MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

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

SP 30

Het uitzetten van waarnemingen op waarschijnlijkheidspapier

door

A.Benard en E.C.Bos-Levenbach

1954

Summary

The plotting of observations on probability paper

The mathematical foundation of probability paper for a variate with cumulative distribution function $F(ax+b)$ 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. Inleiding.

Is $\underline{x}$ een normaal verdeelde stochastische grootheid [1]) met gemiddelde μ en spreiding σ, waarvan

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

de verdelingsfunctie voorstelt, dan heeft de meetkundige plaats van de punten $(x, F(x))$ op gewoon millimeterpapier uitgezet een gedaante als in figuur 1

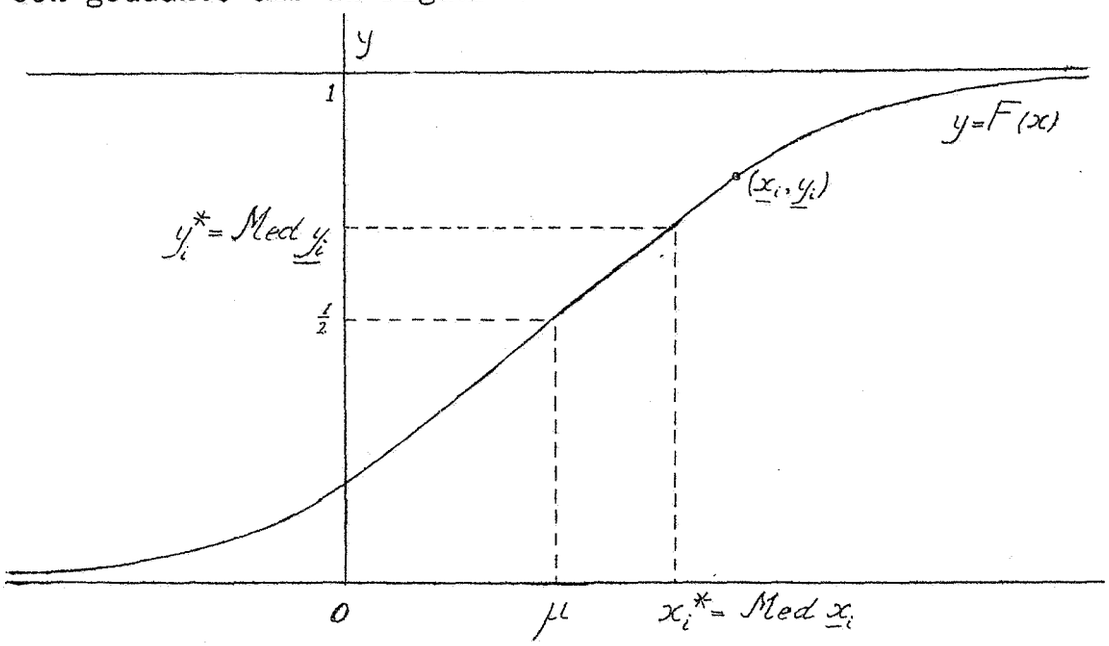

Fig. 1. Normale verdelingsfunctie op gewoon grafiekenpapier

Door transformatie van de verticale schaal kan men gedaan krijgen, dat deze kromme lijn (voor iedere μ en σ) in een rechte lijn overgaat. Deze transformatie bestaat daaruit, dat men op de verticale schaal bij het punt met ordinaat y als nieuwe ordinaat y' het getal $\Phi(y)$ zet, waarin Φ de verdelingsfunctie van de gestandaardiseerde normale verdeling voorstelt:

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

Daar

$$F(x) = P[\underline{x} \leq x] = P\left[\frac{\underline{x}-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

is, is de bij $y' = F(x)$ behorende waarde van y gelijk aan $\frac{x-\mu}{\sigma}$. De meetkundige plaats van de punten (x, y') met $y' = F(x)$ wordt dus op de oorspronkelijke lineaire y-schaal de meetkundige

---

1) Onderstreping geeft aan, dat de desbetreffende grootheid stochastisch is, d.w.z. een waarschijnlijkheidsverdeling bezit

plaats der punten $(x; \frac{x-\mu}{\sigma})$, hetgeen een rechte lijn is. Hierbij is dan μ de abscis van het punt met ordinaat $y' = \frac{1}{2}$, terwijl $1/\sigma$ de richtingscoëfficiënt van de lijn is.

