Dieses White Paper ist Teil einer Reihe von Veröffentlichungen, welche die Forschungsarbeiten der Minitab-Statistiker erläutern, in deren Rahmen die im Assistenten der Minitab Statistical Software verwendeten Methoden und Datenprüfungen entwickelt wurden.

Einfache ANOVA

Übersicht

Mit der einfachen ANOVA werden die Mittelwerte von drei oder mehr Gruppen verglichen, um zu ermitteln, ob sie signifikant voneinander abweichen. Eine weitere wichtige Funktion besteht darin, die Differenzen zwischen bestimmten Gruppen zu schätzen.

Die gängigste Methode zum Erkennen von Differenzen zwischen Gruppen in der einfachen ANOVA ist der F-Test. Dieser Test basiert auf der Annahme, dass die Grundgesamtheiten aller Stichproben eine gemeinsame Standardabweichung aufweisen, die jedoch unbekannt ist. Wir haben festgestellt, dass Stichproben in der Praxis häufig unterschiedliche Standardabweichungen haben. Daher sollte die Welch-Methode untersucht werden. Diese bietet eine Alternative zum F-Test, bei der auch mit ungleichen Standardabweichungen gearbeitet werden kann. Außerdem wollten wir eine Methode zum Berechnen von Mehrfachvergleichen entwickeln, bei der Stichproben mit ungleichen Standardabweichungen berücksichtigt werden. Mit dieser Methode können die einzelnen Intervalle grafisch dargestellt werden, anhand derer auf einfache Weise Gruppen ermittelt werden können, die sich voneinander unterscheiden.

Im vorliegenden White Paper erläutern wir, wie die Methoden entwickelt wurden, die in der einfachen ANOVA im Minitab-Assistenten für folgende Verfahren verwendet werden:

- Welch-Test
- Intervalle für Mehrfachvergleiche

Darüber hinaus betrachten wir die Bedingungen, die sich auf die Gültigkeit der Ergebnisse einer einfachen ANOVA auswirken können, u. a. das Vorhandensein ungewöhnlicher Daten, der Stichprobenumfang und die Trennschärfe des Tests sowie das Vorliegen einer Normalverteilung in den Daten. Auf der Grundlage dieser Bedingungen führt der Assistent automatisch die folgenden Datenprüfungen durch und gibt die Ergebnisse in der Auswertung aus:

- Ungewöhnliche Daten
- Stichprobenumfang

- Vorliegen einer Normalverteilung in den Daten

Im vorliegenden White Paper legen wir dar, wie sich diese Bedingungen in der Praxis auf die einfache ANOVA auswirken. Zudem beschreiben wir, auf welche Weise wir die Richtlinien zum Prüfen dieser Bedingungen im Assistenten entwickelt haben.

Methoden für die einfache ANOVA

F-Test und Welch-Test

Der üblicherweise bei der einfachen ANOVA verwendete F-Test beruht auf der Annahme, dass alle Gruppen die gleiche Standardabweichung (σ) aufweisen, die jedoch unbekannt ist. In der Praxis ist diese Annahme selten richtig, was zu Problemen in Bezug auf den Fehler 1. Art führt. Der Fehler 1. Art ist die Wahrscheinlichkeit, mit der die Nullhypothese fälschlicherweise zurückgewiesen wird (und die Schlussfolgerung gezogen wird, dass die Stichproben signifikant voneinander abweichen, obwohl dies nicht zutrifft). Wenn die Stichproben unterschiedliche Standardabweichungen aufweisen, besteht eine größere Wahrscheinlichkeit, dass der Test zu einer falschen Schlussfolgerung führt. Um dieses Problem zu lösen, wurde als Alternative zum F-Test der Welch-Test entwickelt (Welch, 1951).

Zielstellung

Es sollte ermittelt werden, ob für die einfache ANOVA im Assistenten der F-Test oder der Welch-Test verwendet werden soll. Dazu musste herausgefunden werden, wie gut die tatsächlichen Ergebnisse des F-Tests und des Welch-Tests mit dem Soll-Signifikanzniveau des Tests (Alpha bzw. Wahrscheinlichkeit des Fehlers 1. Art) übereinstimmen, d. h., ob die Nullhypothese mit dem Test bei unterschiedlichen Umfängen und Standardabweichungen der Stichproben häufiger bzw. seltener als vorgesehen fälschlicherweise zurückgewiesen wird.

Methode

Zum Vergleich des F-Tests und des Welch-Tests wurden mehrere Simulationen mit unterschiedlichen Anzahlen, Umfängen und Standardabweichungen der Stichproben durchgeführt. Unter jeder Bedingung wurden 10.000 ANOVA-Tests mit sowohl dem F-Test als auch der Welch-Methode durchgeführt. Für die Tests wurden Zufallsdaten generiert, so dass alle Stichproben den gleichen Mittelwert aufwiesen und daher bei jedem Test die Nullhypothese richtig war. Anschließend wurden die Tests mit Soll-Signifikanzniveaus von 0,05 und 0,01 durchgeführt. Es wurde gezählt, wie häufig die Nullhypothese in 10.000 Tests auf der Grundlage des F-Tests und des Welch-Tests zurückgewiesen wurde, und dieser Anteil wurde mit dem Soll-Signifikanzniveau verglichen. Wenn der Test eine gute Leistung zeigt, sollte der geschätzte Fehler 1. Art sehr nahe am Soll-Signifikanzniveau liegen.

Ergebnisse

Es hat sich herausgestellt, dass die Welch-Methode unter allen untersuchten Bedingungen im Vergleich zum F-Test dieselbe oder eine bessere Leistung zeigt. Beim Vergleich von 5 Stichproben unter Verwendung des Welch-Tests ergab sich beispielsweise eine Wahrscheinlichkeit eines Fehlers 1. Art zwischen 0,0460 und 0,0540, was sehr nahe am Soll-Signifikanzniveau von 0,05 liegt. Dies weist darauf hin, dass die Wahrscheinlichkeit eines Fehlers 1. Art bei der Welch-Methode dem Sollwert entspricht, auch wenn sich die Umfänge und Standardabweichungen bei den Stichproben unterscheiden.

Andererseits lagen die Wahrscheinlichkeiten eines Fehlers 1. Art für den F-Test zwischen 0,0273 und 0,2277. Der F-Test zeigte insbesondere unter folgenden Bedingungen eine schlechte Leistung:

- Die Wahrscheinlichkeiten eines Fehlers 1. Art fielen in der Situation unter 0,05, in der die größte Stichprobe auch die größte Standardabweichung aufwies. Diese Bedingung bewirkt einen konservativeren Test und veranschaulicht, dass das einfache Vergrößern des Stichprobenumfangs keine gangbare Lösung ist, wenn die Standardabweichungen für die Stichproben ungleich sind.
- Die Wahrscheinlichkeiten eines Fehlers 1. Art lagen über 0,05, wenn die Stichprobenumfänge gleich, die Standardabweichungen jedoch unterschiedlich waren. Die Wahrscheinlichkeiten waren ebenfalls größer als 0,05, wenn die Stichprobe mit einer größeren Standardabweichung einen kleineren Umfang als die anderen Stichproben aufwies. Insbesondere in Situationen, in denen kleinere Stichproben größere Standardabweichungen aufweisen, ist ein wesentlicher Anstieg des Risikos zu verzeichnen, dass dieser Test die Nullhypothese fälschlicherweise zurückweist.

Weitere Informationen zur Methodologie der Simulation und zu deren Ergebnissen finden Sie in Anhang A.

Da die Welch-Methode bei ungleichen Standardabweichungen und Umfängen der Stichproben eine gute Leistung zeigte, nutzen wir die Welch-Methode für das Verfahren der einfachen ANOVA im Assistenten.

Vergleichsintervalle

Wenn ein ANOVA-Test statistisch signifikant ist und somit darauf hinweist, dass sich mindestens ein Stichprobenmittelwert von den anderen unterscheidet, wird im nächsten Schritt der Analyse bestimmt, welche Stichproben sich signifikant voneinander unterscheiden. Eine intuitive Vergleichsmöglichkeit besteht darin, die Konfidenzintervalle grafisch darzustellen und die Stichproben zu bestimmen, deren Intervalle einander nicht überlappen. Die aus der grafischen Darstellung gezogenen Schlussfolgerungen entsprechen jedoch u. U. nicht den Testergebnissen, da die einzelnen Konfidenzintervalle nicht auf Vergleiche ausgelegt sind. Es gibt zwar eine veröffentlichte Methode für Mehrfachvergleiche von Stichproben mit gleichen Standardabweichungen, wir mussten diese Methode jedoch so erweitern, dass auch Stichproben mit ungleichen Standardabweichungen berücksichtigt werden.

Zielstellung

Es sollte eine Methode zum Berechnen von einzelnen Vergleichsintervallen entwickelt werden, mit denen alle Stichproben miteinander verglichen werden können und die zudem so weit wie möglich den Testergebnissen entsprechen. Außerdem sollte eine visuelle Methode bereitgestellt werden, mit der bestimmt werden kann, welche Stichproben sich statistisch von den anderen unterscheiden.

Methode

Standardmethoden für Mehrfachvergleiche (Hsu 1996) liefern für jedes Paar von Mittelwerten ein Intervall für die Differenz, wobei für den Einfluss des beim Durchführen von Mehrfachvergleichen höheren Fehlers kontrolliert wird. In besonderen Fall von gleichen Stichprobenumfängen und unter der Annahme gleicher Standardabweichungen können einzelne Intervalle für jeden Mittelwert auf eine Weise angezeigt werden, die genau den Intervallen für die Differenzen aller Paare entspricht. Für ungleiche Stichprobenumfängen und unter der Annahme gleicher Standardabweichungen haben Hochberg, Weiss und Hart (1982) Einzelintervalle entwickelt, die annähernd den Intervallen für Differenzen zwischen Paaren nach der Tukey-Kramer-Methode für Mehrfachvergleiche entsprechen. Im Assistenten wenden wir denselben Ansatz auf die Games-Howell-Methode für Mehrfachvergleiche an, bei der keine Gleichheit der Standardabweichungen angenommen wird. Der im Assistenten in Minitab Release 16 verfolgte Ansatz war konzeptionell sehr ähnlich, er basierte jedoch nicht direkt auf dem Games-Howell-Ansatz. Weitere Informationen finden Sie in Anhang A.

Ergebnisse

Der Assistent zeigt die Vergleichsintervalle im Vergleichsdiagramm für die Mittelwerte in der Zusammenfassung der einfachen ANOVA an. Wenn der ANOVA-Test statistisch signifikant ist, wird jedes Vergleichsintervall, das nicht mit mindestens einem anderen Intervall überlappt, rot gekennzeichnet. Test und Vergleichsintervalle können einander widersprechen. Ein solches Ergebnis ist jedoch selten, da beide Methoden dieselbe Wahrscheinlichkeit aufweisen, dass die Nullhypothese zurückgewiesen wird, wenn sie tatsächlich wahr ist. Wenn der ANOVA-Test signifikant ist, sich jedoch alle Intervalle überlappen, wird das Paar mit der geringsten Überlappung rot gekennzeichnet. Wenn der ANOVA-Test statistisch nicht signifikant ist, werden keine der Intervalle rot gekennzeichnet, selbst wenn für einige Intervalle keine Überlappung vorliegt.

Datenprüfungen

Ungewöhnliche Daten

Ungewöhnliche Daten sind extrem große oder kleine Datenwerte, die auch als Ausreißer bezeichnet werden. Ungewöhnliche Daten können einen starken Einfluss auf die Ergebnisse der Analyse ausüben, und sie können sich auf die Wahrscheinlichkeiten auswirken, dass statistisch signifikante Ergebnisse gefunden werden. Dies gilt insbesondere für kleine Stichproben. Ungewöhnliche Daten können auf Probleme bei der Datenerfassung hinweisen, sie können aber auch auf ein ungewöhnliches Verhalten des untersuchten Prozesses zurückzuführen sein. Daher ist es häufig unverzichtbar, diese Datenpunkte zu untersuchen und nach Möglichkeit zu korrigieren.

Zielstellung

Es sollte eine Methode zum Überprüfen von Datenwerten entwickelt werden, die relativ zur Gesamtstichprobe sehr groß bzw. sehr klein sind und sich auf die Ergebnisse der Analyse auswirken können.

Methode

Wir haben eine Methode zum Prüfen auf ungewöhnliche Daten entwickelt, die auf der von Hoaglin, Iglewicz und Tukey (1986) beschriebenen Methode zum Identifizieren von Ausreißern in Boxplots basiert.

Ergebnisse

Der Assistent identifiziert einen Datenpunkt als ungewöhnlich, wenn er um mehr als das 1,5-fache des Interquartilbereichs jenseits des unteren oder oberen Quartils der Verteilung liegt. Das untere und das obere Quartil stellen das 25. und das 75. Perzentil der Daten dar. Der Interquartilbereich gibt die Differenz zwischen den beiden Quartilen an. Diese Methode liefert selbst dann gute Ergebnisse, wenn mehrere Ausreißer vorhanden sind, da damit jeder einzelne Ausreißer erkannt werden kann.

Für die Prüfung auf ungewöhnliche Daten werden in der Auswertung des Assistenten die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Es gibt keine ungewöhnlichen Datenpunkte.
	Mindestens ein Datenpunkt ist ungewöhnlich und wirkt sich möglicherweise stark auf die Ergebnisse aus.

