ÜBER DAS PROBLEM DER MESSUNG DER STÖRUNG BEI STATISTISCHEN REIHEN MIT ANWENDUNG AUF DIE KLIMATOLOGIE

VON

AARNE KÄRSNA

TARTU 1938

ÜBER DAS PROBLEM DER MESSUNG DER STÖRUNG BEI STATISTISCHEN REIHEN MIT ANWENDUNG AUF DIE KLIMATOLOGIE

VON

AARNE KÄRSNA

TARTU 1938

Acta et Commentationes Universitatis Tartuensis (Dorpatensis) A XXXIV. 2

K. Mattiesens Buchdruckerei Ant.-Ges., Tartu 1938.

§ 1. Einleitung.

1. In einer meiner früheren Arbeiten[1]) habe ich auf Grund klimatologischer Beobachtungen empirisch gezeigt, dass der Charliersche Störungskoeffizient zur Messung der Störung bei statistischen Reihen nicht immer anwendbar ist, denn er führt manchmal zu solchen Resultaten, die dem intuitiven Denken widersprechen und sich nicht physikalisch begründen lassen.

Um festzustellen, ob eine solche Annahme auf (durch das Beobachtungsmaterial bedingtem) Zufall beruhe oder ob dem Charlierschen Störungskoeffizienten ein prinzipieller Fehler innewohne, ist in der folgenden Arbeit das Problem der Messung der Störung bei statistischen Reihen theoretisch untersucht worden, wobei die Berechtigung des Zweifels am Charlierschen Störungskoeffizienten seine Bestätigung gefunden hat.

2. Wenn die Wahrscheinlichkeit des Erscheinens eines Merkmals sich von einem Kollektiv zum anderen ändert (jedes Kollektiv bestehe aus s Elementen), so bilden die entsprechenden Häufigkeitszahlen $m_1, m_2, \ldots m_n$ eine Lexissche Reihe. Bezeichnen wir die veränderlichen Wahrscheinlichkeiten mit $p_1, p_2, \ldots p_n$ und deren arithmetisches Mittel mit

$$p_0 = \frac{p_1 + p_2 + \cdot \cdot p_n}{n},$$

so beträgt bei der Lexisschen Reihe die mittlere Häufigkeitszahl

$$m_0 = s p_0$$

---

[1]) „Ühemodaalsete sageduskõverate süsteemist Lexise ridade puhul ühes rakendusnäidetega klimatoloogias", Tartu 1936.

Diese Arbeit ist nicht gedruckt. Die Handschrift ist in der Bibliothek der Universität Tartu deponiert.

und das Streuungsmass

$$\sigma_L = \sqrt{sp_o(1-p_o) + \frac{s^2-s}{n} \sum_{i=1}^{n} (p_i - p_o)^2}. \tag{1}$$

Die Grösse $sp_o(1-p_o)$ ist das Quadrat des Streuungsmasses bei der Bernoullischen Reihe, wenn das Merkmal in jedem Kollektiv mit der Wahrscheinlichkeit p_o erscheint. Damit ist

$$sp_o\,(1-p_o) = \sigma_B^2. \tag{2}$$

Bei der Wahrscheinlichkeit p_o ist die mittlere Häufigkeitszahl sp_o und bei der Wahrscheinlichkeit p_i ist sie sp_i. Verändert sich die Wahrscheinlichkeit von p_o auf p_i, so beträgt die hiervon abhängige Veränderung der mittleren Häufigkeitszahl

$$s\,(p_i - p_o).$$

Das quadratische Mittel bei n Kollektiven ist

$$\sqrt{\frac{s^2}{n} \sum_{i=1}^{n} (p_i - p_o)^2}, \tag{3}$$

was wir mit δ bezeichnen und die **Störungsstreuung** nennen.

Setzen wir die entsprechenden Ausdrücke von (2) und 3) in die Formel (1) ein, so erhalten wir

$$\sigma_L^2 = \sigma_B^2 + \left(1 - \frac{1}{s}\right)\delta^2. \tag{4}$$

Die Grösse $\left(1 - \frac{1}{s}\right)$ können wir gleich eins annehmen, denn der relative Fehler $\frac{1}{s}$ beeinflusst σ_L weniger als der von der Zahl der Kollektive n abhängige mittlere Fehler von σ_L[1]).

---

[1]) Bei n Kollektiven ist der relative Fehler von σ_L gleich $\frac{1}{\sqrt{2n}}$, oder der relative Fehler von σ_L^2 ist

$$\sim \frac{2}{\sqrt{2\,n}}.$$

Wenn

$$\frac{2}{\sqrt{2\,n}} < \frac{1}{s}, \text{ dann ist}$$

$$n > 2\,s^2.$$

Die letztgegebene Bedingung wird praktisch niemals erfüllt. Sogar bei einem sehr kleinen Kollektiv, z. B. bei $s = 50$, müssten wir den Versuch mindestens 5000 Mal wiederholen, bevor wir mit der weggelassenen Grösse $\frac{1}{s}$ zu rechnen hätten. Bei $s = 1\,000$ muss $n > 2\,000\,000$ sein.

Das Gesagte in Betracht ziehend, können wir auf Grund der Formel (4) schreiben

$$\sigma_L^2 = \sigma_B^2 + \delta^2. \tag{5}$$

Aus letzterer Formel können wir die Störungsstreuung

$$\delta = \sqrt{\sigma_L^2 - \sigma_B^2} \tag{6}$$

bestimmen, in der Form, in welcher sie oft gebraucht wird.

Messen wir die Störungsstreuung in Beziehung zum Kollektivumfang, so erhalten wir

$$\frac{\delta}{s} = \frac{\sqrt{\sigma_L^2 - \sigma_B^2}}{s} = \xi, \tag{7}$$

welchen Wert wir die **mittlere Störung** der Lexisschen Reihe nennen und mit ξ bezeichnen.

3. Wenn bei mancher statistischen Reihe eine von Null abweichende mittlere Störung bemerkbar ist, so weist das auf ein Auftreten äusserer Einflüsse hin. Schätzt man die Grösse der äusseren Einflüsse ab, so scheint es natürlich, sie proportional der mittleren Störung anzunehmen. Wird z. B. die regelmässige Häufigkeit der Geburten von Jahr zu Jahr durch Kriege, Krisen, politische Veränderungen usw. gestört, so wird, je grösser diese Veränderungen sind, desto grösser auch die mittlere Störung der entsprechenden Reihe sein. Die Grösse der mittleren Störung kann man als einen Indikator bezeichnen, der die Gesamtsumme der äusseren Einflüsse darstellt. Die genannten äusseren Einflüsse sind direkt nicht messbar, denn der Einfluss der Kriege ist z. B. nicht mit dem Einfluss der Propaganda vergleichbar usw. Die Bedeutung dieser Faktoren bei der Beeinflussung regelmässiger Erscheinungen kann nur auf Grund der Grösse der mittleren Störung geschätzt werden.

Das Problem ist aber in seiner Realität komplizierter. Die statistischen Beobachtungen zeigen, dass die mittlere Störung und die relative Häufigkeit der Erscheinung (die Wahrscheinlichkeit) miteinander korrelativ verbunden sind, wobei $r > 0$ ist (r ist der Korrelationskoeffizient), d. h. grössere relative Häufigkeiten weisen meistens eine grössere mittlere Störung auf als kleinere relative Häufigkeiten. Nun könnte man ja denken, dass man bei Erscheinungen mit grossen relativen Häufigkeiten nur zufällig grosse Störungen antreffe; aber erstens wäre solch ein oft auftretendes Zusammenfallen, wenn es nur auf Zufall beruhen sollte, nicht wahrscheinlich, und zweitens kann man theoretisch solche Beispiele anführen, aus denen klar hervorgeht, dass kleine relative Häufigkeiten unter gleichen Bedingungen kleinere Störungen aufweisen als entsprechende grosse.

4. Zur normalen Messung der Störung wird von Charlier der sogenannte Störungskoeffizient in folgender Form benutzt:

$$C = \frac{\delta}{m_0} = \frac{\sqrt{\sigma_L^2 - \sigma_B^2}}{m_0}, \tag{8}$$

wo die Buchstaben die obengenannten Bedeutungen haben.

Bei konstantem Kollektivumfang (s = const.) wächst m_0 proportional der relativen Häufigkeit, und daraus folgt, dass der Charliersche Koeffizient den grösseren relativen Häufigkeiten auch die grössere Störungsstreuung zuschreibt, wobei letztere proportional der relativen Häufigkeit wächst.

Ein solcher Koeffizient scheint anfänglich zu genügen, denn er wird als Variationsfaktor[1]) beim Bestimmen der Stabilität der Grössen sehr oft benutzt (z. B. in der Anthropologie). Doch ist ohne weiteres inhaltlich nicht klar, warum die Störungsstreuung bei konstanten Störungsursachen proportional der relativen Häufigkeit wächst, denn das Wachsen könnte doch auch nach anderen Gesetzen vor sich gehen. Charlier führt aber eine Reihe von Beispielen an, die das Gesagte empirisch ganz glaubhaft machen. So bringt er ein Beispiel über die Geburten in Schweden. Die Störungsstreuung bei der Zahl der Gesamtgeburten ist viel grösser als bei Zwillingsgeburten. Es ist begreiflich, dass alle Einflüsse, die auf die Gesamtgeburten wirken, auch auf die Zwillingsgeburten wirken müssen, bei letzteren können aber einige spezielle Einflüsse noch hinzukommen. Benutzt man den Variationsfaktor zur Messung der Störung, so ergibt sich, dass die Variationsfaktoren in beiden Fällen gleich gross sind. Die Gesamtgeburten bringen also durch ihre grössere relative Häufigkeit von selbst die grössere Störungsstreuung mit sich.

[1]) Unter dem Variationsfaktor versteht man in der Statistik im allgemeinen das Verhältnis der Streuung zur mittleren Grösse.

In der vorliegenden Arbeit ist die genannte Frage näher erforscht worden, wobei versucht wurde, für die Messung der Störung einige annehmbare Grundlagen zu finden. Die Ergebnisse führen uns nun recht weit vom Charlierschen Faktor ab, dieser wird aber dennoch nicht unbrauchbar, sondern bildet nur einen Spezialfall.

§ 2. Die Abhängigkeit der Störungsstreuung von der relativen Häufigkeit.

1. Betrachten wir in einem Kollektiv (die Zahl der Elemente sei s) zwei Merkmale, die voneinander unabhängig sind, d. h. bei denen es kein Element gibt, welches beide Merkmale besitzt (z. B. eine Temperatur von 5 bis 6 und eine von 6 bis 7 Grad). Nehmen wir nun als neues Merkmal die Summe der vorigen Merkmale (im genannten Beispiel — die Temperaturen von 5 bis 7 Grad). Die relative Häufigkeit des neuen Merkmals ist gleich der Summe der relativen Häufigkeiten der einzelnen Merkmale. Was aber geschieht mit der Streuung bei der neuen Häufigkeit?

Nehmen wir als den Kollektivumfang s und als die mittleren Häufigkeiten beider Merkmale m_1 und m_2 an und bezeichnen wir die relativen Häufigkeiten

$$\frac{m_1}{s} = p_1 \text{ und } \frac{m_2}{s} = p_2.$$

Die Störungsstreuungen für beide Klassen seien δ_1 und δ_2, die aus den allgemeinen Streuungsmassen σ_1 bzw. σ_2 in folgender Weise abgeleitet werden können:

$$\delta_1 = \sqrt{\sigma_1{}^2 - sp_1(1-p_1)} \quad \text{und} \tag{9}$$

$$\delta_2 = \sqrt{\sigma_2{}^2 - sp_2(1-p_2)}.$$

Betrachten wir nun, wie sich die ganze Erscheinung bei Gültigkeit des Charlierschen Prinzips entwickelt.

Der Charliersche Faktor ist

$$C_1 = \frac{\delta_1}{m_1} \text{ und } C_2 = \frac{\delta_2}{m_2}.$$

Zur Vereinfachung der Untersuchung wählen wir die beiden Häufigkeiten und Störungsstreuungen gleich gross, d. h.

$$m_1 = m_2 \quad \text{und}$$

$$\delta_1 = \delta_2.$$

Dadurch ist auch

$$C_1 = C_2.$$

Wenn zwei Erscheinungen in gleichem Masse gestört werden, ist es selbstverständlich, dass auch die Störung der neuen, durch Zusammenfassung der Einzelerscheinungen entstandenen Erscheinung ebenso gross sein wird (diese Folgerung legt auch Charlier seiner Ausführung zugrunde, wenn er die Störungsstreuung von Zwillingsgeburten und Nichtzwillingsgeburten mit derjenigen der Gesamtgeburten vergleicht).

Die Häufigkeit des Auftretens der neuen Erscheinung ist

$$m_1 + m_2 = 2\, m_1.$$

Die ihr zugehörige Störungsstreuung bezeichnen wir mit δ. Nach dem Vorhergesagten ist

$$\frac{\delta}{2m_1} = C_1 = \frac{\delta_1}{m_1}, \tag{10}$$

woraus

$$\delta = 2\delta_1. \tag{11}$$

Das bedeutet: damit der Charliersche Faktor gelte, müssen beim Zusammenfassen der Häufigkeiten auch die Störungsstreuungen sich addieren. Ist das immer so? Die Antwort braucht nicht weit gesucht zu werden, denn man kann ohne weiteres sagen, dass die obengenannte Bedingung nur im speziellen Falle gilt. Dieses kann auf folgende Weise erklärt werden.

Das Auftreten der Störungen in den beiden Erscheinungen ist immer miteinander korrelativ verbunden, was daraus folgt, dass die Lexissche Reihe der Häufigkeitszahlen, deren arithmetisches Mittel m_1 ist, mit der Reihe, deren mittlere Häufigkeitszahl m_2 ist, korrelativ verbunden ist. Wenn eine solche Beziehung fehlt, so sagt man: der Korrelationskoeffizient sei gleich Null. Darum können wir auf Grund der Korrelationstheorie schreiben

(12) $$\delta^2 = \delta_1{}^2 + \delta_2{}^2 + 2r\,\delta_1\,\delta_2,$$

wo r den Korrelationskoeffizienten zwischen den Störungen bedeutet.

In unserem Falle ist

$$\delta_1 = \delta_2 \text{ und}$$

(13) $$\delta^2 = 2\delta_1{}^2 + 2r\,\delta_1{}^2 = \\ = 2\delta_1{}^2\,(1 + r).$$

Damit der Charliersche Faktor gelte, muss nach (11)

(14) $$\delta^2 = 4\delta_1{}^2.$$

Wir setzen den Ausdruck für δ^2 in (13) ein und erhalten

$$4\delta_1{}^2 = 2\delta_1{}^2\,(1 + r),$$

wo

(15) $$r = 1.$$

Dieses bedeutet, dass der Charliersche Faktor nur dann gilt, wenn die Störungen in absoluter und positiver Korrelation stehen.

2. Entwickeln wir denselben Gedankengang weiter, so bekommen wir bei mehreren Erscheinungen dasselbe Resultat. Es sei

$$p_1 = p_2 = \ldots\ldots p_k$$

und damit

$$m_1 = m_2 = \ldots\ldots m_k.$$

Wenn die Merkmale voneinander unabhängig sind, bekommen wir beim Addieren der Merkmale in obengenannter Weise

$$m = m_1 + m_2 + \ldots\ldots m_k = \\ = km_1.$$

Bezeichnen wir die Störungsstreuungen entsprechend mit

$$\delta_1 = \delta_2 = \ldots\ldots \delta_k,$$

dann muss, falls die Störungen absolut positiv miteinander in Korrelation stehen,

$$(16) \qquad \begin{aligned} \delta^2 = \delta_1^2 + \delta_2^2 + \ldots . . \delta_k^2 + \\ + 2r_{12}\delta_1\delta_2 + 2r_{13}\delta_1\delta_3 + \ldots 2r_{1k}\delta_1\delta_k + \\ + 2r_{23}\delta_2\delta_3 + 2r_{24}\delta_2\delta_4 + \ldots 2r_{2k}\delta_2\delta_k + \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ + 2r_{k-1,\,k}\,\delta_{k-1}\,\delta_k = \\ = \delta_1^2[k + 2(k-1) + 2(k-2) + \ldots . 2] = \\ = k^2\delta_1^2 \end{aligned}$$

sein, denn

$$r_{ij} = 1.$$

Oder

$$\delta = k\delta_1.$$

Damit ist der Charliersche Faktor

$$(17) \qquad C = \frac{\delta}{m} = \frac{k\delta_1}{km_1} = C_1,$$

d. h. er bleibt konstant.

