Kardiotechnik 2023(3): 93–98 | https://doi.org/10.47624/kt.032.RUCI8451

TUTORIAL

Kardiotechnik stellt in der Rubrik Tutorials relevante Methoden für wissenschaftliche Arbeiten zur Perfusion vor.

Statistik Teil 7: Statistische Signifikanztests

M. Kohl
Prof. Dr. Matthias Kohl
Department of Medical and Life Sciences, Institute of Precision Medicine
Hochschule Furtwangen, Jakob-Kienzle-Str. 17
78054 Villingen-Schwenningen (Germany)
Phone: +49 (0) 7720 307-4635, E-Mail: kohl@hs-furtwangen.de
www.hs-furtwangen.de, www.life-data-science.org

M. Kohl, F. Münch

Fazitbox

Pro und Contra statistische Signifikanztests:

Pro

- Korrekt angewendet, stellen statistische Signifikanztests eine gute Möglichkeit dar, unter Einhaltung vorgegebener Fehlerwahrscheinlichkeiten (Forschungs-) Hypothesen zu prüfen bzw. zu falsifizieren.

Contra

- Wird die methodisch korrekte Vorgehensweise nicht eingehalten oder sind notwendige theoretische Voraussetzungen nicht erfüllt, ist nicht gewährleistet, dass die Fehlerwahrscheinlichkeiten beim statistischen Testen eingehalten werden. Entsprechend muss von höheren Wahrscheinlichkeiten für falsch positive, wie auch falsch negative Ergebnisse ausgegangen werden.
- Bei sehr großen Fallzahlen werden auch kleinste, nicht relevante Unterschiede statistisch signifikant.

Einführung

Statistische Tests gehören zu den am häufigsten verwendeten statistischen Verfahren. Sie spielen eine zentrale Rolle, um (Forschungs-)Hypothesen zu prüfen bzw. zu falsifizieren, weshalb man auch von Hypothesentests spricht. Die heute üblicherweise verwendeten statistischen Tests basieren auf der Theorie von Neyman und Pearson (1933) [1] und werden auch als Signifikanztests bezeichnet. Entscheidend für die Gültigkeit der gewonnenen Signifikanz ist hierbei eine methodisch korrekte Vorgehensweise. Da diese nicht immer gegeben ist und da statistische Tests, wie jedes statistische Verfahren, auch Schwächen haben, stehen diese seit einigen Jahren auch stark in der Kritik [2,3]. Im Folgenden werden wir die methodisch korrekte Vorgehensweise für Signifikanztests vorstellen und auf typische Fehler hinweisen. Außerdem werden wir die praktische Anwendung am Beispiel des t-Tests demonstrieren. Hierbei wird auch auf die dafür nötige Fallzahlplanung eingegangen.

Statistische Hypothesen

Damit eine (Forschungs-)Hypothese statistisch untersucht werden kann, muss diese entsprechend mit statistischen Ausdrücken formuliert werden. Insbesondere muss hierbei überlegt werden, welches statistische Modell geeignet ist, die vorliegende Situation zu beschreiben. Dies ist nicht ohne entsprechende fachlich-theoretischen Kenntnisse und Überlegungen möglich.

Man nennt die (Forschungs-)Hypothese im Kontext von Signifikanztests dann auch die **Alternative** oder **Alternativhypothese**, kurz H_1. Das Gegenteil hierzu wird **Nullhypothese**, kurz H_0, genannt. Die beiden Hypothesen sind demnach so konstruiert, dass sie sich gegenseitig ausschließen, womit nur eine der beiden Hypothesen wahr sein kann. Unser Ziel beim Signifikanztesten ist es, die Nullhypothese zu falsifizieren und so indirekt die Alternative zu bestätigen.

Als Beispiel betrachten wir die Troponinwerte (ng/ml) nach einer Operation mit Herz-Lungen-Maschine (HLM), wobei wir zwei Kardioplegieverfahren, nämlich die Kardioplegie nach Bretschneider (Custodiol) und die Mikroplegie, modifiziert nach Calafiore, vergleichen wollen. Unsere Forschungshypothese ist in diesem Fall, dass sich die post-operativen Troponinwerte für die beiden Verfahren unterscheiden. Wir gehen weiter davon aus, dass die Troponinwerte einer log-Normalverteilung folgen, d. h., die log-transformierten Troponinwerte folgen einer Normalverteilung. Wir nehmen daher an, dass die log-transformierten Troponinwerte nach Custodiol einer Normalverteilung mit Mittelwert μ_1 und Standardabweichung σ_1 folgen und die log-transformierten Troponinwerte der Mikroplegie einer Normalverteilung mit Mittelwert μ_2 und Standardabweichung σ_2. Wir nehmen außerdem an, dass die Standardabweichungen unterschiedlich sind ($\sigma_1 \neq \sigma_2$). Unsere Forschungshypothese, dass sich die post-operativen Troponinwerte unterscheiden, können wir somit statistisch wie folgt ausdrücken:

$$H_1: \mu_1 \neq \mu_2$$

Man spricht in diesem Fall auch von einer **zweiseitigen Hypothese**, da $\mu_1 < \mu_2$ oder $\mu_1 > \mu_2$ zutreffend sein könnte. Es ist auch möglich, **einseitige Hypothesen** zu testen. In diesem Fall wäre dies:

$$H_1: \mu_1 < \mu_2 \quad \text{oder} \quad H_1: \mu_1 > \mu_2$$

Die zweiseitige Testsituation lautet demnach:

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_1: \mu_1 \neq \mu_2$$

und die einseitigen Testsituationen wären:

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_1: \mu_1 < \mu_2$$

bzw.

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_1: \mu_1 > \mu_2$$

Man könnte in den einseitigen Fällen die Nullhypothesen auch

umformulieren zu H_0: $\mu_1 \geq \mu_2$ bzw. H_0: $\mu_1 \leq \mu_2$. Diese Umformulierung der Nullhypothese ändert jedoch den zugehörigen Signifikanztest nicht.

Die Entscheidung, ob man eine einseitige oder zweiseitige Alternative betrachtet, muss immer vor der Testdurchführung festgelegt werden. In der Medizin werden selten einseitige Alternativen verwendet. Zwar ist die (Forschungs-)Hypothese meist einseitig, z. B. ob eine neue Behandlung eine Verbesserung bringt, jedoch hätte eine Verschlechterung weitreichende Konsequenzen. Deshalb müssen aus ethischen Gründen und um die Sicherheit der Patient:innen zu gewährleisten, in der Regel zweiseitige Alternativen gewählt werden.

Am obigen Beispiel wird auch eine häufige Kritik gegenüber Signifikanztests deutlich. Es ist sicher in der Praxis aus verschiedensten Gründen nur sehr selten der Fall, dass zwei unterschiedliche Gruppen vollständig identische Werte haben [3]. D. h., die Nullhypothese H_0 ist in erster Linie eine künstliche Konstruktion und nur sehr selten tatsächlich wahr. Wird also im Rahmen eines Signifikanztests eine Entscheidung für die Nullhypothese getroffen, so geschieht dies eher aufgrund eines Mangels an Beweisen für die Alternative als wegen einer erwiesenen Richtigkeit der Nullhypothese.

Entscheidungssituation

Da wir Testentscheidungen auf Basis von Stichproben, die unvermeidbaren zufälligen Schwankungen unterliegen, treffen müssen, können wir Fehlentscheidungen nie völlig ausschließen. Es ist immer möglich, dass wir uns mit einer (hoffentlich) kleinen, aber in jedem Fall positiven Wahrscheinlichkeit falsch entscheiden werden. Im Fall der Signifikanztests ergibt sich die in Tabelle 1 dargestellte Entscheidungssituation.

	H_0 ist wahr	H_1 ist wahr
Entscheidung für H_0	**richtige Entscheidung 1-α** **richtig negativ**	**Fehler 2. Art β** **falsch negativ**
Entscheidung für H_1	**Fehler 1. Art α (Signifikanzniveau)** **falsch positiv**	**Richtige Entscheidung 1-β (Power)** **richtig positiv**

Tab. 1: Entscheidungssituation beim Signifikanztest

Die möglichen Fehlentscheidungen werden bei den Signifikanztests nach Neyman und Pearson (1933) [1] nicht in gleicher Weise behandelt. Die optimale Testfunktion wird bestimmt, indem der Fehler 2. Art unter einer Schranke an den Fehler 1. Art minimiert wird. Damit kann der Fehler 1. Art strikt kontrolliert werden und wird eine vorgegebene Fehlerwahrscheinlichkeit α, das sogenannte **Signifikanzniveau**, nicht überschreiten. Wir wissen außerdem, dass der Fehler 2. Art für die vorgegebene Situation kleinstmöglich bzw. die **Power** des Tests größtmöglich sein wird. Damit ist aber nicht automatisch sichergestellt, dass dieser Fehler tatsächlich klein bzw. die Power tatsächlich groß ist. Dies kann nur durch eine wohlüberlegte Fallzahlplanung erreicht werden. Der Fehler 2. Art wird nämlich insbesondere mit wachsender Fallzahl n kleiner und damit umgekehrt die Power entsprechend größer. Hierbei sollte man aber auch beachten, dass bei einem vorgegebenen Signifikanzniveau α mit wachsender Stichprobengröße n immer kleinere und damit oftmals auch irrelevante Unterschiede durch die Tests als signifikant identifiziert werden.