Stichprobenumfang

Die Trennschärfe ist eine wichtige Eigenschaft jedes Hypothesentests, da sie die Wahrscheinlichkeit angibt, mit der Sie einen signifikanten Effekt oder eine signifikante Differenz erkennen, sofern dieser bzw. diese tatsächlich vorhanden ist. Die Trennschärfe ist die Wahrscheinlichkeit, mit der die Nullhypothese zugunsten der Alternativhypothese zurückgewiesen wird. Häufig kann die Trennschärfe eines Tests am einfachsten erhöht werden, indem der Stichprobenumfang vergrößert wird. Im Assistenten wird für Tests mit niedriger Trennschärfe angegeben, wie umfassend die Stichprobe sein muss, damit die angegebene Differenz erkannt wird. Ist keine Differenz angegeben, so wird die Differenz ausgegeben, die mit adäquater Trennschärfe erkannt werden könnte. Für die Ausgabe dieser Informationen mussten wir eine Methode zum Berechnen der Trennschärfe entwickeln, da für die im Assistenten verwendete Welch-Methode keine exakte Formel für die Trennschärfe verfügbar ist.

Zielstellung

Beim Entwickeln einer Methode zum Berechnen der Trennschärfe mussten zwei Probleme gelöst werden. Erstens fordert der Assistent nicht, dass der Benutzer einen kompletten Satz von Mittelwerten eingibt. Stattdessen muss lediglich eine Differenz zwischen Mittelwerten eingegeben werden, die praktische Konsequenzen hat. Für jede angegebene Differenz gibt es eine unendliche Anzahl möglicher Konfigurationen der Mittelwerte, die die betreffende Differenz aufweisen. Da eine Berechnung der Trennschärfe für alle möglichen Konfigurationen der Mittelwerte nicht möglich ist, mussten wir einen sinnvollen Ansatz entwickeln, mit dem die relevanten Mittelwerte zum Berechnen der Trennschärfe bestimmt werden können. Zweitens musste eine Methode zum Berechnen der Trennschärfe entwickelt werden, da der Assistent die Welch-Methode verwendet, bei der keine gleichen Stichprobenumfänge oder Standardabweichungen erforderlich sind.

Methode

Um mit der unendlichen Anzahl möglicher Konfigurationen der Mittelwerte umgehen zu können, haben wir eine Methode entwickelt, die auf dem Ansatz in der regulären einfachen ANOVA in Minitab (**Statistik** > **Varianzanalyse (ANOVA)** > **Einfache ANOVA**) basiert. Der Schwerpunkt lag dabei auf den Fällen, in denen sich nur zwei der Mittelwerte um den angegebenen Betrag unterscheiden und die übrigen Mittelwerte gleich sind (festgelegt auf den gewichteten Durchschnitt der Mittelwerte). Da angenommen wird, dass nur zwei Mittelwerte vom Gesamtmittelwert abweichen (und nicht mehr als zwei), liefert der Ansatz einen konservativen Schätzwert der Trennschärfe. Da die Stichproben jedoch unterschiedliche Umfänge oder Standardabweichungen aufweisen können, hängt die Berechnung der Trennschärfe immer noch davon ab, für welche zwei Mittelwerte eine Differenz angenommen wird.

Zum Beheben dieses Problems werden die zwei Paare von Mittelwerten identifiziert, die den besten und den schlechtesten Fall darstellen. Der schlechteste Fall tritt ein, wenn der Stichprobenumfang relativ zur Stichprobenvarianz klein ist und die Trennschärfe minimiert wird. Der beste Fall tritt ein, wenn der Stichprobenumfang relativ zur Stichprobenvarianz groß ist und die Trennschärfe maximiert wird. In sämtlichen Trennschärfeberechnungen

werden diese beiden Extremfälle betrachtet, welche die Trennschärfe unter der Annahme, dass genau zwei Mittelwerte vom gewichteten Gesamtdurchschnitt der Mittelwerte abweichen, minimieren bzw. maximieren.

Die Trennschärfeberechnung wurde anhand einer Methode entwickelt, die in Kulinskaya et al. (2003) beschrieben wird. Wir haben die Trennschärfeberechnungen aus unserer Simulation, der von uns entwickelten Methode zum Bewältigen der Konfiguration der Mittelwerte und der in Kulinskaya et al. (2003) beschriebenen Methode verglichen. Zudem wurde eine weitere Trennschärfe-Approximation untersucht, die deutlicher aufzeigt, wie die Trennschärfe von der Konfiguration der Mittelwerte abhängt. Weitere Informationen zur Trennschärfeberechnung finden Sie in Anhang C.

Ergebnisse

Ein Vergleich dieser Methoden zeigte, dass die Methode nach Kulinskaya eine gute Approximation der Trennschärfe liefert und unsere Methode zum Umgang mit der Konfiguration von Mittelwerten geeignet ist.

Wenn die Daten keine ausreichenden Hinweise liefern, die gegen die Nullhypothese sprechen, berechnet der Assistent Differenzen mit praktischen Konsequenzen, die für die angegebenen Stichprobenumfänge mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden können. Wenn Sie zudem eine Differenz mit praktischen Konsequenzen angeben, berechnet der Assistent die minimale und die maximale Trennschärfe für die betreffende Differenz. Wenn die Trennschärfewerte unter 90 % liegen, berechnet der Assistent einen Stichprobenumfang auf der Grundlage der angegebenen Differenz und den beobachteten Standardabweichungen der Stichproben. Um sicherzustellen, dass der Stichprobenumfang eine minimale und eine maximale Trennschärfe ergibt, die beide über 90 % liegen, wird angenommen, dass es sich bei der angegebenen Differenz um die Differenz zwischen den beiden Mittelwerten mit der größten Streuung handelt.

Wenn der Benutzer keine Differenz angibt, bestimmt der Assistent die größte Differenz, bei der das Maximum des Bereichs von Trennschärfewerten 60 % beträgt. Dieser Wert wird an der Grenze zwischen dem roten und dem gelben Balken im Trennschärfebericht beschriftet und entspricht einer Trennschärfe von 60 %. Darüber hinaus wird die kleinste Differenz bestimmt, bei der das Minimum des Bereichs von Trennschärfewerten 90 % beträgt. Dieser Wert wird an der Grenze zwischen dem gelben und dem grünen Balken im Trennschärfebericht beschriftet und entspricht einer Trennschärfe von 90 %.

Für die Prüfung auf die Trennschärfe und den Stichprobenumfang werden in der Auswertung des Assistenten die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Die Daten bieten keine ausreichenden Anzeichen für die Schlussfolgerung, dass eine Differenz zwischen den Mittelwerten besteht. Es wurde keine Differenz angegeben.

Status	Bedingung
	Im Test wird eine Differenz zwischen den Mittelwerten festgestellt, daher stellt die Trennschärfe kein Problem dar. ODER Die Trennschärfe ist ausreichend. Im Test wurde keine Differenz zwischen den Mittelwerten festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 90 % erkannt wird.
	Die Trennschärfe ist möglicherweise ausreichend. Im Test wurde keine Differenz zwischen den Mittelwerten festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 80 % bis 90 % erkannt wird. Der erforderliche Stichprobenumfang zum Erzielen einer Trennschärfe von 90 % wird ausgegeben.
	Die Trennschärfe ist möglicherweise nicht ausreichend. Im Test wurde keine Differenz zwischen den Mittelwerten festgestellt, und die Stichprobe ist umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 60 % bis 80 % erkannt wird. Die erforderlichen Stichprobenumfänge zum Erzielen einer Trennschärfe von 80 % und 90 % werden ausgegeben.
	Die Trennschärfe ist nicht ausreichend. Im Test wurde keine Differenz zwischen den Mittelwerten festgestellt, und die Stichprobe ist nicht groß genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 60 % erkannt wird. Die erforderlichen Stichprobenumfänge zum Erzielen einer Trennschärfe von 80 % und 90 % werden ausgegeben.

Vorliegen einer Normalverteilung

In vielen statistischen Methoden wird angenommen, dass die Daten normalverteilt sind. Erfreulicherweise können auf der Annahme der Normalverteilung basierende Methoden selbst dann eine gute Leistung liefern, wenn die Daten nicht normalverteilt sind. Dieser Umstand wird teilweise durch den zentralen Grenzwertsatz erklärt, der besagt, dass die Verteilung jedes Stichprobenmittelwerts einer annähernden Normalverteilung folgt und dass diese Approximation bei zunehmendem Stichprobenumfang nahezu zu einer Normalverteilung wird.

Zielstellung

Es sollte bestimmt werden, welchen Umfang die Stichprobe aufweisen muss, um eine angemessen gute Approximation der Normalverteilung zu erzielen. Es sollten der Welch-Test und Vergleichsintervalle bei kleinen bis mittleren Stichproben mit verschiedenen Nicht-Normalverteilungen untersucht werden. Dabei sollte ermittelt werden, wie genau die tatsächlichen Testergebnisse für die Welch-Methode und die Vergleichsintervalle dem gewählten Signifikanzniveau (Alpha oder Wahrscheinlichkeit eines Fehlers 1. Art) für den Test entsprechen, d. h. ob der Test die Nullhypothese für unterschiedliche Stichprobenumfänge, Anzahlen der Stufen und Nicht-Normalverteilungen häufiger oder seltener als erwartet fälschlicherweise zurückweist.

Methode

Zum Schätzen des Fehlers 1. Art wurden mehrere Simulationen durchgeführt, wobei die Anzahl der Stichproben, der Stichprobenumfang und die Verteilung der Daten variiert wurden. Die Simulationen umfassten schiefe Verteilungen und Verteilungen mit stärker

besetzten Randbereichen, die erheblich von der Normalverteilung abweichen. Umfang und Standardabweichung waren in jedem Test für alle Stichproben konstant.

Für jede Bedingung wurden 10.000 ANOVA-Tests mit der Welch-Methode und den Vergleichsintervallen durchgeführt. Für die Tests wurden Zufallsdaten generiert, so dass alle Stichproben den gleichen Mittelwert aufwiesen und daher bei jedem Test die Nullhypothese richtig war. Dann wurden die Tests mit einem Soll-Signifikanzniveau von 0,05 durchgeführt. Es wurde gezählt, wie häufig die Nullhypothese in 10.000 Tests tatsächlich zurückgewiesen wurde, und dieser Anteil wurde mit dem Soll-Signifikanzniveau verglichen. Für die Vergleichsintervalle wurde gezählt, wie oft die Intervalle in 10.000 Tests eine oder mehrere Differenzen angaben. Wenn der Test eine gute Leistung zeigt, sollten der Fehler 1. Art sehr nahe am Soll-Signifikanzniveau liegen.

Ergebnisse

Insgesamt bieten die Tests und die Vergleichsintervalle für alle Bedingungen selbst bei kleinen Stichprobenumfängen von 10 oder 15 eine sehr gute Leistung. Für Tests mit bis zu 9 Stufen liegen die Ergebnisse in fast jedem Fall bei einem Stichprobenumfang von 10 innerhalb von 3 Prozentpunkten und bei einem Stichprobenumfang von 15 innerhalb von 2 Prozentpunkten vom Soll-Signifikanzniveau. Für Tests mit 10 oder mehr Stufen liegen die Ergebnisse in den meisten Fällen bei einem Stichprobenumfang von 15 innerhalb von 3 Prozentpunkten und bei einem Stichprobenumfang von 20 innerhalb von 2 Prozentpunkten. Weitere Informationen finden Sie in Anhang D.

Da diese Tests bei relativ kleinen Stichproben eine gute Leistung zeigen, testet der Assistent die Daten nicht auf eine Normalverteilung. Stattdessen prüft der Assistent den Stichprobenumfang und gibt einen entsprechenden Hinweis aus, wenn die Stichproben für 2 bis 9 Stufen kleiner als 15 und für 10 bis 12 Stufen kleiner als 20 sind. Auf der Grundlage dieser Ergebnisse zeigt der Assistent in der Auswertung die folgenden Statusindikatoren an:

Status	Bedingung
	Die Stichprobenumfänge betragen mindestens 15 oder 20, daher ist es kein Problem, wenn keine Normalverteilung vorliegt.
	Da einige Stichprobenumfänge kleiner als 15 oder 20 sind, kann es ein Problem sein, wenn keine Normalverteilung vorliegt.

Literaturhinweise

Dunnet, C. W. (1980). Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association, 75, 796-800*.

Hoaglin, D. C., Iglewicz, B. und Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81, 991-999*.

Hochberg, Y., Weiss, G. und Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association, 77, 767-772*.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall.

Kulinskaya, E., Staudte, R. G. und Gao, H. (2003). Power approximations in testing for unequal means in a One-Way ANOVA weighted for unequal variances, *Communication in Statistics, 32 (12), 2353-2371*.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34, 28-35*

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38, 330-336*.

Anhang A: F-Test und Welch-Test

Der F-Test kann einen Anstieg der Wahrscheinlichkeit eines Fehlers 1. Art nach sich ziehen, wenn die Annahme gleicher Standardabweichungen verletzt wird. Der Welch-Test ist so konzipiert, dass derartige Probleme vermieden werden.

Welch-Test

Es werden Zufallsstichproben der Umfänge n_1, ..., n_k aus k Grundgesamtheiten beobachtet. Seien $\mu_1,\ldots,\mu_k$ die Mittelwerte der Grundgesamtheit und $\sigma_1^2, \ldots, \sigma_k^2$ die Varianzen der Grundgesamtheit. Seien $\bar{x}_1, \ldots, \bar{x}_k$ die Mittelwerte der Stichproben und $s_1^2, \ldots, s_k^2$ die Varianzen der Stichproben. Die folgende Hypothese soll getestet werden:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

H_1: $\mu_i \neq \mu_j$ für einige i, j.

