

Mehrdimensionale Zufallsvariablen

Im Folgenden Beschränkung auf den diskreten Fall und zweidimensionale Zufallsvariablen.

Vorstellung: Auswerten eines mehrdimensionalen Merkmals

$$\begin{pmatrix} \tilde{X} \\ \tilde{Y} \end{pmatrix}$$

also z.B. $\omega \in \tilde{\Omega}$, zufällig gezogene Person und damit $\tilde{X}(\omega)$ und $\tilde{Y}(\omega)$ Auswertung der Merkmale jeweils an *derselben* Person.

\Rightarrow zweidimensionale Zufallsvariable $\begin{pmatrix} \tilde{X} \\ \tilde{Y} \end{pmatrix}$ (wie bei Zusammenhangsanalyse in Statistik I)

Das Hauptinteresse gilt (entsprechend der Kontingenztafel in Statistik I) der gemeinsamen Verteilung

$$P(\{X = x_i\} \cap \{Y = y_j\})$$



Zweidimensionale Verteilungen

Betrachtet werden zwei eindimensionale diskrete Zufallselemente X und Y (zu demselben Zufallsexperiment). Die Wahrscheinlichkeit

$$P(X = x_i, Y = y_j) := P(\{X = x_i\} \cap \{Y = y_j\})$$

in Abhängigkeit von x_i und y_j heißt *gemeinsame Verteilung* der mehrdimensionalen Zufallsvariable $\begin{pmatrix} X \\ Y \end{pmatrix}$ bzw. der Variablen X und Y . Randwahrscheinlichkeiten:

$$p_{i\bullet} = P(X = x_i) = \sum_{j=1}^m P(X = x_i, Y = y_j)$$

$$p_{\bullet j} = P(Y = y_j) = \sum_{i=1}^k P(X = x_i, Y = y_j)$$



$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$
$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

Stetiger Fall: Zufallsvariable mit zweidimensionaler Dichtefunktion $f(x, y)$:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \left(\int_c^d f(x, y) dy \right) dx$$

Seien X und Y zwei Zufallsvariablen. Dann heißt

$$\sigma_{X,Y} := \text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

Kovarianz von X und Y .

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist

$$\text{Cov}(\tilde{X}, \tilde{Y}) = a_X \cdot a_Y \cdot \text{Cov}(X, Y)$$

- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$

Definition

Zwei Zufallsvariablen X und Y mit $\text{Cov}(X, Y) = 0$ heißen *unkorreliert*.

Satz

Stochastisch unabhängige Zufallsvariablen sind unkorreliert. Die Umkehrung gilt jedoch im allgemeinen nicht.
Vergleiche Statistik I: Kovarianz misst nur lineare Zusammenhänge.

der Korrelationskoeffizient

Definition

Gegeben seien zwei Zufallsvariablen X und Y . Dann heißt

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Korrelationskoeffizient von X und Y .



Eigenschaften des Korrelationskoeffizienten

- Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist

$$|\rho(\tilde{X}, \tilde{Y})| = |\rho(X, Y)|.$$

- $-1 \leq \rho(X, Y) \leq 1$.
- $|\rho(X, Y)| = 1 \iff Y = aX + b$
- Sind $\text{Var}(X) > 0$ und $\text{Var}(Y) > 0$, so gilt $\rho(X, Y) = 0$ genau dann, wenn $\text{Cov}(X, Y) = 0$.



Beispiel: Chuck a Luck

X_1 Gewinn, wenn beim ersten Wurf ein Einsatz auf 1 gesetzt wird.

X_6 Gewinn, wenn beim ersten Wurf ein Einsatz auf 6 gesetzt wird.

Kovarianz zwischen X_1 und X_6

(x_1, x_6)	$P(X_1 = x_1, X_6 = x_6)$	(x_1, x_6)	$P(X_1 = x_1, X_6 = x_6)$
$(-1, -1)$	$\frac{64}{216}$	$(-1, 3)$	$\frac{1}{216}$
$(-1, 1)$	$\frac{48}{216}$	$(3, -1)$	$\frac{1}{216}$
$(1, -1)$	$\frac{48}{216}$	$(1, 1)$	$\frac{24}{216}$
$(-1, 2)$	$\frac{12}{216}$	$(1, 2)$	$\frac{3}{216}$
$(2, -1)$	$\frac{12}{216}$	$(1, 2)$	$\frac{3}{216}$

$$\Rightarrow E(X_1 \cdot X_6) = -50/216 = -0.23148$$

$$\text{Cov}(X_1, X_6) = -0.23148 - (-0.0787) \cdot (-0.0787) = -0.23768$$

X_1 und X_6 sind negativ korreliert.



