

Die KARDIOTECHNIK stellt in der Rubrik Tutorials relevante Methoden für wissenschaftliche Arbeiten in der Perfusionologie vor.

Statistik Teil 8: t-Tests und Alternativen

Einführung

Die Analyse und der Vergleich von Mittelwerten gehören zu den am häufigsten durchgeführten statistischen Analysen. Allgemeiner geht es darum, die Lage (Lokation) der Werte von einer, zwei oder mehr Populationen zu bestimmen und mit einem Referenzwert bzw. untereinander zu vergleichen. In den meisten Fällen kommen hierfür statistische Signifikanztests zum Einsatz [1]. Man spricht in diesem Zusammenhang auch von 1-, 2- oder k-Stichproben Tests.

Wir werden uns im Folgenden genauer mit der 1- und 2-Stichproben-Situation beschäftigen und hierfür verschiedene Signifikanztests vorstellen sowie Kriterien angeben, die bei der Testauswahl helfen sollen. Neben verschiedenen t-Tests werden nachstehend Tests vorgestellt, die Permutationen oder Bootstrap nutzen. Des Weiteren werden auch sogenannte Rangtests mit einbezogen, die anstelle der beobachteten Werte „nur“ deren Rangfolge für die Berechnungen verwenden. Diese Tests werden an Beispielen genauer demonstriert.

1-Stichproben Tests

Im 1-Stichprobenfall vergleichen wir den Mittelwert (arithmetisches Mittel) einer Population μ mit einem Referenzwert μ_0 . Der Ausgangspunkt einer Studie ist demnach eine der folgenden Alternativen H_1 :

$$H_1: \mu < \mu_0 \quad \text{oder} \quad H_1: \mu > \mu_0 \quad \text{oder} \quad H_1: \mu \neq \mu_0$$

mit zugehöriger Nullhypothese $H_0: \mu = \mu_0$. Gehen wir davon aus, dass die Werte der Population einer Normalverteilung folgen, so führt uns dies auf den **1-Stichproben (Student) t-Test** als optimalen Test [2]. Ist es unklar, ob die Werte der Population wirklich einer Normalverteilung folgen, so kann anstelle des 1-Stichproben t-Tests eine **Permutations- oder Bootstrap-Variante** des t-Tests verwendet werden [3,4]. In beiden Fällen spricht man auch von **nichtparametrischen Tests**, da die genaue Verteilung der Daten nicht näher bekannt sein muss, sondern in diesem Fall mittels Permutationen bzw. Bootstrap approximiert wird. Eine weitere nichtparametrische Alternative zum 1-Stichproben t-Test ist der **Wilcoxon Vorzeichenrangtest**, der lediglich die Ränge der Werte für die Berechnungen heranzieht [5]. In diesem Fall muss ebenfalls die genaue Verteilung der Daten nicht näher bekannt sein. Es tritt jedoch an die Stelle des Mittelwertes der sogenannte **Pseudomedian**, der mittels des **Hodges-Lehmann Schätzers**



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 t-Tests und Alternativen:

Pro

- Kann man von normalverteilten Werten ausgehen und sind keine Ausreißer zu erwarten, so stellen der 1-Stichproben t-Test und der Welch t-Test die besten Tests dar.
- Permutations- und Bootstrap-Varianten der t-Tests sollten verwendet werden, falls die genaue Verteilung der Werte unbekannt ist und keine Ausreißer zu erwarten sind.

Contra

- Rechnet man mit Ausreißern, können Mittelwerte und darauf basierende t-Tests deutlich verfälschte Ergebnisse liefern. Dies gilt auch für deren Permutations- und Bootstrap-Varianten. In diesem Fall sollte man auf robuste Alternativen wie den Wilcoxon Vorzeichenrangtest oder den Wilcoxon-Mann-Whitney (WMW) Test ausweichen.

geschätzt werden kann [6]. Man bildet hierzu aus den vorliegenden Werten alle möglichen Paare und berechnet den Median der Mittelwerte aller dieser Paare. Der Pseudomedian unterscheidet sich in der Regel vom Mittelwert und auch vom Median. Liegt jedoch eine symmetrische (stetige) Verteilung vor, so sind Pseudomedian, Median und Mittelwert identisch. In diesem Fall kann der Wilcoxon Vorzeichenrangtest auch als Median- oder Mittelwert-Test angesehen werden. Der **Vorzeichen-Test**, welcher dem Binomialtest für eine Erfolgswahrscheinlichkeit von 50 % entspricht, stellt ebenfalls einen Test für den Median dar. Bei Vorliegen einer symmetrischen (stetigen) Verteilung könnte dieser auch als ein Mittelwert-Test angesehen werden. Da der Wilcoxon Vorzeichen-Test eine höhere Power besitzt, werden wir im Folgenden nicht weiter auf den Vorzeichentest eingehen.

