Grundbegriffe der statistischen Modellierung

4

Die Denkweise der Statistik haben wir schon im Einführungsbeispiel in Kap. 1 kennengelernt. Es geht im Folgenden darum, diese Denkweise zu strukturieren und insbesondere die Begriffe rund um die schließende Statistik einzuführen. Das führt uns zum statistischen Modell in Abschn. 4.1 und zum Begriff der Statistik (Abschn. 4.2). In beiden Abschnitten denken wir zunächst an eine feste Anzahl n von Beobachtungen. In Abschn. 4.3 werden asymptotische Erweiterungen ($n \to \infty$) formuliert. Wer sich stärker für die maßtheoretische Formulierung der Begrifflichkeiten interessiert, sei beispielsweise auf Georgii (2009) verwiesen.

Wir denken bei der statistischen Analyse an einen Dreischritt:

1. Wahre Welt: Statistische Analysen sind motiviert durch eine Frage an oder eine Behauptung über eine Population. Wir sammeln dann Beobachtungen, also einen Ausschnitt der Population. Wir fragen: Sind die Beobachtungen mit der Behauptung verträglich?
2. Statistisches Modell: Dazu interpretieren wir die Population, die Behauptung und die Beobachtung im Rahmen eines theoretischen statistischen Modells. Insbesondere verstehen wir die Beobachtungen dabei als Realisierungen von Zufallsvariablen. Damit ist ein Modell immer eine Vereinfachung der Realität. Andererseits sollte es die Möglichkeit bieten, die Unverträglichkeit der Beobachtungen mit der Behauptung zu beurteilen.
3. Wahre Welt: Werden die Beobachtungen im Rahmen des Modells als unwahrscheinlich eingestuft, so interpretieren wir sie in der wahren Welt als nur schwer mit der Behauptung verträglich.

Schritt 1 ist in Abb. 4.1a dargestellt. Wir denken an folgende vier Aspekte, die wir im Kontext des Einführungsbeispiels aus Kap. 1 diskutieren. Zur Erinnerung: Der Organisator der Party steht vor dem Problem, dass der angedachte Raum möglicherweise nicht ausreichen

© Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2019

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_4

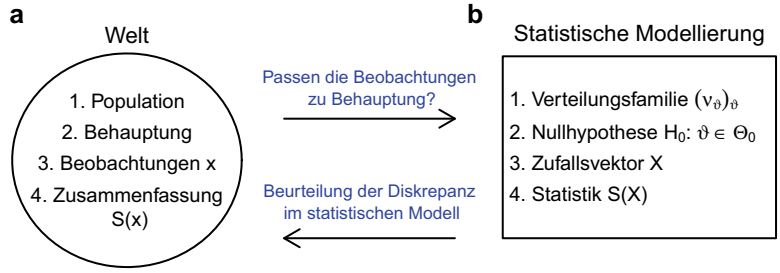

Abb. 4.1 Analogie zwischen angewandter Fragestellung (**a**) und statistischem Modell (**b**)

könnte. Dabei geht es erstens um eine unüberschaubare *Population,* zum Beispiel alle Studierenden aus dem Studiengang, die potenziell zur Party kommen könnten. Zweitens gibt es eine *Behauptung* über die Population: „Wie im Vorjahr nehmen 40% ($p^{(0)} = 0.4$) der Personen aus dieser Population an der Party teil." Dazu werden drittens *Beobachtungen* $\mathbf{x} = (x_1, \ldots, x_n)^t$ gesammelt. Der Organisator befragt Studierende, ob sie teilnehmen. Er erhält einen Ausschnitt aus der Population. Viertens werden die Beobachtungen zusammengefasst in einer *Statistik.* Der Organisator bestimmt den Anteil $\hat{p}(\mathbf{x}) = 0.58$ derjenigen, die zur Party erscheinen werden, unter den Befragten.

Der Anteil $\hat{p}(\mathbf{x}) = 0.58$ der Partyteilnehmer in der Stichprobe war größer als der behauptete Anteil $p^{(0)} = 0.4$. Gibt uns die beobachtete Diskrepanz $|\hat{p}(\mathbf{x}) - p^{(0)}| = 0.18$ einen Anlass, an der Behauptung zu zweifeln? Ist dieser Wert von 0.18 groß? Idee der statistischen Modellierung ist es, einen simplen Mechanismus zu formulieren, der beschreibt, wie die Beobachtungen zustande gekommen sein könnten – die Theorie des Zufalls ist hier das entscheidende Hilfsmittel: Wir interpretieren die Beobachtungen als Ausgang eines Zufallsexperiments. Das erlaubt die Beurteilung der Diskrepanz in Form von Wahrscheinlichkeitsaussagen.

