Statistik I

Inhalt

Statistik	Aufgaben	Techniken
Deskriptive Statistik	Beschreibung, graphische Darstellung und Validierung von Daten. ❗Keine Rückschlüsse auf Grundgesamtheit möglich.	Grafiken, Tabellen, Kennzahlen
Explorative Statistik	Suche nach Struktur in den Daten (ohne stochastische Methoden). Formulierung von Hypothesen für das den Daten zugrunde liegende stochastische Modell.	Iterative und interaktive Anwendung von Techniken aus der deskriptiven und induktiven Statistik.
Induktive Statistik	Ziehung von Schlüssen von den Daten (Stichprobe) auf Grundgesamtheit. Basierend auf stochastischen Modellen.	Statistische Modellierung, statistische Tests, Konfidenzintervalle, Schätzer

Datenerhebung & Messung

Ein Merkmal

Matrikelnummer	Name	Vorname	Geburtsdatum	Hauptfach	Nebenfach
xxxxx 234	Muster	Peter	01.01.2001	Statistik	Informatik
xxxxx 556	Schmid	Lena	31.10.2002	Informatik	Statistik
xxxxx 123	Müller	Jonas	27.08.1999	Mathematik	NA
xxxxx 767	Nguyen	Cho	24.12.2000	Medizin	Soziologie
xxxxx 111	Nagel	Cosima	26.10.1996	Jura	Ethik

} Alle Merkmale

Merkmalsausprägung von Merkmal "Nebenfach" bei der zweiten statistischen Einheit.

Eine Beobachtung

Grundgesamtheit: Studenten der LMU (über welche "Objekte" erhebe ich Daten?)

Stichprobe: z.B. Alle Statistik Studenten ❗Stichproben müssen nicht per Definition zufällig gewählt sein.

Statistische Einheit / Untersuchungseinheit (UE): Ein Student bzw. ein Element der Grundgesamtheit

Merkmal: Messbare Eigenschaft einer statistischen Einheit. In der Tabelle quasi der (sinnvolle) Spaltenname. z.B. Hauptfach ist ein Merkmal

Merkmalsausprägung: Der tatsächliche Wert des Merkmals bei einer statistischen Einheit. In der Tabelle ist das ein Wert in einer Zelle.

Beobachtung: Alle Merkmalsausprägungen einer statistischen Einheit zu einem Zeitpunkt. In der Tabelle sind das die Werte in einer Zeile.

Unterscheidung nach ...		
... Quantifizierbarkeit der Ausprägungen	**Qualitative Merkmale**: • nur zuordenbar (einstufig) • Beispiele: Wohnort, Name	**Quantitative Merkmale**: • mess- oder zählbar • Beispiele: Alter, Körpergröße
... Anzahl der Ausprägungen*	**Diskrete Merkmale**: • höchstens abzählbar unendlich viele mögliche Ausprägungen • Beispiele: Gehaltsklassen, Kaufverhalten	**Stetige Merkmale**: • überabzählbar unendlich viele mögliche Ausprägungen • Beispiele: Geschwindigkeit, Gewicht
... Direktheit der Informationsgewinnung	**Beobachtbare Merkmale**: • können direkt erhoben werden • Beispiel: Abiturnote	**Latente Merkmale**: • Operationalisierung über Indikatoren/Items notwendig • Beispiele: Bildungsgrad, Kreativität, Nutzen

(*) Merkmale die eigentlich diskret sind, aber so viele Ausprägungen haben, dass sie wie stetige Merkmale behandelt werden können, nennt man auch quasi-stetig (z.B. Einkommen)

(*) Stetige Merkmale können durch Klassenbildung in diskrete Merkmal umgewandelt werden.

Skalenniveaus

Skalenniveau	Beispiele	Erlaubte Transformationen um Strukturen zu erhalten	natürliche Ordnung	sinnvolle Abstände	natürliche Null	natürliche Einheit	Berechenbare Kennzahlen
Nominalskala	Wohnort, Farbe	Bijektionen	✗	✗	✗	✗	Mode
Ordinal - Rangskala	Noten, Michelin-Sterne Platzierung bei Sportevent	str. monoton steig. Abb.	✓	✗	✗	✗	Median
Intervallskala	Temperatur in C° Jahreszahlen	affin lin. str. mon. steig. Abb.	✓	✓	✗	✗	Arithm. Mittel
Verhältnisskala	Preis, Länge, Gewicht, Temp. in K°	lineare str. mon. steig. Abb.	✓	✓	✓	✗	Geom. Mittel Harm. Mittel
Absolutskala	Häufigkeit, Anzahl, Prozentpunkte	Identität	✓	✓	✓	✓	Alle

Datenerhebung

Methoden:

Beobachtung
Datengewinnung durch Erfassen von ungesteuerten Sachverhalten

Befragung
Fragebögen für mündliche/schriftliche/online Umfrage.

Experiment
Erzeugung der Daten durch Simulation von Situationen.

Umfang:

Vollerhebung
Alle stat. Einheiten einer GG werden untersucht.

Stichprobe (Teilerhebung)
Ein Teil der UE in einer GG wird untersucht.

Datenform:

Querschnittdaten
Eine Beobachtung pro UE.

- Noten, Aktivitäten, Geschlecht, können zu bestimmtem Zeitpunkt von UE erhoben werden und z.B. mittels Regression auf Zusammenhänge untersucht werden.

Zeitreihe
Mehrere Beobachtungen einer UE.

