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.

Tests auf Standardabweichungen (bei zwei oder mehr Stichproben)

Übersicht

Der Minitab-Assistent bietet zwei Analysen für Vergleiche unabhängiger Stichproben, anhand derer bestimmt wird, ob ihre Streuungen signifikant voneinander abweichen. Beim Test auf Standardabweichung bei zwei Stichproben werden die Standardabweichungen von zwei Stichproben miteinander verglichen, und beim Test auf gleiche Standardabweichungen werden die Standardabweichungen von mehr als zwei Stichproben miteinander verglichen. Im vorliegenden White Paper bezeichnen wir Designs mit k Stichproben bei $k = 2$ als Designs mit zwei Stichproben und Designs mit k Stichproben bei $k > 2$ als Designs mit mehreren Stichproben. Im Allgemeinen werden diese zwei Arten von Designs gesondert untersucht (siehe Anhang A).

Da die Standardabweichung die Quadratwurzel der Varianz darstellt, entspricht ein Hypothesentest zum Vergleichen von Standardabweichungen einem Hypothesentest zum Vergleichen von Varianzen. Es gibt eine Vielzahl von statistischen Methoden zum Vergleichen der Varianzen aus zwei oder mehr Grundgesamtheiten. Unter diesen Tests zählt der Levene/Brown-Forsythe-Test zu den robustesten und am häufigsten angewendeten. In Designs mit zwei Stichproben ist die Trennschärfe des Levene/Brown-Forsythe-Tests jedoch weniger zufriedenstellend als seine Eigenschaften für Fehler 1. Art. Pan (1999) zeigt auf, dass die Trennschärfe des Tests in Designs mit zwei Stichproben bei bestimmten Grundgesamtheiten, darunter auch die normalverteilte Grundgesamtheit, eine Obergrenze aufweist, die weit unter 1 liegen kann, ungeachtet der Größe der Differenz zwischen den

Standardabweichungen. Mit anderen Worten: Für diese Arten von Daten wird im Test mit größerer Wahrscheinlichkeit geschlussfolgert, dass keine Differenz zwischen den Standardabweichungen vorliegt, wobei die tatsächliche Größe der Differenz keine Rolle spielt. Daher verwendet der Assistent für den Test auf Standardabweichung bei zwei Stichproben einen neuen Test, den Bonett-Test. Für den Test auf gleiche Standardabweichungen bei Designs mit mehreren Stichproben nutzt der Assistent ein Mehrfachvergleichsverfahren.

Der Bonett-Test (2006) stellt eine modifizierte Version des Layard-Tests auf Gleichheit von zwei Varianzen (1978) dar und verbessert die Trennschärfe des Tests bei kleinen Stichproben. Banga und Fox (2013A) leiten die Konfidenzintervalle für den Bonett-Test ab und zeigen, dass diese ebenso genau wie die Konfidenzintervalle für den Levene/Brown-Forsythe-Test und für die meisten Verteilungen genauer sind. Darüber hinaus haben Banga und Fox (2013A) festgestellt, dass der Bonett-Test ebenso robust wie der Levene/Brown-Forsythe-Test und für die meisten Verteilungen trennschärfer ist.

Das Mehrfachvergleichsverfahren umfasst einen Gesamttest der Homogenität (Gleichheit) der Standardabweichungen (oder Varianzen) für mehrere Stichproben, der auf den Vergleichsintervallen jedes Paars von Standardabweichungen basiert. Die Vergleichsintervalle werden so hergeleitet, dass der Mehrfachvergleichstest nur dann signifikant ist, wenn für mindestens ein Paar von Vergleichsintervallen keine Überlappung vorliegt. Banga und Fox (2013B) zeigen, dass der Mehrfachvergleichstest Eigenschaften für Fehler 1. Art und 2. Art aufweist, die bei den meisten Verteilungen denen des Levene/Brown-Forsythe-Tests ähneln. Ein wichtiger Vorteil des Mehrfachvergleichstests ist die grafische Darstellung der Vergleichsintervalle, die ein effektives visuelles Werkzeug zum Bestimmen der Stichproben mit unterschiedlichen Standardabweichungen bietet. Wenn das Design nur zwei Stichproben enthält, entspricht der Mehrfachvergleichstest dem Bonett-Test.

In diesem White Paper wird die Gültigkeit des Bonett-Tests und des Mehrfachvergleichstests für verschiedene Datenverteilungen und Stichprobenumfänge untersucht. Zudem untersuchen wir die Analyse der Trennschärfe und des Stichprobenumfangs für den Bonett-Test, die auf einer Approximationsmethode für große Stichproben basiert. Auf der Grundlage dieser Faktoren haben wir die folgenden Prüfungen entwickelt, die der Assistent automatisch für Ihre Daten ausführt und in der Auswertung anzeigt:

- Ungewöhnliche Daten
- Vorliegen einer Normalverteilung
- Gültigkeit des Tests
- Stichprobenumfang (nur Test auf Standardabweichung bei zwei Stichproben)

Methoden für Tests auf gleiche Standardabweichungen

Gültigkeit des Bonett-Tests und des Mehrfachvergleichstests

In ihrer Vergleichsstudie von Tests auf gleiche Varianzen haben Conover et al. (1981) festgestellt, dass der Levene/Brown-Forsythe-Test auf der Grundlage der Wahrscheinlichkeiten eines Fehlers 1. Art und 2. Art unter den Tests mit der besten Leistung war. Seitdem wurden weitere Methoden für Tests auf gleiche Varianzen in Designs mit zwei Stichproben und Designs mit mehreren Stichproben vorgeschlagen (Pan, 1999; Shoemaker, 2003; Bonett, 2006). Pan zeigt beispielsweise, dass der Levene/Brown-Forsythe-Test trotz seiner Robustheit und seiner unkomplizierten Interpretation keine ausreichende Trennschärfe aufweist, um wichtige Differenzen zwischen zwei Standardabweichungen zu erkennen, wenn die Stichproben aus gewissen Grundgesamtheiten (u. a. aus der normalverteilten Grundgesamtheit) stammen. Wegen dieser wesentlichen Einschränkung nutzt der Assistent den Bonett-Test als Test auf Standardabweichung bei zwei Stichproben (siehe Anhang A oder Banga und Fox, 2013A). Für den Test auf gleiche Standardabweichungen bei mehr als zwei Stichproben verwendet der Assistent ein Mehrfachvergleichsverfahren mit Vergleichsintervallen, das eine grafische Darstellung zum Identifizieren von Stichproben mit abweichenden Standardabweichungen liefert, wenn der Mehrfachvergleichstest signifikant ist (siehe Anhang A sowie Banga und Fox, 2013B).

Zielstellung

Zunächst sollte die Leistung des Bonett-Tests beim Vergleich der Standardabweichungen von zwei Grundgesamtheiten ausgewertet werden. Zweitens sollte die Leistung des Mehrfachvergleichstests beim Vergleich der Standardabweichungen für mehr als zwei Grundgesamtheiten ausgewertet werden. Insbesondere sollte die Gültigkeit dieser Tests bei Stichproben mit unterschiedlichem Umfang aus verschiedenen Arten von Verteilungen ausgewertet werden.

Methode

Definitionen der für den Bonett-Test und den Mehrfachvergleichstest verwendeten statistischen Methoden werden in Anhang A aufgeführt. Zum Auswerten der Gültigkeit der Tests musste untersucht werden, ob deren Wahrscheinlichkeiten eines Fehlers 1. Art unter verschiedenen Bedingungen nahe dem Ziel-Konfidenzniveau (Alpha) bleibt. Hierfür wurde eine Reihe von Simulationen durchgeführt, um die Gültigkeit des Bonett-Tests beim Vergleichen der Standardabweichungen aus zwei unabhängigen Stichproben auszuwerten, sowie weitere Reihen von Simulationen, um die Gültigkeit des Mehrfachvergleichstests beim Vergleichen der Standardabweichungen aus mehreren (k) unabhängigen Stichproben auszuwerten, wenn $k > 2$.

Dafür wurden 10.000 Paare von mehreren (k) Zufallsstichproben mit unterschiedlichem Umfang aus mehreren Verteilungen unter Verwendung von balancierten und nicht balancierten Designs generiert. Anschließend wurde ein zweiseitiger Bonett-Test zum Vergleichen der Standardabweichungen der zwei Stichproben oder ein Mehrfachvergleichstest zum Vergleichen der Standardabweichungen der k Stichproben in jedem Experiment durchgeführt, wobei ein Soll-Signifikanzniveau von $\alpha = 0{,}05$ zugrunde gelegt wurde. Es wurde gezählt, wie häufig die Nullhypothese in 10.000 Replikationen zurückgewiesen wurde (wenn die wahren Standardabweichungen tatsächlich gleich waren), und dieser Anteil, der als simuliertes Signifikanzniveau bezeichnet wird, wurde mit dem Soll-Signifikanzniveau verglichen. Wenn der Test eine gute Leistung zeigt, sollte das simulierte Signifikanzniveau, das die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art darstellt, sehr nahe am Soll-Signifikanzniveau liegen. Weitere Einzelheiten zu den spezifischen Methoden, die für Simulationen mit zwei Stichproben und k Stichproben verwendet wurden, finden Sie in Anhang B.

