Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Mittelwertvergleiche

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Mittelwertvergleiche: Übersicht

- Zwei Gruppen: t-Test
  - Unabhängige Stichproben,
    - gleiche bzw.
    - ungleiche Varianzen
  - Abhängige Stichproben
- Mehr als zwei Gruppen: Varianzanalyse
  - Unabhängige Stichproben, einfache (einfaktorielle) Varianzanalyse: Globaltest und Einzelvergleiche
  - Mehrfaktorielle Varianzanalyse
  - Varianzanalyse mit Messwiederholungen

Lernmaterialien: Neben Fahrmeir bspw. die Bücher von Bortz & Schuster (Statistik für Human- und Sozialwissenschaftler) oder Ramsey & Schafer; Vorlesungsunterlagen

t-Test: Auswahl des (richtigen) Verfahrens

- Sind die Gruppen (Stichproben, Messwerte) unabhängig oder abhängig?

  Entscheidung nach inhaltlichem Wissen über Daten und ihre Entstehung.

  Abhängigkeit besteht vor allem, wenn Daten zweimal an den gleichen Personen erhoben wurden oder wenn eine zweite Gruppe in Abhängigkeit von einer anderen Gruppe ausgewählt wurde (z. B. Einkommen von Ehepartnern).

- Varianz gleich oder ungleich? Entscheidung nach Dateninspektion und Tests.

- Daten normalverteilt? Wenn nein, evtl. Alternativen anwenden.

Anwendungsvoraussetzungen

- Der t-Test für unabhängige Stichproben funktioniert am besten bei gleich großen Gruppen mit gleicher Varianz und einer normalverteilten Variablen.
- Bei gleichen Varianzen sind auch unterschiedlich große Gruppen wenig problematisch.
- Wenn sowohl Stichprobenumfänge als auch Varianzen deutlich verschieden sind, dann ist mit mehr Fehlentscheidungen rechnen.
- Das gleiche gilt (auch bei gleich großen Stichproben), wenn die Daten nicht normal verteilt sind und ungleiche Varianzen besitzen.

t-Test für unabhängige Stichproben

Beispiel 1: Altersunterschied?

Männer	$\bar{x} = 42{,}6$	$\widehat{\sigma}^2 = 119{,}5$
Frauen	$\bar{x} = 40{,}6$	$\widehat{\sigma}^2 = 136{,}4$

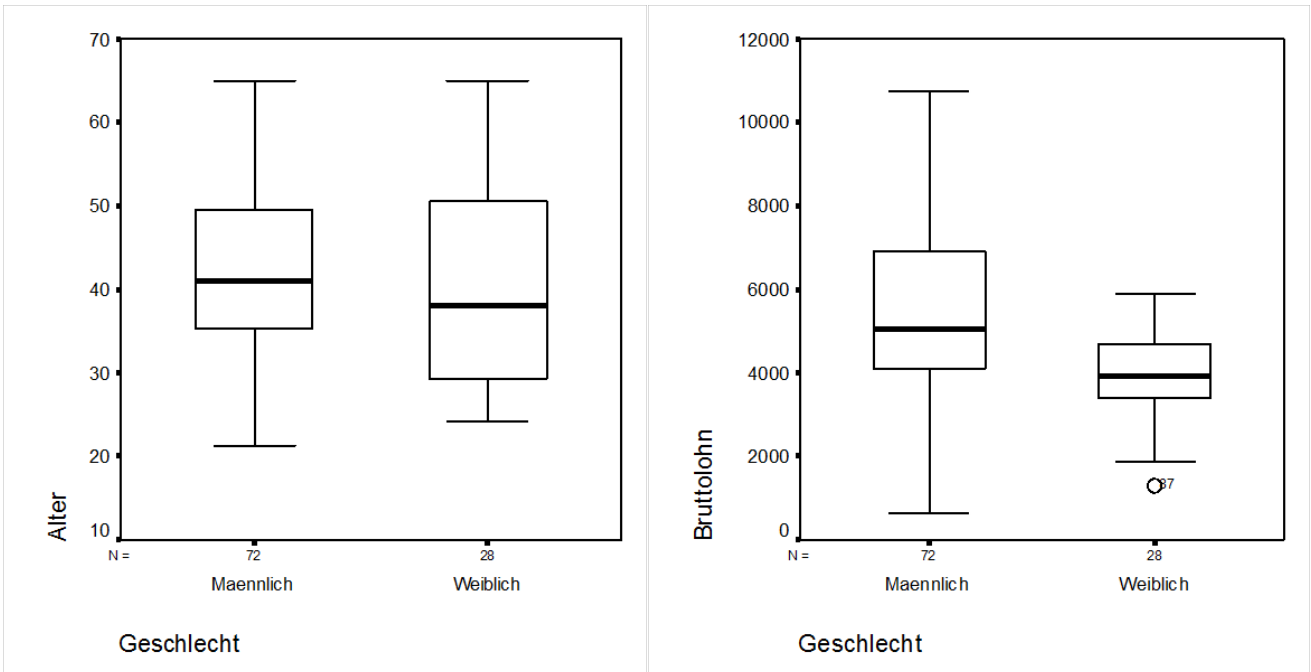
Beispiel 2: Einkommensunterschied?