Für die Auswahl des am besten geeigneten Tests für eine Studie sollten die folgenden Fragen herangezogen werden:

- 1) Ist mit vereinzelten Ausreißern zu rechnen?
- 2) Welche Größe der Stichprobe (Fallzahl) ist zu erwarten?
- 3) Kann von einer Normalverteilung ausgegangen werden?

Für ein methodisch korrektes Vorgehen müssen diese Fragen bereits während der Planung einer Studie beantwortet werden und nicht erst, wenn die Daten der Studie vorliegen [1].

Falls mit vereinzelten Ausreißern zu rechnen ist (1: ja), stellt der Wilcoxon Vorzeichenrangtest einen geeigneten Test dar, unabhängig davon, ob eine Normalverteilung vorliegt oder nicht (3: egal). Jedoch ist zu beachten, dass dieser Test für sehr kleine Stichproben nicht geeignet ist (2: nicht sehr klein). So ergibt sich etwa für eine Fallzahl $n \leq 5$ nie ein p -Wert $< 0,05$. Falls bei sehr kleinen Stichproben (2: sehr klein) zudem vereinzelte Ausreißer zu erwarten sind (1: ja), stellt keiner der oben genannten Tests eine gute Option dar. Liegen keine Ausreißer vor (1: nein) und ist die Verteilung unbekannt (3: unbekannt), so sind die Permutations- und die Bootstrap-Variante des t-Tests gute Optionen, wobei für kleine Stichproben (2: klein) der (exakte) Permutationstest dem Bootstrap-Test vorgezogen werden sollte. Rechnen wir nicht mit Ausreißern (1: nein) und können von einer Normalverteilung ausgehen (3: ja), so stellt der 1-Stichproben t-Test die beste Wahl dar, unabhän-

gig von der Größe der Stichprobe (2: egal). Dieses Vorgehen zur Auswahl eines geeigneten Tests ist in Abbildung 1 grafisch dargestellt. Eine weitere Möglichkeit ergibt sich, wenn man von einer großen Stichprobe (2: groß) ausgeht und keine Ausreißer (1: nein) erwartet. In diesem Fall kann der 1-Stichproben t-Test sogar dann verlässliche Ergebnisse liefern, wenn die Werte der Population nicht normalverteilt sind (3: unbekannt). Dies liegt am sogenannten **zentralen Grenzwertsatz**, welcher besagt, dass sich die Verteilung des Mittelwertes unter recht allgemeinen Voraussetzungen einer Normalverteilung annähert. Dabei gilt, je größer die Stichprobe ist, umso größer dürfen die Abweichungen von der Normalverteilung sein. Bei Abweichungen von der Symmetrie (d. h. schiefen Verteilungen) sollte man sich jedoch immer fragen, ob der Mittelwert einen guten Referenzwert darstellt und man darauf seine statistischen Schlussfolgerungen basieren sollte.