Ebenso bleibt bei $r = 1$ der Charliersche Faktor auch dann konstant, wenn die relativen Häufigkeiten verschiedene Grössen haben.

Es sei die Häufigkeit der einen Erscheinung m_1 und die ihr zugehörige Störungsstreuung δ_1. Damit ist

$$C_1 = \frac{\delta_1}{m_1}.$$

Bei einer anderen Erscheinung sei die Häufigkeit m_2, wobei wir

$$m_2 = nm_1$$

annehmen, und die Störungsstreuung sei δ_2.

Somit ist

$$C_2 = \frac{\delta_2}{nm_1}.$$

Damit

$$\frac{\delta_1}{m_1} = \frac{\delta_2}{nm_1}$$

sei, muss

$$\delta_2 = n\delta_1$$

sein. Fassen wir die beiden Erscheinungen zusammen, so bekommen wir

$$m = m_1 + m_2 =$$
$$= m_1 (1 + n)$$

und

$$\delta^2 = \delta_1^2 + \delta_2^2 + 2r\,\delta_1\,\delta_2 = \tag{18}$$
$$= \delta_1^2 + n^2\,\delta_1^2 + 2n\,\delta_1^2 =$$
$$= \delta_1^2 (1 + n)^2.$$

Hier ist

$$\delta = \delta_1 (1 + n).$$

Daraus folgt

$$C = \frac{\delta}{m} = \frac{\delta_1 (1 + n)}{m_1 (1 + n)} = C_1. \tag{19}$$

Aus allem Dargelegten ist zu ersehen, dass bei absoluter Korrelation der Störungen der einzelnen Erscheinungen, sowie bei Gleichheit ihres Charlierschen Faktors die aus den Einzelerscheinungen zusammengesetzten neuen Erscheinungen denselben Charlierschen Faktor aufweisen.

Da in den von Charlier angeführten Beispielen die beiden Erscheinungen miteinander fast absolut korrelativ verbunden waren, so musste auch der Störungskoeffizient bei den Gesamt- und Zwillingsgeburten dieselbe Grösse besitzen. Praktisch können aber Erscheinungen mit verschiedenen Korrelationskoeffizienten vorkommen, wodurch das ganze Problem viel komplizierter wird.

3. Nehmen wir nochmals zwei Erscheinungen mit gleichen Häufigkeiten und Störungsstreuungen, so erhalten wir nach (13)

$$\delta^2 = 2\delta_1^2 (1 + r).$$

Im Falle $r = 0$ ist (die Störungen sind voneinander unabhängig), erhalten wir

$$\delta^2 = 2\delta_1^2 \text{ und}$$

$$C = \frac{\sqrt{2}\,\delta_1}{2m_1} = \frac{C_1}{\sqrt{2}}. \tag{20}$$

Aus letzterem ersehen wir, dass die Störung dieselbe geblieben, die Störungsstreuung gewachsen ist, dagegen weist aber der Charliersche Faktor auf eine Verminderung der Störung hin.

Im Falle $r = -1$ ist (die Störungen stehen in negativer Korrelation), bekommen wir

$$\delta^2 = 0$$

und

$$C = 0. \tag{21}$$

In diesem Falle gibt es keine Störungsstreuung und damit auch keine Störung, was man auch aus dem Charlierschen Faktor ersieht.

Wenn wir eine Erscheinung finden, deren Störungsstreuung gleich Null ist, sind wir niemals sicher, ob diese Erscheinung sich nicht auf irgendeine Weise so in zwei andere zerlegen lasse, dass diesen letzteren Störungsstreuungen eignen, welche in negativer Korrelation stehen. In solchen Fällen können nach Charlier die Partialerscheinungen grosse Störungsstreuungen und auch einen grossen Störungskoeffizienten aufweisen. Es wird aber der Sinn des Störungskoeffizienten fraglich, denn während bei einer Gruppierung die grossen Störungen verschwinden, können sie bei andersartigen Gruppierungen neu hervorkommen.

Die Grösse des Charlierschen Faktors, welcher beim Zusammenfassen zweier Erscheinungen mit gleicher relativer Häufigkeit und gleicher Störungsstreuung entsteht, ist

$$C = \frac{\delta_1}{m_1}\sqrt{\frac{1+r}{2}} = \tag{22}$$

$$= C_1 \sqrt{\frac{1+r}{2}}.$$

Diese Formel erhalten wir, wenn wir den Wert für δ aus (18) in das Verhältnis $\frac{\delta}{m}$ einsetzen.

Während der Faktor $\sqrt{\frac{1+r}{2}}$ einen Wert, der zwischen 0 und 1 liegt, haben kann, ändert sich C in den Grenzen zwischen 0 und C_1 und es braucht keineswegs immer $C = C_1$ zu sein.

Betrachten wir ferner den allgemeinen Fall, in dem

$$m_2 = k m_1 \tag{23}$$

ist (früher war $m_1 = m_2$).

Wir nehmen den gleichen Charlierschen Faktor für beide Erscheinungen an. Somit ist

(24) $$C_1 = \frac{\delta_1}{m_1} = C_2 = \frac{\delta_2}{m_2}.$$

Aus letzterem folgt:

$$\delta_1 = C_1 m_1$$

und

$$\delta_2 = C_1 m_2.$$

Fassen wir die Klassen mit den Häufigkeiten m_1 und m_2 zusammen, so erhalten wir die Häufigkeit der neuen Klasse

(25) $$m = m_1 + m_2$$

und die zugehörige Störungsstreuung

(26) $$\delta = \sqrt{\delta_1^2 + \delta_2^2 + 2r\,\delta_1\,\delta_2} =$$

$$= C\sqrt{m_1^2 + m_2^2 + 2r\,m_1\,m_2}.$$

Berechnen wir den CharlierschenFaktor der neuen Klasse, so erhalten wir

(27) $$C = \frac{\delta}{m} = C_1 \sqrt{\frac{m_1^2 + m_2^2 + 2r\,m_1\,m_2}{(m_1 + m_2)^2}} =$$

$$= C_1 \sqrt{1 - \frac{2m_1\,m_2\,(1-r)}{(m_1 + m_2)^2}}.$$

Setzen wir statt m_2 seinen Wert km_1 aus (23) ein, so ist

(28) $$C = C_1 \sqrt{1 - \frac{2k\,(1-r)}{(1+k)^2}}.$$

Hieraus ist zu ersehen: wenn $r = 1$, so ist $C = C_1$, d. h. der Charliersche Faktor bleibt derselbe. In jedem anderen Falle ist $C < C_1$, und bei $k = 1$ und $r = -1$ ist $C = 0$.

Hier bleibt, wie auch im vorigen Falle, C immer zwischen 0 und 1.

Um zu untersuchen, welche Werte C bei verschiedenen Werten von k und r annimmt, schreiben wir

$$(29) \qquad \frac{C}{C_1} = \mu = \sqrt{1 - \frac{2k(1-r)}{(1+k)^2}}$$

und stellen ein Nomogramm her (Fig. 1), in dem die Grössen k und r auf den Koordinatenachsen abgetragen werden.

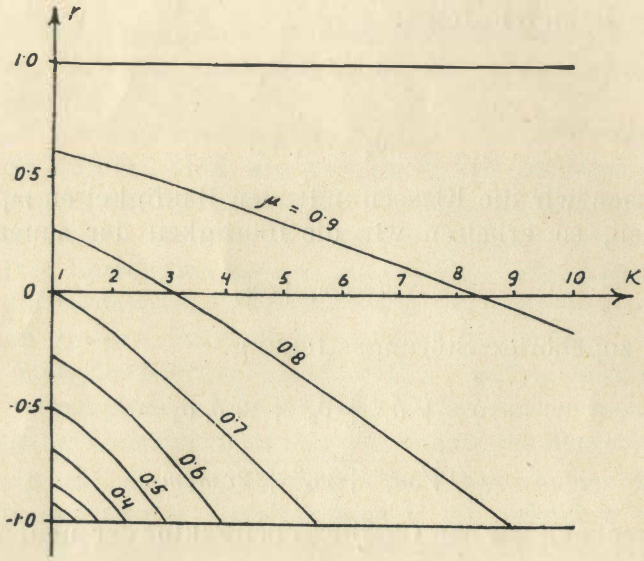

Fig. 1.

Aus dem Nomogramm ersehen wir, dass in einem Streifen für die am häufigsten vorkommenden r-Werte $(-0 \cdot 5 \leqslant r \leqslant 0 \cdot 5)$ μ die Werte von 0·5 bis 0·9, durchschnittlich den Wert von 0·8 hat. Der Umstand, dass dieser Wert der Eins genügend nahe kommt, hat oft dazu geführt, den Charlierschen Faktor als zweckmässig zu gebrauchen.

Weiter können wir, wenn wir gewisse Annahmen über r und k vorangehen lassen, Schlüsse über die Häufigkeit der Grösse μ ziehen. Der Korrelationskoeffizient r kann in den Grenzen -1 und $+1$ vorkommen, k kann aber theoretisch zwischen 1 und ∞ liegen. Praktisch kommt jedoch das Zusammenfassen der Klassen nur bei nicht allzugrossen k-Werten in Frage. Gewöhnlich ändert sich k zwischen 1 und 5 bis 6, selten bis 10.

Setzen wir nun voraus, dass der Korrelationskoeffizient r zwischen -1 und 1, und k zwischen 1 und 10 mit konstanten

Häufigkeiten liegen. Dann geben die Grössen der Flächen zwischen den Nomogrammkurven jene Häufigkeiten an, die jeder μ-Klasse entsprechen. Um sich über diese Häufigkeitsverteilung ein Bild zu verschaffen, schreiben wie die Gleichung (29) in folgender Form:

$$r = 1 - (1 - \mu^2)\,\frac{(1+k)^2}{2k}. \tag{30}$$

Zum Berechnen der Grösse der Flächen müssen erst die Schnittpunkte der Kurven mit der Geraden

$$r = -1$$

bestimmt werden. Setzen wir $r = -1$ in (30) ein und erhalten wir daraus für jedes μ das ihm entsprechende k_μ. Die folgende Tabelle (Tab. 1) enthält die entsprechenden Werte der genannten Grössen (die Zeilen μ und k_μ). Darauf berechnen wir die Fläche, die von den Geraden $k = 1$, $k = 10$, $r = -1$ und von einer Nomogrammkurve begrenzt ist. Falls die Nomogrammkurve die Gerade $r = -1$ in einem Punkte schneidet, für den $k < 10$, so muss die zu berechnende Fläche nur bis zu diesem Schnittpunkte genommen werden. Eine solche Fläche ist durch folgende Formel bestimmt:

$$Q_\mu = \int_1^k \left\{ 2 - (1 - \mu^2)\,\frac{(1+k)^2}{2k} \right\} dk. \tag{31}$$

In der folgenden Tabelle sind für jedes μ die entsprechenden Q-Werte berechnet, wobei k die obere Grenze des Integrals bedeutet.

Tab. 1.

μ	k_μ	k	Q_μ	ΔQ	p %
1·0	∞	10·00	18·00		
				6·63	36·9
0·9	19·00	10·00	11·37		
				5·83	32·4
0·8	9 00	9·00	5·54		
				3·00	16·7
0·7	5·67	5·67	2·54		
				1·30	7·2
0·6	4·00	4·00	1·24		
				0·65	3·6
0·5	3·00	3·00	0·59		
				0·33	1·8
0·4	2·33	2·33	0·26		
				0·17	0·9
0·3	1·86	1·86	0·09		
				0·06	0·3
0·2	1·50	1·50	0·03		
				0·03	0·2
0·1	1·25	1·25	0·00		
				0·00	0 0
0·0	1·00	1·00	0·00		

In der Zeile ΔQ sind die Areale zwischen den Flächen Q gegeben und diese sind somit proportional den Wahrscheinlichkeiten für jedes μ-Intervall. Die genannten Wahrscheinlichkeiten sind in der letzten Zeile, $p\,^0/_0$, in Prozenten angegeben.

Fig. 2 stellt die entsprechende Häufigkeitsverteilung dar. Aus ihr ist zu ersehen, dass die Wahrscheinlichkeit des Auftretens der Klassen mit höheren μ-Werten sehr gross ist.

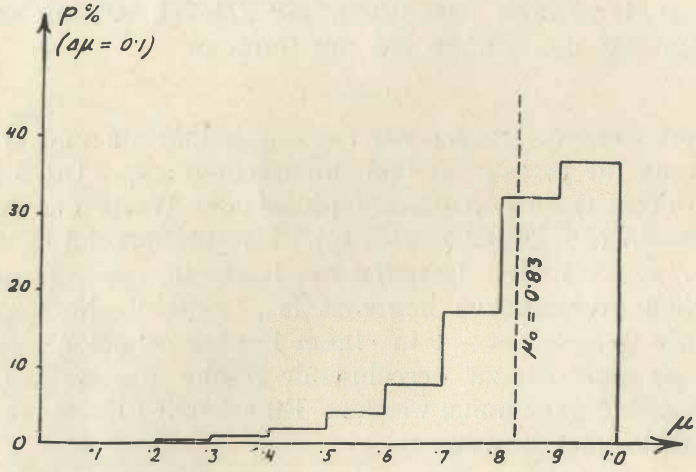

Fig. 2.

Das arithmetische Mittel der μ-Werte ist auf Grund dieser Häufigkeitsverteilung

$$\mu_0 = 0{\cdot}83.$$

Wenn wir unser Nomogramm auch für k-Werte, die grösser als 10 sind, erweitern (in unserem Beispiel war $k = 10$), dann wachsen die Wahrscheinlichkeiten der höheren μ-Werte und wächst somit auch das Mittel, und wenn wir dagegen die obere Grenze von k niedriger nehmen, tritt das Entgegengesetzte ein.

Im Falle, wo $k = 1$ ist, bekommen wir aus (30)

$$r = 2\mu^2 - 1. \tag{32}$$

Wenn die Korrelationskoeffizienten im Intervall $-1 \leqslant r \leqslant 1$ mit konstanter Häufigkeit auftreten, so ermöglicht die letzte Gleichung analog dem oben Gesagten die Häufigkeiten für jede μ-Klasse zu bestimmen, die in der folgenden Tabelle (Tab. 2) in Prozenten, $p\,^0/_0$, gegeben sind (relative Häufigkeit).

Tab. 2.

μ	r	$p\,^0/_0$	μ	r	$p\,^0/_0$
1·0	1·000		0·5	—0·500	
		19·0			9·0
0·9	0·620		0·4	—0·680	
		17·0			7·0
0·8	0·280		0·3	—0·820	
		15·0			5·0
0·7	—0·020		0·2	—0·920	
		13·0			3·0
0·6	—0·280		0·1	—0·980	
		11·0			1·0
			0·0	—1·000	

Fig. 3 stellt die entsprechende Häufigkeitsverteilung dar.

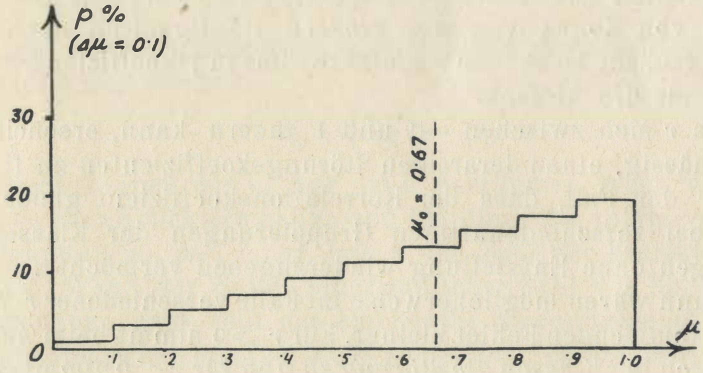

Fig. 3.

Das arithmetische Mittel ist $\mu_0 = 0{\cdot}67$, was auch aus der Formel (32) leicht abzuleiten ist[1]).

