

Kapitel 1

Einführung: Datenanalyse und mathematische Statistik

Die **explorative Datenanalyse (EDA)** ist ein Teilgebiet der Statistik. Sie verfolgt die Aufgabe, in vorhandenen Daten Strukturen zu erkennen, Hypothesen über Ursache und Grund der Daten zu bilden und Grundlagen für eingehendere statistische Modellbildung zu liefern. John W. Tukey hatte in den 1970er Jahren diese Bedeutung der EDA als Kritik und Ergänzung zur (mathematischen) Statistik, in der ein zu großes Gewicht auf das Auswerten und Testen von gegebenen Hypothesen gelegt wird, hervorgehoben. So ist neben den traditionellen statistischen Analysen ein kreativer Impuls gesetzt worden, der mit dem Schlagwort *Let the data speak for themselves* einen Anspruch auf eine bedeutsame und bisher vernachlässigte Aufgabe der Statistik erhob.

Mit der Entwicklung von geeigneten Softwarepaketen ist dieser Teil der Statistik in vielen Anwendungsbereichen in den Vordergrund gerückt. Datenanalytische Verfahren wie z.B. Boxplots, Histogramme, QQ-Plots, Scatterplots und Projection Pursuit sind zum Standard bei Anwendungen geworden und gehören auch zum Repertoire des verwandten **Data-Mining**. Dessen Hauptaufgabenstellung und Ziel ist es, unter Verwendung von Verfahren der multivariaten Statistik neue Muster in großen Datenmengen zu entdecken. Der Fokus des **machine learning** ist dagegen eher auf dem Entdecken bekannter Muster in vorhandenen Datenmengen, z.B. dem Auffinden von Personenbildern im Internet.

Typische Aufgabenstellungen sind die Erkennung von Ausreißern in Datenmengen, die Gruppierung von Objekten nach Ähnlichkeiten (**Clusteranalyse**), die Einteilung oder Einordnung in Klassen (**Klassifikation**), die Identifikation von Zusammenhängen in Daten (**Assoziationsanalyse**) und spezifischer die Beschreibung von funktionalen Zusammenhängen in Datenmengen (**Regressionsanalyse**). Allgemeines Ziel dieser Verfahren ist es, eine Reduktion der Datenmenge auf eine kompaktere Beschreibung ohne wesentlichen Informationsverlust vorzunehmen. Die oben genannten Verfahren können sowohl einen diagnostischen Charakter als auch einen prognostischen Charakter tragen.

Die **mathematische Statistik** basiert dagegen wesentlich auf der Modellierung eines Experimentes, einer Datenmenge durch ein statistisches Modell $(\mathfrak{X}, \mathcal{A}, \{P_\vartheta; \vartheta \in \Theta\})$. Kernaufgaben sind die Modellwahl und Modellevaluation und die begründete Konstruktion und Bewertung von statistischen Verfahren für Hypothesen über das Experiment. Standardaufgaben sind Test- und Schätzprobleme sowie Konfidenzintervalle und Klassifikationsverfahren. Die besondere Qualität der mathematischen Statistik besteht in der quantitativen Evaluation der angewendeten statistischen Verfahren. Diese ist nur möglich auf Grund des statistischen Modells, das die Daten beschreibt. Sie ist auch nur so präzise möglich, wie es die Modellbeschreibung des Experiments ist. Für viele grundlegende Aufgaben, z.B. Medikamententests sind präzise Modellbeschreibungen vorhanden und ermöglichen daher abgesicherte statistische Analysen. In komplexen Datensituationen, z.B. bei Finanzdaten oder räumlichen Datenmustern, können auch intrinsische Verfahren zur Abschätzung der Qualität einer Prognose oder eines Verfahrens der statistischen Datenanalyse wie z.B. Bootstrapsimulationen verwendet werden. Eine zuverlässige Einschätzung dieser Verfahren ist jedoch nur basierend auf Modellen möglich.

In allgemeiner Form besteht die Aufgabe der mathematischen Statistik darin, begründete Entscheidungen in Situationen unter Unsicherheit zu treffen. Dieser Aspekt spiegelt sich in dem engen Zusammenhang der mathematischen Statistik mit der **statistischen Entscheidungstheorie** und insbesondere mit der **Spieltheorie**. Für ein grundlegendes Verständnis statistischer Fragestellungen sind diese Verbindungen fruchtbringend und machen auch einen Teil des Reizes der mathematischen Statistik aus.

1.1 Regressionsanalyse

Als Beispiel für ein datenanalytisches Verfahren behandeln wir in diesem Abschnitt verschiedene Varianten von Regressionsverfahren.

1.1.1 Lineare Regression

Seien $(x_1, y_1), \dots, (x_n, y_n)$ zweidimensionale Daten. Eine Regressionsgerade $y = bx + a$ beschreibt eine lineare Abhängigkeit der Messvariable y von der Einflussvariablen x , z.B. eine zeitliche Abhängigkeit.

