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

Test für Prozentsatz fehlerhafter Einheiten bei zwei Stichproben

Übersicht

Mit einem Test von Anteilen bei zwei Stichproben wird festgestellt, ob eine signifikante Differenz zwischen zwei Stichproben vorliegt. In der Qualitätsanalyse kommt der Test häufig zur Anwendung, wenn ein Produkt oder eine Dienstleistung als fehlerhaft oder nicht fehlerhaft eingestuft wird, um zu bestimmen, ob eine signifikante Differenz für den Prozentsatz fehlerhafter Einheiten von Stichproben vorliegt, die in zwei unabhängigen Prozessen erfasst wurden.

Der Minitab-Assistent bietet einen Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben. Für die Daten des Tests wird die Anzahl der fehlerhaften Einheiten in jeder der beiden unabhängigen Stichproben erfasst. Es wird angenommen, dass es sich hierbei um die beobachteten Werte einer binomial verteilten Zufallsvariablen handelt. Der Assistent nutzt zum Berechnen der Ergebnisse des Hypothesentests exakte Methoden. Daher sollte die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art nahe dem für den Test angegebenen Signifikanzniveau (Alpha) liegen, so dass keine weitere Untersuchung erforderlich ist. Der Assistent verwendet jedoch eine Methode der Normal-Approximation, um das Konfidenzintervall (KI) für die Differenz im Prozentsatz fehlerhafter Einheiten zu berechnen, sowie eine theoretische Trennschärfefunktion des Tests auf Normal-Approximation, um die Analyse von Trennschärfe und Stichprobenumfang auszuführen. Da es sich hierbei um Approximationsmethoden handelt, müssen diese in Bezug auf ihre Genauigkeit untersucht werden.

Im vorliegenden White Paper werden die Bedingungen untersucht, unter denen die approximierten Konfidenzintervalle genau sind. Darüber hinaus wird die Methode zum Auswerten von Trennschärfe und Stichprobenumfang des Tests für den Prozentsatz

fehlerhafter Einheiten bei zwei Stichproben untersucht, indem die theoretische Trennschärfe der Approximationsmethode mit der tatsächlichen Trennschärfe des exakten Tests verglichen wird. Schließlich 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:

- Gültigkeit des KI
- Stichprobenumfang

Der Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben hängt zudem von anderen Annahmen ab. Weitere Informationen finden Sie in Anhang A.

Methoden des Tests für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben

Genauigkeit der Konfidenzintervalle

Der Assistent verwendet zwar Fishers exakten Test, um auszuwerten, ob eine signifikante Differenz zwischen den Prozentsätzen fehlerhafter Einheiten der zwei Stichproben vorliegt, das Konfidenzintervall für die Differenz basiert jedoch auf der Methode der Normal-Approximation. Gemäß der häufig in der Fachliteratur zur Statistik anzutreffenden allgemeinen Regel ist dieses approximierte Konfidenzintervall dann genau, wenn die beobachtete Anzahl der fehlerhaften Einheiten und die beobachtete Anzahl der nicht fehlerhaften Einheiten in jeder Stichprobe mindestens 5 beträgt.

Zielstellung

Wir wollten die Bedingungen untersuchen, unter denen die auf der Normal-Approximation basierenden Konfidenzintervalle genau sind. Insbesondere sollte festgestellt werden, wie sich die allgemeine Regel in Bezug auf die Anzahl der fehlerhaften Einheiten und die Anzahl der nicht fehlerhaften Einheiten in jeder Stichprobe auf die Genauigkeit der approximierten Konfidenzintervalle auswirkt.

Methode

De Formel zum Berechnen des Konfidenzintervalls für die Differenz zwischen den zwei Anteilen und die allgemeine Regel, mit der seine Genauigkeit sichergestellt wird, werden in Anhang D beschrieben. Außerdem erläutern wir eine weniger strikte, abgewandelte Regel, die wir im Verlauf unserer Untersuchung entwickelt haben.

Wir haben Simulationen zum Auswerten der Genauigkeit des approximierten Konfidenzintervalls unter verschiedenen Bedingungen durchgeführt. Für die Simulationen wurden zufällige Paare von Stichproben unterschiedlicher Umfänge aus mehreren Bernoulli-verteilten Grundgesamtheiten generiert. Für jeden Typ von Bernoulli-verteilter Grundgesamtheit wurde ein approximiertes Konfidenzintervall für die Differenz zwischen den zwei Anteilen für jedes Paar der 10.000 Bernoulli-verteilten Stichprobenreplikationen berechnet. Anschließend wurde der Anteil der 10.000 Intervalle berechnet, der die tatsächliche Differenz zwischen den zwei Anteilen enthält. Dieser Anteil wird als simulierte Überdeckungswahrscheinlichkeit bezeichnet. Wenn das approximierte Intervall genau ist, sollte die simulierte Überdeckungswahrscheinlichkeit nahe der Soll-Überdeckungswahrscheinlichkeit von 0,95 liegen. Zum Auswerten der Genauigkeit des approximierten Intervalls in Bezug auf die ursprüngliche und die abgewandelte Regel für die in jeder Stichprobe erforderliche minimale Anzahl von fehlerhaften Einheiten und nicht fehlerhaften Einheiten wurde zudem der Prozentsatz der 10.000 Paare von Stichproben

berechnet, für die die jeweilige Regel erfüllt wurde. Weitere Informationen finden Sie in Anhang D.

Ergebnisse

Das approximierte Konfidenzintervall für die Differenz zwischen zwei Anteilen ist generell genau, wenn die Stichproben ausreichend groß sind, d. h., wenn die beobachtete Anzahl von fehlerhaften Einheiten und die beobachtete Anzahl von nicht fehlerhaften Einheiten in jeder Stichprobe mindestens 5 beträgt. Daher haben wir diese Regel für die Prüfung der Gültigkeit des KI in die Auswertung übernommen. Obwohl diese Regel im Allgemeinen eine gute Leistung zeigt, kann sie in einigen Fällen übermäßig konservativ sein, und u. U. fällt sie etwas zu locker aus, wenn die beiden Anteile nahe 0 oder 1 liegen. Weitere Informationen hierzu finden Sie im Abschnitt „Datenprüfungen“ und in Anhang D.

Leistung der theoretischen Trennschärfefunktion

Der Assistent führt den Hypothesentest durch, um die Anteile aus zwei Bernoulli-verteilten Grundgesamtheiten (Prozentsatz fehlerhafter Einheiten in zwei Stichproben) mit Fishers Test zu vergleichen. Da die Trennschärfefunktion dieses exakten Tests jedoch nicht auf einfache Weise abgeleitet werden kann, muss die Trennschärfefunktion anhand der theoretischen Trennschärfefunktion des entsprechenden Tests auf Normal-Approximation approximiert werden.