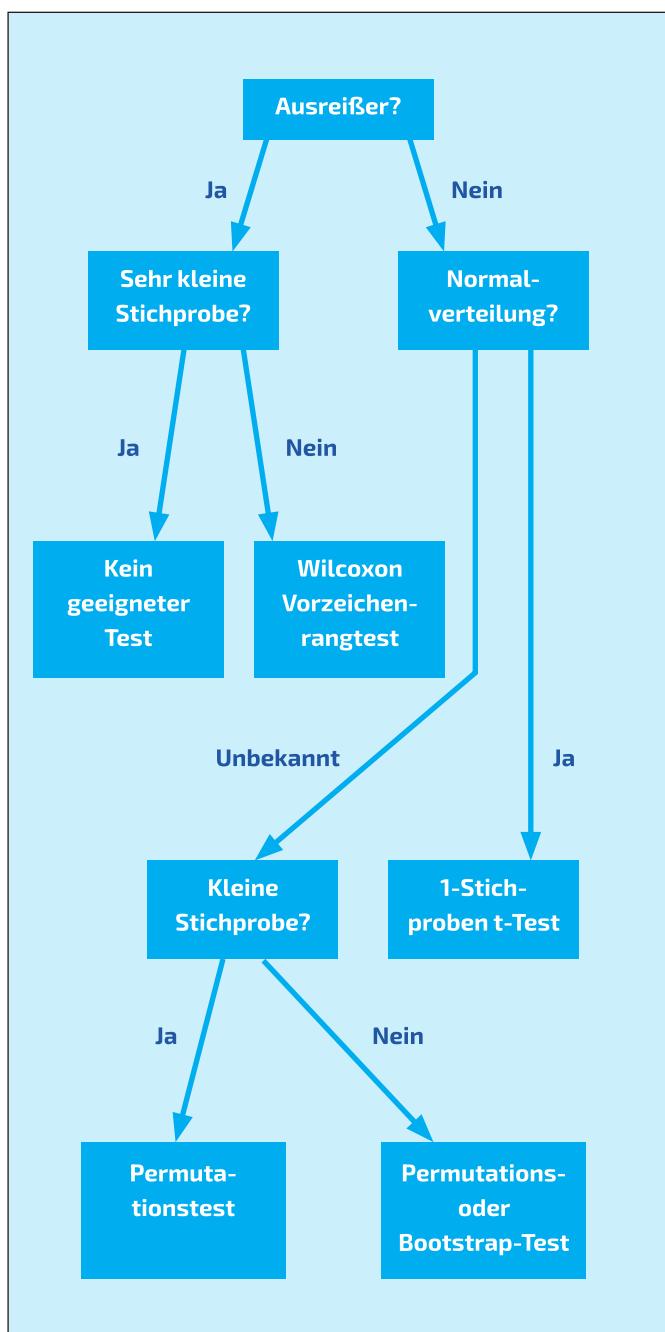


Abb. 1: Auswahl eines geeigneten Tests im 1-Stichprobenfall

Wir betrachten Daten zur venösen Sauerstoffsättigung (SvO_2) von 30 Patient:innen (eigene Daten) [7]. Zur Demonstration, dass sich bereits einzelne Ausreißer sehr negativ auf manche statistische Verfahren auswirken können, wurde in einer zweiten Variante des Datensatzes der kleinste Wert von 63,8 % auf 53,8 % reduziert. Wir gehen von einer Normalverteilung der Daten aus, bzw. von einer symmetrischen Verteilung, die einer Normalverteilung recht ähnlich ist. Bei kleinen bis mittelgroßen Stichproben sind derartige Verteilungsannahmen aber nur schwer zu verifizieren. Wir sehen anhand der Abbildung 2, dass es gewisse Abweichungen von einer idealen Normalverteilung gibt und dass eine leicht linksschiefe Verteilung vorliegt. Es ist aber nicht zu erkennen, ob diese Abweichungen auf Zufallsschwankungen beruhen oder einer Systematik folgen. Auch ist der künstlich eingefügte Ausreißer im Box- und Whisker-Plot deutlich sichtbar.

Wir untersuchen die einseitige Alternative, dass die mittlere venöse Sauerstoffsättigung der Patient:innen größer als 75 % ist (d. h. $H_1: \mu > 75\%$) und verwenden ein Signifikanzniveau von 5 %. Für die Berechnungen verwenden wir die Statistiksoftware R [10] in Kombination mit dem Paket stats [10] für den 1-Stichproben t-Test, dem Paket exactRank Tests [11] für den Wilcoxon Vorzeichenrangtest und dem Paket MKinfer für die Permutations- und Bootstrap-Variante des t-Tests [12]. Im Fall der Originaldaten erhalten wir mit allen vier oben vorgestellten Tests signifikante und ähnliche Ergebnisse (1-Stichproben t-Test: $p = 0,017$; Wilcoxon Vorzeichenrangtest: $p = 0,017$; Permutationstest: $p = 0,018$; Bootstrap-Test: $p = 0,029$), weshalb wir uns für die Alternative entscheiden. Dies spiegelt sich auch in den zugehörigen, einseitigen 95 %-Konfidenzintervallen wider (1-Stichproben t-Test: 75,5 %–100,0 %; Wilcoxon Vorzeichenrangtest: 75,8 %–100,0 %; Permutationstest: 75,7–100,0 %; Bootstrap-Test: 75,6 %–100,0 %), welche alle bei Werten jenseits von 75,0 % starten. Die Ähnlichkeit der Ergebnisse ist außerdem ein Indiz dafür, dass die Annahme einer näherungsweise Normalverteilung zutreffend sein könnte. Wir wiederholen die Analyse mit dem Datensatz, bei dem der kleinste Zahlenwert zu einem Ausreißer verfälscht wurde. Dies führt dazu, dass der 1-Stichproben t-Test sowie die Permutations-

