Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

Kapitel III - Grundannahmen der schließenden Statistik

Induktive Statistik

Prof. Dr. W.-D. Heller
Hartwig Senska
Carlo Siebenschuh

- In der schließenden Statistik erhalten wir wie in dem Beispiel der Warenpartie die Information durch die Auswertung einer Stichprobe.
- Damit ist die Menge aller möglichen Informationen ("Signale") gegeben durch die Menge aller möglichen Stichprobenergebnisse.

Stichproben (mit Zurücklegen) haben wir beschrieben als Wiederholungen eines Experiments, dessen Ausgang wir mittels eines Messwertes beobachten.

Das Experiment besteht aus einer Zufallsvariablen auf einem Wahrscheinlichkeitsraum.

Damit unterstellen wir, dass der relevante Umweltzustand mit diesem Experiment zusammenhängt. In der schließenden Statistik nimmt man an, dass der Zusammenhang zwischen dem relevanten Umweltzustand und der Zufallsvariablen umkehrbar eindeutig ist.

Der für die Entscheidung wesentliche Umweltzustand kann durch eine
Zufallsvariable beschrieben werden.

Beispiele:

1. *Kontrolle einer Warenpartie*

   Relevanter Umweltzustand:

      Ausschussanteil p einer Warenpartie

Beschreibung der Zufallsvariable durch zufällige Entnahme einer Wareneinheit ω

formal:

$$X(\omega) = \begin{cases} 0 & \omega \text{ gut} \\ \\ 1 & \omega \text{ schlecht} \end{cases}$$

$P(X = 1) = p, P(X = 0) = 1 - p$

X Bernoulli-verteilt mit Parameter p (hier: unbekannt)

2. *Abfüllmaschine* - Verteilung unbekannt

Ein Probelauf mit 300 Einheiten hat eine Häufigkeitstabelle entprechend folgender Tabelle erbracht. Treten an der Anlage keine
Änderungen auf, kann davon ausgegangen werden, dass jede weitere
produzierte Einheit eine Füllmenge aufweist, wie eine zufällig herausgegriffene Einheit aus den Einheiten des Probelaufs.

Die Abbildung suggeriert, als Vereinfachung (Idealisierung, Näherung) der Wahrscheinlichkeitsverteilung von einer Normalverteilung
auszugehen, über deren Parameter Unklarheit besteht. Eine Rechtfertigung dafür liefert auch der Zentrale Grenzwertsatz.