Zielstellung

Wir wollten feststellen, ob die theoretische Trennschärfefunktion auf der Grundlage des Tests auf Normal-Approximation zum Auswerten der Anforderungen an Trennschärfe und Stichprobenumfang für den Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben im Assistenten verwendet werden kann. Hierfür musste untersucht werden, ob diese theoretische Trennschärfefunktion die tatsächliche Trennschärfe von Fishers exaktem Test genau abbildet.

Methode

Die Methodologie für Fishers exakten Test, einschließlich der Berechnung des zugehörigen p-Werts, wird ausführlich in Anhang B beschrieben. Eine Definition der theoretischen Trennschärfefunktion auf der Grundlage des Tests auf Normal-Approximation wird in Anhang C gegeben. Auf der Grundlage dieser Definitionen wurden Simulationen zum Schätzen der tatsächlichen Trennschärfen von Fishers exaktem Test (die wir als simulierte Trennschärfen bezeichnen) durchgeführt, wenn dieser Test zum Analysieren der Differenz zwischen den Prozentsätzen fehlerhafter Einheiten in zwei Stichproben verwendet wird.

Für die Simulationen wurden zufällige Paare von Stichproben unterschiedlicher Umfänge aus mehreren Bernoulli-verteilten Grundgesamtheiten generiert. Für jede Kategorie von Bernoulli-verteilter Grundgesamtheit wurde Fishers exakter Test für jedes Paar der 10.000 Stichprobenreplikationen durchgeführt. Für jeden Stichprobenumfang wurde die simulierte Trennschärfe des Tests zum Erkennen einer gegebenen Differenz als Anteil der 10.000 Stichprobenpaare berechnet, bei denen der Test signifikant ist. Zum Vergleich wurde auch die entsprechende theoretische Trennschärfe auf der Grundlage des Tests auf Normal-

Approximation berechnet. Wenn die Approximation gute Ergebnisse liefert, liegen die theoretischen und simulierten Trennschärfen nah beieinander. Weitere Informationen finden Sie in Anhang E.

Ergebnisse

Unsere Simulationen haben gezeigt, dass die theoretische Trennschärfefunktion des Tests auf Normal-Approximation und die simulierte Trennschärfefunktion von Fishers exaktem Test im Allgemeinen annähernd gleich sind. Daher nutzt der Assistent die theoretische Trennschärfefunktion des Tests auf Normal-Approximation, um die Stichprobenumfänge zu schätzen, mit denen bei Fishers exaktem Test Differenzen mit praktischen Konsequenzen erkannt werden können.

Datenprüfungen

Gültigkeit des KI

Da beim Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben ein exakter Test zum Auswerten der Differenz zwischen den Prozentsätzen fehlerhafter Einheiten genutzt wird, wirken sich die Anzahl von fehlerhaften Einheiten und die Anzahl nicht fehlerhaften Einheiten in den einzelnen Stichproben nicht wesentlich auf seine Genauigkeit aus. Das Konfidenzintervall für die Differenz zwischen den Prozentsätzen fehlerhafter Einheiten basiert jedoch auf einer Normal-Approximation. Steigt die Anzahl der fehlerhaften Einheiten und der nicht fehlerhaften Einheiten in den einzelnen Stichproben, nimmt auch die Genauigkeit des approximierten Konfidenzintervalls zu (siehe Anhang D).

Zielstellung

Wir wollten herausfinden, ob die Anzahl der fehlerhaften Einheiten und die Anzahl der nicht fehlerhaften Einheiten in den Stichproben ausreichen, um die Genauigkeit des approximierten Konfidenzintervalls für die Different zwischen den Prozentsätzen fehlerhafter Einheiten sicherzustellen.

Methode

Wir haben die allgemeine, in den meisten statistischen Fachbüchern angeführte Regel verwendet. Wenn jede Stichprobe mindestens 5 fehlerhafte Einheiten und 5 nicht fehlerhafte Einheiten enthält, ist das approximierte Konfidenzintervall für den Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben genau. Weitere Einzelheiten finden Sie im obigen Abschnitt „Methoden des Tests für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben".

Ergebnisse

Wie in den Simulationen veranschaulicht, die im Abschnitt „Methoden des Tests für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben" kurz erläutert werden, hängt die Genauigkeit des Konfidenzintervalls von der Mindestanzahl fehlerhafter und nicht fehlerhafter Einheiten in den einzelnen Stichproben ab. Daher zeigt der des Assistent in der Auswertung die folgenden Statusindikatoren an, anhand derer Sie die Genauigkeit des Konfidenzintervalls für die Different zwischen den zwei Prozentsätzen fehlerhafter Einheiten auswerten können:

Status	Bedingung
	Beide Stichproben enthalten mindestens 5 fehlerhafte Einheiten und 5 nicht fehlerhafte Einheiten. Das Konfidenzintervall für die Differenz sollte genau sein.
	Die Anzahl der fehlerhaften Einheiten oder die Anzahl der nicht fehlerhaften Einheiten in mindestens einer Stichprobe ist kleiner als 5. Das Konfidenzintervall für die Differenz ist möglicherweise nicht genau.

Stichprobenumfang

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

Zielstellung

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

Methode

Die Analyse von Trennschärfe und Stichprobenumfang für den Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben basiert auf der theoretischen Trennschärfefunktion des Tests auf Normal-Approximation, die einen guten Schätzwert der tatsächlichen Trennschärfe von Fishers exaktem Test liefert (siehe die in „Leistung der theoretischen Trennschärfefunktion“ zusammengefassten Simulationsergebnisse im Abschnitt „Methoden des Tests für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben“). Die theoretische Trennschärfefunktion kann als Funktion der Solldifferenz zwischen den Prozentsätzen fehlerhafter Einheiten oder dem Gesamtprozentsatz fehlerhafter Einheiten in den kombinierten Stichproben ausgedrückt werden.

Ergebnisse

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

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

Status	Bedingung
	Im Test wird eine Differenz zwischen den Prozentsätzen fehlerhafter Einheiten festgestellt, daher stellt die Trennschärfe kein Problem dar. ODER Die Trennschärfe ist ausreichend. Im Test wurde keine Differenz zwischen den Prozentsätzen fehlerhafter Einheiten festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 90 % erkannt wird (Trennschärfe $\geq$ 0,90).
	Die Trennschärfe ist möglicherweise ausreichend. Im Test wurde keine Differenz zwischen den Prozentsätzen fehlerhafter Einheiten festgestellt, die Stichprobe ist jedoch umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 80 % bis 90 % erkannt wird (0,80 $\leq$ Trennschärfe < 0,90). Der erforderliche Stichprobenumfang zum Erzielen einer Trennschärfe von 90 % wird ausgegeben.
	Die Trennschärfe ist möglicherweise nicht ausreichend. Im Test wurde keine Differenz zwischen den Prozentsätzen fehlerhafter Einheiten festgestellt, und die Stichprobe ist umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von 60 % bis 80 % erkannt wird (0,60 $\leq$ Trennschärfe < 0,80). 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 Prozentsätzen fehlerhafter Einheiten festgestellt, und die Stichprobe ist nicht umfassend genug, dass die angegebene Differenz mit einer Wahrscheinlichkeit von mindestens 60 % erkannt wird (Trennschärfe < 0,60). Die erforderlichen Stichprobenumfänge zum Erzielen einer Trennschärfe von 80 % und 90 % werden ausgegeben.
	Im Test wurde keine Differenz zwischen den Prozentsätzen fehlerhafter Einheiten festgestellt. Sie haben keine zu erkennende Differenz mit praktischen Konsequenzen angegeben. Abhängig von Ihren Daten werden in der Auswertung u. U. die Differenzen angegeben, die die bei Ihrem dem Stichprobenumfang und dem Alpha mit einer Wahrscheinlichkeit von 80 % und 90 % erkannt werden können.