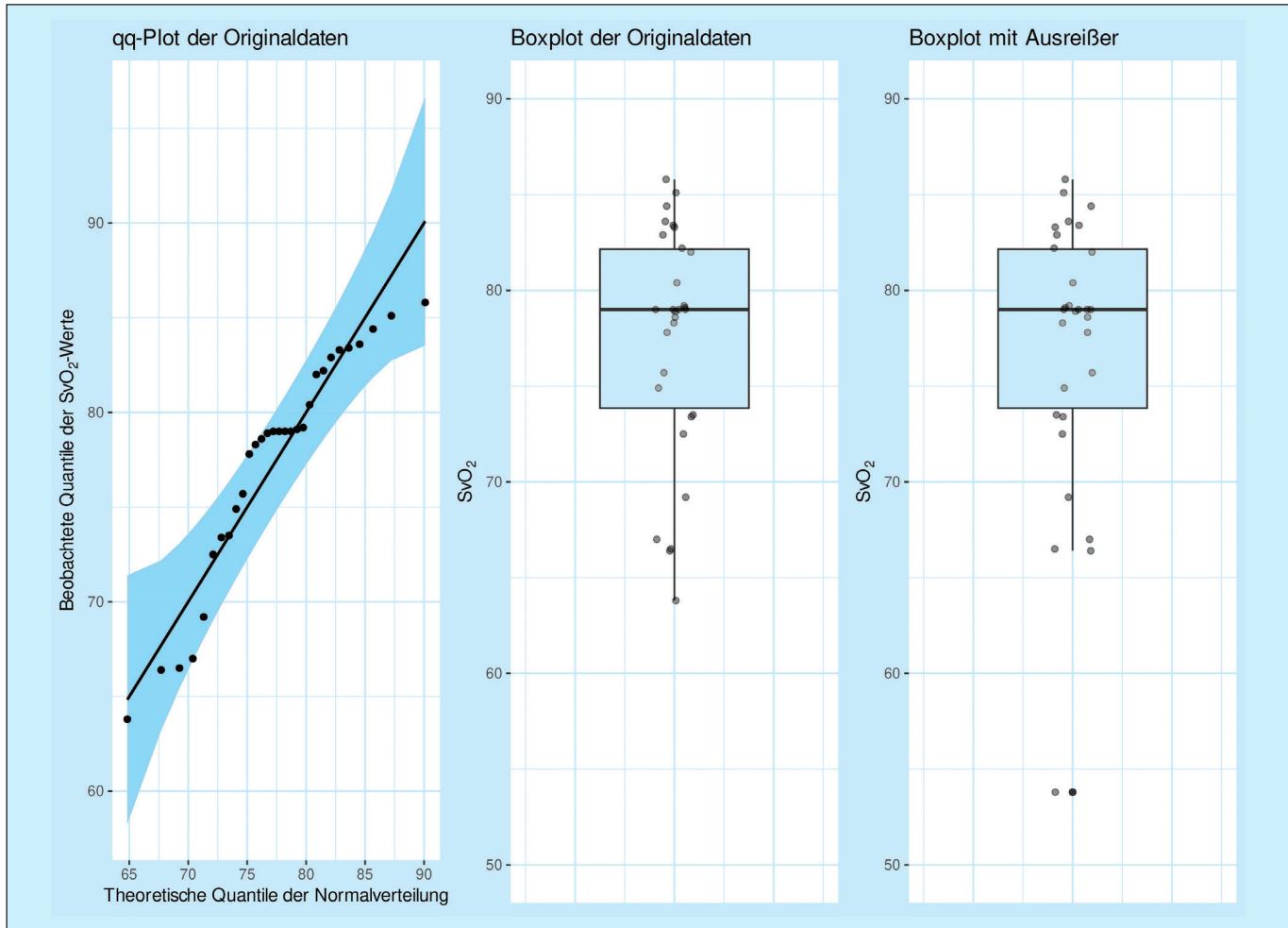


Abb. 2: qq-Plot und Boxplots der venösen Sauerstoffsättigung (SvO₂) von 30 Patient:innen (erstellt mit den Paketen ggplot2 [8] und qqplotr [9] der Statistiksoftware R [10])

und Bootstrap-Variante des t-Tests keine signifikanten Ergebnisse mehr liefern (1-Stichproben t-Test: $p = 0,054$; Permutationstest: $p = 0,055$; Bootstrap-Test: $p = 0,093$). Im Fall des Wilcoxon Vorzeichenrangtests erhalten wir im Gegensatz dazu weiterhin ein signifikantes Ergebnis, wobei der p-Wert sogar unverändert ist, da wir gerade den kleinsten Wert der Stichprobe durch den Ausreißer ersetzt haben. Wir sehen, dass im ungünstigsten Fall bereits ein einzelner Ausreißer das Ergebnis einer statistischen Analyse deutlich verfälschen kann. Wird erwartet, dass bei einer Studie Ausreißer auftreten, empfiehlt es sich daher, immer auf statistische Verfahren zurückzugreifen, die robust gegenüber Ausreißern sind.

