Mehrfachvergleichsmethode

EIN GRAFISCHES VERFAHREN FÜR MEHRFACHVERGLEICHE FÜR MEHRERE STANDARDABWEICHUNGEN

Senin J. Banga und Gregory D. Fox
18. Juni 2013

ZUSAMMENFASSUNG

Ein neues grafisches Verfahren für Mehrfachvergleiche von k Standardabweichungen wird vorgestellt. Als Test für die Homogenität von Varianzen weist das neue Verfahren ähnliche Eigenschaften für Fehler 1. Art und 2. Art wie Levenes Test (1960) in der Ausführung nach Brown and Forsythe (1974), W_{50}, auf. Die grafische Darstellung des Mehrfachvergleichstests jedoch bietet ein hilfreiches visuelles Werkzeug für das Screening von Stichproben mit unterschiedlichen Standardabweichungen.

Indexbegriffe: Homogenität der Varianzen, Levenes Test, Brown-Forsythe-Test, Layard-Test, Mehrfachvergleiche

1. Einführung

Die durch Brown and Forsythe (1974) ausgearbeitete Abwandlung des Tests von Levene (1960), die allgemein als Test W_{50} bezeichnet wird, ist vielleicht eines der gängigsten Verfahren zum Testen der Homogenität (Gleichheit) von Varianzen. Test W_{50} ist zum Teil so populär, weil er robust und asymptotisch verteilungsfrei ist. Im Vergleich mit anderen Tests der Homogenität von Varianzen ist Test W_{50} zudem einfach zu berechnen. (Einen Vergleich solcher Tests finden Sie bei Conover et al. (1981).) Darüber hinaus ist Test W_{50} leicht zugänglich, da er in vielen statistischen Softwarepaketen wie SAS, Minitab, R und JMP enthalten ist.

Für einige Verteilungen kann die Trennschärfe von Test W_{50} jedoch sehr niedrig sein, insbesondere bei kleinen Stichproben. Pan (1999) zeigt beispielsweise auf, dass Test W_{50} für einige Verteilungen (u. a. für die Normalverteilung) u. U. keine ausreichende Trennschärfe aufweist, um Differenzen zwischen zwei Standardabweichungen erkennen zu können, und zwar ungeachtet der Größe der Differenzen. Aus Pans Analyse geht jedoch nicht hervor, ob diese Einschränkung auch für Designs mit mehreren Stichproben gilt. Dass sich diese Einschränkung nicht auf Designs mit mehr als zwei Stichproben erstreckt, könnte man wegen

des einfachen Umstands erwarten, dass derartige Designs mehr Daten als Designs mit zwei Stichproben enthalten. Test W_{50} weist in Bezug auf große Stichproben nachweislich gute Eigenschaften auf (Miller, 1968; Brown und Forsythe, 1974; Conover et al., 1981).

Es ist gängige Praxis geworden, an einen signifikanten Test W_{50} ein simultanes paarweises Vergleichsverfahren auf der Grundlage einer Multiplizitätskorrektur nach Bonferroni anzuschließen. Wie jedoch von Pan (1999) konstatiert, schlägt ein solcher Ansatz aufgrund der niedrigen Trennschärfe von Test W_{50} bei Designs mit zwei Stichproben wahrscheinlich fehl oder liefert irreführende Ergebnisse. Das Problem wird durch die Anwendung der Bonferroni-Korrektur noch verschlimmert, da diese konservativ ist, insbesondere bei einer großen Anzahl von paarweisen Vergleichen. Im Gegensatz dazu sind viele effektive Mehrfachvergleichsverfahren verfügbar, um Mittelwerte im Abschluss an eine einfache ANOVA zu vergleichen. Beispiele finden Sie in Tukey (1953), Hochberg et al. (1982) und Stoline (1981). Eine analoge Post-hoc-Analyse für Vergleiche zwischen Stichprobenvarianzen wäre hilfreich.

Im vorliegenden White Paper schlagen wir eine grafische Methode zum Vergleichen der Varianzen (bzw. Standardabweichungen) mehrerer Stichproben vor. Die Analyse basiert auf „Unsicherheitsintervallen" für Varianzen, die den Unsicherheitsintervallen ähneln, die von Hochberg et al. (1982) für Mittelwerte beschrieben werden. Zunächst beruht ein paarweises Mehrfachvergleichsverfahren auf der abgewandelten Version von Bonett (2006) für den Layard-Test (1973) auf Gleichheit der Varianzen für Designs mit zwei Stichproben. Die Multiplizitätskorrektur in den paarweisen Vergleichen basiert auf einer Generalisierung für große Stichproben gemäß der Tukey-Kramer-Methode (Tukey, 1953; Kramer, 1956), vorgeschlagen von Nakayama (2009). Die Unsicherheitsintervalle, die bei uns als „Mehrfachsvergleichs-Intervalle" bzw. „MV-Intervalle" bezeichnet werden, werden vom paarweisen Vergleichsverfahren mit dem Verfahren der besten Approximation abgeleitet, das von Hochberg et al. (1982) beschrieben wurde. Der resultierende Mehrfachvergleichstest weist die Nullhypothese nur dann zurück, wenn für mindestens ein Paar von MV-Intervallen keine Überlappung vorliegt. Einander nicht überlappende MV-Intervalle geben die Stichproben an, die signifikant unterschiedliche Varianzen (oder Standardabweichungen) aufweisen.

Wir führten Simulationsstudien durch, um die Eigenschaften des Mehrfachvergleichstests in Bezug auf kleine Stichproben zu untersuchen. Zu Vergleichszwecken wird auch Test W_{50} in die Simulationsstudien eingebunden.

2. Grafisches Verfahren für Mehrfachvergleiche

$Y_{i1}, \ldots, Y_{in_i}, \ldots, Y_{k1}, \ldots, Y_{kn_k}$ seien k unabhängige Stichproben, wobei jede Stichprobe unabhängig und identisch mit dem Mittelwert $E(Y_{il}) = \mu_i$ und der Varianz $\mathrm{Var}(Y_{il}) = \sigma_i^2 > 0$ verteilt ist. Außerdem wird angenommen, dass die Stichproben aus Grundgesamtheiten mit einer gemeinsamen Kurtosis $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$ stammen.