[1]) Die gesuchte Häufigkeit für jede μ-Klasse $d\mu$ ist proportional der entsprechenden Grösse dr. Die Wahrscheinlichkeitskurve ist somit durch die Ableitung von r gegeben. Es ist

$$p = a\,\frac{dr}{d\mu} = 4a\mu.$$

Da $\int_0^1 4a\mu d\mu = 1$ ist, finden wir, dass $a = {}^1/_2$ ist. Das arithmetische Mittel beträgt nach der bekannten Formel

$$\mu_0 = \frac{\int_0^1 2\mu^2\,d\mu}{\int_0^1 2\mu\,d\mu} = {}^2/_3.$$

Aus der Figur ist zu ersehen, dass, wenn $k = 1$ ist, auch die kleineren μ-Werte genügend stark mitsprechen, aber mit dem Wachsen der Grösse k wird der Einfluss der kleineren μ-Werte vermindert. Diese Tatsache erlaubt bei vielen Erscheinungen, bei denen k genügend gross ist, den Charlierschen Störungskoeffizienten mit genügendem Erfolg zu verwenden.

4. Aus dem Obengesagten folgt, dass in Fällen, wo die Partialerscheinungen nicht den Korrelationskoeffizienten $r = 1$ haben, der Charliersche Koeffizient unbefriedigend ist, weil er die Störung mit einem veränderlichen Mass schätzt. Beim Zusammenfassen der Klassen wird der Charliersche Koeffizient immer kleiner und dieses hat zur Folge, dass bei einer grösseren Anzahl von Kollektiven die grossen Häufigkeiten nach dem Charlierschen Mass stets kleinere Störungskoeffizienten aufweisen als die kleinen.

Da r sich zwischen -1 und 1 ändern kann, erscheint es zweckmässig, einen derartigen Störungskoeffizienten zu finden, der für den Fall, dass der Korrelationskoeffizient gleich Null wäre, bei verschiedenartigen Gruppierungen der Klassen die Störungen ohne Entstellung wiederzugeben vermöchte.

Dann wären möglicherweise im Falle verschiedener r-Werte die vorkommenden Fehler kleiner. Für $r > 0$ nimmt beim Zusammenfassen der Klassen die Störung zu und für $r < 0$ nimmt sie ab.

Ein solcher Zustand scheint auch natürlich zu sein, denn jene Einflüsse, welche die positive Korrelation verursachen, vergrössern auch die Störungsstreuung. Zur Addition der Störungen auf natürlichem Wege sind somit noch spezielle Einflüsse hinzugekommen, welche die Störungsstreuung noch vergrössern müssen. Bei negativer Korrelation kommen ebenfalls derartige Einflüsse hinzu, die die normale Summe der Störungen vermindern.

Auf Grund solcher Bedingungen, die natürlicher zu sein scheinen als beim Charlierschen Koeffizienten, kann analog ein Störungskoeffizient festgesetzt und mit diesem die Grösse der Störung gemessen werden.

Ein solcher Koeffizient hätte, ebenso wie auch der Charliersche Koeffizient, jenen schwerwiegenden Nachteil, dass beide nur gewissen Anforderungen zu genügen vermögen, also gewissermassen als „Geschmackssache“ aufgefasst werden könnten, da ihnen die inhaltliche Begründung fehlt und es schwer fiele zu sagen, warum der eine oder der andere besonders wesentlich sei.

Im folgenden ist zur Messung der Störungsstreuung ein neuer Störungskoeffizient eingeführt worden, welchem die Wahrscheinlichkeit zugrunde gelegt ist, die für jede Störungsstreuung die Existenz der Störung bestimmt.

§ 3. Die Wahrscheinlichkeit der Existenz der Störung.

Wenn eine Erscheinung mit der Wahrscheinlichkeit p auftritt, dann ist bei dem Kollektivumfang s die mittlere Häufigkeit

$$m = sp.$$

Die normale Streuung dieser Häufigkeit ist nach der Bernoullischen Formel

$$\sigma_B = \sqrt{sp\,(1-p)}. \tag{33}$$

Empirisch können wir σ_B nur dann bestimmen, wenn wir eine Reihe solcher Kollektive haben und für diese die entsprechenden Häufigkeiten $m_1, m_2 \ldots m_n$ (wo n die Zahl der Kollektive bedeutet). Wenn wir es mit einer endlichen Zahl von Kollektiven zu tun haben, dann fällt das empirische Streuungsmass σ_B mit dem theoretischen aus (33) nicht zusammen, es kommen immer Abweichungen zustande, deren quadratisches Mittel durch die Formel

$$\sigma_{\sigma_B} = \sqrt{\frac{sp\,(1-p)}{2n}} \tag{34}$$

ausgedrückt wird.

Nehmen wir den Variationskoeffizienten von σ, so bekommen wir

$$V_\sigma = \frac{\sigma_{\sigma_B}}{\sigma_B} = \frac{1}{\sqrt{2n}}. \tag{35}$$

Wenn man mit einer bestimmen Zahl von Kollektiven zu tun hat, ist $n = \text{const.}$ und

$$\frac{\sigma_{\sigma_B}}{\sigma_B} = \text{const.} \tag{36}$$

Wenn die empirisch bestimmte Streuung, die gewöhnlich mit σ_L (die Streuung der Lexisschen Reihe) bezeichnet wird (denn im allgemeinen Fall kann man die Existenz der Lexisschen Reihe voraussetzen und auch eine Bernoullische Reihe kann als Lexissche mit dem Lexisschen Faktor $L = 1$ aufgefasst werden), grösser als die theoretische Streuung σ_B ist, so kann die Differenz

$$\sigma_L - \sigma_B$$

als eine zufällig entstandene oder eine durch die Störungseinflüsse verursachte betrachtet werden. Darum kann geschrieben werden

$$\sigma_L - \sigma_B = c\sigma_{\sigma_B}, \tag{37}$$

wo die Grösse von c jene Wahrscheinlichkeiten bestimmt, mit welchen wir behaupten können, dass die Abweichung $\sigma_L - \sigma_B$ zufällig entsteht oder durch äussere Gründe verursacht wird.

Aus (37) bekommen wir

$$\frac{\sigma_L - \sigma_B}{\sigma_B} = c\,\frac{\sigma_{\sigma_B}}{\sigma_B}$$

oder

$$L = 1 + \frac{c}{\sqrt{2n}}, \tag{38}$$

denn $\frac{\sigma_L}{\sigma_B} = L$, den sog. Lexisschen Faktor.

Während $n = \text{const.}$ ist, zeigt L, in welchem Masse man an die Existenz von äusseren Einflüssen glauben kann. Je grösser L ist, um so eher kann angenommen werden, dass äussere Einflüsse auf die Erscheinung einwirken.

§ 4. Der normale Störungskoeffizient.

Wenn wir bei Erscheinungen mit verschiedenen relativen Häufigkeiten die Störungsstreuungen vergleichen, so schätzen wir die Störung gleich in dem Falle, wenn die Grössen

$$\frac{\sigma_L - \sigma_B}{\sigma_B}$$

gleich sind, denn dann haben alle Erscheinungen die gleiche Wahrscheinlichkeit für die Existenz von Störung.

Es sei der Kollektivumfang s, die relative Häufigkeit des Auftretens der Erscheinung p und die Störungsstreuung δ.

Da

$$\sigma_L^2 = \sigma_B^2 + \delta^2$$

und

$$\sigma_B{}^2 = sp\,(1 - p),$$

ist

(39) $$\sigma_L = \sqrt{sp\,(1 - p) + \delta^2}$$

und

(40) $$L = \sqrt{1 + \frac{\delta^2}{sp\,(1 - p)}}.$$

Bei konstanter Störung muss auch L konstant sein. Damit ist

(41) $$\frac{\delta}{\sqrt{sp\,(1 - p)}} = \text{const.} = \gamma.$$

Wenn wir bei verschiedenen p-Werten die Störung als konstant betrachten, dann ist auch γ konstant. Wenn aber die Wahrscheinlichkeit der Existenz von Störung bei grösserem γ grösser ist und bei kleinerem γ kleiner, dann können wir die Grösse der Störung mit γ messen, was deshalb besonders zu empfehlen ist, weil in der Formel (41) δ in erster Potenz auftritt, d. h. γ bei konstantem p, wie auch bei dem Charlierschen Koeffizienten, proportional der Störungsstreuung wächst.

Beim Übergang zu s als zu einer veränderlichen Grösse ist der weitere Entwicklungsgang der Berechnungen ganz einfach. Wenn wir bei einem bestimmten s-Werte die Grösse der Störung festgestellt haben, muss der entsprechende Störungskoeffizient so beschaffen sein, dass er sich bei Veränderung des Kollektivumfangs nicht ändert.

Nach (7) ist

$$\delta = s\xi.$$

Setzen wir dieses in (41) ein, so bekommen wir

(42) $$\gamma = \sqrt{s}\,\frac{\xi}{\sqrt{p\,(1 - p)}}.$$

Hieraus ist zu ersehen, dass γ eine noch unbrauchbare Grösse ist, da sie von s abhängig ist. Um dieses zu vermeiden, führen wir eine neue Grösse K ein:

(43) $$K = \frac{\gamma}{\sqrt{s}}.$$

Setzen wir in (43) den Wert von γ aus (42) ein, so erhalten wir

$$\boxed{K = \frac{\xi}{\sqrt{p(1-p)}}} \tag{44}$$

Da die Grösse K von s unabhängig ist und denselben Anforderungen genügt wie auch γ, so ist sie als zur Messung der Störung brauchbare Grösse zu bezeichnen.

In dem Falle, wenn ein Merkmal eine sehr kleine relative Häufigkeit hat (bei sogenannten „kleinen Zahlen"), ist

$$1 - p \approx 1$$

und

$$K \approx \frac{\xi}{\sqrt{p}}. \tag{45}$$

Im Falle $p = {}^1/_2 = 1 - p$, ist

$$K = \frac{\xi}{p} = C \quad (= 2\,\xi). \tag{46}$$

Das bedeutet, dass der Störungskoeffizient gleich dem Charlierschen ist.

§ 5. Die Beziehung zwischen *K* und *C*.

Nach (8) ist

$$C = \frac{\xi}{p}$$

und damit

$$\frac{K}{C} = \sqrt{\frac{p}{1-p}}. \tag{47}$$

Wenn $p < {}^1/_2$ ist, dann ist $K < C$
und wenn $p > {}^1/_2$ ist, dann ist $K > C$.

Wenn wir K als konstant betrachten, so ist ξ nach der Formel (44) proportional der Grösse $\sqrt{p(1-p)}$, und wenn wir C als konstant annehmen, ist ξ proportional p. Beide entsprechenden Funktionen sind in Fig. 4 dargestellt ($K = 1$ und $C = 1$), woraus zu ersehen ist, dass bei $p = {}^1/_2$ C und K gleich sind.

Die graphische Darstellung bringt noch eine interessante Tatsache zum Vorschein. Die eine Kurve ($K = \mathrm{const.}$) ist symmetrisch auf der Geraden $p = {}^1/_2$, die andere ($C = \mathrm{const.}$) nicht. Der letztere Fall bietet sehr wesentliche Argumente gegen den Charlierschen Faktor — eine sehr einfache Betrachtung zeigt nämlich, dass diese Kurve symmetrisch sein muss.

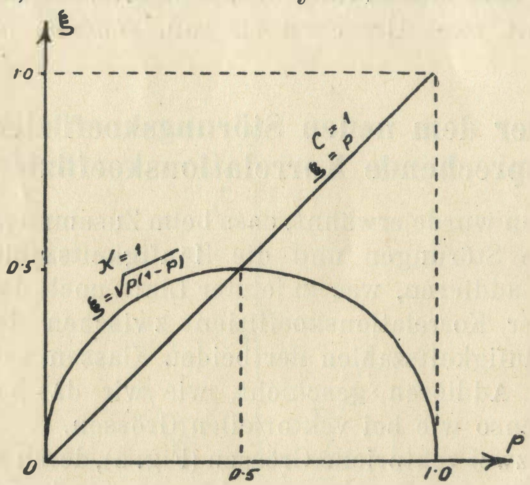

Fig. 4.

Wenn ein Kollektiv aus s Elementen besteht und von diesen m Elemente ein gewisses Merkmal haben, dann fehlt den übrigen $s - m$ Elementen dieses Merkmal. Das Fehlen dieses Merkmals ist für die genannten Elemente auch ein Merkmal. Wenn die eine Klasse mit der relativen Häufigkeit $p_1 = \frac{m}{s}$ eine Störung hat, dann muss auf Grund der absoluten Korrelation ($r = -1$) auch die andere Klasse, deren relative Häufigkeit $p_2 = 1 - p_1 = \frac{s - m}{m}$ ist, genau dieselbe Störung haben. Die Einflüsse, welche das Wachsen von p_1 verursachen, sind dieselben, welche gleichzeitig auch die Verminderung von p_2 hervorrufen. Wenn wir in einem Kartenpaket das Verhältnis der roten und schwarzen Karten verändern wollen, ist es gleichgültig, ob wir die Zahl der schwarzen Karten vergrössern oder die Zahl der roten vermindern. Eine Ursache — die Veränderung des Verhältnisses — trifft gleichzeitig beide Merkmale. **Die**

Grösse, mit welcher die Resultate dieses Einflusses gemessen werden, muss für jedes p und $1-p$ gleich gross sein. Wenn z. B. $p_1 = 0{\cdot}9$ ist, ist damit $p_2 = 0{\cdot}1$, und wenn eine Störung stattfindet, muss diese für beide Erscheinungen gleich sein. Der Charliersche Störungskoeffizient gibt aber im zweiten Fall eine 9 mal grössere Störung an als im ersten. Doch haben wir es nicht mit zwei Ursachen zu tun, sondern nur mit einer.

§ 6. Der dem neuen Störungskoeffizienten entsprechende Korrelationskoeffizient.

1. Oben wurde erwähnt, dass beim Zusammenfassen zweier Klassen die Störungen und die Häufigkeitszahlen sich nicht algebraisch addieren, was in letzter Linie noch davon abhängt, wie sich der Korrelationskoeffizient zwischen den Störungen und den Häufigkeitszahlen der beiden Klassen verhält.

Dieses Addieren geschieht, wie wir das sogleich sehen werden, ebenso wie bei vektoriellen Grössen.

Wenn zwei vektorielle Grössen (Fig. 5), deren skalare Werte δ_1 bzw. δ_2 sind, addiert werden, können wir den skalaren Wert ihrer vektoriellen Summe folgenderweise ausdrücken:

$$\delta^2 = \delta_1^2 + \delta_2^2 + 2\,\delta_1\,\delta_2 \cos\,\alpha, \tag{48}$$

wo α der Winkel zwischen δ_1 und δ_2 ist.

Vergleichen wir jetzt (48) mit (12), so sehen wir, dass beide ganz identisch sind, nur dass wir statt r im letzteren Fall $\cos\alpha$ haben. Diese Feststellung liefert eine sehr einfache Regel zum Berechnen der Störungsstreuung beim Zusammenfassen zweier Klassen, wenn die Störungsstreuungen der beiden Klassen bekannt sind: man muss die Störungsstreuungen vektoriell addieren, wobei der Winkel α zwischen den vektoriellen Grössen

$$\alpha = \text{arc cos } r \tag{49}$$

ist (r ist hier der Korrelationskoeffizient zwischen den Störungen beider Reihen).

Gehen wir von der obengenannten Eigenschaft bei Addition der Störungen aus, so wird auch der Sinn der allgemeinbekannten Formel

$$\sigma_L^2 = \sigma_B^2 + \delta^2$$

klar, denn die Streuung der Bernoullischen Reihe kann auch als von einer Störung (einer Störung, deren Entstehungsgründe uns nicht bekannt sind) beeinflusst betrachtet werden. Da der Korrelationskoeffizient zwischen der δ_B-Reihe (der Reihe, aus der σ_B berechnet ist) und der δ-Reihe gleich Null ist, d. h. $\alpha = \frac{\pi}{2}$, werden die Grössen σ_B und δ nach dem pythagoreischen Lehrsatz addiert.