Behalve dit normale waarschijnlijkheidspapier kan men op analoge wijze, uitgaande van een willekeurige verdelingsfunctie $F(x)$, waarschijnlijkheidspapier vervaardigen, waarop verdelingsfuncties van het type $F(\alpha x+\beta)$ (met α en β parameters) door een rechte lijn worden voorgesteld. Ook is het b.v. mogelijk waarschijnlijkheidspapier te maken, behorende bij verdelingen van het type $F(e^{\alpha x+\beta})$, door de ene schaal logarithmisch te transformeren en de andere volgens $F(x)$. Hetzelfde geldt trouwens voor verdelingen van het type $F(\varphi(\alpha x+\beta))$, waarin φ een willekeurige monotone functie is; de ene schaal wordt dan volgens φ en de andere volgens F getransformeerd.

De beschouwingen van de volgende paragrafen gelden alle voor ieder van deze soorten waarschijnlijkheidspapier.

2. Het doel van waarschijnlijkheidspapier.

Waarschijnlijkheidspapier kan o.a. gebruikt worden om op grond van een gegeven steekproef $x_1, \ldots, x_n$ op snelle wijze een schatting van de beide parameters μ en σ te verkrijgen, indien men de vorm der waarschijnlijkheidsverdeling van $\underline{x}$ kent, of om een visuele indruk te krijgen ter beantwoording van de vraag, òf deze grootheid $\underline{x}$ een verdeling van het beschouwde type bezit.

In beide gevallen zal men de waarnemingen $x_1, \ldots, x_n$ op het waarschijnlijkheidspapier uit moeten zetten. Dit kan nu op verschillende manieren geschieden. De moeilijkheid zit daarin, dat van ieder der punten $(x_i, F(x_i))$, die op het waarschijnlijkheidspapier uitgezet precies op een rechte lijn zouden liggen, alleen de coördinaat x_i bekend is, terwijl $F(x_i)$ onbekend is, daar μ en σ niet bekend zijn. Men gebruikt daarom in plaats van $F(x_i)$ zelf een schatting van $F(x_i)$ en hiervoor kan men verschillende functies gebruiken.

3. Enkele gangbare methoden.

Wij zullen in het volgende onderstellen, dat de waarnemingen $x_1, \ldots, x_n$ genummerd zijn volgens opklimmende grootte, terwijl er geen gelijke waarnemingen zijn (bij een continue verdeling van $\underline{x}$ is de kans op gelijke waarnemingen gelijk aan 0), zodat wij dus hebben

$$x_1 < x_2 < \cdots < x_n. \tag{3}$$

De verschillende schattingen, die voor $F(x_i)$ gebruikt worden, geven wij aan met $\varphi_1(i)$, $\varphi_2(i)$, etc., daar zij alle alleen van het rangnummer i van de beschouwde waarneming afhangen.

Wij beschouwen eerst de veelal gebruikelijke functie

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

Deze heeft het nadeel, dat $\varphi_1(n) = 1$ is, zodat het punt $(x_n, \varphi_1(n))$ buiten het papier valt. Immers $y' = 1$ staat op de verticale schaal bij $y = \infty$, daar $\Phi(\infty) = 1$ is en dit punt bevindt zich dus niet op het papier. Hetzelfde bezwaar, maar nu voor $i = 1$, geldt, indien men

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

gebruikt.

Dit bezwaar kan men op verschillende wijzen ondervangen, b.v. door de volgende functies te gebruiken

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

of

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

Beide functies worden wel gebruikt en er zijn natuurlijk nog talloze andere mogelijkheden. Voor een keuze uit de verschillende mogelijkheden dienen de eigenschappen van deze verschillende methoden onderzocht te worden. In alle gevallen zet men dus de punten

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

op het waarschijnlijkheidspapier uit en uit de vergelijkingen (4),...,(7) valt te zien, dat de 4 boven beschreven methoden asymptotisch voor $n \to \infty$ aequivalent zijn. Dit geldt ook voor de verderop nog in te voeren functies φ_5 en φ_6.

4. <u>De ligging van het punt $(x_i, \varphi(i))$</u>.

Daar $\underline{x}_i$, de i^{de} waarneming naar orde van grootte, een stochastische grootheid is, geldt hetzelfde ook voor iedere functie van $\underline{x}_i$ en dus in het bijzonder voor $F(\underline{x}_i)$. Het punt $(x_i, F(x_i))$ ligt voor iedere waarde van x_i op de kromme $y = F(x)$ in figuur 1, daar deze kromme de meetkundige plaats van deze punten is. Het stochastische punt $(\underline{x}_i, F(\underline{x}_i))$ heeft dus een waarschijnlijkheidsverdeling over deze kromme [2]). De stochastische grootheid

---

2) Deze waarschijnlijkheidsverdeling is niet dezelfde als de oorspronkelijke waarschijnlijkheidsverdeling van $\underline{x}$, daar $\underline{x}_i$ die i^e waarneming bij rangschikking naar grootte voorstelt.

