t-Test bei zwei Stichproben

Übersicht

Mit einem t-Test bei zwei Stichproben kann festgestellt werden, ob sich zwei unabhängige Gruppen voneinander unterscheiden. Dieser Test wird unter der Annahme abgeleitet, dass die Grundgesamtheiten gleiche Varianzen aufweisen und normalverteilt sind. Die Annahme der Normalverteilung ist nicht kritisch (Pearson, 1931; Barlett, 1935; Geary, 1947), die Annahme der gleichen Varianzen hingegen ist kritisch, wenn sich die Stichprobenumfänge erheblich voneinander unterscheiden (Welch, 1937; Horsnell, 1953).

Einige Fachleute führen zunächst einen Vorabtest durch, um die Gleichheit der Varianzen auszuwerten, ehe sie einen klassischen t-Test bei zwei Stichproben nutzen. Eine derartige Vorgehensweise birgt jedoch schwerwiegende Nachteile, da solche Tests auf Gleichheit der Varianzen wichtigen Annahmen und Einschränkungen unterliegen. Viele Tests auf Gleichheit der Varianzen, z. B. der klassische f-Test, sind empfindlich gegenüber Abweichungen von der Normalverteilung. Andere Tests, bei denen die Annahme der Normalverteilung keine Rolle spielt (z. B. Levene/Brown-Forsythe), sind beim Erkennen einer Differenz zwischen den Varianzen wenig trennscharf.

B. L. Welch hat eine Approximationsmethode entwickelt, mit der die Mittelwerte zweier unabhängiger normalverteilter Grundgesamtheiten verglichen werden können, wenn ihre Varianzen nicht zwangsläufig gleich sind (Welch, 1947). Da der modifizierte t-Test nach Welch nicht unter der Annahme gleicher Varianzen abgeleitet ist, können Benutzer damit die Mittelwerte zweier Grundgesamtheiten vergleichen, ohne dass zuvor ein Test auf Gleichheit der Varianzen ausgeführt werden muss.

Im vorliegenden White Paper wird die modifizierte t-Methode nach Welch mit dem klassischen t-Test bei zwei Stichproben verglichen und ermittelt, welches Verfahren zuverlässiger ist. Darüber hinaus werden die folgenden Datenprüfungen beschrieben, die automatisch ausgeführt und in der Auswertung des Assistenten angezeigt werden; dabei wird erklärt, wie sich diese auf die Analyseergebnisse auswirken:

- Vorliegen einer Normalverteilung
- Ungewöhnliche Daten

- Stichprobenumfang

Methode des t-Tests bei zwei Stichproben

Klassischer t-Test bei zwei Stichproben im Vergleich mit dem t-Test nach Welch

Wenn Daten aus zwei normalverteilten Grundgesamtheiten mit den gleichen Varianzen stammen, ist der klassische t-Test bei zwei Stichproben genau so aussagekräftig oder sogar leistungsfähiger als der t-Test nach Welch. Die Annahme der Normalverteilung ist für das klassische Verfahren nicht kritisch (Pearson, 1931; Barlett, 1935; Geary, 1947), die Annahme der gleichen Varianzen hingegen ist wichtig, um gültige Ergebnisse zu gewährleisten. Das klassische Verfahren ist insbesondere empfindlich gegenüber der Annahme der gleichen Varianzen, wenn sich die Stichprobenumfänge unterscheiden, wobei deren Größe keine Rolle spielt (Welch, 1937; Horsnell, 1953). In der Praxis trifft die Annahme der gleichen Varianzen jedoch selten zu, was höhere Wahrscheinlichkeiten eines Fehlers 1. Art nach sich ziehen kann. Daher gilt, dass der klassische t-Test bei zwei Stichproben bei Stichproben mit unterschiedlichen Varianzen mit größerer Wahrscheinlichkeit falsche Ergebnisse liefert.

Der t-Test nach Welch ist eine praktikable Alternative zum klassischen t-Test, da keine gleichen Varianzen angenommen werden und demzufolge bei allen Stichprobenumfängen keine Empfindlichkeit gegenüber ungleichen Varianzen besteht. Der t-Test nach Welch basiert jedoch auf der Approximation, und seine Leistung in Bezug auf kleine Stichproben ist u. U. fraglich. Wir wollten ermitteln, ob der t-Test nach Welch oder der klassische t-Test bei zwei Stichproben zuverlässiger und praxisrelevanter ist und daher im Assistenten verwendet werden sollte.

Zielstellung

Anhand von Simulationsstudien und theoretischen Ableitungen sollte bestimmt werden, ob der t-Test nach Welch oder der klassische t-Test bei zwei Stichproben zuverlässiger ist. Konkret sollte Folgendes untersucht werden:

- Die Wahrscheinlichkeiten eines Fehlers 1. Art und 2. Art sowohl des klassischen t-Tests bei zwei Stichproben als auch t-Tests nach Welch bei verschiedenen Stichprobenumfängen bei normalverteilten Daten und Gleichheit der Varianzen.
- Die Wahrscheinlichkeiten eines Fehlers 1. Art und 2. Art des t-Tests nach Welch für nicht balancierte Designs mit ungleichen Varianzen, für die der klassische t-Test bei zwei Stichproben fehlschlägt.

Methode

Der Schwerpunkt der Simulationen lag auf drei Bereichen:

- Es wurden simulierte Testergebnisse des klassischen t-Tests bei zwei Stichproben und des t-Tests nach Welch unter diversen Modellannahmen verglichen, u. a. Normalverteilung, fehlende Normalverteilung, Gleichheit der Varianzen, Ungleichheit

der Varianzen, balancierte und nicht balancierte Designs. Weitere Informationen finden Sie in Anhang A.

- Die Trennschärfefunktion für den t-Test nach Welch wurde abgeleitet und mit der Trennschärfefunktion des klassischen t-Tests bei zwei Stichproben verglichen. Weitere Informationen finden Sie in Anhang B.
- Die Auswirkung einer fehlenden Normalverteilung auf die theoretische Trennschärfefunktion des t-Tests nach Welch wurde untersucht.

Ergebnisse

Wenn die Annahmen für das Modell des klassischen t-Tests bei zwei Stichproben zutreffen, zeigt der t-Test nach Welch außer bei kleinen, nicht balancierten Designs die gleiche oder nahezu die gleiche Leistung wie der klassische t-Test bei zwei Stichproben. Die Leistung des klassischen t-Tests bei zwei Stichproben kann jedoch aufgrund seiner Empfindlichkeit gegenüber der Annahme gleicher Varianzen ebenfalls schlecht ausfallen, wenn die Designs klein und nicht balanciert sind. Zudem kann in praktischen Anwendungen nur mit Schwierigkeit festgestellt werden, dass zwei Grundgesamtheiten genau die gleiche Varianz aufweisen. Daher hat die theoretische Überlegenheit des klassischen t-Tests bei zwei Stichproben gegenüber dem t-Test nach Welch nur geringen oder überhaupt keinen praktischen Wert. Daher wird im Assistenten der t-Test nach Welch zum Vergleichen der zwei Grundgesamtheiten verwendet. Die ausführlichen Simulationsergebnisse sind in den Anhängen A, B und C enthalten.

Datenprüfungen

Vorliegen einer Normalverteilung

Der t-Test nach Welch, die im Assistenten verwendete Methode zum Vergleichen der Mittelwerte zweier unabhängiger Grundgesamtheiten, wird unter der Annahme abgeleitet, dass die Grundgesamtheiten einer Normalverteilung folgen. Doch selbst wenn die Daten nicht normalverteilt sind, funktioniert der t-Test nach Welch gut, sofern die Stichproben einen ausreichend großen Umfang aufweisen.

Zielstellung

Wir wollten bestimmen, wie genau die simulierten Signifikanzniveaus für die Welch-Methode und den klassischen t-Test bei zwei Stichproben mit dem Soll-Signifikanzniveau (Wahrscheinlichkeit eines Fehlers 1. Art) von 0,05 übereinstimmen.

Methode

Es wurden Simulationen des t-Tests nach Welch und des klassischen t-Tests bei zwei Stichproben für 10.000 Paare von unabhängigen Stichproben durchgeführt, die aus normalverteilten, schiefen und kontaminierten normalverteilten (mit gleichen und ungleichen Varianzen) Grundgesamtheiten generiert wurden. Die Stichproben wiesen unterschiedliche Stichprobenumfänge auf. Die normalverteilte Grundgesamtheit dient als Kontroll-Grundgesamtheit zu Vergleichszwecken. Für jede Bedingung 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.

Ergebnisse

Bei mittleren oder großen Stichproben bleiben die Wahrscheinlichkeiten eines Fehlers 1. Art des t-Tests nach Welch bei normalverteilten und nicht normalverteilten Daten gleich. Die simulierten Signifikanzniveaus liegen nahe beim Soll-Signifikanzniveau, sofern beide Stichprobenumfänge mindestens 15 betragen. Weitere Informationen finden Sie in Anhang A.

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

Status	Bedingung
	Beide Stichprobenumfänge betragen mindestens 15, daher ist es kein Problem, wenn keine Normalverteilung vorliegt.
	Da mindestens ein Stichprobenumfang < 15, könnte es ein Problem sein, wenn keine Normalverteilung vorliegt.

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 oder das ungewöhnliche Verhalten eines Prozesses hinweisen. Daher ist es häufig unverzichtbar, diese Datenpunkte zu untersuchen und nach Möglichkeit zu korrigieren.

Zielstellung

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

Methode

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

Ergebnisse

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

Ausreißer haben tendenziell nur dann einen Einfluss auf die Trennschärfefunktion, wenn die Stichprobenumfänge sehr klein sind. Wenn Ausreißer vorliegen, sind die beobachteten Trennschärfewerte tendenziell etwas höher als die theoretischen Soll-Trennschärfewerte. Dieses Muster ist in Abbildung 10 in Anhang C ersichtlich, in der die simulierte und die theoretische Trennschärfekurve erst ab einem minimalen Stichprobenumfang von 15 relativ dicht beieinander liegen.

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

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

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 die Stichprobenumfänge ausreichend groß sind, um mit einer hohen Wahrscheinlichkeit Differenzen mit praktischen Konsequenzen zu erkennen.

Zielstellung

Wenn die aktuellen Daten keine ausreichenden Hinweise zum Zurückweisen der Nullhypothese liefern, wollten wir ermitteln können, ob die Stichprobenumfänge groß genug für den Test sind, so dass dieser mit hoher Wahrscheinlichkeit Differenzen mit praktischen Konsequenzen erkennt. Bei der Planung 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. Für den t-Test nach Welch hängt diese Trennschärfefunktion von den Stichprobenumfängen, der Differenz zwischen den Mittelwerten der beiden Grundgesamtheiten und den tatsächlichen Varianzen der beiden Grundgesamtheiten 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 die Stichprobenumfänge, bei denen 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 Stichproben des Benutzers abhängen. In den Anhängen B und C finden Sie jedoch weitere Informationen zur Trennschärfefunktion für den Welch-Test.

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

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

Aspin, A. A. (1949). Tables for Use in Comparisons whose Accuracy Involves Two Variances, Separately Estimated, *Biometrika, 36*, *290-296*.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, *31*, *223-231*.

Box, G. E. P. (1953). Non-normality and Tests on Variances, *Biometrika, 40*, *318-335*.

Geary, R. C. (1947). Testing for Normality, *Biometrika, 34*, *209-242*.

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*.

Horsnell, G. (1953). The effect of unequal group variances on the F test for homogeneity of group means. *Biometrika, 40*, *128-136*.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the populations variances are unknown, *Biometrika, 38*, *324-329*.

