Prof. Dr. Klaus Eckhardt

Varianzanalyse

veröffentlicht im Internet unter aufgabomat.de

Inhalte: ANOVA, ein- und zweifaktorielle Varianzanalyse mit festen Effekten, Faktoren, Faktorstufen, Faktorstufenmittelwerte, Faktorstufenkombinationen, Streuungszerlegung, empirische Varianz, Freiheitsgrade

Gliederung:

1 Einleitung

Die Varianzanalyse ist ein Verfahren der Zusammenhangsanalyse. Abgeleitet von der Abkürzung der englischen Bezeichnung „analysis of variance“ wird sie auch als **ANOVA** bezeichnet. Ziel einer Varianzanalyse ist es zu klären, ob ein oder mehrere **Faktoren** Einfluss auf eine oder mehrere interessierende Größen haben. Die Faktoren sind nominalskalierte Merkmale bzw. Variablen. Die möglicherweise beeinflussten Größen sind höher skalierte, d. h. mindestens intervallskalierte Merkmale. Gegenstand einer Varianzanalyse könnten beispielsweise sein:

- Auswirkung unterschiedlicher Lehrmethoden auf den Lernerfolg
- Wirkung der Art der Warenplatzierung in Supermärkten auf den Verkauf eines Produkts
- Einfluss von Studienfach und Geschlecht auf das berufliche Einkommen
- Einfluss unterschiedlicher Formen der Düngung und Bodenbearbeitung auf Ertrag und Proteingehalt eines Getreides.

Untersucht wird, ob die Varianz, die in einem Datensatz durch einzelne Faktoren oder durch Wechselwirkungen zwischen diesen Faktoren hervorgerufen wird, größer ist als der Varianzanteil, der auf zufällige Einflüsse zurückzuführen ist. Die Varianzanteile werden berechnet und dann paarweise mittels des

F-Tests verglichen. Da sich die Varianz aus der Summe der quadrierten Abweichungen ableitet und zu deren Berechnung verschiedene Mittelwerte gebildet worden sein müssen, lässt sich die Varianzanalyse in die folgenden Schritte unterteilen: Berechnung von Mittelwerten - Berechnung von Summen quadrierter Abweichungen - Berechnung von Varianzen - F-Tests. Dies wird nachfolgend für die so genannte ein- und zweifaktorielle Varianzanalyse erläutert.

Die Varianzanalyse wird im vorliegenden Skript nur auszugsweise behandelt. Erstens beschränkt sich die Darstellung auf die so genannte **Varianzanalyse mit festen Effekten**. Bei dieser werden die Ausprägungen der Faktoren gezielt vorgegeben, um etwas über ihre mittleren Effekte in Erfahrung zu bringen. Die Daten entstammen unterschiedlichen Grundgesamtheiten, deren Erwartungswerte verglichen werden sollen. Zweitens werden maximal zwei Faktoren berücksichtigt, auch wenn die Anzahl der Faktoren in der mehrfaktoriellen Varianzanalyse durchaus auch größer als zwei sein kann. Drittens wird nur die **univariate Varianzanalyse** behandelt, bei der genau ein potenziell abhängiges Merkmal vorliegt. Auch hier könnte die Anzahl der Merkmale größer sein. Man würde dann von einer multivariaten Varianzanalyse sprechen.

Zwei abschließende Anmerkungen noch:

- Die mathematische Formalisierung der folgenden Ausführungen wird auf Sie möglicherweise abschreckend wirken. Es handelt sich dabei aber lediglich um die kompakte Darstellung von mathematischen Ausdrücken, die zu berechnen Ihnen nicht mehr abverlangt, als die Grundrechenarten zu beherrschen. Dies wird jeweils an einem Beispiel erläutert.

- Wieso sollen Sie sich überhaupt mit den mathematischen Details befassen, wo es doch eine Vielzahl von Programmen gibt, in denen Funktionen zur Durchführung von Varianzanalysen implementiert sind? Die Antwort lautet: Ein Statistikprogramm bedienen zu können, ist nicht gleichbedeutend damit, die Statistik dahinter zu verstehen. In einer Studie in den USA wurden behandelnde Ärzte, Ärzte im Praktikum und Medizin-Studierende zur Deutung statistischer Kennzahlen eines medizinischen Tests befragt[1]. Rund drei Viertel der Befragten gaben eine falsche Antwort. Wäre das Ergebnis unter Ärzten in Deutschland wesentlich besser? Vermutlich nicht[2]. Und was würde wohl eine entsprechende Befragung von Absolventen anderer, nichtmedizinischer Studiengänge zur Beurteilung von Daten aus ihrem jeweiligen Fachgebiet ergeben?

  Um statistische Verfahren richtig einsetzen und ihre Ergebnisse korrekt deuten zu können, muss man ihre mathematischen Hintergründe verstanden haben.

2 Einfaktorielle Varianzanalyse mit festen Effekten

2.1 Ausgangssituation

Gegeben:
- Daten zu einer nominalskalierten Variable (einem Faktor) A
- Daten zu einer möglicherweise von diesem Faktor abhängigen, mindestens intervallskalierten Variable X

[1] Manrai, A. K., Bhatia, G., Strymish, J., Kohane, I. S., Jain, S. H., 2014: Medicine’s uncomfortable relationship with math: Calculating positive predictive value. JAMA Internal Medicine 174(6), 991-993, doi:10.1001/jamainternmed.2014.1059.
[2] Jenny, M. A., Keller, N., Gigerenzer, G., 2018: Assessing minimal medical statistical literacy using the Quick Risk Test: A prospective observational study in Germany. BMJ Open 8(8), doi:10.1136/bmjopen-2017-020847.

