Mathematisches Unterrichtswerk
für das Gymnasium
Ausgabe Nordrhein-Westfalen

erarbeitet von
Manfred Baum
Wolfgang Riemer
Hartmut Schermuly
Jörg Stark
Ingo Weidig
Peter Zimmermann

unter Mitwirkung von
Detlef Lind
Günther Taetz

Ernst Klett Verlag
Stuttgart Leipzig

An der Entstehung des Gesamtwerks waren weiterhin beteiligt:
Gerhard Brüstle, Heidi Buck, Günther Dopfer, Rolf Dürr, Hans Freudigmann, Rolf Reimer, Maximilian Selinka

Bildquellenverzeichnis:
Archiv für Kunst und Geschichte, Berlin: S. 50 b) (© The Munch Museum/ The Munch Ellingsen Grou VG Bild-Kunst, Bonn 1999), 50 c), 50 d), 107 (unten) – Alexander Weltatlas, Klett-Perthes, Gotha: S. 23 (rechts) – Bavaria, München: S. 86 (SSI, oben), 127 (Vega) – bild der wissenschaft, Mai 1983, S. 102, Stuttgart: S. 87 – Deutsches Museum, München: S. 100 (unten), 107 (oben), 108, 110 (Mitte), 110 (unten), 130 – Fischerwerke, Tumlingen: S. 110 (oben) – Fotoagentur Helga Lade, Frankfurt: S. 17 (G. Schneider), 23 (Ott, unten links), 145 (Bramaz) – IVB -Report, Kappelrodeck: S. 9 (Mitte) – Lande vermessungsamt Nordrhein-Westfalen: S 19 (unten) – Mauritius, Stuttgart: S. 115 (Scott) – Moro, C., Stuttgart: S. 8, 48, 50 a), 52, 83, 86 (Mitte und unten) – Riemer, Köln: S. 47, 49, 54, 58, 61, 63, 65, 66, – Stadtvermessungsamt Tübingen: S. 19 (oben) – Vehrenberg KG, Düsseldorf: S. 32 – Verkehrsverbund Rhein-Ruhr GmbH, Gelsenkirchen: S. 98 – Weidig, Landau: S. 9 (oben), 164 (oben rechts und Mitte) – zefa visual media, Düsseldorf: S. 89 (Rose)

Nicht in allen Fällen war es uns möglich, den uns bekannten Rechtsinhaber ausfindig zu machen. Berechtigte Ansprüche werden selbstverständlich im Rahmen der üblichen Vereinbarungen abgegolten.

1. Auflage 1 16 15 14 13 | 2013 2012 2011 2010

Alle Drucke dieser Auflage können im Unterricht nebeneinander benutzt werden, sie sind unt einander unverändert. Die letzte Zahl bezeichnet das Jahr dieses Druckes.

Internetadresse: http://www.klett.de

Zeichnungen: U. Bartl, Weil der Stadt; H. Günthner, Stuttgart; R. Hungreder, Leinfelden.
Umschlaggestaltung: Alfred Marzell, Schwäbisch Gmünd.
DTP-Satz: topset Computersatz, Nürtingen.
Druck: Druckhaus Götz GmbH, 71636 Ludwigsburg. Printed in Germany.

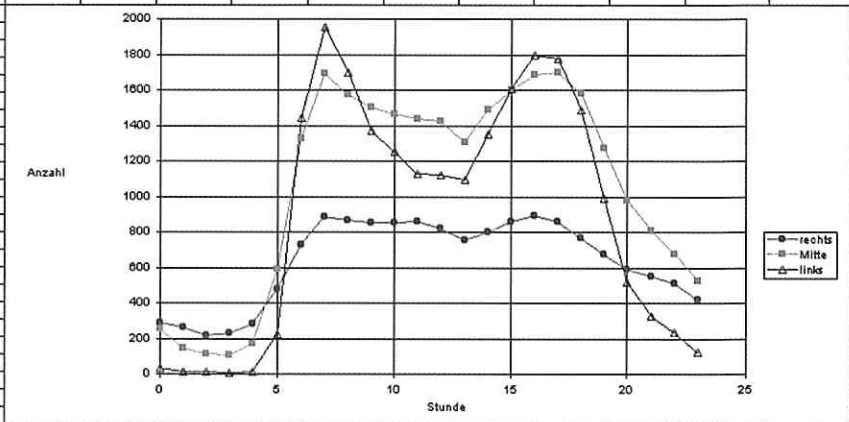
ISBN 3-12-732210-0

1 Daten erheben und darstellen

Die Urlisten finden Sie in Autobahn.xls.

1 Auf Autobahnen wird laufend die Verkehrsdichte (Anzahl der Autos je Stunde) gemessen Fig. 1 zeigt die Verkehrsdichte im Verlaufe eines Tages auf den drei Spuren des Kölner Autobahnrings. Erläutern Sie die dargestellten Informationen.

	I	J	K	L
1		Fahrspur		
2	Std.	rechts	Mitte	links
3	0	287	253	30
4	1	261	141	10
5	2	217	111	11
6	3	229	107	9
7	4	279	171	12
8	5	480	590	220
9	6	725	1325	1440
10	7	887	1694	1956
11	8	868	1571	1700
12	9	850	1501	1373
13	10	852	1461	1255
14	11	860	1436	1130
15	12	818	1420	1122
16	13	757	1305	1094
17	14	801	1487	1349
18	15	862	1600	1606
19	16	891	1686	1795
20	17	856	1697	1778
21	18	768	1582	1487
22	19	675	1270	987
23	20	588	975	520
24	21	550	808	330
25	22	511	673	235
26	23	418	523	122
27		15290	25387	21571

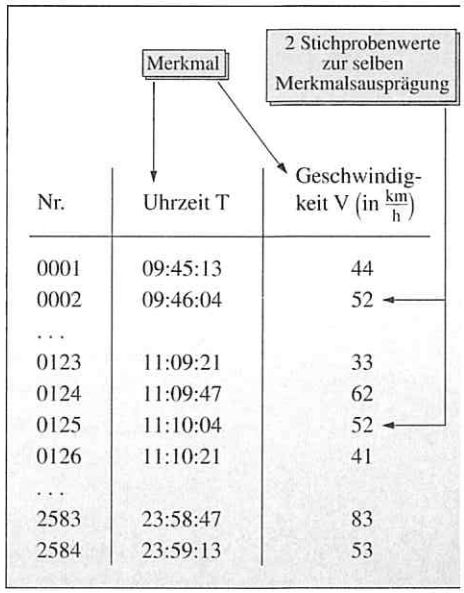


Fig

Die in diesem Kapitel genannten Dateien stehen im Mathematikbereich der Klett-Homepage. Internetadresse: http://www.klett.de

Befragungen von Personen oder das Zählen von Gegenständen sind Beispiele **statistischer Erhebungen**. Bei einer solchen Erhebung wird an **Merkmalsträgern** (z. B. Personen oder Autos) ein bestimmtes Merkmal (z. B. Körpergröße oder gefahrene Geschwindigkeit) untersucht. Alle Merkmalsträger zusammen nennt man Grundgesamtheit. Ausgewählte Merkmalsträger bilden eine Stichprobe. Die Anzahl n der Merkmalsträger in der Stichprobe heißt **Stichprobenumfang**.

Bei einer statistischen Erhebung werden die Daten in einer **Urliste** wie in Fig. 2 festgehalten. Die einzelnen Beobachtungswerte der Urliste heißen **Stichprobenwerte**. Sie können mehrfach vorkommen. Davon zu unterscheiden sind die **Merkmalsausprägungen**. Diese sind alle voneinander verschieden.



Fig

In der Urliste von Fig. 2 sind die Merkmalsträger Autos, bei denen die Merkmale Zeit (T) und Geschwindigkeit (V) gemessen wurden. Der Stichprobenumfang ist $n = 2584$.
Merkmalsausprägungen sind hier die ganzzahligen Geschwindigkeiten (etwa zwischen $0\,\frac{km}{h}$ un $200\,\frac{km}{h}$). Natürlich müssen bei 2584 Stichprobenwerten etliche mehrfach auftreten. Beispielsweise gilt $v_{0002} = v_{0125} = 52\,\frac{km}{h}$.

Die Auswertung einer statistischen Erhebung kann durch Angabe absoluter bzw. relativer Häufigkeiten erfolgen. Da sich diese Häufigkeiten auf die Merkmalsausprägungen verteilen, spricht man von **Häufigkeitsverteilungen**.

Zur Erinnerung:
relative Häufigkeit = $\frac{\text{absolute Häufigkeit}}{\text{Versuchszahl}}$
Z. B.: $\frac{409}{2548} \approx 0{,}1583 = 15{,}83\,\%$

Fasst man in einer Urliste bei dem Merkmal Geschwindigkeit „benachbarte" Geschwindigkeiten zu sog. **Klassen** zusammen, erhält man z. B. die Häufigkeitsverteilung von Fig. 1.
Man kann sie beispielsweise durch ein **Säulendiagramm**, ein **Punktdiagramm** oder ein **Kreisdiagramm** grafisch darstellen (Fig. 2).

Häufigkeitsverteilung der Geschwindigkeiten bei einer Polizeimessung im Ortsbereich

Klasse: V (in $\frac{km}{h}$)	Klassenmitte	abs. H.	rel. H.
$20 < V \leqq 30$	25	409	15,83 %
$30 < V \leqq 40$	35	707	27,36 %
$40 < V \leqq 50$	45	785	30,38 %
$50 < V \leqq 60$	55	499	19,31 %
$60 < V \leqq 70$	65	155	6,00 %
$70 < V \leqq 80$	75	29	1,12 %
		2584	100,00 %

Fig. 1

In Zeitungen werden zur grafischen Darstellung oft auch Piktogramme verwendet. Wie man hiermit oder z. B. mit irreführenden Skalierungen der Achsen „mogeln" kann, erfahren Sie in der Exkursion auf S. 80/81.

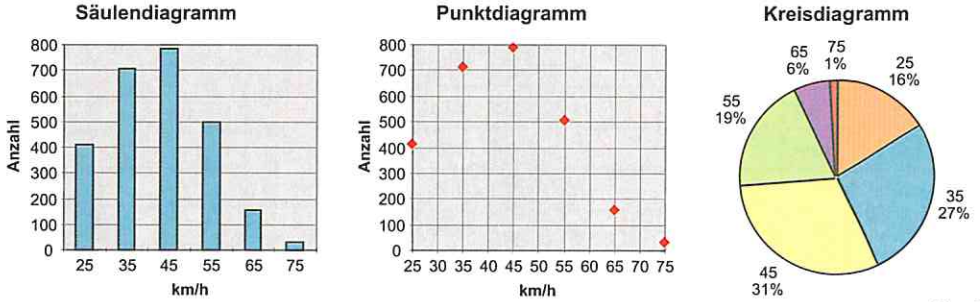


Fig. 2

Vorgehen bei der Durchführung und Auswertung einer statistischen Erhebung:
1. Merkmal X festlegen.
2. Urliste mit den Stichprobenwerten $x_1, x_2, \ldots, x_n$ notieren; Stichprobenumfang n.
3. Evtl. Klassen mit Klassenmitten $k_1, \ldots, k_m$ festlegen; Anzahl der Klassen m.
4. Häufigkeiten der Merkmalsausprägungen bzw. der Klassen bestimmen.
5. Diagramm zeichnen.

1,7	2,7	3,5	2,0	2,7
2,4	3,3	2,9	3,2	3,0
2,5	2,7	3,2	2,6	2,2
1,8	3,1	3,2	1,6	2,2
2,3	3,3	2,2	3,1	1,0
3,0	1,8	2,7	3,0	1,9
3,1	3,3	1,8	3,3	3,2
3,6	1,8	2,4	3,6	3,3
3,3	2,7	3,3	3,0	3,2
3,1	2,2	2,3	1,8	3,2
2,2	1,5	3,0	2,6	3,1
1,9	2,7			

Fig. 3

Beispiel 1: (Ermittlung und Darstellung einer Häufigkeitsverteilung)
In einer Stichprobe aus 57 Abiturienten eines Jahrgangs wurde das Merkmal „Abiturnote" erhoben. Es ergab sich die Urliste von Fig. 3.
Ermitteln Sie die relativen Häufigkeiten, die zu den Merkmalsklassen sehr gut ($1{,}0 \leqq X \leqq 1{,}5$), gut ($1{,}6 \leqq X \leqq 2{,}5$), befriedigend ($2{,}6 \leqq X \leqq 3{,}5$) und ausreichend ($3{,}6 \leqq X \leqq 4{,}0$) gehören.
Stellen Sie die Verteilung der relativen Häufigkeiten in einem Säulendiagramm dar.
Lösung:

Klasse	sehr gut	gut	befr.	ausr.
Strichliste	||	𝍸 𝍸 𝍸 𝍸	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	||
abs. H.	2	20	33	2
rel. H.	3,51 %	35,09 %	57,89 %	3,51 %

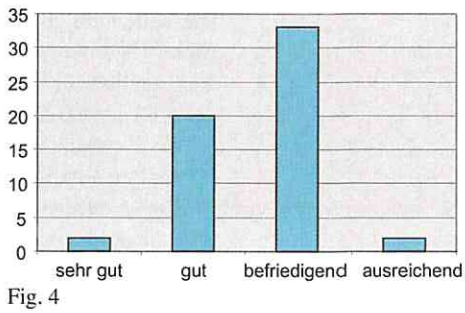


Fig. 4

Zum Umgang mit Tabellenkalkulation

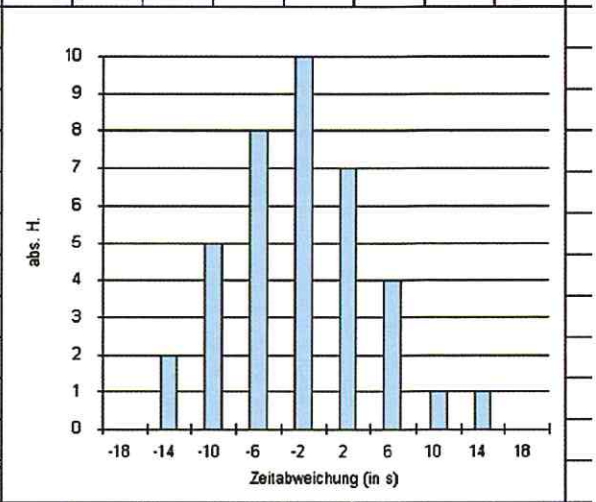
In der Praxis verwendet man zur Auswertung und Präsentation statistischer Erhebungen Tabellenkalkulationsprogramme wie z. B. EXCEL. Sie besitzen Befehle, mit denen man Urlisten (auc mit Klasseneinteilungen) auszählen und Häufigkeitsverteilungen grafisch darstellen kann.

Beispiel 2: (Auswertung einer statistischen Erhebung mit Tabellenkalkulation)
Bei einem Experiment wurden 38 Schülerinnen und Schüler gebeten, nach einem Startsignal den Zeitpunkt zu benennen, an dem ihrem Gefühl nach eine Minute verstrichen war. Als Merk mal X enthält Fig. 1 in den Zellen B2 bis B39 die (vom Versuchspartner gestoppte) Zeit in Sekunden, um die man sich „verschätzt" hatte.

Die Datei Zeitschätzung.xls enthält alle Daten zu Beispiel 2. In Fig. 1 sind die Zeilen 13 bis 37 ausgeblendet.

a) Ermitteln Sie mit Hilfe von EXCEL die absoluten und die relativen Häufigkeiten, mit denen d Zeitabweichungen in den Klassen $-20 < X \leqq -16$, $-16 < X \leqq -12, \ldots, 16 < X \leqq 20$ liegen.
b) Erstellen Sie ein Säulendiagramm.

	A	B	C	D	E	F	G
1			rechte Grenze	Klassenmitte	abs. H.	rel. H.	
2	Moritz	-3,4	-16	-18	0	0,0%	
3	Dominic	-4,8	-12	-14	2	5,3%	
4	Tim	-12,6	-8	-10	5	13,2%	
5	Dennis	-11,5	-4	-6	8	21,1%	
6	Gabi	-6,7	0	-2	10	26,3%	
7	Mattias	-2,1	4	2	7	18,4%	
8	MarkusH	-1,0	8	6	4	10,5%	
9	Daniel	-5,7	12	10	1	2,6%	
10	Simone	-9,7	16	14	1	2,6%	
11	MarkusK	-9,7	20	18	0	0,0%	
12	Paul	-15,4		Summe ->	38	100,0%	
38	Darius	5,0					
39	Volker	14,7					



Fig

Lösung:

Beachten Sie: Andere Tabellenkalkulationsprogramme haben gegenüber EXCEL etwas abweichende, aber ähnliche Formeln und Befehle.

a) Die rechten Klassengrenzen $-16, -12, \ldots, 20$ werden untereinander notiert. In Fig. 1 stehen sie im Bereich C2 bis C11. Ein gleich großer Bereich (E2 bis E11) wird für die Verteilung der absoluten Häufigkeiten markiert und mit der Formel {= HÄUFIGKEIT(B2:B39;C2:C11)} belegt. Die geschweiften Klammern werden nicht eingegeben, sie entstehen, wenn man die Eingabe der Formel durch die Tastenkombination „⇑ Strg ↵" beendet. Die Klammern deuten an, dass sich die Formel auf den markierten Bereich E2 bis E11 bezieht („Vektorformel").
Die Zellen aus solchen Bereichen können nicht einzeln geändert werden, man kann sie nur als Einheit bearbeiten. Im Unterschied dazu kann man die Klassengrenzen im Bereich C2 bis C11 auch nachträglich einzeln ändern und die Veränderungen in der Häufigkeitstabelle studieren. In Zelle E12 steht der Befehl = SUMME(E2:E11), der die Zahl der Versuchsteilnehmer liefert. Zelle F2 enthält eine relative Häufigkeit, die mit der Formel = E2/E$12 berechnet wird und sich in den Bereich F2 bis F12 kopieren lässt.
b) Zur Erstellung eines Diagramms wird der Bereich der Klassenmitten und Häufigkeiten D2 bis E11 markiert. Dann wählt man „Einfügen, Diagramm" und lässt sich von EXCELS Diagramm-Assistenten führen.

Aufgaben

2 Geben Sie die relativen Häufigkeiten als Bruch, als Dezimalzahl und in Prozent an.

a) 38 von 100 b) 12 von 50 c) 3 von 27
d) 4 von 28 e) 32 von 67 f) 45 von 30

3 Die Stichprobenumfänge zweier Untersuchungen sind 80 bzw. 531. Wie groß ist jeweils die absolute Häufigkeit einer Merkmalsausprägung, wenn für ihre gerundete relative Häufigkeit angegeben wird:

a) 50 %, b) 47 %, c) 82 %, d) 0,16,
e) 0,01, f) ein Siebentel, g) zwei von drei?

4 a) Nach der Untersuchung von 2325 Rindern wird die relative Häufigkeit der an Brucellose erkrankten Tiere mit 0,04 angegeben. Wie viele der untersuchten Tiere waren von der Krankheit befallen?

b) Wie genau ist die von Ihnen in a) berechnete Anzahl?

5 Fig. 1 zeigt die Ergebnisse einer Umfrage unter 18 773 Führerscheinbesitzern.

a) Erläutern Sie an diesem Beispiel die Begriffe Grundgesamtheit, Stichprobe, Stichprobenumfang, Merkmal, Merkmalsausprägung, Stichprobenwert.

b) Die Summe der Prozentangaben in den einzelnen Zeilen ist kleiner als 100 %. Wie viel Prozent fehlen jeweils?

c) Welche Bedeutung haben diese fehlenden Angaben und warum werden sie mit steigendem Einkommen kleiner?

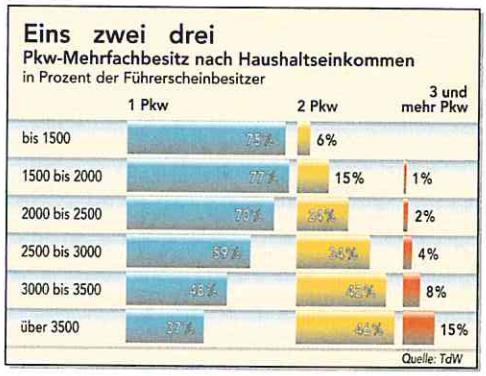


Fig. 1

6 Eine Bürgerinitiative ließ einen Tag lang die Geschwindigkeiten in einem Wohngebiet am Ortsausgang von Köln-Auweiler ($50\frac{\text{km}}{\text{h}}$-Zone) messen.

a) Wie viel Prozent der Autos fuhren schneller als erlaubt? Wie beurteilen Sie die Messergebnisse im Hinblick auf das vermutliche Ziel der Bürgerinitiative?

b) Wie lässt sich unter Beachtung des Kartenausschnittes die Tatsache erklären, dass die „Geschwindigkeitsverteilung" zwei Gipfel besitzt?

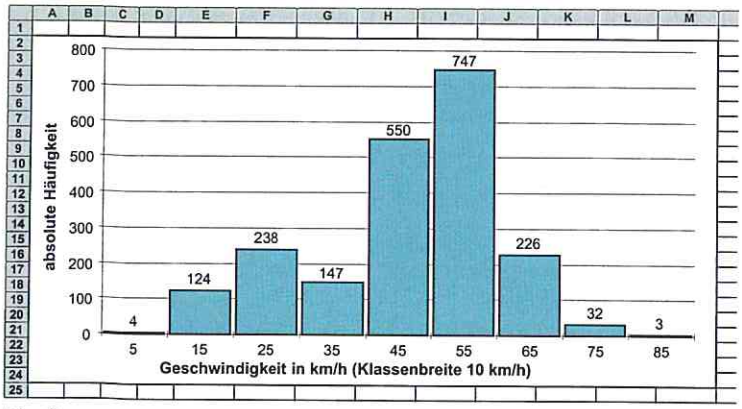


Fig. 2

7 Das Säulendiagramm (Fig. 1) zeigt, wie sich die 5518 von der Kölner Kripo im Jahr 1998 registrierten Wohnungseinbrüche auf die Tageszeiten verteilen.

a) Kommentieren Sie das Diagramm (auch im Vergleich zu Fig. 1, Seite 44). Welche Informationen entsprechen (widersprechen) Ihren Erwartungen?

b) Welche absoluten Häufigkeiten gehören vermutlich zu den einzelnen Säulen des Diagramms?

Fig

c) Wie könnte man sich erklären, dass die Summe der angegebenen relativen Häufigkeiten vor 100 % verschieden ist?

8 Die amtlichen Ergebnisse der Kommunalwahlen zum Rat der Stadt Köln lauten:

	wahl-berechtigt	gültige Stimmen	ungültige Stimmen	SPD	CDU	Grüne	FDP	PDS
1994	668 111	520 702	6862	221 520	176 408	84 392	18 462	(
1999	711 252	324 174	1447	98 295	146 694	51 073	13 197	694

Wenn ein Kursteilnehmer die Wahlergebnisse Ihrer Kommune recherchiert, bietet sich – auch im Hinblick auf Aufgabenteil d) – ein Vergleich mit den Kölner Verhältnissen an. Die entsprechenden Daten sind teilweise im Internet verfügbar.

a) Erläutern Sie an diesem Beispiel die Begriffe Grundgesamtheit, Merkmal, Merkmalsausprägung, Stichprobenwert.

b) Berechnen Sie die Wahlbeteiligung in Prozent.

c) Welche Parteien haben 1999 im Vergleich zu 1994 relativ (absolut) an Stimmen gewonnen? (Der relative Anteil einer Partei wird bezogen auf die Anzahl gültiger Stimmen.)

d) Stellen Sie die Wahlergebnisse durch ein Säulendiagramm (Kreisdiagramm) dar.

e) Zwei Wochen vor der Wahl trat der Oberbürgermeisterkandidat der SPD aufgrund gegen ihr erhobener Vorwürfe von seiner Kandidatur zurück.
Wie beurteilen Sie den Einfluss dieser Tatsache auf den Wahlausgang?

