7 Statistik

Im Allgemeinen haftet der Mathematik der Nimbus der untadeligen Korrektheit an. Wenn es aber um das Thema Statistik geht, ist der gute Ruf dahin. Wir kennen die Grafiken, die bewusst einen falschen Eindruck vermitteln, Prognosen vor Wahlen, die weit daneben gelegen sind, und Sprüche wie „Trau keiner Statistik, die du nicht selbst gefälscht hast!". Dieses Misstrauen ist wohl darauf zurückzuführen, dass statistische Methoden sehr oft nicht korrekt angewandt werden bzw. aus Zahlenwerten falsche Schlüsse gezogen werden.

Der Begriff **Statistik** (latein: „status" = Stand, Umstand) umfasst alle Methoden der Erfassung und Auswertung von Daten. Ziel der **beschreibenden Statistik** ist neben dem Erfassen und Veranschaulichen von Daten deren Auswertung mithilfe von Kennzahlen, die möglichst viel Information über die Originaldaten in einigen wenigen Zahlen zum Ausdruck bringen sollen.

7.1 Beschreibende Statistik

7.1.1 Grundbegriffe der statistischen Erhebung – Darstellung von Daten

Wie in vielen Fachgebieten wurde auch in der Statistik eine Fachsprache entwickelt. Sie soll sicherstellen, dass mit bestimmten Begriffen auch exakt die gleichen Bedeutungen verbunden bzw. Verwechslungen mit Begriffen der Alltagssprache vermieden werden.

AB **7.1** Ein Klassenraum einer Schule soll neu ausgemalt werden. Jede Schülerin bzw. jeder Schüler kann sich für eine der Farben gelb, hellgrün, hellblau oder weiß entscheiden. Jener Vorschlag, der die meisten Stimmen erhält, gilt als angenommen.
Überlegt, wie man die Daten erhebt und die Auswertung durchführt und führt sie in eurer Klasse durch.

Untersucht man zum Beispiel, wie viele Personen pro Haushalt in Österreich gemeldet sind, so nennt man die Objekte der Untersuchung – also die österreichischen Haushalte – die **Erhebungseinheiten**. Die Gesamtheit aller Erhebungseinheiten bildet die **Grundgesamtheit**. Aus praktischen Gründen kann man aber oft nur auf eine Auswahl, die **Stichprobe**, zurückgreifen. Die Eigenschaft, die man untersucht, nennt man **Merkmal**, deren möglichen Werte die **Merkmalsausprägungen**.

In obigem Beispiel ist also die Anzahl der im Haushalt lebenden Personen das Merkmal, die Merkmalsausprägungen sind die Werte 1, 2, 3 ... .

In der folgenden Tabelle werden einige Beispiele für die oben angeführten Begriffe genannt.

Grundgesamtheit	Merkmal	Merkmalsausprägungen
Österreichische Haushalte	Personenanzahl	1, 2, 3, 4 ...
Schülerinnen und Schüler einer HLW	Note in Mathematik	1, 2, 3, 4, 5
Würfe mit einer Münze	Seite	Zahl, Wappen
Angemeldete PKWs in Österreich	Antriebsart	Benzinmotor, Dieselmotor, anderer Antrieb

Die Möglichkeiten, Daten auszuwerten, hängen von deren Art ab. Man kann zum Beispiel die durchschnittliche Anzahl der in einem österreichischen Haushalt lebenden Personen errechnen, diesen Vorgang aber nicht sinnvoll auf die Merkmalsausprägungen Benzinmotor und Dieselmotor der PKWs übertragen. Man unterscheidet daher verschiedene Merkmalsarten:

- **Metrische** oder **quantitative Merkmale** sind zähl- oder messbar. Das Bilden von Differenzen ist sinnvoll.
  Zum Beispiel ist die Differenz zwischen einem 4-Personen-Haushalt und einem 5-Personen-Haushalt ebenso groß wie die Differenz zwischen einem 3-Personen-Haushalt und einem 4-Personen-Haushalt.

- **Ordinale Merkmale** oder **Rangmerkmale** sind Merkmale, deren Merkmalsausprägungen eine natürliche Reihenfolge haben. Am Beispiel von Schulnoten erkennt man, dass das Bilden von Differenzen hier nicht sinnvoll ist. Die Rangordnung (besser – schlechter) ist vorgegeben, der Unterschied zwischen den Noten 1 und 2 ist aber nicht unmittelbar mit dem zwischen den Noten 4 und 5 vergleichbar.

- **Nominale** oder **qualitative Merkmale** sind Merkmale, deren Merkmalsausprägungen keinerlei Vergleichbarkeit oder Reihenfolge zulassen, die also nur Namen (latein: „nomen“ = Name) sind.
  ZB: Antriebsart, Energieträger, Augenfarbe, Geschlecht, Religionszugehörigkeit ... .

Bemerkung: Mit den Formulierungen „durchschnittlich“ bzw. „im Mittel“ meint man oft das arithmetische Mittel, das schon aus Band 1 bekannt ist. Wir werden aber auch noch andere Mittelwerte kennenlernen. Wo Verwechslungen möglich sind, ist die Formulierung „durchschnittlich“ daher zu vermeiden.

Eine weitere für die Verarbeitung der Daten relevante Überlegung ist die Unterscheidung zwischen **diskreten** und **stetigen Merkmalen**. Können die Merkmalsausprägungen nur bestimmte Werte annehmen, zum Beispiel ganze Zahlen, so spricht man von diskreten Merkmalen. Als stetig werden Merkmale bezeichnet, die in einem gewissen Bereich jeden beliebigen Wert annehmen können, zum Beispiel die Körpergröße von Schülerinnen und Schülern.

7.2 Gib jeweils an, ob es sich um ein metrisches, ordinales oder nominales Merkmal handelt. C
1) Güteklassen von Äpfeln
2) Religionszugehörigkeiten von Personen
3) Inflationsraten verschiedener Länder
4) erzielte Weiten beim Kugelstoßen

7.3 Gib an, ob das Merkmal diskret oder stetig ist. Begründe deine Entscheidung. CD
1) Klassenschülerzahl
2) Masse von Hühnereiern
3) Verspätungen im Zugverkehr
4) Anzahl der PKWs pro Haushalt

7.4 Vervollständige die Tabelle mit eigenen Beispielen. Wenn es sich um ein metrisches Merkmal handelt, unterscheide zusätzlich zwischen stetig und diskret. C

Grundgesamtheit	Merkmal	Merkmalsart	stetig/diskret
Arbeitnehmer/innen			
	Farbe		
		ordinal	
			stetig

7.1.2 Tabellarische Darstellung und Häufigkeitsverteilung

Bei Erhebungen fallen Daten oft in ungeordneter und damit unübersichtlicher Form an. Aus dieser so genannten **Urliste** kann zum Beispiel durch Tabellieren eine **Strichliste** erzeugt werden. Die Häufigkeit des Auftretens der einzelnen Merkmalsausprägungen kann dann abgelesen werden. Man unterscheidet:

- Die **absolute Häufigkeit h_i** gibt an, wie oft eine bestimmte Merkmalsausprägung vorkommt. Die Summe der absoluten Häufigkeiten muss gleich dem Umfang n der Stichprobe bzw. der Grundgesamtheit sein.
  ZB: Von 100 untersuchten Personen haben 40 die Blutgruppe A, 15 die Blutgruppe B ...
  Daher sind die absoluten Häufigkeiten: $h_1 = 40$, $h_2 = 15$ ...
- Die **relative Häufigkeit r_i** gibt den Anteil einer bestimmten Merkmalsausprägung an. Bis auf Rundungsfehler ergibt die Summe der relativen Häufigkeiten 1.

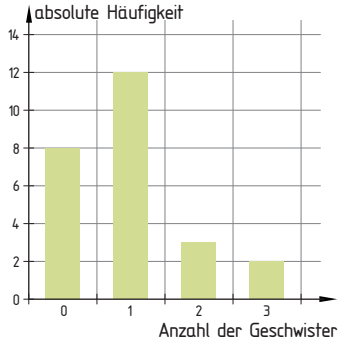
  $$r_i = \frac{\text{absolute Häufigkeit}}{\text{Gesamtzahl}} = \frac{h_i}{n}$$

  Die relative Häufigkeit wird meist als Dezimalzahl angeschrieben. Es sind aber auch Darstellungen als Bruchzahl bzw. Prozentsatz möglich. Die Darstellung der relativen Häufigkeit als Prozentsatz wird auch als **prozentuelle Häufigkeit p_i** bezeichnet.

Zur grafischen Veranschaulichung werden oft **Säulendiagramme** oder **Balkendiagramme** verwendet. Die Merkmalsausprägungen werden auf der waagrechten Achse aufgetragen, darüber Säulen, deren Höhe jeweils der (absoluten oder relativen) Häufigkeit entspricht.

ZB: Erfasst man die Anzahl der Geschwister, die die 25 Schülerinnen und Schüler einer Schulklasse haben, so kann folgende Urliste entstehen:
1, 0, 0, 2, 1, 0, 1, 1, 2, 0, 0, 1, 1, 1, 3, 0, 1, 0, 1, 1, 0, 3, 2, 1, 1

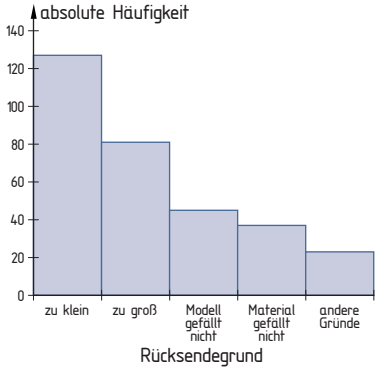
Merkmals-ausprägung	Strichliste	absolute Häufigkeit h_i	relative Häufigkeit r_i	prozentuelle Häufigkeit p_i
0	𝍸 |||	8	0,32	32 %
1	𝍸 𝍸 ||	12	0,48	48 %
2	|||	3	0,12	12 %
3	||	2	0,08	8 %
Summe		25	1	100 %



Wird die Häufigkeitsverteilung eines nominalen Merkmals mithilfe eines Säulendiagramms dargestellt, werden die Säulen oft nach fallenden Häufigkeiten geordnet. Diese Anordnung nennt man **Pareto-Diagramm** (benannt nach dem italienischen Ingenieur, Ökonomen und Soziologen Vilfredo Pareto, 1848 – 1923). Pareto-Diagramme sind vor allem in der Fehleranalyse gebräuchlich und liefern rasch einen guten Überblick über die wichtigsten Einflussgrößen.

ZB: Ein Versandhaus registriert bei zurückgeschickten Waren den Grund der Rücksendung.