Männer	$\bar{x} = 5588$	$\widehat{\sigma}^2 = 5\,184\,729$
Frauen	$\bar{x} = 3898$	$\widehat{\sigma}^2 = 1\,386\,506$

Es gilt jeweils: $n = 72$ Männer und $n = 28$ Frauen

Inspektion der Daten

Die Varianzhomogenität kann formal geprüft werden; am Anfang sollte jedoch immer visuelle Inspektion der Daten stehen (links: Alter, rechts: Einkommen).



Formaler Test auf Varianzgleichheit

Meist verwendet: der Levene-Test (wird hier nicht besprochen; siehe ILMES).

- Ist der Test nicht signifikant, heißt das, dass Nullhypothese gleicher Varianzen nicht abgelehnt werden kann $\rightarrow$ t-Test für gleiche Varianzen
- Ist der Test signifikant, wird Nullhypothese gleicher Varianzen abgelehnt $\rightarrow$ t-Test für ungleiche Varianzen

Ergebnisse des Levene-Tests:

Alter: $F = 0{,}932, p = 0{,}337 \rightarrow$ Varianzen gleich

Einkommen: $F = 8{,}6, p = 0{,}004 \rightarrow$ Varianzen ungleich.

Problem bei Einkommen außerdem: Normalverteilung fraglich. (Mögliche Lösung: Variable transformieren; oder nicht-parametrischer Test. Problem ist v. a. angesichts unterschiedlich großer Gruppen gravierend.)

t-Test: Schritt 1 und 2

Formulieren der Hypothese, Festlegen des Signifikanzniveaus (hier: 0,05) und des kritischen Wertes

Zahl der Freiheitsgrade (bei gleichen Varianzen): $n - 2$, hier also: 98.

$H_0: \quad \mu_1 = \mu_2; \quad H_1: \quad \mu_1 \neq \mu_2$

$\rightarrow$ Kritischer Wert: $t < -1{,}98$ oder $t > +1{,}98$

$H_0: \quad \mu_1 \leq \mu_2; \quad H_1: \quad \mu_1 > \mu_2$

$\rightarrow$ Kritischer Wert: $t > +1{,}66$

$H_0: \quad \mu_1 \geq \mu_2; \quad H_1: \quad \mu_1 < \mu_2$

$\rightarrow$ Kritischer Wert: $t < -1{,}66$

t-Test: Gleiche Varianzen

Die Teststatistik bei gleichen Varianzen:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \mu}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1-1)\cdot\hat{\sigma}_1^2+(n_2-1)\cdot\hat{\sigma}_2^2}{n_1+n_2-2}}}$$

mit μ als einem Wert für die Differenz, der laut H_1 überschritten werden soll (in der Regel: 0 – muss aber nicht zwingend so sein).

Im Beispiel (Alter):

$$T = \frac{(42{,}6 - 40{,}6) - 0}{\sqrt{\left(\frac{1}{72} + \frac{1}{28}\right) \cdot \frac{71\cdot 119{,}5+27\cdot 136{,}4}{72+28-2}}} = 0{,}806$$

Gleichgültig, welche H_0 formuliert wurde – die Teststatistik liegt nicht im Ablehnungsbereich.

t-Test: Ungleiche Varianzen

Die Teststatistik bei ungleichen Varianzen:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \mu}{\sqrt{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)}}$$

ist bereits aus der letzten Vorlesung bekannt.

