TARTU ÜLIKOOLI METEOROLOOGIA OBSERVATOORIUMI
TEADUSLIKUD VÄLJAANDED

SCIENTIFIC PAPERS
OF THE METEOROLOGICAL OBSERVATORY OF THE UNIVERSITY OF TARTU

№ 4

ÜBER DAS SYSTEM DER EINMODIGEN HÄUFIGKEITSKURVEN

VON

AARNE KÄRSNA

TARTU 1939

TARTU ÜLIKOOLI METEOROLOOGIA OBSERVATOORIUMI
TEADUSLIKUD VÄLJAANDED
SCIENTIFIC PAPERS
OF THE METEOROLOGICAL OBSERVATORY OF THE UNIVERSITY OF TARTU
№ 4

ÜBER DAS SYSTEM DER EINMODIGEN HÄUFIGKEITSKURVEN

VON

AARNE KÄRSNA

TARTU 1939

Acta et Commentationes Universitatis Tartuensis (Dorpatensis) A XXXV. 1

K. Mattiesens Buchdruckerei Ant.-Ges., Tartu 1939

I. Einleitung.

Bei der Untersuchung der statistischen Kollektive mittels eines veränderlichen Merkmals benutzt man die diesem Merkmale zugehörige Häufigkeitsverteilung. Da die Häufigkeitsverteilung deutlich die Eigenschaften des Kollektivs zum Vorschein bringt, ist das Problem, die analytische Darstellung der Häufigkeitsverteilung zu finden, immer interessant gewesen. Eine solche Verteilungsfunktion ermöglicht, erstens, die typischen Eigenschaften des Kollektivs zusammenzufassen und, zweitens, die mit den Häufigkeitsverteilungen verknüpften Probleme in in allgemeiner Form zu lösen.

Unter den gebräuchlichen Systemen sind das Pearson'sche und das Charlier'sche die bekanntesten; von diesen besteht das erstere aus 7, das letztere aus 2 Kurventypen. Die Wahl des passenden Kurventypus ist beim Charlier'schen System frei, und es kann bei seiner Anwendung praktisch ausprobiert werden, welcher Typus der passendste ist. Beim Pearson'schen System kann man den Kurventypus nicht freiwillig wählen, da dieser durch eine von der Häufigkeitsverteilung abhängige Charakteristik bestimmt wird. Zum Berechnen dieser Charakteristik sowie der Konstanten der Gleichung wendet man bei beiden Systemen statt der gewöhnlichen Methode der kleinsten Quadrate die bekannte Methode der Momente an. Da beide Systeme bei ihrer praktischen Anwendung viel Rechenarbeit erfordern und die entsprechenden Gleichungen keine direkt klare Vorstellung vom Charakter der Häufigkeitsverteilung bieten, wird in der vorliegenden Arbeit ein neues System vorgeschlagen, in dem die Gleichung der Häufigkeitskurve so gewählt wird, dass die Koeffizienten der Gleichung gleichzeitig auch die Charakteristiken der Häufigkeitsverteilung darstellen. Damit ist auch für die Charakteristiken ein neues System gefunden, weil das bisherige, auf den Momenten beruhende System bei Verteilungen, die von der Normalverteilung stark abweichen, ganz unbrauchbare Resultate lieferte. Beim vorge-

schlagenen System ist das Bestimmen der Koeffizienten und auch das Berechnen der für die graphische Darstellung der Kurve nötigen Koordinaten sehr einfach und selbst dem Nichtmathematiker verständlich. Für den letztgenannten sind am Schluss der Arbeit entsprechend ausgearbeitete Standardmethoden angeführt, die bei nur kleinem Zeitaufwand das Durchführen der nötigen Berechnungen ermöglichen.

Bevor wir unser System näher betrachten, wollen wir einige Mängel der bisher bekannten Systeme erörtern.

II. Die Nachteile der nach den Momenten berechneten Charakteristiken.

Der bekanntesten zum Charakterisieren der Häufigkeitsverteilung benutzten Methode liegen das arithmetische Mittel und die in bezug auf dieses berechneten Momente höherer Ordnung zugrunde. Dieselben Momente bilden auch die Grundlage des Pearson'schen und des Charlier'schen Systems. Gewöhnlich werden nur die vier ersten Momente benutzt, denn die Momente noch höherer Ordnung sind, wie Pearson gezeigt hat, mit sehr grossen Fehlern behaftet. Das Gesagte ist hauptsächlich dadurch bedingt, dass bei gegebenem Kollektivumfang s der Variationsfaktor der Häufigkeit, wenn die Häufigkeit abnimmt, unbegrenzt wächst.

Es sei beim Kollektivumfang s die relative Häufigkeit einer Klasse p. Dann ist die absolute Häufigkeit m

$$m = sp. \tag{1}$$

Bei der Bernoulli'schen Reihe ist das Streuungsmass der Häufigkeit dieser Klasse

$$\sigma_B = \sqrt{sp\,(1-p)}. \tag{2}$$

Danach ist der Variationsfaktor

$$V = \frac{\sigma_B}{m} = \sqrt{\frac{1-p}{sp}}. \tag{3}$$

Aus der Formel (3) ist ersichtlich, dass kleinere Häufigkeiten relativ ungenauer sind und, wenn sie sich an den Enden der Kurve der Häufigkeitsverteilung befinden und damit grosse

Abszissenwerte besitzen (vom arithmetischen Mittel gemessen), bei höheren Momenten zu einem grossen Gesamtfehler führen.

Bei den Lexis'schen Reihen, die am häufigsten vorkommen, wird das Ergebnis noch schlechter. Beim Lexis'schen Faktor L ist

$$V = \frac{L\sigma_B}{m} = \sqrt{\frac{L^2\,(1-p)}{sp}}\,. \tag{4}$$

Dieses zeigt, dass die Streuung leicht grösser werden kann als die Häufigkeitszahl, denn die Bedingung

$$L^2\,(1-p) > sp \tag{5}$$

kann leicht befriedigt werden.

Noch wichtiger als der Fehler der Momente (bei der Bernoulli'schen Reihe kann man diesen Fehler durch Vergrössern des Kollektivumfanges vermindern) ist die Beziehung zwischen den Momenten und der Häufigkeitsverteilung. Beim Bestimmen der Grösse der Momente beeinflusst das typische Gebiet der Häufigkeitsverteilung (die grossen Häufigkeiten) diese viel weniger als die äusseren Teile der Kurve der Häufigkeitsverteilung, wo die Häufigkeiten klein sind und relativ grössere Fehler aufweisen.

Unter diesem wesentlichen Mangel leiden alle Momente und damit auch die nach ihnen berechneten Charakteristiken (auch das arithmetische Mittel), und zwar desto mehr, je höher die Ordnung der Momente ist. Ferner ist bekannt, dass die Abweichungen der Momente in korrelativer Beziehung stehen, und zwar einige (z. B. die des zweiten und vierten Moments) besonders stark. Mit dem Wachsen des einen Moments wächst auch das andere, und infolgedessen haben wir es nicht nur mit grossen Fehlern hinsichtlich der Momente, sondern auch mit ihrer Zusammenwirkung zu tun.

Die nach den Momenten berechneten Charakteristiken sind nur dann befriedigend, wenn die Häufigkeitsverteilung sich in der Nähe der Normalverteilung hält. In vielen Fällen sind die Charakteristiken überhaupt nicht charakteristisch und ist ein Aufzeichnen der Verteilung mit Hilfe derselben ganz unmöglich.

Im folgenden wollen wir die wesentlichen Nachteile der wichtigsten Charakteristiken näher betrachten.

a. Das arithmetische Mittel.

Das arithmetische Mittel ist die am meisten gebräuchliche Charakteristik. Beim Bezeichnen des Charakters eines Kollektivs durch eine Zahl wird das arithmetische Mittel gebraucht. Wenn die Häufigkeitsverteilung der Normalkurve folgt, fällt das arithmetische Mittel mit der Mode zusammen, und damit ist dieses auch die typische Grösse im Kollektiv. Wie aber bekannt ist, kommen die Häufigkeitsverteilungen in verschiedenartiger Form vor und gibt es auch solche, bei denen die Häufigkeit des arithmetischen Mittels im Vergleich zu der Häufigkeit anderer Argumentswerte sehr klein ist. Es ist z. B. das arithmetische Mittel der Flächengrösse der Seen (1612 Seen) Estlands ohne den Võrtsjärv 14 ha[1]), mit dem Võrtsjärv zusammen aber 32 ha. Das Hinzunehmen nur eines einzigen Elementes ins Kollektiv führt wegen dessen grossem Argumentwert das arithmetische Mittel weit über die häufig vorkommenden Argumentwerte hinaus. Was müsste man hierbei für typisch halten: 32 oder 14? Ergänzend sei noch gesagt, dass die Zahl der Seen mit einer Flächengrösse von unter 14 ha 86% und von unter 32 ha sogar 92% der Gesamtzahl beträgt. Da bezüglich der kleineren Seen nicht genügend statistisches Material vorliegt, kann die Lage der Mode nicht genau bestimmt werden. Wenn wir jedoch die Mode gleich 1 ha annehmen, so erhalten wir eine Häufigkeit, die mehr als 30 mal grösser ist, als die für das erste und 120 mal grösser, als die für das zweite arithmetische Mittel. Es scheint demnach, dass das Charakterisieren eines solchen Kollektivs durch das arithmetische Mittel nicht richtig ist, denn in den beiden Häufigkeitsverteilungen kommen ja fast gar keine Unterschiede vor. Beim Charakterisieren des Kollektivs sind alle Elemente gleichwertig, und infolgedessen bietet das arithmetische Mittel nicht immer das Wesentliche und Typische.

b. Das Streuungsmass.

Weiter ist auf Grund desselben Systems die folgende Charakteristik — das Streuungsmass — bestimmt, welches mit Hilfe des zweiten Moments berechnet wird und manchmal auch mittlere

---

1) H. Riikoja and A. Kärsna, On the Distribution of Lakes in Estonia. Loodusuurijate Seltsi aruanded XLII 3—4, 1935. Tartu.

oder quadratische Abweichung genannt wird. Das Streuungsmass wird in statistischen Arbeiten sehr oft benutzt, weil ihm nur ein kleiner Fehler anhaftet. Wenn früher z. B. in biologischen Forschungen zum Charakterisieren der Streuung die Extremwerte mancher Individuen benutzt wurden, wurde doch bald klar, dass die Extremwerte sich bei Veränderung des Kollektivumfangs stark veränderten, das Streuungsmass aber schon bei einer kleinen Zahl von Individuen mehr oder minder die wirkliche Grösse besass.

Das Gesagte gilt jedoch nur solange die Häufigkeitsverteilung in der Nähe der Normalverteilung bleibt, in einem gewissen Grade auch noch dann, wenn die Verteilung symmetrisch ist. Im allgemeinen Fall kann aber die Zahl der Elemente in den Grenzen zwischen $x = -\sigma$ bis $x = \sigma$ sehr schwanken. Das kann durch folgendes Beispiel erläutert werden.

1. Betrachten wir zuerst die konstante Häufigkeitsverteilung (Fig. 1)

(6) $$y = a$$

im Intervall $-b \leqslant x \leqslant b$.

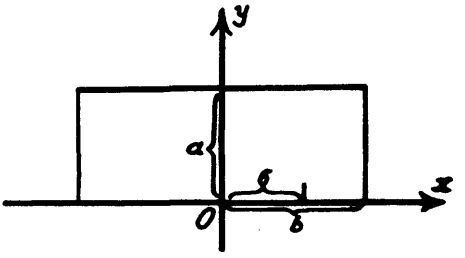


Fig. 1.

Das Streuungsmass (σ) können wir leicht bestimmen:

(7) $$\sigma^2 = \frac{\int_{-b}^{b} ax^2 dx}{\int_{-b}^{b} a dx} = \frac{b^2}{3},$$

woraus

(8) $$\sigma = \frac{b}{\sqrt{3}},$$

so dass die relative Häufigkeit der Klasse von $x = -\sigma$ bis $x = \sigma$

(9) $$S = \frac{\int_{-b}^{\sigma} a\,dx}{\int_{-b}^{b} a\,dx} = \frac{1}{\sqrt{3}} = 57{\cdot}7\,\%$$

beträgt (sie ist nicht von a und b abhängig).

2. Zweitens betrachten wir solch eine Häufigkeitsverteilung, bei der sich die Werte in der Mitte stark anhäufen (Fig. 2).

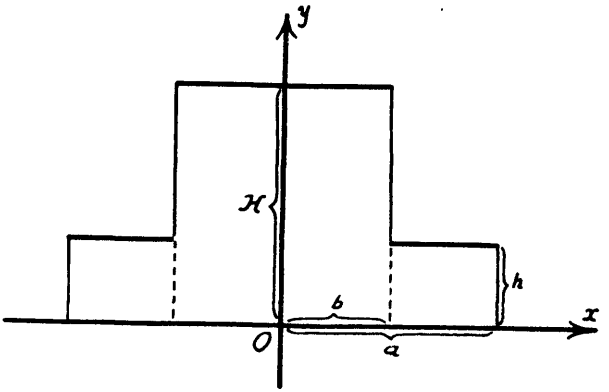


Fig. 2.

Bezeichnen wir

(10) $$\sigma^2 = \frac{\int_{-a}^{a} yx^2\,dx}{\int_{-a}^{a} y\,dx} = \frac{L}{N}.$$

Es sei die Gleichung der Häufigkeitskurve

(11) $$y = H,$$

wenn $-b \leqslant x \leqslant b$ ist, und

(12) $$y = h,$$

wenn $b \leqslant x \leqslant a$ oder $-a \leqslant x \leqslant -b$ ist.

Dementsprechend finden wir

(13) $$N = 2\int_{0}^{b} H dx + 2\int_{b}^{a} h dx = 2Hb + 2h(a-b)$$

und

(14) $$L = 2\int_{0}^{b} Hx^2\,dx + 2\int_{b}^{a} hx^2\,dx = \frac{2Hb^3}{3} + \frac{2h}{3}(a^3 - b^3).$$

Bezeichnen wir

(15) $$H = mh$$

und

(16) $$a = nb$$

und setzen diese Werte in die Formeln (13) und (14) ein, so erhalten wir

(17) $$L = \frac{2hb^3}{3}(m + n^3 - 1)$$

und

(18) $$N = 2hb(m + n - 1).$$

Setzen wir dieses in die Formel (10) ein, so bekommen wir

(19) $$\sigma^2 = \frac{b^2(m + n^3 - 1)}{3(m + n - 1)}.$$

Bemerkung: m und n sind positive Zahlen, und zwar grösser als Eins.

Betrachten wir nur die m- und n-Werte, bei welchen $\sigma = b$ ist, so muss wegen Gleichung (19) auch die Gleichung

(20) $$m + n^3 - 1 = 3(m + n - 1)$$

befriedigt werden, oder

(21) $$m = \frac{n^3 - 3n + 2}{2}.$$

Aus letzterem ersehen wir, dass beim Anwachsen von n auch m, und zwar in einem viel stärkeren Masse, zunimmt. Die folgende Tabelle zeigt die Beziehungen zwischen m und n.

Tab. 1.

n	2	3	4	5	6	7	8
m	2	10	27	56	100	162	245

Weiter wollen wir die Häufigkeit der Klasse von $x = -\sigma$ bis $x = \sigma$ bestimmen; wir bezeichnen diese mit M und erhalten

(22) $$M = \int_{-\sigma}^{\sigma} H dx = 2\int_{0}^{b} mh dx = 2mhb.$$

Die relative Häufigkeit derselben Klasse ist

$$S = \frac{M}{N} = \frac{2mhb}{2hb(m+n-1)} = \frac{m}{(m+n-1)}, \tag{23}$$

wobei die Formel (18) benutzt worden ist.

Da wir den Fall, wo $\sigma = b$ ist, betrachten, können wir aus der Formel (21) den Ausdruck für m in Gleichung (23) einsetzen und erhalten dann

$$S = \frac{n^3 - 3n + 2}{n^3 - n}. \tag{24}$$

Die folgende Tabelle enthält für einige n-Werte die S-Werte in Prozenten.

Tab. 2.

n	2	3	4	5	6	7	8	10	20
$S(\%)$	66·7	83·3	90·0	93·3	95·3	96·4	97·4	98·2	99·6

Aus der Tabelle ist zu ersehen, dass die relative Häufigkeit der genannten Klasse beim Wachsen von n schnell zunimmt. Aus der Formel (24) wird auch ohne weiteres klar, dass

$$\lim_{n \to \infty} S = 1.$$

Zusammenfassend können wir sagen, dass in den gegebenen Fällen die relative Häufigkeit der Klasse von $x = -\sigma$ bis $x = \sigma$ in den Grenzen von 58%—100% schwankt. Die Lage wird schlimmer, wenn wir zu schiefen Häufigkeitskurven übergehen. Bei einer sehr schiefen Kurve kann die σ-Weite über die Kurve hinausreichen. Als Beispiel kann die obenerwähnte Häufigkeitsverteilung der Flächengrössen der Seen Estlands genannt werden, wo die Berechnung das folgende Resultat ergab: ohne den Võrtsjärv ist $\sigma = 67$ ha und mit dem Võrtsjärv ist $\sigma = 700$ ha. Im ersteren Falle liegen in den Grenzen $+\sigma$ 96·6 % aller Elemente und im lezteren Falle 99·8%. Kann man hier von einem Charakterisieren der Streuung sprechen? Ausserdem erreicht im ersten Falle die Häufigkeitsverteilung einerseits des arithmetischen Mittels nur bis $0{\cdot}21\,\sigma$ und im zweiten Falle nur bis $0{\cdot}04\,\sigma$. Wären für das genannte Kollektiv nur das arithmetische Mittel (m) und das Streuungsmass (σ) gegeben ($m = 14$ ha, $\sigma = 67$ ha oder $m = 32$ ha, $\sigma = 700$ ha), so könnte man sich auf Grund dieser

Charakteristiken kein richtiges Bild von den Verhältnissen schaffen.

c. Die höheren Charakteristiken.

Im Pearson'schen und im Charlier'schen System benutzt man für ein genaueres Charakterisieren der Häufigkeitsverteilung noch die Momente höheren Grades und die aus ihnen abgeleiteten Charakteristiken. So benutzt man das mit Hilfe der Momente dritten Grades berechnete Mass der Schiefheit und das mit Hilfe der Momente vierten Grades berechnete Mass des Exzesses.

