

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.

# Test auf Standardabweichung bei einer Stichprobe

## Übersicht

Mit dem Test auf Standardabweichung bei einer Stichprobe kann die Streuung Ihres Prozesses geschätzt und die Streuung mit einem Sollwert verglichen werden. Typischerweise wird die Streuung anhand der Varianz bzw. (äquivalent) der Standardabweichung gemessen.

Es gibt eine Vielzahl von statistischen Methoden zum Auswerten der Varianz einer Grundgesamtheit, von denen jede spezifische Vorteile und Einschränkungen aufweist. Die klassische Chi-Quadrat-Methode zum Testen der Varianz ist wahrscheinlich die gängigste Methode. Sie ist jedoch extrem empfindlich gegenüber der Annahme der Normalverteilung und kann äußerst ungenaue Ergebnisse liefern, wenn die Daten schief sind oder stärker besetzte Randbereiche aufweisen. Auch die anderen Methoden haben Nachteile. Einige Methoden können beispielsweise nur für große Stichproben oder für Daten aus einer symmetrischen Verteilung verwendet werden (siehe Anhang A).

In Minitab 15 wird eine alternative Methode für große Stichproben verwendet, die aus einer skalierten Chi-Quadrat-Approximation an die Verteilung der Stichprobenvarianz nach Box (1953) abgeleitet wurde. Diese Methode, die als Methode der korrigierten Freiheitsgrade (KorDF) bezeichnet wird, ist bei ausreichend großen Stichproben weniger empfindlich gegenüber der Annahme der Normalverteilung, und sie liefert nachweislich genauere Konfidenzintervalle als andere Methoden (Hummel, Banga und Hettmansperger, 2005). Vor Kurzem hat jedoch Bonett (2006) eine überarbeitete statistische Methode entwickelt, die anscheinend bessere approximiertere Konfidenzintervalle produziert.

Im vorliegenden Artikel soll die Leistung der Bonett-Methode ausgewertet werden. Für die Planung der Stichprobenumfänge wird darüber hinaus die Trennschärfefunktion für das äquivalente Testverfahren in Bezug auf Bonett-Konfidenzintervalle untersucht. Auf der Grundlage unserer Ergebnisse wird die Bonett-Methode für den Test auf Standardabweichung bei einer Stichprobe im Assistenten verwendet. Zudem werden die folgenden Datenprüfungen untersucht, die automatisch ausgeführt und in der Auswertung des Assistenten angezeigt werden, und wir erläutern deren Auswirkungen auf die Ergebnisse:

- Ungewöhnliche Daten
- Gültigkeit des Tests
- Stichprobenumfang

# Methoden für den Test auf Standardabweichung bei einer Stichprobe

## Bonett-Methode im Vergleich mit der KorDF-Methode

Vor der Veröffentlichung der Bonett-Methode (2006) war wahrscheinlich die KorDF-Methode das robusteste Verfahren zum Ziehen von Rückschlüssen auf die Varianz einer Grundgesamtheit. Die von Bonett veröffentlichten Ergebnisse zeigen jedoch, dass die Bonett-Methode stabile Konfidenzniveaus liefert, die nahe dem Sollniveau liegen, wenn Stichproben mit mittlerem Umfang aus nicht normalverteilten Grundgesamtheiten gezogen werden. Daher ist ggf. die Bonett-Methode zu bevorzugen, wenn Rückschlüsse bezüglich der Standardabweichung oder Varianz einer Grundgesamtheit gezogen werden sollen.

### Zielstellung

Wir wollten die Leistung der Bonett-Methode in Bezug auf das Ziehen von Rückschlüssen über die Varianz einer einzelnen Grundgesamtheit mit der der KorDF-Methode vergleichen. Insbesondere sollte dabei festgestellt werden, welche Methode genauere Konfidenzintervalle für die Varianz (bzw. die Standardabweichung) produziert, wenn Stichproben unterschiedlicher Umfänge aus nicht normalverteilten Grundgesamtheiten generiert werden.

Es werden Konfidenzintervalle verglichen, da sich die Bonett-Methode direkt auf Konfidenzintervalle bezieht. Das äquivalente hypothetische Testverfahren für Bonett-Konfidenzintervalle kann abgeleitet werden. Um jedoch einen direkten Vergleich unserer Ergebnisse mit den in Bonett (2006) veröffentlichten zu ermöglichen, wurden die Konfidenzintervalle anstelle der Hypothesentests untersucht.

### Methode

Formale Definitionen der KorDF-Methode und der Bonett-Methode werden in Anhang B dargestellt. Zum Vergleichen der Genauigkeit der Konfidenzintervalle für die beiden Methoden wurden die folgenden Simulationen durchgeführt. Zunächst wurden Zufallsstichproben diverser Umfänge aus Verteilungen mit unterschiedlichen Eigenschaften generiert, z. B. aus schiefen Verteilungen mit stärker besetzten Randbereichen, symmetrischen Verteilungen mit stärker besetzten Randbereichen sowie symmetrischen Verteilungen mit schwächer besetzten Randbereichen. Für jeden Stichprobenumfang wurden 10.000 Stichprobenreplikationen aus den einzelnen Verteilungen gezogen, und mit jeder der beiden Methoden wurden beidseitige 95%-Konfidenzintervalle für die tatsächliche Varianz der Verteilung berechnet. Anschließend wurde der Anteil der 10.000 Intervalle berechnet, die die tatsächliche Varianz enthielten. Dieser wird als simulierte Überdeckungswahrscheinlichkeit bezeichnet. Wenn die Konfidenzintervalle genau sind, sollte

die simulierte Überdeckungswahrscheinlichkeit nahe ihrem Sollwert von 0,95 liegen. Darüber hinaus wurden die durchschnittlichen Breiten für die Konfidenzintervalle der einzelnen Methoden berechnet. Wenn die Konfidenzintervalle der beiden Methoden etwa die gleiche simulierte Überdeckungswahrscheinlichkeit aufweisen, ist die Methode präziser, die (im Durchschnitt) schmalere Intervalle produziert. Weitere Einzelheiten finden Sie in Anhang C.

## Ergebnisse

Die Bonett-Methode liefert generell bessere Überdeckungswahrscheinlichkeiten und präzisere Konfidenzintervalle als die KorDF-Methode. Demzufolge weisen die statistischen Tests auf Varianz auf der Grundlage der Bonett-Methode niedrigere Wahrscheinlichkeiten eines Fehlers 1. Art und 2. Art auf. Daher basiert der Test auf Standardabweichung bei einer Stichprobe im Assistenten auf der Bonett-Methode.

Zudem zeigen unsere Ergebnisse, dass die Bonett-Methode bei mittleren bis stärker besetzten Randbereichen der Verteilung größere Stichprobenumfänge benötigt, um die Sollgenauigkeit zu erzielen:

- Bei Verteilungen mit normal oder schwächer besetzten Randbereichen ist ein Stichprobenumfang von 20 ausreichend.
- Bei Verteilungen mit mittelstark besetzten Randbereichen sollte der Stichprobenumfang mindestens 80 betragen.
- Bei Verteilungen mit stärker besetzten Randbereichen sollte der Stichprobenumfang mindestens 200 betragen.

Um also sicherzustellen, dass die Ergebnisse des Tests auf Standardabweichung bei einer Stichprobe bzw. die Konfidenzintervalle für Ihre Daten gültig sind, enthält der Assistent eine Datenprüfung, bei der gleichzeitig der Stichprobenumfang und die Randbereiche der Datenverteilung untersucht werden (siehe „Gültigkeit des Tests“ im nachfolgenden Abschnitt „Datenprüfungen“).

## Leistung der theoretischen Trennschärfe

Die Bonett-Methode bezieht sich direkt auf Konfidenzintervalle für die Varianz (bzw. Standardabweichung). Anhand der statistischen Beziehung zwischen Hypothesentest und Konfidenzintervallen können wir jedoch den äquivalenten Test für die approximierten Konfidenzintervalle nach Bonett ableiten. Da keine exakte Trennschärfefunktion für diesen Test verfügbar ist, musste sie abgeleitet werden. Außerdem sollte die Empfindlichkeit der theoretischen Trennschärfefunktion gegenüber der Annahme der Normalverteilung untersucht werden.

## Zielstellung

Wir wollten feststellen, ob die Anforderungen an Trennschärfe und Stichprobenumfang für den Test auf Standardabweichung bei einer Stichprobe im Assistenten anhand der theoretischen Trennschärfefunktion des Tests für die Bonett-Konfidenzintervalle ausgewertet werden können. Hierzu musste ausgewertet werden, ob diese theoretische Trennschärfefunktion die tatsächliche Trennschärfe des Tests bei der Analyse normalverteilter und nicht normalverteilter Daten genau widerspiegelt.

## Methode

Die Ableitung der theoretischen Trennschärfefunktion des Tests unter Verwendung der Bonett-Methode wird in Anhang C dargestellt. Wir haben Simulationen zum Schätzen der tatsächlichen Trennschärfen (die wir als simulierte Trennschärfen bezeichnen) mit Hilfe der Bonett-Methode durchgeführt. Zunächst wurden Zufallsstichproben mit unterschiedlichen Stichprobenumfängen aus den in der vorherigen Studie beschriebenen Verteilungen generiert: schiefe Verteilungen mit stärker besetzten Randbereichen, symmetrische Verteilungen mit stärker besetzten Randbereichen sowie symmetrische Verteilungen mit schwächer besetzten Randbereichen. Für jede Verteilung wurde der Test an jeder der 10.000 Stichprobenreplikationen durchgeführt. Für jeden Stichprobenumfang wurde die simulierte Trennschärfe des Tests, eine gegebene Differenz zu erkennen, als der Anteil der 10.000 Stichproben berechnet, für die der Test signifikant ist. Zu Vergleichszwecken wurde auch die entsprechende Trennschärfe mit der theoretischen Trennschärfefunktion des Tests berechnet. Wenn die theoretische Trennschärfefunktion nicht zu empfindlich gegenüber der Normalverteilung ist, sollten die theoretische und die simulierte Trennschärfe für normalverteilte und nicht normalverteilte Daten nahe beieinander liegen. Weitere Einzelheiten finden Sie in Anhang D.

## Ergebnisse

Unsere Simulationen zeigen Folgendes: Wenn die Stichproben aus einer Verteilung mit normal oder schwächer besetzten Randbereichen stammen, sind die theoretische und die simulierte Trennschärfe des Tests mit der Bonett-Methode nahezu gleich. Wenn die Stichprobe hingegen aus einer Verteilung mit stärker besetzten Randbereichen stammt, ist die theoretische Trennschärfefunktion möglicherweise konservativ und setzt den erforderlichen Stichprobenumfang zum Erzielen einer bestimmten Trennschärfe zu hoch an. Daher stellt die theoretische Trennschärfefunktion des Tests sicher, dass der Stichprobenumfang ausreichend groß ist, dass eine Differenz mit praktischen Konsequenzen in der Standardabweichung unabhängig von der Verteilung erkannt wird. Wenn die Daten jedoch aus Verteilungen mit stärker besetzten Randbereichen stammen, ist der geschätzte Stichprobenumfang u. U. größer als der tatsächlich benötigte Stichprobenumfang, wodurch die Stichprobengewinnung möglicherweise zu aufwändig wird.

# 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 wichtig, diese Datenpunkte zu untersuchen und nach Möglichkeit zu korrigieren.

### Zielstellung

Wir wollten eine Methode zum Überprüfen von Datenwerten entwickeln, die in Bezug auf die Gesamtstichprobe sehr groß bzw. sehr klein sind und die 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 basiert, mit der Ausreißer in Boxplots bestimmt werden.