- 0 Einführung
- 1 Wahrscheinlichkeitsrechnung
- 2 Zufallsvariablen und ihre Verteilung
- 3 Statistische Inferenz**
- 4 Hypothesentests
- 5 Regression

Ziel: Inferenzschluss, Repräsentationsschluss: Schluss von einer Stichprobe auf Eigenschaften der Grundgesamtheit, aus der sie stammt.

- Von Interesse sei ein Merkmal \tilde{X} in der Grundgesamtheit $\tilde{\Omega}$.
- Ziehe eine Stichprobe $(\omega_1, \dots, \omega_n)$ von Elementen aus $\tilde{\Omega}$ und werte \tilde{X} jeweils aus.
- Man erhält Werte x_1, \dots, x_n . Diese sind Realisationen der i.i.d. Zufallsvariablen oder Zufallselemente X_1, \dots, X_n , wobei die Wahrscheinlichkeitsverteilung der X_1, \dots, X_n genau die Häufigkeitsverhältnisse in der Grundgesamtheit widerspiegelt.

Ziel: Schlüsse von Stichprobe auf Grundgesamtheit
Schlüsse von Experiment auf allgemeines Phänomen

Zentrale Fragen:

- Wie kann die Zufälligkeit in korrekter Weise berücksichtigt werden?
- Wann sind Ergebnisse in der Stichprobe zufallsbedingt?
- Wie sind korrekte Schlüsse möglich?

- 1 Schätzen:
Von Interesse ist der Wert eines Parameters in der Grundgesamtheit, z.B. Mittelwert oder Anteil
Punktschätzung: Angabe eines Wertes
Intervallschätzung (Konfidenzintervall): Angabe eines Bereiches, in dem der Wert mit hoher Sicherheit liegt
- 2 Testen (Signifikanztest):
Untersuchung, ob eine bestimmte Hypothese mit Hilfe der Daten widerlegt werden kann
z.B. Gewisse Satzkonstruktionen führen zu schnellerer Reaktion

Beispiele:

- Punktschätzung: z.B. wahrer Anteil 0.4751
- Intervallschätzung: z.B. wahrer Anteil liegt zwischen 0.46 und 0.48
- Hypothesentest: Die Annahme, der Anteil liegt höchstens bei 50% kann nicht aufrecht erhalten werden

- Stichprobe sollte zufällig sein
- Experimentelle Situation
- Nicht nötig (geeignet) bei Vollerhebungen
- Nicht geeignet bei Vollerhebungen mit geringem Rücklauf

Zentrale Fragestellung

Wie kommt man von Realisationen x_1, \dots, x_n von i.i.d. Zufallsvariablen X_1, \dots, X_n auf die Verteilung der X_i ?

- Dazu nimmt man häufig an, man kenne den Grundtyp der Verteilung der X_1, \dots, X_n . Unbekannt seien nur einzelne Parameter davon.
Beispiel: X_i sei normalverteilt, unbekannt seien nur μ, σ^2 .
 \implies *parametrische Verteilungsannahme* (meist im Folgenden)
- Alternativ: Verteilungstyp nicht oder nur schwach festgelegt (z.B. symmetrische Verteilung)
 \implies *nichtparametrische Modelle*
- Klarerweise gilt im Allgemeinen (generelles Problem bei der Modellierung): Parametrische Modelle liefern schärfere Aussagen – wenn ihre Annahmen zutreffen. Wenn ihre Annahmen nicht zutreffen, dann existiert die große Gefahr von Fehlschlüssen.



Beispiel:

Parameter: Mittelwert der täglichen Fernsehdauer von Jugendlichen in Deutschland

Schätzung: Mittelwert der Fernsehdauer in der Stichprobe
oder: Median aus der Stichprobe?
oder: Mittelwert ohne größten und kleinsten Wert?

