

## Teil B

# Datenanalyse

Die Beschreibung von diskreten und kontinuierlichen Zufallsvariablen durch Wahrscheinlichkeiten und Wahrscheinlichkeitsdichteverteilungen bildet die Grundlage der mathematischen Statistik (B.1). Eventuelle Zusammenhänge zwischen mehreren Zufallsvariablen werden durch Kovarianz und Korrelation beschrieben (B.2).

Durch Parameterschätzung lassen sich Eigenschaften von Verteilungen aus einer Stichprobe bestimmen. Die Maximum-Likelihood-Methode liefert eine Vorschrift für die Konstruktion von Schätzfunktionen und für die Anpassung von Modellen an Messdaten (B.3).

Ein statistischer Test liefert eine quantitative Aussage darüber, inwiefern eine Beobachtung mit einer Hypothese verträglich ist (B.4). Damit verknüpft ist die Angabe von Ergebnissen in Form von Konfidenzintervallen zu bestimmten Konfidenzniveaus (B.5).

Die Fehlerfortpflanzung wird so verallgemeinert, dass in linearer Näherung die Varianzen und Kovarianzen der Zielgrößen unter Berücksichtigung von Korrelationen zwischen den Eingangsgrößen berechnet werden können (B.6).

Die lineare Regression liefert eine analytische Lösung für die Anpassung von Modellen, die in den Parametern linear sind, wobei auch Unsicherheiten berücksichtigt werden können. Die Kovarianzmatrix erlaubt die Konstruktion eines Konfidenzbands (B.7). Das allgemeinste Verfahren zur Anpassung von Modellen ist die numerische Minimierung von  $\chi^2$ , die gleichzeitig ein quantitatives Maß für die Anpassungsgüte liefert (B.8).

Für ein freiwilliges vertieftes Studium empfehlen wir die folgenden Lehrbücher:

- [1] Taylor, An introduction to error analysis, University Science Books, 1997
- [2] Cowan, Statistical data analysis, Oxford Science Publications, 1998
- [3] Waldi, Statistische Datenanalyse, <http://www.redi-bw.de/start/unifr/EBooks-springer/10.1007/978-3-662-47145-6>
- [4] Brandt, Datenanalyse für Naturwissenschaftler und Ingenieure, <http://www.redi-bw.de/start/unifr/EBooks-springer/10.1007/978-3-642-37664-1>

## B.1 Verteilungen

### Verteilungen für diskrete Zufallsvariablen

Für eine diskrete Zufallsvariable  $x$  mit endlich oder unendlich vielen möglichen Werten  $x_i$  ist die Verteilung charakterisiert durch die Wahrscheinlichkeiten für alle  $x_i$ :

$$f(x_i) = \text{Wahrscheinlichkeit für den Wert } x_i \quad (\text{B.1})$$

Äquivalent können kumulierte Wahrscheinlichkeiten angegeben werden:

$$F(x) = \sum_{x_i \leq x} f(x_i) = \text{Wahrscheinlichkeit für einen Wert } \leq x \quad (\text{B.2})$$

Die Normierung entspricht der Forderung, mit Sicherheit irgendeinen Wert zu finden:

$$\sum f(x_i) = 1 \quad (\text{B.3})$$

### Verteilungen für kontinuierliche Zufallsvariablen

Für eine kontinuierliche Zufallsvariable  $x$  mit Werten in  $\mathbb{R}$  ist die Verteilung charakterisiert durch die Wahrscheinlichkeitsdichtefunktion  $f$ , so dass:

$$f(x) dx = \text{Wahrscheinlichkeit für einen Wert im Intervall } [x, x + dx] \quad (\text{B.4})$$

Äquivalent kann die kumulierte Verteilungsfunktion angegeben werden:

$$F(x) = \int_{-\infty}^x f(x') dx' = \text{Wahrscheinlichkeit für einen Wert } \leq x \quad (\text{B.5})$$

Die Normierung entspricht der Forderung, mit Sicherheit irgendeinen Wert zu finden:

$$\int f(x) dx = 1 \quad (\text{B.6})$$

### Quantile, Median, Interquartilsabstand

Das  $\alpha$ -Quantil  $q_\alpha$  ist derjenige  $x$ -Wert, der die Verteilung so teilt, dass der Anteil  $\alpha$  der Werte  $\leq q_\alpha$  ist. Das entspricht der Umkehrung der kumulierten Verteilungsfunktion:

$$q_\alpha = F^{-1}(\alpha) \quad (\text{B.7})$$

Als Spezialfall erhält man den Median  $q_{1/2}$ , der die Verteilung in zwei gleich große Hälften teilt, als ein alternatives Lagemaß. Die Quartile  $q_{1/4}$  und  $q_{3/4}$  definieren den Interquartilsabstand  $q_{3/4} - q_{1/4}$  als ein alternatives Streuungsmaß.

**Erwartungswerte, Mittelwert, Varianz, Standardabweichung, Momente**

Für eine Zufallsvariable  $x$ , die nach der Wahrscheinlichkeitsdichtefunktion  $f(x)$  verteilt ist, und eine Funktion  $g(x)$  definiert man den Erwartungswert der Funktionswerte:

$$E[g(x)] = \sum g(x_i) f(x_i) \quad \text{bzw.} \quad E[g(x)] = \int g(x) f(x) dx \quad (\text{B.8})$$

Den Erwartungswert der Zufallsvariablen selbst nennt man den Mittelwert der Verteilung:

$$\mu = E[x] \quad (\text{B.9})$$

Den Erwartungswert der mittleren quadratischen Abweichung der Zufallsvariablen von ihrem Mittelwert bezeichnet man als die Varianz der Verteilung:

$$\sigma^2 = E[(x - \mu)^2] = E[x^2] - \mu^2 \quad (\text{B.10})$$

Die Wurzel aus der Varianz ist die Standardabweichung der Verteilung  $\sigma$ .

