SAS-Macros zum Hotellingschen T^2-Test und zur Stichprobenplanung bei einfaktorieller MANOVA

Erich Schumacher

Institut für Angewandte Mathematik und Statistik
Universität Hohenheim, 70599 Stuttgart
eMail: schumach@uni-hohenheim.de

Einleitung

Das SAS-Macro *hotellin.mac* gestattet, neben der Durchführung des Hotellingschen T^2-Tests unter Verwendung einer geeigneten nichtzentralen F-Verteilung approvimativ die Güte des Tests zu berechnen, siehe Läuter [1978b]. Ausserdem werden Tests zur Überprüfung der Modellvoraussetzungen durchgeführt. Nach Mardia [1979] wird anhand der *multivariaten Schiefe* und *Kurtosis* die multivariate Normalverteilungsannahme überprüft. Als Test zur Prüfung der Homoskedastizitätsannahme (gleiche Kovarianzmatrizen) wird nach Korin [1968, 1969] der Bartlett-Test durchgeführt.
Das SAS-Macro *stichumf.mac* dient zur Stichprobenplanung im Falle der multivariaten einfaktoriellen Varianzanalyse. Ausserdem kann ein Plot der Gütefunktion angezeigt werden. Die Analysen basieren auf Approximationen der Null- und Nichtnullverteilung des Hotellingschen Spurkriteriums mittels nichtzentralen F-Verteilungen, siehe Läuter [1974, 1978b] sowie Ahrens, Läuter [1981].
Im Spezialfall der univariaten Varianzanalyse führen die Resultate auf dieselben Stichprobenumfänge, wie sie im SAS-Modul Analyst Application berechnet werden können.

Das SAS-Macro hotellin.mac

Das SAS-Macro *hotellin.mac* gestattet neben der Durchführung des Hotellingschen T^2-Tests die Güte des Tests zu berechnen. Ausserdem werden Tests zur Überprüfung der Modellvoraussetzungen - multivariate Normalverteilungsannahme und Homoskedastizitätsannahme – durchgeführt.

1.1. Zweistichprobenvergleich mit Hotellings T^2

Hier wird der Vergleich der p-dimensionalen Erwartungswertvektoren $\boldsymbol{\mu}_1$ *und* $\boldsymbol{\mu}_2$ von k = 2 p-dimensional normalverteilten multivariaten Gesamtheiten behandelt.

Daten

Es liegen zwei p-dimensionale Stichproben, repräsentiert durch die (n_1,p)-Datenmatrix $\mathbf{Y_1}$ sowie die (n_2,p)-Datenmatrix $\mathbf{Y_2}$ vor, welche man durch Untereinanderschreiben zur dann (n_1+n_2,p) - Gesamtdatenmatrix $\mathbf{Y}$ erweitern kann. Mit $N= n_1 + n_2$ wird der Gesamtstichprobenumfang bezeichnet.

Modell

Es wird unterstellt, daß
die n_1 Zeilenvektoren von $\mathbf{Y_1}$ stochastisch unabhängig $\mathbf{N_p}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) - \textit{verteilt}$,
die n_2 Zeilenvektoren von $\mathbf{Y_2}$ stochastisch unabhängig $\mathbf{N_p}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) - \textit{verteilt}$ sind.
Außerdem seien die beiden Stichproben untereinander ebenfalls stochastisch unabhängig, d.h. alle N p-dimensionalen Zeilenvektoren sind stochastisch unabhängig.
Zusätzlich gelte analog zum univariaten (p=1) Fall die Homoskedastizitätsannahme, also

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}.$$

In kompakter Form lautet die Modellgleichung

$$\boxed{\mathbf{y_{ij}} = \boldsymbol{\mu}_\mathbf{i} + \boldsymbol{\varepsilon}_\mathbf{ij} \quad i = 1,2 \; und \; j = 1,2,...,n_i}$$

Die p-dimensionalen Fehlerzufallsvektoren $\boldsymbol{\varepsilon}_{1j}$ bzw. $\boldsymbol{\varepsilon}_{2j}$ gehorchen $\mathbf{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ - bzw. $\mathbf{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_2)$ - Verteilungen.
Die multivariate Normalverteilungsannahme wird in Abschnitt 1.2, die Homoskedastizitätsannahme (gleiche Kovarianzmatrizen) wird im Abschnitt 1.3 näher untersucht.

Hypothesen

Es soll die Null-Hypothese $\mathbf{H_0}: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ gegen die Alternative $\mathbf{H_A}: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$
bei vorgegebener Irrtumswahrscheinlichkeit α getestet werden.