Der Welch-Test zum Testen auf Gleichheit von k Mittelwerten vergleicht die Statistik

$$W^* = \frac{\sum_{j=1}^{k} w_j (\bar{x}_j - \hat{\mu})^2 / (k-1)}{1 + [2(k-2)/(k^2-1)] \sum_{j=1}^{k} h_j}$$

mit der Verteilung F(k – 1; f), wobei

$$w_j = \frac{n_j}{s_j^2},$$

$$W = \sum_{j=1}^{k} w_j,$$

$$\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{W},$$

$$h_j = \frac{(1 - w_j/W)^2}{n_j - 1} \text{ und}$$

$$f = \frac{k^2 - 1}{3 \sum_{j=1}^{k} h_j}.$$

Der Welch-Test weist die Nullhypothese zurück, wenn $W^* \geq F_{k-1, f, 1-\alpha}$, das Perzentil der F-Verteilung, das mit der Wahrscheinlichkeit α überschritten wird.

Ungleiche Standardabweichungen

In diesem Abschnitt wird die Empfindlichkeit des F-Tests in Bezug auf Verletzungen der Annahme gleicher Standardabweichungen veranschaulicht, und es wird ein Vergleich mit dem Welch-Test vorgenommen.

Die nachfolgenden Ergebnisse beziehen sich auf Tests der einfachen ANOVA mit 5 Stichproben von $N(0; \sigma^2)$. Jede Zeile basiert auf 10.000 Simulationen mit dem F-Test und dem Welch-Test. Es wurden zwei Bedingungen für die Standardabweichung getestet. Dazu wurde die Standardabweichung der fünften Stichprobe in Bezug auf die anderen Stichproben verdoppelt und vervierfacht. Für den Stichprobenumfang wurden drei unterschiedliche Bedingungen getestet: Die Stichprobenumfänge sind gleich, die fünfte Stichprobe ist größer als die anderen, und die fünfte Stichprobe ist kleiner als die anderen.

Tabelle 1 Wahrscheinlichkeiten eines Fehlers 1. Art für simulierte F-Tests und Welch-Tests mit 5 Stichproben mit dem Soll-Signifikanzniveau $\alpha = 0{,}05$

Standardabweichung ($\sigma 1$; $\sigma 2$; $\sigma 3$; $\sigma 4$; $\sigma 5$)	Stichprobenumfang (n1; n2; n3; n4; n5)	F-Test	Welch-Test
1; 1; 1; 1; 2	10; 10; 10; 10; 20	0,0273	0,0524
1; 1; 1; 1; 2	20; 20; 20; 20; 20	0,0678	0,0462
1; 1; 1; 1; 2	20; 20; 20; 20; 10	0,1258	0,0540
1; 1; 1; 1; 4	10; 10; 10; 10; 20	0,0312	0,0460
1; 1; 1; 1; 4	20; 20; 20; 20; 20	0,1065	0,0533
1; 1; 1; 1; 4	20; 20; 20; 20; 10	0,2277	0,0503

Bei Gleichheit der Stichprobenumfänge (Zeilen 2 und 5) ist die Wahrscheinlichkeit, dass der F-Test die Nullhypothese fälschlicherweise zurückweist, größer als das Soll 0,05, und die Wahrscheinlichkeit steigt bei größerer Ungleichheit der Standardabweichungen. Dieses Problem vergrößert sich, wenn der Umfang der Stichprobe mit der größten Standardabweichung verringert wird. Wenn hingegen der Umfang der Stichprobe mit der größten Standardabweichung vergrößert wird, nimmt die Wahrscheinlichkeit der Zurückweisung ab. Eine zu starke Vergrößerung des Stichprobenumfangs führt jedoch zu einer zu geringen Wahrscheinlichkeit der Zurückweisung. Hierdurch wird nicht nur der Test unter der Nullhypothese konservativer als notwendig, auch die Trennschärfe des Tests unter der Alternativhypothese wird beeinträchtigt. Vergleichen Sie diese Ergebnisse mit dem Welch-Test, der dem Soll-Signifikanzniveau von 0,05 in jedem Fall gut entspricht.

Anschließend wurde eine Simulation für Fälle mit k = 7 Stichproben durchgeführt. In jeder Zeile der Tabelle sind 10.000 simulierte F-Tests zusammengefasst. Die Standardabweichungen und die Umfänge der Stichproben wurden variiert. Die Soll-Signifikanzniveaus sind $\alpha = 0{,}05$ und $\alpha = 0{,}01$. Wie oben wurden teils sehr starke Abweichungen von den Sollwerten festgestellt. Das Verkleinern des Stichprobenumfangs bei einer höheren Streuung führt zu einer sehr großen Wahrscheinlichkeit eines Fehlers 1. Art, während das Vergrößern des Stichprobenumfangs zu einem extrem konservativen Test führen kann. Die Ergebnisse finden Sie in der nachfolgenden Tabelle 2.

Tabelle 2 Wahrscheinlichkeiten eines Fehlers 1. Art für simulierte F-Tests mit 7 Stichproben

Standardabweichung ($\sigma 1$; $\sigma 2$; $\sigma 3$; $\sigma 4$; $\sigma 5$; $\sigma 6$; $\sigma 7$)	Stichprobenumfänge (n1; n2; n3; n4; n5; n6; n7)	Soll-α = 0,05	Soll-α = 0,01
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	21; 21; 21; 21; 22; 22; 12	0,0795	0,0233
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	20; 21; 21; 21; 21; 24; 12	0,0785	0,0226
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	20; 21; 21; 21; 21; 21; 15	0,0712	0,0199
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	20; 20; 20; 21; 21; 23; 15	0,0719	0,0172

Standardabweichung (σ1; σ2; σ3; σ4; σ5; σ6; σ7)	Stichprobenumfänge (n1; n2; n3; n4; n5; n6; n7)	Soll-α = 0,05	Soll-α = 0,01
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	20; 20; 20; 20; 21; 21; 18	0,0632	0,0166
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	20; 20; 20; 20; 20; 20; 20	0,0576	0,0138
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	18; 19; 19; 20; 20; 20; 24	0,0474	0,0133
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	18; 18; 18; 18; 18; 18; 32	0,0314	0,0057
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	15; 18; 18; 19; 20; 20; 30	0,0400	0,0085
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	12; 18; 18; 18; 19; 19; 36	0,0288	0,0064
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	15; 15; 15; 15; 15; 15; 50	0,0163	0,0025
1,85; 1,85; 1,85; 1,85; 1,85; 1,85; 2,9	12; 12; 12; 12; 12; 12; 68	0,0052	0,0002
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	21; 21; 21; 21; 22; 22; 12	0,1097	0,0436
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	20; 21; 21; 21; 21; 24; 12	0,1119	0,0452
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	20; 21; 21; 21; 21; 21; 15	0,0996	0,0376
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	20; 20; 20; 21; 21; 23; 15	0,0657	0,0345
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	20; 20; 20; 20; 21; 21; 18	0,0779	0,0283
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	20; 20; 20; 20; 20; 20; 20	0,0737	0,0264
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	18; 19; 19; 20; 20; 20; 24	0,0604	0,0204
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	18; 18; 18; 18; 18; 18; 32	0,0368	0,0122
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	15; 18; 18; 19; 20; 20; 30	0,0390	0,0117
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	12; 18; 18; 18; 19; 19; 36	0,0232	0,0046
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	15; 15; 15; 15; 15; 15; 50	0,0124	0,0026
1,75; 1,75; 1,75; 1,75; 1,75; 1,75; 3,5	12; 12; 12; 12; 12; 12; 68	0,0027	0,0004
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	21; 21; 21; 21; 22; 22; 12	0,1340	0,0630
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	20; 21; 21; 21; 21; 24; 12	0,1329	0,0654
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	20; 21; 21; 21; 21; 21; 15	0,1101	0,0484
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	20; 20; 20; 21; 21; 23; 15	0,1121	0,0495
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	20; 20; 20; 20; 21; 21; 18	0,0876	0,0374

Standardabweichung (σ1; σ2; σ3; σ4; σ5; σ6; σ7)	Stichprobenumfänge (n1; n2; n3; n4; n5; n6; n7)	Soll-α = 0,05	Soll-α = 0,01
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	20; 20; 20; 20; 20; 20; 20	0,0808	0,0317
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	18; 19; 19; 20; 20; 20; 24	0,0606	0,0243
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	18; 18; 18; 18; 18; 18; 32	0,0356	0,0119
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	15; 18; 18; 19; 20; 20; 30	0,0412	0,0134
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	12; 18; 18; 18; 19; 19; 36	0,0261	0,0068
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	15; 15; 15; 15; 15; 15; 50	0,0100	0,0023
1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 1,68333; 3,9	12; 12; 12; 12; 12; 12; 68	0,0017	0,0003
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	21; 21; 21; 21; 22; 22; 12	0,1773	0,1006
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	20; 21; 21; 21; 21; 24; 12	0,1811	0,1040
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	20; 21; 21; 21; 21; 21; 15	0,1445	0,0760
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	20; 20; 20; 21; 21; 23; 15	0,1448	0,0786
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	20; 20; 20; 20; 21; 21; 18	0,1164	0,0572
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	20; 20; 20; 20; 20; 20; 20	0,1020	0,0503
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	18; 19; 19; 20; 20; 20; 24	0,0834	0,0369
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	18; 18; 18; 18; 18; 18; 32	0,0425	0,0159
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	15; 18; 18; 19; 20; 20; 30	0,0463	0,0168
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	12; 18; 18; 18; 19; 19; 36	0,0305	0,0103
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	15; 15; 15; 15; 15; 15; 50	0,0082	0,0021
1,55; 1,55; 1,55; 1,55; 1,55; 1,55; 4,7	12; 12; 12; 12; 12; 12; 68	0,0013	0,0001

Anhang B: Vergleichsintervalle

Anhand des Vergleichsdiagramms für die Mittelwerte können Sie die statistische Signifikanz der Differenzen zwischen den Mittelwerten der Grundgesamtheit auswerten.

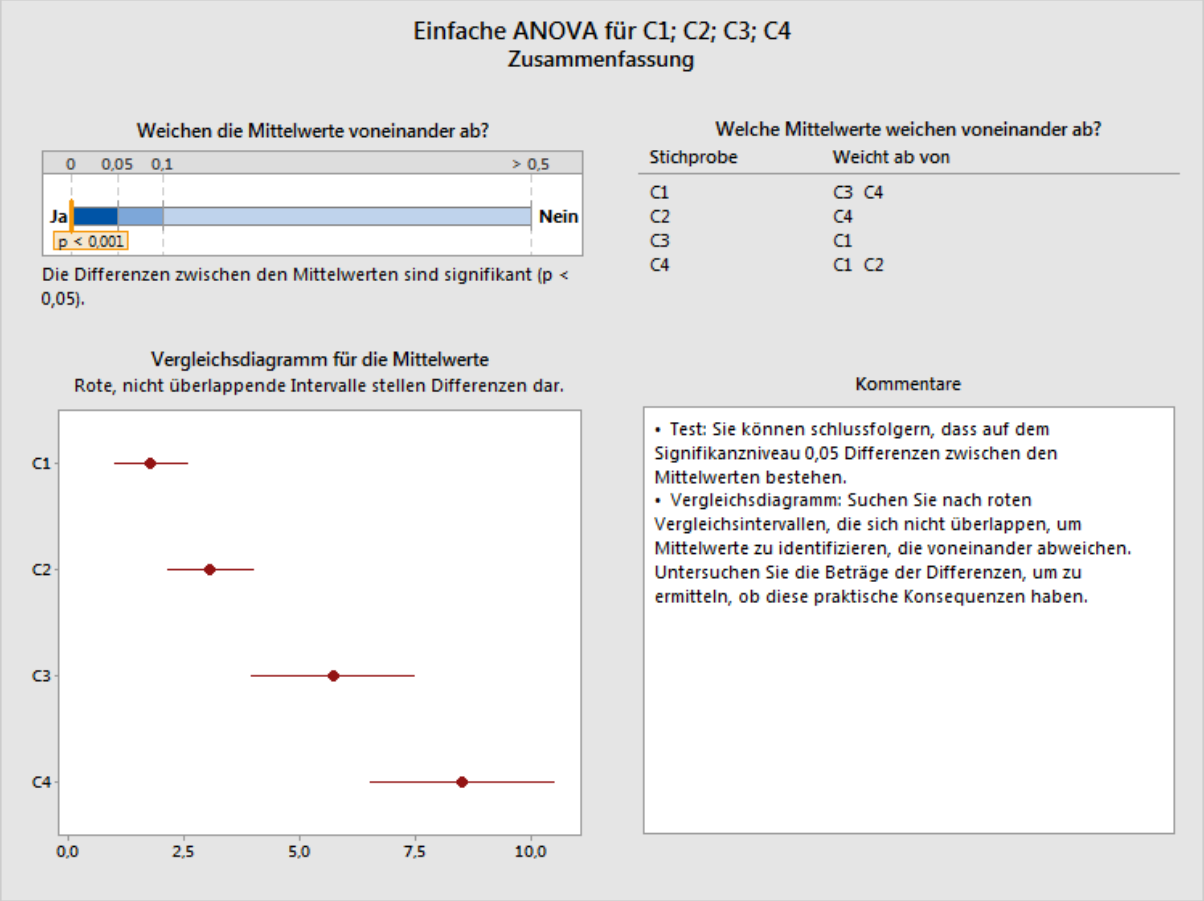
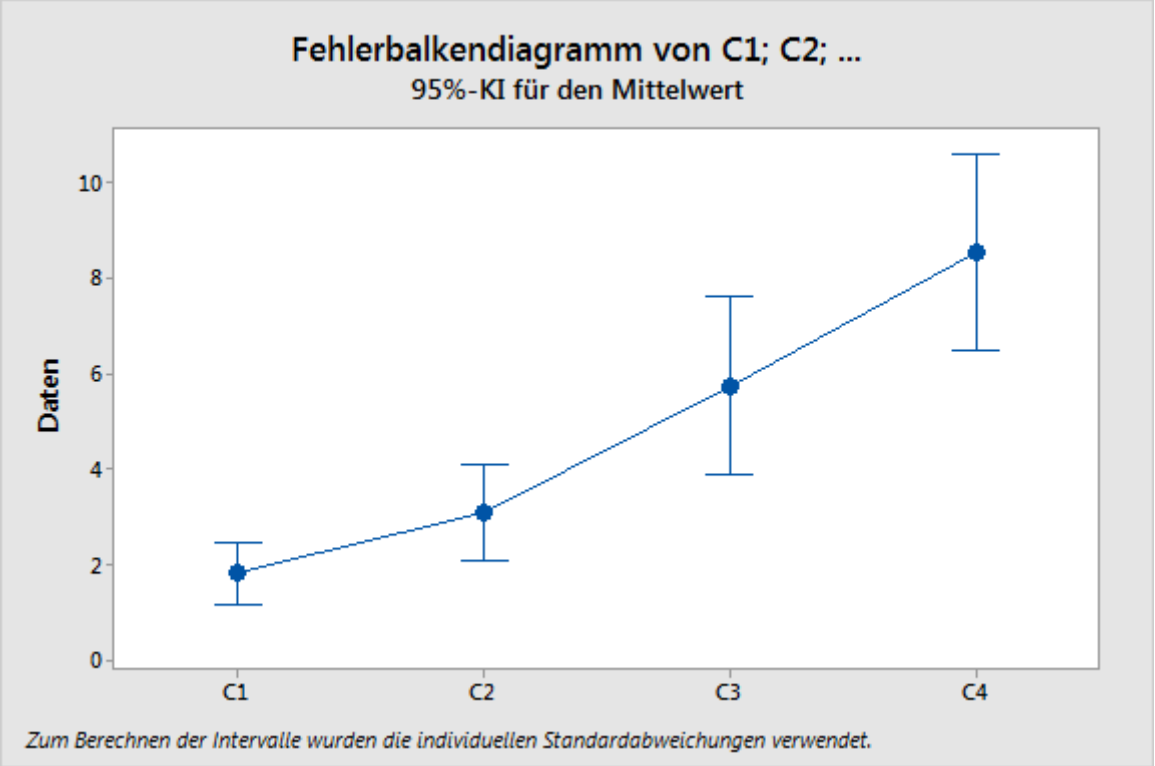