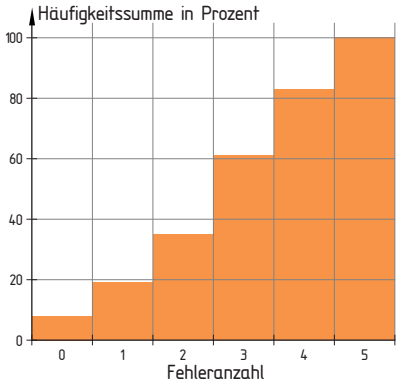
Rücksendegrund	absolute Häufigkeit
Modell gefällt nicht	45
Material gefällt nicht	37
zu groß	81
zu klein	127
andere Gründe	23



Bei quantitativen Merkmalen und Rangmerkmalen sind auch die **Häufigkeitssummen** aussagekräftig. Dabei werden jeweils die prozentuellen Häufigkeiten bis zu einer bestimmten Merkmalsausprägung aufsummiert. Die grafische Darstellung zeigt einen treppenförmigen Verlauf.

ZB: In einer Textilfabrik werden Stoffballen auf die Anzahl von Webfehlern hin untersucht. Sie werden in einer Häufigkeitstabelle erfasst, die um eine Spalte mit den aufsummierten Häufigkeiten ergänzt wird.

Fehleranzahl	h_i	r_i	p_i	Häufigkeitssumme (in Prozent)
0	16	0,08	8 %	8 %
1	22	0,11	11 %	19 %
2	32	0,16	16 %	35 %
3	52	0,26	26 %	61 %
4	44	0,22	22 %	83 %
5	34	0,17	17 %	100 %
Summe	200	1	100 %	



Aus der letzten Spalte kann man nun ablesen:
8 % der Stoffballen sind fehlerlos.
19 % der Stoffballen haben 0 oder 1 Fehler, also höchstens 1 Fehler.
35 % der Stoffballen haben 0, 1 oder 2 Fehler, also höchstens 2 Fehler, usw.

7.5 Die Höhen von 15 gleich alten Kastanienbäumen, die unter genau festgelegten Bedingungen gepflanzt wurden, sind in folgender Urliste angegeben (Werte in Meter): B

10,6 9,8 10,6 11,2 12,5 9,8 9,8 10,3 8,7 10,3 11,2 11,2 10,6 9,8 11,8

Berechne die absoluten, relativen und prozentuellen Häufigkeiten.

7.6 In einer Gemeinde wurde die Anzahl der TV-Geräte pro Haushalt erhoben: BC

Anzahl der TV-Geräte	0	1	2	3	4
Anzahl der Haushalte	14	203	135	51	18

1) In wie vielen Haushalten gibt es höchstens ein Fernsehgerät?
2) Berechne die relativen und prozentuellen Häufigkeiten.
3) Erstelle ein Säulendiagramm mit den absoluten Häufigkeiten.

7.7 Anlässlich einer Erhebung der Verkehrsbetriebe wurden 1 500 Personen befragt, an wie vielen Tagen der vergangenen Woche sie ein öffentliches Verkehrsmittel benutzt hatten. BC

Tage	0	1	2	3	4	5	6	7
Personen	220	185	96	124	178	412	208	77

1) Wie viele Personen haben nie, wie viele an 2 Tagen und wie viele an mindestens 6 Tagen öffentliche Verkehrsmittel benutzt?
2) Berechne die relativen Häufigkeiten und die Häufigkeitssummen.
3) Erstelle je ein Diagramm mit den absoluten Häufigkeiten und den Häufigkeitssummen.
4) Lies aus dem Diagramm der Häufigkeitssummen ab, wie viel Prozent der befragten Personen an höchstens fünf Tagen mit den „Öffis“ gefahren sind.

7.8 Die erste Nationalratswahl der 2. Republik in Österreich fand am 25. November 1945 statt und ergab folgende Stimmenverteilung: B

ÖVP	SPÖ	KPÖ	Sonstige
1 602 227	1 434 898	174 257	5 972

Ermittle die relativen Häufigkeiten in Prozent und erstelle ein Säulendiagramm.

Technologieeinsatz: Beschreibende Statistik

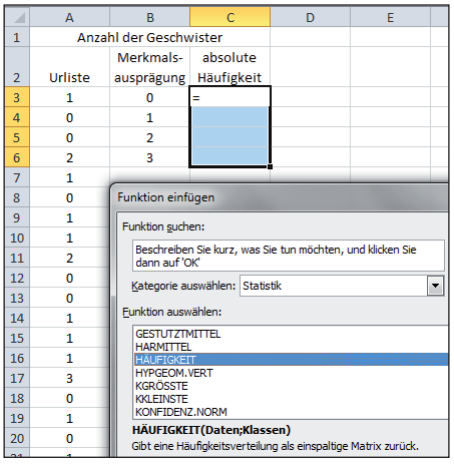
Sehr viele in der Statistik gebräuchliche Funktionen sind in Tabellenkalkulationsprogrammen bereits vordefiniert.

Tabellenkalkulationsprogramm (Excel 2010)

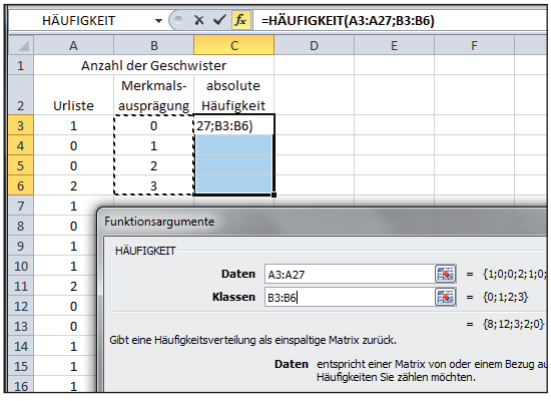
Sind die Werte einer Urliste bereits als Daten erfasst, so kann man deren absolute Häufigkeiten mithilfe des Befehls **HÄUFIGKEIT** ermitteln.

ZB: Von einer gegebenen Urliste (Anzahl der Geschwister, vergleiche Seite 260) sollen die absoluten und relativen Häufigkeiten ermittelt und anschließend ein Säulendiagramm erstellt werden.

Urliste: 1, 0, 0, 2, 1, 0, 1, 1, 2, 0, 0, 1, 1, 1, 3, 0, 1, 0, 1, 1, 0, 3, 2, 1, 1



- In der ersten Spalte werden die Daten der Urliste eingetragen. Da die Merkmalsausprägungen 0, 1, 2 und 3 vorkommen, werden diese in die zweite Spalte eingetragen.
- Um die absolute Häufigkeit jeder Merkmalsausprägung zu ermitteln, werden zuerst die Zellen neben den Ausprägungen markiert. Das ist notwendig, da die Funktion **HÄUFIGKEIT** eine Matrixfunktion ist, also eine Funktion, die mehrere Werte in die vorgesehenen Zellen ausgibt. Die Funktion **HÄUFIGKEIT** wird aus der Kategorie **Statistik** gewählt.



- Der Funktionsassistent bietet nun eine Eingabemaske an.
- Im Feld **Daten** werden die Werte der Urliste eingetragen. Dazu kann auch der Bereich (hier **A3:A27**) markiert werden.
- Im Feld **Klassen** werden die Merkmalsausprägungen eingegeben, also der Bereich **B3:B6** markiert.
- Die gesuchten Häufigkeiten erscheinen als Liste in der Eingabemaske.

	A	B	C
1	Anzahl der Geschwister		
2	Urliste	Merkmals-ausprägung	absolute Häufigkeit
3	1	0	8
4	0	1	12
5	0	2	3
6	2	3	2
7	1		

- Das Übertragen dieser Werte in die zuvor markierten Zellen erfolgt mit der Tastenkombination Strg + Shift + Enter. Beachte: Das Drücken von **OK** oder Betätigen der Enter-Taste alleine hätte zur Folge, dass nur der erste Wert angezeigt wird.

	A	B	C	D
1	Anzahl der Geschwister			
2	Urliste	Merkmals-ausprägung	absolute Häufigkeit	
3	1	0	8	
4	0	1	12	
5	0	2	3	
6	2	3	2	
7	1		=SUMME(C3:C6)	
8	0		SUMME(Zahl1; [Zahl2]; ...)	

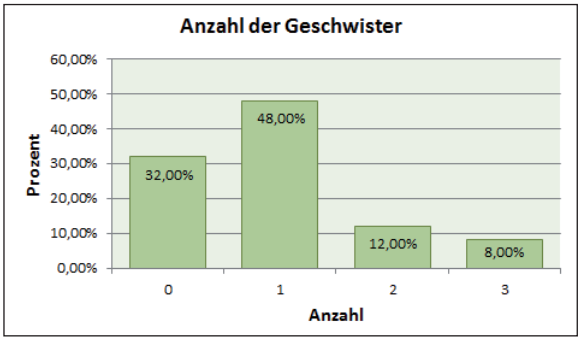
- Um für die weiteren Berechnungen die Gesamtzahl der erfassten Werte in einer Zelle zur Verfügung zu haben, bildet man die Summe der absoluten Häufigkeiten.

absolute Häufigkeit	relative Häufigkeit
8	=C3/C7
12	
3	
2	
25	

- Bei der Formel für die relative Häufigkeit muss die Zelle, in der die Gesamtanzahl der Werte steht, durch das **$**-Zeichen als absolute Adresse gekennzeichnet werden. Danach wird die Formel in die Zellen darunter kopiert.

absolute Häufigkeit	relative Häufigkeit	prozentuelle Häufigkeit
8	0,32	32,00%
12	0,48	48,00%
3	0,12	12,00%
2	0,08	8,00%
25		

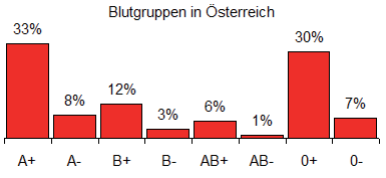
- Die gleichen Werte werden in die nächste Spalte übertragen und mithilfe der Formatangabe als Prozentangaben dargestellt.



- Das Säulendiagramm wird über **Einfügen** – **Diagramme** – **Säule** erstellt.
- Die Achsen- und Datenbeschriftung erfolgt über die **Diagrammtools**.

Sollen Buchstaben oder Wörter abgezählt werden, verwendet man den Befehl **ZÄHLENWENN** statt des Befehls **HÄUFIGKEIT**.

7.9 Die Häufigkeiten der Blutgruppen sind in verschiedenen Regionen unterschiedlich verteilt. Im nebenstehenden Diagramm ist die Verteilung der Blutgruppen inklusive Rhesusfaktor für Österreich abgebildet.