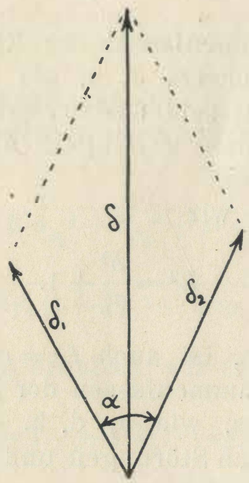

Fig. 5.

2. Aus (15) kann man folgern, dass wenn der Korrelationskoeffizient zwischen den Störungen zweier Reihen gleich eins ist, der Charliersche Störungskoeffizient bei Addition der Reihen konstant bleibt, falls er für jede einzelne Reihe konstant war. Jetzt können wir fragen: bei welchem Korrelationskoeffizienten zwischen den Störungen müssen die Klassen addiert werden, damit K konstant bleibt?

Es sei bemerkt, dass im allgemeinen Fall der Korrelationskoeffizient zwischen den Störungen nicht derselbe ist, wie derjenige zwischen den Häufigkeitszahlen. Aus dem Beobachtungsmaterial können wir nur den Korrelationskoeffizienten zwischen den Häufigkeitszahlen berechnen, nicht aber jenen zwischen den Störungen, denn die Grösse der letzteren ist uns fast immer unbekannt. Nur in solchen Fällen, wo wir selbst die Störungen hervorrufen können, wie z. B. beim Würfelversuch u. s. w., können wir sie errechnen.

Bei unserem Problem, wo K konstant sein muss, ist es aber leicht zu zeigen, dass die beiden Korrelationskoeffizienten gleich sind. Wir haben gesehen, dass nach (38) bei konstantem s auch L konstant sein muss, damit K konstant sei. Da

$$L = \frac{\sigma_L}{\sigma_B}$$

ist, muss beim Zusammenfassen der Klassen σ_L in demselben Verhältnis wie σ_B wachsen, d. h. der Korrelationskoeffizient zwischen den Häufigkeitszahlen ist derselbe wie der Korrelationskoeffizient zwischen den zufälligen Abweichungen in zwei Klassen eines Kollektivs.

Aus (40) erhalten wir

$$L^2 = \frac{\delta^2}{\sigma_B^2} + 1. \tag{50}$$

Da $L =$ konst. ist, ist auch $L^2 =$ const., und wir sehen, dass auch δ beim Zusammenfassen der Klassen im selben Verhältnis wachsen muss, wie σ_B, d. h. dass der Korrelationskoeffizient zwischen den Störungen und den Häufigkeitszahlen zweier Klassen derselbe ist.

Der Korrelationskoeffizient zwischen zufälligen Abweichungen kann in folgender Weise berechnet werden.

3. Es habe eine Klasse die relative Häufigkeit von p_1 und eine andere diejenige von p_2. Bezeichnen wir

$$p_2 = kp_1. \tag{51}$$

Die der ersten Klasse zugehörige Bernoullische Streuung beträgt

$$\sigma_{B_1} = \sqrt{sp_1(1 - p_1)} \tag{52}$$

und die der zweiten

$$\sigma_{B_2} = \sqrt{sp_2(1 - p_2)} = \tag{53}$$

$$= \sqrt{skp_1(1 - kp_1)}.$$

Fassen wir beide Klassen zusammen, so ergibt sich, dass die relative Häufigkeit

$$p = p_1 + p_2 = \tag{54}$$
$$= p_1 (1 + k)$$

und die Bernoullische Streuung

$$\sigma_B = \sqrt{sp(1-p)} = \tag{55}$$
$$= \sqrt{sp_1 (1 + k)(1 - p - kp_1)}$$

ist.

Wenn die Häufigkeitszahlen der beiden Klassen durch den Korrelationskoeffizienten r verbunden sind, so können wir schreiben

$$\sigma_B = \sqrt{\sigma_{B_1}{}^2 + \sigma_{B_2}{}^2 + 2r\,\sigma_{B_1}\sigma_{B_2}} = \tag{56}$$
$$= \sqrt{sp_1 (1 - p_1) + skp_1 (1 - kp_1) + 2rsp_1 \sqrt{k(1-p_1)(1-kp_1)}}.$$

Aus (55) und (56) erhalten wir die folgende Gleichung:

$$(1+k)(1-p_1-kp_1) = 1 - p_1 + k(1-kp_1) + 2r\sqrt{k(1-p_1(1-kp_1)}. \tag{57}$$

Die Gleichung (57) können wir in bezug auf r lösen und erhalten

$$r = -\frac{kp_1}{\sqrt{k(1-p_1)(1-kp_1)}}, \tag{58}$$

oder einfacher

$$r = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}}. \tag{59}$$

Bei solch einem Korrelationskoeffizienten zwischen den Störungen (oder auch zwischen den Häufigkeitszahlen)der beiden Klassen bleibt beim Zusammenfassen der Klassen K konstant. Was stellt dieser Koeffizient dar? Er ist nichts anderes, als der normale Korrelationskoeffizient[1]) zwischen den Häufigkeits-

[1]) Normal kann er genannt werden, weil zu seiner Entstehung keine äusseren Gründe nötig sind und er nur dadurch verursacht wird, dass die Summe aller relativen Häufigkeiten bei aller Art Abweichungen immer gleich eins ist.

zahlen irgendwelcher zweier Klassen in einem Kollektiv. Das Gesagte wird durch Folgendes noch erläutert.

3. Es seien in einem Kollektiv zwei Klassen vorhanden mit der relativen Häufigkeit p_1 bzw. p_2. Wenn die Häufigkeitszahl einer Klasse (mit der relativen Häufigkeit p_1) eine Abweichung d_1 erhält, dann erhält infolge der absoluten Korrelation ($r = -1$) die Häufigkeitszahl der Ergänzungsklasse (mit der relativen Häufigkeit $1 - p_1$) die Abweichung $-d_1$. Die mittlere Abweichung der zweiten Klasse (mit der relativen Häufigkeit p_2) ist dann d_2, die nach folgender Formel bestimmt wird:

$$d_2 = -\frac{p_2}{1 - p_1} d_1. \tag{60}$$

Die erhaltene Gleichung ist eine Regressionsgleichung für d_1 und d_2. Führen wir in diese Gleichung den Korrelationskoeffizienten ein, so ist nach der Korrelationstheorie

$$d_2 = r \cdot \frac{\sigma_{B_2}}{\sigma_{B_1}} d_1. \tag{61}$$

Aus (60) und (61) können wir r bestimmen:

$$r = -\frac{p_2}{1 - p_1} \cdot \frac{\sigma_{B_1}}{\sigma_{B_2}}. \tag{62}$$

Da

$$\sigma_{B_1} = \sqrt{s p_1 (1 - p_1)}$$

und

$$\sigma_{B_2} = \sqrt{s p_2 (1 - p_2)}$$

sind, ist

$$r = -\frac{p_2}{1 - p_1} \sqrt{\frac{p_1 (1 - p_1)}{p_2 (1 - p_2)}} = \tag{63}$$

$$= -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}},$$

wie wir das auch aus (59) ersehen haben.

§ 7. Die Korrelationskoeffizienten zwischen Häufigkeitszahlen und zwischen Störungen.

1. Oben ist gesagt worden, dass im allgemeinen Fall der Korrelationskoeffizient zwischen den Störungen mit demjenigen zwischen den Häufigkeitszahlen nicht zusammenfällt. Im folgenden wollen wir die Beziehung zwischen diesen Korrelationskoeffizienten untersuchen.

Es seien zwei Klassen gegeben, deren Bernoullische Streuungen σ_{B_1} bzw. σ_{B_2}, deren Störungsstreuungen δ_1 bzw. δ_2 und die Lexisschen Streuungen σ_1 bzw. σ_2 sind. Fassen wir beide Klassen zusammen, so erhalten wir eine neue Klasse, für welche wir die obengenannten Streuungen mit σ_B, δ und σ bezeichnen.

Es sei r der Korrelationskoeffizient zwischen den Häufigkeitszahlen und r^* der Korrelationskoeffizient zwischen den Störungen der beiden Klassen.

Im vorigen Kapitel ist gezeigt worden, dass der Korrelationskoeffizient zwischen den Häufigkeitszahlen dei Bernoullischen Reihen gleich

$$-\sqrt{\frac{p_1\, p_2}{(1-p_1)\,(1-p_2)}}$$

ist.

Daher können wir schreiben:

$$\sigma_B^2 = \sigma_{B_1}^2 + \sigma_{B_2}^2 - 2\,\sigma_{B_1}\,\sigma_{B_2}\sqrt{\frac{p_1\, p_2}{(1-p_1)\,(1-p_2)}} = \tag{64}$$

$$= \sigma_{B_1}^2 + \sigma_{B_2}^2 - 2\,sp_1p_2,$$

da

$$\sigma_{B_1} = \sqrt{sp_1\,(1-p_1)}$$

und

$$\sigma_{B_2} = \sqrt{sp_2\,(1-p_2)}$$

ist.

Ebenfalls können wir schreiben

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\,r\,\sigma_1\,\sigma_2: \tag{65}$$

ein Addieren der Lexischen Streuungen).

Da nach (5)

$$\sigma_1^2 = \delta_1^2 + \sigma_{B_1}^2,$$

$$\sigma_2^2 = \delta_2^2 + \sigma_{B_2}^2$$

und

$$\sigma^2 = \delta^2 + \sigma_B^2$$

ist, so bekommen wir aus (65)

$$\delta^2 + \sigma_B^2 = \delta_1^2 + \sigma_{B_1}^2 + \delta_2^2 + \sigma_{B_2}^2 + 2\, r\, \sigma_1\, \sigma_2. \tag{66}$$

Setzen wir hier für σ_B^2 seinen Ausdruck aus (64) ein, so erhalten wir

$$\delta^2 = \delta_1^2 + \delta_2^2 + 2\; (s p_1 p_2 + r\, \sigma_1\, \sigma_2). \tag{67}$$

Da der Korrelationskoeffizient zwischen den Störungen r^* ist, können wir nach der Korrelationstheorie schreiben:

$$\delta^2 = \delta_1^2 + \delta_2^2 + 2\, r^*\, \delta_1\, \delta_2. \tag{68}$$

Aus den Gleichungen (67) und (68) erhalten wir

$$r^*\, \delta_1\, \delta_2 = s p_1 p_2 + r\, \sigma_1\, \sigma_2, \tag{69}$$

wo

$$\boxed{r^* = \frac{s p_1 p_2 + r\, \sigma_1\, \sigma_2}{\delta_1\, \delta_2}} \tag{70}$$

ist, oder

$$\boxed{r = \frac{r^*\, \delta_1\, \delta_2 - s p_1 p_2}{\sigma_1\, \sigma_2}} \tag{71}$$

ist.

2. Der maximale Wert von r^* ist $r^* = 1$. Damit ist

$$r_{max} = \frac{\delta_1\, \delta_2 - s p_1 p_2}{\sigma_1\, \sigma_2} < 1, \tag{72}$$

denn schon

$$\frac{\delta_1\, \delta_2}{\sigma_1\, \sigma_2} < 1,$$

weil $\delta_1 < \sigma_1$ und $\delta_2 < \sigma_2$ sind.

Aus (72) ist ersichtlich, dass der Korrelationskoeffizient zwischen den Häufigkeitszahlen niemals gleich eins ist, auch dann nicht, wenn der Korrelationskoeffizient zwischen den Störungen gleich eins ist. Der Grund hierfür ist darin zu suchen, dass die den Störungen zukommenden zufälligen Abweichungen (die Abweichungen der Bernoullischen Reihe) eine Verminderung des Korrelationskoeffizienten verursachen.

Der minimale Wert von r^* ist $r^* = -1$. Damit ist

(73) $$r_{min} = -\frac{\delta_1 \delta_2 + s p_1 p_2}{\sigma_1 \sigma_2}.$$

Hier ist zu beweisen, dass

$$\frac{\delta_1 \delta_2 + s p_1 p_2}{\sigma_1 \sigma_2} \leqq 1$$

ist, denn aus (73) ist nicht sogleich zu ersehen, dass der Zähler niemals grösser als der Nenner sein kann.

Wir müssen zeigen, dass

(74) $$\delta_1 \delta_2 + s p_1 p_2 \leqq \sigma_1 \sigma_2.$$

a) Betrachten wir zuerst den Fall, wo

$$\delta_1 = \delta_2 = 0 \text{ ist,}$$

d. h. wo wir es nur mit Bernoullischen Reihen zu tun haben. In diesem Fall ist

$$\sigma_1 = \sigma_{B_1}$$

und

$$\sigma_2 = \sigma_{B_2}.$$

Weil die Summe der relativen Häufigkeiten aller Klassen nicht grösser als eins sein kann, können wir schreiben

(75) $$p_1 + p_2 \leqq 1$$

oder

(76) $$0 \leqq 1 - p_1 - p_2.$$

Addieren wir zu den beiden Seiten $p_1\ p_2$, so ist

$$p_1 p_2 \leqslant 1 - p_1 - p_2 + p_1 p_2 \tag{77}$$

oder

$$p_1 p_2 \leqslant (1 - p_1)\ (1 - p_2). \tag{78}$$

Nun multiplizieren wir die beiden Seiten mit

$$s^2\ p_1\ p_2,$$

dann ist

$$s^2 p_1^{\ 2} p_2^{\ 2} \leqslant s p_1 (1 - p_1)\, s p_2 (1 - p_2). \tag{79}$$

Aus beiden Seiten ziehen wir nun die Quadratwurzeln und bekommen

$$s p_1 p_2 \leqslant \sigma_{B_1} \sigma_{B_2}, \tag{80}$$

was auch zu beweisen war.

b) Nun gehen wir zum allgemeinen Fall über, in dem

$$\delta_1 \neq 0 \text{ und } \delta_2 \neq 0.$$

Wir gehen von folgendem Ausdruck aus:

$$(\sigma_{B_1} \delta_2 - \sigma_{B_2} \delta_1)^2 \geqslant 0. \tag{81}$$

In anderer Form lautet das:

$$\sigma_{B_1}^2 \delta_2^{\ 2} + \sigma_{B_2}^2 \delta_1^{\ 2} \geqslant 2\, \sigma_{B_1} \sigma_{B_2} \delta_1 \delta_2. \tag{82}$$

Addieren wir zu den beiden Seiten

$$\sigma_{B_1}^2 \sigma_{B_2}^2 + \delta_1^{\ 2} \delta_2^{\ 2},$$

so ist

$$(\sigma_{B_1}^2 + \delta_1^{\ 2})(\sigma_{B_2}^2 + \delta_2^{\ 2}) \geqslant (\sigma_{B_1} \sigma_{B_2} + \delta_1 \delta_2)^2. \tag{83}$$

Ziehen wir die Quadratwurzeln, dann ist

$$\sigma_1 \sigma_2 \geqslant \sigma_{\ _1} \sigma_{B_2} + \delta_1 \delta_2. \tag{84}$$

Setzen wir in dieser Formel für $\sigma_{B_1} \sigma_{B_2}$ seinen Ausdruck aus (80) ein, so ist

$$\sigma_1 \sigma_2 \geqslant s p_1 p_2 + \delta_1 \delta_2, \tag{85}$$

was wir auch beweisen wollten.

Damit ist gezeigt, dass

$$|r| \leqslant 1$$

ist.

3. Jetzt können wir auch prüfen, ob unser Schluss nach (50) richtig ist: dass im Falle $K_1 = K_2$ ist (K_1 und K_2 sind die Störungskoeffizienten für beide Klassen), die beiden Korrelationskoeffizienten r und r^* gleich sind.

Da nach (59) der Korrelationskoeffizient zwischen den Häufigkeitszahlen

$$r = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}}$$

ist, so können wir nach (70) den Korrelationskoeffizienten zwischen den Störungen bestimmen. Somit ist

(86)
$$r^* = \frac{s p_1 p_2 - \sigma_1 \sigma_2 \sqrt{\dfrac{p_1 p_2}{(1-p_1)(1-p_2)}}}{\delta_1 \delta_2} =$$

$$= \frac{s p_1 p_2 \left(1 - \dfrac{\sigma_1 \sigma_2}{\sigma_{B_1} \sigma_{B_2}}\right)}{\delta_1 \delta_2} =$$

$$= \frac{s p_1 p_2 (\sigma_{B_1} \sigma_{B_2} - \sigma_1 \sigma_2)}{\delta_1 \delta_2 \sigma_{B_1} \sigma_{B_2}}.$$

Um diese Gleichung zu vereinfachen, nehmen wir die Beziehungen zwischen σ, σ_B und δ zur Hilfe.