Voraussetzungen:

- Das Merkmal X wird nur durch den einen Faktor A systematisch beeinflusst, der in I verschiedenen Ausprägungen bzw. Faktorstufen vorgegeben ist.
- Zu den I Faktorstufen sind jeweils mehrere Messwerte x_{ij} erfasst worden, d. h. es liegen I Stichproben vor. Es handelt sich dabei jeweils um die Stichprobe einer normalverteilten Variable. Die Variable auf der Faktorstufe i wird im Folgenden mit X_i bezeichnet.
- Jede der I normalverteilten Variablen X_i hat dieselbe Varianz: $Var(X_i) = \sigma^2$ für alle i = 1, ..., I.
- Die Variablen X_i sind unabhängig voneinander.

Null- und Alternativhypothese:

H_0: Die Erwartungswerte $E(X_i)$ der Variablen X_i stimmen überein, d. h. $E(X_i) = \mu$ für alle i = 1, ..., I.

H_1: Für mindestens ein i ist $E(X_i) \neq \mu$.

Beispiel: Drei Weizensorten sollen hinsichtlich ihres Ertrags X verglichen werden. Dazu werden die Weizensorten auf Versuchsparzellen angebaut. Am Ende der Wachstumsperiode liegen Daten von 11 Parzellen vor. Der Ertrag ist jeweils in Dezitonnen pro Hektar angegeben:

Sorte	Stichproben-umfang N_i	Messwerte (dt/ha)				Mittelwert (dt/ha)
1	4	81	72	68	69	73
2	4	84	82	68	72	77
3	3	57	62	61		60

Faktorstufen i = 1, ..., I   Merkmalswerte x_{ij} (j = 1, ..., N_i)   Faktorstufenmittelwerte $\bar{x}_i$

Der Faktor ist hier die Weizensorte. Dieser liegt in I = 3 Stufen vor. Für jede Faktorstufe i gibt es mehrere Messwerte x_{ij}. Mit dem zweiten Index j werden die Messwerte in den einzelnen Faktorstufen durchnummeriert. Die Anzahl der Messwerte in der Faktorstufe i wird mit N_i bezeichnet. Es ist $N_1 = 4$, $N_2 = 4$ und $N_3 = 3$. Der empirische Mittelwert (das arithmetische Mittel) der N_i Messwerte in der Faktorstufe i wird auch **Faktorstufenmittelwert** genannt und hier mit $\bar{x}_i$ bezeichnet. Da es drei Faktorstufen gibt, gibt es drei Faktorstufenmittelwerte. Diese fallen unterschiedlich aus, was Anlass zu der Vermutung gibt, dass sich die Erwartungswerte des Ertrags für die drei Weizensorten unterscheiden könnten.

H_0: Der Erwartungswert des Ertrags ist unabhängig von der Sorte.

H_1: Der Erwartungswert des Ertrags mindestens einer Sorte unterscheidet sich von dem der anderen Sorten.

2.2 Durchführung der Analyse

2.2.1 Empirische Mittelwerte berechnen

Der Gesamtmittelwert bzw. das arithmetische Mittel aller Merkmalswerte x_{ij} ist

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{N_i} x_{ij} \,. \tag{1}$$

Das doppelte Summenzeichen steht hier, weil zwei Indizes, i und j, ihre Werte durchlaufen müssen, um alle Werte zu erfassen. Der Index i bezeichnet die Faktorstufen, der Index j die Messwerte auf der jeweiligen Faktorstufe:

$$\frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\,(x_{11} + x_{12} + \ldots + x_{1N_1} \quad \leftarrow \text{Messwerte zur Faktorstufe 1 bzw. in der Stichprobe 1}$$

$$+\, x_{21} + x_{22} + \ldots + x_{2N_2} \quad \leftarrow \text{Messwerte zur Faktorstufe 2 bzw. in der Stichprobe 2}$$

$$+\, \ldots$$

$$+\, x_{I1} + x_{I2} + \ldots + x_{IN_I}) \quad \leftarrow \text{Messwerte zur Faktorstufe I bzw. in der Stichprobe I}$$

Die Faktorstufenmittelwerte bzw. empirischen Mittelwerte der Stichproben auf den I Stufen des Faktors ergeben sich als

$$\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}\,. \tag{2}$$

Beispiel: $\bar{x} = \frac{1}{11}\,(81 + 72 + 68 + 69 + 84 + 82 + 68 + 72 + 57 + 62 + 61)$ dt/ha
$= 71$ dt/ha

$\bar{x}_1 = 73$ dt/ha, $\bar{x}_2 = 77$ dt/ha, $\bar{x}_3 = 60$ dt/ha

2.2.2 Streuungszerlegung vornehmen

Im Folgenden steht SQ als Abkürzung für die **Summe der quadrierten Abweichungen** und zwar der Abweichungen der Messwerte von ihrem Mittelwert. Die Summe der quadrierten Abweichungen ist ein Maß für die Streuung der Messwerte. Die **Streuungszerlegung** besteht darin, die Gesamtstreuung SQ der Messwerte aufzuteilen in einen Anteil SQ_A, der durch den Faktor A verursacht wird, und einen Anteil SQ_R, der nicht durch die Wirkung des Faktors erklärt wird und folglich zufällig entstanden ist. SQ_R wird auch als die **Reststreuung** bezeichnet. Mathematisch ausgedrückt lautet die Streuungszerlegung:

$$SQ = SQ_A + SQ_R. \tag{3}$$

Darin sind:

- SQ: Gesamtstreuung

$$SQ = \sum_{i=1}^{I}\sum_{j=1}^{N_i}(x_{ij} - \bar{x})^2 \tag{4}$$

Hier werden die Abweichungen sämtlicher Merkmalswerte x_{ij} von ihrem Mittelwert eingerechnet. Sollten Sie das Tabellenkalkulationsprogramm Excel verwenden, können Sie für die Berechnung die Funktion SUMQUADABW verwenden.

