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.

t-Test bei einer Stichprobe

Übersicht

Mit dem t-Test bei einer Stichprobe wird der Mittelwert eines Prozesses geschätzt und mit einem Sollwert verglichen. Dieser Test wird als robust erachtet, da er gegenüber der Annahme der Normalverteilung äußerst unempfindlich ist, sofern die Stichprobe einen relativ großen Stichprobenumfang aufweist. Gemäß den meisten statistischen Fachbüchern sind der t-Test bei einer Stichprobe und das t-Konfidenzintervall für den Mittelwert für jede Stichprobe mit einem Stichprobenumfang ab 30 geeignet.

Im vorliegenden White Paper werden die Simulationen beschrieben, mit denen wir diese allgemeine Regel eines Minimums von 30 Einheiten pro Stichprobe untersucht haben. Der Schwerpunkt der Simulationen lag auf der Auswirkung einer fehlenden Normalverteilung auf den t-Test bei einer Stichprobe. Zudem sollte die Auswirkung ungewöhnlicher Daten auf die Testergebnisse ausgewertet werden.

Auf der Grundlage unserer Untersuchungen führt der Assistent automatisch die folgenden Prüfungen Ihrer Daten durch und zeigt die Ergebnisse in der Auswertung an:

- Ungewöhnliche Daten
- Vorliegen einer Normalverteilung (Ist die Stichprobe groß genug, dass eine fehlende Normalverteilung kein Problem darstellt?)
- Stichprobenumfang

Allgemeine Informationen zur Methodologie für den t-Test bei einer Stichprobe finden Sie in Arnold (1990), Casella und Berger (1990), Moore und McCabe (1993) sowie Srivastava (1958).

Hinweis Die Ergebnisse in diesem White Paper gelten auch für den t-Test bei verbundenen Stichproben im Assistenten, da beim t-Test bei verbundenen Stichproben die Methode für den t-Test bei einer Stichprobe auf eine Stichprobe von paarweisen Differenzen angewendet 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. Bei einem kleinen Stichprobenumfang können sie sich auf die Wahrscheinlichkeiten auswirken, dass statistisch signifikante Ergebnisse gefunden werden. Ungewöhnliche Daten können auf Probleme bei der Datenerfassung hinweisen, sie können aber auch auf ein ungewöhnliches Verhalten des untersuchten Prozesses zurückzuführen sein. Häufig ist es unverzichtbar, diese Datenpunkte zu untersuchen und nach Möglichkeit zu korrigieren.

Zielstellung

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

Methode

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

Ergebnisse

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

Für die Prüfung auf ungewöhnliche Daten werden in der Auswertung des Assistenten für den t-Test bei einer Stichprobe die folgenden Statusindikatoren angezeigt:

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

Vorliegen einer Normalverteilung

Der t-Test bei einer Stichprobe wird unter der Annahme abgeleitet, dass die Grundgesamtheit normalverteilt ist. Doch selbst wenn die Daten nicht normalverteilt sind, funktioniert dieses Verfahren gut, sofern die Stichprobe umfassend genug ist.

Zielstellung

Wir wollten den Effekt einer fehlenden Normalverteilung auf den Fehler 1. Art und 2. Art des Tests bestimmen, um Richtlinien in Bezug auf Stichprobenumfang und Normalverteilung zu erhalten.

Methode

Wir haben Simulationen durchgeführt, um den Stichprobenumfang zu bestimmen, für den die Annahme der Normalverteilung beim Durchführen eines t-Tests bei einer Stichprobe oder beim Berechnen eines t-Konfidenzintervalls für den Mittelwert einer Grundgesamtheit ignoriert werden kann.

Die erste Studie wurde zum Auswerten des Effekts einer fehlenden Normalverteilung auf die Wahrscheinlichkeit eines Fehlers 1. Art des Tests konzipiert. Insbesondere sollte der Stichprobenumfang ermittelt werden, der mindestens erforderlich ist, damit der Test in Bezug auf die Verteilung der Grundgesamtheit unempfindlich ist. Der t-Test bei einer Stichprobe wurde für kleine, mittlere und große Stichproben durchgeführt, die aus normalverteilten und nicht normalverteilten Grundgesamtheiten generiert wurden. Zu den nicht normalverteilten Grundgesamtheiten zählten leicht bis stark schiefe Grundgesamtheiten, symmetrische Grundgesamtheiten mit schwächer und stärker besetzten Randbereichen sowie Grundgesamtheiten mit kontaminierter Normalverteilung. Die normalverteilte Grundgesamtheit diente als Kontroll-Grundgesamtheit zu Vergleichszwecken. Für jeden Fall wurden die simulierten Signifikanzniveaus berechnet und mit dem Soll-Signifikanzniveau (dem nominalen Signifikanzniveau) von 0,05 verglichen. Wenn der Test eine gute Leistung zeigt, sollten die simulierten Signifikanzniveaus nahe bei 0,05 liegen. Die simulierten Signifikanzniveaus für alle unterschiedlichen Bedingungen wurden untersucht, um den kleinsten Stichprobenumfang zu ermitteln, für den sie ungeachtet der Verteilung nahe dem Soll-Signifikanzniveau verbleiben. Weitere Informationen finden Sie in Anhang A.

In der zweiten Studie wurde der Effekt einer fehlenden Normalverteilung auf den Fehler 2. Art des Tests untersucht. Die Simulation wurde wie in der ersten Studie eingerichtet. Es wurden jedoch simulierte Trennschärfen unter unterschiedlichen Bedingungen mit den Soll-Trennschärfen verglichen, die mit der theoretischen Trennschärfefunktion des t-Tests bei einer Stichprobe berechnet wurden. Weitere Informationen finden Sie in Anhang B.