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

bezit een waarschijnlijkheidsverdeling met als gemiddelde [3])

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

en als modus

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

Verder geldt voor de mediaan y_i^* van $\underline{y}_i$ bij benadering:

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

en wij zien, dat dit laatste getal steeds tussen de beide andere in ligt. Men kan bewijzen, dat dit niet alleen voor deze benadering van de mediaan geldt, maar ook voor de mediaan zelf [4]).

Wij beschouwen nu waarden van $i > \frac{n+1}{2}$ (voor waarden $< \frac{n+1}{2}$ gelden analoge conclusies). Voor een dergelijke i geldt dus

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

Dit betekent, dat de bij $\underline{x}_i$ behorende waarde $\underline{y}_i$ (zie (8)) vaker boven dan onder zijn gemiddelde $\mathcal{E}\underline{y}_i$ ligt, terwijl het omgekeerd de voor de modus geldt. Daar het punt $(\underline{x}_i, \underline{y}_i)$ steeds op de in figuur 1 getekende kromme ligt, zal het punt $(\underline{x}_i, \frac{i}{n+1})$ in meer dan de helft der gevallen onder deze kromme liggen en het punt $(\underline{x}_i, \frac{i-1}{n-1})$ in meer dan de helft der gevallen erboven. Dit blijft gelden, als men de verticale schaal volgens Φ (zie (2)) transformeert, omdat deze transformatie monotoon is, d.w.z. de volgorde van de punten in verticale richting niet verandert.

De functie φ_4 (zie (7)) heeft dus het bezwaar, dat voor $i > \frac{n+1}{2}$ de op het waarschijnlijkheidspapier uitgezette punten in meer dan de helft der gevallen onder de lijn liggen, die de onbekende waarschijnlijkheidsverdeling voorstelt, terwijl dit voor $i < \frac{n+1}{2}$ juist andersom is. Deze wijze van uitzetten heeft dus tot gevolg, dat men in meer dan de helft der gevallen de helling van de lijn te laag zal schatten, dus de spreiding te hoog. Daar komt nog bij, dat dit effect het sterkst is voor kleine en voor grote waarden van i, waardoor de op deze wijze uitgezette punten de neiging hebben op het waarschijnlijkheidspapier min of meer in een S-bocht te gaan liggen. Precies hetzelfde bezwaar,

---

3) Zie voor bewijzen en literatuurverwijzingen de appendix van dit artikel.

4) Het bewijs hiervan wordt in dit artikel niet gegeven. Het volgt op eenvoudige wijze uit C.G.LEKKERKERKER [4] (zie literatuurlijst).

maar nu juist andersom, geldt voor φ_3 (zie (6)), daar

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

en andersom voor $i < \frac{n+1}{2}$. Deze wijze van uitzetten leidt dus tot een onderschatting van σ in meer dan de helft der gevallen en eveneens tot S-bochten, maar nu in de andere richting.

Gebruikt men nu echter

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

dan bestaan deze bezwaren niet, daar de punten $(x_i, \frac{i-0,3}{n+0,4})$ voor iedere i ongeveer even vaak boven als onder de gezochte lijn zullen liggen.

Het bovenstaande wordt geïllustreerd door fig. 2. Daarin stelt de rechte lijn de verdelingsfunctie $y = F(x)$ voor (men moet de figuur dus op waarschijnlijkheidspapier getekend denken), x_i^* stelt de mediaan van $\underline{x}_i$ en y_i^* die van $\underline{y}_i$ voor. Het punt (x_i^*, y_i^*) ligt op de rechte $y = F(x)$ (zie ook opmerking 2 hieronder) en het punt $(\underline{x}_i, \underline{y}_i)$, dat een waarschijnlijkheidsverdeling op de lijn $y = F(x)$ bezit, ligt op deze lijn even vaak links als rechts van dit punt.