Der Mangel aller genannten Charakteristiken besteht darin, dass zwar einer gegebenen Häufigkeitsverteilung nur ein Komplex von Charakteristiken entspricht, einem gegebenen Komplex von Charakteristiken jedoch nicht nur eine Verteilung. Bei zweckmässiger Wahl können die Häufigkeitszahlen einer Verteilung so verändert werden, dass alle Momente unverändert bleiben. Fig. 3 stellt zwei Häufigkeitsverteilungen dar, bei denen die Zahl der Elemente und die Grössen aller vier Momente gleich gross sind, aber als kongruent können wir sie dennoch nicht betrachten.

Die Berechnung ergab für die erste Verteilung (die y-Achse ist durch das arithmetische Mittel gezogen)

$$\mu_2 = 4{\cdot}64,$$
$$\mu_3 = 4{\cdot}47 \text{ und}$$
$$\mu_4 = 60{\cdot}1$$

(μ_i bedeutet das Moment i-ten Grades) und für die zweite Verteilung

$$\mu_2 = 4{\cdot}68,$$
$$\mu_3 = 4{\cdot}86 \text{ und}$$
$$\mu_4 = 61{\cdot}1.$$

Aus der mathematischen Statistik wissen wir, dass die mittleren Fehler der Momente ($\sigma\mu_i$)

$$\sigma\mu_2 = \sigma^2 \sqrt{\frac{2}{s}},$$

$$\sigma\mu_3 = \sigma^3 \sqrt{\frac{6}{s}} \text{ und}$$

$$\sigma\mu_4 = \sigma^4 \sqrt{\frac{96}{s}}$$

sind (wo σ das Streuungsmass und s die Zahl der Elemente

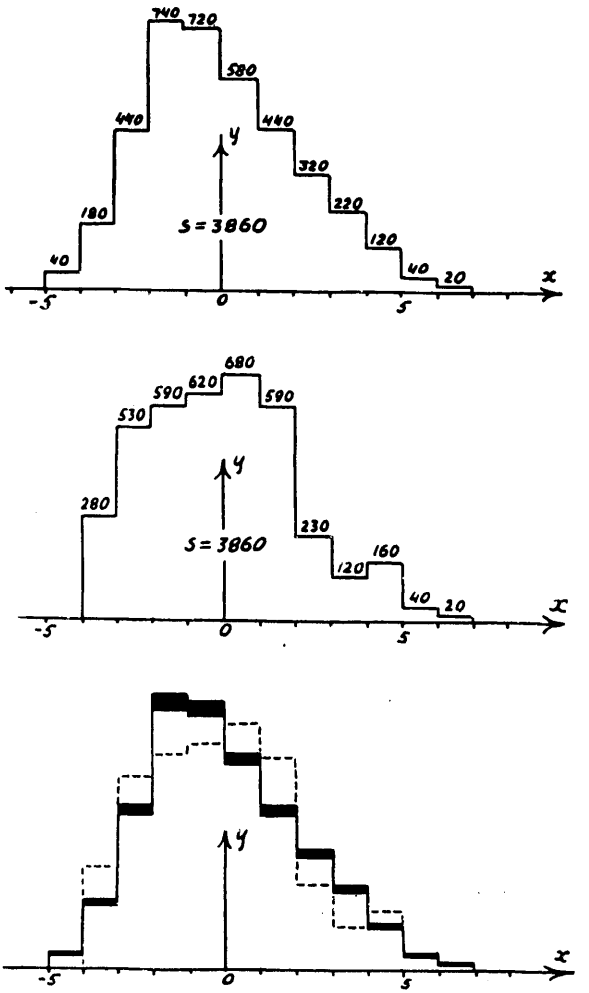


Fig. 3.

angibt), und nehmen wir an, dass die Differenzen zwischen den Momenten der beiden Verteilungen gleich dem mittleren Fehler seien, so können wir die Zahl der Elemente s bestimmen, bei denen solche mittlere Fehler vorkommen können. Nach den Momenten zweiter Ordnung bekommen wir $s = 108\,000$, nach den Momenten dritter Ordnung $s = 16\,000$ und nach den Momenten vierter Ordnung $s = 181\,000$. Die grösste Differenz besteht zwischen den Momenten zweiter Ordnung, beim gegebenen Kollektivumfang $s = 3860$ kann man aber alle Momente im Fehlergebiet als gleich gross betrachten.

Einer besseren Übersicht wegen sind (in derselben Figur) die beiden Häufigkeitsverteilungen in ein und demselben Koordinatensystem dargestellt. Für die eine Verteilung ist der Bernoullische Streuungsstreifen hinzugezeichnet (schwarzer Streifen). Dieses bringt zum Vorschein, dass die beiden Verteilungen nie aus einem und demselben Kollektiv stammen können. Wenn z. B. die vorliegenden Häufigkeitsverteilungen sich auf die Temperaturen eines Beobachtungsortes in zwei aufeinander folgenden Zeitintervallen beziehen, dann hätten wir es mit einer sehr sicheren Klimaänderung zu tun, obwohl keine Charakteristik darauf deutlich hinweist.

III. Die Nachteile des Pearson'schen Systems.

a. Allgemeine Bemerkungen.

Es muss zugegeben werden, dass die Basis des Pearson'schen Systems, der Begriff des erweiterten Elementarfehlers, in vielen Fällen sehr annehmbar ist. Bei einer sehr grossen Menge von empirischen Kollektiven kann man aber eine endliche Zahl konkreter Gründe heraussuchen, bei denen eine Zunahme oder ein Weglassen eines derselben die Form der Häufigkeitsverteilung ansehnlich verändert. Ebenso sind bei empirischen Kollektiven die Merkmale oft miteinander korrelativ verbunden, so dass die dem einen Merkmale zugehörige Häufigkeitsverteilung auch die anderen in einem gewissen Masse beeinflusst.

Bei der Häufigkeitsverteilung der Lufttemperatur sind beispielsweise die Gaskonstante der Luft, die innere Reibung, die spezifische Wärme, die Wärme- und Temperaturleitung, die Schmelz- und Verdampfungswärme des Wassers und noch einige andere Momente massgebend. Könnten wir diese Konstanten verändern, so würde sich auch die Häufigkeitsverteilung verändern. So liegt z.B. in Tartu (Estland) bei der Häufigkeitsverteilung der Temperaturen im Januar die grösste Häufigkeit bei 1^0 und 2^0 C, was hauptsächlich durch die Schmelzwärme des Wassers verursacht wird. Wenn die Temperaturen über Null steigen, so beginnt der Schnee zu schmelzen und bedarf wegen seiner grossen Schmelzwärme zur Veränderung seines Aggregatzustandes der aus den Luftmassen hinzukommenden Wärme. Wäre die Schmelzwärme ganz gering, so würden die Temperaturen nicht bei 1^0 bis 2^0 stehen

bleiben, sondern sie würden viel höher steigen und die Häufigkeitsverteilung hätte eine viel symmetrischere Form angenommen.

Dieses bedeutet, dass die Häufigkeitsverteilung in diesem Falle von einer Reihe spezieller Gründe bestimmt wird, die wir als zeitweise konstant annehmen können. Da diese speziellen Gründe aber nicht in der ganzen Variationsbreite zu wirken brauchen, ist die im allgemeinen Fall mittels des Elementarfehlers abgeleitete Häufigkeitskurve nicht annehmbar. Z. B. werden in Estland für den Polizeidienst Männer mit einer Körperlänge von mindestens 172 cm gewählt, und dadurch ist die Häufigkeitsverteilung der Körperlänge der Polizisten ganz abweichend von der des ganzen Volkes, obwohl jeder Polizist auch zum Volk gehört.

Die Störungsstreuung der Lexis'schen Reihe wird ja durch solche konkrete Gründe verursacht. Wenn das Prinzip des Elementarfehlers allgemeingültig wäre, so wäre eine Entstehung der Lexis'schen Reihe unmöglich, denn die Wahrscheinlichkeit des Auftretens eines Merkmals wäre die ganze Zeit konstant. In einem solchen Falle fehlt jede Evolution, die nur auf Grund der Lexis'schen Reihe möglich ist.

Der Glaube an das Prinzip des Elementarfehlers ist sogar so gross gewesen, dass die Gauss'sche Normalkurve als Anzeiger der Typen biologischer Elemente betrachtet wurde. Einen Anstoss hierzu gab der Umstand, dass bei einigen Typen die Häufigkeiten mancher Merkmale mehr oder weniger der Normalkurve folgten. Daraus folgerte man, dass es so mit jedem Merkmale eines jeden Typus sein müsse, und in dem Falle, wo die Verteilung von der Normalkurve abwich, wurde gedacht, dass dort eine Vermischung von Typen stattgefunden habe. Dieses glaubte auch Pearson; er hat sogar eine Arbeit über das Einteilen der Häufigkeitskurven in Normalkurven veröffentlicht.

Da der allgemeine Charakter der Häufigkeitsverteilung nicht nur durch Elementarfehler, sondern auch durch andere konkrete Gründe verursacht wird, muss das Bestimmen des Typus auf Grund von biologischen Merkmalen stattfinden. Die Häufigkeitsverteilung kann mehr oder weniger normal sein, sie kann es aber auch nicht sein.

Hierher gehört noch die Frage, welches Merkmal der Normalkurve folgt, ob es die Körperlänge, das Gewicht usw. tun. Die Häufigkeitsverteilung der Körperlänge der Menschen ist

annähernd normal, die des Körpergewichts aber ist eine Verteilung mit positiver Schiefe (das Moment dritter Ordnung ist positiv). Es wäre ja ein Paradoxon zu behaupten, dass dieselben Menschen ihrer Körperlänge nach nur zu einer Rasse, ihrem Körpergewicht nach aber zu vielen Rassen gehörten, denn die Häufigkeitsverteilung des Gewichts kann in mehrere Normaltypen zerfallen.

Warum die Häufigkeitsverteilungen der Länge und des Gewichts verschieden sind, wird im folgenden in grossen Zügen erklärt.

Durchschnittlich verändert sich das Körpergewicht proportional der dritten Potenz der Körperlänge. Damit entspricht jeder Länge L normal ein Gewicht M so, dass

$$M = kL^3 \tag{25}$$

ist, wo k eine Konstante ist. Die Zunahme der Länge dL ruft die Zunahme des Gewichts dM hervor, so dass

$$dM = 3kL^2 dL \tag{26}$$

ist. Aus letzterem ersehen wir, dass dM von L abhängig, und bei grösserem L grösser als bei kleinerem L ist. Wenn die Häufigkeitsverteilung von L symmetrisch ist, wirkt sie auf diejenige von M so ein, dass bei letzterer der rechtsliegende Zweig länger wird als der linksliegende.

Da der Variationsfaktor der Länge klein ist (ca 3%), ist auch die Asymmetrie der Häufigkeitsverteilung des Gewichts gering und nicht direkt sichtbar. Es fällt aber die Tatsache auf, dass der Variationsfaktor des Gewichts auffallend grösser ist als derjenige der Länge. Teilen wir die Glieder der Gleichung (26) durch diejenigen von (25), so erhalten wir

$$\frac{dM}{M} = 3\,\frac{dL}{L}. \tag{27}$$

Aus letzterem wird ersichtlich, dass die relativen Abweichungen des Gewichts dreimal grösser sind als die der Länge. Nehmen wir als Abweichungen solche von den Mittelwerten L_0 und M_0, so gilt in erster Annäherung dieselbe Beziehung und muss der Variationsfaktor des Gewichts (V_M) dreimal grösser als derjenige der Länge (V_L) genommen werden. Nach 1700 Messungen in Estland (die Daten des Anthropologen J. Aul) betrug

$$V_M = 9{\cdot}68\,\% \pm 0{\cdot}17\,\% \text{ und}$$
$$V_L = 3{\cdot}27\,\% \pm 0{\cdot}06\,\%.$$

Damit ist

$$\frac{V_M}{V_L} = 2{\cdot}96 \pm 0{\cdot}03$$

(bei der Berechnung des Fehlers muss man wissen, dass der Korrelationsfaktor zwischen der Länge und dem Gewicht $r = 0{\cdot}65$ ist), was eine gute Übereinstimmung ergibt.

Nach gegebener Häufigkeitsverteilung der Länge L kann auch theoretisch die Häufigkeitsverteilung des Gewichts M konstruiert werden.

Wir bezeichnen die zusammengehörigen Ordinaten der beiden Kurven der L und M bzw. durch H_L und H_M. Wir wissen, dass die Häufigkeiten der beiden Verteilungen für jede dL- und dM-Klasse gleich gross sind. Darum können wir schreiben

$$H_L \cdot dL = H_M \cdot dM\,, \tag{28}$$

woraus

$$H_M = H_L \cdot \frac{dL}{dM}. \tag{29}$$

Setzen wir an Stelle von dM den entsprechenden Ausdruck aus (26), so bekommen wir

$$H_M = H_L \cdot \frac{1}{3kL^2}\,, \tag{30}$$

was uns ermöglicht, für jedes k nach der Häufigkeitsverteilung der Länge die Häufigkeitsverteilung des Gewichts zu konstruieren, wobei k die Einheit, mit der M gemessen wird, bestimmt. In Fig. 4

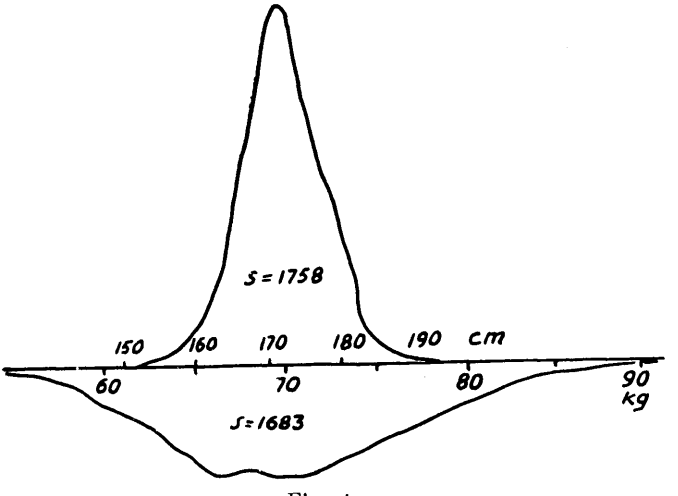


Fig. 4.

sind nach den Messungen von J. Aul die Häufigkeitsverteilungen der Länge und des Gewichts der erwachsenen Männer in Westestland angeführt, woraus ersichtlich ist, in welchem Masse sich diese voneinander unterscheiden (in der Figur fallen die Nullpunkte der Skalen der Länge und des Gewichts zusammen). In der nächsten Figur (Fig. 5) ist die nach der Häufigkeitsverteilung der Länge konstruierte Häufigkeitsverteilung des Gewichts (punktierte Linie), die eine sehr gute Übereinstimmung mit der wirklichen Verteilung gibt, dargestellt. Die nebeneinander gezeichneten Skalen ermöglichen die normale Beziehung zwischen

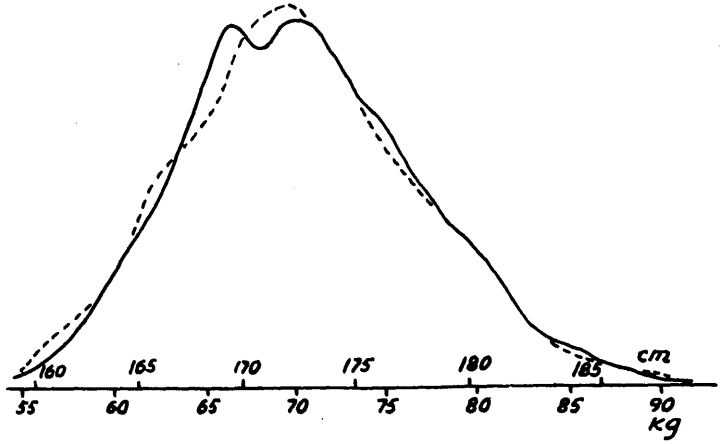


Fig. 5.

Länge und Gewicht zu bestimmen. Wird M in kg und L in cm gemessen, so ist

$$k = 1{\cdot}37.10^{-5}$$

und die entsprechende Beziehung ist

$$M_{kg} = 1{\cdot}37.10^{-5} L^3_{cm},$$

oder

$$M_{kg} = 13{\cdot}7 \; L^3_m.$$

Das Gesagte zeigt, dass die Normalkurve nicht zum Charakterisieren der Rasse dienen kann, denn die Merkmale einer Rasse stehen miteinander in korrelativer Beziehung, und diese ist nicht immer linear. Warum soll der Einfluss der Elementarfehler sich nur in der Körperlänge zeigen, nicht aber beim Körpergewicht? Der Organismus wächst ja in drei Richtungen, und das Wachstum jeder Zelle vergrössert die Masse.

b. Über die technischen Nachteile des Systems.

Die aus den in bezug auf das arithmetische Mittel berechneten Momenten zweiter, dritter und vierter Ordnung (μ_2, μ_3 und μ_4) gebildeten Charakteristiken

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{(das Mass der Schiefheit)}$$

und

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad \text{(das Mass des Exzesses)}$$

bestimmen im Pearson'schen System eine Charakteristik k, welche ihrerseits den Typus der Häufigkeitskurve bestimmt. Dabei ist

$$k = \frac{\beta_1 (\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}.$$

Im Falle

1. $k < 0$, ist die Gleichung der Häufigkeitskurve

$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2} \quad \text{(Typus I)}.$$

2. Wenn $k = 0$ ist, ist

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^m \quad \text{(Typus II)}.$$

Aus diesem folgt der Spezialfall für $\beta = 3$:

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}} \quad \text{(Typus VII, Normalkurve)}.$$

3. Im Falle $0 < k < 1$, ist

$$y = y_0 \left(1 + \frac{x_2}{a_2}\right)^{-m} e^{-v \arctan \frac{x}{a}} \quad \text{(Typus IV)}.$$

4. Wenn $k = 1$ ist, ist

$$y = y_0 x^{-p} e^{-\frac{\gamma}{x}} \quad \text{(Typus V)}.$$

5. Im Falle $1 < k < \infty$, ist

$$y = y_0 (x - a)^{-q_2} x^{-q_1} \quad \text{(Typus VI)},$$

und wenn

6. $k = \infty$ ist (praktisch, wenn k genügend gross ist), ist

$$y = y_0 e^{-\gamma x} \left(1 + \frac{x}{a}\right)^{\gamma a} \quad \text{(Typus III)}.$$

Hier sind x und y die Koordinaten der Kurvenpunkte, e die Basis der natürlichen Logarithmen, und alle übrigen Buchstaben sind Konstanten, die mit Hilfe der Momente berechnet werden können.