Beispiel 1: Schätzer \bar{X}

Grundgesamtheit

1	2	3	4	5
1.30	1.31	1.32	1.40	1.42

Wahrer Wert: 1.35

Ziehe Stichprobe vom Umfang $n=2$ und berechne \bar{X}

S_1	S_2	\bar{X}	P
1	2	1.305	0.1
1	3	1.310	0.1
1	4	1.350	0.1
1	5	1.360	0.1
2	3	1.315	0.1
2	4	1.355	0.1
2	5	1.365	0.1
3	4	1.360	0.1
3	5	1.370	0.1
4	5	1.410	0.1

„Pech“

Beispiel 2: Würfeln mit potentiell gefälschtem Würfel

Wie groß ist der Erwartungswert beim Würfeln mit potentiell gefälschtem Würfel?

Ziehe Stichprobe und berechne Mittelwert \bar{X}

\bar{X} liefert plausible Schätzung für den wahren (theoretischen) Mittelwert.

Simulation mit R



Beachte: Auswahl zufällig \Rightarrow Schätzung zufällig

Die Merkmale der gezogenen n Einheiten sind also Zufallsgrößen.

Wir bezeichnen sie mit X_1, \dots, X_n .

Wird der Parameter einer Merkmalsverteilung durch eine Funktion der Zufallsgrößen X_1, \dots, X_n der Stichprobe geschätzt, so spricht man bei diesem Vorgang von **Punktschätzung**.

Die dabei benutzte Funktion wird auch **Schätzfunktion**, **Schätzstatistik** oder kurz **Schätzer** genannt.

Definition

Sei X_1, \dots, X_n i.i.d. Stichprobe. Eine Funktion

$$T = g(X_1, \dots, X_n)$$

heißt *Schätzer* oder *Schätzfunktion*.

Inhaltlich ist $g(\cdot)$ eine Auswertungsregel der Stichprobe: „Welche Werte sich auch in der Stichprobe ergeben, ich wene das durch $g(\cdot)$ beschriebene Verfahren auf sie an.(z.B. Mittelwert)“

Beispiele für Schätzfunktionen

- Arithmetisches Mittel der Stichprobe:

$$\bar{X} = g(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

Für binäre, dummy-kodierte X_i ist \bar{X} auch die relative Häufigkeit des Auftretens von „ $X_i = 1$ “ in der Stichprobe

- Stichprobenvarianz:

$$\tilde{S}^2 = g(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2$$

- Korrigierte Stichprobenvarianz:

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

Beispiele für Schätzfunktionen (2)

- Größter Stichprobenwert:

$$X_{(n)} = g(X_1, \dots, X_n) = \max_{i=1, \dots, n} X_i$$

- Kleinster Stichprobenwert:

$$X_{(1)} = g(X_1, \dots, X_n) = \min_{i=1, \dots, n} X_i$$



Qualitätsmerkmal eines Schätzers: Erwartungstreue

Erwartungstreue, Bias: Gegeben sei eine Stichprobe X_1, \dots, X_n und eine Schätzfunktion $T = g(X_1, \dots, X_n)$ (mit existierendem Erwartungswert).

- T heißt *erwartungstreu für den Parameter ϑ* , falls gilt

$$\mathbb{E}_{\vartheta}(T) = \vartheta$$

für alle ϑ .

- Die Größe

$$\text{Bias}_{\vartheta}(T) = \mathbb{E}_{\vartheta}(T) - \vartheta$$

heißt *Bias* (oder *Verzerrung*) der Schätzfunktion. Erwartungstreue Schätzfunktionen haben per Definition einen Bias von 0.

Man schreibt $\mathbb{E}_{\vartheta}(T)$ und $\text{Bias}_{\vartheta}(T)$, um deutlich zu machen, dass die Größen von dem wahren ϑ abhängen.

