Tutorial

KARDIOTECHNIK 2022; 031(3):095–098
https://doi.org/10.47624/kt.031.QQOV9624

Statistik Teil 3: Konfidenzintervalle

M. Kohl, F. Münch

Die KARDIOTECHNIK stellt in der Rubrik Tutorials relevante Methoden für wissenschaftliche Arbeiten zur klinischen Perfusion vor.

Fazitbox

Pro und Contra Konfidenzintervalle:

Pro
- Konfidenzintervalle stellen eine sehr gute Möglichkeit dar, die Unsicherheit bei der Schätzung unbekannter Parameter auszudrücken.
- Moderne datenbasierte statistische Ansätze, wie das Bootstrapping, ermöglichen die Berechnung von Konfidenzintervallen, auch ohne umfangreiche (theoretische) Annahmen und sollten vermehrt eingesetzt werden.

Contra
- Konfidenzintervalle werden häufig missinterpretiert.
- Die exakte Überdeckungswahrscheinlichkeit berechneter Konfidenzintervalle ist in praktischen Anwendungen unbekannt.

Einführung

In den meisten Fällen geht es bei statistischen Datenanalysen darum, gewisse unbekannte Kenngrößen einer Population (auch Grundgesamtheit genannt) zu bestimmen. In der Statistik nennt man dies Schätzen oder Fitten unbekannter Parameter. Beispiele sind etwa Wahrscheinlichkeiten, Mittelwerte, Mediane, Standardabweichungen, Korrelationen oder auch Parameter von Regressionsmodellen. Da die Ergebnisse einer Schätzung auf einer repräsentativen Stichprobe basieren, unterliegen die Schätzungen der unbekannten Kenngröße der zufälligen Schwankung. D. h.: Würde die Studie mit einer neuen repräsentativen Stichprobe wiederholt, so könnte man davon ausgehen, dass die neuen Schätzungen mehr oder weniger stark von den ersten Schätzungen abweichen. Auch muss man annehmen, dass die Schätzung nicht exakt dem unbekannten Wert des Parameters für die Population entspricht. Es ist daher wichtig einzuschätzen, wie genau der unbekannte Wert der gesuchten Kenngröße geschätzt wurde. Ein wichtiges Werkzeug hierfür sind die so genannten Konfidenzintervalle, die von Neyman bereits 1937 eingeführt wurden [1]. Ein Konfidenzintervall ist demnach ein wichtiges Hilfsmittel, um die Ungenauigkeit von Schätzungen unbekannter Parameter in der Statistik zu veranschaulichen.

Unter einem Konfidenzintervall (confidence interval = CI) versteht man ein Intervall, welches auf Basis einer gegebenen Stichprobe berechnet wird und den wahren unbekannten Wert des Parameters mit einer vorgegebenen Wahrscheinlichkeit 1-α überdeckt (= Überdeckungswahrscheinlichkeit), siehe auch Abbildung 1.

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

Ein $(1-\alpha)$-Konfidenzintervall (oder auch Vertrauensintervall) ist eine Intervallschätzung $\mathrm{CI} = [S_u, S_o]$, für die gilt

$$P(S_u \leq \theta \leq S_o) \geq 1-\alpha \qquad \alpha \in [0,1] \tag{1}$$

mit P = Wahrscheinlichkeit. Das Konfidenzintervall CI reicht demnach von der unteren Grenze S_u bis zur oberen Grenze S_o und überdeckt den wahren unbekannten Wert des Parameters θ mit einer Wahrscheinlichkeit von mindestens $1-\alpha$.

