11. STATISTIK

11.1. Begriffsbestimmung

Die Statistik ist wie auch die Wahrscheinlichkeitsrechnung ein Wissensgebiet der sogenannten Stochastik. Die Stochastik kann man als die Lehre von zufälligen Vorgängen bzw. Ereignissen beschreiben.

Als „zufällige Ereignisse“ bezeichnet man Vorgänge, deren Ausgänge nicht genau vorhersagbar sind. Es ist nicht einmal sicher, ob die Ereignisse eintreten werden. Der Begriff „zufällig“ im Sinn der Stochastik muß jedoch genau festgelegt werden.

Eine **zufällige** Auswahl ist eine Auswahl, bei der jedes Element die gleiche Chance hat, ausgewählt zu werden; d.h. keines der Elemente darf bevorzugt oder benachteiligt werden. Das Ergebnis darf nicht von subjektiven Eindrücken des Auswählenden abhängen.

Beispiele:

- *Es dürfen repräsentative Umfragen in Haushalten nicht nur vormittags durchgeführt werden, da zu diesem Zeitpunkt nur ein bestimmter Personenkreis (Berufstätige fehlen) erfaßt würde.*

- *Eine zufällige Auswahl aus einer Personengruppe sollte nicht aus den Personen direkt getroffen werden (sympathisches oder weniger sympathisches Aussehen), sondern z.B. durch Zuordnen von Nummern und Ziehen aus einer Urne getätigt werden.*

Die Statistik kann nun folgendermaßen beschrieben werden:

Die Statistik ist die **Gesamtheit aller Methoden** zur Untersuchung von Massenerscheinungen und umfaßt die Bereiche beschreibende Statistik und beurteilende Statistik.

Die beschreibende Statistik hat die Aufgabe, Datenmaterial zu sammeln, zu ordnen, übersichtlich darzustellen und daraus bestimmte Kennzahlen zu berechnen. Weiters sollen aus dem gesammelten Datenmaterial einer möglichst umfangreichen Stichprobe Wahrscheinlichkeiten für die Gesamtheit geschätzt werden. Somit ist die beschreibende Statistik eine Hilfswissenschaft der Wahrscheinlich-keitsrechnung.

Die beurteilende Statistik hat die Aufgabe, mit Hilfe der Wahrscheinlichkeitsrechnung abzuschätzen, wie gerechtfertigt ein Rückschluß aus einer Stichprobe auf die Gesamtheit ist („Testen von Hypothesen"). Somit ist in diesem Fall die Wahrscheinlichkeitsrechnung eine Hilfswissenschaft der beurteilenden Statistik.

Im folgenden Abschnitt wird nur auf die beschreibende Statistik eingegangen.

11.2. Methoden der Statistik

(a) Erhebung und Aufbereitung von Datenmengen

In jeder statistischen Untersuchung muß eine große Zahl von Daten erhoben und ausgewertet werden. Dies kann auf zwei Arten erfolgen; nämlich einerseits durch sekundärstatistische Erhebungen, d.h. Zurückgreifen auf bereits vorhandene Daten aus statistischen Jahrbüchern, amtlichen Statistiken, Fachliteratur, etc. oder andererseits durch primärstatistische Erhebungen, wenn z.B. neue Untersuchungen nötig sind.

***Beispiele**:*

- *In einer Schule sollen zur Durchführung von Schilanglaufkursen zentral die Ausrüstungen angeschafft werden. Um zu wissen, wieviele Stück pro Größe für die nächsten Jahre gekauft werden sollen, wird ein Fragebogen an alle 560 Schüler ausgegeben, in dem nach Körpergröße und Schuhnummer gefragt wird.*

- *Die durchschnittliche Lebensdauer einer neuentwickelten Glühbirne soll bestimmt werden. Dazu werden 1000 Glühbirnen einer Dauerbelastung unterzogen und aus den gemessenen Brennzeiten die durchschnittliche Lebensdauer errechnet.*

- *Ein neues Medikament gegen Allergien soll auf den Markt gebracht werden. Um dessen Wirksamkeit zu testen, wird eine Gruppe von 300 Allergikern gezielt unterschiedlichen Dosen ihres Allergens ausgesetzt und anschließend die Wirkung auf Augenrötung und Nasenschleimhautschwellung mit und ohne Medikament genau beobachtet und aufgezeichnet. Die Methoden der Datenerhebung sind dabei: schriftliche oder mündliche Befragung; Experiment (Messung); Beobachtung.*

Aus jedem der obigen Beispiele können nun Aussagen gewonnen werden, die bestenfalls so gut sind, wie die Daten, auf die sie sich beziehen. Aus ungenauen Erhebungen sind keine sinnvollen Aussagen möglich. So müssen z.B. beim Glühbirnentest alle Lampen mit gleicher Spannung und Stromstärke versorgt werden.

Allen Beispielen ist folgendes gemeinsam:

In jeder Datenerhebung wird eine Gesamtheit von n Elementen auf ein (oder mehrere) Merkmal(e) x, (y, ...) hin untersucht.

***Beispiele**:*

Gesamtheit n:	*Merkmal x (Variable)*
- Menge von 560 Schülern n = 560	*Körpergröße, Schuhnummer*
- Menge von 1000 Glühbirnen n = 1000	*Brenndauer*
- Menge von 300 Allergikern n = 300	*Augenrötung, Schleimhautschwellung*

(b) Ordnen der Daten und Ermitteln von Häufigkeiten

Zum Ordnen der Daten muß zunächst festgestellt werden, ob das untersuchte Merkmal x in einer endlichen Zahl (diskrete Variable) oder unendlichen bzw. sehr großen Zahl (kontinuierliche Variable) von Merkmalausprägungen (Variablenwerten) vorkommt. Am Beginn des Datenordnens steht das Ordnen der Daten nach ihrer Größe, sofern es sich um quantitative Daten handelt.

Zusätzlich ist es meist notwendig, eine Klasseneinteilung durchzuführen. Eine Klasseneinteilung ist die Unterteilung des Variablenwertesbereichs in zueinander elementfremde Teilbereiche. Ein solcher Teilbereich wird als Klasse bezeichnet. Bei quantitativen Daten nennt man die Differenz zwischen größtem und kleinsten Wert einer Klasse die Klassenbreite. Bei Klasseneinteilungen müssen die einzelnen Klassen darüberhinaus nicht gleiche Klassenbreite aufweisen.

Die Einteilung in elementfremde Bereiche einer Datenmenge bezeichnet man als **Klasseneinteilung**.

***Beispiele**:*

- *In der Befragung von 560 Schülern einer Schule nach ihren Schuhgrößen traten die Größen 36 bis 46 auf. D.h. das Merkmal „Schuhgröße" trat in 11 Ausprägungen in den Variablenwerten $x_1 = 36$; $x_2 = 37$; $x_3 = 38$; ... , $x_{11} = 46$ auf und somit in einer endlichen Anzahl.*

- *Bei der Brennzeit von Glühbirnen können beliebige Zeiten zwischen wenigen Minuten und über 15000 Stunden auftreten, bis ein Defekt eintritt. Man muß daher zunächst eine Mindestbrenndauer festlegen, unter der alle Testglühbirnen als defekt gelten und daher für die Bewertung der Brenndauer einer intakten Glühbirne nicht in Betracht kommen und aus der Wertung genommen werden. Für die verbleibenden n Brennzeiten, die in die Statistik aufgenommen werden, wird es auch nicht sinnvoll sein, jede Zeit als eigene Merkmalsausprägung auszuwerten, sondern man wird eine geeignete Einteilung (Klasseneinteilung) vornehmen. Das Merkmal Brenndauer wird im folgenden mit x bezeichnet.*

Klasse 1	*13000 Std. $\leq x <$ 13200 Std.*
Klasse 2	*13200 Std. $\leq x <$ 13400 Std.*
Klasse 3	*13400 Std. $\leq x <$ 13600 Std.*
Klasse 4	*13600 Std. $\leq x <$ 13800 Std.*
Klasse 5	*13800 Std. $\leq x <$ 14000 Std.*
...	*...*
Klasse 15	*15800 Std. $\leq x <$ 16000 Std.*

Somit hat man die zunächst überaus zahlreichen Zeiten (aber nicht unendlich vielen) auf 15 Klassen, d.h. 15 Variablenwerte, reduziert. Zweckmäßigerweise wählt man dazu eine konstante Klassenbreite - hier 200 Stunden.

Liegen nach Auswertung der Meßergebnisse vielleicht in den Klassen 11 bis 15 nur noch sehr wenige Elemente (Glühbirnenbrenndauern), so könnten diese zu einer Klasse 11 mit $x \geq 15000$ Std. zusammengefaßt werden. Auf diese unterschiedliche Klassenbreite muß jedoch in einer graphischen Darstellung des Sachverhalts Rücksicht genommen werden, um nicht zu einer falschen Interpretation des Ergebnisses zu kommen.

- *Der Grad der Augenrötung von untersuchten Allergikern tritt kontinuierlich in unendlich vielen Rottönen (unendliche Variablenwertanzahl, also kontinuierliche Variable) auf und kann daher nur auf Grund der Meßgenauigkeit der optischen Analysegeräte in Klassen eingeteilt werden, z.B.*

Klasse 1	*keine Rötung*
Klasse 2	*leichte Rötung*
Klasse 3	*mittlere Rötung*
Klasse 4	*starke Rötung*

D.h. die Variable x = „Augenrötung“ tritt somit nur noch in 4 Variablenwerten x_1, x_2, x_3, x_4 auf.

Ist das Einteilen und Ordnen der Daten aus der sogenannten Urliste abgeschlossen, kann mit der Auszählung der einzelnen Daten mit jeweils einer bestimmten Merkmalsausprägung x_i ($i \in \mathbb{N}$, $1 \leq i \leq k$; k ist die Anzahl der verschiedenen Variablenwerte des Merkmals x) begonnen werden. Diese Anzahl wird als absolute Häufigkeit H_i bezeichnet.

Die Anzahl der Elemente mit jeweils der gleichen Merkmalsausprägung wird als die **absolute Häufigkeit H_i** bezeichnet.

Die Ermittlung der absoluten Häufigkeit kann für die einzelnen Daten oder aber auch für die einzelnen Klassen erfolgen.

***Beispiel**:* *ad Beispiel Schuhgröße*

x_i ... Schuhgröße		H_i ... Schülerzahl	
x_1	36	6	H_1
x_2	37	45	H_2
x_3	38	82	H_3
x_4	39	98	H_4
x_5	40	87	H_5
x_6	41	84	H_6
x_7	42	51	H_7
x_8	43	23	H_8
x_9	44	48	H_9
x_{10}	45	31	H_{10}
x_{11}	46	5	H_{11}
insgesamt			$560 = n$

Bei diesem Vorgang gilt:

Die Summe aller absoluten Häufigkeiten muß die Anzahl der Elemente der Gesamtheit ergeben: $\sum_{i=1}^{k} H_i = n$

Die absolute Häufigkeit eines bestimmten Variablenwertes ist als alleinige Angabe ohne die Kenntnis der Gesamtzahl n nicht aussagekräftig, denn 90 von 100 sind sehr viele, 90 von 1000 sind relativ wenige; daher sagt die Zahl 90 alleine nichts aus. Aus diesem Grund wählt man für die Angabe der Häufigkeit einer Merkmalsausprägung x_i üblicherweise die relative Häufigkeit h_i .

