WEIHENSTEPHAN · TRIESDORF
University of Applied Sciences

Prof. Dr. Klaus Eckhardt

Multiple Mittelwertvergleiche

veröffentlicht im Internet unter aufgabomat.de

Inhalt

1 Einleitung 1
2 Drei Verfahren für paarweise Vergleiche 2
2.1 Multipler t-Test mit Bonferroni-Korrektur 2
2.2 Scheffé-Test 5
2.3 Tukey-Kramer-Test 5

1 Einleitung

Gegeben seien I Stichproben der Umfänge N_i (i = 1, …, I). Selbst falls die Werte aller dieser Stichproben ein und derselben Grundgesamtheit entstammen, werden die empirischen Mittelwerte $\overline{x}_i$ im Allgemeinen von Stichprobe zu Stichprobe variieren. Unterschiedliche empirische Mittelwerte können allerdings auch darauf hinweisen, dass die Stichproben mehreren Grundgesamtheiten entstammen, deren Erwartungswerte $E(X_i)$ sich unterscheiden. Mittelwertvergleiche sind Verfahren, mit denen geklärt werden soll, von welcher dieser Alternativen auszugehen ist.

Eines dieser Verfahren sollte Ihnen bereits bekannt sein, der t-Test. Er wird angewendet, falls genau zwei Stichproben vorliegen, die unterschiedliche empirische Mittelwerte aufweisen, und falls davon ausgegangen werden kann, dass die Daten in beiden Stichproben normalverteilt mit derselben Varianz sind. Könnte der t-Test nicht auch bei Vorliegen einer größeren Anzahl von Stichproben angewendet werden, um jeweils paarweise Vergleiche durchzuführen?

Nehmen wir an, aus I = 3 Stichproben ergeben sich drei unterschiedliche empirische Mittelwerte und es werden daher drei t-Tests mit den Nullhypothesen $E(X_1) = E(X_2)$, $E(X_1) = E(X_3)$ und $E(X_2) = E(X_3)$ durchgeführt. Die Irrtumswahrscheinlichkeit sei beispielsweise $\alpha = 0{,}05$. Als Ereignis A_i sei definiert, dass die Nullhypothese im i-ten Test irrtümlich verworfen wird. Die Wahrscheinlichkeit dafür ist in jedem Test α: $P(A_1) = 0{,}05$, $P(A_2) = 0{,}05$, $P(A_3) = 0{,}05$. Die Wahrscheinlichkeit dafür, die Nullhypothese in mindestens einem der Tests irrtümlich zu verwerfen (die **multiple Irrtumswahrscheinlichkeit**), berechnet man am besten als 1 minus die Wahrscheinlichkeit, die Nullhypothese in keinem der Tests irrtümlich zu verwerfen. Unter Zuhilfenahme des Multiplikationssatzes der Wahrscheinlichkeitsrechnung ergibt sich:

$$\begin{aligned} 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3) &= 1 - P(\overline{A}_1)\, P(\overline{A}_2)\, P(\overline{A}_3) \\ &= 1 - (1 - 0{,}05)^3 \\ &= 0{,}14. \end{aligned}$$

Die Irrtumswahrscheinlichkeit für die Gesamtheit der drei Tests ist also deutlich größer als 0,05. Allgemein gilt:

- Liegen I Stichproben vor, so beträgt die Anzahl m der paarweise durchzuführenden t-Tests bis zu $\binom{I}{2} = \frac{I\,(I-1)}{2}$ ($\rightarrow$ Kombinatorik, Anzahl Kombinationen ohne Wiederholung).
- Die multiple Irrtumswahrscheinlichkeit bei m paarweisen Vergleichen ist $1 - (1 - \alpha)^m$ und geht damit für $m \rightarrow \infty$ gegen 1.

