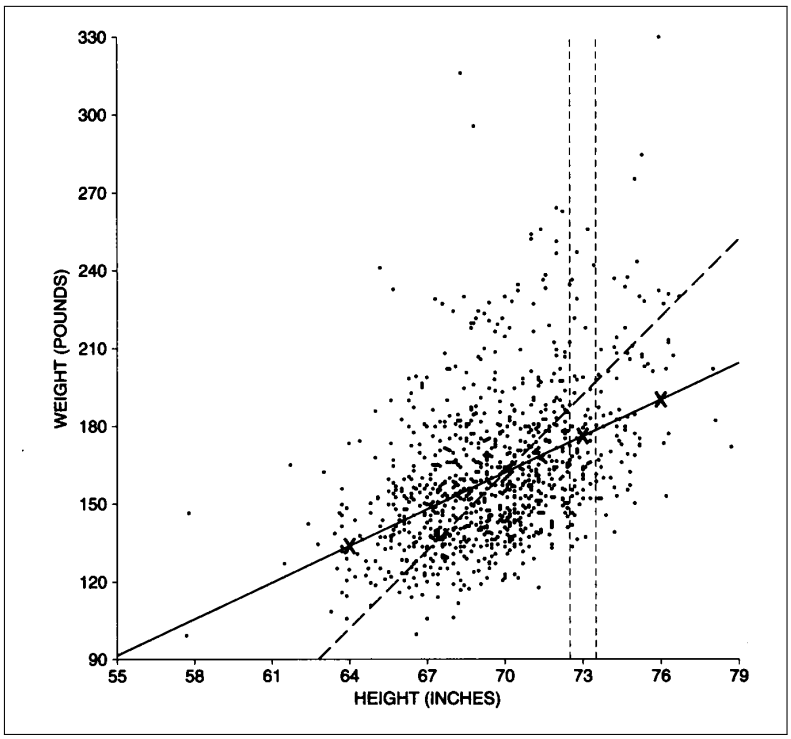
# Statistik

M. Kriener

15. September 2017


Drawing by Dana Fradon; © 1977 The New Yorker Magazine, Inc.



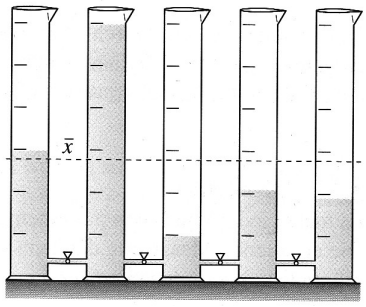
# Inhaltsverzeichnis

# 1 Lage und Streuung von Daten

## 1.1 Der Mittelwert - das arithmetische Mittel

> *It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.*
>
> —SIR FRANCIS GALTON (ENGLAND, 1822–1911)[1]

In der Statistik geht es um Daten. Im einfachsten Fall bestehen diese aus einer Liste von Zahlen $x_1, x_2, x_3, \ldots, x_n$ (zum Beispiel aus Ihren Mathematiknoten aus dem letzten Schuljahr). Die Anzahl $n$ dieser Zahlen kann sehr gross sein, man fasst die Daten deshalb mit Hilfe einer Zahl zusammen: Der **Mittelwert** (auch **Durchschnitt** oder **arithmetisches Mittel**) einer Liste von Zahlen ist ihre Summe, geteilt durch ihre Anzahl:




$$\mu := \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad (1)$$

Die Summe der $n$ Einzelwerte $x_k$ kann man sich ersetzt denken durch $n$ gleich grosse Werte von der Grösse des Mittelwertes.

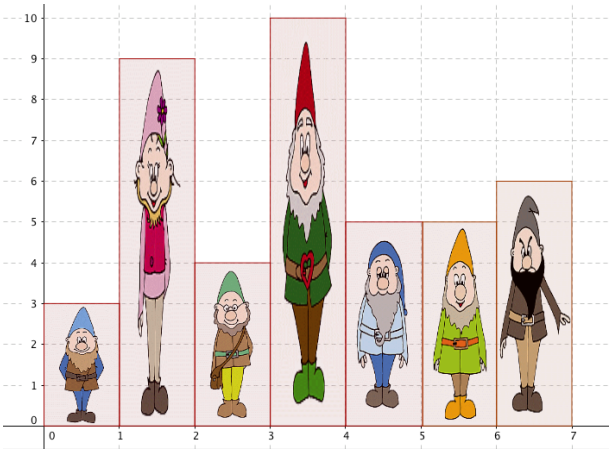
Das $\mu$ ist der griechische Buchstabe "mü". Manchmal bezeichnet man den Mittelwert einer Liste $\{x_k\}$ auch mit $\bar{x}$. Eine charakteristische Eigenschaft des Mittelswertes ist folgende: Die **Abweichungen** $x_k - \mu$ der Daten von ihrem Mittelwert summieren sich zu null auf. Das wird in der Figur oben veranschaulicht. Es gilt die sogenannte **Schwerpunkteigenschaft**:

$$\sum \text{Abweichungen} = \sum_{k=1}^{n}(x_k - \mu) = 0. \qquad (2)$$

Man könnte (2) auch die **Robin-Hood-Gleichung** nennen: Der Mittelwert "nimmt von den Reichen und gibt den Armen".

**Übung 1.1.** Die Wasserstände in den Glasröhren im Bild oben sind 30, 60, 10, 21 und 19 mm. Überprüfen Sie die Aussagen im Kasten anhand dieser Daten. Wie könnte man Gleichung (2) allgemein beweisen?

**Übung 1.2.** a) Bestimmen Sie den Mittelwert der sieben Zwerge im Bild rechts, indem Sie ihn erst schätzen, und dann versuchen, die Blöcke oberhalb auf die Lücken unterhalb von ihrem Schätzwert zu verteilen.




b) Ist der Mittelwert der folgenden Liste genau 8 oder eher etwas grösser? Antworten Sie, ohne den Mittelwert auszurechnen.

9 9 6 9 11 4 8 9 8

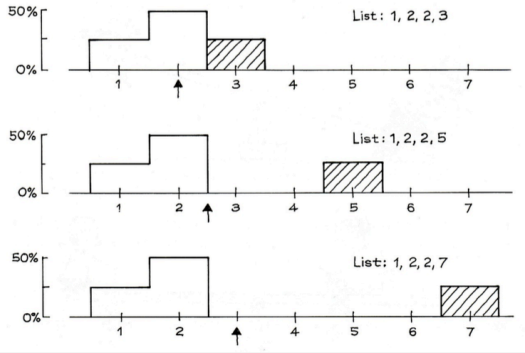
c) Wie muss man die Grösse des 7-ten Zwerges abändern, damit sich der Mittelwert um exakt 1 dm erhöht? Wie lautet die Antwort auf diese Frage für den vierten Zwerg?

**Übung 1.3.** Beantworten Sie die folgenden Fragen zunächst ohne Rechnung, dann überprüfen Sie Ihre Antwort mit Hilfe einer Rechnung.

a) Zehn Personen in einem Raum haben eine Durchschnittsgrösse von 1.67 m. Eine 11. Person betritt den Raum. Sie ist 1.78 m gross. Bestimmen Sie den Mittelwert der Körpergrössen der 11 Personen.

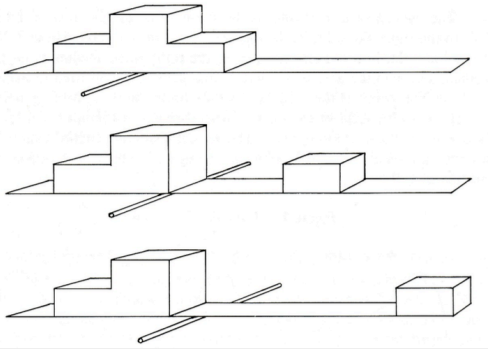
b) 21 Personen in einem Raum haben eine Durchschnittsgrösse von 1.72. Eine 22. Person betritt den Raum. Wie gross müsste sie sein, damit die Durchschnittgrösse um genau einen Zentimeter erhöht wird?

**Übung 1.4.** Unten sehen Sie eine weitere Möglichkeit zur Veranschaulichung des Mittelwertes. Veranschaulichen Sie die Listen rechts analog (die linke Version genügt).



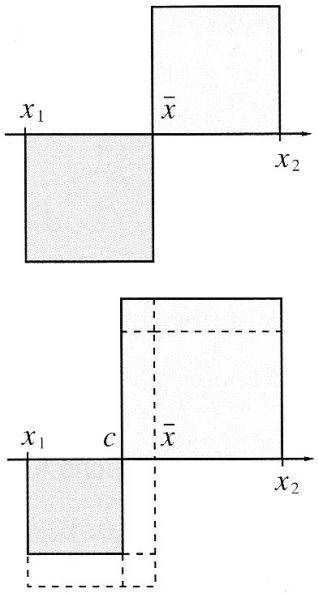

a) 1, 3, 3, 5 b) 1, 3, 3, 7

c) 1, 1, 1, 1, 6



**Übung 1.5.** Gegeben seien zwei Listen $(x_k)$ und $(y_k)$ der Länge $n$ mit den Mittelwerten $\mu_x$ und $\mu_y$. Was ist der Mittelwerte der Liste $(x_k + 2y_k)$?

**Übung 1.6.** Der Mittelwert hat eine interessante geometrische Eigenschaft, die später eine wichtige Rolle spielen wird.




a) Angenommen, man hat zwei Datenpunkte $x_1$ und $x_2$ und trägt diese auf die $x$-Achse eines Koordinatensystems auf. Dann liegt der Mittelwert $\mu$ genau in der Mitte zwischen $x_1$ und $x_2$. Warum ist das so?

b) Man bildet für irgendeinen Wert $c$ die Quadrate über den Strecken zwischen $c$ und $x_1$ bzw. $x_2$. Dann ist die Summe der Flächen dieser Quadrate am kleinsten, wenn $c$ gleich dem Mittelwert ist. Begründen Sie diese Aussage geometrisch mit Hilfe der Figuren rechts.

c) Bestätigen Sie nun die Aussage aus a) auch algebraisch. Anleitung: Sie müssen zeigen, dass

$$(x_1 - c)^2 + (x_2 - c)^2 = \Big((x_1 - \mu) + (\mu - c)\Big)^2 + \Big(x_2 - \mu) + (\mu - c)\Big)^2$$

grösser ist als $(x_1 - \mu)^2 + (x_2 - \mu)^2$. Hierbei wird die "Robin-Hood-Gleichung" (2) auf Seite 3 nützlich sein.

d) Zeigen Sie analog, dass allgemein gilt:

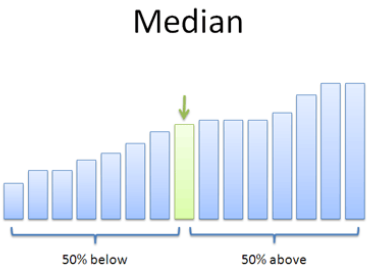
$$\sum_{k=1}^{n}(x_k - c)^2 > \sum_{k=1}^{n}(x_k - \mu)^2 \quad \text{für } c \neq \mu$$

d.h.: **Der Mittelwert minimiert die Summe der quadrierten Abweichungen**.

## 1.2 Noch ein Mittelwert - der Median

Manchmal ist ein anderer Mittelwert geeigneter, der **Median** oder **Zentralwert**. Der Median $\tilde{x}$ einer Liste $x_1, x_2, \ldots, x_n$ ist der Wert, der an der mittleren (zentralen) Stelle steht, wenn man die Werte der Grösse nach sortiert.

**Beispiel:** Der Median der Liste $4, 7, 3, 9, 1$ ist die Zahl 4, nämlich die mittlere Zahl in $1, 3, \underline{4}, 7, 9$. Bei einer Liste mit einer geraden Anzahl von Einträgen wie zum Beispiel $1, 2, 3, 4$ ist der Median das arithmetische Mittel der beiden mittleren, hier also $\tilde{x} = 2.5$.




**Bemerkungen:**

1) Der Median unterteilt eine Liste in zwei gleich grosse Hälften - rechts und links vom Median stehen gleich viele Datenwerte.
2) Neben dem arithmetischen Mittel und dem Median gibt es noch weitere Mittelwerte, zum Beispiel das **geometrische Mittel** oder das **harmonische Mittel**. Wenn einfach vom "Mittelwert" die Rede ist, meint man meist das arithmetische Mittel.

**Übung 1.7.** Ob eher der Mittelwert oder der Median als "typisch" angesehen wird, hängt von dem Datensatz ab. Wir betrachten als Beispiel die Einkommensverteilung dreier hypothetischer Miniländer $A, B$ und $C$ mit jeweils genau 5 Einwohnern, deren Einkommen in der Tabelle rechts aufgelistet ist. Bestimmen Sie in jedem Fall das arithmetische Mittel $\mu$ und den Median $\tilde{x}$. Welchen der beiden Mittelwerte würde man jeweils als typisch auffassen? Charakterisieren Sie Situationen, in denen man den Median als Mittelwert vorziehen wird.

| | A | B | C |
|---|---|---|---|
| $x_1$ | 3000 | 2500 | 80 |
| $x_2$ | 6000 | 2700 | 90 |
| $x_3$ | 1000 | 2800 | 120 |
| $x_4$ | 2100 | 2900 | 110 |
| $x_5$ | 1900 | 3100 | 13600 |
| $\mu$ | | | |
| $\tilde{x}$ | | | |

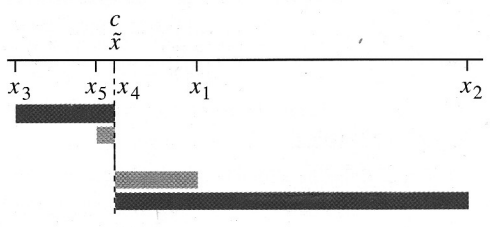
**Übung 1.8.** Wir haben gesehen, dass der Mittelwert die Summe der quadratischen Abweichungen minimiert, vgl. Übung 1.6. Auch der Median lässt sich durch eine Minimierungseigenschaft charakterisieren:

**Der Median minimiert die Summe der absoluten Abweichungen.**

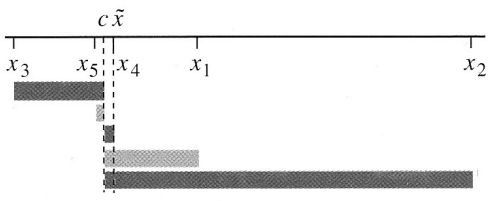
Mit den *absoluten Abweichungen* sind die Beträge (Absolutwerte) der Abweichungen gemeint. Die Summe

$$g(c) = \sum_{k=1}^{n} |x_k - c|$$

wird also dann minimal, wenn $c = \tilde{x}$. Begründen Sie diese Aussage anhand der beiden Diagramme rechts.




