Grafische Veranschaulichung von Statistischen Daten mittels box-plot

Michael Spielmann

1 Zusammenfassung

Der von John W. Tukey eingeführte box-plot zur grafischen Darstellung eindimensionaler statistischer Daten hat inzwischen auch Berücksichtigung in einigen Schulbüchern gefunden. Er ist in grafikfähigen Taschenrechnern Ti83 und Ti92 implementiert. Der box-plot ist sehr einfach zu erstellen, leichter und schneller als ein Histogramm, und er vermittelt dennoch einen guten Eindruck von der Verteilung der Daten. Richtig angewandt liefert er eine Fülle von Informationen.

2 Median contra Arithmetisches Mittel

Der box-plot beruht im wesentlichen auf der Bestimmung des Medianes. Der Median teilt die der Größe nach sortierte Stichprobe in Bezug auf die Anzahl der Daten in zwei Hälften; er ist die Zahl, für die die Hälfte der Daten kleiner oder gleich dieser Zahl und die Hälfte der Daten größer oder gleich dieser Zahl ist. Zwar hat sich das arithmetische Mittel als Ausdruck eines mittleren Wertes sehr stark durchgesetzt, so stark, dass man oft auch Zeugnisnoten durch arithmetische Mittelwerte vergleicht. Doch wir wissen, dass man bei Rangskalen tunlichst den Median benutzen sollte; die Abstände zwischen zwei Noten sind nun einmal nicht gleichmäßig. Der Median hat nicht nur den Vorteil, gleichermaßen auf Ordinalskalen und Rangskalen zu passen, er ist auch leicht zu ermitteln. Man findet den Median prinzipiell durch einfaches Abzählen. Bei ungerader Daten-Anzahl ist der Median der mittlere Wert, bei gerader Anzahl das arithmetische Mittel der beiden mittleren Werte.

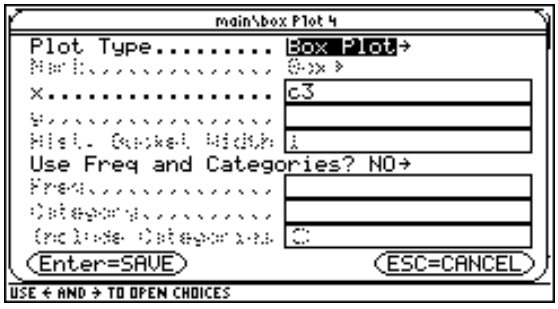
3 Streuung und Quartile

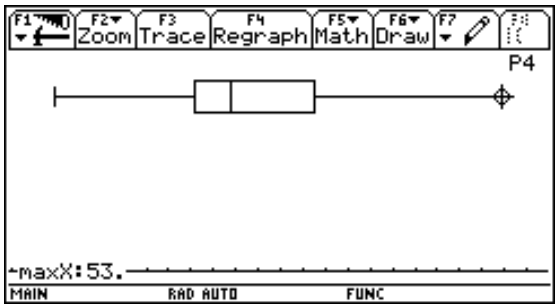
Zum Arithmetischen Mittel wurde die mittels quadrierter Differenzen berechnete Standardabweichung als Streuungsmaß definiert. Wenn man es als Extremalproblem betrachtet, passt zum Median in optimaler Hinsicht die Summe der absoluten Differenzen. Die Summe der quadrierten Abweichungen ist minimal, wenn man sie auf das arithmetische Mittel bezieht. Die Summe der absoluten Abweichungen ist minimal, wenn man sie auf den Median bezieht. Will man sich den Rechenaufwand ersparen, greift man zur Bestimmung der sogenannten Quartile. Das erste Quartil ist der Median der unteren Hälfte, das dritte Quartil ist der Median der oberen Hälfte. In einem späteren Abschnitt erläutern wir den Zusammenhang zwischen Quartil und Standardabweichung. Wie bestimmt man die Quartile? Das Schulbuch Mathematik 11.Schuljahr [1] definiert die Quartile so, als sei die Verteilung stetig. Dies aber ist selten der Fall, ja im allgemeinen stellt sich die Frage, ob bei der Bestimmung des Quartils der Median, also das zweite Quartil mitgezählt oder weggelassen wird. John Tukey zählte den Median mit.
Wenn also zum Beispiel die Stichprobe aus den Zahlen 15 26 28 29 30 32 34 40 53 besteht, ermitteln wir 30 als Median. Nach Tukey sind die beiden Hälften 15 26 28 29 30 und 30 32 34 40 53, das erste Quartil ist also 28 und das zweite ist 34.
Aus dem Beispiel in [1] (S. 127) geht nicht hervor, wie die Quartile bestimmt wurden. Elemente der Mathematik [2] (S. 119) legt sich sprachlich nicht fest, gibt aber ein nachvollziehbares Beispiel. Es folgt der Anweisung Tukeys.

