# Einheit 4

# Statistische Tests

STUDIENZIELE

Nach Abschluss dieser Einheit werden Sie Folgendes gelernt haben...

- den allgemeinen Rahmen und die Interpretation der Ergebnisse von Hypothesentests.

- wie man einige gängige nicht-parametrische Hypothesentests durchführt und interpretiert.

- wie man einige gängige parametrische Ein- und Zweistichprobentests durchführt und interpretiert.

- wie man p-Werte definiert und interpretiert.

- wie man Konfidenzintervalle konstruiert und interpretiert.

- wie man zwei verschiedene Fehlermaße im Zusammenhang mit multiplen Hypothesentests kontrolliert.

# 4. Statistische Tests

### Einführung

Statistische Tests sind das Kernstück der statistischen Inferenz. Statistische Tests, auch Hypothesentest genannt, umfassen strukturierte Paradigmen, die Ergebnisse aus der Wahrscheinlichkeitstheorie verwenden, um die Evidenz oder das Fehlen von Evidenz zu quantifizieren, die eine beobachtete Stichprobe für eine Behauptung liefert. Mit statistischen Tests kann eine Vielzahl von Behauptungen untersucht werden, weshalb sie in fast allen Branchen eingesetzt werden. Zu den Anwendungen gehören Marketing, Finanzen, Medizin und Psychologie.

In dieser Einheit werden wir die gebräuchlichsten statistischen Tests besprechen, wir werden lernen, wie man sie durchführt, ihre Ergebnisse interpretiert und, was am wichtigsten ist, wir werden ihre Grenzen und die Annahmen verstehen, unter denen diese Tests angewendet werden können. Die Tests, die wir hier ausgewählt haben, sind weitverbreitet und erfordern außerdem nur die grundlegenden mathematischen Kenntnisse, die in diesem Kurs erwartet werden.

Angenommen, die Prävalenz von Brustkrebs bei Frauen im Alter von 54-65 Jahren beträgt zwei Prozent. Wir interessieren uns für die Prävalenz von Brustkrebs bei Frauen im Alter von 54-65 Jahren, bei deren Müttern Brustkrebs diagnostiziert wurde. Zu diesem Zweck nehmen wir eine Stichprobe von 10.000 Frauen, die diese Kriterien erfüllen. Von dieser Stichprobe sind 400 an Brustkrebs erkrankt. Die Prävalenz von Brustkrebs in der Stichprobe beträgt also 400/10.000 = 0 . 04 = 4 %. Wir wollen prüfen, ob sich die tatsächliche Prävalenz von Brustkrebs bei Frauen, deren Mütter Brustkrebs hatten, von der Prävalenz bei allen Frauen im Alter von 54-65 Jahren unterscheidet. Der statistische Test, mit dem wir diese Behauptungen überprüfen können, ist Gegenstand von Abschnitt 4.1. In diesem Abschnitt wird die Idee des Hypothesentests im Hinblick auf Behauptungen über einen unbekannten Parameter für eine einzelne Population von Interesse vorgestellt.

Viele der in dieser Einheit behandelten Hypothesentests sind parametrische Tests, d. h. statistische Tests über Parameter einer Verteilung der zugrundeliegenden Grundgesamtheit(en). In Abschnitt 4.2 werden wir einige gängige parametrische Tests besprechen. Bei vielen statistischen Modellen wird davon ausgegangen, dass die zugrundeliegende Grundgesamtheit eine Gauß-Verteilung aufweist. Zu diesem Zweck stellen wir den Kolmogorov-Smirnov-Normalitätstest vor, einen nicht-parametrischen Test, der bestimmt, wie gut die Stichprobendaten einer vorgegebenen Gauß-Verteilung entsprechen. Ein weiterer Anpassungsgütetest verwendet die χ2-Verteilung, um zu beurteilen, wie gut die Zähldaten (Häufigkeitsverteilung) einer vorgeschlagenen kategorialen Verteilung entsprechen. Schließlich wird ein weiterer χ2-Test erörtert, mit dessen Hilfe beurteilt werden kann, ob zwei kategoriale Variablen voneinander abhängig sind.

In den Abschnitten 4.1 und 4.2 werden auf hohem Niveau Hypothesentests behandelt, die eine Behauptung über Daten aufstellen, die aus einer einzigen Grundgesamtheit generiert wurden, so genannte "Einstichprobentests". In Abschnitt 4.3 stellen wir Hypothesentests vor, die Parameter aus zwei Populationen vergleichen. Beispielsweise lernen wir, wie man bei zwei Stichproben, jeweils eine aus zwei unabhängigen Stichproben, einen Hypothesentest durchführt, der die wahren Mittelwerte der Populationen vergleicht. Als Beispiel aus dem Bereich der Benutzerschnittstellen (UI) und der Benutzererfahrung (UX), angewandt auf ein Geschäftsumfeld, betrachten wir das folgende Beispiel.

Die Click-Through-Rate (CTR) ist definiert als der Anteil der Benutzer, die auf einen Call-to-Action-Link auf einer Webseite klicken, im Verhältnis zur Gesamtzahl der Benutzer, die die Seite angesehen haben. In dem Bemühen, die CTR einer bestimmten Seite zu erhöhen, beschließt ein Webdesigner, ein neues Design für die Call-to-Action-Schaltfläche zu testen. Die aktuelle Hintergrundfarbe der Schaltfläche ist hellblau. Nennen wir das aktuelle Design A. Bei dem vorgeschlagenen Design der Call-to-Action-Schaltfläche soll die Hintergrundfarbe hellgrün sein. Dieses vorgeschlagene Design wird mit B bezeichnet. Um das Experiment nicht durch andere Variablen zu stören, wird alles auf der Webseite, außer dem Hintergrund des Call-to-Action-Buttons, unverändert gelassen. Die Seite wird nun jedem Besucher nach dem Zufallsprinzip eine der beiden Varianten A oder B anbieten. Mit anderen Worten: Etwa 50 % der Besucher sehen Design A, während die anderen Design B sehen. Nach einiger Zeit wird die CTR für alle Besucher berechnet, die jedes der Designs gesehen haben; nennen wir diese Anteile πA und πB. Diese beiden Werte dienen als Schätzwerte für den wahren Anteil der CTRs der jeweiligen Designs πA und πB. Wir wollen herausfinden, ob die gesammelten Daten statistisch signifikante Beweise für die Behauptung liefern, dass πA < πB ist. Mit anderen Worten, sollten wir erwarten, dass das neue Design in Bezug auf die CTR besser abschneidet? In Abschnitt 4.3 werden wir dieses Szenario mit Hilfe von A/B-Tests bewerten, einem Marketingbegriff, der "Hypothesentests mit zwei Stichproben" bedeutet. Im Allgemeinen hilft uns dieser Abschnitt bei der Anwendung von Hypothesentests in Bezug auf Behauptungen darüber, wie analoge unbekannte Parameter von Interesse miteinander in Beziehung stehen.

In Abschnitt 4.4 setzen wir unsere Erläuterungen zu Hypothesentests fort. Wir erörtern im Detail den Kompromiss zwischen Fehlern vom Typ I und vom Typ II und definieren die Teststärke (Power) eines Tests, eine häufig übersehene, aber wichtige Überlegung. In diesem Abschnitt führen wir auch die Idee der p-Werte ein, ein wichtiger, aber häufig falsch interpretierter Wert. Wir lernen, wie man p-Werte korrekt berechnet und interpretiert. Außerdem lernen wir in diesem Abschnitt die Intervallschätzung als überlegene Alternative zu Punktschätzungen kennen. Während eine Punktschätzung einen einzigen Wert als Schätzung für einen unbekannten Parameter von Interesse liefert, ist eine Intervallschätzung ein Bereich von Werten, der die mit dem Schätzungsprozess verbundene Unsicherheit berücksichtigt. Wir erörtern auch, wie Konfidenzintervalle verwendet werden können, um Hypothesen zu bewerten, die genau denjenigen entsprechen, die wir in den Abschnitten 4.1 und 4.2 behandelt haben.

Im letzten Abschnitt, Abschnitt 4.5, werden multiple Hypothesentests behandelt. Nehmen wir an, wir wollen auf einem Feld Tomaten anbauen und haben vier verschiedene Pläne für die Bodenhydration. Wir möchten wissen, ob sich der durchschnittliche Ertrag über einen bestimmten Zeitraum, gemessen in Kilogramm, auf einem statistisch signifikanten Niveau zwischen den verschiedenen Hydrationsplänen unterscheidet. In diesem Szenario haben wir viele Hypothesen. Bezeichnen wir μ1, μ2, μ3, μ4 als die wahren Durchschnittserträge aus den vier verschiedenen Hydrationsplänen. Der Status quo, die Nullhypothese, würde besagen, dass alle Mittelwerte gleich sind: kein Effekt. Die Nullhypothese enthält die Aussage H0:μ1 = μ2, μ1 = μ3, μ1 = μ4, μ2 = μ3, μ2 = μ4, und μ3 = μ4. Die Alternativhypothese, die nach einem Effekt sucht, kann durch Ersetzen von = durch ≠ in jeder dieser Aussagen gefunden werden. Wenn wir eine Fehlerquote von fünf Prozent für die Ablehnung einer wahren Nullhypothese akzeptieren, dann ist die Wahrscheinlichkeit, mindestens eine wahre Nullhypothese abzulehnen, viel höher, nämlich mehr als 20 Prozent! In diesem Abschnitt lernen wir zwei Methoden kennen, um Fehler dieser Art zu kontrollieren.

4.1 Hypothesentests und Teststatistiken

Der Hypothesentest ist ein vierteiliges Paradigma zur Bewertung der statistischen Signifikanz beobachteter Daten im Hinblick auf ein Paar konkurrierender Hypothesen (Behauptungen). Der erste Teil dieses Paradigmas besteht darin, die Hypothesen zu bestimmen. Der Status quo, d. h. die Aussage, dass es keine Wirkung oder keine Veränderung gibt, wird als Nullhypothese zusammengefasst und mit H0 bezeichnet. Die Testhypothese, also die Aussage über das Vorhandensein eines Effekts oder einer Veränderung, wird als Alternativhypothese zusammengefasst und mit H1 bezeichnet. In dem Beispiel aus der Einleitung über die Prävalenz von Brustkrebs besagt die Nullhypothese, dass die Prävalenz von Brustkrebs bei Frauen im Alter von 54-65 Jahren, deren Mütter an Brustkrebs erkrankt sind, die gleiche ist wie bei allen Frauen im Alter von 54-65 Jahren (2 %). Die Alternativhypothese besagt, dass die Prävalenz unterschiedlich ist. Mit π bezeichnen wir die Prävalenz von Brustkrebs in unserer interessierenden Population. Wir können die beiden Hypothesen wie folgt formulieren

H0:π = 0 . 02;H1:π ≠ 0 . 02

Im Allgemeinen ist die Prüfung von Behauptungen über einen Bevölkerungsanteil gegen einen bekannten Wert π0 H0:π = π0;H1:π ≠ π0

**Fehler 1. Art**

Dieser Fehler tritt auf durch Ablehnung einer wahren Nullhypothese.

In unserem Fall ist π0 = 0 . 02. Wenn man eine Stichprobe erhält und die Stichprobenprävalenz (Anteil) berechnet, ist es aufgrund von Zufälligkeiten unwahrscheinlich, dass man genau 0,02 erhält, selbst wenn die wahre Prävalenz in der interessierenden Population tatsächlich 0,02 beträgt. Die Stichprobenprävalenz kann 0,03, 0,021, 0,04 oder theoretisch jede beliebige Zahl zwischen 0 und 1 betragen. Wenn die wahre Prävalenz tatsächlich 0,02 beträgt, ist es natürlich wahrscheinlicher, dass der Stichprobenanteil nahe bei 0,02 liegt als weit davon entfernt. Je weiter die Stichprobenprävalenz bei einem gegebenen Stichprobenumfang von 0,02 entfernt ist, desto signifikanter ist der statistische Beweis gegen unsere Annahme H0:π = 0 . 02. Mit anderen Worten: Bei einem gegebenen Stichprobenumfang ist die Differenz zwischen dem Stichprobenanteil und dem angenommenen Anteil, π - π0, ein Indikator für die Evidenz gegen H0:π = π0. Je größer diese Differenz ist, desto wahrscheinlicher ist unsere Annahme (die Nullhypothese) falsch. Doch wie hoch ist der Grenzwert? Damit kommen wir zum zweiten Teil des Paradigmas der Hypothesentests: dem Signifikanzniveau.

**Fehler 2.Art**

Dieser Fehler entsteht, wenn eine falsche Nullhypothese nicht abgelehnt werden kann.

Das Signifikanzniveau eines Hypothesentests ist die höchste Wahrscheinlichkeit, dass die Nullhypothese fälschlicherweise abgelehnt wird. Das heißt, das höchste Risiko, das wir bereit sind einzugehen, wenn wir uns für die Ablehnung der Nullhypothese entscheiden. Formal ist das Signifikanzniveau α die Wahrscheinlichkeit der Ablehnung einer wahren Nullhypothese α = ℙ H0ablehnen H0 ist wahr.

Die Entscheidung, eine wahre Nullhypothese abzulehnen, ist sicherlich ein Fehler. Bei statistischen Tests wird dieser Fehler als Fehler 1. Art bezeichnet. Der Grenzwert für die Differenz zwischen dem Stichprobenanteil und dem angenommenen Anteil, der zu einer Entscheidung für die Ablehnung der Nullhypothese führt, ergibt sich aus dem Signifikanzniveau α. Damit wird die Wahrscheinlichkeit eines Fehlers **1. Art kontrolliert**. Bei der Durchführung eines Hypothesentests ist auch ein anderer Fehler möglich: das Nichtverwerfen einer falschen Nullhypothese. Dieser Fehler ist als Fehler **2. Art bekannt**, und die damit verbundene Wahrscheinlichkeit wird mit β angegeben. Die folgende Tabelle fasst diese beiden Fehlerarten zusammen. Das Schwierige daran ist, dass α und β in einem umgekehrten Verhältnis zueinander stehen. Wird der eine Wert niedrig angesetzt, ist der andere tendenziell hoch. Wir werden den Kompromiss zwischen α und β in einem späteren Abschnitt erörtern.