Mit Hinblick auf oben genannten Dreischritt gehen wir in Schritt 2 über zum statistischen Modell, siehe auch Abb. 4.1b. Dessen Formulierung ist Teil von Abschn. 4.1.

4.1 Statistisches Modell

Ein statistisches Modell ist ein Mittel der Stochastik, das die Komplexität der realen Gegebenheiten auf wenige mathematische Annahmen reduziert. Im Einführungsbeispiel nehme man vielleicht an, dass die Antworten aller Befragten unabhängig gemäß eines Münzwurfs mit Erfolgswahrscheinlichkeit p generiert wurden. Modellannahmen sind grundsätzlich inkorrekt. Hier könnte die Frage nach der Teilnahme etwa auch vom Freundeskreis oder von parallelen Veranstaltungen abhängen. Zufall ist ein theoretisches Konzept. Der Vorteil ist, dass wir damit in der Lage sein werden, die Abweichung der Beobachtungen von den Modellannahmen zu beurteilen.

Die vier angesprochenen Aspekte der wahren Welt, siehe Abb. 4.1a finden Analogien im Rahmen des statistischen Modells, siehe Abb. 4.1b. Erstens wird die Population durch eine Verteilung ν_ϑ beschrieben. Genauer legt man eine ganze *Familie von Verteilungen* $(\nu_\vartheta)_{\vartheta\in\Theta}$ zugrunde, was zumindest zum Teil das mangelnde Wissen über die Population ausdrückt. Der andere Teil dieses mangelnden Wissens schlägt sich in den vereinfachten Annahmen wie etwa der Wahl der Verteilungsfamilie selbst nieder. Zweitens interpretiert man die Beobachtungen $\mathbf{x} = (x_1, \ldots, x_n)^t$ als Realisierung eines *Zufallsvektors* $\mathfrak{X} = (X_1, \ldots, X_n)^t$. Im Einführungsbeispiel etwa hatten wir die Komponenten $X_1, \ldots, X_n$ als unabhängige und identisch $ber(p)$-verteilte Zufallsvariable angenommen. Wir lassen also die Familie $(ber(p))_{p\in\Theta}$ sämtlicher Bernoulli-Verteilungen zu, mit $\Theta := [0, 1]$. Drittens nennen wir eine Teilmenge Θ_0 von Θ eine *Nullhypothese*. Die Nullhypothese fungiert als Analogon zur gemachten Behauptung. Im Einführungsbeispiel ist dies die einelementige Menge $\Theta_0 = \{0.4\}$. Wir verbinden damit die $ber(0.4)$-Verteilung. Viertens ist eine *Statistik* S eine Funktion des Zufallsvektors $\mathfrak{X}$. Im Einführungsbeispiel war dies die relative Häufigkeit $S = \hat{p}$. Der Vorteil: Im Rahmen des Modells wissen wir, wie sich $\hat{p}(\mathfrak{X})$ verteilt, wenn die Nullhypothese zutrifft, wenn also die Zufallsvariablen tatsächlich einen Erfolgsparameter von $p^{(0)} = 0.4$ aufweisen, vgl. Abb. 1.2. Eine Diskrepanz, die mindestens so groß ist wie die beobachtete, $|\hat{p}(\mathfrak{X}) - p^{(0)}| \geq 0.18$, tritt nur in etwa 0.2 % der Fälle auf.

Allgemein verstehen wir ein statistisches Modell gegeben durch

$$\begin{aligned}\text{Modell} \;\widehat{=}\;\; & 1.\ \text{Zufallsvektor } \mathfrak{X} = (X_1, \ldots, X_n)^t \text{ mit Bildraum } \mathcal{X} \subseteq \mathbb{R}^n \text{ und} \\ & 2.\ \text{Familie } (\nu_\vartheta)_{\vartheta\in\Theta} \text{ von Verteilungen auf } \mathcal{X}. \end{aligned} \tag{4.1}$$

Hier betrachten wir Verteilungen auf einer Teilmenge des $\mathbb{R}^n$ und nennen i. Allg. Θ die Indexmenge und ϑ den Index. Jedes Familienmitglied ist ein möglicher Kandidat für die Verteilung von $\mathfrak{X}$. Wir wissen nicht, welches die wahre Verteilung ist.

