Eine Wahrscheinlichkeitsaufgabe in der Pflanzensoziologie.

Von

G. PÓLYA (Zürich).

(Als Manuskript eingegangen am 14. Juli 1930.)

1. Problemstellung. In seinen grundlegenden statistisch-floristischen Untersuchungen hat P. JACCARD den Begriff des generischen Koeffizienten (coefficient générique) eingeführt[1]): Sind in einem Gebiet insgesamt s Spezies vorhanden, die g Genera vertreten, so definiert JACCARD $\frac{100\,g}{s}$ als den generischen Koeffizienten des betreffenden Gebietes.

In einer vor kurzem erschienenen interessanten Mitteilung[2]) hat A. MAILLEFER einen sehr beachtenswerten Ansatz gemacht, den generischen Koeffizienten wahrscheinlichkeitstheoretisch vorauszuberechnen. Der MAILLEFER'sche Ansatz wird, so scheint es mir, am besten verständlich, wenn man die Aufgabe folgendermassen stellt:

Es sei genau bekannt und gegeben, wie viele Arten in einem grösseren Grundgebiet vorhanden sind und auch, wie diese Arten sich auf die vertretenen Genera verteilen; es sei ferner bekannt, wie viele Arten in einem Teilgebiet des Grundgebietes vorhanden sind. Es wird verlangt, auf Grund dieser Daten die Anzahl der Genera im besagten Teilgebiet (mindestens näherungsweise) zu berechnen.

Führen wir Bezeichnungen ein. Es sei

S die Anzahl der Spezies im Grundgebiet,
G „ „ „ Genera „ „
s „ „ „ Spezies im Teilgebiet
g „ „ „ Genera „ „

[1]) Vgl. die zusammenfassende Darstellung P. JACCARD, Die statistisch-floristische Methode als Grundlage der Pflanzensoziologie; Abderhalden's Handbuch der biolog. Arbeitsmethoden, Abt. XI, Teil 5, S. 165—202.

[2]) Le coefficient générique de P. JACCARD et sa signification, Mém. de la Soc. Vaudoise des Sciences Naturelles, vol. 3 (1929) S. 113—183.

g ist also unbekannt und gesucht, hingegen sind S, G, s bekannt, ferner ist bekannt wie viele von den G Genera nur durch eine Art im Grundgebiet vertreten sind, wie viele durch zwei, wie viele durch drei Arten usw.

Wären die s Spezies des Teilgebietes unter den S Spezies des Grundgebietes „mit Absicht" ausgewählt (etwa von einem Feind mathematischer Methoden), so wäre natürlich eine Vorausberechnung von g aussichtslos. Aber die Spezies des Teilgebietes sind durch die Natur ausgewählt worden, und mir scheint, dass folgende Arbeitshypothese nicht von vornherein zu verwerfen ist: Die Natur hat die Spezies des Teilgebietes, insofern es auf die Einteilung in Genera ankommt, zufallsartig ausgewählt. Wir präzisieren also die gestellte Aufgabe folgendermassen: Es ist zu ermitteln die wahrscheinliche oder mittlere Anzahl der Genera unter der Annahme, dass die Spezies des Teilgebietes unter denen des Grundgebietes „durch den Zufall" ausgewählt worden sind.

Wir können eine zufallsartige Auswahl selbst und zwar folgendermassen vornehmen: Wir stellen jede Art durch eine Kugel dar, wir werfen die Kugeln durcheinandergemischt in eine „Urne" (in einen Sack oder derartiges) und ziehen die verlangte Anzahl blindlings heraus. Wir wollen die Kugeln, welche Arten desselben Genus repräsentieren, mit der gleichen Farbe versehen, und die verschiedenen Genera durch verschiedene Farben unterscheiden. Wir haben so insgesamt S Kugeln, die G verschiedene Farben tragen. Wenn wir von diesen Kugeln s blindlings ziehen, so werden wir das eine Mal mehr, das andere Mal weniger verschiedene Farben an den gezogenen Kugeln finden; wenn wir diesen Versuch mit den Kugeln genügend oft wiederholen, können wir die mittlere Anzahl von verschiedenen Farben an s blindlings gezogenen Kugeln ermitteln. Diese mittlere Anzahl wird, wenn unsere Arbeitshypothese zutrifft, auch in der Natur zu beobachten sein, als die mittlere Anzahl der Genera in Teilgebieten mit s Spezies. Darum nennen wir, mit MAILLEFER, diese durch Urnenexperimente ermittelte mittlere Anzahl die wahrscheinliche Genuszahl für s Spezies; wir bezeichnen diese wahrscheinliche Genuszahl mit $\overline{g}$, oder ausführlicher, um die Abhängigkeit von der Spezieszahl s hervorzuheben, mit $\overline{g}_s$. Diese Zahl $\overline{g}_s$, die wir nicht bloss durch Urnenexperimente, sondern, wie wir später sehen werden, auch durch Wahrscheinlichkeitsrechnung bestimmen können, scheint mir eine nicht unvernünftige Lösung der eingangs gestellten Aufgabe zu sein.

