Vorlesung Statistik WS 2008/09
1. Grundbegriffe

Andreas Eberle
Institut für angewandte Mathematik

Oktober 2008

Drei Stufen der Analyse von Beobachtungsdaten

I. Beschreibende Statistik

- Aufbereitung der Daten
- Graphische Darstellung (Stabdiagramm, Histogramm, Boxplot,...)
- Berechnung von Kennzahlen
  (Quantile, Mittelwert, Standardabweichung,...)

Drei Stufen der Analyse von Beobachtungsdaten

II. Modellierung und Wahrscheinlichkeitsrechnung

- Erstellen von mathematischen Modellen für das betrachtete Zufallsexperiment.
- In der Regel kennen wir das "richtige" Modell bestenfalls bis auf einige unbekannte Parameter. Bei einer Meinungsumfrage könnten wir z.B. annehmen, daß die Antworten verschiedener Personen unabhängig sind, und jeweils mit Wahrscheinlichkeit p ein bestimmter Kandidat bevorzugt wird. Den Wert von p kennen wir aber zunächst nicht.
- Berechnung von Wahrscheinlichkeiten für die beobachteten Daten unter Annahme der (verschiedenen) Modelle.

Drei Stufen der Analyse von Beobachtungsdaten

III. Schließende Statistik

- Schätzung der unbekannten Parameter aus den Beobachtungsdaten (mit Aussagen über den Schätzfehler). Rückschluss auf das "richtige" zugrundeliegende mathematische Modell.
- Überprüfen der Plausibilität verschiedener Hypothesen zum zugrundeliegenden Modell (z.B. "Die Erfolgswahrscheinlichkeit von Kandidat A ist größer als 50 %")
- Rückschluss von der Stichprobe auf die Grundgesamtheit.

1. Grundbegriffe der beschreibenden Statistik

Statistische Einheiten, Grundgesamtheit

Definition

Die **Grundgesamtheit** oder **statistische Masse** Ω ist die Menge aller uns interessierenden **statistischen Einheiten** ω.

Beispiel.

$\Omega =$ *alle Hörer im Saal, alle Agrar-/ELW-Studenten im 2. Studienjahr, alle Bäume im Kottenforst, alle Fische im Poppelsdorfer Weiher, ...*

Definition

Die **Stichprobe** $\chi = \{\omega_1, \omega_2, ..., \omega_n\}$ ist der von uns erfasste Teil der Grundgesamtheit.

Eine Stichprobe ist also eine Teilmenge von Ω, z.B.

χ = alle Hörer, die heute in der 1. Reihe sitzen

χ = 2000 Bäume auf 10 repräsentativ ausgewählten Flächenstücken

Grundbegriffe der beschreibenden Statistik

Merkmale

Definition

Ein **Merkmal** $X(\omega)$ ist eine uns interessierende Eigenschaft einer statistischen Einheit ω. Die **Merkmalsausprägungen** sind alle möglichen Werte, die ein Merkmal annehmen kann.

Beispiel.

- $X(\omega) =$ *Körpergröße des Studenten* ω *(in cm)*
  Merkmalsausprägungen: alle positiven reellen Zahlen
- $X(\omega) =$ *von Wähler* ω *bevorzugte Partei*
  Merkmalsausprägungen: CDU, SPD, Grüne, FDP, Linke, keine, k.A.
- $(X(\omega), Y(\omega)) = ($*Alter von Wähler* ω*, von* ω *bevorzugte Partei*$)$
  mehrdimensionales Merkmal

Definition

Die in der Stichprobe beobachteten Merkmalsausprägungen

$$x_i = X(\omega_i), \quad i = 1, 2, ..., n,$$

heißen **Merkmalswerte, Beobachtungswerte, oder Daten.**

Beispiel.

$$\begin{aligned} (x_1, x_2, ..., x_n) &= (185, 193, 160, ..., 168) \quad \textit{(Größe in cm)} \\ (y_1, y_2, ..., y_n) &= (m, w, w, ..., m) \qquad\quad \textit{(Geschlecht)} \\ ((x_1, y_1), ..., (x_n, y_n)) &= ((185, m), (193, w), ..., (168, m)) \end{aligned}$$

Beispiel. (Meinungsumfrage zur Präsidentenwahl in den USA)

- $\Omega =$ *alle Wahlberechtigten in Florida*
- $\chi =$ *"repräsentativ ausgewählte" Stichprobe von 5000 Wählern*
- *Beobachtetes Merkmal: Antwort auf Sonntagsfrage*
- *Merkmalsausprägungen: Obama, McCain, Nader, Enthaltung, unentschlossen*
- $(x_1, x_2, \ldots, x_{5000}) = (\textit{Obama, McCain, Obama, unentschl.,....})$

Beispiel. (Klausurergebnis)

- $\Omega =$ *alle Schüler in Klasse 7c des EMA*
- $\chi = \Omega$
- *Beobachtetes Merkmal: Note in erster Matheklausur 2008/09*
- *Merkmalsausprägungen: 1,2,3,4,5,6*
- $(x_1, x_2, ..., x_{27}) = (2, 1, 1, 2, 3, 2, 4, 1, ...)$

Standarddarstellung von Datensätzen
(z.B. in Textdatei oder Excel-Tabelle):

Student	Geschlecht	Alter	Größe
1	m	23	195
2	w	25	193
3	w	18	165

Welche Merkmalsarten gibt es ?