Als Nachteile des Systems technischer Art können genannt werden:

1. Die äusserliche Verschiedenheit der Kurventypen. Sind zwei Häufigkeitsverteilungen durch verschiedene Gleichungen gegeben, so fehlt die Möglichkeit des Vergleichens der Verteilungen. Die Gestalt der Gleichung und die Koeffizienten können keine Vorstellung vom Charakter der Häufigkeitsverteilung geben.

2. Die Koeffizienten der Gleichung sind für die Häufigkeitsverteilung nicht charakteristisch. Die wichtigsten Eigenschaften der Häufigkeitsverteilung, wie die Variationsbreite, die Schiefheit und der Exzess, sind nach den Koeffizienten sehr schwer, bei einigen Typen fast unmöglich zu erkennen.

3. Die zum Feststellen des Typus und zum Bestimmen der Koeffizienten erforderliche Rechenarbeit ist zeitraubend und mit schwierigen mathematischen Operationen verbunden (das Berechnen der Γ-Funktion). Für Nichtmathematiker ist sie fast undurchführbar.

4. Die Gleichung gibt keine Aufklärung über den Charakter der entsprechenden Kurve, und die zur graphischen Darstellung nötige Vorarbeit ist ebenfalls kompliziert und zeitraubend.

Zusammenfassend kann gesagt werden, dass die Anwendung des Pearson'schen Systems eine gewisse mathematische Vorbildung voraussetzt. Für Nichtmathematiker ist die Anwendung der Gleichungen fast unmöglich, und dadurch ist es zu erklären, dass über das Pearson'sche System in der Literatur genug geschrieben, praktisch aber dieses System in der Naturwissenschaft und Wirtschaftskunde wenig angewandt worden ist. Die allerweiteste Anwendung könnten die statistischen Methoden jedoch gerade in der Biologie und der Sozialmassenkunde finden.

c. Über die inhaltlichen Nachteile des Systems.

1. Das Prinzip des Elementarfehlers, welches infolge des Aufbaues der Kurventypen auf deduktivem Wege die inhaltliche Grundlage bilden soll, scheint bei einer grossen Menge von empirischen Kollektiven nicht annehmbar zu sein, und dadurch wird

der ganze Aufbau in den meisten Fällen zu einer Interpolationslösung. Das System unterscheidet sich nicht vom Approximationssystem.

2. Inhaltlich wird der detaillierte Aufbau des Systems fast überflüssig, wenn die Bernoulli'sche Reihe durch eine Lexis'sche ersetzt werden kann, und die klimatologische, biologische und Sozialstatistik zeigen, dass die tatsächlich vorkommenden statistischen Reihen meist Lexis'sche Reihen sind.

Schon bei der Bernoulli'schen Reihe werden bei einem endlichen Kollektivumfang die Fehler der Momente höherer Ordnung sehr gross, aber dort tröstet uns die Tatsache, dass bei einem unendlich grossen Kollektivumfang sich die Verteilung einem bestimmten Grenzwert nähert. Bei der Lexis'schen Reihe sind die Fehler der Momente höherer Ordnung so gross, dass schon bei Veränderung der Häufigkeit mancher Klasse in ihrem Fehlergebiet der empirischen Kurve mehrere Typen im Pearson'schen System entsprechen können. Diese Erscheinung ist dadurch bedingt, dass die den Typus bestimmende Charakteristik k nicht von konstanter „Empfindlichkeit" ist, sondern in einigen Gebieten der $\beta_1\beta_2$-Ebene ihre Veränderung eine sehr grosse Geschwindigkeit aufweist. Infolgedessen kann wegen der kleinen Veränderungen von β_1 und β_2 einerlei welcher von allen Kurventypen der Gleichung entsprechen, und damit ist auch die folgende Frage berechtigt: ist eine so grosse Menge von Kurventypen überhaupt nötig, oder könnte man im Falle der Lexis'schen Reihe durch einen einzigen Kurventypus dieselben Verteilungsformen gewinnen? Im folgenden Beispiel wird gezeigt, in welcher Weise kleine Veränderungen der Häufigkeiten ein Wechseln der Kurventypen hervorrufen können.

Nach der in Fig. 6 gegebenen Häufigkeitsverteilung erhalten wir

$$\beta_1 = 0{\cdot}01971 \text{ und}$$
$$\beta_2 = 3{\cdot}0247 ;$$

danach ist

$$k = -1{\cdot}53 \text{ (Typus I)}.$$

Addiert man zu der mittleren Ordinate ($y = 40$) die Häufigkeit 0·5 hinzu, (dann ist $y = 40{\cdot}5$) so erhält man

$$\beta_1 = 0{\cdot}01976 \text{ und}$$
$$\beta_2 = 3{\cdot}0332,$$

woraus

$$k = 2 \cdot 10 \text{ (Typus VI).}$$

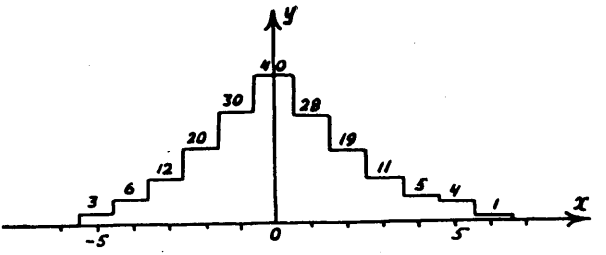


Fig. 6.

Addieren wir zu derselben Ordinate nochmals die Häufigkeit 0·5 hinzu (dann ist $y = 41$), so bekommen wir

$$\beta_1 = 0 \cdot 01982 \text{ und}$$
$$\beta_2 = 3 \cdot 0416,$$

woraus

$$k = 0 \cdot 631 \quad \text{(Typus IV).}$$

Beim Übergang von Typus I zu Typus VI überschreiten wir $k = \infty$ (Typus III), und zwischen dem Typus VI und Typus IV ist $k = 1$ (Typus V). Damit entstehen beim Übergang der Häufigkeitsveränderung von 40 zu 41 fünf Kurventypen.

Setzen wir voraus, dass diese Abweichung gleich der mittleren Abweichung der Bernoulli'schen Reihe ist, so können wir den entsprechenden Kollektivumfang bestimmen, denn der Variationsfaktor ist $\frac{1}{40}$. Benutzen wir die Formel des Variationsfaktors V

$$V = \sqrt{\frac{1-p}{sp}}$$

(wo s den Kollektivumfang und p die relative Häufigkeit der zu betrachtenden Klasse bedeuten), so erhalten wir $s = 5560$, denn aus der Figur ist ersichtlich, dass $p = 0 \cdot 223$ ist.

Dieses bedeutet: bei $s = 5560$ kann diese Abweichung als eine zufällige betrachtet werden, und zwar desto sicherer, je kleiner s ist. Bei der Lexis'schen Reihe kann s noch L^2 mal grösser sein (L ist der Lexis'sche Faktor).

Da praktisch sehr grosse L-Werte vorkommen, wird dadurch das Bedürfnis nach dem Pearson'schen System fraglich, denn wir

können bei einem Kollektiv für seine Partialkollektive verschiedene Kurventypen erhalten. Das Gesagte bedeutet gewiss nicht, dass dieses System unbrauchbar ist, — es ist aber nicht zweckmässig. Darum scheint es zweckmässiger, wenn das System der Häufigkeitskurven induktiv geschaffen wird, indem eine Gleichung von einer passenden Form gewählt wird, deren Koeffizienten die typische Eigenschaften der Häufigkeitsverteilung zum Ausdruck bringen. Dieselben Koeffizienten können dann auch als Charakteristiken der Häufigkeitsverteilung verwendet werden. Da keine inhaltliche Begründung angegeben werden kann, weshalb gerade die genannte Gleichung gewählt worden ist, so gilt ein solches System als ein Abkommen, wie das in der Wissenschaft bei einer grossen Menge von Normen der fall ist.

IV. Die symmetrische Häufigkeitsverteilung.

a. Die symmetrische Häufigkeitskurve.

Wie oben erwähnt wurde, ist es fast unmöglich, auf deduktivem Wege eine analytische Darstellung für die Häufigkeitsverteilungen zu finden, und zwar weil solche allgemeine Gründe, welche die Häufigkeitsverteilungen bestimmen, nicht bekannt sind und man es in einem jeden Falle mit speziellen Gründen zu tun hat. Darum führt die induktive Methode bei der Aufgabe, für die gegebene Häufigkeitsverteilung eine passende Annäherungsgleichung zu finden, schneller zum Ziel. Eine passende Wahl der Gleichungsform kann zu gutem Erfolge führen, was z. B. die Häufigkeitskurve der Flächengrösse der estnischen Seen, bei der eine Potenzfunktion als Näherung angenommen wurde, gezeigt hat, während das Pearson'sche System eine ganz unpassende Näherung ergab.

Auf Grund des Gesagten erscheint uns, dass das Finden der Häufigkeitskurve in jedem Fall als Spezialproblem bestehen bleiben muss. Weil aber eine grosse Menge von Häufigkeitsverteilungen eine Reihe gemeinsamer Züge besitzt, so ist es darum möglich, im Gebiete solcher Verteilungen ein System der Häufigkeitskurven zu schaffen, was auch in der vorliegenden Arbeit durchgeführt worden ist.

Die typischen Formen der einmodigen Häufigkeitsverteilungen in Betracht ziehend, ist zur Grundlage des Systems die Sinuskurve im Umfange einer Welle gewählt worden.

Nehmen wir die Ordinatenachse durch den Gipfel der Welle, so ist die Gleichung

(31) $$y = 1 + \cos x,$$

und genügt in unserem Falle das Intervall $-\pi \leqslant x \leqslant \pi$.

Bei einer passenden Transformation der Abszissenskala können wir die Form der Kurve nach Bedarf verändern. Diese Funktion, welche den nötigen Anforderungen zu genügen hat, ist nach einer empirischen Ermittelung gewählt worden.

(32) $$x_1 = c\sqrt[n]{x},$$

wo c und n Konstanten sind. Setzen wir daraus den Wert von x in die Gleichung (31) ein, so bekommen wir

(33) $$y = 1 + \cos\left(\frac{x_1}{c}\right)^n.$$

Zum Bestimmen der Konstante c sei der Berührungspunkt der x-Achse mit der Kurve gegeben. Die Abszisse dieses Punktes sei A und der entsprechende x-Wert π. Auf Grund dieser Bedingungen ändert sich die Gleichung (32) zu

(34) $$A = c\sqrt[n]{\pi},$$

woraus

(35) $$c = \frac{A}{\sqrt[n]{\pi}}.$$

Setzen wir den letzteren Ausdruck für c in die Gleichung (33) ein, so bekommen wir

(36) $$y = 1 + \cos\pi\left(\frac{x_1}{A}\right)^n.$$

Nehmen wir A als eine neue Einheit und bezeichnen

$$\frac{x_1}{A} = X,$$

so erhalten wir

(37) $$y = 1 + \cos\pi X^n.$$

In welchem Masse der Wert von n die Gestalt der Kurve verändert, zeigt die folgende Figur (Fig. 7).

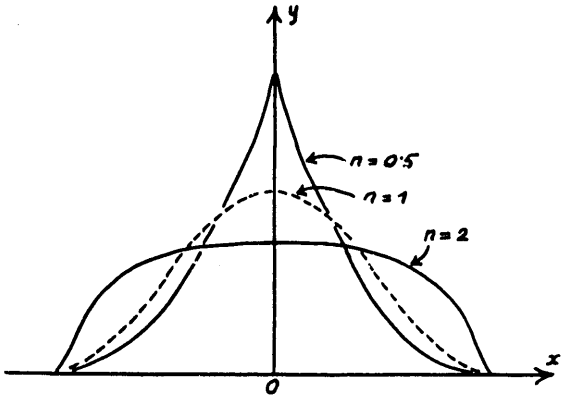


Fig. 7.

b. Der Vergleich mit der Normalkurve.

Da bei der Anwendung der genannten Kurve der Bedarf nach der Normalkurve fortfällt, muss nun festgestellt werden, bei welchem n die Kurve sich der Normalkurve am besten nähert. Da der Endpunkt der gegebenen Kurve bei $X = 1$ liegt, die Normalkurve aber bis in die Unendlichkeit läuft, kommt bei einer Näherung nur jener Teil der Normalkurve in Frage, der praktisch noch brauchbar ist. Dieser Teil ist bekanntlich durch die Breite $\pm 3\,\sigma$ gegeben und erfasst 99·7% aller Elemente. Da die Normalkurve bei $x = \sigma$ ihren Wendepunkt hat, können wir auch diese Bedingung in Betracht ziehen. Beim Bestimmen der Näherung sind drei Bedingungen beachtet worden:

1. es sollte bei gesuchtem n das Streuungsmass $\sigma = \frac{1}{3}$ sein;
2. bei gesuchtem n sollte die Abszisse des Wendepunktes $x = \frac{1}{3}$ sein;
3. bei gesuchtem n sollte die Abszisse des Wendepunktes gleich dem Streuungsmass sein.

Die Resultate sind folgendermassen errechnet:

1. Das Streuungsmass ist durch die folgende Formel bestimmbar:

$$\sigma = \sqrt{\frac{\int yx^2 dx}{\int y dx}}.$$

Setzen wir in dieser Formel statt y den Ausdruck aus (37) ein, so ist

$$\sigma = \sqrt{\frac{\int_0^1 (1+\cos \pi x^n) x^2 dx}{\int_0^1 (1+\cos \pi x^n) dx}} = \frac{1}{3} \tag{38}$$

(statt X ist x geschrieben). Die Lösung ergibt, dass

$$n = 0{\cdot}746$$

ist.

2. Zum Bestimmen des Wendepunktes kann die zweite Ableitung gleich Null gesetzt werden. Nach zweimaligem Differenzieren erhält man

$$\pi n x^{n-2} [\cos \pi x^n \cdot \pi n x^n + (n-1) \sin \pi x^n] = 0, \tag{39}$$

oder

$$\tan \pi x^n - \frac{n \pi x}{1-n} = 0. \tag{40}$$

Setzen wir hier $x = \frac{1}{3}$ ein, so ist

$$n = 0{\cdot}768.$$

3. Zur Bestimmung des n-Wertes, bei dem die Abszisse des Wendepunktes gleich dem Streuungsmass σ ist, müssen wir statt x in die Gleichung (40) den Ausdruck für σ aus (38) einsetzen. Die Berechnungen wurden praktisch so durchgeführt, dass mit Hilfe der Formel (38) für jeden n-Wert der Wert von σ bestimmt und letzterer in die Gleichung (40) eingesetzt wurde, bis die Gleichung erfüllt wurde. Die Berechnung ergab

$$n = 0{\cdot}771.$$

Um die Güte der drei erhaltenen n-Werte zu schätzen, wollen wir für einige Punkte die Ordinaten y mit denjenigen der Normalkurve vergleichen. Es seien die $x = 0$, $x = \sigma$ und $x = 2\sigma$ entsprechenden Punkte gewählt. Gleichzeitig berechnen wir die Fläche

$$N(x) = \int_0^x (1+\cos \pi x^n) dx \tag{41}$$

und die Fläche

$$N = \int_{-1}^{1} (1+\cos \pi x^n) dx. \tag{42}$$

Bezeichnen wir das Verhältnis

$$\frac{N(x)}{N} = Q(x),$$

so können wir es mit demselben Verhältnis bei der Normalkurve vergleichen. Beim Vergleichen der Ordinaten muss man wissen, dass bei der Normalkurve σ und ebenfalls die unter der Kurve liegende Fläche den Wert 1 haben. Darum muss man, um einen Vergleich anstellen zu können, die beiden Achsen entsprechend transformieren und darauf, was leicht zu ersehen ist, die aus Gleichung (37) berechneten $y(x)$-Werte mit dem Faktor $\frac{\sigma}{N}$ multiplizieren. Die erhaltenen Ordinatenwerte bezeichnen wir mit $Y(x)$. Die Ergebnisse sind in der folgenden Tabelle (Tab. 3) gegeben.

Tab. 3.

n	0·746	0·768	0·771	Normalkurve
σ	0·3333	0·3359	0·3362	
$y(\sigma)$	1·186	1·210	1·213	
$y(2\sigma)$	0·318	0·323	0·324	
$N(\sigma)$	0·551	0·561	0·562	
$N(2\sigma)$	0·790	0·807	0·810	
N	1·652	1·686	1·690	
$Y(0)$	0·403	0·399	0·398	0·399
$Y(\sigma)$	0·239	0·241	0·241	0·242
$Y(2\sigma)$	0·064	0·064	0·064	0·054
$Q(\sigma)$	0·336	0·333	0·333	0·341
$Q(2\sigma)$	0·479	0·479	0·479	0·477

Aus der Tabelle ist ersichtlich, dass die Ergebnisse, mit der Normalkurve verglichen, ganz geringe Unterschiede aufweisen und die Kurve als für eine Näherung zur Normalkurve genügend angenommen werden kann in dem Falle, wenn n zwischen 0·75 und 0·77 liegt.

c. Über das Bestimmen des n-Wertes bei gegebener Verteilung.

Das Bestimmen der Gleichung für eine empirische Häufigkeitsverteilung ist dann einfach, wenn die Form der Verteilung der Gleichung entspricht. Bestimmen wir aus Gleichung (36) n,

so erhalten wir

$$n = \frac{\log \pi - \log \arccos (y-1)}{\log A - \log x}, \tag{43}$$

wo statt x und y die Koordinaten eines jeden Punktes eingesetzt werden können. Falls aber die gegebene Verteilung der Gleichung nicht entspricht, so muss die Gleichung als eine Näherung aufgefasst werden und das Bestimmen der besten Näherung ist ein Problem für sich. Gewöhnlich wird zu diesem Zwecke die Methode der kleinsten Quadrate angewandt, bei Häufigkeitsverteilungen aber wird die Methode der Momente bevorzugt. Zwecks Erzielung der besten Annäherung ist in der vorliegenden Arbeit von einer einfachen Methode Gebrauch gemacht worden, die allerdings nicht ganz mechanisch anwendbar ist, da sie eine gewisse Freiheit in der Wahl zulässt. Wenn man aber das Vorhandensein des Streuungsstreifens bei den Lexis'schen Reihen und die schlechte Annäherungsmöglichkeit bei vielen Häufigkeitsverteilungen mit irgendwelchen Gleichungsformen in Betracht zieht, so kann die genannte Methode in den Dienst der gestellten Aufgabe genommen werden. Auf die nähere Betrachtung der Methode werden wir später zurückkommen, dort, wo die Annäherungsaufgabe gleichzeitig auch für die schiefen Häufigkeitskurven gelöst wird. Vorher wollen wir aber das Problem der Schiefheit und die der genannten Häufigkeitsverteilung entsprechende Gleichungsform betrachten.