2-Stichproben Tests

Im 2-Stichprobenfall geht es darum, die Mittelwerte μ_1 und μ_2 von zwei Gruppen zu vergleichen. Es gibt demnach die folgenden Alternativen H_1

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

mit zugehöriger Nullhypothese $H_0: \mu_1 = \mu_2$ (siehe auch [1]). Die beiden Gruppen können hierbei abhängig oder unabhängig sein. Im Fall von abhängigen Gruppen spricht man auch von einem **gepaarten Test**. Typischerweise liegt dieser Fall vor, wenn von einer Person Wertepaare also z. B. zwei Messungen in einem zeitlichen Abstand vorliegen, wie etwa vor und nach einer Operation. Die gepaarte Situation lässt sich auf den 1-Stichproben-

fall zurückführen, indem man die Differenz der Wertepaare betrachtet und für diese Differenz den entsprechenden 1-Stichproben Test berechnet. Die möglichen Tests und deren Auswahl sind folglich analog zum 1-Stichprobenfall.

Wir betrachten im Folgenden den Fall unabhängiger Gruppen genauer. Gehen wir davon aus, dass die Werte der beiden Gruppen einer Normalverteilung folgen, so kommen als mögliche Signifikanztests der **2-Stichproben Student t-Test** (beide Gruppen besitzen die gleiche Varianz) [2] oder der **Welch t-Test** (die Varianzen der beiden Gruppen sind verschieden) [13,14] in Frage. Wie bereits in [1] genauer ausgeführt, sollte man in der Praxis auf Nummer sicher gehen und immer von unterschiedlichen Varianzen (**Heteroskedastizität**) ausgehen, d. h. den Welch t-Test verwenden. Ist es unklar, ob die Werte der beiden Gruppen wirklich Normalverteilungen folgen und nehmen wir weiter an, dass die Varianzen der beiden Gruppen verschieden sind, so kann man eine **Permutations-** oder eine **Bootstrap-Variante** des Welch t-Tests verwenden [15,16,4]. Eine weitere Alternative ist der **Wilcoxon-Mann-Whitney (WMW)** Test, der auch als **Wilcoxon Rangsummentest** [5] und **Mann-Whitney U-Test** [17] bekannt ist. Unter der Annahme, dass die Verteilungen der beiden Gruppen die gleiche Form besitzen und sich nur durch eine Verschiebung in der Lokation unterscheiden, liefert der Test einen Vergleich der Mediane [18]. Der Schätzer für die Gruppenunterschiede, welcher zum WMW Test gehört, ist der

Hodges-Lehmann Schätzer. Es handelt sich hierbei um den Median aller paarweisen Differenzen der Werte beider Gruppen. Die Annahme gleicher Verteilungsform impliziert insbesondere, dass auch die Varianzen der beiden Gruppen gleich sind. Simulationsstudien zeigen, dass zumindest kleine Unterschiede bei den Varianzen noch akzeptabel sind [19,20]. Insgesamt schränkt die Annahme gleicher Verteilungsformen aber die Anwendbarkeit des WMW Tests als Mediantest stark ein [18]. Allgemeiner, ohne dass die Annahme gleicher Verteilungsformen benötigt wird, untersucht der WMW Test die Frage, ob die Werte der einen Gruppe stochastisch größer sind als die Werte der anderen Gruppe [17].

Für die Auswahl des am besten geeigneten Tests für eine Studie sollten die folgenden Fragen herangezogen werden, die ähnlich zum 1-Stichprobenfall sind und bereits während der Planung einer Studie beantwortet werden sollten [1]:

- 1) Ist mit vereinzelten Ausreißern zu rechnen?
- 2) Welche Gruppengrößen (Fallzahlen) sind zu erwarten?
- 3) Kann von Normalverteilungen ausgegangen werden?
- 4) Sind die Verteilungsformen für beide Gruppen gleich?
- 5) Sind die Varianzen beider Gruppen gleich (**Homoskedastizität**)?

Wir gehen auf Nummer sicher und gehen deshalb immer davon aus, dass die Varianzen beider Gruppen nicht gleich sind (5: nein). Dies impliziert, dass auch die Verteilungsformen beider Gruppen ungleich sind (4: nein).