Abweichungen im Fall $c = \tilde{x}$

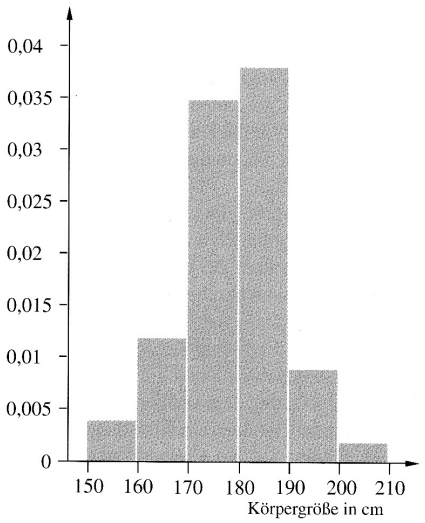



Abweichungen im Fall $c \neq \tilde{x}$

## 1.3 Histogramme

Ein **Histogramm** stellt ähnlich wie ein Säulendiagramm eine Häufigkeitsverteilung von Daten durch Rechtecke dar, allerdings mit dem Unterschied, dass hier die relative Häufigkeit einer Kategorie nicht mehr durch die Höhe eines Rechtecks ausgedrückt wird, sondern durch seinen Flächeninhalt. Die Gesamtfläche eines Histogramms entspricht also immer 100%. Das Histogramm rechts stellt die Verteilung der Körpergrössen von 100 Schülerinnen und Schülern eines deutschen Gymnasiums im Jahr 1998 dar.

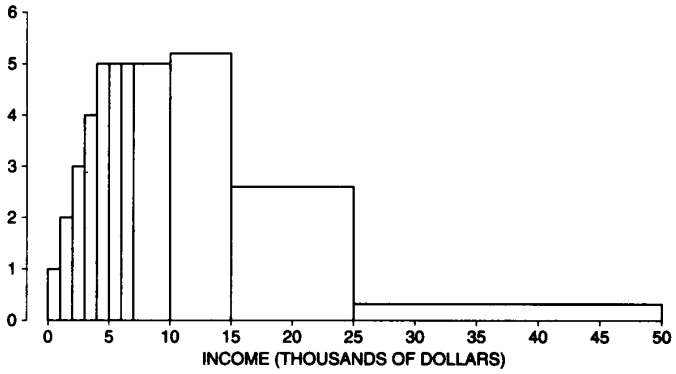
Wie liest man ein solches Histogramm? Der Flächeninhalt des linken Rechtecks ergibt sich als Produkt seiner Breite, also 10, und seiner Höhe, also 0.004. Dieser Flächeninhalt $10 \cdot 0.004 = 0.04$ - ausgedrückt in den Einheiten des Koordinatensystems - entspricht genau der relativen Häufigkeit von 4%.
Die Höhe eines Rechtecks steht in einem Histogramm nicht mehr für die relative Häufigkeit, sondern für die sogenannte **Häufigkeitsdichte**, d.h. die prozentuale Häufigkeit pro Einheit. Diese muss noch mit der Intervallbreite der betreffenden Kategorie oder Klasse multipliziert werden, um deren relative Häufigkeit zu ergeben.




**Übung 1.9.** a) Ungefähr 1% der Familien in dem Histogramm rechts hatten ein Einkommen unter \$1′000. Schätzen Sie den Prozentsatz der Familien mit einem Einkommen –

i) zwischen \$1′000 und \$2′000
ii) zwischen \$2′000 und \$3′000
iii) zwischen \$4′000 und \$7′000
iv) zwischen \$7′000 und \$10′000

b) Gab es in dem Histogramm mehr Familien mit Einkommen zwischen \$10′000 und \$11′000 oder zwischen \$15′000 und \$16′000? Oder waren die Zahlen ungefähr gleichgross. Geben Sie eine möglichst gute Schätzung ab.

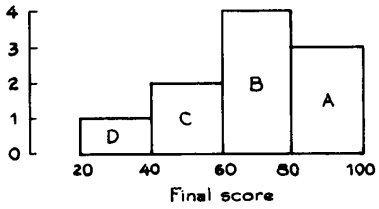



Verteilung der Familien in den USA nach dem Einkommen im Jahr 1973.

**Übung 1.10.** Das Histogramm rechts zeigt die Verteilung der Abschlussnoten in einer bestimmten Klasse. Es sind jeweils 2 Punkte eine Einheit.

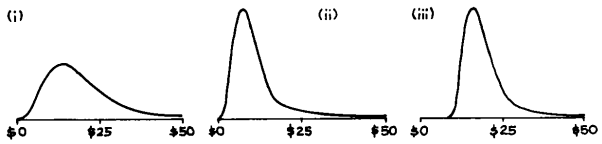
a) 10% erreichten zwischen 20 und 40 Punkten. Wie gross ist der Prozentsatz derjenigen, die zwischen 40 und 60 Punkten erreichten?

b) Wieviel Prozent erreichten mehr als 60 Punkte?

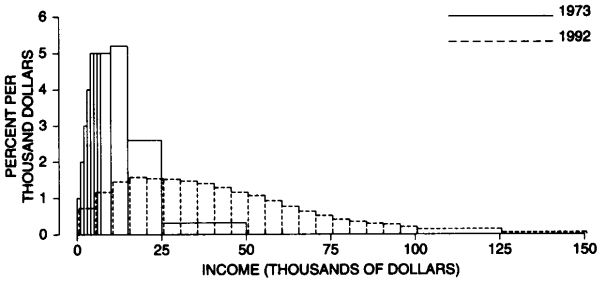



**Übung 1.11.** Überzeugen Sie sich davon, dass die Zeichnungen in Übung 1.4 Histogramme sind. Was ist also vermutlich der Zusammenhang zwischen Histogrammen und dem Mittelwert?

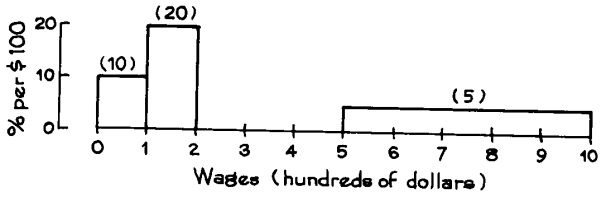
**Übung 1.12.** Jemand sammelt Daten über den Stundenlohn von drei Gruppen von Arbeitern. Die Arbeiter in Gruppe B verdienen ungefähr doppelt so viel wie die in Gruppe A; die Arbeiter in Gruppe C verdienen etwa \$10 in der Stunde mehr als die in Gruppe A. Welches Histogramm gehört zu welcher Gruppe?



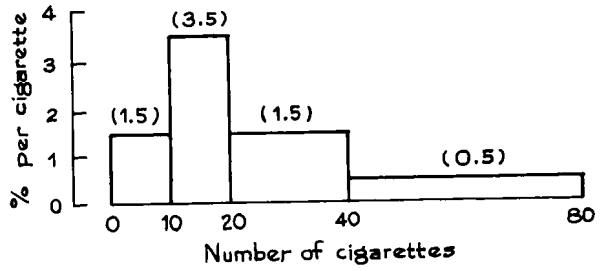

**Übung 1.13.** Die Abbildung unten vergleicht die Histogramme für die Familieneinkommen in den USA in den Jahren 1973 und 1992. Es sieht so aus, als hätte sich das Familieneinkommen in den 20 Jahren etwa verdreifacht. Oder nicht? Diskutieren Sie das.




**Übung 1.14.** Rechts finden Sie ein Histogramm der monatlichen Löhne für Teilzeitangestellte in den USA (Häufigkeitsdichten sind in Klammern angegeben). Niemand verdiente mehr als \$1000 im Monat. Das Rechteck über dem Klassenintervall von \$200 bis \$500 fehlt. Wie hoch muss es sein?




**Übung 1.15.** In einer *Public Health Study* wurde das Histogramm unten gezeichnet, das die Zahl der Zigaretten anzeigt, die die Probanden (männlich) täglich rauchen. Die Dichte ist in Klammern angegeben. Endpunktkonvention: Die Klasseninteralle enthalten die linken Endpunkte, aber nicht die rechten. Wie gross ist der Prozentsatz derjenigen, die täglich




a) zehn Zigaretten oder weniger rauchen? (Vorsicht, die Antwort ist *nicht* 15%!)

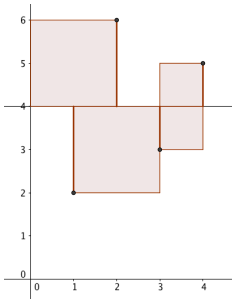
b) ein Päckchen (enthält 20 Zigaretten) oder mehr rauchen?

**Übung 1.16.** Die Tabelle rechts gibt die Verteilung des Bildungsniveaus (Damit ist die Anzahl der Jahre gemeint, die jemand in öffentlichen Schulen zugebracht hat. Kindergarten zählt nicht) der 25-jährigen in den USA in den Jahren 1960, 1970 und 1991 wieder. Endpunktkonvention: Die Klassenintervalle enthalten den linken Endpunkt, aber nicht den rechten. Zeichnen Sie (mit verschiedenen Farben) die drei Histogramme in ein Koordinatensystem wie in Übung 1.13.

| Anzahl Schuljahre | 1960 | 1970 | 1991 |
|---|---|---|---|
| 0-5 | 8 | 6 | 2 |
| 5-8 | 14 | 10 | 4 |
| 8-9 | 18 | 13 | 4 |
| 9-12 | 19 | 19 | 11 |
| 12-13 | 25 | 31 | 39 |
| 13-16 | 9 | 11 | 18 |
| 16 und mehr | 8 | 11 | 21 |

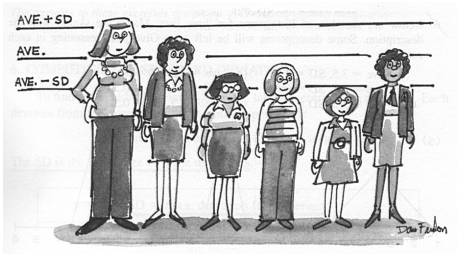
## 1.4 Die Standardabweichung

Bei einer Liste von Zahlen $\{x_1, x_2, x_3, x_4\} = \{2, 6, 3, 5\}$ (das könnten zum Beispiel Ihre letzten Mathematiknoten sein) sind wir nicht nur am Mittelwert interessiert (der ist hier $\mu = 4$), sondern auch an der Streuung der Daten - hier hat man den Eindruck, dass die Noten ziemlich hin- und herspringen. Bei einem anderen Schüler sind die Noten $(4.5, 3.5, 4, 4)$. Der gleiche Mittelwert, aber viel weniger Streuung. Wie können wir das Ausmass der Streuung mit einer Zahl erfassen? Einfach den Durchschnitt der **Abweichungen** $x_k - e\mu$ berechnen funktioniert nicht, denn $\Sigma(x_k - \mu) = 0$ (das ist die Robin-Hood-Gleichung). Wir möchten ausserdem, dass grosse Abweichungen stärker "zählen" sollen als kleine.




Wir haben in Übung 1.6 gesehen, dass der Mittelwert $\mu$ die Summe der quadratischen Abweichungen $\sum(x_k - \mu)^2$ minimiert. Wir nennen den Mittelwert dieser Summe die **Varianz** und die Wurzel der Varianz die **Standardabweichung** einer Liste von Daten:

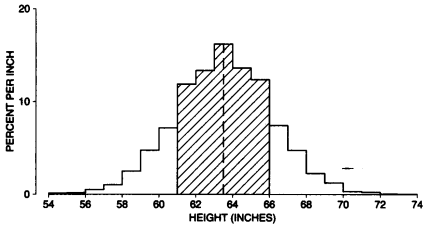
$$\sigma := \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2} \tag{3}$$



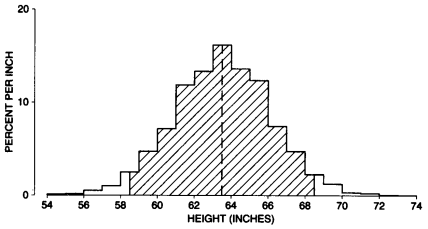

**Beispiel:** Für die Liste $(x_1, x_2, x_3, x_4) = (2, 6, 3, 5)$ von oben (Noten mit dem Mittelwert $\mu = 4$) ist die Standardabweichung

$$\sigma = \sqrt{\frac{2^2 + 2^2 + 1^2 + 1^2}{4}} = \sqrt{2.5} \approx 1.58$$

Was bedeutet diese Zahl? Ganz grob kann man sagen, dass die Standardabweichung "Normalität" misst - alles innerhalb *einer* Standardabweichung $\sigma$ entfernt vom Durchschnitt ist noch "normal".




**Beispiel:** Die beiden Histogramme rechts zeigen zweimal die Grössenverteilung amerikanischer Frauen. Die gestrichelte Linie ist der Mittelwert, und der schattierte Bereich im oberen Histogramm ist der Bereich, der maximal eine Standardabweichung $\sigma$ von $\mu$ entfernt ist - das sind etwa zwei Drittel aller Frauen. Der schattierte Bereich im unteren Histogramm ist der Bereich, der maximal zwei Standardabweichungen $2\sigma$ von $\mu$ entfernt ist - das sind schon 95% der Frauen.




Es gibt eine Daumenregel, die bei sehr vielen Daten gut funktioniert, nämlich bei **normalverteilten** Daten. Eigenschaften sind dann normalverteilt, wenn sie das Resultat von vielen voneinander unabhängigen Einflüssen sind. Körpergrösse und Intelligenz sind zum Beispiel normalverteilt. Die Daumenregel geht so:

- etwa 68% der Daten liegen innerhalb einer Standardabweichung vom Mittelwert - d.h. im Intervall $[\mu - \sigma; \mu + \sigma]$ liegen etwa 68% der gesamten Daten. Das entspricht der schattierten Fläche im oberen Histogramm.
- etwa 95% liegen innerhalb von zwei Standardabweichungen vom Mittelwert - d.h. im Intervall $[\mu - 2\sigma; \mu + 2\sigma]$ liegen etwa 95% der gesamten Daten. Das entspricht der schattierten Fläche im unteren Histogramm.
- etwa 99.7% einer Population liegen innerhalb von drei Standardabweichungen vom Mittelwert. Hier müsste man praktisch die gesamte Fläche schattieren.