BCD

a) Welche Blutgruppe kommt in Österreich am häufigsten vor?

b) Frage mindestens 30 Personen nach ihrer Blutgruppe. Ermittle aus den Daten die absoluten und relativen Häufigkeiten und erstelle ein Säulendiagramm mit den prozentuellen Häufigkeiten. Präsentiere dein Ergebnis und vergleiche es mit dem angegebenen Diagramm.

7.1.3 Klassenbildung

Bisher haben wir uns mit der Verteilung der Häufigkeiten jeder in einer Erhebung auftretenden Merkmalsausprägung befasst. Dies ist jedoch nicht immer sinnvoll, zum Beispiel, wenn die Anzahl der Merkmalsausprägungen sehr groß ist, bzw. nicht möglich, wenn es sich um ein stetiges Merkmal handelt. In diesem Fall bilden wir Intervalle, so genannte **Klassen**. Man spricht dann von klassifizierten Daten. Die **Klasseneinteilung** wird im Allgemeinen nach folgenden Richtlinien getroffen:

- Als Richtwert für die Anzahl der Klassen geht man oft von $\sqrt{n}$ (n ... Anzahl der Daten) aus, mehr als 20 Klassen sind jedoch unüblich.
- Die Klassen sollten nach Möglichkeit **gleich breit** sein.
- Jeder Wert muss genau einer Klasse zugeordnet werden können. Daher verwendet man im Allgemeinen halboffene Intervalle (siehe Band 1, Abschnitt 1.4).
- Zur grafischen Darstellung verwendet man ein **Histogramm**. Dabei werden Rechtecke verwendet, deren Höhe jeweils der Klassenhäufigkeit entspricht, falls alle Klassen gleich breit sind. Andernfalls sollten die Rechtecksflächen den jeweiligen Häufigkeiten entsprechen.

AB

7.10 In einer HLW wurden 92 Schülerinnen und Schüler der 2. Jahrgänge gewogen (Werte in kg):
47,6 50,1 50,6 51,3 52,4 51,9 52,5 52,8 54,5 55,2 57,3 58,2 58,5 58,5 58,9 59,3 59,4 59,5 59,6 59,7 60,2 60,3 60,4 60,5 60,6 60,7 60,8 60,9 61,3 61,5 61,5 61,5 61,7 61,9 62,4 62,8 63,2 63,5 64,4 64,7 65,4 65,4 65,8 65,9 66,0 66,2 66,7 67,3 67,3 67,4 67,5 68,2 68,2 68,3 69,1 69,2 69,3 69,5 69,5 69,7 69,8 70,0 72,3 72,4 74,0 74,1 75,4 75,4 75,6 75,7 75,7 76,1 77,1 78,4 78,5 78,9 79,6 79,7 80,1 82,0 82,9 83,2 83,5 85,8 87,1 88,5 89,4 91,0 94,2 95,6 100,8 101,2

1) Bilde eine Klasseneinteilung.

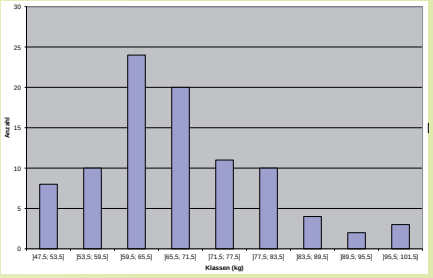
2) Erstelle eine Häufigkeitstabelle mit absoluten, relativen und prozentuellen Häufigkeiten sowie Häufigkeitssummen in Prozent und veranschauliche sie in einem Histogramm.

Lösung:

1) 92 Werte $\Rightarrow$ 9 oder 10 Klassen, wir wählen 9 Klassen.
Klassenbreite:
Max.: 101,2 kg; Min.: 47,6 kg $\Rightarrow \frac{101{,}2 - 47{,}6}{9} = 5{,}95\ldots \approx 6 \Rightarrow$ Klassenbreite 6 kg
Klassengrenzen:
Wir beginnen mit der Klassengrenze 47,5 und verwenden links offene Intervalle:
]47,5; 53,5],]53,3; 59,5],]59,5; 65,5] ...

2)

Klassen (Werte in kg)	absolute Häufigkeit	relative Häufigkeit	prozentuelle Häufigkeit	Häufigkeits-summe
]47,5; 53,5]	8	0,0870	8,70%	8,70%
]53,5; 59,5]	10	0,1087	10,87%	19,57%
]59,5; 65,5]	24	0,2609	26,09%	45,65%
]65,5; 71,5]	20	0,2174	21,74%	67,39%
]71,5; 77,5]	11	0,1196	11,96%	79,35%
]77,5; 83,5]	10	0,1087	10,87%	90,22%
]83,5; 89,5]	4	0,0435	4,35%	94,57%
]89,5; 95,5]	2	0,0217	2,17%	96,74%
]95,5; 101,5]	3	0,0326	3,26%	100,00%
	92			

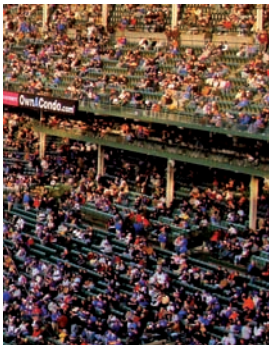


BC

7.11 Erhebe bei der Statistik Austria die aktuellste Altersverteilung der Österreicherinnen und Österreicher. Nimm eine Klasseneinteilung vor und stelle die Verteilung mithilfe eines Histogramms dar.

7.2 Kennzahlen statistischer Verteilungen

7.2.1 Lagemaße



Um die wesentliche Information von Häufigkeitsverteilungen gebündelt zu erfassen, verwendet man Kennzahlen. Dabei ist zu beachten, dass nicht alle Kennzahlen für alle Arten von Daten bzw. Merkmalen geeignet sind. Bei einigen Kennzahlen ist zu unterscheiden, ob mit einer **Grundgesamtheit** oder einer **Stichprobe** gearbeitet wird. Um die Unterscheidung in den Formeln zu erleichtern, werden meist folgende Unterschiede in der Schreibweise gemacht:
Die Anzahl der Daten aus einer **Grundgesamtheit** wird **N** genannt, die einer **Stichprobe n**. Gelten Kennzahlen für Grundgesamtheiten, so werden sie zur leichteren Unterscheidbarkeit mit griechischen Buchstaben abgekürzt.

Lagemaße ermöglichen es, die Lage des „Zentrums" einer Verteilung mit einer Zahl möglichst gut zu erfassen.

Das arithmetische Mittel

7.12 Am Flohmarkt hat Katrin am Freitag 30,00 € eingenommen, am Samstag 65,00 € und am Sonntag 40,00 €. Wie viel hat sie im Schnitt pro Tag eingenommen? AB

Im Alltag empfinden wir oft jenen Wert als „Durchschnitt", der in der Mathematik als **arithmetisches Mittel** folgendermaßen definiert ist:

$$\text{arithmetisches Mittel} = \frac{\text{Summe der Einzelwerte}}{\text{Anzahl der Einzelwerte}}$$

Das Berechnen der Summe der Einzelwerte und damit des arithmetischen Mittels ist nur für metrische (quantitative) Merkmale sinnvoll und zulässig. Im Allgemeinen ist das arithmetische Mittel kein Teil der Urliste.

Zum Beispiel: Die in der Praxis oft durchgeführte Berechnung eines (Schul-)Notendurchschnitts ist statistisch nicht korrekt, da es sich dabei um ein Rangmerkmal handelt. Es ist zum Beispiel die Differenz zwischen den Noten 1 und 2 nicht die gleiche wie zwischen den Noten 4 und 5, die Berechnung des arithmetischen Mittels ist daher streng genommen nicht sinnvoll.

Schreibweisen für das arithmetische Mittel:

Grundgesamtheit vom Umfang N:

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Stichprobe vom Umfang n:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Wurden die Häufigkeiten bereits ermittelt, so kann die Summe der Einzelwerte mit deren Hilfe rascher ermittelt werden. Anstelle der Addition aller Einzelwerte werden die Merkmalsausprägungen mit den jeweiligen Häufigkeiten multipliziert. Bei Verwendung der relativen Häufigkeiten entfällt die Division durch die Anzahl der Werte.

In Abschnitt 7.1.2 haben wir die Anzahl der Geschwister der Schülerinnen und Schüler einer Schulklasse untersucht.

$$\bar{x} = \frac{0 \cdot 8 + 1 \cdot 12 + 2 \cdot 3 + 3 \cdot 2}{25} = 0{,}96 \quad \text{bzw.}$$

$$\bar{x} = 0 \cdot 0{,}32 + 1 \cdot 0{,}48 + 2 \cdot 0{,}12 + 3 \cdot 0{,}08 = 0{,}96$$

Das heißt, im Mittel hat eine Gruppe von 100 Schülerinnen und Schüler in Summe 96 Geschwister.

Anzahl der Geschwister		
Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit
0	8	0,32
1	12	0,48
2	3	0,12
3	2	0,08
	25	

Mithilfe des Summenzeichens Σ (Σ ... „Sigma“, griechischer Großbuchstabe) kann man Summen kürzer anschreiben, zB: $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$

Das **arithmetische Mittel** (μ bzw. $\bar{x}$) ist das am häufigsten verwendete Lagemaß für metrische Merkmale.

Grundgesamtheit (Umfang N)

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

Stichprobe (Umfang n)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Berechnung mithilfe der **absoluten Häufigkeiten** h_i bzw. der **realitiven Häufigkeiten** r_i: (k ... Anzahl der verschiedenen Merkmalsausprägungen):

$$\mu = \frac{1}{N} \cdot \sum_{i=1}^{k} x_i \cdot h_i = \sum_{i=1}^{k} x_i \cdot r_i \qquad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot h_i = \sum_{i=1}^{k} x_i \cdot r_i$$

Quantile: Median, Quartile, Perzentile

BC **7.13** Ein Dorf besteht aus 10 Häusern, wobei 5 Häuser eine Wohnfläche von 100 m^2, 3 Häuser eine Wohnfläche von 130 m^2 und 2 Häuser eine Wohnfläche von 200 m^2 haben. In der Nähe steht ein Schloss mit einer Wohnfläche von 7 650 m^2. Ermittle das arithmetische Mittel der Wohnflächen ***1)*** nur des Dorfs, ***2)*** des Dorfs inklusive Schloss. Welcher Eindruck entsteht jeweils?

Die Berechnung des Mittelwerts ist zwar für metrische Merkmale immer zulässig, jedoch nicht immer sinnvoll. Mitunter beeinflusst ein stark abweichender Wert den Mittelwert so, dass ein falscher Eindruck entsteht. Einen solchen Wert bezeichnet man als **Ausreißer**.

