Werkzeuge der empirischen Forschung

Wolfgang Kössler

Institut für Informatik, Humboldt-Universität zu Berlin SS2008

Übersicht

1. Einleitung
Statistik und Wahrscheinlichkeitsrechnung

Stochastik

- befasst sich mit zufälligen Erscheinungen
  Häufigkeit, Wahrscheinlichkeit und Zufall
  grch: Kunst des geschickten Vermutens
- Teilgebiete
  - Wahrscheinlichkeitsrechnung
  - Statistik

Wahrscheinlichkeitsrechnung

gegebene Grundgesamtheit (Verteilung) $\rightarrow$ Aussagen über Realisierungen einer Zufallsvariablen treffen.

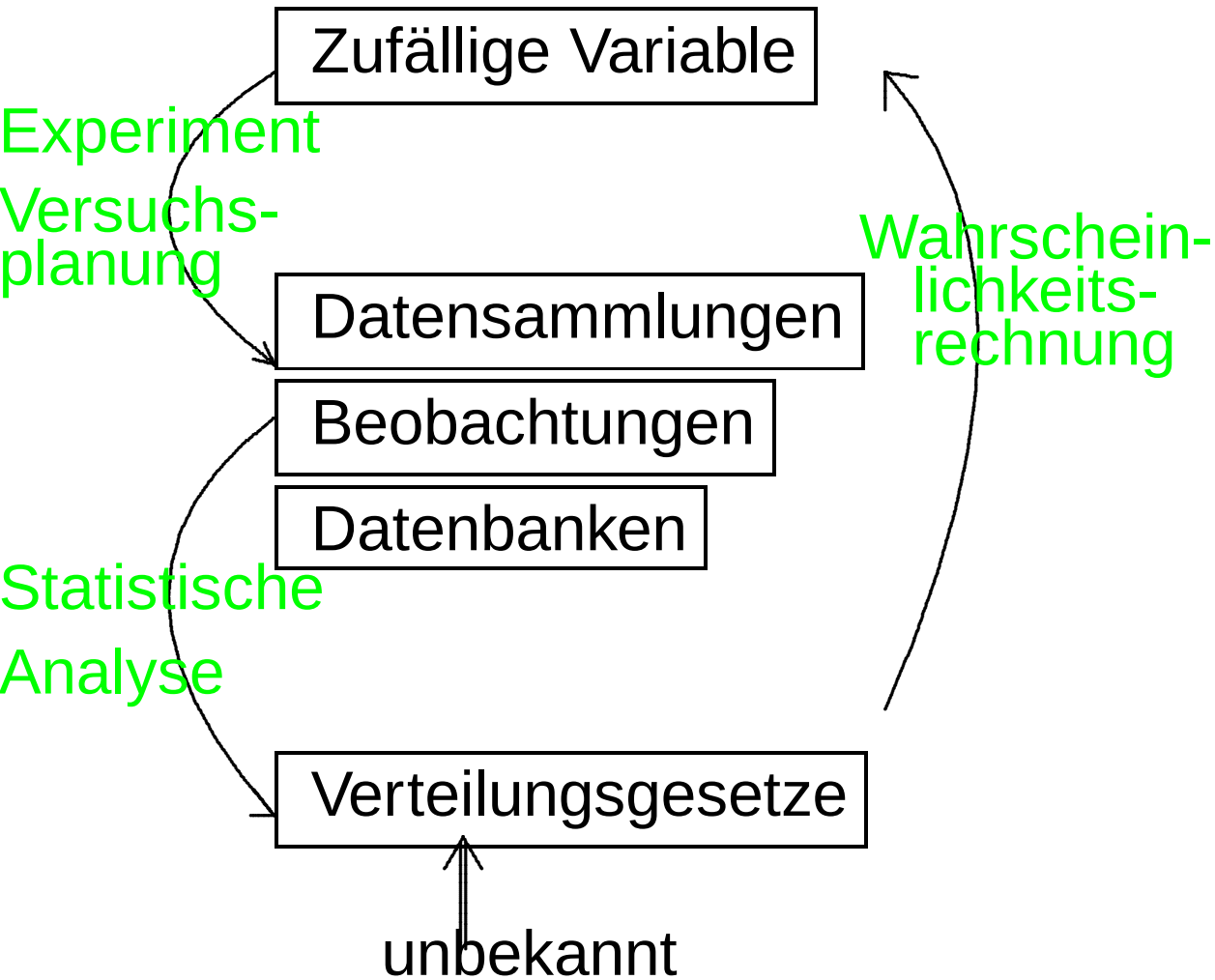
Einleitung
Statistik

Statistik

- Gesamtheit aller Methoden zur Analyse zufallsbehafteter Datenmengen
- Gegeben: (Besondere) zufallsbehaftete Datenmengen
- Gesucht: (Allgemeine) Aussagen über die zugrundeliegende Grundgesamtheit
- Teilgebiete:
  - Beschreibende oder Deskriptive Statistik
  - Induktive Statistik
  - Exporative oder Hyothesen-generierende Statistik (data mining)

Einleitung

Überblick: Statistik



Einleitung
Beschreibene Statistik

Beschreibene Statistik

- statistische Maßzahlen: Mittelwerte, Streuungen, Quantile, ...
- Box-Blots
- Stamm-und Blattdiagramme
- Balkendiagramme
- Zusammenhangsmaße
- Punktediagramme (Scatterplots)

Einleitung
Schließende Statistik

Schließende Statistik

- Vergleich von Behandlungen, Grundgesamtheiten, Effekten
  $\rightarrow$ **t-Test, Wilcoxon-Test, ANOVA, Kruskal-Wallis-Test, Friedman-Test**
- Ursache-Wirkungsanalysen, Vorhersagen, Bestimmen funktionaler Beziehungen, Trendbestimmungen
  $\rightarrow$ **lineare, nichtlineare Regression**
  $\rightarrow$ **Kurvenschätzung**
  $\rightarrow$ **logistische Regression**
  $\rightarrow$ **Korrelation und Unabhängigkeit**

Einleitung
Schließende Statistik

Schließende Statistik

- Klassifikation
  - → **Clusteranalyse**
  - → **Hauptkomponentenanalyse**
  - → **Faktorenanalyse**
  - → **Diskriminanzanalyse**
- weitere Verfahren
  - → **Lebensdaueranalyse (Zuverlässigkeit)**
  - → **Qualitätskontrolle**
  - → **Zeitreihenanalyse**

Einleitung

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

- Einstichprobenproblem
  Messungen sollen mit einem vorgegebenen Wert verglichen werden:
- Zweistichprobenproblem
  - Vergleich zweier unabhängiger Stichproben
  - Vergleich zweier abhängiger Stichproben
- Vergleich mehrerer unabhängiger Stichproben
- Vergleich mehrerer abhängiger Stichproben

Einleitung
Ein- und Zweistichprobenproblem

Eine Stichprobe

- Banknoten: vorgegebene Länge eingehalten?

