unsere Meßergebnisse verfälschen; jeder einzelne dieser „Dämonen“ verursacht einen Fehler derselben Größe ε in positiver oder negativer Richtung, wobei die Wahrscheinlichkeit für $+\varepsilon$ *bzw.* $-\varepsilon$ für jeden der „Dämonen“ jeweils 50% sein soll und die einzelnen „Dämonen“ unabhängig voneinander handeln sollen.

Im Falle eines einzigen „Dämonen“ wäre der Fehler also mit gleicher Wahrscheinlichkeit $+\varepsilon$ oder $-\varepsilon$, bei zwei „Dämonen“ wäre er in jeweils 25% aller Fälle $+2\varepsilon$ oder -2ε, während sich in 50% der Fälle die beiden Fehler aufheben würden.

Allgemein gibt es bei n „Dämonen“ 2^n gleichwahrscheinliche Möglichkeiten für deren Verhalten; die folgende Tabelle zeigt für $n \leq 5$ jeweils die Anzahl der Fälle, die zu dem in der Kopfzeile angegebenen Gesamtfehler führen:

	-5ε	-4ε	-3ε	-2ε	$-\varepsilon$	0	$+\varepsilon$	$+2\varepsilon$	$+3\varepsilon$	$+4\varepsilon$	$+5\varepsilon$
$n = 0$						1					
$n = 1$					1		1				
$n = 2$				1		2		1			
$n = 3$			1		3		3		1		
$n = 4$		1		4		6		4		1	
$n = 5$	1		5		10		10		5		1

Diese dreiecksförmige Anordnung von Zahlen bezeichnet man als PASCAL*sches Dreieck.* Offenbar kann man es dadurch rekursiv zeilenweise berechnen, daß man an jede Stelle die Summe der beiden links und rechts darüberstehenden Zahlen schreibt: Die n-te Störung bringt den Fehler genau dann auf $i \cdot \varepsilon$, wenn sie entweder gleich $+\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i-1)\cdot\varepsilon$ produziert haben, oder aber wenn sie gleich $-\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i+1)\cdot\varepsilon$ produziert haben. Entsprechend ist auch klar, daß die Summe aller Zahlen in der n-ten Zeile gleich 2^n ist, denn in der nullten Zeile haben wir Summe eins, und da die jeweils neu hinzukommende Störung genau zwei Möglichkeiten hat, verdoppelt sich die Summe von Zeile zu Zeile. Die Wahrscheinlichkeit dafür, daß sich n Störungen zu $i \cdot \epsilon$ aufsummieren, ist also gerade gleich der Zahl, die in der n-ten Spalte unter $i \cdot \varepsilon$ steht (beziehungsweise Null, wenn dort keine Zahl steht), dividiert durch 2^n.

Bekanntlich kann man die Zahlen in diesem Dreieck auch explizit berechnen: An der n-ten Zeile stehen die $n+1$ Zahlen

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \quad \text{für} \quad i = 0, \ldots, n\,.$$

Wer diese Formel nicht kennt, kann sie leicht durch vollständige Induktion beweisen: Für $n = 1$ sowie allgemein für $i = 0$ oder $i = n$ ist alles klar; für $n > 1$ und $0 < i < n$ stehen über $\binom{n}{i}$ die beiden Zahlen $\binom{n-1}{i-1}$ und $\binom{n-1}{i}$, für die in der Tat gilt

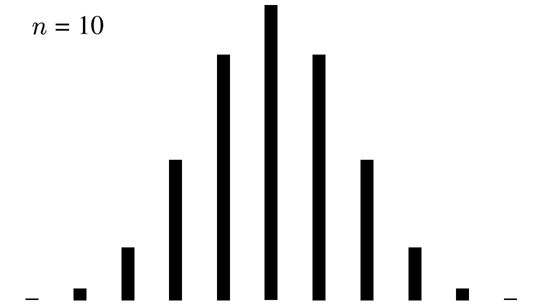
$$\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i)!} + \frac{(n-1)!}{i!(n-i-1)!} = \frac{(n-1)!}{i!(n-i)!}\cdot\Bigl(i + (n-i)\Bigr)$$
$$= \frac{n!}{i!(n-i)!} = \binom{n}{i}\,.$$

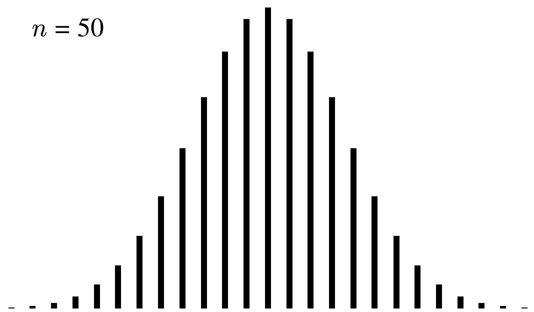
Mit dieser Formel lassen sich die Fallanzahlen für einen bestimmten Fehler im Prinzip berechnen; allerdings ist die Berechnung für große n schnell sehr aufwendig und die Binomialkoeffizienten werden auch schnell sehr groß. Um trotzdem einen Eindruck davon zu bekommen, was für größere n passiert, sind auf den beiden folgenden Seiten die Binomialkoeffizienten für $n = 5, 10, 50, 100, 500$ und 1000 graphisch dargestellt. (Die Tatsache, daß ab $n = 50$ deutlich weniger als $n+1$ Balken zu sehen sind, erklärt sich daraus, daß die restlichen Binomialkoeffizienten zu klein sind, um noch darstellbar zu sein: Die Diagramme sind so skaliert, daß der größte (mittlere) Balken jeweils eine feste Höhe hat.)

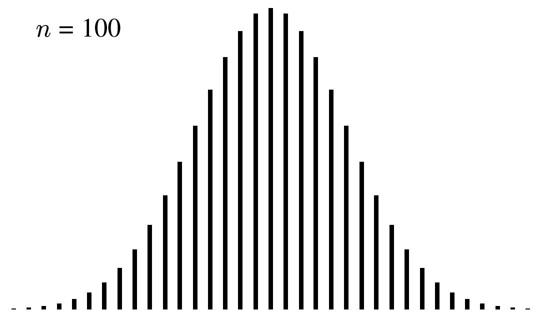
Betrachten wir als nächstes die Größe des Gesamtfehlers. Falls n gerade ist, treten nur Vielfache von 2ε auf und alle diese Vielfachen zwischen $-n\varepsilon$ und $n\varepsilon$ kommen tatsächlich vor; ensprechend sind für ungerades n nur ungeradzahlige Vielfache von ε möglich, und auch hier werden wieder alle solchen Werte zwischen $-n\varepsilon$ und $n\varepsilon$ angenommen. Wir können dies dadurch zusammenfassen, daß in beiden Fällen genau die Werte $(n-2k)\varepsilon$ mit $k = 0, \ldots, n$ angenommen werden, und das PASCALsche Dreieck zeigt, daß der Fehler $(n-2k)\varepsilon$ in

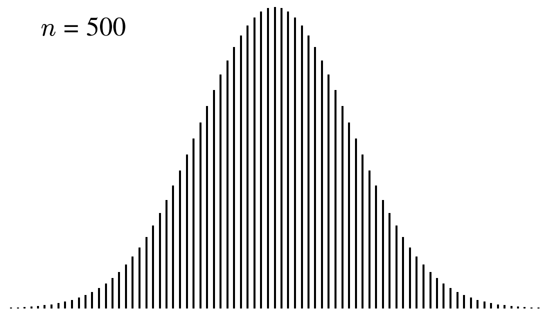
$$\binom{n}{n-k} = \binom{n}{k}$$

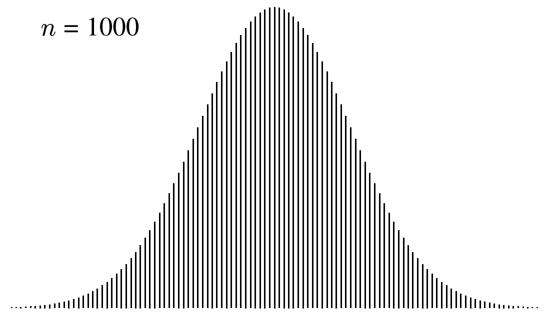

$n = 5$


$n = 10$


$n = 50$


$n = 100$


$n = 500$


$n = 1000$

Fällen auftritt. Da n „Dämonen“ insgesamt 2^n Möglichkeiten zur Fehlererzeugung haben, ist die Wahrscheinlichkeit für den Gesamtfehler $(n-2k)\varepsilon$ also

$$\binom{n}{k} \cdot 2^{-n} .$$

Diese Wahrscheinlichkeit sollte für einen festen Fehlerbetrag im wesentlichen unabhängig von n sein: Da wir nicht wirklich an Dämonen glauben, können wir deren Anzahl schließlich nicht in ein realistisches Fehlermodell einfließen lassen.

Auch die Balkendiagramme zeigen, daß sich die Verteilung der Fehlerwahrscheinlichkeiten für große n einer festen Kurve annähern sollte, der in Abbildung 64 und auf jedem Zehnmarkschein zu findenden *Glockenkurve* oder GAUSS-Kurve.

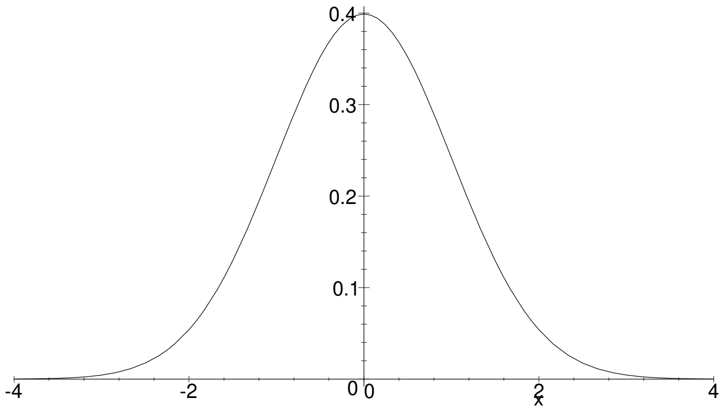


Abb. 64: Die „Glockenkurve“ $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Wenn sich Fehler oder auch beliebige Daten so veteilen, wie es dieser Kurve entspricht, redet man von *normalverteilten* Daten. Damit haben wir also zumindest graphisch gesehen, daß Meßfehler nach dem LAPLACEschen Fehlermodell normalverteilt sind. Das sagt noch nicht unbedingt etwas über die Verteilung realer Meßfehler, da das

LAPLACEsche Fehlermodell von unrealistisch einfachen Annahmen ausgeht; nach einem der fundamentale Gesetze der Statistik, dem *zentralen Grenzwertsatz,* führen aber auch realistischere Annahmen zu genau derselben Verteilung: Sind $u_1, \dots, u_n$ beliebige Quellen von Zufallsfehlern, über deren Verteilung wir (fast) nichts voraussetzen müssen, so ist ihre Summe für hinreichend großes n annähernd normalverteilt. Das eingeklammerte Wort „fast“ ist dabei für praktische Zwecke bedeutungslos, und als „groß“ kann man sich ein n ab etwa dreißig oder vierzig vorstellen.

Im nächsten Paragraphen werden wir uns überlegen, wie man zu einer Gleichung für die Glockenkurve kommt.

b) Statistische Kenngrößen

Die übliche Strategie zum Umgang mit Zufallsfehlern ist wohlbekannt: Man begnügt sich nicht mit einer einzigen Messung, sondern mißt dieselbe Größe mehrmals, so daß man eine ganze Meßreihe

$$x_1, x_2, \dots, x_N$$

erhält. Dann bildet man das *arithmetische Mittel*

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

der Meßreihe in der Hoffnung, daß sich hierbei die Fehler „ausmitteln“, so daß $\overline{x}$ dem theoretisch korrekten Wert $\hat{x}$ nahekommt.

Die Wahl des arithmetischen Mittels läßt sich auch geometrisch begründen: Eine Meßreihe $x_1, \dots, x_N$ für eine Meßgröße mit exaktem Wert $\hat{x}$ definiert einen Vektor im $\mathbb{R}^n$. Falls es keine Meßfehler gäbe, hätte dieser lauter identische Komponenten $\hat{x}$. Tatsächlich ist dies natürlich nicht der Fall; wir können aber nach einem Vektor mit identischen Komponenten suchen, der möglichst nahe am Vektor der Meßwerte liegt. Für einen Vektor, dessen sämtliche Komponenten gleich x sind, ist der EUKLIDische Abstands zum Vektor der Meßwerte gleich

$$d(x) = \sqrt{\sum_{i=1}^{N} (x - x_i)^2} = \sqrt{Nx^2 - 2x \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i^2} .$$

Die quadratische Funktion $d(x)^2$ hat ein eindeutig bestimmtes Minimum bei der Nullstelle ihrer Ableitung

$$2Nx - 2\sum_{i=1}^{N} x_i \,,$$

also beim arithmetischen Mittel $\overline{x}$, und dieses ist auch das einzige Minimum von $d(x)$. Wir nehmen daher das arithmetische Mittel $\overline{x}$ als besten verfügbaren Schätzwert für den unbekannten korrekten Wert $\hat{x}$.

Als Maß für die Schwankungen innerhalb der Meßreihe und damit für die Meßfehler könnte man versucht sein, den Abstand $d(\hat{x})$ zu nehmen; er hat aber den Nachteil, daß er mit steigendem N immer größer wird, d.h. die Schwankungen würden umso größer, je mehr man mißt. Das ist natürlich absurd; daher dividieren wir das Abstandsquadrat noch durch N und definieren die *mittlere quadratische Abweichung* oder *Varianz* der Meßreihe als

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x_i)^2 \,.$$

Die (nichtnegative) Quadratwurzel σ hieraus heißt *Standardabweichung* der Meßreihe.

