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 X in der Grundgesamtheit $\mathcal{G}$.
- Ziehe eine Stichprobe $(g_1, \ldots, g_n)$ von Elementen aus $\mathcal{G}$ und werte 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 (vgl. Bem. 1.50).

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 ("verteilungsfreie Verfahren")* (hier kaum behandelt)

- 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:
Treffe mittels der Auswertung einer

- Zufallsstichprobe
- möglichst gute $\star\star\star$
- Aussagen $\star\star$
- über bestimmte Charakteristika $\star$
- der Grundgesamtheit.

$\star$ Welche Charakteristika sind für die Fragestellung relevant? Natürlich werden für die Inferenz bezüglich des Erwartungswerts andere Methoden als für Schlüsse über die Varianz benötigt.

$\star\star$ verschiedene Formen:

- Punktschätzung: z.B. wahrer Anteil 0.4751
- Interhvallschätzung: 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

$\star\star\star$ Was heißt gut?

- Festlegung von „Gütekriterien“ (Genauigkeit? Wahrscheinlichkeit eines Fehlers gering?)
- Wie konstruiert man ein gutes/optimales Verfahren?
- Sicherheitsstellung der „Objektivität der statistischen Analyse“. Jeder wendet das beste Verfahren an $\Rightarrow$ gleiche Auswertung

Wichtig: Festgelegt werden *vor* dem Ziehen der Stichprobe (bzw. vor dem Bekanntwerden der Daten) *Auswertungsverfahren*, also Zufallsvariablen. Ihre Eigenschaften werden mittels der Wahrscheinlichkeitsrechnung ermittelt.

Man beachte, dass die ganze Argumentation auf der Zufälligkeit der Stichprobenziehung aufbaut. **Methoden der statistischen Inferenz sind also nicht geeignet für nicht zufällige Auswahlen und auch nicht für Vollerhebungen**. Bei letzteren sind sie nicht notwendig, da man ja hier die Grundgesamtheit kennt. Bei nicht zufälligen Auswahlen greift die Grundidee, den nicht ausschließbaren Induktionsfehler durch die Wahrscheinlichkeitsrechnung zu kontrollieren, nicht. Die entsprechenden Schlüsse weisen also einen unkontrollierten Fehler auf. Nicht zufällige Auswahlen entstehen z.B. durch Auswahl auf das Geratewohl (z.B. im Internet; Fenster popt auf: „Haben Sie Zeit, uns ein paar Fragen zu beantworten?“). Problematisch in diesem Kontext sind auch Untersuchungen, bei denen die Teilnahmeverweigerung von Personen nicht zufällig ist, sondern mit inhaltlich interessierenden Merkmalen der Verweigerenden zusammenhängt. Hier sind mindestens aufwändige, modellbasierte Korrekturverfahren nötig; oft ist aber auch eine auf absolut präzise Ergebnisse zielende induktive Verallgemeinerung der Ergebnisse der Auswertung schlicht nicht zulässig.

2.2 Punktschätzung

Ziel: Finde ein möglichst gutes Schätzverfahren und damit 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. "Jede" 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 den Mittelwert der Daten)"

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:

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

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:

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