Tabelle 15: Fehler 1. Art und 2. Art (Wahrscheinlichkeiten)

|  |  |  |
| --- | --- | --- |
|  | Wahre H0 | Falsche H0 |
| H0 ablehnen | Fehler 1. Art (α) | Kein Fehler |
| H nicht ablehnen0 | Kein Fehler | Fehler 2. Art (β) |

**Teststatistik**

Diese Zufallsvariable standardisiert die Größe, die eine Abweichung von der Nullhypothese misst.

In der Regel wird eine solche Zufallsvariable so gewählt, dass ihre Verteilung zumindest approximativ bekannt ist.

Statt mit der beobachteten Differenz π - π0 zu arbeiten und zu versuchen, einen Grenzwert zu finden, arbeiten wir in der Praxis mit einer standardisierten Größe, für die wir die (approximative) Verteilung kennen. Dieser standardisierte Wert wird als **Teststatistik bezeichnet**. Der zentrale Grenzwertsatz besagt, dass bei einem großen Stichprobenumfang die Größe

xxx

(bei π = π0) näherungsweise einer Gaußschen Standardverteilung folgt (Hogg et al., 2019), U H0 N 0, 1 . Dabei ist π = 1nΣi = 1n Xi der Schätzer des Stichprobenanteils, wobei Xi = 1 ist, wenn die i-te Person in unserer Stichprobe an Brustkrebs erkrankt ist, und andernfalls Null. Anstatt einen Grenzwert für die Differenz π - π0 zu finden, finden wir einen Grenzwert für u, so dass die Ablehnung von H0, wenn H0 wahr ist, α ist. In unserem Fall neigen wir dazu, H0 abzulehnen, wenn U groß ist. Unser Grenzwert(e) entspricht uc, wobei

**Ablehnungsbereich**

Dieser Menge von Werten der Teststatistik ist mit einer Gesamtwahrscheinlichkeit gleich dem Signifikanzniveau verbunden und gibt die Abweichung von der Nullhypothese in Richtung der Alternativhypothese(n) an.

Xxx

Anders ausgedrückt: ℙ U < uc H0 = α/2 oder Φ uc = α/2, wobei Φ die CDF der Gauß-Verteilung ist. Sobald dieser Grenzwert bestimmt ist, ist der **Ablehnungsbereich** eine Menge von Werten u, so dass u > uc ist. Die erste Abbildung unten zeigt die Gauß-Verteilung zusammen mit dem Ablehnungsbereich (in orange). Die zweite Abbildung unten zeigt verschiedene Grenzwerte und entsprechende Ablehnungsbereiche für verschiedene Werte von α.

Abbildung 35: Zweiseitige Ablehnungsbereiche einer Gauß-Teststatistik

Abbildung 36: Zweiseitige Ablehnungsbereiche einer Gauß-Teststatistik mit verschiedenen Signifikanzniveaus

Je kleiner das Signifikanzniveau α ist, desto kleiner ist erwartungsgemäß der Ablehnungsbereich. Es ist sehr wichtig zu beachten, dass die Hypothesen sowie das Signifikanzniveau (und damit der Ablehnungsbereich) festgelegt werden, bevor Daten erhoben werden. Andernfalls könnten die Daten die Wahl der Hypothesen und des Signifikanzniveaus beeinflussen, um eine Entscheidung gegenüber einer anderen zu begünstigen. Mit anderen Worten: Die Wahl der Hypothesen und/oder des Signifikanzniveaus nach der Analyse der beobachteten Daten macht die Integrität des statistischen Tests völlig zunichte. In der Praxis ist es nicht zu rechtfertigen, Daten zu sammeln/zu analysieren, bevor diese Entscheidungen getroffen werden!

In unserem Fall wurde die Wahl der Hypothesen bereits getroffen; wir wählen ein Signifikanzniveau von einem Prozent, d. h. α = 0 . 01. Der dritte Teil des statistischen Tests besteht darin, Daten zu erfassen und die beobachteten Werte der interessierenden Größen zu berechnen. Letztendlich den beobachteten Wert der Teststatistik, uobs.

Angenommen, in einer Stichprobe von n = 10 000 Frauen im Alter von 54-65 Jahren, deren Mütter an Brustkrebs erkrankt sind, wird festgestellt, dass 400 von ihnen Brustkrebs haben. Die Stichprobenprävalenz ist π = 400/10.000 = 0 . 04. Der beobachtete Wert der Teststatistik beträgt

xxx

Aus der obigen Abbildung, die verschiedene zweiseitige Ablehnungsbereiche zeigt, geht hervor, dass dieser beobachtete Wert in dem Ablehnungsbereich liegt, der α = 0 . 01entspricht. Damit kommen wir zum vierten und letzten Teil: der Entscheidung. Da der beobachtete Wert der Teststatistik im Ablehnungsbereich liegt, lehnen wir die Nullhypothese mit einem Signifikanzniveau von fünf Prozent ab. Das bedeutet, dass die Daten bei dem fünfprozentigen Signifikanzniveau den Beweis (Evidenz) liefern, dass die Nullhypothese falsch ist.

**Zweiseitiger Test**

Bei diesem statistischen Test wird die

Die Nullhypothese besagt, dass der wahre Parameter

anders als behauptet, was zu einem Ablehnungsbereich führt, der sich aus Intervallen zusammensetzt, die sich in entgegengesetzte Richtungen erstrecken.

Fassen wir die vier Teile für dieses Beispiel zusammen:

1. Hypothesen aufstellen: H0:π = π0 = 0,02 gegenüber H1:π ≠ π0 = 0,02

2. α = 0,01 und die Teststatistik

Xxx

festlegen. Der Ablehnungsbereich beträgt u > uc = 2,576

**Einseitiger Test**

Bei diesem statistischen Test gibt die Alternativhypothese einen Richtungsunterschied an (weniger als oder mehr als).

3. Den beobachteten Wert der Teststatistik anhand der Stichprobendaten: uobs = 14,29 berechnen

4. Der beobachtete Wert uobs fällt in den Ablehnungsbereich, so dass wir H0 bei dem Signifikanzniveau von fünf Prozent ablehnen.

Der statistische Test, den wir für das Szenario Brustkrebs durchgeführt haben, war ein **zweiseitiger Test**. Denn die Alternativhypothese π ≠ π0 ist gleichbedeutend mit der Aussage π < π0 oder π > π0. Anders ausgedrückt: Der Ablehnungsbereich u > uc setzt sich aus den beiden Bereichen u < uc und u > uc zusammen. Manchmal beinhaltet der statistische Test eine gewisse vorherige Vorstellung über die Richtung des Effekts. In diesem Fall kann ein Forscher einen **einseitigen Test** verwenden. Es gibt zwei mögliche einseitige Tests: einen linksseitigen Test, bei dem die Alternativhypothese die Form θ < θ0 annimmt, und einen rechtsseitigen Test, bei dem H1:θ > θ0. Wie wir bereits erwähnt haben, muss die Wahl der Hypothesen getroffen werden, ohne dass die Daten erfasst oder analysiert werden. Der Ablehnungsbereich eines einseitigen Tests ist ein einziges Intervall von Werten, das sich nur in eine Richtung erstreckt. Der Ablehnungsbereich eines linksseitigen Tests erstreckt sich nach links und der eines rechtsseitigen Tests nach rechts. Die folgende Abbildung zeigt die Ablehnungsbereiche der drei Arten von Alternativhypothesen zusammen mit dem Signifikanzniveau und dem/den Grenzwert(en).

#### Beispiel 4.1.1

X1, …,Xn seien unabhängig mit unbekanntem Mittelwert μ und bekannter Standardabweichung σ. Wenn n groß ist, versichert uns der zentrale Grenzwertsatz, dass die Teststatistik, die definiert ist durch

xxx

näherungsweise der Gaußschen Standardverteilung folgt, d. h. Uapprox.N 0, 1 . Ermitteln Sie unter Verwendung von α = 0,05 den Ablehnungsbereich für jede der folgenden Alternativhypothesen unter der Annahme, dass H0:μ = μ0 wahr ist.

1. H1:μ > μ0

2. H1:μ < μ0

3. H1:μ ≠ μ0

#### Lösung

zα bezeichne das 1 - α 100-Quantil der Gauß-Verteilung. Mit anderen Worten: Φ zα = 1 - α oder äquivalent dazu zα = Φ-1 1 - α . Zunächst muss der Grenzwert uc die Bedingung ℙ U > uc = α = 0 . 05 erfüllen, also uc = z0 . 05 = Φ-1 0 . 95 = 1 . 645. Der Ablehnungsbereich lautet 1 . 645, ∞ . Der Grenzwert uc muss ℙ U < uc = α = 0 . 05 erfüllen. Durch Symmetrie ergibt sich uc = - z0 . 05 = - 1 . 645. Der Ablehnungsbereich lautet - ∞ , - 1 . 645 . Es handelt sich um einen zweiseitigen Test. Die Grenzwerte (Cutoff-Werte) uc müssen ℙ U > uc = α = 0 . 05 erfüllen. Mit anderen Worten: ℙ U < ucL = α/2 = 0 . 025 also ucL = - z0 . 025 = - Φ-1 0 . 975 = - 1 . 96. Analog dazu ist ucR = 1 . 96. Der Ablehnungsbereich lautet also - ∞ , - 1 . 96 ∪ 1 . 96, ∞ . Die folgende Abbildung zeigt die drei Ablehnungsbereiche aus dem vorherigen Beispiel.

Abbildung 37: Ablehnungsbereiche für eine Gauß-Teststatistik aus Beispiel 4.1.1

#### Beispiel 4.1.2

X1, …,Xn seien unabhängig mit unbekanntem Mittelwert μ und unbekannter Standardabweichung σ.

Die Teststatistik, die definiert ist durch

xxx,

folgt der Student-t-Verteilung mit n - 1 Freiheitsgraden (Hogg et al., 2019), d. h. U T n - 1 , vorausgesetzt, dass entweder n groß genug ist, um den zentralen Grenzwertsatz anzuwenden, oder die zugrundeliegende Verteilung annähernd gaußförmig ist. Dabei ist S2 die Stichprobenvarianz (der unverzerrte Schätzer der Varianz). Ermitteln Sie unter Verwendung von α = 0,05 und eines Stichprobenumfangs von n = 10 den Ablehnungsbereich für jede der folgenden Alternativhypothesen unter der Annahme, dass H0:μ = μ0 wahr ist.

- H1:μ > μ0

- H1:μ < μ0

- H1:μ ≠ μ0

#### Lösung

tn - 1, α bezeichne das 1 - α 100 Quantil der t-Verteilung mit n - 1 Freiheitsgraden. Mit anderen Worten: ℙ T n - 1 < tn - 1, α = 1 - α. Bei dieser Aufgabenstellung haben wir n = 10, also ist die Verteilung, mit der wir arbeiten werden, T 9 .

1. Der Grenzwert uc muss die Bedingung ℙ U > uc = α = 0 . 05 erfüllen, also uc = t9, 0 . 05 = 1 . 833. Der Ablehnungsbereich lautet 1 . 833, ∞ .

2. Der Grenzwert uc muss ℙ U < uc = α = 0. 05 erfüllen. Durch Symmetrie ergibt sich uc = - t9, 0 . 05 = - 1 . 833. Der Ablehnungsbereich lautet - ∞ , - 1 . 833 .

3. Es handelt sich um einen zweiseitigen Test. Die Grenzwerte uc müssen ℙ U > uc = α = 0. 05 erfüllen. Mit anderen Worten: ℙ U < ucL = α/2 = 0 . 025 also ucL = - t9, 0 . 025 = - 2 . 262. Analog dazu beträgt ucR = 2 . 262. Der Ablehnungsbereich lautet also - ∞ , - 2 . 262 ∪ 2 . 262, ∞ . Die folgende Abbildung zeigt die Ablehnungsbereiche aus Beispiel 4.1.2.

Abbildung 38: Ablehnungsbereiche für eine T(9)-Teststatistik aus Beispiel 4.1.2

Bisher haben wir drei Teststatistiken zur Durchführung von Hypothesentests gesehen. Die erste wurde für einen statistischen Test verwendet, bei dem es um π, den wahren Anteil einer interessierenden Grundgesamtheit, geht. Die beiden anderen wurden für statistische Tests verwendet, bei denen es um μ, den wahren Mittelwert einer interessierenden Grundgesamtheit, geht. Untersuchen Sie die nachstehende Tabelle, in der diese drei Teststatistiken zusammengefasst sind, und wann es angebracht ist, sie zu verwenden.

Tabelle 16: Übliche Teststatistiken für interessierende Parameter

|  |  |  |
| --- | --- | --- |
| Parameter | Teststatistik | Annahmen |
| π (wahrer Anteil) | xxx | np > 10 und n 1 - p > 10 |
| μ (wahrer Mittelwert) | xxx | σ ist bekannt, und entweder ist n ausreichend groß für CLT, oder die die zugrundeliegende Verteilung ist gaußförmig. |
| μ (wahrer Mittelwert) | xxx | S2 ist die Stichprobenvarianz und entweder ist n ausreichend groß für CLT oder die zugrunde liegende Verteilung ist gaußförmig. |

#### Beispiel 4.1.3

Eine Fabrik wird beschuldigt, Industrieabfälle unsachgemäß in Wasserreservoirs zu leiten, die von örtlichen Landwirten genutzt werden. Es wird vermutet, dass die Chemikalien in diesen Industrieabfällen Geburtsfehler verursachen, darunter auch ein niedriges Geburtsgewicht. Ein unabhängiger Rat wird damit beauftragt, die Behauptung zu prüfen, dass das Durchschnittsgewicht der Neugeborenen in dieser Stadt unter dem nationalen Durchschnitt von 3480 Gramm liegt. Das Signifikanzniveau wird auf α = 0 . 01 festgelegt. Die Geburtsgewichte von zwölf Neugeborenen werden in dem einzigen Krankenhaus der Stadt stichprobenartig ermittelt. Der Stichprobendurchschnitt dieser Gewichte beträgt 3250 Gramm mit einer Standardabweichung von 250 Gramm. Verwenden Sie diese Daten zusammen mit t11, 0 . 01 = - 2 . 718, um den entsprechenden statistischen Test durchzuführen. Nehmen Sie an, dass die Gewichte näherungsweise einer Gauß-Verteilung folgen.

