Unterlagen zum Weiterbildungskurs "Stochastik" vom 24. Juni 2009

Hansruedi Künsch
Seminar für Statistik
ETH Zürich

Version vom 25. Juni 2009

1 Statistikunterricht am Gymnasium: Die Sicht eines Hochschullehrers

Jede Person mit Maturität sollte etwas über Wahrscheinlichkeit *und* Statistik gehört haben. Dies ist wichtig sowohl für die Allgemeinbildung als auch für die Vorbereitung des Studiums. Prinzipiell ist *Denkweise der Stochastik* wichtiger als spezifische Kenntnisse.

Etwas Erfahrung mit Daten gehört zum Statistikunterricht (Datengewinnung, Allgegenwart von Variabilität, Messung und Modellierung von Variabilität), und ich empfehle, unbedingt echte Daten und Beispiele mit Bezug zur Wirklichkeit verwenden. Hinweise auf mögliche Quellen gibt es im Abschnitt 4. Schön wäre es, wenn in einem interdisziplinären Unterricht Daten von einem andern Fach dargestellt und analysiert werden könnten. Der Einsatz von Informatikhilfsmitteln für deskriptive Statistik und für Simulationen ist wünschenswert.

Diskrete Modelle genügen, um die entscheidenden und interessanten Phänomene der Stochastik zu zeigen. Zufallsvariable mit Dichten würde ich höchstens am Rande bringen. Auch die Varianz, bzw. Standardabweichung, von Zufallsvariablen würde ich eher weglassen, weil die Regeln (z. B. die Additivität bei Unabhängigkeit) schwierig herzuleiten sind. Die Stichprobenvarianz sollte jedoch in der deskriptiven Statistik behandelt werden im Zusammenhang mit Streuungsmassen.

Bedingte Wahrscheinlichkeiten würde ich eher nicht formal behandeln, obwohl alle Wahrscheinlichkeiten eigentlich bedingt sind auf etwas und sie bei mehrstufigen Versuchen natürlich auftreten (Baumdarstellung !). Bedingte Wahrscheinlichkeiten sind auch ein unentbehrliches Konzept in der fortgeschrittenen Stochastik, aber das Verständnis von bedingten Wahrscheinlichkeiten ist meist sehr schwierig. Viele Paradoxa betreffen bedingte Wahrscheinlichkeiten.

Unabhängigkeit von Ereignissen (oder Zufallsvariablen) ist ein zentraler Begriff. Ich finde jedoch, dass ein intuitives Verständnis von Unabhängigkeit genügt. Wichtig wäre es, dass die Schüler sehen, dass Unabhängigkeit oft postuliert wird.

Testen von Hypothesen ist im Schulunterricht ein beliebtes Thema, aber ich bin der Meinung, dass Vertrauensintervalle mindestens ebenso gut geeignet sind. Vertrauensintervalle

sind informativer als Test, bei Ablehnung der Nullhypothese sieht man insbesondere, welche Alternativen dann in Frage kommen und kann so die Relevanz des Resultats besser beurteilen. Ausserdem muss man eigentlich weniger Terminologie einführen, als beim Testen (die Alternative und der Fehler zweiter Art treten nicht auf).

Die Behandlung von t-Test und Chiquadrat-Test empfehle ich eher nicht. Die Herleitung der Verteilung der Teststatistik ist zu schwierig für die Mittelschule, und dadurch ist die Argumentationskette nicht mehr völlig durchsichtig. Die Binomialverteilung eignet sich viel besser, weil die Teststatistik die Beobachtung selber ist.

Vorlesungen über Wahrscheinlichkeitsrechnung und Statistik gelten bei Studierenden als schwierig, verglichen z.B. mit Analysis, und ich frage mich oft, warum das so ist. Mögliche Ursachen sind, dass in der Stochastik die Verbindung von Mathematik und Wirklichkeit besonders schwierig ist und sich nicht ignorieren lässt, und dass wir von der Evolution her darauf trainiert sind, Regelmässigkeiten zu entdecken, weshalb wir mit der Variabilität des Zufalls schlecht umgehen können. Ein weiterer Faktor ist wohl auch die Zeit, die diesem Thema gewidmet ist. Im Gymnasium kommt die Stochastik meist nur kurz am Ende vor, und an der Hochschule ist es nicht viel besser.

Die Korrektur von falschen Vorstellungen über das Wirken des Zufalls könnte man als spannendes Unterrichtsziel formulieren. Eine unvollständige Liste von solchen Fehlvorstellungen ist

- Scheinbarer Widerspruch zwischen dem Gesetz der grossen Zahlen und der Gedächtnislosigkeit des Zufalls.
- Geburtstagsproblem, das heisst Unverständnis für den Unterschied zwischen "Es gibt zwei Personen mit gleichem Geburtstag" und "Es gibt eine 2. Person mit dem gleichen Geburtstag wie ich".
- Sukzessives Verdoppeln des Einsatzes beim Roulette, warum funktioniert das nicht ?
- Warum ist die Wahrscheinlichkeit, bei n Würfen eine Sechs zu haben, nicht $n \cdot \frac{1}{6}$?
- Beispiele von Kahneman und Tversky.

Da die Beispiele von Kahneman und Tversky (zwei Psychologen) nicht allgemein bekannt sind, gebe ich sie hier an: Man nimmt an, dass Knaben- und Mädchengeburten je Wahrscheinlichkeit $\frac{1}{2}$ haben. Im ersten Beispiel wird gefragt, welche der folgenden Sequenzen bei 6 Geburten wahrscheinlicher ist:
a) KMMKMK b) KKKKMK c) beide gleich wahrscheinlich.
Im 2. Beispiel wird gefragt, wie gross die Wahrscheinlichkeit ist, dass es bei 6 Kindern genau 3 Mädchen gibt:
a) $\frac{1}{2}$ b) $\frac{20}{64}$ c) ein anderer Wert.
Im 3. Beispiel wird gefragt, wo es im Verlauf eines Jahres mehr Tage gibt, an denen mindestens 60% der Neugeborenen Knaben sind:
a) in einem grossen Spital, b) in einem kleinen Spital, c) kein Unterschied.

Soweit die allgemeinen Überlegungen. Im Folgenden werde ich als Erstes eine Behandlung der Statistik der Binomialverteilung vorstellen, von der ich glaube, dass sie sich für das Grundlagenfach eignet und nicht viel Zeit braucht. Im zweiten Teil werden dann einige Aspekte der Regressionsrechnung diskutiert, die im Rahmen eines Ergänzungs- oder

Schwerpunktfachs behandelt werden könnten. Die Regression ist die am häufigsten verwendete Methode in der angewandten Statistik, was auch für SchülerInnen motivierend sein könnte. Insbesondere kann dies für die Auswertung von Daten in Maturaarbeiten nützlich sein. Schliesslich gibt es am Schluss noch eine Liste von Quellen und Materialien.

2 Statistik der Binomialverteilung

2.1 Ziele dieses Kapitels

Wir werden die folgenden typischen Fragestellungen behandeln

- In der Mendel'schen Theorie sollten in einem bestimmten Kreuzungsversuch von Erbsen 25% der Pflanzen grüne und 75% gelbe Keimblätter haben. In einem Versuch erhielt Mendel unter 8023 Pflanzen 2001 mit grünen Keimblättern. Spricht das für oder gegen Mendel's Theorie ?
- Eine Umfrage der New York Times gab bei 1154 befragten Erwachsenen 646 Befürworter, also einen Anteil von 56%. Wie viel anders käme das Ergebnis heraus, wenn man die Befragung mit andern Personen wiederholt? Kann man aus dem Ergebnis Schlüsse ziehen über den Anteil von Befürwortern in der ganzen Population erwachsener AmerikanerInnen ? Dazu steht in einem kleinen Kasten: "In theory, in 19 cases out of 20 the results based on such a sample (of 1154 adults) will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults". Wie kommt man zu so einer Aussage ?

2.2 Die Binomialverteilung

Die Binomialverteilung beschreibt die Anzahl "Erfolge" bei n unabhängigen Wiederholungen eines Experimentes mit zwei möglichen Ausgängen "Erfolg/Misserfolg" unter gleichen Bedingungen. Durch geeignete Interpretation eines "Erfolges" kann man diese Verteilung in sehr vielen Situationen anwenden, unter anderem auch bei obigen Beispielen:

- Mendel's Experiment: Jede Pflanze stellt eine Wiederholung dar (n=8023), und ein "Erfolg" ist eine mit grünen Keimblättern. Man hat also 2001 Erfolge beobachtet.
- Meinungsumfragen: Die Befragung jeder Person stellt eine Wiederholung dar, und ein "Erfolg" ist eine "Ja"-Antwort. Die Meinungsumfrage ergab 56% Befürworter, also etwa 646 Erfolge.

Ob die Voraussetzungen "gleiche Bedingungen" und "unabhängige Wiederholungen" erfüllt sind, sollte bei jeder Anwendung hinterfragt werden. Bei den Mendelschen Experimenten ist das einleuchtend. Für Meinungsumfragen holen wir es später nach.

Die Binomialverteilung tritt an erstaunlich vielen Stellen auf. Die folgende Liste enthält eine Reihe weiterer möglicher Beispiele:

- Geschlecht von Kälbern, siehe Artikel aus dem Tages-Anzeiger.
- Geschlecht von neugeborenen Menschen, siehe z.B. Engel, Abschnitt 12.3.5.

- Hellsehen, siehe z.B. Freedman et al., Kap. 26.5, oder Utts, Case Study 22.1.
- Über-/Untervertretung bestimmter Bevölkerungsgruppen in einem Beruf, Gremium etc..
- Rechnungsrevisionen, siehe z.B. Utts, Case Study 20.1, Moore, Kap. 1 (Examples 5 und 19).
- Schuhe mit Linksdrall, siehe Artikel aus dem Tages-Anzeiger.
- Analyse eines Versuchs zur Polio-Impfung, siehe Freedman et al., Kap. 1.1.
- Vorzeichentest beim gepaarten Vergleich zweier Methoden oder Behandlungen.
- Anzahl Wechsel beim Münzwurf.

Weniger geeignet sind zum Beispiel binary choice Prüfungen. Die Binomialverteilung gilt nur, wenn jemand absolut nichts weiss, und das ist (glücklicherweise) selten der Fall.

Beginnen wir jetzt mit der mathematischen Behandlung. Mit p bezeichnen wir die Wahrscheinlichkeit für einen Erfolg bei einer bestimmten Wiederholung. Dann gilt die folgende grundlegende Formel, die ich hier als bekannt voraussetze:

$$p_n(k) = P[\text{ genau } k \text{ Erfolge }] = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

Weniger bekannt ist folgender rekursiver Zusammenhang

$$p_n(k) = \frac{n-k+1}{k} \; \frac{p}{1-p} \quad p_n(k-1). \tag{2}$$

Dieser erlaubt nicht nur eine rasche numerische Berechnung der Wahrscheinlichkeiten $p_n(k)$ (man startet mit $p_n(0) = (1-p)^n$), sondern gibt auch Einblick in die Form der Verteilung. Der Quotient

$$\frac{p_n(k)}{p_n(k-1)} = \frac{n-k+1}{k} \; \frac{p}{1-p} \tag{3}$$

ist nämlich monoton fallend in k für festes p und n. Im Bereich $1 \le k \le (n+1)p$ ist dieser Quotient grösser oder gleich eins ist, und damit ist $p_n(k)$ dort monoton wachsend. Analog ist $p_n(k)$ monoton fallend im Bereich $(n+1)p \le k \le n$, weil der Quotient dort kleiner oder gleich eins ist. Auf beiden Seiten des Maximums fällt $p_n(k)$ sogar schneller als exponentiell ab (beim exponentiellen Abfall wäre der Quotient konstant).