Gibt man die Daten in einen Ti92 ein, so erhält man andere Quartilswerte. In der Tat streicht der Ti92 den Median aus der Rangliste und bestimmt das Quartil so:
15 26 28 29 30 32 34 40 53
15 26 28 29 32 34 40 53
Das erste Quartil ist 27 und das dritte Quartil ist 37. Eine erhebliche Abweichung von den Werten nach Tukey. Wenn wir also den Ti92 im Unterricht einsetzen, sollten wir die Schüler auf diese Problematik hinweisen und besser sofort abweichend vom Lehrbuch zum Beispiel das erste Quartil folgendermaßen definieren.
Bestimme den Median; dann bestimme den Median der Daten, deren Nummern ***kleiner*** *als die Nummer des Medians sind; das ist ein Datenwert oder die Mitte zwischen zwei Datenwerten.*
Die Unterschiede in den unterschiedlich erklärten Quartilswerten werden geringer, wenn größere Stichprobenumfänge gegeben sind, sie sind aber auch bei 20 Daten noch auffallend.

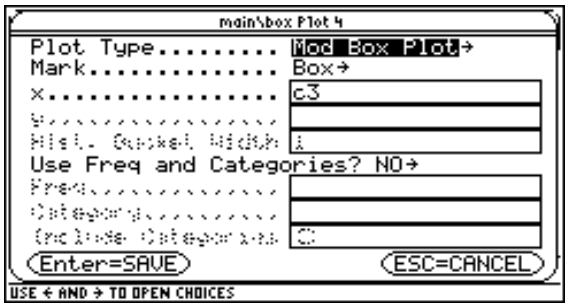
4 box-plot

Der box-plot ist auf dem Taschencomputer Ti92 implementiert, auf dem Ti92 plus sogar in zwei Varianten, die hier erläutert werden sollen.
In der statistischen Literatur existieren diverse Erklärungen, wie ein box-plot zu erstellen ist.
Ein box-plot besteht aus zwei Teilen. Der Median und das erste und dritte Quartil legen die box fest. Darüber hinaus umspannen sogenannte whiskers (Schnurrbarthaare) die komplette restliche Verteilung. Bei der Box besteht weitgehend Einigkeit in der Definition, bei den whiskers nicht.

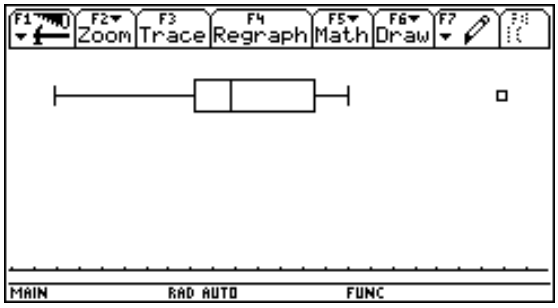




Der Ti92 zeichnet mit dem Plot-Befehl box-plot die Box nach der oben beschriebenen Methode und nimmt als Begrenzung für die whiskers den größten und den kleinsten Wert der Stichprobe.



Der Ti92plus bietet einen modifizierten box-plot-Befehl an.
Er erzeugt folgende Grafik.



Was ist passiert? Die whiskers enden jetzt bei 15 bzw. 40. Und außerhalb des Boxplots ist die 53 markiert: als Ausreißer. Das Programm hat alle Berechnungen der Quartile inklusive des Wertes 53 durchgeführt, dann aber die Daten um den Wert 53 verkürzt, und ihn gesondert markiert.
Der modifizierte box-plot des Ti92plus bestimmt die Länge der whiskers mit dem 1,5-fachen des Quartilabstandes. In der amerikanischen Literatur findet man den Abstand unter dem Namen Inter-Quartil-Range, abgekürzt IQR. Bei unserem Zahlenbeispiel beträgt der Abstand 10, drittes Quartil 37 plus 1,5 mal 10 ergibt 52. Der Datenwert 53 liegt darüber und wird als Ausreißer angesehen.
[1] setzt die Grenzen für die whiskers bei 5% und 95%, [2] wählt den einfachen Quartilabstand.