Teststatistik

Die (1,p)-Mittelwertsvektoren $\bar{\mathbf{y}}_1 \; und \; \bar{\mathbf{y}}_2$ sowie die empirischen (p,p)- Kovarianzmatrizen $\mathbf{S}_1 \; und \; \mathbf{S}_2$ sind erwartungstreue Schätzungen der unbekannten (1,p)-Erwartungswertvektoren $\boldsymbol{\mu}_1 \; und \; \boldsymbol{\mu}_2$ sowie der unbekannten (p,p)-Kovarianzmatrizen $\boldsymbol{\Sigma}_1 \; und \; \boldsymbol{\Sigma}_2$.
Ein geeignetes statistisches Abstandsmass zwischen zwei p-dimensionalen Erwartungswertvektoren $\boldsymbol{\mu}_1 \; und \; \boldsymbol{\mu}_2$ ist die sog. „*Quadrierte Mahalanobis-Distanz*"

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'.$$

Eine erwartungstreue Schätzung von Δ^2 ergibt sich gemäss

$$D^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)',$$

wobei

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)\mathbf{S_1} + (n_2 - 1)\mathbf{S_2})$$

die *gepoolte* Kovarianzmatrix bezeichnet.

„*Hotellings* T^2 " wird als Teststatistik verwendet, wobei nach Definition gilt

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2.$$

Hotellings T^2 ist auch nahe verwandt mit dem in der multivariaten Varianzanalyse verwendeten Hotelling-Lawley-Spur-Kriterium U, siehe Abschnitt 2.1, im Falle des zwei-Stichprobenvergleichs gilt

$$U = \frac{1}{n_1 + n_2 - 2} T^2 \quad .$$

Unter der Null-Hypothese $\mathbf{H_0 : \mu_1 = \mu_2}$ folgt die Prüfgrösse

$$\mathrm{F} = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2 \quad \textit{einer zentralen } F_{p, n_1 + n_2 - p - 1} - \textit{Verteilung}$$

Entscheidungsvorschrift

$$\boxed{\text{Verwerfe } \mathrm{H}_0 \text{ , falls } \mathrm{T}^2_{\mathrm{ber}} \; > \; \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{1-\alpha, p, n_1 + n_2 - p - 1}}$$

Eine äquivalente Form der Entscheidungsvorschrift in SAS wird mittels der sog. „Überschreitungswahrscheinlichkeit" geliefert:

$$\boxed{\mathrm{Pr\_}T2 \; < \; \alpha \quad , \quad \textit{dann verwerfe } H_0}$$

Güte des Hotelling-Tests

Unter der Alternative $\mathbf{H_A : \mu_1 \neq \mu_2}$ folgt die oben eingeführte Prüfgrösse

$$\boxed{\begin{array}{l} \mathrm{F} = \dfrac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2 \quad \textit{einer nichtzentralen } F_{p, n_1 + n_2 - p - 1} - \textit{Verteilung} \\ \textit{mit Nichtzentralitätsparameter} \quad \delta^2 = \dfrac{n_1 n_2}{n_1 + n_2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \end{array}}$$

Bei vorgegebener quadrierter Mahalanobis-Distanz $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'$ kann die Güte (Power) des Hotelling-Tests zum Niveau α in SAS berechnet werden gemäss

$$\boxed{guete = 1 - PROBF(F_{ber}, \, p \, , n_1 + n_2 - p - 1 \, , \delta^2).}$$

In der Praxis muss die Kovarianzmatrix $\boldsymbol{\Sigma}$ durch die Schätzung $\mathbf{S}$ ersetzt werden.

1.2. Test auf multivariate Normalverteilung

Zunächst werden die Tests auf multivariate Normalverteilung bei Vorliegen einer Datengruppe entwickelt.

Im Falle des Hotellingschen Zweistichprobenvergleichs wird der Test auf multivariate Normalverteilung auf die $n_1 + n_2$ beobachteten Residuenvektoren $\mathbf{e_{ij}}$ angewendet.

Setzt man voraus, daß die n Zeilen einer (n,p)-Datenmatrix $\mathbf{Y}$ Realisationen von $\mathbf{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \textit{verteilten}$ Zufallsvektoren sind, dann sollte man diese Modellannahme an den vorliegenden Daten nachprüfen. Mardia [1979] stellt eine Menge von Methoden vor, wie die multivariate Normalverteilungsannahme zu verifizieren ist. Eine einfache Methode ist, anhand der *multivariaten Schiefe* und *Kurtosis* die multivariate Normalverteilungsannahme nachzuprüfen. Unter Verwendung der invarianten quadratischen Formen

$$\boxed{g_{rs} = (\mathbf{y}_r - \overline{\mathbf{y}})\mathbf{S}^{-1}(\mathbf{y}_s - \overline{\mathbf{y}})'; \quad r, s = 1, 2, ..., n}$$

kann man nach Mardia die multivariate Schiefe durch

$$\boxed{\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{r,s}^{n} g_{rs}^3}$$