**Übung 1.17.** Berechnen sie die Standardabweichung der folgenden Listen:

a) $\{4,\ 5,\ 4.2,\ 4.8\}$

b) $\{10, 8, 13, 11, 12, 9, 7\}$

**Übung 1.18.** Was macht man, wenn die Listen sehr lang sind? Eine Schachtel enthält eine unbekannte Anzahl Lose. Wenn Sie zufällig ein Los ziehen, gewinnen Sie den Betrag $x_k$, der darauf notiert ist. Die Gewinne sind folgendermassen verteilt:

| $x_k$ = Betrag auf dem Los | 0\$ | 1\$ | 10\$ |
|---|---|---|---|
| $p(x_k)$ = Prozentsatz der Lose mit $x_k$ | 50% | 40% | 10% |

a) Finden sie eine konkrete Schachtel (also mit einer selbstgewählten Anzahl von Losen) mit der obigen Verteilung.

b) Berechnen Sie Mittelwert $\mu$ und Standardabweichung $\sigma$ für die beiden Schachtel.

c) Wie könnte man $\mu$ und $\sigma$ direkt aus der Tabelle berechnen, ohne den Umweg über eine Schachtel?

d) Probieren Sie das anhand der folgenden Tabelle aus:

| $x_k$ | 2 | 5 | 1000 |
|---|---|---|---|
| $p(x_k)$ | 43% | 55% | 2% |

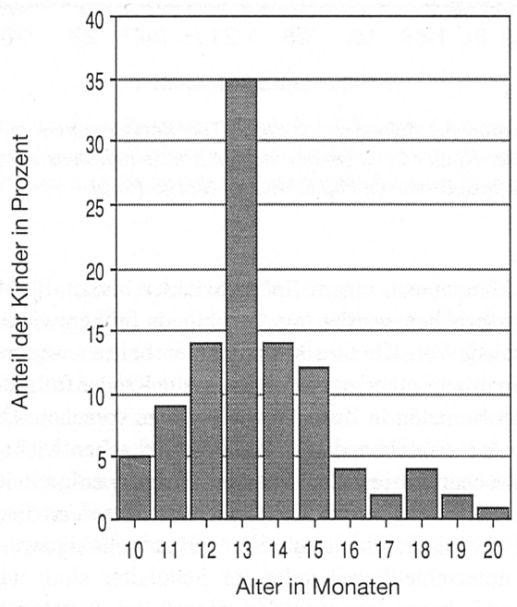
e) Beschreiben Sie die Bedeutung von $\mu$ und $\sigma$ in eigenen Worten.

**Übung 1.19.** Sie haben sicher schon von Intelligenztests gehört. Dabei wird einer Person eine Zahl zugeordnet (der IQ), diese Zahl soll ein Mass für die Intelligenz der Person sein. Jedem IQ, zum Beispiel $k$, entspricht ein bestimmter Prozentsatz $p_k$ einer Population (z.B. der Schüler einer bestimmten Schule), welche den IQ $k$ haben. Man erhält also eine Tabelle, die einem sagt, wie die Intelligenz innerhalb der Population verteilt ist. Technisch nennt man das eine **Wahrscheinlichkeitsverteilung**.

| $k$ = IQ | 1 | 2 | ... | 115 | ... |
|---|---|---|---|---|---|
| $p_k$ = Prozentsatz mit dem IQ $k$ | $p_1$ | $p_2$ | ... | $p_{115}$ | ... |

IQ-Tests sind so konstruiert, dass die Ergebnisse für eine hinreichend grosse Bevölkerungsstichprobe annähernd normalverteilt sind. Und zwar so, dass der Mittelwert immer 100 ist, und die Standardabweichung 15. Verwenden Sie die Daumenregel über die Standardabweichung, um folgende Fragen zu beantworten: Wie viele von 1000 zufällig ausgewählten Personen haben vermutlich einen IQ über 130? Über 145? (Vorsicht: Diese 1000 Personen müssen aus einer repräsentativen Gruppe aus allen Bevölkerungsschichten gewählt werden, bei 1000 SuS der KSWE sieht die Antwort vermutlich anders aus!)

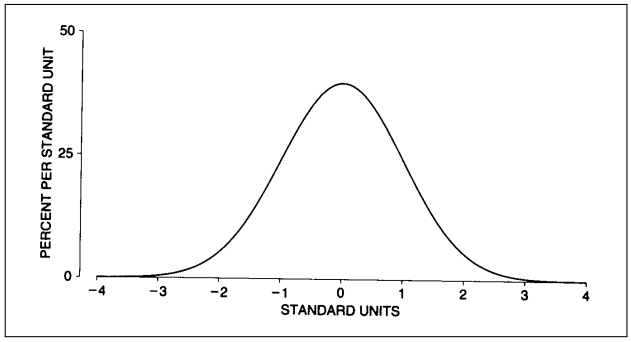
**Übung 1.20. Geh-Alter:** Läuft er schon? Wie, immer noch nicht? *Remo Largo*, Professor für Kinderpsychologie in Zürich, hat 1985 eine Studie durchgeführt, bei der das Alter untersucht wurde, in dem Kleinkinder ihre ersten Schritte machen. Rechts sehen Sie die Ergebnisse in Form eines Histogramms.

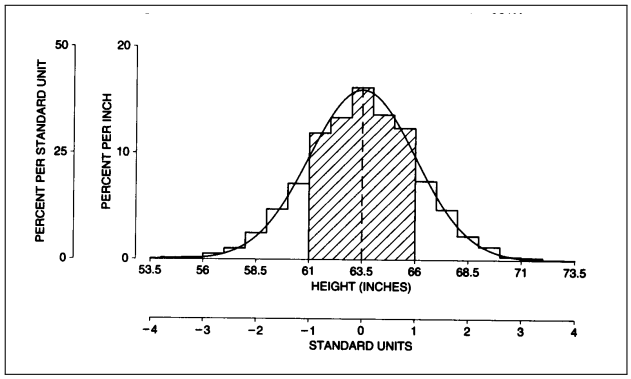



Berechnen Sie MW und Standardabweichung für diese Daten. Beachten Sie dabei, dass die Angaben in Prozent gemacht werden. Man kann diese Daten aber leicht als Liste interpretieren, indem man sich vorstellt, dass genau 100 Kinder untersucht worden sind, von denen also 5 schon im Alter von 10 Monaten ihre ersten Schritte gemacht haben und so weiter.
Hinweis: Die Graphik ist leider nicht ganz korrekt - wenn man alles zusammenzählt, kommt man auf 102. Lesen Sie bitte die Einträge bei 16 und 18 jeweils als 3%.

# 2 Die Normalapproximation von Daten

Etwa 1870 hatte der belgische Mathematiker Adolph Quetelet die Idee, die **Normalkurve** oder **Gauss'sche Glockenkurve** (die schon 150 früher von dem Franzosen Abraham de Moivre entdeckt wurde) als "Standardhistogramm" aufzufassen, mit der man die Histogramme **normalverteilter** Daten vergleichen kann. Eigenschaften sind dann normalverteilt, wenn sie das Resultat von vielen voneinander unabhängigen Einflüssen sind.


Die Normalkurve hat die Gleichung

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Diese Gleichung ist insofern bemerkenswert, als sie drei der bekanntesten mathematischen Konstanten beinhaltet: $\sqrt{2}$, $\pi$ und die **Eulerkonstante** $e = 2.71828\ldots$. Sie werden sehen, dass man mit der Normalkurve wunderbar arbeiten kann (indem man Tabellen und Graphiken benutzt), ohne die Gleichung zu verwenden.

Das obere Diagramm zeigt den Graphen der Normalkurve, darunter sehen Sie, dass das Histogramm von Seite 8 über die Grösse amerikanischer Frauen sehr gut durch die Normalkurve angenähert wird - vorausgesetzt man zeichnet es im richtigen Massstab. Dies gelingt, indem man die Datenpunkte ($x_k$) in **Standardeinheiten** umrechnet:

$$x_k^* := \frac{x_k - \mu}{\sigma} \tag{4}$$

Die Standardeinheit $x_k^*$ sagt einem, wie viele Standardabweichungen $\sigma$ der Datenpunkt $x_k$ vom Mittelwert $\mu$ entfernt ist.

**Übung 2.1.** In einer Abschlussprüfung wurden durchschnittlich 50 Punkte erreicht und die Standardabweichung lag bei $\sigma = 10$ Punkten.

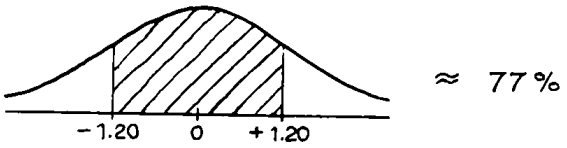
a) Rechnen Sie die folgenden Ergebnisse in Standardabweichungen um: $60, 45, 75, 89$

b) Finden Sie die Ergebnisse zu den folgenden Standardeinheiten: $0, +1.5, -2.8$.

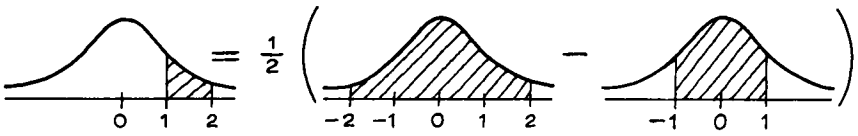
**Übung 2.2.** a) Rechnen Sie jeden Datenpunkt der Liste $\{x_1, x_2, x_3, x_4, x_5\} = \{13, 9, 11, 7, 10\}$ in Standardeinheiten um. Dazu müssen Sie natürlich zunächst das $\mu$ und das $\sigma$ berechnen.

b) Finden Sie Mittelwert und Standardabweichung der umgerechneten Liste $\{x_1^*, x_2^*, x_3^*, x_4^*, x_5^*\}$.

**Übung 2.3.** Die Tabelle auf Seite 12 erlaubt es Ihnen, den Prozentsatz der Daten abzulesen, die $\pm 1.20$ Standardabweichungen vom Mittelwert entfernt sind - dies sind ca. $A(1.20) = 77\%$. Was sind $A(0.45)$ und $A(3.15)$? Für welches $z$ gilt $A(z) = 85.29\%$?




**Übung 2.4.** Mit den Daten der Tabelle auf Seite 12 kann man auch den Prozentsatz für andere Bereiche unter dem Graphen der Normalkurve bestimmen. Mit Hilfe des folgenden Bildes dürfte es Ihnen nicht schwerfallen, zu berechnen, wie viele Prozente im Bereich zwischen 1 und 2 Standardabweichungen vom Mittelwert liegen:




Zeichnen Sie danach ein ähnliches Bild, um die Berechnung des Prozentsatzes zu illustrieren, der zwischen $-1$ und 2 Standardabweichungen liegen - und berechnen Sie ihn dann auch.

**Übung 2.5.** Bestimmen Sie jeweils die Prozentzahlen, die zu den beschriebenen Bereichen gehören und skizzieren Sie die zugehörige Fläche:

a) alles rechts von 1.25

b) alles links von $-0.40$

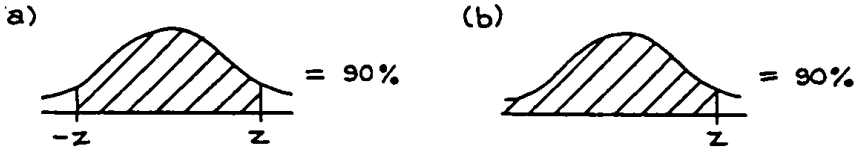
c) alles links von 0.80

d) zwischen 0.40 und 1.30

e) zwischen $-0.30$ und 0.90

f) ausserhalb von $-1.5$ bis 1.5

**Übung 2.6.** Finden Sie jeweils das zugehörige $z$:

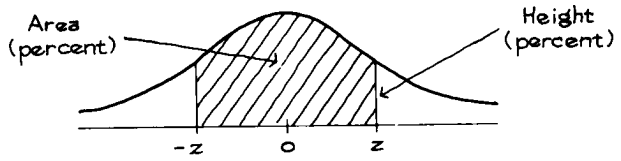



**Übung 2.7.** Finden Sie das $z$, so dass die Fläche zwischen $z$ und 1 unter der Normalkurve den folgenden Prozentzahlen entsprechen:

a) 68.27%

b) 75%

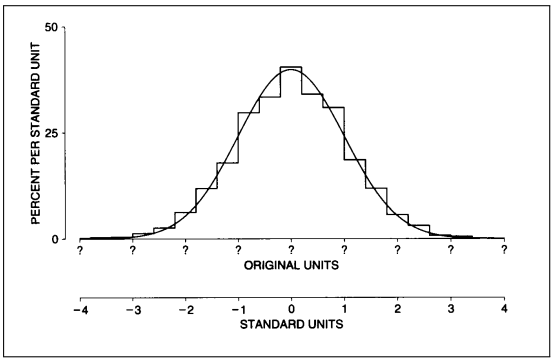
Wir bezeichnen den Prozentsatz zwischen $-z$ und $z$ (im Bild ist das der schraffierte Bereich) mit $A(z)$. Zum Beispiel ist $A(1) = 68.27\%$. Die Höhe benötigen wir nicht.