#### Lösung

μ bezeichne das wahre durchschnittliche Geburtsgewicht der Babys in dieser Stadt. Legen Sie μ0 = 3480 als das wahre durchschnittliche Geburtsgewicht der Babys in dem Land fest, mit dem wir vergleichen möchten. Die Nullhypothese lautet H0:μ = μ0 = 3480. Da der Rat den Auftrag hat, die Behauptung zu bewerten, dass das wahre Geburtsgewicht in dieser Stadt unter dem Bevölkerungsdurchschnitt liegt, haben wir eine einseitige (linksseitige) Alternativhypothese: H1:μ < μ0 = 3480. Da die Standardabweichung des Geburtsgewichts in dem Land nicht angegeben ist, nehmen wir an, dass sie unbekannt ist, und verwenden stattdessen S = 250 Gramm, die Standardabweichung der Stichprobe. Aus der obigen Tabelle ergibt sich die geeignete Teststatistik wie folgt

xxx

und ihr beobachteter Wert beträgt

xxx

Der Ablehnungsbereich lautet u < uc = t11, 0,01 = -2,718. Da der beobachtete Wert der Teststatistik in den Ablehnungsbereich fällt, lehnen wir die Nullhypothese mit einem Signifikanzniveau von 1 % ab, dass das durchschnittliche Geburtsgewicht der Neugeborenen in dieser Stadt mit dem nationalen Durchschnitt übereinstimmt. Mit anderen Worten: Die Daten liefern Evidenz zugunsten der Alternativhypothese.

Statistische Tests helfen uns, objektive Schlussfolgerungen über die statistische Signifikanz der Daten zu ziehen. Es ist jedoch nicht angebracht, allein auf der Grundlage dieser Schlussfolgerungen Empfehlungen abzugeben und/oder Maßnahmen zu ergreifen (oder nicht zu ergreifen), wenn es um ein reales Problem geht. Bei Anwendungen in der realen Welt müssen wir auch die wissenschaftliche und praktische Bedeutung der beobachteten Daten berücksichtigen. Aus dem vorangegangenen Beispiel geht hervor, dass wahrscheinlich ein Geburtshelfer zu Rate gezogen werden sollte, um festzustellen, ob der Stichprobendurchschnitt innerhalb der Grenzen des "normalen" Geburtsgewichts liegt. Darüber hinaus kann der Arzt anhand der Standardabweichung der Stichprobe um eine Intervallschätzung des Durchschnitts anstelle einer Punktschätzung bitten und entscheiden, ob diese Intervallschätzung vollständig innerhalb der Grenzen des normalen Geburtsgewichts liegt. Dies sind nur einige Überlegungen, die den Rahmen dessen sprengen, was ein statistischer Test ermitteln kann.

Bislang haben wir die folgenden drei statistischen Tests besprochen:

1. Einstichproben-Gauß-Test zum Testen von Behauptungen über den wahren Anteil π einer Grundgesamtheit im Vergleich zu einer Basislinie.

2. Einstichproben-Gauß-Test zum Testen von Behauptungen über den wahren Mittelwert μ einer Grundgesamtheit gegenüber einer Basislinie, wenn die Standardabweichung σ der Grundgesamtheit bekannt ist. Dieser Test setzt voraus, dass die Verteilung der Grundgesamtheit (annähernd) gaußförmig ist. Wenn diese Annahme nicht zutrifft, muss der Stichprobenumfang groß genug sein, damit der zentrale Grenzwertsatz die Normalität der Verteilung des Stichprobenmittelwerts garantieren kann.

3. Einstichproben-T-Test zum Testen von Behauptungen über den wahren Mittelwert μ einer Grundgesamtheit gegenüber einer Basislinie, wenn die Standardabweichung unbekannt ist und durch die Standardabweichung der Stichprobe ersetzt wird. Für diesen Test gelten die gleichen Annahmen wie für den vorherigen Test.

4.2 Einige gängige nicht-parametrische Tests

Die im vorigen Abschnitt besprochenen statistischen Tests werden als parametrische Tests bezeichnet, weil die Hypothesen Aussagen über Verteilungsparameter beinhalten. In diesem Abschnitt werden wir einige gängige nicht-parametrische Tests besprechen. Dabei handelt es sich um Tests, bei denen die Hypothese keine Aussagen über Populationsparameter enthält.

### Der Chi-Quadrat-Anpassungstest (Goodness-Of-Fit)

Angenommen, ein Unternehmen ist an der Verteilung der Abwesenheiten seiner Mitarbeiter auf die Arbeitstage Montag bis Freitag interessiert. Zu diesem Zweck soll die Hypothese, dass die Verteilung der Abwesenheiten über die fünf Arbeitstage variiert, gegenüber der Nullhypothese getestet werden, dass die Verteilung gleichmäßig ist. Die Nullhypothese besagt, dass die Verteilung gleichmäßig ist

xxx

Die Alternativhypothese, H1, besagt, dass die Verteilung unterschiedlich ist. Wir legen ein Signifikanzniveau von α = 0 . 05 fest und erheben dann einige Daten. Wir erheben Daten über erwartete und beobachtete Abwesenheiten. Die folgende Tabelle zeigt eine Zusammenfassung dieser Daten.

Tabelle 17: Beobachtete und erwartete Abwesenheiten an verschiedenen Wochentagen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Wochentag |  |  |  |  |
| Anzahl der Abwesenheiten | Montag | Dienstag | Mittwoch | Donnerstag | Freitag |

Jede der fünf Zellen in dieser Tabelle enthält die beobachtete und die erwartete Anzahl. Die Gesamtzahl der beobachteten Abwesenheiten beträgt 140 + 158 + ⋯ + 200 = 800. Laut der Nullhypothese beträgt die erwartete Anzahl der Abwesenheiten am Montag also 15 . 800 = EMon. Die beobachtete Zahl der Abwesenheiten am Montag beträgt OMon = 140. Tatsächlich sind die erwarteten Zählwerte gemäß der Nullhypothese alle gleich: EMon = ⋯EFri. Keine der beobachteten Zahlen stimmt mit den erwarteten Zahlen überein. Dies könnte jedoch aufgrund von Zufälligkeiten geschehen, selbst wenn die Nullhypothese wahr wäre. Wir wollen wissen, ob die Abweichungen der beobachteten Zählwerte von den erwarteten Zählwerten statistisch signifikant sind. Zu diesem Zweck wird eine standardisierte Größe, eine Teststatistik, verwendet, die die Gesamtabweichung der beobachteten Zählwerte von den erwarteten Zählwerten quantifiziert. Diese Teststatistik ist wie folgt definiert

xxx

Diese Größe ist natürlich zufällig, da sie von den beobachteten Zählwerten Oi abhängt. Die Verteilung von U ist χ2 4 : eine Chi-Quadrat-Verteilung mit ν = 5 - 1 = 4 Freiheitsgraden. Je größer die Differenz zwischen der beobachteten und der erwarteten Anzahl ist, desto größer ist der beobachtete Wert dieser Teststatistik. Große Werte von U sind also Evidenz gegen die Nullhypothese. Das 1 - α = 1 - 0 . 05 = 0 . 95 Quantil von U χ2 4 ist uc = 9 . 488. Wenn der beobachtete Wert uobs größer als dieser Wert ist, wird die Nullhypothese bei dem Signifikanzniveau α = 0 . 05 abgelehnt. Andernfalls wird die Nullhypothese nicht abgelehnt. Der beobachtete Wert wird wie folgt berechnet

xxx

Somit beträgt uobs > uc = 9 . 488 und wir können daraus schließen, dass die Daten die Nullhypothese widerlegen. Wir lehnen die Hypothese, dass die Verteilung der Abwesenheiten über die Wochentage gleichmäßig ist, zugunsten der Alternativhypothese ab, dass die Verteilung nicht gleichmäßig ist (bei dem Signifikanzniveau von 5 %).

Eine χ2-Verteilung ist durch einen Parameter, ν, gekennzeichnet, der als Freiheitsgrad bezeichnet wird. Obwohl wir die PDF nicht direkt verwenden werden, wird sie im Folgenden als Referenz angegeben. Die gebräuchlichsten Statistikpakete (Excel, R, Python (SciPy)) verfügen über numerische Implementierungen der PDF, CDF und Inverse CDF der χ2-Verteilung. Es ist zu beachten, dass die Unterstützung der χ2-Verteilung nicht-negativ ist und zur Modellierung von Zufallsvariablen verwendet wird, die nicht-negativ sind.

f x =

1

2ν/2Γ ν/2

xν/2 - 1e-x/2 x > 0undν = 1oderx ≥ 0undν > 1

0 sonst

Die folgende Abbildung zeigt einen Graphen der oben verwendeten PDF von χ2 4 sowie die Werte von uc und uobs.

Abbildung 39: PDF der Chi-Quadrat-Verteilung mit vier Freiheitsgraden und einem rechtsseitigen Ablehnungsbereich bei einem Signifikanzniveau von fünf Prozent

Der χ2-Anpassungstest verwendet beobachtete Zählwerte, die in vordefinierte und erschöpfende Kategorien fallen. Um die empirische Verteilung mit der in der Nullhypothese definierten theoretischen Verteilung zu vergleichen, nehmen wir an, dass eine kategoriale Variable K mögliche Werte (Klassen) hat, die mit k = 1, 2, ...,K bezeichnet werden. Die Nullhypothese spezifiziert die (diskrete) kategoriale Verteilung p1 = ℙ A1 , die Wahrscheinlichkeit, dass die Variable zur Klasse k = 1 gehört, p2 = ℙ A2 , die Wahrscheinlichkeit, dass die Variable zur Klasse k = 2 gehört, und so weiter bis pk = ℙ Ak . In der vorangegangenen Diskussion waren die Kategorien die Wochentage. Wir hätten die Klassen auch so umkodieren können, dass k = 1 für Montag, k = 2 für Dienstag usw. steht. Die beobachteten Zählwerte werden mit Ok für k = 1, 2,...K bezeichnet. Die erwarteten Zählwerte werden anhand der Gesamtzählwerte n = Σ

k = 1

K

Ok und der in der Nullhypothese angegebenen Verteilung berechnet. Die erwartete Anzahl für die Klasse k ist gegeben durch Ek = npk, k = 1, 2, ...,K.

Die unten angegebene Teststatistik wird verwendet, um die empirische Verteilung (beobachtete Zählwerte) mit der theoretischen Verteilung der Nullhypothese (erwartete Zählwerte) zu vergleichen:

xxx

Mit anderen Worten: U folgt einer χ2-Verteilung mit K - 1 Freiheitsgraden. In der Diskussion über die Abwesenheitstage in der Woche haben wir K = 5 Klassen (Arbeitstage), so dass die Teststatistik einer χ2 4-Verteilung folgt. Wenn wir auf einem α-Signifikanzniveau testen wollen, ist der Ablehnungsbereich definiert als Werte größer als uc, der Wert, der die obersten fünf Prozent der χ2 K - 1-Verteilung vom Rest trennt. Mit anderen Worten: Der χ2-Anpassungstest ist immer ein rechtsseitiger Test. Wenn der beobachtete Wert der Teststatistik uobs größer ist als der kritische Wert uc, uobs > uc, dann lehnen wir die Nullhypothese bei einem Signifikanzniveau von α ab. Andernfalls würden wir sagen, dass die Daten keine Evidenz für eine Abweichung von der vorgeschlagenen Verteilung liefern.

#### Beispiel 4.2.1

Angenommen, in der deutschen Bevölkerung beträgt die Verteilung der Blutphänotypen AB, B, O und A jeweils 5, 11, 41 und 43 Prozent. Fünfhundert zufällig ausgewählte Universitätsstudenten geben ihre Blutgruppe an. Die Zählwerte sind in der nachstehenden Tabelle zusammengefasst. Führen Sie einen χ2-Anpassungstest durch, um die Behauptung zu prüfen, dass sich die Anteile der Blutphänotypen von den Basiswerten der deutschen Bevölkerung unterscheiden. Verwenden Sie α = 0 . 05 als Signifikanzniveau und den 1 - α = 0 . 95 Quantilwert von χ2 3 , uc = u3, 0 . 95 = 7 . 815.

Tabelle 18: Blutphänotyp von 500 Hochschulstudenten

#### Lösung

Die Nullhypothese besagt, dass die Verteilung der Blutgruppen der Hochschulstudenten dieselbe ist wie die der deutschen Bevölkerung: H0:πAB = 0 . 05, πB = 0 . 11, πO = 0 . 41, πA = 0 . 43. Die Alternativhypothese besagt, dass die Verteilung anders ist. Wir berechnen zunächst die erwarteten Zählwerte für jede der einzelnen Blutgruppen für diese Studenten.

Xxx

Als Nächstes berechnen wir den beobachteten Wert der Teststatistik U χ2 3

Xxx

Da der beobachtete Wert nicht größer ist als der kritische Wert uc = 7 . 815 liegt, kann die Nullhypothese bei einem Signifikanzniveau von 0,05 nicht abgelehnt werden. Mit anderen Worten: Die Daten liefern keine Evidenz für eine Abweichung von der in der Nullhypothese angegebenen Verteilung. Im Zusammenhang mit dieser Aufgabenstellung interpretieren wir dies als: "Die Daten liefern keine statistisch signifikante Evidenz bei dem 5-Prozent-Niveau, dass sich die Verteilung der Blutgruppe von Studenten von der Grundverteilung der deutschen Bevölkerung unterscheidet."