$\bar{Y}_i$ und S_i seien der Mittelwert und die Standardabweichung von Stichprobe i. m_i sei der getrimmte Mittelwert von Stichprobe i mit dem Trim-Anteil $1/[2\sqrt{n_i - 4}]$, und $\hat{\gamma}_{ij}$ sei ein zusammengefasster Kurtosis-Schätzwert der Stichproben $(i; j)$, angegeben als

$$\hat{\gamma}_{ij} = (n_i + n_j)\frac{\sum_{l=1}^{n_i}(Y_{il} - m_i)^4 + \sum_{l=1}^{n_j}(Y_{jl} - m_j)^4}{\left[\sum_{l=1}^{n_i}(Y_{il} - \bar{Y}_i)^2 + \sum_{l=1}^{n_j}(Y_{jl} - \bar{Y}_j)^2\right]^2}$$

$$= (n_i + n_j)\frac{\sum_{l=1}^{n_i}(Y_{il} - m_i)^4 + \sum_{l=1}^{n_j}(Y_{jl} - m_j)^4}{\left[(n_i - 1)S_i^2 + (n_j - 1)S_j^2\right]^2}$$

Beachten Sie, dass $\hat{\gamma}_{ij}$ asymptotisch äquivalent zum zusammengefassten Kurtosis-Schätzwert nach Layard (1973) ist, wobei der Stichprobenmittelwert $\bar{Y}_i$ durch den getrimmten Mittelwert m_i ersetzt wurde. Damit ist $\hat{\gamma}_{ij}$ ein konsistenter Schätzwert der unbekannten gemeinsamen Kurtosis γ, so lange die Varianzen der Grundgesamtheiten gleich sind. Bonett (2006) schlägt diesen Schätzwert anstelle des zusammengefassten Kurtosis-Schätzwerts nach Layard vor, um die Leistung des Layard-Tests in Bezug auf kleine Stichproben bei Fragestellungen mit zwei Stichproben zu verbessern. Wir bezeichnen die abgewandelte Version des Layard-Tests nach Bonett (2006) in diesem Artikel einfach als Bonett-Test.

Angenommen, es sind mehr als zwei unabhängige Gruppen oder Stichproben vorhanden, die verglichen werden sollen ($k > 2$). Das von uns vorgeschlagene grafische Mehrfachvergleichsverfahren ist von den multiplen paarweisen Vergleichen abgeleitet, die auf dem Bonett-Test basieren. Ein alternativer Ansatz besteht darin, die paarweisen Vergleichen von Test W_{50} herzuleiten. Bei Designs mit zwei Stichproben ist jedoch die Trennschärfe von Test W_{50} für einige Verteilungen problematisch, u. a. für die Normalverteilung (Pan, 1999). Außerdem zeigten Banga und Fox (2013) auf, dass Konfidenzintervalle für das Verhältnis der Varianzen, die auf dem Bonett-Test basieren, generell denjenigen überlegen sind, die auf Test W_{50} beruhen.

Bei einem beliebigen Paar $(i; j)$ von Stichproben weist ein beidseitiger Bonett-Test mit dem Signifikanzniveau α' die Nullhypothese der Gleichheit von Varianzen nur dann zurück, wenn

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > z_{\alpha'/2}\sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

Hierbei ist $z_{\alpha'/2}$ der $\alpha'/2 \times 100$. obere Perzentilpunkt der Standardnormalverteilung:

$$k_i = \frac{n_i - 3}{n_i}, k_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha/2}}, c_j = \frac{n_j}{n_j - z_{\alpha/2}}$$

Da mehrere paarweise Vergleiche vorhanden sind, also genau $k(k-1)/2$ Vergleiche, ist eine Multiplizitätskorrektur erforderlich. Wenn beispielsweise ein Soll-Gesamtsignifikanzniveau bzw. simultanes Signifikanzniveau α gegeben ist, besteht ein häufig angewendeter Ansatz, die Bonferroni-Korrektur, darin, für das Signifikanzniveau der einzelnen $k(k-1)/2$ paarweisen Vergleiche $\alpha' = 2\alpha/(k(k-1))$ auszuwählen. Von der Bonferroni-Korrektur ist jedoch bekannt, dass mit steigender Anzahl der zu vergleichenden Stichproben zunehmend konservative paarweise Vergleichsverfahren liefert. Ein alternativer und besser geeigneter Ansatz wird von Nakayama (2009) vorgeschlagen. Diesem liegt eine Approximation für große Stichproben der Tukey-Kramer-Methode (Tukey, 1953; Kramer, 1956) zugrunde. Insbesondere gilt, dass der Gesamttest der mehrfachen paarweisen Vergleiche nur dann signifikant ist, wenn auf ein Paar $(i; j)$ von Stichproben Folgendes zutrifft:

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > \frac{q_{k,\alpha}}{\sqrt{2}} \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

Hierbei ist $q_{\alpha,k}$ der obere α. Punkt des Bereichs von k unabhängigen und identisch verteilten Zufallsvariablen einer Standardnormalverteilung. Das heißt, $q_{\alpha,k}$ erfüllt

$$\Pr\left(\max_{1 \le i < j \le k} \left|Z_i - Z_j\right| \le q_{\alpha,k}\right) = 1 - \alpha$$

Hierbei sind $Z_1; \ldots; Z_k$ unabhängige und identisch verteilte Zufallsvariablen einer Standardnormalverteilung. Barnard (1978) gibt einen einfachen numerischen Algorithmus an, der auf einer Gaußschen Quadratur von 16 Punkten zum Berechnen der Verteilungsfunktion des Normalverteilungsbereichs basiert.