**A NORMAL TABLE**

| *z* | *Height* | *Area* | *z* | *Height* | *Area* | *z* | *Height* | *Area* |
|---|---|---|---|---|---|---|---|---|
| 0.00 | 39.89 | 0 | 1.50 | 12.95 | 86.64 | 3.00 | 0.443 | 99.730 |
| 0.05 | 39.84 | 3.99 | 1.55 | 12.00 | 87.89 | 3.05 | 0.381 | 99.771 |
| 0.10 | 39.69 | 7.97 | 1.60 | 11.09 | 89.04 | 3.10 | 0.327 | 99.806 |
| 0.15 | 39.45 | 11.92 | 1.65 | 10.23 | 90.11 | 3.15 | 0.279 | 99.837 |
| 0.20 | 39.10 | 15.85 | 1.70 | 9.40 | 91.09 | 3.20 | 0.238 | 99.863 |
| 0.25 | 38.67 | 19.74 | 1.75 | 8.63 | 91.99 | 3.25 | 0.203 | 99.885 |
| 0.30 | 38.14 | 23.58 | 1.80 | 7.90 | 92.81 | 3.30 | 0.172 | 99.903 |
| 0.35 | 37.52 | 27.37 | 1.85 | 7.21 | 93.57 | 3.35 | 0.146 | 99.919 |
| 0.40 | 36.83 | 31.08 | 1.90 | 6.56 | 94.26 | 3.40 | 0.123 | 99.933 |
| 0.45 | 36.05 | 34.73 | 1.95 | 5.96 | 94.88 | 3.45 | 0.104 | 99.944 |
| 0.50 | 35.21 | 38.29 | 2.00 | 5.40 | 95.45 | 3.50 | 0.087 | 99.953 |
| 0.55 | 34.29 | 41.77 | 2.05 | 4.88 | 95.96 | 3.55 | 0.073 | 99.961 |
| 0.60 | 33.32 | 45.15 | 2.10 | 4.40 | 96.43 | 3.60 | 0.061 | 99.968 |
| 0.65 | 32.30 | 48.43 | 2.15 | 3.96 | 96.84 | 3.65 | 0.051 | 99.974 |
| 0.70 | 31.23 | 51.61 | 2.20 | 3.55 | 97.22 | 3.70 | 0.042 | 99.978 |
| 0.75 | 30.11 | 54.67 | 2.25 | 3.17 | 97.56 | 3.75 | 0.035 | 99.982 |
| 0.80 | 28.97 | 57.63 | 2.30 | 2.83 | 97.86 | 3.80 | 0.029 | 99.986 |
| 0.85 | 27.80 | 60.47 | 2.35 | 2.52 | 98.12 | 3.85 | 0.024 | 99.988 |
| 0.90 | 26.61 | 63.19 | 2.40 | 2.24 | 98.36 | 3.90 | 0.020 | 99.990 |
| 0.95 | 25.41 | 65.79 | 2.45 | 1.98 | 98.57 | 3.95 | 0.016 | 99.992 |
| 1.00 | 24.20 | 68.27 | 2.50 | 1.75 | 98.76 | 4.00 | 0.013 | 99.9937 |
| 1.05 | 22.99 | 70.63 | 2.55 | 1.54 | 98.92 | 4.05 | 0.011 | 99.9949 |
| 1.10 | 21.79 | 72.87 | 2.60 | 1.36 | 99.07 | 4.10 | 0.009 | 99.9959 |
| 1.15 | 20.59 | 74.99 | 2.65 | 1.19 | 99.20 | 4.15 | 0.007 | 99.9967 |
| 1.20 | 19.42 | 76.99 | 2.70 | 1.04 | 99.31 | 4.20 | 0.006 | 99.9973 |
| 1.25 | 18.26 | 78.87 | 2.75 | 0.91 | 99.40 | 4.25 | 0.005 | 99.9979 |
| 1.30 | 17.14 | 80.64 | 2.80 | 0.79 | 99.49 | 4.30 | 0.004 | 99.9983 |
| 1.35 | 16.04 | 82.30 | 2.85 | 0.69 | 99.56 | 4.35 | 0.003 | 99.9986 |
| 1.40 | 14.97 | 83.85 | 2.90 | 0.60 | 99.63 | 4.40 | 0.002 | 99.9989 |
| 1.45 | 13.94 | 85.29 | 2.95 | 0.51 | 99.68 | 4.45 | 0.002 | 99.9991 |

**Übung 2.8.** In den Jahren 1976 – 80 wurde in den USA eine grossangelegte Studie mit 20′322 Teilnehmern im Alter zwischen 1 und 74 Jahren durchgeführt, die HANES (Health and Nutrition Examination Survey). Die Männer im Alter zwischen 18 und 74 waren im Durchschnitt 175.3 cm gross mit einer Standardabweichung von 7.6 cm. Ersetzen Sie die Fragezeichen im Histogramm rechts durch Körpergrössen und berechnen Sie den Prozentsatz der Männer zwischen 160 cm und 180 cm.




**Übung 2.9.** Die Durchschnittsgrösse der Frauen im Alter von 18 – 24 in HANES betrug 163.3 cm, das $\sigma$ war 6.6 cm. Berechnen Sie den Prozentsatz der Frauen :

a) kleiner als 155 cm; b) zwischen 160 cm und 175 cm; c) grösser als 180 cm.

**Übung 2.10.** Die Tabelle enthält nicht alle Werte - zum Beispiel kann man $A(0.27)$ nicht direkt ablesen. In solchen Fällen verwendet man eine sogenannte **lineare Interpolation**. Dazu geht man folgendermassen vor: 0.27 liegt zwischen 0.25 und 0.30, für diese $z$-Werte findet man $A(z)$ in der Tabelle.

a) Tragen Sie die Punktepaare $(0.25, A(0.25))$ und $(0.30, A(0.30))$ in ein (entsprechend skaliertes) Koordinatensystem ein und verbinden Sie die beiden Punkte durch eine Strecke. Machen Sie sich dann klar, dass folgendes Vorgehen einen vernünftigen Näherungswert für $A(0.27)$ liefern sollte:

$$A(0.27) = 0.6 \cdot A(0.25) + 0.4 \cdot A(0.30) = 0.6 \cdot 19.74 + 0.4 \cdot 23.58 = 21.28$$

b) Bestimmen Sie analog Näherungen für $A(1.61)$ und $A(1.63)$.

**Übung 2.11.** Von den Mitt-60iger Jahren bis in die frühen 90iger Jahren beobachtete man einen langsamen aber stetigen Abstieg in den Ergebnissen des "Scholastic Aptitude Test" (SAT) in den USA, der über die Zulassung zu den Universitäten entscheidet. In der Sprachprüfung war der Mittelwert der erreichten Punkte 1967 noch 466; im Jahr 1994 war der Mittelwert auf 423 gesunken. Die Standardabweichung änderte sich nicht, sie war nahe 110. Das Absinken der Durchschnitte hat einen grossen Effekt auf die Enden der Histogramme:

a) Berechnen Sie den Prozentsatz der Studierenden mit über 600 Punkten im Jahr 1967.

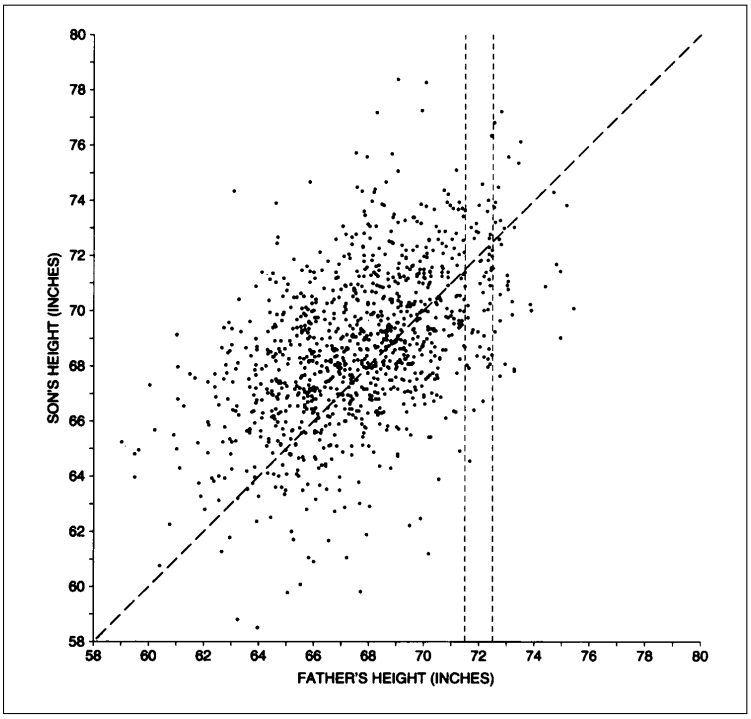
b) Berechnen Sie den Prozentsatz der Studierenden mit über 600 Punkten im Jahr 1994.

Man kann davon ausgehen, dass das Histogramm durch eine Normalkurve angenähert werden kann. SAT Punkte sind im Bereich 200 bis 800. Man kann ausschliessen, dass der SAT schwieriger geworden ist; einen grossen Teil des Absinkens in den 1960iger Jahren lässt sich auf einen Wechsel in den Studentenpopulationen zurückführen (deutlich höhere Studentenzahlen); das Absinken in den 1970iger Jahren lässt sich so nicht erklären; von 1990 bis 1994 haben sich die Durchschnitte stabilisiert.

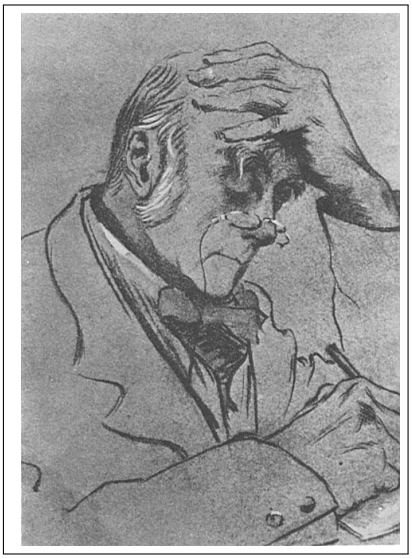
# 3 Korrelation

## 3.1 Streudiagramme

Die bisherigen Methoden waren bestens geeignet für eindimensionale Daten (man spricht auch von **univariaten Daten**). Sir Francis Galton wollte die Beziehung zwischen zwei Variablen (also von **bivariaten Daten**) untersuchen, als er über den Grad der Ähnlichkeit zwischen Eltern und ihren Kindern nachdachte. Statistiker im viktorianischen England waren fasziniert von der Vorstellung, Erbeinflüsse quantitativ zu erforschen und sammelten gewaltige Mengen an Daten in dieser Hinsicht. Wir schauen uns die Resultate einer Studie an, die Galtons Schüler Karl Pearson durchgeführt hat.




Als ein Teil seiner Studien mass Pearson die Körpergrösse von 1′078 Vätern und ihrer erwachsenen Söhnen. Eine Liste von 1′078 Zahlenpaaren wäre nicht sehr aufschlussreich. Man kann aber die Beziehung von zwei Listen sehr gut graphisch in Form eines **Streudiagramms** darstellen, siehe die Figur oben. Jeder Punkt in dem Diagramm repräsentiert ein Vater-Sohn-Paar.



Sir Francis Galton
(1822 - 1911)

Galton and his disciple Karl Pearson
(1857 - 1936)

**Übung 3.1.** Mit Pearsons Streudiagramm oben können Sie folgende Fragen beantworten. Geben Sie Ihre Antworten in a), b) und c) in cm an ( 1 inch = 1 Zoll = 2.54 cm).

a) Wie gross ist der kleinste Vater? Und dessen Sohn?

b) Wie gross ist der grösste Vater? Und dessen Sohn?

c) Schauen Sie nur die Punkte an, bei der der Vater $72 \pm 0.25$ Zoll gross war (das sind die Punkte in dem markierten Streifen). Wie gross war hier der grösste Sohn? Der kleinste Sohn?

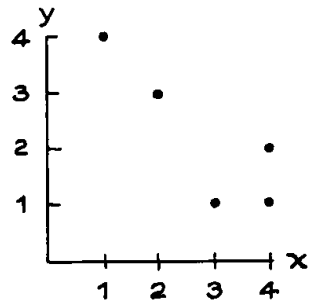
d) In wie vielen Familien war der Sohn grösser als 78 Zoll? Wie gross waren die zugehörigen Väter?

e) Lag der Mittelwert der Grösse der Väter eher bei 64, 68 oder 72 Zoll?

f) Lag die Standardabweichung der Väter eher bei 3, 6 oder 9 Zoll? Die Körpergrösse ist normalverteilt.

**Übung 3.2.** Rechts finden Sie eine Wertetabelle und das zugehörige Streudiagramm. Es sollte Ihnen keine grosse Mühe bereiten, die Lücken in der Wertetabelle zu füllen.

| $x$ | $y$ |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | – |
| – | 1 |
| – | – |



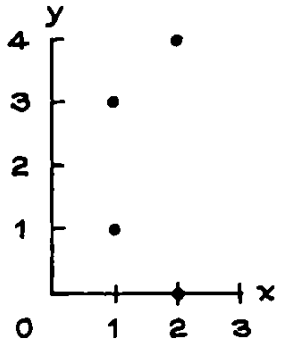

**Übung 3.3.** Rechts sehen Sie das Streudiagramm eines hypothetischen Datensatzes mit vier Datenpunkten. Schätzen Sie:

a) Ist $\mu_x$ (der Mittelwert der $x$-Werte) eher 1, 1.5 oder 2?

b) Ist $\sigma_x$ (die Standardabweichung der $x$-Werte) eher 0.1, 0.5 oder 1?

c) Liegt $\mu_y$ (der Mittelwert der $y$-Werte) bei 1, 1.5 oder 2?

d) Liegt $\sigma_y$ (die Standardabweichung der $y$-Werte) bei 0.5, 1.5 oder 3?



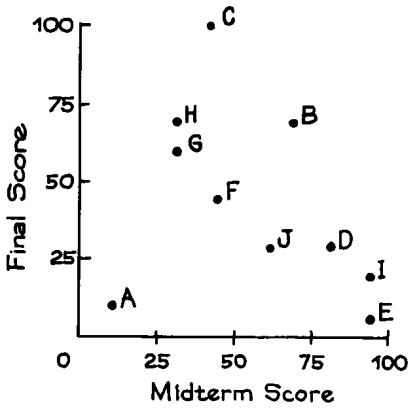

**Übung 3.4.** Studierende namens $A, B, C, \ldots H, I, J$ legten in einem Kurs in der Mitte (midterm) und am Ende (final) eines Semesters Prüfungen ab. Ein Streudiagramm mit den Ergebnissen sehen Sie rechts. Schätzen Sie wieder:

a) Welche Studenten schnitten bei beiden Prüfungen etwa gleich gut ab? Welche waren in der am Ende des Semesters besser?

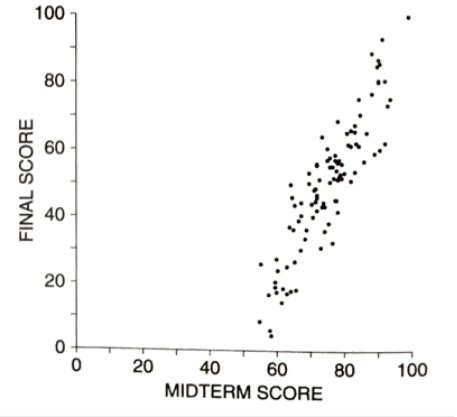
b) Lag der Mittelwert der zweiten Prüfung eher bei 10, 25, oder 50?

c) War $\sigma_y$ eher 10, 25, oder 50?

d) Lag der Mittelwert für die Studierenden, die mehr als 50 Punkte in der ersten Prüfung machten dann bei der zweiten Prüfung eher um die 30, 50, oder 70 Punkte?