Das Konfidenzintervall $\mathrm{CI} = [S_u, S_o]$ konkretisiert sich oft auf die folgende Form (vgl. Abschnitt 6.5 in [5])

$$\mathrm{CI} = [\theta_n - k_1\sigma_{\theta_n}, \theta_n + k_2\sigma_{\theta_n}] \tag{2}$$

Hierbei steht θ_n für den (Punkt-)Schätzer des gesuchten Parameters θ und σ_{θ_n} für die Standardabweichung dieses Schätzers, die auch Standardfehler genannt wird. Bei den Konstanten $k_1 > 0$ und $k_2 > 0$ handelt es sich um geeignete Quantile, die von α, dem verwendeten Schätzer θ_n, dem zugrundeliegenden Wahrscheinlichkeitsmodell und der Stichprobengröße n abhängen können.

Die Größen θ_n und σ_{θ_n} hängen u.a. von der Fallzahl n ab. Es ist zu erwarten, dass der Schätzer θ_n mit wachsenden n immer näher beim wahren unbekannten Wert des Parameters θ liegt (Gesetz der großen Zahlen) und der Standardfehler σ_{θ_n} immer kleiner wird. In der Regel schrumpft der Standardfehler σ_{θ_n} mit einer Rate von $\sqrt{n}$. In diesem Fall halbiert sich also die Länge des Intervalls, wenn man die Stichprobengröße vervierfacht $(4n)$.

Abb. 1: Mathematische Definition des Konfidenzintervalls

Für α werden hierbei sinnvollerweise kleine Werte gewählt, meistens 5 %, womit sich in diesem Fall ein 95 %-CI (CI95) ergibt. Dabei ist zu beachten, dass ein CI umso länger wird, je kleiner α gewählt wird. Generell ist es das Ziel, die Grenzen so zu wählen, dass die Überdeckungswahrscheinlichkeit genau 1-α ist. In diesem Fall spricht man auch von einem exakten CI.

Es ist sehr wichtig, den für die Definition des CI zentralen Begriff der Überdeckungswahrscheinlichkeit richtig zu interpretieren. Eine Überdeckungswahrscheinlichkeit von 1-α bedeutet, dass in (mindestens) (1-α) % der Fälle, in denen wir auf Basis einer Stichprobe ein CI berechnen, dieses Intervall tatsächlich den wahren unbekannten Wert des Parameters enthält. Es ist folglich nicht richtig, dass der wahre Wert des Parameters mit (mindestens) (1-α) % Wahrscheinlichkeit im berechneten CI liegt. Denn nachdem das CI auf Basis der vorliegenden Daten berechnet wurde, liegt der gesuchte unbekannte Wert des Parameters entweder im Intervall oder eben nicht. Die Wahrscheinlichkeit, dass der unbekannte Wert des Parameters im berechneten CI liegt, ist demnach nicht 1-α, sondern eben 1 oder 0. Diese häufige und fundamentale Fehlinterpretation des CI bezeichnen Morey et al. als „The Fundamental Confidence Fallacy" [2]. In Abbildung 2 findet sich ein Beispiel hierfür. Es wurden 100 normalverteilte Zufallsstichproben mit jeweils 20 Beobachtungen erzeugt und anschließend für jede Stichprobe jeweils der Mittelwert (arithmetisches Mittel) und das CI95 für den Mittelwert berechnet. Es wäre demnach zu erwarten (Erwartungswert), dass in 95 Fällen die CIs den wahren unbekannten Mit-

telwert der Normalverteilung überdecken. Tatsächlich kommen wir mit der Simulation dem erwarteten Wert sehr nahe und erhalten in 94 Fällen eine Überdeckung.

Bei der unteren und oberen Grenze des CI handelt es sich um Schätzungen, weshalb jede neue Stichprobe zu einer jeweils mehr oder weniger anderen unteren und oberen Grenze führt (Abb. 2). Es ist auch möglich, einseitige CIs zu betrachten, bei denen dann eben nur die untere oder obere Grenze geschätzt werden muss, während die jeweils andere Grenze auf den maximal bzw. minimal möglichen Wert des jeweiligen Parameters gesetzt wird. Ist der Parameter θ, der geschätzt werden soll, zum Beispiel eine Wahrscheinlichkeit (Minimum = 0, Maximum = 1), so ergäben sich die folgenden einseitigen (1-α)-Konfidenzintervalle CI = [0, S_o] bzw. CI = [S_u, 1].

