Theoretical and Applied Genetics 38, 301—303 (1968)

Die Schätzung der Varianzkomponenten beim Fixwertmodell der Varianzanalyse

ERIK SCHWARZBACH, Brumov/ČSSR *

Estimating Variance Components in a Fixed Model of Variance Analysis

Summary. The partition of a variance into components in experiments with fixed effects, i.e. experiments with varieties, treatment etc., is meaningful only if "components" are defined as additive parts of a whole. This also demands that the sum of the components of variance be equal to the total variance. Such additivity of the components of variance is achieved by appropriate definitions of the components of variance. Unbiassed estimates of the additively defined components of variance are obtained by a simple correction which eliminates the influence of degrees of freedom on the calculation of mean squares. This correction is not needed for the estimation of variances of random effects.

The expected values of mean squares for the partition of variance and formulas for unbiassed estimates of additively defined components of variance are demonstrated in an experiment with two fixed effects and one random effect.

Einleitung

Es ist oft wichtig, die Frage zu beantworten, welchen Anteil eine bestimmte Variabilitätsursache an der Gesamtvariabilität eines untersuchten Objektes besitzt. In Fällen, wo sämtliche Variabilitätsursachen Zufallseinwirkungen sind, läßt sich die Frage oft mit der Varianzanalyse[1] beantworten. So zum Beispiel, wenn aus einer großen Herde zufällig einige Zuchttiere ausgelesen werden und bei den Nachkommen der Anteil der erblich bedingten Variabilität der Herde festgestellt werden soll. Das Verfahren wurde oft beschrieben, siehe z. B. LE ROY (1960). Die eingangs erwähnte Frage soll jedoch oft auch dann beantwortet werden, wenn die gefragten Variabilitätsursachen keine Zufallswirkungen sind. So zum Beispiel, wenn die Wirkung verschiedener Behandlungen geprüft oder in Sortenversuchen nach dem Anteil der Sortenunterschiede an der Gesamtvariabilität eines Versuches gefragt wird. Dies trifft meist in der Pflanzenzüchtung zu. Statt mit Zufallswirkungen hat man es dann mit fixen, meist sorgfältig gewählten, Wirkungsfaktoren (Behandlungen, Sorten, Anbauorten usw.) zu tun. In solchen Fällen ist das zuständige Modell das klassische Fixwertemodell der Varianzanalyse, das ursprünglich ausschließlich zum Prüfen von Differenzen zwischen Gruppenmittelwerten, also für eine andere Fragestellung, bestimmt war (EISENHART 1947). Im Folgenden soll gezeigt werden, daß das für das Zufallswertemodell ausgearbeitete Verfahren zum Schätzen von Varianzkomponenten nicht unverändert auf das Fixwertemodell übertragen werden kann, was bisher nicht berücksichtigt wurde (siehe z. B. SNEDECOR 1962, KEMPTHORNE 1957, WEBER 1964, 1967 u. a.), sondern daß beim Fixwertemodell das Verfahren modifiziert werden muß. Das Problem sei an einem der häufigsten Fälle dargestellt, an der Varianzanalyse mit 2 konstanten Variabilitätsursachen und einer Zufallswirkung.

* Jetzige Anschrift: Erlangen, Sperlingstraße 45.

[1] Das zum Beantworten dieser Frage entwickelte Verfahren nennt EISENHART (1947) die Varianzanalyse im eigentlichen Sinn des Wortes. Das Verfahren darf nicht mit dem Verfahren zum Prüfen von Gruppenmittelwerten, das gemeinhin auch als Varianzanalyse bezeichnet wird, verwechselt werden. Letzteres soll hier nicht behandelt werden.