### Der Chi-Quadrat-Test auf Unabhängigkeit

Forschende möchten wissen, ob es einen Zusammenhang zwischen der Häufigkeit des Kirchenbesuchs und der politischen Zugehörigkeit in einem überwiegend christlichen Viertel in den Vereinigten Staaten gibt. Zu diesem Zweck werden Daten von 500 Personen erhoben. Bei den Daten handelt es sich um die Antworten auf zwei Fragen: (i) Wie oft besuchen Sie pro Jahr den Gottesdienst? (ii) Welcher politischen Partei gehören Sie an? Die Antworten auf die erste Frage werden in 4 Kategorien eingeteilt: C1: weniger als dreimal pro Jahr, C2: zwischen 4- und 8-mal pro Jahr, C3: zwischen 9- und 12-mal pro Jahr und C4: mehr als 12-mal pro Jahr. Die Antwortmöglichkeiten für die zweite Frage sind P1: Republikanische Partei, P2: Demokratische Partei, P3: Unabhängig. Die Ergebnisse der Umfrage sind in der nachstehenden Tabelle zusammengefasst.

Tabelle 19: Beobachtete Zählwerte der politischen Parteizugehörigkeit und der Kirchenbesuche

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Politische Parteizugehörigkeit |  |  |  |
| Kirchenbesuche | Republikaner | Demokratisch | Unabhängig | Insgesamt |
| Weniger als 3 |  |  |  |  |
| Zwischen 4 und 8 |  |  |  |  |
| Zwischen 9 und 12 |  |  |  |  |
| Mehr als 12 |  |  |  |  |
| Insgesamt |  |  |  |  |

Der Status quo ist die Annahme, dass die beiden Variablen unabhängig sind. Dies ist die Nullhypothese, H0. Aus der Wahrscheinlichkeitstheorie wissen wir, dass, wenn zwei Ereignisse unabhängig sind, die Gesamtwahrscheinlichkeit das Produkt der Randwahrscheinlichkeiten ist. Verwenden wir die beobachteten Daten, um die Wahrscheinlichkeit zu berechnen, dass eine zufällig ausgewählte Person ein Demokrat ist und zwischen vier und acht Mal die Kirche besucht, unter der Annahme, dass Parteizugehörigkeit und Kirchenbesuch unabhängig sind. Die Randwerte sind ℙ Demokrat = ℙ C2 = 180 500 und ℙ zwischen 8 und 8 = ℙ P = P2 = 140 500

Xxx

Betrachtet man die beobachtete (unbedingte) Wahrscheinlichkeit, so ergibt sich

Xxx

Natürlich sind diese beiden Wahrscheinlichkeiten unterschiedlich, aber ist dieser Unterschied signifikant? Was ist mit den anderen Zellen? Sie können überprüfen, dass ℙ P1,C3 = 0 . 0800 und ℙ P1,C3 H0 = 0 . 14880 betragen, eine größere Differenz. Wenn wir analoge Wahrscheinlichkeiten für die anderen Zellen in der Tabelle berechnen, werden wir feststellen, dass die gemeinsame bedingte Wahrscheinlichkeit nicht mit der gemeinsamen (unbedingten) Wahrscheinlichkeit identisch ist. Wir möchten prüfen, ob der Gesamtunterschied zwischen dem, was beobachtet wird (unbedingte Wahrscheinlichkeit) und dem, was (unter der Bedingung der Unabhängigkeit) erwartet wird, statistisch signifikant ist. Ähnlich wie beim χ2-Anpassungstest werden wir Zählwerte anstelle von Wahrscheinlichkeiten verwenden.

Zunächst werden die erwarteten Zählwerte, Eij, für alle Zellen laut H0 berechnen. Zum Beispiel,

Xxx

In ähnlicher Weise

Xxx

Oij bezeichne den beobachtete Zählwert für die Zelle in der i-ten Zeile und j-ten Spalte (wobei i = 1, ..., 4 und j = 1,⋯3). Dann ist die Gesamtzahl der Beobachtungen n = ΣijOij = 500. Oi: = ΣjOij bezeichne die Zeilensumme der i-ten Zeile und O: j = ΣiOij die Spaltensumme der Spalte j. Aus unseren obigen Berechnungen ergibt sich der erwartete Zählwert Eij in der i-ten Zeile und j-ten Spalte wie folgt

xxx

Es kann hilfreich sein, sich diese Formel in Worten zu merken: erwarteter Zählwert in Zelleij = Summe Zeilei. Summe Spaltej Gesamtsumme

Anhand dieser Formel können wir die erwarteten Zählwerte aller 12 Zellen errechnen. In der folgenden Tabelle sind sowohl die beobachteten als auch die erwarteten Zählwerte (in Klammern) zusammengefasst.

Tabelle 20: Beobachtete und erwartete Zählwerte der Parteizugehörigkeit gegenüber Kirchenbesuchen

Wie beim χ2-Anpassungstest vergleichen wir die Differenzen zwischen den beobachteten Zählwerten, Oij, und den erwarteten Zählwerten, Eij, und aggregieren alle Differenzen, um unsere standardisierte Größe zu bilden, die uns als Teststatistik dient.

Xxx

Wenn wir I Zeilen und J Spalten haben, lautet die Teststatistik im Allgemeinen

Xxx

In Worten: Diese Teststatistik folgt einer χ2-Verteilung, und die Freiheitsgrade sind das Produkt aus der Anzahl der Zeilen minus eins mal der Anzahl der Spalten minus eins. In unserem Beispiel haben wir I = 4 Zeilen und J = 3 Spalten, also sind die Freiheitsgrade 4 - 1 3 - 1 = 6.

Wie beim Anpassungstest gilt auch hier: Je höher der Wert von U, desto mehr Evidenz liefern die Daten gegen H0. Daher handelt es sich um einen rechtsseitigen Test, bei dem sich der Ablehnungsbereich im rechten Ausläufer befindet. Bei einem Signifikanzniveau von einem Prozent ergibt der kritische Wert (aus einer χ2-Tabelle oder mit Hilfe einer Computersoftware) uc = 16 . 812. Der beobachtete Wert wird wie in der obigen Gleichung angegeben berechnet, genauer gesagt

xxx

Da uobs > uc ist, lehnen wir die Nullhypothese bei einem Signifikanzniveau von einem Prozent ab. Mit anderen Worten: Die Daten liefern bei diesem Signifikanzniveau Evidenz, dass politische Zugehörigkeit und Häufigkeit des Kirchenbesuchs voneinander abhängig sind.

Die folgende Tabelle enthält die Befehle in verschiedenen gängigen Softwareprogrammen zur Berechnung des kritischen Wertes der χ2-Verteilung mit df Freiheitsgraden und Signifikanzniveau.

Tabelle 21: Befehle zur Berechnung des kritischen Wertes einer Chi-Quadrat-Verteilung in verschiedenen Softwarepaketen

|  |  |
| --- | --- |
| SoftwarepPaket | Befehl |

Bei beiden von uns untersuchten χ2-Tests ist zu beachten, dass die Verteilung der Teststatistik nicht genau die χ2-Verteilung ist. Stattdessen ist sie asymptotisch die χ2-Verteilung, d. h. für eine große Anzahl von Beobachtungen. Die Zweckmäßigkeit des Tests kann in Frage gestellt werden, wenn fast alle Zellen (mehr als 80 %) erwartete Zählwerte von weniger als fünf enthalten. In solchen Fällen sind andere Tests besser geeignet.

### Kolmogorov-Smirnov-Test auf Normalverteilung

In diesem Abschnitt wird erörtert, wie man einen Anpassungstest speziell für die Gauß-Verteilung durchführt. Wie üblich besagt die Nullhypothese, dass die Verteilung, aus der die Daten gezogen werden, einer Gauß-Verteilung mit bekanntem Mittelwert μ und bekannter Standardabweichung σ folgt. Das heißt, X1, ...,Xn iidN μ, σ . Die Alternativhypothese besagt, dass dies nicht der Fall ist. Die Teststatistik, mit der bewertet wird, inwieweit die Verteilung der Daten von der angegebenen Normalverteilung abweicht, berücksichtigt die maximale beobachtete Differenz zwischen der (theoretischen) CDF Φ . und der empirischen CDF F ., die anhand der Zufallsstichprobe ausgewertet wird:

xxx

Auf den ersten Blick ist es nicht klar, wo die Zufallsstichprobe in der Teststatistik verwendet wird, da die Abhängigkeit von den Xi explizit ist. Es ist zu beachten, dass das Maximum über alle möglichen Werte von x gilt, nicht nur über die beobachteten Werte. Wie bei allen Teststatistiken hängt D jedoch tatsächlich von der Zufallsstichprobe über die empirische CDF,F , ab, die definiert ist durch

xxx

wobei X i der i-te größte Wert in der Zufallsstichprobe ist (in der Reihenfolge vom kleinsten zum größten Wert) und # X i ≤ x die Anzahl der Werte in der Zufallsstichprobe zählt, die höchstens so groß wie x sind. Sobald wir Werte der Zufallsstichprobe x1, ..., xn beobachtet haben, ist die empirische CDF Fobs x, wobei X i durch x i für jedes i = 1,...n ersetzt wird. In der Praxis werden wir daher den Wert Dobs = maxx Φ x - Fobs x berechnen. Die nachstehende Tabelle enthält die kritischen Werte (Grenzwerte) für verschiedene Werte von α für n = 10 (zur Verwendung bei Stichproben mit einem Umfang von 10).

Tabelle 22: Kritische Werte für den Kolmogorov-Smirnov-Test auf Normalverteilung für verschiedene Signifikanzniveaus und Stichproben des Umfangs zehn

Obwohl wir den beobachteten Wert der Teststatistik nicht von Hand berechnen können, betrachten wir zur Veranschaulichung der Ideen, die diesem Test zugrunde liegen, einen Datensatz der Größe zehn:

- 2 . 43, - 1 . 51, - 1 . 09, - 0 . 87, - 0 . 58, - 0 . 43, 0 . 28, 1 . 00, 1 . 27, 1 . 65 . Wir wollen sehen, ob diese Stichprobe aus einer Gaußschen Standardverteilung stammt: N 0, 1 . Der erste Schritt besteht darin, die empirische CDF für diese Stichprobe aufzuschreiben: Fobs x . Diese stückweise Funktion ist gegeben durch

xxx

Der beobachtete Wert der Teststatistik ist max

x

Φ x - Fobs x ; grafisch gesehen ist dies der maximale vertikale Abstand zwischen den Graphen von Φ und Fobs. Die nachstehende Abbildung zeigt eine Darstellung der beiden Graphen und hebt den Maximalwert als Dobs = 0 . 266 hervor. Der Tabelle zufolge entspricht der kritische Wert, α = 0 . 05, Dc = 0 . 41. Da der beobachtete Wert nicht größer ist als der kritische Wert, können wir die Nullhypothese bei dem Signifikanzniveau α = 0 . 05 nicht ablehnen. Mit anderen Worten: Die Daten liefern bei diesem Signifikanzniveau keine Evidenz dafür, dass die Daten aus einer anderen Verteilung als der Standardnormalverteilung stammen.

Abbildung 40: Empirische und theoretische CDFs für den Kolmogorov-Smirnov-Test auf Gleichheit

In unserer obigen Diskussion haben wir die Gaußsche Standardverteilung verwendet. Der Kolmogorov-Smirnov-Test funktioniert jedoch mit jeder beliebigen Normalverteilung, solange sie vollständig spezifiziert ist. Mit anderen Worten: Wir müssen den Mittelwert und die Standardabweichung der Normalverteilung, auf die wir testen wollen, vorher bestimmen, ohne die Daten zu verwenden. Es ist verlockend, den Stichprobenmittelwert und die Stichprobenstandardabweichung aus den Daten für diese Parameterwerte zu verwenden, aber der Test ist dann ungültig. Wenn die Parameter der Zielverteilung nicht bekannt sind, sollte ein anderer Test angewendet werden. Der Kolmogorov-Smirnov-Test kann auch verwendet werden, um einen Anpassungstest anderer kontinuierlicher Verteilungen als der Gauß-Verteilung durchzuführen. Wir würden einfach die entsprechende CDF wählen. In der Praxis wird bei vielen Algorithmen für statistische Inferenz und statistisches Lernen davon ausgegangen, dass die zugrundeliegende Verteilung gaußförmig ist. Aus diesem Grund konzentrieren wir uns in unserer Präsentation auf den Kolmogorov-Smirnov-Test für diese Verteilung.

4.3 ZweisStichprobentests

Im vorigen Abschnitt haben wir gelernt, wie man einige statistische Tests durchführt, bei denen ein Parameter einer interessierenden Population mit dem bekannten entsprechenden Parameter einer Referenz-Grundgesamtheit verglichen wird. In diesem Abschnitt werden wir lernen, wie man statistische Tests durchführt, bei denen wir Parameter aus zwei interessierenden Populationen vergleichen, deren wahre Parameter beide unbekannt sind. Für solche statistischen Tests werden zwei Stichproben benötigt, eine aus jeder Population. Es ist zu beachten, dass wir in diesem Abschnitt davon ausgehen, dass die beiden Grundgesamtheiten unabhängig voneinander sind.

### Vergleich von zwei Anteilen: Ein Z-Test

Beachten Sie, dass A/B-Tests manchmal anstelle von Zweistichprobentests verwendet werden. In jedem Fall handelt es sich um einen statistischen Test mit zwei Stichproben, bei dem die beiden Stichproben (idealerweise) in jeder Hinsicht identisch sind, außer in Bezug auf die Art der "Behandlung". Bei solchen Tests versuchen wir festzustellen, ob der interessierende Parameter (Anteil oder Mittelwert) mit der Art der Behandlung assoziiert ist.

