Graphische Darstellung von Messresultaten in der Biologie

Martin A. Hefti

Die Visualisierung von Daten in aussagekräftigen Grafiken, wie auch die kritische Interpretation von graphischen Darstellungen, gehört zu den zentralen Fertigkeiten wissenschaftlichen Arbeitens. Im Folgenden schauen wir uns einige grundlegende Aspekte an und greifen uns dann mit dem sog. BoxPlot eine leistungsfähige Anwendung heraus.

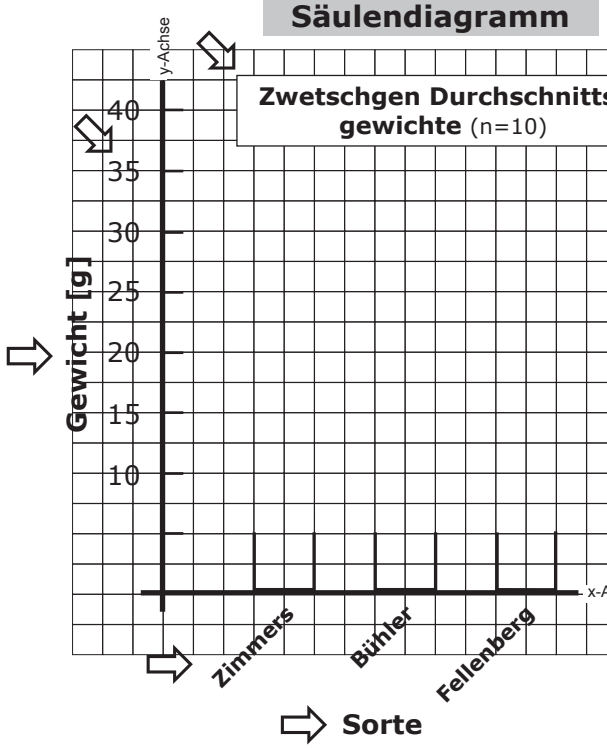
Beispiel 1
Obstbauer Minder erntet Zwetschgen. Auf seinem Land baut er drei Sorten an: die früh reifenden "Zimmers", die "Bühler" und die späteren - wie ich finde besonders feinen - "Fellenberg". 650 kg Zwetschgen hat er im letzten Jahr geerntet, wobei 235 kg von der Sorte Zimmers, 115 kg Bühler und der Rest Fellenberg war. Im letzten Jahr hat sich Herr Minder einmal die Mühe genommen, je eine Stichprobe der verschiedenen Sorten genauer zu untersuchen. Dabei erhielt er folgende Resultate bezüglich Gewicht [in Gramm, g]:

Aufgaben

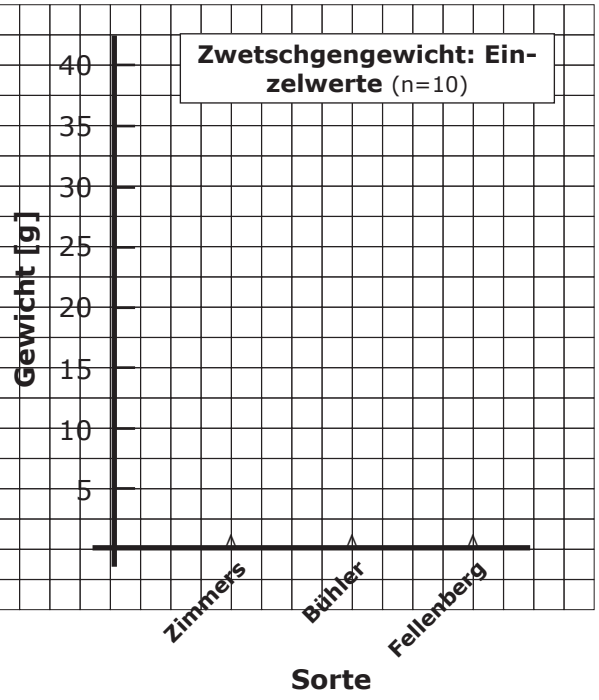
1) Berechne das Durchschnittsgewicht (ø = Mittelwert) jeder der drei Sorten.
2) Notiere die Anteile der Sorten an der Gesamternte (kg) und berechne den %-Anteil.
3) Stelle die drei Mittelwerte in einem **Säulendiagramm** dar.
4) Notiere die einzelnen Messwerte für die drei Sorten in einem **Streudiagramm**.
5) Stelle die %-Anteile der drei Sorten an der Gesamternte in einem **Säulen- und Kuchendiagramm** dar (Seite 2).