Die Varianzanalyse eines Versuches mit m Sorten, r Düngungsstufen und n Wiederholungen

a) Additiver Fall (ohne Interaktionen)

Es sei eine Ertragsprüfung mit m Sorten, r Düngungsstufen und n Wiederholungen je Kombination mit zufälliger Anordnung der Teilstücke angenommen. Ein Versuch also, wie er in der Praxis häufig vorkommt. Die mnr Teilstücke unterscheiden sich in ihrem Ertrag, weisen also eine Variabilität auf. Die uns hier interessierende Fragestellung lautet: Welchen Anteil haben die einzelnen Variabilitätsursachen (Sortenunterschiede, Düngungsstufen, Zufallswirkung) an der Gesamtvariabilität der Teilstückerträge?

Diese Frage läßt sich nicht beantworten, wenn sich die Anteile nicht additiv verhalten, mit anderen Worten: wenn nicht die Summe der Anteile gleich dem Ganzen ist. Symbolisch ausgedrückt lautet diese Forderung, wenn als Maß der Variabilität die Varianz, σ^2, benützt wird:

$$\sigma^2_{\text{Gesamt}} = \sigma^2_{\text{Sorten}} + \sigma^2_{\text{Düngung}} + \sigma^2_{\text{Zufallswirkung}} \qquad (1)$$

Hieraus ergibt sich die Notwendigkeit, die Varianzen eindeutig so zu definieren, daß sie sich additiv verhalten. Die Definitionen ergeben sich aus dem zuständigen varianzanalytischen Modell. Das Fixwertemodell wird im hier vorliegenden Falle dadurch charakterisiert, daß angenommen wird, daß von jeder Kombination von Sorte mit Düngung nicht n Wiederholungen, sondern unendlich viele vorhanden sind, deren Einzelwerte normal verteilt sind und innerhalb aller Kombinationen die gleiche Varianz, σ^2_z, besitzen. Ferner, daß sich der Mittelwert jeder Kombination linear aus dem entsprechenden Mittelwert der Sorte und der Düngungsstufe ergibt. Mathematisch kann man das Modell durch die bekannte einfache Formel

$$x_{ijk} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + z_{ijk}$$

ausdrücken, wobei

x_{ijk} = Teilstückertrag
μ = unbekannter Gesamtmittelwert des Modells
μ_i = unbekannter Mittelwert der i-ten Sorte ($i = 1 \cdots m$)
μ_j = unbekannter Mittelwert der j-ten Düngung ($j = 1 \cdots r$)

z_{ijk} = zufallsbedingte Abweichung des k-ten Teilstückertrages vom ij-ten Gruppenmittelwert, μ_{ij}, des Modells ($k = 1 \cdots n$)

Die je n Wiederholungen in den Kombinationen des realen Versuches werden als zufällige, voneinander unabhängige Stichproben aus den unendlich großen Grundgesamtheiten des Modells angesehen.

Entsprechend der Forderung (1) ergeben sich nun die Definitionen der Varianzen wie folgt:

Gesamtvarianz: (2)

$$\sigma_G^2 = E(x_{ijk} - \mu)^2$$

Varianz der Sorten

$$\sigma_S^2 = E(\mu_i - \mu)^2$$

Varianz der Düngungen

$$\sigma_D^2 = E(\mu_j - \mu)^2$$

Varianz der Zufallswirkung

$$\sigma_Z^2 = E(z_{ijk})^2 = E(x_{ijk} - u_{ij})^2$$

wobei das Symbol E Erwartungswert bedeutet.

Unter den so definierten Varianzen besteht bei dem ausgeführten Modell die einfache algebraische Beziehung

$$\sigma_G^2 = \sigma_S^2 + \sigma_D^2 + \sigma_Z^2$$

wodurch die Forderung (1) erfüllt ist.