Wie von Hochberg et al. (1982) behauptet, würde ein grafisches Mehrfachvergleichsverfahren, das sich an das oben beschriebene paarweise Mehrfachvergleichsverfahren annähert, die Nullhypothese nur dann zurückweisen, wenn

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > q_{\alpha,k}(V_i + V_j)/\sqrt{2}$$

Hierbei werden die V_i ausgewählt, um Folgendes zu minimieren:

$$\sum_{i \ne j} \sum \left(V_i + V_j - b_{ij}\right)^2$$

Dabei gilt Folgendes:

$$b_{ij} = \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

Die Lösung dieses Problems, wie bei Hochberg et al. (1982) veranschaulicht, besteht in der Auswahl von

$$V_i = \frac{(k-1)\sum_{j \ne i} b_{ij} - \sum\sum_{1 \le j < l \le k} b_{jl}}{(k-1)(k-2)}$$

Daraus folgt, dass ein Test auf Homogenität von Varianzen auf der Grundlage dieses Approximationsverfahrens die Nullhypothese nur dann zurückweist, wenn für mindestens ein Paar der unten angegebenen Intervalle keine Überlappung vorliegt:

$$\left[S_i\sqrt{c_i\exp(-q_{\alpha,k}V_i/\sqrt{2}\,)}\,,S_i\sqrt{c_i\exp(q_{\alpha,k}V_i/\sqrt{2})}\right],i=1;\ldots;k$$

Das grafische Mehrfachvergleichsverfahren besteht aus der Darstellung dieser Intervalle in einem Diagramm, in dem die Stichproben mit einander nicht überlappenden Intervallen ermittelt werden können. Darüber hinaus kann der p-Wert des Gesamttests auf Homogenität der Varianzen (oder Standardabweichungen) bestimmt werden. Im nächsten Abschnitt werden ausführliche Algorithmen zum Berechnen des p-Werts vorgestellt. Zunächst werden jedoch einige einfache Tatsachen zum Mehrfachvergleichsverfahren erläutert.

ANMERKUNG

1. Der zusammengefasste Kurtosis-Schätzwert $\hat{\gamma}_{ij}$, der auf dem Paar $(i;j)$ von Stichproben basiert, könnte durch den zusammengefassten Kurtosis-Gesamtschätzwert ersetzt werden, der auf allen k Stichproben basiert. Bei diesem Ansatz werden zwar die Berechnungen etwas vereinfacht, hier nicht vorgestellte Simulationsergebnisse verweisen jedoch darauf, dass die Verwendung von $\hat{\gamma}_{ij}$ zu besseren Ergebnissen führt.
2. Das Intervall, das Stichprobe i entspricht, ist kein Konfidenzintervall für die Standardabweichung der übergeordneten Grundgesamtheit der Stichproben. Hochberg et al. (1982) bezeichnen ein derartiges Intervall als „Unsicherheitsintervall". Wir hingegen bezeichnen es als „Mehrfachvergleichsintervall" oder „MV-Intervall". MV-Intervalle sind nur nützlich bei Vergleichen der Standardabweichungen bzw. Varianzen für Designs mit mehreren Stichproben.
3. Anhand der im vorliegenden Artikel beschriebenen MV-Intervalle können nur mehr als zwei Standardabweichungen miteinander verglichen werden. Wenn nur zwei Stichproben vorhanden sind, können Vergleichsintervalle konstruiert werden, sie vermitteln jedoch dieselben Informationen, die auch von den Testergebnissen geliefert werden. Viel aufschlussreicher ist es, ein Konfidenzintervall für das Verhältnis der Standardabweichungen aufzustellen, z. B. wie das von Banga und Fox (2013) beschriebene. Dieses wird über den Minitab-Befehl „Test auf Varianzen, 2 Stichproben" bereitgestellt.

3. p-Wert der grafischen Methode für Mehrfachvergleiche

Bevor der Algorithmus zum Berechnen des p-Werts der grafischen Mehrfachvergleichsmethode beschrieben wird, leiten wir zunächst den p-Wert für die Abwandlung des Layard-Tests nach Bonett (2006) in Designs mit zwei Stichproben ab. Anschließend wird veranschaulicht, wie die Ergebnisse für das Design mit zwei Stichproben auf das Mehrfachvergleichsverfahren übertragen werden.

3.1 p-Wert in Designs mit zwei Stichproben

Wie bereits erwähnt, weist die Abwandlung des Layard-Tests nach Bonett (2006) in Designs mit zwei Stichproben die Nullhypothese der Homogenität von Varianzen nur dann zurück, wenn

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

oder äquivalent

$$\left|\ln(c_{\alpha/2}\, S_1^2 / S_2^2)\right| > z_{\alpha/2} se$$

Dabei gilt Folgendes:

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Bonett führte die Konstante $c_{\alpha/2}$ als Korrektur für kleine Stichproben ein, um den Effekt ungleicher Fehlerwahrscheinlichkeiten in den Randbereichen von unbalancierten Designs mit zwei Stichproben zu mindern. Der Effekt der Konstante ist in unbalancierten Designs mit großen Stichproben jedoch vernachlässigbar, und die Konstante hat keinen Effekt in balancierten Designs.

Daraus folgt, dass bei einem balancierten Design der p-Wert des beidseitigen Tests auf Homogenität der Varianzen einfach berechnet werden kann als

$$P = 2\Pr(Z > |Z_0|)$$

Dabei gilt Folgendes:

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

Wenn das Design unbalanciert ist, dann ist $P = 2\min(\alpha_L; \alpha_U)$,, wobei α_L die kleinste Lösung für α in der folgenden Gleichung ist:

$\exp[\ln(c_\alpha S_1^2/S_2^2) - z_\alpha se] = 1$ (1)

und α_U die kleinste Lösung für α in der folgenden Gleichung ist:

$\exp[\ln(c_\alpha S_1^2/S_2^2) + z_\alpha se] = 1$ (2)

Algorithmen zum Bestimmen von α_L und α_U sind unten angegeben. Auf die mathematischen Details der Algorithmen wird erst im Anhang eingegangen.

Sei

$$L(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1; n_2)$$

Sei außerdem

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Die Lösungen α_L und α_U werden in den folgenden Schritten berechnet:

Fall 1: $n_1 < n_2$

- z_m wird wie im Ergebnis oben berechnet, und $L(z_m, n_1, n_2, S_1, S_2)$ wird ausgewertet.
- Wenn $L(z_m) \leq 0$, wird die Wurzel z_L von $L(z, n_1, n_2, S_1, S_2)$ im Intervall $(-\infty; z_m)$ bestimmt, und $\alpha_L = \Pr(Z > z_L)$ wird berechnet.
- Wenn $L(z_m) > 0$, dann hat die Funktion $L(z, n_1, n_2, S_1, S_2)$ keine Wurzel. Es wird $\alpha_L = 0{,}0$ festgelegt.

Fall 2: $n_1 > n_2$

- $L(0, n_1, n_2, S_1, S_2) = \ln S_1^2/S_2^2$ wird berechnet.
- Wenn $L(0, n_1, n_2, S_1, S_2) \geq 0$, wird die Wurzel z_o von $L(z, n_1, n_2, S_1, S_2)$ im Intervall $(0;\ n_2)$ bestimmt. Andernfalls wird die Wurzel z_L im Intervall $(-\infty;\ 0)$ bestimmt.
- $\alpha_L = \Pr(Z > z_L)$ wird berechnet.

