ulm university universität ulm

Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs |

1. Thema

Heutiges Thema

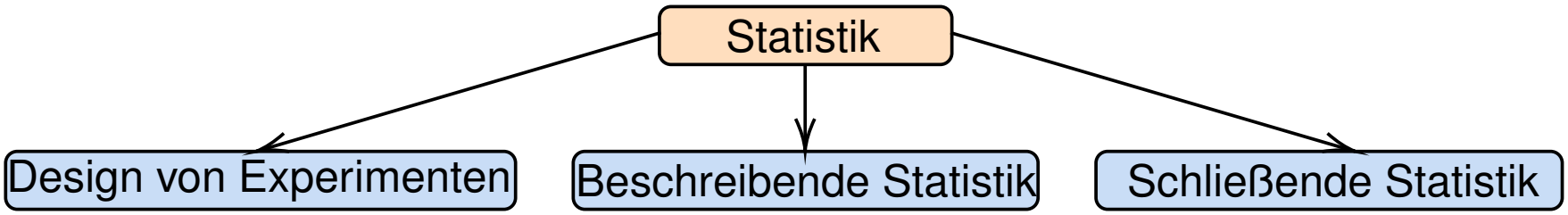
- Einführung in die Statistik

Einleitung

- Im alltäglichen Sprachgebrauch versteht man unter ”Statistik” eine Darstellung von Ergebnissen des Zusammenzählens von Daten und Fakten jeglicher Art, wie z.B. ökonomischen Kenngrößen, politischen Umfragen, Daten der Marktforschung, klinischen Studien in der Biologie und Medizin, usw.

Einleitung

- Die *mathematische Statistik* jedoch kann viel mehr.
- Sie arbeitet mit *Daten-Stichproben,* die nach einem bestimmten Zufallsmechanismus gezogen werden aus der *Grundgesamtheit* aller Daten, die in Folge von Beobachtung, Experimenten (reale Daten) oder Computersimulation (synthetische Daten) erhoben wurden.
- Dabei beschäftigt sich die mathematische Statistik mit folgenden Fragestellungen:



1. Wie sollen die Daten gewonnen werden?
   (Design von Experimenten)
2. Wie sollen (insbesondere riesengroße) Datensätze beschrieben werden, um die Gesetzmäßigkeiten und Strukturen in ihnen entecken zu können?
   (Beschreibende (deskriptive) und explorative Statistik)
3. Welche Schlüsse kann man aus den Daten ziehen?
   (Schließende oder induktive Statistik)

In dieser einführenden Vorlesung werden wir Teile der beschreibenden und schließenden Statistik kennenlernen, wobei die Datenerhebung aus Platzgründen ausgelassen wird. Die *Arbeitsweise eines Statistikers* sieht folgendermaßen aus:

(a) *Datenerhebung*

(b) *Visualisierung und beschreibende Datenanalyse*

(c) *Datenbereinigung* (z.B. Erkennung fehlerhafter Messungen, Ausreißern, usw.)

(d) *Explorative Datenanalyse* (Suche nach Gesetzmäßigkeiten)

(e) *Modellierung der Daten* mit Methoden der Stochastik

(f) *Modellanpassung* (Schätzung der Modellparameter)

(g) *Modellvalidierung* (wie gut war die Modellanpassung?)

(h) *Schließende Datenanalyse:*

- Konstruktion von *Vertrauensintervallen* (Konfidenzintervallen) für Modellparameter und deren Funktionen,
- Tests statistischer Hypothesen,
- Vorhersage von Zielgrößen (z.B. auf Basis modellbezogener Computersimulation).

Uns werden auf den folgenden Foliensätzen vor allem die Arbeitspunkte (b), (d)–(f) und (h) beschäftigen.

Beispiel 6.1.1.

Nachfolgend geben wir einige typische Fragestellungen der Statistik an Beispielen von Datensätzen an:

(a) *Statistische Herleitung von Grundsätzen der biologischen Evolution (Mendel, 1865):*

- Es wurden Nachkommen von zwei Erbsensorten, die sich in der Samenform unterscheiden, gezüchtet: die erste Sorte hat runde, die zweite kantige Erbsen.
- Johann Gregor Mendel hat festgestellt, dass sich runde Samen dominant vererben.
- Dabei werden bei einer Bestäubung von Pflanzen der einen Sorte mit Pollen der anderen alle Nachkommen runde Samen zeigen, die genetisch heterozygot sind, d.h., beide Allele aufweisen.

