

Einführung in die Statistik

Andreas Eberle

15. September 2024

Inhaltsverzeichnis

| | |
|--|------------|
| Inhaltsverzeichnis | 2 |
| 1. Statistische Verfahren: Grundbegriffe und Beispiele | 1 |
| 1.1. Zwei Hypothesentests | 1 |
| 1.2. Schätzen von Populationsgrößen | 5 |
| 1.3. Statistische Modelle und Verfahren | 9 |
| 1.4. Statistische Verfahren | 11 |
| 2. Likelihood | 20 |
| 2.1. Das Maximum-Likelihood-Prinzip | 20 |
| 2.2. Suffiziente Statistiken | 23 |
| 2.3. Exponentielle Familien | 26 |
| 2.4. Likelihood-Quotienten-Tests | 30 |
| 2.5. Studentsche Konfidenzintervalle und t-Test | 36 |
| 3. Relative Entropie, Information und statistische Unterscheidbarkeit | 42 |
| 3.1. Entropie und relative Entropie | 42 |
| 3.2. Anwendungen in der Statistik | 48 |
| 3.3. Fisher-Information | 51 |
| 3.4. Weitere Anwendungen von Entropie und relativer Entropie | 59 |
| 4. Empirische Verteilungen | 63 |
| 4.1. Empirische Verteilungen | 63 |
| 4.2. Plug-in-Schätzer und Bootstrap | 67 |
| 4.3. Anpassungstests | 72 |
| 4.4. Empirische Verteilungen numerischer Merkmale | 78 |
| 4.5. Robuste Verfahren | 82 |
| 4.6. Graphische Überprüfung von Verteilungsfunktionen | 89 |
| 5. Zusammenhang mehrerer Merkmale | 91 |
| 5.1. Binäre Merkmale: Chancenquotienten und Vierfeldertafeln | 91 |
| 5.2. Test auf Unabhängigkeit | 94 |
| 5.3. Permutationstests | 97 |
| 6. Regression | 101 |
| 6.1. Einfache Lineare Regression | 102 |
| 6.2. Lineare Modelle | 105 |
| 6.3. Andere Regressionsverfahren | 111 |
| 7. Bayes-Statistik | 114 |
| 7.1. Ansatz der Bayesschen Statistik | 114 |
| 7.2. Markov Chain Monte Carlo (MCMC)-Verfahren | 118 |

| | |
|---|------------|
| A. Ergänzungen aus der Wahrscheinlichkeitstheorie | 121 |
| A.1. Kovarianz, Korrelation und lineare Prognosen | 121 |
| A.2. Wahrscheinlichkeitsverteilungen im \mathbb{R}^n | 125 |
| A.3. Charakteristische Funktionen und mehrdimensionaler zentraler Grenzwertsatz | 129 |

1. Statistische Verfahren: Grundbegriffe und Beispiele

1.1. Zwei Hypothesentests

a) STILLES MINERALWASSER ODER LEITUNGSWASSER?

Eine Person behauptet, dass sie anhand des Geschmacks unterscheiden kann, ob es sich um stilles Mineralwasser oder Leitungswasser handelt. Um diese Behauptung zu überprüfen, wird der folgende Test durchgeführt:

Dem Probanden werden insgesamt $2l$ Gläser präsentiert, wovon l mit Leitungswasser und die restlichen l mit stillem Mineralwasser gefüllt sind. Der Proband wählt l Gläser aus, von denen er annimmt, dass sie Leitungswasser enthalten. Sei X die Anzahl der korrekt ausgewählten Gläser.

Das Ziel des Tests ist es zu zeigen, ob die Person tatsächlich über die behauptete Fähigkeit verfügt. Es ist jedoch einfacher, die Nullhypothese zu widerlegen, die besagt, dass die Person keinen Unterschied erkennen kann. Unter der Annahme der Nullhypothese ist die Verteilung von X hypergeometrisch. Wir möchten herausfinden, wie groß X sein muss, damit die Nullhypothese H_0 verworfen werden kann. Für einen beobachteten Wert x betrachten wir den p -Wert.

$$p := \mathbb{P}_0 [X \geq x] = 1 - \mathbb{P}_0 [X \leq x - 1] = 1 - F_{2l,l}(x - 1)$$

Bei gegebenem *Signifikanzniveau* α ergibt sich die Entscheidungsregel:

Verwerfe Nullhypothese H_0 falls $p \leq \alpha \Leftrightarrow F_{2l,l}(x - 1) \geq 1 - \alpha$.

Beispiel. Betrachten wir den beschriebenen Test mit $l = 5$. Bei $x = 4$ korrekt ausgewählten Gläsern ergibt sich ein p -Wert von

$$p = \mathbb{P}_0 [X \geq 4] = \frac{\binom{5}{4} \binom{5}{1}}{\binom{10}{5}} + \frac{\binom{5}{5} \binom{5}{0}}{\binom{10}{5}} = \frac{26}{252} > 0,1$$

Folglich können wir H_0 nicht zum Signifikanzniveau 10% verwerfen. Für den Fall $l = 5, x = 5$ ergibt sich $p = \mathbb{P}_0 [X \geq 5] = \frac{1}{252} \approx 0,4\%$. Somit können wir H_0 sogar zum Signifikanzniveau 0,5% verwerfen.

Bemerkung. 1) BEDEUTUNG DES SIGNIFIKANZNIVEAUS Sei beispielsweise $\alpha = 5\%$. Bei 100 Wiederholungen desselben Experiments würden wir unter H_0 durchschnittlich höchstens 5 mal einen so hohen p -Wert sehen, dass wir H_0 tatsächlich verwerfen.

2) Wenn wir H_0 nicht signifikant verwerfen können, bedeutet das nicht, dass wir davon ausgehen sollten, dass H_0 wahr ist. Möglicherweise haben wir nur zu wenige Daten, um eine Entscheidung zu treffen.

Beispiel. Sei nun jedes Glas unabhängig, zufällig mit Wahrscheinlichkeit $\frac{1}{2}$ mit Mineral- bzw. Leitungswasser gefüllt. Unter H_0 gilt dann $X \sim \text{Bin}(2l, \frac{1}{2})$. Wir erhalten für $l = 5$ und $x = 4$:

$$p = \mathbb{P}_0 [X \geq 4] = \frac{\binom{5}{4} + \binom{5}{5}}{2^5} = \frac{6}{32} \approx 0,19$$

1. Statistische Verfahren: Grundbegriffe und Beispiele

Einen nicht signifikanten p -Wert für $\alpha = 10\%$. Andererseits ist $l = 5, x = 5$ mit

$$p = \mathbb{P}_0 [X \geq 5] = \frac{1}{32} \approx 3\%$$

Schon signifikant zum Wert $\alpha = 5\%$.

b) FISCHERS EXAKTER TEST

Angenommen man möchte die Wirksamkeit eines Medikaments nachweisen. Um dies gegebenenfalls zu zeigen, werden $N = n_1 + n_2$ Probanden zufällig in zwei Gruppen eingeteilt: Die n_1 Individuen der 1. Gruppe erhalten das Medikament, die n_2 Individuen in Gruppe 2 erhalten ein Placebo. Nach einer gewissen Zeit wird ermittelt, wie viele Behandlungserfolge und -misserfolge in den beiden Gruppen auftraten. Die Ergebnisse lassen sich als *Viefeldertafel* zusammenfassen.

| | | | |
|------------|-------------------|-------------|-------|
| | Erfolg | Misserfolg | |
| Medikament | H_1 | $n_1 - H_1$ | n_1 |
| Placebo | H_2 | $n_2 - H_2$ | n_2 |
| | $H_+ = H_1 + H_2$ | $N - H_+$ | N |

Unter der Nullhypothese H_0 , dass das Medikament genau wie das Placebo wirkt (kein Effekt), sind die Verteilungen von H_1, H_2 und H_+ unbekannt. Jedoch gilt unter H_0 , dass die bedingte Verteilung von H_1 gegeben $H_+ = l$ hypergeometrisch verteilt ist: $H_1 | H_+ = l \sim \text{Hyp}(N, l, n_1)$. Damit gilt:

$$\mathbb{P}_0 [H_1 > q \mid H_+ = l] = 1 - F_{N,l,n_1}(q) \leq \alpha$$

Die Frage ist nun, wie groß H_1 sein sollte, damit wir die Nullhypothese verwerfen können. Zu diesem Zweck fixieren wir ein Testniveau $\alpha \in (0, 1)$ und betrachten den kleinsten Wert von q , welcher die obige Ungleichung erfüllt:

$$q_{1-\alpha} = \min \{x \in \mathbb{R} : F_{N,l,n_1}(x) \geq 1 - \alpha\}$$

Im Falle von $H_1 > q_{1-\alpha}$ können wir mit einer Sicherheit von $1 - \alpha$ behaupten, dass die Nullhypothese nicht stimmt. Dies motiviert die Entscheidungsregel:

Verwerfe H_0 zum Signifikanzniveau α falls $H_1 > q_{1-\alpha}$, wobei l der realisierte Wert von H_+ ist.

Beispiel. FISCHERS EXAKTER TEST Um die Wirksamkeit eines Medikamentes zu testen wird wie beschrieben ein Test durchgeführt mit $N = 40$ Probanden, $n_1 = n_2 = 20$ und einem Signifikanzniveau von $\alpha = 5\%$. Bei beobachteten Häufigkeiten von $H_+ = 26, H_1 = 15, H_2 = 11$ und dem 95%-Quantil $q_{0,95;40,26,20} = 15$ kann die Nullhypothese nicht verworfen werden, da $H_1 \not> 15$. Folglich kann keine Aussage getroffen werden.

Quantile

Quantile sind Punkte, an denen die Verteilungsfunktion einen bestimmten Wert überschreitet. Diese Punkte sind besonders in praktischen Anwendungen wie der Qualitätskontrolle von Bedeutung. Quantile ermöglichen es, verallgemeinerte Umkehrfunktionen der im Allgemeinen nicht bijektiven Verteilungsfunktion zu definieren.

Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion

$$F(c) = \mu [(-\infty, c]], \quad c \in \mathbb{R}.$$

Die Funktion F ist monoton wachsend und rechtsstetig. Den linksseitigen Limes bezeichnen wir als $F(c_-) = \lim_{\varepsilon \downarrow 0} F(c - \varepsilon) = \mu [(-\infty, c)]$.

Definition 1.1 (Quantile). Sei $u \in [0, 1]$. Ein u -Quantil $q \in \mathbb{R}$ der Wahrscheinlichkeitsverteilung μ erfüllt die Bedingungen:

$$F(q_-) = \mu[(-\infty, q)] \leq u \quad \text{und} \quad F(q) = \mu[(-\infty, q]] \geq u.$$

Ein *Median* ist ein $\frac{1}{2}$ -Quantil.

Wenn die Verteilungsfunktion streng monoton wachsend ist, ist $q = F^{-1}(u)$ für $u \in (0, 1)$ das eindeutige u -Quantil. Im Allgemeinen kann es jedoch mehrere u -Quantile geben, die denselben Wert u haben. Wir definieren nun zwei verallgemeinerte Inverse einer Verteilungsfunktion F , da diese im Allgemeinen nicht bijektiv ist. Für $u \in (0, 1)$ sei

$$\begin{aligned} \underline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) \geq u\} = \sup\{x \in \mathbb{R} : F(x) < u\}, & \text{und} \\ \overline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) > u\} = \sup\{x \in \mathbb{R} : F(x) \leq u\}. \end{aligned}$$

Wie die folgende Abbildung zeigt, sind Quantile im Allgemeinen nicht eindeutig.

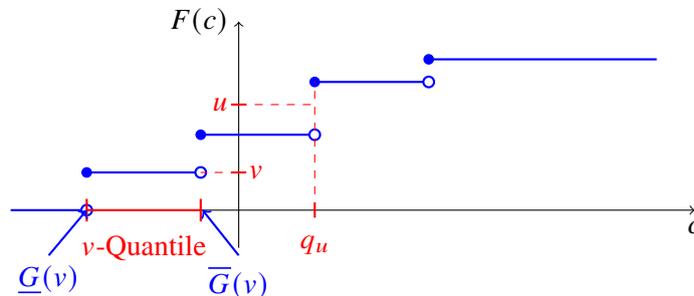


Abbildung 1.1.: Quantilillustration mit Verteilungsschritten und Quantilen.

Offensichtlich gilt $\underline{G}(u) \leq \overline{G}(u)$. Die Funktionen \underline{G} bzw. \overline{G} sind links- bzw. rechtsstetig. Ist F stetig und streng monoton wachsend, also eine Bijektion von \mathbb{R} nach $(0, 1)$, dann ist $\underline{G}(u) = \overline{G}(u) = F^{-1}(u)$. Die Funktion \underline{G} wird daher auch als die *linksstetige verallgemeinerte Inverse* von F bezeichnet. Das folgende Lemma zeigt, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u -Quantil ist:

Lemma 1.2. Für $u \in (0, 1)$ und $q \in \mathbb{R}$ sind die folgenden Aussagen äquivalent:

- (i) q ist ein u -Quantil.
- (ii) $F(q_-) \leq u \leq F(q)$.
- (iii) $\underline{G}(u) \leq q \leq \overline{G}(u)$.

Beweis. Nach Definition ist q genau dann ein u -Quantil, wenn $P[X < q] \leq u \leq 1 - P[X > q] = P[X \leq q]$ gilt. Hieraus folgt die Äquivalenz von (i) und (ii).

Um zu zeigen, dass (iii) äquivalent zu diesen Bedingungen ist, müssen wir zeigen, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u -Quantil ist. Wir bemerken zunächst, dass $\underline{G}(u)$ ein u -Quantil ist, da

$$F(\underline{G}(u)-) = \lim_{x \nearrow \underline{G}(u)} F(x) \leq u, \quad \text{und} \quad F(\underline{G}(u)) = \lim_{x \searrow \underline{G}(u)} F(x) \geq u.$$

Andererseits ist $x < \underline{G}(u)$ kein u -Quantil, denn es gilt $F(x) < u$. Somit ist $\underline{G}(u)$ das kleinste u -Quantil. Auf ähnliche Weise folgt, dass $\overline{G}(u)$ das größte u -Quantil ist. ■

Satz 1.3 (Quantiltransformation). Sei $U \sim \text{Unif}(0, 1)$ eine auf $(0, 1)$ gleichverteilte Zufallsvariable. Dann ist durch

$$X := \underline{G}(u)$$

eine Zufallsvariable mit Verteilung μ definiert.

Beweis (Beweisskizze). Sei $c \in \mathbb{R}$ und F_X , bzw. F_μ die jeweiligen Verteilungsfunktionen. Da Verteilungen eindeutig über ihre Verteilungsfunktionen bestimmt sind, reicht es, deren Gleichheit zu zeigen:

$$F_X(c) = \mathbb{P}[\underline{G}(u) \leq c] = \mathbb{P}[U \leq F_\mu(c)] = F_\mu(c).$$

Die mittlere Gleichheit ist eine Übungsaufgabe. ■

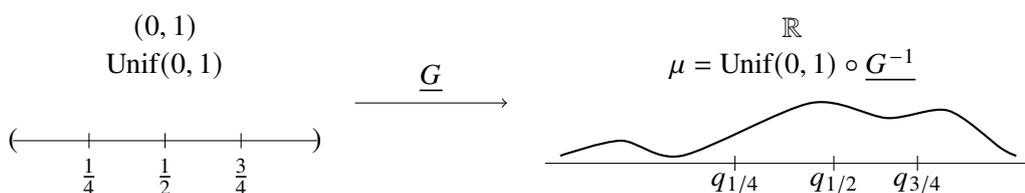


Abbildung 1.2.: Quantiltransformation

ANWENDUNG Da Stichproben der Gleichverteilung auf $(0, 1)$ einfacher zu simulieren sind, ist die Quantiltransformation besonders hilfreich bei der Simulation von Stichproben einer Wahrscheinlichkeitsverteilung μ . Bei gegebenen unabhängigen Stichproben u_1, \dots, u_n von $\text{Unif}(0, 1)$ erhält man durch Quantiltransformation unabhängige Stichproben $\underline{G}(u_1), \dots, \underline{G}(u_n)$ von μ .

Sei u nun eine Stichprobe der Gleichverteilung $\text{Unif}(0, 1)$, dann erhalten wir beispielsweise Stichproben aus gegebenen Verteilungen wie

$$\text{Bernoulli}(p) \sim 1_{u > 1-p}, \quad \text{Exp}(\lambda) \sim -\frac{1}{\lambda} \log(1 - u).$$

***p*-Werte**

Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir beobachten den Wert $x = X(\omega)$. Nun stellt sich die Frage, ob X tatsächlich der Verteilung folgt oder ob der beobachtete Wert „verdächtig klein“ bzw. „verdächtig groß“ ist. Dies motiviert die folgende Definition:

Definition 1.4 (*p*-Wert). Sei F die Verteilungsfunktion der Zufallsvariable X . Dann definieren wir

$$\begin{aligned} p_l &:= \mathbb{P}[X \leq x] = F(x), && \text{linksseitiger } p\text{-Wert,} \\ p_r &:= \mathbb{P}[X \geq x] = 1 - F(x_-), && \text{rechtsseitiger } p\text{-Wert,} \\ p &:= 2 \min(p_l, p_r), && \text{beidseitiger } p\text{-Wert.} \end{aligned}$$

Kleine *p*-Werte sprechen dafür, dass X nicht gemäß \mathbb{P} verteilt ist. Dies wird im folgenden Lemma präzisiert:

Lemma 1.5. Sei $\alpha \in [0, 1]$. Dann gilt für die drei *p*-Werte:

- (i) $\mathbb{P}[p_l \leq \alpha] = \mathbb{P}[F(X) \leq \alpha] \leq \alpha,$
(ii) $\mathbb{P}[p_r \leq \alpha] = \mathbb{P}[1 - F(X_-) \leq \alpha] \leq \alpha,$
(iii) $\mathbb{P}[p \leq \alpha] = \mathbb{P}[2 \min(F(X), 1 - F(X_-)) \leq \alpha] \leq \alpha.$

Für den Fall, dass F bijektiv ist, gilt jeweils Gleichheit.

Beweis. Sei \underline{G} die verallgemeinerte inverse Funktion von F , $U \sim \text{Unif}(0, 1)$ und wir definieren die Zufallsvariable $X = \underline{G}(U)$. Dann gilt:

(i)

$$\mathbb{P}[F(X) \leq \alpha] = \mathbb{P}[F(\underline{G}(U)) \leq \alpha] \leq \mathbb{P}[U \leq \alpha] = \alpha$$

Da $F(\underline{G}(U)) \geq U$ gemäß der Definition von \underline{G} und der Rechtsstetigkeit von F .

(ii)

$$\mathbb{P}[1 - F(X_-) \geq 1 - \alpha] = \mathbb{P}[F(\underline{G}(U)_-) \geq 1 - \alpha] \geq \mathbb{P}[U \leq 1 - \alpha] = \alpha$$

Da $F(\underline{G}(U)_-) \leq U$ gemäß der Definition von \underline{G} .

(iii) Folgt aus (i) und (ii), da:

$$\min(F(X), 1 - F(X_-)) \leq \frac{\alpha}{2} \Rightarrow F(X) \leq \frac{\alpha}{2} \text{ oder } 1 - F(X_-) \leq \frac{\alpha}{2}$$

Nach (i) und (ii) ist die Wahrscheinlichkeit für beide Ausdrücke $\leq \frac{\alpha}{2}$ und somit für den Gesamtdruck $\leq \alpha$. ■

Bemerkung (Bedeutung des p -Wertes). Wenn das Zufallsexperiment oft wiederholt wird und jedes Mal der p -Wert berechnet wird, dann sehen wir nur selten einen kleinen p -Wert. Zum Beispiel für $\alpha = 0,05$ nur in durchschnittlich 5 von 100 Fällen. Ein kleiner p -Wert legt daher nahe, die Verteilungshypothese zu verwerfen.

Beispiel (Fishers exakter Test). Betrachten wir erneut den oben beschriebenen Test zur Prüfung der Wirksamkeit eines Medikaments. Wir hatten gesehen, dass wir die Nullhypothese verwerfen können, falls

$$H_1 > q_{1-\alpha; N, H_+, n_1} \stackrel{\text{Übung}}{\Leftrightarrow} \underbrace{1 - F_{N, H_+, n_1}(H_{1-})}_{p_r} \leq \alpha.$$

Nach Lemma 1.5 folgt: Die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, obwohl sie wahr ist, beträgt höchstens α .

1.2. Schätzen von Populationsgrößen

In vielen statistischen Anwendungen analysiert man Stichproben aus einer bestimmten Population. Ziel ist es, anhand der Stichprobe Rückschlüsse auf die Zusammensetzung der gesamten Population zu ziehen. Die Frage nach der Größe der Population ist dabei besonders nützlich für verschiedene statistische Verfahren, wie das folgende illustrierte „Taxiproblem“ zeigt.

Taxiproblem

In einer Stadt gibt es eine unbekannte Anzahl N an Taxis, die von 1 bis N durchnummeriert sind. Um die Anzahl der Taxis zu bestimmen, beobachten wir n verschiedene Taxis mit den Nummern $\omega_1, \dots, \omega_n$. Konkret lässt sich das Beispiel durch das folgende Modell darstellen: Sei

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \mathbb{N}, \omega_i \neq \omega_j \text{ für } i \neq j\} \subset \mathbb{N}^n$$

mit der Stichprobe $X(\omega) = \omega$ und den Stichprobenwerten $X_i(\omega) = \omega_i$. Abhängig vom unbekanntem Parameter N definieren wir die Wahrscheinlichkeitsverteilung $\mathbb{P}_N = \text{Unif}(\Omega_N)$, wobei $\Omega_N = \{\omega \in \Omega : \omega_i \leq N \forall i \in \{1, \dots, n\}\}$.

Im Folgenden betrachten wir zwei Schätzer für den unbekanntem Parameter N . Nach dem Gesetz der großen Zahlen konvergiert das arithmetische Mittel \bar{X}_n der Stichprobenwerte für große n gegen $(N + 1)/2$. Daher ist der Schätzer

$$Y = 2 \cdot \frac{X_1 + \dots + X_n}{n} - 1$$

naheliegender. Ein weiterer interessanter Schätzer ist die Zahl $M = \max(X_1, \dots, X_n)$. Um diese beiden Schätzer zu vergleichen, betrachten wir ihre Nähe zum unbekanntem Parameter N .

Zu Y : Mit dem Erwartungswert $\mathbb{E}_N[X_1] = (N + 1)/2$ folgt

$$\mathbb{E}_N[Y] = 2\mathbb{E}_N[X_1] - 1 = N.$$

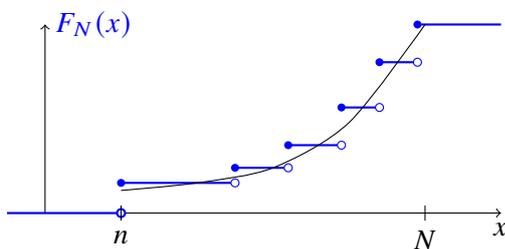
Man sagt, dass der Schätzer Y *erwartungstreu* ist. Ein weiteres gängiges Maß für die Ungenauigkeit eines Schätzers ist der mittlere quadratische Fehler (MSE):

$$\mathbb{E}_N[(Y - N)^2] = \text{Var}_N[Y] = \frac{4}{n^2} \text{Var}[X_1 + \dots + X_n] \approx \frac{4n}{n^2} \text{Var}_N[X_1] \quad \text{für } n \ll N,$$

was in $O\left(\frac{N^2}{n}\right)$ liegt, falls N deutlich größer als n ist. Die Ungenauigkeit von Y nimmt also linear mit n ab.

Zu M : Nach Definition des Maximumschätzers gilt $\mathbb{E}_N[M] < N$ für alle $N > 1$. Der Schätzer M ist *nicht erwartungstreu*. Um eine genauere Fehlerabschätzung zu erhalten, ist die Berechnung der Verteilung von M unter \mathbb{P}_N erforderlich. Für $n \leq x \leq N$ gilt dann

$$F_N(x) := \mathbb{P}_N[M \leq x] = \frac{x(x-1)(x-2) \dots (x-n+1)}{N(N-1)(N-2) \dots (N-n+1)} =: \frac{[x]_n}{[N]_n}. \quad (*)$$



Lemma 1.6. Seien N, n, M wie oben. Dann gilt:

(i) $\mathbb{E}_N[M] = \frac{n}{n+1}(N+1) = N - \frac{N-n}{n+1} \in O\left(\frac{N}{n}\right)$

(ii) $\text{Var}_N[M] = \frac{n(N-n)(N+1)}{(n+1)^2(n+2)} \in O\left(\frac{N^2}{n^2}\right)$

Beweis. Nach (*) gilt

$$\mathbb{P}_N[M = x] = F_N(x) - F_N(x-1) \stackrel{(*)}{=} \frac{n[x-1]_{n-1}}{[N]_n}.$$

Somit folgt

$$\mathbb{E}_N[M] = \sum_{x=n}^N x \mathbb{P}_N[M = x] = \frac{n}{[N]_n} \sum_{x=n}^N [x-1]_{n-1} = \frac{n}{n+1}(N+1),$$

da

$$\sum_{x=n}^N \mathbb{P}[M = x] = 1 \Rightarrow \sum_{x=n}^N [x-1]_{n-1} = \frac{[N]_n}{n}.$$

Analog erhält man

$$\mathbb{E}_N[M(M+1)] = \frac{n}{n+2}(N+1)(N+2) \Rightarrow \text{Var}_N[M] = \mathbb{E}_N[M^2] - (\mathbb{E}_N[M])^2 = \dots = \frac{n(N-n)(N+1)}{(n+1)^2(n+2)}.$$

Die Varianz von M fällt schneller ab als die von Y , aber M ist nicht erwartungstreu. Daher betrachten wir den folgenden modifizierten Schätzer:

Korollar 1.7. Für den modifizierten Maximumschätzer

$$\hat{N} := \frac{n+1}{n}M - 1$$

gilt:

$$\mathbb{E}_N[\hat{N}] = N \quad \text{und} \quad \mathbb{E}_N[(\hat{N} - N)^2] < \frac{N^2}{n^2}.$$

Beweis. Nach Definition von \hat{N} gilt

$$\mathbb{E}_N[\hat{N}] = \frac{n+1}{n} \mathbb{E}_N[M] - 1 \stackrel{(1.6)}{=} N.$$

Für die mittlere quadratische Abweichung erhalten wir somit

$$\mathbb{E}_N[(\hat{N} - N)^2] = \text{Var}_N[\hat{N}] = \left(\frac{n+1}{n}\right)^2 \text{Var}_N[M] \stackrel{(1.6)}{=} \frac{(N-n)(N+1)}{n(n+2)} < \frac{N^2}{n^2}.$$

Der modifizierte Schätzer \hat{N} ist erwartungstreu und seine Ungenauigkeit fällt schneller mit n ab.

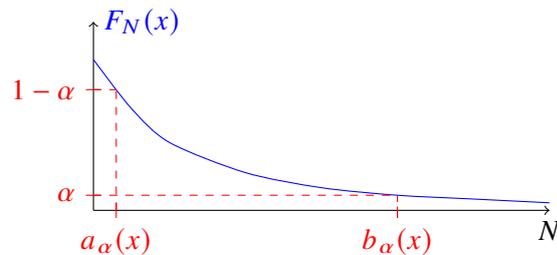
Konfidenzschranken für N

Anstatt eines konkreten Schätzers kann man auch Schranken für N angeben, die mit vorgegebener Sicherheit korrekt sind. Wir wollen für mögliche Werte von N bestimmen, ob der Wert x für die Verteilungsfunktion F_N „verdächtig klein“ bzw. „verdächtig groß“ ist. Beispielsweise liefert uns der Maximumschätzer eine sichere untere Schranke, denn $N \geq M$ gilt immer. Sei also $\alpha \in (0, 1)$ gegeben, so suchen wir eine obere Schranke für N mit Sicherheit bzw. Konfidenzniveau $1 - \alpha$.

Für die Verteilungsfunktion gilt nach (*)

$$F_N(x) = \mathbb{P}_N[M \leq x] = \frac{[x]_n}{[N]_n} \quad \text{für } n \leq x \leq N.$$

1. Statistische Verfahren: Grundbegriffe und Beispiele



Die Verteilungsfunktion F_N ist monoton wachsend in x , aber monoton fallend in N . Wenn N größer wird, verschiebt sich die Verteilung von M zu größeren Werten, wodurch die Verteilungsfunktion später ansteigt.

Wie die Abbildung zeigt, betrachten wir die beiden Werte

$$a_\alpha := \min\{N \geq n : F_N(x_-) < 1 - \alpha\} \quad \text{und} \quad b_\alpha := \max\{N \geq n : F_N(x) > \alpha\}.$$

Nach Lemma 1.5 gibt uns $b_\alpha(x)$ eine obere Schranke mit Sicherheit $1 - \alpha$. Dies wird im folgenden Lemma konkretisiert.

Lemma 1.8. Die datenabhängige Zahl $b_\alpha(x)$ ist eine obere Konfidenzschranke für N zum Konfidenzniveau $1 - \alpha$ bzw. Signifikanzniveau α . Das heißt:

$$\mathbb{P}[N \leq b_\alpha(M)] \geq 1 - \alpha \quad \text{für alle } N \in \mathbb{N}.$$

Beweis. Sei $\alpha \in (0, 1)$. Nach Lemma 1.5 gilt:

$$\mathbb{P}_N[F_N(M) \leq \alpha] \leq \alpha.$$

Da die Abbildung $N \rightarrow F_N(M)$ monoton fallend ist, ist dies äquivalent zu:

$$\mathbb{P}_N[N \leq b_\alpha(M)] = \mathbb{P}_N[F_N(M) > \alpha] \geq 1 - \alpha.$$

Wie erwähnt liefert uns M zudem eine untere Konfidenzschranke für N zum Konfidenzniveau $1 - \alpha$. Eine untere Konfidenzschranke zu einem anderen Konfidenzniveau $1 - \alpha$ können wir analog zu Lemma 1.7 mithilfe des rechtsseitigen p -Wertes p_r erhalten:

$$\mathbb{P}_N[N \geq a_\alpha(M)] = \mathbb{P}_N[F_N(M_-) < 1 - \alpha] = \mathbb{P}_N[1 - F_N(M_-) > \alpha] \geq 1 - \alpha.$$

Insgesamt erhalten wir:

Satz 1.9. Die Intervalle $[M, b_\alpha(M)]$ sowie $[a_{\frac{\alpha}{2}}, b_{\frac{\alpha}{2}}(M)]$ sind jeweils Konfidenzintervalle für N zum Konfidenzniveau $1 - \alpha$, das heißt:

$$\mathbb{P}_N[N \notin [M, b_\alpha(M)]] \leq \alpha \quad \text{für alle } N \in \mathbb{N},$$

$$\mathbb{P}_N[N \notin [a_{\frac{\alpha}{2}}(M), b_{\frac{\alpha}{2}}(M)]] \leq \alpha \quad \text{für alle } N \in \mathbb{N}.$$

Bemerkung. (i) Die Konfidenzintervalle hängen von M ab. Bei jeder Durchführung des Zufallsexperiments ergibt sich also ein anderes Konfidenzintervall.

(ii) Führen wir mehrere unabhängige Untersuchungen durch und erhalten jeweils ein $(1 - \alpha)$ -Konfidenzintervall, dann liegen die wahren Parameterwerte in durchschnittlich mindestens $100(1 - \alpha)\%$ der Fälle im jeweiligen Konfidenzintervall.

(iii) Bei einer Durchführung erhält man eine konkrete Realisierung des Konfidenzintervalls. Die Aussage „Der wahre Parameter liegt mit Wahrscheinlichkeit $1 - \alpha$ in dem (festen, realisierten) Intervall“ ist falsch und sinnlos, denn N ist nicht zufällig.

Capture-Recapture-Verfahren

Ein weiteres Verfahren zur Bestimmung einer unbekanntem Populationsgröße N ist das sogenannte Capture-Recapture-Verfahren. Im einfachsten Fall handelt es sich um ein zweistufiges Experiment:

1. Capture Entnehme eine Zufallsstichprobe der Größe $l \leq N$. Markiere diese und entlasse sie wieder.

2. Recapture Nehme eine unabhängige Zufallsstichprobe der Größe $n \leq N$.

Sei nun H die Anzahl der Markierten in der 2. Stichprobe. Unter der Wahrscheinlichkeitsverteilung \mathbb{P}_N ist H dann hypergeometrisch verteilt mit den Parametern N, l, n . Ein möglicher Schätzer für N wäre:

$$\hat{N} := \frac{nl}{H}$$

Denn wenn wir davon ausgehen, dass der Anteil der Markierten in der 2. Stichprobe annähernd dem Anteil in der gesamten Population entspricht, ergibt sich:

$$\frac{l}{N} \approx \frac{H}{n} \Rightarrow N \approx \frac{nl}{H} = \hat{N}.$$

Konfidenzintervall

Wie in der Übungsaufgabe bewiesen, ist die Abbildung $N \mapsto F_{N,l,n}(x)$ monoton wachsend. Nach Lemma 1.5 ergibt sich die untere Konfidenzschranke:

$$1 - \alpha \leq \mathbb{P}_N[F_{N,l,n}(H) > \alpha] = \mathbb{P}_N[N \geq a_\alpha(H)] \quad \text{wobei} \quad a_\alpha(x) := \min\{N : F_{N,l,n}(H) > \alpha\}.$$

Eine obere Konfidenzschranke erhält man analog.

Beispiel. Wir betrachten das Capture-Recapture-Verfahren mit $l = n = 20$ und dem Beobachtungswert $H = 2$. Somit erhalten wir den Schätzer:

$$\hat{N} = 20 \cdot \frac{20}{2} = 200.$$

Gegeben sei das Signifikanzniveau $\alpha = 5\%$:

$$F_{N,20,20} : \quad N = 77 \sim 0,0495 \quad ; \quad N = 78 \sim 0,0537.$$

Somit ergibt sich eine untere 95%-Konfidenzschranke von 78 für N .

1.3. Statistische Modelle und Verfahren

In diesem Abschnitt befassen wir uns mit der schließenden Statistik. Hierbei werden aus empirischen Daten Rückschlüsse auf zugrundeliegende Phänomene gezogen, selbst wenn die Daten fehlerbehaftet oder unvollständig sind. Wir betrachten den folgenden allgemeinen Rahmen für statistische Modelle.

Definition 1.10 (Statistisches Modell). 1) Ein *statistisches Modell* besteht aus:

- Einer Menge $\Omega \neq \emptyset$ zusammen mit einer σ -Algebra \mathcal{A} .
- Einer Parametermenge $\Theta \neq \emptyset$.
- Einer Familie $(P_\theta)_{\theta \in \Theta}$ von Wahrscheinlichkeitsverteilungen auf (Ω, \mathcal{A}) , wobei $\theta \in \Theta$ der unbekannte Parameter ist.
- Einer messbaren Abbildung $X : \Omega \rightarrow S$ (Stichprobe), wobei (S, \mathcal{B}) ein messbarer Raum ist.

2) Eine *Statistik* ist eine Abbildung $T(X)$, wobei $T : S \rightarrow \mathbb{R}$ eine messbare Abbildung ist.

1. Statistische Verfahren: Grundbegriffe und Beispiele

Der realisierte Wert $T(x)$ einer Statistik ist eine Kenngröße, die wir aus den Beobachtungsdaten $x = X(\omega)$ berechnen können.

Beispiel (Capture-Recapture). Betrachten wir das beschriebene Capture-Recapture-Verfahren zur Bestimmung einer Populationsgröße. Hierbei gilt $\Theta = \mathbb{N}_{\geq \max(l,n)}$ mit

$$\Omega = \left\{ (\omega_1^{(1)}, \dots, \omega_1^{(l)}, \omega_2^{(1)}, \dots, \omega_2^{(n)}) : \omega_i^{(k)} \neq \omega_i^{(l)} \text{ für } k \neq l, i \in \{1, 2\} \right\} \subset \mathbb{N}_0^{l+n}$$

und $X(\omega) = \omega$. Mit der Familie $\mathbb{P}_N = \text{Unif}(\Omega_N)$, wobei $\Omega_N = \{\omega \in \Omega : \omega_i^{(k)} \leq N \text{ für alle } i, k\}$. Die Anzahl der Markierten in der 2. Stichprobe ist beispielsweise eine Statistik, die durch

$$H(\omega) = \left| \{\omega_1^{(1)}, \dots, \omega_1^{(l)}\} \cap \{\omega_2^{(1)}, \dots, \omega_2^{(n)}\} \right|$$

gegeben ist.

Einige grundlegende Modelle

1) BERNOULLI-MODELL, SCHÄTZEN VON WAHRSCHEINLICHKEITEN

Wir betrachten ein Bernoulli-Experiment mit unbekannter Erfolgswahrscheinlichkeit θ und dem eindimensionalen Parameterraum $\Theta = [0, 1]$. Die n -fache Durchführung des Experiments liefert die Stichprobe $X = (X_1, \dots, X_n)$, wobei $X_i \sim \text{Bernoulli}(\theta)$ unabhängig unter \mathbb{P}_θ verteilt sind. Eine interessante Statistik hierbei ist die relative Häufigkeit

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Nach dem Gesetz der großen Zahlen gilt $\bar{X}_n \approx \theta$ für große n . Eine Fehlerabschätzung kann beispielsweise über die Tschebyscheff-Ungleichung oder exponentielle Ungleichungen erfolgen.

2) GAUß-MODELL, PARAMETERSCHÄTZUNG

Das Schätzen einer Wahrscheinlichkeitsverteilung kann komplex sein. Unter der Annahme, dass die Verteilung annähernd normal ist, lässt sich dieses Problem jedoch erheblich vereinfachen. Bei einer Normalverteilung, die durch Erwartungswert und Varianz eindeutig bestimmt ist, betrachten wir den zweidimensionalen Parameterbereich $\Theta = \mathbb{R} \times (0, \infty) \sim (m, v)$. Die Stichproben sind $X = (X_1, \dots, X_n)$, wobei $X_i \sim N(m, v)$ unabhängig unter $\mathbb{P}_{m,v}$ verteilt sind. Interessante Statistiken sind hierbei \bar{X}_n und

$$V_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad T_n = \frac{\bar{X}_n - m_0}{\sqrt{V_n/n}}.$$

Hierbei ist T_n die Student-t-Statistik mit der Nullhypothese $m = m_0$. Insbesondere werden wir später zeigen, dass der Schätzer $V_n^* = \frac{n}{n-1} V_n$ genauer ist als V_n .

3) NICHPARAMETRISCHE SCHÄTZUNG DER VERTEILUNG BZW. VERTEILUNGSFUNKTION

Hierbei ist Θ die Menge aller Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Stichprobe $X = (X_1, \dots, X_n)$, wobei $X_i \sim \mu$ unabhängig unter \mathbb{P}_μ verteilt sind. Statistiken sind:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \quad \hat{F}_n(c) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq c\}}.$$

Die *empirische Verteilung* $\hat{\mu}_n[B]$ gibt die relative Häufigkeit der Werte in B unter X_1, \dots, X_n an. Analog beschreibt die *empirische Verteilungsfunktion* die relative Häufigkeit der Werte $\leq c$ in der Stichprobe.

4) NICHTPARAMETRISCHE DICHTESCHÄTZUNG

Sei

$$\Theta = \left\{ f : \mathbb{R} \rightarrow [0, \infty) : \int f(x) dx = 1, \exists \text{ schwache Ableitung } f'' \text{ mit } \int (f''(x))^2 dx < \infty \right\}.$$

Hierbei bedeutet schwache Ableitung, dass f geschrieben werden kann als $f(x) = \int_0^x \int_0^t g(s) ds dt$, wobei $\int g^2 < \infty$. Die Stichproben X_1, \dots, X_n sind unabhängig mit Dichte f unter \mathbb{P}_f . Mögliche Statistiken sind:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \varphi_h(x - X_i), \quad \varphi_h(x) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{x^2}{2h}}, \quad \text{für } h > 0.$$

5) REGRESSION

Betrachten wir nun die Regression. Gegeben seien $(X_1, Y_1), \dots, (X_n, Y_n) \subset \Omega \rightarrow \mathbb{R}^d \times \mathbb{R}$, wobei X_i Kontrollgrößen und Y_i abhängige Variablen darstellen. Die Abhängigkeit sei gegeben durch

$$Y_i = f(X_i) + \sqrt{v}\varepsilon_i$$

Hierbei sind ε_i identisch und unabhängig verteilt und unabhängig von X . Im Folgenden betrachten wir einige grundlegende Modelle der Regression.

- **Nichtparametrisches Modell:**

$$\Theta = \{(v, f) \mid v \in (0, \infty), f : \mathbb{R}^d \rightarrow \mathbb{R}\}$$

Ein nichtparametrisches Modell ist *unendlichdimensional* und bietet große Flexibilität, da keine spezifische Form für f angenommen wird.

- **Lineares Modell:**

$$f(x) = \omega^T x$$

Hierbei ist der Parameterraum:

$$\Theta = \{(v, \omega) \mid v \in (0, \infty), \omega \in \mathbb{R}^d\}$$

Ein lineares Modell ist relativ einfach und hat die Dimension $d + 1$, wobei d die Anzahl der Kontrollgrößen ist.

- **Neuronales Netzwerk:**

$$f(x)$$

In diesem Fall ist $f(x)$ eine nichtlineare Funktion, die sich über ein vorgegebenes neuronales Netzwerk darstellen lässt. Diese Modelle sind *hochdimensional* und können sehr komplexe Zusammenhänge abbilden.

1.4. Statistische Verfahren

In diesem Abschnitt befassen wir uns mit verschiedenen statistischen Verfahren, die zur Schätzung unbekannter Parameter und zur Beurteilung der Genauigkeit dieser Schätzungen verwendet werden. Wir werden wichtige Konzepte wie Schätzer, systematische Fehler (Bias) und mittlere quadratische Fehler (MSE) einführen und an Beispielen veranschaulichen.

Schätzer

Gesucht sei $g(\theta)$, wobei $g : \Theta \rightarrow \mathbb{R}$ eine Funktion ist. Zum Beispiel hatten wir im Gaußmodell $\theta = (m, v)$, und gesucht war m .

Ein *Schätzer für $g(\theta)$* ist durch eine Statistik $\hat{g} = T(X)$ gegeben. Um verschiedene Schätzer zu vergleichen, betrachten wir unter anderem die folgenden Größen:

Definition 1.11. (i) Der *systematische Fehler* (Bias) von \hat{g} ist

$$\text{Bias}_\theta(\hat{g}) := \mathbb{E}_\theta[\hat{g}] - g(\theta).$$

Der Schätzer \hat{g} heißt *erwartungstreu* (unbiased), falls

$$\text{Bias}_\theta(\hat{g}) = 0 \quad \text{für alle } \theta \in \Theta.$$

(ii) Der *mittlere quadratische Fehler* (mean squared error, MSE) von \hat{g} ist

$$\text{MSE}_\theta(\hat{g}) := \mathbb{E}_\theta [(\hat{g} - g(\theta))^2].$$

Das folgende Lemma liefert eine praktische Konkretisierung des mittleren quadratischen Fehlers.

Lemma 1.12. Für einen Schätzer \hat{g} von $g(\theta)$ gilt:

$$\text{MSE}_\theta(\hat{g}) = \text{Var}_\theta[\hat{g}] + \text{Bias}_\theta(\hat{g})^2.$$

Beweis.

$$\text{MSE}_\theta(\hat{g}) = \mathbb{E}_\theta [(\hat{g} - \mathbb{E}_\theta[\hat{g}] + \overbrace{\mathbb{E}_\theta[\hat{g}] - g(\theta)}^{\text{Bias}_\theta(\hat{g})})^2].$$

Die Behauptung folgt, weil

$$\mathbb{E}_\theta [(\hat{g} - \mathbb{E}_\theta[\hat{g}]) \cdot \text{Bias}_\theta(\hat{g})] = 0.$$

Beispiel (NICHPARAMETRISCHES SCHÄTZEN VON ERWARTUNGSWERT, VARIANZ UND VERTEILUNG).

Wir betrachten Θ als die Menge aller Wahrscheinlichkeitsverteilungen μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit $\int_{\mathbb{R}} x^2 \mu(dx) < \infty$. Mit der Stichprobe $X = (X_1, \dots, X_n)$, wobei $X_i \sim \mu$ unabhängig unter \mathbb{P}_μ sind, sind Erwartungswert und Varianz durch die folgenden Abbildungen gegeben:

$$m(\mu) = \int_{\mathbb{R}} x \mu(dx), \quad v(\mu) = \int_{\mathbb{R}} (x - m(\mu))^2 \mu(dx).$$

- 1) SCHÄTZEN VON m : Die relative Häufigkeit $\bar{X}_n = \frac{1}{n} \sum X_i$ ist ein erwartungstreuer Schätzer für den Erwartungswert, denn

$$\mathbb{E}_\mu[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = m(\mu) \quad \text{für alle } \mu \in \Theta.$$

Da \bar{X}_n erwartungstreu ist und die Stichproben X_i unabhängig sind, folgt:

$$\text{MSE}_\mu(\bar{X}_n) = \text{Var}_\mu[\bar{X}_n] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_\mu[X_i] = \frac{v(\mu)}{n} \in O\left(\frac{1}{n}\right).$$

- 2) SCHÄTZEN VON v : Die Stichprobenvarianz $V_n = \frac{1}{n} \sum (X_i - \bar{X}_n)^2$ ist nicht erwartungstreu. Zuerst wollen wir die folgende Behauptung beweisen:

$$\mathbb{E}_\mu \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] = (n-1)v(\mu).$$

Sei ohne Beschränkung der Allgemeinheit $\mathbb{E}_\mu[X_i] = 0$, ansonsten betrachte $\tilde{X}_i = X_i - \mathbb{E}_\mu[X_i]$. Nun gilt:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i \bar{X}_n + \sum_{i=1}^n \bar{X}_n^2 \\ &= \sum_{i=1}^n X_i^2 - n \bar{X}_n^2 \\ \Rightarrow \mathbb{E}_\mu \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] &= \underbrace{n \mathbb{E}_\mu[X_1^2]}_{=v(\mu)} - n \underbrace{\mathbb{E}_\mu[\bar{X}_n^2]}_{=\text{Var}_\mu[\bar{X}_n] = \frac{v(\mu)}{n}} = (n-1)v(\mu). \end{aligned}$$

Es folgt also, dass $V_n^* := \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$ ein erwartungstreuer Schätzer von v ist.

- 3) SCHÄTZEN VON μ : Als Schätzer der Verteilung μ betrachten wir die empirische Verteilung

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \quad \hat{\mu}_n(B) = \frac{1}{n} \sum_{i=1}^n 1_B(X_i),$$

die die relative Häufigkeit von Werten in $B \in \mathcal{B}(\mathbb{R})$ angibt. Es ergibt sich insbesondere:

$$\mathbb{E}_\mu[\hat{\mu}_n(B)] = \frac{1}{n} \sum_{i=1}^n \mathbb{P}_\mu[X_i \in B] = \mu(B) \quad \text{für alle } B \in \mathcal{B}(\mathbb{R}).$$

Somit ist die empirische Verteilung $\hat{\mu}_n$ ein erwartungstreuer Schätzer für μ .

Konfidenzbereiche

Definition 1.13. Sei $\alpha \in (0, 1)$. Ein *Konfidenzbereich* für $g(\theta)$ mit *Konfidenzniveau* $1 - \alpha$ ist eine Abbildung $C(X)$, so dass

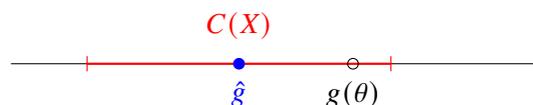
$$C : S \rightarrow \mathcal{P}(G),$$

wobei $\mathcal{P}(G)$ die Menge aller Teilmengen von G ist, und

$$\mathbb{P}_\theta[g(\theta) \in C(X)] \geq 1 - \alpha \quad \text{für alle } \theta \in \Theta.$$

Bemerkung. Insbesondere soll $\{x \in S : g(\theta) \in C(x)\}$ für jedes $\theta \in \Theta$ eine messbare Teilmenge von S sein.

Für $G = \mathbb{R}$ folgt meistens, dass die Konfidenzbereiche durch $C(X) = [a(X), b(X)]$ gegeben sind, wobei a und b Statistiken sind.



Konstruktionsverfahren für Konfidenzintervalle

- 1) ÜBER DIE VERTEILUNGSFUNKTION EINER STATISTIK: Sei $T(X) : \Omega \rightarrow \mathbb{R}$ eine Statistik mit der Verteilungsfunktion $F_\theta(c) = \mathbb{P}_\theta[T(X) \leq c]$ für $c \in \mathbb{R}$. Nach Lemma 1.3 gilt dann: $\mathbb{P}_\theta[F_\theta(T(X)) > \alpha] \geq 1 - \alpha$. Dies liefert den Konfidenzbereich

$$C(x) := \{g(\theta) : \theta \in \Theta \text{ mit } F_\theta(T(X)) > \alpha\},$$

welcher jene Parameterwerte ausschließt, für die $T(X)$ „verdächtig klein“ ist. Folglich gilt mit Lemma 1.3: $\mathbb{P}_\theta[g(\theta) \in C(X)] \geq 1 - \alpha$ für alle $\theta \in \Theta$. Analog erhalten wir

$$C'(x) := \{g(\theta) : \theta \in \Theta \text{ mit } F_\theta(T(X)_-) < 1 - \alpha\},$$

welcher Werte ausschließt, für die $T(X)$ „verdächtig groß“ ist. Schließlich definiert man

$$C''(x) := \left\{g(\theta) : \theta \in \Theta \text{ mit } F_\theta(T(X)_-) < 1 - \frac{\alpha}{2} \text{ und } F_\theta(T(X)_-) > \frac{\alpha}{2}\right\},$$

welcher Werte ausschließt, für die $T(X)$ „verdächtig klein oder groß“ ist. Diese Konfidenzintervalle sind wie folgt konkretisiert:

Korollar 1.14. Die beschriebenen Konfidenzbereiche $C(X), C'(X)$ und $C''(X)$ sind $(1 - \alpha)$ -Konfidenzintervalle für $g(\theta)$.

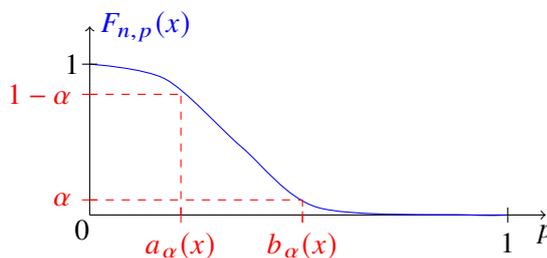
Beispiel (Bernoulli-Modell). Sei $\Theta = [0, 1]$ und $p \in \Theta$ mit den Stichproben $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ unabhängig unter \mathbb{P}_p . Dann ist die Stichprobensumme $H_n = X_1 + \dots + X_n$ unter \mathbb{P}_p binomialverteilt mit Parametern (n, p) und der Verteilungsfunktion

$$F_{n,p}(c) = \mathbb{P}_p[H_n \leq c] = \sum_{k=0}^c \binom{n}{k} p^k (1-p)^{n-k} \quad \text{für } c = 0, 1, \dots, n.$$

Insbesondere ist die Abbildung $p \mapsto F_{n,p}(c)$ für $c < n$ stetig differenzierbar mit

$$\frac{d}{dp} F_{n,p}(c) = -n \binom{n-1}{c} p^c (1-p)^{n-1-c} < 0,$$

somit ist sie streng monoton fallend.



Nach Korollar folgt, dass $C(X) = \{p : F(n, p)(H_n) > \alpha\} = [0, b_\alpha(H_n))$ ein Konfidenzintervall zum Niveau $1 - \alpha$ und $(a_\alpha(H_n), b_\alpha(H_n))$ ein Konfidenzintervall zum Niveau $1 - 2\alpha$ ist.

- 2) ÜBER DIE LIKELIHOOD-FUNKTION: Gegeben den Beobachtungswert $x = X(\omega)$ wollen wir Parameterwerte ausschließen, unter denen x eher unwahrscheinlich ist. Hierfür setzen wir folgende Annahmen voraus:

- a) S ist abzählbar mit der Massenfunktion $f_\theta(x) = \mathbb{P}_\theta[X = x]$,
- b) $S = \mathbb{R}^d$ mit der Dichtefunktion $\mathbb{P}_\theta[X \in B] = \int_B f_\theta(x) dx$.

Die Funktion $\theta \mapsto f_\theta(x)$ wird als *Likelihood* oder Plausibilität bezeichnet. Für einen festen Schwellenwert c_θ definieren wir $C(x) = \{g(\theta) : f_\theta(x) \geq c_\theta\}$. Damit ergibt sich:

$$\mathbb{P}_\theta[g(\theta) \in C(X)] \geq \mathbb{P}_\theta[f_\theta(X) \geq c_\theta] \geq 1 - \alpha \quad \text{für alle } \theta \in \Theta. \quad (**)$$

Abhängig von θ sollte c_θ so groß wie möglich gewählt werden, damit (*) gerade noch erfüllt ist.

Beispiel. Ein Energieunternehmen besitzt $N = 10$ Heizkraftwerke, davon werden $n = 4$ auf ihre Schadstoffwerte überprüft. Bei der Überprüfung wird bei x Heizkraftwerken ein zu hoher Schadstoffwert festgestellt. Wir möchten anhand dieser Stichprobe den Parameter $\theta =$ „Anzahl der Kraftwerke mit zu hohen Werten“ bestimmen. Die Zufallsvariable X ist hypergeometrisch verteilt mit Parametern (N, θ, n) , wobei $S = \{0, 1, \dots, n\}$. Die Massenfunktion ist gegeben durch

$$f_\theta(x) = \frac{\binom{\theta}{x} \binom{N-\theta}{n-x}}{\binom{N}{n}} \quad \text{für } x \leq \theta, \quad 0 \text{ sonst.}$$

Die folgende Tabelle veranschaulicht die Verteilung sowohl abhängig von θ als auch von x .

| | | $\binom{\theta}{x} \binom{N-\theta}{n-x}$ | | | | | |
|----------|----|---|-----|-----|-----|-----|-------------|
| | | 0 | 1 | 2 | 3 | 4 | |
| θ | 0 | 210 | 0 | 0 | 0 | 0 | $\geq 80\%$ |
| | 1 | 126 | 84 | 0 | 0 | 0 | $\geq 80\%$ |
| | 2 | 70 | 112 | 28 | 0 | 0 | $\geq 80\%$ |
| | 3 | 35 | 105 | 63 | 7 | 0 | $\geq 80\%$ |
| | 4 | 15 | 80 | 90 | 24 | 1 | $\geq 80\%$ |
| | 5 | 5 | 50 | 100 | 50 | 5 | $\geq 80\%$ |
| | 6 | 1 | 24 | 90 | 80 | 15 | $\geq 80\%$ |
| | 7 | 0 | 7 | 63 | 105 | 35 | $\geq 80\%$ |
| | 8 | 0 | 0 | 28 | 112 | 70 | $\geq 80\%$ |
| | 9 | 0 | 0 | 0 | 84 | 126 | $\geq 80\%$ |
| | 10 | 0 | 0 | 0 | 0 | 210 | $\geq 80\%$ |

Für den Wert $X = 0$ ergibt sich der Konfidenzbereich $C(X) = \{0, 1, 2\}$ mit Konfidenzniveau 0,8.

3) ÜBER EINE PIVOT-STATISTIK:

Beispiel (Gauß-Modell mit fester Varianz). Gegeben sei die Varianz $v > 0$ einer Normalverteilung und der Erwartungswert $m \in \mathbb{R}$ sei unbekannt. Wir betrachten ein statistisches Modell mit dem Parameterbereich $\Theta = \mathbb{R}$ und unabhängigen Stichproben $X_1, \dots, X_n \sim N(m, v)$ unter $\mathbb{P}_{m,v}$. Die Summe $\sum X_i$ ist wieder normalverteilt mit $N(nm, nv)$ und somit ist $\bar{X}_n = \frac{1}{n} \sum X_i \sim N(m, \frac{v}{n})$. Um ein Konfidenzintervall für m zu bestimmen, standardisieren wir:

$$Z := \frac{\bar{X}_n - m}{\sqrt{v/n}} \sim N(0, 1).$$

Man sagt, dass Z ein *Pivot* ist, was bedeutet, dass die Verteilung nicht vom Parameter m abhängt. Damit können wir ein Konfidenzintervall konstruieren:

$$\mathbb{P}_{m,v} \left[|\bar{X}_n - m| \geq \sqrt{\frac{v}{n}} c \right] = \mathbb{P}_{m,v} [|Z| \geq c] \stackrel{!}{\leq} \alpha \quad \text{mit} \quad \Phi(c) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Diese Ungleichung ist erfüllt, falls $c = \Phi^{-1}(1 - \frac{\alpha}{2})$. Das Konfidenzintervall wird wie folgt konstruiert:

$$\left(\bar{X}_n - \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\frac{v}{n}}, \bar{X}_n + \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\frac{v}{n}} \right)$$

zum Niveau $1 - \alpha$. Allgemeiner definieren wir:

Definition 1.15 (Pivot). Ein *Pivot* für $g(\theta)$ ist eine Statistik $T(X, g(\theta))$, wobei $T : S \times \mathbb{R} \rightarrow \mathbb{R}$ messbar ist und deren Verteilung μ unter \mathbb{P}_θ nicht von θ abhängt.

Ein Pivot hat die besondere Eigenschaft, dass seine Verteilung unter der Nullhypothese nicht vom Parameter θ abhängt, was es zu einem robusten Werkzeug in der statistischen Inferenz macht. Diese Eigenschaft wird im folgenden Satz demonstriert:

Satz 1.16. Ist $T(X, g(\theta))$ ein Pivot und $B \in \mathcal{B}(\mathbb{R})$, dann ist

$$C(X) = \{\gamma \in \mathbb{R} : T(X, \gamma) \in B\}$$

ein Konfidenzbereich für $g(\theta)$ zum Niveau $\mu(B)$.

Beweis. Falls $T(X, g(\theta)) \in B$ ist, dann ist $g(\theta) \in C(X)$. Daher gilt

$$\mathbb{P}_\theta [g(\theta) \in C(X)] \geq \mathbb{P}_\theta [T(X, g(\theta)) \in B] = \mu(B).$$

Beispiel (Gauß-Modell mit m, v unbekannt). Betrachten wir erneut das statistische Modell $\Theta = \{(m, v) : m \in \mathbb{R}, v > 0\}$ mit den unabhängigen Stichproben X_1, \dots, X_n , welche unter $\mathbb{P}_{m,v}$ als $N(m, v)$ verteilt angenommen werden. Gesucht ist ein Konfidenzintervall für den Erwartungswert m . Seien \bar{X}_n und V_n^* wie oben, dann ist

$$T_n = \frac{\bar{X}_n - m}{\sqrt{V_n^*/n}}$$

ein Pivot für m .

Beweis. Die standardisierten Zufallsvariablen $Y_i := \frac{X_i - m}{\sqrt{v}}$ sind unter $\mathbb{P}_{m,v}$ unabhängig und folgen der Verteilung $N(0, 1)$. Folglich hängt die Verteilung von $Y = (Y_1, \dots, Y_n)$ nicht von m und v ab. Wir können T_n wie folgt als Funktion von Y darstellen:

$$T_n = \sqrt{n} \frac{\frac{\bar{X}_n - m}{\sqrt{v}}}{\sqrt{\frac{V_n^*}{v}}} = \sqrt{n} \frac{\bar{Y}_n}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

Da

$$\frac{V_n^*}{v} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sqrt{v}} \right)^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}_n}{\sqrt{1}} \right)^2,$$

hängt auch die Verteilung von T_n nicht von m und v ab, und die Behauptung folgt. ■

Damit erhalten wir analog zum vorherigen Beispiel:

$$\{|T_n| < q_{1-\frac{\alpha}{2}}\} = \left(\bar{X}_n - q_{1-\frac{\alpha}{2}} \sqrt{\frac{V_n^*}{n}}, \bar{X}_n + q_{1-\frac{\alpha}{2}} \sqrt{\frac{V_n^*}{n}} \right)$$

mit $q_n := F_{T_n}^{-1}(c)$ ein Konfidenzintervall für m zum Niveau $1 - \alpha$.

Die Verteilung von T_n wird als *Student'sche t-Verteilung mit $n - 1$ Freiheitsgraden* bezeichnet. Die Dichte und die Verteilungsfunktion dieser Verteilung werden später berechnet.

