Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Funktionen `oneway.test()` und `aov()`

R stellt mehrere Funktionen zur Durchführung einer einfaktoriellen Varianzanalyse zur Verfügung

- `aov()`: für ein- und mehrfaktorielle Varianzanalyse, besteht im Aufruf der Funktionen `anova(lm(...))`
- `oneway.test()`: nur für einfaktorielle Varianzanalyse, kann auch den Fall inhomogener Varianzen behandeln

Grundlegend für die Anwendung beider Funktionen ist eine adäquate Eingabe der Daten, die die einzelnen Fälle zeilenweise charakterisiert

Teststärke und Versuchsplanung

Bestimmung des Stichprobenumfangs *n* pro Gruppe, um mit ca. 80% Power (Irrtumswahrscheinlichkeit 5%) einen Effekt von ± 3 Scorepunkten *relativ zur Kontrollgruppe* bei $\sigma = 4$ zeigen zu können, H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$

```
n <- 10                    # Stichprobenumfang pro Gruppe
grp <- factor(rep(1:3, c(n, n, n)),
              labels = c("control", "pro", "contra"))
beta <- c(mu = 10, a2 = 3, a3 = -3)                # a1 := 0
means <- model.matrix(~ grp) %*% beta

pval <- replicate(2000, {
  y <- means + rnorm(n + n + n, sd = 4)            # y = mu + a + e
  m <- aov(y ~ grp)
  summary(m)[[1]]$"Pr(>F)"[1]
})
mean(pval < 0.05)
```

Dateneingabe

```
dat <- read.table(header = TRUE,
                  text = "
  id group   score
   1 control  8
   2 control 12
   3 control  7
   4 control 10
   5 control 11
   6 control 12
   7 pro      7
   8 pro      9
   9 pro     15
  10 pro     13
  11 pro     11
  12 pro     16
  13 pro     12
  14 pro      8
  15 pro     13
  16 pro     16
  17 contra   4
  18 contra   5
  19 contra   6
  20 contra   3
  21 contra   8
  22 contra  10
  23 contra   3
  24 contra   9
")
```

Dateneingabe

Der Aufruf

```
read.table(filename, header = TRUE)
```

liest die Daten von Datei ein, wobei die Angabe `header = TRUE` anzeigt, dass die Datei als erste Zeile die Namen der in den darauf folgenden Zeilen aufgelisteten Variablen enthält

Durch das Argument `text` statt `filename` wird die unmittelbare Eingabe der Daten ermöglicht

Festlegen der Faktorstufen mit Funktion `factor()`

Mit dem Aufruf

```
dat$group <- factor(dat$group,
                    levels = c("control", "pro", "contra"))
```

wird die Anordnung der Faktorstufen festgelegt, deren Ausprägungen als Teil des Dataframes `dat` durch `dat$group` angesprochen werden

Diese Festlegung ist erforderlich, da R die Faktorstufen ansonsten anhand ihrer Namen alphabetisch sortiert

Funktion `oneway.test()`

Die varianzanalytische Auswertung der Daten mit `oneway.test()` erhält man über den Aufruf

```
oneway.test(score ~ group, data = dat, var.equal = TRUE)
```

Mit der Option `var.equal = TRUE` wird die Annahme der Varianzhomogenität zugrunde gelegt

Man erhält folgende Ausgabe

```
    One-way analysis of means

data:  score and group
F = 10.125, num df = 2, denom df = 21, p-value = 0.0008
```

Funktion aov()

Die varianzanalytische Auswertung der Daten mit `aov()` und die Ausgabe der Ergebnisse erhält man über die Kommandos