ZB: Dem Einkommensbericht der Statistik Austria kann man entnehmen, dass im Jahr 2010 der (arithmetische) Mittelwert des Bruttojahreseinkommens der unselbständig erwerbstätigen Österreicherinnen und Österreicher 28 715,00 € betrug. Da wenige sehr hohe Einkommen diesen Wert jedoch stark beeinflussen, wird üblicherweise eine andere Kennzahl angegeben: Der Betrag in der Spalte Median bedeutet, dass 50 % der Erwerbstätigen 24 516,00 € oder weniger verdient haben.

Bruttojahreseinkommen der unselbständig Erwerbstätigen 2010

Jahr	Median	Arithmetisches Mittel
	EUR	
2010	24 516	28 715

Der Wert in der „Mitte“ einer geordneten Liste von Werten heißt **Median** oder **Zentralwert** $\tilde{x}$. Ist die Anzahl der Werte gerade, so ist der Median das arithmetische Mittel der beiden mittleren Werte. Der Median „teilt“ die geordnete Liste in zwei gleich große Teile. Mindestens 50 % aller Werte sind kleiner gleich dem Median, mindestens 50 % aller Werte sind größer gleich dem Median. Da als „Rechenschritt“ nur erforderlich ist, die Liste der Urwerte zu ordnen, ist der Median auch für Rangmerkmale angebbar. Einzelne, weit von den anderen Werten entfernt liegende Merkmalsausprägungen, beeinflussen den Median im Gegensatz zum arithmetischen Mittel nicht.

Als **Median** oder **Zentralwert $\tilde{x}$** einer Verteilung bezeichnet man den **mittleren Wert der geordneten Liste** bzw. das arithmetische Mittel der beiden mittleren Werte, falls die Anzahl der Werte gerade ist.
(Mindestens) 50 % aller Werte sind kleiner gleich $\tilde{x}$, (mindestens) 50 % sind größer gleich $\tilde{x}$.

7.14 Ermittle den Median der angegebenen Daten. B

a) 8 4 7 6 4 9 11 7 5 ***b)*** 78 45 32 66 81 55 90 60 62 71 49 84

Lösung:

a) Geordnete Liste:

4 4 5 6 **7** 7 8 9 11

Median: $\tilde{x} = 7$

- Die Liste enthält neun Elemente, der Wert in der Mitte ist daher das 5. Listenelement.

b) Geordnete Liste:

32 45 49 55 60 **62** **66** 71 78 81 84 90

Median: $\tilde{x} = \frac{62 + 66}{2} = 64$

- Die Liste enthält zwölf Elemente. Der Median ist das arithmetische Mittel der beiden „mittleren" Werte.

7.15 Bei einer Telefonumfrage werden 15 Personen befragt, wie oft sie im vergangenen Jahr ein Theater besucht haben. Ermittle aus der Urliste das arithmetische Mittel und den Median. Welcher Wert beschreibt die Verteilung besser? Begründe deine Antwort. BCD

Urliste: 0 1 1 5 3 2 0 27 3 0 8 7 6 0 2

Lösung:

Arithmetisches Mittel: $\bar{x} = \frac{0 \cdot 4 + 1 \cdot 2 + 2 \cdot 2 + 3 \cdot 2 + 5 \cdot 1 + 6 \cdot 1 + 7 \cdot 1 + 8 \cdot 1 + 27 \cdot 1}{15} = 4{,}33\ldots \approx 4{,}3$

Median:

Geordnete Liste: 0 0 0 0 1 1 2 **2** 3 3 5 6 7 8 27

$\tilde{x} = 2$

Der Median beschreibt die Verteilung besser, weil das arithmetische Mittel durch den Ausreißer 27 stark beeinflusst wird.

Detailliertere Informationen als der Median liefern die **Quartile**. Das zweite Quartil q_2 entspricht dem Median. Aus den vor dem Median liegenden Teil der geordneten Liste wird – analog zum Median – der „mittlere" Wert bestimmt. Dieser wird als erstes Quartil q_1 bezeichnet. (Mindestens) 25 % alle Werte sind kleiner gleich q_1. Ebenso kann in der oberen Hälfte das dritte Quartil q_3 ermittelt werden.

ZB: In einem Kindergarten wurde die Größe von 15 Dreijährigen erhoben.
Geordnete Liste der Werte (in cm):

88,3 91,6 92,4 **93,5** 94,9 96,4 97,0 **97,6** 98,1 98,7 99,3 **99,8** 100,3 103,8 105,1

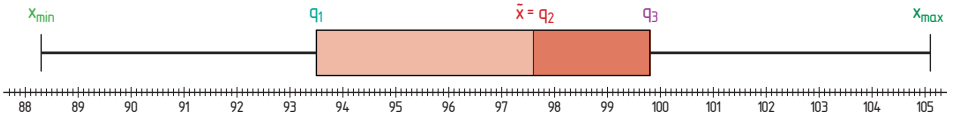
q_1 — $\tilde{x} = q_2$ — q_3

Mindestens 25 % aller Kinder sind 93,5 cm groß oder kleiner.

Mindestens 50 % aller Kinder sind höchstens 97,6 cm groß.

Mindestens 75 % aller Kinder sind höchstens 99,8 cm groß.

Zur grafischen Veranschaulichung verwendet man einen **Boxplot** (**Kastenschaubild**).



Für eine noch genauere Unterteilung verwendet man **Perzentile**. Sie geben für vorgegebene Prozentsätze p % den Wert an, für den p % aller Werte kleiner gleich dieser Grenze sind. Üblich ist diese Angabe zum Beispiel, um Größen- und Gewichtsentwicklung von Kindern darzustellen.

Aus Abbildung 7.1 kann man ablesen:

- 3 % der 12-jährigen Mädchen haben 35 kg oder weniger.
- 90 % der 13-jährigen Mädchen haben maximal 75 kg, also nur 10 % der 13-jährigen haben 75 kg oder mehr.

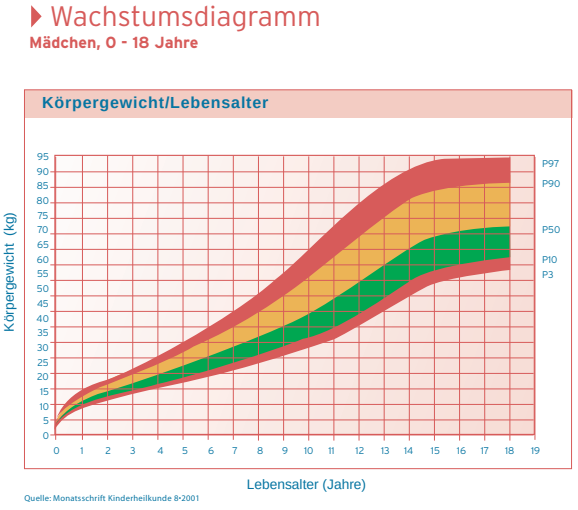
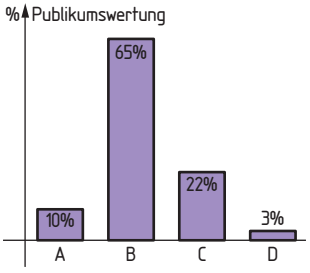


Abb. 7.1

Modalwert (Modus)

In einer bekannten Fernsehquiz-Show, bei der Kandidatinnen bzw. Kandidaten aus vier möglichen Antworten A, B, C und D die richtige auswählen müssen, kann man die Hilfe der Zuschauer/innen in Form eines „Publikumsjokers“ in Anspruch nehmen. Oft entscheiden sich die Kandidatinnen bzw. Kandidaten dann für jene Antwort, die die meisten Zuschauer/innen gewählt haben. Dieser am **häufigsten vorkommende Wert** heißt **Modalwert** oder **Modus** der Verteilung. Treten mehrere Werte gleich häufig auf, so gibt es mehrere Modalwerte. Dieses Lagemaß ist das einzige, das auch für nominale (qualitative) Merkmale ermittelt werden kann.



Der häufigste Wert einer Liste heißt **Modalwert** oder **Modus**, bei gleicher Häufigkeit gibt es mehrere Modalwerte.

BC **7.16** An der Kassa eines Baumarkts wurden die Kunden nach der Postleitzahl ihres Wohnorts befragt. Ermittle den Modalwert der Urliste:

1210 1190 1210 1220 1220 2202 1200 1110 1190 1210 1180 2230
1190 2202 1200 1190 1180 2230 1220 1210 2211 1190 1210

Lösung:

1110: | 1200: || 2202: ||
1180: || 1210: 𝍸||| 2211: |
1190: 𝍸||| 1220: ||| 2230: ||

Es gibt zwei Modalwerte: 1190 und 1210

Technologieeinsatz: Lagemaße

Tabellenkalkulationsprogramm (Excel 2010)

In der Funktionengruppe Statistik gibt es die Befehle **MITTELWERT** zur Bestimmung des arithmetischen Mittels, **MEDIAN** und **MODUS.EINF**.
Als Eingabe ist jeweils die Liste der Daten anzugeben.

4	=MITTELWERT(A1:A9)
5	
7	=MEDIAN(A1:A9)
10	
9	=MODUS.EINF(A1:A9)
1	
2	
5	
3	

4	5,11	Mittelwert
5		
7	5	Median
10		
9	5	Modalwert
1		
2		
5		
3		

7.17 Ermittle das arithmetische Mittel der Daten aus der gegebenen Häufigkeitsverteilung. B

a)

Länge in mm	absolute Häufigkeit
2	17
2,5	24
3	26
3,5	15
4	22
4,5	9
5	2

b)

Profiltiefe in mm	prozentuelle Häufigkeit
3,6	2,4 %
3,7	7,3 %
3,8	10,8 %
3,9	8,2 %
4	9,5 %
4,1	15,6 %
4,2	22,4 %
4,3	15,3 %
4,4	5,1 %
4,5	3,4 %

7.18 In Abbildung 7.2 sind die Einwohnerzahlen der EU-Mitgliedsstaaten angegeben. BC
1) Ermittle das arithmetische Mittel und den Median.
2) Beim Übertragen der Werte wird für Frankreich irrtümlich ein Wert von 654 Mio. eingegeben. Wie wirkt sich dieser Fehler auf das arithmetische Mittel bzw. den Median aus?

Mitgliedsstaat	Bevölkerung (Mio.)
Malta	0,4
Luxemburg	0,5
Zypern	0,9
Estland	1,3
Lettland	2,0*
Slowenien	2,1
Litauen	3,2
Irland	4,5
Finnland	5,4
Slowakei	5,4
Dänemark	5,6
Bulgarien	7,3
Österreich	8,4
Schweden	9,5
Ungarn	10,0
Tschechien	10,5
Portugal	10,5
Belgien	11,0
Griechenland	11,3
Niederlande	16,7
Rumänien	21,4
Polen	38,2
Spanien	46,2
Italien	60,9
Großbritannien	63,0
Frankreich	65,4
Deutschland	81,8
Gesamt	503,5

Quelle: EUROSTAT, Stand Jänner 2012, * ... vorläufig

Abb. 7.2

7.19 Bei einer Prüfung erreichten die Kandidaten folgende Punktezahlen: 22 15 9 18 12 23 25 17 16 12 19 21 20 3 19 20 14 16 16 22 23 9 11 B
1) Berechne das arithmetische Mittel.
2) Ermittle den Median, die Quartile q_1 und q_3 und erstelle einen Boxplot.