Abbildung 1 Das Vergleichsdiagramm für die Mittelwerte in der Auswertung der einfachen ANOVA im Assistenten

Ein ähnlicher Satz von Intervallen wird in der Ausgabe für die reguläre einfache ANOVA in Minitab angezeigt (**Statistik** > **Varianzanalyse (ANOVA)** > **Einfache ANOVA**):



Beachten Sie jedoch, dass es sich bei den oben gezeigten Intervallen lediglich um individuelle Konfidenzintervalle für die Mittelwerte handelt. Wenn aufgrund des ANOVA-Tests (entweder F-Test oder Welch-Test) geschlussfolgert wird, dass sich einige Mittelwerte voneinander unterscheiden, liegt es nahe, nach Intervallen zu suchen, die einander nicht überlappen, und aus diesen zu schließen, welche Mittelwerte abweichen. Eine solche informelle Analyse der individuellen Konfidenzintervalle führt häufig zu sinnvollen Schlussfolgerungen, bietet jedoch nicht dieselbe Kontrolle über die Fehlerwahrscheinlichkeit wie der ANOVA-Test. Je nach Anzahl der Grundgesamtheiten wird anhand der Intervalle möglicherweise mit erheblich größerer oder geringerer Wahrscheinlichkeit als beim Test geschlussfolgert, dass Differenzen vorliegen. Daher ist es leicht möglich, dass mit den beiden Methoden widersprüchliche Schlussfolgerungen gezogen werden. Das Vergleichsdiagramm ist so konzipiert, dass bei Mehrfachvergleichen beständiger eine Übereinstimmung mit den Ergebnissen des Welch-Tests erzielt wird, wobei es nicht immer möglich ist, eine vollständige Übereinstimmung zu erzielen.

Mehrfachvergleichsmethoden wie die Vergleiche nach Tukey-Kramer und Games-Howell in Minitab (**Statistik** > **Varianzanalyse (ANOVA)** > **Einfache ANOVA**) ermöglichen es Ihnen, statistisch gültige Schlussfolgerungen zu Differenzen zwischen den einzelnen Mittelwerten zu ziehen. Bei diesen beiden Methoden handelt es sich um paarweise Vergleichsmethoden, die ein Intervall für die Differenz zwischen jedem Paar von Mittelwerten ausgeben. Die Wahrscheinlichkeit, dass alle Intervalle gleichzeitig die geschätzten Differenzen enthalten, liegt bei mindestens $1 - \alpha$. Die Tukey-Kramer-Methode hängt von der Annahme gleicher Varianzen ab, während für die Games-Howell-Methode keine Gleichheit der Varianzen erforderlich ist. Wenn die Nullhypothese gleicher Mittelwerte zutreffend ist, sind alle

Differenzen null, und die Wahrscheinlichkeit, dass eines der Games-Howell-Intervalle nicht null enthält, beträgt höchstens α. Daher kann mit den Intervallen ein Hypothesentest mit dem Signifikanzniveau α ausgeführt werden. Wir nutzen Games-Howell-Intervalle als Ausgangspunkt zum Ableiten der Intervalle im Vergleichsdiagramm des Assistenten.

Gegeben sei eine Gruppe von Intervallen [L_{ij}, U_{ij}] für alle Differenzen $\mu_i - \mu_j$, $1 \leq i < j \leq k$, davon ausgehend soll eine Gruppe von Intervallen [L_i, U_i] für die einzelnen Mittelwerte μ_i, $1 \leq i \leq k$ gefunden werden, die dieselben Informationen liefert. Dies erfordert, dass sich jede Differenz d nur dann im Intervall [L_{ij}, U_{ij}] befindet, wenn $\mu_i \in [L_i, U_i]$ und $\mu_j \in [L_j, U_j]$ vorhanden sind, so dass $\mu_i - \mu_j = d$. Die Endpunkte der Intervalle müssen die durch die folgenden Gleichungen dargestellte Beziehung aufweisen:

$$U_i - L_j = U_{ij} \text{ und}$$

$$L_i - U_j = L_{ij}.$$

Für k = 2 ist nur eine Differenz vorhanden, jedoch zwei einzelne Intervalle. Daher ist es möglich, exakte Vergleichsintervalle zu erhalten. Tatsächlich besteht ein gewisses Maß an Flexibilität hinsichtlich der Breite der Intervalle, die diese Bedingung erfüllen. Für k = 3 liegen drei Differenzen und drei einzelne Intervalle vor. Damit ist es auch in diesem Fall möglich, die Bedingung zu erfüllen, nun jedoch ohne die Flexibilität beim Festlegen der Breite der Intervalle. Für k = 4 liegen sechs Differenzen, jedoch nur vier einzelne Intervalle vor. Die Vergleichsintervalle müssen dieselben Informationen mit einer kleineren Anzahl von Intervallen vermitteln. Im Allgemeinen sind für k ≥ 4 mehr Differenzen als einzelne Mittelwerte vorhanden, so dass keine exakte Lösung gegeben ist, es sei denn, es werden zusätzliche Bedingungen für die Intervalle der Differenzen festgelegt, z. B. Gleichheit der Breiten.

Intervalle nach Tukey-Kramer weisen nur dann gleiche Breiten auf, wenn alle Stichprobenumfänge gleich sind. Die gleichen Breiten sind zudem eine Folge der Annahme gleicher Varianzen. Für Intervalle nach Games-Howell wird keine Gleichheit der Varianzen angenommen, daher weisen sie keine gleichen Breiten auf. Im Assistenten müssen wir uns auf Annäherungsmethoden stützen, um Vergleichsintervalle zu erhalten.

Das Intervall nach Games-Howell für $\mu_i - \mu_j$ lautet:

$$\bar{x}_i - \bar{x}_j \pm |q^*(k, \hat{v}_{ij})| \sqrt{s_i^2/n_i + s_j^2/n_j}$$

Hierbei ist $q^*(k, \hat{v}_{ij})$ das entsprechende Perzentil der Verteilung der studentisierten Spannweiten. Dieses hängt ab von k, der Anzahl der verglichenen Mittelwerte, sowie von ν_{ij}, den Freiheitsgraden für das Paar (i, j):

$$\hat{v}_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\left(\frac{s_i^2}{n_i}\right)^2 \frac{1}{n_i - 1} + \left(\frac{s_j^2}{n_j}\right)^2 \frac{1}{n_j - 1}}.$$

Hochberg, Weiss und Hart (1982) haben mit folgender Formel einzelne Intervalle berechnet, die annähernd Äquivalent zu diesen paarweisen Vergleichen sind:

$$\bar{x}_i \pm |q^*(k, v)| s_p X_i.$$

Die Werte X_i werden gewählt, um Folgendes zu minimieren:

$$\sum\sum_{i \neq j}\left(X_i + X_j - a_{ij}\right)^2,$$

Dabei gilt Folgendes:

$$a_{ij} = \sqrt{1/n_i + 1/n_j}\,.$$

Wir haben diesen Ansatz für den Fall bei ungleichen Varianzen übernommen, indem Intervalle aus Vergleichen nach Games-Howell der folgenden Form abgeleitet werden:

$$\bar{x}_i \pm d_i.$$

Die Werte d_i werden gewählt, um Folgendes zu minimieren:

$$\sum\sum_{i \neq j}\left(d_i + d_j - b_{ij}\right)^2,$$

Dabei gilt Folgendes:

$$b_{ij} = \left|q^*(k, \hat{v}_{ij})\right|\sqrt{s_i^2/n_i + s_j^2/n_j}\,.$$

Die Lösung lautet:

$$d_i = \frac{1}{k-1}\sum_{j \neq i} b_{ij} - \frac{1}{(k-1)(k-2)}\sum_{j \neq i, l \neq i, j<l} b_{jl}.$$

In den unten stehenden Diagrammen werden die Simulationsergebnisse für den Welch-Test mit den Ergebnissen für Vergleichsintervalle mit Hilfe von zwei Methoden verglichen: mit der Methode nach Games-Howell, die wir derzeit verwenden, sowie mit der Methode, die in Minitab Release 16 verwendet wird und auf der Berechnung des Durchschnitts von Freiheitsgraden beruht. Auf der vertikalen Achse wird dargestellt, wie häufig in 10.000 Simulationen die Nullhypothese vom Welch-Test fälschlicherweise zurückgewiesen wird bzw. sich nicht alle Vergleichsintervalle überlappen. Der Sollwert von Alpha in diesen Beispielen ist $\alpha = 0{,}05$. Diese Simulationen decken diverse Fälle ungleicher Standardabweichungen und Stichprobenumfänge ab; jede Position auf der horizontalen Achse stellt einen anderen Fall dar.

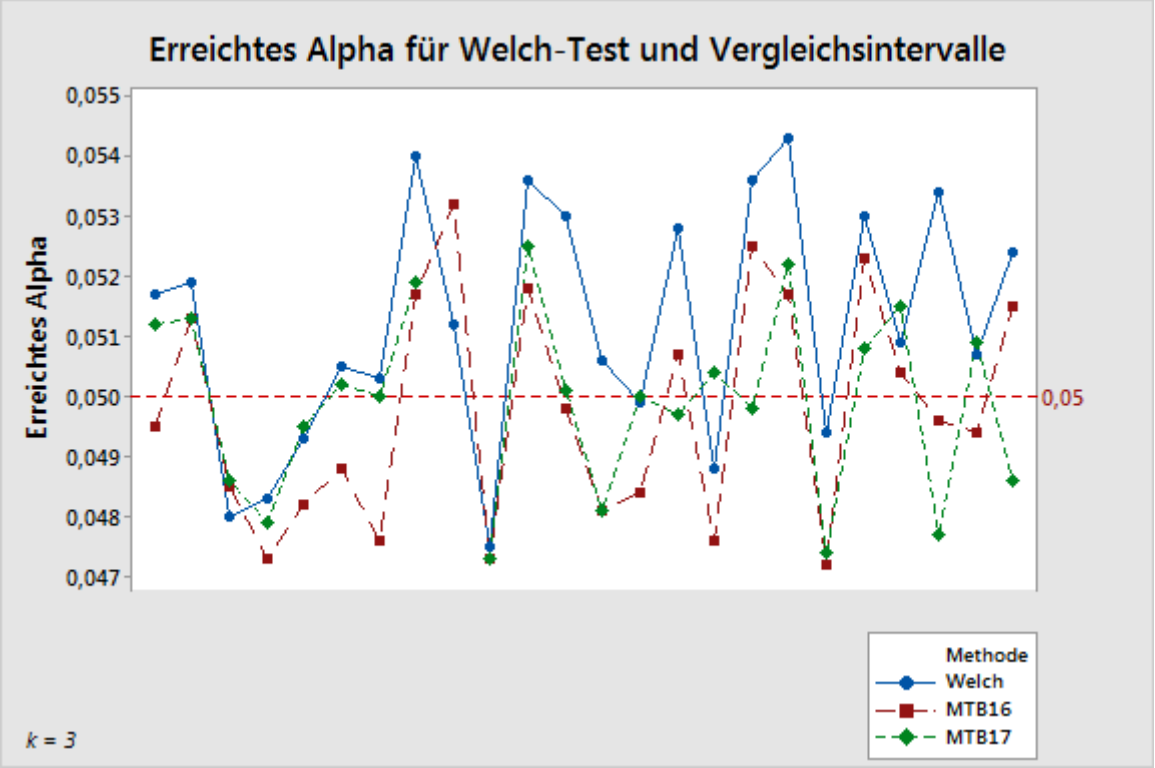


Abbildung 2 Welch-Test im Vergleich mit zwei Methoden zum Berechnen von Vergleichsintervallen für 3 Stichproben

