5. Schließende Statistik (Inferenzstatistik, konfirmatorische Verfahren)

5.1. Einführung

- **Schätzen** unbekannter Parameter im Modell, z.B.
  Wahrscheinlichkeiten p_i (Anteile in der Gesamtmenge),
  Erwartungswerte (Durchschnittswerte in der Gesamtmenge),
  Parameter von Verteilungen auftretender ZV $(\mu, \sigma, \dots)$
- **Testen** von Hypothesen über diese Parameter bzw. Verteilungen, d.h. über die Angepasstheit eines Modells und damit schließlich über die interessierenden Verhältnisse in der Grundmenge (Population)

Jeder Schluss von einer Teilerhebung (Stichprobe) auf die Grundmenge ist mit Unsicherheiten verbunden.
Die wahrscheinlichkeitstheoretischen Modelle ermöglichen es, diese Unsicherheiten zu quantifizieren.

Statistische Grundbegriffe

"konkrete Natur"	stochastisches Modell
1) Population Studenten, Produkte, Werkstücke = Merkmalsträger interessierende Merkmale: Meinungen, Haltbarkeitsdauer, Abmessung oder (einfacher) 2) zufälliger Versuch Wurf mit einem Würfel mögliche Versuchsausgänge $\{1, 2, \ldots, 6\}$	Grundgesamtheit $(\Omega, \mathcal{A}, P)$ 1) Menge der möglichen Merkmalsausprägungen $\omega \in \Omega$ $\Omega = \{ja, nein\}$ $\Omega = [0, \infty)$ $\mathcal{A}$ geeignetes Ereignisfeld $P(A)$ so, dass z.B. $P(\{ja\})$ den Anteil der Befürworter, bzw. P([a,b]) den Anteil der Teile mit Abmessung $\in [a, b]$ usw. angibt, P unbekannt! 2) klar
Alle Merkmalsausprägungen bzw. Versuchsausg. als Zahlen codieren ($ja = 0, nein = 1$)	$\rightarrow$ ZV X über $(\Omega, \mathcal{A}, P)$ mit Verteilungsfunktion F_X (unbekannt!)
1) Ziehen von n Elementen aus der Population (Zurücklegen?) Feststellung der Merkmalsausprägung $x_i,\ i = 1, \ldots, n$ 2) n Versuche mit Ausgängen $x_1, x_2, \ldots, x_n$	n (unabhängige) Zufallsvariablen $X_1, X_2, \ldots, X_n$, alle mit der gleichen Verteilungsfunktion F_X
$(x_1, x_2, \ldots, x_n)$	$(X_1, X_2, \ldots, X_n)$
konkrete Stichprobe	mathematische Stichprobe

- n heißt Stichprobenumfang
- übliche Sprechweise für Modellannahmen:
  "Die SP $(x_1, \ldots, x_n)$ entstamme einer nach F_X verteilten Grundgesamtheit."
- Problem: Auswahl aus der Population
  - Ziel: Stichprobe als möglichst getreues Abbild der Realität (repräsentativ)
  - Hauptmethode: Einfache Zufallsstichprobe, jedes Objekt hat die gleiche Chance gezogen zu werden.
  - andere Varianten: systematische Ziehungen, geschichtete Zufallsstichproben, ...
  - Sonderfall: Totalerhebung
- Praktisch hat man es stets mit der konkreten Stichprobe $(x_1, \ldots, x_n)$ zu tun, mit deren Hilfe man Informationen über die Population gewinnen will.
  Die mathematische Stichprobe dient zur wahrscheinlichkeitstheoretischen Begründung der Schlussweisen.
- Werden mehrere Merkmale registriert, bzw. besteht das Anliegen im Vergleich verschiedener Merkmale bzw. verschiedener Populationen, sind entsprechend bei der Modellbildung verschiedene Zufallsvariablen $(X, Y, \ldots)$ einzuführen.