Ergebnisse

Für Vergleiche mit zwei Stichproben lagen die simulierten Wahrscheinlichkeiten eines Fehlers 1. Art des Bonett-Tests bei mittleren oder großen Stichproben nahe dem Soll-Signifikanzniveau, ungeachtet der Verteilung und ungeachtet davon, ob das Design balanciert oder nicht balanciert war. Wenn jedoch kleine Stichproben aus extrem schiefen Grundgesamtheiten gezogen wurden, war der Bonett-Test generell konservativ und wies Wahrscheinlichkeiten eines Fehlers 1. Art auf, die etwas niedriger als das Soll-Signifikanzniveau (d. h. die Soll-Wahrscheinlichkeit eines Fehlers 1. Art) lagen.

Für Vergleiche mit mehreren Stichproben lagen die Wahrscheinlichkeiten eines Fehlers 1 Art des Mehrfachvergleichstests bei mittleren oder großen Stichproben nahe dem Soll-Signifikanzniveau, ungeachtet der Verteilung und ungeachtet davon, ob das Design balanciert oder nicht balanciert war. Bei kleinen und extrem schiefen Stichproben war der Test jedoch generell weniger konservativ und lieferte Wahrscheinlichkeiten eines Fehlers 1. Art, die größer als das Soll-Signifikanzniveau waren, wenn das Design eine große Anzahl von Stichproben enthielt.

Die Ergebnisse unserer Untersuchungen stimmten mit denen von Banga und Fox (2013A) und (2013B) überein. Wir haben die Schlussfolgerung gezogen, dass der Bonett-Test und der Mehrfachvergleichstest eine gute Leistung zeigen, wenn der kleinste Stichprobenumfang mindestens 20 beträgt. Daher wird diese Anforderung eines minimalen Stichprobenumfangs für die Gültigkeit des Tests geprüft und in der Auswertung des Assistenten aufgeführt (siehe Abschnitt „Datenprüfungen“).

Vergleichsintervalle

Wenn ein Test zum Vergleichen von zwei oder mehr Standardabweichungen statistisch signifikant ist und darauf hinweist, dass sich mindestens eine Standardabweichung von den anderen unterscheidet, muss im nächsten Schritt der Analyse bestimmt werden, welche Stichproben statistisch betrachtet Abweichungen aufweisen. Eine intuitive Vergleichsmöglichkeit besteht darin, die Konfidenzintervalle der einzelnen Stichproben 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.

Zielstellung

Es sollte eine Methode zum Berechnen einzelner Vergleichsintervalle entwickelt werden, die sowohl als Gesamttest der Homogenität der Varianzen als auch als Methode zum Identifizieren der Stichproben mit abweichenden Varianzen verwendet werden kann, wenn der Gesamttest signifikant ist. Eine kritische Anforderung in Bezug auf das Mehrfachvergleichsverfahren besteht darin, dass der Gesamttest nur dann signifikant ist, wenn für mindestens ein Paar der Vergleichsintervalle keine Überlappung festgestellt wird. Dies weist darauf hin, dass sich die Standardabweichungen von mindestens zwei Stichproben voneinander unterscheiden.

Methode

Das verwendete Mehrfachvergleichsverfahren zum Vergleichen mehrerer Standardabweichungen wird aus mehreren paarweisen Vergleichen abgeleitet. Jedes Paar von Stichproben wird mit dem Bonett-Test (2006) auf Gleichheit der Standardabweichungen der zwei Grundgesamtheiten verglichen. Bei den paarweisen Vergleichen kommt eine Multiplizitätskorrektur zur Anwendung, die auf einer Approximation für große Stichproben beruht, wie in Nakayama (2009) gezeigt. Die Approximation für große Stichproben wird gegenüber der häufiger verwendeten Bonferroni-Korrektur bevorzugt, da die Bonferroni-Korrektur mit steigender Anzahl von Stichproben zunehmend konservativ wird. Die Vergleichsintervalle schließlich sind Ergebnis der paarweisen Vergleiche, die auf dem Verfahren der besten Approximation von Hochberg et al. (1982) beruhen. Einzelheiten können Sie Anhang A entnehmen.

Ergebnisse

Das Mehrfachvergleichsverfahren erfüllt die Anforderung, dass der Gesamttest der Gleichheit von Standardabweichungen nur dann signifikant ist, wenn für mindestens zwei Vergleichsintervalle keine Überlappung vorliegt. Wenn der Gesamttest nicht signifikant ist, müssen alle Vergleichsintervalle überlappen.

Der Assistent zeigt die Vergleichsintervalle im Vergleichsdiagramm für Standardabweichungen im Zusammenfassungsbericht an. Neben diesem Diagramm wird vom Assistenten der p-Wert des Mehrfachvergleichstests angezeigt, der der Gesamttest auf Homogenität der Standardabweichungen ist. Wenn der Test auf Standardabweichungen statistisch signifikant ist, wird jedes Vergleichsintervall rot markiert, das nicht mindestens ein anderes Intervall überlappt. Wenn der Test auf gleiche Standardabweichungen hingegen nicht statistisch signifikant ist, werden keine der Intervalle rot gekennzeichnet.

Leistung der theoretischen Trennschärfe (nur Designs mit zwei Stichproben)

Die theoretischen Trennschärfefunktionen der Bonett- und Mehrfachvergleichstests werden zum Planen von Stichprobenumfängen benötigt. Bei Designs mit zwei Stichproben kann eine

approximierte theoretische Trennschärfefunktion des Tests mit theoretischen Methoden für große Stichproben abgeleitet werden. Da diese Funktion auf Approximationsmethoden für große Stichproben zurückgeht, müssen ihre Eigenschaften ausgewertet werden, wenn der Test mit kleinen Stichproben durchgeführt wird, die aus Normalverteilungen und Nicht-Normalverteilungen generiert wurden. Beim Vergleichen der Standardabweichungen von mehr als zwei Gruppen kann die theoretische Trennschärfefunktion des Mehrfachvergleichstests jedoch nicht auf einfache Weise ermittelt werden.

Zielstellung

Wir wollten feststellen, ob die theoretische Trennschärfefunktion auf Grundlage der Approximation für große Stichproben zum Auswerten der Anforderungen an Trennschärfe und Stichprobenumfang für den Test auf Standardabweichung bei zwei Stichproben im Assistenten verwendet werden kann. Hierfür musste untersucht werden, ob die approximierte theoretische Trennschärfefunktion die vom Bonett-Test tatsächlich erzielte Trennschärfe darstellt, wenn dieser mit Daten aus mehreren Arten von Verteilungen (Normalverteilungen und Nicht-Normalverteilungen) durchgeführt wird.

Methode

Die Ableitung der approximierten theoretischen Trennschärfefunktion des Bonett-Tests für Designs mit zwei Stichproben wird in Anhang C gezeigt.

Mit dem Bonett-Test wurden Simulationen zum Schätzen der tatsächlichen Trennschärfen durchgeführt (die wir als simulierte Trennschärfen bezeichnen). Zuerst wurden Paare von Zufallsstichproben mit unterschiedlichem Umfang aus mehreren Verteilungen (Normalverteilungen und Nicht-Normalverteilungen) generiert. Für jede Verteilung wurde der Bonett-Test mit jedem der 10.000 Paare von Stichprobenreplikationen durchgeführt. Für jedes Paar von Stichprobenumfängen wurde die simulierte Trennschärfe des Tests zum Erkennen einer gegebenen Differenz als Anteil der 10.000 Paare von Stichproben berechnet, bei denen der Test signifikant ist. Zum Vergleich wurde auch die entsprechende Trennschärfe mit der approximierten theoretischen Trennschärfefunktion des Tests berechnet. Wenn die Approximation eine gute Leistung zeigt, liegen die theoretischen und simulierten Trennschärfen nahe beieinander. Weitere Informationen finden Sie in Anhang D.

Ergebnisse

Die Simulationen zeigten, dass die theoretische und die simulierte Trennschärfefunktion des Bonett-Tests für die meisten Verteilungen bei kleinen Stichprobenumfängen nahezu gleich sind und näher beieinander liegen, wenn der minimale Stichprobenumfang den Wert 20 erreicht. Bei symmetrischen und nahezu symmetrischen Verteilungen mit schwächer bis gemäßigt besetzten Randbereichen liegen die theoretischen Trennschärfen etwas höher als die simulierten (tatsächlichen) Trennschärfen. Bei schiefen Verteilungen und Verteilungen mit stärker gewichteten Randbereichen sind sie jedoch kleiner als die simulierten (tatsächlichen) Trennschärfen. Weitere Informationen finden Sie in Anhang D.