Dass ν die Verteilung des Vektors $\mathfrak{X}$ ist, ist festgelegt dadurch, dass $\nu(B) = \mathbb{P}(\mathfrak{X} \in B)$ für sämtliche Quader $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$, mit $a_i < b_i$. Nehmen wir weiter an, dass die Komponenten von $\mathfrak{X}$ unabhängig $(*)$ und identisch verteilt $(**)$ sind, so finden wir

$$\nu(B) = \mathbb{P}(\mathfrak{X} \in B) \overset{(*)}{=} \prod_{i=1}^{n} \mathbb{P}(X_i \in [a_i, b_i]) \overset{(**)}{=} \prod_{i=1}^{n} \mathbb{P}(X_1 \in [a_i, b_i]). \tag{4.2}$$

Unter Unabhängigkeit ist also die gemeinsame Verteilung der Komponenten schon durch die Angabe der Verteilung der einzelnen Komponenten festgelegt.

Sind die einzelnen Komponenten zudem identisch verteilt, so reicht die Angabe der Verteilung der ersten Komponente. Ein solches Modell beschreiben wir häufig durch die Formulierung der Komponenten anstelle des Vektors selbst, etwa

$$\begin{aligned}\text{Modell} \;\widehat{=}\;\; & 1.\ \text{Unabhängige und identisch verteilte Zufallsvariable } X_1, \ldots, X_n \\ & \quad \text{mit Bildraum } \mathcal{X}_1 \subseteq \mathbb{R} \text{ und} \\ & 2.\ \text{Familie } (\nu_\vartheta)_{\vartheta\in\Theta} \text{ von Verteilungen auf } \mathcal{X}_1. \end{aligned} \tag{4.3}$$

Der Bildraum des Vektors $\mathfrak{X} = (X_1, \ldots, X_n)^t$ ist dann der Produktraum $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_1$. Jede Verteilung ν_ϑ beschreibt eine Kandidatenverteilung der Komponente X_1, welche wiederum die Verteilung des Vektors $\mathfrak{X}$ gemäß (4.2) eindeutig festlegt.

Oft verwenden wir eine konkrete Verteilungsfamilie $(\nu_\vartheta)_{\vartheta\in\Theta}$, bei der jedes Mitglied ν_ϑ eine Dichte f_ϑ oder Gewichte g_ϑ besitzt, und sprechen dann von einem *parametrischen* Modell. In parametrischen Modellen nennen wir Θ Parameterraum und ϑ den (durch die Parametrisierung gegebenen) Parameter. Ginge es etwa um die Beobachtungen in Abb. 3.1, die sich annähernd glockenförmig verteilen, wäre es unter der Annahme, dass die Beobachtungen unabhängig und identisch verteilt sind, vielleicht naheliegend, sämtliche Normalverteilungen als potenzielle Kandidatenverteilungen für die erste Komponente zuzulassen. Deren Dichte ist gegeben durch

$$f_{(\mu,\sigma^2)}(x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_1-\mu)^2\right) \qquad \text{mit} \qquad (\mu,\sigma^2) \in \Theta := \mathbb{R}\times\mathbb{R}^+,$$

und so ist der Parameterraum mit $(\mu, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}^+$ zweidimensional, d. h. $d = 2$. Andererseits würde die schiefe Verteilung der Beobachtungen in Abb. 3.2 möglicherweise die Familie der Exponentialverteilungen als Kandidatenverteilungen für die erste Komponente nahelegen,

$$f_\lambda(x_1) = \lambda \exp(-\lambda x_1) \qquad \text{mit} \qquad \lambda \in \Theta := \mathbb{R}^+.$$

Hier ist der Parameterraum mit $\lambda \in \Theta := \mathbb{R}^+$ eindimensional, d. h. $d = 1$.

