1.8 Mehrdimensionale Zufallsvariablen

Im Folgenden Beschränkung auf den diskreten Fall und zweidimensionale Zufallsvariablen.
„Schnelldurchgang unter Bezug auf das Eindimensionale und Statistik I“

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

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

Definition 1.79.

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 $\binom{X}{Y}$ bzw. der Variablen X und Y.

Randwahrscheinlichkeiten:

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

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

Bedingte Verteilungen:

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

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

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

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

Definition 1.80.

Seien X und Y zwei Zufallsvariablen. Dann heißt

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

Kovarianz von X und Y.

Rechenregeln:

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

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

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

Definition 1.81.

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

Satz 1.82.

Stochastisch unabhängige Zufallsvariablen sind unkorreliert. Die Umkehrung gilt jedoch im allgemeinen nicht.

Definition 1.83.

Gegeben seien zwei Zufallsvariablen X und Y. Dann heißt

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

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

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

- $-1 \leq \rho(X, Y) \leq 1$.

- $|\rho(X, Y)| = 1 \Longleftrightarrow Y = aX + b$

- Sind $\mathrm{Var}(X) > 0$ und $\mathrm{Var}(Y) > 0$, so gilt $\rho(X, Y) = 0$ genau dann, wenn $\mathrm{Cov}(X, Y) = 0$.

Bsp. 1.84. [Chuckk-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)	$P(X_1 = x_1, X_6 = x_6)$
$(-1,-1)$	$\frac{64}{216}$
$(-1,1)$	$\frac{48}{216}$
$(1,-1)$	$\frac{48}{216}$
$(-1,2)$	$\frac{12}{216}$
$(2,-1)$	$\frac{12}{216}$
$(-1,3)$	$\frac{1}{216}$
$(3,-1)$	$\frac{1}{216}$
$(1,1)$	$\frac{24}{216}$
$(1,2)$	$\frac{3}{216}$
$(1,2)$	$\frac{3}{216}$

$$\begin{aligned} \Rightarrow \quad E(X_1 \cdot X_6) &= -50/216 = -0.23148 \\ \mathrm{Cov}(X_1, X_6) &= -0.23148 - (-0.0787) \cdot (-0.0787) = -0.23768 \end{aligned}$$

X_1 und X_6 sind negativ korreliert.

2 Induktive Statistik

2.1 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.

Die Frage lautet also: 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 (vgl. Kap. 1.6).

  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.

Wichtige Fragestellungen der induktiven Statistik:

2.2 Punktschätzung

Ziel: Finde einen möglichst guten Schätzwert für eine bestimmte Kenngröße ϑ (Parameter) der Grundgesamtheit, z.B. den wahren Anteil der rot/grün-Wähler, den wahren Mittelwert, die wahre Varianz, aber auch z.B. das wahre Maximum (z.B. von Windgeschwindigkeit).

2.2.1 Schätzfunktionen

Gegeben sei die in Kapitel 2.1 beschriebene Situation, also eine i.i.d. Stichprobe $X_1, \ldots, X_n$ eines Merkmales $\tilde{X}$.

Definition 2.1.

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. ich bilde Mittelwert)“

Typische Beispiele für Schätzfunktionen:

1. Arithmetisches Mittel der Stichprobe:

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

   Für binäre X_i mit $X_i \in \{0, 1\}$ ist $\bar{X}$ auch die relative Häufigkeit des Auftretens von „$X_i = 1$“ in der Stichprobe

2. Stichprobenvarianz:

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

3. Korrigierte Stichprobenvarianz:

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

4. Größter Stichprobenwert:

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

5. Kleinster Stichprobenwert:

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

Schätzfunktion und Schätzwert: Da $X_1, \ldots, X_n$ zufällig sind, ist auch die Schätzfunktion $T = g(X_1, \ldots, X_n)$ *zufällig.* Zieht man mehrere Stichproben, so erhält man jeweils andere Realisationen von $X_1, \ldots, X_n$, und damit auch von T.

Die Realisation t (konkreter Wert) der Zufallsvariable T (Variable) heißt $Schätzwert$.

Man hat in der Praxis meist nur eine konkrete Stichprobe und damit auch nur einen konkreten Wert t von T. Zur Beurteilung der mathematischen Eigenschaften werden aber alle denkbaren Stichproben und die zugehörigen Realisationen der Schätzfunktion T herangezogen.