In vielen statistischen Lehrbüchern werden beim Fixwertemodell die Varianzen der konstanten Faktoren definiert (mit der hier verwendeten Symbolik am Beispiel der Sortenvarianz) als

$$\sigma_S^2 = E\,(\mu_i - \mu)^2 \frac{m}{m-1}\,.$$

So z. B. KEMPTHORNE (1957, S. 252). Dagegen ist nichts einzuwenden, solange nicht gefordert wird, daß es sich um eine Varianzkomponente handeln soll. Denn eine so definierte Größe verhält sich natürlich nicht additiv zu den übrigen Varianzkomponenten und kann daher nicht als Varianzkomponente im eigentlichen Sinn des Wortes angesehen werden. Auch die von WEBER (1964) beim Fixwertemodell benützte Größe σ_1^2 (S. 167 ff) wird in diesem Sinne verstanden. Oft wird beim Fixwertemodell statt des Symbols σ^2 das Symbol k^2 benützt, um zu zeigen, daß es sich um Konstanten handelt, wobei aber diese Größe im selben Sinne verstanden wird wie bei KEMPTHORNE (1957) und als Varianzkomponente bezeichnet wird (SNEDECOR 1962 u. a.). Deshalb können die auf dieser Definition fußenden Verfahren zum Schätzen der Varianzkomponenten zu keiner sinnvollen Antwort auf die eingangs erwähnte Frage bei Versuchen mit festen Variabilitätsursachen führen.

Die unter (2) definierten Varianzen können natürlich nur geschätzt werden. Die Grundlage der Schätzung bei der Varianzanalyse bildet die Berechnung der mittleren Quadrate aus den aufgeteilten Quadratsummen. Die Erwartungswerte der mittleren Quadrate lassen sich mit einfacher Algebra berechnen und finden sich z. T. bereits bei EISENHART (1947). Sie sind in Tabelle 1 zusammengestellt. Die Berechnungsformeln sind in der allgemeinen Form wiedergegeben und entsprechen dem bei der Varianzanalyse üblichen Zerlegungsschema.

Aus den Erwartungswerten in der Tabelle ergeben sich die erwartungstreuen Schätzwerte der Varianzen folgendermaßen:

$$\hat{\sigma}_Z^2 = MQ_i$$
$$\hat{\sigma}_Z^2 = MQ_r$$

(Die zwei unabhängigen Schätzwerte derselben Varianz dürfen sich nicht signifikant unterscheiden (F-Test). Anderenfalls liegen Interaktionen vor und es tritt Modellfall b ein.)

$$\hat{\sigma}_S^2 = \frac{MQ_s - MQ_i}{nr} \cdot \frac{m-1}{m}\,,$$

$$\hat{\sigma}_D^2 = \frac{MQ_d - MQ_i}{mn} \cdot \frac{r-1}{r}\,,$$

$$\hat{\sigma}_G^2 = MQ_g \frac{mnr-1}{mnr} + \frac{MQ_i}{mnr}\,.$$

Diese Schätzungen unterscheiden sich von den Schätzungen beim Zufallswertemodell (vgl. z. B. WEBER 1967) lediglich dadurch, daß die Überhöhung der mittleren Quadrate, die durch Division mit Freiheitsgraden in der Berechnungsformel entstanden ist, wieder korrigiert wird. Nur bei der Schätzung der Zufallswirkung, σ_Z^2, bleiben im Nenner der Berechnungsformel unverändert die Freiheitsgrade.

Tabelle 1. *Erwartungswerte der mittleren Quadrate beim Fixwertemodell der Varianzanalyse*

Variabilität	Mittleres Quadrat	Berechnungsformel	Erwartungswert
Zwischen Sorten	MQ_s	$\dfrac{n\,r \sum_i (\bar{x}_i - \bar{x})^2}{m-1}$	$\sigma_Z^2 + n\,r\,\sigma_S^2 \dfrac{m}{m-1}$
Zwischen Düngungsstufen	MQ_d	$\dfrac{n\,m \sum_j (\bar{x}_j - \bar{x})^2}{r-1}$	$\sigma_Z^2 + n\,m\,\sigma_D^2 \dfrac{r}{r-1}$
Innerhalb Gruppen	MQ_i	$\dfrac{\sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ij})^2}{m\,r\,(n-1)}$	σ_Z^2
Rest	MQ_r	$\dfrac{n \sum_i \sum_j (\bar{x}_i - \bar{x}_j - \bar{x}_{ij} + \bar{x})^2}{(m-1)\,(r-1)}$	a) bei Additivität: σ_Z^2 b) bei Nichtadditivität: $\sigma_Z^2 + n\,\sigma_{S\times D}^2 \dfrac{m\,r}{(m-1)\,(r-1)}$
Gesamt	MQ_g	$\dfrac{\sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2}{m\,n\,r - 1}$	$\sigma_G^2 \dfrac{m\,n\,r}{m\,n\,r-1} - \dfrac{\sigma_Z^2}{m\,n\,r-1}$