Die **Methode der kleinsten Quadrate** bestimmt die Regressionsgerade als Lösung des Minimierungsproblems:

$$S(a, b) := \sum_{i=1}^n (y_i - a - bx_i)^2 = \min! \quad (1.1)$$

Mit der notwendigen Bedingung

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

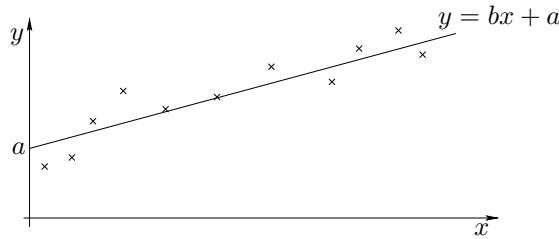


Abbildung 1.1 Regressionsgerade

ergibt sich $a = \bar{y}_n - b\bar{x}_n$ mit $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$. Weiter folgt aus

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0$$

$$b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i.$$

Einsetzen von a ergibt dann die **Regressionsgerade**:

$$y = \hat{a} + \hat{b}x \quad \text{mit} \quad \hat{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}_n \bar{y}_n}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2}, \quad \hat{a} = \bar{y}_n - \hat{b}\bar{x}_n. \quad (1.2)$$

Der Regressionskoeffizient \hat{b} hat die alternative Darstellung

$$\hat{b} = \frac{s_{x,y}}{s_x^2} \quad (1.3)$$

mit der **Stichprobenkovarianz**

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

und der **Stichprobenvarianz**

$$s_x^2 = s_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Die normierte Größe $r_{x,y} := \frac{s_{x,y}}{s_x s_y}$ heißt **empirischer Korrelationskoeffizient**. Nach der Cauchy-Schwarz-Ungleichung gilt

$$-1 \leq r_{x,y} \leq 1. \quad (1.4)$$

Ist $r_{x,y} \approx 0$ so sind die Daten annähernd linear unabhängig. Ist $r_{x,y} > 0$ so ist die Steigung $\hat{b} = r_{x,y} \frac{s_y}{s_x}$ der Regressionsgeraden $y = \hat{a} + \hat{b}x$ positiv, d.h. die Daten sind positiv linear abhängig; ist $r_{x,y} < 0$, dann negativ linear abhängig. Für

$r_{x,y} \approx 1$ oder -1 konzentrieren sich die Daten stark in der Nähe der Regressionsgeraden. Die Regressionsgerade liefert dann eine recht präzise Beschreibung der Datenmengen.

Ein Problem der Methode der kleinsten Quadrate zeigt die folgende Abbildung 1.2. Die Regressionsgerade ist nicht stabil. Ein einziger Ausreißerpunkt kann die Lage der Regressionsgerade stark verändern.

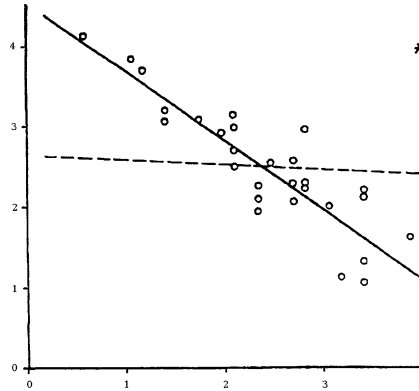


Abbildung 1.2 Regressionsgerade — — mit Zusatzpunkt * im Vergleich zur Regressionsgerade ohne Zusatzpunkt

Eine gegen Ausreißer stabile Version der Regressionsgerade wurde von Tukey eingeführt. Dazu werden die x -Werte gleichmäßig in drei Gruppe (kleine, mittlere, große) eingeteilt.

n	kleine	mittlere	große
$3l$	l	l	l
$3l + 1$	l	$l + 1$	l
$3l + 2$	$l + 1$	l	$l + 1$

Bilde nun die Mediane \tilde{x}_L , \tilde{x}_M , \tilde{x}_R der x -Werte dieser Gruppen und \tilde{y}_L , \tilde{y}_M , \tilde{y}_R der y -Werte dieser Gruppen. Sei $b_T = \frac{\tilde{y}_R - \tilde{y}_L}{\tilde{x}_R - \tilde{x}_L}$ der Anstieg der Geraden durch $(\tilde{x}_L, \tilde{y}_L)$, $(\tilde{x}_R, \tilde{y}_R)$ mit Achsenabschnitt $a_L = a_R$. Sei schließlich a_M der Achsenabschnitt der hierzu parallelen Gerade durch $(\tilde{x}_M, \tilde{y}_M)$. Mit $a_T := \frac{1}{3}(a_L + a_M + a_R)$ heißt dann

$$y = a_T + b_T x \quad (1.5)$$

Tukey-Gerade der Daten $(x_1, y_1), \dots, (x_n, y_n)$. Für das Beispiel aus Abbildung 1.2 mit Ausreißerpunkt * bleibt die Tukey-Gerade stabil (vgl. Abbildungen 1.3 und 1.4).

Die Abweichungen

$$r_i = y_i - a - b x_i \quad (1.6)$$

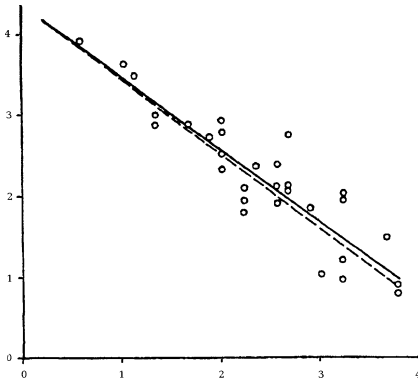


Abbildung 1.3 Tukey-Gerade — — im Vergleich zur Regressionsgeraden

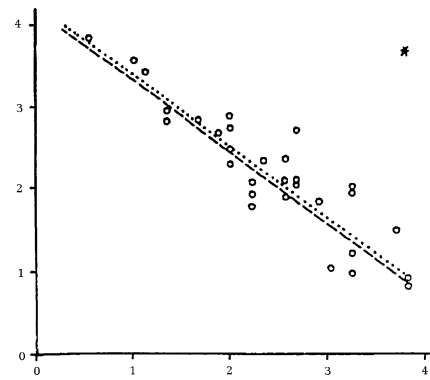


Abbildung 1.4 Tukey-Gerade — — mit Zusatzpunkt * im Vergleich zur Tukey-Geraden ohne Zusatzpunkt

der Variable y_i von den Werten der Regressionsgeraden heißen **Residuen**. Ist die Datenmenge gut durch eine lineare Regression darstellbar, dann sollten die Residuen unsystematisch um Null variieren. Das folgende Bild der Residuen würde deutlich gegen eine lineare Regression sprechen:

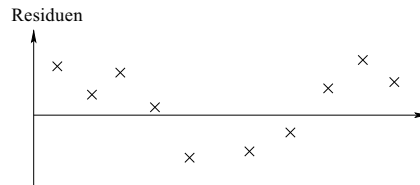


Abbildung 1.5 Residuen

1.1.2 Nichtlineare Abhängigkeit

Sei $y = f(a, b, c, x)$ eine nichtlineare Funktion einer Einflussvariablen x mit drei Parametern a, b, c . Bei quadratischer Abhängigkeit ist z.B. $y = a + bx + cx^2$.