Multiple t-Tests allein sind also nicht die Lösung. Außerdem sind auch dann Mittelwertvergleiche gefragt, wenn die Daten nicht normalverteilt sind und/oder die Voraussetzung der Varianzhomogenität nicht erfüllt ist. Aus diesen Gründen ist eine große Anzahl unterschiedlicher Verfahren für Mittelwertvergleiche entwickelt worden. Nur drei Verfahren werden nachfolgend vorgestellt: der multiple t-Test mit Bonferroni-Korrektur (Abschnitt 2.1), der Scheffé-Test (Abschnitt 2.2) und der Tukey-Kramer-Test (Abschnitt 2.3). Dies geschieht wiederum nur für paarweise Vergleiche, d. h. für Vergleiche von jeweils genau zwei Stichproben. Auf den ebenfalls möglichen Vergleich ganzer Gruppen von Stichproben wird nicht eingegangen.

Das Anwendungsbeispiel ist dasselbe, das im Skript „Varianzanalyse“ unter aufgabomat.de zur Veranschaulichung der einfaktoriellen Varianzanalyse (ANOVA) angeführt wird. Überhaupt werden Mittelwertvergleiche typischerweise im Anschluss an eine Varianzanalyse vorgenommen, bei der es zu einer Ablehnung der Nullhypothese kam. Man spricht in diesem Fall von **a-posteriori-Mittelwertvergleichen** oder **post-hoc-Mittelwertvergleichen**.

Warum aber wird auf die Varianzanalyse nicht ganz verzichtet, warum wird sie nicht einfach durch einen Mittelwertvergleich ersetzt? Die Antwort lautet: Die sehr zahlreichen Verfahren, die für Mittelwertvergleiche entwickelt worden sind (Scheffé, Tukey-Kramer, Student-Newman-Keuls, Duncan, Holm, Games-Howell, Dunnett, …), können durchaus voneinander abweichende Ergebnisse liefern. Sie sind unterschiedlich konservativ, d. h. neigen mehr oder weniger stark dazu, Mittelwertunterschiede – eventuell auch fälschlicherweise – anzuzeigen oder aber – eventuell fälschlicherweise – nicht zu erkennen. Eine vorangehende Varianzanalyse schafft eine erhöhte Sicherheit, dass sich die Erwartungswerte mindestens zweier Stichprobenvariablen tatsächlich unterscheiden, nämlich mindestens derjenigen Stichproben, deren empirische Mittelwerte am stärksten voneinander abweichen. Es lässt sich dann eher vertreten, den Mittelwertvergleich mit einem der weniger konservativen Verfahren durchzuführen, die stärker dazu tendieren, Unterschiede auszuweisen. Dazu kommt, dass im Scheffé- und Tukey-Kramer-Test ohnehin auf eine Größe zurückgegriffen werden muss (die zufallsbedingte empirische Varianz bzw. Restvarianz), deren Berechnung bereits einen Großteil der Varianzanalyse ausmacht.

2 Drei Verfahren für paarweise Vergleiche

2.1 Multipler t-Test mit Bonferroni-Korrektur

Beim multiplen bzw. m-fachen t-Test mit Bonferroni-Korrektur gelten dieselben Voraussetzungen wie beim einfachen t-Test. Gegeben seien I Stichproben normalverteilter Variablen X_i gleicher Varianz (i = 1, …, I). Die empirischen Mittelwerte $\bar{x}_i$ und $\bar{x}_j$ mindestens zweier der Stichproben seien unterschiedlich. Bei der m-fachen paarweisen Durchführung des t-Tests soll die multiple Irrtumswahrscheinlichkeit $\alpha = 1 - (1 - \alpha')^m$ eingehalten werden[1]. Die Bonferroni-Korrektur besteht darin, jeden der m t-Tests mit der Irrtumswahrscheinlichkeit

$$\begin{aligned} \alpha' &= 1 - (1 - \alpha)^{1/m} \\ &\approx \alpha/m \end{aligned} \qquad (1)$$

durchzuführen.

Beispiel: Es liegen I = 3 Stichproben mit unterschiedlichen empirischen Mittelwerten vor.

