

EINE NEUE METHODE IN DER THEORIE DER GEORDNETEN STICHPROBEN

Von ALFRED RÉNYI in Budapest

In der mathematischen Statistik erwirbt die Theorie der geordneten Stichproben, die sog. Ordnungsstatistik oder, anders ausgedrückt: die Theorie der Variationsreihen, eine ständig wachsende Bedeutung. Die Bedeutung der Methoden der Ordnungsstatistik wird durch den Umstand gegeben, daß ihre Resultate auf statistische Gesamtheiten mit beliebiger (kontinuierlicher) Verteilung angewendet werden können; die Methoden der Ordnungsstatistik sind daher sog. nichtparametrische Methoden, deren Anwendungsbereich ein viel weiterer ist als der der parametrischen Methoden, die nur im Falle von Verteilungen eines gewissen angegebenen Typs verwendet werden können.

Seit dem Beginn des gegenwärtigen Jahrhunderts haben sich mehrere Autoren, insbesondere PEARSON [1], v. BORTKIEWICZ [2], DODD [3], TIPPET [4], FRÉCHET [5], mit solchen Fragen beschäftigt, die als zu dem Bereich der Ordnungsstatistik gehörig betrachtet werden können. Zu einer wirklich umfassenden Theorie wurde die Ordnungsstatistik von sowjetischen Mathematikern, so in erster Linie von KOLMOGOROV [6], GLIVENKO [7], SMIRNOV [8], GNEDENKO [9] und ihren Schülern entwickelt.

Besonders in den letzten drei Jahren beschäftigten sich zahlreiche Artikel mit diesem Problemkreis; unter diesen wollen wir die Arbeiten von GNEDENKO und KOROLJUK [10], GNEDENKO und RVAČEVA [11], GNEDENKO und MICHALEVIČ [12], MICHALEVIČ [13], KVIČ [14], MANIJA [15], GICHMAN [16] hervorheben. An der Weiterentwicklung der Ergebnisse der sowjetischen Mathematiker haben sich auch zahlreiche amerikanische Mathematiker beteiligt: unter diesen möchten wir die Arbeiten von FELLER [17], DOOB [18], MASSEY [19], DONSKER [20], ANDERSON und DARLING [21] hervorheben.

Die Theorie der geordneten Stichproben hat mannigfaltige wichtige praktische Anwendungen, z. B. im Zusammenhang mit der statistischen Qualitätskontrolle; aber darauf kann ich hier nicht eingehen.

Der Zweck meines Vortrages ist, eine neue Methode mitzuteilen, mit deren Hilfe die heute bereits klassischen Resultate der Theorie der Ordnungsstatistik mit überraschender Einfachheit gewonnen werden können und die auch den Beweis neuer Sätze ermöglicht. Das Wesen der Methode besteht darin, daß sie die Probleme der Ordnungsstatistik auf die Untersuchung gewisser Markoffscher Vorgänge bzw. Markoffscher Ketten zurückführt.

Im folgenden benütze ich die wohlbekannte Kolmogoroffsche Theorie der Wahrscheinlichkeitsrechnung, welche heute ganz allgemein angenommen ist, d. h., ich nehme

an, es sei gegeben: 1. eine Grundmenge M von abstrakten Elementen a, b, c, \dots , Elementarereignisse genannt; 2. ein Borelscher Mengenkörper K von Untermengen A, B, C, \dots von M , schlechthin Ereignisse genannt, unter denen M vorkommt; 3. eine total additive und nichtnegative Mengenfunktion $P(A)$, definiert auf K , die Wahrscheinlichkeit von A , mit $P(M) = 1$. Mit anderen Worten: es ist ein Wahrscheinlichkeitsfeld (M, K, P) gegeben. Zufällige Größen oder Zufallsveränderliche nennt man die Funktionen $\xi = \xi(a)$ (a wird nicht angeschrieben), die auf M definiert und bezüglich K meßbar sind. Jede solche zufällige Veränderliche besitzt eine Verteilungsfunktion $F(x) = P(\xi < x)$; der Mittelwert $M(\xi)$ von ξ wird als (abstraktes) Lebesguesches Integral von ξ bezüglich P definiert: $M(\xi) = \int_M \xi(a) dP$ und das Streuungsquadrat $D^2(\xi)$ von ξ als $M[(\xi - M(\xi))^2] = D^2(\xi)$. Ferner heißen ξ und η unabhängig, wenn $P(\xi < x, \eta < y) = P(\xi < x) P(\eta < y)$ für alle x und y gilt. Mit $P(A | B)$ bezeichnen wir die bedingte Wahrscheinlichkeit von A unter der Bedingung B .