Ergebnisse

Der Effekt einer fehlenden Normalverteilung auf die Wahrscheinlichkeiten eines Fehlers 1. Art und 2. Art des Tests ist minimal, selbst wenn die Stichprobenumfänge lediglich 20 betragen. Wenn die übergeordnete Grundgesamtheit der Stichprobe jedoch extrem schief ist, sind u. U. größere Stichproben erforderlich. Wir empfehlen, für derartige Fälle einen Stichprobenumfang von etwa 40 zu wählen. Weitere Informationen finden Sie in Anhang A und Anhang B.

Da der Test bei relativ kleinen Stichproben eine gute Leistung zeigt, testet der Assistent die Daten nicht auf eine Normalverteilung. Stattdessen wird der Stichprobenumfang der Stichprobe überprüft, und in der Auswertung werden die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Der Stichprobenumfang beträgt mindestens 20, daher ist es kein Problem, wenn keine Normalverteilung vorliegt.
	Da der Stichprobenumfang kleiner als 20 ist, könnte es ein Problem sein, wenn keine Normalverteilung vorliegt.

Stichprobenumfang

Normalerweise wird ein Hypothesentest durchgeführt, um einen Beleg für die Zurückweisung der Nullhypothese („keine Differenz“) zu erhalten. Wenn die Stichproben zu klein sind, reicht die Trennschärfe des Tests u. U. nicht aus, um eine tatsächlich vorhandene Differenz zwischen den Mittelwerten zu erkennen; hierbei handelt es sich um einen Fehler 2. Art. Daher muss unbedingt sichergestellt werden, dass der Stichprobenumfang ausreichend groß ist, 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 der Stichprobenumfang groß genug für den Test ist, so dass dieser mit hoher Wahrscheinlichkeit Differenzen mit praktischen Konsequenzen erkennt. Bei der Planung des Stichprobenumfangs soll zwar sichergestellt werden, dass dieser ausreichend groß ist, um mit hoher Wahrscheinlichkeit wichtige Differenzen zu erkennen; andererseits darf er aber nicht so groß sein, dass bedeutungslose Differenzen mit hoher Wahrscheinlichkeit statistisch signifikant werden.

Methode

Die Analyse der Trennschärfe und des Stichprobenumfangs basiert auf der theoretischen Trennschärfefunktion des spezifischen Tests, mit dem die statistische Analyse durchgeführt wird. Wie bereits weiter oben erläutert, ist die Trennschärfefunktion des t-Tests bei einer Stichprobe gegenüber der Annahme der Normalverteilung unempfindlich, wenn der Stichprobenumfang mindestens 20 beträgt. Die Trennschärfefunktion hängt vom Stichprobenumfang, der Differenz zwischen dem Soll-Mittelwert und dem Mittelwert der Grundgesamtheit sowie von der Varianz der Grundgesamtheit ab. Weitere Informationen finden Sie in Anhang B.

Ergebnisse

Wenn die Daten keine ausreichenden Hinweise liefern, die gegen die Nullhypothese sprechen, berechnet der Assistent Differenzen mit praktischen Konsequenzen, die für die angegebenen Stichprobenumfänge mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt

werden können. Wenn der Benutzer zudem eine konkrete Differenz mit praktischen Konsequenzen angibt, berechnet der Assistent den Stichprobenumfang, bei dem die Differenz mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt wird.

Wir können an dieser Stelle keine allgemeingültigen Ergebnisse aufführen, da die Ergebnisse von der spezifischen Stichprobe des Benutzers abhängen. In Anhang B finden Sie jedoch weitere Informationen zur Trennschärfe für den t-Test bei einer Stichprobe.

Für die Prüfung auf die Trennschärfe und den Stichprobenumfang werden in der Auswertung des Assistenten für den t-Test bei einer Stichprobe die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Im Test wird eine Differenz zwischen dem Mittelwert und dem Sollwert festgestellt, daher stellt die Trennschärfe kein Problem dar. ODER Die Trennschärfe ist ausreichend. Im Test wurde keine Differenz zwischen dem Mittelwert und dem Sollwert 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 dem Mittelwert und dem Sollwert festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 80 % bis 90 % erkannt wird. Der erforderliche Stichprobenumfang zum Erzielen einer Trennschärfe von 90 % wird ausgegeben.
	Die Trennschärfe ist möglicherweise nicht ausreichend. Im Test wurde keine Differenz zwischen dem Mittelwert und dem Sollwert festgestellt, und die Stichprobe ist umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 60 % bis 80 % erkannt wird. Die erforderlichen Stichprobenumfänge zum Erzielen einer Trennschärfe von 80 % und 90 % werden ausgegeben.
	Die Trennschärfe ist nicht ausreichend. Im Test wurde keine Differenz zwischen dem Mittelwert 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 Erzielen einer Trennschärfe von 80 % und 90 % werden ausgegeben.
	Im Test wurde keine Differenz zwischen dem Mittelwert und dem Sollwert erkannt. Sie haben keine zu erkennende Differenz mit praktischen Konsequenzen zwischen dem Mittelwert und dem Sollwert angegeben; daher werden in der Auswertung die Differenzen angegeben, die bei Ihren Stichprobenumfängen, Standardabweichungen und Alpha mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden.

Literaturhinweise

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

Casella, G. und Berger, R. L. (1990). *Statistical inference*. Pacific Grove, CA: Wadsworth, Inc.

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

Moore, D.S. und McCabe, G.P. (1993). *Introduction to the practice of statistics*, 2nd ed. New York, NY: W. H. Freeman and Company.

Neyman, J., Iwaszkiewicz, K. und Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation, *Journal of the Royal Statistical Society, Series B, 2, 107-180*.