V. Die schiefe Häufigkeitsverteilung.

Die schiefe Häufigkeitskurve können wir aus der symmetrischen erhalten, indem wir bei der letzteren die Abszissenskala mit Hilfe einer entsprechenden Funktion transformieren. Dabei ist die Bedingung aufgestellt worden, dass das Transformationsmass sich in der ganzen Variationsbreite nur in einer Richtung ändere. Dieses in Betracht ziehend, ist nach der Prüfung verschiedener Gleichungsformen vieler Häufigkeitsverteilungen zur Transformation eine Exponentialfunktion gewählt worden. Dieses bedeutet, dass eine schiefe Häufigkeitsverteilung durch eine zweckmässige Wahl der logarithmischen Skala zu einer symmetrischen wird.

Zur analogen Transformation gelangen wir, wenn wir das Bestimmen des oft gebrauchten geometrischen Mittels betrachten. Wenn wir es mit einer positiven Schiefe zu tun haben, dann weicht das arithmetische Mittel von der Mode ab, und zwar ist es grösser als die letztere. Bei Benutzung eines geometrischen Mittels, welches kleiner als das arithmetische ist, will man die Mode erreichen oder mindestens sich ihr nähern.

Wenn die Klassen des Merkmals $x_1, x_2, \ldots x_k$ die Häufigkeiten bzw. $m_1, m_2, \ldots m_k$ haben, dann wird das geometrische Mittel G durch folgende Formel ausgedrückt:

$$G = \sqrt[n]{x_1^{m_1} \cdot x_2^{m_2} \ldots x_k^{m_k}}, \tag{44}$$

wo $m_1 + m_2 + \ldots m_k = n$ ist.

Nach dem Logarithmieren bekommen wir

$$\log G = g = \frac{m_1 \log x_1 + m_2 \log x_2 + \ldots m_k \log x_k}{n}. \tag{45}$$

Ersetzen wir die x-Achse durch die Funktionsskala

$$\xi = \log x, \tag{46}$$

so ist

$$g = \frac{m_1 \xi_1 + m_2 \xi_2 + \ldots m_k \xi_k}{n}. \tag{47}$$

Das bedeutet, dass die Berechnung im neuen Koordinatensystem zum arithmetischen Mittel führt. Damit das erhaltene Mittel die Mode sei ($G = M_0$), muss die Häufigkeitskurve im neuen Koordinatensystem symmetrisch sein. Nehmen wir den Anfang des Koordinatensystems in der Mode und bezeichnen

$$\xi - g = X, \tag{48}$$

so können wir nach (45) und (46) schreiben:

$$X = \log x - \log G = \log \frac{x}{G}. \tag{49}$$

Nehmen wir bei gleichen Ordinaten die Abszissenwerte X_1 und X_2, denen auf der gegebenen Kurve irgendwelche x_1 und x_2 entsprechen, so können wir infolge der Symmetrie schreiben:

$$X_1 = -X_2, \tag{50}$$

oder

(51) $$\log \frac{x_1}{Mo} = -\log \frac{x_2}{Mo},$$

oder

(52) $$x_1 x_2 = Mo^2.$$

Da diese Beziehung nur dann gilt, wenn die Häufigkeitsverteilung der Logarithmen des Merkmals symmetrisch ist, so wird in der vorliegenden Arbeit der genannte Gedankengang so erweitert, dass der Ausgangspunkt, von welchem aus die Längen x gemessen werden, nicht im Koordinatenanfangspunkt gelassen wird, sondern ihm zwecks Erreichung symmetrischer Verteilung eine entsprechende Lage ausserhalb der Variationsbreite (damit die Beziehung $x_1 x_2 \quad Mo^2$ gelte) zugewiesen wird. Wenn dieser Punkt (nennen wir ihn den **Pol** und bezeichnen wir ihn mit P) auf der linken Seite der Verteilung liegt, so nennen wir die Schiefe positiv, und wenn der Pol auf der rechten Seite der Verteilung liegt, ist die Schiefe negativ. Je grösser die Schiefe ist, desto näher zur Kurve liegt P und umgekehrt.

Nehmen wir den Koordinatenanfangspunkt in der Mode (Fig. 8) und bezeichnen wir die Abszisse von P mit p (bei positi-

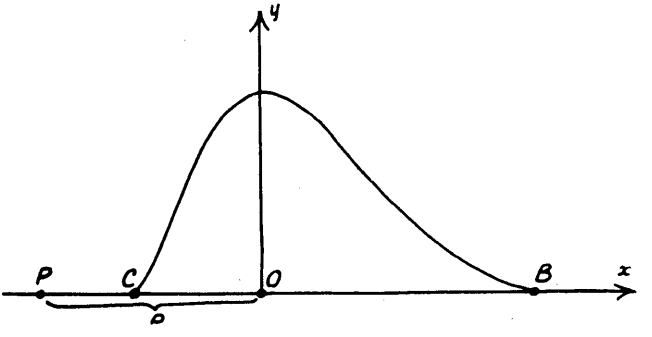


Fig. 8.

ver Schiefe ist $p < 0$ und bei negativer Schiefe $p > 0$), so können wir nach (49) schreiben:

(53) $$X = c \log \frac{-p+x}{-p} = c \log \frac{p-x}{p},$$

wo c eine Konstante ist.

Da negative Zahlen keinen Logarithmus haben, muss das Verhältnis

$$\frac{-p+x}{-p}$$

immer positiv sein. Dieses trifft tatsächlich zu, denn die Grösse $-p+x$ bezeichnet die Entfernung des Punktes x von P und $-p$ diejenige des Koordinatenanfangspunktes O von P. Da die beiden Punkte x und O zu einer Seite von P liegen, haben sie auch dieselben Vorzeichen.

Zum Bestimmen der Konstante c setzen wir $X=1$, dann ist $x=B$ (der rechtseitige Endpunkt der Kurve). Setzen wir diese Werte in die Gleichung (53) ein, so bekommen wir

$$c=\frac{1}{\log\frac{p-B}{p}}.$$

Damit ist

$$X=\frac{\log\frac{p-x}{p}}{\log\frac{p-B}{p}}. \tag{54}$$

Setzen wir dieses in Gleichung (37) ein, so erhalten wir die gesuchte Form der Gleichung:

$$y=1+\cos\pi\left[\frac{\log\frac{p-x}{p}}{\log\frac{p-B}{p}}\right]^n. \tag{55}$$

Statt der Grösse B kann die Koordinate des linksliegenden Endpunktes der Kurve, die wir mit C bezeichnen, eingeführt werden.

Es ist bekannt, dass $X=-1$ ist, wenn $x=C$ ist. Nach (53) bekommen wir, dass

$$c=-\frac{1}{\log\frac{p-C}{p}}=\frac{1}{\log\frac{p}{p-C}} \tag{56}$$

ist, und die Gleichung (55) erhält dann die Form

$$y=1+\cos\pi\left[\frac{\log\frac{p-x}{p}}{\log\frac{p}{p-C}}\right]^n. \tag{57}$$

Aus den Gleichungen (54) und (56) folgt, dass

$$\frac{p-B}{p}=\frac{p}{p-C} \tag{58}$$

ist, woraus

$$C=\frac{Bp}{B-p}, \tag{59}$$

oder

$$B=\frac{Cp}{C-p}, \tag{60}$$

oder

$$p=\frac{BC}{B+C} \tag{61}$$

folgt. (Hierbei muss im Auge behalten werden, dass immer $B>0$ und $C<0$ sind.)

Bei negativer Schiefe (Fig. 9) müssen die Distanzen von P nach links gemessen werden. Obwohl beide Distanzen (von P bis zur Mode und bis zum Punkte x) negativ sind, ist ihr Verhältnis positiv und erlaubt dieses die Anwendung der Formel (49).

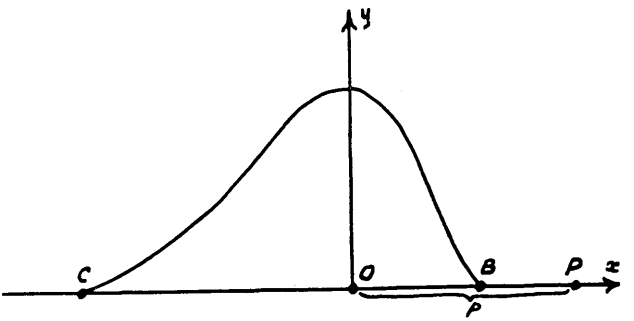


Fig. 9.

Setzen wir die entsprechenden Grössen ein, so erhalten wir genau die Gleichung (53):

$$X=c\log\frac{-p+x}{-p}. \tag{62}$$

Damit bleibt die Form der Gleichung dieselbe, und die Gleichung kann für beide Arten der Schiefe gebraucht werden. Der Unterschied besteht nur darin, dass im ersten Falle $p<0$ und im zweiten $p>0$ ist.

Die Grösse

$$\frac{p-B}{p}=\beta \tag{63}$$

kann als Mass der Schiefheit gewählt werden, denn bei positiver Schiefe, wo $p<0$, ist, wächst mit der Zunahme von B auch β und umgekehrt. Bei negativer Schiefe, wo $p>0$ ist, ist die Erscheinung umgekehrt.

Nach Formel (58) ist ebenso

$$\frac{p}{p-C}=\beta. \tag{64}$$

Aus (63) erhalten wir

$$B=p(1-\beta) \tag{65}$$

und aus (64)

$$C=\frac{p(\beta-1)}{\beta}. \tag{66}$$

Nach Division der Gleichungen (65) und (66) erhält man

$$\frac{B}{C}=-\beta. \tag{67}$$

Da $B>0$ und $C<0$ ist, so ist das Verhältnis der absoluten Grössen der Längen der Zweige der Häufigkeitsverteilung β (der rechtsliegende dividiert durch den linksliegenden).

Infolge von $\cos a = \cos(-a)$ sind in der Formel (55) das Zeichen des Bruches in den Klammern und der Charakter von n (eine gerade oder ungerade Zahl) nicht wesentlich. Darum kann in der Formel (55) im Nenner des Bruches statt $\frac{p-B}{p}$ auch $\frac{p}{p-B}$ stehen, oder nach (58) die allgemeine Gleichung in folgender Form geschrieben werden:

$$y=1+\cos\pi\left[\frac{\log\frac{p-x}{p}}{\log\frac{p-C}{p}}\right]^n.$$

Dieses bedeutet, dass es einerlei ist, in bezug auf welchen Endpunkt die Gleichung (58) gilt. Ob wir es mit positiver oder negativer Schiefe zu tun haben, kann an dem Zähler des Bruches, der folgendermassen geschrieben werden kann:

$$\log\frac{p-x}{p}=\log\left(1-\frac{x}{p}\right), \tag{68}$$

erkannt werden. Da bei positiver Schiefe $p<0$ ist, so ist im Zähler der Koeffizient vor x positiv. Bei negativer Schiefe, wo $p>$ ist, ist derselbe Koeffizient negativ.

VI. Die Beziehung zwischen der symmetrischen und der schiefen Häufigkeitskurve.

Aus der Formel (63) ist ersichtlich, dass mit dem Wachsen von p die Schiefe abnimmt, und praktisch kann die Kurve, wenn p genügend gross ist, als symmetrisch angesehen werden. Die Gleichung der symmetrischen Kurve (37) kann nicht als eine Ausnahme betrachtet werden, denn bei genügend grossem p ist der Unterschied zwischen den beiden Kurven wirklich gering und nähert sich beim Wachsen von p der Null. Um dieses zu erklären, schreiben wir den Bruchausdruck aus Formel (55) in Form einer Reihe, wobei wir die bekannte Reihenentwicklung durchführen. Auf diese Art erhalten wir

$$\frac{\log\frac{p-x}{p}}{\log\frac{p-B}{p}} = \frac{x+\frac{x^2}{2p}+\frac{x^3}{3p^2}+\dots}{B+\frac{B^2}{2p}+\frac{B^3}{3p^2}+\dots}. \tag{69}$$

Daraus ist ersichtlich, dass bei $p \to \infty$ der Grenzwert des Bruches $\frac{x}{B}$ ist; die Gleichung (55) bekommt dann die Form

$$y = 1+\cos\pi\left(\frac{x}{B}\right)^n, \tag{70}$$

was die uns bekannte Gleichung (36) darstellt.

In dem Falle, wenn $\beta \approx 1$ ist, kann man ebenfalls zur Reihenentwicklung greifen, und durch Fortfallen der Logarithmen wird die weitere Rechenarbeit erleichtert. In der neuen Form ist die Gleichung die folgende:

$$y = 1+\cos\pi\left\{\frac{x+\frac{x^2}{2p}+\frac{x^3}{3p^2}+\dots}{B+\frac{B^2}{2p}+\frac{B^3}{3p^3}+\dots}\right\}^n. \tag{71}$$

VII. Das Bestimmen der Charakteristiken.

Im allgemeinen Fall (wenn der Koordinatenanfangspunkt nicht in der Mode liegt) hat die Gleichung (55) die Form

$$y = A\left\{1+\cos\pi\left[\frac{\log(a+bx)}{\log\beta}\right]^n\right\}. \tag{72}$$

Die Gleichung enthält fünf Konstanten, deren Bedeutungen im folgenden erklärt werden.

1. β ist das früher erwähnte Mass der Schiefe.

2. n ist das ebenfalls früher erwähnte Mass der Grösse der Häufung.

3. A ist die Hälfte der Länge der Ordinate der Mode, denn der maximale Wert von y ist $y_{max} = 2A$.

Weiter hat der Ausdruck im Zähler $a + bx$ einige wichtige Bedeutungen:

4. Die Wurzel der Gleichung $a + bx = 1$ bestimmt die Mode Mo, wo

$$Mo = \frac{1 - a}{b} \tag{73}$$

ist.

5. Die Wurzel der Gleichung $a + bx = \beta$ bestimmt den rechtseitigen Endpunkt der Kurve, den Wert von B, wo

$$B = \frac{\beta - a}{b} \tag{74}$$

ist.

6. Die Wurzel der Gleichung $a + bx = \frac{1}{\beta}$ bestimmt den linkseitigen Endpunkt der Kurve, den Wert von C, wo

$$C = \frac{1 - a\beta}{b\beta} \tag{75}$$

ist.

7. Der Abstand der beiden Endpunkte B und C, die Variationsbreite v, ist nach (74) und (75)

$$v = \frac{\beta - a}{b} - \frac{1 - a\beta}{b\beta} = \frac{\beta^2 - 1}{b\beta}. \tag{76}$$

Beispiele:

1. Es sei die Gleichung gegeben

$$y = 4\left\{1 + \cos \pi \left[\frac{\log (-0{\cdot}37 + 0{\cdot}25\, x)}{\log 4{\cdot}00}\right]^{0{\cdot}80}\right\}.$$

Nach der Gleichung (73) erhalten wir

$$Mo = \frac{1 + 0{\cdot}37}{0{\cdot}25} = 5{\cdot}48.$$

Nach (74) erhält man

$$B = \frac{4{\cdot}00 + 0{\cdot}37}{0{\cdot}25} = 17{\cdot}48.$$

Nach (75) ist

$$C = \frac{1 + 0{\cdot}37.4{\cdot}00}{0{\cdot}25.4{\cdot}00} = 2{\cdot}48.$$

Nach (76) ist

$$v = \frac{4{\cdot}00^2 - 1}{0{\cdot}25.4{\cdot}00} = 15{\cdot}00.$$

Dasselbe ergibt auch eine folgende Nachprüfung, denn

$$v = B - C = 17{\cdot}48 - 2{\cdot}48 = 15{\cdot}00$$

und

$$\beta = \frac{B - Mo}{Mo - C} = \frac{17{\cdot}48 - 5{\cdot}48}{5{\cdot}48 - 2{\cdot}48} = 4{\cdot}00.$$

Ferner ist

$$p = -\frac{1}{b} = -\frac{1}{0{\cdot}25} = -\ 4{\cdot}00,$$

woraus

$$P = Mo + p = 5{\cdot}48 - 4{\cdot}00 = 1{\cdot}48.$$

Damit sind alle wichtigen zum Zeichnen der Kurve notwendigen Punkte gefunden.

Ebenso einfach ist die Lösung der umgekehrten Aufgabe.

2. Es seien die beiden Endpunkte der Kurve und die Lage der Mode gegeben. Man muss die Gleichung aufschreiben.

Es sei

$$Mo = 2, \quad B = 16 \quad \text{und} \quad C = -5.$$

Zum Bestimmen der Grössen a, b und β haben wir vier Gleichungen, von denen wir irgendwelche drei benutzen können. Der einfachste Weg im gegebenen Falle ist aber folgender:

β ist schon durch drei gegebene Punkte bestimmt, denn

$$\beta = \frac{B - Mo}{Mo - C} = \frac{16 - 2}{2 + 5} = 2.$$

Ebenso ist auch v bestimmt, denn

$$v = B - C = 16 + 5 = 21.$$

Aus der Gleichung (76) erhalten wir, dass

$$b = \frac{\beta^2 - 1}{\beta v}$$

ist. Setzen wir hier die gegebenen Zahlen ein, so ist

$$b = \frac{2^2 - 1}{2.21} = \frac{1}{14}.$$

Damit ist

$$p = -14$$

und

$$P = Mo + p = 2 - 14 = -12.$$

Aus der Gleichung (73) erhält man

$$a = 1 - bMo = \frac{6}{7}.$$

Nehmen wir z. B. an, dass $A = 4$ und $n = 0{\cdot}80$ ist, dann gilt die Gleichung

$$y = 4\left\{1 + \cos \pi \left[\frac{\log\left(\frac{6}{7} + \frac{1}{14}x\right)}{\log 2}\right]^{0{\cdot}80}\right\}.$$

Bemerkung: Es ist zu empfehlen, die Variationsbreite v immer positiv zu nehmen, d. h. $v = B - C$. Dann sieht man aus der Gleichung (76), dass bei negativer Schiefe, wo $b < 0$ ist, der Nenner negativ ist und deshalb auch der Zähler negativ sein muss. Wenn wir die Schiefe β als Verhältnis der Länge des rechtseitigen Zweiges zu derjenigen des linkseitigen annehmen, dann ist es tatsächlich so, denn dann ist $\beta < 1$. Ohne diese Forderung (dass $v > 0$ ist) können wir immer, wenn $\beta < 1$ ist, statt β auch $\frac{1}{\beta}$ schreiben, denn in der Gleichung verändert sich nichts, da der absolute Wert von $\log\beta$ derselbe bleibt.