	Zimmers	Bühler	Fellenberg
1	16.8	11.2	28
2	17.7	12.8	31.2
3	18.5	14	32.2
4	18.8	15.1	33.1
5	20.2	15.8	35.5
6	15.9	10.5	16.2
7	18.2	13.8	31.7
8	16.6	11.1	26.9
9	17.2	12	28.9
10	17.5	12.5	29.9
ø			
Anteil Ernte [kg]			
Anteil Ernte [%]			

Säulendiagramm

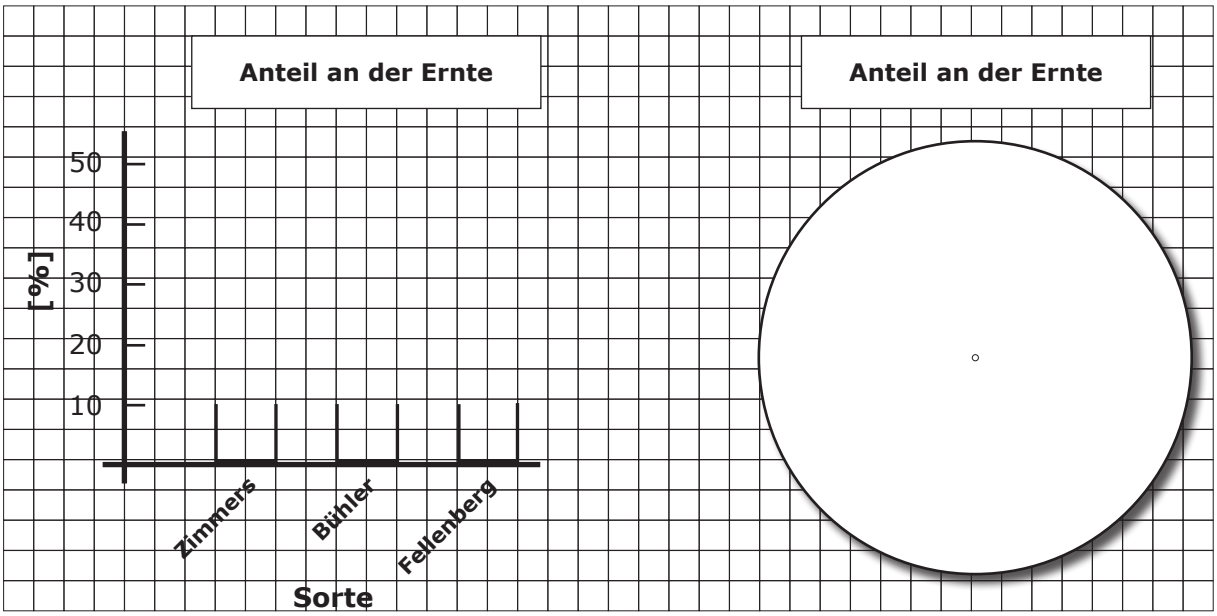


Streudiagramm



Graphische Darstellung von Messresultaten in der Biologie © M. A. Hefti • Fachdidaktik Biologie UZH

Kuchendiagramm



Die Mittelwerte in einem Säulendiagramm hat man zB in Excel schnell gezeichnet. Allerdings können extreme Einzelwerte einen Durchschnitt massiv verfälschen (siehe Beispiel 2). Zudem werden im Säulendiagramm der Mittelwerte Details unter den Tisch gewischt: ich sehe die Verteilung der Einzelwerte (also zB wie stark die Werte streuen) nicht mehr. Im Streudiagramm hingegen wird dies sichtbar und keine Werte sind unterschlagen. Allerdings ist das Zeichnen eines Streudiagramms aufwändig und bei vielen Datenpunkten nicht mehr machbar und auch nicht mehr übersichtlich.

Für eine weitergehende Datenanalyse führen wir zuerst einmal neben dem Durchschnitt (Mittelwert, arithmetisches Mittel) einen weiteren wichtigen statistischen Wert ein, den sog. **Median**.

Beispiel 2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
4	**8**	**7**	**4**	**3**	**5**	**3**	**112**	**5**	**4**

Damit die Datentabelle einfacher zu interpretieren ist, schreibe ich die 10 Daten in aufsteigender Form:

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
3	**3**	**4**	**4**	**4**	**5**	**5**	**7**	**8**	**112**

Der Mittelwert (15.5) lässt sich zwar schnell berechnen (Summe dividiert durch Anzahl Werte), ist hier aber gar nicht repräsentativ: 9 von 10 Datenpunkten meiner Verteilung liegen deutlich unter dem Mittelwert; der Datenpunkt "112" passt offensichtlich nicht in diese Reihe und verfälscht den Durchschnitt "nach oben".