Die Freiheitsgrade müssen wie folgt berechnet werden:

$$\text{d.f.} = \frac{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\widehat{\sigma}_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{\widehat{\sigma}_2^2}{n_2}\right)^2}$$

t-Test: Ungleiche Varianzen

Im Beispiel (Einkommen):

$$T = \frac{(5\,588 - 3\,898) - 0}{\sqrt{\left(\frac{5\,184\,729}{72} + \frac{1\,386\,505}{28}\right)}} = 4{,}85$$

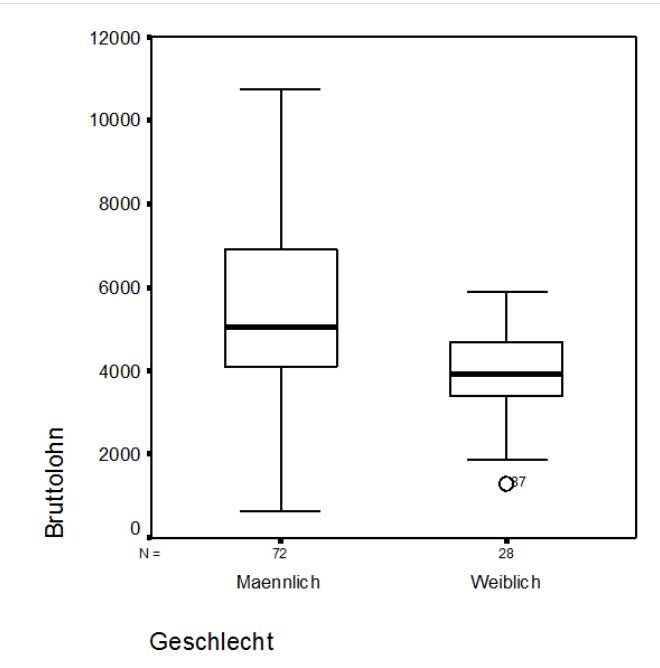
mit

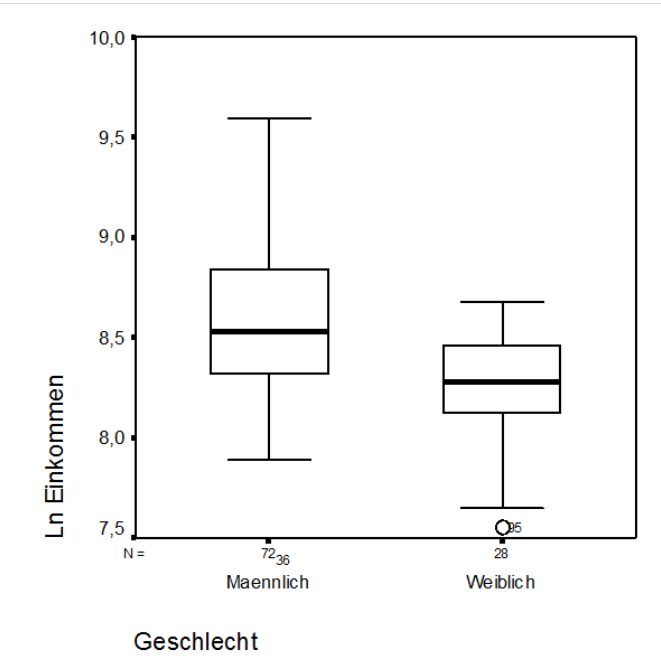
$$\frac{14\,769\,102\,527}{73\,034\,621 + 90\,816\,274} = 90{,}14 \approx 90 \text{ d. f.}$$

Die Teststatistik liegt im Ablehnungsbereich für die H_0 $\mu_1 = \mu_2$ und $\mu_1 \leq \mu_2$; für die H_0 $\mu_1 \geq \mu_2$ liegt sie dagegen nicht im Ablehnungsbereich.

Varianzungleichheit und Abweichung von Normalverteilung

Einkommensdaten: Durch Logarithmieren kann möglicherweise das Problem der Varianzungleichheit und der Abweichung von der Normalverteilung gelöst werden (links: vorher, rechts: nachher).