Beispiel: $SQ = [(81 - 71)^2 + (72 - 71)^2 + \ldots + (61 - 71)^2]\ (\text{dt/ha})^2$
$= 789\ (\text{dt/ha})^2$

- SQ_A: durch den Faktor A verursachte Streuung

$$SQ_A = \sum_{i=1}^{I} N_i\,(\bar{x}_i - \bar{x})^2 \tag{5}$$

Hat der Faktor keine Wirkung, so unterscheiden sich die Faktorstufenmittelwerte $\bar{x}_i$ im Allgemeinen wenig voneinander und stimmen weitgehend mit dem Gesamtmittelwert $\bar{x}$ überein. Dem Faktor kann nur dann eine Wirkung zugeschrieben werden, wenn es auf mindestens einer Stufe eine vom Betrag her größere Differenz $\bar{x}_i - \bar{x}$ gibt. Daher sind es diese Differenzen, auf denen die Berechnung von SQ_A basiert.

Würde in Gleichung 5 hinter dem Summenzeichen nur $(\bar{x}_i - \bar{x})^2$ stehen, so wäre SQ_A die Summe von lediglich I quadrierten Abweichungen. SQ dagegen (Gleichung 4) wird aus einer weit größeren Anzahl von Summanden gebildet, den quadrierten Abweichungen aller Messwerte x_{ij} vom Mittelwert $\bar{x}$. Darum wird bei der Berechnung von SQ_A jede quadrierte Abweichung $(\bar{x}_i - \bar{x})^2$ mit dem Stichprobenumfang N_i auf der betreffenden Faktorstufe multipliziert. Dies bewirkt, dass SQ_A und SQ der Summe einer gleich großen Anzahl von quadrierten Abweichungen entsprechen.

Beispiel: $SQ_A = [4\,(73 - 71)^2 + 4\,(77 - 71)^2 + 3\,(60 - 71)^2]\ (\text{dt/ha})^2$
$= 523\ (\text{dt/ha})^2$

- SQ_R: zufällige Streuung bzw. Reststreuung

Nach der Berechnung von SQ und SQ_A ergibt sich die Reststreuung als

$$SQ_R = SQ - SQ_A. \tag{6}$$

Beispiel: $SQ_R = (789 - 523)\ (\text{dt/ha})^2$
$= 266\ (\text{dt/ha})^2$

2.2.3 Streuungen in empirische Varianzen umrechnen

Von der Summe der quadrierten Abweichungen, die im vorangehenden Abschnitt behandelt wurde, leitet sich das in der Statistik meistgebrauchte Streuungsmaß ab, die Varianz. Im vorliegenden Fall wird die Varianz aus Stichprobenwerten abgeleitet und darum als **empirische Varianz** bezeichnet. Es ist:

$$\text{empirische Varianz} = \frac{\text{Summe der quadrierten Abweichungen}}{\text{Anzahl der Freiheitsgrade}}\ . \tag{7}$$

Die Anzahl der **Freiheitsgrade** berechnet sich wie folgt:

$$\begin{array}{c} \text{Anzahl der} \\ \text{Freiheitsgrade} \end{array} = \begin{array}{c} \text{Anzahl der voneinander unabhängigen Werte, die in} \\ \text{die Berechnung der betreffenden Größe eingehen} \end{array} - \begin{array}{c} \text{Anzahl der ab-} \\ \text{hängigen Werte.} \end{array} \tag{8}$$

Dies sollte Ihnen schon einmal begegnet sein, nämlich in Form der empirischen Varianz der Einzelwerte in einer Stichprobe des Umfangs N:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2\ .$$

In diese Berechnung gehen die N Stichprobenwerte und eine aus diesen Stichprobenwerten abgeleitete Größe ein, nämlich der empirische Mittelwert. Die Anzahl der Freiheitsgrade ist daher f = N – 1, die Summe der quadrierten Abweichungen wird durch N – 1 geteilt.

Aus den in Abschnitt 2.2.2 berechneten Summen der quadrierten Abweichungen werden die folgenden Varianzen abgeleitet:

- S_A^2: unabhängig: I Faktorstufenmittelwerte (Gleichung 5)
  abhängig: Mittelwert $\bar{x}$
  Freiheitsgrade: $f_A = I - 1$

$$S_A^2 = \frac{SQ_A}{I - 1} \tag{9}$$

- S_R^2: $SQ_R = SQ - SQ_A$ (Gleichung 6)
  $\Rightarrow f_R = f - f_A$
  $= (N - 1) - (I - 1)$
  $= N - I$

$$S_R^2 = \frac{SQ_R}{N - I} \tag{10}$$

Beispiel: $s_A^2 = \dfrac{523\,(\mathrm{dt/ha})^2}{3-1}$ $= 262\ (\mathrm{dt/ha})^2$