Insgesamt zeigen unsere Ergebnisse, dass die theoretische Trennschärfefunktion eine gute Grundlage für die Planung von Stichprobenumfängen darstellt.

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. Die Simulationsuntersuchungen zeigen, dass der Bonett-Test und der Mehrfachvergleichstest konservativ sind, wenn die Daten Ausreißer enthalten (siehe Anhang B). Die tatsächlichen Signifikanzniveaus der Tests sind erheblich kleiner als die Soll-Niveaus, insbesondere wenn die Analyse mit kleinen Stichproben durchgeführt wird.

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.

Vorliegen einer Normalverteilung

Im Unterschied zu den meisten Tests auf Gleichheit der Varianzen, die unter der Annahme der Normalverteilung abgeleitet werden, wird für den Bonett-Test und den Mehrfachvergleichstest auf Gleichheit der Standardabweichungen keine Annahme zur spezifischen Verteilung der Daten getroffen.

Zielstellung

Dem Bonett-Test und dem Mehrfachvergleichstest liegen zwar Methoden der Approximation für große Stichproben zugrunde, wir wollten jedoch nachweisen, dass sie bei kleinen Stichproben aus normalverteilten und nicht normalverteilten Daten eine gute Leistung zeigen. Zudem wollten wir die Benutzer darüber informieren können, in welcher Beziehung das Vorliegen einer Normalverteilung in den Daten zu den Ergebnissen der Tests auf gleiche Standardabweichungen steht.

Methode

Zum Auswerten der Gültigkeit der Tests unter verschiedenen Bedingungen wurden Simulationen durchgeführt, um die Wahrscheinlichkeit eines Fehlers 1. Art des Bonett-Tests und des Mehrfachvergleichstests bei normalverteilten und nicht normalverteilten Daten unterschiedlicher Stichprobenumfänge zu untersuchen. Weitere Einzelheiten finden Sie im Abschnitt „Methoden für Tests auf gleiche Standardabweichungen“ und in Anhang B.

Ergebnisse

Die Simulationen zeigten, dass die Verteilung der Daten bei ausreichend großen Stichproben (minimaler Stichprobenumfang ≥ 20) keine wesentliche Auswirkung auf die Eigenschaften für den Fehler 1. Art des Bonett-Tests oder des Mehrfachvergleichstests hatte. Die Tests weisen Wahrscheinlichkeiten eines Fehlers 1. Art auf, die für normalverteilte und nicht normalverteilte Daten durchgehend nahe der Soll-Wahrscheinlichkeit liegen.

Basierend auf diesen Ergebnissen für die Wahrscheinlichkeit eines Fehlers 1. Art zeigt der Assistent die folgenden Informationen zur Normalverteilung in der Auswertung an.

Für Designs mit zwei Stichproben zeigt der Assistent den folgenden Indikator an:

Status	Bedingung
	Für diese Analyse wird der Bonett-Test verwendet. Bei ausreichend großen Stichproben ergibt der Test sowohl für normalverteilte als auch für nicht normalverteilte Daten gute Ergebnisse.

Für Designs mit mehreren Stichproben zeigt der Assistent den folgenden Indikator an:

Status	Bedingung
	Für diese Analyse wird ein Mehrfachvergleichstest verwendet. Bei ausreichend großen Stichproben ergibt der Test sowohl für normalverteilte als auch für nicht normalverteilte Daten gute Ergebnisse.

Gültigkeit des Tests

Im Abschnitt „Methoden für Tests auf gleiche Standardabweichungen“ wurde gezeigt, dass der Bonett-Test und der Mehrfachvergleichstest für Vergleiche mit zwei Stichproben und für mehrere (k) Stichproben Wahrscheinlichkeiten eines Fehlers 1. Art aufweisen, die sowohl für normalverteilte als auch für nicht normalverteilte Daten in balancierten und nicht balancierten Designs nahe der Soll-Wahrscheinlichkeit liegen, sofern die Stichproben einen mittleren oder großen Stichprobenumfang aufweisen. Bei kleinen Stichproben zeigen der Bonett-Test und der Mehrfachvergleichstest hingegen nicht generell eine gute Leistung.

Zielstellung

Wir wollten eine Regel erarbeiten, mit der die Gültigkeit der Ergebnisse des Tests auf Standardabweichungen bei zwei Stichproben und bei mehreren (k) Stichproben auf der Grundlage der Daten des Benutzers ausgewertet werden kann.

Methode

Zum Auswerten der Gültigkeit der Tests unter verschiedenen Bedingungen wurden Simulationen durchgeführt, um die Wahrscheinlichkeit eines Fehlers 1. Art des Bonett-Tests und des Mehrfachvergleichstests bei verschiedenen Datenverteilungen, Stichprobenanzahlen und Stichprobenumfängen zu untersuchen, wie bereits im Abschnitt „Methoden für Tests auf gleiche Standardabweichungen“ beschrieben. Weitere Informationen finden Sie in Anhang B.

Ergebnisse

Der Bonett-Test und der Mehrfachvergleichstest zeigen eine gute Leistung, wenn der Stichprobenumfang der kleinsten Stichprobe mindestens 20 beträgt. Daher zeigt der Assistent für die Gültigkeit der Tests auf gleiche Standardabweichungen die folgenden Statusindikatoren in der Auswertung an.

Status	Bedingung
	Die Stichprobenumfänge betragen mindestens 20, so dass der p-Wert genau sein sollte.
	Einige Stichprobenumfänge liegen unter 20, so dass der p-Wert u. U. nicht genau ist. Erwägen Sie, die Stichprobenumfänge auf mindestens 20 zu erhöhen.

Stichprobenumfang (nur für Test auf Standardabweichungen bei zwei Stichproben)

Normalerweise wird ein statistischer Hypothesentest durchgeführt, um einen Beleg für die Zurückweisung der Nullhypothese („keine Differenz“) zu erhalten. Wenn die Stichprobe zu klein ist, reicht die Trennschärfe des Tests u. U. nicht aus, um eine tatsächlich vorhandene Differenz zu erkennen; hierbei handelt es sich um einen Fehler 2. Art. Daher muss unbedingt sichergestellt werden, dass die Stichprobenumfänge ausreichend groß sind, um mit einer hohen Wahrscheinlichkeit Differenzen mit praktischen Konsequenzen zu erkennen.

Zielstellung

Wenn die Daten keine ausreichenden Hinweise zum Zurückweisen der Nullhypothese liefern, wollten wir ermitteln können, ob die Stichprobenumfänge groß genug für den Test sind, so dass dieser mit hoher Wahrscheinlichkeit Differenzen mit praktischen Konsequenzen erkennt. Bei der Planung der Stichprobenumfänge soll zwar sichergestellt werden, dass die Stichprobenumfänge ausreichend groß sind, um mit hoher Wahrscheinlichkeit wichtige Differenzen zu erkennen; andererseits dürfen sie aber nicht so groß sein, dass bedeutungslose Differenzen mit hoher Wahrscheinlichkeit statistisch signifikant werden.

Methode

Die Analyse von Trennschärfe und Stichprobenumfang für den Test auf Standardabweichung bei zwei Stichproben basiert auf einer Approximation der Trennschärfefunktion des Bonett-Tests, der i. d. R. gute Schätzwerte der tatsächlichen Trennschärfefunktion des Tests liefert (siehe die Zusammenfassung der Simulationsergebnisse unter „Leistung der theoretischen Trennschärfe“ im Abschnitt „Methoden“).

Ergebnisse

Wenn die Daten keine ausreichenden Hinweise liefern, die gegen die Nullhypothese sprechen, berechnet der Assistent mit Hilfe der approximierten Trennschärfefunktion des Bonett-Tests die Differenzen mit praktischen Konsequenzen, die für den gegebenen Stichprobenumfang mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden können. Wenn der Benutzer zudem eine konkrete Differenz mit praktischen Konsequenzen angibt, berechnet der Assistent mit der Trennschärfefunktion des Tests auf Normal-Approximation Stichprobenumfänge, bei denen die Differenz mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt wird.

Um die Interpretation der Ergebnisse zu erleichtern, werden für die Prüfung auf die Trennschärfe und den Stichprobenumfang in der Auswertung des Assistenten für den Test auf Standardabweichung bei zwei Stichproben die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Im Test wird eine Differenz der Standardabweichungen festgestellt, so dass die Trennschärfe kein Problem darstellt. ODER Die Trennschärfe ist ausreichend. Im Test wurde keine Differenz zwischen den Standardabweichungen 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 Standardabweichungen festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 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 Standardabweichungen 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.

