Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Varianzanalyse

Einführung

Mit dem t-Test in seinen verschiedenen Varianten haben wir ein Verfahren kennengelernt, bei dem – unter der Annahme normalverteilter Merkmale – zwei Stichproben miteinander verglichen werden

- Dabei wird statistisch geprüft, ob sich die Erwartungswerte der zugehörigen Populationen unterscheiden

Häufig ist man aber am Vergleich von mehr als zwei Gruppen interessiert

Die notwendige Verallgemeinerung wird im Rahmen der *Varianzanalyse* (*analysis of variance*, ANOVA) zur Verfügung gestellt

- Die Varianzanalyse vergleicht – ebenfalls unter der Annahme normalverteilter Merkmale – die Erwartungswerte mehrerer Gruppen miteinander

Einführung: Beispiel

Es wird eine Untersuchung zur Einstellung von Jugendlichen hinsichtlich der Nutzung der Kernenergie geplant

Jugendliche sollen zufällig einer von drei Gruppen zugewiesen werden

- Die "Pro-Gruppe" sieht einen Film, der die Vorteile der Kernkraft zeigt und diese befürwortet
- Die "Kontra-Gruppe" sieht einen Film, der die Nachteile und Risiken der Kernkraft darstellt
- Die Kontrollgruppe sieht einen ausgewogenen Film zu Vor- und Nachteilen der Kernkraft

Anschließend soll die Einstellung der Jugendlichen zur Kernkraft über einen Fragebogen erfasst werden, aus dem ein zusammenfassender Wert ("Score") resultiert

Einführung: Beispiel

Nach einer Poweranalyse (s. u.) wurden folgende Scores an einer Stichprobe von Jugendlichen erhoben

	Scores									
Kontrollgruppe	8	12	7	10	11	12				
Pro-Gruppe	7	9	15	13	11	16	12	8	13	16
Kontra-Gruppe	4	5	6	3	8	10	3	9		

Anhand der Unterschiede in den Scores der Gruppen soll dann die Frage untersucht werden, wie groß der Einfluss der Informationsfilme auf die Einstellung der Jugendlichen ist

Einführung

Das obige Beispiel ist in verschiedener Hinsicht ein spezifischer Fall der im Rahmen einer Varianzanalyse zu betrachtenden Situationen

- Es wird nur eine metrische abhängige Variable betrachtet
- Es wird nur eine Einflussgröße (Art des gezeigten Films) betrachtet, die in drei Ausprägungen vorliegt
- Es interessieren die Unterschiede bezüglich der von den drei gezeigten Filmen induzierten Effekte
- Die Daten stammen aus einem Gruppenversuchsplan (Between-Subjects-Design, ohne Messwiederholung)

Varianzanalytische Verfahren

Die Varianzanalyse kann die mit der Charakterisierung des Beispiels angedeuteten allgemeinen Fälle abdecken

Univariate bzw. multivariate Varianzanalyse

- Die unvariate Varianzanalyse untersucht Effekte auf eine einzige metrische Variable
- Die multivariate Varianzanalyse untersucht simultan die Effekte auf mehrere metrische Variablen (wird im Rahmen der Multivariaten Statistik betrachtet)

Ein- und mehrfaktorielle Varianzanalyse

- Die einfaktorielle oder einfache Varianzanalyse untersucht den Einfluss eines Faktors
- Die mehrfaktorielle, oder mehrfache Varianzanalyse untersucht den Einfluss mehrerer Faktoren

Die Faktoren können dabei zwei oder mehrere Ausprägungen (Faktorstufen) aufweisen

Varianzanalytische Verfahren

Feste vs. zufällige Effekte (fixed effects vs. random effects)

- Bei festen Effekten werden ausschließlich die Effekte der in der Untersuchung verwendeten Faktorstufen betrachtet (z. B. gezeigte Filme zur Kernkraft)
- Bei zufälligen Effekten werden die in der Untersuchung verwendeten Faktorstufen aus einer Grundgesamtheit gezogen (z. B. Wörter, Rater, ...)
  Aussagen sollen hier über die unterliegende Grundgesamtheit der Faktorstufen gemacht werden

Kombinationen von Faktoren mit festen Effekten und solchen mit zufälligen Effekten werden als gemischtes Modell (mixed model) bezeichnet

Varianzanalytische Verfahren

Gruppenplan (Between-Subjects-Design) vs. Versuchsplan mit Messwiederholung (Within-Subjects-Design)