Nach (7) und (44) schreiben wir:

$$K = \frac{\delta}{s\sqrt{p(1-p)}}.$$

Falls $K_1 = K_2$ ist, ist

(87)
$$\frac{\delta_1}{s\sqrt{p_1(1-p_1)}} = \frac{\delta_2}{s\sqrt{p_2(1-p_2)}}.$$

Multiplizieren wir die beiden Seiten mit $\sqrt{s}$, so bekommen wir:

(88) $$\frac{\delta_1}{\sigma_{B_1}} = \frac{\delta_2}{\sigma_{B_2}},$$

oder:

(89) $$\frac{\delta_1^2}{\sigma_{B_1}^2} = \frac{\delta_2^2}{\sigma_{B_2}^2}.$$

Etwas umgeformt:

(90) $$\frac{\delta_1^2}{\delta_1^2 + \sigma_{B_1}^2} = \frac{\delta_2^2}{\delta_2^2 + \sigma_{B_2}^2},$$

oder:

(91) $$\sigma_1^2 \delta_2^2 - \sigma_2^2 \delta_1^2 = 0.$$

Addieren wir jetzt zu den beiden Seiten

$$\sigma_1^2 \sigma_2^2 + \delta_1^2 \delta_2^2 - 2\, \sigma_1^2 \delta_2^2$$

hinzu, so erhalten wir:

(92) $$(\sigma_1^2 - \delta_1^2)(\sigma_2^2 - \delta_2^2) = \sigma_1^2 \sigma_2^2 + \delta_1^2 \delta_2^2 - 2\, \sigma_1^2 \delta_2^2.$$

Da nach (91)

$$\sigma_1 \delta_2 = \sigma_2 \delta_1$$

ist, so ist

(93) $$2\, \sigma_1^2 \delta_2^2 = 2\, \sigma_1 \sigma_2 \delta_1 \delta_2.$$

Setzen wir diesen Ausdruck in (92) ein, dann ist

(94) $$(\sigma_1^2 - \delta_1^2)(\sigma_2^2 - \delta_2^2) = (\sigma_1 \sigma_2 - \delta_1 \delta_2)^2,$$

oder:

(95) $$\sigma_{B_1} \sigma_{B_2} = \sigma_1 \sigma_2 - \delta_1 \delta_2,$$

oder:

(96) $$\sigma_{B_1} \sigma_{B_2} - \sigma_1 \sigma_2 = -\delta_1 \delta_2.$$

Setzen wir diese Ausdrücke in (86) ein, dann ist

(97) $$r^* = -\frac{s p_1 p_2}{\sigma_{B_1} \sigma_{B_2}} =$$

$$= -\frac{s p_1 p_2}{\sqrt{s^2 p_1 p_2 (1 - p_1)(1 - p_2)}} =$$

$$= -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}.$$

Damit ist auch bewiesen, dass bei $K_1 = K_2$ auch $r = r^*$ ist, falls

$$r = r^* = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}}$$

ist.

Jetzt können wir fragen, bei welchem r im allgemeinen Fall $r^* = r$ ist, wenn $K_1 \neq K_2$ ist.

Nach (70) können wir schreiben:

$$r = \frac{s p_1 p_2 + r \sigma_1 \sigma_2}{\delta_1 \delta_2}, \tag{98}$$

wo

$$r = -\frac{s p_1 p_2}{\sigma_1 \sigma_2 - \delta_1 \delta_2} \text{ [1]}. \tag{99}$$

In unserem Fall, wo $K_1 = K_2$ ist, erhalten wir aus (95):

$$\sigma_1 \sigma_2 - \delta_1 \delta_2 = \sigma_{B_1} \sigma_{B_2}$$

und setzen diesen Ausdruck in (99) ein. Dann erhalten wir:

$$r = -\frac{s p_1 p_2}{\sigma_{B_1} \sigma_{B_2}} = \tag{100}$$

$$= -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}}.$$

4. Die Formel (71) zeigt klar, dass im Falle $r^* > 0$, immer $r < r^*$ ist. Deshalb können solche Fälle vorkommen, wo r^* sehr gross und daher auch der Charliersche Störungskoeffizient gut anwendbar ist, wo aber der Korrelationskoeffizient zwischen den Häufigkeitszahlen zweier Klassen nicht besonders hoch ist. Dieses ist meistens dann der Fall, wenn der Lexissche Faktor sich nicht viel von Eins unterscheidet, denn dann sind δ_1 und δ_2 klein und damit auch r. Deshalb ist auch aus den Häufigkeitszahlen niemals ersichtlich, mit einem wie grossen Korrelationskoeffizienten zwischen den Störungen man es zu tun hat. Dass die genannten Differenzen zwischen r und r^* ganz gross sein können, zeigt uns das folgende Beispiel.

[1] Aus (85) ist zu ersehen, dass auch jetzt $|r| \leqq 1$ ist.

Es sei für zwei Klassen

$$p_1 = p_2 = 0{\cdot}1,$$
$$s = 1000 \quad \text{und}$$
$$r^* = 1.$$

a) Nehmen wir an:

$$\delta_1 = \delta_2 = 20.$$

Dann ist nach (5)

$$\sigma^2 = s\,p\,(1 - p) + \delta^2 =$$
$$= 90 + 400 = 490$$

und nach (71)

$$r = \frac{400 - 10}{490} = 0{\cdot}796$$

(genügend weit von eins).

b) Nehmen wir nun an:

$$\delta_1 = \delta_2 = 10.$$

Dann ist

$$\sigma^2 = 90 + 100 = 190$$

und

$$r = \frac{100 - 10}{190} = 0{\cdot}473$$

(schon ganz klein).

c) Zuletzt nehmen wir an:

$$\delta_1 = \delta_2 = 4.$$

Dann ist

$$\sigma^2 = 90 + 16 = 106$$

und

$$r = \frac{16 - 10}{106} = 0{\cdot}057.$$

Im letzten Falle kann man aus den empirischen Daten nicht herauslesen, dass zwischen den Störungen eine so hohe Korrelation herrscht ($r^* = 1$), denn der Korrelationsfaktor zwischen den Häufigkeitszahlen gleicht fast Null.

Bei noch kleineren δ-Werten kann dieser Korrelationskoeffizient sogar negativ werden und erreicht bei $\delta_1 = \delta_2 = 0$ seinen Grenzwert, den normalen Korrelationskoeffizienten

$$r = -\sqrt{\frac{p_1 \, p_2}{(1-p_1)(1-p_2)}} = -0{\cdot}111.$$

§ 8. Die Verallgemeinerung des Störungskoeffizienten.

Das Ergebnis der zwei letzten Kapitel zusammenfassend können wir sagen: wenn zwei Erscheinungen, deren Störungen gleich sind, zusammengefasst werden und wenn der Korrelationskoeffizient zwischen den entsprechenden Häufigkeitszahlen oder zwischen den Störungen gleich dem normalen Korrelationskoeffizienten ist, dann bleibt die Störung konstant. Ein solches Kollektiv, in dem zwischen den einzelnen Klassen besondere Beziehungen fehlen, kann man ein **normales Kollektiv** nennen.

Es kann ja andersartige Kollektive geben, und in der praktischen Statistik kann man genug solche finden, in denen die Häufigkeitszahlen einzelner Klassen miteinander in Korrelation stehen. Wie muss dort beim Zusammenfassen der Klassen die Störung geschätzt werden? Wenn zwei Klassen, deren relative Häufigkeiten gleich sind, d. h. $p_1 = p_2$, und ebenfalls auch die Störungsstreuungen, d. h. $\delta_1 = \delta_2$, zusammengefasst werden, dann hängt die Störungsstreuung der neuen Klasse, wie wir das schon gesehen haben, vom Korrelationskoeffizienten zwischen den Störungen ab. Wenn dieser Korrelationskoeffizient $r = -1$ ist, dann wird die Störungsstreuung gleich Null sein. Man muss nun annehmen, dass die Störung, die mit Hilfe des Störungskoeffizienten C oder K gemessen wird, gleich Null ist, — oder ist sie ebenso gross, wie diejenige in den Partialklassen, wo die Störungsstreuung gleich $\delta_1 = \delta_2$ war? In unserem Falle ist es offenbar klar, dass die Störung gleich Null zu schätzen ist. Wir haben es demnach mit einer Klasse zu tun, deren Störungsstreuung gleich Null ist, und der Umstand, dass diese Klasse sich so in zwei teilen lässt, dass die beiden Partialklassen von Null abweichende Störungen haben, verändert daran nichts. Es kann ja leicht ein Fall vorkommen, wo eine Klasse sich auf viele Arten in zwei zerlegen lässt und ein jedes Paar seine eigene Störungsstreuung besitzt.

Wenn es Gründe gibt, die verursachen, dass die Abweichungen der beiden Klassen den Korrelationsfaktor $r = -1$ haben, kann man diese Gründe in bezug auf die durch das Zusammenfassen erhaltene Klasse als solche, die die Störung vermindern, betrachten. Das Umgekehrte ist der Fall, wenn $r > 0$. Dann sind die die Korrelation verursachenden Gründe in bezug auf die neue Klasse solche, die die Störung vergrössern, und in diesem Fall ist es natürlich, dass die Störung der neuen Klasse grösser sein muss als die ihrer beiden Komponenten.

Ob eine zusammengesetzte Klasse (und als eine Zusammenfassung von zwei Klassen können wir eine jede Klasse betrachten) eine grosse Störungsstreuung deshalb besitzt, weil irgendwelche zwei ihrer Komponenten grosse Störungsstreuungen haben, oder weil die geringen Störungsstreuungen der Komponenten infolge der korrelativen Beziehungen gross geworden sind, ist gleichgültig; wir konstatieren nur die Tatsache. Darum ist es auch selbstverständlich, dass die Störung nicht nur im normalen, sondern auch in jedem Kollektiv mit Hilfe des Störungskoeffizienten K gemessen wird.

§ 9. Die Anwendung der Resultate auf die Klimatologie.

1. Da die Häufigkeitsverteilungen der klimatologischen Grössen (Temperatur, Luftdruck, Feuchtigkeit usw.) für einen bestimmten Ort und ein bestimmtes Zeitintervall charakteristisch sind, d. h. eine Stetigkeit aufweisen, hat auch jede Klasse einer klimatologischen Grösse eine bestimmte relative Häufigkeit. Wenn das genannte Zeitintervall sich wiederholt, bilden die Häufigkeitszahlen jeder Klasse eine statistische Reihe, und man kann mit Hilfe der statistischen Methoden untersuchen, ob die Häufigkeitszahlen jeder Klasse einer klimatologischen Grösse eine Streuung von zufälligem Charakter (die Bernoullische Reihe) haben oder nicht. Wenn die Streuung grösser als die der Bernoullischen Reihe ist, dann kann man von einer Störung sprechen, und es ist unsere Aufgabe festzustellen, welche klimatologische Elemente, in welchem Intervall und welche Störungen haben.

Hier haben wir aber auch eine Gelegenheit, die beiden Störungskoeffizienten (den Charlierschen und den in vorliegender Arbeit gebotenen) anzuwenden und die Resultate zu vergleichen. Darauf kann auch nach intuitiver Beurteilung entschieden werden, welcher Störungskoeffizient grössere Vorzüge bietet.

Zu diesem Zweck sind die Beobachtungen des Meteorologischen Observatoriums in Tartu der Jahre 1926—1933 (incl.) herangezogen worden, wobei als zu betrachtende Elemente Temperatur, Luftdruck, Windgeschwindigkeit und relative

Tab. 3.

t	m	σ_L	σ_B	δ	$K\%$	$C\%$	t	m	σ_L	σ_B	δ	$K\%$	$C\%$
29	4	5	2·0	5	2·7	125	0	427	85	20·6	83	4·4	19
28	10	15	3·2	15	5·1	150	— 0	298	41	17·3	37	2·3	12
27	14	13	3·7	12	3·4	86	— 1	254	33	15·9	34	2·3	13
26	28	21	5·3	20	4·0	71	— 2	251	61	15·8	59	4·0	24
25	36	22	6·0	21	3·7	58	— 3	219	69	14·8	67	4·9	31
24	46	21	6·8	20	3·2	44	— 4	205	49	14·3	47	3·6	23
23	61	22	7·8	20	2·7	33	— 5	193	66	13·9	64	5·0	33
22	78	25	8·8	23	2·8	30	— 6	169	46	13·0	44	3·7	26
21	85	24	9·2	22	2·6	26	— 7	142	33	11·9	31	2·7	22
20	111	26	10·5	24	2·1	23	— 8	132	45	11·5	44	4·2	33
19	135	26	11·6	23	2·1	17	— 9	132	40	11·5	38	3·6	29
18	179	34	13·4	31	2·5	17	—10	110	33	10·5	31	3·2	28
17	209	26	14·4	22	1·6	10	—11	101	26	10·0	24	2·6	24
16	261	18	16·1	8	0·5	3·1	—12	92	41	9·5	40	4·5	43
15	274	33	16·5	29	1·9	11	—13	74	33	8·6	32	4·0	43
14	296	35	17·2	31	1·9	10	—14	70	37	8·4	36	4·7	52
13	314	64	17·7	61	3·7	19	—15	50	16	7·1	15	2·3	30
12	302	63	17·4	61	3·8	20	—16	46	21	6·8	20	3·2	43
11	300	62	17·3	60	3·7	20	—17	35	17	5·9	16	2·9	46
10	274	48	16·5	45	3·0	16	—18	25	10	5·0	9	1·9	36
9	282	43	16·8	40	2·6	14	—19	22	14	4·7	13	3·0	59
8	283	43	16·8	40	2·5	14	—20	12	6	3·5	5	1·5	43
7	241	38	15·5	34	2·4	14	—21	10	9	3·2	8	2·7	80
6	265	17	16·3	6	0·4	2·3	—22	9	7	3·0	6	2·1	67
5	269	46	16·4	43	2·8	16	—23	10	11	3·2	11	3·7	110
4	263	55	16·2	53	3·5	20	—24	7	10	2·6	10	4·0	143
3	288	58	16·9	55	3.5	19	—25	6	8	2·4	8	3·5	125
2	315	73	17·8	71	4·3	22	—26	4	8	2·0	8	4·3	200
1	436	139	20·9	132	7·0	30	—27	2	7	1·4	7	5·3	350

Feuchtigkeit genommen worden sind. Die Häufigkeitszahlen sind in Klassenbreiten bei einer Temperatur von je 1° C, bei einem Luftdruck von je 4 mb, bei einer Windgeschwindigkeit von je 1 m/sec und bei einer relativen Feuchtigkeit von je 1 % gezählt worden. Beim Zählen sind die stündlichen Beobachtungs-

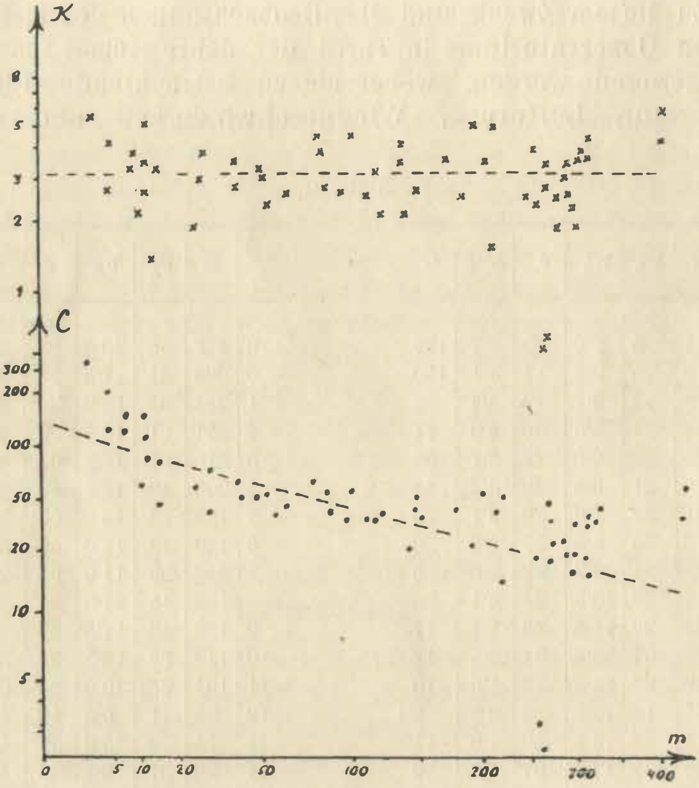

Fig. 6.

daten verwandt und als Kollektivumfang ist ein ganzes Jahr ($s = 365^1/_4 . 24 = 8766$) genommen worden.