**Übung 3.5.** Das Streudiagramm rechts zeigt wieder die Ergebnisse in den Prüfungen in der Mitte und am Ende eines Semesters.




a) Lag der Mittelwert der ersten Prüfung $\mu_x$ eher bei 25, 50, oder 75 Punkten?

b) War $\sigma_x$ eher 5, 10, oder 20?

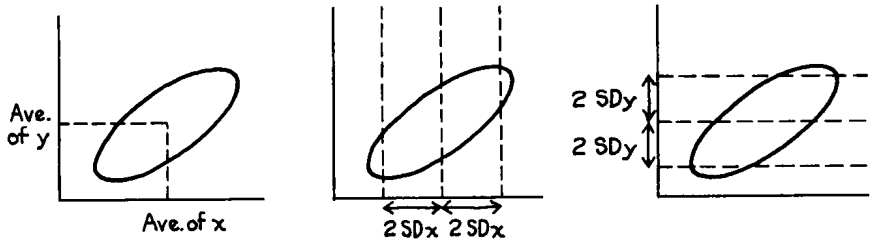
c) War $\sigma_y$ eher 5, 10, or 20?

d) Welche Prüfung war schwieriger - die erste oder die zweite?

e) War die Streuung der Ergebnisse der ersten oder bei der zweiten Prüfung grösser?

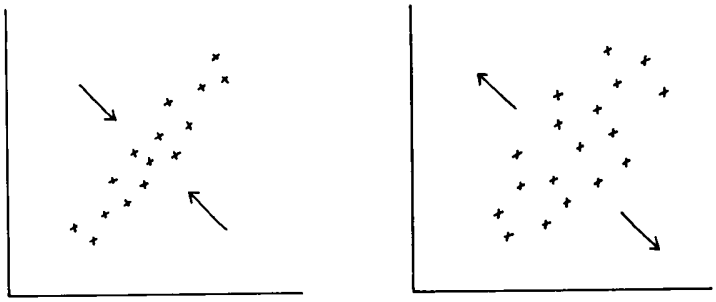
f) Wahr oder falsch: Es gab einen starken Zusammenhang zwischen den Ergebnissen der beiden Prüfungen. Begründen Sie.

**Übung 3.6.** Angenommen Sie interessieren sich für den Zusammenhang zwischen zwei Variablen und haben schon das Streudiagramm gezeichnet. Der Graph ist eine ellipsenförmige Punktwolke. Wie könnte man diese numerisch erfassen? Ein erster Schritt wäre, den Mittelwert $\mu_x$ der $x$-Werte und den Mittelwert $\mu_y$ der $y$-Werte zu markieren. Dies ergibt den **Schwerpunkt** $(\mu_x, \mu_y)$ der Punktwolke, siehe die erste der drei Skizzen unten.



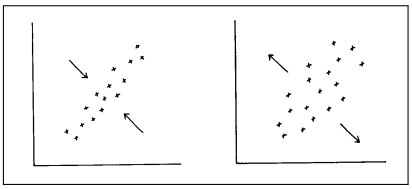

In einem zweiten Schritt würde man die Streuung der Punktwolke in horizontaler und vertikaler Richtung messen. Dazu kann man $\sigma_x$ verwenden (in den Skizzen wird die Notation $\mathrm{SD}_x$ verwendet). Die meisten der Punkte werden sich innerhalb von 2 horizontalen Standardabweichungen links und rechts vom Schwerpunkt befinden. Genauso kann man $\sigma_y$ verwenden, um die vertikale Streuung der Daten zu bestimmen.

Nun schauen Sie sich die beiden Streudiagramme unten an (und ignorieren Sie dabei für den Moment die Pfeile). Markieren Sie den Schwerpunkt und die gestrichelten Linien, die angeben, wie weit die beiden $\sigma_x$ und die beiden $\sigma_y$ reichen. Vergleichen Sie die Resultate. Was stellen Sie fest? Was folgern Sie daraus?

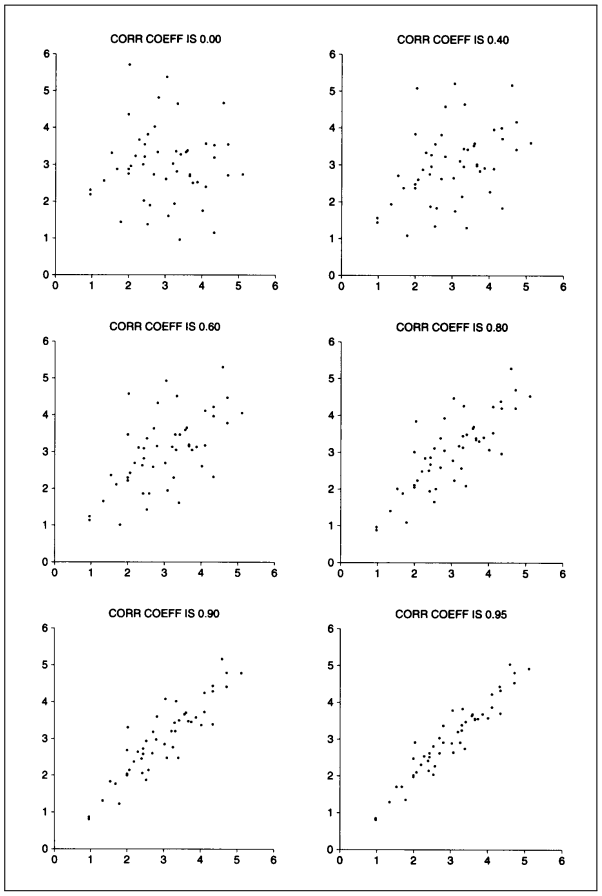


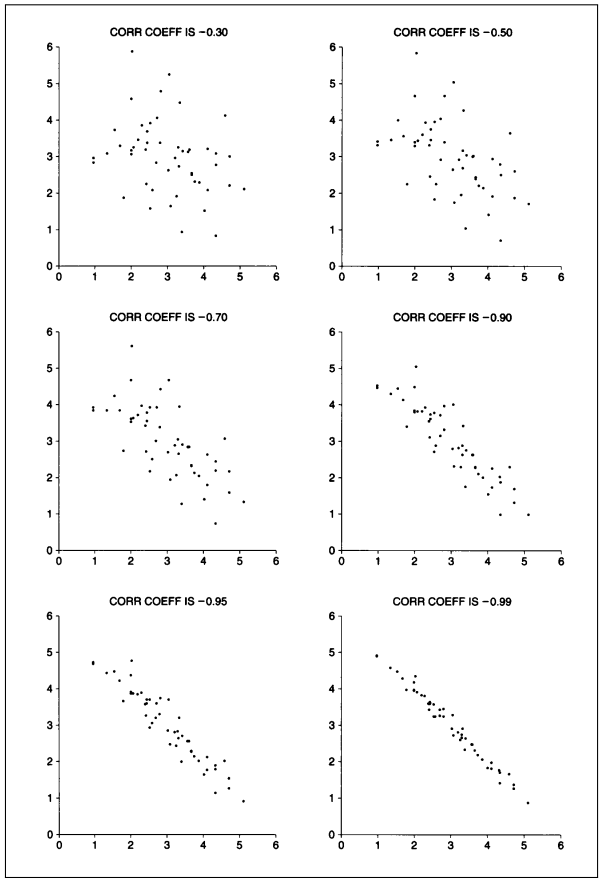
## 3.2 Der Korrelationskoeffizient r

Wie die letzte Übung 3.6 gezeigt hat, können zwei Streudiagramme dieselben Werte für $\mu_x, \mu_y, \sigma_x$ und $\sigma_y$ haben, also denselben Schwerpunkt und dieselben horizontalen und vertikalen Streuungen, und trotzdem sehr verschieden aussehen: Die Punkte in der ersten liegen dicht an einer Geraden; es besteht eine starke lineare Beziehung zwischen den beiden. In der zweiten Wolke liegen die Punkte viel lockerer.



Um diese Beziehung zu messen, benötigt man eine weitere Zahl – den **Korrelationskoeffizienten**. Dieser Koeffizient wird üblicherweise mit $r$ bezeichnet. Ohne gute Gründe – vielleicht weil zwei r's in "Korrelation" auftreten? Unten sehen Sie zwölf verschiedene Streudiagramme - sie sind so skaliert, dass in allen der Schwerpunkt $(\mu_x, \mu_y) = (3, 3)$ ist und $\sigma_x = \sigma_y = 2$, sie unterscheiden sich also nur in $r$.

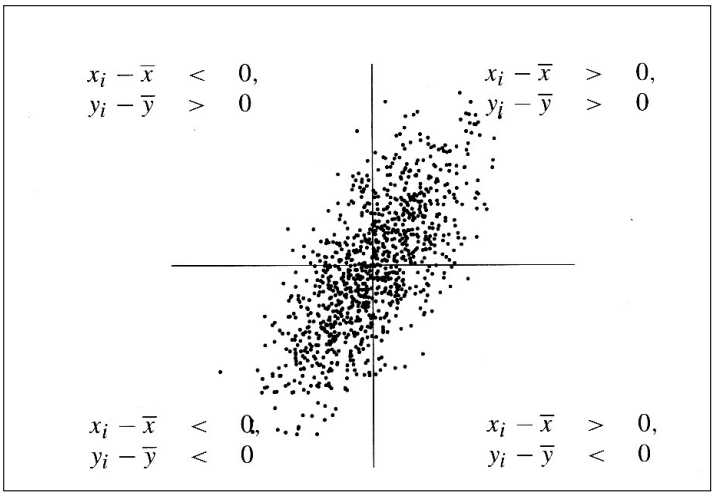



Sie sehen wie unterschiedliche Werte von $r$ die Gestalt der Wolke beeinflussen:

- Der Korrelationskoeffizient $r$ liegt immer zwischen $-1$ und 1, er kann jeden Wert dazwischen annehmen.
- Ein positives $r$ bedeutet, dass die Wolke ansteigt; werden die Werte der einen Variablen grösser, dann auch die der anderen. Je näher $r$ bei 1 liegt, desto dichter liegen die Punkte an einer Geraden.
- Für $r = 0$ haben die Punkte keinen ersichtlichen Trend - es gibt also keine Korrelation zwischen den Variablen.
- Eine negative Korrelation bedeutet, dass die Wolke eine negative Steigung hat; werden die Werte der einen Variablen grösser, dann werden die der anderen kleiner. Je näher $r$ bei $-1$, desto näher liegen die Punkte an einer fallenden Geraden.

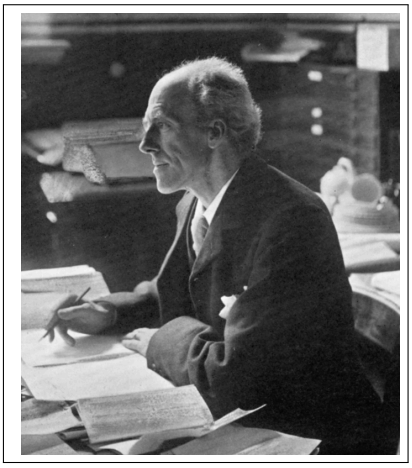
Wie können wir dieses $r$ berechnen? Zeichnen Sie eine vertikale und eine horizontale Gerade durch den Schwerpunkt wie in der Figur.




Wir berechnen alle Abweichungen $x_k - \mu_x$ und $y_k - \mu_y$ (in der Figur wird die Notation $\bar{x}$ statt $\mu_x$ und $\bar{y}$ statt $\mu_y$ benutzt) und betrachten den **Durchschnitt der Produkte der Abweichungen**, dabei nehmen wir an, dass wir $n$ Datenpunkte $(x_k, y_k)$ vorliegen haben:

$$\text{cov} := \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu_x)(y_k - \mu_y)$$

Diese **Kovarianz** hat schon einige der gewünschten Eigenschaften des Korrelationskoeffizienten:



Karl Pearson (1857 - 1936)

Er führte den Korrelationskoeffizienten in Anlehnung an verwandte Ideen von Francis Galton um 1880 ein. Als der 23-jährige Albert Einstein um 1902 zusammen mit seinen Freunden Maurice Solovine und Conrad Habicht eine Studiengruppe gründete, war sein erster Lektürevorschlag Pearsons **The Grammar of Science**. Das Buch behandelte diverse Themen, die dann auch in Einsteins bahnbrechenden Arbeiten von 1905 eine Rolle spielen sollten.

- Wenn die Datenwolke eine positive Steigung hat wie im Bild, dann sind die meisten der Produkte $(x_k - \mu_x)(y_k - \mu_y)$ positiv, denn bei den meisten Datenpunkten haben beide Faktoren dasselbe Vorzeichen. Also wird cov eine positive Zahl sein. Entsprechend wird cov negativ, wenn die Datenwolke eine negative Steigung hat.

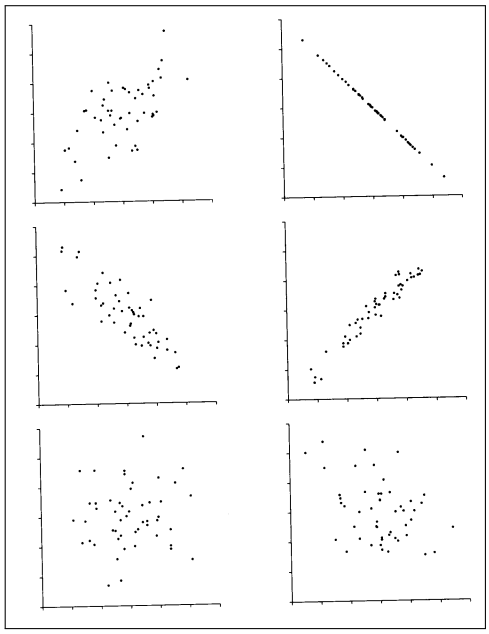
- Je enger sich die Datenpunkte an eine Gerade anschmiegen (sagen wir eine mit positiver Steigung), desto weniger Datenpunkte liegen in den Bereichen, in denen das Produkt $(x_k - \mu_x)(y_k - \mu_y)$ negativ ist. Daher wird deren Summe umso grösser, je besser die Datenpunkte (linear) korrelieren.