Der **Median** (=Zentralwert) ist der **mittlere Datenpunkt** einer Verteilung: ich habe zehn Datenpunkte, also liegt der Median beim 5. oder 6. Wert. Bei einer ungeraden Anzahl Datenpunkte ist der Median sofort klar, bei einer geraden Anzahl Daten nimmt man den Durchschnitt der beiden mittleren Datenpunkte, hier also ((4+5)/2) = 4.5. Der Median teilt meine Datenpunkte in zwei gleich grosse Hälften. Der Median ist hier ein wesentlich robusterer Kennwert meiner Verteilung. Wenn statt 112 nämlich der Wert 1120 stünde, wäre der Median immer noch 4.5, der Mittelwert aber 116.

Box Plot (Box and whisker plot)
Ein leitungsfähiges Werkzeug der "exploratory data analysis" - der "explorativen Datenanalyse" (was etwa so viel bedeutet wie die "kritisch-erkundende/begutachtende Analyse der Daten") - ist der sogenannte Box Plot. Es ist die graphische Darstellung einiger wichtiger Parameter einer Anzahl Daten. Wenn Du diese Form der Darstellung verstanden hast, hast Du durch Dein ganzes Studium hindurch bis zur Doktorarbeit einen robusten graphischen Begleiter zur Hand, wenn es um Datenanalyse geht. Dazu müssen wir allerdings einen weiteren Begriff klären, das sog. **Quartil**. Hinter diesem Begriff versteckt sich der "Viertel". Am Einfachsten lassen sich die Parameter des Box Plots und deren Berechnung an einem konkreten Beispiel erläutern.

Beispiel 3 - So zeichne ich einen BoxPlot

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(unsortiert)	9	6	7	7	3	9	10	1	8	7	9	9	8	10	5	9	10	9	10	8
(sortiert)	**1**	**3**	**5**	**6**	**7**	**7**	**7**	**8**	**8**	**8**	**9**	**9**	**9**	**9**	**9**	**9**	**10**	**10**	**10**	**10**

Q2: der mittlere Quartilspunkt, entspricht dem Median der gesamten Datenreihe.

Hier: liegt zwischen dem 10. und 11. Datenpunkt, entspricht also 8.5.

Q1: der untere Quartilspunkt, entspricht dem Median der Daten unterhalb des Q2-Medians (also der ersten 50% der Daten). a)

*Hier: liegt zwischen dem 5. und 6. Datenpunkt, entspricht also **7**. Q_2 Median wird in diesem Beispiel nicht mitgezählt.*

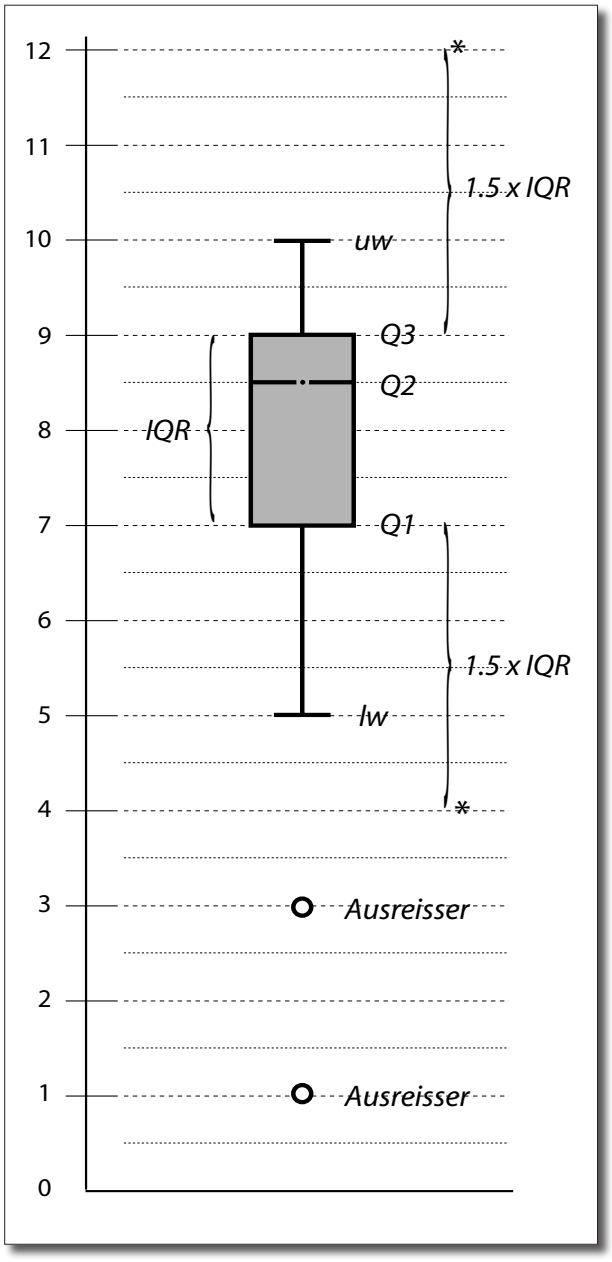
Q3: der obere Quartilspunkt, entspricht dem Median der Daten oberhalb des Q2-Medians (also der zweiten 50% der Daten). a)