Die Berechnung von CIs, speziell exakten CIs, ist in der Praxis in vielen Fällen sehr schwierig bzw. sogar unmöglich, weshalb man stattdessen oft auf approximative CIs zurückgreifen muss. In der klassischen Statistik werden solche genäherten CIs mit Hilfe asymptotischer (n wächst ins Unendliche) Ergebnisse berechnet. Das wichtigste Hilfsmittel dafür sind sogenannte zentrale Grenzwertsätze, welche auf eine Normalverteilung als Grenzwertverteilung für den Schätzer führen. In diesem Fall gilt in Gleichung (2) aus Abbildung 1: $k_1 = k_2 = z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das (1-α/2)-Quantil der Normalverteilung mit Mittelwert 0 und Standardabweichung 1 ist. Im Fall α = 5 % ergibt sich $z_{0,975} = 1{,}96 \approx 2$ (vgl. Abschnitt 6.5 in [5]).

In der modernen datengestützten Statistik kommen für die Berechnung angenäherter (approximativer) Konfidenzintervalle sogenannte Permutations- oder Resamplingverfahren und im speziellen das sogenannte Bootstrapping [6] zum Einsatz (nähere Informationen zum Thema Bootstrap-Konfidenzintervall sind im Supplement über den nebenstehenden QR-Code zu finden). Im Unterschied zur klassischen Statistik sind bei diesen Verfahren weniger (theoretische) Annahmen nötig. So kann ein entsprechendes CI auch ohne ein konkretes Wahrscheinlichkeitsmodell und ohne Berechnung des Standardfehlers ermittelt werden. Alle benötigten Informationen werden direkt aus den Daten gewonnen. Im einfachsten Fall wird beim Bootstrapping aus der vorliegenden repräsentativen Stichprobe durch Ziehen mit Zurücklegen eine neue Zufallsstichprobe identischer Größe erzeugt und der