Bias und Erwartungstreue für \bar{X}

Das arithmetische Mittel $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ist erwartungstreu für den Mittelwert μ einer Grundgesamtheit: Aus X_1, \dots, X_n i.i.d. und $\mathbb{E}_\mu(X_1) = \mathbb{E}_\mu(X_2) = \dots = \mu$ folgt:

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}_\mu \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n} \mathbb{E}_\mu \left(\sum_{i=1}^n X_i \right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu = \mu\end{aligned}$$

Bias und Erwartungstreue bei einigen typischen Schätzfunktionen

Sei σ^2 die Varianz in der Grundgesamtheit. Es gilt

$$\mathbb{E}_{\sigma^2}(\tilde{S}^2) = \frac{n-1}{n}\sigma^2,$$

also ist \tilde{S}^2 *nicht* erwartungstreu für σ^2 .

$$\text{Bias}_{\sigma^2}(\tilde{S}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

(Für $n \rightarrow \infty$ geht $\text{Bias}_{\sigma^2}(\tilde{S}^2)$ gegen 0, \tilde{S}^2 ist „asymptotisch erwartungstreu“.)



Für die korrigierte Stichprobenvarianz gilt dagegen:

$$\begin{aligned}\mathbb{E}_{\sigma^2}(S^2) &= \mathbb{E}_{\sigma^2} \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\ &= \mathbb{E}_{\sigma^2} \left(\frac{1}{n-1} \cdot \frac{n}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\ &= \mathbb{E}_{\sigma^2} \left(\frac{n}{n-1} S^2 \right) = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2\end{aligned}$$

Also ist S^2 erwartungstreu für σ^2 . Diese Eigenschaft ist auch die Motivation für die Korrektur der Stichprobenvarianz.

Vorsicht: Im Allgemeinen gilt für beliebige, nichtlineare Funktionen g

$$\mathbb{E} g(X) \neq g(\mathbb{E}(X)).$$

Man kann also nicht einfach z.B. $\sqrt{\cdot}$ und \mathbb{E} vertauschen. In der Tat gilt: S^2 ist zwar erwartungstreu für σ^2 , aber $\sqrt{S^2}$ ist nicht erwartungstreu für $\sqrt{\sigma^2} = \sigma$.

Eigenschaften des Schätzers \bar{X} für den unbekanntem Erwartungswert μ

- Plausibles Vorgehen: Schätze den unbekanntem Mittelwert der Grundgesamtheit durch den Mittelwert aus der Stichprobe
- \bar{X} ist erwartungstreuer Schätzer
- \bar{X} hat die Standardabweichung σ/\sqrt{n}
Diese wird häufig als Standardfehler bezeichnet
- „Um die Schätzgenauigkeit zu verdoppeln, ist eine Erhöhung des Stichprobenumfangs n um den Faktor 4 nötig“



Gegeben sei eine Stichprobe der wahlberechtigten Bundesbürger. Geben Sie einen erwartungstreuen Schätzer des Anteils der rot-grün Wähler an. Grundgesamtheit: Dichotomes Merkmal

$$\tilde{X} = \begin{cases} 1 & \text{rot/grün: ja} \\ 0 & \text{rot/grün: nein} \end{cases}$$

Der Mittelwert π von \tilde{X} ist der Anteil der rot/grün-Wähler in der Grundgesamtheit.

Stichprobe X_1, \dots, X_n vom Umfang n :

$$X_i = \begin{cases} 1 & i\text{-te Person wählt rot/grün} \\ 0 & \text{sonst} \end{cases}$$

Aus den Überlegungen zum arithmetischen Mittel folgt, dass

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

ein erwartungstreuer Schätzer für den hier betrachteten Parameter π ist. Also verwendet man die relative Häufigkeit in der Stichprobe, um den wahren Anteil π in der Grundgesamtheit zu schätzen.



Bedeutung der Erwartungstreue:

Erwartungstreue ist ein schwaches Kriterium!
Betrachte die offensichtlich unsinnige Schätzfunktion

$$T_2 = g_2(X_1, \dots, X_n) = X_1,$$

d.h. $T_2 = 100\%$, falls der erste Befragte rot-grün wählt und $T_2 = 0\%$ sonst.