Die relative Häufigkeit beträgt: $h_i = \frac{H_i}{n}$ $0 \le h_i \le 1$

Häufig wird die relative Häufigkeit auch in Prozent von der Gesamtheit n ausgedrückt:

Die relative prozentuelle Häufigkeit beträgt: $h_i(\%) = \frac{H_i}{n} \cdot 100$ $0\% \le h_i(\%) \le 100\%$

***Beispiele**:*

- *ad Beispiel Schuhgröße*

$$h_1 = \frac{H_1}{560} = \frac{6}{560} = 0{,}0107 \mathrel{\hat{=}} 1{,}07\%$$
$$h_2 = \frac{H_2}{560} = \frac{45}{560} = 0{,}0804 \mathrel{\hat{=}} 8{,}04\%$$
$$h_3 = \frac{H_3}{560} = \frac{82}{560} = 0{,}1464 \mathrel{\hat{=}} 14{,}64\%$$
$$h_4 = \frac{H_4}{560} = \frac{98}{560} = 0{,}175 \mathrel{\hat{=}} 17{,}50\%$$
$$h_5 = \frac{H_5}{560} = \frac{87}{560} = 0{,}1554 \mathrel{\hat{=}} 15{,}54\%$$
$$h_6 = \frac{H_6}{560} = \frac{84}{560} = 0{,}15 \mathrel{\hat{=}} 15{,}00\%$$
$$h_7 = \frac{H_7}{560} = \frac{51}{560} = 0{,}0911 \mathrel{\hat{=}} 9{,}11\%$$
$$h_8 = \frac{H_8}{560} = \frac{23}{560} = 0{,}0411 \mathrel{\hat{=}} 4{,}11\%$$
$$h_9 = \frac{H_9}{560} = \frac{48}{560} = 0{,}0857 \mathrel{\hat{=}} 8{,}57\%$$
$$h_{10} = \frac{H_{10}}{560} = \frac{31}{560} = 0{,}0554 \mathrel{\hat{=}} 5{,}54\%$$
$$h_{11} = \frac{H_{11}}{560} = \frac{5}{560} = 0{,}0089 \mathrel{\hat{=}} 0{,}89\%$$
$$\sum_{i=1}^{11} h_i = 1 \mathrel{\hat{=}} 100\%$$

- *ad Beispiel Brenndauer*

Brenndauer in Stunden x_i	H_i	h_i(%)
x_1 = Klasse 1 = [13000;13200)	7	0,007 ≙ 0,7%
x_2 = Klasse 2 = [13200;13400)	8	0,008 ≙ 0,8%
x_3 = Klasse 3 = [13400;13600)	11	0,011 ≙ 1,1%
$x_4 = K_4$.	42	0,042 ≙ 4,2%
$x_5 = K_5$.	65	0,065 ≙ 6,5%
$x_6 = K_6$.	96	0,096 ≙ 9,6%
$x_7 = K_7$.	104	0,104 ≙ 10,4%
$x_8 = K_8$.	173	0,173 ≙ 17,3%
$x_9 = K_9$.	105	0,105 ≙ 10,5%
$x_{10} = K_{10}$.	143	0,143 ≙ 14,3%
$x_{11} = K_{11}$.	108	0,108 ≙ 10,8%
$x_{12} = K_{12}$.	72	0,072 ≙ 7,2%
$x_{13} = K_{13}$.	34	0,034 ≙ 3,4%
$x_{14} = K_{14}$.	22	0,022 ≙ 2,2%
x_{15} = Klasse 15 = [15800;16000)	10	0,010 ≙ 1,0%
$n = \sum_{i=1}^{15} H_i = 1000$		$\sum_{i=1}^{15} h_i = 1 \mathrel{\hat{=}} 100\%$

- *ad Beispiel Allergiker*

	ohne Einnahme des Medikamentes		*nach Einnahme des Medikamentes*	
Augenrötung x_i	*H_i*	*$h_i \triangleq h_i(\%)$*	*H_i*	*$h_i \triangleq h_i(\%)$*
x_1 keine Rötung	*3*	*0,01 ≙ 1%*	*42*	*0,14 ≙ 14%*
x_2 leichte Rötung	*12*	*0,04 ≙ 4%*	*102*	*0,34 ≙ 34%*
x_3 mittlere Rötung	*129*	*0,43 ≙ 43%*	*96*	*0,32 ≙ 32%*
x_4 starke Rötung	*156*	*0,52 ≙ 52%*	*60*	*0,2 ≙ 20%*
	n = 300	*1 ≙ 100%*	*n = 300*	*1 ≙ 100%*

Abgesehen von ungenauen Rundungen muß immer gelten:

$$\sum_{i=1}^{k} h_i = 1 \qquad \sum_{i=1}^{k} h_i(\%) = 100\%$$

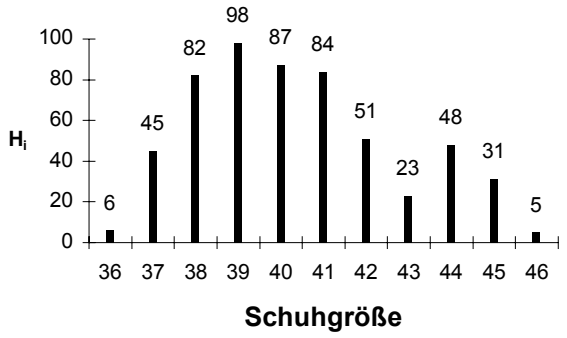
(c) Graphische Darstellungen

Die relativen Häufigkeiten lassen sich in verschiedenen Diagrammen, sogenannten Histogrammen, darstellen, um einen Überblick über die Häufigkeitsverteilung zu gewinnen.

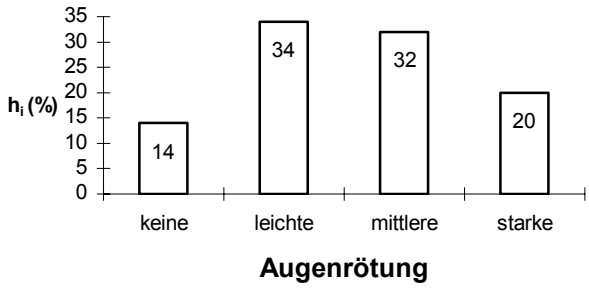
Stabdiagramm

***Beispiele**:*

Anzahl der Schüler mit der jeweiligen Schuhgröße

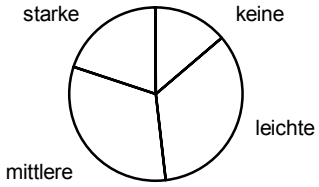


Augenrötung nach Einnahme des Medikamentes



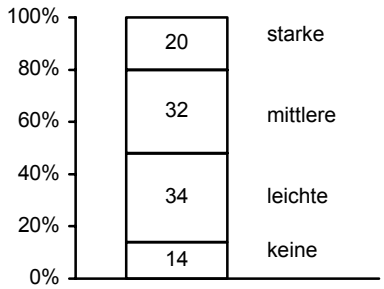
Kreisdiagramm

***Beispiel**:*
Augenrötung nach Einnahme des Medikamentes



Säulendiagramm

***Beispiel**:*
Augenrötung nach Einnahme des Medikamentes



In allen Darstellungen sind die Höhen der Rechtecke, Länge der „Stäbe“, Winkel der Kreisausschnitte bzw. Höhen der Säulenabschnitte proportional zu den Häufigkeiten.

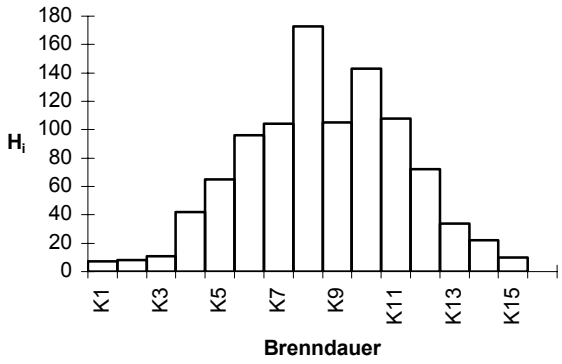
Längen und Winkel in Histogrammen sind proportional zu den Häufigkeiten.

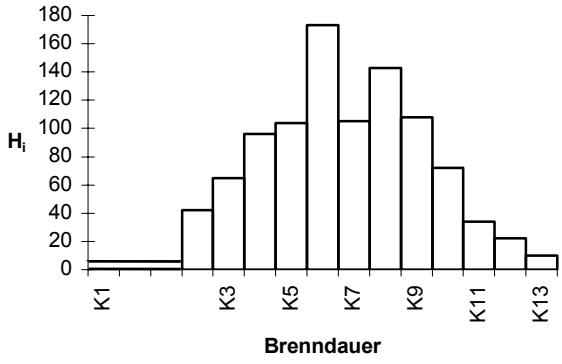
Sind die Variablenwerte x_i Klassen von quantitativen Daten wie im Beispiel Brenndauer und wählt man als graphische Darstellung der Häufigkeiten Rechtecke, wobei die Rechtecksbreite maßstabsgetreu der Klassenbreite entspricht, so sind die Häufigkeiten nur dann proportional zu den Rechteckshöhen, wenn die Klassen alle gleich breit sind. Bei unterschiedlicher Klassenbreite muß immer beachtet werden, daß die Flächeninhalte der Rechtecke in einem Histogramm proportional zu den entsprechenden Häufigkeiten sein müssen.

Histogramm bei Klasseneinteilung	$\text{Höhe} = \frac{\text{Häufigkeit}}{\text{Klassenbreite}}$

***Beispiel**:* *ad Beispiel Brenndauer*

Die Klassenbreiten haben alle die gleiche Breite, nämlich 200 Stunden. Im Histogramm sind die Höhen aller Rechtecke also proportional zu den Häufigkeiten. Faßt man die ersten drei Klassen zusammen, also zu [13000;13600), zu sind in dieser neuen Klasse mit der Klassenbreite 600 nun 26 Lampen enthalten. Im Histogramm ist die Höhe der neuen Klasse jedoch nicht 26 Einheiten hoch einzuzeichnen, sondern durch 3 zu dividieren, da die Klasse dreimal so breit ist.





Das zweite Histogramm weist aufgrund der Zusammenfassung dreier Klassen nur mehr 13 Klassen auf. Im speziellen ist die Klasse K_1 nur 26 : 3 = 8,67 Einheiten hoch gezeichnet.

Wird bei der Darstellung in Histogrammen die Klassenbreite nicht entsprechend berücksichtigt, so vermitteln die Histogramme einen falschen Eindruck. Oftmals wird dies jedoch bewußt zur Manipulation des Betrachters verwendet.