VIII. Über die Wahl der Methode zum Bestimmen der Charakteristiken.

Zum Aufschreiben der Gleichung brauchen wir fünf Konstanten. Da die Zahl der Häufigkeitszahlen gewöhnlich grösser ist, ist das Problem nicht eindeutig lösbar. Es gibt unendlich viele Lösungen. Unter diesen muss man eine solche auswählen, die eine

„möglichst gute" Annäherung gibt. Das Bestimmen der „allerbesten Annäherung" ist ein Problem für sich, und es gibt keine Methode, die es mit genügendem Erfolg lösen könnte. Die klassische Methode ist diejenige der kleinsten Quadrate, deren Anwendung aber bei komplizierten Gleichungsformen sehr ungünstig ist. Pearson und Charlier haben die Methode der Momente angewandt, die aber mit zeitraubender Rechenarbeit verbunden ist. Andere Nachteile sind in dieser Arbeit schon analysiert worden. Zur Lösung des Problems ist in der vorliegenden Arbeit eine einfache Methode, die wir die Flächenmethode nennen, angewandt worden, die mit einfacher Rechenarbeit verbunden und bei verschiedenartigen Gleichungsformen brauchbar ist. Die Methode besteht darin, dass ein Flächenstück unter der Häufigkeitskurve (die Häufigkeit irgendwelcher Klasse) dem entsprechenden Teil der empirischen Kurve gleich gesetzt wird. Die Zahl der Flächenstücke wird nach der Zahl der zu bestimmenden Charakteristiken gewählt. Die Methode ist bei Häufigkeitsverteilungen leicht anwendbar, denn da sind die Grössen der Flächenstücke schon durch die Häufigkeitszahlen gegeben. Ein Vergleich mit der Methode der kleinsten Quadrate ergab, dass die Genauigkeit der beiden Methoden ungefähr dieselbe ist. Die neue Methode ist teilweise besser, teilweise schlechter, je nachdem, welche Flächenstücke genommen werden. Im folgenden vergleichen wir, um Raum zu sparen, die genannte Methode mit derjenigen der kleinsten Quadrate nur bei einer linearen Funktion.

IX. Vergleich mit der Methode der kleinsten Quadrate.

1. Es stelle im Koordinatensystem xy die fett ausgezogene horizontale Linie (Fig. 10) den Gang der Häufigkeitsverteilung in der Variationsbreite l dar. Es sei der Kollektivumfang s. Nehmen wir an beiden Enden der Verteilung Flächenstücke mit der Grundlinie k (k ist die Klassenbreite) und ziehen wir aus den Mittelpunkten der Grundlinien die Höhen h_1 und h_2. Die oberen Punkte der Höhen A und B verbinden wir durch Gerade, die wir als Näherungsgerade wählen.

Nach dem Bernoulli'schen Gesetz hat jedes Flächenstück,

welches die Häufigkeit eines gewissen Merkmals anzeigt, die mittlere Abweichung σ, die nach der Formel

$$\sigma = \sqrt{sp\,(1-p)}$$

bestimmt werden kann (s — Kollektivumfang, p — die relative Häufigkeit der zu betrachtenden Klasse).

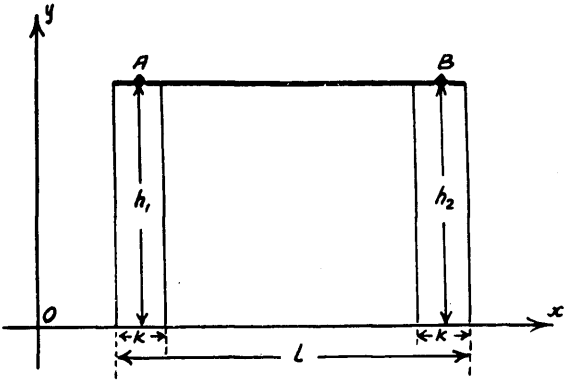


Fig. 10.

Nach der angenommenen Verteilungsfunktion ist

$$p = \frac{k}{l}.$$

Damit ist die mittlere Abweichung der beiden Flächenstücke

$$\sigma_1 = \sigma_2 = \sqrt{\frac{sk\,(l-k)}{l^2}} \tag{77}$$

und die mittlere Abweichung der Höhen h

$$\sigma h_1 = \sigma h_2 = \frac{\sigma_1}{k} = \frac{\sigma_2}{k} = \sqrt{\frac{s(l-k)}{kl^2}}. \tag{78}$$

Die mittlere Abweichung des Steigungsmasses der Näherungsgeraden wird durch die mittlere Abweichung der Differenz der beiden Höhen ($\Delta h = h_1 - h_2$) bestimmt, wobei die letztgenannte mittlere Abweichung noch durch die Differenz der Abszissen der Punkte A und B $l - k$ dividiert werden muss.

Beim Bestimmen der mittleren Abweichung der Höhendifferenz muss in Betracht gezogen werden, dass die Abweichungen der Flächenstücke und damit auch die Abweichungen der

Höhen miteinander korrelativ verbunden sind. Die mittlere Abweichung der Differenz der Höhen $\sigma\Delta h$ wird durch folgende Formel bestimmt:

$$\sigma^2\Delta h = \sigma^2 h_1 + \sigma^2 h_2 - 2r \,.\, \sigma h_1 \,.\, \sigma h_2 . \tag{79}$$

Der Korrelationsfaktor r ist eine Funktion von k (bei kleinem k ist $r \approx o$, ist aber $k = \frac{l}{2}$, so ist $r = -1$), und wir können ihn folgendermassen finden:

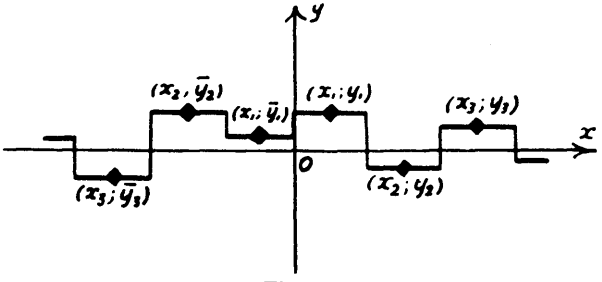


Fig. 11.

Wenn im ersten Flächenstück eine Abweichung δ_1 stattgefunden hat, dann zeigt das übriggebliebene Flächenstück die Abweichung $-\delta_1$. Daher fällt auf das zweite Flächenstück die Abweichung

$$\delta_2 = -\frac{k}{l-k}\delta_1 . \tag{80}$$

Aus der Korrelationstheorie wissen wir, dass die Regressionsgleichung für diese Abweichungen die folgende ist:

$$\delta_2 = r\frac{\sigma_2}{\sigma_1}\delta_1 , \tag{81}$$

wo σ_1 und σ_2 die in der Formel (77) vorkommenden Grössen sind.

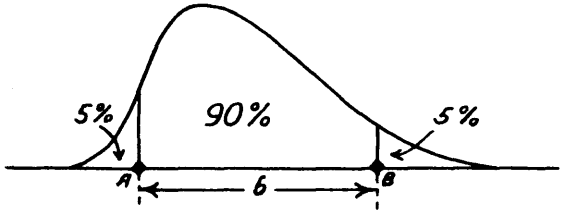


Fig. 12.

Dividieren wir die Gleichung (80) durch (81), so bekommen wir

$$r = -\frac{k}{l-k}. \tag{82}$$

Hieraus ersehen wir, dass wenn $k = \frac{l}{2}$ ist, dann $r = -1$ ist.

Im Falle $k > \frac{l}{2}$ ist

$$r = -\frac{l-k}{k} \tag{83}$$

(der reziproke Wert des früheren r), was leicht ableitbar ist.

Setzen wir aus (78) und (82) die entsprechenden Grössen in (79) ein, so erhalten wir

$$\sigma^2 \Delta h = \frac{2s}{kl}, \tag{84}$$

denn $\sigma h_1 = \sigma h_2$, und der Korrelationsfaktor zwischen den Abweichungen der Grössen der Flächenstücke ist derselbe wie der Korrelationsfaktor zwischen den Abweichungen der Höhen.

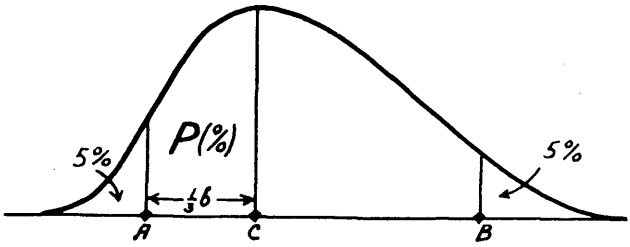


Fig. 13.

Zum Bestimmen der mittleren Abweichung des Steigungsmasses m dividieren wir den erhaltenen $\sigma\Delta h$-Wert noch durch die Grösse $l - k$. Dann bekommen wir

$$\sigma^2 m = \frac{2s}{kl(l-k)^2}. \tag{85}$$

Ist $k > \frac{l}{2}$, so erhält man mit Hilfe der Formel (83) analog

$$\sigma^2 m = \frac{2s}{k^2 l(l-k)}. \tag{86}$$

Wenn wir beim Wählen der Flächenstücke die kleinere von den Grössen k und $l-k$ mit γ und die grössere mit $l-\gamma$ bezeichnen, dann ist im ersteren Falle (wo $k<\frac{l}{2}$ ist)

$$k=\gamma$$

und

$$l-k=l-\gamma,$$

und im letzteren Falle (wo $k>\frac{l}{2}$ ist)

$$k=l-\gamma$$

und

$$l-k=\gamma.$$

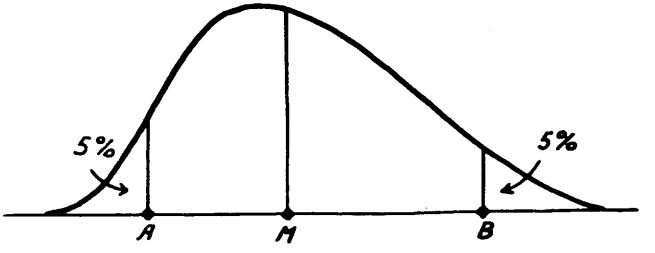


Fig. 14.

Setzen wir diese Grössen bzw. in Formel (85) und (86) ein, dann erhalten wir für beide Fälle

$$\sigma^2 m=\frac{2s}{\gamma l(l-\gamma)^2}. \tag{87}$$

Dieses bedeutet, dass die Formel (85) immer angewandt werden kann, es muss nur unter k der kleinere Teil von l verstanden werden.

Bezeichnen wir

$$k=\varkappa l,$$

dann ist nach (85)

$$\sigma^2 m=\frac{2s}{l^4\varkappa(1-\varkappa)^2}. \tag{88}$$

Berechnen wir den Minimalwert von σm, so finden wir, dass dieser bei $x=\frac{1}{3}$ zutage tritt und seine Grösse

(89) $$\sigma m_{min} = 3{\cdot}67 \frac{\sqrt{s}}{l^2}$$

beträgt.

2. Zweitens wollen wir betrachten, mit einer wie grossen Genauigkeit das Steigungsmass der Näherungsgeraden bei der Methode der kleinsten Quadrate bestimmt wird.

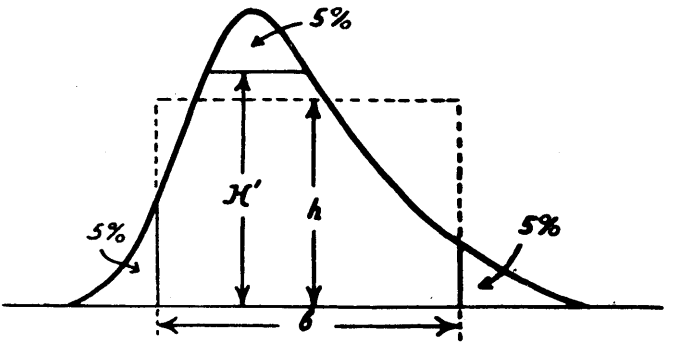


Fig. 15.

Es sei ebenso die Zahl der Elemente s, die Klassenbreite k und die Variationsbreite l (Fig. 9).

Wenn das Steigungsmass der Regressionsgeraden gleich Null ist, dann ist

(90) $$\sigma m = \frac{\sigma y}{\sigma x} \sigma r$$

(darin bedeutet: σm die gesuchte mittlere Abweichung des Steigungsmasses m, σy und σx die Streuungsmasse der Koordinaten der die Regressionsgleichung bestimmenden Punkte A, B und anderer dazwischen liegender Punkte, sowie σr die mittlere Abweichung des Korrelationsfaktors r), denn nach der Korrelationstheorie ist

$$m = r \frac{\sigma y}{\sigma x}$$

und nach Differenzierung erhalten wir

(91) $$dm = \frac{\sigma y}{\sigma x} dr + r \frac{d\sigma y}{\sigma x} - r \frac{\sigma y \, d\sigma x}{\sigma x^2}.$$

Da $r = 0$ ist, bleibt übrig

$$dm = \frac{\sigma y}{\sigma x} dr,$$

woraus die Formel (90) ableitbar ist. Damit müssen zum Bestimmen von σm die Grössen σy, σr und σx berechnet werden.

a. σy ist die mittlere Abweichung der Höhe, die schon nach der Formel (78) bestimmt worden ist. Damit ist

$$\sigma y = \sqrt{\frac{s(l-k)}{kl^2}}. \tag{92}$$

b. Nach der Korrelationstheorie ist

$$\sigma r = \frac{1 - r^2}{\sqrt{n}},$$

wo r der Korrelationsfaktor und n die Zahl der Klassen ist.

Da $r = 0$ und $n = \frac{l}{k}$ ist, bekommen wir

$$\sigma r = \sqrt{\frac{k}{l}}.$$

Das erhaltene σr gilt nur dann, wenn die Abweichungen der Höhen nicht in korrelativem Zusammenhang stehen. In unse-

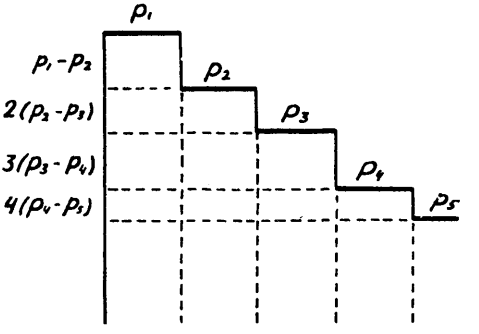


Fig. 16.

rem Falle, wo die genannten Abweichungen in korrelativer Beziehung stehen, wird σr infolge einer negativen Korrelation grösser.

Um das gesuchte σr zu bestimmen, schreiben wir in der Formel des Korrelationsfaktors

$$r = \frac{\Sigma xy}{n \sigma x \, \sigma y}$$

(wo x und y vom arithmetischen Mittel aus genommen werden

und n die Zahl der Punkte, in unserem Falle die Zahl der Klassen bedeutet) den Zähler Σxy in der Form

$$\Sigma xy = x_1(y_1 - \bar{y}_1) + x_2(y_2 - \bar{y}_2) + \ldots$$

(das Bezeichnungsschema ist in Fig. 11 gegeben).

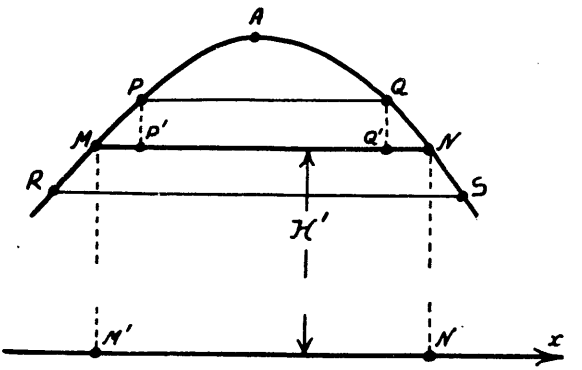


Fig. 17.

Da die x-Werte unverändlich sind, können, wenn ein korrelativer Zusammenhang mit dem Korrelationsfaktor r zwischen den Häufigkeitszahlen der Klassen vorhanden ist, nur die mittleren Abweichungen der Ausdrücke in den Klammern ($y_i - \bar{y}_i = \Delta y_i$) sich verändern.

Nach der Korrelationstheorie ist

$$\sigma^2 \Delta y = 2\sigma y (1 - r), \tag{93}$$

woraus ersichtlich ist, dass bei dem Korrelationsfaktor r der Wert von $\sigma \Delta y$ sich verändert hat, wobei der Veränderungskoeffizient

$$\lambda = \sqrt{1 - r} \tag{94}$$

ist. Damit ist auch σr λ-mal grösser geworden. Setzen wir den Ausdruck von r aus (82) ein, so finden wir, dass

$$\lambda = \sqrt{\frac{l}{l-k}}$$

ist, und damit

$$\sigma r = \sqrt{\frac{k}{l-k}}. \tag{95}$$

c. Wenn alle Klassen die gleiche Breite haben, befinden

sich die Mittelpunkte der Klassen mit konstanter Häufigkeit auf der x-Achse, und nach (8) ist

$$\sigma x = \frac{l}{2\sqrt{3}}. \tag{96}$$

Setzen wir die erhaltenen Werte für σy, σr und σx aus (92), (95) und (96) in die Gleichung (90) ein, so erhalten wir, dass

$$\sigma^2 m = \frac{12s}{l^4} \tag{97}$$

ist, oder

$$\sigma m = 3{\cdot}46\,\frac{\sqrt{s}}{l^2}. \tag{98}$$

Durch einen Vergleich des Erhaltenen mit (89) stellen wir eine ganz kleine Verminderung von σm fest.

In der Tat ist aber σm bei Anwendung der Methode der kleinsten Quadrate grösser, denn die Formel (90) gilt für den Fall, wenn $r = 0$ ist. Bei einem endlichen Kollektivumfang hat r einen von Null abweichenden Wert und kommen dann in der Formel (91) die weggelassenen Glieder zur Geltung.

Betrachten wir z. B. einen Spezialfall, wo

$$r = \sigma r$$

ist. Da die Abweichungen von r, σy und σx miteinander nicht in korrelativem Zusammenhang stehen, können wir nach (91) schreiben:

$$\sigma^2 m = \frac{\sigma^2 y}{\sigma^2 x}\sigma^2 r + \frac{\sigma^2 r}{\sigma^2 x}\sigma^2{}_{\sigma y} + \frac{\sigma^2 r\,\sigma^2 y}{\sigma^4 x}\sigma^2{}_{\sigma x} = \tag{99}$$

$$= \frac{\sigma^2 y}{\sigma^2 x}\sigma^2 r\left(1 + \frac{\sigma^2{}_{\sigma y}}{\sigma^2 y} + \frac{\sigma^2{}_{\sigma x}}{\sigma^2 x}\right) = \frac{\sigma^2 y}{\sigma^2 x}\sigma^2 r\left(1 + \frac{1}{n}\right),$$

denn der Variationsfaktor von σx und σy ist $\frac{1}{\sqrt{2n}}$.

