Tutorial

Statistik Teil 1: Der Box- und Whisker-Plot

M. Kohl, F. Münch

Die KARDIOTECHNIK wird in der Rubrik Tutorials in Folge relevante Methoden für wissenschaftliche Arbeiten zur klinischen Perfusion vorstellen.

Einführung

Eine der wichtigsten Graphiken zur explorativen Darstellung von univariaten Daten ist der sogenannte Box- and Whisker-Plot, welcher zu Beginn der 1970er Jahre von John Tukey [1] eingeführt wurde. Die Illustration in Abbildung 1 zeigt, wie ein Box- und Whisker-Plot aufgebaut ist und welche Informationen von diesem abgelesen werden können (Abb. 1).

Den Ausgangspunkt bildet die Box, die vom ersten (Q1 = 25 % Quantil) bis zum dritten Quartil (Q3 = 75 % Quantil) reicht und somit den Median (50 % Quantil) beinhaltet, der üblicherweise durch eine Linie oder einen Punkt in der Box markiert ist. Neben dem Lagemaß Median, ist auch der Interquartilsabstand (IQR) als wichtiges Streuungsmaß im Box- und Whisker-Plot dargestellt. Der IQR entspricht gerade der Länge der Box (Q3–Q1). Die Box beschreibt demnach die mittleren 50 % der Datenwerte, und es liegen jeweils 25 % der gemessenen Werte unterhalb und oberhalb der Box. Ausgehend von den Grenzen der Box werden der untere und obere Fence („Zaun“) definiert. Diese entsprechen dem kleinsten bzw. größten Datenwert, der nicht weiter als das 1,5-fache des Interquartilsabstandes vom unteren bzw. oberen Ende der Box entfernt ist. Die Whisker verbinden das untere bzw. obere Ende der Box mit dem unteren bzw. oberen Fence. Das Ende eines Whiskers entspricht folglich immer einem gemessenen Datenpunkt.

Gibt es Werte, die weiter als das 1,5-fache des Interquartilsabstandes von den Enden der Box entfernt sind, werden diese durch einzelne Punkte dargestellt. Man bezeichnet diese Messwerte auch als Ausreißer. Der Box- und Whisker-Plot kann folglich auch dazu verwendet werden, um ungewöhnlich kleine oder große Datenpunkte zu identifizieren, die dann einer genaueren Überprüfung unterzogen werden können. Hier ist zu beachten, dass sich hinter einem einzelnen Ausreißerpunkt auch mehrere identische Messwerte verbergen können.

Neben dem Lagemaß Median und dem Streuungsmaß IQR kann auch zu einem gewissen Maß die Schiefe (Stärke der Asymmetrie) der Datenverteilung aus einem Box- und Whisker-Plot abgeleitet werden. Bei einer nahezu symmetrischen Verteilung ist zu erwarten, dass der Median sehr gut in der Mitte der Box liegt und auch die Whisker nahezu gleich lang sind. Liegt der Median eher am unteren Ende der Box, wobei der obere Whisker länger ist als der untere Whisker, so liegt eine sogenannte rechts-schiefe Verteilung vor (Median kleiner als der Mittelwert, Mittelwert = arithmetisches Mittel). Ist es gerade umgekehrt, so kann man von einer links-schiefen Verteilung (Median größer als der Mittelwert) ausgehen.

Die Abbildung 2 zeigt einen Box- und Whisker-Plot der venösen Sauerstoffsätti-

Fazitbox

Pro und Contra Box- und Whisker-Plot:

Pro

- Zeigt Lage (Median), Streuung (IQR) und Schiefe der Verteilung univariater Daten an, wobei Median und IQR auch beim Vorliegen einzelner Ausreißer verlässliche Ergebnisse für die Lage und die Streuung der Daten liefern.
- Eignet sich dazu, Ausreißer zu identifizieren.
- Auch Daten mit einer sehr schiefen Verteilung können mittels Transformationen oder Adjustierung des klassischen Box- und Whisker-Plots adäquat dargestellt werden.

Contra

Viele klassische statistische Verfahren (z. B. 1- und 2-Stichproben t-Test, ANOVA) zielen auf den Vergleich von Mittelwerten ab. In diesem Fall sollte anstelle des Box- und Whisker-Plots besser eine Darstellung gewählt werden, welche die Mittelwerte und die Streuung der Daten oder die Streuung der Mittelwerte veranschaulicht, wie etwa Mittelwert +/-Standardabweichung, Mittelwert +/-Standardfehler oder Mittelwert und Konfidenzintervall des Mittelwertes.