Die Anzahl der paarweise durchzuführenden t-Tests ist

$$\begin{aligned} m &= \binom{3}{2} \\ &= \frac{3!}{2!\,(3-2)!} \\ &= 3. \end{aligned}$$

Gewünscht wird die multiple Irrtumswahrscheinlichkeit α = 0,05. Gemäß Bonferroni-Korrektur wird dazu jeder einzelne t-Test mit der Irrtumswahrscheinlichkeit

$$\alpha' = 1 - (1 - 0{,}05)^{1/3} = 0{,}017$$

durchgeführt. Vereinfacht lässt sich α' auch berechnen als

$$\alpha' = 0{,}05/3 = 0{,}01\overline{6}.$$

[1] Zur Irrtumswahrscheinlichkeit beim multiplen t-Test siehe Abschnitt 1.

Sei A_i das Ereignis, dass die Nullhypothese im i-ten Test irrtümlich verworfen wird. Dann ergibt sich als multiple Irrtumswahrscheinlichkeit wie gefordert

$$1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3) = 1 - (1 - 0{,}017)^3 = 0{,}05.$$

Nun zur Durchführung der einzelnen t-Tests. Seien $\bar{x}_i \neq \bar{x}_j$ die empirischen Mittelwerte in zwei Stichproben i und j mit den Umfängen N_i und N_j und den empirischen Varianzen s_i^2 und s_j^2. Null- und Alternativhypothese des t-Tests lauten
H_0: $E(X_i) = E(X_j)$
H_1: $E(X_i) \neq E(X_j)$.

Der Wert der Teststatistik bzw. Prüfwert beim Zweistichproben-t-Test ist

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j}}}. \qquad (2)$$

Das Annahmeintervall $[t_{\alpha'/2}; t_{1-\alpha'/2}]$ wird durch zwei Quantile der t-Verteilung mit dem Freiheitsgrad

$$f = N_i + N_j - 2 \qquad (3)$$

begrenzt. Die Nullhypothese H_0 wird beibehalten, falls der Wert der Teststatistik im Annahmeintervall liegt bzw. wenn

$$t_{\frac{\alpha'}{2}} \leq \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j}}} \leq t_{1-\frac{\alpha'}{2}}.$$

Da $t_{\alpha'/2} = -\,t_{1-\alpha'/2}$ ist (Abbildung 1), lässt sich diese Bedingung auch schreiben als

$$\frac{|\,\bar{x}_i - \bar{x}_j\,|}{\sqrt{\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j}}} \leq t_{1-\frac{\alpha'}{2}}.$$

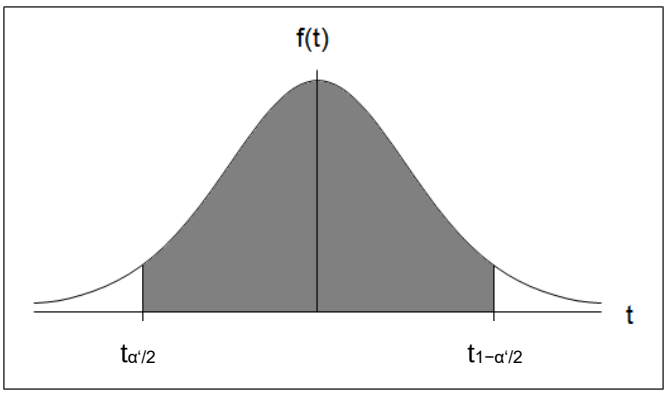

Abbildung 1: Wahrscheinlichkeitsdichtefunktion der t-Verteilung.

Umgekehrt lässt sich formulieren: H_0 wird verworfen, falls

$$|\bar{x}_i - \bar{x}_j| > t_{1-\frac{\alpha'}{2}} \sqrt{\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j}} \,. \tag{4}$$

Man sagt auch, dass die Nullhypothese verworfen wird, falls $|\bar{x}_i - \bar{x}_j|$ größer als die **Grenzdifferenz**

$$g_{ij} = t_{1-\frac{\alpha'}{2}} \sqrt{\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j}} \tag{5}$$

ist.