Kulinskaya, E. Staudte, R. G. und Gao, H. (2003). Power Approximations in Testing for unequal Means in a One-Way Anova Weighted for Unequal Variances, *Communication in Statistics, 32(12), 2353-2371*.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley.

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. (1931). The Analysis of variance in case of non-normal variation, *Biometrika, 23*, *114-133*.

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*.

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

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal, *Biometrika, 29*, *350-362*.

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

Anhang A: Auswirkungen einer fehlenden Normalverteilung und der Heterogenität auf den klassischen t-Test bei zwei Stichproben und den t-Test nach Welch

Wir haben eine Reihe von Simulationsstudien durchgeführt, bei denen der klassische t-Test bei zwei Stichproben und der t-Test nach Welch unter verschiedenen Modellannahmen verglichen wurden.

Simulationsstudie A

Die Studie wurde in drei Teilen durchgeführt:

- Im ersten Teil der Studie wurde die Empfindlichkeit des klassischen t-Tests bei zwei Stichproben und des t-Tests nach Welch in Bezug auf die Annahme gleicher Varianzen untersucht, wenn die Annahme der Normalverteilung zutrifft. Es wurden zwei Stichproben aus zwei unabhängigen normalverteilten Grundgesamtheiten generiert. Die erste Stichprobe, die Basisstichprobe, wurde aus einer normalverteilten Grundgesamtheit mit dem Mittelwert 0 und der Standardabweichung $\sigma_1 = 2$, $N(0; 2)$ gezogen. Die zweite Stichprobe wurde ebenfalls aus einer normalverteilten Stichprobe mit dem Mittelwert 0 gezogen, als Standardabweichung wurde jedoch σ_2 gewählt, so dass das Verhältnis $\rho = \sigma_2/\sigma_1$ 0,5; 1,0; 1,5 und 2 vorliegt. Mit anderen Worten: Die zweiten Stichproben wurden aus den Grundgesamtheiten $N(0;\ 1)$, $N(0;\ 2)$, $N(0;\ 3)$ und $N(0;\ 4)$ gezogen. Darüber hinaus wurde der Basisstichprobenumfang in jedem Fall auf $n_1 = 5, 10, 15, 20$ festgelegt, und für jedes gegebene n_1 wurde der zweite Stichprobenumfang n_2 derart gewählt, dass das Verhältnis der Stichprobenumfänge $r = n_2/n_1$ etwa gleich 0,5; 1,0; 1,5 und 2,0 war.

  Für jedes dieser zwei Designs mit zwei Stichproben wurden 10.000 Paare von unabhängigen Stichproben aus den jeweiligen Grundgesamtheiten generiert. Anschließend wurde der klassische t-Test bei zwei Stichproben und der t-Test nach Welch für jedes der 10.000 Paare von Stichproben ausgeführt, um die Nullhypothese einer fehlenden Differenz zwischen den Mittelwerten zu testen. Da die tatsächliche Differenz zwischen den Mittelwerten 0 ist, stellt der Anteil der 10.000 Replikationen, für die die Nullhypothese zurückgewiesen wird, das simulierte Signifikanzniveau des Tests dar. Da das Soll-Signifikanzniveau für jeden der Tests $\alpha = 0{,}05$ ist, beträgt der Simulationsfehler der Tests und jedes Experiments ca. 0,2 %.

- Im zweiten Teil wurde die Auswirkung einer fehlenden Normalverteilung, insbesondere der Schiefe, auf die simulierten Signifikanzniveaus der beiden Tests untersucht. Die Einrichtung dieser Simulation entsprach dem der vorherigen

Simulation, die Basisstichprobe wurde jedoch aus der Chi-Quadrat-Verteilung mit zwei Freiheitsgraden Chi(2) gezogen, und die zweiten Stichproben wurden aus anderen Chi-Quadrat-Verteilungen gezogen, so dass $\rho = \sigma_2/\sigma_1$ die Werte 0,5; 1,0; 1,5 und 2 annimmt. Die Hypothesendifferenz zwischen den Mittelwerten wurde auf die tatsächliche Differenz zwischen den Mittelwerten der übergeordneten Grundgesamtheiten festgelegt.

- Im dritten Teil wurde der Effekt von Ausreißern auf die Leistung der zwei t-Tests untersucht. Hierfür wurden die zwei Stichproben aus kontaminierten Normalverteilungen gezogen. Eine kontaminierte Normalverteilung $\mathrm{CN}(\mathrm{p};\sigma)$ ist eine Mischung von zwei Normalverteilungen: Grundgesamtheit $\mathrm{N}(0;1)$ und normalverteilte Grundgesamtheit $\mathrm{N}(0;\sigma)$. Eine kontaminierte Normalverteilung wird wie folgt definiert:

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

Hierbei ist p der Mischparameter und $1 - \mathrm{p}$ der Anteil der Kontamination (bzw. Anteil der Ausreißer). Wenn X die Verteilung $CN(p;\sigma)$ aufweist, kann problemlos gezeigt werden, dass der entsprechende Mittelwert $\mu_X = 0$ und die entsprechende Standardabweichung $\sigma_X = \sqrt{p + (1-p)\sigma^2}$ ist.

Die Basisstichprobe wurde aus $CN(0{,}8;4)$ gezogen, und die zweite Stichprobe wurde aus der kontaminierten Normalverteilung $CN(0{,}8;\sigma)$ gezogen. Der Parameter σ wurde so gewählt, dass das Verhältnis der Standardabweichungen der beiden (kontaminierten) Grundgesamtheiten $\rho = \sigma_2/\sigma_1$ gleich 0,5; 1,0; 1,5 und 2 ist, wie dies auch in Teil I und Teil II der Fall war. Da $\sigma_1 = \sqrt{0{,}8 + (1-0{,}8) * 16} = 2{,}0$, wird entsprechend $\sigma = 1;\ 4;\ 6{,}40;\ 8{,}72$ gewählt. Mit anderen Worten: Die zweiten Stichproben wurden aus $CN(0{,}8;\ 1)$, $CN(0{,}8;4)$, $CN(0{,}8;6{,}4)$ und $CN(0{,}8;8{,}72)$ gezogen. Anschließend wurden die Simulationen wie in Teil I beschrieben ausgeführt.

Die Ergebnisse der Studie sind in Tabelle 1 aufgeführt und in den Abbildungen 1, 2 und 3 veranschaulicht.

Ergebnisse und Zusammenfassung

Die Simulationsergebnisse stützen im Allgemeinen die theoretischen Ergebnisse, dass der klassische t-Test bei zwei Stichproben unter Annahme der Normalverteilung und der Gleichheit der Varianzen selbst bei kleinen Stichproben Signifikanzniveaus nahe dem Sollniveau erzielt. Die zweite Spalte von Diagrammen in Abbildung 1 stellt die simulierten Signifikanzniveaus in Designs dar, bei denen die Varianzen der beiden normalverteilten Grundgesamtheiten gleich sind. Die Kurven der simulierten Signifikanzniveaus für den klassischen t-Test bei zwei Stichproben können von den Kurven der Sollniveaus nicht unterschieden werden.

In den nachfolgenden Tabellen werden die simulierten Signifikanzniveaus der beidseitigen Tests sowohl für den klassischen t-Test bei zwei Stichproben als auch für den t-Test nach Welch aufgeführt, jeweils mit $\alpha = 0{,}05$ und auf der Grundlage von Stichproben, die aus einer normalverteilten Grundgesamtheit, schiefen Grundgesamtheiten (Chi-Quadrat) und kontaminierten normalverteilten Grundgesamtheiten generiert wurden. Die Paare von Stichproben stammen aus derselben Klasse von Verteilungen, die Varianzen der jeweiligen übergeordneten Grundgesamtheiten sind jedoch nicht unbedingt gleich.

Tabelle 1 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch, jeweils mit $\alpha = 0{,}05$) für n = 5.

			Basis-Grundges.: N(0;2) 2. Grundges.: N(0; σ_2)				**Basis-Grundges.: Chi(2) 2. Grundges.: Chi-Quadrat**				**Basis-Grundges.: CN(0,8;4) 2. Grundges.: CN(0,8; σ)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$	**Meth.**	$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	**0,6**	**2T**	0,035	0,050	0,079	0,105	0,058	0,042	0,078	0,113	0,031	0,036	0,035	0,034
		Welch	0,035	0,039	0,049	0,055	0,048	0,029	0,055	0,063	0,029	0,024	0,021	0,020
5	**1,0**	**2T**	0,061	0,052	0,054	0,058	0,086	0,036	0,054	0,064	0,035	0,031	0,025	0,023
		Welch	0,048	0,042	0,044	0,047	0,066	0,021	0,040	0,050	0,027	0,023	0,018	0,016
8	**1,6**	**2T**	0,096	0,048	0,033	0,027	0,133	0,041	0,033	0,032	0,059	0,037	0,029	0,024
		Welch	0,050	0,045	0,043	0,042	0,094	0,034	0,032	0,041	0,034	0,029	0,026	0,022
10	**2,0**	**2T**	0,118	0,055	0,034	0,025	0,139	0,041	0,028	0,024	0,073	0,041	0,028	0,023
		Welch	0,052	0,051	0,050	0,051	0,097	0,041	0,033	0,042	0,035	0,032	0,028	0,025

Tabelle 2 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch, jeweils mit $\alpha = 0{,}05$) für n = 10

			Basis-Grundges.: N(0;2) 2. Grundges.: N(0; σ_2)				**Basis-Grundges.: Chi(2) 2. Grundges.: Chi-Quadrat**				**Basis-Grundges.: CN(0,8;4) 2. Grundges.: CN(0,8; σ)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$	**Meth.**	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	**0,5**	**2T**	0,020	0,050	0,081	0,112	0,039	0,044	0,091	0,123	0,021	0,035	0,045	0,047
		Welch	0,046	0,048	0,050	0,050	0,043	0,047	0,067	0,063	0,034	0,028	0,022	0,019
10	**1,0**	**2T**	0,057	0,051	0,053	0,055	0,068	0,044	0,053	0,054	0,043	0,042	0,037	0,032
		Welch	0,051	0,049	0,049	0,049	0,062	0,037	0,046	0,049	0,039	0,038	0,032	0,027
15	**1,5**	**2T**	0,088	0,048	0,034	0,029	0,100	0,043	0,032	0,032	0,064	0,040	0,028	0,021
		Welch	0,050	0,048	0,047	0,048	0,074	0,044	0,041	0,046	0,035	0,037	0,035	0,031
20	**2**	**2T**	0,110	0,048	0,026	0,019	0,133	0,042	0,026	0,022	0,093	0,046	0,029	0,019

			Basis-Grundges.: N(0;2) **2. Grundges.: N(0; σ_2)**				**Basis-Grundges.: Chi(2)** **2. Grundges.: Chi-Quadrat**				**Basis-Grundges.: CN(0,8;4)** **2. Grundges.: CN(0,8; σ)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$	**Meth.**	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
		Welch	0,048	0,047	0,045	0,046	0,083	0,050	0,044	0,049	0,036	0,039	0,040	0,038

Tabelle 3 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch, jeweils mit $\alpha = 0{,}05$) für n = 15