$s_R^2 = \dfrac{262\,(\mathrm{dt/ha})^2}{11-3}$ $= 33\ (\mathrm{dt/ha})^2$

2.2.4 Einseitigen F-Test durchführen

Im Beispiel ist die auf den Faktor zurückzuführende empirische Varianz größer als diejenige, die als zufällig anzusehen ist ($s_A^2 > s_R^2$). Ist dies nur in dieser Stichprobe so, oder deutet dies darauf hin, dass in den zugehörigen Grundgesamtheiten $\sigma_A^2 > \sigma_R^2$ ist?[3] Um die Varianzen zweier normalverteilter Zufallsvariablen zu vergleichen, wird üblicherweise der **F-Test** verwendet, so auch hier. Die **Teststatistik** des F-Tests ist

$$F = \frac{S_A^2}{S_R^2}. \tag{11}$$

Die Teststatistik ist eine Variable, die dann, wenn die Nullhypothese des Tests gilt, bekannte Eigenschaften aufweist. Im Fall des F-Tests ist sie F-verteilt. Die F-Verteilung hat zwei Parameter, den Freiheitsgrad f_1 des Zählers und den Freiheitsgrad f_2 des Nenners in Gleichung 11. Im vorliegenden Zusammenhang ist $f_1 = f_A = I - 1$ und $f_2 = f_R = N - I$.

Der Wertebereich der Teststatistik ist [0; ∞[. Kleine Werte der Teststatistik ergeben sich, wenn die faktorbedingte empirische Varianz klein und daher davon auszugehen ist, dass der Faktor keine Wirkung

[3] Zur Erinnerung: Die Varianz Var(X) ist ein Maß für die Streuung einer Zufallsvariable X um ihren Erwartungswert. Normalverteilte Variablen werden durch zwei Parameter beschrieben, den Mittelwert μ und die Standardabweichung σ. Speziell im Fall normalverteilter Variablen ist $\mathrm{Var}(X) = \sigma^2$. Die Varianzen σ_A^2 und σ_R^2 sind also Maße für die Streuung zweier normalverteilter Variablen. Diese Variablen sind aber nicht vollständig bekannt. Statt ihrer Varianzen σ_A^2 und σ_R^2 sind nur Schätzwerte für diese Varianzen gegeben, die empirischen Varianzen s_A^2 und s_R^2. Daher die Frage nach den Grundgesamtheiten, d. h. nach den Variablen in ihrer vollständigen Beschreibung.

hat. Ein Einfluss des Faktors ist nur im Fall großer Werte der Teststatistik anzunehmen. Der Test wird daher rechtsseitig mit den folgenden Hypothesen durchgeführt:

H_0: $\sigma_A^2 \leq \sigma_R^2$
H_1: $\sigma_A^2 > \sigma_R^2$.

Beispiel: Als Irrtumswahrscheinlichkeit wird $\alpha = 0{,}05$ gewählt. Der Wert der Teststatistik ist

$$F = \frac{262}{33} = 7{,}9.$$

Die rechtsseitige Begrenzung des Annahmeintervalls wird durch das 0,95-Quantil der F-Verteilung mit $f_1 = 2$ und $f_2 = 8$ gebildet. Dieses hat den Wert $F_{0,95} = 4{,}5$, wie sich einer Quantiltabelle der F-Verteilung entnehmen oder mit einer entsprechenden Funktion eines Tabellenkalkulations- oder Statistikprogramm berechnen lässt, in Excel beispielsweise als F.INV(0,95;2;8).

$F > F_{0,95} \Rightarrow H_0$ wird verworfen.

Wird die Nullhypothese des Tests verworfen, so wird zwangsläufig von der Gültigkeit der Alternativhypothese ausgegangen. $\sigma_A^2 > \sigma_R^2$ aber bedeutet: Dem Faktor A wird eine Wirkung auf die Zielvariable X zugeschrieben. Diese besteht darin, dass sich der Erwartungswert der Zielvariable für mindestens eine Ausprägung des Faktors von denjenigen Erwartungswerten unterscheidet, die auf den anderen Faktorstufen auftreten. Im Beispiel: Der Erwartungswert des Ertrags mindestens einer Sorte unterscheidet sich von dem der anderen Sorten. Wie viele Sorten und welche Sorten sich in dieser Weise auszeichnen, darüber gibt die Varianzanalyse keine Information, sodass an die Varianzanalyse bei Ablehnung der Nullhypothese üblicherweise ein multipler Mittelwertvergleich angeschlossen wird[4].

Weitere Übungsaufgaben finden sich beispielweise im Internet unter der Adresse aufgabomat.de in der Rubrik Statistik.

3 Zweifaktorielle Varianzanalyse mit festen Effekten

3.1 Ausgangssituation

Gegeben:

- Daten zu zwei nominalskalierten Variablen (Faktoren) A und B
- Daten zu einer möglicherweise von diesen Faktoren abhängigen, mindestens intervallskalierten Variable X

Voraussetzungen:

- Das Merkmal X wird nur durch die zwei Faktoren A und B systematisch beeinflusst, die in I und J Stufen und damit in $I \cdot J$ Faktorstufenkombinationen vorgegeben sind.
- Zu den I Stufen des Faktors A und den J Stufen des Faktors B sind jeweils K Messwerte x_{ijk} erfasst worden, d. h. es liegen $I \cdot J$ gleichgroße Stichproben vor. Es handelt sich dabei jeweils um die Stichprobe einer normalverteilten Variable X_{ij}. Die Gesamtzahl der Messwerte beträgt $N = I \cdot J \cdot K$.

[4] Siehe beispielsweise Skript „Multiple Mittelwertvergleiche" unter aufgabomat.de.

- Jede der I · J normalverteilten Variablen X_{ij} hat dieselbe Varianz: $Var(X_{ij}) = \sigma^2$ für alle i = 1, ..., I und j = 1, ..., J.
- Die Variablen X_{ij} sind unabhängig voneinander.