4) APPROXIMATIVE KONFIDENZINTERVALLE (über eine Normalapproximation)

Beispiel (Bernoulli- bzw. Binomialmodell)

Wir betrachten das Bernoulli- bzw. Binomialmodell mit der unbekanntem Wahrscheinlichkeit $p \in [0, 1] = \Theta$. Sei $H_n = X_1 + \dots + X_n$, wobei $X_i \sim \text{Bernoulli}(p)$ unabhängig unter \mathbb{P}_p sind. H_n ist binomialverteilt, daher gilt $\mathbb{E}_p[H_n] = np$ und $\text{Var}_p[H_n] = np(1-p)$. Sei nun $\hat{p} = \frac{H_n}{n}$. Nach dem zentralen Grenzwertsatz bzw. dem Satz von de Moivre/Laplace gilt:

$$\mathbb{P}_p \left[\frac{H_n - np}{\sqrt{np(1-p)}} \in (-c, c) \right] = \mathbb{P}_p \left[|p - \hat{p}_n| < c \sqrt{\frac{p(1-p)}{n}} \right] \xrightarrow{n \rightarrow \infty} N(0, 1)(-c, c) = 2 \left(\Phi(c) - \frac{1}{2} \right)$$

Für große n gilt also approximativ:

$$\mathbb{P}_p \left[|p - \hat{p}_n| < c \sqrt{\frac{p(1-p)}{n}} \right] \approx 2 \left(\Phi(c) - \frac{1}{2} \right) \geq 1 - \alpha \text{ für } c \geq \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$$

Die Breite dieses Konfidenzintervalls hängt jedoch vom Parameter p ab, was die Berechnung und Interpretation der Intervalle kompliziert macht. Im Folgenden betrachten wir Möglichkeiten, diese Abhängigkeit zu umgehen.

1. Für alle $p \in [0, 1]$ gilt die Ungleichung $p(1-p) \leq \frac{1}{4}$. Dies ergibt das Konfidenzintervall:

$$\left(\hat{p}_n \pm \frac{c}{\sqrt{4n}} \right)$$

Dieses Intervall ist zwar einfach zu berechnen, aber konservativ.

2. Nach dem Gesetz der großen Zahlen gilt $\hat{p}_n \approx p$ für große n . Daraus ergibt sich das approximative Konfidenzintervall:

$$\left(\hat{p}_n \pm c \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right) \text{ praktisch, aber ungenau.}$$

3. Der Ansatz von Abraham Wald zeigt, dass das Konfidenzintervall auch als Lösung einer Ungleichung der Form

$$\begin{aligned} |p - \hat{p}_n| \leq \frac{c}{\sqrt{n}} \sqrt{p(1-p)} &\Leftrightarrow p^2 - 2p\hat{p}_n + \hat{p}_n^2 \leq \frac{c^2}{n} (p - p^2) \\ &\Leftrightarrow \left(1 + \frac{c^2}{n} \right) p^2 - 2p \left(\hat{p}_n + \frac{c^2}{2n} \right) + \hat{p}_n^2 \leq 0 \\ &\Leftrightarrow p \in \left(\frac{\hat{p}_n + \frac{c^2}{2n} \pm \frac{c}{\sqrt{n}} \sqrt{\hat{p}_n(1-\hat{p}_n) + \frac{c^2}{4n}}}{1 + \frac{c^2}{n}} \right) \end{aligned}$$

beschrieben werden kann. Durch quadratische Ergänzung erhalten wir im letzten Schritt ein approximatives Konfidenzintervall. Heutzutage können exakte Konfidenzintervalle oft numerisch sehr genau berechnet werden, wodurch eine Normalapproximation nicht mehr erforderlich ist.

Hypothesentests

Für $\theta \in \Theta$ betrachten wir den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ mit $X : \Omega \rightarrow S$. Seien $\Theta_0, \Theta_1 \subseteq \Theta$ disjunkte Teilmengen. Durch sie sind die Nullhypothese H_0 und die alternative Hypothese H_1 bestimmt:

$$H_0 : \theta \in \Theta_0 \quad \text{und} \quad H_1 : \theta \in \Theta_1$$

Definition 1.17 (Hypothesentest). Ein *Hypothesentest* für das obige Testproblem ist gegeben durch eine messbare Funktion

$$\varphi : S \rightarrow \{0, 1\} \quad (\text{nicht-randomisierter Test}), \quad \text{bzw.} \quad \varphi : S \rightarrow [0, 1] \quad (\text{randomisierter Test})$$

mit der Entscheidungsregel:

- Verwerfe H_0 falls $\varphi(x) = 1$,
- Verwerfe H_0 nicht falls $\varphi(x) = 0$,
- Verwerfe H_0 mit Wahrscheinlichkeit $\varphi(x)$ falls $\varphi(x) \in (0, 1)$.

Der *Verwerfungsbereich* des Tests ist die Menge

$$R = \{x \in S : \varphi(x) = 1\}.$$

Die Funktion

$$\beta(\theta) = \mathbb{E}_\theta[\varphi(X)] \quad (\theta \in \Theta)$$

heißt *Gütefunktion*. Sie beschreibt die Verwerfungswahrscheinlichkeit in Abhängigkeit von θ .

Der Test hat *Signifikanzniveau* α , falls die Niveaubedingung

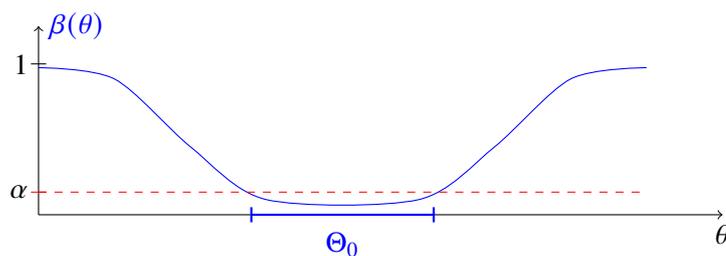
$$\beta(\theta) \leq \alpha \quad \text{für alle } \theta \in \Theta_0$$

erfüllt ist. Die auf Θ_1 eingeschränkte Gütefunktion

$$\beta(\theta), \quad \theta \in \Theta_1,$$

heißt *Macht* des Hypothesentests.

Ziel sollte es sein, dass $\beta(\theta)$ für $\theta \in \Theta_1$ möglichst groß ist, um Fehler 2. Art zu vermeiden, aber die Nebenbedingung $\beta(\theta) \leq \alpha$ für alle $\theta \in \Theta_0$ noch immer erfüllt ist.



Ein Fehler 1. Art liegt vor, wenn die Nullhypothese H_0 verworfen wird, obwohl sie in Wirklichkeit wahr ist. Bei gegebenem Signifikanzniveau α tritt dies mit einer Wahrscheinlichkeit $\leq \alpha$ ein. Im Gegensatz dazu bedeutet ein Fehler 2. Art, dass der Test die Nullhypothese H_0 nicht verwirft, obwohl die Alternative zutrifft. Dies geschieht mit einer Wahrscheinlichkeit $1 - \beta(\theta)$, $\theta \in \Theta_1$, welche möglichst klein sein sollte.

Hypothesentests hängen eng mit Konfidenzbereichen zusammen, wie der folgende Satz zeigt:

Satz 1.18. Sei $C(X)$ ein $(1 - \alpha)$ -Konfidenzbereich für θ . Dann ist für jedes $\theta_0 \in \Theta$ ein Hypothesentest mit Signifikanzniveau α für die Nullhypothese $H_0 : \theta = \theta_0$ gegeben durch

$$\varphi = \begin{cases} 0 & \text{falls } \theta_0 \in C(X) \\ 1 & \text{falls } \theta_0 \notin C(X). \end{cases}$$

Beweis. Für $\theta_0 \in \Theta$ gilt nach Definition des Konfidenzbereichs

$$\beta(\theta_0) = \mathbb{E}_{\theta_0}[\varphi(X)] = \mathbb{P}_{\theta_0}[\theta_0 \notin C(X)] \leq \alpha.$$

Dies zeigt das Signifikanzniveau α . Die Umkehrung ist in Übung. ■

Der Satz zeigt, dass wir Tests mit analogen Methoden konstruieren können wie Konfidenzintervalle.

Beispiel (Gauß-Modell, t -Test). Seien X_1, \dots, X_n unabhängig und normalverteilt mit $N(m, \nu)$, und $m_0 \in \mathbb{R}$.

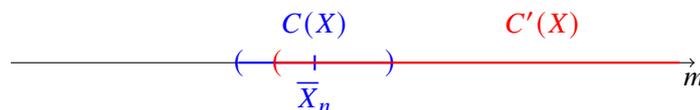
- a) Wir betrachten die Nullhypothese $H_0 : m = m_0$ mit der Alternative $H_1 : m \neq m_0$. Wir hatten gezeigt, dass wir mit der Student'schen t -Statistik T_n den folgenden $(1 - \alpha)$ -Konfidenzbereich erhalten:

$$C(X) = \left\{ m \in \mathbb{R} : \left| \frac{\bar{X}_n - m}{\sqrt{V_n^*/n}} \right| < q_{1-\frac{\alpha}{2}} \right\}.$$

Mit dem Satz erhalten wir einen α -Niveau-Test $\varphi(X) = 1$, falls $|T_n(m_0)| \geq q_{1-\frac{\alpha}{2}}$. Dieser Test wird als *Student's t -Test* bezeichnet.

- b) Sei $H_0 : m = m_0$ nun jedoch mit der linksseitigen Alternative $H_1 : m < m_0$. Analog zu 1) liefert dies den $(1 - \alpha)$ -Konfidenzbereich:

$$C'(X) = \{m \in \mathbb{R} : T_n(m) < q_{1-\alpha}\}.$$



Welcher den Niveau α -Test $\varphi(X) = 1$ falls $T_n(m_0) \geq q_{1-\alpha}$. ($\Leftrightarrow m_0 \notin C'(X)$)

- c) Mit der rechtsseitigen alternative $H_1 : m > m_0$ erhalten wir dann den $(1 - \alpha)$ Konfidenzbereich

$$C'''(X) = \{m \in \mathbb{R} : T_n(m) > q_\alpha\}$$

Analog ergibt sich der Niveau α -Test $\varphi(X) = 1$ falls $T_n(m_0) < q_{1-\alpha}$

2. Likelihood

In diesem Kapitel betrachten wir stets ein reguläres statistisches Modell mit Dichten bzw. Massenfunktionen $f_\theta(x)$ ($x \in S, \theta \in \Theta$). Angenommen, wir beobachten einen Wert x . Dann können wir $f_\theta(x)$ als Maß für die „Plausibilität“ des Wertes x bezüglich des Parameters θ interpretieren.

Definition 2.1 (Likelihood). Die Funktion

$$L(\theta; x) = f_\theta(x), \quad \theta \in \Theta,$$

heißt *Likelihood-Funktion* des statistischen Modells bei Beobachtungswert x . Die *log-Likelihood-Funktion* ist definiert als natürlicher Logarithmus der Likelihood-Funktion, d.h.

$$\ell(\theta; x) = \log f_\theta(x), \quad \theta \in \Theta,$$

wobei wir $\log 0 := -\infty$ setzen.

Wichtig ist, dass wir hier die Dichte bzw. Massenfunktion als Funktion des unbekanntem Parameters θ (bzw. als Funktion von θ und x) betrachten, während in der Wahrscheinlichkeitstheorie θ üblicherweise einen festen Wert hat. Die Likelihood-Funktion ermöglicht uns eine einfache ad hoc Konstruktion von verschiedenen statistischen Verfahren.

2.1. Das Maximum-Likelihood-Prinzip

Das Maximum-Likelihood-Prinzip ist ein einfaches und allgemeines ad hoc Verfahren zur Konstruktion von Parameterschätzern. Beim Beobachtungswert x wählen wir als Schätzwert für θ denjenigen Parameterwert $T(x)$, bezüglich dessen x „am plausibelsten“ ist, das heißt für den

$$L(T(x); x) = \max_{\theta \in \Theta} L(\theta; x)$$

gilt. Im Allgemeinen ist dieser Wert nicht eindeutig, da es mehrere globale Maxima geben kann. Daher definieren wir:

Definition 2.2 (Maximum-Likelihood-Schätzer). Eine Statistik $\hat{\theta} = T(X)$ heißt *Maximum-Likelihood-Schätzer* (engl. Maximum Likelihood Estimator, kurz MLE) für den Parameter θ falls

$$T(x) \in \operatorname{argmax}_{\theta \in \Theta} L(\theta; x) \quad \text{für alle } x \in S$$

gilt, wobei $\operatorname{argmax}_{\theta \in \Theta} L(\theta; x)$ die Menge aller globalen Maxima der Funktion $\theta \mapsto L(\theta; x)$ bezeichnet.

Da der Logarithmus eine strikt monoton wachsende Funktion ist, können wir zur Berechnung des MLE statt der Likelihood-Funktion auch die Log-Likelihood-Funktion maximieren.

Beispiele. a) TAXIPROBLEM. Die Taxis in einer Stadt sind von 1 bis N durchnummeriert. Wir beobachten n verschiedene Taxis, und wollen den Wert des unbekanntem Parameters $\theta = N$ schätzen. Sei $X = (X_1, \dots, X_n)$ das zufällige n -Tupel mit den beobachteten Taxinummern. Sei S die Menge aller n -Tupel $x = (x_1, \dots, x_n)$ mit $x_i \in \mathbb{N}$ und $x_i \neq x_j$ für $i \neq j$. Wir nehmen an, dass X unter P_N gleichverteilt ist auf der Teilmenge

$$S_N = \{x \in S : x_1, \dots, x_n \in \{1, 2, \dots, N\}\} = \{x \in S : \max(x_1, \dots, x_n) \leq N\}.$$

Die entsprechende Likelihood-Funktion

$$L(N; x) = f_N(x) = \begin{cases} 1/|S_N| & \text{für } \max(x_1, \dots, x_n) \leq N, \\ 0 & \text{sonst,} \end{cases}$$

ist maximal für $N = \max(x_1, \dots, x_n)$. Also ist

$$\hat{N} = \max(X_1, \dots, X_n)$$

der (eindeutige) Maximum-Likelihood-Schätzer für N .

- b) CAPTURE-RECAPTURE-VERFAHREN. Wir wollen die Größe N einer unbekanntem Population schätzen. Dazu markieren wir im ersten Schritt ℓ zufällig ausgewählte Objekte, und entnehmen im zweiten Schritt eine unabhängige Zufallsstichprobe der Größe n . Die zufällige Anzahl X der markierten Objekte in der zweiten Stichprobe ist dann unter P_N hypergeometrisch verteilt mit Parametern N, ℓ, n . Nachrechnen zeigt, dass der ganzzahlige Anteil

$$\hat{N} = \left\lfloor \frac{n\ell}{X} \right\rfloor$$

ein Maximum-Likelihood-Schätzer für N ist. Die Details sind eine Übungsaufgabe.

- c) SCHÄTZEN EINER UNBEKANNTEN WAHRSCHEINLICHKEIT. Sei $p \in [0, 1]$ eine unbekanntem Wahrscheinlichkeit. Beispielsweise ist p bei einer Wahlprognose der relative Stimmenanteil einer Kandidatin unter allen Wählern. Diese relative Häufigkeit können wir als Wahrscheinlichkeit bezüglich der entsprechenden empirischen Verteilung deuten. Um p zu schätzen, führen wir eine Befragung von n zufällig ausgewählten Wählern durch. Sei $X_i = 1$ falls der i -te Wähler beabsichtigt, für die Kandidatin zu stimmen, und $X_i = 0$ sonst. Da die Gesamtpopulation aller Wähler sehr viel größer als unsere Stichprobe ist, nehmen wir der Einfachheit halber an, dass X_1, \dots, X_n unter P_p unabhängige und zum Parameter p Bernoulli-verteilte Zufallsvariablen sind. Die Likelihood und die Log-Likelihood sind dann für $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ gegeben durch

$$\begin{aligned} L(p; x) &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}, \\ \ell(p; x) &= \log(p) \sum_{i=1}^n x_i + \log(1-p) \sum_{i=1}^n (1-x_i). \end{aligned}$$

Um das Maximum zu bestimmen, berechnen wir die erste Ableitung der Log-Likelihood nach p :

$$\frac{\partial \ell}{\partial p}(p; x) = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (1-x_i).$$

An der Stelle $p = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n$ hat die erste Ableitung einen Vorzeichenwechsel von positiven zu negativen Werten. Also ist die Log-Likelihood an dieser Stelle maximal, und somit ist der Stichprobenmittelwert

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

der (eindeutige) Maximum-Likelihood-Schätzer für p .

2. Likelihood

- d) **SCHÄTZEN EINER UNBEKANNTEN WAHRSCHEINLICHKEITSVERTEILUNG.** Das letzte Beispiel können wir folgendermaßen verallgemeinern: Angenommen, wir haben nicht zwei mögliche Werte (0 und 1), sondern k mögliche Werte a_1, a_2, \dots, a_k mit unbekanntem Wahrscheinlichkeiten p_1, \dots, p_k . Beispielsweise sind p_1, \dots, p_k die Stimmenanteile von k verschiedenen Parteien unter allen Wählern. Der Parameterraum ist dann die Menge aller Wahrscheinlichkeitsverteilungen auf $\{a_1, \dots, a_k\}$, d.h.

$$\Theta = \text{WV}(\{a_1, \dots, a_k\}) = \left\{ p = (p_1, \dots, p_k) : p_i \geq 0 \forall i, \sum_i p_i = 1 \right\}.$$

Geometrisch ist Θ ein $(k - 1)$ -dimensionales Simplex im \mathbb{R}^k . Gegeben seien wieder die Werte x_1, \dots, x_n von n unabhängigen Einzelstichproben. Wir nehmen an, dass diese Beobachtungswerte Realisierungen von unabhängigen Zufallsvariablen X_1, \dots, X_n mit Verteilung

$$P_p[X_i = a_l] = p_l \quad \text{für } l = 1, \dots, k$$

sind. Die Likelihood-Funktion ist dann gegeben durch

$$L(p; x_1, \dots, x_n) = \prod_{i=1}^n p_{x_i} = \prod_{l=1}^k p_l^{h_l},$$

wobei h_l die Häufigkeit des Werts a_l in der Stichprobe $x = (x_1, \dots, x_n)$ ist. Wir behaupten, dass der Maximum-Likelihood-Schätzer \hat{p} für p durch die empirische Verteilung gegeben ist, d.h.

$$\hat{p}_l = \frac{H_l}{n} = \text{relative Häufigkeit von } a_l \text{ unter } X_1, \dots, X_n.$$

Zum Beweis bestimmt man das Maximum der Log-Likelihood

$$\ell(p; x) = \sum_{l=1}^k h_l \log(p_l)$$

unter der Nebenbedingung $\sum_{l=1}^k p_l = 1$ mithilfe von Lagrange-Multiplikatoren. Wegen

$$\frac{\partial \ell}{\partial p_l}(p; x) = \frac{h_l}{p_l} \quad \text{und} \quad \frac{\partial}{\partial p_l} \sum_{l=1}^k p_l = 1$$

führt dies unter Beachtung der Nebenbedingung auf die notwendige Bedingung $p_l = h_l/n$ für ein Maximum im Inneren des Simplex, und wegen

$$\frac{\partial^2 \ell}{\partial p_j \partial p_l}(p; x) = -\frac{h_l}{p_l^2} \delta_{jl}$$

ist die Log-Likelihood strikt konkav, und damit ist die notwendige Bedingung auch hinreichend.

- e) **GAUßSCHES PRODUKTMODELL.** In einem Experiment beobachten wir n unabhängige Messwerte x_1, \dots, x_n einer reellwertigen Messgröße. Wir gehen davon aus, dass die Fluktuationen der Messwerte durch die additive Überlagerung vieler kleiner unabhängiger Zufallseffekte entstehen. Aufgrund des zentralen Grenzwertsatzes liegt es dann nahe, die Werte x_1, \dots, x_n als Realisierungen von unabhängigen Zufallsvariablen X_1, \dots, X_n mit Normalverteilung $N(m, v)$ zu deuten. Dabei sind im Allgemeinen sowohl der Mittelwert m als auch die Varianz v unbekannt Parameter, d.h. die Parametermenge ist

$$\Theta = \{\theta = (m, v) : m \in \mathbb{R}, v \in [0, \infty)\}.$$

Die Likelihood-Funktion ist dann durch die Produktdichte

$$L(m, v; x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x_i - m)^2}{2v}}$$

gegeben, und für die Log-Likelihood erhalten wir

$$\begin{aligned}\ell(m, v; x_1, \dots, x_n) &= -\frac{n}{2} \log(2\pi v) - \frac{1}{2v} \sum_{i=1}^n (x_i - m)^2 \\ &= -\frac{n}{2} \log(2\pi v) - \frac{1}{2v} \sum_{i=1}^n (x_i - \bar{x}_n)^2 - \frac{n}{2v} (\bar{x}_n - m)^2.\end{aligned}\quad (2.1)$$

- (i) *Schätzen des Erwartungswerts m bei bekannter Varianz v .* Wenn v fest ist, dann ist die Log-Likelihood maximal für $m = \bar{x}_n$. Der MLE ist also

$$\hat{m} = \bar{X}_n.$$

- (ii) *Schätzen der Varianz v bei bekanntem Erwartungswert m .* Wenn m fest ist, dann können wir das Maximum der Log-Likelihood bestimmen, indem wir die partielle Ableitung

$$\frac{\partial}{\partial v} \ell(m, v; x_1, \dots, x_n) = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^n (x_i - m)^2$$

betrachten. Diese verschwindet für $v = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$, und zum Beispiel durch Betrachten der zweiten Ableitung sieht man, dass dies das eindeutige Maximum ist. Also ist

$$\hat{v} = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$$

der eindeutige MLE für v .

- (iii) *Schätzen von Erwartungswert und Varianz.* Üblicherweise kennt man weder den Erwartungswert noch die Varianz. Der unbekannte Parameter $\theta = (m, v)$ ist dann zweidimensional. Das Maximum können wir auch in diesem Fall einfach bestimmen: Für jeden Wert von v ist ℓ maximal für $m = \bar{x}_n$, und der entsprechende Wert der Log-Likelihood wird am größten, wenn wir $v = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ wählen. Der MLE ist also

$$\hat{\theta} = \left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right).$$

2.2. Suffiziente Statistiken

Datensätze können oft sehr groß sein. Wenn wir beispielsweise bei einer Wahlbefragung 5000 Wählerinnen und Wähler befragen, dann ist unser Beobachtungswert zunächst ein 5000 dimensionaler Vektor, der als Komponenten die Angaben aller Befragten enthält. Tatsächlich ist aber intuitiv klar, dass für die Wahlprognose nur relevant ist, wieviele der Befragten jeweils für die einzelnen Parteien stimmen wollen. Für statistische Rückschlüsse sollte die entsprechende Häufigkeitsverteilung $H = (H_1, \dots, H_k)$ ausreichend sein, wobei H_1, \dots, H_k die Häufigkeiten der k Parteien in der Stichprobe sind. Eine solche Statistik nennt man *suffizient*; für Rückschlüsse auf den unbekannt Parameter sollte nur der Wert der suffizienten Statistik relevant sein. Wir wollen den Begriff nun mathematisch formalisieren. Wir beginnen mit einer praktischen Definition, die sich in Anwendungsbeispielen leicht nachprüfen lässt. Im Anschluss zeigen wir, dass diese Definition äquivalent zu einer anderen anschaulichen Bedingung ist.

Definition 2.3 (Suffiziente Statistik). Eine Statistik $T(X)$ heißt *suffizient* für den unbekannt Parameter θ , falls sich die Likelihood-Funktion in der Form

$$L(\theta; x) = g_\theta(T(x)) h(x) \quad \text{für alle } \theta \in \Theta \text{ und } x \in S \quad (2.2)$$

mit messbaren Funktionen $g_\theta : \mathbb{R} \rightarrow [0, \infty)$ ($\theta \in \Theta$) und $h : S \rightarrow [0, \infty)$ darstellen lässt.

2. Likelihood

Anschaulich hängt die Likelihood also nur über die suffiziente Statistik $T(x)$ vom Parameter θ ab. Diese Anschauung werden wir gleich noch präzisieren. Gilt $h(x) > 0$, dann folgt aus (2.2) unmittelbar, dass der Maximum-Likelihood-Schätzer nur von der suffizienten Statistik $T(x)$ abhängt, vorausgesetzt, das Maximum der Likelihood-Funktion ist eindeutig. Ebenso hängen andere Likelihood-basierte statistische Verfahren (z.B. Likelihood-Quotienten-Tests, s.u.) nur von den Werten der suffizienten Statistik ab.

Beispiel (Schätzen einer unbekanntem Wahrscheinlichkeit). In Beispiel c) von oben ist

$$L(p; x) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = (1-p)^n \left(\frac{p}{1-p} \right)^{\sum_{i=1}^n x_i}.$$

Also ist die Häufigkeit $H_1 = \sum_{i=1}^n X_i$ von „1“ eine suffiziente Statistik, und ebenso die relative Häufigkeit $\bar{X}_n = H_1/n$.

Beispiel (Schätzen einer unbekanntem Wahrscheinlichkeitsverteilung). Entsprechend hängt in Beispiel d) von oben die Likelihood-Funktion nur von den Häufigkeiten der möglichen Werte a_1, \dots, a_k ab. Daher ist der Histogrammvektor $H = (H_1, \dots, H_k)$ der Stichprobe eine suffiziente Statistik, und ebenso die empirische Verteilung, d.h. der Vektor H/n der relativen Häufigkeiten.

Wir geben nun eine äquivalente Charakterisierung von suffizienten Statistiken. Dabei beschränken wir uns der Einfachheit halber auf den diskreten Fall. Eine entsprechende Aussage gilt aber auch allgemein, wenn man sie mithilfe von allgemeinen bedingten Erwartungen formuliert.

Lemma 2.4 (Charakterisierung von Suffizienz). *Ist S abzählbar, und $T : S \rightarrow \mathbb{R}$ eine Abbildung, dann ist die Statistik $T(X)$ genau dann suffizient für den unbekanntem Parameter θ , wenn die bedingten Wahrscheinlichkeiten $P_\theta[X = x | T(X) = t]$ für x, t mit $P[T(X) = t] \neq 0$ nicht von θ abhängen.*

Mit anderen Worten: *Eine Statistik $T(X)$ ist genau dann suffizient, wenn die bedingte Verteilung von X gegeben $T(X)$ nicht von θ abhängt.*

Beweis. “ \Leftarrow ”: Wir nehmen zunächst an, dass die bedingten Wahrscheinlichkeiten nicht von θ abhängen. Dann erhalten wir für die Likelihood

$$\begin{aligned} L(\theta; x) &= P_\theta[X = x] = P_\theta[X = x \text{ und } T(X) = T(x)] \\ &= P_\theta[X = x | T(X) = T(x)] \cdot P_\theta[T(X) = T(x)] = h(x) \cdot g_\theta(T(x)) \end{aligned}$$

mit geeigneten Funktionen h und g_θ .

“ \Rightarrow ”: Gilt umgekehrt $L(\theta; x) = g_\theta(T(x))h(x)$ mit Funktionen g_θ und h , dann folgt

$$\begin{aligned} P_\theta[X = x | T(X) = t] &= \frac{P_\theta[X = x, T(X) = t]}{P_\theta[T(X) = t]} = \frac{P_\theta[X = x] \cdot 1_{T(x)=t}}{\sum_{a: T(a)=t} P_\theta[X = a]} \\ &= \frac{g_\theta(t)h(x)}{\sum_{a: T(a)=t} g_\theta(t)h(a)} = \frac{h(x)}{\sum_{a: T(a)=t} h(a)}, \end{aligned}$$

und somit hängen die bedingten Wahrscheinlichkeiten nicht von θ ab. ■

Darstellung als Zweistufenmodell

Ist $T(X)$ eine suffiziente Statistik, dann können wir das zugrundeliegende Zufallsexperiment “Ziehen einer Stichprobe x ” als ein Zweistufenmodell darstellen, in dem der unbekanntem Parameter θ nur in der zweiten Stufe eingeht:

1. Ziehe Stichprobe t von der Verteilung der suffizienten Statistik $T(X)$ (hängt von θ ab)
2. Ziehe Stichprobe x von der bedingten Verteilung von X gegeben $T(X)$ (hängt **nicht** von θ ab)

Wir betrachten unter diesem Aspekt noch einmal einige der Beispiele von oben.

Beispiele. c) SCHÄTZEN EINER UNBEKANNTEN WAHRSCHEINLICHKEIT. Hier ist $T(X) = \sum_{i=1}^n X_i$ eine suffiziente Statistik. $T(X)$ ist die Häufigkeit von "1", also binomialverteilt mit Parametern n und p . Die bedingte Verteilung von X gegeben $T(X) = t$ ist die Gleichverteilung auf der Menge S_t aller $x \in \{0, 1\}^n$ mit $\sum_{i=1}^n x_i = t$. Also erhalten wir die folgende Darstellung als Zweistufenmodell:

1. Ziehe $t \sim \text{Bin}(n, p)$.
2. Ziehe $x = (x_1, \dots, x_n) \sim \text{Unif}(S_t)$.

Der unbekannte Parameter p geht nur im ersten Schritt ein.

- d) SCHÄTZEN EINER UNBEKANNTEN WAHRSCHEINLICHKEITSVERTEILUNG. In diesem Modell ist der Histogrammvektor $H = (H_1, \dots, H_k)$ eine suffiziente Statistik, die multinomialverteilt ist mit Parametern n und p . Überlegen Sie sich selbst, wie die Darstellung als Zweistufenmodell aussieht.
- e) GAUßSCHES PRODUKTMODELL. Aus der Formel (2.1) für die log-Likelihood folgt, dass

$$T(X_1, \dots, X_n) = \left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right)$$

eine suffiziente Statistik im Gauß-Modell ist. Diese Statistik enthält sowohl den empirischen Mittelwert als auch die empirische Varianz. Die Verteilung von $T(X)$ werden wir in Abschnitt 2.5 berechnen. Die bedingte Verteilung von X gegeben $T(X) = (m, v)$ ist eine Gleichverteilung auf dem Schnitt der Sphäre mit Mittelpunkt $(m, \dots, m) \in \mathbb{R}^n$ und Radius \sqrt{nv} mit der Hyperebene $\{x \in \mathbb{R}^n : \sum x_i = nm\}$. Eine andere suffiziente Statistik ist $\tilde{T}(X) = (\sum X_i, \sum X_i^2)$.

- f) ALLGEMEINES PRODUKTMODELL. In einem allgemeinen Produktmodell mit n identischen reellen Faktoren mit stetiger Verteilung können wir die Likelihood schreiben als

$$L(\theta; x) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n f_\theta(x_{(i)})$$

wobei $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die *Ordnungsstatistiken*, d.h. die der Größe nach geordneten Werte x_1, \dots, x_n sind. Also ist

$$T(X) = (X_{(1)}, X_{(2)}, \dots, X_{(n)})$$

eine suffiziente Statistik. Diese enthält nur die Informationen über die vorkommenden Datenwerte, aber nicht über deren Reihenfolge. Das Zweistufenmodell sieht in diesem Fall folgendermaßen aus:

1. Ziehe $y = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$ von der Verteilung mit Dichte $n! \prod_{i=1}^n f_\theta(y_i)$ auf der Menge aller $(y_1, \dots, y_n) \in \mathbb{R}^n$ mit $y_1 \leq y_2 \leq \dots \leq y_n$.
2. Wähle die Positionen (Ränge) $R(1), \dots, R(n)$ von x_1, \dots, x_n zufällig gemäß der Gleichverteilung auf der Menge \mathcal{S}_n aller Permutationen von $\{1, \dots, n\}$, und setze $x_i := y_{R(i)}$.

Die Beispiele zeigen: Je größer die betrachtete Klasse von Wahrscheinlichkeitsverteilungen in unserem Modell ist, desto mehr Informationen müssen in einer suffizienten Statistik enthalten sein. Im Extremfall, in dem wir alle Wahrscheinlichkeitsverteilungen auf dem Grundraum in Betracht ziehen, enthält eine suffiziente Statistik die komplette Information.

Verbessern von Schätzern

Sei $T(X)$ eine suffiziente Statistik für den Parameter $\theta \in \Theta$ und $g : \Theta \rightarrow \mathbb{R}$ eine Funktion. Ist der Maximum-Likelihood-Schätzer eindeutig, dann hängt dieser nur von $T(X)$ ab. Wir zeigen nun, dass wir einen beliebigen Schätzer, der keine Funktion von $T(X)$ ist, mithilfe von $T(X)$ verbessern können. Dazu

2. Likelihood

nehmen wir der Einfachheit halber an, dass der Wertebereich $T(S)$ abzählbar ist. In diesem Fall ist die bedingte Erwartung einer reellwertigen Zufallsvariable Y gegeben $T(X)$ bezüglich P_θ definiert als

$$E_\theta [Y | T(X)] = \sum_{t: P_\theta[T(X)=t] > 0} E_\theta [Y | T(X) = t] \cdot 1_{\{T(X)=t\}}. \quad (2.3)$$

Die bedingte Erwartung von Y gegeben $T(X)$ ist also wieder eine *Zufallsvariable*, deren Wert auf der Menge $\{T(X) = t\}$ die bedingte Erwartung von Y gegeben $T(X) = t$ ist. Insbesondere ist die bedingte Erwartung eine Funktion von $T(X)$.

Satz 2.5 (Rao-Blackwell). Sei \hat{g} ein Schätzer für $g(\theta)$ und $T(X)$ eine suffiziente Statistik mit abzählbarem Wertebereich. Dann ist die bedingte Erwartung $\tilde{g} := E_\theta[\hat{g}|T(X)]$ ein Schätzer für $g(\theta)$ mit

$$\text{Bias}_\theta(\tilde{g}) = \text{Bias}_\theta(\hat{g}) \quad \text{für alle } \theta \in \Theta, \quad \text{und} \quad (2.4)$$

$$\text{MSE}_\theta(\tilde{g}) \leq \text{MSE}_\theta(\hat{g}) \quad \text{für alle } \theta \in \Theta. \quad (2.5)$$

Beweis. Nach Lemma 2.4 folgt aus der Suffizienz von $T(X)$, dass \tilde{g} nicht von θ abhängt. Außerdem ist \tilde{g} nach Definition eine Funktion von $T(X)$, also insbesondere eine Statistik. Mithilfe der Definition der bedingten Erwartung und der Cauchy-Schwarz-Ungleichung erhalten wir

$$\begin{aligned} E_\theta[\tilde{g}] &= \sum_{t: P_\theta[T(X)=t] > 0} E_\theta[\hat{g} | T(X) = t] \cdot P_\theta[T(X) = t] = E_\theta[\hat{g}], \quad \text{und} \\ \text{MSE}_\theta[\tilde{g}] &= E_\theta[(\tilde{g} - g(\theta))^2] = \sum_{t: P_\theta[T(X)=t] > 0} (E_\theta[\hat{g} | T(X) = t] - g(\theta))^2 \cdot P_\theta[T(X) = t] \\ &\leq \sum_{t: P_\theta[T(X)=t] > 0} E_\theta[(\hat{g} - g(\theta))^2 | T(X) = t] \cdot P_\theta[T(X) = t] = \text{MSE}_\theta[\hat{g}] \end{aligned}$$

für alle $\theta \in \Theta$. ■

Bedingte Erwartungen werden systematisch in der Vorlesung »Stochastic Processes« behandelt. Mit der dort gegebenen allgemeinen Definition kann man den Satz von Rao-Blackwell auf den Fall erweitern, dass der Wertebereich von $T(X)$ nicht abzählbar ist.

Beispiel (Bernoulli-Modell). Seien X_1, \dots, X_n unter P_p unabhängig und Bernoulli(p) verteilt. Wegen $E_p[X_1] = p$ ist $\hat{p} = X_1$ ein erwartungstreuer Schätzer für p . Dieser Schätzer ist selbst nicht sonderlich interessant, da er nur die Werte 0 oder 1 annimmt. Mithilfe des Satzes von Rao-Blackwell können wir aber daraus einen besseren erwartungstreuen Schätzer \tilde{p} konstruieren, der auf der suffizienten Statistik $T(X) = X_1 + \dots + X_n$ basiert. Wir erhalten

$$\tilde{p} = E_p[X_1 | T(X)] = \frac{X_1 + \dots + X_n}{n} = \bar{X}_n.$$

Hierbei haben wir benutzt, dass aus Symmetriegründen die bedingte Erwartung von X_i gegeben $T(X)$ nicht von i abhängt. Daher gilt

$$E_p[X_1 | T(X)] = \frac{1}{n} E_p[X_1 + \dots + X_n | T(X)] = \frac{1}{n} E_p[T(X) | T(X)] = \frac{1}{n} T(X).$$

2.3. Exponentielle Familien

Wir betrachten nun eine wichtige Klasse von statistischen Modellen, die viele der üblichen Modelle umfasst.

Definition 2.6 (Exponentielle Familie).

- (i) Eine *exponentielle Familie* ist ein reguläres statistisches Modell mit Dichten bzw. Massenfunktionen, die sich in der Form

$$f_{\theta}(x) = e^{c(\theta) \cdot T(x) + d(\theta) + U(x)} 1_A(x) \quad (2.6)$$

mit $l \in \mathbb{N}$, messbaren Funktionen $T : S \rightarrow \mathbb{R}^l$ und $U : S \rightarrow \mathbb{R}$, einer messbaren Menge $A \subseteq S$, und Funktionen $c : \Theta \rightarrow \mathbb{R}^l, d : \Theta \rightarrow \mathbb{R}$ darstellen lässt.

- (ii) Eine exponentielle Familie ist *in natürlicher Form*, falls $c(\theta) = \theta$ gilt.

Die Dichte bzw. Massenfunktion einer exponentiellen Familie können wir auch schreiben als

$$f_{\theta}(x) = \frac{1}{\mathcal{Z}(\theta)} e^{c(\theta) \cdot T(x)} h(x) \quad (2.7)$$

mit der *Normierungskonstanten* $Z(\theta) := e^{-d(\theta)}$ und der *Referenzdichte* $h(x) := e^{U(x)} 1_A(x)$. Die Funktionen T und U sind nicht eindeutig festgelegt. In vielen Fällen kann man exponentielle Familien durch eine Substitution im Parameterraum in natürliche Form bringen. Daher werden wir oft nur diesen Fall betrachten.

Bemerkung (Suffiziente Statistik). In einer exponentiellen Familie ist $T(X)$ eine suffiziente Statistik.

Bemerkung (Boltzmann-Verteilung). In der statistischen Physik treten exponentielle Familien als Gleichgewichtsverteilungen auf. Beispielsweise hat die Verteilung im thermodynamischen Gleichgewicht in einem abgeschlossenen System bei inverser Temperatur $\beta = 1/T$ die Dichte bzw. Massenfunktion

$$f_{\beta}(x) = \frac{1}{\mathcal{Z}(\beta)} e^{-\beta H(x)},$$

wobei $H(x)$ die Energie des Zustands x ist. Die Normierungskonstante $Z(\beta)$ heißt in der statistischen Physik *Partitionsfunktion*.

Wir betrachten nun zunächst einige elementare Beispiele von exponentiellen Familien:

Beispiele. a) EXPONENTIALVERTEILUNGEN. Die Dichte der Exponentialverteilung mit Parameter $\lambda \in (0, \infty)$ ist

$$f_{\lambda}(x) = \lambda e^{-\lambda x} 1_{(0, \infty)}(x) = e^{-\lambda x + \log \lambda} 1_{(0, \infty)}(x).$$

Die Exponentialverteilungen bilden also eine exponentielle Familie mit $T(x) = x$, $U(x) = 0$, $A = (0, \infty)$, $c(\lambda) = -\lambda$, und $d(\lambda) = \log \lambda$.

b) BERNOULLI-, BINOMIAL- UND POISSON-VERTEILUNGEN. Die Massenfunktion der Bernoulli-Verteilung mit Parameter p ist

$$f_p(x) = p^x (1-p)^{1-x} = e^{\log(p/(1-p))x + \log(1-p)}, \quad x \in \{0, 1\}.$$

Dies ist eine exponentielle Familie mit $T(x) = x$, $U(x) = 0$, $c(p) = \log(p/(1-p))$ und $d(p) = \log(1-p)$. Entsprechend gilt für die Binomialverteilungen

$$f_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x} = e^{\log(p/(1-p))x + n \log(1-p) + \log \binom{n}{x}}, \quad x \in \{0, 1, \dots, n\}.$$

Für festes n und variables p bilden diese eine exponentielle Familie mit $T(x) = x$ und $U(x) = \log \binom{n}{x}$. Die Poisson-Verteilungen bilden eine exponentielle Familie mit $T(x) = x$ und $U(x) = -\log(x!)$, denn

$$f_{\lambda}(x) = \frac{\lambda^x}{x!} e^{-\lambda} = e^{\log(\lambda)x - \lambda - \log(x!)}, \quad x \in \mathbb{N}_0.$$

2. Likelihood

c) *Normalverteilungen.* Die Dichte der eindimensionalen Normalverteilung $N(m, v)$ ist

$$f_{m,v}(x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} = e^{c_1(m,v)x + c_2(m,v)x^2 + d(m,v)}$$

mit $c_1(m, v) = m/v$, $c_2(m, v) = -1/(2v)$ und $d(m, v) = -\frac{1}{2} \left(\frac{m^2}{2v} + \log(2\pi v) \right)$. Die Normalverteilungen bilden also eine *zweiparametrische* exponentielle Familie (d.h. $l = 2$) mit $T(x) = (x, x^2)$ und $U(x) = 0$.

Faktorisierung

Eine wichtige Eigenschaft exponentieller Familien ist die Stabilität unter Produktbildung. Sind X_1, \dots, X_n unter P_θ unabhängige identisch verteilte Zufallsvariablen mit Dichten bzw. Massenfunktionen $f_\theta(x_i)$, wobei $(f_\theta)_{\theta \in \Theta}$ eine exponentielle Familie ist, dann ist die Likelihood gegeben durch

$$L(\theta; x) = \prod_{i=1}^n f_\theta(x_i) = \exp \left(c(\theta) \cdot \sum_{i=1}^n T(x_i) + nd(\theta) + \sum_{i=1}^n U(x_i) \right) 1_{A_1 \times \dots \times A_n}(x_1, \dots, x_n). \quad (2.8)$$

Das Produktmodell ist also wieder eine exponentielle Familie mit

$$\tilde{T}(x) = \sum_{i=1}^n T(x_i), \quad \tilde{U}(x) = \sum_{i=1}^n U(x_i), \quad \tilde{A} = A_1 \times \dots \times A_n. \quad (2.9)$$

Erwartungswerte und Kovarianzen

Wir betrachten nun eine exponentielle Familie in natürlicher Form. Sei $\Theta \subseteq \mathbb{R}^l$ und

$$f_\theta(x) = e^{\theta \cdot T(x)} e^{d(\theta)} h(x), \quad \theta \in \Theta.$$

Wie nehmen an, dass die Verteilungen absolutstetig sind - im diskreten Fall gelten entsprechende Aussagen mit Summen statt Integralen. Wie zuvor bezeichnen wir mit $\mathcal{Z}(\theta)$ die Normierungskonstante der Verteilung mit Dichte f_θ , d.h.

$$\mathcal{Z}(\theta) = e^{-d(\theta)} = \int e^{\theta \cdot T(x)} h(x) dx.$$

Sei $\overset{\circ}{\Theta} := \Theta \setminus \partial\Theta$ das Innere des Parameterbereichs.

Lemma 2.7 (Berechnung der Momente in exponentiellen Familien). *Es gilt $d \in C^2(\overset{\circ}{\Theta})$, und für $\theta \in \overset{\circ}{\Theta}$,*

$$E_\theta [T_i(X)] = -\frac{\partial d}{\partial \theta_i}(\theta), \quad (2.10)$$

$$\text{Cov}_\theta [T_i(X), T_j(X)] = -\frac{\partial^2 d}{\partial \theta_i \partial \theta_j}(\theta). \quad (2.11)$$

Beweis. Zum Beweis betrachten wir für ein festes $\theta \in \overset{\circ}{\Theta}$ die momentenerzeugende Funktion

$$M(s) := E_\theta [e^{s \cdot T(x)}] = \int e^{s \cdot T(x)} e^{\theta \cdot T(x)} h(x) dx, \quad s \in \mathbb{R}^l.$$

Für $\theta \in \overset{\circ}{\Theta}$ gilt $\int e^{(s+\theta) \cdot T(x)} h(x) dx = \mathcal{Z}(s+\theta) < \infty$ falls $|s|$ hinreichend klein ist. In diesem Fall erhalten wir

$$M(s) = \frac{\mathcal{Z}(s+\theta)}{\mathcal{Z}(\theta)} = e^{d(\theta) - d(s+\theta)} < \infty.$$

Hieraus folgt, dass die momentenerzeugende Funktion in einer Umgebung der 0 beliebig oft stetig differenzierbar ist, da die entsprechende Potenzreihe absolut konvergiert. Die Momente ergeben sich dann durch Ableiten der momentenerzeugenden Funktion. Insbesondere erhalten wir

$$\begin{aligned} E_{\theta}[T_i(X)] &= \frac{\partial M}{\partial s_i}(0) = -\frac{\partial d}{\partial \theta_i}(\theta), \\ E_{\theta}[T_i(X)T_j(X)] &= \frac{\partial^2 M}{\partial s_i \partial s_j}(0) = -\frac{\partial^2 d}{\partial \theta_i \partial \theta_j}(\theta) + \frac{\partial d}{\partial \theta_i}(\theta) \frac{\partial d}{\partial \theta_j}(\theta), \end{aligned}$$

und damit $\text{Cov}_{\theta}[T_i(X), T_j(X)] = -\frac{\partial^2 d}{\partial \theta_i \partial \theta_j}(\theta)$. ■

Die Aussage aus dem Lemma können wir benutzen, um den Maximum-Likelihood-Schätzer in einer exponentiellen Familie zu charakterisieren. Dazu beschränken wir uns der Einfachheit halber auf den Fall einer einparametrischen exponentiellen Familie.

Satz 2.8 (Maximum-Likelihood-Schätzer in exponentiellen Familien). Sei $\Theta \subseteq \mathbb{R}$ ein offenes Intervall. Dann ist eine Statistik $\hat{\theta}$ mit Werten in Θ genau dann ein Maximum-Likelihood-Schätzer für θ , wenn

$$E_{\hat{\theta}}[T(X)] = -d'(\theta) = T(X).$$

Eine entsprechende Aussage folgt mit analogem Beweis auch für mehrparametrische exponentielle Familien.

Beweis. Die log-Likelihood ist gegeben durch

$$\ell(\theta; x) = \theta \cdot T(x) + d(\theta) + \log h(x).$$

Damit erhalten wir für $\theta \in \Theta$ nach dem Lemma und der Voraussetzung

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell(\theta; x) &= T(x) + d'(\theta) = T(x) - E_{\theta}[T(X)], \\ \frac{\partial^2}{\partial \theta^2} \ell(\theta; x) &= d''(\theta) = -\text{Var}_{\theta}[T(X)] \leq 0. \end{aligned}$$

Insbesondere ist die log-Likelihood-Funktion konkav. Also ist θ genau dann eine Maximalstelle von $\theta \mapsto \ell(\theta; x)$, wenn $E_{\theta}[T(X)] = T(x)$ gilt. ■

Ist $T(X)$ für $\theta \in \Theta$ nicht P_{θ} -fast sicher konstant, dann ist die log-Likelihood sogar strikt konkav, und die Funktion $\theta \mapsto E_{\theta}[T(X)]$ ist streng monoton. In diesem Fall ist der Maximum-Likelihood-Schätzer eindeutig. Die Existenz ist hingegen nicht gewährleistet, da das Maximum zum Beispiel auch auf dem Rand des Intervalls Θ angenommen werden könnte, welcher hier nicht in der Parametermenge enthalten ist. Im Allgemeinen ist der MLE nicht explizit, sondern nur numerisch berechenbar.

Beispiel (Schätzen der Intensität einer Exponentialverteilung). Sind X_1, \dots, X_n unter P_{θ} unabhängige Einzelstichproben von einer Exponentialverteilung mit unbekannter Intensität θ , dann gilt

$$f_{\theta}(x) = \prod_{i=1}^n \theta e^{-\theta x_i} = e^{d(\theta) + \theta T(x)}$$

mit $d(\theta) = n \log \theta$ und $T(x) = -\sum x_i$. Aus der Gleichung $d'(\hat{\theta}) = -T(X)$ ergibt sich der Maximum-Likelihood-Schätzer

$$\hat{\theta} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}_n}.$$

2. Likelihood

Beispiel (MLE im Produktfall). Wie in (2.8) betrachten wir wieder eine exponentielle Familie basierend auf n unabhängigen identisch verteilten Stichproben. Nach (2.9) ist in diesem Fall $\tilde{T}(x) = \sum_{i=1}^n T(x_i)$, und damit

$$E_{\theta}[\tilde{T}(X_1, \dots, X_n)] = \sum_{i=1}^n E_{\theta}[T(X_i)] = n \cdot E_{\theta}[T(X_1)].$$

Nach Satz A.7 folgt, dass der Maximum-Likelihood-Schätzer $\tilde{\theta}$ im Produktmodell durch die Gleichung

$$E_{\tilde{\theta}}[T(X_1)] = \frac{1}{n} \sum_{i=1}^n T(X_i)$$

charakterisiert ist. Somit ergibt sich im Produktmodell derselbe Maximum-Likelihood-Schätzer wie im Einkomponentenmodell mit gemitteltem Beobachtungswert $\frac{1}{n} \sum_{i=1}^n T(X_i)$.

2.4. Likelihood-Quotienten-Tests

Wir führen nun einen allgemeinen Rahmen für Hypothesentests ein. Wie zuvor betrachten wir ein reguläres statistisches Modell mit Wahrscheinlichkeitsräumen $(\Omega, \mathcal{A}, P_{\theta})$, $\theta \in \Theta$, Beobachtungsabbildung $X : \Omega \rightarrow S$, und Massen- bzw. Dichtefunktionen $f_{\theta}(x)$. Die Nullhypothese H_0 und die Alternative H_1 sind durch disjunkte Teilmengen Θ_0 und Θ_1 des Parameterbereichs Θ bestimmt:

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1.$$

Definition 2.9 (Hypothesentest, Gütefunktion, Signifikanzniveau und Macht).

- (i) Ein *Hypothesentest* für das obige Testproblem ist gegeben durch eine messbare Funktion

$$\begin{aligned} \varphi : S &\rightarrow \{0, 1\} && \text{(nicht-randomisierter Test),} && \text{bzw.} \\ \varphi : S &\rightarrow [0, 1] && \text{(randomisierter Test)} \end{aligned}$$

mit der Entscheidungsregel

- verwerfe H_0 falls $\varphi(x) = 1$,
- verwerfe H_0 nicht falls $\varphi(x) = 0$,
- verwerfe H_0 mit Wahrscheinlichkeit $\varphi(x)$ falls $\varphi(x) \in (0, 1)$.

Der *Verwerfungsbereich* des Tests ist die Menge

$$R = \{x \in S : \varphi(x) = 1\}.$$

- (ii) Die Funktion

$$\beta(\theta) = E_{\theta}[\varphi(X)] \quad (\theta \in \Theta)$$

heißt *Gütefunktion*. Sie beschreibt die Verwerfungswahrscheinlichkeit in Abhängigkeit von θ .

- (iii) Der Test hat *Signifikanzniveau* α , falls die Niveaubedingung

$$\beta(\theta) \leq \alpha \quad \text{für alle } \theta \in \Theta_0$$

erfüllt ist. Die auf Θ_1 eingeschränkte Gütefunktion

$$\beta(\theta), \quad \theta \in \Theta_1,$$

heißt *Macht* des Hypothesentests.

Man kann sich nun fragen, ob es zu einem vorgegebenen Signifikanzniveau α einen mächtigsten Test gibt. Dies ist im Allgemeinen nicht der Fall, aber in einigen einfachen Fällen können wir die Frage positiv beantworten.

Einfache Hypothese und Alternative

Seien $\theta_0, \theta_1 \in \Theta$ mit $\theta_0 \neq \theta_1$. Angenommen, wir wissen, dass die Stichproben von einer der beiden Verteilungen mit Dichten bzw. Massenfunktionen f_{θ_0} und f_{θ_1} stammen, und wir wollen entscheiden zwischen der

$$\text{Nullhypothese } H_0: \quad \gg \theta = \theta_0 \ll$$

und der

$$\text{Alternative } H_1: \quad \gg \theta = \theta_1 \ll.$$

Ein solches Problem tritt in Anwendungen zwar selten auf, bildet aber einen ersten Schritt zum Verständnis allgemeinerer Testprobleme. Sei

$$\lambda(x) = \frac{L(\theta_1; x)}{L(\theta_0; x)} = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$$

der Quotient der Likelihood-Funktionen, also die relative Dichte der beiden Wahrscheinlichkeitsverteilungen.

Definition 2.10 (Likelihood-Quotienten-Test). Seien $c \in (0, \infty)$ und $p \in [0, 1]$. Ein Test mit Entscheidungsregel

- verwerfe H_0 , falls $\lambda(x) > c$;
- verwerfe H_0 nicht, falls $\lambda(x) < c$;
- verwerfe H_0 mit Wahrscheinlichkeit p , falls $\lambda(x) = c$;

heißt *Likelihood-Quotienten-Test* mit *Schwellenwert* c und *Randomisierungswahrscheinlichkeit* p .

Der Verwerfungsbereich eines Likelihood-Quotienten-Tests ist also $R = \{\lambda > c\}$, und die Entscheidungsfunktion ist

$$\varphi = 1_{\lambda > c} + p \cdot 1_{\lambda = c}.$$

Die Verwerfungswahrscheinlichkeit beträgt

$$\beta(\theta) = E_{\theta}[\varphi(X)] = \int \varphi(x) f_{\theta}(x) dx \quad \text{bzw.} \quad \sum_{x \in S} \varphi(x) f_{\theta}(x)$$

im absolutstetigen bzw. diskreten Fall. Insbesondere hat der Test das Niveau $\alpha = \beta(\theta_0)$. Der Zweck der Randomisierung ist, dass man so zu jedem Niveau α einen Likelihood-Quotienten-Test finden kann, der dieses Niveau genau erreicht. Dies ist für die Praxis nicht relevant, aber es ermöglicht uns, zu jedem vorgegebenen Niveau einen mächtigsten Test zu konstruieren.

Satz 2.11 (Neyman-Pearson-Lemma). Der Likelihood-Quotienten-Test ist der *mächtigste Test zu seinem Niveau* α , d.h. jeder Test mit

$$\beta(\theta_0) = \text{Wahrscheinlichkeit (Fehler 1. Art)} \leq \alpha$$

hat eine kleinere Macht (d.h. eine höhere Wahrscheinlichkeit für den Fehler 2. Art).

2. Likelihood

Beweis. Wir zeigen die Aussage im absolutstetigen Fall; der Beweis im diskreten Fall verläuft analog. Sei $\psi : S \rightarrow [0, 1]$ die Entscheidungsfunktion eines Tests mit Niveau α . Dann gilt

$$\int \psi f_{\theta_0} dx \leq \alpha = \int \varphi f_{\theta_0} dx,$$

und damit

$$\int (\varphi - \psi) f_{\theta_0} dx \geq 0. \quad (2.12)$$

Zu zeigen ist, dass ψ eine kleinere Macht als φ hat, d.h.

$$\int (\varphi - \psi) f_{\theta_1} dx \geq 0. \quad (2.13)$$

Da φ ein Likelihood-Quotienten-Test ist, gilt

$$\varphi - \psi = \begin{cases} 1 - \psi \geq 0 & \text{auf } \{\lambda > c\}, \\ 0 - \psi \leq 0 & \text{auf } \{\lambda < c\}. \end{cases}$$

Also ist $(\varphi - \psi) \cdot (\lambda - c) \geq 0$, und wegen $\lambda = f_{\theta_1} / f_{\theta_0}$ und (2.12) erhalten wir

$$\begin{aligned} 0 &\leq \int (\varphi - \psi) \cdot (\lambda - c) f_{\theta_0} dx \\ &= \int (\varphi - \psi) f_{\theta_1} dx - c \int (\varphi - \psi) f_{\theta_0} dx \\ &\leq \int (\varphi - \psi) f_{\theta_1} dx. \end{aligned}$$

Damit ist die Behauptung (2.13) gezeigt. ■

Bemerkung (Existenz eines Likelihood-Quotienten-Tests mit vorgegebenem Niveau). Die Randomisierung stellt sicher, dass zu jedem Niveau $\alpha \in (0, 1)$ ein Likelihood-Quotienten-Test existiert, der das Niveau genau erreicht (und somit ein mächtigster Test zu dem vorgegebenen Niveau ist). Für den Likelihood-Quotienten-Test mit Schwellenwert c und Randomisierungswahrscheinlichkeit p beträgt die Wahrscheinlichkeit für den Fehler 1. Art nämlich

$$E_{\theta_0}[\varphi(X)] = P_{\theta_0}[\lambda(X) > c] + p \cdot P_{\theta_0}[\lambda(X) = c].$$

Ist nun c ein α -Quantil der Verteilung von $\lambda(X)$ unter der Nullhypothese, dann ist dieser Wert für $p = 0$ kleiner oder gleich α , und für $p = 1$ größer oder gleich α . Also gibt es einen Zwischenwert $p \in [0, 1]$ für den die Wahrscheinlichkeit für den Fehler 1. Art exakt gleich α ist.

Beispiel (Signal oder Rauschen ?). Wir wollen mithilfe von n unabhängigen reellen Beobachtungswerten entscheiden, ob ein Signal vorliegt, oder nur zufälliges Rauschen. Als Nullhypothese („nur Rauschen“) nehmen wir an, dass die Beobachtungswerte Stichproben von unabhängigen Zufallsvariablen X_1, \dots, X_n mit Verteilung $\mathcal{N}(0, \nu)$ sind, die Alternative („Signal eingetroffen“) modellieren wir durch unabhängige Zufallsvariablen X_1, \dots, X_n mit Verteilung $\mathcal{N}(m, \nu)$. Wir nehmen an, dass sowohl die Signalstärke m als auch die Varianz ν der zufälligen Fluktuationen bekannt sind. Als Likelihood-Quotient ergibt sich in diesem Fall

$$\lambda(x) = \prod_{i=1}^n \exp\left(\frac{x_i^2}{2\nu} - \frac{(x_i - m)^2}{2\nu}\right) = \exp\left(\frac{m}{\nu} \sum_{i=1}^n x_i - \frac{nm^2}{2\nu}\right).$$

Insbesondere ist der Likelihood-Quotient für $m > 0$ eine streng monoton wachsende Funktion des Stichprobenmittelwerts \bar{x}_n . Der Likelihood-Quotienten-Test (ohne Randomisierung) verwirft daher die Nullhypothese, falls $\bar{x}_n > c$ für einen vorgegebenen Schwellenwert c gilt. Randomisierung ist hier nicht

nötig, da die Verteilung absolutstetig ist, und das Ereignis $\bar{X}_n = c$ daher die Wahrscheinlichkeit Null hat. Unter der Nullhypothese hat \bar{X}_n die Verteilung $\mathcal{N}(0, v/n)$, also ist $\bar{X}_n/\sqrt{v/n}$ standardnormalverteilt. Damit erhalten wir als Wahrscheinlichkeit für den Fehler 1. Art

$$P_0[\bar{X}_n > c] = P_0[\bar{X}_n/\sqrt{v/n} > c/\sqrt{v/n}] = 1 - \Phi(c/\sqrt{v/n}).$$

Insbesondere hat der Test genau das Niveau α , falls wir $c = q_{1-\alpha}\sqrt{v/n}$ wählen, wobei $q_{1-\alpha} = \Phi^{-1}(1-\alpha)$ das $(1-\alpha)$ -Quantil der Standardnormalverteilung ist. Um die Macht zu berechnen, bemerken wir, dass unter der Alternative die Zufallsvariable $(\bar{X}_n - m)/\sqrt{v/n}$ standardnormalverteilt ist. Daher erhalten wir

$$\begin{aligned} \beta(m) &= P_1[\bar{X}_n > c] = P_1[(\bar{X}_n - m)/\sqrt{v/n} > (c - m)/\sqrt{v/n}] = 1 - \Phi\left((c - m)/\sqrt{v/n}\right) \\ &= 1 - \Phi\left(q_{1-\alpha} - m\sqrt{n/v}\right) = \Phi\left(q_\alpha + m\sqrt{n/v}\right). \end{aligned}$$

Die Gütefunktion ist also eine reskalierte Verteilungsfunktion der Standardnormalverteilung. Um eine gewisse Macht zu erreichen, muss das "signal-to-noise-ratio" m/\sqrt{v} von der Größenordnung $\Omega(1/\sqrt{n})$ sein, die Anzahl der n der nötigen Stichproben ist also von der Größenordnung $O((m/\sqrt{v})^{-2})$.

Monotone Likelihood-Quotienten

Wir betrachten nun ein Testproblem

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \quad (2.14)$$

mit einer zusammengesetzten Hypothese und Alternative. Das Neyman-Pearson-Lemma liefert für feste Werte $\theta_0 \in \Theta_0$ und $\theta_1 \in \Theta_1$ einen mächtigsten Test. Im Allgemeinen hängt dieser aber von θ_0 und θ_1 ab. Es stellt sich daher die Frage, ob es unter bestimmten Voraussetzungen einen gleichmäßig mächtigsten Test gibt.