9 Die Teilnehmer eines Kurses wurden befragt, an wie viele der Begriffe „relative Häufigkeit", „Baumdiagramm", „Pfadregel" sie sich erinnern. Es ergab sich die folgende Urliste.

3 1 1 3 3 1 1 0 0 3 1 3 3 2 0 0 1 3 1 3 3 2 1 1 0 0 1 3 0

a) Bestimmen Sie die absoluten und die relativen Häufigkeiten der Merkmalsausprägungen.

b) Zeichnen Sie ein Säulendiagramm (Kreisdiagramm).

10 Marie (M) und Lena (L) spielen noch nicht sehr gut Federball.
Sie notieren, wie oft sie den Ball mit ihrem Schläger treffen, bevor er zu Boden fällt, und ein neuer Aufschlag fällig wird.
Die Angabe „0" bedeutet, dass der Ball beim Aufschlag nicht getroffen wird, bei „1" gelingt de Aufschlag, aber die Gegnerin kann den Ball nicht zurückschlagen.

3, 1, **0**, 6, **5**, 1, **10**, 2, **4**, 12, **1**, 9, **3**, 3, **0**, 1, **6**, 8, **2**, 18, **12**, 1, **0**, 6, **3**, 3, **2**, 8, **5**, 2, **2**, 8, **4**, 9, **7**, 0, **1**, 1, **8**, 9, **3**, 2, **7**, 12, **5**, 2, **5**, 6, **14**, 12, **14**, 16, **10**, 7, **9**, 6, **4**, 1, **0**, 4, **3**, 8, **2**

a) Bestimmen Sie die absoluten und die relativen Häufigkeiten der Merkmalsausprägungen.

b) Stellen Sie die Ergebnisse durch ein Säulendiagramm mit Klassenbreite 3 dar, wobei (0, 1, 2 zur ersten, (3, 4, 5) zur zweiten Klasse . . . gehören.

c) Bei den fett gedruckten Zahlen der Urliste hatte Marie Aufschlag. Wie oft wurde das Spiel durch einen Fehler von Marie, wie oft durch einen Fehler von Lena beendet?
(Tipp: **3** bedeutet: M L M (L) Fehler von Lena, 1 bedeutet L (M) Fehler von Marie.)

11 In einer Klasse 5 wurden die Merkmale Fehlstunden F und Zeugnisdurchschnittsnote N erhoben.

	F	N		F	N		F	N		F	N		F	N		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	Patrick	2	3,38	Marius	0	2,88	Melek	16	3,57	Nino	9	2,38

Wenn Sie mit einem Kalkulationsprogramm arbeiten wollen, laden Sie Noten-5.xls.

a) Werten Sie die Urliste nach dem Merkmal Fehlstunden aus. Erstellen Sie dazu eine Tabelle der absoluten (relativen) Häufigkeiten für die Klasseneinteilung $0 \leqq F \leqq 10$; $11 \leqq F \leqq 20$; . . .; $41 \leqq F \leqq 50$. Zeichnen Sie ein Säulendiagramm (Kreisdiagramm).
b) Werten Sie die Urliste nach dem Merkmal Zeugnisnote aus. Erstellen Sie eine Tabelle der absoluten (relativen) Häufigkeiten für die Klasseneinteilung $1 \leqq N \leqq 1{,}50$; $1{,}50 < N \leqq 2{,}50$; . . .; $3{,}50 < N \leqq 5{,}50$. Zeichnen Sie ein Säulendiagramm (Kreisdiagramm).

Hausnummern.xls enthält 1152 Hausnummern für alle Schüler einer Schule. Sie können auch diese Datei bearbeiten. Dort hat die Verteilung der relativen Häufigkeiten starke Ähnlichkeiten mit einer Exponentialfunktion. Ob das nur ein Zufall ist?

12 Eine Erhebung von Hausnummern X in den Adressen von 60 Schülern einer Jahrgangsstufe 11 lieferte die folgende Urliste.

9, 145, 16, 51, 39, 15, 7, 4, 3, 15, 68, 11, 1, 89, 20, 5, 1, 11, 9, 7, 1, 24, 22, 4, 2, 26, 7, 26, 205, 140, 96, 60, 55, 48, 43, 41, 26, 25, 20, 9, 6, 5, 3, 2, 1, 41, 27, 20, 20, 16, 3, 6, 297, 26, 8, 184, 5, 11, 15, 5

a) Erstellen Sie eine Tabelle der relativen Häufigkeiten für die Klasseneinteilung $0 < X \leqq 10$; $10 < X \leqq 20$; . . . und zeichnen Sie ein Säulendiagramm.
b) Begründen Sie, warum kleine Hausnummern in der Regel häufiger auftreten als große.
c) Werten Sie entsprechend die Seite des Telefonbuchs aus, auf der Ihre Nummer (die Nummer „Ihres Freundes" oder „Ihrer Freundin") steht.
d) Fassen Sie die Ergebnisse aus c) zusammen und prüfen Sie die Vermutung aus Aufgabenteil b).

Experiment

13 a) Legen Sie in der Gemüseabteilung (nach *höflicher* Rücksprache mit dem Filialleiter) viele Beutel Kartoffeln auf die elektronische Waage. Notieren Sie, um wie viel die tatsächlichen Gewichte von den auf der Packung angegebenen abweichen.
b) Werten Sie die Urliste aus und zeichnen Sie ein Säulendiagramm; verwenden Sie dabei eine Klasseneinteilung der Breite 10 Gramm.

14 An einem Gymnasium wurde für jeden Schüler ermittelt, an welchem Wochentag er geboren wurde. Das Säulendiagramm zeigt, wie sich das Merkmal auf die einzelnen Wochentage verteilt.
a) Nennen Sie plausible Gründe dafür, dass am Wochenende weniger Kinder geboren werden als am Wochenanfang.
b) Wenn an Ihrer Schule die Schülerdaten elektronisch gespeichert werden: Lassen Sie sich die Geburtstage als Tabelle ausgeben und untersuchen Sie mit einem Kalkulationsprogramm, ob auch an Ihrer Schule „Samstags-" und „Sonntagskinder" seltener sind als „Montags-" oder „Dienstagskinder".

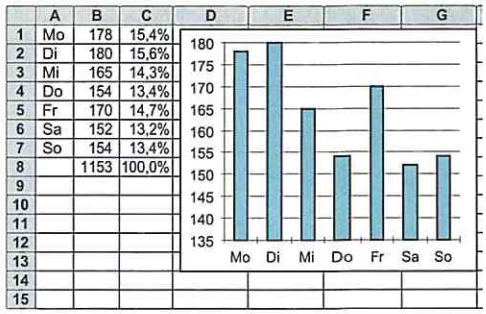


Fig. 1

Experiment

Excel kennt den Befehl Wochentag.

2 Mittelwert – Erwartungswert

1 Im Eiscafe Bez gibt es kleine Kugeln zu 0,60 Euro und große zu 1,00 Euro. Herr Bez erläutert: „Die kleinen Kugeln sind halb so groß, da ich aber mehr Arbeit habe, mache ich sie 10 Cents teurer."

a) Sind die kleinen Kugeln wirklich halb so „groß" wie die großen?

b) Ursula kauft für 3 Euro drei große, Dorothea fünf kleine Kugeln. Wer bekommt vermutlich mehr Eis?

kleine Kugeln:
34 g, 36 g, 35 g, 32 g, 34 g, 33 g, 37 g,
40 g, 40 g, 38 g, 33 g, 36 g, 34 g, 35 g,
33 g, 33 g, 36 g, 37 g, 34 g, 35 g

große Kugeln:
50 g, 56 g, 54 g, 58 g, 50 g, 54 g, 49 g,
58 g, 54 g, 60 g, 62 g, 64 g, 57 g, 57 g,
55 g, 62 g, 59 g, 59 g, 54 g, 58 g

Z. B.:
Munch – Kopf (f)
van Gogh – Zahl (r)
Sisley – Zahl (f)
ein Treffer

Reale Datenerhebung – Mittelwert

In einem „Allgemeinbildungstest" sollen drei Gemälde den Stilrichtungen Impressionismus bzw. Expressionismus zugeordnet werden. Als Merkmal X wird die Anzahl der Treffer erhoben.

Urliste:

Andy:	$x_1 = 3$ Treffer	Memo:	$x_6 = 3$ Treffer
Oliver:	$x_2 = 2$ Treffer	Dietmar:	$x_7 = 1$ Treffer
Pjotre:	$x_3 = 2$ Treffer	Jens:	$x_8 = 0$ Treffer
Mirjam:	$x_4 = 3$ Treffer	Robert:	$x_9 = 1$ Treffer
Michael:	$x_5 = 2$ Treffer	Melanie:	$x_{10} = 2$ Treffer

Häufigkeitstabelle:

Trefferzahl	abs. H.	rel. H.
0	1	10 %
1	2	20 %
2	4	40 %
3	3	30 %

Die in der Urliste enthaltene Information über die Sachkenntnis der gesamten Testgruppe lässt sich durch den Mittelwert $\bar{x}$ der Anzahl der Treffer charakterisieren:

$\bar{x} = \frac{1}{10}(3 + 2 + 2 + 3 + 2 + 3 + 1 + 0 + 1 + 2) = \frac{19}{10} = 1{,}9$ (Treffer).

Wenn man die Summanden in der Klammer der Größe nach anordnet und gleiche Summanden zusammenfasst wird deutlich, dass man den Mittelwert auch unter Benutzung der relativen Häufigkeiten berechnen kann. Man multipliziert die Trefferzahlen mit den relativen Häufigkeiten und addiert die Produkte.

$\bar{x} = \frac{1}{10}(0 + 1 + 1 + 2 + 2 + 2 + 2 + 3 + 3 + 3) = 0 \cdot \frac{1}{10} + 1 \cdot \frac{2}{10} + 2 \cdot \frac{4}{10} + 3 \cdot \frac{3}{10} = \frac{19}{10} = 1{,}9$ (Treffer

Wenn die Häufigkeitsverteilung mithilfe von Merkmalsklassen erstellt wurde, dann berechnet man den Mittelwert näherungsweise unter Benutzung der Klassenmitten.

Simulation

Man kann reale statistische Erhebungen wie den Allgemeinbildungstest auch mit Münzen, Karten oder Zufallszahlen nachspielen (**„simulieren"**).

Die Häufigkeitstabelle zeigt die Ergebnisse zweier Testgruppen, die ihre Antworten ohne Sachkenntnis durch das Werfen von Münzen gaben („Kopf" = Impressionismus, „Zahl" = Expressionismus).

Anzahl der „Treffer"	rel. H. bei Testgruppe I	rel. H. bei Testgruppe II
0	10 %	20 %
1	40 %	30 %
2	30 %	50 %
3	20 %	0 %
Mittel:	1,6	1,3

Diese simulierten Testgruppen erreichten mit $\bar{x}_I = 1{,}6$ bzw. $\bar{x}_{II} = 1{,}3$ im Mittel deutlich weniger Treffer als die reale Testgruppe. Das spricht für eine (wohl nicht sehr ausgeprägte) „Kompetenz" der realen Testgruppe.

Wahrscheinlichkeit – Erwartungswert

Mitunter kann man schon vor einer realen oder simulierten statistischen Untersuchung durch eine **Wahrscheinlichkeitsverteilung** beschreiben, welche relativen Häufigkeiten man für die einzelnen Merkmalsausprägungen erwartet.

Zum Beispiel erwartet man bei einer Testperson, die ihre Antworten durch Werfen einer Münze findet, „drei Treffer" mit der Wahrscheinlichkeit $p_3 = \frac{1}{8}$ (vgl. Fig. 1). Bei der Bestimmung von Wahrscheinlichkeiten können ein Baumdiagramm oder eine Tabelle hilfreich sein.

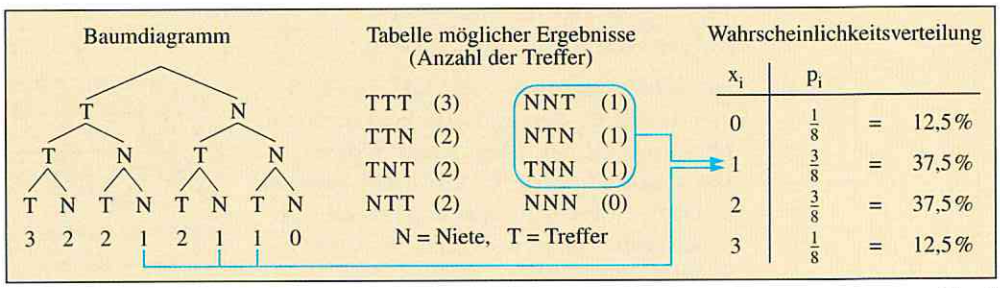


Fig. 1

*Im 17. und 18. Jh. untersuchte man in der Stochastik zunächst Glücksspiele, bei denen „Gewinne" x_i mit den Wahrscheinlichkeiten p_i auftraten. Der Erwartungswert μ misst dann, welchen Gewinn man im Mittel auf lange Sicht **erwarten** kann. Im Allgemeinen wird μ keiner der Werte x_i sein.*

Den analog zum **Mittelwert** einer **Häufigkeitsverteilung** aufgebauten Ausdruck

$$\mu = 0 \cdot \tfrac{1}{8} + 1 \cdot \tfrac{3}{8} + 2 \cdot \tfrac{3}{8} + 3 \cdot \tfrac{1}{8} = 1{,}5$$

bezeichnet man als **Erwartungswert** der **Wahrscheinlichkeitsverteilung**.

Da bei großem Stichprobenumfang die relativen Häufigkeiten h_i in der Nähe der Wahrscheinlichkeiten p_i liegen, wird auch der Mittelwert in der Nähe des Erwartungswertes liegen, es gilt also $\bar{x} \approx \mu$. Mit dem Mittelwert $\bar{x}$ kann man den oft unbekannten Erwartungswert μ schätzen.

Sind $x_1, x_2, x_3, \ldots, x_n$ Stichprobenwerte eines Merkmals X, so nennt man
$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$ den **Mittelwert** von $x_1, x_2, x_3, \ldots, x_n$.
Kommen die Merkmalsausprägungen $x_1, x_2, x_3, \ldots, x_r$ mit den relativen Häufigkeiten $h_1, h_2, h_3, \ldots, h_r$ vor, so kann man einfacher rechnen:
$\bar{x} = x_1 \cdot h_1 + x_2 \cdot h_2 + x_3 \cdot h_3 + \ldots + x_r \cdot h_r$
(Mittelwert einer Häufigkeitsverteilung).
Kann man auch die zugehörigen Wahrscheinlichkeiten $p_1, p_2, p_3, \ldots, p_r$ angeben, so nennt man $\mu = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + \ldots + x_r \cdot p_r$
den **Erwartungswert der Wahrscheinlichkeitsverteilung**.

Tipp:
BAHL (f)
IBBU (r)
XOOX (f)
1 Treffer

Beispiel: (Mittelwert, Erwartungswert)

In drei Bechern befinden sich drei verschiedene Sorten Chips. Die Becher sollen nach einer Geschmacksprobe den Sorten zugeordnet werden. Als Merkmal gilt die Anzahl der richtigen Zuordnungen (Treffer).

Bei einer Stichprobe mit 40 Personen ergab sich folgende Häufigkeitsverteilung.

Anzahl der Treffer	0	1	2	3
absolute Häufigkeit	9	13	0	18

a) Berechnen Sie den Mittelwert (kurz: die „mittlere Trefferzahl").

b) Simulieren Sie (mithilfe von Spielkarten) Versuchsgruppen aus 40 Personen, die ihre Zuordnungen zufällig vornehmen. Ermitteln Sie die mittlere Trefferzahl.

c) Bestimmen Sie die zu b) gehörende Wahrscheinlichkeitsverteilung und den Erwartungswert.

Lösung:

a) Mittelwert der Anzahl der Treffer: $\bar{x} = 0 \cdot \frac{9}{40} + 1 \cdot \frac{13}{40} + 2 \cdot \frac{0}{40} + 3 \cdot \frac{18}{40} = 1{,}675$ (Treffer).

Versuchsleiter: ♣, ♥, ♦
Teilnehmer: ♦, ♥, ♣
Also: 1 Treffer

b) Ein mögliches Vorgehen: Der Versuchsleiter und der Versuchsteilnehmer mischen jeweils d Karten mit den Farben ♣, ♦, ♥ und legen die Karten verdeckt untereinander. Dann werden di Karten aufgedeckt und die Übereinstimmungen gezählt.
Die Tabelle zeigt die Häufigkeitsverteilung zweier Versuchsreihen.

Anzahl der Treffer	0	1	2	3
Absolute Häufigkeit bei Versuchsreihe I	10	24	0	6
Absolute Häufigkeit bei Versuchsreihe II	14	21	0	5

Als Mittelwerte ergeben sich hier $\bar{x}_I = 1{,}05$ und $\bar{x}_{II} = 0{,}90$. Sie liegen deutlich unter dem Mittelwert von 1,675 der „realen" Gruppe. Das spricht für deren „Geschmackskompetenz".

c) Durch Betrachtung aller möglichen Ergebnisse findet man die in der Tabelle angegebene Wahrscheinlichkeitsverteilung. Daraus ergibt sich als Erwartungswert

$\mu = 0 \cdot \frac{1}{6} + 1 \cdot \frac{3}{6} + 2 \cdot \frac{0}{6} + 3 \cdot \frac{1}{6} = 1$ (Treffer).

Die Mittelwerte $\bar{x}$ liegen bei den Simulationen in der Nähe des Erwartungswertes μ.

Versuchsleiter: ♣ ♦ ♥	Wahrscheinlichkeitsverteilung	
Mögl. Ergebnisse des Teilnehmers:	Anzahl der Treffer	Wahrsch
♣ ♦ ♥ 3 Treffer	0	$\frac{2}{6}$
♣ ♥ ♦ 1 Treffer		
♦ ♣ ♥ 1 Treffer	1	$\frac{3}{6}$
♦ ♣ ♥ 0 Treffer	2	$\frac{0}{6}$
♥ ♣ ♦ 0 Treffer		
♥ ♦ ♣ 1 Treffer	3	$\frac{1}{6}$

Aufgaben

2 Paul erhielt im Fach Erdkunde die Einzelnoten 2; 3; 1,5 und 2. Berechnen Sie den Mittelwert. Welche Gesamtnote erhält er vermutlich?

3 Petra hat in den ersten drei Klassenarbeiten im Fach Deutsch die Noten 4; 1; 3 geschrieben. Sie schreibt noch eine Klassenarbeit. Welche Note muss sie mindestens schreiben, damit der Mittelwert der Klassenarbeitsnoten besser als 2,5 wird?

4 In einer Streichholzschachtel sollen sich gemäß Packungsaufdruck 38 Hölzchen befinden Sabrina untersucht eine Stichprobe und erhält folgende Tabelle.

Zahl der Hölzer	35	37	38	39	40	41	42	44
absolute Häufigkeit	1	4	5	5	6	5	3	1

Wie viele Hölzer befanden sich durchschnittlich in einer Schachtel?

5 In der Jahrgangsstufe 11 ergab sich am Schuljahresanfang die „Altersverteilung" aus Fig. 1.

a) Berechnen Sie näherungsweise das mittlere Alter $\bar{x}$ unter Benutzung der Klassenmitten $k_1 = 15{,}3, \ldots, k_9 = 20{,}1$.

b) Vergleichen Sie Ihr Ergebnis mit dem aus der Urliste berechneten Mittelwert $\bar{x} = 18{,}9789$ Jahre.

c) Begründen Sie: Der Unterschied ist stets höchstens so groß wie die Klassenbreite (hier 0,6 Jahre).

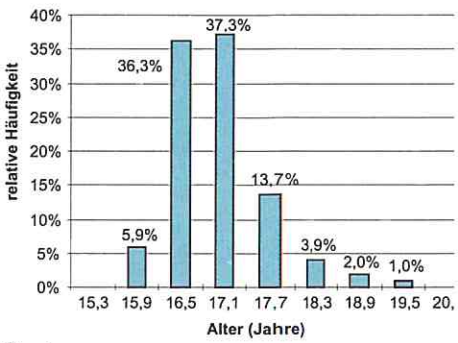


Fig. 1

6 Michael und Mario haben im Berufsverkehr die Anzahl der Personen in PKWs gezählt.
Um wie viel Prozent würde der PKW-Verkehr abnehmen, wenn alle Autos 4 Personen befördern würden?

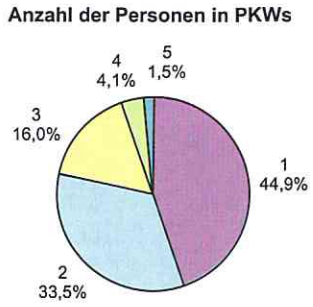


Fig. 1

7 Frau Kolmetz fährt mit ihrem PKW die ersten 100 km mit der Geschwindigkeit 80 $\frac{km}{h}$, die zweiten 100 km mit 120 $\frac{km}{h}$. Frau Kaiser fährt die erste Stunde mit genau 80 $\frac{km}{h}$, eine Stunde mit 120 $\frac{km}{h}$. Mit welcher konstanten Geschwindigkeit hätten die beiden Damen fahren müssen, ohne ihre jeweiligen Gesamtfahrzeiten zu ändern?

8 Wolfgang führt beim Tanken Protokoll.
a) Bestimmen Sie die fehlenden Werte in der letzten Spalte der Tabelle. Berechnen Sie den mittleren Verbrauch, indem Sie den Mittelwert der 5 Werte in der letzten Spalte ermitteln.
b) Frank meint, man solle den mittleren Verbrauch auf 100 km in einem Schritt aus Kilometerstand am Anfang und am Ende berechnen. Welchen Wert erhält Frank?
c) Nehmen Sie zu dem Ergebnis Stellung.

km-Stand	Liter	Liter/100 km
38996		
39461	34,01	7,31
39908	50,96	11,40
40451	42,91	
40881	45,40	
41305	36,73	

9 Marcellina: „Der Mittelwert aller Abweichungen vom Mittelwert ist immer Null."
a) Was meint Marcellina mit ihrer „Entdeckung"?
Erläutern Sie die Aussage am Beispiel der Urliste 1, 5, 0, 2, 1, 8, 0, 3.
b) Begründen Sie, dass Marcellinas Aussage für jede Urliste stimmt.

10 a) Würfeln Sie 10-mal. Bilden Sie den Mittelwert $\bar{x}$ aus den erhaltenen Augenzahlen.
b) Vergleichen Sie $\bar{x}$ mit dem Erwartungswert der Augenzahl beim Werfen eines Würfels.

11 a) Werfen Sie 10-mal zwei (vier, fünf) Münzen. Als Merkmal X gelte die Anzahl der gefallenen „Köpfe". Bilden Sie den Mittelwert.
b) Zeigen Sie: X hat bei zwei Münzen die folgende Wahrscheinlichkeitsverteilung.

Ergebnis x_i	0	1	2
Wahrscheinlichkeit p_i	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

c) Berechnen Sie den Erwartungswert und vergleichen Sie ihn mit dem Mittelwert aus a).
d) Ermitteln Sie die Wahrscheinlichkeitsverteilung, die zum Wurf von vier (fünf) Münzen gehört, und berechnen Sie den Erwartungswert.