Beispiel 6.1.1.

- Kreuzt man diese hybriden Pflanzen, so zeigen sie runde und kantige Samen im Verhältnis $3 : 1$ (Spaltungs- und Dominanzregeln von Mendel).
- Bei der statistischen Überprüfung seiner Vermutungen erhielt Mendel 5475 runde und 1850 kantige Samen, die somit im Verhältnis $2,96 : 1$ stehen. In der folgenden Tabelle sind Ergebnisse für die ersten 10 Pflanzen gezeigt.
- Man sieht, dass das oben genannte Verhältnis zufällig um $3 : 1$ schwankt.
- Durch die Bildung des Mittels über das Gesamtkollektiv der Daten wird die Gesetzmäßigkeit $3 : 1$ gefunden (explorative Statistik).

Beispiel 6.1.1.

Pflanze	1	2	3	4	5	6	7	8	9	10
rund	45	27	24	19	32	26	88	22	28	25
kantig	12	8	7	10	11	6	24	10	6	7
Verhältnis										
. . . : 1	3,8	3,4	3,4	1,9	2,9	4,3	3,7	2,2	4,7	3,6

Table: Ergebnisse von Mendel

Beispiel 6.1.1.

(b) *Kreditwürdigkeit bei Kreditvergabe*

- Die Banken sind offensichtlich daran interessiert, Bankkredite an Kunden zu vergeben, die in der Zukunft solvent bleiben, also die Kreditraten regelmäßig zurückzahlen können.
- Um die Kreditwürdigkeit zu überprüfen, werden Umfragen gemacht, wobei die Antworten unter anderem in folgenden Variablen kodiert werden:

Beispiel 6.1.1.

- X_1 Laufendes Konto bei der Bank (1 $=$ nein, 2 $=$ ja und durchschnittlich geführt, 3 $=$ ja und gut geführt)
- X_2 Laufzeit des Kredits in Monaten
- X_3 Kredithöhe in Euro
- X_4 Rückzahlung früherer Kredite (gut/ schlecht)
- X_5 Verwendungszweck (privat / geschäftlich)
- X_6 Geschlecht (weiblich / männlich)

Beispiel 6.1.1.

- Um an Hand eines ausgefüllten Fragebogens wie diesem eine Entscheidung über die Vergabe des Kredits treffen zu können, werden *Lernstichproben* herangezogen, bei denen das Ergebnis Y der erfolgten Kreditvergabe bekannt ist.
- Dabei bedeutet $Y = 0$ gut und $Y = 1$ schlecht.
- Betrachten wir eine solche Stichprobe einer süddeutschen Bank, die 1000 Umfragebögen umfasst.
- Dabei sind 700 kreditwürdig und 300 davon nicht kreditwürdig gewesen.

Beispiel 6.1.1.

- Die folgende Tabelle zeigt Prozentzahlen dieses Datensatzes für ausgewählte Merkmale X_i.
- Dabei ist es möglich, mit Hilfe statistischer Methoden (Regression) eine Kreditentscheidung bei einem Kunden an Hand dieser Lernprobe automatisch treffen zu können.
- Dieser Vorgang wird manchmal auch ”statistisches Lernen” genannt.
- Fragestellungen wie diese werden erst in der Vorlesung Mathematical Statistics (verallgemeinerte lineare Modelle) behandelt.

Beispiel 6.1.1.

	Y	
X_1: laufendes Konto	1	0
nein	45, 0	19, 9
gut	15, 3	49, 7
mittel	39, 7	30, 4

Table: Lernstichprobe zur Vergabe von Krediten

Beispiel 6.1.1.

X_3: Kredithöhe in €	1	0
$0 < \ldots \leq 500$	$1,00$	$2,14$
$500 < \ldots \leq 1000$	$11,33$	$9,14$
$1000 < \ldots \leq 1500$	$17,00$	$19,86$
$1500 < \ldots \leq 2500$	$19,67$	$24,57$
$2500 < \ldots \leq 5000$	$25,00$	$28,57$
$5000 < \ldots \leq 7500$	$11,33$	$9,71$
$7500 < \ldots \leq 10000$	$6,67$	$3,71$
$10000 < \ldots \leq 15000$	$7,00$	$2,00$
$15000 < \ldots \leq 20000$	$1,00$	$0,29$

Table: Lernstichprobe zur Vergabe von Krediten

Beispiel 6.1.1.