### 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 übt möglicherweise einen starken Einfluss auf die Ergebnisse aus.

## Gültigkeit des Tests

Im vorausgegangenen Abschnitt „Methoden für den Test auf Standardabweichung bei einer Stichprobe“ wurde gezeigt, dass die Bonett-Methode generell bessere Ergebnisse als die KorDF-Methode liefert. Wenn die Randbereiche einer Verteilung jedoch stärker besetzt sind, werden für die Bonett-Methode größere Stichprobenumfänge benötigt, um genaue Ergebnisse zu erzielen. Daher muss einer Methode zum Auswerten der Gültigkeit des Tests nicht nur der Stichprobenumfang zugrunde liegen, sondern auch die Besetzung der Randbereiche der übergeordneten Verteilung. Gel et al. (2007) haben einen Test entwickelt, mit dem bestimmt wird, ob eine Stichprobe aus einer Verteilung mit stärker besetzten Randbereichen stammt. Dieser Test wird als SJ-Test bezeichnet. Er basiert auf dem Verhältnis der Standardabweichung der Stichprobe ( $s$ ) und des Schätzwerts für die Randbereiche  $J$  (Einzelheiten finden Sie in Anhang E).

### Zielstellung

Für eine gegebene Stichprobe von Daten musste eine Regel entwickelt werden, anhand derer die Gültigkeit der Bonett-Methode durch Auswerten der Besetzung der Randbereiche in den Daten beurteilt werden kann.



### Methode

Wir haben Simulationen durchgeführt, um die Trennschärfe des SJ-Tests zum Identifizieren von Verteilungen mit stärker besetzten Randbereichen zu untersuchen. Wenn der SJ-Test eine gute Trennschärfe für mittelgroße Stichproben aufweist, kann er für unsere Zwecke genutzt werden, um zwischen Verteilungen mit stärker und schwächer besetzten Randbereichen zu unterscheiden. Weitere Einzelheiten finden Sie in Anhang F.

### Ergebnisse

Unsere Simulationen zeigten, dass der SJ-Test bei ausreichendem Umfang der Stichproben verwendet werden kann, um zwischen Verteilungen mit stärker und schwächer besetzten Randbereichen zu unterscheiden. Bei mittelgroßen bzw. großen Stichproben weisen kleinere  $p$ -Werte auf stärker besetzte Randbereiche und größere  $p$ -Werte auf schwächer besetzte Randbereiche hin. Da größere Stichproben jedoch tendenziell kleinere  $p$ -Werte als kleinere Stichproben aufweisen, wird beim Bestimmen der Besetzung von Randbereichen auch der Stichprobenumfang betrachtet. Daher wurde eine Reihe von Regeln für den Assistenten entwickelt, anhand derer die Randbereiche der Verteilung für jede Stichprobe auf der Grundlage sowohl des Stichprobenumfangs als auch des  $p$ -Werts des SJ-Tests klassifiziert werden. Die konkreten Bereiche der  $p$ -Werte und Stichprobenumfänge für Verteilungen mit schwächer, mittelstark und stärker besetzten Randbereichen finden Sie in Anhang F.

Auf der Grundlage dieser Ergebnisse werden in der Auswertung des Assistenten die folgenden Statusindikatoren angezeigt, um die Gültigkeit des Tests auf Standardabweichung bei einer Stichprobe (Bonett-Methode) für Ihre Stichprobendaten auszuwerten:

Status	Bedingung
	<p>Es liegen keine Anzeichen dafür vor, dass Ihre Stichprobe stärker besetzte Randbereiche aufweist. Der Stichprobenumfang ist ausreichend groß, um eine zuverlässige Prüfung auf diesen Zustand durchzuführen.</p> <p>ODER</p> <p>Ihre Stichprobe weist mittelstark bzw. stärker besetzte Randbereiche auf. Der Stichprobenumfang ist jedoch ausreichend groß, um dies auszugleichen. Daher sollte der p-Wert genau sein.</p>
	<p>Ihre Stichprobe weist mittelstark bzw. stärker besetzte Randbereiche auf. Der Stichprobenumfang ist nicht groß genug, um dies auszugleichen. Interpretieren Sie die Ergebnisse mit Vorsicht.</p> <p>ODER</p> <p>Die Stichprobe ist nicht groß genug, um eine zuverlässige Prüfung auf stärker besetzte Randbereiche durchzuführen. Interpretieren Sie die Ergebnisse mit Vorsicht.</p>

## Stichprobenumfang

Normalerweise wird ein statistischer Hypothesentest durchgeführt, um Belege zum Zurückweisen der Nullhypothese („keine Differenz“) zu sammeln. Wenn die Stichprobe zu klein ist, reicht die Trennschärfe des Tests u. U. nicht aus, um eine tatsächlich vorhandene Differenz zu erkennen; dies ist ein Fehler 2. Art. Daher muss sichergestellt werden, dass die Stichprobenumfänge ausreichend groß sind, um Differenzen mit praktischen Konsequenzen mit hoher Wahrscheinlichkeit zu erkennen.

### Zielstellung

Wenn die Daten keine ausreichenden Hinweise zum Zurückweisen der Nullhypothese liefern, sollte bestimmt werden, ob die Stichprobenumfänge für den Test groß genug sind, um Differenzen mit praktischen Konsequenzen mit hoher Wahrscheinlichkeit zu erkennen. Bei der Planung der Stichprobenumfänge soll zwar sichergestellt werden, dass die Stichprobenumfänge ausreichend groß sind, um wichtige Differenzen mit hoher Wahrscheinlichkeit zu erkennen; andererseits dürfen sie 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 einer Stichprobe basiert auf der theoretischen Trennschärfefunktion des Tests. Diese Trennschärfefunktion erzielt gute Schätzwerte, wenn Daten nahezu normal oder schwächer besetzte Randbereiche aufweisen; sie liefert jedoch möglicherweise konservative Schätzwerte, wenn die Daten stärker besetzte Randbereiche aufweisen (siehe Zusammenfassung der Simulationsergebnisse unter „Leistung der theoretischen



Trennschärfe“ im obigen Abschnitt „Methoden für den Test auf Standardabweichung bei einer Stichprobe“).

## Ergebnisse

Wenn die Daten keine ausreichenden Hinweise zum Zurückweisen der Nullhypothese liefern, nutzt der Assistent die Trennschärfefunktion des Tests auf Normal-Approximation, um die Differenzen mit praktischen Konsequenzen zu berechnen, die für den angegebenen Stichprobenumfang mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden können. Wenn der Benutzer eine bestimmte Differenz mit praktischen Konsequenzen angibt, nutzt der Assistent zudem die Trennschärfefunktion des Tests auf Normal-Approximation, um Stichprobenumfänge zu berechnen, bei der die betreffende Differenz mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden kann.

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

Status	Bedingung
	Im Test wird eine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt, so dass die Trennschärfe unproblematisch ist.  ODER  Die Trennschärfe ist ausreichend. Im Test wurde keine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt, die Stichprobe ist jedoch groß genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 90 % erkannt wird.
	Die Trennschärfe ist möglicherweise ausreichend. Im Test wurde keine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt, die Stichprobe ist jedoch groß genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 80 % bis 90 % erkannt wird. Der erforderliche Stichprobenumfang zum Erreichen einer Trennschärfe von 90 % wird ausgegeben.
	Die Trennschärfe ist möglicherweise nicht ausreichend. Im Test wurde keine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt, und die Stichprobe ist groß genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 60 % bis 80 % erkannt wird. Die erforderlichen Stichprobenumfänge zum Erreichen einer Trennschärfe von 80 % und 90 % werden gemeldet.
	Die Trennschärfe ist nicht ausreichend. Im Test wurde keine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt, und die Stichprobe ist nicht groß genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 60 % erkannt wird. Die erforderlichen Stichprobenumfänge zum Erreichen einer Trennschärfe von 80 % und 90 % werden ausgegeben.
	Im Test wurde keine Differenz zwischen der Standardabweichung und dem Sollwert festgestellt. Sie haben keine zu erkennende Differenz mit praktischen Konsequenzen angegeben. Abhängig von Ihren Daten werden in der Auswertung möglicherweise die Differenzen angegeben, die mit Ihrem Stichprobenumfang und dem Alpha mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden können.

# Literaturhinweise

Bonett, D. G. (2006). Approximate confidence interval for standard deviation of nonnormal distributions. *Computational Statistics & Data Analysis*, 50, 775-782.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40, 318.

Efron, B. und Tibshirani, R. J. Lawless (2002). (1993). An introduction to the bootstrap. Boca Raton, FL: Chapman and Hall/CRC.

Gel, Y. R., Miao, W. und Gastwirth, J. L. (2007). Robust directed tests of normality against heavy-tailed alternatives. *Computational Statistics & Data Analysis*, 51, 2734-2746.

Hummel, R., Banga, S. und Hettmansperger, T. P. (2005). Better confidence intervals for the variance in a random sample. *Minitab Technical Report*.

Lee, S. J. und Ping, S. (1996). Testing the variance of symmetric heavy-tailed distributions. *Journal of Statistical Computation and Simulation*, 56, 39-52.

# Anhang A: Methoden zum Testen auf Varianz (oder Standardabweichung)

In der nachfolgenden Tabelle werden die Vorteile und Nachteile der verschiedenen Methoden zum Testen auf Varianz zusammengefasst.

Methode	Kommentar
Klassisches Chi-Quadrat-Verfahren	Extrem empfindlich gegenüber der Annahme der Normalverteilung. Selbst geringfügige Abweichungen von der Normalverteilung können ungeachtet des Stichprobenumfangs ungenaue Ergebnisse produzieren. Wenn die Daten von der Normalverteilung abweichen, führt das Vergrößern des Stichprobenumfangs sogar zu einer Abnahme der Genauigkeit des Verfahrens.
Methode für große Stichproben auf der Grundlage der asymptotischen Normalverteilung der logarithmischen Transformation der Stichprobenvarianz	Im Allgemeinen besser als die klassische Chi-Quadrat-Methode, erfordert jedoch größere Stichprobenumfänge, um eine Unempfindlichkeit gegenüber der Annahme der Normalverteilung zu erreichen.
Methode für große Stichproben auf Grundlage der Edgeworth-Erweiterung für einseitige Tests (oberer Randbereich)  Siehe Lee und Ping (1996).	Weist eine akzeptable Wahrscheinlichkeit eines Fehlers 1. Art auf, setzt jedoch voraus, dass die Daten aus einer symmetrischen Verteilung stammen.
Methode für große Stichproben auf der Grundlage einer Approximation der Verteilung der Stichprobenvarianz durch eine skalierte Chi-Quadrat-Verteilung. Die Methode wird als Methode der korrigierten Freiheitsgrade (KorDF) bezeichnet.  Siehe Hummel, Banga und Hettmansperger (2005).	Liefert eine bessere Überdeckungswahrscheinlichkeit als die Methode auf der Grundlage der asymptotischen Normalverteilung der logarithmischen Transformation der Stichprobenvarianz und die nichtparametrische ABC-Bootstrap-Approximationsmethode für Konfidenzintervalle (Efron und Tibshirani, 1993).  Wird für den Test auf Varianzen bei einer Stichprobe in Minitab 15 verwendet.
Korrigierte asymptotische Verteilung der logarithmischen Transformation der Stichprobenvarianz nach Bonett  Siehe Bonett (2006).	Liefert eine gute Überdeckungswahrscheinlichkeit für Konfidenzintervalle selbst für mittelgroße Stichproben. Es sind jedoch viel größere Stichproben erforderlich, wenn die Daten aus Verteilungen mit stärker besetzten Randbereichen stammen.  Wird für den Test auf Varianzen bei einer Stichprobe und den Test auf Standardabweichung bei einer Stichprobe im Assistenten in Minitab 16 verwendet.