Der Erwartungswert  $E[x^k]$  definiert das  $k$ -te Moment, der Erwartungswert  $E[(x - \mu)^k]$  das  $k$ -te zentrale Moment der Verteilung. Das erste Moment entspricht somit dem Mittelwert, das zweite zentrale Moment der Varianz. Das dritte zentrale Moment beschreibt die Asymmetrie der Verteilung, es ist proportional zur Schiefe. Das vierte zentrale Moment beschreibt die Häufung der Werte um den Mittelwert, es ist proportional zur Wölbung oder Kurtosis.

**Binomialverteilung**

Führt man ein Experiment, bei dem mit der Wahrscheinlichkeit  $p$  ein Erfolg eintritt,  $n$  mal durch, dann ist die Wahrscheinlichkeit für  $x$  Erfolge gegeben durch die Binomialverteilung:

$$f_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{mit} \quad \binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (\text{B.11})$$

Der Mittelwert der Binomialverteilung ist  $np$ , die Varianz ist  $np(1-p)$ .

**Poisson-Verteilung**

Für  $n \rightarrow \infty$ ,  $p \rightarrow 0$ ,  $np = \mu$  geht die Binomialverteilung über in eine Poisson-Verteilung:

$$f_{\mu}(x) = \frac{\mu^x}{x!} e^{-\mu} \quad (\text{B.12})$$

Der Mittelwert und die Varianz der Poisson-Verteilung sind  $\mu$ .

Für  $\mu \gg 1$  geht die Poisson-Verteilung in eine diskrete Normalverteilung mit Mittelwert  $\mu$  und Varianz  $\mu$  bzw. Standardabweichung  $\sqrt{\mu}$  über.

**Rechteck- und Dreiecksverteilung**

Für eine Rechteckverteilung der Breite  $2a$  ergibt sich aus (B.10) die Varianz  $\sigma^2 = a^2/3$ , für eine Dreiecksverteilung der Breite  $2a$  ergibt sich die Varianz  $\sigma^2 = a^2/6$ .

## Chi-Quadrat-Verteilung

Wir betrachten eine Stichprobe der Länge  $n$  für eine Zufallsvariable, deren Einzelwerte  $x_i$  einer Normalverteilung mit Mittelwert  $\mu$  und Standardabweichung  $\sigma$  folgen. Die Summe der durch  $\sigma^2$  normierten quadratischen Abweichungen vom wahren Wert  $\mu$

$$\chi^2 = \sum \frac{(x_i - \mu)^2}{\sigma^2} \quad (\text{B.13})$$

folgt einer Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden:

$$f_n(x) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad (\text{B.14})$$

Der Mittelwert der Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden ist  $n$ , die Varianz  $2n$ .

Die Summe der durch  $\sigma^2$  normierten quadratischen Abweichungen vom Mittelwert  $\bar{x}$

$$\chi^2 = \sum \frac{(x_i - \bar{x})^2}{\sigma^2} \quad (\text{B.15})$$

folgt ebenfalls einer Chi-Quadrat-Verteilung, allerdings mit  $n - 1$  Freiheitsgraden.

## Student-t-Verteilung

Die durch die wahre Standardabweichung der Verteilung der Mittelwerte  $\sigma/\sqrt{n}$  normierte („standardisierte“) Abweichung der Mittelwerte vom wahren Wert

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (\text{B.16})$$

folgt einer Standardnormalverteilung mit Mittelwert 0 und Standardabweichung 1.

Die durch die geschätzte Standardabweichung der Verteilung der Mittelwerte  $s_{\bar{x}} = s_x/\sqrt{n}$  normierte („studentisierte“) Abweichung der Mittelwerte vom wahren Wert

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad (\text{B.17})$$

folgt einer t-Verteilung mit  $n - 1$  Freiheitsgraden, definiert durch (für  $n$  Freiheitsgrade):

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (\text{B.18})$$

Der Mittelwert der t-Verteilung mit  $n$  Freiheitsgraden ist 0, die Varianz  $n/(n - 2)$ .

Für  $n \rightarrow \infty$  geht die t-Verteilung in eine Standardnormalverteilung über.

## B.2 Kovarianz und Korrelation

Wir betrachten zwei Zufallsvariablen  $x, y$  mit der gemeinsamen Wahrscheinlichkeitsdichtefunktion  $f(x, y)$ . Die einzelnen Verteilungen der  $x, y$  entsprechen den Randverteilungen:

$$f(x) = \int f(x, y) dy \quad \text{und} \quad f(y) = \int f(x, y) dx \quad (\text{B.19})$$

Falls die Variablen  $x, y$  voneinander unabhängig sind, dann ist die gemeinsame Wahrscheinlichkeitsdichtefunktion das Produkt der Randverteilungen:

$$f(x, y) = f(x) \cdot f(y) \quad (\text{B.20})$$

Falls die Variablen nicht voneinander unabhängig sind, dann kann es einen Zusammenhang geben in dem Sinne, dass der Wert von  $x$  die Wahrscheinlichkeitsverteilung für  $y$  beeinflusst und umgekehrt. Ein solcher Zusammenhang wird beschrieben durch die Kovarianz:

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)] \quad (\text{B.21})$$

Die Kovarianz ist positiv (negativ), wenn mit  $x > \mu_x$  eine erhöhte (erniedrigte) Wahrscheinlichkeit für  $y > \mu_y$  einhergeht.