			Basis-Grundges.: N(0;2) **2. Grundges.: N(0; σ_2)**				**Basis-Grundges.: Chi(2)** **2. Grundges.: Chi-Quadrat**				**Basis-Grundges.: CN(0,8;4)** **2. Grundges.: CN(0,8; σ)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$	**Meth.**	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	**0,53**	**2T**	0,021	0,050	0,083	0,110	0,036	0,041	0,089	0,114	0,022	0,044	0,056	0,062
		Welch	0,050	0,051	0,051	0,050	0,047	0,049	0,067	0,062	0,044	0,036	0,027	0,022
15	**1,0**	**2T**	0,049	0,047	0,050	0,053	0,064	0,046	0,051	0,061	0,045	0,045	0,041	0,037
		Welch	0,045	0,046	0,049	0,048	0,060	0,042	0,048	0,057	0,042	0,043	0,039	0,033
23	**1,53**	**2T**	0,081	0,049	0,033	0,028	0,103	0,042	0,036	0,030	0,075	0,048	0,033	0,024
		Welch	0,048	0,049	0,048	0,050	0,071	0,042	0,048	0,050	0,042	0,045	0,044	0,041
30	**2,0**	**2T**	0,111	0,050	0,028	0,018	0,123	0,049	0,027	0,020	0,100	0,046	0,025	0,016
		Welch	0,049	0,051	0,051	0,053	0,074	0,056	0,045	0,047	0,039	0,044	0,042	0,040

Tabelle 4 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch, jeweils mit $\alpha = 0{,}05$) für n = 20

			Basis-Grundges.: N(0;2) **2. Grundges.: N(0; σ_2)**				**Basis-Grundges.: Chi(2)** **2. Grundges.: Chi-Quadrat**				**Basis-Grundges.: CN(0,8;4)** **2. Grundges.: CN(0,8; σ)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$	**Meth.**	$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	**0,5**	**2T**	0,019	0,052	0,087	0,115	0,028	0,048	0,087	0,119	0,021	0,048	0,067	0,079
		Welch	0,050	0,054	0,053	0,053	0,044	0,054	0,061	0,061	0,048	0,042	0,035	0,028
20	**1,0**	**2T**	0,048	0,049	0,052	0,053	0,057	0,046	0,052	0,056	0,049	0,044	0,042	0,040
		Welch	0,045	0,049	0,051	0,050	0,055	0,044	0,050	0,052	0,047	0,042	0,040	0,037
30	**1,5**	**2T**	0,086	0,054	0,039	0,032	0,098	0,047	0,035	0,033	0,075	0,047	0,033	0,022
		Welch	0,054	0,054	0,053	0,052	0,068	0,047	0,051	0,053	0,041	0,043	0,044	0,042
40	**2,0**	**2T**	0,107	0,049	0,026	0,016	0,123	0,046	0,027	0,019	0,107	0,047	0,026	0,016
		Welch	0,048	0,049	0,046	0,047	0,070	0,054	0,046	0,045	0,044	0,043	0,043	0,042

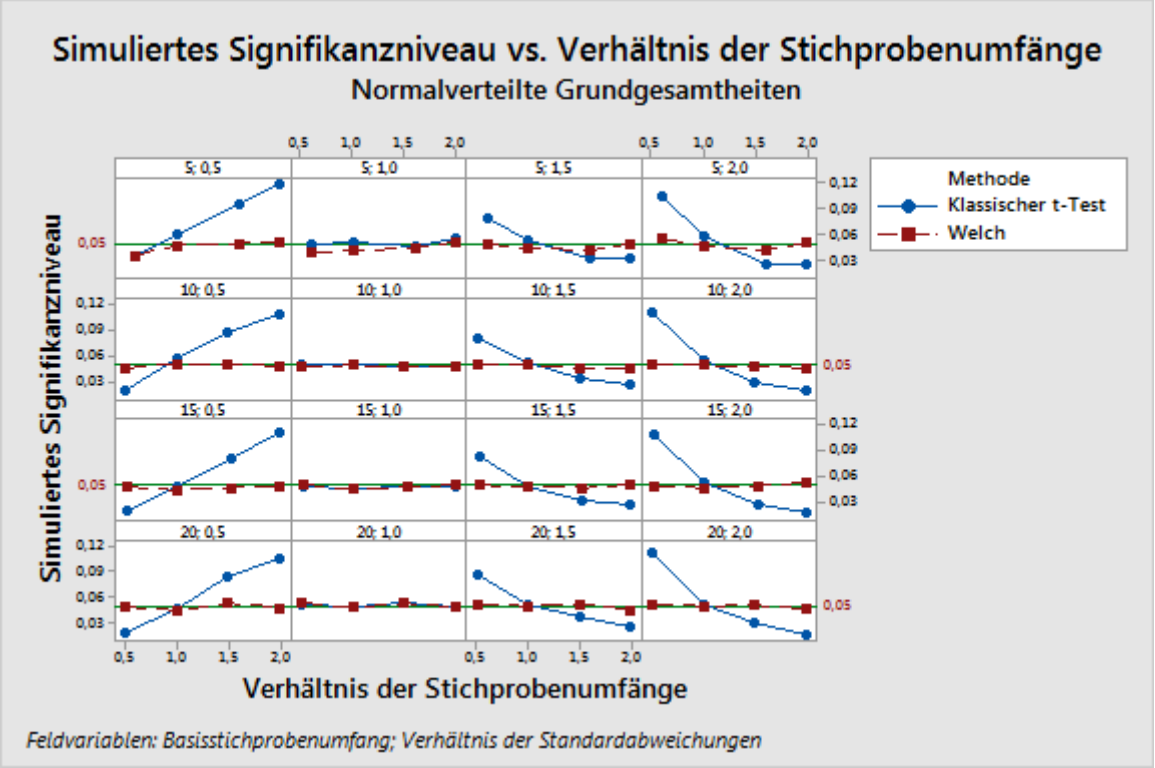

Abbildung 1 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch, jeweils mit $\alpha = 0{,}05$) basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

Die Simulationsergebnisse zeigen, dass der klassische t-Test bei zwei Stichproben für relativ kleine Stichproben robust in Bezug auf eine fehlende Normalverteilung, jedoch empfindlich gegenüber der Annahme gleicher Varianzen ist, es sei denn, das Design mit zwei Stichproben ist nahezu balanciert. Dies wird in den Abbildungen 1, 2 und 3 grafisch veranschaulicht. Die Kurven der simulierten Signifikanzniveaus für den klassischen t-Test bei zwei Stichproben schneiden die Linie des Sollniveaus an dem Punkt, an dem das Verhältnis der Stichprobenumfänge 1,0 beträgt, selbst wenn sich die Varianzen sehr unterscheiden. Für alle drei Klassen von Verteilungen (Normalverteilung, Chi-Quadrat-Verteilung und kontaminierte normalverteilte Grundgesamtheiten) gilt Folgendes: Bei unterschiedlichen Stichprobenumfängen liegen die simulierten Signifikanzniveaus des klassischen t-Tests bei zwei Stichproben nur dann nahe dem Sollniveau, wenn die Varianzen gleich sind. Dies wird in der zweiten Spalte von Diagrammen in den Abbildungen 1, 2 und 3 veranschaulicht.

Die Leistung des klassischen t-Tests ist nicht wünschenswert, wenn das Design nicht balanciert ist und die Varianzen ungleich sind. Selbst geringfügige Ungleichheiten zwischen den Varianzen sind problematisch. Für derartige nicht balancierte Designs mit ungleichen Varianzen bewirkt eine Normalverteilung der Daten keine Verbesserung der simulierten Signifikanzniveaus. Tatsächlich entfernen sich die simulierten Signifikanzniveaus mit zunehmendem Stichprobenumfang vom Sollniveau, ungeachtet der übergeordneten Grundgesamtheit. Wenn die größere Stichprobe aus der Grundgesamtheit mit der größeren Varianz gezogen wird, sind die simulierten Signifikanzniveaus kleiner als das Sollniveau.

Wenn die größere Stichprobe aus der Grundgesamtheit mit der kleineren Varianz gezogen wird, sind die simulierten Niveaus größer als die Sollniveaus. Arnold (1990, Seite 372) zog einen ähnlichen Schluss bei der Untersuchung der asymptotischen Verteilung des klassischen t-Tests bei zwei Stichproben unter Annahme der Ungleichheit der Varianzen.

Der t-Test bei zwei Stichproben nach Welch hingegen ist unempfindlich gegenüber Abweichungen von der Annahme gleicher Varianzen, wie in den Abbildungen 1, 2 und 3 veranschaulicht. Dies ist nicht überraschend, da der t-Test nach Welch nicht unter der Annahme gleicher Varianzen abgeleitet wird. Die Annahme der Normalverteilung, anhand derer der t-Test nach Welch abgeleitet ist, scheint nur dann wichtig zu sein, wenn das Minimum der beiden Stichprobenumfänge sehr klein ist. Bei größeren Stichproben jedoch wird der Test immun gegenüber Abweichungen von der Annahme der Normalverteilung. Dies wird in den Abbildungen 2 und 3 veranschaulicht, in denen die simulierten Signifikanzniveaus durchgehend nahe dem Sollniveau bleiben, wenn der minimale Umfang der beiden Stichproben 15 beträgt. Wenn beide Stichproben aus der Chi-Quadrat-Verteilung mit zwei Freiheitsgraden generiert werden und ihr Stichprobenumfang jeweils 15 beträgt, ist das simulierte Signifikanzniveau 0,042 (siehe Tabelle 3).

Ausreißer scheinen sich ebenfalls nicht auf die Leistung des t-Tests nach Welch auszuwirken, sofern der minimale Umfang der zwei Stichproben ausreichend groß gewählt ist. In Tabelle 3 und Abbildung 3 wird gezeigt, dass die simulierten Signifikanzniveaus ab einem minimalen Umfang der beiden Stichproben von 15 nahe dem Sollniveau liegen (die simulierten Signifikanzniveaus sind 0,045; 0,045; 0,041 und 0,037, wenn das Verhältnis der Standardabweichungen 0,5; 1,0; 1,5 bzw. 2,0 beträgt).

Diese Ergebnisse zeigen, dass der t-Test bei zwei Stichproben nach Welch für die meisten praktischen Anwendungen hinsichtlich der simulierten Signifikanzniveaus oder der Wahrscheinlichkeiten eines Fehlers 1. Art eine bessere Leistung als der klassische t-Test bei zwei Stichproben aufweist.

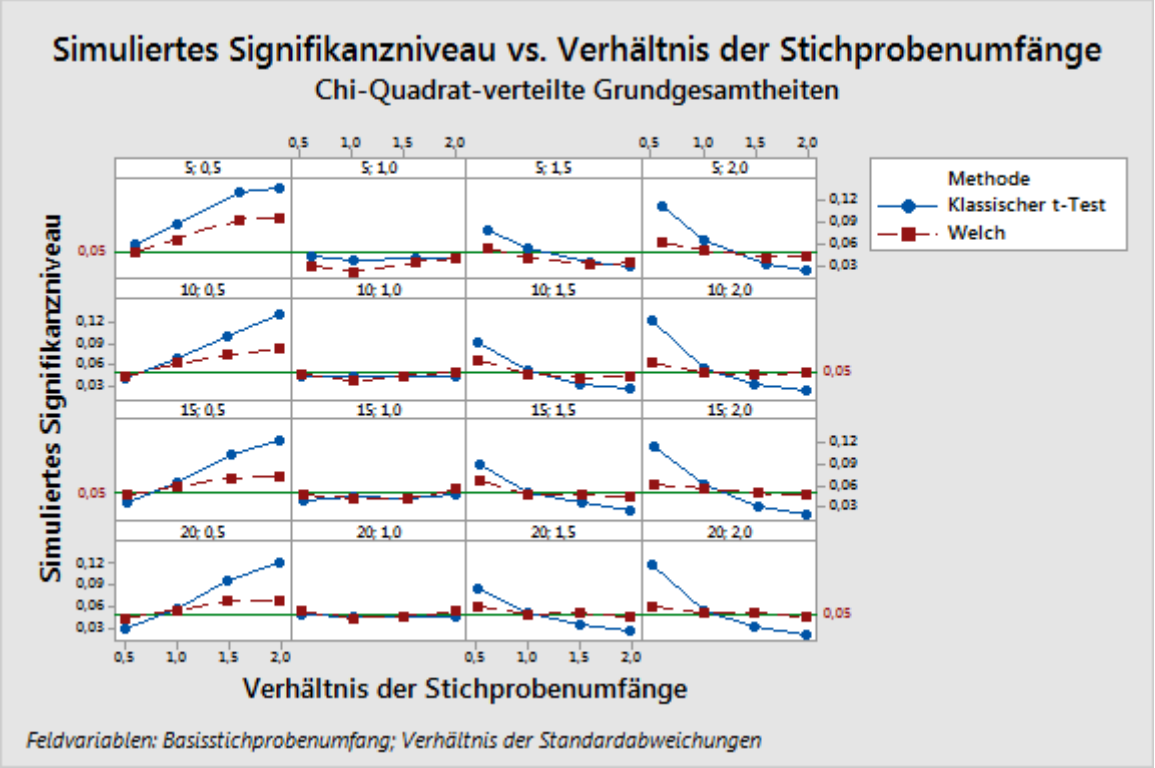

Abbildung 2 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch), basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