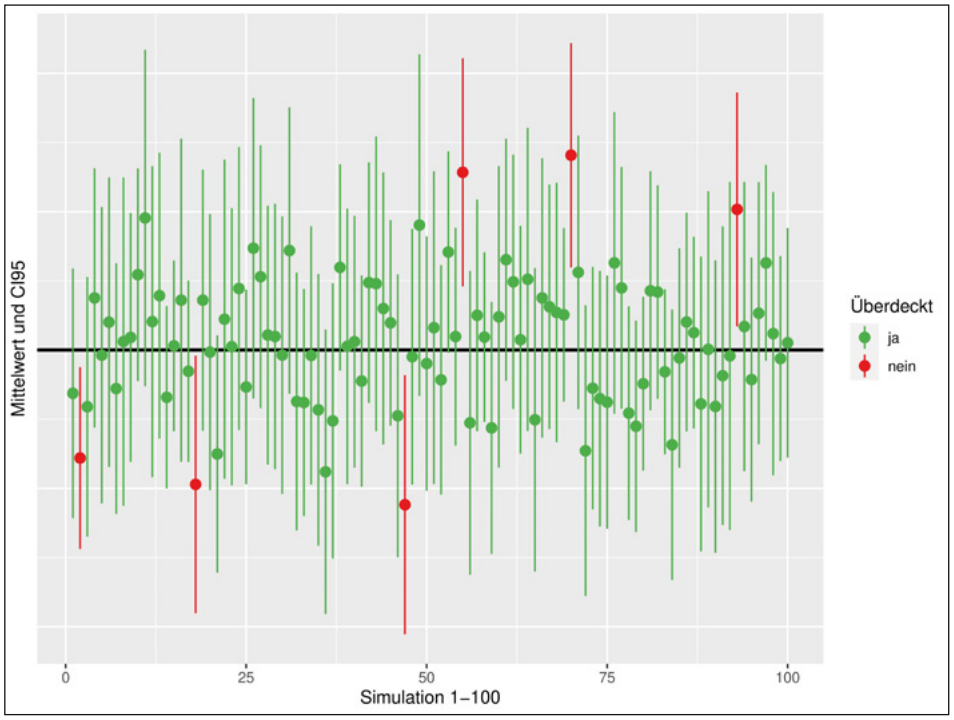

Abb.2: Mittelwert und CI95 für 100 normalverteilte Zufallsstichproben; (erstellt mit der Statistiksoftware R [3] und dem Paket ggplot2 [4])

gesuchte Parameter auf Basis dieser neuen Zufallsstichprobe geschätzt. Dieses Vorgehen wird dann *tausende Male wiederholt*, woraus sich entsprechend *tausende Schätzwerte* für den gesuchten Parameter ergeben. Die einfachste Möglichkeit zur Berechnung des Boostrap-CI besteht darin, das α/2- und (1-α/2)-Quantil dieser Schätzwerte als Unter- und Obergrenze des Intervalls zu nehmen. Das Bootstrapping funktioniert erfahrungsgemäß auch bei kleinen Stichprobengrößen sehr gut, wobei Bootstrap-CIs für kleine bis moderate Fallzahlen ($10 \leq n \leq 50$) und schiefe Verteilungen tendenziell etwas zu kurz sind (vgl. Kapitel 3 in [7]).

Der generelle Nachteil der approximativen CIs liegt darin, dass die genaue Überdeckungswahrscheinlichkeit unbekannt ist. Insbesondere ist auch nicht klar, ob diese, wie in der Definition des CIs gefordert, größer oder gleich 1-α ist. Es ist bekannt, dass approximative CIs für kleine bis moderate Fallzahlen ($n \leq 50$) tendenziell zu kurz sind und nicht die vorgegebene Überdeckungswahrscheinlichkeit von 1-α erreichen. Jedoch ist dies auch bei den theoretisch exakten CIs nicht garantiert, da nicht klar ist, ob die für die Herleitung notwendigen (theoretischen) Annahmen in der Praxis auch tatsächlich erfüllt sind. Entsprechend ist es in jedem Fall nötig, CIs, wie jedes statistische Resultat, mit gebührender Vorsicht zu interpretieren. Bei größeren Fallzahlen ($n \geq 100$) sind die Unterschiede zwischen den verschiedenen Ansätzen hingegen üblicherweise klein bis sehr klein und auffällige Unterschiede können darauf hindeuten, dass gemachte Annahmen nicht zutreffend sind. Insbesondere sollte in diesem Fall hinterfragt werden, ob das statistische Model richtig gewählt wurde. CIs sind auch eng mit statistischen Hypothesentests verwandt, wobei es einige Argumente dafür gibt, warum CIs den statistischen Tests vorzuziehen sind [8,9]. Darauf wollen wir hier aber nicht näher eingehen, sondern betrachten zwei Beispiele, wobei wir im ersten Beispiel zeigen, wie ein CI für die Fallzahlplanung herangezogen werden kann.

Beispiel 1: Venöse Sauerstoffsättigung an der EKZ

Wir gehen davon aus, dass die vorliegenden Daten zur venösen Sauerstoffsättigung (S_vO_2) von 30 Patienten einer Normalverteilung unterliegen, wobei wir am Mittel-

Es seien AM_n das arithmetische Mittel (AM) und SD_n die geschätzte Standardabweichung (SD) zur Stichprobengröße n. Dann ergibt sich das folgende exakte CI für den Mittelwert der angenommenen Normalverteilung (vgl. Abschnitt 6.8.2 in [5])

$$\left[\mathrm{AM}_n - t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}},\ \mathrm{AM}_n + t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}}\right] \quad (3)$$

wobei $t_{n-1;1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ist. Der Ausdruck $\frac{\mathrm{SD}_n}{\sqrt{n}}$ entspricht dem Standardfehler des Mittelwertes (standard error of mean = SEM).