11.3. Zentralmaße

Meist versucht man in der Statistik die Vielzahl der aufgenommenen Daten durch eine Zahl zu ersetzen, welche die ganze Liste möglichst gut repräsentiert. Solche Zahlen bezeichnet man als Zentralmaße. Es gibt verschiedene solche Zentralmaße: **Minimum, Maximum, Spannweite, Modus, Median, Quartilen** und diverse **Mittelwerte.** Im folgenden wird zwischen direkt ablesbaren Zentralmaßen, die ohne Berechnung aus der Datenmenge ermittelt werden können, und den Mittelwerten, die sich erst nach Berechnung ergeben, unterschieden.

(a) Direkt ablesbare Zentralmaße

Das **Minimum min** ist naheliegender Weise der kleinste Wert der Datenmenge, das **Maximum max** entsprechend der größte Wert der Datenmenge. Als **Spannweite S** bezeichnet man die Differenz zwischen Maximum und Minimum.

Der **Modus** oder **Modalwert M** ist der Variablenwert mit der größten Häufigkeit. Er wird dann eine Liste gut repräsentieren, wenn die Häufigkeit des Modus viel größer als die Häufigkeit der übrigen Werte ist und außerdem die meisten auftretenden Variablenwerte in der Nähe des Modus liegen.

Der **Median** oder **Zentralwert Z** ist der in der Mitte stehende Variablenwert der der Größe nach geordneten Liste der Variablenwerte. Der Median ist also nicht geeignet als Zentralmaß für rein qualitative Variablen (z.B. Haarfarbe von n Personen). Zu Ermittlung muß man zunächst alle n Werte der Größe nach ordnen, wobei gleiche Werte mehrmals ihrer Häufigkeit entsprechend angeschrieben werden $x_1 \le x_2 \le x_3 \le ... \le x_n$. Dann gilt:

$$Z = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \text{ falls n gerade bzw. } Z = x_{\frac{n+1}{2}} \text{ falls n ungerade}$$

Bei einer geraden Anzahl von Werten ergibt sich also der Median aus dem Mittelwert der beiden in der Mitte stehenden Werte. Bei einer ungeraden Anzahl von Werten ist der Median der in der Mitte stehende Wert.

Der Median teilt die Liste aller Werte in zwei gleich große Teile, nämlich in die Menge der darunter-liegenden und die der darüberliegenden Werte.

Manchmal ist es von Interesse, die Datenliste nicht nur in zwei gleich große Teile zu teilen, sondern in vier Bereiche, in denen jeweils gleich viele Werte liegen. Die Grenzen dieser Viertel sind durch die sogenannten

Quartilen gegeben. Der untere Quartil oder 1. Quartil Q_1 ist der Median der 1. Hälfte der Werte; der obere Quartil oder 3. Quartil Q_3 ist der Median der 2. Hälfte der Werte.

***Beispiel**:* *Bei einem Wettrennen von 20 Läufern wurden folgende Zeiten gemessen (Liste in geordneter aufsteigender Reihenfolge).*

9,2s; 9,3s; 9,3s; 9,4s; 9,4s; 9,5s; 9,6s; 9,6s; 9,6s; 9,8s; 9,9s; 9,9s; 10,0s; 10,0s; 10,0s; 10,0s; 10,1s; 10,1s; 10,2s; 10,3s.

Ein Läufer mit der Zeit 9,6s will wissen, ob er im besten Viertel liegt. Ermitteln Sie zusätzlich alle bisher bekannten Zentralmaße.

In diesem Fall sucht man also die Quartilen. Da n=20 gerade ist, ist der Median der Mittelwert zwischen 10. und 11. Wert der Liste, also

$$Z = ½ \cdot (x_{10} + x_{11}) = ½ \cdot (9,8 + 9,9) = 9,85$$

Der 1. Quartil ist der Median der unteren Hälfte, und da nun n=10, ist der 1. Quartil der Mittelwert zwischen 5. und 6. Wert der Liste, also

$$Q_1 = ½ \cdot (x_5 + x_6) = ½ \cdot (9,4 + 9,5) = 9,45$$

Der Läufer liegt nicht im besten Viertel.

Der Läufer liegt im zweiten Viertel, d.h. in der besseren Hälfte, aber nicht im besten Viertel. Er ist aber wesentlich besser als der Modus M = 10,0s. Abschließend die fehlenden Zentralmaße:

$$Q_3 = ½ \cdot (x_{15} + x_{16}) = ½ \cdot (10,0 + 10,0) = 10,0$$

$$min = 9,2;\ max = 10,3;\ S = 1,1;\ M = 10$$

Zu den bisherigen Zentralmaßen ist anzumerken, daß kein Wert alleine die Datenliste gut repräsentieren kann. Erst durch das Wissen mehrerer Zentralmaße kann man einen vereinfachten Überblick über die Datenliste erhalten.

Zur Repräsentation einer Datenliste von quantitativen Daten durch einen alleinigen Wert verwendet die Statistik meist einen der sogenannten Mittelwerte. Da sich diese aus der Datenliste errechnen lassen und daher jeder Wert der Datenliste verwendet wird, sind diese ungleich repräsentativer als die bisherigen Zentralmaße. Trotzdem muß klar bleiben, daß ein einzelner Wert keinen Überblick über eine Datenliste geben kann.

(b) Mittelwerte

Zur Repräsentation einer Datenliste von quantitativen Daten verwendet die Statistik meist den **arithmetischen Mittelwert** $\bar{x}$**.** Die Merkmalsausprägungen müssen durch Zahlen angegeben sein, welche nicht nur Verschiedenartigkeit und Rangordnung ausdrücken, sondern mit deren Hilfe auch Abstände zwischen den Merkmalsausprägungen angegeben werden können.

Man nennt $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ das **arithmetische Mittel** der reellen Zahlen x_i.

Zur Berechung des arithmetischen Mittels werden also alle Wert addiert und diese Summe durch die Anzahl der Werte dividiert. Sind die x_i nicht alle verschieden, sondern treten mehrere gleiche Daten auf, so fertigt man zunächst eine Häufigkeitstabelle an.
Sind dann x_1, x_2, ..., x_k alle verschiedenen Variablenwerte mit den absoluten Häufigkeiten H_1, H_2, ..., H_k und den relativen Häufigkeiten h_1, h_2, ..., h_k, dann gilt:

$$\bar{x} = \frac{x_1 \cdot H_1 + x_2 \cdot H_2 + \ldots + x_k \cdot H_k}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot H_i \quad \text{bzw.}$$

$$\bar{x} = x_1 \cdot \frac{H_1}{n} + x_2 \cdot \frac{H_2}{n} + \ldots + x_k \cdot \frac{H_k}{n} = x_1 \cdot h_1 + x_2 \cdot h_2 + \ldots + x_k \cdot h_k = \sum_{i=1}^{k} x_i \cdot h_i$$

Das arithmetische Mittel ist vor allem dann das geeignete Zentralmaß, wenn es um Summenbildung geht,

denn $\bar{x}$ ist jene Zahl, für die gilt: $n \cdot \bar{x} = x_1 \cdot H_1 + x_2 \cdot H_2 + \ldots + x_k \cdot H_k = \sum_{i=1}^{n} x_i$

Das arithmetische Mittel ist also jene Zahl, die man n-mal addieren könnte, um die gleiche Summe aller tatsächlichen Werte x_1, x_2, x_3, ..., x_n zu erhalten.

Die Formel für den arithmetischen Mittelwert unter Berücksichtigung der Häufigkeiten der Variablenwerte wird auch **gewogenes arithmetisches Mittel** der Variablenwerte x_1, x_2, x_3, ..., x_k mit den **Gewichten** h_1, h_2, h_3, ..., h_k genannt.

Gewogenes arithmetisches Mittel: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot H_i = \sum_{i=1}^{k} x_i \cdot h_i$

***Beispiele**:*

- *ad Beispiel Schuhgröße*

n = 560 Schuhgrößen von x_1 = 36 bis x_{11} = 46

$$\bar{x} = \frac{6 \cdot 36 + 45 \cdot 37 + 82 \cdot 38 + 98 \cdot 39 + 87 \cdot 40 + 84 \cdot 41 + 51 \cdot 42 + 23 \cdot 43 + 48 \cdot 44 + 31 \cdot 45 + 5 \cdot 46}{560}$$

$\bar{x}$ *= 40,38*

- *ad Beispiel Brenndauer*

n = 1000 Brenndauer von Glühbirnen von 13000 Stunden bis 16000 Stunden in 15 Klassen eingeteilt.

Zur Berechnung des Mittelwertes ersetzt man die einzelnen Klassen durch den Mittelwert der jeweiligen Klassengrenzen:

$$\bar{x} = \frac{1}{1000} \cdot (13100 \cdot 7 + 13300 \cdot 8 + 13500 \cdot 11 + 13700 \cdot 42 + 13900 \cdot 65 + 14100 \cdot 96 + 14300 \cdot 104 + 14500 \cdot 173 + 14700 \cdot 105 + 14900 \cdot 143 + 15100 \cdot 108 + 15300 \cdot 72 + 15500 \cdot 34 + 15700 \cdot 22 + 15900 \cdot 10)$$

$\bar{x}$ *=14612,8 Stunden*

Es wird daher sinnvoll sein, die mittlere Brenndauer der Glühbirne mit ca. 14600 Stunden anzugeben.

- *ad Beispiel Allergiker*

n = 300 Augenrötung mit Medikament in 4 Klassen eingeteilt

Um den Mittelwert zu berechnen, ersetzt man die qualitativen Werte der einzelnen Klassen durch Zahlen z.B. $x_1 = 1$; $x_2 = 2$; $x_3 = 3$; $x_4 = 4$

$$\bar{x} = \frac{42 \cdot 1 + 102 \cdot 2 + 96 \cdot 3 + 60 \cdot 4}{300}$$

$\bar{x}$ *=2,58; also beim Übergang von der 2. zur 3. Klasse*

Im Schnitt hatten die Patienten mit Medikamenteneinnahme eine leichte bis mittlere Augenrötung.

Der Vollständigkeit halber sei erwähnt, daß es zur Mittelwertbildung nicht nur die Möglichkeit des arithmetischen Mittels gibt, sondern auch noch das geometrische Mittel und das harmonische Mittel. In der Statistik ist jedoch meist das arithmetische Mittel von Bedeutung.

Den **geometrischen Mittelwert** $\hat{x}$ benötigt man zur Durchschnittsberechnung bei exponentiellen Wachstums- oder Abnahmeprozessen.

Man nennt $\hat{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ das **geometrische Mittel** der reellen Zahlen x_i.