Status	Bedingung
	Die Trennschärfe ist nicht ausreichend. Im Test wurde keine Differenz zwischen den Standardabweichungen 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.
	Im Test wurde keine Differenz zwischen den Standardabweichungen festgestellt. Sie haben keine zu erkennende Differenz mit praktischen Konsequenzen angegeben; daher wird in der Auswertung die Differenz angegeben, die bei Ihrem Stichprobenumfang und Alpha mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt wird.

Literaturhinweise

Arnold, S.F. (1990). *Mathematical statistics*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Banga, S.J. und Fox, G.D. (2013A). On Bonett's Robust Confidence Interval for a Ratio of Standard Deviations. *White Paper, Minitab Inc*.

Banga, S.J. und Fox, G.D. (2013B) A graphical multiple comparison procedure for several standard deviations. *White Paper, Minitab Inc*.

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.

Gastwirth, J. L. (1982). Statistical properties of a measure of tax assessment uniformity. *Journal of Statistical Planning and Inference, 6*, 1-12.

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

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 (Hrsg.), *Probability and statistics* (278-292). Stanford University Press, Palo Alto, California.

Nakayama, M.K. (2009). Asymptotically valid single-stage multiple-comparison procedures. *Journal of Statistical Planning and Inference*, 139, 1348-1356.

Pan, G. (1999) On a Levene type test for equality of two variances. *Journal of Statistical Computation and Simulation*, 63, 59-71.

Shoemaker, L. H. (2003). Fixing the F test for equal variances. *The American Statistician, 57 (2)*, 105-114.

Anhang A: Methode für den Bonett-Test und den Mehrfachvergleichstest

Die zugrunde liegenden Annahmen für das Ziehen von Rückschlüssen zu den Standardabweichungen oder Varianzen mit der Bonett-Methode (Designs mit zwei Stichproben) oder dem Mehrfachvergleichsverfahren (Designs mit mehreren Stichproben) können wie folgt beschrieben werden. Angenommen, $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$ sind k ($k \geq 2$) unabhängige Zufallsstichproben, wobei jede Stichprobe aus einer Verteilung mit dem unbekannten Mittelwert μ_i und der unbekannten Varianz σ_i^2 gezogen wurde, für $i = 1, \ldots, k$. Angenommen, die Verteilung der Grundgesamtheit der Stichproben weist eine gemeinsame endliche Kurtosis $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$ auf. Diese Annahme ist zwar für die theoretischen Ableitungen unerlässlich, für die meisten praktischen Anwendungen, bei denen die Stichproben ausreichend groß sind, ist sie jedoch weniger wichtig (Banga und Fox, 2013A).

Methode A1: Bonett-Test auf Gleichheit von zwei Varianzen

Der Bonett-Test gilt nur für Designs mit zwei Stichproben, bei denen zwei Varianzen oder Standardabweichungen verglichen werden. Der Test stellt eine abgewandelte Version des Layard-Tests auf Gleichheit von Varianzen (1978) in Designs mit zwei Stichproben dar. In einem beidseitiger Bonett-Test auf Gleichheit von zwei Varianzen mit dem Signifikanzniveau α wird die Nullhypothese der Gleichheit nur in folgendem Fall zurückgewiesen:

$$\left|\ln(c\, S_1^2/S_2^2)\right| > z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

Dabei gilt Folgendes:

S_i ist die Stichproben-Standardabweichung von Stichprobe i

$g_i = (n_i - 3)/n_i, i = 1{,}2$

$z_{\alpha/2}$ bezieht sich auf das obere $\alpha/2$. Perzentil der Standardnormalverteilung

$\hat{\gamma}_P$ ist der zusammengefasste Kurtosis-Schätzwert, angegeben wie folgt:

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

Im Ausdruck des zusammengefassten Kurtosis-Schätzwerts ist m_i der getrimmte Mittelwert für Stichprobe i mit dem Trim-Anteil $1/[2(n_i - 4)^{1/2}]$.

Der oben genannte Ausdruck enthält die Konstante c als Korrektur für kleine Stichproben, um die Auswirkung ungleicher Fehlerwahrscheinlichkeiten in den Randbereichen nicht

balancierter Designs zu reduzieren. Diese Konstante ist als $c = c_1/c_2$ angegeben, wobei Folgendes gilt:

$$c_{\mathrm{i}} = \frac{n_{\mathrm{i}}}{n_{\mathrm{i}} - z_{\alpha/2}}, i = 1; 2$$

Wenn das Design balanciert ist, d. h. wenn $n_1 = n_2$, wird der p-Wert des Tests berechnet als:

$$\mathrm{P} = 2\,\Pr(Z > z)$$

Hierbei ist Z eine Zufallsvariable, die einer Standardnormalverteilung folgt, und z der beobachtete Wert der folgenden Statistik, die auf den vorhandenen Daten beruht. Die Statistik lautet:

$$Z = \frac{\ln(C\ S_1^2/S_2^2)}{se}$$

Dabei gilt Folgendes:

$$se = \sqrt{\frac{\hat{Y}_P - g_1}{n_1 - 1} + \frac{\hat{Y}_P - g_2}{n_2 - 1}}$$

Wenn das Design hingegen nicht balanciert ist, wird der p-Wert des Tests berechnet als:

$$P = 2\min(\alpha_L; \alpha_U)$$

Hierbei ist $\alpha_L = \Pr(Z > z_L)$ und $\alpha_U = \Pr(Z > z_U)$. Die Variable z_L ist die kleinste Wurzel der Funktion

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

und z_U ist die kleinste Wurzel der Funktion $L(z, S_2, S_1, n_2, n_1)$.

Methode A2: Mehrfachvergleichstest und Vergleichsintervalle

Angenommen, es liegen k $(k \geq 2)$ unabhängige Gruppen oder Stichproben vor. Unser Ziel bestand darin, ein System von k Intervallen für die Standardabweichungen der Grundgesamtheit derart aufzustellen, dass der Test auf Gleichheit der Standardabweichungen nur dann signifikant ist, wenn für mindestens zwei der k Intervalle keine Überlappung vorliegt. Diese Intervalle werden als Vergleichsintervalle bezeichnet. Diese Vergleichsmethode ähnelt den Verfahren für Mehrfachvergleiche der Mittelwerte in Modelle für einfache Varianzanalysen (ANOVA), die ursprünglich von Tukey-Kramer entwickelt und später von Hochberg et al. (1982) verallgemeinert wurden.

Vergleichen von zwei Standardabweichungen

Bei Designs mit zwei Stichproben können die Konfidenzintervalle für das Verhältnis der Standardabweichungen für den Bonett-Test direkt berechnet werden, um die Größe der Differenz zwischen den Standardabweichungen zu ermitteln (Banga und Fox, 2013A). Tatsächlich nutzen wir diesen Ansatz für „Statistik > Statistische Standardverfahren > Test auf Varianzen, 2 Stichproben“ in Release 17 von Minitab. Im Assistenten sollten jedoch Vergleichsintervalle bereitgestellt werden, die leichter als das Konfidenzintervall des

Verhältnisses der Standardabweichungen interpretiert werden können. Hierfür wurden mit dem in Methode A1 beschriebenen Bonett-Verfahren die Vergleichsintervalle für die zwei Stichproben bestimmt.

Wenn zwei Stichproben vorhanden sind, ist der Bonett-Test auf Gleichheit von Varianzen nur dann signifikant, wenn das folgende Annahmeintervall für den Bonett-Test auf Gleichheit von Varianzen nicht 0 enthält:

$$\ln(c_1 S_1^2) - \ln(c_2 S_2^2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

Hierbei entsprechen der zusammengefasste Kurtosis-Schätzwert $\hat{\gamma}_P$ und $g_i, i = 1; 2$ den oben gegebenen Definitionen.

Aus diesem Intervall werden die folgenden zwei Vergleichsintervalle derart abgeleitet, dass der Test auf Gleichheit von Varianzen oder Standardabweichungen nur dann signifikant ist, wenn diese einander nicht überlappen. Diese zwei Intervalle lauten:

$$\left[S_i\sqrt{\mathrm{C_i}\exp(-z_{\alpha/2}V_i)}\,, S_i\sqrt{C_i\exp(z_{\alpha/2}V_i)}\right], i = 1; 2$$

Dabei gilt Folgendes:

$$V_i = \frac{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}}}{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}} + \sqrt{\frac{\hat{\gamma}_P - g_j}{n_j - 1}}}\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}, i = 1; 2, j = 1; 2, i \neq j$$

Die Verwendung dieser Intervalle als Verfahren für Tests auf Gleichheit der Standardabweichungen entspricht dem Bonett-Test auf Gleichheit von Standardabweichungen. Insbesondere liegt nur dann keine Überlappung der Intervalle vor, wenn der Bonett-Test auf Gleichheit von Standardabweichungen signifikant ist. Beachten Sie jedoch, dass diese Intervalle keine Konfidenzintervalle von Standardabweichungen darstellen, sondern lediglich für Mehrfachvergleiche von Standardabweichungen geeignet sind. Hochberg et al. bezeichnen aus eben diesem Grund ähnliche Intervalle für Vergleiche von Mittelwerten als Unsicherheitsintervalle. Wir bezeichnen diese Intervalle als Vergleichsintervalle.

