Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(6) Einfaktorielle Varianzanalyse

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Anwendungsszenario und Datendeskription

Anwendungsszenario

- Zwei oder mehr Gruppen experimenteller Einheiten mit Datendimension $m > 1$
- Annahme der unabhängigen und identischen Normalverteilung $N(\mu_i, \Sigma)$ der Daten
- $\mu_i \in \mathbb{R}^m, i = 1, ..., p$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd unbekannt
- Absicht des inferentiellen Testens der Nullhypothese identischer Gruppenerwartungswerte

Generalisierung des Zweistichproben-T^2-Tests bei unabhängigen Stichproben mit einfacher Nullhypothese für mehr als zwei Stichproben

Anwendungsbeispiel

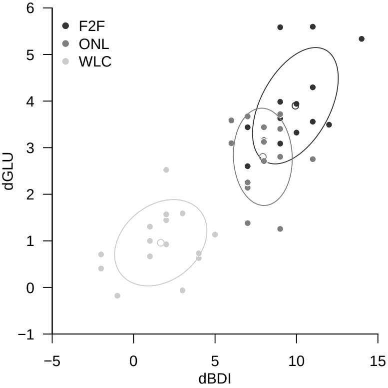
- Analyse von Daten dreier Therapiegruppen (F2F, ONL, WLC) von `dBDI` und `dGLU` Daten
- Testen der Nullhypothese $\mu_1 = \mu_2 = \mu_3$

Table 1: Prä-Post-Interventions-BDI-II-Score und -Glukokortikoidplasmalevel Differenzenwerte von drei Studiengruppen (F2F: Face-to-face, ONL: online, WLC: waitlist control) jeweils 15 Patient:innen. Die Tabelle zeigt exemplarisch die ersten fünf Datenpunkte jeder Gruppe.

	COND	dBDI	dGLU
1	F2F	11	4.3
2	F2F	10	3.9
3	F2F	12	3.5
4	F2F	7	2.6
5	F2F	10	3.3
16	ONL	6	3.1
17	ONL	8	2.7
18	ONL	7	2.1
19	ONL	8	3.1
20	ONL	11	2.8
31	WLC	-2	0.7
32	WLC	2	1.4
33	WLC	1	1.0
34	WLC	2	0.9
35	WLC	3	1.6

Anwendungsszenario und Datendeskription

Anwendungsbeispiel



Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Definition (Modell der einfaktoriellen multivariaten Varianzanalyse)

Für $i = 1, ..., p$ und $j = 1, ..., n_i$ seien υ_{ij} m-dimensionale Zufallsvektoren, die die $n := \sum_{i=1}^{p} n_i$ m-dimensionalen Datenpunkte eines einfaktoriellen multivariaten Varianzanalyseszenarios modellieren. Dann hat das Modell der einfaktoriellen multivariaten Varianzanalyse die strukturelle Form

$$\upsilon_{ij} = \mu_i + \varepsilon_{ij} \text{ mit } \varepsilon_{ij} \sim N(0_m, \Sigma) \text{ u.i.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd} \tag{1}$$

und die Datenverteilungsform

$$\upsilon_{ij} \sim N(\mu_i, \Sigma) \text{ u.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd.} \tag{2}$$

Bemerkungen

- Der Einfachheit halber setzen wir meist identische Gruppengrößen $k := n_i$ voraus.
- Die Gesamtheit aller Datenzufallsvektoren bezeichnen wir mit Υ.

Theorem (Parameterschätzer)

Gegeben sei das Modell der einfaktoriellen multivariaten Varianzanalyse. Dann ist für $i = 1, ..., p$

$$\hat{\mu}_i := \frac{1}{n_i}\sum_{j=1}^{n_i} \upsilon_{ij} \tag{3}$$

ein unverzerrte Schätzer des gruppenspezifischen Erwartungswertparameters μ_i *und*

$$\hat{\Sigma} := \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\upsilon_{ij} - \hat{\mu}_i\right)\left(\upsilon_{ij} - \hat{\mu}_i\right)^T \tag{4}$$

ein unverzerrrter Schätzer des Kovarianzmatrixparameters Σ.

