STATISTIK I – Übung 07
Box-Plots und Stem-and-Leaf-Diagramme

1 Kurze Wiederholung

Warum nur zwei grafische Darstellungsformen?

Im Rahmen der Vorlesungen haben wir – kurz – eine ganze Reihe grafischer Darstellungsformen für statistische Daten betrachtet – so etwa die Balken- und Kreisdiagramme, die Histogramme oder die 2D- und 3D-Streudiagramme. Aus praktischen Erwägungen und im Sinne der Eingrenzung des Klausurstoffs, müssen aber lediglich zwei dieser grafischen Darstellungsformen in der Prüfung konstruiert werden können: Der (erweiterte) Box-Plot sowie das Stem-and-Leaf-Diagramm.

(Erweiterter) Box-Plot

Der Box-Plot ist eine der wohl spannendsten grafischen Darstellungsformen, welche die deskriptive Statistik zu bieten hat. In dieser einen Grafik finden sich komprimiert Angaben zu einer Vielzahl von Verteilungsparametern wieder, die wir in den vorangegangenen Übungseinheiten betrachtet haben. So kann man neben Lagemaßen (Median, Quartilswerte) auch Streuungsmaße (Spannweite, IQR) sowie die Form der Verteilung (linkssteil, symmetrisch oder rechtssteil) direkt aus dem Box-Plot ablesen – und sogar über das Vorhandensein von Ausreißern im Datensatz lässt sich auf Basis der Konstruktionsvorschrift für den Box-Plot eine Feststellung treffen. Der Box-Plot gestattet also Aussagen über Zentrum, Streuung, Form und Ausreißer einer Verteilung und ist damit die informationsdichteste grafische Darstellungsform, die wir im Rahmen der Vorlesung kennenlernen werden. Ein noch größeres Informationspotential entfaltet der Box-Plot übrigens beim grafischen Vergleich von Verteilungen durch das Nebeneinanderstellen mehrerer Box-Plots.

Bei der Konstruktion von Box-Plots wird in einfache Box-Plots (bei denen die Zäune jeweils bis zum größten sowie bis zum kleinsten Wert im Datensatz reichen) und in sogenannte erweiterte Box-Plots (bei deren Konstruktion die Grenzen der Zäune über den Interquartilsabstand berechnet und in denen Ausreißer und Extremwerte ausgewiesen werden) unterschieden. Nachfolgend wird in dieser Übungseinheit nur noch der erweiterte Box-Plot betrachtet. Ein solcher erweiterter Box-Plot besteht aus drei Komponenten: Der eigentlichen Box, den Zäunen der Box sowie möglicherweise einzuzeichnenden Ausreißern oder Extremwerden, sollten solche im Datensatz auftauchen. Die Konstruktion eines erweiterten Box-Plots erfolgt demnach ebenfalls in drei Schritten.