MAILLEFER nahm als Grundgebiet die Schweiz an und ermittelte die wahrscheinliche Genuszahl $\overline{g}_s$ für mehrere Artanzahlen s durch Urnenexperimente. (Er experimentierte nicht mit farbigen Kugeln, sondern mit beschriebenen Zetteln, was aber offenbar auf dasselbe herauskommt.) MAILLEFER hat die wahrscheinlichen, durch Urnenexperimente ermittelten Genuszahlen mit zahlreichen wirklichen, durch floristisch-statistische Beobachtungen gewonnenen Genuszahlen verglichen, sowohl in Teilgebieten der Schweiz, wie in Gebieten ausserhalb der Schweiz. Ohne eine mir in botanischen Fragen nicht zukommende Kompetenz zu beanspruchen, darf ich wohl sagen, dass, insofern auf Grund des von MAILLEFER vereinigten Zahlenmaterials geurteilt werden kann, die Übereinstimmung mir eine recht gute zu sein scheint, sodass die wirklichen Genuszahlen durch die wahrscheinlichen Genuszahlen, mindestens in ihrem überwiegenden Teil, als erklärt zu betrachten wären.

Ich will im folgenden die noch ausstehende mathematische Aufgabe lösen, deren Lösung auch MAILLEFER als wünschenswert bezeichnet hat[3]): Die durch Urnenexperimente ermittelten Zahlen $\overline{g}_s$ wahrscheinlichkeitstheoretisch berechnen.

Ich will die genaue Lösung der mathematischen Aufgabe in Nr. 2 geben und daraus in Nr. 3 eine für kleinere Werte von s gut brauchbare Näherungsformel herleiten. Zum Schluss stelle ich in Nr. 4 einige numerische Resultate zusammen.

Es wäre möglich, die Lösung in verschiedenen Richtungen weiter zu verfolgen, aber ich habe darauf insbesondere aus dem Grunde verzichtet, weil ich als Nichtbotaniker nicht beurteilen kann, ob die Abgrenzung von Spezies und Genera denjenigen Grad von Festigkeit besitzt, der bei der Abgrenzung der Merkmale in statistischen Untersuchungen im allgemeinen wünschbar ist.

2. Lösung der wahrscheinlichkeitstheoretischen Aufgabe. Wir sind also zur folgenden Aufgabe geführt worden: Eine Urne enthält S Kugeln, die G verschiedene Farben tragen. Es seien k_1 Kugeln von der ersten, k_2 von der zweiten, ... k_G von der letzten, G-ten Farbe vorhanden, so dass

$$k_1 + k_2 + k_3 + \cdots + k_G = S . \tag{1}$$

[3]) A. a. O. S. 119. Ich werde, im völligen Einverständnis mit MAILLEFER, die Aufgabe auf die Art, die er „manière b)" nennt, auffassen.

Gesucht ist die mittlere Anzahl oder mathematische Erwartung $\bar{g}_s$ der verschiedenen Farben, die auf s der Urne zugleich entnommenen Kugeln auftreten.

Suchen wir zuerst die Wahrscheinlichkeit dafür, dass genau g Farben auf den s der Urne entnommenen Kugeln auftreten. Die Anzahl der möglichen Fälle ist offenbar $\binom{S}{s}$. Die Anzahl der günstigen Fälle, nennen wir sie A_{sg}, ist die Summe aller Produkte

$$\binom{k_1}{r_1}\binom{k_2}{r_2}\ldots\binom{k_G}{r_G}$$

von der Art, dass

$$r_1 + r_2 + \ldots + r_G = s$$
$$0 \leqq r_1 \leqq k_1\,,\quad 0 \leqq r_2 \leqq k_2\,,\ \ldots\quad 0 \leqq r_G \leqq k_G$$

und unter den Zahlen r_1, r_2, $\ldots$ r_G genau g von 0 verschieden sind. Wir erhalten A_{sg} am bequemsten als Koeffizienten in einer erzeugenden Funktion:

$$(2)\quad f(x,y) = \prod_{\nu=1}^{G}\left\{1 + \binom{k_\nu}{1} xy + \binom{k_\nu}{2} x^2 y + \ldots + \binom{k_\nu}{k_\nu} x^{k_\nu} y\right\}$$
$$= \prod_{\nu=1}^{G}\left\{1 + \left[(1+x)^{k_\nu} - 1\right] y\right\}$$
$$= \sum_s \sum_g A_{sg}\, x^s\, y^g\,.$$