Definition 2.12 (Gleichmäßig mächtigster Test). Ein Test mit Entscheidungsfunktion $\varphi : S \rightarrow [0, 1]$ heißt *gleichmäßig mächtigster Test* für das Testproblem (2.18) zum Niveau $\alpha \in [0, 1]$, falls der Test das Niveau α hat, und für jeden Test $\psi : S \rightarrow [0, 1]$ mit Niveau α gilt:

$$\beta_\psi(\theta) := E_\theta[\psi(X)] \leq E_\theta[\varphi(X)] =: \beta_\varphi(\theta) \quad \text{für alle } \theta \in \Theta_1.$$

Im Allgemeinen kann man die Existenz eines gleichmäßig mächtigsten Tests nicht erwarten. Eine wichtige Ausnahme bilden Modelle mit Likelihood-Quotienten, die monoton von einer Teststatistik abhängen. Beispielsweise waren die Likelihood-Quotienten im Beispiel oben (Signal oder Rauschen) monotone Funktionen des Stichprobenmittelwerts. Allgemeiner gilt in einer exponentiellen Familie mit Parameterbereich $\Theta \subseteq \mathbb{R}$ und Likelihood-Funktion $L(\theta; x) = \mathcal{Z}(\theta)^{-1} e^{c(\theta)T(x)} h(x)$ für $\theta < \tilde{\theta}$:

$$\frac{L(\tilde{\theta}; x)}{L(\theta; x)} = \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\tilde{\theta})} e^{c(\tilde{\theta}) - c(\theta)T(x)}.$$

Ist die Funktion c streng monoton wachsend, dann ist der Likelihood-Quotient also eine streng monoton wachsende Funktion von $T(x)$. Somit hat ein Likelihood-Quotienten-Test für $\{\theta\}$ vs. $\{\tilde{\theta}\}$ unabhängig von θ und $\tilde{\theta}$ die einfache Form

$$\varphi(x) = \begin{cases} 1 & \text{für } T(x) > c, \\ 0 & \text{für } T(x) < c, \\ p & \text{für } T(x) = c. \end{cases} \quad (2.15)$$

Da die Form der Likelihood-Quotienten-Tests nicht von den Parametern abhängt, gibt es in diesem Fall einen gleichmäßig mächtigsten Test.

Satz 2.13 (Gleichmäßig mächtigster Test bei monotonen Likelihood-Quotienten). Ist $\Theta \subseteq \mathbb{R}$, und sind die Likelihood-Quotienten in obigem Sinne streng monoton wachsend in einer Statistik $T(x)$, dann existiert für jedes $\theta_0 \in \Theta$ und für jedes Niveau $\alpha \in (0, 1)$ ein gleichmäßig mächtigster Test von der Form (2.15) für das *einseitige* Testproblem

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0. \quad (2.16)$$

Beweis. Sei zunächst $\theta > \theta_0$ fest. Dann gibt es einen mächtigsten Test zum Niveau α für $\{\theta_0\}$ vs. $\{\theta\}$ von der Form (2.15). Dabei sind der Schwellenwert $c \in \mathbb{R}$ und die Randomisierungswahrscheinlichkeit $p \in [0, 1]$ eindeutig bestimmt durch die Bedingung

$$\alpha = E_{\theta_0}[\varphi(X)] = P_{\theta_0}[T(X) > c] + p \cdot P_{\theta_0}[T(X) = c].$$

Insbesondere sind c und p unabhängig von θ . Also ist φ gleichmäßig mächtigster Test zum Niveau α für das Testproblem mit einfacher Hypothese $\theta = \theta_0$ und zusammengesetzter Alternative $\theta > \theta_0$.

Für den Beweis der Behauptung bleibt nur noch zu zeigen, dass φ auch ein Test zum Niveau α für das Testproblem mit zusammengesetzter Hypothese $\theta \leq \theta_0$ und Alternative $\theta > \theta_0$ ist. Sei also $\theta \leq \theta_0$, und sei $\tilde{\alpha} := E_{\theta}[\varphi(X)]$. Dann ist φ auch ein Likelihood-Quotienten-Test für das Testproblem mit einfacher Hypothese $\{\theta\}$ und einfacher Alternative $\{\theta_0\}$. Also ist φ nach dem Neyman-Pearson-Lemma ein mächtigster Test zum Niveau $\tilde{\alpha}$ für dieses Testproblem. Durch Vergleich mit dem Test mit konstanter Entscheidungsfunktion $\psi \equiv \tilde{\alpha}$ folgt daher

$$\alpha = E_{\theta_0}[\varphi(X)] \geq E_{\theta_0}[\psi(X)] = \tilde{\alpha} = E_{\theta}[\varphi(X)].$$

Also erfüllt φ die Niveaubedingung auch für die zusammengesetzte Hypothese $\theta \leq \theta_0$, und ist somit auch gleichmäßig mächtigster Test für das Testproblem mit linksseitiger Hypothese und rechtsseitiger Alternative. ■

Beispiel (Tests in einparametrischen Gauß-Modellen). Sei $X = (X_1, \dots, X_n)$ mit unabhängigen Zufallsvariablen $X_1, \dots, X_n \sim \mathcal{N}(m, v)$. Die Likelihood-Funktion ist dann gegeben durch

$$L(m, v; x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x_i - m)^2}{2v}}, \quad (2.17)$$

und für die Log-Likelihood erhalten wir die Darstellungen

$$\ell(m, v; x_1, \dots, x_n) = -\frac{n}{2} \log(2\pi v) - \frac{1}{2v} \sum_{i=1}^n (x_i - m)^2 = -\frac{n}{2} \left(\log(2\pi v) + \frac{v_n}{v} + \frac{1}{v} (\bar{x}_n - m)^2 \right)$$

mit dem Stichprobenmittelwert \bar{x}_n und der (nicht renormierten) Stichprobenvarianz v_n . Meistens sind sowohl der Mittelwert m als auch die Varianz v unbekannt. Mit diesem Fall beschäftigen wir uns im nächsten Abschnitt. An dieser Stelle betrachten wir zunächst die einfacheren einparametrischen Fälle, in denen entweder die Varianz oder der Mittelwert bekannt ist.

- (i) *Testproblem* $H_0 : m \leq m_0$ vs. $H_1 : m > m_0$ *bei bekannter Varianz* $v > 0$. Wenn v fest ist, dann können wir die Log-Likelihood schreiben als

$$\ell(m, v; x_1, \dots, x_n) = \text{const.}(m, v) + \frac{nm}{v} \bar{x}_n - \frac{1}{2v} \sum_{i=1}^n x_i^2.$$

Der Likelihood-Quotient $L(\tilde{m}, v; x)/L(m, v; x)$ ist daher für $m < \tilde{m}$ eine streng monoton wachsende Funktion von \bar{x}_n . Also ist der Test mit Verwerfungsbereich $\{\bar{X}_n > c\}$ für jeden Schwellenwert $c \in \mathbb{R}$ gleichmäßig mächtigster Test zu seinem Niveau α . Um ein festes Niveau α genau zu erreichen wählt man $c = m_0 + q_{1-\alpha} \sqrt{v/n}$ mit dem $(1-\alpha)$ -Quantil $q_{1-\alpha}$ der Standardnormalverteilung.

Auf analoge Weise wie im Beispiel von oben mit einfacher Hypothese und Alternative erhalten wir als Gütefunktion

$$\beta(m) = \Phi\left(q_\alpha + \frac{m - m_0}{\sqrt{v}}\sqrt{n}\right).$$

- (ii) *Testproblem* $H_0 : v \geq v_0$ vs. $H_1 : v < v_0$ bei bekanntem Mittelwert m . Der Likelihood-Quotient $L(m, \tilde{v}; x)/L(m, v; x)$ ist für $\tilde{v} < v$ eine *streng monoton fallende* Funktion von $\sum_{i=1}^n (x_i - m)^2$. Daher ist der Test mit Verwerfungsbereich $\sum_{i=1}^n (X_i - m)^2 < c$ für jeden Schwellenwert $c > 0$ ein gleichmäßig mächtigster Test zu seinem Niveau. Die Wahl des Schwellenwerts zu vorgegebenem Niveau ergibt sich aus

$$P_{m,v} \left[\sum_{i=1}^n (X_i - m)^2 < c \right] = P_{m,v} \left[\sum_{i=1}^n \left(\frac{X_i - m}{\sqrt{v}} \right)^2 < \frac{c}{v} \right] = F_{\chi^2(n)} \left(\frac{c}{v} \right) \leq F_{\chi^2(n)} \left(\frac{c}{v_0} \right)$$

für $v \geq v_0$, wobei $F_{\chi^2(n)}$ die Verteilungsfunktion der Chiquadrat-Verteilung mit n Freiheitsgraden ist. Hierbei haben wir benutzt, dass $Z_1^2 + \dots + Z_n^2$ Chiquadrat-verteilt ist, falls Z_1, \dots, Z_n unabhängige standardnormalverteilte Zufallsvariablen sind, siehe unten. Um das Niveau α genau zu erreichen, muss man also $c = v_0 \cdot q_{\chi^2(n), \alpha}$ wählen.

Im letzten Beispiel sehen wir auch, dass der Verwerfungsbereich des gleichmäßig mächtigsten Tests von m abhängt. Daher existiert kein gleichmäßig mächtigster Test für das Testproblem (ii) falls m und v beide unbekannt sind. Welchen Test sollten wir dann verwenden?

Der oben verwendete Test φ_m mit Verwerfungsbereich $\sum_{i=1}^n (X_i - m)^2 < c$ ist für jeden festen Wert von m unbrauchbar, wenn der Mittelwert unbekannt ist, denn für jeden Schwellenwert c und jedes $v > 0$ gilt

$$\lim_{\tilde{m} \rightarrow \infty} \beta_{\varphi_m}(\tilde{m}, v) = \lim_{\tilde{m} \rightarrow \infty} P_{\tilde{m}, v} \left[\sum_{i=1}^n (X_i - \tilde{m})^2 < c \right] = 0.$$

Selbst wenn $v < v_0$ gilt, wird die Verwerfungswahrscheinlichkeit also beliebig klein, wenn \tilde{m} groß wird. Insbesondere ist die Niveaubedingung für große \tilde{m} auch auf der Alternative erfüllt. Einen Test mit dieser Eigenschaft nennt man *verfälscht*.

Als Ausweg liegt es nahe, den Mittelwert durch \bar{x}_n zu schätzen, und die Nullhypothese zu verwerfen, falls der Wert der Teststatistik $\sum_{i=1}^n (x_i - \bar{x}_n)^2$ unterhalb eines Schwellenwerts c liegt. Tatsächlich kann man zeigen, dass ein solcher Test ein *gleichmäßig mächtigster unverfälschter Test* zu seinem Niveau ist, siehe zum Beispiel Abschnitt 10.4 in [Georgii].

Allgemeine Likelihood-Quotienten-Tests

Wir betrachten wieder ein allgemeines Testproblem

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \quad (2.18)$$

mit einer zusammengesetzten Hypothese und Alternative. Da der Likelihood-Quotient von den Parameterwerten abhängt, können wir ihn im Allgemeinen nicht zur Konstruktion eines Tests benutzen. Stattdessen betrachtet man als ad-hoc-Verfahren zur Konzeption von Hypothesentests den *verallgemeinerten Likelihood-Quotienten*

$$\lambda(x) := \frac{\sup \{L(\theta; x) : \theta \in \Theta_1\}}{\sup \{L(\theta; x) : \theta \in \Theta_0\}} = \frac{\text{max. likelihood von } x \text{ falls } H_1 \text{ wahr}}{\text{max. likelihood von } x \text{ falls } H_0 \text{ wahr}}.$$

Definition 2.14 (Verallgemeinerter Likelihood-Quotienten-Test). Seien $c \in (0, \infty)$ und $p \in [0, 1]$. Ein Test mit Entscheidungsregel

- verwerfe H_0 , falls $\lambda(x) > c$;

2. Likelihood

- verwerfe H_0 nicht, falls $\lambda(x) < c$;
- verwerfe H_0 mit Wahrscheinlichkeit p , falls $\lambda(x) = c$;

heißt verallgemeinerter Likelihood-Quotienten-Test mit Schwellenwert c und Randomisierungswahrscheinlichkeit p .

Im oben diskutierten Beispiel (ii) erhält man falls m und v beide unbekannt sind:

$$\lambda(x) = \frac{\sup \{L(m, v; x) : v < v_0\}}{\sup \{L(m, v; x) : v \geq v_0\}} = \begin{cases} \exp\left(-\frac{n}{2} \left(\frac{v_n}{v_0} - 1 - \log \frac{v_n}{v_0}\right)\right) & \text{für } v_n \geq v_0, \\ \exp\left(+\frac{n}{2} \left(\frac{v_n}{v_0} - 1 - \log \frac{v_n}{v_0}\right)\right) & \text{für } v_n \leq v_0. \end{cases}$$

Da der verallgemeinerte Likelihood-Quotient eine streng monoton fallende Funktion von v_n ist, verwirft ein verallgemeinerter Likelihood-Quotienten-Test die Nullhypothese genau dann, wenn die Stichprobenvarianz $V_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ unterhalb eines Schwellenwerts c liegt. Ein verallgemeinerter Likelihood-Quotienten-Test stimmt damit in diesem Beispiel mit dem oben erwähnten gleichmäßig mächtigsten unverfälschten Test überein.

2.5. Studentsche Konfidenzintervalle und t-Test

Angenommen, wir beobachten reellwertige Messwerte (Stichproben, Daten), die von einer unbekanntem Wahrscheinlichkeitsverteilung μ auf \mathbb{R} stammen. Ziel der Statistik ist es, Rückschlüsse auf die zugrundeliegende Verteilung aus den Daten zu erhalten. Im einfachsten Modell (Gauß-Modell) nimmt man an, dass die Daten unabhängige Stichproben von einer Normalverteilung mit unbekanntem Mittelwert und/oder Varianz sind:

$$\mu = N(m, v), \quad m, v \text{ unbekannt.}$$

Eine partielle Rechtfertigung für die Normalverteilungsannahme liefert der zentrale Grenzwertsatz. Letztendlich muss man aber in jedem Fall überprüfen, ob eine solche Annahme gerechtfertigt ist. Ein erstes Ziel ist es nun, den Wert von m auf der Basis von n unabhängigen Stichproben $X_1(\omega) = x_1, \dots, X_n(\omega) = x_n$ zu schätzen, und zu quantifizieren.

Problemstellung: Schätzung des Erwartungswerts

- Schätze m auf der Basis von n unabhängigen Stichproben $X_1(\omega), \dots, X_n(\omega)$ von μ .
- Herleitung von Konfidenzintervallen.

Im mathematischen Modell interpretieren wir die Beobachtungswerte als Realisierungen von unabhängigen Zufallsvariablen X_1, \dots, X_n . Da wir die tatsächliche Verteilung nicht kennen, untersuchen wir alle in Betracht gezogenen Verteilungen simultan:

$$X_1, \dots, X_n \sim N(m, v) \quad \text{unabhängig unter } P_{m,v}. \quad (2.19)$$

Ein naheliegender Schätzer für m ist der *empirische Mittelwert*

$$\bar{X}_n(\omega) := \frac{X_1(\omega) + \dots + X_n(\omega)}{n}.$$

Wir haben oben bereits gezeigt, dass dieser Schätzer *erwartungstreu* (*unbiased*) und *konsistent* ist, d.h. für alle m, v gilt:

$$E_{m,v}[\bar{X}_n] = m$$

und

$$\bar{X}_n \rightarrow m \quad P_{m,v}\text{-stochastisch für } n \rightarrow \infty.$$

Wie wir den Schätzfehler quantifizieren hängt davon ab, ob wir die Varianz kennen.

Schätzung von m bei bekannter Varianz v .

Um den Schätzfehler zu kontrollieren, berechnen wir die Verteilung von \bar{X}_n :

$$\begin{aligned} X_i \sim N(m, v) \text{ unabhängig} &\Rightarrow X_1 + \dots + X_n \sim N(nm, nv) \\ &\Rightarrow \bar{X}_n \sim N(m, v/n) \\ &\Rightarrow \frac{\bar{X}_n - m}{\sqrt{v/n}} \sim N(0, 1) \end{aligned}$$

Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung, dann erhalten wir

$$P_{m,v} \left[|\bar{X}_n - m| < q \sqrt{\frac{v}{n}} \right] = N(0, 1)(-q, q) = 2 \left(\Phi(q) - \frac{1}{2} \right) \quad \text{für alle } m \in \mathbb{R}.$$

Satz 2.15 (Konfidenzintervalle bei bekannter Varianz). Im Gaußmodell (2.19) mit bekannter Varianz v ist das zufällige Intervall

$$\left(\bar{X}_n - \Phi^{-1}(\alpha) \sqrt{\frac{v}{n}}, \bar{X}_n + \Phi^{-1}(\alpha) \sqrt{\frac{v}{n}} \right)$$

ein *Konfidenzintervall* zum *Konfidenzniveau* $2\alpha - 1$ für m , d.h.

$$P_{m,v}[m \in \text{Intervall}] \geq 2\alpha - 1 \quad \text{für alle } m \in \mathbb{R}.$$

Man beachte, dass die Länge des Konfidenzintervalls in diesem Fall nicht von den beobachteten Stichproben abhängt!

Schätzung von m bei unbekannter Varianz v . In Anwendungen ist meistens die Varianz unbekannt. In diesem Fall können wir das Intervall oben nicht verwenden, da es von der unbekanntem Varianz v abhängt. Stattdessen schätzen wir m und v simultan, und konstruieren ein Konfidenzintervall für m mithilfe beider Schätzwerte. Erwartungstreue Schätzer für m und v sind

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{und} \quad V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Um ein Konfidenzintervall für m zu erhalten, bestimmen wir mithilfe des Transformationssatzes die gemeinsame Verteilung von \bar{X}_n und V_n :

Lemma 2.16. \bar{X}_n und V_n sind unabhängig unter $P_{m,v}$ mit Verteilung

$$\bar{X}_n \sim N\left(m, \frac{v}{n}\right), \quad \frac{n-1}{v} V_n \sim \chi^2(n-1).$$

Beweis. Wir führen eine lineare Koordinatentransformation $Y = OX$ durch, wobei O eine orthogonale $n \times n$ -Matrix vom Typ

$$O = \begin{pmatrix} \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \\ \text{beliebig} & & \end{pmatrix}$$

2. Likelihood

ist. Eine solche Matrix erhalten wir durch Ergänzen des normierten Vektors $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ zu einer Orthonormalbasis des \mathbb{R}^n . Da die Matrix O orthogonal ist, gilt in den neuen Koordinaten

$$\begin{aligned}\bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{\sqrt{n}} Y_1, \quad \text{und} \\ (n-1)V_n &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 = \|X\|_{\mathbb{R}^n}^2 - n\bar{X}_n^2 = \|Y\|_{\mathbb{R}^n}^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.\end{aligned}$$

Da die Zufallsvariablen X_i ($1 \leq i \leq n$) unabhängig und $N(m, v)$ -verteilt sind, ist der Zufallsvektor $X = (X_1, \dots, X_n)$ multivariat normalverteilt mit Mittel (m, \dots, m) und Kovarianzmatrix $v \cdot I_n$. Nach dem Transformationssatz oder mithilfe von charakteristischen Funktionen folgt

$$Y \sim N\left(O \begin{pmatrix} m \\ \vdots \\ m \end{pmatrix}, v \cdot O I_n O^T\right) = N\left(\begin{pmatrix} m\sqrt{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v \cdot I_n\right).$$

Also sind Y_1, \dots, Y_n unabhängige Zufallsvariablen mit Verteilungen

$$Y_1 \sim N(m\sqrt{n}, v) \quad , \quad Y_i \sim N(0, v) \quad \text{für } i \geq 2.$$

Es folgt, dass

$$\bar{X}_n = \frac{Y_1}{\sqrt{n}} \quad \text{und} \quad \frac{n-1}{v} V_n = \sum_{i=2}^n \left(\frac{Y_i}{\sqrt{v}}\right)^2$$

unabhängige Zufallsvariablen mit Verteilungen $N(m, v/n)$ bzw. $\chi^2(n-1)$ sind. ■

Bei bekannter Varianz v hatten wir Konfidenzintervalle für m vom Typ $\bar{X}_n \pm q \cdot \sqrt{v/n}$ erhalten, wobei q ein geeignetes Quantil der Standardnormalverteilung ist. Daher liegt es nahe, zu versuchen, bei unbekannter Varianz Konfidenzintervalle vom Typ $\bar{X}_n \pm q \cdot \sqrt{V_n/n}$ herzuleiten. Es gilt

$$P_{m,v} \left[|\bar{X}_n - m| \geq q \sqrt{V_n/n} \right] = P_{m,v} [|T_{n-1}(X)| \geq q] \quad \text{mit}$$

$$T_{n-1}(X) := \frac{\sqrt{n} \cdot (\bar{X}_n - m)}{\sqrt{V_n}}.$$

Die Statistik $T_{n-1}(X)$ heißt **Studentsche t -Statistik mit $n-1$ Freiheitsgraden**. Unsere Überlegungen zeigen, dass wir aus Quantilen der Studentschen t -Statistik Konfidenzintervalle für das Gauß-Modell herleiten können, falls die t -Statistik ein Pivot ist, das heißt ihre Verteilung nicht von den unbekanntem Parametern m und v abhängt. Dies ist in der Tat der Fall.

Satz 2.17 (Student). Die Verteilung von $T_{n-1}(X)$ ist absolutstetig mit Dichte

$$f_{t(n-1)}(t) = B\left(\frac{1}{2}, \frac{n-1}{2}\right)^{-1} \cdot (n-1)^{-1/2} \cdot \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \quad (t \in \mathbb{R}). \quad (2.20)$$

Hierbei ist der Normierungsfaktor $B(a, b) = \int_0^1 s^{a-1} (1-s)^{b-1} ds$ die *Eulersche Beta-Funktion*.

Beweis. Direkt oder mithilfe des Transformationssatzes zeigt man: Sind Z und Y unabhängige Zufallsvariablen mit Verteilungen $N(0, 1)$ bzw. $\chi^2(n)$, dann ist $Z/\sqrt{\frac{1}{n}Y}$ absolutstetig mit Dichte f_{T_n} . Der Satz folgt dann nach Lemma 2.16 mit $Z := \frac{\bar{X}_n - m}{\sqrt{v/n}}$ und $Y := \frac{n-1}{v}V_n$. ■

Definition 2.18 (Studentische t -Verteilung). Die Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichtefunktion (2.20) nennt man »Studentische t -Verteilung mit $n - 1$ Freiheitsgraden«.

Anwendung auf Konfidenzintervalle und Tests

Aus Satz 2.17 ergibt sich unmittelbar die folgende Aussage.

Korollar 2.19 (Studentische Konfidenzintervalle und t -Test). Sei $\alpha \in (0, 1)$, und sei

$$q = F_{t(n-1)}^{-1}(1 - \alpha/2)$$

das $(1 - \alpha/2)$ -Quantil der t -Verteilung mit $n - 1$ Freiheitsgraden. Dann gilt:

- (i) Das zufällige Intervall

$$\left(\bar{X}_n - q \cdot \sqrt{V_n/n}, \bar{X}_n + q \cdot \sqrt{V_n/n} \right)$$

ist ein Konfidenzintervall für m zum Konfidenzniveau $1 - \alpha$.

- (ii) Für $m_0 \in \mathbb{R}$ ist

$$|\bar{X}_n - m_0| \geq q \cdot \sqrt{V_n/n}$$

der Verwerfungsbereich eines Hypothesentest für das beidseitige Testproblem

$$H_0 : m = m_0 \quad \text{vs.} \quad H_1 : m \neq m_0$$

zum Signifikanzniveau α .

Entsprechend erhält man auch einseitige Konfidenzintervalle sowie Verwerfungsbereiche für Hypothesentests mit einseitiger Alternative mithilfe der Studentischen t -Statistik. Im Korollar zeigt sich auch ein Zusammenhang von Konfidenzintervallen und Hypothesentests, der auch allgemeiner gilt.

Übung (Dualität von Konfidenzbereichen und Hypothesentests). Sei $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, und $X : \Omega \rightarrow S$ die Beobachtung.

- a) Zeigen Sie: Ist $x \mapsto C(x) \subseteq \Theta$ ein Konfidenzbereich für ϑ zum Niveau $1 - \alpha$ und $\vartheta_0 \in \Theta$, dann ist

$$R(\vartheta_0) = \{x \in S : \vartheta_0 \notin C(x)\}$$

der Verwerfungsbereich eines Tests zum Niveau α der Hypothese $\Theta_0 = \{\vartheta_0\}$.

- b) Umgekehrt sei $R(\vartheta_0)$ für jedes $\vartheta_0 \in \Theta$ der Verwerfungsbereich eines Tests der Hypothese $\Theta_0 = \{\vartheta_0\}$ zum Niveau α . Konstruieren Sie einen Konfidenzbereich für ϑ zum Niveau $1 - \alpha$.
- c) Illustrieren Sie die Aussagen anhand eines Beispiels.

Hier sieht man auch, dass die Angabe eines Konfidenzbereichs eine allgemeinere Aussage liefert als die 0-1-Entscheidung eines Hypothesentests, welche nur eine feste Hypothese betrifft.

Der t -Test als Likelihood-Quotienten-Test

Wir zeigen nun, dass der t -Test im Gauß-Modell der Likelihood-Quotienten-Test für das Testproblem

$$H_0 : m = m_0 \quad \text{vs.} \quad H_1 : m \neq m_0$$

ist. Für die Likelihood gilt nach (2.17)

$$\ell(m, v; x_1, \dots, x_n) = -\frac{n}{2} \log(2\pi v) - \frac{1}{2v} \sum_{i=1}^n (x_i - m)^2 = -\frac{n}{2} \left(\log(2\pi v) + \frac{v_n}{v} + \frac{1}{v} (\bar{x}_n - m)^2 \right).$$

Mit $\tilde{v}_n := \frac{1}{n} \sum_{i=1}^n (x_i - m_0)^2$ und $\hat{v}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ erhalten wir

$$\begin{aligned} \sup_{m=m_0} \ell(m, v; x) &= -\frac{n}{2} (\log(2\pi\tilde{v}_n) + 1), \\ \sup_{m \neq m_0} \ell(m, v; x) &= -\frac{n}{2} (\log(2\pi\hat{v}_n) + 1), \end{aligned}$$

und damit als Likelihood-Quotienten

$$\lambda(x) = e^{\frac{n}{2} (\log \tilde{v}_n - \log \hat{v}_n)} = \left(\frac{\tilde{v}_n}{\hat{v}_n} \right)^{n/2} = \left(1 + \frac{(\bar{x}_n - m_0)^2}{\hat{v}_n} \right)^{n/2} = \left(\frac{1 + T_{n-1}(x)^2}{n-1} \right)^{n/2},$$

wobei $T_{n-1}(x) = (\bar{x}_n - m_0) / \sqrt{v_n/n}$ mit $v_n = n\hat{v}_n / (n-1)$ die Studentsche t -Statistik zur Nullhypothese ist. Also verwirft der Likelihood-Quotienten-Test in der Tat genau dann, wenn $|T_{n-1}(x)|$ oberhalb eines festen Schwellenwerts $c > 0$ liegt.

Optimalität des t -Tests

Wir betrachten nun das einseitige Testproblem

$$H_0 : m \leq m_0 \quad \text{vs.} \quad H_1 : m > m_0.$$

Der t -Test zum Niveau $\alpha \in (0, 1)$ für dieses Testproblem hat die Entscheidungsregel

$$\varphi(X) = 1_{T_{n-1}(X) > q} \quad \text{wobei} \quad T_{n-1}(X) = \sqrt{n}(\bar{X}_n - m_0) / \sqrt{V_n}$$

die Studentsche t -Statistik, und $q = q_{1-\alpha; t(n-1)}$ das $(1-\alpha)$ -Quantil der t -Verteilung mit $n-1$ Freiheitsgraden ist. Für dieses Testproblem gibt es keinen gleichmäßig mächtigsten Test, aber wir werden sehen, dass der t -Test der beste *unverfälschte* Test ist.

Lemma 2.20. *Der t -Test mit Entscheidungsregel φ ist unverfälscht zum Niveau α , das heißt*

$$P_{m,v} [T_{n-1}(X) > q] \begin{cases} \leq \alpha & \text{für } m \leq m_0, \\ \geq \alpha & \text{für } m \geq m_0. \end{cases}$$

Beweis. Wir können ohne Beschränkung der Allgemeinheit $m_0 = 0$ annehmen. Da $T_{n-1}(X)$ unter der Nullhypothese die Verteilung $t(n-1)$ hat, gilt

$$P_{0,v} [T_{n-1}(X) > q] = \alpha.$$

Für $m \neq 0$ hat $X = (X_1, \dots, X_n)$ unter $P_{m,v}$ dieselbe Verteilung wie $\tilde{X} = (X_1 + m, \dots, X_n + m)$ unter $P_{0,v}$. Wegen $T_{n-1}(\tilde{X}) = T_{n-1}(X) + m\sqrt{n/V_n}$ folgt

$$P_{m,v} [T_{n-1}(X) > q] = P_{0,v} [T_{n-1}(\tilde{X}) > q] \begin{cases} \geq P_{0,v} [T_{n-1}(X) > q] = \alpha & \text{für } m \geq 0, \\ \leq P_{0,v} [T_{n-1}(X) > q] = \alpha & \text{für } m \leq 0, \end{cases}$$

und damit die Behauptung. ■

Satz 2.21 (Optimalität des t -Tests). Der t -Test mit Entscheidungsregel φ ist *bester unverfälschter Test* zum Niveau α für H_0 vs. H_1 , das heißt für jeden unverfälschten Niveau- α -Test ψ gilt

$$E_{m,v}[\psi(X)] \leq E_{m,v}[\varphi(X)] \quad \text{für alle } m > m_0 \text{ und } v > 0.$$

Beweis (Skizze). Für $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ setzen wir $z := \sqrt{n}\bar{x}_n$, und $s := \sum_{i=1}^n x_i^2$. Damit gilt $s - z^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2$, und somit

$$T_{n-1}(x) = \sqrt{n-1} \frac{z}{s - z^2}.$$

Der t -Test verwirft also, falls $z/\sqrt{s - z^2} > r$ für einen festen Schwellenwert r gilt, bzw., äquivalent, falls $z > \tilde{r}\sqrt{s}$ mit $\tilde{r} := r/\sqrt{1+r^2}$. Für die Likelihood gilt

$$L(m, v; x) = \text{const.}(m, v) \cdot e^{az - bs}$$

mit $a := m\sqrt{n}/v$ und $b := 1/(2v)$, und somit hat der Likelihood-Quotient die Form

$$\lambda(m, v, 0, v; x) := \frac{L(m, v; x)}{L(0, v; x)} = \text{const.}(m, v) \cdot e^{az}.$$

Der t -Test verwirft also genau dann, wenn

$$\lambda(m, v, 0, v; x) > h(s)$$

gilt, wobei h eine Funktion der Form $h(s) = ce^{d\sqrt{s}}$ mit positiven reellen Konstanten c und d ist. Für einen beliebigen unverfälschten Test mit Entscheidungsregel ψ folgt

$$(\varphi(x) - \psi(x)) \cdot (\lambda(m, v, 0, v; x) - h(s)) \geq 0 \quad \text{für alle } x,$$

und daher

$$\begin{aligned} \int (\varphi(x) - \psi(x)) f_{m,v}(x) dx &= \int (\varphi(x) - \psi(x)) \lambda(m, v, 0, v; x) f_{0,v}(x) dx \\ &\geq \int (\varphi(x) - \psi(x)) h(s) f_{0,v}(x) dx. \end{aligned} \quad (2.21)$$

Die Behauptung folgt, wenn wir zeigen können, dass dieser Ausdruck für jede sub-exponentiell wachsende Funktion $h \in C(\mathbb{R}_+, \mathbb{R})$ nicht negativ ist. Für $h \equiv 1$ ergibt sich dies aus der Unverfälschtheit von φ und ψ . Allgemeiner folgt für $h(s) = e^{-ks}$ mit $k \in \mathbb{N}$

$$\int (\varphi(x) - \psi(x)) h(s) f_{0,v}(x) dx = \int (\varphi(x) - \psi(x)) h(s) f_{0,\tilde{v}}(x) dx = 0$$

mit einer Konstanten $\tilde{v} > 0$. Daher verschwindet die rechte Seite von (2.21), falls $h(s)$ ein Polynom von e^{-s} ist. Die Behauptung folgt nun durch ein Approximationsargument. ■

3. Relative Entropie, Information und statistische Unterscheidbarkeit

Asymptotische Aussagen über Wahrscheinlichkeiten, die wir in diesem Abschnitt herleiten werden, betreffen meistens nur die exponentielle Abfallrate. Subexponentiell fallend oder wachsende Faktoren werden ignoriert. Wir führen einen entsprechenden Äquivalenzbegriff für Folgen auf der exponentiellen Skala ein:

Definition 3.1 (Asymptotische exponentielle Äquivalenz von Folgen). Zwei Folgen $(a_n)_{n \in \mathbb{N}}$ und $(b_n)_{n \in \mathbb{N}}$ von positiven reellen Zahlen heißen *asymptotisch exponentiell äquivalent* ($a_n \simeq b_n$), falls

$$\frac{1}{n} \log \frac{a_n}{b_n} = \frac{1}{n} (\log a_n - \log b_n) \longrightarrow 0 \quad \text{für } n \rightarrow \infty.$$

Beispielsweise gilt $n^{-k} \exp(-cn) \simeq \exp(-cn)$ für alle $k, c \in \mathbb{R}$.

Um exponentielle Äquivalenz zu zeigen werden wir häufig separat eine Abschätzung nach oben („obere Schranke“) und eine Abschätzung nach unten („untere Schranke“) beweisen. Entsprechend schreiben wir „ $a_n \preceq b_n$ “, falls

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0$$

ist. Beispielsweise gilt für reelle Zahlen c, d, k, l :

$$n^{-k} \exp(-cn) \preceq n^{-l} \exp(-dn) \iff c \geq d.$$

3.1. Entropie und relative Entropie

Wir definieren zunächst die Entropie einer diskreten Wahrscheinlichkeitsverteilung und die relative Entropie zweier beliebiger Wahrscheinlichkeitsmaße. Mithilfe des Gesetzes der großen Zahlen können wir statistische Interpretationen dieser Größen geben. Insbesondere misst die relative Entropie die Unterscheidbarkeit zweier Wahrscheinlichkeitsmaße durch Folgen von unabhängigen Stichproben.

Entropie und Information

Wir bemerken zunächst, dass die auf $[0, \infty)$ definierte Funktion

$$u(x) := \begin{cases} x \log x & \text{für } x > 0 \\ 0 & \text{für } x = 0 \end{cases}$$

stetig und strikt konvex ist mit $u'(x) = 1 + \log x$ und $u''(x) = 1/x$ für $x > 0$. Insbesondere gilt

$$u(x) \leq 0 \quad \text{für alle } x \in [0, 1], \quad (3.1)$$

$$u(x) \geq x - 1 \quad \text{für alle } x \geq 0, \quad (3.2)$$

und $u(1/e) = -1/e$ ist das absolute Minimum der Funktion u .

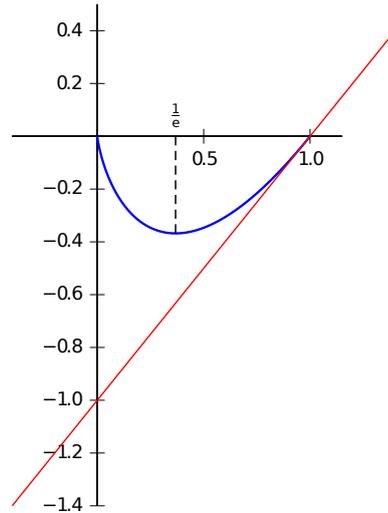


Abbildung 3.1.: Graph der Funktion $u(x)$ (blau) und ihrer unteren Schranke $x - 1$ (rot)

Sei nun S eine abzählbare Menge, und $\mu = (\mu(x))_{x \in S}$ eine Wahrscheinlichkeitsverteilung auf S .

Definition 3.2 (Entropie einer diskreten Wahrscheinlichkeitsverteilung). Die Größe

$$H(\mu) := - \sum_{\substack{x \in S \\ \mu(x) \neq 0}} \mu(x) \log \mu(x) = - \sum_{x \in S} u(\mu(x)) \in [0, \infty]$$

heißt **Entropie** der Wahrscheinlichkeitsverteilung μ .

Anschaulich können wir $-\log \mu(x)$ interpretieren als Maß für die »Überraschung« bzw. den »Informationsgewinn«, falls eine Stichprobe von der Verteilung μ den Wert x hat. Die »Überraschung« ist umso größer, je unwahrscheinlicher x ist. Die Entropie $H(\mu)$ ist dann die »mittlere Überraschung« bzw. der »mittlere Informationsgewinn« beim Ziehen einer Stichprobe von ν . Eine wichtige Eigenschaft der Entropie, die auch die Wahl des Logarithmus erklärt, ist:

Satz 3.3 (Faktorisierungseigenschaft). Für beliebige diskrete Wahrscheinlichkeitsverteilungen ν und μ gilt:

$$H(\nu \otimes \mu) = H(\nu) + H(\mu).$$

Der mittlere Informationszuwachs in einem aus zwei unabhängigen Experimenten zusammengesetzten Zufallsexperiment ist also die Summe der einzelnen mittleren Informationszuwächse.

Beweis. Nach Definition der Entropie gilt:

$$\begin{aligned} H(\nu \otimes \mu) &= \sum_{\substack{x,y \\ \nu(x)\mu(y) \neq 0}} \nu(x)\mu(y) \log(\nu(x)\mu(y)) \\ &= - \sum_{x:\nu(x) \neq 0} \nu(x) \log(\nu(x)) - \sum_{y:\mu(y) \neq 0} \mu(y) \log(\mu(y)) \\ &= H(\nu) + H(\mu). \end{aligned}$$

■

3. Relative Entropie, Information und statistische Unterscheidbarkeit

Wir bestimmen nun auf einer gegebenen abzählbaren Menge S die Wahrscheinlichkeitsverteilungen mit minimaler bzw. maximaler Entropie.

Entropieminima

Nach (3.1) ist die Entropie stets nicht-negativ. Zudem gilt:

$$H(\mu) = 0 \iff \mu(x) \in \{0, 1\} \quad \forall x \in S \iff \mu \text{ ist ein Dirac-Ma\ss.}$$

Die Dirac-Ma\ss e sind also die Entropieminima. Ist das Zufallsexperiment deterministisch, d.h. μ ein Diracma\ss, dann tritt bei Ziehen einer Stichprobe von μ keine \u00c4berraschung bzw. kein Informationszuwachs auf.

Entropiemaximum

Ist S endlich, dann gilt f\u00fcr alle Wahrscheinlichkeitsverteilungen μ auf S :

$$H(\mu) \leq \log(|S|) = H(\text{Unif}_S),$$

wobei Unif_S die Gleichverteilung auf S ist. Nach der Jensenschen Ungleichung gilt n\u00e4mlich

$$\begin{aligned} -\sum_{x \in S} u(\mu(x)) &= -|S| \cdot \int u(\mu(x)) \text{Unif}_S(dx) \\ &\leq -|S| \cdot u\left(\int \mu(x) \text{Unif}_S(dx)\right) \\ &= -|S| \cdot u(1/|S|) = \log|S| \end{aligned}$$

mit Gleichheit genau dann, wenn μ die Gleichverteilung ist.

Die Gleichverteilung maximiert also die Entropie auf einem endlichen Zustandsraum. Anschaulich k\u00f6nnen wir die Gleichverteilung als eine »v\u00f6llig zuf\u00e4llige« Verteilung auffassen – d.h. wir verwenden die Gleichverteilung als Modell, wenn wir keinen Grund haben, einen der Zust\u00e4nde zu bevorzugen. Die Entropie ist in diesem Sinne ein Ma\ss f\u00fcr die »Zuf\u00e4lligkeit« (bzw. »Unordnung«) der Wahrscheinlichkeitsverteilung μ .

Ist S abz\u00e4hlbar unendlich, dann gibt es Wahrscheinlichkeitsverteilungen auf S mit unendlicher Entropie.

Als n\u00e4chstes geben wir eine statistische Interpretation der Entropie. Sei μ eine Wahrscheinlichkeitsverteilung auf einer abz\u00e4hlbaren Menge S . Die Wahrscheinlichkeit einer Folge von Ausg\u00e4ngen x_1, \dots, x_n bei Entnehmen einer Stichprobe aus n unabh\u00e4ngigen Zufallsgr\u00f6\ss en mit Verteilung μ betr\u00e4gt

$$L(\mu; x_1, \dots, x_n) = \prod_{i=1}^n \mu(x_i).$$

Der gemittelte Informationszuwachs durch Auswertung der Werte x_1, \dots, x_n ist also

$$-\frac{1}{n} \log L(\mu; x_1, \dots, x_n).$$

Mithilfe des Gesetzes der gro\ss en Zahlen k\u00f6nnen wir die Asymptotik dieser Gr\u00f6\ss en f\u00fcr $n \rightarrow \infty$ untersuchen:

Satz 3.4 (Entropie als asymptotische Informationszuwachsrate). Seien $X_1, X_2, \dots : \Omega \rightarrow S$ unter P unabh\u00e4ngige Zufallsvariablen mit Verteilung μ . Dann gilt P -fast sicher :

$$-\frac{1}{n} \log L(\mu; X_1, \dots, X_n) \longrightarrow H(\mu) \quad \text{f\u00fcr } n \rightarrow \infty.$$

In der zu Beginn dieses Kapitels eingeführten Notation zur Äquivalenz auf der exponentiellen Skala besagt die Aussage des Satzes, dass fast sicher

$$L(\mu; X_1, \dots, X_n) \simeq e^{-nH(\mu)} \quad \text{gilt.}$$

Beweis. Mit Wahrscheinlichkeit 1 gilt $0 < \mu(X_i) \leq 1$, also $-\log \mu(X_i) \in [0, \infty)$ für alle i . Nach dem Gesetz der großen Zahlen folgt, dass fast sicher

$$-\frac{1}{n} \log L(\mu; X_1, \dots, X_n) = -\frac{1}{n} \sum_{i=1}^n \log \mu(X_i) \xrightarrow{n \rightarrow \infty} -\int \log \mu \, d\mu = H(\mu)$$

gilt. ■

Relative Entropie

Die Entropie ist nur für diskrete Wahrscheinlichkeitsmaße definiert. Eine Übertragung der Definition auf absolutstetige Wahrscheinlichkeitsmaße auf \mathbb{R}^d ist möglich, indem man

$$H(\mu) = -\int_{\mathbb{R}^d} f \log f \, dx \quad \text{mit} \quad f = d\mu/dx$$

setzt. Allerdings kann die so definierte Entropie sowohl positive als auch negative Werte annehmen. Wir führen jetzt den allgemeineren Begriff der *relativen Entropie* zweier Wahrscheinlichkeitsmaße auf einem beliebigen meßbaren Raum (S, \mathcal{B}) ein:

Definition 3.5 (Relative Entropie zweier Wahrscheinlichkeitsmaße). Seien μ und ν Wahrscheinlichkeitsmaße auf (S, \mathcal{B}) . Die durch

$$H(\mu | \nu) := \begin{cases} \int \log w \, d\mu = \int w \log w \, d\nu & \text{falls } \mu \ll \nu \text{ mit relativer Dichte } w, \\ \infty & \text{sonst,} \end{cases} \quad (3.3)$$

definierte Größe $H(\mu | \nu) \in [0, \infty]$ heißt *relative Entropie* (oder *Kullback-Leibler Information*) von μ bzgl. ν .

Um eine anschauliche Interpretation der relativen Entropie zu geben, nehmen wir an, dass μ und ν Wahrscheinlichkeitsmaße auf $S = \mathbb{R}^d$ oder einem diskreten Raum mit Dichten (bzw. Massenfunktionen) $f, g > 0$ sind. Die relative Dichte w von μ bzgl. ν ist dann

$$w(x) = \frac{d\mu}{d\nu}(x) = \frac{f(x)}{g(x)} \quad \text{für } \nu\text{-fast alle } x \in S,$$

und

$$H(\mu | \nu) = \int \log \frac{f}{g} \, d\mu = \int (-\log g(x) - (-\log f(x))) \mu(dx).$$

Wir können $-\log g(x)$ und $-\log f(x)$ als Maß für die Überraschung (den Informationsgewinn) bei Eintreten von x interpretieren, falls ν bzw. μ das zugrundeliegende Modell ist. Wenn wir also ν als Modell annehmen, aber tatsächlich μ die zugrundeliegende Verteilung ist, dann erhöht sich die Überraschung (der Informationszuwachs) bei Ziehen einer Stichprobe im Vergleich zum korrekten Modell im Mittel um $H(\mu | \nu)$.

3. Relative Entropie, Information und statistische Unterscheidbarkeit

Bemerkung (Entropie als Spezialfall der relativen Entropie). Ist ν das Zählmaß auf einer abzählbaren Menge S , dann gilt

$$H(\mu | \nu) = \sum_{\mu(x) \neq 0} \mu(x) \log \mu(x) = -H(\mu). \quad (3.4)$$

Ist S endlich, und ν die Gleichverteilung (also das normierte Zählmaß) auf S , dann folgt entsprechend

$$H(\mu | \nu) = \sum_{\mu(x) \neq 0} \mu(x) \log(\mu(x) \cdot |S|) = \log |S| - H(\mu).$$

Aussagen über die relative Entropie liefern also als Spezialfall entsprechende Aussagen für die Entropie (wobei sich aber das Vorzeichen umkehrt!)

Das folgende Lemma fasst elementare Eigenschaften der relativen Entropie zusammen:

Lemma 3.6 (Eigenschaften der relativen Entropie).

(i) Es gilt $H(\mu | \nu) \geq 0$, mit Gleichheit genau dann, wenn $\mu = \nu$.

(ii)
$$H(\mu_1 \otimes \dots \otimes \mu_n | \nu_1 \otimes \dots \otimes \nu_n) = \sum_{i=1}^n H(\mu_i | \nu_i).$$

Beweis. Sei $\mu \ll \nu$ mit relativer Dichte w . Wegen $x \log x \geq x - 1$ folgt

$$H(\mu | \nu) = \int w \log w \, d\nu \geq \int (w - 1) \, d\nu = \int w \, d\nu - 1 = 0.$$

Gleichheit gilt genau dann, wenn w ν -fast sicher gleich 1, also $\mu = \nu$ ist. Der Beweis der zweiten Aussage ist eine Übungsaufgabe. ■

Die folgenden Beispiele zeigen, dass die relative Entropie im Allgemeinen *nicht symmetrisch* ist.

Beispiele (Relative Entropie von Bernoulli- und Binomialverteilungen).

(i) Für die Bernoulli-Verteilungen mit $\nu_p(1) = p$ und $\nu_p(0) = 1 - p$ gilt:

$$H(\nu_a | \nu_p) = a \log \left(\frac{a}{p} \right) + (1 - a) \log \left(\frac{1 - a}{1 - p} \right) \quad \text{für alle } a, p \in (0, 1).$$

(ii) Für Normalverteilungen mit Mittelwerten $m, \tilde{m} \in \mathbb{R}$ und Varianzen $v, \tilde{v} > 0$ gilt

$$H(N(\tilde{m}, \tilde{v}) | N(m, v)) = \frac{1}{2} \left(\log \left(+ \frac{(\tilde{m} - m)^2}{v} + \frac{\tilde{v}}{v} - 1 + \frac{v}{\tilde{v}} \right) \right), \quad \text{also insbesondere}$$

$$H(N(\tilde{m}, v) | N(m, v)) = \frac{(\tilde{m} - m)^2}{2v}.$$

Die Chiquadrat-Divergenz wird häufig als Approximation der relativen Entropie verwendet. Im Gegensatz zur relativen Entropie erfüllt sie aber keine Faktorisierungseigenschaft, und kann daher in hohen Dimensionen sehr große Werte annehmen.

Wir geben nun eine statistische Interpretation der relativen Entropie. Dazu nehmen wir wieder an, dass μ und ν Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}^d$ oder einem diskreten Raum mit Dichten (bzw. Massenfunktionen) $f, g > 0$, und relativer Dichte $w = f/g$ sind.

Minimierung unter Nebenbedingungen

Gegeben sei eine Wahrscheinlichkeitsverteilung μ_0 mit Dichte $f_{\mu_0}(x)$ und einer Statistik $T : S \rightarrow \mathbb{R}$. Wir wollen eine Wahrscheinlichkeitsverteilung μ finden, welche die relative Entropie $H(\mu|\mu_0)$ minimiert unter der Nebenbedingung, dass $\int T d\mu = m$ wobei $m \in \mathbb{R}$ gegeben ist. Hierfür betrachten wir eine exponentielle Familie mit Wahrscheinlichkeitsverteilungen μ_{ϑ} ($\vartheta \in \Theta$), $0 \in \Theta$ und $\Theta \subseteq \mathbb{R}$ offen, wobei die Dichtefunktionen gegeben sind durch

$$f_{\vartheta}(x) = \frac{1}{Z(\vartheta)} e^{\vartheta T(x)} \quad \text{mit} \quad Z(\vartheta) \in (0, \infty).$$

Um die Nebenbedingung zu beschreiben definieren wir die folgende Funktion

$$m(\vartheta) = \int T d\mu_{\vartheta} \quad \text{für} \quad \vartheta \in \mathbb{R}$$

Nach Lemma 2.7 wissen wir, dass gilt $m(\vartheta) = -d'(\vartheta) = (\log Z)'(\vartheta)$. Somit gilt insbesondere

$$\mu_{\vartheta} \ll \mu_0 \quad \text{mit} \quad \frac{d\mu_{\vartheta}}{d\mu_0}(x) = \frac{Z(0)}{Z(\vartheta)} e^{\vartheta T(x)} \quad (*)$$

Nun liefert und der folgende Satz das gewünschte Ergebnis.

Satz 3.7 (Gibbs'sches Variationsprinzip).

- 1) Die Wahrscheinlichkeitsverteilung μ_{ϑ} ist ein Minimum der Funktion

$$F(\mu) := H(\mu|\mu_0) - \vartheta \int T d\mu = H(\mu|\mu_0) - \vartheta m(\vartheta)$$

- 2) Somit ist die Relative Entropie gegeben durch

$$H(\mu_{\vartheta}|\mu_0) = \min \left\{ H(\mu|\mu_0) : \mu \text{ ist Wahrscheinlichkeitsverteilung mit } \int T d\mu = m(\vartheta) \right\}$$

Beweis.

- 1) Mit (*) folgt:

$$\begin{aligned} H(\mu|\mu_0) &= \int \log \frac{d\mu}{d\mu_0} d\mu = \int \log \frac{d\mu}{d\mu_{\vartheta}} d\mu + \int \log \frac{d\mu_{\vartheta}}{d\mu_0} d\mu \\ &= H(\mu|\mu_{\vartheta}) + \log \frac{Z(0)}{Z(\vartheta)} + \vartheta \int T d\mu \end{aligned}$$

Es folgt also, dass gilt $F(\mu) = H(\mu|\mu_{\vartheta}) + \log \frac{Z(0)}{Z(\vartheta)}$. Da $\mu \mapsto H(\mu|\mu_{\vartheta})$ durch μ_{ϑ} minimiert wird, folgt die Aussage. Die 2. Behauptung folgt nun aus der 1. Aussage. ■

Bemerkung (Statistische Physik). In der statistischen Physik spielt die Boltzmann-Gibbs-Verteilung μ_{ϑ} eine zentrale Rolle. Sie beschreibt die Wahrscheinlichkeitsverteilung von Zuständen eines Systems im thermischen Gleichgewicht. Hierbei ist $\vartheta = \frac{1}{T}$ die inverse Temperatur und $U(x)$ die Energie an der Stelle x . Das Prinzip der minimalen freien Energie besagt, dass ein System im Gleichgewicht die freie Energie minimiert. Die freie Energie ist gegeben durch

$$\frac{1}{\vartheta} F(\mu) = \int U d\mu - T \cdot H(\mu|\mu_0)$$

Ziel ist es also, eine Wahrscheinlichkeitsverteilung μ zu finden, welche die freie Energie minimiert, da bei dieser das System im Gleichgewicht ist. Die Boltzmann-Gibbs-Verteilung ist eben jene Verteilung.

3.2. Anwendungen in der Statistik

Wie kann man anhand von unabhängigen Stichproben erkennen, welche der beiden Verteilungen ν und μ in einem Zufallsexperiment vorliegt? Dazu betrachten wir den *Likelihood-Quotienten*

$$\lambda_n(x_1, \dots, x_n) := \frac{L_n(\mu; x_1, \dots, x_n)}{L_n(\nu; x_1, \dots, x_n)} = \frac{\prod_{i=1}^n f(x_i)}{\prod_{i=1}^n g(x_i)} = \prod_{i=1}^n w(x_i),$$

wobei $w = \frac{f}{g}$ die relative Dichte von μ bzgl. ν ist. Analog zu Satz 3.4 erhalten wir die folgende (allgemeinere) Aussage:

Satz 3.8 (Relative Entropie als statistische Unterscheidbarkeit; Shannon-McMillan). Seien $X_1, X_2, \dots : \Omega \rightarrow S$ unabhängige Zufallsvariablen unter P_ν bzw. P_μ mit Verteilung ν bzw. μ . Dann gilt für $n \rightarrow \infty$:

$$\frac{1}{n} \log \lambda_n(X_1, \dots, X_n) \rightarrow H(\mu | \nu) \quad P_\mu\text{-fast sicher, und} \quad (3.5)$$

$$\frac{1}{n} \log \lambda_n(X_1, \dots, X_n) \rightarrow -H(\nu | \mu) \quad P_\nu\text{-fast sicher.} \quad (3.6)$$

Beweis. (i) Für $n \rightarrow \infty$ gilt nach dem Gesetz der großen Zahlen

$$\frac{1}{n} \log \lambda_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \log w(X_i) \rightarrow \int \log w \, d\mu \quad P_\mu\text{-fast sicher.}$$

Das Gesetz der großen Zahlen ist anwendbar, da

$$\int (\log w)^- \, d\mu = \int (w \log w)^- \, d\nu \leq \frac{1}{e} < \infty.$$

(ii) Da ν absolutstetig bzgl. μ mit Dichte $1/w$ ist, gilt entsprechend

$$\begin{aligned} \frac{1}{n} \log \lambda_n(X_1, \dots, X_n) &= -\frac{1}{n} \sum_{i=1}^n \log \frac{1}{w(X_i)} \\ &\xrightarrow{\text{GGZ}} -\int \log \frac{1}{w} \, d\nu = -H(\nu | \mu) \quad P_\nu\text{-fast sicher.} \quad \blacksquare \end{aligned}$$

Der Satz von Shannon-Mc Millan zeigt, dass sich die Produktdichte (der Likelihood-Quotient) asymptotisch auf der exponentiellen Skala (d.h. unter Vernachlässigung subexponentiell wachsender Faktoren) folgendermaßen verhält:

$$\lambda_n(X_1, \dots, X_n) \simeq \begin{cases} e^{nH(\mu | \nu)} & P_\mu\text{-fast sicher,} \\ e^{-nH(\nu | \mu)} & P_\nu\text{-fast sicher.} \end{cases}$$

Damit erhalten wir eine statistische Interpretation der relativen Entropie als natürlichen (*nicht-symmetrischen!*) Abstandsbegriff für Wahrscheinlichkeitsmaße. Wir werden diese statistische Interpretation im folgenden noch weiter präzisieren.

Konsistenz von Maximum-Likelihood-Schätzern

Als erste Folgerung aus Satz 3.8 zeigen wir, dass Maximum-Likelihood-Schätzer unter geeigneten Voraussetzungen konsistent sind. Sei dazu ν_ϑ ($\vartheta \in \Theta$) eine einparametrische (d.h. $\Theta \subseteq \mathbb{R}$) Familie von Wahrscheinlichkeitsverteilungen mit Dichten bzw. Massenfunktionen f_ϑ . Es gelte:

Annahme (Unimodalität): Für alle $n \in \mathbb{N}$ und $x \in S^n$ existiert ein $\hat{\vartheta}_n(x_1, \dots, x_n)$, sodass

$$\vartheta \mapsto L_n(\vartheta; x_1, \dots, x_n) \begin{cases} \text{ist monoton wachsend für } \vartheta \leq \hat{\vartheta}_n(x_1, \dots, x_n). \\ \text{ist monoton fallend für } \vartheta \geq \hat{\vartheta}_n(x_1, \dots, x_n). \end{cases}$$

Bemerkung. (i) Die Annahme ist z.B. erfüllt, falls $\vartheta \mapsto \log f_\vartheta(x)$ für jedes x konkav ist - denn dann ist auch $\log L_n(\vartheta, x_1, \dots, x_n) = \sum_{i=1}^n \log f_\vartheta(x_i)$ konkav in ϑ .

(ii) $\hat{\vartheta}_n(X_1, \dots, X_n)$ ist im unimodalen Fall eindeutiger Maximum-Likelihood-Schätzer für ϑ .

Satz 3.9 (Konsistenz von Maximum-Likelihood-Schätzern). Es gelte die Annahme, sowie $\nu_\vartheta \neq \nu_{\tilde{\vartheta}}$ für $\vartheta \neq \tilde{\vartheta}$. Dann ist $\hat{\vartheta}_n(X_1, \dots, X_n)$ ($n \in \mathbb{N}$) eine *konsistente* Folge von Schätzern für ϑ , d.h. für jedes $\varepsilon > 0$ gilt

$$P_\vartheta [|\hat{\vartheta}_n(X_1, \dots, X_n) - \vartheta| < \varepsilon] \rightarrow 1 \quad \text{für } n \rightarrow \infty.$$

Beweis. Wegen der Unimodalität gilt $\hat{\vartheta}_n(x_1, \dots, x_n) \in (\vartheta - \varepsilon, \vartheta + \varepsilon)$ falls

$$L_n(\vartheta; x_1, \dots, x_n) > L_n(\vartheta \pm \varepsilon; x_1, \dots, x_n).$$

Also:

$$P_\vartheta [|\hat{\vartheta}_n(X_1, \dots, X_n) - \vartheta| < \varepsilon] \geq P_\vartheta \left[\frac{L_n(\vartheta; X_1, \dots, X_n)}{L_n(\vartheta \pm \varepsilon; X_1, \dots, X_n)} > 1 \right].$$

Die rechte Seite konvergiert aber für $n \rightarrow \infty$ nach Satz 3.8 für jedes ϑ gegen 1. ■

Asymptotische Macht von Likelihood-Quotienten-Tests

Wir nehmen nun wieder an, dass μ und ν Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}^d$ oder einem diskreten Raum mit Dichten (bzw. Massenfunktionen) $f, g > 0$, und relativer Dichte $w = f/g$ sind.

Seien X_1, X_2, \dots unter P_ν bzw. P_μ unabhängige Zufallsvariablen mit Verteilung ν bzw. μ . Wir betrachten den *Likelihood-Quotienten*

$$\lambda_n(x_1, \dots, x_n) := \frac{L_n(\mu; x_1, \dots, x_n)}{L_n(\nu; x_1, \dots, x_n)} = \frac{\prod_{i=1}^n f(x_i)}{\prod_{i=1}^n g(x_i)} = \prod_{i=1}^n w(x_i)$$

für n unabhängige Stichproben x_1, \dots, x_n . Nach dem Neyman-Pearson-Lemma ist der Likelihood-Quotienten-Test mit Verwerfungsbereich

$$C = \{w_n > c\}$$

für jedes $c \in (0, \infty)$ der mächtigste Test für das Testproblem

$$H_0: \quad \gg X_i \sim \nu \ll \quad \text{vs.} \quad H_1: \quad \gg X_i \sim \mu \ll$$

zum Niveau $\alpha = \nu^n(w_n > c)$. Wie mächtig ist dieser Test asymptotisch für große n ?

Satz 3.10 (Asymptotische Macht des Likelihoodquotiententests). Sei $\alpha \in (0, 1)$ ein festes Niveau, und sei $c_n \in (0, \infty)$ ($n \in \mathbb{N}$) mit

$$\nu^n(w_n > c_n) \leq \alpha \leq \nu^n(w_n \geq c_n) \quad (3.7)$$

Dann gilt:

(i)

$$\frac{1}{n} \log c_n \longrightarrow -H(\nu|\mu) \quad \text{für } n \rightarrow \infty.$$

(ii)

$$\frac{1}{n} \log \mu^n(w_n \leq c_n) \longrightarrow -H(\nu|\mu) \quad \text{für } n \rightarrow \infty,$$

d.h. die Wahrscheinlichkeit für den Fehler 2. Art fällt exponentiell mit Rate $H(\nu|\mu)$.