Varianzungleichheit und Abweichung von Normalverteilung

Der t-Test für die logarithmierten Werte:

Männer:	$\bar{x} = 8{,}542$	$\widehat{\sigma}^2 = 0{,}2009$
Frauen:	$\bar{x} = 8{,}214$	$\widehat{\sigma}^2 = 0{,}1284$

Test auf Varianzhomogenität: $F = 0{,}682, p = 0{,}411$

t-Test: $t = 3{,}46$ (d. f. $= 98$)

Die Alternative: Verteilungsfreie Tests (folgen später).

t-Test bei bekannter Varianz

Ist die Varianz des untersuchten Merkmals in der Grundgesamtheit bekannt (etwa bei standardisierten Tests), folgt die Testgröße einer Standardnormalverteilung.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

t-Test für abhängige Stichproben

Bei abhängigen Stichproben lautet die Teststatistik:

$$T = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

mit $n - 1$ Freiheitsgraden (n entspricht Zahl der Fälle, nicht der Messwerte) und μ_d als dem Wert für die Differenz, der laut H_1 überschritten werden soll. Dabei ist

$\bar{x}_d$ das arithmetische Mittel der Differenzen und

$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{x}_d)^2}{n-1}}$ die Standardabweichung der Differenzen d_i.

t-Test für abhängige Stichproben

Ein fiktives Beispiel: Punkte im Mathematiktest nach alter und nach neuer Unterrichtsmethode.

Fall-Nr.	vorher	nachher
1	1	2
2	4	6
3	6	8
4	11	12
5	17	19
6	20	22

$$T = \frac{-1{,}66667}{\frac{\mathbf{0{,}5164}}{\mathbf{2{,}449}}} = -7{,}906$$

$$s_d = \sqrt{\frac{1{,}3333}{5}} = 0{,}5164$$

Da die H_1 lautet, dass die alte Methode zu schlechteren Ergebnissen führt als die neue (also: $\mu_{\text{alt}} < \mu_{\text{neu}}$), kann die H_0 verworfen werden.

t-Test für abhängige Stichproben

Fortsetzung des Beispiels:

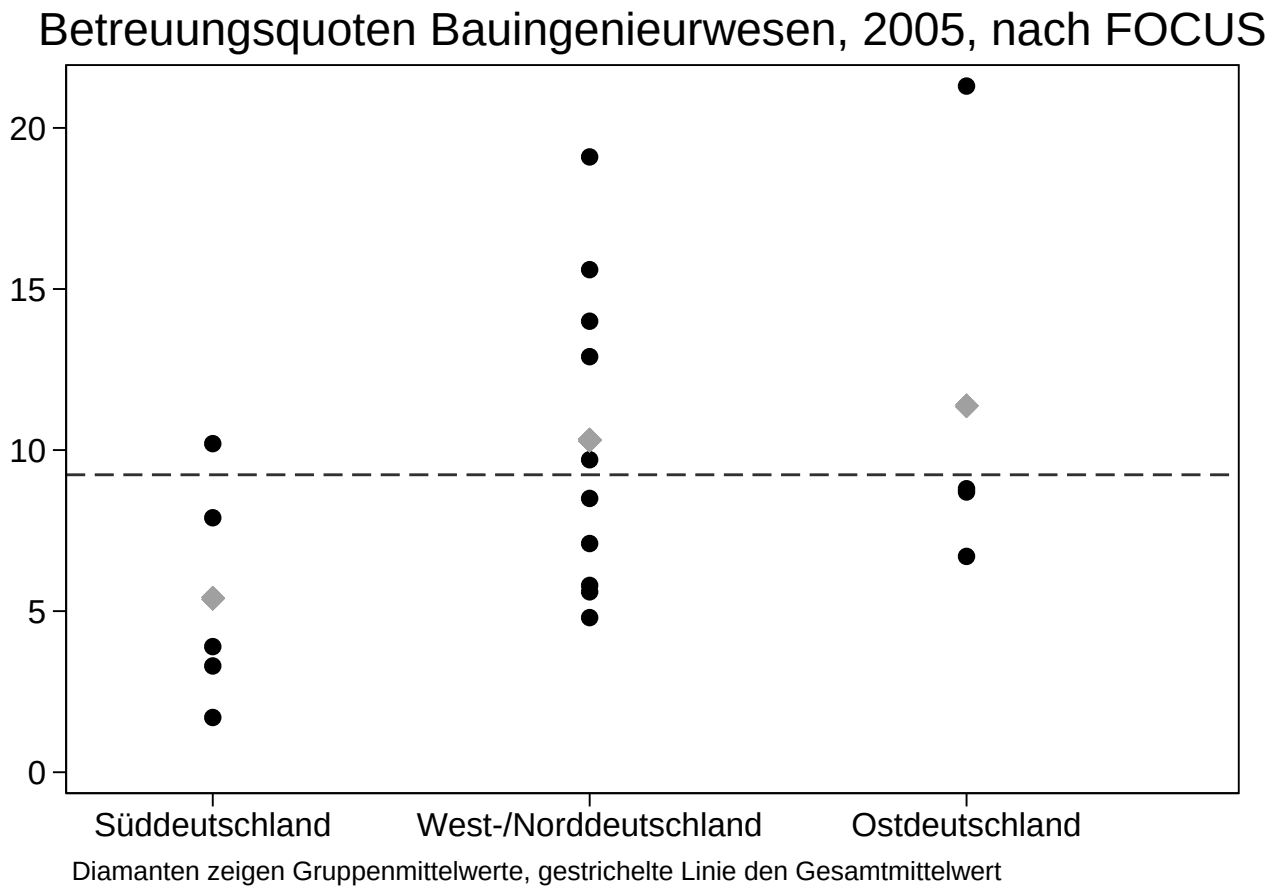
Die gleiche Datenkonstellation bei unabhängigen Stichproben bringt ein T von $-0{,}378 \rightarrow$ nicht im geringsten signifikant.