Da das Verfahren der Vergleichsintervalle dem Bonett-Test auf Gleichheit von Standardabweichungen entspricht, ist der p-Wert für die Vergleichsintervalle identisch mit dem p-Wert des bereits an früherer Stelle beschriebenen Bonett-Tests auf Gleichheit von zwei Standardabweichungen.

Vergleichen von mehreren Standardabweichungen

Wenn mehr als zwei Gruppen oder Stichproben vorliegen, werden die k Vergleichsintervalle von $k(k-1)/2$ paarweisen Simultantests auf Gleichheit von Standardabweichungen mit simultanem Signifikanzniveau α abgeleitet. Konkreter: Seien $X_{i1}, \ldots, X_{in_i}$ und $X_{j1}, \ldots, X_{jn_j}$ die Stichprobendaten für jedes Paar (i, j) von Stichproben. Ebenso wie bei zwei Stichproben ist der Test auf Gleichheit der Standardabweichungen für das spezifische Paar (i, j) von Stichproben nur dann auf einem α'-Niveau signifikant, wenn das Intervall

$$\ln(c_i S_i^2) - \ln(c_j S_j^2) \pm z_{\alpha'/2}\sqrt{\frac{\hat{\gamma}_{ij} - g_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - g_j}{n_j - 1}}$$

nicht 0 enthält. Im oben aufgeführten $\hat{\gamma}_{ij}$ basiert der zusammengefasste Kurtosis-Schätzwert auf dem Paar $(i; j)$ von Stichproben, und er wird wie folgt ausgedrückt:

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

Darüber hinaus ist, wie bereits definiert, m_i der getrimmte Mittelwert für Stichprobe i mit dem Trim-Anteil $1/[2(n_i - 4)^{1/2}]$ und

$$g_i = \frac{n_i - 3}{n_i}, g_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha'/2}}, c_j = \frac{n_j}{n_j - z_{\alpha'/2}}$$

Da $k(k-1)/2$ paarweise Simultantests vorhanden sind, muss das α'-Niveau so gewählt werden, dass die tatsächliche simultane Fehlerrate nahe dem Soll-Signifikanzniveau α liegt. Eine mögliche Korrektur basiert auf der Approximation nach Bonferroni. Der konservative Charakter von Bonferroni-Korrekturen nimmt jedoch mit steigender Anzahl von Stichproben im Design nachweislich zu. Ein besserer Ansatz basiert auf einer Normal-Approximation, wie bei Nakayama (2008) beschrieben. Bei diesem Ansatz wird lediglich $z_{\alpha'/2}$ durch $q_{\alpha,k}/\sqrt{2}$ ersetzt, wobei $q_{\alpha,k}$ der obere α-Punkt des Bereichs von k unabhängigen und identisch verteilten Zufallsvariablen mit Standardnormalverteilung ist; d. h.

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

wobei $Z_1, \ldots, Z_k$ unabhängige und identisch verteilte Zufallsvariablen mit Standardnormalverteilung sind.

In Anlehnung an eine Methode von Hochberg et al. (1982) weist das Verfahren, das das oben beschriebene paarweise Verfahren am besten approximiert, die Nullhypothese der Gleichheit von Standardabweichungen nur zurück, wenn für ein bestimmtes Paar $(i; j)$ von Stichproben folgendes gilt:

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

Hierbei wird V_i derart gewählt, dass folgender Betrag minimiert wird:

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

mit

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

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 der auf dem Approximationsverfahren basierende Test nur dann signifikant ist, wenn für mindestens ein Paar der folgenden k Intervalle keine Überlappung vorliegt.

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

Zum Berechnen des p-Gesamtwerts für den Mehrfachvergleichstest wird als P_{ij} p-Wert für ein beliebiges Paar $(i;j)$ von Stichproben festgelegt. Daraus folgt dieser p-Gesamtwert für den Mehrfachvergleichstest:

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

Zum Berechnen von P_{ij} wird der Algorithmus des Designs mit zwei Stichproben aus Methode A1 ausgeführt mit

$$se=V_i+V_j$$

Hierbei entspricht V_i dem oben aufgeführten Ausdruck.

Konkreter: Wenn $n_i\neq n_j$

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

wobei $\alpha_L=\Pr(Q>z_L\sqrt{2})$, $\alpha_U=\Pr(Q>z_U\sqrt{2})$, die Variable z_L die kleinste Wurzel der Funktion $L\big(z,S_i,S_j,n_i,n_j\big)$, die Variable z_U die kleinste Wurzel der Funktion $L\big(z,S_j,S_i,n_j,n_i\big)$ und Q eine Zufallsvariable mit der bereits definierten Bereichsverteilung ist.

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

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

Anhang B: Gültigkeit des Bonett-Tests und des Mehrfachvergleichstests

Simulation B1: Gültigkeit des Bonett-Tests (Modelle mit zwei Stichproben, balancierte und nicht balancierte Designs)

Es wurden Paare von Zufallsstichproben mit kleinem bis mittlerem Umfang aus Verteilungen mit unterschiedlichen Eigenschaften generiert. Hierzu zählten folgende Verteilungen:

- Standardnormalverteilung (N(0;1))
- Symmetrische Verteilungen mit schwächer besetzten Randbereichen, darunter die Gleichverteilung (U(0;1)) und die Betaverteilung, bei der beide Parameter auf 3 festgelegt sind (B(3;3))
- Symmetrische Verteilungen mit stärker besetzten Randbereichen, darunter t-Verteilungen mit 5 und 10 Freiheitsgraden (t(5), t(10)) und die Laplace-Verteilung mit Lage 0 und Skala 1 (Lpl))
- Schiefe Verteilungen mit stärker besetzten Randbereichen, darunter die Exponentialverteilung mit Skala 1 (Exp) und Chi-Quadrat-Verteilungen mit 5 und 10 Freiheitsgraden (Chi(5), Chi(10))
- Linksschiefe Verteilungen mit stärker besetzten Randbereichen, insbesondere die Betaverteilung, deren Parameters auf 8 bzw. 1 festgelegt sind (B(8;1))

Zum Untersuchen der direkten Auswirkung von Ausreißern wurden Paare von Stichproben aus kontaminierten Normalverteilungen generiert, die folgendermaßen definiert wurden:

$$CN(p;\sigma) = pN(0;1) + (1-p)N(0;\sigma)$$

Hierbei ist p der Mischparameter und $1-p$ der Anteil der Kontamination (der dem Anteil der Ausreißer entspricht). Es wurden zwei kontaminierte Normalverteilungen für die Untersuchung ausgewählt: $CN(0{,}9;3)$, bei der 10 % der Grundgesamtheit Ausreißer darstellen, und $CN(0{,}8;3)$, bei der 20 % der Grundgesamtheit Ausreißer sind. Diese zwei Verteilungen sind symmetrisch und haben aufgrund der Ausreißer lange Randbereiche.

Für jedes Paar von Stichproben aus jeder Verteilung wurde ein beidseitiger Bonett-Test mit einem Soll-Signifikanzniveau von $\alpha = 0{,}05$ ausgeführt. Da die simulierten Signifikanzniveaus in jedem Fall auf 10.000 Paaren von Stichprobenreplikationen basierten und ein Soll-Signifikanzniveau von 5 % angesetzt wurde, betrug der Simulationsfehler $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$.

Die Simulationsergebnisse sind unten in Tabelle 1 zusammengefasst.

Tabelle 1 Simulierte Signifikanzniveaus für einen beidseitigen Bonett-Test in balancierten und nicht balancierten Designs mit zwei Stichproben. Das Soll-Signifikanzniveau ist 0,05.