Die Regressionsfunktion f wird wieder nach der Methode der kleinsten Quadrate angepasst.

$$F(a, b, c) := \sum_{i=1}^n (y_i - f(a, b, c, x_i))^2 = \min!_{a, b, c} \quad (1.7)$$

Die zugehörigen **kleinste Quadratgleichungen** lauten dann

$$\frac{\partial F}{\partial a} = 0, \quad \frac{\partial F}{\partial b} = 0, \quad \frac{\partial F}{\partial c} = 0 \quad (1.8)$$

und liefern Kandidaten \hat{a} , \hat{b} , \hat{c} für die Lösung von (1.7) wie im Fall der linearen Regression.

$y = f(\hat{a}, \hat{b}, \hat{c}, x)$ heißt dann **(nichtlineare) Regressionsfunktion** für die Daten $(x_1, y_1), \dots, (x_n, y_n)$.

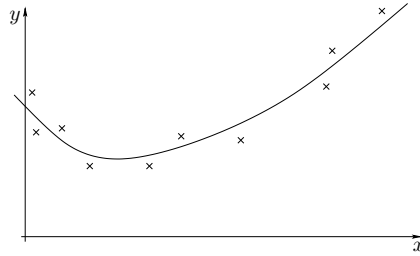


Abbildung 1.6 quadratische Regression

1.1.3 Lineare Modelle

Liegen multivariate Einflussgrößen $x = (x_1, \dots, x_m)^\top \in \mathbb{R}^m$ und multivariate Beobachtungsgrößen $y = (y_1, \dots, y_k)^\top \in \mathbb{R}^k$ vor, dann heißt

$$y = b + Ax \quad (1.9)$$

ein lineares Modell mit der **Designmatrix** $A = (a_{ij}) \in \mathbb{R}^{k \times m}$ und $b = (b_1, \dots, b_k) \in \mathbb{R}^k$, d.h. es gilt

$$y_i = \sum_{j=1}^m a_{ij} x_j + b_i, \quad 1 \leq i \leq k. \quad (1.10)$$

Bei gegebenen Daten (y^i, x^i) , $1 \leq i \leq n$ lautet die **Methode kleinster Quadrate**

$$F(A, b) := \sum_{i=1}^n \|y^i - (Ax^i + b)\|^2 = \min_{A, b}! \quad (1.11)$$

Die Lösungen \hat{A} , \hat{b} dieser Gleichungen lassen sich explizit angeben (vgl. Abschnitt 8.4 über Gauß-Markov-Schätzer) und bestimmen die multivariate Regressionsgerade

$$y = \hat{A}x + \hat{b}. \quad (1.12)$$

Einige Klassen nichtlinearer Regressionsgeraden wie in Abschnitt 1.1.2 z.B. Regressionen der Form $y = \sum a_i f_i(x) + b$ lassen sich als Spezialfall des linearen Modells (1.9) einordnen, indem als neue Einflussvariable $z_i = f_i(x)$ gewählt werden.

1.1.4 Nichtparametrische Regression

Gesucht wird ein funktionaler Zusammenhang $y = f(x)$ zwischen der Einflussvariablen x und der Beobachtungsvariablen y . Im Unterschied zu Abschnitt 1.1.2 ist f jedoch nicht nur bis auf einige Parameter a, b, c, \dots bestimmt sondern gänzlich bis auf evtl. qualitative Eigenschaften unbekannt. Ein vielfach verwendetes Verfahren zur Bestimmung einer Regressionsfunktion f bei gegebener Datenmenge (x_i, y_i) , $1 \leq i \leq n$, sind **Kernschätzer**. Sie basieren auf einem Kern $k: \mathbb{R}^1 \rightarrow \mathbb{R}_+$ (im Falle $x \in \mathbb{R}^1$) wie z.B. dem **Histogramm-Kern** $k(y) = \frac{1}{2}1_{[-1,1]}(y)$ oder dem **Gaußkern** $k(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$. Durch einen reellen Parameter $h > 0$, die **Bandweite**, lässt sich aus einem Kern k eine Klasse von Kernen erzeugen

$$k_h(y) := \frac{1}{h}k\left(\frac{y}{h}\right). \quad (1.13)$$

Damit erhalten wir nichtparametrische **Regressionsschätzer**

$$\hat{f}(x) = \hat{f}_h(x) = \frac{\frac{1}{n} \sum_{i=1}^n k_h(x - x_i) y_i}{\frac{1}{n} \sum_{i=1}^n k_h(x - x_i)}. \quad (1.14)$$

$\hat{f}(x)$ ist ein gewichteter Mittelwert der y_i -Werte zu x_i -Werten in der ‘Nähe’ von x . Der Gewichtungsfaktor $k_h(x - x_i)$ beschreibt den Einfluss der x_i in der Nähe von x . Die Nähe hängt einerseits vom Kern ab. Als gravierender stellt sich aber der Einfluss der Bandweite h heraus. Für kleine Bandweiten h ($h \downarrow 0$) wird nur über kleine Umgebungen von x gemittelt, der Regressionsschätzer $\hat{f}(x)$ wird irregulärer und passt sich mehr den Daten an, für große Bandweiten h wird $\hat{f}(x)$ glatter und gibt eine ‘weitsichtigere’ Interpolation der Daten (vgl. Abbildung 1.7).