Das Ergebnis einer Messung wird meist angegeben in der Form

$$x = \overline{x} \pm \sigma \,,$$

man betrachtet also die Standardabweichung der Meßreihe als Maß für den Meßfehler. Da deren Definition allerdings vom (im allgemeinen unbekannten) korrekten Wert $\hat{x}$ abhängt, können wir sie nicht berechnen, sondern müssen im folgenden sehen, wie wir sie zumindest schätzen können.

Als einfachste Möglichkeit bietet sich an, σ^2 durch

$$\frac{1}{N}\sum_{i=1}^{N}(\overline{x} - x_i)^2$$

zu schätzen, aber das führt sicherlich zu einem zu kleinen Ergebnis: Schließlich ist $d(\overline{x})$ das eindeutig bestimmte Minimum der Abstandsfunktion d, so daß der korrekte Wert $d(\hat{x})$ für $\hat{x} \neq \overline{x}$ notwendigerweise größer sein muß.

In Abschnitt *d)* werden wir aus dem Fehlerfortpflanzungsgesetz einen besseren Schätzwert für σ herleiten.

Warum betrachten wir eigentlich quadratische Abweichungen und nicht die einfacheren linearen Abweichungen? Nun, der Mittelwert aller Abweichungen ist

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}N\overline{x} = \overline{x} - \overline{x} = 0 \,,$$

also ist dies keine geeignete Maßzahl. Möglich wäre die mittlere *betragsmäßige* Abweichung

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}| \,,$$

allerdings wird die im allgemeinen nicht für das arithmetische Mittel $\bar{x}$ minimal, sondern, wie man sich leicht überlegen kann, für jede Zahl $\overline{x}$ mit der Eigenschaft, daß gleichviele Meßwerte größer und kleiner als $\overline{x}$ sind; einen solchen Wert $\overline{x}$ bezeichnet man als *Median* der Meßreihe. Für die Beschreibung wirtschafts- und sozialwissenschaftlicher Daten ist dieser Median meist eine aussagekräftigere Kennzahl als das arithmetische Mittel; in den Naturwissenschaften und der Technik spielt er allerdings keine große Rolle. Im nächsten Paragraphen werden wir sehen, daß auch das LAPLACEsche Fehlermodell in natürlicher Weise auf quadratische Abweichungen führt.

c) Das Fehlerfortpflanzungsgesetz

Gegeben seien zwei Größen

$$x = \hat{x} \pm \sigma_x \quad \text{und} \quad y = \hat{y} \pm \sigma_y$$

(die Verallgemeinerung auf mehr als zwei Größen erfordert, wie man sich bei der folgenden Rechnung leicht klarmacht, nur etwas mehr Schreibaufwand; sie ist nicht prinzipiell schwieriger), und eine Größe

$$w = f(x, y) \,,$$

die von diesen beiden abhängt. Um vernünftige Aussagen machen zu können, setzen wir dabei f als stetig differenzierbar voraus.

Für x seien N Meßwerte $x_1, \ldots, x_N$ gegeben, und für y entsprechend M Werte $y_1, \ldots, y_M$. Wenn wir echte Zufallsfehler haben, können wir davon ausgehen, daß die Fehler der x-Werte und die der y-Werte voneinander unabhängig sind, und das wollen wir im folgenden auch annehmen.

Für w haben wir dann NM Werte $w_{ij} = f(x_i, y_j)$, deren Mittelwert die beste Schätzung für den „wahren“ Wert $\widehat{w} = f(\hat{x}, \hat{y})$ ist. Dieser Mittelwert ist für komplizierte Funktionen f und/oder große Werte von n und m umständlich auszurechnen; günstiger wäre es, einfach den Mittelwert $\bar{x}$ der x_i und den Mittelwert $\bar{y}$ der y_j zu berechnen, um dann $f(\bar{x}, \bar{y})$ als Schätzung für $\widehat{w}$ zu benutzen. Zur Abschätzung des dadurch bedingten Fehler setzen wir

$$x_i = \bar{x} + h_i \quad \text{und} \quad y_j = \bar{y}_j + k_j ;$$

dann ist wegen der Differenzierbarkeit von f

$$\begin{aligned} w_{ij} &= f(x_i, y_j) = f(\bar{x} + h_i, \bar{y} + k_j) \\ &= f(\bar{x}, \bar{y}) + f_x(\bar{x}, \bar{y})h_i + f_y(\bar{x}, \bar{y})k_j + o\left(\sqrt{h_i^2 + k_j^2}\right) , \end{aligned}$$

wobei

$$f_x = \frac{\partial f}{\partial x} \quad \text{und} \quad f_y = \frac{\partial f}{\partial y}$$

die partiellen Ableitungen von f bezeichnen. Da die h_i und die k_j als Abweichungen vom Mittelwert die Summe null haben, ist also der Mittelwert der w_{ij} bis auf einen Fehler der Größenordnung $o\left(\sqrt{h^2 + k^2}\right)$ gleich $f(\bar{x}, \bar{y})$, wobei h, k die Batragsmaxima der h_i, k_j sind.

Als nächstes müssen wir den Fehler von $\widehat{w}$ berechnen, also den Erwartungswert der $(w_{ij} - \widehat{w})^2$. Dazu schreiben wir zunächst

$$x_i = \hat{x} + u_i \quad \text{und} \quad y_j = \hat{y}_j + v_j ,$$

betrachten also anstelle der Abweichungen vom Mittelwert die echten Meßfehler, und erhalten genau wie eben

$$w_{ij} - \widehat{w} = f(x_i, y_j) - f(\hat{x}, \hat{y}) \approx u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})$$

mit Quadrat

$$\begin{aligned} (w_{ij} - \widehat{w})^2 &\approx \big(u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})\big)^2 \\ &= u_i^2 \cdot f_x(\hat{x}, \hat{y})^2 + v_j^2 \cdot f_y(\hat{x}, \hat{y})^2 + 2u_i \cdot v_j \cdot f_x(\hat{x}, \hat{y}) \cdot f_y(\hat{x}, \hat{y}) . \end{aligned}$$

Hier sind die Werte u_i^2, v_j^2 und $u_i v_j$ jeweils Zufallsgrößen, über deren Werte wir nichts sagen können. Wir haben aber gewisse Erwartungen darüber, wie sie sich *im Mittel* verhalten: u_i^2 sollte, da σ_x^2 die mittlere quadratische Abweichung von $\hat{x}$ ist, im Mittel gleich σ_x^2 sein und v_j^2 entsprechend σ_y^2. Genauso sollten u_i und v_j im Mittel gleich null sein, und wenn wir annehmen, daß die Fehler u_i und v_j voneinander unabhängig sind, sollte auch ihr Produkt im Mittel verschwinden. Diese sogenannten *Erwartungswerte* sind offensichtlich die bestmöglichen Schätzwerte für die jeweiligen Größen; als beste Schätzung für σ_w^2 erhalten wir damit das GAUSS*sche Fehlerfortpflanzungsgesetz*

$$\sigma_w^2 = f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2$$

oder

$$\sigma_w = \sqrt{f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2} .$$

Genauso gilt dieses Gesetz auch für Funktionen von mehr als zwei Größen; für $w = f(x_1, \dots, x_n)$ ist

$$\sigma_w = \sqrt{f_{x_1}(\hat{x}_1, \dots \hat{x}_n)^2 \cdot \sigma_{x_1}^2 + \cdots + f_{x_n}(\hat{x}_1, \dots \hat{x}_n)^2 \cdot \sigma_{x_n}^2} .$$

d) Die Standardabweichung des Mittelwerts und die Schätzung der Varianz

Als einfache Anwendung des Fehlerfortpflanzungsgesetzes betrachten wir die Funktion

$$\bar{x} = f(x_1, \dots, x_N) = \frac{x_1 + \cdots + x_N}{N} ,$$

also den Mittelwert der x_i. Jede Messung x_i sei mit demselben erwarteten Fehler σ behaftet; da alle partiellen Ableitungen von f gleich $1/N$ sind, folgt für den Fehler des Mittelwerts

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} .$$

Dies bestätigt die implizit stets angewandte Regel, daß man durch mehrfaches Messen ein zuverlässigeres Ergebnis erhält; durch 25 Messungen beispielsweise läßt sich der Fehler auf ein Fünftel reduzieren, und für $N \to \infty$ geht er gegen Null *(Gesetz der großen Zahl).*

Damit wissen wir, wie man aus den Meßwerten auf den Fehler des Mittelwerts schließen kann – sofern man die Fehler der Meßwerte kennt. Wie lassen sich diese schätzen?

Zunächst ist

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}-x_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\big((\hat{x}-\bar{x})+(\bar{x}-x_i)\big)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}(\hat{x}-\bar{x})^2 + \frac{1}{N}\sum_{i=1}^{N}(\bar{x}-x_i)^2 + \frac{2}{N}\sum_{i=1}^{N}(\hat{x}-\bar{x})\cdot(\bar{x}-x_i)\,.$$

Die letzte dieser drei Summen ist

$$2\cdot\frac{(\hat{x}-\bar{x})}{N}\sum_{i=1}^{N}(\bar{x}-x_i) = 0\,,$$

da $\bar{x}$ der Mittelwert der x_i ist. Die zweite Summe ist der Mittelwert der $(\bar{x}-x_i)^2$, also die Varianz der Meßreihe, und von der ersten schließlich wissen wir, daß $(\hat{x}-\bar{x})^2$, das Quadrat des Fehlers des Mittelwerts, den Erwartungswert σ^2/N hat. Die gesamte erste Summe ist somit

$$\frac{1}{N}\cdot N\cdot\frac{\sigma^2}{N} = \frac{\sigma^2}{N}\,,$$

und obige Formel wird zu

$$\sigma^2 = \frac{\sigma^2}{N} + \frac{1}{N}\sum_{i=1}^{N}(\bar{x}-x_i)^2\,.$$

Bringt man hier noch den Term σ^2/N auf die linke Seite, so folgt

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

oder

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}-x_i)^2\,,$$

also

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\bar{x}-x_i)^2}{N-1}}\,.$$

Somit läßt sich auch σ aus den Meßdaten berechnen, der Meßfehler kann also ohne Kenntnis des „wahren“ Werts anhand der gemessenen Werte geschätzt werden.

e) Die Methode der kleinsten Quadrate

Oftmals ist zu gegebenen Beobachtungsdaten grundsätzlich bekannt, welcher *Art* von Gesetz sie genügen sollten, und das Problem besteht „nur“ noch darin, die in diesem Gesetz vorkommenden *Parameter* zu bestimmen. Im einfachsten Fall könnte man etwa an einen Widerstand denken, der dadurch gemessen wird, daß man verschiedene Spannungen U_i anlegt und die zugehörigen Stromstärken I_i mißt. Nach dem Ohmschen Gesetz ist dann $U_i = R\cdot I_i$, aber aufgrund der unvermeidlichen Meßfehler werden die verschiedenen Quotienten U_i/I_i natürlich nicht alle gleich sein. Die Lösung dieses Problems ist klar: Man nimmt den Mittelwert der Quotienten. Schwieriger wird es, wenn mehrere Parameter ins Spiel kommen.

Betrachten wir als Beispiel etwa die Entwicklung der Weltbevölkerung. Da (vor allem auf dem Stand von 1960) vieles für ein stärker als exponentielles Bevölkerungswachstum sprach, enwickelten damals HEINZ VON FOERSTER, PATRICIA M. MORA und LAWRENCE W. AMIOT, allesamt Elektrotechniker an der University of Illinois in Urbana, ein überexponentielles Wachstumsmodell. Die Wachstumsdifferentialgleichung

$$\dot{y}(t) = \beta y(t)^r$$

für die Weltbevölkerung $y(t)$ zum Zeitpunkt t kann durch Trennung der Veränderlichen leicht gelöst werden; die Lösung ist

$$y(t) = \left(\frac{\beta}{t_0-t}\right)^s \quad \text{mit} \quad s = \frac{1}{r-1}\,.$$

β, t_0 und r sind Parameter, die aufgrund der bekannten Zahlen zur Bevölkerungsentwicklung geschätzt werden müssen. Besonders interessant ist dabei natürlich der Parameter t_0, denn zum Zeitpunkt $t = t_0$

wird die Weltbevölkerung unendlich, was man wohl mit dem Weltuntergang gleichsetzen kann.

Für zahlreiche Zeitpunkte t_i sind die Bevölkerungszahlen y_i oder zumindest Schätzungen dafür bekannt, aber natürlich gibt es auch hier keine Parameterwerte, für die *alle* y_i gleich $f(t_i)$ sind; Ziel kann es nur sein, die Parameter so zu wählen, daß die Unterschiede zwischen y_i und $y(t_i)$ nicht zu groß werden.

FOERSTER, MORA und AMIOT taten genau das und kamen auf das Ergebnis

$$\begin{aligned} \beta &= (179 \pm 14) \cdot 10^9 \\ s &= 0{,}990 \pm 0{,}009 \\ t_0 &= 2026{,}87 \pm 5{,}5 \text{ Jahre} \\ r &= 2{,}01 \pm 0{,}009\,; \end{aligned}$$

der Weltuntergang ist also etwa 2026 fällig. Der beste Schätzwert für das Datum ist, wenn wir 0,87 Jahre in Tage, Stunden und Minuten umrechnen, Freitag, der 13. November gegen 13^{13} uhr – sofern man das bei einer Standardabweichung von fünfeinhalb Jahren überhaupt so genau wissen will. (Für Einzelheiten sei auf die Arbeit *Doomsday: Friday, 13 November, A.D. 2026*, Science **138**, November 1960, Seite 1291–1295, verwiesen.)