7.20 Welches der behandelten Lagemaße ist immer ein Wert der Urliste, welches nicht? Begründe deine Antwort. D

7.21 In der 2A und in der 2B wurde der gleiche Test abgehalten. Die 27 Schülerinnen und Schüler der 2A erreichten im Mittel 34 Punkte. In der 2B mit 23 Schülerinnen und Schülern betrug das arithmetische Mittel 38 Punkte. Wie groß ist das arithmetische Mittel der Punktezahlen aller 50 Teilnehmenden? Welche der beiden Klassen hat das arithmetische Mittel mehr beeinflusst? ABC

7.22 Für das arithmetische Mittel gilt, dass die Summe der Differenzen aller Werte vom Mittelwert null ergibt. ABD
1) Prüfe die Behauptung an einem selbst gewählten Beispiel mit fünf Werten nach.
2) Beweise, dass diese Aussage allgemein gültig ist.

7.2.2 Streuungsmaße

BC **7.23** Zwei Gruppen zu je fünf Personen erreichten bei einem Test folgende Punkteanzahlen: Gruppe 1: 2 3 3 3 4 Gruppe 2: 1 1 3 5 5

1) Gib für beide Gruppen das arithmetische Mittel an.

2) Was unterscheidet die Testergebnisse der beiden Gruppen voneinander?

Mithilfe der Lagemaße können wir gewisse Informationen über die Größe der Werte eines Datensatzes angeben. Diese Zahlen sagen jedoch nichts darüber aus, wie weit die Werte voneinander oder von einem gewählten Lagemaß entfernt liegen. Die Streuungsmaße beschreiben diese Abweichung voneinander. Die Berechnung dieser Maßzahlen ist im Allgemeinen nur für metrische Merkmale möglich.

Spannweite (Range)

Die Differenz zwischen dem größten und dem kleinsten Wert von Daten wird als **Spannweite** $R = x_{max} - x_{min}$ bezeichnet. Die Spannweite ist leicht zu ermitteln und von Lagemaßen unabhängig, hat aber den Nachteil, durch einzelne Ausreißer stark beeinflusst zu werden.

Interquartilsabstand

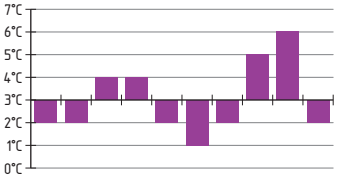
Weniger anfällig für Ausreißer ist die Differenz zwischen den Quartilsgrenzen q_1 und q_3. Der **Interquartilsabstand** $d = q_3 - q_1$ gibt an, in welchem Bereich die mittleren 50 % aller Werte liegen.

Varianz und Standardabweichung

Die wichtigste Kennzahl zur Beschreibung der Streuung von metrischen Merkmalen ist die **Varianz** σ^2 bzw. die Wurzel aus der Varianz, die **Standardabweichung** σ. Wir suchen eine Maßzahl, die beschreibt, wie weit die Merkmalsausprägungen „im Durchschnitt" vom Mittelwert μ entfernt sind. Die Summe aller Differenzen vom Mittelwert ist jedoch ungeeignet, da sie immer null ist (siehe Aufgabe 7.22).

ZB: Die Temperatur T um 12:00 Uhr wird zehn Tage lang aufgezeichnet:

2 °C 2 °C 4 °C 4 °C 2 °C 1 °C 2 °C 5 °C 6 °C 2 °C



Wir berechnen den Mittelwert $\mu = 3$ °C. Mithilfe eines Diagramms können die jeweiligen Abweichungen vom Mittelwert veranschaulicht werden.

Um ein Maß für die Streuung zu erhalten, quadrieren wir die Differenzen vom Mittelwert μ, hier also $(T_i - 3)^2$. Den Mittelwert dieser Abweichungsquadrate bezeichnet man als Varianz σ^2.

$$\sigma^2 = \frac{(2-3)^2 + (2-3)^2 + (4-3)^2 + (4-3)^2 + (2-3)^2 + (1-3)^2 + (2-3)^2 + (5-3)^2 + (6-3)^2 + (2-3)^2}{10} = 2{,}4$$

Um die Berechnung zu vereinfachen, kann man bei gleichen Werten die Abweichungsquadrate jeweils mit der Häufigkeit multiplizieren: $\sigma^2 = \frac{(1-3)^2 + (2-3)^2 \cdot 5 + (4-3)^2 \cdot 2 + (5-3)^2 + (6-3)^2}{10} = 2{,}4$

Die Standardabweichung $\sigma = \sqrt{2{,}4} \approx 1{,}55$ °C gibt an, wie stark die Werte in Bezug auf den Mittelwert streuen.

Arbeitet man hingegen mit einer Stichprobe mit dem Mittelwert $\bar{x}$, so wird die Varianz mit s^2 und mit folgender Formel berechnet:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Es wird dabei nicht durch den Stichprobenumfang n, sondern durch (n – 1) dividiert. Damit ergeben sich bessere Möglichkeiten, um von der Stichprobe auf die Grundgesamtheit zu schließen. Die mathematische Begründung erfordert allerdings Kenntnisse, die über an dieses Kapitel zu stellenden Anforderungen weit hinausgehen.

Sind umfangreiche Datenmengen annähernd „normalverteilt", liegen etwa $\frac{2}{3}$ aller Daten im Intervall $[\mu - \sigma, \mu + \sigma]$ bzw. $[\bar{x} - s, \bar{x} + s]$ (einfache Standardabweichung-Umgebung), ca. 95 % aller Daten im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$ bzw. $[\bar{x} - 2s, \bar{x} + 2s]$ (doppelte Standardabweichung-Umgebung) und „fast alle" Daten im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ bzw. $[\bar{x} - 3s, \bar{x} + 3s]$ (dreifache Standardabweichung-Umgebung).

„Normalverteilte Datenmengen" werden in Band 4 ausführlich behandelt werden.

Das wichtigste **Streuungsmaß** in der Statistik ist die **Varianz**.
Man unterscheidet bei der Berechnung zwischen Grundgesamtheiten und Stichproben.

Varianz einer Grundgesamtheit mit N Werten:

$$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$$

Varianz einer Stichprobe mit n Werten:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Varianz von klassifizierten Daten:
Es wird mit den Klassenmitten x_i und deren Klassenhäufigkeiten h_i gearbeitet.

$$s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{k} h_i \cdot x_i^2 - n \cdot \bar{x}^2 \right) \quad \text{mit} \quad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} h_i \cdot x_i \qquad k \ldots \text{Anzahl der Klassen}$$

Die Quadratwurzel aus der Varianz wird als **Standardabweichung σ** bzw. **s** bezeichnet. Sie gibt an, wie stark die Werte in Bezug auf den Mittelwert streuen.

Bemerkung: Für die numerische Berechnung kann man auch folgende, durch Umformen entstandene Formeln verwenden (siehe Aufgabe 7.30):

$$\sigma^2 = \frac{1}{N} \cdot \left(\sum_{i=1}^{N} x_i^2 - N \cdot \mu^2 \right) \quad \text{bzw.} \quad s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right)$$

7.24 Ermittle die Varianz bzw. die Standardabweichung der Stichprobe auf zwei Arten. **B**

11 8 10 10 12 7 11 13 9 10 8 8 7 11 12 13

Lösung:

1) $\bar{x} = \frac{7 \cdot 2 + 8 \cdot 3 + 9 + 10 \cdot 3 + 11 \cdot 3 + 12 \cdot 2 + 13 \cdot 2}{16} = 10$

Mittelwert: $\bar{x} = 10$

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

$$= \frac{1}{16-1} \cdot [2 \cdot (7-10)^2 + 3 \cdot (8-10)^2 + (9-10)^2 + 3 \cdot (10-10)^2 + 3 \cdot (11-10)^2 + 2 \cdot (12-10)^2 + 2 \cdot (13-10)^2] =$$

$$= \frac{1}{15} \cdot (18 + 12 + 1 + 0 + 3 + 8 + 18) = \frac{60}{15} = 4$$

2) $s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right) =$

$$= \frac{1}{16-1} \cdot [(2 \cdot 7^2 + 3 \cdot 8^2 + 9^2 + 3 \cdot 10^2 + 3 \cdot 11^2 + 2 \cdot 12^2 + 2 \cdot 13^2) - 16 \cdot 10^2] =$$

$$= \frac{1}{15} \cdot [(98 + 192 + 81 + 300 + 363 + 288 + 338) - 1\,600] = \frac{1\,660 - 1\,600}{15} = 4$$

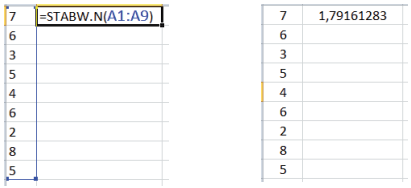
Varianz $s^2 = 4$ bzw. Standardabweichung $s = 2$

Technologieeinsatz: Standardabweichung

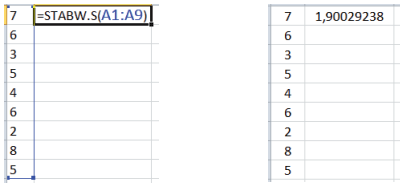
Tabellenkalkulationsprogramm (Excel 2010)

Standardabweichung

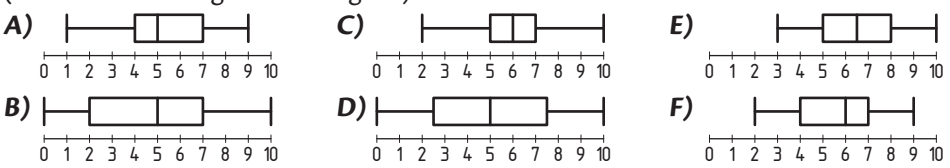
einer Grundgesamtheit ... **STABW.N**



einer Stichprobe ... **STABW.S**



C **7.25** Ordne den gegebenen sechs Boxplots die unten angegebenen Aussagen zu (Mehrfachnennungen sind möglich).



1) Der Interquartilabstand beträgt 3.
2) Die Verteilung ist symmetrisch.
3) Ein Viertel der Werte ist größer gleich 7.
4) Die Spannweite beträgt 7.
5) Die Hälfte der Werte liegt im Bereich [2; 7].
6) Ein Viertel der Werte ist kleiner gleich 5.
7) Die Daten streuen stark.
8) Die Hälfte der Werte ist größer 5.

