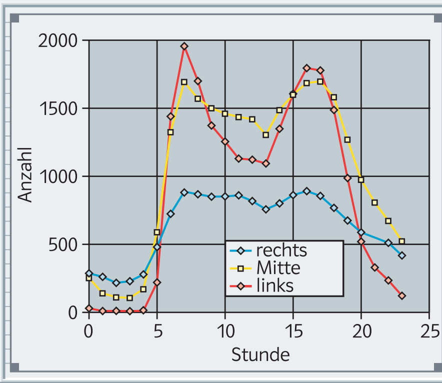
Mittelwert und empirische Standardabweichung

Die Abbildung zeigt die „Verkehrsdichte" (Anzahl der Kraftfahrzeuge je Stunde) im Verlaufe eines Tages auf drei Autobahn-Spuren. Fassen Sie die dargestellten Informationen in Worte.



Std.	rechts	Mitte	links
0	287	253	30
1	261	141	10
2	217	111	11
3	229	107	9
4	279	171	12
5	480	590	220
6	725	1325	1440
7	887	1694	1956
8	868	1571	1700
9	850	1501	1373
10	852	1461	1255
11	860	1436	1130
12	818	1420	1122
13	757	1305	1094
14	801	1487	1349
15	862	1600	1606
16	891	1686	1795
17	856	1697	1778
18	768	1582	1487
19	675	1270	987
20	588	975	520
21	550	808	330
22	511	673	235
23	418	523	122
	15290	25387	21571

Bevor man die Realität durch Wahrscheinlichkeitsmodelle beschreibt, muss man sie durch Messen, Zählen und Visualisieren „erfassen". Das geschieht in der **Beschreibenden Statistik**.

Umfragen, Verkehrszählungen, Beobachtungen zur Lebensdauer von Produkten sind Beispiele **statistischer Erhebungen**. Meist kann man nicht alle interessierenden Personen oder Produkte (die **Grundgesamtheit**) untersuchen. Man ist auf eine repräsentative **Stichprobe** angewiesen, deren Ergebnisse man in einer **Urliste** festhält.
In der Tabelle ist eine Urliste dargestellt, in der Geschwindigkeiten in einer „Tempo-30-Zone" gemessen wurden. Der **Stichprobenumfang** betrug $n = 2586$.

Urliste

Nr.	Uhrzeit T	$V\left(\frac{km}{h}\right)$	gerundet
0001	09:45:13	44,2	40
0002	09:46:04	32,8	30
…			
0123	11:09:21	52,0	50
0124	11:09:47	61,8	60
0125	11:10:04	28,2	30
0126	11:10:21	41,1	40
…			
2585	23:58:47	74,2	70
2586	23:59:13	52,6	50

Fig. 1

Eine Stichprobe heißt **repräsentativ**, wenn sie die Grundgesamtheit ausgewogen widerspiegelt. Würde man nur Jungwähler unter 22 befragen, wäre das sicher keine für alle Wähler repräsentative Stichprobe.

In Fig. 2 wurde die **Urliste** aus Fig. 1 nach Geschwindigkeiten ausgezählt, wobei „benachbarte" Geschwindigkeiten zu **Klassen** zusammengefasst wurden. So gehören alle Geschwindigkeiten zwischen $15\frac{km}{h}$ und $25\frac{km}{h}$ zur Klasse mit **Klassenmitte** $20\frac{km}{h}$. Das entspricht hier dem Runden auf Zehner. Fig. 3 zeigt das zugehörige **Säulendiagramm**.
Wenn man die relativen Häufigkeiten nicht durch Säulenhöhen, sondern durch geschlossene Rechteck**flächen** darstellt, spricht man von einem **Histogramm**. (Fig 4).