Die Veränderung ist relativ gering im Vergleich zur Unterschiedlichkeit (Varianz) der Untersuchungsobjekte . . . aber diese Unterschiedlichkeit interessiert hier nicht, sondern nur die (geringe) Änderung aufgrund der Wirkung der neuen Unterrichtsmethode. Und diese Änderung ist sehr gleichartig (hat wenig Streuung).

Varianzanalyse: die Fragestellung

- Die Varianzanalyse kann Mittelwerte von mehr als zwei Gruppen vergleichen.
- Bei mehr als zwei Gruppen sind u. U. Fragen des Testens globaler Unterschiede (unterscheiden sich irgendwelche Mittelwerte) vs. spezifischer Unterschiede zu lösen.
- Mit der Varianzanalyse können auch Einflüsse mehrerer Gruppierungsvariablen analysiert werden (z. B. zwei Therapien bei zwei verschiedenen Krankheitsformen) $\rightarrow$ mehrfaktorielle Varianzanalyse, hier nicht behandelt.
- Es gibt auch Verfahren für abhängige Stichproben (Messwiederholungen); ebenfalls hier nicht besprochen.

Varianzanalyse – illustriert



Achtung – Daten dienen nur der Illustration, Gültigkeit ist fraglich!

Varianzanalyse – die Grundidee verbal

Die Unterschiedlichkeit (Varianz) der Datenwerte kann in zwei Teile „zerlegt“ werden:

- Die Unterschiedlichkeit, die auf die Gruppenzugehörigkeit zurückgeht, ausgedrückt in den Abweichungen der Gruppenmittelwerte $\bar{y}_i$ vom Gesamtmittelwert $\bar{y}$.
- Die Unterschiedlichkeit, die nicht auf die Gruppenzugehörigkeit zurückgeht, ausgedrückt in den Abweichungen der individuellen Messwerte vom jeweiligen Gruppenmittelwert.

Varianzanalyse – die Grundidee formal

- Gegeben sind i, $i = 1 \ldots r$ Gruppen
- In jeder Gruppe werden Daten von j, $j = 1 \ldots m$ Personen erhoben (d. h., pro Gruppe gleich viele Personen – wichtige Vereinfachung, die bei experimentellen Studien oft befolgt wird).
- Die Summe aller quadrierten Abweichungen vom Mittelwert („Quadratsumme“, QS) lässt sich zerlegen in

$$\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2 = m\sum_{i=1}^{r}(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2$$

oder: $QS_{\text{total}} = QS_{\text{zwischen}} + QS_{\text{innerhalb}}$

Zur Erinnerung: η^2

Das Verhältnis der durch die Gruppenzugehörigkeit bedingten Abweichungen vom Mittelwert (QS_{zwischen}) zur Gesamtheit der Abweichungen ist ein Maß für die Stärke des Einflusses der Gruppenzugehörigkeit:

$$\eta^2 \text{ (Eta-Quadrat)} = \frac{\text{QS}_{\text{zwischen}}}{\text{QS}_{\text{total}}}$$

Es handelt sich mithin um ein PRE-Maß.