Da Weltuntergangsszenarien nicht zum Stoff dieser Vorlesung gehören, wollen wir auf die genaue Behandlung dieses Beispiels verzichten; wir wollen aber zumindest grundsätzlich verstehen lernen, wie man solche Parameterschätzungen durchführen kann, und wir wollen diese Techniken auch auf etwas alltäglichere Probleme anwenden.

Der allgemeine Ansatz ist folgender: Sei $u = f(a, b, \ldots; t)$ der von den Parametern $a, b, \ldots$ abhängige Zusammenhang zwischen den Größen t und u, und seien $(t_1, u_1), \ldots, (t_N, u_N)$ gegebene (Meß-)Werte, an die die Parameter angepaßt werden sollen.

Nehmen wir an, es gäbe Parameter $a, b, \ldots$, für die $u = f(a, b, \ldots; t)$ den exakten Zusammenhang zwischen den Größen u und t wiedergibt, daß aber die Werte u_i aufgrund von Beobachtungsfehlern von den

$f(a, b, \ldots; t_i)$ abweichen – das ist genau die Situation, mit denen man es in den Naturwissenschaften und in der Technik meist zu tun hat.

Bezüglich der weiteren Vorgehensweise können wir wieder wie beim arithmetischen Mittel und der Standardabweichung argumentieren: Wir versuchen, den EUKLIDischen Abstand zwischen dem Vektor der tatsächlichen Beobachtungswerte und dem der vorhergesagten Werte zu minimieren, d.h. wir wählen die Parameter $a, b, \ldots$ so, daß

$$Q(a, b, \ldots) \underset{\text{def}}{=} \sum_{i=1}^{N} \big(u_i - f(a, b, \ldots; t_i)\big)^2$$

minimal wird – das ist die von CARL FRIEDRICH GAUSS eingeführte *Methode der kleinsten Quadrate.*

Obiger Ausdruck ist eine Funktion der Parameter $a, b, \ldots$; falls sie differenzierbar ist, kann sie nur dort ein Minimum annehmen, wo die partiellen Ableitungen

$$\frac{\partial Q}{\partial a}, \quad \frac{\partial Q}{\partial b}, \quad \ldots$$

allesamt verschwinden. Dies wird im allgemeinen ein nichtlineares Gleichungssystem definieren, dessen Lösung recht kompliziert sein kann; wir wollen uns hier auf den Fall beschränken, daß die Funktion f *linear* von den Parametern $a, b, \ldots$ abhängt. In diesem Fall bezeichnet man die Suche nach den optimalen Parametern als *lineare Regression;* sie hat sowohl in Naturwissenschaft und Technik als auch in den Wirtschafts- und Sozialwissenschaften zahlreiche interessante Anwendungen.

Falls alle Parameter linear in f eingehen, kommen sie in Q höchstens quadratisch vor; die partiellen Ableitungen sind also wieder linear, so daß insgesamt nur ein lineares Gleichungssystem gelöst werden muß.

Wir betrachten hier nur beispielhaft den einfachsten Fall einer Funktion, die von zwei Parametern abhängt: die lineare Funktion

$$u = f(a, b; t) = at + b\,;$$

die Rechnungen, die wir dafür im folgenden durchführen, funktionieren aber genauso auch für allgemeinere Funktionen f, solange f linear in den Parametern $a, b, \ldots,$ ist; Nichtlinearität in t ist problemlos.

Gesucht sind also reelle Zahlen a und b, bei denen die Funktion

$$Q(a,b)=\sum_{i=1}^{N}(u_i-at_i-b)^2$$

minimal wird. Dort müssen dann insbesondere die beiden partiellen Ableitungen $Q_a(a,b)$ und $Q_b(a,b)$ verschwinden. Bevor wir diese ausrechnen, empfiehlt es sich, die Funktion zunächst durch Ausmultiplizieren nach Termen in a und b zu sortieren:

$$Q(a,b)=\sum_{i=1}^{N}u_i^2+a^2\sum_{i=1}^{N}t_i^2+Nb^2-2a\sum_{i=1}^{N}u_it_i$$
$$-2b\sum_{i=1}^{N}u_i+2ab\sum_{i=1}^{N}t_i\,.$$

Damit ist

$$Q_a(a,b)=2a\sum_{i=1}^{N}t_i^2-2\sum_{i=1}^{N}u_it_i+2b\sum_{i=1}^{N}t_i$$

und

$$Q_b(a,b)=2Nb-2\sum_{i=1}^{N}u_i+2a\sum_{i=1}^{N}t_i\,.$$

Division durch zwei ergibt die beiden Gleichungen für das Verschwinden dieser partiellen Ableitungen:

$$a\sum_{i=1}^{N}t_i^2+b\sum_{i=1}^{N}t_i=\sum_{i=1}^{N}t_iu_i$$

und $$a\sum_{i=1}^{N}t_i+bN=\sum_{i=1}^{N}u_i\,.$$

Um b zu eliminieren, multiplizieren wir die zweite Gleichung mit $\frac{1}{N}\sum_{i=1}^{N}t_i$

$$a\cdot\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)^2+b\sum_{i=1}^{N}t_i=\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)\cdot\left(\sum_{i=1}^{N}u_i\right)$$

und subtrahieren das von der ersten Gleichung:

$$a\left(\sum_{i=1}^{N}t_i^2-\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)^2\right)=\sum_{i=1}^{N}t_iu_i-\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)\cdot\left(\sum_{i=1}^{N}u_i\right),$$

d.h.

$$a=\frac{\sum_{i=1}^{N}t_iu_i-\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)\cdot\left(\sum_{i=1}^{N}u_i\right)}{\sum_{i=1}^{N}t_i^2-\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)^2}\,.$$

Entsprechend läßt sich a eliminieren, wenn wir die erste Gleichung mit $\sum t_i$ und die zweite mit $\sum t_i^2$ multiplizieren; wir erhalten

$$b=\frac{\left(\sum_{i=1}^{N}t_i\right)\cdot\left(\sum_{i=1}^{N}t_iu_i\right)-\left(\sum_{i=1}^{N}t_i^2\right)\cdot\left(\sum_{i=1}^{N}u_i\right)}{\left(\sum_{i=1}^{N}t_i\right)^2-N\sum_{i=1}^{N}t_i^2}\,.$$

Als Beispiel betrachten wir den Zusammenhang zwischen dem Grad der Korruption in einem Land und dessen Bruttosozialprodukt pro Einwohner. Jedes Jahr veröffentlicht die Organisation *Transparancy International* ihren *corruption perceptions index (CPI),* in dem jedem Land eine Zahl zwischen null und zehn zugeordnet wird, je nachdem, wie stark Geschäftsleute, Risikospezialisten und die Bevölkerung die Korruption im betreffenden Land einschätzen: Ein Index von zehn bedeutet, daß es praktische keine Korruption gibt, während bei null nichts läuft ohne Bimbes. Die neuesten Daten stammen von Juni 2001 und sind unter

`http://www.transparency.org/documents/cpi/2001/cpi2001.de.html`

zu finden. Die Zahlen werden als Mittelwerte über die letzten drei Jahren berechnet, so daß singuläre Ereignisse eines Jahres nicht zu sehr ins Gewicht fallen.

Wir vergleichen diese Zahlen mit dem Bruttosozialprodukt pro Einwohner, wie es die Weltbank für 1998 festgestellt hat. Dies sind die neuesten verfügbaren Zahlen; sie sind beispielsweise auf dem Server des Statistischen Bundesamtes unter

`http://www.statistik-bund.de/basis/d/ausl/ausllae5.htm`

zu finden. In der folgenden Tabelle sind Staaten aufgelistet, für die sowohl das Bruttosozialprodukt pro Einwohner als auch der CPI für 2001 vorliegt; das Bruttosozialprodukt pro Einwohner in US-$ ist kursiv gedruckt, der Korruptionsindex fett.

Ägypten	*1290*	**3,6**
Argentinien	*8030*	**3.5**
Aserbaidschan	*480*	**2,0**
Australien	*20640*	**8,5**
Bangladesch	*350*	**0,4**
Belgien	*25380*	**6,6**
Bolivien	*1010*	**2,0**
Botswana	*3070*	**6,0**
Brasilien	*4630*	**4,0**
Bulgarien	*1220*	**3,9**
Chile	*4990*	**7,5**
China (ohne Hongkong & Taiwan)	*750*	**3,5**
Costa Rica	*2720*	**4,5**
Dänemark	*33040*	**9,5**
Deutschland	*28570*	**7,4**
Dominikanische Republik	*1770*	**3.1**
Ecuador	*1520*	**2,3**
Elfenbeinküste	*700*	**2,4**
El Salvador	*1850*	**3,6**
Estland	*3360*	**5,6**
Finnland	*24280*	**9,9**
Frankreich	*24210*	**6,7**
Ghana	*390*	**3,4**
Griechenland	*11740*	**4,2**
Großbritannien und Nordirland	*21410*	**8,3**
Guatemala	*1640*	**2,9**

Honduras	*740*	**2,7**
Hongkong	*23660*	**7,9**
Indien	*440*	**2,7**
Indonesien	*640*	**1,9**
Irland	*18710*	**7,5**
Island	*27830*	**9,2**
Israel	*16180*	**7,6**
Italien	*20090*	**5,5**
Japan	*32350*	**7,1**
Jordanien	*1150*	**4,9**
Kamerun	*610*	**2,0**
Kanada	*19170*	**8,9**
Kasachstan	*1340*	**2,7**
Kenia	*350*	**2,0**
Kolumbien	*2470*	**3,8**
Korea, Republik	*8600*	**4,2**
Kroatien	*4620*	**3,9**
Lettland	*2420*	**3,4**
Litauen	*2540*	**4,8**
Luxemburg *(BSP von 1996)*	*45100*	**8,7**
Malawi	*210*	**3,2**
Malaysia	*3670*	**5,0**
Mauritius	*3730*	**4,5**
Mexiko	*3840*	**3,7**
Moldau	*380*	**3,1**
Namibia	*1940*	**5,4**
Neuseeland	*14600*	**9,4**
Nicaragua	*370*	**2,4**
Niederlande	*24780*	**8,8**
Nigeria	*300*	**1,0**
Norwegen	*34310*	**8,6**
Österreich	*26830*	**7,8**
Pakistan	*470*	**2,3**
Panama	*2990*	**3,7**
Peru	*2440*	**4,1**
Philippinen	*1050*	**2,9**

Polen	*3910*	**4,1**
Portugal	*10670*	**6,3**
Rumänien	*1360*	**2,8**
Rußland	*2260*	**2,3**
Sambia	*330*	**2,6**
Schweden	*25580*	**9,0**
Schweiz	*39980*	**8,4**
Senegal	*520*	**2,9**
Simbabwe	*620*	**2.9**
Singapur	*30170*	**9,2**
Slowakei	*3700*	**3,7**
Slowenien	*9780*	**5,2**
Spanien	*14100*	**7,0**
Südafrika	*3310*	**4,8**
Taiwan	*1233*	**5,9**
Tansania, Vereinigte Republik	*220*	**2,2**
Thailand	*2740*	**3,2**
Trinidad und Tobago	*4520*	**5.3**
Tschechien	*5150*	**3,9**
Tunesien	*2060*	**5,3**
Türkei	*3160*	**3,6**
Uganda	*310*	**1,9**
Ungarn	*4510*	**5,3**
Ukraine	*980*	**2,1**
Uruguay	*6070*	**5,1**
Usbekistan	*950*	**2,7**
Venezuela	*3530*	**2,8**
Vereinigte Staaten	*29240*	**7,6**
Vietnam	*350*	**2,6**

Der erste Augenschein zeigt, daß korruptionsärmere Länder oftmals reicher sind: Das weitgehend korruptionsfreie Dänemark hat ein Bruttosozialprodukt von 33 040 $ pro Einwohner, das deutlich korruptere Deutschland nur 28 570 $ und ein stark korruptes Land wie Bangladesch nur 350 $. Allerdings gibt es auch Ausnahmen, denn Chile hat, obwohl geringfügig weniger korrupt als Deutschland, nur einBruttosozialpro-

dukt von 4 990 $pro Einwohner. Es gibt also sicherlich keinen deterministischen Zusammenhang zwischen Korruption und Wohlstand, aber es lohnt sich doch, eine Ausgleichsgerade zu berechnen. Obige Formeln führen auf

$$\text{CPI} = 3{,}18723 + 0{,}18105 \cdot 10^{-3} \cdot \text{BSP}\,;$$

Abbildung 55 zeigt letztere Gerade zusammen mit den Datenpunkten.

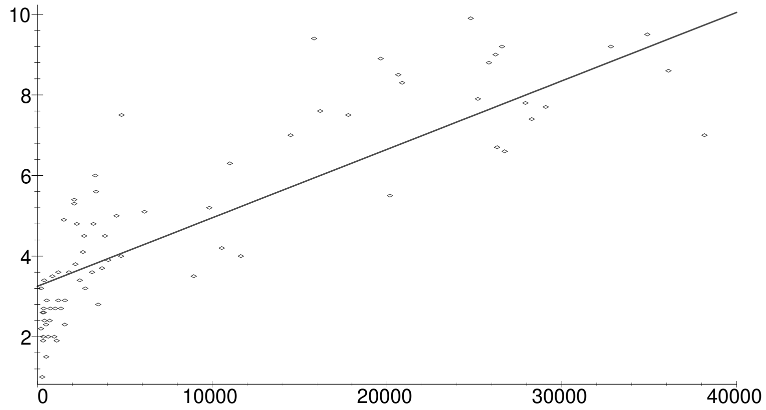


Abb. 55: Korruption 2001 über Bruttosozialprodukt 1998 für 91 Staaten

Natürlich liegen nicht alle Datenpunkte auf der Ausgleichsgeraden: Beispielsweise sind Japan, Belgien und Italien deutlich korrupter, als ihre Wirtschaftskraft vermuten läßt, wohingegen Dänemark, Finnland und Neuseeland deutlich ärmer sind, als man aufgrund der geringen Korruption erwarten sollte. Trotzdem beschreibt die Ausgleichsgerade zumindest einen gewissen Trend.