Falls mit vereinzelten Ausreißern zu rechnen ist (1: ja), stellt der WMW Test einen geeigneten Test dar, unabhängig davon, ob Normalverteilungen vorliegen oder nicht (3: egal). Jedoch ist zu beachten, dass dieser Test für sehr kleine Stichproben nicht geeignet ist (2: nicht sehr klein) und nur dann einen Median-test liefert, falls die Verteilungsformen (inkl. Varianzen) beider Gruppen zumindest sehr ähnlich sind (4 + 5: zumindest näherungsweise). Falls bei sehr kleinen Stichproben (2: sehr klein) zudem vereinzelte Ausreißer zu erwarten sind (1: ja), stellt keiner der oben genannten Tests eine gute Option dar. Liegen keine Ausreißer vor (1: nein) und ist die Verteilung unbekannt (3: unbekannt), so sind die Permutations- und die Bootstrap-Variante des Welch t-Tests gute Optionen, wobei für kleine Stichproben (2: klein) der (exakte) Permutationstest dem Bootstrap-Test vorgezogen werden sollte. Rechnen wir nicht mit Ausreißern (1: nein) und können von Normalverteilungen ausgehen (3: ja), so stellt der Welch t-Test die beste Wahl dar, unabhängig von der Größe der Stichprobe (2: egal). Dieses Vorgehen zur Auswahl eines geeigneten Tests ist in Abbildung 3 nochmals grafisch dargestellt. Simulationsstudien zeigen außerdem, dass der Welch t-Test sogar unter Abweichungen von der Normalverteilung gute Ergebnisse liefert [21], was zumindest zum Teil sicher auch durch den zentralen Grenzwertsatz zu erklären ist.

Zur Demonstration betrachten wir die Troponinwerte (ng/ml) nach einer Operation mit Herz-Lungen-Maschine (HLM), wobei wir zwei Kardioplegieverfahren, nämlich die Kardioplegie nach Bretschneider (CCC) und die Mikroplegie, modifiziert nach Calafiore (MBC), vergleichen wollen. Wir haben hierzu aus ei-

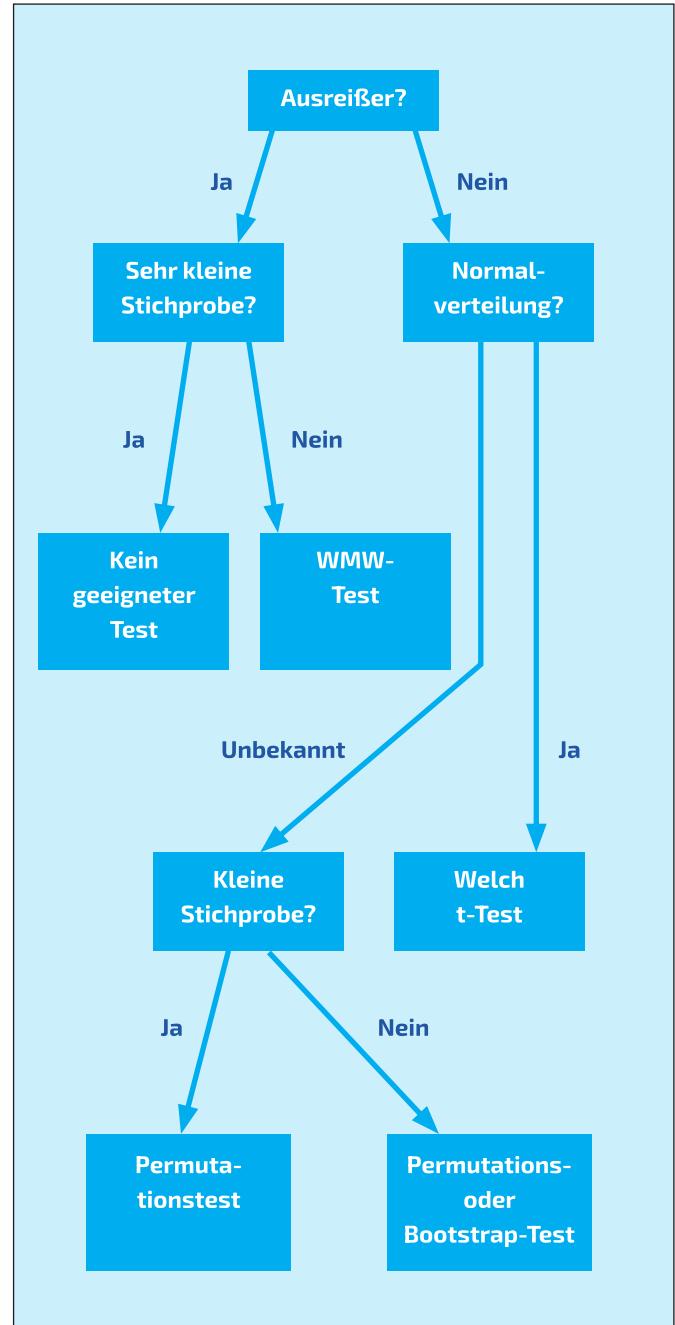


Abb. 3: Auswahl eines geeigneten Tests im 2-Stichprobenfall unter der Annahme, dass die Varianzen verschieden sind (Heteroskedastizität)