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.

Anhang A: Zusätzliche Annahmen für den Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben

Dem Test für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben liegen die folgenden Annahmen zugrunde:

- Die Daten in jeder Stichprobe bestehen aus n verschiedenen Einheiten, wobei jede Einheit als fehlerhaft oder als nicht fehlerhaft klassifiziert ist.
- Die Wahrscheinlichkeit, dass eine Einheit fehlerhaft ist, ist für jede Einheit in einer Stichprobe gleich.
- Die Wahrscheinlichkeit, dass eine Einheit fehlerhaft ist, wird nicht dadurch beeinflusst, ob eine andere Einheit fehlerhaft ist.

Die Richtigkeit dieser Annahmen kann in den Datenprüfungen der Auswertung im Assistenten nicht bestätigt werden, da für diesen Test Zusammenfassungsdaten und keine Rohdaten erfasst werden.

Anhang B: Fishers exakter Test

Angenommen, es werden zwei unabhängige Zufallsstichproben $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ aus Bernoulli-Verteilungen beobachtet, so dass

$$p_1 = \Pr(X_i = 1) = 1 - \Pr(X_i = 0) \text{ und } p_2 = \Pr(Y_j = 1) = 1 - \Pr(Y_j = 0)$$

In den folgenden Abschnitten werden die Verfahren beschrieben, mit denen Rückschlüsse auf die Differenz zwischen den Anteilen $\delta = p_1 - p_2$ gezogen werden.

Formel B1: Fishers exakter Test und p-Wert

Eine Beschreibung von Fishers exaktem Test findet sich in Arnold (1994). Wir geben eine kurze Beschreibung des Tests.

Sei V die Anzahl der Erfolge in der ersten Stichprobe und $v = n_1\hat{p}_1$ die beobachtete Anzahl der Erfolge in der ersten Stichprobe bei Durchführung eines Experiments. Sei außerdem W die Gesamtzahl der Erfolge in den zwei Stichproben und $w = n_1\hat{p}_1 + n_2\hat{p}_2$ die Anzahl der beobachteten Erfolge bei Durchführung eines Experiments. Beachten Sie, dass $\hat{p}_1$ und $\hat{p}_2$ die the Stichproben-Punktschätzungen von p_1 und p_2 sind.

Unter der Nullhypothese $\delta = p_1 - p_2 = 0$ ist die bedingte Verteilung von V bei W die hypergeometrische Verteilung mit der Wahrscheinlichkeitsbelegungsfunktion

$$f(v|w) = \frac{\binom{n_1}{v}\binom{n_2}{w}}{\binom{n_1 + n_2}{w}}$$

Sei $F(v|w)$ die kumulative Verteilungsfunktion der Verteilung. Dann lauten die p-Werte für den einseitigen und den beidseitigen Test:

- **Beim Testen gegen $\delta < 0$ bzw. als Äquivalent $p_1 < p_2$**

  Der p-Wert wird berechnet als $F(v|w)$, wobei v der beobachtete Wert von V bzw. die beobachtete Anzahl der Erfolge in der ersten Stichprobe und w der beobachtete Wert von W bzw. die beobachtete Anzahl der Erfolge in beiden Stichproben ist.

- **Beim Testen gegen $\delta > 0$ bzw. als Äquivalent $p_1 > p_2$**

  Der p-Wert wird berechnet als $1 - F(v - 1|w)$, wobei v der beobachtete Wert von V bzw. die beobachtete Anzahl der Erfolge in der ersten Stichprobe und w der beobachtete Wert von W bzw. die beobachtete Anzahl der Erfolge in beiden Stichproben ist.

- **Beim Testen gegen $\delta \neq 0$ bzw. als Äquivalent $p_1 \neq p_2$**

  Der p-Wert wird entsprechend dem folgenden Algorithmus berechnet, wobei m der Modalwert der oben beschriebenen hypergeometrischen Verteilung ist.

  - Wenn $v < m$, dann wird der p-Wert als $1 - F(y-1|w) + F(v|w)$ berechnet, wobei v und w wie oben definiert lauten und $y = \min\{k \geq m: f(k|w) \leq f(v|W)\}$
  - Wenn $v = m$, dann ist der p-Wert 1,0
  - Wenn $v > m$, dann wird der p-Wert als $1 - F(v-1|w) + F(y|w)$ berechnet, wobei v und w wie oben definiert lauten und $y = \max\{k \leq m: f(k|w) \leq f(v|W)\}$

Anhang C: Theoretische Trennschärfefunktion

Zum Vergleichen von zwei Anteilen (oder konkreter von zwei Prozentsätzen fehlerhafter Einheiten) wird Fishers exakter Test entsprechend der Beschreibung in Anhang B verwendet. Da eine theoretische Trennschärfefunktion dieses Tests zu komplex ist, um sie abzuleiten, verwenden wir eine approximierte Trennschärfefunktion. Konkret wird die Trennschärfefunktion des hinreichend bekannten Tests auf Normal-Approximation für zwei Anteile verwendet, um die Trennschärfe von Fishers exaktem Test zu approximieren.

Die Trennschärfefunktion der Normal-Approximation für den beidseitigen Test lautet

$$\pi(n_1, n_2, \delta) = 1 - \Phi\left(\frac{-\delta + z_{\frac{\alpha}{2}}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right) + \Phi\left(\frac{-\delta - z_{\alpha/2}\sqrt{p_c(1-p_c)(1/n_1\ +1/n_2)}}{se}\right)$$

Hierbei ist $\delta = p_1 - p_2$,

$$se = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

und $p_c = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$.

Beim Testen von $p_1 = p_2$ gegen $p_1 > p_2$ lautet die Trennschärfefunktion

$$\pi(n_1, n_2, \delta) = 1 - \Phi\left(\frac{-\delta + z_{\alpha}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right)$$

Beim Testen von $p_1 = p_2$ gegen $p_1 < p_2$ lautet die Trennschärfefunktion

$$\pi(n_1, n_2, \delta) = \Phi\left(\frac{-\delta - z_{\alpha}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right)$$

Anhang D: Approximierte Konfidenzintervalle

Formel D1: Berechnen eines approximierten Konfidenzintervalls für die Differenz zwischen zwei Anteilen

Ein asymptotisches $100(1-\alpha)\%$-Konfidenzintervall für $\delta = p_1 - p_2$ auf Grundlage der Normal-Approximation lautet:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

Eine hinreichend bekannte allgemeine Regel zum Auswerten der Zuverlässigkeit dieses approximierten Konfidenzintervalls besagt $n_1\hat{p}_1 \geq 5$; $n_1(1-\hat{p}_1) \geq 5$; $n_2\hat{p}_2 \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$. Anders ausgedrückt: Das Konfidenzintervall ist genau, wenn die beobachtete Anzahl von Erfolgen und Ausfällen in jeder Stichprobe mindestens 5 beträgt.