Verteilung	n_1, n_2	Simuliertes Niveau	Verteilung	n_1, n_2	Simuliertes Niveau
N(0;1)	10; 10	0,038	Exp	10; 10	0,052
	20; 10	0,043		20; 10	0,051
	20; 20	0,045		20; 20	0,049
	30; 10	0,044		30; 10	0,044
	30; 20	0,046		30; 20	0,042
	25; 25	0,048		25; 25	0,043
	30; 30	0,048		30; 30	0,042
	40; 40	0,051		40; 40	0,042
	50; 50	0,047		50; 50	0,039
t(5)	10; 10	0,044	Chi(5)	10; 10	0,040
	20; 10	0,042		20; 10	0,043
	20; 20	0,046		20; 20	0,040
	30; 10	0,041		30; 10	0,039
	30; 20	0,046		30; 20	0,043
	25; 25	0,048		25; 25	0,042
	30; 30	0,043		30; 30	0,043
	40; 40	0,046		40; 40	0,040
	50; 50	0,050		50; 50	0,039

Verteilung	n_1, n_2	Simuliertes Niveau	Verteilung	n_1, n_2	Simuliertes Niveau
t(10)	10; 10	0,041	Chi(10)	10; 10	0,044
	20; 10	0,040		20; 10	0,042
	20; 20	0,045		20; 20	0,041
	30; 10	0,046		30; 10	0,043
	30; 20	0,045		30; 20	0,045
	25; 25	0,046		25; 25	0,046
	30; 30	0,048		30; 30	0,038
	40; 40	0,045		40; 40	0,042
	50; 50	0,051		50; 50	0,049
Lpl	10; 10	0,054	B(8;1)	10; 10	0,053
	20; 10	0,056		20; 10	0,045
	20; 20	0,055		20; 20	0,048
	30; 10	0,057		30; 10	0,042
	30; 20	0,058		30; 20	0,047
	25; 25	0,057		25; 25	0,041
	30; 30	0,053		30; 30	0,040
	40; 40	0,047		40; 40	0,042
	50; 50	0,048		50; 50	0,038
B(3;3)	10; 10	0,032	CN(0,9;3)	10; 10	0,024
	20; 10	0,037		20; 10	0,022
	20; 20	0,042		20; 20	0,018
	30; 10	0,039		30; 10	0,019
	30; 20	0,038		30; 20	0,020
	25; 25	0,039		25; 25	0,019
	30; 30	0,041		30; 30	0,015
	40; 40	0,044		40; 40	0,020
	50; 50	0,046		50; 50	0,017

Verteilung	n_1, n_2	Simuliertes Niveau	Verteilung	n_1, n_2	Simuliertes Niveau
U(0;1)	10; 10	0,030	CN(0,8;3)	10; 10	0,022
	20; 10	0,032		20; 10	0,019
	20; 20	0,031		20; 20	0,020
	30; 10	0,034		30; 10	0,017
	30; 20	0,034		30; 20	0,020
	25; 25	0,034		25; 25	0,021
	30; 30	0,037		30; 30	0,017
	40; 40	0,043		40; 40	0,023
	50; 50	0,043		50; 50	0,020

Wie in Tabelle 1 gezeigt, sind die simulierten Signifikanzniveaus des Bonett-Tests bei Stichproben mit kleineren Umfängen aus symmetrischen oder nahezu symmetrischen Verteilungen mit schwächer bis gemäßigt besetzten Randbereichen niedriger als das Soll-Signifikanzniveau (0,05). Andererseits sind die simulierten Niveaus tendenziell etwas größer als das Soll-Niveau, wenn kleine Stichproben aus stark schiefen Verteilungen stammen.

Bei mittleren bzw. großen Stichproben liegen die simulierten Signifikanzniveaus für alle Verteilungen nage beim Soll-Niveau. Tatsächlich zeigt der Test selbst für stark schiefe Verteilungen eine recht gute Leistung, z. B. für die Exponentialverteilung und die Betaverteilung (8;1).

Außerdem haben die Ausreißer in kleinen Stichproben anscheinend eine stärkere Auswirkung als in großen Stichproben. Die simulierten Signifikanzniveaus für die kontaminierten normalverteilten Grundgesamtheiten stabilisierten sich bei etwa 0,020, als der minimale Umfang der beiden Stichproben 20 erreichte.

Wenn der minimale Umfang der zwei Stichproben 20 ist, fallen die simulierten Signifikanzniveaus durchgängig in das Intervall [0,038; 0,058], außer bei der flachen Gleichverteilung und den kontaminierten Normalverteilungen. Obwohl ein simuliertes Signifikanzniveau von 0,040 für ein Soll-Niveau von 0,05 etwas konservativ ist, kann diese Wahrscheinlichkeit eines Fehlers 1. Art für die meisten praktischen Zwecke als akzeptabel erachtet werden. Daher kann geschlussfolgert werden, dass der Bonett-Test gültig ist, wenn der minimale Umfang der zwei Stichproben mindestens 20 beträgt.

Simulation B2: Gültigkeit des Mehrfachvergleichstests (Modelle mit mehreren Stichproben)

Teil I: Balancierte Designs

Wir haben eine Simulation durchgeführt, um die Leistung des Mehrfachvergleichstests in Modellen mit mehreren Stichproben mit balancierten Designs zu untersuchen. Dabei wurden unter Verwendung der bereits in Simulation B1 aufgeführten Verteilungen jeweils k Stichproben mit gleichem Umfang aus derselben Verteilung generiert. Für die Anzahl der Stichproben in einem Design wurde $k = 3$, $k = 4$ und $k = 6$ gewählt, und der Umfang der k Stichproben in jedem Experiment wurde auf 10, 15, 20, 25, 50 und 100 festgelegt.

Für dieselben Stichproben jedes Design-Falls wurde jeweils ein beidseitiger Mehrfachvergleichstest mit einem Soll-Signifikanzniveau von $\alpha = 0{,}05$ ausgeführt. Da die simulierten Signifikanzniveaus in jedem Fall auf 10.000 Paaren von Stichprobenreplikationen basierten und ein Soll-Signifikanzniveau von 5 % angesetzt wurde, betrug der Simulationsfehler $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$.

Die nachfolgenden Tabellen 2a und 2b enthalten eine Zusammenfassung der Simulationsergebnisse.

Tabelle 2a Simulierte Signifikanzniveaus für einen beidseitigen Mehrfachvergleichstest in balancierten Designs mit mehreren Stichproben. Das Soll-Signifikanzniveau für den Test ist 0,05.

Verteilung	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau
N(0;1)	10	0,038	10	0,038	10	0,036
	15	0,040	15	0,041	15	0,039
	20	0,039	20	0,040	20	0,041
	25	0,045	25	0,047	25	0,047
	50	0,046	50	0,046	50	0,052
	100	0,049	100	0,049	100	0,052
t(5)	10	0,042	10	0,044	10	0,042
	15	0,041	15	0,044	15	0,046
	20	0,043	20	0,045	20	0,045
	25	0,046	25	0,048	25	0,046

Verteilung	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau
	50	0,040	50	0,039	50	0,038
	100	0,038	100	0,040	100	0,040
T(10)	10	0,033	10	0,037	10	0,038
	15	0,040	15	0,042	15	0,041
	20	0,042	20	0,043	20	0,043
	25	0,041	25	0,042	25	0,045
	50	0,047	50	0,044	50	0,047
	100	0,048	100	0,046	100	0,047
Lpl	10	0,056	10	0,063	10	0,071
	15	0,056	15	0,061	15	0,063
	20	0,054	20	0,058	20	0,059
	25	0,051	25	0,056	25	0,58
	50	0,045	50	0,051	50	0,049
	100	0,044	100	0,047	100	0,050
B(3;3)	10	0,031	10	0,031	10	0,031
	15	0,037	15	0,036	15	0,034
	20	0,035	20	0,036	20	0,037
	25	0,039	25	0,038	25	0,040
	50	0,044	50	0,044	50	0,044
	100	0,044	100	0,046	100	0,043
U(0;1)	10	0,029	10	0,025	10	0,023
	15	0,026	15	0,027	15	0,026
	20	0,028	20	0,030	20	0,028
	25	0,034	25	0,033	25	0,032
	50	0,041	50	0,036	50	0,036
	100	0,048	100	0,047	100	0,045

Verteilung	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau
Exp	10	0,063	10	0,073	10	0,076
	15	0,056	15	0,058	15	0,064
	20	0,051	20	0,053	20	0,057
	25	0,043	25	0,045	25	0,050
	50	0,033	50	0,037	50	0,038
	100	0,033	100	0,035	100	0,035

Tabelle 2b Simulierte Signifikanzniveaus für einen beidseitigen Mehrfachvergleichstest in balancierten Designs mit mehreren Stichproben. Das Soll-Signifikanzniveau für den Test ist 0,05.