Betrachten wir das in der Einleitung beschriebene CTR-Szenario. πA und πB sollen die wahre CTR für die Designs A bzw. B bezeichnen. Wie bei Einstichprobentests besagt die Nullhypothese "kein Effekt"; die CTRs sind für beide Designs gleich oder H0:πA = πB. Wie bei den Einstichprobentests ist die Alternativhypothese eine Aussage über eine Veränderung der CTR. Sie kann eine von drei Formen annehmen: (i) Design A hat eine niedrigere CTR als Design B, (ii) Design A hat eine höhere CTR als Design B, (iii) die beiden Designs haben unterschiedliche CTRs. Diese drei möglichen Alternativhypothesen können wie folgt dargestellt werden

1. H1:πA < πB

2. H1:πA > πB

3. H1:πA ≠ πB

Ähnlich wie bei Einstichprobentests sind die ersten beiden für einseitige Tests und der letzte für einen zweiseitigen Test vorgesehen. Die Entscheidung, welche der drei alternativen Hypothesen verwendet werden soll, muss getroffen werden, bevor Daten gesammelt oder betrachtet werden. In unserem Fall wollen wir herausfinden, ob das neue Design, B, eine höhere Click-Through-Rate ergibt. Daher wählen wir H1:πA < πB als unsere Alternativhypothese.

Lassen Sie uns diese Symbole näher erläutern. In unserem Szenario haben wir zwei unabhängige Grundgesamtheiten, A und B. Grundgesamtheit A repräsentiert alle Besucher, die mit Design A konfrontiert würden, und Grundgesamtheit B die Besucher, die mit Design B konfrontiert würden. Das Ergebnis jedes Mitglieds der Grundgesamtheitenkann als Zufallsvariable X Bernoulli πA und Y Bernoulli πB dargestellt werden, wobei X = 1 das Ereignis bezeichnet, dass ein Benutzer aus der ersten Grundgesamtheit auf die Zielschaltfläche geklickt hat, und entsprechend für Y = 1.

Für jeden statistischen Test werden Daten benötigt. Für A/B-Tests ist es unerlässlich, dass wir Daten aus einem randomisierten Experiment gewinnen. Das heißt, dass wir jedem Besucher, der die Ziel-Webseite aufruft, nach dem Zufallsprinzip (mit gleicher Wahrscheinlichkeit) entweder Design A oder B zeigen. Je nachdem, welche Seite er anschaut, stellt der Besucher einen Stichprobenpunkt aus der GrundgesamtheitA oder B dar. Daher nehmen wir zwei Stichproben auf, eine aus jeder Grundgesamtheit. X1, ...,Xm sollen eine Zufallsstichprobe aus Grundgesamtheit A und Y1, ...,Yn eine Zufallsstichprobe aus Grundgesamtheit B bezeichnen. Es ist zu beachten, dass die beiden Stichproben unabhängig voneinander sind.

Sobald wir die beobachteten Werte dieser Stichproben x1, ..., xm und y1, ..., yn haben, können wir die Schätzungen der CTRs berechnen: πA = 1m

Σ

i = 1

m

xi und πB = 1n

Σ

j = 1

n

yj. Theoretisch sind zwei Extremfälle denkbar, die beobachtet werden könnten. Einerseits könnten wir πA - πB = 0 erhalten, in diesem Fall gäbe es keinen Grund, H0 abzulehnen, und andererseits könnten wir πA - πB = 1 erhalten, in diesem Fall könnten wir sicherlich für die Ablehnung von H0 argumentieren. Selbst wenn H0 wahr wäre, ist es aufgrund der inhärenten Zufälligkeit unwahrscheinlich, dass wir πA - πB = 0 sehen. Ebenso ist es unwahrscheinlich, dass wir einen bestimmten Unterschied in den Stichprobenanteilen beobachten, wenn H0 falsch wäre: πA - πB = 0 . 1. Es ist sinnvoller, die Wahrscheinlichkeiten für einen Bereich von Werten zu berechnen.

Wenn H0 wahr wäre, wie wahrscheinlich ist es, dass wir πA - πB ≤ d beobachten? Wir können diese Frage mit Hilfe der Schätzer πA = 1m

Σ

i = 1

m

Xi und πB = 1n

Σ

j = 1

n

Yj wie folgt stellen:

Xxx

Um eine solche Wahrscheinlichkeit zu berechnen, müssen wir die Verteilung von πA - πB kennen. Wenn die Stichproben groß sind (m, n groß sind), besagt der zentrale Grenzwertsatz, dass die standardisierte Differenz der Stichprobenanteile annähernd normalverteilt ist (unter der Voraussetzung von H0:πA = πB). Symbolisch gilt

xxx

Wir erinnern daran, dass V πA - πB = V πA + V πB =πA 1 - πAm +πB 1 - πBn . Da wir die wahren Proportionen nicht kennen, können wir die Varianz approximieren, indem wir sie durch ihre Schätzer ersetzen: V πA - πB ≈πA 1 - πAm +πB 1 - πBn . Wir erhalten also

xxx

Nun kann die Wahrscheinlichkeit von ℙ πA - πB ≤ d πA = πB geschrieben werden als

Xxx

wobei Φ . die CDF der Gaußschen Standardverteilung ist.

Damit kommen wir zum zweiten Teil der Hypothesentests: die Teststatistik (U) und der Ablehnungsbereich. Die Wahrscheinlichkeit misst, wie wahrscheinlich es ist, dass wir einen Datensatz beobachten, der einem Unterschied in den Anteilen von mindestens d entspricht, wenn die Nullhypothese wahr ist. Intuitiv gesprochen, wäre die Nullhypothese wahr (die wahren Anteile sind gleich), wäre die Wahrscheinlichkeit, einen großen (negativen) Unterschied in den Anteilen zu beobachten, gering. Analog dazu ist die Wahrscheinlichkeit, kleine Unterschiede zu beobachten, groß.

Das Signifikanzniveau α ist die maximale Wahrscheinlichkeit eines Fehlers 1.Art, der wir ausgesetzt sein wollen. Mit anderen Worten: Wenn die Anteile gleich sind, soll die Nullhypothese in höchstens 100α% der Fälle (fälschlich) abgelehnt werden. Wenn α = 0 . 01 beträgt, bedeutet dies, dass die Wahrscheinlichkeit, einen Fehler 1. Art zu machen, nicht mehr als ein Prozent betragen sollte. In diesem Sinne entscheiden wir uns für eine Ablehnung, wenn die Wahrscheinlichkeit kleiner ist als α = 0 . 01. Die Menge der Werte von u, für die ℙ U < u < α gilt, wird als Ablehnungsbereich bezeichnet. Der Grenzwert wird als kritischer Wert bezeichnet, der mit ucr bezeichnet wird und sich aus ℙ U = ucr = α ergibt. In unserem Fall ist bei 01 ucr einfach das Quantil der Gauß-Verteilung, das einem Prozent entspricht, ϕ uc = 0 . 01 oder uc ≈ - 2 . 33. Daher umfasst der Ablehnungsbereich, der einem Signifikanzniveau von α = 0 . 01 entspricht, die Werte in der Menge RR = u u < 2 . 33 . Wenn wir Daten beobachten, die einem Wert der Teststatistik (u) von weniger als 2,33 entsprechen, dann lautet unsere Entscheidung, die Nullhypothese bei dem Signifikanzniveau von 1 % abzulehnen. Dies würde darauf hindeuten, dass Design B möglicherweise zu einer höheren CTR führt als Design A.

#### Beispiel 4.3.1

Anknüpfend an unsere Diskussion über den Vergleich von CTRs nehmen wir an, dass wir 567 Klicks von 900 und 650 Klicks von 950 aus Stichproben beobachtet haben, die aus den Designs A bzw. B gewonnen wurden. Berechnen Sie die beobachtete Teststatistik uobs von U und notieren Sie die Entscheidung des A/B-Tests bei dem Signifikanzniveau von einem Prozent.

#### Lösung

Die Anteile der Stichprobe sind gegeben durch πA = 567

900 = 0 . 6300 und πB = 650

950a ≈ 0 . 6842. Der Standardfehler (Standardabweichung der Differenz der Stichprobenanteile) wird wie folgt errechnet

xxx

Setzt man diese Zahlen zusammen, kann man den beobachteten Wert der Teststatistik aus der obigen Gleichung für U wie folgt berechnen

xxx

Wie in der Diskussion erwähnt, handelt es sich um einen linksseitigen Test mit dem Ablehnungsbereich RR = u u < - 2 . 33 . Da die uobs in den Ablehnungsbereich fällt, wird die Nullhypothese bei einem Signifikanzniveau von einem Prozent abgelehnt. Daher liefern die Daten Evidenz für die Alternative H1: πA < πB; Design B hat möglicherweise eine höhere CTR.

Das Ergebnis eines statistischen Tests allein reicht nicht aus, um das Design zu ändern oder Empfehlungen auszusprechen. Wie im ersten Abschnitt dieser Einheit erörtert, müssen andere Überlegungen, wie die praktische Bedeutung, analysiert werden. Hier ist der beobachtete Effekt πA - πB ≈ 0 . 6300 - 0 . 6842 = - 0 . 0542. Mit anderen Worten: Die Veränderung der CTR begünstigt Design B mit einer Marge von etwa fünf Prozent. Ein Unternehmensanalytiker muss möglicherweise zusammen mit dem Designteam entscheiden, ob ein solcher Anstieg der CTR den Einsatz zusätzlicher Ressourcen rechtfertigt, um die Änderung dauerhaft zu machen. Diese Überlegung liegt außerhalb des Rahmens des von uns durchgeführten statistischen Tests und wird den entsprechenden Experten überlassen.

### Vergleich zweier Mittelwerte

In Abschnitt 4.1 haben wir erörtert, wie man eine Behauptung über den wahren Mittelwert einer interessierenden Grundgesamtheitmit einem Basiswert prüft. In diesem Teil lernen wir, wie man Behauptungen über den Vergleich der wahren Mittelwerte zweier unabhängiger Grundgesamtheiten prüft. Die Nullhypothese, der Status quo, besagt, dass die beiden Mittelwerte gleich sind: H0:μ1 = μ2, während die Alternative verschiedene Formen annehmen kann:

- zweiseitiger Test (Ablehnungsbereich mit zwei Ausläufern): H1:μ1 ≠ μ2

- einseitiger Test

- linksseitiger Ablehnungsbereich: H1:μ1 < μ2

- rechtsseitiger Ablehnungsbereich:H1:μ1 > μ2

Wie zuvor müssen die Alternativhypothese sowie das Signifikanzniveau α vor der Analyse von Daten festgelegt werden. Der Rahmen für den Vergleich zweier Mittelwerte aus unabhängigen Stichproben ist ähnlich wie bei jedem anderen Hypothesentest. Er besteht aus vier Teilen:

1. Hypothesen

2. α und Wahl der Teststatistik

3. Beobachteter Wert der Teststatistik und Festlegung des Ablehnungsbereichs

4. Entscheidung

Um die drei häufigsten Fälle des Vergleichs von Mittelwerten aus unabhängigen Grundgesamtheiten zu erörtern, ist der allgemeine Rahmen derselbe. Das Einzige, was sich ändert, ist die Teststatistik und ihre zugehörige Verteilung. Bezeichnen wir mit μ1 und μ2 die wahren, aber unbekannten Mittelwerte der interessierenden (unabhängigen) Grundgesamtheiten und mit X1, ...,Xn1 und Y1, ...,Yn2 zwei Zufallsstichproben, eine aus jeder Grundgesamtheit. Es ist zu beachten, dass der Stichprobenumfang mit n1 bzw. n2 bezeichnet wird. Die beobachteten Werte dieser Stichproben werden mit x1, ..., xn1 bzw. y1, ..., yn2 bezeichnet. Die einzige Annahme, die alle drei Fälle, die wir im Folgenden erörtern, gemeinsam haben, ist, dass entweder sowohl n1 als auch n2 groß genug sind, um eine Approximation der Normalverteilung der Stichprobenmittelwerte X und Y nach dem zentralen Grenzwertsatz zu ermöglichen, oder dass die zugrundeliegende Verteilung, aus der beide Stichproben gezogen werden, (zumindest annähernd) gaußförmig ist. Wenn keine dieser Bedingungen erfüllt ist, sind die Tests möglicherweise nicht gültig. Die drei von uns erörterten Fälle beruhen auf Informationen über die Varianzen (Standardabweichungen) der zugrundeliegenden Verteilungen:

1. Die Varianzen der beiden Populationen, σ1 2 und σ2, sind bekannt.

2. Die Varianzen der beiden Populationen sind unbekannt, werden aber als gleich angenommen.

3. Die Varianzen der beiden Populationen sind unbekannt und werden als ungleich angenommen.

Im ersten Fall wird ein Z-Test verwendet, d. h. ein Test auf der Grundlage der Gauß-Verteilung. In den beiden anderen Fällen verwenden wir einen T-Test, d. h. einen Test auf der Grundlage der Student-t-Verteilung.

#### Bekannte Varianzen: Ein Z-Test

Die Nullhypothese H0:μ1 = μ2 kann auch als H0:μ1 - μ2 = 0 geschrieben werden. Auf diese Weise motiviert sie uns, die Differenz der Stichprobenmittelwerte X - Y zu berechnen und diese Differenz mit Null zu vergleichen. Wenn die Differenz nämlich weit von Null entfernt ist, neigen wir dazu, die Nullhypothese abzulehnen. Anstatt mit dieser Differenz zu arbeiten, würden wir es vorziehen, mit einem standardisierten Wert zu arbeiten. Wenn die Stichproben aus unabhängigen Gauß-Verteilungen gezogen werden, dann

xxx

Laut der Nullhypothese wissen wir, dass μ1 - μ2 = 0, also

Xxx

Daher beträgt die Größe

Xxx

Diese Zufallsvariable ist unsere Teststatistik. Der beobachtete Wert dieser Teststatistik ersetzt X und Y durch die beobachteten Stichprobenmittelwerte x bzw. y. Schließlich wird der Ablehnungsbereich auf genau dieselbe Weise entwickelt wie bei jedem anderen Z-Test.