Inferenzstatistik global I

Zur Prüfung, ob sich die Gruppen überzufällig unterscheiden, werden nicht die Quadratsummen, sondern die Varianzen zu einander in Beziehung gesetzt. Diese heißen hier auch „mittlere Quadratsummen“ (MQS).

$$\text{MQS}_{\text{zwischen}} = \frac{\text{QS}_{\text{zwischen}}}{r-1} = \frac{m \sum_{i=1}^{r} (\bar{y}_i - \bar{y})^2}{r-1}$$

$$\text{MQS}_{\text{innerhalb}} = \frac{\text{QS}_{\text{innerhalb}}}{n-r} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2}{n-r}$$

Inferenzstatistik global II

Die Größe

$$F = \frac{\text{MQS}_{\text{zwischen}}}{\text{MQS}_{\text{innerhalb}}}$$

folgt einer F-Verteilung mit r-1 und n-r Freiheitsgraden. Sie prüft, ob sich irgendwelche Gruppenmittelwerte voneinander unterscheiden, also

$$\text{H}_0: \quad \mu_1 = \mu_2 = \ldots = \mu_r = \mu$$

$$\text{H}_1: \quad \mu_i \neq \mu \quad \text{für mindestens ein i}$$

Einzelvergleiche: Das Problem

Der globale F-Test sagt nur aus, dass sich irgendwelche Gruppen in irgendeiner Art unterscheiden. Für speziellere Hypothesen können herangezogen werden:

- A priori-Vergleiche durch Bildung von Kontrasten
- A posteriori-Vergleiche durch spezielle Teststatistiken, die für multiple Vergleiche geeignet sind.

Einzelvergleiche a priori

Vergleich einzelner Gruppen aufgrund theoretischer Annahmen.

Vorgehen: Bilden von Kontrasten durch Linearkombinationen

$$g = c_1\bar{y}_1 + c_2\bar{y}_2 + \ldots c_r\bar{y}_r \text{ mit } \sum_{i=1}^{r} c_i = 0$$

Beispiele:

$$g = -1 \cdot \bar{y}_1 + 1 \cdot \bar{y}_2 + 0 \cdot \bar{y}_3 \text{ für } H_1: \quad \mu_1 < \mu; \mu_2 > \mu; \mu_3 = \mu$$

$$g = -0{,}5 \cdot \bar{y}_1 - 0{,}5 \cdot \bar{y}_2 + 1 \cdot \bar{y}_3 \text{ für } H_1: \quad \mu_1, \mu_2 < \mu; \mu_3 > \mu$$

Einzelvergleiche a priori

Der Standardfehler von g beträgt:

$$SE(g) = \sqrt{\text{MQS}_{\text{innerhalb}}} \cdot \sqrt{\frac{c_1^2}{m} + \frac{c_2^2}{m} + \ldots \frac{c_r^2}{m}}$$

Die Statistik $\frac{g}{SE(g)}$ folgt einer t-Verteilung mit n-r Freiheitsgraden.

Einzelvergleiche a priori: Ein Beispiel

Daten und Hypothesen nach Ramsey & Schafer, S. 150. Geprüft werden soll folgende Annahme: Gruppe 1 unterscheidet sich nicht vom Gesamtmittelwert, Gruppe 2 und 4 liegen unter diesem, Gruppe 3 und 5 darüber.

Display 6.1 Stem-and-leaf diagrams of applicant qualification scores given to applicants simulating five different handicap conditions

	None	Amputee	Crutches	Hearing	Wheelchair
0					
1	9	9		4	7
2	5	56		149	8
3	06	268	7	479	5
4	129	06	033	237	78
5	149	3589	18	589	03
6	17	1	0234	5	1124
7	48	2	445		246
8			5		
9					

Legend: 7 | 4 represents a score of 7.4 on the Applicant Qualification Scale.