Der Satz demonstriert erneut, daß die relative Entropie ein gutes Maß für die Unterscheidbarkeit zweier Wahrscheinlichkeitsverteilungen ist. Der Beweis basiert auf dem Satz von Shannon-McMillan und einem Lemma zu unteren Schranken für Wahrscheinlichkeiten bei Maßwechsel, dass wir im Anschluss beweisen werden.

Beweis. (i) Sei $\varepsilon > 0$. Für große n gilt nach dem Satz von Shannon-McMillan

$$\nu^n(w_n > e^{-n(H(\nu|\mu)+\varepsilon)}) > \alpha \stackrel{(3.7)}{\geq} \nu^n(w_n > c_n).$$

Es folgt $e^{-n(H(\nu|\mu)+\varepsilon)} < c_n$. Analog zeigt man $e^{-n(H(\nu|\mu)-\varepsilon)} > c_n$. Die Behauptung folgt dann für $\varepsilon \rightarrow 0$.

(ii) *Untere Schranke:* Wegen

$$\nu^n(w_n \leq c_n) \geq 1 - \alpha > 0 \quad \forall n \in \mathbb{N}$$

folgt nach Lemma 3.12, dass

$$\liminf \frac{1}{n} \log \mu^n(w_n \leq c_n) \geq -H(\nu|\mu).$$

Obere Schranke: Wegen

$$\mu^n(w_n \leq c_n) = \int_{w_n \leq c_n} w_n d\nu^n \leq c_n$$

folgt nach (i) die Abschätzung

$$\limsup \frac{1}{n} \log \mu^n(w_n \leq c_n) \leq \limsup \frac{1}{n} \log c_n = -H(\nu|\mu),$$

und damit die Behauptung. ■

Wir kommen nun zum Beweis des oben verwendeten Lemmas. Dieses spielt auch allgemein in der Theorie großer Abweichungen eine wichtige Rolle. Wir beginnen mit einer Definition.

Definition 3.11 (Wesentliche Mengen in Produktmodellen). Eine Folge von Mengen $B_n \subseteq S^n$ ($n \in \mathbb{N}$) heißt *wesentlich bzgl. ν* , falls

$$\liminf_{n \rightarrow \infty} P_\nu[(X_1, \dots, X_n) \in B_n] = \liminf_{n \rightarrow \infty} \nu^n[B_n] > 0.$$

Wie wahrscheinlich muss eine bzgl. ν wesentliche Folge von Mengen unter μ mindestens sein? Das folgende Lemma beantwortet diese Frage auf der exponentiellen Skala.

Lemma 3.12 (Untere Schranken für Wahrscheinlichkeiten bei Maßwechsel). (i) Für $\varepsilon > 0$ sei

$$B_{n,\varepsilon} := \{(x_1, \dots, x_n) : e^{-n(H(\nu|\mu)+\varepsilon)} \leq \lambda_n(x_1, \dots, x_n) \leq e^{-n(H(\nu|\mu)-\varepsilon)}\} \subseteq S^n.$$

Dann gilt $\lim_{n \rightarrow \infty} \nu[B_{n,\varepsilon}] = 1$ (insbesondere ist $(B_{n,\varepsilon})_{n \in \mathbb{N}}$ wesentlich bzgl. ν), und

$$\mu^n[B_{n,\varepsilon}] \leq e^{-n(H(\nu|\mu)-\varepsilon)} \quad \text{für alle } n \in \mathbb{N}. \quad (3.8)$$

(ii) Für beliebige messbare Mengen $A_n \subseteq S^n$ mit

$$\liminf \nu^n[A_n] > 0 \quad (3.9)$$

gilt

$$\liminf \frac{1}{n} \log \mu^n[A_n] \geq -H(\nu|\mu). \quad (3.10)$$

Beweis. (i) Die erste Aussage folgt nach Satz 3.8. Zudem gilt

$$1 \geq \nu^n[B_{n,\varepsilon}] = \int_{B_{n,\varepsilon}} \frac{1}{w_n} d\mu^n \geq \mu^n[B_{n,\varepsilon}] \cdot e^{n(H(\nu|\mu)-\varepsilon)}.$$

(ii) Aus

$$\mu^n[A_n] = \int_{A_n} w_n d\nu_n \geq e^{-n(H(\nu|\mu)+\varepsilon)} \nu^n[A_n \cap B_{n,\varepsilon}]$$

folgt

$$\begin{aligned} \liminf \frac{1}{n} \log \mu^n[A_n] &\geq -(H(\nu|\mu) + \varepsilon) + \liminf \frac{1}{n} \log \nu^n[A_n \cap B_{n,\varepsilon}] \\ &= -(H(\nu|\mu) + \varepsilon), \end{aligned}$$

da $\liminf \nu^n[A_n \cap B_{n,\varepsilon}] = \liminf \nu^n[A_n] > 0$ nach (i) gilt. Die Behauptung folgt für $\varepsilon \rightarrow 0$. \blacksquare

Die zweite Aussage der Satzes können wir als eine allgemeine untere Schranke für große Abweichungen interpretieren: Ist $A_n \subseteq S^n$ eine beliebige Folge von Ereignissen, dann liefert uns (3.10) für jede Wahrscheinlichkeitsverteilung ν mit (3.9) eine asymptotische untere Schranke für die Wahrscheinlichkeiten

$$P_\mu[(X_1, \dots, X_n) \in A_n] = \mu^n[A_n]$$

auf der exponentiellen Skala.

3.3. Fisher-Information

Fisher-Informationsungleichung

Wir betrachten nun wieder ein parametrisches Modell. Der Einfachheit halber nehmen wir an, dass der Parameterraum Θ ein offenes Teilintervall von \mathbb{R} ist. Entsprechende Aussagen gelten aber auch für mehrdimensionale Parametermengen. Sei μ_ϑ ($\vartheta \in \Theta$) eine Familie von Wahrscheinlichkeitsverteilungen auf $S \subseteq \mathbb{R}^d$ (bzw. auf einem diskreten Raum S) mit Dichten (bzw. Massenfunktionen) f_ϑ . Für die folgenden Aussagen benötigen wir zudem unterschiedliche Regularitätsannahmen. Die folgende (relativ starke) Annahme ist für alle Aussagen hinreichend, kann aber noch deutlich abgeschwächt werden.

Annahme (Regularität): Die folgenden Bedingungen sind erfüllt:

3. Relative Entropie, Information und statistische Unterscheidbarkeit

- *Identifizierbarkeit:* Für $\vartheta \neq \bar{\vartheta}$ gilt $\mu_{\vartheta} \neq \mu_{\bar{\vartheta}}$.
- *Nicht-degeneriert:* Für alle $x \in S$ und $\vartheta \in \Theta$ ist $f_{\vartheta}(x) > 0$.
- *Glattheit der Log-Likelihood:* Für jedes $x \in S$ ist $\vartheta \mapsto f_{\vartheta}(x)$ dreimal stetig differenzierbar, und für $k = 2$ sowie $k = 3$ gilt

$$\sup_{x, \vartheta} \left| \frac{\partial^k}{\partial \vartheta^k} \log f_{\vartheta}(x) \right| < \infty.$$

- *Vertauschen von Ableitung und Integral:* Für $k = 1$ und $k = 2$ gilt

$$\int \frac{\partial^k}{\partial \vartheta^k} f_{\vartheta}(x) dx = \frac{\partial^k}{\partial \vartheta^k} \int f_{\vartheta}(x) dx = 0.$$

Unter den Annahmen ist die log-Likelihood

$$\ell(\vartheta; x) = \log f_{\vartheta}(x)$$

eine glatte Funktion, deren erste Ableitung die **Score-Funktion**

$$\ell'(\vartheta) = \frac{\partial}{\partial \vartheta} \ell(\vartheta; x) = \frac{1}{f_{\vartheta}(x)} \frac{\partial}{\partial \vartheta} f_{\vartheta}(x)$$

ist. Die Score-Funktion misst also, wie stark die log-Likelihood vom Parameter ϑ abhängt. Um zu quantifizieren wie stark die Verteilung von ϑ abhängt, betrachten wir die Fisher-Information.

Definition 3.13 (Fisher-Information). Die **Fisher-Information** des parametrischen Modells ist für $\vartheta \in \Theta$ definiert als

$$I(\vartheta) = \int \ell'(\vartheta)^2 f_{\vartheta} dx = \int \frac{1}{f_{\vartheta}(x)} \left| \frac{\partial f_{\vartheta}(x)}{\partial \vartheta} \right|^2 dx.$$

Bemerkung. Es gilt

$$I(\vartheta) = \int \ell'(\vartheta, x)^2 f_{\vartheta}(x) dx = \int \frac{1}{f_{\vartheta}} \left(\frac{\partial f_{\vartheta}}{\partial \vartheta} \right)^2 dx$$

Lemma 3.14. Aus den Regularitätsannahmen folgt unmittelbar

1)

$$\int \ell'(\vartheta) f_{\vartheta} dx = 0, \quad \text{und} \quad I(\vartheta) = \text{Var}[\ell''(\vartheta, x)], \quad (3.11)$$

2)

$$\mathbb{E}_{\vartheta}[\ell''(\vartheta, x)] = - \int \ell''(\vartheta) f_{\vartheta} dx = I(\vartheta) \quad (3.12)$$

Beweis. Denn

$$\int f_{\vartheta} dx = 1 \quad \forall \vartheta \in \Theta \quad \Rightarrow \quad 0 = \frac{\partial}{\partial \vartheta} \int f_{\vartheta} dx = \int \frac{\partial f_{\vartheta}}{\partial \vartheta} dx = \int \ell'(\vartheta) f_{\vartheta} dx,$$

und somit

$$0 = \int \frac{\partial^2}{\partial \vartheta^2} f_{\vartheta} dx = \int \ell''(\vartheta) f_{\vartheta} dx + \int \ell'(\vartheta) \frac{\partial}{\partial \vartheta} f_{\vartheta} dx = \int \ell''(\vartheta) f_{\vartheta} dx + I(\vartheta)$$

denn

$$\ell'' f_{\vartheta} + \ell' \frac{\partial f_{\vartheta}}{\partial \vartheta} = \left(\frac{f'_{\vartheta}}{f_{\vartheta}} \right)' f_{\vartheta} + \frac{f''_{\vartheta}}{f_{\vartheta}} = \frac{f''_{\vartheta} f_{\vartheta} - f'_{\vartheta}{}^2}{f_{\vartheta}} + \frac{f''_{\vartheta}}{f_{\vartheta}} = f''_{\vartheta}$$

■

Insbesondere ist die Fisher-Information sowohl die Varianz der Score-Funktion als auch der Erwartungswert der negativen zweiten Ableitung der log-Likelihood. Letztere kann geometrisch als Krümmung des Modells interpretiert werden.

Beispiele.

- 1) Für $p \in (0, 1)$ betrachten wir die Bernoulliverteilung $\mu_p = \text{Bernoulli}(p)$, mit $S = \{0, 1\}$. Dann ist die log-Likelihoodfunktion gegeben durch

$$\ell(p, 0) = \log(1 - p), \quad \ell(p, 1) = \log(p) \quad \text{mit} \quad \ell'(p, 0) = -\frac{1}{1-p}, \quad \ell'(p, 1) = \frac{1}{p}$$

Wir erhalten die folgende Fisher-Information

$$I(p) = \mathbb{E}_p[\ell'(p, X)^2] = \left(\frac{1}{1-p}\right)^2 (1-p) + \frac{1}{p^2} p = \frac{1}{1-p} + \frac{1}{p} = \frac{1}{p(1-p)}$$

- 2) Sei nun $\mu_m = N(m, v)$ die Normalverteilung mit fester Varianz $v > 0$ und der log-Likelihoodfunktion

$$\ell(m, v) = -\frac{1}{2} \log(2\pi v) - \frac{(x-m)^2}{2v} \quad \Rightarrow \quad \ell'(m, x) = \frac{x-m}{v}$$

Wir erhalten die Fisher-Information

$$I(m) = \mathbb{E}_m \left[\left(\frac{x-m}{v} \right)^2 \right] = \frac{1}{v^2} \underbrace{\mathbb{E}[(x-m)^2]}_{=v} = \frac{1}{v}$$

Der folgende Satz zeigt, dass die Fisher-Information die Unterscheidbarkeit von den Verteilungen μ_ϑ und μ_{ϑ_0} für $\vartheta \approx \vartheta_0$ quantifiziert.

Satz 3.15 (Zusammenhang von relativer Entropie und Fisher-Information). Sei $\vartheta_0 \in \Theta$ ein fester Parameterwert. Dann gilt unter Regularitätsvoraussetzungen

$$H(\mu_{\vartheta_0} | \mu_\vartheta) = \frac{1}{2} I(\vartheta_0) \cdot (\vartheta - \vartheta_0)^2 + O(|\vartheta - \vartheta_0|^3), \quad \text{und} \quad (3.13)$$

$$H(\mu_\vartheta | \mu_{\vartheta_0}) = \frac{1}{2} I(\vartheta_0) \cdot (\vartheta - \vartheta_0)^2 + O(|\vartheta - \vartheta_0|^3). \quad (3.14)$$

Beweis. Die Aussagen folgen durch eine Taylor-Entwicklung der relativen Entropie. Sei

$$h(\vartheta) := H(\mu_{\vartheta_0} | \mu_\vartheta) = \int \log(f_{\vartheta_0}/f_\vartheta) f_{\vartheta_0}.$$

Dann gilt $h(\vartheta_0) = 0$. Wegen $\log(f_{\vartheta_0}/f_\vartheta) = \ell(\vartheta_0) - \ell(\vartheta)$ folgt zudem

$$h'(\vartheta) = - \int \ell'(\vartheta) f_{\vartheta_0} \quad \text{und} \quad h''(\vartheta) = - \int \ell''(\vartheta) f_{\vartheta_0},$$

und somit $h'(\vartheta_0) = 0$ und $h''(\vartheta_0) = I(\vartheta_0)$ nach (3.12). Also erhalten wir unter den obigen Regularitätsvoraussetzungen die Taylor-Entwicklung

$$h(\vartheta) = h(\vartheta_0) + (\vartheta - \vartheta_0)h'(\vartheta_0) + \frac{1}{2}(\vartheta - \vartheta_0)^2 h''(\vartheta_0) + O(|\vartheta - \vartheta_0|^3) = \frac{1}{2} I(\vartheta_0) \cdot (\vartheta - \vartheta_0)^2 + O(|\vartheta - \vartheta_0|^3).$$

Die zweite Aussage zeigt man auf ähnliche Weise. ■

Bemerkung (Mehrdimensionale Verallgemeinerung; Fisher-Informationsmatrix und Informationsgeometrie).

Für $\Theta \subseteq \mathbb{R}^d$ folgt auf ähnliche Weise $D_{\vartheta}^2 H(\mu_{\vartheta_0} | \mu_{\vartheta}) = I(\vartheta_0)$ wobei $I(\vartheta_0) \in \mathbb{R}^{d \times d}$ die durch

$$I(\vartheta)_{kl} = \int \frac{\partial}{\partial \vartheta_k} (\log f_{\vartheta}) \frac{\partial}{\partial \vartheta_l} (\log f_{\vartheta}) f_{\vartheta}$$

definierte *Fisher-Informations-Matrix* ist. Unter Regularitätsvoraussetzungen folgt wie im eindimensionalen Fall $I(\vartheta) = - \int D^2 \ell(\vartheta) f_{\vartheta}$. Ist die log-Likelihood konkav, dann definiert die Fisher-Information eine Riemannsche Metrik auf dem Parameterraum, die auf natürliche Weise mit dem zugrundeliegenden statistischen Modell verbunden ist.

Wenn die Fisher-Information groß ist, dann hängt die Verteilung stark von ϑ ab, und es sollte einfacher sein, den Parameter zu schätzen. Umgekehrt sollte es schwierig sein, den Parameter zu schätzen, wenn $I(\vartheta)$ klein ist. Dies wird durch die Informationsungleichung bestätigt.

Satz 3.16 (Informationsungleichung von Cramér-Rao). Sei $g : \Theta \rightarrow \mathbb{R}$ differenzierbar. Ist das Modell regulär, dann gilt für jeden erwartungstreuer Schätzer \hat{g} für $g(\vartheta)$ die untere Schranke

$$\text{MSE}(\hat{g}) = \text{Var}_{\vartheta}[\hat{g}] \geq \frac{g'(\vartheta)^2}{I(\vartheta)} \quad \text{für alle } \vartheta \in \Theta.$$

Beweis. Die Behauptung ergibt sich (nach kurzer Rechnung) durch Anwenden der Cauchy-Schwarz-Ungleichung auf die Kovarianz von \hat{g} und der Score-Funktion $\frac{\partial \ell}{\partial \vartheta}(\vartheta; X)$. Nach dieser gilt:

$$\underbrace{\text{Cov}_{\vartheta}(\hat{g}, \ell'(\vartheta, x))}_{\stackrel{!}{=} g'(\vartheta)} \leq \sqrt{\text{Var}(\hat{g}) \text{Var}(\ell'(\vartheta, x))}$$

Nach (3.11) gilt $\text{Var}(\ell'(\vartheta, x)) = I(\vartheta)$, somit reicht es die angeschriebene Gleichheit zu zeigen. Wir betrachten hierfür

$$\begin{aligned} \hat{g}(\vartheta) &= \frac{\partial}{\partial \vartheta} \mathbb{E}_{\vartheta}[\hat{g}] \stackrel{\hat{g}:=T(x)}{=} \frac{\partial}{\partial \vartheta} \int T(x) f_{\vartheta}(x) dx = \int T(x) \ell'(\vartheta, x) f_{\vartheta}(x) dx \\ &= \mathbb{E}_{\vartheta}[T(x) \ell'(\vartheta, x)] \stackrel{\text{Lem.}}{=} \text{Cov}_{\vartheta}(\hat{g}, \ell'(\vartheta, x)) \end{aligned}$$

■

Auf ähnliche Weise sieht man auch, dass in der Informationsungleichung für einen *erwartungstreuen* Schätzer $\hat{g} = T(X)$ genau dann Gleichheit für alle ϑ gilt, wenn $(\mu_{\vartheta})_{\vartheta \in \Theta}$ eine exponentielle Familie zur Statistik $T(X)$ ist. In diesem Fall nennt man den Schätzer *effizient*. Auf die Voraussetzungen Erwartungstreue und Regularität kann man nicht verzichten. Wir sehen, dass genau dann Gleichheit in der Cauchy-Schwarz-Ungleichung für die Kovarianz gilt, wenn

$$\begin{aligned} &\exists a(\vartheta), b(\vartheta) : \ell'(\vartheta, x) = a(\vartheta)T(x) + b(\vartheta) \quad \mathbb{P}_{\vartheta} \text{ fast sicher} \\ \Leftrightarrow &\ell(\vartheta, x) = c(\vartheta)T(x) + d(\vartheta) + U(x) \quad \text{für alle } \vartheta \in \Theta, \text{ fast alle } x, \text{ wobei } c' = a, d' = b \quad (3.15) \\ \Leftrightarrow &f_{\vartheta}(x) = e^{c(\vartheta)T(x) + d(\vartheta) + U(x)} \quad (\text{exponentielle Familie}) \end{aligned}$$

Korollar 3.17. $T(X)$ ist genau dann ein gleichmäßig varianzminimierender, erwartungstreuer Schätzer für $g(\vartheta)$, falls f_{ϑ} die Form wie in (3.15) hat und es gilt

$$c'(\vartheta)g(\vartheta) = -d'(\vartheta)$$

Das also gilt $g(\vartheta) = -\frac{d'(\vartheta)}{c'(\vartheta)} = \mathbb{E}_{\vartheta}[T(X)]$.

Beispiel (Produktmodell). In einem regulären Produktmodell mit n unabhängigen, identisch verteilten Faktoren ist die Fisher-Information durch $I_n(\vartheta) = nI(\vartheta)$ gegeben, wobei I die Fisher-Information der einzelnen Komponenten ist. Aufgrund der Informationsungleichung gilt also

$$\text{Var}_\vartheta[\hat{g}] \geq \frac{g'(\vartheta)^2}{nI(\vartheta)}$$

für jeden erwartungstreuen Schätzer. Andererseits haben wir in den Übungen „supereffiziente“ erwartungstreue Schätzer $\hat{\vartheta}_n$ in einem (nicht-regulären) Produktmodell betrachtet, deren Varianz mit Ordnung $O(1/n^2)$ abfällt.

Vergleich erwartungstreuer Schätzer

Wir betrachten wieder ein parametrisches Modell mit Parameterraum $\Theta = \mathbb{R}$. Sei dann μ_ϑ eine Familie von Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}$ mit Dichtefunktionen

$$f_\vartheta = g(x - \vartheta), \quad g : \mathbb{R} \rightarrow (0, \infty) \quad \text{wobei } g \text{ stetig und symmetrisch ist.}$$

Für $\vartheta \in \Theta$ ist folglich ϑ der Erwartungswert der Wahrscheinlichkeitsverteilung μ_ϑ . Dies nennt man ein Lokationsmodell, für das insbesondere die Regularitätsannahmen gelten. Seien X_1, \dots, X_n zufällige unabhängige Stichproben der Verteilung μ_ϑ . Für einen erwartungstreuen Schätzer $\hat{\vartheta}_n = T(X_1, \dots, X_n)$ gilt die Informationsungleichung, welche aus den Regularitätsannahmen folgt:

$$\text{Var}(\hat{\vartheta}_n) \geq \frac{1}{nI(\vartheta)} \quad \text{für alle } \vartheta \in \Theta.$$

Die Fisher-Information ist gegeben durch

$$I(\vartheta) = \mathbb{E}_\vartheta[\ell'(\vartheta, X)^2] = \int \frac{g'(x - \vartheta)^2}{g(x - \vartheta)} dx = \int \frac{g'(x)^2}{g(x)} dx \quad \text{für alle } \vartheta \in \Theta.$$

Eine große Fisher-Information bedeutet, dass die Daten viel Information über den Parameter ϑ enthalten, was zu einer kleinen Varianz des Schätzers führt. Die folgende Definition liefert wichtige Erkenntnisse, wie gut ein Schätzer in der Praxis funktioniert.

Definition 3.18 (Asymptotische Effizienz). Seien X_1, X_2, \dots unabhängige Stichproben der Verteilung μ_ϑ und $(\hat{\vartheta}_n)_{n \in \mathbb{N}}$ ein erwartungstreuer Schätzer. Der Schätzer $\hat{\vartheta}_n$ heißt *asymptotisch effizient*, falls

$$\text{Var}_\vartheta(\hat{\vartheta}_n) \sim \frac{1}{nI(\vartheta)} \quad \text{für alle } \vartheta.$$

Die Definition gibt uns ein Kriterium, um Schätzer zu vergleichen. Ein Schätzer, der asymptotisch effizient ist, erreicht die, durch die Informationsungleichung gegebene, untere Schranke im Grenzfall. Solche Schätzer sind in der Praxis besonders wünschenswert, weil sie die präzisesten Schätzungen liefern, die theoretisch möglich sind. Wir betrachten im Folgenden zwei asymptotisch effiziente Schätzer.

Beispiele.

- 1) Ein Beispiel für einen erwartungstreuen Schätzer ist das arithmetische Mittel $\hat{\vartheta}_n = \bar{X}_n$. Die Varianz des Mittelwerts ist gegeben durch

$$\text{Var}_\vartheta(\bar{X}_n) = \frac{1}{n} \text{Var}(X) = \frac{v}{n}, \quad \text{mit } v = \int x^2 g(x) dx.$$

Es gibt verschiedene Modelle, die die Fisher-Information und die Effizienz eines Schätzers illustrieren:

3. Relative Entropie, Information und statistische Unterscheidbarkeit

- (i) Für die Normalverteilung $N(\vartheta, \nu)$ ist die Fisher-Information $I(\vartheta) = \frac{1}{\nu}$ (siehe oben). Hierfür ist das arithmetische Mittel effizient.
- (ii) Bei der Doppelexponentialverteilung mit Dichtefunktion $f_{\vartheta}(x) = \frac{1}{2}e^{-|x-\vartheta|}$ gilt $\nu = 2$. In diesem Fall ist die Fisher-Information $I(\vartheta) = 1$, und somit $\nu > \frac{1}{I(\vartheta)}$. Das arithmetische Mittel ist nicht asymptotisch effizient, jedoch trotzdem akzeptabel.
- (iii) Für eine Verteilung mit schweren Rändern (heavy tails) ist die Funktion $g(x)$ proportional zu $\frac{1}{1+|x|^{\alpha}}$ mit $\alpha \in (2, 3)$. Wir erhalten eine Varianz von

$$\nu \propto \int \frac{x^2}{1+|x|^{\alpha}} dx = \infty, \quad I(\vartheta) < \infty.$$

In diesem Fall ist der Schätzer ganz schlecht.

- 2) Ein weiteres Beispiel für einen erwartungstreuen Schätzer ist der Stichprobenmedian

$$\hat{\vartheta}_n = M_n = \begin{cases} X_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} (X_{(n/2)} + X_{(n/2+1)}) & \text{falls } n \text{ gerade} \end{cases}.$$

Wir betrachten hierbei die geordneten Stichproben $X_{(i)} \leq X_{(j)}$ für $i \leq j$.

Für den Median liefert der folgende Satz eine konkrete Beschreibung seines asymptotischen Verhaltens.

Satz 3.19. Für den Stichprobenmedian gilt

$$\text{Var}_{\vartheta}(M_n) \sim \frac{1}{4g(0)^2 n} \quad \text{für } n \rightarrow \infty.$$

Konkreter gilt nach dem Zentralen Grenzwertsatz:

$$(M_n - q_{1/2}) \cdot \sqrt{n} \xrightarrow{\mathcal{D}} N\left(0, \frac{1}{4g(0)^2}\right) \quad \text{für alle } \vartheta.$$

Wie bereits oben für das arithmetische Mittel, betrachten wir im Folgenden die asymptotische Effizienz des Stichprobenmedians in verschiedenen Modellen.

Beispiele.

- (i) Für die Normalverteilung $N(\vartheta, 1)$ ist die Dichtefunktion an der Stelle 0 gegeben durch $g(0) = \frac{1}{\sqrt{2\pi}}$. Mit dem Satz ergibt sich das folgende asymptotische Verhalten der Varianz

$$\text{Var}_{\vartheta}(M_n) \sim \frac{\pi}{2n} > \frac{1}{n}.$$

Obwohl der Schätzer für die Normalverteilung nicht asymptotisch effizient ist, ist die Abweichung jedoch noch immer akzeptabel.

- (ii) Für die Doppelexponentialverteilung mit Parameter ϑ erhalten wir $g(0) = \frac{1}{2}$. Dies führt zur Varianz des Stichprobenmedians

$$\text{Var}_{\vartheta}(M_n) \sim \frac{1}{n} = \frac{1}{nI(\vartheta)}.$$

In diesem Fall ist der Stichprobenmedian asymptotisch effizient.

- (iii) Bei Verteilungen mit heavy tails verhält sich die Varianz wie $\text{Var}_{\vartheta}(M_n) \sim \frac{c\alpha}{n}$. Dieses Verhalten ist zwar immer noch nicht ideal, dennoch bleibt der Median aufgrund seiner Robustheit der bevorzugte Schätzer.

Beweis (Skizze). Der Einfachheit halber nehmen wir an, dass n ungerade ist. Sei $c \in \mathbb{R}$, dann gilt

$$\begin{aligned} \mathbb{P}_\vartheta [\sqrt{v}(M_n - q_{1/2}) \leq c] &= \mathbb{P}_\vartheta \left[M_n \leq q_{1/2} + \frac{c}{\sqrt{n}} \right] = \mathbb{P}_\vartheta \left[X_i \leq q_{1/2} + \frac{c}{\sqrt{n}} \text{ mindestens } \frac{n+1}{2} \text{ mal} \right] \\ &= \mathbb{P}_\vartheta \left[H \geq \frac{n+1}{2} \right] \end{aligned}$$

Wobei $H \sim \text{Bin}(n, p)$ und $p = F\left(q_{1/2} + \frac{c}{\sqrt{n}}\right) = F(q_{1/2}) + F'(q_{1/2})\frac{c}{\sqrt{n}} + O\left(\frac{1}{n}\right)$. Sei Z eine standardnormalverteilte Zufallsvariable, dann gilt nach dem zentralen Grenzwertsatz

$$\begin{aligned} \mathbb{P} \left[\frac{H - np}{\sqrt{np(1-p)}} \geq \frac{n(1/2 - p) + 1/2}{\sqrt{n}\sqrt{p(1-p)}} \right] &\approx \mathbb{P} \left[Z \geq \frac{\sqrt{n}\left(\frac{1}{2} - p\right) + \frac{1}{\sqrt{n}}}{\sqrt{p(1-p)}} \right] \\ &\approx \mathbb{P} [Z \geq -2g(0)c] = \mathbb{P} \left[\frac{Z}{2g(0)} \right] \end{aligned}$$

Nun ist $\frac{Z}{2g(0)}$ eine Zufallsvariable mit Verteilung $N\left(0, \frac{1}{4g(0)^2}\right)$. ■

Asymptotische Effizienz von Maximum-Likelihood-Schätzern

Wir betrachten nun ein reguläres Produktmodell mit Dichte f_ϑ , $\vartheta \in \Theta \subseteq \mathbb{R}$. Die Log-Likelihood ℓ_n für n unabhängigen Stichproben ist dann gegeben durch

$$\ell_n(\vartheta; x_1, \dots, x_n) = \sum_{i=1}^n \log f_\vartheta(x_i).$$

Ist $\hat{\vartheta}_n = T_n(X_1, \dots, X_n)$ ein Maximum-Likelihood-Schätzer, dann gilt die *Likelihood-Gleichung*

$$\frac{\partial \ell_n}{\partial \vartheta}(T_n(x_1, \dots, x_n); x_1, \dots, x_n) = 0 \quad (3.16)$$

für alle Beobachtungswerte x_1, \dots, x_n . Mithilfe der relativen Entropie konnten wir bereits zeigen, dass Maximum-Likelihood-Schätzer unter Regularitätsannahmen konsistent sind. Mithilfe der Fisher-Information

$$I(\vartheta) = \int \left| \frac{\partial}{\partial \vartheta} \log f_\vartheta(x) \right|^2 f_\vartheta(x) dx$$

kann man nun die asymptotische Varianz identifizieren, und einen zentralen Grenzwertsatz für Maximum-Likelihood-Schätzer beweisen.

Satz 3.20 (Fisher, Wilks, Wald). Sei $T_n(X_1, \dots, X_n)$ ($n \in \mathbb{N}$) eine konsistente Folge von Schätzern, die die Likelihood-Gleichung (3.16) erfüllt. Dann gilt unter Regularitätsvoraussetzungen für jeden Parameterwert $\vartheta \in \Theta$

$$\sqrt{n} \left(\hat{\vartheta}_n(X_1, \dots, X_n) - \vartheta \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{I(\vartheta)}\right),$$

wobei „ $\xrightarrow{\mathcal{D}}$ “ für Konvergenz in Verteilung bezüglich P_ϑ für $n \rightarrow \infty$ steht.

Da andererseits nach der Informationsungleichung die Varianz eines erwartungstreuen Schätzers für ϑ basierend auf n unabhängigen Stichproben unter Regularitätsbedingungen stets größer als $\frac{1}{nI(\vartheta)}$ ist, folgt, dass Maximum-Likelihood-Schätzer *asymptotisch effizient* sind.

3. Relative Entropie, Information und statistische Unterscheidbarkeit

Beweis. Wir skizzieren hier nur die Beweisidee. Ein vollständiger Beweis erfordert eine sorgfältige Rechtfertigung der einzelnen Schritte inklusive der Kontrolle der Restterme in den folgenden Approximationen mithilfe der Regularitätsannahmen.

Nach (3.16) und einer Taylor-Approximation gilt

$$0 = \ell'_n(T_n) \approx \ell'_n(\vartheta) + (T_n - \vartheta) \ell''_n(\vartheta),$$

und damit

$$\sqrt{n}(T_n - \vartheta) \approx \frac{n^{-1/2} \ell'_n(\vartheta)}{-\frac{1}{n} \ell''_n(\vartheta)}.$$

Wir betrachten nun den Zähler und Nenner separat. Für den Zähler erhalten wir unter P_ϑ nach dem zentralen Grenzwertsatz für Summen von unabhängigen identisch verteilten Zufallsvariablen

$$\frac{1}{\sqrt{n}} \ell'_n(\vartheta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \vartheta} \log f_\vartheta(X_i) \xrightarrow{\mathcal{D}} N(0, I(\vartheta)).$$

Hierbei haben wir benutzt, dass die Summanden nach (3.11) und der Definition der Fisher-Information unabhängige zentrierte Zufallsvariablen mit Varianz $I(\vartheta)$ sind. Für den Nenner erhalten wir entsprechend nach dem Gesetz der großen Zahlen und (3.11)

$$\frac{1}{\sqrt{n}} \ell''_n(\vartheta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n -\frac{\partial^2}{\partial \vartheta^2} \log f_\vartheta(X_i) \xrightarrow{\text{f.s.}} I(\vartheta).$$

Insgesamt folgt damit nach dem Satz von Slutsky, dass

$$\sqrt{n}(T_n - \vartheta) \xrightarrow{\mathcal{D}} W/I(\vartheta),$$

wobei W eine Zufallsvariable mit Verteilung $\mathcal{N}(0, \vartheta)$ ist, also $W/I(\vartheta) \sim \mathcal{N}(0, 1/I(\vartheta))$. ■

Der Satz von Fisher, Wilks und Wald ermöglicht es, approximative Konfidenzintervalle basierend auf Maximum-Likelihood-Schätzern anzugeben. Für „große“ n gilt nach dem Satz näherungsweise

$$\frac{\hat{\vartheta}_n - \vartheta}{\sqrt{\frac{1}{nI(\vartheta)}}} \approx \mathcal{N}(0, 1).$$

Nach dem Satz von Slutsky folgt auch

$$\frac{\hat{\vartheta}_n - \vartheta}{\sqrt{\frac{1}{nI(\hat{\vartheta}_n)}}} \approx \mathcal{N}(0, 1).$$

Also ist

$$\hat{\vartheta}_n \pm \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{1}{nI(\hat{\vartheta}_n)}}$$

ein *approximatives Konfidenzintervall* für ϑ zum Konfidenzniveau $1 - \alpha$. Beispielsweise ist $\hat{\vartheta}_n \pm 2/\sqrt{nI(\hat{\vartheta}_n)}$ ein approximatives 95%-Konfidenzintervall. Allerdings ist unklar, wie groß man n wählen muss, damit die Approximation hinreichend genau ist.

3.4. Weitere Anwendungen von Entropie und relativer Entropie

Entropie und Kodierung

Als Spezialfall der Aussagen (1) und (2) in Satz 3.12 erhalten wir, wenn S endlich und ν die Gleichverteilung ist, zwei bekannte Aussagen aus der Informationstheorie: die „asymptotische Gleichverteilungseigenschaft“ und den Quellenkodierungssatz von Shannon:

Wir betrachten die *möglichst effiziente Beschreibung/Kodierung einer Zufallsfolge*. Eine unbekannte Signalfolge mit Werten in einer endlichen Menge S (dem zugrundeliegenden „Alphabet“) beschreibt man im einfachsten A-Priori-Modell durch unabhängige Zufallsvariablen X_1, X_2, \dots mit Verteilung μ , wobei $\mu(x)$ die relative Häufigkeit des Buchstabens x in der verwendeten Sprache ist. Eine „perfekte“ Kodierung ordnet jedem Wort mit einer vorgegebenen Anzahl n von Buchstaben, also jedem Element des Produktraums S^n , eine Binärfolge zu. Will man alle Wörter mit n Buchstaben perfekt kodieren, werden $n \cdot \log_2 |S|$ Bits benötigt. Wir betrachten stattdessen „effiziente“ Kodierungen, die nur den „meisten“ Wörtern mit n Buchstaben eindeutig eine Binärfolge zuordnen.

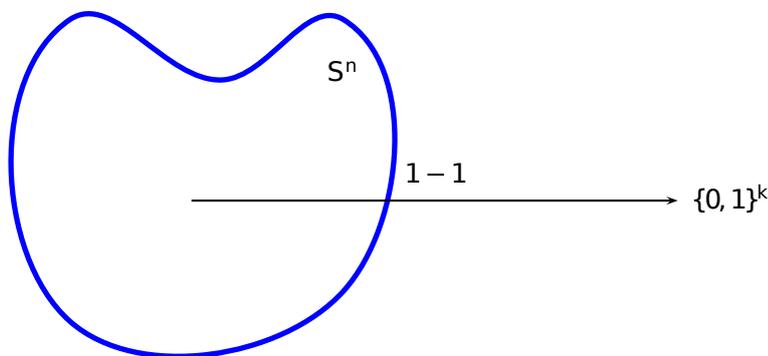


Abbildung 3.2.: Perfekte Kodierung

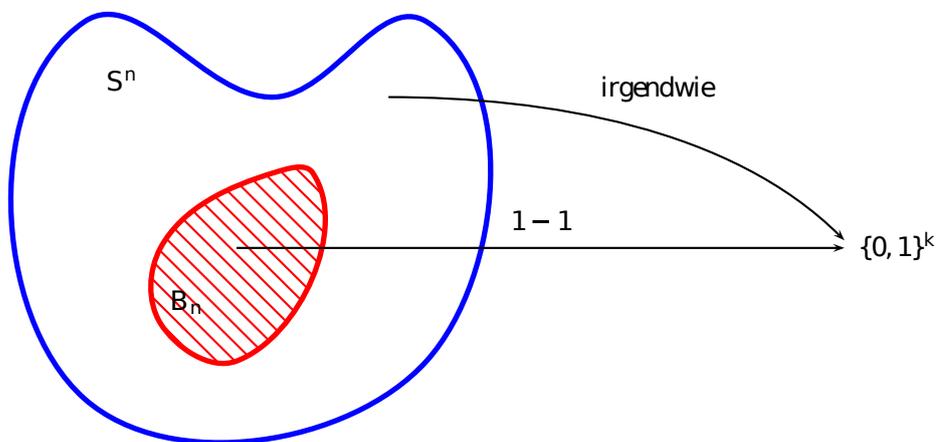


Abbildung 3.3.: Effiziente Kodierung bzgl. einer Folge von wesentlichen Mengen B_n .

Sei $L(\mu; x_1, \dots, x_n) = \prod_{i=1}^n \mu(x_i)$ die A-Priori-Wahrscheinlichkeit von (x_1, \dots, x_n) unter dem Produktmaß μ^n . Wählen wir für ν die Gleichverteilung auf S , dann gilt

$$|A_n| = \nu^n[A_n] \cdot |S|^n \quad \text{für alle Mengen } A_n \subseteq S^n, n \in \mathbb{N}.$$

Damit können wir die Aussage von Satz 3.12 (1) folgendermaßen umformulieren:

Korollar 3.21 (Asymptotische Gleichverteilungseigenschaft). Für jedes $\varepsilon > 0$ ist die Folge

$$B_{n,\varepsilon} := \{(x_1, \dots, x_n) \in S^n : e^{-n(H(\mu)+\varepsilon)} \leq L(\mu; x_1, \dots, x_n) \leq e^{-n(H(\mu)-\varepsilon)}\}, \quad n \in \mathbb{N},$$

wesentlich bzgl. μ , und es gilt

$$|B_{n,\varepsilon}| \leq e^{n(H(\mu)+\varepsilon)} \quad \text{für alle } n \in \mathbb{N}.$$

Beweis. Die Aussage folgt aus Satz 3.12 (1) wegen $H(\mu|\nu) = \log |S| - H(\mu)$ und $d\mu^n/d\nu^n = |S|^n \cdot p_n$. ■

Die asymptotische Gleichverteilungseigenschaft zeigt, dass Folgen von wesentlichen Mengen existieren, deren Mächtigkeit auf der exponentiellen Skala nicht viel schneller als $\exp(n \cdot H(\mu))$ wächst.

Wieviele Elemente enthalten wesentliche Mengen mindestens? Für $p \in (0, 1)$ sei

$$K(n, p) = \inf \{|A_n| : A_n \subseteq S^n \text{ mit } P[(X_1, \dots, X_n) \in A_n] \geq p\}$$

die mindestens benötigte Anzahl von Wörtern, um den Text (X_1, \dots, X_n) mit Wahrscheinlichkeit $\geq p$ korrekt zu erfassen. Dann ist $\log_2 K(n, p)$ die für eine korrekte binäre Kodierung von (X_1, \dots, X_n) mit Wahrscheinlichkeit $\geq p$ mindestens benötigte Anzahl von Bits. Aus dem zweiten Teil von Satz 3.12 ergibt sich:

Korollar 3.22 (Quellenkodierungssatz von Shannon). Für alle $p \in (0, 1)$ gilt:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) &= H(\mu), \quad \text{bzw.} \\ \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 K(n, p) &= H_2(\mu) := - \sum_{x:\mu(x) \neq 0} \mu(x) \log_2 \mu(x). \end{aligned}$$

Insbesondere gilt: Ist A_n ($n \in \mathbb{N}$) wesentlich bzgl. μ , so ist $|A_n| \geq \exp(nH(\mu))$.

Die Größe $\frac{1}{n} \log_2 K(n, p)$ kann als die für eine mit Wahrscheinlichkeit $\geq p$ korrekte Kodierung benötigte Zahl von Bits pro gesendetem Buchstaben interpretiert werden.

Bemerkung. Der Quellenkodierungssatz zeigt, dass es keine Folge von wesentlichen Mengen gibt, die auf der exponentiellen Skala deutlich langsamer wächst als die in Korollar 3.21 angegebenen Folgen $B_{n,\varepsilon}$ ($n \in \mathbb{N}$).

Beweis. Die Aussage ergibt sich wieder als Spezialfall von Satz 3.12 wenn wir $\nu = \text{Unif}_S$ setzen:

Oberer Schranke: $\limsup \frac{1}{n} \log K(n, p) \leq H(\mu) :$

Sei $\varepsilon > 0$. Nach Korollar 3.21 erfüllt die dort konstruierte Folge $B_{n,\varepsilon}$ ($n \in \mathbb{N}$) die Bedingung

$$\lim_{n \rightarrow \infty} P[(X_1, \dots, X_n) \in B_{n,\varepsilon}] = 1 > p, \quad \blacksquare$$

und es gilt $\frac{1}{n} \log |B_{n,\varepsilon}| \leq H(\mu) + \varepsilon$. Damit folgt

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log |B_{n,\varepsilon}| \leq H(\mu) + \varepsilon.$$

Die Behauptung ergibt sich für $\varepsilon \rightarrow 0$.

Untere Schranke: $\liminf \frac{1}{n} \log K(n, p) \geq H(\mu)$:

Für Mengen $A_n \subseteq S^n$ mit $P[(X_1, \dots, X_n) \in A_n] \geq p$ gilt nach Satz 3.12 (2):

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log |A_n| = \liminf_{n \rightarrow \infty} \frac{1}{n} \log (\nu^n(A_n) \cdot S^n) \geq \log |S| - H(\mu|\nu) = H(\mu).$$

Entropie von Markovketten

Sei $p(x, y)$ ($x, y \in S$) eine stochastische Matrix auf einer endlichen Menge S mit Gleichgewichtsverteilung $\nu \in WV(S)$, d.h. für alle $y \in S$ gelte

$$\sum_{x \in S} \nu(x) p(x, y) = \nu(y). \tag{3.17}$$

Der folgende wichtige Satz zeigt, dass die relative Entropie $H(\mu p^n | \nu)$ der Verteilung zur Zeit n einer Markovkette mit Startverteilung μ und Übergangsmatrix p bezüglich des Gleichgewichts ν monoton fällt:

Satz 3.23 (Abfall der relativen Entropie). Ist p eine stochastische Matrix auf S und ν ein Gleichgewicht von p , dann gilt für jede Wahrscheinlichkeitsverteilung μ auf S :

$$H(\mu p | \nu) \leq H(\mu | \nu). \tag{3.18}$$

Insbesondere ist $n \mapsto H(\mu p^n | \nu)$ stets monoton fallend.

Beweis. Ist μ nicht absolutstetig bezüglich ν , dann ist die Aussage (3.18) automatisch erfüllt. Andernfalls sei w eine Version der relativen Dichte $d\mu/d\nu$. Dann gilt

$$(\mu p)(y) = \sum_{x \in S} \mu(x) p(x, y) = \sum_{x \in S} w(x) \nu(x) p(x, y). \tag{3.19}$$

Aus der Gleichgewichtsbedingung (3.17) folgt $\nu(x)p(x, y) \leq \nu(y)$ für alle $x, y \in S$. Also ist auch μp absolutstetig bzgl. ν , mit relativer Dichte

$$\frac{(\mu p)(y)}{\nu(y)} = \sum_{x \in S} w(x) \frac{\nu(x) p(x, y)}{\nu(y)}. \tag{3.20}$$

Aus der Gleichgewichtsbedingung folgt auch, dass $x \mapsto \nu(x)p(x, y)/\nu(y)$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf S ist. Darum können wir die Jensensche Ungleichung auf die konvexe Funktion $u(x) = x \log_+ x$ anwenden, und erhalten

$$u \left(\sum_{x \in S} w(x) \frac{\nu(x) p(x, y)}{\nu(y)} \right) \leq \sum_{x \in S} u(w(x)) \frac{\nu(x) p(x, y)}{\nu(y)}.$$

Zusammen mit (3.20) ergibt sich

$$\begin{aligned} H(\mu p | \nu) &= \sum_{y: \nu(y) \neq 0} u \left(\sum_{x \in S} w(x) \nu(x) p(x, y) / \nu(y) \right) \nu(y) \\ &\leq \sum_{y \in S} \sum_{x \in S} u(w(x)) \nu(x) p(x, y) \\ &= \sum_{x \in S} u(w(x)) \nu(x) = H(\mu | \nu). \end{aligned} \quad \blacksquare$$

Bemerkung (Zunahme der Entropie; thermodynamische Irreversibilität). Als Spezialfall ergibt sich wegen $H(\mu) = \log |S| - H(\mu|\nu)$ die Aussage, dass die Entropie $H(\mu p^n)$ der Verteilung einer Markovkette zur Zeit n monoton wächst, falls der Zustandsraum endlich und die Gleichverteilung ν ein Gleichgewicht ist. In der Interpretation der statistischen Physik geht die zeitliche Entwicklung auf makroskopischer Ebene (Thermodynamik) von einem geordneten hin zu einem ungeordneten Zustand mit (lokal) maximaler Entropie (»thermodynamische Irreversibilität«). Trotzdem ist auf mikroskopischer Ebene die Dynamik rekurrent, d.h. jeder Zustand $x \in S$ wird von der Markovkette mit Wahrscheinlichkeit 1 unendlich oft besucht – dies dauert nur eventuell astronomisch lange. Die Einführung eines Markov-Modells durch die österreichischen Physiker Tatjana und Paul Ehrenfest konnte eine entsprechende Kontroverse von Zermelo („Dynamik kehrt immer wieder zurück“) und Boltzmann („soll solange warten“) lösen.

Beispiel (Irrfahrten). Ist p die Übergangsmatrix eines symmetrischen Random Walks auf dem diskreten Kreis $\mathbb{Z}_k = \mathbb{Z}/(k\mathbb{Z})$, der symmetrischen Gruppe S_n („Mischen eines Kartenspiels“), oder dem diskreten Hyperwürfel $\{0, 1\}^n$ („Ehrenfest-Modell“), dann ist die Gleichverteilung ein Gleichgewicht, und die Entropie wächst monoton.

4. Empirische Verteilungen

4.1. Empirische Verteilung

Sei μ eine unbekannte Wahrscheinlichkeitsverteilung auf S mit einem messbaren Raum (S, \mathcal{B}) . Wir arbeiten abermals mit einer endlichen Stichprobe X_1, \dots, X_n von Zufallsvariablen, welche unter \mathbb{P}_μ unabhängig und identisch zu μ verteilt sind. Wir betrachten die *empirische Verteilung* der Stichproben, gegeben durch:

$$L_n(\omega) := \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)} \qquad L_n(B) = \frac{1}{n} \sum_{i=1}^n \overbrace{1_{\{X_i \in B\}}}^{H_n(B)}$$

welche die relative Häufigkeit von Werten der Stichproben in der Menge B angibt. Hierbei definieren wir wie angezeigt die absoluten Häufigkeiten durch $H_n(B)$. Für eine Abbildung $f : S \rightarrow \mathbb{R}$ erhalten wir somit den empirischen Mittelwert von f :

$$\int f dL_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

Die empirische Verteilung L_n ist eine *zufällige Wahrscheinlichkeitsverteilung*, das heißt, L_n ist eine Zufallsvariable mit Werten im Raum der Wahrscheinlichkeitsverteilungen auf S .

Die empirische Verteilung ist gegeben durch $L_n = \frac{1}{n} H_n$. Um also die Verteilung der empirischen Verteilung zu bestimmen, betrachten wir H_n . Für eine feste Menge $B \in \mathcal{B}$ gilt:

$$H_n(B) \sim \text{Bin}(n, \mu(B))$$

Allgemeiner gilt für jede disjunkte Zerlegung $S = B_1 \cup B_2 \cup \dots \cup B_k$ mit $B_i \in \mathcal{B}$:

$$(H_n(B_1), H_n(B_2), \dots, H_n(B_k)) \sim \text{Mult}(n, \mu(B_1), \dots, \mu(B_k))$$

Somit erhalten wir für $h \in \mathbb{N}^k$ mit $\sum h_l = n$, dass gilt

$$\mathbb{P}_\mu[H_n(B_l) = h_l \quad \forall l = 1, \dots, k] = \frac{n!}{h_1! h_2! \dots h_k!} \prod_{l=1}^k \mu(B_l)^{h_l}$$

Hierdurch ist die Verteilung der maßwertigen Zufallsvariablen H_n und $L_n = \frac{1}{n} H_n$ eindeutig festgelegt.

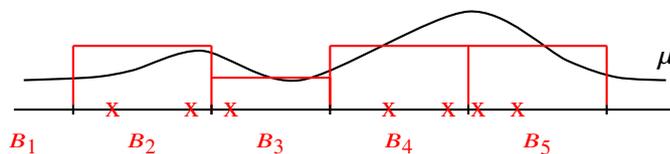


Abbildung 4.1.: Histogramm der Bins B_1, \dots, B_5

Die empirische Verteilung können wir als Schätzer für die unbekannte Wahrscheinlichkeitsverteilung μ verwenden:

$$\hat{\mu}_n := L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

4. Empirische Verteilungen

Wegen $L_n(B) = \frac{1}{n} \sum_{i=1}^n 1_B(X_i)$ gilt für $B \in \mathcal{B}$:

$$\begin{aligned} \mathbb{E}_\mu[L_n(B)] &= \mu(B) && \text{erwartungstreu} \\ \text{Var}_\mu[L_n(B)] &= \frac{\mu(B)(1-\mu(B))}{n} \leq \frac{1}{4n} \\ \mathbb{P}_\mu[|L_n(B) - \mu(B)| \geq \varepsilon] &\leq 2e^{-2\varepsilon^2 n} && \text{für alle } \varepsilon > 0 \end{aligned}$$

Die Abschätzung der Varianz folgt nach der Bernstein-Ungleichung, siehe Stochastik bzw. Algorithmische Mathematik 2.

Gesetz der großen Zahlen

Aus dem Gesetz der großen Zahlen folgt die fast sichere Konvergenz der relativen Häufigkeiten gegen die unbekannte Wahrscheinlichkeit einer Menge, das heißt:

$$L_n(B) \xrightarrow{n \uparrow \infty} \mu(B) \quad \mathbb{P}_\mu\text{-fast sicher für alle } B \in \mathcal{B}$$

Eine wichtige Frage ist, ob man die Annahmемenge unabhängig von B wählen kann. Im Allgemeinen gilt die Aussage jedoch nicht. Wenn die Anzahl der Teilmengen von S endlich ist, kann man die Annahmемenge unabhängig wählen, da es nur endlich viele Teilmengen gibt, für die die Konvergenz zu überprüfen ist. Aus dem Gesetz der großen Zahlen folgt insbesondere, dass die Erwartungswerte einer integrierbaren Funktion $f \in \mathcal{L}^1(\mu)$ bezüglich der zufälligen Maße $L_n(B)$ für fast alle $B \in \mathcal{B}$ gegen die Erwartungswerte bzgl. μ konvergieren:

$$\int f dL_n(B) = \frac{1}{n} \sum_{i=1}^n f(X_i(B)) \longrightarrow \int f d\mu \quad \text{für } \mathbb{P}_\mu\text{-fast alle } B \in \mathcal{B}. \quad (4.1)$$

Die Ausnahmemenge kann dabei aber von der Funktion f abhängen. Da der Raum der stetigen beschränkten Funktionen $F : S \rightarrow \mathbb{R}$ im Allgemeinen nicht mehr separabel ist, ist die folgende Erweiterung des Gesetzes der großen Zahlen nicht offensichtlich:

Satz 4.1 (GGZ für empirische Verteilungen; Varadarajan). Ist S separabel und $\mathcal{B} = \mathcal{B}(S)$ die Borelsche σ -Algebra, dann gilt:

$$\mathbb{P}_\mu[L_n \xrightarrow{w} \mu] = 1$$

Eine Menge ist separabel, falls es eine abzählbare Teilmenge gibt, welche dicht in der Menge selbst ist.

Bemerkung. ERINNERUNG: SCHWACHE KONVERGENZ VON WAHRSCHEINLICHKEITSVERTEILUNGEN Nach dem Portmanteau-Theorem (siehe Einführung in die Wahrscheinlichkeitstheorie) gelten die folgenden Äquivalenzen:

$$\begin{aligned} L_n \xrightarrow{w} \mu &\Leftrightarrow \int f dL_n = \frac{1}{n} \sum_{i=1}^n f(X_i) \rightarrow \int f d\mu && \text{für alle } f \in C_b(S) \\ &\Leftrightarrow L_n(B) \rightarrow \mu(B) && \text{für alle } B \text{ mit } \mu(\partial B) = 0 \end{aligned}$$

Beweis. Nach dem Portmanteau-Theorem genügt es zu zeigen, dass \mathbb{P}_μ -fast sicher gilt:

$$\int f dL_n \rightarrow \int f d\mu \quad \text{für alle } f \in C_{b,L}(S) \quad (*)$$

wobei $C_{b,L}(S)$ die beschränkten lipschitzstetigen Funktionen enthält, siehe Einführung in die Wahrscheinlichkeitstheorie. Man kann beweisen, dass im Raum $C_{b,L}(S)$ eine bezüglich der Supremums-Norm dichte, abzählbare Teilmenge A existiert, wenn S separabel ist, siehe z.B. [Dudley:Real Analysis and Probability]. Nach dem Gesetz der großen Zahlen gilt \mathbb{P}_μ -fast sicher:

$$\int f dL_n \rightarrow \int f d\mu \text{ für alle } f \in A$$

denn die Vereinigung von abzählbar vielen Nullmengen ist wieder eine Nullmenge. Sei nun $f \in C_{b,L}(S)$ und $\varepsilon > 0$. Dann existiert $\tilde{f} \in A$ mit $\|f - \tilde{f}\|_{\text{sup}} \leq \varepsilon/2$. Damit folgt:

$$\left| \int f dL_n - \int f d\mu \right| \leq \underbrace{\left| \int f dL_n - \int \tilde{f} dL_n \right|}_{\leq \|f - \tilde{f}\|_{\text{sup}} \leq \frac{\varepsilon}{2}} + \underbrace{\left| \int \tilde{f} dL_n - \int \tilde{f} d\mu \right|}_{\rightarrow 0} + \underbrace{\left| \int \tilde{f} d\mu - \int f d\mu \right|}_{\leq \frac{\varepsilon}{2}}$$

also $\limsup_{n \rightarrow \infty} \left| \int f dL_n - \int f d\mu \right| \leq \varepsilon$ \mathbb{P}_μ -fast sicher. Die Behauptung (*) folgt dann für $\varepsilon = 1/k$ gegen 0. ■

Bemerkung. Für $S = \mathbb{R}$ gilt eine stärkere, nicht asymptotische Abschätzung (Dvoretzky-Kiefer-Wolfowitz-Ungleichung):

$$\mathbb{P}_\mu \left[\sup_{c \in \mathbb{R}} |L_n((-\infty, c]) - \mu((-\infty, c])| \geq \varepsilon \right] \leq 2e^{-2\varepsilon^2 n}$$

für alle $n \in \mathbb{N}$ und $\varepsilon > 0$, siehe Abschnitt 4.3.

Große Abweichungen für empirische Verteilungen

Die Konvergenzgeschwindigkeit auf der exponentiellen Skala lässt sich durch ein Prinzip der großen Abweichungen auf dem Raum $\text{WV}(S)$ der Wahrscheinlichkeitsverteilungen auf (S, \mathcal{B}) mit der Topologie der schwachen Konvergenz beschreiben:

Satz 4.2 (Sanov). Die empirischen Verteilungen $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ erfüllen das folgende Prinzip der großen Abweichungen:

(i) *Obere Schranke:* Für jede abgeschlossene Menge $\mathcal{A} \subseteq \text{WV}(S)$ gilt:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P_\nu[L_n \in \mathcal{A}] \leq - \inf_{\mu \in \mathcal{A}} H(\mu | \nu).$$

(ii) *Untere Schranke:* Für jede offene Menge $\mathcal{O} \subseteq \text{WV}(S)$ gilt:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_\nu[L_n \in \mathcal{O}] \geq - \inf_{\mu \in \mathcal{O}} H(\mu | \nu).$$

Beweis. (ii) Zum Beweis der unteren Schranke wechseln wir wieder das zugrundeliegende Maß, und wenden Satz 3.12 an. Sei $\mathcal{O} \subseteq \text{WV}(S)$ offen und $\mu \in \mathcal{O}$. Nach (??) ist dann die Folge

$$A_n = \left\{ (x_1, \dots, x_n) \in S^n : \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \in \mathcal{O} \right\}$$

4. Empirische Verteilungen

wesentlich bzgl. μ , denn

$$\mu^n[A_n] = P_\mu[L_n \in O] \longrightarrow 1$$

für $n \rightarrow \infty$. Daher folgt nach Korollar 3.12(2):

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_\nu[L_n \in O] = \liminf_{n \rightarrow \infty} \frac{1}{n} \log \nu^n[A_n] \geq -H(\mu | \nu).$$

Die Behauptung ergibt sich, da dies für alle $\mu \in O$ gilt.

- (i) Die obere Schranke beweisen wir hier nur für endliche Zustandsräume S , s. z.B. [Dembo, Zeitouni: Large Deviations] für den Beweis im allgemeinen Fall. Ist S endlich, und μ eine bzgl. ν absolutstetige Wahrscheinlichkeitsverteilung mit Dichte $w = d\mu/d\nu$, dann gilt für alle $(x_1, \dots, x_n) \in S^n$ mit empirischer Verteilung $\frac{1}{n} \sum_{i=1}^n \delta_{x_i} = \mu$:

$$\begin{aligned} \frac{d\mu^n}{d\nu^n}(x_1, \dots, x_n) &= \prod_{i=1}^n \frac{d\mu}{d\nu}(x_i) = \exp\left(\sum_{i=1}^n \log\left(\frac{d\mu}{d\nu}(x_i)\right)\right) \\ &= \exp\left(n \int \log\left(\frac{d\mu}{d\nu}\right) d\mu\right) = \exp(nH(\mu | \nu)). \end{aligned}$$

Damit folgt

$$\begin{aligned} P_\nu[L_n = \mu] &= \nu^n \left[\left\{ (x_1, \dots, x_n) : \frac{1}{n} \sum_{i=1}^n \delta_{x_i} = \mu \right\} \right] \\ &= e^{-nH(\mu | \nu)} \cdot \mu^n \left[\left\{ (x_1, \dots, x_n) : \frac{1}{n} \sum_{i=1}^n \delta_{x_i} = \mu \right\} \right] \\ &\leq e^{-nH(\mu | \nu)}. \end{aligned} \tag{4.2}$$

Jeder empirischen Verteilung von n Elementen $x_1, \dots, x_n \in S$ entspricht ein Histogramm $\vec{h} = (h_a)_{a \in S} \in \{0, 1, \dots, n\}^S$. Für die Anzahl der möglichen empirischen Verteilungen gilt daher

$$\left| \left\{ \frac{1}{n} \sum_{i=1}^n \delta_{x_i} : (x_1, \dots, x_n) \in S^n \right\} \right| \leq (n+1)^{|S|}.$$

Nach (4.2) erhalten wir nun für eine beliebige Menge $\mathcal{A} \subseteq \text{WV}(S)$ die (nicht-asymptotische) Abschätzung

$$P_\nu[L_n \in \mathcal{A}] = \sum_{\mu \in \mathcal{A}} P_\nu[L_n = \mu] \leq (n+1)^{|S|} \cdot \exp\left(-n \inf_{\mu \in \mathcal{A}} H(\mu | \nu)\right),$$

aus der die asymptotische obere Schranke wegen $|S| < \infty$ folgt. ■

Bemerkung. Wie der Beweis schon andeutet, gilt auch die obere Schranke in diesem Fall nur noch asymptotisch und modulo subexponentiell wachsender Faktoren. Der Übergang von endlichen zu allgemeinen Zustandsräumen ist bei der oberen Schranke nicht trivial, s. [Dembo, Zeitouni: Large deviations].