Die Schätzfunktion ignoriert fast alle Daten, ist aber erwartungstreu:

$$\mathbb{E}(T_2) = \mathbb{E}(X_1) = \mu$$

Deshalb betrachtet man zusätzlich die Effizienz eines Schätzers



Beispiel Wahlumfrage: Gegeben sind zwei erwartungstreue Schätzer (n sei gerade):

$$T_1 = \frac{1}{n} \sum_{i=1}^n X_i$$

$$T_2 = \frac{1}{n/2} \sum_{i=1}^{n/2} X_i$$

Was unterscheidet formal T_1 von dem unsinnigen Schätzer T_2 , der die in der Stichprobe enthaltene Information nicht vollständig ausnutzt? Vergleiche die Schätzer über ihre Varianz, nicht nur über den Erwartungswert!

Wenn n so groß ist, dass der zentrale Grenzwertsatz angewendet werden kann, dann gilt approximativ

$$\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n (X_i - \pi)}{\sqrt{\pi(1-\pi)}} = \frac{\sum_{i=1}^n X_i - n \cdot \pi}{\sqrt{n} \sqrt{\pi(1-\pi)}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0; 1)$$

und damit

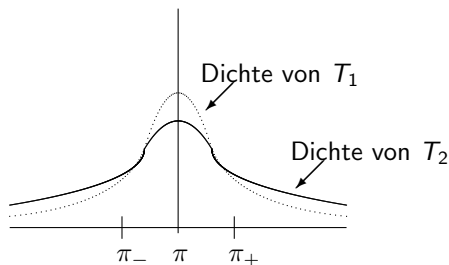
$$T_1 = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\pi; \frac{\pi(1-\pi)}{n}\right).$$

Analog kann man zeigen:

$$T_2 = \frac{1}{n/2} \sum_{i=1}^{n/2} X_i \sim N\left(\pi, \frac{\pi(1-\pi)}{n/2}\right).$$

T_1 und T_2 sind approximativ normalverteilt, wobei T_1 eine deutlich kleinere Varianz als T_2 hat.

T_1 und T_2 treffen beide im Durchschnitt den richtigen Wert π . T_1 schwankt aber weniger um das wahre π , ist also „im Durchschnitt genauer“.



Für jeden Punkt $\pi_+ > \pi$ ist damit $P(T_1 > \pi_+) < P(T_2 > \pi_+)$
und für jeden Punkt $\pi_- < \pi$ ist $P(T_1 < \pi_-) < P(T_2 < \pi_-)$.

Es ist also die Wahrscheinlichkeit, mindestens um $\pi_+ - \pi$ bzw. $\pi - \pi_-$ daneben zu liegen, bei T_2 stets größer als bei T_1 . Umgekehrt gesagt: Ein konkreter Wert ist damit verlässlicher, wenn er von T_1 , als wenn er von T_2 stammt.

Diese Überlegung gilt ganz allgemein: Ein erwartungstreuer Schätzer ist umso besser, je kleiner seine Varianz ist.

$$\text{Var}(T) = \text{Erwartete quadratische Abweichung von } T \text{ von } \underbrace{\mathbb{E}(T)}_{=\vartheta!}$$

Je kleiner die Varianz, umso mehr konzentriert sich die Verteilung eines erwartungstreuen Schätzers um den wahren Wert. Dies ist umso wichtiger, da der Schätzer den wahren Wert i.A. nur selten exakt trifft.

- Gegeben seien zwei erwartungstreue Schätzfunktionen T_1 und T_2 für einen Parameter ϑ . Gilt

$$\text{Var}_{\vartheta}(T_1) \leq \text{Var}_{\vartheta}(T_2) \text{ für alle } \vartheta$$

und

$$\text{Var}_{\vartheta^*}(T_1) < \text{Var}_{\vartheta^*}(T_2) \text{ für mindestens ein } \vartheta^*$$

so heißt T_1 *effizienter als* T_2 .

- Eine für ϑ erwartungstreue Schätzfunktion T heißt *UMVU-Schätzfunktion* für ϑ (*uniformly minimum variance unbiased*), falls

$$\text{Var}_{\vartheta}(T) \leq \text{Var}_{\vartheta}(T^*)$$

für alle ϑ und für alle erwartungstreuen Schätzfunktionen T^* .

- *Inhaltliche Bemerkung:* Der (tiefere) Sinn von Optimalitätskriterien wird klassischerweise insbesondere auch in der *Gewährleistung von Objektivität* gesehen.
- Ist X_1, \dots, X_n eine i.i.d. Stichprobe mit $X_i \sim N(\mu, \sigma^2)$, dann ist
 - \bar{X} UMVU-Schätzfunktion für μ und
 - S^2 UMVU-Schätzfunktion für σ^2 .