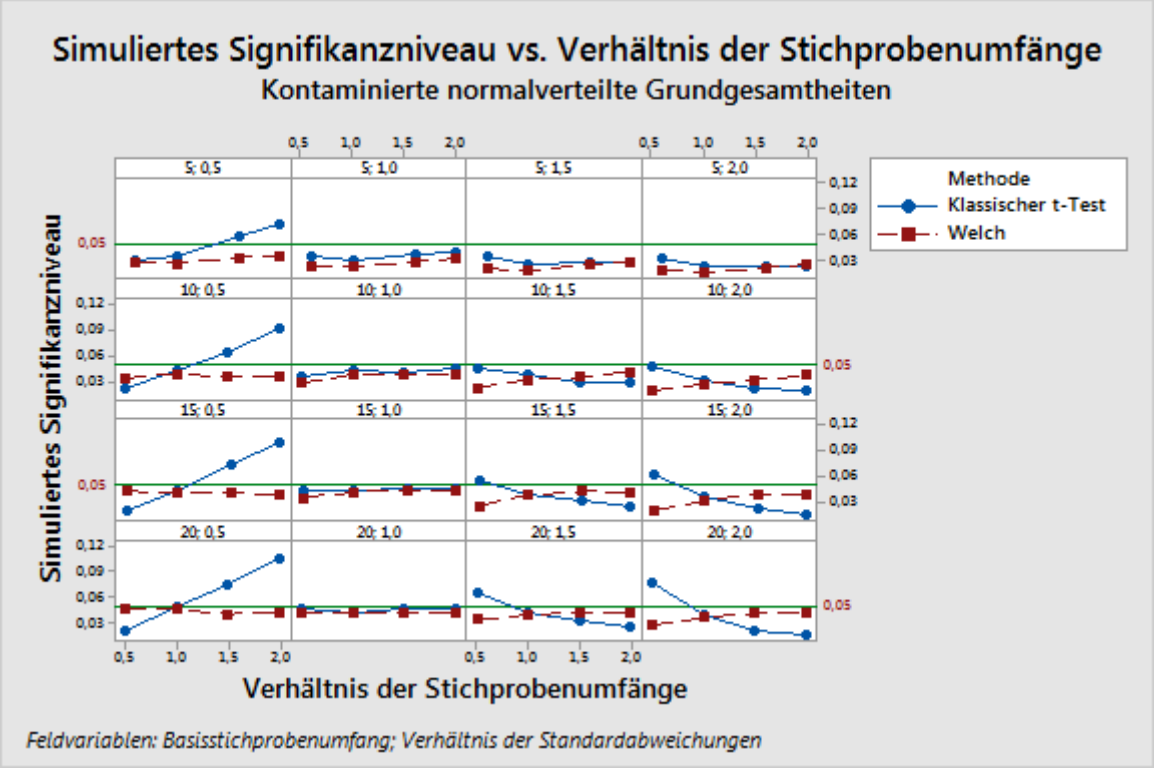

Abbildung 3 Simulierte Signifikanzniveaus der beidseitigen Tests (klassischer t-Test bei zwei Stichproben und t-Test nach Welch), basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

Anhang B: Vergleich der Trennschärfefunktionen der beiden Tests

Wir wollten die Bedingungen bestimmen, unter denen die Trennschärfefunktion für den t-Test nach Welch mit der Trennschärfefunktion des klassischen t-Tests bei zwei Stichproben übereinstimmt bzw. nahezu übereinstimmt.

Die Trennschärfefunktionen der t-Tests (bei einer Stichprobe oder zwei Stichproben) sind im Allgemeinen hinreichend bekannt und werden in einer Vielzahl von Publikationen (Pearson und Hartley, 1952; Neyman et al., 1935; Srivastava, 1958) erörtert. Das folgende Theorem gibt die Trennschärfefunktion für jede der drei verschiedenen Alternativhypothesen in Designs mit zwei Stichproben an.

THEOREM B1

Unter den Annahmen der Normalverteilung und der Gleichheit der Varianzen kann die Trennschärfefunktion eines beidseitigen t-Tests bei zwei Stichproben mit dem nominalen Niveau α als Funktion der Stichprobenumfänge und der Differenz$\delta = \mu_1 - \mu_2$ ausgedrückt werden als

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha/2}\right) + F_{d_C,\lambda}\left(-t_{d_C}^{\alpha/2}\right)$$

Hierbei ist $F_{d_C,\lambda}(.)$ die kumulative Verteilungsfunktion der nicht zentralen t-Verteilung mit $d_C = n_1 + n_2 - 2$ Freiheitsgraden und dem Nichtzentralitätsparameter

$$\lambda = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Zudem wird die Trennschärfefunktion für die Alternativhypothese $\mu_1 > \mu_2$ angegeben als

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha}\right)$$

Beim Testen gegen die Alternative $\mu_1 < \mu_2$ wird die Trennschärfe hingegen ausgedrückt als

$$\pi(n_1, n_2, \delta) = F_{d_C,\lambda}\left(-t_{d_C}^{\alpha}\right)$$

Das Ergebnis im oben aufgeführten Theorem ist zwar ausreichend dokumentiert, die Trennschärfefunktion des Tests auf Grundlage des modifizierten t-Tests nach Welch wurde jedoch bisher nicht gesondert in der Fachliteratur diskutiert. Eine Approximation kann von der approximierten Trennschärfefunktion für das einfache ANOVA-Modell abgeleitet werden (siehe Kulinskaya et. al, 2003). Leider gilt diese Trennschärfefunktion lediglich für beidseitige Alternativen. Das Design mit zwei Stichproben ist jedoch so ein spezieller Fall, dass ein anderer Ansatz verfolgt werden kann, um die (genaue) Trennschärfefunktion des t-Tests nach Welch für jede der drei Alternativen zu bestimmen. Diese Funktionen werden im folgenden Theorem angegeben.

THEOREM B2

Unter der Annahme, dass die Grundgesamtheiten normalverteilt sind (jedoch nicht unbedingt mit der gleichen Varianz), kann die Trennschärfefunktion eines beidseitigen t-Tests nach Welch mit einem nominalen Niveau α als Funktion der Stichprobenumfänge und der Differenz $\delta = \mu_1 - \mu_2$ ausgedrückt werden als

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

Hierbei ist $G_{d,\lambda}(.)$ die kumulative Verteilungsfunktion der nicht zentralen t-Verteilung mit d_W Freiheitsgraden, angegeben als

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

und dem Nichtzentralitätsparameter

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Für die einseitigen Alternativen werden die Trennschärfefunktionen angegeben als

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha}\right)$$

und

$$\pi_W(n_1, n_2, \delta) = G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha}\right)$$

zum Testen der Nullhypothese gegen die Alternative $\mu_1 > \mu_2$ bzw. zum Testen der Nullhypothese gegen die Alternative $\mu_1 < \mu_2$.

Der Beweis des Ergebnisses wird in Anhang D aufgeführt.

Bedenken Sie vor dem Vergleich der beiden Trennschärfefunktionen Folgendes: Wegen der Ableitung des klassischen t-Tests bei zwei Stichproben unter der zusätzlichen Annahme der Gleichheit der Varianzen der Grundgesamtheiten müssen die theoretischen Trennschärfefunktionen der beiden Tests verglichen werden, wenn diese zweite Annahme für den t-Test nach Welch ebenfalls gültig ist.

Theoretisch ist bekannt, dass unter den Annahmen der Normalverteilung und der Gleichheit der Varianzen Folgendes gilt:

$\pi(n_1, n_2, \delta) \geq \pi_W(n_1, n_2, \delta)$ für alle n_1, n_2, δ

Im nächsten Ergebnis werden Bedingungen ausgewiesen, unter denen die beiden Funktionen (annähernd) gleich sind.

THEOREM B3

Unter den Annahmen der Normalverteilung der Gleichheit der Varianzen kann Folgendes festgestellt werden:

1. Wenn $n_1 \sim n_2$, dann ist $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ für jede Differenz δ. Insbesondere wenn $n_1 = n_2$, dann ist $\pi(n_1, n_2, \delta) = \pi_W(n_1, n_2, \delta)$ für jede Differenz δ, so dass der t-

Test nach Welch die gleiche Trennschärfe wie der klassische t-Test bei zwei Stichproben aufweist.

2. Wenn n_1 und n_2 klein sind und $n_1 \neq n_2$, dann weist der t-Test nach Welch eine geringere Trennschärfe als der klassische t-Test bei zwei Stichproben auf. Wenn jedoch n_1 und n_2 groß sind, dann ist $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ (ungeachtet der Differenz zwischen den Stichprobenumfängen).

Der Beweis des Ergebnisses wird in Anhang E aufgeführt.

Unter der Annahme der Gleichheit der Varianzen sind die Nichtzentralitätsparameter der Trennschärfefunktionen der beiden Tests identisch. Die Differenz zwischen den Trennschärfefunktionen kann lediglich auf die Differenz zwischen ihren jeweiligen Freiheitsgraden zurückgeführt werden. Aus der Theorie ist bekannt, dass der klassische t-Test unter den besagten Annahmen ein gleichmäßig trennschärfster Test (uniformly most powerful, UMP) und daher durch höhere Freiheitsgrade gekennzeichnet ist. Die oben aufgeführten Ergebnisse haben jedoch folgende Kernaussage: Wenn das Design balanciert oder nahezu balanciert ist, sind auch die Trennschärfefunktionen identisch oder nahezu identisch. Der klassische t-Test weist nur in einem Fall eine erheblich größere Trennschärfe als der t-Test nach Welch auf, nämlich wenn das Design stark unbalanciert ist und die Stichproben klein sind. Leider ist dies auch genau die Situation, in der der klassische t-Test bei zwei Stichproben besonders empfindlich gegenüber der Annahme gleicher Varianzen ist, wie in Anhang A veranschaulicht. Daher ist die Trennschärfefunktion des t-Tests nach Welch für praktische Zwecke als zuverlässiger zu erachten.

Die Ergebnisse von Theorem B3 werden anhand des folgenden Beispiels veranschaulicht, in dem die beiden normalverteilten Grundgesamtheiten die gleiche Standardabweichung 3 aufweisen. Trennschärfewerte auf der Grundlage der (beidseitigen) Trennschärfefunktionen von Theorem B1 und Theorem B2 werden gemäß den folgenden vier Szenarios berechnet:

1. Beide Stichproben sind klein, weisen jedoch den gleichen Umfang auf ($n_1 = n_2 = 10$).
2. Beide Stichproben sind klein, eine Stichprobe ist jedoch zwei Mal größer als die andere ($n_1 = 10, n_2 = 20$).
3. Eine Stichprobe ist klein, und die andere Stichprobe weist einen mittleren Umfang auf; die mittlere Stichprobe ist jedoch vier Mal größer als die kleinere Stichprobe ($n_1 = 10, n_2 = 40$).
4. Eine Stichprobe weist einen mittleren Umfang auf, während die andere groß ist; die größere Stichprobe ist jedoch vier Mal so groß wie die mittlere Stichprobe ($n_1 = 50, n_2 = 200$).