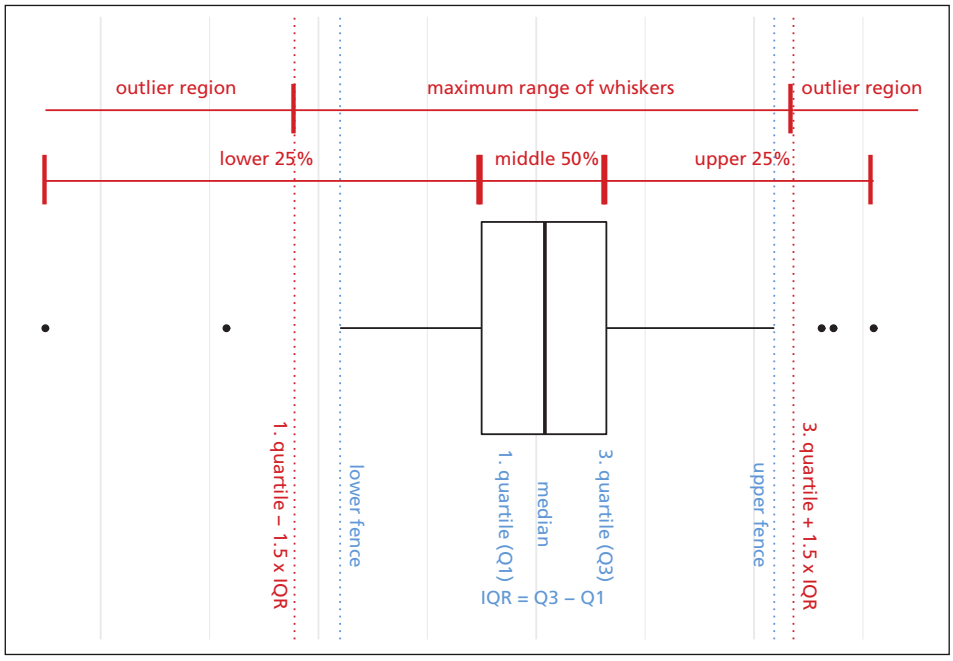

Abb.1: Illustration des Box- and Whisker-Plots. Erstellt mit dem Paket MKdescr [2] der Statistiksoftware R [3]

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

gung (SvO_2) von 30 erwachsenen Patienten an der extrakorporalen Zirkulation (EKZ). Es deutet sich eine leicht links-schiefe Verteilung an. Die Werte unterhalb des Medians streuen demnach stärker als die Werte oberhalb des Medians.

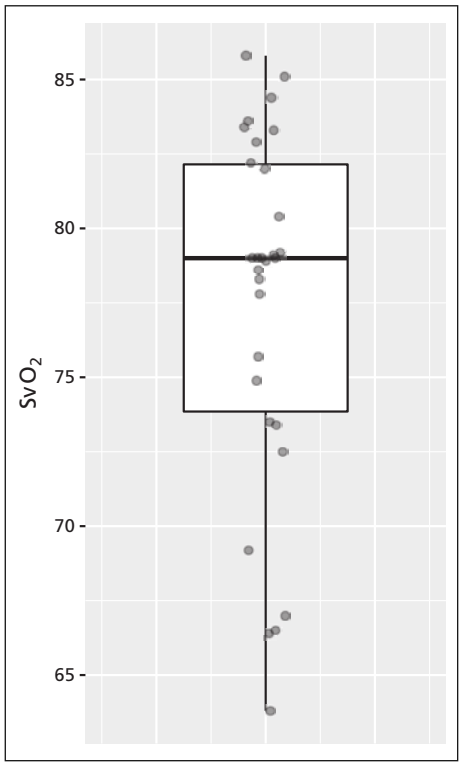

Abb.2: Box- und Whisker-Plot der venösen Sauerstoffsättigung (SvO_2) von 30 Patienten (eigene Daten). Erstellt mit dem R Paket ggplot2 [4]