2. Zuerst wollen wir die Temperatur betrachten. In der umstehenden Tabelle (Tab. 3) sind die nötigen Charakteristiken der statistischen Reihen gegeben. In Kolumne (t) sind alle zu betrachtenden Temperaturklassen, in Kolumne (m) die mittleren Häufigkeitszahlen während der acht Jahre, in Kolumne (σ_L) das Streuungsmass der Häufigkeitszahlen, in Kolumne (σ_B) das Streu-

ungsmass der Bernoullischen Reihe, in Kolumne (δ) die Störungsstreuungen ($\delta = \sqrt{\sigma_L^2 - \sigma_B^2}$), in Kolumne ($K$) die in dieser Arbeit gebotenen Störungskoeffizienten und in Kolumne (C) die Charlierschen Störungskoeffizienten angeführt.

Schon ein oberflächlicher Blick auf die Kolumnen (K) und (C) zeigt, dass K für alle Klassen mehr oder weniger konstant bleibt, C aber für die Klassen mit kleinen relativen Häufigkeiten gross und für solche mit grossen relativen Häufigkeiten klein ist. Besonders deutlich tritt das Gesagte in Fig. 6 zum Vorschein, wo die Abhängigkeit der Grössen K und C von der Häufigkeit m graphisch dargestellt ist.

Bemerkung: Die Skalen auf den Achsen m, K und C sind so gewählt worden, dass die Häufigkeitsverteilungen von m, K und C möglichst symmetrisch und damit der Normalverteilung möglichst nahe seien. Damit nähert sich das Korrelationsfeld dem linearen Normalfelde, und die korrelative Abhängigkeit kann durch den gewöhnlichen Korrelationsfaktor bestimmt werden.

Im ersten Fall (die Abhängigkeit der Grössen K und m voneinander betreffend) ist der Korrelationsfaktor gleich Null, was auch in Fig. 6 sichtbar ist[1]).

Im zweiten Fall (die Abhängigkeit der Grössen C und m voneinander betreffend) ist

$$r = -0{\cdot}78 \pm 0{\cdot}05,$$

was die Abhängigkeit von C und m voneinander sehr deutlich hervorhebt.

Da K und C den inhaltlichen Wert der Störung angeben müssen, zeigt K, dass die Störung über die ganze Variationsbreite konstant ist; C aber zeigt, dass die Temperaturklassen mit grosser Häufigkeit viele Male (über 10 mal) weniger als die Temperaturklassen mit kleiner Häufigkeit gestört werden — oder dass die sehr niedrigen und sehr hohen Temperaturen viel stärker gestört sind als die mittleren.

Wenn man nun fragen sollte, welche Behauptung natürlicher erscheine, so müsste man der ersteren den Vorzug geben. Erstens ist sie einfacher und zweitens sind in der Meteorologie keine Gründe bekannt, die zu zeigen vermöchten, dass bei der

---

[1]) Tatsächlich war $r = -0{\cdot}02$, der Fehler des Korrelationskoeffizienten $\sigma r = 0{\cdot}13$, wodurch praktisch r gleich Null ist.

Temperatur die Störung über die Variationsbreite so verteilt ist, wie es der Charliersche Störungskoeffizient verlangt.

Unter Benutzung der in Fig. 6 gegebenen Skalen ($x = \sqrt{m}$ und $y = \log (C - 15)$) erhalten wir nach Durchführung der entsprechenden Berechnungen die Regressionsgleichungen

$$C = 120\, e^{-0{\cdot}150 \sqrt{m}} + 1{\cdot}5$$

und

$$K = 3{\cdot}2.$$

Um die Abhängigkeit zwischen K und t (^{0}C) zu veranschaulichen, ist in Fig. 7 der entsprechende Gang von K wiedergegeben.

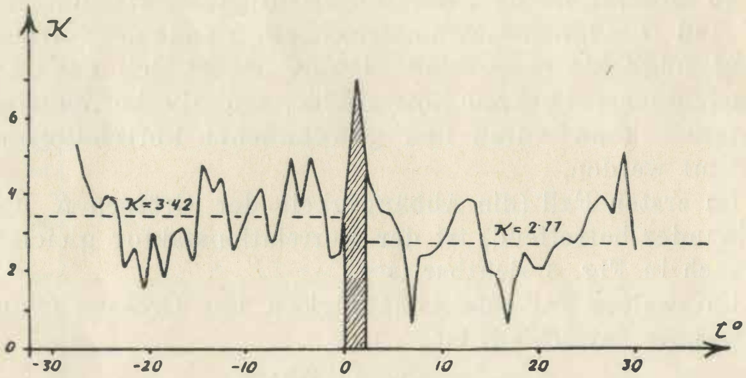

Fig. 7.

Es ist zu ersehen, dass K im grossen und ganzen konstant ist (im Mittel ist $K = 3{\cdot}20 \pm 0{\cdot}16$), denn wie wir später zeigen werden, bleiben die Abweichungen vom Mittelwerte in den Grenzen des Fehlergebiets. Nur im Umkreis von 1^0 ist die Störung etwas grösser, was aber auch in der Meteorologie eine entsprechende Begründung hat. Wenn im Winter die Temperatur über null Grad steigt, bedürfen die grossen Schneemassen zum Schmelzen einer zufliessenden Wärme, und dadurch hält sich die Temperatur längere Zeit etwas über Null und entsteht bei diesem Temperaturwert eine sehr grosse Häufigkeit. Da aber solche Tauwetterperioden einen zufälligen Charakter tragen und in manchen Jahren sehr selten vorkommen, ist es selbstverständlich, dass an dieser Stelle die Störung gross sein muss. Diese Störung wäre noch viel grösser, wenn wir die Winterperiode gesondert betrachten wollten. Die Frühlings- und Herbstperioden, wo die Temperatur diese Höhe erreicht, haben eine mässige Störung, und dadurch wird die Jahresstörung vermindert.

Lassen wir die drei Klassen (0, 1 und 2) fort und berechnen wir die mittlere Störung für negative und positive Temperaturen, so bekommen wir im ersten Fall:

$$K_{-} = 3{\cdot}42 \pm 0{\cdot}27$$

und im zweiten Fall:

$$K_{+} = 2{\cdot}77 \pm 0{\cdot}19.$$

Die Differenz ist

$$K_{-} - K_{+} = 0{\cdot}65 \pm 0{\cdot}33.$$

Das Ergebnis sagt uns, dass die Störung bei niedrigen Temperaturen grösser als bei hohen ist, obwohl das Übergewicht ganz gering ist. Aber aus der Meteorologie ist es bekannt, dass in unserem Klima die Temperatur (auch der Luftdruck) im Sommer eine grössere Stabilität aufweist als im Winter, und daraufhin kann man den erhaltenen Resultaten glauben.

Im Umkreis von 1^0 gibt der Charliersche Störungskoeffizient aber keine ausserordentliche Störung ($C = 20$ bis 30%) an, während er bei höheren und niedrigeren Temperaturen eine viel grössere Störung liefert ($C =$ über 100%).

3. Im folgenden wollen wir betrachten, in welchem Masse die Abweichungen der Grösse K vom arithmetischen Mittel (Tab. 3 und Fig. 7) als zufällig und in welchem Masse sie als klimatologisch charakteristisch aufgefasst werden können. Dazu müssen wir für K den ihm entsprechenden Fehler berechnen und dann feststellen, ob die Abweichungen durch den Fehler hervorgerufen werden oder nicht. Im letzteren Fall könnte man sagen, dass die Störung in der Variationsbreite nicht ganz konstant sei.

Nach (7) und (44) ist

$$K = \frac{\sqrt{\sigma_L^2 - \sigma_B^2}}{s\sqrt{p\,(1-p)}} = \tag{101}$$

$$= \sqrt{\frac{\sigma_L^2 - \sigma_B^2}{s\,\sigma_B^2}} =$$

$$= \sqrt{\frac{L^2-1}{s}}\,.$$

Nach der Fehlertheorie können wir schreiben:

$$\sigma K = \sqrt{\left(\frac{dK}{dL}\cdot\sigma L\right)^2} = \tag{102}$$

$$= \frac{L}{\sqrt{s\,(L^2-1)}}\cdot\sigma L,$$

wo σK und σL die quadratischen Abweichungen von K bzw. L sind.

Weiter können wir schreiben:

$$\frac{\sigma K}{K} = \frac{L}{L^2-1}\cdot\sigma L, \tag{103}$$

wobei die Formel (101) benutzt worden ist.

Aus letzterem erhalten wir:

$$\frac{\sigma K}{K} = \frac{L^2}{L^2-1}\cdot\frac{\sigma L}{L}\,. \tag{104}$$

Hieraus ersehen wir, dass der Variationsfaktor von K nur etwas grösser als der von L ist. Bei grossen L-Werten werden beide gleich gross. In unserem Beispiel, wo $L \approx 3$ ist, ist

$$\frac{L^2}{L^2-1} = 1{\cdot}12,$$

d. h. der Variationsfaktor von K ist ungefähr um 12% grösser als derjenige von L. Bei $L = 10$ beträgt der Unterschied nur 1%.

Aus (102) sieht man, dass wir zum Bestimmen von σK σL kennen müssen, oder auch zum Berechnen von $\frac{\sigma K}{K}$ nach (104) — die Grösse $\frac{\sigma L}{L}$ [1].

[1]) Weiter unten wird gezeigt werden, dass gerade diese Grösse $\left(\frac{\sigma L}{L}\right)$ für alle Klassen fast konstant ist.

Zum Bestimmen von σL nehmen wir den bekannten Ausdruck

$$L = \frac{\sigma_L}{\sigma_B}$$

zu Hilfe.

Wir bezeichnen die quadratischen Abweichungen der Grössen L, σ_L und σ_B bzw. σL, $\sigma\sigma_L$ und $\sigma\sigma_B$ und können dann nach der Fehlertheorie schreiben:

$$(105) \qquad \sigma L = \sqrt{\left(\frac{dL}{d\sigma_L} \cdot \sigma\sigma_L\right)^2 + \left(\frac{dL}{d\sigma_B} \cdot \sigma\sigma_B\right)^2} = \\ = \sqrt{\left(\frac{\sigma\sigma_L}{\sigma_B}\right)^2 + \left(\frac{\sigma_L \ \sigma\sigma_B}{\sigma_B^2}\right)^2}.$$

Multiplizieren wir jetzt die beiden Seiten mit

$$\frac{1}{L} = \frac{\sigma_B}{\sigma_L},$$

so erhalten wir:

$$(106) \qquad \frac{\sigma L}{L} = \sqrt{\left(\frac{\sigma\sigma_L}{\sigma_L}\right)^2 + \left(\frac{\sigma\sigma_B}{\sigma_B}\right)^2}.$$

Damit ist der Variationsfaktor von L durch die Variationsfaktoren σ_L und σ_B gegeben.

a) Aus der Wahrscheinlichkeitstheorie ist bekannt, dass

$$(107) \qquad \frac{\sigma\sigma_L}{\sigma_L} = \frac{1}{\sqrt{2n}}$$

ist, wo n die Anzahl der Häufigkeitszahlen, aus welchen σ_L berechnet worden ist, bedeutet.

b) Ebenso wäre

$$\frac{\sigma\sigma_B}{\sigma_B} = \frac{1}{\sqrt{2n}},$$

falls σ_B aus den Häufigkeitszahlen der Bernoullischen Reihe berechnet wäre. Da wir es aber nur mit der Lexisschen Reihe zu tun haben, müssen wir σ_B nach der mittleren Häufigkeitszahl theoretisch bestimmen, und daher konnte auch ihr Variationsfaktor oben nicht gegeben werden. Wegen des Gesagten wird aber die Genauigkeit nicht schlechter, denn weiter werden wir sehen, dass das theoretisch berechnete σ_B meist genauer ist.

Nach der Bernoullischen Formel ist

$$\sigma_B = \sqrt{s(p - p^2)} = \tag{108}$$

$$= \sqrt{m - \frac{m^2}{s}},$$

wo $m = sp$ ist.

Nach der Fehlertheorie ist

$$\sigma_{\sigma_B} = \sqrt{\left(\frac{d\,\sigma_B}{dm} \cdot \sigma m\right)^2} = \tag{109}$$

$$= \frac{\left|1 - \frac{2m}{s}\right|}{2\sqrt{m - \frac{m^2}{s}}}\,\sigma m =$$

$$= \frac{|1 - 2p|}{2\,\sigma_B}\,\sigma m\,.$$

Aus der Wahrscheinlichkeitstheorie ist bekannt, dass

$$\sigma m = \frac{\sigma_L}{\sqrt{n}} \tag{110}$$

ist, wo n dieselbe Bedeutung wie in Formel (107) hat.

Setzen wir diesen Ausdruck in (109) ein, so ist

$$\sigma_{\sigma_B} = \frac{|1 - 2p|\,\sigma_L}{2\sqrt{n}\,\sigma_B} = \tag{111}$$

$$= \frac{|1 - 2p|\,L}{2\sqrt{n}}\ [1].$$

[1]) Im gegebenen Fall ist σ_B mit Hilfe der mittleren Häufigkeitszahl m bestimmt, welche ihrerseits aus der Lexisschen Reihe berechnet worden ist. Wenn aber die empirisch gegebene Reihe eine Bernoullische ist, dann ist $L = 1$, und wir bekommen aus (111):

$$\sigma_{\sigma_B} = \frac{|1 - 2p|}{2\sqrt{n}}\,.$$

Daraus ersieht man:

1. σ_{σ_B} ist von der Grösse s unabhängig, obwohl σ_B selbst von s abhängt. Dieses kommt daher, dass bei grossen s-Werten die Häufigkeitszahl m relativ

Damit ist

$$\frac{\sigma_{\sigma_B}}{\sigma_B} = \frac{|1 - 2p|\, L}{2\sqrt{ns\,(p - p^2)}}. \tag{112}$$

Setzen wir jetzt (107) und (112) in (106) ein, so erhalten wir:

$$\frac{\sigma L}{L} = \sqrt{\frac{1}{2n} + \frac{(1-2p)^2 L^2}{4ns\,(p-p^2)}} = \tag{113}$$

$$= \sqrt{\frac{1}{2n}\left\{1 + \frac{(1-2p)^2 L^2}{2m\,(1-p)}\right\}}.$$

Statt

$$\frac{(1-2p)^2}{1-p}$$

können wir schreiben:

$$\frac{(1-2p)^2}{1-p} = 1 - 3p,$$

weil p für alle Klassen kleiner als 0·05 ist. Der durch die Vereinfachung entstandene Fehler kann höchstens 0·4‰ betra-

---

genauer bestimmt worden ist, und dadurch auch p. Deshalb ist bei grossen s-Werten auch σ_B relativ genauer.

Das Gesagte sei durch ein Beispiel illustriert.

I. Es sei

$$s = 100$$
$$n = 8$$
$$m = 30, \quad \text{damit ist}$$
$$p = 0{\cdot}30$$
$$\sigma_B = \sqrt{100 \,.\, 0{\cdot}30 \,.\, 0{\cdot}70} = 4{\cdot}58$$
$$\sigma m = \frac{4{\cdot}58}{\sqrt{8}} = 1{\cdot}62.$$

Nehmen wir an, m habe eine Abweichung, deren Grösse σm ist. Dann ist

$$m = 31{\cdot}6$$
$$p = 0{\cdot}316$$
$$\sigma_B = \sqrt{100 \,.\, 0{\cdot}316 \,.\, 0{\cdot}684} = 4{\cdot}65.$$

Hier sehen wir, dass σ_B grösser geworden ist. Die entsprechende Abweichung ist

$$\Delta\sigma_B = 4{\cdot}65 - 4{\cdot}58 = 0{\cdot}07.$$

gen und braucht nicht in Betracht gezogen zu werden, da das zweite Glied in der Klammer selbst sehr klein ist.