- Zufallsvariablen $X_1, X_2, \ldots, X_n$ sind unabhängig, wenn die Ereignisse

$$\{X_1 < x_1\}, \{X_2 < x_2\}, \ldots, \{X_n < x_n\}$$

für beliebige $x_1, x_2, \ldots, x_n \in R$ unabhängig sind.

- Es gilt:

Sind die ZV $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2$, und unabhängig,

dann $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Stichprobenfunktionen

Die Anliegen der schließenden Statistik werden mit

Stichprobenfunktionen

realisiert.

Stichprobenfunktion T, eine Funktion von n Veränderlichen

$$(X_1, X_2, \ldots, X_n) \longrightarrow T = T(X_1, X_2, \ldots, X_n)$$

math. Stichpr. **Zufallsvariable**

$$(x_1, x_2, \ldots, x_n) \longrightarrow t = T(x_1, x_2, \ldots, x_n)$$

konkrete Stichpr. **Zahl**

Bemerkungen

- T bzw. t sind allgemein übliche Bezeichnungen, für spezielle Stichprobenfunktionen sind aber auch andere Bezeichnungen üblich.
  Beispiel:
  $$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \bar{x} = \ldots$$
- Stichprobenfunktionen begegnen uns also als Formeln:
  Setzen wir die Werte der konkreten SP ein, kommt eine Zahl t heraus. Setzen wir die ZV der mathematischen SP ein, kommt eine Zufallsvariable T heraus.
- t kann als Realisierung der Zufallsvariable T verstanden werden.

5.2. Punktschätzungen

Beispiel:

GSTAT (Fred Böker, *Statistik lernen am PC*, Vandenhoeck & Ruprecht 1989)

enthält (u.a.) für das Jahr 1974 die Altersverteilung aller Personen, die in diesem Jahr in der Bundesrepublik Deutschland gemeldet waren, sowie die Möglichkeit, das Ziehen einer Stichprobe zu simulieren und deren Verteilung mit der tatsächlichen (über Histogramme und Mittelwerte) zu vergleichen.

- gesucht: Durchschnittsalter μ der Bevölkerung
- gegeben: konkrete SP.: $x_1, \ldots, x_n$
- plausibel:

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

  als Schätzung für μ
- Wie gut ist diese Schätzung?
- Dazu:
  - ZV X, die die unbekannte Altersverteilung beschreibt, dann

$$\mu \;=\; E(X)$$

  - $X_1, \ldots, X_n$ unabhängige ZV, verteilt wie X

– Die Stichprobenfunktion $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

heißt Punktschätzung für μ

$\bar{x}$: konkrete Punktschätzung

- Eigenschaften dieser Punktschätzung:
  - $E(\bar{X}) = \mu =$ gesuchter Parameter

    Eine Punktschätzung mit dieser Eigenschaft heißt erwartungstreu.

  - $var(\bar{X}) \to 0$ $(n \to \infty)$. Eine Punktschätzung mit dieser Eigenschaft heißt konsistent.
  - Für große n ist $\bar{X}$ näherungsweise $N(\mu, \sigma^2/n)$–verteilt (Zentraler Grenzwertsatz).

- Die empirische Varianz $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ist eine erwartungstreue und konsistente Punktschätzung für $var(X)$.
- Die relative Häufigkeit des Eintretens eines Ereignisses A bei n unabhängigen Versuchen ist eine erwartungstreue und konsistente Punktschätzung für $P(A)$.

Problem: $P(\bar{X} = \mu) = 0$ (bei stet. ZV)

Ausweg: Intervallschätzungen

5.3. Intervallschätzungen

Wieder: γ ein zu schätzender Parameter

Ziel: Berechnung eines **Intervalls** I aus den Werten der konkreten SP so, dass einerseits

- I möglichst klein, aber andererseits

- I die Zahl γ mit großer Wahrscheinlichkeit enthält.