Wir gehen von dem folgenden Spezialfall aus: es seien die Ergebnisse von n unabhängigen, auf den Wert einer exponentiell verteilten Zufallsveränderlichen ζ bezüglichen – oder, mit anderen Worten, aus einer exponentiell verteilten Gesamtheit genommenen – Beobachtungen gegeben, die durch $\zeta_1, \zeta_2, \dots, \zeta_n$ bezeichnet werden mögen; mit anderen Worten: es seien $\zeta_1, \zeta_2, \dots, \zeta_n$ unabhängige und die gleiche exponentielle Verteilungsfunktion besitzende Zufallsveränderliche, d. h.

$$P(\zeta_k < x) = 1 - e^{-\lambda x} \quad (x \geq 0; k = 1, 2, \dots, n).$$

Wir werden die folgende wohlbekannte Eigenschaft der exponentiellen Verteilung benötigen: besitzt ζ eine exponentielle Verteilung, so ist (wenn $x > 0$ und $y \geq 0$)

$$(1) \quad P(\zeta < y + x | \zeta \geq y) = P(\zeta < x).$$

Diese Eigenschaft kennzeichnet übrigens die exponentielle Verteilungen eindeutig. Es besitze nämlich ζ die Verteilungsfunktion $F(x)$, d. h. $F(x) = P(\zeta < x)$, so ist nach der Definition der bedingten Wahrscheinlichkeit

$$P(\zeta < y + x | \zeta \geq y) = \frac{F(y + x) - F(y)}{1 - F(y)}$$

und daher (1) mit der folgenden Relation äquivalent:

$$(2) \quad \Phi(x + y) = \Phi(x) \Phi(y),$$

worin $\Phi(x) = 1 - F(x)$ ist. Andererseits gilt jedoch (2) unter allen die Bedingung $0 \leq \Phi(x) \leq 1$ erfüllenden Funktionen, abgesehen von den trivialen Fällen $\Phi(x) \equiv 0$ und $\Phi(x) \equiv 1$, nur für die Funktionen $\Phi(x) = e^{-\lambda x}$ ($\lambda > 0$).

Besonders einleuchtend wird die Bedeutung von (1), wenn man die Zufallsveränderliche ζ als die zufällige Dauer eines Ereignisses interpretiert. Bei dieser Interpretation läßt sich die Behauptung (1) so formulieren, daß im Falle eines exponentiell verteilte Dauer besitzenden Ereignisses, das in einem Zeitpunkt $y \geq 0$ noch andauert, die Frage, wie lange dasselbe noch dauern wird, von y , d. h. von dem Umstande, wie lange dasselbe bereits andauert, unabhängig ist. Als Beispiel soll erwähnt werden, daß für die Zeitdauer zwischen zwei nacheinander folgenden Ereignissen in einem Poissonschen Prozeß (z. B. in einem radioaktiven Zerfallsprozeß) die

Bedingung erfüllt ist. Auf diese Weise ergibt es sich, daß, wenn wir die Zahlen $\zeta_1, \zeta_2, \dots, \zeta_n$ ihrer Größe nach ordnen und die Bezeichnungen

$$(3) \quad \zeta_k^* = O_k(\zeta_1, \zeta_2, \dots, \zeta_n) \quad (k = 1, 2, \dots, n)$$

einführen, wo die Funktion $O_k(x_1, x_2, \dots, x_n)$ von n Veränderlichen die der Größe nach k -te unter den Zahlen x_1, x_2, \dots, x_n bedeutet ($k = 1, 2, \dots, n$), die Verteilung der Elemente und die Verteilung der Gesamtheit der Elemente der *geordneten Stichprobe* $\zeta_1^* \leq \zeta_2^* \leq \dots \leq \zeta_n^*$ sehr einfach bestimmt werden kann.

Zu diesem Zwecke wollen wir die Veränderlichen ζ_k als die Dauer von n voneinander unabhängigen und die gleiche exponentielle Verteilung besitzenden zufälligen Ereignissen definieren; ζ_k^* bedeutet die Dauer jenes Ereignisses, das als k -tes endet. Es sei $t > 0$ ein beliebiger Zeitpunkt, und es bedeute ϑ_t die Anzahl der Ereignisse, die noch im Zeitpunkt t im Gange sind. Dann ist $\{\vartheta_t\}$ ein Markoffscher Prozeß, denn

$$(4) \quad P(\vartheta_{t+\tau} = j \mid \vartheta_t = k, \vartheta_{t-\delta} = l) = \binom{k}{j} (1 - e^{-\lambda\tau})^j e^{-\lambda\tau(k-j)}$$