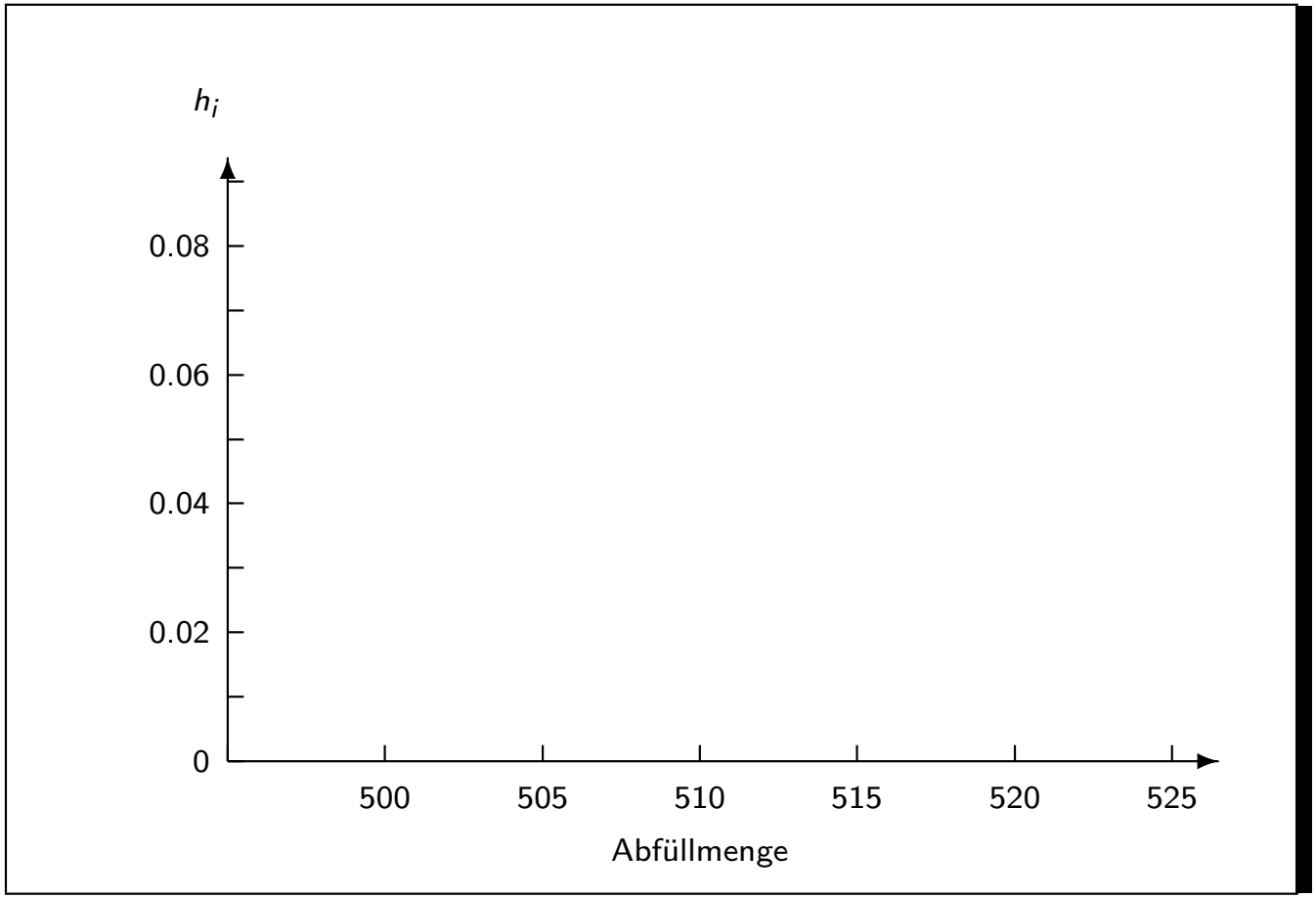
Füllmenge (in cm^3)			i	h_i	$100p_i$
495	bis unter	496		1	0.33
496	"	497	2	0	0
497	"	498	3	1	0.33
498	"	499	4	2	0.67
499	"	500	5	1	0.33
500	"	501	6	3	1.00
501	"	502	7	6	2.00
502	"	503	8	8	2.67
503	"	504	9	10	3.33
504	"	505	10	13	4.33
505	"	506	11	15	5.00
506	"	507	12	19	6.33
507	"	508	13	21	7.00
508	"	509	14	23	7.67
509	"	510	15	24	8.00
			$\sum_1$	147	

Füllmenge (in cm^3)			i	h_i	$100p_i$
510	bis unter	511	16	25	8.33
511	"	512	17	22	7.33
512	"	513	18	22	7.33
513	"	514	19	19	6.33
514	"	515	20	16	5.33
515	"	516	21	12	4.00
516	"	517	22	10	3.33
517	"	518	23	8	2.67
518	"	519	24	7	2.33
519	"	520	25	3	1.00
520	"	521	26	4	1.33
521	"	522	27	2	0.67
522	"	523	28	1	0.33
523	"	524	29	1	0.33
524	"	525	30	1	0.33
			$\sum_2$	153	
			$\sum_1 + \sum_2$	300	

h_i : absolute Häufigkeit von Beobachtungen in Klasse i

$p_i = \frac{h_i}{n}$: relative Häufigkeit von Beobachtungen in Klasse i

$n = 300$: Stichprobenumfang (Gesamtzahl der Beobachtungen)


h_i
0.08
0.06
0.04
0.02
0
500
505
510
515
520
525
Abfüllmenge

In beiden Beispielen liegt eine Wahrscheinlichkeitsverteilung vor, die zu einer bestimmten Klasse von Wahrscheinlichkeitsverteilungen gehört,

1. die Klasse der Bernoulli-Verteilungen bei 1.
2. die Klasse der Normalverteilungen bei 2.

Für die den Umweltzustand beschreibende Zufallsvariable Y kann eine
Klasse von Wahrscheinlichkeitsverteilungen angegeben werden, der die
Wahrscheinlichkeitsverteilung von Y angehört.