Dieses Verhalten ist auch in der Abbildung 1 in der rechten Spalte ersichtlich. Wir stellen fest: **Die Wahrscheinlichkeiten $p_n(k)$ sind nur in einem relativ kleinen Bereich um $k = np$ herum wesentlich verschieden von Null** (klein relativ zur Anzahl $n+1$ der prinzipiell möglichen Werte, absolut gesehen wächst dieser Bereich mit n).

Um das etwas präziser zu fassen, definieren wir einen Teilbereich

$$A(n,p) = \{k_1(n,p), k_1(n,p)+1, \ldots k_2(n,p)\}$$

des Wertebereichs $\{0, 1, \ldots, n\}$ der Binomialverteilung mit der Eigenschaft, dass

$$P[\text{ Anzahl Erfolge liegt in } A(n,p)] \approx 0.95.$$

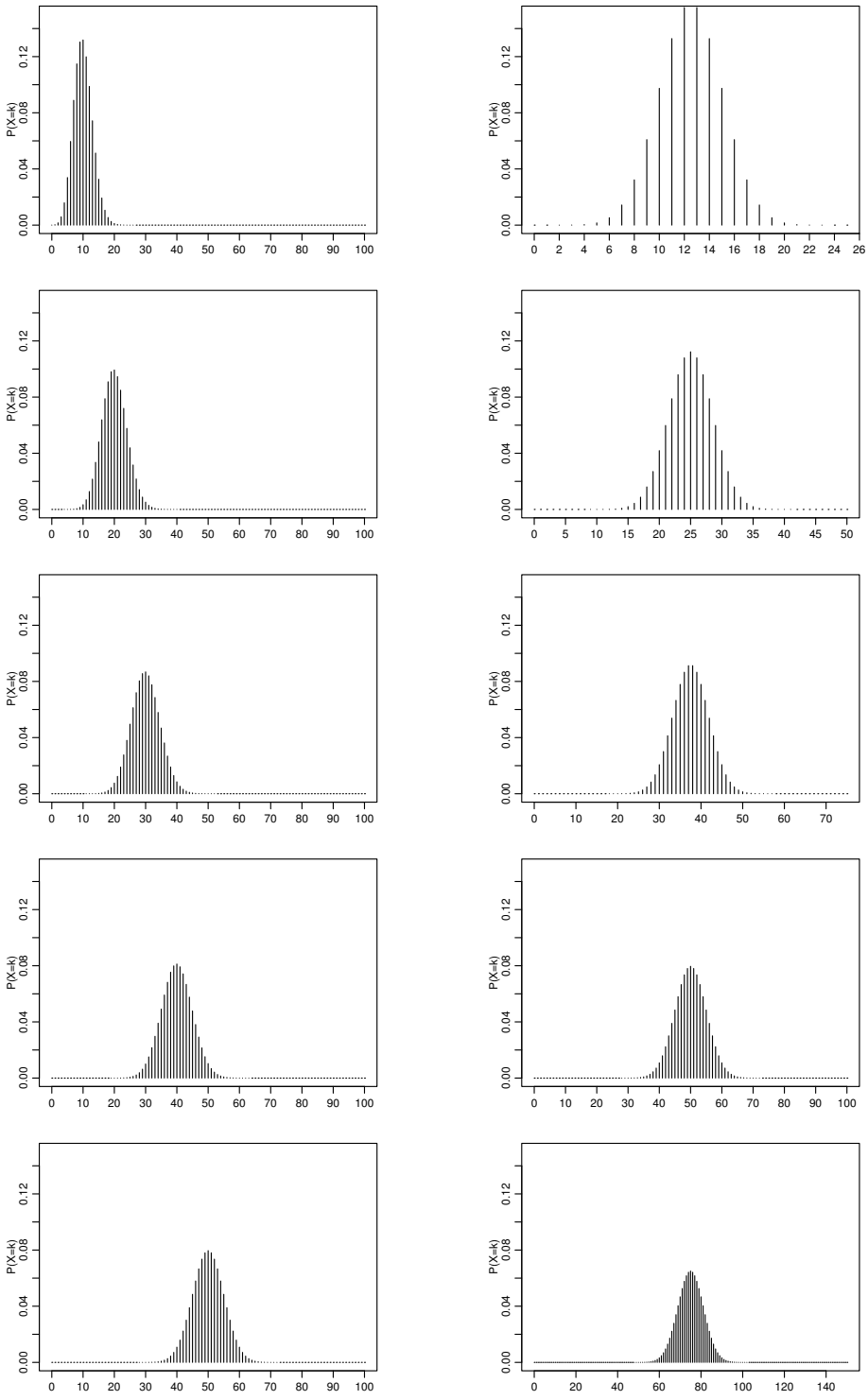

Abbildung 1: Die Binomialwahrscheinlichkeiten $P(X = k)$ als Funktion von k für verschiedene n's und p's. Links ist $n = 100$ und $p = 0.1, 0.2, 0.3, 0.4, 0.5$ und rechts ist $p = 0.5$ und $n = 25, 50, 75, 100, 150$.

Dazu schneiden wir bei der Binomialverteilung auf beiden Seiten je 2.5% Wahrscheinlichkeiten ab, bzw. etwas weniger, wenn wir die 2.5% nicht genau erreichen können:

$$\sum_{j=0}^{k_1(n,p)-1} p_n(j) \leq 0.025 < \sum_{j=0}^{k_1(n,p)} p_n(j) \tag{4}$$

bzw.

$$\sum_{j=k_2(n,p)+1}^{n} p_n(j) \leq 0.025 < \sum_{j=k_2(n,p)}^{n} p_n(j). \tag{5}$$

Aus Symmetriegründen (Vertauschung von Erfolg und Misserfolg) gilt
$k_2(n,p) = n - k_1(n, 1-p)$.

Wir nennen $A(n,p)$ den **Bereich von plausiblen Werten für die Anzahl Erfolge** k bei vorgegebenem n und p. Wie wir bereits empirisch festgestellt haben, ist $A(n,p)$ viel kleiner als der ganze Bereich. Weiter unten werden wir noch sehen, dass die Länge von $A(n,p)$ ungefähr gleich $const(p) \cdot \sqrt{n}$ ist.

Der Wert 0.95 beruht auf einer Konvention und könnte durch einen andern Wert nahe bei 1 ersetzt werden. Um die Notation einfach zu halten, arbeiten wir hier aber stets mit 0.95.

Im Beispiel von Mendel's Erbsenexperiment ist $n = 8023$ und $p = 0.25$. Man erhält mit obiger Rekursionsformel $A(n,p) = \{1930, \ldots, 2082\}$. Die 2001 beobachteten Erfolge liegen also klar innerhalb des plausiblen Bereichs. Das Experiment zeigt daher keinen Widerspruch zur Mendel'schen Theorie auf. Im Gegenteil, die Abweichungen sind eher zu klein. Dies ist auch bei allen andern Experimenten von Mendel der Fall; vermutlich wurden die Daten "geschönt" (siehe Freedman et al., Kap. 25.2.)

Was wir hier gemacht haben, ist nichts anderes als ein **zweiseitiger Test** der Nullhypothese $p = 0.25$ auf dem Niveau $\alpha = 0.05$. Dies kann geschehen, ohne viel Terminologie einzuführen. Trotzdem sieht man das Wesentliche.

2.3 Vom Test zum Vertrauensintervall

Bei der Meinungsumfrage nützen uns die bisherigen Überlegungen noch nicht viel, da wir p nicht kennen. Die Erfolgswahrscheinlichkeit p ist der Anteil Befürworter unter allen amerikanischen Erwachsenen, und den will man ja gerade mit der Umfrage herausfinden ! Die relative Häufigkeit 646/1154=0.56 ist zwar hoffentlich in der Nähe von p, aber mit Abweichungen muss man rechnen. Hätte man eine andere Stichprobe von 1154 Personen befragt, hätte man vermutlich eine leicht andere relative Häufigkeit bekommen. Das heisst, die relative Häufigkeit schwankt zufällig, während die Wahrscheinlichkeit fest, aber unbekannt ist. Das eine ist eine Schätzung, das andere ein Parameter, und man möchte etwas über den Unterschied sagen, ohne dass man den Parameter kennt.

Hier haben wir also ein Beispiel des typischen Umkehrschlusses der Statistik: Ausgehend von Beobachtungen will man etwas über die zu Grunde liegende Verteilung aussagen. Dies ist möglich, wenn wir annehmen, dass die beobachtete Anzahl von 646 Erfolgen im plausiblen Bereich $A(n,p)$ liegt. Dann können wir nämlich das unbekannte p einschränken auf

$$I(1154, 646) = \{p \mid A(1154,p) \text{ enthält } 646\} = \{p \mid k_1(1154,p) \leq 646 \leq k_2(1154,p)\}.$$

Die obere Grenze von $I(1154, 646)$ ist derjenige Wert p, bei dem $k_1(1154, p)$ von 646 nach 647 springt: Unterhalb dieses Werts ist nämlich $k_1(1154, p) \leq 646$, während oberhalb $k_1(1154, p) > 646$ ist. Analog liegt die untere Grenze dort, wo $k_2(1154, p)$ von 646 nach 645 springt. Damit können wir den gesuchten Bereich angeben:

$$I(1154, 646) = (0.5306, 0.5887),$$

was recht genau mit der von der New York Times angegebenen Unsicherheit von 3% übereinstimmt. Bevor wir diskutieren, wie man das konkret berechnet, machen wir noch einige grundsätzliche Betrachtungen.

Offensichtlich können wir die gleiche Überlegung für jeden beobachteten Wert x von Erfolgen durchführen. Wir erhalten so ein Intervall $I(n, x)$ von **plausiblen Werten für die unbekannte Wahrscheinlichkeit** p. Dieses Intervall heisst das **Vertrauensintervall** für p zum 95%-Niveau.

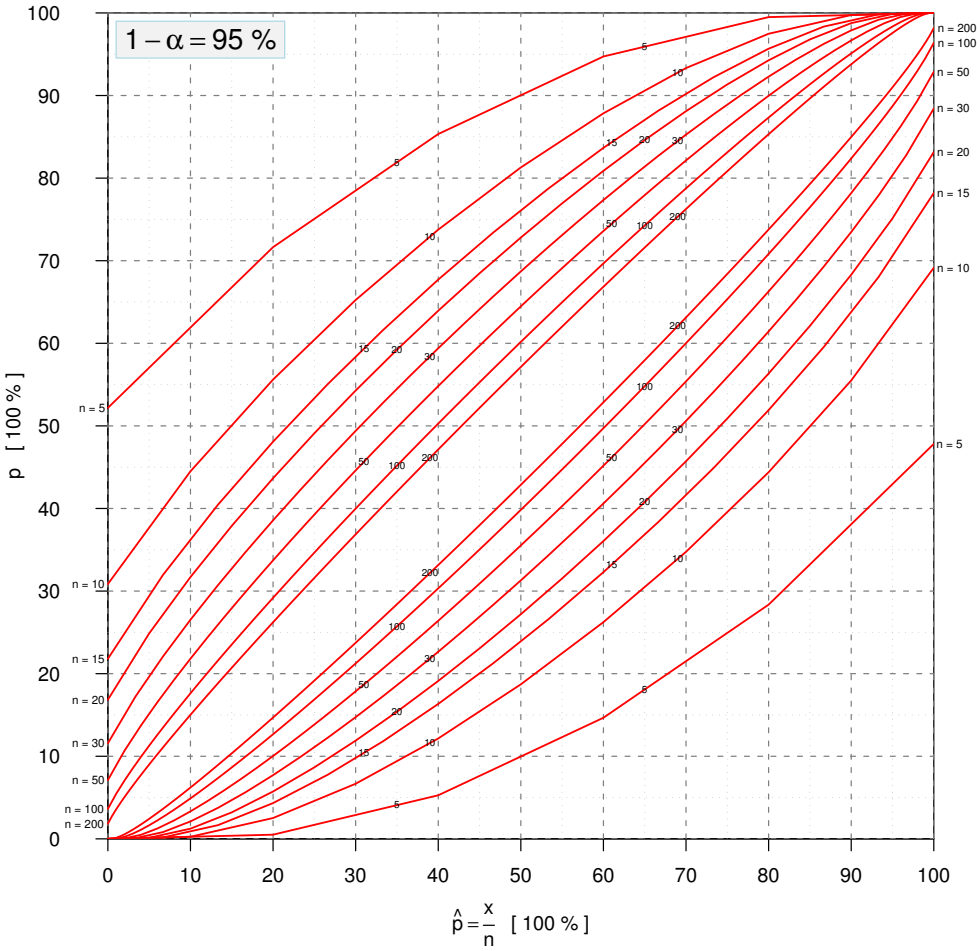

Abbildung 2: Grafik zur Bestimmung des Annahmebereichs, bzw. des Vertrauensintervalls bei der Binomialverteilung.