	A	B	C	D
1	Klasse	Mitte	abs. H.	rel. H.
2	$15 \leq X < 25$	20	409	15,83 %
3	$25 \leq X < 35$	30	707	27,36 %
4	$35 \leq X < 45$	40	787	30,38 %
5	$45 \leq X < 55$	50	499	19,31 %
6	$55 \leq X < 65$	60	155	6,00 %
7	$65 \leq X < 75$	70	29	1,12 %
8			2586	100,00 %
9				
10				

Fig. 2

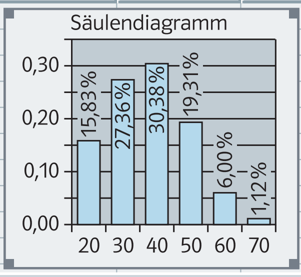


Fig. 3

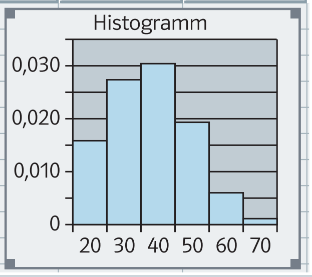


Fig. 4

Wenn man die Höhe der Säule durch die Klassenbreite teilt, erhält man die Höhe des Histogrammrechtecks.

Histogramme gestatten Häufigkeiten durch Flächen darzustellen.

Seite aus: Lambacher Schweizer Stochastik
ISBN-13: 978-3-12-735710-3

Vielfach sind die Details der Häufigkeitsverteilung von untergeordnetem Interesse, es reicht, wenn man den **Mittelwert** kennt und weiß, wie stark die Daten um den Mittelwert **streuen**.

Der Mittelwert der Geschwindigkeiten ist $\bar{x} = \frac{1}{2586}(44{,}2 + 32{,}8 + \ldots + 52{,}6) \approx 37{,}58$
Man kann ihn näherungsweise auch aus den relativen Häufigkeiten der Klassen berechnen
$\bar{x} \approx 20 \cdot \frac{409}{2586} + 30 \cdot \frac{707}{2586} + 40 \cdot \frac{785}{2586} + 50 \cdot \frac{499}{2586} + 60 \cdot \frac{155}{2586} + 70 \cdot \frac{29}{2586} \approx 37{,}53.$
Dabei ersetzt man jeden der 2584 Messwerte durch den auf die zugehörige Klassenmitte gerundeten Wert und fasst gleiche gerundete Werte zusammen.
Zum Beispiel hatte man nach dem Runden 409-mal die Geschwindigkeit 20 $\frac{\text{km}}{\text{h}}$.

Der Name Standardabweichung kommt daher, dass bei Daten wie Körperlänge von Schulanfängern oder der Intelligenzquotienten, die einer Vielzahl unabhängiger Einflüsse unterliegen, „standardmäßig" ca. 68 % aller Daten um höchstens eine Standardabweichung vom Mittelwert abweichen. **Ca. 68 % aller Daten liegen also im Intervall [$\bar{x}$ – s; $\bar{x}$ + s].** Dieser fundamentale Zusammenhang wird die folgenden Kapitel wie ein roter Faden durchziehen.

Als Maß für die Streuung der Daten um den Mittelwert benutzt man die empirische Standardabweichung
$s = \sqrt{\frac{1}{2586}((44{,}2 - \bar{x})^2 + (32{,}8 - \bar{x})^2 + \ldots + (52{,}6 - \bar{x})^2)} \approx 11{,}75$

Man kann auch hier näherungsweise mit den relativen Häufigkeiten arbeiten:

$s \approx \sqrt{(20 - \bar{x})^2 \cdot \frac{409}{2586} + (30 - \bar{x})^2 \cdot \frac{707}{2586} + \ldots + (70 - \bar{x})^2 \cdot \frac{29}{2586}} \approx 11{,}75$

Gegeben ist eine Urliste $x_1, x_2, x_3, \ldots, x_n$. Die zugehörigen Kenngrößen sind:
der **Mittelwert** $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$ und
die empirische **Standardabweichung** $s = \sqrt{\frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)}$

Wenn eine relative Häufigkeitsverteilung mit Klassenmitten $m_1, m_2, m_3, \ldots, m_k$ und den relativen Häufigkeiten $h_1, h_2, \ldots, h_k$ vorliegt, so gilt näherungsweise auch
$\bar{x} \approx m_1 h_1 + m_2 h_2 + m_3 h_3 + \ldots + m_k h_k$ und
$s \approx \sqrt{(m_1 - \bar{x})^2 \cdot h_1 + (m_2 - \bar{x})^2 \cdot h_2 + \ldots + (m_k - \bar{x})^2 \cdot h_k}$