Verteilung	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau
Chi(5)	10	0,040	10	0,046	10	0,048
	15	0,043	15	0,046	15	0,049
	20	0,040	20	0,040	20	0,042
	25	0,040	25	0,045	25	0,042
	50	0,037	50	0,038	50	0,040
	100	0,036	100	0,037	100	0,038
Chi(10)	10	0,042	10	0,045	10	0,045
	15	0,038	15	0,044	15	0,047
	20	0,036	20	0,039	20	0,040
	25	0,043	25	0,044	25	0,045
	50	0,041	50	0,040	50	0,042
	100	0,038	100	0,040	100	0,042
B(8;1)	10	0,058	10	0,060	10	0,066
	15	0,057	15	0,061	15	0,064

Verteilung	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau	n_i	Simuliertes Niveau
	20	0,049	20	0,051	20	0,055
	25	0,044	25	0,046	25	0,050
	50	0,037	50	0,037	50	0,039
	100	0,037	100	0,038	100	0,039
CN(0,9;3)	10	0,020	10	0,018	10	0,016
	15	0,022	15	0,020	15	0,017
	20	0,014	20	0,012	20	0,008
	25	0,011	25	0,011	25	0,008
	50	0,009	50	0,007	50	0,006
	100	0,010	100	0,008	100	0,008
CN(0,8;3)	10	0,017	10	0,015	10	0,011
	15	0,013	15	0,011	15	0,008
	20	0,012	20	0,012	20	0,009
	25	0,013	25	0,010	25	0,009
	50	0,011	50	0,011	50	0,009
	100	0,014	100	0,012	100	0,010

Wie in den Tabellen 2a und 2b gezeigt, ist der Mehrfachvergleichstest für symmetrische und nahezu symmetrische Verteilungen in balancierten Designs bei kleinem Stichprobenumfang generell konservativ. Anderseits ist der Test liberal für kleine Stichproben, die aus stark schiefen Verteilungen stammen, z. B. aus der Exponentialverteilung und der Betaverteilung (8;1). Mit zunehmendem Stichprobenumfang nähern sich die simulierten Signifikanzniveaus jedoch dem Soll-Signifikanzniveau (0,05). Außerdem scheint die Anzahl der Stichproben bei Stichproben mittleren Umfangs keinen starken Effekt auf die Leistung des Tests zu haben. Wenn die Daten mit Ausreißern kontaminiert sind, ist jedoch eine bedeutsame Auswirkung auf die Leistung des Tests zu verzeichnen. Der Test ist durchgehend und übermäßig konservativ, wenn Ausreißer in den Daten vorhanden sind.

Teil II: Nicht balancierte Designs

Wir haben eine Simulation durchgeführt, um die Leistung des Mehrfachvergleichstests in nicht balancierten Designs zu untersuchen. Dabei wurden unter Verwendung der bereits in Simulation B1 aufgeführten Verteilungen jeweils 3 Stichproben aus derselben Verteilung

generiert. In der ersten Reihe von Experimenten betrug der Umfang der ersten beiden Stichproben $n_1 = n_2 = 10$, während der Umfang der dritten Stichprobe $n_3 = 15, 20, 25, 50, 100$ betrug. In der zweiten Reihe von Experimenten betrug der Umfang der ersten beiden Stichproben $n_1 = n_2 = 15$, und der Umfang der dritten Gruppe von Stichproben war $n_3 = 20, 25, 30, 50, 100$. In der dritten Reihe von Experimenten wurde der minimale Stichprobenumfang auf 20 festgelegt, wobei der Umfang der ersten zwei Stichproben auf $n_1 = n_2 = 20$ und der Umfang der dritten Stichprobe auf $n_3 = 25, 30, 40, 50, 100$. festgelegt wurde.

Für die gleichen drei Stichproben aus jeder Verteilung wurde ein beidseitiger Mehrfachvergleichstest mit einem Soll-Signifikanzniveau von $\alpha = 0{,}05$ durchgeführt. Da die simulierten Signifikanzniveaus in jedem Fall auf 10.000 Paaren von Stichprobenreplikationen basierten und ein Soll-Signifikanzniveau von 5 % angesetzt wurde, betrug der Simulationsfehler $\sqrt{0{,}95(0{,}05)/10.000} = 0{,}2\%$..

Die nachfolgenden Tabellen 3a und 3b enthalten eine Zusammenfassung der Simulationsergebnisse.

Tabelle 3a Simulierte Signifikanzniveaus für den Mehrfachvergleichstest in nicht balancierten Designs mit mehreren Stichproben. Das Soll-Signifikanzniveau des Tests ist 0,05.

Verteilung	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau
N(0;1)	15	0,032	20	0,040	25	0,045
	20	0,037	25	0,039	30	0,041
	25	0,038	30	0,037	40	0,043
	50	0,041	50	0,044	50	0,041
	100	0,042	100	0,042	100	0,044
t(5)	15	0,040	20	0,042	25	0,043
	20	0,036	25	0,040	30	0,037
	25	0,044	30	0,036	40	0,038
	50	0,033	50	0,036	50	0,035
	100	0,032	100	0,031	100	0,032
t(10)	15	0,039	20	0,042	25	0,042
	20	0,038	25	0,041	30	0,040
	25	0,040	30	0,041	40	0,041
	50	0,037	50	0,043	50	0,042
	100	0,036	100	0,039	100	0,040

Verteilung	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau
Lpl	15	0,059	20	0,060	25	0,054
	20	0,057	25	0,054	30	0,051
	25	0,056	30	0,051	40	0,050
	50	0,049	50	0,051	50	0,050
	100	0,048	100	0,047	100	0,046
B(3;3)	15	0,034	20	0,033	25	0,037
	20	0,031	25	0,035	30	0,039
	25	0,031	30	0,034	40	0,039
	50	0,036	50	0,039	50	0,038
	100	0,035	100	0,039	100	0,039
U(0;1)	15	0,027	20	0,030	25	0,032
	20	0,030	25	0,030	30	0,031
	25	0,028	30	0,032	40	0,036
	50	0,039	50	0,034	50	0,037
	100	0,042	100	0,038	100	0,042
Exp	15	0,061	20	0,053	25	0,042
	20	0,060	25	0,052	30	0,047
	25	0,054	30	0,049	40	0,043
	50	0,050	50	0,046	50	0,041
	100	0,044	100	0,040	100	0,040

Tabelle 3b Simulierte Signifikanzniveaus für den Mehrfachvergleichstest in nicht balancierten Designs mit mehreren Stichproben. Das Soll-Signifikanzniveau des Tests ist 0,05.

Verteilung	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau
Chi(5)	15	0,047	20	0,045	25	0,041
	20	0,043	25	0,042	30	0,039

Verteilung	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau	n_3	Simuliertes Niveau
	25	0,043	30	0,039	40	0,040
	50	0,039	50	0,037	50	0,040
	100	0,034	100	0,035	100	0,034
Chi(10)	15	0,043	20	0,042	25	0,042
	20	0,039	25	0,038	30	0,041
	25	0,040	30	0,041	40	0,038
	50	0,038	50	0,041	50	0,042
	100	0,035	100	0,034	100	0,035
B(8;1)	15	0,056	20	0,052	25	0,048
	20	0,054	25	0,046	30	0,044
	25	0,050	30	0,047	40	0,046
	50	0,046	50	0,043	50	0,043
	100	0,043	100	0,042	100	0,044
CN(0,9;3)	15	0,017	20	0,020	25	0,017
	20	0,020	25	0,019	30	0,012
	25	0,017	30	0,016	40	0,013
	50	0,019	50	0,016	50	0,012
	100	0,014	100	0,016	100	0,010
CN(0,8;3)	15	0,012	20	0,013	25	0,013
	20	0,016	25	0,012	30	0,012
	25	0,014	30	0,010	40	0,010
	50	0,015	50	0,010	50	0,013
	100	0,012	100	0,011	100	0,010

Die in den Tabellen 3a und 3b aufgeführten simulierten Signifikanzniveaus stimmen mit denen überein, die zuvor für mehrere Stichproben mit balancierten Designs berechnet wurden. Daher haben nicht balancierte Designs anscheinend keine Auswirkung auf die Leistung des Mehrfachvergleichstests. Wenn der minimale Stichprobenumfang 20 beträgt,

liegen die simulierten Signifikanzniveaus zudem nahe dem Soll-Niveau; dies gilt allerdings nicht für die kontaminierten Daten.

Fazit: Wenn die kleinste Stichprobe einen Umfang von mindestens 20 aufweist, zeigt der Mehrfachvergleichstest für mehrere (k) Stichproben sowohl in balancierten als auch in nicht balancierten Designs eine gute Leistung. Bei kleineren Stichproben ist der Test jedoch konservativer für symmetrische und nahezu symmetrische Daten sowie liberal für stark schiefe Daten.

Anhang C: Theoretische Trennschärfefunktion

Die genaue theoretische Trennschärfefunktion des Mehrfachvergleichstests ist nicht verfügbar. Bei Designs mit zwei Stichproben kann jedoch anhand theoretischer Methoden für große Stichproben eine approximierte Trennschärfefunktion abgeleitet werden. Bei Designs mit mehreren Stichproben bedarf es noch weiterer Forschung, um eine vergleichbare Approximation herzuleiten.