Als anschaulicheres Maß definiert man den dimensionslosen Korrelationskoeffizienten:

$$\varrho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{mit} \quad -1 \leq \varrho_{xy} \leq +1 \quad (\text{B.22})$$

Für eine Darstellung in Matrixschreibweise definiert man die Kovarianzmatrix

$$V = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \quad (\text{B.23})$$

mit den Varianzen als Diagonal- und den Kovarianzen als Nichtdiagonalelemente.

Analog definiert man die Korrelationsmatrix:

$$R = \begin{pmatrix} 1 & \varrho_{xy} \\ \varrho_{xy} & 1 \end{pmatrix} \quad (\text{B.24})$$

Unabhängige Variablen  $x, y$  mit (B.20) sind immer unkorreliert, aber umgekehrt sind linear unkorrelierte Variablen mit  $\varrho_{xy} = 0$  nicht immer unabhängig.

Für die Schätzung von  $\sigma_{xy}$  aus Stichproben definiert man die empirische Kovarianz

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \quad (\text{B.25})$$

und für die Schätzung von  $\varrho_{xy}$  den empirischen Korrelationskoeffizienten:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \text{mit} \quad -1 \leq r_{xy} \leq +1 \quad (\text{B.26})$$

## B.3 Parameterschätzung

### Allgemeines Konzept, Schätzfunktion, Eigenschaften

Wir betrachten eine Zufallsvariable  $x$  mit Verteilung  $f(x)$ . Für diese Zufallsvariable kennen wir eine Stichprobe  $(x_1 \dots x_n)$  der Länge  $n$ . Die zentrale Aufgabe der induktiven Statistik ist die Schätzung der Eigenschaften von  $f(x)$  auf Basis der beobachteten Stichprobe.

Die mathematische Form der Verteilung  $f_p(x)$  mit Parameter(n)  $p$  wird vorgegeben. Für die Parameterschätzung müssen dann Schätzfunktionen  $\hat{p} = \hat{p}(x_1 \dots x_n)$  konstruiert werden.

Dabei werden folgende Eigenschaften angestrebt:

- Konsistenz, d.h. für  $n \rightarrow \infty$  konvergiert  $\hat{p}$  gegen den wahren Wert  $p$ .
- Erwartungstreue, d.h. für alle  $n$  ist  $E[\hat{p}] = p$ .
- Effizienz, d.h. der mittlere quadratische Fehler  $E[(\hat{p} - p)^2]$  ist möglichst klein.

### Maximum Likelihood (ML)

Unter der Annahme, dass die  $x$ -Werte entsprechend  $f_p(x)$  verteilt sind, ist die Wahrscheinlichkeit dafür, die beobachtete Stichprobe zu finden, proportional zur Likelihood-Funktion:

$$L(p) = \prod f_p(x_i) \quad (\text{B.27})$$

Diese entspricht einfach der kombinierten Wahrscheinlichkeitsdichtefunktion der  $x_i$ . Allerdings betrachtet man die  $x_i$  hier als gegeben, also  $L$  als Funktion der Parameter  $p$ .

Der Maximum-Likelihood-Schätzer  $\hat{p}$  maximiert die Likelihood-Funktion

$$L(\hat{p}) = \max \quad (\text{B.28})$$

oder äquivalent ihren Logarithmus, die Log-Likelihood-Funktion:

$$\ln L(\hat{p}) = \max \quad (\text{B.29})$$

Die Konstruktion von Schätzfunktionen mit der Maximum-Likelihood-Methode ist einfach und plausibel, garantiert aber nicht optimale Eigenschaften.

### Varianz von ML-Schätzern

Aus der Verteilung der Werte eines ML-Schätzers  $\hat{p}$  lässt sich nach der Definition (B.10) die Varianz  $\sigma_{\hat{p}}^2$  bzw. die Standardabweichung  $\sigma_{\hat{p}}$  berechnen:

$$\sigma_{\hat{p}}^2 = E[(\hat{p} - E[\hat{p}])^2] = E[\hat{p}^2] - E[\hat{p}]^2 \quad (\text{B.30})$$

Alternativ kann die Varianz aus der Log-Likelihood-Funktion  $\ln L(p)$ , die um  $p = \hat{p}$  näherungsweise parabelförmig ist, geschätzt werden, entweder aus der zweiten Ableitung

$$\hat{\sigma}_{\hat{p}}^2 = - \left( \frac{\partial^2 \ln L}{\partial p^2} \right)_{p=\hat{p}}^{-1} \quad (\text{B.31})$$

oder aus der Abnahme um  $1/2$  gegenüber dem Maximum:

$$\ln L(\hat{p} \pm \hat{\sigma}_{\hat{p}}) = \ln L(\hat{p}) - \frac{1}{2} \quad (\text{B.32})$$

Für mehrere Parameter ( $k > 1$ ) erhält man aus (B.31) entsprechend die Kovarianzmatrix. Die Bedingung (B.32) liefert eine Kontur konstanter Wahrscheinlichkeit im Parameterraum, die einen bestimmten Konfidenzbereich einschließt, wobei das Konfidenzniveau kritisch von  $k$  abhängt. Die Werte  $\hat{p} \pm \hat{\sigma}_{\hat{p}}$  bilden Tangenten an diesen Bereich.