Für Deutschland ergibt die Formel einen erwarteten Index von 8,05, was zwar praktisch perfekt mit den 1999er CPI von 8,0 übereinstimmt, aber deutlich über dem 2001er CPI von 7,4 liegt. Da 1999 noch kaum jemand das Wort „Bimbes“ kannte, während 2000 „Schwarzgeldaffäre“ zum Wort des Jahres wurde, ist dieser Rückgang nicht weiter verwunderlich und wird wohl auch nächstes Jahr weitergehen, denn der CPI beruht ja auf Studien der vergangenen drei Jahre.

Ähnlich sieht es aus, wenn man das Klausurergebnis der Scheinklausur zu dieser Vorlesung mit den bearbeiteten Übungsaufgaben in Verbindung bringt: Hier ergibt sich die Gerade

$$\textit{Klausurpunkte} = 11{,}78054131 + 0{,}05813492738 \cdot \textit{Übungsprozente}\,,$$

die in Abbildung 56 zusammen mit den Datenpunkten eingezeichnet ist.

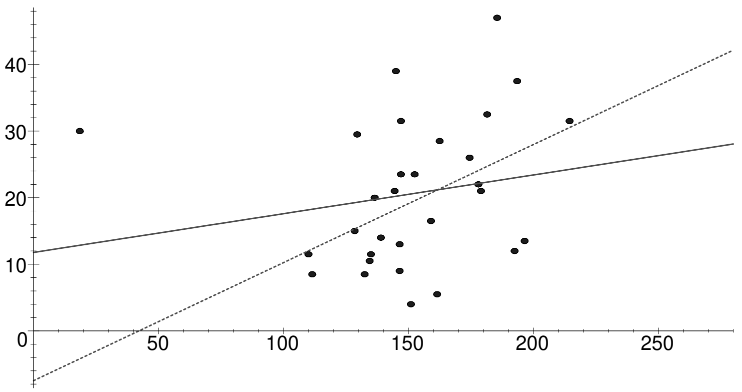


Abb. 56: Zusammenhang zwischen Klausurergebnis und Übungen

Wieder sieht man, daß im großen und ganzen höhere Punktezahlen bei den Übungsaufgaben einem besseren Klausurergebnis entsprechen, aber es gibt auch recht deutliche Ausreißer, deren Interpretation dem Leser überlassen sei.

Wie sehr Ausgleichsgeraden von einzelnen Datenpunkten abhängen können sieht man an der gestrichelt eingezeichneten Geraden

$$\textit{Klausurpunkte} = -7{,}473678952 + 0{,}1772199512 \cdot \textit{Übungsprozente}\,,$$

die fast anhand derselben Daten berechnet wurde: Lediglich der eine Datenpunkt für einen Studenten, der nur ein einziges Übungsblatt bearbeitete und trotzdem die siebtbeste Klausur schrieb, blieb unbeachtet.

Unser nächstes Ziel ist ein Maß, mit dem wir die Qualität einer Ausgleichsgerade $u = at + b$ zu N Datenpaaren (t_i, u_i) beurteilen können.

Zu seiner Definition gehen wir ähnlich vor wie bei der Herleitung des GAUSSschen Fehlerfortpflanzungsgesetzes: $\bar t$ sei der Mittelwert der t_i und $\bar u$ der der u_i. Falls die u_i und die t_i voneinander unabhängig sind, erwarten wir, daß

$$\sum_{i=1}^N (t_i - \bar t)(u_i - \bar u)$$

verschwinden sollte, denn die Abweichungen $t_i - \bar t$ der t_i von ihrem Mittelwert sollten nichts mit den Abweichungen $u_i - \bar u$ der u_i von deren Mittelwert zu tun haben, und im Mittel sind beide Abweichungen null.

Tatsächlich ist obiges Argument kein Beweis, sondern die Definition des Wortes „unabhängig“: Wir bezeichnen

$$\operatorname{cov}(\boldsymbol{t}, \boldsymbol{u}) \underset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^N (t_i - \bar t)(u_i - \bar u)$$

als die *Kovarianz* der Meßreihen

$$\boldsymbol{t} = (t_1, \dots, t_N) \quad \text{und} \quad \boldsymbol{u} = (u_1, \dots, u_N)$$

und bezeichnen die beiden Meßreihen als statistisch unabhängig, falls diese Kovarianz verschwindet.

Falls im anderen Extrem $u_i = at_i + b$ ist für alle i, so ist $\bar u = a\bar t + b$ und

$$\begin{aligned}\sum_{i=1}^N (t_i - \bar t)(u_i - \bar u) &= \sum_{i=1}^N (t_i - \bar t)(at_i - a\bar t)\\ &= a\sum_{i=1}^N (t_i - \bar t)^2\end{aligned}$$

eine (fast immer) positive Zahl, die eng mit der Standardabweichung der Verteilung der t_i zusammenhängt. Um sie weiter auszurechnen, beachten wir, daß

$$\sum_{i=1}^N (u_i - \bar u)^2 = a^2 \sum_{i=1}^N (t_i - \bar t)^2$$

ist und damit

$$\sum_{i=1}^{N}(t_i-\bar{t})^2\cdot\sum_{i=1}^{N}(u_i-\bar{u})^2=a^2\left(\sum_{i=1}^{N}(t_i-\bar{t})^2\right)^2$$

oder

$$\sqrt{\sum_{i=1}^{N}(t_i-\bar{t})^2\cdot\sum_{i=1}^{N}(u_i-\bar{u})^2}=a\sum_{i=1}^{N}(t_i-\bar{t})^2\,.$$

Damit ist also der Quotient

$$\rho=\frac{\displaystyle\sum_{i=1}^{N}(t_i-\bar{t})(u_i-\bar{u})}{\sqrt{\displaystyle\sum_{i=1}^{N}(t_i-\bar{t})^2\cdot\sum_{i=1}^{N}(u_i-\bar{u})^2}}$$

je nach Vorzeichen von a gleich ± 1, wenn alle $u_i=at_i+b$ sind, und $r=0$, wenn die u_i unabhängig von den t_i sind. Für beliebige Wertepaare (t_i,u_i) sagt uns die CAUCHY-SCHWARZsche Ungleichung aus **[HM1]**, Kap. 2, §1*a)*, daß stets

$$-1\le\rho\le 1$$

ist.

Diesen Wert ρ nennen wir den *Korrelationskoeffizienten* der beiden Meßreihen $\boldsymbol{t}$ und $\boldsymbol{u}$; sein Betrag gibt an, wie gut die Beziehung zwischen den u_i und den t_i durch eine Gerade beschrieben werden kann, während sein Vorzeichen angibt, ob diese Gerade ansteigt (die Daten sind *positiv korreliert*) oder sinkt (*negative Korrelation*).

Kürzer läßt er sich auch schreiben als

$$\rho=\frac{\operatorname{cov}(\boldsymbol{t},\boldsymbol{u})}{\sigma_{\boldsymbol{t}}\sigma_{\boldsymbol{u}}}\,,$$

wobei $\sigma_{\boldsymbol{t}}$ und $\sigma_{\boldsymbol{u}}$ die Standardabweichungen der beiden Meßreihen bezeichnen.

Bei den obigen Zahlenbeispielen ergeben sich die Korrelationskoeffizienten zu

$$\rho\approx 0{,}8483022667$$

für den Zusammenhang zwischen Korruption und Bruttosozialprodukt; ;die Korrelation ist also recht gut. Bei den Klausurergebnissen dagegen erhalten wir nur

$$\rho\approx 0{,}1910296117 \quad \textit{bzw.} \quad \rho\approx 0{,}4218151723\,,$$

falls der eine Ausreißer nicht mitgerechnet wird – beides nicht gerade sehr berauschend. Letztes Jahr lag der entsprechende Korrelationskoeffizient übrigens bei

$$\rho\approx 0{,}7465571442\,;$$

wer will, kann über mögliche Gründe für diesen Unterschied spekulieren.

Zum Abschluß sei noch darauf hingewiesen, daß auch ein Korrelationskoeffizient nahe bei ± 1 nicht unbedingt bedeutet, daß ein *inhaltlicher* Zusammenhang zwischen den verglichenen Größen besteht: Die Korrelation könnte auch rein formal sein. Ein beliebtes Beispiel, zu dem leider nirgends Zahlenwerte zu finden sind, ist die gute Korrelation zwischen dem Rückgang der Storchenpopulationen und der Geburtenrückgang zur Zeit der Industrialisierung: Die meisten Bevölkerungsexperten gehen hier nicht von einem kausalen Zusammenhang aus, sondern eher davon, daß die Industrialisierung den Lebensraum der Störche zerstörte und wegen der gleichzeitigen Einführung der Rentenversicherung die Familien nicht mehr so viele Kinder zur Alterssicherung brauchten.

§5: Die Gaußsche Normalverteilung und die Maximum Likelihood Methode

a) Der Grenzfall des Laplaceschen Fehlermodells

Leider ist in dieser Vorlesung nicht genug Zeit, um auch nur annähernd die notwendigen Grundlagen für einen Beweis des zentralen Grenzwertsatzes bereitszustellen. Wir haben aber immerhin schon am Beispiel des LAPLACEschen Fehlermodell graphisch gesehen, daß für die Verteilungen mit $u_i=\pm\varepsilon$, jeweils mit Wahrscheinlichkeit $\frac{1}{2}$, die Verteilung der Summen gegen eine Normalverteilung konvergiert.

Zumindest in diesem einfachen Fall können wir dies auch rechnerisch einsehen und auf diesem Weg insbesondere auch die Gleichung der Glockenkurve herleiten. Dazu müssen wir uns zunächst überlegen, was ε tun soll, wenn n gegen unendlich geht.

Wir kennen zwei statistische Kennzahlen zur Beschreibung der Fehlerverteilung: Das arithmetische Mittel und die Varianz. Das arithmetische Mittel ist (z.B. aus Symmetriegründen) null, bleibt also noch die Varianz.

Zu deren Berechnung müssen wir die Fehlerquadrate über die 2^n möglichen Verhaltensweisen der „Dämonen" summieren, d.h. wir müssen die Streuung

$$\sigma^2 = \frac{1}{2^n} \sum (\varepsilon_1 + \ldots + \varepsilon_n)^2$$

berechnen, wobei sich die Summation über alle n-tupel

$$(\varepsilon_1, \cdots, \varepsilon_n) \quad \text{mit} \quad \varepsilon_i = \pm\varepsilon$$

erstreckt. Beim Ausmultiplizieren heben sich alle gemischten Terme der Form $\varepsilon_i\varepsilon_j$ gegenseitig weg, denn das Tupel, bei dem nur an der i-ten Stelle das Vorzeichen geändert wurde, liefert einen Summanden $-\varepsilon_i\varepsilon_j$. Also bleiben nur die Quadrate; diese sind alle gleich ε^2, und es sind pro Summand n Stück. Da die Anzahl der Summanden gleich dem Nenner des Vorfaktors ist, berechnet sich die Varianz daher zu

$$\sigma^2 = n\varepsilon^2$$

Somit müssen wir für $n \to \infty$ den Einfluß ε jedes einzelnen „Dämonen" so gegen null gehen lassen, daß $n\varepsilon^2$ konstant bleibt, d.h. wir setzen

$$\varepsilon = \frac{\sigma}{\sqrt{n}}$$

für eine geeignet zu wählende Konstante $\sigma > 0$, die Standardabweichung.

Für festes n kann der Fehler einen der $n + 1$ Werte

$$-n\varepsilon, \quad -(n-2)\varepsilon, \quad \ldots, \quad (n-2)\varepsilon, \quad n\varepsilon$$

annehmen, was wir in der Form

$$u = (n-2k)\varepsilon = \frac{(n-2k)\sigma}{\sqrt{n}} \quad \text{mit} \quad k = -n, \ldots, n$$

schreiben wollen. Dieser Fehler tritt genau dann auf, wenn k der Dämonen den Fehler ε erzeugen und die restlichen $n - k$ den Fehler $-\varepsilon$. Dies geschieht in $\binom{n}{k}$ der 2^n möglichen Fälle; die Wahrscheinlichkeit dafür ist also $\binom{n}{k}2^{-n}$.

Für $n \to \infty$ geht dieser Ausdruck gegen null, denn mit n geht schließlich auch die Anzahl der zu betrachtenden Fälle gegen unendlich. Falls es ein Intervall gäbe, in dem die Wahrscheinlichkeit für jeden darin liegenden Fehler größer als irgendein $\alpha > 0$ wäre, ginge allein schon die Summe der Wahrscheinlichkeiten für Fehler aus diesem Teilintervall mit n gegen unendlich, da die Anzahl der dort liegenden möglichen u-Werte wegen der $\sqrt{n}$ im Nenner von u gegen unendlich geht. Da die Summe aller Wahrscheinlichkeiten aber nicht größer als eins werden kann, muß die Wahrscheinlichkeit also in jedem einzelnen Punkt für $n \to \infty$ gegen null gehen.

Wenn wir n variieren lassen, ist es allerdings ohnehin sinnlos, einen genauen Wert des Fehlers zu betrachten: Für jedes n gibt es nur $n + 1$ mögliche Werte, und bei den meisten größeren Werten von n werden diese Zahlen – abgesehen von der Null – nicht auftreten. Wenn wir etwas von n unabhängiges definieren möchten (und sofern wir nicht an Dämonen glauben, bleibt uns kaum etwas anderes übrig) dürfen wir also nicht den genauen *Wert* des Fehlers festlegen, sondern müssen ein *Fehlerintervall* betrachten.