Unter der Annahme, dass für beide Tests $\alpha = 0{,}05$, werden die Trennschärfefunktionen in jedem Szenario bei der Differenz $\delta = 0{,}0; 0{,}5; 1{,}0; 1{,}5; 2{,}0; \ldots 5{,}0$ ausgewertet. Die Ergebnisse werden in Tabelle 5 aufgeführt, und die Funktionen werden in Abbildung 4 dargestellt.

Tabelle 5 Vergleich der theoretischen Trennschärfefunktionen der beidseitigen klassischen t-Tests bei zwei Stichproben und der beidseitigen t-Tests nach Welch bei $\alpha = 0{,}05$. Die Stichprobenumfänge n_1 und n_2 sind festgelegt, und die Trennschärfefunktionen werden bei den Differenzen δ im Bereich von 0,0 bis 5,0 ausgewertet.

δ	**0,0**	**0,5**	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**
$n_1 = n_2 = 10$											
$\pi(n_1, n_2, \delta)$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$\pi_W(n_1, n_2, \delta)$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$n_1 = 10, n_2 = 20$											
$\pi(n_1, n_2, \delta)$	0,05	0,070	0,132	0,239	0,383	0,547	0,703	0,828	0,913	0,962	0,986
$\pi_W(n_1, n_2, \delta)$	0,05	0,070	0,129	0,231	0,371	0,531	0,686	0,813	0,902	0,955	0,982
$n_1 = 10, n_2 = 40$											
$\pi(n_1, n_2, \delta)$	0,05	0,075	0,152	0,283	0,455	0,637	0,791	0,899	0,959	0,986	0,996
$\pi_W(n_1, n_2, \delta)$	0,05	0,072	0,142	0,261	0,419	0,592	0,748	0,865	0,938	0,976	0,992
$n_1 = 50, n_2 = 200$											
$\pi(n_1, n_2, \delta)$	0,05	0,182	0,556	0,883	0,987	0,999	1,0	1,0	1,0	1,0	1,0
$\pi_W(n_1, n_2, \delta)$	0,05	0,180	0,548	0,877	0,986	0,999	1,0	1,0	1,0	1,0	1,0

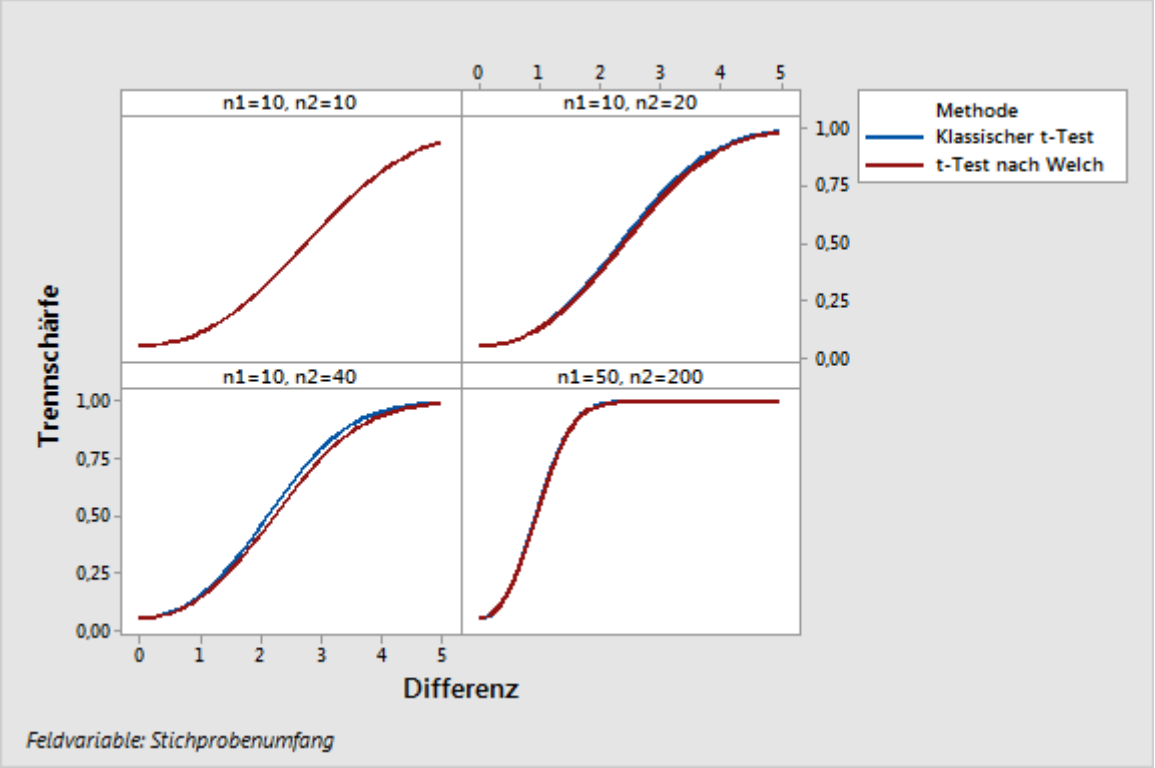

Abbildung 4 Diagramme der theoretischen Trennschärfefunktionen der beidseitigen klassischen t-Tests bei zwei Stichproben und der beidseitigen t-Tests nach Welch im Vergleich zu δ, der zu erkennenden Differenz zwischen den Mittelwerten. Beide Tests verwenden $\alpha = 0{,}05$. Die angenommenen Grundgesamtheiten sind normalverteilt mit der gleichen Standardabweichung 3.

Simulationsstudie B

Der Zweck dieser Simulationsstudie besteht darin, die Trennschärfen des klassischen t-Tests bei zwei Stichproben mit den Trennschärfen des t-Tests bei zwei Stichproben nach Welch in balancierten Designs zu vergleichen, wobei eine Ungleichheit der Varianzen angenommen wird. Die Experimente in diesen Studien ähneln denen, die in Anhang A erläutert werden.

In der ersten Gruppe von Experimenten wurden Paare von Stichproben mit gleichem Umfang aus den normalverteilten Grundgesamtheiten mit ungleichen Varianzen generiert. Die Basis-Grundgesamtheit wurde auf $N(0; 2)$ festgelegt, während die zweiten normalverteilten Grundgesamtheiten so gewählt wurden, dass das Verhältnis der Standardabweichungen $\rho = \sigma_2/\sigma_1$ gleich 0,5; 1,5 und 2 war. Analog dazu wurden in einer zweiten Gruppe die zwei Stichproben aus Chi-Quadrat-Verteilungen mit ungleichen Varianzen gezogen (Basis-Grundgesamtheit ist Chi(2)). In der letzten Gruppe von Experimenten wurden die Paare von Stichproben aus der kontaminierten Normalverteilung gezogen (Basis-Grundgesamtheit CN(0,8;4)), wie bereits in Anhang A definiert.

Für jede Gruppe von Experimenten wurden die simulierten Trennschärfen (bei einer angegebenen erkennbaren Differenz δ) der einzelnen Tests für die Stichprobenumfänge $n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$ berechnet. In jedem Experiment wurde die simulierte Trennschärfe als Anteil der Instanzen berechnet, bei denen die Nullhypothese

zurückgewiesen wurde, wenn sie nicht zutreffend war. Für alle Experimente wurde die Differenz zwischen den Mittelwerten in einer Einheit des Standards in der Basis-Grundgesamtheit (der ersten der zwei Stichproben) angegeben. Konkret: Wir haben $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ fixiert, da der Wert für alle drei Klassen von Verteilungen in dieser Studie relativ klein ist. Die Simulationsergebnisse werden in Tabelle 2.2 aufgeführt und in Abbildung 2.2a, Abbildung 2.2b und Abbildung 2.2c grafisch dargestellt.

Ergebnisse und Zusammenfassung

Die Ergebnisse in Tabelle 6 und Abbildung 4 zeigen, dass die theoretischen Trennschärfefunktionen unter der Annahme gleicher Varianzen in balancierten Designs identisch sind, wie in Theorem 2.3 angegeben. Wenn die Stichprobenumfänge zudem relativ klein und nahezu gleich sind, liefern die zwei Funktionen Trennschärfewerte, die annähernd gleich sind. Nur wenn die Stichproben relativ klein sind und eine Stichprobe etwa vier Mal größer als die andere Stichprobe ist, zeichnen sich erkennbare Differenzen zwischen den Trennschärfefunktionen ab (z. B. bei $n_1 = 10, n_2 = 40$). Selbst in diesem Fall sind die theoretischen Trennschärfewerte aus dem klassischen t-Test bei zwei Stichproben nur geringfügig höher als die Trennschärfewerte aus dem t-Test nach Welch. Wenn die Designs stark unbalanciert, die Stichproben hingegen (relativ) groß sind, sind die zwei Trennschärfefunktionen im Wesentlichen identisch, wie in Theorem B3 behauptet.

In balancierten Designs mit ungleichen Varianzen liefern die beiden Tests zudem Trennschärfewerte, die praktisch identisch sind. Bei sehr kleinen Stichproben ($n < 10$) ist für den klassischen t-Test bei zwei Stichproben jedoch eine etwas bessere Leistung zu verzeichnen.

Tabelle 6 Vergleich der simulierten Trennschärfen des klassischen t-Tests bei zwei Stichproben und des t-Tests nach Welch in balancierten Designs mit ungleichen Varianzen

		Basis-Grundgesamtheit: N(0;2)			**Basis-Grundgesamtheit: Chi(2)**			**Basis-Grundgesamtheit: CN(0,8;4)**		
n	$\frac{\sigma_2}{\sigma_1}$	0,5	1,5	2,0	0,5	1,5	2,0	0,5	1,5	2,0
5	2T	0,431	0,196	0,152	0,555	0,281	0,215	0,579	0,373	0,335
	Welch	0,366	0,166	0,119	0,424	0,250	0,184	0,521	0,320	0,283
10	2T	0,770	0,385	0,270	0,846	0,438	0,324	0,790	0,510	0,435
	Welch	0,747	0,372	0,253	0,832	0,427	0,308	0,776	0,493	0,417
15	2T	0,916	0,539	0,387	0,948	0,565	0,424	0,898	0,615	0,508
	Welch	0,908	0,532	0,375	0,945	0,557	0,413	0,891	0,605	0,497
20	2T	0,971	0,682	0,497	0,982	0,680	0,521	0,952	0,702	0,573
	Welch	0,969	0,677	0,487	0,981	0,676	0,511	0,947	0,697	0,563
25	2T	0,990	0,779	0,591	0,994	0,765	0,605	0,980	0,783	0,641

		Basis-Grundgesamtheit: N(0;2)			Basis-Grundgesamtheit: Chi(2)			Basis-Grundgesamtheit: CN(0,8;4)		
	Welch	0,990	0,777	0,582	0,994	0,762	0,597	0,979	0,778	0,636
30	2T	0,998	0,851	0,675	0,998	0,826	0,676	0,994	0,839	0,699
	Welch	0,998	0,849	0,670	0,998	0,824	0,668	0,994	0,836	0,694