Pearson, E.S. und Hartley, H.O. (Hrsg.). (1954). *Biometrika tables for statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of t-test, *Biometrika*, *45, 421-429*.

Anhang A: Auswirkung einer fehlenden Normalverteilung auf das Signifikanzniveau (Gültigkeit des Tests)

Unter der Annahme der Normalverteilung ist der t-Test bei einer Stichprobe ein *gleichmäßig trennschärfster* (uniformly most powerful, UMP) *und unverzerrter* Test mit Niveau α. Das heißt, der Test weist die gleiche oder eine bessere Trennschärfe als jeder andere unverzerrte Test mit Niveau α um den Mittelwert auf. Doch auch wenn die übergeordnete Grundgesamtheit der Stichprobe nicht normalverteilt ist, sind die oben genannten Optimalitätseigenschaften gültig, sofern die Stichprobe ausreichend groß ist. Mit anderen Worten, für ausreichend große Stichproben entspricht das tatsächliche Signifikanzniveau des t-Tests bei einer Stichprobe sowohl für normalverteilte als auch für nicht normalverteilte Daten ungefähr dem Sollniveau, und die Trennschärfefunktion des Tests ist ebenso unempfindlich in Bezug auf die Annahme der Normalverteilung (Srivastava, 1958).

Wir wollten untersuchen, ab welchem Umfang eine Stichprobe als ausreichend groß erachtet werden kann, dass der t-Test gegenüber der Annahme der Normalverteilung unempfindlich ist. In vielen Fachbüchern wird empfohlen, dass bei einem Stichprobenumfang von $n \geq 30$ die Annahme der Normalverteilung für die meisten praktischen Zwecke ignoriert werden kann (Arnold, 1990; Casella und Berger, 1990 sowie Moore und McCabe, 1993). Der Zweck der in diesen Anhängen beschriebenen Untersuchung ist die Durchführung von Simulationsstudien, in denen diese allgemeine Regel auswertet werden soll. Dazu wird die Auswirkung unterschiedlicher Nicht-Normalverteilungen auf den t-Test bei einer Stichprobe untersucht.

Simulationsstudie A

Wir wollten die Auswirkung einer fehlenden Normalverteilung auf die Wahrscheinlichkeit eines Fehlers 1. Art des Tests untersuchen, um einen mindestens erforderlichen Stichprobenumfang zu bestimmen, bei der sie ungeachtet der Verteilung stabil und nahe der Soll-Fehlerwahrscheinlichkeit bleibt.

Hierzu wurden zwei beidseitige t-Tests mit $\alpha = 0{,}05$ unter Verwendung von Zufallsstichproben von unterschiedlichem Umfang ($n = 10, 15, 20, 25, 30, 35, 40, 50, 60, 80, 100$) durchgeführt, die aus verschiedenen Verteilungen mit unterschiedlichen Eigenschaften generiert wurden. Hierzu zählten folgende Verteilungen:

- Standardnormalverteilung (N(0;1))
- Symmetrische Verteilungen mit stärker besetzten Randbereichen, z. B. die t-Verteilung mit 5 und 10 Freiheitsgraden (t(5), t(10))
- Laplace-Verteilung mit Lage 0 und Skala 1 (Lpl))

- Schiefe Verteilungen mit stärker besetzten Randbereichen, vertreten durch die Exponentialverteilung mit Skala 1 (Exp), die Chi-Quadrat-Verteilungen mit 3, 5 und 10 Freiheitsgraden (Chi(3), Chi(5), Chi(10))
- 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))
- Eine linksschiefe Verteilung mit stärker besetzten Randbereichen (B(8;1))

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

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

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

Für jeden Stichprobenumfang wurden aus jeder Grundgesamtheit 10.000 Stichprobenreplikationen gezogen, und für jede der 10.000 Stichproben wurde ein t-Test bei einer Stichprobe mit der Nullhypothese $\mu = \mu_o$ und der Alternativhypothese $\mu \neq \mu_o$ durchgeführt. Für jeden Test wird der Hypothesenmittelwert μ_o auf den tatsächlichen Mittelwert der übergeordneten Grundgesamtheit der Stichprobe festgelegt. Deshalb stellt für einen bestimmten Stichprobenumfang der Anteil der 10.000 Stichprobenreplikationen, die zu einer Zurückweisung der Nullhypothese führen, die simulierte Wahrscheinlichkeit eines Fehlers 1. Art bzw. das Signifikanzniveau des Tests dar. Da das Soll-Signifikanzniveau 5 % ist, beträgt der Simulationsfehler ca. 0,2 %.

Die Simulationsergebnisse werden in den Tabellen 1 und 2 aufgeführt und in den Abbildungen 1 und 2 grafisch veranschaulicht.

Ergebnisse und Zusammenfassung

Die Ergebnisse (siehe Tabelle 1 und Abbildung 1) zeigen, dass die simulierten Signifikanzniveaus des Tests nahe dem Soll-Signifikanzniveau liegen, wenn die Stichproben aus symmetrischen Grundgesamtheiten stammen. Dies gilt selbst für kleine Stichproben. Die Testergebnisse sind jedoch leicht konservativ bei symmetrischen Verteilungen mit stärker besetzten Randbereichen, bei denen die Stichproben klein sind. Dies gilt auch für kleine Stichproben, die aus den kontaminierten Normalverteilungen gezogen wurden. Zudem scheint es, dass Ausreißer bei kleinen Stichproben das Signifikanzniveau des Tests verringern. Dieser Effekt wird jedoch umgekehrt, wenn kleine Stichproben aus übergeordneten symmetrischen Grundgesamtheiten mit schwächer besetzten Randbereichen (Betaverteilung (3;3) und Gleichverteilung) generiert werden. Für diese Fälle sind die simulierten Signifikanzniveaus etwas höher.