Hinweis: In diesem Abschnitt und den nachfolgenden Abschnitten formulieren wir die Regel für das Konfidenzintervall in ihrer allgemeinsten Form anhand der Anzahl der Erfolge und der Anzahl der Ausfälle in jeder Stichprobe. Ein Erfolg ist das relevante Ereignis, und ein Ausfall ist das Komplement des relevanten Ereignisses. Daher entspricht im speziellen Kontext des Tests für den Prozentsatz fehlerhafter Einheiten bei zwei Stichproben die Anzahl der „Erfolge" der Anzahl der fehlerhaften Einheiten, während die Anzahl der „Ausfälle" gleich der Anzahl der nicht fehlerhaften Einheiten ist.

Formel D2: Regeln für approximierte Konfidenzintervalle

Die allgemeine Regel für auf der Normal-Approximation basierende Konfidenzintervalle besagt, dass die Konfidenzintervalle genau sind, wenn $n_1\hat{p}_1 \geq 5$; $n_1(1-\hat{p}_1) \geq 5$; $n_2\hat{p}_2 \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$. Das heißt, das tatsächliche Konfidenzniveau des Intervalls ist gleich bzw. annähernd gleich dem Soll-Konfidenzniveau, wenn jede Stichprobe mindestens 5 Erfolge (fehlerhafte Einheiten) und 5 Ausfälle (nicht fehlerhafte Einheiten) enthält.

Die Regel wird anhand der geschätzten Anteile der Erfolge und Ausfälle und nicht mit den tatsächlichen Anteilen ausgedrückt, da in der Praxis die tatsächlichen Anteile unbekannt sind. In einer theoretischen Situation, wenn die tatsächlichen Anteile angenommen werden oder bekannt sind, kann die Regel hingegen direkt mit den tatsächlichen Anteilen ausgedrückt werden. In diesen Fällen kann direkt auswertet werden, wie sich die tatsächliche erwartete Anzahl der Erfolge und die tatsächliche erwartete Anzahl der Ausfälle, n_1p_1; n_2p_2; $n_1(1-p_1)$ und $n_2(1-p_2)$, auf die tatsächliche Überdeckungswahrscheinlichkeit des Konfidenzintervalls für die Differenz zwischen den Anteilen auswirken.

Die tatsächliche Überdeckungswahrscheinlichkeit kann ermittelt werden, indem eine große Anzahl von Stichprobenpaaren mit den Umfängen n_1 und n_2 aus den zwei Bernoulli-verteilten Grundgesamtheiten mit den Erfolgswahrscheinlichkeiten p_1 und p_2 gezogen werden. Die tatsächliche Überdeckungswahrscheinlichkeit wird dann als relative Häufigkeit der Paare von Stichproben berechnet, die Konfidenzintervalle ergeben, die die tatsächliche Differenz zwischen den zwei Anteilen enthalten. Wenn die tatsächliche Überdeckungswahrscheinlichkeit bei $n_1p_1 \geq 5$; $n_2p_2 \geq 5$; $n_1(1-p_1) \geq 5$ und $n_2(1-p_2) \geq 5$ genau, ist die Überdeckungswahrscheinlichkeit gemäß dem starken Gesetz der großen Zahlen genau, wenn $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$. Wenn also diese Regel gültig ist und das tatsächliche Konfidenzniveau und das Soll-Konfidenzniveau nahe beieinander liegen, ist zu erwarten, dass für einen großen Anteil der Paare der aus den zwei Bernoulli-verteilten Grundgesamtheiten generierten Stichproben $n_1\hat{p}_1 \geq 5$; $n_1(1-\hat{p}_1) \geq 5$; $n_2\hat{p}_2 \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$ gilt. In der folgenden Simulation wird auf diese Regel als Regel 1 Bezug genommen.

Darüber hinaus haben wir im Verlauf dieser Untersuchung in vielen Fällen festgestellt, dass bei entweder $n_1p_1 \geq 5$ und $n_2p_2 \geq 5$ oder $n_1(1-p_1) \geq 5$ und $n_2(1-p_2) \geq 5$ die simulierte Überdeckungswahrscheinlichkeit des Intervalls nahe der Sollüberdeckung liegt. Dies führte zu einer alternativen und weniger strikten Regel, die besagt, dass die approximierten Konfidenzintervalle genau sind, wenn $n_1\hat{p}_1 \geq 5$ und $n_2\hat{p}_2 \geq 5$ bzw. $n_1(1-\hat{p}_1) \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$. In der nachfolgenden Simulation wird auf diese abgewandelte Regel als Regel 2 Bezug genommen.

Simulation D1: Auswerten der Genauigkeit von approximierten Konfidenzintervallen

Wir haben Simulationen zum Auswerten der Bedingungen durchgeführt, unter denen das approximierte Konfidenzintervall für die Differenz zwischen zwei Anteilen genau ist. Dabei wurde insbesondere die Genauigkeit des Intervalls in Bezug auf die folgenden allgemeinen Regeln untersucht:

Regel 1 (ursprünglich) $n_1p_1 \geq 5$; $n_2p_2 \geq 5$; $n_1(1-p_1) \geq 5$ und $n_2(1-p_2) \geq 5$

Regel 2 (abgewandelt) $n_1\hat{p}_1 \geq 5$ und $n_2\hat{p}_2 \geq 5$ **ODER** $n_1(1-\hat{p}_1) \geq 5$ und $n_2(1-\hat{p}_2) \geq 5$

In jedem Experiment wurden 10.000 Stichprobenpaare aus Paaren von Bernoulli-verteilten Grundgesamtheiten generiert, die durch die folgenden Anteile definiert sind:

- **A-Anteile: Sowohl p_1 als auch p_2 liegen nahe 1,0 (oder nahe 0).** Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten in der Simulation wurden $p_1 = 0{,}8$ und $p_2 = 0{,}9$ verwendet.
- **B-Anteile: p_1 und p_2 liegen nahe 0,5.** Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten in der Simulation wurden $p_1 = 0{,}4$ und $p_2 = 0{,}55$ verwendet.
- **C-Anteile: p_1 liegt nahe 0,5 und p_2 nahe 1,0**. Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten wurden $p_1 = 0{,}4$ und $p_2 = 0{,}9$ verwendet.