Punkte einer Klausur
12; 9; 7; 11; 10; 10; 7; 8; 8; 8; 10; 9; 8; 10; 13; 12; 8; 8; 13; 11; 11; 7; 11; 11; 15; 8; 13; 10; 8; 13; 8; 7; 12; 7; 11; 7; 12; 10; 5; 7; 4; 10; 7; 8; 7; 6; 5; 7; 10; 6; 11; 5; 7; 6; 5; 14; 13

Beispiel Eine Stichprobe aus 57 Klausuren einer Stufe lieferte die links stehenden Punkte.
a) Ermitteln Sie die relativen Häufigkeiten, die zu den Klassen ausreichend (zwischen 4 Punkte und 6 Punkte), befriedigend (zwischen 7P und 9P), gut (zwischen 10P und 12P) und sehr gut (zwischen 13P und 15P) gehören.
b) Stellen Sie die relativen Häufigkeiten grafisch dar.
c) Berechnen Sie Mittelwert und Standardabweichung der Bewertungspunkte.
■ Lösung:

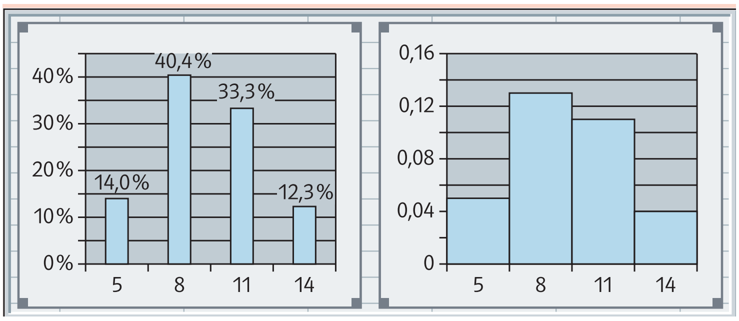


Fig. 1: Säulendiagramm Fig. 2: Histogramm

a), b) siehe Fig. 1 und Fig. 2
c) Mittelwert über Urliste:
$\bar{x} = \frac{12 + 9 + 7 + \ldots + 13}{57} = 9{,}05$
oder über die Klassenmitten
$\bar{x} \approx 5 \cdot 0{,}140 + 8 \cdot 0{,}404 + \ldots + 14 \cdot 0{,}123 = 9{,}32$

Standardabweichung über Urliste:
$s = \sqrt{\frac{(12 - 9{,}05)^2 + \ldots + (14 - 9{,}05)^2}{57}} = 2{,}59$
oder über Klassenmitten
$s \approx \sqrt{(5 - 9{,}32)^2 \cdot 0{,}14 + \ldots + (14 - 9{,}32)^2 \cdot 0{,}123}$
$= 2{,}64$

Seite aus: Lambacher Schweizer Stochastik
ISBN-13: 978-3-12-735710-3

Aufgaben

1 a) Gegeben sind zehn rote und zehn blaue Zahlen. Schätzen Sie den Mittelwert und die Standardabweichung und kontrollieren Sie Ihre Schätzungen durch Nachrechnen.
b) Welcher Mittelwert, welche Standardabweichung ergibt sich, wenn man die einzelnen Zahlen ganzzahlig rundet?
10 6,1 10,3 7,1 8,2 12,6 10,4 5,6 1,1 10,6
9,7 5,4 3,9 8,1 3,3 4,5 4,8 12,2 8,4 7,3

2 a) Berechnen Sie Mittelwert $\overline{x}$ und Standardabweichung s der folgenden Paketgewichte.
2,7 2,5 2,3 1,1 2,0 2,4 2,6 2,6 2,2 1,7 1,0 2,7 2,9 1,8 2,9 1,8 1,6 0,6 1,6 2,4
b) Wie viel Prozent der Gewichte liegen im Intervall $[\overline{x} - s;\ \overline{x} + s]$?