Nun wollen wir aber natürlich als Ergebnis keine Funktion, die von einem Intervall abhängt, sondern eine gewöhnliche Funktion von u. Um eine solche zu bekommen, betrachten wir nicht die Wahrscheinlichkeit, sondern die *Wahrscheinlichkeitsdichte:* Für eine kontinuierlich variierende zufällige Größe definieren wir die Wahrscheinlichkeitsdichte $\varphi(u_0)$ im Punkt u_0 als

$$\varphi(u_0) = \lim_{\varepsilon \to 0} \frac{\text{Wahrscheinlichkeit für } u_0 - \varepsilon \le u \le u_0 - \varepsilon}{2\varepsilon},$$

d.h. also als Wahrscheinlichkeit dividiert durch die Intervallbreite. Falls diese Dichte existiert, folgt mehr oder weniger sofort aus der Definition des RIEMANN-Integrals, daß

$$(\text{Wahrscheinlichkeit für } a \le u \le b) = \int_a^b \varphi(u)\,du$$

ist. Unsere Ziel ist also, diese Wahrscheinlichkeitsdichte φ für $n \to \infty$ zu berechnen.

Für festes n haben die möglichen Fehlerwerte den Abstand 2ε, wir betrachten daher Intervalle der Länge 2ε mit den Werten $(n-2k)\varepsilon$ als Mittelpunkten; diese überdecken den möglichen Fehlerbereich lückenlos, und die Wahrscheinlichkeit für einen Fehler in diesem Intervall ist $\binom{n}{k}2^{-n}$.

Wir interessieren uns daher für den Grenzwert von

$$\frac{\binom{n}{k}2^{-n}}{2\varepsilon} = \frac{\binom{n}{k}2^{-n}}{2\sigma/\sqrt{n}} = \binom{n}{k}2^{-n-1}\frac{\sqrt{n}}{\sigma}$$

für $n \to \infty$.

b) Die Eulersche Summenformel

Das Problem bei der Berechnung dieses Grenzwerts ist der Binomialkoeffizient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}\,;$$

um diesen abzuschätzen brauchen wir einen handhabbaren Ausdruck für $n!$.

Dazu schreiben wir

$$\ln n! = \sum_{k=1}^{n} \ln k$$

und berechnen dies nach einer Methode von EULER, die nicht nur für Summen von Logarithmen anwendbar ist.

Wir betrachten allgemeiner irgendeine reellwertige differenzierbare Funktion f, deren Definitionsbereich das Intervall $[1,\ n]$ enthält.

Für eine reelle Zahl x bezeichnen wir wie üblich mit $[x]$ die größte ganze Zahl kleiner oder gleich x und mit

$$\{x\} \underset{\text{def}}{=} x - [x]$$

den gebrochenen Anteil von x; ist k eine ganze Zahl, ist somit für $x \in [k,\ k+1)$

$$\{x\} = x - k\,.$$

Partielle Integration führt auf die Gleichung

$$\int_k^{k+1} \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \left(x - k - \tfrac{1}{2}\right) f(x)\,\Big|_k^{k+1} - \int_k^{k+1} f(x)\,dx$$

$$= \frac{f(k+1)+f(k)}{2} - \int_k^{k+1} f(x)\,dx\,.$$

Addition aller solcher Gleichungen von $k=1$ bis $k=n-1$ liefert

$$\int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \frac{f(1)}{2} + \sum_{k=2}^{n-1} f(k) + \frac{f(n)}{2} - \int_1^n f(x)\,dx\,,$$

womit man die Summe der $f(k)$ berechnen kann:

Satz (EULERsche Summenformel): Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}$, deren Definitionsbereich das Intervall $[1,\ n]$ umfaßt, ist

$$\sum_{k=1}^{n} f(k) = \int_1^n f(x)\,dx + \frac{f(1)+f(n)}{2} + \int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx\,.$$

■

Für die Abschätzung der Binomialkoeffizienten und Fakultäten interessiert uns speziell der Fall $f(x) = \ln x$; hierfür wird die EULERsche

Summenformel zu

$$\ln n! = \int_1^n \ln x\,dx + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx$$
$$= x(\ln x - 1)\Big|_1^n + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx$$
$$= n(\ln n - 1) + 1 + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac12}{x}\,dx\,.$$

In dieser Formel stört noch das rechte Integral; dieses können wir wie folgt abschätzen: Für eine natürliche Zahl k ist

$$\int_k^{k+1} \frac{\{x\} - \frac12}{x}\,dx = \int_{-\frac12}^{\frac12} \frac{x}{k+\frac12+x}\,dx$$
$$= \int_0^{\frac12} \left(\frac{x}{k+\frac12+x} - \frac{x}{k+\frac12-x}\right)dx$$
$$= \int_0^{\frac12} \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2}\,dx\,.$$

Im Intervall von 0 bis $\frac12$ ist der Integrand monoton fallend, d.h.

$$0 \ge \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2} \ge \frac{-\frac12}{\left(k+\frac12\right)^2 - \frac14} = \frac{-2}{(2k+1)^2 - 1} \ge -\frac{1}{4k^2}\,,$$

und damit ist

$$0 \ge \int_k^{k+1} \frac{\{x\} - \frac12}{x}\,dx = \int_0^{\frac12} \frac{-2x^2}{\left(k+\frac12\right)^2 - x^2}\,dx \ge -\frac{1}{8k^2}\,,$$

denn wir können das Integral abschätzen durch das Produkt aus der Länge des Integrationsintervalls und dem Minimum des Integranden. Summation von $k = 1$ bis $n-1$ schließlich gibt die Abschätzung

$$0 \ge \int_1^n \frac{\{x\} - \frac12}{x}\,dx \ge -\sum_{k=1}^{n-1} \frac{1}{4k^2}$$

für das störende Integral aus der obigen Formel.

Wie wohl jeder schon einmal in einer Analysis I Übungsaufgabe zeigen mußte, konvergiert die rechtsstehende Summe (egal ob mit oder ohne acht im Nenner) für $n \to \infty$; aus Kapitel III, §3*f)* wissen wir sogar, daß der Grenzwert $\pi^2/48$ ist. Auf jeden Fall können wir folgern, daß das uneigentliche Integral

$$\int_1^\infty \frac{\{x\} - \frac12}{x}\,dx$$

konvergiert; den uns bislang noch unbekannten Grenzwert wollen wir mit I bezeichnen. Damit ist

$$\ln n! = n(\ln n - 1) + \frac{\ln n}{2} + C + o(1) \quad \text{mit} \quad C = I + 1$$

oder

$$n! \approx e^C \cdot n^n e^{-n} \sqrt{n}\,.$$

c) Die Stirlingsche Formel und die Normalverteilung

Mit dieser Formel können wir nun daran gehen, den Binomialkoeffizienten $\binom{n}{k}$ abzuschätzen:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{e^C n^n e^{-n}\sqrt{n}}{e^C k^k e^{-k}\sqrt{k}\, e^C (n-k)^{n-k} e^{-(n-k)}\sqrt{(n-k)}}$$
$$= \frac{1}{e^C} \frac{n^n}{k^k\,(n-k)^{n-k}} \sqrt{\frac{n}{k\cdot(n-k)}}\,.$$

Ausgedrückt durch $u = (n-2k)\varepsilon = (n-2k)\sigma/\sqrt{n}$ ist

$$k = \frac{n}{2} - \frac{u\sqrt{n}}{2\sigma}\,,$$

und setzen wir zur Vereinfachung der Schreibweise

$$m = \frac{n}{2}, \quad v = \frac{u}{2\sigma} \quad \text{und} \quad \ell = \sqrt{2m}\cdot v\,,$$

so ergeben sich die Formeln

$$k = m - \sqrt{2m}\cdot v = m - \ell \quad \text{und} \quad n - k = m + \sqrt{2m}\cdot v = m + \ell\,.$$

Setzen wir dies alles in die Formel für die Wahrscheinlichkeitsdichte ein, erhalten wir

$$\begin{aligned}
&\binom{n}{k} 2^{-n-1}\frac{\sqrt{n}}{\sigma}\\
\approx\ & \frac{1}{e^C}\frac{n^n}{k^k\,(n-k)^{n-k}}\sqrt{\frac{n}{k\cdot(n-k)}}\,2^{-n-1}\frac{\sqrt{n}}{\sigma}\\
=\ & \frac{1}{e^C\sigma}\frac{n^n\cdot 2^{-n}}{k^k\,(n-k)^{n-k}}\frac{n\cdot 2^{-1}}{\sqrt{k\cdot(n-k)}}\\
=\ & \frac{1}{e^C\sigma}\frac{(2m)^{2m}\cdot 2^{-2m}}{(m-\ell)^{m-\ell}(m+\ell)^{m+\ell}}\frac{2m\cdot 2^{-1}}{\sqrt{(m-\ell)(m+\ell)}}\\
=\ & \frac{1}{e^C\sigma}\frac{m^{2m}}{(m-\ell)^m(m+\ell)^m}\left(\frac{m-\ell}{m+\ell}\right)^{\ell}\frac{m}{\sqrt{m^2-\ell^2}}\\
=\ & \frac{1}{e^C\sigma}\frac{m^{2m}}{(m^2-\ell^2)^m}\left(\frac{m-\ell}{m+\ell}\right)^{\ell}\frac{m}{\sqrt{m^2-\ell^2}}\\
=\ & \frac{1}{e^C\sigma}\frac{1}{\left(1-\frac{\ell^2}{m^2}\right)^m}\left(\frac{1-\ell/m}{1+\ell/m}\right)^{\ell}\frac{1}{\sqrt{1-\ell^2/m^2}}\\
=\ & \frac{1}{e^C\sigma}\frac{1}{\left(1-\frac{2v^2}{m}\right)^m}\left(\frac{1-\sqrt{2/m}\cdot v}{1+\sqrt{2/m}\cdot v}\right)^{\sqrt{2m}\cdot v}\frac{1}{\sqrt{1-2v^2/m}}\,.
\end{aligned}$$

Nun können wir langsam daran denken, n (und damit auch m) gegen unendlich gehen zu lassen; wir verwenden dazu die aus der Analysis I und wahrscheinlich auch aus der Schule bekannte Beziehung

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n\,.$$

Danach ist insbesondere

$$\lim_{m\to\infty}\left(1-\frac{2v^2}{m}\right)^m = e^{-2v^2}$$

und

$$\begin{aligned}
\lim_{m\to\infty}\left(1\pm\sqrt{\frac{2}{m}}\cdot v\right)^{\sqrt{2m}\cdot v} &= \lim_{m\to\infty}\left(1\pm\frac{\sqrt{2}\cdot v}{\sqrt{m}}\right)^{\sqrt{m}\cdot\sqrt{2}\cdot v}\\
= \lim_{q\to\infty}\left(1\pm\frac{\sqrt{2}\cdot v}{q}\right)^{q\cdot\sqrt{2}\cdot v} &= \left(e^{\pm\sqrt{2}\cdot v}\right)^{\sqrt{2}\cdot v} = e^{\pm 2v^2}\,,
\end{aligned}$$

denn es bleibt sich natürlich gleich, ob m oder $q = \sqrt{m}$ gegen unendlich geht. Da der Term v^2/m gegen null geht, erhalten wir somit als Grenzwert des gesamten obigen Ausdrucks

$$\frac{1}{e^C\sigma}\cdot\frac{1}{e^{-2v^2}}\cdot\frac{e^{-2v^2}}{e^{+2v^2}}\cdot 1 = \frac{1}{e^C\sigma}e^{-2v^2}\,.$$

Beachten wir nun noch, daß $v = u/2\sigma$ war, erhalten wir

$$\frac{1}{e^C\sigma}e^{-u^2/2\sigma^2}\,.$$

Damit sind wir fast am Ziel; das einzige, was noch fehlt, ist die Konstante C. Diese können wir bestimmen, indem wir ausnutzen, daß jeder Fehler mit Wahrscheinlichkeit eins zwischen $-\infty$ und ∞ liegt, d.h.

$$\frac{1}{e^C\sigma}\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\,du = 1\,.$$

Aus **[HM1]**, *Kap.* 2, §6*c),* wissen wir, daß

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

ist; mit der Substitution $x = u/\sqrt{2}\sigma$ folgt, daß dann

$$\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\,du = \sqrt{2\pi}\sigma$$

ist und

$$e^C = \sqrt{2\pi} \quad \text{oder} \quad C = \tfrac{1}{2}\ln(2\pi)\,.$$

Damit haben wir die Wahrscheinlichkeitsdichte endlich vollständig berechnet; das Endergebnis ist

$$\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-u^2}{2\sigma^2}}\,.$$

Auch die Formel für $n!$ können wir nach der Bestimmung von C nun vollständig hinschreiben:

$$\ln n! = n(\ln n - 1) + \ln\sqrt{2\pi n} + o(1) \quad \text{oder} \quad n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\,.$$

Beides bezeichnet man als STIRLINGsche Formel.