Die oben ausgeführte Klassifikation der Anteile basiert auf der Normal-Approximation nach DeMoivre-Laplace an die Binomialverteilung, aus der die approximierten Konfidenzintervalle abgeleitet sind. Diese Normal-Approximation ist bekanntermaßen genau, wenn die Bernoulli-verteilte Stichprobe größer als 10 ist und die Erfolgswahrscheinlichkeit nahe 0,5 liegt. Wenn die Erfolgswahrscheinlichkeit nahe 0 oder 1 liegt, ist generell eine größere Bernoulli-verteilte Stichprobe erforderlich.

Die Stichprobenumfänge für beide Paare wurden auf einen einzelnen Wert von n festgelegt, wobei $n = 10, 15, 20, 30, \ldots, 100$. Wir haben die Studie auf balancierte Designs ($n_1 = n_2 = n$) beschränkt, ohne dass die Allgemeingültigkeit beeinträchtigt wird, da beide Regeln von der beobachteten Anzahl von Erfolgen und Ausfällen abhängen, die anhand des Stichprobenumfangs und des Anteils der Erfolge kontrolliert werden können.

Zum Schätzen des tatsächlichen Konfidenzniveaus des Konfidenzintervalls für die Differenz zwischen den beiden Grundgesamtheiten (als simuliertes Konfidenzniveau bezeichnet) wurde der Anteil der 10.000 Intervalle berechnet, die die tatsächliche Differenz zwischen den beiden Anteilen enthalten. Die Soll-Überdeckungswahrscheinlichkeit in jedem Experiment betrug 0,95. Darüber hinaus wurde der Prozentsatz der 10.000 Stichproben bestimmt, für den die Bedingungen gemäß zwei Regeln erfüllt wurden.

Hinweis: Für einige kleine Stichproben betrug der geschätzte Standardfehler der Differenz zwischen den Anteilen 0. Diese Stichproben wurden als „degeneriert" eingestuft und aus dem Experiment ausgeschlossen. Daher betrug die Anzahl der Stichprobenreplikationen in einigen wenigen Fällen etwas weniger als 10.000.

Die Ergebnisse werden unten in den Tabellen 1-11 aufgeführt und in Abbildung 1 grafisch veranschaulicht.

Tabelle 1 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=10 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 10$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	8,00	2,00	0,907	0,0	99,1
	p_2	0,90	9,00	1,00			
B	p_1	0,40	4,00	6,00	0,928	4,4	63,0
	p_2	0,55	5,50	4,50			
C	p_1	0,45	4,50	5,50	0,919	0,0	48,3
	p_2	0,90	9,00	1,00			

Tabelle 2 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=15 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 15$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	12,00	3,00	0,938	0,2	100,0
	p_2	0,90	13,50	1,50			
B	p_1	0,40	6,00	9,00	0,914	65,0	97,3
	p_2	0,55	8,25	6,75			
C	p_1	0,45	6,75	8,25	0,930	1,2	86,90
	p_2	0,90	13,50	1,50			

Tabelle 3 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=20 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 20$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	16,00	4,00	0,942	1,5	100,0
	p_2	0,90	18,00	2,00			
B	p_1	0,40	8,00	12,00	0,943	92,8	99,9
	p_2	0,55	11,00	9,00			
C	p_1	0,45	9,00	11,00	0,934	4,1	98,2
	p_2	0,90	18,00	2,00			

Tabelle 4 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=30 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 30$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	24,00	6,00	0,941	4,3	100,0
	p_2	0,90	27,00	3,00			
B	p_1	0,40	12,00	18,00	0,944	99,7	100,0
	p_2	0,55	16,50	13,50			

$n = 30$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
C	p_1	0,45	13,50	16,50	0,938	7,2	100,0
	p_2	0,90	27,00	3,00			

Tabelle 5 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=40 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 40$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	32,00	8,00	0,941	35,1	100,0
	p_2	0,90	36,00	4,00			
B	p_1	0,40	16,00	24,00	0,945	100,0	100,0
	p_2	0,55	22,00	18,00			
C	p_1	0,45	18,00	22,00	0,945	37,7	100,0
	p_2	0,90	36,00	4,00			

Tabelle 6 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=50 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 50$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	40,00	10,00	0,942	36,4	100,0
	p_2	0,90	45,00	5,00			
B	p_1	0,40	20,00	30,00	0,944	100,0	100,0
	p_2	0,55	27,50	22,50			
C	p_1	0,45	22,50	27,50	0,935	38,3	100,0
	p_2	0,90	45,00	5,00			

Tabelle 7 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=60 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 60$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	48,00	12,00	0,947	72,8	100,0
	p_2	0,90	54,00	6,00			
B	p_1	0,40	24,00	36,00	0,947	100,0	100,0
	p_2	0,55	33,00	27,00			
C	p_1	0,45	27,00	33,00	0,949	73,1	100,0
	p_2	0,90	54,00	6,00			

Tabelle 8 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=70 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 70$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	56,00	14,00	0,939	71,70	100,0
	p_2	0,90	63,00	7,00			
B	p_1	0,40	28,00	42,00	0,945	100,0	100,0
	p_2	0,55	38,50	31,50			
C	p_1	0,45	31,50	38,50	0,944	71,8	100,0
	p_2	0,90	63,00	7,00			

Tabelle 9 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=80 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 80$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	64,00	16,00	0,947	91,3	100,0
	p_2	0,90	72,00	8,00			
B	p_1	0,40	32,00	48,00	0,947	100,0	100,0
	p_2	0,55	44,00	36,00			

$n = 80$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
C	p_1	0,45	36,00	44,00	0,948	91,3	100,0
	p_2	0,90	72,00	8,00			

Tabelle 10 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=90 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 90$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	72,00	18,00	0,947	95,18	100,0
	p_2	0,90	81,00	9,00			
B	p_1	0,40	36,00	54,00	0,951	100,0	100,0
	p_2	0,55	49,50	40,50			
C	p_1	0,45	40,50	49,50	0,945	95,2	100,0
	p_2	0,90	81,00	9,00			

Tabelle 11 Simulierte Überdeckungswahrscheinlichkeiten und Prozentsatz der Stichproben, die Regel 1 und Regel 2 für n=100 erfüllen. Die Soll-Überdeckungswahrscheinlichkeit ist 0,95.