Beispiel: Es wird dasselbe Beispiel betrachtet, das im Skript „Varianzanalyse“ unter aufgabomat.de zur Veranschaulichung der einfaktoriellen Varianzanalyse angeführt wird. Drei Weizensorten sollen hinsichtlich ihres Ertrags X verglichen werden. Dazu werden die Weizensorten auf Versuchsparzellen angebaut. Am Ende der Wachstumsperiode liegen Daten von 11 Parzellen vor (Tabelle 1).

Sorte	Ertrag (dt/ha)			
1	81	72	67	69
2	85	85	64	72
3	55	61	58	

Tabelle 1: Daten des Sortenversuchs.

Die empirischen Mittelwerte des Ertrags sind in allen drei Stichproben unterschiedlich. Es wird eine einfaktorielle Varianzanalyse mit der Nullhypothese H_0 durchgeführt, dass der Erwartungswert des Ertrags unabhängig von der Sorte ist.

Die Varianzanalyse führt zur Ablehnung von H_0. Der Erwartungswert des Ertrags mindestens einer Sorte unterscheidet sich von dem der anderen Sorten. Wie viele Sorten und welche Sorten sich in dieser Weise auszeichnen, darüber gibt die Varianzanalyse keine Information, sodass ein multipler Mittelwertvergleich angeschlossen wird, hier ein multipler t-Test mit Bonferroni-Korrektur. Neben den empirischen Mittelwerten müssen, um die Grenzdifferenz ermitteln zu können (Gleichung 5), auch die empirischen Varianzen s_i^2 bzw. die empirischen Standardabweichungen s_i berechnet werden. Die Ergebnisse sind Tabelle 2 zu entnehmen.

Sorte	Stichprobenumfang N_i	$\bar{x}_i$ (dt/ha)	s_i (dt/ha)
1	4	72	6,18
2	4	77	10,3
3	3	58	3,00

Tabelle 2: Empirische Mittelwerte und empirische Standardabweichungen.

Die multiple Irrtumswahrscheinlichkeit soll $\alpha = 0{,}05$ betragen. Bei $m = 3$ paarweisen Vergleichen müssen die einzelnen Vergleiche gemäß Gleichung 1 mit $\alpha' = 0{,}05/3 = 0{,}017$ durchgeführt werden. Dann ist $1 - \alpha'/2 = 0{,}992$. Für den Vergleich zwischen den Sorten 1 und 2 ist das 0,992-Quantil der t-Verteilung mit $f = 6$ zu bestimmen. Dieses hat den Wert 3,57, wie sich beispielsweise mit der Excel-Funktion T.INV berechnen lässt. Für die Vergleiche zwischen den Sorten 1 und 3 sowie 2 und 3 wird das 0,992-Quantil der t-Verteilung mit $f = 5$ benötigt. Sein Wert beträgt 3,32. Damit ergeben sich die in Tabelle 3 aufgeführten Werte der Grenzdifferenz.

Sorten	$\lvert \bar{x}_i - \bar{x}_j \rvert$ (dt/ha)	Grenzdifferenz (dt/ha)	signifikanter Unterschied	
			ja	nein
1 und 2	5	21		x
1 und 3	14	12	x	
2 und 3	19	18	x	

Tabelle 3: Mittelwertvergleich (multipler t-Test mit Bonferroni-Korrektur) für die Daten des Parzellenversuchs.

Bei zwei Vergleichen ist der Betrag der Differenz zwischen den empirischen Mittelwerten größer als die Grenzdifferenz, sodass die Schlussfolgerung lautet, dass sowohl Sorte 1 als auch Sorte 2 hinsichtlich des Ertrags der Sorte 3 vorzuziehen sind.

2.2 Scheffé-Test

Der Scheffé-Test für paarweise Vergleiche wird mit der Grenzdifferenz

$$g_{ij} = \sqrt{(I-1)\left(\frac{1}{N_i}+\frac{1}{N_j}\right) s_R{}^2 \, F_{1-\alpha}} \tag{6}$$

durchgeführt. Darin sind:

- N_i, N_j: Umfänge der beiden miteinander zu vergleichenden Stichproben i und j
- s_R^2: zufallsbedingte empirische Varianz. Diese wird im Rahmen der Varianzanalyse berechnet[2].
- $F_{1-\alpha}$: (1−α)-Quantil der F-Verteilung mit den Freiheitsgraden $f_1 = I - 1$ und $f_2 = N - I$.