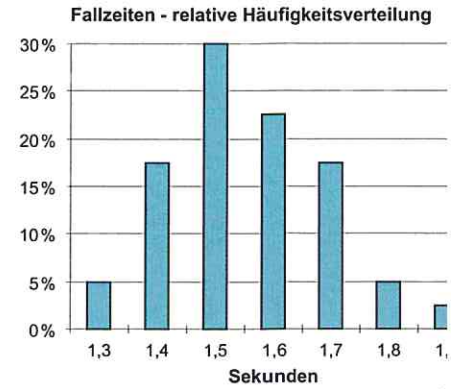
Gemeinsames Experiment

12 Wählen Sie in Ihrem Kurs einen Versuchsleiter, der mit Ihnen einen selbst ausgedachten Test nach dem Muster „Stilrichtung von Gemälden erkennen" oder einen Geschmackstest (wie im Beispiel) durchführt.
Beurteilen Sie anschließend die Kompetenz Ihrer Gruppe, indem Sie für eine Simulation Ihres Tests die Wahrscheinlichkeitsverteilung bestimmen, den Erwartungswert berechnen und ihn mit der mittleren Trefferzahl der Gruppe vergleichen.

3 Varianz, Standardabweichung

1 Die Höhe h eines Turms kann man näherungsweise aus der Fallzeit t eines Steinchens na der Formel $h = 5t^2$ bestimmen. Wenn der Stein 1 s braucht, beträgt die Höhe 5 m, bei 2 s Fall zeit 20 m. Neun Schüler versuchten gemeinsam, mithilfe von Stoppuhren, die Höhe ihres Kla senraumes über dem Schulhof zu ermitteln. Jeder stoppte mit seiner Uhr die Fallzeiten von fü Kieselsteinen. (Bei leeren Feldern versagten die Uhren.)

	1	2	3	4	5
Cristian	1,3	1,5	1,5	1,6	
Hamed	1,8	1,4	1,7	1,6	1,8
Nicole	1,6	1,4	1,7	1,6	1,9
Sabine	1,7	1,5		1,6	1,7
Nuri	1,7	1,5	1,6	1,6	1,7
Stefan	1,6	1,5	1,6		
Catcher	1,7	1,5	1,4	1,4	1,4
Daniela	1,5	1,5		1,4	1,5
Bianca	1,5	1,5	1,5	1,3	1,4



Fi

a) Wie hoch lag der Klassenraum über dem Schulhof?
b) Können Sie eine grobe Aussage über die Genauigkeit der Höhenangabe aus a) machen?

Jan	Frank
x_i	y_i
0,6	0,68
0,5	0,63
0,4	0,64
0,5	0,54
0,7	0,66
0,8	0,54
0,6	0,59
0,7	0,54
0,6	0,58
0,7	0,60
0,9	0,60
0,5	0,61
0,9	0,58
0,5	0,66
0,7	0,62
0,7	0,58
0,3	0,63
0,8	0,58
0,7	0,57
0,2	0,49
$\bar{x} \approx 0{,}62$	$\bar{y} \approx 0{,}60$
$s_X \approx 0{,}18$	$s_Y \approx 0{,}05$

Fig. 2

Jan wirft 10 Reißzwecken; 6 davon liegen mit der Spitze nach oben (relative Häufigkeit $x_1 = 0{,}6$; vgl. Fig. 2). Er führt das Experiment insgesamt 20-mal durch, der zweite Wurf liefer $x_2 = 0{,}5$. Frank wirft stets 100 Reißzwecken. Er erhält $y_1 = 0{,}68$, $y_2 = 0{,}63$, . . .
Fig. 2 vermittelt den Eindruck, dass (bei annähernd gleichem Mittelwert) die Ergebnisse von stärker „streuen" als diejenigen von Frank. Man kann die unterschiedliche „Streuung" auf ver schiedene Weisen „messen".

Als **Spannweite** einer Urliste bezeichnet man die Differenz aus dem größten und dem kleinst Wert. Bei Jan ist sie mit $0{,}9 - 0{,}2 = 0{,}7$ viel größer als bei Frank: $0{,}68 - 0{,}49 = 0{,}19$.
Auch die **mittlere absolute Abweichung** von $\bar{x}$ ist als Streumaß geeignet.
Bei Jan ergibt sich mit

$$\frac{1}{20}(|0{,}6 - 0{,}62| + |0{,}5 - 0{,}62| + |0{,}4 - 0{,}62| + \ldots + |0{,}2 - 0{,}62|) = 0{,}145$$

wiederum ein größerer Wert als bei Frank:

$$\frac{1}{20}(|0{,}68 - 0{,}60| + |0{,}63 - 0{,}60| + |0{,}64 - 0{,}60| + \ldots + |0{,}49 - 0{,}60|) = 0{,}0370.$$

(Würde man auf die Betragsstriche verzichten und nur den Mittelwert aus den Abweichungen berechnen, so ergäbe sich stets der Wert 0. Man erhielte kein brauchbares „Streuungsmaß".)

In der Praxis verwendet man statt der Beträge meist die Quadrate der einzelnen Differenzen. Das zugehörige Streuungsmaß V_X nennt man die **mittlere quadratische Abweichung** oder **Varianz**. Die Wurzel aus der Varianz heißt **Standardabweichung**. Sie wird mit s_X bezeichnet So erhält man für Jan

$$V_X = \frac{1}{20}((0{,}6 - 0{,}62)^2 + (0{,}5 - 0{,}62)^2 + \ldots + (0{,}2 - 0{,}62)^2) \approx 0{,}0323, \quad s_X \approx \sqrt{V_X} = 0{,}1797.$$

Für Frank ergibt sich $V_Y \approx 0{,}0021$ und $s_Y \approx 0{,}0462$.

Sind $x_1, x_2, \ldots, x_n$ Stichprobenwerte eines Merkmals X mit dem Mittelwert $\bar{x}$, so nennt man

$$V_X = \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

die **Varianz** von X.

Die Quadratwurzel aus der Varianz V_X wird **Standardabweichung** s_X genannt:

$$s_X = \sqrt{V_X} = \sqrt{\frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)}.$$

Kommen die Merkmalsausprägungen $x_1, x_2, \ldots, x_r$ mit den relativen Häufigkeiten $h_1, h_2, \ldots, h_r$ vor, so kann man einfacher rechnen:

$$s_X = \sqrt{V_X} = \sqrt{(x_1 - \bar{x})^2 \cdot h_1 + (x_2 - \bar{x})^2 \cdot h_2 + \ldots + (x_r - \bar{x})^2 \cdot h_r}.$$

Kann man auch die zugehörigen Wahrscheinlichkeiten $p_1, p_2, p_3, \ldots, p_r$ angeben, dann nennt man $\sigma_X = \sqrt{(x_1 - \mu)^2 \cdot p_1 + (x_2 - \mu)^2 \cdot p_2 + \ldots + (x_r - \mu)^2 \cdot p_r}$ die theoretische Standardabweichung.

Taschenrechner und Tabellenkalkulation teilen zur Berechnung der empirischen Varianz mitunter nicht durch n, sondern durch n – 1. Der Hintergrund dafür wird in Aufgabe 8 angesprochen.

Zur Abgrenzung von der theoretischen Standardabweichung σ_X nennt man s_X auch die **empirische** Standardabweichung. Da bei *großem* Stichprobenumfang die relativen Häufigkeiten h_i in der Nähe der Wahrscheinlichkeiten p_i liegen und $\bar{x} \approx \mu_X$ gilt, gilt auch $s_X \approx \sigma_X$. Mittels der empirischen Standardabweichung s_X kann man die oft unbekannte theoretische Standardabweichung σ_X **schätzen.**

Der Name Standardabweichung kommt daher, dass bei vielen Erhebungen das Merkmal X, wie z. B. die Körperlänge von Schulanfängern oder der Intelligenzquotient, einer Vielzahl von unabhängigen Einflüssen unterliegt. Nach einer groben Faustregel weichen „standardmäßig" ca. 68 % aller Daten um höchstens eine Standardabweichung vom Mittelwert ab.
Ca. 68 % aller Daten liegen also im Intervall $[\bar{x} - s_X; \bar{x} + s_X]$.

Beispiel 1: (Berechnung einer empirischen Standardabweichung)

a) (Urliste gegeben)
Frau Tribler wiegt sich jeden Morgen auf ihrer Präzisionswaage. Ihre Ergebnisse sind in Fig. 1, Spalten A bis C, enthalten. Berechnen Sie die Standardabweichung.
Lösung:

	A	B	C	D
1		x_i (kg)	$(x_i-\bar{x})^2$	
2	Mo	67,98	0,004	
3	Di	67,98	0,0279	
4	Mi	67,98	0,0825	
5	Do	67,98	0,0334	
6	Fr	67,98	0,1874	
7	Sa	67,98	0,0033	
8	So	67,98	0,0279	
9		$\bar{x}$=67,98	V_x=0,05	=Mittelwert(C2:C8)
10			s_x=0,2288	=Wurzel(C9)
11				

Fig. 1

Man erhält $s_X = 0{,}229$ (kg).

b) (Häufigkeitsverteilung gegeben)
In Fig. 2, Spalten A und B, ist die Häufigkeitsverteilung der Fallzeiten von Aufgabe 1 angegeben.
Berechnen Sie die Standardabweichung.
Lösung:

	A	B	C	D
1	x_i (sec)	h_i	x_i*h_i	$(x_i-\bar{x})^2*h_i$
2	1,3	5,0 %	0,0650	0,00325
3	1,4	17,5 %	0,2450	0,00420
4	1,5	30,0 %	0,4500	0,00091
5	1,6	22,5 %	0,3600	0,00046
6	1,7	17,5 %	0,2975	0,00368
7	1,8	5,0 %	0,0900	0,00300
8	1,9	2,5 %	0,0475	0,00298
9		100,0 %	$\bar{x}$=1,5550	V_x=0,01848
10				s_x=0,14
11				

Fig. 2

Man erhält $s_X = 0{,}14$ (s).

Hinweise zu EXCEL
Zu a):
Ohne Zwischenrechnung erhält man Varianz und Standardabweichung über die Befehle =VARIANZEN(B2:B8) bzw. =STABWN(B2:B8). Dagegen berechnen die Befehle =VARIANZ bzw. =STABW die empirische Standardabweichung mit dem Nenner n – 1 statt n.
Zu b):
*Man kann die Zwischenergebnisse im Bereich D2 bis D8 zellenweise oder mit der Vektorformel {=(A2:A8–C9)^2*B2:B8} berechnen.*

Beispiel 2: (Theoretische Standardabweichung)
Ein normaler Würfel wird geworfen und die Augenzahl X beobachtet.
a) Berechnen Sie die theoretische Standardabweichung σ_X.
b) Vergleichen Sie σ_X mit den empirischen Standardabweichungen s_X aus den Urlisten
5 2 6 3 6 1 5 3 1 4 6 3 3 1 2 2 4 2 2 4 bzw. 2 1 6 4 1 1 3 3 3 3 5 5 6 4 4 5 2 1 1
Lösung:
a) Die Augenzahlen treten alle mit der Wahrscheinlichkeit $\frac{1}{6}$ auf. Es gilt
$\mu_X = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3{,}5;$
$\sigma_X = \sqrt{\left(1-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(2-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(3-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(4-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(5-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(6-\frac{7}{2}\right)^2\cdot\frac{1}{6}} = \sqrt{\frac{35}{12}} \approx 1{,}708.$
b) Für die Urlisten erhält man $\bar{x} \approx 3{,}25$, $s_X \approx 1{,}64$ bzw. $\bar{x} \approx 3{,}30$, $s_X \approx 1{,}76$. Die empirische Standardabweichungen liegen in der Nähe der theoretischen Standardabweichung.

Aufgaben

2 Berechnen Sie die mittlere absolute Abweichung, die Varianz und die Standardabweichu der Urliste.
a) 5, 0, 8, 6, 5, 9 b) 3, 3, 3, 3, 3 c) 1, 3, 1, 3, 1, 3

3 Für die Teile einer Klausuraufgabe erhielten 20 Schüler die folgenden Punktzahlen.

Schüler	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
Teil a:	8	6	8	7	7	7	7	8	3	8	6	5	8	5	5	6	8	8	8	
Teil b:	5	2	5	8	4	5	1	7	0	3	8	2	7	2	2	4	3	5	7	
Teil c:	6	7	4	4	5	3	4	7	0	5	5	6	2	3	2	4	3	4	10	

a) Berechnen Sie zeilenweise die Spannweite, die mittlere absolute Abweichung, die Varianz und die Standardabweichung der in den drei Aufgabenteilen erreichten Punktzahlen. Bei welchem Aufgabenteil ist die Varianz besonders hoch? (Man nennt solche Aufgaben „trennscharf".)
b) Berechnen Sie spaltenweise die von jedem der 20 Schüler erreichte Punktsumme. Wie groß sind Varianz und Standardabweichung der Punktsumme?

4 Mit einer elektronischen Waage wurde wiederholt das Gewicht eines Briefes (in g) gemessen. Man erhielt die folgende Urliste.

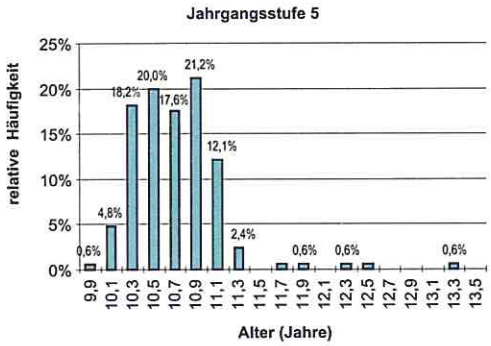
22,94 22,90 22,92 22,76 22,80 22,85 22,84 22,86 22,83 22,87
22,88 22,86 22,88 22,85 22,96 22,93 22,95 22,61 22,70 22,69

a) Berechnen Sie Mittelwert und Standardabweichung.
b) Die Angabe einer Messgenauigkeit (0,1 %) bedeutet: Die Standardabweichung der Messwerte beträgt höchstens 0,1 % des „wahren Wertes". Darf die Waage das „Gütesiegel 0,1 %" tragen

5 a) Berechnen Sie die Standardabweichungen der gemessenen Fallzeiten für jeden der 5 Fallversuche aus Fig. 1 von Aufgabe 1 (je Spalte eine Rechnung). Bei welchem Fallversuch „streuen" die Messwerte besonders wenig?
b) Berechnen Sie die Mittelwerte der Fallzeiten für jeden der 9 Schüler aus Fig. 1 von Aufgabe 1 (je Zeile eine Rechnung). Berechnen Sie anschließend die Standardabweichung dieser Mittelwerte.
c) „Mittelwerte streuen i. Allg. weniger als Einzelmessungen." Belegen Sie diese Aussage mithilfe der Ergebnisse aus den Aufgabenteilen a) und b).

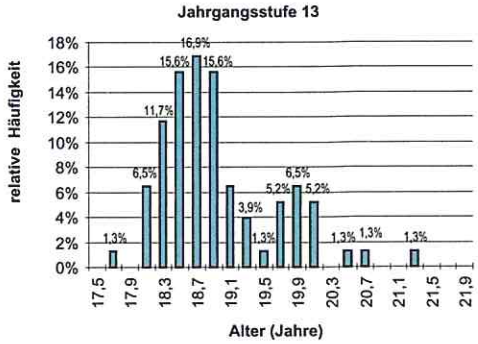
6 a) Berechnen Sie aus den Angaben von Fig. 1 und von Fig. 2 die Mittelwerte und die Standardabweichungen der Altersverteilungen in der Jahrgangsstufe 5 und in der Jahrgangsstufe 13.
b) Wie erklären Sie inhaltlich, dass die Standardabweichung in der Jahrgangsstufe 13 größer ist als in der Jahrgangsstufe 5?

In dieser Stichprobe liegen von den 165 Schülern der 5. Klassen 78,8 % mit ihrem Alter zwischen $\bar{x} - s_X$ und $\bar{x} + s_X$. In Stufe 13 (77 Schüler) liegt der Prozentsatz mit 70,1 % näher am 68 %-Wert der Faustregel. Sie können diese Angaben durch Auswerten der Urliste Alter.xls kontrollieren.



Altersverteilung Jahrgangsstufe 5

Fig. 1



Altersverteilung Jahrgangsstufe 13

Fig. 2

Experiment

7 a) Zählen Sie bei 5 (bei 10) Telefongesprächen, wie oft Sie anklingeln müssen, bis der Gesprächspartner abhebt.
Anrufbeantworter und erfolglose Verbindungsversuche bleiben unberücksichtigt.
Ermitteln Sie Mittelwert und Standardabweichung Ihrer Urliste.
b) Fassen Sie gemeinsam die Urlisten aller Kursteilnehmer zusammen und berechnen Sie Mittelwert und Standardabweichung der gemeinsamen Urliste.
Untersuchen Sie, ob die 68 %-Regel ein brauchbares Ergebnis liefert.

Zur Theorie:
Die Abweichungen der Stichprobenwerte vom Mittelwert $\bar{x}$ sind im Mittel kleiner als die Abweichungen vom Erwartungswert μ. Das liegt daran, dass sich der Mittelwert jeder einzelnen Stichprobe besonders gut anpasst. Bei der Urliste „1, 1, 1" sind z. B. alle Abweichungen vom Mittelwert 0, alle Abweichungen vom Erwartungswert 0,5. Daher liefert Tims Formel im Mittel zu kleine Schätzwerte für die theoretische Varianz. Wenn man durch n – 1 teilt statt durch n, wird dieser „Schätzfehler" kompensiert.
Das rechtfertigt, dass viele Rechner zwei Formeln für die Varianz anbieten.

8 Varianz-Formel „mit Nenner n – 1"

a) Begründen Sie durch eine Rechnung:
Die Wahrscheinlichkeitsverteilung in Fig. 3, die den Wurf einer Münze beschreibt, hat den Erwartungswert 0,5 und die Varianz 0,25.

Ergebnis x_i	0	1
Wahrscheinlichkeit p_i	0,5	0,5

Fig. 3

b) Tim, Tom und Gabi versuchen, aus kleinen Stichproben (n = 3) die theoretische Varianz zu schätzen.
Sie rechnen so:
Tim teilt durch n = 3:
$V_X = \frac{1}{3}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2)$;
Tom teilt durch n – 1 = 2:
$V_X = \frac{1}{2}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2)$;
Gabi benutzt μ_X statt $\bar{x}$:
$V_X = \frac{1}{3}((x_1 - \mu_X)^2 + (x_2 - \mu_X)^2 + (x_3 - \mu_X)^2)$.
Welche Schätzwerte liefern die drei Formeln für die Urlisten „0, 1, 0" und „0, 1, 1"?

	A	B	C	D	E	F	G
1	x_1	x_2	x_3	$\bar{x}$	Tim	Tom	Gabi
2	0	0	0	0	0,000	0,000	0,25
3	0	0	1	0,333	0,222	0,333	0,25
4	0	1	0	0,333	0,222	0,333	0,25
5	0	1	1	0,667	0,222	0,333	0,25
6	1	0	0	0,333	0,222	0,333	0,25
7	1	0	1	0,667	0,222	0,333	0,25
8	1	1	0	0,667	0,222	0,333	0,25
9	1	1	1	1	0,000	0,000	0,25
10	Mittelwert			0,5	0,167	0,250	0,25
11							

Fig. 4

c) Karl hat alle Rechnungen für die 8 möglichen Urlisten der Länge 3 in Fig. 4 durchgespielt. Er behauptet: Toms und Gabis Formeln schätzen im Durchschnitt richtig, Tims Formel liefert zu kleine Werte.
Überprüfen Sie Karls Behauptung.
d) Untersuchen Sie entsprechend die drei Formeln für den Stichprobenumfang 4.

4 Punktdiagramme, Kovarianz

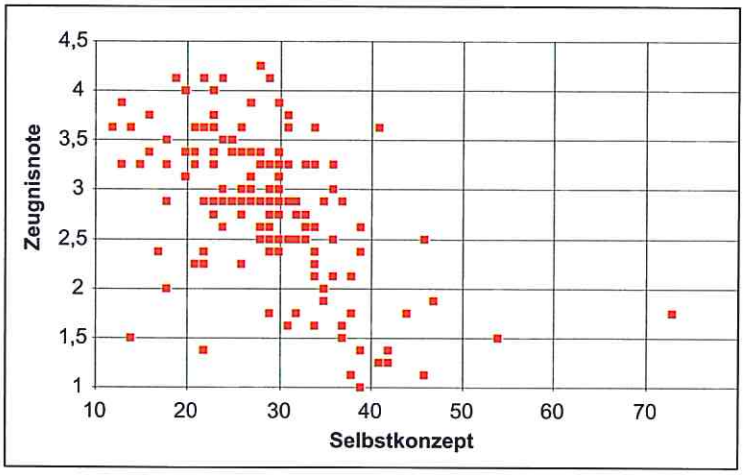


Fig. 1

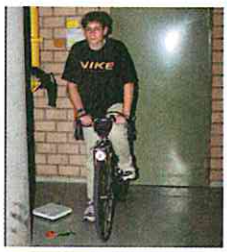
Den kompletten Fragebogen und die Testergebnisse finden Sie in Selbstkonzept.doc bzw. Selbstkonzept.xls.

1 In der Psychologie „misst" man Persö lichkeitsmerkmale wie z. B. das Selbstkonz (X) über Fragen der folgenden Form.

– „Ich habe ein gutes Gefühl, was meine Arbeit angeht."	nein ja 1 – 2 – 3 – 4 – 5 – 6
– „Ich kann Sachen selbst rauskriegen."	nein ja 1 – 2 – 3 – 4 – 5 – 6

Die Summe der angekreuzten Punktzahlen eines ganzen Fragebogens gilt als Messwer für das Merkmal X. Bei einer Gruppe von Schülern wurde deren Selbstkonzept X unte sucht. In Fig. 1 ist für jeden dieser Schüler Punkt gezeichnet. Dabei entspricht die x-Ko dinate dem Merkmal X, die y-Koordinate der Schulnote Y. Welche Information über den Zusa menhang zwischen Selbstkonzept und Schulnote kann man der „Punktwolke" entnehmen?

Häufig untersucht man gleichzeitig zwei Merkmale (X; Y). Man interessiert sich für Abhängig keiten. Sind große Stichprobenwerte von X eher mit großen oder eher mit kleinen Stichprobe werten des Merkmals Y verknüpft?
Kurz: Besteht eine Tendenz der Form: „Je größer X, desto größer (desto kleiner) Y"?



Rollen „Dicke" weiter als „Dünne"?
Sollten Sie dieses Experiment durchführen wollen, so achten Sie auf eine ebene Rollstrecke und Windstille. Besonders geeignet sind lange Flure oder Turnhallen. Um die Varianz der Körpergewichte zu erhöhen, können Sie eine Klasse 5 zur Teilnahme am Experiment einladen.

Punktdiagramm

Eine erste Antwort auf diese Frage kann man einem Punktdiagramm entnehmen: Die Stichpro benwertepaare (kurz: „Datenpunkte") $(x_1; y_1)$, $(x_2; y_2)$, . . . , $(x_n; y_n)$ werden in einem Koordina tensystem als „Punktwolke" dargestellt.

x_i(kg)	y_i(m)
51	16,3
87	25,3
53	23,7
69	25,5
74	30,5
62	26,7
60	21,2
64	19,4
63	21,9
76	25,3
35	20,5
$\bar{x}$=62	$\bar{y}$=23,1

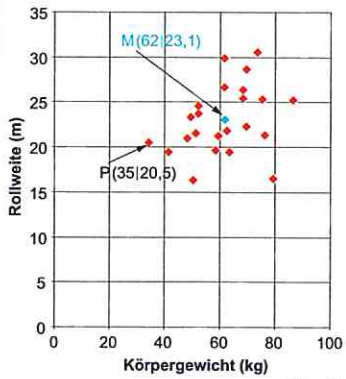


Fig. 2

Ein Experiment zur Erläuterung:
Sie stellen das Pedal eines Fahrrades nach oben, steigen auf und drücken das Pedal nu mit Ihrem Körpergewicht nach unten. Gem sen werden Körpergewicht X (in kg) und di mit dem „Schwung" aus einer halben Pedal umdrehung erreichte Rollweite Y (in m).
P(35|20,5) bedeutet z. B.: Ein Fahrer mit 35 kg Körpergewicht rollte 20,5 m weit. Rol len schwere Personen im Mittel weiter (wei sie mehr „Schwung" bekommen) oder weni ger weit (weil durch das hohe Gewicht das Rad schlechter rollt)?