Null- und Alternativhypothese:

H_0: Die Erwartungswerte $E(X_{ij})$ der Variablen X_{ij} in den I · J Faktorstufenkombinationen stimmen überein, d. h. $E(X_{ij}) = \mu$ für alle i = 1, ..., I und j = 1, ..., J.

H_1: Für mindestens eine Kombination ij ist $E(X_{ij}) \neq \mu$.

Beispiel: Zwei Getreidesorten werden mit drei verschiedenen Düngern behandelt. Jede Faktorstufenkombination wird auf drei Parzellen getestet, oder, wie man auch sagt, in drei **Wiederholungen** (Tabelle 1). Wie sind die Effekte von Sorte und Dünger auf den Ertrag zu beurteilen?

Sorte (Faktor A)	Wiederholungen	Dünger (Faktor B) j = 1	j = 2	j = 3
i = 1	k = 1	62	65	76
	k = 2	68	69	80
	k = 3	70	74	71
i = 2	k = 1	57	69	76
	k = 2	62	62	76
	k = 3	61	60	71

Tabelle 1: Messwerte x_{ijk} des Ertrags (Einheit: 1 dt/ha).

3.2 Durchführung der Analyse

3.2.1 Empirische Mittelwerte berechnen

Der Gesamtmittelwert bzw. das arithmetische Mittel aller Merkmalswerte x_{ijk} ist

$$\bar{x} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} \ . \tag{12}$$

Es sind drei Summenzeichen erforderlich, da jetzt mit drei Indizes i, j und k gearbeitet werden muss. Der Index i bezeichnet die Stufen des Faktors A und der Index j die Stufen des Faktors B. Die jeweils K Messwerte in jeder Faktorstufenkombination werden mit dem Index k durchnummeriert.

Beispiel: Faktor A: Sorte, I = 2
Faktor B: Dünger, J = 3
K = 3

$$\bar{x} = \frac{1}{2\cdot 3\cdot 3}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3} x_{ijk}$$

$$= \frac{1}{18}(x_{111} + x_{112} + x_{113} + x_{121} + x_{122} + x_{123} + x_{131} + x_{132} + x_{133}$$

$$+ x_{211} + x_{212} + x_{213} + x_{221} + x_{222} + x_{223} + x_{231} + x_{232} + x_{233})$$

$$= \frac{1}{18}(62 + 68 + 70 + 65 + 69 + 74 + 76 + 80 + 71$$

$$+ 75 + 62 + 61 + 69 + 62 + 60 + 76 + 76 + 71) \text{ dt/ha}$$

$= 68$ dt/ha

In den I Stufen des Faktors A gibt es jeweils K Werte zu den J Stufen des Faktors B, das sind jeweils J · K Werte. Die empirischen Mittelwerte in den Stufen des Faktors A ergeben sich als

$$\bar{x}_{A\,i} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} \;. \tag{13}$$

Beispiel: Die Sorte 1 ist auf jeweils drei Parzellen mit Dünger 1, 2 und 3 behandelt worden, sodass zur Sorte 1 neun Messwerte vorliegen.

$$\begin{aligned}\bar{x}_{A\,1} &= \frac{1}{3\cdot 3}\sum_{j=1}^{3}\sum_{k=1}^{3} x_{1jk}\\ &= \frac{1}{9}\,(x_{111} + x_{112} + x_{113} + x_{121} + x_{122} + x_{123} + x_{131} + x_{132} + x_{133})\\ &= \frac{1}{9}\,(62 + 68 + 70 + 65 + 69 + 74 + 76 + 80 + 71)\text{ dt/ha}\\ &= 71\text{ dt/ha}\end{aligned}$$

Analog ergibt sich für die Sorte 2 $\bar{x}_{A\,2} = 66$ dt/ha. Die beiden Faktorstufenmittelwerte sind unterschiedlich, was ein Hinweis darauf sein könnte, dass die Sorte den mittleren Ertrag beeinflusst.

In den J Stufen des Faktors B gibt es jeweils K Werte zu den I Stufen des Faktors A, das sind jeweils I · K Werte. Die empirischen Mittelwerte in den Stufen des Faktors B ergeben sich als

$$\bar{x}_{B\,j} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk} \;. \tag{14}$$

Beispiel:
$$\begin{aligned}\bar{x}_{B\,1} &= \frac{1}{2\cdot 3}\sum_{i=1}^{2}\sum_{k=1}^{3} x_{i1k}\\ &= \frac{1}{6}\,(x_{111} + x_{112} + x_{113} + x_{211} + x_{212} + x_{213})\\ &= \frac{1}{6}\,(62 + 68 + 70 + 57 + 62 + 61)\text{ dt/ha}\\ &= 63\text{ dt/ha}\end{aligned}$$

$$\begin{aligned}\bar{x}_{B\,2} &= 67\text{ dt/ha}\\ \bar{x}_{B\,3} &= 75\text{ dt/ha}\end{aligned}$$

Die Faktorstufenmittelwerte für die verschiedenen Dünger unterscheiden sich ebenfalls. Der mittlere Ertrag könnte auch vom Dünger abhängen.

Neu bei der mehrfaktoriellen Varianzanalyse im Vergleich zur einfaktoriellen ist, dass auch die Faktorstufenkombinationen zu analysieren sind. Der empirische Mittelwert in der Faktorstufenkombination ij ist

$$\bar{x}_{ij} = \frac{1}{K}\sum_{k=1}^{K} x_{ijk} \,. \tag{15}$$

Beispiel:

Sorte (Faktor A)	Dünger (Faktor B) j = 1	j = 2	j = 3
i = 1	67	69	76
i = 2	60	64	74

Tabelle 2: Empirische Mittelwerte $\bar{x}_{ij}$ des Ertrags (Einheit: 1 dt/ha) in den Faktorstufenkombinationen der Tabelle 1.