Zum Berechnen von α_U werden einfach die obigen Schritte mit der Funktion $L(z, n_2, n_1, S_2, S_1)$ anstelle der Funktion $L(z, n_1, n_2, S_1, S_2)$ angewendet.

3.2 p-Wert der grafischen Mehrfachvergleiche

Angenommen, es sind k ($k > 2$) Stichproben im Design vorhanden. Dann sei P_{ij} der p-Wert des Tests für ein beliebiges Paar $(i; j)$ von Stichproben. Rufen Sie sich ins Gedächtnis zurück, dass die Nullhypothese der Homogenität von Varianzen beim Mehrfachvergleichstest nur dann zurückgewiesen wird, wenn für mindestens ein Paar der k Vergleichsintervalle keine Überlappung vorliegt. Daraus folgt, dass der p-Gesamtwert für das Mehrfachvergleichsverfahren folgendermaßen lautet:

$$P = \min\{ P_{ij};\ 1 \leq i < j \leq k\}$$

Zum Berechnen von P_{ij} wird der Algorithmus für Designs mit zwei Stichproben ausgeführt mit:

$$se = V_i + V_j$$

Hierbei entspricht V_i der vorausgegangenen Definition.

Wenn $n_i \neq n_j$, dann

$$P_{ij} = \min(\alpha_L; \alpha_U)$$

Hierbei ist $\alpha_L = \Pr(Q > z_L\sqrt{2})$; $\alpha_U = \Pr(Q > z_U\sqrt{2})$; z_L ist die kleinste Wurzel der Funktion $L\left(z, n_i, n_j, S_i, S_j\right)$, z_U ist die kleinste Wurzel der Funktion $L\left(z, n_j, n_i, S_j, S_i\right)$, und Q ist eine Zufallsvariable entsprechend der vorausgegangenen Definition. Die Größen z_L und z_U werden durch Anwenden des vorher erläuterten Algorithmus für Designs mit zwei Stichproben auf das Paar $(i; j)$ von Stichproben ermittelt.

Wenn $n_i = n_j$, dann $P_{ij} = \Pr(Q > |z_o|\sqrt{2})$, wobei

$$z_o = \frac{\ln S_i^2 - \ln S_j^2}{V_i + V_j}$$

4. Simulationsstudie und Ergebnisse

Zwei umfassende Simulationsstudien werden durchgeführt, um die Leistung des Mehrfachvergleichstests in Bezug auf kleine Stichproben als Gesamttest auf Homogenität der Varianzen zu untersuchen. Alle Simulationen wurden mit Version 8 des Mathematica-Softwarepakets durchgeführt.

Studie 1

Mit der ersten Studie werden die Eigenschaften im Hinblick auf den Fehler 1. Art des Mehrfachvergleichstests und des Tests W_{50} ausgewertet und verglichen. Dabei wurde die Leistung der beiden Tests mit Stichproben aus diversen Verteilungen in drei unterschiedlichen Designs verglichen: einem Design mit drei Stichproben, einem Design mit vier Stichproben und einem Design mit mit sechs Stichproben. In jedem Design variieren die Stichprobenumfänge zwischen 10 und 50 in Schritten von 10. Stichproben werden aus den folgenden übergeordneten Verteilungen gezogen:

- Normalverteilung
- symmetrische Verteilungen mit schwächer besetzten Randbereichen, dargestellt durch die Gleichverteilung und eine Beta-Verteilung mit den Parametern (3;3)
- symmetrische Verteilungen mit stärker besetzten Randbereichen, dargestellt durch eine t-Verteilung mit 5 Freiheitsgraden ($t(5)$) und die Laplace-Verteilung
- schiefe Verteilungen mit stärker besetzten Randbereichen, dargestellt durch die Exponentialverteilung, eine Chi-Quadrat-Verteilung mit 1 Freiheitsgrad ($\chi^2(1)$) und eine Chi-Quadrat-Verteilung mit 5 Freiheitsgraden ($\chi^2(5)$)
- eine kontaminierte Normalverteilung (CN(0,9;3)), für die 90 % der Beobachtungen aus der Standardnormalverteilung und die übrigen 10 % aus einer Normalverteilung mit dem Mittelwert 0 und der Standardabweichung 3 gezogen wurden.

Jede Simulation besteht aus 10.000 Stichprobenreplikationen. Das nominale α-Sollniveau ist 0,05. Der Simulationsfehler liegt bei ca. 0,002. Die simulierten Signifikanzniveaus für die einzelnen Tests sind in Tabelle 1 aufgeführt.

Tabelle 1 Vergleich der simulierten Signifikanzniveaus ($\alpha = 0{,}05$)

Beschreibung	Verteilung [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MV	W_{50}	MV	W_{50}	MV	W_{50}
Normal	Normal [3,0]	10	0,038	0,033	0,038	0,031	0,036	0,029
		20	0,039	0,038	0,040	0,038	0,041	0,033
		30	0,043	0,041	0,044	0,038	0,046	0,039
		40	0,046	0,043	0,046	0,041	0,048	0,041
		50	0,046	0,046	0,046	0,044	0,052	0,047