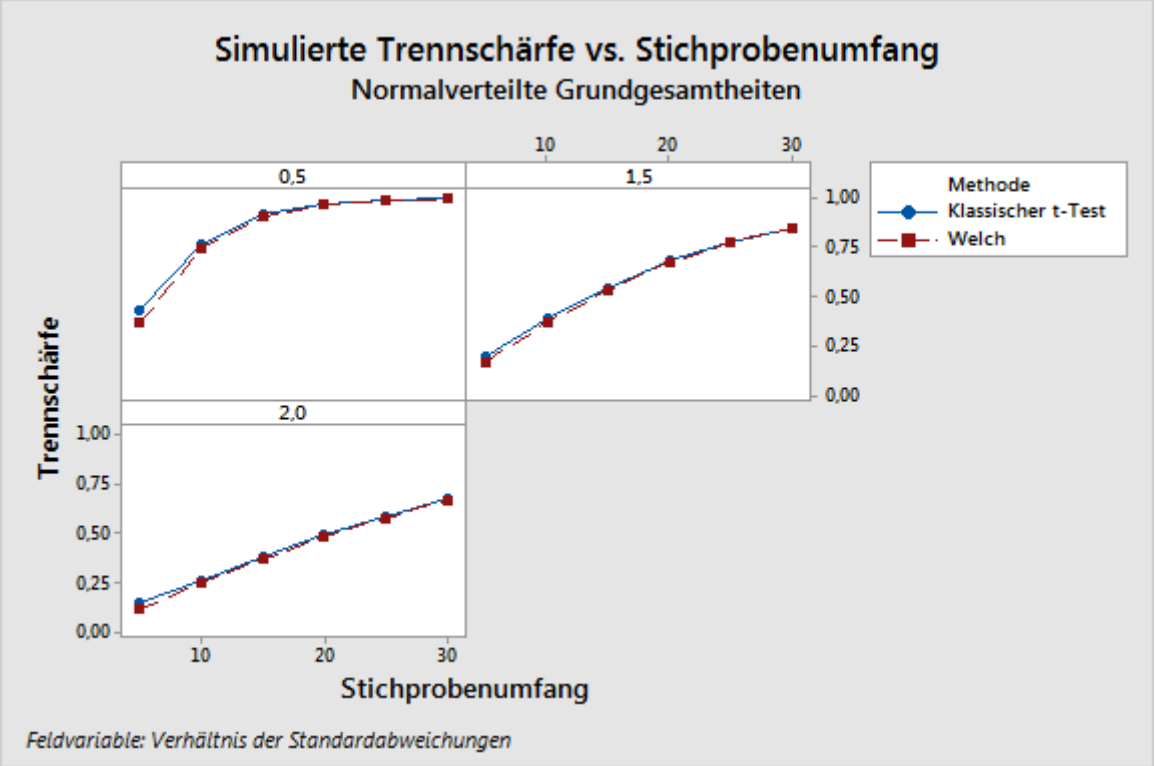

Abbildung 5 Vergleich der simulierten Trennschärfen des klassischen t-Tests bei zwei Stichproben und des t-Tests bei zwei Stichproben nach Welch in balancierten Designs mit ungleichen Varianzen. Stichproben wurden so aus normalverteilten Grundgesamtheiten mit ungleichen Varianzen gezogen, dass das Verhältnis der Standardabweichungen 0,5; 1,5 und 2,0 war.

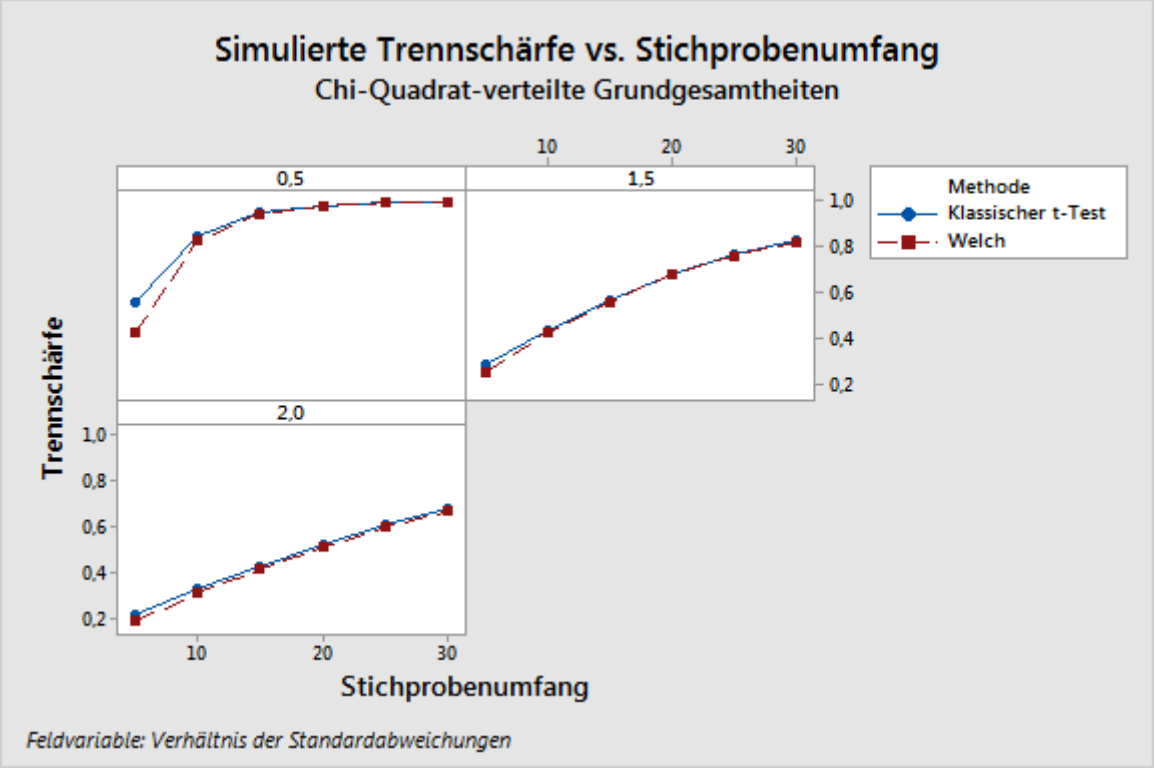

Abbildung 6 Vergleich der simulierten Trennschärfen des klassischen t-Tests bei zwei Stichproben und des t-Tests bei zwei Stichproben nach Welch in balancierten Designs mit ungleichen Varianzen. Stichproben wurden so aus Chi-Quadrat-Grundgesamtheiten mit ungleichen Varianzen gezogen, dass das Verhältnis der Standardabweichungen 0,5; 1,5 und 2,0 war.

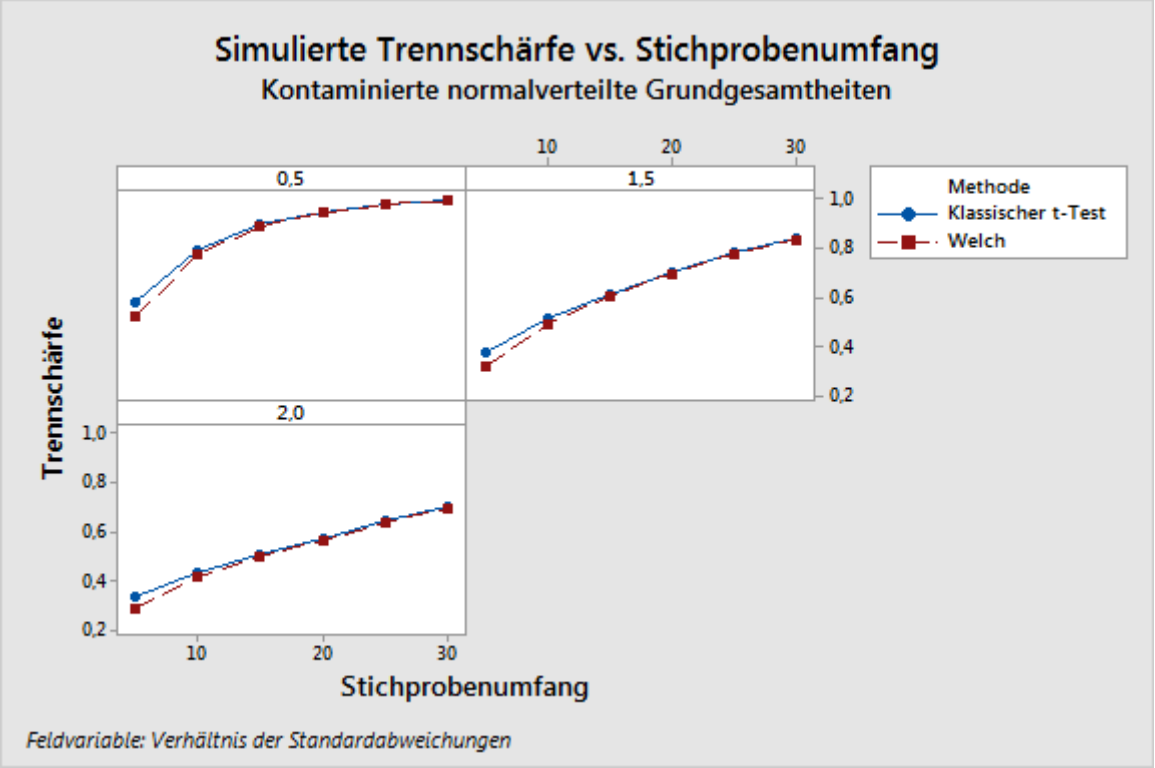

Abbildung 7 Vergleich der simulierten Trennschärfen des klassischen t-Tests bei zwei Stichproben und des t-Tests bei zwei Stichproben nach Welch in balancierten Designs mit ungleichen Varianzen. Stichproben wurden so aus kontaminierten normalverteilten Grundgesamtheiten mit ungleichen Varianzen gezogen, dass das Verhältnis der Standardabweichungen 0,5; 1,5 und 2,0 war.

Anhang C: Trennschärfe und Stichprobenumfang und Empfindlichkeit gegenüber der Annahme einer Normalverteilung

Im Assistenten basiert die Trennschärfeanalyse zum Vergleichen der Mittelwerte zweier Grundgesamtheiten auf der Trennschärfefunktion des t-Tests nach Welch. Sollte diese Funktion empfindlich gegenüber der Annahme der Normalverteilung sein, unter der sie abgeleitet wurde, kann die Trennschärfeanalyse zu fehlerhaften Schlussfolgerungen führen. Daher haben wir eine Simulationsstudie ausgeführt, um die Empfindlichkeit dieser Funktion in Bezug auf die Annahme der Normalverteilung zu untersuchen. Die Empfindlichkeit wird als Übereinstimmung der simulierten Trennschärfen und der Trennschärfen bewertet, die aus der theoretischen Trennschärfefunktion berechnet werden, wenn Stichproben aus Nicht-Normalverteilungen generiert wurden. Die Normalverteilung fungiert als Kontroll-Grundgesamtheit, da die simulierten Trennschärfen und die theoretischen Trennschärfen laut Theorem B2 am dichtesten beieinander liegen, wenn Stichproben aus normalverteilten Grundgesamtheiten generiert werden.

Simulationsstudie C

Die Studie wird in drei Teilen mit drei Verteilungen durchgeführt: Normalverteilung, Chi-Quadrat-Verteilung und kontaminierte Normalverteilung. Weitere Informationen finden Sie in Anhang A. Für jeden Teil der Studie wird die simulierte Trennschärfe (für die angegebenen Stichprobenumfänge n_1 und n_2 bei einer angegebenen erkennbaren Differenz δ) als Anteil der Instanzen berechnet, für die die Nullhypothese zurückgewiesen wurde, wenn sie nicht zutreffend war. In allen Fällen ist die zu erkennende Differenz in einer Einheit des Standards in der Basis-Grundgesamtheit angegeben. Dies ist $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ für alle drei Klassen von Verteilungen in dieser Studie. Die theoretischen Trennschärfewerte des t-Tests nach Welch werden zu Vergleichszwecken ebenfalls berechnet.