# Anhang B: Definition der Bonett-Methode und der KorDF-Methode

Sei  $x_1, \dots, x_n$  eine beobachtete Zufallsstichprobe mit dem Umfang  $n$  aus einer Grundgesamtheit mit einem endlichen Moment vierter Ordnung. Seien  $\bar{x}$  und  $s$  der beobachtete Mittelwert und die beobachtete Standardabweichung der Stichprobe. Seien darüber hinaus  $\gamma$  und  $\gamma_e$  die Kurtosis der Grundgesamtheit und der Kurtosis-Überschuss, so dass  $\gamma_e = \gamma - 3$ . Damit gilt für eine normalverteilte Grundgesamtheit  $\gamma = 3$  und  $\gamma_e = 0$ . Sei  $\sigma^2$  die unbekannte Varianz der Grundgesamtheit. In den nachfolgenden Abschnitten stellen wir zwei Methoden vor, mit denen ein Rückschluss auf  $\sigma^2$  gezogen werden kann, die Methode der korrigierten Freiheitsgrade (KorDF) und die Bonett-Methode.

## Formel B1: KorDF-Methode

Die KorDF-Methode basiert auf einer Approximation der Verteilung der Stichprobenvarianz durch eine skalierte Chi-Quadrat-Verteilung (siehe Box, 1953). Konkreter: Die Momente der ersten beiden Ordnungen der Stichprobenvarianz werden mit den Momenten einer skalierten Chi-Quadrat-Verteilung verglichen, um die unbekannte Skala und die Freiheitsgrade zu bestimmen. Dieser Ansatz liefert das folgende approximierete beidseitige  $(1 - \alpha)100\%$ -Konfidenzintervall für die Varianz:

$$\left[ \frac{rs^2}{\chi_{r,\alpha/2}^2}, \frac{rs^2}{\chi_{r,1-\alpha/2}^2} \right]$$

Dabei gilt Folgendes:

$$r = \frac{2n}{\hat{\gamma}_e + 2n/(n-1)}$$
$$\hat{\gamma}_e = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Dieser Schätzwert des Kurtosis-Überschusses ist identisch mit dem, der für die Befehle in „Statistische Standardverfahren“ in Minitab verwendet wird.

## Formel B2: Bonett-Methode

Die Bonett-Methode stützt sich auf den hinreichend bekannten klassischen Ansatz, bei dem anhand des zentralen Grenzwertsatzes und der Cramer- $\delta$ -Methode eine asymptotische Verteilung der Log-Transformation der Stichprobenvarianz berechnet wird. Mit der Log-Transformation wird die Konvergenz mit der Normalverteilung beschleunigt. Mit diesem Ansatz wird das approximierete beidseitige  $(1 - \alpha)100\%$ -Konfidenzintervall für die Varianz wie folgt definiert:

$$\left[ s^2 \exp(-z_{\alpha/2} se), s^2 \exp(z_{\alpha/2} se) \right]$$

Hierbei ist  $z_\alpha$  das obere Perzentil der Standardnormalverteilung und  $se$  ein asymptotischer Schätzwert des Standardfehlers der log-transformierten Stichprobenvarianz, angegeben als:

$$se = \sqrt{\frac{\hat{\gamma} - (n-3)/n}{n-1}} = \sqrt{\frac{\hat{\gamma}_e + 2 + 3/n}{n-1}}$$

Hummel et al. (2005) haben Simulationsstudien durchgeführt, mit denen veranschaulicht wurde, dass die KorDF-Methode diesem klassischen Ansatz überlegen ist. Bonett hat jedoch zwei Korrekturen des klassischen Ansatzes vorgenommen, um dessen Einschränkungen zu beseitigen. Die erste Korrektur betrifft den Schätzwert der Kurtosis. Bonett schätzt die Kurtosis anhand der folgenden Formel:

$$\hat{\gamma}_e = \frac{n}{(n-1)^2} \sum_{i=1}^n \left( \frac{x_i - m}{s} \right)^4 - 3$$

Hierbei ist  $m$  ein getrimmter Mittelwert, dessen Trim-Anteil gleich  $1/2\sqrt{n-4}$  ist. Dieser Schätzwert der Kurtosis verbessert tendenziell die Genauigkeit der Konfidenzniveaus für symmetrische oder schiefe Verteilungen mit stärker besetzten Randbereichen.

Als zweite Korrektur bestimmt Bonett auf empirische Weise einen konstanten Multiplikator für die Stichprobenvarianz und den Standardfehler. Dieser konstante Multiplikator gleicht annähernd die Randbereich-Wahrscheinlichkeiten aus, wenn der Stichprobenumfang klein ist; er wird wie folgt angegeben:

$$c = \frac{n}{n - z_{\alpha/2}}$$

Diese Korrekturen liefern das approximierte beidseitige  $(1 - \alpha)100\%$ -Konfidenzintervall von Bonett für die Varianz:

$$[cs^2 \exp(-c z_{\alpha/2} se), cs^2 \exp(c z_{\alpha/2} se)]$$

# Anhang C: Genauigkeit der Bonett-Methode im Vergleich mit der Genauigkeit der KorDF-Methode

## Simulation C1: Vergleich von Konfidenzintervallen

Wir wollten die Genauigkeit von Konfidenzintervallen für die Varianz vergleichen, die mit der KorDF-Methode und der Bonett-Methode berechnet wurden. Es wurden Zufallsstichproben mit unterschiedlichen Stichprobenumfängen ( $n = 20, 30, 40, 50, 60, 80, 100, 150, 200, 250, 300$ ) aus verschiedenen Verteilungen generiert, und mit beiden Methoden wurden Konfidenzintervalle berechnet. Zu den Verteilungen zählten:

- 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 3, 5 und 10 Freiheitsgraden ( $Chi(3), Chi(5), Chi(10)$ )
- Linksschiefe Verteilung mit stärker besetzten Randbereichen, konkret die Betaverteilung, deren Parameter auf 8 und 1 festgelegt sind ( $B(8;1)$ )

Um die direkte Auswirkung von Ausreißern auszuwerten, wurden zudem Stichproben aus kontaminierten Normalverteilungen generiert, die wie folgt definiert sind:

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

Hierbei  $p$  ist Mischungsparameter und  $1 - p$  der Anteil der Kontamination (der dem Anteil der Ausreißer entspricht). Es wurden zwei kontaminierte normalverteilte Grundgesamtheiten für die Studie ausgewählt:  $CN(0,9; 3)$ , in der 10 % der Grundgesamtheit Ausreißer darstellen, und  $CN(0,8; 3)$ , in der 20 % der Grundgesamtheit Ausreißer sind. Diese beiden Verteilungen sind symmetrisch und weisen aufgrund der Ausreißer lange Randbereiche auf.

Für jeden Stichprobenumfang wurden 10.000 Stichprobenreplikationen aus jeder Verteilung gezogen, und mit jeder Methode wurden die beidseitigen 95%-Konfidenzintervalle berechnet. Der Startwert des Zufallszahlengenerators wurde festgelegt, damit beide Modelle auf exakt gleiche Stichproben angewendet wurden. Auf der Grundlage dieser Konfidenzintervalle wurden anschließend die simulierten Überdeckungswahrscheinlichkeiten (ÜW) und durchschnittlichen Intervallbreiten (DB) für jede Methode berechnet. Wenn die Konfidenzintervalle der beiden Methoden etwa die gleiche simulierte Überdeckungswahrscheinlichkeit aufweisen, ist die Methode präziser, die (im Durchschnitt)

schmalere Intervalle produziert. Da ein Soll-Konfidenzniveau von 95 % verwendet wurde, betrug der Simulationsfehler  $\sqrt{0,95(0,05)/10.000} = 0,2\%$ .

Die Simulationsergebnisse werden in den nachfolgenden Tabellen 1 und 2 aufgeführt.

**Tabelle 1** Simulierte Überdeckungswahrscheinlichkeiten von beidseitigen 95%-Konfidenzintervallen für die Varianz, die mit der KorDF-Methode und der Bonett-Methode berechnet wurden. Diese Stichproben wurden aus symmetrischen Verteilungen mit schwächer, normal, nahezu normal und stärker besetzten Randbereichen generiert.

Verteilung	Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
	U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	T(5)
Schiefte	0	0	0	0	0	0	0	0
Kurtosis ( $\gamma_e$ )	-1,200	-0,667	0	1,000	3,000	4,544	5,333	6,000

**n = 10**

KorDF	ÜW	0,910	0,909	0,903	0,883	0,853	0,793	0,815	0,858
	DB	0,154	0,087	3,276	5,160	13,924	21,658	14,913	11,742
Bonett	ÜW	0,972	0,967	0,962	0,952	0,919	0,891	0,920	0,935
	DB	0,242	0,115	3,710	5,134	10,566	15,335	10,367	8,578

**n = 20**

KorDF	ÜW	0,937	0,937	0,923	0,909	0,881	0,819	0,817	0,868
	DB	0,080	0,045	1,572	2,463	5,781	9,265	6,539	5,151
Bonett	ÜW	0,953	0,954	0,946	0,934	0,909	0,856	0,864	0,904
	DB	0,100	0,051	1,683	2,422	4,932	7,282	4,945	4,026

**n = 30**

KorDF	ÜW	0,946	0,942	0,933	0,917	0,894	0,851	0,823	0,882
	DB	0,061	0,034	1,170	1,764	4,117	6,330	4,557	3,667
Bonett	ÜW	0,951	0,950	0,947	0,933	0,909	0,869	0,852	0,907
	DB	0,070	0,037	1,221	1,750	3,654	5,383	3,736	2,997

**n = 40**

KorDF	ÜW	0,953	0,947	0,932	0,922	0,904	0,867	0,833	0,890
	DB	0,051	0,028	0,971	1,489	3,246	5,131	3,654	3,024
Bonett	ÜW	0,954	0,951	0,941	0,936	0,914	0,879	0,856	0,907
	DB	0,057	0,030	1,002	1,469	2,994	4,519	3,128	2,542

Verteilung	Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
	U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	T(5)
Schiefe	0	0	0	0	0	0	0	0
Kurtosis ( $\gamma_e$ )	-1,200	-0,667	0	1,000	3,000	4,544	5,333	6,000

**n = 50**

KorDF	ÜW	0,951	0,945	0,937	0,925	0,911	0,878	0,838	0,893
	DB	0,045	0,025	0,849	1,291	2,789	4,357	3,091	2,603
Bonett	ÜW	0,951	0,947	0,944	0,938	0,918	0,888	0,855	0,908
	DB	0,049	0,026	0,870	1,280	2,613	3,939	2,729	2,240

**n = 60**

KorDF	ÜW	0,949	0,943	0,938	0,926	0,913	0,890	0,853	0,899
	DB	0,040	0,022	0,766	1,155	2,490	3,857	2,768	2,283
Bonett	ÜW	0,949	0,947	0,943	0,935	0,918	0,896	0,868	0,910
	DB	0,043	0,023	0,781	1,147	2,354	3,552	2,498	2,023

**n = 70**

KorDF	ÜW	0,948	0,945	0,940	0,930	0,913	0,890	0,858	0,896
	DB	0,037	0,020	0,701	1,056	2,283	3,458	2,475	2,049
Bonett	ÜW	0,947	0,946	0,944	0,938	0,918	0,894	0,868	0,905
	DB	0,039	0,021	0,713	1,049	2,174	3,227	2,272	1,828