Nun bekommen wir:

$$\frac{\sigma L}{L} = \sqrt{\frac{1}{2n}\left\{1 + \frac{(1-3p)L^2}{2m}\right\}} = \tag{114}$$

$$= \sqrt{\frac{1}{2n}\left\{1 + \frac{L^2}{2}\left(\frac{1}{m} - \frac{3}{s}\right)\right\}} =$$

$$= 0{\cdot}250\sqrt{1 + \frac{L^2}{2}\left(\frac{1}{m} - 0{\cdot}000\,342\right)},$$

weil $n = 8$ und $s = 8766$ ist.

Wenn m wächst, dann

$$\frac{\sigma L}{L} \to 0{\cdot}250,$$

was auch selbstverständlich ist, denn bei grossem m kommt

II. Es sei

$$s = 1000$$
$$n = 8$$
$$m = 300, \quad \text{damit ist}$$
$$p = 0{\cdot}30$$
$$\sigma_B = \sqrt{1000\,.\,0{\cdot}30\,.\,0{\cdot}70} = 14{\cdot}50$$
$$\sigma m = \frac{14{\cdot}50}{\sqrt{8}} = 5{\cdot}12.$$

Nehmen wir jetzt analog für m die Abweichung σm an. Dann ist

$$m = 305{\cdot}1 \quad \text{und}$$
$$p = 0{\cdot}3051$$
$$\sigma_B = \sqrt{1000\,.\,0{\cdot}3051\,.\,0{\cdot}6949} = 14{\cdot}57.$$

Die entsprechende Abweichung von σ_B ist

$$\Delta\sigma_B = 14{\cdot}57 - 14{\cdot}50 = 0{\cdot}07.$$

Nach der Formel ist

$$\sigma_{\sigma_B} = \frac{1 - 0{\cdot}6}{2\sqrt{8}} = 0{\cdot}071.$$

2. Wenn $p = \frac{1}{2}$ ist, dann ist

$$\sigma_{\sigma_B} = 0,$$

d. h. in der Umgebung von $p = \frac{1}{2}$ ist das theoretisch bestimmte σ_B sehr genau.

$\frac{\sigma_{\sigma_B}}{\sigma_B}$ nicht mehr in Betracht und in der Formel (106) bleibt nur

$$\frac{\sigma L}{L} = \frac{\sigma_{\sigma_L}}{\sigma_L} = \frac{1}{2\sqrt{n}} = 0{\cdot}250.$$

Aber eine einfache Berechnung zeigt auch, dass bei den in Betracht kommenden m-Werten die Grösse $\frac{\sigma L}{L}$ nicht viel von 0·250 abweicht.

Bei	$m = 400$	ist	$\frac{\sigma L}{L} = 0{\cdot}2503$
„	$m = 100$	„	$\frac{\sigma L}{L} = 0{\cdot}252$
„	$m = 50$	„	$\frac{\sigma L}{L} = 0{\cdot}254$
„	$m = 10$	„	$\frac{\sigma L}{L} = 0{\cdot}268.$

---

3. Die maximale Grösse σ_{σ_B} (wenn $p \rightarrow 0$ oder $p \rightarrow 1$) ist

$$\sigma_{\sigma_{B_{max}}} = \frac{1}{2\sqrt{n}}.$$

Da die letztere Grösse bei jedem n kleiner als $\frac{1}{2}$ ist, zeigt dies die grosse Genauigkeit der Bernoullischen Formel. Wie im Folgenden gezeigt wird, ist die Genauigkeit bei der Bestimmung von σ_B nach der Bernoullischen Formel fast immer grösser.

Für das aus empirischen Häufigkeitszahlen berechnete σ_B ist

$$o\sigma_B = \frac{\sigma_B}{\sqrt{2n}} =$$

$$= \sqrt{\frac{s\,(p - p^2)}{2\,n}}.$$

Damit das nach der Bernoullischen Formel berechnete σ_B genauer wäre, muss

$$\frac{|1 - 2\,p\,|}{2\sqrt{n}} < \sqrt{\frac{s\,(p - p^2)}{2\,n}}$$

sein, wo

$$s > \frac{(1 - 2\,p)^2}{2\,(p - p^2)} \quad \text{ist.}$$

Folgende Tabelle zeigt die entsprechenden s-Werte für einige p-Werte.

p	0·2	0·1	0·01	0·001	0·0001	0·00001
s	2	5	50	500	5000	50000

Der mittlere Wert von $\frac{\sigma L}{L}$ für alle Klassen beträgt

$$\frac{\sigma L}{L} = 0{\cdot}256. \tag{115}$$

Setzen wir diesen Wert in (104) ein, so ist

$$\frac{\sigma K}{K} = 0{\cdot}256 \frac{L^2}{L^2 - 1}. \tag{116}$$

Nach (101) ist

$$L^2 - 1 = sK^2 \tag{117}$$

und

$$L^2 = 1 + sK^2. \tag{118}$$

Setzen wir jetzt letzteren Ausdruck in (116) ein, so ist

$$\frac{\sigma K}{K} = 0{\cdot}256 \left(\frac{1}{sK^2} + 1\right). \tag{119}$$

Wenn wir voraussetzen, dass K konstant ist, ermöglicht die Gleichung (119) die mittlere Abweichung für K zu bestimmen.

Eine Betrachtung der Tabelle zeigt, dass bei sog. „kleinen Zahlen", wo $p \approx 0$ ist, man doch genauer prüfen müsste, welche Berechnungsart sich für σ_B bessereignet — die auf der Bernoullischen Formel beruhende oder jene nach den Häufigkeitszahlen. Um ein zweckmässiges Kriterium zu finden, das zeigen würde, welche Methode im gegebenen Fall die bequemere sei, multiplizieren wir die beide Seiten der letzten Formel mit p. Dann bekommen wir:

$$sp > \frac{(1 - 2p)^2}{2(1 - p)}.$$

Da immer

$$1 - 2p < 1 - p$$

und

$$(1 - 2p)^2 < 1 - 2p,$$

so ist

$$(1 - 2p)^2 < 1 - p.$$

Daraus ist zu ersehen, dass die obengegebene Bedingung dann erfüllt ist, wenn

$$sp > \tfrac{1}{2}$$

ist, oder

$$m > \tfrac{1}{2}$$

ist.

Dadurch ist erklärt, dass wenn die mittlere Häufigkeitszahl grösser als $\frac{1}{2}$ ist, auch das mit Hilfe der Bernoullischen Formel berechnete σ_B genauer ist als das aus den Häufigkeitszahlen erhaltene. Bei praktisch vorkommenden statistischen Reihen ist diese Forderung fast immer erfüllt.

Nach unseren Daten ist durchschnittlich

$$K = 3{\cdot}20\% \pm 0{\cdot}16\%.$$

Damit ist

(120) $$\frac{\sigma K}{K} = 0{\cdot}284 \pm 0{\cdot}002,$$

wo

(121) $$\sigma K = 0{\cdot}91\% \pm 0{\cdot}04\%$$

ist.

Aus den gegebenen K-Werten berechnen wir, dass

(122) $$\sigma K = 1{\cdot}20\%$$

ist.

Da das letzte σK aus 58 Zahlen berechnet worden ist, ist sein Fehler

(123) $$\sigma_{\sigma K} = 0{\cdot}11\%.$$

Die Differenz zwischen den beiden σK beträgt $0{\cdot}29\% \pm 0{\cdot}12\%$, ist also fast $2\frac{1}{2}$ mal grösser als der mittlere Fehler von σK. Darum ist die Wahrscheinlichkeit, dass σK zufällig so gross ist, ganz gering ($\approx 1\%$), und es ist richtiger anzunehmen, dass auf der Temperaturskala alle Stellen nicht gleich gestört sind (z. B. der Umkreis von 1^0 und der Unterschied zwischen höheren und niedrigeren Temperaturen).

4. Die angeführten Berechnungen verlangen noch eine Korrektur. Das zur Bestimmung der Störungsstreuung verwendete σ_B ist unter der Voraussetzung berechnet, dass die Wahrscheinlichkeit für jede Temperaturklasse konstant ist. Wie bekannt, ist jedoch das Auftreten der Temperaturen während eines Jahres sehr veränderlich. So kommen im Sommer niedrige Temperaturen (unter 0^0) fast nie vor, und ebenso im Winter fast nie hohe Temperaturen (über $+10^0$). Damit haben wir es innerhalb eines „Versuches“ mit einer veränderlichen Wahrscheinlichkeit zu tun, und die Streuung muss nicht nach der Bernoullischen, sondern nach der Poissonschen Formel berechnet werden.

Die Streuung der Poissonschen Reihe ist nach der bekannten Formel folgende:

(124) $$\sigma_P^2 = \sigma_B^2 - \Sigma (p_i - p_0)^2.$$

Hier bedeuten: σ_P — das Streuungsmass der Poissonschen Reihe, σ_B — dasselbe Mass der Bernoullischen Reihe, p_0 — die mittlere Wahrscheinlichkeit für das ganze Jahr und p_i — die Wahrscheinlichkeit für jede einzelne Beobachtung.

Um zu schätzen, wie gross der Unterschied zwischen σ_P und σ_B ist, schreiben wir die letzte Formel in folgender Form:

(125) $$\sigma_P^2 = sp_0 q_0 - s\Sigma \frac{(p_i - p_0)^2}{s} = $$
$$= s p_0 q_0 - s\,\sigma_*^2 = $$
$$= s p_0 q_0 \left(1 - \frac{\sigma_*^2}{p_0 q_0}\right),$$

wo

$$\sigma_* = \sqrt{\frac{(p_i - p_0)^2}{s}}\,^{1)} \quad \text{und} \quad q_0 = 1 - p_0$$

ist.

Weil nach den Beobachtungsdaten q_0 immer grösser als 0·95 ist (durchschnittlich 0·98), so können wir in der Formel (125) $q_0 = 1$ nehmen, und wir erhalten dann:

(126) $$\sigma_P^2 = \sigma_B^2 \left(1 - \frac{\sigma_*^2}{p_0}\right).$$

Bezeichnen wir die Grösse

$$\frac{\sigma_*^2}{p_0} = 2\varkappa$$

und berechnen sie für jede Temperaturklasse, so sehen wir, dass $2\varkappa$ immer kleiner als 0·05 ist (durchschnittlich 0·024). Darum können wir auf Grund von (126) schreiben:

(127) $$\sigma_P = \sigma_B (1 - \varkappa).$$

Die folgende Tabelle (Tab. 4) gibt für jede Temperaturklasse die entsprechenden p_0-, σ_*- und $\varkappa$-Werte in Prozenten.

---

1) Die Berechnung von σ_* ist dadurch vereinfacht worden, dass die Wahrscheinlichkeit p_i für jeden Monat als konstant angenommen worden ist. Dadurch ist

$$\sigma_*^2 = \sum_{i=1}^{12} \frac{(p_i - p_0)^2}{12},$$

wo p_1 die entsprechende Wahrscheinlichkeit für den Januar, p_2 für den Februar u. s. w. ist.

Tab. 4.

t	p_0	σ_*	$\varkappa$	t	p_0	σ_*	$\varkappa$
28	0·12	0·36	0·5	0	4·94	4·04	1·7
27	0·16	0·39	0·5	— 0	3·43	2·80	1·2
26	0·32	0·62	0·6	— 1	2·94	2·51	1·1
25	0·41	0·68	0·6	— 2	2·90	2·77	1·3
24	0·53	0·89	0·8	— 3	2·53	2·61	1·4
23	0·70	1·08	0·8	— 4	2·37	2·62	1·4
22	0·91	1·49	1·2	— 5	2·24	2·48	1·4
21	0·98	1·58	1·3	— 6	1·96	2·22	1·3
20	1·28	1·86	1·4	— 7	1·64	2·00	1·2
19	1·56	2·22	1·6	— 8	1·53	2·00	1·3
18	2·07	2·79	1·9	— 9	1·53	2·05	1·4
17	2·42	3·38	2·4	—10	1·27	1·69	1·1
16	3 02	3·88	2·5	—11	1·07	1·63	1·2
15	3·17	3·86	2·4	—12	1·06	1·58	1·2
14	3·42	3·90	2·2	—13	0·86	1·25	0·9
13	3·63	3·98	2·2	—14	0·81	1·27	1·0
12	3·48	3·75	2·0	—15	0·58	0·97	0·8
11	3·47	3·42	1·7	—16	0·53	0·91	0·8
10	3·17	3·22	1·6	—17	0·41	0·58	0·4
9	3·25	3·32	1·7	—18	0·29	0·47	0·4
8	3·27	3·54	1·9	—19	0·25	0·47	0·4
7	2·79	3·85	2·7	—20	0·14	0·26	0·2
6	3·07	3·02	1·5	—21	0·12	0·22	0·2
5	3·11	3·05	1·5	—22	0·10	0·21	0·2
4	3·04	2·86	1·4	—23	0·12	0·26	0·3
3	3·32	3·09	1·4	—24	0·08	0·19	0·2
2	3·63	3·38	1·6	—25	0·07	0·17	0·2
1	5·05	4·29	1·8				

Aus der Tabelle ersehen wir, dass σ_P durchschnittlich nur um 1·0% von σ_B abweicht (maximal um 2·7%).

Nach (111) ist

$$\sigma_{\sigma_B} = \frac{|1-2p|\,L}{2\sqrt{n}}.$$

Weil $n = 8$ ist und durchschnittlich $p = 0{\cdot}02$ und $L = 3{\cdot}3$ sind, so beträgt durchschnittlich

$$\sigma_{\sigma_B} = \frac{0{\cdot}96\,.\,3{\cdot}3}{2\sqrt{8}} = 0{\cdot}56.$$

Da die Grösse σ_B zwischen 2·0 und 20·9 schwankt (die durchschnittliche Grösse von σ_B ist 11·0), verändert sich der relative Fehler von σ_B in den Grenzen von 2·7 % bis 28 % und hat die durchschnittliche Grösse von 5·1 %. Im Vergleich zu dieser Zahl ist die Differenz zwischen σ_B und σ_P viele Male kleiner, und wir können ruhig statt σ_P σ_B schreiben.

Bei anderen meteorologischen Elementen (Windgeschwindigkeit, relative Feuchtigkeit und Luftdruck) ist diese Differenz noch kleiner, denn jene weisen in ihrem jährlichen Verlaufe eine viel grössere Stetigkeit auf als die Temperatur.

5. Im folgenden wollen wir die Grösse des Störungskoeffizienten beim Luftdruck (P_{mb}) betrachten. Die folgende Tabelle (Tab. 5) gibt analog jener für die Temperatur die notwendigen Grössen.

Tab. 5.

P_{mb}	m	σ_L	σ_B	δ	$K\%$	$C\%$
40—43	3	7	1·8	7	4·1	212
36—39	18	21	4·2	21	5·3	117
32—35	40	32	6·3	31	5·3	78
28—31	90	55	9·4	54	6·2	60
24—27	178	99	13·2	97	7·8	54
20—23	341	75	18·1	73	4·3	22
16—19	472	106	21·1	103	5·2	22
12—15	708	76	25·5	71	3·0	10
08—11	1060	140	30·5	137	4·8	13
04—07	1342	199	33·7	196	6·2	15
00—03	1303	170	33·3	167	5·4	13
96—99	1144	141	31·5	138	4·7	12
92—95	831	44	27·4	33	1.3	4
88—91	546	100	22·6	97	4·6	18
84—87	291	50	16·8	47	3·0	16
80—83	174	64	13·1	63	5·1	36
76—79	108	26	10·3	24	2·4	22
72—75	64	41	8·0	40	5·4	62
68—71	23	18	4·8	18	4·0	78
64—67	8	8	2·8	8	3·0	100
60—63	2	5	1·6	5	3·5	250
56—59	5	10	2·2	10	4·7	200
52—55	2	5	1·3	4	3·7	200

Aus dieser Tabelle ist ebenfalls ersichtlich, dass K mehr oder weniger konstant ist, C aber in einer gewissen Beziehung zu der Grösse m steht. Das Gesagte wird noch deutlicher, wenn wir die Fig. 8 betrachten.