Möchten wir uns andererseits nicht auf eine konkrete Verteilungsfamilie festlegen, so sprechen wir von *nichtparametrischen* Modellen. Beispielsweise könnten wir die Familie *aller* reellwertigen Verteilungen zugrunde legen, vgl. Beispiel 4.2. In nichtparametrischen Modellen bezeichnen wir die Verteilungsfamilie allgemein durch $(\nu_\vartheta)_{\vartheta\in\Theta}$. Eine mögliche einfache Form der Indizierung wäre dann etwa, Θ als die Menge der relevanten Verteilungen selbst zu wählen und jede Verteilung mit sich selbst zu indizieren. Einfache Beispiele statistischer Modelle sind:

Beispiel 4.1 (Das Bernoulli-Modell des Einführungsbeispiels)
Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim ber(p)$ und $p \in \Theta := [0, 1]$. Der Bildraum ist hier $\mathcal{X}_1 = \{0, 1\}$, und die zugehörige Verteilungsfamilie ist $(ber(p))_{p\in[0,1]}$.

Beispiel 4.2 (Ein allgemeines nichtparametrisches Modell)
Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ ist Mitglied der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller reellwertigen Verteilungen.

Beispiel 4.3 (Ein Modell mit Normalverteilungsannahme)
Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim N(\vartheta, 1)$ und $\vartheta \in \Theta = \mathbb{R}$.

Wir finden dann für $a < b$

$$\nu_\vartheta([a,b]) = \mathbb{P}_\vartheta(X_1 \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\vartheta)^2\right) dx.$$

Dass wir uns dabei auf den Parameter ϑ beziehen, lesen wir wie folgt: „Unter der Annahme, dass ν_ϑ die wahre zugrunde liegende Verteilung ist, entspricht das Ereignis $\{X_1 \in [a,b]\}$ gerade dem Integral der rechten Seite." Entsprechend wird die Schreibweise und die Interpretation auf sämtliche Kenngrößen der Verteilung ν_ϑ vererbt. Wir schreiben also zum Beispiel $\mathbb{E}_\vartheta[X_1] = \vartheta$, sowie $\mathbb{V}ar_\vartheta(X_1) = 1$ und sagen: „Wenn ϑ der wahre Parameter ist, dann ist der Erwartungswert von X_1 gerade ϑ", bzw. „Unter sämtlichen Verteilungen ist die Varianz von X_1 konstant 1".

4.2 Statistik und Schätzer

Wir erinnern an Abb. 4.1. Im Rahmen des Modells formulieren wir nun die Nullhypothese und die Statistik. Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X} \subseteq \mathbb{R}^n$ und eine Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ von Verteilungen auf $\mathcal{X}$. Eine *Nullhypothese* ist eine Teilmenge

$$\Theta_0 \subseteq \Theta,$$

und fungiert als Analogon zu einer Behauptung über die Population. Eine *Statistik* ist eine Abbildung

$$S : \mathcal{X} \to \Gamma,$$

wobei wir beim Bildraum Γ in der Regel auch an die reellen Zahlen oder eine geeignete Teilmenge von $\mathbb{R}$ denken. Die Aufgabe der Statistik ist es, die Zufallsvariablen problemabhängig zusammenzufassen.

Wir betrachten das Modell aus Beispiel 4.3, dargestellt in Abb. 4.2a, b. Dort ist die Nullhypothese einelementig, d. h. $\Theta_0 = \{\vartheta^{(0)}\}$. In der Abbildung erkennen wir die zugehörige Dichte $f_{\vartheta^{(0)}}$ (magentafarben). Die Beobachtungen x_i (blau) liegen zum großen Teil tief in den Flanken von $f_{\vartheta^{(0)}}$, und das ist untypisch, falls die Nullhypothese zutrifft, d. h., falls die Beobachtungen x_i tatsächlich Realisierungen unabhängiger Zufallsvariablen X_i mit Dichte $f_{\vartheta^{(0)}}$ sind. Um die Unverträglichkeit der Beobachtungen x_i und $f_{\vartheta^{(0)}}$ zu beschreiben, werden die Beobachtungen zusammengefasst in der Statistik S, hier formuliert als der empirische Mittelwert $S(\mathbf{x}) = \bar{x}_n$. Neben den Beobachtungen kann nun auch der Zufallsvektor $\mathfrak{X}$ weiter verarbeitet werden zu $S(\mathfrak{X}) = \bar{X}_n$. Durch jede Kandidatenverteilung ν_ϑ