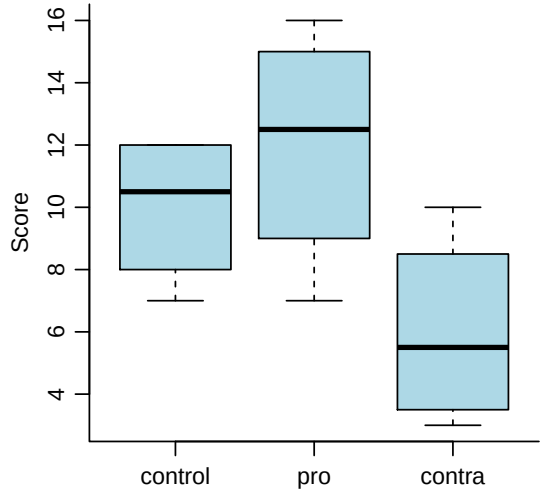
```
m <- aov(score ~ group, data = dat)
summary(m)
```

Man erhält folgende Ausgabe

```
             Df Sum Sq Mean Sq F value Pr(>F)
group         2    162      81  10.125 0.0008
Residuals    21    168       8
```

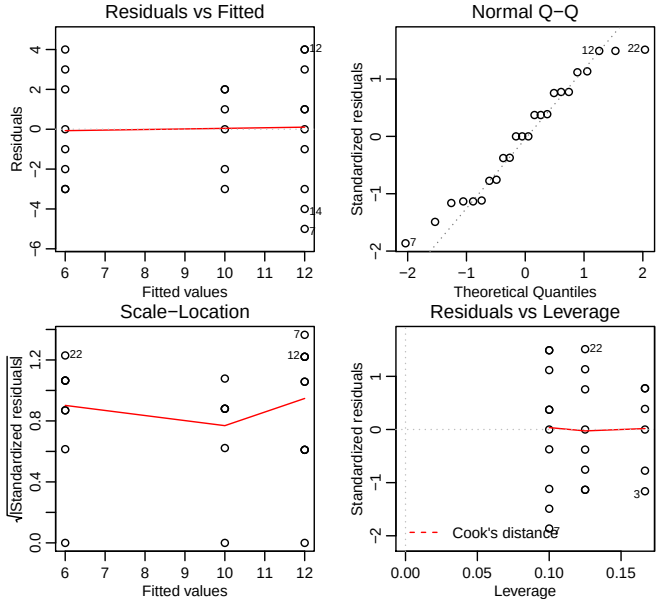
Grafische Darstellung der Ergebnisse

```
boxplot(score ~ group,
        data = dat)
```



Modelldiagnose

```
par(mfrow = c(2, 2))
plot(m)
```



Voraussetzungen

Die Annahme der Unabhängigkeit kann verletzt sein, beispielsweise bei wiederholter Messung derselben Variable an derselben Untersuchungseinheit (z. B. Versuchsperson)

- Dem ist gegebenenfalls durch Einführung eines Messwiederholungsfaktors (Within-Subjects-Faktor) Rechnung zu tragen (wird noch besprochen)

Bei einer antizipierten Verletzung der Normalverteilungsannahme gibt es entsprechende verteilungsfreie Tests, wie beispielsweise den Kruskal-Wallis-Test als Verallgemeinerung des Wilcoxon-Rangsummen-Tests (Funktion `kruskal.test()` in R)

Bei der einfaktoriellen Varianzanalyse erhält man unter der Annahme zufälliger Effekte rechnerisch dieselben Ergebnisse, lediglich die Interpretation unterscheidet sich vom betrachteten Fall mit festen Effekten

Einzelvergleiche

Die Ablehnung der Nullhypothese belegt, dass zwischen den Gruppen Unterschiede bestehen

Häufig möchte man diese Unterschiede genauer lokalisieren, d. h. feststellen, zwischen welchen Gruppen statistisch signifikante Unterschiede bestehen

Diese paarweisen Vergleiche sind entweder

- geplante Vergleiche (stehen bereits vor der Datenerhebung fest)
- Post-Hoc-Vergleiche (nach Kenntnis der Daten ausgewählt)

Gültige Schlussfolgerungen sind nur möglich, wenn die Vergleiche vor der Datenerhebung festgelgt werden

