Wann gebraucht man Boxplots?

Von Gary Kader und Mike Perry
übersetzt und bearbeitet von Gerhard König

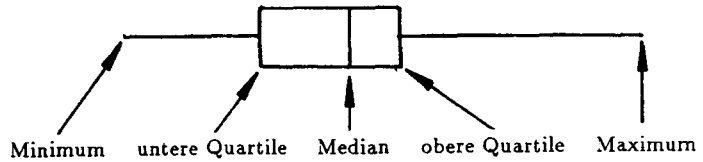
Zusammenfassung: Nach einer Beschreibung der Technik der Boxploterstellung in der Explorativen Datenanalyse werden Verwendungsmöglichkeiten diskutiert. Speziell wird aufgezeigt, wie Mißbräuche zu vermeiden sind.

Boxplots (Box-plot-Diagramme, verdeutscht auch Kastendiagramme, Kastenschaubilder) benutzt man, um die Verteilung von Daten geeignet graphisch darzustellen. Dabei werden nicht nur die einzelnen Daten dargestellt, sondern es wird auch ihre Streuung sichtbar. Besonders gut geeignet sind diese Darstellungen für den Vergleich von zwei verschiedenen Datensätzen.

Boxplots werden nicht nur in der sog. Explorativen Datenanalyse eingesetzt, sondern haben auch wegen ihrer Einfachheit und Vielseitigkeit Eingang in Curricula und Lehrplänen gefunden. Ein zu häufiger Einsatz in Büchern und im Unterricht bedeutet aber auch, daß die Gefahr eines Mißbrauchs besteht. Die Konstruktion von Boxplots mag einfach sein, deren geignete Anwendung und Interpretation ist jedoch vom statistischen Standpunkt aus vielschichtig und komplex. In diesem Beitrag wird daher etwas konkreter über mißbräuchliche Verwendungen von Boxplots berichtet und diskutiert, warum dies geschah und wie wir im Unterricht solche Mißbräuche vermeiden können.

Was sind Boxplots?

Ein Boxplot ist eine graphische Darstellung, die die Lage und Ausdehnung jedes Quartils in der Verteilung der Daten graphisch darstellt. Fünf wesentliche Kenngrößen bestimmen die Grundform eines Boxplots: Minimum, Unteres Quartil, Median, Oberes Quartil, Maximum (s.Abbildung). Manchmal werden auch noch sog. Ausreißer besonders gekennzeichnet.



Stochastik in der Schule 17(1997)1, S.3-7
Übersetzung aus Teaching Statistics 18(1996)2, S.39-41

Wie wir der Abbildung entnehmen können, werden die Werte des ersten und dritten Quartils zu einer Box vervollständigt und die Extremwerte werden durch waagerechte Linien mit der Box verbunden.

Folgende Texte führen in diese graphische Darstellungsmethode ein:

1. Biehler, R.; Steinbring, H.: Entdeckende Statistik, Stengel und Blätter, Boxplots: Konzepte, Begründungen und Erfahrungen eines Unterrichtsversuchs. In: MU 37(Nov.1991)6, S.5-32
2. Biehler, R.: Explorative Datenanalyse als Impuls für fächerverbindende Datenanalyse in der Schule. In: Computer und Unterricht 5(1994)17, S.56-66
3. Borovcnik, M.: Eine Einführung in die explorative Datenanalyse. In: Stochastik in der Schule 9(1989)3, S.5-20
4. Borovcnik, M.: Explorative Datenanalyse- Techniken und Leitideen. In: Didaktik der Mathematik 18(1990)1, S.61-80
5. Jambu, M.: Explorative Datenanalyse. Stuttgart: Fischer, 1992
6. Polasek, W.: EDA:Explorative Datenanlyse. Einführung in die deskriptive Statistik. Berlin: Springer, 1994

Wann werden Boxplots richtig genutzt?

Professionelle Statistiker benutzen Boxplots als informelle Technik zur ersten Sichtung der Datenverteilung. Boxplots sollten nur benutzt werden, wenn die zu verdichtenden Daten aus Meßwerten einer Variablen bestehen, oder wenn die zu vergleichenden Gruppen von Daten Beobachtungen derselben Variablen sind.

Die vier Boxplots in der Abb.1 sind daher z.B.geeignet, wenn es um den Vergleich des Benzinverbrauchs von vier verschiedenen Autotypen geht. Die Daten stellen dar, wieviel Meilen per Gallon (MPG) 51 verschiedene Autotypen schaffen. Jeder der Boxplots ist die Verdichtung einer Menge von Werten (13 Kompaktautos, 15 Mittelklassewagen, 12 Kleinwagen, 11 Sportautos) über derselben Variablen: MPG.

