

Quantile, empirische Verteilungsfunktion und Box Plot

St. Lange¹, R. Bender²

¹ Abteilung für Medizinische Informatik, Biometrie und Epidemiologie der Ruhr-Universität Bochum

² Fakultät für Gesundheitswissenschaften, AG Epidemiologie und medizinische Statistik, Universität Bielefeld

Neben der Darstellung des Schwerpunkts bzw. der Mitte der Daten durch Mittelwert oder Median können für eine adäquate Beschreibung noch weitere Lagepunkte relevant sein. Verallgemeinerungen des Medians sind die Quantile, die die Daten ebenfalls in zwei (allerdings nicht gleich große) Hälften teilen. Das 25%-Quantil zum Beispiel markiert einen Punkt, unterhalb dessen (mindestens [s.u.]) 25% und darüber (höchstens [s.u.]) 75% der Werte liegen.

Die Einschränkung »mindestens« bzw. »höchstens« im vorigen Absatz ist bei diskreten (nur endlich viele Merkmalausprägungen möglich, z.B. Scores) Daten und endlichen Stichproben relevant; hier werden die Quantile formal zumeist als der kleinste Wert der Stichprobe definiert, für den mindestens ein vorgegebener Anteil – also zum Beispiel 25% – kleiner oder gleich diesem Wert ist. Bei einem Stichprobenumfang von beispielsweise 11 (unterschiedlichen) Werten sind ca. 18% der Werte kleiner oder gleich dem zweitkleinsten, ca. 27% kleiner oder gleich dem dritt kleinsten und ca. 36% kleiner oder gleich dem viert kleinsten Wert. Damit ist der dritt kleinste Wert als 25%-Quantil definiert.

Das **25%-Quantil**, der Median und das **75%-Quantil** werden auch als Quartile bezeichnet, da sie die Stichprobe in 4 (zumindest annähernd) gleich große Bereiche unterteilen. In ähnlicher Weise werden auch Terzile, Quintile, Dezile oder Perzentile angegeben, die die Daten in drei, fünf, zehn bzw. hundert gleich große Bereiche untergliedern. Weiterhin relevant sind noch die 2,5%-, 5%-, 95%- bzw. 97,5%-Quantile, da sie häufig für die Bildung von Referenzbereichen Verwendung finden (1).

Für die Bestimmung von Quantilen ist es zunächst erforderlich, die Werte der Stichprobe vom kleinsten zum größten Wert zu sortieren. Anschließend wird für jede Beobachtung die kumulative (relative) Häufigkeit angegeben, das heißt der Anteil von Werten, die kleiner (oder gleich) dieser Beobachtung sind. Das kann dann graphisch als Treppenfunktion dargestellt werden, wobei jede Beobachtung einer Stufe entspricht. Die Höhe jeder Stufe hängt von der Zahl gleicher Beobachtungen ab. Falls keine gleichen Werte in der Stichprobe vorliegen, beträgt die Höhe der Stufen 1 (dividiert durch den Stichprobenumfang). Die Breite der Stufen entspricht dem Abstand zweier benachbarter beobachteter Merkmalausprägungen. Diese Funktion (treppenartige Kurve) wird als **empirische Verteilungsfunktion** bezeichnet und stellt die wichtigste, nicht aggregierende, das heißt, noch jede Einzelinformation enthaltende Darstellung von quantitativen Daten dar. In **Abb. 1** ist zur Illustration die empirische Verteilungsfunktion der sys-

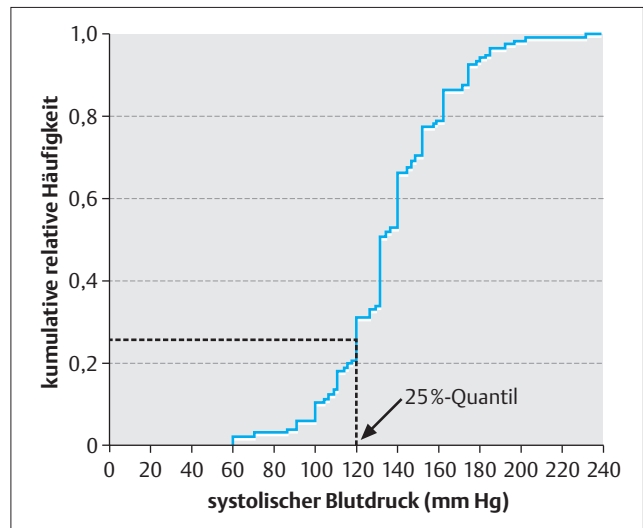


Abb.1 Empirische Verteilungsfunktion der systolischen Blutdruckwerte von 150 Patienten mit akutem Myokardinfarkt zum Zeitpunkt der Krankenhausaufnahme. Eingezeichnet ist das 25%-Quantil.

tolischen Blutdruckwerte zum Zeitpunkt der Krankenhausaufnahme von 150 Patienten mit einem akuten Myokardinfarkt dargestellt.

Die Quantile können nun anhand der Verteilungsfunktion zeichnerisch ermittelt werden, indem von dem interessierenden Quantilwert auf der Ordinate (zum Beispiel 25% bzw. 0,25) eine Parallele zur Abszisse gezogen und am (ersten) Schnittpunkt mit dem Graphen der Verteilungsfunktion das Lot auf die Abszisse gefällt wird. Ein nicht graphischer Zugangsweg ist der folgende: Bezeichnet man mit n den Stichprobenumfang, ist das $q\%$ -Quantil der $(q/100 \cdot n)$ -kleinste Wert der Stichprobe. Bei einem Stichprobenumfang von beispielsweise 76 ist das 25%-Quantil der 19-kleinste Wert ($25/100 \cdot 76 = 19$).