Aus (99) wird ersichtlich, dass σm sich um das $\sqrt{1 + \frac{1}{n}}$-fache vergrössert hat, und z. B. im Falle $n < 9$ (bei einer kleinen Zahl der Klassen) ist die Genauigkeit der Methode der kleinsten Quadrate geringer als diejenige der Flächenmethode.

Weiter ist ebenso leicht zu beweisen, dass in dem Falle, wo das Steigungsmass der Regressionsgeraden von Null verschieden ist ($m \neq 0$), die Genauigkeit der Methode der kleinsten Quadrate kleiner (σm wird grösser), bei der Flächenmethode aber grösser wird (σm wird kleiner). Z. B. in dem Falle, wo die Gerade den Koordinatenanfangspunkt schneidet, erhalten wir nach einfacher Berechnung analog der Formel (88)

$$\sigma^2 m = \frac{2s}{l^4 \varkappa (1-\varkappa)^2}(1 - 2\varkappa + 4\varkappa^2 - 2\varkappa^3). \tag{100}$$

Hier ist σm kleiner als dasselbe in Formel (88), denn

$$1 - 2\varkappa + 4\varkappa^2 - 2\varkappa^3 = 1 - 2\varkappa(1-\varkappa)^2 < 1.$$

σm hat ein Minimum bei demselben Argumentwert, d. h. bei $\varkappa = \frac{1}{3}$ und

$$\sigma m_{min} = 3{\cdot}08 \frac{\sqrt{s}}{l^2}, \tag{101}$$

was kleiner als dasselbe in (98) ist.

Die bei linearer Häufigkeitsverteilung kurz durchgeführte Analyse hat gezeigt, dass die Flächenmethode bei zweckmässiger Wahl der Flächenstücke nicht als schlechter denn die Methode der kleinsten Quadrate gelten kann. Ihre Anwendung ist aber viel einfacher. Bei der linearen Approximation brauchen z. B. nur die Häufigkeitszahlen summiert zu werden, und das Steigungsmass der Geraden ist bestimmt.

X. Über die Anwendung der Flächenmethode beim Bestimmen der Charakteristiken.

Bei der in der vorliegenden Arbeit untersuchten Gleichung der Häufigkeitskurve ist die Anwendung der Flächenmethode nicht so einfach wie bei dem oben gegebenen Beispiel, weil die Gleichung nach einer elementaren Methode nicht integrierbar ist. Darum sind die entsprechenden Berechnungen nach graphischen Methoden durchgeführt und die Resultate auch graphisch dargestellt worden. Da die Genauigkeit der Koeffizienten von der Wahl der Flächenstücke abhängt, mussten zum Aussuchen der passenden Flächenstücke (Indikatoren), die ein genaueres Bestimmen von β

und n ermöglichten, verschiedene Fälle untersucht und die entsprechenden Fehlerberechnungen durchgeführt werden. Durch diese zwei Koeffizienten ist der Charakter der Häufigkeitsverteilung gegeben, denn die drei übrigen bestimmen nur den Mass-

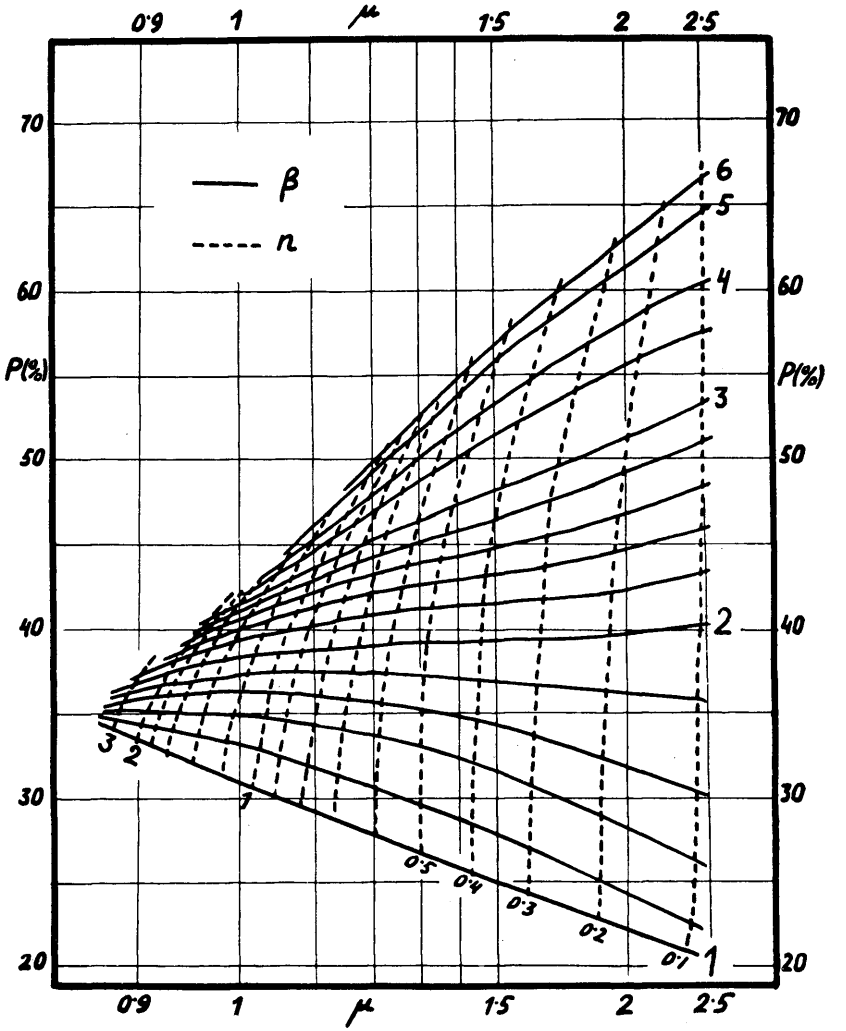


Fig. 18.

stab der x- und y-Achse und die Lage der Verteilung in bezug auf die x-Achse.

Bevor wir zur Lösung der Aufgabe übergehen, müssen wir

eine Möglichkeit zum Fixieren zweier durch die Häufigkeitsverteilung verbundener Punkte schaffen, denn nur dadurch kann die Lage der Verteilung auf der x-Achse bestimmt werden. Es hat sich als am zweckmässigsten erwiesen, einen 5%-gen Abschnitt zu wählen (die Punkte A und B in Fig. 12), d. h. einen solchen Punkt, dass an der einen Seite von ihm 5% aller Elemente liegen. Den Abschnitt zwischen den genannten Punkten, die Länge AB, werden wir die Basis nennen und mit b bezeichnen.

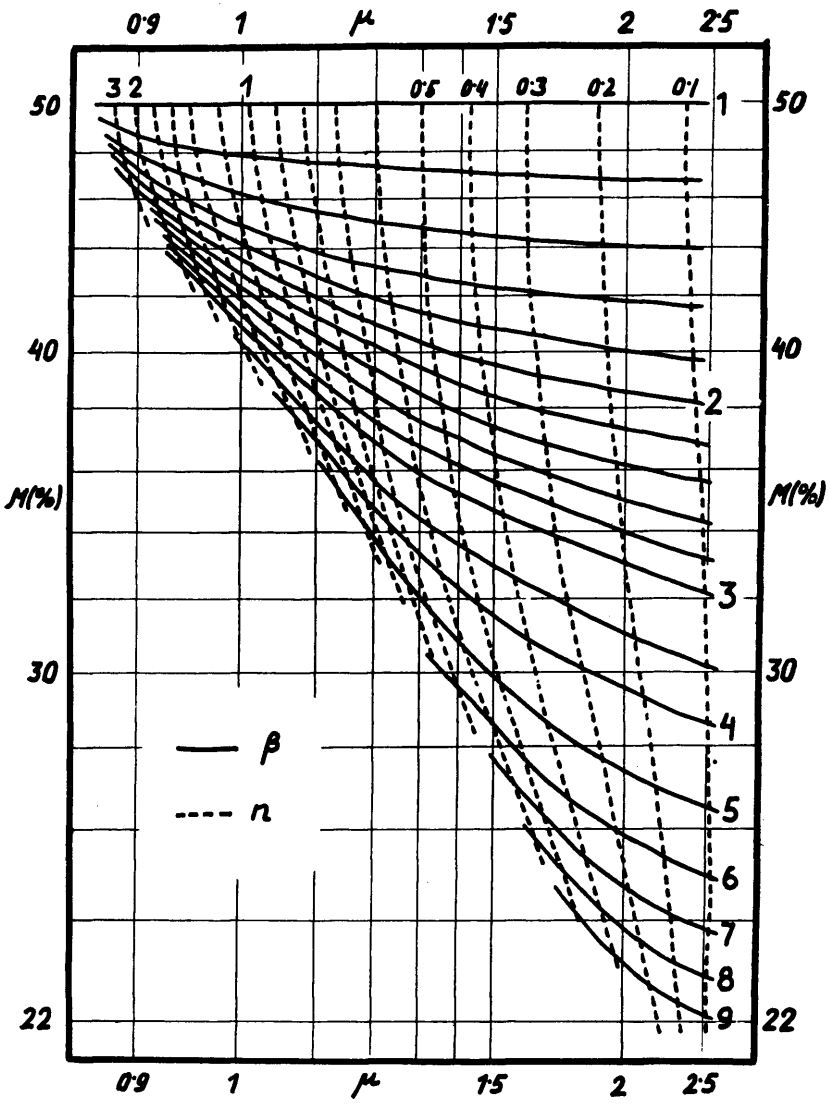
Nach hinreichender Erwägung bleiben von den Indikatoren drei übrig.

1. Der erste Indikator ist die relative Häufigkeit einer mit der Basis in Verbindung stehenden Klasse, wobei die Klassenbreite ein Drittel der Basis ausmacht und die Klasse selbst auf derjenigen Seite der Basis liegt, wo sich die Mode befindet (in Fig. 13 ist die genannte Klassenbreite AC). Es ist am einfachsten, die gesamte Zahl der Elemente bis zum Punkte C zu zählen, um den 5%-gen Teil nicht abziehen zu müssen. Diesen Indikator bezeichnen wir mit $P(\%)$, da die relative Häufigkeit in Prozenten angegeben wird.

2. Der zweite Indikator ist der Abstand der Mediane von dem der Mode näher liegenden Endpunkt der Basis (in Fig. 14 die Grösse AM), wobei diese Distanz als Verhältnis zu der Länge der Basis genommen wird. Diesen Indikator bezeichnen wir mit $M\ (\%)$. Damit ist $M\,\% = \dfrac{AM}{AB}$.

Bemerkung 1. Hier besteht ein Unterschied zwischen diesen Charakteristiken: die erste $P\ (\%)$ ist die Häufigkeit einer bestimmten Klasse, die zweite $M\ (\%)$ aber die Klassenbreite einer bestimmten Häufigkeit.

Bemerkung 2. Wenn die Verteilung ganz nahe einer symmetrischen liegt, ist es schwer zu bestimmen, zu welchem Endpunkte der Basis die Mode näher ist. Für den richtig gewählten Punkt gilt $M(\%) < {}^1/_2$ oder $P_A\ (\%) > P_B\ (\%)$, wenn der Punkt A der richtige ist (hier bedeutet $P_A(\%)$ den Wert von $P\ (\%)$, wobei die Klassenbreite vom Punkte A aus gemessen wird, und $P_B(\%)$ analog den Wert von $P\ (\%)$ für die Klasse, die vom Punkte B aus gemessen wird). Wenn die Verteilung eine solche ist, dass wir z. B. mit Hilfe von $P\ (\%)$ die positive Schiefe und mit Hilfe



- Fig. 19.

von M (%) die negative Schiefe erhalten, dann kann für die endgültige Schiefe das Mittel aus diesen zwei Werten genommen werden.

3. Die beiden gewählten Indikatoren sind empfindlich gegen Veränderungen von β, aber sie reagieren ganz schwach auf Veränderungen von n. Zum besseren Charakterisieren von n

ist ein dritter Indikator μ geschaffen worden, der folgendermassen bestimmt wird.

Schneiden wir bei der Häufigkeitsverteilung durch eine der x-Achse parallele Gerade einen Teil ab, dessen Fläche 5% der unter der Häufigkeitskurve liegenden Fläche beträgt (Fig. 15). Den Abstand dieser Geraden von der x-Achse bezeichnen wir mit H^1. Ferner berechnen wir die auf die Basis b reduzierte mittlere Häufigkeit h, d. h. $h = \frac{s}{b}$, wobei s die Zahl der Elemente ist. Durch diese zwei Grössen H^1 und h ist μ folgendermassen bestimmt: $\mu = \frac{H^1}{h}$.

Da H^1 aus den Häufigkeitzahlen nicht direkt bestimmt werden kann, müssen wir seine Entstehung näher betrachten.

Wenn der Kollektivumfang praktisch unendlich gross wäre, wäre die relative Häufigkeit jeder Klasse praktisch genau bestimmbar. In einem solchen Falle ist es sehr einfach, den genannten Schnitt zu machen. Wir schreiben die relativen Häufigkeiten in eine Reihe nach ihren Grössen, z. B. p_1, p_2, p_3 ... (Fig. 16), und finden die Flächengrösse bis zu der Höhe p_2, weiter bis p_3 usw., bis die Grösse der abgeschnittenen Fläche 5% der Gesamtfläche ausmacht, wobei nach Bedarf ein Interpolieren stattfinden kann. Auf diese Art stufenweise weitergehend, stellen wir fest, dass die abgeschnittene Fläche nach folgendem Gesetz zunimmt:

$$Q = p_1 - p_2 + 2(p_2 - p_3) + 3(p_3 - p_4) + \ldots n(p_n - p_{n+1}) =$$
$$= p_1 + p_2 + p_3 + \ldots p_n - np_{n+1} =$$
$$= \sum_{i=1}^{n} p_i - np_{n+1}.$$

Beispiel:

i	1	2	3	4	5	6	7	8
p_i (%)	7·8	7·3	6·9	6·7	6·5	6·1	5·7	5·4
Σp_i	7·8	15·1	22·0	28·7	35·2	41·3	47·0	52·4
ip_{i+1}	7·3	13·8	20·1	26·0	30·5	34·2	37·8	
Q	0·5	1·3	1·9	2·7	4·7	7·1	9·2	

Wie aus dem Schema ersichtlich ist, wächst die Fläche Q um 0·5%, wenn die Schnittlinie die Höhe p_2 erreicht hat. Wenn die Schnittlinie die Höhe p_3 erreicht, ist die Fläche Q bis 1·3% gewachsen, usw. Um für $Q = 5\%$ die entsprechende Höhe H^1 zu

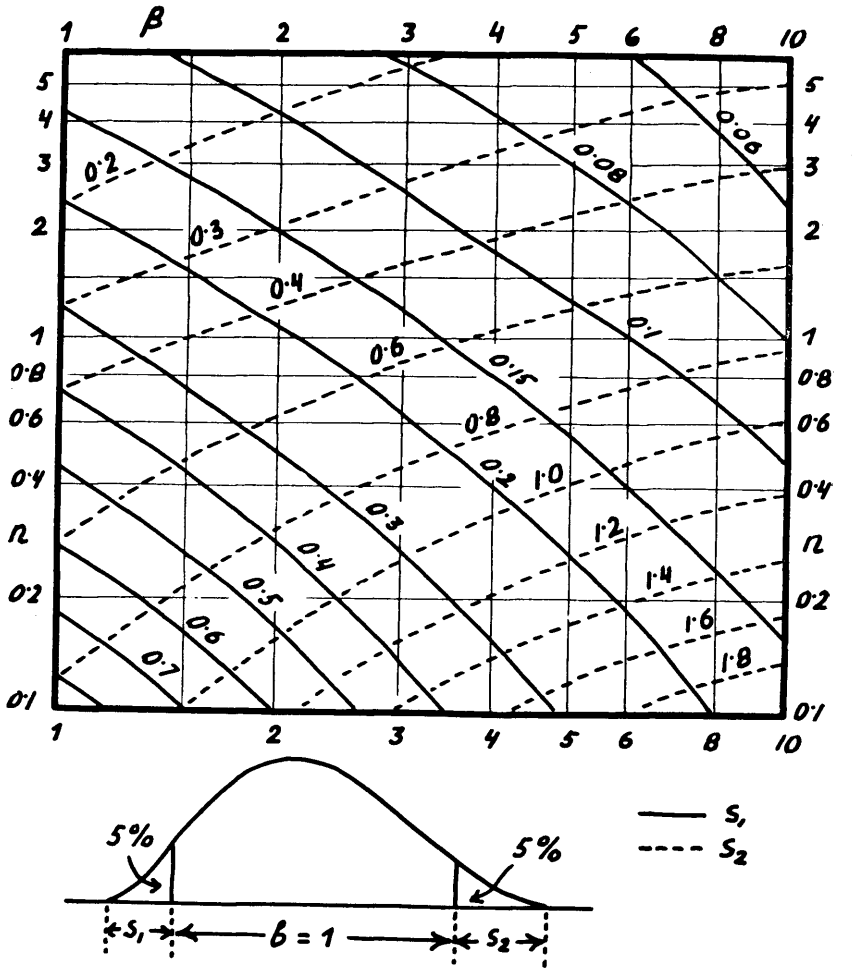


Fig. 20.

finden, müssen wir feststellen, bei welchem p die Fläche Q auf 5% wächst. Bis zur sechsten Stufe (der Wert von p ist dann 6·1 %) ist $Q = 4{\cdot}7\%$, weiter aber schon 7·1%. Bei dem ersten Q-Wert fehlen noch 0·3%, und da die Schnittlinie jetzt durch 6 Klassen geht, muss er um $\frac{0{\cdot}3\,\%}{6} = 0{\cdot}05\%$ sinken. Damit ist $H^1 = 6{\cdot}10 - 0{\cdot}05 = 6{\cdot}05\,\%$. Da H^1 nicht in Prozenten, sondern in Längeneinheiten gemessen wird, bedeutet das Resultat, dass H nach der auf die y-Achse aufgetragenen Skala einen solchen Wert besitzt, der gleich ist der Höhe der entsprechenden Klasse mit der Häufigkeit 6·05%.