***Beispiel**:* *In einem Betrieb wurden in 5 aufeinanderfolgenden Jahren die Produktionszahlen jeweils um 5%; 35%; 105%; 25% und 30% gesteigert. Wie groß ist die durchschnittliche jährliche Steigerung?*

Angenommen man berechnet das arithmetische Mittel aus den Steigerungen:

$$\overline{x} = \frac{5 + 35 + 105 + 25 + 30}{5} = \frac{200}{5} = 40$$

So müßte also eine durchschnittliche Steigerung um 40% durch alle 5 Jahre dasselbe Resultat liefern, wie die tatsächlichen Steigerungen. Bei einer Ausgangsproduktion P_0 ergibt sich dann:

$$P_1 \text{ (nach einem Jahr)} = P_o + \frac{5 \cdot P_0}{100} = P_o \cdot 1{,}05$$

$$P_2 \text{ (nach zwei Jahren)} = P_1 + \frac{35 \cdot P_1}{100} = P_1 \cdot 1{,}35 = P_o \cdot 1{,}05 \cdot 1{,}35 \text{ usw.}$$

und letztendlich für $P_5 = P_o \cdot 1{,}05 \cdot 1{,}35 \cdot 2{,}05 \cdot 1{,}25 \cdot 1{,}3 = P_o \cdot 4{,}722$

Mit 40% jährlicher Steigerung erhält man für $P_5 = P_0 \cdot 1{,}40^5 = P_o \cdot 5{,}378$, also ein wesentlich größeres Endresultat. Der arithmetische Mittelwert ist daher nicht zufriedenstellend. Ist p die gesuchte prozentuelle Steigerung, dann muß gelten:

$$P_o \cdot 1{,}05 \cdot 1{,}35 \cdot 2{,}05 \cdot 1{,}25 \cdot 1{,}3 = P_o \cdot \left(1 + \frac{p}{100}\right)^5$$

$$\left(1 + \frac{p}{100}\right) = \sqrt[5]{1{,}05 \cdot 1{,}35 \cdot 2{,}05 \cdot 1{,}25 \cdot 1{,}3}$$

$$1 + \frac{p}{100} = 1{,}364$$

$$p = 36{,}4$$

Die durchschnittliche prozentuelle Steigerung durch alle 5 Jahre betrug 36,4%.

Die Werte x_1, x_2, ..., x_5 sind also nicht die Prozentzahlen 5, 35, 105, 25 und 30 sondern die Wachstumsfaktoren 1,05; 1,35; 2,05; 1,25 und 1,3.

Beim arithmetischen Mittel ändert sich die Summe der Datenliste nicht, wenn man den Mittelwert n-mal anstatt der einzelnen x_i addiert. Entsprechend ändert sich beim geometrischen Mittel das Produkt nicht, wenn man den Mittelwert n-mal anstatt der einzelnen x_i multipliziert. Der **harmonische Mittelwert** $\tilde{x}$ ist immer dann anzuwenden, wenn die Variablenwerte verkehrt proportional zu jener Größe sind, die sich durch die Durchschnittsbildung nicht verändern darf.

Man nennt $\tilde{x} = \dfrac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$ das **harmonische Mittel** der reellen Zahlen x_i.

***Beispiel**: Ein Rennfahrer fährt hintereinander 6 Runden mit folgenden mittleren Geschwindigkeiten: 190 km/h, 205 km/h, 185 km/h, 208 km/h 201 km/h und 198 km/h. Wie groß ist die mittlere Geschwindigkeit für alle 6 Runden?*

Die aufgewendete Gesamtzeit darf sich durch Verwendung des Mittelwertes anstatt der Einzelgeschwindigkeiten nicht ändern. Es gilt:

$$t = \frac{s}{v} \quad \textit{(t ... Zeit, s ... Weg, v ... Geschwindigkeit)}$$

v ist also verkehrt proportional zu t

$$\frac{s}{190} + \frac{s}{205} + \frac{s}{185} + \frac{s}{208} + \frac{s}{201} + \frac{s}{198} = \frac{6s}{\tilde{v}}$$

$$\tilde{v} = \frac{6}{\frac{1}{190} + \frac{1}{205} + \frac{1}{185} + \frac{1}{208} + \frac{1}{201} + \frac{1}{198}}$$

$$\tilde{v} = 197{,}5 \text{ km/h}$$

Mit einer Durchschnittsgeschwindigkeit von 197,5 km/h über alle 6 Runden hätte der Rennfahrer dieselbe Gesamtzeit erreicht. Oftmals wird gerade bei der Berechnung der durchschnittlichen Geschwindigkeit im Alltagsleben fehlerhaft vorgegangen.

11.4. Streuungsmaße

Die Angabe eines Zentralmaßes alleine besagt meist sehr wenig über die vorliegende Datenliste, wenn nicht bekannt ist, wie stark die einzelnen Werte der Liste vom Zentralmaß abweichen bzw. um das Zentralmaß streuen.

Eine Möglichkeit, die Streuung von Daten auszudrücken, ist die **mittlere (absolute) Abweichung s^*** vom Zentralmaß. Man versteht darunter den arithmetischen Mittelwert aller Absolutbeträge der Differenzen aller Listenwerte vom Zentralmaß (Zm). Die Absolutbeträge sind nötig, da sich sonst im Durchschnitt negative und positive Abweichungen aufheben würden.

Man nennt $s^* = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - Zm|$ die mittlere absolute Abweichung der reellen Zahlen x_i von einem Zentralmaß Zm.

Wählt man als Zentralmaß den Mittelwert $\bar{x}$, dann gilt : $s^* = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ \dots \ + |x_n - \bar{x}|}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$

Bei x_1, x_2, ..., x_k verschiedenen Variablenwerten mit den absoluten Häufigkeiten H_1, H_2, ..., H_k und den relativen Häufigkeiten h_1, h_2, ... h_k ergibt sich dann: $s^* = \frac{1}{n} \cdot \sum_{i=1}^{k} H_i \cdot |x_i - \bar{x}| = \sum_{i=1}^{k} h_i \cdot |x_i - \bar{x}|$

Zur Berechnung der absoluten Abweichungen ist das geeignetste Zentralmaß nicht das arithmetische Mittel, sondern der Median, da für den Median die mittlere Abweichung $s^* = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - Z|$ den kleinsten Wert erreicht.

Für den arithmetischen Mittelwert erreicht die Summe der Quadrate der Abweichungen ihren kleinsten Wert. Diese Behauptung, wie auch der Beweis ist erst nach dem Kapitel Differentialrechnung einsichtig.

Aus diesem Grund ist für den arithmetischen Mittelwert das geeignete Streumaß der Mittelwert der Quadrate der Abweichungen. Man nennt diese mittlere quadratische Abweichung vom arithmetischen Mittel **(empirische) Varianz.**

Man nennt $V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ bzw. $V = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot H_i = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot h_i$ die **empirische Varianz** der reellen Zahlen x_i.

Diese Formel läßt sich noch vereinfachen zu:

$$V = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} H_i \cdot x_i^2 \right) - \bar{x}^2 = \left(\sum_{i=1}^{n} h_i \cdot x_i \right) - \bar{x}^2$$ „Steinerscher Verschiebungssatz“

Beweis:

$$V = \frac{(x_1 - \bar{x})^2 \cdot H_1 + (x_2 - \bar{x})^2 \cdot H_2 + \ \dots \ + (x_k - \bar{x})^2 \cdot H_k}{n} =$$

$$= \frac{\left(x_1^2 - 2x_1\bar{x} + \bar{x}^2\right) \cdot H_1 + \left(x_2^2 - 2x_2\bar{x} + \bar{x}^2\right) \cdot H_2 + \ \dots \ + \left(x_k^2 - 2x_k\bar{x} + \bar{x}^2\right) \cdot H_k}{n} =$$

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

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

Aus der Varianz läßt sich durch Wurzelziehen die sogenannten **empirische Standardabweichung s** einer Liste von n Werten mit dem arithmetischen Mittel $\bar{x}$ errechnen.

Die **empirische Standardabweichung s** beträgt:

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

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

Die empirische Standardabweichung ist das gebräuchlichste Streuungsmaß.

Um Listen miteinander vergleichen zu können, muß immer das Streuungsmaß im Zusammenhang mit dem Zentralmaß angegeben werden. Will man beurteilen, bei welcher Liste die Werte stärker streuen, so genügt es nur dann, die absoluten mittleren Abweichungen bzw. die Standardabweichungen der Liste miteinander zu vergleichen, wenn die Listen ungefähr denselben Median bzw. arithmetischen Mittelwert aufweisen. Denn es ist zum Beispiel bei einer mittleren Länge von 10 m eine Abweichung von 1 mm sehr gering; aber bei einer mittleren Länge von 10 mm wäre die Abweichung von 1 mm extrem groß.

Streuungen können daher nur verglichen werden, wenn sie durch den **Variabilitätskoeffizienten** $v^* = \frac{s^*}{Z}$ bzw. durch den **Variationskoeffizienten** $v = \frac{s}{\bar{x}}$ angegeben werden. In beiden Fällen wird üblicherweise die Abweichung (s^* bzw. s) in Prozent vom Zentralmaß (Z bzw. $\bar{x}$) angegeben.

Man nennt $v^* = \frac{s^*}{Z}$ den **Variabilitätskoeffizienten** und $v = \frac{s}{\bar{x}}$ den **Variationskoeffizienten**.

Nachfolgend sollen für die Beispiele Schuhgröße und Brenndauer die Streuungsmaße berechnet werden.

***Beispiele**:*

- *ad Beispiel Schuhgröße*

x_i	H_i	$H_i \cdot \lvert x_i - Z \rvert$	$H_i \cdot (x_i - \bar{x})^2$
36	*6*	*24*	*114,94*
37	*45*	*135*	*513,12*
38	*82*	*164*	*463,23*
39	*98*	*98*	*185,76*
40	*87*	*0*	*12,35*
41	*84*	*84*	*32,63*
42	*51*	*102*	*134,38*
43	*23*	*69*	*158,27*
44	*48*	*192*	*630,13*
45	*31*	*155*	*662,60*
46	*5*	*30*	*158,10*
Z = 40	*560*	*1053*	*3065,50*
$\bar{x}$ *= 40,38*		s^* *= 1,88*	*s = 2,34*

Die Standardabweichung läßt sich aufgrund der gegebenen Häufigkeiten leichter mit der entsprechenden Formel berechnen: $s = \sqrt{\left(\frac{1}{n}\cdot\sum_{i=1}^{k} x_i^2 \cdot H_i\right) - \bar{x}^2} = \sqrt{\left(\frac{1}{560}\cdot\left(36^2\cdot 6 + 37^2\cdot 45 + ... + 46^2\cdot 5\right)\right) - 40{,}38^2} = 2{,}34$

Nun lassen sich Variabilitätskoeffizient und Variationskoeffizient berechnen.

$$v^* = \frac{1{,}88}{40} = 0{,}047 \equiv 4{,}7\% \quad \textit{Streuung um Z}$$

$$v = \frac{2{,}45}{40{,}37} = 0{,}060 \equiv 6\% \quad \textit{Streuung um } \bar{x}$$

Die durchschnittliche Schuhgröße der Schüler beträgt 40,37 ± 6%

bzw. die Schuhgrößen liegen mit ± 4,7% um die Größe 40.

Welches Maß das aussagekräftigste bzw. das sinnvollste ist, hängt prinzipiell immer vom konkreten Beispiel ab. Im obigen Fall ist die Streuung um den Median Z, der ja selbst eine Schuhgröße darstellt, sicher die interessantere Aussage.