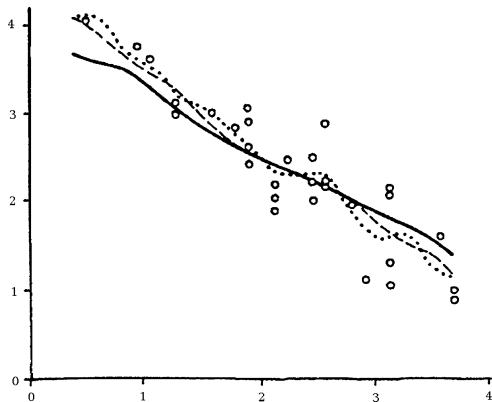


Abbildung 1.7 Regressionsschätzer mit Bandweiten $h = 0,25$ (\cdots), $h = 0,5$ ($- - -$) und $h = 1$ ($—$).

Für zu kleine Bandweiten h ist der Bias gering aber die Varianz von $\hat{f}(x)$ groß und daher die Prognosefähigkeit von \hat{f} in Frage gestellt, für zu große Bandweiten h

ist die Lage umgekehrt. Dieses ist das berühmte **Bias-Varianz-Dilemma**. Es lässt sich am besten in einem stochastischen Modell für die Daten (x_i, y_i) beschreiben. Dann ist der Schätzfehler gegeben durch $E(\hat{f}(x) - f(x))^2$. Dieser lässt sich in einem Bias-Term und einen Varianz-Term zerlegen

$$E(\hat{f}(x) - f(x))^2 = E(\hat{f}(x) - E\hat{f}(x))^2 + (E\hat{f}(x) - f(x))^2. \quad (1.15)$$

Unter sehr allgemeinen Bedingungen an das stochastische Modell gilt für $n \rightarrow \infty$ und $h = h(n) \rightarrow 0$, so dass $n \cdot h \rightarrow \infty$

$$E\hat{f}_h(x) \xrightarrow{h \rightarrow 0} f(x), \quad (1.16)$$

d.h. der Bias-Term $(E\hat{f}_h(x) - f(x))^2$ ist klein für (nicht zu) kleines h , aber der Varianzterm $E(\hat{f}_h(x) - E\hat{f}_h(x))^2$ ist groß für kleines h . Umgekehrt ist für ‘große’ Bandweite h der Varianzterm klein, dafür aber der Bias-Fehler groß.

Ein wichtiges Problem bei der Anwendung von nichtparametrischen Regressions-schätzern ist daher die Wahl einer geeigneten Bandweite h .

1.2 Mathematische Statistik – Entscheidung unter Unsicherheit

Zentrale Aufgabe der mathematischen Statistik ist es, geeignete statistische Modelle für ein Experiment zu konstruieren und zu evaluieren und basierend auf diesen Modellen geeignete statistische Verfahren zu konstruieren und zu bewerten. Ziel dieser Verfahren ist es, Entscheidungen unter Unsicherheit zu treffen und deren Risiko zu beschreiben. Im Folgenden behandeln wir einige Beispiele, um die Vielfalt dieser Fragestellung aufzuzeigen.

1.2.1 Ein Auswahlproblem

Zwei Zettel tragen die Zahlen x bzw. y und werden vermischt. Einem ‘Spieler’ dem die Zahlen x und y nicht bekannt sind, wird ein Zettel (zufällig ausgewählt) angeboten. Er kann den nun zufälligen Inhalt X dieses Zettels, $P(X = x) = P(X = y) = \frac{1}{2}$, akzeptieren oder zu dem Angebot Y des anderen Zettels wechseln und muss dann diesen akzeptieren. Sein Ziel ist es, mit möglichst großer Wahrscheinlichkeit eine Entscheidung für die größere der beiden Zahlen zu treffen.

Eine natürliche Frage ist es, ob der Spieler eine Entscheidungsregel finden kann, die ihm mit größerer Wahrscheinlichkeit als $\frac{1}{2}$, also bei zufälliger Auswahl, basierend auf der Kenntnis von X den größeren Gewinn sichert.

Die Unsicherheit in diesem Entscheidungsproblem wird durch das statistische Modell $\{P_{x,y}; x, y \in \mathbb{R}, x \neq y\}$ beschrieben mit $P_{x,y}(\{X = x\}) = P_{x,y}(\{X = y\}) = \frac{1}{2}$ und $P_{x,y}(\{Y = y \mid X = x\}) = P_{x,y}(\{Y = x \mid X = y\}) = 1$. Es gilt auch $P_{x,y}(\{Y = x\}) = P_{x,y}(\{Y = y\}) = \frac{1}{2}$.