Zur Erklärung der sehr grossen Korrelationskoeffizienten ($|r| > 0{\cdot}9$) zwischen m und C kann man keine physikalischen Ursachen angeben.

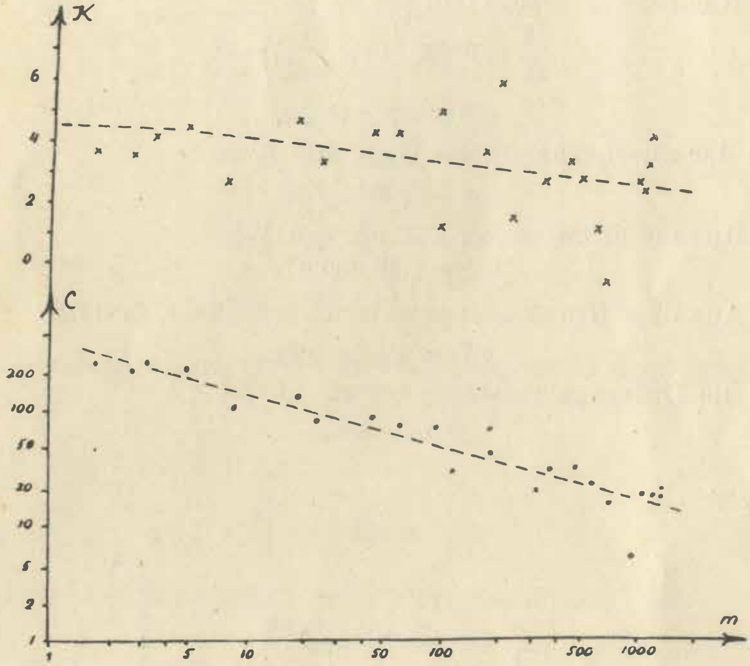

Fig. 8.

In Fig. 9 ist die Abhängigkeit zwischen Luftdruck und K dargestellt worden.

Aus der Fig. 9 ersieht man, dass zwischen K und dem Luftdruck ein geringer positiv-korrelativer Zusammenhang besteht; die Abweichungen der Grössen K von ihrem mittleren Werte sind aber so gross, dass dieser Zusammenhang nur scheinbar sein kann. Um zu beurteilen, ob die genannten Abweichungen zufällige sind oder nicht, berechnen wir, wie auch bei der Temperatur, die Grösse σK und vergleichen sie mit dem aus dem Beobachtungsmaterial berechneten σK.

Berechnen wir nach der Formel (114) für jede Klasse $\frac{\sigma L}{L}$ und deren arithmetisches Mittel, so erhalten wir für das letztere:

$$\frac{\sigma L}{L} = 0{\cdot}257.$$

Nach der Formel (119) ist

$$\frac{\sigma K}{K} = 0{\cdot}257 \left(1 + \frac{1}{s K^2}\right) =$$

$$= 0{\cdot}272 \pm 0{\cdot}002,$$

denn der durchschnittliche Wert von K ist

$$K = 4{\cdot}48 \pm 0{\cdot}29.$$

Hieraus erhalten wir für σK den Wert

$$\sigma K = 1{\cdot}22 \pm 0{\cdot}07.$$

Aus dem Beobachtungsmaterial berechnet, beträgt

$$\sigma K = 1{\cdot}40 \pm 0{\cdot}21.$$

Die Differenz zwischen beiden σK ist

$$0{\cdot}18 \pm 0{\cdot}22.$$

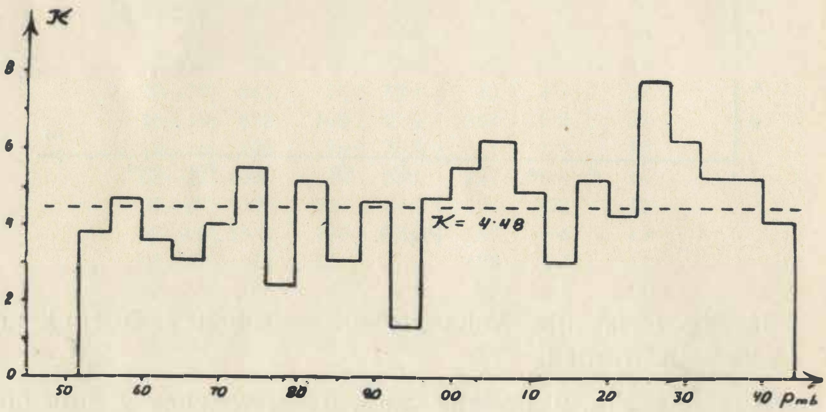

Fig. 9.

Darum kann man sagen, dass die Störung für jede Luftdruckklasse konstant ist, was auch einer intuitiven Kalkulation gemäss annehmbar ist.

6. Ebenso steht es mit der Windgeschwindigkeit. Die folgende Tabelle (Tab. 6) bietet analog den früheren Tabellen

die entsprechenden Daten für jede Geschwindigkeitsklasse (v m/sec).

Tab. 6.

v m/sec	m	σ_L	σ_B	δ	$K\%$	$C\%$
0	143	34	11·9	32	2·9	22·4
1	1011	103	29·9	98	3·5	9·7
2	2519	151	42·3	145	3·7	5·8
3	1839	117	38·2	111	3·1	6·0
4	1305	105	33·3	100	3·2	7·7
5	901	79	28·4	74	2·8	8·2
6	496	73	21·6	70	3·5	14·1
7	303	62	17·1	60	3·8	19·8
8	150	23	12·1	20	1·8	13·3
9	57	13	7·5	10	1·4	17·5
10	23	7	4·8	5	1·1	21·7

Wir kommen zu dem Schluss, dass durchschnittlich

$$\frac{\sigma L}{L} = 0{\cdot}253$$

und

$$\frac{\sigma K}{K} = 0{\cdot}290 \pm 0{\cdot}007$$

ist, denn der durchschnittliche Wert von K ist

$$K = 2{\cdot}80 \pm 0{\cdot}28.$$

Hieraus erhalten wir:

$$\sigma K = 0{\cdot}81 \pm 0{\cdot}06.$$

Aus den Beobachtungsdaten berechnet, ist

$$\sigma K = 0{\cdot}93 \pm 0{\cdot}20.$$

Die Differenz zwischen den beiden σK beträgt hiernach

$$0{\cdot}12 \pm 21.$$

Die Berechnungsresultate zeigen, dass die Störung für alle Geschwindigkeitsklassen als konstant angenommen werden kann.

7. Zuletzt wollen wir die Störung der relativen Feuchtigkeit betrachten. Analog dem Vorhergehenden bietet die

folgende Tabelle (Tab. 7.) für jede Feuchtigkeitsklasse (R %) die entsprechenden m-, σ_L-, σ_B-, δ-, K- und C-Werte.

Tab. 7.

R%	m	σ_L	σ_B	δ	K%	C%
100	221	244	14·7	244	17·8	110
99	145	96	12·0	95	8·6	66
98	226	88	14·8	87	6·3	38
97	273	67	16·2	65	4·3	24
96	317	69	17·5	67	4·1	21
95	353	63	18·4	60	3·5	17
94	375	48	19·0	44	2·5	12
93	371	83	18·9	81	4·6	22
92	342	55	18·1	52	3·1	15
91	334	62	17·9	59	3·5	18
90	353	67	18·4	65	3·8	18
89	297	36	16·9	32	2·1	11
88	284	20	16·6	11	0·7	4
87	273	28	16·3	23	1·5	8
86	268	48	16·1	45	3·0	17
85	237	49	15·2	46	3·3	19
84	233	61	15·0	59	4·2	25
83	203	40	14·1	37	2·7	18
82	197	23	13·9	18	1·4	9
81	181	18	13·3	12	1·0	7
80	180	27	13·2	24	1.9	13
79	154	29	12·3	26	2·3	17
78	153	31	12·3	28	2·5	18
77	146	28	12·0	25	2·2	17
76	135	23	11·5	20	1·9	15
75	134	15	11·5	10	0·9	8
74	130	22	11·3	19	1·8	15
73	124	22	11·0	19	1·7	15
72	122	18	11·0	14	1·4	12
71	100	11	9·9	5	0·5	5
70	108	18	10·3	15	1·6	14
69	96	22	9·7	20	2·2	21
68	94	14	9·6	10	1·1	11
67	90	20	9·4	18	2·1	20
66	84	20	9·1	18	2·1	21
65	81	13	8·9	10	1·2	12
64	73	15	8·5	12	1·5	16
63	76	14	8·7	11	1·4	14
62	72	16	8·5	14	1·8	20
61	65	16	8·0	14	1·9	22
60	73	16	8·5	14	1·8	19
59	66	12	8·1	9	1·2	14
58	62	16	7·8	14	1·9	23
57	65	13	8·0	10	1·3	15
56	55	12	7·4	9	1·8	16
55	59	16	7·6	14	2·0	24
54	54	15	7·3	13	1·9	24
53	46	13	6·7	11	1·7	24
52	51	9	7·1	5	0·8	10
51	44	9	6·6	6	1·0	14
50	48	13	6·9	11	1·7	23
49	44	10	6·6	8	1·3	23
48	40	9	6·3	6	1·0	15
47	38	11	6·1	9	1·6	24
46	36	8	6·0	5	0·9	14
45	34	13	5·8	12	2·2	35
44	29	7	5·4	4	0·8	14
43	27	8	5·2	6	1·3	22
42	25	8	5·0	6	1·3	24
41	24	10	4·9	9	2·0	38
40	23	11	4·8	10	2·3	44
39	21	10	4·6	9	2·1	43
38	19	8	4·3	7	1·7	37
37	17	7	4·1	6	1·6	35
36	13	8	3·6	7	2·1	56
35	13	11	3·6	10	3·0	78
34	10	5	3·2	4	1·4	40
33	7	7	2·6	6	2·4	86
32	4	5	2·0	5	2·7	125

Hier ist das Problem ein wenig komplizierter als in den vorigen Fällen. Aus der Tabelle ersieht man, dass die K-Werte nicht als konstant angenommen werden können, denn dazu ist der Unterschied zwischen einigen Werten zu gross. Daraus folgt, dass einige Feuchtigkeitsklassen mehr gestört sind als die anderen. Das Gesagte wird durch Fig. 10 veranschaulicht, in welcher der Zusammenhang zwischen der relativen Feuchtigkeit (R%) und K gegeben wird.

Aus der Figur 10 ersieht man, dass K bis zu 95—96% mehr oder weniger konstant ist. Um das zu beweisen, berech-

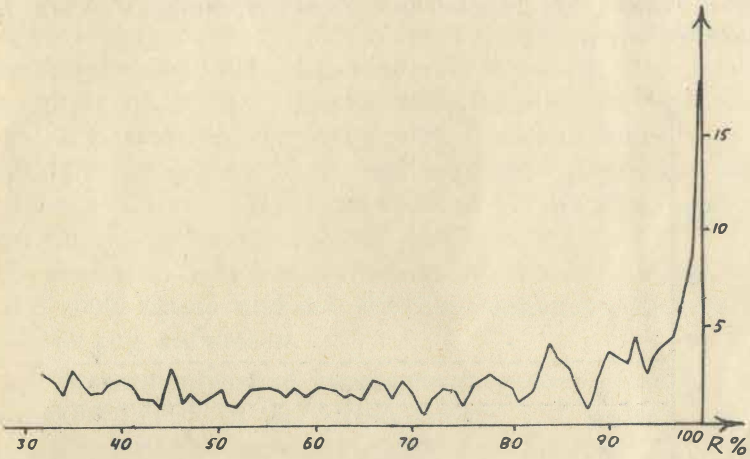

Fig. 10.

nen wir analog dem Vorhergehenden den mittleren Fehler K und vergleichen ihn mit dem aus den K-Daten berechneten.

Nach (114) erhalten wir, dass

$$\frac{\sigma L}{L} = 0{\cdot}258$$

und dass damit

$$\frac{\sigma K}{K} = 0{\cdot}337 \pm 0{\cdot}008$$

ist, denn der mittlere Wert von K ist

$$K = 1{\cdot}93 \pm 0{\cdot}10.$$

Hieraus erhalten wir:

$$\sigma K = 0{\cdot}65 \pm 0{\cdot}02.$$

Aus den Beobachtungsdaten berechnet, ist

$$\sigma K = 0{\cdot}85 \pm 0{\cdot}07.$$

Die Differenz zwischen den beiden σK beträgt

$$0{\cdot}20 \pm 0{\cdot}07.$$

Auf Grund dieser Resultate kann man sagen, dass bei relativen Feuchtigkeiten bis zu 95% (incl.) alle Klassen nicht gleich gestört sind; aber die Abweichung von einem konstanten Wert ist ganz gering. Dies ersieht man auch aus der folgenden Figur (Fig. 11), wo der Zusammenhang zwischen K und m gezeigt wird.

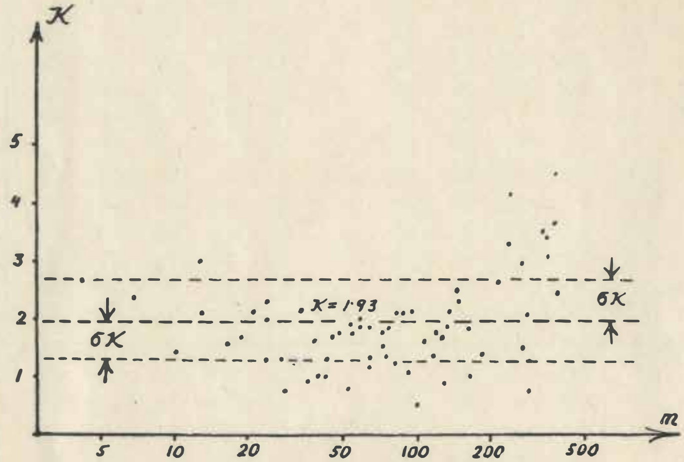

Fig. 11.

Warum die grossen relativen Feuchtigkeiten mehr gestört sind als die kleinen, besonders diejenigen in der Nähe von 100%, kann durch mehrere Ursachen erklärt werden. Die Hauptursache ist die, dass diese Feuchtigkeiten in der Nähe vom Kondensationspunkte des Wasserdampfes liegen. Wegen der klimatologischen Unterschiede zwischen den verschiedenen Jahren können die Feuchtigkeiten in einem Jahre 100% nicht erreichen, in anderen aber sie weit übersteigen. Da jedoch aus physikalischen Gründen (die Kondensation) dieses letztere nicht möglich ist, so lassen sich alle Klassen mit relativer Feuchtigkeit über 100% zu einer Klasse ($R = 100\%$) zusammenfassen, und dadurch wird auch ein hoher Wert der Störung erreicht. Zweitens ist bei ein und derselben Temperatur die maximale Spannkraft des Wasserdampfes nicht immer konstant, und deshalb müsste man auch relative Feuchtigkeiten über 100% einführen. Das wird aber nicht getan, sondern man nimmt die bei der Kondensation wirklich vorhandene Spannkraft als 100% und die anderen proportional an. Dadurch werden die Klassen verschoben, und zwar desto mehr, je näher sie zu 100% stehen, was die Streuung ebenfalls vergrössert. Drittens werden die Beobachtungsdaten der relativen Feuchtigkeit dem Hygrogramm entnommen, wobei fünfmal täglich mit Hilfe des Assmannschen Aspirationspsychrometers Korrektionen angebracht werden. Für die dazwischenliegenden Stunden, in welchen die relative Feuchtigkeit nach dem Hygrogramm wegen der sich immer ändernden Korrektion über 100% steigt, nimmt man die grösste Feuchtigkeit gleich 100% an und korrigiert dementsprechend die anderen Daten. Das verschiebt ebenfalls die Lage der Klassen und vergrössert dadurch die Streuung.

Aus diesen Gründen ist schon im Voraus zu erwarten, dass die 100%-Klasse und die ihr nahestehenden Klassen eine grosse Störung aufweisen.

8. Zusammenfassend kann gesagt werden, dass der neue Störungskoeffizient in der Klimatologie gut anwendbar ist, wobei alle mit Hilfe desselben gezogenen Schlüsse sich physikalisch begründen lassen.