Tabelle 1 Simulierte Signifikanzniveaus für den beidseitigen t-Test bei einer Stichprobe für Stichproben, die aus symmetrischen Grundgesamtheiten generiert wurden. Das Soll-Signifikanzniveau ist $\alpha = 0{,}05$.

Vert.	N(0;1)	t(5)	t(10)	Lpl	CN(0,9;3)	CN(0,8;3)	B(3;3)	U(0;1)
N	Symmetrisch und stärker besetzte Randbereiche						Symmetrisch und schwächer besetzte Randbereiche	
10	0,050	0,046	0,048	0,044	0,043	0,039	0,057	0,057
15	0,051	0,050	0,049	0,049	0,043	0,043	0,053	0,054
20	0,047	0,051	0,051	0,047	0,043	0,044	0,051	0,052
25	0,050	0,047	0,050	0,046	0,046	0,046	0,048	0,050
30	0,053	0,050	0,048	0,043	0,049	0,046	0,050	0,048
35	0,052	0,047	0,049	0,050	0,047	0,045	0,051	0,054
40	0,046	0,052	0,054	0,048	0,046	0,049	0,044	0,050
50	0,050	0,049	0,051	0,048	0,047	0,051	0,053	0,050
60	0,052	0,049	0,053	0,050	0,051	0,056	0,054	0,052
80	0,049	0,050	0,051	0,047	0,047	0,052	0,049	0,049
100	0,050	0,052	0,049	0,051	0,052	0,054	0,051	0,054

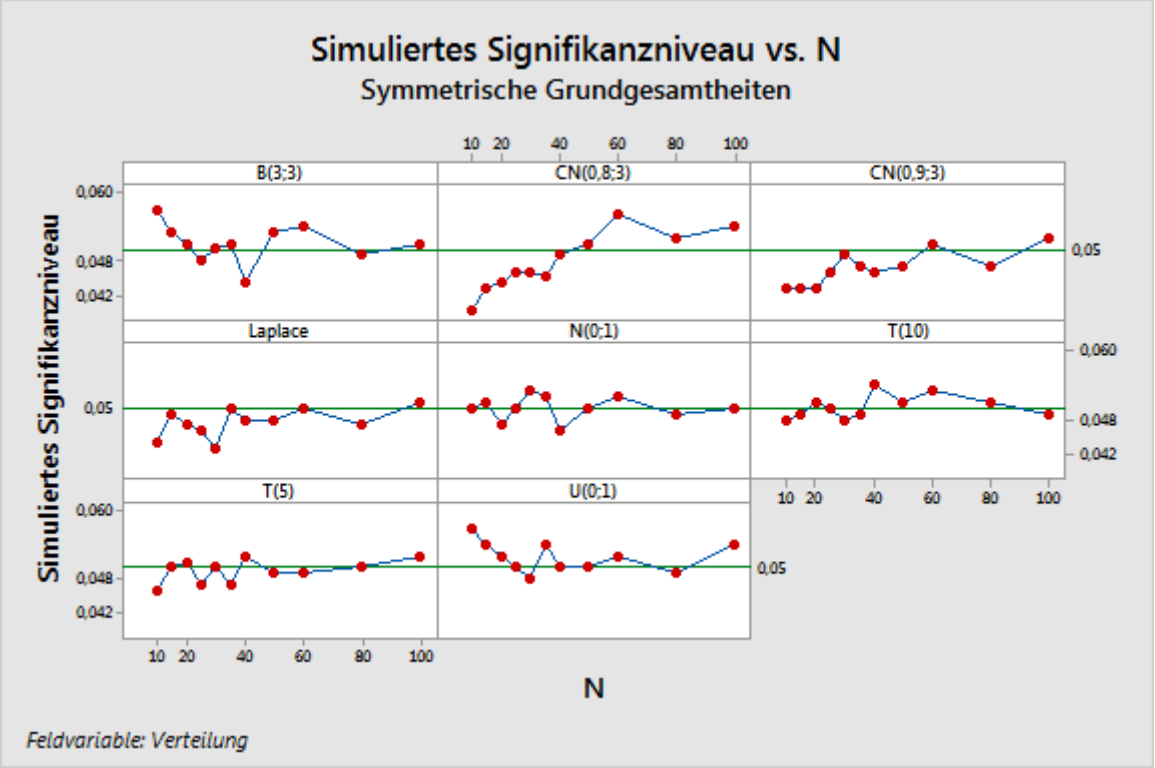

Abbildung 1 Diagramm der simulierten Trennschärfen für beidseitige t-Tests bei einer Stichprobe im Vergleich zum Umfang der Stichproben, die aus symmetrischen Grundgesamtheiten generiert wurden. Das Soll-Signifikanzniveau ist $\alpha = 0{,}05$.

Wenn Stichproben hingegen aus schiefen Verteilungen generiert werden, hängt die Leistung des Tests vom Ausmaß der Schiefe ab. Die Ergebnisse in Tabelle 2 und Abbildung 2 zeigen, dass der t-Test bei einer Stichprobe empfindlich gegenüber Schiefe in kleinen Stichproben ist. Bei stark schiefen Grundgesamtheiten (Exponential, Chi(3) und Beta(8;1)) sind größere Stichproben erforderlich, damit die simulierten Signifikanzniveaus nahe dem Soll-Signifikanzniveau liegen. Bei mäßig schiefen Grundgesamtheiten wie (Chi(5) und Chi(10)) reicht jedoch ein minimaler Stichprobenumfang von 20 aus, damit die simulierten Signifikanzniveaus nahe dem Sollniveau liegen. Beim einem Stichprobenumfang von 20 beträgt das simulierte Signifikanzniveau für die Chi-Quadrat-Verteilung mit 5 Freiheitsgraden ca. 0,063 und für die Chi-Quadrat-Verteilung mit 10 Freiheitsgraden ca. 0,056.