D.h. beurteilt wird $nicht$ der einzelne Schätzwert als solcher, sondern die Schätzfunktion, als $Methode$, d.h. als $Regel$ zur Berechnung des Schätzwerts aus der Stichprobe.

Andere Notation in der Literatur: $\hat{\vartheta}$ Schätzer für ϑ.

Dabei wird nicht mehr direkt unterschieden zwischen Zufallsvariable (bei uns Großbuchstaben) und Realisation (bei uns klein). $\Longrightarrow$ Schreibe $\hat{\vartheta}(X_1, \ldots, X_n)$ bzw. $\hat{\vartheta}(x_1, \ldots, x_n)$ wenn die Unterscheidung benötigt wird.

Bsp. 2.2.

Durchschnittliche Anzahl der Statistikbücher in einer Grundgesamtheit von Studierenden schätzen.

- Grundgesamtheit: Drei Personen $\tilde{\Omega} = \{\tilde{\omega_1}, \tilde{\omega_2}, \tilde{\omega_3}\}$.
- Merkmal $\tilde{X}$: Anzahl der Statistikbücher

$$\tilde{X}(\tilde{\omega_1}) = 3 \quad \tilde{X}(\tilde{\omega_2}) = 1 \quad \tilde{X}(\tilde{\omega_3}) = 2.$$

  Wahrer Durchschnittswert: $\mu = 2$.

- Stichprobe X_1, X_2 ohne Zurücklegen (Stichprobenumfang $n = 2$):

$$X_1 = \tilde{X}(\omega_1) \qquad X_2 = \tilde{X}(\omega_2)$$

  wobei

ω_1 erste gezogene Person, ω_2 zweite gezogene Person.

Betrachte folgende möglichen Schätzer:

$$
\begin{aligned}
T_1 &= g_1(X_1, X_2) = \bar{X} = \frac{X_1 + X_2}{2} \\
T_2 &= X_1 \\
T_3 &= g(X_1, X_2) = \frac{1}{2} X_{(2)} = \frac{2}{3} \max(X_1, X_2)
\end{aligned}
$$

2.2.2 Gütekriterien

Beurteile die Schätzfunktionen, also das Verfahren $an\ sich$, nicht den einzelnen Schätzwert. Besonders bei komplexeren Schätzproblemen sind klar festgelegte Güteeigenschaften wichtig.

Natürlich ist auch zu Beginn genau festzulegen, was geschätzt werden soll. Im Folgenden sei der Parameter ϑ stets eine eindimensionale Kenngröße der Grundgesamtheit (z.B. Mittelwert, Varianz, Maximum)

Der Punkt ist, dass T zufällig ist; der Wert schwankt mit der konkreten Stichprobe.

- Man kann also nicht erwarten, dass man immer den richtigen Wert trifft.
- Die Beurteilung der Güte des Schätzers bezieht sich auf Kenngrößen seiner Verteilung (v.a. Erwartungswert und Varianz)

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

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

  für alle ϑ.

- Die Größe

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

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

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

Anschauliche Interpretation:

Bsp. 2.3. [Fortsetzung des Beispiels]

Nehmen Sie an, die Stichprobenziehung sei gemäß einer reinen Zufallsauswahl erfolgt, d.h. jede Stichprobe hat dieselbe Wahrscheinlichkeit gezogen zu werden (hier $\frac{1}{6}$). Sind die oben betrachteten Schätzfunktionen T_1, T_2, T_3 erwartungstreu?

Für die Träger $\mathcal{T}_i$ von $T_i,\ i = 1, 2, 3$ gilt:

$$\begin{aligned} \mathcal{T}_1 &= \{1.5, 2, 2.5\} \\ \mathcal{T}_2 &= \{1, 2, 3\} \\ \mathcal{T}_3 &= \{1.\bar{3}, 2\} \end{aligned}$$

Bei T_1 gilt: $P(\{T_1 = 1.5\}) = P(\{T_1 = 2\}) = P(\{T_1 = 2.5\}) = \frac{2}{6} = \frac{1}{3}$

Bei T_2 gilt: $P(\{T_2 = 1\}) = P(\{T_2 = 2\}) = P(\{T_2 = 3\}) = \frac{2}{6} = \frac{1}{3}$