Der Gedankengang lässt sich grafisch leicht illustrieren. In der Abbildung 2 sind die Grenzen k_1, bzw. k_2 für verschiedene n's gegeben als Funktion von p (allerdings mit vertauschter Anordnung der Achsen). Schneidet man die beiden Kurven horizontal für ein festes p, erhält man nach Multiplikation mit n die beiden Grenzen von $A(n, p)$. Schneidet man

die beiden Kurven vertikal für eine beobachtete relative Häufigkeit $\frac{x}{n}$, so erhält man das Vertrauensintervall $I(n, x)$.

Wegen der Diskretheit der Binomialverteilung sind k_1 und k_2 genau genommen Treppenfunktionen (für festes n, mit p als Argument). In der Abbildung wurde der umhüllende Polygonzug gezeichnet.

Aus der Konstruktion ergibt sich auch ganz natürlich die **Interpretation des Vertrauensintervalls**. Wir können nämlich nicht garantieren, dass das Intervall $I(1154, 646)$ den unbekannten Anteil p der Befürworter enthält. Wenn zufällig die Beobachtung 646 aus dem Intervall $A(1154, p)$ herausgefallen ist, dann ist dies nicht der Fall. Dies geschieht aber höchstens mit Wahrscheinlichkeit 5%, und wir nehmen diese Irrtumswahrscheinlichkeit in Kauf. **Das zufällige Vertrauensintervall enthält die unbekannte Wahrscheinlichkeit p in etwa 19 von 20 Fällen**

Die Interpretation dieser Wahrscheinlichkeit ist etwas heikel. Zufällig ist das Vertrauensintervall und nicht der Parameter p. Man betrachtet explizit andere Stichproben, die man hätte ziehen können, aber nicht gezogen hat. Für viele Leute ist das etwas suspekt. Bei einer ersten Behandlung würde ich nicht zu stark auf diese Feinheit eingehen.

Nun aber zur Berechnung der Grenzen des Vertrauensintervall. Für die obere Grenze müssen wir denjenigen Wert p suchen, bei dem $k_1(n, p)$ von x nach $x + 1$ springt. Wegen der Formel (4) ist für $x < n$ die obere Grenze von $I(n, x)$ daher die Lösung der Gleichung

$$\sum_{j=0}^{x} p_n(j) = \sum_{j=0}^{x} \binom{n}{j} p^j (1-p)^{n-j} = 0.025 \tag{6}$$

bezüglich p im Intervall $(0, 1)$. Für $x = n$ ist die obere Grenze gleich eins. Analog ist die untere Grenze der Wert p, bei dem $k_2(n, p)$ von x nach $x - 1$ springt. Wegen der Formel (5) ist für $x > 0$ die untere Grenze von $I(n, x)$ daher die Lösung der Gleichung

$$\sum_{j=x}^{n} p_n(j) = \sum_{j=x}^{n} \binom{n}{j} p^j (1-p)^{n-j} = 0.025 \tag{7}$$

bezüglich p im Intervall $(0, 1)$. Für $x = 0$ ist die untere Grenze gleich null.

Es handelt sich bei (6) und (7) um polynomiale Gleichungen vom Grad n, welche keine explizite Lösungen haben. Auch Näherungen (numerisch, bzw. analytisch) erfordern etwas Aufwand. Im nächsten Abschnitt gehen wir darauf noch etwas näher ein.

Für kleines n gibt es Tabellen für das Vertrauensintervall. Damit kann man leicht mit einer Simulation illustrieren, dass das Vertrauensintervall das unbekannte p in etwa 19 von 20 Fällen enthält. Für $n = 40$ ist die Tabelle hier wiedergegeben.

2.4 Berechnung der Grenzen k_1, k_2, bzw. von Vertrauensintervallen

Exakte Berechnungen sind mit der Rekursionsformel (2) auch für n in der Grössenordnung von einigen Tausend ohne Weiteres möglich. Mit einem programmierbaren Taschenrechner kann man $k_1(n, p)$ und $k_2(n, p)$ mit dem folgenden Algorithmus berechnen:

- Runde np auf die naheliegendste ganze Zahl k_0.

0	1	2	3	4	5	6	7	8	9
0.088	0.132	0.169	0.204	0.237	0.268	0.298	0.328	0.356	0.385
10	11	12	13	14	15	16	17	18	19
0.412	0.439	0.465	0.491	0.517	0.542	0.567	0.591	0.615	0.639
20	21	22	23	24	25	26	27	28	29
0.662	0.685	0.707	0.730	0.751	0.773	0.794	0.814	0.834	0.854
30	31	32	33	34	35	36	37	38	39
0.873	0.892	0.910	0.927	0.943	0.958	0.972	0.984	0.994	0.999

Tabelle 1: Annahmebereich und Vertrauensintervall bei Binomialverteilung, $n = 40$. Für $x \in \{0, \ldots, 39\}$ ist die obere Grenze des Vertrauensintervalls $I(40, k)$ angegeben. Dies ist gleich dem Wert p, an dem $k_1(40, p)$ von x auf $x + 1$ springt. Die untere Grenze, bzw. die Stellen, wo $k_2(40, p)$ springt, erhält man aus Symmetrie.

- Berechne $p_n(k_0)$ aus der Rekursionsformel, unter Verwendung von Logarithmen, um Unterfluss zu vermeiden:

$$\ln(p_n(k_0)) = n \ln(1-p) + \sum_{j=1}^{k_0} \ln(\frac{n-j+1}{j}) + k_0 \ln(\frac{p}{1-p}).$$

  Setze $S = p_n(k_0)$ und $j = 0$.

- Solange $S < 0.95$, erhöhe j um 1, berechne $p_n(k_0 \pm j)$ aus $p(n, k_0 \pm (j-1))$ mit der Rekursionsformel und setze $S = S + p_n(k_0 + j) + p_n(k_0 - j)$.

- Bei Abbruch ist $k_1(n, p) = k_0 - j$ und $k_2(n, p) = k_0 + j$.

Dies macht man für $p = 0.05,\ 0.10, \ldots 0.95$ (in einer Klasse kann man dies gut als Team durchführen). Durch Interpolation erhält man ein ähnliches Bild wie in Abbildung 2, und daraus kann man dann das Vertrauensintervall genähert ablesen.

Viele Programmpakete und Taschenrechner haben die sogenannte kumulative Verteilungsfunktion

$$F_n(x) = \sum_{j=0}^{x} p_n(j)$$

fest eingebaut. Dann findet man $k_1(n, p)$ durch Probieren, vgl. Formel (4). Die Grenzen des Vertrauensintervalls erhält man durch Auflösen von $F_n(x) = 0.025$, bzw. $F_n(x-1) = 0.975$ nach p (vgl. Formeln (6) und (7)), was iterativ gemacht werden kann mit der Bisektion. Einige Programmpakete geben auch direkt $k_1(n, p)$ als das sogenannte 2.5%-Quantil der Binomialverteilung.

Für sehr grosse n's (z.B. $n \geq 1000$) muss man die Normalapproximation der Binomialverteilung zu Hilfe ziehen. Programmpakete wechseln von der exakten zur asymptotischen Berechnung, ohne dass dies der Benutzer merkt. Wenn man diese Approximation nicht behandelt hat, kann man auch empirisch argumentieren. Auf eine der oben beschriebenen Arten kann man z.B. Abbildung 3 bzw.4 erzeugen. Diese zeigt, dass bereits für relativ kleine n's in guter Näherung gilt

- $A(n, p)$ ist ungefähr symmetrisch um np.

- $np - k_1(n,p)$ ist am grössten für $p = 0.5$.
- $np - k_1(n,p) \approx const.\sqrt{n}$, und für $p = 0.5$ ist $const. \approx 1$.

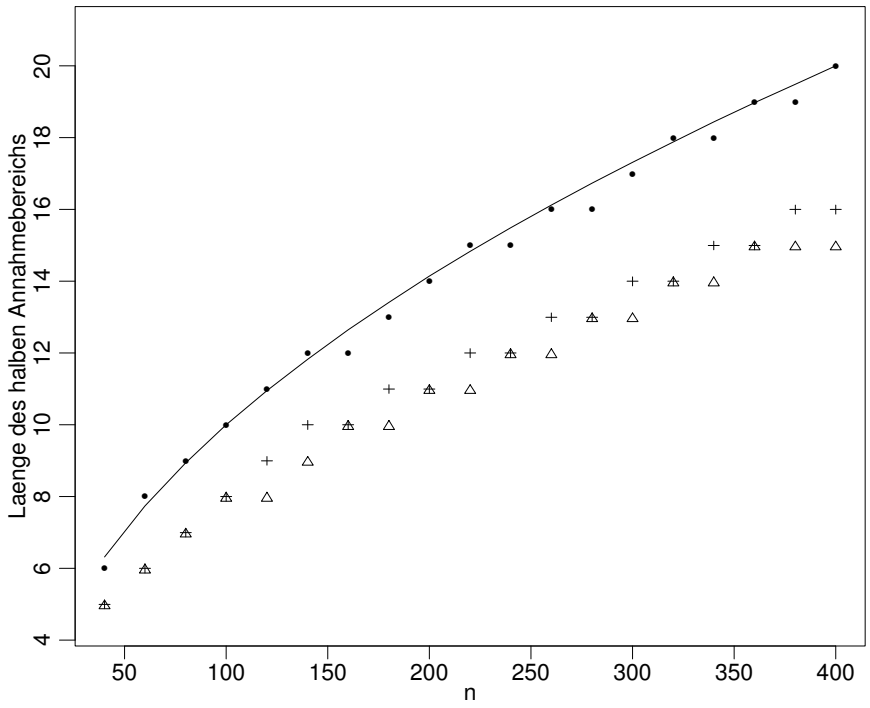

Abbildung 3: Länge des linken halben Annahmebereichs $np - c_1(n,p)$ bei der Binomialverteilung in Abhängigkeit von n. Kreise gehören zu $p = 0.5$, Kreuze zu $p = 0.8$, Dreiecke zu $p = 0.2$. Zusätzlich ist $\sqrt{n}$ eingetragen.