hängt nicht davon ab, was für einen Wert $\vartheta_{t-\delta}$ besitzt, d. h., die Vergangenheit des Prozesses ist von Einfluß auf die Zukunft nur durch den gegenwärtigen Zustand. Anders ausgedrückt: die Veränderlichen $\zeta_1^*, \zeta_2^*, \dots, \zeta_n^*$ selbst bilden eine Markoffsche Kette; um das zu zeigen, werden wir vor allem die Verteilungen der Differenzen $\zeta_{k+1}^* - \zeta_k^*$ berechnen. Ist $\zeta_k^* = y$, so bedeutet in

$$(5) \quad P(\zeta_{k+1}^* - \zeta_k^* > x \mid \zeta_k^* = y) = P(\zeta_{k+1}^* > x + y \mid \zeta_k^* = y),$$

die auf der rechten Seite stehende Wahrscheinlichkeit, daß unter den im Zeitpunkte y noch andauernden $(n - k)$ Ereignissen kein einziges bis zu dem Zeitpunkt $x + y$ endigt. Andererseits ist die Wahrscheinlichkeit dieses Ereignisses gemäß (1):

$$P(\zeta > x)^{n-k} = e^{-(n-k)\lambda x},$$

und somit ist die bedingte Verteilungsfunktion von $\zeta_{k+1}^* - \zeta_k^*$ unter der Bedingung $\zeta_k^* = y$:

$$(6) \quad P(\zeta_{k+1}^* - \zeta_k^* < x \mid \zeta_k^* = y) = 1 - e^{-(n-k)\lambda x}.$$

Da die Verteilung (6) von y nicht abhängt, ergibt sie gleichzeitig auch die unbedingte Verteilungsfunktion von $\zeta_{k+1}^* - \zeta_k^*$, d. h.

$$(7) \quad P(\zeta_{k+1}^* - \zeta_k^* < x) = 1 - e^{-(n-k)\lambda x}.$$

Die Differenzen $\zeta_{k+1}^* - \zeta_k^*$ selbst besitzen somit ebenfalls eine exponentielle Verteilung mit einem Erwartungswert $\frac{1}{(n-k)\lambda}$ und die Veränderlichen

$$(8) \quad \delta_{k+1} = (n - k)(\zeta_{k+1}^* - \zeta_k^*) \quad (k = 0, 1, 2, \dots, n - 1; \zeta_0^* = 0)$$

insbesondere mit dem gleichen Erwartungswert $\frac{1}{2}$.

Aus dem Gesagten folgt auch, daß die Veränderlichen

$$\delta_{k+1} = (n - k)(\zeta_{k+1}^* - \zeta_k^*)$$

unabhängig sind. Somit können die Veränderlichen ζ_k^* in der Form

$$(9) \quad \zeta_k^* = \frac{\delta_1}{n} + \frac{\delta_2}{n-1} + \cdots + \frac{\delta_k}{n-k+1} \quad (k = 1, 2, \dots, n)$$

als aus unabhängigen und die gleiche Verteilung besitzenden Veränderlichen gebildete lineare Ausdrücke dargestellt werden. (9) läßt sich auch so ausdrücken, daß die Veränderlichen $\zeta_1^*, \zeta_2^*, \dots, \zeta_n^*$ eine Markoffsche Kette, und zwar eine sog. additive Kette oder – anders ausgedrückt – eine Markoffsche Kette mit unabhängigen Zuwächsen bilden. Mit Hilfe von (9) läßt sich die Verteilung von ζ_k^* bzw. die gemeinsame Verteilung einer beliebigen Anzahl von Veränderlichen unter den ζ_k^* auch explizit bestimmen, ferner können die auf die Verteilung der Veränderlichen bezüglichen Grenzwertsätze mit Hilfe von (9) auf den zentralen Grenzwertungssatz der Wahrscheinlichkeitsrechnung zurückgeführt werden.

Betrachten wir nun die Frage, auf welche Weise das Gesagte auf die Untersuchung von geordneten Stichproben aus Gesamtheiten mit beliebiger Verteilung angewendet werden kann. Es sei ξ eine beliebige, kontinuierliche Verteilung besitzende Zufallsveränderliche, deren Verteilungsfunktion wir mit $F(x)$ bezeichnen wollen. Es sei $(\xi_1, \xi_2, \dots, \xi_n)$ eine aus unabhängigen Beobachtungen bestehende (auf den Wert der Veränderlichen ξ bezügliche), n Elemente enthaltende Stichprobe, d. h., es seien $\xi_1, \xi_2, \dots, \xi_n$ voneinander in ihrer Gesamtheit unabhängige und die gleiche (kontinuierliche) Verteilungsfunktion $F(x)$ besitzende Zufallsveränderliche. Wir ordnen nun die Stichprobenelemente ξ_k ihrer Größe nach, d. h., wir bilden die neuen Veränderlichen $O_k(\xi_1, \xi_2, \dots, \xi_n) = \xi_k^*$ ($k = 1, 2, \dots, n$).