- In Between-Subjects-Designs können sämtliche Werte in den einzelnen Gruppen als voneinander unabhängig betrachtet werden
- In Within-Subjects-Designs stammen die Werte in dem betrachteten Messwiederholungsfaktor von denselben Untersuchungseinheiten und können daher nicht als stochastisch unabhängig betrachtet werden

Grundlegende Idee

Die Varianzanalyse untersucht, ob sich die Erwartungswerte verschiedener Gruppen unterscheiden, wobei sich die Gruppen durch die Stufen eines Faktors ergeben oder durch die systematische Kombination der verschiedenen Stufen mehrerer Faktoren

Die Bezeichnung Varianzanalyse ist folgendermaßen begründet

- Man betrachtet die Variabilität zwischen den Gruppen und setzt diese in Beziehung zur Variabilität innerhalb der Gruppen
- Die Variabilität innerhalb der Gruppen wird als Fehlervarianz betrachtet, die durch etwaige Gruppeneffekte nicht erklärt wird
- Die Variabilität zwischen den Gruppen hängt von der Wirksamkeit der Faktorstufen ab, sodass eine Erhöhung dieser Variabilität relativ zu der innerhalb der Gruppen für das Vorliegen von Unterschieden bei den Erwartungswerten spricht

Einfaktorielle Varianzanalyse mit festen Effekten

Wir betrachten den Einfluss eines Faktors A mit m Faktorstufen auf eine metrische Variable Y

Für jede der Faktorstufen $i = 1, \ldots, m$ wird eine unabhängige Stichprobe von n_i Untersuchungseinheiten (z. B. Versuchspersonen) gezogen

- Man hat somit die Datenwerte

$$y_{ij} \quad \text{mit} \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i$$

- Die Werte y_{ij} werden als Realisierungen von Zufallsvariablen

$$Y_{ij} \quad \text{mit} \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i$$

  betrachtet

Datenschema

Die Stichproben können in folgendem Schema veranschaulicht werden

Faktorstufe	Stichprobenwerte				$\sum$
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$y_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$y_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{in_i}	$y_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	y_{m1}	y_{m2}	$\cdots$	y_{mn_m}	$y_{m.}$
$\sum$					$y_{..}$

Zufallsvariablen

Für die zugrunde liegenden Zufallsvariablen Y_{ij} hat man dann das Schema

Faktorstufe					$\sum$
1	Y_{11}	Y_{12}	$\dots$	Y_{1n_1}	$Y_{1.}$
2	Y_{21}	Y_{22}	$\dots$	Y_{2n_2}	$Y_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	Y_{i1}	Y_{i2}	$\dots$	Y_{in_i}	$Y_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	$\dots$	Y_{mn_m}	$Y_{m.}$
$\sum$					$Y_{..}$

Modelle

Im Rahmen der Varianzanalyse wird nun ein lineares Modell formuliert, das den Einfluss des Faktors A auf die Variablen Y_{ij} mit $i = 1, \ldots, m$ und $j = 1, \ldots, n_i$ beschreibt

Es existieren mehrere verschiedene aber äquivalente Formulierungen des Modells

Alle Modellformulierungen beinhalten einen Fehlerterm ε_{ij} für $i = 1, \ldots, m$ und $j = 1, \ldots, n_i$, wie wir ihn bereits aus der Regressionsanalyse kennen

- Ebenso wie dort wird auch hier angenommen, dass die Fehler
  - stochastisch unabhängig sind
  - identisch normalverteilt gemäß $N(0, \sigma^2)$ sind (homogene Varianzen)
- Diese Annahmen übertragen sich wiederum entsprechend auf die Zufallsvariablen Y_{ij}

Modell: Alternative I

Die Y_{ij} ergeben sich aus einer durch die jeweilige Faktorstufe $i = 1, \ldots, m$ induzierten mittleren Ausprägung μ_i und individuellen Streuungen um diesen Wert, die für $j = 1, \ldots, n_i$ durch ε_{ij} bezeichnet werden

Damit ergibt sich das Modell

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{mit} \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i$$

und $\varepsilon_{ij} \sim N(0, \sigma^2)$

Daraus folgt

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

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

Modell: Alternative I

Für diese Modellformulierung hat man folgende Hypothesen

- Nullhypothese

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_m$$

- Alternativhypothese

$$H_1: \ \mu_i \neq \mu_{i'} \ \text{ für mindestens ein Paar } \ i, i' \in \{1, \ldots, m\}$$