Aus der Normalapproximation weiss man, dass

$$const = 1.96 \cdot \sqrt{p(1-p)} \leq 1.96 \cdot \sqrt{0.5 \cdot 0.5} \approx 1.$$

Daraus folgt, dass die "Linse" von Abbildung 2 enthalten ist im Parallelstreifen um $k = np \pm \sqrt{n}$. Also gilt für das Vertrauensintervall

$$I(n,k) \subseteq \left[\frac{k}{n} - \frac{1}{\sqrt{n}}, \frac{k}{n} + \frac{1}{\sqrt{n}}\right]. \tag{8}$$

Insbesondere ist $1/\sqrt{1154} = 0.03$, was die am Anfang angegebene Genauigkeit erklärt. Man kann also im Voraus sagen, wie gross die Stichprobe sein muss, bevor man die Anzahl Befürworter kennt.

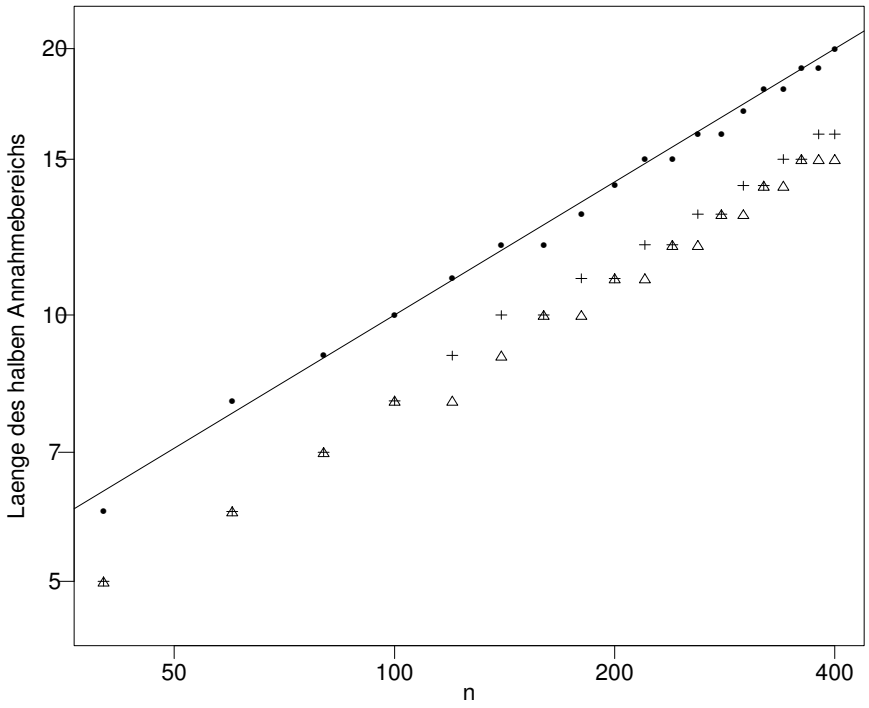

Abbildung 4: Gleiche Darstellung wie in Abbildung 3, aber jetzt in logarithmischer Skala auf beiden Achsen.

2.5 Bemerkungen zu Meinungsumfragen

Für mehr Details siehe z.B. Moore, Kap. 1.

Ist die Anzahl Befürworter tatsächlich binomialverteilt ? In sehr guter Näherung ja, *falls die befragten Personen zufällig ausgewählt wurden.*

Sei N die Grösse der Population, K die Anzahl Befürworter in der Population und n die Grösse der Stichprobe. Falls die Stichprobe durch Ziehen mit Zurücklegen ausgewählt wird, ist die Anzahl Befürworter sogar exakt binomialverteilt mit $p = K/N$.

Da man aber die gleiche Person nicht zweimal befragen will, verwendet man stets Ziehen ohne Zurücklegen. Dadurch werden die Ergebnisse der einzelnen Befragungen ganz leicht abhängig, und statt der Binomial- müsste man die hypergeometrische Verteilung verwenden.

Intuitiv vermutet man, dass die Unterschiede klein sind, wenn n klein gegenüber N ist. Dies ist tatsächlich richtig (Faustregeln geben als Bedingung $n < 0.1\ N$ an). Eine strenge mathematische Begründung ist aber nicht ganz einfach.

Intuitiv sollte es einleuchten, dass man beim Ziehen mit Zurücklegen eine *grössere* Variabilität hat als beim Ziehen ohne Zurücklegen (wenn eine Person zweimal vorkommt, zählt

sie entweder zweimal als Befürworter oder zweimal als Gegner). Die Vertrauensintervalle, die auf der Binomialverteilung beruhen, sind also eher zu gross als zu klein. Aus all diesen Gründen genügt es, mit der Binomialverteilung zu arbeiten.

Das Ganze hat aber eine Konsequenz, die den meisten Leuten gar nicht einleuchtet: Die Länge des Vertrauensintervalls ist nur abhängig von n und nicht von N! Das heisst also, dass eine Stichprobengrösse von $n \approx 1000$ sowohl in der Schweiz als auch in den USA genügend ist, um das Resultat mit einem Fehler von 3 Prozent zu erhalten. Man muss nicht einen bestimmten Prozentsatz der Population befragen.

Warum stimmen dann aber so viele Meinungsumfragen nicht ? Es ist gar nicht einfach, eine Zufallsauswahl zu realisieren. Dazu bräuchte man eine Liste der ganzen Population, und nach der Auswahl ist es sehr aufwendig, die Befragungen durchzuführen. Einige ausgewählte Personen kann man gar nicht erreichen, oder sie weigern sich, zu antworten.

Häufig begnügt man sich daher mit andern Auswahlverfahren, z.B. mit sogenannten Quotenstichproben, bei denen der Interviewer innerhalb gewisser Randbedingungen die befragten Personen frei auswählen kann. Die Randbedingungen sind Quoten bezüglich Alter, Geschlecht, Wohnort, etc.. Bei diesen Stichproben ist es aber schwierig, den Stichprobenfehler abzuschätzen, und es gibt berühmte Beispiele, wo sie versagten.

Unbedingt zu vermeiden sind sogenannte Willkürstichproben, wo man die am leichtesten zu erreichenden Personen befragt oder diejenigen, die sich selber melden. Solche Umfragen sind völlig wertlos, auch wenn sie auf sehr vielen Antworten beruhen.

Weiter sind auch der Effekt der Fragestellung, der Einfluss des Interviewers und absichtliche Falschantworten nicht zu vernachlässigen. All dies führt zu systematischen Fehlern. In der Praxis sind Meinungsumfragen ein dauernder Kampf gegen solche systematischen Fehler.

Sind Meinungsumfragen sinnvoll ? Sicher nicht alle. Sie bilden aber eine Möglichkeit, wie die Bevölkerung die Politik beeinflussen kann, auch in Situationen, wo keine Abstimmungsmöglichkeiten gegeben sind.

Meinungsumfragen als Projekte oder Maturarbeiten an der Mittelschule ? Ich rate eher davon ab, weil der Aufwand ausserordentlich gross ist (Formulierung der Fragen, Auswahl der zu Befragenden, Durchführung der Befragung, Eingabe und Kontrolle der Antworten), und weil die statistische Auswertung von komplexen Fragebögen schwierig ist (Gruppenbildung von ähnlichen Antwortprofilen, Finden von wesentlichen Einflussgrössen).

3 Einführung in die Regression

3.1 Die Methode der Kleinsten Quadrate

3.1.1 Einführung und Beispiele

In vielen Anwendungen hat man Beobachtungen oder Messwerte $((x_i, y_i); i = 1, 2, \ldots, n)$ von zwei Grössen X und Y, und man interessiert sich für den Zusammenhang untereinander. Stellt man die Werte als Punkte in der Ebene dar, so liegen sie genähert auf einer Geraden, d.h. der Zusammenhang ist ungefähr linear.

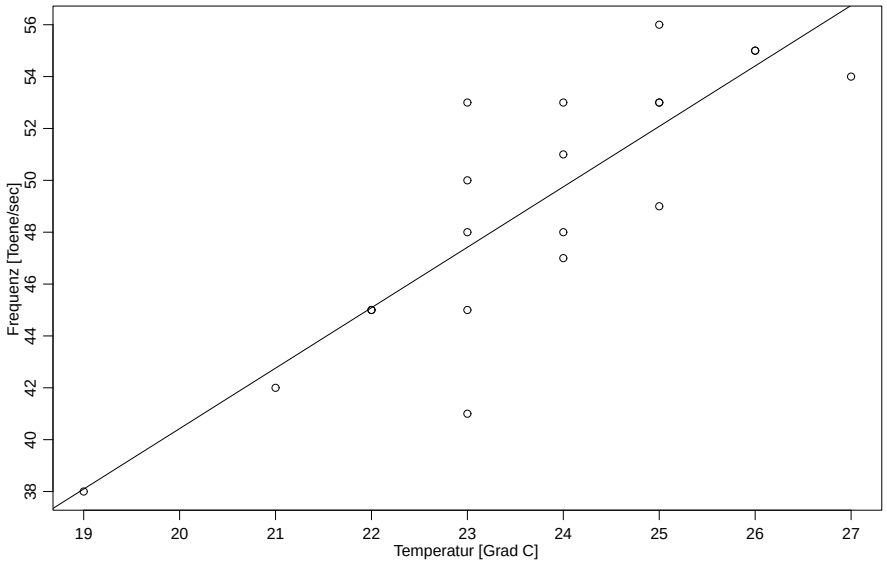

Abbildung 5: Frequenz beim Balzruf und Umgebungstemperatur bei 20 Fröschen der Art Hyla chrisoscelis

Wir geben hier einige Beispiele. Frösche sind Kaltblütler und daher ist ihre Aktivität stark von der Umgebungstemperatur abhängig. In einer biologischen Studie wurde untersucht, wie die Temperatur X die Frequenz Y im Balzruf der Froschart Hyla chrisoscelis beeinflusst. Für 20 Frösche in der natürlichen Umgebung wurde die Frequenz y_i (Anzahl Töne pro Sekunde) und die herrschende Lufttemperatur x_i (in Grad Celsius) gemessen $(i = 1, 2, \ldots, 20)$. Die Daten und eine approximierende Gerade sind in Abbildung 5 dargestellt.

Eine Methode zur Bestimmung des Gehaltes von Lebensmittelfarbstoffen ist Hochdruckchromatografie. In der Untersuchung der Qualität dieser Messmethode wurde bei 21 Proben mit unterschiedlichem Gehalt x_i des Farbstoffs FD&C Yellow No. 5 (in Prozent) die Fläche unter dem Gipfel im Chromatogramm (in cm^2) gemessen. Diese Daten sind in Abbildung 6 dargestellt.

Der schottische Physiker James David Forbes (1809-1868) hat den Zusammenhang zwischen Luftdruck und Siedepunkt des Wassers untersucht. Seine Daten sind in Abbildung 7 dargestellt.