und die multivariate Kurtosis durch

$$\boxed{\hat{\beta}_{2,p} = \frac{1}{n}\sum_{r}^{n} g_{rr}^2}$$

definieren.
Unter der Hypothese H_{NV} der multivariaten Normalverteilung ist

$$\boxed{E(\hat{\beta}_{1,p}) = 0 \text{ und } E(\hat{\beta}_{2,p}) = p(p+2).}$$

Unter H_{NV} und für grosse Stichprobenumfänge n folgt

$$\boxed{\kappa appa_1 = n\hat{\beta}_{1,p}/6}$$

einer $\chi^2-Verteilung\ mit\ p(p+1)(p+2)/6\ Freiheitsgraden.$

Unter H_{NV} und für grosse Stichprobenumfänge n folgt

$$\boxed{\kappa appa_2 = (\hat{\beta}_{2,p} - p(p+2))/\sqrt{8p(p+2)/n}}$$

einer standardisierten N(0,1)-Verteilung.

Bei Vorliegen von Nichtnormalität ist das vorgegebene Niveau α des Hotelling- Tests vor allem sensitiv gegenüber der multivariaten Schiefe, während der in Abschnitt 1.3 besprochene Bartlett-Test durch die multivariate Kurtosis beeinflusst wird, siehe Khattree, Naik [1995].

Des weiteren sind grafische Verfahren wie z.B. Q-Q-Plots zur Prüfung der multivariaten Normalverteilungsannahme möglich, siehe Mardia [1979] und Gnanadesikan [1980].
Einfache zwei- oder drei-dimensionale Scatterplots zur Begutachtung der Daten lassen sich mit Hilfe der SAS-Prozeduren PROC GPLOT und PROC G3D durchführen.

Im Falle des Hotellingschen Zweistichprobenvergleichs wird der Test auf multivariate Normalverteilung auf die $n_1 + n_2$ beobachteten Residuenvektoren $\mathbf{e}_{ij} = \mathbf{y}_{ij} - \overline{\mathbf{y}}_{i.}$ angewendet, obwohl diese nicht stochastisch unabhängig sind. In der Regel sind beide oben erwähnte Tests erst anzuwenden, wenn beide Stichprobenumfänge mindestens 10 sind, so dass mindestens 20 Residuenvektoren vorliegen.

Bemerkung
Führt man Vortests zur Prüfung einer Verteilungsannahme durch, dann kann das nominelle Testniveau der darauf folgenden Tests verfälscht werden. Die obigen Tests sind als explorative Datenanalyse anzusehen.

1.3. Test auf homogene Kovarianzmatrizen

Zur Überprüfung der Nullhypothese $\mathbf{H}_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ verwendet man den Bartlett-Test, siehe Korin [1968, 1969].

$\mathbf{S}_1$ *und* $\mathbf{S}_2$ sind erwartungstreue Schätzungen der unbekannten (p,p)-Kovarianzmatrizen $\mathbf{\Sigma}_1$ *und* $\mathbf{\Sigma}_2$, die gepoolte Kovarianzmatrix $\mathbf{S}$ schätzt $\mathbf{\Sigma}$.

Die verwendete Teststatistik basiert auf der Grösse

$$\boxed{M = (N-2)\log_e(\det(\mathbf{S})) \; - \; \sum_{\mathbf{i=1}}^{\mathbf{2}} (n_i - 1)\log_e(\det(\mathbf{S_i}))}$$

Fall 1: Falls beide Stichprobenumfänge $n_{1,} \; n_2 \geq 20$ *und die Dimension* $p \leq 5$, dann berechnet man einen geeigneten Korrekturfaktor C, siehe Korin [1968,1969)].

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)} \left(\sum_{i=1}^{2} \frac{1}{n_i - 1} \; - \; \frac{1}{N-2} \right) \quad , \quad v_1 = \frac{1}{2} p(p+1)$$

Entscheidungsvorschrift:

$$\boxed{Ist \; PR\_BART = 1 - PROBCHI(CM \; , v_1) \;\; < \;\; \alpha \quad , \quad dann \; verwerfe \; H_0}$$

Fall 2: Falls ein Stichprobenumfang kleiner als 20 oder die Dimension grösser als 5, dann können bei Anwendung weiterer Fallunterscheidungen appproximative F-Tests zur Prüfung obiger Nullhypothese verwendet werden.

Nähere Datails findet man bei Korin [1968,1969] bzw. im Source-Code des SAS-Macros hotellin.mac.

$$A_1 = 1 - C \quad , \quad A_2 = \frac{(p-1)(p+2)}{6} \left(\sum_{i=1}^{2} \frac{1}{(n_i - 1)^2} \; - \; \frac{1}{(N-2)^2} \right) \quad , \quad v_2 = \frac{v_1 + 2}{\left| A_2 - A_1^2 \right|}$$

$$B_1 = \frac{1 - A_1 - \dfrac{v_1}{v_2}}{v_1} \quad , \quad B_2 = \frac{v_2}{1 - A_1 + \dfrac{2}{v_2}}$$