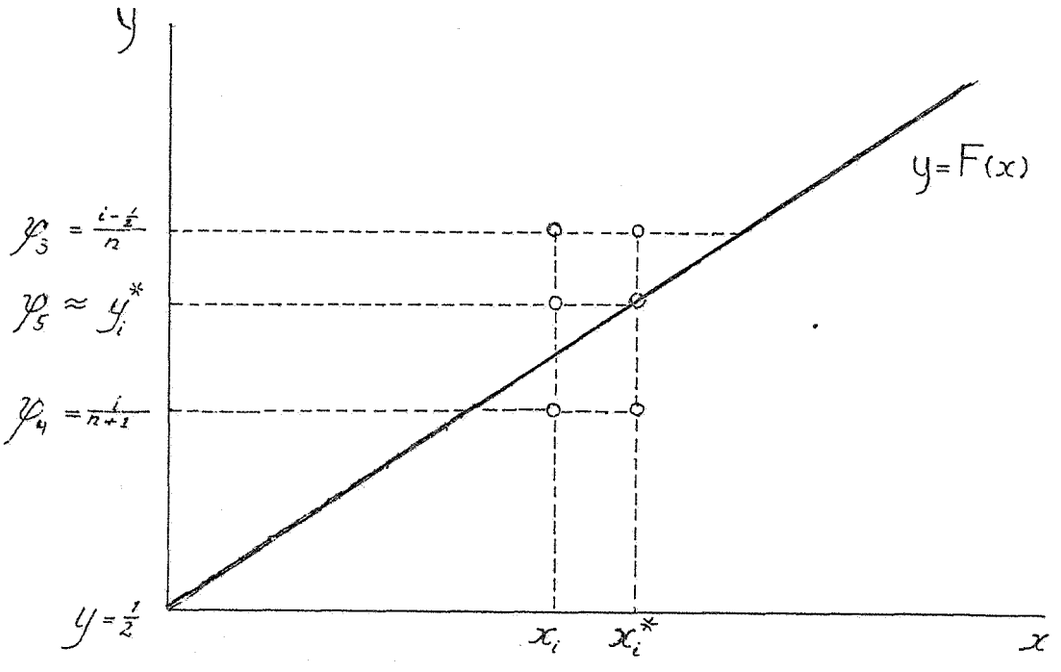

Fig. 2. Drie methoden voor het uitzetten van waarnemingen op waarschijnlijkheidspapier.

In de figuur ziet men, dat het punt $(x_i, y_3(i))$ even vaak links van $(x_i^*, y_3(i))$ als rechts daarvan komt te liggen en dus vaker boven $y = F(x)$ dan daaronder ligt en omgekeerd voor $y_4(i)$. Het in de figuur getekende punt x_i ligt links van x_i^*, terwijl toch $(x_i, y_4(i))$ nog onder de lijn ligt in plaats van erboven.

Opmerkingen.

1) Een zesde mogelijkheid

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

waarin het rechterlid dus de modus van y_i voorstelt, verenigt de bezwaren van φ_1, φ_2 en φ_3 in zich en kan dus niet aanbevolen worden.

2) Daar y_i een monotone functie is van x_i (immers $y_i = F(x_i)$) vinden we een benadering voor de mediaan van x_i door de waarde $y_i = \frac{i-0.3}{n+0.4}$ te substitueren in de inverse functie $x_i = F^{-1}(y_i)$. Dus $\mathrm{Med}\, x_i \approx F^{-1}\left(\frac{i-0.3}{n+0.4}\right)$. Vgl. fig. 1.

3) Treden onder de waarnemingen door groepering of afronding gelijken op, dan kan men wellicht het beste de methode der gemiddelde rangnummers toepassen. Deze bestaat daarin, dat men alle waarnemingen uit een groep van gelijken als rangnummers het gemiddelde toekent van de rangnummers, die deze waarnemingen gehad zouden hebben, indien zij ongelijk waren geweest, maar ten opzichte van alle niet tot die groep behorende waarnemingen dezelfde positie zouden hebben ingenomen bij rangschikking naar grootte als nu het geval is. Deze gemiddelde rangnummers vult men dan in de formules voor φ in.

4) De hier beschreven methoden putten natuurlijk de mogelijkheden geenszins uit. E.J.GUMBEL [2] b.v. geeft voor het door hem ontworpen waarschijnlijkheidspapier voor de uiterste waarden (de dubbel-exponentiële verdeling) een methode aan, waarbij in verticale richting F (mod x_i) wordt uitgezet. Deze methode is voor het door hem beschouwde geval asymptotisch aequivalent met het gebruik van φ_6, echter zonder het bezwaar, dat de eerste en de laatste waarneming niet uitgezet kunnen worden.

5. Appendix.

Geven wij de verdelingsfunctie van y_i aan met G_i, dan geldt (zie b.v. D.van DANTZIG [1], hoofdstuk IV, par. 2):

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

hetgeen ook geschreven kan worden in de vorm

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

(zie b.v. M.G.KENDALL [3], p. 120). Door de tweede afgeleide van G_i naar y gelijk aan nul te stellen, vindt men de modus van y_i, dus (10). Ook de verwachting van y_i, (9), volgt op de gebruikelijke wijze uit (15).

