

Statistik I

Inhalt

Einführung

Kennwerte & Verteilungseigenschaften

Datenerhebung und Messung

Wichtige parametrische Verteilungen

**Wahrscheinlichkeitsrechnung: Grundlagen
und Definitionen**

Schätzung & Grenzwertsätze

Zufallsvariablen, Verteilungen & Häufigkeiten

**Zufallsvektoren und multivariate
Verteilungen**

**Stochastische Unabhängigkeit und
Zusammenhangsmaße für diskrete Merkmale**

**Zusammenhangsmaße für metrische
Merkmale**

Statistische Grafiken

Korrelation und Kausalität

Statistik	Aufgaben	Techniken
Descriptive 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	Sozialologie
xxxxx 111	Nagel	Cosima	26.10.1996	Jura	Ethik

} Alle Merkmale

Merkmaalsausprägung vom 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 das (sinnvolle) Spaltenname. z.B. Hauptfach ist ein Merkmal

Merkmaalsausprägung: 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: <ul style="list-style-type: none"> nur zuordnbar (einstufig) Beispiele: Wohnort, Name 	Quantitative Merkmale: <ul style="list-style-type: none"> mess- oderzählbar Beispiele: Alter, Körpergröße
... Anzahl der Ausprägungen*	Diskrete Merkmale: <ul style="list-style-type: none"> höchstens abzählbar unendlich viele mögliche Ausprägungen Beispiele: Gehaltsklassen, Kaufverhalten 	Stetige Merkmale: <ul style="list-style-type: none"> überabzählbar unendlich viele mögliche Ausprägungen Beispiele: Geschwindigkeit, Gewicht
... Direktheit der Informationsgewinnung	Beobachtbare Merkmale: <ul style="list-style-type: none"> können direkt erhoben werden Beispiel: Abiturnote 	Latente Merkmale: <ul style="list-style-type: none"> 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 Merkmale umgewandelt werden.

Skalen niveaus

Skalen niveau	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 ungesteuertem 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 $\{w\} \subseteq \Omega$ 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 $P(A) := \frac{|A|}{|\Omega|}$

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) $\forall A, B \subseteq \Omega : A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Bedingte Wahrscheinlichkeit

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

Folgerungen

Korollar

- $P(\emptyset) = 0$
 - $P(\bar{A}) = 1 - P(A)$
 - $B \subseteq A \Rightarrow P(A|B) = P(A) - P(B)$
 - $B \subseteq A \Rightarrow P(B) \leq P(A)$
 - $P\left[\bigcup_{i=1}^n A_i\right] = \sum_{k=1}^n (-1)^{k+1} \cdot \sum_{1 \leq i_1 < \dots < i_k \leq n} P[A_{i_1} \cap \dots \cap A_{i_k}]$ ← Sichformel 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: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Satz

- $P\left[\bigcap_{i=1}^n A_i\right] = \prod_{i=1}^n P[A_i] \quad P[A_i] = P[A_i] \cdot P[A_2 | A_1] \cdot P[A_3 | A_1, A_2] \cdot \dots \cdot P[A_n | A_1, \dots, A_{n-1}]$ (Multiplikationssatz)
- Sei $(A_i)_{i \in I}$ eine disjunkte Zerlegung von Ω . D.h. $\Omega = \bigcup_{i \in I} A_i$. 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: $P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

Stochastische Unabhängigkeit

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

- $A \perp B \Leftrightarrow A, B$ stochastisch unabhängig
- Paarweise Unabhängig \Rightarrow Unabhängigkeit
- $A \perp B \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$

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	$n!$	n^m
Anzahl Permutationen	$n!$	$\frac{n!}{n_1! \cdot \dots \cdot n_k!}$

Satz

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

Satz von Bayes

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

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

Wettverhältnis (Odds-Update)

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

Zufallsvariablen, Verteilungen & Häufigkeiten

Zufallsvariable

Eine Zufallsvariable X ist eine Abb. $X: \Omega \rightarrow \mathbb{R}$.

$T := X(\Omega)$ nennen wir Träger von X .