**n = 80**

KorDF	ÜW	0,947	0,949	0,938	0,929	0,918	0,905	0,869	0,902
	DB	0,034	0,019	0,652	0,988	2,089	3,205	2,300	1,906
Bonett	ÜW	0,946	0,950	0,942	0,935	0,923	0,907	0,877	0,911
	DB	0,036	0,019	0,662	0,982	2,005	3,014	2,133	1,716

**n = 90**

KorDF	ÜW	0,946	0,947	0,948	0,929	0,918	0,908	0,869	0,901
	DB	0,032	0,018	0,611	0,921	1,951	2,982	2,124	1,874
Bonett	ÜW	0,945	0,948	0,952	0,936	0,920	0,910	0,874	0,909
	DB	0,034	0,018	0,618	0,916	1,882	2,822	1,984	1,646



Verteilung	Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
	U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	T(5)
Schiefe	0	0	0	0	0	0	0	0
Kurtosis ( $\gamma_e$ )	-1,200	-0,667	0	1,000	3,000	4,544	5,333	6,000

**n = 100**

KorDF	ÜW	0,947	0,951	0,945	0,933	0,920	0,910	0,885	0,912
	DB	0,030	0,017	0,576	0,873	1,830	2,801	2,017	1,658
Bonett	ÜW	0,946	0,953	0,948	0,937	0,923	0,912	0,891	0,916
	DB	0,032	0,017	0,583	0,869	1,772	2,666	1,899	1,522

**n = 150**

KorDF	ÜW	0,949	0,951	0,947	0,936	0,932	0,925	0,896	0,912
	DB	0,024	0,014	0,464	0,700	1,470	2,228	1,602	1,325
Bonett	ÜW	0,948	0,952	0,949	0,939	0,933	0,924	0,898	0,915
	DB	0,025	0,014	0,467	0,698	1,438	2,156	1,539	1,251

**n = 200**

KorDF	ÜW	0,943	0,949	0,948	0,938	0,927	0,930	0,914	0,918
	DB	0,021	0,012	0,400	0,605	1,265	1,906	1,373	1,178
Bonett	ÜW	0,942	0,951	0,949	0,940	0,928	0,930	0,915	0,920
	DB	0,021	0,012	0,402	0,603	1,245	1,860	1,333	1,106

**n = 250**

KorDF	ÜW	0,952	0,952	0,949	0,942	0,938	0,929	0,909	0,915
	DB	0,019	0,010	0,355	0,538	1,120	1,690	1,219	1,037
Bonett	ÜW	0,951	0,952	0,949	0,944	0,941	0,929	0,909	0,916
	DB	0,019	0,010	0,357	0,537	1,106	1,657	1,190	0,986

**n = 300**

KorDF	ÜW	0,950	0,948	0,951	0,940	0,938	0,936	0,920	0,914
	DB	0,017	0,009	0,324	0,490	1,019	1,544	1,115	0,933
Bonett	ÜW	0,950	0,947	0,951	0,942	0,937	0,929	0,920	0,916
	DB	0,017	0,010	0,325	0,489	1,009	1,657	1,093	0,897

**Tabelle 2** Simulierte Überdeckungswahrscheinlichkeiten von beidseitigen 95%-Konfidenzintervallen für die Varianzen, die mit der KorDF-Methode und der Bonett-Methode berechnet wurden. Diese Stichproben wurden aus schiefen Verteilungen mit nahezu normal, mittelstark und stärker besetzten Randbereichen generiert.

Verteilung		Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8:1)	Chi(5)	Chi(3)	Exp
Schiefe		0,894	-1,423	1,265	1,633	2
Kurtosis ( $\gamma_e$ )		1,200	2,284	2,400	4,000	6
<b>n = 10</b>						
KorDF	ÜW	0,869	0,815	0,836	0,797	0,758
	DB	93,383	0,065	61,994	47,821	10,711
Bonett	ÜW	0,950	0,917	0,938	0,911	0,882
	DB	91,006	0,058	53,830	38,137	7,498
<b>n = 20</b>						
KorDF	ÜW	0,889	0,862	0,862	0,833	0,811
	DB	41,497	0,026	25,479	20,099	4,293
Bonett	ÜW	0,932	0,912	0,913	0,893	0,877
	DB	41,600	0,026	24,094	17,232	3,370
<b>n = 30</b>						
KorDF	ÜW	0,901	0,881	0,880	0,864	0,838
	DB	30,021	0,018	18,182	13,630	2,844
Bonett	ÜW	0,931	0,920	0,914	0,906	0,885
	DB	30,462	0,019	17,858	12,634	2,441
<b>n = 40</b>						
KorDF	ÜW	0,909	0,882	0,885	0,867	0,862
	DB	24,459	0,015	14,577	10,649	2,193
Bonett	ÜW	0,930	0,915	0,913	0,904	0,898
	DB	24,952	0,015	14,504	1,991	1,991
<b>n = 50</b>						
KorDF	ÜW	0,912	0,900	0,892	0,871	0,868
	DB	21,373	0,013	12,694	9,115	1,861

Verteilung		Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8:1)	Chi(5)	Chi(3)	Exp
Schiefe		0,894	-1,423	1,265	1,633	2
Kurtosis ( $\gamma_e$ )		1,200	2,284	2,400	4,000	6
Bonett	ÜW	0,930	0,927	0,916	0,903	0,901
	DB	21,814	0,013	12,741	8,897	1,735

**n = 60**

KorDF	ÜW	0,915	0,908	0,901	0,890	0,875
	DB	18,928	0,011	11,338	8,211	1,645
Bonett	ÜW	0,930	0,933	0,923	0,917	0,900
	DB	19,369	0,012	11,456	8,093	1,554

**n = 70**

KorDF	ÜW	0,915	0,910	0,904	0,898	0,881
	DB	17,513	0,010	10,307	7,461	1,488
Bonett	ÜW	0,932	0,932	0,922	0,919	0,906
	DB	17,906	0,011	10,464	7,408	1,429

**n = 80**

KorDF	ÜW	0,920	0,916	0,911	0,904	0,890
	DB	16,157	0,009	9,604	6,892	1,349
Bonett	ÜW	0,935	0,936	0,929	0,924	0,915
	DB	16,537	0,010	9,765	6,882	1,314

**n = 90**

KorDF	ÜW	0,924	0,918	0,911	0,897	0,894
	DB	15,250	0,009	9,007	6,323	1,255
Bonett	ÜW	0,938	0,936	0,929	0,918	0,913
	DB	15,609	0,009	9,175	6,366	1,230

**n = 100**

KorDF	ÜW	0,926	0,919	0,915	0,908	0,895
	DB	14,332	0,008	8,451	6,016	1,171
Bonett	ÜW	0,935	0,936	0,931	0,924	0,916
	DB	14,664	0,009	8,625	6,063	1,158

Verteilung	Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
	Chi(10)	B(8:1)	Chi(5)	Chi(3)	Exp
Schiefe	0,894	-1,423	1,265	1,633	2
Kurtosis ( $\gamma_e$ )	1,200	2,284	2,400	4,000	6

**n = 150**

KorDF	ÜW	0,933	0,925	0,923	0,913	0,911
	DB	11,606	0,007	6,781	4,792	0,933
Bonett	ÜW	0,943	0,941	0,936	0,929	0,928
	DB	11,846	0,007	6,942	4,875	0,937

**n = 200**

KorDF	ÜW	0,935	0,934	0,926	0,916	0,915
	DB	9,973	0,006	5,849	4,127	0,799
Bonett	ÜW	0,942	0,948	0,936	0,930	0,931
	DB	10,185	0,006	5,991	4,212	0,808

**n = 250**

KorDF	ÜW	0,938	0,939	0,934	0,926	0,922
	DB	8,899	0,005	5,231	3,652	0,705
Bonett	ÜW	0,946	0,951	0,944	0,936	0,931
	DB	9,078	0,005	5,355	3,735	0,716

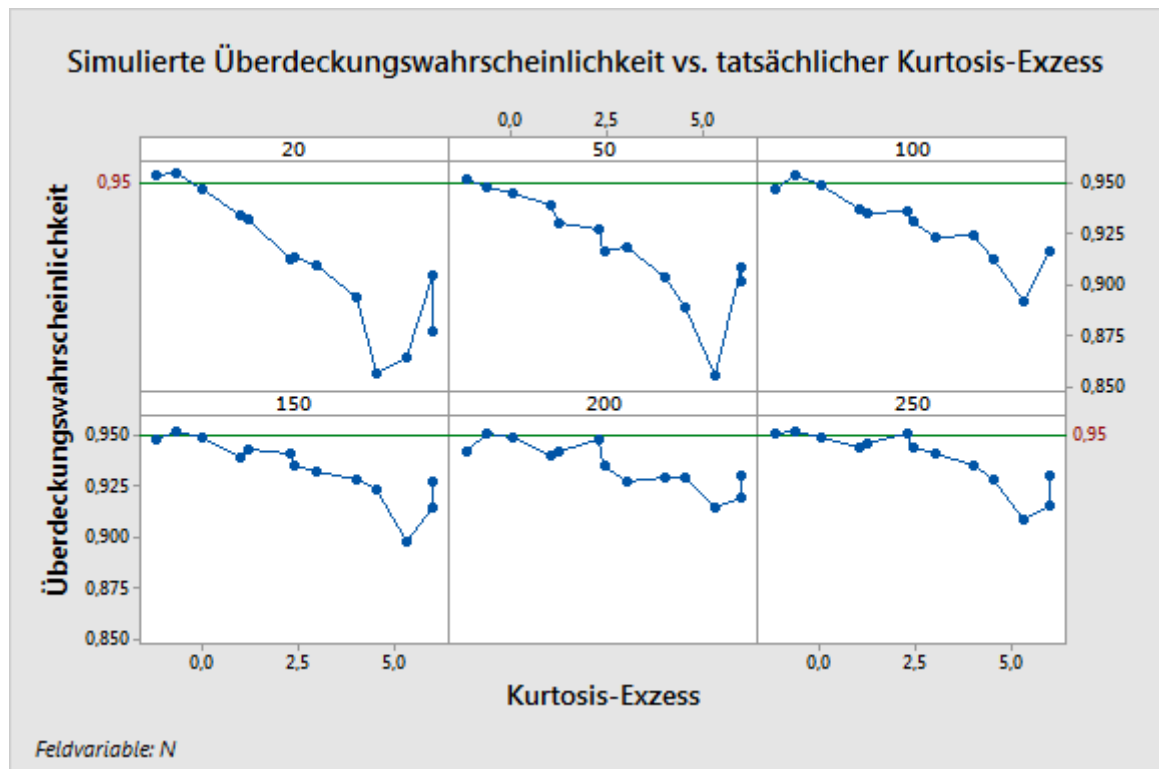
**n = 300**

KorDF	ÜW	0,942	0,938	0,934	0,931	0,922
	DB	8,156	0,005	4,749	3,344	0,640
Bonett	ÜW	0,947	0,948	0,943	0,941	0,933
	DB	8,314	0,005	4,862	3,419	0,651

Unsere Ergebnisse entsprechen weitestgehend denen, die von Bonett (2006) veröffentlicht wurden. Wie in den Tabellen 1 und 2 ersichtlich, sind die mit der Bonett-Methode berechneten Konfidenzintervalle den mit der KorDF-Methode berechneten überlegen, da sie Überdeckungswahrscheinlichkeiten ergeben, die näher am Sollwert von 0,95 liegen, und durchschnittlich schmalere sind. Wenn die Konfidenzintervalle der beiden Methoden etwa die gleiche simulierte Überdeckungswahrscheinlichkeit aufweisen, ist die Methode präziser, die (im Durchschnitt) schmalere Intervalle produziert. Daher weist der statistische Test auf Varianz auf der Grundlage der Bonett-Methode eine bessere Leistung und eine niedrigere Wahrscheinlichkeit eines Fehlers 1. Art und 2. Art auf. Bei großen Stichprobenumfängen

liefern die beiden Methoden nahezu identische Ergebnisse, bei kleinen bis mittleren Stichprobenumfängen hingegen ist die Bonett-Methode überlegen.