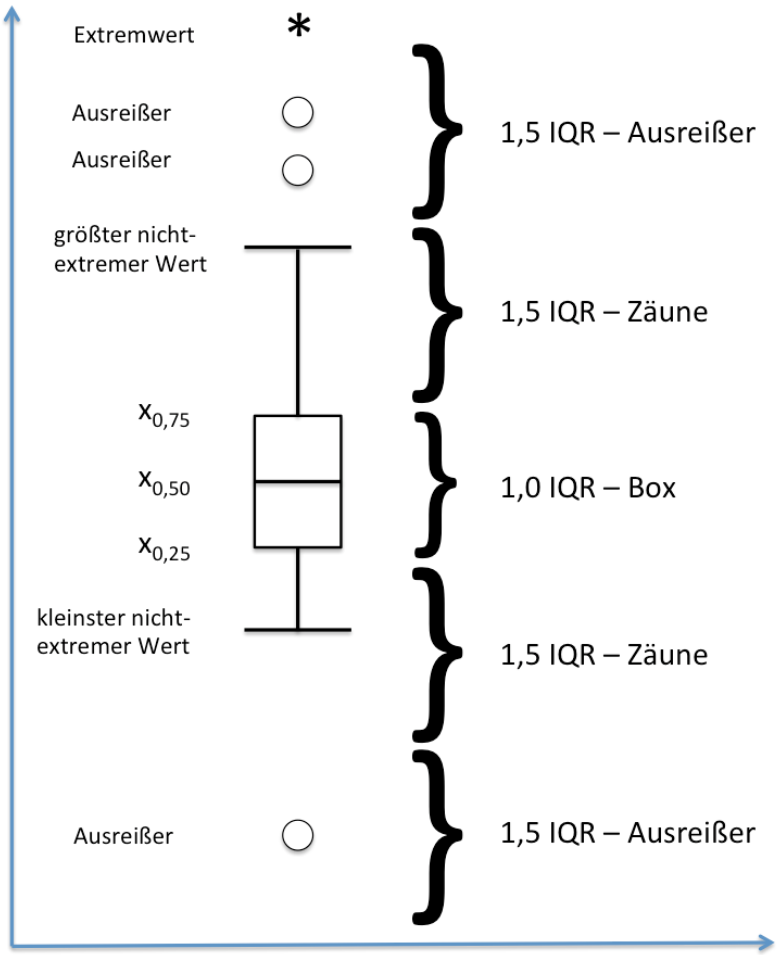
Schritt 1: Konstruktion der Box. Hierfür werden drei Werte benötigt: Das obere Quartil (obere Grenze der Box), das untere Quartil (untere Grenze der Box) sowie der Median (wird als zusätzliche Linie in die Box eingezeichnet). Sollte der Median mit einem der beiden Quartile identisch sein, wird die entsprechende Grenze mit doppelter Strichstärke gekennzeichnet. Sind die drei Quartilswerte identisch, kann keine Box konstruiert werden – in diesem Falle wird der Box durch eine dicke Linie an der Stelle $x_{0,75} = x_{0,50} = x_{0,25}$ ersetzt (Zäune und Ausreißer könnten in diesem Fall aber noch vorkommen).

Aus der Lage des Medians innerhalb der Box lässt sich übrigens eine Aussage über die Form der Verteilung herauslesen: Liegt der Median (ungefähr) in der Mitte, handelt es sich um eine symmetrische Verteilung, liegt der Median dagegen nahe der unteren Grenze der Box, so ist die Verteilung rechtsschief und linkssteil. Liegt der Median nahe an der oberen Grenze der Box, so ist die Verteilung dementsprechend rechtssteil und linksschief. Der Box-Plot kann daher (etwa in der Klausur) auch als visuelle Kontrolle für die Richtigkeit der Berechnung des Momentenkoeffizienten der Schiefe oder auch des Quartilskoeffizienten der Schiefe herangezogen werden.

Schritt 2: Konstruktion der Zäune. Da die Box vom oberen zum unteren Quartil verläuft, entspricht ihre Höhe genau dem Interquartilsabstand (IQR; Inter Quartile Range – siehe Übungseinheiten zu den Streuungsmaßen). Die Berechnung des IQR liefert uns die benötigten Angaben für das Einzeichnen der Zäune. Der 1,5-fache Wert des IQR wird nämlich zum oberen Quartilswert addiert bzw. vom unteren Quartilswert subtrahiert, um die virtuellen (nicht einzuzeichnenden) Maximal- bzw. Minimalwerte für die Grenzen der Zäune zu ermitteln. Anschließend wird der größte Wert bzw. der kleinste Werte im Datensatz ermittelt, der noch in den Bereich $x_{0,75}$ + 1,5 IQR bzw. in den Bereich $x_{0,25}$ - 1,5 IQR fällt. Der obere bzw. der untere Zaun werden dann bis zu diesen Werten gezeichnet.

In der Praxis kann es vorkommen, dass der obere Zaun, der untere Zaun oder auch beide Zäune entfallen, da keine Werte aus dem Datensatz in den benannten Bereichen liegen. Auch kann der Fall eintreten, dass Zäune genau bis zu den Maximal- bzw. Minimalwerten reichen, weil reale Werte im Datensatz exakt an der Stelle $x_{0,75}$ + 1,5 IQR bzw. $x_{0,25}$ - 1,5 IQR liegen. Von beiden Fällen sollte man sich also keinesfalls irritieren lassen – insbesondere nicht in der Klausur.