Bei T_3 gilt: $P(\{T_3 = 1.5\}) = \frac{2}{6} = \frac{1}{3}; \qquad P(\{T_3 = 3\}) = \frac{4}{6} = \frac{2}{3}$

und damit bei $\vartheta = \mu = 2$

$$\mathbb{E}_2(T_1) = \sum_{t_1 \in \mathcal{T}_1} t_1 \cdot P(\{T_1 = t_1\}) = \frac{1}{3}(1.5 + 2 + 2.5) = 2$$

In der Tat gilt allgemein: Das arithmetische Mittel ist erwartungstreu für den Erwartungswert.

$$\mathbb{E}_2(T_2) = \sum_{t_2 \in \mathcal{T}_2} t_2 \cdot P(\{T_2 = t_2\}) = \frac{1}{3}(1+2+3) = 2$$

Wieder gilt allgemein: Einzelne Stichprobenvariablen ergeben erwartungstreue Schätzer für den Erwartungswert.

$$\mathbb{E}_2(T_3) = \sum_{t_3 \in \mathcal{T}_3} t_3 \cdot P(\{T_3 = t_3\}) = \frac{1}{3} \cdot 1.\bar{3} + \frac{2}{3} \cdot 2 = \frac{16}{9} \neq 2$$

T_3 ist also nicht erwartungstreu. Es gilt

$$\text{Bias}(T_3) = \mathbb{E}_2(T_3) - 2 = \frac{16}{9} - \frac{18}{9} = -\frac{2}{9}$$

Bias und Erwartungstreue bei einigen typischen Schätzfunktionen

- 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 $\mathrm{E}_\mu(X_1) = \mathrm{E}_\mu(X_2) = \ldots = \mu$ folgt:

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

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

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

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

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

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

- Für die korrigierte Stichprobenvarianz gilt dagegen:

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

Also ist S^2 erwartungstreu für σ^2. Diese Eigenschaft ist auch die Motivation dafür, von einer Korrektur der Stichprobenvarianz zu sprechen.

- *Vorsicht:* Im Allgemeinen gilt, wie in Kapitel 1.5.3 ausgeführt, für beliebige, nichtlineare Funktionen g

$$\mathrm{E}\, g(X) \neq g(\mathrm{E}(X)).$$

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

Bsp. 2.4. [Wahlumfrage]

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

Bedeutung der Erwartungstreue: Erwartungstreue ist in gewisser Weise ein schwaches Kriterium, denn es gibt viele einsinnige erwartungstreue Schätzer!

Deshalb betrachtet man zusätzlich die Effizienz eines Schätzers, s.u.

2.2.3 Effizienz

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

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ T_2 &= \frac{1}{n/2}\sum_{i=1}^{n/2} X_i \end{aligned}$$

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

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

und damit

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

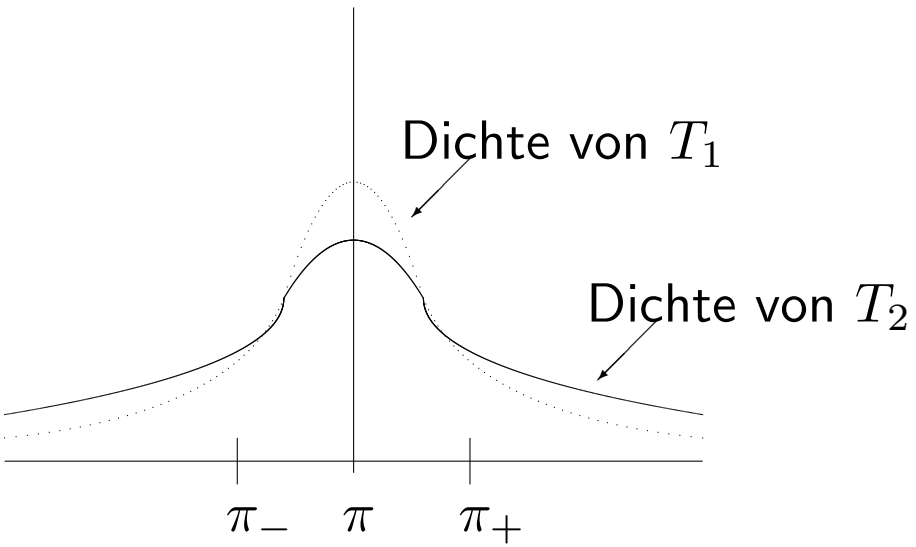
Analog kann man zeigen:

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

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“.