B **7.26** In einer KFZ-Werkstätte wurde eine Stichprobe über den Zeitaufwand bei der Reparatur eines bestimmten Schadens erhoben (Angaben in Stunden):

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

1) Ermittle den Median, die Quartile q_1 und q_3, den Interquartilsabstand, den kleinsten und größten Wert sowie die Spannweite und zeichne einen Boxplot.
2) Ermittle das arithmetische Mittel und die Varianz.

B **7.27** Bei einer Telefonumfrage wurde die Anzahl der Mobiltelefone pro Haushalt erfragt:

Anzahl	0	1	2	3	4
Häufigkeit	18	156	243	161	87

1) Erstelle ein Histogramm.
2) Ermittle das arithmetische Mittel.
3) Berechne die Varianz und die Standardabweichung.

B **7.28** Berechne für die Daten aus 7.10 die Varianz und die Standardabweichung.

AB **7.29** Bei einer Abfüllanlage wurden folgende Messungen vorgenommen (Werte in Milliliter):

434 423 501 509 423 499 500 421 471 456 461 456 499 485 452 437 457 464 475 480
425 425 491 471 483 491 421 502 422 465 480 449 479 450 480 425 499 475 433 461
434 423 501 509 423 499 500 421 471 456 461 456 499 485 452 437 457 464 475 480
444 488 511 489 475 482 476 465 458 449 429 490 471 450 472 474 457 481 462 429

1) Klassifiziere die Daten und erstelle ein Histogramm.
2) Berechne aus den klassifizierten Daten den Mittelwert und die Standardabweichung.

BD **7.30** Zeige die Richtigkeit der Umformung der Formel für die Varianz einer Stichprobe:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right)$$

Hinweis: Beachte, dass die Summe aller Merkmalsausprägungen x_i das n-fache des Mittelwerts beträgt.

7.3 Korrelation und Regression

7.3.1 Lineare Korrelation

ABD

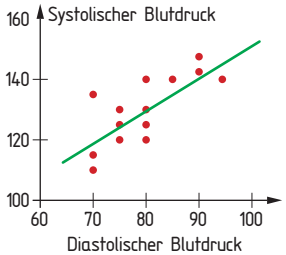
7.31 Haben große Eltern große Kinder? Sammle die Daten der Körpergröße der Eltern und deren Kinder für deine Schulklasse. Lässt sich ein Zusammenhang erkennen?

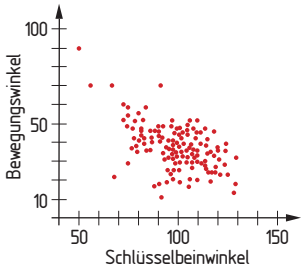
In vielen wissenschaftlichen Bereichen ist es notwendig, zwei Größen in Zusammenhang zueinander zu bringen.

So kann man zB bei quantitativen chemischen Analysen aus der Schwächung der Strahlungsintensität beim Durchgang durch eine absorbierende Lösung auf deren Konzentration schließen. In der Medizin möchte man den Zusammenhang zwischen Lebensgewohnheiten und gesundheitlichem Befinden erforschen. Dazu gehören Fragen wie „Leben verheiratete Menschen länger?“ oder „Verursacht Rauchen Lungenkrebs?“ Um solche Fragen beantworten zu können, muss man zwei Merkmale vergleichen. So können die Anzahl der täglich gerauchten Zigaretten und das Ergebnis eines Lungenfunktionstests (Spirometrie) miteinander in Beziehung gebracht werden.

Die **Korrelation** (latein: relatio = Beziehung) beschreibt die Beziehung zwischen zwei oder mehreren Größen. Allerdings lassen sich aus der Korrelation keine Schlüsse ziehen, ob eine der Größen die andere kausal beeinflusst, das heißt, ob sie diese Größe bzw. ihre Ausprägung verursacht. So lässt sich zB das gemeinsame Auftreten von Störchen und Geburten rechnerisch zeigen, ohne dass man daraus einen kausalen Zusammenhang ableiten könnte. Stellt man den Zusammenhang zwischen zwei Größen in einem Koordinatensystem dar, erhält man ein so genanntes **Punktwolken-Diagramm**. Bei linearer Korrelation liegen diese Punkte annähernd auf einer Geraden.

Das nebenstehende Diagramm gibt den diastolischen und systolischen Wert des Blutdrucks verschiedener Personen wieder. Es zeigt, dass ein hoher Wert der einen Größe häufig gleichzeitig mit einem hohen Wert der zweiten Größe gemessen wurde. Kennt man nur einen der beiden Werte, so liegt der zweite Wert vermutlich in einem eingeschränkten Bereich. Die durchgezogene Linie symbolisiert diese Beziehung.





Ein solcher Zusammenhang muss nicht immer gegeben sein. Die Untersuchung der nebenstehenden Grafik befasst sich mit der Abhängigkeit der Beweglichkeit der Schulter von der Struktur des Schlüsselbeins. Die Punkte sind regellos verteilt, eine Korrelation zwischen Schlüsselbeinwinkel und Bewegungswinkel lässt sich mithilfe dieser Grafik nicht erkennen.

Karl Pearson (britischer Mathematiker, 1857 – 1936) entwickelte eine Maßzahl, deren Wert r ein Schätzwert für die Richtung und Ausprägung eines linearen Zusammenhangs zwischen zwei Messgrößen darstellt.

Pearson'scher Korrelationskoeffizient (Empirischer Korrelationskoeffizient)

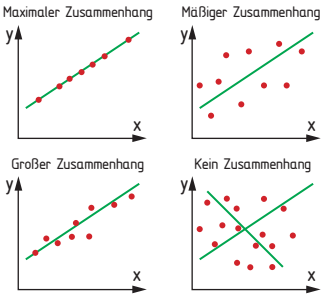
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x_i, y_i ... Koordinaten von i Messpunkten
$\bar{x}$, $\bar{y}$ ... arithmetisches Mittel

Der lineare Zusammenhang ist umso besser, je näher $|r|$ bei 1 liegt.

- $|r| = 1$ ... Alle Punkte liegen auf einer Geraden, dies stellt den maximalen Zusammenhang dar.
- $r = 0$ ... Die Punkte liegen verstreut, es gibt keinen linearen Zusammenhang.
- $0 < |r| < 1$ ... Je näher der Wert von $|r|$ bei 1 liegt, desto größer ist der lineare Zusammenhang.

Das Vorzeichen von r gibt die Richtung der Geraden an, für $r < 0$ ist sie fallend, für $r > 0$ ist sie steigend.



B **7.32** In einer Gesundheitsbefragung wurden fünf Personen verschiedenen Alters nach ihrem „subjektiven Gesundheitszustand" befragt. Die Antworten variieren dabei von 1,0 (Sehr gut) bis 5,0 (Sehr schlecht). Berechne den Korrelationskoeffizienten.

Alter in Jahren	22	37	45	48	62
Gesundheitszustand	1,2	1,0	1,8	2,7	3,4

Lösung:

$\bar{x} = \frac{214}{5} = 42{,}8$ und $\bar{y} = \frac{10{,}1}{5} = 2{,}02$

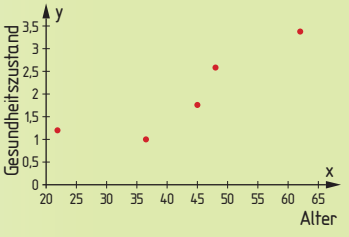
- Zur Vereinfachung der Berechnungen wird die Tabelle angelegt.

Alter	Gesundheitszustand	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
22	1,2	−20,8	432,64	−0,82	0,672 4	17,056
37	1,0	−5,8	33,64	−1,02	1,040 4	5,916
45	1,8	2,2	4,84	−0,22	0,048 4	−0,484
48	2,7	5,2	27,04	0,68	0,462 4	3,536
62	3,4	19,2	368,64	1,38	1,904 4	26,496
Summe 214	10,1		866,8		4,128	52,52

$$\sum_{i=1}^{n}(x_i - 42{,}8) \cdot (y_i - 2{,}02) = 52{,}52$$

$$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{866{,}8 \cdot 4{,}128} = 59{,}817\ldots$$

$$r = \frac{52{,}52}{59{,}817\ldots} = 0{,}878\,0\ldots \Rightarrow r \approx 0{,}878$$



Rein rechnerisch haben wir also den Korrelationskoeffizienten $r \approx 0{,}878$ ermittelt. Hier handelt es sich offensichtlich um eine hohe positive Korrelation. Da der Wert jedoch nur auf einer Befragung von fünf Personen beruht, sind jegliche Verallgemeinerungen wie zB „Je älter die Person ist, desto schlechter ist ihr subjektiver Gesundheitszustand." nicht zulässig.

Bei der Interpretation von Korrelationskoeffizienten ist zu beachten, dass eine einheitliche Aussage über die Höhe des Zusammenhangs nicht definiert ist. So wird zB im Bereich der Medizin oder Pharmazie ein Korrelationskoeffizient von 0,3 mitunter schon als sehr hoch gewertet, während in den Wirtschaftswissenschaften von hoher Korrelation meist erst ab 0,9 gesprochen wird.

Tabellenkalkulationsprogramm (Excel 2010)

	A	B	C
1	x	y	
2	22	1,2	
3	37	1,0	
4	45	1,8	
5	48	2,7	
6	62	3,4	
7	Korrelationskoeffizient:		0,878

Der Korrelationskoeffizient kann mithilfe eines Tabellenkalkulationsprogramms berechnet werden.

Die Berechnung erfolgt mit dem Befehl
`=KORREL(Matrix1;Matrix2)`

7.33 Berechne den Korrelationskoeffizienten der Datenpaare M = {(2; 4), (3; 3), (4; 5), (6; 6)} ohne Verwendung eines Computers. B

7.34 ***1)*** Stelle die Datenmenge M grafisch dar. B
2) Wie gut korrelieren die Daten? Begründe deine Antwort.
3) Überprüfe deine Vermutung durch Berechnung des Korrelationskoeffizienten.
a) M = {(2; 4), (3; 3), (4; 5), (6; 6), (7; 8), (8; 7), (10; 9), (12; 13)}
b) M = {(1; 10), (2; 9), (3; 12), (5; 15), (6; 14), (7; 15), (9; 18), (11; 23), (14; 27), (15; 30)}

7.35 Zehn internationale Konzerne einer Branche hatten im Jahr 2012 die in der Tabelle ausgewiesenen Werbeausgaben und Jahresumsätze. BC

Werbeausgaben (in Millionen €)	3,15	3,05	1,75	0,78	1,52	1,60	2,12	0,81	0,91	2,12
Jahresumsatz (in Milliarden €)	12,04	11,05	6,45	1,25	5,25	4,65	8,90	1,62	2,24	7,32

Berechne den Korrelationskoeffizienten und interpretiere das Ergebnis.