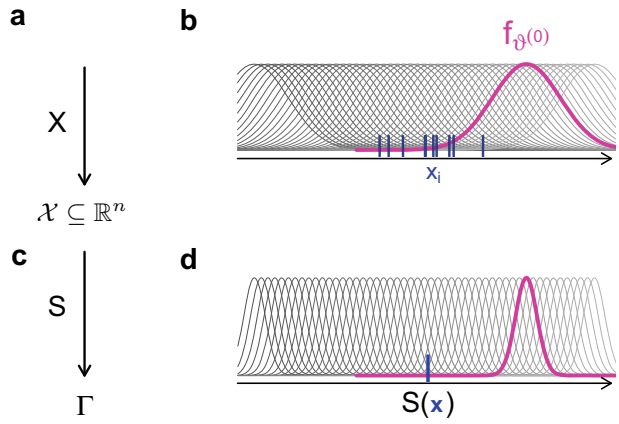

Abb. 4.2 Statistisches Modell, bestehend aus **a** Zufallsvektor $\mathfrak{X}$ und **b** Familie der Normalverteilungen mit Varianz 1. Beobachtungen **x** aufgefasst als Realisierung von $\mathfrak{X}$. **c** Statistik als Abbildung. **d** Die Verteilungen der Statistiken $S(\mathfrak{X}) = (1/n)\sum X_i$ unter sämtlichen Mitgliedern der Familie

wird dann eine Verteilung der Statistik $S(\mathfrak{X})$ induziert, Abb. 4.2c, d: Für jedes $\vartheta \in \Theta$ gilt, dass $S(\mathfrak{X}) \sim N(\vartheta, 1/n)$, denn die Summe unabhängiger und identisch normalverteilter Zufallsvariablen mit Erwartungswert ϑ und Varianz 1 ist wieder normalverteilt mit Erwartungswert $n\vartheta$ und Varianz n. In Abb. 4.2 erkennen wir, dass jede Kandidatenverteilung der Beobachtungen (b) eine zugehörige Verteilung der Statistik $S(\mathfrak{X})$ (d) induziert. Insbesondere gilt, dass unter der Annahme, dass $f_{\vartheta(0)}$ die Dichte der X_i beschreibt, die Zufallsvariable $S(\mathfrak{X})$ der $N(\vartheta^{(0)}, 1/n)$-Verteilung folgt (d, magentafarben). Wir können nun die Statistik $S(\mathbf{x})$ basierend auf den Daten **x** vergleichen mit der magentafarbenen Verteilung von $S(\mathfrak{X})$. Wieder erkennen wir: „Unter der Annahme, dass die Nullhypothese zutrifft, ist etwas Unwahrscheinliches eingetreten." Insbesondere können wir diese Aussage – durch die Zusammenfassung mit der Statistik S – nun quantifizieren, zum Beispiel durch die Wahrscheinlichkeit, dass $S(\mathfrak{X})$ im gegebenen Modell mindestens so tief in der linken Flanke sitzt. In Abb. 4.2 ist etwa

$$P(\mathbf{x}) := \mathbb{P}_{\vartheta(0)}(S(\mathfrak{X}) < S(\mathbf{x})) < 10^{-20}.$$

Ausdrücke dieser Art werden wir später als den P-Wert kennenlernen. Der P-Wert kann Werte zwischen null und eins annehmen. Hier ist er verschwindend klein, und auf Basis dessen kommen wir schließlich von der statistischen Modellierung zurück zur Realität und interpretieren die Daten **x** aufgrund des winzigen P-Wertes als kaum mit der Behauptung verträglich. Diese Interpretation der Unverträglichkeit der Beobachtungen mit der Behauptung schließt den letzten Schritt des Dreischritts ab.

