



K2 Beschreibende Statistik

2.1 Fragen

Was für Fälle?

Personen (z.B. Studenten im dritten Semester)

Standorte (z.B. Kreisstädte)

Wiederholte Versuche (z.B. Laufzeiten)

Wieviele im Datensatz?

Gibt es mehr?

Wären mehr leicht zu finden?

Wieviele Variablen?

z. B. Studenten: Alter, Geschlecht, Fachgebiet

z.B. Städte: Fläche, Bevölkerungsdichte,
Verbrechensraten, Krankenhausbetten, Anzahl
Autos

Was für Variablen?



2.2 Beschreibung eines Datensatzes

Die einfachste Form ist eine Datentabelle:

n Fälle in Zeilen

m Variablen in Spalten

⇒ eine $n \times m$ Datenmatrix

x_{ij} = Wert der Variable j für Fall i

Kompliziertere Strukturen gibt es aus verknüpften Tabellen, z.B.

Patient i wird vom Arzt d_j behandelt

im Krankenhaus K_k

wegen Krankheit I_l

mit Medizin M_r

Arzt d_j arbeitet in den Krankenhäusern K_k und K_s

Krankenhaus K_k hat B_k Betten

u.s.w.



Datensatz aus einem irischen Touristenumfrage

Tage Irland	Anzahl Pers	Ziel	AutoVerm	Geld	Nur Irland
21	1	6	Yes	1606	Only
4	1	4	No	200	No
4	1	4	No	100	No
10	2	3	Yes	1000	Only
4	1	Business	No	100	No
7	1	3	No	900	Only
2	1	4	No	113	No
9	1	•	No		No
7	1	4	No	124	No
2	2	4	No	590	No
7	2	4	No	670	No
7	1	6	No	100	No
7	3	3	No		Only
2	1	2	No	350	No
3.5	1	4	No	64	No
3	11	7	Yes	300	No
6	1	3	Yes	1770	Only
3	2	6	Yes	539	No
4	1	4	No	100	Only
5	1	2	Yes	960	No
	2	6	Yes	126	No
5	1	3	No	456	Only
2	1	6	No	329	No
11	1	5	No		Only
2	1	6	No	415	Only
6	1	6	No	900	Only
10	1	11	No	317	Only
19	1	2	No	500	Only
8	2	3	Yes	1586	No
10	1	3	No	1000	Only



2.3 Datentypen

Kategorial

Nominal (z.B. männlich oder weiblich)

Ordinal (z.B. Eiskunstlaufwertungen)

Diskret (z.B. Anzahl Kinder)

Stetig (z.B. Regen)

Datenvariabilität

Niedrig oder hoch

(z.B. Butterpreise v. Computerpreise)

Systematisch oder zufällig

(z.B. tägliche Temperatur v. Aktienpreise)

Datenqualität

Alt oder modern

(z.B. 1983 Verkäufe oder 2006 Verkäufe)

Zuverlässig oder zweifelhaft

(z.B. Umfragedaten oder Anekdoten)



2.4 Schwierigkeiten mit Daten

(1) Datenqualität

- Fehler
- Ungenauigkeiten
- Widersprüche

(2) Ausreißer

„Wahre“ Werte, die sich von den anderen stark unterscheiden

(3) Fehlende Werte

- Daten existieren, sind aber nicht verfügbar
(z. B. Verkäufe der Konkurrenten)
- Daten existierten, sind aber nie gemessen und werden nie verfügbar
(z. B. Temperatur einer Patientin gestern)
- Daten existierten nie
(z. B. Aktienpreise am Sonntag)



2.5 Beschreibung kategorialer Variablen

Häufigkeitstabelle

n_i Anzahl in der Kategorie i (absolute Häufigkeit)

h_i Proportion in der Kategorie i (relative Häufigkeit)

$$\sum n_i = n \text{ (Gesamtzahl der Fälle)}$$

$$h_i = n_i / n$$

$$\sum h_i = 1$$

z.B. Autovermietung

Kategorie

n_i

h_i

Ja

Nein



Häufigkeitstabelle für viele Kategorien (z.B. Ziel des Besuchs)

<u>Kategorie</u>	n_i	h_i
2		
3		
4		
5		
6		
7		
11		
Geschäft		
Fehlend		
Gesamt		

Um einen Überblick zu gewinnen, können wir Kategorien zusammenfassen. Auf der einen Seite möchten wir wichtige Sonderfälle nicht aus dem Auge verlieren, auf der anderen Seite möchten wir uns auf die größten Gruppen konzentrieren.