Ergebnisse und Zusammenfassung der Simulation

Die Ergebnisse zeigen, dass die Trennschärfefunktion des t-Tests nach Welch für relativ kleine Stichprobenumfänge unempfindlich gegenüber der Annahme der Normalverteilung ist. Wenn der minimale Umfang der beiden Stichproben lediglich 15 beträgt, liegen die simulierten Trennschärfewerte im Allgemeinen immer noch dicht bei ihren entsprechenden theoretischen Soll-Trennschärfen (siehe Tabellen 7-10 und Abbildungen 8-10).

Die Tabellen 7-10 zeigen die simulierten Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$, basierend auf Paaren von Stichproben, die aus einer normalverteilten Grundgesamtheit, schiefen Grundgesamtheiten (Chi-Quadrat) und kontaminierten normalverteilten Grundgesamtheiten generiert wurden. Die Paare von Stichproben stammen aus derselben Klasse von Verteilungen, die Varianzen der übergeordneten

Grundgesamtheiten sind jedoch nicht unbedingt gleich. Zu Vergleichszwecken wurden die theoretischen Trennschärfewerte berechnet.

Tabelle 7 Simulierte Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$ für n=5

			Basis-Grundgesamtheit: N(0;2)				**Basis-Grundgesamtheit: Chi(2)**				**Basis-Grundgesamtheit: CN(0,8;4)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	0,6	Beob.	0,288	0,158	0,113	0,091	0,432	0,305	0,211	0,149	0,361	0,257	0,234	0,220
		Soll	0,353	0,192	0,116	0,092	0,353	0,192	0,116	0,092	0,353	0,192	0,116	0,092
5	1,0	Beob.	0,370	0,252	0,169	0,121	0,427	0,334	0,248	0,189	0,522	0,380	0,319	0,284
		Soll	0,389	0,286	0,190	0,137	0,389	0,286	0,190	0,137	0,389	0,286	0,190	0,137
8	1,6	Beob.	0,387	0,326	0,242	0,179	0,427	0,364	0,286	0,225	0,573	0,453	0,374	0,319
		Soll	0,400	0,345	0,260	0,193	0,400	0,345	0,260	0,193	0,400	0,345	0,260	0,193
10	2,0	Beob.	0,390	0,351	0,272	0,208	0,421	0,373	0,296	0,235	0,590	0,483	0,394	0,336
		Soll	0,402	0,364	0,291	0,223	0,402	0,364	0,291	0,223	0,402	0,364	0,291	0,223

Tabelle 8 Simulierte Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$ für n=10

			Basis-Grundgesamtheit: N(0;2)				**Basis-Grundgesamtheit: Chi(2)**				**Basis-Grundgesamtheit: CN(0,8;4)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	0,5	Beob.	0,651	0,346	0,197	0,131	0,768	0,493	0,320	0,221	0,689	0,484	0,404	0,358

			Basis-Grundgesamtheit: N(0;2)				**Basis-Grundgesamtheit: Chi(2)**				**Basis-Grundgesamtheit: CN(0,8;4)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
		Soll	0,666	0,364	0,206	0,139	0,666	0,364	0,206	0,139	0,666	0,364	0,206	0,139
10	1,0	Beob.	0,742	0,556	0,369	0,254	0,831	0,612	0,430	0,308	0,776	0,619	0,496	0,419
		Soll	0,745	0,562	0,337	0,259	0,745	0,562	0,337	0,259	0,745	0,562	0,337	0,259
15	1,5	Beob.	0,765	0,641	0,483	0,358	0,865	0,679	0,511	0,377	0,792	0,679	0,547	0,456
		Soll	0,767	0,643	0,483	0,352	0,767	0,643	0,483	0,352	0,767	0,643	0,483	0,352
20	2	Beob.	0,774	0,683	0,549	0,417	0,898	0,737	0,565	0,448	0,797	0,716	0,596	0,490
		Soll	0,777	0,686	0,551	0,422	0,777	0,686	0,551	0,422	0,777	0,686	0,551	0,422

Tabelle 9 Simulierte Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$ für n=15

			Basis-Grundgesamtheit: N(0;2)				**Basis-Grundgesamtheit: Chi(2)**				**Basis-Grundgesamtheit: CN(0,8;4)**			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	0,53	Beob.	0,857	0,569	0,342	0,229	0,871	0,651	0,421	0,293	0,853	0,632	0,505	0,428
		Soll	0,861	0,568	0,338	0,221	0,861	0,568	0,338	0,221	0,861	0,568	0,338	0,221
15	1,0	Beob.	0,906	0,745	0,535	0,368	0,942	0,763	0,563	0,415	0,891	0,760	0,611	0,500
		Soll	0,910	0,753	0,541	0,379	0,910	0,753	0,541	0,379	0,910	0,753	0,541	0,379
23	1,53	Beob.	0,928	0,831	0,667	0,502	0,975	0,858	0,676	0,517	0,898	0,825	0,698	0,572

		Soll	0,925	0,830	0,670	0,509	0,925	0,830	0,670	0,509	0,925	0,830	0,670	0,509
30	2,0	Beob.	0,933	0,861	0,737	0,589	0,984	0,903	0,750	0,598	0,902	0,847	0,742	0,619
		Soll	0,931	0,863	0,736	0,589	0,931	0,863	0,736	0,589	0,931	0,863	0,736	0,589

Tabelle 10 Simulierte Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$ für n=20

			Basis-Grundgesamtheit: N(0;2)				**Basis-Grundgesamtheit: Chi(2)**				**Basis-Grundgesamtheit: CN(0,8;4)**			
		$\frac{\sigma_2}{\sigma_1}$	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**	**0,5**	**1,0**	**1,5**	**2,0**
n_2	$\frac{n_2}{n_1}$		$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	0,5	Beob.	0,938	0,687	0,426	0,275	0,920	0,698	0,486	0,333	0,923	0,716	0,568	0,476
		Soll	0,941	0,686	0,424	0,277	0,941	0,686	0,424	0,277	0,941	0,686	0,424	0,277
20	1,0	Beob.	0,971	0,866	0,672	0,485	0,981	0,858	0,670	0,506	0,952	0,856	0,696	0,567
		Soll	0,971	0,869	0,673	0,489	0,971	0,869	0,673	0,489	0,971	0,869	0,673	0,489
30	1,5	Beob.	0,977	0,923	0,791	0,629	0,995	0,932	0,785	0,631	0,960	0,908	0,798	0,662
		Soll	0,978	0,922	0,791	0,628	0,978	0,922	0,791	0,628	0,978	0,922	0,791	0,628
40	2,0	Beob.	0,983	0,950	0,858	0,724	0,998	0,966	0,864	0,726	0.958	0,929	0,845	0,725
		Soll	0,981	0,945	0,854	0,719	0,981	0,945	0,854	0,719	0,981	0,945	0,854	0,719

Wenn die zwei Stichproben aus normalverteilten Grundgesamtheiten generiert werden, stimmen die simulierten Trennschärfewerte mit den theoretischen Trennschärfewerten überein, selbst bei sehr kleinen Stichproben. Wie in Abbildung 7 veranschaulicht, sind die Kurven der theoretischen und der simulierten Trennschärfe praktisch nicht zu unterscheiden. Diese Ergebnisse stimmen mit Theorem B2 überein.

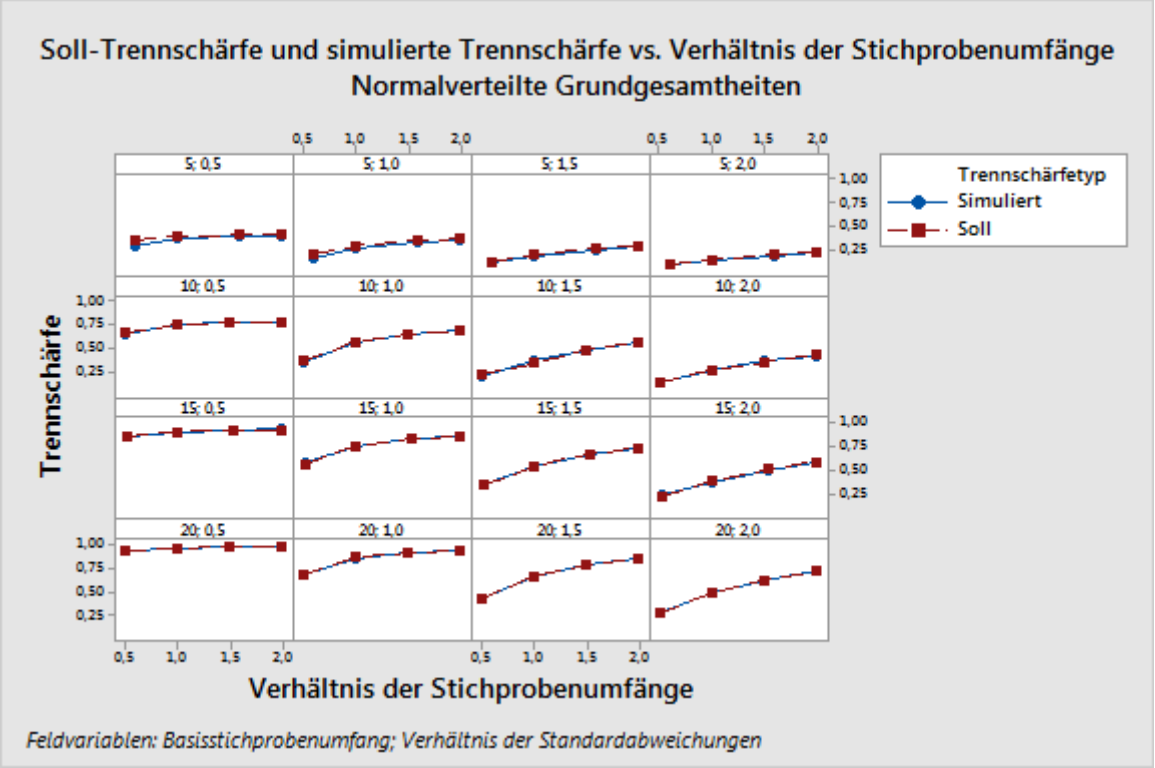

Abbildung 8 Simulierte Trennschärfen und theoretische Soll-Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$, basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

Werden die Stichproben aus den schiefen Chi-Quadrat-Verteilungen generiert, sind die simulierten Trennschärfewerte für sehr kleine Stichproben höher als die theoretischen Trennschärfewerte; die Trennschärfewerte nähern sich jedoch bei steigenden Stichprobenumfängen aneinander an. Abbildung 9 zeigt, dass die Kurven der theoretischen Soll-Trennschärfen und der simulierten Trennschärfen durchgehend nah beieinander liegen, wenn der minimale Umfang der beiden Stichproben mindestens 10 beträgt. Dies veranschaulicht, dass schiefe Daten keinen erkennbaren Effekt auf die Trennschärfefunktion des t-Tests nach Welch haben, selbst bei relativ kleinen Stichprobenumfängen.