Tabelle 2 Simulierte Signifikanzniveaus für den beidseitigen t-Test bei einer Stichprobe für Stichproben, die aus schiefen Grundgesamtheiten generiert wurden. Das Soll-Signifikanzniveau ist $\alpha = 0{,}05$.

N	Exp	Chi(3)	B(8;1)	Chi(5)	Chi(10)
	Schiefe der Grundgesamtheit				
	2,0	1,633	-1,423	1,265	0,894
	Simulierte Signifikanzniveaus				
10	0,101	0,089	0,087	0,069	0,060
15	0,088	0,076	0,072	0,068	0,057
20	0,083	0,073	0,069	0,063	0,056
25	0,075	0,068	0,067	0,067	0,056
30	0,069	0,067	0,066	0,058	0,054
35	0,075	0,067	0,062	0,062	0,056
40	0,067	0,067	0,061	0,059	0,056
50	0,064	0,057	0,062	0,057	0,054
60	0,063	0,056	0,061	0,054	0,055
80	0,059	0,058	0,053	0,052	0,052
100	0,060	0,055	0,055	0,047	0,053

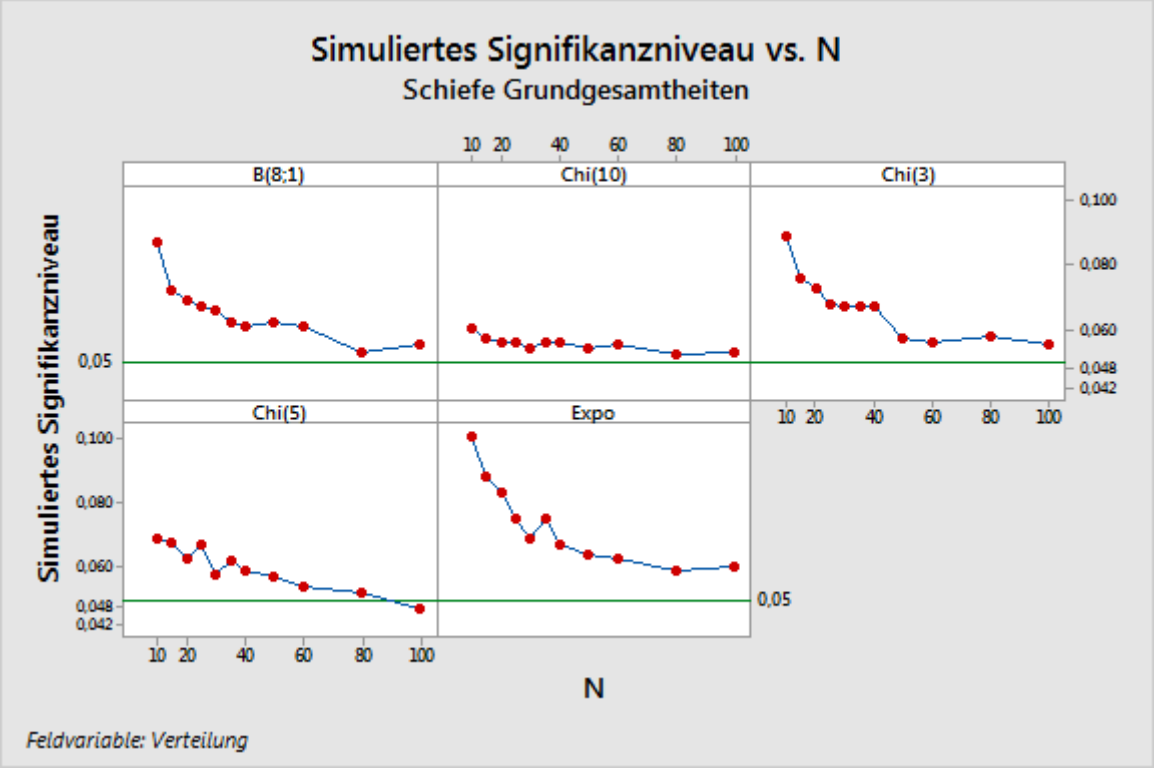

Abbildung 2 Diagramm der simulierten Trennschärfen für beidseitige t-Tests bei einer Stichprobe im Vergleich zum Umfang der Stichproben, die aus schiefen Grundgesamtheiten generiert wurden. Das Soll-Signifikanzniveau ist $\alpha = 0{,}05$.

Der Schwerpunkt dieser Untersuchung lag auf den Hypothesentests und nicht auf den Konfidenzintervallen. Die Ergebnisse beziehen sich aber natürlich auch auf Konfidenzintervalle, da die statistische Signifikanz sowohl mit Hypothesentests als auch mit Konfidenzintervallen bestimmt werden kann.

Anhang B: Stichprobenumfang und Trennschärfe des Tests

Wir wollten die Empfindlichkeit der Trennschärfefunktion gegenüber der Annahme der Normalverteilung, unter der sie abgeleitet wird, untersuchen. Beachten Sie zunächst: Wenn β der Fehler 2. Art eines Tests ist, dann ist $1 - \beta$ die Trennschärfe des Tests. Deshalb wird der geplante Stichprobenumfang so berechnet, dass die Wahrscheinlichkeit eines Fehlers 2. Art klein bzw. die Trennschärfe angemessen hoch ist (was denselben Sachverhalt ausdrückt).

Die Trennschärfefunktionen für t-Tests sind hinreichend bekannt und dokumentiert. Pearson und Hartley (1954) sowie Neyman, Iwaszkiewicz und Kolodziejczyk (1935) haben Diagramme und Tabellen der Trennschärfefunktionen veröffentlicht.