Die Daten zu diesem und einem weiteren Experiment finden Sie in Rollweiten.xls.

Es scheint sich in dieser Stichprobe (Mountainbike auf einem Sportplatz) eine schwache Tendenz der Form „Je schwerer, desto weiter" abzuzeichnen. Das mittlere Gewicht der Versuchspersonen betrug 62,0 kg, im Mittel wurde die Rollweite 23,1 m erreicht.
$M(\bar{x}|\bar{y}) = M(62|23{,}1)$ bezeichnet man als den **Mittelpunkt** der Wolke.

Kovarianz – Urlisten

Bei einer Tendenz „Je größer X, desto größer Y" müssten über dem Durchschnitt liegende Stichprobenwerte von X $(x_i - \bar{x} > 0)$ häufig zusammen mit überdurchschnittlich großen Stichprobenwerten für Y auftreten: $(y_i - \bar{y} > 0)$. Entsprechendes gilt für unter dem Durchschnitt liegende Werte von X und Y. In beiden Fällen sind die Produkte $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ positiv. Um diese Tendenz zahlenmäßig zu erfassen, bildet man den Mittelwert dieser Produkte

$c_{XY} = \frac{1}{n}((x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + (x_2 - \bar{x}) \cdot (y_2 - \bar{y}) + \ldots + (x_n - \bar{x}) \cdot (y_n - \bar{y}))$.

Für das Rollexperiment erhält man

$c_{XY} = \frac{1}{24}((51 - 62) \cdot (16{,}3 - 23{,}1) + (87 - 62) \cdot (25{,}3 - 23{,}1) + \ldots + (35 - 62) \cdot (20{,}5 - 23{,}1)) \approx 15{,}5$.

c_{XY} heißt (empirische) **Kovarianz** der Merkmale X und Y. Das positive Vorzeichen von c_{XY} stützt die Vermutung, dass schwere Personen tendenziell weiter rollen.

Urliste

	x_i	y_i
1 Christina	2	2
2 Thilo	5	3
3 Emek	4	3
4 Marcel	3	4
5 Jacqueline	5	3
6 Matthias	4	3
...		
80 Daniel	5	4
81 Marc	3	4
Mittelwert	3,1	2,8

Fig. 1

Kovarianz – Häufigkeitstabellen

Nach gängiger Meinung sind „gute Mathematiker" oft musikalisch. Von 81 Schülern der 8. Klasse wurden die Noten notiert. Unter den 81 Wertepaaren $(x_1; y_1), \ldots, (x_{81}; y_{81})$ der Stichprobe kommen viele mehrfach vor, da es nur 36 verschiedene „Notenpaare" geben kann. Durch Auszählen der Urliste erhält man die Häufigkeitstabelle in Fig. 3. Das Notenpaar $(x_4; y_4) = (3; 4)$ z. B. tritt mit der relativen Häufigkeit $\frac{9}{81}$ auf. Im Punktdiagramm (Fig. 2) liegen folglich viele Datenpunkte aufeinander. Das kennzeichnet man durch Häufigkeitsangaben.

Das Stichprobenwertepaar $(x_1; y_1) = (2; 2)$ beschreibt Christinas Noten.
Das Merkmalsausprägungspaar $(x_1; y_1) = (1; 1)$ tritt mit der relativen Häufigkeit $\frac{6}{81}$ auf.

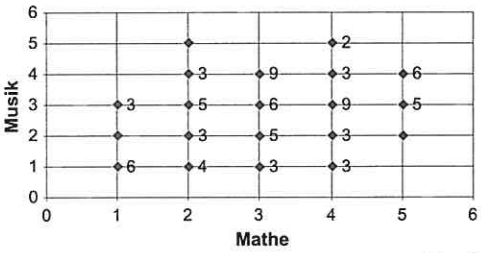


Fig. 2

y \ x	1	2	3	4	5	6	
1	6	4	3	3			16
2	1	3	5	3	1		13
3	3	5	6	9	5		28
4		3	9	3	6		21
5		1		2			3
6							0
	10	16	23	20	12	0	81

Fig. 3

Mit der Häufigkeitstabelle kann man so die Kovarianz berechnen:

$c_{XY} = (x_1 - \bar{x}) \cdot (y_1 - \bar{y}) \cdot h_{11} + (x_1 - \bar{x}) \cdot (y_2 - \bar{y}) \cdot h_{12} + \ldots + (x_r - \bar{x}) \cdot (y_s - \bar{y}) \cdot h_{rs}$

$= (1 - 3{,}1) \cdot (1 - 2{,}8) \cdot \frac{6}{81} + (1 - 3{,}1) \cdot (2 - 2{,}8) \cdot \frac{1}{81} + \ldots + (6 - 3{,}1) \cdot (6 - 2{,}8) \cdot \frac{0}{81} \approx 0{,}516 > 0$.

Die durch Fig. 2 und 3 nahe gelegte Vermutung, dass gute Mathematiknoten tendenziell mit guten Musiknoten einhergehen, wird durch das positive Vorzeichen von c_{XY} gestützt.

Beachten Sie:
Bei der Kovarianz c_{XY} interessiert nur das Vorzeichen. Die numerische Größe ist bedeutungslos. Wenn man z. B. im Rollexperiment die Rollweite in cm statt in m misst, steigt der Wert der Kovarianz von 15,5 auf 1550.
Teilt man c_{XY} durch s_X und s_Y, erhält man den ***Korrelationskoeffizienten****, der das gleiche Vorzeichen hat wie die Kovarianz. Er liegt zwischen –1 und +1 und misst tatsächlich die Stärke des Zusammenhanges (Lerneinheit 7).*

Sind $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ die Stichprobenwertepaare zweier Merkmale X, Y mit den Mittelwerten $\bar{x}$ bzw. $\bar{y}$, so nennt man

$$c_{XY} = \frac{1}{n}((x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + (x_2 - \bar{x}) \cdot (y_2 - \bar{y}) + \ldots + (x_n - \bar{x}) \cdot (y_n - \bar{y}))$$

die empirische **Kovarianz** zwischen X und Y.

Kommen die r · s Paare von Merkmalsausprägungen $(x_1; y_1), (x_1; y_2), \ldots, (x_r; y_s)$ mit den relativen Häufigkeiten $h_{11}, h_{12}, \ldots, h_{rs}$ vor, so kann man einfacher rechnen:

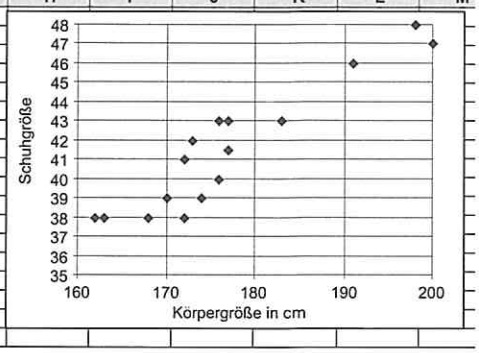
$$c_{XY} = (x_1 - \bar{x}) \cdot (y_1 - \bar{y}) \cdot h_{11} + (x_1 - \bar{x}) \cdot (y_2 - \bar{y}) \cdot h_{12} + \ldots + (x_r - \bar{x}) \cdot (y_s - \bar{y}) \cdot h_{rs}.$$

Kann man auch die zugehörigen Wahrscheinlichkeiten $p_{11}, p_{12}, \ldots, p_{rs}$ angeben, dann nennt man $\gamma_{XY} = (x_1 - \mu_X) \cdot (y_1 - \mu_Y) \cdot p_{11} + (x_1 - \mu_X) \cdot (y_2 - \mu_Y) \cdot p_{12} + \ldots + (x_r - \mu_X) \cdot (y_s - \mu_Y) \cdot p_{rs}$ die theoretische Kovarianz.

Positive Kovarianz spricht eher für einen Zusammenhang der Form: „Je größer X, desto größer Y", negative Kovarianz eher für eine Tendenz: „Je größer X, desto kleiner Y".

Beispiel 1: (Empirische Kovarianz)
In einer Stichprobe (n = 15) wurden Körpergröße X (in cm) und Schuhgröße Y von Personen erhoben. Erstellen Sie ein Punktdiagramm und berechnen Sie die Kovarianz.
(183; 43), (191; 46), (198; 48), (177; 43), (162; 38), (200; 47), (172; 38), (173; 42), (176; 43), (172; 41), (168; 38), (174; 39), (176; 40), (163; 38), (170; 39)
Lösung:
Man notiert die Wertepaare untereinander, berechnet die Mittelwerte $\overline{x} = 177$, $\overline{y} = 41{,}5$ und notiert die Differenzen $(x_i - \overline{x})$, $(y_i - \overline{y})$ und deren Produkte $(x_i - \overline{x}) \cdot (y_i - \overline{y})$. Der Mittelwer der Produkte ist die empirische Kovarianz $c_{XY} = 34{,}1$.
Obwohl es einige Wertepaare, z. B. (173; 42), (174; 39), gibt, bei denen trotz größerer Körperlän kleinere Schuhe getragen werden, ist insgesamt der Trend „je größer die Körperlänge, desto größ die Schuhe" unverkennbar. Das vereinbart sich mit dem positiven Vorzeichen der Kovarianz. In Excel (Fig. 1) erhält man die Kovarianz auch über den Befehl „=KOVARIANZ(B2:B16;C2:C16

	A	B	C	D	E	F	G
1	i	x_i	y_i	(x_i-x)	(y_i-y)	$(x_i-x)(y_i-y)$	
2	1	183	43	6	1,467	8,8	
3	2	191	46	14	4,467	62,5	
4	3	198	48	21	6,467	135,8	
5	4	177	43	0	1,467	0,0	
6	5	162	38	–15	–3,553	53,0	
7	6	200	47	23	5,467	125,7	
8	7	172	38	–5	–3,553	17,7	
9	8	173	42	–4	0,467	–1,9	
10	9	176	43	–1	1,467	-1,5	
11	10	172	41	–5	–0,553	2,7	
12	11	168	38	–9	–3,553	31,8	
13	12	174	39	–3	–2,553	7,6	
14	13	176	40	–1	–1,553	1,5	
15	14	163	38	–14	–3,553	49,5	
16	15	170	39	–7	–2,553	17,7	
17		$\overline{x}$ = 177	$\overline{y}$ = 41,5			c_{XY} = 34,1	
18							



Fig

Beispiel 2: (Theoretische Kovarianz)
Xander spielt mit seiner kleinen Schwester Yvonne ein Würfelspiel. Zuerst würfelt Xander (Merkmal X). Dann würfelt Yvonne. Sie darf aber um die größere der Augenzahlen, die Xand und sie gewürfelt haben, weiterrücken (Merkmal Y). So liefern z. B. die Augenzahlen (5; 3) da Wertepaar $(x_i; y_i) = (5; 5)$, die Augenzahlen (1; 3) das Wertepaar $(x_i; y_i) = (1; 3)$.
Berechnen Sie die theoretische Kovarianz γ_{XY}.
Lösung:
Es gibt 36 Paare von Augenzahlen, die zu insgesamt 21 verschiedenen Wertepaaren führen. Deren Wahrscheinlichkeiten erhält man durch Betrachtung aller möglichen Ergebnisse.

Das Wertepaar (4; 4) hat die Wahrscheinlichkeit $\frac{4}{36}$, denn die zugehörigen Augenzahlen sind (4; 1), (4; 2), (4; 3), (4; 4).

(1; 1)	(2; 1)	(3; 1)	(4; 1)	(5; 1)	(6; 1)
(1; 2)	(2; 2)	(3; 2)	(4; 2)	(5; 2)	(6; 2)
(1; 3)	(2; 3)	(3; 3)	(4; 3)	(5; 3)	(6; 3)
(1; 4)	(2; 4)	(3; 4)	(4; 4)	(5; 4)	(6; 4)
(1; 5)	(2; 5)	(3; 5)	(4; 5)	(5; 5)	(6; 5)
(1; 6)	(2; 6)	(3; 6)	(4; 6)	(5; 6)	(6; 6)

y \ x	1	2	3	4	5	6
1	$\frac{1}{36}$					
2	$\frac{1}{36}$	$\frac{2}{36}$				
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{3}{36}$			
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{4}{36}$		
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{5}{36}$	
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{6}{36}$

Die Wahrscheinlichkeiten für Y erhält man durch Addition der „Zeilen".

Man erhält $\mu_X = 3{,}5$ und $\mu_Y = 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36} + 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36} = \frac{161}{36} \approx 4{,}47$;
$\gamma_{XY} = (1 - \mu_X)(1 - \mu_Y)\frac{1}{36} + (1 - \mu_X)(2 - \mu_Y)\frac{1}{36} + \ldots + (6 - \mu_X)(6 - \mu_Y)\frac{6}{36} = \frac{35}{24} \approx 1{,}458$.
Die positive Kovarianz bestätigt die nahe liegende Vermutung: Je größer X, desto größer Y.

Aufgaben

2 Berechnen Sie die Kovarianz c_{XY} zu den folgenden (X; Y)-Wertepaaren.

a) (2; 0), (3; 2), (4; 0)
b) (0; 0), (1; –1), (2; –4)
c) (0; 3), (1; 3), (2; 3)
d) (3; 0), (3; 1), (3; 2)

3 Berechnen Sie die Kovarianz c_{XY} für die „Wolken" aus je 4 Punkten.

a) b) c)

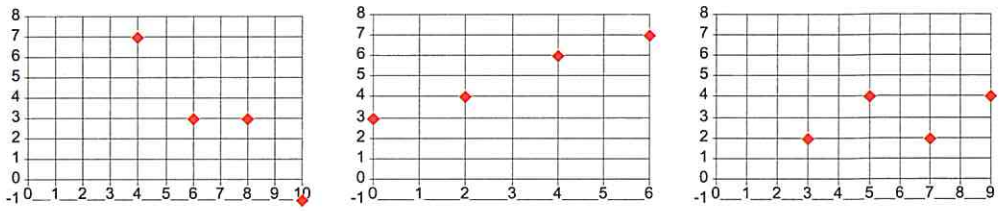


Fig. 1

4 15 Schüler liefen eine Treppe hinauf. Angegeben sind Paare aus Körpergewicht X (in kg) und benötigter Zeit Y (in s).

(53; 6,02), (66; 5,5), (46; 5), (50; 4,88), (44; 5,36), (72; 4,1), (62; 4,02),
(84; 4,1), (65; 4,2), (80; 4,3), (71; 4,15), (62; 4,8), (61; 4,45), (70; 6,4),
(50; 4,75), (64; 4,55), (67; 4,05), (68; 4,75), (65; 4,0), (84; 4,4)

a) Zeichnen Sie ein Punktdiagramm und beantworten Sie, ob in dieser Stichprobe schwerere Schüler tendenziell auch langsamer sind.
b) Zeigen Sie: $\bar{x} = 64{,}2$ (kg), $\bar{y} = 4{,}689$ (s).
c) Berechnen Sie die empirische Kovarianz und prüfen Sie Ihre Antwort aus a) rechnerisch.

5 In einer Untersuchung wurden die Mathematik- (X) und Französisch- (Y) Notenpaare von 51 Schülerinnen und Schülern einer Klassenstufe 8 erhoben.
a) Zeigen Sie: $\bar{x} = 3{,}2$ und $\bar{y} = 3{,}5$.
b) Berechnen Sie die Kovarianz zwischen den Mathematik- und Französisch-Noten. Kommentieren Sie Ihr Ergebnis.

y \ x	1	2	3	4	5	6	
1		3	2				5
2	3	2	3	4	1		13
3		1	5	3	3		12
4				9	1	1	11
5			2	5	3		10
	3	6	12	21	8	1	51

6 Zwanzig Doppelwürfe mit einem roten und einem weißen Würfel lieferten für die Merkmale X (Augenzahl des roten Würfels) und Y (Minimum der Augenzahlen beider Würfel) die folgende Urliste.

(5; 4), (3; 3), (6; 6), (2; 2), (6; 6), (2; 2), (6; 5), (6; 2), (1; 1), (3; 2), (4; 4),
(4; 4), (4; 3), (4; 2), (6; 1), (2; 2), (3; 2), (4; 3), (2; 2), (3; 3)

a) Zeigen Sie: $\bar{x} = 3{,}8$ und $\bar{y} = 2{,}95$. Berechnen Sie die empirische Kovarianz.
b) Berechnen Sie wie im Beispiel 2 die theoretische Kovarianz. Vergleichen Sie mit Ihrem Ergebnis von a).

Wenn Sie mit EXCEL „würfeln" wollen, finden Sie eine Anleitung in Würfel-Kovarianz.xls.

7 „Werfen" Sie ein Paar „verschiedenfarbiger" Würfel 20-mal. A sei die Augenzahl des ersten, B die des zweiten Würfels. Untersuchen Sie die empirische und die theoretische Kovarianz zwischen den Merkmalen X und Y. Orientieren Sie sich am Beispiel 2.

a) X = A, Y = B
b) X = A, Y = A + B
c) X = A + B und Y = A – B

5 Lineare Regression

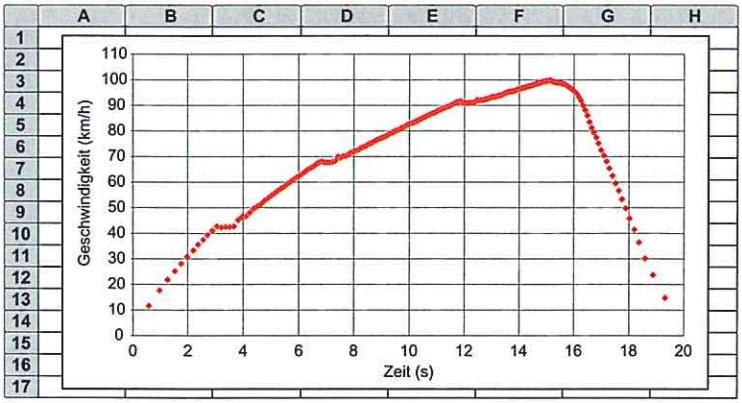


Fig. 1

Für eine genauere Analyse finden Sie die Daten in der Datei VW-Passat.xls.

1 Während eines Belastungstests wurde ein PKW mit Vollgas beschleunigt und nach Erreichen einer Geschwindigkeit von $100\,\frac{km}{h}$ „hart abgebremst".

a) Fassen Sie die Informationen, die Sie der Grafik in Fig. 1 entnehmen können, in eigene Worte.

b) Nach welcher Zeit hätte der PKW theoretisch die Geschwindigkeit $100\,\frac{km}{h}$ erreichen können, wenn der Fahrer während der ganzen Zeit das Auto im ersten (zweiten) Gang hätte beschleunigen können ohne schalten zu müssen?

Lineare Modelle

In vielen Anwendungssituationen scheint es gerechtfertigt, zwischen zwei Merkmalen X und Y wie z. B. Körpergröße und Körpergewicht (zumindest in gewissen Bereichen), eine lineare Beziehung $Y = \alpha \cdot X + \beta$ zu unterstellen.

Mit dieser Formel beschreibt man, welchen Wert man für das Merkmal Y erwartet, wenn der Wert des Merkmals X bekannt ist. So stammt die „Normalgewichtsbeziehung" $Y = 1 \cdot X - 10$ zwischen Körpergröße X (in cm) und Körpergewicht Y (in kg) von dem französischen Arzt BROCA (1824–1880). Hier ist $\alpha = 1$ und $\beta = -100$. Für eine Person mit der Körpergröße 172 cm erwartet man nach BROCA das Körpergewicht 72 kg.

Natürlich sind gleich große Menschen in der Regel nicht gleich schwer. Daher denkt man sich in jedem Einzelfall das tatsächliche Gewicht Y zusammengesetzt aus dem Normalgewicht $\alpha X + \beta$ und einer zufälligen **Störgröße D**, die bei „Übergewichtigen" positiv, bei „Untergewichtigen" negativ ist.

Eike (172; 65,2) hat Untergewicht: $D = -6{,}8$ (kg).

Sven (185; 90) hat Übergewicht: $D = +5$ (kg).

Über die Gesamtbevölkerung gemittelt sollte aber gelten $\bar{d} \approx 0$ bzw. $\mu_D = 0$.

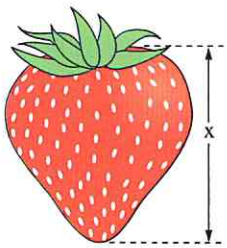
Andernfalls wäre die Bezeichnung „Normalgewicht" nicht gerechtfertigt.

Die Beziehung $\mathbf{Y = \alpha \cdot X + \beta}$ bezeichnet man als ein **„lineares Modell"**. Für jedes einzelne Wertepaar gilt jedoch $Y = \alpha X + \beta + D$.

Der Steigungsfaktor α heißt theoretischer **Regressionskoeffizient**.

Die Gleichung $Y = \alpha \cdot X + \beta + D$ bedeutet: Für jedes Wertepaar $(x_i; y_i)$ gilt $y_i = \alpha x_i + \beta + d_i$. Dabei ist $\alpha x_i + \beta$ der zu $X = x_i$ gehörige erwartete Wert von Y und d_i die Abweichung des beobachteten vom erwarteten Wert.

***regredere** (lat.): zurückschreiten. Man führt das Merkmal Y auf das Merkmal X zurück. Die Werte d_i der Störgröße D bezeichnet man als Residuen (das, was „übrig bleibt").*



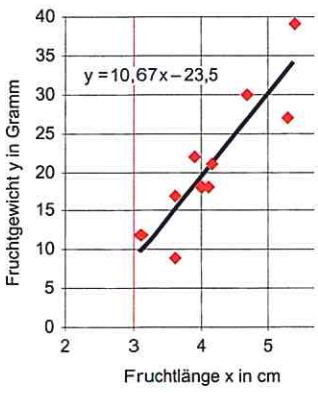
Schätzen der Parameter α und β

In der Regel sind die Parameter α und β unbekannt, man muss sie aus Messwerten schätzen.

Das sei am Beispiel von Erdbeeren erläutert, für die man ebenso wie für Menschen eine „Normalgewichtsformel" aufstellen kann.

In der Stichprobe (Fig. 2) gehört zur mittleren Fruchtlänge $\bar{x} = 4{,}17$ (in cm) das mittlere Gewicht $\bar{y} = 21$ (in g).

x_i	y_i
5,4	39
5,3	27
3,9	22
3,1	12
4	18
4,1	18
3,6	9
4,7	30
3,6	17
4	18
$\bar{x} = 4{,}17$	$\bar{y} = 21$



Fig

Größere Erdbeeren $(x - \bar{x} > 0)$ sind tendenziell schwerer $(y - \bar{y} > 0)$. Man nimmt dem linearen Modell entsprechend an, dass Gewichts- und Größenzunahme proportional sind, dass also im Mittel gilt $(y_i - \bar{y}) \approx \alpha \cdot (x_i - \bar{x})$. Dann erhält man für die empirische Kovarianz

$$c_{XY} = \frac{1}{n}((x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + (x_2 - \bar{x}) \cdot (y_2 - \bar{y}) + \ldots + (x_n - \bar{x}) \cdot (y_n - \bar{y}))$$
$$\approx \frac{1}{n}(\alpha \cdot (x_1 - \bar{x})^2 + \alpha \cdot (x_2 - \bar{x})^2 + \ldots + \alpha \cdot (x_n - \bar{x})^2)$$
$$= \alpha \cdot \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2) = \alpha \cdot V_X.$$

Damit ergibt sich $\alpha \approx \frac{c_{XY}}{V_X}$.