Intervallschätzung: $I(X_1, \ldots, X_n) = [A, B]$ Intervall,

Konfidenzschätzung, Konfidenzintervall

A und B sind Stichprobenfunktionen. Für eine konkrete SP $x_1, \ldots, x_n$ erhält man ein konkretes Intervall

$$[\, a, b\,] \;=\; [\, A(x_1, \ldots, x_n)\,,\; B(x_1, \ldots, x_n)\,]\,.$$

Vorgehen am Beispiel eines Konfidenzintervalls für den Erwartungswert:

- Modellannahme:
  X sei normalverteilt mit
  - unbekanntem Erwartungswert μ und
  - bekannter Standardabweichung σ.

- Vorgabe der Wahrscheinlichkeit $P(\mu \in I) \;=\; \varepsilon$
  $=$ Konfidenzniveau $(0 < \varepsilon < 1)$

  also I so, dass es den unbekannten Parameter mit Wkt. ε überdeckt;

  Die Zahl μ ist fest aber unbekannt. Zufällig ist das Intervall I, das von der Stichprobe abhängt.

- Sei $\sigma = 2$, $n = 36$, $\bar{x} = 10$, $\varepsilon = 0,95$
- Konstruktion eines (möglichst kleinen) Intervalls I, mit $P(\mu \in I) = 0,95$:

  $\bar{x}$ ist eine erwartungstreue Punktschätzung für μ, und $\bar{X}$ ist

  $N(\mu, \frac{\sigma^2}{n}) = N(\mu, \frac{1}{9})$ – verteilt.

  (Wegen $P\,(\,\mu\,<\,\bar{X}\,) = P\,(\,\mu\,>\,\bar{X}\,)$) sinnvoll:

  symmetrisches Intervall um $\bar{x}$: $[a, b] = [\bar{x} - c, \bar{x} + c]$.

  Wie groß muss c sein?

$$\begin{aligned} 0,95 &= P\,(\,\bar{X} - c \,\leq\, \mu \,\leq\, \bar{X} + c\,) \\ &= P\,(\,-c \,\leq\, \mu - \bar{X} \,\leq\, c\,) \\ &= P\,(\,-c \,\leq\, \bar{X} - \mu \,\leq\, c\,) \\ &= P\left(\,-\frac{c}{1/3} \leq \frac{\bar{X} - \mu}{1/3} \,\leq\, \frac{c}{1/3}\,\right) \\ &= 2\,\Phi(3c) \;-\; 1 \end{aligned}$$

Also:

$$\Phi(3c) \;=\; \frac{1 + 0,95}{2} \;=\; 0,975 \qquad \Rightarrow \quad 3c \;=\; z_{0,975} = 1,96$$

$c = 1,96/3 \approx 0,6533\ldots$

$$I \;=\; [10 - 0,653\ldots\,,\; 10 + 0,653\ldots] \;\subset\; [9,346\,;\; 10,654]$$

(richtig runden!)

Allgemein: Konfidenzintervall für μ bei bekanntem σ,
$X \sim N(\mu,\, \sigma^2)$

Bezeichnung: $\alpha = 1 - \varepsilon$ ... Irrtumswahrscheinlichkeit

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \,,\; \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Beispiel:

konkrete Stichprobe $(2,\ 3,\ 1,\ 0)$ vom Umfang $n = 4$, $(\bar{x} = 1,5)$

Es wird angenommen, dass diese Stichprobe einer normalverteilten Grundgesamtheit mit der bekannten Standardabweichung $\sigma = 1,2$ entstammt.

Sei $\alpha = 0,05$

Tafel: $z_{0,975} = 1,96$

konkretes Konfidenzintervall:

$$= \left[1,5 - 1,96 \cdot \frac{1,2}{\sqrt{4}} \,,\; 1,5 + 1,96 \cdot \frac{1,2}{\sqrt{4}} \right] = [\,0,324\,,\; 2,676\,]$$

Interpretation: Unter den oben getroffenen Annahmen (Normalverteilung, $\sigma = 1.2$) überdeckt das Intervall $[\,0,324\,,\,2,676\,]$ mit 95%-iger "Sicherheit" den unbekannten Erwartungswert μ.

Was heißt das?

Im konkreten Fall wird μ entweder überdeckt, oder es wird nicht überdeckt!