Leider liegt cov nicht immer zwischen $-1$ and 1, sondern ist stark abhängig von der Art der Daten. Wieder hilft der alte Trick - man rechnet die Originaldaten $(x_k, y_k)$ in Standardeinheiten um und erhält so einen reskalierten Datensatz $(x_k^*, y_k^*)$. Das Streudiagramm hat jetzt $(0, 0)$ als Schwerpunkt. Da die $x$- und $y$-Koordinaten der Datenpunkte mit evt. unterschiedlichen Standardabweichungen reskaliert werden, liegt die Symmetrieachse dieser reskalierten Daten immer auf der positiven oder negativen Winkelhalbierenden. Das "Streuverhalten" der Punktwolke ändert sich aber nicht. Wegen $\mu_x^* = \mu_y^* = 0$ sind bequemerweise die Abweichungen die Datenpunkte $x_k^*$ und $y_k^*$ selbst. Jetzt definieren wir den **Korrelationskoeffizienten** als die Kovarianz der Daten in Standardeinheiten:

$$r := \frac{1}{n}\sum_{k=1}^{n} x_k^* y_k^* \tag{5}$$

**Übung 3.7.** In der Figur rechts sehen Sie sechs Streudiagramme. Ordnen Sie sie den folgenden Korrelationskoeffizienten zu:

$$-0.85 \quad -0.38 \quad -1 \quad 0.06 \quad 0.97 \quad 0.62$$



**Übung 3.8.** Berechnen Sie den Korrelationskoeffizienten $r$ für den folgenden Datensatz $\{(x_k, y_k)\}$:

$$(1|5) \qquad (3|9) \qquad (4|7) \qquad (5|1) \qquad (7|13)$$

Dazu müssen Sie zunächst $\mu_x, \mu_y, \sigma_x$ und $\sigma_y$ bestimmen, denn Sie müssen die Datensätze $\{x_n\}$ und $\{y_n\}$ ja zunächst in Datensätze $\{x_n^*\}$ und $\{y_n^*\}$ in Standardeinheiten umwandeln (siehe S. 10). Als Resultat sollten Sie $r = 0.4$ erhalten. Am besten listen Sie Zwischenergebnisse tabellarisch auf, zur Kontrolle sind hier ist die ersten zwei Zeilen der Tabelle. Zur Berechnung von $r$ müssen Sie nur noch den Mittelwert der letzten Spalte berechnen.

| $x_i$ | $y_i$ | $x_i^*$ | $y_i^*$ | $x_i^* y_i^*$ |
|---|---|---|---|---|
| 1 | 5 | $-1.5$ | $-0.5$ | 0.75 |
| 3 | 9 | $-0.5$ | 0.5 | $-0.25$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

**Übung 3.9.** Im Jahre 1955 publizierte *R. Doll* eine bahnbrechende Arbeit über Zigarettenkonsum und Lungenkrebs im Jahre 1930 in 11 Ländern[a]. In der Tabelle rechts sehen Sie seine Ergebnisse.

a) Zeichnen Sie ein Streudiagramm für diese Daten. Tun Sie dies sorgfältig, denn Sie werden es in späteren Aufgaben benötigen.

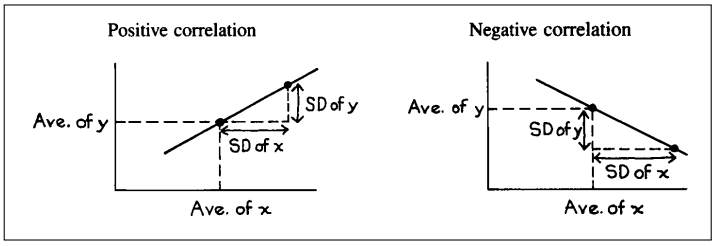
b) Berechnen Sie $\mu_x, \mu_y, \sigma_x, \sigma_y$ und $r$ für diese Daten.

[a] vgl. *Etiology of lung cancer*, Advances in Cancer Research, vol. 3, 1955, $1 - 50$.

| Land | Zigaretten-konsum | Lungenkrebstote pro Million Einwohner |
|---|---|---|
| Island | 230 | 60 |
| Norwegen | 250 | 90 |
| Schweden | 300 | 110 |
| Dänemark | 380 | 170 |
| Australien | 480 | 180 |
| Niederlande | 490 | 240 |
| Kanada | 500 | 150 |
| Schweiz | 510 | 250 |
| Finnland | 1′100 | 350 |
| Grossbritannien | 1′100 | 460 |
| USA | 1′300 | 200 |

**Übung 3.10.** Die (im normalverteilten Fall) ellipsenförmige Punktwolke eines Streudiagramms hat eine Symmetrieachse. Wie kann man diese Gerade berechnen? Diese $\sigma-$**Gerade** (das ist keine Standardbezeichnung) hat zwei Eigenschaften, mit deren Hilfe Sie ihre Gleichung finden können:

- Die $\sigma-$Gerade verläuft durch den Schwerpunkt $(\mu_x|\mu_y)$.
- Ihre Steigung ist $m = \pm\frac{\sigma_y}{\sigma_x}$. Das Vorzeichen von $m$ hängt natürlich davon ab, ob $r$ positiv oder negativ ist.




a) Berechnen Sie die $\sigma-$Gerade für einen hypothetischen Datensatz mit

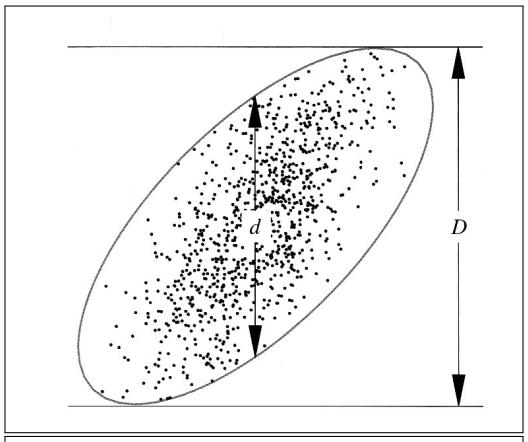
$$\mu_x = 2 \qquad \mu_y = 5 \qquad \sigma_x = 3 \qquad \sigma_y = 2 \qquad r = -0.8$$

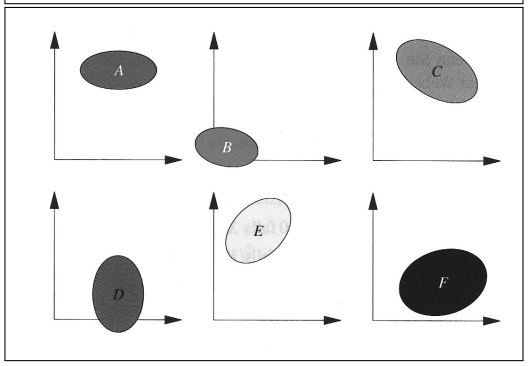
und zeichnen Sie sie in ein Koordinatensystem.

b) Berechnen Sie die $\sigma-$Gerade für die Daten in Übung 3.9 und zeichnen Sie sie in Ihr dort erstelltes Streudiagramm ein.

**Übung 3.11.** Aus einem Streudiagramm kann man grob Mittelwerte und Standardabweichungen ablesen. Man kann auch den Korrelationskoeffizienten $r$ mit Hilfe der sogenannten **Ellipsenregel** abschätzen:

- Man zeichnet eine Ellipse, mit der man ca. 95% der Daten umfasst - Ausreisser sollte man also ignorieren.
- Dann misst man die Länge der Strecken $d$ und $D$, so wie in der oberen Figur rechts illustriert.
- Nun gilt $r^2 \approx 1 - \left(\frac{d}{D}\right)^2$.
- Schliesslich berechnet man noch die Wurzel der erhaltenen Zahl und dies ergibt eine recht gute Näherung von $|r|$. (Das Vorzeichen von $r$ sollte offensichtlich sein.)




Der Beweis der Ellipsenregel involviert die sogenannte **Kovarianzmatrix** und ist ausserhalb unserer Reichweite. Aber das sollte Sie nicht daran hindern, die Regel anzuwenden:

a) Verwenden Sie die Ellipsenregel, um den Korrelationskoeffizienten für das Streudiagramm auf Seite 14 abzuschätzen.

b) Ordnen Sie die hypothetischen Daten auf dem unteren Bild rechts nach der Grösse von $r$. Beachten Sie dabei das Vorzeichen von $r$. Sie müssen dazu nichts rechnen.

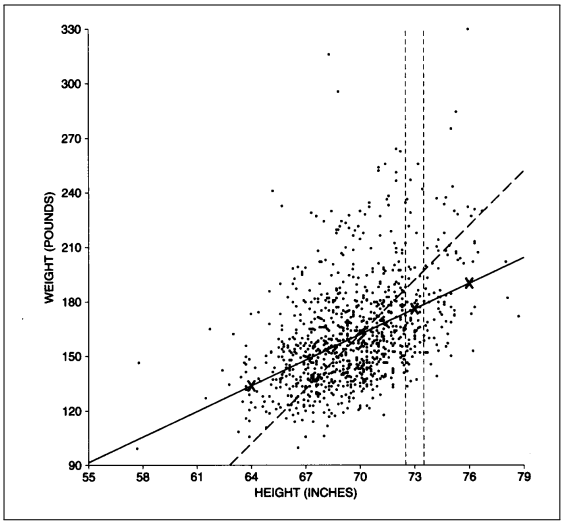
# 4 Regression

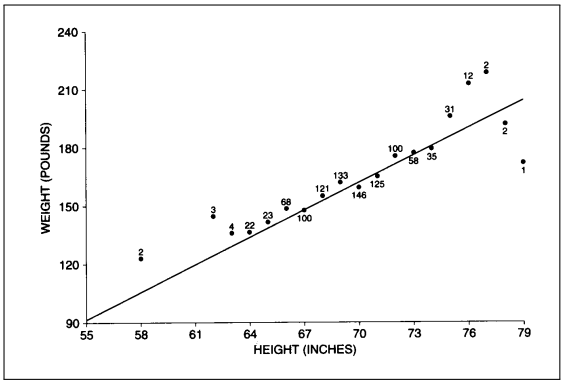
## 4.1 Die Regressionsgerade

Die Regressionsgerade beschreibt, wie eine Variable von einer zweiten abhängt. Man nehme zum Beispiel Grösse und Gewicht. Dazu verwenden wir die Daten von 988 Männern im Alter von $18-24$ (aus der schon erwähnten HANES Studie). Die Kennzahlen dieses Datensatzes sind:

| | $\mu$ | $\sigma$ | r |
|---|---|---|---|
| Grösse (in Zoll) | 70 | 3 | |
| Gewicht (in Pfund) | 162 | 30 | |
| | | | 0.47 |

Der vertikale Streifen in der Figur rechts zeigt die Männer, deren Grösse eine Standardabweichung $\sigma_x$ über dem Mittelwert liegt. Die Männer, deren Gewicht dann genau eine Standardabweichung $\sigma_y$ über dem Mittel liegt, würden dann auf der $\sigma-$Gerade liegen (das ist die gestrichelte Gerade). Aber die meisten Punkte in dem Streifen liegen deutlich unter der $\sigma-$Gerade. Wenn wir jeden Mittelwert beginnend von dem Streifen 58 Zoll (das sind $-4\sigma_x$) bis 79 Zoll (das sind $+3\sigma_x$), so erhalten wir das zweite Bild - den **Graphen der Mittelwerte**. Die **Regressionsgerade** in diesem Bild ist die Ausgleichsgerade für den Graphen der Mittelwerte. Wir können sie im Prinzip mit der **Methode der kleinsten Quadrate** berechnen (die auf Karl Friedrich Gauss zurückgeht). Diese sollte Sie an die Berechnung der Standardabweichung erinnern: Wir suchen nach einer Gerade, so dass die Summe der quadratischen Abweichungen minimal wird. Es stellt sich heraus, dass wir die Regressionsgerade ganz einfach mit Hilfe der fünf Kennzahlen von oben berechnen können:




Die Steigung der Regressionsgerade lässt sich durch die folgende Formel berechnen (siehe auch das unterste Bild rechts):

$$m = \frac{r \cdot \sigma_y}{\sigma_x} = \frac{0.47 \cdot 30}{3} = 4.7.$$

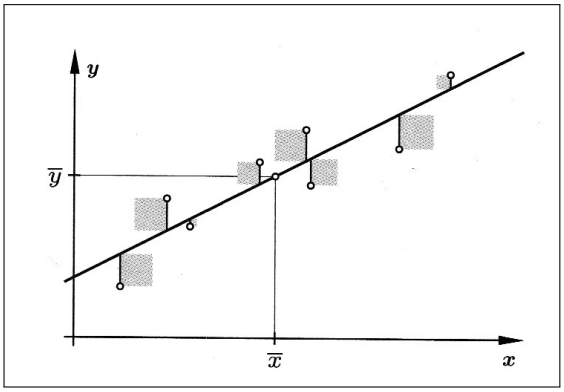
Weil die Regressionsgerade durch den Schwerpunkt $(\mu_x|\mu_y) = (70|162)$ verlaufen muss, können wir diesen Punkt in den folgenden Ansatz einsetzen:

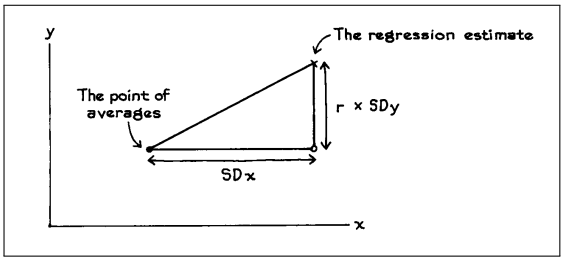
$$y = 4.7x + b$$

Das ergibt $162 = 4.7 \cdot 70 + b$ oder $b = -167$. Die Regressionsgerade für die obigen Daten lautet also

$$y = 4.7x - 167.$$

Wozu braucht man diese Geradengleichung? Wenn Sie eine möglichst gute Schätzung der durchschnittlichen Grösse der Männer im Alter von $18-24$ benötigen, die 73 Zoll gross sind, dann können Sie einfach $x = 73$ in die Gleichung der Regressionsgeraden einsetzen und erhalten die Antwort $4.7 \cdot 73 - 167 = 176.1$ Pfund.