Gibt es ein Entscheidungsverfahren $d : \mathbb{R} \rightarrow \{0, 1\}$ mit $d(X) = 0 \simeq$ Entscheidung für X , $d(X) = 1 \simeq$ Entscheidung für Y , so dass die Erfolgswahrscheinlichkeit E_d größer als $\frac{1}{2}$ ist:

$$\begin{aligned} E_d &= P(d(X) = 0, X > Y) + P(d(X) = 1, X < Y) \\ &> \frac{1}{2}. \end{aligned} \quad (1.17)$$

Es ist überraschend, dass die Antwort auf diese Frage positiv ist. Eine Lösung basiert auf einem randomisierten Entscheidungsverfahren. Sei Z eine Zufallsvariable mit positiver Dichte $f > 0$ auf \mathbb{R} , stochastisch unabhängig von X, Y ; z.B. sei Z normalverteilt, $Z \sim N(\mu, 1)$. Wir nennen Z eine ‘Splitvariable’ und definieren die Entscheidungsregel

$$d(x) = \begin{cases} 1, & x \leq Z, \\ 0, & x > Z. \end{cases} \quad (1.18)$$

Dann gilt

$$\begin{aligned} E_d &= P(d(X) = 1_{\{Y \geq X\}}) \\ &= P(X \geq Z, X \geq Y) + P(X < Y, X < Z) \\ &= \frac{1}{2}(P(X, Y < Z) + P(X, Y \geq Z)) + P(X < Z \leq Y) + P(Y < Z \leq X) \\ &= \frac{1}{2} + \frac{1}{2}(P(X < Z \leq Y) + P(Y < Z \leq X)). \end{aligned} \quad (1.19)$$

Es ist also

$$E_d - \frac{1}{2} = \frac{1}{2}P(Z \text{ ist ein Split von } X, Y) > 0, \quad (1.20)$$

da Z eine überall positive Dichte hat. Für jede Splitstrategie ist also die Erfolgswahrscheinlichkeit größer als $\frac{1}{2}$. Hat man keine weiteren Informationen, so gibt es keine Möglichkeit einen ‘guten’ oder ‘optimalen’ Split auszuwählen.

Hat man zusätzliche Informationen über die Verteilung von (X, Y) , so lässt sich ein guter Split konstruieren und die Erfolgswahrscheinlichkeit vergrößern. Ist der bedingte Median $m_x = \text{med}(Y \mid X = x)$ bekannt, dann ist $Z = m_X = \text{med}(Y \mid X)$ ein optimaler Split. Es gilt für eine Entscheidungsregel d

$$\begin{aligned} E_d &= P(d(X) = 1_{(Y \geq X)}) \\ &= P(d(X) = 1, Y \geq X) + P(d(X) = 0, Y < X) \\ &= \int (d(X)1_{(Y \geq X)} + (1 - d(X))1_{(Y < X)})dP \\ &= \int d(X)(1_{(Y \geq X)} - 1_{(Y < X)})dP + P(Y < X) \\ &= \int d(x)(P(Y \geq x \mid X = x) - P(Y < x \mid X = x))dP^X(x) + P(Y < X). \end{aligned} \quad (1.21)$$

Es gilt:

$$P(Y \geq x \mid X = x) - P(Y < x \mid X = x)$$

ist ≥ 0 wenn $x \leq m_x$ und ≤ 0 wenn $x > m_x$.

Daraus folgt, dass

$$d^*(x) = \begin{cases} 1, & \text{für } x \leq m_x, \\ 0, & \text{für } x > m_x, \end{cases} \quad (1.22)$$

eine **optimale Entscheidungsfunktion** ist, d.h. die Erfolgswahrscheinlichkeit E_{d^*} ist maximal.

Sind X, Y stochastisch unabhängig, ist $P^X = P^Y = Q$ stetig und $c = m_X = \text{med}(Q)$, dann gilt

$$E_{d^*} = \frac{3}{4}. \quad (1.23)$$

Mit Wahrscheinlichkeit $\frac{3}{4}$ lässt sich die größere Zahl finden. Der Median $Z = \text{med}(Q)$ ist eine optimale Split-Variable. In der in diesem Beispiel vorliegenden Entscheidungssituation unter Unsicherheit ist es möglich eine ‘optimale’ Entscheidung zu treffen und Erfolgs-/Misserfolgswahrscheinlichkeit genau zu quantifizieren.

1.2.2 Zufällige Folgen

a) Musterverteilungen

Gegeben sei eine 0-1-Folge x_1, \dots, x_n . Gesucht ist eine Entscheidungsverfahren um festzustellen, ob die Folge zufällig erzeugt ist, d.h. die Realisierung eines Bernoulliexperiments $P = \otimes_{i=1}^n \mathcal{B}(1, \frac{1}{2})$ ist. Die Idee eines solchen Testverfahrens ist es, zu prüfen, ob geeignete Muster in x_1, \dots, x_n in der Häufigkeit vorhanden ist, die man in einer Bernoullifolge erwarten würde.

Bezeichne etwa R_n die **Anzahl der Runs** in der 0-1-Folge x_1, \dots, x_n , d.h. die Anzahl der Wechsel von 0- und 1-Sequenzen. Z.B. hat die Folge 1100010000110 eine Anzahl von 6 Runs. Die **maximale Runlänge** M_n ist 4. Ein geeigneter Test zur Überprüfung der Hypothese, dass x_1, \dots, x_n Realisierungen eines Bernoulliexperiments sind ist es, die beobachteten Runzahlen r_n und m_n mit den Verteilungen von R_n, M_n im Bernoulli-Fall zu vergleichen.

Wir betrachten z.B. die folgenden 0-1-Folgen x, y der Länge $n = 36$. Welche der beiden Folgen ist zufällig erzeugt? Eine von ihnen stammt aus einem echten Zufallsgenerator, die andere Folge ist z.B. eine ‘ausgedachte Zufallsfolge’.

$$x = 111001100010010101000001001100001110$$

$$y = 101011010011000101100110010110101011$$

Um die zufällige Folge zu identifizieren bestimmen wir die maximalen Runlängen und die Anzahl der Runs, $M_n(x) = 5$, $R_n(x) = 18$, $M_n(y) = 3$, $R_n(y) = 25$. Wir vergleichen diese mit der Verteilung von R_n, M_n unter der Hypothese. Die Verteilung der maximalen Runlänge M_n (vgl. Abbildung 1.8) und der Anzahl der Runs

R_n (vgl. Abbildung 1.9) einer Bernoullifolge hat die folgende Form (Simulation mit 10.000 Wiederholungen, Länge $n = 36$).