#### Beispiel 4.3.2

Die Stundenlöhne von zwei Unternehmen, A und B, sind normalverteilt. Die Varianzen der Stundenlöhne betragen σ1 2 = 20 und σ2 2 = 18 für die Unternehmen A bzw. B. Forschende möchten wissen, ob die durchschnittlichen Stundenlöhne der beiden Unternehmen unterschiedlich sind. Zu diesem Zweck erheben sie die Stundenlöhne von 20 Arbeitnehmern, 10 aus jedem der beiden Unternehmen. Die Mittelwerte der Stichprobe betragen x =15,23€ und y =14,15€ für Unternehmen A bzw. B. Führen Sie einen Hypothesentest mit einem Signifikanzniveau von fünf Prozent durch, um das Interesse der Forschenden zu zu bedienen.

#### Lösung

Die Verteilung der Stundenlöhne ist für beide Unternehmen gaußförmig und die Varianzen sind bekannt. Daher können wir einen Z-Test durchführen.

1. H0:μ1 = μ2, die mittleren Stundenlöhne der beiden Unternehmen sind gleich. H1:μ1 ≠ μ2, die mittleren Stundenlöhne der beiden Unternehmen sind unterschiedlich.

2. α = 0 . 05 und

Xxx

3. Die Grenzwerte sind uc = Å} z0.025 = Å} 1,96, d. h. der Ablehnungsbereich ist RR = u u < - 1 . 96oderu > 1 . 96 . Der beobachtete Wert der Teststatistik beträgt

Xxx

4. Da uobs nicht im Ablehnungsbereich liegt, kann die Nullhypothese bei einem Signifikanzniveau von fünf Prozent nicht abgelehnt werden.

Gemäß unserer Entscheidungserklärung können wir sagen, dass die Daten keine Evidenz für unterschiedliche Durchschnittslöhne zwischen den beiden Unternehmen liefern (bei dem gegebenen Signifikanzniveau)liefern.

#### Unbekannte/gleiche Varianzen: Ein t-Test

Der zweite Fall beim Testen auf Mittelwertgleichheit tritt auf, wenn die Varianzen der Grundgesamtheiten unbekannt sind, aber als gleich angenommen werden. In diesem Fall müssen wir die Teststatistik anpassen, da wir die Werte von σ1 2 und σ2 2 nicht kennen. Daher müssen diese Größen durch eine Schätzung einer gemeinsamen Stichprobenvarianz ersetzt werden. Diese Größe wird als gepoolte Varianz, Sp 2 , bezeichnet und berechnet sich wie folgt

xxx

wobei S1 2 und S2 2 die Stichprobenvarianz der Xi bzw. Yj sind (i = 1, ..., n1 und j = 1, ..., n2). Die zugehörige Teststatistik folgt einer t-Verteilung:

xxx

Die Größe ν = n1 + n2 - 2 ist die Anzahl der Freiheitsgrade der t-Verteilung. Der beobachtete Wert dieser Teststatistik ergibt sich aus der Ersetzung von X, Y, S1 2 und S2 2 durch die beobachteten Werte x, y, s1 2 bzw. s2 2.

#### Beispiel 4.3.3

Forschende möchten herausfinden, ob das durchschnittliche (jährliche) Gehalt der männlichen Manager eines bestimmten Unternehmens höher ist als das durchschnittliche Gehalt der weiblichen Manager. Sie gehen davon aus, dass die Gehälter von männlichen und weiblichen Managern annähernd gaußförmig sind und gleiche (aber unbekannte) Varianzen aufweisen. Die Forschenden erheben eine Zufallsstichprobe von 32 Managern, 16 männlichen und 16 weiblichen. Die folgende Tabelle fasst die Daten zusammen.

Tabelle 23: Zusammenfassung der Daten für Beispiel 4.3.3

Testen Sie die Behauptung der Forschenden mit einem Signifikanzniveau von α = 0 . 01 .

#### Lösung

Da die Gehälter der männlichen und weiblichen Manager normalverteilt sind und die Varianzen unbekannt sind, aber als gleich angenommen werden, wird ein t-Test mit der gepoolten Varianzversion der Teststatistik verwendet. Bezeichnen wir mit μ1 und μ2 die Durchschnittsgehälter der männlichen bzw. weiblichen Manager.

1. H0:μ1 = μ2, das Durchschnittsgehalt der männlichen und weiblichen Führungskräfte ist gleich. H1:μ1 > μ2, das Durchschnittsgehalt der männlichen Manager ist höher als das Durchschnittsgehalt der weiblichen Manager.

2. α = 0 . 01 und

Xxx

3. Der Grenzwert beträgt uc = t30, 0 . 01 = 2 . 457, und wir haben einen rechtsseitigen Ablehnungsbereich: RR = u u > 2 . 457 . Die beobachteten Werte der gepoolten Varianz und der gepoolten Standardabweichung lauten

xxx

Der beobachtete Wert der Teststatistik lautet dann

Xxx

4. Da der beobachtete Wert im Ablehnungsbereich liegt, lehnen wir die Nullhypothese bei einem Signifikanzniveau von einem Prozent ab.

Unserer Entscheidung folgend können wir sagen, dass die Daten Evidenz liefern, dass das Durchschnittsgehalt männlicher Manager höher ist als das Durchschnittsgehalt weiblicher Manager (bei dem gegebenen Signifikanzniveau).

### Unbekannte/ungleiche Varianzen: Ein t-Test

Der letzte Fall, den wir erörtern werden, tritt auf, wenn die Varianz der Grundgesamtheitenunbekannt ist und nicht als gleich angenommen werden kann. In diesem Fall können wir σ1 2 und σ2 2 aus der Teststatistik des Z-Tests (bekannte Varianzen) durch S1 2 bzw. S2 2 ersetzen. Die Verteilung der resultierenden Teststatistik ist annähernd t (Welch, 1947):

xxx

wobei die Freiheitsgrade, ν, gegeben sind durch

xxx

Obwohl die Berechnung von νW, den Freiheitsgraden, recht aufwändig ist, verfügen fast alle gängigen Statistikpakete über Implementierungen zur Durchführung dieses t-Tests. Die folgende Tabelle zeigt, wie der t-Test mit zwei unabhängigen Stichproben, x und y, unter der Annahme, dass die Varianzen unbekannt und ungleich sind, durchgeführt wird. Alle Befehle verwenden standardmäßig eine zweiseitige Alternativhypothese. Dieser Test wird manchmal auch Welch-Test genannt, nach dem berühmten Statistiker, der ihn entwickelt hat.

Tabelle 24: Befehle für einen t-Test mit unbekannten und ungleichen Varianzen in verschiedenen Software-Paketen

## 4.4 Leistung, P-Werte und Konfidenzintervalle

### Die Teststärke (Power) eines Tests

In Abschnitt 4.1 haben wir erörtert, dass bei der Durchführung eines Hypothesentests je nach Entscheidung zwei Fehler möglich sind: (i) ein Fehler 1. Art, bei dem wir eine wahre Nullhypothese ablehnen, und (ii) ein Fehler 2. Art, bei dem wir eine falsche Nullhypothese nicht ablehnen. Die Wahrscheinlichkeiten für solche Fehler werden mit α bzw. β bezeichnet.

Nämlich α = ℙ H0 ablehnen H0 ist wahr und β = ℙ H0 nicht ablehnenH0 ist falsch

Eine eng verwandte Größe ist die Teststärke (Power) eines Tests. Informell ist diese Größe die Wahrscheinlichkeit, einen Effekt zu erkennen, wenn es tatsächlich einen Effekt gibt. Mit anderen Worten, die Power eines Tests ist das Komplement des Fehlers 2. Art:

Teststärke = 1 - β

Wenn wir eine dieser beiden Größen, β oder Teststärke, berechnen können, dann ist die andere Größe einfach das Komplement. Für allgemeine statistische Tests ist es sehr schwierig, diese Wahrscheinlichkeiten zu berechnen. Daher konzentrieren wir uns auf einen Einstichproben-Z-Test.

Wir erinnern uns an die Rahmenbedingungen für einen Einstichproben-Z-Test. X1, ...,Xn sind unabhängige Gaußsche N μ, σ mit bekannter σ. Die Wahrscheinlichkeit eines Fehlers 1. Art, α, ist festgelegt und die Nullhypothese lautet H0:μ = μ0 (für einen Wert von μ0. Betrachten wir eine rechtsseitige Alternative, H1:μ > μ0. Die zugehörige Teststatistik lautet

xxx

Der kritische Wert ist gegeben durch uc = zα und der Ablehnungsbereich ist RR = u u > uc . Der Ablehnungsbereich kann auch in Form des Stichprobenmittelwerts angegeben werden: RR = x x > xc mit

xc = μ0 + uc . σn

Die Wahrscheinlichkeit eines Fehlers 2. Art sowie die Teststärke eines Tests werden für einen festen Alternativwert μ = μ1 mit μ1 > μ0 berechnet:

xxx

Wir haben die Abhängigkeit von μ1 explizit gemacht, indem wir die Schreibweise β μ1 anstelle von β verwenden. Analog dazu ist die Teststärke des Tests für diese Alternative gegeben durch

Teststärke μ1 = ℙ X > xc μ = μ1

Die folgende Abbildung zeigt die beiden Verteilungen von X, eine gemäß H0:μ = μ0 und eine gemäß H1′:μ = μ1. Zusätzlich haben wir die Bereiche hervorgehoben, die den Fehlern 1. Art und 2. Art, α bzw. β, entsprechen.

Abbildung 41: Fehler 2. Art und Teststärke eines Tests, aufgetragen gegen die Wahrscheinlichkeit eines Fehlers 1. Art

Für einen festen Wert der Alternative, μ = μ1 (gleichbedeutend mit einer festen Effektgröße μ1 - μ0), gibt es einen Kompromiss zwischen β und α. Wenn wir α klein machen, dann wird der Ablehnungsbereich kleiner und β μ1 = ℙ X ∉ RR μ = μ1 wird größer. Dieser Kompromiss wird in der nachstehenden Abbildung veranschaulicht, die die Graphen von β μ1 und Teststärke μ1 gegen α bei festem μ1 (und festem Stichprobenumfang, n) zeigt.

Abbildung 42: Verteilung des Stichprobenmittelwerts gemäß zwei konkurrierenden Hypothesen zusammen mit den Wahrscheinlichkeiten von Fehlern 1. Art und 2. Art

Wir wollen natürlich, dass die Teststärke eines Tests so groß wie möglich ist. Daher ist es wichtig, auch seine Beziehung zur Effektgröße μ1 - μ0 sowie zum Stichprobenumfang n zu verstehen. Für ein festes α gilt

xxx

Bei einem festen Stichprobenumfang n ist zu erkennen, dass bei einer größeren Effektgröße die Wahrscheinlichkeit zunimmt, wenn μ1 größer und größer als μ0 wird. Dies liegt daran, dass die rechte Seite der Ungleichung, die das Ereignis beschreibt, einen größeren negativen Wert annimmt und somit das Ereignis Z > ... größer ist. Mit anderen Worten: Je größer die Effektgröße (bei festem α und n), desto höher ist die Teststärke des Tests. Das folgende Diagramm zeigt die Beziehung zwischen der Teststärke eines Tests und der Effektgröße für verschiedene Werte von α.

Abbildung 43: Teststärke eines Tests gegenüber der Effektgröße (feste Stichprobengröße und verschiedene Signifikanzniveaus)

Legen wir dagegen die Effektgröße fest (d. h. μ1), dann ist die Teststärke bei einem gegebenen α umso geringer, je größer die Stichprobengröße ist. Die folgende Abbildung zeigt die Beziehung zwischen der Teststärke eines Tests und dem Stichprobenumfang.

Abbildung 44: Teststärke eines Tests gegenüber der Effektgröße (feste Effektgröße und verschiedene Werte des

Signifikanzniveaus)

In der Planungsphase der Durchführung eines statistischen Tests müssen wir verschiedene Entscheidungen treffen. Wir müssen die Richtung der Alternativhypothese (zweiseitig, linksseitig, rechtsseitig), unsere Toleranz für einen Fehler 1. Art, α, und, bei einer gewünschten Effektgröße, unsere gewünschte Teststärke bestimmen. So können wir versuchen, unsere Anforderungen durch die Wahl des richtigen Stichprobenumfangs zu erfüllen. Um zu demonstrieren, wie das geht, nehmen wir an, dass H0:μ = μ0 ist, und planen wir, eine Effektgröße zu ermitteln, die μ = μ1 für einen rechtsseitigen Test entspricht, d. h. μ1 > μ0. Wir wollen einen Fehler 1. Art von α und einen Fehler 2. Art von β haben (Teststärke = 1 - β). Wir erhalten

xxx

wobei Z N 0, 1 . Aus den obigen Gleichungen ergibt sich

xxx

oder gleichwertig

xxx

#### Beispiel 4.4.1

Forschende möchten einen Hypothesentest mit H0:μ = μ0 = 0 und H1:μ > 0 für eine Größe durchführen, von der bekannt ist, dass sie einer Gauß-Verteilung mit bekannter Standardabweichung σ = 1 folgt. Die Forschenden möchten in der Lage sein, einen Effekt nachzuweisen, der μ1 = 0. 5 entspricht, und ihre Toleranz für den Fehler 1. Art beträgt α = 0 . 10 und die gewünschte Teststärke beträgt = 0 . 80. Welchen Stichprobenumfang sollten sie wählen?

#### Lösung

Nach der obigen Formel ergibt sich zα = z0 . 10 = 1 . 28. Da die Teststärke = 0 . 80 ist, erhalten wir β = 1 - 0 . 80 = 0 . 20, und zβ = z0 . 20 = 0 . 84. Mit σ = 1 und μ1 = 0 . 5 erhalten wir

xxx

Bei einem Stichprobenumfang von n = 18 erfüllt der Test daher α ≈ 0 . 05 und Teststärke ≈ 0 . 80.