### Beispiele für die Normalverteilung

Wir betrachten Messwerte, die einer Normalverteilung mit Mittelwert  $\mu$  und Varianz  $\sigma^2$  folgen. Die Parameter  $\mu$  und  $\sigma^2$  sollen aus einer Stichprobe der Länge  $n$  geschätzt werden.

Als ML-Schätzer für  $\mu$  erhält man den bekannten arithmetischen Mittelwert,  $\hat{\mu} = \bar{x}$ . Dieser Schätzer ist konsistent und erwartungstreu. Seine Varianz  $\sigma^2/n$  entspricht der bekannten Standardabweichung des Mittelwerts  $s_{\bar{x}}$ .

Als ML-Schätzer für  $\sigma^2$  ergibt sich:

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \quad (\text{B.33})$$

Dieser Schätzer ist nur asymptotisch erwartungstreu, denn sein Erwartungswert ist:

$$E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2 \quad (\text{B.34})$$

Durch Multiplikation mit  $n/(n-1)$ , die sogenannte Bessel-Korrektur, wird aus dem ML-Schätzer  $\hat{\sigma}^2$  der bekannte erwartungstreue Schätzer  $s_x^2$ . Seine Varianz  $2\sigma^4/\sqrt{n-1}$  entspricht der bekannten Standardabweichung der Standardabweichung.

### Kurvenanpassung mit ML

Die ML-Methode liefert auch eine Vorschrift für die Anpassung eines Modells  $y = m_p(x)$  an gegebene Messwerte  $(x_i, y_i \pm \Delta y_i)$ . Unter der Annahme, dass die  $y_i$  normalverteilt sind mit Mittelwerten  $\mu_i = m_p(x_i)$  und Standardabweichungen  $\sigma_i = \Delta y_i$ , gilt für jeden einzelnen Punkt  $(x_i, y_i)$  die Wahrscheinlichkeitsdichte:

$$f_{p,x_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[y_i - m_p(x_i)]^2}{2\sigma_i^2}\right) \quad (\text{B.35})$$

Der ML-Schätzer  $\hat{p}$  maximiert die Likelihood-Funktion

$$L(p) = \prod \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[y_i - m_p(x_i)]^2}{2\sigma_i^2}\right) \quad (\text{B.36})$$

oder äquivalent die Log-Likelihood-Funktion:

$$\ln L(p) = \text{const} - \frac{1}{2} \sum \frac{[y_i - m_p(x_i)]^2}{\sigma_i^2} \quad (\text{B.37})$$

Die Likelihood-Maximierung entspricht also der Minimierung von Chi-Quadrat

$$\chi^2 = \sum \frac{[y_i - m_p(x_i)]^2}{\sigma_i^2} \quad (\text{B.38})$$

bzw. für  $\sigma_i = \sigma$  der Minimierung der Fehlerquadratsumme:

$$S = \sum [y_i - m_p(x_i)]^2 \quad (\text{B.39})$$

Die Schätzung der Parameter-Unsicherheit  $\hat{\sigma}_{\hat{p}}$  nach (B.32) führt auf:

$$\chi^2(\hat{p} \pm \hat{\sigma}_{\hat{p}}) = \chi^2(\hat{p}) + 1 \quad \text{bzw.} \quad S(\hat{p} \pm \hat{\sigma}_{\hat{p}}) = S(\hat{p}) + \sigma^2 \quad (\text{B.40})$$

### Kombination von Ergebnissen

Die Kombination mehrerer Messwerte  $y_i \pm \Delta y_i$  zu einem gemeinsamen Ergebnis kann als Spezialfall der Kurvenanpassung mit der ML-Methode betrachtet werden. Die Annahme, dass die  $y_i$  normalverteilt sind mit dem für alle Messungen identischen Mittelwert  $\mu$  und den Standardabweichungen  $\sigma_i = \Delta y_i$ , führt auf die Minimierung von

$$\chi^2(\mu) = \sum \frac{(y_i - \mu)^2}{\sigma_i^2} \quad (\text{B.41})$$

und damit auf die bekannten Formeln für den gewichteten Mittelwert und seine Unsicherheit.

## B.4 Hypothesentests

### Hypothesen, Nullhypothese, P-Wert

Das Ziel eines statistischen Tests ist eine Aussage über die Wahrscheinlichkeiten, mit denen eine Beobachtung unter Annahme von bestimmten Hypothesen erklärt werden kann.

Im einfachsten Fall überprüft man nur eine Hypothese, die sogenannte Nullhypothese  $H_0$ . Diese kann mithilfe des Tests abgelehnt oder akzeptiert (nicht abgelehnt) werden.

Zunächst konstruiert man eine Funktion  $t = t(x)$ , die beliebigen Messwerten  $x$  einen (skalaren) Testwert  $t$  zuordnet. Unter Annahme der Nullhypothese  $H_0$  berechnet man die Verteilung der Messwerte  $f(x|H_0)$  und daraus die Verteilung der Testwerte  $f(t|H_0)$ .

Für den beobachteten Messwert  $x_o$  ergibt sich ein Testwert  $t_o = t(x_o)$ . Die  $t$ -Werte aller Messungen, die mit der Nullhypothese mindestens so unverträglich sind wie die Beobachtung, bilden den Bereich  $Q$ . Je nach Formulierung kann dieser Bereich ein- oder zweiseitig sein.