nem vorliegenden größeren Datensatz, bestehend aus 542 Patient:innen (187 Patient:innen mit CCC, 355 Patient:innen mit MBC) zufällig 67 Patient:innen für jede Gruppe ausgewählt. Wir gehen weiter davon aus, dass die Troponinwerte zumindest näherungsweise einer log-Normalverteilung folgen, wobei wir von ähnlichen Varianzen und ähnlichen Verteilungsformen ausgehen. Wir vermuten, dass sich die postoperativen Troponinwerte unterscheiden und verwenden einen zweiseitigen Test zum Signifikanzniveau von 5 %, um dies genauer zu untersuchen, wobei wir die Tests auf die log10-transformierten Troponinwerte anwenden. Für weitere Einzelheiten verweisen wir auf [1]. Für die Berechnungen verwenden wir die Statistiksoftware R [10] in Kombination mit dem Paket stats [10] für den Welch t-Test, dem Paket exactRank Tests [11] für den WMW-Test und dem Paket MKinfer für die Permutations- und Bootstrap-Variante des Welch t-Tests [12]. Wir erhalten für alle

vier vorgestellten Tests ein signifikantes Ergebnis mit ähnlichen p-Werten (Welch t-Test: $p = 0,0011$, WMW Test: $p = 0,0005$, Permutationstest: $p = 0,0014$, Bootstrap-Test: $p = 0,0014$) und ähnlichen 95 %-Konfidenzintervallen (Welch t-Test: 0,082–0,322; WMW-Test: 0,098–0,306; Permutationstest: 0,079–0,326; Bootstrap-Test: 0,084–0,322). Die postoperativen Troponinwerte sind demnach bei CCC signifikant höher im Vergleich zur Gruppe MBC. Die etwas größeren p-Werte und die etwas längeren 95 %-Konfidenzintervalle im Fall vom Welch t-Test sowie dem Permutations- und Bootstrap-Test sind vermutlich auf die Ausreißer zurückzuführen, die in Abbildung 4 zu sehen sind, und zu einer Vergrößerung der Varianz führen.

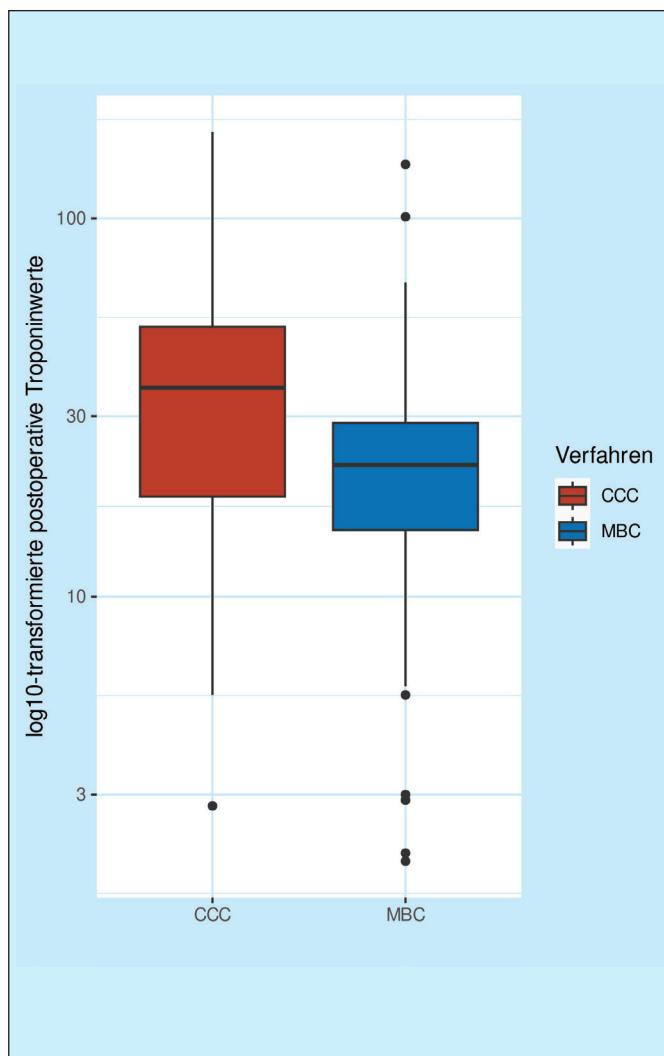


Abb. 4: Boxplots der log10-transformierten postoperativen Troponinwerte (ng/ml) für 67 Patient:innen mit CCC-Verfahren und 67 Patient:innen mit MBC-Verfahren (erstellt mit dem Paket ggplot2 [8] der Statistiksoftware R [10])