Abbildung 1 zeigt die Werte aus Tabelle 2 in grafischer Darstellung. Die Sorte 2 liefert im Mittel niedrigere Erträge. Wäre dieser Effekt vom verwendeten Dünger unabhängig, so würden die beiden Kurvenzüge in der Abbildung - bis auf zufällige Abweichungen - parallel verlaufen.

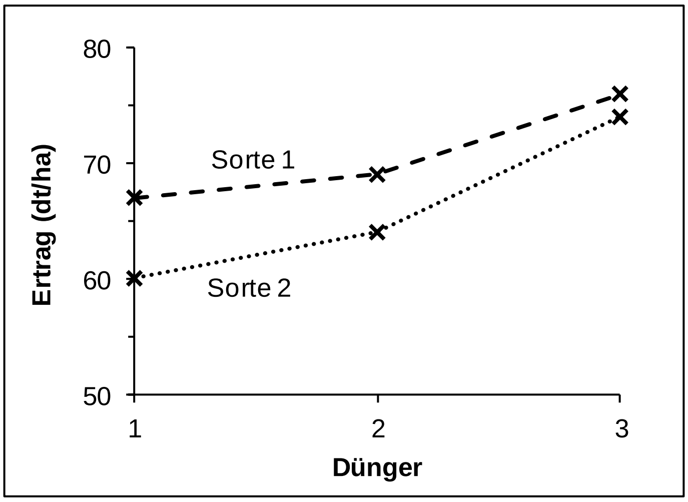

Abbildung 1: Grafische Darstellung der Werte aus Tabelle 2.

Ein paralleler Verlauf tritt im vorliegenden Fall nicht auf. Die Ertragsunterschiede zwischen den Sorten unterscheiden sich je nach Dünger. Zwischen Sorte und Dünger deutet sich eine **Wechselwirkung** an: Sorte 2 scheint stärker auf die Änderung der Düngung zu reagieren als Sorte 1.

3.2.2 Streuungszerlegung vornehmen

Im Fall der zweifaktoriellen Varianzanalyse lautet die Streuungszerlegung

$$SQ = SQ_A + SQ_B + SQ_{AB} + SQ_R. \tag{16}$$

- SQ: Gesamtstreuung

$$SQ = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x})^2 \tag{17}$$

Beispiel: $SQ = 747\ dt^2/ha^2$

Mit der Excel-Funktion SUMQUADABW ergibt sich aufgrund einer anderen Rundung des empirischen Mittelwerts SQ = 746 dt^2/ha^2.

- SQ_A: durch den Faktor A verursachte Streuung

$$SQ_A = J\,K \sum_{i=1}^{I} (\bar{x}_{A\,i} - \bar{x})^2 \tag{18}$$

Eine Wirkung des Faktors A kann nur dann vermutet werden, wenn mindestens ein Faktorstufenmittelwert $\bar{x}_{A\,i}$ eine größere Abweichung vom Gesamtmittelwert $\bar{x}$ zeigt. Daher basiert die Berechnung von SQ_A auf den quadrierten Abweichungen $(\bar{x}_{A\,i} - \bar{x})^2$. Um zu bewirken, dass sowohl SQ_A als auch SQ der Summe von N = I · J · K quadrierten Abweichungen entsprechen, wird die I Terme umfassende Summe in Gleichung 18 mit J · K multipliziert.

Beispiel: SQ_A = 3 · 3 [(71 – 68)² + (66 – 68)²] dt^2/ha^2
= 117 dt^2/ha^2

- SQ_B: durch den Faktor B verursachte Streuung

$$SQ_B = I\,K \sum_{j=1}^{J} (\bar{x}_{B\,j} - \bar{x})^2 \tag{19}$$

Eine Wirkung des Faktors B kann nur dann vermutet werden, wenn mindestens ein Faktorstufenmittelwert $\bar{x}_{B\,j}$ eine größere Abweichung vom Gesamtmittelwert $\bar{x}$ zeigt. Daher basiert die Berechnung von SQ_B auf den quadrierten Abweichungen $(\bar{x}_{B\,j} - \bar{x})^2$. Um zu bewirken, dass SQ_B ebenso wie SQ_A und SQ der Summe von N = I · J · K quadrierten Abweichungen entspricht, wird die J Terme umfassende Summe in Gleichung 19 mit I · K multipliziert.

Beispiel: SQ_B = 2 · 3 [(63 – 68)² + (67 – 68)² + (75 – 68)²] dt^2/ha^2
= 450 dt^2/ha^2

- SQ_{AB}: durch eine Wechselwirkung der Faktoren A und B verursachte Streuung

$$SQ_{AB} = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{x}_{ij} - \bar{x}_{A\,i} - \bar{x}_{B\,j} + \bar{x})^2 \tag{20}$$

Der Ausdruck $\bar{x}_{ij} - \bar{x}_{A\,i} - \bar{x}_{B\,j} + \bar{x}$ hinter dem zweifachen Summenzeichen ist wie folgt zu deuten: Es ist

$$\bar{x}_{ij} - \bar{x}_{A\,i} - \bar{x}_{B\,j} + \bar{x} = (\bar{x}_{ij} - \bar{x}) - (\bar{x}_{A\,i} - \bar{x}) - (\bar{x}_{B\,j} - \bar{x}).$$

Die Differenzen $\bar{x}_{A\,i} - \bar{x}$ sind ein Maß für die Wirkung des Faktors A, die Differenzen $\bar{x}_{B\,j} - \bar{x}$ ein Maß für die Wirkung des Faktors B. Der Ausdruck $(\bar{x}_{ij} - \bar{x}) - (\bar{x}_{A\,i} - \bar{x}) - (\bar{x}_{B\,j} - \bar{x})$ stellt daher die Abweichung des Mittelwerts $\bar{x}_{ij}$ vom Gesamtmittel $\bar{x}$ dar, die sich nicht durch die Effekte der Faktoren A und B alleine erklären lässt.