Schritt 3: Identifikation von Ausreißern und Extremwerten. Liegen Werte im Datensatz oberhalb von $x_{0,75}$ + 1,5 IQR bzw. unterhalb von $x_{0,25}$ - 1,5 IQR, handelt es sich um Ausreißer. Beim erweiterten Box-Plot wird dabei noch in Ausreißer und „extreme“ Ausreißer – die sogenannten Extremwerte - unterschieden, indem eine weitere „virtuelle“ Grenze basierend auf dem IQR errichtet wird. Werte, die zwischen $x_{0,75}$ + 1,5 IQR und $x_{0,75}$ + 3 IQR bzw. zwischen $x_{0,25}$ – 1,5 IQR und $x_{0,25}$ - 3 IQR liegen, werden als „normale“ Ausreißer mit einem Kreis markiert. Werte, die sogar noch außerhalb dieser Bereiche liegen, gelten als Extremwerte und sind mit einem Sternchen zu markieren. Sowohl die Ausreißer als auch die Extremwerte werden in der Regel noch mit der fortlaufenden Nummer des Datensatzes versehen, um diesen in nachfolgenden Untersuchungen schneller auffinden zu können.



Wie schon bei der Konstruktion der Zäune, kann es auch bei der Identifikation von Ausreißern und Extremwerten vorkommen, dass in eine oder auch beide Richtungen keine entsprechende Werte zu finden sind und daher nichts in den Box-Plot eingezeichnet wird. Auch hiervon sollte man sich – sollte ein solcher Fall in der Klausur vorkommen – also nicht irritieren lassen.

Da in der deskriptiven Statistik keine allgemeingültige Definition für Ausreißer existiert, kann es im übrigen auch an anderer Stelle (etwa bei Unklarheiten über die Einordnung eines Wertes als Ausreißer) sinnvoll sein, auf das Konstruktionsprinzip des Box-Plots zurückzugreifen. Die Unterscheidung in Ausreißer und extreme Ausreißer ist außerhalb der Box-Plot-Konstruktion allerdings eher unüblich.

Stem-and-Leaf-Diagramm

Die zweite, ebenfalls recht spezielle Form der grafischen Darstellung von Daten, die wir im Rahmen dieser Vorlesung näher betrachten, ist das sogenannte Stem-and-Leaf- oder Stamm-Blatt-Diagramm. Während die Besonderheit des Box-Plots in der Fülle der enthaltenen Informationen (Zentrum, Streuung, Form, Ausreißer) liegt, zeichnet sich das Stem-and-Leaf-Diargamm dadurch aus, dass man aus diesem als einzige grafische Darstellungsform die Originaldaten wieder herauslesen kann.

Dies ist deshalb möglich, da das Stem-and-Leaf-Diagramm nicht gezeichnet, sondern unmittelbar mit den Zahlenwerten der Originaldaten konstruiert wird. Diese werden hierfür in einen „Stamm“ sowie in an diesen angeheftete „Blätter“ zerlegt. Liegen beispielsweise nur Werte zwischen 10 und 30 vor, so bietet es sich an, den Stamm in 10er-Schritten aufzubauen, d.h.

10 |
20 |
30 |

Die konkreten Werte 11, 18, 22, 34 und 39 würden nun wie folgt in das Diagramm eingetragen:

10 | 1 8
20 | 2
30 | 4 9

Zu notieren wäre außerdem die Stammbreite (in diesem Fall 10), die Anzahl der Werte, die jeweils einem Blatt entsprechen (in diesem Fall 1) sowie ggf. noch die Anzahl der Ausreißer (in diesem Fall 0), die allein schon aus praktischen Erwägungen – ein einzelner Ausreißer beim Wert 240 würde im vorliegenden Beispiel 20 „leere“ Stammzeilen erzeugen – nicht mit eingezeichnet werden. Da in unserem Beispielfall kein Ausreißer existiert, sieht das vollständige Diagramm also wie folgt aus:

10 | 1 8
20 | 2
30 | 4 9

Stammbreite: 10
Jedes Blatt ein Fall

Betrachtet man das Stem-and-Leaf-Diagramm genauer, so fällt auf, dass es sich im Grunde um ein auf die Seite gekipptes Histogramm handelt – immer vorausgesetzt, man verwendet eine nichtproportionale Schriftart, bei der alle Zeichen exakt die gleiche Breite aufweisen (was in diesem Dokument übrigens nicht der Fall ist). Durch Abänderungen an Stammbreite und Fallzahl lassen sich Stamm-Blatt-Diagramme - insbesondere bei Verteilungen mit vielen Werten – auf äußerst vielfältige Art und Weise konstruieren. Zur Demonstration soll uns an dieser Stelle das bereits bekannte Mini-Beispiel bei einer Stammbreite von 5 dienen:

10 | 1
10 | 8
20 | 2
20 |
30 | 4
30 | 9

Stammbreite: 5
Jedes Blatt ein Fall

2 Beispielgrafiken

Auf dem Campus der Hochschule Harz haben wir 20 willkürlich ausgewählte Studierende nach ihrem Alter (in ganzen Jahren) befragt. Dabei ergab sich die folgende Verteilung:

Student	Alter	Student	Alter
1	24	11	21
2	22	12	24
3	23	13	22
4	32	14	26
5	28	15	26
6	62	16	28
7	31	17	31
8	36	18	22
9	22	19	21
10	22	20	26

Konstruktion des Box-Plots

Da für die Konstruktion des Box-Plots die Quartilswerte und der Interquartilsabstand berechnet werden müssen, lohnt sich im ersten Schritt das Festhalten der geordneten Verteilung:

21; 21; 22; 22; 22; 22; 22; 23; 24; 24; 26; 26; 26; 28; 28; 31; 31; 32; 36; 62

Ergibt (n * p) einen ganzzahligen Wert (k), berechnet sich das Perzentil wie folgt:

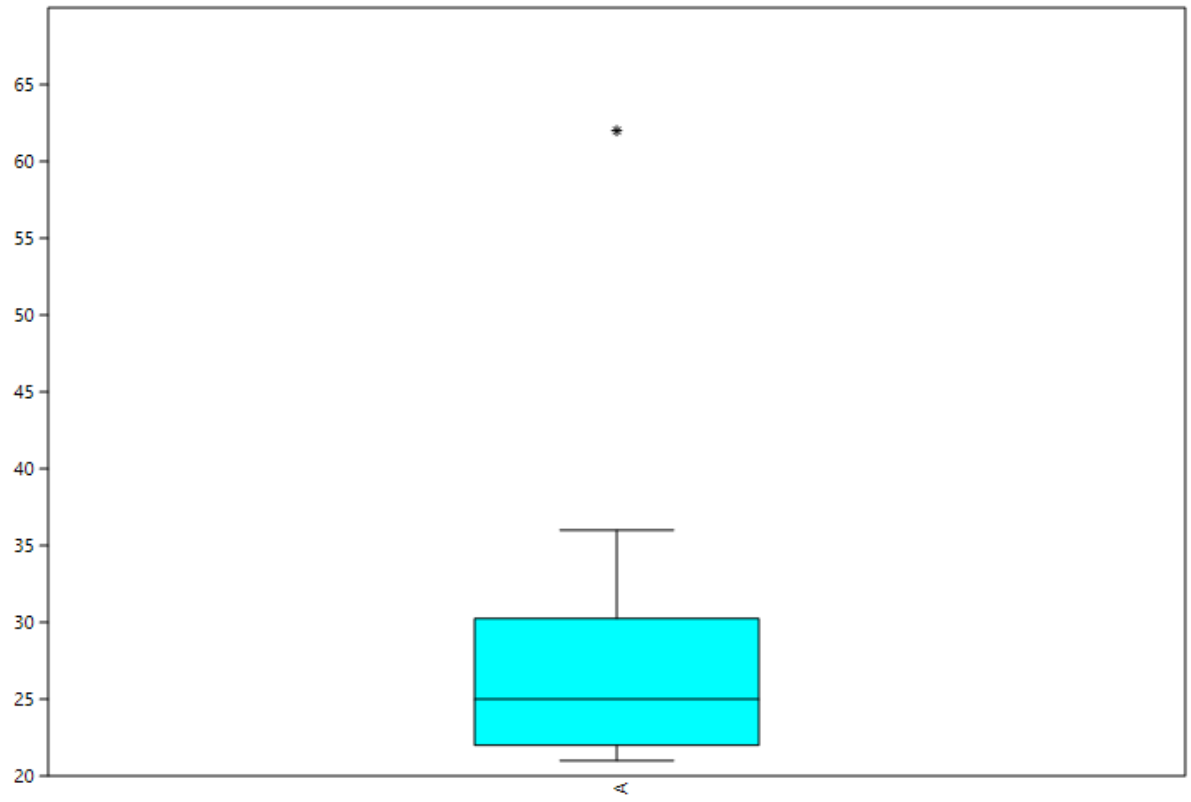
$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