ε ist die Wahrscheinlichkeit dafür, dass die zufällige Stichprobenauswahl zu einer Stichprobe führt, so dass das aus dieser Stichprobe berechnete Intervall den wahren Wert des Parameters enthält.

Wird also das Verfahren der Intervallschätzung sehr oft wiederholt, dann erhält man in etwa $\varepsilon \cdot 100\%$ der Fälle ein Schätzintervall, das den gesuchten Parameter enthält.

Problem: Wahl von ε

Diese Entscheidung muss vom Anwender getroffen werden.
Üblich sind folgende Niveaus: $\varepsilon = 0,90,\ = 0,95,\ = 0,99$
je nach Problemstellung.

Bemerkungen:

- Mit wachsendem Stichprobenumfang n (und festem σ) wird das Konfidenzintervall kleiner.
- Für große Standardabweichung σ (und feste n) ist das Konfidenzintervall größer.
- Mit wachsendem ε (Sicherheitsbedürfnis) wird das Konfidenzintervall größer (für $\varepsilon = 1$ unsinnig).

Wichtige Verteilungen in der Statistik

Verteilung	Beispiel	(Vor.: Normalvert. Grundgesamtheiten)
t-Vert.	$\dfrac{\bar{X} - \mu}{\sqrt{\dfrac{S^2}{n}}}$	"Quotient aus unabh. NV-ZV und χ^2-verteilter ZV"
χ^2-Vert.	S^2	"Quadrate normalvert. ZV"
F-Vert.	$\dfrac{{S_1}^2}{{S_2}^2}$	"Quotienten unabh. χ^2-vert. ZV"

Beispiel:

Stichprobe aus normalverteilter Grundgesamtheit:

i	1	2	3	4	5	6	7	8	9	10
x_i	3,2	2,1	-1,7	-4,7	-1,0	-4,2	1,1	3,8	-3,1	-0,8

i	11	12	13	14	15	16	17	18	19	20
x_i	-2,6	3,6	-1,5	0,5	-2,5	2,8	-2,8	1,1	-1,8	3,7

Es gilt: $n = 20,\ \bar{x} = -0,24,\ s^2 = 7,65,\ s = 2,765$

Konfidenzintervall für μ bei unbekannter Varianz σ^2:

Konfidenzniveau sei $\varepsilon = 0,95 \quad \Rightarrow\ \alpha = 0,05$

Tafel:

$$t_{n-1,1-\frac{\alpha}{2}} \ = \ t_{19,\ 0,975} \ = \ 2,09$$

Konkretes Konfidenzintervall:

$$\left[\ \bar{x} \ - \ t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\ ,\ \bar{x} \ + \ t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\ \right]$$

$$= \ \left[\ -0,24 \ - \ 2,09 \cdot \frac{2,765}{\sqrt{20}}\ ,\ -0,24 \ + \ 2,09 \cdot \frac{2,765}{\sqrt{20}}\ \right]$$

$$= \ [\ -1,53\ ,\ +1,05\ ]$$

Bemerkungen:

- Voraussetzung für die Konstruktion eines Konfidenzintervalls ist die Kenntnis der Verteilung der zugrundeliegenden Stichprobenfunktion.
  Dazu müssen Verteilungsannahmen getroffen werden.

  Bisher stets: NV

- Unter sehr allgemeinen Voraussetzungen ist das Stichprobenmittel $\bar{X}$ asymptotisch normalverteilt, d.h. die Verteilung nähert sich mit wachsendem n einer Normalverteilung an (Zentraler Grenzwertsatz). Bereits bei $n \geq 30$ kann im allgemeinen von einer "hinreichend guten" Approximation ausgegangen werden.

- Konfidenzintervalle müssen nicht notwendig zweiseitig sein. Soll ε z.B. die Sicherheit sein, mit der ein Parameter eine angegebene Grenze g nicht überschreitet (unterschreitet), wird man Konfidenzintervalle der Form $(-\infty, g]$
  (bzw. $[g, \infty)$) bestimmen.