Um zu bewirken, dass SQ_{AB} ebenso wie SQ_A, SQ_B und SQ der Summe von N = I · J · K quadrierten Abweichungen entspricht, wird die I · J Terme umfassende Summe in Gleichung 20 mit K multipliziert.

Beispiel: $SQ_{AB} = 3\,[\,(\bar{x}_{11} - \bar{x}_{A\,1} - \bar{x}_{B\,1} + \bar{x}\,)^2 + (\bar{x}_{12} - \bar{x}_{A\,1} - \bar{x}_{B\,2} + \bar{x}\,)^2 + (\bar{x}_{13} - \bar{x}_{A\,1} - \bar{x}_{B\,3} + \bar{x}\,)^2 +$
$(\bar{x}_{21} - \bar{x}_{A\,2} - \bar{x}_{B\,1} + \bar{x}\,)^2 + (\bar{x}_{22} - \bar{x}_{A\,2} - \bar{x}_{B\,2} + \bar{x}\,)^2 + (\bar{x}_{23} - \bar{x}_{A\,2} - \bar{x}_{B\,3} + \bar{x}\,)^2\,]$
$= 27\ dt^2/ha^2$

- SQ_R: zufällige Streuung bzw. Reststreuung

  Nach der Berechnung von SQ, SQ_A, SQ_B und SQ_{AB} ergibt sich die Reststreuung als

  $$SQ_R = SQ - SQ_A - SQ_B - SQ_{AB}. \qquad (21)$$

  Beispiel: $SQ_R = (747 - 117 - 450 - 27)\ dt^2/ha^2$
  $= 153\ dt^2/ha^2$

3.2.3 Streuungen in empirische Varianzen umrechnen

Wie schon im Abschnitt 2.2.3 angeführt gilt

$$\text{empirische Varianz} = \frac{\text{Summe der quadrierten Abweichungen}}{\text{Anzahl der Freiheitsgrade}}$$

und

$$\begin{matrix}\text{Anzahl der}\\ \text{Freiheitsgrade}\end{matrix} = \begin{matrix}\text{Anzahl der voneinander unabhängigen Werte, die in}\\ \text{die Berechnung der betreffenden Größe eingehen}\end{matrix} - \begin{matrix}\text{Anzahl der ab-}\\ \text{hängigen Werte.}\end{matrix}$$

Im Fall der zweifaktoriellen Varianzanalyse ergeben sich die folgenden Varianzen:

- S^2: unabhängig: I · J · K Merkmalswerte x_{ijk}
  abhängig: Mittelwert $\bar{x}$
  Freiheitsgrade: f = I J K – 1

- S_A^2: unabhängig: I Faktorstufenmittelwerte $\bar{x}_{A\,i}$
  abhängig: Mittelwert $\bar{x}$
  Freiheitsgrade: f_A = I – 1

  $$S_A^2 = \frac{SQ_A}{I-1} \qquad (22)$$

- S_B^2: unabhängig: J Faktorstufenmittelwerte $\bar{x}_{B\,j}$
  abhängig: Mittelwert $\bar{x}$
  Freiheitsgrade: f_B = J – 1

  $$S_B^2 = \frac{SQ_B}{J-1} \qquad (23)$$

- S_{AB}^2: unabhängig: I · J Mittelwerte $\bar{x}_{ij}$ in den Faktorstufenkombinationen
  abhängig: Faktorstufenmittelwerte $\bar{x}_{A\,i}$ und $\bar{x}_{B\,j}$, Mittelwert $\bar{x}$
  Freiheitsgrade: $f_{AB} = I\,J - f_A - f_B - 1$
  $= I\,J - (I - 1) - (J - 1) - 1$
  $= I\,J - I - J + 1$

$$= (I - 1)\,(J - 1)$$

$$s_{AB}^2 = \frac{SQ_{AB}}{(I-1)\,(J-1)} \qquad (24)$$

- s_R^2: $SQ_R = SQ - SQ_A - SQ_B - SQ_{AB}$ (Gleichung 21)

  $\Rightarrow f_R = f - f_A - f_B - f_{AB}$
  $= (I\,J\,K - 1) - (I - 1) - (J - 1) - (I - 1)\,(J - 1)$
  $= I\,J\,K - 1 - I + 1 - J + 1 - I\,J + I + J - 1$
  $= I\,J\,K - I\,J$
  $= I\,J\,(K - 1)$

$$s_R^2 = \frac{SQ_R}{I\,J\,(K-1)} \qquad (25)$$

Beispiel:

$s_A^2 = \dfrac{117\ (\mathrm{dt/ha})^2}{2 - 1} = 117\ \mathrm{dt^2/ha^2}$

$s_B^2 = \dfrac{450\ (\mathrm{dt/ha})^2}{3 - 1} = 225\ \mathrm{dt^2/ha^2}$

$s_{AB}^2 = \dfrac{27\ (\mathrm{dt/ha})^2}{(2 - 1)\,(3 - 1)} = 14\ \mathrm{dt^2/ha^2}$