→ **Einstichproben t-Test, Signed-Wilcoxon-Test**

Abhängige und Unabhängige Stichproben

- Vergleich zweier unabhängiger Stichproben
  - echte - gefälschte Banknoten
  - Schädel aus verschiedenen Gegenden Tibets

  → **t-Test, Wilcoxon-Test**
- Vergleich zweier abhängiger Stichproben
  Länge des Scheines oben und unten
  → **Einstichproben t-Test,**
  Vorzeichen-Wilcoxon-Test

Einleitung

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

Abhängige und Unabhängige Stichproben

- Vergleich mehrerer unabhängiger Stichproben: Ägypt. Schädel: mehrere Grundgesamtheiten, Epochen
  $\rightarrow$ **ANOVA, Kruskal-Wallis-Test**
- Vergleich mehrerer abhängiger Stichproben Blutdruck von Patienten an mehreren aufeinanderfolgenden Tagen, (Faktoren: Patient, Tag)
  Preisrichter beim Synchronschwimmen
  $\rightarrow$ **2 fakt. Varianzanalyse, Friedman-Test**

Einleitung
Ursache - Wirkungsanalysen

Ursache - Wirkungsanalysen

- Ursache - Wirkungsanalysen
  - Zusammenhangsanalyse
  - Bestimmen funktionaler Beziehungen
  - Trends, Vorhersagen
- Beispiele:
  - Bluthochdruck - Rauchgewohnheiten
  - Blutdruck - Proteinuria
  - Größe - Gewicht
  - Sterblichkeit - Wasserhärte

$\longrightarrow$ **Lineare, Nichtlineare und Nichtparametrische Regression**

$\longrightarrow$ **Korrelation**

Einleitung
Klassifiaktion

Klassifikation

- Auffinden von Gruppen in Daten
  → **Clusteranalyse**
- Individuen sollen einer von vorgegebenen Klassen zugeordnet werden
  → **Diskriminanzanalyse**
- Datensatz hat Variablen, die mehr oder weniger voneinander abhängen.
  Welche Struktur besteht zwischen den Variablen?
  → **Hauptkomponentenanalyse**
  → **Faktorenanalyse**

Einleitung
Literatur

Literatur (1)

Dufner, Jensen, Schumacher (1992). Statistik mit SAS, Teubner.

Falk, Becker, Marohn (1995). Angewandte Statistik mit SAS, Springer.

Graf, Ortseifen (1995). Statistische und grafische Datenanalyse mit SAS, Spektrum akademischer Verlag Heidelberg.

Krämer, Schoffer, Tschiersch (2004). Datenanalyse mit SAS, Springer.

SAS-Oneline Dokumentation, SAS-Handbücher

Einleitung
Literatur

Literatur (2)

Hartung (1993). Statistik, Lehr- und Handbuch, Oldenbourg.

Sachs (1999). Angewandte Statistik, Springer.

Muche, Habel, Rohlmann (2000). Medizinische Statistik mit SAS Analyst, Springer.

Graf, Bundschuh, Kruse (1993). Effektives Arbeiten mit SAS, Wissenschaftsverlag.

Gogolok, Schuemer, Ströhlein (1990). Datenverarbeitung und statistische Auswertung mit SAS, Fischer

Einleitung
Literatur

Literatur (3)

Nagel, W, (1992). Statistische Datenanalyse mit SAS. Campus Verlag Frankfurt/M.

Steinhausen, Zörkendörfer (1992). Informationsverarbeitung und Datenanalyse mit dem Programmsystem SAS, Oldenbourg.

Göttsche (1992). SAS-kompakt, Fischer.

Einleitung
Statistik Software

Statistik-Software

SAS	-	sehr umfangreich, universell
	-	weit verbreitet
SPSS	-	umfangreich
	-	Anwendung vor allem in Biowiss., Medizin, Sozialwiss.
SYSTAT	-	ähnlich wie SPSS
	-	sehr gut
BMDP	-	umfangreich
S, S^+, R	-	funktionale Sprachen
	-	R: frei verfügbar

STATGRAPHICS, XPLORE, MATHEMATICA ...

Einleitung
Starten und Beenden von SAS

Account für Mathepool beantragen

- Die Software ist im Mathepool R. 2.212 und R. 2.213 installiert.
- vor erster Benutzung Nutzerkennzeichen beantragen

`https://www.math.hu-berlin.de/account`

- Passwort abholen (Dr. Gehne, 2.2.05)

Einleitung
Starten und Beenden von SAS

Starten und Beenden von SAS

- Starten von SAS
  1. Sitzungsart KDE anmelden
  2. Behelfsfenster- **Konsole** starten
  3. beim Windows-Server einloggen:
     `rdesktop -f idun`
     Passwort angeben;
     log on to: **localmath** (nicht: idun)
  4. Start von SAS: **All Programs** $>$ **SAS-System** $>$ **The SAS-System 9.1 (English)**
- Beenden der Sitzung
  All Desktop exit $>$ **Logoff** $>$ **Abmelden**

Einleitung
Allgemeine Struktur von SAS

SAS-Fenster

- Nach dem Starten erscheinen 3 Fenster
  - Log-Fenster
  - Editor-Fenster
  - Output-Fenster (verdeckt)
- weitere Fenster:
  - Results: Ergebnisse aus der Sitzung
  - Grafik-Fenster (gegebenfalls)
  - Hilfen

Einleitung
Allgemeine Struktur von SAS

Hilfen

- help > SAS Help and Documentation
- SAS Products
- BASE SAS
  - > SAS Language Concepts
    - > Data Step Concepts
  - > SAS STAT
  - > SAS STAT User's Guide

Einleitung
Allgemeine Struktur eines SAS-Programms

Aufbau einer SAS-file

DATA

PROC

DATA

PROC

PROC

. . .

- DATA-Schritte:
  - Erstellen der SAS-Dateien
  - Einlesen, Erstellen, Modifikation der Daten
- PROC-Schritte: Auswertung der Dateien

Einleitung
Daten

Daten

Ausgangspunkt sind die Daten, die für die Analyse relevant sind.
Die Struktur der Daten hat die folgende allgemeine Form: x_{ij}

Objekte	Merkmale						
	1	2	3	..	j	..	p
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Wert oder Ausprägung des Merkmals j am Objekt i

Einleitung
Daten

Daten

p: Anzahl der Merkmale

N: Gesamtanzahl der einbezogenen Objekte (Individuen)

Objekte	Merkmale						
	1	2	3	..	j	..	p
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Qualität des Datenmaterials wird im Wesentlichen durch die Auswahl der Objekte aus einer größeren Grundgesamtheit bestimmt.

Einleitung
Daten

Beispiele

- Objekte: Patienten einer Klinik
  Merkmale: Alter, Geschlecht, Krankheiten
- Objekte: Bäckereien in einer bestimmten Region
  Merkmale: Anzahl der Beschäftigten, Geräteausstattung, Umsatz, Produktpalette
- Objekte: Banknoten
  Merkmale: Längenparameter

Einleitung
Daten

Datenmatrix

- Zeilen: Individuen, Objekte, Beobachtungen
- Spalten: Merkmalsausprägungen, -werte, -realisierungen

Banknote	Merkmale						
	laenge	oben	unten	..	j	..	gr
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Einleitung
Daten

Merkmale

- Definition: **Merkmale** sind Zufallsvariablen, die für jedes Individuum (Objekt) eine bestimmte Realisierung (Merkmalsausprägung) haben.
- Stetige Merkmale: laenge, oben
- Diskrete Merkmale: gr (Gruppe)

Banknote	Merkmale						
	laenge	oben	unten	..	j	..	gr
1							
2							
..							