Der schottische Mathematiker JAMES STIRLING (1692–1770) war Anhänger des gestürzten Königs Jakob II Stuart und hatte deshalb große politische Probleme bei seinem Studium; unter anderem wurde er deshalb von der Universität Oxford ausgeschlossen. 1717–1722 lebte er in Venedig und hatte auch gute Kontakte zu NICOLAUS BERNOULLI an der Universität von Padua; außerdem brachte er aus Venedig die Produktionsgeheimnisse der dortigen Glasbläser mit. Ab 1724 arbeitete er zehn Jahre lang als Mathematiklehrer in London, wo er viel mit NEWTON zusammentraf; 1735 wurde er Direktor einer schottischen Bergbaugesellschaft. In seine Londoner Zeit fällt die Veröffentlichung seines bedeutendsten Werks *Methodus Differentialis sive Tractatus de Summatione er Interpolatione Serierum Infinitarum* im Jahre 1730, das die obige Formel als Beispiel zwei zu Proposition 28 enthält. Ebenfalls ziemlich bekannt wurde seine 1735 veröffentlichte Arbeit über die Gestalt der Erde.

d) Eigenschaften der Normalverteilung

Oft interessiert nicht so sehr die Verteilung der Fehler, sondern die der Meßwerte selbst. Ist $\hat{x}$ der korrekte Wert und x_i der i-te Meßwert dafür, der gemäß $x_i = \hat{x} + u_i$ mit dem Fehler u_i behaftet ist, so können wir mit $x = \hat{x} + u$ die obigen Wahrscheinlichkeitsdichte auch als

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\hat{x})^2/2\sigma^2}$$

schreiben.

Als *Normalverteilung mit Mittelwert a und Standardabweichung σ* bezeichnen wir daher die Verteilung mit Wahrscheinlichkeitsdichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}\,.$$

Diese Wahrscheinlichkeitsdichte hängt offensichtlich nur von der *normierten* Variablen

$$z = \frac{x-a}{\sigma}$$

ab; diese hat Mittelwert null und Standardabweichung eins. Daher gibt es für die Normalverteilung nicht – wie für viele andere statistische Verteilungen – je nach Parameterwerten verschiedene Tabellen, sondern man findet in allen Tabellenwerken nur die Normalverteilung mit Mittelwert null und Standardabweichung eins, man findet also die Wahrscheinlichkeitsdichte

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

und deren Integral

$$F(z) = \int_{-\infty}^{z} f(u)\,du = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du\,.$$

Dieses Integral läßt sich nicht weiter vereinfachen, da sich die Stammfunktion von $e^{-u^2/2}$ nicht durch elementare Funktionen ausdrücken läßt. Für die Bestimmung von $F(z)$ ist man daher auf Tabellen oder Computerprogramme angewiesen; eine graphische Darstellung von $F(z)$ ist in Abbildung 57 zu sehen. Mit dieser Funktion läßt sich die Wahrscheinlichkeit dafür, daß

$$c \le z = \frac{x-a}{\sigma} \le d$$

ist berechnen als $F(d) - F(c)$, und damit läßt sich auch leicht die Wahrscheinlichkeit berechnen, daß x selbst zwischen zwei gegebenen Schranken liegt.

Mißt man beispielsweise die Temperatur eines Wasserbads eine Viertelstunde lang jede Minute und erhält dabei 15 Meßwerte mit Mittelw-

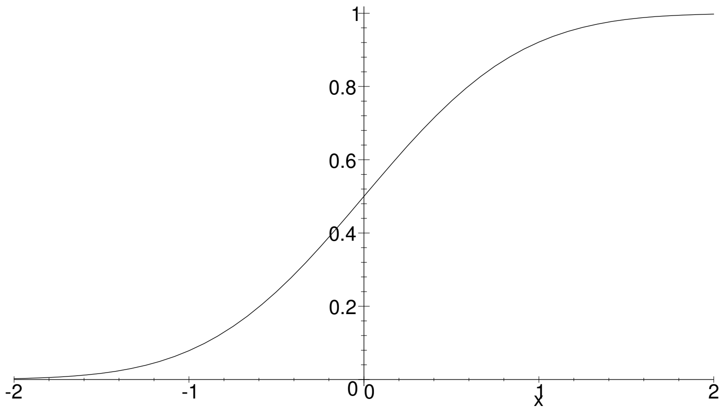


Abb. 57: Das Integral $F(z)$ über die „Glockenkurve“

ert 20,1°C und Standardabweichung 0,2°C, so ist die Standardabweichung des Mittelwerts

$$\sigma_{\overline{\vartheta}} = \frac{0,2^\circ\mathrm{C}}{\sqrt{14}} \approx 0{,}053^\circ\mathrm{C}\,.$$

Wenn wird dann beispielsweise wissen wollen, mit welcher Wahrscheinlichkeit die „tatsächliche“ mittlere Temperatur zwischen 20,0°C und 20,2°C liegt, müssen wir dazu zunächst die normalisierten Werte berechnen:

$$z_1 = \frac{20{,}0 - 20{,}1}{0{,}053} \approx -1{,}89 \quad \text{und} \quad z_2 = \frac{20{,}2 - 20{,}1}{0{,}053} \approx 1{,}89\,.$$

Die Wahrscheinlichkeit ist also

$$F(1{,}89) - F(-1{,}89) \approx 0{,}94\,;$$

oder rund 94%.

Schaut man in einer Tabelle nach, wird man dort allerdings im allgemeinen nur den Wert $F(1{,}89)$ finden, nicht aber $F(-1{,}89)$. Der Grund dafür liegt in der Symmetrie des Graphen von F bezüglich des Punktes $(0, \frac{1}{2})$. Was dahinter steckt, sieht man am besten, wenn man die Dichtefunktion der Normalverteilung betrachtet, also die Glockenkurve:

Für $z > 0$ ist $F(-z)$ die in Abbildung 58 links eingezeichnete schraffierte Fläche. Diese Fläche ist wegen der Symmetrie der Glockenkurve zur senkrechten Achse gleich der rechts eingezeichneten schraffierten Fläche, und deren Komplement ist $F(z)$. Also ist

$$F(-z) = 1 - F(z)\,,$$

und es reicht, wenn wir die Werte von F im positiven Bereich kennen.

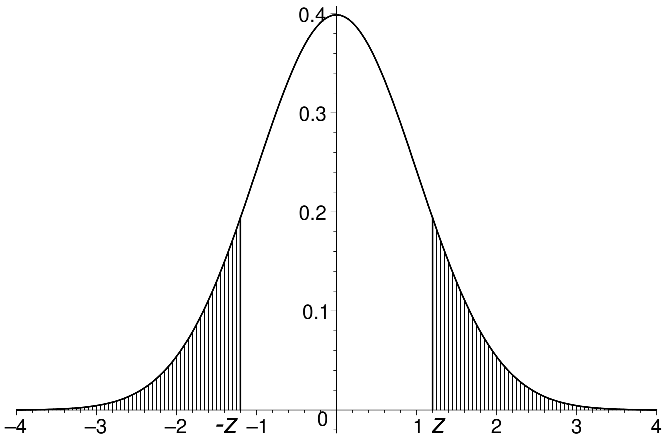


Abb. 58: Zusammenhang zwischen $F(z)$ und $F(-z)$

Oft interessiert auch die Wahrscheinlichkeit dafür, daß der Betrag des Fehlers unterhalb einer bestimmten Schranke liegt, etwa $z \cdot \sigma$; in Abbildung 58 wäre dies der nichtschraffierte Bereich unter der Glockenkurve.

Wie man sich anhand der Abbildung leicht klarmacht, ist diese Wahrscheinlichkeit gleich

$$F(z) - F(-z) = 2F(z) - 1\,;$$

die Wahrscheinlichkeit, daß wir im obigen Beispiel die mittlere Temperatur mit einem Fehler von höchstens 0,05° gemessen haben, ist also

$$2F\left(\frac{0{,}05}{0{,}053}\right) \approx F(0{,}94) \approx 0{,}83\,.$$

Ein Wasserbad hat üblicherweise den Sinn, ein Experiment unter kontrollierten Temperaturbedingungen durchzuführen; daher interessiert vor allem, inwieweit es gelingt, die Temperatur innerhalb gewisser Schranken zu halten. Die Wahrscheinlichkeit dafür können wir mit denselben Methoden berechnen, allerdings müssen wir dazu mit der Standardabweichung der Meßreihe selbst arbeiten.

Wenn wir etwa wollen, daß die Temperatur immer zwischen 19,5 und 20,5°C liegt, so ist die Wahrscheinlichkeit, daß wir dies mit dem oben ausgemessenen Versuchsaufbau erreichen, gleich

$$F\left(\frac{20{,}5-20{,}1}{0{,}2}\right)-F\left(\frac{29{,}5-20{,}1}{0{,}2}\right)=F(2)-F(-3)\approx 0{,}976\,.$$

In knapp zweieinhalb Prozent aller Fälle, im Schnitt also alle vierzig Minuten, müssen wir also damit rechnen, daß die Toleranzgrenzen überschritten werden.

Wie Abbildung 57 zeigt, liegt $F(-2)$ sehr nahe bei null und $F(2)$ sehr nahe bei eins. In der Tat ist die Wahrscheinlichkeit dafür, daß ein Wert z Betrag größer z hat, nach obiger Diskussion gleich

$$1-(2F(z)-1)=2F(z)-2\,,$$

was für $z=2$ zu -0,0455 wird; die Wahrscheinlichkeit ist also kleiner als 5%. Allgemein gilt für eine beliebige Normalverteilung, daß der Wert der Variablen mit folgenden Wahrscheinlichkeiten um höchstens $i\sigma$ vom Mittelwert abweicht:

$i =$	1	2	3	4
Wahrscheinlichkeit:	0, 683	0, 954	0, 9973	0, 99994

Damit liegen also etwa zwei Drittel aller Fehler zwischen $-\sigma$ und σ, 95% liegen zwischen -2σ und 2σ und 99,7% zwischen -3σ und 3σ; die Wahrscheinlichkeit dafür, daß der Fehler größer als 3σ ist, beträgt nur etwa 0,27%. Da Ereignisse mit einer so geringen Wahrscheinlichkeit seltener als in einem von 300 Fällen auftreten, betrachtet man Fehler, die außerhalb des 3σ-Bereichs liegen, oft als „Ausreißer“, d.h. als grobe Meßfehler, die bei der Bestimmung des Ergebnisses nicht berücksichtigt werden. Sehr vorsichtige Leute reden allerdings erst ab einer Abweichung von 4σ von Ausreißern; solche Fehler treten zufällig weniger als einmal pro 15 000 Messungen auf.

Für Leser, die ihren Computer selbst programmieren und keine spezielle Statistiksoftware haben, sei hier eine Näherungsformel für $F(z)$ angegeben: Mit einem Fehler von höchstens $7{,}5\cdot 10^{-8}$ ist

$$F(z)=1-\varphi(z)\cdot\left(a_1t+a_2t^2+a_3a^3+a_4t^4+a_5t^5\right)$$

mit $t=\dfrac{1}{1+pz}$ und $\varphi(z)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ sowie

$a_1 =\ 0{,}319\,381\,530\quad a_2=-0{,}356\,563\,782\quad a_3=1{,}781\,477\,973$
$a_4=-1{,}821\,255\,978\quad a_5=\ 1{,}330\,274\,429\quad p=0{,}231\,641\,9\quad .$

Beim Rechnen mit dem Taschenrechner kann man sich auch mit einer vereinfachten Version begnügen, bei der $a_4=a_5=0$ ist und

$a_1=0{,}436\,1836\quad a_2=-0{,}120\,1676\quad a_3=0{,}937\,2980\quad p=0{,}332\,67\,;$

hier kann der Fehler bis zu 10^{-5} betragen.

e) Die Maximum Likelihood Methode

GAUSS gab im Laufe seines Lebens mehrere Begründungen für die Methode der kleinsten Quadrate (die er bei sowohl bei seinen astronomischen Arbeiten wie auch bei der von ihm geleiteten Vermessung des Königreichs Hannover zwischen 1818 und 1832 ständig benutzte); die unter dem Gesichtspunkt einer in sich geschlossenen Fehlertheorie interessanteste beruht auf dem LAPLACEschen Fehlermodell:

Danach sollte der Wert u_i für die korrekten Parameterwerte $a,b,\dots$ aus einer Normalverteilung mit Mittelwert $f(a,b,\dots;t_i)$ kommen, deren Standardabweichung σ_i von der Genauigkeit abhängt, mit der u_i bestimmt werden kann. Die Wahrscheinlichkeit dafür, daß u_i zwischen zwei Werten a und b liegt, ist damit

$$\int_a^b e^{-\left(u-f(a,b,\dots;t_i)\right)^2/2\sigma_i^2}\,.$$

Von der Wahrscheinlichkeit, daß u_i gleich einem Wert c ist, können wir natürlich nicht reden, da diese nach obiger Formel ein Integral von c nach c wäre, also Null. Aber die Wahrscheinlichkeit dafür, daß u_i in einem kleinen Intervall der Länge ε_i um einen Wert c_i liegt, ist ungefähr proportional zum ε_i-fachen Wert des Integranden an der Stelle c_i, also zu

$$\varepsilon_i \cdot e^{-\left(c_i - f(a,b,\ldots;\,t_i)\right)^2/2\sigma_i^2}\,.$$

Entsprechend ist

$$\varepsilon_j \cdot e^{-\left(c_j - f(a,b,\ldots;\,t_j)\right)^2/2\sigma_j^2}$$

ungefähr gleich der Wahrscheinlichkeit dafür, daß u_j in einem Intervall der Breite ε_j um c_j liegt.