Für einen beidseitigen t-Test bei einer Stichprobe mit dem Niveau α kann diese Funktion mit dem Stichprobenumfang und der Differenz δ zwischen dem tatsächlichen Mittelwert μ und dem Hypothesenmittelwert μ_o ausgedrückt werden als

$$\pi(n;\delta) = 1 - F_{n-1,\lambda}\left(t_{n-1}^{\alpha/2}\right) + F_{n-1,\lambda}\left(-t_{n-1}^{\alpha/2}\right)$$

Hierbei ist $\mathrm{F_{d,\lambda}(.)}$ die kumulative Verteilungsfunktion der nicht zentralen t-Verteilung mit $\mathrm{d} = \mathrm{n} - 1$ Freiheitsgraden und dem Nichtzentralitätsparameter

$$\lambda = \frac{\delta\sqrt{n}}{\sigma}$$

Außerdem ist $\mathrm{t_d^{\alpha}}$ der obere 100α. Perzentilpunkt der t-Verteilung mit d Freiheitsgraden.

Für einseitige Alternativen wird die Trennschärfe angegeben als

$$\pi(n;\delta) = 1 - F_{n-1,\lambda}(t_{n-1}^{\alpha})$$

zum Überprüfen der Nullhypothese im Vergleich zu $\mu > \mu_o$, und sie wird angegeben als

$$\pi(n;\delta) = F_{n-1,\lambda}(-t_{n-1}^{\alpha})$$

zum Überprüfen der Nullhypothese im Vergleich zu $\mu < \mu_o$.

Diese Funktionen werden unter der Annahme abgeleitet, dass die Daten normalverteilt sind und dass das Signifikanzniveau des Tests auf einen bestimmten Wert von α festgelegt ist.

Simulationsstudie B

Diese Simulation dient zum Untersuchen des Effekts einer fehlenden Normalverteilung auf die theoretische Trennschärfefunktion des t-Tests bei einer Stichprobe. Zum Auswerten des Effekts einer fehlenden Normalverteilung wurden die simulierten Trennschärfen mit den Soll-Trennschärfen verglichen, die mit der theoretischen Trennschärfefunktion des Tests berechnet wurden.

Es wurden beidseitige t-Tests bei $\alpha = 0{,}05$ mit Zufallsstichproben verschiedener Umfänge $(\mathrm{n} = 10, 15, 20, 25, 30, 35, 40, 50, 60, 80, 100)$ durchgeführt, die aus den gleichen Grundgesamtheiten generiert wurden, die in der ersten Simulationsstudie beschrieben wurden (siehe Anhang A).

Für jede der Grundgesamtheiten lautet die Nullhypothese des Tests $\mu = \mu_o - \delta$ und die zugehörige Alternativhypothese $\neq \mu_o - \delta$, wobei μ_o auf den tatsächlichen Mittelwert der Grundgesamtheit festgelegt ist und $\delta = \sigma/2$ (σ ist die Standardabweichung der übergeordneten Grundgesamtheit). Damit ist die Differenz zwischen dem tatsächlichen Mittelwert und dem Hypothesenmittelwert 0, so dass die richtige Entscheidung darin besteht, die Nullhypothese zurückzuweisen.

Für jeden angegebenen Stichprobenumfang werden 10.000 Stichprobenreplikationen aus jeder der Verteilungen gezogen. Für jeden angegebenen Stichprobenumfang stellt der Anteil der 10.000 Replikationen, für die die Nullhypothese zurückgewiesen wird, die simulierte Trennschärfe des Tests für den jeweils angegebenen Stichprobenumfang und die Differenz δ dar. Beachten Sie, dass wir diese spezifische Differenz gewählt haben, da damit Trennschärfewerte erzielt werden, die bei kleinen Stichprobenumfänge relativ klein sind.

Darüber hinaus werden die entsprechenden theoretischen Trennschärfewerte (die als Soll-Trennschärfewerte bezeichnet werden) bei der Differenz δ und den unterschiedlichen Stichprobenumfängen für den Vergleich mit den simulierten Trennschärfewerten berechnet.

Die Simulationsergebnisse werden Tabelle 3 und Tabelle 4 aufgeführt und in Abbildung 3 und Abbildung 4 grafisch veranschaulicht.

Ergebnisse und Zusammenfassung

Die Ergebnisse bestätigen, dass die Trennschärfe des t-Tests bei einer Stichprobe im Allgemeinen gegenüber der Annahme der Normalverteilung unempfindlich ist, sofern der Stichprobenumfang ausreichend groß gewählt wurde. Für Stichproben aus symmetrischen Grundgesamtheiten zeigen die Ergebnisse in Tabelle 3, dass die Soll-Trennschärfe und die simulierte Trennschärfe selbst bei kleinen Stichproben nah beieinander liegen. Die entsprechenden Trennschärfekurven in Abbildung 3 sind praktisch nicht voneinander zu unterscheiden. Für Stichproben aus kontaminierten Normalverteilungen sind die Trennschärfewerte für kleine bis mittlere Stichproben etwas konservativ. Dies kann darauf zurückzuführen sein, dass das tatsächliche Signifikanzniveau des Tests für diese Grundgesamtheiten etwas höher als das festgelegte Soll-Signifikanzniveau α ist.