Damit berechnet man den P-Wert für die Beobachtung  $x_o$ :

$$P = \int_Q f(t|H_0) dt \quad (\text{B.42})$$

Dieser entspricht der Wahrscheinlichkeit, bei korrekter Nullhypothese zufällig ein Ergebnis zu erhalten, das mindestens so unverträglich ist wie die Beobachtung. Der P-Wert entspricht *nicht* der Wahrscheinlichkeit dafür, dass die Nullhypothese wahr ist.



## Test, Ablehnungsbereich, Signifikanzniveau

Für die Formulierung eines statistischen Tests definiert man einen Ablehnungsbereich  $R$  so, dass für Beobachtungen  $x_o$  mit  $t_o \in R$  die Nullhypothese abgelehnt wird. Dieser Bereich kann ein- oder zweiseitig sein.

Aus dem Ablehnungsbereich ergibt sich das Signifikanzniveau des Tests:

$$\alpha = \int_R f(t|H_0) dt \quad (\text{B.43})$$

Dieses entspricht der Wahrscheinlichkeit, bei korrekter Nullhypothese (also rein zufällig) einen Testwert im Ablehnungsbereich zu erhalten („Irrtumswahrscheinlichkeit“).

Häufig wird umgekehrt das Signifikanzniveau vorgegeben (beispielsweise  $\alpha = 0,05$ ) und der Ablehnungsbereich entsprechend bestimmt.

## Chi-Quadrat-Test

Der Chi-Quadrat-Test wird zur Überprüfung der Anpassungsgüte eingesetzt.

Man betrachtet Messwerte  $(x_i, y_i \pm \Delta y_i)$  und ein Modell  $y = m_p(x)$  mit  $k$  Parametern. Als Testwert wählt man den Wert  $\chi^2(\hat{p})$  für die angepassten Parameter  $\hat{p}$ .

Unter der Nullhypothese, dass die Daten normalverteilt sind mit  $\mu_i = m_{\hat{p}}(x_i)$  und  $\sigma_i = \Delta y_i$ , folgt der Testwert  $\chi^2(\hat{p})$  einer Chi-Quadrat-Verteilung  $f_{n-k}(x)$  mit  $n - k$  Freiheitsgraden.

Der Ablehnungsbereich wird einseitig (rechts) festgelegt. Der P-Wert ist:

$$P = \int_{\chi^2(\hat{p})}^{\infty} f_{n-k}(x) dx \quad (\text{B.44})$$

Die Nullhypothese wird auf dem Signifikanzniveau  $\alpha$  abgelehnt, wenn  $P \leq \alpha$  ist.

## t-Test

Der t-Test prüft, ob ein Messwert von einem erwarteten Wert signifikant abweicht.

Man betrachtet eine Stichprobe  $(x_1 \dots x_n)$  mit Mittelwert  $\bar{x}$ , Standardabweichung  $s_x$  und Standardabweichung des Mittelwerts  $s_{\bar{x}}$ . Als Testwert wählt man die Abweichung vom erwarteten Wert  $y$  in Einheiten der Standardunsicherheit,  $t = (\bar{x} - y)/s_{\bar{x}}$ .

Unter der Nullhypothese, dass die Einzelwerte normalverteilt sind mit Mittelwert  $\mu = y$ , folgt der Testwert  $t$  einer t-Verteilung  $f_{n-1}(t)$  mit  $n - 1$  Freiheitsgraden.

Der Ablehnungsbereich wird zweiseitig (symmetrisch) festgelegt. Der P-Wert ist:

$$P = \int_{-\infty}^{-|t|} f_{n-1}(x) dx + \int_{|t|}^{\infty} f_{n-1}(x) dx \quad (\text{B.45})$$

Die Nullhypothese wird auf dem Signifikanzniveau  $\alpha$  abgelehnt, wenn  $P \leq \alpha$  ist.

Für  $\alpha = 0,05$  und  $n \gg 1$  ergibt sich ungefähr der Ablehnungsbereich  $t \geq 2$ .

## B.5 Konfidenzintervalle

Das Ergebnis einer Messung wird häufig in der Form  $\hat{x}_o \pm \hat{\sigma}_{\hat{x}}$  angegeben. Dabei sind  $\hat{x}_o$  der beobachtete Wert des Schätzers  $\hat{x}$  und  $\hat{\sigma}_{\hat{x}}$  seine geschätzte Standardabweichung, die Standardunsicherheit. Anschaulich schätzt man dabei Mittelwert  $x$  und Standardabweichung  $\sigma_{\hat{x}}$  einer Verteilung, der die Werte  $\hat{x}$  bei identisch wiederholten Messungen folgen (würden).

Alternativ kann das Ergebnis als Konfidenzintervall  $[a, b]$  zum Konfidenzniveau  $\gamma$  angegeben werden. Das Konfidenzniveau entspricht dabei der Wahrscheinlichkeit, mit der das Konfidenzintervall den wahren Wert überdeckt. Es bleibt also eine Irrtumswahrscheinlichkeit  $\alpha = 1 - \gamma$ , mit der das Konfidenzintervall den wahren Wert nicht überdeckt.

### Allgemeine Konstruktion

Für die allgemeine Konstruktion des Konfidenzintervalls  $[a, b]$  wird vorausgesetzt, dass die Wahrscheinlichkeitsdichtefunktion der Schätzwerte  $f_x(\hat{x})$  bekannt ist, die den (unbekannten) wahren Wert  $x$  als Parameter enthält.