Die gesuchte Wahrscheinlichkeit ist der Quotient $A_{sg}\Big/\binom{S}{s}$ und die gesuchte mittlere Anzahl der Farben

$$(3)\qquad \bar{g}_s = \binom{S}{s}^{-1} \sum_g g\, A_{sg}\,.$$

Es ist, nach (2),

$$\frac{\partial f}{\partial y} = \sum_s \sum_g g\, A_{sg}\, x^s\, y^{g-1}$$

und daher, gemäss (3),

$$(4)\qquad \left(\frac{\partial f}{\partial y}\right)_{y=1} = \sum_s \sum_g g\, A_{sg}\, x^s = \sum_s \binom{S}{s} \bar{g}_s\, x^s\,.$$

Wir können, ausgehend von (2), die Funktion (4) noch anders ausrechnen:

$$\frac{\partial f}{\partial y} = f\,\frac{\partial \log f}{\partial y} = f(x,y) \sum_{\nu=1}^{G} \frac{(1+x)^{k_\nu} - 1}{1 + \left[(1+x)^{k_\nu} - 1\right] y}\,,$$

$$(5) \qquad \left(\frac{\partial f}{\partial y}\right)_{y=1} = f(x\ ,\ 1) \sum_{\nu=1}^{G} \frac{(1+x)^{k_\nu} - 1}{(1+x)^{k_\nu}}$$

$$= (1+x)^S \sum_{\nu=1}^{G} [1 - (1+x)^{-k_\nu}]$$

$$= G\,(1+x)^S - \sum_{\nu=1}^{G} (1+x)^{S-k_\nu}\ ;$$

um $f(x\ ,1)$ zu bestimmen, haben wir (1) benützt. Der Vergleich des Koeffizienten von x^s in (4) und (5) ergibt die gewünschte mittlere Anzahl:

$$(6) \qquad \bar{g}_s = G - \sum_{\nu=1}^{G} \binom{S-k_\nu}{s} \Big/ \binom{S}{s} = G - \sum_{\nu=1}^{G} \frac{(S-k_\nu)!\,(S-s)!}{(S-k_\nu-s)!\,S!}\ .$$

Durch die Betrachtung von $\left(\frac{\partial^2 f}{\partial y^2}\right)_{y=1}$ könnten wir eine ähnliche Formel für die mittlere Abweichung der Farbenzahl (Genuszahl) g von dem Mittelwert $\bar{g}_s$ finden.

3. Diskussion und Näherungsformeln. Zur Lösung unserer Aufgabe müssen wir wissen, wie die S im Grundgebiet vorhandenen Spezies auf die G daselbst vertretenen Genera verteilt sind. Wir müssen also insbesondere kennen die Anzahl derjenigen Genera, die im Grundgebiet durch nur eine Art vertreten sind; diese Anzahl sei mit H_1 bezeichnet. Ähnlicherweise sei H_2 die Anzahl derjenigen Genera, die im Grundgebiet durch zwei, H_3 die Anzahl derjenigen, die durch drei Spezies vertreten sind, usw. Wir müssen also alle diese Häufigkeitszahlen

$$H_1\ ,\ H_2\ ,\ H_3\ ,\ H_4\ ,\ \ldots$$

kennen, die die Verteilung der Gattungen nach der Anzahl der von ihnen umfassten, im Grundgebiet heimischen Arten angeben. Sie bilden eine Tabelle, die man als Gattungumfangstabelle bezeichnen könnte. Mit solchen Tabellen hat sich J. C. WILLIS beschäftigt; er untersuchte die Gattungumfänge auch innerhalb von Familien, die Umfänge von endemischen Gattungen usw.[4]) Am Schlusse dieser Arbeit findet der Leser die von MAILLEFER benutzte Gattungumfangstabelle der Schweizer Flora (Tabelle I); sie ist von ähnlicher Struktur, wie die

[4]) J. C. WILLIS, Age and Area, Cambridge 1922. Für eine mathematische Behandlung vgl. insbesondere G. UDNY YULE, A mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, Phil. Transactions R. S. London, Ser. B. Vol. 213 (1924).

von WILLIS betrachteten Gattungumfangstabellen; H_1, die Anzahl der innerhalb der Schweiz monotypen Genera, ist am grössten, und allgemein nimmt H_k ab, als k wächst, und zwar anfänglich recht glatt und regelmässig.