H_0 wird verworfen, falls $\lvert \bar{x}_i - \bar{x}_j \rvert > g_{ij}$.

Beispiel: Ertrag dreier unterschiedlicher Weizensorten auf insgesamt 11 Versuchsparzellen (Tabelle 1)
$s_R^2 = 54$ (dt/ha)^2
$\alpha = 0{,}05$
0,95-Quantil der F-Verteilung mit $f_1 = 2$ und $f_2 = 8$: $F_{0,95} = 4{,}5$

Sorten	$\lvert \bar{x}_i - \bar{x}_j \rvert$ (dt/ha)	Grenzdifferenz (dt/ha)	signifikanter Unterschied	
			ja	nein
1 und 2	5	16		x
1 und 3	14	17		x
2 und 3	19	17	x	

Tabelle 4: Mittelwertvergleich (Scheffé-Test) für die Daten des Parzellenversuchs.

Der Scheffé-Test ist generell sehr konservativ. Im Beispiel wird nur ein Unterschied der Sorten 2 und 3 ausgewiesen.

2.3 Tukey-Kramer-Test

Der Tukey-Kramer-Test für paarweise Vergleiche wird mit der Grenzdifferenz

[2] Siehe Abschnitt 2 im Skript „Varianzanalyse“ unter aufgabomat.de.

$$g_{ij} = q_{1-\alpha} \sqrt{\frac{1}{2}\left(\frac{1}{N_i} + \frac{1}{N_j}\right) {s_R}^2} \qquad (7)$$

durchgeführt. Darin sind:

- N_i, N_j: Umfänge der beiden miteinander zu vergleichenden Stichproben i und j
- s_R^2: zufallsbedingte empirische Varianz. Diese wird im Rahmen der Varianzanalyse berechnet[3].
- $q_{1-\alpha}$: (1−α)-Quantil der Verteilung der studentisierten Variationsbreite mit den Parametern I und f = N – I

H_0 wird verworfen, falls $|\bar{x}_i - \bar{x}_j| > g_{ij}$.

Werte für die Quantile der Verteilung der studentisierten Variationsbreite können Tabelle 5 entnommen werden. Eine Excel-Funktion zu ihrer Berechnung gibt es nicht.