- *ad Beispiel Brenndauer*

Klassenmitte x_i	*H_i*	*$H_i\cdot\lvert x_i - Z\rvert$*	*$x_i^2\cdot H_i$ in 10^4*	*Klassenmitte x_i*	*H_i*	*$H_i\cdot\lvert x_i - Z\rvert$*	*$x_i^2\cdot H_i$ in 10^4*
13100	*7*	*9800*	*120127*	*14700*	*105*	*21000*	*2268945*
13300	*8*	*9600*	*261639*	*14900*	*143*	*57200*	*3174743*
13500	*11*	*11000*	*200475*	*15100*	*108*	*64800*	*2462508*
13700	*42*	*33600*	*788298*	*15300*	*72*	*57600*	*1685448*
13900	*65*	*39000*	*1255865*	*15500*	*34*	*34000*	*816850*
14100	*96*	*38400*	*1908576*	*15700*	*22*	*26400*	*542278*
14300	*104*	*20800*	*2126696*	*15900*	*10*	*14000*	*252810*
14500	*173*	*0*	*3637325*				
Z=14500	*1000*					*437200*	*21502580*
$\bar{x}$ =14600							

s^* = 437,2 und s = 1365,95

v^* = 3% und v = 9,35%

Die mittlere Brenndauer der Glühlampen beträgt also 14600 Stunden ± 9,35%.

11.5. Zusammenhänge zwischen Datenmengen

(a) Regressionsanalyse

Bei vielen statistischen Erhebungen wird die Gesamtheit der n Elemente auf mehrere Merkmale x, y, ... hin untersucht, wobei im Anschluß nicht nur die Auswertung der einzelnen Variablen von Bedeutung ist, sondern es wird von Interesse sein, Zusammenhänge zwischen den einzelnen Variablen zu untersuchen. Zum Beispiel wird oft ein Zusammenhang hergestellt zwischen Körpergröße und Körpermasse, zwischen Luftdruck und Niederschlagsmenge, zwischen Werbungskosten und Umsatzsteigerung, u.v.a.

***Beispiel**:* *ad Beispiel Schuhgröße*

Die Schüler wurden auch nach ihrer Körpergröße befragt. Im folgenden werden 20 solcher zusammenhängender Daten herausgegriffen:

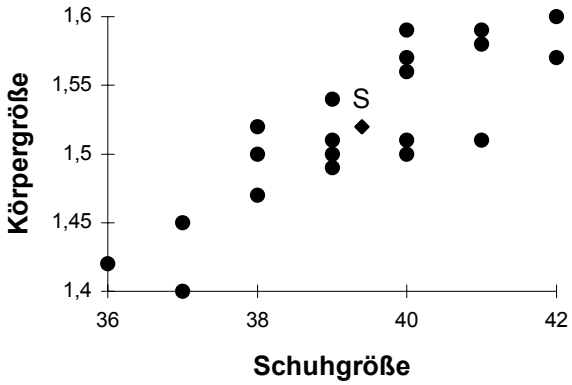
x_i (Schuhnummer)	y_i (Körpergröße in m)	x_i (Schuhnummer)	y_i (Körpergröße in m)
36	1,42	40	1,59
37	1,40	40	1,56
37	1,45	40	1,57
38	1,47	40	1,50
38	1,50	40	1,51
38	1,52	41	1,59
39	1,50	41	1,51
39	1,54	41	1,58
39	1,49	42	1,57
39	1,51	42	1,60
Z = 39,5	Z = 1,51	$\overline{x}$ = 39,35 ≙ 39,4	$\overline{y}$ = 1,519 ≙ 1,52

Aus den Daten läßt sich ein tendenzieller Zusammenhang erkennen; nämlich mit zunehmender Größe der Schuhnummer wird auch die Körpergröße größer. Diese Tendenz muß jedoch nicht in jedem Einzelfall stimmen (z.B. 41/1,51). Darüberhinaus kann auch keine Aussage über die Stärke dieses Zusammenhangs gegeben werden.

Besser als in einer Tabelle läßt sich ein eventueller Zusammenhang in einem **Streudiagramm** erkennen. In einem solchen Diagramm werden alle Wertepaare $(x_i;y_i)$ als Punkte $P_i(x_i \mid y_i)$ dargestellt. Auf diese Weise entsteht eine Punktwolke, deren Schwerpunkt durch $S(\overline{x} \mid \overline{y})$ gegeben ist.

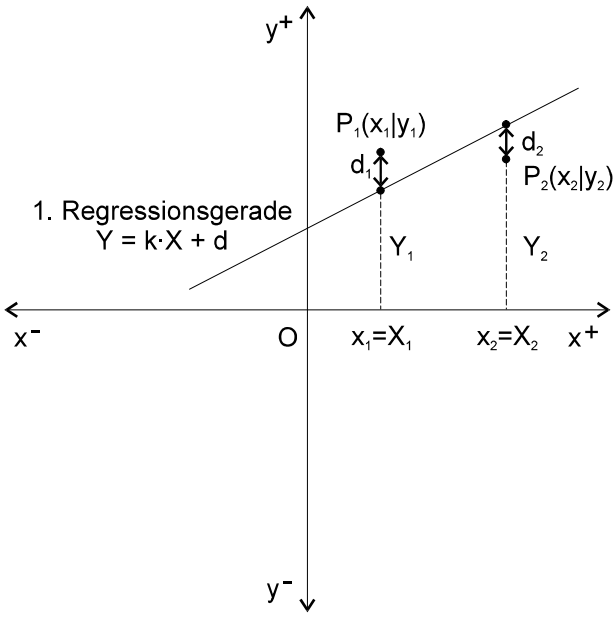
***Beispiel**:*

Streudiagramm zum Beispiel
Schuhgröße / Körpergröße



Die **Regressionsanalyse** versucht den Zusammenhang zwischen zwei Variablen durch eine Funktion zu beschreiben. Wie gut dieser Zusammenhang tatsächlich gegeben ist, wird durch die **Korrelationsanalyse** ausgedrückt.

Im einfachsten Fall kann der Zusammenhang durch eine lineare Funktion beschrieben werden. In diesem Fall spricht man von **linearer Regression.** Man versucht bei diesem Verfahren, die Punktewolke durch eine Gerade vereinfacht darzustellen. Der Graph der linearen Regressionsfunktion ist also die sogenannte Regressionsgerade.



Die Regressionsgerade muß den Schwerpunkt der Punktwolke $S(\overline{x}/\overline{y})$ enthalten und soll „möglichst nahe“ bei den einzelnen Punkten liegen.

Will man durch lineare Regression aus den x-Werten die y-Werte näherungsweise berechnen, so müssen die Abstände $d_i = Y_i - y_i$ der tatsächlichen Punkte von der 1. Regressionsgeraden in y-Richtung möglichst gering sein.

Man nimmt als Maß für diese Abweichung nicht d_1, d_2, etc., da diese für jede Gerade durch S einander aufheben würden (Vorzeichen); auch die Beträge der Abweichungen führen zu Schwierig-

keiten beim Festlegen der Regressionsgeraden. Die übliche Methode ist die **Fehlerquadratmethode von C.F. GAUSS (1777-1855)**. Die Methode der kleinsten Quadrate verlangt, daß die Summe aller Abweichungen $d_1^2+d_2^2+ ... +d_n^2$ ein Minimum annimmt, wenn die Regressionsgerade richtig festgelegt wird.

Es sei $Y = k \cdot X + d$ die Regressionsfunktion, $P_i(x_i \mid y_i)$ sind die tatsächlichen Punkte; auf der 1. Regressionsgeraden liegen die Punkte $R_i(X_i = x_i \mid Y_i = k \cdot x_i + d)$. Die Summe der Abstandsquadrate ergibt sich somit:

$$\sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} (kx_i + d - y_i)^2 = F(k,d)$$

Die Werte für k und d sollen nun so bestimmt werden, daß F(k,d) einen Minimalwert annimmt. Die genaue Berechnung erfolgt mit den Mitteln der Differentialrechnung, an dieser Stelle sei nur das Ergebnis angeführt.

Für k ergibt sich:
$$k = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Für d ergibt sich:
$$d = \frac{\sum_{i=1}^{n} y_i}{n} - k \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

Die **1. Regressionsgerade** lautet

$$Y = k \cdot X + d \text{ mit } k = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \text{ und } d = \frac{\sum_{i=1}^{n} y_i}{n} - k \cdot \frac{\sum_{i=1}^{n} x_i}{n}.$$

Unter Anwendung des Zusammenhangs $\frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \overline{y}$ und $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \overline{x}$ lassen sich obige Ausdrücke noch umformen:
$$d = \overline{y} - k \cdot \overline{x}.$$

Dies ist gleichzeitig ein Beweis dafür, daß der Schwerpunkt $S(\overline{x} \mid \overline{y})$ auf der Regressionsgeraden liegt. Auch das Ergebnis für k läßt sich unter Verwendung des Steinerschen Verschiebungssatzes weiter vereinfachen.

$$k = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot n \cdot \bar{y}}{n \cdot \sum_{i=1}^{n} x_i^2 - (n \cdot \bar{x})^2} = \frac{n^2 \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \overline{xy} \right)}{n^2 \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right)} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

In diesem Zusammenhang bezeichnet man $s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$ als Kovarianz von x und y und $s_x = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ wie bisher als Varianz von x. Damit läßt sich k angeben als $k = \frac{\text{Kovarianz von x und y}}{\text{Varianz von x}}$.

Da die Abstände d_i in y-Richtung minimiert wurden, stellt die 1. Regressionsgerade nun eine Möglichkeit zur Abschätzung von y-Werten aus gegebenen x-Werten dar. Einige Taschenrechner mit statistischen Funktionen sind in der Lage, nach Eingabe der Datenlisten x_i und y_i die Zentralmaße, Streuungsmaße und Koeffizienten der Regressionsgeraden zu berechnen. Für die händische Berechnung geht man wie bisher mit Tabellen vor.

***Beispiel**:* *ad Beispiel Schuhgröße / Körpergröße*

Welche Körpergröße hat ein Schüler voraussichtlich mit Schuhgröße 39 bzw. 40?

In diesem Beispiel soll nun eine Körpergröße (y) abgeschätzt werden, wenn die Schuhnummer des Schülers (x) bekannt ist. Dazu werden zunächst die Koeffizienten der 1. Regressionsgeraden errechnet.

$$k = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{20 \cdot 1196{,}97 - 787 \cdot 30{,}38}{20 \cdot 31021 - 787^2} = 0{,}02886$$

$$d = \frac{30{,}38}{20} - k \cdot \frac{787}{20} = 0{,}383$$

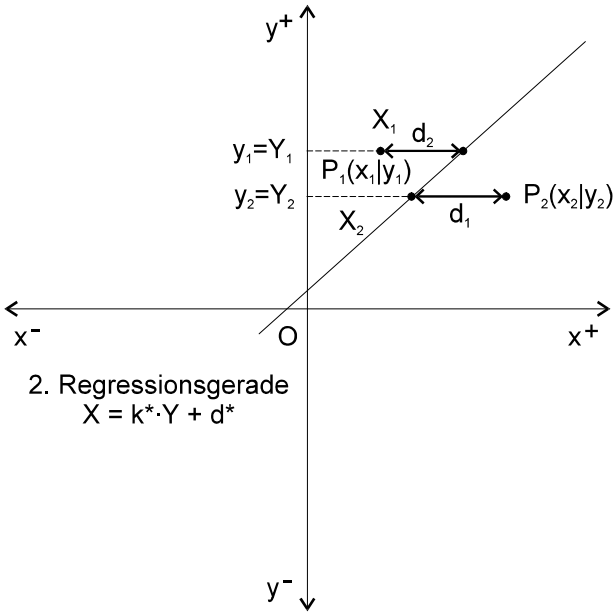
$$Y = 0{,}02886 \cdot X + 0{,}383$$

Nun wird mit den Werten X = 39 und X = 40 in der Funktion eingesetzt, um die voraussichtliche Körpergröße Y zu erhalten.