X_4: Frühere Kredite	1	0
gut	82, 33	94, 85
schlecht	17, 66	5, 15
X_5: Verwendungszweck	1	0
privat	57, 53	69, 29
beruflich	42, 47	30, 71

Table: Lernstichprobe zur Vergabe von Krediten

Beispiel 6.1.1.

(c) *Korrosion von Legierungen*

- In diesem Beispiel wurde der Korrosionsgrad einer Kupfer-Nickel-Legierung in Abhängigkeit ihres Eisengehalts untersucht.
- Dazu wurden 13 verschiedene Räder mit dieser Legierung beschichtet und 60 Tage lang in Meerwasser gedreht.
- Danach wurde der Gewichtsverlust in mg pro dm^2 und Tag bestimmt.
- Aus dem folgenden Bild ist zu sehen, dass die Korrosion in Abhängigkeit vom Eisengehalt linear abnimmt.
- Mit statistischen Methoden (einfache lineare Regression) kann die Geschwindigkeit dieser Abnahme geschätzt werden.

Beispiel 6.1.1.

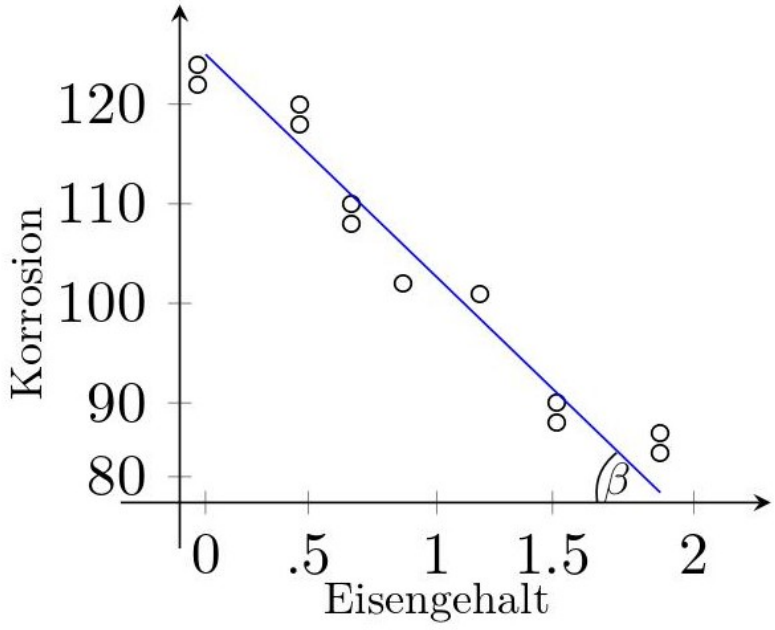
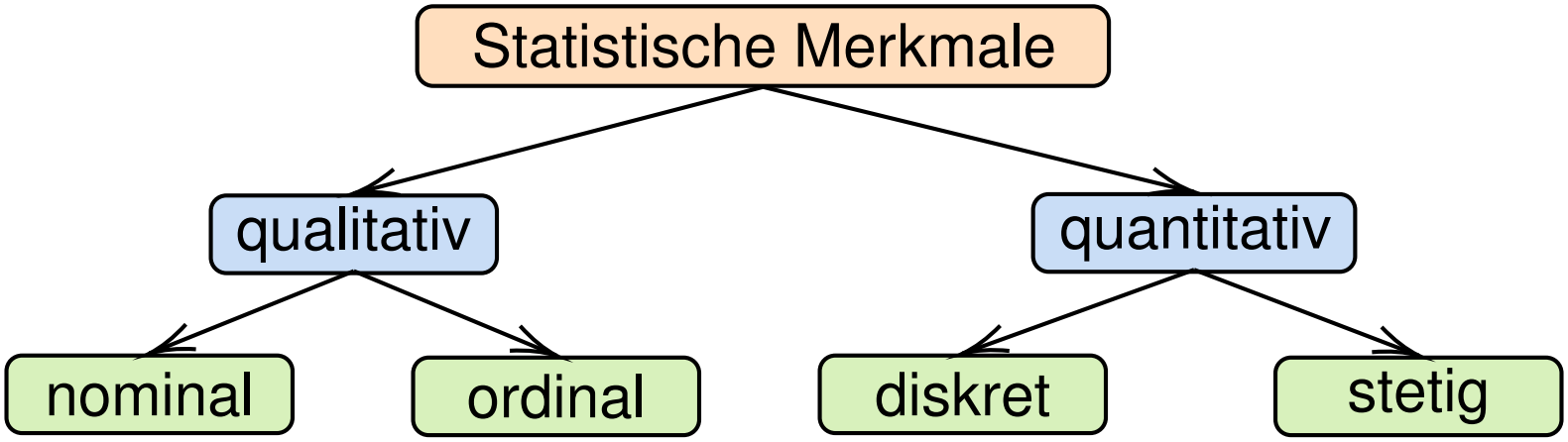