In praktischen Anwendungen wird üblicherweise ein Signifikanzniveau α von 5 % gewählt, aber auch kleinere Werte von 1 % oder 0,1 % kommen vor. Natürlich könnten auch Werte größer 5 % gewählt werden. Dies wäre jedoch sehr ungewöhnlich, man muss daher davon ausgehen, dass dies zu einer geringeren Akzeptanz der gewonnen Ergebnisse führen könnte. Im Fall der Ablehnung der Nullhypothese spricht man in Abhängigkeit vom Signifikanzniveau auch von einem **signifikanten** ($\alpha = 0{,}05$), **hoch** ($\alpha = 0{,}01$) oder **höchst** ($\alpha = 0{,}001$) **signifikanten** Ergebnis. Typische Werte für den Fehler 2. Art sind $\beta = 0{,}20$ bzw. $\beta = 0{,}10$, was einer Power von 80 % bzw. 90 % entspricht. In Fällen, in denen ein Fehler 2. Art schwerwiegende Konsequenzen hätte, kann β aber auch deutlich kleiner gewählt werden. So soll etwa die Wirkung eines neuen Medikaments, dessen Entwicklung hunderte Millionen Euro verschlingen kann [4], nur mit möglichst kleiner Wahrscheinlichkeit fälschlicherweise unerkannt bleiben.

Es sollte für die Praxis unbedingt beachtet werden, dass die statistische Signifikanz eine rein formale mathematische Konstruktion ist und damit nicht automatisch auch eine fachliche Relevanz anzeigt. Es ist demnach zwingend erforderlich, ein statistisch signifikantes Ergebnis auch immer hinsichtlich seiner fachlichen Relevanz zu hinterfragen. Der sogenannte Publikationsbias, der besagt, dass positive (d. h. statistisch signifikante) Ergebnisse deutlich häufiger publiziert werden als negative [5], hat leider dazu geführt, dass in der Forschung häufig nur noch nach statistischer Signifikanz „gejagt" wurde und wird und die Relevanz der Ergebnisse oftmals vernachlässigt wird [6]. Wir müssen daher davon ausgehen, dass viele, wenn nicht sogar die meisten publizierten Ergebnisse falsch positiv sind [7, 8]. Dieser Tatsache ist man sich heute durchaus bewusst, weshalb man sich vermehrt mit den Stärken und Schwächen des statistischen Testens auseinandersetzt [9, 10, 11, 12]. So wird heute etwa in vielen Leitfäden zur Berichterstattung (vgl. https://www.equator-network.org/) zusätzlich die Angabe von Konfidenzintervallen [13] gefordert.

Methodisch korrekte Vorgehensweise

Die beiden oben beschriebenen unvermeidbaren Fehler beim Testen sind aber nur dann unter Kontrolle, wenn methodisch korrekt vorgegangen wird. Dies geschieht, indem man die im Folgenden angegebene Reihenfolge der notwendigen Schritte einhält:

1. Formulierung der statistischen Hypothesen H_0 und H_1 (ein/zweiseitig?)
2. Auswahl eines geeigneten Tests.
3. Festlegung des gewünschten Signifikanzniveaus α und der angestrebten Power 1-β.
4. Berechnung der benötigten Fallzahl n (Fallzahlplanung).
5. Durchführung der Studie bzw. Erzeugung/Erhebung der Daten.
6. Anwendung des Tests auf die Daten und Entscheidung für H_0 oder H_1.

Im Punkt 2 wird üblicherweise der Test mit der größten Power für die vorgegebene Situation ausgewählt. Die Berechnung der benötigten Fallzahl in Schritt 4 erfordert neben der Festlegung von α und β auch eine sehr gute Abschätzung der zu erwartenden Ergebnisse (z. B. Effektstärke, Größe der Varianz). Hierauf wird im Abschnitt Fallzahlplanung noch genauer eingegangen werden. Mit Abschluss von Schritt 4 ist außerdem der Annahme- und Ablehnungsbereich der Nullhypothese H_0 festgelegt. Dies bedeutet, das statistische Verfahren ist fixiert und man weiß bereits vor Schritt 5, welches Testergebnis zu welcher Entscheidung führen wird. **Es ist methodisch entschieden abzulehnen, irgendeine Komponente der Schritte 1–4 nach Kenntnis der Daten noch zu verändern.** Nur so kann gewährleistet werden, dass der durchgeführte Signifikanztest die Fehlerwahrscheinlichkeiten einhält. In klinischen Studien wird dies durch die Forderung einer Registrierung der Studien unterstützt. Es kann damit nämlich geprüft werden, ob die berichteten Ergebnisse zur ursprünglichen Studienplanung passen und man daher davon ausgehen kann, dass die vorgegebenen Fehlerwahrscheinlichkeiten eingehalten wurden. Jedoch ist in vielen Bereichen eine Registrierung von Studien immer noch unüblich.