Für Designs mit zwei Stichproben kann die theoretische Trennschärfefunktion des Bonett-Tests jedoch mit theoretischen Methoden für große Stichproben bestimmt werden. Konkreter heißt dies: Die unten angegebene Teststatistik T ist asymptotisch nach einer Chi-Quadrat-Verteilung mit 1 Freiheitsgrad verteilt:

$$T = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

In diesem Ausdruck von T ist $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, $g_i = (n_i - 3)/n_i$, und γ ist die unbekannte gemeinsame Kurtosis der beiden Grundgesamtheiten.

Daraus folgt, dass die theoretische Trennschärfefunktion eines beidseitigen Bonett-Tests auf Gleichheit von Varianzen mit einem approximierten Signifikanzniveau α wie folgt angegeben werden kann:

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

Dabei gilt Folgendes:

$$se = \sqrt{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

Für einseitige Tests ist die approximierte Trennschärfefunktion beim Testen auf $\sigma_1 > \sigma_2$

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

Beim Testen auf $\sigma_1 < \sigma_2$ lautet die approximierte Trennschärfefunktion

$$\pi(n_1, n_2, \rho) = \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

Beachten Sie, dass die gemeinsame Kurtosis der Grundgesamtheiten γ in der Phase zur Planung des Stichprobenumfangs für die Datenanalyse unbekannt ist. Daher muss man sich i. d. R. auf Fachwissen oder die Ergebnisse früherer Experimente stützen, um einen Planwert für γ zu erhalten. Sollten derartige Informationen nicht zur Verfügung stehen, empfiehlt es sich häufig, eine kleine Pilotstudie durchzuführen, um die Pläne für die Hauptuntersuchung auszuarbeiten. Auf der Grundlage der Stichproben aus der Pilotstudie wird der Planwert von γberechnet als zusammengefasste Kurtosis, die ausgedrückt wird als

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

Im Menü des Assistenten wird der Planschätzwert von γ rückwirkend auf Grundlage der verfügbaren Daten des Benutzers bestimmt.

Anhang D: Vergleich der theoretischen und der simulierten Trennschärfe

Simulation D1: Simulierte (tatsächliche) Trennschärfe des Bonett-Tests

Wir haben eine Simulation durchgeführt, um die simulierten Trennschärfen des Bonett-Tests mit den Trennschärfen zu vergleichen, die auf der approximierten Trennschärfefunktion basieren, deren Ableitung in Anhang C gezeigt wird.

Dabei wurden 10.000 Paare von Stichproben für jede der oben beschriebenen Verteilungen generiert (siehe Simulation B1). Im Allgemeinen waren die ausgewählten Stichprobenumfänge auf der Grundlage der Ergebnisse aus der vorausgegangenen Simulation B1 groß genug, dass das simulierte Signifikanzniveau des Tests hinreichend nahe am Soll-Signifikanzniveau lag.

Zum Untersuchen der simulierten Trennschärfen bei einem Verhältnis der Standardabweichungen von $\rho = \sigma_1/\sigma_2 = 1/2$ wurde die zweite Stichprobe in jedem Paar von Stichproben mit der Konstanten 2 multipliziert. Als Ergebnis wurde die simulierte Trennschärfe für eine bestimmte Verteilung und für die angegebenen Stichprobenumfänge n_1 und n_2 als der Anteil der 10.000 Paare von Stichprobenreplikationen berechnet, für den der beidseitige Bonett-Test signifikant war. Das Soll-Signifikanzniveau des Tests wurde auf $\alpha = 0{,}05$ festgelegt. Zum Vergleich wurden die entsprechenden theoretischen Trennschärfen auf der Grundlage der approximierten Trennschärfefunktion berechnet, deren Ableitung in Anhang C gezeigt wird.

Die Ergebnisse werden in der nachfolgenden Tabelle 4 aufgeführt.

Tabelle 4 Vergleich der simulierten Trennschärfen mit den approximierten Trennschärfen eines beidseitigen Bonett-Tests. Das Soll-Signifikanzniveau ist 0,05.

Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe	Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe
N(0;1)	20; 10	0,627	0,527	Exp	20; 10	0,222	0,227
	20; 20	0,830	0,765		20; 20	0,322	0,368
	20; 30	0,896	0,846		20; 30	0,377	0,434
	20; 40	0,925	0,886		20; 40	0,412	0,475
	30; 15	0,825	0,771		30; 15	0,320	0,307
	30; 30	0,954	0,925		30; 30	0,458	0,500

Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe	Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe
	30; 45	0,980	0,970		30; 45	0,531	0,579
	30; 60	0,989	0,984		30; 60	0,575	0,622
t(5)	20; 10	0,222	0,379	Chi(5)	20; 10	0,355	0,347
	20; 20	0,322	0,569		20; 20	0,517	0,530
	20; 30	0,377	0,637		20; 30	0,597	0,616
	20; 40	0,412	0,690		20; 40	0,644	0,661
	30; 15	0,320	0,545		30; 15	0,513	0,510
	30; 30	0,458	0,733		30; 30	0,701	0,711
	30; 45	0,531	0,795		30; 45	0,781	0,793
	30; 60	0,575	0,828		30; 60	0,823	0,833
t(10)	20; 10	0,476	0,450	Chi(10)	20; 10	0,454	0,414
	20; 20	0,673	0,673		20; 20	0,646	0,631
	20; 30	0,756	0,749		20; 30	0,730	0,717
	20; 40	0,800	0,803		20; 40	0,776	0,771
	30; 15	0,668	0,659		30; 15	0,641	0,618
	30; 30	0,850	0,852		30; 30	0,828	0,819
	30; 45	0,910	0,911		30; 45	0,892	0,882
	30; 60	0,936	0,937		30; 60	0,921	0,912
Lpl	20; 10	0,321	0,330	B(8;1)	20; 10	0,363	0,278
	20; 20	0,469	0,519		20; 20	0,528	0,463
	20; 30	0,545	0,585		20; 30	0,609	0,549
	20; 40	0,590	0,632		20; 40	0,655	0,600
	30; 15	0,466	0,475		30; 15	0,524	0,419
	30; 30	0,647	0,673		30; 30	0,713	0,634
	30; 45	0,729	0,758		30; 45	0,792	0,737
	30; 60	0,773	0,800		30; 60	0,833	0,777
B(3;3)	20; 10	0,777	0,628	CN(0,9;3)	20; 10	0,238	0,284
	20; 20	0,939	0,869		20; 20	0,346	0,452

Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe	Verteilung	n_1, n_2	Approx. Trennschärfe	Simulierte Trennschärfe
	20; 30	0,973	0,936		20; 30	0,405	0,517
	20; 40	0,984	0,964		20; 40	0,442	0,561
	30; 15	0,935	0,871		30; 15	0,343	0,374
	30; 30	0,993	0,980		30; 30	0,491	0,598
	30; 45	0,998	0,995		30; 45	0,567	0,700
	30; 60	0,999	0,999		30; 60	0,612	0,719
U(0;1)	20; 10	0,916	0,740	CN(0,8;3)	20; 10	0,260	0,223
	20; 20	0,992	0,950		20; 20	0,379	0,396
	20; 30	0,998	0,985		20; 30	0,444	0,467
	20; 40	0,999	0,995		20; 40	0,484	0,520
	30; 15	0,991	0,941		30; 15	0,376	0,354
	30; 30	1,0	0,996		30; 30	0,535	0,549
	30; 45	1,0	1,0		30; 45	0,614	0,650
	30; 60	1,0	1,0		30; 60	0,661	0,706

Die Ergebnisse zeigen, dass die approximierten Trennschärfen und die simulierten Trennschärfen im Allgemeinen nahe beieinander liegen. Mit zunehmendem Stichprobenumfang nähern sie sich einander an. Bei symmetrischen und nahezu symmetrischen Verteilungen mit gemäßigt bis schwach besetzten Randbereichen sind die approximierten Trennschärfen i. d. R. etwas größer als die simulierten Trennschärfen. Bei symmetrischen Verteilungen mit stärker besetzten Randbereichen oder für stark schiefe Verteilungen sind sie jedoch etwas kleiner als die simulierten Trennschärfen. Die Differenz zwischen den beiden Trennschärfefunktionen ist i. d. R. nicht wichtig; eine Ausnahme stellt der Fall dar, in dem die Stichproben aus der t-Verteilung mit 5 Freiheitsgraden generiert wurden.

Generell wurde Folgendes festgestellt: Wenn der minimale Stichprobenumfang 20 erreicht, liegen die approximierten Trennschärfen und die simulierten Trennschärfen auffallend dicht beieinander. Somit können der Planung der Stichprobenumfänge die approximierten Trennschärfefunktionen zugrunde gelegt werden.