Die Aufgabe der Ordnungsstatistik besteht in der Untersuchung der Veränderlichen ξ_k^* ; dies läßt sich jedoch auf den Spezialfall zurückführen, in dem die Veränderlichen ξ_k^* eine exponentielle Verteilung besitzen, und somit – gemäß Formel (9) – auf die Untersuchung von Summen unabhängiger Zufallsveränderlichen, und zwar folgendermaßen: es sei

$$(10) \quad \zeta_k = \log \frac{1}{F(\xi_k)} \quad (k = 1, 2, \dots, n)$$

und

$$(11) \quad \zeta_k^* = \log \frac{1}{F(\xi_{n+1-k}^*)} \quad (k = 1, 2, \dots, n).$$

Da $\log \frac{1}{F(x)}$ eine monoton abnehmende Funktion ist, so ist

$$(12) \quad \zeta_k^* = O_k(\zeta_1, \zeta_2, \dots, \zeta_n).$$

Untersuchen wir nun die Verteilung der Veränderlichen ζ_k . Da im Falle $x > 0$, wenn man die inverse Funktion von $x = F(y)$ mit $y = F^{-1}(x)$ bezeichnet,

$$(13) \quad P(\zeta_k < x) = P(F(\xi_{n+1-k}) > e^{-x}) = P(\xi_{n+1-k} > F^{-1}(e^{-x}))$$

und somit

$$(14) \quad P(\zeta_k < x) = 1 - F(F^{-1}(e^{-x})) = 1 - e^{-x}$$

ist, sind die Veränderlichen ζ_k ($k = 1, 2, \dots, n$) unabhängige und die gleiche exponentielle Verteilung besitzende Zufallsveränderliche, und zwar mit einem Er-

wartungswert $\frac{1}{\lambda} = 1$. Auf diese Weise können die Zufallsveränderlichen ξ_k^* selbst in der Form

$$(15) \quad \xi_k^* = F^{-1} \left[e^{-\left(\frac{\delta_1}{n} + \frac{\delta_2}{n-1} + \dots + \frac{\delta_{n+1-k}}{k} \right)} \right]$$

dargestellt werden, wobei $\delta_1, \delta_2, \dots, \delta_n$ unabhängige und die gleiche exponentielle Verteilungsfunktion $1 - e^{-x}$ besitzende Zufallsveränderliche sind. Aus unserem Resultat folgt auch, daß die Quotienten

$$(16) \quad \frac{F(\xi_{k+1}^*)}{F(\xi_k^*)}$$

voneinander unabhängig sind und daß auf diese Weise die Veränderlichen $\eta_k^* = F(\xi_k^*)$ und somit die Veränderlichen ξ_k^* selbst eine (endliche) Markoffsche Kette bilden. Die Veränderlichen $\eta_k = F(\xi_k)$ besitzen offenbar eine gleichförmige Verteilung im Intervall $(0, 1)$, und somit sind die Veränderlichen η_k^* nichts anderes als die ihrer Größe nach geordneten Elemente einer aus n Elementen bestehenden, aus einer in dem Intervall $(0, 1)$ gleichförmige Verteilung besitzenden Gesamtheit genommenen Stichprobe.

Um zu illustrieren, auf welche Weise man auf Grund des Gesagten die bekannten Ergebnisse der Ordnungsstatistik auf sehr einfache Weise erhalten kann, beweisen wir den folgenden

Satz. Wenn $F'(x) = f(x)$ existiert, $0 < p < 1$, $F(q_p) = p$ und $f(q_p) > 0$ und ferner $f(x)$ in einer Umgebung des Punktes q_p kontinuierlich ist, ferner wenn $n \rightarrow \infty$ und $|k(n) - np| < \text{konst.}$ ist, so besitzt $\xi_{k(n)}^*$, d. h. das p -Quantil der Stichprobe im Grenzwert, eine normale Verteilung mit einem Erwartungswert q_p und mit einer Streuung $\frac{1}{f(q_p)} \sqrt{\frac{p(1-p)}{n}}$, mit anderen Worten

$$(17) \quad \lim_{n \rightarrow \infty} P \left(\frac{\xi_{k(n)}^* - q_p}{\frac{1}{f(q_p)} \sqrt{\frac{p(1-p)}{n}}} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Es ist nämlich

$$(18) \quad M \left(\frac{\delta_1}{n} + \dots + \frac{\delta_{n+1-k(n)}}{k(n)} \right) = \frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{k(n)} \approx \log \frac{n}{k(n)} \approx \log \frac{1}{p},$$

ferner (da auch die Streuung der einen Erwartungswert 1 besitzenden exponentiellen Verteilung 1 beträgt):