- Allgemeine Syntax
- Eingabe über die Tastatur
- Transformationen
- Eingabe durch externes File
- Wichtige Varianten der INPUT-Anweisung
- Ein- u. Ausgabe von SAS-Files
- Zusammenfügen von Files
- Output-Anweisung
- DO-Schleifen im DATA-Step

2. Dateneingabe und Transformation

2.0 Allgemeine Syntax

DATA <dateiname <(dateioptionen)>>;

...

RUN;

<... > kennzeichnet optionale Parameter

Externes File INFILE ' ... ' ; INPUT ... ;	Tastatur INPUT ... ; CARDS; Daten ;
SAS-System-File SET SAS-dateiname;	

+ zusätzliche Anweisungen

Programmbeispiele: Eingabe... .sas

Dateneingabe und Transformation

2.1 Eingabe über die Tastatur

```
DATA Eingabe1;
INPUT a $ x y z;
s = x + y + z;
CARDS;
b 1 2 3
c 4 5 6
d 7 8 9 ;
RUN;
/* Erläuterung dazu: siehe Datei Eingabe.sas. */
PROC PRINT; RUN;
```

Mit PROC PRINT wird die gesamte erzeugte Datei ausgedruckt ins Output-Fenster.

Dateneingabe und Transformation

Aktivierung des Programms

- klicken auf MännchenLogoGrafik oder
- klicken auf ’run’ $\rightarrow$ ’submit’ oder
- F3-Taste

Die Datei Eingabe1 hat

3 Beobachtungen (Individuen, Wertesätze)
5 Merkmale (Variablen) a, x, y, z und s.

Dateneingabe und Transformation

Alternative Besichtigung der Daten

Solutions

Analysis

Interactive Data Analysis

je nach DATA-Kommando:

	Bibliothek	Dateiname
DATA Eingabe1;	WORK	Eingabe1
DATA sasuser.Eing1;	SASAUSER	Eing1
DATA;	WORK	DATA1 DATA2 ...

Bemerkung:

Dateien, die sich im Arbeitsverzeichnis WORK befinden, werden am Ende der Sitzung gelöscht.

Die Variante “DATA sasuser.Eing1;” nicht verwenden.

Dateneingabe und Transformation

Automatisch generierte Variablen

$\_N\_$ oder obs

gibt die aktuelle Beobachtungsnummer an.

$\_ERROR\_$

- Nichtzulässige mathematische Operationen führen zu $\_ERROR\_ = 1$ und das Ergebnis wird auf "." (missing value) gesetzt. (vgl. Beispiel Eingabe2)
- Schlimmere Fehler führen zu höherem $\_ERROR\_$-Wert.

2.2 Dateneingabe und Transformation

Transformationen

- immer nach der INPUT-Anweisung angeben!

IF THEN ELSE und logische Operationen

vgl. Programm Eingabe2

Funktionen

vgl. Programm Eingabe3

Arithmetische Operationen

+,-,*,/,**

IF(log. Ausdruck)

nur bestimmte Wertesätze einlesen

Es werden nur die Wertesätze eingelesen, die die logische Bedingung erfüllen.

Dateneingabe und Transformation
IF THEN ELSE

jeweils nur eine Anweisung ausführen

IF (log. Ausdruck) THEN Anweisung;
ELSE Anweisung;

jeweils mehrere Anweisungen ausführen

- IF (log. Ausdruck) THEN Anweisung;
  ELSE DO
  Anweisung1; Anweisung2; ... END;
- IF (log. Ausdruck) THEN DO
  Anweisung1; ... END;
  ELSE DO
  Anweisung1; Anweisung2; ... END;

2.3 Eingabe durch externes File (ASCII)

DATA Eingabe4;
INFILE ’Pfadname’;
INPUT Variablen;
evtl. Transformationen;
RUN;

- Diese Eingabe ist formatfrei, d.h. die Variablen sind im Rohdatenfile durch Leerzeichen getrennt.
- Sind die Eingabedaten durch ein anderes Zeichen, z.B. ‘;’, getrennt, dann ist in der INFILE-Anweisung die Option DELIMITER=’;’ (oder DLM=’;’) anzugeben.
  Tabulatorzeichen: `DLM='09'X;`

- Bedingungen:
  fehlende Werte: . (Punkt)
  alphanumerische Zeichenketten dürfen keine Leerzeichen enthalten.
- Die INPUT-Anweisung kann auch abgekürzt werden, z.B.
  INPUT V1-V7;

Eingabe durch externes File (EXCEL)

```
PROC IMPORT datafile=”... .xls”;
out Dateiname; /*SAS-Datei*/
getnames=no; /*Variablennamen werden nicht
                                    übernommen*/
sheet=spreadsheetname;
RUN;
```

2.4 Wichtige Varianten der INPUT-Anweisung

- bisher: formatfrei
  INPUT a $ b $ c d;
- formatiert-spaltenorientiert
  INPUT a $ 1-10 b $ 11 c 13-14 .1;
- formatiert-über die Zeichenlänge
  INPUT a $10. b $ 1. c 2. d 5.1;

Eingabeformate

w.	2.	standard numerisch
w.d	2.1	standard numerisch mit Dezimalstelle
$w.	$10	Zeichenlänge

Nachgestelltes $-Zeichen steht für Zeichenketten.

Eingabe5

Eingabe6 (komplexere Formate)

Weitere Formatierungselemente

Spaltenzeiger

@n: Zeige auf Spalte n (z.B. @12)
+n: Setze den Zeiger n Positionen weiter

Zeilenzeiger

n: Zeige auf Spalte 1 der n-ten Zeile

Zeilenhalter

@ (nachgestellt)	Datenzeile wird von mehreren INPUT-Anweisungen gelesen
@@ (nachgestellt)	Aus einer Eingabezeile werden mehrere Beobachtungen gelesen

2.5 Ein- u. Ausgabe von SAS-Files

Abspeichern einer permanenten SAS-Datei

```
DATA sasuser.banknote; /* Eine Datei mit
    dem Namen 'banknote' wird im SAS-internen
       Verzeichnis 'sasuser' gespeichert */
<INFILE ' Pfadname der einzulesenden Datei;>
    INPUT Formatangaben;
    <CARDS;
       Daten (zeilenweise); >
RUN;
```

Einlesen einer SAS-Datei

```
DATA banknote1;
       SET sasuser.banknote < (Optionen)>;
RUN;
```

Ein- u. Ausgabe von SAS- Files

Einige Optionen

DROP = Varname(n);	Weglassen von Variablen
KEEP = Varname(n);	nur diese Variablen werden verwendet
FIRSTOBS=integer;	1. zu verarbeitender Wertesatz
OBS = integer;	letzter zu verarbeitender Wertesatz

RENAME = (alter Varname = neuer Varname);

Ausgabe

Formatierte Ausgabe

```
DATA;
Pi=3.141592;
FORMAT Pi 5.3;
OUTPUT;
STOP;
RUN;
```

Standard: 8 Zeichen.