Da streng monotone Transformationen die Ränge nicht verändern, könnten wir im Fall des WMW Tests auf die log10-Transformation verzichten, da dies zu einer identischen Teststatistik und damit einem identischen p-Wert führt. Es hätte zudem den Vorteil, dass das 95 %-Konfidenzintervall die Werte auf der Originalskala zeigt. Wir erhalten so als 95 %-Konfidenzintervall für den Median der Differenzen (Hodges-Lehmann-Schätzer): 5,60–18,60 ng/ml.

Zusammenfassung

Da die Annahme einer Normalverteilung eine recht starke Voraussetzung darstellt, sollten in der Praxis verstärkt Permutations- und Bootstrap-Varianten der t-Tests zum Einsatz kommen [19], wobei auch der Welch t-Tests recht robust gegenüber Abweichungen von der Normalverteilung ist [19,21]. Beim Vorliegen von Ausreißern können sowohl die t-Tests als auch deren Permutations- und Bootstrap-Varianten verfälschte Ergebnisse liefern. In diesem Fall sollte auf robuste Tests, wie den Wilcoxon Vorzeichenrangtest und den Wilcoxon-Mann-Whitney Test ausgewichen werden. Dabei ist zu beachten, dass der Wilcoxon-Mann-Whitney Test nur bei Vorliegen gleicher Verteilungsformen einen Mediantest liefert und andernfalls allgemeiner untersucht, ob die Werte der einen Gruppe stochastisch größer sind als die Werte der anderen Gruppe. Bei sehr kleinen Stichproben (ohne Ausreißer) bleiben als mögliche Wahl nur die t-Tests, wobei man in diesem Fall vom Vorliegen einer Normalverteilung ausgehen muss und dies nicht sinnvoll geprüft werden kann. Außerdem sollte man sich in diesem Zusammenhang bewusst sein, dass die Power von sehr kleinen Studien nur gering ist und bei signifikanten Ergebnissen der tatsächliche Effekt meist über-schätzt wird [22].

Literatur

1. Kohl M, Münch F (2023). Statistik Teil 7: Statistische Signifikanztests. Kardiotechnik, 2023(3): 93-98.
2. Student (1908). The Probable Error of a Mean. Biometrika, 6(1): 1-25.
3. Pitman EJG (1937). Significance tests which may be applied to samples from any population. Royal Statistical Society Supplement, 4: 119-130 and 225-32 (parts I and II).
4. Efron B, Tibshirani R (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC.
5. Wilcoxon F (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6): 80-83.
6. Hodges JL, Lehmann EL (1963). Estimation of location based on ranks. Annals of Mathematical Statistics, 34(2): 598-611.
7. Kohl M, Münch F (2022). Statistik Teil 1: Der Box- und Whisker-Plot. Kardiotechnik 31(1): 15-17.
8. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
9. Almeida A, Loy A, Hofmann H (2018). ggplot2 Compatible Quantile-Quantile Plots in R. The R Journal, 10(2): 248-261. URL 248-261.
10. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
11. Hothorn T, Hornik K (2022). exactRankTests: Exact Distributions for Rank and Permutation Tests. R package version 0.8-35.
12. Kohl M (2023). MKinfer: Inferential Statistics. R package version 1.2.
13. Welch BL (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29(3/4): 350-62.
14. Welch BL (1947). The generalization of "Student's" problem when several different population variances are involved. Biometrika, 34(1-2): 28-35.
15. Janssen A (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. Statistics and Probability Letters, 36: 9-21.
16. Chung E, Romano JP (2013). Exact and asymptotically robust permutation tests. The Annals of Statistics, 41(2): 484-507.
17. Mann HB, Whitney DR (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18(1): 50-60.
18. Divine GW, Norton HJ, Barón AE, Juarez-Colunga E (2018). The Wilcoxon-Mann-Whitney Procedure Fails as a Test of Medians. The American Statistician, 72(3): 278-286.
19. Skovlund E, Fenstad GU (2001). Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? Journal of Clinical Epidemiology, 54(1): 86-92.
20. Dwivedi AK, Mallawaarachchi I, Alvarado LA (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. Statist. Med., 36: 2187-2205.
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. Ioannidis JPA (2008). Why Most Discovered True Associations Are Inflated. Epidemiology, 19(5): 640-648.