Für Schuhgröße 39 schätzt man eine Körpergröße 1,508 ≡ 1,51m.

Für Schuhgröße 40 schätzt man eine Körpergröße 1,537 ≡ 1,54m.

Besteht ein Zusammenhang zwischen den Variablen x und y, so kann natürlich nicht nur von x auf y geschlossen werden, sondern auch aus bekannten y-Werten auf x. Dazu sollte dann allerdings eine Regressionsgerade verwendet werden, für die die Summe der Quadrate der Abstände der tatsächlichen Punkte von der Regressionsgeraden in x-Richtung möglichst klein wird. Dies ist dann die 2. Regressionsgerade.



Die gesuchte Regressionsgerade für diesen Fall lautet $X = k^* \cdot Y + d^*$. In diesem Fall müssen die Abstände $d_i = X_i - x_i$ bzw. genauer die Summe deren Quadrate minimiert werden.

Daher muß also

$$F(k^*, d^*) = \sum_{i=1}^{n} (X_i - x_i)^2 = \sum_{i=1}^{n} \left(k^* \cdot y_i + d^* - x_i\right)^2$$

einen Minimalwert annehmen.

Völlig analog zur 1. Regressionsgeraden ergibt sich:

$$k^* = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - \overline{y}^2} = \frac{s_{xy}}{s_y^2} \text{ und } d^* = \frac{\sum_{i=1}^{n} x_i - k^* \cdot \sum_{i=1}^{n} y_i}{n} = \overline{x} - k^* \cdot \overline{y}$$

Die **2. Regressionsgerade** lautet

$$X = k^* \cdot Y + d^* \text{ mit } k^* = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \text{ und } d^* = \frac{\sum_{i=1}^{n} x_i - k^* \cdot \sum_{i=1}^{n} y_i}{n}.$$

Wie die obigen Formeln zeigen, liegt $S(\overline{x}|\overline{y})$ auch auf der 2.Regressionsgeraden. Die zweite Regressionsgerade erlaubt es nun, für ein Y ein voraussichtliches X zu schätzen.

***Beispiel**:* *ad Beispiel Schuhgröße / Körpergröße*

Welche Schuhgröße hat ein Schüler voraussichtlich mit Körpergröße 1,50 m bzw. 1,54 m?

Für dieses Beispiel ergibt sich folgende 2. Regressionsgerade:

$$k^* = \frac{20 \cdot 1196{,}97 - 787 \cdot 30{,}38}{20 \cdot 46{,}2082 - (30{,}38)^2} = 24{,}877$$

$$d^* = \frac{787}{20} - k^* \cdot \frac{30{,}38}{20} = 1{,}5618$$

$$X = 24{,}877 \cdot Y + 1{,}5618$$

Nun wird mit den Werten Y = 1,50 und Y = 1,54 in der Funktion eingesetzt, um die voraussichtliche Schuhgröße X zu erhalten.

Für Körpergröße 1,50 m kann damit die Schuhgröße 38,87 ≡ 39 abgeschätzt werden.

Für Körpergröße 1,54 m kann damit die Schuhgröße 39,87 ≡ 40 abgeschätzt werden.

Um beide Regressionsgeraden im selben Koordinatensystem einzuzeichnen und miteinander vergleichen zu können, ist es günstiger, bei beiden Geraden Y explizit auszudrücken:

1. Regressionsgerade g_1: $$Y = k \cdot X + d = \frac{s_{xy}}{s_x^2} \cdot X + \overline{y} - \frac{s_{xy}}{s_x^2} \cdot \overline{x}$$

2. Regressionsgerade g_2: $$Y = \frac{1}{k^*} \cdot X - \frac{d^*}{k^*} = \frac{s_y{}^2}{s_{xy}} \cdot X - \frac{\overline{x} - k^* \overline{y}}{k^*} = \frac{s_y{}^2}{s_{xy}} \cdot X + \overline{y} - \frac{s_y{}^2}{s_{xy}} \cdot \overline{x}$$

Im Falle eines perfekten linearen Zusammenhanges zwischen den Variablen x und y müssen die beiden Regressionsgeraden zusammenfallen.

Der nächste Abschnitt beschäftigt sich genauer mit der Untersuchung des Zusammenhangs zwischen den beiden Regressionsgeraden und daher mit dem Zusammenhang zwischen den beiden Datenlisten. Diese Untersuchung bezeichnet man als **Korrelationsanalyse**.

(b) Korrelationsanalyse

Naheliegenderweise fallen die Regressionsgeraden zusammen, wenn zwischen den Variablen x und y ein perfekter linearer Zusammenhang besteht. Das bedeutet, daß man in die Regressionsgeraden mit einem Wert einsetzen kann und dann den tatsächlichen Wert, und nicht nur eine Schätzung, als Ergebnis bekommt. Darüberhinaus bedeutet es auch, daß alle Wertepaare der Datenliste auf einer Geraden liegen.

Wenn die Geraden zusammenfallen, heißt das, daß sie auch den gleichen Anstieg haben. Es gilt also:

$$k = \frac{1}{k^*} \text{ und somit } k \cdot k^* = 1$$

Im anderen Extremfall, wenn also überhaupt kein Zusammenhang zwischen den Variablen x und y besteht, ergeben sich aufeinander normal stehende Regressionsgeraden. In diesem Fall ist sowohl k=0 also auch k^*=0 und es gilt: $k \cdot k^* = 0$

In jedem anderen Fall schließen die beiden Regressionsgeraden einen spitzen (und einen stumpfen) Winkel α ein, der umso größer sein wird, je weniger der tatsächliche Zusammenhang linear ist ($0° < \alpha < 90°$). Die Untersuchung, wie gut nun die lineare Regression dem tatsächlichen Zusammenhang angepaßt ist, nennt man **Korrelationsanalyse.**

Das rechte Maß dafür ist also der Winkel α zwischen g_1 und g_2. Aus $k = \tan(\alpha_1)$ und $\frac{1}{k^*} = \tan(\alpha_2)$ läßt sich α_1 und α_2 errechnen. Dann ist $\arctan(k) = \alpha_1$ und $\arctan\left(\frac{1}{k^*}\right) = \alpha_2$, wobei α_1 der Winkel zwischen g_1 und der x-Achse und α_2 der Winkel zwischen g_2 und der x-Achse ist. Es ergibt sich dann der Winkel zwischen den beiden Geraden als $\alpha = \alpha_2 - \alpha_1$.

Eine zweite Berechnungsmethode unter Verwendung der Vektorrechnung liefert folgendes:

$$\cos(\alpha) = \frac{\begin{pmatrix}1\\k\end{pmatrix} \cdot \begin{pmatrix}1\\\frac{1}{k^*}\end{pmatrix}}{\left|\begin{pmatrix}1\\k\end{pmatrix}\right| \cdot \left|\begin{pmatrix}1\\\frac{1}{k^*}\end{pmatrix}\right|} = \frac{1+\frac{k}{k^*}}{\sqrt{1+k^2} \cdot \sqrt{1+\left(\frac{1}{k^*}\right)^2}}$$

Üblicherweise verzichtet man jedoch auf die Berechnung der Winkel und verwendet als **Bestimmtheitsmaß der Korrelation** das Verhältnis der Steigungen der beiden Regressionsgeraden.

Bestimmtheitsmaß der Korrelation: $r_{xy}^2 = k : \frac{1}{k^*} = k \cdot k^* = \frac{s_{xy}}{s_x^2} \cdot \frac{s_{xy}}{s_y^2}$

Da sich aus diesem Wert problemlos die Wurzel ziehen läßt, verwendet man in der Praxis meist den Pearsonschen Korrelationskoeffizienten r_{xy} , den sogenannten linearen Korrelationskoeffizient.

Pearsonscher Korrelationskoeffizient r_{xy} $r_{xy} = \sqrt{k \cdot k^*} = \frac{s_{xy}}{s_x \cdot s_y}$

Setzt man mit der Kovarianz $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ und den Standardabweichungen s_x und s_y mit $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ und $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ in die Formel für den Korrelationskoeffizienten ein, so erhält man:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{bzw.} \quad r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Für den linearen Korrelationskoeffizienten gilt dabei stets: $-1 \le r_{xy} \le +1$

Hierbei haben die Werte von r_{xy} folgende Bedeutung:

$r_{xy} = 1$	perfekter direkter Zusammenhang (d.h. z.B. bei Verdopplung von x auch Verdopplung von y)
$0 < r_{xy} < 1$	direkter Zusammenhang (d.h. mit zunehmenden Werten von x auch Zunahme der Werte von y)
$r_{xy} = 0$	kein Zusammenhang zwischen x und y oder zumindest kein linearer
$-1 < r_{xy} < 0$	indirekter (umgekehrter) Zusammenhang (d.h. mit zunehmenden Werten von x Abnahme der Werte von y)
$rxy = -1$	perfekter indirekter (umgekehrter) Zusammenhang (d.h. z.B bei Verdopplung von x folgt Halbierung von y)

Wenn man r_{xy} über die Anstiege berechnet, muß das Vorzeichen dieser Anstiege erst nach dem Wurzelziehen berücksichtigt werden, d.h., daß man unter der Wurzel vorerst den Absolutbetrag nimmt und nach dem Wurzelziehen die Vorzeichen der Anstiege wieder hinzufügt.

***Beispiel**:* *ad Beispiel Schuhgröße / Körpergröße*

Bestimmen Sie den Zusammenhang zwischen Schuh- und Körpergröße.

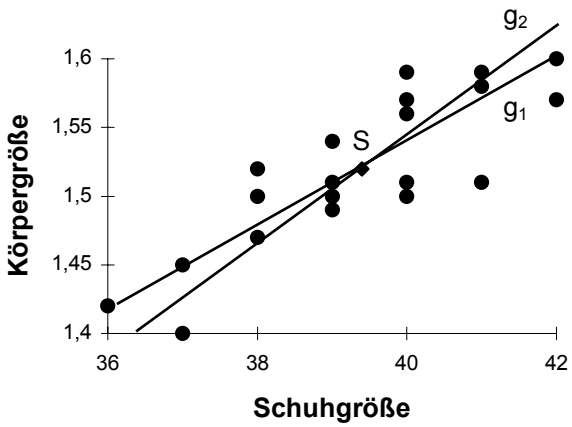
Die Regressionsanalyse hat die beiden Regressionsgeraden g_1 und g_2 ergeben:

g_1: $Y = 0{,}02886 \cdot X + 0{,}383$

g_2: $X = 24{,}877 \cdot Y + 1{,}5618$

$$r_{xy} = \sqrt{0{,}02886 \cdot 24{,}877} = 0{,}847$$

Da r_{xy} in der Nähe von 1 liegt besteht ein ziemlich guter linearer Zusammenhang zwischen x (Schuhgröße) und y (Körpergröße). Das Streudiagramm mit den beiden Regressionsgeraden spiegelt auch diesen Zusammenhang wider.