Den Satz von Sanov bezeichnet man gelegentlich auch als ein „Prinzip der großen Abweichungen auf Level II“, d.h. für die empirischen Verteilungen. Wir bemerken abschließend, dass sich eine Version des Satzes von Cramér, d.h. ein „Prinzip der großen Abweichungen auf Level I“ als Spezialfall ergibt:

Für eine stetige beschränkte Funktion $f : S \rightarrow \mathbb{R}$ und eine offene Menge $B \subseteq \mathbb{R}$ gilt nach dem Satz von Sanov:

$$P_\nu \left[\frac{1}{n} \sum_{i=1}^n f(X_i) \in B \right] = P_\nu[L_n \in O] \geq \exp\left(-\inf_{\mu \in O} H(\mu | \nu)\right)$$

mit $O = \{\mu \in \text{WV}(S) : \int f d\mu \in B\}$. Entsprechend ergibt sich eine analoge obere Schranke, falls B abgeschlossen ist.

4.2. Plug-in-Schätzer und Bootstrap

Das folgende Kapitel basiert zum großen Teil auf den Kapiteln 7 und 8 des Buches "All of Statistics" von L. Wasserman.

Definition 4.3. 1) Ein *statistisches Funktional* ist eine Abbildung

$$g : \mathcal{P} \rightarrow \mathbb{R}, \mu \mapsto g(\mu)$$

wobei \mathcal{P} eine Teilmenge der Menge $WV(S)$ der Wahrscheinlichkeitsverteilungen auf $(S, \mathcal{B}(S))$ ist.

2) Die Statistik $\hat{g}_n := g(L_n)$ heißt *Plug-in-Schätzer* für $g(\mu)$.

Wir haben bereits einige statistische Funktionale bzw. Plug-in-Schätzer gesehen. Im Folgenden werden wir noch einige weitere betrachten.

Beispiele. LINEARE FUNKTIONALE

a) Wir sehen, dass das Auswerten an einer Menge $B \in \mathcal{B}$ ein lineares Funktional liefert:

$$g(\mu) = \mu(B) \quad : \quad \hat{g}_n = L_n(B) = \frac{1}{n} H_n(B)$$

Wie oben ist hierbei H_n die Häufigkeit der Werte von X_i in B .

b) Analog gibt der Erwartungswert einer Funktion f unter μ ein lineares Funktional:

$$g(\mu) = \int f d\mu \quad : \quad \hat{g}_n = \int f dL_n = \frac{1}{n} \sum_1^n f(X_i)$$

NUMERISCHE MERKMALE In den folgenden Fällen sei $S = \mathbb{R}$.

a) Der bereits bekannte Mittelwert \bar{X}_n ist auch Plug-in-Schätzer

$$m(\mu) = \int x d\mu \quad : \quad \hat{m}_n = \bar{X}_n$$

b) Auch die Standardabweichung können wir als Funktional betrachten:

$$\sigma(\mu) = \sqrt{\int (x - m(\mu))^2 d\mu} \quad : \quad \hat{\sigma}_n = \sqrt{\frac{1}{n} \sum_1^n (X_i - \bar{X}_n)^2}$$

c) Das statistische Funktional $K(\mu)$ heißt *Schiefte* der Verteilung μ . Die Schiefe ist ein Maß für die Asymmetrie der Verteilung. Sie ist gegeben durch:

$$K(\mu) = \left(\int (x - m(\mu))^3 d\mu \right) / \sigma(\mu)^3 \quad : \quad \hat{K}_n = \frac{1}{n} \left(\sum_1^n (X_i - \bar{X}_n)^3 \right) / \hat{\sigma}_n^3$$

d) Wir können auch die Verteilungsfunktion, ausgewertet an einem Wert $c \in \mathbb{R}$, als statistisches Funktional betrachten:

$$F_\mu(c) = \mu((-\infty, c]) \quad : \quad \hat{F}_n(c) = L_n((-\infty, c]) = |\{i \leq n : X_i \leq c\}|/n$$

\hat{F}_n heißt *empirische Verteilungsfunktion*. Die empirische Verteilungsfunktion ist stückweise konstant mit Sprüngen an den Stellen X_1, \dots, X_n .

4. Empirische Verteilungen

e) Auch die verallgemeinerten Inversen können wir als statistische Funktionale betrachten.

$$\underline{q}_\alpha(\mu) = \underline{G}_\mu(\alpha) \quad : \quad \underline{q}_{\alpha,n} = \underline{G}_{L_n}(\alpha) = \inf\{x : \hat{F}_n(x) \geq \alpha\}$$

Sind $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ die Ordnungsstatistiken, also die der Reihe nach geordneten Werte X_1, \dots, X_n , dann gilt:

$$\underline{q}_{\alpha,n} = X_{(\lceil n\alpha \rceil)} \quad \text{das untere Stichprobenquantil}$$

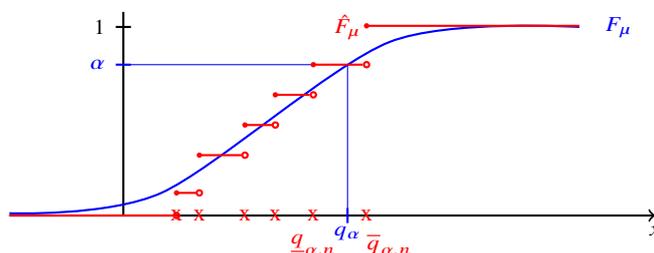
Entsprechend können wir die oberen Quantile betrachten:

$$\bar{q}_\alpha(\mu) = \bar{G}_\mu(\alpha) \quad : \quad \hat{\bar{q}}_{\alpha,n} = \bar{G}_{L_n}(\alpha) = \sup\{x : \hat{F}_n(x) \leq \alpha\}, \quad \hat{\bar{q}}_{\alpha,n} = X_{(\lfloor n\alpha + 1 \rfloor)}$$

Mit dem gegebenen oberen Stichprobenquantil. Als (zentrales) Stichprobenquantil definieren wir

$$q_\alpha(\mu) = \frac{\bar{q}_\alpha(\mu) + \underline{q}_\alpha(\mu)}{2} \quad : \quad \hat{q}_{\alpha,n} = \frac{X_{(\lfloor n\alpha + 1 \rfloor)} + X_{(\lceil n\alpha \rceil)}}{2}$$

Insbesondere erhalten wir für $\alpha = 1/2$ den Median.



Als graphische Darstellung verwendet man häufig einen *Boxplot*, in dem die Quantile $\hat{q}_{1/4}, \hat{q}_{1/2}, \hat{q}_{3/4}$ sowie die Extremwerte aufgetragen sind:

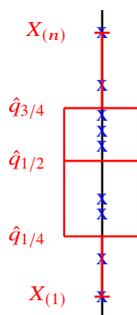


Abbildung 4.2.: Boxplot einer empirischen Verteilung auf \mathbb{R}

ZUSAMMENHANG ZWEIER NUMERISCHER MERKMALE In den folgenden Beispielen sei $S = \mathbb{R}^2$, mit den zwei Stichproben X_1, \dots, X_n und Y_1, \dots, Y_n welche μ verteilt und jeweils unabhängig sind unter \mathbb{P}_μ .

a) Dann können wir die Kovarianz betrachten

$$c(\mu) = \int (x - \int x d\mu)(y - \int y d\mu) d\mu \quad : \quad \hat{c}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

Welche uns die empirische Kovarianz als Plug-in Schätzer liefert.

b) Analog ergibt sich er Korrelationskoeffizient.

$$\varrho(\mu) = \frac{c(\mu)}{\sigma_X(\mu)\sigma_Y(\mu)} \quad : \quad \hat{\varrho}_n = \frac{\hat{c}_n}{\hat{\sigma}_{X,n} \cdot \hat{\sigma}_{Y,n}}$$

Insbesondere folgt nach Cauchy-Schwarz, dass gilt $\varrho(\mu) \in [-1, 1]$.

Bootstrap

Die nun betrachtete Bootstrap-Methode geht auf den US-amerikanischen Statistiker B. Efron zurück (1979). Der Name der Methode entstammt dem englischen Spruch "Pulling yourself up by your own bootstraps", was auf Deutsch so viel heißt wie "Sich am eigenen Schopf aus dem Sumpf ziehen".

Sei $\hat{g}_n = g(L_n)$ ein Plug-in-Schätzer. Wir wollen die Varianz von \hat{g}_n bestimmen und basierend auf \hat{g}_n Konfidenzintervalle für $g(\mu)$ angeben. Diese ist aber nicht bekannt, da wir die zugrundeliegende Wahrscheinlichkeitsverteilung μ nicht kennen. Die grundlegende Idee des Bootstrap-Verfahrens ist, dass die empirische Verteilung L_n unter geeigneten Voraussetzungen für große n eine hinreichend gute Approximation der unbekanntenen Verteilung μ liefern sollte. Wir beginnen mit zwei einfachen Fällen:

Beispiele (Lineare Funktionale).

- a) Für $g(\mu) = \mu(B)$ ist der Plug-in-Schätzer die relative Häufigkeit $\hat{g}_n = L_n(B) = H_n(B)/n$. Die absolute Häufigkeit $H_n(B)$ ist eine binomialverteilte Zufallsvariable mit Parametern n und $\mu(B)$. Daher können wir die in Kapitel 1 hergeleiteten Konfidenzintervalle für das Binomialmodell verwenden (z.B. exaktes Konfidenzintervall, Wilson, Wald). Die auf der Normalapproximation basierenden Wald-Intervalle haben die Form $\hat{g}_n \pm (\Phi^{-1}(1-\alpha/2)\hat{\sigma}_n)/\sqrt{n}$, wobei $\hat{\sigma}_n = \sqrt{\hat{g}_n(1-\hat{g}_n)}$ mit dem Plug-in-Schätzer für die Standardabweichung übereinstimmt.
- b) Für $g(\mu) = \int f d\mu$ ist $\hat{g}_n = \frac{1}{n} \sum f(X_i)$ der Plug-in-Schätzer. Die Standardabweichung von \hat{g}_n unter \mathbb{P}_n ist

$$\sigma(\hat{g}_n) = \frac{1}{\sqrt{n}\sigma_\mu(f)}$$

Da wir μ nicht kennen, können wir $\sigma_\mu(f)$ nicht berechnen. Stattdessen approximieren wir μ durch die empirische Verteilung L_n und erhalten:

$$\sigma_\mu(f)^2 \approx \sigma_{L_n}(f)^2 = \int (f - \hat{g}_n)^2 dL_n = \frac{1}{n} \sum_{i=1}^n (f(X_i) - \hat{g}_n)^2$$

Wenn die Voraussetzungen des zentralen Grenzwertsatzes erfüllt sind, dann erhalten wir approximative Konfidenzintervalle

$$\hat{g}_n \pm \frac{\Phi^{-1}(1-\frac{\alpha}{2})\sigma_{L_n}(f)}{\sqrt{n}}$$

für $g(\mu)$. Möglicherweise sind diese aufgrund der verwendeten Approximation allerdings ungenau.

Für nicht-lineare Funktionale $g(\mu)$ gibt es keine so elementare Schätzmethode für die Varianz und die Verteilung von \hat{g}_n . Eine erste Idee wäre, ein Monte Carlo Verfahren zu verwenden:

Angenommen, wir können viele unabhängige Kopien $X^{(1)}, X^{(2)}, \dots, X^{(B)}$ von der Stichprobe $X = (X_1, \dots, X_n)$ erzeugen. Jede dieser Kopien wäre also ein Datenvektor mit n Komponenten: $X^{(b)} = (X_1^{(b)}, \dots, X_n^{(b)})$. Dann würde nach dem Gesetz der großen Zahlen und den entsprechenden Fehlerabschätzungen gelten:

$$F(c) = \mathbb{P}_\mu[\hat{g}_n \leq c] \approx \frac{1}{B} \sum_{i=1}^B 1_{\hat{g}_n^{(b)} \leq c} =: \hat{F}_B(c) \quad (*)$$

$$\mathbb{P}_\mu \circ \hat{g}_n^{-1} \approx \frac{1}{B} \sum_{i=1}^B \delta_{\hat{g}_n^{(b)}} =: L_n^B \quad (**)$$

$$\text{Var}_\mu[\hat{g}_n] \approx \frac{1}{B-1} \sum_{b=1}^B \left(\hat{g}_n^{(b)} - \frac{1}{B} \sum_{r=1}^B \hat{g}_n^{(r)} \right)^2 =: V_B \quad (***)$$

Zum Beispiel würde die Bernstein-Ungleichung eine sehr gute Fehlerabschätzung für die Monte-Carlo-Approximation in (*) liefern. Wir stoßen jedoch auf das Problem, dass wir nur eine Stichprobe (X_1, \dots, X_n)

4. Empirische Verteilungen

haben.

BOOTSTRAP-IDEE: Falls n groß genug ist, folgt $\mu \approx L_n$ mit großer Wahrscheinlichkeit. Anstatt unabhängige Stichproben von μ zu erzeugen, simulieren wir *gegeben den Wert von X* (bedingt) unabhängige Stichproben

$$X^{(1)} = (X_1^{(1)}, \dots, X_n^{(1)}), \dots, X^{(B)} = (X_1^{(B)}, \dots, X_n^{(B)}) \sim \bigotimes_{i=1}^n L_n$$

durch Ziehen mit Zurücklegen von der empirischen Verteilung L_n , und schätzen

$$F(c) = \mathbb{P}_\mu[\hat{g}_n \leq c] \approx \mathbb{P}_{L_n}[\hat{g}_n \leq c] \approx \frac{1}{B} \sum_{b=1}^B 1_{\hat{g}_n^{(b)} \leq c} =: \hat{F}_B(c)$$

Wir wollen die beschriebenen Schritte im Folgenden konkretisieren.

BOOTSTRAP-VERFAHREN

Gegeben einer Stichprobe X erhalten wir die empirische Verteilung L_n .

- 1) Erzeuge unabhängige Stichproben $X_i^{(b)} \sim L_n$ mit $1 \leq b \leq B, 1 \leq i \leq n$.
- 2) Berechne $\hat{g}_n^{(b)} = g(L_n^{(b)})$, $L_n^{(b)} = 1/n \cdot \sum_i \delta_{X_i^{(b)}}$, für $b = 1, \dots, B$.
- 3) Schätze Verteilung bzw. Varianz von \hat{g}_n durch (*), (**), (***) .

Die erzeugten Stichproben heißen *Bootstrap-Stichproben* $X^* = (X_1^*, \dots, X_n^*)$. Bezüglich der bedingten Verteilungen gegeben X sind X_1^*, \dots, X_n^* unabhängig mit Verteilung L_n . Wir können die Stichproben explizit konstruieren durch eine Simulation von Ziehen mit Zurücklegen:

$$X_j^* = X_{I_j}, \quad \text{wobei } I_1, \dots, I_n \text{ unabhängig gleichverteilt sind auf } \{1, \dots, n\}$$

Analog zu oben erhalten wir $\hat{g}_n^* = g(L_n)$ mit $L_n^* = \frac{1}{n} \sum \delta_{X_i^*}$.

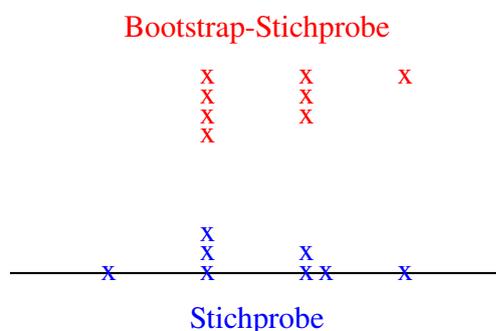


Abbildung 4.3.: Multinomiales Resampling

SCHÄTZEN DER VERTEILUNG VON \hat{g}_N Seien $X^{(b)} = (X_1^{(b)}, \dots, X_n^{(b)})$, $b = 1, \dots, B$ unabhängige (bedingt X) Bootstrap-Stichproben mit

$$\hat{g}_n^{(b)} = g(L_n^{(b)}), \quad L_n^{(b)} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i^{(b)}}$$

Unter \mathbb{P}_μ hat \hat{g}_n durch Bootstrap-Approximation die Verteilung $\approx \frac{1}{B} \sum_{b=1}^B \delta_{\hat{g}_n^{(b)}}$ auch Bootstrap Verteilung. Analog erhalten wir

$$F(c) \approx \hat{F}_B(c) = \frac{1}{B} \sum_{b=1}^B 1_{\hat{g}_n^{(b)} \leq c}$$

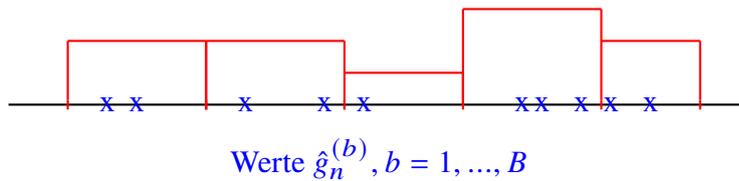


Abbildung 4.4.: Histogramm der Bootstrap Verteilung

Graphische Darstellung der Bootstrap-Verteilung über ein Histogramm:

HEURISTISCHE RECHTFERTIGUNG DES BOOTSTRAP-VERFAHRENS Im Folgenden wollen wir eine praktikable Anschauung über die Sinnhaftigkeit der Methoden des Bootstrap-Verfahrens geben. Hierfür betrachten wir die folgenden Approximationen

$$\hat{F}_B(c) \stackrel{\textcircled{1}}{\approx} \mathbb{P}_\mu [\hat{g}_n^* \leq c | X] \stackrel{\textcircled{2}}{=} \mathbb{P}_{L_n} [g(L_n) \leq c] \stackrel{\textcircled{3}}{\approx} \mathbb{P}_\mu [g(L_n) \leq c] = F(c)$$

- ① Dieser Schritt erfolgt durch klassische Monte-Carlo-Approximation. Dies ist in Ordnung, da $\hat{g}_n^{(b)}$ bedingt X unabhängig sind mit Verteilung $\hat{g}_n^* = g(L_n^*)$. Insbesondere erhalten wir eine präzise Fehlerabschätzung über die Bernstein-Ungleichung.
- ② Diese Gleichung stimmt, da L_n^* die empirische Verteilung von X_1^*, \dots, X_n^* ist, und die X_i^* unter der bedingten Verteilung gegeben X unabhängig sind mit Verteilung L_n .
- ③ Dies ist der problematische Schritt. Denn dann ist für

$$\phi_n(\mu) := \mathbb{P}_\mu [g(L_n) \leq c] \quad \text{zu zeigen, dass} \quad \phi_n(L_n) \approx \phi_n(\mu)$$

für n ausreichend groß. Diese Approximation ist nicht trivial und kann auch schiefgehen, siehe Beispiel unten. Würde ϕ_n nicht explizit von n abhängen, dann würde es ausreichen, dass L_n in einer gewissen Topologie gegen μ konvergiert und ϕ in dieser Topologie stetig ist. Da ϕ_n aber auch von n abhängt, benötigt man eine kompliziertere Approximation zur Rechtfertigung. Zum Beispiel:

$$\phi_n(\mu) \approx \phi_\infty(\mu) \approx \phi_\infty(L_n) \approx \phi_n(L_n)$$

Eine solche Schlussweise ist anwendbar, wenn folgende Bedingungen bezüglich eines geeigneten Konvergenzbegriffs erfüllt sind:

1. $L_n \rightarrow \mu$
2. $\phi_n \rightarrow \phi_\infty$ gleichmäßig in der Umgebung von μ
3. ϕ_∞ stetig

siehe zum Beispiel "Davidson, Hinkley: Bootstrap methods and their applications." Betrachten wir nun ein Beispiel, in dem dieser Schritt schiefgeht.

Beispiel. Seien X_1, \dots, X_n unabhängig mit Verteilung $\text{Unif}(0, \vartheta)$. Der MLE für ϑ ist $\hat{\vartheta}_n = \max(X_1, \dots, X_n)$, siehe oben. Wir können die exakte Verteilung von $\hat{\vartheta}_n$ berechnen:

$$\mathbb{P}_\vartheta \left[\hat{\vartheta}_n \leq \left(1 - \frac{c}{n}\right) \vartheta \right] = \left(1 - \frac{c}{n}\right)^n \sim e^{-c} \quad \text{für alle } c > 0$$

Insbesondere hat $\hat{\vartheta}_n$ eine absolutstetige Verteilung, genauer nähert sich die Verteilung von $n(1 - \hat{\vartheta}_n)$ für n gegen unendlich einer $\text{EXP}(1)$ -Verteilung an.

Sei nun (x_1, \dots, x_n) eine feste Realisierung von X mit $x_i \neq x_j$ für alle $i \neq j$. Dann gilt für die exakte Bootstrap-Verteilung:

$$\mathbb{P}_{L_n} [\hat{\vartheta}_n = x_{(n)}] = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow{n \uparrow \infty} 1 - e^{-1}$$

Mit $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Die Bootstrap-Verteilung enthält also einen Dirac-Anteil bei $x_{(n)}$, dessen Gewicht für n gegen unendlich nicht gegen 0 geht. Mit anderen Worten: Die Bootstrap-Verteilung hat einen diskreten Anteil, der auch für n gegen unendlich nicht verschwindet. Daher liefert diese Verteilung auch für große n keine Approximation der tatsächlichen Verteilung von $\hat{\vartheta}_n$.

Konfidenzintervalle für $g(\mu)$

NORMALAPPROXIMATION: Eine sehr grobe Methode, um mithilfe des Bootstrap-Verfahrens Konfidenzintervalle zu erhalten, geht davon aus, dass \hat{g}_n unter \mathbb{P}_μ näherungsweise normalverteilt ist. Dies ist nicht immer der Fall und sollte zuvor empirisch getestet werden, zum Beispiel mithilfe des Bootstrap-Histogramms. Geht man von einer approximativen Normalverteilung aus, dann kann man das Bootstrap-Konfidenzintervall

$$\hat{g}_n \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}_B} \quad \text{mit} \quad \hat{V}_B = \frac{1}{B-1} \cdot \sum_{n=1}^B \left(\hat{g}_n^{(b)} - \sum_{r=1}^B \hat{g}_n^{(r)} \right)^2$$

verwenden, wobei \hat{V}_B der Bootstrap-Schätzwert der Varianz von \hat{g}_n ist.

BOOTSTRAP-INTERVALLE Sei $R_n := \hat{g}_n - g(\mu)$. Sind C und D Statistiken, dann ist das Intervall (C, D) ein Konfidenzintervall für $g(\mu)$ zum Niveau $1 - \alpha$, falls

$$\mathbb{P}_\mu[g(\mu) \notin (C, D)] = \mathbb{P}_\mu[R_n \notin (\hat{g}_n - D, \hat{g}_n - C)] \leq \alpha$$

Dies ist erfüllt, falls

$$\begin{aligned} \hat{g}_n - C &= q_{1-\alpha/2} & \text{also} & & C &= \hat{g}_n - q_{1-\alpha/2} \quad , \text{ und} \\ \hat{g}_n - D &= q_{\alpha/2} & \text{also} & & D &= \hat{g}_n - q_{\alpha/2} \end{aligned}$$

wobei q_β das β -Quantil der Verteilung von R_n unter \mathbb{P}_μ ist. Da wir diese Quantile nicht kennen, schätzen wir sie mit einem Bootstrap-Verfahren. Dabei ersetzen wir wieder die exakte Verteilung μ durch die empirische Verteilung L_n . Die Bootstrap-Replikationen von $R_n = \hat{g}_n - g(\mu)$ sind

$$R^{(b)} := \hat{g}_n^{(b)} - g(L_n) = \hat{g}_n^{(b)} - \hat{g}_n \quad \text{mit } b = 1, \dots, B$$

und wir schätzen

$$q_\beta \approx \hat{q}_\beta(R_n^{(1)}, \dots, R_n^{(B)}) = \hat{q}_\beta(\hat{g}_n^{(1)}, \dots, \hat{g}_n^{(B)}) - \hat{g}_n$$

wobei \hat{q}_β das jeweilige Stichprobenquantil bezeichnet. Damit ergibt sich das Bootstrap-Konfidenzintervall

$$\begin{aligned} (\hat{C}, \hat{D}) &= \left(\hat{g}_n - \hat{q}_{1-\alpha/2}(R_n^{(1)}, \dots, R_n^{(B)}), \hat{g}_n - \hat{q}_{\alpha/2}(R_n^{(1)}, \dots, R_n^{(B)}) \right) \\ &= \left(2\hat{g}_n - \hat{q}_{1-\alpha/2}(\hat{g}_n^{(1)}, \dots, \hat{g}_n^{(B)}), 2\hat{g}_n - \hat{q}_{\alpha/2}(\hat{g}_n^{(1)}, \dots, \hat{g}_n^{(B)}) \right) \end{aligned}$$

4.3. Anpassungstests

Sei μ eine unbekannte Wahrscheinlichkeitsverteilung auf S . X_1, \dots, X_n sind unabhängige Stichproben mit $X_i \sim \mu$ unter \mathbb{P}_μ . Anhand der Stichproben soll eine gewisse Verteilung μ_0 auf die Verteilung der Stichproben getestet werden. Wir betrachten also das Testproblem:

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Beispielsweise können wir uns fragen, ob die Augenzahlen eines Würfels gleichverteilt auf der Menge $\{1, \dots, 6\}$ sind. Ebenso könnte man einen Zufallsgenerator anhand seiner Ausgaben U_i auf seine Genauigkeit testen wollen, also $(U_i, U_{i+1}, \dots, U_{i+k}) \sim \text{Unif}(0, 1)^{k+1}$?

Wie bereits oben erwähnt, wollen wir in einigen Fällen die Normalverteilung testen, um herauszufinden, ob gewisse Modellierungsannahmen gerechtfertigt sind. Wir betrachten zunächst kategorielle Merkmale, das heißt, dass die Menge S endlich ist. Sei also $S = \{a_1, \dots, a_k\}$. Wir setzen

$$p_l := \mu(a_l) \quad \text{und} \quad p_l^0 := \mu_0(a_l).$$

Bemerkung. Ist S nicht endlich, sondern zum Beispiel \mathbb{R} , dann können wir die folgenden Verfahren trotzdem anwenden, indem wir S in endlich viele disjunkte Teilmengen ("bins") B_1, \dots, B_k unterteilen, und dann die kategoriellen Merkmale $\tilde{X}_i := \sum_{l=1}^k l \cdot 1_{\{X_i=l\}}$ betrachten. Dabei ist es natürlich wichtig, die Unterteilung möglichst geschickt zu wählen.

Eine suffiziente Statistik ist die empirische Verteilung L_n sowie die Häufigkeitsverteilung nL_n , die wir graphisch als Histogramm darstellen können. Es gilt

$$nL_n \sim \text{Mult}(n, p_1, \dots, p_n) \quad \text{unter} \quad \mathbb{P}_\mu$$

Die Likelihood-Funktion können wir mithilfe der (relativen) Entropie ausdrücken:

Lemma 4.4. Sei μ eine Wahrscheinlichkeitsverteilung auf S mit $\mu(a_l) > 0$ für $l = 1, \dots, k$. Dann gilt

$$L(\mu; X_1, \dots, X_n) = e^{-n(H(L_n) + H(L_n|\mu))}$$

Beweis. Wegen der Unabhängigkeit von X_1, \dots, X_n gilt:

$$\begin{aligned} L(\mu; X) &= \prod_{i=1}^n P_{X_i} = \prod_{l=1}^k p_l^{nL_n(a_l)} \\ \Rightarrow \frac{1}{n} \log L(\mu; X) &= \sum_{l=1: L_n(a_l) \neq 0}^k L_n(a_l) \log p_l \\ &= \sum_{l=1}^k L_n(a_l) \log L_n(a_l) - \sum_{l=1}^k L_n(a_l) \log \frac{L_n(a_l)}{p_l} \\ &= -H(L_n) - H(L_n|\mu) \end{aligned}$$

Insbesondere ist $\hat{\mu} = L_n$ das globale Maximum der Likelihood-Funktion, also der MLE, und

$$\sup_{\mu} L(\mu; X) = L(L_n, X) = e^{-nH(L_n)}$$

Damit erhalten wir den Likelihood-Quotienten

$$\lambda(X) = \frac{\sup_{\mu \neq \mu_0} L(\mu, X)}{L(\mu_0; X)} = \frac{\sup_{\mu} L(\mu; X)}{L(\mu_0; X)} = \frac{e^{-nH(L_n)}}{e^{-n(H(L_n) + H(L_n|\mu))}} = e^{nH(L_n|\mu)}$$

Im Folgenden werden wir verschiedene Anpassungstests betrachten.

G-Test (Hoeffdings Entropietest)

Der G-Test ist der Likelihood-Quotienten-Test für das Testproblem

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 = \mu \neq \mu_0.$$

Die Entscheidungsregel zum Schwellenwert c lautet:

$$\text{Verwerfe } H_0, \text{ falls } G := nH(L_n|\mu_0) \geq c.$$

Explizit erhalten wir:

$$G = n \sum_{l=1}^k L_n(a_l) \log \frac{L_n(a_l)}{\mu_0(a_l)} = \sum_{l=1}^k H_l \log \frac{H_l}{np_l^0}$$

4. Empirische Verteilungen

mit $H_l := nL_n(a_l)$ und $p_l^0 := \mu_0(a_l)$. Dabei ist H_l die Häufigkeit von a_l in der Stichprobe $X = (X_1, \dots, X_n)$ und np_l^0 ist die erwartete Häufigkeit.

Sei nun g der beobachtete Wert der Teststatistik G . Der (rechtsseitige) p -Wert ist dann

$$p = \mathbb{P}_0[G \geq g].$$

Der exakte G -Test verwirft die Nullhypothese zum Niveau α , falls $p \leq \alpha$ gilt. Normalerweise können wir den p -Wert jedoch nicht exakt berechnen. Stattdessen können wir ein Monte-Carlo-Verfahren zur Approximation von p verwenden. Sind G_1, \dots, G_B unabhängige Stichproben der Verteilung von G unter \mathbb{P}_0 , dann gilt

$$p \approx \frac{|\{b \in \{1, \dots, B\} : G_b \geq g\}| + 1}{B + 1} =: \hat{p}_{MC}.$$

Der Schätzer \hat{p}_{MC} heißt *Monte-Carlo- p -Wert*. Dieser liefert den folgenden Test:

MONTE-CARLO-G-TEST:

- (i) Simuliere $X_i^{(b)} \sim \mu_0$ mit $i = 1, \dots, n$ und $b = 1, \dots, B$ unabhängig.
- (ii) Für $b = 1, \dots, B$ berechne jeweils den Wert G_b der G -Statistik für $X^{(b)} = (X_1^{(b)}, \dots, X_n^{(b)})$.
- (iii) Berechne \hat{p}_{MC} und verwirfe H_0 , falls $\hat{p}_{MC} \leq \alpha$.

Das folgende Lemma zeigt, dass der Monte-Carlo-Test die Niveaubedingung nicht nur asymptotisch für B gegen unendlich, sondern sogar für jedes feste $B \in \mathbb{N}$ erfüllt.

Lemma 4.5 (Niveau bei Monte-Carlo-Test). Seien $G_0, \dots, G_B : \Omega \rightarrow \mathbb{R}$ austauschbare Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Das heißt, für jede Permutation σ von $\{0, 1, \dots, B\}$ gilt:

$$(G_{\sigma(0)}, G_{\sigma(1)}, \dots, G_{\sigma(B)}) \sim (G_0, G_1, \dots, G_B) \quad \text{und sei} \quad \hat{p} := \frac{|\{b \in \{1, \dots, B\} : G_b \geq G_0\}| + 1}{B + 1}.$$

Dann gilt

$$\mathbb{P}[\hat{p} \leq \alpha] \leq \frac{\lfloor (B+1)\alpha \rfloor}{B+1} \leq \alpha \quad \text{für alle } \alpha \in (0, 1), B \in \mathbb{N}.$$

Bemerkung. Insbesondere sind unabhängige und identisch verteilte Zufallsvariablen austauschbar.

Beispiel (Monte-Carlo-Anpassungstest). Sei $G_0 = G$ der beobachtete G -Wert und G_1, \dots, G_B simulierte G -Werte. Diese sind unabhängig und identisch verteilt, also gilt:

$$\mathbb{P}_0[\hat{p}_{MC} \leq \alpha] \leq \alpha.$$

Das heißt, dass der Monte-Carlo-Test ein Niveau- α -Test ist. Dies gilt für jedes $B \in \mathbb{N}$, also sogar für $B = 1$. Ist aber $B + 1 < \frac{1}{\alpha}$, dann ist $\frac{\lfloor (B+1)\alpha \rfloor}{B+1} = 0$, das heißt, der Test verwirft nie. Damit eine sinnvolle Anwendung möglich ist, sollte zumindest $B + 1 \geq \frac{2}{\alpha}$ gelten, denn dann folgt $\frac{\lfloor (B+1)\alpha \rfloor}{B+1} \geq \frac{\alpha}{2}$.

Beweis. Für $i \in \{0, \dots, B\}$ sei

$$\hat{p}_i := \frac{|\{b \in \{0, \dots, B\} : G_b \geq G_i\}|}{B + 1}.$$

Da G_0, \dots, G_B austauschbar sind, haben die Zufallsvariablen alle dieselbe Verteilung, also $\hat{p}_i \sim \hat{p}_0 = \hat{p}$ für alle $i \in \{0, \dots, B\}$. Damit erhalten wir:

$$\mathbb{P}_0[\hat{p} \leq \alpha] = \frac{1}{B + 1} \sum_{i=0}^B \mathbb{P}_0[\hat{p}_i \leq \alpha] = \mathbb{E}_0 \left[\frac{1}{B + 1} \sum_{i=0}^B 1_{\hat{p}_i \leq \alpha} \right].$$

Der Beweis ist abgeschlossen, wenn wir zeigen können, dass die Anzahl aller $i \in \{0, \dots, B\}$ mit $\hat{p}_i \leq \alpha$ kleiner gleich $\lfloor (B + 1)\alpha \rfloor$ ist. Dazu bemerken wir:

$$\hat{p}_i \leq \alpha \Leftrightarrow |\{b \in \{0, \dots, B\} : G_b \geq G_i\}| \leq \lfloor (B + 1)\alpha \rfloor.$$

Dies ist äquivalent dazu, dass es höchstens $\lfloor (B + 1)\alpha \rfloor$ Werte $G_b \geq G_i$ gibt. Dies kann aber nur für die $\lfloor (B + 1)\alpha \rfloor$ größten Werte gelten, also höchstens $\lfloor (B + 1)\alpha \rfloor$ mal. ■

Chiquadrat-Anpassungstest

Da es früher nicht möglich war, Monte-Carlo-Tests durchzuführen, verwendete man approximative Tests. Der bekannteste approximative Anpassungstest ist der Chiquadrat-Test, bei dem sowohl die relative Entropie durch eine einfache (quadratische) Funktion als auch die Verteilung der Teststatistik mithilfe einer Normalapproximation approximiert wird. Die Taylor-Approximation

$$x \log x = x - 1 + \frac{1}{2}(x - 1)^2 + O(|x - 1|^3)$$

liefert für die relative Entropie zweier Wahrscheinlichkeitsverteilungen μ, ν mit relativer Dichte $\varrho = \frac{d\mu}{d\nu} \approx 1$ die Näherung

$$H(\mu|\nu) = \int \varrho \log \varrho \, d\nu \approx \underbrace{\int (\varrho - 1) \, d\nu}_{=0} + \frac{1}{2} \int (\varrho - 1)^2 \, d\nu$$

Definition 4.6. Die *Chiquadrat-Divergenz* von μ bezüglich ν ist gegeben durch

$$D_2(\mu|\nu) := \begin{cases} \int (\varrho - 1)^2 \, d\nu & \text{falls } \mu \ll \nu \text{ mit Dichte } \varrho \\ +\infty & \text{sonst} \end{cases}$$

Ersetzt man die relative Entropie (Kullback-Leibler-Divergenz) durch die Chiquadrat-Divergenz, ergibt sich die folgende Entscheidungsregel des approximativen Likelihood-Quotienten-Tests:

$$\text{Verwerfe } H_0, \text{ falls } T := nD_2(L_n|\mu_0) \geq c$$

Explizit erhalten wir:

$$T = n \sum_{l=1}^k \left(\frac{L_n(a_l)}{\mu_0(a_l)} - 1 \right)^2 \mu_0(a_l) = n \sum_{l=1}^k \frac{(L_n(a_l) - \mu_0(a_l))^2}{\mu_0(a_l)} = \sum_{l=1}^k \frac{(H_l - np_l^0)^2}{np_l^0} = \sum_{l=1}^k \frac{H_l^2}{np_l^0} - n$$

also

$$T = \sum_{l=1}^k \frac{(\text{Häufigkeit von } a_l - \text{erwartete Häufigkeit von } a_l)^2}{\text{erwartete Häufigkeit von } a_l}$$

Die Verteilung dieser Teststatistik unter H_0 berechnen wir approximativ mithilfe einer Normalapproximation der Multinomialverteilung. Unter H_0 gilt:

$$H = (H_1, \dots, H_k) \sim \text{Mult}(n, p^0), \quad H_l \sim \text{Bin}(n, p_l^0)$$

Wir betrachten nun $Y = (Y_1, \dots, Y_k)$ mit $Y_l := \frac{H_l - np_l^0}{\sqrt{np_l^0}}$. Nach Definition der Teststatistik gilt

$$T = \sum_{l=1}^k Y_l^2 = \|Y\|_{\mathbb{R}^k}^2$$

Wegen $\sum H_l = n$ kann der Vektor nur Werte in einer $(k - 1)$ -dimensionalen Hyperebene im \mathbb{R}^k annehmen. In der Tat gilt

$$\sum_l \sqrt{p_l^0} Y_l = \frac{1}{\sqrt{n}} \sum_l (H_l - np_l^0) = \frac{n - n}{\sqrt{n}} = 0$$

4. Empirische Verteilungen

also

$$Y \in (\sqrt{p_1^0}, \dots, \sqrt{p_k^0})^\perp =: \mathbb{H} \subseteq \mathbb{R}^k$$

Seien e_1, \dots, e_{k-1} eine Orthonormalbasis von \mathbb{H} und $\tilde{Y} := (e_1 \cdot Y, \dots, e_{k-1} \cdot Y) \in \mathbb{R}^{k-1}$ die entsprechende Koordinatendarstellung von Y in dieser Basis. Dann folgt wegen $Y \in \mathbb{H}$:

$$T = \|Y\|_{\mathbb{R}^k}^2 = \|\tilde{Y}\|_{\mathbb{R}^{k-1}}^2 = \sum_{l=1}^{k-1} \tilde{Y}_l^2$$

Der zentrale Grenzwertsatz liefert eine Normalapproximation für \tilde{Y} :

Satz 4.7. Im Grenzwert n gegen unendlich gilt:

- 1) Die Verteilung von \tilde{Y} unter \mathbb{P}_0 konvergiert schwach gegen eine $(k-1)$ -dimensionale Standardnormalverteilung.
- 2) Die Verteilung von T unter \mathbb{P}_0 konvergiert schwach gegen eine Chiquadrat-Verteilung mit $k-1$ Freiheitsgraden.

Bemerkung. Die Verteilung von Y konvergiert gegen eine Normalverteilung im \mathbb{R}^k mit degenerativer Kovarianzmatrix C .

Aus dem Satz ergibt sich als Approximation des p -Werts für große n :

$$p = \mathbb{P}_0[T \geq t] \approx 1 - F_{\chi^2(k-1)}(t)$$

wobei t der beobachtete Wert von T ist. Damit erhalten wir den folgenden approximativen Test zum Niveau α :

CHIQUADRAT-TEST ZUM NIVEAU α (PEARSON 1900):

Verwerfe H_0 , falls $t \geq q_{1-\alpha, \chi^2(k-1)}$.

Beweis. 1) Für $l = 1, \dots, k$ gilt

$$Y_l = \frac{H_l - np_l^0}{\sqrt{np_l^0}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{il} \text{ mit } V_{il} := \frac{1_{X_i=l} - p_l^0}{\sqrt{p_l^0}}$$

also $Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i$. Die Zufallsvektoren V_i sind unabhängig und identisch verteilt mit Werten in der Hyperebene $\mathbb{H} \subseteq \mathbb{R}^k$. Sei

$$\tilde{V}_i := (e_1 \cdot V_i, \dots, e_{k-1} \cdot V_i) \in \mathbb{R}^{k-1}$$

die Darstellung in der Orthonormalbasis $\{e_1, \dots, e_{k-1}\}$ von \mathbb{H} . Damit gilt dann

$$\tilde{Y} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{V}_i$$

und auch die \tilde{V}_i sind wieder unabhängig und identisch verteilt. Wir rechnen nun nach, dass diese Zufallsvektoren standardisiert sind und wenden dann den zentralen Grenzwertsatz im \mathbb{R}^{k-1} an. In der Tat gilt

$$\mathbb{E}_0[V_{il}] = 0 \text{ für } l = 1, \dots, k \quad \Rightarrow \quad \mathbb{E}_0[\tilde{V}_{ir}] = e_r \cdot \mathbb{E}_0[V_i] = 0 \text{ für } r = 1, \dots, k-1$$

Für die Kovarianzmatrizen erhalten wir

$$\begin{aligned} C_{lm} &:= \text{Cov}_0(V_{il}, V_{im}) = \frac{1}{\sqrt{p_l^0 p_m^0}} \text{Cov}_0(1_{X_i=l}, 1_{X_i=m}) \\ &= \frac{1}{\sqrt{p_l^0 p_m^0}} (\delta_{lm} p_l^0 - p_l^0 p_m^0) = \delta_{lm} - \sqrt{p_l^0 p_m^0} \end{aligned}$$

also

$$\begin{aligned} \text{Cov}_0(\tilde{V}_{ir}, \tilde{V}_{is}) &= \sum_{l,m} e_{rl} e_{sm} \cdot \text{Cov}_0(V_{il}, V_{im}) \\ &= \sum_l e_{rl} e_{sl} - \sum_l e_{rl} \sqrt{p_l^0} \cdot \sum_s e_{sm} \sqrt{p_m^0} = e_r \cdot e_s = \delta_{rs} \end{aligned}$$

wobei wir benutzt haben, dass die Vektoren e_r und e_s in der Hyperebene \mathbb{H} liegen. Also sind V_1, V_2, \dots unabhängige, identisch verteilte, standardisierte Zufallsvektoren im \mathbb{R}^{k-1} . Nach dem zentralen Grenzwertsatz folgt, dass $\tilde{Y} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{V}_i$ in Verteilung gegen $N(0, I_{k-1})$ konvergiert.

2) Wegen $T = \sum_{l=1}^k Y_l^2 = \|Y\|_{\mathbb{R}^k}^2 = \|\tilde{Y}\|_{\mathbb{R}^{k-1}}^2$ folgt

$$\mathbb{P}_0[t \leq c] = \mathbb{P}_0[\|\tilde{Y}\|_{\mathbb{R}^{k-1}}^2 \leq c] \xrightarrow{n \uparrow \infty} \mathbb{P}_0[\|Z\|^2 \leq c]$$

mit $Z \sim N(0, I_{k-1})$ für alle $c \in \mathbb{R}$. Die Behauptung folgt, da $\|Z\|^2$ die Verteilung $\chi^2(k-1)$ hat. ■

Beispiel (Mendels Erbsen). Nach dem 2. Mendelschen Gesetz sollte sich in der 2. Generation eine Verteilung $9 : 3 : 3 : 1$ zwischen den vier möglichen Typen von Erbsen ergeben, das heißt $p^0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$. Ein empirischer Test liefert die Häufigkeitsverteilung $h = (272, 89, 90, 29)$. Als beobachteter Wert der Chiquadrat-Statistik ergibt sich

$$t = \frac{1}{480} \left(\frac{272^2}{9/16} + \frac{89^2}{3/16} + \frac{90^2}{3/16} + \frac{29^2}{1/16} \right) - 400 = 0,0059.$$

Der empirische p -Wert ist

$$p_r = \mathbb{P}_0[T \geq t] \approx 1 - F_{\chi^2(3)}(0,0059) > 0,995.$$

Der beobachtete Wert t ist also nicht auffällig groß, aber auffällig klein. Genauer gilt $p_l = \mathbb{P}_0[T \leq t] < 0,005$. Die empirische Verteilung liegt also näher an der tatsächlichen Verteilung, als man es bei einem Zufallsexperiment erwarten würde! Dies deutet darauf hin, dass die beobachteten Werte möglicherweise manipuliert sind.

Die asymptotische Chiquadrat-Verteilung gilt nicht nur für die Teststatistik $T_n = nD_2(L_n|\mu)$ im Chiquadrat-Test, sondern auch für die Teststatistik $G_n = nH(L_n|\mu)$ im G-Test:

Lemma 4.8. Sei $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ die empirische Verteilung von unabhängigen, identisch verteilten Zufallsvariablen X_i mit Verteilung μ_0 . Dann gilt für $n \rightarrow \infty$:

$$1) \quad 2G_n - T_n \rightarrow 0 \quad \mathbb{P}_0\text{-stochastisch,}$$

$$2) \quad \mathbb{P}_0 \circ (2G_n)^{-1} \xrightarrow{w} \chi^2(k-1).$$

Der Beweis ist eine Übungsaufgabe.

Anpassungstest für parametrische Familie

Oft will man nicht testen, ob eine bestimmte Verteilung vorliegt, sondern ob die zugrunde liegende Verteilung aus einer bestimmten Familie kommt (um dann das entsprechende parametrische Modell zu betrachten). Die Nullhypothese hat dann die Form

$$H_0 : \mu \in \{\mu_\vartheta : \vartheta \in \Theta_0\},$$

wobei Θ_0 eine d -dimensionale Parametermenge ist. In diesem Fall kann man den G-Test (oder Chiquadrat-Test) mit Parameterschätzung durchführen. Die Verwerfungsregel lautet: Verwerfe H_0 , falls $G := nH(L_n | \hat{\mu}_\vartheta) \geq c/2$, wobei $\hat{\vartheta}$ der Maximum-Likelihood-Schätzer für ϑ ist. Das *Chiquadrat-Prinzip von Wilks* besagt, dass unter geeigneten Voraussetzungen die modifizierte G-Statistik unter der Nullhypothese asymptotisch Chiquadrat-verteilt ist mit $k - 1 - d$ Freiheitsgraden.

Konfidenzbereich für μ

Als Alternative zu Anpassungstests kann man einen Konfidenzbereich für die unbekannte Wahrscheinlichkeitsverteilung μ angeben. Seien X_1, \dots, X_n unabhängig unter \mathbb{P}_0 mit Verteilung μ . Dabei sei μ eine Wahrscheinlichkeitsverteilung auf $\{a_1, \dots, a_k\}$ mit Gewichten $p_l = \mu(a_l)$ und $p = (p_1, \dots, p_k) \in \mathbb{R}^k$. Ein einfacher Konfidenzbereich für die Massenfunktion p hat die Form

$$C = (A_1, B_1) \times (A_2, B_2) \times \dots \times (A_n, B_n),$$

wobei A_l und B_l für $l = 1, \dots, n$ Statistiken $A_l \leq B_l$ sind. Es gilt:

$$\mathbb{P}_\mu[p \notin C] = \mathbb{P}_\mu[\text{Es gibt ein } l \in \{1, \dots, k\} : p_l \notin (A_l, B_l)] \leq \sum_{l=1}^k \mathbb{P}_\mu[p_l \notin (A_l, B_l)].$$

Damit folgt unmittelbar das Lemma:

Lemma 4.9. *Ist (A_l, B_l) für jedes $l = 1, \dots, k$ ein Konfidenzintervall für p_l zum Niveau $1 - \alpha/k$, dann ist C ein Konfidenzbereich für p zum Niveau $1 - \alpha$.*

Da H_l unter \mathbb{P}_μ binomialverteilt ist mit Parametern n und p_l , können wir für die Komponenten p_l die Konfidenzintervalle aus dem Binomialmodell verwenden.

VORTEIL: Kontrolle aller Werte p_1, \dots, p_k , nicht nur im quadratischen Mittel.

NACHTEIL: Wir benötigen Konfidenzintervalle zum Niveau $\frac{\alpha}{k}$ statt α ("Bonferroni-Korrektur"), das heißt wir brauchen mehr Stichprobenwerte.

4.4. Empirische Verteilungen numerischer Merkmale

Wir betrachten nun numerische Merkmale, das heißt reellwertige Stichprobenwerte. Sei also μ eine unbekannte Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, und sei $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ die empirische Verteilung von unabhängigen Stichprobenwerten $X_1, \dots, X_n \sim \mu$. In diesem Fall bilden die Ordnungsstatistiken $(X_{(1)}, \dots, X_{(n)})$ eine suffiziente Statistik, das heißt es ist irrelevant, in welcher Reihenfolge die Werte $X_{(1)}, \dots, X_{(n)}$ beobachtet wurden. Die empirische Verteilung können wir durch ihre Verteilungsfunktion

$$F_n(c) = L_n((-\infty, c]) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq c} \quad \text{für alle } c \in \mathbb{R}$$

beschreiben. Diese ist durch die Ordnungsstatistiken festgelegt:

$$F_n(c) = \frac{i}{n} \quad \text{für } c \in [X_{(i)}, X_{(i+1)}],$$

wobei wir $X_{(0)} := -\infty$ und $X_{(n+1)} := \infty$ setzen. Die empirische Verteilungsfunktion ist ein erwartungstreuer Schätzer für die Verteilungsfunktion F von μ :

$$\mathbb{E}_\mu[F_n(c)] = F(c) \quad \text{für alle } c \in \mathbb{R}.$$

Konfidenzbereich für F

Wir zeigen nun, dass Konfidenzbereiche für die unbekannte Verteilungsfunktion F durch ε -Bänder um die empirische Verteilungsfunktion gegeben sind. Die benötigte Abschätzung können wir mithilfe einer Quantiltransformation auf den Fall unabhängiger Zufallsvariablen $U_1, \dots, U_n \sim \text{Unif}(0, 1)$ auf $(\Omega, \mathcal{A}, \mathbb{P})$ zurückführen. Für $v \in [0, 1]$ sei

$$F_n^U(v) = \frac{1}{n} \sum_{i=1}^n 1_{U_i \leq v}$$

die entsprechende empirische Verteilungsfunktion.

Satz 4.10 (Dvoretzky-Kiefer-Wolfowitz-Ungleichung; Massart 1990). Für alle $n \in \mathbb{N}, \varepsilon > 0$ und alle Wahrscheinlichkeitsverteilungen μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gilt

$$\mathbb{P}_\mu \left[\sup_{c \in \mathbb{R}} |F_n(c) - F(c)| \geq \varepsilon \right] \leq \mathbb{P} \left[\sup_{v \in [0,1]} |F_n^U(v) - v| \geq \varepsilon \right] \leq 2e^{-2n\varepsilon^2}. \quad (\text{DKW})$$

Aus dem Satz folgt unmittelbar, dass für $2e^{-2n\varepsilon^2} \leq \alpha$ die Verteilungsfunktion F mit Sicherheit $1 - \alpha$ in einem ε -Band (ε -Umgebung bezüglich der sup-Norm) um F_n liegt.

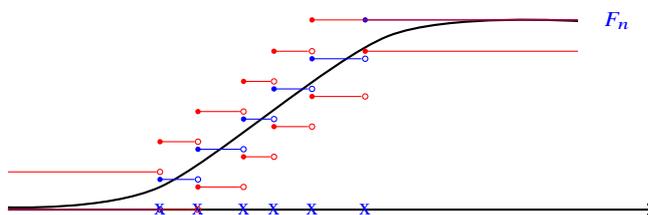


Abbildung 4.5.: Konfidenzband für F

Bemerkung. Da Verteilungsfunktionen rechtsstetig sind, gilt

$$\sup_{c \in \mathbb{Q}} |F_n(c) - F(c)| = \sup_{c \in \mathbb{R}} |F_n(c) - F(c)| \quad (**)$$

Hieraus folgt, dass das Supremum eine Zufallsvariable ist.

Beweis. Der Beweis der ersten Ungleichung beruht auf einer Quantiltransformation. Mit $\underline{G}(u) = \inf\{c \in \mathbb{R} : F(c) \geq u\}$ gilt

$$(X_1, \dots, X_n) \sim (\underline{G}(U_1), \dots, \underline{G}(U_n)) \quad (*)$$

Da das Supremum in (**) eine Funktion von X_1, \dots, X_n ist, folgt

$$\begin{aligned} \mathbb{P}_\mu \left[\sup_{c \in \mathbb{R}} |F_n(c) - F(c)| \geq \varepsilon \right] &= \mathbb{P} \left[\sup_{c \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n 1_{\underline{G}(U_i) \leq c} - F(c) \right| \geq \varepsilon \right] \\ &\stackrel{v:=F(c)}{\leq} \mathbb{P} \left[\sup_{v \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n 1_{U_i \leq v} - v \right| \geq \varepsilon \right] \end{aligned}$$

4. Empirische Verteilungen

Dies ist die erste Ungleichung in (DKW).

Statt der zweiten Ungleichung zeigen wir hier nur eine etwas schwächere Abschätzung. Der Beweis der vollen Aussage findet sich in "Massart, *Annals of Probability* 1990". Zum Beweis der schwächeren Abschätzung sei $k \in \mathbb{N}$ und

$$M_k := \max_{j=0, \dots, k} \left| F_n^U \left(\frac{j}{k} \right) - \frac{j}{k} \right|$$

aus der Bernstein-Ungleichung folgt:

$$\mathbb{P}[M_k \geq \varepsilon] \leq \sum_{j=1}^{k-1} \mathbb{P} \left[\left| F_n^U \left(\frac{j}{k} \right) - \frac{j}{k} \right| \geq \varepsilon \right] \leq 2(k-1)e^{-2n\varepsilon^2}$$

Für $v \in [0, 1]$ sei $j \in \{0, \dots, n-1\}$ mit $v \in \left[\frac{j-1}{k}, \frac{j}{k} \right]$. Dann folgt:

$$\begin{aligned} F_n^U(v) - v &\leq F_n^U \left(\frac{j}{k} \right) - \frac{j-1}{k} \leq M_k + \frac{1}{k} && \text{und} \\ v - F_n^U(v) &\leq \frac{j}{k} - F_n^U \left(\frac{j-1}{k} \right) \leq M_k + \frac{1}{k} \end{aligned}$$

Hierbei haben wir benutzt, dass Verteilungsfunktionen monoton wachsend sind. Da die Abschätzungen für alle $v \in [0, 1]$ gelten, erhalten wir:

$$\sup_{v \in [0,1]} |F_n^U(v) - v| \leq M_k + \frac{1}{k} \text{ und damit } \mathbb{P} \left[\sup_{v \in [0,1]} |F_n^U(v) - v| \geq \varepsilon \right] \leq \mathbb{P} \left[M_k \geq \varepsilon - \frac{1}{k} \right]$$

Wählen wir nun $k := \lceil \frac{2}{\varepsilon} \rceil$, dann erhalten wir

$$\mathbb{P} \left[\sum_{v \in [0,1]} |F_n^U(v) - v| \geq \varepsilon \right] \geq \mathbb{P} \left[M_k \geq \frac{\varepsilon}{2} \right] \leq 2(k-1)e^{-\frac{n\varepsilon^2}{2}} \leq \frac{4}{\varepsilon} e^{-\frac{n\varepsilon^2}{2}}$$

In dieser Abschätzung ist der Exponent etwas kleiner als in der optimalen Abschätzung, und wir haben zusätzlich den Faktor $\frac{1}{\varepsilon}$ vor dem Exponenten erhalten. ■

Aus Satz 4.8 (oder auch aus der bewiesenen schwächeren Abschätzung) folgt, dass die empirischen Verteilungsfunktionen für $n \rightarrow \infty$ fast sicher gleichmäßig konvergieren:

Korollar 4.11 (Satz von Glivenko-Cantelli). Für jede Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gilt:

$$\mathbb{P}_\mu[\sup |F_n - F| \rightarrow 0] = 1$$

Beweis. Nach Satz 4.8 ist $\sum_{n=1}^{\infty} \mathbb{P}_\mu[\sup |F_n - F| > \varepsilon] < \infty$ für alle $\varepsilon > 0$. Die Folge der Suprema konvergiert also schnell stochastisch gegen Null, und die Behauptung folgt aus dem Borel-Cantelli-Lemma. ■

Anpassungstests

Satz 4.8 lieferte einen Konfidenzbereich für die Verteilungsfunktion F . Ebenso können wir den Satz benutzen, um Hypothesentests zu konstruieren. A) KOLMOGOROV-SMIRNOV-TEST

Hier testen wir die Nullhypothese $\mu = \mu_0$, wobei μ_0 eine feste Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist. Die Nullhypothese wird verworfen, falls

$$S_n := \sup_{c \in \mathbb{R}} |F_n(c) - F_0(c)| \geq \varepsilon$$

gilt, wobei ε ein vorgegebener Schwellenwert und F_0 die Verteilungsfunktion von μ_0 ist. S_n heißt Kolmogorov-Smirnov-Statistik. Aus Satz 4.8 folgt insbesondere:

Korollar 4.12. Der Kolmogorov-Smirnov-Test hat Niveau α , falls $2e^{-2n\varepsilon^2} \geq \alpha$ gilt. Dies ist genau dann erfüllt, wenn

$$\varepsilon \geq \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)} \quad \text{bzw.} \quad n \geq \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\alpha}\right)$$

B) LILLIEFORS-TEST AUF NORMALVERTEILUNG

Um zu testen, ob die zugrunde liegende Verteilung eine Normalverteilung ist, kann man eine Variante des Kolmogorov-Smirnov-Tests verwenden. Dabei werden die Parameter geschätzt, und die Kolmogorov-Smirnov-Statistik zu den geschätzten Parametern betrachtet. Wir betrachten die Nullhypothese

$$H_0 : \mu \in \{\mu_{m,\sigma^2} : m \in \mathbb{R}, \sigma > 0\}$$

Hierbei ist $\mu_{m,\sigma^2} \sim N(m, \sigma^2)$ die Normalverteilung, und Φ_{m,σ^2} deren Verteilungsfunktion. Die Nullhypothese wird verworfen, falls

$$\tilde{S}_n := \sup_{c \in \mathbb{R}} |F_n(c) - \Phi_{\bar{X}_n, V_n}(c)| \geq \varepsilon$$

für einen Schwellenwert $\varepsilon > 0$ gilt. Wesentlich ist, dass aufgrund der Skalierungseigenschaften der Normalverteilungen die Verwerfungswahrscheinlichkeit unter der Nullhypothese nicht von den unbekanntem Parametern m und σ abhängt.

Lemma 4.13 (Skaleninvarianz). Die Verwerfungswahrscheinlichkeit $\beta = \mathbb{P}_{m,\sigma^2}[\tilde{S}_n \geq \varepsilon]$ hängt nicht von m und σ ab.

Beweis. Es gilt $X_i = \sigma Z_i + m$ mit unabhängigen unter \mathbb{P}_{m,σ^2} standardnormalverteilten Zufallsvariablen Z_i , und damit $\bar{X}_n = \sigma \bar{Z}_n + m$ und $V_n = \sigma^2 V_n^Z$. Für die Verteilungsfunktion erhalten wir

$$F_n(c) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq c} = \frac{1}{n} \sum_{i=1}^n 1_{Z_i \leq \frac{c-m}{\sigma}} = F_n^Z\left(\frac{c-m}{\sigma}\right)$$

$$\Phi_{\bar{X}_n, V_n}(c) = \Phi_{0,1}\left(\frac{c - \bar{X}_n}{\sqrt{V_n}}\right) = \Phi_{0,1}\left(\frac{c - m - \sigma \bar{Z}_n}{\sigma \sqrt{V_n^Z}}\right) = \Phi_{\bar{Z}_n, V_n^Z}\left(\frac{c-m}{\sigma}\right)$$

und damit

$$\tilde{S}_n = \sup |F_n - \Phi_{\bar{X}_n, V_n}| = \sup |F_n^Z - \Phi_{\bar{Z}_n, V_n^Z}| =: \tilde{S}_n^Z$$

Wegen $Z_i \sim N(0, 1)$ folgt:

$$\mathbb{P}_{m,\sigma^2}[\tilde{S}_n \geq \varepsilon] = \mathbb{P}_{m,\sigma^2}[\tilde{S}_n^Z \geq \varepsilon] = \mathbb{P}_{0,1}[\tilde{S}_n \geq \varepsilon]$$

Ist s ein beobachteter Wert der Statistik \tilde{S}_n , dann ist nach Lemma 4.11 $p = \mathbb{P}_{0,1}[\tilde{S}_n \geq s]$ der rechtsseitige p -Wert des Tests mit zusammengesetzter Nullhypothese H_0 . Wie in Lemma 4.4 können wir diesen nicht explizit berechenbaren p -Wert durch den entsprechenden Monte-Carlo- p -Wert ersetzen und erhalten so einen Hypothesentest zu einem vorgegebenen Niveau α .