Die möglicherweise große Zahl dieser paarweisen Vergleiche führt zur Problematik multipler Tests

Zweifaktorielle Varianzanalyse mit festen Effekten

Einführung

Bei der Betrachtung zweier Einflussfaktoren stellt sich die Frage, ob diese sich gemeinsam anders auswirken, als jede der Variablen einzeln

Das varianzanalytische Modell hat daher neben den Effekten der einzelnen Variablen auch die Möglichkeit von deren gegenseitigen Beeinflussung zu erfassen

- Eine derartige Wechselwirkung (Interaktion) liegt vor, wenn sich die Stufe eines Faktors je nach der vorliegenden Ausprägung des anderen Faktors unterschiedlich auswirkt
- Davon zu unterscheiden ist die Wirksamkeit eines Faktors unabhängig vom jeweils anderen Faktor (Haupteffekte)

Der Grund der mehrfaktoriellen Auswertung ist in der Regel das Interesse an der Stärke des Interaktionseffekts

Beispiel

Als Einflussfaktoren auf die Zufriedenheit mit dem Studium werden betrachtet

- Motivation im Studium (Faktor A) mit den Stufen "motiviert" (A_1) und "unmotiviert" (A_2)
- Familiäre Situation (Faktor B) mit den Stufen "mit einem Partner lebend" (B_1) und "allein lebend" (B_2)

Die Zufriedenheit mit dem Studium wird über einen Score gemessen, der die Werte 0 bis 100 annimmt und so konstruiert ist, dass er als approximativ normalverteilt angesehen werden kann

Es soll überprüft werden, ob der Motivationseffekt auf die Zufriedenheit bei mit einem Partner Lebenden (minimal 15 Scorepunkte) stärker ist als bei allein Lebenden

Beispiel

Nach einer Poweranalyse (s. u.) wurden pro Kombination der beiden Faktoren fünf Studierende zufällig ausgewählt und deren Scores erhoben

Motivation	familiäre Situation	
	Partner	allein lebend
motiviert	85, 89, 91, 95, 80	50, 52, 65, 71, 72
unmotiviert	34, 30, 28, 23, 40	30, 28, 33, 16, 23

Datenschema

Faktor A	Faktor B				
	1	$\dots$	j	$\dots$	J
1					
$\vdots$					
i			$y_{ij1}, \dots,$ $y_{ijk},$ $\dots, y_{ijK}$		
$\vdots$					
I					

Modell: Alternative I

Die Y_{ijk} ergeben sich aus einer durch die jeweilige Faktorkombination (i, j) mit $i = 1, \ldots, I$, $j = 1, \ldots, J$ induzierten mittleren Ausprägung μ_{ij} und unabhängigen individuellen Abweichungen von diesem Wert, die für $k = 1, \ldots, K$ durch ε_{ijk} bezeichnet werden

Damit ergibt sich das Modell

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

mit $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ und $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Daraus folgt

$$E(Y_{ijk}) = E(\mu_{ij} + \varepsilon_{ijk}) = \mu_{ij}$$

und $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$

Modell: Alternative II

Der Erwartungswert μ_{ij} wird zerlegt in den Einfluss des globalen Erwartungswerts μ, die Effekte der einzelnen Faktoren und den durch deren gegenseitigen Beeinflussung bewirkten Effekt

Damit ergibt sich das Modell

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

mit $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ und $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Die ε_{ijk} werden auch in dieser Modellformulierung als unabhängig vorausgesetzt

Modell: Alternative II

Analog zum Modell im einfaktoriellen Fall (bei gleichen Zellenbesetzungen) ergibt sich

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{i=1}^{I} (\alpha\beta)_{ij} = 0, \quad \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$$

Weiter gilt

$$\mu = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$$

Haupteffekte

Der Effekt des Faktors A (Haupteffekt) wird durch α_i beschrieben, d. h. es gilt

$$\alpha_i = \mu_{i.} - \mu$$

mit $\mu_{i.} = \frac{1}{J} \sum_{j=1}^{J} \mu_{ij}$ der Erwartungswert des Faktors A auf der Stufe i