- Temperatur, Wind & Luftfeuchtigkeit werden in regelmässigen Abständen gemessen um Prognosen über die zeitliche Entwicklung der UE 'Wetter' zu machen

Längsschnittdaten
Mehrere Beobachtungen mehrerer UE.

- Kohortenstudien in Medizin
- Mikrozensus

Wahrscheinlichkeitsrechnung

Elementarereignisse

Für eine Grundmenge Ω wird die ein-elementige Teilmenge $\{\omega\} \subseteq \Omega$ als Elementarereignis bezeichnet

Ereignisse

Für eine Grundmenge Ω wird $A \subseteq \Omega$ als Ereignis bezeichnet.

Laplace - Wahrscheinlichkeit

Für eine abzählbare Grundmenge Ω und ein Ereignis $A \subseteq \Omega$ ist die Laplace-Wahrscheinlichkeit $\mathbb{P}(A) := \frac{|A|}{|\Omega|}$

Wahrscheinlichkeitsverteilung (Axiome von Kolmogorov) (vereinfacht)

Sei Ω eine Grundmenge und P ein Funktion auf $\mathcal{P}(\Omega)$. P heißt Wahrscheinlichkeitsverteilung oder Wahrscheinlichkeitsmaß auf Ω, wenn sie folgende Eigenschaft erfüllt: 1) $P(\Omega) = 1$ 2) $\forall A \subseteq \Omega: P(A) \geq 0$ 3) $\forall A,B \subseteq \Omega: A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Folgerungen

Korollar

- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- $B \subseteq A \Rightarrow P(A \setminus B) = P(A) - P(B)$
- $B \subseteq A \Rightarrow P(B) \leq P(A)$
- $P[\bigcup_{i=1}^n A_i] = \sum_{k=1}^n (-1)^{k+1} \cdot \sum_{1 \leq i_1 < \dots < i_k \leq n} P[A_{i_1} \cap \dots \cap A_{i_k}]$ ← Siebformel von Sylvester-Poincaré

$$= \sum_{i=1}^n P[A_i] - \sum_{1 \leq i < j \leq n} P[A_i \cap A_j] + \sum_{1 \leq i < j < k \leq n} P[A_i \cap A_j \cap A_k] - \dots + (-1)^{n+1} \cdot P[\bigcap_{i=1}^n A_i]$$

Spezialfall: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit von A gegeben B für Ereignisse $A, B \subseteq \Omega$ mit $P(B) > 0$ ist $P(A|B) := \frac{P(A \cap B)}{P(B)}$

Satz

- $P[\overset{\neq\emptyset}{\bigcap_{i=1}^n A_i}] = \prod_{i=1}^n P[A_i \mid \bigcap_{j=1}^{i-1} A_j] = P[A_1] \cdot P[A_2|A_1] \cdot P[A_3|A_1 \cap A_2] \cdot \ldots \cdot P[A_n|A_1 \cap \dots \cap A_{n-1}]$ (Multiplikationssatz)
- Sei $(A_i)_{i \in I}$ eine disjunkte Zerlegung von Ω. D.h. $\Omega = \dot{\bigcup}_{i \in I} A_i$. Dann gilt für beliebiges B
  $P[B] = \sum_{i: P[A_i] > 0} P[B|A_i] \cdot P[A_i]$ (Satz von totaler Wahrscheinlichkeit)
  Spezialfall: $P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

Kombinatorik	mit Wiederholung/ mit Zurücklegen	ohne Wiederholung/ ohne Zurücklegen
Anzahl Kombinationen ohne Reihenfolge	$\binom{n}{m}$	$\binom{n+m-1}{m}$
Anzahl Kombinationen mit Reihenfolge	$\frac{n!}{(n-m)!}$	n^m
Anzahl Permutationen	$n!$	$\frac{n!}{n_1! \cdot \ldots \cdot n_s!}$

Stochastische Unabhängigkeit

Eine Kollektion von Ereignissen $(A_i)_{i \in I}$ heisst (stochastisch) unabhängig, wenn für jede endliche Kollektion $J \subseteq I$ gilt: $P[\bigcap_{i \in J} A_i] = \prod_{i \in J} P[A_i]$

- $A \perp B \Leftrightarrow A, B$ stochastisch unabhängig
- Paarweise Unabhängig $\not\Rightarrow$ Unabhängigkeit
- $A \perp B \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$

Satz

Die Ereignisse $(A_i)_{i \in I}$ seien unabhängig. Für jedes i sei $B_i = A_i \lor B_i = \bar{A}_i$. Dann sind die Ereignisse $(B_i)_{i \in I}$ unabhängig.

Satz von Bayes

$(A_i)_{i \in I}$ sei so, dass $\Omega = \dot{\bigcup}_{i \in I} A_i$. B sei so, dass $P[B] \neq 0$.

Dann ist $P[A_i|B] = \frac{P[B|A_i] \cdot P[A_i]}{\sum_j P[B|A_j] \cdot P[A_j]}$ Spezialfall: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$

Wettverhältnis (Odds-Update)

$$\underbrace{\frac{P[B|A]}{P[B^c|A]}}_{\text{a-posteriori Verhältnis}} = \overbrace{\frac{P[A|B]}{P[A|B^c]}}^{\text{Faktor der neuen Information}} \cdot \underbrace{\frac{P[B]}{P[B^c]}}_{\text{a-priori-Verhältnis}}$$