b) Nichtadditiver Fall (mit Interaktionen)

Das hier zutreffende Modell unterscheidet sich vom vorangegangenen dadurch, daß sich die Wirkungen der konstanten Faktoren nicht additiv verhalten (hier Wirkung der Sorte und Düngung). Die Abweichung von der Additivität wird als Interaktion bezeichnet. Die Interaktionen wirken sich als zusätzliche Variabilitätsursache auf die Gesamtvariabilität aus. Wenn die Interaktionen als Varianzkomponente erfaßt werden sollen, muß diese so definiert werden, daß sie sich additiv zu den übrigen Varianzkomponenten verhält. Dieser Forderung genügt die Definition

$$\sigma^2_{S\times D} = E\,(\mu_{ij} - \mu_i - \mu_j + \mu)^2 \qquad (3).$$

Unter den Varianzkomponenten besteht dann die einfache algebraische Beziehung

$$\sigma^2_G = \sigma^2_S + \sigma^2_D + \sigma^2_{S\times D} + \sigma^2_Z \qquad (4).$$

Ähnlich wie die Varianzen der einzelnen konstanten Faktoren wird oft die Varianz der Interaktionen konstanter Faktoren verstanden als

$$\sigma^2_{S\times D} = E(\mu_{ij} - \mu_i - \mu_j + \mu)^2 \frac{m\,r}{(m-1)\,(r-1)}.$$

Eine so definierte Varianz verhält sich jedoch nicht additiv und ist somit keine Varianzkomponente im eigentlichen Sinn des Wortes. Falls nach Varianzkomponenten gefragt wird, muß daher die Definition (3) benützt werden.

Die erwartungstreuen Schätzungen der Varianzkomponenten im Modellfall b) ergeben sich aus den Erwartungswerten in Tab. 1. Die erwartungstreue Schätzung der Varianz der Interaktionen, $\sigma^2_{S\times D}$, ergibt sich als

$$\hat\sigma^2_{S\times D} = (M\,Q_r - M\,Q_i)\,\frac{(m-1)\,(r-1)}{m\,n\,r}.$$

Auch hier muß also eine Korrektion vorgenommen werden, um die durch Division mit Freiheitsgraden verursachte Überhöhung des $M\,Q_r$ zu korrigieren.

Die erwartungstreue Schätzung der übrigen Varianzen erfolgt wie unter a).

Diskussion

Die Mißachtung der Notwendigkeit, die Varianzkomponenten additiv zu definieren, führt u. U. zu erheblichen Überschätzungen der Varianzkomponenten, besonders bei geringer Anzahl von Gruppen. Wenn z. B. die Zahl der Sorten 4, die der Behandlungen 3 und die Zahl der Wiederholungen beliebig ist, wird bei üblicher Definition der Varianzen die Varianz der Sorten 4/3-fach, d. h. um 33%, die Varianz der Behandlungen 3/2-fach, d. h. um 50%, und die Varianz der Interaktionen sogar $4\cdot 3/3\cdot 2$-fach, d. h. um 100%, überschätzt. Dies führt dann dazu, daß die Gesamtvarianz stets geringer ist als die Summe der Schätzwerte der Komponenten. Werden hingegen die Varianzen so definiert, daß sie additive Varianzkomponenten darstellen, gilt nicht nur für die unbekannten Varianzen des Modells die Beziehung (4), sondern auch die Schätzwerte der Varianzen verhalten sich dann additiv:

$$\hat\sigma^2_G = \hat\sigma^2_S + \hat\sigma^2_D + \hat\sigma^2_{S\times D} + \hat\sigma^2_Z$$

Man könnte daher auf die Schätzung einer der Varianzen verzichten und den Schätzwert durch Subtraktion gewinnen. Als Rechenkontrolle empfiehlt es sich jedoch, alle Varianzen getrennt zu schätzen.

