Kapitel 6

Ideen und Begriffe aus der Statistik

6.2

6.1 Grundlegende Begriffe, Schätzen von Parametern

Beispiel 6.1. Bei einer biologischen Expedition wurden $n = 53$ Krebse einer gewissen Art gefangen, davon $k = 23$ Weibchen.

Was sagt uns dies über den Weibchenanteil in der Population?
Gibt die Beobachtung (Weibchenanteil in der Stichprobe $\frac{23}{53} \approxeq 0{,}434$) Anlass, an einem ausgeglichenen Geschlechterverhältnis in dieser Population zu zweifeln?

Vorstellung: Eine sehr große Population von Krebsen mit (uns unbekanntem) Weibchenanteil $\vartheta \in (0, 1)$,
der naheliegendste Schätzwert für ϑ (angesichts der Beobachtungen) ist

$$\widehat{\vartheta} = \frac{23}{53} \ (\approxeq 0{,}434).$$

Bei einer biologischen Expedition wurden $n = 53$ Krebse einer gewissen Art gefangen, davon $k = 23$ Weibchen.

Was sagt uns dies über den Weibchenanteil ϑ in der Population?

Modell: $X = (X_1, X_2, \ldots, X_n)$,

$$X_i = \begin{cases} 1, & i\text{-ter gefangener Krebs ist Weibchen,} \\ 0, & i\text{-ter gefangener Krebs ist Männchen} \end{cases}$$

X_i sind u.a., $\sim \mathrm{Ber}_\vartheta$ (Wir tun hier so, als ob wir mit Zurücklegen rein zufällig aus der Gesamtpopulation gezogen hätten – diese Approximation ist für große Populationen gerechtfertigt.)

Wir interpretieren $\frac{k}{n} = \frac{23}{53}$ als Realisierung der Zufallsvariable

$$\widehat{\vartheta} := \frac{1}{n}\big(X_1 + \cdots + X_n\big)$$

Modell:

$$X_i = \begin{cases} 1, & i\text{-ter gefangener Krebs ist Weibchen,} \\ 0, & i\text{-ter gefangener Krebs ist Männchen} \end{cases}$$

X_i sind u.a., $\sim \mathrm{Ber}_\vartheta$, wobei ϑ der (uns unbekannte) tatsächliche Anteil der Weibchen in der Population ist.

$\hat{\vartheta} = \frac{1}{n}(X_1 + \cdots + X_n)$

1. (Punktschätzung) Was auch immer ϑ ist, es gilt

$$\mathbb{E}_\vartheta\big[\widehat{\vartheta}\big] = \vartheta,$$
$$\mathrm{Var}_\vartheta\big[\widehat{\vartheta}\big] = \frac{1}{n^2} n\vartheta(1-\vartheta) = \frac{1}{n}\vartheta(1-\vartheta) = \frac{\sigma^2}{n}$$

mit $\sigma = \sigma(\vartheta) = \sqrt{\vartheta(1-\vartheta)}$. ($\mathbb{E}_\vartheta$, etc. bezieht sich auf Erwartungswerte bezüglich dem Wahrscheinlichkeitsmaß, unter dem $X_i \sim \mathrm{Ber}_\vartheta$ und u.a. sind.)

$$\widehat{\sigma} := \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$$

ist ein naheliegender Schätzer für σ.

Ein Schätzer für die Standardabweichung von $\widehat{\vartheta}$ ist $\dfrac{\widehat{\sigma}}{\sqrt{n}}$.

Im Statistik-Jargon heißt dies auch der „Standardfehler“ (englisch: standard error [of the mean], SEM), dies ist eine naheliegende Maßzahl für die „Genauigkeit der Schätzung“.

(Im Beispiel:

$$\frac{23}{53}\cdot\frac{30}{53} \approxeq 0{,}246, \quad \widehat{\sigma} \approxeq 0{,}496, \quad \frac{\widehat{\sigma}}{\sqrt{53}} \approxeq 0{,}0681$$

man gibt also an: geschätzter Weibchenanteil $0{,}43 \pm 0{,}068$.)

Modell:

$$X_i = \begin{cases} 1, & i\text{-ter gefangener Krebs ist Weibchen,} \\ 0, & i\text{-ter gefangener Krebs ist Männchen} \end{cases}$$

X_i sind u.a., $\sim \mathrm{Ber}_\vartheta$, wobei ϑ der (uns unbekannte) tatsächliche Anteil der Weibchen in der Population ist.

2. („Wie genau ist die Schätzung?": **Konfidenzintervall**) Wenn der wahre Weibchenanteil ϑ ist, so ist $X_1 + \cdots + X_n \sim \mathrm{Bin}_{n,\vartheta}$, also

$$\frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} \approx \frac{\widehat{\vartheta} - \vartheta}{\sigma(\vartheta)/\sqrt{n}} \overset{d}{\approx} \mathcal{N}_{0,1}$$

gemäß dem Satz von de Moivre-Laplace (Satz 5.1 und Korollar 5.2), also

$$P_\vartheta\Big(-1{,}96 \le \frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} \le 1{,}96 \Big) \approx P(-1{,}96 \le Z \le 1{,}96) \approx 0{,}95$$

mit $Z \sim \mathcal{N}_{0,1}$, d.h. das (zufällige) Intervall

$$I := \Big[\widehat{\vartheta} - 1{,}96\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + 1{,}96\frac{\widehat{\sigma}}{\sqrt{n}}\Big] \quad \text{erfüllt } P_\vartheta(\vartheta \in I) \approx 0{,}95$$

(für jede Wahl von $\vartheta \in [0,1]$).