Längere Variablennamen

vor die INPUT-Anweisung:

```
LENGTH Var.name $länge;
```

z.B. LENGTH Var.name $12;

2.6 Zusammenfügen von Files

Files ‘untereinander’

SASfile_1

...

SASfile_n

DATA; /* `Eingabe_Banknote13.sas` */
SET SASfile_1 <(options)>
... SASfile_n<(options)>;
RUN;

Files ‘nebeneinander’

SASfile_1 ... SASfile_n

DATA; /* `Eingabe_Banknote34.sas` */
SET SASfile_1; SET SASfile_2;
... SET SASfile_n; **RUN**;

Sortieren und Zusammenfügen von Dateien

Sortieren von Dateien

PROC SORT DATA=SASfile; BY nr; **RUN**;
nr gibt das Merkmal an, nach dem sortiert werden soll.

Zusammenfügen von Dateien

MERGE SASfile_1 SASfile_2; BY nr; **RUN**;
Die Dateien müssen nach dem Merkmal nr sortiert sein!
Wie bei SET sind auch hier Optionen möglich.

2.7 Output-Anweisung

- dient der Ausgabe von Dateien
- es können mehrere Dateien gleichzeitig ausgegeben werden
- die Namen der auszugebenden Dateien erscheinen im DATA-Step.

`Eingabe12.sas`

2.8 DO-Schleifen im DATA-Step

Allgemeine Syntax

- DO Indexvariable = Anfangswert
  <TO Endwert> <BY Schrittweite>;
  END;
- DO WHILE (Ausdruck) | UNTIL (Ausdruck);

`Eingabe13.sas`
`Eingabe14.sas`
`Eingabe15.sas`

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung (1)
- Erwartungswert
- Varianz
- Normalverteilung (2)

3. Wahrscheinlichkeitsrechnung

3.1 Grundbegriffe

Eine Grundgesamtheit (oder Population)

ist eine Menge von Objekten, die gewissen Kriterien genügen.
Die einzelnen Objekte heißen Individuen.

- Menge aller Haushalte
- Menge aller Studenten
- Menge aller Studenten der HUB
- Menge aller Einwohner von GB
- Menge aller Heroin-Abhängigen
- Menge aller Bewohner Tibets
- Menge aller verschiedenen Computer
- Menge aller Schweizer Franken

Grundbegriffe
Zufällige Stichprobe

Die gesamte Population zu erfassen und zu untersuchen ist meist zu aufwendig, deshalb beschränkt man sich auf zufällige Stichproben.

Zufällige Stichprobe

Eine zufällige Stichprobe ist eine zufällige Teilmenge der Grundgesamtheit, bei der jedes Element mit ‘der gleichen Wahrscheinlichkeit’ ausgewählt wird.

Grundbegriffe

Klassifikation von Merkmalen

Nominale Merkmale

Die Ausprägungen sind lediglich Bezeichnungen für Zustände oder Sachverhalte.
Sie können auch durch Zahlen kodiert sein!

Bsp: Familienstand, Nationalität, Beruf

Dichotome Merkmale

Hat das (nominale) Merkmal nur 2 Ausprägungen, so heißt es auch binär oder dichotom.

gut - schlecht
männlich - weiblich
wahr - falsch

Klassifikation von Merkmalen

Ordinale und metrische Merkmale

Ordinale Merkmale (Rangskala)

Die Menge der Merkmalsausprägungen besitzt eine Rangordnung!

Rangzahlen einer Rangliste (z.B. beim Sport)
Härtegrade
Schulzensuren

Metrische Merkmale (kardinale/quantitative M.)

Werte können auf der Zahlengeraden aufgetragen werden (metrische Skala)

Meßwerte, Längen, Größen, Gewichte, Alter

Klassifikation von Merkmalen

Metrische Merkmale

Metrische Merkmale werden unterschieden nach:

Diskrete Merkmale

nehmen höchstens abzählbar viele Werte an.

Alter, Länge einer Warteschlange

Stetige Merkmale

können Werte in jedem Punkt eines Intervalls annehmen, z.B. $x \in [a, b]$, $x \in (-\infty, \infty)$.

Metrische Merkmale sind immer auch ordinal.

Grundbegriffe

Stichprobenraum

Der Stichprobenraum Ω eines zufälligen Experiments

ist die Menge aller möglichen Versuchsausgänge
Die Elemente ω des Stichprobenraums Ω heißen Elementarereignisse.

- Münzwurf $\Omega = \{Z, B\}$
- Würfel $\Omega = \{1, ..., 6\}$
- Qualitätskontrolle $\Omega = \{$gut, schlecht$\}$
- Lebensdauer einer Glühlampe $\Omega = [0, \infty)$
- 100m - Zeit $\Omega =$[9.81, 20)
- Blutdruck, Herzfrequenz
- Länge einer Warteschlange $\Omega = \{0, 1, 2, ...\}$
- Anzahl der radioaktiven Teilchen beim Zerfall
- Wasserstand eines Flusses $\Omega = [0, \ldots)$

Grundbegriffe

Ein Ereignis ist eine Teilmenge $A, A \subseteq \Omega$

Lebensdauer ≤ 10 min.
Augensumme gerade.
Warteschlange hat Länge von ≤ 10 Personen.

Realisierungen sind die Ergebnisse des Experiments

(die realisierten Elemente von Ω)

Verknüpfungen von Ereignissen werden durch entsprechende Mengenverknüpfungen beschrieben

$A \cup B$	A oder B tritt ein
$A \cap B$	A und B tritt ein
$\overline{A} = \Omega \setminus A$	A tritt nicht ein.

Grundbegriffe
Ereignisfeld

Forderung (damit die Verknüpfungen auch immer ausgeführt werden können):
Die Ereignisse liegen in einem Ereignisfeld (σ-Algebra) $\mathfrak{E}$.

Ereignisfeld

Das Mengensystem $\mathfrak{E} \subseteq \mathfrak{P}(\Omega)$ heißt Ereignisfeld, falls gilt:
1. $\Omega \in \mathfrak{E}$
2. $A \in \mathfrak{E} \Longrightarrow \overline{A} \in \mathfrak{E}$
3. $A_i \in \mathfrak{E}, i = 1, 2, \ldots \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathfrak{E}$.

3.2 Wahrscheinlichkeit

Das Axiomsystem von Kolmogorov

Sei $\mathfrak{E}$ ein Ereignisfeld. Die Abbildung

$$P : \mathfrak{E} \longrightarrow \mathbb{R}$$

heißt Wahrscheinlichkeit, falls sie folgende Eigenschaften hat:

1. Für alle $A \in \mathfrak{E}$ gilt: $0 \leq P(A) \leq 1$.
2. $P(\Omega) = 1$.
3. Sei A_i eine Folge von Ereignissen, $A_i \in \mathfrak{E}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

falls $A_i \cap A_j = \emptyset \quad \forall i, i \neq j$

Wahrscheinlichkeit

Eigenschaften (1)

$P(\overline{A}) = 1 - P(A)$.