Einzelvergleiche a priori: Ein Beispiel

Daten und Hypothesen nach Ramsey & Schafer, S. 150. Geprüft werden soll folgende Annahme: Gruppe 1 unterscheidet sich nicht vom Gesamtmittelwert, Gruppe 2 und 4 liegen unter diesem, Gruppe 3 und 5 darüber.

$$\begin{aligned} g &= 0\bar{y}_1 + (-0{,}5\bar{y}_2) + 0{,}5\bar{y}_3 + (-0{,}5\bar{y}_4) + 0{,}5\bar{y}_5 \\ &= 0 \cdot 4{,}900 - 0{,}5 \cdot 4{,}429 + \ 0{,}5 \cdot 5{,}921 - 0{,}5 \cdot 4{,}050 + 0{,}5 \cdot 5{,}343 \\ &= 1{,}393 \end{aligned}$$

$$SE(g) = \sqrt{2{,}666} \cdot \sqrt{\frac{0^2 + (-0{,}5)^2 + 0,5^2 + (-0{,}5)^2 + 0,5^2}{14}} = 0{,}436$$

$$t = \frac{1{,}393}{0{,}436} = 3{,}192 > 1{,}67 \quad \text{(t-Verteilung, 65 d. f.)}$$

Es handelt sich formal um einen einseitigen Test (H_0: Die postulierten Unterschiede treffen nicht zu) mit Ablehnungsbereich $1 - \alpha$.

Einzelvergleiche a posteriori

Das Problem: Wird jede Gruppe mit jeder anderen verglichen, werden viele [genauer: $(n-1)!$] Tests durchgeführt $\rightarrow$ gesucht ist eine Korrektur für multiples Testen.

Das Vorgehen: Es wird eine „kritische Differenz“ (oder Grenzdifferenz) berechnet, die einen Korrekturfaktor für das mehrfache Testen enthält. Überschreitet die Stichprobendifferenz zwischen zwei Messwerten diese kritische Differenz, so wird angenommen, dass auch die betreffende Differenz in der Grundgesamtheit von 0 verschieden ist.

Es gibt eine erhebliche Menge von Vorschlägen zur Berechnung dieser kritischen Differenz. Der im Folgenden besprochene Scheffé-Test gilt als konservativ, d. h., er stellt hohe Anforderungen an die Anerkennung einer Differenz als signifikant.

Einzelvergleiche a posteriori: Der Scheffé-Test

Die kritische Differenz wird berechnet als

$$D_{\text{Scheffe}} = \sqrt{MQS_{\text{innerhalb}} \cdot \left(\frac{1}{m} + \frac{1}{m}\right) \cdot (r-1) \cdot F_{r-1,\, n-r;\, 1-\alpha}}$$

Im Beispiel (α wieder mal 0,05):

$$D_{\text{Scheffe}} = \sqrt{2{,}666 \cdot \left(\frac{1}{14} + \frac{1}{14}\right) \cdot 4 \cdot 2{,}513} = 1{,}957$$

Der größte Abstand zwischen zwei Gruppen beträgt 1,871. Kein Abstand überschreitet also die kritische Differenz $\rightarrow$ kein signifikanter Einzelunterschied!

Abschließende Bemerkungen

Die meisten vorgestellten Berechnungsmethoden funktionieren auch bei unterschiedlichen Gruppengrößen (statt einheitlicher Größe m werden dann n_1, n_2 ... n_r verwendet).

Das gilt aber nicht grundsätzlich. Auch werden Probleme fehlender Normalverteilung bzw. Varianzhomogenität der Daten durch ungleiche Gruppengrößen verstärkt.

Ausführliche Diskussionen der Anwendungsvoraussetzungen bei Bortz & Schuster.

Literatur

Ramsey, Fred L. & Schafer, Daniel W.: The Statistical Sleuth. A Course in Methods of Data Analysis. Pacific Grove, CA: Duxbury, 2. Aufl. 2002.

Die Original-Untersuchung, auf die Ramsey & Schafer sich beziehen: Cesare, S. J. et al.: Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy, in: Human Performance 3 (3), 1990: 157-171.