(n*p) = (20 * 0,25) = 5 -> k = 5; k+1 = 6 -> xp = (22+22)/2 = 22
(n*p) = (20 * 0,50) = 10 -> k = 10; k+1 = 11 -> xp = (24+26)/2 = 25
(n*p) = (20 * 0,75) = 15 -> k = 15; k+1 = 16 -> xp = (28+31)/2 = 29,5

IQR = 29,5 – 22 = 7,5

Die Box wird im ersten Schritt also von 29,5 (obere Grenze) zu 22 (untere Grenze) gezeichnet, der Median wird bis 25 eingetragen. Da der 1,5-fache IQR bei 11,25 liegt, endet der obere Zaun beim größten Wert zwischen 29,5 und 40,75 (36), der untere Zaun beim kleinsten Wert zwischen 22 und 10,75 (21). Da die 21 bereits den kleinsten Wert der Verteilung darstellt, können im unteren Bereich der Verteilung weder Ausreißer noch Extremwerte liegen. Ausreißer im oberen Bereich der Verteilung müssten zwischen 40,75 und 52 liegen – hier finden sich in der Tabelle allerdings keine Werte. Der noch verbliebene Wert von 62 stellt damit einen Extremwert dar.

Der mit Hilfe der Software PAST berechnete erweiterte Box-Plot sieht wie folgt aus:



Konstruktion des Stem-and-Leaf-Diagramms

Auch für die Konstruktion des Stem-and-Leaf-Diagramms ist die geordnete Verteilung von Nutzen:

21; 21; 22; 22; 22; 22; 22; 23; 24; 24; 26; 26; 26; 28; 28; 31; 31; 32; 36; 62

Da alle Werte (bis auf einen) zwischen 20 und 40 liegen, führt eine Stammbreite von 10 in diesem Fall (obwohl sie durchaus nicht „falsch“ wäre) zu einem recht kurzen Diagramm, weshalb wir nachfolgend auf eine Stammbreite von 5 ausweichen wollen.

2 | 1 1 2 2 2 2 2 3 4 4
2 | 6 6 6 8 8
3 | 1 1 2
3 | 6

Stammbreite: 5
Jedes Blatt ein Fall
Ein Ausreißer (62)

3 Übungsaufgaben (Lösungen folgen in der kommenden Woche)

Parallel zur Befragung der 20 Studierenden, wurden auch 20 willkürlich ausgewählte Professorinnen und Professoren der Hochschule Harz nach ihrem Alter befragt. Dabei ergab sich folgendes Bild:

Prof.	Alter	Prof.	Alter
1	44	11	48
2	61	12	56
3	62	13	66
4	54	14	53
5	55	15	39
6	50	16	42
7	51	17	46
8	44	18	45
9	40	19	60
10	33	20	52

1) Konstruieren Sie einen erweiterten Box-Plot.

2) Konstruieren Sie ein Stem-and-Leaf-Diagramm.