Obleich die Bonett-Methode im Allgemeinen eine bessere Leistung als die KorDF-Methode zeigt, liefert sie bei symmetrischen oder schiefen Verteilungen mit stärker besetzten Randbereichen durchgehend Überdeckungswahrscheinlichkeiten unter dem Sollwert von 0,95, selbst bei außergewöhnlich großen Stichproben ( $n > 100$ ). Dies wird unten in Abbildung 1 veranschaulicht, in der die simulierten Überdeckungswahrscheinlichkeiten für die Bonett-Methode im Vergleich zum tatsächlichen Kurtosis-Überschuss der Grundgesamtheit für kleine, mittlere und große Stichprobenumfänge dargestellt werden.



**Abbildung 1** Simulierte Überdeckungswahrscheinlichkeiten für 95%-Konfidenzintervalle nach Bonett, dargestellt im Vergleich zum Kurtosis-Überschuss der einzelnen Verteilungen bei unterschiedlichen Stichprobenumfängen.

Wie in Abbildung 1 ersichtlich, steigt mit Zunahme der Kurtosis auch der Stichprobenumfang, der erforderlich ist, damit sich die simulierten Überdeckungswahrscheinlichkeiten dem Sollwert annähern. Wie bereits erwähnt, sind die simulierten Überdeckungswahrscheinlichkeiten für die Bonett-Methode bei Verteilungen mit stärker besetzten Randbereichen niedrig. Bei Verteilungen mit schwächer besetzten Randbereichen, z. B. der Gleichverteilung und der Beta(3;3)-Verteilung, sind die simulierten Überdeckungswahrscheinlichkeiten jedoch stabil und entsprechen dem Sollwert, auch wenn die Stichprobenumfänge lediglich 20 betragen. Deshalb legen wir unserem Kriterium zum Bewerten der Gültigkeit der Bonett-Methode sowohl den Stichprobenumfang als auch die Besetzung der Randbereiche der Verteilung zugrunde, aus der die jeweilige Stichprobe gezogen wird.

Im ersten Schritt bei der Entwicklung dieses Kriteriums unterteilen wir die Verteilungen entsprechend der Besetzung ihrer Randbereiche in drei Kategorien:

- **Verteilungen mit schwächer und normal besetzten Randbereichen (Typ „Schwach“):** Dies sind die Verteilungen, für die Bonett-Konfidenzintervalle stabile Überdeckungswahrscheinlichkeiten nahe der Sollüberdeckung liefern. Bei diesen Verteilungen sind die Ergebnisse genau, selbst wenn die Stichprobenumfänge lediglich 20 betragen. Beispiele sind die Gleichverteilung, die Beta(3;3)-Verteilung, die Normalverteilung, die t-Verteilung mit 10 Freiheitsgraden und die Chi-Quadrat-Verteilung mit 10 Freiheitsgraden.
- **Verteilungen mit mittelstark besetzten Randbereichen (Typ „Mittel“):** Bei diesen Verteilungen benötigt die Bonett-Methode einen Stichprobenumfang von mindestens 80, damit die simulierten Überdeckungswahrscheinlichkeiten nahe der Sollüberdeckung liegen. Beispiele sind die Chi-Quadrat-Verteilung mit 5 Freiheitsgraden und die Beta(8;1)-Verteilung.
- **Verteilungen mit stärker besetzten Randbereichen (Typ „Stark“):** Dies sind Verteilungen, für die die Bonett-Konfidenzintervalle Überdeckungswahrscheinlichkeiten liefern, die weit unter der Sollüberdeckung liegen, es sei denn, die Stichprobenumfänge sind extrem groß ( $n \geq 200$ ). Beispiele sind die t-Verteilung mit 5 Freiheitsgraden, die Laplace-Verteilung, die Chi-Quadrat-Verteilung mit 3 Freiheitsgraden, die Exponentialverteilung und die zwei kontaminierten Normalverteilungen  $CN(0,9;3)$  und  $CN(0,8;3)$ .

Damit erfordert eine allgemeine Regel zum Auswerten der Gültigkeit der Bonett-Methode ein Verfahren, mit dem bestimmt werden kann, aus welchem der drei Verteilungstypen die Stichprobendaten stammen. Ein solches Verfahren wurde im Rahmen der Prüfung auf Gültigkeit der Testdaten entwickelt. Weitere Einzelheiten finden Sie in Anhang E.

# Anhang D: Theoretische Trennschärfe

Wir haben die theoretische Trennschärfefunktion des Tests für die Bonett-Methode abgeleitet und Simulationen durchgeführt, um die theoretische und die simulierte Trennschärfe des Tests miteinander zu vergleichen. Wenn die Kurven der theoretischen und der simulierten Trennschärfe dicht beieinander liegen, sollte die Analyse von Trennschärfe und Stichprobenumfang auf der Grundlage der theoretischen Trennschärfefunktion genaue Ergebnisse liefern.

## Formel D1: Theoretische Trennschärfefunktion für die Bonett-Methode

Wie bereits erläutert, basiert die Bonett-Methode auf dem hinreichend bekannten klassischen Ansatz, bei dem anhand des zentralen Grenzwertsatzes und der Cramer- $\delta$ -Methode eine asymptotische Verteilung der log-transformierten Stichprobenvarianz bestimmt wird. Konkreter: Es wird dargestellt, dass die Verteilung von  $\frac{\ln S^2 - \ln \sigma^2}{se}$  in großen Stichproben annähernd der Standardnormalverteilung folgt. Der Nenner  $se$  ist der Standardfehler der log-transformierten Stichprobenvarianz für große Stichproben, der angegeben wird als

$$se = \sqrt{\frac{\gamma - (n-3)/n}{n-1}}$$

Hierbei ist  $\gamma$  die Kurtosis der unbekanntes übergeordneten Grundgesamtheit.

Daraus folgt, dass eine approximierte Trennschärfefunktion mit einem approximierten Alpha-Niveau für den beidseitigen Test unter Verwendung der Bonett-Methode folgendermaßen als Funktion des Stichprobenumfangs, des Verhältnisses  $\rho = \sigma/\sigma_0$  und der Kurtosis der übergeordneten Grundgesamtheit  $\gamma$  angegeben werden kann:

$$\pi(n, \rho, \gamma) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{\sqrt{\frac{\gamma - 1 + 3/n}{n-1}}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{\sqrt{\frac{\gamma - 1 + 3/n}{n-1}}}\right)$$

Hierbei ist  $\sigma_0$  der hypothetische Wert der unbekanntes Standardabweichung,  $\Phi$  die Verteilungsfunktion der Standardnormalverteilung und  $z_\alpha$  der obere  $\alpha$ . Perzentilpunkt der Standardnormalverteilung. Die einseitigen Trennschärfefunktionen können ebenfalls aus diesen Berechnungen abgeleitet werden.

Beachten Sie, dass beim Planen des Stichprobenumfangs für eine Studie anstelle der tatsächlichen Kurtosis ein Schätzwert der Kurtosis verwendet werden kann. Diesem Schätzwert werden i. d. R. Expertenmeinungen oder die Ergebnisse früherer Experimente zugrunde liegen. Wenn derartige Informationen nicht verfügbar sind, empfiehlt es sich häufig, eine kleine Pilotstudie durchzuführen, um die Pläne für die umfassendere Studie zu

entwickeln. Die Kurtosis kann anhand einer Stichprobe aus der Pilotstudie geschätzt werden als

$$\hat{\gamma} = \frac{n}{(n-1)^2} \sum_{i=1}^n \left( \frac{x_i - m}{s} \right)^4$$

Hierbei ist  $m$  ein getrimmter Mittelwert, dessen Trim-Anteil gleich  $1/2\sqrt{n-4}$  ist.

## Simulation D1: Vergleich der tatsächlichen Trennschärfe mit der theoretischen Trennschärfe

Wir haben eine Simulation entworfen, um die tatsächlichen Trennschärfen (die wir als simulierte Trennschärfen bezeichnen) mit den theoretischen Trennschärfen (die wir als approximierte Trennschärfen bezeichnen) miteinander zu vergleichen, wenn ein Varianztest auf der Grundlage der Bonett-Methode ausgeführt wird.

In jedem Experiment wurden aus den in Simulation C1 beschriebenen Verteilungen (siehe Anhang C) 10.000 Stichprobenreplikationen jeweils mit dem Umfang  $n$  generiert, wobei  $n = 20, 30, 40, 50, \dots, 120$ . Für jede Verteilung und jeden Stichprobenumfang  $n$  wurden die simulierten Trennschärfen als Anteil der 10.000 zufälligen Stichprobenreplikationen berechnet, für die der beidseitige Test mit dem Alpha-Niveau  $\alpha = 0,05$  signifikant war. Beim Berechnen der simulierten Trennschärfe wurde  $\rho = \sigma/\sigma_0 = 1,25$  verwendet, um relativ geringe Trennschärfen zu erhalten. Anschließend wurden die entsprechenden Trennschärfen zu Vergleichszwecken mit Hilfe der theoretischen Trennschärfefunktion berechnet.

Die Ergebnisse sind in den Tabellen 3 und 4 aufgeführt und in der nachfolgenden Abbildung 2 grafisch dargestellt.

**Tabelle 3** Simulierte Trennschärfen (ausgewertet bei  $\rho = \sigma/\sigma_0 = 1,25$ ) eines beidseitigen Tests auf Varianz auf der Grundlage der Bonett-Methode im Vergleich mit den theoretischen Trennschärfen (Normal-Approximation). Die Stichproben wurden aus symmetrischen Verteilungen mit schwächer, normal, nahezu normal und stärker besetzten Randbereichen generiert.