Der unbekannte theoretische Regressionskoeffizient α des linearen Modells lässt sich folglich durch den aus der Stichprobe ermittelten Wert $\alpha = \frac{c_{XY}}{V_X}$ schätzen, es gilt $\alpha \approx a$.

Daher bezeichnet man a als den **empirischen Regressionskoeffizienten**.

Die Urlisten zum Erdbeer-Beispiel finden sich in Erdbeere.xls.

Für die Erdbeeren (Fig. 3 der Seite gegenüber) gilt $c_{XY} = 5{,}34$ und $V_X = 0{,}5$. Der empirische Regressionskoeffizient ist damit $a = \frac{5{,}34}{0{,}50} \approx 10{,}67$. Die empirische Regressionsgerade hat die Gleichung $y = a \cdot (x - \bar{x}) + \bar{y} = 10{,}67\,(x - 4{,}17) + 21 = 10{,}67\,x - 23{,}5$.

Der Achsenabschnitt $b = -23{,}5$ ist eine Schätzung für β.

Eine andere Stichprobe aus 45 Erdbeeren lieferte $y = 11{,}0\,x - 25{,}8$.

Beide Geraden beschreiben im Bereich von 2 cm bis 5 cm den tendenziellen Zusammenhang zwischen Länge und Gewicht. Je Zentimeter Länge nimmt damit das Gewicht im Mittel um ca. 11 Gramm zu.

Die Gleichung der theoretischen Regressionsgeraden ist unbekannt.

Gegeben sind zwei Merkmale X und Y mit den Mittelwerten $\bar{x}$ bzw. $\bar{y}$, der empirischen Varianz V_X und der empirischen Kovarianz c_{XY}.

Der Quotient $a = \frac{c_{XY}}{V_X} = \frac{c_{XY}}{s_X^2}$ heißt empirischer **Regressionskoeffizient**.

Die Gerade durch den Mittelpunkt $M(\bar{x}\,|\,\bar{y})$ der Punktwolke mit Steigung a heißt *empirische* **Regressionsgerade** oder **Trendgerade**.

Sie hat die Gleichung $y = a \cdot (x - \bar{x}) + \bar{y}$.

Theoretische Regressionsgerade
$y = \alpha \cdot (x - \mu_X) + \mu_Y$
(meist unbekannt)

Empirische Regressionsgerade
$y = a \cdot (x - \bar{x}) + \bar{y}$
(hängt von der Stichprobe ab)

Bei verschiedenen Stichproben ergeben sich in der Regel unterschiedliche Punktwolken und damit unterschiedliche empirische Regressionsgeraden.

Die empirischen Regressionsgeraden $y = a \cdot (x - \bar{x}) + \bar{y}$ schwanken wegen $a \approx \alpha$, $\bar{x} \approx \mu_X$, $\bar{y} \approx \mu_Y$ um die theoretische Regressionsgerade.

In der Praxis ist die theoretische Regressionsgerade selten bekannt, man arbeitet ersatzweise mit der geschätzten empirischen Regressionsgeraden.

Beispiel 1: (Schuhgröße – Schuhlänge)
Mit einem Stahlband wurde von 18 Schuhen zwischen Größe $X = 19$ und $X = 42$ die Innenlänge Y gemessen.

(42; 25,7), (34; 22), (35; 22,1), (33; 19,5), (35; 21,8), (33; 20,5), (34; 22),
(33; 20,8), (34; 21,5), (34; 21), (33; 20,5), (39; 24,5), (39; 24,8), (39; 24,3),
(42; 25,8), (42; 26), (19; 12,2), (21; 13)

a) Ermitteln Sie die Gleichung der empirischen Regressionsgeraden.

b) Eine andere Stichprobe lieferte die empirische Regressionsgerade $y = 0{,}669\,x - 0{,}732$.
Eine Verkäuferin erinnert sich an die Formel „Schuhlänge = doppelte Schuhgröße durch drei". Erläutern Sie den Zusammenhang.

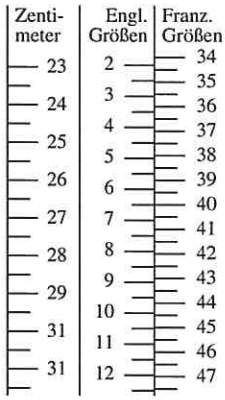


Fig. 1

Die Urlisten zu Beispiel 1 finden Sie in Schuhe.xls.

Lösung:

Für die Stichprobe gilt $\bar{x} \approx 34{,}5$ (Schuhgröße) und $\bar{y} \approx 21{,}6$ (Innenlänge in cm). Man erhält

$c_{XY} \approx \frac{1}{18}((42-34{,}5)(25{,}7-21{,}6)+(34-34{,}5)(22-21{,}6)+\ldots+(21-34{,}5)(13-21{,}6)) \approx 22{,}478;$

$V_X \approx \frac{1}{18}((42-34{,}5)^2+(34-34{,}5)^2+\ldots+(21-34{,}5)^2) \approx 36{,}806$ und $a \approx \frac{22{,}478}{36{,}806} \approx 0{,}611.$

Die empirische Regressionsgerade hat damit die Gleichung

$y \approx 0{,}611\,(x-34{,}5)+21{,}6 \approx 0{,}611\,x-0{,}486.$

b) Die im Schuhgeschäft genannte Gleichung $y = \frac{2}{3}x$ ist die in der Produktion vorgeschriebe theoretische Regressionsgerade. Schuhe gleicher Größe fallen aber aufgrund von Toleranzen unterschiedlich groß aus, auch die Innenlänge lässt sich nur bis auf einige Millimeter genau messen. Wegen dieser „Störgrößen" „streuen" die Messpunkte und damit auch die empirische Regressionsgeraden um die theoretische Regressionsgerade.

Beispiel 2: (EXCEL)

Zeichnen Sie Punktdiagramm und Regressionsgerade zu der Stichprobe aus Schuhgröße und Innenlänge (Beispiel 1) mit EXCEL.

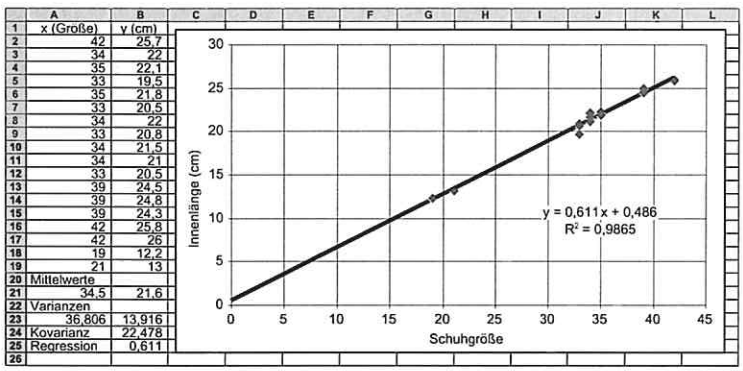


Fig. 2

Lösung:

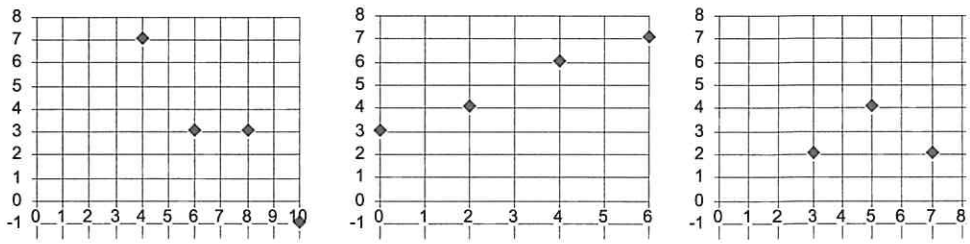
Man markiert die Urliste (Fig. 2) im Bereich A2 bis B19, wählt „Einfügen", „Diagramm" „X-Y-Punktdiagramm" und lässt sich vom Diagramm-Assistenten bei der Erstellung führen. Das Diagramm wird anschließend geöffnet, die Punktwolke wird markiert. Mi der rechten Maustaste öffnet man ein Menü in dem man „Trendlinie zeichnen" anklickt und die Option „Formel einblenden" wählt. Die empirische Regressionsgerade wird mit der Geradengleichung in das Punktdiagram eingetragen.

Den empirischen Regressionskoeffizienten kann man auch über den Befehl „=Steigung(B2:B19;A2:A19)" (Fig. 2, Zelle B25) berechnen.

Ebenso erhält man den y-Achsenabschnitt über „=Achsenabschnitt(B2:B19;A2:A19)".

Aufgaben

2 a) Schätzen Sie nach Augenmaß die Gleichungen der Regressionsgeraden zu den folgenden Punktwolken.



Fig

b) Berechnen Sie die Gleichungen der Regressionsgeraden. Vergleichen Sie mit Ihrer Schätzung.

3 Gegeben sind die folgenden Wertepaare. Zeichnen Sie je ein Punktdiagramm, ermitteln und zeichnen Sie die empirische Regressionsgerade.

a) (3; 5), (4; 7), (5; 12) b) (5; 3), (7; 4), (12; 5) c) (0; 6), (2; 2), (4; 1)
d) (3; 4), (5; 4), (7; 4) e) (2; 6), (4; 5), (6; 4) f) (–2; 4), (–1; 1), (0; 0), (1; 1), (2; 4)

4 Mirjam, Iman und Roxana sollen bei einem „Schülerexperiment" mehrere Schokoladetafeln an Federn oder Gummibändern aufhängen und die Länge Y der Feder (in cm) ablesen.

Anzahl der Tafeln	X	1	2	3	4	5
Mirjam	Y_1	12,7	13,6	14,8	16,3	17,9
Iman	Y_2	11,4	11,9	12,7	13,7	14,6
Roxana	Y_3	6,4	10,1	17,6	21,9	22,4

a) Zeichnen Sie die (X; Y)-Datenpaare in ein Koordinatensystem, ermitteln und zeichnen Sie jeweils die Regressionsgerade.
b) Wie lang waren vermutlich die Federn bzw. Gummibänder, als noch keine Tafel angehängt war (X = 0)?
c) Welche Länge erwarten Sie bei einer Belastung durch 8 Tafeln?
d) Welche inhaltliche Bedeutung besitzt der Regressionskoeffizient?
e) Bei dem Experiment wurden tatsächlich zwei Federn und ein Gummiband verwendet. Wer benutzte vermutlich das Gummiband? Begründen Sie Ihre Antwort.

Wenn Sie ohne Tabellenkalkulation arbeiten, beschränken Sie sich auf die kursiv gedruckten Daten. Andernfalls können Sie auch den vollen Datensatz aus VW-Passat.xls auswerten.

5 Bei den folgenden Datenpaaren handelt es sich um Zeiten (in Sekunden) und die zugehörigen Geschwindigkeiten (in $\frac{km}{h}$) der in Fig. 1 von Aufgabe 1 dargestellten PKW-Testfahrt. Ermitteln Sie jeweils die Regressionsgeraden.
a) Erster Gang: *(0,57; 11,70)*, (0,95; 17,70), (1,25; 21,83), *(1,52; 25,13)*, (1,75; 28,14), (1,97; 30,86), *(2,17; 33,34)*, (2,35; 35,57), (2,53; 37,60), *(2,70; 39,44)*.
b) Zweiter Gang: *(4,08; 46,70)*, (4,35; 49,66), (4,61; 51,52), *(4,86; 53,68)*, (5,10; 55,69), (5,34; 57,51), *(5,56; 59,31)*, (5,78; 60,97), (6,00; 62,57).
c) Vollbremsung: *(16,54; 83,51)*, (16,71; 79,42), (16,88; 75,20), *(17,07; 70,42)*, (17,27; 65,44), (17,49; 59,48), *(17,73; 53,37)*, (18,01; 45,88), (18,35; 36,48), *(18,85; 23,77)*.
d) Welche inhaltliche Bedeutung haben die Regressionskoeffizienten und die Schnittpunkte der Regressionsgeraden mit den Koordinatenachsen?

Schwankungen der Regressionsgeraden

Die Urliste findet sich in der Datei Kartoffeln.xls.

6 In einer Stichprobe wurden die Längen X (in cm) und die Gewichte Y (in g) von 39 Kartoffeln gemessen.
a) Die „ersten" 4 Kartoffeln der Stichprobe hatten die Maße (7,9; 110), (7,8; 77), (6,4; 69), (8,7; 79). Ermitteln Sie die Regressionsgerade zu dieser Teil-Stichprobe.
b) Die „letzten" 4 Kartoffeln lieferten (5,7; 59), (8,3; 64), (6,4; 54), (7,7; 97). Ermitteln Sie die Regressionsgerade zu dieser Teil-Stichprobe und vergleichen Sie mit a).
c) Lesen Sie aus dem Diagramm 4 Punkte ab, die als Teil-Stichprobe sogar einen negativen Regressionskoeffizienten geliefert hätten.

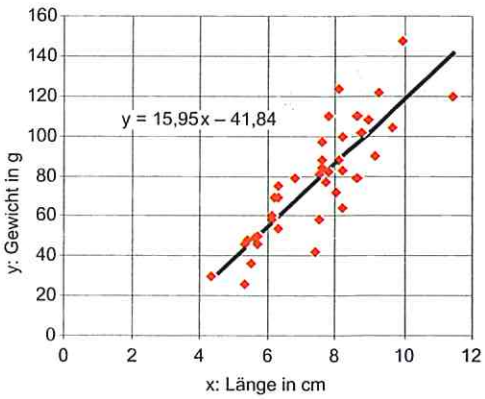
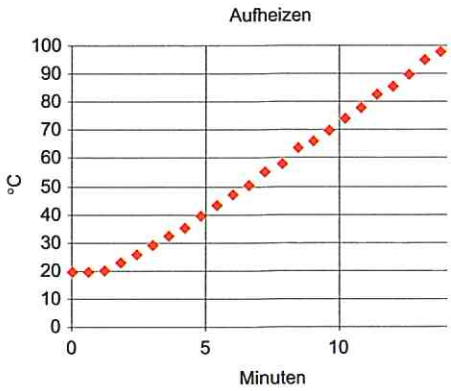


Fig. 1

Grenzen linearer Modelle

d) Wenn man die gesamte Stichprobe untersucht, erhält man die Regressionsgerade $y = 15{,}95\,x - 41{,}84$. Dieter wundert sich, dass für $x = 2\,cm$ ein negatives Gewicht vorhergesagt wird. Helfen Sie ihm bei der Deutung.

7 Mirjam hat Wasser auf dem Herd zum Kochen gebracht und den Temperaturanstieg protokolliert. Die Messung beginnt zur Zeit t = 0 min mit der Raumtemperatur Y = 19,60 (in °C).
Die folgenden Zeit-Temperatur-Wertepaare sind Auszüge aus dem von Mirjam angefertigten Messprotokoll.
a) (0,6; 19,61), (1,2; 20,08), (1,8; 22,56), (2,4; 25,55)
b) (3; 29,01), (3,6; 32,05), (4,2; 34,9), (4,8; 39,46)
c) (5,4; 43,29), (6; 46,84), (6,6; 50,13), (7,2; 55,05)



Fi

Ermitteln Sie jeweils die Regressionsgerade und bestimmen Sie den Zeitpunkt, zu dem man c Kochen des Wassers (Y = 100) erwartet hätte.
Bewerten Sie die Brauchbarkeit des linearen Modells im Vergleich mit den tatsächlichen Mes werten.

8 Weltrekorde

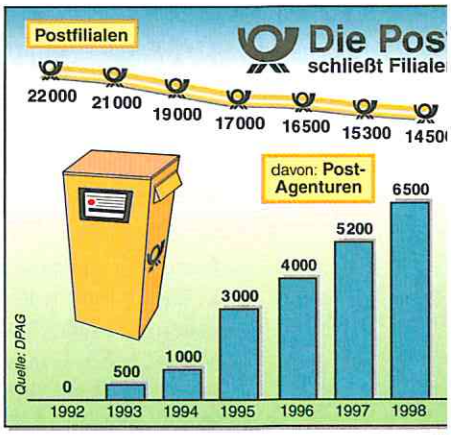
a) Mirko möchte mithilfe einer Regressionsgeraden vorhersagen, in welchem Jahr der 100 m-Weltrekord bei 9,5 Sekunden liegen wird. Zu welchem Ergebnis kommt er?
b) Wie bewerten Sie Mirkos Ansatz und das Rechenergebnis?

		Sekund
21.06.1960	Armin Hary	10,0
14.10.1968	Jim Hines	9,95
03.07.1983	Calvin Smith	9,93
25.08.1991	Carl Lewis	9,86
06.07.1994	Leroy Burell	9,85
27.07.1996	Donovan Bailey	9,84
16.06.1999	Maurice Greene	9,79

Eine „Verbesserung" (?) der Formel von Broca

9 Sterben der Postämter

a) Michael schätzt mithilfe der Regressionsrechnung, in welchem Jahr die letzte Postfiliale schließen wird.
Zu welchem Ergebnis kommt er?
b) Mario berechnet ebenfalls mit Hilfe der Regressionsrechnung, wann alle Postfilialen Postagenturen sein werden.
Zu welchem Ergebnis kommt Michael?
c) Bewerten Sie die Rechenergebnisse.



Fi

10 Regressionsgeraden einmal anders

Frank schätzt: Ein durchschnittlicher Mann ist 180 cm groß, wiegt 80 kg und hat die Bundweite 90 cm.
a) Mit den Kreisformeln $U = 2\pi r$ und $A = r^2\pi$ berechnet er den Körperquerschnitt (in cm^2) in Bundhöhe. Zu welchem Ergebnis kommt Frank?
b) Um das Gewicht eines 181 cm großen Mannes zu schätzen, denkt sich Frank in Bundhöhe „eine Scheibe von der Dicke 1 cm" eingeschoben.
Zu welchem Schätzergebnis kommt Frank, wenn 1000 cm^3 menschliches Gewebe 1 Kilogrra wiegen?
c) Welche lineare Beziehung zwischen Körpergröße X und Körpergewicht Y ergibt sich aus Franks Ansatz?

Lineare Modelle und quadratische Abhängigkeit

Züge.xls enthält zum Vergleich die Messwerte zu weiteren Messungen.

11 Bahnhofsmathematik

Beim Anfahren eines IC in Köln Hbf. wurden von 12 Schülern die Zeiten gestoppt, zu denen die 12 Wagen die Messstelle (am hinteren Puffer der stehenden Lok) passierten.
Der vom Zug zurückgelegte Weg Y (in m) scheint **quadratisch** von der verstrichenen Zeit X (in s) abzuhängen: $Y = c\,X^2$. Welchen Wert liefert die lineare Regressionsrechnung für die Konstante c?

Die empirische Regressionsgerade geht bei solchen Messungen selten genau durch den Koordinatenursprung. Ihre Schnittstelle x_0 mit der Zeitachse kann als Startzeitpunkt der Lok gelten, der in der Praxis wegen des langsamen Anfahrens nur ungenau festgelegt werden kann.
Wer genauer arbeiten möchte, subtrahiert von allen Zeitangaben den Wert x_0.

Mit $Z = \sqrt{Y}$ ist die quadratische Abhängigkeit gleichwertig zu einer linearen Abhängigkeit $Z = \sqrt{c}\,X$ zwischen der Wurzel aus dem Weg Z und der Zeit X. Man kann die Frage daher durch Berechnung einer linearen Regression beantworten.

a) Berechnen und zeichnen Sie die 12 (X; Z)-Wertepaare. Sie können Ihre Ergebnisse mit dem Anfang der dritten Spalte von Fig. 1 vergleichen.

b) Berechnen und zeichnen Sie die zugehörige lineare Regressionsgerade ($Z = a\,X + b$).

c) Welchen Wert liefert der empirische Regressionskoeffizient a für die Konstante c? Zeichnen Sie die (X; Y)-Punktwolke (Fig. 1) zusammen mit der zugehörigen „Regressionsparabel" $Y = c\,X^2$ in ein Koordinatensystem.

d) Wann hätte die Lok den Bahnhof Köln-Deutz (Y = 2200 m) erreicht, wenn der Zug mit gleicher Beschleunigung hätte weiterfahren können?

	x_i	y_i	z_i
0	0	0	0,00
1	20,39	26,4	5,14
2	28,38	52,8	
3	34,14	79,2	
4	38,93	105,6	
5	43,22	132	
6	46,94	158,4	
7	50,31	184,8	
8	53,43	211,2	
9	56,35	237,6	
10	59,07	264	
11	61,54	290,4	
12	64	316,8	

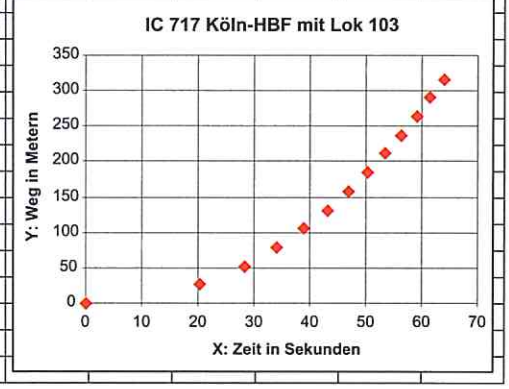


Fig. 1

Lineare Modelle und exponentielle Abhängigkeit

12 Frau Sterk gießt Tee aus der Thermoskanne in eine Tasse. Da wird sie zum Vertretungsunterricht in ihre 5d gerufen. Herr Konietzko, ein engagierter Mathelehrer, misst die Temperatur in der Tasse im Abstand von 10 Minuten (Fig. 2). Die Temperaturdifferenz Y zur Raumtemperatur (20 °C) nimmt in jeder Minute um einen festen Faktor c ab. Welche exponentielle Beziehung $Y = d\,c^X$ liefert die Regressionsrechnung?
Wenn man definiert $Z = \lg(Y)$, ist die exponentielle Abhängigkeit gleichwertig zur linearen Abhängigkeit $Z = \lg(c)\,X + \lg(d)$ zwischen dem Logarithmus der Temperaturdifferenz Z und der verstrichenen Zeit X.

Zur Erinnerung:
$\lg(d \cdot c^x)$
$= \lg(d) + \lg(c^x)$
$= \lg(d) + x \cdot \lg(c)$

a) Berechnen und zeichnen Sie die (X; Z)-Wertepaare als „Punktwolke".

b) Berechnen und zeichnen Sie die zugehörige lineare Regressionsgerade $Z = a\,X + b$ ins gleiche Koordinatensystem.

c) Welcher Wert ergibt sich aus der empirischen Regressionsgeraden für die Konstanten c und d?
Zeichnen Sie die (X; Y)-Punktwolke (Fig. 2) und die zugehörige Exponentialfunktion mit $Y = d\,c^X$.

d) Welche Temperatur hatte der Tee zum Zeitpunkt X = 0, als er eingegossen wurde?

x	y	z
0		
10	31,1	1,49
20	17,4	1,24
30	10,5	
40	6,6	
50	4,3	
60	2,9	

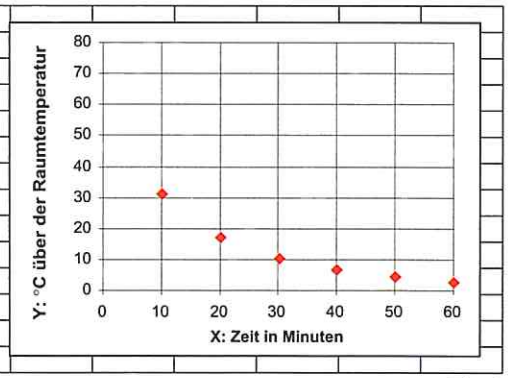


Fig. 2

6 Minimalitätseigenschaft der Regressionsgeraden

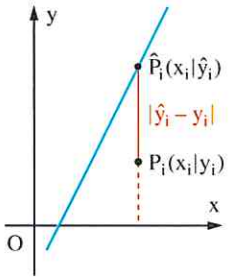
1 Gegeben sind die Zahlen 1, 4, 7.

a) Bestimmen Sie die Zahl a, für welche die „Summe der Abstandsquadrate“ $(1-x)^2+(4-x)^2+(7-x)^2$ – und damit die „Streuung“ um x – am kleinsten wird.

b) Wie ändert sich das Ergebnis in a), wenn man die Zahl 7 durch 6 (durch 8) ersetzt?

c) Verallgemeinern Sie Ihre Entdeckung auf mehr als drei Zahlen.

