underline{x}_j, \underline{y}_j)$ always is positioned on the in figure 1 drawn curve, the point $(\underline{x}_j, i/(n + 1))$ will in more than half of the cases be positioned below this curve and the point $(\underline{x}_j, (i - 1)/(n - 1))$ in more than half above. This stays valid when the vertical scale is transformed according to Φ (see (2)) because this transformation is monotone, meaning that the order of the points in vertical direction does not change.

The function φ_4 (see (7)) thus has the objection that, for $i > (n + 1)/2$ the points, plotted on the probability paper, in more than half of the cases are below the line, representing the unknown probability distribution, while this is just the opposite for $i < (n + 1)/2$.

So, this way of plotting gives as results that in more than half of the cases the slope of the line will be estimated too low, so the spread too high. On top of that, this effect is the most for small and large values of i, consequently the plotted points are tending to form an S-shape figure on the probability paper. Exactly the same objection, but just the other way around, applies for φ_3 (see (6)) as

$$\frac{i - \frac{1}{2}}{n} > \frac{i - 0{,}3}{n + 0{,}4} \quad \text{for} \quad i > \frac{n + 1}{2}$$

and, the other way around, for $i < (n + 1)/2$. Thus this way of plotting leads to an underestimation of σ in more than half of the cases and also to S-shapes curved plots, but than in the other direction.

When one uses

$$\varphi_5(i) = \frac{i - 0{,}3}{n + 0{,}4}, \qquad (13)$$

then these objections are not existing, as the points $(x_i, (i - 0{,}3)/(n + 0{,}4))$ for each i shall be located approximately equal often above as below the searched line.

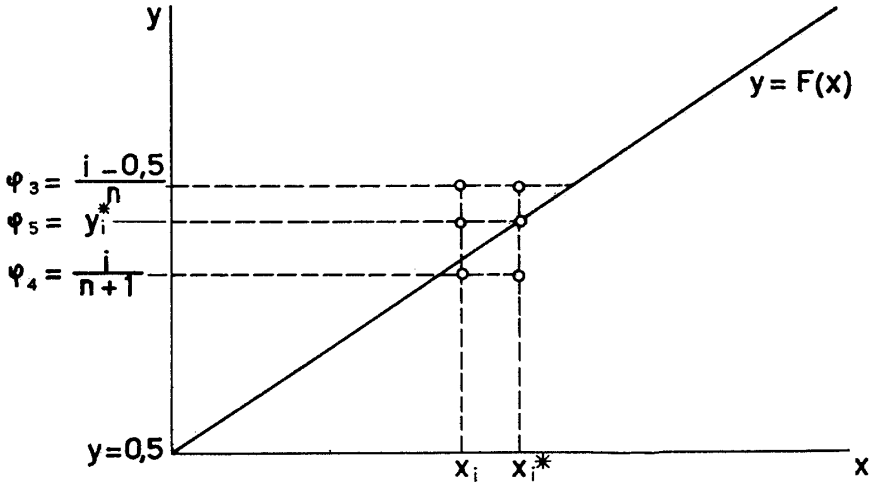

Fig. 2. Drie methoden voor het uitzetten van waarnemingen op waarschijnlijkheids-papier.

Figure 2. Three methods for plotting observations on probability paper.

All above is illustrated in figure 2. In this the straight line represents the distribution function $y = F(x)$, (the figure is supposed to be drawn on probability paper) x_i^* is supposed to be the median of $\underline{x}_j$ and y_i^* the median of $\underline{y}_j$. The point (x_i^*, y_i^*) is positioned on the straight line $y = F(x)$, (see also remark 2 below) and the point $(\underline{x}_i, \underline{y}_i)$, that owns a probability distribution on the line $y = F(x)$, lies on this line equally often left as right of this point.

In the figure one can see that the point $(x_i, \varphi_3(i))$ equally often is positioned on the left of $(x_i^*, \varphi_3(i))$ as on the right, and so thus more often above $y = F(x)$ than below, and opposite for $\varphi_4(i)$. The in the figure drawn point x_i is positioned left from x_i^* while still $(x_i, \varphi_4(i))$ is positioned below the line instead of above.

Remarks.

1. A sixth possibility

$$\varphi_6(i) = \frac{i - 1}{n - 1}, \qquad (14)$$

in which the right part of the equation represents the modus *of* $\underline{y}_j$, combines the objections of φ_1, φ_2 and φ_3 and thus can not be recommended.

2. As y_i is a monotone function of x_i, (after all, $y_i = F(x_i)$) we find a method of approaching the median of $\underline{x}_j$ by substituting the value $y_i = (i - 0{,}3)/(n + 0{,}4)$ in the inverse function of $x_i = F^{-1}(y_i)$. Thus

$$Med\ \underline{x}_i \approx F^{-1}\left(\frac{i - 0{,}3}{n + 0{,}4}\right)$$ according to figure 1.

3. When in the observations, due to grouping or rounding-off, equal numbers appear; probably the method of average rank-numbers can be used. This method consist of appointing the mean of the rank-numbers that these observations would have had, if they were not equal, but compared with all not to the group belonging observations would have had the same position when ranking to magnitude as now done, to all observations of a group with equal rank-numbers. These median rank-numbers are substituted in the formulas for φ.

4. The described methods here are certainly not exhausting all the possibilities. E. J. Gumbel [2] for example gives for the by him designed probability paper for extreme values (the double exponential distribution) a method in which the $F(mod\ \underline{x}_j)$ is plotted in vertical direction. This method is, for the case he observed, asymptotic equivalent with the use of φ_6, however without the objection that the first and the last observation can not be plotted.