Das asymptotische CI ergibt sich, indem man in Gleichung (3) $t_{n-1;1-\alpha/2}$ durch $z_{1-\alpha/2}$, das $(1-\alpha/2)$-Quantil der Standardnormalverteilung (Mittelwert = 0 und Standardabweichung = 1), ersetzt

$$\left[\mathrm{AM}_n - z_{1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}},\ \mathrm{AM}_n + z_{1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}}\right] \quad (4)$$

wobei $z_{1-\alpha/2} < t_{n-1;1-\alpha/2}$. Folglich ist das asymptotische Konfidenzintervall für den Mittelwert immer kürzer als das exakte Intervall.

Abb. 3: Exaktes und asymptotisches Konfidenzintervall für den Mittelwert einer Normalverteilung

wert der Verteilung interessiert sind und die Standardabweichung unbekannt ist. Die Formeln für das exakte und asymptotische CI für den Mittelwert finden sich in Abbildung 3. In Abbildung 4 sehen wir das exakte, das asymptotische und das Bootstrap-CI auf Basis von 10.000 Wiederholungen.

Da $z_{0,975} = 1{,}96 < 2{,}05 = t_{29;0,975}$, ist das asymptotische CI etwas kürzer als das exakte CI. Das Boostrap-CI ist am kürzesten und zudem leicht asymmetrisch. Auf Basis der vorliegenden Daten erscheint demnach eine mittlere venöse Sauerstoffsättigung im Bereich von 75–80 % für das beobachtete EKZ-Patientenkollektiv plausibel.

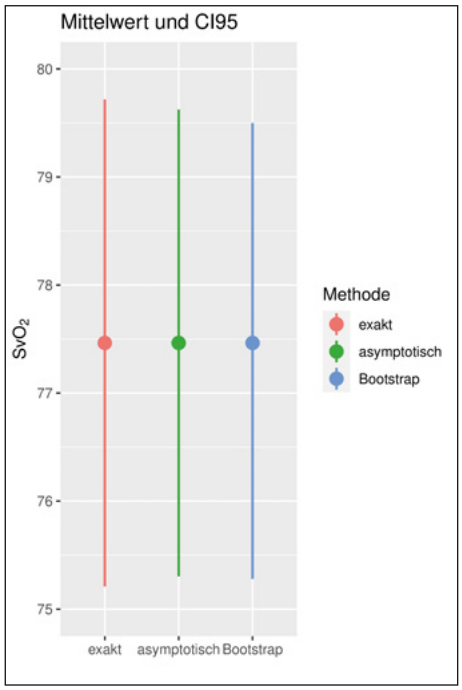

Abb. 4: Mittelwert und CI95 der venösen Sauerstoffsättigung (S_vO_2) von 30 Patienten; (Erstellt mit den R Paketen ggplot2 [4] und MKinfer [10])

Fallzahlberechnung mittels CI

Hierfür betrachten wir nun die obigen Daten als Daten einer Pilotstudie, deren Ergebnisse wir für die Fallzahlplanung einer neuen, größer angelegten Studie nutzen wollen. Der mittlere SvO_2 lag in der Studie bei 77,5 %, die Standardabweichung bei 6,0 %. Für die Fallzahlberechnung müssen wir außerdem α und die Länge des CI festlegen. Wir wählen α = 5 % und eine Länge des CI95 von 3 % (Mittelwert +/-1,5 %). Da die Nullstelle für die Funktion f aus Gleichung (7) zwischen n = 63 und n = 64 liegt, ist die benötigte Fallzahl für die geplante Studie n = 64. Bei der Anwendung der asymptotischen Formel aus Gleichung (9), ergibt sich eine Fallzahl von n = 62. Das etwas kürzere asymptotische CI führt demnach zu einer leichten Unterschätzung der Fallzahl (Abb. 5).