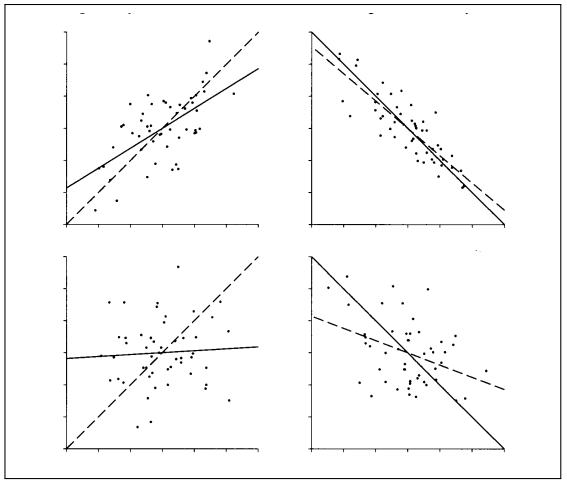



**Übung 4.1.** Die Kennzahlen in dem Beispiel auf Seite 21 sind in Zoll (1 Zoll = 2.54 cm) und Pfund (1 Pfund = 0.5 kg) angegeben.

a) Berechnen Sie die Regressionsgerade für die Einheiten cm und kg. Sie werden die Mittelwerte und Standardabweichungen in cm and kg umrechnen müssen, aber da diese Umrechnung die Punkte in dem Streudiagramm nur reskaliert, wird sich $r$ nicht ändern.

b) Geben Sie ausserdem eine Schätzung für das Gewicht eines Mannes im Alter von $18 - 24$ (in den USA, etwas im Jahr 1980) an, wenn Sie zudem wissen, das dieser 180 cm gross ist.

**Übung 4.2.** In der Figur rechts sehen Sie vier Streudiagramme, jedes mit einer durchgezogenen und einer gestrichelten Geraden. Entscheiden Sie bei jedem Diagramm, welches die $\sigma-$Gerade und welches die Regressionsgerade ist.



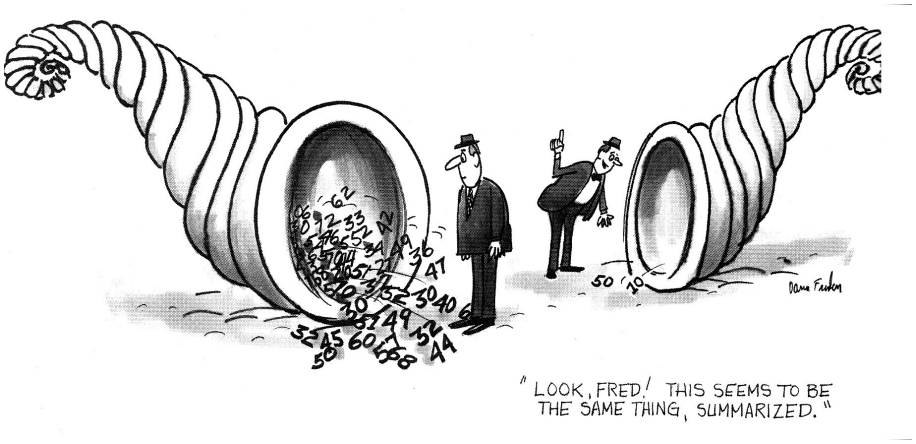
**Übung 4.3.** In einer Studie über die Stabilität von Intelligenzquotienten wurde eine grosse Gruppe von Individuen einmal im Alter von 18 und dann nochmals im Alter von 35 getestet. Die Kennzahlen der Resultate finden Sie in der Tabelle rechts. Wenn jemand im Alter von 18 Jahren einen IQ von 115 hatte, was ist dann die beste Schätzung für seinen IQ im Alter von 35?

| | $\mu$ | $\sigma$ | r |
|---|---|---|---|
| mit 18 | 100 | 15 | |
| mit 35 | 100 | 15 | |
| | | | 0.8 |

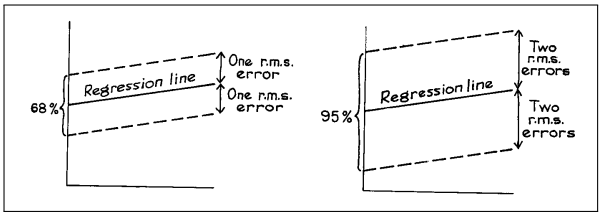
**Übung 4.4.** IQs werden so skaliert, dass der Mittelwert bei 100 liegt und die Standardabweichung bei 15. Dies gilt gleichermassen für Männer und Frauen. Die Korrelation des IQ zwischen Ehemännern und ihren Ehegattinen liegt bei $r = 0.50$. Bei einer grossen Studie fand man heraus, dass Männer, deren *IQ* bei 140 lag, Ehefrauen hatten, deren IQ bei 120 lag. Betrachten wir nun die Frauen mit einem IQ von 120; sollte der durchschnittliche IQ ihrer Ehemänner grösser oder kleiner sein als 120? Antworten Sie mit "ja" oder "nein" und erklären Sie kurz.

**Übung 4.5.** As part of their training, air force pilots make two practice landings with instructors, and are rated on performance. The instructors discuss the ratings with the pilots after each landing. Statistical analysis shows that pilots who make poor landings the first time tend to do better the second time. Conversely, pilots who make good landings the first time tend to do worse the second time. The conclusion: criticism helps the pilots while praise makes them do worse. As a result, instructors were ordered to criticize all landings, good or bad. Was this warranted by the facts? Answer yes or no, and explain briefly.

## 4.2 Der Regressionsfehler




Wir rekapitulieren: Eine eindimensionale Liste von Daten kann man durch zwei (nulldimensionale?) Zahlen zusammenfassen - den Mittelwert $\mu$ und die Standardabweichung $\sigma$, siehe der Cartoon oben. Ein zweidimensionales Streudiagramm können wir durch ein eindimensionales Objekt zusammenfassen - die Regressionsgerade.




Diese Gerade sagt die Position von Datenpunkten voraus. Natürlich gibt es Vorhersagefehler. Wir möchten ein Mass für diese Vorhersagefehler - und nehmen dafür einfach die Standardabweichung der Vorhersagefehler, den **RF**, d.h. den **Regressionsfehler**. Glücklicherweise können wir im Falle von ellipsenförmigen (d.h. normalverteilten) Datenwolken diesen RF ganz einfach berechnen:

$$\mathrm{RF} = \sqrt{1 - r^2} \cdot \sigma_y.$$

Hier bezeichnet $r$ den Korrelationskoeffizient der Daten und $\sigma_y$ ist die Standardabweichung der $y$-Werte. Genau wie im Fall der Standardabweichung liegen etwa 68% der Daten im Bereich von $\pm 1$ RF von der Regressionsgeraden. Etwa 95% liegen im Bereich von $\pm 2$ RF, siehe die Figur oben.

**Übung 4.6.** Zwei Datensätze stimmen in $\mu_x, \mu_y, \sigma_x$ und $\sigma_y$ überein, aber der erste hat die Korrelation $r = 0.7$ und die zweite hat die Korrelation $r = 0.3$. Welche von ihnen hat den grösseren RF und warum? Was ist besser, ein grosser oder ein kleiner RF?

**Übung 4.7.** Ein Datensatz hat einen RF von $0.866\sigma_y$. Wie lautet der zugehörige Korrelationskoeffizient $r$? Für welches $r$ ist der RF halb so gross, also $0.433\sigma_y$?

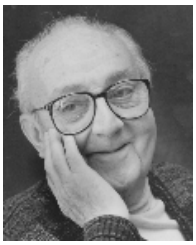
**Übung 4.8.** Tragen Sie den RF in das Streudiagramm von Aufgabe 3.9 ein. Markieren Sie sowohl den Bereich von 68% als auch den von 95%.

**Übung 4.9.** Es gibt Hinweise darauf, dass ein moderater Weinkonsum vor Herzinfarkten schützt, die Daten sind in der Tabelle rechts. Bestimmen Sie die fünf Kennzahlen dieser Daten und zeichnen Sie das zugehörige Streudiagramm. Markieren Sie ausserdem den RF in dem Streudiagramm.

| Country | Alcohol from wine | Heart disease deaths | Country | Alcohol from wine | Heart disease deaths |
|---|---|---|---|---|---|
| Australia | 2.5 | 211 | Netherlands | 1.8 | 167 |
| Austria | 3.9 | 167 | New Zealand | 1.9 | 266 |
| Belgium | 2.9 | 131 | Norway | 0.8 | 227 |
| Canada | 2.4 | 191 | Spain | 6.5 | 86 |
| Denmark | 2.9 | 220 | Sweden | 1.6 | 207 |
| Finland | 0.8 | 297 | Switzerland | 5.8 | 115 |
| France | 9.1 | 71 | United Kingdom | 1.3 | 285 |
| Iceland | 0.8 | 211 | United States | 1.2 | 199 |
| Ireland | 0.7 | 300 | West Germany | 2.7 | 172 |
| Italy | 7.9 | 107 | | | |

## 4.3 Das Bestimmtheitsmass

Der Statistiker *George Box* sagte einmal:



*Alle Modelle sind falsch - aber einige sind nützlicher als andere.*

Das Modell der linearen Regression ist offenbar bei dem Datensatz in dem ersten Bild rechts nicht sehr nützlich. Wir benötigen ein Mass dafür, wie viele Prozent der Änderung einer Variablen $y$ tatsächlich mit Hilfe der Regressionsgeraden $y = mx + b$ durch eine Änderung der Variablen $x$ erklärt werden kann.

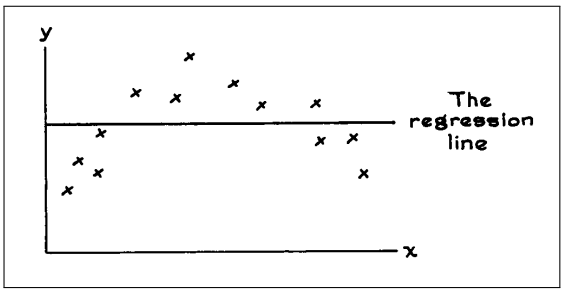
Als konkretes Beispiel betrachten wir die Beziehung zwischen Einkommen und Bildung für eine Stichprobe von 555 kalifornischen Männern im Alter zwischen 25 und 29 Jahren im Jahre 1993, siehe das Streudiagramm rechts. Die Kennzahlen für dieses Streudiagramm sind:

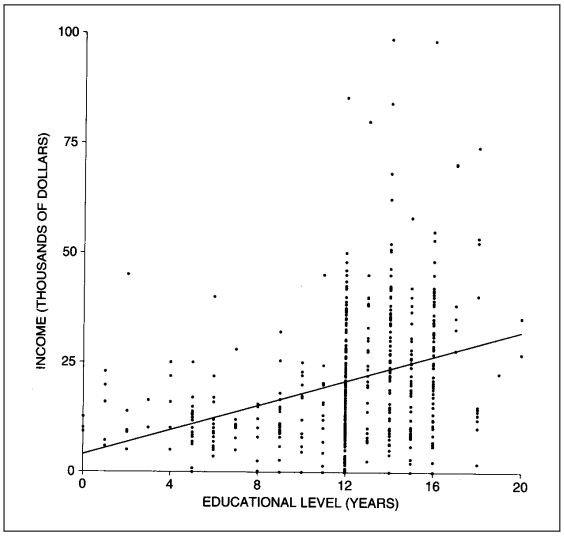
| | $\mu$ | $\sigma$ | r |
|---|---|---|---|
| Anzahl Schuljahre | 12.5 Jahre | 4 Jahre | |
| Einkommen | \$21′500 | \$16′000 | |
| | | | 0.35 |

Die Regressionsgerade ist $y = 1400x + 4000$.

Man muss aufpassen, dass man die Steigung nicht überinterpretiert. Sie gilt nur für einen Ist-Zustand, also für die Gruppe von Männern, die man untersucht hat. Man darf daraus nicht folgern:

*Wenn ich vier Jahre länger zur Schule gehe, dann werde ich* \$5′600 *mehr verdienen.*




Beschrieben wird eine Korrelation, kein Kausalzusammenhang. Vermutlich sind hier verborgene Variablen am Werk: Zum Beispiel könnten die, die länger zur Schule gehen, das hauptsächlich tun weil sie aus einem Elternhaus stammen, in dem man sich einen längeren Schulbesuch leisten kann. Und mit diesem familiären Hintergrund ist man vielleicht sowieso schon dazu prädestiniert, mehr zu verdienen als andere. Die 1′400 Dollar mehr pro zusätzlichem Schuljahr lassen sich nur zu einem gewissen Prozentsatz durch die verlängerte Ausbildung erklären – für den Rest sind andere Gründe verantwortlich. Aber zu welchem Prozentsatz? Wieder hilft der Korrelationskoeffizient weiter. Tatsächlich gibt das *Quadrat* von $r$ diesen Prozentsatz an – in unserem Beispiel ist $r = 0.35$, also ist $r^2 = 0.1225$, demnach lassen sich nur etwa 12% des Einkommenszuwachses durch eine längere Ausbildung erklären. Nicht sehr viel, aber immerhin.

Das **Bestimmtheitsmass** $r^2$ gibt an, wieviel Prozent der Änderung (genauer gesagt der Varianz = Quadrat der Standardabweichung, vgl. S. 28) der y-Werte durch die Änderung der $x$-Werte erklärt werden kann.

**Übung 4.10.** Die Korrelation zwischen dem Zuckergehalt und der Kalorienanzahl bei Frühstücksflocken liegt bei $r = 0.56$. Wie viele Prozent einer erhöhten Kalorienanzahl lassen sich durch erhöhte Zuckerzusätze erklären?

# 5 Anhang: Beweise

## 5.1 Warum funktioniert r als Mass für die Korrelation?

Mit Hilfe der $\sigma-$Geraden $y = \frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y$ kann man auch rein algebraisch verstehen, warum der Korrelationskoeffizient als Mass für die Korrelation funktioniert:

Wir betrachten einen schon standardisierten Datensatz $\{(x_i, y_i)\}$, wobei $i = 1, 2, \ldots, n$. Also gilt

$$\mu_x = 0 \qquad \mu_y = 0 \qquad \sigma_x = 1 \qquad \sigma_y = 1$$

Wir betrachten nur den Fall einer positiven Tendenz, also $r > 0$. Damit hat die $\sigma-$Gerade die Funktionsgleichung $y = x$.