4. Empirische Verteilungen

c) ANDERSON-DARLING-TEST

Ein Nachteil des Kolmogorov-Smirnov- bzw. Lilliefors-Tests ist, dass wir an beliebigen Stellen eine Abweichung ε der Verteilungsfunktion und der empirischen Verteilungsfunktion tolerieren. Dies gilt auch in den Tails, das heißt, wenn der Wert von $F(c)$ nahe bei 0 liegt, obwohl in diesem Fall die relative Abweichung sehr groß ist. Eine Abweichung der Verteilung in den Tails wird daher eventuell nicht erkannt. Eine mögliche Alternative, bei der die Tails der Verteilung entsprechend stärker gewichtet werden, ist die Anderson-Darling-Teststatistik

$$W_n = n \int_{-\infty}^{\infty} \frac{F_n(c) - F_0(c)}{F_0(c)(1 - F_0(c))} \mu_0(dc)$$

wobei μ_0 die Verteilung ist, auf die getestet wird, und F_0 die Verteilungsfunktion von μ_0 ist. Indem man das Integral als Summe der Integrale über die Intervalle $[X_{(i)}, X_{(i+1)}]$ schreibt und partiell integriert, kann man nachrechnen, dass

$$W_n = -n - \sum_{j=1}^n \frac{2j-1}{n} (\log(F_0(X_{(j)})) + \log(1 - F_0(X_{(n-j+1)})))$$

gilt. Um auf Normalverteilung zu testen, schätzt man wieder die Parameter und verwendet die entsprechend modifizierte Teststatistik \tilde{W}_n .

Neben dem Shapiro-Wilk-Test, den wir hier nicht besprechen, sind der Lilliefors- und der Anderson-Darling-Test häufig verwendete Normalitätstests. Auch die graphische Analyse mit QQ-Plots wird oft verwendet, siehe unten. Dabei ist zu beachten, dass keiner dieser Anpassungstests perfekt ist, da jeder Test nur auf einer bestimmten Teststatistik basiert. Es ist daher sinnvoll, verschiedene Normalitätstests durchzuführen, bevor man von einer Normalverteilungsannahme ausgeht.

4.5. Robuste Verfahren

„Viele der weiter oben betrachteten statistischen Verfahren haben einen gravierenden Nachteil: Sie sind sehr anfällig für Ausreißer in den Daten oder kleine Abweichungen in den Modellannahmen. Zum Beispiel kann der Stichprobenmittelwert \bar{X}_n durch einen einzigen falschen Datenwert beliebig stark verzerrt werden. Robuste Verfahren ändern sich dagegen nur begrenzt, wenn ein hinreichend kleiner Anteil der Daten verändert wird. Ein Beispiel für einen sehr robusten Schätzer ist der Stichprobenmedian. Ein Austausch eines einzelnen Datenwertes verschiebt diesen höchstens zur nächstgrößeren oder nächstkleineren Ordnungsstatistik.

Konfidenzintervalle für Quantile

Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion F . Sei $\gamma \in (0, 1)$ und sei q_γ ein (beliebiges) γ -Quantil von μ . Gegeben seien unabhängige Stichprobenwerte X_1, \dots, X_n mit Verteilung μ . Ein Plug-in-Schätzer für q_γ ist das Stichprobenquantil

$$\hat{q}_\gamma = \begin{cases} X_{(\lceil n\gamma \rceil)} & \text{falls } n\gamma \notin \mathbb{N}, \\ (X_{(n\gamma)} + X_{(n\gamma+1)})/2 & \text{falls } n\gamma \in \mathbb{N} \end{cases}$$

Wir wollen ein Konfidenzintervall für q_γ bestimmen, das von zwei Ordnungsstatistiken begrenzt wird. Dazu berechnen wir die Verteilung der Ordnungsstatistiken.

Satz 4.14. Sei $1 \leq k < l \leq n$. Dann gilt:

1) Die Ordnungsstatistik $X_{(k)}$ hat die Verteilungsfunktion

$$F_k(c) = \text{Bin}(n, F(c)) [\{k, k+1, \dots, n\}] = \sum_{j=k}^n \binom{n}{j} F(c)^j (1-F(c))^{n-j} \quad (**)$$

2) Für jede Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gilt

$$\mathbb{P}_\mu [q_\gamma \in [X_{(k)}, X_{(l)}]] \geq \text{Bin}(n, \gamma) [\{k, \dots, l-1\}] \quad (***)$$

mit Gleichheit, falls die Verteilungsfunktion von μ stetig ist.

Beweis. Die Ereignisse $\{X_i \leq c\}$ mit $1 \leq i \leq n$ sind unter \mathbb{P}_μ unabhängig mit Wahrscheinlichkeit $F(c)$. Also gilt:

$$\begin{aligned} F_k(c) &= \mathbb{P}_\mu [X_{(k)} \leq c] \\ &= \mathbb{P}_\mu [X_i \leq c \text{ für mindestens } k \text{ verschiedene } i \in \{1, \dots, n\}] \\ &= \text{Bin}(n, F(c)) [\{k, k+1, \dots, n\}] \end{aligned}$$

Wegen $F(q_\gamma) \geq \gamma$ und $F(q_{\gamma-}) \leq \gamma$ erhalten wir

$$\begin{aligned} \mathbb{P}_\mu [X_{(k)} \leq q_\gamma \leq X_{(l)}] &= F_k(q_\gamma) - F_l(q_{\gamma-}) \\ &= \text{Bin}(n, F(q_\gamma)) [\{k, \dots, n\}] - \text{Bin}(n, F(q_{\gamma-})) [\{l, \dots, n\}] \\ &\leq \text{Bin}(n, \gamma) [\{k, \dots, l-1\}] \end{aligned}$$

Ist F stetig, dann ist $F(q_\gamma) = F(q_{\gamma-})$ und es ergibt sich Gleichheit. ■

Bemerkenswert ist, dass die rechte Seite in (***) nicht von μ abhängt. Daher erhalten wir ein Konfidenzintervall für q_γ ohne eine Annahme an die zugrundeliegende Verteilung μ . Sei $F_{n,\gamma}$ die Verteilungsfunktion der Binomialverteilung $\text{Bin}(n, \gamma)$.

Korollar 4.15. Sei $\alpha \in (0, 1)$ und seien $1 \leq k < l \leq n$ mit

$$F_{n,\gamma}(l-1) - F_{n,\gamma}(k-1) \geq 1 - \alpha$$

Dann ist $[X_{(k)}, X_{(l)}]$ ein Konfidenzintervall für q_γ zum Niveau $1 - \alpha$.

Beispiel (Konfidenzintervall für den Median). Sei $n = 129$ und $\gamma = 1/2$. Für die Verteilungsfunktion von $\text{Bin}(n, \gamma)$ gilt dann:

$$F_{n,\gamma}(52) < 2.5\% \quad \text{und} \quad F_{n,\gamma}(76) > 97.5\%$$

Also ist $[X_{(53)}, X_{(77)}]$ ein verteilungsunabhängiges Konfidenzintervall für den Median zum Niveau 95%.

"

Testproblem für den Median

Auf ähnliche Weise wie oben können wir auch robuste Testverfahren herleiten. Seien $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$ unabhängige, identisch verteilte Beobachtungspaare ("gepaarte" bzw. "verbundene" Stichprobenwerte). Beispielsweise sind Y_i und Z_i die Ernteerträge auf denselben Flächen bei verschiedenen Arten von Saatgut. Die Differenzen $X_i = Y_i - Z_i$ ($i = 1, \dots, n$) sind dann ebenfalls unabhängig mit einer identischen Verteilung μ . Wir wollen testen, ob die X_i tendenziell eher größer als 0 sind (das erste Saatgut liefert also höhere Erträge). Dafür gibt es mehrere Optionen, zum Beispiel:

Option 1: **t-Test** Wir betrachten das Testproblem

$$H_0 : \text{Mittelwert } (\mu) = 0 \quad H_1 : \text{Mittelwert } (\mu) > 0$$

und verwerfen die Nullhypothese zum Niveau α , falls

$$T = \frac{\bar{X}_n}{\sqrt{V_n^*/n}} > q_{1-\alpha, t(n-1)}$$

Problematisch ist dabei, dass die Niveaubedingung nur bei einer Normalverteilungsannahme exakt erfüllt ist, und dass Ausreißer den Wert von T sehr stark beeinflussen können. Daher suchen wir nach robusteren Testverfahren.

Option 2: **Median-Test** Hier betrachten wir das Testproblem

$$H_0 : \text{Median } (\mu) = 0 \quad H_1 : \text{Median } (\mu) > 0$$

Wir verwerfen H_0 , falls $X_{(k)} > 0$ für ein bestimmtes $k \in \mathbb{N}$. Um ein geeignetes k zu bestimmen, können wir Satz 4.11 anwenden. Danach gilt

$$\mathbb{P}_\mu[q_{1/2} < X_{(k)}] = 1 - \mathbb{P}_\mu[q_{1/2} \geq X_{(k)}] \geq F_{n,1/2}(k-1)$$

Unter H_0 ist $q_{1/2} = 0$. Also wählen wir: $k = q_{\alpha, \text{Bin}(n, 1/2)}$, um sicherzustellen, dass die Niveaubedingung eingehalten wird. Der entsprechende Test heißt *Pearsons Vorzeichentest*.

Beispiel. Für $n = 129$ und $\alpha = 2.5\%$ verwerfen wir H_0 , falls $X_{(53)} > 0$, siehe das entsprechende Konfidenzintervall oben.

Bemerkung. Die Bedingung $X_{(k)} > 0$ ist äquivalent zu $\sum_1^n 1_{X_i > 0} \geq n - k + 1$, das heißt, wir verwerfen H_0 , falls mindestens $n - k + 1$ Differenzen ein positives Vorzeichen haben.

Der Vorzeichentest ist verteilungsunabhängig, und das Testergebnis wird nur wenig durch Ausreißer beeinflusst. Ein Nachteil ist, dass die Größe der X_i überhaupt keine Rolle für die Testentscheidung spielt, sondern nur deren Vorzeichen. Der folgende Test ist ein Kompromiss, der auch die Größe der Werte in begrenztem Umfang berücksichtigt, aber trotzdem gute Robustheitseigenschaften hat.

Option 3: **Wilcoxon Signed-Rank-Test**

Definition 4.16. Sind X_1, \dots, X_n reellwertig, dann ist der *Rang* von X_i unter X_1, \dots, X_n definiert als

$$R_i := \frac{j+k}{2} \quad \text{falls} \quad X_i = X_{(j)} \quad \text{und} \quad X_{(j-1)} < X_{(j)} \leq \dots \leq X_{(k)} < X_{(k+1)}$$

gilt.

Sind die Werte X_1, \dots, X_n alle verschieden, dann gilt $R_i = j$ genau dann, wenn $X_i = X_{(j)}$.
Wir betrachten nun die Nullhypothese

$$H_0 : \text{Verteilung}(X_i) = \text{Verteilung}(-X_i)$$

Um diese zu testen, betrachten wir die Ränge \tilde{R}_i der Absolutbeträge $|X_1|, \dots, |X_n|$. Die Entscheidungsregel in *Wilcoxon's Signed-Rank-Test* lautet:

$$\text{Verwerfe } H_0, \text{ falls } \sum_{i=1}^n 1_{X_i > 0} \tilde{R}_i > c$$

für einen Schwellenwert $c \in (0, \infty)$. Betragsmäßig größere Werte werden dabei stärker berücksichtigt. Mehr zu diesem Test siehe Übungen bzw. [Dümbgen, Kapitel 4.3].

Robuste Schätzer

Als Nächstes überlegen wir, wie man systematischer robuste Schätzer konstruieren kann. Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und X eine Zufallsvariable mit Verteilung μ .

Lemma 4.17 (Median und Mittelwert als beste Prognosewerte). 1) Die Zahl $m \in \mathbb{R}$ ist genau dann ein Median bzw. Mittelwert von μ , wenn m ein Minimierer der Funktion

$$H(r) = \mathbb{E}[|X - r|] \quad \text{bzw.} \quad H(r) = \mathbb{E}[|X - r|^2] \quad \text{ist}$$

2) Die Zahl $\hat{m} \in \mathbb{R}$ ist genau dann ein Stichprobenmedian bzw. Stichprobenmittelwert von X_1, \dots, X_n , wenn \hat{m} ein Minimierer von

$$\hat{H}(r) = \frac{1}{n} \sum_{i=1}^n |X_i - m| \quad \text{bzw.} \quad \hat{H}(r) = \frac{1}{n} \sum_{i=1}^n |X_i - m|^2 \quad \text{ist}$$

Beweis. Die zweite Aussage folgt durch Anwenden der ersten Aussage auf die empirische Verteilung $\mu = L_n$. Wir zeigen nun die erste Aussage für den Median, die entsprechende Aussage für den Mittelwert ist eine Übungsaufgabe. Sei also $r < s$. Dann gilt nach dem Hauptsatz der Differential- und Integralrechnung und dem Satz von Fubini:

$$\begin{aligned} H(s) - H(r) &= \mathbb{E}[|X - s| - |X - r|] = \mathbb{E}\left[-\int_r^s \text{sgn}(X - u) \, du\right] \\ &= -\int_r^s \mathbb{E}[\text{sgn}(X - u)] \, du = \int_r^s (\mathbb{P}[X \geq u] - \mathbb{P}[X > u]) \, du \\ &= \int_r^s (2F(u) - 1) \, du \end{aligned}$$

wobei wir $\text{sgn}(x) := 1$ für $x > 0$ und -1 für $x \leq 0$ gesetzt haben. Es folgt, dass H bei m genau dann minimal ist, wenn $F(m) \geq \frac{1}{2}$ und $F(m_-) \leq \frac{1}{2}$ gilt, das heißt, wenn m ein Median ist. ■

Da Mittelwert und Median beide Minimierer von gewissen Funktionen sind, liegt es nahe, andere Schätzer zu betrachten, die ähnliche Funktionen minimieren. Sei dazu

$$\varrho : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty), (x, v) \mapsto \varrho(x, v)$$

eine nicht-negative Funktion.

Definition 4.18 (Huber M-Schätzer). Eine Statistik \hat{m} heißt *M-Schätzer*, falls \hat{m} ein globales Minimum von

$$\hat{H}(\nu) = \sum_{i=1}^n \varrho(X_i, \nu)$$

ist.

Bemerkung. 1) \hat{m} ist ein Plug-in-Schätzer für $m = \underset{\nu}{\operatorname{argmin}} \mathbb{E}[\varrho(X, \nu)]$.

2) Ist ϱ differenzierbar, dann gilt

$$0 = \hat{H}'(\hat{m}) = \sum_{i=1}^n \frac{\partial \varrho}{\partial \nu}(X_i, \hat{m})$$

Diese Gleichung kann man benutzen, um einen M-Schätzer zu berechnen. Sie ist allerdings im Allgemeinen nur numerisch lösbar, zum Beispiel mit dem Newton-Verfahren oder einem gewichteten kleinste-Quadrate-Verfahren.

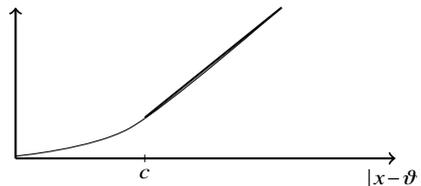
Beispiele. 1) Ist $\varrho(x, \nu) = -l(\nu, x)$ die negative log-Likelihood, dann ist \hat{H} die negative log-Likelihood des Produktmodells. In diesem Fall ist ein M-Schätzer ein Maximum-Likelihood-Schätzer.

2) Nach Lemma 4.14 sind der Median und der Mittelwert M-Schätzer zu $\varrho(x, \nu) = |x - \nu|$ bzw. $\varrho(x, \nu) = |x - \nu|^2$.

3) Als robusten Kompromiss zwischen Mittelwert und Median betrachtet man für $c > 0$ die Funktion

$$\varrho(x, \nu) = \begin{cases} \frac{1}{2}|x - \nu|^2 & \text{für } |x - \nu| \leq c \\ c(|x - \nu| - \frac{c}{2}) & \text{für } |x - \nu| \geq c \end{cases}$$

Sobald also $|x - \nu| \geq c$ verhält sich ϱ in $|x - \nu|$ linear. Das wird durch die folgende Abbildung nochmals verdeutlicht



Ein entsprechender M-Schätzer heißt *Huber-Schätzer*. Da $\frac{\partial \varrho}{\partial \nu}(x, \nu)$ für $|x - \nu| \geq c$ konstant ist, haben einzelne sehr große Werte nur einen begrenzten Einfluss auf den Huber-Schätzer.

Analog zur Aussage über Maximum-Likelihood-Schätzer (Satz von Fisher, Wilks, Wald) kann man unter geeigneten Voraussetzungen zeigen, dass auch für allgemeinere Schätzer ein zentraler Grenzwertsatz gilt. Die asymptotische Varianz ist zwar größer als die asymptotische Varianz $\frac{1}{I(\nu)}$ vom Maximum-Likelihood-Schätzer (das heißt, die Schätzer sind nicht asymptotisch effizient), aber man verliert nur einen festen Faktor bei der Effizienz und gewinnt Robustheit.

Robustheitsmaße

Wir betrachten nun einige Größen, mit denen man Robustheit quantifizieren kann. Seien $X_1, X_2, \dots : \Omega \rightarrow S$ unabhängig mit Verteilung μ , und sei

$$T_n(X_1, \dots, X_n) = g(L_n), \quad L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

der Plug-in-Schätzer für ein statistisches Funktional $g(\mu)$.

Definition 4.19 (Bruchpunkt, Sensitivität, Einflussfunktion). 1) Der *Bruchpunkt* (*breaking point*) von T_n bei (x_1, \dots, x_n) ist definiert als

$$\varepsilon_n(x_1, \dots, x_n) := \frac{1}{n} \max\{k \in \{0, 1, \dots, n\} : \sup_{y_i \neq x_i \text{ max } k \text{ mal}} |T_n(y_1, \dots, y_n)| < \infty\}$$

2) Der *asymptotische Bruchpunkt* ist definiert durch

$$\varepsilon := \liminf_{n \rightarrow \infty} \varepsilon_n(X_1, \dots, X_n)$$

3) Die *Sensitivitätsfunktion* von T_n bei (x_1, \dots, x_{n-1}) ist definiert durch

$$s(x) := \frac{T_n(x_1, \dots, x_{n-1}, x) - T_n(x_1, \dots, x_{n-1})}{1/n}$$

4) Die *Einflussfunktion* des statistischen Funktionals g ist

$$\text{IF}(g; x) := \lim_{h \downarrow 0} \frac{g((1-h)\mu + h\delta_x) - g(\mu)}{h}$$

Bemerkung. 1) Der Bruchpunkt ist der maximale Anteil der Daten, der verändert werden kann, ohne dass T_n beliebig stark verfälscht wird.

2) Ein von Null verschiedener asymptotischer Bruchpunkt entspricht Robustheit.

3) Die Sensitivitätsfunktion misst den Einfluss eines zusätzlichen Beobachtungswertes x auf den Schätzer.

4) Die Einflussfunktion ist die Ableitung des Funktionals g in Richtung $\delta_x - \mu$. Sie misst den Einfluss einer Störung der Verteilung durch ein Dirac-Maß auf g .

5) Für große n gilt

$$s(x) = \frac{g\left(\left(1 - \frac{1}{n}\right)L_{n-1} + \frac{1}{n}\delta_x\right) - g(L_{n-1})}{1/n} \approx \text{IF}(L_{n+1}; x)$$

Beispiele (Robustheit bei Lage-, Skalen- und Formparametern). 1) Ein *Lageparameter* ist eine Statistik $T_n(X_1, \dots, X_n)$, die

$$T_n(a + bX_1, \dots, a + bX_n) = a + bT(X_1, \dots, X_n)$$

für alle $a \in \mathbb{R}$ und $b > 0$ erfüllt. Beispiele sind:

a) *Mittelwert* \bar{X}_n : Hier ist $\varepsilon_n = \frac{1}{n}$ und $\varepsilon = 0$. Der Mittelwert ist nicht robust - er kann durch einen Ausreißer beliebig stark verfälscht werden. Es gilt

$$s(x) = x - \bar{x}_{n-1} \quad \text{und} \quad \text{IF}(x) = x - m(\mu)$$

b) *Median* $M_n = \hat{q}_{1/2}$: Hier ist $\varepsilon_n = \frac{\lfloor \frac{n-1}{2} \rfloor}{n}$ und $\varepsilon = \frac{1}{2}$. Das heißt, man kann bis zu der Hälfte der Werte verändern, ohne dass der Median beliebig stark verfälscht wird. Für ungerade n gilt $M_n = \frac{1}{2}(X_{(\frac{n+1}{2})} + X_{(\frac{n-1}{2})})$. Die Sensitivitätsfunktion wächst zwischen $X_{(\frac{n-1}{2})}$ und $X_{(\frac{n+1}{2})}$ linear an, ist aber außerhalb dieses Intervalls konstant und damit insbesondere beschränkt. Auch die Einflussfunktion ist für $x < q_{1/2}$ und $x > q_{1/2}$ jeweils konstant und springt nur an der Stelle $q_{1/2}$, siehe Übung.

4. Empirische Verteilungen

- c) *Getrimmter Mittelwert*: Ein weiterer Kompromiss zwischen Mittelwert und Median ist es, einen gewissen Anteil von Extremwerten wegzulassen und die verbliebenen Werte zu mitteln. Für $\tau \in (0, 1/2)$ definiert man:

$$\bar{X}_\tau := \frac{2}{n-2k} \sum_{i=k+1}^{n-k} X_{(i)} \quad \text{mit } k := \lfloor \tau n \rfloor$$

Der Schätzer ist robust mit asymptotischem Bruchpunkt τ und beschränkter Sensitivitätsfunktion.

- d) *Huber-Schätzer*: Man kann zeigen, dass der Huber-Schätzer robust ist mit asymptotischem Bruchpunkt $\frac{1}{2}$ und beschränkter Sensitivität.
- 2) Ein *Skalenparameter* ist eine Statistik $T_n(X_1, \dots, X_n)$, die

$$T_n(a - bX_1, \dots, a + bX_n) = bT_n(X_1, \dots, X_n)$$

für alle $a \in \mathbb{R}$ und $b > 0$ erfüllt. Beispiele sind:

- a) *Standardabweichung*:

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} = \sqrt{\frac{1}{n(n-1)} \sum_{i < j} (X_i - X_j)^2}$$

Der asymptotische Bruchpunkt ist 0 und die Sensitivitätsfunktion ist unbeschränkt. Die Standardabweichung ist nicht robust.

- b) *Gini-Parameter*:

$$G_n = \frac{1}{\binom{n}{2}} \sum_{i < j} |X_i - X_j|$$

Auch hier gilt $\varepsilon = 0$ und $s(x)$ ist unbeschränkt.

- c) *Median absolute deviation*: Dies ist eine robuste Alternative zur Standardabweichung:

$$MAD_n := \text{Median}_n(|X_1 - M_n|, \dots, |X_n - M_n|)$$

Asymptotischer Bruchpunkt und Sensitivitätsfunktion siehe Übungsaufgabe.

- d) *Spannweite*:

$$R_n = X_{(n)} - X_{(1)}$$

Auch die Spannweite ist nicht robust - es gilt: $\varepsilon_n = \frac{1}{n}$ und $\varepsilon = 0$.

- e) *Interquartilsabstand*:

$$IQR_n = \hat{q}_{0.75} - \hat{q}_{0.5}$$

Diese Größe ist robust mit asymptotischem Bruchpunkt $\frac{1}{4}$.

- 3) Ein *Formparameter* erfüllt:

$$T_n(a + bX_1, \dots, a + bX_n) = T_n(X_1, \dots, X_n)$$

für alle $a \in \mathbb{R}$ und $b > 0$, das heißt, T_n ist *affin invariant*. Nicht-robuste Beispiele sind:

- a) *Schiefe*:

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{S_n} \right)^3$$

Diese misst die Asymmetrie der Verteilung.

- b) *Kurtose*:

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{S_n} \right)^4 - 3$$

Diese misst den Anteil der Masse in Extrembereichen. Beim Jarque-Bera-Test verwendet man eine Kombination von Schiefe und Kurtose als Teststatistik, um auf Normalverteilung zu testen.

4.6. Graphische Überprüfung von Verteilungsfunktionen

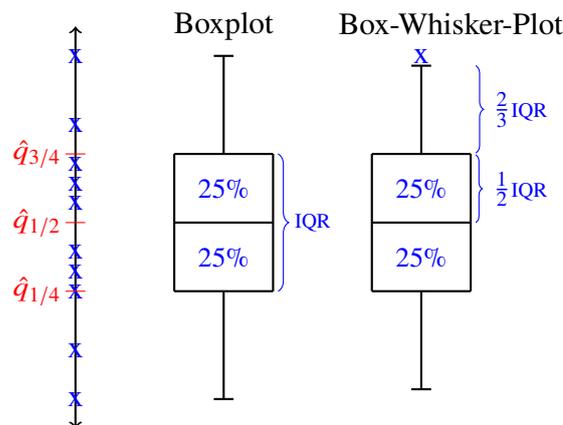
Sei μ eine unbekannte Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir nehmen der Einfachheit halber an, dass die Verteilungsfunktion F stetig ist. Dann gilt für unabhängige Zufallsvariablen X_1, \dots, X_n mit Verteilung μ :

$$\mathbb{P}_\mu[\exists i \neq j : X_i = X_j] = 0$$

Das heißt, mit Wahrscheinlichkeit 1 sind die Werte alle verschieden, und für die Ordnungsstatistiken gilt $X_{(1)} < X_{(2)} < \dots < X_{(n)}$.

Boxplots

Um einen ersten Überblick über die Daten zu erhalten, kann man einen *Boxplot* oder einen *Box-Whisker-Plot* erstellen.



In beiden Plots werden die Quantile $\hat{q}_{1/4}$, $\hat{q}_{1/2}$ und $\hat{q}_{3/4}$ aufgetragen, im Boxplot zudem die Spannweite. Im Box-Whisker-Plot werden dagegen Datenwerte außerhalb des Intervalls $[\hat{q}_{1/4} - \frac{3}{2} IQR, \hat{q}_{1/4} + \frac{3}{2} IQR]$ separat aufgetragen.

Quantil-Quantil-Plots

Um die Nullhypothese $H_0 : \text{Verteilung}(X_i) = \mu$ grafisch zu testen, kann man die Stichprobenquantile mit den theoretisch zu erwartenden Quantilen vergleichen. Seien U_1, \dots, U_n unabhängige auf $(0, 1)$ gleichverteilte Zufallsvariablen. Unter der Nullhypothese gilt:

$$(X_1, \dots, X_n) \sim (G(U_1), \dots, G(U_n)) \quad (X_{(1)}, \dots, X_{(n)}) \sim (G(U_{(1)}), \dots, G(U_{(n)}))$$

wobei $G(u) = \inf\{c \in \mathbb{R} : F(c) \geq u\}$ die untere Quantilsfunktion ist. Die Verteilung der Ordnungsstatistik $U_{(*)}$ wird in der Übung berechnet. Insbesondere gilt:

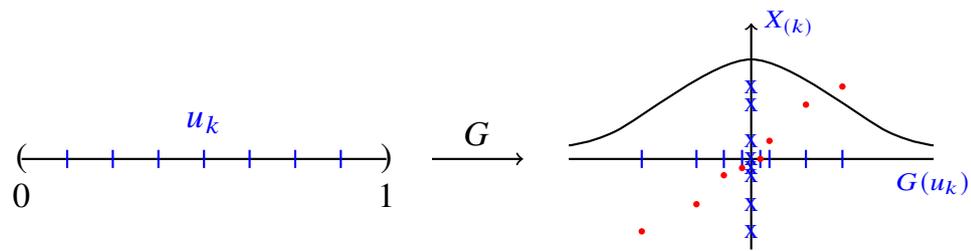
$$\mathbb{E}[U_{(k)}] = \frac{k}{n+1} =: u_k, \quad \sigma[U_{(k)}] = \sqrt{\frac{u_k(1-u_k)}{n+2}}$$

also $U_{(k)} \approx u_k$ für große n , und damit

$$(X_{(1)}, \dots, X_{(n)}) \approx (G(u_1), \dots, G(u_n))$$

Einfacher QQ-Plot

Ein QQ-Plot ist ein Streudiagramm der Werte $(G(u_k), X_{(k)})$, $k = 1, \dots, n$. Die Werte $G(u_k)$ sind Quantile der theoretischen Verteilung, und die Werte $X_{(k)}$ entsprechen Quantilen der empirischen Verteilung. Gilt die Nullhypothese $X_i \sim \mu$, dann sollten die Punkte „in der Nähe“ der Diagonalen liegen.



QQ-Plot mit Parameterschätzung

Will man testen, ob die Verteilung der X_i aus einer parametrischen Familie $\{\mu_\vartheta : \vartheta \in \Theta\}$ kommt, dann kann man zunächst ϑ schätzen, und dann einen QQ-Plot mithilfe der Quantilfunktion $G_{\hat{\vartheta}} = F_{\hat{\vartheta}}^{-1}$ erstellen.

Normal-QQ-Plot

Um auf Normalverteilung zu testen, gibt es eine einfachere Möglichkeit, die die Skalierungseigenschaften der Normalverteilung ausnutzt. Diese bilden eine *Lokations-Skalenfamilie*, das heißt:

$$\phi_{m,\sigma^2}(c) = \phi_{0,1}\left(\frac{c-m}{\sigma}\right) \quad \text{für alle } c \in \mathbb{R}$$

Für die Quantilsfunktionen folgt:

$$\phi_{m,\sigma^2}^{-1}(u) = m + \sigma\phi^{-1}(u) \quad \text{für alle } u \in (0, 1)$$

Ein *Normal-QQ-Plot* ist ein Streudiagramm der Werte

$$\left(\phi_{0,1}^{-1}(u_k), X_{(k)}\right), k = 1, \dots, n$$

Gilt $X_i \sim N(m, \sigma^2)$, dann folgt für große n :

$$X_{(k)} \approx \phi_{m,\sigma^2}^{-1}(u_k) = m + \sigma\phi^{-1}(u_k)$$

Das heißt, die Punkte liegen „nahe“ der Geraden $y = m + \sigma x$. Ein analoges Verfahren funktioniert für eine beliebige Lokations-Skalenfamilie.

PP-Plot

In einem PP-Plot werden statt der Werte $(G(u_k), X_{(k)})$ die Werte $(u_k, F(X_{(k)}))$ aufgetragen. Auch diese sollten nach der Diagonalen liegen, aber im QQ-Plot sind Abweichungen meist besser erkennbar.

Normale bzw. auffällige Abweichungen

Um nun die Diagramme zu quantifizieren, ist es wichtig, zwischen normalen bzw. auffälligen Abweichungen von der Diagonalen / Geraden zu differenzieren. Dies kann man auf verschiedene Arten testen. Eine einfache Möglichkeit ist es, zum Vergleich eine Stichprobe $(\tilde{X}_1, \dots, \tilde{X}_n)$ von der Verteilung μ zu simulieren, und die QQ-Plots für X und \tilde{X} zu vergleichen. Alternativ kann man Anpassungstests durchführen, die auf dem QQ-Plot basieren. Um auf Normalverteilung zu testen, kann man zum Beispiel eine Teststatistik der Form

$$T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \left| \frac{X_{(i)} - \hat{m}}{\hat{\sigma}} - \phi^{-1}(u_k) \right|$$

betrachten. Ist der Schätzer \hat{m} ein Lageparameter (zum Beispiel Mittelwert, Median, getrimmter Mittelwert) und $\hat{\sigma}$ ein Skalenparameter (zum Beispiel Standardabweichung, *MAD*), dann ist $T(X_1, \dots, X_n)$ ein Formparameter. Hieraus folgt, dass die Verteilung nicht von m und σ abhängt.

5. Zusammenhang mehrerer Merkmale

Bisher haben wir meist einzelne kategorielle oder numerische Merkmale betrachtet. In den meisten statistischen Problemen geht es jedoch um den Zusammenhang verschiedener Größen, also um multivariable Verfahren.

Beispiel (Zusammenhang von Haar- und Augenfarbe). Eine Stichprobe liefert die folgende Häufigkeitsverteilung für verschiedene Kombinationen von Haar- und Augenfarbe. Was lässt sich daraus über den Zusammenhang dieser beiden Merkmale schließen?

| | | AUGENFARBE | | | |
|-----------|---------|------------|------|-------|------|
| | | Braun | Blau | Hazel | Grün |
| HAARFARBE | Schwarz | 68 | 20 | 15 | 5 |
| | Braun | 119 | 84 | 54 | 29 |
| | Rot | 26 | 17 | 14 | 14 |
| | Blond | 7 | 94 | 10 | 16 |

5.1. Binäre Merkmale: Chancenquotienten und Vierfeldertafeln

Das folgende Kapitel basiert zu großen Teilen auf Kapitel 7 des Buches "Einführung in die Statistik" von L. Dümbgen. Wir betrachten zunächst den Fall, dass es für jedes Merkmal nur zwei mögliche Werte gibt (0 und 1). Im Beispiel oben könnten wir etwa die Fragestellung darauf reduzieren, ob es einen Zusammenhang zwischen roten Haaren und grünen Augen gibt, und dementsprechend bei der Haar- und Augenfarbe nur zwischen *rot/nicht rot* sowie *grün/nicht grün* unterscheiden. Ist A ein Ereignis mit $\mathbb{P}[A] < 1$, dann definieren wir

$$\text{Chancen}[A] := \frac{\mathbb{P}[A]}{\mathbb{P}[A^c]} = \frac{\mathbb{P}[A]}{1 - \mathbb{P}[A]} \in [0, \infty).$$

Zum Beispiel gilt für $\mathbb{P}[A] = \frac{1}{4}$: $\text{Chancen}[A] = \frac{1}{3}$. Da die Funktion $p \mapsto \frac{p}{1-p}$ von $[0, 1)$ nach $[0, \infty)$ streng monoton wachsend und damit bijektiv ist, ist $\text{Chancen}[A]$ einfach eine nichtlineare Transformation von $\mathbb{P}[A]$.

Vierfeldertafel und Chancenquotient

Seien nun $X, Y : \Omega \rightarrow \{0, 1\}$ binäre Zufallsvariablen, und seien

$$p_{kl} := \mathbb{P}[X = k, Y = l], \quad k, l \in \{0, 1\},$$

die Gewichte der gemeinsamen Verteilung $p = (p_{kl})$. Die Randverteilungen von X und Y haben dann die Gewichte

$$p_{k+} = p_{k0} + p_{k1} = \mathbb{P}[X = k],$$

$$p_{+l} = p_{0l} + p_{1l} = \mathbb{P}[Y = l].$$

Diese theoretische Verteilung können wir auch als Vierfeldertafel darstellen:

| | | | |
|----------|----------|----------|---------------|
| p_{11} | p_{10} | p_{1+} | Vert. von X |
| p_{01} | p_{00} | p_{0+} | |
| p_{+1} | p_{+0} | 1 | |
| | | | Vert. von Y |

5. Zusammenhang mehrerer Merkmale

Wir setzen im Folgenden $p_{10} \neq 0$ und $p_{01} \neq 0$ voraus.

Definition 5.1 (Chancenquotient). Der Quotient

$$\varrho := \frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{\mathbb{P}[X = 1, Y = 1] \cdot \mathbb{P}[X = 0, Y = 0]}{\mathbb{P}[X = 1, Y = 0] \cdot \mathbb{P}[X = 0, Y = 1]}$$

heißt *Chancenquotient*.

Bemerkung. Es gilt

$$\varrho = \frac{\mathbb{P}[X = 1 | Y = 1]}{\mathbb{P}[X = 0 | Y = 1]} \bigg/ \frac{\mathbb{P}[X = 1 | Y = 0]}{\mathbb{P}[X = 0 | Y = 0]} = \frac{\text{Chancen}[X = 1 | Y = 1]}{\text{Chancen}[X = 1 | Y = 0]},$$

das heißt ϱ misst die Abhängigkeit der Chancen für $X = 1$ von Y bzw. umgekehrt die Abhängigkeit der Chancen für $Y = 1$ von X .

Lemma 5.2. *Es gilt*

$$\varrho = 1 \Leftrightarrow X, Y \text{ unabhängig} \Leftrightarrow p_{kl} = p_{k+}p_{+l} \quad \text{für alle } k, l.$$

Achtung, diese Aussage gilt nur für binäre Merkmale.

Beweis. $\varrho = 1$ ist äquivalent zu

$$\text{Chancen}[X = 1 | Y = 1] = \text{Chancen}[X = 1 | Y = 0], \quad (*)$$

was gleichbedeutend ist mit

$$\mathbb{P}[X = 1 | Y = 1] = \mathbb{P}[X = 1 | Y = 0], \quad (**)$$

also die Unabhängigkeit von $\{X = 1\}$ und Y . Da X nur die Werte 0 oder 1 annehmen kann, ist dies äquivalent zur Unabhängigkeit von X und Y . ■

Schätzer und Konfidenzintervall für ϱ

Wir wollen nun die Chancenquotienten benutzen, um zu testen, ob X und Y unabhängig sind. Dazu schätzen wir zunächst ϱ . Seien $(X_1, Y_1), \dots, (X_N, Y_N)$ unabhängige Kopien von (X, Y) , und sei

$$H_{kl} = |\{1 \leq i \leq N : X_i = k, Y_i = l\}|$$

die Häufigkeit von (k, l) in der Stichprobe. Im Folgenden werden diese empirischen Häufigkeiten in einer empirischen Verteilung dargestellt.

| | | | |
|------------------------------|----------|----------|------------------------------|
| H_{11} | H_{10} | H_{1+} | Häufigkeit der Werte von X |
| H_{01} | H_{00} | H_{0+} | |
| H_{+1} | H_{+0} | N | |
| Häufigkeit der Werte von Y | | | |

Als Plug-in-Schätzer für ϱ erhalten wir den empirischen Chancenquotienten

$$\hat{\varrho} = \frac{\frac{H_{11}}{N} \cdot \frac{H_{00}}{N}}{\frac{H_{10}}{N} \cdot \frac{H_{01}}{N}} = \frac{H_{11}H_{00}}{H_{10}H_{01}}.$$

Als nächsten Schritt wollen wir Konfidenzintervalle für ϱ herleiten, die auf den Statistiken $H_{kl}, k, l \in \{0, 1\}$, basieren. Die gemeinsamen Verteilungen dieser Statistiken ist eine Multinomialverteilung:

$$(H_{11}, H_{10}, H_{01}, H_{00}) \sim \text{Mult}(N, p_{11}, p_{10}, p_{01}, p_{00}).$$

Uns interessieren aber nicht alle Parameter p_{kl} , sondern nur der Chancenquotient ϱ . Wenn wir ϱ kennen, dann können wir die bedingte Verteilung von H_{11} gegeben H_{1+} und H_{+1} berechnen. Umgekehrt können wir ein Konfidenzintervall für ϱ erhalten, das auf H_{11}, H_{1+} und H_{+1} basiert. Das Vorgehen ist analog zur Herleitung von Fishers exaktem Test in Kapitel 1.

Lemma 5.3. *Es existiert eine Konstante $C_{n,l,N,\varrho} > 0$, die nicht von x abhängt, so dass für alle $l, n \in \{0, \dots, N\}$ und $x \in \{0, \dots, \min(l, n)\}$.*

$$\mathbb{P}[H_{11}|H_{+1} = l, H_{1+} = n] = C_{n,l,N,\varrho}^{-1} \binom{n}{x} \binom{N-n}{l-x} \varrho^x \quad (*)$$

Beweis. Es gilt

$$\begin{aligned} \mathbb{P}[H_{11}|H_{+1} = l, H_{1+} = n] &= \frac{\mathbb{P}[H_{11} = x, H_{01} = l - x, H_{10} = n - x, H_{00} = N - l - n + x]}{\mathbb{P}[H_{+1} = l, H_{1+} = n]} \\ &\propto \frac{N!}{x!(l-x)!(n-x)!(N-l-n+x)!} p_{11}^x p_{01}^{l-x} p_{10}^{n-x} p_{00}^{N-l-n+x} \\ &\propto \binom{n}{x} \binom{N-n}{l-x} \left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)^x = \binom{n}{x} \binom{N-n}{l-x} \varrho^x \end{aligned}$$

Hierbei bedeutet \propto dass beide Seiten sich nur um eine multiplikative Konstante unterscheiden, die von n, l, N und p aber nicht abhängt. Daraus ergibt sich (*), wobei $C_{n,l,N,\varrho}$ die Normierungskonstante der bedingten Verteilung ist. ■

Mithilfe des Lemmas erhalten wir *Konfidenzintervalle* für den Chancenquotienten ϱ . Sei dazu $\alpha \in (0, 1)$, und sei $F_{\varrho,N,l,n}$ die Verteilungsfunktion der Wahrscheinlichkeitsverteilung mit den Gewichten aus (*). Nach Lemma 1.5 gilt

$$\mathbb{P}[F_{\varrho,N,H_{+1},H_{1+}}(H_{11}) \leq \alpha | H_{+1} = l, H_{1+} = n] = \mathbb{P}[F_{\varrho,N,l,n}(H_{11}) \leq \alpha | H_{+1} = l, H_{1+} = n] \leq \alpha$$

für alle $l, n \in \{0, \dots, N\}$. Nach dem Satz von der totalen Wahrscheinlichkeit folgt

$$\mathbb{P}[F_{\varrho,N,H_{+1},H_{1+}}(H_{11}) \leq \alpha] \leq \alpha \quad (**)$$

Wir betrachten nun die Funktion

$$\varrho \quad \mapsto \quad \tilde{F}_{N,H_{+1},H_{1+},H_{11}}(\varrho) := F_{\varrho,N,H_{+1},H_{1+}}(H_{11}).$$

Diese ist streng monoton fallend (Übung, siehe „Dümbegen, Lemma 2.7“), also invertierbar. Daher folgt aus (**) unmittelbar die erste Aussage des folgenden Satzes

Satz 5.4 (Konfidenzschranken für Chancenquotienten).

- 1) $b_\alpha := \tilde{F}_{N,H_{+1},H_{1+},H_{11}}^{-1}(\alpha)$ ist eine obere Konfidenzschranke für ϱ zum Niveau $1 - \alpha$.
- 2) $a_\alpha := \tilde{F}_{N,H_{+1},H_{1+},H_{11}}^{-1}(1 - \alpha)$ ist eine untere Konfidenzschranke für ϱ zum Niveau $1 - \alpha$.

5. Zusammenhang mehrerer Merkmale

Beweis. Nach (**) gilt $\mathbb{P}[\varrho \geq b_\alpha] \leq \alpha$. Die zweite Aussage folgt auf ähnliche Weise. ■

Bemerkung (Fishers exakter Test). Da wir Konfidenzschranken erhalten, können wir auch die entsprechenden Testprobleme betrachten, zum Beispiel $H_0 : \varrho = 1$ vs. $H_1 : \varrho > 1$. Wir verwerfen H_0 , falls $\hat{\varrho}$ auffällig groß ist. Dies ist der Fall, wenn $a_\alpha > 1$ gilt, das heißt

$$1 - \alpha < F_{1, N, H_{+1}, H_{1+}}(H_{11} - 1).$$

Dies ist genau der rechtsseitige p -Wert bei Fishers exaktem Test.

Vierfeldertafel für modifiziertes Modell

Auch in einer etwas anderen Situation kann man analog vorgehen:

Beispiel (Randomisierte Studie, siehe Abschnitt 1.1). Eine randomisierte Studie für ein neues Medikament liefert folgende Ergebnisse

| | Besserung | keine Besserung | |
|------------|-----------|-----------------|-------|
| Medikament | H_{11} | H_{12} | n_1 |
| Placebo | H_{21} | H_{22} | n_2 |
| | H_{+1} | H_{+2} | N |

Im Unterschied zur oben betrachteten Situation sind n_1 und n_2 fest vorgegeben.

Modell: Die Zufallsgrößen $H_{11} \sim \text{Bin}(n_1, p_1)$, $H_{21} \sim \text{Bin}(n_2, p_2)$ sind unabhängig und es ergibt sich $H_{12} = n_1 - H_{11}$, $H_{22} = n_2 - H_{21}$. Die theoretische Verteilung ist nun

$$\begin{matrix} p_1 & 1 - p_1 \\ p_2 & 1 - p_2 \end{matrix}$$

Dementsprechend betrachten wir

$$\varrho := \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

Wie oben können wir ϱ schätzen durch

$$\hat{\varrho} = \frac{H_{11}H_{22}}{H_{12}H_{21}}$$

Man kann zeigen, dass Lemma 5.3 auch in dieser Situation gilt, siehe „Dümbegen, Lemma 7.2“. Damit erhält man wieder analoge Konfidenzintervalle wie oben.

5.2. Test auf Unabhängigkeit

Wir betrachten nun zwei allgemeine kategorielle Merkmale mit Werten in endlichen Mengen $S = \{a_1, \dots, a_K\}$ und $T = \{b_1, \dots, b_L\}$ mit $|S| = K$ und $|T| = L$. Seien

$$X_1, \dots, X_N : \Omega \rightarrow S \quad \text{und} \quad Y_1, \dots, Y_N : \Omega \rightarrow T$$

jeweils unabhängig und identisch verteilt mit gemeinsamer Verteilung $p = (p_{kl})$ von X_i und Y_i , das heißt

$$p_{kl} = \mathbb{P}[X_i = a_k, Y_i = b_l] \quad \text{für } 1 \leq k \leq K, \text{ und } 1 \leq l \leq L.$$

Dann sind $X = (X_1, \dots, X_n)$ und $Y = (Y_1, \dots, Y_n)$ genau dann unabhängig, wenn die gemeinsame Verteilung das Produkt der Randverteilungen ist, das heißt wenn

$$p_{kl} = p_{k+}p_{+l} =: \bar{p}_{kl} \quad \text{für alle } k, l.$$

Wir wollen nun diese Nullhypothese testen:

$$H_0 : p = \bar{p}$$

Multiple Tests

Eine Möglichkeit ist es, Fishers exakten Test für jeden festen Wert von k und l durchzuführen. Dies liefert sehr genau Informationen über den Zusammenhang von X und Y , erfordert aber eine große Anzahl von Stichproben um ein vorgegebenes Niveau zu erreichen, da wir $K \cdot L$ Test durchführen (Bonferroni- Korrektur).

Chiquadrat-Test auf Unabhängigkeit

Wir betrachten nun eine andere Option analog zu Anpassungstests: Wir schätzen die Chiquadrat Divergenz

$$\chi^2(p|\bar{p}) = \sum_{k,l} \left(\frac{p_{kl}}{\bar{p}_{kl}} - 1 \right)^2 \cdot \bar{p}_{kl} = \sum_{k,l} \frac{p_{kl}^2}{\bar{p}_{kl}} - 1$$

zwischen der Verteilung und dem Produkt der Randverteilungen. Statt der Chiquadrat-Divergenz kann man auch wieder die relative Entropie betrachten. Ein Schätzer für $N \cdot \chi^2(p|\bar{p})$ ist die entsprechende Chiquadrat-Divergenz der empirischen Verteilungen:

$$T(X, Y) = N \cdot \chi^2(L_N^{X,Y} | L_N^X \otimes L_N^Y) = \sum_{kl} \left(\frac{H_{kl}(X, Y)}{\bar{H}_{kl}(X, Y)} - 1 \right)^2 \bar{H}_{kl}(X, Y) = \sum_{kl} \frac{H_{kl}(X, Y)^2}{\bar{H}_{kl}(X, Y)} - N,$$

wobei $H_{kl}(X, Y) = \sum_{i=1}^N 1_k(X_i) 1_l(Y_i)$ die Häufigkeit von (k, l) , und

$$\bar{H}_{kl}(X, Y) := \frac{H_k(X) \cdot H_l(Y)}{N}$$

ist. Unter der Nullhypothese sollte für große N gelten:

$$\frac{H_{kl}(X, Y)}{N} \stackrel{N \text{ groß}}{\approx} p_{kl} \stackrel{H_0}{=} p_{k+p+l} \stackrel{N \text{ groß}}{\approx} \frac{H_k(X)}{N} \cdot \frac{H_l(Y)}{N} = \frac{\bar{H}_{kl}(X, Y)}{N}.$$

Die Chiquadrat-Statistik $T(X, Y)$ misst die mittlere quadratische Abweichung dieser Größen, welche wir in einer Kontingenztafel darstellen können:

| Kontingenztafel der Häufigkeiten | | | | | Erwartete Häufigkeiten im Produktfall | | | | |
|----------------------------------|----------|-----|----------|----------|---------------------------------------|----------------|-----|----------------|----------|
| | b_1 | ... | b_L | | | b_1 | ... | b_L | |
| a_1 | H_{11} | ... | H_{1L} | $H_1(X)$ | a_1 | \bar{H}_{11} | ... | \bar{H}_{1L} | $H_1(X)$ |
| \vdots | \vdots | | \vdots | \vdots | \vdots | \vdots | | \vdots | \vdots |
| a_K | H_{K1} | ... | H_{KL} | $H_K(X)$ | a_K | \bar{H}_{K1} | ... | \bar{H}_{KL} | $H_K(X)$ |
| | $H_1(Y)$ | ... | $H_L(Y)$ | N | | $H_1(Y)$ | ... | $H_L(Y)$ | N |

Analog zum χ^2 -Anpassungstest betrachten wir nun den folgenden Unabhängigkeitstest, wobei $c > 0$ ein vorzugebender Schwellenwert ist.

Chiquadrat-Test auf Unabhängigkeit: Verwerfe H_0 , falls $T(X, Y) \geq c$. Sei $t = T(x, y)$ der beobachtete Wert der χ^2 -Statistik. Eine exakte Berechnung des entsprechenden p-Werts ist normalerweise nicht möglich. Stattdessen bieten sich zwei approximative Verfahren an:

a) **Bootstrap**

Unter H_0 gilt $p_{kl} = p_{k+p+l} \approx \frac{H_{k+}}{N} \frac{H_{+l}}{N}$ für große N . Daher kann man wie folgt vorgehen, um den p-Wert zu schätzen:

- Simuliere $B \cdot N$ Stichproben vom Produkt der empirischen Verteilungen von X und Y .

5. Zusammenhang mehrerer Merkmale

- Erstelle daraus B Simulationen von Kontingenztafeln, und berechne Werte t_1, \dots, t_B der χ^2 -Statistik.
- Berechne den Monte-Carlo-p-Wert

$$\hat{p}_{MC} = \frac{|\{i \in \{1, \dots, B\} \mid t_i \geq t\}| + 1}{B + 1}$$

b) Chiquadrat-Approximation (N groß)

Man kann zeigen, dass die Verteilung von $T(X, Y)$ unter der Nullhypothese für $N \rightarrow \infty$ gegen eine χ^2 -Verteilung mit $(K - 1)(L - 1)$ Freiheitsgraden konvergiert. Damit ergibt sich ein approximativer Test, der H_0 zum Niveau α verwirft, falls

$$T(X, Y) \geq q_{1-\alpha, \chi^2((K-1)(L-1))}$$

gilt. Der approximative p-Wert ist

$$p_r \approx 1 - F_{\chi^2((K-1)(L-1))}(T(X, Y)).$$

Der vollständige Beweis der Konvergenz gegen die χ^2 -Verteilung findet sich beispielsweise in „Georgii, Satz 11.17“. Wir motivieren hier nur anschaulich, warum die Anzahl der Freiheitsgrade gleich $(K - 1)(L - 1)$ ist.

Heuristik: In diesem Fall ist der Parameterraum Θ die Menge aller Wahrscheinlichkeitsverteilungen auf $\{a_1, \dots, a_K\} \times \{b_1, \dots, b_L\}$. Diese Menge ist ein Simplex im $\mathbb{R}^{K \cdot L}$:

$$\Theta = \{p = (p_{kl}) \mid p_{kl} \geq 0 \text{ für alle } k, l, \sum p_{kl} = 1\},$$

hat also Dimension $K \cdot L - 1$. Die Parametermenge der Nullhypothese ist

$$\Theta_0 = \{p_X \otimes p_Y \mid p_X \in WV(\{a_1, \dots, a_K\}), p_Y \in WV(\{b_1, \dots, b_L\})\}.$$

Diese Menge hat Dimension $K - 1 + L - 1$. Die Anzahl der im orthogonalen Komplement von Θ_0 verbleibenden Freiheitsgrade ist also

$$\dim(\Theta) - \dim(\Theta_0) = KL - 1 - K - L + 2 = (K - 1)(L - 1).$$

Hat man dies realisiert, dann kann man die Konvergenz gegen die Verteilung $\chi^2((K - 1)(L - 1))$ auf ähnliche Weise zeigen wie beim χ^2 -Anpassungstest, siehe „Georgii, Satz 11.17“.

Die Nullhypothese der Unabhängigkeit kann man noch etwas abschwächen.

Verallgemeinerte Nullhypothese

Die alternative Nullhypothese H'_0 geht von der **Bedingten Austauschbarkeit von Y gegeben X** aus, das heißt, dass für jede Permutation $\pi \in \mathcal{S}_N$ gilt:

$$(X, \pi Y) \sim (X, Y), \quad \text{wobei } \pi Y := (Y_{\pi(1)}, \dots, Y_{\pi(N)})$$

Aus (H_0) folgt (H'_0) , denn da Y_1, \dots, Y_N nach Voraussetzung unabhängig und identisch verteilt sind, gilt $\pi Y \sim Y$ für alle $\pi \in \mathcal{S}_N$. Unter (H_0) ist X unabhängig von Y , und damit gilt auch $(X, \pi Y) \sim (X, Y)$ für alle $\pi \in \mathcal{S}_N$. Die folgenden Aussagen gelten sogar unter der schwächeren Annahme (H'_0) .

Lemma 5.5. Sei $\pi \sim \text{Unif}(\mathcal{S}_N)$ eine zufällige Permutation der Indizes. Dann gilt für alle $X = (x_1, \dots, x_N)$ und $y = (y_1, \dots, y_N)$:

- 1) $\mathbb{E}[H_{kl}(x, \pi y)] = \bar{H}_{kl}(x, y)$ für alle k, l
- 2) $\mathbb{E}[T(x, \pi y)] = \frac{N}{N - 1}(K - 1)(L - 1)$

Zum Beweis betrachtet man die Darstellung

$$H_{kl}(x, \pi y) = \sum_{i=1}^N 1_k(x_i) 1_l(y_{\pi(i)}),$$

und berechnet damit die Erwartungswerte von $H_{kl}(x, \pi y)$ und $H_{kl}(x, \pi y)^2$, siehe Übung.

Satz 5.6. Unter (H'_0) gilt:

- 1) $\mathbb{E}[E_{kl}(X, Y)] = \mathbb{E}[\bar{H}_{kl}(X, Y)]$ für alle k, L
- 2) $\mathbb{E}[T(X, Y)] = \frac{N}{N-1} (K-1)(L-1)$
- 3) $\mathbb{P}[T(X, Y) \leq c] \stackrel{N \uparrow \infty}{\rightarrow} F_{\chi^2((K-1)(L-1))}(c)$ für alle $c \in \mathbb{R}$

Die Aussagen 1) und 2) folgen aus dem Lemma, siehe Übung. Zum Beweis von 3) siehe „Georgii, Satz 11.17“.

5.3. Permutationstests

Das folgende Kapitel basiert zu großen Teilen auf dem Abschnitt 8.1 - 8-3 des Buches „Einführung in die Statistik“ von L. Dümbgen.

Wir betrachten nun ein allgemeines Verfahren um auf Symmetrien (zum Beispiel Invarianz unter Permutation) zu testen. Sei

$$(X, Y) = (X_1, \dots, X_n, Y_1, \dots, Y_m) : \Omega \rightarrow S^{n+m}$$

ein Stichprobenvektor mit Werten in X^{m+n} wobei (S, \mathcal{B}) ein messbarer Raum ist. Weiterhin sei G eine endliche Gruppe von Bijektionen $g : S^{m+n} \rightarrow S^{m+n}$. Wir wollen die Nullhypothese testen, dass die Verteilung von (X, Y) invariant unter G ist:

$$(H_0) \quad \mathbf{G\text{-Invarianz}} \quad \text{Für alle } g \in G \text{ gilt} \quad g(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim (X_1, \dots, X_n, Y_1, \dots, Y_m)$$

In diesem allgemeinen Rahmen lassen sich viele interessante Fälle einbetten.

Beispiele. 1) **Test auf Austauschbarkeit**

Sei $G = S_{n+m}$ die Gruppe aller Permutationen von $\{1, \dots, n+m\}$. Wir setzen $X_{n+k} := Y_k$ für $k = 1, \dots, m$ und definieren für $\pi \in G$:

$$\pi(X, Y) = \pi(X_1, \dots, X_{n+m}) = (X_{\pi(1)}, \dots, X_{\pi(n+m)}),$$

das heißt die Datenwerte X_i und Y_i werden permutiert. Die Nullhypothese lautet dann

$$(H_0) : \quad X_1, \dots, X_n, Y_1, \dots, Y_m \quad \text{sind austauschbar.}$$

Sie ist insbesondere erfüllt wenn diese Zufallsvariablen unabhängig und identisch verteilt sind. Einen Test von H_0 können wir daher verwenden, um zu wieder legen, dann die X_i und Y_i *identisch verteilt* sind.

2) **Test auf bedingte Austauschbarkeit**

Sei $G = S_m$. Für $\pi \in G$ setzen wir

$$\pi(X, Y) := (X, \pi Y),$$

5. Zusammenhang mehrerer Merkmale

das heißt es werden nur die Werte von Y permutiert. Die Nullhypothese lautet dann:

$$(H_0) : \quad Y_1, \dots, Y_m \quad \text{sind bedingt austauschbar gegebne } X_1, \dots, X_n.$$

Dies ist insbesondere erfüllt, wenn Y_1, \dots, Y_m austauschbar sind, und unabhängig von X_1, \dots, X_n . Mit einem Test von H_0 können wir also wiederlegen, dass X und Y *unabhängig* sind. Dies haben wir auch bereits im letzten Abschnitt verwendet

3) **Test auf Vorzeichensymmetrie** Sei $n = 0$ und

$$G = \{+1, -1\}^m = \{s = (s_i)_{i=1, \dots, m} \mid s_i \in \{+1, -1\}\}$$

mit der Komponentenweisen Multiplikation

$$s \circ \tilde{s} := (s_i \cdot \tilde{s}_i)_{i=1, \dots, m}$$

die Nullhypothese lautet nun

(H_0) Die Verteilung von $Y = (Y_1, \dots, Y_m)$ ist invariant unter beliebigen Vorzeichenwechseln der Y_i .

Einen Test von H_0 kann man zum Beispiel verwenden, um eine Tendenz für positive oder negative vorzeichen aufzuzeigen.

Die allgemeinere Nullhypothese lässt sich auch noch etwas anders formulieren:

Lemma 5.7. Sei $\pi \sim \text{Unif}(G)$ ein Zufälliges Element aus G , unabhängig von X und Y . Dann ist H_0 genau dann erfüllt, wenn

$$(H'_0)\pi(X, Y) \quad \sim \quad (X, Y).$$

beispielsweise ist Austauschbarkeit äquivalent zur Invarianz der gemeinsamen Verteilung unter einer unabhängigen Zufallspermutation.

Beweis.

„ $H_0 \Rightarrow H'_0$ “ Ist H_0 erfüllt, dann gilt $g(X, Y) \sim (X, Y)$ für alle $g \in G$. Mit dem Satz von der totalen Wahrscheinlichkeit folgt:

$$\mathbb{P}[\pi(X, Y) \in B] = \sum_{g \in G} \frac{1}{|G|} \mathbb{P}[g(X, Y) \in B] = \mathbb{P}[(X, Y) \in B]$$

für alle $B \in \mathcal{B}$.

„ $H'_0 \Rightarrow H_0$ “ Sei $g \in G$ beliebig. Dann ist die Abbildung $\pi \mapsto g \circ \pi$ eine Bijektion auf G . Daher ist mit π auch $g \circ \pi$ gleichverteilt auf G und unabhängig von (X, Y) . Somit gilt

$$\mathbb{P}[g(\pi(X, Y)) \in B] = \mathbb{P}[\pi(X, Y) \in B] \quad \text{für alle } B \in \mathcal{B}.$$

Ist H'_0 erfüllt, dann hat $\pi(X, Y)$ dieselbe Verteilung wie (X, Y) , und somit erhalten wir.

$$\mathbb{P}[g(X, Y) \in B] = \mathbb{P}[(X, Y) \in B] \quad \text{für alle } B \in \mathcal{B}.$$

Konstruktion von Tests via G -Symmetrie

Um Tests für die Nullhypothese H_0 bzw. H'_0 zu konstruieren, wählen wir eine beliebige Statistik $T : S^{n+m} \rightarrow \mathbb{R}$. Unter der Nullhypothese H'_0 hat $T(\pi(X, Y))$ dieselbe Verteilung wie $T(X, Y)$. Daher können wir erwarten, dass das Ergebnis, dass der Wert der Statistik für die permutierten Daten nur selten kleiner ist als der Wert für die Ausgangsdaten, eine kleine Wahrscheinlichkeit hat. Um dies zu präzisieren definieren wir einen modifizierten p-Wert.