5 Ausreißer

Bei allen statistischen Untersuchungen stellen Ausreißer ein besonderes Problem dar. Oft werden sie als unbedeutend für die Gesamtheit der Stichprobe angesehen und deshalb ignoriert. Wenn sie nicht bei der groben Sichtung der Daten auffallen, können sie sich allerdings auf den arithmetischen Mittelwert und die Standardabweichung zum Teil erheblich auswirken. Sie können also Daten stark verfälschen. Der Median reagiert auf Ausreißer nicht so empfindlich.
Allerdings können vermeintliche Ausreißer auch ein Indiz sein, dass die Stichprobe weiter ausgedehnt werden sollte. Bedrückendes Beispiel ist das 1987 passierte Unglück der Raumfähre Challenger, das auf einen spröden Dichtungsring zurückzuführen ist. Die Stabilität dieses Ringes war abhängig von der Außentemperatur. Messreihen untersuchten den Zusammenhang im Bereich von 12° bis 27° Celsius. Bei 12° wurde eine hohe Empfindlichkeit festgestellt, bei allen anderen Temperaturen eine geringe. Beim Start der Challenger betrug die Außentemperatur 0°! Man hätte den vermeintlichen Ausreißer bei 12° ernst nehmen und die Reihe auf kältere Temperaturen hin fortsetzen müssen.

6 Grenzen für die whiskers

Anscheinend ist die Wahl der Grenzen für die whiskers ziemlich subjektiv. Wir überprüfen sie an einer Standard-Normalverteilung. Wie groß ist die Wahrscheinlichkeit für Werte außerhalb der whiskers-Grenzen bei den verschiedenen Festlegungen?
Die beiden Quartile entsprechen einem z-Wert von
$z_1 = -0{,}67$ und $z_2 = 0{,}67$. Daher entspricht der Quartilabstand $z_2 - z_1 = 1{,}34$.

a) **Ti92plus**
Die Grenze für die whiskers findet man bei $0{,}67 + 1{,}5*1{,}34 = 2{,}68$; die Wahrscheinlichkeit dafür ist 0,0037. Die Summe beider Seiten ergibt eine Wahrscheinlichkeit von etwa 1%. Daten mit einer Wahrscheinlichkeit von höchstens 1% gelten daher als Ausreißer.
Das entspricht einer Absicherung mit dem dreifachen Streuungsintervall bei Standardabweichung.

b) **Elemente der Mathematik**
Die Grenze für die whiskers findet man bei $0{,}67 + 1*1{,}34 = 2{,}00$; die Wahrscheinlichkeit dafür ist 0,0228. Die Summe beider Seiten ergibt eine Wahrscheinlichkeit von etwa 5%. Daten mit einer Wahrscheinlichkeit von höchstens 5% gelten daher als Ausreißer.
Das entspricht einer Absicherung mit dem zweifachen Streuungsintervall bei Standardabweichung.

c) **Mathematik 11.Schuljahr** setzt die Grenzen so, dass 10% der Daten als Ausreißer gewertet werden.

7 Wertung

Insgesamt erscheint der box-plot als eine in vieler Hinsicht informative Möglichkeit der Datenanalyse.
Er liefert auf leicht bestimmbare Art einen brauchbaren mittleren Wert. Er läßt Ausreißer sehr deutlich in Erscheinung treten. Schade, dass man sich noch nicht auf einen einzigen Modus der Darstellung geeinigt hat.

8 Literatur

[1] *Jahnke, Th. / Wuttke, H. (Hrsg.)*: Mathematik 11.Schuljahr NRW, Cornelsen, 2000
[2] *Griesel, H. / Postel, H. (Hrsg.)*: Elemente der Mathematik, Bd. 11 NRW, Schroedel, 1999

Anschrift des Verfassers:
StD Michael Spielmann, Wolfgangstr. 14, 42655 Solingen.
email: spielmann@wtal.de