Die Grenzen des Konfidenzintervalls  $[a, b]$  ergeben sich dann als diejenigen hypothetischen Werte des Parameters  $x$ , für die der Anteil  $\alpha/2$  der Schätzwerte  $\hat{x}$  größer bzw. kleiner wäre als der beobachtete Wert  $\hat{x}_o$ :

$$\frac{\alpha}{2} = \int_{\hat{x}_o}^{\infty} f_a(\hat{x}) d\hat{x} = \int_{-\infty}^{\hat{x}_o} f_b(\hat{x}) d\hat{x} \quad (\text{B.46})$$

### Konfidenzintervalle für normalverteilte Schätzwerte

Folgt der Schätzwert  $\hat{x}$  einer Normalverteilung  $f_{\mu,\sigma}(\hat{x})$  mit Mittelwert  $\mu$  und (bekannter) Standardabweichung  $\sigma$ , dann vereinfachen sich die Gleichungen (B.46) zu:

$$\frac{\alpha}{2} = \int_{\hat{x}_o}^{\infty} f_{a,\sigma}(\hat{x}) d\hat{x} = 1 - F_{a,\sigma}(\hat{x}_o) = 1 - \Phi\left(\frac{\hat{x}_o - a}{\sigma}\right) \quad (\text{B.47})$$

$$\frac{\alpha}{2} = \int_{-\infty}^{\hat{x}_o} f_{b,\sigma}(\hat{x}) d\hat{x} = F_{b,\sigma}(\hat{x}_o) = \Phi\left(\frac{\hat{x}_o - b}{\sigma}\right) \quad (\text{B.48})$$

Dabei sind  $F_{\mu,\sigma}$  die kumulierte Normalverteilung mit den Parametern  $\mu, \sigma$  und  $\Phi = F_{0,1}$  die kumulierte Standardnormalverteilung mit  $\mu = 0$  und  $\sigma = 1$ .

Mit den Quantilen  $\Phi^{-1}$  lässt sich das Konfidenzintervall explizit angeben:

$$[a, b] = \hat{x}_o \pm \sigma \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (\text{B.49})$$

Für normalverteilte Schätzgrößen entspricht die Angabe  $\hat{x}_o \pm \sigma$  also dem Konfidenzintervall zum Konfidenzniveau  $\gamma \approx 0,68$ . Die mit den Erweiterungsfaktoren  $k_\gamma = 2, 3$  gebildeten Intervalle  $\hat{x}_o \pm k_\gamma \cdot \sigma$  entsprechen den Konfidenzniveaus  $\gamma \approx 0,95$  bzw.  $\gamma \approx 0,997$ .

Normalerweise wird die unbekannte Standardabweichung  $\sigma$  aus der Stichprobe geschätzt. Bei der Berechnung der Intervallgrenzen nach (B.49) müssen die Quantile der Standardnormalverteilung  $\Phi^{-1}$  dann durch die Quantile der t-Verteilung ersetzt werden.

## B.6 Fehlerfortpflanzung

### Wahrscheinlichkeitsdichte der Zielgröße

Wir betrachten eine Zufallsvariable  $x$  (Eingangsgröße), die entsprechend der Wahrscheinlichkeitsdichte  $f(x)$  verteilt ist. Mit der Funktion  $z = z(x)$  berechnet man daraus eine neue Variable  $z$  (Zielgröße). Dabei ergibt sich für  $z$  eine Wahrscheinlichkeitsdichte  $f(z)$ .

Für eine lineare Funktion  $z = a + bx$  bleibt dabei die Form der Verteilung erhalten, d.h. für  $z$  gilt die gleiche Wahrscheinlichkeitsdichte wie für  $x$ , abgesehen von einer Verschiebung um  $a$  und Skalierung mit  $b$ .

Für eine nichtlineare Funktion  $z \neq a + bx$  bleibt die Form der Verteilung näherungsweise erhalten, wenn die Verteilung  $f(x)$  so schmal ist, dass im Bereich der meisten Werte, also etwa  $\mu_x \pm \sigma_x$ , eine lineare Näherung akzeptabel ist.

Falls diese Voraussetzung nicht erfüllt ist, kann die Verteilung  $f(z)$  numerisch durch eine Monte-Carlo-Simulation berechnet und dann statistisch analysiert werden.

### Mittelwert und Varianz der Zielgröße

In vielen Fällen reduziert man die Beschreibung auf die wichtigsten Kennzahlen der Zielgröße. Ohne Kenntnis der Verteilung  $f(z)$  ergibt sich der Mittelwert von  $z$  als

$$\mu_z = \int z(x) f(x) dx \quad (\text{B.50})$$

und die Varianz von  $z$  als:

$$\sigma_z^2 = \int [z(x) - \mu_z]^2 f(x) dx \quad (\text{B.51})$$

Die lineare Näherung (Taylor-Entwicklung 1. Ordnung) der Funktion  $z = z(x)$  um  $\mu_x$  führt auf die bekannten Formeln für  $\mu_z$  und  $\sigma_z$ . Für nichtlineare Funktionen können durch Taylor-Entwicklung in höheren Ordnungen bessere Näherungen erreicht werden.