Definition 5.8 (p-Wert für Test auf G-Symmetrie). Für einen Datenvektor $(x, y) \in S^{n+m}$ setzen wir

$$p_l^G(x, y) := \frac{|\{\pi \in G : T(\pi(x, y)) \leq T(x, y)\}|}{|G|} = \mathbb{P}[T(\pi(x, y)) \leq T(x, y)]$$

$$p_r^G(x, y) := \frac{|\{\pi \in G : T(\pi(x, y)) \geq T(x, y)\}|}{|G|} = \mathbb{P}[T(\pi(x, y)) \geq T(x, y)]$$

Bemerkung. 1) Für $(x, y) \in S^{n+m}$ gilt:

$$p_l^G(x, y) = F_{T(\pi(x, y))}(T(x, y)),$$

$$p_r^G(x, y) = 1 - F_{T(\pi(x, y))}(T(x, y)-).$$

Hierbei ist nur die Permutation π zufällig! Insbesondere hängt die Verteilungsfunktion nicht von der Verteilung von (X, Y) ab. Das heißt der entsprechende Test ist verteilungsunabhängig.

2) Im Gegensatz dazu sind die klassischen p-Werte definiert als

$$p_l(x, y) = \mathbb{P}[T(X, Y) \leq T(x, y)] = F_{T(X, Y)}(T(x, y)),$$

$$p_r(x, y) = \mathbb{P}[T(X, Y) \geq T(x, y)] = 1 - F_{T(X, Y)}(T(x, y)-).$$

Hier ist (X, Y) zufällig und die Verteilungsfunktion hängt somit von der Verteilung von (X, Y) ab.

3) Ist $|G|$ klein, dann kann man die p-Werte durch Abzählen ermitteln. Normalerweise ist dies aber nicht praktikabel, zum Beispiel gilt $|\mathcal{S}_n| = n!$. Stattdessen berechnet man denn die Monte-Carlo p-Werte.

$$p_l^{MC}(x, y) = \frac{|\{i \in \{1, \dots, B\} : T(\pi_i(x, y)) \leq T(x, y)\}| + 1}{B + 1}$$

und entsprechend $p_r^{MC}(x, y)$, wobei $\pi_1, \dots, \pi_B \sim \text{Unif}(G)$ unabhängige Stichproben von der Gleichverteilung auf G sind.

Satz 5.9. Unter H_0 gilt

$$\mathbb{P}[p_l^G(X, Y) \leq \alpha] \leq \alpha \quad \text{und} \quad \mathbb{P}[p_r^G(X, Y) \leq \alpha] \leq \alpha$$

für alle $\alpha \in [0, 1]$.

Beweis. Nach Lemma 5.7 folgt aus H_0 :

$$\mathbb{P}[p_l^G(X, Y) \leq \alpha] = \mathbb{P}[p_l^G(\pi(X, Y)) \leq \alpha]. \quad (*)$$

Wir wollen zeigen, dass die rechte Seite kleiner oder gleich α ist. Nach dem Satz von der totalen Wahrscheinlichkeit ist dies der Fall, wenn

$$\mathbb{P}[p_l^G(\pi(x, y)) \leq \alpha] \leq \alpha \quad \text{für alle } (x, y) \in S^{n+m} \quad (**)$$

gilt. Wegen $p_l^G(\pi(x, y)) = F_{T(\pi(x, y))}(\pi(x, y))$ folgt (*) aus Lemma 1.5. Der Beweis für den rechtsseitigen p-Wert verläuft analog. ■

5. Zusammenhang mehrerer Merkmale

Nach Satz 5.9 erhalten wir einen Hypothesentest für H_0 , der die Niveaubedingung exakt erhält:
Exakter Symmetrietest

Verwerfe H_0 , falls $p_i^G(X, Y) \leq \alpha$ (bzw. falls $p_r^G(X, Y) \leq \alpha$).

In der Praxis ersetzen wir den p-Wert durch den Monte-Carlo p-Wert.

Monte Carlo Symmetrietest

Verwerfe H_0 , falls $p_i^{MC}(X, Y) \leq \alpha$ (bzw. falls $p_r^{MC}(X, Y) \leq \alpha$).

Auch der Monte Carlo Test erfüllt die Niveaubedingung exakt. Dies zeigt man auf ähnliche Weise wie im Beweis von Lemma 4.4, siehe „*Dümbegen, Lemma 8.3*“.

Beispiele. 1) Test auf identische Verteilung der X_i und Y_i

Hier wählen wir $G = \mathcal{S}_{n+m}$, siehe oben, und betrachten eine Statistik der Form

$$T(X, Y) = |M(X) - M(Y)|.$$

Dabei ist M eine Statistik bzgl. der wir den Unterschied zwischen den Verteilungen erkennen möchten, zum Beispiel der Mittelwert, oder eine beliebige andere Kenngröße. Welches M wir betrachten hängt von der Form der Alternativen ab.

2) Test auf Unabhängigkeit bzw. bedingte Austauschbarkeit

Wir betrachten beispielsweise gepaarte Daten:

$$(X_i, Y_i), \quad i = 1, \dots, n \quad \text{mit Werten in } \{0, 1\} \times \mathbb{R}.$$

Um zu testen ob X und Y unabhängig sind, wählen wir $G = \mathcal{S}_n$ und setzen $\pi(X, Y) := (X, \pi Y)$. Wieder kann man viele Arten von Teststatistiken betrachten, beispielsweise

$$T(X, Y) = \sum_{i=1}^n 1_{\{X_i=1\}} R_i(Y),$$

wobei $R(Y)$ der Rang von Y_i unter Y_1, \dots, Y_n ist. Hier würde man testen, ob die Y_i tendenziell größer oder kleinere Ränge haben, falls $X_i = 1$ gilt.

3) Zeitreihen

Sei $n = 0$ und $Y_1, \dots, Y_m : \Omega \rightarrow \{0, 1\}$. Um zu testen, ob die Y_i austauschbar sind, oder ob es eine Abhängigkeit der Werte vom (Zeit-)Index i gibt, wählt man $G = \mathcal{S}_m$. Teststatistiken für verschiedene Alternativen sind zum Beispiel

$$T(Y) = \sum_{i=1}^m Y_i \cdot i \quad (\text{Test auf monotonen Trend}),$$

$$T(Y) = \sum_{i=1}^{m-1} 1_{Y_i \neq Y_{i+1}} \quad (\text{Test auf Clustering}).$$

4) Vorzeichentests

Sei $n = 0$ und $Y_1, \dots, Y_m : \Omega \rightarrow \mathbb{R}$. Um Vorzeichensymmetrie zu testen, betrachten wir $G = +1, -1^m$. Teststatistiken sind etwa

$$TT(Y) = \sum_{i=1}^m \text{sign}(Y_i) \quad (\text{Pearson Vorzeichentest}),$$

$$T(Y) = \sum_{i=1}^{m-1} \text{sign}(Y_i) \cdot \tilde{R}_i \quad (\text{Wilcoxon signedrank}).$$

Diese Tests haben wir schon oben betrachtet.

6. Regression

In diesem Kapitel werden wir die Grundlagen und Methoden der Regression behandeln. Als Literaturgrundlage dient Kapitel 12 aus *Stochastik, H. Georgii* oder Kapitel 13 aus Wasserman sowie das Werk von Bickel und Doksum.

Oft hängen Beobachtungswerte von bestimmten Kontrollparametern ab, die im Experiment nach Belieben eingestellt werden können. Diese Abhängigkeit wird jedoch durch zufällige Beobachtungsfehler gestört, und die zugehörigen Koeffizienten sind unbekannt. Wie kann man sie trotzdem aus den Beobachtungen ermitteln? Dies ist das Ziel der Regression, bei der aus gestörten Beobachtungen auf die zugrunde liegende Beziehung geschlossen wird. Im einfachsten Fall wird eine Reihe von Messpunkten in der Ebene durch eine Kurve angenähert.

Wir betrachten Daten der Form $(X_1, Y_1), \dots, (X_n, Y_n) : \Omega \rightarrow \mathbb{R}^d \times \mathbb{R}$ (bzw. $\mathbb{R}^d \times \mathbb{R}^{d'}$).

- X_i sind die *covariates* bzw. Kovariaten (explanatory/predictor variables, features, Kontrollgrößen, Ausgangs-/Regressorvariablen).
- Y_i sind die *response variables* bzw. abhängigen Variablen.

Des Weiteren sind die Beobachtungen Y_1, \dots, Y_n bedingt unabhängig gegeben X . Dies bedeutet, dass die Y_i unter der Bedingung der X_i unabhängig voneinander sind.

Ein häufig betrachtetes Szenario ist, dass die X_i deterministisch sind, also $X_1 = x_1, \dots, X_n = x_n$ und die Y_i unabhängig sind. Um die Abhängigkeit der Y_i von den X_i zu bestimmen, betrachten wir die bedingten Dichtefunktionen von Y_i gegeben durch $f_\vartheta(y_i|x_i)$, $i = 1, \dots, n$ mit $\vartheta \in \Theta$. Dann erhalten wir unter

$$Y_i \sim f_\vartheta(y_i|x_i) \quad \text{die Likelihoodfunktion} \quad L(\theta; y) = \prod_{i=1}^n f_\vartheta(y_i|x_i)$$

wobei $y = (y_1, \dots, y_n)$.

Beispiele. 1) **Nichtparametrisches Regressionsmodell:** Seien die abhängigen Variablen gegeben durch $Y_i = g(X_i) + \varepsilon_i$. Unter der Annahme, dass die ε unabhängig und identisch verteilt sind, und unabhängig von X , mit $\mathbb{E}[\varepsilon_i] = 0$, $\text{DomVar}[\varepsilon_i] = v$, wobei $v > 0$ und $g : \mathbb{R}^d \rightarrow \mathbb{R}$ eine unbekannte Funktion ist.

2) **Lineares Modell:** Im linearen Modell gehen wir davon aus, dass die unbekannte Funktion g gegeben ist durch:

$$g(x) = w^T \cdot x = \sum_{j=1}^d w_j x_j \quad \text{wobei } x = (x_i) \in \mathbb{R}^d.$$

Hierbei heißt $w \in \mathbb{R}^d$ Gewichtsvektor. Wir erhalten also

$$Y_i = w^T X_i + \varepsilon_i \quad \text{und} \quad Y = Aw + \varepsilon \quad \text{wobei } A = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

Die Matrix A heißt Design-Matrix.

3) **Two-Layer Neural Network:** Wir betrachten ein zweischichtiges Neuronales Netzwerk mit Eingabeschicht $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ und Ausgabeschicht $y = (y_1, \dots, y_{d_2}) \in \mathbb{R}^{d_2}$ und einer d_1

6. Regression

dimensionalen Zwischenschicht. Wir gehen davon aus, dass der Output beschrieben werden kann durch

$$g(x) = \sigma(w_2^T \cdot \sigma(w_1^T \cdot x)),$$

wobei $w_1 \in \mathbb{R}^{d \times d_1}$, $w_2 \in \mathbb{R}^{d_1 \times d_2}$ in Abhängigkeit der Funktion σ . Beispielsweise wählt man

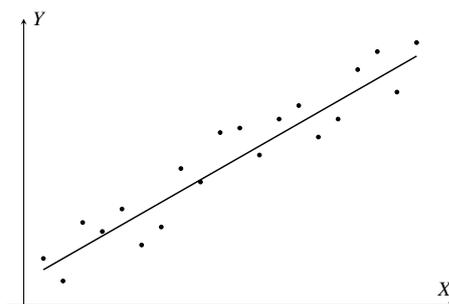
$$\sigma(x) = \begin{pmatrix} s(x_1) \\ \vdots \\ s(x_d) \end{pmatrix} \quad \text{wobei } s(x) = \frac{1}{1 + e^{-x}} \quad (\text{Sigmoid-Funktion})$$

6.1. Einfache Lineare Regression

Die einfache Lineare Regression untersucht die Beziehung zwischen den Kontrollgrößen X_i und den davon linear abhängigen Variablen Y_i . Es wird also angenommen, dass die Abhängigkeit gegeben ist durch

$$Y_i = a + bX_i + \varepsilon_i \quad \text{mit } i = 1, \dots, n,$$

wobei $a, b \in \mathbb{R}$ zwei unbekannte Koeffizienten sind, die ermittelt werden sollen, und ε_i geeignete, von X unabhängige Zufallsvariablen sind, welche den zufälligen Messfehler beschreiben. Insbesondere sind diese wieder zentriert, das heißt $\mathbb{E}[\varepsilon_i] = 0$ und $\text{DomCov}[\varepsilon_i, \varepsilon_j] = v \cdot \delta_{ij}$. Ziel der linearen Regression ist es, die unbekannt Parameter $\vartheta = (a, b, v)$ zu bestimmen. Anders gesagt, wollen wir die Parameter so schätzen, dass die dadurch entstandene *Regressionsgerade* $y = \hat{a} + \hat{b}x$ die erhaltenen Messpunkte $(X_1, Y_1), \dots, (X_n, Y_n)$ möglichst gut repräsentiert.



Insbesondere können wir die lineare Abhängigkeit auch wie folgt beschreiben:

$$Y = a \mathbf{1} + bX + \varepsilon, \quad \text{wobei } \mathbf{1} \text{ der Einheitsvektor im } \mathbb{R}^n \text{ ist.}$$

Beispiel (Gauß-Modell). In diesem Beispiel gehen wir davon aus, dass die bedingte Likelihood bei $X = x$ gegeben ist durch

$$L(\vartheta, y) = (2\pi v)^{-\frac{n}{2}} \exp\left(-\frac{1}{2v} \sum_{i=1}^n (y_i - a - x_i)^2\right).$$

Um einen Maximum-Likelihood-Schätzer für (a, b) bei gegebenem v zu finden, wollen wir die folgende Summe minimieren:

$$\underbrace{\sum_{i=1}^n (y_i - a - x_i)^2}_{=||y - a\mathbf{1} - bx||^2} \stackrel{!}{=} \min.$$

Die Schätzer \hat{a}, \hat{b} , die wir dadurch erhalten, heißen *Kleinste-Quadrate-Schätzer* (auch LSE: Least Squares Estimator). Dies liefert die *Regressionsgerade* $y = \hat{a} + \hat{b}x$.

Bemerkung. Der LSE wird oft auch im nicht-Gauß-Fall verwendet (obwohl er dann nicht mit dem MLE übereinstimmt!).

Definition 6.1. Seien $\tilde{X}_i := X_i - \bar{X}$ und $\tilde{Y}_i := Y_i - \bar{Y}$ die zentrierten Zufallsvariablen mit $i = 1, \dots, n$. Dann definieren wir

(i)

$$s_X := \sqrt{\frac{1}{n-1} \sum_{i=1}^n \tilde{X}_i^2} \quad \text{bzw.} \quad s_Y := \sqrt{\frac{1}{n-1} \sum_{i=1}^n \tilde{Y}_i^2}$$

als die *Stichprobenstandardabweichungen* für X bzw. Y .

(ii)

$$\hat{\rho} := \frac{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{Y}_i}{\sqrt{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \tilde{Y}_i^2}} = \frac{\tilde{X}_i \cdot \tilde{Y}_i}{\|\tilde{X}_i\| \cdot \|\tilde{Y}_i\|}$$

ist die *Stichprobenkorrelation* von X und Y .

Korollar 6.2. 1) Der LSE bzw. die Regressionsgerade ist gegeben durch $\hat{Y} = \hat{a} + \hat{b}X$, wobei

$$\hat{b} = \hat{\rho} \cdot \frac{s_Y}{s_X} = \frac{\sum \tilde{X}_i \tilde{Y}_i}{\sum \tilde{X}_i^2} = \frac{\tilde{X}_i \cdot \tilde{Y}_i}{\|\tilde{X}\|^2}, \quad \hat{a} = \bar{Y} - \hat{b}\bar{X}$$

2) Der Schätzungsfehler ist gegeben durch

$$\|Y - \hat{Y}\|^2 = (1 - \hat{\rho}^2) \|\tilde{Y}\|^2.$$

Beweis. 1) Die Behauptung ist ein Spezialfall des Satzes von der linearen Prognose („Dümbgen, Lemma 8.5“ bzw. Satz ?? im Skript), wenn die zugrunde liegende Wahrscheinlichkeitsverteilung die empirische Verteilung ist:

$$\mathbb{P} = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$$

2) Wir sehen, dass die Regressionsgerade die Funktion $\|Y - (\hat{a}1 + \hat{b}X)\|$ minimiert. Aus der Linearen Algebra wissen wir somit, dass \hat{Y} die orthogonale Projektion von $Y \in \mathbb{R}^n$ auf den Unterraum

$$L = \text{span}\left(\overbrace{1, X}^{\text{i.A. nicht orth.}} \right) = \text{span}\left(\overbrace{1, \tilde{X}}^{\text{orthogonal}} \right)$$

ist. Also gilt

$$\hat{Y} = \frac{Y \cdot 1}{\|1\|^2} 1 + \frac{Y \cdot \tilde{X}}{\|\tilde{X}\|^2} \tilde{X} = (\bar{Y} - \hat{b}\bar{X})1 + \hat{b}X$$

6. Regression

Dies ist der erste Teil der Behauptung. Außerdem erhalten wir

$$\begin{aligned} Y - \hat{Y} &= Y - \bar{Y}1 - \hat{b}(X - \bar{X}1) = \tilde{Y} - \hat{b}\tilde{X}, \\ \Rightarrow \|Y - \tilde{Y}\|^2 &= \|\tilde{Y} - \hat{b}\tilde{X}\|^2 = \|\tilde{Y}^2\| - \frac{(\tilde{Y} \cdot \tilde{X})^2}{\|\tilde{X}\|^2}, \\ &= (1 - \hat{\rho}^2)\|\tilde{Y}\|^2 \end{aligned}$$

Bemerkung. 1) Die Gleichung für die Regressionsgerade kann man auch schreiben als

$$\frac{\hat{Y} - \bar{Y}}{s_Y} = \hat{\rho} \frac{X - \bar{X}}{s_X}.$$

Insbesondere liegt der Schwerpunkt (\bar{X}, \bar{Y}) auf der Regressionsgerade.

2) Die Abweichung der einzelnen Beobachtungswerte gegeben durch $\hat{\varepsilon}_i := Y_i - \hat{Y}_i$ heißt *Residuum*. Es gilt:

$$\hat{\varepsilon} = a1 + bX + \varepsilon - \hat{a} - \hat{b}X = a - \hat{a} + (b - \hat{b})X + \varepsilon.$$

Eigenschaften des LSE

Satz 6.3.

Unter der bedingten Verteilung gegeben X gilt:

1) \hat{a} und \hat{b} sind erwartungstreue Schätzer für a und b mit

$$\text{Var}[\hat{b}] = \frac{v}{\sum \tilde{X}_i^2}, \quad \text{Var}[\hat{a}] = v \frac{\frac{1}{n} \sum X_i^2}{\sum \tilde{X}_i^2}, \quad \text{Cov}[\hat{a}, \hat{b}] = -v \frac{\bar{X}_n}{\sum \tilde{X}_i^2}$$

2) Der Schätzer

$$\hat{V} = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

ist Erwartungstreu für v .

Beweis. 1) Seien $a, b \in \mathbb{R}$ die Koeffizienten der Geraden $Y = a1 + bX + \varepsilon$, dann folgt

$$\begin{aligned} \Rightarrow \tilde{Y} &= Y - \bar{Y}1 = b\tilde{X} + \varepsilon - \bar{\varepsilon}1 \\ \Rightarrow \hat{b} &= \frac{\tilde{X} \cdot \tilde{Y}}{\|\tilde{X}\|^2} = b + \frac{\tilde{X} \cdot (\varepsilon - \bar{\varepsilon}1)}{\|\tilde{X}\|^2} = b + \frac{\tilde{X} \cdot \varepsilon}{\|\tilde{X}\|^2}, \\ \Rightarrow \mathbb{E}[\hat{b}] &= b, \quad \text{Var}[\hat{b}] = \frac{1}{\|\tilde{X}\|^4} \text{Var}[\tilde{X} \cdot \varepsilon] = \frac{v}{\|\tilde{X}\|^2}. \end{aligned}$$

Die letzte Gleichung folgt, da $\text{Var}[\tilde{X} \cdot \varepsilon] = \sum_{i,j} \tilde{X}_i \tilde{X}_j \text{Cov}[\varepsilon_i, \varepsilon_j] = v \sum \tilde{X}_i^2$. Mit $\bar{Y} = a + b\bar{X} + \bar{\varepsilon}$ folgt

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} = a + (b - \hat{b})\bar{X} + \bar{\varepsilon}.$$

Da sowohl $(b - \hat{b})$ als auch $\bar{\varepsilon}$ Erwartungswert 0 haben (siehe oben), folgt $\mathbb{E}[\hat{a}] = 0$ sowie die Behauptung der Varianz und Covarianz.

2) Da \hat{Y} die orthogonale Projektion von Y auf L ist, folgt

$$\hat{\varepsilon} = Y - \hat{Y} = Y - \Pi^L Y = \Pi^{L^\perp} Y = \Pi^{L^\perp} \varepsilon$$

. Somit folgt $\mathbb{E}[\hat{\varepsilon}] = 0$. Wir betrachten

$$\begin{aligned} \mathbb{E}[|\hat{\varepsilon}|^2] &= \mathbb{E}[|\Pi^{L^\perp} \varepsilon|^2] = \sum_{i,j,k} \Pi_{ij}^{L^\perp} \Pi_{ik}^{L^\perp} \mathbb{E}[\varepsilon_j \varepsilon_k] \\ &= \sum_{i,j,k} \Pi_{ij}^{L^\perp} \Pi_{ik}^{L^\perp} \delta_{jk} \cdot v = v \sum_{i,j} \Pi_{ij}^{L^\perp} \Pi_{ij}^{L^\perp} \\ &= v \cdot \text{Tr}(\Pi_{L^\perp} \Pi_{L^\perp}) = v \cdot (n - 2) \end{aligned}$$

Die letzte Gleichung folgt, da $\dim(L^\perp) = n - \dim(L) = n - 2$. ■

Um Konfidenzintervalle zu berechnen, benötigen wir die gemeinsame Verteilung von \hat{a} , \hat{b} , $\hat{\varepsilon}$. Um diese berechnen zu können, müssen wir voraussetzen, dass ε multivariat normalverteilt ist, das heißt

$$\varepsilon \sim N(0, v \cdot I_n) \quad (\text{Lineares Gauß-Modell}).$$

Wir wollen die Berechnung im nächsten Abschnitt in größerer Allgemeinheit durchführen.

6.2. Lineare Modelle

Das folgende Kapitel basiert zu großen Teilen auf den Abschnitten 12.2, 12.3 aus *Stochastik, H. Georgii*.

Wir wollen eine Verallgemeinerung der bereits gesehenen einfachen linearen Regressionsgleichung betrachten. Sei hierfür $A \in \mathbb{R}^{n \times d}$ eine Matrix mit Rang d , also $\ker(A) = \{0\}$, wie oben auch Design-Matrix. Die zufälligen Störgrößen seien gegeben durch $\varepsilon : \Omega \rightarrow \mathbb{R}^n$, wobei wie oben $\mathbb{E}[\varepsilon_i] = 0$ und $\text{Cov}[\varepsilon_i, \varepsilon_j] = v \cdot \delta_{ij}$, $v > 0$. Dann sei die Beobachtung gegeben durch

$$Y : \Omega \rightarrow \mathbb{R}^n \quad Y = Aw + \varepsilon$$

mit den unbekanntem Parametern $w_i \in \mathbb{R}$.

Beispiele. 1) Insbesondere ist die einfache lineare Regression ein Spezialfall der linearen Regression. Denn mit $d = 2$ und

$$w \begin{pmatrix} a \\ b \end{pmatrix}, \quad A = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \Rightarrow \quad Y = Aw + \varepsilon = 1a + bX + \varepsilon$$

2) Ein weiteres Beispiel ist die multiple lineare Regression, bei der man von mehreren Einflussfaktoren / Kovarianten $A = (X_{i,j})_{i,j}$, $i = 1, \dots, n, j = 1, \dots, d$ ausgeht. Wir erhalten also die Regressionsgleichung

$$Y_i = \sum_{j=1}^d w_j X_{i,j} + \varepsilon_i$$

3) In der polynomialen Regression wird die Abhängigkeit durch ein Polynom höherer Ordnung beschrieben. Um nichtlineare Zusammenhänge abzubilden, werden zusätzlich die Terme X^2, X^3, \dots berücksichtigt. Wir betrachten also

$$A = \begin{pmatrix} 1 & X_1 & X_1^2 & \dots & X_1^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^k \end{pmatrix} \quad \text{und erhalten die Abhängigkeit} \quad Y_i = \sum_{j=0}^k w_j X_i^j + \varepsilon_i$$

6. Regression

- 4) Das Einstichproben-Lokationsmodell ist ein Spezialfall der linearen Regression, bei dem überprüft wird, ob der Mittelwert einer Population einem bestimmten Wert w entspricht. Hierbei wird die abhängige Variable modelliert durch $w = m$, $A = 1$ und wir erhalten $Y_i = m + \varepsilon_i$. Falls die Y_i unabhängig und normalverteilt sind nach $N(m, \nu)$, ergibt sich das Standard-Gauß-Modell, siehe oben.
- 5) Das Mehrstichproben-Lokationsmodell vergleicht die Mittelwerte mehrerer Gruppen, um gegebenenfalls signifikante Unterschiede festzustellen. Beispielsweise der Weizenertrag bei verschiedenen Düngerarten. Das Modell beschreibt Y_{ik} als den Mittelwert der i -ten Gruppe. Also $Y_{ik} = m_i + \varepsilon_{ik}$ für $i = 1, \dots, p$ und $k = 1, \dots, n$. Wir erhalten

$$d = p, \quad n = n_1 + \dots + n_p, \quad w = m, \quad \text{und} \quad A = \begin{pmatrix} 1_{n_1} & 0 & \dots & \dots & 0 \\ 0 & 1_{n_2} & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1_{n_p} \end{pmatrix}$$

Hierbei ist 1_{n_i} die Einheitsmatrix in \mathbb{R}^{n_i} .

Eine Anwendung des Mehrstichproben-Lokationsmodells ist die Varianzanalyse (ANOVA). Sie vergleicht die Varianz zwischen den Gruppen mit der Varianz innerhalb der Gruppen.

Ein Test, der diese Varianzanalyse verwendet, ist der *F-Test*. Dieser testet die Hypothese

$$H_0 : m_1 = \dots = m_p$$

. Die Nullhypothese wird verworfen, falls $\hat{V}_{ZG} > c \cdot \hat{V}_{iG}$, wobei

$$\hat{V}_{ZG} := \frac{1}{p-1} \sum_{i=1}^p n_i (M_i - M)^2, \quad \hat{V}_{iG} := \frac{1}{n-p} \sum_{i,k} (Y_{ik} - M_i)^2$$

Hierbei sind M_i die Stichprobenmittel in Gruppe i und M das allgemeine Stichprobenmittel. Im Gauß-Fall erhalten wir

$$\frac{\hat{V}_{ZG}}{\hat{V}_{iG}} \sim \frac{\frac{1}{p-1} \sum_{i=1}^{p-1} Z_i^2}{\frac{1}{n-p} \sum_{i=p}^n Z_i^2}$$

mit $Z_i \sim N(0, 1)$ unabhängig. Die Verteilung dieser Zufallsgröße heißt Fisher-Verteilung mit $p-1$ und $n-p$ Freiheitsgraden.

Schätzer für w

Wie im einfachen linearen Modell interessieren wir uns für die unbekannt Parameter w der Regressionsgleichung $Y = Aw + \varepsilon$. Sei nun

$$L := \text{Range}(A) = \{Aw : w \in \mathbb{R}^d\}, \quad \text{wobei} \quad \dim(L) = d$$

Die Aussage folgt, da $\text{Rang}(A) = d$. Sei $\Pi : \mathbb{R}^n \rightarrow L$ die orthogonale Projektion auf L . Dann gilt:

$$\Pi Y = Aw + \Pi \varepsilon \stackrel{\text{Lem}}{=} A\hat{w} \quad \text{mit} \quad \hat{w} := (A^T A)^{-1} A^T Y \quad (**)$$

Lemma 6.4. Die Matrix $A^T A \in \mathbb{R}^{d \times d}$ ist invertierbar, und die orthogonale Projektion ist gegeben durch Multiplikation mit

$$\Pi = A(A^T A)^{-1} A^T$$

Beweis. (i) Wir wollen zeigen, dass die Matrix $A^T A$ vollen Rang hat. Sei also $u \in \mathbb{R}^d$, dann gilt

$$A^T A u = 0 \Rightarrow \|Au\|^2 = u \cdot A^T A u = 0 \stackrel{\ker(A)=\{0\}}{\Rightarrow} u = 0$$

Folglich ist die Matrix invertierbar.

(ii) Wir sehen, dass $A(A^T A)^{-1} A^T Y \in L$ für alle $Y \in \mathbb{R}^n$, und

$$(Y - A(A^T A)^{-1} A^T Y) \cdot Au = (A^T Y - A^T Y) \cdot u = 0 \quad \text{für alle } u \in \mathbb{R}^d$$

Also ist $A(A^T A)^{-1} A^T Y$ die orthogonale Projektion von Y auf L . ■

Satz 6.5. 1) Der Schätzer \hat{w} ist erwartungstreu für w .

2) Für jedes $u \in \mathbb{R}^d$ ist $u \cdot \hat{w}$ ein erwartungstreuer Schätzer für $u \cdot w$ mit minimaler Varianz unter allen linearen erwartungstreuen Schätzern.

3) Der Schätzer $\hat{V} := \frac{\|Y - \Pi Y\|^2}{n-d}$ ist ein erwartungstreuer Schätzer für v .

Beweis. 1) Wir nutzen die Injektivität der Abbildung, die durch die Matrix A induziert wird. Es gilt:

$$A\mathbb{E}[\hat{w}] = \mathbb{E}[Aw] = Aw$$

Die letzte Gleichheit folgt aus (**). Folglich ist der Schätzer erwartungstreu.

2) Der Schätzer $u \cdot \hat{w} = [A(A^T A)^{-1} u] \cdot Y$ ist nach 1) erwartungstreu. Sei nun $b \cdot Y$ mit $b \cdot \mathbb{R}^n$ ein beliebiger erwartungstreuer linearer Schätzer für $u \cdot w$. Mit $a := A(A^T A)^{-1} u$ gilt dann:

$$\begin{aligned} b \cdot Aw &= \mathbb{E}[b \cdot Y] = \mathbb{E}[a \cdot Y] = a \cdot Aw \quad \text{für alle } w \in \mathbb{R}^d \\ &\Rightarrow b - a \in \perp L \quad \Rightarrow a = \Pi b \\ &\Rightarrow \text{Var}[b \cdot Y] = \text{Var}[b \cdot \varepsilon] = v \cdot \|b\|^2 \geq v \cdot \|a\|^2 = \text{Var}[a \cdot Y] \end{aligned}$$

Denn $\text{Var}[b \cdot \varepsilon] = v \cdot \|b\|^2$ gilt nach den Gauß-Markov-Annahmen.

3) Sei nun e_1, \dots, e_d eine Orthonormalbasis von L , ergänze diese zu einer Orthonormalbasis e_1, \dots, e_n von \mathbb{R}^n . Dann gilt

$$\hat{V} = \frac{\|Y - \Pi Y\|^2}{n-d} = \frac{\|Aw - \varepsilon - Aw + \Pi \varepsilon\|^2}{n-d} = \frac{\|\varepsilon - \Pi \varepsilon\|^2}{n-d} = \frac{1}{n-d} \sum_{i=d+1}^n (e_i \cdot \varepsilon)^2 \quad (***)$$

Mit $\mathbb{E}[(e_i \cdot \varepsilon)^2] = v \cdot \|e_i\|^2 = v$ folgt die Aussage schon mit

$$\mathbb{E}[\hat{V}] = \frac{1}{n-d} \sum_{i=d+1}^n \mathbb{E}[(e_i \cdot \varepsilon)^2] = v$$
■

Konfidenzbereiche / Tests

Um Konfidenzbereiche angeben zu können, ist eine Spezifikation der Verteilung von ε erforderlich. Wir betrachten also im Folgenden das Gauß-Modell, das heißt $\varepsilon \sim N(0, I_n)$.

Satz 6.6. Im linearen Gauß-Modell erhalten wir die folgenden Verteilungen:

- 1) Die Verteilung von \hat{w} ist gegeben durch die multivariate Normalverteilung

$$\hat{w} \sim N(w, v \cdot (A^T A)^{-1})$$

mit dem Mittelwertvektor w und der Kovarianzmatrix $v(A^T A)^{-1}$.

- 2)

$$(n-d) \frac{\hat{V}}{v} \sim \chi^2(n-d)$$

ist Chi-Quadrat-verteilt mit $(n-d)$ Freiheitsgraden.

- 3)

$$\frac{\|A(\hat{w} - w)\|^2}{v} \sim \chi^2(d)$$

Insbesondere ist diese Zufallsgröße unabhängig von \hat{V} .

- 4)

$$\frac{\|A(\hat{w} - w)\|^2}{d\hat{V}} \sim F(d, n-d)$$

F ist hierbei die Fisher-Verteilung.

Beweis. 1) Wir wissen, dass $Y = Aw + \varepsilon \sim N(Aw, v \cdot I_n)$, somit folgt:

$$\Rightarrow \hat{w} = (A^T A)^{-1} A^T Y \sim N(\underbrace{BAw}_{=w}, \underbrace{B(vI_n)N^T}_{=(A^T A)^{-1}})$$

- 2) Nach (***) gilt

$$(n-d) \frac{\hat{V}}{v} = \sum_{i=d+1}^n \left(\frac{e_i \cdot \varepsilon}{\sqrt{v}} \right)^2 \sim \chi^2(n-d)$$

Da nämlich gilt $\varepsilon_i / \sqrt{v} =: Z_i \sim N(0, 1)$.

- 3) Wir betrachten

$$\frac{1}{v} \|\Pi Y - \mathbb{E}[Y]\|^2 = \frac{1}{v} \|Aw + \Pi \varepsilon - Aw\|^2 = \frac{1}{v} \|\Pi \varepsilon\|^2 = \sum_{i=1}^d \left(\frac{e_i \cdot \varepsilon}{\sqrt{v}} \right)^2$$

Analog zu 2) erhalten wir die Verteilung $\chi^2(d)$, und insbesondere folgt auch die gegenseitige Unabhängigkeit.

- 4) Mit 2) und 3) erhalten wir

$$\frac{\|A(\hat{w} - w)\|^2}{d\hat{V}} = \frac{1}{d} \sum_{i=1}^d Z_i^2 \bigg/ \frac{1}{v} \sum_{i=d+1}^n Z_i^2 \sim F(d, n-d)$$

■

Korollar 6.7 (Konfidenzbereiche).

1)

$$C = \left\{ w \in \mathbb{R}^d : \frac{\|A(\hat{w} - w)\|^2}{d\hat{V}} < f_{d, n-d; 1-\alpha} \right\}$$

ist ein Konfidenzellipsoid für w zum Konfidenzniveau $1 - \alpha$.

2)

$$I = \left(\frac{(n-d)\hat{V}}{q_+}, \frac{(n-d)\hat{V}}{q_-} \right)$$

mit $q_+ = \chi_{n-d, 1-\alpha/2}^2$, $q_- = \chi_{n-d, \alpha/2}^2$, ist ein Konfidenzintervall für v zum Konfidenzniveau $1 - \alpha$.

Die Aussage folgt unmittelbar aus Satz 6.6.

Hypothesentests

In diesem Abschnitt beschäftigen wir uns mit der Frage, welche Kovariaten relevant sind, um die Abhängigkeit von Y zu X zu bestimmen. Wir werden im Folgenden Hypothesentests durchführen, um die Kovariaten auf ihre Signifikanz zu bestimmen. Hierbei sei

$$\begin{aligned} w_j &= e_j \cdot w && \text{Gewicht des } j\text{-ten Kovariaten} \\ \hat{w}_j &= e_j \cdot \hat{w} && \text{erwartungstreuer Schätzer für } w_j \end{aligned}$$

wobei e_j der j -te Einheitsvektor ist. Wir wollen also auf die Nullhypothese

$$H_0 : w_j = 0$$

testen. Eine erste Idee für eine Teststatistik ist \hat{w}_j durch seine Standardabweichung zu teilen, also

$$\frac{\hat{w}_j}{\sigma[\hat{w}_j]}, \quad \sigma[\hat{w}_j] = \sqrt{v} \sqrt{e_j \cdot (A^T A)^{-1} e_j}$$

Da jedoch v unbekannt ist, ist dieser Schätzer nicht praktikabel. Eine bessere Idee ist es, die geschätzte Standardabweichung zu betrachten. Wir definieren also die folgende Teststatistik

$$T(Y) = \frac{\hat{w}_j}{\hat{\sigma}[\hat{w}_j]} = \frac{\hat{w}_j}{\sqrt{\hat{V}} \sqrt{e_j \cdot (A^T A)^{-1} e_j}}$$

Um die Verteilung berechnen zu können, betrachten wir wieder das Gauß-Modell und nehmen an, dass $\varepsilon \sim N(0, v \cdot I_n)$.

Korollar 6.8. Im linearen Gauß-Modell gilt

$$T(Y) \sim t(n-d)$$

Beweis. Nach Satz 6.6 folgt

$$T(Y) = \frac{\overbrace{\hat{w}_j}^{\sim N(0,1)}}{\sigma[\hat{w}_j]} \bigg/ \sqrt{\frac{1}{n-d} \overbrace{(n-d)}^{\sim \chi^2(n-d)} \frac{\hat{v}}{v}}$$

wobei die beiden Zufallsvariablen jeweils unabhängig sind. ■

Wir erhalten also den **t-Test** zum Niveau α , welcher H_0 verwirft, falls $|T(Y)| > t_{n-d; 1-\frac{\alpha}{2}}$. Wir wollen im Folgenden jedoch noch eine zweite Methode betrachten, bei der mehrere Kovariaten gleichzeitig getestet werden. Wir betrachten also die Nullhypothese

$$H_0 : w_1 = w_2 = \dots = w_l = 0$$

wobei $l < d$. Dies bedeutet, dass wir prüfen, ob mehrere Kovariaten simultan keinen Einfluss auf die abhängige Variable haben. Wir können diese Nullhypothese auch beschreiben durch:

$$H_0 \Leftrightarrow Aw \in H := \{Au : u \in \mathbb{R}^d \text{ mit } u_1 = u_2 = \dots = u_l = 0\}$$

Hierbei ist $H \subseteq L$ ein linearer Unterraum mit $\dim(H) = d - l$. Wir betrachten die folgende Teststatistik:

$$F(Y) = \frac{\frac{1}{l} \left\| \overbrace{A\hat{w}}^{=\Pi_L Y} - \overbrace{\Pi_H A\hat{w}}^{=\Pi_H Y} \right\|^2}{\hat{v}} = \frac{\frac{1}{l} \|\Pi_L Y - \Pi_H Y\|^2}{\frac{1}{n-d} \|Y - \Pi_L Y\|^2}$$

wobei Π_H die Projektion auf den Unterraum H ist. Die Statistik heißt Fisher-Statistik.

Korollar 6.9. Im linearen Gauß-Modell gilt

$$F(Y) \sim F(l, n-d)$$

Beweis. Sei e_1, \dots, e_{d-l} eine Orthonormalbasis von H , ergänze diese zu einer Orthonormalbasis e_1, \dots, e_l von L , und diese dann zu einer Orthonormalbasis e_1, \dots, e_n von \mathbb{R}^n . Dann gilt unter der Nullhypothese

$$F(Y) = \frac{1}{l} \sum_{i=d-l+1}^d \left(\frac{e_i \cdot \varepsilon}{\sqrt{v}} \right)^2 \bigg/ \frac{1}{n-d} \sum_{i=d+1}^n \left(\frac{e_i \cdot \varepsilon}{\sqrt{v}} \right)^2 \sim F(l, n-d)$$

Analog zum obigen Beweis sind $Z_i = \frac{e_i \cdot \varepsilon}{\sqrt{v}} \sim N(0, 1)$ unabhängig. ■

Wir erhalten also den **F-Test** zum Niveau α , welcher H_0 verwirft, falls $F(Y) > f_{l, n-d; 1-\alpha}$.

Es ist jedoch entscheidend, die Ergebnisse mit Vorsicht zu interpretieren. Dies gilt insbesondere, wenn zusätzliche Einflussfaktoren in die Analyse einbezogen werden. Eine bekannte Problematik in diesem Zusammenhang ist das *Simpson-Paradoxon*.

Simpson-Paradoxon: Das Simpson-Paradoxon beschreibt eine Situation, in der eine Beobachtung verschwindet oder sich umkehrt, wenn weitere Daten aggregiert werden. In unserem Fall drückt sich dies in den folgenden Zusammenhängen aus.

- (i) Es kann gelten $w_j > 0$, aber $w_j < 0$ bei der Berücksichtigung zusätzlicher Einflussfaktoren (Kovariaten).

- (ii) Falls $w_j \approx 0$ bedeutet dies nicht zwangsläufig, dass Y nicht von $X^{(j)}$ abhängt, sondern nur, dass $X^{(j)}$ für eine optimale lineare Prognose nicht benötigt wird. Zum Beispiel, weil der Effekt schon durch andere Einflussfaktoren prognostiziert werden kann, die wiederum mit $X^{(j)}$ korreliert sind.
- (iii) Analog impliziert $w_j \neq 0$ nicht automatisch, dass eine kausale Beziehung existiert. Zum Beispiel können $X^{(j)}$ und Y von ähnlichen Faktoren beeinflusst werden.

Diese Aspekte unterstreichen die Wichtigkeit, bei der Interpretation der Ergebnisse die zugrunde liegenden Annahmen und die möglichen Einflüsse von Kovariaten zu berücksichtigen.

Beispiel (ANOVA). Wie im obigen Beispiel betrachten wir die Ernteerträge von verschiedenen Düngarten. Hierbei sei das Modell gegeben durch

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, \dots, p, \quad j = 1, \dots, n$$

wobei Y_{ij} die Beobachtung der j -ten Stichprobe in der i -ten Gruppe ist, μ_i der mittlere Ertrag der i -ten Gruppe und ε_{ij} der gruppenspezifische Fehlerterm ist. Dann gilt:

$$d = p, \quad n = n_1, \dots, n_p, \quad w = \mu, \quad A = \begin{pmatrix} 1_{n_1} & 0 & \dots & \dots & 0 \\ 0 & 1_{n_2} & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1_{n_p} \end{pmatrix}$$

Die Nullhypothese der ANOVA besagt, dass die Mittelwerte aller Gruppen gleich sind. Wir betrachten also

$$H_0 : \quad \mu_1 = \mu_2 = \dots = \mu_p$$

Dies kann auch äquivalent dargestellt werden durch:

$$H_0 \Leftrightarrow A\mu \in H := \{A\mu : \mu_1 = \mu_2 = \dots = \mu_p\}, \quad \dim(H) = 1$$

Zur Überprüfung der Nullhypothese kann der F-Test ähnlich wie oben angewendet werden.

6.3. Andere Regressionsverfahren

In diesem Kapitel werden verschiedene Regressionsmethoden vorgestellt, die über das klassische lineare Modell hinausgehen. Diese erweiterten Verfahren sind besonders nützlich, wenn das klassische Modell aufgrund von Problemen wie Überanpassung (Overfitting) oder einem Bedarf an robusteren Schätzverfahren nicht ausreicht. Dieses Kapitel basiert auf den Ausführungen von Murphy sowie Hastie, Tibshirani und Friedman, die in ihren Werken *Machine Learning* und *The Elements of Statistical Learning* eine detaillierte Einführung in diese Methoden geben.

Robustheit

Ein häufiges Problem bei der linearen Regression ist die Anfälligkeit gegenüber Ausreißern. Das klassische Verfahren der kleinsten Quadrate (Least Squares), das die Residuen minimiert, kann durch solche abweichenden Datenpunkte stark beeinflusst werden. Dieses Verfahren wird durch das folgende Minimum beschrieben:

$$LSE = \min_w = \sum_{i=1}^n |Y_i - w^T X_i|^2$$

Betrachten wir die Summe als Funktion von w , spricht man von der Verlustfunktion oder auch dem RRS (residual sum of squares), wobei X_i die i -te Zeile der Design-Matrix A ist. Im Folgenden betrachten wir zwei Alternativen, die weniger anfällig für Ausreißer sind:

6. Regression

- (i) Um die Robustheit zu erhöhen, können anstelle der quadrierten Abweichungen die absoluten Abweichungen minimiert werden. Wir wollen also die l^1 -Norm minimieren:

$$\sum_{i=1}^n |Y_i - w^T X_i|$$

Diese Methode ist weniger anfällig gegenüber Ausreißern, da größere Fehler nicht so stark gewichtet werden. Insbesondere entspricht dies dem Maximum-Likelihood-Schätzer im Laplace-Modell mit den Dichtefunktionen

$$f_w(y_i|x_i) \propto e^{-|y_i - w^T x_i|}$$

Hierbei handelt es sich um ein lineares Optimierungsproblem.

- (ii) Eine weitere Möglichkeit, die Robustheit zu erhöhen, besteht darin, den Huber-Loss zu minimieren. Dieser ist gegeben durch:

$$\sum_{i=1}^n f_\delta(Y_i - w^T X_i) \quad \text{wobei} \quad f_\delta(r) := \begin{cases} r^2/2 & \text{für } |r| \leq \delta \\ \delta r - \frac{\delta^2}{2} & \text{für } |r| \geq \delta \end{cases}$$

Der Vorteil hierbei ist, dass wir dies durch numerische Verfahren schneller optimieren können, beispielsweise durch das Gauß-Newton-Verfahren.

Overfitting

Overfitting tritt auf, wenn ein Modell zu stark an die Trainingsdaten angepasst wird und dadurch bei neuen, unbekanntem Daten schlecht generalisiert. Dies ist besonders problematisch, wenn das Modell zu komplex ist und zu viele Parameter enthält.

Beispiel (Polynomielle Regression). In der polynomiellen Regression wird ein Polynom höherer Ordnung verwendet, um die Daten anzupassen:

$$y_i = \sum_{j=1}^k w_j x_i^j + \varepsilon_i$$

Während ein höhergradiges Polynom die Trainingsdaten sehr genau modellieren kann, besteht die Gefahr, dass es den Daten zu stark folgt und die zugrunde liegenden Trends nicht mehr zuverlässig erfasst. Dies führt zu einem Modell mit hoher Varianz, das schlecht bei neuen Daten generalisiert.

Eine Möglichkeit, Overfitting zu vermeiden, besteht darin, eine Regularisierung einzuführen, wie es bei der Ridge-Regression der Fall ist. Hierbei wird eine Strafe für große Koeffizienten hinzugefügt. Für einen festen Parameter $\lambda > 0$ betrachten wir:

$$\min_w g(w) := \frac{1}{n} \sum_{i=1}^n |Y_i - w^T X_i|^2 + \lambda \|w\|^2$$

Der Regularisierungsparameter λ steuert das Ausmaß der Bestrafung für große Koeffizienten. Dadurch wird das Modell gezwungen, einfacher und glatter zu bleiben, was die Generalisierungsfähigkeit erhöht und Overfitting reduziert. Äquivalent wäre es, für ein gegebenes $B > 0$ den RSS unter der Nebenbedingung $\|w\| \leq B$ zu minimieren. Die Korrespondenz lässt sich also durch ein indirekt proportionales Verhältnis ausdrücken, im Sinne, dass B klein ist, falls λ groß ist, und andersherum.

Die Lösung der Ridge-Regression ergibt sich durch Minimierung der regularisierten Verlustfunktion, also durch Minimierung des Gradienten:

$$\nabla g(w) = -\frac{2}{n} A^T Y + \frac{2}{n} A^T A w + 2\lambda w \stackrel{!}{=} 0$$

Dies führt zur geschlossenen Form der Lösung:

$$\hat{w}_{ridge} = (\lambda I_d + A^T A)^{-1} A^T Y$$

Ein größeres λ führt hierbei zu mehr Regularität. Insbesondere ist dies numerisch vorteilhaft, denn wir erhalten für $\lambda > 0$ eine einfache Berechnung von \hat{w}_{ridge} via einer QR -Zerlegung.

Hochdimensionaler Parameterraum

In vielen modernen Anwendungen stehen extrem viele Prädiktoren zur Verfügung. Diese Situation, in der die Anzahl der Prädiktoren d größer ist als die Anzahl der Beobachtungen n , führt zu einer Reihe von Schwierigkeiten. Denn falls $d > n$ folgt $\text{Rang}(A) < d$, und somit hat die Gleichung $Aw = \Pi Y$ keine eindeutige Lösung für w .

Eine mögliche Lösung ist es, die Sparsity (Sparsamkeit) zu fördern, also nur wenige, aber relevante Prädiktoren auszuwählen, während alle anderen auf null gesetzt werden. Sei also $s \in \{1, \dots, d\}$, dann betrachten wir:

$$\text{Minimiere } \|Y - Aw\|^2 \text{ unter der Nebenbedingung } \|w\|_0 := |\{j \in \{1, \dots, d\} : w_j \neq 0\}| \leq s$$

Analog können wir, wie oben bemerkt, in diesem Fall die Ridge-Regression betrachten, wobei wir $g(w) := \|Y - Aw\|^2 + \lambda \|w\|_0$ minimieren. Dieses Verfahren bringt jedoch gewisse Probleme mit sich:

- (i) Einerseits ist die Abbildung $w \mapsto \|w\|_0$ nicht stetig und insbesondere nicht konvex. Dies erschwert die Optimierung von g .
- (ii) Insbesondere wird die Größe der Gewichte nicht berücksichtigt.

Eine alternative Methode, die speziell für hochdimensionale Daten entwickelt wurde, ist LASSO (Least Absolute Shrinkage and Selection Operator). Im Gegensatz zur Ridge-Regression minimiert LASSO die Summe der absoluten Werte der Koeffizienten. Wir betrachten also die Abbildung $w \mapsto \|w\|_1$, welche insbesondere lipschitzstetig und konvex ist. Zusätzlich wird auch die Größe der Gewichte berücksichtigt. Es ergibt sich das konvexe Optimierungsproblem:

$$\text{Minimiere } \|Y - Aw\|^2 \text{ unter der Nebenbedingung } \|w\|_1 \leq B.$$

Binäre Response-Variable

Wir betrachten nun eine spezielle Art von Regressionsproblem, bei dem die abhängige Variable $Y_i \in \{0, 1\}$ binär ist. Solche Probleme treten häufig in der Klassifikation auf, wo es darum geht, eine Zugehörigkeit zu einer von zwei Klassen vorherzusagen.

Um binäre Response-Variablen zu modellieren, wird häufig die logistische Regression verwendet. Diese basiert auf der Annahme, dass die abhängige Variable Y_i Bernoulli-verteilt mit Parameter p_i ist.

$$Y_i \sim \text{Bernoulli}(p_i) \quad \text{mit} \quad p_i = \text{sigm}(w^T X_i)$$

Die Sigmoid-Funktion, die p_i berechnet, ist definiert als $\text{sigm}(x) = 1/(1 + e^{-x})$. Es ergibt sich also die Likelihood-Funktion:

$$L(w; Y) = \prod_{i=1}^n \text{sigm}(w^T X_i)^{Y_i} \cdot (1 - \text{sigm}(w^T X_i))^{1-Y_i}$$

Um den Maximum-Likelihood-Schätzer zu berechnen, können verschiedene numerische Optimierungsverfahren eingesetzt werden, beispielsweise das Newton-Verfahren.

7. Bayes-Statistik

Die Bayessche Statistik bietet eine grundlegend andere Herangehensweise an die Schätzung unbekannter Parameter als die klassische frequentistische Statistik. Der entscheidende Unterschied liegt darin, dass bei der Bayes-Statistik Parameter als Zufallsvariablen betrachtet werden. Dies erlaubt es, Unsicherheit über den wahren Wert eines Parameters explizit durch Wahrscheinlichkeitsverteilungen zu modellieren.

Erinnerung: Bayessche Formel

Bevor wir tiefer in die Details der Bayesschen Statistik eintauchen, beginnen wir mit einer Wiederholung der grundlegenden Konzepte der bedingten Wahrscheinlichkeiten, die der Bayesschen Formel zugrunde liegen.

Gegeben seien zwei Zufallsvariablen X und Y auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Die gemeinsame Dichte bzw. Massenfunktion dieser Zufallsvariablen bezeichnen wir mit $f_{X,Y}(x, y)$. Aus der gemeinsamen Dichte lassen sich die Randdichten ableiten:

$$f_X(x) = \int f_{X,Y}(x, y) dy \quad \text{und} \quad f_Y(y) = \int f_{X,Y}(x, y) dx \quad (**)$$

Es folgt die bedingte Dichte von Y gegeben X durch:

$$f_{X|Y}(y, x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{für} \quad f_Y(y) > 0 \quad (*)$$

Satz 7.1 (Satz von Bayes). Seien X, Y zwei Zufallsvariablen mit bedingter Dichte $f_{Y|X}$. So gilt:

$$f_{X|Y}(x|y) = \frac{1}{C(y)} f_X(x) f_{Y|X}(y|x) \quad \text{falls} \quad f_Y(y) > 0 \quad (***)$$

wobei $C(y) = f_Y(y) = \int f_X(x) \cdot f_{Y|X}(y|x) dx$.

Der Satz folgt unmittelbar aus (*) und (**).

Der Satz von Bayes ist eine fundamentale Regel der Wahrscheinlichkeitsrechnung und bildet das Herzstück der Bayesschen Statistik, indem er es ermöglicht, a priori Wahrscheinlichkeiten in a posteriori Wahrscheinlichkeiten umzuwandeln.

7.1. Ansatz der Bayesschen Statistik

In der Bayesschen Statistik betrachten wir den Parameter ϑ als Zufallsvariable. Anstatt einen festen, aber unbekannt Parameterwert zu schätzen, modellieren wir unsere Unsicherheit über diesen Parameter mittels einer Wahrscheinlichkeitsverteilung. Diese Methode erlaubt es, vorhandenes Vorwissen systematisch in den Analyseprozess einzubeziehen.

A-priori

Die *a-priori-Verteilung*, gegeben durch die Dichte $f(\vartheta)$, beschreibt unser Wissen oder unsere Annahmen über den Parameter ϑ vor der Beobachtung von Daten. Diese Verteilung kann subjektive Einschätzungen oder objektives Vorwissen beinhalten und dient als Ausgangspunkt für unsere weitere Analyse.

Likelihood

Die Likelihood-Funktion $f(x|\vartheta) = L(\vartheta, x)$ beschreibt die Wahrscheinlichkeit, die beobachteten Daten x zu erhalten, gegeben den Parameter ϑ . Sie quantifiziert, wie gut verschiedene Werte von ϑ die Daten erklären können.

A-posteriori

Nach der Beobachtung von Daten aktualisieren wir unser Wissen über den Parameter. Dies geschieht durch die Berechnung der a-posteriori-Verteilung $f(\vartheta|x)$, die unser aktualisiertes Wissen nach der Beobachtung von x beschreibt. Diese Verteilung wird durch den Bayesschen Satz berechnet:

$$f(\vartheta) \propto f(x|\vartheta) \cdot f(\vartheta)$$

Die a-posteriori-Verteilung kombiniert somit die Informationen aus den Daten (Likelihood) mit dem Vorwissen (a-priori-Verteilung), um eine aktualisierte Schätzung des Parameters zu liefern.

Schätzer

In der Bayesschen Statistik existieren verschiedene Ansätze zur Schätzung des Parameters ϑ . Zwei gebräuchliche Schätzer sind:

1) MAP = Maximum a-posteriori

Dieser Schätzer $\hat{\vartheta}_{MAP}$ wählt den Wert von ϑ , welcher die a-posteriori-Verteilung maximiert.

$$\hat{\vartheta}_{MAP} \in \text{Domarg max}_{\vartheta} f(\vartheta|x)$$

In der Praxis wird dieser durch numerische Berechnung über Optimierungsverfahren ermittelt.

2) MMS = Minimum Mean Squared

Hier wird der Erwartungswert der a-posteriori-Verteilung als Schätzer verwendet.

$$\hat{\vartheta}_{MMS} = \mathbb{E}[\vartheta|X] = \int \vartheta f(\vartheta|X) d\vartheta = T(X)$$

Dieser minimiert $\mathbb{E}[|T(X) - \vartheta|^2]$ (das Bayes-Risiko). Zur Berechnung betrachten wir:

$$\hat{\vartheta}_{MMS} = \frac{\int \vartheta f(\vartheta) f(X|\vartheta) d\vartheta}{\int f(\vartheta) f(X|\vartheta) d\vartheta}.$$

Die Integrale sind oft nicht einfach zu berechnen (komplizierte Verteilung, hochdimensionaler Parameterraum, ...). Die Berechnung erfolgt also meistens über numerische Verfahren wie (Markov Chain) Monte Carlo-Verfahren.

Posteriori-Intervalle/Bereiche

Neben Punkt-Schätzungen sind Posteriori-Intervalle ein zentrales Konzept in der Bayesschen Statistik. Sei hierfür $\alpha \in (0, 1)$, dann betrachten wir das $1 - \alpha$ Konfidenzintervall der posteriori-Verteilung:

$$I(X) \subseteq \Theta \quad \text{mit} \quad \mathbb{P}[\vartheta \in I(X)|X] = \int_{I(X)} f(\vartheta|X) d\vartheta \geq 1 - \alpha.$$

Der Parameter liegt mit Wahrscheinlichkeit $1 - \alpha$ im Intervall (im Gegensatz zu Konfidenzintervallen).

Posteriori-Verteilung

$$\mu(d\vartheta) = f(\vartheta|x) d\vartheta$$

Die Posteriori-Verteilung enthält die komplette Information für die Bayessche Inferenz.

Beispiel. Seien X_1, \dots, X_n unabhängig und identisch Bernoulli- p -verteilte Zufallsvariablen mit $p \in (0, 1)$. So erhalten wir die Likelihood-Funktion:

$$f(x|p) = p^s(1-p)^{n-s}, \quad \text{wobei } s = \sum_{i=1}^n x_i.$$

Als a-priori-Verteilung lässt sich beispielsweise die Gleichverteilung wählen, gegeben durch:

$$f(p) = 1_{(0,1)}(p) \sim \text{Unif}(0, 1).$$

Dies ist gleichbedeutend damit, dass wir kein Vorwissen besitzen; man spricht auch von einem „flat prior“. Folglich erhalten wir die a-posteriori-Verteilung:

$$f(p|x) \stackrel{\text{Bayessche Formel}}{\propto} f(p)f(x|p) = p^s(1-p)^{n-s} 1_{(0,1)}(p).$$

Wir erhalten also die Beta-Verteilung mit Parametern $s+1$ und $n+1-s$. Der Minimum Mean Square-Schätzer für p ist:

$$\hat{\vartheta}_{MMS} = \text{Mittelwert von } \beta(s+1, n+1-s) = \frac{s+1}{n+2} = \frac{n}{n+2} \cdot \frac{s}{n} + \frac{2}{n+2} \cdot \frac{1}{2}.$$

Hierbei ist s/n der Maximum-Likelihood-Schätzer $\hat{\vartheta}$ und $1/2$ der Mittelwert der a-priori-Verteilung. Als MMS erhalten wir also eine Konvexkombination der beiden. Für große n gilt $\hat{\vartheta}_{MMS} \approx \hat{\vartheta}$, für kleine n unterscheiden sich die Schätzer, da $\hat{\vartheta}_{MMS}$ das Vorwissen mit berücksichtigt. Allgemeiner gilt, dass bei gegebener a-priori-Verteilung $\beta(a, b)$ sich die a-posteriori-Verteilung $\beta(a+s, b+n-s)$ ergibt.