$n = 100$							
Kategorie		Anteil (p)	np	$n(1-p)$	Überdeckungs-wahrscheinlichkeit	% Stichproben, die Regel 1 erfüllen	% Stichproben, die Regel 2 erfüllen
A	p_1	0,80	80,00	20,00	0,952	97,7	100,0
	p_2	0,90	90,00	10,00			
B	p_1	0,40	40,00	60,00	0,945	100,0	100,0
	p_2	0,55	55,00	45,00			
C	p_1	0,45	45,00	55,00	0,948	97,7	100,0
	p_2	0,90	90,00	10,00			

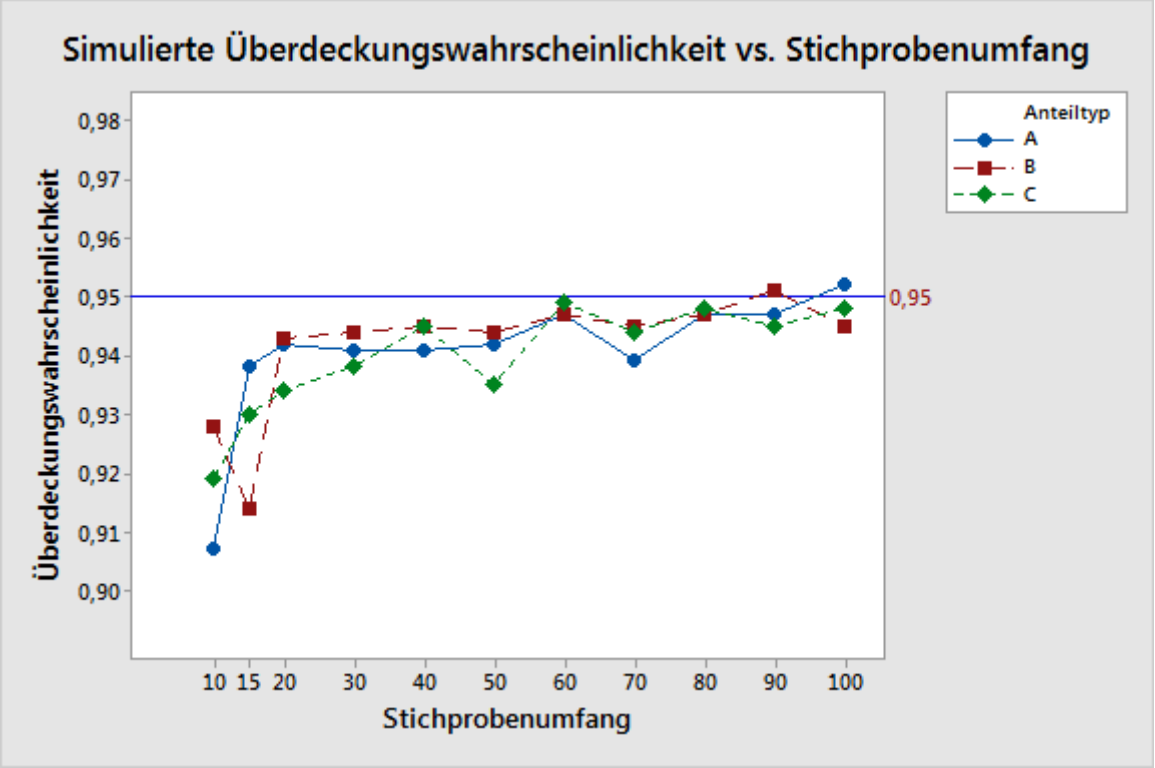

Abbildung 1 Simulierte Überdeckungswahrscheinlichkeiten im Vergleich zum Stichprobenumfang für jede Kategorie von Bernoulli-verteilten Grundgesamtheiten.

Die Ergebnisse in den Tabellen 1-11 und Abbildung 1 zeigen, dass aus Bernoulli-verteilten Grundgesamtheiten in Kategorie B generierte Stichproben (bei denen beide Anteile nahe 0,5 liegen) durchgehend simulierte Überdeckungswahrscheinlichkeiten ergeben, die stabiler sind und nahe der Sollüberdeckung von 0,95 liegen. In dieser Kategorie sind die erwartete Anzahl der Erfolge und die erwartete Anzahl der Ausfälle in beiden Grundgesamtheiten größer als in den anderen Kategorien; dies gilt selbst für kleine Stichproben.

Für die aus Paaren von Bernoulli-verteilten Grundgesamtheiten generierten Stichproben in Kategorie A (bei denen beide Anteile nahe 1,0 liegen) oder in Kategorie C (bei denen ein Anteil nahe 1,0 und der andere nahe 0 liegt) hingegen weichen die simulierten Überdeckungswahrscheinlichkeiten in den kleineren Stichproben vom Sollwert ab, außer wenn die erwartete Anzahl der Erfolge (*np*) oder die erwartete Anzahl der Ausfälle (*n(1-p)*) groß genug ist.

Betrachten Sie beispielsweise die aus den Bernoulli-verteilten Grundgesamtheiten generierten Stichproben in Kategorie A bei $n = 15$. Die erwarteten Anzahlen der Erfolge für jede Grundgesamtheit sind 12,0 und 13,5, während die erwarteten Anzahlen der Ausfälle 3,0 und 1,5 betragen. Obwohl die erwartete Anzahl der Ausfälle für beide Grundgesamtheiten kleiner als 5 ist, beträgt die simulierte Überdeckungswahrscheinlichkeit etwa 0,94. Ergebnisse wie diese veranlassten uns, Regel 2 aufzustellen, die lediglich vorschreibt, dass *entweder* die erwartete Anzahl der Erfolge *oder* die erwartete Anzahl der Ausfälle für jede Stichprobe größer oder gleich 5 sein muss.

Um umfassender beurteilen zu können, wie effektiv Regel 1 oder Regel 2 die Approximation für das Konfidenzintervall auswerten, wurde der Prozentsatz der Stichproben, die Regel 1

erfüllen und der Prozentsatz der Stichproben, die Regel 2 erfüllen im Vergleich zu den simulierten Überdeckungswahrscheinlichkeiten in den Experimenten grafisch dargestellt. Die Diagramme sind in Abbildung 2 gezeigt.