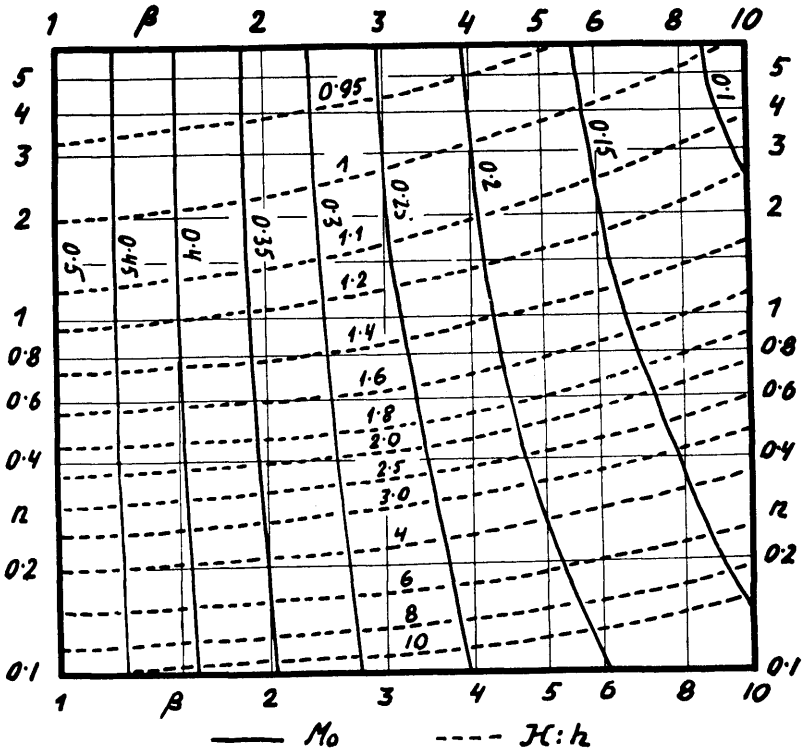


Fig. 21.

Wenn der Kollektivumfang nicht genügend gross ist, werden die bei der Bernoulli'schen oder der Lexis'schen Reihe in jeder Klasse entstandenen Abweichungen störend wirken. Wenn wir dann die Häufigkeiten nach ihren Grössen ordnen, kann es der Fall sein, dass wir den Wert von H^1 grösser erhalten, als er tatsächlich ist, denn es ist sehr wahrscheinlich, dass wir nur Klassen mit positiven Abweichungen genommen haben. Wenn die Häufigkeiten nach beiden Seiten gleichmässig abnehmen, ist die genannte Gefahr nicht so gross, denn dann ist es nicht wahrscheinlich, dass alle Klassen positive Abweichungen haben. Wenn aber bei der Abnahme der Häufigkeiten der Klassen gewisse Schwankungen entstehen, so ist es sehr gefährlich, nur die grösseren Häufigkeiten auszuwählen. Im folgenden ist eine Methode gegeben, deren man sich in solchen Fällen mit Erfolg bedienen kann.

Setzen wir zuerst voraus, dass der Kollektivumfang praktisch unendlich gross ist. Weiter nehmen wir an, dass der 5%-ge

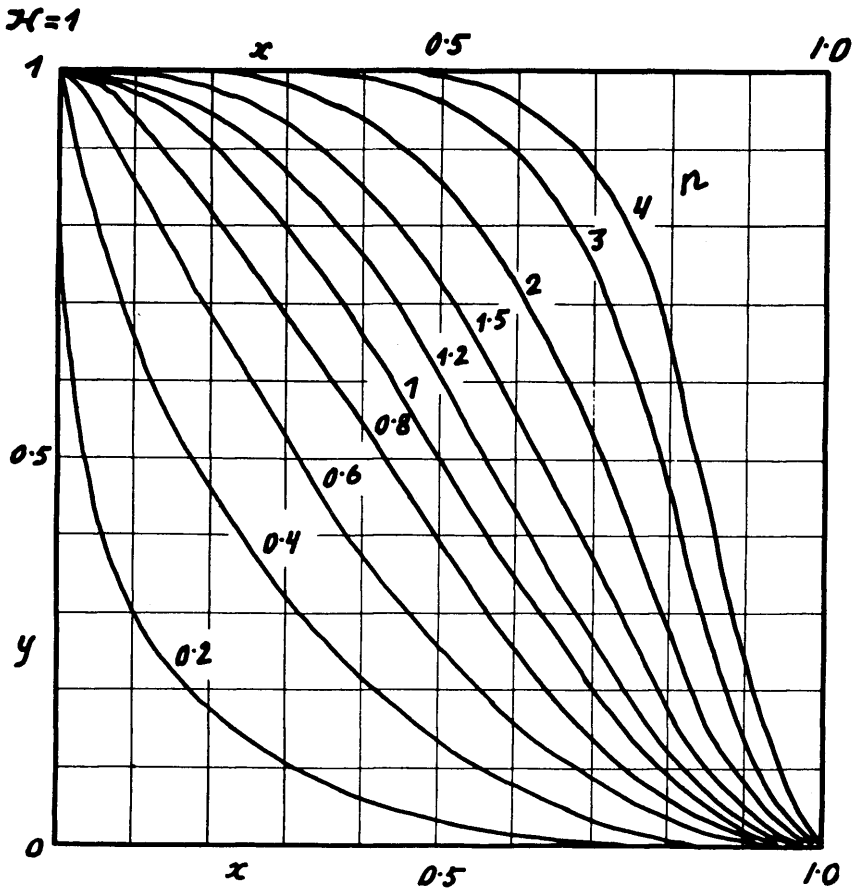


Fig. 22.

Teil schon abgeschnitten ist, wobei die Schnittgerade die Häufigkeitskurve in den Punkten M und N schneidet (Fig. 17). Wenn wir die Strecke zwischen den Projektionen der genannten Punkte von M^1 bis N^1 als eine Klasse betrachten, ist H^1 die der Häufigkeit dieser Klasse entsprechende Höhe, wobei von der relativen Häufigkeit der Klasse schon vorher 5% abgezogen worden sind. Wenn wir eine andere Klasse betrachten, die kleiner oder grösser als M^1N^1 ist, von ihrer relativen Häufigkeit 5% abziehen und dann die mittlere Höhe H^* finden, kann man beweisen, dass immer $H^* < H^1$ ist. Das Gesagte bedeutet, dass wenn wir den Gang der Grösse H^* als Funktion der Klassenbreite darstellen, $H^1 = H^*_{max}$ ist. Dieses kann folgendermassen gezeigt werden.

Wenn die Klassenbreite kleiner ist als MN, z. B. irgendein PQ (siehe Figur), dann ist die oberhalb der Geraden MN liegende Fläche P^1PAQQ^1 kleiner als 5% (wenn die ganze unter der Häufigkeitskurve liegende Fläche 100% beträgt) und wir müssen die Gerade MN nach unten verschieben, bis der 5%-ge Teil abgeschnitten ist. Die neue Höhe H^* ist dann kleiner als H^1.

Wenn die Klassenbreite grösser ist als MN, z. B. irgendein RS (siehe Figur), dann ist H^* die mittlere Höhe des unter der Geraden MN verbliebenen Teils, was wieder kleiner als H^1 ist, denn die Strecke MN ist kleiner als RS.

Bei empirischen Verteilungen ist eine stetige Veränderung der Klassenbreite nicht möglich, denn die Häufigkeitskurve ist nicht stetig. Wir können aber dann so verfahren, dass wir zuerst die grösste Klasse nehmen, dann zu ihr die grössere unter den danebenstehenden Klassen hinzuaddieren usw. Das Gesagte wird durch folgendes Beispiel erläutert, wobei die früher gegebenen Häufigkeitszahlen verwendet werden.

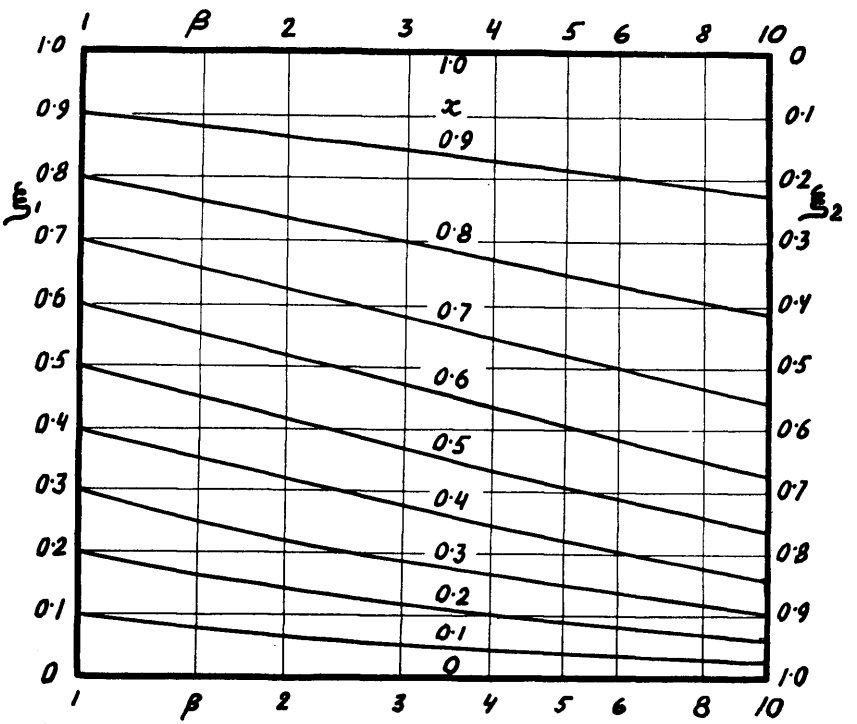


Fig. 23.

Beispiel:

i	1	2	3	4	5	6	7	8
$p_i(\%)$	7·8	7·3	6·9	6·7	6·5	6·1	5·7	5·4
Σp_i	7·8	15·1	22·0	28·7	35·2	41·3	47·0	52·4
$S_i = \Sigma p_i - 5{\cdot}0$	2·8	10·1	17·0	23·7	30·2	36·3	42·0	47·4
$H^* = \frac{S_i}{i}$	2·80	5·05	5·67	5·92	6·04	6.05	6·00	5·92

Hieraus ersehen wir, dass $H_{max} = 6{\cdot}05$ ist, was mit dem nach der vorigen Berechnungsmethode erhaltenen Resultate übereinstimmt. Da die letztgegebene Berechnungsmethode beinahe einfacher als die vorher gegebene ist, empfiehlt es sich, von ihr Gebrauch zu machen.

Die folgende Tabelle (Tab. 4) gibt die Werte für die drei Indikatoren bei einigen Wertkomplexen von n und β wieder.

Tab. 4.

$P(\%)$-Werte.

β \ n	0·1	0·3	0·5	0·7	1·0	1·5	2·0	4·0	6·0
1·0	21·0	24·1	26·9	28·8	30·6	32·3	33·3	34·6	34·8
1·2	22·5	26·9	29·3	31·2	32·8	33·8	34·4	34·8	34·9
1·5	27·5	32·2	34·2	35·0	35·6	35·7	35·6	35·2	35·1
2·0	40·2	39·4	39·0	38·7	38·3	37·7	37·3	36·3	35·5
3·0	53·0	49·2	47·0	45·3	43·2	41·1	39·6	36·6	35.6
5·0	65·0	60·0	55·9	52·2	48·5	44·6	42·3	38·3	35·7
10·0	72·3	66·7	62·9	59·1	54·3	49·1	46·1	40·0	35·8

$M(\%)$-Werte.

β \ n	0·1	0·3	0·5	0·7	1·0	1·5	2·0	4·0	6·0
1·0	50·0	50·0	50·0	50·0	50·0	50·0	50·0	50·0	50·0
1·2	46·8	47·0	47·3	47·6	47·9	48·4	48·7	49·5	49·8
1·5	42·7	43·0	43·6	44·2	45·1	46·2	47·0	48·9	49·4
2·0	38·1	39·2	40·3	41·2	42·6	44·1	45·6	48·2	49·0
3·0	32·0	33·8	35·3	36·6	38·6	41·0	43·0	46·5	48·0
5·0	26·5	27·6	29·6	31·5	34·1	37.4	40·0	44·9	47·0
10·0	20·0	21·0	23·9	26·0	29·8	33·6	36·6	42·7	45·2

μ-Werte.

β \ n	0·1	0·3	0·5	0·7	1·0	1·5	2·0	4·0	6·0
1·0	2·35	1·59	1·30	1·13	1·01	0·94	0·90	0·86	0·86
1·2	2·38	1·59	1·30	1·13	1·02	0·94	0·90	0·86	0·86
1·5	2·41	1·61	1·30	1·15	1·03	0·95	0·91	0·86	0·86
2.0	2·46	1·65	1·32	1·17	1·05	0·97	0·92	0·87	0·86
3·0	2·50	1·73	1·38	1·23	1·10	1·00	0·95	0·88	0·86
5·0	2·56	1·91	1·51	1·34	1·19	1·07	1·00	0·91	0·88
10·0	2·53	2·28	1·83	1·59	1·37	1·19	1·09	0·99	0·92

Aus diesen Tabellen ersieht man, dass jeder Indikator von den beiden Grössen β und n abhängig ist. Darum können wir auch nicht auf Grund der gegebenen Grösse nur des einen Indikators

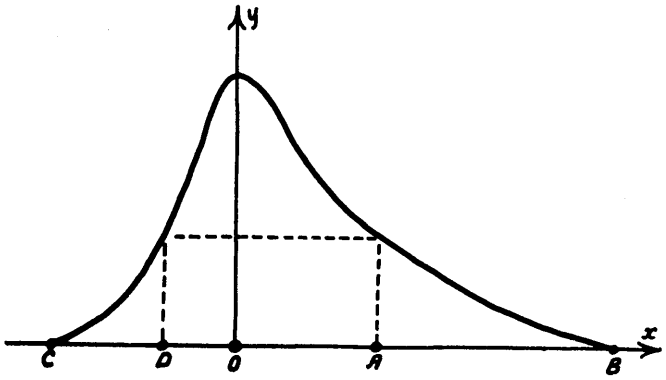


Fig. 24.

β oder n bestimmen; wir brauchen dazu zwei Indikatoren. Als zu diesem Zweck passend können das $P(\%)$- und μ- oder das $M(\%)$- und μ-Paar genommen werden (diese zwei Paare wurden unter 28 Paaren gewählt, denn ausser $P(\%)$, $M(\%)$ und μ war noch eine Reihe anderer Indikatoren vorhanden, die jedoch mehr oder weniger ungünstig waren). Diesem entsprechend sind Nomogramme konstruiert worden, welche für die erhaltenen Werte der Indikatoren die gesuchten β und n zu finden sofort ermöglichen (Fig. 18 und 19). Es ist empfehlenswert, die beiden Nomogramme zu benutzen und aus den erhaltenen β- und n-Werten das Mittel zu nehmen, entweder das einfache oder das gewägte. Wenn die beiden Nomogramme verschiedene β- oder n-Werte geben, so bedeutet dieses, dass die Verteilung sehr schlecht durch irgendwelche

Gleichung darstellbar ist, und es können zum Bestimmen der Gleichung auch andere dazwischen liegende β- und n-Werte genommen werden, denn die Güte der Approximation kann ja verschiedenartig definiert werden.

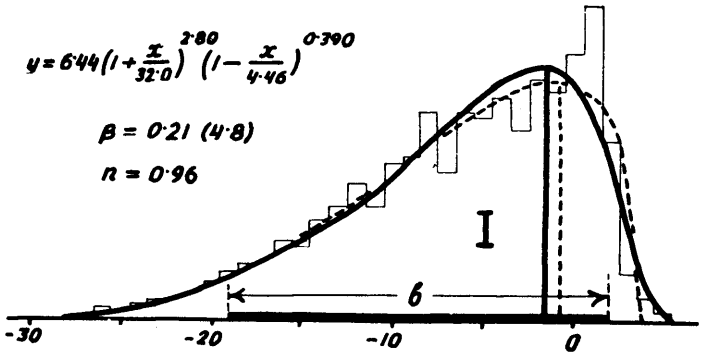


Fig. 25.

XI. Das Zeichnen der Kurve.

Zum Zeichnen der Kurve bedürfen wir der beiden Endpunkte, der Mode und der Länge der maximalen Ordinate H. Zum Bestimmen der Endpunkte ist in Fig. 20 das entsprechende Nomogramm konstruiert worden, welches nach gegebenen β- und n-Werten den Abstand der gesuchten Punkte von den Endpunkten der Basis, s_1 und s_2, angibt, wobei als Einheit der Zahlen s_1 und s_2 die Länge der Basis gilt.

Das folgende Nomogramm (Fig. 21) gibt analog die Länge der maximalen Ordinate H an, wobei diese als Verhältnis zu der in bezug auf die Basis reduzierten mittleren Höhe h gegeben ist. Ebenfalls gibt dasselbe Nomogramm die Lagef der Mode Mo gerechnet von dem der Mode näher liegenden Endpunkt der Basis (die Einheit der Zahlen Mo ist die Länge der Basis).

Zuletzt sind zur Bestimmung der anderen Kurvenpunkte Nomogramme angefertigt worden, die es ermöglichen, für jede Abszisse die entsprechende Ordinate zu finden. Da wir es hier mit vier Veränderlichen zu tun haben (β, n, x, y), so ist das Problem der Einfachheit halber in zwei zerlegt worden:

1. das Bestimmen bei gegebenem n und bei symmetrischer Kurve ($\beta = 1$) für jedes x des ihm entsprechenden y und

2. das Bestimmen der Lage desselben x (d. h. mit derselben Ordinate y) bei irgendwelchem β.

Zur Lösung des ersten Teiles ist das Nomogramm Fig. 22 gezeichnet worden, das nichts anderes vorstellt, als eine Schar von symmetrischen Kurven bei verschiedenen n-Werten.

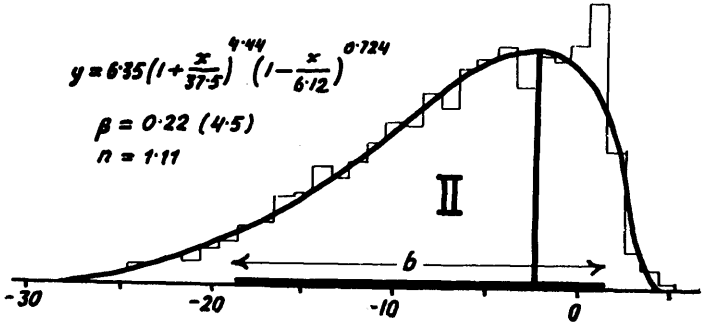


Fig. 26.