$$(19) \quad D^2 \left(\frac{\delta_1}{n} + \dots + \frac{\delta_{n+1-k(n)}}{k(n)} \right) = \frac{1}{n^2} + \dots + \frac{1}{k^2(n)} \approx \frac{1}{k(n)} - \frac{1}{n} \approx \frac{1-p}{np},$$

und schließlich, da für die Summe $\frac{\delta_1}{n} + \dots + \frac{\delta_{n+1-k(n)}}{k(n)}$ die Bedingungen des Satzes von LJAPUNOFF erfüllt sind, so besitzt

$$\sqrt{\frac{np}{1-p}} \log \frac{F(\xi_{k(n)}^*)}{p} = \sqrt{\frac{np}{1-p}} \log \frac{F(\xi_{k(n)}^*)}{F(q_p)}$$

im Grenzwert eine normale Verteilung mit einem Erwartungswert 0 und einer Streuung 1. Andererseits ist aber

$$\frac{F(\xi_k^*(n))}{F(q_p)} = 1 + \frac{(\xi_k^*(n) - q_p)}{p} f(q_p + \vartheta(\xi_k^*(n) - q_p)) \quad (0 < \vartheta < 1),$$

und so besitzt im Sinne unserer Voraussetzung auch

$$\frac{\xi_k^*(n) - q_p}{\frac{1}{f(q_p)} \sqrt{\frac{p(1-p)}{n}}}$$

im Grenzwert eine normale Verteilung mit dem Erwartungswert 0 und der Streuung 1. (Auf gewisse triviale Einzelheiten der Berechnung gehen wir nicht ein.) Speziell, wenn $p = \frac{1}{2}$ ist, so besagt dieser Satz, auf welche Weise man aus dem empirischen Median der Stichprobe auf den unbekanntem Median der Verteilung schließen kann. Auf ähnliche Weise läßt sich auch der bekannte Satz erhalten, laut dessen zwischen $\xi_{k_1}^*(n)$ und $\xi_{k_2}^*(n)$ im Grenzwert eine normale Korrelation besteht, wenn $\lim_{n \rightarrow \infty} \frac{k_1(n)}{n} = p_1$ und $\lim_{n \rightarrow \infty} \frac{k_2(n)}{n} = p_2$ ($0 < p_1 < p_2 < 1$) ist, ferner, daß $\xi_{k_1}^*(n)$ und $\xi_{k_2}^*(n)$ im Grenzwert unabhängig sind, wenn $\lim_{n \rightarrow \infty} \frac{k_1(n)}{n} = 0$ und $\lim_{n \rightarrow \infty} \frac{k_2(n)}{n} = 1$ ist. Außerdem können noch zahlreiche andere bekannte Sätze und deren Verschärfungen erhalten werden. Auf Grund von (12) werden die sich auf die Grenzverteilung der sogenannten „extremen“ Glieder beziehenden Sätze, d.h. die auf die Grenzverteilung von ξ_k^* und ξ_{n+1-k}^* bezüglichen Sätze, wo $n \rightarrow \infty$ und k konstant ist, geradezu evident.

Der wesentliche Vorteil unserer Methode besteht darin, daß sie uns ermöglicht, alle diese Sätze mit Hilfe der Formel (9) aus den bekannten Sätzen des am weitesten entwickelten Kapitels der Wahrscheinlichkeitsrechnung, der Theorie der Grenzverteilung für Summen unabhängiger Zufallsveränderlicher, abzuleiten.

Mit dem bisher Gesagten haben wir nur gezeigt, auf welche Weise es möglich ist, mittels unserer Methode gewisse wohlbekannte klassische Resultate der Ordnungsstatistik mittels eines einheitlichen Verfahrens auf einfache Art zu beweisen und zu verschärfen. Jetzt möchte ich noch einige Worte darüber sagen, welche neuen Ergebnisse mit der gleichen Methode gewonnen werden können.

A. N. KOLMOGOROV und N. V. SMIRNOV haben zwei grundlegende Sätze bewiesen, die Kriterien hinsichtlich der Frage bieten, ob eine Stichprobe aus einer Gesamtheit mit gegebener Verteilung stammen kann, d.h. auf welche Weise man aus der Verteilung der Elemente der Stichprobe auf die unbekanntem Verteilung der ganzen Gesamtheit folgern kann. Es sei

$$(20) \quad F_n(x) = \begin{cases} 0 & (x \leq \xi_1^*), \\ \frac{k}{n} & (\xi_k^* < x \leq \xi_{k+1}^*; k = 1, 2, \dots, n-1), \\ 1 & (\xi_n^* < x), \end{cases}$$