- Ist X_1, \dots, X_n mit $X_i \in \{0, 1\}$ eine i.i.d. Stichprobe mit $\pi = P(X_i = 1)$, dann ist die relative Häufigkeit \bar{X} UMVU-Schätzfunktion für π .
- Bei nicht erwartungstreuen Schätzern macht es keinen Sinn, sich ausschließlich auf die Varianz zu konzentrieren.
- Z.B. hat der unsinnige Schätzer $T = g(X_1, \dots, X_n) = 42$, der die Stichprobe nicht beachtet, Varianz 0.

- Man zieht dann den sogenannten *Mean Squared Error*

$$\text{MSE}_{\vartheta}(T) = \mathbb{E}_{\vartheta}(T - \vartheta)^2$$

zur Beurteilung heran. Es gilt

$$\text{MSE}_{\vartheta}(T) = \text{Var}_{\vartheta}(T) + (\text{Bias}_{\vartheta}(T))^2.$$

Der MSE kann als Kompromiss zwischen zwei Auffassungen von Präzision gesehen werden: möglichst geringe systematische Verzerrung (Bias) und möglichst geringe Schwankung (Varianz).

Asymptotische Erwartungstreue

- * Eine Schätzfunktion heißt asymptotisch erwartungstreu, falls

$$\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$$

bzw.

$$\lim_{n \rightarrow \infty} \text{Bias}(\hat{\theta}) = 0$$

gelten.

- * Abschwächung des Begriffs der Erwartungstreue: Gilt nur noch bei einer unendlich großen Stichprobe.
- * Erwartungstreue Schätzer sind auch asymptotisch erwartungstreu.
- * Sowohl S^2 als auch \tilde{S}^2 sind asymptotisch erwartungstreu.



- Für komplexere Modelle ist oft die Erwartungstreue der Verfahren ein zu restriktives Kriterium. Man fordert deshalb oft nur, dass sich der Schätzer wenigstens für große Stichproben gut verhält. Hierzu gibt es v.a. zwei verwandte aber „etwas“ unterschiedliche Kriterien.
- Ein Schätzer heißt (MSE-)konsistent oder konsistent im quadratischen Mittel, wenn gilt

$$\lim_{n \rightarrow \infty} (\text{MSE}(T)) = 0.$$

Der MSE von \bar{X} ist gegeben durch

$$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) + \text{Bias}^2(\bar{X}) = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n} \rightarrow 0.$$

\bar{X} ist also ein MSE-konsistente Schätzer für den Erwartungswert. Anschaulich bedeutet die Konsistenz, dass sich die Verteilung des Schätzers für wachsenden Stichprobenumfang n immer stärker beim richtigen Wert „zusammenzieht“. Er trifft also für unendlich große Stichproben praktisch sicher den wahren Wert. (Dies gilt als eine Minimalanforderung an statistische Verfahren.)

Maximum Likelihood

Das Maximum-Likelihood-Prinzip Sie wissen als Wirt, dass heute die Lokalparteien ihre Busausflüge unternehmen: Es werden Busse mit je 100 Personen von der jeweiliger Partei organisiert. Bus I: 85% Partei A, 15% Partei B

Bus II: 15% Partei A, 85% Partei B

Bus fährt vor, anhand Stichprobe ermitteln, ob Bild von ... von der Wand genommen werden soll oder nicht.

Stichprobe von 10 Personen ergibt 80% Anhänger der Partei A, welche Partei: wohl A, aber B nicht ausgeschlossen bei unglücklicher Auswahl.

Warum: A ist plausibler, da die Wahrscheinlichkeit, ungefähr den in der Stichprobe beobachteten Wert zu erhalten (bzw. erhalten zu haben) bei Bus I wesentlich größer ist als bei Bus II.