Fall 2a: $A_2 - A_1^2 > 0$

Entscheidungsvorschrift:

$$\boxed{Ist \; PR\_BART = 1 - PROBF(B_1 M \; , v_1, v_2) \;\; < \;\; \alpha \quad , \quad dann \; verwerfe \; H_0}$$

Fall 2b: $A_2 - A_1^2 < 0$

Entscheidungsvorschrift:

$$\boxed{Ist \; PR\_BART = 1 - PROBF\left(\frac{v_2 M}{v_1 (B_2 - M)}, v_1, v_2 \right) \;\; < \;\; \alpha \quad , \quad dann \; verwerfe \; H_0}$$

Fall 2c: $A_2 - A_1^2 = 0$

Entscheidungsvorschrift:

$$\boxed{Ist \; PR\_BART = 1 - PROBCHI(CM \; , v_1) \;\; < \;\; \alpha \quad , \quad dann \; verwerfe \; H_0}$$

1.4. Durchführung in SAS mittels Macro hotellin.mac

Wir führen Hotellings T^2-Test samt den Tests der Modellannahmen beispielhaft an den Daten *daten1* und *daten2* durch.
Aus k = 2 Grundgesamtheiten wurden jeweils n = 5 Personen zufällig ausgewählt und an jeder Person Körpergewicht [in kg] und Körpergröße [im cm] gemessen.
Die Versuchsfrage lautet: Sind die beiden Gesamtheiten bezüglich des (p=2)-dimensionalen Merkmals (*gew,lae*) unterschiedlich ?
Um diese Fragestellung zu beantworten, wird ein statistischer Test durchgeführt.

Es wird getestet die Null-Hypothese

$$\boxed{H_0 : (\mu_{11} \quad \mu_{12}) = (\mu_{21} \quad \mu_{22})}$$

zum vorgegebenem Niveau $\alpha = 0.05$.

Durch Vorgabe einer quadrierten Mahalanobis-Distanz von beispielsweise Δ^2=2 kann ausserdem die Güte (= Aufdeckungswahrscheinlichkeit des vorgegebenem Δ^2) berechnet werden.
Zuerst werden die beiden SAS-Dateien *daten1* und *daten2* mit den Variablen *gew* und *lae* in einem SAS-DATA Step erzeugt.

Data Step

```
DATA daten1;
INPUT gew lae@@;
CARDS;
69 178 54 162 68 174 78 190 71 181
;
DATA daten2;
INPUT gew lae@@;
CARDS;
78 180 62 160 80 175 85 185 75 165
;
```

Aufruf des Macros

```
%INC 'D:\KSFE\KSFE4\hotellin.mac';
%hotellin(daten1,daten2,alpha=0.05,delta_2=2);
```

Der Aufruf des Macros über das `%INC-Statement` hängt natürlich vom Unterverzeichnis ab, in dem *hotellin.mac* abgespeichert ist.
Im `%hotellin-Statement` sind die beiden (zulässigen) SAS-Dateinamen, die Irrtumswahrscheinlickeit α und die aufzudeckende quadrierte Mahalanobis-Distanz Δ^2 anzugeben.
Die Durchführung des Macros bewirkt , dass im Output zuerst beschreibende Statistiken, dann Hotellings Test angezeigt werden. Anschliessend folgt die Güteberechnung und die beiden explorativen Tests auf homogene Kovarianzmatrizen und multivariate Normalverteilung.

Output

```
          HOTELLINGS T_Quadrat TEST                        1
NV-Pruefung, Bartlett-Test und Gueteberechnung

          BESCHREIBENDE STATISTIKEN
          -------------------------
                                 N1        N2
Stichprobenumfänge                5         5

                                       P
     Dimension                         2

(n,p)-Datenmatrix Y1             69       178
                                 54       162
                                 68       174
                                 78       190
                                 71       181

                                 Y2
(n,p)-Datenmatrix Y2             78       180
                                 62       160
                                 80       175
                                 85       185
                                 75       165

                                      D
(1,p)-Differenzvektor y1-y2      -8         4

                                 S1
(p,p)-Kovarianzmatrix S1       76.5     88.25
                              88.25       105

                                 S2
(p,p)-Kovarianzmatrix S2       74.5        80
                                 80     107.5


                                  S
(p,p)-Gepoolte Kov.-matrix     75.5    84.125
                             84.125    106.25
```