Falls der Ausdruck $(q/100 \cdot n)$ einen nicht ganzzahligen Wert ergibt, wird das $q\%$ -Quantil zumeist als der nächstgrößere Wert der Stichprobe definiert. Allerdings finden sich in der Literatur und insbesondere in Statistik-Software andere Definitionen, bei denen die Quantile nicht notwendigerweise tatsächlich beobachtete Werte der Stichprobe sind, sondern zum Beispiel durch Interpolation entstehen.

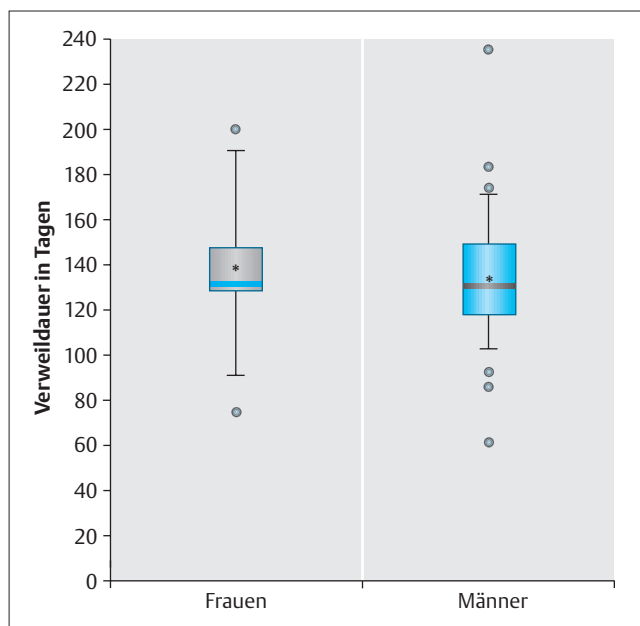


Abb. 2 Box-and-Whisker-Plots der systolischen Blutdruckwerte von 150 Patienten mit akutem Myokardinfarkt zum Zeitpunkt der Krankenhausaufnahme, getrennt nach Männern und Frauen. Die Whiskers kennzeichnen das 5%- bzw. das 95%-Quantil. Die Mittelwerte sind durch einen * symbolisiert.

Tab. 1 Übersetzungen (deutsch – englisch)

Quantil	quantile
Referenzbereich	reference interval
kumulative (relative) Häufigkeit	cumulative (relative) frequency
Verteilungsfunktion	distribution function

kurzgefasst: Quantile dienen der Beschreibung einer nach der Größe der Werte geordneten Reihe. Je nach Fragestellung können unterschiedliche Quantile eingesetzt werden, z. B. die Quartile (25%-, Median, 75%-Quantil) oder auch die 2,5%-, 5%- 95%-, 97,5%-Quantile, die bei der Bestimmung von Referenzbereichen verwendet werden können. Stellt man die Verteilung der geordneten Werte graphisch dar, so ergibt sich die empirische Verteilungsfunktion.

Eine einfache, aber außerordentlich nützliche und anschauliche Möglichkeit, Daten anhand von Quantilen zusammenfassend darzustellen, bieten so genannte »Box-and-Whisker-Plots« (oder kurz Box Plot) (2). Die Box im Box-and-Whisker-Plot wird durch das 25%- und 75%-Quantil begrenzt. In den Kästen wird der Median als waagerechte Linie eingezeichnet und häufig zusätzlich der Mittelwert als Punkt oder Stern, während senkrechte Linien zu den »Schnurrhaaren« (»whiskers«, ebenfalls wieder waagerechte Linien) gezogen werden. Häufig stellen die »whiskers« das 10%- (bzw. 90%-) oder das 5%- (95%-) Quantil dar. Sinnvoll ist es, jedoch ebenfalls nicht einheitlich verwendet, die außerhalb dieses Bereichs liegenden Extremwerte als separate Punkte mit in die graphische Darstellung aufzunehmen.

Box-and-Whisker-Plots sind insbesondere dann hilfreich, wenn mehrere Gruppen von Patienten hinsichtlich der Verteilung ihrer Daten verglichen werden sollen. Neben der zentralen Lage erhält man einen groben Überblick über die Symmetrie der Verteilungen. In **Abb. 2** sind Box-and-Whisker Plots für die systolischen Blutdruckwerte der 150 Patienten mit akutem Myokardinfarkt aus Abbildung 1, getrennt nach Männern und Frauen, dargestellt.

kurzgefasst: Im Box-and-Whisker-Plot können Daten anhand von Quantilen zusammenfassend dargestellt werden. Die Box wird begrenzt durch das 25%- und das 75%-Quantil, der Median in der Mitte eingezeichnet, und die Whiskers (Schnurrhaare) begrenzen beispielsweise das 10%- und das 90%-Quantil.

Die englischsprachigen Bezeichnungen der wichtigsten in diesem Beitrag diskutierten Begriffe finden Sie in **Tab. 1**.

Literatur

- 1 Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. *Brit med J* 1994; 309: 996
- 2 Wilson APR. Box-plots for microbiologists? *Lancet* 1993; 341: 282

Korrespondenz

Dr. Stefan Lange
Abteilung für Medizinische Informatik, Biometrie und Epidemiologie
Ruhr-Universität
Universitätsstraße 150
44780 Bochum
E-Mail: stefan.f.lange@ruhr-uni-bochum.de