Wenn wir wie üblich davon ausgehen, daß keine systematischen Fehler auftreten, sind die Fehler von u_i und u_j voneinander unabhängig, die Wahrscheinlichkeit dafür, daß (u_i, u_j) in einem Rechteck mit Seiten ε_i und ε_j um (c_i, c_j) liegt, ist also proportional zum Produkt der beiden obigen Einzelwahrscheinlichkeiten, d.h. zu

$$\varepsilon_i\varepsilon_j\, e^{-\left(c_i - f(a,b,\ldots;\,t_i)\right)^2/2\sigma_i^2 - \left(c_j - f(a,b,\ldots;\,t_j)\right)^2/2\sigma_j^2}\,.$$

Entsprechend kann auch die Wahrscheinlichkeit dafür berechnet werden, daß der Punkt $(u_1, \ldots, u_n)$ in einem kleinen gegebenen Quader mit Kantenlängen $\varepsilon_1, \ldots, \varepsilon_n$ liegt; sie ergibt sich zu

$$L(a,b,\ldots)\cdot\prod_{i=1}^{n}\varepsilon_i$$

mit

$$L(a,b,\ldots) \underset{\text{def}}{=} e^{\sum_{i=1}^{n}\left(u_i - f(a,b,\ldots;\,t_i)\right)^2/2\sigma_i^2}\,.$$

Diese Größe ist selbst keine Wahrscheinlichkeit, sondern der Quotient aus einer Wahrscheinlichkeit und einem Volumen; man spricht daher von einer *Wahrscheinlichkeitsdichte.*

Wenn wir diese Wahrscheinlichkeitsdichte als Funktion von $a, b, \ldots$ betrachten, macht sie eine Aussage über die Güte der Parameter:

Schließlich wird man einem Modell, das dem beobachteten Ausgang eines Experiments eine hohe Wahrscheinlichkeit zuweist, eher glauben als einem alternativen Modell, das die beobachteten Daten zu Ausreißern erklärt. Aus diesem Grund kann die Funktion L auch als Maß dafür betrachtet werden, wie „wahrscheinlich“ in irgendeinem umgangssprachlichen (und schwer präzisierbaren) Sinne die Parameter $a, b, \ldots$ sind.

Im englischen gibt es zwei Wörter für Wahrscheinlichkeit: Das romanische Wort *probability* und das germanische Wort *likelihood.* Für den mathematisch exakten Wahrscheinlichkeitsbegriff verwendet man probability, für „Wahrscheinlichkeit“ im Sinne der Funktion L likelihood. Da es im deutschen kein zweites Wort für Wahrscheinlichkeit gibt, spricht man hier in Anlehnung an das Englische von einer *Likelihoodfunktion.*

Die Maximum Likelihood Methode besteht nun genau in dem, was ihr Name besagt: *Man wähle die Parameter a, b, … so, daß die Likelihoodfunktion maximal wird.*

Da $L(a, b, \ldots)$ durch eine Exponentialfunktion beschrieben wird, wird die Likelihoodfunktion genau dann maximal, wenn ihr Exponent maximal wird. Dieser Exponent ist eine negative Zahl, wird also genau dann maximal, wenn sein Betrag *minimal* wird, das heißt, wenn die Quadratsumme

$$\sum_{i=1}^{n}\frac{\left(u_i - f(a,b,\ldots;\,t_i)\right)^2}{2\sigma_i^2}$$

minimal wird.

In vielen Fällen wird die Zuverlässigkeit der einzelnen Paare (t_i, u_i) miteinander vergleichbar sein, so daß alle σ_i gleich sind; in diesem Fall kann man die σ_i ignorieren und einfach die Quadratsumme

$$\sum_{i=1}^{n}\left(u_i - f(a,b,\ldots;\,t_i)\right)^2$$

minimieren, d.h. wir kommen wieder zur klassischen Methode der kleinsten Quadrate. Es gibt aber auch Anwendungen, wie etwa oben beim überexponentiellen Bevölkerungswachstum, bei denen die Verschiedenheit des σ_i sehr wesentlich ist: Sicherlich wird man etwa der auf Volkszählungen beruhenden Weltbevölkerungszahl, die die Vereinten Nationen für 1995 veröffentlichten, mehr Vertrauen entgegenbringen als der Schätzung eines Historikers für die Weltbevölkerung des Jahres Null, und selbst bei ein und derselben Meßreihe im Labor kommt es gelegentlich vor, daß (beispielsweise aufgrund unterschiedlicher Genauigkeit eines Meßinstruments in verschiedenen Bereichen) manche Daten zuverlässiger sind als andere.

§6: Kompression von Bild- und Audiodaten

Zum Abschluß der Vorlesung wollen wir wenigstens kurz eine praktische Anwendung kennenlernen, in der mit Eigenwerten und Eigenvektoren symmetrischer Matrizen, FOURIER-Transformationen und Statistik gleich mehrere der Methoden aus diesem Semester gleichzeitig benötigt werden: die Komprimierung von Bild- und Audiodaten.

a) Datenkompression

Ziel der Datenkompression ist es, eine Datei für Zwecke der Speicherung oder Übertragung möglichst stark zu verkleinern, das aber in einer solchen Weise, daß sich die ursprüngliche Datei aus der verkleinerten wieder exakt rekonstruieren läßt.

Es ist klar, daß es keinen universellen Algorithmus zur Datenkompression geben kann: Gäbe es nämlich ein Verfahren, das für beliebige Dateien einen Kompressionsfaktor $\alpha < 1$ garantieren würde, so könnte man dieses Verfahren iterativ anwenden und nach n Anwendungen eine Kompressionsrate von α^n erreichen. Wenn man n nur hinreichend groß wählt, könnte man daher jede Datei auf weniger als ein Bit komprimieren, was natürlich absurd ist.

Ein Kompressionsverfahren kann also nur auf Dateien mit spezieller Struktur erfolgreich angewandt werden und muß die spezielle Redundanz in diesen Dateien ausnutzen. In Textdateien beispielsweise ist dies die Redundanz der Sprache, die schon bei bloßer Beachtung der höchst unterschiedlichen Buchstabenhäufigkeiten Kompressionen von rund 50% gestattet.

Bilddaten werden typischerweise als Matrizen aus ganzen Zahlen zwischen 0 und 255 digitalisiert; bei Audiodaten nimmt man Vektoren von ganzen Zahlen zwischen 0 und $65\,535 = 2^{16} - 1$ oder $16\,777\,215 = 2^{24} - 1$. (Der Unterschied zwischen den Wertenbereichen liegt darin begründet, daß unser Auge selbst bei gedruckten Bildern mit nur 64 Graustufen praktisch keine Artefakte mehr erkennen kann, wohingegen unser Gehör noch auf sehr feine Unterschiede reagiert.)

Bei einer Musik-CD etwa wird das Signal 44 100-mal pro Sekunde abgetastet (dies bedeutet nach dem Abtasttheorem von NYQUIST, daß ein auf den Bereich von 0 bis 22,05kHz bandbegrenztes Signal fehlerfrei rekonstruiert werden kann), und das Ergebnis wird dann so skaliert und quantisiert (d.h. gerundet), daß eine Zahl zwischen 0 und 65535 entsteht. Bei Bilddaten werden je nach Auflösung und Seitenverhältnis zwischen etwa 256×256 und 1024×1024 Bildpunkte abgetastet, für Schwarzweißbilder nur nach Helligkeit, für Farbbildern nach insgesamt drei Größen, die vom jeweiligen Farbmodell abhängen. Das Ergebnis dieser Abtastungen wird dann entsprechend skaliert und quantisiert.

Typische Komprimierungsverfahren arbeiten daher mit Vektoren oder Matrizen aus Zahlen zwischen 0 und einer geeigneten Zahl M, die aus praktischen Gründen meist von der Form $2^{8r} - 1$ ist, wobei die Zahl r der Empfindlichkeit unserer Sinne angepaßt zwischen eins und drei liegt.

Da man zur eindeutigen Festlegung von N beliebigen Zahlen zwischen 0 und 2^{8r} nicht mit weniger als den $8Nr$ Bit auskommen kann, die man zum Hinschreiben der Zahlen braucht, sehen wir auch hier wieder, daß kein Verfahren *alle* solchen Vektoren komprimieren kann; wir müssen also eine Teilmenge auszeichnen.

Die ideale solche Teilmenge wäre hier natürlich die Menge aller möglicher Bilder (oder Audiosequenzen), aber diese Menge dürfte mathematisch kaum definierbar sein: Schließlich hängt es sehr vom Betrachter ab, welches Pixelmuster er noch als „Bild“ gelten läßt und welches nicht. Sinnvoll läßt sich eine solche Menge daher höchstens definieren, wenn von vornherein feststeht, welche Bilder berücksichtigt werden sollen – und dann ist wohl ein Verfahren, das statt vom Bildinhalt von einer Bildnummer ausgeht, unschlagbar.

Die meisten klassischen Verfahren, die beliebige, aber realistische Bilder komprimieren sollen, gehen aus von einem *statistischen Modell,* das zwar auch viele Matrizen produziert, die niemand als „Bilder“ anerkennen würde, das aber dennoch genügend viele Eigenschaften realer Bilder reproduziert, um eine große Anzahl von „Nichtbildern“ auszuschließen.

Ausgangspunkt ist die Beobachtung, daß es in einem Bild oder Musikstück nur wenige aprupte Übergänge gibt. Zwar gibt es natürlich immer wieder ein plötzliches *fortissimo,* das auf eine leise Stelle folgt, aber da das Signal 44 100-mal pro Sekunde abgetastet wird und solche Übergänge selbst bei der schrägsten Musik deutlich seltener als im Sekundenrythmus erfolgen, sind diese Sprünge innerhalb des zu behandelnden Datenstroms in der Tat sehr seltene Ereignisse. Wir können daher davon ausgehen, daß sich die unmittelbaren Nachbarn eines Datums *im Mittel* nur wenig vom gegebenen Datum unterscheiden.

Dasselbe gilt auch für Bilddaten: Falls das Bild digital hinreichend fein dargestellt wird, so daß keine Rastereffekte erkennbar sind, kommen große Sprünge in den Helligkeitswerten nur selten vor.

Bei diesem engen Zusammenhang zwischen benachbarten Werten setzen die gängigen Komprimierungsalgorithmen an: Wenn zwei Größen typischerweise sehr ähnlich sind, wird bei der Übertragung oder Speicherung *beider* Werte ein großer Teil der Information doppelt betrachtet; die Informationsdichte kann also deutlich erhöht werden, wenn man nur Informationen bertrachtet, die weitgehend unabhängig voneinander sind.

Bevor wir das in die Praxis umsetzen können, müssen wir zunächst mathematisch fassen, was die gegenseitige Abhängigkeit benachbartet Daten bedeutet. Aus dem letzten Paragraphen kennen wir schon ein Maß für die gegenseitige Abhängigkeit zweier Meßreihen, den Korrelationskoeffizienten; wir wollen etwas entsprechendes nun für eine beliebige Anzahl von Meßreihen oder Datenvektoren definieren.

Zur Klärung der Begriffe beginnen wir mit einem Beispiel, das zwar nicht das geringste mit Bild- oder Audiodaten zu tun hat, das dafür aber einfach und übersichtlich ist:

b) Kovarianzen und Korrelationen bei der Europawahl 1999

Wir betrachten das Ergebnis der Europawahl vom 13. Juni 1999 in den verschiedenen Mannheimer Stadtteilen. Laut Mannheimer Morgen vom 15. Juni gab es folgende Prozentzahlen:

Bezirk	Bet.	CDU	SPD	Grüne	Rep.	FDP	PDS	Sonst.
(A)	37,7	40,7	31,8	13,0	2,4	5,0	3,4	4,2
(B)	30,9	33,1	39,8	12,9	3,9	2,1	4,1	4,1
(D)	36,1	39,5	36,1	10,8	2,8	3,1	3,7	4,0
(D)	44,9	46,0	26,6	12,8	1,9	6,6	2,5	3,6
(E)	46,0	44,2	30,0	12,0	2,4	5,6	2,1	3,7
(F)	48,6	51,1	23,7	10,8	2,6	6,3	2,5	3,0
(G)	36,8	43,3	42,0	4,6	3,1	2,2	1,3	3,5
(H)	28,3	39,2	46,0	4,0	4,1	1,6	1,5	3,6
(I)	37,9	40,1	42,7	5,7	3,7	2,8	1,7	3,3
(J)	37,9	44,9	37,3	6,4	3,3	2,6	1,8	3,7
(K)	40,9	44,8	40,9	3,8	3,2	2,5	1,6	3,2
(L)	49,1	46,3	35,9	7,7	2,2	3,7	1,2	3,0
(M)	47,4	47,0	30,2	11,9	1,6	4,2	1,8	3,3
(N)	42,5	48,8	32,7	9,1	1,9	3,5	1,4	2,6
(O)	40,9	44,8	38,1	7,4	1,8	3,2	1,8	2,9
(P)	42,5	47,8	30,3	10,2	2,2	3,8	2,1	3,5
(Q)	38,4	47,6	34,9	6,0	3,4	2,9	1,2	4,0
MA	**39,8**	**44,2**	**34,6**	**9,0**	**2,7**	**3,7**	**2,1**	**3,7**

Die Stadtteile sind, damit alles in eine Zeile paßt, durch Buchstaben bezeichnet; für Interessenten seien auch die richtigen Namen angegeben: (A) steht für Innenstadt/Jungbusch, (B) für Neckarstadt West, (C) für Neckarstadt Ost/Wohlgelegen, (D) ist die Oststadt/Schwetzingerstadt, (E) der Lindenhof, (F) Neuostheim/Neuhermsheim, (G) Sandhofen, (H) Schönau, (I) Waldhof, (J) Käfertal, (K) Vogelstang, (L) Wallstadt, (M) Feudenheim, (N) Seckenheim, (O) Friedrichsfeld, (P) Neckarau, (Q) Rheinau und MA schließlich Mannheim insgesamt.

Wie man sieht, haben die einzelnen Parteien in den verschiedenen Stadtteilen recht unterschiedlich abgeschnitten; bei der FDP etwa ergibt sich mehr als ein Faktor vier zwischen den 1,6% in der Schönau und den

6,6% in Neuostheim/Neuhermsheim. Es ist auch klar, daß die CDU dort unterdurchschnittlich viele Stimmen hat, wo die SPD überdurchschnittlich abgeschnitten hat; anders geht das bei den Stimmenzahlen dieser beiden Parteien gar nicht. Wie steht es aber mit den kleinen Parteien?