Beweis:

$$\begin{aligned} 1 &= P(\Omega) && \text{Axiom 2} \\ &= P(A \cup \overline{A}) \\ &= P(A) + P(\overline{A}) && \text{Axiom 3} \end{aligned}$$

Wahrscheinlichkeit

Eigenschaften (2)

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

Beweis:

$$\begin{aligned}
P(A \cup B) &= P((A \cap B) \cup (A \cap \overline{B}) \cup (B \cap \overline{A})) \\
&= \underbrace{P(A \cap B) + P(A \cap \overline{B})}_{} \\
&\quad + P(B \cap \overline{A}) \quad \text{Axiom 3} \\
&= P(A) + \underbrace{P(B \cap \overline{A}) + P(A \cap B)}_{} \\
&\quad - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

3.3 Zufallsvariablen

Eine (meßbare) Abbildung heißt Zufallsvariable.

$$\begin{aligned} X: \quad \Omega &\longrightarrow \mathbb{R} \\ \omega &\longrightarrow r \end{aligned}$$

Diskrete Zufallsvariable

Die Zufallsvariable X heißt diskret, wenn X nur endlich viele oder abzählbar unendlich viele Werte x_i annehmen kann. Jeder dieser Werte kann mit einer gewissen Wkt. $p_i = P(X = x_i)$ auftreten. ($p_i > 0$)

- geografische Lage (N,O,S,W)
- Länge einer Warteschlange
- Anzahl der erreichten Punkte in der Klausur.

Stetige Zufallsvariable

Stetige Zufallsvariable

Die Zufallsvariable X heißt stetig, falls X beliebige Werte in einem Intervall (a, b), $[a, b]$, $\overline{(a, b]}$, $(a, b]$, $(-\infty, a)$, (b, ∞), $(-\infty, a]$, $[b, \infty)$, $(-\infty, \infty)$ annehmen kann.

- Wassergehalt von Butter
- Messgrößen (z.B. bei der Banknote)
- Lebensdauer von Kühlschränken

Verteilungsfunktion

Diskrete Zufallsvariable

$$F_X(x) := P(X \leq x) = \sum_{i:i \leq x} p_i = \sum_{i=0}^{x} p_i$$

heißt Verteilungsfunktion der diskreten zufälligen Variable X

Stetige Zufallsvariable

Die Zufallsvariable X wird mit Hilfe der sogen. Dichtefunktion f beschrieben,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

3.4 Diskrete Zufallsvariablen

Bezeichnung

$$X \in \{x_1, x_2, x_3, \ldots\}$$

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots p_n & \cdots \end{pmatrix}$$

$$p_i = P(X = x_i) > 0, \quad i = 1, 2, 3, \ldots$$

$$\sum_{i=1}^{\infty} p_i = 1$$

Diskrete Zufallsvariablen

Beispiele

Zweimaliges Werfen einer Münze

$\Omega = \{ZZ, ZB, BZ, BB\}$, $X :=$ Anzahl von Blatt

$$X : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Erfolge bei n Versuchen

X: Anzahl der “Erfolge” bei n Versuchen, wobei jeder der n Versuche eine Erfolgswahrscheinlichkeit p hat.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Binomialwkt.}$$

$$F_X(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \quad \text{Vf.}$$

Diskrete Zufallsvariablen

Übungsaufgabe

Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

X: Anzahl der Sechsen.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - \sum_{i=0}^{3} P(X = i)$$

$$= 1 - \left(\frac{5}{6}\right)^{20} - 20\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{19} - \frac{20 \cdot 19}{2}\left(\frac{1}{6}\right)^{2}\left(\frac{5}{6}\right)^{18} -$$

$$- \frac{20 \cdot 19 \cdot 18}{6}\left(\frac{1}{6}\right)^{3}\left(\frac{5}{6}\right)^{17}$$

= 1 - CDF('Binomial',3,$1/6$,20)

= SDF('Binomial',3,$1/6$,20)

$\approx$ 0.43.

Diskrete Zufallsvariablen

Poisson (1)

X: Anzahl der Anrufe pro Zeiteinheit

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \end{pmatrix}$$

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \quad \lambda > 0$$

$$\sum_{i=0}^{\infty} p_i = \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}}_{e^{\lambda}} e^{-\lambda} = 1.$$

Bez.: $X \sim Poi(\lambda)$, wobei λ ein noch unbestimmter Parameter ist. Er kann als mittlere Rate aufgefaßt werden.

Diskrete Zufallsvariablen

Poisson (2), Motivation

Sei $\{N_t\}_{t \in T}$ eine Menge von Zufallsvariablen (ein stochastischer Prozeß) mit den Eigenschaften:

- V1: Zuwächse sind unabhängig, dh. die Zufallsvar. $N_{t+h} - N_t$ und $N_t - N_{t-h}$ sind unabhängig
- V2: es ist egal wo wir das Zeitintervall betrachten, dh. N_{t+h} und N_t haben dieselbe Verteilung
- V3: Wkt., daß mindestens ein Ereignis in der Zeit h eintritt, z.B. ein Kunde ankommt. $p(h) = a \cdot h + o(h), \quad a > 0, h \to 0$
- V4: Wkt. für $k \geq 2$ Ereignisse in der Zeit h: $o(h)$

Diskrete Zufallsvariablen

Poisson (3)

Frage: Wkt. bis zum Zeitpunkt t genau i Ereignisse?
(eingetroffene Kunden, zerfallene Teilchen)

$$P_k(t) := P(N_t = k), \qquad P_k(t) = 0 \quad \text{für} \quad k < 0$$

$$P_k(t) = \frac{a^k t^k}{k!} e^{-at}, \qquad k \geq 0$$

Poisson-Verteilung mit Parameter $\lambda = at$.

Beweis: Stochastik-Vorlesung.

Diskrete Zufallsvariablen

Poisson (4)

Binomial und Poisson

Seien $X_n \sim Bi(n,p)$ $\quad Y \sim Poi(\lambda)$

Für $n \cdot p = \lambda$ gilt: $\quad P(X_n = k) \longrightarrow_{n\to\infty} P(Y = k)$.

Beweis:

$$
\begin{aligned}
P(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\
&= \frac{1}{k!} \underbrace{\frac{n(n-1)\cdots(n-k+1)}{(n-\lambda)^k}}_{\to 1} \lambda^k \underbrace{(1-\frac{\lambda}{n})^n}_{\to e^{-\lambda}}
\end{aligned}
$$

Diskrete Zufallsvariablen

Geometrische Verteilung

Münzwurf solange bis B(Blatt) kommt

Ω = {B, ZB, ZZB, ...}

X := Anzahl der Würfe bis zum ersten Blatt.

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ (1/2) & (1/2)^2 & (1/2)^3 & (1/2)^4 & \cdots & (1/2)^n & \cdots \end{pmatrix}$$

$$\sum_{i=1}^{\infty} p_i = \sum_{i=1}^{\infty} (1/2)^i = \frac{1}{1-\frac{1}{2}} - 1 = 1$$

geometrische Reihe

geometrische Verteilung mit p=1/2, $p_i = (1/2)^i$.

allgemeiner: $p_i = p^{i-1}(1-p)$.