Die Nullhypothese besagt also, dass keinerlei Unterschiede zwischen den durch die Faktorstufen gebildeten Gruppen existieren

Modell: Alternative II

Nach dieser Formulierung werden die Effekte der Faktorstufe $i = 1, \ldots, m$ auf die Y_{ij} als Abweichung von einer mittleren Ausprägung betrachtet, die über alle Gruppen hinweg ermittelt wird

Damit ergibt sich das Modell

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{mit} \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i$$

und $\varepsilon_{ij} \sim N(0, \sigma^2)$

Man bezeichnet diese Formulierung als *Modell in Effektdarstellung*

Modell: Alternative II

Den Parameter μ bezeichnet man als *grand mean* oder globalen Erwartungswert, da mit $N = n_1 + \ldots + n_m$ gilt

$$\mu = \frac{1}{N} \sum_{i=1}^{m} n_i \mu_i$$

Den Parameter α_i bezeichnet man als *Effekt* der i-ten Faktorstufe, für den gilt

$$\alpha_i = \mu_i - \mu$$

Einsetzen dieser Beziehung in die Modellgleichung zeigt die Äquivalenz zur Formulierung des Modells in Alternative I

Modell: Alternative II

Es gilt allgemein

$$\sum_{i=1}^{m} n_i \alpha_i = \sum_{i=1}^{m} n_i(\mu_i - \mu) = \sum_{i=1}^{m} n_i \mu_i - n\mu = n\mu - n\mu = 0$$

Für gleich große Gruppen $n_1 = \ldots = n_m := n$ hat man dann

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \mu_i \quad \text{und} \quad \sum_{i=1}^{m} \alpha_i = 0$$

Hypothesen

Die Hypothesen der Alternative I lassen sich unter der Effektdarstellung entsprechend reformulieren

- Nullhypothese

$$H_0: \ \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$$

- Alternativhypothese

$$H_1: \ \alpha_i \neq 0 \ \text{ für mindestens ein } \ i \in \{1, \ldots, m\}$$

Unter der Nullhypothese existieren also keinerlei Effekte der einzelnen Faktorstufen

Parameterschätzung

Grand Mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N} Y_{\cdot\cdot} = \bar{Y}$$

Gruppenspezifische Erwartungswerte

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} Y_{i\cdot} = \bar{Y}_i$$

Effekte

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{Y}_i - \bar{Y}$$

Residuen

$$\hat{\varepsilon}_{ij} = Y_{ij} - (\hat{\mu} + \hat{\alpha}_i) = Y_{ij} - (\bar{Y} + \bar{Y}_i - \bar{Y}) = Y_{ij} - \bar{Y}_i$$

Beispiel

Schätzung der gruppenspezifischen Mittelwerte $\hat{\mu}_i$ für $i = 1, 2, 3$

$$\bar{y}_1 = 10, \quad \bar{y}_2 = 12, \quad \bar{y}_3 = 6$$

Schätzung des Grand Mean $\hat{\mu}$

$$\bar{y} = \frac{1}{24}\left(n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3\right) = \frac{1}{24}\left(6 \cdot 10 + 10 \cdot 12 + 8 \cdot 6\right) = 9.5$$

Schätzung der Effekte $\hat{\alpha}_i = \bar{y}_i - \bar{y}$

$$\hat{\alpha}_1 = 10 - 9.5 = 0.5, \quad \hat{\alpha}_2 = 12 - 9.5 = 2.5, \quad \hat{\alpha}_3 = 6 - 9.5 = -3.5$$

Beispiel

Schätzung der Fehlerterme $\hat{\varepsilon}_{ij}$ für $i = 1, 2, 3$, $j = 1, \ldots, 8$, z. B.

$$\hat{\varepsilon}_{11} = y_{11} - \bar{y}_1 = 8 - 10 = -2$$

Damit ergibt sich folgende Modelldarstellung

(i,j)	y_{ij}	$=$	$\hat{\mu}$	$+$	$\hat{\alpha}_i$	$+$	$\hat{\varepsilon}_{ij}$
$(1,1)$	8	$=$	9.5	$+$	0.5	$-$	2
$(1,2)$	12	$=$	9.5	$+$	0.5	$+$	2
$\vdots$							
$(3,7)$	3	$=$	9.5	$-$	3.5	$-$	3
$(3,8)$	9	$=$	9.5	$-$	3.5	$+$	3