Weitere Beispiele sind der Zusammenhang zwischen der Erschütterung Y und der Distanz X zum Sprengort beim Bau eines Tunnels, der Zusammenhang zwischen der Inflationsrate

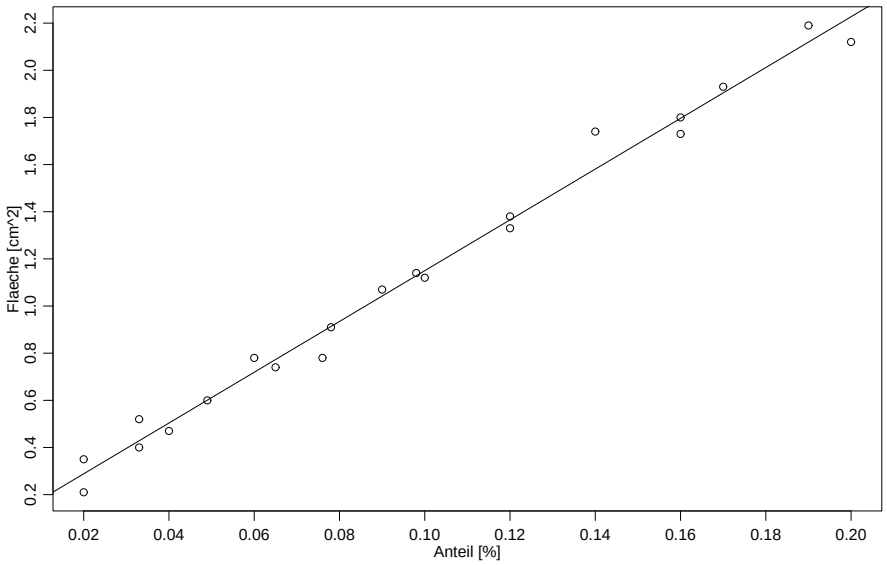

Abbildung 6: Anteil des Lebensmittelfarbstoffs FD&C Yellow No. 5 und Fläche unter dem Gipfel im Chromatogramm bei 21 Proben

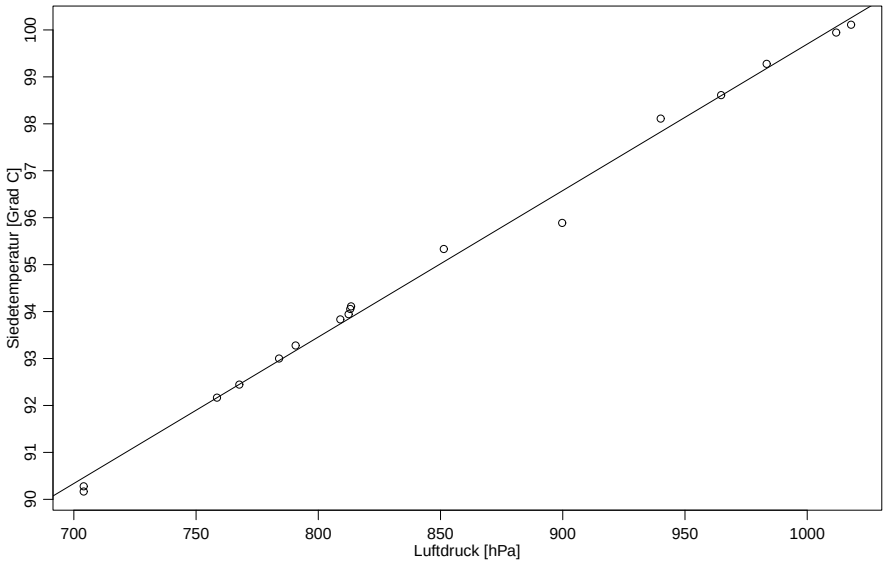

Abbildung 7: Siedepunkt von Wasser in Abhängigkeit vom Luftdruck

X und der Rendite von Staatsanleihen Y, oder der Zusammenhang zwischen psychischem Stress durch sich verändernde Lebensumstände (gemessen auf Grund eines Fragebogens) und der Schwere von chronischen Krankheiten. Man findet in der Literatur Studien aus praktisch allen Gebieten, von der Biologie und Medizin über Chemie und Physik bis zu Ökonomie und Psychologie. Häufig hat man mehr als zwei Grössen; die hier besprochenen Verfahren lassen sich verallgemeinern, wobei aber zusätzliche Phänomene und Schwierigkeiten auftreten.

Von Auge kann man auch eine gut passende Gerade durch die Punktwolke legen. Die Methode der Kleinsten Quadrate ist das einfachste (aber nicht das einzig mögliche) reproduzierbare Verfahren, eine solche Gerade zu bestimmen. Die Steigung b und der Achsenab-

schnitt a der Geraden werden so bestimmt, dass die Summe der quadrierten Abweichungen in y-Richtung

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

minimal wird. Wenn man Funktionen von mehreren Variablen und partielle Ableitungen kennt, kann man leicht die Bedingungen herleiten, welche die optimalen Werte a_0 und b_0 erfüllen müssen:

$$\sum_{i=1}^{n}(y_i - a_0 - b_0x_i) = 0, \quad \sum_{i=1}^{n} x_i(y_i - a_0 - b_0x_i) = 0.$$

Die Lösung dieser Gleichungen lautet

$$a_0 = \bar{y} - b_0\bar{x}, \quad b_0 = \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i(x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

Dabei bezeichnen $\bar{x}$, bzw. $\bar{y}$ die arithmetischen Mittel der x_i bzw. y_i. Das zweite Gleichheitszeichen folgt wegen

$$\sum_i (x_i - \bar{x}) = \sum_i (y_i - \bar{y}) = 0.$$

Der Nenner von b_0 ist nichts anderes als $n-1$ mal die Stichproben-Varianz der x_i

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

Im Folgenden werden wir eine Herleitung geben, welche nur Funktionen einer Variablen benutzen.

3.1.2 Anpassen einer horizontalen Geraden

Wenn die Steigung b null ist, besteht kein Zusammenhang: Die Werte x_i werden ignoriert. Man sucht dann einfach eine Konstante, welche die n Werte y_i möglichst gut approximiert. Minimieren von

$$\sum_{i=1}^{n}(y_i - a)^2$$

gibt als Lösung das arithmetische Mittel

$$a_0 = \bar{y}.$$

In andern Worten: Das arithmetische Mittel ist die beste Approximation einer Stichprobe bezüglich des Kriteriums der Kleinsten Quadrate. Dieses Ergebnis erklärt auch teilweise den Nenner $n-1$ in der Formel für die Stichprobenvarianz

$$s(y)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Der Zähler wäre grösser, wenn wir statt $\bar{y}$ den unbekannten Erwartungswert $E(Y)$ einsetzen würden. Diese systematische Unterschätzung im Zähler wird nun durch den Nenner $n-1$ kompensiert. Dass dies gerade der richtige Faktor ist, muss natürlich bewiesen werden. Wir verzichten darauf, bemerken aber, dass es unmittelbar einleuchtet, dass der Nenner bei $n = 1$ Null ist: Aus einer Beobachtung erhält man keine Angabe über die Abweichungen.

3.1.3 Anpassen einer Geraden durch den Ursprung

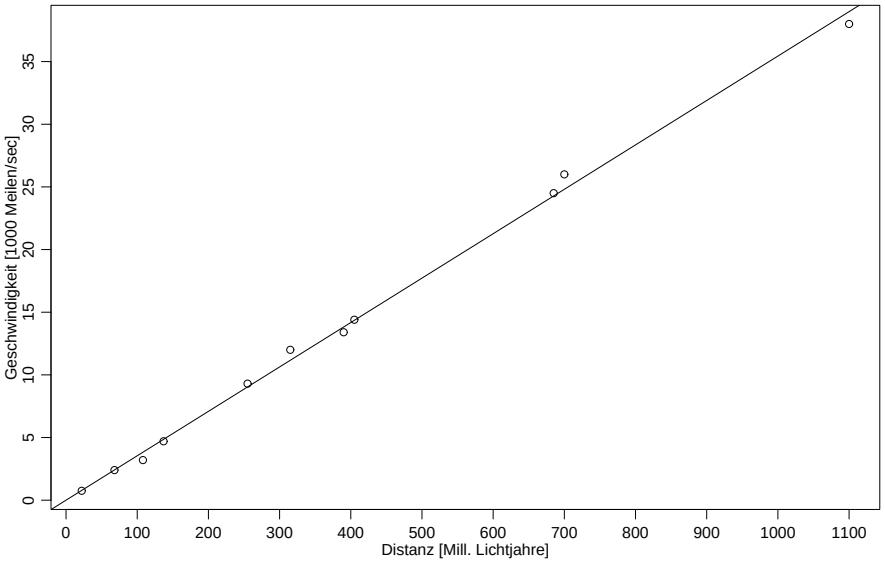

Abbildung 8: Distanz und Entfernungsgeschwindigkeit von Galaxien

Wenn der Achsenabschnitt a_0 gleich null ist, hat man eine Gerade durch den Ursprung. Das ist in gewissen Anwendungen sinnvoll, wo alle x_i und y_i positiv sind und man weiss, dass für $x = 0$ auch $y = 0$ sein muss. Als Beispiel betrachten wir die Entdeckung von Hubble aus dem Jahr 1929, dass sich das Universum nicht im Gleichgewicht befindet, sondern sich ausdehnt. Diese Entdeckung beruht auf der Untersuchung des Zusammenhangs zwischen der Geschwindigkeit V, mit der sich eine Galaxie von der unsern entfernt, und der Distanz D dieser Galaxie von der unsern. Hubble's Gesetz besagt, dass

$$V = H \cdot D$$

wobei H die sogenannte Hubble-Konstante ist. Daten von 11 Galaxienhaufen sind in Abbildung 8 dargestellt.

Wir verwenden wieder die Standardnotation mit Y statt V, X statt D und b statt H. Minimieren von

$$\sum_{i=1}^{n}(y_i - bx_i)^2$$

bezüglich b ergibt

$$b_0 = \frac{\sum_i y_i x_i}{\sum x_i^2}.$$

3.1.4 Die Herleitung der Kleinste Quadrate Gerade

Für beliebige, aber feste Steigung b erhält man analog wie oben den optimalen Achsenabschnitt: Wir setzen $z_i = y_i - bx_i$ und minimieren

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}(z_i - a)^2$$

bezüglich a. Dies ergibt

$$a_0 = a_0(b) = \bar{z} = \bar{y} - b\bar{x}.$$

Um die optimale Steigung zu berechnen, müssen wir also

$$\sum_{i=1}^{n}(y_i - a_0(b) - bx_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y} - b(x_i - \bar{x}))^2$$

minimieren, was auf

$$b_0 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

führt.

3.1.5 Andere Kriterien zur Anpassung von Geraden

Die Methode der Kleinsten Quadrate behandelt die beiden Variablen nicht gleichwertig, da nur die Abstände in y-Richtung betrachtet werden. In vielen Anwendungen sind die beiden Grössen X und Y aber nicht symmetrisch. Häufig liegt ein Kausalzusammenhang von X nach Y vor, oder wir wollen Y mit Hilfe von X vorhersagen. Deshalb bezeichnen wir Y auch häufig als Zielgrösse und X als erklärende Grösse.

Eine best-passende Gerade würde anschaulich eher so bestimmt, dass die quadrierten Abstände der Punkte von der Geraden, senkrecht zur Geraden gemessen, möglichst klein würden. Man nennt diese Methode orthogonale Regression. Die beiden Methoden beruhen auf unterschiedlichen Modellvorstellungen, wie die Abweichungen von einer Geraden zustandekommen.

Die Verwendung der quadrierten Abstände ist ebenfalls nicht zwingend. Man kann zum Beispiel auch die Absolutbeträge statt der Quadrate verwenden, d.h. man minimiert

$$\sum_{i=1}^{n} |y_i - a - bx_i|,$$

was historisch älter ist als die Methode der Kleinsten Quadrate. Für die Lösung gibt es keine einfache Formel, aber dafür schnelle Algorithmen. Das qualitative Verhalten der beiden Lösungen ist ziemlich unterschiedlich, vor allem, wenn einzelne Punkte stark von der Geraden abweichen. Im Fall der Anpassung einer Konstanten $b = 0$ werden wir das in einer Übungsaufgabe sehen.