Aufgabe: Schätze den Parameter ϑ eines parametrischen Modells anhand einer i.i.d. Stichprobe X_1, \dots, X_n mit der konkreten Realisation x_1, \dots, x_n .
Idee der Maximum-Likelihood (ML) Schätzung für diskrete Verteilungen:

- Man kann für jedes ϑ die Wahrscheinlichkeit ausrechnen, genau die Stichprobe x_1, \dots, x_n zu erhalten:

$$P_{\vartheta}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P_{\vartheta}(X_i = x_i)$$

- Je größer für ein gegebenes ϑ_0 die Wahrscheinlichkeit ist, die konkrete Stichprobe erhalten zu haben, umso plausibler ist es, dass tatsächlich ϑ_0 der wahre Wert ist (gute Übereinstimmung zwischen Modell und Daten).

Beispiel

I.i.d. Stichprobe vom Umfang $n = 5$ aus einer $B(10, \pi)$ -Verteilung:

6 5 3 4 4

Wahrscheinlichkeit der Stichprobe für gegebenes π :

$$\begin{aligned}P(X_1 = 6, \dots, X_5 = 4 \mid \pi) &= P(X_1 = 6 \mid \pi) \cdot \dots \cdot P(X_5 = 4 \mid \pi) \\ &= \binom{10}{6} \pi^6 (1 - \pi)^4 \cdot \dots \cdot \binom{10}{4} \pi^4 (1 - \pi)^6.\end{aligned}$$

„ $P(\dots \mid \pi)$ Wahrscheinlichkeit, wenn π der wahre Parameter ist“



Wahrscheinlichkeit für einige Werte von π :

π	$P(X_1 = 6, \dots, X_5 = 4 \pi)$
0.1	0.00000000000001
0.2	0.0000000227200
0.3	0.0000040425220
0.4	0.0003025481000
0.5	0.0002487367000
0.6	0.0000026561150
0.7	0.0000000250490
0.8	0.00000000000055
0.9	0.00000000000000

Man nennt daher $L(\vartheta) = P_{\vartheta}(X_1 = x_1, \dots, X_n = x_n)$, nun als Funktion von ϑ gesehen, die *Likelihood* (deutsch: Plausibilität, Mutmaßlichkeit) von ϑ gegeben die Realisation x_1, \dots, x_n . Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \dots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \dots, X_n)$ *Maximum-Likelihood-Schätzer*

$P_{\vartheta}(X_1 = x_1, \dots, X_n = x_n) :$

- Deduktiv (Wahrscheinlichkeitsrechnung): ϑ bekannt, x_1, \dots, x_n zufällig („unbekannt“).
- Induktiv (Statistik): ϑ unbekannt, x_1, \dots, x_n bekannt.

Deduktiv

Induktiv

geg: Parameter bekannt

ges: Plausibilität des Parameters

$$P_{\vartheta}(X_1 = x_1, \dots, X_n = x_n)$$

Funktion von x_1, \dots, x_n
bei festem ϑ

$$P_{\vartheta}(X_1 = x_1, \dots, X_n = x_n)$$

Funktion von ϑ
bei festem x_1, \dots, x_n

geg: Wskt von Beobachtungen ges: Beobachtung bekannt

Definition ML

Gegeben sei die Realisation x_1, \dots, x_n einer i.i.d. Stichprobe. Die Funktion in ϑ

$$L(\vartheta) = \begin{cases} \prod_{i=1}^n P_{\vartheta}(X_i = x_i) & \text{falls } X_i \text{ diskret} \\ \prod_{i=1}^n f_{\vartheta}(x_i) & \text{falls } X_i \text{ stetig.} \end{cases}$$

heißt *Likelihood* des Parameters ϑ bei der Beobachtung x_1, \dots, x_n . Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \dots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \dots, X_n)$ *Maximum-Likelihood-Schätzer*.



- In diesem Fall verwendet man die Dichte

$$f_{\vartheta}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\vartheta}(x_i)$$

als Maß für die Plausibilität von ϑ .