Tipp zu a):
Bestimmen Sie den Scheitelpunkt einer Parabel.



		$g_{1,5}$	g_2
x_i	y_i	$\hat{y}_i$	$\hat{y}_i$
1	1	0,5	−1
3	2	3,5	3
8	12	11	13

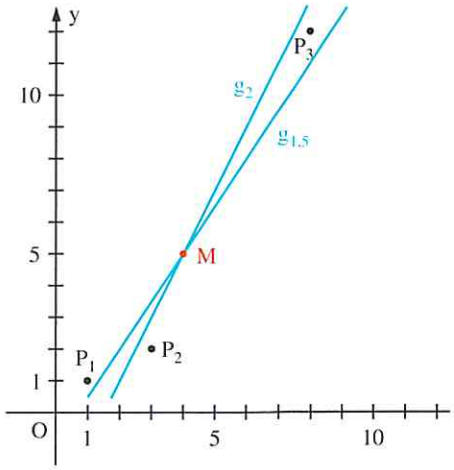
Mit $\hat{y}_i$ werden hier die y-Koordinaten der auf der Geraden g_a liegenden Punkte bezeichnet.

Fig. 1 zeigt eine „Wolke“ aus drei Punkten $P_1(1|1)$, $P_2(3|2)$, $P_3(8|12)$ mit dem Mittelpunkt $M(\bar{x}|\bar{y}) = M(4|5)$. Die Geraden $g_{1,5}$: $y = 1{,}5\cdot(x-4)+5 = 1{,}5x-1$ und g_2: $y = 2\cdot(x-4)+5 = 2x-3$ mit den Steigungen 1,5 bzw. 2 gehen durch M. Sie „passen“ beide gut zu der Punktwolke. Um zu entscheiden, welche der Geraden besser „passt“, misst man die Güte der Anpassung durch den Mittelwert der „vertikalen Abstandsquadrate“, also durch die Varianz der Punktwolke um die Gerade $g_{1,5}$ bzw. g_2:

$V_{1,5} = \frac{1}{3}[(0{,}5-1)^2+(3{,}5-2)^2+(11-12)^2] \approx 1{,}17$;

$V_2 = \frac{1}{3}[(-1-1)^2+(3-2)^2+(13-12)^2] = 2$.

Damit „passt“ $g_{1,5}$ besser als g_2.



Fi

Die folgende Rechnung zeigt, dass die Varianz V_a am kleinsten ist, wenn g_a die Regressionsgerade ist. Für die Gerade g_a: $y = a\cdot(x-\bar{x})+\bar{y}$ durch $M(\bar{x}|\bar{y})$ mit der Steigung a gilt:

x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - y_i)^2$
1	1	$a(1-\bar{x})+\bar{y}$	$[a(1-\bar{x})+\bar{y}-1]^2 = [a(1-\bar{x})-(1-\bar{y})]^2 = a^2\cdot(1-\bar{x})^2 - 2a\cdot(1-\bar{x})(1-\bar{y}) + (1-\bar{y})^2$
3	2	$a(3-\bar{x})+\bar{y}$	$[a(3-\bar{x})+\bar{y}-2]^2 = [a(3-\bar{x})-(2-\bar{y})]^2 = a^2\cdot(3-\bar{x})^2 - 2a\cdot(3-\bar{x})(2-\bar{y}) + (2-\bar{y})^2$
8	12	$a(8-\bar{x})+\bar{y}$	$[a(8-\bar{x})+\bar{y}-12]^2 = [a(8-\bar{x})-(12-\bar{y})]^2 = a^2\cdot(8-\bar{x})^2 - 2a\cdot(8-\bar{x})(12-\bar{y}) + (12-\bar{y})^2$
			Mittelwert: $V_a = a^2\cdot V_X - 2a\cdot c_{XY} + V_Y$

Fasst man nun $V_a = a^2\cdot V_X - 2a\cdot c_{XY} + V_Y$ als eine Funktion von a auf, so ist ihr Graph eine Parabel mit dem Scheitel $S\left(\frac{c_{XY}}{V_X}\middle|V_Y - \frac{c_{XY}^2}{V_X}\right)$. Damit hat die Varianz V_a ihr Minimum bei $a = \frac{c_{XY}}{V_X}$, wenn also die Gerade g_a mit der empirischen Regressionsgeraden übereinstimmt.

Für die betrachtete Punktwolke gilt näherungsweise S(1,65|0,96): Die Regressionsgerade hat die Steigung $a \approx 1{,}65$, die minimale Varianz beträgt 0,96. Allgemein gilt:

Die Parabel der quadratischen Funktion f mit $f(x) = ax^2+bx+c$ hat den Scheitelpunkt
$S\left(\frac{-\frac{b}{2}}{a}\middle|c-\frac{\left(\frac{b}{2}\right)^2}{a}\right)$.

Setzen Sie:
$a = V_X$,
$b = -2c_{XY}$,
$c = V_Y$.

Berechnungen von V_a für verschiedene Werte von a und die zugehörige Parabel finden sich in der Datei Minimalität.xls.

Gegeben ist eine Punktwolke mit dem Mittelpunkt $M(\bar{x}|\bar{y})$. Die Gerade durch M, um die di Punktwolke am wenigsten „streut“, ist die Regressionsgerade.
Für die Varianz V_D der Punktwolke um die Regressionsgerade gilt: $V_D = V_Y - \frac{c_{XY}^2}{V_X}$.

Da die Regressionsgerade „optimal“ zur Punktwolke passt, d. h. die Streuungen der Punktwol besonders „ausgleicht“, bezeichnet man sie auch als **Ausgleichsgerade**.

Beispiel: (Bestimmung einer Ausgleichsgeraden)
Gegeben ist die Punktwolke $P_1(8|1)$, $P_2(5|4)$, $P_3(4|4{,}5)$, $P_4(2|6)$.
a) Berechnen Sie die Varianz der Punktwolke um die Gerade durch $M(4{,}75|3{,}875)$ mit der Steigung -1.
b) Bestimmen Sie die Gleichung der Regressionsgeraden bzw. Ausgleichsgeraden g.
c) Berechnen Sie die Varianz V_D der Punktwolke um die Regressionsgerade.

Zu b):

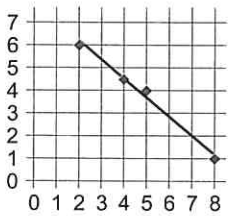


Fig. 1

Lösung:
a) Geradengleichung von g_{-1}: $y = -1 \cdot (x - 4{,}75) + 3{,}875 = -x + 8{,}625$;
$V = \frac{1}{4}[(0{,}625 - 1)^2 + (3{,}625 - 4)^2 + (4{,}625 - 4{,}5)^2 + (6{,}625 - 6)^2] \approx 0{,}17$.
b) $a = \frac{c_{XY}}{V_X} = \frac{-3{,}906\,25}{4{,}6875} \approx -0{,}833$; also g: $y = -0{,}833\,(x - 4{,}75) + 3{,}875 = -0{,}833\,x + 7{,}833$.
c) $V_D = V_Y - \frac{c_{XY}^2}{V_X} = 3{,}296\,875 - \frac{(-3{,}906\,25)^2}{4{,}6875} \approx 0{,}042$.

Aufgaben

2 Gegeben ist eine „Wolke" $P_1(-1|1)$, $P_2(4|2)$, $P_3(9|9)$ mit dem Mittelpunkt $M(4|4)$.
a) Zeichnen Sie die Punktwolke und die Geraden durch M mit den Steigungen 0,5; 1 und 1,5.
b) Berechnen Sie die Varianz der Punktwolke um diese drei Geraden.
c) Berechnen Sie die Gleichung der Regressionsgeraden und die Varianz der Punktwolke um die Regressionsgerade. Vergleichen Sie mit den Ergebnissen von Aufgabenteil b).

3 Die Erhebung von Körpergröße X (in cm) und Gewicht Y (in kg) in einer Klassenstufe 11 lieferte die folgende Urliste.

(197; 87), (186; 76), (183; 66), (167; 55), (188; 60), (188; 72), (170; 60), (174; 59), (180; 66), (171; 75), (161; 66), (180; 73), (181; 63), (184; 79), (160; 49), (181; 76), (175; 65), (168; 68), (176; 60)

a) Berechnen Sie unter Benutzung der Zwischenergebnisse $\bar{x} = 177{,}4$; $\bar{y} = 67{,}1$ und $s_X = 9{,}4$ die Gleichung der Regressionsgeraden.
Zeichnen Sie diese Gerade zusammen mit der Punktwolke in ein Koordinatensystem.
b) Wenn man das „Normalgewicht" $y_{normal} = 1 \cdot (x - 100)$ um 10 % reduziert, erhält man das Idealgewicht $y_{ideal} = 0{,}9 \cdot (x - 100)$. Ergänzen Sie Ihre Zeichnung durch diese beiden Geraden.
c) Welche der drei Geraden passt am besten zu der Punktwolke? Berechnen und vergleichen Sie die zugehörigen Varianzen.

4 Zeigen Sie, dass der Term $V_D = V_Y - \frac{c_{XY}^2}{V_X}$ für die Varianz der Punktwolke um die Regressionsgerade den Wert 0 annimmt, wenn die Wolke nur aus zwei Punkten besteht. Rechnen Sie mit $P(2|-1)$ und $Q(4|1)$ und allgemein. Deuten Sie das Ergebnis.

5 Ursprungs-Ausgleichsgerade
Gegeben sind die Punkte $P_1(0|-2)$, $P_2(2|2)$, $P_3(4|3)$, $P_4(6|5)$.
a) Berechnen Sie die Varianz der Punktwolke um die Ursprungsgeraden mit $y = a \cdot x$ für die Steigungen 0,7; 1 und 3 und allgemein für die Steigung a.
b) Für welche Steigung a wird die Varianz minimal? Wie groß ist sie dann?
c) Beweisen Sie: Die optimale Ursprungs-Ausgleichsgerade zu der Punktwolke $P_1(x_1|y_1)$, $P_2(x_2|y_2)$, ..., $P_n(x_n|y_n)$ hat die Steigung $a = \frac{x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n}{x_1^2 + x_2^2 + \ldots + x_n^2}$.
Berechnen Sie a für die gegebene Punktwolke. Vergleichen Sie mit Ihrem Ergebnis aus b).

7 Korrelation, Bestimmtheitsmaß

1 Tabellenkalkulationsprogramme können neben Regressionsgeraden auch ein „Bestimmtheitsmaß r^2" berechnen. Fig. 1 bis 3 zeigen „Punktwolken" mit gleichen empirischen Regressionsgeraden und den von EXCEL berechneten „r^2-Werten".
Mit welcher Eigenschaft der Punktwolke bringen Sie r^2 in Zusammenhang?

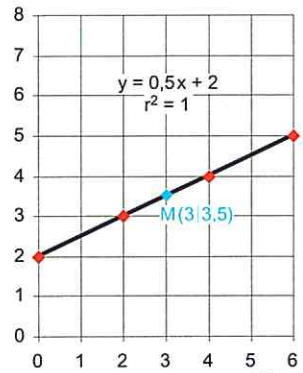


Fig. 1

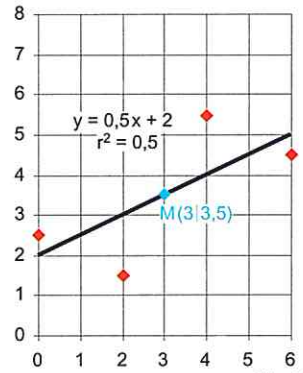


Fig. 2

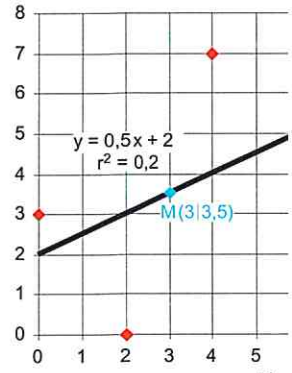


Fig.

Die empirische Regressionsgerade beschreibt, welche lineare Beziehung am besten zu einer gegebenen Punktwolke passt. Damit hat man aber noch keine Information, wie gut sie sich für praktische Vorhersagen eignet. Im Folgenden wird ein Maß für die Güte der Anpassung definiert.

Zerlegung der Varianz des Merkmals Y
Für die Varianz der Punktwolke um die zugehörige empirische Regressionsgerade mit der Steigung a gilt

$$V_D = V_Y - \frac{c_{XY}^2}{V_X} = V_Y - \frac{c_{XY}^2}{V_X^2} \cdot V_X = V_Y - a^2 \cdot V_X \text{ und damit } V_Y = V_D + a^2 \cdot V_X.$$

Die Varianz des Merkmals Y kann man also in eine Summe zerlegen, deren Summanden sich inhaltlich deuten lassen:

Fig.	V_X	a	$a^2 V_X$	V_D	V_Y	r^2	r
1	5	0,5	1,25	0	1,25	1	1
2	5	0,5	1,25	1,25	2,50	0,5	0,71
3	5	0,5	1,25	5	6,25	0,2	0,45

- Der Summand $a^2 \cdot V_X$ ist die Varianz, die das Merkmal Y besitzen würde, wenn alle Punkte der Wolke wie in Fig. 1 genau auf der Regressionsgeraden liegen würden. V_D hat dann nämlich den Wert 0 und es gilt $V_Y = 0 + a^2 \cdot V_X$.
  Da sich in diesem Fall die Varianz des Merkmals Y vollständig auf die Varianz des Merkmals X zurückführen lässt, heißt der Summand $a^2 \cdot V_X$ der (durch das lineare Modell) **„erklärbare"** Teil der **Varianz** des Merkmals Y.

- Wenn die Punkte um die Regressionsgerade streuen (Fig. 2 und 3), dann ist die Varianz V_D der Punktwolke um die Regressionsgerade ungleich null. Man deutet V_D als Varianz einer den strengen linearen Zusammenhang überlagernden Störgröße D und bezeichnet sie in diesem Zusammenhang als **„Störvarianz"**.

Regression:
Welche Steigung passt zur Wolke?
Korrelation:
Wie gut beschreibt die Steigung die Wolke?

Bestimmtheitsmaß

Die Punktwolken in Fig. 1 und 2 haben die gleiche Störvarianz $V_D = 5$.
Dennoch scheint die Wolke aus Fig. 1 weniger um die Regressionsgerade zu streuen als die aus Fig. 2.
Dies liegt an den unterschiedlichen Anteilen der Störvarianz V_D an der gesamten Varianz V_Y:
$\frac{V_D}{V_Y} = \frac{5}{25} = 20\,\%$ bei Fig. 1,
$\frac{V_D}{V_Y} = \frac{5}{6{,}25} = 80\,\%$ bei Fig. 2.

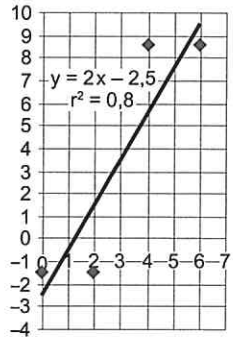


Fig. 1

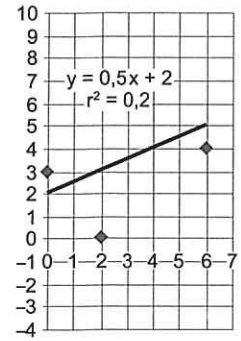
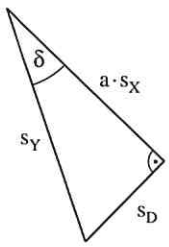


Fig. 2

„Eselsbrücke":
Im rechtwinkligen Dreieck der Standardabweichungen gilt $s_Y^2 = (a \cdot s_X)^2 + s_D^2$



Wegen $r = \frac{a \cdot s_X}{s_Y}$
kann man den Korrelationskoeffizienten deuten als Verhältnis von „erklärter" zu „gesamter" Standardabweichung des Merkmals Y.
Es gilt $r = \sin\delta$.

Es ist daher sinnvoll, die absoluten Größen der Stör- und erklärbaren Varianz im Verhältnis zur gesamten Varianz V_Y zu betrachten. Den Anteil der erklärbaren Varianz an der gesamten Varianz nennt man **Bestimmtheitsmaß** und bezeichnet ihn mit r^2. Somit ist $r^2 = \frac{a^2 \cdot V_X}{V_Y} = \frac{c_{XY}^2}{V_X^2} \cdot \frac{V_X}{V_Y} = \frac{c_{XY}^2}{V_X \cdot V_Y}$.
Die Zahl r selbst heißt **Korrelationskoeffizient**: $r = \frac{c_{XY}}{s_X \cdot s_Y}$.
Während das Bestimmtheitsmaß immer positiv ist, ist der Korrelationskoeffizient positiv, wenn die Regressionsgerade steigt, und negativ, wenn sie fällt (vgl. Beispiel 1). Der Grund dafür liegt im Vorzeichen von c_{XY}. Unter Verwendung von r kann man für die Störvarianz schreiben:
$V_D = V_Y - a^2 \cdot V_X = \left(1 - \frac{a^2 \cdot V_X}{V_Y}\right) \cdot V_Y = (1 - r^2) \cdot V_Y$. Der Korrelationskoeffizient liegt umso näher bei 1 oder –1, je weniger die Punktwolke um die Regressionsgerade streut. Im Fall $r = 1$ oder $r = -1$ hängen die Merkmale streng linear voneinander ab. Bei $0{,}8 \leqq |r|$ spricht man von **hoher**, bei $0{,}3 \leqq |r| < 0{,}8$ von **schwacher** Korrelation zwischen den Merkmalen. Bei noch kleineren Werten von $|r|$ nennt man die Merkmale „praktisch unkorreliert".

Für Punktwolken der Merkmalspaare (X; Y) gilt:

Korrelationskoeffizient:	Bestimmtheitsmaß:	Störvarianz:
$r = \frac{c_{XY}}{s_X \cdot s_Y}$, es gilt $-1 \leqq r \leqq 1$,	$r^2 = \frac{c_{XY}^2}{V_X \cdot V_Y}$, es gilt $0 \leqq r^2 \leqq 1$,	$V_D = (1 - r^2) \cdot V_Y$.

Beispiel 1: (Zerlegung der Varianz V_Y)
Gegeben ist die Punktwolke $P_1(0|11)$, $P_2(2|-1)$, $P_3(4|7)$, $P_4(6|-5)$ mit $V_X = 5$, $V_Y = 40$ und $c_{XY} = -10$. Die Regressionsgerade hat die Gleichung $y = -2x + 9$.
a) Bestimmen Sie unabhängig voneinander den durch das lineare Modell erklärbaren Teil der Varianz und die Störvarianz V_D.
b) Berechnen Sie das Bestimmtheitsmaß und den Korrelationskoeffizienten.

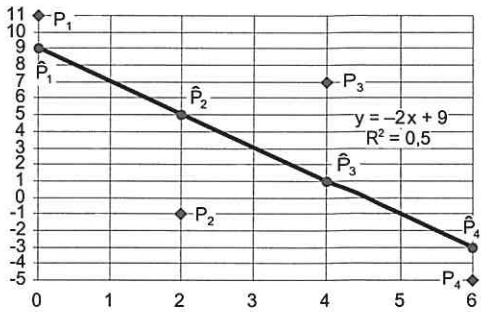


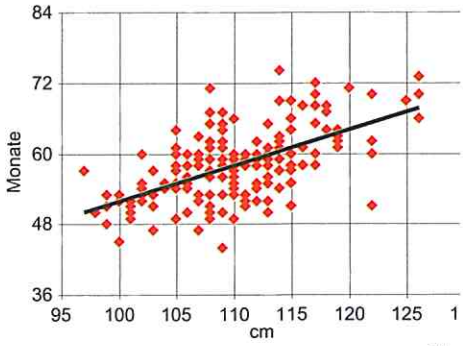
Fig. 3

Lösung:
a) Wenn die Punkte auf der Regressionsgeraden liegen würden ($\hat{P}_1$ bis $\hat{P}_4$), dann hätten sie die y-Koordinaten 9; 5; 1; –3 mit dem Mittelwert 3 und der Varianz 20. Also ist 20 der durch das lineare Modell erklärbare Teil von V_Y. (Die Formel $a^2 V_X = (-2)^2 \cdot 5$ liefert ebenso 20.)
Störvarianz: $V_D = \frac{1}{4}[2^2 + (-6)^2 + 6^2 + (-2)^2] = 20$. (Zusammen ergibt sich $20 + 20 = V_Y$.)
b) Da die Hälfte von V_Y „aufgeklärt" wird, gilt $r^2 = 0{,}5$. Dies ist das Quadrat des Korrelationskoeffizienten $r = \frac{c_{XY}}{s_X \cdot s_Y} = \frac{-10}{\sqrt{5} \cdot \sqrt{40}} = \frac{-1}{\sqrt{2}} \approx -0{,}707$.

Die Störvarianz V_D ermittelt man über die „vertikalen Abstandsquadrate" der Punkte von der Regressionsgeraden.

Die Urliste finden Sie in der Datei Vorschulkinder.xls.

Beispiel 2: (Alter schätzen)
Es ist schwer, aus der Körpergröße X (in cm) eines Kindes das Alter Y (in Monaten) zu bestimmen.
Eine Untersuchung von 166 Vorschulkindern durch das Gesundheitsamt Köln lieferte für die Kovarianz zwischen Körpergröße ($\bar{x} = 110{,}2$, $s_X = 5{,}85$) und Alter ($\bar{y} = 58$, $s_Y = 6{,}13$) den Wert 20,88.



Fig

Im Intervall $[\bar{y} - s_Y; \bar{y} + s_Y]$ mit einer Standardabweichung als „Radius" liegen nach einer Faustregel etwa 68 % aller Stichprobenwerte.

a) Wie groß sind Korrelation und Regression?
b) Ein Kind ist 130 cm groß. Welches Alter erwarten Sie?
c) In welchem Intervall wird das Alter mit 68%iger Wahrscheinlichkeit liegen?
d) Wie alt müsste ein Kind der Größe 50 cm (180 cm) sein?

Lösung:

Der Regressionskoeffizient a ist zugleich die Steigung der Regressionsgeraden (vgl. Seite 63).

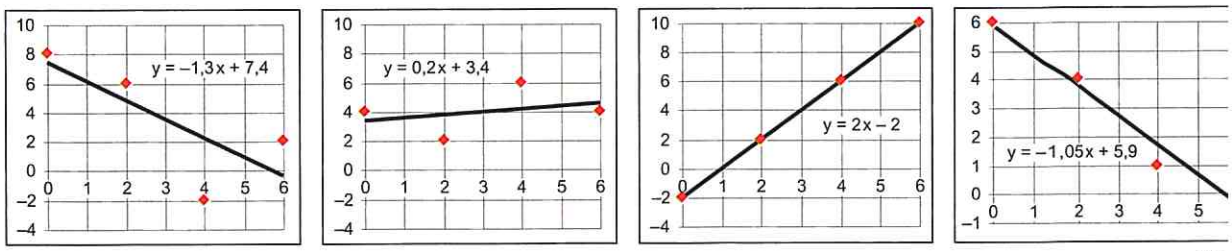
a) Es gilt $r = \frac{c_{XY}}{s_X \cdot s_Y} = \frac{20{,}88}{5{,}85 \cdot 6{,}13} \approx 0{,}58$ und $a = \frac{c_{XY}}{s_X^2} = \frac{20{,}88}{5{,}85^2} \approx 0{,}61$.