*Hier: liegt zwischen dem 15. und 16. Datenpunkt, entspricht also **9**. Q_2 Median wird nicht mitgezählt.*

a) Bei einer ungeraden Zahl Daten der ***ganzen Verteilung*** zählt man den Q2-Median zur Berechnung von Q1 *und* von Q3 jeweils mit; bei einer geraden Zahl Daten der ***ganzen Verteilung*** zählt man den Q2-Median für Q1 und Q3 nicht mit.

Es herrscht übrigens in der Welt der Statistiker keine einheitliche Meinung darüber, wie die Quartilspunkte genau berechnet werden sollen. Aktuell werden mindestens fünf Varianten propagiert. Der Ti-83 zB oder Excel berechnen die Quartilspunkte je auf eine andere Weise - also nicht verwirren lassen! Die hier vorgeschlagene Methode scheint mir einfach und nachvollziehbar.

IQR: Interquartilsabstand (inter quartil range) entspricht Q3-Q1: In diesem Bereich sitzen also die mittleren 50% der Daten.
*Hier: Q3 minus Q1 (**9-7**), entspricht also **2**. Der Bereich zwischen Q1 und Q3 wird in der Grafik als **Box** dargestellt und dort der Median wie im Bsp. nebenan eingezeichnet.*



uw: upper whisker, entspricht dem letzten realen Datenpunkt innerhalb des Bereichs von 1.5x IQR oberhalb von Q3.

Hier: 1.5xIQR sind 3; ich addiere 3 zu Q3 (=12). Welcher Datenpunkt ist der letzte zwischen Q3 und 12? Es ist "10", also zeichne ich den uw bei 10.

lw: lower whisker, entspricht dem letzten realen Datenpunkt innerhalb des Bereichs von 1.5x IQR unterhalb von Q1.

Hier: 1.5xIQR sind 3; ich subtrahiere 3 von Q1 (=4). Welcher Datenpunkt ist der letzte zwischen Q1 und 4? Es ist "5", also steht der lw bei 5. Damit stehen "3" und "1" als Ausreisser fest: mit ⚬ markieren.

Ausreisser: Datenpunkte die ausserhalb von IQR ± 1.5xIQR liegen werden so markiert.

> Es gibt zum BoxPlot auch noch andere Versionen. Eine Variante zeichnet zB die lw und uw schlicht bis zu den letzten Datenpunkten (so werden keine Ausreisser definiert). Uw und lw werden zT auch anders berechnet als hier vorgeschlagen, und wiederum eine andere Variante unterscheidet zwischen "normalen" Ausreissern und "extremen" Ausreissern.

Beispiel 4

Die Stichproben der Zwetschgen von Obstbauer Minder umfassten 21 Stück pro Sorte (in Beispiel 1 haben wir nur jeweils 10 Zwetschgen untersucht); hier ist die vollständige Tabelle. Berechne zuerst die für einen BoxPlot relevanten Daten und zeichne sodann auf mm-Papier nebeneinander die BoxPlots für die drei Sorten. Skala y-Achse (wichtig!): **1cm = 2g.** Zur Vereinfachung sind die Daten bereits in absteigender Reihe notiert.

	Zimmers	**Bühler**	**Fellenberg**
1	24.8	16.6	36.2
2	20.4	16.2	36
3	20.2	15.8	35.5
4	19.2	15.3	33.3
5	18.8	15.1	33.1
6	18.6	14.5	33
7	18.5	14.0	32.2
8	18.2	13.9	31.9
9	18.2	13.8	31.7
10	17.7	13.3	31.5
11	17.7	12.8	31.2
12	17.6	12.8	30.2
13	17.5	12.5	29.9
14	17.4	12.2	29.8
15	17.2	12.0	28.9
16	17	11.5	28.7
17	16.8	11.2	28
18	16.8	11.2	27.7
19	16.6	11.1	26.9
20	15.9	10.9	25.4
21	15.9	10.5	16.2

Median (Q2)			
Q3			
Q1			
IQR			
1.5x IQR			
(Mittelwert)	—	—	—

***