Bemerkungen

- $\hat{\mu}_i$ ist das Stichprobenmittel der $i$ten Gruppe.
- $\hat{\Sigma}$ ist die mit $1/(n-p)$ skalierte Within-Group Sum-of-Squares Matrix (siehe unten).
- Anstelle eines Beweis validieren wir die Aussage des Theorems mithilfe einer Simulation.

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Modellevaluation mit der Wilks'-Λ-Statistik

Überblick

- Ziel einer einfaktoriellen multivariaten Varianzanalyse ist meist das Testen der Nullhypothese

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{5}$$

- Es ergibt sich damit die Alternativhypothese

$$\begin{aligned} H_1 : \mu_{i_l} \neq \mu_{j_l} \text{ für mindestens ein Paar } i, j \text{ mit } i \neq j, 1 \leq i, j \leq p \\ \text{und mindestens ein } l \text{ mit } 1 \leq l \leq m. \end{aligned} \tag{6}$$

- Zur Evaluation von H_0 wurden eine Reihe von Teststatistiken vorgeschlagen.
- Alle Teststatistiken beruhen auf der Kreuzproduktsummenmatrixzerlegung.
- Wir betrachten hier exemplarisch die Wilks'-Λ-Statistik.
- Die Verteilungen der Wilks'-Λ-Statistik unter der Nullhypothese sind zum Teil analytisch angebbar, zum Teil müssen sie approximiert werden und haben die Form von f-Verteilungen.

Theorem (Kreuzproduktsummenmatrizenzerlegung)

Für $i = 1, ..., p$ und $j = 1, ..., n_i$ bezeichne υ_{ij} den $j$ten Stichprobenvektor der $i$ten Stichprobengruppe eines einfaktoriellen multivariaten Varianzanalysemodells. Weiterhin seien

$$\bar{\upsilon} := \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\upsilon_{ij} \text{ und } \bar{\upsilon}_i := \frac{1}{n_i}\sum_{j=1}^{n_i}\upsilon_{ij} \tag{7}$$

das *Gesamtstichprobenmittel* und das i*te Gruppenstichprobenmittel*, respektive. Schließlich seien

$T := \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\upsilon_{ij} - \bar{\upsilon}\right)\left(\upsilon_{ij} - \bar{\upsilon}\right)^T$ die *Totale Sum-of-Squares Matrix*

$B := \sum_{i=1}^{p} n_i\left(\bar{\upsilon}_i - \bar{\upsilon}\right)\left(\bar{\upsilon}_i - \bar{\upsilon}\right)^T$ die *Between-Group Sum-of-Squares Matrix*

$W := \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\upsilon_{ij} - \bar{\upsilon}_i\right)\left(\upsilon_{ij} - \bar{\upsilon}_i\right)^T$ die *Within-Group Sum-of-Squares Matrix*.

Dann gilt

$$T = B + W. \tag{8}$$

Bemerkungen

$T \in \mathbb{R}^{m\times m}$ misst die totale Variabilität der Datenvektoren um das Gesamtstichprobenmittel.

$B \in \mathbb{R}^{m\times m}$ misst die Variabilität der Gruppenstichprobenmittel um das Gesamtstichprobenmittel.

$W \in \mathbb{R}^{m\times m}$ misst die Variabilität der Datenvektoren um ihre jeweiligen Gruppenstichprobenmittel.

Die totale Variabilität wird hier also in zwei unabhängige Beiträge von Variabilität zerlegt.

W heißt auch *Residualvariabiliät*, weil sie die verbleibende Variabilität nach Schätzung der Gruppenerwartungswertparameter quantifiziert und gilt das

$$W = (n-p)\hat{\Sigma}. \tag{9}$$

Ein Beweis gibt sich durch algebraische Umformung.