Im ersten Teil des Outputs werden nochmal die beiden Stichprobenumfänge ($n_1 = n_2 = 5$), die Dimension (p=2) und die beiden (5,2)-Datenmatrizen $\mathbf{Y_1}$ *und* $\mathbf{Y_2}$ aufgelistet. Aus dem (1,2)-Differenzen-Mittelwertsvektor D lesen wir ab, dass die mittlere Differenz im Körpergewicht zwischen erster und zweiter Gruppe –8 [kg] , die mittlere Differenz der Körpergrösse 4 [cm] beträgt. Ausserdem werden die beiden empirischen (2,2)-Kovarianzmatrizen **S1** und **S2** sowie die gepoolte Kovarianzmatrix **S** ausgegeben. Aus der Matrix **S1** entnehmen wir, dass die Stichprobenvarianz der Variablen *gew* $s_{11} = 76.5$ beträgt, die Stichprobenvarianz der Variablen *lae* $s_{22} = 105$ und die Stichprobenkovarianz der Variablen *gew* und *lae* $s_{12} = 88.25$ beträgt. Analog ind die anderen Kovarianzmatrizen zu lesen. Die gepoolte Kovarianzmatrix berechnet sich gemäss $\mathbf{S} = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$.

```
                    HOTELLINGS T-QUADRAT TEST                    2
                    -------------------------
                        HO: MUE1=MUE2
                                              T2
Hotellings T2                          35.433844
                                               U
Hotelling Lawleys U                    4.4292305
                                              MA
Beob. Mahalanobis Distanz              14.173538
                                               F
F-Statitik                             15.502307
                                             DF1
Zählerfreiheitsgrad                            2
                                             DF2
Nennerfreiheitsgrad                            7
                                           PR_T2
Überschr.- Wahrsch.                    0.0026817
                                            FINV
(1-alpha)-Quantil                      4.7374141
```

In diesem Teil des Output wird die empirische quadrierte Mahalanobis-Distanz berechnet:

$$MA = D^2 = \begin{pmatrix} -8 & 4 \end{pmatrix} \begin{pmatrix} 75.5 & 84.125 \\ 84.125 & 106.25 \end{pmatrix}^{-1} \begin{pmatrix} -8 \\ 4 \end{pmatrix} = 14.173538$$

Aus Abschnitt 1.1 ist zu entnehmen, dass

$$T2 = T^2 = \tfrac{5 \cdot 5}{5+5} D^2 = 2.5 \cdot D^2 = 35.433844$$
$$U = \tfrac{1}{5+5-2} T^2 = \tfrac{1}{8} T^2 = 4.4292305$$
$$F = \tfrac{5+5-2-1}{(5+5-2)2} T^2 = \tfrac{7}{16} = 15.502307$$
$$DF1 = p = 2\,,\;\; DF2 = 5+5-2-1 = 7$$
$$\Pr\_T2 = 1 - PROBF(15.502307, 2, 7) = 0.0026817$$
$$FINV = F_{1-0.05,2,7} = 4.7374141$$

Es liegen signifikante Unterschiede auf dem 0.05-Niveau vor, da Pr_T2=0.0026817 < 0.05 ist.

```
                      BERECHNUNG DER GUETE                          3
                      --------------------
                                         DELTA_2
Vorgegebene Mahalanobisdistanz                 2
                                           GUETE
Guete für diese Distanz                0.3460654
```

Die Güte bei vorgegebenem $\Delta^2 = 2$ ergibt sich gemäss Abschnitt 1.1 zu

$$guete = 1 - PROBF(4.7374141,\;\; 2\;,\;\; 7\;,\;\; 5) = 0.3460654,$$
$$\textit{der Nichtzentralitätsparameter ist } \tfrac{5 \cdot 5}{5+5} \cdot 2 = 5$$

Die Aufdeckungswahrscheinlichkeit von $\Delta^2 = 2$ (bei $\alpha = 0.05$) liegt also bei 0.3460654.

```
                TEST AUF HOMOGENE KOVARIANZMATRIZEN         4
                -----------------------------------
                          Bartlett-Test
                        HO: SIGMA1=SIGMA2
                                   PR_BART
Überschr.- Wahrsch.              0.4950703
```

Bei diesem Test verwenden wir ein Niveau von $\alpha = 0.10$, um den Test trennschärfer zu gestalten. Da die Überschreitungswahrscheinlichkeit PR_BART = 0.4950703 > 0.10, wird nach Abschnitt 1.3 die Homoskedastizitätsannahme beibehalten.

```
           TEST AUF NORMALVERTEILUNG                  5
          --------------------------
                        HO:Schiefe=0
                               BETA1HAT
Schiefe                       1.8670757
                                 KAPPA1
Teststatistik                 3.1117929
                               PVALSKEW
Überschr.-Wahrsch.            0.5392945


                     HO:Kurtosis=p(p+2)

                               BETA2HAT
Kurtosis                      6.2782486
                                 KAPPA2
Teststatistik                 -0.680582
                               PVALKURT
Überschr.-Wahrsch.             0.496136
```

Bei diesen Tests verwenden wir je ein Niveau von $\alpha = 0.10$, das multiple Niveau bei simultaner Durchführung beträgt dann maximal 0.20. Da die Überschreitungswahrscheinlichkeiten PVALSKEW = 0.5392945 > 0.10 und PVALKURT=0.496136 > 0.10, wird nach Abschnitt 1.2 die Normalverteilungsannahme beibehalten. Die geschätzte multivariate Schiefe hat einen Wert von $\hat{\beta}_{1,p} = BETA1HAT = 1.8670757$ und die multiple Kurtosis einen Wert von $\hat{\beta}_{2,p} = BETA2HAT = 6.2782486$.