Aus der Formel (3) für das exakte CI des Mittelwertes folgt für die Länge L des Intervalls

$$L = \mathrm{AM}_n + t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}} - \left(\mathrm{AM}_n - t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}}\right) \quad (5)$$

$$= 2t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}} \quad (6)$$

Demnach spielt der angenommen Mittelwert für die Berechnung der Fallzahl keine Rolle. Da das Quantil der t-Verteilung von n abhängt, lässt sich die Gleichung nicht direkt nach n auflösen. Wir können die Gleichung aber numerisch nach n lösen. Hierzu definieren wir die Funktion

$$f(n) := 2t_{n-1;1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}} - L \quad (7)$$

und bestimmen deren Nullstelle.

Ersetzen wir die exakte Formel (3) durch die asymptotische Formel (4), so ergibt sich

$$L = 2z_{1-\alpha/2}\frac{\mathrm{SD}_n}{\sqrt{n}} \quad (8)$$

Diese Gleichung können wir nach n auflösen und erhalten

$$n = \left(\frac{2z_{1-\alpha/2}\mathrm{SD}_n}{L}\right)^2 \quad (9)$$

Abb. 5: Fallzahlberechnung für den Mittelwert

Beispiel 2: Urinmenge an der EKZ

Wir untersuchen die ausgeschiedene Urinmenge in ml pro kg Körpergewicht und h EKZ-Zeit anhand der Daten von 627 Patienten [11]. Aufgrund der schiefen Verteilung der Daten (vgl. Histogramm in Abb. 6), wählen wir den Median, um die Lage der Daten zu beschreiben. Der Median und die CI95s für den Median sind in Abbildung 6 dargestellt. Wir sehen, dass die CIs für alle drei Methoden sehr ähnlich sind, was wegen der recht hohen Fallzahl nicht überrascht. Genauer gesagt sind die oberen Grenzen der Intervalle in allen drei Fällen identisch bei 5,28 ml. Die unteren Grenzen liegen bei 4,243 ml (exakt), 4,242 ml (asymptotisch) und 4,263 ml (Bootstrap) und unterscheiden sich somit nur minimal. Wir können demnach auf Basis der vorliegenden Daten davon ausgehen, dass die ausgeschiedene Urinmenge für die vorliegende Patientenpopulation im Median zwischen 4,2 und 5,3 ml pro kg Körpergewicht und h EKZ-Zeit liegt.

Zusammenfassung

Konfidenzintervalle sind heute ein unverzichtbares Hilfsmittel der angewandten Statistik, um die Unsicherheit in der Schätzung von unbekannten Parametern auszudrücken, und können mittels moderner Statistiksoftware für beliebige Parameter (Wahrscheinlichkeiten, Mittelwerte, Mediane, Standardabweichungen, Korrelationen, Parameter von Regressionsmodellen etc.) berechnet werden. Ihre Verwendung wird in den aktuellen Empfehlungen zur Berichterstattung von Studienergebnissen empfohlen (vgl. etwa CONSORT 2010 [12] bzw. https://www.equator-network.org/). Insbesondere die modernen datenbasierten Ansätze wie etwa Bootstrap-Konfidenzintervalle, welche weniger (theoretische) Annahmen als exakte oder asymptotische Intervalle benötigen, sollten vermehrt eingesetzt werden.