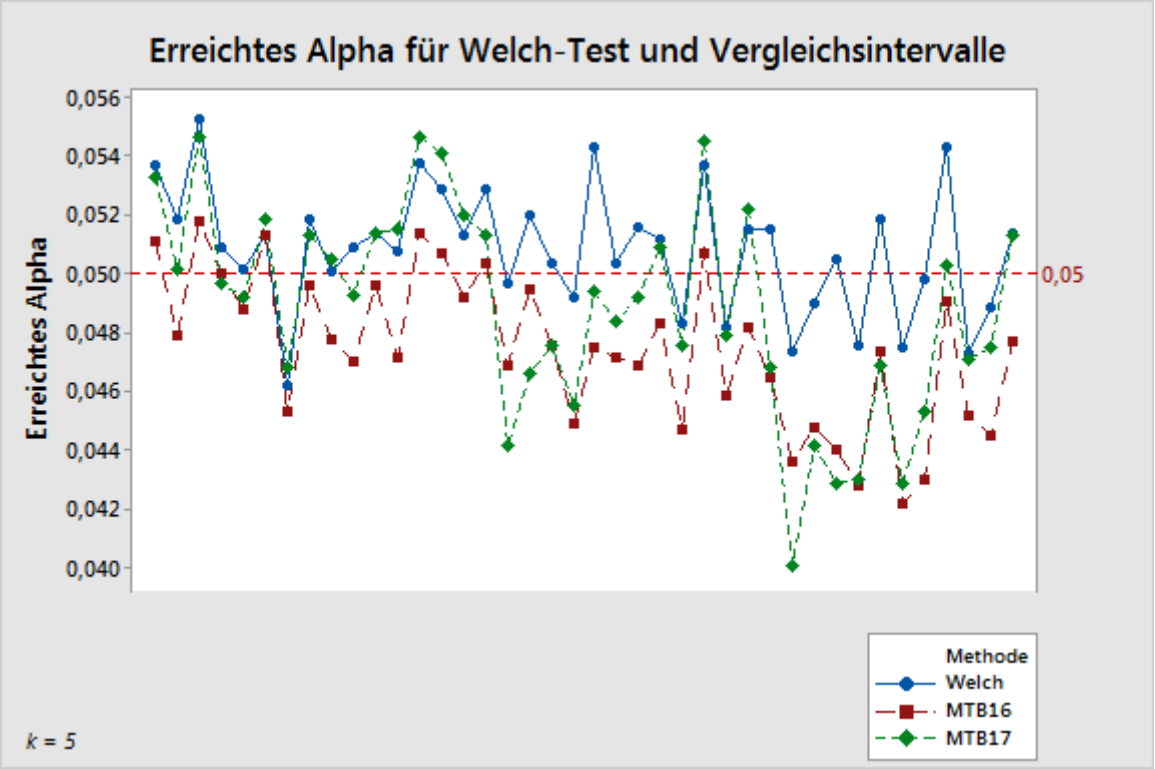


Abbildung 3 Welch-Test im Vergleich mit zwei Methoden zum Berechnen von Vergleichsintervallen für 5 Stichproben

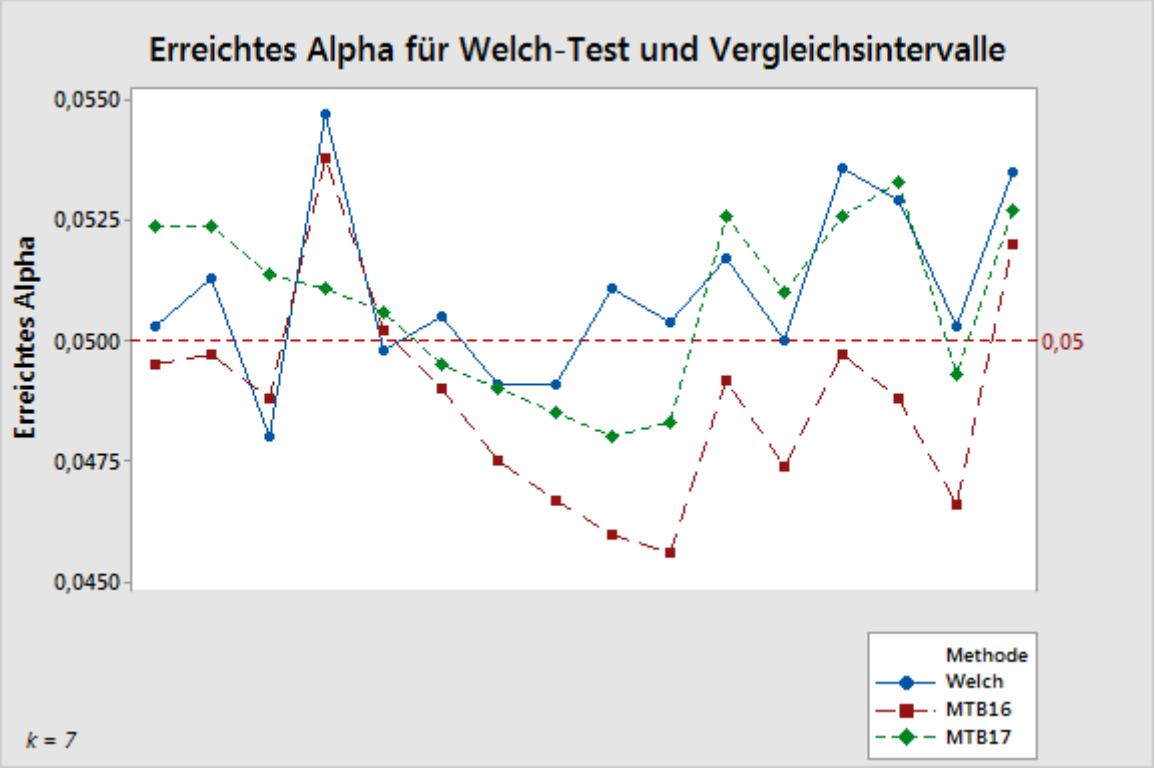


Abbildung 4 Welch-Test im Vergleich mit zwei Methoden zum Berechnen von Vergleichsintervallen für 7 Stichproben

Diese Ergebnisse zeigen simulierte Alpha-Werte in einem engen Bereich um den Sollwert von 0,05. Die mit der in Minitab Release 17 implementierten Games-Howell-Methode erzielten Ergebnisse kommen den Ergebnissen für den Welch-Test wohl näher als die Ergebnisse der Methode, die in Minitab Release 16 verwendet wurde.

Es liegen Anzeichen dafür vor, dass die Überdeckungswahrscheinlichkeit von Intervallen möglicherweise empfindlich gegenüber ungleichen Standardabweichungen ist. Diese Empfindlichkeit ist jedoch weit geringer als die des F-Tests. Im unten stehenden Diagramm wird diese Abhängigkeit für den Fall k = 5 veranschaulicht.

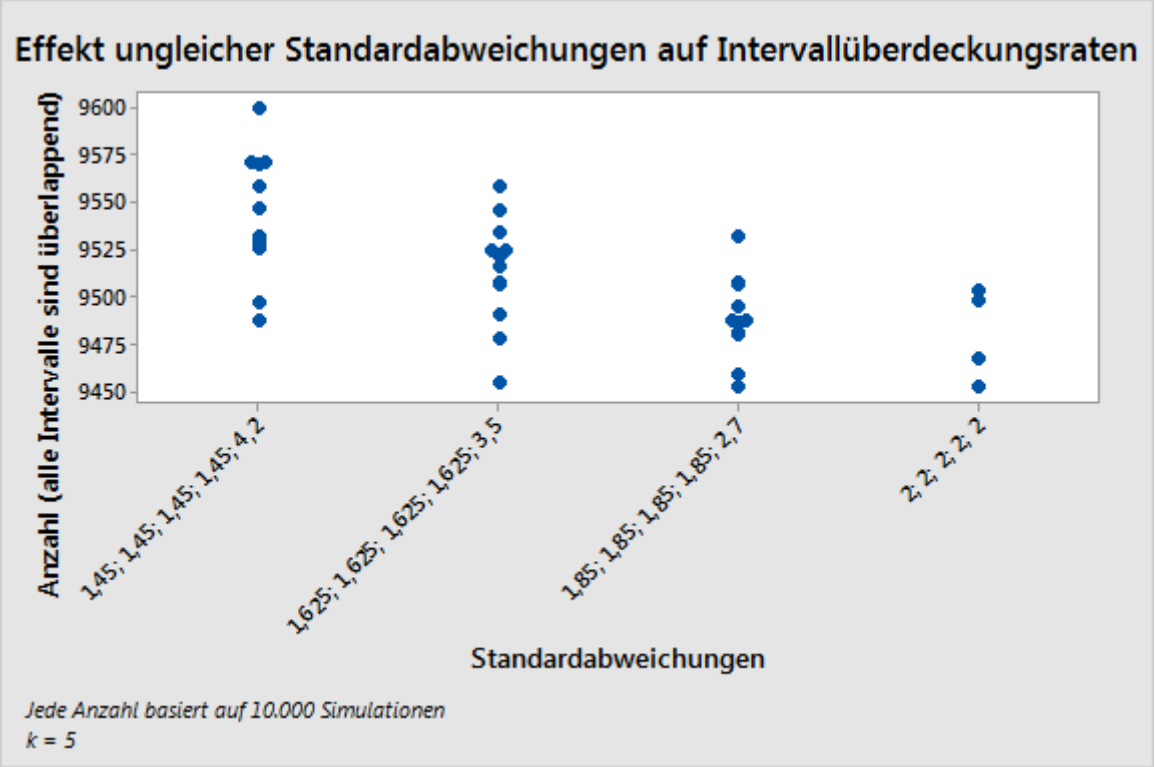


Abbildung 5 Ergebnisse der Simulation mit ungleichen Standardabweichungen

Kombinierte Verwendung des Hypothesentests und der Vergleichsintervalle

In seltenen Fällen kann es vorkommen, dass der Hypothesentest und der Vergleich einander beim Zurückweisen der Nullhypothese widersprechen. Der Test weist die Nullhypothese zurück, während bei sämtlichen Vergleichsintervallen Überlappungen zu verzeichnen sind. Umgekehrt kann es vorkommen, dass der Test die Nullhypothese nicht zurückweist, während Intervalle ohne Überlappung vorhanden sind. Diese Widersprüche treten selten auf, weil beide Methoden die Nullhypothese mit der gleichen Wahrscheinlichkeit zurückweisen, wenn diese tatsächlich wahr ist.

Ist dies jedoch der Fall, werden zunächst die Testergebnisse betrachtet, und bei einem signifikanten Test werden mit den Vergleichen weitere Untersuchungen durchgeführt. Wenn der Test beim Signifikanzniveau α die Nullhypothese zurückweist, wird jedes Vergleichsintervall rot gekennzeichnet, das nicht mit mindestens einem anderen Vergleichsintervall überlappt. Damit liegt ein visuelles Anzeichen darauf vor, dass der entsprechende Gruppenmittelwert sich von mindestens einem anderen unterscheidet. Selbst wenn alle Intervalle einander überlappen, wird bei einem signifikanten Test das Paar mit der geringsten Überlappung rot gekennzeichnet, um die „wahrscheinlichste“ Differenz

anzugeben (siehe Abbildung 6 unten). Diese Auswahl ist etwas willkürlich, insbesondere dann, wenn andere Paare mit sehr geringer Überlappung vorhanden sind. Es gibt jedoch kein anderes Paar, dessen Differenz näher an null liegt.

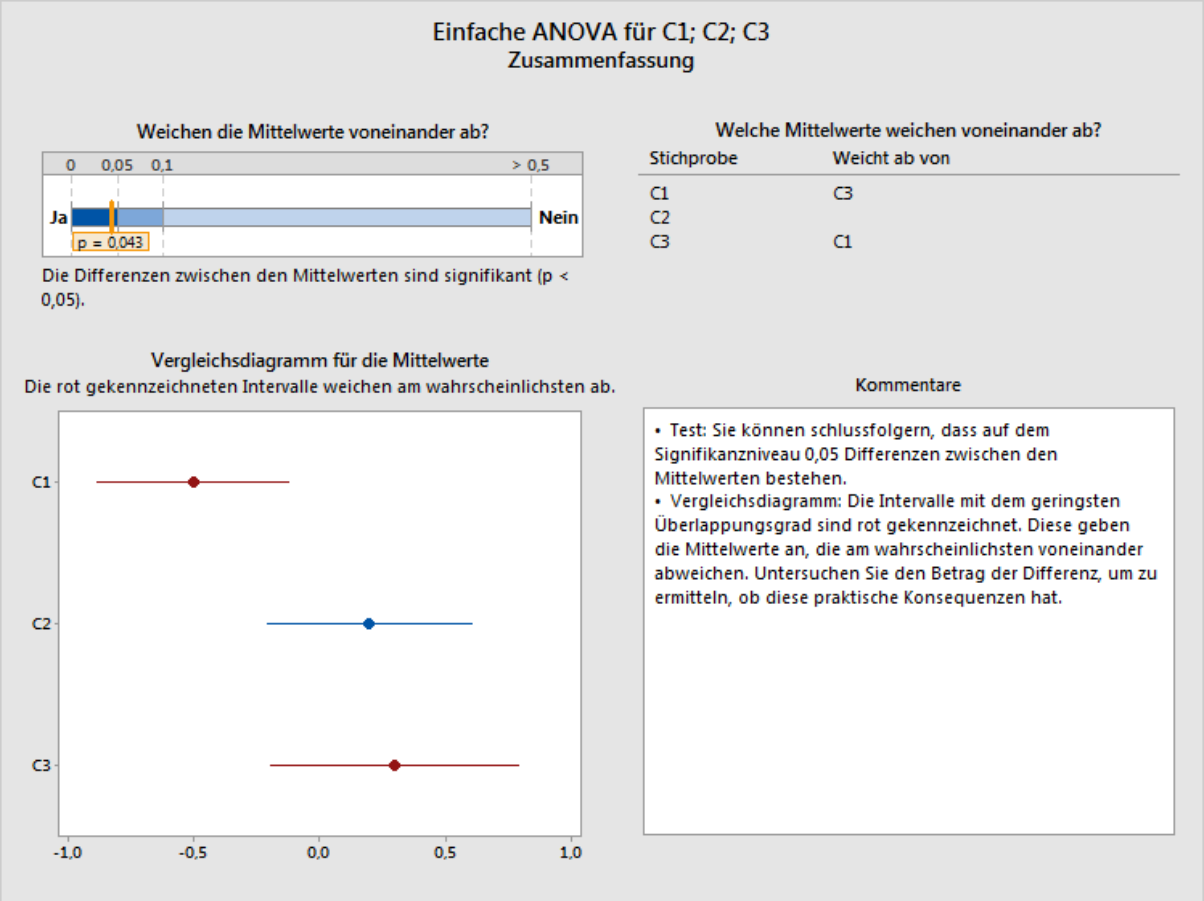


Abbildung 6 Signifikanter Test, Intervalle sind selbst bei einer Überlappung zwischen Stichproben rot gekennzeichnet