Streuungszerlegung

Der statistische Test der Effekte beruht – vollkommen analog zur Regressionanalyse – auf einer Zerlegung der Gesamtsumme der quadratischen Abweichungen

$$SS_T = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2$$

Die Notation SS_T bezieht sich auf die international übliche Bezeichnung *total sum of squares*

Die Gesamtsumme der quadratischen Abweichungen wird zerlegt in die Abweichungen der Gruppen voneinander ("zwischen den Gruppen") und die Abweichungen innerhalb der Gruppen

Streuungszerlegung

Man erhält

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[(Y_{ij}-\bar{Y}_i)+(\bar{Y}_i-\bar{Y})\right]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2 \\
&\quad + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)(\bar{Y}_i-\bar{Y})
\end{aligned}
$$

Streuungszerlegung

Für den letzten Summanden lässt sich zeigen, dass gilt

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)(\bar{Y}_i-\bar{Y}) &= \sum_{i=1}^{m}(\bar{Y}_i-\bar{Y})\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i) \\
&= \sum_{i=1}^{m}(\bar{Y}_i-\bar{Y})\left(\sum_{j=1}^{n_i}Y_{ij}-n_i\bar{Y}_i\right) \\
&= \sum_{i=1}^{m}(\bar{Y}_i-\bar{Y})\left(n_i\bar{Y}_i-n_i\bar{Y}_i\right) \\
&= 0
\end{aligned}$$

Streuungszerlegung

Damit folgt mit der Gleichung

$$\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = n_i(\bar{Y}_i - \bar{Y})^2$$

dann

$$\begin{aligned} SS_T &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{m} n_i(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 \\ &= \quad SS_B \quad + \quad SS_W \end{aligned}$$

Streuungszerlegung

In dieser Zerlegung ist

$$SS_B = \sum_{i=1}^{m} n_i \, (\bar{Y}_i - \bar{Y})^2$$

die gewichtete Summe der quadratischen Abweichungen der Gruppenmittel $\bar{Y}_i$ vom Gesamtmittel $\bar{Y}$ und wird daher auch als Summe der Abweichungsquadrate *zwischen* den Gruppen bezeichnet (*sum of squares between groups*)

Weiter ist

$$SS_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

die Summe der quadratischen Abweichungen der Stichprobenwerte Y_{ij} von den jeweiligen Gruppenmitteln $\bar{Y}_i$ und wird daher auch als Summe der Abweichungsquadrate *innerhalb* der Gruppen bezeichnet (*sum of squares within groups*)

Mittlere Quadrate

Aus den Summen der quadratischen Abweichungen SS_B zwischen den Gruppen und SS_W innerhalb der Gruppen bildet man durch Normierung auf die jeweilige Zahl der Freiheitsgrade die mittleren Abweichungsquadrate (*mean squares between groups* bzw. *mean squares within groups*)

Es ergibt sich

$$MS_B = \frac{SS_B}{m-1} = \frac{1}{m-1} \sum_{i=1}^{m} n_i \, (\bar{Y}_i - \bar{Y})^2$$

und

$$MS_W = \frac{SS_W}{N-m} = \frac{1}{N-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Erwartete mittlere Quadrate

Für die Erwartungswerte der mittleren Abweichungsquadrate MS_B und MS_W kann man zeigen

$$E(MS_B) = \sigma^2 + \frac{1}{m-1} \sum_{i=1}^{m} n_i \, \alpha_i^2$$

und

$$E(MS_W) = \sigma^2$$

Unter der Nullhypothese H_0: $\alpha_1 = \ldots = \alpha_m = 0$ sind MS_B und MS_W voneinander stochastisch unabhängige und erwartungstreue Schätzfunktionen für die Varianz σ^2

Für MS_W gilt dies auch unter der Alternativhypothese H_1, die Schätzfunktion MS_B wird dann aber größere Werte annehmen, die auf den von den Effekten α_i abhängigen Term zurück gehen