Figure: Korrosion von Kupfer-Nickel-Legierung

Stichproben und ihre Funktionen

- Die Daten, die zur statistischen Analyse vorliegen, können eine oder mehrere interessierende Größen (die auch *Variablen* oder *Merkmale* genannt werden) umfassen.
- Ihre Werte werden *Merkmalsausprägungen* genannt.
- In dem nachfolgenden Diagramm werden mögliche Typen der statistischen Merkmale gegeben.

Stichproben und ihre Funktionen



Stichproben und ihre Funktionen

- Diese Typen entstehen in Folge der Klassifikation von Wertebereichen (Skalen) der Merkmale.
- Dennoch ist diese Einteilung nicht vollständig und kann bei Bedarf erweitert werden.
- Man unterscheidet *qualitative* und *quantitative* Merkmale.
- *Quantitative Merkmale* lassen sich inhaltlich gut durch Zahlen darstellen (z.B. Kredithöhe in €, Körpergewicht und Körpergröße, Blutdruck usw.).
- Sie können *diskrete* oder *stetige* Wertebereiche haben, wobei diskrete Merkmale isolierte Werte annehmen können (z.B. Anzahl der Schäden eines Versicherers pro Jahr).
- Stetige Wertebereiche hingegen sind überabzählbar.

Stichproben und ihre Funktionen

- Dennoch liegen in der Praxis stetige Merkmale in gerundeter Form vor (z.B. Körpergröße auf cm gerundet, Geldbeträge auf € gerundet usw.).
- Im Gegensatz zu den quantitativen Merkmalen sind die Inhalte der *qualitativen Merkmale*, wie z.B. Blutgruppe (0, *A*, *B* und *AB*) oder Familienstand (ledig, verheiratet, verwitwet), nicht sinnvoll durch Zahlen darzustellen.
- Sie können zwar formell mit Zahlen kodiert werden (z.B. bei Blutgruppen $0 = 0$, $A = 1$, $B = 2$, $AB = 3$), aber solche Kodierungen stellen keinen inhaltlichen Zusammenhang zwischen Ausprägungen und Zahlen-Codes dar sondern dienen lediglich der besseren Identifikation der Merkmale auf einem Rechner.
- Es ist insbesondere unsinnig, Mittelwerte und ähnliches von solchen Codes zu bilden.

Stichproben und ihre Funktionen

- Ein qualitatives Merkmal mit nur 2 Ausprägungen (z.B. männlich / weiblich, Raucher / Nichtraucher) heißt *alternativ*.
- Ein qualitatives Merkmal kann *ordinal* (wenn sich eine natürliche lineare Ordnung in den Merkmalsausprägungen finden lässt, wie z.B. gut / mittel / schlecht bei Qualitätsbewertung in Umfragen oder sehr gut / gut / befriedigend / ausreichend / mangelhaft / ungenügend bei Schulnoten) oder *nominal* (wenn eine solche Ordnung nicht vorhanden ist) sein.
- Beispiele von nominalen Merkmalen sind Fahrzeugmarken in der KFZ-Versicherung (z.B. BMW, Peugeot, Volvo, usw.) oder Führerscheinklassen (A, B, C, ...).
- Datenmerkmale können auch mehrdimensionale Ausprägungen haben.