Beschreibung	Verteilung [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MV	W_{50}	MV	W_{50}	MV	W_{50}
Symmetrisch mit schwächer besetzten Randbereichen	Gleichverteilung [1,8]	10	0,029	0,029	0,025	0,024	0,023	0,020
		20	0,028	0,026	0,030	0,026	0,028	0,023
		30	0,037	0,035	0,034	0,032	0,034	0,030
		40	0,038	0,037	0,037	0,037	0,035	0,033
		50	0,041	0,041	0,036	0,036	0,036	0,036
	Beta (3;3) [2,5]	10	0,031	0,032	0,031	0,029	0,031	0,025
		20	0,035	0,031	0,036	0,027	0,037	0,026
		30	0,041	0,035	0,037	0,034	0,037	0,032
		40	0,040	0,036	0,039	0,035	0,040	0,033
		50	0,044	0,039	0,044	0,037	0,044	0,035
Symmetrisch mit stärker besetzten Randbereichen	Laplace [6,0]	10	0,056	0,038	0,063	0,041	0,071	0,039
		20	0,054	0,044	0,058	0,043	0,059	0,041
		30	0,051	0,042	0,053	0,043	0,052	0,044
		40	0,048	0,045	0,048	0,045	0,048	0,046
		50	0,045	0,045	0,051	0,046	0,049	0,047
	$t(5)$ [9,0]	10	0,042	0,032	0,044	0,031	0,042	0,031
		20	0,043	0,039	0,045	0,038	0,045	0,040
		30	0,039	0,040	0,040	0,040	0,041	0,040
		40	0,041	0,042	0,040	0,041	0,039	0,038
		50	0,040	0,050	0,039	0,046	0,038	0,046
Schief mit stärker besetzten Randbereichen	$\chi^2(5)$ [5,4]	10	0,040	0,039	0,046	0,040	0,048	0,039
		20	0,040	0,043	0,040	0,040	0,042	0,039
		30	0,039	0,047	0,042	0,044	0,043	0,042
		40	0,040	0,046	0,041	0,044	0,039	0,042
		50	0,037	0,047	0,038	0,047	0,040	0,048
	Exponential [9,0]	10	0,063	0,051	0,073	0,049	0,076	0,048
		20	0,051	0,049	0,053	0,048	0,057	0,046
		30	0,042	0,048	0,046	0,051	0,049	0,049
		40	0,034	0,050	0,038	0,046	0,037	0,049
		50	0,033	0,045	0,037	0,047	0,038	0,046

Beschreibung	Verteilung [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MV	W_{50}	MV	W_{50}	MV	W_{50}
	$\chi^2(1)$ [15,0]	10	0,084	0,048	0,098	0,050	0,118	0,050
		20	0,053	0,046	0,060	0,047	0,068	0,046
		30	0,041	0,041	0,045	0,045	0,050	0,047
		40	0,044	0,049	0,046	0,047	0,045	0,047
		50	0,038	0,050	0,037	0,049	0,040	0,049
Kontaminierte Normalverteilung	CN(0,9;3) [8,3]	10	0,020	0,016	0,018	0,012	0,016	0,010
		20	0,014	0,015	0,012	0,013	0,008	0,007
		30	0,012	0,014	0,010	0,011	0,007	0,008
		40	0,009	0,017	0,009	0,014	0,006	0,008
		50	0,009	0,016	0,007	0,012	0,006	0,009

Die Ergebnisse zeigen, dass beide Tests für die meisten Verteilungen eine gute Leistung bieten. Die Mehrzahl der simulierten Signifikanzniveaus liegen nahe dem Sollwert von 0,05. Die simulierten Signifikanzniveaus für beide Tests sind jedoch tendenziell konservativ (niedriger als 0,05), wenn kleine Stichproben aus Normalverteilungen und symmetrischen Verteilungen mit schwächer besetzten Randbereichen gezogen wurden. Für diese Verteilungen liegen die simulierten Signifikanzniveaus für den Mehrfachvergleichstests näher am Soll-Signifikanzniveau als diejenigen für Test W_{50}.

Wenn kleine Stichproben aus Verteilungen mit stärker besetzten Randbereichen gezogen werden, ist Test W_{50} tendenziell konservativ, während der Mehrfachvergleichstest tendenziell liberal ist. Der Mehrfachvergleichstest ist noch stärker liberal ausgeprägt, wenn kleine Stichproben aus extrem schiefen Verteilungen gezogen werden. Wenn beispielsweise Stichproben mit dem Umfang 10 aus einer Chi-Quadrat-Verteilung mit 1 Freiheitsgrad gezogen werden, betragen die simulierten Signifikanzniveaus für den Mehrfachvergleichstest für das Design mit 2, 4 bzw. 6 Stichproben 0,084; 0,098 und 0,118.

Beide Tests werden durch Ausreißer beeinflusst. Die Signifikanzniveaus für die kontaminierte Normalverteilung sind extrem konservativ, selbst wenn die Stichproben Umfänge von bis zu 50 erreichen.

Studie 2

In der zweiten Studie werden die Eigenschaften im Hinblick auf den Fehler 2. Art (Trennschärfe) der beiden Verfahren in einem Design mit 4 Stichproben untersucht und miteinander verglichen. Für diese Studie werden die gleichen Stichproben wie für die Stichproben des Umfangs 20 und die Bedingung $k = 4$ in Studie 1 verwendet. Die Beobachtungen sind mit einem Faktor von 1, 2, 3 oder 4 skaliert. Unter der Bedingung 1:1:4:4 beispielsweise sind die Beobachtungen für die Stichproben 1 und 2 identisch mit denen aus Studie 1. Die Beobachtungen in den Stichproben 3 und 4 sind mit dem Faktor 4 skaliert.

Die Bedingung 1:1:1:1 wird zu Vergleichszwecken aufgeführt. Beachten Sie, dass die Ergebnisse für diese Bedingung mit denen für Stichproben mit einem Umfang von 20 und $k = 4$ aus Studie 1 übereinstimmen. Der Stichprobenumfang 20 wurde gewählt, da die Ergebnisse von Studie 1 nahelegen, dass für beide Tests, für die meisten Verteilungen und für Stichproben des Umfangs 20 Signifikanzniveaus erhalten werden, die nahe dem Sollniveau liegen.

Die simulierten Trennschärfen in diesen Experimenten werden als der Anteil der Stichprobenreplikationen berechnet, bei denen die Nullhypothese der Homogenität von Varianzen zurückgewiesen wird.

Die Ergebnisse werden in Tabelle 2 aufgeführt.

Tabelle 2 Vergleich der simulierten Trennschärfen ($\alpha = 0{,}05$)

Beschreibung	Verteilung	Verhältnis der Standardabweichungen							
		1:1:1:1		1:1:2:2		1:2:3:4		1:1:4:4	
		MV	W_{50}	MV	W_{50}	MV	W_{50}	MV	W_{50}
	Normal	0,040	0,038	0,846	0,853	0,998	0,994	1,000	1,000
Symmetrisch mit schwächer besetzten Randbereichen	Gleichverteilung	0,030	0,026	0,985	0,962	1,000	0,999	1,000	1,000
	Beta (3;3)	0,036	0,027	0,938	0,916	1,000	0,999	1,000	1,000
Symmetrisch mit stärker besetzten Randbereichen	Laplace	0,058	0,043	0,597	0,629	0,931	0,921	0,996	0,998
	$t(5)$	0,045	0,038	0,657	0,703	0,952	0,949	0,997	0,998
Schief mit stärker besetzten Randbereichen	$\chi^2(5)$	0,040	0,040	0,625	0,704	0,949	0,949	0,996	0,999
	Exponential	0,053	0,048	0,431	0,507	0,804	0,779	0,963	0,978
	$\chi^2(1)$	0,060	0,047	0,298	0,291	0,602	0,504	0,838	0,824
Kontaminiert	CN(0,9;3)	0,012	0,013	0,499	0,612	0,889	0,917	0,989	0,998

Die Ergebnisse legen nahe, dass die Eigenschaften in Bezug auf den Fehler 2. Art (Trennschärfe) für den Mehrfachvergleichstest und den Test W_{50} einander ähneln. Im Allgemeinen weisen die simulierten Trennschärfen, die mit beiden Tests erzielt werden, die gleiche Größenordnung auf. Nur in einem einzigen Fall unterscheiden sich die Trennschärfen der beiden Tests um mehr als 0,1.