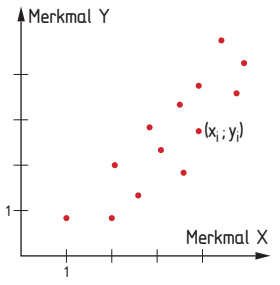
7.36 Schreibe die Schuhgrößen und die Körpergrößen von fünf Freunden oder Verwandten an. Berechne den Korrelationskoeffizienten und interpretiere das Ergebnis. BC

7.37 Bei einem Asynchronmotor wurde das Drehmoment M in Abhängigkeit von der Drehzahl n an 4 verschiedenen Messpunkten ermittelt: BC

	1	**2**	**3**	**4**
Drehzahl n in $\frac{1}{\text{min}}$	100	500	2 300	2 900
Drehmoment M in Nm	19,5	18,9	23	10

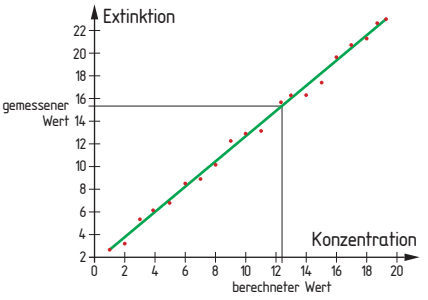
1) Zeichne ein Punktwolken-Diagramm.
2) Bestimme den Korrelationskoeffizienten und interpretiere das Ergebnis.

7.3.2 Regression



B **7.38** ***1)*** Lege durch die dargestellte Punktmenge eine Gerade, sodass alle Punkte möglichst „nahe“ bei dieser Geraden liegen.
2) Gib die Gleichung der Geraden y = ax + b durch Ablesen der Werte für a und b an.

In vielen naturwissenschaftlichen Anwendungen ist es nötig, aus einer bekannten Datenmenge auf einen unbekannten Wert zu schließen. So kann man zB die unbekannte Konzentration einer Lösung mithilfe einer Verdünnungsreihe bestimmen. Dabei werden verschiedene Verdünnungen bekannter Konzentration angefertigt und eine korrelierende Eigenschaft wie zB die Extinktion gemessen.



Aus den gemessenen Werten kann die so genannte **Regressionsgerade** erstellt werden, mit deren Hilfe die unbekannte Konzentration berechnet werden kann.

Ermittlung einer **Regressionsgeraden** y = ax + b durch die Punkte $P_i(x_i|y_i)$

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{und} \quad b = \bar{y} - a \cdot \bar{x}$$

$\bar{x}$, $\bar{y}$ ... arithmetische Mittel

Bemerkung: Die Koeffizienten der Regressionsgeraden werden üblicherweise mithilfe von Technologieeinsatz berechnet (vergleiche Seiten 179 und 180).

B **7.39** ***1)*** Gib die Gleichung der Regressionsgeraden an.
2) Berechne die jeweils fehlenden Werte.
a) M = {(2; 4), (3; 3), (4; 5), (6; 6), (7; 8), (8; 7), (10; 9), (12; 13)}, A(9|y), B(x|10)
b) M = {(1; 10), (3; 12), (5; 15), (6; 14), (9; 18), (11; 23), (14; 27), (15; 30)}, A(x|12), B(8|y)

BC **7.40** Bei der Messung der Extinktion einer Verdünnungsreihe ergaben sich folgende Werte:

Konzentration in $\frac{mol}{L}$	0,01	0,05	0,1	0,5	1,0	1,5	2,0
Extinktion	0,003	0,016	0,03	0,15	0,3	0,45	0,6

1) Welche Konzentration erwartet man für eine Lösung mit der Extinktion E = 0,22?
2) Welche Extinktion erwartet man für eine Lösung mit der Konzentration c = 0,75 $\frac{mol}{L}$?

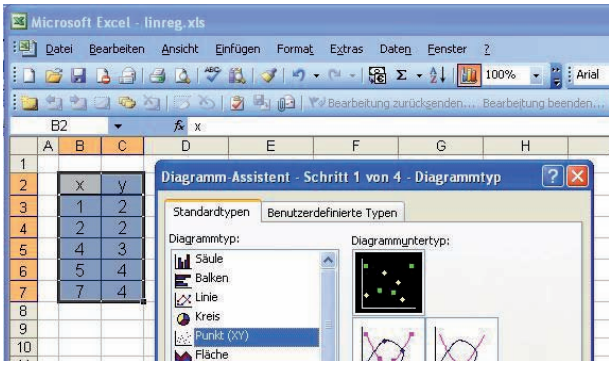
BC **7.41** Der Betriebsmanager eines Unternehmens, das Radios in Form von Computermäusen produziert, hat folgende anfallende Gesamtkosten bei der Produktion festgestellt:

Anzahl der Radios	100	200	300	400	500
Kosten in €	3 450,00	5 210,00	7 400,00	9 180,00	10 940,00

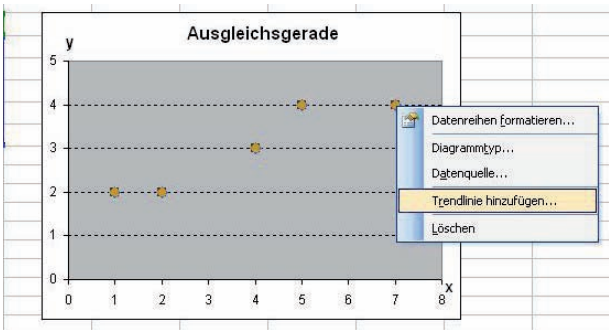
1) Bestimme mithilfe der Methode der kleinsten Quadrate eine lineare Kostenfunktion.
2) Wie hoch wären gemäß dieser Kostenfunktion die Gesamtkosten bei einer Produktion von 600 Radios?

Tabellenkalkulationsprogramm

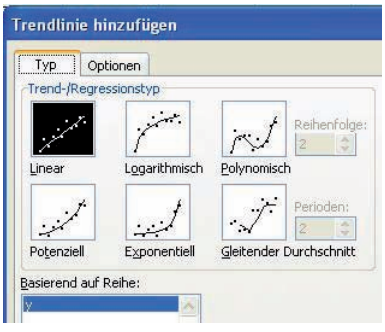
ZB: Regressionsgerade, gegeben sind: P(1|2), Q(2|2), R(4|3), S(5|4), T(7|4)



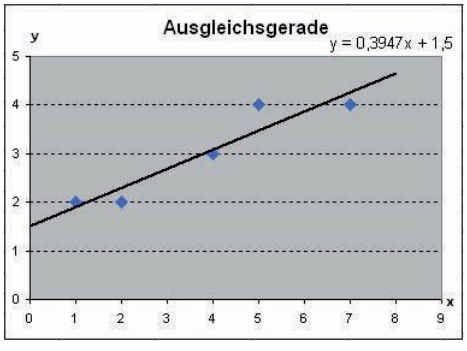
Die x- und y-Werte werden eingegeben und die Punkte in einem Diagramm (**Punkt (XY)**) dargestellt.
Das Diagramm kann dann wie gewünscht formatiert werden.



Um nun die Ausgleichsgerade berechnen und darstellen zu lassen, klickt man mit der rechten Maustaste auf einen Datenpunkt und wählt **Trendlinie hinzufügen ...** .



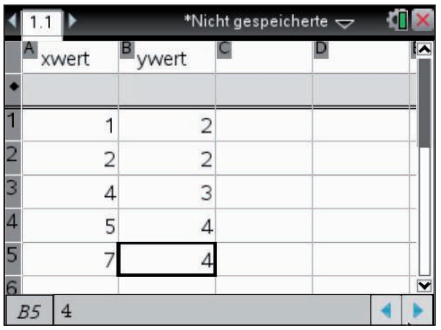
Als **Typ** wird **Linear** ausgewählt.
In den **Optionen** kann bei **Trend** angegeben werden, ob die Gerade „verlängert“ dargestellt wird, also einen Trend angibt. Zusätzlich wird **Gleichung im Diagramm darstellen** aktiviert.



Nach Bestätigen mit **OK** erhält man die Regressionsgerade. Diese wird sowohl grafisch dargestellt als auch als Gleichung angegeben.

TI-Nspire

ZB: Regressionsgerade, gegeben sind: P(1|2), Q(2|2), R(4|3), S(5|4), T(7|4)

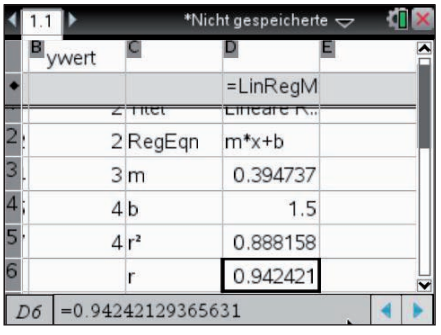


Die Messpunkte werden in der Applikation **Lists & Spreadsheet** eingegeben.

In die nun erscheinende Tabelle werden in der Spalte **A** die x-Werte und bei **B** die y-Werte der Punkte eingegeben. Um die Werte anschließend in einem Streudiagramm grafisch darstellen zu können, müssen die Spalten Namen erhalten.

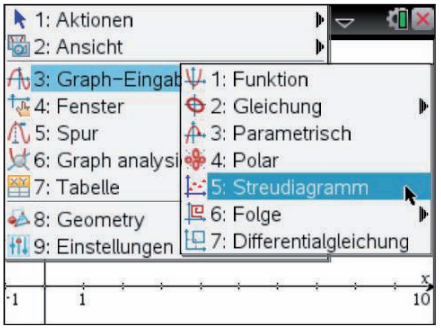
Die Berechnung der Regressionsgeraden erfolgt über Menü **4: Statistik**, **1: Statistische Berechnung**, **3: Lineare Regression (mx + b)** oder **4: Lineare Regression (a + bx)**.

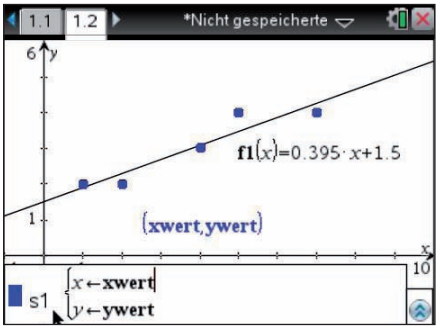
Bei **x-Liste:** wird der Name der Spalte mit den x-Werten, bei **y-Liste:** jener der y-Werte eingegeben. Unter **RegEqn speichern unter:** kann die entstehende Gerade gespeichert werden.