Verteilungsfunktion (diskret)

$$F(x) = P(X \leq x) = \sum_{i|x_i \leq x} f(x_i) = \sum_{i|x_i \leq x} P(X=x_i)$$

Verteilungsfunktion (stetig)

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

! Stringentere Definition später in Wahrscheinlichkeitstheorie (2. Semester)

Wahrscheinlichkeitsfunktion einer diskreten ZV

$$f(x_i) = P(X=x_i) = P(\{\omega \in \Omega | X(\omega) = x_i\}) \quad \forall x_i \in \mathbb{R}$$

f heißt auch **Wahrscheinlichkeitsdichte**.

Satz:

Jede Verteilungsfunktion F erfüllt folgende Eigenschaften:

- i) Monotonie: $a \leq b \Rightarrow F(a) \leq F(b)$
- ii) Rechtsseitig: $F(a) = \lim_{h \rightarrow 0^+} F(a+h) \quad \forall a \in \mathbb{R}$
- iii) Normierung: $\lim_{a \rightarrow -\infty} F(a) = 0$ und $\lim_{a \rightarrow \infty} F(a) = 1$

Umgekehrt ist jede Funktion F mit diesen drei Eigenschaften eine Verteilungsfunktion einer ZV. D.h. $\exists X: F = F_X$.

Verteilung von X

$$\begin{aligned} P(X \in B) &:= P(X^{-1}(B)) \\ &= P(\{\omega \in \Omega | X(\omega) \in B\}) \end{aligned}$$

(diskret)

X ist eine diskrete ZV, wenn der Träger von X eine abzählbare Menge ist.

(stetig)

X ist eine stetige ZV, wenn es eine Funktion f gibt, für die gilt

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x) dx = 1$
- $\forall b \in \mathbb{R}: F(b) = \int_{-\infty}^b f(x) dx$

Indikatorfunktion

Die Indikatorfunktion dient dazu zu checken ob ein Element in einer bestimmten Menge enthalten ist.

Sei: $A \subseteq \mathbb{R}$.

$$I_A(x) = \begin{cases} 0, & \text{wenn } x \notin A \\ 1, & \text{wenn } x \in A \end{cases}$$

Andere Schreibweisen sind $\mathbb{1}_A$, \mathbb{I}_A , $I(x \in A)$

Notation & Terminologie

Bezeichnung in der Empirie	Notation/Berechnung in der Theorie
Merkmal	Zufallsvariable X, Y, \dots
Anzahl Untersuchungseinheiten	n
Merkmaalsausprägung von Merkmal X der i -ten UE.	$x_i, i \in \{1, \dots, n\}$
Rohdaten / Urliste	x_1, \dots, x_n
(evtl. geordnete) verschiedene Werte aus der Urliste	$a_1, \dots, a_k, k \leq n, a_i \in \{x_1, \dots, x_n\}$ (Eindeutige Elemente der Urliste)
relative Häufigkeit f_j	Wahrsch.fkt. bzw. Dichte $f_X(a_j)$
kum. rel. Häufigkeit $F(x)$	Verteilungsfunktion $F_X(x)$

Absolute Häufigkeit

Die absolute Häufigkeit von a_j ist die Anzahl der x_i aus der Urliste mit $x_i = a_j$
 $h(a_j) = h_j$

Relative Häufigkeit

Die relative Häufigkeit von a_j ist der Anteil der x_i an der Urliste für die gilt $x_i = a_j$. $f(a_j) = f_j = \frac{h_j}{n}$

Absolute / relative Häufigkeitsverteilung

h_1, \dots, h_k heißt absolute Häufigkeitsverteilung
 f_1, \dots, f_k heißt relative Häufigkeitsverteilung

Kumulative relative Häufigkeit/empirische Verteilungsfunktion

(Sinnvoll bei ordinal oder metrisch)

$$F(x) = (\text{Anteil UE mit } x_i \leq x) = \sum_{a_i \leq x} f(a_i) \quad (\text{ECDF})$$

- monoton wachsende Treppenfkt. mit Sprüngen bei a_1, \dots, a_k .
- Sprunghöhe f_1, \dots, f_k
- rechtsseitig stetig
- $F(x) = 0$ für $x < a_1$, $F(x) = 1$ für $x \geq a_k$