Einen Zentimeter größere Kinder sind im Mittel 0,61 Monate älter.
b) Die Regressionsgerade hat die Gleichung $y = 0{,}61\,(x - 110{,}2) + 58 \approx 0{,}61\,x - 9{,}24$.
Für $x = 130$ (cm) erwartet man das Alter $y \approx 70$ (Monate), also 5 Jahre und 10 Monate.
c) Für die Störvarianz erhält man $V_D = (1 - r^2)\,V_Y \approx 24{,}93$. Die Standardabweichung beträgt $s_D = \sqrt{V_D} \approx 5$ (Monate).
Mit 68 % Wahrscheinlichkeit wird das Alter des Kindes zwischen 5 Jahren, 5 Monaten und 6 Jahren, 3 Monaten liegen.
d) Für $x = 50$ (cm) erwartet man das Alter ≈ 21 (Monate), für $x = 180$ (cm) das Alter ≈ 101 (Monate), also 8 Jahre und 5 Monate. Beide Werte sind völlig unrealistisch, womit die Grenzen linearer Modelle für Prognosen deutlich werden.

Aufgaben

2 Zur Punktwolke $P_1(0|4)$, $P_2(2|7)$, $P_3(4|0)$, $P_4(6|3)$ gehört die Regressionsgerade mit der Gleichung $y = -0{,}5\,x + 5$.
a) Zeichnen Sie die Punkte und die Regressionsgerade in ein Koordinatensystem.
b) Berechnen Sie V_Y und zerlegen Sie V_Y unter Benutzung der Zeichnung in die Bestandteile „erklärbare Varianz" und Störvarianz V_D.
c) Berechnen Sie die Korrelation r und das Bestimmtheitsmaß r^2.

3 a) Ordnen Sie die Punktwolken „gefühlsmäßig" nach steigender Korrelation.



Fig

b) Kontrollieren Sie Ihre Antwort zu a) rechnerisch.

4 Die Noten in Deutsch (X) und Mathematik (Y) fielen in einer Stichprobe (Klasse 5) wie folgt aus: (3; 2), (2; 2), (3; 4), (2; 1), (2; 2), (2; 2), (3; 2), (4; 3), (2; 2), (2; 1), (4; 4). Berechnen Sie die Korrelation zwischen Deutschnote und Mathematiknote.

5 Am Ende einer Spielsaison bot die Fußball-Bundesliga das folgende Bild.

Sie können statt Fig. 1 natürlich auch die aktuelle Bundesliga-Tabelle auswerten.

	A	B	C	D	E	F	G	H	I	J	K	L
1			Heim		Auswärts				Tore			
2	Tabellenplatz		Sieg	Un.	Nied.	Sieg	Un.	Nied.	+	–	Punkte	
3		1	13	2	2	6	9	2	63	39	68	
4		2	11	4	2	8	5	4	69	37	66	
5		3	9	7	1	5	5	6	66	39	54	
6		4	10	5	2	4	5	8	55	49	52	
7		5	10	5	2	3	8	6	38	32	52	
8		6	9	3	5	5	6	6	54	46	51	
9		7	8	6	3	6	2	9	43	47	50	
10		8	6	6	5	5	5	7	43	44	44	
11		9	6	5	6	5	6	6	38	46	44	
12		10	8	4	5	3	6	8	57	55	43	
13		11	7	6	4	4	1	11	41	53	40	
14		12	8	4	5	3	4	10	41	49	41	
15		13	7	3	7	4	5	8	43	54	41	
16		14	7	4	6	4	2	11	38	54	39	
17		15	5	8	4	4	3	10	54	59	38	
18		16	6	5	6	3	6	8	48	60	38	
19		17	8	3	6	2	3	12	49	64	36	
20		18	7	5	5	1	3	13	43	56	32	
21	Mittelwert	9,50	8,06	4,72	4,22	4,17	4,67	8,06	49,06	49,06	46,06	
22	Standardabw.	5,19	1,96	1,48	1,75	1,57	2,03	2,76	9,63	8,42	9,54	
23	Kovarianzen	X									X	–46,36
24			X			X						1,27
25		X							X			–26,64
26									X	X		–20,84
27		X								X		38,25
28												

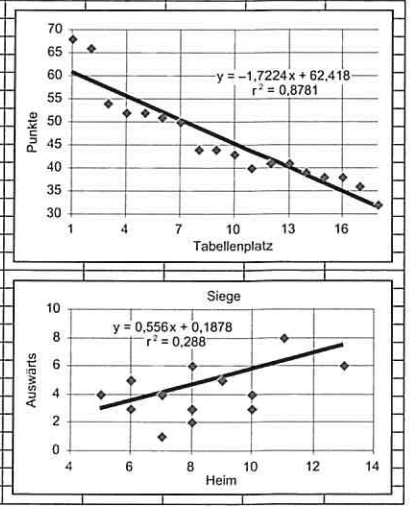


Fig. 1

Mittelwerte, Standardabweichungen und Kovarianzen interessierender Merkmalspaare sind bereits angegeben.

a) Berechnen Sie Regression und Korrelation zwischen dem Tabellenplatz und erreichter Punktzahl. Vergleichen Sie mit Fig. 1.

b) Berechnen Sie Regression und Korrelation zwischen den Anzahlen der Heim- und Auswärtssiege. Vergleichen Sie mit Fig. 2. Welche inhaltliche Bedeutung hat die Tatsache, dass der Regressionskoeffizient kleiner als 1 ist?

c) Berechnen Sie Regression und Korrelation zwischen Tabellenplatz und geschossenen (+) Toren (zwischen geschossenen (+) und kassierten (–) Toren).

Zeichnen Sie ein Punktdiagramm mit der dazu gehörenden Regressionsgeraden.

6 1955 publizierte R. DOLL eine Arbeit über Zigarettenkonsum und Lungenkrebs in 11 Ländern. Zeichnen Sie ein Punktdiagramm und berechnen Sie die Korrelation zwischen dem Merkmal X (Zigarettenverbrauch pro Kopf 1930) und Y (Todesfälle an Lungenkrebs 1950 je Million Einwohner).

Land	x_i	y_i	Land	x_i	y_i
Island	230	60	Kanada	500	150
Norwegen	250	90	Schweiz	510	250
Schweden	300	110	Finnland	1100	350
Dänemark	380	170	England	1100	460
Australien	480	180	USA	1300	200
Holland	490	240			

7 Für jedes der Jahre 1930 bis 1936 wurden in Oldenburg die Anzahl der Storchenpaare und die Einwohnerzahl ermittelt. Stellen Sie die Daten grafisch mit Regressionsgerade dar und berechnen Sie die Korrelation.

Also doch?

Jahr	1930	1931	1932	1933	1934	1935	1936
Storchenpaare	132	142	166	188	240	250	252
Einwohner	55400	55400	65000	67700	69800	72300	76000

8 Eine Untersuchung unter Abiturienten lieferte für die Kovarianz zwischen der Durchschnittsnote im Zeugnis der Klasse 5 ($\bar{x} = 2{,}53$, $s_X = 0{,}81$) und der Abiturnote ($\bar{y} = 2{,}65$, $s_Y = 0{,}62$) den Wert 0,1728.
a) Wie groß sind Korrelation und Regression?
b) Ein Abiturient hatte in der Klasse 5 den Zeugnisdurchschnitt 2,0. Welche Abiturnote erwarte Sie?
c) In welchem Intervall wird die Abiturnote mit 68%iger Wahrscheinlichkeit liegen?

9 Eine Untersuchung unter Vorschulkindern lieferte für die Kovarianz zwischen Körpergröße in cm ($\bar{x} = 110{,}2$, $s_X = 5{,}85$) und Gewicht in kg ($\bar{y} = 19{,}02$, $s_Y = 3{,}70$) den Wert $c_{XY} = 15{,}92$
a) Wie groß sind Korrelation und Regression?
b) Ein Vorschulkind ist 120 cm groß. Welches Körpergewicht erwarten Sie?
c) In welchem Intervall wird das Körpergewicht dieses Kindes mit 68%iger Wahrscheinlichkeit liegen?

Aufgabenteil a) sollten Sie nur bearbeiten, wenn Ihnen ein PC mit Tabellenkalkulation zur Verfügung steht.

10 In einer Tageszeitung fanden sich im Anzeigenteil folgende Angebote für gebrauchte VW-Golf (X = „Fahrleistung" in tausend Kilometer, Y = Preis in tausend Euro):
(58; 8,8), (32; 8,5), (70; 8,5), (50; 9,0), (54; 9,0), (84; 9,0), (63; 9,5), (97; 7,6), (69; 8,0), (29,5; 10,0), (40; 9,8), (54; 7,9), (53; 7,8), (60; 8,3), (44,5; 8,3), (120; 2,3), (126; 3,8), (160; 6,0), (166; 2,5), (130; 4,9), (120; 2,7), (103; 4,9), (120; 2,7), (103; 6,0), (87; 6,5), (79; 3,0), (85; 7,0), (230; 1,9), (110; 7,0).
a) Kontrollieren Sie: $\bar{x} = 87{,}9$; $s_X = 45{,}7$; $\bar{y} = 6{,}8$; $s_Y = 2{,}5$; $c_{XY} = -92$.
b) Bestimmen Sie die Gleichung der Regressionsgeraden und den Korrelationskoeffizienten.
c) Welche inhaltliche Bedeutung könnte man den Schnittpunkten der Regressionsgeraden mit den Koordinatenachsen zuschreiben?
d) Ein Gebrauchtwagen ist 100 000 km gelaufen. Welchen Preis erwarten Sie?
In welchem Intervall wird der Preis mit einer Wahrscheinlichkeit von ca. 68 % liegen?

Ggf. sollten Sie eine eigene Untersuchung durchführen (PKW-Baujahr, Preis; Preis und Fläche angebotener Wohnungen usw.).

Kausalität

11 Mitunter wird eine hohe Korrelation als Indiz gedeutet für eine kausale Beziehung zwische zwei Merkmalen in dem Sinne, dass hohe Merkmalswerte von X auch hohe (niedrige) Werte voı Y „verursachen". Vermuten Sie zwischen folgenden Merkmalen positive bzw. negative Korrelation? Liegt Ihres Wissens ein kausaler Zusammenhang vor? Finden Sie weitere Beispiele.

		Korrelation: pos./neg.	Kausalität: j/n
Autos in einer Stadt	verkaufte Benzinmenge		
Ausbildungsdauer	Jahreseinkommen		
Berufstätigkeit der Eltern	Fernsehkonsum der Kinder		
Duschgelkonsum	Ausgaben für Kleidung		
Alkoholkonsum	Tabakkonsum		
Freizeit	Einkommen		
Bierkonsum	mittlere Tagestemperatur		
Alter des Ehemannes	Alter der Ehefrau		
Anzahl der Störche je km²	Bevölkerungszahl je km²		

Vergleich Kovarianz – Korrelation

12 a) Dagobert Duck übt sich im „Geldbeutel-Weitwurf". Seine Ergebnisse waren heute: (3 kg; 8 m), (5 kg; 7 m), (7 kg; 3 m). Berechnen Sie Kovarianz und Korrelation zwischen Gewicht und Wurfweite.
b) Die gleichen Stichprobenwerte werden in anderen Einheiten gemessen: (3 kg; 800 cm), (5 kg; 700 cm), (7 kg; 300 cm) bzw. (3000 g; 800 cm), (5000 g; 700 cm), (7000 g; 300 cm). Berechnen Sie erneut Kovarianz und Korrelation.
c) Kommentieren Sie Ihre Erkenntnisse zum unterschiedlichen Verhalten von Kovarianz und Korrelation beim Wechsel der Maßskalen.

13 Frau Görgen sucht eine Wolke aus drei oder vier Punkten, bei der die Regression positiv, aber die Korrelation den Wert Null hat. Helfen Sie ihr.

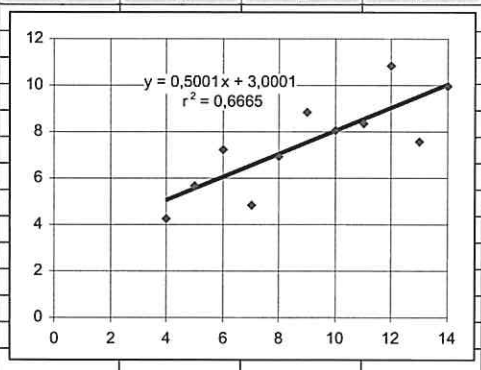
14 Das Bestimmtheitsmaß r^2 hängt von der Steigung a der Regressionsgeraden und der Störvarianz V_D ab: $r^2 = \frac{1}{1 + \frac{V_D}{a^2 \cdot V_X}}$. Wie ändert sich das Bestimmtheitsmaß r^2

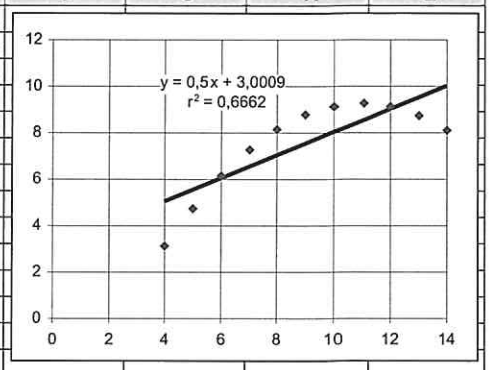
a) bei konstanter Störvarianz V_D und wachsender Steigung a,
b) bei konstanter Steigung a und wachsender Störvarianz V_D?
c) Begründen Sie die angegebene Formel durch eine Termumformung.

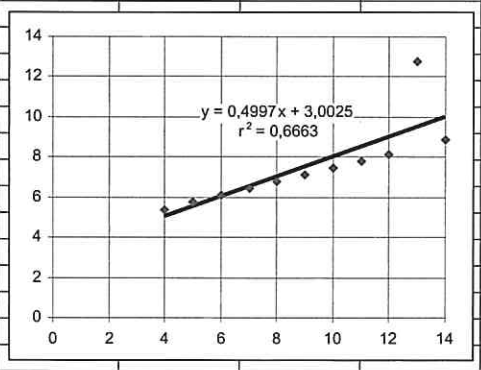
15 Die Erhebung zweier Merkmale X und Y lieferte (2; 4), (3; 5) und (4; 9).
a) Berechnen Sie Korrelation und Regression.
b) Welche Werte erhält man für Korrelation und Regression, wenn man die Merkmale vertauscht, also die Paare (4; 2), (5; 3) und (9; 4) untersucht?
c) Kontrollieren und beweisen Sie allgemein: Wenn man die beiden Regressionskoeffizienten aus a) und b) multipliziert, erhält man das Bestimmtheitsmaß.

16 Fig. 1 zeigt Punktwolken sehr unterschiedlicher Gestalt, die alle (nahezu) die gleiche Regressionsgerade ($y = 0{,}5\,x + 3$) und den gleichen Korrelationskoeffizienten $r = 0{,}816$ besitzen. Kommentieren Sie.

	A	B	G	H
1	10	8,04	10	9,14
2	8	6,95	8	8,14
3	13	7,58	13	8,74
4	9	8,81	9	8,77
5	11	8,33	11	9,26
6	14	9,96	14	8,1
7	6	7,24	6	6,13
8	4	4,26	4	3,1
9	12	10,84	12	9,13
10	7	4,82	7	7,26
11	5	5,68	5	4,74
12				
13				
14				
15				
16				
17	10	7,46	8	6,58
18	8	6,77	8	5,76
19	13	12,74	8	7,71
20	9	7,11	8	8,84
21	11	7,81	8	8,47
22	14	8,84	8	7,04
23	6	6,08	8	5,25
24	4	5,39	19	12,5
25	12	8,15	8	5,56
26	7	6,42	8	7,91
27	5	5,73	8	6,89
28				
29				
30				
31				
32				







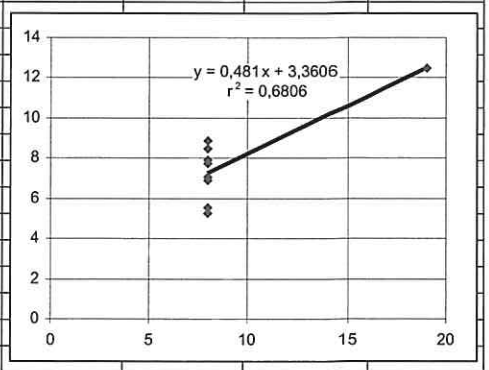


Fig. 1

(Quelle: Amer. Statistican 27, 1973)

8 Wirklichkeit erkennen – spannende Experimente mit dem Computer

Regression-Korrelation.xls

Nützliches Kalkulationsblatt

Bei folgenden Aufgaben und Fragestellungen der Regressions- und Korrelationsrechnung erweist sich ein „universell verwendbares" Kalkulationsblatt als sehr hilfreich.

	A	B	C	D
1	Regression-Korrelation.xls			
2				
3				
4	Mittelwerte		sX	2,278
5	4,75	3,3	sY	1,603
6	X	Y	cXY	2,375
7	6	2,2	a	0,458
8	5	3	b	1,125
9	7	6	r	0,650
10	1	2	sD	1,218
11				
12				
13				
14				
15				

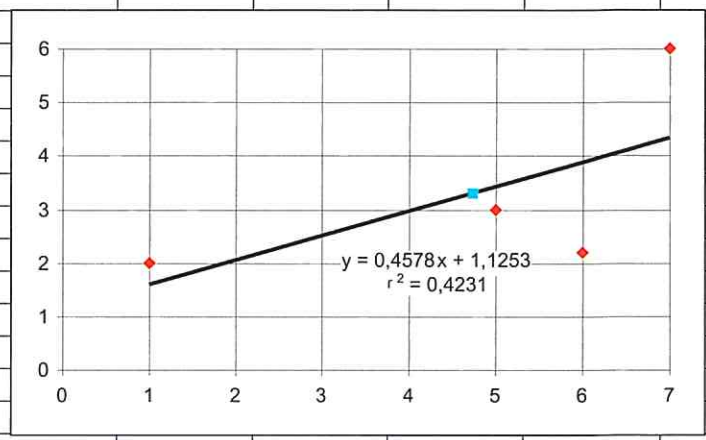


Fig.

Hinweise zur Verwendung

- Die (X; Y)-Merkmalspaare werden im Bereich A7:B100 eingegeben.

Automatisch werden bereitgestellt

- die Mittelwerte $\bar{x}$ und $\bar{y}$ in A5 und B5, berechnet mit den Formeln =MITTELWERT(A7:A100) bzw. =MITTELWERT(B7:B100),
- das Punktdiagramm mit dem Mittelpunkt M($\bar{x}|\bar{y}$) und der empirischen Regressionsgeraden samt Gleichung und Bestimmtheitsmaß r^2,
- die empirischen Standardabweichungen s_X und s_Y in D4 und D5, berechnet mit den Formeln =STABWN(A7:A100) bzw. =STABWN(B7:B100),
- die empirische Kovarianz c_{XY} in D6, berechnet mit der Formel =KOVAR(A7:A100;B7:B100),
- der empirische Regressionskoeffizient a in D7, berechnet mit der Formel =STEIGUNG(B7:B100;A7:A100) oder mit =D6/D4^2,
- der y-Achsenabschnitt b der empirischen Regressionsgeraden in D8, berechnet mit der Form =ACHSENABSCHNITT(B7:B100;A7:A100) oder mit =B5–D7*A5,
- der empirische Korrelationskoeffizient r in D9, berechnet mit der Formel =KORREL(A7:A100;B7:B100) oder mit =D6/(D4*D5),
- die empirische Standardabweichung s_D der Störgröße D in D10, berechnet mit der Formel =WURZEL(1–D9^2)*D5 oder =WURZEL(D5^2–D7^2*D4^2).

Wenn keine Merkmalspaare eingegeben sind oder mehrere Paare mit gleichen x-Koordinaten und verschiedenen y-Koordinaten vorliegen, dann sind manche Berechnungen sinnlos. EXCEL reagiert mit entsprechenden Meldungen in den Zellen.

Aufgaben

1 Interpretieren

Beschreiben und interpretieren Sie in eigenen Worten die Ausgaben des Rechenblattes, wenn Sie schrittweise folgende Merkmalspaare eingeben:

a) (–2; 9), (1; 3), (3; –1), (10; –15),

b) (2; 3), (2; 4), (2; 5), (4; 5), (4; 4), (4; 3),

c) (1; 2), (2; 4), (3; 6), (1; 4), (2; 6), (3; 8).

Kontrollieren Sie, ob die in die Grafik eingeblendete Regressionsgleichung und das Bestimmtheitsmaß mit den in der Tabelle berechneten Kenngrößen übereinstimmen.

2 Experiment „Springen große (schwere) Leute höher?“

Statt der Körpergröße bietet sich als Merkmal X auch das Körpergewicht an. Die Angabe des Körpergewichtes muss freiwillig geschehen.

Um die Varianz des Merkmals Y zu erhöhen, bietet es sich an, eine Klasse 6 zur Teilnahme am Experiment einzuladen.

Untersuchen Sie den Zusammenhang zwischen Körpergröße und Hochsprungleistung.

Versuchsanleitung: Jeder Versuchsteilnehmer markiert mit ausgestrecktem Arm seine „Ausgangshöhe“ an der Tafel oder auf einem an der Wand befestigten Tapetenstreifen. Anschließend springt jeder aus dem Stand möglichst hoch und versucht dabei, eine Markierung in maximaler Höhe anzubringen. Die Differenz der Markierungen gilt als Sprunghöhe.

a) Erstellen Sie in einem gemeinsamen Experiment eine Urliste aus den Merkmalspaaren (X = Körpergröße, Y = Sprunghöhe).

b) Berechnen Sie Regression, Korrelation und Standardabweichung s_D der Störgröße D. Zeichnen Sie ein Punktdiagramm mit der zugehörigen Regressionsgeraden. Beantworten Sie die in der Überschrift formulierte Frage.

3 Springen gute Sprinter weiter?

Bei den Bundesjugendspielen werden Daten erhoben in den Disziplinen Sprint, Weitsprung und Ballwurf (200 g). Besorgen Sie sich die (anonymen) Ergebnisse einer Klasse und untersuchen Sie Regression und Korrelation zwischen den Merkmalen

a) Zeit beim Sprint und Sprungweite,

b) Zeit beim Sprint und Wurfweite.

Eine andere Urliste enthält Bundesjugendspiele.xls.

c) Wenn Sie über einen Internet-Zugang verfügen: Laden Sie die (von EXCEL lesbaren) Daten einer Klasse 5 (http://www.learn-line.nrw.de/Themen/EDA/medio/bjsp/bjsphome.htm). Berechnen Sie für diese Daten Regression und Korrelation und vergleichen Sie mit den Ergebnissen Ihrer eigenen Untersuchung.