Nach Bestätigen mit **OK** werden in der Tabelle die berechneten Werte ausgegeben. Für die Gleichung der Regressionsgeraden sind die Werte **m** und **b** relevant. Weiters werden das Bestimmtheitsmaß r^2 und der Korrelationskoeffizeint **r** ausgegeben.

Wechselt man anschließend in die Applikation **Graphs**, so kann nach Aufruf der Funktion **f1** die Regressionsgerade dargestellt werden. Um die Messpunkte anzuzeigen, muss unter Menü **3: Graph-Eingabe/Bearbeitung**, **5: Streudiagramm** ausgewählt werden. In die Eingabezeile werden die Spaltennamen der Liste eingegeben.





Zusammenfassung

Statistik befasst sich mit der Erhebung, Auswertung und Darstellung von Daten. Man unterscheidet **metrische (quantitative) Merkmale, ordinale Merkmale (Rangmerkmale)** und **nominale (qualitative) Merkmale**.

Aus praktischen Gründen wird oft statt der **Grundgesamtheit** N nur eine Auswahl, eine **Stichprobe** n, verwendet.

Auswertung von Daten:

Häufigkeitsverteilung
Die absoluten, relativen oder prozentuellen Häufigkeiten von Datenmengen werden meist in Tabellenform angegeben.

Klassenbildung
Große Datenmengen werden zur Übersichtlichkeit klassifiziert und dabei in maximal 20 wenn möglich gleich breite Klassen zusammengefasst.

Lagemaße
Arithmetisches Mittel: Summe aller Werte, dividiert durch deren Anzahl
Median: Mittlerer Wert der geordneten Liste
Quartile: Teilen die Liste in vier Bereiche zu ca. 25 %
Modalwert: Häufigster Wert einer Liste

Streuungsmaße
Spannweite (Range): Differenz zwischen Maximum und Minimum
Interquartilsabstand: Differenz zwischen den Quartilen q_3 und q_1
Varianz σ^2 bzw. s^2: Mittlere quadratische Abweichung vom Mittelwert

$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$ bei Grundgesamtheiten bzw. $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ bei Stichproben

Standardabweichung = $\sqrt{\text{Varianz}}$

Grafische Darstellung: **Säulendiagramm, Pareto-Diagramm, Histogramm, Boxplot**

Korrelation und Regression

Pearson'scher Korrelationskoeffizient (Empirischer Korrelationskoeffizient)

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

x_i, y_i ... Koordinaten von i Messpunkten
$\bar{x}$, $\bar{y}$ ... arithmetische Mittel

Regressionsgerade $y = a \cdot x + b$

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{und} \quad b = \bar{y} - a \cdot \bar{x}$$

x_i, y_i ... Koordinaten von i Messpunkten
$\bar{x}$, $\bar{y}$ ... arithmetische Mittel

Weitere Aufgaben

C **7.42** Gib an, welche der genannten Merkmale metrische Merkmale, Rangmerkmale bzw. nominale Merkmale sind: Größenklassen von Hühnereiern, Geburtsgewicht von Säuglingen, Staatsangehörigkeit, Sterne von Hotels, Inflationsraten von Staaten, Religionszugehörigkeit

B **7.43** Der Durchmesser von Werkstücken wurde gemesessen (Werte in Millimeter):

73	70	68	73	63	67	71	67	69	73
64	65	75	69	74	65	68	69	64	72
66	78	69	75	68	77	69	74	68	65
67	71	68	64	75	67	63	70	69	71

1) Erstelle eine Tabelle mit der Häufigkeitsverteilung, ermittle auch die Häufigkeitssummen.
2) Erstelle ein Säulendiagramm und stelle die Häufigkeitssummen grafisch dar.
3) Berechne folgende Lagemaße: Mittelwert, Median, Quartile q_1 und q_3
4) Erstelle einen Boxplot und gib die Spannweite und den Interquartilsabstand an.
5) Berechne die Varianz und die Standardabweichung.

BC **7.44** Folgende Statistik zeigt die Anzahl der Touristen in einer Kleinstadt in den Jahren 2002 und 2012.

	2002	2012
Touristen	121 731	234 287
Inländische Gäste	21 456	32 017
Ausländische Gäste aus	100 275	202 270
Deutschland	40 130	77 402
Frankreich	7 455	12 898
Großbritannien	6 989	11 697
Italien	12 961	34 002
Japan	4 108	12 214
Schweden	4 477	4 637
Schweiz und Liechtenstein	6 802	13 974
Spanien	3 514	13 517
USA	13 208	18 655
andere	631	3 274

Erstelle für jedes der beiden Jahre ein Pareto-Diagramm und interpretiere das Ergebnis.

AB **7.45** 100 Personen wurden während einer Diät medizinisch betreut und ihre Abnehmerfolge aufgezeichnet (Werte in kg).

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

1) Nimm eine Klasseneinteilung vor und erstelle eine Häufigkeitstabelle und ein Histogramm.
2) Ermittle aus den klassifizierten Daten das arithmetische Mittel und die Varianz.

7.46 Gegeben sind folgende Messpunkte: P(2|3), Q(4|7), R(5|8), S(6|10) und T(7|11)
Bestimme die Gleichung der Regressionsgeraden. Zeichne die Punkte und die ermittelte Gerade in ein Koordinatensystem ein.

B

7.47 Aus einer Versuchsreihe erhält man folgende Messwerte:

BC

x	1	2	3	4	5	6	7	8	9
y	9	8	?	6	4	4	2	1	0

1) Berechne die Gleichung der Ausgleichsgeraden.
2) Bestimme den y-Wert für x = 3.

7.48 Eine Firma produziert einen Spezialstoff für Ballkleider. Die Gesamtkosten K(x) für eine tägliche Produktionsmenge x betragen:

BC

x in m	10	30	40	70	90
K in €	1 500,00	2 000,00	3 000,00	4 000,00	5 000,00

1) Ermittle die Gleichung der angenäherten linearen Kostenfunktion.
2) Welche Kosten werden bei einer Produktionsmenge von 55 m Stoff erwartet?

7.49 Beim Testen eines Schiffsmotors wurde dessen Leistung P in Abhängigkeit von der Drehzahl n gemessen. Dabei ergaben sich die folgenden Werte:

BC

Drehzahl in $\frac{1}{\text{min}}$	2 400	2 800	3 100	3 800	4 200
P in kW	18,4	24,8	30,6	38,5	42,4

1) Ermittle die Ausgleichsgerade.
2) Welche Leistung kann man bei 1 200 Umdrehungen erwarten?

7.50 Bei einer Verdünnungsreihe wurden folgende Extinktionen E (Absorption des Lichts bei bestimmten Wellenlängen) bei einer Wellenlänge von 440 nm gemessen:

BC

$c_{\text{Lösung}}$ in $\frac{\text{mol}}{\ell}$	0,1	0,05	0,01	0,005	0,001
E	0,60	0,32	0,18	0,08	0,04

Bestimme die Konzentration einer Lösung mit E = 0,25, wenn der Zusammenhang zwischen Extinktion und Konzentration durch eine lineare Funktion beschrieben werden kann.

7.51 Bei einem Axialventilator wurden die Totaldruckerhöhungen Δp bei unterschiedlichen Volumenströmen V an 4 Messpunkten gemessen:

BC

	1	2	3	4
Volumenstrom V in $\frac{\text{m}^3}{\text{s}}$	0,0	17,8	26,9	35,5
Totaldruckerhöhung Δp in Pa	2 454	1 955	1 651	1 013

1) Gib die Gleichung der Regressionsgeraden zur Annäherung der Kennlinie des Axialventilators an.
2) Ermittle den Korrelationskoeffizienten zu dieser Datenmenge.
3) Gib an, ob es empfehlenswert ist, mit der Regressionsgeraden zu arbeiten.

Wissens-Check

		gelöst
1	Welche Merkmalsarten von Daten kennst du? Gib jeweils ein Beispiel an.	
2	Gib an, ob das Merkmal diskret oder stetig ist. A) Anzahl Handys pro Haushalt B) Temperaturwerte C) Fußgrößen D) Schuhgrößen	
3	Ein Pareto-Diagramm ordnet die Häufigkeiten ...	
4	Ich kenne die Richtlinien für eine Klasseneinteilung und weiß, wie die grafische Darstellung heißt. Zum Beispiel würde ich für 100 Daten ... Klassen wählen.	
5	Warum sind mindestens 50 % aller Werte kleiner gleich dem Median?	
6	Was kann man aus einem Boxplot ablesen?	
7	Bei welcher Merkmalsart ist der Modalwert das einzige Lagemaß, das ermittelt werden kann? Nenne ein Beispiel.	
8	Welchen Vorteil hat der Interquartilsabstand gegenüber der Spannweite? Erkläre dies anhand folgender Datenreihe. 102,6; 101,5; 101,0; 102,1; 80,0; 102,2; 101,9; 102,4; 102,7; 102,8; 102,1; 101,9	
9	Warum ist die Summe aller Differenzen vom Mittelwert als Streuungsmaß ungeeignet?	
10	Gib die unterschiedlichen Formeln für die Varianz einer Grundgesamtheit und einer Stichprobe an.	
11	Berechne die Standardabweichung s der Daten: 23,6; 22,8; 21,9; 24,8; 23,9; 23,5; 23,3; 22,8; 23,8; 21,9 Wäre der Wert für σ größer oder kleiner? Begründe deine Antwort.	
12	Ermittle für 10 Autos das arithmetische Mittel und den Median des Benzinverbrauchs (Werte in Liter pro 100 km). Welcher Wert beschreibt die Verteilung besser? Begründe kurz deine Antwort. 6,6; 3,9; 8,7; 3,8; 24,1; 7,5; 8,4; 7,3; 9,4; 21,5	

Lösung:
1) siehe Seite 161 2) A) diskret, B) stetig, C) stetig, D) diskret 3) fallend an.
4) siehe Seite 166; Histogramm; 10 5) Es gibt zwei Möglichkeiten: Bei einer geraden Anzahl von Werten ist der Median nicht Teil der Urliste und es sind genau 50 % der Werte kleiner. Bei einer ungeraden Anzahl ist der Median Teil der Urliste und es bleiben somit weniger als 50 % der Werte, die größer sind, übrig. 6) siehe Seite 169
7) siehe Seite 170 8) Der Interquartilsabstand ist nicht so anfällig für den Ausreißer 80,0.
9) siehe Aufgabe 7.22 10) siehe Seite 172f 11) $s = 0{,}904...$; σ ist kleiner ($\sigma = 0{,}857...$), da das Ergebnis wegen der Division durch $(n - 1)$ größer ist. 12) $\bar{x} = 10{,}12$; $\tilde{x} = 7{,}95$; Der Median, da der Mittelwert durch die zwei Ausreißer nach oben verschoben wird.