Die Testentscheidung in Schritt 6 geschieht in der Praxis üblicherweise nicht durch das Heranziehen des Annahme- und Ablehnungsbereichs der Nullhypothese H_0, sondern durch Berechnung des sogenannten p-Wertes. Es handelt sich hierbei um eine bedingte Wahrscheinlichkeit. Man nimmt an, dass die Nullhypothese H_0 wahr ist und berechnet unter dieser Bedingung die Wahrscheinlichkeit, dass das berechnete Testergebnis oder ein noch extremeres Ergebnis zustande kommt. Man stellt sich gewissermaßen die Frage, ob das berechnete Testergebnis bei Vorliegen der Nullhypothese H_0 eine hohe Wahrscheinlichkeit hätte. Leider wird der p-Wert oft missverstanden oder fehlinterpretiert [14]. Ist der p-Wert kleiner als das vorgegebene Signifikanzniveau α, wird H_0 abgelehnt. Ist der p-Wert größer oder gleich α, liegt keine ausreichende Evidenz für die Forschungshypothese H_1 vor, weshalb H_0 beibehalten werden muss. Wie oben bereits diskutiert, ist dies nicht gleichbedeutend damit, dass die Nullhypothese richtig und entsprechend die Forschungshypothese falsch ist [15]. Es gibt demnach nicht nur das Problem von falsch positiven Testergebnissen, sondern in ähnlichem Maße auch das Problem von falsch negativen Testergebnissen. Des Weiteren muss beachtet werden, dass sich das Ausmaß der Signifikanz nicht vom p-Wert, sondern vom gewählten Signifikanzniveau ableitet. D.h., wird ein Signifikanzniveau von 5 % gewählt und man erhält einen p-Wert von weniger als 0,01, oder sogar weniger als 0,001, so sollte das Testergebnis trotzdem nicht als hoch oder höchst signifikant bezeichnet werden, was in der Praxis leider häufig der Fall ist, **sondern lediglich als signifikant**.

Wir wollen die vorgestellte methodisch korrekte Vorgehensweise am obigen Beispiel zum Vergleich der beiden Kardioplegieverfahren Custodiol (CCC) und Mikroplegie (MBC) hinsichtlich der postoperativen Troponinwerte demonstrieren. Die statistischen Modelle und die Hypothesen wurden bereits formuliert. Wir wählen zwar die zweiseitige Alternative, erwarten aber, dass die MBC-Methode zu niedrigeren postoperativen Troponinwerten als die CCC-Methode führt.

Für die Planung unserer Studie ziehen wir eigene Daten einer Pilotstudie bestehend aus 65 Kindern im Alter von 0 bis 1 Jahren heran. Bei 42 Kindern wurde CCC und bei 23 MBC verwendet. Die qq-Plots in Abbildung 1, welche einen Vergleich zwischen den theoretischen Quantilen der Normalverteilung und den beobachteten Quantilen zeigen, sprechen dafür, dass sich die log10-transformierten postoperativen Troponinwerte tatsächlich gut durch eine Normalverteilung beschreiben lassen. Wir gehen daher davon aus, dass die obige statistische Formulierung der Hypothesen passend ist. In der Pilotstudie erhalten wir für die log10-transformierten postoperativen Troponinwerte (ng/ml) die folgenden Mittelwerte und Standardabweichungen: 1,45 ± 0,27 (CCC) und 1,24 ± 0,35 (MBC).

Als Signifikanztest wählen wir den Welch t-Test [19,20], welcher genau genommen nur ein näherungsweise optimaler Test für die vorliegende Situation ist, die auch als Behrens-Fisher Problem bekannt ist. Wir wählen weiterhin ein