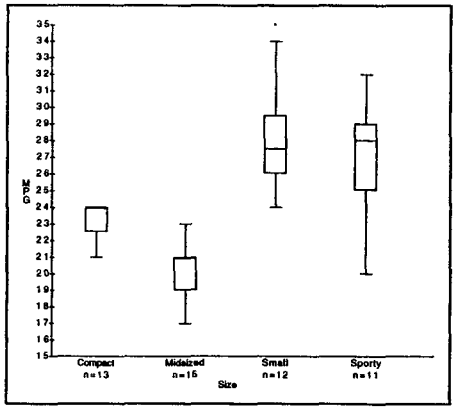


Abb.1 Miles per Gallon (MPG) und Autotypen

Mißbrauch von Boxplots

Das folgende Beispiel aus Hirsch (1992) zeigt, wie man es nicht machen soll. In Tabelle 1 wurden acht Autos einer gewissen Preisklasse von Experten in eine der dargestellten 11 Kategorien einsortiert, um das „beste" Auto herauszufinden.

Car	Engine	Transmission	Brakes	Handling	Ergonomics	Comfort	Ride	Utility	Style	Value	Fun to Drive
1	22	28	27	22	29	27	22	27	25	35	23
2	38	38	30	37	33	26	28	33	37	32	35
3	28	26	22	26	30	28	30	29	23	26	24
4	27	29	25	25	32	36	35	33	28	33	25
5	27	33	25	20	30	27	29	31	29	23	23
6	13	21	21	17	23	18	16	26	18	21	11
7	26	33	29	34	31	31	31	31	21	28	29
8	35	29	32	38	37	34	32	37	26	35	36

Tabelle 1: Einschätzungen verschiedener Autos

Jede Spalte enthält eine Menge von Beobachtungen über derselben Variablen. Jede Reihe ist ein spezieller Fall und enthält die Einschätzung der Experten in jeder der elf Variablen für ein bestimmtes Auto. Die Boxplots der elf Eingruppierungen für jedes Auto können wie in Abb.2 dargestellt werden.

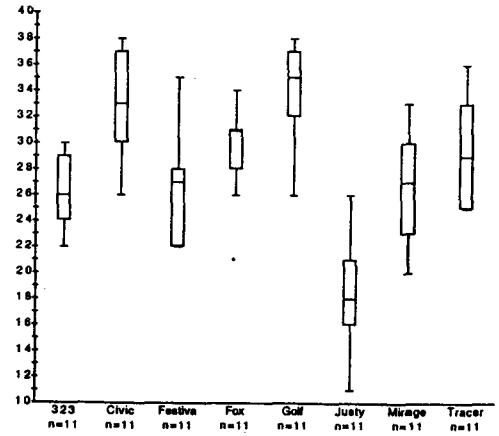


Abb.2: Boxplots der Experteneinschätzungen für die acht ausgewählten Autos

Doch ist diese Boxplotdarstellung so hilfreich und der Situation angemessen? Für jede der elf Variablen wird dieselbe Einschätzung verwendet. Trotzdem ist es nicht gut, die Eingruppierungen für jedes Auto in Boxplots darzustellen. Eine Motivation, Boxplots für jedes Auto zu benutzen, wäre, diese miteinander zu vergleichen. Auf den ersten Blick scheint dies Sinn zu machen. Wenn man jedoch genauer hinschaut, erkennt man das Problem. Um genauer zu analysieren, warum solche Darstellungen keine richtige Interpretation des gewünschten Sachverhalts liefern können, analysieren wir die beiden Boxplots der Autos x und y in Abb.3.

Jede vernünftige Interpretation würde suggerieren, daß beide Autos gleich wert seien. Die beiden Boxplots stellen die hypothetischen Daten der Tab. 2 graphisch dar. Die Mediane sind „gleich", sind aber doch wieder „verschieden". Beide Mediane haben den Wert 20. Aber der Median für das Auto x ist die Komfort-Einschätzung, für das Auto y dagegen die Freude am Fahren. Ähnliche Unterschiede gibt es für die Quartile sowie für Maximum und Minimum, obwohl die entsprechenden Werte gleich sind.

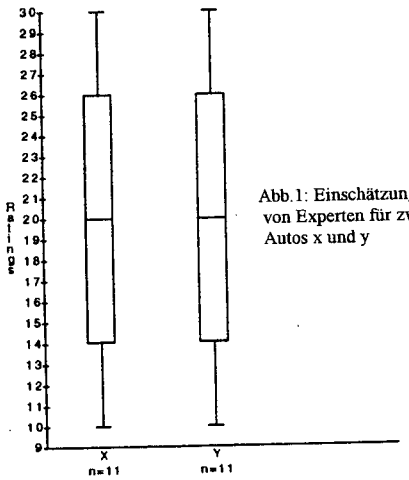


Abb.1: Einschätzungen von Experten für zwei Autos x und y

Schaut man sich Tabelle 2 genauer an, stellt man fest, daß für jede Kategorie die Werte der Einschätzungen für beide Autos sich um 10-12 Punkte unterscheiden. Das Auto x wird z.B beim Aussehen um 10 Punkte höher als das Auto y eingeschätzt, während das Auto y bei den Bremsen um 12 Punkte besser abschnitt.