Wie angeführt, unterscheidet sich das hier ausgeführte Schätzungsverfahren vom bisher üblichen lediglich dadurch, daß die Überhöhung der mittleren Quadrate der konstanten Faktoren, entstanden durch Division mit Freiheitsgraden in der Berechnungsformel, anschließend wieder zurückkorrigiert wird. Man könnte meinen, daß es einfacher wäre, im Divisor der Berechnungsformel des mittleren Quadrates statt der Freiheitsgrade direkt die Gruppenzahl einzusetzen, etwa

$$M\,Q_s = \frac{n\,r \sum_i (\bar x_i - \bar x)^2}{m}.$$

Dabei ist allerdings zu bedenken, daß die Frage nach der Größe der Varianzkomponenten meist erst nach der Prüfung der Nullhypothesen $\sigma^2_S = 0$, $\sigma^2_D = 0$ und $\sigma^2_{S\times D} = 0$ Sinn hat. Die Prüfverfahren sind jedoch gerade auf mittleren Quadraten, die mit Hilfe von Freiheitsgraden berechnet wurden, aufgebaut. Es scheint daher besser, bei der Berechnung der mittleren Quadrate wie bisher Freiheitsgrade zu verwenden. Anders bei den Definitionen der Varianzen. Hier ist zu bemerken, daß bereits P. Lorenz (1959) die Notwendigkeit der Definition der Varianz als arithmetisches Mittel der Quadrate der Abweichungen vom Mittelwert auch für andere als hier angeführte Zwecke betonte, was allerdings viel zu wenig gewürdigt wurde.

Zusammenfassung

Die Schätzung der Varianzkomponenten bei Versuchen mit festen Variabilitätsursachen, z. B. bei der Prüfung von Sorten, Behandlungen usw., ist nur dann sinnvoll, wenn die Varianzkomponenten so definiert sind, daß ihre Summe der Gesamtvarianz gleich ist, d. h. daß sie sich additiv verhalten. Will man die so definierten Varianzkomponenten erwartungstreu schätzen, muß man die Überhöhung der mittleren Quadrate, die durch Division mit Freiheitsgraden statt mit der Gruppenzahl entstanden ist, wieder korrigieren. Diese Korrektion entfällt bei der Schätzung der Varianz von Zufallswirkungen.

Die Erwartungswerte der mittleren Quadrate der üblichen varianzanalytischen Zerlegung und die erwartungstreuen Schätzungen der additiv definierten Varianzkomponenten werden am Beispiel eines Versuches mit zwei konstanten Variabilitätsursachen und einer Zufallseinwirkung dargelegt.

Literatur

1. Eisenhart, C.: The assumptions underlying the analysis of variance. Biometrics 3, 1–21 (1947). — 2. Kempthorne, O.: An introduction to genetic statistics. New York: J. Wiley and Sons, Inc. 1957. — 3. Le Roy, H. L.: Statistische Methoden der Populationsgenetik. Basel u. Stuttgart: Birkhäuser-Verlag 1960. — 4. Lorenz, P.: Anschauungsunterricht in mathematischer Statistik. II. Der Schluss vom Teil aufs Ganze. Leipzig: S. Hirzel-Verlag 1959. — 5. Snedecor, G. W.: Statistical methods. Ames, Iowa: The Iowa State University 1962. — 6. Weber, Erna: Grundriß der biologischen Statistik. 5. Aufl. Jena: G. Fischer-Verlag 1964. — 7. Weber, Erna: Mathematische Grundlagen der Genetik. Jena: G. Fischer-Verlag 1967.