3.2 Die Grundideen der schliessenden Statistik in der Regression

3.2.1 Das lineare Modell

Wir gehen nun einen Schritt weiter mit der Interpretation der Kleinsten Quadrate Geraden. Wenn wir akzeptieren, dass alle Ergebnisse von Experimenten und Beobachtungen "auch ein bisschen anders hätten herauskommen können", dann stellen sich sofort weitere Fragen

- Wie genau ist die Steigung der Geraden bestimmt ?
- Können wir die Regressionsgerade benutzen um vorauszusagen, was der Wert der Zielvariable bei gegebenem Wert der erklärenden Variable ist ?

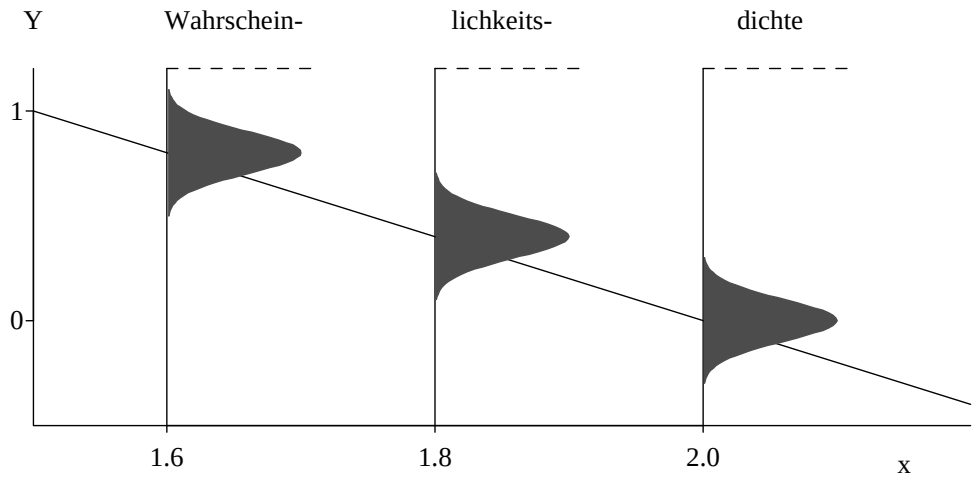

Abbildung 9: Veranschaulichung des Wahrscheinlichkeitsmodells $Y_i = 4 - 2x_i + E_i$ für drei Beobachtungen Y_1, Y_2 und Y_3 zu den x-Werten $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$.

- Können wir entscheiden, ob der Zusammenhang zwischen den Variablen tatsächlich linear ist ?

Beim Hubble'schen Gesetz ist die Hubble-Konstante (d.h. die Steigung) gleich dem Kehrwert des Alters des Universums. Damit ist es offensichtlich zentral anzugeben, wie genau man diese Konstante aus den Daten bestimmen kann. Im Beispiel des Balzrufes von Fröschen ist die Antwort auf die erste Frage wichtig, wenn man verschiedene Arten vergleichen will bezüglich der Sensitivität auf die Umgebungstemperatur. Im Beispiel der Messung des Lebensmittelfarbstoffs will man umgekehrt aus der Fläche auf den unbekannten Anteil zurückschliessen. Dafür ist es wichtig zu wissen, wie genau die Gerade bestimmt ist. Im Beispiel der Sprengungen möchte man den Wert der Erschütterung für einen vorgegebene Distanz vorhersagen. Die Frage der Linearität stellt sich zum Beispiel beim Zusammenhang zwischen Luftdruck und Siedetemperatur.

Die Antwort auf diese Fragen gibt die Schliessende oder Analytische Statistik, die auf der Wahrscheinlichkeitsrechnung beruht. Um sie zu verstehen, müssen wir zunächst eine Modellvorstellung entwickeln, die sagt, welche anderen Datensätze ebenso gut möglich gewesen wären wie die tatsächlichen Daten. In andern Worten, wir überlegen uns ein Wahrscheinlichkeitsmodell, das beschreibt, wie Daten zustande kommen.

Das Modell, das der Regression zu Grunde liegt, besagt, dass es eine unbekannte wahre Beziehung

$$y = \alpha + \beta x$$

gibt und dass bei gegebenem Wert x_i der erklärenden Variablen der Wert y_i der Zielgrösse zufällig um den Funktionswert $\alpha + \beta x_i$ herum streut:

$$y_i = \alpha + \beta x_i + E_i.$$

Die Verteilung der Abweichungen E_i soll zudem die gleiche sein für alle Werte x_i. Das Ganze kann man sich so vorstellen, dass der Experimentator den Wert x_i wählt, während die "Natur" aus einer Urne von möglichen Abweichungen einen Wert E_i zieht und die Summe $\alpha + \beta x_i + E_i$ als Ergebnis y_i liefert.

Eine Modell-Vorstellung entsteht in unseren Köpfen. Das Modell wird erst dann konkret, wenn wir die drei Zahlen α, β und eine Verteilung für die E_i festlegen. Abbildung 9 veranschaulicht das Modell der linearen Regression mit den Parameter-Werten $\alpha = 4$, $\beta = -2$ und einer Normalverteilung mit Erwartungswert Null und Standardabweichung 0.1 für die E_i. Die Wahrscheinlichkeit en, mit denen bestimmte Werte für die Y-Variable erwartet werden, sind mit den Wahrscheinlichkeitsdichten dargestellt.

3.2.2 Die Verteilung der Kleinste-Quadrate Schätzungen

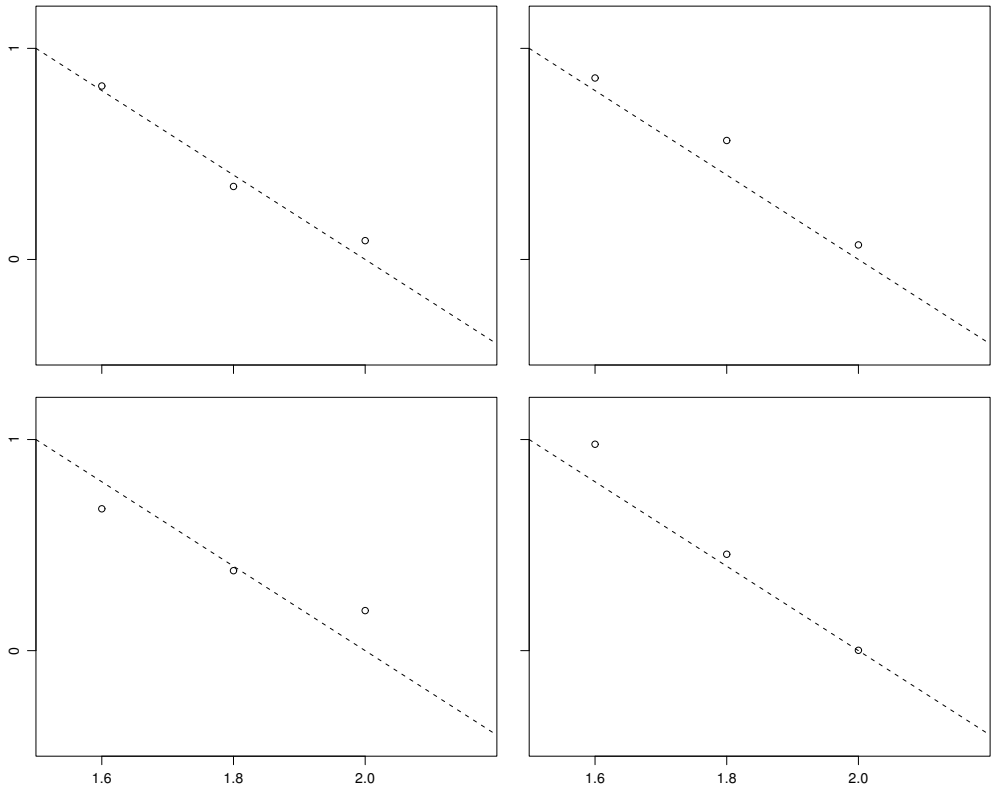

Abbildung 10: Vier simulierte Ergebnisse für drei Messungen gemäss dem Modell $y_i = 4 - 2x_i + E_i$ (gestrichelte Geraden)

Als zweite Veranschaulichung wollen wir Zufallszahlen gemäss unserem Modell ziehen und darstellen. Drei $\mathcal{N}(0, 0.1^2)$-verteilte Zufallszahlen bilden ein mögliches Ergebnis für die drei zufälligen Abweichungen E_1, E_2 und E_3. Ein Zufallszahlen-Generator lieferte die vier Dreiergruppen

$$(-0.0419, -0.1536, -0.0671), \qquad (0.0253, -0.0587, -0.0065),$$
$$(0.1287, 0.1623, -0.1442), \qquad (-0.0417, 0.1427, 0.0897).$$

Wenn $4 - 2x_i$ mit $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$ dazugezählt werden, erhält man je die entsprechenden Werte für y_1, y_2 und y_3. In Abbildung 10 sind die so "simulierten" Ergebnisse dargestellt. Jede der vier Figuren stellt also einen möglichen Versuchen mit je einer Messung an den Stellen $x = 1.6, 1.8, 2.0$ dar.

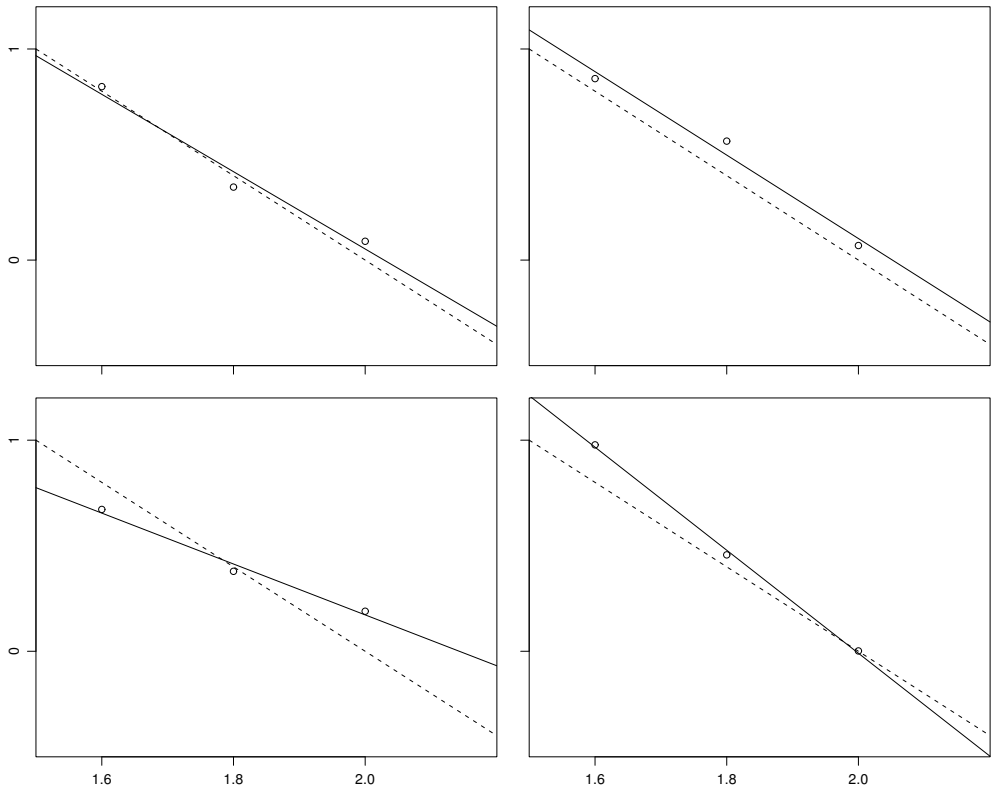

Abbildung 11: Vier simulierte Ergebnisse für drei Messungen mit den geschätzten (ausgezogenen) und den wahren (gestrichelten) Geraden.