In der Praxis muss eine Statistik geeignet gewählt werden und kann verschiedene Funktionen erfüllen. Zum einen sollte sie, wie oben beschrieben, die für die Fragestellung relevanten Abweichungen der Beobachtungen vom Modell gut quantifizieren. Ein weiterer Einsatz von Statistiken ist das *Schätzen.* Wenn eine Statistik eine Kenngröße $\tau(\vartheta)$ der zugrunde liegenden Verteilungsfamilie $(\nu_\vartheta)_{\vartheta\in\Theta}$ schätzen soll, wird sie auch als *Schätzer* von $\tau(\vartheta)$ bezeichnet. In obigem Beispiel der Normalverteilungen verstehen wir etwa den

Erwartungswert der Individualbeobachtung $\tau(\vartheta) = \mathbb{E}_\vartheta[X_1] = \vartheta$ als eine Kenngröße der Verteilung. In parametrischen Modellen nennen wir $\tau(\vartheta)$ auch einen abgeleiteten Parameter. Jedenfalls ist τ eine Abbildung von Θ nach Γ, sodass die zu schätzende Kenngröße im Bildraum des Schätzers liegt. Um den Erwartungswert zu schätzen, könnte man etwa den Mittelwert betrachten, also $S(\mathbf{x}) = \bar{x}_n$.

Beispiel 4.4 (Schätzer im Bernoulli-Modell)
Wir betrachten Beispiel 4.1. Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim ber(p)$ und $p \in \Theta := [0, 1]$.

i. Dann verstehen wir die relative Häufigkeit (den Mittelwert) $\hat{p}(\boldsymbol{x}) = \bar{x}_n$ als Schätzer für p. Da $\mathbb{E}_p[X_1] = p$ für alle $p \in [0, 1]$, nutzen wir hier den Mittelwert als Schätzer für den Erwartungswert von X_1.
ii. Wir nutzen die empirische Varianz $s^2(\boldsymbol{x}) = (1/(n-1)) \sum_{i=1}^n (x_i - \bar{x}_n)^2$ als Schätzer für den abgeleiteten Parameter $\tau(p) = \mathbb{V}ar_p(X_1) = p(1-p)$, d. h. für die Varianz von X_1.

Analog verstehen wir in nichtparametrischen Modellen die empirischen Momente als Schätzer für die theoretischen Momente.

Beispiel 4.5 (Schätzer in einem nichtparametrischen Modell)
Es seien $X_1, \ldots, X_n$, unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ ist Mitglied der Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ aller reellwertigen Verteilungen, deren erste beiden Momente existieren.
In diesem nichtparametrischen Modell interpretieren wir den Erwartungswert $\mathbb{E}_\vartheta[X_1]$ und die Varianz $\mathbb{V}ar_\vartheta(X_1)$ als abgeleitete Kenngrößen der Verteilung ν_ϑ und verstehen wieder den Mittelwert $\bar{x}_n$ bzw. die empirische Varianz $s^2(\boldsymbol{x})$ als deren Schätzer.

In beiden Beispielen haben wir den Mittelwert und die empirische Varianz als Schätzer genutzt. Es gibt aber auch andere Möglichkeiten, denn ein Schätzer ist im Grunde lediglich eine Abbildung vom Raum $\mathcal{X}$. Damit sind grundsätzlich viele Funktionen als Schätzer zugelassen, zum Beispiel auch die wenig sinnvolle konstante Funktion $S \equiv 1/2$. Daher ist es wichtig, Schätzer miteinander vergleichen zu können, um für das jeweilige Problem einen ‚guten" Schätzer auszuwählen. Die Bewertung der Güte von Schätzern ist Thema von Abschn. 5.1.

4.3 Folgen von Modellen und Statistiken

Bisher haben wir Statistiken betrachtet, bei denen die Verteilung von $S(\mathfrak{X})$ zu vorgegebener Kandidatenverteilung von $\mathfrak{X}$ bekannt ist. Oft lässt sich die Verteilung aber nicht so einfach bestimmen. Man versucht dann gerne, die Verteilung durch asymptotische Betrachtungen zu

approximieren. Seien etwa $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim U[0, b]$ und $b \in \Theta := (0, \infty)$ und $S(\mathfrak{X}) = \bar{X}_n$. Laut des Zentralen Grenzwertsatzes 2.11 ist $S(\mathfrak{X})$ für große n näherungsweise normalverteilt mit Erwartungswert $\mathbb{E}[X_1]$ und Varianz $\mathbb{V}ar(X_1)/n$. Dann gilt approximativ, dass $\bar{X}_n \sim N(b^{(0)}/2, (b^{(0)})^2/(12n))$, falls $b^{(0)} \in \Theta$ dem wahren Parameter entspricht, siehe auch Beispiel 2.6iv. Diese approximative Verteilung von $\bar{X}_n$ kann dann für große n zum Vergleich mit dem auf den Beobachtungen basierenden Mittelwert $\bar{x}_n$ herangezogen werden.