Stichproben und ihre Funktionen

- Aus den obigen Beispielen wird klar, dass ein Statistiker mit Datensätzen der Form $(x_1, \ldots, x_n)$ arbeitet, wobei die Einzeleinträge x_i aus einer Grundgesamtheit $G \subset \mathbb{R}^k$ stammen, die hypothetisch unendlich groß ist.
- Der vorliegende Datensatz $(x_1, \ldots, x_n)$ wird auch *(konkrete) Stichprobe* von Umfang n genannt.
- Die Menge B aller potentiell möglichen Stichproben bezeichnen wir als *Stichprobenraum* und setzen zur Vereinfachung der Notation $B = \mathbb{R}^{kn}$.
- Auf diesen Folien werden wir meistens die univariate statistische Analyse (also $k = 1$, ein eindimensionales Merkmal) betreiben.

Stichproben und ihre Funktionen

- In der beschreibenden Statistik arbeitet man mit Stichproben $(x_1, \ldots, x_n)$ und ihren Funktionen, um diese Daten visualisieren zu können.
- Für die Aufgabe der schließenden Statistik jedoch reicht diese Datenebene nicht mehr aus.
- Daher wird die zweite Ebene der Betrachtung eingeführt, die sogenannte *Modellebene*.
- Dabei wird angenommen, dass die konkrete Stichprobe $(x_1, \ldots, x_n)$ eine *Realisierung* eines stochastischen Modells $(X_1, \ldots, X_n)$ darstellt, wobei $X_1, \ldots, X_n$ (meistens unabhängige identisch verteilte) Zufallsvariablen auf einem (nicht näher spezifizierten) Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ sind.
- Diese Zufallsvariablen $X_i, \quad i = 1, \ldots, n$ können als konsequente Beobachtungen eines Merkmals interpretiert werden.

Stichproben und ihre Funktionen

- In Bsp. 6.1.1, 3a) z.B. die Erbsenform mit

$$X_i = \begin{cases} 0, & \text{falls Erbse } i \text{ rund,} \\ 1, & \text{falls Erbse } i \text{ eckig,} \end{cases} \qquad i = 1, \ldots, n.$$

- Der Vektor $(X_1, \ldots, X_n)$ wird dabei *Zufallsstichprobe* genannt.
- Man setzt weiter voraus, dass $EX_i^2 < \infty \; \forall i = 1, \ldots, n$, damit man von der Varianz Var X_i der Einzeleinträge sprechen kann.
- Es wird außerdem angenommen, dass ein $\omega \in \Omega$ existiert, sodass $X_i(\omega) = x_i \quad \forall i = 1, \ldots, n$.
- Sei F die Verteilungsfunktion der Zufallsvariablen X_i.
- Eine der wichtigsten Aufgaben der Statistik ist die Bestimmung von F (man sagt, ”Schätzung von F”) aus den konkreten Daten $(x_1, \ldots, x_n)$.

Stichproben und ihre Funktionen

- Dabei können auch Momente von F und ihre Funktionen (Erwartungswert, Varianz, Schiefe, usw.) von Interesse sein.
- Um die obigen Aufgaben erfüllen zu können, braucht man gewisse Funktionen $\varphi : \mathbb{R}^n \to \mathbb{R}^m, \quad m \in \mathbb{N}$ auf dem Stichprobenraum, die diese Stichprobe bewerten.

Definition 6.2.1.

Eine Borel-meßbare Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion*. Wenn man auf der Modellebene mit einer Zufallsstichprobe $(X_1, \ldots, X_n)$ arbeitet, so heißt die Zufallsvariable

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

eine *Statistik*. In der Schätztheorie spricht man dabei von *Schätzern* und bei statistischen Tests wird $\varphi(X_1, \ldots, X_n)$ *Teststatistik* genannt.

Stichproben und ihre Funktionen

Beispiele für Stichprobenfunktionen sind (unter Anderem) das *Stichprobenmittel*

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

die *Stichprobenvarianz*

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2$$

und die *Ordnungsstatistiken*

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \,,$$

welche entstehen, wenn man eine Stichprobe, die aus quantitativen Merkmalen besteht, linear ordnet

$$(x_{(1)} = \min_{i=1,\ldots,n} x_i, \ldots, x_{(n)} = \max_{i=1\ldots n} x_i).$$