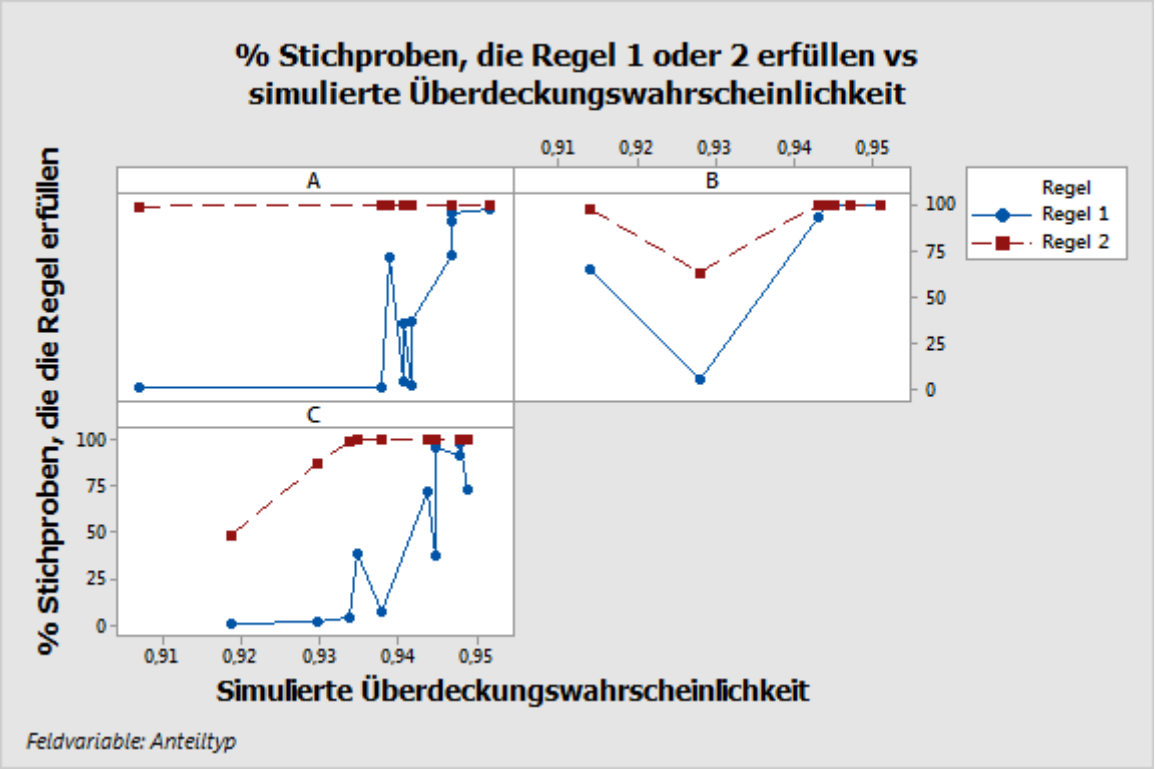

Abbildung 2 Prozentsatz der Stichproben, die Regel 1 und Regel 2 erfüllen, dargestellt im Vergleich zu den simulierten Überdeckungswahrscheinlichkeit für jede Kategorie von Bernoulli-verteilten Grundgesamtheiten.

Die Diagramme veranschaulichen Folgendes: Wenn sich die simulierten Überdeckungswahrscheinlichkeiten der Sollüberdeckung von 0,95 nähern, nähert sich der Prozentsatz der Stichproben, die die Anforderungen beider Regeln erfüllen, generell 100 % an. Für die Stichproben, die aus Bernoulli-verteilten Grundgesamtheiten in den Kategorien A und C generiert wurden, ist Regel 1 bei kleinen Stichproben stringent, wie durch den außerordentlich kleinen Prozentsatz von Stichproben gezeigt, die die Regel erfüllen; dies gilt selbst dann, wenn die simulierten Überdeckungswahrscheinlichkeiten nahe dem Sollwert liegen. Wenn z. B. $n = 20$ und die Stichproben aus den Bernoulli-verteilten Grundgesamtheiten in Kategorie A generiert wurden, beträgt die simulierte Überdeckungswahrscheinlichkeit 0,942 (siehe Tabelle 3). Der Anteil der Stichproben, die die Regel erfüllen, liegt jedoch nahe 0 (0,015) (siehe Abbildung 2). Daher ist die Regel in diesen Fällen möglicherweise übermäßig konservativ.

Regel 2 hingegen ist weniger stringent für kleine Stichproben, die aus den Bernoulli-verteilten Grundgesamtheiten in Kategorie A generiert wurden. In Tabelle 1 wird z. B. ersichtlich, dass die simulierte Überdeckungswahrscheinlichkeit bei $n = 10$ und Stichproben aus den Bernoulli-verteilten Grundgesamtheiten in Kategorie A 0,907 beträgt und 99,1 % der Stichproben die Regel erfüllen.

Fazit: Regel 1 ist bei kleinen Stichproben tendenziell übermäßig konservativ. Regel 2 ist weniger konservativ und ist bei kleinen Stichproben möglicherweise vorzuziehen. Regel 1 ist jedoch weithin bekannt und akzeptiert. Obwohl Regel 2 vielversprechend ist, kann sie in einigen Fällen zu liberal sein, wie oben gezeigt. Möglicherweise können beide Regeln kombiniert werden, um die Stärken beider Regeln auszunutzen. Ein solcher Ansatz erfordert jedoch weitere Untersuchungen, ehe er verfolgt werden kann.

Anhang E: Vergleich der tatsächlichen Trennschärfe und der theoretischen Trennschärfe

Simulation E1: Schätzen der tatsächlichen Trennschärfe mit Fishers exaktem Test

Wir haben eine Simulation zum Vergleichen der geschätzten tatsächlichen Trennschärfen (die als simulierte Trennschärfen bezeichnet werden) von Fishers exaktem Test mit den theoretischen Trennschärfen auf Grundlage der Trennschärfefunktion des Tests auf Normal-Approximation (die als approximierte Trennschärfen bezeichnet werden) konzipiert. In jedem Experiment wurden 10.000 Paare von Stichproben aus Paaren von Bernoulli-verteilten Grundgesamtheiten generiert. Für jedes Paar von Stichproben wurden die Anteile so gewählt, dass die Differenz zwischen den Anteilen $p_1 - p_2 = -0{,}20$ betrug.

- **A-Anteile: Sowohl p_1 als auch p_2 liegen nahe 1,0 (oder nahe 0).** Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten in der Simulation wurden $p_1 = 0{,}70$ und $p_2 = 0{,}90$ verwendet.
- **B-Anteile: Sowohl p_1 als auch p_2 liegen nahe 0,5.** Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten in der Simulation wurden $p_1 = 0{,}40$ und $p_2 = 0{,}60$ verwendet.
- **C-Anteile: p_1 liegt nahe 0,5 und p_2 nahe 1,0.** Zum Darstellen dieses Paares von Bernoulli-verteilten Grundgesamtheiten in der Simulation wurden $p_1 = 0{,}55$ und $p_2 = 0{,}75$ verwendet.

Die Stichprobenumfänge für beide Paare wurden auf einen einzelnen Wert von n festgelegt, wobei $n = 10, 15, 20, 30, \ldots, 100$. Die Studie wurde auf balancierte Designs ($n_1 = n_2 = n$) beschränkt, weil normalerweise anzunehmen ist, dass beide Stichproben den gleichen Umfang aufweisen. Es wurde ein gemeinsamer Stichprobenumfang berechnet, der erforderlich ist, um eine Differenz mit praktischen Konsequenzen mit einer bestimmten Trennschärfe erkennen zu können.

Zum Schätzen der tatsächlichen Trennschärfe für Fishers exakten Test basierend auf den Ergebnissen der einzelnen Simulationen wurde der Anteil der 10.000 Stichprobenpaare berechnet, für die der beidseitige Test beim Soll-Signifikanzniveau $\alpha = 0{,}05$ signifikant war. Anschließend wurden zu Vergleichszwecken die entsprechenden theoretischen Trennschärfen auf der Grundlage des Tests auf Normal-Approximation berechnet. Die Ergebnisse werden unten in Tabelle 12 aufgeführt.

