Dr. Katharina Best
Sommersemester 2011
9. Juni 2011

ulm university universität **uulm**

Reader – Teil 8: Varianzanalyse

WiMa-Praktikum

Bei der Varianzanalyse gehen wir davon aus, dass ein faktoriell vorliegendes Merkmal Effekt auf unsere Messungen hat. Diese Effekte können einzeln oder gemeinsam untersucht werden. Wir betrachten zuerst den eindimensionalen Fall.

1 Varianzanalyse eines Faktors

1.1 Modellierung

Wir gehen von Beobachtungen aus, welche in einem Merkmal A kategoriale Ausprägungen aufweisen, kodiert mit $1, \ldots, I$. Desweiteren haben wir pro Kategorie jeweils $J \geq 2$ Beobachtungen. Wir sprechen hierbei von *balancierten Daten*. Damit können wir sie tabellarisch anordnen und bezeichnen mit Y_{ij} die Zufallsstichprobenvariable der j-ten Beobachtung zur i-ten Ausprägung. Haben die unterschiedlichen Ausprägungen Effekte auf die Messung, so sind die Erwartungswerte innerhalb der Gruppen unterschiedlich. Deswegen definieren wir uns sowohl diese als μ_i als auch den Gesamterwartungswert μ, für den gilt $\mu = \mathbb{E}\left(\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}\right) = \frac{1}{I}\sum_{i=1}^{I}\mu_i$. Außerdem wollen wir die gruppenweise Abweichung $\alpha_i := \mu - \mu_i$ definieren. Es ist klar, dass $\sum_{i=1}^{I}\alpha_i = 0$ Wir können dann unsere Zufallsstichprobenvariablen aufstellen als $Y_{ij} = \mu + \alpha_i + Z_{ij}$, wobei Z_{ij} die zufällige Abweichung ist. Bevor wir uns einer statistischen Untersuchung unserer Fragestellung zuwenden, benötigen wir noch Schätzer für die Parameter. So ist $\hat{\mu} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} =: \overline{Y}_{\bullet\bullet}$ und $\hat{\mu}_i = \frac{1}{J}\sum_{j=1} JY_{ij} =: \overline{Y}_{i\bullet}$, womit wir den Schätzer $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$ haben. Um die Variation untersuchen zu können, definieren wir noch $SS_A = J\sum_{i=1}^{I}\hat{\alpha}_i^2$ als die Streuung zwischen den Gruppen, $SS_R = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{\mu}_i)^2$ als die Streuung innerhalb der Gruppen und die Gesamtstreuung $SS_T = SS_A + SS_R$.

Dieser Aufbau erlaubt es uns, einen statistischen Test mit der Nullhypothese $H_0 : \alpha_1 = \alpha_2 = \ldots = 0$ zu formulieren. Eine geeignete Teststatistik ist $T := \frac{SS_A/I-1}{SS_R/I(J-1)}$, welche unter der Annahme der identischen Normalverteilung der $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$ F-verteilt ist.

Alle Größen werden in einer Tabelle zusammengefasst, die ANOVA-Tabelle von *analysis of variance* heißt.

	Freiheitsgrade	Streuung	F-Statistik
SS_A	$I-1$	$J\sum_{i=1}^{I}(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})^2$	$\frac{SS_A/I-1}{SS_R/I(J-1)}$
SS_R	$I(J-1)$	$\sum_{j=1}^{J}\sum_{i=1}^{I}(\overline{Y}_{ij}-\overline{Y}_{i\bullet})^2$	
SS_T	$IJ-1$	$\sum_{j=1}^{J}\sum_{i=1}^{I}(\overline{Y}_{ij}-\overline{Y}_{\bullet\bullet})^2$	

1.2 Der Tukey-Test

Der oben beschriebene Test kann abgeändert werden. Eine weitere Möglichket, die Nullhypothese zu formulieren, ist $H_0 : \alpha_{i_1} = \alpha_{i_2}$ für $i_1, i_2 = 1, \ldots, I$, was mit der Bedingung $\sum_{i=1}^{I}\alpha_i = 0$ auf das Gleiche hinausführt. Der Vorteil ist, dass wir hier gleich Paare von Effekten erhalten, die nicht zu übereinstimmen erscheinen. Die Teststatistik $T = \frac{\max_{1\le i_1,i_2\le I}|\overline{Y}_{i_1\bullet}-\overline{Y}_{i_2\bullet}|}{\sqrt{\frac{1}{I(J-1)}\sum_{j=1}^{J}\sum_{i=1}^{I}(\overline{Y}_{ij}-\overline{Y}_{i\bullet})^2}/\sqrt{J}}$, wieder unter der Annahme der identischen Normalverteilung der $Z_{ij} \sim \mathcal{N}(0,\sigma^2)$, folgt der *studentisierten Spannweitenverteilung* mit den Parametern m und n.