Entsprechend beschreibt β_j den Effekt des Faktors B (Haupteffekt) durch

$$\beta_j = \mu_{.j} - \mu$$

mit $\mu_{.j} = \frac{1}{I} \sum_{i=1}^{I} \mu_{ij}$ der Erwartungswert des Faktors B auf der Stufe j

Interaktionseffekt

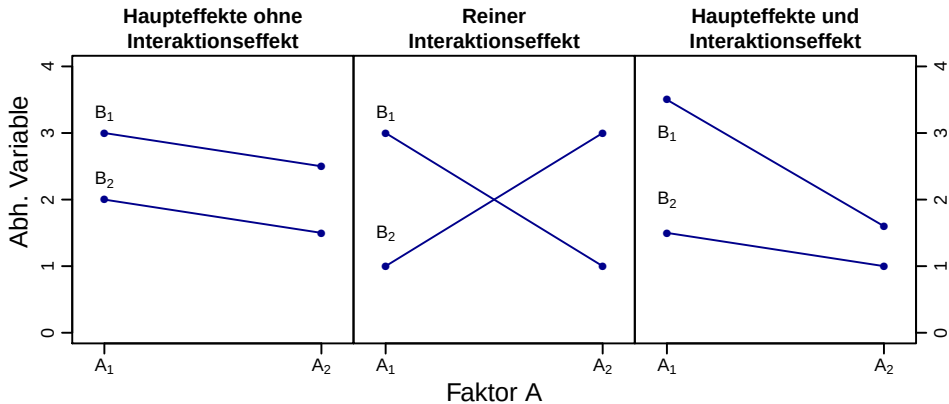
Die Wechselwirkung der Faktoren A und B (Interaktionseffekt) auf der Faktorstufenkombination (i, j) wird durch $(\alpha\beta)_{ij}$ beschrieben, d. h. es gilt

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$$

Damit beschreibt der Parameter $(\alpha\beta)_{ij}$ die Abweichung des Erwartungswerts μ_{ij} der Zelle (i, j) vom Grand Mean μ, die nicht durch Addition der beiden Haupteffekte α_i und β_j erklärt wird

Im Folgenden werden im Rahmen eines (2×2)-Designs reine Haupteffekte, reine Interaktionseffekte und die Kombination beider Effekte schematisch dargestellt

Effekte im (2×2)-Design



Parameterschätzung

Grand Mean μ

$$\hat{\mu} = \frac{1}{I\,J\,K} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk} = \bar{Y}_{...}$$

Haupteffekt α_i

$$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \quad \text{mit} \quad \bar{Y}_{i..} = \frac{1}{J\,K} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk}$$

Haupteffekt β_j

$$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...} \quad \text{mit} \quad \bar{Y}_{.j.} = \frac{1}{I\,K} \sum_{i=1}^{I} \sum_{k=1}^{K} Y_{ijk}$$

Parameterschätzung

Interaktionseffekt $(\alpha\beta)_{ij}$

Mit

$$\bar{Y}_{ij.} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijk}$$

gelten folgende Gleichungen

$$\begin{aligned}
\widehat{(\alpha\beta)}_{ij} &= \bar{Y}_{ij.} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) \\
&= \bar{Y}_{ij.} - (\bar{Y}_{...} + \bar{Y}_{i..} - \bar{Y}_{...} + \bar{Y}_{.j.} - \bar{Y}_{...}) \\
&= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}
\end{aligned}$$

Residuen ε_{ijk}

$$\hat{\varepsilon}_{ijk} = Y_{ijk} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij})$$

Beispiel

Schätzung des Grand Mean μ

$$\hat{\mu} = \frac{1}{I\,J\,K}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} = \frac{1}{2\cdot 2\cdot 5}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} = 51.75$$