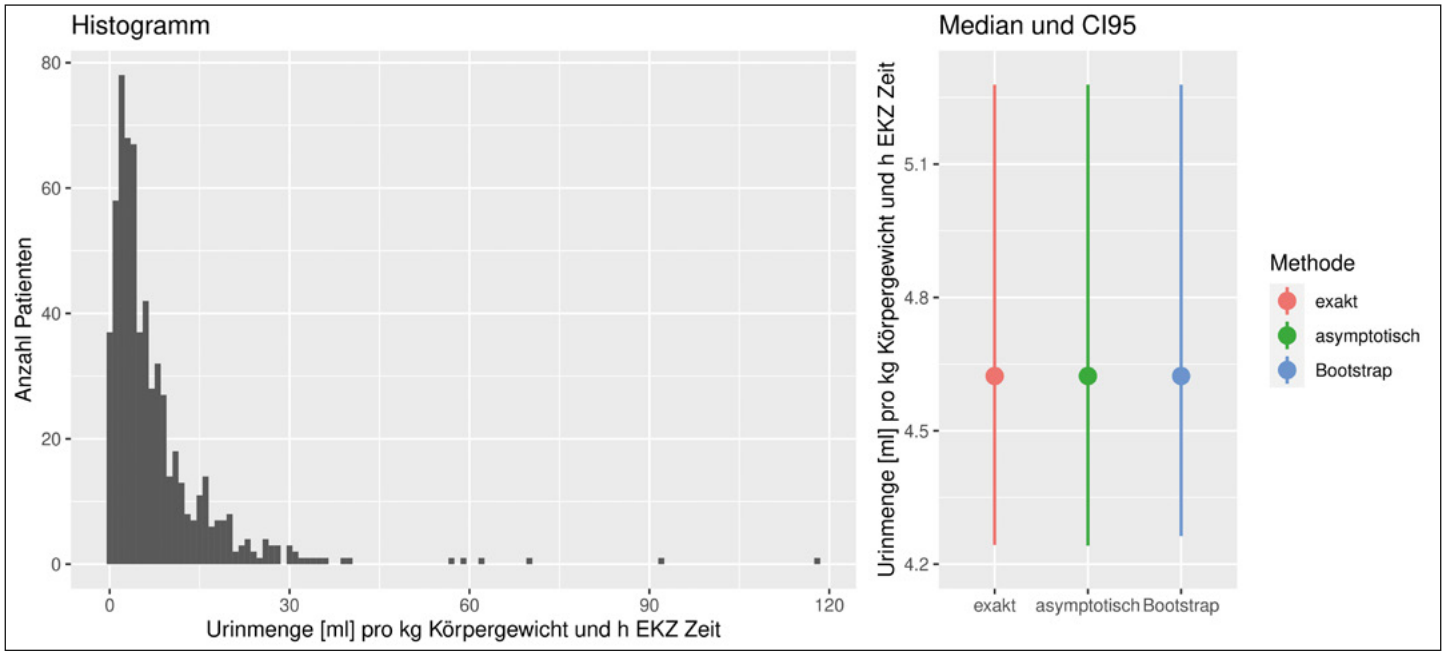

Abb. 6: Histogramm sowie Median und CI95 für die ausgeschiedene Urinmenge von 627 Patienten in ml pro kg Körpergewicht und h EKZ-Zeit; (erstellt mit den R Paketen ggplot 2[4] und MKinfer [10])

LITERATUR

1. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London. Series A. Mathematical and Physical Sciences 1937, 236:333-380.

2. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev. 2016, 23(1):103-23.

3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2022, Vienna, Austria. URL https://www.R-project.org/.

4. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2022, Springer-Verlag New York.

5. Hedderich J, Sachs L. Angewandte Statistik. Methodensammlung mit R. 2020, 17. Auflage, Springer-Verlag.

6. Efron B, Tibshirani, RJ. An Introduction to the Bootstrap. 1993, New York, NY: Chapman and Hall.

7. Chernick MR, LaBudde RA. An introduction to bootstrap methods with applications to R. Wiley & Sons Ltd. 2011.

8. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986; 292(6522):746-750.

9. du Prel JB, Hommel G, Röhrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106(19):335-339.

10. Kohl M. MKinfer: Inferential Statistics. R package version 0.6. 2020

11. Kohl M, Münch F. Statistik Teil 1: Der Box- und Whisker-Plot. Kardiotechnik. 2022; 31(1):15-17.

12. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332.