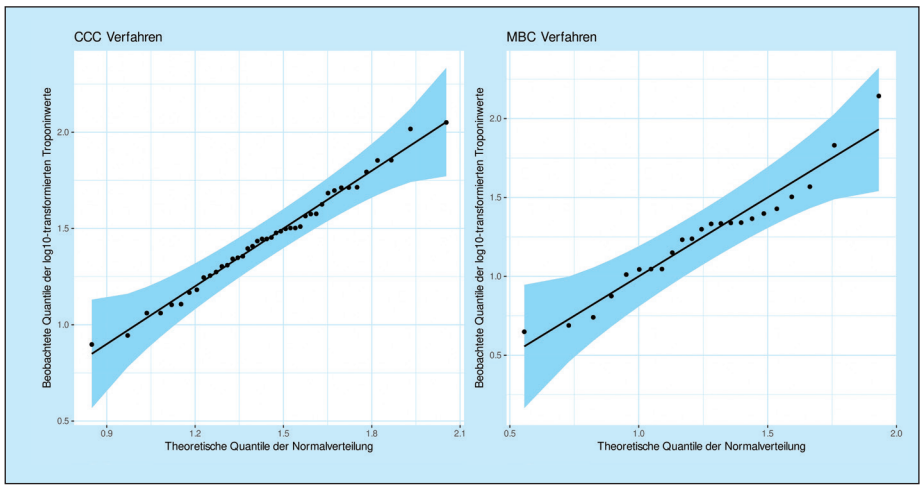

Abb. 1: qq-Plot der log10-transformierten postoperativen Troponinwerte (ng/ml) für 42 Patienten mit CCC-Verfahren und 23 Patienten mit MBC-Verfahren (erstellt mit der Statistiksoftware R [16] und den R Paketen ggplot2 [17] und qqplotr [18])

Signifikanzniveau von 5 % (α = 0,05) und eine Power von 90 % (β = 0,10). Wären die Standardabweichungen beider Gruppen gleich groß, wäre der Student t-Test der optimale Test. Da man davon ausgehen kann, dass die Standardabweichungen von zwei Gruppen in der Praxis nur sehr selten exakt gleich sein werden und da der Student t-Test bei ungleichen Standardabweichungen den Fehler 1. Art nicht einhält, sollte man in der Praxis auf die Anwendung des **Student t-Tests besser verzichten** und stattdessen auf Nummer sicher gehen und immer den **Welch t-Test** wählen [21, 22].

Fallzahlplanung

Wird ein Signifikanztest für die Fallzahlplanung herangezogen, so erhält man die Fallzahl, indem man die Powerfunktion des Tests gleich der gewünschten Power setzt und die entstehende Gleichung nach der Fallzahl n auflöst. Diese Gleichung enthält neben α und β aber auch einen oder mehrere weitere unbekannte Parameter (z. B. Mittelwert, Varianz, Erfolgswahrscheinlichkeit etc.). Die Werte dieser Parameter werden üblicherweise aus Pilotstudien oder der Literatur gewonnen. Die Fallzahl bzw. die Genauigkeit der Fallzahlberechnung hängt dabei entscheidend davon ab, wie gut diese unbekannten Parameter eingeschätzt werden. Liegen ungünstige Schätzungen vor, kann die benötigte Fallzahl, um die vorgegebene Power zu erreichen, auch deutlich unter- oder überschätzt werden. Daher sollten im Rahmen der Fallzahlplanung auch immer verschiedene Werte für diese Parameter zum Einsatz kommen und so die Sensitivität der Fallzahl gegenüber diesen Parametern untersucht werden.

In unserem Beispiel zu den Kardioplegieverfahren benötigen wir zusätzlich die Werte μ_1, σ_1, μ_2 und σ_2. Da die beiden Standardabweichungen in der Pilotstudie von ungefähr der gleichen Größenordnung sind, wählen wir zur Vereinfachung ein sogenanntes ausgewogenes Design. D. h. wir wählen die beiden Gruppen gleich groß. Unter Umständen kann man eine etwas kleinere Gesamtfallzahl erreichen, wenn man unterschiedlich große Gruppen erlauben würde. In diesem Fall würde sich für die Gruppe mit der größeren Standardabweichung eine etwas größere Fallzahl als für die Gruppe mit der kleineren Standardabweichung ergeben. Abbildung 2 zeigt die benötigte Fallzahl pro Gruppe, wobei wir $\mu_1 - \mu_2$, σ_1 und σ_2 in den Unterabbildungen etwas um die Werte aus der Pilotstudie variiert haben.

Wir sehen insbesondere, dass eine Verkleinerung der Mittelwertdifferenz bzw. eine Erhöhung der Standardabweichungen zu einer Erhöhung der Fallzahl führt. Um auf Nummer sicher zu gehen, empfiehlt es sich immer, die Ergebnisse aus einer (kleinen) Pilotstudie mit Vorsicht zu betrachten und die Parameter lieber etwas konservativer zu wählen. Zur Festlegung der endgültigen Fallzahl für unsere Beispielstudie wählen wir $\mu_1 - \mu_2 = 0{,}2$, $\sigma_1 = 0{,}3$ und $\sigma_2 = 0{,}4$. Dies führt uns auf eine Fallzahl von 67 Patienten pro Gruppe.