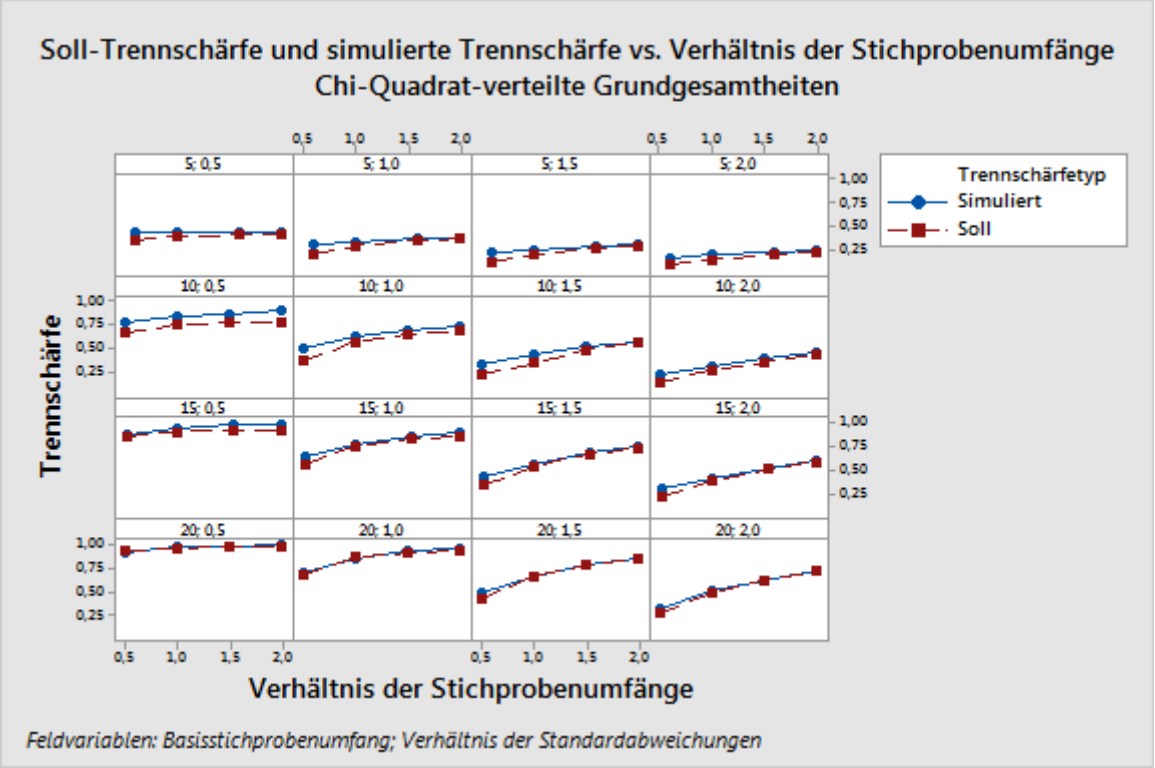

Abbildung 9 Simulierte Trennschärfen und theoretische Soll-Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$, basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

Darüber hinaus haben Ausreißer tendenziell nur dann einen Einfluss auf die Trennschärfefunktion, wenn die Stichprobenumfänge sehr klein sind. Wenn Ausreißer vorliegen, sind die simulierten Trennschärfewerte tendenziell etwas höher als die theoretischen Soll-Trennschärfewerte. Dies wird in Abbildung 10 veranschaulicht, in der die Kurven der simulierten und der theoretischen Trennschärfen erst ab einem minimalen Stichprobenumfang von 15 relativ dicht beieinander liegen.

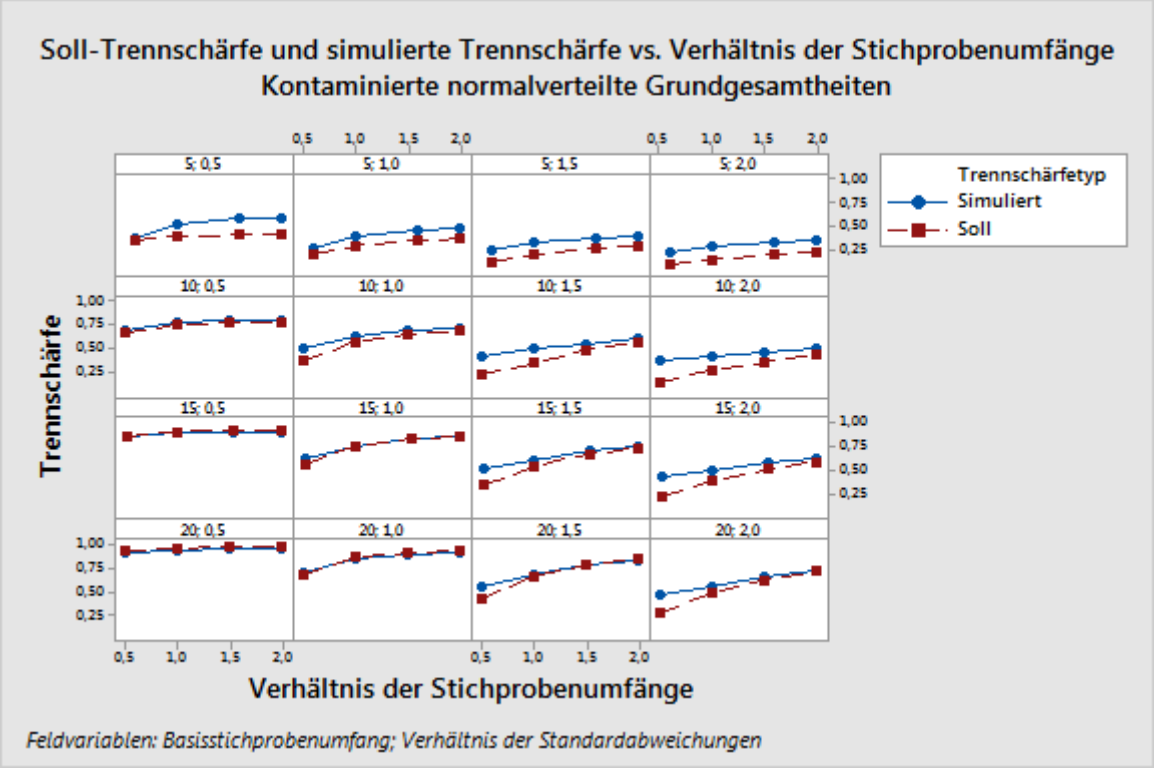

Abbildung 10 Simulierte Trennschärfen und theoretische Soll-Trennschärfen eines beidseitigen t-Tests nach Welch mit $\alpha = 0{,}05$, basierend auf Paaren von Stichproben, die aus zwei normalverteilten Grundgesamtheiten mit gleichen oder ungleichen Varianzen generiert wurden, dargestellt im Vergleich zum Verhältnis der Stichprobenumfänge.

Anhang D: Beweis von Theorem B2

Für das Modell mit zwei Stichproben basiert der Welch-Ansatz zum Ableiten der Verteilung der Teststatistik

$$t_w(x,y) = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

unter der Nullhypothese auf einer Approximation der Verteilung von

$$V = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

proportional zu einer Chi-Quadrat-Verteilung. Konkreter:

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

ist annähernd als Chi-Quadrat-Verteilung mit d_W Freiheitsgraden verteilt, wobei

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)}}$$

(Beachten Sie, dass sich dies in einem Fall mit einer Stichprobe auf das allgemein bekannte klassische Ergebnis $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ reduziert.)

Betrachten Sie den Test der Nullhypothese H_o: $\mu_1 = \mu_2$ (oder äquivalent $\delta = 0$) gegen die Alternative H_A: $\mu_1 \neq \mu_2$ (oder äquivalent $\delta \neq 0$)

Unter der Nullhypothese ist die Trennschärfefunktion

$$\pi(n_1, n_2, \delta) = \pi(n_1, n_2, 0) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \leq \frac{\bar{x} - \bar{y}}{\sqrt{V}} \leq t_{d_W}^{\alpha/2}\right) \approx \alpha$$

Hierbei ist t_d^α der obere 100 α. Perzentilpunkt der t-Verteilung mit d Freiheitsgraden.

Unter der Alternativhypothese besitzt

$$\frac{\bar{x} - \bar{y}}{\sqrt{V}} = \frac{\frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{d_W V}{d_W\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}}$$

die approximierte nicht zentrale t-Verteilung mit d_W Freiheitsgraden mit Nichtzentralitätsparameter

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

da, wie bereits erklärt,

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

annähernd als Chi-Quadrat-Verteilung mit d_W-Freiheitsgraden verteilt ist und

$$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

als Standardnormalverteilung verteilt ist.

Daraus folgt, dass unter der Alternative

$$\pi(n_1, n_2, \delta) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \leq \frac{\bar{x}-\bar{y}}{\sqrt{V}} \leq t_{d_W}^{\alpha/2}\right) \approx 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

Hierbei ist $G_{d_W,\lambda}(.)$ die kumulative Verteilungsfunktion der nicht zentralen t-Verteilung mit d_W Freiheitsgraden und Nichtzentralitätsparameter λ, wie oben angegeben.

Anhang E: Beweis von Theorem B3

Beachten Sie zunächst, dass d_W umformuliert werden kann als

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{\rho^2}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{\rho^4}{n_2^2(n_2-1)}}$$

Hierbei ist $\rho = \sigma_1/\sigma_2$.

Ebenso kann der Nichtzentralitätsparameter für die Trennschärfefunktion des t-Tests nach Welch wie folgt geschrieben werden:

$$\lambda_W = \frac{\delta/\sigma_1}{\sqrt{1/n_1 + \rho^2/n_2}}$$

Unter der Annahme der Gleichheit der Varianzen stimmen die Nichtzentralitätsparameter für die Trennschärfefunktionen des klassischen t-Tests bei zwei Stichproben und des t-Tests nach Welch überein. Das heißt

$$\lambda = \lambda_W = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Hierbei ist σ die gemeinsame Varianz der beiden Grundgesamtheiten. Damit beschränkt sich die Differenz der Trennschärfefunktionen der beiden Tests auf die Differenz zwischen ihren jeweiligen Freiheitsgraden. Unter der Annahme gleicher Varianzen werden die Freiheitsgrade für die Trennschärfefunktion des t-Tests nach Welch jedoch zu

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{1}{n_2^2(n_2-1)}} = \frac{(n_1+n_2)^2(n_1-1)(n_2-1)}{n_1^2(n_1-1) + n_2^2(n_2-1)}$$

Laut Theorem 1 sind die Freiheitsgrade für die Trennschärfefunktion des klassischen t-Tests bei zwei Stichproben $d_C = n_1 + n_2 - 2$. Nach algebraischem Umformen ergibt sich

$$d_C - d_W = \frac{(n_1-n_2)^2(n_1+n_2-1)^2}{n_1^2(n_1-1) + n_2^2(n_2-1)} \geq 0$$

Der Umstand $d - d_W \geq 0$ ist nicht überraschend, da der klassische t-Test bei zwei Stichproben unter der Annahme der Gleichheit der Varianzen bekanntlich ein gleichmäßig trennschärfster Test (uniformly most powerful, UMP) ist; daher können höhere Freiheitsgrade für diese Trennschärfefunktion erwartet werden.

Wenn also $n_1 \sim n_2$, dann $d \sim d_W$, und infolgedessen weisen die Trennschärfefunktionen die gleiche Größenordnung auf. Die Trennschärfefunktionen der beiden Tests sind insbesondere dann identisch, wenn $n_1 = n_2$. Dies beweist den ersten Teil von Theorem 2.3.

Wenn $n_1 \neq n_2$, dann $d_C - d_W > 0$, so dass der t-Test nach Welch eine geringere Trennschärfe als der klassische t-Test bei zwei Stichproben aufweist.

Wenn die Stichproben zudem groß sind, d. h., wenn $n_1 \to \infty$ und $n_2 \to \infty$, dann $d_C \to \infty$ und $d_W \to \infty$, so dass die asymptotische Verteilung der Teststatistik für beide Tests die

Standardnormalverteilung ist. Damit sind die Tests asymptotisch äquivalent und liefern die gleiche asymptotische Trennschärfefunktion.