Tabelle 12 Simulierte Trennschärfen von Fishers exaktem Test im Vergleich mit den approximierten Trennschärfen für die drei Kategorien von Bernoulli-verteilten Grundgesamtheiten. Das Soll-Signifikanzniveau ist $\alpha = 0{,}05$.

n	A-Anteile		B-Anteile		C-Anteile	
	$p_1 = 0{,}70$ $p_2 = 0{,}90$		$p_1 = 0{,}40$ $p_2 = 0{,}60$		$p_1 = 0{,}55$ $p_2 = 0{,}75$	
	Simulierte Trennschärfe	Approx. Trennschärfe	Simulierte Trennschärfe	Approx. Trennschärfe	Simulierte Trennschärfe	Approx. Trennschärfe
10	0,063	0,193	0,056	0,140	0,056	0,149
15	0,151	0,271	0,097	0,190	0,101	0,204
20	0,244	0,348	0,146	0,240	0,183	0,259
30	0,370	0,490	0,256	0,338	0,272	0,366
40	0,534	0,612	0,371	0,431	0,381	0,466
50	0,641	0,711	0,477	0,516	0,491	0,556
60	0,726	0,789	0,536	0,593	0,560	0,635
70	0,814	0,849	0,610	0,661	0,649	0,703
80	0,870	0,893	0,660	0,720	0,716	0,760
90	0,907	0,925	0,716	0,770	0,772	0,808
100	0,939	0,948	0,792	0,812	0,812	0,848

Die Ergebnisse in Tabelle 12 zeigen, dass die approximierte Trennschärfe für alle drei Kategorien von Bernoulli-verteilten Grundgesamtheiten (A, B und C) tendenziell höher als die simulierte Trennschärfe ist. Für die Anteile in Kategorie A beträgt der tatsächliche Stichprobenumfang, der zum Erkennen einer absoluten Differenz von -0,20 mit einer approximierten Trennschärfe von 0,91 erforderlich ist, etwa 90. Im Gegensatz dazu beträgt der entsprechende Stichprobenumfang basierend auf der approximierten theoretischen Trennschärfefunktion etwa 85. Daher ist die geschätzte Trennschärfe gemäß der approximierten Trennschärfefunktion im Allgemeinen etwas kleiner als der tatsächliche Stichprobenumfang, der zum Erzielen einer bestimmten Trennschärfe erforderlich ist.

Diese Beziehung wird noch deutlicher, wenn die Ergebnisse als Trennschärfekurven wie in der nachfolgenden Abbildung 3 dargestellt werden.

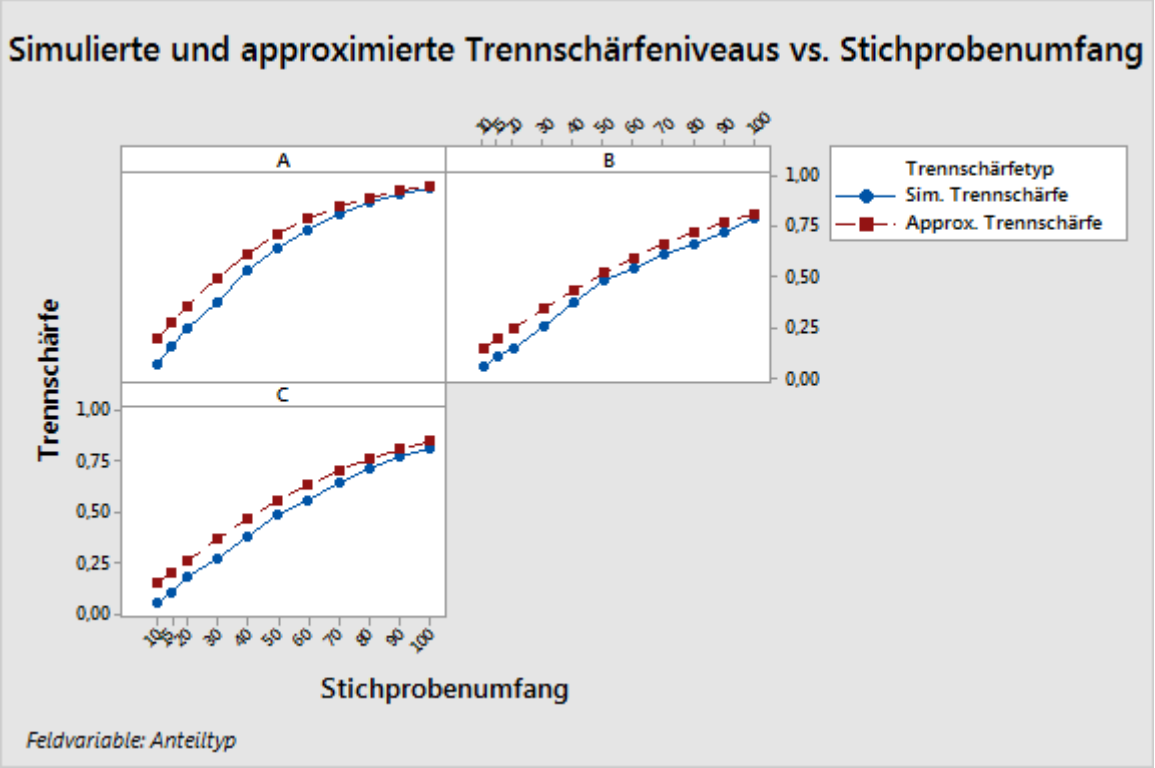

Abbildung 3 Diagramme der simulierten und approximierten Trennschärfen des beidseitigen Tests zum Vergleichen von zwei Anteilen. Die Trennschärfen werden im Vergleich zum Stichprobenumfang in separaten Feldern für jede Kategorie von Bernoulli-verteilten Grundgesamtheiten dargestellt.

Beachten Sie Folgendes: Obwohl die Kurven der simulierten Trennschärfe für alle drei Kategorien von Bernoulli-verteilten Grundgesamtheiten (A, B und C) niedriger als die Kurven der approximierten Trennschärfe sind, hängt die Größe der Differenz zwischen den Kurven von den tatsächlichen Anteilen der Bernoulli-verteilten Grundgesamtheiten ab, aus denen die Stichproben gezogen wurden. Wenn die beiden Anteile beispielsweise nahe 0,5 (Kategorie B) liegen, liegen die beiden Trennschärfen durchgehend dicht beieinander. Der Unterschied zwischen den beiden Trennschärfekurven wird jedoch deutlicher für die Anteile in kleinen Stichproben aus den Grundgesamtheiten der Kategorien A und C.

Diese Ergebnisse zeigen, dass die theoretische Trennschärfefunktion des Tests auf Normal-Approximation und die simulierte Trennschärfefunktion von Fishers exaktem Test im Allgemeinen annähernd gleich sind. Daher nutzt der Assistent die theoretische Trennschärfefunktion des Tests auf Normal-Approximation, um die Stichprobenumfänge zu schätzen, bevor Fishers exakter Test ausgeführt wird. Die mit der approximierten Trennschärfefunktion berechneten Stichprobenumfänge können jedoch u. U. etwas kleiner als die Stichprobenumfänge sein, die tatsächlich erforderlich sind, um eine Differenz zwischen den beiden Anteilen (Prozentsatz fehlerhafter Einheiten) mit einer bestimmten Trennschärfe erkennen zu können.