Diskrete Zufallsvariablen

Hypergeometrische Verteilung (1)

Qualitätskontrolle

Warenlieferung mit N Stücken, davon genau n schlecht. Frage: Wkt., daß in einer Stichprobe vom Umfang m höchstens k Stück schlecht sind?

X: Anzahl der schlechten Stücke in der Stichprobe.

$$P(X = k) = \frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}}$$

$\binom{N}{n}$: $\#$ möglichen Stichproben.

$\binom{n}{k}$: $\#$ Möglichkeiten, aus n schlechten Stücken in der Population k schlechte Stücke zu ziehen.

$\binom{N-n}{m-k}$: $\#$ Möglichkeiten, aus $N-n$ guten Stücken in der Population $m-k$ gute Stücke zu ziehen.

Diskrete Zufallsvariablen

Hypergeometrische Verteilung (2)

Offenbar:

$0 \leq x \leq \min(n, m)$

$m - x \leq N - n$.

Eine Zufallsvariable mit der Verteilungsfunktion

$$F(k|H_{N,n,m}) = \sum_{x=0}^{k} \frac{\binom{n}{x} \cdot \binom{N-n}{m-x}}{\binom{N}{m}}$$

heißt hypergeometrisch verteilt.

Bemerkung: Für $N \to \infty$, $n \to \infty$, $\frac{n}{N} \to p$ gilt:

$$f(x|H_{N,n,m}) \to \binom{m}{x} p^x (1-p)^{m-x} = f(x|Bi(m,p))$$

SAS-Anweisungen

CDF('Binomial',m,p,n)	PDF('Binomial',m,p,n)
CDF('Poisson',m,λ)	PDF('Poisson',m,λ)
CDF('Geometric',m,p)	PDF('Geometric',i,p)
CDF('Hyper',K,N,n,m)	PDF('Hyper',k,N,n,m)

```
Descr_Binomial_neu.sas
Descr_Poisson.sas
Descr_Geometr.sas
Descr_Hypergeom.sas
```

In den Wahrscheinlichkeiten können Parameter auftreten, die in der Regel unbekannt sind.

Die Parameter sind anhand der Beobachtungen (der Daten) zu bestimmen/zu schätzen!

$\longrightarrow$ Aufgabe der Statistik

3.5 Stetige Zufallsvariablen

Sei X stetig auf (a,b), wobei a, b unendlich sein können,
$a \leq x_0 < x_1 \leq b$
$P(X = x_0) = 0, \quad P(x_0 < X < x_1) > 0$ (wenn $f > 0$).

Die Funktion f heißt Dichtefunktion (von X) falls:

1. $f(x) \geq 0, \quad a < x < b$.
2. $\int_a^b f(x)\,dx = 1$.

Die stetige Zufallsvariable X wird also durch seine Dichtefunktion beschrieben.

$$P(c < X < d) = \int_c^d f(x)\,dx.$$

Die Dichtefunktion hängt i.A. von unbekannten Parametern ab, die geschätzt werden müssen.

Beispiele

Gleich- und Exponentialverteilung

Gleichverteilung auf [a,b], $X \sim R(a,b)$, $a < b$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{falls } a \leq x \leq b \text{ ,} \\ 0 & \text{sonst.} \end{cases}$$

- Referenzverteilung - Zufallszahlen

Exponentialverteilung, $X \sim Exp(\lambda)$, ($\lambda > 0$)

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{falls } x \leq 0 \\ 1 - e^{-\frac{x}{\lambda}} & \text{falls } x > 0. \end{cases}$$

- Lebensdauer - Zeitdauer zwischen Ankünften

Beispiele

Exponentialverteilung (2)

Gedächtnislosigkeit

Eine Verteilung P (mit Verteilungsfunktion F) heißt gedächtnislos, wenn für alle $s, t \geq 0$, gilt:

$$P(X \geq s + t | X \geq t) = P(X \geq s).$$

Es gilt (Definition der bedingten Wahrscheinlichkeit)

$$\begin{aligned} P(X \geq s + t | X \geq t) &= \frac{P(\{X \geq s + t\} \cap \{X \geq t\})}{P(X \geq t)} \\ &= \frac{P(X \geq s + t)}{P(X \geq t)}. \end{aligned}$$

Gedächtnislosigkeit

Cauchy-Funtionalgleichung

Eine Verteilung ist also gedächtnislos, gdw.

$$\frac{P(X \geq s+t)}{P(X \geq t)} = P(X \geq s)$$

bzw.

$$\frac{1-F(s+t)}{1-F(t)} = 1 - F(s).$$

Überlebensfunktion (oder Zuverlässigkeitsfunktion)

$$G(t) = 1 - F(t)$$

Die Vf. F (mit der Überlebensfunktion G) ist also gedächtnislos gdw

$$G(s+t) = G(s) \cdot G(t) \quad \text{für alle} \quad s, t \geq 0$$

Cauchy-Funktionalgleichung

Eine Lösung

Satz: Die Exponentialverteilung ist gedächtnislos.

Beweis: Die Verteilungsfunktion ist (sei $\lambda' := \frac{1}{\lambda}$)

$$F(t) = P(X < t) = \begin{cases} 1 - e^{-\lambda' t} & \text{falls} \quad t \geq 0 \\ 0 & \text{sonst,} \end{cases}$$

und die Überlebensfunktion

$$G(t) = 1 - F(t) = 1 - (1 - e^{-\lambda' t}) = e^{-\lambda' t}.$$

Folglich erhalten wir

$$G(s+t) = e^{-\lambda'(s+t)} = e^{-\lambda' s} e^{-\lambda' t} = G(s) \cdot G(t).$$

Cauchy-Funktionalgleichung

Die einzige Lösung

Satz:

Sei F eine stetige Verteilungsfunktion mit
$F(0) = 0$ und $G(t) = 1 - F(t)$.
Es gelte die Cauchy-Funktionalgleichung

$$G(s + t) = G(s) \cdot G(t) \quad \text{für alle} \quad s, t \geq 0.$$

Dann gilt für alle t, $t > 0$,

$$F(t) = 1 - e^{-\lambda t},$$

wobei $\lambda > 0$. D.h. F ist Exponential-Verteilungsfunktion.

Beweis: Stochastik-Vorlesung.

Beispiele

Normalverteilung (NV)

Dichtefunktion und Verteilungsfunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)} \tag{1}$$

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \tag{2}$$

$(-\infty < x < \infty), \quad -\infty < \mu < \infty, \sigma^2 > 0.$

Bez.: $X \sim N(\mu, \sigma^2)$
μ: Lageparameter, σ: Skalenparameter
NV: wichtigste Verteilung in der Statistik
warum? $\longrightarrow$ später.