Testergebnisse

Ist die geplante Studie abgeschlossen, werden die Daten in die gewählte Testfunktion eingesetzt und der p-Wert für das Testergebnis berechnet. Leider wird in der Praxis häufig von der oben vorgestellten methodisch korrekten Vorgehensweise abgewichen. So ist es in vielen Bereichen immer noch üblich, dass Experimente oder Studien ohne eine formale Fallzahlplanung, geschweige denn generelle Überlegungen zur Fallzahl durchgeführt werden. Entsprechend ist damit auch die Power bzw. der Fehler 2. Art solcher Studien unklar. Eine nachträgliche Berechnung der Power unter Verwendung der Studienergebnisse, wie manchmal gefordert, ist methodisch nicht korrekt, und man muss davon ausgehen, dass das Ergebnis nicht der wahren Power entspricht [24]. Eine andere gängige Praxis besteht darin, die Daten zunächst hinsichtlich statistischer Annahmen zu testen und auf Basis der Ergebnisse dieser Tests einen geeigneten Test für die eigentliche Fragestellung zu wählen. So wird etwa ein Test auf Normalverteilung durchgeführt und falls dieser Test die Normalverteilungsannahme nicht ablehnt, ein t-Test zum Vergleich der Mittelwerte gewählt. Im Fall von zwei unabhängigen Gruppen wird dann in einem nächsten Schritt oft noch die Gleichheit der Varianzen getestet. Lehnt der Test die Gleichheit der Vari-

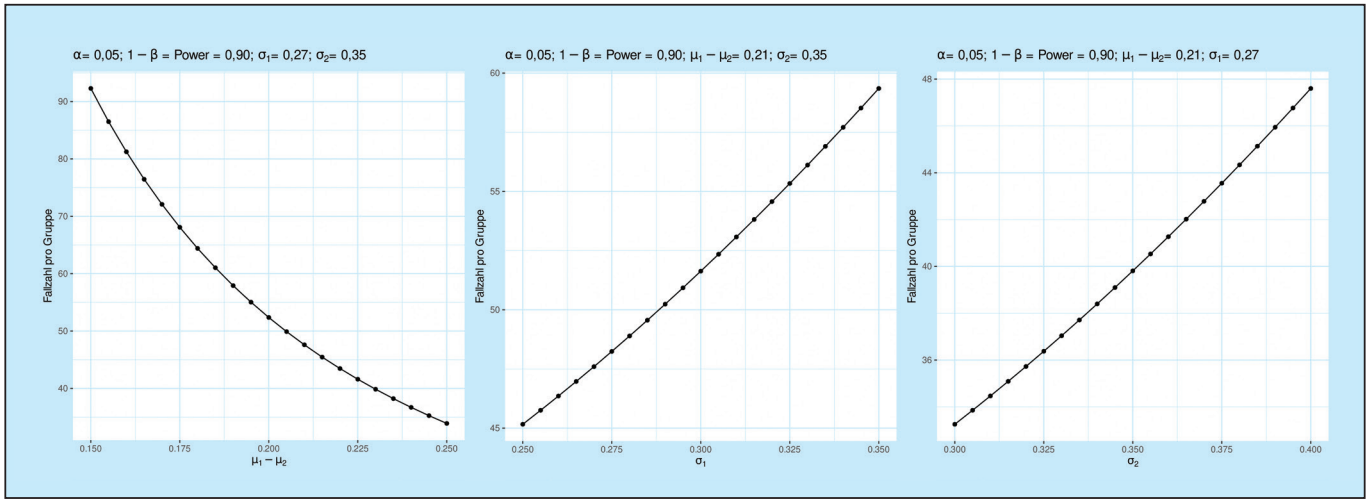

Abb. 2: Fallzahl pro Gruppe in Abhängigkeit der Parameter $\mu_1 - \mu_2$, σ_1 und σ_2 (erstellt mit der Statistiksoftware R [16] und den R Paketen ggplot2 [17] und MKpower [23])

anzen nicht ab, wird der Student t-Test durchgeführt, andernfalls der Welch t-Test. Diese Vorgehensweise ist methodisch gesehen äußerst fragwürdig und kann folglich auch nicht die Einhaltung der Fehlerwahrscheinlichkeiten gewährleisten [22]. Wie bereits erwähnt, ist generell von einer Verwendung des Student t-Tests eher abzuraten [21, 22].