$s_R^2 = \dfrac{153\ (\mathrm{dt/ha})^2}{2 \cdot 3\,(3 - 1)} = 13\ \mathrm{dt^2/ha^2}$

s_A^2, s_B^2 und s_{AB}^2 sind größer als s_R^2. Dies ist zunächst einmal nur ein Befund in der vorliegenden Stichprobe, der auch zufällig zustande gekommen sein könnte. Es könnte sich aber auch eine wesentlich weitergehende Schlussfolgerung andeuten, nämlich, dass der Erwartungswert des Ertrags von der Sorte und vom Dünger abhängt und möglicherweise eine Wechselwirkung zwischen Sorte und Dünger vorliegt, d. h. eine unterschiedlich starke Wirkung der Dünger je nach Sorte.

3.2.4 Einseitige F-Tests durchführen

Bei der zweifaktoriellen Varianzanalyse ist nicht nur die Frage zu klären, ob die beiden Faktoren jeweils für sich genommen die Zielvariable beeinflussen, sondern auch, ob es möglicherweise eine Wechselwirkung zwischen den beiden Faktoren gibt. Daher sind drei F-Tests durchzuführen.

- Hat der Faktor A einen Effekt?

  Ergibt sich auf Basis der vorliegenden Messwerte, dass die auf den Faktor zurückzuführende empirische Varianz größer als die Restvarianz ist ($s_A^2 > s_R^2$), so könnte dies darauf hinweisen, dass $\sigma_A^2 > \sigma_R^2$ gilt. Dieser Verdacht wird zur Alternativhypothese des Tests gemacht:

  H_0: $\sigma_A^2 \leq \sigma_R^2$
  H_1: $\sigma_A^2 > \sigma_R^2$

  Teststatistik: $$F_A = \frac{s_A^2}{s_R^2} \qquad (26)$$

  Falls die Nullhypothese erfüllt ist, ist die Teststatistik F-verteilt mit den Freiheitgraden $f_1 = f_A = I - 1$ und $f_2 = f_R = I\,J\,(K - 1)$.

Beispiel: $\alpha = 0{,}05$

$$F_A = \frac{117}{13} = 9{,}0$$

0,95-Quantil der F-Verteilung mit $f_1 = 2 - 1$ und $f_2 = 2 \cdot 3\,(3 - 1)$: $F_{0,95}(1;\,12)$[5] $= 4{,}7$

$F_A > F_{0,95}(1;\,12) \Rightarrow H_0$ wird verworfen.

- Hat der Faktor B einen Effekt?

H_0: $\sigma_B^2 \leq \sigma_R^2$

H_1: $\sigma_B^2 > \sigma_R^2$

Teststatistik: $$F_B = \frac{S_B^2}{S_R^2} \qquad (27)$$

Falls die Nullhypothese erfüllt ist, ist die Teststatistik F-verteilt mit den Freiheitgraden $f_1 = f_B = J - 1$ und $f_2 = f_R = I\,J\,(K - 1)$.

Beispiel: $\alpha = 0{,}05$

$$F_B = \frac{225}{13} = 17$$

0,95-Quantil der F-Verteilung mit $f_1 = 3 - 1$ und $f_2 = 2 \cdot 3\,(3 - 1)$: $F_{0,95}(2;\,12) = 3{,}9$

$F_B > F_{0,95}(2;\,12) \Rightarrow H_0$ wird verworfen.

- Gibt es einen Effekt durch eine Wechselwirkung der Faktoren A und B?

H_0: $\sigma_{AB}^2 \leq \sigma_R^2$

H_1: $\sigma_{AB}^2 > \sigma_R^2$

Teststatistik: $$F_{AB} = \frac{S_{AB}^2}{S_R^2} \qquad (28)$$

Falls die Nullhypothese erfüllt ist, ist die Teststatistik F-verteilt mit den Freiheitgraden $f_1 = f_{AB} = (I - 1)\,(J - 1)$ und $f_2 = f_R = I\,J\,(K - 1)$.

Beispiel: $\alpha = 0{,}05$

$$F_{AB} = \frac{14}{13} = 1{,}1$$

0,95-Quantil der F-Verteilung mit $f_1 = (2 - 1)\,(3 - 1)$ und $f_2 = 2 \cdot 3\,(3 - 1)$: $F_{0,95}(2;\,12) = 3{,}9$

$F_{AB} \leq F_{0,95}(2;\,12) \Rightarrow H_0$ wird beibehalten.

[5] Da in den drei F-Tests als rechte Begrenzung der Annahmeintervalle drei im Allgemeinen verschiedene Quantile der F-Verteilung zu bestimmen sind, werden diese in ihrer symbolischen Bezeichnung durch die zusätzliche Angabe der beiden Freiheitsgrade in runden Klammern unterschieden.

In diesem Beispiel aus dem Pflanzenbau ergibt sich, dass von einem Einfluss sowohl der Sorte als auch des Düngers auf den Ertrag auszugehen ist, dass die beiden Sorten im Ertrag aber nicht unterschiedlich stark darauf reagieren, welcher der drei Dünger verwendet wird.

Wie schon bei der einfaktoriellen Varianzanalyse müssten sich hier multiple Mittelwertvergleiche anschließen, um zu analysieren, welche Sorten und welche Dünger sich hinsichtlich des Ertrags signifikant von den anderen unterscheiden.

veröffentlicht im Internet unter aufgabomat.de