Sie können den Taschenrechner nutzen (s. Infokasten Seite 4).

3 a) Berechnen Sie aus den Angaben von Fig. 1 und von Fig. 2 die Mittelwerte $\overline{x}$ und die Standardabweichungen s der Altersverteilungen in der Jahrgangsstufe 6 und in der Jahrgangsstufe 12.
b) Wie erklären Sie, dass die Standardabweichung in der Jahrgangsstufe 12 größer ist?
c) Wie müssten Sie die Säulendiagramme verändern, um „echte Histogramme“ zu erhalten, bei denen relative Häufigkeiten Flächen und nicht Säulenlängen entsprechen?

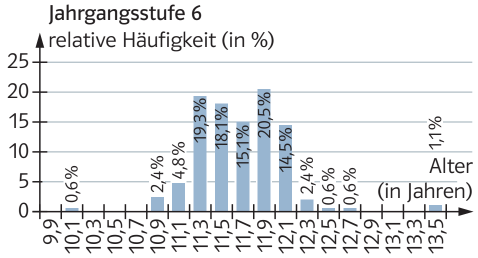


Fig. 1

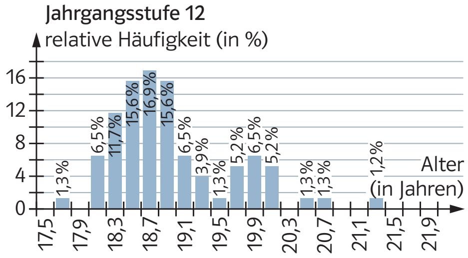


Fig. 2

In dieser Stichprobe liegt das Alter von 73 % der 166 Schüler der 6. Klassen zwischen $\overline{x} - s$ und $\overline{x} + s$.
In Stufe 12 (77 Schüler) liegt der Prozentsatz mit 71,5 % näher am 68 %-Wert.

4 In einer Klasse 5 wurden Fehlstunden F und Zeugnisnote N erhoben.

Name	F	N	Name	F	N	Name	F	N	Name	F	N	Name	F	N	Name	F	N
Nadja	19	2,43	Nadine	0	2,50	Simone	18	3,25	Gülsah	15	2,71	Leonard	22	3,29	Onur	0	3,29
Songül	2	2,57	Elif	6	2,57	Stephanie	19	2,88	Markus	48	2,38	Stefan	0	2,88	Tina	12	2,88
Dirk	4	2,00	Janina	17	2,38	Cigdem	5	2,43	Michael	14	2,75	Dominique	32	2,25	Esther	6	2,50
Ataelahi	6	3,00	Tim	0	3,25	Sezen	9	3,00	Sabine	0	2,50	Sascha	0	2,63	Sebastian	20	1,88
Florian	2	2,13	Paul	0	2,25	Patrik	2	3,38	Marius	0	2,88	Melek	16	3,57	Nino	9	2,38

a) Berechnen Sie die zugehörigen Kennwerte Mittelwert und Standardabweichung.
b) Stellen Sie die Fehlstunden in einem Säulendiagramm dar (Klasseneinteilung $0 \le F \le 10$, $11 \le F \le 20$, …, $41 \le F \le 50$).
c) Visualisieren Sie die Zeugnisnoten durch ein Histogramm (Klasseneinteilung $1 \le N \le 1{,}5$, $1{,}5 < N \le 2$, …, $2{,}5 < N \le 3$).

Auch hier ist der Einsatz des Taschenrechners sinnvoll.

Zeit zu überprüfen

5 Liana hat 10-mal gewürfelt: 3-1-5-3-4-6-2-6-1-2. Berechnen Sie den Mittelwert und die Standardabweichung der Augenzahl. Zeichnen Sie ein Säulendiagramm.

6 a) Würfeln Sie 10-mal und berechnen Sie den Mittelwert und die Standardabweichung.
b) Lassen Sie einen GTR oder Excel 100-mal würfeln. Stellen Sie die Ergebnisse grafisch dar. Berechnen Sie den Mittelwert und die Standardabweichung.