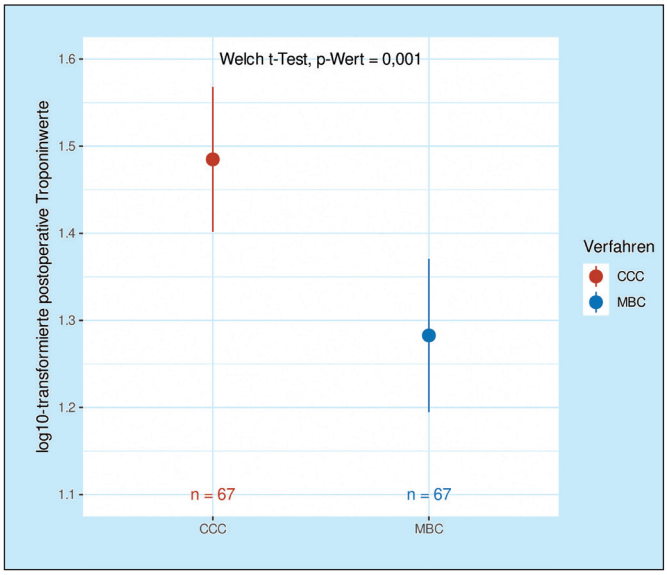

Abb. 3: Mittelwerte und 95 %-Konfidenzintervalle der log10-transformierten postoperartiven Troponinwerte für die Kardioplegieverfahren CCC und MBC und das Testergebnis für deren Vergleich mittels Welch t-Test (erstellt mit der Statistiksoftware R [16] und den R Paketen ggplot2 [17] und MKinfer [25])

Wir wollen abschließend das Testergebnis für unsere geplante Studie ermitteln. Natürlich handelt es sich beim gewählten Beispiel um eine hypothetische Studie zum Vergleich der beiden Kardioplegieverfahren CCC und MBC. Diese wurde faktisch nie so durchgeführt. Wir stellen diese (prospektive, randomisierte) Studie nach, indem wir aus einem vorliegenden, größeren Datensatz bestehend aus 542 Patient:innen (187 Patienten mit CCC, 355 Patienten mit MBC) zufällig 67 Patient:innen für jede Gruppe auswählen. Wir erhalten damit in unserer Beispielstudie einen p-Wert von 0,001 und somit einen signifikanten Unterschied für die Mittelwerte der log10-transformierten postoperativen Troponinwerte. Wie oben ausgeführt, sollten wir unser Ergebnis trotz des kleinen p-Wertes nicht als hoch signifikant bezeichnen, da wir ein Signifikanzniveau von 5 % gewählt hatten. Es ergeben sich außerdem die folgenden Gruppenmittelwerte und Standardabweichungen: 1,48 ± 0,34 (CCC) und 1,28 ± 0,36 (MBC). Die geschätzte Mittelwertdifferenz liegt folglich bei 0,20 und das 95 %-Konfidenzintervall (CI95) der Differenz der Mittelwerte beträgt 0,08–0,32. Die Mittelwerte und zugehörigen 95 %-Konfidenzintervalle für die beiden Gruppen sowie das Testergebnis sind in Abbildung 3 dargestellt.

Anstelle eines ein- oder zweiseitigen Signifikanztests zum Signifikanzniveau α kann in vielen Situationen auch ein äquivalentes ein- oder zweiseitiges (1-α)-Konfidenzintervall berechnet werden, für welches dann geprüft werden muss, ob der Wert der Nullhypothese im Intervall enthalten ist oder nicht. In unserem Beispiel muss demnach kontrolliert werden, ob eine Mittelwertdifferenz von 0 im CI95 enthalten ist. Da dies nicht der Fall ist, könnten wir auch mit Hilfe dieses Konfidenzintervalls die Nullhypothese zum Signifikanzniveau von 5 % ablehnen. Darüber hinaus ist das Konfidenzintervall informativer als die Angabe, ob ein signifikantes Ergebnis vorliegt oder nicht. Es zeigt zusätzlich an, welcher Effekt auf Basis der vorliegenden Daten möglich sein könnte. In unserem Beispiel liegt die geschätzte Mittelwertdifferenz bei 0,2. Auf Basis des 95 %-Konfidenzintervalls könnte jedoch auch eine kleinere Differenz von nur 0,08 bzw. eine größere Differenz von bis zu 0,32 möglich sein. Abschließend bleibt noch zu prüfen, ob eine Mittelwertdifferenz von 0,2 (CI95: 0,08–0,32) für die log10-transformierten postoperativen Troponinwerte einen klinisch relevanten Unterschied darstellt. Da ein log10-facher Unterschied von 0,2 einer relativen Veränderung von $10^{0,2}$ = 1,58 (CI95: 1,20–2,09) und somit einem Unterschied von mehr als 50 % entspricht und da außerdem das CI95 auf einen Unterschied von mindestens 20 % hindeutet, gehen wir auch von einer klinischen Relevanz des gefundenen signifikanten Unterschieds aus*.