Ermittlung der Verteilungsannahme aus:

- theoretischen Überlegungen (siehe Bsp. 1)
- dem Datenmaterial und Plausibilitätsbetrachtungen (siehe Bsp. 2)

Konflikt:

Klasse klein: Gefahr der Fehlspezifikation, falls wahre Verteilung nicht vorhanden, Aufgabenstellung leichter zu handhaben, präzisere Aussagen möglich.

Klasse groß: Geringe Gefahr der Fehlspezifikation, Aufgabenstellung schwieriger zu bearbeiten, nur ungenaue Aussagen möglich.

Art der Verteilungsannahme meist:
Festlegung eines Verteilungstyps mit nicht festgelegtem Parameter

1. Bernoulli-Verteilung mit Parameter $p \in [0,1]$(s.o.)

2. Binomialverteilung $B(n,p)$ mit Parameter $p \in [0,1]$

3. Poissonverteilung $Poi(\lambda)$ mit Parameter $\lambda \in \mathbb{R}$ und $\lambda > 0$
   (Ankunftsverhalten von Kunden, Fehlerzahlen, radioaktive Pozesse,...)

4. Exponentialverteilung mit Parameter $\lambda, \lambda \in \mathbb{R}$ und $\lambda > 0$
   (Lebensdauer, Zeitdauer zwischen Ereignissen)

5. Normalverteilung mit

   a. Parameterpaar (μ, σ^2), $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$
   b. Parameter μ, $\mu \in \mathbb{R}$ (σ^2 wird als feste Größe angesehen, u.U. ist σ^2 auch bekannt.)
   c. Parameter σ^2, $\sigma^2 \in \mathbb{R}$ und $\sigma^2 > 0$ (hier ist μ fest bzw. bekannt)

6. Gleichverteilung auf dem Intervall $[a, b]$ mit den Parameter(n) $a \in \mathbb{R}$ und/oder $b \in \mathbb{R}$ mit $a < b$

zu 1.: $[0,1]$ bzw. $\{\frac{k}{N}|k = 0, 1, \ \ldots\ , N\}$ bei N unabhängigen Versuchen

zu 2.: wie bei 1.

zu 3.: $\mathbb{R}_+ = \{\lambda \in \mathbb{R},\ \lambda > 0\}$

zu 4.: " "

zu 5.:

a) $\mathbb{R} \times \mathbb{R}_+$

b) $\mathbb{R}$

c) $\mathbb{R}_+$

zu 6.: $\{(a, b)|a, b \in \mathbb{R}, a < b\}$

Parametrische Verteilungsannahme:

Die Wahrscheinlichkeitsverteilung variiert mit dem Parameter.
Für den Parameter gibt es eine Menge von möglichen Werten,
den Parameterraum Γ.
Klasse der Wahrscheinlichkeitsverteilungen: **Verteilungsannahme**

$W = \{P_\gamma | \gamma \in \Gamma\}$ mit $P_\gamma \neq P_{\tilde{\gamma}}$ für $\gamma \neq \tilde{\gamma}$
mit P_γ Wahrscheinlichkeitsverteilung in W für alle γ
W heißt *parametrische Verteilungsannahme mit Parameterraum* Γ.

Nichtparametrische Verteilungsannahme:

Allgemeingehaltene Klasse von Wahrscheinlichkeitsverteilungen ohne Einschränkung auf den Verteilungstyp

Beispiel:

- Klasse aller stetigen Verteilungen
- Klasse aller diskreten Verteilungen auf der Menge der natürlichen Zahlen
- ...

Zustandsraum des Entscheidungsproblems:
Klasse der Wahrscheinlichkeitsverteilungen = Verteilungsannahme

Akionenraum des Entscheidungsproblems:
Menge der Entscheidungsmöglichkeiten, die direkt oder indirekt die Verteilungen der Verteilungsannahme betreffen.
Bei einer parametrischen Verteilungsannahme betreffen die möglichen Entscheidungen also die Parameter des Parameterraums.