Tabelle 3 Simulierte Trennschärfen bei der Differenz $\delta = \sigma/2$ für einen beidseitigen t-Test bei einer Stichprobe mit Niveau $\alpha = 0{,}05$, wenn die Stichproben aus symmetrischen Grundgesamtheiten generiert wurden. Die simulierten Trennschärfen werden mit den theoretischen Soll-Trennschärfen verglichen, die unter der Annahme der Normalverteilung abgeleitet wurden.

n	Soll-Trennschärfe	N(0;1)	t(5)	t(10)	Lpl	CN (0,9;3)	CN (0,8;3)	B(3;3)	U(0;1)
		Simulierte Trennschärfe bei $\boldsymbol{\delta = \sigma/2}$ (symmetrische Grundgesamtheiten)							
10	0,293	0,299	0,334	0,311	0,357	0,361	0,385	0,280	0,269
15	0,438	0,438	0,480	0,450	0,491	0,512	0,511	0,423	0,421
20	0,565	0,570	0,603	0,578	0,600	0,629	0,623	0,557	0,548
25	0,670	0,674	0,695	0,680	0,691	0,712	0,700	0,665	0,670
30	0,754	0,756	0,770	0,756	0,767	0,768	0,765	0,754	0,750
35	0,820	0,819	0,827	0,815	0,820	0,819	0,812	0,822	0,818
40	0,869	0,870	0,871	0,868	0,862	0,869	0,868	0,875	0,867
50	0,934	0,933	0,929	0,930	0,929	0,923	0,925	0,932	0,940
60	0,968	0,967	0,963	0,965	0,964	0,955	0,955	0,968	0,971
80	0,993	0,993	0,989	0,992	0,991	0,988	0,989	0,994	0,994
100	0,999	0,998	0,996	0,998	0,999	0,998	0,996	0,999	0,999

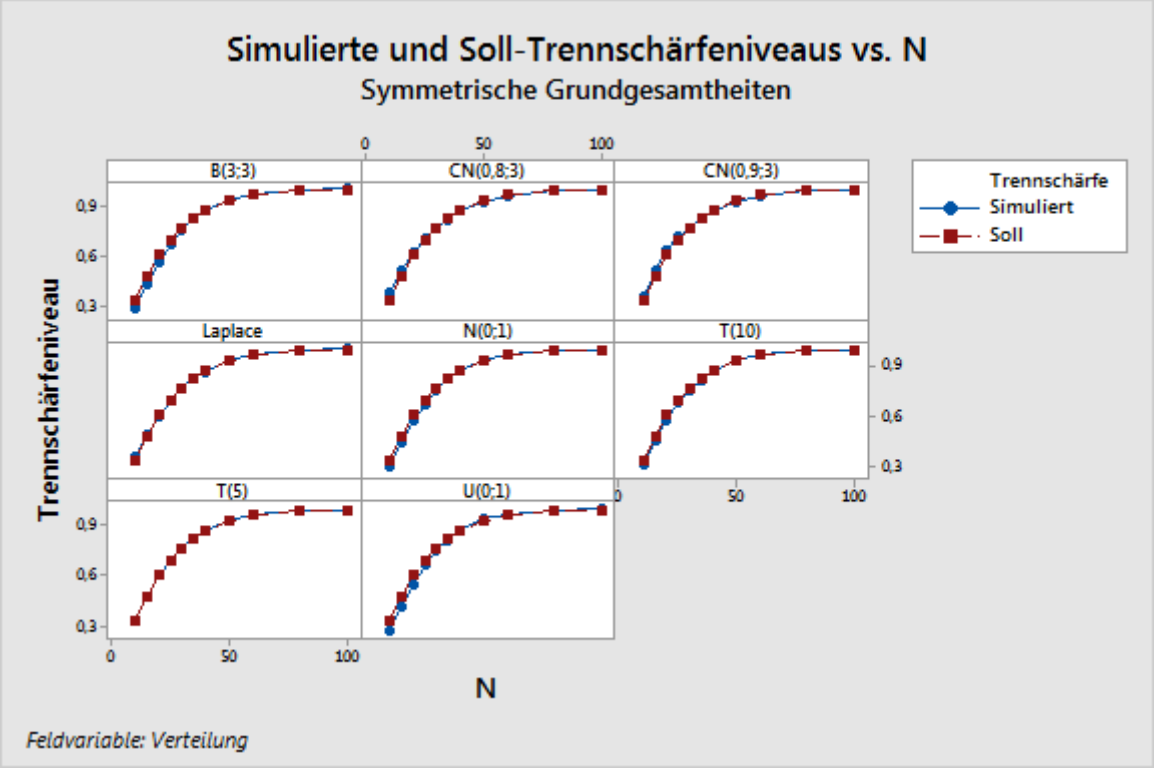

Abbildung 3 Kurven der simulierten Trennschärfen im Vergleich mit den Kurven der theoretischen Soll-Trennschärfen für einen beidseitigen t-Test bei einer Stichprobe bei $\alpha = 0{,}05$, wenn die Stichproben aus symmetrischen Grundgesamtheiten generiert wurden. Die Trennschärfewerte werden bei einer Differenz von $\delta = \sigma/2$ ausgewertet.

Wenn jedoch die Stichproben aus schiefen Verteilungen stammen, weichen die simulierten Trennschärfewerte für kleine Stichproben vom Sollwert ab, wie in Tabelle 4 und Abbildung 4 veranschaulicht. Bei mäßig schiefen Grundgesamtheiten wie der Chi-Quadrat-Verteilung mit 5 Freiheitsgraden und der Chi-Quadrat-Verteilung mit 10 Freiheitsgraden liegen die Soll-Trennschärfe und die simulierte Trennschärfe bei Stichprobenumfängen von mindestens 20 nah beieinander. Für $n = 20$ ist die Soll-Trennschärfe beispielsweise 0,565, wenn die simulierten Trennschärfewerte für die Chi-Quadrat-Verteilung mit 5 Freiheitsgraden und die Chi-Quadrat-Verteilung mit 10 Freiheitsgraden 0,576 bzw. 0,577 sind. Bei stark schiefen Verteilungen sind größere Stichproben erforderlich, damit sich die simulierten Trennschärfen dem Soll-Signifikanzniveau annähern. Dies kann daran liegen, dass beim t-Test bei einer Stichprobe keine ordnungsgemäße Kontrolle für einen Fehler 1. Art erfolgt, wenn die Stichproben klein und die übergeordneten Grundgesamtheiten stark schief sind.