Die simulierten Trennschärfen für den Mehrfachvergleichstest sind etwas besser als die für Test W_{50}, wenn Stichproben aus symmetrischen Verteilungen mit schwächer bis gemäßigt besetzten Randbereichen gezogen werden. Andererseits scheint Test W_{50} eine etwas bessere Trennschärfe als der Mehrfachvergleichstest aufzuweisen, wenn Stichproben aus Verteilungen mit stärker besetzten Randbereichen gezogen werden.

5. Beispiel

In diesem Abschnitt werden das grafische Mehrfachvergleichsverfahren und Test W_{50} auf einen Datensatz angewendet, der Ott et al. (2010), Seite 397, entnommen wurde. Die Daten werden wie folgt beschrieben:

> *Ein Gussteilhersteller verfügt über mehrere Öfen, in denen das Rohmaterial geschmolzen wird, ehe es in eine Wachsgussform gegossen wird. Es ist unerlässlich, dass die Metalle auf eine genaue Temperatur mit geringstmöglicher Streuung erhitzt werden. Drei Öfen werden nach dem Zufallsprinzip ausgewählt, und ihre Temperatur (°C) in 10 aufeinander folgenden Erhitzungsvorgängen wird äußerst genau aufgezeichnet. Die folgenden Daten wurden erfasst:*

Ofen 1	1670,87	1670,88	1671,51	1672,01	1669,63	1670,95	1668,70	1671,86	1669,12	1672,52
Ofen 2	1669,16	1669,60	1669,76	1669,18	1671,92	1669,69	1669,45	1669,35	1671,89	1673,45
Ofen 3	1673,08	1672,75	1675,14	1674,94	1671,33	1660,38	1679,94	1660,51	1668,78	1664,32

Abbildung 1 zeigt Boxplots der Temperaturen für die einzelnen Öfen. Die Boxplots weisen darauf hin, dass in den aufgezeichneten Temperaturen keine Ausreißer vorhanden sind und sich die Temperaturstreuung für Ofen 3 von der von Ofen 1 bzw. Ofen 2 unterscheidet.

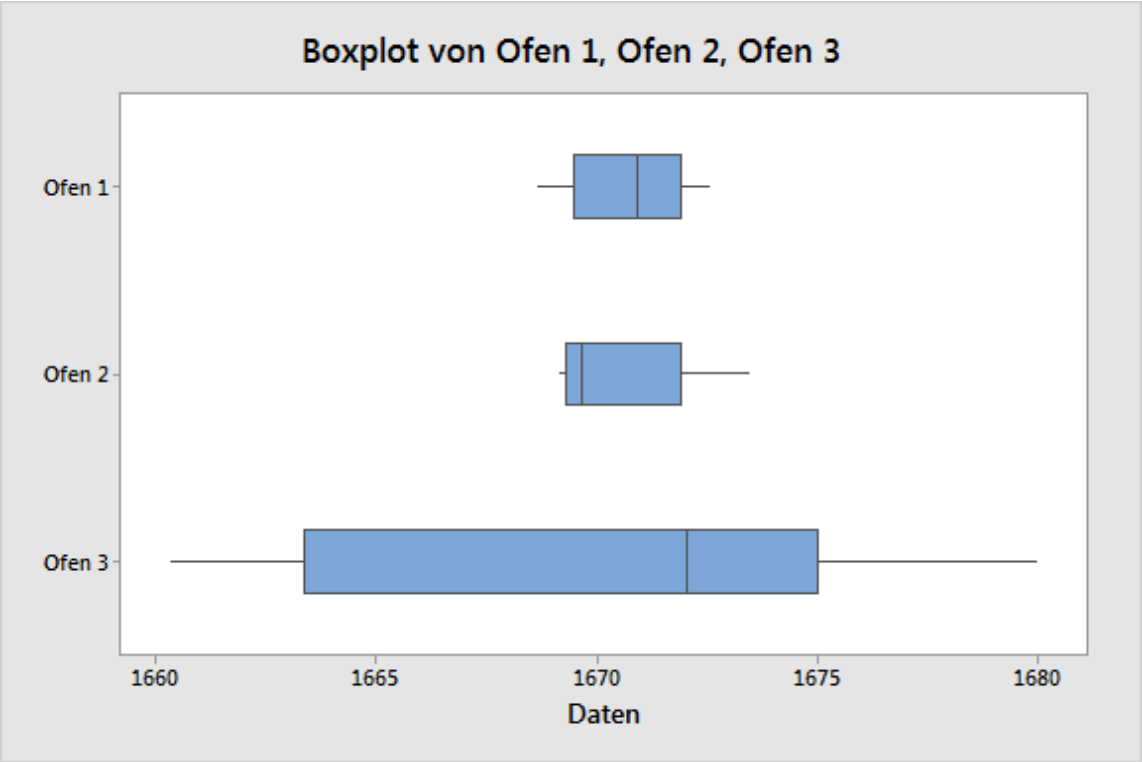

Abbildung 1 Boxplots der Ofentemperatur (°C)

Abbildung 2 zeigt die Mehrfachvergleichsintervalle für die gleichen Daten sowie die Ergebnisse des Mehrfachvergleich-Gesamttests und von Test W_{50}, der in der Legende als Levenes Test bezeichnet wird. Die signifikanten p-Werte für beide Tests geben an, dass sich die Streuungen der Temperatur für die drei Öfen voneinander unterscheiden. Die einander nicht überlappenden MV-Intervalle bestätigen, dass sich die Streuung für Ofen 3 von der für Ofen 2 bzw. Ofen 1 unterscheidet. Die MV-Intervalle für die Öfen 1, 2 und 3 sind $(0{,}896; 2{,}378)$; $(1{,}072;\ 2{,}760)$ und $(4{,}366; 12{,}787)$.