Modellevaluation mit der Wilks'-Λ-Statistik

<u>Beweis</u>

$$
\begin{aligned}
T &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon})(\upsilon_{ij}-\bar{\upsilon})^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i+\bar{\upsilon}_i-\bar{\upsilon})(\upsilon_{ij}-\bar{\upsilon}_i+\bar{\upsilon}_i-\bar{\upsilon})^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\big((\upsilon_{ij}-\bar{\upsilon}_i)+(\bar{\upsilon}_i-\bar{\upsilon})\big)\big((\upsilon_{ij}-\bar{\upsilon}_i)+(\bar{\upsilon}_i-\bar{\upsilon})\big)^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\big((\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+2(\upsilon_{ij}-\bar{\upsilon}_i)(\bar{\upsilon}_i-\bar{\upsilon})^T+(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\big) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+\sum_{j=1}^{n_i}2(\upsilon_{ij}-\bar{\upsilon}_i)(\bar{\upsilon}_i-\bar{\upsilon})^T+\sum_{j=1}^{n_i}(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+2\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)\right)(\bar{\upsilon}_i-\bar{\upsilon})^T+n_i(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+2\left(\sum_{j=1}^{n_i}\left(\upsilon_{ij}-\frac{1}{n_i}\sum_{j=1}^{n_i}\upsilon_{ij}\right)\right)(\bar{\upsilon}_i-\bar{\upsilon})^T+n_i(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+2\left(\sum_{j=1}^{n_i}\upsilon_{ij}-\sum_{j=1}^{n_i}\upsilon_{ij}\right)(\bar{\upsilon}_i-\bar{\upsilon})^T+n_i(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T+n_i(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T\right) \\
&= \sum_{i=1}^{p}n_i(\bar{\upsilon}_i-\bar{\upsilon})(\bar{\upsilon}_i-\bar{\upsilon})^T+\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\upsilon_{ij}-\bar{\upsilon}_i)(\upsilon_{ij}-\bar{\upsilon}_i)^T \\
&= B+W.
\end{aligned}
\tag{10}
$$

Modellevaluation mit der Wilks'-Λ-Statistik

Definition (Wilks'-Λ-Statistik)

Es seien das Modell der einfaktoriellen Varianzanalyse sowie die Between Sum of Squares Matrix B und die Within Sum-of-Squares Matrix W definiert wie oben. Dann ist die Wilks'-Λ-Statistik Teststatistik definiert als

$$\Lambda := \frac{|W|}{|B+W|}, \tag{11}$$

wobei $|\cdot|$ die Determinante bezeichnet.

Bemerkungen

- Intuitiv misst Λ das Verhältnis von Residualvariabilität und Gesamtvariabilität.
- Ohne Beweis halten wir fest, dass $\Lambda \in [0,1]$
- Für $\bar{\upsilon}_1 = \cdots = \bar{\upsilon}_p = \bar{\upsilon}$ gilt $B = 0_{mm}$ und damit $\Lambda = 1$.
- Für steigende Unterschiede zwischen den $\bar{\upsilon}_i$ nimmt $|B+W|$ gegenüber $|W|$ zu, Λ also ab.
- Kleine Werte von Λ sprechen also für eine Abweichung von der Nullhypothese.

Theorem (Spezielle H_0 Verteilungen von Λ Transformationen)

Es seien das Modell der einfaktoriellen Varianzanalyse und Wilks'-Λ-Statistik definiert wie oben und es gelte außerdem

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{12}$$

Dann sind für die in den ersten beiden Tabellenspalten aufgeführten Spezialfällen die in der dritten Tabellenspalte aufgeführten Statistiken f-Zufallsvariablen und zwar mit den in der vierten Tabellenspalte aufgeführten Parametern.

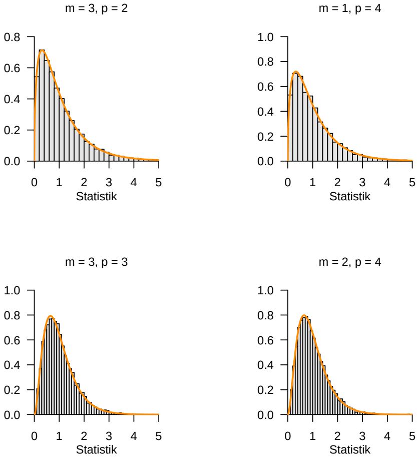
Datendimension m	Gruppenanzahl p	Statistik	f-Verteilungsparameter
Beliebig	2	$\frac{1-\Lambda}{\Lambda}\frac{n-p-m+1}{m}$	$m, n-p-m+1$
Beliebig	3	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-p-m+1}{m}$	$2m, 2(n-p-m+1)$
1	Beliebig	$\frac{1-\Lambda}{\Lambda}\frac{n-p}{p-1}$	$p-1, n-p$
2	Beliebig	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-p-1}{p-1}$	$2(p-1), 2(n-p-1)$

Bemerkungen

- Die Verteilungen gehen zurück auf Wilks (1932).