Wenn wir für jeden der vier möglichen Versuche die Steigung und den Achsenabschnitt mit Hilfe der Kleinsten-Quadrate-Methode berechnen, dann erhalten wir nicht die wahren Werte 4, bzw. -2, sondern jedesmal einen andern Wert, nämlich

$$a_0 = 4.03, b_0 = -2.06, \qquad a_0 = 4.12, b_0 = -2.08,$$
$$a_0 = 5.28, b_0 = -2.68, \qquad a_0 = 3.47, b_0 = -1.67.$$

Dies ist in Abbildung 11 illustriert, in der jeweils die zu den Punkten aus Abbildung 10 best-passenden Geraden eingezeichnet sind. Die so erhaltenen Geraden streuen um die "wahre" Gerade. Die berechneten Werte a_0 und b_0 sind also Schätzungen (Näherungen) für die "wahren" Werte α, bzw. β. In der Statistik bevorzugt man deshalb die Notation $\widehat{\alpha}$ und $\widehat{\beta}$ an Stelle von a_0 und b_0. Wir bleiben hier jedoch bei den Bezeichnungen a_0 und b_0.

Der Grund für diese unterschiedlichen Ergebnisse ist leicht einzusehen: Die Werte y_i gehen linear in die durch die Kleinste Quadrate Methode bestimmte Steigung a_0 , bzw. den Achsenabschnitt b_0 ein. Weil die y_i zufällig streuen, ist dies daher auch bei den berechneten Parametern a_0 und b_0 der Fall. Sie sind also wie die Abweichungen E_i auch wieder Zufallsvariablen. Es stellt sich daher die Frage nach der Verteilung von a_0 und b_0. Diese Verteilung kann mit Hilfe der Wahrscheinlichkeitstheorie bestimmt werden. Dies ist jedoch zu aufwändig für den Gymnasialunterricht. Anschaulicher ist es, wenn wir Modell-Experimente betrachten. Dazu werden Zufallszahlen gemäss dem Modell gezogen analog dem Beispiel in Abbildung 10. Dann werden die Steigung und der Achsenabschnitt für diese simulierten

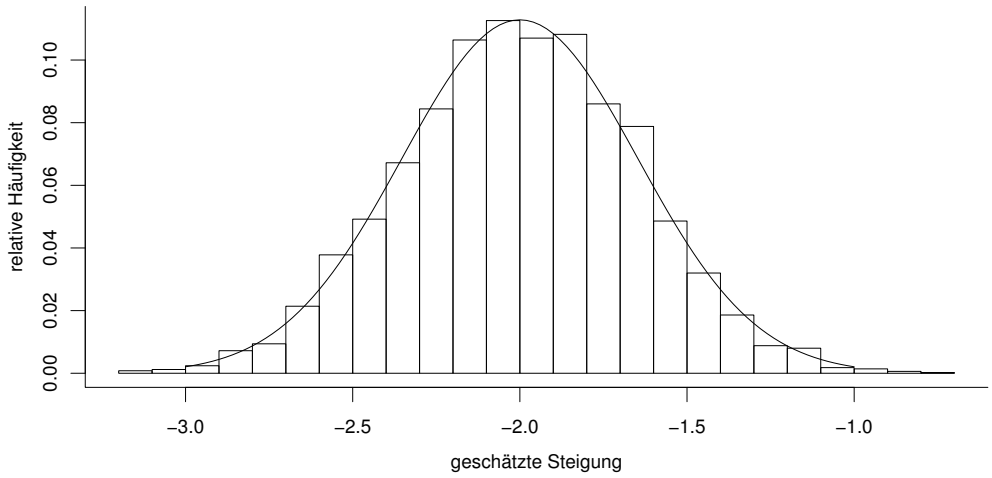

Abbildung 12: Simulierte und theoretische Verteilung der geschätzten Steigung im Fall $\alpha = 4$, $\beta = -2$ und $\sigma(E) = 0.1$

Beobachtungen mit der Kleinste Quadrate Methode geschätzt. Dieses Vorgehen wird nun m mal wiederholt und wir erhalten somit m Schätzwerte für die Parameter α und β. In Abbildung 12 sind 1000 Schätzwerte der Steigung β in einem Histogramm zusammengefasst.

Wie gesagt, die Verteilungen der Schätzungen lassen sich mit Hilfe der Wahrscheinlichkeitsrechnung direkt aus den Annahmen über die Verteilung der Abweichungen E_i bestimmen. Wenn wir annehmen, dass diese unabhängig und $\mathcal{N}(0, \sigma(E)^2)$-verteilt sind, dann kann man zeigen, dass die Kleinste-Quadrate-Schätzungen a_0 und b_0 ebenfalls normalverteilt sind, nämlich

$$b_0 \sim \mathcal{N}(\beta, \sigma(b_0)^2) \text{ und } a_0 \sim \mathcal{N}(\alpha, \sigma(a_0)^2) ,$$

mit

$$\sigma(b_0)^2 = \frac{\sigma(E)^2}{(n-1)s(x)^2} \quad \sigma(a_0)^2 = \sigma(E)^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s(x)^2} \right).$$

Wenn die Messfehler nicht normalverteilt sind, sind die Kleinste-Quadrate-Schätzungen a_0 und b_0 wenigstens noch genähert normalverteilt mit den gleichen Erwartungswerten und den gleichen Varianzen.

3.2.3 Das Vertrauensintervall für die unbekannte Steigung β

Auf den ersten Blick scheinen die obigen Resultate für die Verteilung von a_0 und b_0 nicht besonders nützlich, weil wir weder die wahren Werte α und β noch die Varianz $\sigma(E)^2$ der Abweichungen kennen. Dies ist aber nicht der Fall, wie wir gleich sehen werden. Zunächst nehmen wir an, dass wir wenigstens $\sigma(E)^2$ kennen; später werden wir zeigen, dass dies auch umgangen werden kann.

Eine normalverteilte Zufallsvariable liegt bekanntlich mit Wahrscheinlichkeit 95% zwischen dem Erwartungswert plus/minus 1.96 mal die Standardabweichung. Damit gilt mit Wahrscheinlichkeit 95%

$$\beta - 1.96 \frac{\sigma(E)}{\sqrt{n-1}\ \ s(x)} \leq b_0 \leq \beta + 1.96 \frac{\sigma(E)}{\sqrt{n-1}\ \ s(x)}.$$

$f = n-2$	1	2	3	4	5	6	7	8	9	10	
$q_f(0.975)$	12.71	4.30	3.18	2.78	2.57	2.45	2.37	2.31	2.26	2.23	
$f = n-2$	11	12	13	14	15	20	30	40	50	100	∞
$q_f(0.975)$	2.20	2.18	2.16	2.15	2.13	2.09	2.04	2.02	2.01	1.98	1.96

Tabelle 2: Quantile der t-Verteilung mit f Freiheitsgraden

Lösen wir diese Ungleichungen auf nach β, so sehen wir, dass das Intervall

$$b_0 \pm 1.96 \frac{\sigma(E)}{\sqrt{n-1}\ \ s(x)}$$

die unbekannte wahre Steigung β mit Wahrscheinlichkeit 95% "einfängt". Mit andern Worten, wir haben ein Vertrauensintervall für β gefunden!

Wenn $\sigma(E)^2$ nicht bekannt ist, dann gehen wir wie folgt vor: Wir schätzen $\sigma(E)^2$ aus den Daten durch

$$s(E)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a_0 - b_0 x_i)^2$$

Die Grössen $y_i - a_0 - b_0 x_i$ geben die Abweichungen von der angepassten Gerade in y-Richtung an und heissen auch die Residuen. Es sind Approximationen für die Abweichungen $E_i = y_i - \alpha - \beta x_i$ von der wahren Geraden, die wir nicht kennen.

Wir berücksichtigen die zusätzliche Unsicherheit, die dadurch entsteht, dass wir die Residuen an Stelle der E_i verwenden, indem wir die Konstante 1.96 ersetzen durch eine andere, leicht grössere Konstante. Da die zusätzliche Unsicherheit abnimmt, wenn man mehr Beobachtungen hat, muss die Konstante von n abhängen. Man kann zeigen, dass die richtige Konstante gegeben ist durch das 97.5%-Quantil der t-Verteilung mit $n-2$ Freiheitsgraden. Einige Werte sind in Tabelle 2 gegeben:

Zur Erklärung des Nenners gelten die analogen Bemerkungen wie im Abschnitt 3.1.2 zur Stichprobenvarianz.

3.3 Die Rolle von Transformationen

Die Annahme eines linearen Zusammenhangs ist nicht so restriktiv, wie es zuerst aussieht. Mit Hilfe von Transformationen der x_i und/oder der y_i lassen sich viele Beziehungen linearisieren. Die wichtigsten Beispiele sind der Potenzzusammenhang

$$y = \alpha x^\beta \Leftrightarrow \log y = \log \alpha + \beta \log x$$

und der Exponentialzusammenhang

$$y = \alpha e^{\beta x} \Leftrightarrow \log y = \log \alpha + \beta x.$$

Die Idee, einen Potenzzusammenhang durch Logarithmieren zu linearisieren, hatten wir bereits in Abbildung 4 verwendet.

Es spielt jedoch eine Rolle, ob man die Annahme von additiven Abweichungen mit konstanter Streuung vor oder nach der Transformation macht, z.B.

$$\log y_i = \log \alpha + \beta \log x_i + E_i \Leftrightarrow y_i = \alpha x_i^b \cdot \exp(E_i)$$

Das heisst additive Fehler auf der logarithmischen Skala werden zu multiplikativen Fehler in der ursprünglichen Skala. Insbesondere ist damit auf der ursprünglichen Skala die Streuung der Abweichungen proportional zu den Y-Werten. In vielen Anwendungen zeigt es sich zum Glück, dass Transformationen, welche den Zusammenhang linearisieren, häufig auch die Annahme von additiven Abweichungen gleicher Streuung plausibler machen.

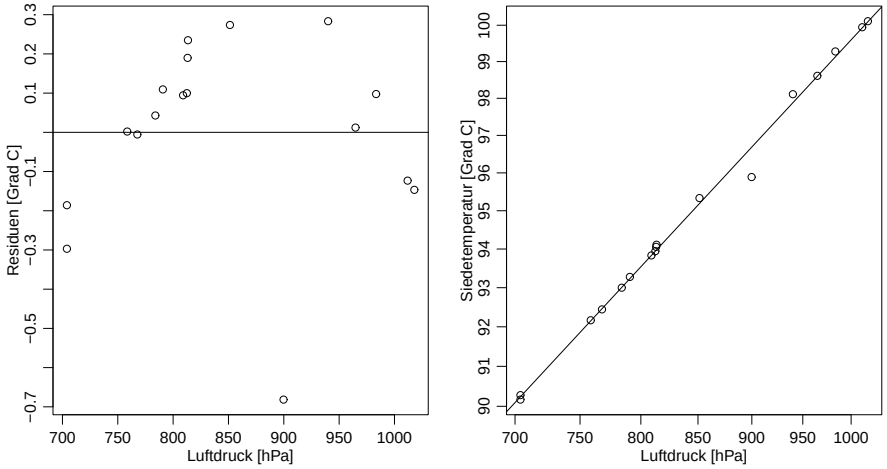

Abbildung 13: Siedetemperatur in Abhängigkeit des Luftdrucks: Residuen gegen erklärende Variable ohne Transformation (links). Anpassung einer Geraden nach Logarithmus-Transformation beider Variablen

Betrachten wir die Siedetemperatur des Wassers in Abhängigkeit des Luftdrucks. Um genauer zu sehen, ob die Beziehung linear ist, empfiehlt es sich, die Abweichungen von der angepassten Geraden als Funktion von X darzustellen: Das heisst, man zeichnet die Residuen $y_i - a_0 - b_0 x_i$ gegen x_i. Die Abbildung 13 links gibt Evidenz auf eine Krümmung bei den Daten zu Siedepunkt und Luftdruck, welche in Abbildung 7 schwieriger zu entdecken ist. Abbildung 13 rechts zeigt die Situation, wenn man beide Variablen logarithmiert. Ein Wert fällt nun deutlich heraus, und bei den übrigen Beobachtungen ist die Übereinstimmung eher besser.