Tabelle 4 Simulierte Trennschärfewerte bei der Differenz $\delta = \sigma/2$ für einen beidseitigen t-Test bei einer Stichprobe mit Niveau $\alpha = 0{,}05$, wenn die Stichproben aus schiefen Grundgesamtheiten generiert wurden. Die simulierten Trennschärfewerte werden mit den Soll-Trennschärfewerten verglichen, die unter der Annahme der Normalverteilung abgeleitet wurden.

N	Soll-Trennschärfe	Exp		Chi(3)	B(8;1)	Chi(5)	Chi(10)
			Schiefe der Grundgesamtheit				
		2,0		1,633	-1,423	1,265	0,894
			Simulierte Trennschärfen				
10	0,293	0,206		0,212	0,390	0,225	0,238
15	0,438	0,416		0,413	0,484	0,409	0,407
20	0,565	0,604		0,591	0,566	0,576	0,577
25	0,670	0,763		0,734	0,657	0,709	0,695
30	0,754	0,859		0,834	0,729	0,808	0,785
35	0,820	0,917		0,895	0,776	0,874	0,835
40	0,869	0,955		0,935	0,823	0,925	0,905
50	0,934	0,987		0,981	0,900	0,973	0,960
60	0,968	0,997		0,994	0,937	0,991	0,985
80	0,993	1,000		0,999	0,980	0,999	0,997
100	0,999	1,000		1,000	0,994	1,000	1,000

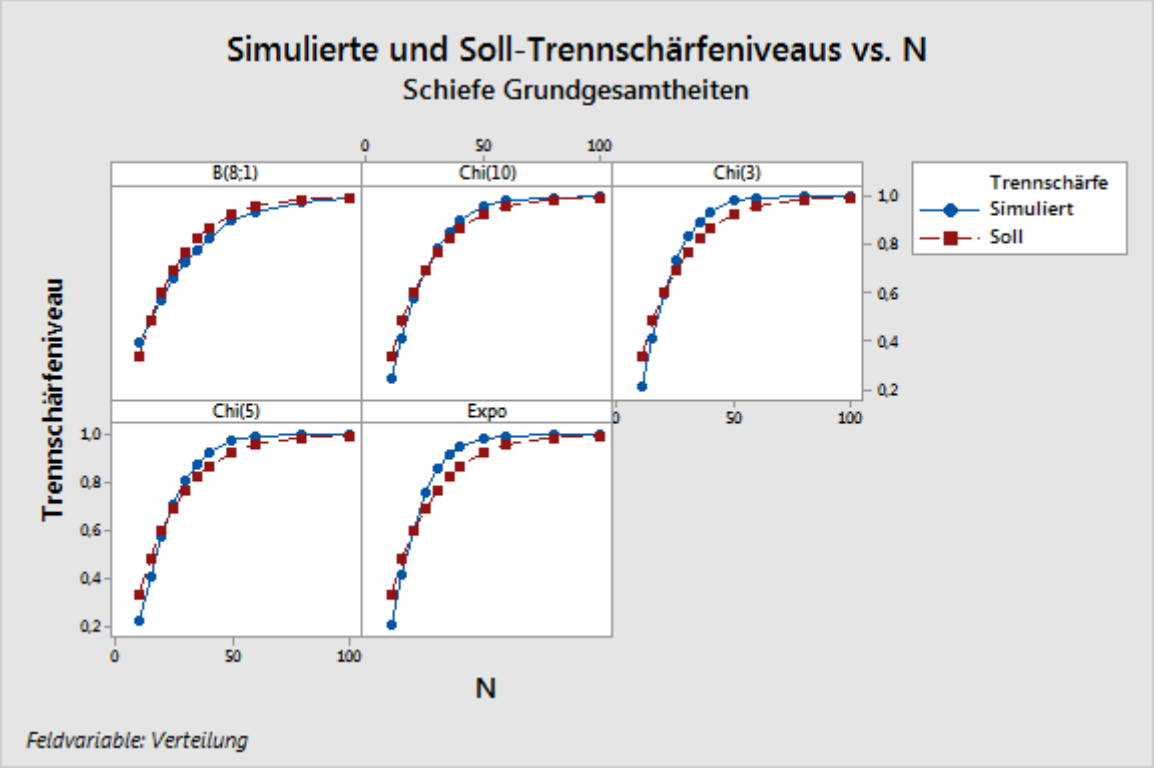

Abbildung 4 Kurven der simulierten Trennschärfen im Vergleich mit den Kurven der theoretischen Soll-Trennschärfen für einen beidseitigen t-Test bei einer Stichprobe bei $\alpha = 0{,}05$, wenn die Stichproben aus symmetrischen Grundgesamtheiten abgeleitet wurden. Die Trennschärfewerte werden bei einer Differenz von $\delta = \sigma/2$ ausgewertet.

Insgesamt ist die Trennschärfefunktion für mäßig schiefe Verteilungen ungeachtet der übergeordneten Grundgesamtheit, aus der die Stichproben gezogen wurden, zuverlässig, sofern der Stichprobenumfang mindestens 20 beträgt. Bei extrem schiefen Grundgesamtheiten ist ein größerer Stichprobenumfang (ca. 40) erforderlich, damit die simulierte Trennschärfe nahe der Soll-Trennschärfe liegt.