Neben dem Box- und Whisker-Plot sind hier auch die Werte der einzelnen Patienten mittels Punkten dargestellt. Dies empfiehlt sich vor allem bei wenigen bis moderat vielen Datenpunkten, da ein Box- und Whisker-Plot auch bereits mit wenigen Datenpunkten erzeugt werden kann und eine gewisse Datenverteilung suggeriert, die bei wenigen Messwerten kaum zuverlässig bestimmt werden kann. Für die Darstellung wurde das sogenannte Jittering („Verwackeln“) verwendet. Hierbei werden die Koordinaten der Punkte zufällig etwas verwackelt; d. h. die Punkte werden zufällig etwas in x- und/oder y-Richtung verschoben. In Abbildung 3 wurde nur eine zufällige Verschiebung in x-Richtung vorgenommen und keine Verschiebung in y-Richtung, da diese die gemessenen Werte verfälscht hätte. Zusätzlich kam das sogenannte Alpha-Blending zum Einsatz, welches dazu dient, die Überlagerung von Objekten in Bildern sichtbar zu machen, indem die Farbgebung der Objekte entsprechend angepasst wird. Im Beispiel führt es dazu, dass sich überlagernde Punkte eine dunklere Farbe bekommen (vgl. im Bereich des Medians). Der Einsatz von Jittering und Alpha-Blending dient dazu, alle Messwerte sichtbar zu machen, auch die Daten von Patienten mit identischen Werten.

Der Box- und Whisker-Plot kann irreführend sein, wenn die Datenverteilung deutlich schief ist. Dies kann dazu führen, dass viele Messwerte im Ausreißerbereich landen. In Abbildung 3 sehen wir auf der linken Seite ein Histogramm für die Messwerte von 627 Kindern mit einem Körpergewicht von bis zu 12 kg, die an der EKZ behandelt wurden. Wir sehen eine deutlich rechts-schiefe Verteilungsform (Mittelwert > Median).

Es gibt demnach eine sehr große Anzahl von Patienten mit einer recht geringen Urinmenge. Anhand des Box- und Whisker-Plots können wir sehen, dass mehr als 75 % der Patienten eine Urinmenge von

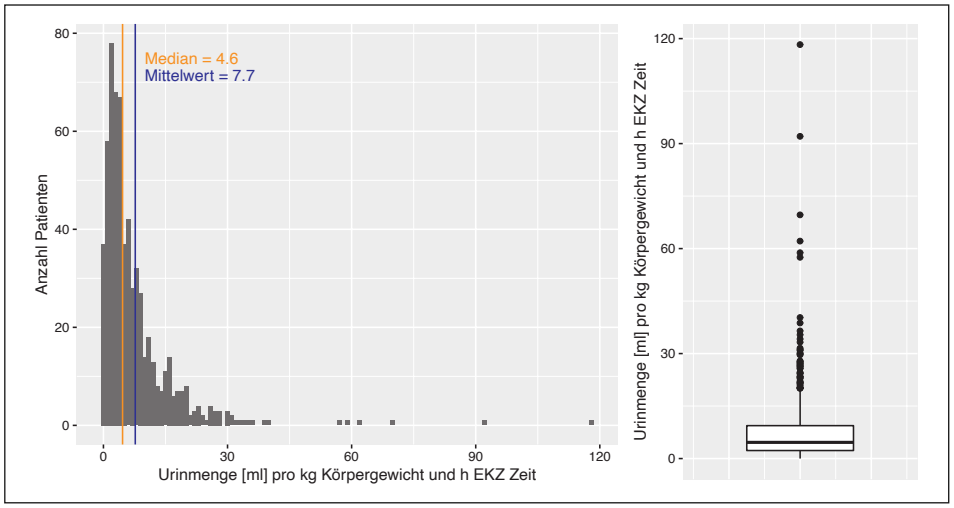

Abb. 3: Histogramm sowie Box- und Whisker-Plot der Urinmenge (in ml) pro kg Körpergewicht und h EKZ-Zeit. Es handelt sich um die Daten von 627 Patienten (eigene Daten). Erstellt mit dem R Paket ggplot2 [4]

weniger als 15 ml pro kg Körpergewicht und h EKZ Zeit aufweisen (oberes Ende der Box ist unterhalb von 15). Bei den verbleibenden knapp 25 % der Patienten sehen wir Werte, die jenseits von 20, 50 oder sogar 100 ml pro kg Körpergewicht und h EKZ-Zeit liegen. Aufgrund der recht großen Anzahl von Patienten, die vergleichsweise hohe Werte aufweisen, ist es eher unwahrscheinlich, dass dies durchweg Ausreißer sind, wie es vom Box- und Whisker-Plot suggeriert wird. Es ist eher plausibel, dass diese schiefe Verteilungsform eine typische Charakteristik der Daten ist, die bei der statistischen Analyse der Daten berücksichtigt werden sollte.