Aus der Definition der Häufigkeitszahlen H_k folgt ohne weiteres, dass, die Summen über alle vorhandenen Werte von k erstreckt,

$$\sum_k H_k = G \,, \tag{7}$$

$$\sum_k k H_k = S \tag{8}$$

ist. Die Formel (6) lässt sich nach Einführung der H_k so schreiben

$$g_s = G - \sum_k H_k \frac{(S-k)\,!\,(S-s)\,!}{(S-k-s)\,!\,S\,!} \,. \tag{9}$$

Durch leichte Umformung erhält man den folgenden Ausdruck für den wahrscheinlichen generischen Koeffizienten:

$$\frac{100\,\bar{g}_s}{s} = 100 \sum_k H_k \frac{1}{s}\left[1-\left(1-\frac{s}{S}\right)\left(1-\frac{s}{S-1}\right)\cdots\left(1-\frac{s}{S-k+1}\right)\right]. \tag{10}$$

Auf Grund von diesem Ausdruck kann man zeigen: Der wahrscheinliche generische Koeffizient, $\frac{100\,\bar{g}_s}{s}$, nimmt ständig ab, und zwar von 100 bis $\frac{100\,G}{S}$, wenn s, die Artanzahl im Teilgebiet, von 1 bis S wächst.

Der Beweis beruht auf der folgenden allgemeinen Bemerkung: Hat das Polynom $f(x)$ lauter positive Wurzeln, von denen a die kleinste ist, und ist $f(0) > 0$, so ist die zweite Derivierte $f''(x)$, da sie links von a nicht verschwindet, positiv im Intervall $0 < x < a$. Somit ist daselbst die Kurve $y = f(x)$ von unten gesehen konvex, und daher nimmt $\frac{f(x)-f(0)}{x}$, als Richtungskoeffizient einer Sehne, ständig zu, wenn x von 0 bis a wächst. Die Anwendung auf das Polynom, dessen Wurzeln $S-k+1, \ldots S-1, S$ sind, und das für $x = 0$ den Wert 1 annimmt, ergibt, dass das allgemeine Glied der Summe rechts in (10) ständig abnimmt, wenn s stetig von 0 bis $S-k+1$ wächst. Von hier bis S fällt aber das Glied, wenn s ganzzahlig wächst, mit der abnehmenden Funktion s^{-1} zusammen.

Die numerische Auswertung der Formel (10) ist recht mühsam. Man gelangt zu einer für kleine Werte von s brauchbaren Näherungsformel, wenn man das im allgemeinen Glied der Summe (10) auftretende Produkt

$$\left(1-\frac{s}{S}\right)\left(1-\frac{s}{S-1}\right)\cdots\left(1-\frac{s}{S-k+1}\right)$$

durch die k-te Potenz

$$\left(1-\frac{s}{S}\right)^k$$

und diese durch ihre ersten 4 Entwicklungsglieder

$$1-\frac{k}{S}s+\frac{k\,(k-1)}{2\,S^2}s^2-\frac{k\,(k-1)\,(k-2)}{6\,S^3}s^3$$

annähernd ersetzt. Es ergibt sich so die Näherungsformel

$$(11)\quad \frac{100\,\bar{g}_s}{s}\sim 100\sum_k H_k\left[\frac{k}{S}-\frac{k\,(k-1)}{2\,S^2}s+\frac{k\,(k-1)\,(k-2)}{6\,S^3}s^2\right]$$
$$=100-b\,s+c\,s^2.$$

Wir benutzten (8) und setzten zur Abkürzung

$$(12)\quad \frac{100}{2\,S^2}\sum_k k\,(k-1)\;H_k=b\;,\quad \frac{100}{6\,S^3}\sum_k k\,(k-1)\,(k-2)\;H_k=c\;.$$

Offenbar ist die Näherungsformel (11) höchstens so lange brauchbar, bis s unterhalb der Abscisse des Minimums bleibt, das heisst solange

$$(13)\qquad s<\frac{b}{2\,c}\;.$$

Eine ähnliche aber etwas genauere Näherungsformel erhält man, wenn man auf die in (9) vorkommenden Fakultäten die STIRLING'sche Formel anwendet und dann nach abnehmenden Potenzen von S, bis zu S^{-3} einschliesslich, entwickelt. Eine längere Rechnung, die ich hier unterdrücke, ergibt

$$(14)\quad \frac{100\,\bar{g}_s}{s}\sim 100\left(1-\frac{1}{6S^2}\right)-\left[b\left(1+\frac{1}{S}\right)+3c\right](s-1)+c\,(s^2-1)\,.$$

Diese Näherungsformel ist ebenfalls höchstens bis zu ihrem Minimum brauchbar.