Lineares Gauss-Modell, Bayes-Variante

Wir erinnern uns an das lineare Gauß-Modell, welches eine lineare Beziehung zwischen den abhängigen Variablen $Y \in \mathbb{R}^n$ und den unabhängigen Variablen $A \in \mathbb{R}^{n \times d}$ beschreibt. Diese Abhängigkeit ist gegeben durch:

$$Y = A\omega + \varepsilon, \quad \text{wobei } \varepsilon \sim N(0, C).$$

Hierbei ist der Fehler $\varepsilon \in \mathbb{R}^n$ normalverteilt mit symmetrischer und positiv definierter Kovarianzmatrix $C \in \mathbb{R}^{n \times n}$. Beispielsweise $C = v \cdot I_n$, womit angenommen wird, dass die Fehler unabhängig und identisch verteilt sind. Wir erhalten die Likelihood:

$$f(y|\omega) \propto \exp\left(-\frac{1}{2}(y - A\omega)^T \cdot C^{-1}(y - A\omega)\right).$$

Mit der a-priori-Verteilung $\omega \sim N(0, C_{\text{prior}})$, wobei $C_{\text{prior}} \in \mathbb{R}^{d \times d}$ symmetrisch und positiv definit ist.

Satz 7.2. Die a-posteriori-Verteilung im linearen Gauss-Modell ist $N(\hat{\omega}, C_{\text{post}})$.

$$C_{\text{post}}^{-1} = C_{\text{prior}}^{-1} + A^T C^{-1} A, \quad \hat{\omega} = C_{\text{post}} A^T C^{-1} Y$$

Man spricht bei den Inversen der jeweiligen Kovarianzmatrizen auch von der „prior“- und der „posterior precision“, wobei C^{-1} auch die „precision of observation“ ist.

Beweis. Der Beweis ist Übungsaufgabe. ■

Wenn $C = v \cdot I_n$ angenommen wird, also die Fehler unabhängig und identisch verteilt sind, vereinfacht sich die Formel für die a-posteriori-Präzision:

$$C_{\text{post}}^{-1} = C_{\text{prior}}^{-1} + \frac{1}{v} A^T A, \quad \hat{\omega} = \left(v \cdot C_{\text{prior}}^{-1} + A^T A \right)^{-1} A^T Y$$

Falls insbesondere $C_{\text{prior}}^{-1} = 0$, wir also keine a-priori-Information haben, stimmt $\hat{\omega}$ mit dem Least-Squares-Schätzer (LSE) überein.

Beispiele. 1) Gaußsches Produktmodell

Betrachten wir das Beispiel einer einfachen Regression, bei der die Abhängigkeit gegeben ist durch

$$Y_i = \omega + \varepsilon_i, \quad \text{mit } \varepsilon_i \sim N(0, v \cdot I_n), \quad A = I_n, \quad A^T \cdot A = n$$

Die a-priori-Verteilung von ω sei ebenfalls normalverteilt mit $\omega \sim N(0, u)$, wobei $u > 0$ ist. Die a-posteriori-Verteilung von ω ist dann:

$$\omega \sim N(\hat{\omega}, u_{\text{post}})$$

Hierbei ist die Präzision der a-posteriori-Verteilung gegeben durch:

$$\frac{1}{u_{\text{post}}} = \underbrace{\frac{1}{u}}_{\text{Präzision des Priors}} + \underbrace{\frac{1}{\sigma^2}}_{\text{Präzision von } \bar{Y}}$$

$$\hat{\omega} = \left(\frac{v}{u} + n \right)^{-1} \bar{Y} + \frac{\frac{v}{n}}{u + \frac{v}{n}} \cdot 0$$

wobei \bar{Y} der Maximum-Likelihood-Schätzer ist und 0 der Mittelwert der a-priori-Verteilung.

2) Bayessche lineare Regression

In der bayesschen linearen Regression wird das Modell

$$Y_i = X_i^T \omega + \varepsilon_i$$

betrachtet, wobei die Fehler ε_i unabhängig und $N(0, v)$ verteilt sind. Eine übliche Wahl für die a-priori-Verteilung ist $\omega \sim N(0, \lambda^{-1} I_d)$, wobei λ die „prior precision“ darstellt. Die a-posteriori-Verteilung ergibt sich dann durch:

$$f(\omega|y) \propto f(\omega) f(y|\omega) \propto \exp \left(-\frac{1}{2} \left(\lambda \|\omega\|^2 + \frac{1}{v} \|y - A\omega\|^2 \right) \right)$$

Maximal wird diese Verteilung für $\omega = \hat{\omega}_{\text{ridge}}$. Das heißt, $\hat{\omega}_{\text{ridge}}$ ist der MAP-Schätzer (Maximum A Posteriori-Schätzer).

Probleme

Die Wahl eines geeigneten a-priori ist oft problematisch. Einige der Herausforderungen sind:

- 1) In der Praxis ist C oft unbekannt, was bedeutet, dass wir a-priori-Verteilungen auch für C annehmen müssen, was das Modell komplexer macht.
- 2) Es ist oft schwierig, den richtigen a-priori auszuwählen. Eine Möglichkeit ist die Verwendung eines uninformativen Priors. Ein Beispiel hierfür wäre der Jeffrey's prior:

$$f(\vartheta) \propto \frac{1}{\sqrt{I(\vartheta)}}$$

- 3) In vielen Fällen ist die Berechnung der a-posteriori-Verteilung nicht einfach, insbesondere wenn der a-priori nicht konjugiert ist. Hier können Methoden wie die (Markov-Chain) Monte-Carlo-Verfahren helfen.

7.2. Markov Chain Monte Carlo (MCMC)-Verfahren

In der bayesschen Statistik ist es häufig schwierig, die posterioren Verteilungen analytisch zu berechnen, insbesondere bei komplizierten Modellen. Um dennoch zuverlässige Schätzungen zu erhalten, greift man auf numerische Verfahren wie die Markov Chain Monte Carlo (MCMC)-Methoden zurück. Diese Techniken ermöglichen es, Stichproben aus der posterioren Verteilung zu ziehen und somit Schätzungen für unbekannte Parameter sowie deren Unsicherheit zu erhalten.

Die Grundidee der MCMC-Verfahren besteht darin, eine Markov-Kette zu konstruieren, deren stationäre Verteilung der gesuchten posterioren Verteilung entspricht. Man startet mit einem zufälligen Wert und generiert durch eine Folge von Schritten eine Kette von Werten, die nach einer gewissen „Einschwingphase“ (burn-in) Verteilungen aus der Zielverteilung entnommen sind.

Wir erhalten also eine Markov-Kette $(\vartheta_n)_{n=0,1,2,\dots}$ auf Θ mit Gleichgewicht μ_{post} . Falls diese insbesondere ergodisch ist, erhalten wir, dass die empirische Verteilung der ϑ_i für große n annähernd der Posteriori-Verteilung μ_{post} entspricht. Um den anfänglichen „burn-in“ nicht mit einzubeziehen, betrachten wir jedoch den abgeänderten Schätzer:

$$\hat{\mu}_{\text{post}} = \frac{1}{B} \sum_{i=b}^{b+B-1} \delta_{\vartheta_i} \quad \text{als Schätzer für } \mu_{\text{post}}$$

Wobei b das „burn-in“ ist und B sowie b „hinreichend groß“ sind.

Quantifizierung der Posteriori-Verteilung

Nehmen wir an, wir möchten die posteriori-Verteilung approximieren. Die MCMC-Verfahren generieren eine Folge von Werten:

$$\vartheta_1, \dots, \vartheta_B \sim f(\vartheta|X) d\vartheta$$

welche unabhängig und posteriori verteilt sind. Der Schätzer für die Verteilung ist dann:

$$\mu_{\text{post}} \approx \frac{1}{B} \sum_{i=1}^B \delta_{\vartheta_i} = \hat{\mu}_{\text{post}}$$

$$\hat{\vartheta}_{\text{MMS}} \approx \frac{1}{B} \sum_{i=1}^B \vartheta_i = \hat{\vartheta}_B^{\text{MC}}$$

Das approximative Posteriori-Intervall $(\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2})$ lässt sich durch die Quantile der gegebenen empirischen Verteilung $\hat{\mu}_{\text{post}}$ bestimmen. Je mehr Stichproben verwendet werden, desto genauer wird die Schätzung der Posterior-Verteilung. Der entscheidende Vorteil der MCMC-Methoden ist ihre Anwendbarkeit auch in komplexen, hochdimensionalen Modellen, bei denen direkte Simulationsverfahren nicht praktikabel sind.

Gibbs-Sampler

Eine der bekanntesten MCMC-Methoden ist der Gibbs-Sampler, der besonders dann effektiv ist, wenn die bedingten Verteilungen der Parameter in einem Modell einfach zu berechnen sind. Der Gibbs-Sampler iteriert dabei über alle Parameter und aktualisiert jeden Parameter abwechselnd, wobei die anderen Parameter jeweils als fest gegeben betrachtet werden. Konkreter können wir den Algorithmus wie folgt beschreiben:

- 0) Wähle einen Startwert $\vartheta^{(0)} \in \mathbb{R}^d$ und setze $n = 0$.
- 1) Simuliere ein $j \sim \text{Unif}\{1, \dots, d\}$.

2) Iteratives Sampling der folgenden $\vartheta^{(n)}$ wie folgt:

$$\vartheta_{n+1,k} := \begin{cases} \vartheta_{n,k} & \text{falls } k \neq j \\ \text{Stichprobe der bedingten posteriori-Verteilung der } k\text{-ten Komponente} & \\ \text{gegeben der } l\text{-ten Komponente } \vartheta_{n,l} \quad \forall l \neq k & \text{falls } k = j \end{cases}$$

3) Setze n auf $n + 1$ und iteriere ab Schritt 2).

Der Gibbs-Sampler hat den Vorteil, dass er relativ einfach zu implementieren ist, insbesondere in Fällen, in denen die bedingten Verteilungen bekannt und leicht zu berechnen sind. Das Verfahren konvergiert zudem oft schnell zur stationären Verteilung, was es zu einem effizienten Werkzeug für die numerische Berechnung von Posterior-Verteilungen macht.

A. Ergänzungen aus der Wahrscheinlichkeitstheorie

A.1. Kovarianz, Korrelation und lineare Prognosen

Kovarianz und Korrelation

Für Zufallsvariablen $X, Y \in \mathcal{L}^2$ können wir die Kovarianz und die Korrelation definieren.

Definition A.1. Seien X und Y Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$.

- (i) Die **Kovarianz** von X und Y ist definiert als

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

- (ii) Gilt $\sigma[X]\sigma[Y] \neq 0$, so heißt

$$\varrho[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma[X]\sigma[Y]}$$

Korrelationskoeffizient von X und Y .

- (iii) Die Zufallsvariablen X und Y heißen **unkorreliert**, falls $\text{Cov}[X, Y] = 0$, d.h. falls

$$E[XY] = E[X] \cdot E[Y].$$

Gilt $\text{Cov}[X, Y] > 0$ bzw. < 0 , dann heißen X und Y **positiv** bzw. **negativ korreliert**.

Satz A.2 (Cauchy-Schwarz-Ungleichung für Kovarianz).

- (i) Die Kovarianz ist eine symmetrische und bilineare Abbildung von $\mathcal{L}^2 \times \mathcal{L}^2$ nach \mathbb{R} mit

$$\text{Cov}[X, X] = \text{Var}[X] \geq 0 \quad \text{für alle } X \in \mathcal{L}^2.$$

- (ii) Für $X, Y \in \mathcal{L}^2$ gilt die *Cauchy-Schwarz-Ungleichung*

$$|\text{Cov}[X, Y]| \leq \sqrt{\text{Var}[X]} \cdot \sqrt{\text{Var}[Y]} = \sigma[X] \cdot \sigma[Y]. \quad (\text{A.1})$$

Insbesondere gilt für den Korrelationskoeffizienten im Fall $\sigma[X] \cdot \sigma[Y] \neq 0$ stets

$$|\varrho[X, Y]| \leq 1. \quad (\text{A.2})$$

- (iii) Gleichheit gilt in den Ungleichungen (A.1) bzw. (A.2) genau dann, wenn Konstanten $a, b \in \mathbb{R}$ existieren, sodass

$$Y = aX + b \quad \text{mit Wahrscheinlichkeit 1.}$$

In diesem Fall ist $\varrho[X, Y] = 1$ falls $a > 0$, und $\varrho[X, Y] = -1$ falls $a < 0$.

Beweis. Nach Definition gilt $\text{Cov}[X, Y] = \text{Cov}[Y, X]$ und $\text{Cov}[X, X] = \text{Var}[X]$. Außerdem folgt aus der Linearität des Erwartungswerts für $X, Y, Z \in \mathcal{L}^2$ und $a \in \mathbb{R}$:

$$\text{Cov}[X, aY + Z] = E[(X - E[X])(aY + Z - E[aY + Z])] = a \text{Cov}[X, Y] + \text{Cov}[X, Z].$$

Somit ist die Kovarianz linear in der zweiten Komponente und damit wegen der Symmetrie auch bilinear. Cov ist also eine nicht-negative definite symmetrische Bilinearform auf dem Vektorraum \mathcal{L}^2 . Damit gilt insbesondere die Cauchy-Schwarz-Ungleichung, siehe die Vorlesung LINEARE ALGEBRA. Den letzten Teil der Aussage und auch die Cauchy-Schwarz-Ungleichung werden wir gleich nebenbei im Rahmen eines Exkurses zu linearen Prognosen beweisen. ■

Lineare Prognosen

Angenommen, wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable $Y : \Omega \rightarrow \mathbb{R}$ beschrieben wird. Welches ist der *beste Prognosewert* b für $Y(\omega)$, wenn uns keine weiteren Informationen zur Verfügung stehen?

Die Antwort hängt offensichtlich davon ab, wie wir den Prognosefehler messen. Häufig verwendet man den mittleren quadratischen Fehler (*Mean Square Error*)

$$\text{MSE} = E[(Y - b)^2].$$

Satz A.3 (Erwartungswert als bester Prognosewert im quadratischen Mittel). Ist Y eine Zufallsvariable in $L^2(\Omega, \mathcal{A}, P)$, dann gilt für alle $b \in \mathbb{R}$:

$$E[(Y - b)^2] = \text{Var}[Y] + (b - E[Y])^2 \geq E[(Y - E[Y])^2].$$

Der mittlere quadratische Fehler des Prognosewertes b ist also die Summe der Varianz von Y und des Quadrats des systematischen bzw. mittleren Prognosefehlers (engl. *Bias*) $b - E[Y]$:

$$\text{MSE} = \text{Varianz} + \text{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für $b = E[Y]$ minimal.

Beweis. Für $b \in \mathbb{R}$ gilt wegen der Linearität des Erwartungswertes:

$$E[(Y - b)^2] = \text{Var}[Y - b] + E[Y - b]^2 = \text{Var}[Y] + (E[Y] - b)^2. \quad \blacksquare$$

Seien nun $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ quadratintegrierbare Zufallsvariablen mit $\sigma[X] \neq 0$. Angenommen, wir kennen bereits den Wert $X(\omega)$ in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) = aX(\omega) + b, \quad (a, b \in \mathbb{R}) \quad (\text{A.3})$$

für $Y(\omega)$ im quadratischen Mittel. Zu minimieren ist jetzt der mittlere quadratischen Fehler

$$\text{MSE} := E[(\hat{Y} - Y)^2]$$

unter allen Zufallsvariablen \hat{Y} , die affine Funktionen von X sind. In diesem Fall erhalten wir

$$\text{MSE} = \text{Var}[Y - \hat{Y}] + E[Y - \hat{Y}]^2 = \text{Var}[Y - aX] + (E[Y] - aE[X] - b)^2.$$

Den zweiten Term können wir für gegebenes a minimieren, indem wir

$$b = E[Y] - aE[X]$$

wählen. Für den ersten Term ergibt sich

$$\begin{aligned}\text{Var}[Y - aX] &= \text{Cov}[Y - aX, Y - aX] = \text{Var}[Y] - 2a \text{Cov}[X, Y] + a^2 \text{Var}[X] \\ &= \left(a \cdot \sigma[X] - \frac{\text{Cov}[X, Y]}{\sigma[X]} \right)^2 + \text{Var}[Y] - \frac{\text{Cov}[X, Y]^2}{\text{Var}[X]}.\end{aligned}\quad (\text{A.4})$$

Dieser Ausdruck wird minimiert, wenn wir $a = \text{Cov}[X, Y]/\sigma[X]^2$ wählen. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für Y gestützt auf X ist dann

$$\hat{Y}_{\text{opt}} = aX + b = E[Y] + a(X - E[X]).$$

Damit haben wir gezeigt:

Satz A.4 (Lineare Prognose/Regression von Y gestützt auf X). Der mittlere quadratische Fehler $E[(\hat{Y} - Y)^2]$ ist minimal unter allen Zufallsvariablen der Form $\hat{Y} = aX + b$ mit $a, b \in \mathbb{R}$ für

$$\hat{Y}(\omega) = E[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \cdot (X(\omega) - E[X]).$$

Das Problem der linearen Prognose steht in engem Zusammenhang mit der Cauchy-Schwarz-Ungleichung für die Kovarianz. In der Tat ergibt sich diese Ungleichung unmittelbar aus Gleichung (A.4):

Beweis (Cauchy-Schwarz-Ungleichung, Satz A.2 (ii) und (iii)). Im Fall $\sigma[X] = 0$ gilt $X = E[X]$ mit Wahrscheinlichkeit 1, und die Ungleichung (A.1) ist trivialerweise erfüllt. Wir nehmen nun an, dass $\sigma[X] \neq 0$ gilt. Wählt man dann wie oben $a = \text{Cov}[X, Y]/\sigma[X]^2$, so folgt aus (A.4) die Cauchy-Schwarz-Ungleichung

$$\text{Var}[Y] - \frac{\text{Cov}[X, Y]^2}{\text{Var}[X]} \geq 0.$$

Die Ungleichung (A.2) folgt unmittelbar. Zudem erhalten wir nach (A.4) genau dann Gleichheit in (A.1) bzw. (A.2), wenn $\text{Var}[Y - aX] = 0$ gilt, also wenn $Y - aX$ mit Wahrscheinlichkeit 1 konstant ist. In diesem Fall folgt $\text{Cov}[X, Y] = \text{Cov}[X, aX] = a \text{Var}[X]$, also hat $\varrho[X, Y]$ dasselbe Vorzeichen wie a . ■

Beispiel (Regressionsgerade, Methode der kleinsten Quadrate). Wenn die gemeinsame Verteilung von X und Y eine empirische Verteilung von Daten $(x_i, y_i) \in \mathbb{R}^2, i = 1, \dots, n$, ist, d.h. wenn

$$(X, Y) = (x_i, y_i) \quad \text{mit Wahrscheinlichkeit } 1/n$$

für $1 \leq i \leq n$ gilt, dann sind die Erwartungswerte und die Kovarianz gegeben durch

$$\begin{aligned}E[X] &= \frac{1}{n} \sum_{i=1}^n x_i =: \bar{x}_n, & E[Y] &= \bar{y}_n, \\ \text{Cov}[X, Y] &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n} \left(\sum_{i=1}^n x_i y_i \right) - \bar{x}_n \bar{y}_n.\end{aligned}$$

Der entsprechende *empirische Korrelationskoeffizient* der Daten $(x_i, y_i), 1 \leq i \leq n$, ist

$$\varrho[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma[X]\sigma[Y]} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)^{1/2} \left(\sum_{i=1}^n (y_i - \bar{y}_n)^2 \right)^{1/2}}$$

Dieses verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekanntem Verteilungen. Die Grafiken in Abbildung A.1 zeigen Datensätze mit verschiedenen Korrelationskoeffizienten ρ .

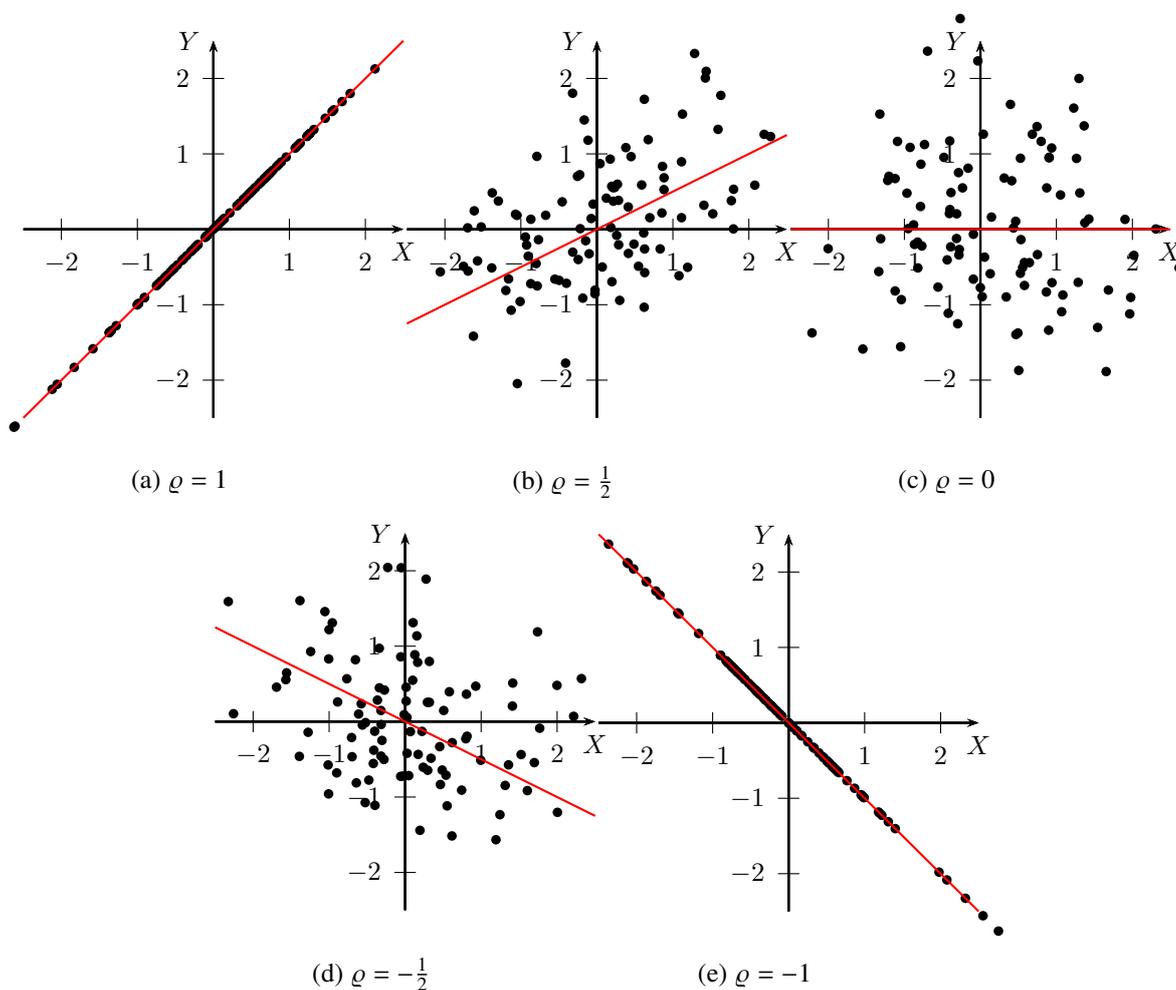


Abbildung A.1.: Korrelationskoeffizienten und Regressionsgeraden für verschiedene Datensätze

Als beste lineare Prognose von Y gestützt auf X im quadratischen Mittel erhalten wir die *Regressionsgerade* $y = ax + b$, die die Quadratsumme

$$\sum_{i=1}^n (ax_i + b - y_i)^2 = n \cdot \text{MSE}$$

der Abweichungen minimiert. Hierbei gilt nach Satz A.4:

$$a = \frac{\text{Cov}[X, Y]}{\sigma[X]^2} = \frac{\sum(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum(x_i - \bar{x}_n)^2} \quad \text{und} \quad b = E[Y] - a \cdot E[X] = \bar{y}_n - a \cdot \bar{x}_n.$$

Die Regressionsgeraden sind in Abbildung A.1 eingezeichnet.

A.2. Wahrscheinlichkeitsverteilungen im \mathbb{R}^n

Transformation von mehrdimensionalen Dichten

Absolutstetige reellwertige Zufallsvariablen X_1, \dots, X_n sind genau dann unabhängig, wenn ihre gemeinsame Verteilung absolutstetig ist mit einer Dichte, die sich in Produktform darstellen lässt:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n g_i(x_i), \quad g_i : \mathbb{R} \rightarrow [0, \infty) \text{ meßbar.}$$

In diesem Fall sind die Dichten der einzelnen Zufallsvariablen X_i proportional zu den Funktionen g_i . Modelle mit komplizierterer Abhängigkeitsstruktur können manchmal durch geeignete Transformationen in eine (vollständige oder partielle) Produktform gebracht werden.

Satz A.5 (Mehrdimensionaler Dichtetransformationssatz). Seien $S, T \subseteq \mathbb{R}^n$ offen, und sei $X : \Omega \rightarrow S$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum (Ω, \mathcal{A}, P) mit absolutstetiger Verteilung μ_X mit Dichte f_X . Ist $\psi : S \rightarrow T$ ein Diffeomorphismus (C^1) mit $\det D\psi(x) \neq 0$ für alle $x \in S$, dann ist die Verteilung von $\psi(X)$ absolutstetig mit Dichte

$$f_{\psi(X)}(y) = f_X(\psi^{-1}(y)) \cdot |\det D\psi^{-1}(y)|, \quad (\text{A.5})$$

wobei $\det D\psi^{-1}(y) = \det\left(\frac{\partial x_i}{\partial y_j}\right)$ die Jacobideterminante der Koordinatentransformation ist.

Beweis. Die Behauptung folgt aus dem Transformationssatz der multivariaten Analysis:

$$\begin{aligned} P[\psi(X) \in B] &= P[X \in \psi^{-1}(B)] \\ &= \int_{\psi^{-1}(B)} f_X(x) dx \stackrel{\text{Subst.}}{=} \int_B f_X(\psi^{-1}(y)) \cdot |\det D\psi^{-1}(y)| dy. \quad \blacksquare \end{aligned}$$

Bemerkung (Volumentransformation). Der Zusatzfaktor $|\det D\psi^{-1}(y)|$ in (A.5) beschreibt die Transformation des Volumens (also des Lebesguemaßes) bei Anwenden der Abbildung ψ^{-1} . Anschaulich wird ein infinitesimaler Quader am Punkt y mit Volumen $dy = dy_1 \cdots dy_n$ durch die Abbildung ψ^{-1} auf ein infinitesimales Parallelepipiped am Punkt $\psi^{-1}(y)$ abgebildet, das von den Vektoren $\frac{\partial \psi^{-1}}{\partial y_i}(y) dy_i$ ($i = 1, \dots, n$) aufgespannt wird. Das Volumen dieses infinitesimalen Parallelepipeds beträgt

$$\left| \det \left(\frac{\partial \psi^{-1}}{\partial y_1}(y) dy_1, \dots, \frac{\partial \psi^{-1}}{\partial y_n}(y) dy_n \right) \right| = |\det D\psi^{-1}(y)| dy.$$

Für einen rigorosen Beweis der mehrdimensionalen Substitutionsformel verweisen wir auf die Analysisvorlesung.

Multivariate Normalverteilungen

Sei $Z = (Z_1, Z_2, \dots, Z_n)$ mit unabhängigen, standardnormalverteilten Zufallsvariablen Z_i . Die Verteilung des Zufallsvektors Z ist dann absolutstetig bzgl. des Lebesguemaßes im \mathbb{R}^n mit Dichte

$$f_Z(x) = \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \right) = (2\pi)^{-n/2} e^{-|x|^2/2} \quad (n\text{-dimensionale Standardnormalverteilung}).$$

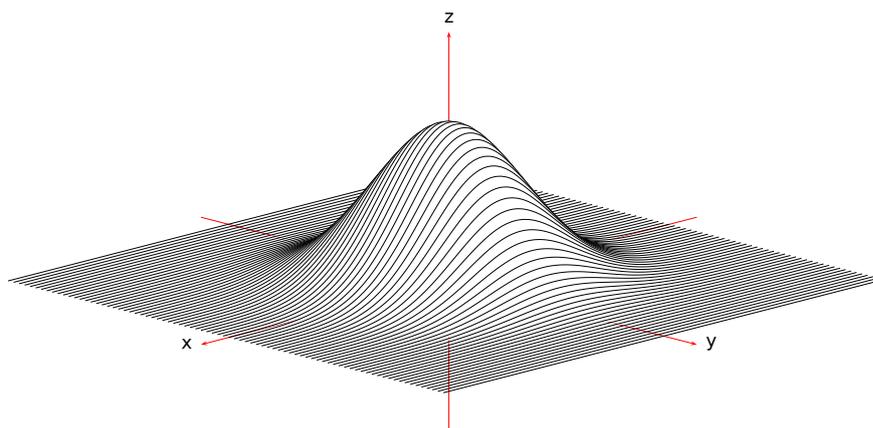


Abbildung A.2.: Dichte der Standardnormalverteilung in \mathbb{R}^2 .

Sei nun $m \in \mathbb{R}^n$, und $\sigma \in \mathbb{R}^{n \times n}$ eine $n \times n$ -Matrix. Wir betrachten den Zufallsvektor

$$Y = \sigma Z + m .$$

Ist σ regulär, dann können wir die Dichte der Verteilung von Y bzgl. des Lebesgue-Maßes im \mathbb{R}^n mithilfe des Transformationssatzes explizit berechnen. Mit $C := \sigma \sigma^T$ erhalten wir

$$\begin{aligned} f_Y(y) &= f_X(\sigma^{-1}(y - m)) \cdot |\det \sigma^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\left(-\frac{1}{2}(y - m) \cdot C^{-1}(y - m)\right) . \end{aligned}$$

Ist σ nicht regulär, dann nimmt X nur Werte in einem echten Unterraum des \mathbb{R}^n an. Die Verteilung von X ist in diesem Fall *nicht absolutstetig* bzgl. des Lebesgue-Maßes im \mathbb{R}^n .

Definition A.6 (Normalverteilung im \mathbb{R}^n). Sei $m \in \mathbb{R}^n$, und sei $C \in \mathbb{R}^{n \times n}$ eine symmetrische, positiv definite Matrix. Die Verteilung $N(m, C)$ im \mathbb{R}^n mit Dichte f_Y heißt **n -dimensionale Normalverteilung** mit Mittelwertvektor m und Kovarianzmatrix C .

Wir werden unten nachrechnen, dass die Kovarianzen der Komponenten eines Zufallsvektors $Y \sim N(m, C)$ tatsächlich durch die Einträge C_{ij} der Matrix C gegeben sind.

Beispiel (Zufällige Punkte in der Ebene). Sind X und Y unabhängige, $N(0, \sigma^2)$ -verteilte Zufallsvariablen auf (Ω, \mathcal{A}, P) mit $\sigma > 0$, dann ist die gemeinsame Verteilung $\mu_{X,Y}$ absolutstetig mit Dichte

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (x, y) \in \mathbb{R}^2.$$

Insbesondere gilt $(X, Y) \neq (0, 0)$ P -fast sicher. Wir definieren den Radial- und Polaranteil

$$R : \Omega \rightarrow (0, \infty), \quad \Phi : \Omega \rightarrow [0, 2\pi)$$

durch

$$X = R \cdot \cos \Phi \quad \text{und} \quad Y = R \cdot \sin \Phi,$$

d.h. $R = \sqrt{X^2 + Y^2}$ und $\Phi = \arg(X + iY)$ falls $(X, Y) \neq (0, 0)$. Auf der Nullmenge $\{(X, Y) = (0, 0)\}$ definieren wir (R, Φ) in beliebiger Weise, sodass sich messbare Funktionen ergeben. Wir berechnen nun

die gemeinsame Verteilung von R und Φ :

$$\begin{aligned} P[R \leq r_0, \Phi \leq \varphi_0] &= P[(X, Y) \in \text{„Kuchenstück“ mit Winkel } \varphi_0 \text{ und Radius } r_0] \\ &= \int\int_{\text{„Kuchenstück“}} f_{X,Y}(x, y) \, dx \, dy \\ &= \int_0^{r_0} \int_0^{\varphi_0} f_{X,Y}(r \cos \varphi, r \sin \varphi) \underbrace{r}_{\substack{\text{Jacobideterminante} \\ \text{der Koordinatentransf.}}} \, d\varphi \, dr \\ &= \int_0^{r_0} \int_0^{\varphi_0} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)} \, d\varphi \, dr. \end{aligned}$$

Hierbei haben wir im 3. Schritt den Transformationssatz (Substitutionsregel) für mehrdimensionale Integrale verwendet - der Faktor r ist die Jacobideterminante der Koordinatentransformation. Es folgt, dass die gemeinsame Verteilung $\mu_{R,\Phi}$ absolutstetig ist mit Dichte

$$f_{R,\Phi}(r, \varphi) = \frac{1}{2\pi} \cdot \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)}.$$

Da die Dichte Produktform hat, sind R und Φ unabhängig. Die Randverteilung μ_Φ ist absolutstetig mit Dichte

$$f_\Phi(\varphi) = \text{const.} = \frac{1}{2\pi} \quad (0 \leq \varphi < 2\pi),$$

d.h. Φ ist gleichverteilt auf $[0, 2\pi)$. Somit ist μ_R absolutstetig mit Dichte

$$f_R(r) = \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)} \quad (r > 0).$$

Die Berechnung können wir verwenden, um Stichproben von der Standardnormalverteilung zu simulieren:

Beispiel (Simulation von normalverteilten Zufallsvariablen). Die Verteilungsfunktion einer $N(0, 1)$ -verteilten Zufallsvariable X ist

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} \, dt.$$

Das Integral ist nicht explizit lösbar und die Inverse F_X^{-1} ist dementsprechend nur approximativ berechenbar. Daher ist die Simulation einer Standardnormalverteilung durch Inversion der Verteilungsfunktion relativ aufwendig. Ein einfacheres Simulationsverfahren ergibt sich, wenn wir eine zweidimensionale Standardnormalverteilung betrachten und auf Polarkoordinaten transformieren. Dann gilt für den Radialanteil:

$$F_R(s) = \int_0^s e^{-r^2/2} r \, dr = 1 - e^{-s^2/2} \quad \text{für alle } s \geq 0.$$

Das Integral ist also explizit berechenbar, und

$$F_R^{-1}(u) = \sqrt{-2 \log(1-u)}, \quad u \in (0, 1).$$

Der Winkelanteil Φ ist unabhängig von R und gleichverteilt auf $[0, 2\pi)$. Wir können Zufallsvariablen mit der entsprechenden gemeinsamen Verteilung erzeugen, indem wir

$$\begin{aligned} \Phi &:= 2\pi U_1, \\ R &:= \sqrt{-2 \log(1-U_2)} \quad \left(\text{bzw.} = \sqrt{-2 \log U_2} \right), \end{aligned}$$

setzen, wobei U_1 und U_2 unabhängige, auf $(0, 1)$ gleichverteilte Zufallsvariablen sind. Stichproben von U_1 und U_2 können durch Pseudozufallszahlen simuliert werden. Die Zufallsvariablen

$$X := R \cos \Phi \quad \text{und} \quad Y := R \sin \Phi$$

sind dann unabhängig und $N(0, 1)$ -verteilt. Für $m \in \mathbb{R}$ und $\sigma > 0$ sind $\sigma X + m$ und $\sigma Y + m$ unabhängige $N(m, \sigma^2)$ -verteilte Zufallsvariable.

Wir erhalten also den folgenden Algorithmus zur Simulation von Stichproben einer Normalverteilung:

Algorithmus 1: Box-Muller-Verfahren

Input : $m \in \mathbb{R}, \sigma > 0$

Output Unabhängige Stichproben \tilde{x}, \tilde{y} von $N(m, \sigma^2)$

- ⋮
- 1 Erzeuge unabhängige Zufallszahlen $u_1, u_2 \sim \text{Unif}_{(0,1)}$;
 - 2 $x := \sqrt{-2 \log u_1} \cos(2\pi u_2), y := \sqrt{-2 \log u_1} \sin(2\pi u_2)$;
 - 3 $\tilde{x} := \sigma x + m, \tilde{y} := \sigma y + m$;
 - 4 **return** x, y ;
-

Ordnungsstatistiken, Beta-Verteilung

Ist die Verteilung von unabhängigen, identisch verteilten Zufallsvariablen X_1, \dots, X_n absolutstetig mit Dichte f , dann kann man die Dichte der gemeinsamen Verteilung der Ordnungsstatistiken $X_{(1)} \leq \dots \leq X_{(n)}$ mit einem Symmetrieargument berechnen. Dazu bemerken wir, dass für beliebige Permutationen $\pi \in \mathcal{S}_n$

$$(X_{\pi(1)}, \dots, X_{\pi(n)}) \sim (X_1, \dots, X_n)$$

gilt, da die gemeinsame Dichte $\prod_{i=1}^n f(x_i)$ von X_1, \dots, X_n invariant unter Permutationen der Koordinaten ist. Wegen $P[X_i = X_j] = 0$ für $i \neq j$ erhalten wir damit

$$\begin{aligned} P[X_{(1)} \leq c_1, \dots, X_{(n)} \leq c_n] &= \sum_{\pi \in \mathcal{S}_n} P[X_{\pi(1)} \leq c_1, \dots, X_{\pi(n)} \leq c_n, X_{\pi(1)} < \dots < X_{\pi(n)}] \\ &= n! P[X_1 \leq c_1, \dots, X_n \leq c_n, X_1 < X_2 < \dots < X_n] \\ &= n! \int_{-\infty}^{c_1} \dots \int_{-\infty}^{c_n} I_{\{y_1 < y_2 < \dots < y_n\}} f(y_1) \dots f(y_n) dy_1 \dots dy_n. \end{aligned}$$

Hieraus folgt, dass die gemeinsame Verteilung von $X_{(1)}, \dots, X_{(n)}$ absolutstetig ist mit Dichte

$$f_{X_{(1)}, \dots, X_{(n)}}(y_1, \dots, y_n) = n! \cdot I_{\{y_1 < y_2 < \dots < y_n\}} f(y_1) \dots f(y_n).$$

Durch Aufintegrieren erhält man daraus mithilfe des Satzes von Fubini und einer erneuten Symmetrieüberlegung die Dichten der Verteilungen der einzelnen Ordnungsstatistiken:

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1} (1-F(y))^{n-k} f(y).$$

Beispiel (Beta-Verteilungen). Sind die Zufallsvariablen X_i auf $(0, 1)$ gleichverteilt, dann hat $X_{(k)}$ die Dichte

$$f_{X_{(k)}}(u) = B(k, n-k+1)^{-1} u^{k-1} (1-u)^{n-k} I_{(0,1)}(u)$$

mit Normierungskonstante

$$B(a, b) = \int_0^1 u^{a-1} (1-u)^{b-1} du \quad \left(= \frac{(a-1)!(b-1)!}{(a+b-1)!} \text{ für } a, b \in \mathbb{N} \right).$$

Die entsprechende Verteilung heißt *Beta-Verteilung mit Parametern* $a, b > 0$, die Funktion B ist die *Eulersche Beta-Funktion*.

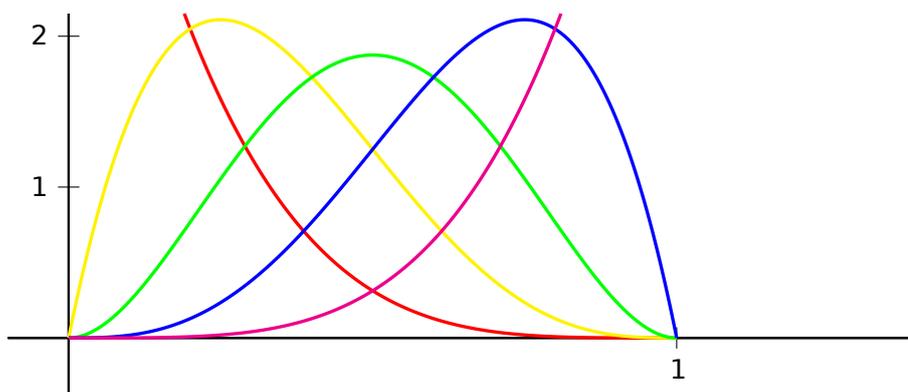


Abbildung A.3.: Abbildung der Dichtefunktionen der Ordnungsstatistiken $X_{(1)}, \dots, X_{(5)}$ (rot, gelb, grün, blau, magenta) bzgl. der Gleichverteilung auf $(0, 1)$.

A.3. Charakteristische Funktionen und mehrdimensionaler zentraler Grenzwertsatz

Momentenerzeugende und charakteristische Funktionen

Sei (Ω, \mathcal{A}, P) ein Wahrscheinlichkeitsraum und $X : \Omega \rightarrow \mathbb{R}^d$ eine Zufallsvariable mit Verteilung μ . Wir definieren den Erwartungswert (bzw. das Lebesgue-Integral bzgl. P) für eine komplexwertige Zufallsvariable $Z = U + iV$ mit Real- und Imaginärteil $U, V : \Omega \rightarrow \mathbb{R}$ durch $E[Z] = E[U] + iE[V]$. Der Erwartungswert ist immer dann definiert, wenn $|Z| = \sqrt{U^2 + V^2}$ integrierbar ist, da dann U und V integrierbare reellwertige Zufallsvariablen sind. Man verifiziert leicht, dass grundlegende Rechenregeln für den Erwartungswert (z.B. Linearität, $|E[Z]| \leq E[|Z|]$, Satz von Lebesgue) sich auf komplexwertige Zufallsvariablen übertragen.

Definition A.7 (Momentenerzeugende und charakteristische Funktion).

Die Funktionen $M : \mathbb{R}^d \rightarrow (0, \infty]$ bzw. $\phi : \mathbb{R}^d \rightarrow \mathbb{C}$,

$$M(t) := E[e^{t \cdot X}] = \int_{\mathbb{R}^d} e^{t \cdot x} \mu(dx),$$

$$\phi(t) := E[e^{it \cdot X}] = \int_{\mathbb{R}^d} e^{it \cdot x} \mu(dx),$$

heißen *momentenerzeugende* bzw. *charakteristische Funktion* der Zufallsvariable X oder der Verteilung μ .

Da die Funktionen $t \mapsto e^{t \cdot x}$ und $t \mapsto e^{it \cdot x}$ für $t \in \mathbb{R}^d$ nichtnegativ bzw. beschränkt sind, sind die Erwartungswerte definiert. Dabei nimmt $M(t)$ den Wert $+\infty$ an, falls $\exp(t \cdot X)$ nicht integrierbar ist. Für die Norm der komplexen Zahl $\phi(t)$ gilt dagegen

$$|\phi(t)| \leq E[|\exp(it \cdot x)|] = 1 \quad \text{für alle } t \in \mathbb{R}^d.$$

Bemerkung (Fourier- und Laplace-Transformation). Die Funktion $\phi(-t) = \int e^{-it \cdot x} \mu(dx)$ ist die *Fourier-Transformation* des Maßes μ . Ist μ absolutstetig bzgl. des Lebesguemaßes mit Dichte f , dann ist $\phi(-t)$ die Fourier-Transformation der Funktion f , d.h.

$$\phi(-t) = \widehat{f}(t) := \int_{\mathbb{R}^d} e^{-it \cdot x} f(x) dx.$$

Entsprechend ist

$$M(-t) = \int_{\mathbb{R}^d} e^{-t \cdot x} \mu(dx) \quad (t > 0)$$

die Laplace-Transformation des Maßes μ bzw. der Dichte f .

Rechenregeln. Die folgenden Rechenregeln ergeben sich unmittelbar aus den Definitionen der momentenerzeugenden bzw. charakteristischen Funktionen:

(i) Sind $X, Y : \Omega \rightarrow \mathbb{R}^d$ unabhängige Zufallsvariablen auf (Ω, \mathcal{A}, P) , dann gilt

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \quad \text{und} \quad \phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$$

für alle $t \in \mathbb{R}^d$.

(ii) Ist $X = (X_1, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$ ein Zufallsvektor mit unabhängigen Komponenten X_1, \dots, X_d , dann gilt für $t = (t_1, \dots, t_d) \in \mathbb{R}^d$:

$$M_X(t) = \prod_{i=1}^d M_{X_i}(t_i) \quad \text{und} \quad \phi_X(t) = \prod_{i=1}^d \phi_{X_i}(t_i).$$

(iii) Für $A \in \mathbb{R}^{d \times d}$ und $b \in \mathbb{R}^d$ gilt

$$M_{AX+b}(t) = e^{t \cdot b} M_X(A^T t) \quad \text{und} \quad \phi_{AX+b}(t) = e^{it \cdot b} \phi_X(A^T t).$$

(iv) Es gilt stets $M(0) = \phi(0) = 1$ und $\phi(-t) = \overline{\phi(t)}$ für alle $t \in \mathbb{R}$.

Beispiel (Binomialverteilung). Die Binomialverteilung $\text{Bin}(n, p)$ ist die Verteilung der Summe $\sum_{i=1}^n Y_i$ von unabhängigen Bernoulli(p)-verteilten Zufallsvariablen Y_1, \dots, Y_n . Also sind

$$\phi(t) = \prod_{i=1}^n \phi_{Y_i}(t) = (1 - p + pe^{it})^n, \quad \text{und} \quad M(t) = (1 - p + pe^t)^n$$

die charakteristische und momentenerzeugende Funktion von $\text{Bin}(n, p)$.

Der Übersichtlichkeit halber beschränken wir uns nun auf den Fall $d = 1$. Wir zeigen, dass sich die Momente $E[X^n]$ einer Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ unter geeigneten Voraussetzungen aus der momentenerzeugenden bzw. charakteristischen Funktion berechnen lassen. Die nötigen Voraussetzungen sind allerdings im Fall der momentenerzeugenden Funktion viel stärker.

Satz A.8 (Momentenerzeugung). (i) Ist $M(t) = E[e^{tX}]$ endlich auf $(-\delta, \delta)$ für ein $\delta > 0$, dann existiert der Erwartungswert $M(z) := E[e^{zX}]$ für alle $z \in \mathbb{C}$ mit $|\text{Re}(z)| < \delta$, und es gilt

$$E[e^{zX}] = \sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] \quad \text{für alle } z \in \mathbb{C} \text{ mit } |z| < \delta.$$

Insbesondere folgt

$$E[X^n] = M^{(n)}(0) \quad \text{für alle } n \in \mathbb{Z}_+.$$

(ii) Ist $E[|X|^n] < \infty$ für ein $n \in \mathbb{N}$, dann gilt $\phi \in C^n(\mathbb{R})$ und

$$\phi^{(n)}(t) = i^n \cdot E[X^n e^{itX}] \quad \text{für alle } t \in \mathbb{R}. \quad (\text{A.6})$$

Man beachte, dass die Voraussetzung im ersten Teil des Satzes erfüllt ist, falls $M(s) < \infty$ und $M(-s) < \infty$ für ein festes $s > 0$ gilt. Nach der Jensenschen Ungleichung folgt nämlich aus $M(s) < \infty$ auch

$$M(t) = E[e^{tX}] \leq E[e^{sX}]^{t/s} < \infty \quad \text{für alle } t \in [0, s].$$

Entsprechend folgt $M < \infty$ auf $[-s, 0]$ aus $M(-s) < \infty$.

Beweis. (i) Aus der Voraussetzung und dem Satz von der monotonen Konvergenz ergibt sich

$$\sum_{n=0}^{\infty} \frac{s^n}{n!} E[|X|^n] = E[e^{s|X|}] \leq E[e^{sX}] + E[e^{-sX}] < \infty \quad \text{für } s \in (0, \delta).$$

Insbesondere existieren alle Momente $E[X^n]$ ($n \in \mathbb{N}$), sowie die exponentiellen Momente $E[e^{zX}]$ für $z \in \mathbb{C}$ mit $|\operatorname{Re}(z)| < \delta$. Nach dem Satz von Lebesgue erhalten wir für $z \in \mathbb{C}$ mit $|z| < \delta$ zudem

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] = \lim_{m \rightarrow \infty} E \left[\sum_{n=0}^m \frac{(zX)^n}{n!} \right] = E \left[\lim_{m \rightarrow \infty} \sum_{n=0}^m \frac{(zX)^n}{n!} \right] = E[e^{zX}],$$

da $e^{s|X|}$ für $s \geq |z|$ eine Majorante der Partialsummen ist.

(ii) Wir zeigen die Behauptung durch Induktion nach n . Für $n = 0$ gilt (A.6) nach Definition von $\phi(t)$. Ist $E[|X|^{n+1}] < \infty$, dann folgt nach Induktionsvoraussetzung und mit dem Satz von Lebesgue:

$$\begin{aligned} \frac{\phi^{(n)}(t+h) - \phi^{(n)}(t)}{h} &= \frac{1}{h} E \left[(iX)^n \left(e^{i(t+h)X} - e^{itX} \right) \right] \\ &= E \left[(iX)^n \frac{1}{h} \int_t^{t+h} iX e^{isX} ds \right] \rightarrow E \left[(iX)^{n+1} e^{itX} \right] \end{aligned}$$

für $h \rightarrow 0$, also

$$\phi^{(n+1)}(t) = E[(iX)^{n+1} e^{itX}].$$

Die Stetigkeit von $\phi^{(n)}(t)$ folgt ebenfalls aus dem Satz von Lebesgue unter der Voraussetzung $E[|X|^n] < \infty$. ■

Beispiele. (i) Für eine Zufallsvariable X mit Dichtefunktion $f(x) \propto e^{-|x|^{1/2}}$ gilt $E[|X|^n] < \infty$ für alle $n \in \mathbb{N}$. Also ist die charakteristische Funktion beliebig oft differenzierbar. Die momentenerzeugende Funktion $M_X(t)$ ist hingegen nur für $t = 0$ endlich.

(ii) Ein Standardbeispiel einer Verteilung, deren Momente nicht existieren, ist die *Cauchy-Verteilung* mit Dichte

$$f(x) = \frac{1}{\pi(1+x^2)} \quad (x \in \mathbb{R}).$$

Für eine Cauchy-verteilte Zufallsvariable X gilt $M_X(t) = \infty$ für alle $t \neq 0$. Trotzdem existiert

$$\phi_X(t) = e^{-|t|} \quad \text{für alle } t \in \mathbb{R}.$$

Die charakteristische Funktion ist allerdings bei 0 nicht differenzierbar.

Bemerkung (Zusammenhang von M und ϕ). Gilt $M < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann hat die Funktion M eine eindeutige analytische Fortsetzung auf den Streifen $\{z \in \mathbb{C} : |\operatorname{Re}(z)| < \delta\}$ in der komplexen Zahlenebene, die durch $M(z) = E[\exp(zX)]$ gegeben ist. In diesem Fall gilt

$$\phi(t) = M(it) \quad \text{für alle } t \in \mathbb{R}.$$

Insbesondere ist die charakteristische Funktion dann durch die momentenerzeugende Funktion eindeutig bestimmt.

Die letzte Bemerkung ermöglicht manchmal eine vereinfachte Berechnung von charakteristischen Funktionen:

Beispiel (Normalverteilungen). (i) Für eine standardnormalverteilte Zufallsvariable Z gilt

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx-x^2/2} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2} < \infty \quad \text{für } t \in \mathbb{R}.$$

Die eindeutige analytische Fortsetzung auf \mathbb{C} ist die als Potenzreihe darstellbare Funktion $M_Z(z) = \exp(z^2/2)$. Also ist die charakteristische Funktion gegeben durch

$$\phi_Z(t) = M_Z(it) = e^{-t^2/2} \quad \text{für alle } t \in \mathbb{R}.$$

(ii) Eine normalverteilte Zufallsvariable X mit Mittel m und Varianz σ^2 können wir darstellen als $X = \sigma Z + m$ mit $Z \sim N(0, 1)$. Also gilt

$$\begin{aligned} M_X(t) &= e^{mt} M_Z(\sigma t) = \exp\left(mt + \sigma^2 t^2/2\right), \quad \text{und} \\ \phi_X(t) &= e^{imt} \phi_Z(\sigma t) = \exp\left(imt - \sigma^2 t^2/2\right). \end{aligned}$$

Bemerkung (Satz von Bochner). Eine Funktion $\phi : \mathbb{R} \rightarrow \mathbb{C}$ ist genau dann eine charakteristische Funktion einer Wahrscheinlichkeitsverteilung auf \mathbb{R} , wenn die folgenden Bedingungen erfüllt sind:

- (i) $\phi(0) = 1$ und $|\phi(t)| \leq 1$ für alle $t \in \mathbb{R}$.
- (ii) ϕ ist gleichmäßig stetig.
- (iii) ϕ ist *nicht-negativ definit*, d.h.

$$\sum_{i,j=1}^n \phi(t_i - t_j) z_i \bar{z}_j \geq 0 \quad \forall n \in \mathbb{N}, t_1, \dots, t_n \in \mathbb{R}, z_1, \dots, z_n \in \mathbb{C}.$$

Dass jede charakteristische Funktion einer Wahrscheinlichkeitsverteilung die Eigenschaften (i)-(iii) hat, prüft man leicht nach. Der Beweis der umgekehrten Aussage findet sich z.B. in Vol. II des Lehrbuchs von Feller [Feller2].

Anwendung auf multivariate Normalverteilungen

Multivariate Normalverteilungen haben wir bereits in Abschnitt A.2 eingeführt. Mithilfe von charakteristischen Funktionen können wir eine etwas allgemeinere Definition geben, die auch degenerierte Normalverteilungen (zum Beispiel Dirac-Maße) einschließt:

Definition A.9 (Normalverteilung im \mathbb{R}^d). Sei $m \in \mathbb{R}^d$, und sei $C \in \mathbb{R}^{d \times d}$ eine symmetrische, nicht-negativ definite Matrix. Die eindeutige Wahrscheinlichkeitsverteilung $N(m, C)$ im \mathbb{R}^d mit charakteristischer Funktion $\phi(t) = \exp\left(-\frac{1}{2}t \cdot Ct + it \cdot m\right)$ heißt **Normalverteilung mit Mittelwertvektor m und Kovarianzmatrix C** .

Die Existenz und Konstruktion einer Zufallsvariable mit Verteilung $N(m, C)$ ergibt sich aus der folgenden Bemerkung (iii).

Bemerkung (Charakterisierungen und Transformationen von Normalverteilungen). Die folgenden Aussagen beweist man mithilfe von charakteristischen Funktionen:

- (i) Ein Zufallsvektor $X : \Omega \rightarrow \mathbb{R}^d$ ist genau dann multivariat normalverteilt, wenn jede Linearkombination $\sum_{i=1}^d t_i X_i$ der Komponenten mit $t_1, \dots, t_d \in \mathbb{R}$ normalverteilt ist. Genauer ist $X \sim N(m, C)$ äquivalent zu

$$t \cdot X \sim N(t \cdot m, t \cdot C t) \quad \text{für alle } t \in \mathbb{R}^d.$$

- (ii) Ist $X \sim N(m, C)$, dann gilt

$$AX + b \sim N(Am + b, ACA^T) \quad \text{für alle } b \in \mathbb{R}^k \text{ und } A \in \mathbb{R}^{k \times d}, k \in \mathbb{N}.$$

- (iii) Sind Z_1, \dots, Z_d unabhängige, standardnormalverteilte Zufallsvariablen, und ist σ eine reelle $d \times d$ -Matrix mit $C = \sigma\sigma^T$, dann hat der Zufallsvektor $\sigma Z + m$ mit $Z = (Z_1, \dots, Z_d)^T$ die Verteilung $N(m, C)$.

- (iv) Im Fall $\det C \neq 0$ ist die Verteilung $N(m, C)$ absolutstetig bzgl. des d -dimensionalen Lebesgue-Maßes mit Dichte

$$f(y) = \frac{1}{\sqrt{(2\pi)^d |\det C|}} \exp\left(-\frac{1}{2}(y - m) \cdot C^{-1}(y - m)\right).$$

Beispiel (χ^2 -Verteilungen). Wir berechnen die Verteilung vom Quadrat des Abstandes vom Ursprung eines standardnormalverteilten Zufallsvektors im \mathbb{R}^d :

$$Z = (Z_1, \dots, Z_d) \sim N(0, I_d), \quad \|Z\|^2 = \sum_{i=1}^d Z_i^2.$$

Wegen $f_{|Z_i|}(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot I_{(0,\infty)}(x)$ folgt durch Anwenden des Dichtetransformationssatzes:

$$f_{Z_i^2}(y) = \sqrt{\frac{2}{\pi}} e^{-\frac{y}{2}} \cdot I_{(0,\infty)}(y) \cdot \frac{1}{2\sqrt{y}},$$

d.h. Z_i^2 ist $\Gamma(\frac{1}{2}, \frac{1}{2})$ -verteilt. Da die Zufallsvariablen Z_i^2 , $1 \leq i \leq d$, unabhängig sind, folgt:

$$\|Z\|^2 = \sum_{i=1}^d Z_i^2 \sim \Gamma\left(\frac{1}{2}, \frac{d}{2}\right).$$

Definition A.10 (χ^2 -Verteilung). Die Gamma-Verteilung mit Parametern $1/2$ und $d/2$ heißt auch **Chi-Quadrat-Verteilung $\chi^2(d)$ mit d Freiheitsgraden.**

Multivariater zentraler Grenzwertsatz

Auch im \mathbb{R}^d gilt ein zentraler Grenzwertsatz:

Satz A.11 (Multivariater zentraler Grenzwertsatz). Seien $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}^d$ unabhängige, identisch verteilte, quadratintegrierbare Zufallsvektoren auf (Ω, \mathcal{A}, P) , und sei $S_n = X_1 + \dots + X_n$. Dann gilt

$$\frac{S_n - E[S_n]}{\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0, C),$$

wobei $C_{jk} = \text{Cov}[X_{1,j}, X_{1,k}]$ die Kovarianzmatrix der Zufallsvektoren X_i ist.

Der Beweis basiert auf folgender Charakterisierung der Verteilungskonvergenz von Zufallsvektoren:

Lemma A.12 (Cramér-Wold Device). Für Zufallsvariablen $Y, Y_1, Y_2, \dots : \Omega \rightarrow \mathbb{R}^d$ gilt:

$$Y_n \xrightarrow{\mathcal{D}} Y \Leftrightarrow p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \quad \forall p \in \mathbb{R}^d.$$

Beweis. Der Beweis der Implikation „ \Rightarrow “ ist eine Übungsaufgabe. Umgekehrt gilt:

$$p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \Rightarrow E[\exp(ip \cdot Y_n)] \rightarrow E[\exp(ip \cdot Y)] \quad \forall p \in \mathbb{R}^d. \quad (\text{A.7})$$

Zudem ist unter dieser Voraussetzung für jedes $i \in \{1, \dots, d\}$ die Folge der Verteilungen der eindimensionalen Zufallsvariablen $Y_{n,i} := e_i \cdot Y_n$ ($n \in \mathbb{N}$) schwach konvergent. Daher existiert zu jedem $\varepsilon > 0$ ein $C \in (0, \infty)$, so dass $P[|Y_{n,i}| > C] \leq \varepsilon/d$ für alle n, i , und damit

$$P[|Y_n| \notin [-C, C]^d] \leq \sum_{i=1}^n P[|Y_{n,i}| > C] \leq \varepsilon \quad \text{für alle } n \in \mathbb{N}$$

gilt. Somit ist die Folge der Verteilungen μ_n der mehrdimensionalen Zufallsvariablen Y_n ($n \in \mathbb{N}$) eine straffe Folge von Wahrscheinlichkeitsmaßen im \mathbb{R}^d . Mit einem ähnlichen Beweis wie im eindimensionalen Fall zeigt man, dass der Satz von Prokhorov auch in diesem Fall gilt. Dabei verwendet man statt der eindimensionalen die multivariaten Verteilungsfunktionen

$$F_n(c_1, \dots, c_d) = P[Y_{n,1} \leq c_1, \dots, Y_{n,d} \leq c_d], \quad (c_1, \dots, c_d) \in \mathbb{R}^d.$$

Es folgt, dass jede Teilfolge von (μ_n) eine schwach konvergente Teilfolge hat, und mithilfe von (A.7) verifiziert man leicht, dass alle Grenzwerte von Teilfolgen mit der Verteilung μ von Y übereinstimmen. Also konvergiert μ_n schwach gegen μ , d.h. Y_n konvergiert in Verteilung gegen Y . ■

Wir beweisen nun den zentralen Grenzwertsatz im \mathbb{R}^d :

Beweis. Für $p \in \mathbb{R}^d$ gilt nach dem eindimensionalen zentralen Grenzwertsatz:

$$p \cdot \left(\frac{S_n - E[S_n]}{\sqrt{n}} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (p \cdot X_i - E[p \cdot X_i]) \xrightarrow{\mathcal{D}} N(0, \text{Var}[p \cdot X_1]) = N(0, p \cdot Cp),$$

da

$$\text{Var}[p \cdot X_1] = \text{Cov} \left[\sum_k p_k X_{1,k}, \sum_l p_l X_{1,l} \right] = \sum_{k,l} p_k p_l C_{kl} = p \cdot Cp.$$

Ist Y ein $N(0, C)$ -verteilter Zufallsvektor, dann ist $N(0, p \cdot Cp)$ die Verteilung von $p \cdot Y$. Mithilfe der Cramér-Wold Device folgt also, dass $(S_n - E[S_n]) / \sqrt{n}$ in Verteilung gegen Y konvergiert. ■