6 Vergleich mehrerer unverbundener Stichproben

6.1 Die einfaktorielle Varianzanalyse

Die einfaktorielle Varianzanalyse dient der Untersuchung des Einflusses einer kategorieller (bzw. nichtmetrischer) Variablen, die die Gruppenzugehörigkeit beschreibt, auf eine (oder im multivariaten Fall auf mehrere) abhängige stetige Variable. In unserer Darstellung der Daten ist die unabhängige Variable nicht explizit zu sehen, da wir die Daten verschiedener Gruppen bzw. Teilstichproben in verschiede Spalten einsortiert haben (siehe Beispiel unten und die Bemerkung am Ende des Kapitels). Im Rahmen der Varianzanalyse soll untersucht werden, ob die Gruppenzugehörigkeit einen Einfluss auf die Erwartungswerte hat, d.h. ob die Gruppen aus Grundgesamtheiten mit unterschiedlichen Erwartungswerten stammen. Dabei wird vorausgesetzt, dass die Gruppen jeweils aus normalverteilten Grundgesamtheiten stammen (die Voraussetzungen werden nach dem Output noch mal beschrieben). Bei k Gruppen kann im Rahmen der Varianzanalyse

H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$
gegen
H_1: Es existiert ein $j \in \{1,2,...,k\}$ mit $\mu_j \neq \mu$

getestet werden.

Wir beginnen mit einem Beispiel aus der Psychologie. Mit drei Gruppen von jeweils 5 Personen wird ein psychologischer Test durchgeführt. Gemessen wird eine stetige Größe, die wir als Testleistung bezeichnen. Wir gehen also von k = 3 Teilstichproben (Subpopulationen) aus mit jeweils gleichen Teilstichprobenumfängen ($n_1 = n_2 = n_3 = 5$), was im Allgemeinen jedoch nicht erforderlich ist. Wir wollen nun einen Unterschied in den Testleistungen zwischen den

Gruppen nachweisen. Unser Beispiel stellt eine einfaktorielle Varianzanalyse dar, denn wir wollen den Einfluss eines einzigen Faktors (die Gruppenzugehörigkeit) auf die Testleistung nachweisen.

Die Daten im Beispiel sind:

v1	v2	v3
10	8	4
15	12	8
14	7	6
12	9	7
8	14	5

Wenn Sie die Variablen v1, v2 und v3 ganz rechts im Menü unter „Welche Spalten sollen vergleichen werden:“ auswählen und dann **→Vergleich mehrerer unverbundener Teilstichproben →einfaktorielle Varianzanalyse** wählen, erhalten Sie den Output:

ANOVA

H0: Die Teilstichproben stammen aus Grundgesamtheiten mit gleichen Erwartungswerten
gegen
H1: Es existieren mindestens (mindestens) zwei Teilstichproben aus Grundgesamtheiten mit unterschiedlichen Erwartungswerten

Die Daten:

	Teilstichprobe 1	Teilstichprobe 2	Teilstichprobe 3
Beobachtung 1	10	8	4
Beobachtung 2	15	12	8
Beobachtung 3	14	7	6
Beobachtung 4	12	9	7
Beobachtung 5	8	14	5
Teilstichprobenumfänge n_j	5	5	5
Mittelwerte der Teilstichproben	11.8	10	6

ANOVA-Tabelle:

	Freiheitsgrade	Quadratsummen	Mittlere Quadratsummen	Prüfgröße ('F-Value')	p-Wert
Modell (Model)	2	88.133333333333	44.066666666667	6.8854166666667	0.0102
Fehler (Error)	12	76.8	6.4		
Gesamt (Total)	14	164.93333333333			

Es sei y_{ij} die i-te Beobachtung (i = 1, 2, …, n_j) in der j-ten Gruppe (j = 1, 2, …, k). Es wird bei der Varianzanalyse vorausgesetzt, dass die Werte y_{ij} Realisierungen von unabhängigen, normalverteilten zufälligen Größen Y_{ij} sind, mit dem Erwartungswert μ_j und der Varianz σ^2. Die Verteilungsvoraussetzungen sind also kurz: $Y_{ij} \sim N(\mu_j, \sigma^2)$. Im Beispiel ist y_{ij} die Testleistung der i-ten Person in der j-ten Gruppe. Der Gesamtstichprobenumfang ist:

$$n = n_1 + n_2 + \ldots + n_k$$

Falls die Normalverteilungsvoraussetzung nicht erfüllt ist (dies kann z.B. mit dem Kolmogorov-Smirnov-Test überprüft werden), so kann ein nichtparametrisches Verfahren verwendet werden (z.B. Kruskal-Wallis, siehe Kapitel 4.4).

Wie Sie oben erkennen können, müssen auch die Varianzen der Teilstichproben alle gleich σ^2 sein. Diese Voraussetzung der Varianzhomogenität wird auch als Homoskedastizität bezeichnet.

Die Varianzanalyse trägt ihren Namen von dem in der klassischen Varianzanalyse gemachten Ansatz der Streuungszerlegung. Dabei wird die Gesamtstreuung (SST) der Beobachtungen y_{ij} um das Gesamtmittel zerlegt in die Summe aus der Streuung zwischen den Gruppen (SSB) und der Streuung innerhalb der Gruppen (SSW).