Excel „würfelt“ mit dem Befehl = ZUFALLSBEREICH (1; 6)

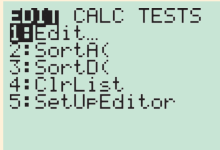
Seite aus: Lambacher Schweizer Stochastik
ISBN-13: 978-3-12-735710-3

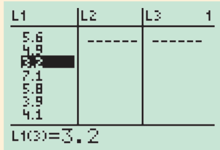
INFO → Aufgabe 7

Arbeiten mit dem GTR

Bei längeren Urlisten lohnt sich der Einsatz von Tabellenkalkulation oder GTR. Als Beispiel untersuchen wir die folgende Urliste:
5,6 4,9 3,2 7,1 5,8 3,9 4,1 5,2 8,6 4,7 5,1 4,0 6,5 5,0 4,8 5,3 3,8 7,2 5,5 4,6

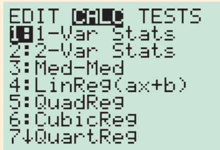
Zunächst wird die Urliste in eine Liste des GTR übertragen (hier L1).

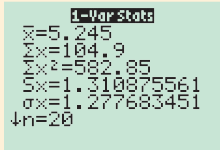




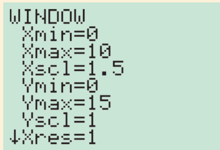
Die empirische Standardabweichung wird bei einigen Taschenrechner-Modellen mit σx bezeichnet.

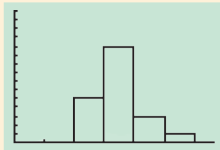
Im Statistik-Menü lassen sich bereits alle benötigten Informationen abrufen:
Mittelwert: $\bar{x} = 5{,}245$
Standardabweichung (σx): $s \approx 1{,}278$





Für ein Säulendiagramm müssen noch die Klassen festgelegt werden. Mit den gezeigten Einstellungen erhält man die Klassen:
$0 \le X < 1{,}5$; $1{,}5 \le X < 3$
usw.





7 Gegeben sind die folgenden fünf blauen und fünf roten Nettolöhne:
7000 €, 300 €, 240 €, 5500 €, 210 €
2600 €, 2700 €, 2650 €, 2800 €, 2500 €
a) Bestimmen Sie für beide Listen den Mittelwert und die Standardabweichung.
Welche Bedeutung hat der Mittelwert hier?
Warum ist die Angabe des Mittelwerts alleine nicht aussagekräftig?
b) Berechnen Sie das „Streumaß“ mit folgender alternativer Formel:
$s_{\text{alternativ}} = \frac{1}{n}\left(|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|\right)$.
Vergleichen Sie die Ergebnisse mit der in Teilaufgabe a) berechneten Standardabweichung.
Welcher Effekt wird durch die Standardabweichung in a) deutlicher betont?

8 Boxplots
Bei einem Boxplot (Fig. 1), wird die der Größe nach geordnete Datenmenge durch q_u (unteres Viertel), den Median (Zentralwert) und q_o (oberes Viertel) in vier Teile geteilt.
Informieren Sie sich über die Darstellung von Daten mithilfe eines Boxplots. Stellen Sie damit die folgende Datenmenge dar:
7,5 6,8 4,2 5,1 10,0 6,9 5,9 4,5 7,2 5,8.
Welche Informationen können aus einem Boxplot abgelesen werden? Bereiten Sie dazu ein kurzes Referat mit selbst gewählten Beispielen vor.

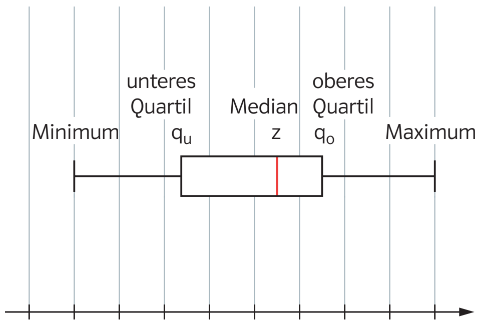


Fig. 1

Seite aus: Lambacher Schweizer Stochastik
ISBN-13: 978-3-12-735710-3