4. Numerisches Beispiel. Die Formel (10) gestattet uns, den wahrscheinlichen generischen Koeffizienten $100\,\bar{g}_s/s$, ausgehend von einem beliebigen bekannten Grundgebiet, für Teilgebiete mit irgend einer gegebenen Spezieszahl s zu berechnen. Die genaue Formel

(10) ist, wenn s genügend klein, durch die viel bequemere Näherungsformel (14) zu ersetzen.

Ich habe die Rechnung, für die Schweiz als Grundgebiet, mit den von MAILLEFER benutzten Daten ausgeführt, die in der unten folgenden Tabelle I zusammengestellt sind. Die recht mühsame genaue Formel (10) wurde nur für die Speziesanzahl $s = 200$ benutzt [5]). Durch Einsetzen der Daten der Tabelle I in (12) erhält man ziemlich leicht die Zahlenwerte für b und c (es handelt sich ja um die dem Statistiker geläufigen „faktoriellen Momente") und durch Einsetzung der Werte von b, c und S in (14) ergibt sich die Näherungsformel

$$\frac{100\,\bar{g}_s}{s} \sim 100{,}27 - 0{,}2672\,s + 0{,}001329\,s^2. \tag{15}$$

Das Minimum der rechten Seite liegt bei $s = 100{,}5$; die Formel (15) bleibt tatsächlich etwa für $s < 90$ gut brauchbar. In der am Schluss der Arbeit folgenden Tabelle II stelle ich die Werte, die ich auf Grund der Formeln (10) (15) ausgerechnet habe, in der mit „theoretisch" überschriebenen dritten Spalte denjenigen gegenüber, die MAILLEFER durch Urnenexperimente gewonnen hat, und die in der zweiten, mit „beobachtet" überschriebenen Spalte stehen. Die Übereinstimmung ist vollständig befriedigend. Da MAILLEFER seine Zahlen mit einem ausgedehnten pflanzengeographischen Beobachtungsmaterial mit Erfolg verglichen hat, schien mir die Mitteilung der vorangehenden Rechnungen von einem gewissen Interesse zu sein, sowohl für die Wahrscheinlichkeitsrechnung wie für das Studium der Pflanzenverteilung, in welchem Studium auf die Wichtigkeit statistischer Aufnahmen und numerischer Beziehungen ja eben die grundlegenden Arbeiten von P. JACCARD hingewiesen haben.

Ich möchte zum Schluss noch meinem verehrten Kollegen, Herrn Professor P. JACCARD, für seinen freundlichen Hinweis auf die eben behandelte Frage herzlich danken.

---

[5]) Die Rechnung wurde durch Herrn Assistenten E. MOECKLIN durchgeführt.

Tabelle I.

Gattungumfangstabelle der Schweizer Flora nach MAILLEFER[6]).

H_k = Anzahl der Genera, die durch k Arten vertreten sind.

k	H_k	k	H_k	k	H_k	k	H_k
1	331	11	8	21	1	31	1
2	133	12	3	22	2		
3	56	13	6	23	3	34	1
4	36	14	4	24	1		
5	27	15	1	25	0	74	1
6	20	16	2	26	1		
7	15	17	4	27	0	85	1
8	8	18	2	28	1		
9	7	19	4	29	2		
10	7	20	5	30	1		

Tabelle II.

Wahrscheinliche generische Koeffizienten, ermittelt durch Urnenexperimente und Rechnung.

s = Anzahl der Arten im Teilgebiet.

s	beobachtet	theoretisch
5	99,7	98,9
10	97,9	97,7
20	95,3	95,5
30	93,4	93,5
40	91,8	91,7
50	89,6	90,2
60	88,4	89,2
80	87,1	87,6
200	72,1	71,7

[6]) A. a. O. S. 118. Man beachte die Bemerkungen auf S. 117.