5 . Appendix.

When the distribution function of y_i is indicated with G_i, than applies (see e.g. D. van Dantzig [1], chapter IV, par. 2):

$$G_i(y) = \sum_{k=i}^{n} \binom{n}{k} y^k (1 - y)^{n-k} \qquad (0 \leqslant y \leqslant 1)$$

which can also be written in the shape

$$G_i(y) = \frac{n!}{(i - 1)!\,(n - i)!} \int_0^y u^{i-1} (1 - u)^{n-i}\, du. \qquad (15)$$

(see e.g. M. G. Kendall [3]. Page 120). By equating the second derivation of G_i to y with zero, the modus of $\underline{y}_j$ is found, so (10). Also the expectation of $\underline{y}_j$ (9) follows on the usual way from (15).
The exact value of y_i^*, the median of y_i , can be found with help of the tables of the incomplete Béta-function (e.g. C. M. Thompson [6]). The approximation formula (11) is deduced as follows: for each value of the distribution function y the following relation is valid:

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

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

meaning that, if the observations are not ranked in rising but in decreasing order, then *i* will be replaced by *(n + 1 - i)*, *y* by *(1 - y)* and *G* by *(1 - G*). It is obvious that also for φ it might be required that this symmetrical relation exists. That means that now it is valid to state:

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

φ_1 and φ_2 do not comply with this, but the other previously mentioned functions of φ do. Now, if y_i^* is the median of $\underline{y}_i^*$, so that

$$G_i(y_i^*) = \tfrac{1}{2},$$

then

$$1 - G_{n+1-i}(1 - y_i^*) = \tfrac{1}{2}, \qquad (16)$$

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

Now we write y_i^* as

$$y_i^* = \frac{i-a}{n+b} \quad [6] \quad (a \text{ and } b \text{ are functions of } i \text{ and } n) \qquad (17)$$

From equations (16) and (17) follows the relation

$$1 - \frac{i-a}{n+b} = \frac{n+1-i-a}{n+b}$$

and from this follows

$$b = 1 - 2a \qquad (18)$$

Formula (17) transforms with this to

$$y_i^* = \frac{i-a}{n+1-2a} \qquad (19)$$

while a still must comply to the relation

$$\sum_{k=i}^{n} \binom{n}{k} \left(\frac{i-a}{n+1-2a}\right)^k \left(1 - \frac{i-a}{n+1-2a}\right)^{n-k} = \tfrac{1}{2},$$

[6] Asymptotic for $n \to \infty$ applies $y_i^* = i/n$.

which can also be written as

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

The value of *a*, complying to this relation, can be found again with help of the tables of the incomplete Béta-function, and the value is obviously depends on both *i* and *n*. We are looking for a constant as approximation for *a*, that is sufficient precise for practical use. Therefore, the limit for $n \to \infty$ is taken in the left part of the equation.

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

$$\lim_{n\to\infty} \sum_{k=0}^{i-1} \frac{n(n-1)\ldots(n-k+1)}{(n+1-2a)^k} \frac{(i-a)^k}{k!} e^{-(i-a)} =$$

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

This limit is now equated to ½.

This means that we are looking for a Poisson distribution with an average of *i* - *a* and a median of *i*. For each value of *i* we can, using the table of Poisson distributions (e.g. E. C. Molina [5]) determine the value belonging to *a*. The result of this is shows in table I.

TABLE I. Values of *a* for different *i* and $n \to \infty$

i	*a*
1	0,307
2	0,321
3	0,326
4	0,328
5	0,329
10	0,331
50	0,332
100	0,333

As we want to use óne fixed value for *a*, and as for *i* = 1 and *i* = *n* the points belonging to $(x_i , \varphi(i))$ normally show the largest deviations from the straight line, we choose the to this point belonging value of *a*, which is, for easier use, rounded to 0,3.

For examining the influence of the above used limit transition for $n \to \infty$, we defined for *i = 1* and for some small values of *n* the exact value of

$G_1\left(\frac{1-a}{n+1-2a}\right)$ with *a = 0,3*, that should be, as *a* was exactly correct, equal to ½.

The results, given in table II, are positive.

TABLE II. Exact values of $G_1\left(\frac{0{,}7}{n+0{,}4}\right)$.

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

Finally we compared, for $n = 10$ and $n = 15$, the values with the exact medians that can be established with the table [5] and when using this for every i, the probability that the point $(\underline{x}_i, \varphi(i))$ lies above the straight line should be exactly equal to ½ (but can not be represented by a constant value of a). These values appeared nearly almost to correspond up to three decimals accurate, while the difference for no single value of i is larger than 1%. From this we can conclude that the approximation is ample sufficient for practical use and that by plotting the points $x_i, (i - 0,3)/(n + 0,4))$ on probability paper one achieves that each of the point owns an equal chance to lie above or below the searched line.

Literature:

[1] D. van Dantzig, Kadercursus Mathematische Statistiek, Mathematisch Centrum, Amsterdam 1947.

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

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

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

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

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

The reference for the original paper in the Dutch language is:

Benard, A and Bos-Levenbach, E. C., "*Het uitzetten van waarnemingen op waarschijnlijkdeids-papier" (The Plotting of Observations on Probability Paper)*, **Statististica Neerlandica**, Volume 7, pages 163-173, 1953.

Please note the European convention of using commas (,) prevails rather than the American convention of a decimal point (.). Thus 0,3 in European convention is equivalent to 0.3 in American convention.

December 28, 2001.