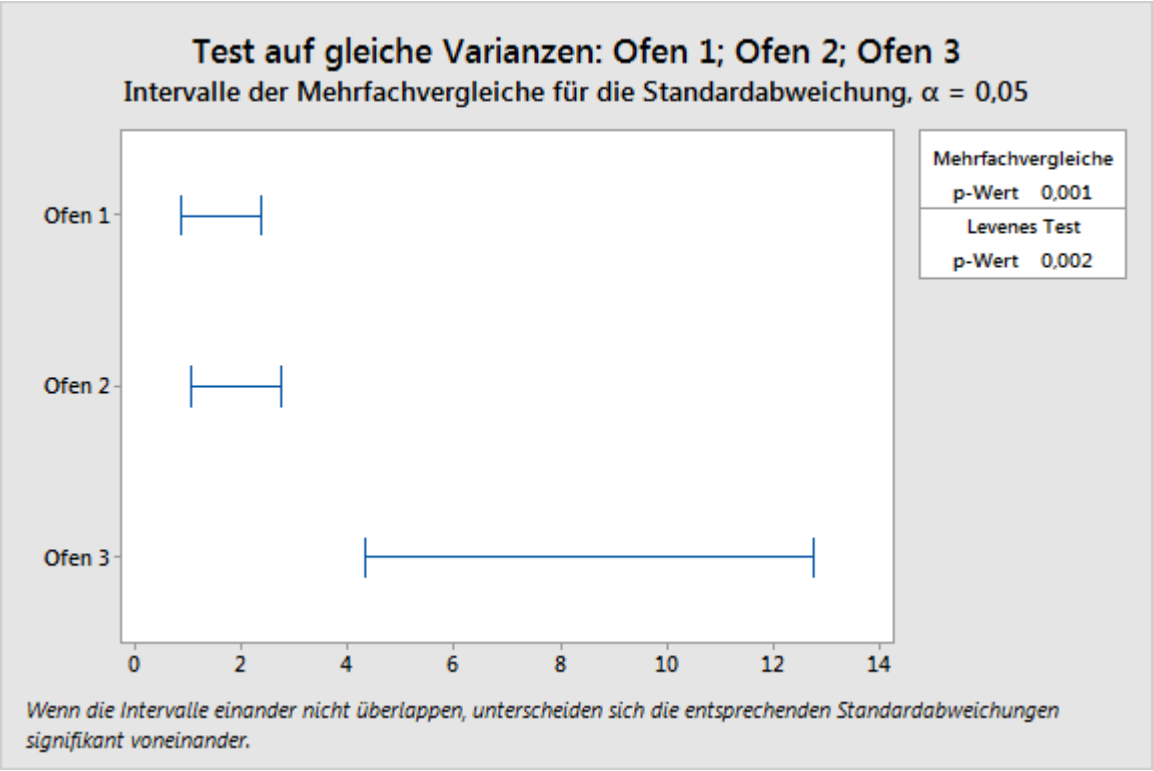

Abbildung 2 MV-Intervalle und p-Werte für den Mehrfachvergleichstest und Test W_{50} (Levenes Test)

6. Schlussfolgerung

Ingesamt zeigen die Simulationsergebnisse, dass die Leistung des Mehrfachvergleichstests für Designs mit mehreren kleinen Stichproben der von Test W_{50} ähnelt. Der Mehrfachvergleichstest ist etwas besser für symmetrische oder nahezu symmetrische Verteilungen mit schwächer bis gemäßigt besetzten Randbereichen geeignet, während sich Test W_{50} eher empfiehlt, wenn Daten aus stark schiefen Verteilungen und Verteilungen mit stärker besetzten Randbereichen gezogen werden. Ein eindeutiger Vorteil des Mehrfachvergleichstests besteht darin, dass er ein effektives visuelles Werkzeug für das Screening von Stichproben mit unterschiedlichen Standardabweichungen oder Varianzen darstellt, wenn der Gesamttest auf Homogenität der Standardabweichungen signifikant ist. Das grafische Mehrfachvergleichsverfahren ist in Minitab Release 17 verfügbar.

7. Anhang

Die Abwandlung des Layard-Tests nach Bonett (2006) in Designs mit zwei Stichproben weist die Nullhypothese der Homogenität von Varianzen nur dann zurück, wenn

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

oder äquivalent

$$\left|\ln(c_{\alpha/2}\, S_1^2 / S_2^2)\right| > z_{\alpha/2} se$$

Dabei gilt Folgendes:

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Damit ist bei einem balancierten Design $c_{\alpha/2} = 1$, und somit ist der p-Wert des Tests einfach

$$P = 2\Pr(Z > |Z_0|)$$

Dabei gilt Folgendes:

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

Wenn das Design unbalanciert ist, dann ist $P = 2\min(\alpha_L; \alpha_U)$: hierbei gilt:

$\alpha_L$ist die kleinste Lösung für α in der Gleichung

$\exp[\ln(c_\alpha S_1^2/S_2^2) - z_\alpha se] = 1$ (1)

und α_U ist die kleinste Lösung für α der Gleichung

$\exp[\ln(c_\alpha S_1^2/S_2^2) + z_\alpha se] = 1$ (2)

Beim Lösen dieser Gleichungen für α werden zunächst die Gleichungen für $z \equiv z_\alpha$ gelöst, und anschließend wird $\alpha = \Pr(Z > z)$ bestimmt, wobei die Zufallsvariable Z die Standardnormalverteilung aufweist. Bevor beschrieben wird, wie diese Gleichungen gelöst werden, ist anzumerken, dass Gleichung (1) als Gleichung $L(z) = 0$ neu formuliert werden kann, wobei

$$L(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1; n_2)$$

Ebenso entspricht Gleichung (2) der Gleichung $U(z) = 0$, wobei

$$U(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} + z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1; n_2)$$

Wir stellen fest, dass $L(z, n_2, n_1, S_2, S_1) = -U(z, n_1, n_2, S_1, S_2)$. Demzufolge müssen nur die Wurzeln einer der beiden Funktionen bestimmt werden.