d.h., es möge $F_n(x)$ die empirische Verteilungsfunktion der Stichprobe bezeichnen. Mit anderen Worten: $F_n(x)$ ist die relative Häufigkeit jener Elemente der Stichprobe,

die kleiner als x sind. $F_n(x)$ ist also eine Treppenfunktion, die in den Punkten $\xi_1, \xi_2, \dots, \xi_n$ die Sprünge $\frac{1}{n}$ hat. Die Sätze SMIRNOVS und KOLMOGOROVs lauten folgendermaßen:

$$(21) \quad \begin{cases} \lim_{n \rightarrow \infty} P \left(\sqrt{n} \sup_{-\infty < x < \infty} [F_n(x) - F(x)] < y \right) = \begin{cases} 1 - e^{-2y^2} & (\text{für } y > 0), \\ 0 & (\text{für } y \leq 0), \end{cases} \\ \lim_{n \rightarrow \infty} P \left(\sqrt{n} \sup_{-\infty < x < \infty} |F_n(x) - F(x)| < y \right) = \begin{cases} \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2} & (\text{für } y > 0), \\ 0 & (\text{für } y \leq 0). \end{cases} \end{cases}$$

Diese Sätze geben daher die Grenzverteilungen des Maximums bzw. des Maximums des absoluten Wertes der Abweichung zwischen der empirischen und der theoretischen Verteilung an, wobei diese Grenzverteilungen nicht von der Verteilung $F(x)$ abhängen, von der man zur Gültigkeit des Satzes nur so viel voraussetzen muß, daß dieselbe kontinuierlich ist. Naturgemäß ergibt sich das Problem, die Grenzverteilung des Maximums bzw. des Maximums des absoluten Wertes, der *relativen* (d. h. auf den Wert der theoretischen Verteilung bezogenen) Abweichung zwischen der empirischen und der theoretischen Verteilung zu untersuchen. Diese Fragestellung ist in mancher Hinsicht noch natürlicher, da man in der Praxis immer relative Fehler betrachtet. Es gelang, diese Probleme mit Hilfe der in meinem Vortrage skizzierten Methode zu lösen, welche zur Lösung dieses Problems durchaus geeignet ist. Im Laufe der Lösung des Problems mußte eine natürliche Beschränkung eingeführt werden: da $F(x)$ auch beliebig kleine Werte annehmen kann, so ist es nicht zweckmäßig, das Maximum von $\frac{F_n(x) - F(x)}{F(x)}$ in dem ganzen Intervall $-\infty < x < +\infty$ zu untersuchen, sondern nur in irgendeinem Intervall $x_a \leq x < +\infty$, wo x_a durch die Bedingung $F(x_a) = a > 0$ definiert ist; der Wert von a kann jedoch eine beliebig kleine positive Zahl sein. Das ist eine Einschränkung, die vom praktischen Standpunkt aus natürlich ist, da wir im allgemeinen wirklich nur ein begrenztes Beobachtungsintervall haben. Hinsichtlich des in Rede stehenden Intervalles habe ich die folgenden Resultate bewiesen:

Satz 1.

$$(22) \quad \lim_{n \rightarrow \infty} P \left(\sqrt{n} \sup_{a \leq F(x) \leq 1} \left[\frac{F_n(x) - F(x)}{F(x)} \right] < y \right) = \begin{cases} \sqrt{\frac{2}{\pi}} \int_0^{y \sqrt{\frac{a}{1-a}}} e^{-\frac{t^2}{2}} dt & (\text{wenn } y > 0), \\ 0 & (\text{wenn } y \leq 0). \end{cases}$$

Satz 2.

$$(23) \quad \lim_{n \rightarrow \infty} P \left(\sqrt{n} \sup_{a \leq F(x) \leq 1} \left| \frac{F_n(x) - F(x)}{F(x)} \right| < y \right) = \begin{cases} \frac{4}{\pi} \sum_{k=0}^{\infty} (-1)^k e^{-\frac{(2k+1)^2 \pi^2 (1-a)}{8 y^2 a}} \frac{1}{2k+1} & (\text{wenn } y > 0), \\ 0 & (\text{wenn } y \leq 0). \end{cases}$$

Ich habe auch die Grenzverteilung des Maximums bzw. des Maximums des absoluten Wertes von $\frac{F_n(x) - F(x)}{F(x)}$ in dem Intervall $x_a \leq x < x_b$ untersucht, worin $F(x_a) = a$ und $F(x_b) = b$ ($0 < a < b \leq 1$) ist. Der auf das Maximum bezügliche Satz ist der folgende

Satz 3.