SAS-Anweisungen

PDF('Exponential',x,λ)	Dichtefkt.
CDF('Exponential',x,λ)	Verteilungsfkt.
PDF('Normal',x,μ, σ)	Dichtefunktion
CDF('Normal',x,μ, σ)	Verteilungsfkt.
PROBNORM(x,μ, λ)	
Quantile('Normal',u,μ,σ)	Quantilfkt.
PROBIT(u,μ,σ)	

Stetige Zufallsvariablen

Weitere wichtige Verteilungen

Weibull-Verteilung	CDF('Weibull',x,a,λ)
Gamma-Verteilung	CDF('Gamma',x,a,λ)
χ^2-Verteilung	CDF('Chisq',x,ν, λ)
t-Verteilung	CDF('t',x,ν, δ)
F-Verteilung	CDF('F',x,ν_1, ν_2, δ)

Die drei letzten Verteilungen werden vor allem bei statistischen Tests benötigt(später).

`Descr_Weibull`
`Descr_Gamma`

Wahrscheinlichkeitsverteilungen in SAS

↪ help
↪ SAS Help and Documentation
↪ SAS Products
↪ BASE SAS
↪ SAS Language Dictionary
↪ Dictionary of Language
↪ Functions and Call Routines
↪ CDF
↪ PDF
↪ Quantile

Wahrscheinlichkeitsverteilungen in SAS

CDF(’Verteilung’,x,Parameterliste) Verteilungsfkt.
PDF(’Verteilung’,x,Parameterliste) Df (Wkt.fkt.)
SDF (’Verteilung’,x,Parameterliste) = 1-CDF
Überlebensfunktion $(1 - F(x))$
Quantile(’Verteilung’,u,Parameterliste) Quantilfkt.

Verteilung: in der obigen Liste nachsehen (s. letzte Folie)

3.6 Normalverteilung (1)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)}$$

Gauß

Normalverteilung

Satz: f aus (1) ist Dichte.

Beweis: 1. $f(x) \geq 0 \; \forall x \in \mathbf{R}$ und $\sigma > 0$.

2. bleibt z.z.

$$\lim_{x \to \infty} F(x) = \int_{-\infty}^{\infty} f(t)\, dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt = 1.$$

Wir bezeichnen

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx =: I.$$

Normalverteilung

Wir betrachten zunächst:

$$
\begin{aligned}
I^2 &= \left(\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right)^2 \\
&= \frac{1}{2\pi\sigma^2} \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \right) \\
&= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right) e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\
&= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dx\, dy
\end{aligned}
$$

Normalverteilung

Substitution:

$$s := \frac{x-\mu}{\sigma} \qquad t := \frac{y-\mu}{\sigma}.$$

$$dx = \sigma\, ds \qquad dy = \sigma\, dt.$$

Wir erhalten damit:

$$\begin{aligned} I^2 &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}s^2} e^{-\frac{1}{2}t^2} \sigma^2\, ds\, dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(s^2+t^2)}\, ds\, dt \end{aligned}$$

Normalverteilung

Weitere Substitution (Polarkoordinaten):

$$s = r\cos\varphi \qquad t = r\sin\varphi.$$

Dann gilt allgemein nach der Substitutionsregel:

$$\int\int g(s,t)\,ds\,dt = \int\int g(r,\varphi)\det J\,dr\,d\varphi,$$

wobei hier:

$$\begin{aligned}
\det J = |J| &= \begin{vmatrix} \frac{\partial s}{\partial r} & \frac{\partial s}{\partial \varphi} \\ \frac{\partial t}{\partial r} & \frac{\partial t}{\partial \varphi} \end{vmatrix} \\
&= \begin{vmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{vmatrix} \\
&= r\cos^2\varphi + r\sin^2\varphi \\
&= r(\cos^2\varphi + \sin^2\varphi) = r
\end{aligned}$$

Normalverteilung

$$
\begin{aligned}
I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}(r^2 \cos^2 \varphi + r^2 \sin^2 \varphi)} r \, dr \, d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2} r^2} r \, dr \, d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} \left[-e^{-\frac{r^2}{2}} \right]_0^{\infty} d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} d\varphi = \frac{1}{2\pi} 2\pi = 1
\end{aligned}
$$

Normalverteilung

Standard-Normalverteilung

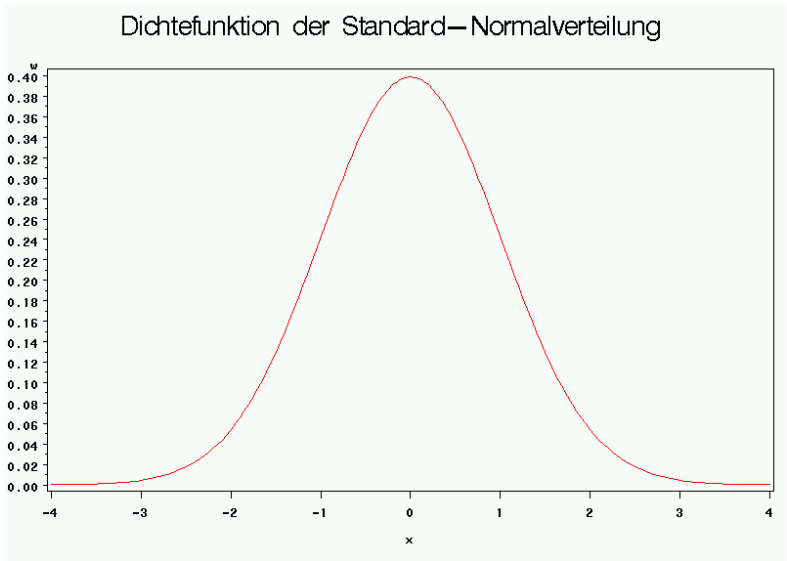
$\mu = 0, \quad \sigma^2 = 1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \qquad \text{Dichte}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \qquad \text{Verteilungsfunktion}$$

$\varphi(x), \Phi(x)$ sind tabelliert.
Es geht auch einfacher mit CDF und PDF.

Dichte der Standardnormalverteilung



$$\varphi(x) = \varphi(-x)$$
$$\Phi(x) = 1 - \Phi(-x)$$

`Programm: Descr_normal.sas`

Frage: Für welches x gilt: $\Phi(x) = \alpha$?

$x = \Phi^{-1}(\alpha)$ α-Quantil.

$\Phi^{-1}(\alpha)$ als Funktion: Quantilfunktion

SAS: QUANTILE('normal',α,0,1)

Normalverteilung

Beziehung zur Standard-Normalverteilung

Sei $X \sim N(0,1)$. Dann $P(a < X < b) = \Phi(b) - \Phi(a)$.