Mittelwerte der Stufen des Faktors A

$$\bar{y}_{1..} = \frac{1}{10}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{1jk} = 75, \qquad \bar{y}_{2..} = \frac{1}{10}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{2jk} = 28.5$$

Mittelwerte der Stufen des Faktors B

$$\bar{y}_{.1.} = \frac{1}{10}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{i1k} = 59.5, \qquad \bar{y}_{.2.} = \frac{1}{10}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{i2k} = 44$$

Beispiel

Mittelwerte der Kombinationen der Faktorstufen von A und B

$$\bar{y}_{11\cdot} = \frac{1}{5}\sum_{k=1}^{K} y_{11k} = 88, \qquad \bar{y}_{12\cdot} = \frac{1}{5}\sum_{k=1}^{K} y_{12k} = 62$$

und

$$\bar{y}_{21\cdot} = \frac{1}{5}\sum_{k=1}^{K} y_{21k} = 31, \qquad \bar{y}_{22\cdot} = \frac{1}{5}\sum_{k=1}^{K} y_{22k} = 26$$

Mit diesen Größen lassen sich nun die Haupt- und Interaktionseffekte schätzen

Beispiel

$$\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{...} = 75 - 51.75 = 23.25$$
$$\hat{\alpha}_2 = \bar{y}_{2..} - \bar{y}_{...} = 28.5 - 51.75 = -23.25$$
$$\hat{\beta}_1 = \bar{y}_{.1.} - \bar{y}_{...} = 59.5 - 51.75 = 7.75$$
$$\hat{\beta}_2 = \bar{y}_{.2.} - \bar{y}_{...} = 44 - 51.75 = -7.75$$
$$\widehat{(\alpha\beta)}_{11} = \bar{y}_{11.} - \bar{y}_{1..} - \bar{y}_{.1.} + \bar{y}_{...} = 88 - 75 - 59.5 + 51.75 = 5.25$$
$$\widehat{(\alpha\beta)}_{12} = \bar{y}_{12.} - \bar{y}_{1..} - \bar{y}_{.2.} + \bar{y}_{...} = 62 - 75 - 44 + 51.75 = -5.25$$
$$\widehat{(\alpha\beta)}_{21} = \bar{y}_{21.} - \bar{y}_{2..} - \bar{y}_{.1.} + \bar{y}_{...} = 31 - 28.5 - 59.5 + 51.75 = -5.25$$
$$\widehat{(\alpha\beta)}_{22} = \bar{y}_{22.} - \bar{y}_{2..} - \bar{y}_{.2.} + \bar{y}_{...} = 26 - 28.5 - 44 + 51.75 = 5.25$$

Hypothesen

Vorliegen von Interaktionen

Nullhypothese

$$H_0^{A\times B}:\ (\alpha\beta)_{ij} = 0 \text{ für alle } i = 1,\ldots,I,\ j = 1,\ldots,J$$

Alternativhypothese

$$H_1^{A\times B}:\ (\alpha\beta)_{ij} \neq 0 \text{ für mindestens ein Paar } (i,j)$$

Hypothesen

Vorliegen von Haupteffekten des Faktors A

Nullhypothese

$$H_0^A: \ \alpha_i = 0 \text{ für alle } i = 1, \ldots, I$$

Alternativhypothese

$$H_1^A: \ \alpha_i \neq 0 \text{ für mindestens ein } i$$

Vorliegen von Haupteffekten des Faktors B

Nullhypothese

$$H_0^B: \ \beta_j = 0 \text{ für alle } j = 1, \ldots, J$$

Alternativhypothese

$$H_1^B: \ \beta_j \neq 0 \text{ für mindestens ein } j$$

Streuungszerlegung

Zur Herleitung geeigneter Teststatistiken zur Überprüfung dieser Hypothesen wird – wie in der einfaktoriellen Varianzanalyse – die Gesamtsumme der Abweichungsquadrate SS_T zerlegt in