- **Qualitative Merkmale**
  - **Nominalskaliert:**
    Merkmalsausprägungen haben keine natürliche Ordnung.
    Z.B. weiblich/männlich, CDU/SPD/Grüne/FDP/Linke/k.A.
  - **Ordinalskaliert:**
    Merkmalsausprägungen haben eine natürliche Ordnung.
    Z.B. super-gut-mittelmäßig-schlecht
    oder links-Mitte-rechts
- **Quantitative Merkmale**
  - **Metrisch skaliert (Kardinalskaliert):**
    Merkmale sind Zahlen, und können auf sinnvolle Weise addiert und subtrahiert werden (d.h. die Abstände der Zahlen haben eine praktische Bedeutung).
    Z.B. Alter, Bevölkerungszahl, Temperaturmeßwert,....

Beipiele.

1. *Präsidentenwahl USA*
   Merkmalsausprägungen: Obama, McCain, Nader, k.A.
   Qualitativ, Nominalskaliert
2. *Klausurnoten in der Schule*
   Das Merkmal ist eigentlich **qualitativ** und **ordinalskaliert**, denn die Ergebnisse werden in geordnete Gruppen eingeteilt ("sehr gut","gut","befriedigend",...). Häufig wird das Merkmal "Klasurnote" aber *quantitativ* interpretiert, und es werden z.B. Mittelwerte von mehreren Klausurnoten gebildet.
3. *Körpergröße von Studenten*
   Metrisch skaliert - kontinuierlich

Beipiele.

1. *Präsidentenwahl USA*
   Merkmalsausprägungen: Obama, McCain, Nader, k.A.
   Qualitativ, Nominalskaliert
2. *Klausurnoten in der Schule*
   Das Merkmal ist eigentlich **qualitativ** und **ordinalskaliert**, denn die Ergebnisse werden in geordnete Gruppen eingeteilt ("sehr gut","gut","befriedigend",...). Häufig wird das Merkmal "Klasurnote" aber *quantitativ* interpretiert, und es werden z.B. Mittelwerte von mehreren Klausurnoten gebildet.
3. *Körpergröße von Studenten*
   Metrisch skaliert - kontinuierlich
4. *Regentage pro Woche*
   Metrisch skaliert - diskret
   (kann man auch als ordinalskaliert interpretieren)

Bemerkungen.

- Es gibt noch andere Arten von Merkmalsausprägungen.
- Nicht immer gibt es eine eindeutige Zuordnung.
- Quantitative Merkmale können durch Klasseneinteilung in ordinalskalierte Merkmale überführt werden.

Beispiel. (Anzahl der Regentage pro Jahr)

0-70 trocken, 70-150 normal, $>$150 nass

Gesichtspunkte bei der Datenerhebung

Welche statistischen Einheiten werden betrachtet ? Wie wird die Stichprobe gewählt ?

- Welche Einheiten betrachtet werden, muß möglichst genau abgegrenzt werden:
  - sachlich (z.B. Hörer der Vorlesung Biometrie und Methodik)
  - räumlich (an der Universität Bonn, landwirtschaftl. Fakutät)
  - zeitlich (erste Vorlesung im WS 08/09)
- Die Auswahl der Stichprobe muß genau durchdacht sein, und wiederum genau abgegrenzt werden.
  - Idealfall: Zufallsstichprobe
  - Oft ist das Ziehen einer Zufallsstichprobe nicht möglich. Klassische wahrscheinlichkeitstheoretische Modelle sind dann nicht mehr unmittelbar anwendbar. Wie erhält man trotzdem eine "repräsentive" Stichprobe, d.h. eine Stichprobe, deren Eigenschaften denen einer Zufallsstichprobe "möglichst nahe kommen" ?

Beispiel zur Wahl der statistischen Einheiten

- Sie wollen die Erträge pro Anbaufläche von Baumwollfarmer/innen im Mittleren Westen der U.S.A. untersuchen. Mithilfe eines "Stichprobendesigns" wählen Sie 75 Farmen aus. Sie telefonieren mit den Farmern, und haben anschließend einen Zettel vor sich, der ungefähr so aussieht (die 75 Farmen haben Sie durchnummeriert):

Farm	Name	Tel.-No.	Anbaufläche in Morgen	Jahresproduktion in Pfund
1.	Jane	0361-5576	150	38250
2.	Bob	0342-43487	100	37300
3.	Gerry	0416-6437	325	81250
4.	Mary	0531-43476	400	130 000
5.	Greg	0327-64764	370	103970
6.	...	...	...	...
...				

- Einen solchen Zettel nennen wir die Urliste.