$n$	Trennschärfe	Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
		U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	t(5)
20	Simul.	0,521	0,390	0,310	0,237	0,178	0,152	0,139	0,172
	Approx.	0,514	0,359	0,264	0,195	0,137	0,117	0,109	0,104
30	Simul.	0,707	0,551	0,441	0,337	0,225	0,186	0,169	0,228
	Approx.	0,717	0,519	0,382	0,276	0,186	0,154	0,143	0,135
40	Simul.	0,831	0,679	0,526	0,427	0,285	0,266	0,203	0,285
	Approx.	0,846	0,651	0,490	0,356	0,236	0,192	0,176	0,165

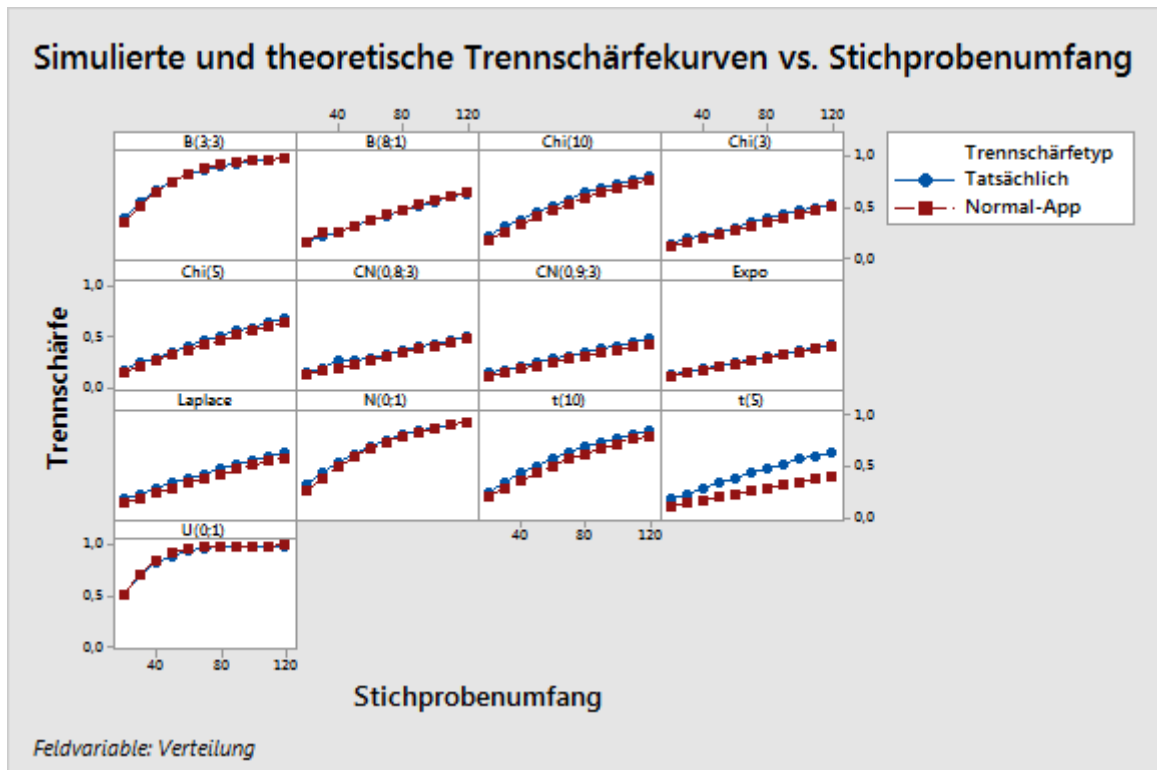


n	Trennschärfe	Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
		U(0:1)	B(3:3)	N(0:1)	t(10)	Lpl	CN (0,8:3)	CN (0,9:3)	t(5)
50	Simul.	0,899	0,753	0,621	0,505	0,332	0,255	0,238	0,340
	Approx.	0,921	0,754	0,586	0,431	0,284	0,229	0,210	0,196
60	Simul.	0,942	0,822	0,701	0,570	0,380	0,285	0,274	0,384
	Approx.	0,961	0,830	0,668	0,501	0,332	0,266	0,243	0,227
70	Simul.	0,964	0,866	0,757	0,632	0,424	0,327	0,314	0,439
	Approx.	0,981	0,885	0,737	0,566	0,379	0,303	0,276	0,257
80	Simul.	0,981	0,909	0,815	0,689	0,481	0,372	0,347	0,483
	Approx.	0,991	0,923	0,794	0,624	0,423	0,340	0,309	0,288
90	Simul.	0,988	0,937	0,851	0,724	0,514	0,400	0,377	0,523
	Approx.	0,996	0,950	0,840	0,676	0,467	0,375	0,342	0,318
100	Simul.	0,994	0,961	0,880	0,779	0,558	0,430	0,411	0,566
	Approx.	0,998	0,967	0,876	0,722	0,508	0,410	0,373	0,347
110	Simul.	0,997	0,967	0,909	0,803	0,591	0,471	0,449	0,592
	Approx.	0,999	0,979	0,905	0,763	0,547	0,443	0,404	0,376
120	Simul.	0,999	0,982	0,929	0,844	0,629	0,502	0,476	0,630
	Approx.	1,000	0,987	0,928	0,799	0,584	0,476	0,434	0,405

**Tabelle 4** Simulierte Trennschärfen (ausgewertet bei  $\rho = \sigma/\sigma_0 = 1,25$ ) eines beidseitigen Tests auf Varianz auf der Grundlage der Bonett-Methode im Vergleich mit den theoretischen Trennschärfen (Normal-Approximation). Die Stichproben wurden aus schiefen Verteilungen mit nahezu normal, mittelstark und stärker besetzten Randbereichen generiert.

n	Trennschärfe	Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8;1)	Chi(5)	Chi(3)	Exp
20	Simul.	0,222	0,166	0,172	0,139	0,128
	Approx.	0,186	0,152	0,149	0,123	0,104
30	Simul.	0,314	0,216	0,234	0,190	0,151
	Approx.	0,263	0,263	0,205	0,164	0,135
40	Simul.	0,387	0,266	0,292	0,223	0,186
	Approx.	0,338	0,266	0,261	0,204	0,165

<b>n</b>	Trennschärfe	Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8;1)	Chi(5)	Chi(3)	Exp
50	Simul.	0,455	0,324	0,349	0,263	0,208
	Approx.	0,409	0,323	0,316	0,245	0,196
60	Simul.	0,521	0,376	0,399	0,302	0,239
	Approx.	0,477	0,377	0,369	0,286	0,227
70	Simul.	0,583	0,419	0,463	0,361	0,269
	Approx.	0,539	0,430	0,420	0,325	0,257
80	Simul.	0,646	0,473	0,499	0,394	0,299
	Approx.	0,597	0,479	0,469	0,365	0,288
90	Simul.	0,688	0,517	0,561	0,428	0,327
	Approx.	0,649	0,526	0,516	0,403	0,318
100	Simul.	0,738	0,561	0,591	0,469	0,368
	Approx.	0,695	0,571	0,560	0,440	0,347
110	Simul.	0,779	0,608	0,637	0,495	0,394
	Approx.	0,737	0,611	0,600	0,475	0,376
120	Simul.	0,810	0,635	0,679	0,538	0,416
	Approx.	0,774	0,650	0,638	0,509	0,405



**Abbildung 2** Simulierte Trennschärfekurven im Vergleich mit theoretischen Trennschärfekurven für unterschiedliche Verteilungen

Die Ergebnisse in den Tabellen 3 und 4 sowie Abbildung 2 zeigen, dass die Werte der theoretischen Trennschärfe und die simulierten Trennschärfen bei Stichproben aus Verteilungen mit schwächer besetzten Randbereichen (Verteilungstyp „Schwach“ gemäß Definition in Anhang C), z. B. der Gleichverteilung, der Beta (3;3)-Verteilung, der Normalverteilung, der t-Verteilung mit 10 Freiheitsgraden und der Chi-Quadrat-Verteilung mit 10 Freiheitsgraden, praktisch nicht voneinander unterschieden werden können.

Bei Verteilungen mit stärker besetzten Randbereichen (Verteilungstyp „Stark“) liegen die Kurven der simulierten Trennschärfen bei kleinen Stichproben hingegen deutlich über den Kurven der theoretischen Trennschärfen. Zu diesen Verteilungen mit stärker besetzten Randbereichen zählen die t-Verteilung mit 5 Freiheitsgraden, die Laplace-Verteilung, die Chi-Quadrat-Verteilung mit 3 Freiheitsgraden, die Exponentialverteilung und die zwei kontaminierten Normalverteilungen CN(0,9;3) und CN(0,8;3). Wenn Sie also den Stichprobenumfang für eine Studie planen und die Stichprobe aus einer Verteilung mit stärker besetzten Randbereichen stammt, kann der von der theoretischen Trennschärfefunktion geschätzte Stichprobenumfang größer als der Stichprobenumfang sein, der zum Erzielen einer bestimmten Solltrennschärfe tatsächlich erforderlich ist.

# Anhang E: Der SJ-Test für normal und stärker besetzte Randbereiche

Aus den Ergebnissen der Simulationsstudie in Anhang C kann geschlossen werden, dass bei stärker besetzten Randbereichen der Verteilung größere Stichprobenumfänge erforderlich sind, damit sich die simulierte Überdeckungswahrscheinlichkeit der Bonett-Konfidenzintervalle dem Sollwert annähert. Die Schiefe hingegen scheint keine signifikante Auswirkung auf die simulierten Überdeckungswahrscheinlichkeiten auszuüben.

Deshalb musste ein Kriterium zum Auswerten der Gültigkeit der Bonett-Methode entwickelt werden, das sowohl auf dem Stichprobenumfang als auch auf der Besetzung der Randbereiche der Verteilung basiert, aus der die Stichprobe gezogen wird. Glücklicherweise haben Gel et al. (2007) einen ausreichend trennscharfen Test entwickelt, mit dem die Nullhypothese von normal besetzten Randbereichen der Verteilung direkt gegen die Alternativhypothese von stärker besetzten Randbereichen getestet werden kann. Der Test, der von uns als SJ-Test bezeichnet wird, basiert auf der folgenden Statistik:

$$\hat{R} = \frac{S}{\hat{j}}$$

Hierbei ist  $S$  die Standardabweichung der Stichprobe und  $\hat{j}$  der Schätzwert der absoluten Abweichung des Stichprobenmittelwerts vom Median,  $m$ , der wie folgt angegeben wird:

$$\hat{j} = \frac{\sqrt{\pi/2}}{n} \sum_{i=1}^n |X_i - m|$$

Ein Test mit ungefährem Niveau  $\alpha$  gegen die Alternativhypothese von stärker besetzten Randbereichen weist die Nullhypothese von normal besetzten Randbereichen zurück, wenn

$$\frac{\sqrt{n}(\hat{R} - 1)}{\sigma_R} \geq z_\alpha$$

Hierbei ist  $z_\alpha$  das obere  $\alpha$ . Perzentil einer Standardnormalverteilung und  $\sigma_R = (\pi - 3)/2$ .

Gel et al. (2007) haben gezeigt, dass durch Ersetzen des oberen  $\alpha$ . Perzentils der Standardnormalverteilung durch das der t-Verteilung mit  $(\sqrt{n} + 3)/2$  Freiheitsgraden für mittlere Stichprobenumfänge bessere Approximationen erzielt werden. Beim Anwenden des SJ-Tests auf die Prüfung auf Gültigkeit des Tests ersetzen wir daher  $z_\alpha$  durch  $t_{d,\alpha}$ , das obere  $\alpha$ . Perzentil der t-Verteilung mit  $d = (\sqrt{n} + 3)/2$  Freiheitsgraden.

# Anhang F: Gültigkeit des Tests

## Simulation F1: Ermitteln von Verteilungsklassifikationen anhand der simulierten Trennschärfe des SJ-Tests

Wir haben Simulationen durchgeführt, um die Trennschärfe des SJ-Tests zu untersuchen. Dabei wurden Stichproben unterschiedlicher Umfänge ( $n = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 180, 200$ ) aus verschiedenen Verteilungen generiert. Die Verteilungen wiesen normal, schwächer, mittelstark oder stärker besetzte Randbereiche auf und sind identisch mit denen, die in Simulation C1 (siehe Anhang C) beschrieben werden. Für jeden gegebenen Stichprobenumfang wurden 10.000 Stichprobenreplikationen aus den einzelnen Verteilungen gezogen. Die simulierte Trennschärfe des SJ-Tests wurde als Anteil der Fälle berechnet, bei denen die Nullhypothese (dass die Verteilung normal besetzte Randbereiche aufweist) zurückgewiesen wurde. Darüber hinaus wurden die durchschnittlichen  $R$ -Werte (DR) und die durchschnittlichen  $p$ -Werte (DP) berechnet.

Die Simulationsergebnisse werden in den nachfolgenden Tabellen 5 und 6 aufgeführt.

**Tabelle 5** Simulierte Trennschärfen des SJ-Tests. Die Stichproben wurden aus symmetrischen Verteilungen mit schwächer, normal, nahezu normal und stärker besetzten Randbereichen generiert.