2. Das SAS-Macro stichumf.mac

2.1. Einfaktorielle multivariate Varianzanalyse

Es werden Versuche analysiert, bei der die Wirkung eines Einflußfaktors - der k verschiedene Stufen annehmen kann - auf ein p-dimensionales Merkmal untersucht werden soll. Hier wird also der Vergleich der p-dimensionalen Erwartungswertvektoren von allgemein k (≥ 2) multivariaten Gesamtheiten behandelt.

Daten

Es liegen k p-dimensionale Stichproben, repräsentiert durch die (n,p)-Datenmatrix $\mathbf{Y_1}$, (n,p)-Datenmatrix $\mathbf{Y_2}$ **, ...,** (n,p)-Datenmatrix $\mathbf{Y_k}$ vor. Der Gesamtstichprobenumfang ist also N = kn. Die (N,p)- Gesamtdatenmatrix **Y** entsteht durch vertikales Untereinanderschreiben der k Gruppen-Datenmatrizen. Die Daten müssen nach den k Gruppen klassifiziert sein.

Modell

Es wird unterstellt, daß folgende additive lineare Modellgleichung gilt:

$$\boxed{\mathbf{y}_{\mathbf{ij}} = \boldsymbol{\mu}_{\mathbf{i}} + \boldsymbol{\varepsilon}_{\mathbf{ij}} \quad i = 1, 2, ..., k \; und \; j = 1, 2, ..., n}$$

Die insgesamt N = kn stochastisch unabhängigen p-dimensionalen Fehlerzufallsvektoren $\boldsymbol{\varepsilon}_{\mathbf{ij}}$ gehorchen $\mathbf{N_p}(\mathbf{0}, \boldsymbol{\Sigma})$ - Verteilungen (Normal- und Homoskedastizitätsannahme).

Hypothesen

Es soll die Null-Hypothese $\mathbf{H_0} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_k$
bei vorgegebener Irrtumswahrscheinlichkeit α getestet werden.

Teststatistik

Bezeichne $\mathbf{y}_{\mathbf{ij}}$ den j-ten p-dimensionalen Beobachtungsvektor in der i-ten Gruppe,

$\bar{\mathbf{y}}_{\mathbf{i}.} = \frac{1}{\mathbf{n}} \sum_{j=1}^{n} \mathbf{y}_{\mathbf{ij}}$ den p-dimensionalen Mittelwertsvektor in der i-ten Gruppe

$\bar{\mathbf{y}}_{..} = \frac{1}{\mathbf{k \cdot n}} \sum_{i=1}^{k} \sum_{j=1}^{n} \mathbf{y}_{\mathbf{ij}}$ den p-dimensionalen Gesamt-Mittelwertsvektor,

$\mathbf{H} = \sum_{i=1}^{k} n(\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})'(\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})$ die (p,p)- Hypothesenmatrix

$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mathbf{y}_{\mathbf{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.})'(\mathbf{y}_{\mathbf{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.})$ die (p,p)- Errormatrix,

dann wird die (p,p)-Matrix $\mathbf{HE^{-1}}$ als Grundbaustein zur Herleitung einer Teststatistik verwendet. Man beachte, daß die Matrix $\mathbf{HE^{-1}}$ im allgemeinen nicht symmetrisch ist, obwohl dies sowohl **E** als auch **H** sind. Um eine Testentscheidung zu ermöglichen, müssen wir die Matrix $\mathbf{HE^{-1}}$ zu einer Zahl zusammenfassen. Dafür gibt es vier übliche Methoden. Damit gibt es vier gebräuchliche Tests im MANOVA-Modell. Die üblichen statistischen Programmpakete (also auch SAS) geben die Werte aller vier Teststatistiken und ihre jeweiligen Überschreitungswahrscheinlichkeiten für alle vier Tests an, da keiner dem anderen eindeutig vorzuziehen ist. Es bleibt den Benutzern überlassen, welchen Test Sie verwenden.

Der hier verwendete "**Hotelling-Lawley-Test**" verwendet als Teststatistik

$$\boxed{T_{HL} = U = spur(\mathbf{H} \bullet \mathbf{E}^{-1}) = \lambda_1 + \lambda_2 + \ldots + \lambda_q}$$

Dabei sind die $\lambda_1, \lambda_2, \ldots, \lambda_q$ die nichtverschwindenden Eigenwerte von $\mathbf{HE}^{-1}$.
Unter der Nullhypothese kann die Verteilung eines gewissen Vielfachen von U durch eine zentrale F-Verteilung approximiert werden, siehe Läuter [1974,1978b] und Ahrens, Läuter [1981].