Schadensfunktion als Konsequenz der Handlung bezüglich des Zustands

$$S : A \times W \longrightarrow \mathbb{R}$$

1. Kontrolle der Warenpartie
   Aktionenraum $= \{\text{Annahme, Ablehnung}\} = \{\text{“}p \leq p_0\text{“}, \text{“}p > p_0\text{“}\}$
   p tatsächlicher Ausschussanteil, p_0 tolerierbarer Ausschussanteil
2. Abfüllanlage mit σ^2 fest, bekannt
   Verteilungsannahme: $W = \{N(\mu, \sigma^2) | \mu \in \mathbb{R}\}$
   a. Aufgabe: Schätzen des Parameters
      Aktionenraum = Parameterraum $= \mathbb{R}$
   b. Aufgabe: Testen, ob Sollwert μ_0 eingehalten
      Aktionenraum $= \{\text{ja, nein}\} = \{\text{“}\mu = \mu_0\text{“}, \text{“}\mu \neq \mu_0\text{“}\}$

Hinweis: Da im Beispiel 2a Füllmengen betrachtet werden, ist der Parameterraum $\mathbb{R}_{>0}$.

Zur Bewertung der Entscheidung im Beispiel 2a) könnte man etwa die sogenannte quadratische Schadensfunktion zugrundelegen:

$$S(\hat{\mu}, N(\mu, \sigma^2)) = (\hat{\mu} - \mu)^2 \text{ für alle } \mu, \hat{\mu} \in \mathbb{R}$$

$\hat{\mu}$ Schätzwert für μ (Aktion), μ wahrer Wert (Zustand).

Bezüglich der Minimaxregel sind dann alle Entscheidungen äquivalent:

$$\sup_{N(\mu,\sigma^2)\in W} S(\hat{\mu}, N(\mu, \sigma^2)) = \sup_{\mu\in\mathbb{R}} (\hat{\mu} - \mu)^2 = +\infty$$

Damit zeigt sich, dass die Minimaxregel nichts zur Lösung des Problems beiträgt.

$\Longrightarrow$ Ansatz: mittels Stichproben Information gewinnen

Machen wir nun allgemein n Beobachtungen der Zufallsvariablen Y:

$$x_1, \ldots, x_n, \quad x_i \quad i\text{-te Beobachtung},$$

so können wir

$$(x_1, \ldots, x_n)$$

als Realisation eines Zufallsvektors

$$(X_1, \ldots, X_n)$$

auffassen.

Es muss bekannt sein, wie sich die gemeinsame Wahrscheinlichkeitsverteilung des Vektors $(X_1, \ldots, X_n)$ aus der Verteilung von Y ergibt.

W sei eine Verteilungsannahme. Es wird angenommen, dass Realisationen $x_1, \ldots, x_n$ von Zufallsvariablen $X_1, \ldots, X_n$ beobachtet werden können, deren gemeinsame Wahrscheinlichkeitsverteilung von der Verteilung der Zufallsvariablen Y_z des wahren Zustands in vollständig bekannter Weise abhängt.

Bezeichnung

Man bezeichnet den Vektor $(X_1, \ldots, X_n)$ als *Stichprobe vom Umfang n zu Y* und $(x_1, \ldots, x_n)$ als eine *Stichprobenrealisation*.

Bezeichnung

Der wahre Zustand z ist nicht bekannt. Bekannt ist der Zusammenhang zwischen der Wahrscheinlichkeitsverteilung von $X_1, \ldots, X_n$ und dem Zustand z für jedes z.
D.h. für ein bestimmtes z hat $X_1, \ldots, X_n$ eine bestimmte Verteilung.

$X_1, \ldots, X_n$ unabhängig und jede von ihnen hat dieselbe Verteilungsfunktion wie Y. Damit:

$$F_X^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_Y(x_i)$$

Daraus folgt für diskretes Y:

$$P^{(n)}(X = x) = \prod_{i=1}^{n} P(Y = x_i)$$

und für stetiges Y:

$$f_X^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_Y(x_i).$$

Hinweis: Die drei Gleichungen gelten jeweils für alle $x \in \mathbb{R}^n$, wobei $x \stackrel{def}{=} (x_1, ..., x_n)$ eine Realisation des Zufallsvektors $X \stackrel{def}{=} (X_1, ..., X_n)$ darstellt.

Bei einer parametrischen Verteilungsannahme ist F_Y durch den Typ und den (wahren aber unbekannten) Parameter γ festgelegt. Damit ergibt sich auch F_X aus dem Typ der Verteilung und dem Parameter γ.

1. Parameterpunktschätzungen:

Aus der Stichprobenrealisation $(x_1, \ldots, x_n)$ ist ein Schätzwert für den Parameter γ der Verteilung des wahren Umweltzustands zu ermitteln. Gesucht ist also eine "optimale Schätzfunktion"

$$\delta : \mathcal{X} \to \Gamma$$

mit $\delta(x_1, \ldots, x_n)$: Schätzwert für den wahren Parameter bei der Stichprobenrealisation $(x_1, \ldots, x_n)$,
$\mathcal{X}$ sei dabei die Menge aller möglichen Stichprobenrealisationen, der *Stichprobenraum*.