### P-Werte

In der Praxis ist der von Forschenden gewählte Wert von α etwas willkürlich. Daher können zwei Forschende mit denselben Daten zu entgegengesetzten Schlussfolgerungen aus demselben Hypothesentest kommen. So kann es sein, dass ein Forschender α = 0 . 05 wählt und eine andere α = 0 . 01. Ersterer kann die Nullhypothese ablehnen, während letztere sie nicht ablehnen kann. Technisch gesehen muss die Wahl von α, wie bereits erwähnt, sorgfältig im Hinblick auf β/Teststärke getroffen werden. In vielen Veröffentlichungen wird verlangt, dass Forschende den kleinsten Wert von α angeben, der zu einer Ablehnung der Nullhypothese führen würde, auch bekannt als p-Wert. Da α jedoch, wie wir in dieser Einheit häufig erwähnt haben, vor der Betrachtung der Daten gewählt werden muss, ist die Interpretation des p-Wertes anfällig für Fehlinterpretationen.

Für eine Teststatistik U ist der p-Wert der kleinste Wert von α, für den die beobachteten Daten eine Ablehnung der Nullhypothese nahelegen. Je kleiner der p-Wert ist, desto wahrscheinlicher ist es, dass die Nullhypothese abgelehnt wird. Um die damit verbundenen Feinheiten zu verstehen, betrachten wir einen Test mit den Hypothesen H0:μ = μ0 und H1:μ > μ0. U soll für die zugehörige Teststatistik stehen und uobs für den zugehörigen beobachteten Wert. Der p-Wert ist die Wahrscheinlichkeit, einen Wert von U zu beobachten, der mindestens so extrem ist wie der beobachtete Wert, wobei die Wahrscheinlichkeit unter der Annahme berechnet wird, dass die Nullhypothese wahr ist. Formal gilt

p-Wert = ℙ U > uobs μ = μ0

Angenommen, U folgt einer t-Verteilung mit 30 Freiheitsgraden: U T 30 . Wenn uobs = 2 . 21, dann

xxx

Mit anderen Worten: 0 . 0174 oder 1 . 74 % ist der kleinste Wert von α, der zu einer Ablehnung der Nullhypothese führen würde.

Wenn die Alternativhypothese zweiseitig war, H1:μ ≠ μ0, bedeutet die Aussage "mindestens so extrem", dass wir beide Richtungen der Extremwerte berücksichtigen müssen, extrem und über dem Gegenteil sowie extrem und darunter. Mit anderen Worten: Wir erhalten

xxx

Da t eine symmetrische Verteilung ist, hätten wir dies auch wie folgt berechnen können

Xxx

Der Schlüssel hierzu ist, dass uobs aus einer Zufallsstichprobe stammt. Als solche ist sie eine Zufallsgröße und macht den p-Wert zu einer Zufallsvariablen. Tatsächlich ist der p-Wert Gleichförmig 0, 1 , d. h. er folgt einer Gleichverteilung auf dem Intervall 0, 1 . Aufgrund dieser Tatsache weist der p-Wert eine sehr hohe Streuung auf.

Es ist ein weit verbreiteter Irrglaube, dass der p-Wert dazu verwendet werden kann, zu bestätigen, dass die Alternativhypothese wahr ist. Stattdessen lässt ein kleiner p-Wert Zweifel an der Nullhypothese aufkommen und ermutigt die Forschenden, die Forschungsarbeit fortzusetzen. Ein weiteres Missverständnis ist, dass ein großer p-Wert darauf hinweist, dass der beobachtete Effekt auf einen Zufall zurückzuführen ist. Stattdessen ist es wichtig zu wissen, dass die Alternativhypothese immer noch zutreffen kann, selbst wenn ein hoher p-Wert ermittelt wurde. Ein großer p-Wert bedeutet nämlich nur, dass es keine Evidenz gegen die Nullhypothese gibt, und wir sollten dies nicht als Evidenz für das Nichtvorhandensein eines Effekts interpretieren.

Was geschieht, wenn zwei Studien, die denselben Test mit unterschiedlichen Daten wiederholen, denselben kleinen p-Wert aufweisen? Bedeutet dies, dass die Nullhypothese falsch ist? Nicht unbedingt, und wir können diese Schlussfolgerung nicht allein ziehen. Wir müssen die von diesen Forschenden angegebene Effektgröße berücksichtigen. Wenn sie nicht vergleichbar sind, dann führt ihr Wortlaut nicht zu derselben Schlussfolgerung. Und selbst wenn sich die Daten nicht allzu sehr unterscheiden, kann es sein, dass eine Studie aufgrund von Zufälligkeiten einen signifikanten Effekt feststellt, während die Replikationsstudie dies nicht tut. Vielleicht liegen die p-Werte so nahe an der Schwelle, dass die eine knapp darunter und die andere knapp darüber liegt.

Die Angabe von p-Werten hat sowohl einen Vorteil als auch einen Nachteil. Der Vorteil besteht darin, dass die Entscheidung, ob die Null abgelehnt wird, dem Leser überlassen werden kann, anstatt sie zu melden. Der Nachteil ist, dass p-Werte in der Praxis häufig falsch interpretiert werden. Daher sind Entscheidungen, die nur auf p-Werten basieren, bedeutungslos. Außerdem bedeutet ein kleiner p-Wert nicht, dass die Effektgröße praktisch signifikant ist.

Die wahrscheinlich wichtigste Eigenschaft jeder wissenschaftlichen Studie ist ihre Replizierbarkeit. Diese Qualität bildet die Grundlage dafür, wie sehr (oder wie wenig) der p-Wert unsere Entscheidung, die Nullhypothese abzulehnen oder nicht abzulehnen, beeinflusst. Der Grund, warum die Replizierbarkeit hier wichtig ist, ergibt sich aus der rechten Seite der Wahrscheinlichkeitsaussage. Das Ereignis U > uobs H0 setzt voraus, dass es theoretisch möglich ist, aus derselben Grundgesamtheit und mit demselben Stichprobenumfang wiederholt Daten zu entnehmen. Wenn dies der Fall ist, ist der p-Wert der langfristige Anteil dieser Stichproben, der zu einer Teststatistik führt, die mindestens so extrem ist wie die, die wir bei unserer einzigen Stichprobe beobachtet haben.

Die Probleme, die in der wissenschaftlichen Gemeinschaft bei der Berechnung und Interpretation von p-Werten bestehen, liegen in der Annahme der Replizierbarkeit. Wenn zum Beispiel Forschende eine Stichprobe sammeln und das Ergebnis p-Wert nicht so klein ist wie gewünscht, können sie immer mehr Daten sammeln, bis der p-Wert klein genug für die Veröffentlichung ist (dies ist als "p-Wert-Hacking" bekannt). Dies ist in zweierlei Hinsicht problematisch: Es verstößt gegen die Annahme der Replizierbarkeit der Stichprobengröße, da die Stichprobengröße nicht im Voraus festgelegt wurde, und die Teststärke nimmt ab! Es ist erwähnenswert, dass eine Teststatistik wie die unten angegebene tendenziell mit größeren Werten von n steigt, was wiederum uobs erhöht und den p-Wert verringert:

xxx

Das folgende Beispiel veranschaulicht eines der Probleme mit p-Werten.

#### Beispiel 4.4.2

Eine Schülerin beantwortet zehn (faktische) Richtig/Falsch-Fragen und hat sieben richtig und drei falsch beantwortet. Die Nullhypothese besagt, dass die Schülerin bei jeder Frage geraten hat. Berechnen Sie den p-Wert für zwei Szenarien:

1. Die Anzahl der Fragen, zehn, war im Voraus festgelegt und fixiert.

2. Der Schülerin werden wiederholt Fragen gestellt, bis sie drei davon falsch beantwortet hat.

#### Lösung

1. Der p-Wert ist die Wahrscheinlichkeit, einen so extremen Wert zu beobachten wie den, den wir beobachtet haben. In diesem Zusammenhang ist dies die Wahrscheinlichkeit, mindestens sieben richtige Antworten zu erhalten. Die Binomialverteilung ist hier angemessen. Sei U Binomial 10, 0 . 5 . Wir können die Binomialverteilung verwenden, weil (i) die Anzahl der Fragen (Versuche) festgelegt ist, (ii) die Fragen (Versuche) unabhängig sind und (iii) die Wahrscheinlichkeit, eine Frage richtig zu beantworten, von einem Versuch zum nächsten festgelegt ist (50 %). Der p-Wert beträt

xxx

2. In diesem Szenario folgt die Teststatistik U einer negativen Binomialverteilung U Neg-Binomial 3, 0 . 5 , die die Anzahl der Fragen modelliert, bis drei falsche Fragen beobachtet werden. Der p-Wert wird wie folgt berechnet

xxx

Wenn wir einen Schwellenwert von α = 0 . 05 verwenden, so ist im ersten Szenario die Nullhypothese nicht sehr zweifelhaft, im zweiten Szenario hingegen schon. Dies ist eine wichtige Überlegung, da wir in beiden Szenarien die gleichen Beobachtungsdaten verwendet haben!

Ein weiteres Problem bei sehr kleinen p-Werten ist, dass wir, selbst wenn alle Annahmen eines Tests erfüllt sind, darauf achten müssen, die praktische Signifikanz der beobachteten Effektgröße gegenüber der statistischen Signifikanz zu bewerten. Wir haben dies bereits bei Entscheidungen auf der Grundlage von Ablehnungsbereichen erwähnt, aber hier ist es noch wichtiger. Nehmen wir an, dass ein bestimmtes Verfahren das Leben von Krebspatienten im Durchschnitt um eine Woche verlängert. Nehmen wir weiter an, dieser Effekt sei statistisch signifikant mit einem sehr kleinen p-Wert. Bevor man eine solche Behandlung empfiehlt, sollte man überlegen, ob sie praktisch sinnvoll ist. Die statistische Theorie ist (selbst bei sehr kleinen p-Werten) nicht in der Lage, solche Entscheidungen zu treffen.

#### Beispiel 4.4.3

Berechnen Sie den p-Wert für Beispiel 4.3.3, in dem es um das durchschnittliche Gehalt von männlichen Managern im Vergleich zu weiblichen Managern ging.

#### Lösung

Wir erinnern uns, dass dieses Beispiel Hypothesen hatte: H0:μ1 = μ2 und H1:μ1 > μ2. Die Teststatistik lautet U = X - YSp116 + 116T 30 und uobs = 2 . 945. Der p-Wert ist gegeben durch

xxx

### Konfidenzintervalle

Konfidenzintervalle bieten einen Wertebereich zur Schätzung eines interessierenden Parameters. Im Gegensatz zu einer Punktschätzung (einem einzelnen Wert) berücksichtigt eine Intervallschätzung auch die mit der Schätzung verbundene Unsicherheit und ist daher einer Punktschätzung überlegen. Darüber hinaus haben Konfidenzintervalle, wie wir weiter unten sehen werden, die zusätzliche Stärke, dass sie uns helfen, Entscheidungen in Bezug auf Hypothesentests als Alternative zu Ablehnungsbereichen oder p-Werten zu treffen.

Wenn der Punktschätzer des interessierenden Parameters einer symmetrischen Verteilung folgt (Gauß- oder t-Verteilung), dann ist das zugehörige Konfidenzintervall ein Zufallsintervall der Form θ Å} ME, wobei θ der Schätzer des interessierenden Parameters θ ist und ME ein positiver Wert ist, der die Breite des Konfidenzintervalls bestimmt. Der Schätzer θ basiert auf einer Zufallsstichprobe und ist eine Zufallsvariable. Die Fehlermarge (ME) enthält Informationen über die Unsicherheit dieses Schätzers. Die Wahrscheinlichkeit, dass ein Konfidenzintervall den wahren Wert des Zielparameters enthält, wird als Konfidenzniveau (CL) bezeichnet. Die Beziehung zwischen CL und dem Signifikanzniveau α wird durch CL = 1 - α für zweiseitige Hypothesentests beschrieben. Das Konfidenzniveau wirkt sich auch auf die Fehlermarge aus. Für eine gegebene Unsicherheit wird die Breite des Konfidenzintervalls über dieME durch das Konfidenzniveau bestimmt. Je größer das Konfidenzniveau, desto größer die Fehlermarge und umgekehrt. Intuitiv würde man sagen, je größer der Korb (Konfidenzintervall) ist, desto wahrscheinlicher ist es, dass er den Ball (den wahren Wert des Zielparameters) fängt. Analog dazu gilt für ein gegebenes Konfidenzniveau: Je größer die Unsicherheit, desto größer die Fehlermarge.

Angenommen, X1, ...,Xn ist eine unabhängige Stichprobe aus einer Gauß-Verteilung N μ, σ mit unbekanntem μ und bekannter σ. Wir wollen ein Konfidenzintervall von 1 - α % für μ finden. Wir verwenden den Stichprobenmittelwert als unsere Punktschätzung. θ = X, ebenfalls eine Gauß-Verteilung mit demselben Mittelwert, aber skalierter Standardabweichung: X N μ, σ/ n . Wir wollen, dass unser Konfidenzintervall μ mit einer Wahrscheinlichkeit von 1 - α enthält. Anders ausgedrückt:

xxx

Dies ist gleichbedeutend mit

Xxx

oder

Xxx

Daher gilt MEσn= zα/2. Die Fehlermarge schließlich ist gegeben durch

Xxx

Unter den oben genannten Bedingungen für die Stichprobe ist das Konfidenzintervall daher gegeben durch

xxx

Sobald wir die beobachteten Werte der Stichprobe x1, ..., xn haben, können wir das beobachtete Konfidenzintervall angeben, indem wir einfach X durch x ersetzen:

xxx

Wenn wir ein Konfidenzintervall von 95 Prozent finden sollen, würden wir den zugehörigen kritischen z-Wert berechnen, z0 . 05/2 = z0 . 025 = 1 . 96, und unser Konfidenzintervall würde lauten

xxx

Für ein Konfidenzintervall von90 Prozent erhalten wir z0.01 = z0.05 = 2.58, so dass unser Konfidenzintervall wie folgt lauten würde

xxx

Nehmen wir nun an, dass die Bedingungen für die Stichprobe dieselben sind, nur dass die Standardabweichung σ unbekannt ist. Wir können σ2 durch S2, den Stichprobenvarianzschätzer , oder σ durch S = S2 ersetzen. In diesem Fall würde der Stichprobenmittelwert einer t-Verteilung mit ν = n - 1 Freiheitsgraden folgen: X T n - 1 . Der Rest der Analyse ist ähnlich, wobei zα/2 durch tn - 1, α/2 ersetzt wird. In diesem Fall würde das (beobachtete) Konfidenzintervall lauten

xxx

Bei einer Stichprobe mit dem Umfang n = 20 würde das 95-prozentige Konfidenzintervall beispielsweise den kritischen t-Wert t19, 0 . 025 = 2 . 09, so dass das Konfidenzintervall wie folgt lautet

xxx

#### Beispiel 4.4.4

Nehmen wir das Szenario aus Beispiel 4.1.3, bei dem es um das Geburtsgewicht in einer Stadt ging. Die Stichprobe von 12 Neugeborenen ergab ein mittleres Geburtsgewicht von 3250 Gramm und eine Standardabweichung von 250 Gramm. Konstruieren Sie ein Konfidenzintervall von 99 Prozent für den wahren Mittelwert der Neugeborenen. Verwenden Sie

t11, 0 . 005 = 3 . 11.