2.6 Beschreibung von stetigen Variablen

Klasseneinteilung in Tabellen

z.B. Dauer des Besuchs in Irland

<u>Tage</u>	n_i	h_i
-------------	-------	-------

$0 < x \leq 5$

$5 < x \leq 10$

$10 < x \leq 15$

$15 < x \leq 20$

$20 < x$

oder $0 < x \leq 1$, $1 < x \leq 2$, $2 < x \leq 3$, u.s.w.

oder $0 < x \leq 15$, $15 < x \leq 30$, $30 < x$

Gute Tabellen

- umfassen alle Beobachtungswerte
- erzwingen eine eindeutige Klassenzuordnung
- haben nicht zu viele Klassen
- haben nicht zu wenige Klassen
- haben „sinnvolle“ Klassen



2.7 Statistiken / Kenngrößen

Datensatz $\{x_1, x_2, \dots, x_n\}$

oder in aufsteigender Reihenfolge $\{x_{(i)}\}$

2.7.1 Stichprobengröße n

Sehr wichtig. Im Prinzip je mehr desto besser, aber alles hängt von der Erhebungsmethode ab.

2.7.2 Ordnungsstatistiken und Quantile

Maximum $x_{(n)}$ z.B. höchste Temperatur in Bayern

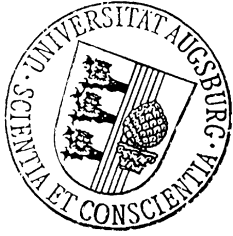
Minimum $x_{(1)}$ z.B. schnellste Zeit in Formel 1

Theoretische Quantile

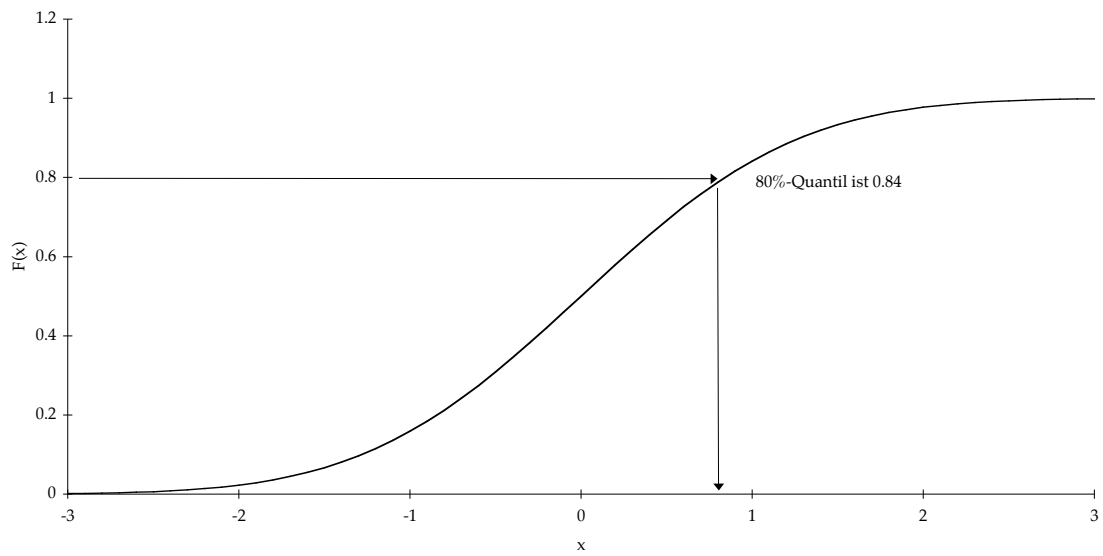
$$Q(p) = \quad (2.7.1)$$

z.B. Krankenwagen sollten 95% aller Unfälle innerhalb von 10 (?) Minuten erreichen