De exacte waarden van y_i^*, de mediaan van y_i, kan men vinden met behulp van tabellen van de onvolledige bêta-functie (b.v. C.M.THOMPSON [6]).

De benaderingsformule (11) leiden we als volgt af: voor iedere waarde van de verdelingsfunctie $\mathcal{G}$ geldt de volgende betrekking:

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

d.w.z. indien de waarnemingen niet naar opklimmende maar naar dalende grootte gerangschikt worden, dan wordt i door $n+1-i$, y door $1-y$ en G door $1-G$ vervangen. Het ligt voor de hand ook van φ te verlangen dat aan deze symmetrierelatie voldaan is, d.w.z. dat geldt:

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

Hieraan wordt door φ_1 en φ_2 niet voldaan, maar wel door de overige in het voorgaande genoemde functies φ.

Is nu y_i^* de mediaan van $\underline{y}_i$, zodat

$$G_i(y_i^*) = 1/2,$$

dan is dus

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

dus $1-y_i^*$ is de mediaan van $\underline{y}_{n+1-i}$ [5]).

Wij schrijven y_i^* nu als

(17) $$y_i^* = \frac{i-a}{n+b}$$ (a en b zijn functies van i en n)

Uit (16) en (17) volgt de relatie

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

en hieruit volgt

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

Formule (17) gaat hierdoor over in

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

terwijl a nog zou moeten voldoen aan de betrekking

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

waarvoor wij ook kunnen schrijven

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

---

5) Asymptotisch voor $n \to \infty$ geldt $y_i^* = \frac{i}{n}$.

De a, die aan deze betrekking voldoet, kan weer gevonden worden met behulp van tabellen van de onvolledige bêta-functie, en deze a hangt uiteraard zowel van i als van n af. Wij zoeken nu een constante als benadering voor a, die voor iedere i en n voldoende nauwkeurig is voor practisch gebruik en nemen daartoe in het linkerlid de limiet voor $n \to \infty$:

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

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

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

Deze limiet stellen wij nu gelijk aan $\frac{1}{2}$.
Dit betekent dus, dat wij een Poisson-verdeling zoeken, met gemiddelde $i-a$ en mediaan i. Voor iedere waarde van i kunnen wij nu, met behulp van een tabel van Poisson-verdelingen (b.v. E.C.MOLINA [5]) de bijbehorende waarden van a bepalen. Het resultaat hiervan vindt men in tabel I.

Tabel I

Waarden van a voor verschillende i en $n \to \infty$.

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

Daar wij één vaste waarde voor a willen gebruiken en daar voor $i=1$ en $i=n$ de bijbehorende punten $(x_i, \varphi(i))$ gewoonlijk de grootste afwijkingen van de rechte lijn vertonen, kiezen we de daarbij behorende waarde van a, die wij voor het gemak op 0,3 afronden.

Om de invloed van de boven gebruikte limietovergang voor $n \to \infty$ na te gaan, hebben wij voor $i=1$ en voor enkele kleine waarden van n de exact waarde van $G_1\left(\frac{1-a}{n+1-2a}\right)$ met $a = 0,3$ bepaald, die dus, als a precies juist was, gelijk aan $\frac{1}{2}$ zou moeten zijn.

De resultaten, samengevat in tabel II, zijn gunstig.

Tabel II

Exacte waarden van $G_1\left(\frac{0,7}{n+0,4}\right)$.

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

Ten slotte hebben wij, voor $n = 10$ en $n = 15$, de waarden vergeleken met de exacte medianen, die uit de tabel [5] bepaald kunnen worden en bij gebruik waarvan voor iedere i de kans, dat het punt $(x_i, \varphi(i))$ boven de rechte lijn komt te liggen, precies gelijk aan $\frac{1}{2}$ zou zijn (maar die niet door een constante α weergegeven kunnen worden). Deze waarden bleken vrijwel steeds tot in drie decimalen overeen te stemmen, terwijl het verschil voor geen enkele i meer dan 1% bedroeg. Hieruit kunnen wij dus concluderen, dat de benadering voor practische doeleinden ruim voldoende is en dat men door de punten $\left(x_i, \frac{i-0,7}{n+0,4}\right)$ op waarschijnlijkheidspapier uit te zetten bereikt, dat ieder der punten gelijke kans bezit om boven of onder de gezochte lijn te komen liggen.

Literatuur.

[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 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.