Modellevaluation mit der Wilks'-Λ-Statistik

Simulation spezieller H_0 Verteilungen von Wilks'-Λ-Statistik Transformationen



Modellevaluation mit der Wilks'-Λ-Statistik

Theorem (Approximative H_0 Verteilungen von Λ Transformationen)

Es seien das Modell der einfaktoriellen Varianzanalyse und Wilks'-Λ-Statistik definiert wie oben und es gelte außerdem

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{13}$$

Dann ist die Statistik

$$\tau := \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\frac{\nu_2}{\nu_1} \tag{14}$$

mit

$$\nu_1 := m(p-1) \text{ und } \nu_2 := wt - \frac{1}{2}(m(p-1)-2) \tag{15}$$

sowie

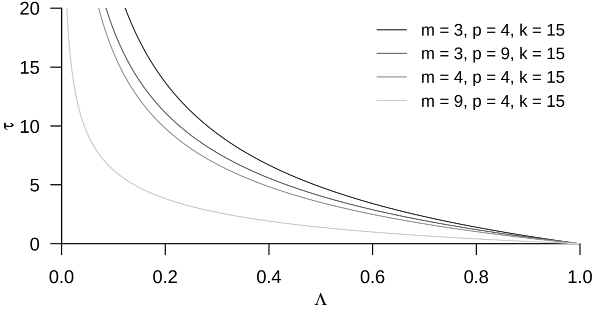
$$w := n-1-\frac{1}{2}(m+k) \text{ und } t := \sqrt{\frac{m^2(p-1)^2-4}{m^2+(p-1)^2-5}} \tag{16}$$

approximativ f-verteilt mit Freiheitsgradparametern ν_1 und ν_2.

Bemerkungen

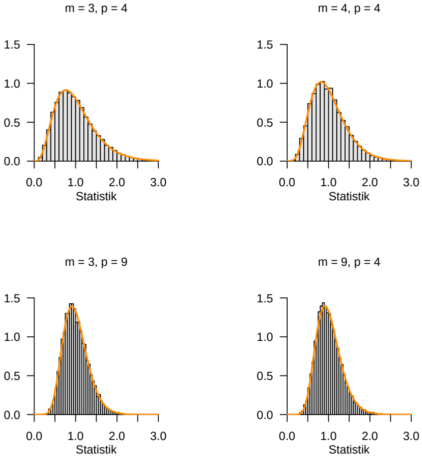
- Die Approximation geht zurück auf Rao (1951).

τ als Funktion von Λ : $\Lambda \downarrow \Rightarrow \tau \uparrow$



Modellevaluation mit der Wilks'-Λ-Statistik

Simulation approximativer H_0 Verteilungen von Wilks'-Λ-Statistik Transformationen



Modellevaluation mit der Wilks'-Λ-Statistik

Theorem (Wilks'-Λ-Statistik-basierter Test)

Es seien das Modell der einfaktoriellen Varianzanalyse und die Wilks'-Λ-Statistik basierte Teststatistik τ mit Verteilungsparametern ν_1, ν_2 wie oben definiert. Weiterhin sei der kritische Wert-basierte Test

$$\phi(\Upsilon) := 1_{\{\tau > k\}} := \begin{cases} 1 & \tau > k \\ 0 & \tau \le k \end{cases} \tag{17}$$

definiert. ϕ ist genau dann ein Level-α_0-Test mit Testumfang α, wenn

$$k := k_{\alpha_0} := F^{-1}(1 - \alpha_0; \nu_1, \nu_2) \tag{18}$$

ist und der p-Wert einer realisiertern τ-Teststatistik $\tilde{\tau}$ ergibt sich zu

$$\text{p-Wert} = \mathbb{P}(\tau \ge \tilde{\tau}) = 1 - F(\tilde{\tau}; \nu_1, \nu_2) \tag{19}$$

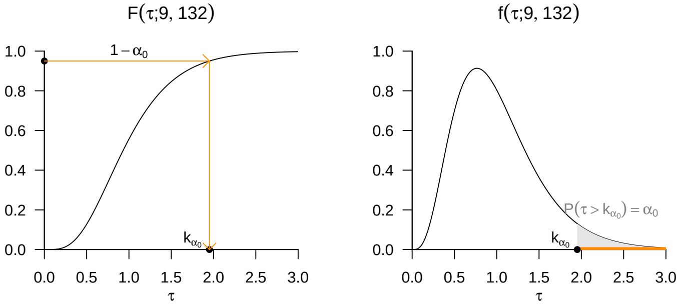
Bemerkungen

- Ein Beweis kann in Analogie zum Einstichproben-T^2-Test Fall geführt werden.
- Wir validieren die Testumfangkontrolle mithilfe von k_{α_0} in untenstehender Simulation.