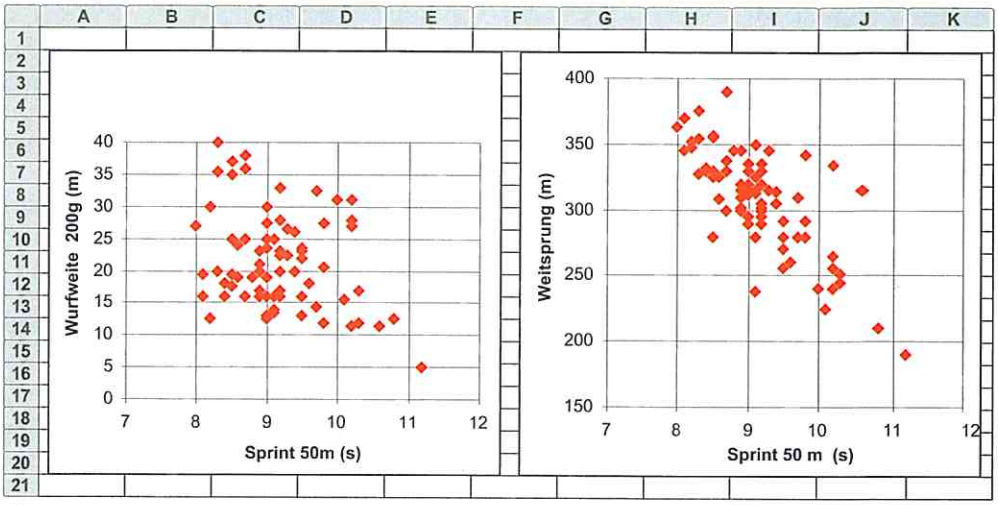


Fig. 1

Wenn Sie nicht selber messen können, nutzen Sie die Daten aus Bundweite.xls.

4 Experiment „Kleidergröße“

a) Messen Sie in einem Kaufhaus bei 10 Damenhosen/Herrenhosen verschiedener Größen die Bundweite. Erstellen Sie eine gemeinsame Urliste aus den Merkmalspaaren Größe – Bundweite.
b) Berechnen Sie die Regression a, die Korrelation r und die Standardabweichung s_D der Störgröße D.
c) Zeichnen Sie ein Punktdiagramm mit der zugehörigen Regressionsgeraden.
d) Gibt es bei den von Ihnen untersuchten Merkmalen eine ähnlich ausgeprägte lineare Abhängigkeit wie zwischen Schuhgröße und Innenlänge?

Beispiel: Otto verwendet $\alpha = 2$, $\beta = -3$. Er würfelt 4, sein zweiter Wurf zeigt 5 ($D = +1$). Das Merkmalspaar lautet $(4; 2 \cdot 4 - 3 + 1) = (4; 6)$.

Vorschlag: Otto würfelt vor der Tür und ruft seine Ergebnisse in die Klasse. Sie arbeiten an Ihren Computern oder projizieren die Ergebnisse eines Computers an die Wand.

5 Lineare Modelle raten I (reales Experiment)

Versuchsanleitung:
Bestimmen Sie eine Versuchsperson („Otto“), die sich eine ganze Zahl α zwischen -4 und $+4$ und eine ganze Zahl β zwischen -10 und $+10$ ausdenkt, aber nicht verrät.
Otto „produziert“ mit einem Würfel Merkmalspaare nach dem geheim gehaltenen linearen Modell $Y = \alpha X + \beta + D$:

- Die x-Koordinate zwischen 1 und 6 wird gewürfelt,
- die y-Koordinate ergibt sich aus $y = \alpha x + \beta$,
- zusätzlich (nochmals würfeln) wird die Störgröße D addiert:
  bei den Augenzahlen 1 oder 2 wird addiert $D = -1$;
  bei den Augenzahlen 3 oder 4 wird addiert $D = 0$;
  bei den Augenzahlen 5 oder 6 wird addiert $D = +1$.

a) Geben Sie die von Otto bekannt gegebenen Merkmalspaare schrittweise so lange in das Kalkulationsblatt „Regression-Korrelation.xls“ ein, bis Sie „hinreichend sicher wissen“, welche „geheimen Parameter“ α und β Otto verwendet.
b) Wiederholen Sie den Versuch mehrfach und tabellieren Sie Ottos „wahre“ Werte α, β, die geschätzten Größen a, b, r und die benötigte Schrittzahl n.

Möglicher Anfang der Tabelle

α	β	a	b	r	n
2	4	2,01	3,6	0,92	21

c) Julia glaubt: „Die empirische Korrelation r zwischen X und Y hängt von der verwendeten Steigung α ab, nicht aber von β.“
Bewerten Sie diese Aussage mithilfe Ihrer Tabelle.
d) Beate findet die Aussage „Je mehr Versuche desto genauer wird es“ einleuchtend. Sie folgert, dass bei hohem n der Korrelationskoeffizient r in der Nähe von ± 1 liegt.
Nehmen Sie Stellung.

X	Y	Wk
1	0	$\frac{1}{18}$
1	1	$\frac{1}{18}$
1	2	$\frac{1}{18}$
2	2	$\frac{1}{18}$
2	3	$\frac{1}{18}$
2	4	$\frac{1}{18}$
3	4	$\frac{1}{18}$
3	5	$\frac{1}{18}$
3	6	$\frac{1}{18}$
4	6	$\frac{1}{18}$
4	7	$\frac{1}{18}$
4	8	$\frac{1}{18}$
5	8	$\frac{1}{18}$
5	9	$\frac{1}{18}$
5	10	$\frac{1}{18}$
6	10	$\frac{1}{18}$
6	11	$\frac{1}{18}$
6	12	$\frac{1}{18}$

Y	Wk
0	$\frac{1}{18}$
1	$\frac{1}{18}$
2	$\frac{2}{18}$
3	$\frac{1}{18}$
4	$\frac{2}{18}$
5	$\frac{1}{18}$
6	$\frac{2}{18}$
7	$\frac{1}{18}$
8	$\frac{2}{18}$
9	$\frac{1}{18}$
10	$\frac{2}{18}$
11	$\frac{1}{18}$
12	$\frac{1}{18}$

D	Wk
-1	$\frac{1}{3}$
0	$\frac{1}{3}$
$+1$	$\frac{1}{3}$

Fig. 1

6 Berechnung der theoretischen Korrelation ϱ („Theorie“ zu Aufgabe 5)

Bei Simulationen kann man den theoretischen Korrelationskoeffizienten ϱ berechnen, da man die Wahrscheinlichkeiten der Merkmalspaare kennt.
Im Experiment der Aufgabe 5 treten 18 Merkmalspaare jeweils mit der Wahrscheinlichkeit $\frac{1}{18}$ auf. Fig. 1 zeigt ein Beispiel für $\alpha = 2$ und $\beta = -1$.
a) Berechnen Sie für dieses Beispiel die Varianzen σ_X^2, σ_Y^2, σ_D^2, die theoretische Kovarianz $\gamma_{X,Y}$ und zeigen Sie, dass der theoretische Korrelationskoeffizient $\varrho = \frac{\gamma_{X,Y}}{\sigma_X \cdot \sigma_Y}$ den Wert $\sqrt{\frac{35}{37}} \approx 0{,}973$ hat.
b) Kontrollieren Sie durch eine Rechnung, dass der aus Kovarianz und Varianz berechnete theoretische Regressionskoeffizient $\frac{\gamma_{X,Y}}{\sigma_X^2}$ den Wert 2 hat, also tatsächlich mit dem Steigungsfaktor des linearen Modells übereinstimmt.
c) Zeigen Sie durch eine Rechnung, dass man auch die theoretische Varianz des Merkmals Y zerlegen kann in einen „erklärbaren“ Teil und die Varianz der Störgröße D: $\sigma_Y^2 = \alpha^2 \cdot \sigma_X^2 + \sigma_D^2$
d) Berechnen Sie die theoretische Korrelation ϱ für $\alpha = -2$ und $\beta = 0$.

7 Lineare Modelle raten II (mit EXCEL simuliertes Experiment)

Hintergrundinformation zu Regression-Korrelation-Simulation-step.xls:
Das Merkmal X, das die ganzen Zahlen 1, 2, …, k je mit Wahrscheinlichkeit $\frac{1}{k}$ annimmt, hat den Erwartungswert
$\mu_X = \frac{(k+1)}{2}$
und die Standardabweichung
$\sigma_X = \sqrt{\frac{k^2-1}{12}}$.
Für die Störgröße D, die die ganzen Zahlen zwischen –m und +m je mit Wahrscheinlichkeit $\frac{1}{2m+1}$ annimmt, gilt $\mu_D = 0$ und
$\sigma_D = \sqrt{\frac{m^2+m}{3}}$.
Für die theoretische Kovarianz gilt
$\varrho_{XY} = \alpha \cdot \sigma_X^2$.

	A	B	C	D	E	F
1	Regression-Korrelation-simulation-step.xls					
2	x zwischen 1 und k				k	6
3	d zwischen –m und m				m	1
4	Mittelwerte		sX	1,54	sigmaX	1,71
5	3,80	6,70	sY	3,26	sigmaY	3,51
6	X	Y	sXY	4,84	gammaXY	5,83
7	3	6	a	2,0508	alpha	2,00
8	6	10	b	–1,0932	beta	–1
9	6	11	r	0,97	rho	0,97
10	4	8	sD	0,83	sigmaD	0,82
11	1	1	n	10		
12	3	4				
13	4	7				
14	2	2				
15	5	10				
16	4	8				
17						
18						

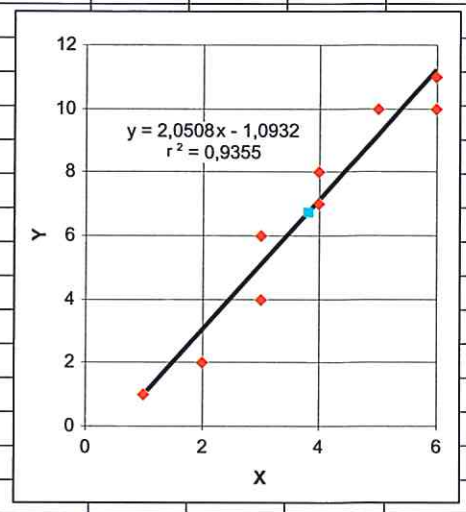


Fig. 1

Das Kalkulationsblatt Regression-Korrelation-Simulation-step.xls erzeugt, wenn man im Menüpunkt Extra das Makro „step" anklickt, zufällige (X; Y)-Merkmalspaare.

- Die x-Koordinaten liegen zwischen 1 und k, wobei k in der Zelle F2 festgelegt wird.
- Die y-Koordinaten ergeben sich aus $y = \alpha x + \beta + d$, wobei α und β in den Zellen F7 und F8 stehen und d eine Zufallszahl zwischen –m und +m ist. m wird in Zelle F3 festgelegt. (Für $k = 6$ und $m = 1$ erhält man „Ottos Versuchsanordnung" aus Aufgabe 5.)

a) Machen Sie sich mit der Funktionsweise des Kalkulationsblattes vertraut, indem Sie die Werte von Aufgabe 6 eingeben ($\alpha = 2$, $\beta = -1$, $k = 6$, $m = 1$) und kontrollieren, dass sich die aus den Merkmalspaaren geschätzten Parameter a und b schrittweise den Modellparametern α und β nähern.

b) Untersuchen Sie experimentell, wie sich eine Vergrößerung von n und damit eine Vergrößerung von σ_D auf den empirischen Korrelationskoeffizienten r auswirkt.

c) Untersuchen Sie experimentell den Einfluss von α auf die empirischen Korrelationskoeffizienten r.

d) Für den theoretischen Korrelationskoeffizienten gilt (analog zum empirischen)

$\varrho^2 = \frac{\alpha^2\sigma_X^2}{\sigma_Y^2} = \frac{\alpha^2\sigma_X^2}{\alpha^2\sigma_X^2 + \sigma_D^2} = \frac{1}{1 + \frac{\sigma_D^2}{\alpha^2\sigma_X^2}}$. Begründen Sie die Umformungsschritte.

Für einen kleinen Stichprobenumfang n ist die aus der Punktwolke geschätzte Standardabweichung s_D meist kleiner als der theoretische Wert σ_D, da die empirische Regressionsgerade besser zu der Punktwolke passt als die theoretische. Besonders deutlich wird das für $n = 2$.
Die empirische Regressionsgerade verläuft im Gegensatz zur theoretischen stets genau durch die beiden Datenpunkte. Für einen größeren Stichprobenumfang n werden die Abweichungen unbedeutend.

Welchen Einfluss hat eine Veränderung von σ_D (von α) auf den theoretischen Korrelationskoeffizienten? Stellen Sie eine Beziehung zu Ihren Untersuchungsergebnissen aus den Aufgabenteilen b) und c) her.

e) **„Wirklichkeit erkennen"**

In den Zellen F4, F5, F6, F9 und F10 werden aus α, β, k und m die „wirklichen" (theoretischen) Größen σ_X, σ_Y, γ_{XY}, ϱ und σ_D berechnet (vgl. die Formeln auf dem Rand). Kontrollieren Sie experimentell, dass sich die empirischen Größen aus Spalte D mit wachsendem Versuchsumfang n diesen „wirklichen" Größen in Spalte F nähern.

Varianz und Standardabweichung

Bei einer statistischen Erhebung werden in einer Stichprobe bestimmte Merkmale untersucht.

Sind $x_1, x_2, \ldots, x_n$ Stichprobenwerte eines Merkmals X mit dem Mittelwert $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$, dann nennt man

$V_X = \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$

die **Varianz** von X.

Die Quadratwurzel aus der Varianz V_X wird **Standardabweichung** s_X genannt: $s_X = \sqrt{V_X}$.

Nach einer Faustregel liegen ca. 68 % aller Daten in dem „Standard-Abweichungs-Intervall" um den Mittelwert.

Kommen bei einer Häufigkeitsverteilung die Merkmalsausprägungen $x_1, x_2, x_3, \ldots, x_r$ mit den relativen Häufigkeiten $h_1, h_2, h_3, \ldots, h_r$ vor, so erhält man den Mittelwert einer **Häufigkeitsverteilung** durch

$\bar{x} = x_1 \cdot h_1 + x_2 \cdot h_2 + x_3 \cdot h_3 + \ldots + x_r \cdot h_r$.

Für die Standardabweichung gilt:

$s_X = \sqrt{V_X} = \sqrt{(x_1 - \bar{x})^2 \cdot h_1 + (x_2 - \bar{x})^2 \cdot h_2 + \ldots + (x_r - \bar{x})^2 \cdot h_r}$.

Regression

In der Regressionsrechnung untersucht man Abhängigkeiten zwischen zwei Merkmalen X und Y. Dazu macht man die Annahme, dass zwischen X und Y ein linearer Zusammenhang der Form $Y = \alpha X + \beta$ besteht, der nur von einer Größe D „gestört" wird.

Sind $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ die Stichprobenwertepaare zweier Merkmale X, Y mit den Mittelwerten $\bar{x}$ und $\bar{y}$, so wird die theoretische Regressionsgerade $g: y = \alpha x + \beta$ geschätzt durch die **empirische Regressionsgerade (Ausgleichsgerade)** mit der Gleichung

$y = a(x - \bar{x}) + \bar{y} = ax + b$.

a heißt empirischer **Regressionskoeffizient**.

Es gilt: $a = \frac{c_{XY}}{V_X} = \frac{c_{XY}}{s_X^2}$, wobei

$c_{XY} = \frac{1}{n}((x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y}))$

die empirische **Kovarianz** zwischen X und Y ist.

Kommen die r · s Paare von Merkmalsausprägungen $(x_1; y_1), (x_1; y_2), \ldots, (x_r; y_s)$ mit den relativen Häufigkeiten $h_{11}, h_{12}, \ldots, h_{rs}$ vor, so gilt

$c_{XY} = (x_1 - \bar{x})(y_1 - \bar{y}) \cdot h_{11} + (x_1 - \bar{x})(y_2 - \bar{y}) \cdot h_{12} + \ldots + (x_r - \bar{x})(y_s - \bar{y}) \cdot h_{rs}$.

Korrelation

Eine Punktwolke mit dem Mittelpunkt $M(\bar{x}|\bar{y})$ „streut" am wenigsten um ihre Regressionsgerade.

Der **Korrelationskoeffizient** r misst, wie gut die Regressionsgerade den Zusammenhang zwischen den Merkmalen X und Y beschreibt.

Es gilt $r = \frac{c_{XY}}{s_X \cdot s_Y}$.

Je näher der Korrelationskoeffizient bei ±1 liegt, desto besser eignet sich die Regressionsgerade für Vorhersagen.

Aus der Varianz des Merkmals Y und dem Korrelationskoeffizienten r kann man die **Varianz der „Störgröße" D** und damit die Genauigkeit der Vorhersage abschätzen. Es gilt $V_D = (1 - r^2) \cdot V_Y$.

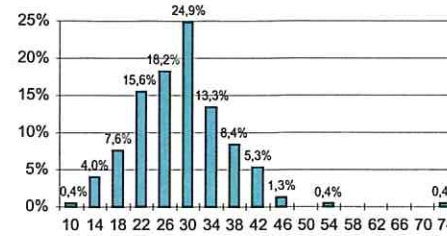
Beispiel:

Merkmal X: Punkte bei einem Eignungstest, Urliste: $x_1 = 36$, $x_2 = 38, \ldots, x_{225} = 23$,

$\bar{x} = \frac{1}{225}(36 + 38 + \ldots + 23) = 29{,}2$,

$s_X = \frac{1}{225}((36 - \bar{x})^2 + \ldots + (23 - \bar{x})^2) = 7{,}9$.

Verteilung der relativen Häufigkeiten:

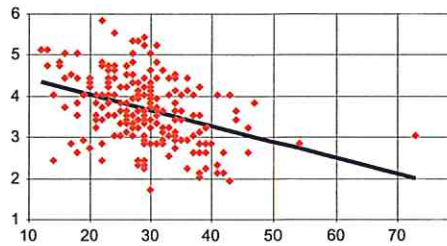


Im Intervall $[\bar{x} - s_X; \bar{x} + s_X] = [21{,}3; 37{,}1]$ liegen $\frac{161}{225} = 71{,}6\,\%$ aller Stichprobenwerte

Merkmal Y: Note beim Abschlusstest, Urliste: $y_1 = 4{,}1$; $y_2 = 2{,}8; \ldots; y_{225} = 4{,}6$,

$\bar{y} = 3{,}7$, $s_Y = 0{,}8$.

(X; Y)-Punktdiagramm:



Kovarianz:

$c_{XY} = \frac{1}{225}((36 - \bar{x})(4{,}1 - \bar{y}) + \ldots + (23 - \bar{x})(4{,}6 - \bar{y}) = -2{,}4$,

Regression:

$a = \frac{c_{XY}}{s_X^2} \approx \frac{-2{,}4}{7{,}9^2} \approx -0{,}038$,

Regressionsgerade:

$y = a(x - \bar{x}) + \bar{y} \approx -0{,}038x + 4{,}8$,

Korrelation:

$r = \frac{c_{XY}}{s_X \cdot s_Y} \approx \frac{-2{,}4}{7{,}9 \cdot 0{,}8} \approx -0{,}37$,

Störgröße:

$s_D = \sqrt{1 - r^2} \cdot s_Y \approx 0{,}74$,

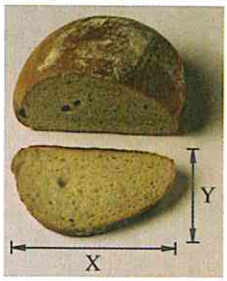
Vorhersage:

Für das Ergebnis $x = 60$ beim Eignungstest erwartet man für den Abschlusstest $y = 2{,}52$. Mit ca. 68%iger Sicherheit gilt $1{,}78 < y < 3{,}26$.

1 In einer Klassenarbeit ergaben sich folgende Noten:
2, 5, 5, 4, 3, 2, 3, 1, 4, 2, 2, 4, 1, 3, 2, 3, 4, 1, 2, 6.
a) Bestimmen Sie den Mittelwert $\bar{x}$ und die Standardabweichung s_x.
b) Ermitteln Sie die Verteilung der relativen Häufigkeiten; zeichnen Sie ein Säulendiagramm.

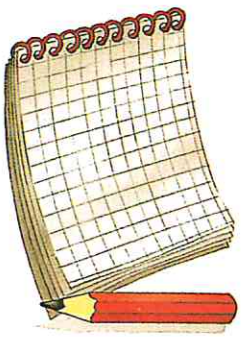
2 Berechnen Sie zur Punktwolke $P_1(2|4)$, $P_2(4|5)$, $P_3(4|7)$, $P_4(6|8)$ den Regressions- und den Korrelationskoeffizienten. Zerlegen Sie die Varianz V_Y in den erklärbaren Teil und die Störvarianz V_D.

3 Die Größe X von Fahrrädern (d. h. der Durchmesser des Laufrades) wird in Zoll (") angegeben. Jan hat bei verschieden großen Fahrrädern den Umfang Y (in cm) der Gummireifen gemessen. Er erhält die folgenden Wertepaare.
(28, 219), (28, 218), (24, 192), (24, 191), (16, 134), (16, 135).
a) Ermitteln Sie die Gleichung der empirischen Regressionsgerade und den Korrelationskoeffizienten.
b) Zeichnen Sie ein Punktdiagramm mitsamt der Regressionsgeraden.
c) Ein Zoll hat die Länge 2,54 cm. Welche Gleichung müsste die Regressionsgerade haben, wenn man die Beziehung $U = \pi \cdot d$ benutzt? Wie erklären Sie sich, dass die empirische Regressionsgerade nicht durch den Ursprung geht?



4 Pit hat einen Laib Brot in Scheiben schneiden lassen und bei jeder Scheibe die Breite X (in cm) und die Höhe Y (in cm) gemessen:
(8,5; 5,2), (13,7; 8,3), (15; 9,4), (16; 9,5), (16,9; 9,6), (15,1; 9,4), (13,9; 8,8), (12; 7,3).
a) Bestimmen Sie die Gleichung der empirischen Regressionsgeraden.
b) Berechnen Sie die Korrelation r und die Standardabweichung s_D der Störgröße D.
c) Eine Scheibe hat die Breite $X = 10$ cm.
In welchem Intervall liegt die Höhe Y mit der Wahrscheinlichkeit 68 %?

5 Meko hat bei Holzkugeln den Durchmesser X (in cm) und das Gewicht Y (in g) gemessen:
(2; 2), (2,3; 4), (3; 9), (3,9; 22), (4,9; 44).
a) Bestimmen Sie die Gleichung der empirischen Regressionsgeraden und den Korrelationskoeffizienten.
b) Zeichnen Sie ein Punktdiagramm mitsamt der Regressionsgeraden.
c) Warum ist ein lineares Modell zur Beschreibung des Zusammenhanges zwischen Durchmesser und Gewicht nicht sinnvoll?



6 Ulla hat bei einer statistischen Untersuchung eine geringe Regression, aber eine hohe Korrelation zwischen den Merkmalen X und Y festgestellt, Doris bei einer anderen Untersuchung eine hohe Regression und eine geringe Korrelation.
a) Wie könnten die zugehörigen Punktdiagramme prinzipiell aussehen?
b) In welchem der beiden Fälle ist das lineare Modell besser zur Prognose geeignet?

7 Bei einer empirischen Untersuchung zweier Merkmale X und Y ergab sich $s_X = 3$ und die Regressionsgerade mit der Gleichung $y = 2{,}4x - 25$.
Bei einem Merkmalsträger ergab sich der Wert $x \approx 20$.
Bestimmen Sie das Intervall, in dem der zugehörige y-Wert mit 68%iger Wahrscheinlichkeit liegt, wenn für die Korrelation gilt
a) $r = 0{,}99$ b) $r = 0{,}9$ c) $r = 0{,}5$ d) $r = 0{,}1$.
Tipp: Nutzen Sie die Beziehung $V_D = \left(\frac{1}{r^2} - 1\right) a^2 V_X$, die Sie aus $V_D = (1 - r^2) V_Y$ herleiten können.

Die Lösungen zu den Aufgaben dieser Seite finden Sie auf Seite 171.