Approximative Betrachtungen dieser Art basieren auf Folgen von Zufallsvariablen. Daher formulieren wir entsprechend *Folgen von Modellen und Statistiken*. In Anlehnung an 4.1 sei

$$\begin{aligned} \text{Modell} \;\widehat{=}\; & 1.\ \text{Zufallsvektor } \mathfrak{X}_\infty = (X_1, X_2, \ldots)^t \text{ mit Bildraum } \mathcal{X} \text{ und} \\ & 2.\ \text{Familie}(\nu_\vartheta)_{\vartheta \in \Theta} \text{ von Verteilungen auf } \mathcal{X}. \end{aligned} \tag{4.4}$$

Für sämtliche $n = 1, 2, \ldots$ bezeichne $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ die Einschränkung von $\mathfrak{X}_\infty$ auf seine ersten n Komponenten. Schränken wir jede Kandidatenverteilung ν_ϑ von $\mathfrak{X}_\infty$ auf die Randverteilung der Komponenten von $\mathfrak{X}_n$ ein, so erhalten wir ein Modell bezüglich der ersten n Beobachtungen. Wir nennen dies das *n-te restringierte Modell*. Bezüglich des n-ten restringierten Modells betrachten wir eine Statistik S_n, wobei wir daran denken, dass die funktionale Form der Statistik S_n unter n gleich bleibt, wie etwa beim Mittelwert.

Im Kontext solcher Folgen können wir dann die Anzahl n an Zufallsvariablen anwachsen lassen und das Verhalten von $S_n(\mathfrak{X}_n)$ für $n \to \infty$ studieren. Wir sprechen dann auch von einer Folge von Modellen und Statistiken.

4.4 Dialog: Statistische Modelle

Nachdem in der Vorlesung statistische Modelle eingeführt wurden, sind für einen Studenten (**S**) noch ziemlich viele Fragen offengeblieben. Manche Grundideen leuchten ihm einfach nicht ganz ein. In der Hoffnung auf Antworten wendet er sich an die Doktorandin (**D**), die seine wöchentliche Übungsgruppe leitet.

S: Ich verstehe nicht, warum in der Vorlesung so ein Aufwand um das statistische Modell betrieben wird. Die Zufallsvariablen aus den Beispielen und ihre Eigenschaften kennen wir doch schon aus der Einführung in die Stochastik.
D: Stimmt, in der Statistik benutzen wir Ideen aus der Stochastik. Wir kennen viele Begriffe schon. Aber was bei der Denkweise der statistischen Modellierung dazu kommt, ist, dass wir echte, in der Realität gemachte Beobachtungen durch ein virtuelles Modell beschreiben wollen. Und dabei ist dann eben auch die Auswahl eines geeigneten Modells wichtig.
Der Student ist immer noch skeptisch.