Wenn der Test die Nullhypothese nicht zurückweist, wird keines der Intervalle rot gekennzeichnet, selbst wenn Intervalle vorhanden sind, die sich nicht überlappen (siehe Abbildung 7 unten). Obwohl die Intervalle implizieren, dass Differenzen zwischen den Mittelwerten vorhanden sind, ist zu beachten, dass eine fehlende Zurückweisung der Nullhypothese nicht gleichbedeutend mit der Schlussfolgerung ist, dass die Nullhypothese wahr ist. Hiermit wird lediglich angegeben, dass die beobachteten Differenzen nicht groß genug sind, um eine zufällige Ursache auszuschließen. Es ist außerdem anzumerken, dass der Abstand zwischen den nicht überlappenden Intervallen in einer derartigen Situation typischerweise sehr klein ist. Daher passen äußerst kleine Differenzen immer noch zu den Intervallen und verweisen nicht zwangsläufig darauf, dass eine Differenz mit praktischen Konsequenzen vorliegt.

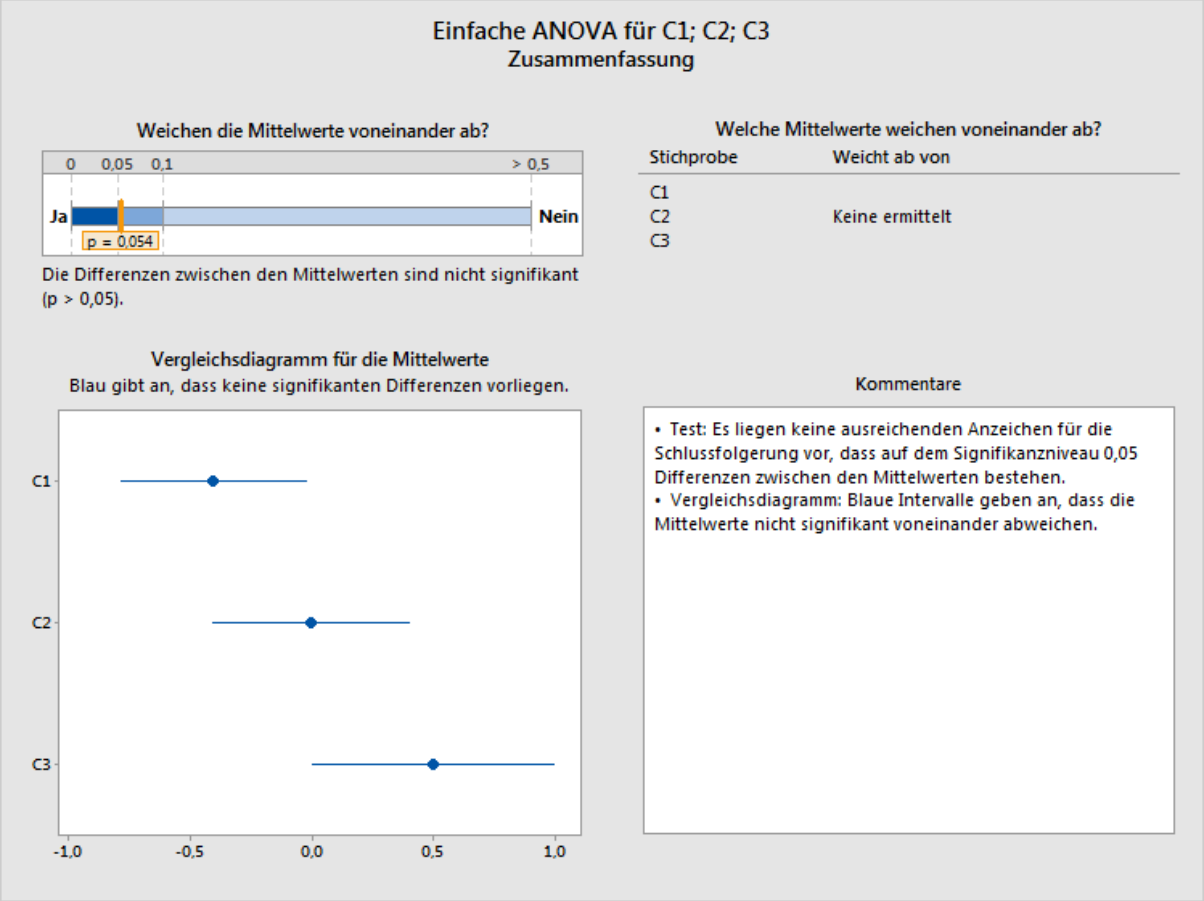


Abbildung 7 Test schlägt fehl, keine Intervalle sind rot gekennzeichnet, selbst wenn keine Überlappungen zwischen den Stichproben vorliegen

Anhang C: Stichprobenumfang

In der einfachen ANOVA werden als Parameter die Grundgesamtheits-Mittelwerte μ_1, μ_2, … μ_k der verschiedenen Gruppen bzw. Grundgesamtheiten getestet. Die Parameter erfüllen die Nullhypothese, wenn sie alle gleich sind. Wenn Differenzen zwischen den Mittelwerten vorliegen, erfüllen sie die Alternativhypothese. Die Wahrscheinlichkeit der Zurückweisung der Nullhypothese darf für Mittelwerte, die die Nullhypothese erfüllen, nicht größer als α sein. Die tatsächlichen Wahrscheinlichkeiten hängen von der Standardabweichung der Verteilungen sowie vom Umfang der Stichproben ab. Die Trennschärfe für die Erkennung von Abweichungen von der Nullhypothese erhöht sich mit kleineren Standardabweichungen oder größeren Stichproben.

Die Trennschärfe des F-Tests unter der Annahme von Normalverteilungen mit gleichen Standardabweichungen kann mit einer nicht zentralen F-Verteilung berechnet werden. Der Nichtzentralitätsparameter lautet:

$$\theta_F = \sum_{i=1}^{k} n_i \, (\mu_i - \mu)^2 / \sigma^2$$

Hierbei ist μ der gewichtete Durchschnitt der Mittelwerte:

$$\mu \; = \; \sum_{i=1}^{k} n_i \mu_i \, / \sum_{i=1}^{k} n_i \; ,$$

und σ ist die Standardabweichung, die als konstant angenommen wird. Bei sonst gleichen Bedingungen wächst die Trennschärfe mit θ_F. Genau in diesem Sinn nimmt die Trennschärfe zu, wenn die Mittelwerte weiter von der Nullhypothese abweichen.

Im Unterschied zum F-Test gibt es beim Welch-Test keine einfache exakte Formel für die Trennschärfe. Betrachten wir im Folgenden jedoch zwei hinreichend gute näherungsweise Formeln. In der ersten wird eine nicht zentrale F-Verteilung ähnlich wie bei der Berechnung der Trennschärfe für den F-Test verwendet. Der verwendete Nichtzentralitätsparameter weist weiterhin die folgende Form auf:

$$\theta_W \; = \; \sum_{i=1}^{k} w_i (\mu_i - \mu)^2$$

Hierbei ist μ der gewichtete Durchschnitt:

$$\mu \; = \sum_{i=1}^{k} w_i \mu_i / \sum_{j=1}^{k} w_j$$

Die Gewichtungen hängen jedoch sowohl von den Standardabweichungen als auch von den Stichprobenumfängen ab, d. h. $w_i = \; n_i / \sigma_i^2$ oder $w_i = \; n_i / s_i^2$, je nachdem, ob die Ergebnisse für bekannte Standardabweichungen σ_i^2 simuliert werden oder die Trennschärfe auf der Grundlage der Standardabweichungen der Stichproben s_i^2 geschätzt wird. Die ungefähre Trennschärfe wird dann wie folgt berechnet:

$$P\left(F_{k-1,f,\theta_w} \geq \; F_{k-1,f,1-\alpha}\right)$$

Hierbei sind die Freiheitsgrade des Nenners:

$$f \; = \; \frac{k^2 - 1}{3 \sum_{i=1}^{k} \left(1 - w_i / \sum_{j=1}^{k} w_j\right) / (n_i - 1)}.$$

Wie wir unten verdeutlichen, erhalten wir damit hinreichend gute Annäherungen an die in den Simulationen beobachtete Trennschärfe. Im Menü „Assistent“ wird zwar eine andere

Annäherung zum Berechnen der Trennschärfe verwendet, die vorliegende liefert jedoch gute Erkenntnisse und ist die Grundlage für die Auswahl der Konfiguration der Mittelwerte, bei der die Trennschärfe im Menü „Assistent“ berechnet wird.

Konfiguration von Mittelwerten

Wie bei dem Ansatz zur Berechnung der Trennschärfe und des Stichprobenumfangs in Minitab (**Statistik** > **Varianzanalyse (ANOVA)** > **Einfache ANOVA**) werden Benutzer vom Assistenten nicht aufgefordert, einen kompletten Satz von Mittelwerten anzugeben, bei denen die Trennschärfe auswertet werden soll. Stattdessen müssen Benutzer eine Differenz zwischen den Mittelwerten angeben, die praktische Konsequenzen hat. Für jede angegebene Differenz gibt es eine unendliche Anzahl möglicher Konfigurationen von Mittelwerten, bei denen sich der kleinste und der größte Mittelwert um den betreffenden Betrag unterscheiden. In jedem der folgenden Fälle beispielsweise gibt es eine maximale Differenz von 10 zwischen einer Gruppe von fünf Mittelwerten:

$\mu_1 = 0; \mu_2 = 5; \mu_3 = 5; \mu_4 = 5; \mu_5 = 10;$

$\mu_1 = 5; \mu_2 = 0; \mu_3 = 10; \mu_4 = 10; \mu_5 = 0;$

$\mu_1 = 0; \mu_2 = 10; \mu_3 = 0; \mu_4 = 0; \mu_5 = 0;$

Darüber hinaus gibt es unendlich viele weitere.

Es wird derselbe Ansatz wie bei der Berechnung von Trennschärfe und Stichprobenumfang in Minitab (**Statistik > Trennschärfe und Stichprobenumfang > Einfache ANOVA**) verfolgt. Dabei wird ein Fall ausgewählt, bei dem alle mit Ausnahme von zwei Mittelwerten beim (gewichteten) Durchschnitt der Mittelwerte liegen und sich die übrigen zwei Mittelwerte um den angegebenen Betrag unterscheiden. Wegen der Möglichkeit ungleicher Varianzen und Stichprobenumfänge hängt der Nichtzentralitätsparameter (und damit die Trennschärfe) jedoch immer noch davon ab, für welche zwei Mittelwerte eine Differenz angenommen wird.

Betrachten Sie die Konfiguration von Mittelwerten $\mu_1, \ldots, \mu_k$, in der alle Mittelwerte mit Ausnahme von zwei Mittelwerten dem gewichteten Gesamtmittelwert μ entsprechen und zwei Mittelwerte (z. B. $\mu_i > \mu_j$) sich sowohl voneinander als auch vom Gesamtmittelwert unterscheiden. Sei $\Delta = \mu_i - \mu_j$ die Differenz zwischen den beiden Mittelwerten. Sei $\Delta_i = \mu_i - \mu$ und $\Delta_j = \mu - \mu_j$. Damit ist $\Delta = \Delta_i + \Delta_j$. Da μ den gewichteten Mittelwert aller k Mittelwerte darstellt und für (k – 2) Mittelwerte angenommen wird, dass sie gleich μ sind, gilt Folgendes:

$$\mu = \left[\sum_{l \neq i,j} w_l\mu_l + w_i(\mu + \Delta_i) + w_j(\mu - \Delta_j)\right] \Big/ \sum_{l=1}^{k} w_l = \mu + \left(w_i\Delta_i - w_j\Delta_j\right) \Big/ \sum_{l=1}^{k} w_l \,.$$

Somit gilt:

$$w_i\Delta_i = w_j\Delta_j = w_j(\Delta - \Delta_i) \,,$$

Daher gilt:

$$\Delta_i = \frac{w_j}{w_i + w_j}\Delta$$

$$\Delta_j = \frac{w_i}{w_i + w_j}\Delta$$

Für diese spezifische Konfiguration von Mittelwerten kann der Nichtzentralitätsparameter in Bezug auf den Welch-Test berechnet werden:

$$\theta_W = w_i(\mu_i - \mu)^2 + w_j(\mu_j - \mu)^2$$
$$= \frac{w_i w_j^2 \Delta^2 + w_j w_i^2 \Delta^2}{(w_i + w_j)^2} = \frac{w_i w_j \Delta^2}{w_i + w_j}$$

Dieser Betrag erhöht sich durch w_i für ein festes w_j und umgekehrt. Daher wird er beim Paar (i, j) mit den beiden größten Gewichtungen maximiert und beim Paar mit den beiden kleinsten Gewichtungen minimiert. In sämtlichen Trennschärfeberechnungen werden diese beiden Extremfälle betrachtet, welche die Trennschärfe unter der Annahme, dass genau zwei Mittelwerte vom gewichteten Gesamtdurchschnitt der Mittelwerte abweichen, minimieren bzw. maximieren.