- Für die praktische Berechnung maximiert man statt der Likelihood typischerweise die Log-Likelihood

$$l(\vartheta) = \ln(L(\vartheta)) = \ln \prod_{i=1}^n P_{\vartheta}(X_i = x_i) = \sum_{i=1}^n \ln P_{\vartheta}(X_i = x_i)$$

bzw.

$$l(\vartheta) = \ln \prod_{i=1}^n f_{\vartheta}(x_i) = \sum_{i=1}^n \ln f_{\vartheta}(x_i).$$

$$X_i = \begin{cases} 1 & \text{falls Rot/Grün} \\ 0 & \text{sonst} \end{cases}$$

Verteilung der X_i : Binomialverteilung $B(1, \pi)$ (Bernoulliverteilung)

$$P(X_i = 1) = \pi$$

$$P(X_i = 0) = 1 - \pi$$

$$P(X_i = x_i) = \pi^{x_i} \cdot (1 - \pi)^{1-x_i}, \quad x_i \in \{0; 1\}.$$

Hier ist π der unbekannte Parameter, der allgemein mit ϑ bezeichnet wird.

Likelihood:

$$\begin{aligned}L(\pi) &= P(X_1 = x_1, \dots, X_n = x_n) \\&= \prod_{i=1}^n \pi^{x_i} (1 - \pi)^{1-x_i} \\&= \pi^{\sum_{i=1}^n x_i} \cdot (1 - \pi)^{n - \sum_{i=1}^n x_i}\end{aligned}$$



Logarithmierte Likelihood:

$$l(\pi) = \ln(P(X_1 = x_1, \dots, X_n = x_n)) = \sum_{i=1}^n x_i \cdot \ln(\pi) + (n - \sum_{i=1}^n x_i) \cdot \ln(1 - \pi)$$

Ableiten (nach π):

$$\frac{\partial}{\partial \pi} l(\pi) = \frac{\sum_{i=1}^n x_i}{\pi} + \frac{n - \sum_{i=1}^n x_i}{1 - \pi} \cdot (-1)$$



Dies liefert denselben Schätzwert $\hat{\vartheta}$ und erspart beim Differenzieren die Anwendung der Produktregel.

Der Logarithmus ist streng monoton wachsend. Allgemein gilt für streng monoton wachsende Funktionen g : x_0 Stelle des Maximums von $L(x)$
 $\iff x_0$ auch Stelle des Maximums von $g(L(x))$.

Nullsetzen und nach π auflösen ergibt:

$$\begin{aligned}\frac{\partial}{\partial \pi} l(\pi) = 0 &\iff \frac{\sum_{i=1}^n x_i}{\pi} = \frac{n - \sum_{i=1}^n x_i}{1 - \pi} \\ &\iff (1 - \pi) \sum_{i=1}^n x_i = n \cdot \pi - \pi \sum_{i=1}^n x_i\end{aligned}$$

$$\iff \sum_{i=1}^n x_i = n \cdot \pi$$

also

$$\hat{\pi} = \frac{\sum_{i=1}^n x_i}{n}$$

Also ist \bar{X} der Maximum-Likelihood-Schätzer für π .



- Bestimme die Likelihoodfunktion

$$\begin{aligned}L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}(\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{2\pi^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right)\end{aligned}$$

- Bestimme die Log-Likelihoodfunktion

$$\begin{aligned}l(\mu, \sigma^2) &= \ln(L(\mu, \sigma^2)) \\ &= \ln(1) - \frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\end{aligned}$$

- Ableiten und Nullsetzen der Loglikelihoodfunktion

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{2\sigma^2} \cdot \sum_{i=1}^n (x_i - \mu) = 0$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

ML-Schätzung bei Normalverteilung

- Auflösen der beiden Gleichungen nach μ und σ^2
Aus der ersten Gleichung erhalten wir

$$\sum_{i=1}^n x_i - n\mu = 0 \quad \text{also} \quad \hat{\mu} = \bar{x}.$$

Aus der zweiten Gleichung erhalten wir durch Einsetzen von $\hat{\mu} = \bar{x}$

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

also

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



- Der ML-Schätzer $\hat{\mu} = \bar{X}$ für μ stimmt mit dem üblichen Schätzer für den Erwartungswert überein.
- Der ML-Schätzer $\hat{\sigma}^2 = \tilde{S}^2$ für σ^2 ist verzerrt, d.h. nicht erwartungstreu.

Einige allgemeine Eigenschaften von ML-Schätzern

- ML-Schätzer $\hat{\theta}$ sind im Allgemeinen nicht erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind asymptotisch erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind konsistent (und meist in einem asymptotischen Sinne effizient).