$$(24) \lim_{n \rightarrow \infty} P \left(\sqrt{n} \sup_{x_a \leq x \leq x_b} \left[\frac{F_n(x) - F(x)}{F(x)} \right] < y \right) = \frac{1}{\pi} \sqrt{\frac{b}{1-b}} \int_{-\infty}^y e^{-\frac{bx^2}{2(1-b)}} \int_0^{(y-x)\sqrt{\frac{ab}{b-a}}} e^{-\frac{t^2}{2}} dt dx.$$

Ich möchte nur auf eine sehr überraschende Konsequenz des Satzes 3 hinweisen. Es folgt aus dem Satz von SMIRNOV, daß

$$(25) \lim_{n \rightarrow \infty} P \left(\sup_{-\infty < x < +\infty} (F_n(x) - F(x)) < 0 \right) = 0,$$

d. h. die Wahrscheinlichkeit dafür, daß die empirische Verteilungsfunktion in ihrem ganzen Verlauf unterhalb der theoretischen Verteilungsfunktion verläuft, strebt gegen 0, wenn $n \rightarrow \infty$. Es folgt aus dem Satz 1, daß

$$(26) \lim_{n \rightarrow \infty} P \left(\sup_{x_a < x < +\infty} (F_n(x) - F(x)) < 0 \right) = 0,$$

d. h., dasselbe besteht auch, wenn man nur das Intervall ($x_a < x < +\infty$) untersucht: durch Vergrößerung des Umfanges der Stichprobe wird es immer seltener, daß $F_n(x)$ im ganzen Verlauf im Intervalle $x_a \leq x < +\infty$ unterhalb (oberhalb) $F(x)$ liegt. Demgegenüber ist im Sinne des Satzes 3:

$$(27) \lim_{n \rightarrow \infty} P \left(\sup_{x_a \leq x \leq x_b} (F_n(x) - F(x)) < 0 \right) = \frac{1}{\pi} \sqrt{\frac{b}{1-b}} \int_{-\infty}^0 e^{-\frac{bx^2}{2(1-b)}} \int_0^{x\sqrt{\frac{ab}{b-a}}} e^{-\frac{t^2}{2}} dt dx > 0,$$

d. h. die Wahrscheinlichkeit dafür, daß auf dem Abschnitt, auf dem der Wert von $F(x)$ zwischen gewissen Zahlen $a > 0$ und $b < 1$ fällt, die empirische Verteilungsfunktion unterhalb der theoretischen Verteilungsfunktion verläuft, wie klein die positive Zahl a auch sei und wie nahe zu 1 sich auch die Zahl b befinden mag, ist im Grenzwert *positiv*. Dieses überraschende Ergebnis ist offenbar auch vom Gesichtspunkte der statistischen Praxis aus von Bedeutung. Es bedeutet, daß, wenn zwischen x_a ($F(x_a) = a$) und x_b ($F(x_b) = b$) $F_n(x)$ unterhalb (aber sehr nahe) $F(x)$ läuft, noch nicht gefolgert werden kann, daß die Hypothese bezüglich $F(x)$ falsch ist, denn durch Vergrößerung von n , d. h. des Stichprobenumfangs, wird das Vorkommen solcher Stichproben nicht seltener; wenn auch selten, kommen solche Fälle doch immer wieder vor.

Daß der auf der linken Seite von (27) stehende Grenzwert positiv ist, ist, wie ich nach Entdeckung dieser Tatsache erfuhr, schon früher von GICHMAN bewiesen worden, der das Resultat

$$(28) \lim_{n \rightarrow \infty} P \left(\sup_{x_a \leq x \leq x_b} (F_n(x) - F(x)) < 0 \right) = \frac{1}{\pi} \arcsin \sqrt{\frac{a(1-b)}{b(1-a)}}$$

erhielt. GICHMAN teilt ferner mit, daß das Resultat (28) – nach einer persönlichen Mitteilung – auch GNEDENKO schon bekannt war. Die auf der rechten Seite von (27) und (28) stehenden Ausdrücke sind natürlich identisch, was durch eine einfache Rechnung, aber auch ohne Berechnung auf Grund von wahrscheinlichkeitstheoretischen Erwägungen eingesehen werden kann; es ist nämlich die linke Seite von (27) gleich dem Integral

$$(29) \quad \frac{2}{\sqrt{2}\pi} \sqrt{\frac{b-a}{(1-b)a}} \int_0^\infty e^{-\frac{x^2(b-a)}{2a(1-b)}} \left(\frac{1}{\sqrt{2}\pi} \int_0^x e^{-\frac{y^2}{2}} dy \right) dx,$$

und dies ist nichts anderes als die doppelte Wahrscheinlichkeit dafür, daß ein auf der Ebene eine normale Verteilung besitzender Punkt, dessen Projektionsverteilungen auf die Achsen x und y Streuungen $\sqrt{\frac{a(1-b)}{b-a}}$ und 1 besitzen, in den Winkelraum $0 < x < \infty, 0 < y < x$ falle, und diese Wahrscheinlichkeit ist nichts anderes als

$$(30) \quad \frac{2 \operatorname{arc} \operatorname{tg} \sqrt{\frac{a(1-b)}{b-a}}}{2\pi} = \frac{1}{\pi} \operatorname{arc} \sin \sqrt{\frac{a(1-b)}{b(1-a)}}.$$