Wenn Sie eine Differenz für den Test angeben, werden der minimale und der maximale Trennschärfenwert für diese Differenz ermittelt. Der Bereich dieser Trennschärfen wird in den Berichten auf einer farblich kodierten Leiste angegeben, auf der Trennschärfen unter 60 % rot, Trennschärfen ab 90 % grün und Trennschärfen zwischen 60 % und 90 % gelb gekennzeichnet sind. Die Ergebnisse in der Auswertung hängen davon ab, in welchen Bereich die Trennschärfen in Bezug auf diese farblich kodierte Skala fallen. Wenn der gesamte Bereich im roten Abschnitt liegt, ist die Trennschärfe für jedes Paar von Gruppen kleiner oder gleich 60 %, und das rote Symbol in der Auswertung zeigt ein Problem aufgrund unzureichender Trennschärfe an. Befindet sich der gesamte Bereich im grünen Abschnitt, beträgt die Trennschärfe für jede Gruppe mindestens 90 %, und das grüne Symbol in der Auswertung gibt an, dass die Trennschärfe ausreichend ist. Alle sonstigen Bedingungen werden als Zwischenstufe behandelt, was durch ein gelbes Symbol in der Auswertung angegeben wird.

In Fällen, in denen die grüne Bedingung nicht erfüllt ist, berechnet der Assistent einen Stichprobenumfang, mit dem bei der vom Benutzer angegebenen Differenz und den beobachteten Standardabweichungen der Stichproben die grüne Bedingung erreicht werden kann. Die geschätzte Trennschärfe hängt über die Gewichtungen $w_i = n_i/s_i^2$. von den Stichprobenumfängen ab. Wenn für alle Stichproben ein gleicher Stichprobenumfang angenommen wird, entsprechen die beiden kleinsten Gewichtungen den beiden Gruppen mit den größten Standardabweichungen der Stichproben. Der Assistent bestimmt einen Stichprobenumfang, bei dem eine Trennschärfe von mindestens 90 % erreicht wird, sofern die angegebene Differenz zwischen den beiden Gruppen mit der größten Streuung vorliegt. Wenn also ein Stichprobenumfang von mindestens dieser Größe für alle Gruppen festgesetzt wird, liegt der komplette Bereich der Trennschärfewerte sämtlicher Gruppen mindestens bei 90 %, womit die grüne Bedingung erfüllt ist.

Wenn der Benutzer keine Differenz für die Trennschärfeberechnung angibt, bestimmt der Assistent die größte Differenz, bei der das Maximum des Bereichs der berechneten Trennschärfen bei 60 % liegt. Dieser Wert wird an der Grenze zwischen dem roten und dem gelben Abschnitt der Leiste beschriftet und entspricht einer Trennschärfe von 60 %. Außerdem wird die kleinste Differenz bestimmt, bei der das Minimum des Bereichs der berechneten Trennschärfen 90 % beträgt. Dieser Wert wird an der Grenze zwischen dem

gelben und dem grünen Abschnitt der Leiste beschriftet und entspricht einer Trennschärfe von 90 %.

Trennschärfeberechnung

Trennschärfeberechnung erfolgt auf der Grundlage der Approximation nach Kulinskaya et al. (2003):

Es werden die folgenden Größen definiert:

$\lambda = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2$,

$A = \sum_{i=1}^{k} h_i$,

$B = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2(1 - w_i/W)/(n_i - 1)$,

$D = \sum_{i=1}^{k} w_i^2(\mu_i - \mu)^4/(n_i - 1)$,

$E = \sum_{i=1}^{k} w_i^3(\mu_i - \mu)^6/(n_i - 1)^2$.

Die ersten drei Kumulanten des Zählers $\sum_{i=1}^{k} w_i(\bar{x}_i - \hat{\mu})^2$ der Welch-Statistik können geschätzt werden als:

$\kappa_1 = k - 1 + \lambda + 2A + 2B$,

$\kappa_2 = 2(k - 1 + 2\lambda + 7A + 14B + D)$,

$\kappa_3 = 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E)$.

Sei $F_{k-1,\,f,\,1-\alpha}$ das Quantil $(1 - \alpha)$ der Verteilung F(k – 1; f). Wie bereits ausgeführt, ist $W^* \geq F_{k-1,\,f,\,1-\alpha}$ das Kriterium für die Zurückweisung der Nullhypothese in einem Welch-Test des Umfangs α.

Seien

$q = (k - 1)\left[1 + \frac{2(k-2)A}{k^2 - 1}\right] F_{k-1,f,1-\alpha}$,

$b = \kappa_1 - 2\kappa_2^2/\kappa_3$,

$c = \kappa_3/(4\kappa_2)$ [Hinweis: Der Ausdruck für c ist in Kulinskaya et al. (2003) ohne Klammern angegeben.]

$\nu = 8\kappa_2^3/\kappa_3^2$.

Die geschätzte approximierte Trennschärfe des Welch-Tests ist dann:

$$P(\chi_\nu^2 \geq \frac{q - b}{c})$$

Hierbei ist χ_ν^2 eine Zufallsvariable mit Chi-Quadrat-Verteilung und ν Freiheitsgraden.

In den folgenden Ergebnissen wird die Trennschärfe für die beiden Approximationsmethoden mit der simulierten Trennschärfe für einen Bereich von Beispielen verglichen; Grundlage sind 10.000 Simulationen.

Tabelle 3 Trennschärfeberechnungen für die beiden Approximationsmethoden im Vergleich mit der simulierten Trennschärfe

Beispiel	Alpha	Simulierte Trennschärfe	Nicht zentrale F-Verteilung	Kulinskaya et al.
μ: 0; 0; 0; -0,1724; 0,8276 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,1372 0,0739 0,0195	0,135702 0,072563 0,016587	0,135795 0,069512 0,012538
μ: 0; 0; 0; -0,3448; 1,6552 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,2498 0,1574 0,0541	0,251064 0,153128 0,045211	0,257455 0,156215 0,042195
μ: 0; 0; 0; -0,5172; 2,4828 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,4534 0,3211 0,1273	0,445570 0,311994 0,121225	0,453506 0,321575 0,125065
μ: 0; 0; 0; -0,6896; 3,3104 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,6620 0,5219 0,2842	0,671317 0,533819 0,271316	0,670296 0,538617 0,282759
μ: 0; 0; 0; -0,8620; 4,1380 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,8417 0,7382 0,4883	0,852589 0,752173 0,487601	0,846697 0,746121 0,49323
μ: 0; 0; 0; -1,0344; 4,9656 σ: 2; 2; 2; 2; 4 n: 12; 12; 12; 12; 10	0,10 0,05 0,01	0,9429 0,8866 0,6910	0,952077 0,901485 0,711055	0,954929 0,897937 0,703379
μ: 0; 0; 0; 0; 0; -0,148148; 1,85185 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,2011 0,1201 0,0385	0,189392 0,108986 0,028986	0,200114 0,117420 0,031456
μ: 0; 0; 0; 0; 0; -0,296296; 3,70370 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,4942 0,3677 0,1770	0,485917 0,351593 0,149041	0,500143 0,375296 0,177189
μ: 0; 0; 0; 0; 0; -0,444444; 5,55556 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,8125 0,7131 0,4876	0,829702 0,727384 0,474291	0,819542 0,720807 0,49469

Beispiel	Alpha	Simulierte Trennschärfe	Nicht zentrale F-Verteilung	Kulinskaya et al.
μ: 0; 0; 0; 0; 0; -0,592593; 7,40741 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,9645 0,9286 0,7938	0,977211 0,949997 0,831174	0,984213 0,949239 0,814067
μ: 0; 0; 0; 0; 0; -0,740741; 9,25926 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,9961 0,9895 0,9528	0,998947 0,996653 0,977536	1,00000 1,00000 0,98705
μ: 0; 0; 0; 0; 0; -0,888889; 11,1111 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,9999 0,9995 0,9943	0,999985 0,999926 0,998910	1,00000 1,00000 1,00000
μ: 0; 0; 0; 0; 0; -0,518519; 6,48148 σ: 2; 2; 2; 2; 2; 2; 5 n: 20; 20; 20; 20; 20; 20; 10	0,10 0,05 0,01	0,9059 0,8403 0,6511	0,929392 0,868721 0,67121	0,924696 0,856720 0,666520
μ: 0; 0; 0; 0; 0; -0,5; 0,5 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,1870 0,1098 0,0315	0,186658 0,106600 0,027773	0,183290 0,100189 0,021332
μ: 0; 0; 0; 0; 0; -1; 1 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,4734 0,3394 0,1378	0,474736 0,338655 0,137788	0,472469 0,334430 0,128693
μ: 0; 0; 0; 0; 0; -1,5; 1,5 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,8228 0,7112 0,4391	0,817355 0,707319 0,441154	0,810181 0,698461 0,431868
μ: 0; 0; 0; 0; 0; -2; 2 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,9691 0,9312 0,7817	0,973246 0,940585 0,799339	0,973319 0,936546 0,785099
μ: 0; 0; 0; 0; 0; -2,5; 2,5 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,9984 0,9936 0,9587	0,998579 0,995330 0,967674	0,999763 0,997481 0,966249
μ: 0; 0; 0; 0; 0; -3; 3 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	1,0000 0,9997 0,9959	0,999975 0,999870 0,997927	1,00000 1,00000 0,99961
μ: 0; 0; 0; 0; 0; -3,5; 3,5 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	1,00000 1,00000 0,99998	1,00000 1,00000 0,99995	1,00000 1,00000 1,00000

Beispiel	Alpha	Simulierte Trennschärfe	Nicht zentrale F-Verteilung	Kulinskaya et al.
μ: 0; 0; 0; 0; 0; -1,75; 1,75 σ: 2; 2; 2; 2; 2; 2; 2 n: 12; 12; 12; 12; 12; 12; 12	0,10 0,05 0,01	0,9140 0,8418 0,6190	0,921225 0,852755 0,633815	0,916652 0,843856 0,620704
μ: 0; -0,5; 0,5 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,2548 0,1549 0,0470	0,259249 0,160861 0,049045	0,257149 0,156251 0,042292
μ: 0; -1; 1 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,6540 0,5205 0,2612	0,659073 0,522885 0,263550	0,654105 0,515816 0,252469
μ: 0; -1,5; 1,5 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,9364 0,8747 0,6614	0,935939 0,875620 0,664478	0,937768 0,872608 0,652563
μ: 0; -1,75; 1,75 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,9810 0,9522 0,8251	0,981434 0,956100 0,830726	0,986815 0,959796 0,823624
μ: 0; -2; 2 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,9953 0,9878 0,9308	0,995969 0,988175 0,931922	0,999332 0,993705 0,933446
μ: 0; -2,5; 2,5 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	0,9999 0,9997 0,9949	0,999923 0,999634 0,994725	1,00000 1,00000 0,99909
μ: 0; -3; 3 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	1,0000 1,0000 0,9999	1,00000 1,00000 0,99985	1,00000 1,00000 1,00000
μ: 0; -3,5; 3,5 σ: 2; 2; 2 n: 12; 12; 12	0,10 0,05 0,01	1,0000 1,0000 0,9999	1,00000 1,00000 1,00000	1,00000 1,00000 1,00000
μ: 0; -0,142857; 0,857143 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,1452 0,0790 0,0223	0,143156 0,077699 0,018200	0,146824 0,077538 0,014338
μ: 0; -0,285714; 1,71429 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,2765 0,1787 0,0624	0,274240 0,170628 0,051588	0,286222 0,179469 0,050335

Beispiel	Alpha	Simulierte Trennschärfe	Nicht zentrale F-Verteilung	Kulinskaya et al.
μ: 0; -0,428571; 2,57143 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,4861 0,3487 0,1467	0,476925 0,338626 0,132405	0,490018 0,355743 0,141352
μ: 0; -0,50000; 3 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,5846 0,4425 0,2107	0,588533 0,444491 0,197290	0,596795 0,460707 0,212798
μ: 0; -0,571429; 3,42857 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,6933 0,5631 0,3052	0,694684 0,555731 0,279131	0,696773 0,567129 0,299302
μ: 0; -0,714286; 4,28571 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,8480 0,7402 0,4871	0,861469 0,759703 0,480052	0,859329 0,759762 0,497421
μ: 0; -0,857143; 5,14286 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,9434 0,8869 0,6649	0,952562 0,898817 0,687058	0,961913 0,902716 0,692591
μ: 0; -1; 6 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,9849 0,9609 0,8294	0,987981 0,967589 0,847436	0,999989 0,985049 0,853787
μ: 0; -1,14286; 6,85714 σ: 2; 2; 4 n: 14; 12; 8	0,10 0,05 0,01	0,9976 0,9890 0,9222	0,997776 0,992220 0,940972	1,00000 1,00000 0,96383
μ: 1; 2; 3 σ: 0,3; 2,4; 3,6 n: 13; 19; 25	0,10 0,05 0,01	0,8838 0,7995 0,5632	0,882194 0,797869 0,556486	0,884649 0,802137 0,563208
μ: 1; 2; 3 σ: 2,77489; 2,77489; 2,77489 n: 13; 19; 25	0,10 0,05 0,01	0,5649 0,4305 0,1994	0,566831 0,431302 0,201329	0,565141 0,428126 0,195734

Die obigen Ergebnisse sind im unten stehenden Diagramm zusammengefasst, das die Unterschiede zwischen den einzelnen Approximationen und die in der Simulation geschätzten Trennschärfewerte darstellt.