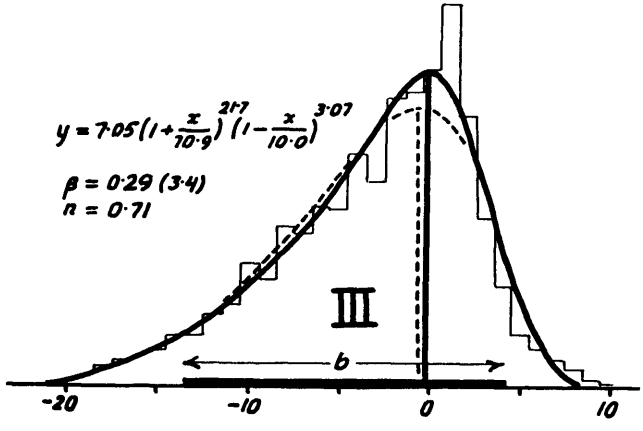


Fig. 27.

Zur Lösung des zweiten Teiles ist ebenso ein Nomogramm konstruiert worden (Fig. 23), mit dessen Hilfe man nach der Abszisse x der symmetrischen Kurve die entsprechende Ab-

szisse der schiefen Kurve bestimmen kann. Um diese Abszisse vom Massstab der Abszissenachse unabhängig zu machen, ist die von der Mode aus genommene Abszissenlänge durch die Länge des entsprechenden Zweiges der Kurve dividiert und mit ξ bezeichnet. Da ξ für die beiden Zweige der Kurve nicht gleich gross ist, sind in dem Nomogramm die Werte von ξ für beide Fälle dargestellt, wobei ξ_1 den Wert für die längere Seite und ξ_2 denjenigen für die kürzere Seite bedeutet. Damit ist: $\xi_1 = \frac{OA}{OB}$ und $\xi_2 = \frac{OD}{OC}$ (Fig. 24).

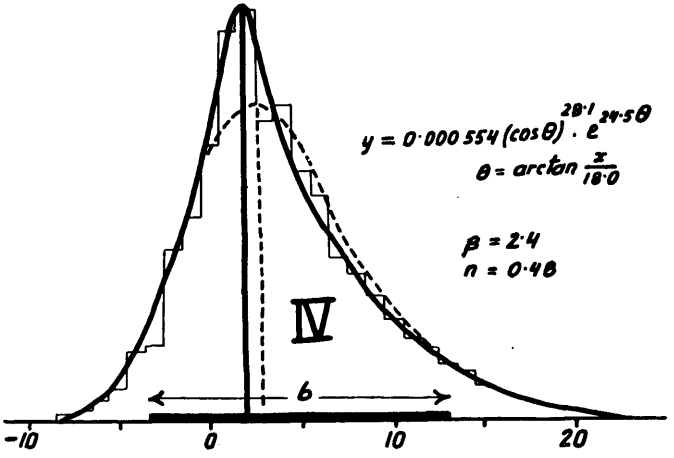


Fig. 28.

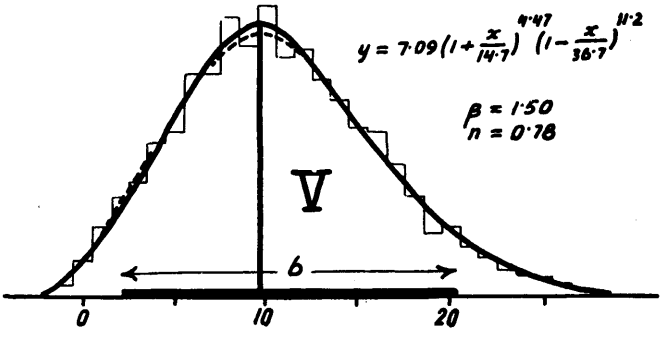


Fig. 29.

XII. Ein Vergleich mit dem Pearson'schen System.

Um die Güte unseres Systems zu beurteilen, ist dieses gleichzeitig mit dem Pearson'schen System bei Untersuchung der Häufigkeitsverteilung der Temperatur in Tartu für alle Monate der Zeitspanne 1866—1935 (incl.) angewandt worden (die Häufigkeitsverteilungen und die ihnen entsprechenden Gleichungen sind von Prof. K. Kirde zusammengestellt worden). In den Figuren

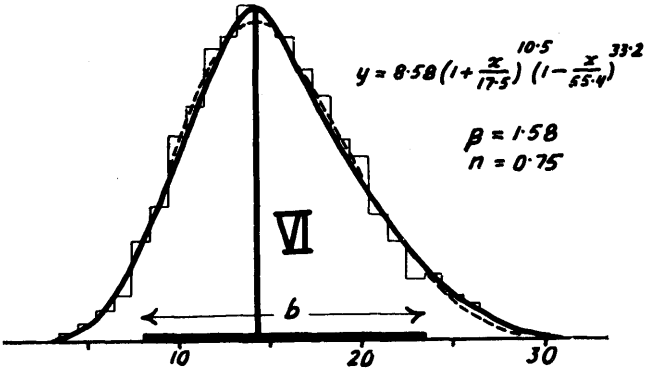


Fig. 30.

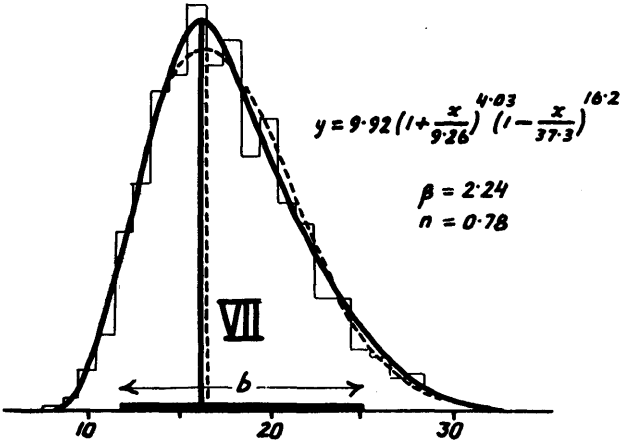


Fig. 31.

(Fig. 25—36) bedeutet die ausgezogene Linie die nach dem von uns vorgeschlagenen System berechnete Häufigkeitskurve und die punktierte die Pearson'sche Kurve. Neben die Kurven sind die

Pearson'sche Gleichung und die Charakteristiken des neuen Systems geschrieben worden. Bei negativer Schiefe ($\beta < 1$) ist in Klammern der reziproke Wert von β gegeben worden, welcher die absolute Grösse der Schiefe wiedergibt.

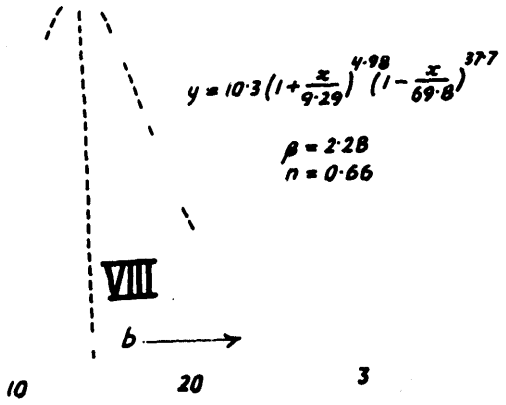


Fig. 32.

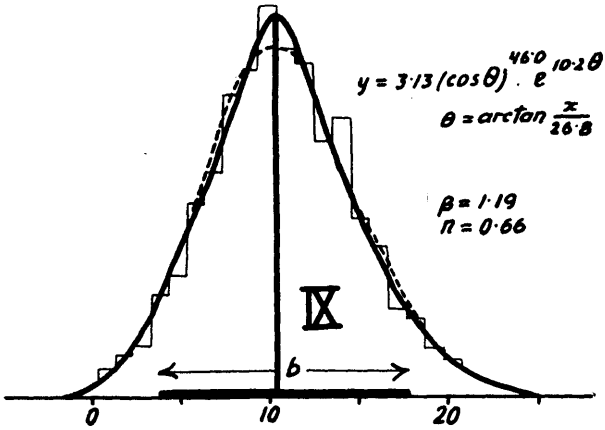


Fig. 33.

Aus den Figuren ersieht man, dass das neue System nicht schlechter als das Pearson'sche ist (mit seinen vielen Kurventypen müsste das Pearson'sche System eine bessere Möglichkeit der Näherung bieten), ja für einige Monate (März, April, November) sogar besser ist. Aber der zu den Berechnungen nötige Zeitauf-

wand und die Arbeitsmenge ist beim neuen System viel geringer. Die Versuche haben gezeigt, dass bei gegebenen Temperaturangaben, wo wir es mit 40 Klassen zu tun haben, bei Anwendung des Pearson'schen Systems (mit Aufzeichnen der Kurve) ein Zeitaufwand von 4—6 Stunden in Betracht kam, während bei Anwendung des neu gegebenen Systems dazu nur 10—15 Minuten erforderlich waren. Dabei ist der Berechnungsgang viel übersichtlicher

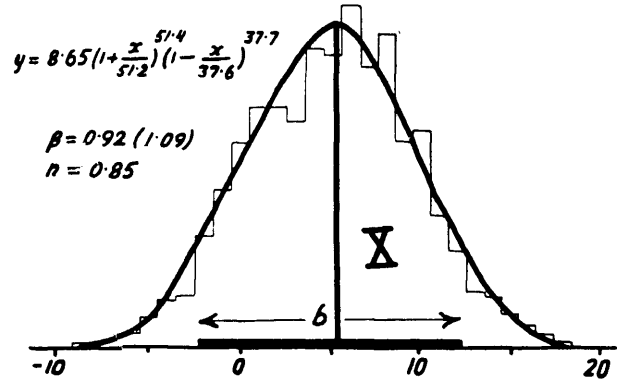


Fig. 34.

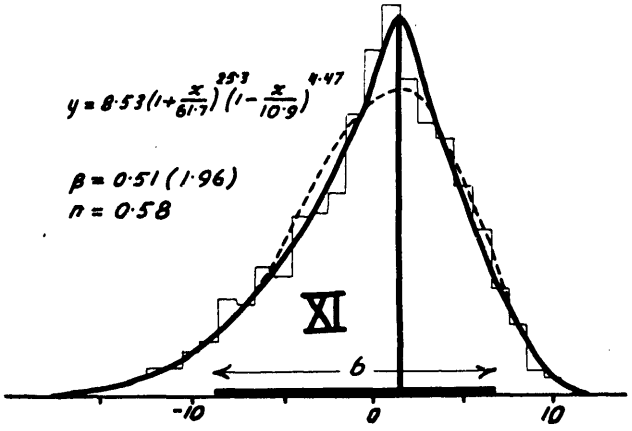


Fig. 35.

und jeder Schritt leicht kontrollierbar. Bei einer genügenden Übung können die Grössen β und n nach Augenmass mit einer Genauigkeit von ca 10% bestimmt werden, so dass das Entstehen grosser Fehler fast unmöglich ist.

Als Beispiel für die Anwendbarkeit des Systems ist die Häufigkeitsverteilung der Flächengrössen der Seen in Estland genommen worden (Fig. 37). Obwohl die Kurve nur bis 23 ha reicht und es auch grössere Seen gibt, sind letztere über grosse Intervalle zerstreut und besitzen eine so kleine Häufigkeit, dass diese im gegebenen linearen Koordinatensystem nicht dargestellt werden kann. Darum eignet sich die Kurve für die Darstellung des reliefartigen Teils der Verteilung, ungeachtet dessen, dass die Momente überhaupt nicht übereinstimmen.

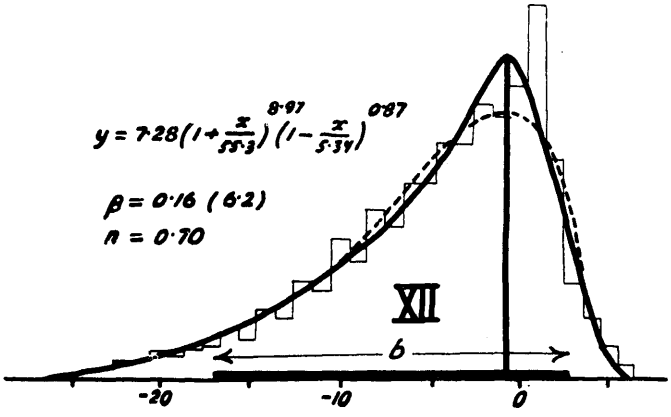


Fig. 36.

Daraus sieht man auch, wie leicht die Eigenschaften der Verteilung aus der Gleichung herauszulesen sind. Die Zahl 2·2 gibt den reziproken Wert der Mode; damit ist $Mo = 0{\cdot}45$. Die Gleichung $2{\cdot}2x = 51$ gibt den rechtsliegenden Endpunkt der Kurve, damit ist $B = 23{\cdot}2$, und die Gleichung $2{\cdot}2.51x = 1$ den linksliegenden Punkt, damit ist $c = 0{\cdot}009$ (praktisch $= 0$). Die Zahl 51 besagt, dass wir es mit einer sehr grossen Schiefe zu tun haben, denn der rechte Zweig der Kurve ist 51 mal länger als der linke. Der Exponent 0·50 besagt, dass wir es mit einer übernormalen Häufung (bei normaler Häufung ist $n \approx 0{\cdot}75$) in der Umgebung der Mode zu tun haben. Hierbei sei gesagt, dass wegen der sehr anormalen Verteilung das Pearson'sche System überhaupt keine passende Näherungskurve ergab.

Über die Anwendbarkeit des Charlier'schen Systems muss man sagen, dass in der Umgebung der normalen Verteilung

die Näherungsfähigkeit des Systems ganz gut ist, dass aber bei genügend schiefen und mit einer anormalen Häufung behafteten Verteilungen das System keine genügende Biegsamkeit

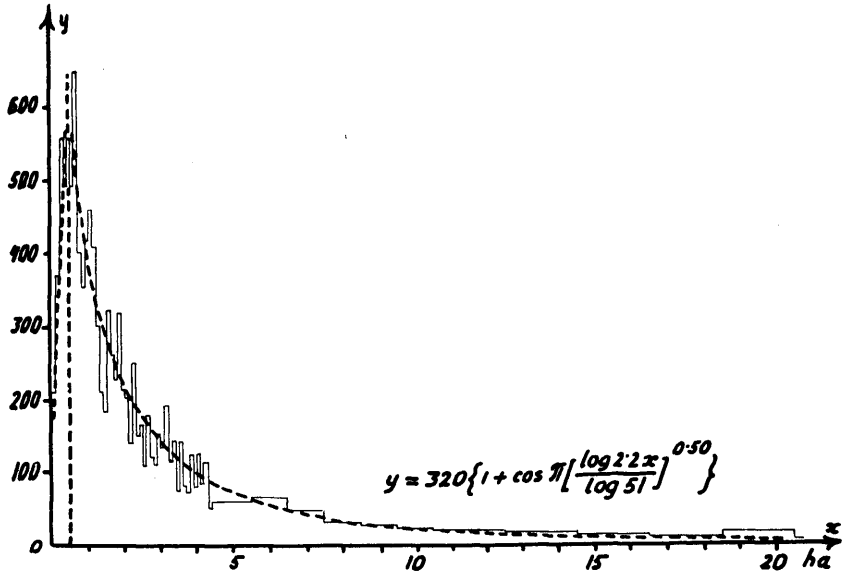


Fig. 37.

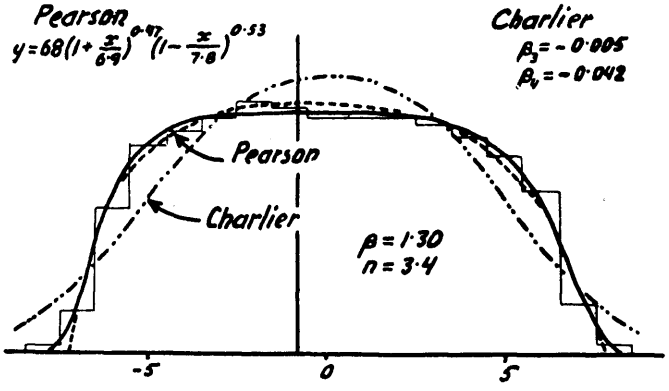


Fig. 38.

besitzt. In Fig. 38 ist eine Verteilung gegeben, welche durch das von uns gegebene System (ununterbrochene Linie) und das Pearson'sche System (punktierte Linie) gut approximiert werden

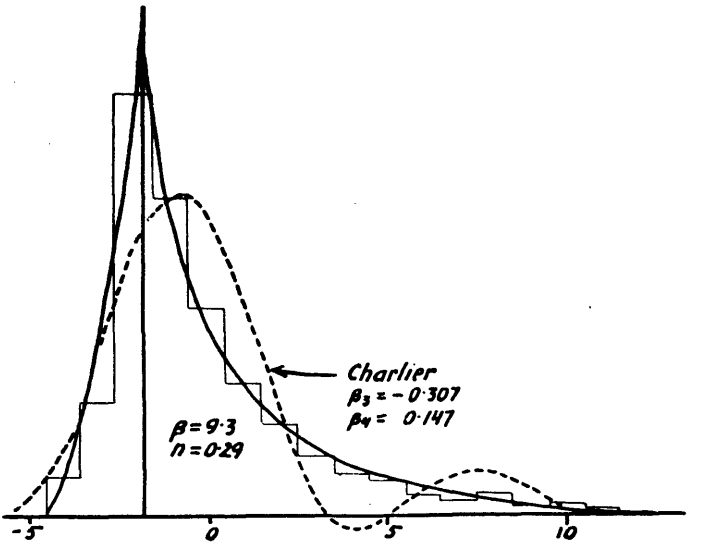


Fig. 39.

kann; die Charlier'sche Kurve (doppeltpunktierte Linie) gibt aber den Charakter der Verteilung nicht mit genügender Güte wieder.

In Fig. 39 sind für eine schiefe Verteilung die Näherungskurven des in der vorliegenden Arbeit entwickelten und die des Charlier'schen Systems dargestellt. Aus ihnen ersieht man, dass in manchen Fällen das Charlier'sche System einen wichtigen Nachteil besitzt — hier kommen nämlich negative Häufigkeiten vor. Dieser schwache Punkt im Charlier'schen System kommt häufig zum Vorschein, wenn die Verteilung genügend schief und eine anormale Häufung in der Umgebung der Mode vorhanden ist. Wenn die Schiefe und die Häufung nicht von Bedeutung sind, kommt häufig eine Nebenmode zum Vorschein.