#### Lösung

Ein Konfidenzintervall von 1 - α 100% für den wahren Mittelwert μ ist gegeben durch

Xxx

mit n = 12, α = 0 . 01, s = 250 und x = 3250. Daher ergibt sich

xxx

oder 3025 . 56 ≤ μ ≤ 3474 . 44.

Konfidenzintervalle sind für sich genommen sehr nützlich, um die Genauigkeit der Punktschätzung zu beschreiben. Sie können jedoch auch verwendet werden, um Entscheidungen über die Nullhypothese zu treffen: H0:θ = θ0. Wenn wir einen zweiseitigen Test durchführen, H1:θ ≠ θ0, dann können wir die Nullhypothese bei einem Signifikanzniveau von α ablehnen, wenn θ0 nicht im Konfidenzintervall 1 - α 100% enthalten ist. Bei einem linksseitigen Test (H1:θ < θ0) wird die Nullhypothese abgelehnt, wenn θ0 größer als die obere Grenze des Konfidenzintervalls ist. Bei einem rechtsseitigen Test, H1:θ > θ0, wird die Nullhypothese abgelehnt, wenn θ0 kleiner als die untere Grenze des Konfidenzintervalls ist.

Zurückblickend auf Beispiel 4.1.3 hatten wir H0:μ = μ0 = 3480 und H1:μ < μ0 = 3480. Da es sich um einen linksseitigen Test handelt, lehnen wir die Nullhypothese ab, wenn μ0 = 3480 größer ist als die obere Grenze des Konfidenzintervalls. Aus dem Ergebnis von Beispiel 4.4.1 geht hervor, dass dies tatsächlich der Fall ist, da die obere Grenze des Konfidenzintervalls 3474,44 beträgt. Da das Konfidenzintervall mit einem Konfidenzniveau von 99 Prozent konstruiert wurde, wird die Entscheidung auf dem Niveau von α = 1 % getroffen.

Bisher haben wir Konfidenzintervalle für unbekannte Parameter einer einzelnen Grundgesamtheit diskutiert. Wir können auch Konfidenzintervalle für die Differenz von zwei Parametern aus unabhängigen Grundgesamtheiten konstruieren. Die Herleitung dieser Ergebnisse ist für den Fall einer einzigen Stichprobe identisch. Wir fassen die Formeln in der folgenden Tabelle zusammen.

Tabelle 25: Konfidenzintervalle für die Differenz von Parametern von zwei unabhängigen Grundgesamtheiten

|  |  |  |
| --- | --- | --- |
| Annahmen | Differenz | Fehlermarge (ME) |
| Große Stichproben |  |  |
| Gauß (σ1, σ2 bekannt) |  |  |
| Gauß (σ1 = σ2 unbekannt) |  |  |
| Gauß (σ1 ≠ σ2 unbekannt) |  |  |

Die Größe sp ist die gepoolte Standardabweichung, die wir in Abschnitt 4.3 erörtert haben, und νW ist der Freiheitsgrad des Welch-Tests.

#### Beispiel 4.4.5

Ermitteln Sie unter Bezugnahme auf Beispiel 4.3.3 ein Konfidenzintervall von 99 Prozent für die Differenz der Mittelwerte μ1 - μ2, wobei μ1 das durchschnittliche Jahresgehalt der männlichen Manager und μ2 das gleiche für die weiblichen Manager ist. Verwenden Sie das Konfidenzintervall, um eine Entscheidung über die Hypothesen H0:μ1 = μ2 und H1:μ1 > μ2 zu treffen.

#### Lösung

In Beispiel 4.3 haben wir angenommen, dass die unbekannten Varianzen der beiden Populationen gleich sind. Daher lautet das geeignete Konfidenzintervall

xxx

Aus der Lösung von Beispiel 4.3.3 ergibt sich sp ≈ 3001 . 67. Wir haben x = 150.000 und y = 125.000, also x - y = 25.000. Außerdem ist n1 = n2 = 16 und somit ν = 16 + 16 - 2 = 30. Da wir ein Konfidenzintervall von 99 Prozent berechnen wollen, ergibt sich α = 1 - 0 . 99 = 0 . 01 also α/2 = 0 . 005 und t30, 0 . 005 = 2 . 75. Setzt man dies alles zusammen, ergibt sich

xxx

Das Konfidenzintervall beträgt also

xxx

Um unsere Entscheidung bezüglich des Hypothesentests zu untermauern, ist die Differenz laut der Nullhypothese μ1 - μ2 = 0 kleiner als die untere Grenze, so dass wir die Nullhypothese mit einem Signifikanzniveau von einem Prozent ablehnen.

4.5 Multiple Tests

Bei multiplen Tests geht es darum, zwei oder mehr Nullhypothesen zu testen. Wenn wir jede der vielen Nullhypothesen separat testen wollen, erhöht sich die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen, und wir müssen Wege finden, um dies zu kontrollieren. In diesem Abschnitt wird zunächst erläutert, warum das Problem der Verstärkung des Fehlers 1. Art auftritt. Anschließend werden zwei Maßnahmen zur Fehlerkontrolle und deren Anwendung erörtert. Lassen Sie uns mit einem Beispiel beginnen.

Angenommen, 100 Schülerinnen und Schüler nehmen an einem Quiz mit zehn Fragen zu Fakten teil, die wahr oder falsch sind. Wir wollen herausfinden, ob die Schüler bei dem Quiz richtig geraten haben. In diesem Zusammenhang führen wir 100 Hypothesentests durch, einen für jeden Schüler. Nehmen wir an, dass wir für jeden Hypothesentest ein fünfprozentiges Signifikanzniveau wählen, d. h. α = 0 . 05. Die Nullhypothesen, H0

1 ,H0

2 , ...,H0

100 besagen jeweils, dass der betreffende Schüler beim Quiz zufällig geraten hat. Die Wahrscheinlichkeit, dass mindestens eine wahre Nullhypothese abgelehnt wird, beträgt 1 - 0 . 95100 = 0 . 994, was ziemlich hoch ist. Dies ist nicht sehr vielversprechend! Welchen Wert von α sollten wir für die einzelnen Tests wählen, damit die Wahrscheinlichkeit, mindestens eine wahre Nullhypothese abzulehnen, klein ist, beispielsweise 0,05? In unserem Fall mit 100 Hypothesen sollten wir α = 0 . 0005 setzen:

xxx

Diese Art von Fehler wird als Family-wise Error Rate (FWE, etwa „familienspezifische Fehlerquote“) bezeichnet. Die typische Methode zur Kontrolle dieser Art von Fehler ist die Bonferroni-Methode, die eine bestimmte Nullhypothese ablehnt, wenn der entsprechende p-Wert kleiner als α/m ist, wobei m die Anzahl der Hypothesen ist. In unserem Fall ist α/m = 0 . 05/100 = 0 . 0005. Dass wir diesen Wert aus unserer obigen Berechnung erhalten haben, sollte überraschen, denn wenn α klein und m groß ist, ergibt sich

1 - 1 - α 1/m ≈ α/m

Das Problem bei dieser Art der Kontrolle besteht darin, dass die Mehrfachtestverfahren zu einer geringen Teststärke führen können. Es sei daran erinnert, dass der Fehler 1. Art (in diesem Fall α/m) und der Fehler 2. Art (β) in einem umgekehrten Verhältnis zueinander stehen. Wenn also α/m zu klein ist, kann β unannehmbar groß sein, und in diesem Fall ist die Teststärke des Tests zu gering. Daher benötigen wir ein alternatives Fehlermaß zur Kontrolle.

Bei vielen Hypothesentests ist es sinnvoll, dass ein kleiner Teil der wahren Nullhypothesen abgelehnt werden kann. Auf diese Weise wird die Teststärke nicht zu gering. Zu diesem Zweck erörtern wir ein Maß, das den Anteil wahrer abgelehnter Nullhypothesen kontrolliert. Dazu benötigen wir vorab einige Größen:

- FP (false positives) ist die Anzahl der wahren Nullhypothesen, die abgelehnt werden.

- TP (true positives) ist die Anzahl der falschen Nullhypothesen, die abgelehnt werden.

- TN (true negatives) ist die Anzahl der wahren Nullhypothesen, die nicht abgelehnt werden.

- FN (false negatives) ist die Anzahl der falschen Nullhypothesen, die nicht abgelehnt werden.

- R ist die Gesamtzahl der abgelehnten Nullhypothesen: R = FP + TP

- m ist die Gesamtzahl der Nullhypothesen.

- m - R ist die Gesamtzahl der Nullhypothesen, die nicht abgelehnt werden.

- m0 ist die Gesamtzahl der wahren Nullhypothesen.

- m - m0 ist die Gesamtzahl der falschen Nullhypothesen.

Eine zweiseitige Tabelle ist eine gute Möglichkeit, sich diese Größem zu merken.

Tabelle 26: Zusammenfassung der Größen aus dem Ergebnis von multiplen Hypothesentests

|  |  |  |  |
| --- | --- | --- | --- |
|  | Wahre H0 | Falsche H0 | Insgesamt |
| H0 abgelehnt |  |  |  |
| H nicht abgelehnt0 |  |  |  |
| Insgesamt |  |  |  |

Bei einem (einzelnen) Hypothesentest möchten wir eine geringe Wahrscheinlichkeit haben, eine wahre Nullhypothese abzulehnen. Mit anderen Worten: Wir wollen eine geringe Wahrscheinlichkeit für ein falsches Positiv oder eine Falscherkennung. Wenn wir diese Interpretation nehmen und multiple Hypothesentests anwenden, wollen wir eine geringe Falsch-Positiv-Rate oder Falscherkennungsrate (FDR). Die FDR ist definiert als der erwartete Anteil der falsch-positiven Ergebnisse im Verhältnis zu allen positiven Ergebnissen:

xxx

Der Nachteil dieser Größe besteht darin, dass der Erwartungswert nicht berechnet werden kann, da die Zufallsvariablen FP und TP nicht beobachtbar sind. Als Abhilfe (für unkorrelierte oder positiv korrelierte Tests) wurde die Verwendung marginaler p-Werte vorgeschlagen, d. h. der p-Werte, die mit jeder Nullhypothese verbunden sind. Wir ordnen die Hypothesen in aufsteigender Reihenfolge hinsichtlich ihrer p-Werte neu an:

xxx

wobei die p-Werte in nicht abnehmender Reihenfolge stehen

xxx

Als Nächstes wählen wir die größte positive ganze Zahl k, so dass

Xxx

Dann werden die Hypothesen H0

i1 ,H0

i2 , ...,H0

ik (mit der neuen Reihenfolge) abgelehnt, was auf einen statistisch signifikanten Effekt hinweist.

### Zusammenfassung

In Abschnitt 4.1 haben wir den allgemeinen Rahmen der Hypothesentests als eine Möglichkeit erörtert, Aussagen über den Mittelwert oder den Anteil der Grundgesamtheit aus einer einzelnen Grundgesamtheit anhand einer beobachteten Stichprobe zu treffen. Abschnitt 4.2 verwendet den allgemeinen Rahmen der Hypothesentests und wendet ihn auf eine nicht-parametrische Umgebung an.

In Abschnitt 4.3 haben wir gelernt, wie man den Rahmen der Hypothesentests auf die Fälle anwendet, in denen wir analoge Parameter zweier Grundgesamtheiten vergleichen müssen. Die Interpretation dieser Tests sowie ihre Einschränkungen entsprechen immer noch denen der Einstichprobentests. Es ist wichtig, bei der Verwendung dieser Tests die zugrundeliegenden Annahmen zu beachten. Die Qualität eines Tests wird durch verschiedene Größen gemessen: Fehler 1. Art, Fehler 2. Art und die Teststärke eines Tests. Der p-Wert steht in engem Zusammenhang mit diesen Größen. Er hat einen Platz in der Interpretation, wird aber oft falsch interpretiert. Daher ist die Angabe des entsprechenden Konfidenzintervalls bei der Durchführung eines Hypothesentests wichtig. Diese Eigenschaften, der p-Wert und das Konfidenzintervall, sind Gegenstand der Diskussion in Abschnitt 4.4.

Schließlich haben wir in Abschnitt 4.5 dargelegt, warum wir möglicherweise multiple Tests durchführen müssen und welche Herausforderungen damit verbunden sind. Wir haben gelernt, wie wir einige dieser Herausforderungen mit Hilfe von zwei verschiedenen Maßnahmen zur Fehlerkontrolle angehen können. In dieser Einheit haben wir immer wieder darauf hingewiesen, dass, selbst wenn wir zu dem Schluss kommen, dass die Nullhypothese abgelehnt werden muss (was bedeutet, dass wir eine Wirkung festgestellt haben), dies allein noch nicht als Grundlage für Strategien, Empfehlungen oder Maßnahmen dienen kann. In realen Anwendungen müssen wir die Tests wiederholen und dann feststellen, ob die festgestellten Effektgrößen von praktischer Bedeutung sind.