- die durch den Faktor A bedingte Summe der Abweichungsquadrate SS_A
- die durch den Faktor B bedingte Summe der Abweichungsquadrate SS_B
- die durch die Interaktion bedingte Summe der Abweichungsquadrate $SS_{A\times B}$
- das Residuum der Summe der Abweichungsquadrate SS_W innerhalb der Gruppen

$$SS_T = SS_A + SS_B + SS_{A\times B} + SS_W$$

Streuungszerlegung

Dabei ist

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{...})^2$$

$$SS_A = K \cdot J \cdot \sum_{i=1}^{I}(\bar{Y}_{i..} - \bar{Y}_{...})^2 = K \cdot J \cdot \sum_{i=1}^{I}\hat{\alpha}_i^2$$

$$SS_B = K \cdot I \cdot \sum_{j=1}^{J}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 = K \cdot I \cdot \sum_{j=1}^{J}\hat{\beta}_j^2$$

Streuungszerlegung

Weiter gilt

$$SS_{A\times B} = K \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 = K \cdot \sum_{i=1}^{I}\sum_{j=1}^{J}(\widehat{\alpha\beta})_{ij}^2$$

$$SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{ij.})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(K-1)S_{ij}^2$$

mit der Stichprobenvarianz S_{ij}^2 der Werte in Zelle (i, j)

Mittlere Quadrate

Normierung auf die Anzahl der Freiheitsgrade liefert dann die mittleren Abweichungsquadrate

$$MS_A = \frac{SS_A}{I-1}, \qquad MS_B = \frac{SS_B}{J-1}$$

und

$$MS_{A\times B} = \frac{SS_{A\times B}}{(I-1)(J-1)}, \qquad MS_W = \frac{SS_W}{I\,J\,(K-1)}$$

Für die Summe der Freiheitsgrade hat man

$$I-1+J-1+(I-1)(J-1)+I\,J\,(K-1) = I\,J\,K-1 = N-1$$

Erwartete mittlere Quadrate

Die Erwartungswerte der mittleren Abweichungsquadrate lauten dann

$$E(MS_A) = \sigma^2 + \frac{K\,J}{I-1}\sum_{i=1}^{I}\alpha_i^2, \quad E(MS_B) = \sigma^2 + \frac{K\,I}{J-1}\sum_{j=1}^{J}\beta_j^2$$

und

$$E(MS_{A\times B}) = \sigma^2 + \frac{K}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha\beta)_{ij}^2, \quad E(MS_W) = \sigma^2$$

Unter H_0^A sind daher MS_A und MS_W unabhängige Schätzer für σ^2, entsprechendes gilt unter H_0^B für MS_B und MS_W und unter $H_0^{A\times B}$ für $MS_{A\times B}$ und MS_W

Teststatistiken

Statistischer Test von $H_0^{A\times B}$ vs. $H_1^{A\times B}$

Teststatistik

$$F_{A\times B} = \frac{MS_{A\times B}}{MS_W}$$

Verteilung unter $H_0^{A\times B}$

$$F_{A\times B} \sim F((I-1)(J-1),\ IJ(K-1))$$

Ablehnungsbereich

$$F_{A\times B} > F_{1-\alpha}((I-1)(J-1),\ IJ(K-1))$$

Teststatistiken

Statistischer Test von H_0^A vs. H_1^A

Teststatistik

$$F_A = \frac{MS_A}{MS_W}$$

Verteilung unter H_0^A

$$F_A \sim F(I-1,\ IJ(K-1))$$

Ablehnungsbereich

$$F_A > F_{1-\alpha}(I-1,\ IJ(K-1))$$

Statistischer Test von H_0^B vs. H_1^B

Teststatistik

$$F_B = \frac{MS_B}{MS_W}$$

Verteilung unter H_0^B

$$F_B \sim F(J-1,\ IJ(K-1))$$

Ablehnungsbereich

$$F_B > F_{1-\alpha}(J-1,\ IJ(K-1))$$