Verteilung		Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
		U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	t(5)
<b>n</b>	Wahres R	0,921	0,965	1,0	1,032	1,128	1,152	1,118	1,085
10	Trennschärfe	0,021	0,041	0,075	0,103	0,249	0,264	0,198	0,161
	DR	1,010	1,036	1,060	1,073	1,129	1,131	1,106	1,096
	DP	0,482	0,401	0,341	0,314	0,219	0,228	0,272	0,278
15	Trennschärfe	0,009	0,027	0,071	0,121	0,350	0,389	0,283	0,215
	DR	0,986	1,018	1,043	1,063	1,130	1,140	1,110	1,093
	DP	0,572	0,440	0,357	0,302	0,171	0,181	0,240	0,247
20	Trennschärfe	0,002	0,016	0,066	0,144	0,428	0,465	0,331	0,253
	DR	0,966	1,001	1,030	1,054	1,127	1,137	1,104	1,086
	DP	0,669	0,503	0,382	0,311	0,147	0,161	0,236	0,244

Verteilung		Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
		U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	t(5)
<b>n</b>	Wahres R	0,921	0,965	1,0	1,032	1,128	1,152	1,118	1,085
25	Trennschärfe	0,002	0,011	0,065	0,153	0,500	0,550	0,397	0,293
	DR	0,959	0,995	1,025	1,050	1,128	1,141	1,107	1,086
	DP	0,721	0,535	0,391	0,305	0,120	0,128	0,208	0,223
30	Trennschärfe	0,001	0,010	0,060	0,170	0,561	0,603	0,431	0,334
	DR	0,951	0,989	1,019	1,046	1,127	1,141	1,106	1,084
	DP	0,773	0,570	0,409	0,304	0,103	0,112	0,197	0,209
40	Trennschärfe	0,000	0,006	0,058	0,190	0,665	0,709	0,513	0,401
	DR	0,944	0,984	1,015	1,043	1,126	1,145	1,109	1,084
	DP	0,840	0,616	0,420	0,287	0,073	0,076	0,162	0,179
50	Trennschärfe	0,000	0,004	0,058	0,208	0,746	0,785	0,590	0,462
	DR	0,939	0,980	1,012	1,040	1,126	1,146	1,111	1,084
	DP	0,886	0,654	0,427	0,279	0,053	0,055	0,131	0,156
60	Trennschärfe	0,000	0,002	0,060	0,231	0,813	0,836	0,647	0,518
	DR	0,936	0,978	1,010	1,039	1,127	1,146	1,112	1,084
	DP	0,913	0,686	0,430	0,267	0,039	0,039	0,109	0,134
70	Trennschärfe	0,000	0,002	0,054	0,247	0,863	0,879	0,702	0,554
	DR	0,934	0,975	1,009	1,037	1,127	1,147	1,112	1,083
	DP	0,935	0,716	0,437	0,259	0,028	0,029	0,091	0,123
80	Trennschärfe	0,000	0,001	0,054	0,265	0,896	0,912	0,729	0,591
	DR	0,933	0,974	1,007	1,037	1,128	1,147	1,111	1,083
	DP	0,950	0,740	0,440	0,241	0,021	0,021	0,079	0,105
90	Trennschärfe	0,000	0,001	0,054	0,281	0,933	0,934	0,771	0,633
	DR	0,932	0,973	1,007	1,036	1,128	1,148	1,113	1,083
	DP	0,962	0,759	0,445	0,237	0,014	0,016	0,067	0,093
100	Trennschärfe	0,000	0,001	0,057	0,301	0,947	0,954	0,805	0,661
	DR	0,930	0,972	1,006	1,036	1,127	1,148	1,113	1,083
	DP	0,971	0,779	0,446	0,224	0,012	0,011	0,055	0,083
120	Trennschärfe	0,000	0,000	0,052	0,334	0,974	0,974	0,852	0,732
	DR	0,929	0,971	1,005	1,035	1,128	1,149	1,114	1,083
	DP	0,982	0,809	0,452	0,206	0,006	0,007	0,041	0,064

Verteilung		Symmetrische Verteilungen mit schwächer, normal oder nahezu normal besetzten Randbereichen				Symmetrische Verteilungen mit stärker besetzten Randbereichen			
		U(0;1)	B(3;3)	N(0;1)	t(10)	Lpl	CN (0,8;3)	CN (0,9;3)	t(5)
<b>n</b>	Wahres R	0,921	0,965	1,0	1,032	1,128	1,152	1,118	1,085
140	Trennschärfe	0,000	0,000	0,052	0,336	0,986	0,988	0,894	0,785
	DR	0,928	0,971	1,004	1,034	1,127	1,150	1,116	1,084
	DP	0,989	0,834	0,454	0,192	0,004	0,003	0,027	0,048
160	Trennschärfe	0,000	0,000	0,054	0,402	0,993	0,992	0,916	0,819
	DR	0,927	0,970	1,004	1,034	1,128	1,150	1,114	1,084
	DP	0,993	0,858	0,457	0,177	0,002	0,002	0,021	0,040
180	Trennschärfe	0,000	0,000	0,052	0,416	0,998	0,996	0,934	0,853
	DR	0,926	0,969	1,003	1,034	1,128	1,149	1,115	1,084
	DP	0,995	0,874	0,461	0,167	0,001	0,001	0,016	0,033
200	Trennschärfe	0,000	0,000	0,053	0,448	0,998	0,998	0,954	0,884
	DR	0,926	0,969	1,003	1,034	1,127	1,150	1,116	1,083
	DP	0,997	0,890	0,461	0,153	0,001	0,001	0,011	0,025

**Tabelle 6** Simulierte Trennschärfen des SJ-Tests. Die Stichproben wurden aus schiefen Verteilungen mit nahezu normal, mittelstark und stärker besetzten Randbereichen generiert.

Verteilungen		Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8;1)	Chi(5)	Chi(3)	Exp
<b>n</b>	Wahres R	1,028	1,075	1,059	1,098	1,151
10	Trennschärfe	0,120	0,213	0,161	0,218	0,283
	DR	1,072	1,105	1,088	1,108	1,136
	DP	0,326	0,284	0,304	0,279	0,251
15	Trennschärfe	0,139	0,270	0,205	0,292	0,377
	DR	1,062	1,105	1,082	1,110	1,141
	DP	0,320	0,261	0,286	0,245	0,209
20	Trennschärfe	0,152	0,295	0,223	0,328	0,449
	DR	1,051	1,089	1,070	1,101	1,142
	DP	0,335	0,260	0,296	0,242	0,186

Verteilungen		Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8:1)	Chi(5)	Chi(3)	Exp
<b>n</b>	Wahres R	1,028	1,075	1,059	1,098	1,151
25	Trennschärf e	0,160	0,336	0,255	0,374	0,515
	DR	1,043	1,084	1,068	1,101	1,144
	DP	0,337	0,236	0,281	0,219	0,156
30	Trennschärf e	0,171	0,370	0,285	0,414	0,564
	DR	1,043	1,084	1,065	1,097	1,142
	DP	0,329	0,228	0,274	0,206	0,139
40	Trennschärf e	0,193	0,440	0,331	0,490	0,651
	DR	1,039	1,085	1,064	1,098	1,143
	DP	0,321	0,188	0,246	0,171	0,106
50	Trennschärf e	0,215	0,484	0,370	0,556	0,720
	DR	1,037	1,081	1,064	1,100	1,143
	DP	0,314	0,173	0,220	0,140	0,080
60	Trennschärf e	0,224	0,527	0,395	0,607	0,778
	DR	1,035	1,079	1,062	1,099	1,146
	DP	0,303	0,152	0,208	0,119	0,062
70	Trennschärf e	0,241	0,568	0,438	0,648	0,822
	DR	1,034	1,079	1,061	1,098	1,146
	DP	0,292	0,134	0,191	0,104	0,048
80	Trennschärf e	0,259	0,612	0,474	0,689	0,855
	DR	1,034	1,079	1,062	1,098	1,148
	DP	0,28	0,115	0,170	0,089	0,036
90	Trennschärf e	0,284	0,643	0,501	0,733	0,890
	DR	1,034	1,079	1,060	1,099	1,148
	DP	0,270	0,104	0,163	0,075	0,028



Verteilungen		Schiefe Verteilungen mit nahezu normal oder mittelstark besetzten Randbereichen			Schiefe Verteilungen mit stärker besetzten Randbereichen	
		Chi(10)	B(8:1)	Chi(5)	Chi(3)	Exp
<i>n</i>	Wahres R	1,028	1,075	1,059	1,098	1,151
100	Trennschärfe	0,285	0,675	0,527	0,757	0,912
	DR	1,032	1,078	1,060	1,098	1,147
	DP	0,267	0,094	0,151	0,067	0,022
120	Trennschärfe	0,323	0,728	0,572	0,816	0,942
	DR	1,032	1,077	1,060	1,098	1,149
	DP	0,246	0,074	0,129	0,050	0,014
140	Trennschärfe	0,344	0,769	0,621	0,852	0,963
	DR	1,031	1,077	1,060	1,099	1,148
	DP	0,232	0,060	0,112	0,036	0,009
160	Trennschärfe	0,363	0,815	0,666	0,887	0,978
	DR	1,031	1,077	1,060	1,098	1,150
	DP	0,217	0,047	0,093	0,027	0,005
180	Trennschärfe	0,385	0,843	0,692	0,910	0,986
	DR	1,031	1,077	1,059	1,099	1,148
	DP	0,209	0,039	0,083	0,021	0,004
200	Trennschärfe	0,410	0,877	0,727	0,931	0,989
	DR	1,030	1,077	1,059	1,098	1,149
	DP	0,196	0,030	0,071	0,016	0,003

Unsere Simulationsergebnisse in den Tabellen 5 und 6 stimmen mit denen überein, die in Gel et al. (2007) veröffentlicht wurden. Wenn die Stichproben aus den normalverteilten Grundgesamtheiten stammen, liegen die simulierten Trennschärfen (die in diesem Fall das tatsächliche Signifikanzniveau des Tests darstellen) selbst bei kleinen Stichprobenumfängen wie 25 nicht weit vom Sollwert entfernt. Wenn die Stichproben aus Verteilungen mit stärker besetzten Randbereichen stammen, ist die Trennschärfe des Tests bei kleinen Stichprobenumfängen gering; sie erhöht sich jedoch auf mindestens 40 %, wenn der Stichprobenumfang den Wert 40 erreicht. Konkret beträgt die Trennschärfe bei einem Stichprobenumfang von 40 etwa 40,1 % für die t-Verteilung mit 5 Freiheitsgraden, 66,5 % für die Laplace-Verteilung und 65,1 % für die Exponentialverteilung.

Bei Verteilungen mit schwächer besetzten Randbereichen (die Beta(3;3)-Verteilung und die Gleichverteilung) liegt die Trennschärfe des Tests bei kleinen Stichproben nahe 0, und sie verringert sich bei zunehmendem Stichprobenumfang weiter. Dies überrascht nicht, da die Ergebnisse für diese Verteilungen tatsächlich die Alternativhypothese einer Verteilung mit schwächer besetzten Randbereichen stützen und nicht die Alternativhypothese einer Verteilung mit stärker besetzten Randbereichen.

Wenn die Stichproben aus Verteilungen mit etwas stärker besetzten Randbereichen gewonnen werden, z. B. aus der t-Verteilung mit 10 Freiheitsgraden oder der Chi-Quadrat-Verteilung mit 10 Freiheitsgraden, sind die Trennschärfen für mittlere bis große Stichprobenumfänge niedrig. Für unsere Zwecke kann dies als gutes Resultat erachtet werden, da der Test auf Varianzen (Standardabweichung) bei einer Stichprobe für diese Verteilungen eine gute Leistung zeigt und diese Verteilungen nicht als Verteilungen mit stärker besetzten Randbereichen identifiziert werden sollen. Mit zunehmendem Stichprobenumfang steigt jedoch auch die Trennschärfe des Tests, so dass diese Verteilungen mit etwas stärker besetzten Randbereichen als Verteilungen mit stärker besetzten Randbereichen erkannt werden.