Modellevaluation mit der Wilks'-Λ-Statistik

Testumfangkontrolle

Wahl von k_{α_0} bei $m = 3, p = 4, k = 15 \Rightarrow \nu_1 = 9, \nu_2 = 132$ und $\alpha_0 = 0.05$.



Modellevaluation mit der Wilks'-Λ-Statistik

Praktisches Vorgehen

- Man nimmt an, dass ein vorliegender Datensatz von p Gruppen von m-dimensionalen Datenvektoren für jeweils $j = 1, ..., n_i$ Realisationen von $\upsilon_{ij} \sim N(\mu_i, \Sigma)$ mit unbekannten Parametern $\mu_i \in \mathbb{R}^m, i = 1, ..., p$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd sind.
- Man möchte entscheiden ob $H_0 : \mu_1 = \cdots = \mu_p$ eher zutrifft oder eher nicht.
- Man wählt ein Signifikanzniveau α_0 und bestimmt den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0}. Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05, m = 3, p = 4, k = 15$ für $i = 1, ...p$ und somit $n = 60$ sowie $\nu_1 = 9, \nu_2 = 132$, dass $k_{\alpha_0} = F^{-1}(1 - 0.05; 9, 132) \approx 1.95$ ist.
- Anhand der Wilk's Λ Statistik sowie m, p und k berechnet man man den realisierten Wert der τ-Teststatistik, den wir hier mit $\tilde{\tau}$ bezeichnen.
- Wenn $\tilde{\tau}$ größer als k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht.
- Die oben entwickelte Theorie garantiert dann, dass man im Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.
- Schließlich ergibt sich der assoziierte p-Wert der realisiertern τ-Teststatistik $\tilde{\tau}$ zu

$$\text{p-Wert} = \mathbb{P}(\tau \geq \tilde{\tau}) = 1 - F(\tilde{\tau}; \nu_1, \nu_2) \tag{20}$$

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Selbstkontrolfragen

1. Erläutern Sie das Anwendungsszenario einer einfaktoriellen multivariaten Varianzanalyse.
2. Geben Sie die Definition des Modells der einfaktoriellen multivariaten Varianzanalyse wieder.
3. Geben Sie das Theorem zu den Parameterschätzern der einfaktoriellen multivariaten Varianzanalyse wieder.
4. Erläutern Sie die Null- und Alternativhypothesen einer einfaktoriellen multivariaten Varianzanalyse.
5. Geben Sie das Theorem zur Kreuzproduktsummenmatrizenzerlegung wieder.
6. Was messen die Totale, Between-Group und die Within-Group Sum-of-Squares Matrizen, respektive?
7. Geben Sie die Definition der Wilks'-Λ-Statistik wieder.
8. Erläutern Sie Gemeinsamkeiten und Unterschiede zwischen speziellen und approximativen H_0 Verteilungen von Wilks-Λ-Transformationen bei der einfaktoriellen multivariaten Varianzanalyse.
9. Geben Sie das Theorem zum Wilks-Λ-Statistik-basierten Test im Rahmen der einfaktoriellen multivariaten Varianzanalyse wieder.
10. Erläutern Sie das praktische Vorgehen zur Durchführung eines Wilks-Λ-Statistik-basierten Tes tim Rahmen der einfaktoriellen multivariaten Varianzanalyse.

Referenzen

Rao, C. Radhakrishna. 1951. "An Asymptotic Expansion of the Distribution of Wilk's Criterion." *Bulletin of the International Statistical Institute* 33 (2): 177–80.

Wilks, S. S. 1932. "Certain Generalizations in the Analysis of Variance." *Biometrika* 24 (3-4): 471–94. https://doi.org/10.1093/biomet/24.3-4.471.