Es ist nämlich, wenn die normale Verteilung eine symmetrische ist (bei der die Streuungen der Projektionen auf der x - und y -Achse gleich sind), die dem eine Öffnung ϑ besitzenden Winkelraum entsprechende Wahrscheinlichkeit $\frac{\vartheta}{2\pi}$, und hieraus läßt sich der allgemeine Fall mittels affiner Transformation erhalten.

Literatur

- [1] K. PEARSON, On the systematic fitting of curves to observations and measurements. *Biometrika* **1** (1902), 265–303.
- [2] L. v. BORTKIEWICZ, Variationsbreite und mittlere Fehler. *Sitzungsber. Berl. Math. Ges.* **21** (1922), 3–11.
- [3] E. L. DODD, The greatest and the least variate under general laws of error. *Trans. Amer. Math. Soc.* **25** (1923), 525–539.
- [4] L. H. C. TIPPET, On the extreme individuals and the range of samples taken from a normal population. *Biometrika* **17** (1925), 364–387.
- [5] M. FRÉCHET, Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math.* **6** (1928), 93–116.
- [6] A. N. KOLMOGOROV, Über die Grenzwertsätze der Wahrscheinlichkeitsrechnung. *Bull. Acad. Sci. Leningrad* **7** (1933), 363–372.
—, Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari* **4** (1933), 83–91.
- [7] V. I. GLIVENKO, Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari* **4** (1933), 92–99.
- [8] N. V. SMIRNOV, Sur la distribution de ω^2 (critérium de M. v. Mises). (Russisch mit französischem Auszug.) *Rec. Math. Moscou* (2) **2** (1937), 973–993.
—, Sur les écarts de la courbe de distribution empirique. (Russisch mit französischem Auszug.) *Rec. Math. Moscou* (2) **6** (1939), 3–26.
—, Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statistics* **19** (1948), 279–281.

- [9] B. V. GNEDENKO, Grenzwertsätze für das maximale Glied der Variationsreihe. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **32** (1941), 7–9.
 —, Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. **44** (1943), 423–453.
- [10] B. V. GNEDENKO und V. S. KOROLJUK, Über die maximale Abweichung zweier empirischer Verteilungen. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **80** (1951), 525–528.
- [11] B. V. GNEDENKO und E. L. RVAČEVA, Über ein Problem des Vergleichs zweier empirischer Verteilungen. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **82** (1952), 513–516.
- [12] B. V. GNEDENKO und V. S. MICHALEVIČ, Über die Verteilung der Anzahl der Stellen, an denen eine empirische Verteilungsfunktion die andere übertrifft. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **82** (1952), 841–843.
- [13] V. S. MICHALEVIČ, Über die gegenseitige Lage zweier empirischer Verteilungsfunktionen. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **85** (1952), 485–488.
- [14] I. D. KVIT, Über einen Satz von N. V. Smirnov betreffs des Vergleichs zweier Stichproben. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **71** (1950), 229–231.
- [15] G. M. MANIJA, Verallgemeinerung eines Kriteriums von A. N. Kolmogorov für die Abschätzung eines Verteilungsgesetzes nach empirischen Daten. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **69** (1949), 495–497.
- [16] I. I. GICHMAN, Über die empirische Verteilungsfunktion im Falle der Gruppierung der Daten. (Russisch.) Doklady Akad. Nauk SSSR, n. Ser. **82** (1952), 837–840.
- [17] W. FELLER, On the Kolmogorov-Smirnov limit theorems for empirical distributions. Ann. Math. Statistics **19** (1948), 177–189.
- [18] J. L. DOOB, Heuristic approach to the Kolmogorov-Smirnov theorems. Ann. Math. Statistics **20** (1949), 393–403.
- [19] F. J. MASSEY, A note on the estimation of a distribution function by confidence limits. Ann. Math. Statistics **21** (1950), 116–119.
 —, A note on the power of a non-parametric test. Ann. Math. Statistics **21** (1950), 440–443.
 Berichtigung *ibid.* **23** (1952), 637–638.
 —, Distribution table for the deviation between two sample cumulatives. Ann. Math. Statistics **23** (1952), 435–441.
- [20] M. D. DONSKER, Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. Ann. Math. Statistics **23** (1952), 277–281.
- [21] T. W. ANDERSON und D. A. DARLING, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Statistics **23** (1952), 193–212.