Car	Engine	Transmission	Brakes	Handling	Ergonomics	Comfort	Ride	Utility	Style	Value	Fun to Drive
X-Car	10	12	14	16	18	20	22	24	26	28	30
Y-Car	22	24	26	28	30	10	12	14	16	18	20

Tabelle 1: Einschätzungen von Experten für zwei Autos x und y

Was wäre nun eine geeignete Darstellung, die verschiedenen Streuungen in den Einschätzungen der Experten bei verschiedenen Autos geeignet darzustellen. Die Profilplots der Abb. 4 vergleichen z.B. übersichtlich Vorteile und Nachteile der beiden Autos 1 und 2 aus Tabelle 1 beim Expertenvergleich.

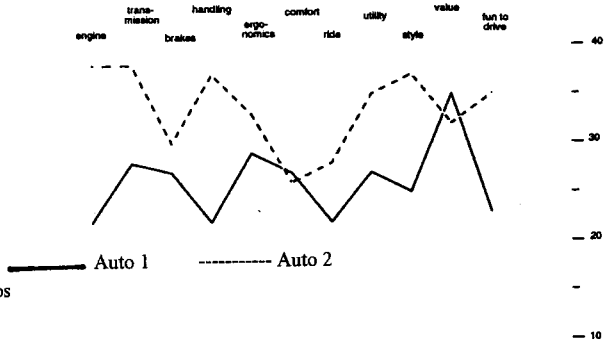


Abb.4: Profilplots zweier Autos

Der Übersetzer möchte allerdings hinzufügen, daß ihm Abb.4 als Ersatz für die sicherlich ungeeigneten Boxplots auch nicht geeignet erscheint. Die Daten sind diskret und Punkte zwischen ihnen sind bedeutungslos. Warum also die „Fieberkurve"? Ist das nicht ein gutes Beispiel für eine weitere Fehldeutung?

Die Rolle der Computer Software

Professionelle wie auch pädagogische Software haben erhebliche Fortschritte gemacht sowohl bezüglich der Anwendungsmöglichkeiten als auch hinsichtlich der Benutzerfreundlichkeit. Aus den englischsprachigen Ländern ist besonders auf das Werkzeug MINITAB hinzuweisen, daß sehr vielfältige Anwendungen der Stochastik ermöglicht, aber auch speziell für Ausbildungszwecke konzipiert ist.
Deutsche Software ist:
--EDA, das Werkzeug zur statistischen Datenanalyse. Duisburg: CoMet Verlag Verlag für Unterrichtssoftware, 1992.
--GSTAT, Statistikprogrammpaket, zusammen mit dem Buch von F.Böker: Statistik lernen am Computer. Göttingen: Vandenhoek&Ruprecht, 1989

Die Vorteile der Computer Software liegen in der praktischen Datenanalyse: Reale Daten können als Beispiel-und Übungsmaterial zur Illustration von Techniken und Strategien für die Explorative Datenanalyse herangezogen werden.

Schlußfolgerungen

Wie können wir potentiellen Mißbräuchen im Unterricht begegnen? Wir müssen die Daten in einen Problemlösungsprozeß eingliedern, könnte eine Antwort lauten. Dieser Prozeß schliesst ein: eine geeignet gestellte Frage, relevante Daten, eine richtige Analyse und sorgfältige Interpretationen der Auswertung. Eine angemessene Analyse setzt ein Verständnis der Daten voraus. Als Lehrer müssen wir sowohl Wert auf sinnvolle Daten als auch auf richtige Analysetechniken legen. Weiter müssen die Lernenden verstehen, wie die Daten gesammelt und gemessen wurden. Ferner müssen die beobachteten Variablen deutlich sein. Nur unter allen diesen Voraussetzunegn können wir erwarten, daß Mißdeutungen, wie in diesem Beitrag geschildert, vermieden werden können.

Literatur

Hirsch, C.(series editor): Curriculum and evaluation standards for school mathematics. Addenda series grades 9-12. Reston (Virginia): National Council of Teachers of Mathematics, 1992
Landwehr, J.M.; Watkins, A.E.: Exploring data. Palo Alto (California): Dale Seymour Publications, 1986 (The Quantitative Literacy Series)