Man lehnt die Nullhypothese ab, wenn $c \cdot U$ größer ist als ein $(1-\alpha)-Quantil$ einer geeigneten F-Verteilung ist.

2.2. Gütefunktion und Stichprobenplanung

Falls die Nullhypothese nicht mehr gilt, gehorcht die (p,p)-Errormatrix **E** weiterhin einer sog. „WISHART-Verteilung“ $W(\mathbf{\Sigma}, f_2)$ mit $f_2 =$ k(n-1)) Freiheitsgraden, während die (p,p)-Hypothesenmatrix **H** nun einer nichtzentralen WISHART-Verteilung $W(\mathbf{\Sigma}, f_1, \mathbf{M})$ folgt mit $f_1 = k-1$ Freiheitsgraden und Nichtzentralitätsparameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$. Die (p,k-1)-Matrix **M** soll Rang 1 haben, d.h. $\gamma_1 = spur(\mathbf{M}'\mathbf{\Sigma}^{-1}\mathbf{M})$ ist der einzige in der Verteilung von **H** nichtverschwindende Nichtzentralitätsparameter.
Um den Stichprobenumfang planen zu können, muss die Irrtumswahrscheinlichkeit α, die Güte $1-\beta$, die Varianz σ^2 und eine gewisse (aufzudeckende) Mindestabweichung von der Nullhypothese vorgegeben werden.

2.2.1. Zwei Möglichkeiten von Alternativen

Folgende Alternativen sind nach Läuter [1978] besonders zu beachten.
Es wird das **j-te Merkmal** (Variable) mit $j \in \{1, 2, ..., p\}$ herausgegriffen.

$$H_1 : \frac{1}{\sigma_j^2} \sum_{i=1}^{k} (\mu_{ij} - \bar{\mu}_{.j})^2 \geq q^2$$

Die Summe der quadrierten Abweichungen des Erwartungswerts μ_{ij} der i-ten Gruppe bezüglich des über die Gruppen gemittelten Totalerwartungswerts $\bar{\mu}_{.j}$ dividiert durch die Varianz σ_j^2 des j-ten Merkmals soll mindestens q^2 sein. Dieses Kriterium wird bei der Stichprobenplanung im eindimensionalen Fall (p = 1) zugrundegelegt.

$$H_2 : \frac{1}{\sigma_j} \left| \mu_{i_1 j} - \mu_{i_2 j} \right| \geq theta$$

Es gibt 2 Gruppen i_1 und i_2, deren Erwartungswerte um mindestens $theta \cdot \sigma_j$ voneinander abweichen.

Für den ersten Nichtzentralitätsparameter γ_1 der Verteilung der Hypothesenmatrix **H** bedeutet dies, dass gilt:

$$\gamma_1 \geq n \cdot q^2 \qquad \textit{im Falle von } H_1$$

$$\gamma_1 \geq n \cdot \frac{theta^2}{2} \qquad \textit{im Falle von } H_2$$

Die anderen Nichtzentralitätsparameter $\gamma_{2,}\gamma_{3,}....$ können 0 sein unter $H_1\, bzw. H_2$. Da die Gütefunktion des Hotelling Lawley Tests monoton in allen Nichtzentralitätsparametern ist, genügt es zur Bestimmung des Stichprobenumfangs, die Abhängigkeit vom ersten Nichtzentralitätsparameter zu kennen und die anderen Null zu setzen, siehe Läuter [1978b].

2.2.2. Approximation der Nichtnullverteilung

Unter den anfangs des Abschnitts erwähnten Bedingungen über die Verteilungen von **H** und **E** wird die Nichtnullverteilung der Teststatistik $U = spur(\mathbf{H} \bullet \mathbf{E}^{-1})$ betrachtet, wobei noch gilt:

$$p \geq 1, \quad k \geq 2, \quad n \geq 1 + \frac{p+2}{k}$$

$$\text{Für } f_1 + f_2 - f_1 p - 1 > 0 \quad ist \quad \tilde{F} = \frac{f_2 - p + 1}{f_1 p} \cdot U$$

approximativ nichtzentral F-verteilt mit den Freiheitsgraden

$$g_1 = \frac{f_1 p (f_2 - p)}{f_1 + f_2 - f_1 p - 1} \quad , \quad g_2 = f_2 - p + 1$$

$$\text{und der Nichtzentralität } nc = \frac{g_1}{f_1 p} \cdot \gamma_1$$

$$\text{Für } f_1 + f_2 - f_1 p - 1 \leq 0 \quad ist \quad \tilde{F} = \frac{f_2 - p - 1}{f_1 p + \gamma_1} \cdot \frac{g_2}{g_2 - 2} \cdot U$$

approximativ zentral F-verteilt mit den Freiheitsgraden

$$g_1 = \infty \quad , \quad g_2 = 4 + \frac{(f_1 p + \gamma_1)^2 (f_2 - p) f_2 - p - 3)}{(f_1 p + 2\gamma_1)(f_2 - 1)(f_1 + f_2 - p - 1) + \gamma_1^2 (f_2 - p)}$$

Die Approximationen gelten in dem Sinne, dass Übereinstimmung der exakten und approximativen Verteilungen in den ersten beiden Momenten vorliegt, siehe Läuter [1978b] und Ahrens, Läuter [1981].