Andere Interpretation:



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.

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

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.

Definition 2.5. *Effizienz*

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

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

und

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

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

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

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

Bem. 2.6.

- *Inhaltliche Bemerkung:* Der (tiefere) „Sinn von Optimalitätskriterien“ bei der Auswertung der Stichprobe wird klassischerweise insbesondere auch in der *Gewährleistung von Objektivität* gesehen. Ohne wissenschaftlichen Konsens darüber, welcher Schätzer in welcher Situation zu wählen ist, wäre die Auswertung einer Stichprobe willkürlich und der Manipulation Tür und Tor geöffnet. Allerdings gibt es wirkliche Eindeutigkeit nur bei „idealen“, sauberen Daten. Z.B. sind ausreißerunempfindliche Verfahren bei „idealen“ Daten weniger effizient, haben aber den Vorteil, stabiler bei kleinen Abweichungen von den Verteilungsannahmen zu sein.

- Ist $X_1, \ldots, X_n$ eine i.i.d. Stichprobe mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$, dann ist

  * $\bar{X}$ UMVU-Schätzfunktion für μ bei bekanntem σ^2 und

  * S^2 UMVU-Schätzfunktion für σ^2 bei bekanntem μ.

- 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.

  Man zieht dann den sogenannten *Mean Squared Error*

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

  zur Beurteilung heran. Es gilt

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

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

2.2.4 Asymptotische Gütekriterien

- **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: Gefordert wird die Erwartungstreue nur 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 verwandte aber „etwas“ unterschiedliche Kriterien, z.B. das folgende:

- Ein Schätzer heißt (MSE-)konsistent oder konsistent im quadratischen Mittel, wenn gilt

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

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

$$\begin{aligned}\mathrm{MSE}(\bar{X}) &= \mathrm{Var}(\bar{X}) + \mathrm{Bias}^2(\bar{X}) \\ &= \frac{\sigma^2}{n} + 0 \\ &= \frac{\sigma^2}{n} \to 0.\end{aligned}$$

$\bar{X}$ ist also ein MSE-konsistenter Schätzer für den Erwartungswert.

- Anschaulich bedeutet die Konsistenz,

2.2.5 Konstruktionsprinzipien guter Schätzer

Die Methode der kleinsten Quadrate

$\Longrightarrow$ Regressionsanalyse

Das Maximum-Likelihood-Prinzip

- 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).

- 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 $Maximum\text{-}Likelihood\text{-}Schätzwert$; die zugehörige Schätzfunktion $T(X_1, \ldots, X_n)$ $Maximum\text{-}Likelihood\text{-}Schätzer$ (siehe genauer Definition 2.9).

Bsp. 2.7.

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:

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

„$P(\ldots || \pi)$ 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

Bem. 2.8.

- Zwei Sichtweisen auf
$$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n):$$
  - Deduktiv (Wahrscheinlichkeitsrechnung): ϑ bekannt, $x_1, \ldots, x_n$ zufällig („unbekannt“).
  - Induktiv (Statistik): ϑ unbekannt, $x_1, \ldots, x_n$ bekannt.
- Für stetige Verteilungen gilt
$$P_\vartheta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = 0$$
für beliebige Werte ϑ. 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, also den natürlichen Logarithmus der 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.

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

Dies liefert denselben Schätzwert $\hat{\vartheta}$ und erspart beim Differenzieren die Anwendung der Produktregel. Manchmal ist es noch geschickter, zunächst $\prod_{i=1}^{n} P_\vartheta(X_i = x_i)$ zu vereinfachen und dann zu logarithmieren.

- Bei den in Statistik II betrachteten „regulären“ Verteilungsmodellen reicht es, die erste Ableitung zu betrachten. Man kann zeigen, dass sie immer ein Maximum führt.

Definition 2.9. *(Zusammenfassung: Maximum-Likelihood-Schätzung)*

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*.

Bsp. 2.10. *(Maximum-Likelihood-Schätzer bei der Binomial und der Normalverteilung)*

ML-Schätzung bei Normalverteilung

- Der ML-Schätzer $\hat{\sigma}^2 = \tilde{S}^2$ für σ^2 ist die Stichprobenvarianz; diese ist nicht erwartungstreu.

Bem. 2.11. [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).