**Spezialfall:**
Wenn die Daten perfekt korreliert sind, liegen alle auf der $\sigma-$Geraden, also gilt $x_i = y_i$. Damit erhält man mit unserer Formel für die Korrelation

$$r = \frac{1}{n}\sum_i x_i^2 = \frac{1}{n}\sum_i (x_i - \mu_x)^2 = \sigma_x^2 = 1$$

**Allgemeiner Fall:**
Wir schreiben die $y$-Variablen als $y_i = x_i + \epsilon_i$ mit einer positiven oder negativen Abweichung $\epsilon_i$. Dann gilt die folgende Formel für die Korrelation:

$$r = 1 - \frac{1}{2n}\sum_{k=1}^{n} \epsilon_i^2 \tag{6}$$

Diese Formel kann man so interpretieren: Je grösser die Abweichungen $\epsilon_i$ sind, desto mehr zieht man bei der Berechnung von $r$ von der 1 ab.

**Beweis:** Nach Definition gilt:

$$\begin{aligned} r &= \frac{1}{n}\sum_i x_i(x_i + \epsilon_i) \\ &= \frac{1}{n}\sum_i x_i^2 + \frac{1}{n}\sum_i x_i\epsilon_i \\ &= 1 + \frac{1}{n}\sum_i x_i\epsilon_i \end{aligned}$$

Die letzte Gleichheit folgt aus $\sigma_x = 1$. Wenn wir zeigen können, dass

$$\frac{1}{n}\sum_i x_i\epsilon_i = -\frac{1}{2n}\sum_i \epsilon_i^2 \tag{7}$$

folgt daraus (6). Die Identität (7) zeigt man so: Aus $\sigma_y = 1$ folgt

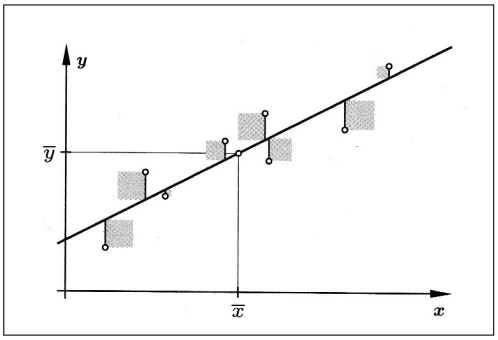
$$\begin{aligned} 1 &= \frac{1}{n}\sum_i (x_i + \epsilon_i)^2 \\ &= \frac{1}{n}\sum_i x_i^2 + \frac{1}{n}\sum_i 2x_i\epsilon_i + \frac{1}{n}\sum_i \epsilon_i^2 \\ &= 1 + \frac{1}{n}\sum_i 2x_i\epsilon_i + \frac{1}{n}\sum_i \epsilon_i^2 \end{aligned}$$

Subtrahiert man von dieser Gleichung 1 und bringt die erste Summe auf die andere Seite, so erhält man (7).

## 5.2 Das Prinzip der kleinsten Quadrate und die Regressionsgerade

In diesem Abschnitt leiten wir die Gleichung der Regressionsgeraden her.

Angenommen, wir suchen die Regressionsgerade für einen Datensatz $(x_i, y_i)$ mit $i = 1, 2, \ldots, n$. Zunächst müssen wir wissen, wie sie definiert ist. So wie der Mittelwert dadurch charakterisiert ist, dass er die Summe der quadrierten Abweichungen minimiert, so soll die Regressionsgerade so gewählt werden, dass die Summe der quadrierten vertikalen Abweichungen von der Geraden minimal wird. Die Idee ist also, die Summe der Fehlerquadrate so klein wie möglich zu halten - im Bild entspricht diese Summe der grauschattierten Fläche.




Unsere Aufgabe lautet demnach, eine Gerade $y = mx + b$ zu finden, so dass die Summe der quadrierten Abweichungen

$$\text{SQA} = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

minimal wird. Um die Rechnung einfach zu halten, machen wir die plausible (und vorläufige) Annahme, dass die Gerade durch den Schwerpunkt $(\bar{x}, \bar{y})$ verläuft.

- "Vorläufig" ist die Annahme, weil Sie in Übung 3 zeigen werden, dass die Regressionsgerade automatisch den Schwerpunkt enthält.
- Plausibel ist sie, weil für jede Gerade, die den Schwerpunkt enthält, ebenfalls analog zum Mittelwert die **vertikale Abweichungseigenschaft** gilt: Ist $y - \mu_y = m(x - \mu_x)$ eine solche Gerade, so betrachtet man die vertikalen Abweichungen $a_i$ von der Geraden:

  $$a_i = y_i - (m(x_i - \mu_x) + \mu_y) = (y_i - \mu_y) - m(x_i - \mu_x)$$

  Im Gegensatz zu den quadrierten Abweichungen können sie positive und negative Vorzeichen haben. Summiert man sie auf, so heben sich die positiven und die negativen Abweichungen gerade auf:

  $$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} (y_i - \mu_y) - m \sum_{i=1}^{n} (x_i - \mu_x) = 0$$

  Die letzte Gleichheit folgt aus der gewöhnlichen Schwerpunkteigenschaft ("Robin-Hood-Gleichung") für die Datensätze $\{x_i\}$ und $\{y_i\}$.

Man kann die vertikale Abweichungseigenschaft auch anders lesen: Die Abweichungen der $y$-Werte eines Datensatzes von einer beliebigen Gerade durch den Schwerpunkt haben den Mittelwert null.

Für die Herleitung der Regressionsgeraden betrachten wir einen schon standardisierten Datensatz $(x_i, y_i)$, wobei $i = 1, 2, \ldots, n$. Also gilt

$$\mu_x = 0 \qquad \mu_y = 0 \qquad \sigma_x = 1 \qquad \sigma_y = 1$$

Gesucht ist also eine Gerade durch den Schwerpunkt - das ist hier der Ursprung. Also eine Gerade der Form $y = mx$. Dabei ist $m$ so gewählt, dass die von $m$ abhängige quadratische Funktion

$$\begin{aligned} \mathrm{SQA}(m) &= \sum_{i=1}^{n}(y_i - mx_i)^2 \\ &= m^2\Big(\sum_{i=1}^{n} x_i^2\Big) + m\Big(-2\sum_{i=1}^{n} x_i y_i\Big) + \Big(\sum_{i=1}^{n} y_i^2\Big) \end{aligned}$$

ihren kleinsten Wert annimmt. Das ist beim Scheitelpunkt der zugehörigen Parabel der Fall. Nach der Scheitelpunktformel also für

$$m = -\frac{-2\sum x_i y_i}{2\sum x_i^2} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = r$$

Die vorletzte Gleichheit folgt wegen $\sigma_x = 1$, d.h. $\sum_{i=1}^{n} x_i^2 = n$, die letzte aus der Definition des Korrelationskoeffizienten $r$ als Kovarianz der standardisierten Daten.

Die Regressionsgerade eines standardisierten Datensatzes $(x_i^*, y_i^*)$ lautet also einfach

$$y^* = rx^*$$

Um die Standardisierung rückgängig zu machen, setzen wir

$$x^* = \frac{x - \mu_x}{\sigma_x} \quad \text{und} \quad y^* = \frac{y - \mu_y}{\sigma_y}$$

in die Regressionsgleichung ein und erhalten (nach einer kürzeren Rechnung, siehe Übung 1) die Gleichung der Regressionsgeraden für einen beliebigen Datensatz:

$$y = \frac{r \cdot \sigma_y}{\sigma_x}(x - \mu_x) + \mu_y$$

**Einige Übungen dazu**

1. Führen Sie die eben erwähnte kürzere Rechnung durch.

2. Wie lautet der Achsenabschnitt der allgemeinen Regressionsgeraden?

3. Wir müssen noch zeigen, dass der Schwerpunkt automatisch auf der Regressionsgeraden liegt. Wir gehen wieder von einem standardisierten Datensatz aus, nehmen jetzt aber an, dass die Regressionsgerade durch den Punkt $(0, c)$ geht. Zu zeigen ist, dass $c = 0$ gilt.

   (a) Zeigen Sie zunächst, dass für die von $m$ und $c$ abhängige Summe der quadrierten Abweichungen gilt:

   $$\begin{aligned} \mathrm{SQA}(m, c) &= \sum_{i=1}^{n}(y_i - (mx_i + c))^2 \\ &= n \cdot m^2 - 2nr \cdot m + n(c^2 + 1) \end{aligned}$$

   (b) Dies kann man wieder als quadratische Funktion in $m$ auffassen. Für welches (von $c$ abhängige) $m$ wird der Scheitelpunkt angenommen? Für welches $c$ ist dann $\mathrm{SQA}(m, c)$ minimal?

4. Wir haben gesehen, dass alle Geraden durch den Schwerpunkt eines Datensatzes die vertikale Abweichungseigenschaft haben. Zeigen Sie, dass Geraden, die *nicht* durch den Schwerpunkt verlaufen, diese Eigenschaft *nicht* haben.

## 5.3 Warum (und wie genau) funktioniert das Bestimmtheitsmass?

Wir vergleichen die realen $y$-Werte mit den durch das Regressionsmodell vorhergesagten Werten $\hat{y}$. Wenn also $x_k$ ein gemessenes Alter ist, dann ist

$$\hat{y}_k = mx_k + b$$

die vorhergesagt Körpergrösse. Offenbar streuen die gemessenen Daten $y_k$ mehr als die vorhergesagten Daten $\hat{y}_k$, für die Varianzen (das waren die Quadrate der Standardabweichungen) gilt also

$$\text{var}(\hat{y}) < \text{var}(y)$$

Wenn man zum Beispiel das Verhältnis

$$\frac{\text{var}(\hat{y})}{\text{var}(y)} = 0.75 \quad \text{bzw.} \quad \text{var}(\hat{y}) = 0.75 \cdot \text{var}(y)$$

erhält, dann verwendet man die Sprechweise, dass 75% der gesamten Varianz durch das lineare Modell "erklärt" werden.

Die Formel

$$\frac{\text{var}(\hat{y})}{\text{var}(y)} = r^2 \quad \text{bzw.} \quad \text{var}(\hat{y}) = r^2 \cdot \text{var}(y)$$

ergibt sich leicht aus den Rechenregel für Varianzen, als Übung sollten Sie folgendes nachrechnen: Für jede Konstante $c$ gilt

$$\text{var}(y + c) = \text{var}(y) \quad \text{und} \quad \text{var}(c \cdot y) = c^2 \cdot \text{var}(y)$$

Mit

$$\hat{y} = \frac{r \cdot \sigma_y}{\sigma_x} \cdot x + b$$

erhält man

$$\begin{aligned}
\text{var}(\hat{y}) &= \text{var}\left(\frac{r \cdot \sigma_y}{\sigma_x} \cdot x + b\right) \\
&= \text{var}\left(\frac{r \cdot \sigma_y}{\sigma_x} \cdot x\right) \\
&= \frac{r^2 \cdot \sigma_y^2}{\sigma_x^2} \cdot \text{var}(x) \\
&= \frac{r^2 \cdot \text{var}(y)}{\text{var}(x)} \cdot \text{var}(x) \\
&= r^2 \cdot \text{var}(y)
\end{aligned}$$

wie behauptet.

## 5.4 Warum funktioniert der Regressionsfehler RF?

Da das Bestimmtheitsmass $r^2$ den Prozentsatz angibt, zu dem das lineare Modell für die Varianz der $y$-Werte verantwortlich ist, ist es nicht weiter überraschend, dass $1-r^2$ in dem Ausdruck vorkommt, der die restliche Streuung der vertikalen Abweichungen von der Regressionsgeraden beschreibt. Wir betrachten die vertikalen Abweichungen

$$a_k = y_k - \hat{y}_k$$

der Punkte von der Regressionsgeraden. Es stellt sich heraus, dass der RF die Standardabweichung dieser Abweichungen $a_k$ ist:

$$\mathrm{RF} = \sigma_{y-\hat{y}} = \sqrt{\mathrm{var}(y-\hat{y})}$$

Zum Beweis genügt es zu zeigen, dass

$$\mathrm{var}(y-\hat{y}) = (1-r^2)\mathrm{var}(y)$$

gilt. Um die Rechnung zu vereinfachen, transformieren wir die Daten (durch Subtrahieren der Mittelwerte), so dass $\bar{x} = \bar{y} = 0$ gilt - dadurch ändern sich die Werte von $a_k$ nicht. Die Regressionsgerade hat dann die Gleichung

$$\hat{y} = \frac{r \cdot \sigma_y}{\sigma_x} \cdot x$$

Wie wir auf Seite 26 gesehen haben, haben die Abweichungen $a_k$ den Mittelwert $\bar{a} = 0$, damit gilt

$$\begin{aligned}
\mathrm{var}(y-\hat{y}) &= \mathrm{var}(a) \\
&= \frac{1}{n}\sum_{k=1}^{n} a_k^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \frac{r\cdot\sigma_y}{\sigma_x}\cdot x_k\right)^2
\end{aligned}$$

Für den letzten Term erhält man durch Ausmultiplizieren

$$\begin{aligned}
& \frac{1}{n}\sum_{k=1}^{n} y_k^2 - 2\frac{r\cdot\sigma_y}{\sigma_x}\frac{1}{n}\sum_{k=1}^{n} x_k y_k + \frac{r^2\cdot\sigma_y^2}{\sigma_x^2}\frac{1}{n}\sum_{k=1}^{n} x_k^2 \\
= \;& \mathrm{var}(y) - 2\frac{r\cdot\sigma_y}{\sigma_x}\frac{1}{n}\sum_{k=1}^{n} x_k y_k + \frac{r^2\mathrm{var}(y)}{\mathrm{var}(x)}\mathrm{var}(x) \\
= \;& \mathrm{var}(y) - 2\frac{r\cdot\sigma_y}{\sigma_x}\cdot\sigma_x\cdot\sigma_y\frac{1}{n}\sum_{k=1}^{n}\frac{x_k}{\sigma_x}\frac{y_k}{\sigma_y} + r^2\mathrm{var}(y) \\
= \;& \mathrm{var}(y) - 2r\sigma_y^2\cdot r + r^2\mathrm{var}(y) \\
= \;& \mathrm{var}(y) - r^2\mathrm{var}(y) \\
= \;& (1-r^2)\mathrm{var}(y)
\end{aligned}$$