1.3 Der Kruskal-Wallis-Test

Können wir die Annahme der Normalverteilung der Z_{ij} nicht treffen, so können wir nichtparametrisch vorgehen. Dazu definieren wir uns die Zufallsvariablen R_{ij}, welche die Ränge der Y_{ij} in der gesamten Stichprobe angeben, und führen dann unsere Untersuchung mit diesen durch. Dazu definieren wir uns $\overline{R}_{i\bullet} := \frac{1}{J}\sum_{j=1}^{J}R_{ij}$ für $i = 1, \ldots, I$ und $\overline{R}_{\bullet\bullet} := \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}R_{ij} = \frac{IJ+1}{2}$. Weiter benötigen wir die Streuung $SRS_A := \sum_{i=1}^{I}J_i(\overline{R}_{i\bullet}-\overline{R}_{\bullet\bullet})$. Unter der Nullhypothese keiner Effekte gilt $\frac{12}{N(N+1)}SRS_A \xrightarrow{D} \chi^2_{I-1}$.

1.4 Unbalancierte Daten

Haben wir für jede Ausprägung I statt der bisher angenommenen identischen J unterschiedlich viele Beobachtungen, sagen wir J_i, so müssen nur die Definitionen entsprechend angepasst werden.

2 Varianzanalyse zweier Faktoren

In diesem Fall haben wir noch ein zusätzliches kategorial auftretendes Merkmal B, welches $K \ge 2$ mögliche Ausprägungen aufweist. Außerdem gehen wir vorerst wieder von balancierten Daten mit jeweils J Beobachtungen aus. Sei μ_{ik} der Erwartungswert der Faktorkombination, schreiben wir $Y_{ikj} = \mu_{ik} + Z_{ikj}$. Für den Gesamterwartungswert erhalten wir $\mu = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\mu_{ik}$ und definieren die Effekte $\alpha_i := \frac{1}{K}\sum_{i=1}^{K}\mu_{ik} - \mu$ des ersten Merkmals, $\beta_k := \frac{1}{I}\sum_{k=1}^{I}\mu_{ik} - \mu$ des zweiten Merkmals und $\gamma_{ik} := \mu_{ik} - \alpha_i - \beta_k - \mu$ der Kombination. Die Effektdarstellung ist dann $Y_{ikj} = \mu + \alpha_i + \beta_k + \gamma_{ik} + Z_{ikj}$.

Die Hypothese wird mehrstufig aufgebaut. Zuerst wird $H_{0,\gamma} : \gamma_{ik} = 0$ für alle $i = 1, \ldots, I$ und $k = 1, \ldots, K$ überprüft. Gibt es keine Interaktionseffekte, so werden $H_{0,\alpha} : \alpha_i = 0$ für alle $i = 1, \ldots, I$ und $H_{0,\beta} : \beta_k = 0$

für alle $k = 1, \ldots, K$ betrachtet.

Die Schätzer werden analog zu oben definiert, $\hat{\mu}_{ik} := \frac{1}{J}\sum_{j=1}^{J} Y_{ikj} = \overline{Y}_{ik\bullet}, \hat{\mu} := \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\hat{\mu}_{ik} = \overline{Y}_{\bullet\bullet\bullet}$ und $\hat{\alpha}_i := \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{ik} - \hat{\mu} = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \hat{\beta}_k := \overline{Y}_{\bullet k\bullet} - \overline{Y}_{\bullet\bullet\bullet}$ sowie $\hat{\gamma}_{ik} := \overline{Y}_{ik\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet k\bullet} + \overline{Y}_{\bullet\bullet\bullet}$. Die Streuungen zwischen den Gruppen werden in $SS_A := KJ\sum_{i=1}^{I}\hat{\alpha}_i^2$, $SS_B := IJ\sum_{k=1}^{K}\hat{\beta}_k^2$ und $SS_{AB} := J\sum_{i=1}^{I}\sum_{k=1}^{K}\gamma_{ik}^2$. festgehalten, die Streuung innerhalb der Gruppen durch $SS_R = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J}(Y_{ikj} - \overline{Y}_{\bullet\bullet\bullet})^2$ und es gilt dann $SS_T = SS_A + SS_B + SS_{AB} + SS_R$. Daraus ergeben sich entsprechende Teststatistiken und die ANOVA-Tabelle:

	Freiheitsgrade	Streuung	F-Statistik
Haupteffekt A	$I-1$	SS_A	$T_\alpha := \frac{SS_A/I-1}{SS_R/IK(J-1)}$
Haupteffekt B	$K-1$	SS_B	$T_\beta := \frac{SS_B/K-1}{SS_R/IK(J-1)}$
Interaktionseffekt	$(I-1)(K-1)$	SS_{AB}	$T_\gamma := \frac{SS_{AB}/(I-1)(K-1)}{SS_R/IK(J-1)}$
Residualstreuung	$IK(J-1)$	SS_R	
Gesamtstreuung	$IKJ-1$	SS_T	