Um die Korrelation direkt aus den Daten zu ermitteln, empfiehlt sich folgende Tabelle:

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
36	1,42	1296	2,0164	51,12	40	1,59	1600	2,5281	63,6
37	1,40	1369	1,96	51,8	40	1,56	1600	2,4336	62,4
37	1,45	1369	2,1025	53,65	40	1,57	1600	2,4649	62,8
38	1,47	1444	2,1609	55,86	40	1,50	1600	2,25	60,0
38	1,50	1444	2,25	57,0	40	1,51	1600	2,2801	60,4
38	1,52	1444	2,3104	57,76	41	1,59	1681	2,5281	65,19
39	1,50	1521	2,25	58,5	41	1,51	1681	2,2801	61,91
39	1,54	1521	2,3716	60,06	41	1,58	1681	2,4964	64,78
39	1,49	1521	2,2201	58,11	42	1,57	1764	2,4649	65,94
39	1,51	1521	2,2801	58,89	42	1,60	1764	2,56	67,2

$$\sum x_i = 787 \qquad \sum y_i = 30{,}38 \qquad \sum x_i^2 = 31021 \qquad \sum y_i^2 = 46{,}2082 \qquad \sum x_i y_i = 1196{,}67$$

$$r_{xy} = \frac{20 \cdot 1196{,}97 - 787 \cdot 30{,}38}{\sqrt{\left(20 \cdot 31021 - 787^2\right) \cdot \left(20 \cdot 46{,}2082 - 30{,}38^2\right)}} = 0{,}847$$

Das Berechnen von Regressionsgeraden und des Korrelationskoeffizienten ist nur für **metrisch skalierte Daten** möglich (d.h. quantitative Daten, aus denen nicht nur Verschiedenartigkeit und Rangordnung, sondern auch Abstände zwischen den Merkmalsausprägungen angegeben werden können) und nur sinnvoll, wenn genügend Daten vorhanden sind (zumindest $n > 5$).

Regression und Korrelation beschreiben einen linearen Zusammenhang zwischen zwei Variablen auf rein mathematischer Grundlage ohne auf die Ursachen oder Sinnhaftigkeit des Zusammenhangs einzugehen. Für $|r_{xy}| > 0{,}6$ besteht rechnerisch ein starker Zusammenhang, der aber völlig sinnlos sein kann, z.B. kann die Anzahl der Autos in einer Stadt in den letzten 3 Jahren stark gestiegen und gleichzeitig die Geburtenrate stark zurückgegangen sein. Rechnerisch können diese beiden Merkmale korrelieren, trotzdem wird man kaum einen ursächlichen Zusammenhang zwischen diesen beiden Variablen herstellen können. Es liegt dann eine **Scheinkorrelation** vor.

Häufig sollen in der Statistik auch Zusammenhänge zwischen **nominal skalierten** Daten, die nur Verschiedenartigkeit ausdrücken wie z.B. Geschlecht, Haarfarbe, usw., oder **ordinal skalierten** Daten, die Verschiedenartigkeit und Rangordnung ausdrücken wie z.B. Klasseneinteilung bei Beispiel Allergiker, Schulnoten, usw., untersucht werden.

Handelt es sich sowohl bei der Variablen x als auch bei y um ordinal skalierte Daten, so ordnet man zunächst die Variablenwerte gemäß ihrer natürlichen Rangfolge und ordnet ihnen dann die Rangzahlen $r = 1, 2, 3, \ldots$ zu. Bei gleichen Variablenwerten wird das arithmetische Mittel der entsprechenden Rangzahlen zugeordnet.

Sind dann d_i die Differenzen aus den Rangzahlen der Werte x_i und y_i, so erhält man durch Vereinfachung des Pearsonschen Korrelationskoeffizienten den sogenannten Spearmanschen Rangkorrelationskoeffizienten r_s.

Spearmanscher Rangkorrelationskoeffizient r_s:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}$$

Auch r_s nimmt die Werte zwischen -1 und $+1$ an mit derselben Bedeutung wie r_{xy}. Da die Differenzen d_i im Zuge der Berechnung quadriert werden, ist es egal, ob die Differenz der Rangzahlen $r(x_i)-r(y_i)$ oder $r(y_i)-r(x_i)$ gebildet wird.

***Beispiel**: Die Korrelation zwischen Aufmerksamkeit von Schülern im Unterricht und Note auf die nächste Schularbeit soll untersucht werden.*

Bewertung Aufmerksamkeit: Sehr gut ≡ 1, Mittelmäßig ≡ 2, Schlecht ≡ 3

Schüler Nr.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
x_i *(Aufmerksamkeit)*	*1*	*2*	*3*	*3*	*2*	*1*	*1*	*1*	*3*	*3*	*3*
y_i *(Schularbeitsnote)*	*1*	*3*	*3*	*4*	*1*	*3*	*2*	*1*	*4*	*5*	*4*

Um die Rangzahlen für x_i und y_i zu ermitteln, müssen die Listen ihrer Größe nach geordnet werden und erhalten „Grundrangzahlen" entsprechend ihrer Reihenfolge. Aus den „Grundrangzahlen" wird nun für jedes x_i und y_i die definitive Rangzahl ermittelt.

Diese ist die Grundrangzahl, sofern der Variablenwert nur einmal auftritt. Tritt er jedoch öfters auf, so ergibt sich die Rangzahl als arithmetisches Mittel der Grundrangzahlen mit diesem Variablenwert. So ist der Rang für die Aufmerksamkeitsnote 1 gleich 2,5, da (1+2+3+4):4=2,5 gilt.

x_i *(Aufmerksamkeit)*	*1*	*1*	*1*	*1*	*2*	*2*	*3*	*3*	*3*	*3*	*3*
Grundrangzahl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Rangzahl	*2,5*	*2,5*	*2,5*	*2,5*	*5,5*	*5,5*	*9*	*9*	*9*	*9*	*9*

y_i *(Schularbeitsnote)*	*1*	*1*	*1*	*2*	*3*	*3*	*3*	*4*	*4*	*4*	*5*
Grundrangzahl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Rangzahl	*2*	*2*	*2*	*4*	*6*	*6*	*6*	*9*	*9*	*9*	*11*

Somit ergibt sich für die einzelnen Schüler mit anschließender Differenzbildung $d_i = r(y_i) - r(x_i)$:

Schüler Nr.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Rang für x_i	*2,5*	*5,5*	*9*	*9*	*5,5*	*2,5*	*2,5*	*2,5*	*9*	*9*	*9*
Rang für y_i	*2*	*6*	*6*	*9*	*2*	*6*	*4*	*2*	*9*	*11*	*9*
d_i	*–0,5*	*0,5*	*–3*	*0*	*–3,5*	*3,5*	*1,5*	*–0,5*	*0*	*2*	*0*

$$r_s = 1 - \frac{6 \cdot 40{,}5}{11 \cdot (121 - 1)} = 0{,}8159$$

Es besteht ein deutlicher Zusammenhang zwischen Aufmerksamkeit und Noten.

Am kompliziertesten ist die Untersuchung des Zusammenhangs von nominal skalierten Daten. Dazu werden die Daten in eine Tafel eingetragen.

	x_1	x_2	x_3	...	x_k	insgesamt
y_1	$H_{1,1}$	$H_{1,2}$	$H_{1,3}$	...	$H_{1,k}$	$\sum_{j=1}^{k} H_{1,j}$
y_2	$H_{2,1}$	$H_{2,2}$	$H_{2,3}$	...	$H_{2,k}$	$\sum_{j=1}^{k} H_{2,j}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_m	$H_{m,1}$	$H_{m,2}$	$H_{m,3}$	...	$H_{m,k}$	$\sum_{j=1}^{k} H_{m,j}$
insgesamt	$\sum_{i=1}^{m} H_{i,1}$	$\sum_{i=1}^{m} H_{i,2}$	$\sum_{i=1}^{m} H_{i,3}$	...	$\sum_{i=1}^{m} H_{i,k}$	$\sum_{i=1}^{m}\left(\sum_{j=1}^{k} H_{i,j}\right)$

Die Werte $H_{i,j}$ sind die beobachteten Anteile für die Variablenwerte y_i und x_j . Bei Unabhängigkeit der beiden Merkmale kann man die Anteile, die zu erwarten wären, berechnen, denn die Häufigkeitsverteilung hinsichtlich der Merkmalsausprägung x_j müßte für alle y_i gleich sein. Diese „erwarteten Anteile“ $E_{r,s}$ für die r-te Zeile und s-te Spalte berechnen sich folgendermaßen:

$$E_{r,s} = \frac{\sum_{i=1}^{m} H_{i,s}}{n} \cdot \sum_{j=1}^{k} H_{r,j} = \frac{(s-\text{te Spaltensumme})}{n} \cdot (r-\text{te Zeilensumme})$$

Das Maß für die Stärke des Zusammenhangs zwischen den Merkmalen x und y ist der sogenannte Kontingenzkoeffizient C.

Kontingenzkoeffizient C: $\quad C = \sqrt{\frac{\chi^2}{n+\chi^2}} \quad$ mit $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(H_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$ (χ ... Chi)

C kann Werte zwischen 0 und 1 annehmen. Der maximal mögliche Wert von C hängt jeweils von der Zeilen- und Spaltenzahl ab. Für quadratische Tafeln m=k gilt $C_{max} = \sqrt{\frac{m-1}{m}} = \sqrt{\frac{k-1}{k}}$, für rechteckige Tafeln $m \neq k$ gilt $C_{max} \approx \frac{1}{2} \cdot \left(\sqrt{\frac{m-1}{m}} + \sqrt{\frac{k-1}{k}} \right)$. Das Verfahren ist in der Regel nur anwendbar, wenn kein erwarteter Anteil kleiner 1 ist und höchstens ein Fünftel der erwarteten Anteile kleiner 5 ist.