f	p	I												
		2	3	4	5	6	7	8	9	10	11	12	15	20
2	0,950	6,1	8,3	9,8	11	12	12	13	14	14	14	15	16	17
	0,990	14	19	22	25	27	28	30	31	32	33	33	35	38
3	0,950	4,5	5,9	6,8	7,5	8,0	8,5	8,9	9,2	9,5	9,7	9,9	11	11
	0,990	8,3	11	12	13	14	15	16	16	17	17	18	19	20
4	0,950	3,9	5,0	5,8	6,3	6,7	7,1	7,3	7,6	7,8	8,0	8,2	8,7	9,1
	0,990	6,5	8,1	9,2	10	11	11	12	12	12	13	13	14	14
5	0,950	3,6	4,6	5,2	5,7	6,0	6,3	6,6	6,8	7,0	7,2	7,3	7,7	8,2
	0,990	5,7	7,0	7,8	8,4	8,9	9,3	9,7	10	10	10	11	11	12
6	0,950	3,5	4,3	4,9	5,3	5,6	5,9	6,1	6,3	6,5	6,6	6,8	7,1	7,6
	0,990	5,2	6,3	7,0	7,6	8,0	8,3	8,6	8,9	9,0	9,3	9,5	10	11
7	0,950	3,3	4,2	4,7	5,1	5,4	5,6	5,8	6,0	6,2	6,3	6,4	6,8	7,2
	0,990	5,0	5,9	6,5	7,0	7,4	7,7	7,9	8,2	8,4	8,5	8,7	9,1	10
8	0,950	3,3	4,0	4,5	4,9	5,2	5,4	5,6	5,8	5,9	6,1	6,2	6,5	6,9
	0,990	4,7	5,6	6,2	6,6	7,0	7,2	7,5	7,7	7,9	8,0	8,2	8,6	9,0
9	0,950	3,2	3,9	4,4	4,8	5,0	5,2	5,4	5,6	5,7	5,9	6,0	6,3	6,6
	0,990	4,6	5,4	6,0	6,3	6,7	6,9	7,1	7,3	7,5	7,6	7,8	8,1	8,6
10	0,950	3,2	3,9	4,3	4,7	4,9	5,1	5,3	5,5	5,6	5,7	5,8	6,1	6,5
	0,990	4,5	5,3	5,8	6,1	6,4	6,7	6,9	7,1	7,2	7,4	7,5	7,8	8,2
11	0,950	3,1	3,8	4,3	4,6	4,8	5,0	5,2	5,4	5,5	5,6	5,7	6,0	6,3
	0,990	4,4	5,1	5,6	6,0	6,2	6,5	5,7	6,8	7,0	7,1	7,3	7,6	8,0
12	0,950	3,1	3,8	4,2	4,5	4,8	5,0	5,1	5,3	5,4	5,5	5,6	5,9	6,2
	0,990	4,3	5,0	5,5	5,8	6,1	6,3	6,5	6,7	6,8	6,9	7,1	7,4	7,7
15	0,950	3,0	3,7	4,1	4,4	4,6	4,8	4,9	5,1	5,2	5,3	5,4	5,6	6,0
	0,990	4,2	4,8	5,3	5,6	5,8	6,0	6,2	6,3	6,4	6,6	6,7	6,9	7,3
20	0,950	3,0	3,6	4,0	4,2	4,4	4,6	4,8	4,9	5,0	5,1	5,2	5,4	5,7
	0,990	4,0	4,6	5,0	5,3	5,5	5,7	5,8	6,0	6,1	6,2	6,3	6,5	6,8
30	0,950	2,9	3,5	3,8	4,1	4,3	4,5	4,6	4,7	4,8	4,9	5,0	5,2	5,5
	0,990	3,9	4,5	4,8	5,0	5,2	5,4	5,5	5,7	5,8	5,8	5,9	6,1	6,4
∞	0,950	2,8	3,3	3,6	3,9	4,0	4,2	4,3	4,4	4,5	4,6	4,6	4,8	5,0
	0,990	3,6	4,1	4,4	4,6	4,8	4,9	5,0	5,1	5,2	5,2	5,3	5,4	5,6

Tabelle 5: Quantile q_p der Verteilung der studentisierten Variationsbreite.

Beispiel: Ertrag dreier unterschiedlicher Weizensorten auf insgesamt 11 Versuchsparzellen (Tabelle 1)

$s_R^2 = 54$ (dt/ha)^2

α = 0,05

0,95-Quantil der Verteilung der studentisierten Variationsbreite mit I = 3 und f = 8: $q_{0,95} = 4,0$

[3] Siehe Abschnitt 2 im Skript „Varianzanalyse“ unter aufgabomat.de.

Sorten	$\lvert \bar{x}_i - \bar{x}_j \rvert$ (dt/ha)	Grenzdifferenz (dt/ha)	signifikanter Unterschied	
			ja	nein
1 und 2	5	15		x
1 und 3	14	16		x
2 und 3	19	16	x	

Tabelle 6: Mittelwertvergleich (Tukey-Kramer-Test) für die Daten des Parzellenversuchs.

Wie schon beim Scheffé-Test (Abschnitt 2.2) wird nur ein Unterschied der Sorten 2 und 3 ausgewiesen. Daran, dass im vorliegenden Beispiel die nach Tukey-Kramer berechneten Grenzdifferenzen kleiner als diejenigen nach Scheffé sind (Tabelle 4), lässt sich allerdings erkennen, dass der Tukey-Kramer-Test weniger konservativ ist, also eher zur Anzeige von Unterschieden führt.

Weitere Übungsaufgaben finden sich beispielweise im Internet unter der Adresse aufgabomat.de in der Rubrik Statistik.

veröffentlicht im Internet unter aufgabomat.de