### Mehrere unabhängige Variablen

Betrachtet man mehrere nach  $f(x)$  verteilte unabhängige Zufallsvariablen  $x_1 \dots x_n$  und eine Funktion  $z = z(x_1 \dots x_n)$ , dann hat die Verteilung  $f(z)$  im Allgemeinen eine andere Form als  $f(x)$ . Dies gilt schon für eine einfache Addition von zwei Zufallszahlen,  $z = x_1 + x_2$ .

Die Normalverteilung spielt dabei eine besondere Rolle. Für zwei unabhängige Zufallszahlen  $x_1$  und  $x_2$ , die normalverteilt sind mit Mittelwerten  $\mu_{1,2}$  und Varianzen  $\sigma_{1,2}^2$ , ist die Summe  $z = x_1 + x_2$  normalverteilt mit Mittelwert  $\mu_1 + \mu_2$  und Varianz  $\sigma_1^2 + \sigma_2^2$ .

Allgemein gilt der zentrale Grenzwertsatz: Für unabhängige Zufallsvariablen  $x_1 \dots x_n$ , die einer (beliebigen) Verteilung  $f(x)$  mit Mittelwert  $\mu$  und Varianz  $\sigma^2$  folgen, ist die Summe  $z = \sum x_i$  im Grenzfall  $n \rightarrow \infty$  normalverteilt mit Mittelwert  $n\mu$  und Varianz  $n\sigma^2$ .

## Zwei korrelierte Variablen

Wir betrachten zwei Eingangsgrößen  $x, y$  mit den Mittelwerten  $\mu_x, \mu_y$  und Varianzen  $\sigma_x^2, \sigma_y^2$  unter Berücksichtigung einer möglichen Kovarianz  $\sigma_{xy}$  bzw. Korrelation  $\varrho_{xy}$ .

Im Rahmen der linearen Näherung ergibt sich für die Zielgröße  $z = z(x, y)$  der Mittelwert:

$$\mu_z = z(\mu_x, \mu_y) \quad (\text{B.52})$$

Für die Varianz der Zielgröße  $z$  erhält man:

$$\sigma_z^2 = \left[ \frac{\partial z}{\partial x}(\mu_x, \mu_y) \right]^2 \sigma_x^2 + \left[ \frac{\partial z}{\partial y}(\mu_x, \mu_y) \right]^2 \sigma_y^2 + 2 \frac{\partial z}{\partial x}(\mu_x, \mu_y) \frac{\partial z}{\partial y}(\mu_x, \mu_y) \sigma_{xy} \quad (\text{B.53})$$

Dabei beschreibt der dritte Summand

$$2 \frac{\partial z}{\partial x}(\mu_x, \mu_y) \frac{\partial z}{\partial y}(\mu_x, \mu_y) \sigma_{xy} = 2 \frac{\partial z}{\partial x}(\mu_x, \mu_y) \frac{\partial z}{\partial y}(\mu_x, \mu_y) \sigma_x \sigma_y \varrho_{xy} \quad (\text{B.54})$$

den Einfluss der Korrelation zwischen  $x$  und  $y$ . Je nach Korrelation und Funktionszusammenhang kann dieser Term die Varianz der Zielgröße  $\sigma_z^2$  vergrößern oder verkleinern. Die einfache Gaußsche Fehlerfortpflanzung (A.25) ist also nur dann korrekt, wenn die Eingangsgrößen unkorreliert sind,  $\varrho_{xy} = 0$ .

## Mehr als zwei korrelierte Variablen, Matrixschreibweise

Das Ergebnis lässt sich leicht auf mehr als zwei Variablen verallgemeinern. Für  $n$  Eingangsgrößen  $x = (x_1 \dots x_n)$  seien die Mittelwerte  $\mu = (\mu_1 \dots \mu_n)$ , die Varianzen  $\sigma_1^2 \dots \sigma_n^2$  und die Kovarianzen  $\sigma_{ij}$  ( $i \neq j$ ) bzw. die Kovarianzmatrix  $V_x$  bekannt.

Mit dem bei  $x = \mu$  ausgewerteten (Zeilen-)Vektor der Ableitungen

$$A = \left[ \frac{\partial z}{\partial x_i} \right] (\mu) \quad (\text{B.55})$$

erhält man die Varianz der Zielgröße in Matrixschreibweise:

$$\sigma_z^2 = A V_x A^T \quad (\text{B.56})$$

Für  $m > 1$  Zielgrößen  $z_1 \dots z_m$  erhält man analog mit der bei  $x = \mu$  ausgewerteten Matrix der Ableitungen (Jacobi-Matrix)

$$A = \left[ \frac{\partial z_i}{\partial x_j} \right] (\mu) \quad (\text{B.57})$$

die Kovarianzmatrix der Zielgrößen:

$$V_z = A V_x A^T \quad (\text{B.58})$$

Diese enthält die Varianzen und Kovarianzen für alle Zielgrößen  $z_i$  unter Berücksichtigung eventueller Korrelationen zwischen den Eingangsgrößen.

## B.7 Lineare Regression

Die Anpassung einer Ausgleichsgeraden  $y = a + bx$  stellt nur einen Spezialfall der linearen Regression dar. Für die allgemeine Formulierung nehmen wir an, dass die Messwerte  $(x_i, y_i)$  und Standardunsicherheiten  $\Delta y_i$  bekannt sind ( $i = 1 \dots n$ ). Die  $x$ -Werte werden als exakt behandelt.