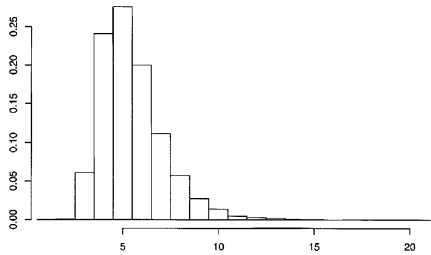


Abbildung 1.8 Dichte der maximalen Runlänge M_n

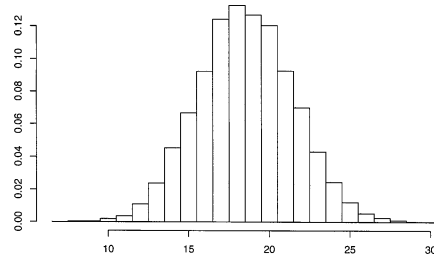


Abbildung 1.9 Dichte der Anzahl der Runs R_n

Es stellt sich heraus, dass $R_n(y)$ zu groß und $M_n(y)$ zu klein ist, während $R_n(x)$ und $M_n(x)$ in den zentralen Bereich der Verteilung unter der Hypothese fällt. Unsere Entscheidung lautet daher: x stammt aus dem echten Zufallsgenerator; die Folge y ist dagegen nicht zufällig konstruiert.

b) χ^2 -Test

Sei nun allgemeiner X_1, \dots, X_n eine zufällige, stochastisch unabhängige, identisch verteilte (iid) Folge mit Werten in $\{1, \dots, k\}$ mit $P(X_i = s) = p_s$, $1 \leq s \leq k$. Um zu prüfen, ob eine vorliegende Folge x_1, \dots, x_n Realisierung einer solchen stochastischen Folge ist, betrachten wir $Z_s := \sum_{i=1}^n 1_{\{s\}}(X_i)$, die Anzahl der Beobachtungen von Kategorie s . Unter der Hypothese zufälliger Folgen ist $EZ_s = np_s$. Die χ^2 -Statistik

$$V_n = \sum_{s=1}^k \frac{(Z_s - np_s)^2}{np_s} \quad (1.24)$$

ist ein gewichteter Vergleich der Anzahlen Z_s zu deren erwarteten Anzahlen.

Mit $(Z_s - np_s)^2 = Z_s^2 - 2np_s Z_s + n^2 p_s^2$, $\sum_{s=1}^k Z_s = n$, $\sum_{i=1}^k p_i = 1$ folgt

$$V_n = \frac{1}{n} \sum_{s=1}^k \frac{Z_s^2}{p_s} - n. \quad (1.25)$$

Aus dem zentralen Grenzwertsatz für (schwach abhängige) Folgen erhält man

$$P^{V_n} \xrightarrow{\mathcal{D}} \chi_{k-1}^2, \quad (1.26)$$

d.h. V_n konvergiert in Verteilung gegen eine χ^2 -Verteilung mit $k - 1$ Freiheitsgraden. Sei für $\alpha \in (0, 1)$ $\chi_{k-1, \alpha}^2$ das α -Fraktile der χ_{k-1}^2 -Verteilung, d.h. $P(Z \in [\chi_{k-1, \alpha}^2, \infty)) = \alpha$ für $Z \sim \chi_{k-1}^2$. Dann gilt unter der Hypothese

$$P(V_n \geq \chi_{k-1, \alpha}^2) \rightarrow P(Z \geq \chi_{k-1, \alpha}^2) = \alpha. \quad (1.27)$$

Typischerweise wird das Fehlerniveau α klein gewählt, z.B. $\alpha = 0,01$ oder $0,05$.

Sei v_n der beobachtete Wert von V_n . Der χ^2 -Test lehnt die Hypothese, dass x_1, \dots, x_n aus einer zufälligen iid Folge ist ab, wenn $v_n \geq \chi_{k-1, \alpha}^2$ ist. Die Fehlerwahrscheinlichkeit für diese Entscheidung ist (approximativ) nur α .

Die Größe

$$p = \chi_{k-1}^2([v_n, \infty)) \quad (1.28)$$

heißt **p -Wert** unseres Tests. Ist p sehr klein, z.B. $p = 0,01$, so spricht das stark gegen die Hypothese; die Abweichungen der Z_s von den erwarteten Häufigkeiten sind zu groß. Ist der p -Wert sehr groß, z.B. $p = 0,99$, so spricht das aber ebenfalls stark gegen die Hypothese einer zufälligen Folge. Die Abweichungen von den Erwartungswerten sind zu klein.