S: Das verstehe ich sowieso nicht. Erstens: Wir haben doch gelernt, dass ein Modell eine grobe Vereinfachung ist und damit letztlich immer falsch. Wieso soll ich denn so ein „falsches“ Modell dann überhaupt verwenden? Und zweitens: Wie soll ich mich zwischen verschiedenen Modellen, die aber in diesem Sinne alle falsch sind, für eines entscheiden?
D: Das sind sehr wichtige Fragen, und ich glaube, vor allem über die zweite Frage streiten Statistiker sehr häufig. Fangen wir mal mit der ersten an: Wieso soll ich ein Modell verwenden? Gegenfrage: Was soll ich denn sonst machen?
Nach einigem Überlegen schlägt der Student vor:
S: Vielleicht sollten wir einfach gar keine Beobachtungen anschauen und nur die statistischen Modelle! Mit denen kann man doch gut Mathematik machen.
D: Ja, warum nicht? In der Mathematischen Statistik macht man das auch meistens so. Wenn Du Dich darauf spezialisieren willst, bist Du in guter Gesellschaft!
Um deutlich zu machen, dass es aber sehr wohl auch Fragestellungen aus der Praxis gibt, um die sich ein Statistiker kümmern wollen könnte, verweist die Doktorandin auf das Beispiel der Fachschaftsfeier.
D: Vielleicht kommen wir noch mal auf das Einführungsbeispiel zurück: Was sagst Du denn jetzt dem Organisator der Party? Soll er einen größeren Raum reservieren oder nicht?
S: Puh, das ist schwer. An diese Frage traue ich mich eigentlich gar nicht ran, wenn ich ehrlich bin. . . Am Ende kommt es dann ganz anders, und dann stehen wir dumm da.
D: Sehr gut! Das ist eigentlich schon die erste wichtige Botschaft: Was wirklich los sein wird, können wir überhaupt nicht sagen! Aber trotzdem möchten wir uns vielleicht zumindest ein grobes Bild machen.
S: Auch wenn wir wissen, dass es eigentlich falsch ist?
D: Ja, auch wenn wir wissen, dass es falsch ist – was wir natürlich immer im Auge behalten müssen, wenn wir Ergebnisse interpretieren!
Der Student hakt weiter nach, er will ja noch eine Antwort auf seine zweite Frage.
S: Aber welches Modell nehme ich denn jetzt? Wieso nehme ich zum Beispiel für jeden Studenten an, dass er unabhängig von den anderen mit gleicher Wahrscheinlichkeit die Party besuchen wird? Ich weiß doch zum Beispiel, dass meine Clique nur gesammelt zur Party erscheint oder gar nicht, und ich weiß auch, dass manche Studenten fast auf keine einzige Party gehen, während andere keine Party auslassen. Da sind doch dann die Wahrscheinlichkeiten völlig unterschiedlich!
D: Moment mal. Die Annahmen der Unabhängigkeit und gleichen Erfolgswahrscheinlichkeit beziehen sich ja nicht auf die gesamte Population, sondern nur auf die Stichprobe. Wir nehmen an, dass jedes befragte Individuum unabhängig von den anderen befragten Individuen mit gleicher Wahrscheinlichkeit p die Party besuchen wird, und weiter nehmen wir an, dass dieses p in der gesamten Population den Anteil der Partybesucher beschreibt. Ob manche Individuen der Population praktisch immer und andere fast nie

auf eine Party gehen, spielt dabei keine Rolle: Die Gesamtpopulation hat einen wahren Anteil an Partybesuchern.
S: Aha, und aus dieser Population muss ich dann nur noch unabhängig und rein zufällig ziehen?
Damit hat der Student eines der zentralen Probleme angesprochen.
D: Jawoll, aber das ist tatsächlich ziemlich schwierig. Hast du eine Idee, warum?
Der Student muss wieder kurz überlegen.
S: Hm... Unabhängigkeit könnte zum Beispiel dann keine vernünftige Annahme sein, wenn ich keine Einzelpersonen, sondern Gruppen befrage, oder?
D: Ja, da hast du völlig recht, man sollte die Leute einzeln befragen. Das scheint aber in unserem Fall auch so gemacht worden zu sein. Dann bleibt aber auch noch das Problem des rein zufälligen Ziehens: Es wurden ja nur Studierende auf dem Campus befragt. Ist es Deiner Erfahrung nach realistisch, dass wir jeden Studierenden etwa gleich häufig auf dem Campus treffen?
Der Student lacht auf.
S: Quatsch! Manche habe ich schon seit Wochen nicht mehr gesehen!
D: Genau. Wenn die Stichprobe aber gar nicht rein zufällig aus der Population gezogen wurde, sondern vielleicht aus einer Teilpopulation derer, die häufig auf dem Campus sind, und der Anteil an Partybesuchern in dieser Teilpopulation anders ist als in der gesamten Population, dann kann es natürlich sein, dass wir mit unserer Schätzung total danebenliegen.
S: Okay, aber das ist ja dann wirklich blöd... Sollte man nicht versuchen, das ins Modell mit aufzunehmen?
D: Das kann man versuchen – wenn man genauere Informationen über die Teilpopulationen hat. Dabei muss man aber immer abwägen, wie viele potentielle Fehler man vielleicht durch zusätzliche Annahmen einbauen könnte, die dann die Vorhersage vielleicht sogar verschlechtern. Oft fährt man mit möglichst einfachen Modellen am besten...
S: Okay, notiert: einfache Modelle!