Bei dem moderneren Ansatz der Varianzanalyse wird ein so genanntes lineares Modell formuliert, mit dem Vorteil, dass man nicht nur, wie in der klassischen Varianzanalyse, einen Einfluss der Faktorvariablen auf die Responsevariablen qualitativ nachweisen kann, sondern darüber hinaus diesen Einfluss sogar quantitativ beschreiben kann. Dabei können Unterschiede auch mit so genannten „allgemeinen linearen Hypothesen“ überprüft werden.

Auf Grund der Verteilungsvoraussetzungen lassen sich die Y_{ij} folgendermaßen darstellen:

$$Y_{ij} = \mu_j + E_{ij}\text{ , mit } E_{ij} \sim N(0, \sigma^2) \quad \text{für } i = 1, ..., n_j \text{ und } j = 1, ..., k\text{ .}$$

Die unabhängigen Zufallsvariablen E_{ij} sind die Fehlerterme bzw. Residuen.

Analog zur Regressionsanalyse verwendet man bei der einfaktoriellen Varianzanalyse ebenfalls die Varianzzerlegung um über diese einen F-Test durchführen zu können

$$SST = SSW + SSB\text{ ,}$$

mit

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2\text{ , } \bar{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}\text{ ,}$$

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2\text{ , } \bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \text{ und}$$

$$SSB = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2\text{ .}$$

Die Bezeichnungen stehen für: SST = Sum of Squares Total (Output: Total), SSW = Sum of Squares within Groups (Output: Error, entspricht SSE bei der Regressionsanalyse), SSB = Sum of Squares between Groups (Output: Model).

Es gilt:
1) SSW/σ^2 ist Realisierung einer Chi-Quadrat-verteilten Zufallsvariablen mit n - k Freiheitsgraden.
2) SSB/σ^2 ist Realisierung einer Chi-Quadrat-verteilten Zufallsvariablen mit k - 1 Freiheitsgraden.
3) SST/σ^2 ist Realisierung einer Chi-Quadrat-verteilten Zufallsvariablen mit n - 1 Freiheitsgraden.

Es folgt damit:

$$f = \frac{SSB/(k-1)}{SSW/(n-k)}$$

ist Realisierung einer F-verteilten Zufallsvariablen mit k - 1 und n - k Freiheitsgraden und somit Prüfgröße des F-Tests. Die Nullhypothese wird dann verworfen, wenn $f \geq F^{-1}_{F_{k-1,n-k}}(1-\alpha)$, bzw. wenn

$$\text{p-Wert} = 1 - F_{F_{k-1,n-k}}(f) \leq \alpha \ .$$

Bei der F-Verteilung wird bei allen Verfahren vom System der p-Wert nur dann exakt berechnet, wenn für die beiden Freiheitsgrade der F-Verteilung m_1 und m_2 die Bedingung $m_1 + m_2 \leq 240$ erfüllt ist, oder wenn $m_1 = 1$ und m_2 beliebig ist, oder wenn $m_2 = 1$ und m_1 beliebig ist.

Im Beispiel ist f = 6,8854….. und p-Wert ≈ 0,0102 ≤ 0,05, womit sich auf einem Signifikanzniveau von 5% mindestens zwei Teilstichproben hinsichtlich der Erwartungswerte unterscheiden.

Bemerkung:
Man hätte die Daten auch im großen Menü in der folgenden Form eingeben können: In der Variable v1 wird die Gruppenzugehörigkeit gespeichert, z.B. mit dem Wert 1 für die erste, 2 für die zweite und 3 für die dritte Gruppe, man könnte aber auch Buchstaben für die Gruppen verwenden. In der Variablen v2 werden dann die entsprechenden Werte in den Gruppen gespeichert.

D.h., statt

v1	v2	v3
10	8	4
15	12	8
14	7	6
12	9	7
8	14	5

wird dann

v1	v2
1	10
1	15
1	14
1	12
1	8
2	8
……..	
3	5

eingegeben. Danach kann man unter dem Button „Vergleich unverbundener Teilstichproben“ die Variable v1 als Faktorvariable und die Variable v2 als abhängige Variable auswählen und dann auf **→Vergleich unverbundener Teilstichproben** und **→Berechnung** klicken. Hier wird dann die Varianzanalyse zur Auswahl angeboten.

Umsetzung mit SAS:

```
Data dat1;
input x y;
datalines;
1  10
1  15
1  14
1  12
1   8
2   8
2  12
2   7
2   9
2  14
3   4
3   8
3   6
3   7
3   5
run;

proc anova data = dat1;
class x;
model y = x;
means x;
run;
```

SAS-Output zur Prozedur ANOVA:

Die Prozedur ANOVA

Klassifizierungsausprägungsinformationen		
Klasse	**Ausprägungen**	**Werte**
x	3	1 2 3

Anzahl gelesener Beobachtungen	15
Anzahl verwendeter Beobachtungen	15

Abhängige Variable: y

Quelle	DF	Summe der Quadrate	Mittleres Quadrat	F-Statistik	Pr > F
Modell	2	88.1333333	44.0666667	6.89	0.0102
Error	12	76.8000000	6.4000000		
Korrigierte Summe	14	164.9333333			

R-Quadrat	Koeff.var	Wurzel MSE	y Mittelwert
0.534357	27.30024	2.529822	9.266667

Quelle	DF	Anova SS	Mittleres Quadrat	F-Statistik	Pr > F
x	2	88.13333333	44.06666667	6.89	0.0102

Im nächsten Kapitel gehen wir auf die Möglichkeit der Verwendung von linearen Modellen im Rahmen der Varianzanalyse ein und wie diese mit der klassischen Varianzanalyse zusammenhängen.