Eine Partei hat in einem Stadtteil dann überdurchschnittlich gut abgeschnitten, wenn sie dort einen höheren Prozentsatz erreicht hat als für Mannheim insgesamt, wenn also die Differenz zwischen dem Prozentsatz dort und dem für ganz Mannheim positiv ist. Bei negativer Differenz hat sie unterdurchschnittlich abgeschnitten.

Wenn wir also vergleichen wollen, ob zwei Parteien ihre Hochburgen und ihre Schwächegebiete jeweils in denselben Stadtteilen haben, geben uns wie in §4 die Produkte ihrer Abweichungen von ihren Mittelwerten einen ersten Hinweis, d.h. also die Ausdrücke

$$\frac{1}{17}\sum_{i=1}^{17}(x_i-\widetilde{x})(y_i-\widetilde{y})\,,$$

wobei die x_i, y_i die Ergebnisse aus den einzelnen Stadtteilen sind und $\widetilde{x}, \widetilde{y}$ die Ergebnisse für ganz Mannheim.

Da die verschiedenen Stadtteile verschieden groß sind, sind $\widetilde{x}$ und $\widetilde{y}$ nicht genau die Mittelwerte der x_i beziehungsweise y_i, sondern aussagekräftigere gewichtete Mittelwerte; da andererseits Wahlbezirke ungefähr gleich groß sein sollten, wollen wir diesen Unterschied nicht allzu ernst nehmen und können obige Summe im wesentlichen mit der Kovarianz

$$\operatorname{cov}(\vec{x},\vec{y}) = \frac{1}{17}\sum_{i=1}^{17}(x_i-\overline{x})(y_i-\overline{y})$$

identifizieren, wobei

$$\overline{x} = \frac{1}{17}\sum_{i=1}^{17}x_i \quad \text{und} \quad \overline{y} = \frac{1}{17}\sum_{i=1}^{17}y_i$$

die Mittelwerte der beiden Datensätze sind.

Bei m Datensätzen $\vec{x}_1,\dots\vec{x}_m \in \mathbb{R}^N$ können wir die Kovarianzen $\operatorname{cov}(\vec{x}_i,\vec{x}_j)$ zu einer $m\times m$-Matrix zusammenfassen, der *Kovarianzmatrix*

$$\operatorname{Cov}(\vec{x}_1,\dots\vec{x}_m) = (c_iij) \quad \text{mit} \quad c_{ij} = \operatorname{cov}(\vec{x}_i,\vec{x}_j)\,.$$

Sie ist offensichtlich eine symmetrische Matrix.

Wenn wir die Wahlbeteiligung formal wie eine Partei namens W behandeln, erhalten wir hier die folgende Kovarianzmatrix:

	W	CDU	SPD	Grüne	Rep.	FDP	PDS	Sonst.
W	32,4	18,8	−24,3	5,66	−3,27	5,79	−1,32	−1,46
CDU	18,8	18,2	−15,7	−0,58	−1,96	3,00	−2,07	−1,08
SPD	−24,3	−15,7	34,9	−13,8	3,11	−7,66	−1,07	0,10
Grüne	5,66	−0,58	−13,8	10,2	−1,21	3,13	1,93	0,44
Rep.	−3,27	−1,96	3,11	−1,21	0,57	−0,69	0,05	0,14
FDP	5,79	3,00	−7,66	3,13	−0,69	2,04	0,26	−0,04
PDS	−1,32	−2,07	−1,07	1,93	0,05	0,26	0,73	0,22
Sonst.	−1,46	−1,08	0,10	0,44	0,14	−0,04	0,23	0,24

Übersichtlicher ist die *Korrelationsmatrix,* deren Einträge nicht die Kovarianzen sondern die Korrelationskoeffizienten sind; diese liegen bekanntlich stets zwischen -1 und 1. Im vorliegenden Fall ist das die Matrix

	W	CDU	SPD	Grüne	Rep.	FDP	PDS	Sonst.
W	1,00	0,77	−0,72	0,31	−0,76	0,71	−0,27	−0,52
CDU	0,77	1,00	−0,62	−0,04	−0,61	0,49	−0,57	−0,52
SPD	−0,72	−0,62	1,00	−0,73	0,70	−0,91	−0,21	0,03
Grüne	0,31	−0,04	−0,73	1,00	−0,50	0,69	0,71	0,28
Rep.	−0,76	−0,61	0,70	−0,50	1,00	−0,64	0,07	0,37
FDP	0,71	0,49	−0,91	0,69	−0,64	1,00	0,21	−0,05
PDS	−0,27	−0,56	−0,21	0,71	0,07	0,21	1,00	0,53
Sonst.	−0,52	−0,52	0,03	0,28	0,37	−0,05	0,53	1,00

Dieser Matrix können wir nun aller gewünschten Informationen entnehmen. Die erste Zeile beispielsweise sagt uns, daß die CDU, FDP und in geringerem Ausmaß auch die Grünen dort überdurchschnittlich gut abgeschnitten haben, wo auch die Wahlbeteilung überdurchschnittlich war, diese drei Parteien konnten also anscheinend ihre Anhänger am besten mobilisieren.

Wo die CDU überdurchschnittlich gut abschnitt, waren alle anderen Parteien mit Ausnahme der FDP unterdurchschnittlich; wo die SPD überdurchschnitt gut abschnitt, waren ebenfalls fast alle anderen unterdurchschnittlich; die Ausnahme hier sind die Repuplikaner. Zwischen SPD und FDP haben wir mit -0,91 den betragsgrößten Korrelationskoeffizienten; die Hochburgen dieser beiden Parteien sind also sehr unterschiedlich. Die Grünen schließlich sind positiv korreliert mit FDP, PDS und Sonstigen.

c) Zufallsvariablen und ihre statistischen Kenngrößen

Die Europawahl vom 13. Juni 1999 ist vorbei und ausgezählt, wir konnten daher im letzten Abschnitt mit wohlbekannten Daten rechnen. Beim Umgang mit Bild- und Tondaten gilt das nur in eingeschränktem Maße: Zwar können Daten erst dann verarbeitet werden, wenn sie da sind, aber eine Digitalkamera beispielsweise muß doch schon heute auf die Bilder vorbereitet sein, die ihr Besitzer erst in ein paar Jahren aufnehmen wird.

Für den Grundalgorithmus zur Datenkompression müssen wir daher Daten zulassen, über die wir noch nichts konkretes wissen – abgesehen von gewissen vagen Gesetzmäßigkeiten, durch die sich echte Bilddaten von beliebigen Matrizen unterscheiden. Als Hilfmittel dazu dient der Begriff der *Zufallsvariablen.*

Definition: Eine (diskrete) Zufallsvariablen ist ein Prozeß, der zufällig einen Wert aus einer vorgegebenen endlichen Menge

$$\{x_0, \ldots, x_m\}$$

liefert.

Dieser „Zufall“ muß natürlich, falls er mathematisch faßbar sein soll, irgendwelchen Regeln genügen, und hier kommt der zweite fundamentale Begriff zum Einsatz: Wir nehmen an, daß für jeden der möglichen Werte x_i feststeht, mit welcher *Wahrscheinlichkeit* p_i er angenommen

wird. Diese „Wahrscheinlichkeit“ definieren wir informell so, daß bei einer großen Anzahl m von Versuchen *ungefähr* $p_i m$-mal der Wert x_i geliefert wird. Das „Gesetz der großen Zahlen“, das wir im nächsten Semester kennenlernen werden, sagt, daß diese Definition sinnvoll ist und man die Wahrscheinlichkeiten p_i damit in wohldefinierter Weise mit beliebiger Genauigkeit bestimmen kann.

Zwei formale Konsequenzen der Definition sind offensichtlich:

$$0 \le p_i \le 1 \quad \text{für alle } i \text{ und} \quad \sum_{i=0}^{m} p_i = 1 \,.$$

Sollen beispielsweise alle x_i mit gleicher Wahrscheinlichkeit angenommen werden (was für Bilddaten nicht unbedingt ein sehr realistisches Modell ist), so sind alle $p_i = 1/(m+1)$.

In unserem Modell soll ein „Bild“ dann produziert werden durch eine Matrix (X_{ij}) von geeigneten Zufallsvariablen; eine „Audiosequenz“ entsprechend durch einen Vektor $(X_1, \ldots, X_N)$.

Welche Zufallsvariablen sind geeignet? In unserem einfachen Modell wollen wir uns nicht darauf festlegen, wie die p_i gewählt werden sollen, sondern stattdessen mit ihrer Hilfe Analoga zu den Kenngrößen aus dem vorigen Abschnitt definieren.

Definition: Der *Erwartungswert* $E(X)$ einer Zufallsvariablen X ist

$$E(X) = \sum_{i=0}^{m} p_i x_i \,.$$

Falls alle x_i mit gleicher Wahrscheinlichkeit angenommen werden, ist der Erwartungswert also einfach das arithmetische Mittel

$$\frac{1}{m+1} \sum_{i=0}^{m} x_i$$

der möglichen Werte; ansonsten ist er ein sogenanntes gewichtetes arithmetisches Mittel. Für einen Würfel etwa, der die Augenzahlen von $x_0 = 1$ bis $x_5 = 6$ mit jeweils gleicher Wahrscheinlichkeit produziert, ist

$$E(X) = \frac{1+2+3+4+5+6}{6} = 3\frac{1}{2} \,.$$

Als nächste wichtige Kenngröße hatten wir imletzten Abschnitt die mittlere quadratiscche Abweichung vom Mittelwert, die Varianz, betrachtet; in völliger Analogie zu dort definieren wir

Definition: Die Varianz einer Zufallsvariablen X mit Erwartungswert $E(X)$ ist

$$\sigma_X^2 = E\big((X - E(X))^2\big) = \sum_{i=0}^{m} p_i\big(x_i - E(X)\big)^2 ;$$

ihre Standardabweichung ist $\sigma_X = \sqrt{\sigma_X^2}$.

Beim Würfel wäre also

$$\sigma_X^2 = \frac{\left(-2\frac{1}{2}\right)^2 + \left(-1\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(1\frac{1}{2}\right)^2 + \left(2\frac{1}{2}\right)^2}{6} = \frac{35}{12}$$

und

$$\sigma_X = \sqrt{\frac{35}{12}} \approx 1{,}7078 .$$

d) Beispiele aus der Bildverarbeitung

Um die Bedeutung dieser Kenngrößen in der Bildverarbeitung zu veranschaulichen, sind auf der nächsten Doppelseit sechs beliebte Testbilder zusammen mit den Werten dieser Kenngrößen abgedruckt. Die Werte sind entnommen aus

P.M. FARELLE: Recursive Block Coding for Image Data Compression, *Springer,* 1990 ;

sie beziehen sich natürlich auf die Originalbilder und nicht auf das, was der Druckvorgang hier im Skriptum daraus gemacht hat. Trotzdem sollte der Vergleich von Bildern und Daten einen einigermaßen korrekten Eindruck zumindest der relativen Situation vermitteln, da hoffentlich alle hier abgedruckte Bilder in derselben Weise verunstaltet sind.

Die mittlere Helligkeit eines Bildes, dessen (viele) Pixel durch je eine Zufallsvariable mit Erwartungswert μ produziert werden, sollte ziemlich nahe bei μ liegen; der beste Schätzwert für den gemeinsamen Erwartungswert der Zufallsvariablen ist also die mittlere Helligkeit

des Bildes. Typischerweise werden Helligkeiten durch Zahlen zwischen 0 und 255 kodiert, wobei schwarz der Zahl Null entspricht und weiß der 255. Dies sieht man gut an den Beispielbildern, wo das mit Abstand hellste Bild „Tiffany" auch den mit Abstand größten Mittelwert μ hat; den kleinsten Wert hat das auch visuell dunkelste Bild „Lenna".

Die nächste wichtige Kenngröße ist die Varianz, welche angibt, wie stark eine Zufallsvariable um ihren Erwartungswert streut. Auch hier wollen wir wieder davon ausgehen, daß alle Zufallsvariablen zu einem gegebenen Bild *bzw.* einer gegebenen Audiosequenz aus N dieselbe Varianz haben. Wir schätzen diese gemeinsame Varianz aufgrund der vorliegenden Daten als

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

wobei $y_1, \ldots, y_N$ die Helligkeits- *bzw.* Lautstärkewerte sind. (Wer sich wundert, daß vor dieser Summe mit N Summanden nur $N-1$ im Nenner steht, sollte zu §4*d)* zurückblättern.)

Was die Varianz und die Standardabweichung bedeuten, sieht man wieder deutlich an den Beispielbildern: Bilder mit geringem Kontrast wie „Tiffany" oder „Lenna" haben deutlich geringere Werte als die kontrastreicheren Bilder „Peppers" und „Sailboat".

e) Kovarianz und Korrelation von Zufallsvariablen

So, wie wir sie bislang definiert haben, ist jede Zufallsvariable ein eigenständiger Prozeß, und zwei verschiedene Zufallsvariablen haben nichts miteinander zu tun. Das ist natürlich nicht das, was wir für die Beschreibung von Bild- und Audiodaten brauchen; hier müssen wir davon ausgehen, daß ein einziger Prozeß gleichzeitig einen ganzen Vektor *bzw.* eine ganze Matrix von Zufallswerten erzeugt, wobei deren einzelne Komponenten dann sehr wohl voneinander abhängig sein können.

Für zwei solche Komponenten X und Y mit jeweiligen Wertebereichen $\{x_0, \ldots, x_m\}$ und $\{y_0, \ldots, y_n\}$ sowie Wahrscheinlichkeiten p_i für x_i und q_j für y_j ist die Wahrscheinlichkeit dafür, daß X den Wert x_i liefert