- Der nächste Schritt ist die Bearbeitung der Urliste. Lassen Sie alle Daten weg, die Sie für Ihre Untersuchung für unwichtig halten, z.B. den Namen des Farmers / der Farmerin, die Tel-No. etc. . .
- Berechnen Sie die interessierende Größe Ertrag/Anbaufläche

Farm	Fläche [Morgen]	Jahresproduktion [Pfund]	Flächenertrag
1.	150	38250	255
2.	100	37300	373
3.	325	81250	250
4.	400	130 000	325
5.	370	103970	281
6.	. . .	. . .	. . .
. . .			

- Da Sie sich nur für die letzte Spalte interessieren, lassen Sie die drei ersten Spalten weg und erhalten, indem Sie nach der Höhe des Flächenertrags sortieren:

Tabelle der Baumwollerträge pro Fläche (in Pfund/(Morgen$\times$Jahr)) von 75 Farmen im Westen der USA

215	217	228	234	235
242	249	250	251	254
255	256	257	258	259
260	260	261	268	268
...	...	...	...	...
...	...	...	...	...
354	358	365	367	373

- Diese Darstellung ist zwar vollständig, aber nicht besonders übersichtlich. Besser ist es daher, die Daten in Klassen aufzuteilen.

Einteilung in Klassen

• In wieviele Klassen sollen wir die Daten einteilen? Einen Anhaltspunkt liefert die Faustformel von Sturge:

$$\boxed{K = 1 + \log_2 n}$$

mit K der Anzahl von Klassen (muss gerundet werden) und n der Anzahl der untersuchten Einheiten. Für $n = 75$ liegt $K = 7.037$ ziemlich nah bei 7.

• Den Wertebereich (215-373) teilen wir also in 7 gleichgroße Klassen der Breite $(373 - 215)/7 \cong 20$ auf. Durch die Rundungsfehler passt der Wertebereich dann schließlich in 8 Klassen. Wir berechnen nun die Häufigkeit, dass der Ertrag/Fläche in einer bestimmten Klasse liegt:

Häufigkeit und Klassenmittelpunkte in Pfund Baumwolle pro Morgen und Jahr

Klasse	Klassenmittelpunkt	Häufigkeit
215–235	225	4
235–255	245	6
255–275	265	13
275-295	285	21
295-315	305	15
315–335	325	7
335–355	345	5
355–375	365	4
Summe		75

- Nach Anschauen der Tabelle könnte man nun sagen:

> Der mittlere Baumwollertrag liegt im mittleren Westen der USA bei ca. 295 Pfund/(Morgen x Jahr)

- Kennt man nun die Gesamtanbaufläche von Baumwolle im Mittleren Westen, könnte man auf die jährliche Gesamtproduktion schließen...

Dieser Schluß ist falsch!!!

• Die vorhandenen Daten wurden falsch interpretiert. Man nehme einen extremen Fall an: Eine Farmerin, Jill, ist eine gefürchtete Baumwollbaronin, mit einer Anbaufläche von 70.000 Morgen und einem Ertrag von 367 Pfund/Morgen. Diesen weit überdurchschnittlichen Ertrag/Morgen erzielt sie, da sie sich nicht an Umweltauflagen zu halten braucht, stattdessen geht sie jede Woche mit dem Gouverneur essen...

- Obwohl die Farm von Jill die Gesamtproduktion sehr stark beeinflusst, geht sie in unserer Statistik nur mit einem Eintrag ein, genau wie die Minifarm von Bob.
- Unsere Daten aus der gruppierten Tabelle erlauben i.a. keine Aussagen über den mittleren Ertrag auf einem Morgen. Es werden in dieser Tabelle nämlich nicht die Morgen, sondern die Farmen als statistische Einheiten untersucht.
- Wollten wir die Morgen untersuchen, müssten wir eine Tabelle aufstellen, in denen der Ertrag eines Morgens auf jeder Farm mit ihrer Größe (also der Anzahl von Morgen) gewichtet wird.

• Sind die Daten in dem Diagramm also nutzlos? Nein, aber wir haben sie so aufbereitet, dass wir nur noch nach dem fragen können, was tatsächlich untersucht wird, die Effizienz der Farmer/innen! Richtige, durch die Daten gestützte Aussagen sind z.B.:

- "Farmerin Jane mit 255 Pfund/(Morgen x Jahr) muss sich sorgen um ihren Ertrag machen, denn die mittlere Farmerin / der mittlere Farmer erzielt deutlich mehr - vielleicht macht sie etwas falsch?"
- "Die Baumwollbaronin Jill kann mit 367 Pfund/(Morgen x Jahr) sehr zufrieden sein."
- "Übertroffen wird sie nur von Farmer Bob mit der Minifarm. Das ist sehr bemerkenswert und man sollte in Erfahrung bringen, was Bob besser macht als die anderen"
- Die Daten sind also durchaus aussagekräftig (für Jane, Jill & Bob, aber nicht für den Minister für Export, der die Gesamtproduktion kennen möchte).