Es gibt im Wesentlichen zwei Ansätze, diesem Problem zu begegnen. Die erste Möglichkeit stellt der Einsatz einer geeigneten Datentransformation dar. Häufig kommt hierbei der Logarithmus, die Quadratwurzel oder auch die Standard-Hyperbel (Kehrwert) zum Einsatz [5]. Im vorliegenden Fall haben wir es mit einer rechts-schiefen Verteilung zu tun. In einem solchen Fall eignet sich oft der Logarithmus als Transformation [6,7]. Leider enthält der vorliegende Datensatz auch 24 Patienten, bei denen eine Urinmenge von 0 ml pro kg Körpergewicht und h EKZ Zeit aufgezeichnet wurde. Diese Werte würden mit dem Logarithmus auf $-\infty$ abgebildet und wären so nur schwierig darstellbar. Wir wählen daher als Transformation den Areasinus hyperbolicus, der manchmal auch als verallgemeinerter Logarithmus bezeichnet wird (g-log). Diese Funktion ist für große Werte dem Logarithmus sehr ähnlich, ist aber auch für Werte kleiner oder gleich Null definiert. Der zweite Ansatz besteht darin, die Daten unverändert zu lassen und statistische Methoden zu wählen, welche die konkrete Datenverteilung berücksichtigen. Eine Variante des Box- und Whisker-Plots, die für schiefe Verteilungen bessere Ergebnisse liefert, wurden zum Beispiel von Hubert und Vandervieren vorgeschlagen [8].

Die beiden Ansätze wurden auf die Daten zur Urinmenge angewendet. Die Ergebnisse sind in Abbildung 4 zu sehen. Die Transformation zeigt wie erhofft eine stark symmetrisierende Wirkung; die Schiefe ist nahezu verschwunden. Es werden nun nur noch sehr wenige Patienten vom Box- und Whisker-Plot als auffällig (Ausreißer) angezeigt. Der an die Schiefe angepasste Box- und Whisker-Plot markiert ebenfalls deutlich weniger Patienten als Ausreißer, wobei hier auch sehr kleine Werte als auffällig markiert werden.

Die angeführten Beispiele zeigen, dass der Box- und Whisker-Plot vielfältig einsetzbar