### Anpassung mit Gewichtung

Allgemein lässt sich die  $\chi^2$ -Minimierung (B.38) für alle Modelle  $y = m_p(x)$  analytisch lösen, die in den Parametern  $p_1 \dots p_k$  linear sind:

$$y = m_p(x) = \sum \phi_j(x) \cdot p_j, \quad j = 1 \dots k \quad (\text{B.59})$$

Mit den Vektoren  $y = [y_i]$ ,  $p = [p_j]$  und der Matrix  $M = [\phi_j(x_i)]$  entspricht dies der Gleichung  $y = Mp$ . Da es normalerweise weniger Parameter als Messwerte gibt,  $k < n$ , existiert keine exakte Lösung. Die beste Näherungslösung ergibt sich aus der  $\chi^2$ -Minimierung, wobei die Unsicherheiten  $\Delta y$  durch eine geeignete Gewichtung berücksichtigt werden.

Mit  $V_y = \text{diag}[\Delta y_i^2]$  erhält man folgende Gleichung für die besten Parameter-Schätzwerte:

$$\hat{p} = (M^T V_y^{-1} M)^{-1} M^T V_y^{-1} y \quad (\text{B.60})$$

Durch Fehlerfortpflanzung (B.58) ergibt sich die Parameter-Kovarianzmatrix:

$$V_p = (M^T V_y^{-1} M)^{-1} \quad (\text{B.61})$$

### Anpassung ohne Gewichtung

Wenn die Unsicherheiten  $\Delta y_i$  nicht bekannt sind, bestimmt man zunächst mit  $V_y = \text{diag}[1]$  die Parameter-Bestwerte  $\hat{p}$  ohne Gewichtung, entsprechend der Minimierung von  $S$  (B.39).

Mit der Streuung der Messwerte um das angepasste Modell

$$s = \sqrt{\frac{1}{n-k} \sum [y_i - m_{\hat{p}}(x_i)]^2} \quad (\text{B.62})$$

setzt man dann  $V_y = \text{diag}[s^2]$  und berechnet damit die Parameter-Kovarianzmatrix  $V_p$ .

### Konfidenzband

Das (angepasste) Modell  $y = m_{\hat{p}}(x)$  ordnet jedem  $x$ -Wert einen  $y$ -Wert zu. Die Unsicherheiten der Parameter, die durch die Kovarianzmatrix  $V_p$  gegeben sind, führen nach (B.56) zu einer Unsicherheit des Modells:

$$\Delta m_p(x) = \sqrt{A V_p A^T} \quad (\text{B.63})$$

Dabei enthält der (Zeilen-)Vektor der Ableitungen wegen (B.59) einfach die Funktionen  $\phi_j$ :

$$A = [\phi_j(x)] \quad (\text{B.64})$$

Für jeden  $x$ -Wert resultiert ein Intervall  $m_{\hat{p}}(x) \pm \Delta m_p(x)$ , insgesamt also ein Bereich um die Ausgleichsfunktion, der näherungsweise dem 68%-Konfidenzband entspricht.

## B.8 Numerische Minimierung

Das allgemeinste Verfahren zur Anpassung von parametrisierten Modellen an Messdaten ist die numerische Minimierung von  $\chi^2$ .

Wir betrachten wieder Messwerte  $(x_i, y_i)$  mit  $i = 1 \dots n$ . Die Standardunsicherheiten  $\Delta y_i$  seien bekannt, die  $x$ -Werte werden als exakt behandelt. Das anzupassende Modell sei  $y = m_p(x)$  mit Parameter(n)  $p_j$ ,  $j = 1 \dots k$ .

### Parameter, Unsicherheiten, Kovarianzen

Die Parameter-Schätzwerte  $\hat{p}$  findet man durch numerische Minimierung von  $\chi^2(p)$  (B.38).

Die Parameter-Kovarianzmatrix  $V_p$  bestimmt man entsprechend (B.31) numerisch aus der Krümmung von  $\chi^2(p)$  in der Umgebung des Minimums bei  $p = \hat{p}$ :

$$V_p = 2 \left[ \frac{\partial^2 \chi^2}{\partial p_i \partial p_j} \right]_{p=\hat{p}}^{-1} \quad (\text{B.65})$$

Für einen einzelnen Parameter ( $k = 1$ ) entspricht das der Zunahme von  $\chi^2$  um 1 (B.32):

$$\chi^2(\hat{p} \pm \Delta p) = \chi^2(\hat{p}) + 1 \quad (\text{B.66})$$

### Anpassungsgüte

Die Anpassungsgüte ist ein (quantitatives) Maß dafür, wie gut das angepasste Modell die experimentellen Daten beschreibt. Die Definition (B.38) legt nahe, dass der minimale Wert  $\chi^2(\hat{p})$  ein geeignetes Maß für die Anpassungsgüte darstellt, denn die Summanden entsprechen den quadrierten Abweichungen vom Modell in Einheiten der Standardunsicherheiten.

Wenn die Abweichungen vom Modell  $y = m_{\hat{p}}(x)$  normalverteilt sind mit  $\mu = 0$  und  $\sigma_i = \Delta y_i$ , dann folgt  $\chi^2$  einer Chi-Quadrat-Verteilung mit  $n - k$  Freiheitsgraden.

Qualitativ bedeutet also  $\chi^2(\hat{p}) \approx n - k$  eine akzeptable Anpassung, während  $\chi^2(\hat{p}) \gg n - k$  auf eine signifikante Abweichung zwischen Daten und Modell hinweist. Quantitativ kann ein P-Wert (B.44) berechnet und ein Chi-Quadrat-Test durchgeführt werden.