2.3. Durchführung in SAS mittels Macro stichumf.mac

Die Bestimmung des gesuchten Stichprobenumfangs erfolgt durch Aufruf des Macros mittels des %INC-Statements, wobei zu beachten ist, in welchem Unterverzeichnis das Macro *stichumf.mac* abgespeichert ist.

1. Beispiel

```
%INC 'D:\KSFE\KSFE4\stichumf.mac';

%stichumf(k=3,p=2,theta=.,smq=12.5,s2=10,alpha=0.05,guete_v=0.95);
```

Im %stichumf - Statement sind zu spezifizieren:
Gruppenzahl k , Dimension p,
Varianz s2 = s^2 , Irrtumswahrscheinlichkeit alpha = α ,
Güte guete_v = $1-\beta$,
Entweder wird die Genauigkeit mittels Vorgabe von theta (Alternative H_2) oder durch Vorgabe von smq (Alternative H_1) spezifiziert, eine der beiden Angaben ist mit einem . (für missing value) zu versehen.

Output

P	K	ALPHA	GUETE_V	S2	THETA	DELTA	SMQ	Q2	UMFANG
2	3	0.05	0.95	10	.	.	12.5	1.25	17

Im Falle k=3, p=2, $\alpha = 0.05, 1-\beta = 0.95, s^2 = 10 \; sowie \; smq = \sum_{i=1}^{3} \left(\mu_{i.} - \bar{\mu}\right)^2 = 12.5$ erhält man einen Stichprobenumfang von 17.

2. Beispiel

```
%INC 'D:\KSFE\KSFE4\stichumf.mac';

%stichumf(k=3,p=2,theta=5,smq= . ,s2=10,alpha=0.05,guete_v=0.95);
```

erhält man den

Output

P	K	ALPHA	GUETE_V	S2	THETA	DELTA	SMQ	Q2	UMFANG
2	3	0.05	0.95	10	5	1.58113	.	.	17

Im Falle k = 3, p = 2, $\alpha = 0.05, 1-\beta = 0.95, s^2 = 10 \; sowie \; theta = \left|\mu_1 - \mu_3\right| = 5$ erhält man ebenfalls einen Stichprobenumfang von 17.
Beispielsweise führen die Verhältnisse $\mu_1 = 7.5, \mu_2 = 10, \mu_3 = 12.5$ auf theta = 5 und auf smq=12.5.

Ein Plot der Gütefunktion (güte*stichprobenumfang) wird automatisch durch das macro miterzeugt.

In SAS ANALYST APPLICATION steht unter SAMPLE SIZE die Stichprobenplanung

unter One Way ANOVA zur Verfügung, aber nur für Dimension p = 1.

Die im Macro $smq = \sum_{i=1}^{k} (\mu_{i.} - \bar{\mu})^2$ genannte Grösse wird mit *CSS of means* bezeichnet, k wird *#treatments* genannt, guete_v heisst *power* und anstelle der Varianz muss die *Standard-abweichung* angegeben werden. Dann erhält man identische Resultate.

Literatur

Ahrens, H., Läuter, J. (1981). Mehrdimensionale Varianzanalyse, Akademie Verlag Berlin

Gnanadesikan, R. (1980). „Graphical Methods for Internal Comparisons in ANOVA and MANOVA“, In Handbook of Variance, Ed. P. R. Krishnaiah, 133-177, Amsterdam: North Holland.

Läuter, J. (1974). Approximation des Hotelling T^2 durch die F-Verteilung, Biometrische Zeitschrift 16, Heft 3

Läuter, J. (1978b). Sample Size Requirements for the T^2-Test of MANOVA, Biometrical Journal 20, Nr. 4

Khattree, R., Naik, D.N. (1995). Applied Multivariate Statistics with SAS Software, SAS Institute Inc., Cary, NC, USA

Korin, B.P. (1968). On the distribution of a statistic used for testing a covariance matrix, Biometrika 55, 171-178

Korin, B.P. (1969). On testing the equality of k covariance matrices. Biometrika 56, 216-218

Mardia, K.V., Kent, J.J. and Bibby, J.M. (1979). Multivariate Analysis. New York, Academic Press

SAS Institute Inc. (1989). SAS/IML Software: Usage and Reference. Version 6. First Edition, Cary NC: SAS Institute Inc.

SAS Institute Inc. (1989), SAS/STAT User's Guide, Version 6. Fourth Edition, Volume 2, Cary NC: SAS Institute Inc.