***Beispiel**:* *Der Zusammenhang zwischen Rauchgewohnheit und Geschlecht soll an 200 Personen untersucht werden.*

	Raucher	*Nichtraucher*	*insgesamt*
männlich	*64*	*42*	*106*
weiblich	*21*	*73*	*94*
insgesamt	*85*	*115*	*200*

Aus obiger Tabelle lassen sich die angesprochenen erwarteten Anteile leicht errechnen:

$$E_{1,1} = \frac{85}{200} \cdot 106 = 45{,}05 \qquad E_{1,2} = \frac{115}{200} \cdot 106 = 60{,}95$$

$$E_{2,1} = \frac{85}{200} \cdot 94 = 39{,}95 \qquad E_{2,2} = \frac{85}{200} \cdot 94 = 54{,}05$$

Nun läßt sich χ berechnen:

$$\chi^2 = \frac{(64-45{,}05)^2}{45{,}05} + \frac{(42-60{,}95)^2}{60{,}95} + \frac{(21-39{,}95)^2}{39{,}95} + \frac{(73-54{,}05)^2}{54{,}05} = 29{,}4956 \cong 29{,}5$$

Und mit χ letztendlich der Kontingenzkoeffizient C:

$$C = \sqrt{\frac{29{,}5}{200+29{,}5}} = 0{,}358 \text{ und } C_{max} = \sqrt{\frac{2-1}{2}} = 0{,}707$$

Der errechnete Kontingenzkoeffizient C ist also etwa halb so groß wie C_{max}. Daher besteht ein mittelstarker Zusammenhang zwischen „Rauchgewohnheit“ und „Geschlecht“.

In diesem Beispiel traten in der Tafel nur vier Felder auf. Für einen solchen Fall kann als Maß für die Stärke des Zusammenhangs auch der wesentlich einfacher zu berechnende **Vierfelderkoeffizient** ϕ verwendet werden. Auch ϕ kann Werte zwischen 0 und 1 annehmen, praktisch kann +1 aber fast nie erreicht werden.

Vierfelderkoeffizient Φ:

$$\Phi = \frac{\left|H_{1,1} \cdot H_{2,2} - H_{1,2} \cdot H_{2,1}\right|}{\sqrt{\left(H_{1,1}+H_{1,2}\right) \cdot \left(H_{2,1}+H_{2,2}\right) \cdot \left(H_{1,1}+H_{2,1}\right) \cdot \left(H_{1,2}+H_{2,2}\right)}}$$

Für das vorangegangene Beispiel ergibt sich:

$$\Phi = \frac{|64 \cdot 73 - 42 \cdot 21|}{\sqrt{106 \cdot 94 \cdot 85 \cdot 115}} = 0{,}384$$

Anhang: Übungsbeispiele zum 11. Kapitel

11/1 Bestimmen Sie die Häufigkeiten, relativen Häufikeiten und prozentuellen Häufigkeiten der folgenden Datenliste:
68, 84, 75, 82, 68, 90, 62, 88, 76, 93, 73, 79, 88, 73, 60, 93, 71, 59, 85, 75, 61, 65, 75, 87, 74, 62, 95, 78, 63, 72, 66, 78, 82, 75, 87, 74, 62, 95, 78, 63, 72, 66, 78, 82, 75, 94, 77, 69, 74, 68, 60, 96, 78, 89, 61, 75, 95, 60, 79, 73

11/2 Bestimmen Sie die Häufigkeiten, relativen Häufikeiten und prozentuellen Häufigkeiten der folgenden Datenliste:
179, 162, 167, 197, 178, 185, 176, 165, 171, 175, 165, 180, 173, 157, 188, 178, 162, 176, 153, 174, 186, 167, 173, 181, 172, 163, 176, 175, 185, 177

11/3 Die Daten aus Beispiel 11/1 sind das Ergebnis eines Tests, bei dem 100 Punkte zu erreichen waren. Legen Sie eine geeignete Klasseneinteilung fest und bestimmen Sie die Häufigkeiten, relativen Häufikeiten und prozentuellen Häufigkeiten der einzelnen Klassen.

11/4 Die Daten aus Beispiel 11/2 sind das Ergebnis eines Bewerbungstests, bei dem 200 Punkte zu erreichen waren. Legen Sie eine geeignete Klasseneinteilung fest und bestimmen Sie die Häufigkeiten, relativen Häufikeiten und prozentuellen Häufigkeiten der einzelnen Klassen.

11/5 Stellen Sie die Daten der Klasseneinteilung aus Beispiel 11/3 in einem geeigneten Diagramm dar.

11/6 Stellen Sie die Daten der Klasseneinteilung aus Beispiel 11/3 in einem geeigneten Diagramm dar.

11/7 Bestimmen Sie für das Beispiel 11/1 folgende Zentralmaße:
Minimum, Maximum, Spannweite, Modus, Median und Quartilen.

11/8 Bestimmen Sie für das Beispiel 11/2 folgende Zentralmaße:
Minimum, Maximum, Spannweite, Modus, Median und Quartilen.

11/9 Bestimmen Sie für das Beispiel 11/1 das arithmetische Mittel.

11/10 Bestimmen Sie für das Beispiel 11/2 das arithmetische Mittel.

11/11 Bestimmen Sie für das Beispiel 11/1 unter Verwendung der Häufigkeiten das arithmetische Mittel.

11/12 Bestimmen Sie für das Beispiel 11/2 unter Verwendung der Häufigkeiten das arithmetische Mittel.

11/13 Der Umsatz eines Betriebes ist in 4 aufeinanderfolgenden Jahren jeweils um 45%, 110%, 30% und 40% gestiegen. Wie groß ist die durchschnittliche jährliche Steigerung?

11/14 Eine Bakterienkultur wächst während der 16 Tagstunden um 20% pro Stunde, in der Nacht um nur 12% pro Stunde. Berechnen Sie das durchschnittliche Wachstum pro Stunde.

11/15 Ein Kapital von ÖS 1000,- wird drei Jahre lang verzinst, und zwar im ersten Jahr mit p = 8%, im zeiten mit p = 7% und im dritten mit i = 6,5%. Berechnen Sie den durchschnittlichen Jahreszinssatz.

11/16 Ein Sportflugzeug fliegt von Graz nach Wien (ca. 150 km) mit einer Geschwindigkeit von 300 km/h und zurück mit 450 km/h. Berechnen Sie die mittlere Geschwindigkeit.

11/17 Bei einem Autorennen sind 5 Runden zu fahren. Die mittleren Geschwindigkeiten für die einzelnen Runden betragen für einen Fahrer 183, 210, 201, 180, 182 km/h. Wie groß ist die mittlere Geschwindigkeit für alle 5 Runden?

11/18 Wie schnell muß man eine zweite Runde auf einer Rennstrecke fahren, wenn die erste mit 120 km/h gefahren wurde und für beide durchschnittlich 200 km/h erzielt werden sollen?

11/19 Berechnen Sie für das Beispiel 11/1 die absolute Abweichung vom Median.

11/20 Berechnen Sie für das Beispiel 11/2 die absolute Abweichung vom Median.

11/21 Berechnen Sie für das Beispiel 11/1 die Varianz und die Standardabweichung vom Mittelwert.

11/22 Berechnen Sie für das Beispiel 11/2 die Varianz und die Standardabweichung vom Mittelwert.

11/23 Berechnen Sie für das Beispiel 11/1 den Variabilitäts- und den Variationskoeffizienten.

11/24 Berechnen Sie für das Beispiel 11/2 den Variabilitäts- und den Variationskoeffizienten.

11/25 In der folgenden Tabelle sind Körpergröße (cm) und Körpermasse (kg) von 12 Personen gegeben. Ermitteln Sie sowohl für Größe als auch Masse den Mittelwert, den Median und die Standardabweichung und vergleichen Sie die Streuungen der beiden Datenlisten mittels der Variationskoeffizienten.

Nr. i	Größe x_i	Masse y_i	Nr. i	Größe x_i	Masse y_i	Nr. i	Größe x_i	Masse y_i
1	164	48	5	165	53	9	171	63
2	169	68	6	165	66	10	167	50
3	160	51	7	170	56	11	154	46
4	171	54	8	164	48	12	156	50

11/26 Folgende Tabelle zeigt den Zusammenhang zwischen Werbungskosten für ein Produkt und dem zugehörigen Jahresumsatz. Ermitteln Sie sowohl für Kosten als auch Umsatz den Mittelwert, den Median und die Standardabweichung und vergleichen Sie die Streuungen der beiden Datenlisten mittels der Variationskoeffizienten.

Kosten x_i (in 10000)	30	35	35	40	50	50	55	65
Umsatz y_i (in Mio.)	50	60	70	90	120	130	140	140

11/27 Der Zusammenhang zwischen den Punkten bei der Aufnahmsprüfung und der Note nach dem 1. Jahr am Wifi ist gegeben durch folgende Tabelle. Ermitteln Sie sowohl für Punkte als auch Noten den Mittelwert, den Median und die Standardabweichung und vergleichen Sie die Streuungen der beiden Datenlisten mittels der Variationskoeffizienten.

Punkte	52	46	41	48	42	47	50	56
Note	2	3	2	2	3	2	2	1

11/28 Bestimmen Sie für das Beispiel 11/25 die 1. Regressionsgerade und schätzen Sie die voraussichtliche Masse für eine Körpergröße von 180 cm.

11/29 Bestimmen Sie für das Beispiel 11/26 die 1. Regressionsgerade und schätzen Sie den voraussichtlichen Umsatz für Werbungskosten in Höhe von ÖS 450000,-.

11/30 Bestimmen Sie für das Beispiel 11/27 die 1. Regressionsgerade und schätzen Sie die voraussichtliche Note für eine Punktezahl von 44 Punkten.

11/31 Bestimmen Sie für das Beispiel 11/25 die 2. Regressionsgerade und schätzen Sie die voraussichtliche Körpergröße für eine Masse von 65 kg.

11/32 Bestimmen Sie für das Beispiel 11/26 die 2. Regressionsgerade und schätzen Sie die voraussichtlichen Werbungskosten für einen Umsatz in Höhe von 10 Millionen.

11/33 Bestimmen Sie für das Beispiel 11/27 die 2. Regressionsgerade und schätzen Sie die voraussichtliche Punktezahl für die Note 3.

11/34 Bestimmen Sie für das Beispiel 11/25 den Korrelationskoeffizienten.

11/35 Bestimmen Sie für das Beispiel 11/26 den Korrelationskoeffizienten.

11/36 Bestimmen Sie für das Beispiel 11/27 den Korrelationskoeffizienten.

11/37 Von 12 Schüler wurden die Noten aus Mathematik und Datenverarbeitung erhoben. Bestimmen Sie, ob ein Zusammenhang zwischen den Noten besteht.

Schüler	1	2	3	4	5	6	7	8	9	10	11	12
M	3	3	4	3	5	2	1	3	4	3	2	4
DV	3	4	2	1	4	1	1	2	5	2	1	3

11/38 Untersuchen Sie, ob sich anhand der nachstehenden Tabelle bei den befragten Personen ein Zusammenhang zwischen Schulbildung und Einstellung gegenüber Ausländern ableiten läßt bzw. wie stark dieser ist.

	Schulbildung		
Einstellung	Hauptschule	Matura	Studium
negativ	84	34	12
neutral	126	42	34
positiv	72	35	14

11/39 Je 100 Männer und Frauen wurden befragt, ob sie lieber einen Sohn oder eine Tochter hätten. Dabei entschieden sich 27 Männer und 52 Frauen für einen Sohn, 48 Männer und 26 Frauen für eine Tochter. Die anderen waren unentschlossen. Besteht ein Zusammenhang zwischen dem Geschlecht des Befragten und dem erwünschten Geschlecht des Kindes und wie groß ist dieser?