Der Algorithmus zum Lösen von Gleichung (1) bzw. (2) wird von folgendem Ergebnis abgeleitet:

Ergebnis

Seien n_1, n_2, S_1 und S_2 vorgegeben und festgelegt. Bei nicht balancierten Designs hat die Funktion $L(z, n_1, n_2, S_1, S_2)$ höchstens zwei Wurzeln.

4. *Wenn $n_1 < n_2$, dann ist $L(z, n_1, n_2, S_1, S_2)$ konvex: $L(-\infty, n_1, n_2, S_1, S_2) = L(n_1, n_1, n_2, S_1, S_2) = +\infty$ wird erfüllt, und das Minimum wird erreicht bei*

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

   Wenn also $L(z_m, n_1, n_2, S_1, S_2) \leq 0$, dann sind zwei Wurzeln vorhanden: eine im Intervall $(-\infty, ; z_m)$ und die andere im Intervall $(z_m, ; n_1)$. Wenn jedoch $L(z_m, n_1, n_2, S_1, S_2) > 0$, dann hat die Funktion $L(z, n_1, n_2, S_1, S_2)$ keine Wurzel.

5. Wenn $n_1 > n_2$, dann verringert sich $L(z, n_1, n_2, S_1, S_2)$ monoton von $+\infty$ auf $-\infty$ und hat daher eine eindeutige Wurzel. Wenn $L(0, n_1, n_2, S_1, S_2) = ln\, S_1^2/S_2^2 \geq 0$, dann liegt die Wurzel im Intervall $(0;\ n_2)$; andernfalls liegt sie im Intervall $(-\infty;\ 0)$.

Beweis

Im Folgenden sei $L(z) \equiv L(z, n_1, n_2, S_1, S_2)$.

Zunächst soll Folgendes bewiesen werden: Wenn $n_1 < n_2$, dann ist $L(z)$ konvex und erreicht sein Minimum bei

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Wie bereits zuvor definiert:

$$L(z) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1;\ n_2)$$

In diesem Fall gilt $\lim_{z \to -\infty} L(z) = +\infty$ und

$$\lim_{z \to \min(n_1, n_2)} L(z) = \begin{cases} +\infty\ (n_1 < n_2) \\ -\infty\ (n_2 < n_1) \end{cases}$$

Beachten Sie zudem, dass das Derivat von $L(z)$ Folgendes erfüllt:

$$-\frac{(n_1 - z)(n_2 - z)}{se} L'(z) = z^2 - (n_1 + n_2)z + n_1 n_2 + \frac{n_1 - n_2}{se}$$

Sei

$$Q(z) = -\frac{(n_1 - z)(n_2 - z)}{se} L'(z)$$

Wenn $n_1 < n_2$, dann hat das quadrierte $Q(z)$ zwei Wurzeln:

$$z_1 = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

und

$$z_2 = \frac{n_1 + n_2 + \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Da $Q(n_1) = \frac{n_1 - n_2}{se} < 0$, gilt $z_1 < n_1 = \min(n_1; n_2) < z_2$, so dass $Q(z) > 0$ für z in $(-\infty; z_1)$ und so dass $Q(z) < 0$ für z in $(z_1;\ n_1)$. Daraus folgt, dass $L'(z) < 0$ für z in $(-\infty; z_1)$ und dass $L'(z) > 0$ für z in $(z_1;\ n_1)$. Damit ist $L(z)$ konvex in der Domäne $(-\infty, \min(n_1; n_2))$, und sein Minimalwert wird erreicht bei $z_1 \equiv z_m$.

Wenn $n_1 > n_2$, liegen zwei Fälle vor: der Fall, bei dem $n_1 - n_2 > 4/se$, und der Fall, bei dem $0 < n_1 - n_2 < 4/se$. Im ersten Fall sind z_1 und z_2 die Wurzeln von $Q(z)$, so dass $n_2 = \min(n_1; n_2) < z_1 < z_2$. (Dies liegt daran, dass $n_2 - \frac{z_1 + z_2}{2} = \frac{n_2 - n_1}{2} < 0$). Damit ist $Q(z) > 0$ für z in der Domäne $(-\infty, \min(n_1; n_2))$. Im zweiten Fall hat $Q(z)$ keine Wurzeln, so dass für die Domäne $Q(z) > 0$ gilt.

Daraus folgt: Wenn $n_1 > n_2$, dann $L'(z) < 0$, so dass $L(z)$ monoton von $+\infty$ bis $-\infty$ abnimmt.

8. Literaturhinweise

Banga, S. J. und Fox, G. D. (2013). On Bonett's Robust Confidence Interval for a Ratio of Standard Deviations. Im Druck.

Barnard, J. (1978). Probability Integral of the Normal Range. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 27, 197–198.

Bonett, D. G. (2006). Robust Confidence Interval for a Ratio of Standard Deviations. *Applied Psychological Measurements*, 30, 432–439.

Brown, M. B. und Forsythe A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69, 364–367.

Conover, W. J., Johnson, M. E. und Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, 23, 351–361.

Hochberg, Y., Weiss, G. und Hart S. (1982). On Graphical Procedures for Multiple Comparisons. *Journal of the American Statistical Association*, 77, 767–772.

Kramer, C. Y. (1956). Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12, 307–310.

Layard, M. W. J. (1973). Robust Large-Sample Tests for Homogeneity of Variances. *Journal of the American Statistical Association*, 68, 195–198.

Levene, H. (1960). "Robust Tests for Equality of Variances," in I. Olkin, ed., *Contributions to Probability and Statistics*, Palo Alto, CA: Stanford University Press, 278–292.

Miller, R. G. (1968). Jackknifing Variances. *Annals of Mathematical Statistics*, 39, 567–582.

Nakayama, M. K. (2009). Asymptotically Valid Single-Stage Multiple-Comparison Procedures. *Journal of Statistical Planning and Inference*, 139, 1348–1356.

Ott, R. L. und Longnecker, M. (2010). *An introduction to Statistical Methods and Data Analysis, sixth edition*, Brooks/Cole, Cengage Learning.

Pan, G. (1999). On a Levene Type Test for Equality of Two Variances. *Journal of Statistical Computation and Simulation*, 63, 59–71.

Stoline, M. R. (1981). The Status of Multiple of Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs. *The American Statistician*, 35, 134–141.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*. Mimeographed monograph.

Wolfram, S. (1999). *The Mathematica Book*, 4th ed. Wolfram Media/Cambridge University Press.