Grundprinzipien der induktiven Statistik

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

Statistische Inferenz

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?

Inferenzprinzipien

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 zuschnellerer Reaktion

Beispiele:

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

Voraussetzungen für das Anwenden statistischer Inferenz

- 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 $x_1, \ldots, x_n$ von i.i.d. Zufallsvariablen $X_1, \ldots, X_n$ auf die Verteilung der X_i?

- Dazu nimmt man häufig an, man kenne den Grundtyp der Verteilung der $X_1, \ldots, X_n$. Unbekannt seien nur einzelne Parameter davon.

  Beispiel: X_i sei normalverteilt, unbekannt seien nur μ, σ^2.

  $\Longrightarrow$ *parametrische Verteilungsannahme* (meist im Folgenden)
- Alternativ: Verteilungstyp nicht oder nur schwach festgelegt (z.B. symmetrische Verteilung)

  $\Longrightarrow$ *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.

Punktschätzung

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

Punktschätzung

Beachte: Auswahl zufällig $\Rightarrow$ Schätzung zufällig

- Die Merkmale der gezogenen n Einheiten sind also Zufallsgrößen.
- Bezeichnung: $X_1, \ldots, X_n$.
- Wird der Parameter einer Merkmalsverteilung durch eine Funktion der Zufallsgrößen $X_1, \ldots, X_n$ der Stichprobe geschätzt, so spricht man bei diesem Vorgang von **Punktschätzung**.
- Die dabei benutzte Funktion wird auch **Schätzfunktion**, **Schätzstatistik** oder kurz **Schätzer** genannt.

Schätzfunktionen

Definition

Sei $X_1, \ldots, X_n$ i.i.d. Stichprobe. Eine Funktion

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

heißt Schätzer oder Schätzfunktion.

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

Beispiele für Schätzfunktionen I

- Arithmetisches Mittel der Stichprobe:

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

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

- Stichprobenvarianz:

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

- Korrigierte Stichprobenvarianz:

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

Beispiele für Schätzfunktionen II

- Größter Stichprobenwert:

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

- Kleinster Stichprobenwert:

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

Qualitätsmerkmal eines Schätzers

Erwartungstreue, Bias:
Gegeben sei eine Stichprobe $X_1, \ldots, X_n$ und eine Schätzfunktion $T = g(X_1, \ldots, X_n)$ (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 $X_1, \ldots, X_n$ i.i.d. und $\mathbb{E}_\mu(X_1) = \mathbb{E}_\mu(X_2) = \ldots = \mu$ 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 für $\tilde{S}^2$

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

Bias und Erwartungstreue für S^2

Für die korrigierte Stichprobenvarianz hingegen gilt:

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

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

Nichtlineare Funktionen

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 $\sqrt{\sigma^2} = \sigma$.

Wahlumfrage

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 $X_1, \ldots, X_n$ vom Umfang n:

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

Anteil als erwartungstreuer Schätzer

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

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

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 erwartungtreu:

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

Deshalb betrachtet man zusätzlich die Effizienz eines Schätzers

Effizienz I

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!

Effizienz II

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

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

und damit

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

Effizienz II

Analog kann man zeigen:

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

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“.

Effizienz IV

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.

Beste Schätzer

- 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 (*u*niformly *m*inimum *v*ariance *u*nbiased), falls

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

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

UMVU-Schätzer

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

Verzerrte Schätzer

- Ist $X_1, \ldots, X_n$ 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 $T = g(X_1, \ldots, X_n) = 42$, der die Stichprobe nicht beachtet, Varianz 0.

MSE

Man zieht dann den sogenannten *Mean Squared Error*

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

zur Beurteilung heran. Es gilt

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

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

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

bzw.

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

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.

Konsistenz

- 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

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

Konsistenz von $\bar{X}$

Der MSE von $\bar{X}$ ist gegeben durch

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

$\bar{X}$ ist also ein MSE-konsistente Schäter 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 praktische sicher den wahren Wert. (Dies gilt als eine Minimalanforderung an statistische Verfahren.)

Maximum–Likelihood–Prinzip I

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.

Maximum–Likelihood–Prinzip II

Aufgabe: Schätze den Parameter ϑ eines parametrischen Modells anhand einer i.i.d. Stichprobe $X_1, \ldots, X_n$ mit der konkreten Realisation $x_1, \ldots, x_n$.

Idee der Maximium-Likelihood (ML) Schätzung für diskrete Verteilungen:

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

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

- 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).

Maximum–Likelihood–Prinzip: Beispiel

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

$$6 \quad 5 \quad 3 \quad 4 \quad 4$$

Wahrscheinlichkeit der Stichprobe für gegebenes π:

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

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

Wahrscheinlichkeit für einige Werte von π:

π	$P(X_1 = 6, \ldots, X_5 = 4 \mid \pi)$
0.1	0.0000000000001
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	0.0000000000055
0.9	0.0000000000000

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

$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n)$:

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

Deduktiv

geg: Parameter bekannt

$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n)$

Funktion von $x_1, \ldots, x_n$
bei festem ϑ

ges: Wskt von Beobachtungen

Induktiv

ges: Plausibilität des Parameters

$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n)$

Funktion von ϑ
bei festem $x_1, \ldots, x_n$

geg: Beobachtung bekannt

Definition Maximum Likelihood

Gegeben sei die Realisation $x_1, \ldots, x_n$ 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 $x_1, \ldots, x_n$.

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

Likelihood bei stetige Verteilungen

- In diesem Fall verwendet man die Dichte

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

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.

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

ML Schätzung für π einer Bernoulliverteilung I

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

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

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

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

ML Schätzung für π einer Bernoulliverteilung I

- Bestimme die Likelihoodfunktion

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

ML Schätzung für π einer Bernoulliverteilung II

- Berechne die logarithmierte Likelihoodfunktion

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

- Ableiten (nach π):

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

ML Schätzung für π einer Bernoulliverteilung III

- Bemerkung zur Loglikelihood

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

ML Schätzung für π einer Bernoulliverteilung IV

- Berechnung des ML-Schätzers durch Nullsetzen der abgeleiteten Loglikelihoodfunktion

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

also

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

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

ML-Schätzung bei Normalverteilung I

- 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

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

ML-Schätzung bei Normalverteilung II

- Ableiten und Nullsetzen der Loglikelihoodfunktion

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

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

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

Fazit

- 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).