Deshalb muss in den Regeln zum Auswerten der Besetzung der Randbereiche der Verteilung für diesen Test auch der Stichprobenumfang berücksichtigt werden. Hierfür kann beispielsweise ein Konfidenzintervall für das Maß der Besetzung der Randbereiche berechnet werden; die Verteilung der SJ-Statistik ist jedoch extrem empfindlich gegenüber der übergeordneten Verteilung der Stichprobe. Ein alternativer Ansatz besteht darin, die Besetzung der Randbereiche der Verteilung sowohl anhand der Stärke der Zurückweisung der Nullhypothese des SJ-Tests als auch anhand des Stichprobenumfangs auszuwerten. Konkret bedeutet dies: Kleinere p-Werte verweisen auf stärker besetzte Randbereiche, und größere p-Werte geben schwächer besetzte Randbereiche an. Größere Stichproben weisen jedoch tendenziell kleinere p-Werte als kleinere Stichproben auf. Daher entwickeln wir auf der Grundlage der simulierten Trennschärfen, Stichprobenumfänge und durchschnittlichen p-Werte in Tabelle 3 eine Reihe von allgemeinen Regeln, anhand derer die Randbereiche einer Verteilung für jede Stichprobe unter Verwendung des SJ-Tests ausgewertet werden.

Wenn für mittlere bis große Stichprobenumfänge ( $40 \leq n \leq 100$ ) der p-Wert zwischen 0,01 und 0,05 liegt, erachten wir dies als schwachen Hinweis auf Zurückweisung der Nullhypothese. Das heißt, die Verteilung der Stichprobe wird als Verteilung mit mittelstark besetzten Randbereichen (Typ „Mittel“) klassifiziert. Wenn der p-Wert hingegen unter 0,01 liegt, gibt es starke Belege gegen die Nullhypothese, und die übergeordnete Verteilung der Stichprobe wird als Verteilung mit stärker besetzten Randbereichen (Typ „Stark“) klassifiziert.

Für große Stichproben ( $n > 100$ ) kategorisieren wir die übergeordnete Verteilung als Typ „Mittel“, wenn der p-Wert in den Bereich zwischen 0,005 und 0,01 fällt, und als Typ „Stark“, wenn der p-Wert außerordentlich klein (unter 0,005) ist. Beachten Sie, dass die Trennschärfe des SJ-Tests bei einem Stichprobenumfang unter 40 generell zu gering ist, um die Verteilung der Stichprobe effektiv zu bestimmen.

Die allgemeinen Klassifikationsregeln für die Gültigkeit des Tests auf Varianzen bei einer Stichprobe mit der Bonett-Methode werden in der nachfolgenden Tabelle 7 zusammengefasst.

**Tabelle 7** Klassifikationsregeln zum Bestimmen der übergeordneten Verteilung der einzelnen Stichproben ( $p$  ist der p-Wert des SJ-Tests)

Bedingung	Verteilungstyp
$n < 40$	Wird nicht bestimmt
$100 \geq n \geq 40$ und $p > 0,05$	Typ „Schwach“
$n > 100$ und $p > 0,01$	Typ „Schwach“
$40 \leq n \leq 100$ und $0,01 < p \leq 0,05$	Typ „Mittel“
$n > 100$ und $0,005 < p \leq 0,01$	Typ „Mittel“
$40 \leq n \leq 100$ und $p \leq 0,01$	Typ „Stark“
$n > 100$ und $p \leq 0,005$	Typ „Stark“

Wie bereits zuvor erwähnt und entsprechend den Ergebnissen aus Tabelle 1 und Tabelle 2 in Simulation C1, beträgt der approximierte Stichprobenumfang mindestens 20, 80 bzw. 200, wenn die Stichproben aus einer Verteilung vom Typ „Schwach“, einer Verteilung vom Typ „Mittel“ bzw. einer Verteilung vom Typ „Stark“ stammen, um eine Überdeckungswahrscheinlichkeit von mindestens 0,93 zu erzielen. Da die Trennschärfe des SJ-Tests für kleine Stichproben jedoch gering ist, wird die Mindestanforderung für den Stichprobenumfang für Verteilungen vom Typ „Schwach“ auf 40 festgelegt.

## Simulation F2: Prüfen der Regeln zum Klassifizieren von Verteilungen

Wir haben Stichproben aus einigen der in Simulation C1 beschriebenen Verteilungen generiert und mit dem SJ-Test die Anteile der Stichproben bestimmt, die in eine von drei Verteilungsgruppen eingestuft wurden: Typ „Schwach“, Typ „Mittel“ und Typ „Stark“. Die Simulationsergebnisse werden in Tabelle 8 aufgeführt.

**Tabelle 8** Anteil der 10.000 Stichproben mit unterschiedlichem Umfang aus verschiedenen Verteilungen, die als Verteilungen vom Typ „Schwach“, Typ „Mittel“ und Typ „Stark“ identifiziert werden

$n$	Verteilungen	Typ „Schwach“				Typ „Mittel“	Typ „Stark“	
		B(3;3)	N(0;1)	t(10)	Chi(10)	Chi(5)	Lpl	Exp
40	% Typ „Schwach“	99,6	94,0	81,5	80,3	66,6	33,0	34,4
	% Typ „Mittel“	0,4	5,5	14,0	14,0	20,0	31,9	22,9
	% Typ „Stark“	0,0	0,5	4,5	5,7	13,4	35,1	42,8

n	Verteilungen	Typ „Schwach“				Typ „Mittel“	Typ „Stark“	
		B(3;3)	N(0;1)	t(10)	Chi(10)	Chi(5)	Lpl	Exp
50	% Typ „Schwach“	99,7	94,4	78,7	79,1	64,0	25,1	28,0
	% Typ „Mittel“	0,3	5,1	15,6	14,2	20,0	29,9	20,7
	% Typ „Stark“	0,0	0,5	5,7	6,7	16,0	45,0	51,3
60	% Typ „Schwach“	99,7	94,5	77,3	77,3	59,1	18,5	22,6
	% Typ „Mittel“	0,3	5,1	16,4	15,0	22,0	27,4	19,2
	% Typ „Stark“	0,0	0,5	6,3	7,7	18,9	54,1	58,2
70	% Typ „Schwach“	99,8	94,4	74,5	75,2	55,9	14,0	18,1
	% Typ „Mittel“	0,2	5,0	18,1	16,0	22,2	24,0	17,5
	% Typ „Stark“	0,0	0,6	7,4	8,8	21,9	62,0	64,4
80	% Typ „Schwach“	99,9	94,3	74,1	74,4	53,0	10,0	13,9
	% Typ „Mittel“	0,1	5,1	17,8	16,7	22,8	21,0	15,5
	% Typ „Stark“	0,0	0,6	8,2	8,9	24,2	69,0	70,6
90	% Typ „Schwach“	99,9	94,4	71,2	72,1	49,5	7,5	11,1
	% Typ „Mittel“	0,1	5,0	19,1	17,2	22,6	16,5	13,7
	% Typ „Stark“	0,0	0,6	9,7	10,7	27,9	76,0	75,3
100	% Typ „Schwach“	99,9	94,5	70,8	70,3	47,3	4,8	8,9
	% Typ „Mittel“	0,1	4,9	19,5	17,9	22,7	14,3	11,8
	% Typ „Stark“	0,0	0,6	9,7	11,8	30,0	80,9	79,4
120	% Typ „Schwach“	100,0	99,4	87,4	87,2	64,8	12,0	14,4
	% Typ „Mittel“	0,0	0,4	5,0	4,5	7,9	7,8	5,6
	% Typ „Stark“	0,0	0,2	7,6	8,4	27,4	80,4	80,0
140	% Typ „Schwach“	100,0	99,3	86,0	85,1	60,5	7,0	9,9
	% Typ „Mittel“	0,0	0,5	5,2	5,0	8,6	5,6	4,1
	% Typ „Stark“	0,0	0,2	8,8	9,9	30,9	87,4	86,0
160	% Typ „Schwach“	100,0	99,4	83,4	83,0	55,6	4,0	6,9
	% Typ „Mittel“	0,0	0,5	6,3	5,8	9,5	3,5	3,0
	% Typ „Stark“	0,0	0,1	10,4	11,2	34,9	92,5	90,1
180	% Typ „Schwach“	100,0	99,3	81,1	81,7	51,0	2,5	4,6
	% Typ „Mittel“	0,0	0,5	6,8	5,9	9,4	1,9	2,2
	% Typ „Stark“	0,0	0,2	12,1	12,4	39,6	95,6	93,2
200	% Typ „Schwach“	100,0	99,5	79,0	80,5	47,2	1,3	3,0
	% Typ „Mittel“	0,0	0,4	7,6	6,1	9,4	1,6	1,7
	% Typ „Stark“	0,0	0,1	13,4	13,4	43,4	97,1	95,3

Die Ergebnisse in Tabelle 8 zeigen Folgendes: Wenn Stichproben aus Verteilungen mit schwächer besetzten (Typ „Schwach“) und stärker besetzten (Typ „Stark“) Randbereichen stammen, wird ein größerer Anteil der Stichproben ordnungsgemäß klassifiziert. Wurden beispielsweise Stichproben mit dem Umfang 40 aus der Beta(3;3)-Verteilung generiert, wurden 99,6 % der Stichproben richtig als Verteilungen mit schwächer besetzten Randbereichen klassifiziert; wenn Stichproben mit dem Umfang 90 aus der Laplace-Verteilung generiert wurden, wurden 76,0 % richtig als Verteilungen mit stärker besetzten Randbereichen klassifiziert. Daher werden in der Auswertung nicht fälschlicherweise Warnmeldungen zur Gültigkeit des Tests ausgegeben, wenn Stichproben tatsächlich aus Verteilungen mit schwächer besetzten Randbereichen stammen, und die Warnmeldungen werden ordnungsgemäß ausgegeben, wenn Stichproben aus Verteilungen mit stärker besetzten Randbereichen stammen und die Anforderung in Bezug auf den minimalen Stichprobenumfang nicht erfüllt ist. Darüber hinaus wurde für Stichproben aus Verteilungen mit mittelstark besetzten Randbereichen (Typ „Mittel“) wie die Chi-Quadrat(5)-Verteilung festgestellt, dass ein größerer Anteil von Stichproben fälschlicherweise als Verteilungen mit schwächer besetzten Randbereichen (Typ „Schwach“) klassifiziert wird, wenn die Stichproben klein sind (bei einem Stichprobenumfang von 40 werden 66 % der Stichproben fälschlicherweise als Verteilung mit schwächer besetzten Randbereichen klassifiziert). Daher werden für diese Fälle u. U. keine Warnmeldungen in der Auswertung ausgegeben, selbst wenn die übergeordneten Verteilungen mittelstark besetzte Randbereiche aufweisen. Wenn der Stichprobenumfang jedoch größer als 80 ist, hat die Fehlklassifikation als Verteilung vom Typ „Schwach“ keine Auswirkungen, da die Anforderung in Bezug auf den minimalen Stichprobenumfang bereits erfüllt wurde.

© 2015, 2017 Minitab Inc. All rights reserved.

Minitab®, Quality. Analysis. Results.® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries. See [minitab.com/legal/trademarks](http://minitab.com/legal/trademarks) for more information.