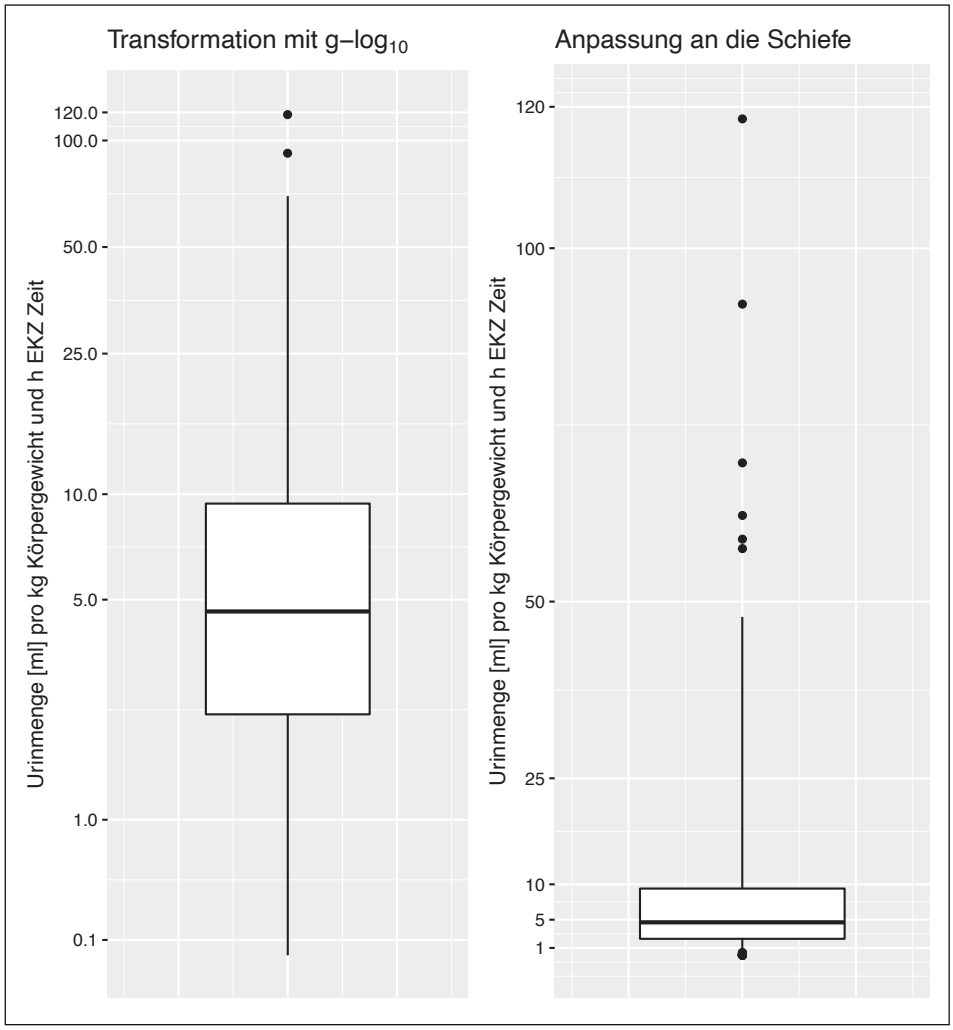

Abb. 4: Box- und Whisker-Plots der Urinmenge (in ml) pro kg Körpergewicht und h EKZ Zeit nach Anwendung des verallgemeinerten Logarithmus (links) bzw. einer Anpassung an die Schiefe (rechts). Erstellt mit den R-Paketen MKdescr [2], ggplot2 [4] und litteR [9]

ist und wichtige statistische Informationen (Lage, Streuung und Schiefe der Daten) in einer sehr übersichtlichen Form graphisch darstellt. Dieser Plot kann daher unter Umständen mit zusätzlichem Einsatz von Transformationen oder Adjustierungen nahezu uneingeschränkt für die graphische Darstellung univariater Daten empfohlen werden. Für eine noch umfangreichere Behandlung des Box- und Whisker-Plots inklusive weiterer Adjustierungen sowie von Erweiterungen auf zweidimensionale Darstellungen verweisen wir auf den technischen Report von Wickham und Stryjewski [10].

LITERATUR

1. Tukey JW. Exploratory Data Analysis. Addison-Wesley. 1977.

2. Kohl M. MKdescr: Descriptive Statistics. R package version 0.8. 2021. https://www.stamats.de/.

3. Core R. Team. A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. 2021. https://www.R-project.org/.

4. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016. Springer-Verlag New York.

5. Bland JM, Altman DG. Statistics Notes: Transforming data. BMJ 1996; 312:770. doi:10.1136/bmj.312.7033.770/.

6. Bland JM, Altman DG. Statistics notes: Transformations, means, and confidence intervals. BMJ 1996; 312 :1079 doi:10.1136/bmj.312.7038.1079/.

7. Bland JM, Altman DG. Statistics Notes: The use of transformation when comparing two means. BMJ 1996; 312:1153 doi:10.1136/bmj.312.7039.1153/.

8. Hubert M, Vandervieren, E. An adjusted boxplot for skewed distribution. Computational Statistics and Data Analysis. 2008. 52(12): 5186–201.

9. Schulz M, Walvoort D, Barry J, Fleet D, van Loon W. Baseline and power analyses for the assessment of beach litter reductions in the European OSPAR region. Environmental Pollution 2019. 248,555-564. doi: 10.1016/j.envpol.2019.02.030/.

10. Wickham H, Stryjewski L. 40 years of boxplots. 2012. https://vita.had.co.nz/papers/boxplots.pdf, letzter Zugriff 19.12.2021.

Matthias Kohl studierte Mathematik an der Universität Bayreuth und promovierte in Robuster Statistik am Lehrstuhl für Mathematische Statistik.
Seit 2010 ist er an der Hochschule Furtwangen als Professor für Mathematik in Biologie und Medizin beschäftigt. Er ist Mitglied des BW-CAR (Baden-Württemberg Center of Applied Research), war Gründungsvorstand des Institute of Precision Medicine und leitet die Forschungsgruppe Data Science for Life Science. Er gründete und leitet das Steinbeistransferzentrum Personalisierte Medizin und hat mehr als 20 Jahre Erfahrung in der statistischen Beratung von Wissenschaftlern und Firmen.
Er ist (Co-)Autor von mehr als 120 begutachteten wissenschaftlichen Artikeln und mehr als 30 Erweiterungspaketen zur Statistiksoftware R.