Beispiel 1.2.1 (n -maliger Wurf von 2 Würfeln)

Für den Wurf von 2 fairen Würfeln X_1 und Y_1 gilt mit $P(X_1 = s) = \frac{1}{6}$, $1 \leq s \leq 6$ und $p_s = P(X_1 + Y_1 = s)$

s	2	3	4	5	6	7	8	9	10	11	12
p_s	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Ein Experiment mit $n = 144$ Würfen von 2 Würfeln liefert

s	2	3	4	5	6	7	8	9	10	11	12
Z_s	2	4	10	12	22	29	21	15	14	9	6
np_s	4	8	12	16	20	24	20	16	12	8	4

Es ergibt sich $V_n = \sum_{s=2}^{12} \frac{(Z_s - np_s)^2}{np_s} = 7 \frac{7}{48}$. Einige α -Fraktile der χ_{10}^2 -Verteilung ersieht man aus folgender Tabelle

α	0,99	0,95	0,75	0,5	0,25	0,05	0,001
$\chi_{10, \alpha}^2$	2,56	3,94	6,74	9,34	12,55	18,31	23,21

Das Experiment ist mit der Verteilung verträglich. Wir können die Hypothese ‘fairer Würfel’ nicht ablehnen. Zwei weitere Experimente mit ‘Zufallszahlengeneratoren’ ergeben

s	2	3	4	5	6	7	8	9	10	11	12
1	4	10	10	13	20	18	18	11	13	14	13
2	3	7	11	15	19	24	21	17	13	9	5

Die zugehörigen χ^2 -Statistiken sind $V_n^1 = 29 \frac{59}{120}$, $V_n^2 = 1 \frac{17}{20}$. Die zugehörigen p -Werte sind $\chi_{10}^2([v_n^1, \infty)) = 0,001$, $\chi_{10}^2([0, v_n^2]) = 0,00003$. v_n^1 ist zu groß, so dass wir die Hypothese ablehnen. v_n^2 ist zu klein; die Folge ist nicht zufällig genug. Wir lehnen ebenfalls die Hypothese ab.

Der obige χ^2 -Test basiert auf dem Vergleich der Häufigkeiten Z_s mit den erwarteten Anzahlen np_s für Kategorie s . Er kann also nur Abweichungen von den Wahrscheinlichkeiten p_s entdecken. Möchte man Abweichungen der Folge x_1, \dots, x_n

von der Unabhängigkeitsannahme überprüfen, so kann man z.B. die Daten gruppieren zu $(x_1, x_2), (x_3, x_4), \dots$ und mit dem χ^2 -Test überprüfen ob die Häufigkeiten für Paarereignisse mit den Wahrscheinlichkeiten $p_{ij} = p_i p_j$ im Einklang sind. Es gibt viele einfallsreiche Mustervorschläge für solche Vergleiche, z.B. den Pokertest, der Muster aus dem Pokerspiel (Drillinge, Flush, Full House, ...) verwendet und deren Wahrscheinlichkeiten mit den beobachteten Häufigkeiten vergleicht.

c) Kolmogorov-Smirnov-Test

Um zu prüfen, ob eine Datenfolge x_1, \dots, x_n Realisierung einer iid Folge X_1, \dots, X_n mit Verteilungsfunktion F ist betrachten wir die empirische Verteilungsfunktion

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, x]}(x_i). \quad (1.29)$$

$\hat{F}_n(x)$ ist ein Schätzer für $F(x)$. Wir betrachten die Statistiken

$$\begin{aligned} K_n^+ &:= \sqrt{n} \max_n (\hat{F}_n(x) - F(x)) \\ K_n^- &:= \sqrt{n} \max_n (F(x) - \hat{F}_n(x)) \end{aligned} \quad (1.30)$$

und

$$K_n = \max(K_n^+, K_n^-) = \sqrt{n} \max |\hat{F}_n(x) - F(x)|, \quad (1.31)$$

Kolmogorov und Smirnov haben für stetige Verteilungsfunktionen F gezeigt:

$$K_n^+ \xrightarrow{\mathcal{D}} K_\infty^+ \quad (1.32)$$

mit $F_{K_\infty^+}(x) = 1 - e^{-2x^2}$, $x \geq 0$, die Kolmogorov-Smirnov-Verteilung. Wie in a) und b) vergleicht der Kolmogorov-Smirnov-Test den Wert der Statistik K_n^+ mit dem α -Fraktile der Kolmogorov-Smirnov-Verteilung.

1.2.3 Bildverarbeitung und Bilderkennung

a) Bildverarbeitung

Wir betrachten ein Bild B mit Graustufenwerten $0 = \text{schwarz}, 1, 2, \dots, 255 = \text{weiß}$. Jedes Pixel repräsentiert einen Graustufenwert, d.h. $B \in \{0, \dots, 255\}^{n \times m}$. $n \times m$ beschreibt das Format des Bildes, z.B. 500×500 . Wir betrachten $B = (b_{ij})$ als Ergebnis eines Zufallsprozesses mit diskreter empirischer Dichte f und empirischer Verteilungsfunktion $H(s)$.

Für eine reelle Zufallsvariable X mit stetiger Verteilungsfunktion $F = F_X$ ist die Transformation

$$Y = F(X) \quad (1.33)$$

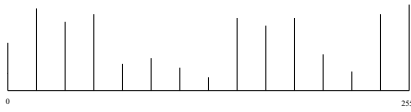


Abbildung 1.10 Graustufenverteilung:
empirische Dichte f

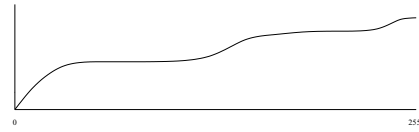


Abbildung 1.11 Graustufenverteilung:
empirische Verteilungsfunktion H (geglättet)

eine auf $[0, 1]$ gleichverteilte Zufallsvariable, denn für $x \in (0, 1)$ gilt

$$\begin{aligned} P(Y \leq x) &= P(F(X) \leq x) \\ &= P(X \leq F^{-1}(x)) = F(F^{-1}(x)) = x. \end{aligned}$$

Für eine nicht stetig verteilte Zufallsvariable X ist die Transformation $Y = F(X)$ nur annähernd gleichverteilt. Sei nun X eine Zufallsvariable mit $F_X = H$, d.h. X repräsentiert die Graustufenverteilung des Bildes B . Dann liefert $s \xrightarrow{h} [255 \cdot H(s)]$ eine angenäherte Gleichverteilung der Graustufenwerte