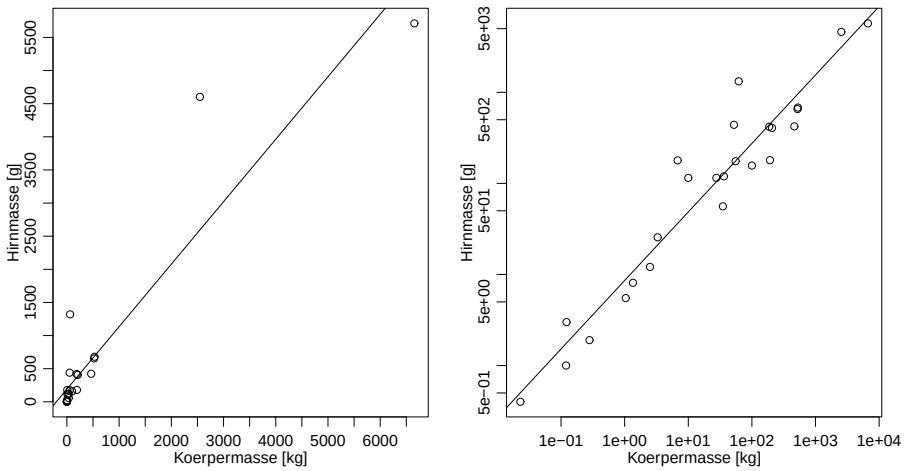

Abbildung 14: Körpermasse und Hirnmasse von 25 Tierarten. Unrsprüngliche Skala (links) und nach Logarithmus-Transformation beider Variablen (rechts).

Ein extremeres Beispiel erhält man, wenn man die Hirnmasse von Säugetierarten in Abhängigkeit von der Körpermasse betrachtet, siehe Abbildung 14. Die ausgewählten Tierarten sind Biber, Kuh, Wolf, Ziege, Meerschweinchen, Asiatischer Elefant, Esel, Pferd, Husarenaffe, Katze, Giraffe, Gorilla, Mensch, Afrikanischer Elefant, Rhesusaffe, Känguru, Goldhamster, Maus, Kaninchen, Schaf, Jaguar, Schimpanse, Ratte, Maulwurf, Schwein.

3.4 Regression als "Rückschritt zum Mittel"

Regression heisst ja übersetzt "Rückschritt", und es erscheint mysteriös, weshalb gerade dieser Name gewählt wurde. Der Grund dafür ist eine einfache mathematische Tatsache, die wir jetzt herleiten. Auf Grund der Formeln für a_0 und b_0 folgt

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - a_0 - b_0 x_i)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} - b_0(x_i - \bar{x}))^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + b_0^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2b_0 \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
&= (n-1)s(y)^2 + \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right)^2}{(n-1)s(x)^2} - 2\frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right)^2}{(n-1)s(x)^2} \\
&= (n-1)s(y)^2 \left(1 - \frac{\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right)^2}{s(y)^2 s(x)^2}\right)
\end{aligned}$$

Am Schluss taucht das Quadrat der Stichprobenkorrelation

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{s(x)s(y)}$$

auf. Weil die linke Seite nicht negativ ist, haben wir also bewiesen, dass $|r_{xy}| \leq 1$ gilt, wobei der Wert ± 1 genau dann auftritt, wenn alle Punkte auf einer Geraden liegen. Ferner sehen wir, dass $1 - r_{xy}^2$ gleich dem Verhältnis der Summe der quadratischen Abweichungen von der best-passendsten Gerade zur Summe der Abweichungen von der best-passendsten horizontalen Gerade ist. In andern Worten: Je grösser der Absolutbetrag von r_{xy} ist, desto stärker werden die Schwankungen der Y-Werte reduziert durch die Berücksichtigung der X-Werte. Die Stichprobenkorrelation misst also die Stärke und Richtung des linearen Zusammenhangs.

Mit Hilfe der Stichprobenkorrelation können wir die Regressionsgerade auch schreiben als

$$\frac{y - \bar{y}}{s(y)} = r_{xy} \cdot \frac{x - \bar{x}}{s(x)}.$$

Wenn wir die Regressionsgerade benutzen, um zu gegebenen x den Wert y zu berechnen, dann taucht das Phänomen des Rückschritts zum Mittel auf: Wenn r_{xy} positiv ist und x z.B. eine Standardabweichung $s(x)$ über dem Mittel $\bar{x}$ liegt, dann liegt der vorhergesagte Wert y nicht eine Standardabweichung $s(y)$ über dem Mittel $\bar{y}$, sondern weniger, denn r_{xy} ist ja kleiner als Eins (ausser wenn der Zusammenhang perfekt ist). Ein analoger Effekt tritt bei Werten x unterhalb von $\bar{x}$ auf. Wenn also X die Mathematiknote eines Schülers in der 5. Klasse bezeichnet und Y die Mathematiknote des gleichen Schülers in der 6. Klasse, dann liegen diejenigen, die in der 5.Klasse über dem Durchschnitt sind, tendenziell

in der 6. Klasse zwar noch immer über dem Mittel, sind aber eher schlechter, während diejenigen, die in der 5. Klasse unter dem Durchschnitt sind, in der 6. Klasse tendenziell besser abschneiden.

Dieser Rückschritt zum Mittel wurde mehrmals unabhängig voneinander in verschiedenen Wissenschaftszweigen entdeckt und meist kulturpessimistisch als ein bedauerliches Hindernis zum allgemeinen Fortschritt interpretiert. Vertauscht man die Rolle von X und Y, so verwandelt sich jedoch der Rückschritt in einen Fortschritt: Die Schüler, deren Note in der 6. Klasse über dem Durchschnitt liegt, haben sich tendenziell verbessert gegenüber ihrer Note in der 5. Klasse. Mit andern Worten: Der Rückschritt zum Mittel hat keine tiefere Bedeutung, sondern reflektiert einfach die Tatsache, dass es bei Unsicherheit besser ist, mit dem Durchschnitt als mit einer grossen Abweichung zu rechnen.

4 Materialien zum Stochastikunterricht

4.1 Bücher für ein interessiertes allgemeines Publikum

- Ivar Ekeland, Zufall, Glück und Chaos : mathematische Expeditionen, München, Hanser, 1992
- Darrel Huff, How to Lie With Statistics, W. W. Norton & Company, New York, 1954 und 1982
- Walter Krämer, So lügt man mit Statistik, Campus Verlag, Frankfurt, 1991
- Walter Krämer, Statistik verstehen; Eine Gebrauchsanweisung, Campus Verlag, Frankfurt, 1992
- Walter Krämer, So überzeugt man mit Statistik, Campus Verlag, Frankfurt, 1994
- Mike Orkin, What are the odds ? Chance in everyday life. Freeman, New York, 2000.
- Mike Orkin, Can you win ? The real odds for casino gambling, sports betting and lotteries. Freeman, New York, 1991.
- John Allen Paulos, Zahlenblind : mathematisches Analphabetentum und seine Konsequenzen, München, Heyne, 1993
- C. Radhakrishna Rao, Was ist Zufall? Statistik und Wahrheit, Prentice Hall, München, 1995
- Hans Riedwyl, Schweizer Zahlenlotto: Spiel, Zufall und Gewinn, Haupt, Bern, 1979
- David Ruelle, Zufall und Chaos, Springer Verlag, 1991
- Gero von Randow, Das Ziegenproblem; Denken in Wahrscheinlichkeiten, Rowolt Taschenbuch Verlag, 1992
- Judith M. Tanur, Frederick Mosteller, William H. Kruskal, Erich L. Lehmann, Richard F. Link, Richard S. Pieters and Gerald R. Rising, Statistics: A Guide to the Unknown, Wadsworth and Brooks Cole, 1989
- Edward R. Tufte, The Visual Display of Quantitative Information, Graphics Press, Cheshire, 1983
- Edward R. Tufte, Envisioning Information, Graphics Press, Cheshire, 1990
- Edward R. Tufte, Visual Explanations, Graphics Press, Cheshire, 1990

4.2 Lehrbücher Statistik

Bücher auf Mittelschulstufe. Ich habe auf diesem Gebiet keinen Überblick. Die Liste ist deshalb höchst unvollständig

- Arthur Engel, Stochastik, Klett 1987. Leider vergriffen.
- Heinz Klaus Strick, Einführung in die beurteilende Statistik. Schroedel 2008.

Bücher auf einfachem mathematischen Niveau, mit guter Diskussion der Konzepte und interessanten Beispielen:

- David Freedman, Robert Pisani and Roger Purves, Statistics, 4. Auflage, W. W. Norton & Company, New York, 2007.
- Richard De Veaux, Paul Velleman und David Bock, Stats: Data and Models. Addison Wesley 2004.
- Larry Gonick and Woollcott Smith, The Cartoon Guide to Statistics, Harper Perennial, HarperCollins Publishers, NY, 1993.
- David S. Moore and William I. Notz. Statistics: Concepts and Controversies, 6. Auflage, Freeman, NY, 2006.
- Jessica M. Utts, Seeing Through Statistics, 3rd edition, Brooks/Cole, 2005.

Etwas schwieriger, eher für Lehrpersonen:

- Richard J. Larsen and Morris L. Marx. An Introduction to Mathematical Statistics and its Applications. 3rd edition, Prentice Hall, 2000.
- David S. Moore and George P. McCabe, The Basic Practice of Statistics, 5th edition, Freeman, NY, 2009.
- Werner A. Stahel, Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler, 4. Auflage, Vieweg, 2002.
- John A. Rice, Mathematical Statistics and Data Analysis, Duxbury Press, 3rd edition, 2006.

4.3 Quellen für Daten und Beispiele

Die meisten der oben erwähnten Lehrbücher enthalten Beispiele von echten Datensätzen, mit Quellenangaben.

UNDERSTANDING UNCERTAINTY. The site that tries to make sense of chance, risk, luck, uncertainty and probability. Mathematics won't tell us what to do, but we think that understanding the numbers can help us deal with our own uncertainty and allow us to look critically at stories in the media.
http://understandinguncertainty.org/

CHANCE. The goal of Chance is to make students more informed, critical readers of current news stories that use probability and statistics.
http://www.dartmouth.edu/ chance/

DASL (pronounced "dazzle") is an online library of datafiles and stories that illustrate the use of basic statistics methods. We hope to provide data from a wide variety of topics so that statistics teachers can find real-world examples that will be interesting to their students. http://lib.stat.cmu.edu/DASL/

Bundesamt für Statistik. Das Portal Statistik Schweiz bietet im Bereich THEMEN eine Fülle von Materialien zu unzähligen Gebieten, die für die Schule relevant sind. "Für den

Unterricht" bringt Einstiege, Hinweise, Orientierungen und Anregungen zu ihrer Verwendung im Schulunterricht.
http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/forumschule/intro.html

J. Albert and R. H. Koning (eds), Statistical Thinking in Sports. Chapman and Hall/CRC, Boca Raton, 2008.

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski, A Handbook of Small Data Sets, Chapman and Hall, London, 1994.