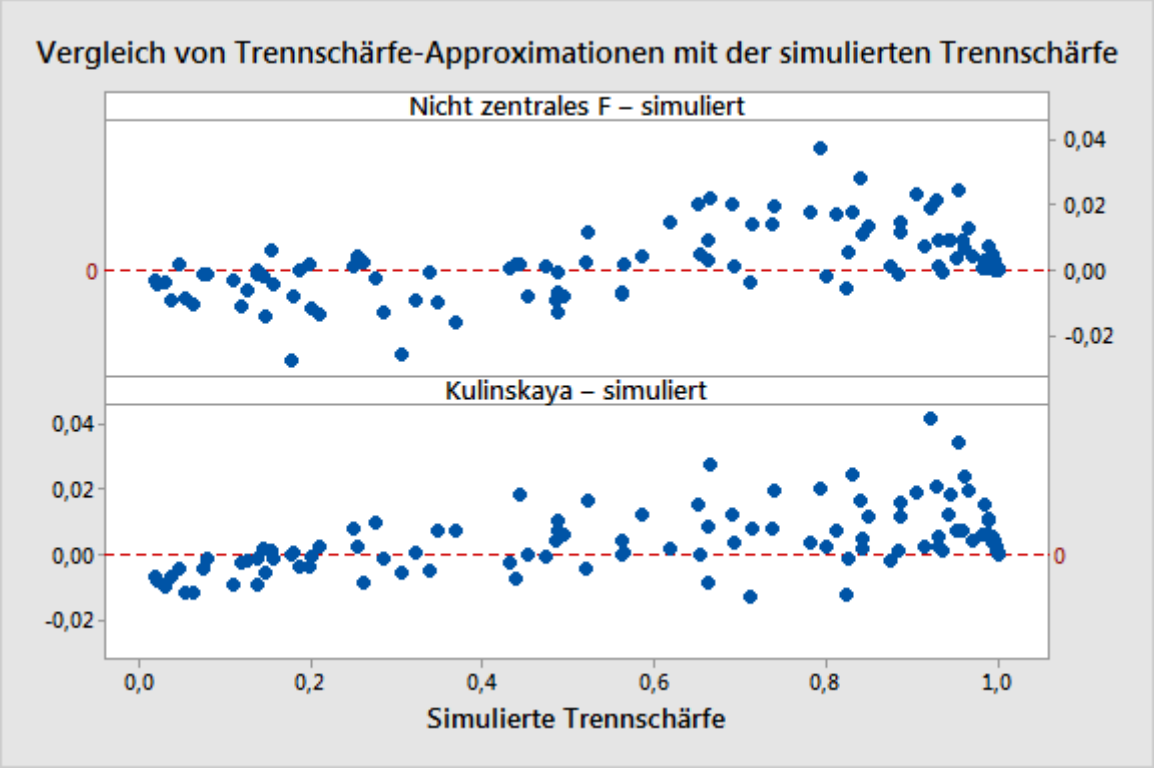


Abbildung 8 Vergleich der beiden Trennschärfe-Approximationen und der durch Simulation geschätzten Trennschärfe

Anhang D: Vorliegen einer Normalverteilung

In diesem Abschnitt werden die Simulationen erläutert, mit denen die Leistung des Welch-Tests und der Vergleichsintervalle bei kleinen bis mittleren Stichproben aus verschiedenen Nicht-Normalverteilungen untersucht wurde.

In den unten stehenden Tabellen sind die Simulationsergebnisse für unterschiedliche Arten von Verteilungen unter der Nullhypothese gleicher Mittelwerte zusammengefasst. Für diese Beispiele sind außerdem alle Standardabweichungen gleich, und sämtliche Stichproben weisen den gleichen Stichprobenumfang auf. Die Anzahl der Stichproben ist k = 3, 5 oder 7.

In jeder Zelle wird der Schätzwert der Wahrscheinlichkeit eines Fehlers 1. Art aus 10.000 Simulationen angezeigt. Das Soll-Signifikanzniveau (Soll-α) ist 0,05.

Tabelle 4 Simulationsergebnisse des Welch-Tests mit gleichen Mittelwerten für unterschiedliche Verteilungen

	Stichprobenumfang n = 10			Stichprobenumfang n = 15		
Verteilung	**k = 3**	**k = 5**	**k = 7**	**k = 3**	**k = 5**	**k = 7**
N(0;1)	0,0490	0,0486	0,0512	0,0534	0,0522	0,0550
T(3)	0,0371	0,0361	0,0348	0,0353	0,0385	0,0365
T(5)	0,0440	0,0425	0,0439	0,0435	0,0428	0,0428
Laplace(0;1)	0,0433	0,0354	0,0345	0,0445	0,0397	0,0407
Gleichverteilung(-1; 1)	0,0544	0,0640	0,0718	0,0517	0,0573	0,0585
Beta(3; 3)	0,0504	0,0577	0,0622	0,0501	0,0538	0,0564
Exponential	0,0508	0,0621	0,0748	0,0483	0,0633	0,0779
Chi-Quadrat(3)	0,0473	0,0579	0,0753	0,0499	0,0588	0,0703
Chi-Quadrat(5)	0,0458	0,0594	0,0643	0,0504	0,0606	0,0679
Chi-Quadrat(10)	0,0463	0,0510	0,0585	0,0463	0,0552	0,0567
Beta(8; 1)	0,0500	0,0622	0,0775	0,0549	0,0653	0,0760

Die Wahrscheinlichkeiten eines Fehlers 1. Art liegen alle innerhalb von 3 Prozentpunkten des Soll-α, selbst für Stichproben mit dem Umfang 10. Größere Abweichungen treten tendenziell bei einer größeren Anzahl von Gruppen und bei Verteilungen auf, die weit von der Normalverteilung entfernt sind. Bei Stichprobenumfängen von 10 liegen die einzigen Fälle, bei denen die Annahmewahrscheinlichkeit um mehr als 2 Prozentpunkte abweicht, bei k = 7 vor. Diese treten für die Gleichverteilung auf, die über viel kürzere Randbereiche als die Normalverteilung verfügt, sowie für die Exponentialverteilung, die Chi-Quadrat-Verteilung (3)

und die Beta-Verteilung (8; 1), die allesamt stark schief sind. Wenn die Stichprobenumfänge auf 15 vergrößert werden, verbessern sich die Ergebnisse für die Gleichverteilung merklich, jedoch nicht für die zwei stark schiefen Verteilungen.

Eine ähnliche Simulation wurde für Vergleichsintervalle durchgeführt. Das simulierte α in diesem Fall ist die Anzahl der Simulationen aus 10.000 Simulationen, in denen sich einige Intervalle nicht überlappen. Das Soll-$\alpha = 0{,}05$.

Tabelle 5 Simulationsergebnisse der Vergleichsintervalle mit gleichen Mittelwerten für unterschiedliche Verteilungen

	Stichprobenumfang n = 10			Stichprobenumfang n = 15		
Verteilung	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0;1)	0,0493	0,0494	0,0469	0,0538	0,0518	0,0561
t(3)	0,0378	0,0321	0,0254	0,0347	0,0343	0,0289
t(5)	0,0449	0,0399	0,0361	0,0447	0,0444	0,0412
Laplace(0;1)	0,0438	0,0305	0,0246	0,0456	0,0366	0,0348
Gleichverteilung(-1; 1)	0,0559	0,0605	0,0699	0,0534	0,0607	0,0590
Beta(3; 3)	0,0515	0,0569	0,0615	0,0510	0,0553	0,0568
Exponential	0,0353	0,0254	0,0207	0,0346	0,0310	0,0275
Chi-Quadrat(3)	0,0375	0,0305	0,0296	0,0384	0,0359	0,0339
Chi-Quadrat(5)	0,0405	0,0390	0,0353	0,0417	0,0433	0,0416
Chi-Quadrat(10)	0,0425	0,0428	0,0447	0,0435	0,0476	0,0464
Beta(8; 1)	0,0381	0,0352	0,0287	0,0459	0,0428	0,0403

Wie beim Welch-Test liegen die Wahrscheinlichkeiten eines Fehlers 1. Art durchgehend innerhalb von 3 Prozentpunkten vom Soll-α, selbst bei Stichproben mit dem Umfang 10. Größere Abweichungen treten tendenziell bei einer größeren Anzahl von Stichproben und bei Verteilungen auf, die weit von der Normalverteilung entfernt sind. Bei Stichprobenumfängen von 10 liegen die Fehlerwahrscheinlichkeiten gelegentlich für k = 7 (und in einem Fall für k = 5) um mehr als 2 Prozentpunkte entfernt. Diese Fälle treten für die t-Verteilung mit 3 Freiheitsgraden und extrem stark besetzten Randbereichen, die Laplace-Verteilung sowie die stark schiefe Exponentialverteilung und die ebenfalls stark schiefe Chi-Quadrat-Verteilung (3) auf. Wenn die Stichprobenumfänge auf 15 vergrößert werden, verbessert dies die Ergebnisse, und es verbleiben nur noch die t-Verteilung (3) und die Exponentialverteilung mit simulierten Werten von α, die um mehr als 2 Prozentpunkte vom Sollwert entfernt liegen. Beachten Sie, dass die größeren Abweichungen für Vergleichsintervalle im Gegensatz zum Welch-Test eher konservativ sind.

Die einfache ANOVA im Assistenten lässt bis zu k = 12 Stichproben zu. Daher werden nun Ergebnisse für mehr als 7 Stichproben betrachtet. In der unten stehenden Tabelle sind die

Wahrscheinlichkeiten eines Fehlers 1. Art beim Welch-Test für nicht normalverteilte Daten in k = 9 Gruppen aufgelistet. Auch hier ist das Soll-$\alpha = 0{,}05$.

Tabelle 6 Simulationsergebnisse des Welch-Tests für unterschiedliche Verteilungen mit 9 Stichproben

Verteilung	k = 9
t(3)	0,0362
t(5)	0,0426
Laplace(0;1)	0,0402
Gleichverteilung(-1; 1)	0,0625
Beta(3; 3)	0,0584
Exponential	0,0885
Chi-Quadrat(3)	0,0774
Chi-Quadrat(5)	0,0686
Chi-Quadrat(10)	0,0581
Beta(8; 1)	0,0863

Erwartungsgemäß zeigen die stark schiefen Verteilungen die größten Abweichungen vom Soll-α. Trotz dieses Umstands weichen keine Fehlerwahrscheinlichkeiten um mehr als 4 Prozentpunkte vom Sollwert ab, obgleich dies für die Abweichung der Exponentialverteilung nahezu zutrifft. In der Auswertung werden Stichproben des Umfangs 15 als ausreichend erachtet, so dass sie nicht wegen einer fehlenden Normalverteilung gekennzeichnet werden, da alle Ergebnisse zumindest relativ nah am Soll-α liegen.

Stichproben des Umfangs n = 15 zeigen eine weniger gute Leistung als k = 12 Stichproben. Im Folgenden werden die simulierten Ergebnisse für den Welch-Test für einen Bereich von Stichprobenumfängen aus extremen Nicht-Normalverteilungen untersucht. Dies erleichtert uns das Entwickeln eines sinnvollen Kriteriums für den Stichprobenumfang.

Tabelle 7 Simulationsergebnisse des Welch-Tests für unterschiedliche Verteilungen mit 12 Stichproben

n	T(3)	Gleichverteilung	Chi-Quadrat(5)
10	0,0397	0,0918	0,0792
15	0,0351	0,0695	0,0717
20	0,0362	0,0622	0,0671
30	0,0408	0,0573	0,0657

Für diese Verteilungen ist n = 15 akzeptabel, wenn eine Abweichung von etwas mehr als 2 Prozentpunkten vom Soll-α als akzeptabel erachtet wird. Um die Abweichung unter 2 Prozentpunkten zu halten, muss ein Stichprobenumfang von 20 gewählt werden. Nun werden die Ergebnisse der Chi-Quadrat-Verteilung (3) und der Exponentialverteilung untersucht, die beide eine stärkere Schiefe aufweisen.

Tabelle 8 Simulationsergebnisse des Welch-Tests für die Chi-Quadrat- und die Exponentialverteilung mit 12 Stichproben

n	Chi-Quadrat(3)	Exponential
10	0,1013	0,1064
15	0,0854	0,1079
20	0,0850	0,0951
30	0,0746	0,0829
40	0,0727	0,0735
50	0,0675	0,0694

Diese stark schiefen Verteilungen stellen eine größere Herausforderung dar. Wenn eine Abweichung von weit mehr als 3 Prozentpunkten vom Soll-$\alpha = 0{,}05$ als akzeptabel erachtet wird, kann n = 15 für die Chi-Quadrat-Verteilung (3) als ausreichend akzeptiert werden; für die Exponentialverteilung hingegen ist ein Stichprobenumfang erforderlich, der näher an n = 30 liegt. Da das Kriterium eines bestimmten Stichprobenumfangs tendenziell willkürlich und n = 20 für viele verschiedene Verteilungen relativ gut und für stark schiefe Verteilungen grenzwertig gut geeignet ist, verwenden wir n = 20 als empfohlenen Mindeststichprobenumfang für 10 bis 12 Stichproben. Wenn die Abweichung selbst für stark schiefe Verteilungen klein gehalten werden soll, empfiehlt es sich offensichtlich, größere Stichprobenumfänge zu wählen.