Zusammenfassung

Die methodisch korrekte Vorgehensweise ist von entscheidender Bedeutung, um die beim statistischen Signifikanztest auftretenden Fehlerwahrscheinlichkeiten (Fehler 1. und 2. Art) zu kontrollieren. Gewisse Vorgehensweisen, wie die Auswahl des statistischen Tests auf Basis anderer statistischer Tests, sind methodisch abzulehnen und garantieren nicht die Einhaltung der Fehler 1. und 2. Art [22]. Um nachweisen zu können, dass man auch nach dem Beginn der Studie nicht von der ursprünglichen Planung abgewichen ist, sollte man seine Studie nach Möglichkeit in einer Datenbank für Studien (wie https://drks.de oder https://www.clinicaltrials.gov/) registrieren. Damit erhöht man unter anderem auch das Vertrauen in die Studienergebnisse. Des Weiteren sollte beim statistischen Testen beachtet werden, dass statistische Signifikanz und damit ein positives Studienergebnis nicht automatisch mit einer fachlichen (klinischen) Relevanz gleichzusetzen ist. Durch die zusätzliche Angabe von Konfidenzintervallen, wie in aktuellen Leitfäden zur Berichterstattung gefordert, kann die mögliche Bandbreite und damit die Größe und Relevanz des Effekts besser abgeschätzt werden. Diese Angabe ist insbesondere auch bei einem nicht signifikanten Test sehr hilfreich, da ein negatives Testergebnis in der Regel nicht mit der Gültigkeit der Nullhypothese, sondern mit einer mangelnden Evidenz für die Alternative gleichzusetzen ist.

* Achtung: Auch wenn reale Daten verwendet wurden, handelt es sich bei der beschriebenen Beispielstudie um keine reale Studie, weshalb die Ergebnisse zu Signifikanz und Relevanz auch keinesfalls als reale Ergebnisse für die beschriebene Situation aufgefasst werden sollten.

Literatur

1. Neyman J, Pearson, ES (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A. 231: 289-337.
2. Amrhein V, Greenland S, McShane B (2019). Retire statistical significance. Nature 567: 305-307.
3. McShane BB, Gal D, Gelman A, Robert C, Tackett JL (2019). Abandon statistical significance. The American Statistician, 73 (S1): 235-245.
4. Wouters OJ, McKee M, Luyten J (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA 323(9): 844-853.
5. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I (2010). Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess, 14(8): 1–193.
6. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015). The extent and consequences of p-hacking in science. PLOS Biology, 13(3): 1–15.
7. Ioannidis JPA (2005). Why most published research findings are false. PLOS Medicine, 2(8).
8. Colquhoun D (2014). An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci, 1(3):140216.
9. Cohen J (1994). The earth is round (p < .05). American Psychologist, 49(12): 997-1003.
10. Sterne JA, Davey Smith G (2001). Sifting the evidence-what's wrong with significance tests? BMJ, 322(7280): 226-231.
11. Wasserstein RL, Lazar NA (2016). The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2): 129-133.
12. Amrhein V, Korner-Nievergelt F, Roth T (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ, 5:e3544.
13. Kohl M, Münch F (2022). Statistik Teil 3: Konfidenzintervalle. Kardiotechnik 31(3): 95-98.
14. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 31(4): 337-350.
15. Altman DG, Bland JM (1995). Absence of evidence is not evidence of absence. BMJ 19;311(7003):485.
16. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
17. Wickham H (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.
18. Almeida A, Loy A, Hofmann H (2018). ggplot2 Compatible quantile-quantile plots in R. The R Journal, 10(2): 248-261. URL 248-261.
19. Welch BL (1938). The significance of the difference between two means when the population variances are unequal. Biometrika. 29 (3/4): 350-62.
20. Welch BL (1947). The generalization of „Student's" problem when several different population variances are involved. Biometrika. 34 (1-2): 28-35.
21. Ruxton GD (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4): 688-690.
22. Rasch D, Kubinger KD, Moder K (2011). The two-sample t test: pre-testing its assumptions does not pay off. Stat. Papers, 52: 219-231.
23. Kohl M (2023). MKpower: Power analysis and sample size calculation. R package version 0.7.
24. Hoenig JM, Heisey DM (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician 55: 19-24.
25. Kohl M (2023). MKinfer: Inferential Statistics. R package version 1.1.