Teststatistik

Unter der Nullhypothese $H_0\colon\ \alpha_1 = \ldots = \alpha_m = 0$ ist dann der Quotient

$$F = \frac{\frac{MS_B}{\sigma^2}}{\frac{MS_W}{\sigma^2}} = \frac{MS_B}{MS_W}$$

F-verteilt mit $m-1$ und $N-m$ Freiheitsgraden

Mit dieser Teststatistik kann man einen Test der Nullhypothese konstruieren

Einfaktorielle Varianzanalyse mit festen Effekten

Modell

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{mit} \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

sowie $\sum_{i=1}^{m} n_i \, \alpha_i = 0$ und stochastisch unabhängigen $\varepsilon_{ij} \sim N(0, \sigma^2)$

Voraussetzungen

$$Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$$

und stochastisch unabhängig für alle $i = 1, \ldots, m$, $j = 1, \ldots, n_i$

Hypothesen

- H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$
- H_1: $\alpha_i \neq 0$ für mindestens ein $i \in \{1, \ldots, m\}$

Einfaktorielle Varianzanalyse mit festen Effekten

Schätzfunktionen

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}, \quad \hat{\alpha}_i = \bar{Y}_i - \bar{Y}, \quad \text{mit} \quad \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

Teststatistik

$$F = \frac{MS_B}{MS_W} = \frac{\frac{1}{m-1}\sum_{i=1}^{m} n_i\,(\bar{Y}_i - \bar{Y})^2}{\frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}$$

Verteilung unter H_0

$$F \sim F(m-1, N-m)$$

Ablehnungbereich

$$F > F_{1-\alpha}(m-1, N-m)$$

Beispiel

Als Schätzer für die Summe der Abweichungsquadrate zwischen den Gruppen erhält man

$$\begin{aligned} ss_B &= n_1 \cdot (\bar{y}_1 - \bar{y})^2 + n_2 \cdot (\bar{y}_2 - \bar{y})^2 + n_3 \cdot (\bar{y}_3 - \bar{y})^2 \\ &= 6 \cdot (10 - 9.5)^2 + 10 \cdot (12 - 9.5)^2 + 8 \cdot (6 - 9.5)^2 \\ &= 162 \end{aligned}$$

Damit ergibt sich für das mittlere Abweichungsquadrat zwischen den Gruppen

$$ms_B = \frac{ss_B}{m-1} = \frac{162}{3-1} = 81$$

Beispiel

Für die Summe der Abweichungsquadrate innerhalb der Gruppen erhält man die Schätzung

$$\begin{aligned} ss_W &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\ &= (8-10)^2 + (12-10)^2 + \ldots + (3-6)^2 + (9-6)^2 \\ &= 168 \end{aligned}$$

Damit ergibt sich für das mittlere Abweichungsquadrat innerhalb der Gruppen

$$ms_W = \frac{ss_W}{N-m} = \frac{168}{24-3} = 8$$

Beispiel

Für die Teststatistik berechnet man dann den Wert

$$F = \frac{ms_B}{ms_W} = \frac{81}{8} = 10.125$$

Für die Irrtumswahrscheinlichkeit $\alpha = 0.05$, sowie 2 Zähler- und 21 Nennerfreiheitsgrade erhält man als kritisches Quantil der F-Verteilung

$$F_{0.95}(2, 21) = 3.47$$

Damit ist die Nullhypothese zu verwerfen, d. h. die gezeigten Filme beeinflussen die erhobene Einstellung zur Kernkraft

Der Anteil der durch den Faktor erklärten Varianz beträgt

$$\eta^2 = ss_B/ss_T = 162/(162 + 168) = 0.49$$

Rechenvereinfachungen

Die Summe der Abstandsquadrate innerhalb der Gruppen kann man auch über die Stichprobenvarianzen S_i der einzelnen Gruppen ermitteln

Es gilt

$$SS_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{m} (n_i - 1) \cdot S_i^2$$

Für das Beispiel erhält man mit

$$s_1^2 = 4.4000, \quad s_2^2 = 10.4444, \quad s_3^2 = 7.4286$$

Damit ergibt sich

$$ss_W = (6-1) \cdot 4.4000 + (10-1) \cdot 10.4444 + (8-1) \cdot 7.4286 = 168$$

Rechenvereinfachungen

Wegen

$$SS_T = SS_B + SS_W$$

kann man auch zunächst die Summe der Abweichungsquadrate SS_T für die "gepoolte" Stichprobe sowie SS_B zwischen den Gruppen berechnen

Die Summe der Abweichungsquadrate SS_W innerhalb der Gruppen erhält man dann als Differenz $SS_T - SS_B$