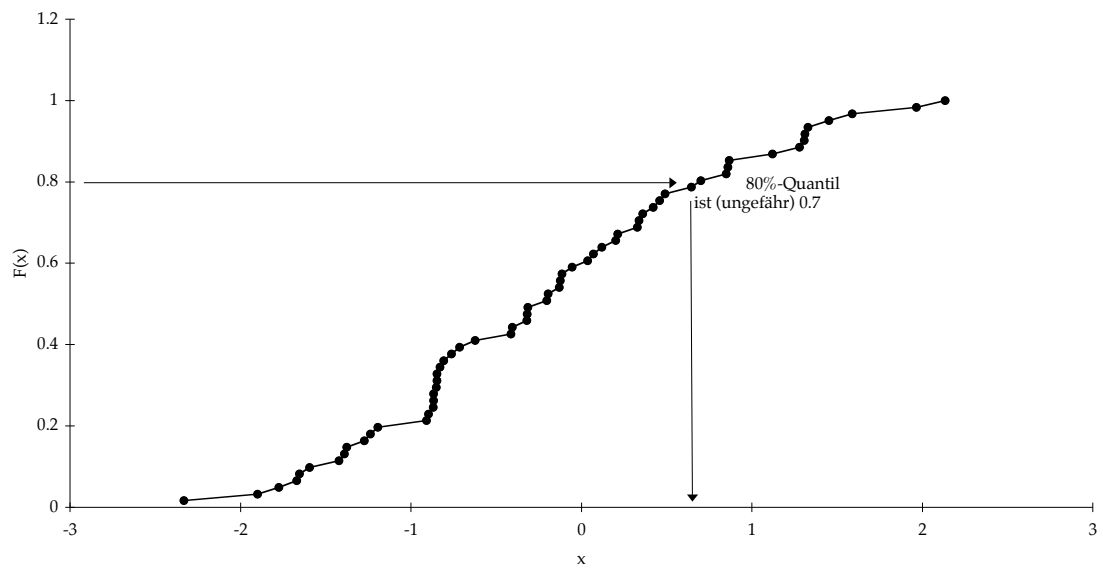
Für Empirische Quantile gibt es keine ideale einfache Definition. Schwierigkeiten tauchen bei kleinen Datensätzen und bei extremen Quantilen auf.



Beispiel eines theoretischen Quantils



Beispiel eines empirischen Quantils





2.7.3 Lage

(a) Mittelwert $\bar{x} =$ (2.7.2)

Nützlich für Schätzer z. B. Wieviel kauft ein Durchschnittsverbraucher pro Woche?

Empfindlich gegen Ausreißer

Mathematisch schön (s. Zentraler Grenzwertsatz)

$\bar{x} \sim N(\mu, \sigma^2)$ für n groß und

$t = (\bar{x} - \mu) \cdot (\sqrt{n}) / s \sim t_{n-1}$ für X Normal

(b) Median (Zentraler Wert) 50%-Quantil

$$m = \quad (n = 2k + 1)$$

und $m = \quad (n = 2k)$ (2.7.3)

Robust, aber mathematisch schwierig

z.B. Wie hoch ist das durchschnittliche Einkommen?

(c) Modus (der häufigste Wert)

Nicht eindeutig

Manchmal nützlich als typischer Wert

z.B. Wie lange bleibt man im Krankenhaus für eine Blinddarmoperation?



2.7.4 Variabilität/Streuung

(a) Spannweite := Maximum - Minimum $x_{(n)} - x_{(1)}$

Sehr empfindlich gegenüber Ausreißern

Mathematisch unangenehm

(b) Interquartilsabstand IQ = (2.7.4)

Robust, mathematisch schwierig

(c) Varianz $\hat{\sigma}^2 =$ (2.7.5)

(wenn $X \sim N$)

$s^2 =$

$(n - 1) \cdot s^2 / \sigma^2 \sim$ (wenn $X \sim N$)

(d) Standardabweichung $s = \sqrt{\text{Varianz}}$

Meistgebraucht

Ausreißerempfindlich

Schwach bei asymmetrischen Daten

Schwer zu schätzen

Gleiche Einheiten wie die Daten selbst