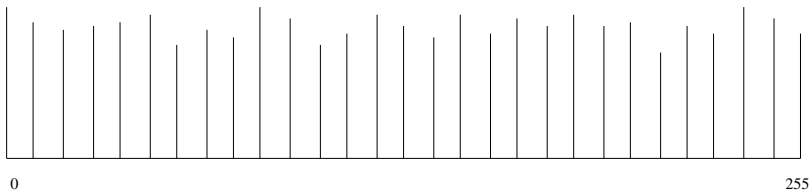


Abbildung 1.12 Graustufenwerte

Die Transformation der Graustufen (Farben) des Bildes B

$$B \longrightarrow \hat{B} = (h(b_{ij})) = \hat{B}_h \quad (1.34)$$

liefert ein deutlich verbessertes Bild mit einer Verstärkung der Kontraste. Dieses ist eine vielfach verwendete Methode der Bildverarbeitung, die z.B. bei Nacht-sichtgeräten Verwendung findet. Das folgende Beispiel einer Luftaufnahme einer Stadt (Straßburg) mittels Aufnahme aus einem Weltraumsatelliten (vgl. Abbildung 1.13) zeigt eindrücklich die erzielte Verbesserung der Kontraststruktur (vgl. Abbildung 1.14).

Grundlage dieser Methode der Bildverarbeitung ist das einfache Transformationsresultat in (1.33) für die Verteilung der Graustufenwerte.

b) Bilderkennung

Ein Bild B_0 , z.B. das Bild einer Person, soll in einer großen Datenbank gefunden werden. Dieses geschieht, indem für eine große Anzahl von Bildern B aus der Datenbank das Testproblem mit den Hypothesen $H_0 : B = B_0$, $H_1 : B \neq B_0$ gelöst

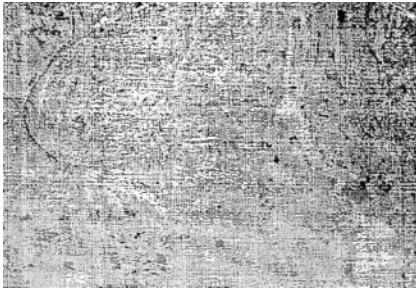


Abbildung 1.13 Satellitenbild von
Straßburg, nicht transformiert¹



Abbildung 1.14 Satellitenbild von
Straßburg, transformiert¹

wird. Die Bilder B , die durch den Test nicht abgelehnt werden, sind mögliche Kandidaten.

Zur Konstruktion eines Anpassungstests für H_0 ist es grundlegend, eine **Datenreduktion** durch die Auswahl geeigneter Merkmale vorzunehmen. Für praktische Zwecke werden diese z.B. rotationsinvariant gewählt um die Person auch in unterschiedlichen Position zu erkennen.

Seien $T(B) = (T_1(B), \dots, T_k(B))$ k geeignete Merkmale (z.B. Breite des Augenabstandes, Länge der Nase, ... bei Personenbildern). Das gesuchte Bild B_0 habe den bekannten Merkmalsvektor $t = T(B_0) = (t_1, \dots, t_k)$. Wir treffen (basierend auf historischen Daten) die Annahme, dass $T(B)$ multivariat normalverteilt ist mit bekannter Kovarianzmatrix Σ , $T(B) \sim N(\mu, \Sigma)$. Das Testproblem reduziert sich dann auf die Hypothesen

$$H_0 : \mu = t, \quad H_1 : \mu \neq t. \quad (1.35)$$

Als Teststatistik verwenden wir den normierten Abstand von $T(B)$ zu t

$$S := (T(B) - t)^\top \Sigma^{-1} (T(B) - t). \quad (1.36)$$

Unter der Nullhypothese H_0 ist

$$Y = \Sigma^{-\frac{1}{2}} (T(B) - t) \sim N(0, I_k). \quad (1.37)$$

Hieraus folgt, dass

$$S = Y^\top Y = \sum_{i=1}^k Y_i^2 \sim \chi_k^2,$$

S ist χ_k^2 -verteilt. Damit können wir als Bilderkennungstest den folgenden Test verwenden:

$$\varphi(B) = \begin{cases} 1, & S \geq \chi_{k,\alpha}^2, \\ 0, & S < \chi_{k,\alpha}^2, \end{cases} \quad (1.38)$$

wobei $\chi_{k,\alpha}^2$ das α -Fraktile der χ_k^2 -Verteilung ist.

¹Quelle: C. Dupuis: How calculators and computers change the field of problems in teaching statistics, in: *Teaching of Statistics in the Computer Age*, L. Råde and T. Speed (eds.); ISBN: 91-44-23631-X; 1985, 45–59, Figures 10a and 10b

φ ist ein Test zum Niveau α ; d.h. wenn $B \hat{=} B_0$ ist, dann wird das mit einer Fehlerwahrscheinlichkeit $\leq \alpha$ nicht erkannt. Ist die Kovarianzmatrix Σ in obigem Normalverteilungsmodell nicht bekannt, so ersetzt man Σ in der Teststatistik durch einen Schätzer $\hat{\Sigma}$ von Σ (**plug-in-Methode**) und kann den Test in ähnlicher Form durchführen. Auf Testverfahren, die ähnlich zu dem obigen aufgebaut sind, basieren viele der sehr schnellen und effektiven Suchverfahren für Bilder im Internet. Sie werden noch gekoppelt mit Indikatoren für interessante Links und Seiten, wo das gesuchte Bild mit größerer Wahrscheinlichkeit zu finden ist. Diese Seiten werden zuerst abgesucht und getestet.

Mathematische Statistik

Rüschendorf, L.

2014, XI, 427 S. 36 Abb., Softcover

ISBN: 978-3-642-41996-6