2. Parameterbereichsschätzungen:

Abhängig von der Stichprobenrealisation $(x_1, \ldots, x_n)$ ist eine Teilmenge, ein "Bereich", des Parameterraums Γ anzugeben. Gesucht ist also eine Funktion $\delta : \mathcal{X} \to P(\Gamma)$ mit

$\delta(x_1, \ldots, x_n)$: geschätzter Bereich für den wahren Parameter.

Üblicherweise wird der geschätzte Bereich bei einparametrigen Verteilungsannahmen ($\Gamma \subset \mathbb{R}$) ein Intervall sein:

$\delta(x_1, \ldots, x_n)$ liefert untere $(\delta_1(x))$ und obere Intervallgrenze $(\delta_2(x))$.

Bei *Konfidenz-* bzw. *Vertrauensintervallen* fordert man von dem Zufallsintervall $[\delta_1(x), \delta_2(x)]$:
wahren Parameterwert mit vorgegebener Wahrscheinlichkeit beinhalten, d.h.
die zufälligen (von dem Stichprobenergebnis abhängigen) Intervallgrenzen sind so anzugeben, dass

$$P(\delta_1(X) \leq \gamma \leq \delta_2(X)) = \alpha$$

wobei

1.) γ der wahre Parameter

2.) X der Zufallsvektor der Stichprobe

3.) α die vorgegebene Überdeckungswahrscheinlichkeit.

3. Tests:

Entscheidung zwischen zwei alternativen, sich gegenseitig ausschließenden “Hypothesen“.
Die Hypothesenbildung entspricht einer Aufteilung des Parameterraums in zwei disjunkte Teilmengen:

$$\Gamma = \Gamma_0 \cup \Gamma_1, \quad \Gamma_0 \cap \Gamma_1 = \emptyset, \quad \Gamma_0, \Gamma_1 \neq \emptyset$$

Die Hypothesen lauten dann:

$$\begin{aligned} H_0 &: \gamma \in \Gamma_0 \quad (\text{“Nullhypothese“}) \\ H_1 &: \gamma \in \Gamma_1 \quad (\text{“Gegenhypothese“}) \end{aligned}$$

Gesucht wird eine Entscheidungsfunktion

$$\delta : \mathcal{X} \to \{d_0, d_1\},$$

wobei

$$\begin{aligned} d_0 &: H_0 \quad (\gamma \in \Gamma_0) \text{ wird als richtig betrachtet} \\ d_1 &: H_1 \quad (\gamma \in \Gamma_1) \text{ wird als richtig betrachtet} \end{aligned}$$

bezeichne. Die Entscheidungsfunktion gibt damit an, für welche Hypothese man sich beim Vorliegen des Stichprobenergebnisses $(x_1, \ldots, x_n)$ entscheidet.

Zusammenfassung der Entscheidungssituationen der schließenden Statistik:

- Zustandsraum ist der Parameterraum Γ.
- Aktionenraum ist die Menge der möglichen Entscheidungen, wie sie etwa in den drei Aufgabengebieten angegeben wurden.
- $S : A \times \Gamma \to \mathbb{R}$ sei die ermittelte Schadensfunktion des Problems.
- $\mathcal{X}$ sei der Stichprobenraum, d.h. die Menge der möglichen Stichprobenergebnisse.

Aus der Menge Δ der Entscheidungsfunktionen

$$\delta : \mathcal{X} \to A$$

ist eine unter Berücksichtigung der Schadensfunktion am besten geeignete herauszufinden.

- Ergebnis der Stichprobe als Information betrachten,
- Entscheidungssituation bei Information,
- Man kann das Erwartungswertprinzip benutzen, also die Risikofunktion bilden.

Risikofunktionen:

Zu $\delta \in \Delta$ sei

$$R(\delta, \gamma) = E(S(\delta(X), \gamma)).$$

Erwarteter Schaden bei der Entscheidungsfunktion δ und dem wahren Parameter (Zustand) γ.

Man erhält dann eine Entscheidungssituation

$$(\Delta, \Gamma, R).$$