I heißt ein Konfidenzintervall (für ϑ) zum (approximativen) Niveau $0{,}95$.

(Im Beispiel: $I \approx [0{,}30, 0{,}57]$)

$\widehat{\vartheta} - 1{,}96 \cdot \frac{\widehat{\sigma}}{\sqrt{n}} \le \vartheta \le \widehat{\vartheta} + 1{,}96 \cdot \frac{\widehat{\sigma}}{\sqrt{n}}$

Modell:

$$X_i = \begin{cases} 1, & i\text{-ter gefangener Krebs ist Weibchen,} \\ 0, & i\text{-ter gefangener Krebs ist Männchen} \end{cases}$$

X_i sind u.a., $\sim \mathrm{Ber}_\vartheta$, wobei ϑ der (uns unbekannte) tatsächliche Anteil der Weibchen in der Population ist.

3. (**Testen von Hypothesen**) Passen die Beobachtungen zur Hypothese, dass der wahre Weibchenanteil in der Population $\vartheta_0 = \frac{1}{2}$ ist?

Wir beobachten

$$\left|\widehat{\vartheta} - \frac{1}{2}\right| = \left|\frac{23}{53} - \frac{1}{2}\right| \approxeq 0{,}07,$$

es ist

$$P_{\vartheta_0}\Big(\big|\widehat{\vartheta} - \vartheta_0\big| \geq 0{,}07\Big) \approx P\Big(|Z| \geq \sqrt{4 \cdot n} \cdot 0{,}07\Big) \approxeq 2\Big(1 - \Phi\big(0{,}96\big)\Big) \approxeq 0{,}336.$$

Demnach: Wenn die Hypothese $\vartheta = \vartheta_0 = \frac{1}{2}$ zutrifft, würden wir in ca. 1/3 der Fälle eine mindestens so große Abweichung $\left|\widehat{\vartheta} - \frac{1}{2}\right|$ wie die tatsächlich anhand der Daten beobachtete finden. Insoweit gibt die Beobachtung keinen Anlass, diese Hypothese anzuzweifeln.

Definition 6.2. Ein *statistisches Modell* ist ein Tripel $(\mathcal{M} =) \big(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta}\big)$, wo $\mathcal{X} \neq \varnothing$ Menge („Beobachtungs- oder Stichprobenraum"), $\mathscr{F} \subset 2^{\mathcal{X}}$ eine σ-Algebra, Θ eine Menge (mit $|\Theta| > 1$) und für jedes $\vartheta \in \Theta$ ist P_ϑ ein W'maß auf $(\mathcal{X}, \mathscr{F})$.

Das Modell $\mathcal{M}$ heißt *parametrisch*, wenn $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$, speziell *einparametrisch*, wenn $d = 1$.

$\mathcal{M}$ heißt *diskret*, wenn $\mathcal{X}$ abzählbar ist, $\mathcal{M}$ heißt *stetig*, wenn $\mathcal{X} \subset \mathbb{R}^n$ und jedes P_ϑ eine Dichte $\rho_\vartheta : \mathcal{X} \to [0, \infty]$ besitzt.

Ein diskretes oder stetiges Modell heißt ein *Standardmodell*.

$\big(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (P_\vartheta^{\otimes n})_{\vartheta\in\Theta}\big)$ heißt das n-fache Produktmodell von $\mathcal{M}$ (für Produktmaße vgl. Def. 2.21).

Definition 6.3. $\big(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta}\big)$ statistisches Modell, $(S, \mathscr{A})$ messbarer Raum.

1. Eine Zufallsvariable X (definiert auf $(\mathcal{X}, \mathscr{F})$ mit Werten in S, d.h. $X : \mathcal{X} \to S$ ist $\mathscr{F}$-$\mathscr{A}$-messbar) heißt eine *Statistik* (manchmal auch: „Stichprobe").

2. Sei $\tau : \Theta \to \mathbb{R}$ eine reelle Kenngröße (oder „Parametermerkmal"), eine Statistik $T : \mathcal{X} \to \mathbb{R}$ heißt ein *Schätzer* (genauer: „Punktschätzer") für τ.

3. Ein Schätzer T für τ heißt *erwartungstreu* (oder „unverzerrt"), wenn gilt

$$\forall\, \vartheta \in \Theta \;:\quad \mathbb{E}_\vartheta[T] = \tau(\vartheta).$$

   $b_\vartheta(T) := \mathbb{E}_\vartheta[T] - \tau(\vartheta)$ heißt die *Verzerrung* (englisch: bias) von T.

Die typische Konstruktion / Situation eines Schätzers ist $T = t(X)$ für eine Funktion $t : S \to \mathbb{R}$.

Man schreibt / benennt einen Schätzer für τ oft $\widehat{\tau}$.