Satz. Es gilt:

$$\begin{aligned} X \sim N(0,1) &\iff \sigma X + \mu \sim N(\mu, \sigma^2) \\ X \sim N(\mu, \sigma^2) &\iff \alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2) \\ X \sim N(\mu, \sigma^2) &\iff \frac{X-\mu}{\sigma} \sim N(0,1) \end{aligned}$$

Beweis: Wir zeigen nur 1. ($\rightarrow$). Sei $X \sim N(0,1)$.

$$P(\sigma X + \mu \le x) = P(X \le \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}) =$$

$$= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)}\, du$$

Normalverteilung

Unterschiedliche Parameter (1)

Vergleichen Sie

a) σ^2 fest, μ verschieden
b) μ fest, σ^2 verschieden

`Descr_Normal_1.sas`

Normalverteilung

Unterschiedliche Parameter (1)

Satz:

Seien $X_1 \sim N(\mu, \sigma_1^2), X_2 \sim N(\mu, \sigma_2^2)$,
$\sigma_1^2 < \sigma_2^2$ und $a > 0$. Dann gilt:

$$P(\mu - a < X_1 < \mu + a) > P(\mu - a < X_2 < \mu + a).$$

Beweis:

$$\begin{aligned} P(\mu - a < X_1 < \mu + a) &= P(\frac{-a}{\sigma_1} < \frac{X_1 - \mu}{\sigma_1} < \frac{a}{\sigma_1}) \\ &= \Phi(\frac{a}{\sigma_1}) - \Phi(-\frac{a}{\sigma_1}) \\ &> \Phi(\frac{a}{\sigma_2}) - \Phi(-\frac{a}{\sigma_2}) \\ &= P(\mu - a < X_2 < \mu + a). \end{aligned}$$

Normalverteilung

Beispiel: $X_1 \sim N(10,4), X_2 \sim N(10,9), a = 1.$

$$
\begin{aligned}
P(9 < X_1 < 11) &= \Phi(\frac{11-10}{2}) - \Phi(\frac{9-10}{2}) \\
&= \Phi(\frac{1}{2}) - \Phi(-\frac{1}{2}) \\
&= 2 \cdot \Phi(\frac{1}{2}) - 1 \\
&= 2 \cdot 0.6915 - 1 = 0.383.
\end{aligned}
$$

$$
\begin{aligned}
P(9 < X_2 < 11) &= \Phi(\frac{11-10}{3}) - \Phi(\frac{9-10}{3}) \\
&= \Phi(\frac{1}{3}) - \Phi(-\frac{1}{3}) \\
&= 2 \cdot \Phi(\frac{1}{3}) - 1 \\
&= 2 \cdot 0.6306 - 1 = 0.26112.
\end{aligned}
$$

Wahrscheinlichkeitsverteilungen

Zusammenfassung (1)

Diskrete Verteilungen

Binomial $X \sim B(n, p)$

X : Anzahl von “Erfolgen”, n Versuche, Erfolgswkt. p.

Poisson $X \sim Poi(\lambda)$

X : Anzahl von “Erfolgen”, n Versuche, Erfolgswkt. p,
n groß und p klein, $n \cdot p = \lambda$.
X : $\#$ Ankünfte in einem Zeitintervall.

Geometrisch, $X \sim Geo(p)$

X :: Zahl der Versuche bis zum ersten “Erfolg”.

Wahrscheinlichkeitsverteilungen

Zusammenfassung (2)

Stetige Verteilungen

Gleichverteilung $X \sim R(a, b)$

Zufallszahlen

Exponential $X \sim Exp(\lambda)$

“gedächtnislose” stetige Verteilung.

Normal $X \sim N(\mu, \sigma^2)$

Zentraler Grenzwertsatz
Fehlergesetz (viele kleine unabh. Fehler)

3.7 Erwartungswert

Einleitende Motivation

Eine Münze wird 3 mal geworfen.
Wie oft können wir erwarten, daß Blatt oben liegt?
Wie oft wird im Mittel Blatt oben liegen?

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1/8 & 3/8 & 3/8 & 1/8 \end{pmatrix}$$

Erwartungswert:
$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5$
D.h. bei 10maliger Durchführung des Experiments können wir im Mittel mit 15mal Blatt rechnen!

Erwartungswert

Diskrete Zufallsvariable

Sei X diskrete Zufallsvariable

$$X : \begin{pmatrix} x_1 & \dots & x_n & \dots \\ p_1 & \dots & p_n & \dots \end{pmatrix}$$

$$\mathbf{E}X = \sum_{i=1}^{\infty} p_i x_i$$

heißt Erwartungswert von X.

Erwartungswert

$X \sim \text{Poisson}(\lambda)$

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \dots \\ p_0 & p_1 & p_2 & p_3 & \dots \end{pmatrix} \qquad p_i = \frac{\lambda^i}{i!} e^{-\lambda}$$

$$\begin{aligned} \mathbf{E}X &= \sum_{i=0}^{\infty} p_i i \\ &= \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \cdot i \\ &= \lambda \underbrace{\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}}_{e^{\lambda}} e^{-\lambda} = \lambda. \end{aligned}$$

z.B. mittlere Ankunftsrate.

Erwartungswert

$X \sim Bi(n,p)$

$$
\begin{aligned}
\mathbf{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k \cdot (1-p)^{n-k} \\
&= p \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= p \cdot n \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= p \cdot n \sum_{i=0}^{n-1} \binom{n-1}{i} p^{i} (1-p)^{n-1-i}, \qquad k = i+1 \\
&= n \cdot p.
\end{aligned}
$$

Erwartungswert

Stetige Verteilung

Sei X stetig mit Dichte f. Die Größe

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

heißt Erwartungswert von X.

$X \sim Exp(\lambda), \qquad \lambda > 0$

$$\mathbf{E}X = \int_{0}^{\infty} x \cdot \frac{1}{\lambda} \cdot e^{-\frac{x}{\lambda}} dx = \lambda$$

Erwartungswert

Normalverteilung

$X \sim N(\mu, \sigma^2)$

$$
\begin{aligned}
\mathbf{E}X &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-(\frac{x-\mu}{\sigma})^2/2} dx \\
&= \int_{-\infty}^{\infty} (\sigma t + \mu) \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \\
&= \mu + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma \cdot t \cdot e^{\frac{-t^2}{2}} dt \\
&= \mu.
\end{aligned}
$$

$\frac{x-\mu}{\sigma} = t, \quad dx = \sigma dt$

Erwartungswert

Gleichverteilung

$X \sim R(a, b)$, gleichverteilt auf dem Intervall (a,b)

$$\begin{aligned} \mathbf{E}X &= \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

Erwartungswert

Eigenschaften des Erwartungswertes

$\mathbf{E}$ ist Linearer Operator

$$\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y.$$

Seien X und Y stochastisch unabhängig. Dann

$$\mathbf{E}(X \cdot Y) = \mathbf{E}X \cdot \mathbf{E}Y.$$

Regel des Faulen Statistikers

Sei X Zufallsvariable, g: $R \longrightarrow R$ (rechtsseitig) stetig $\Rightarrow$

$$\mathbf{E}(g(X)) = \begin{cases} \sum_{i=0}^{\infty} g(x_i)p_i & \text{, falls X diskret} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{, falls X stetig,} \end{cases}$$

vorausgesetzt die Erwartungswerte existieren.

3.8 Die Varianz (Streuung)

Definition

Ang., die betrachteten Erwartungswerte existieren.

$$var(X) = \mathbf{E}(X - \mathbf{E}X)^2$$

heißt Varianz der Zufallsvariable X.

$$\sigma = \sqrt{Var(X)}$$

heißt Standardabweichung der Zufallsvariablen X.

Bez.: $var(X), Var(X), varX, \sigma^2, \sigma_X^2, \sigma, \sigma_X$.

Sei $\mu = \mathbf{E}X$.

Die Varianz

Stetige und diskrete Zufallsvariablen

Wenn X diskret, so gilt:

$$var(X) = \sum_{i=0}^{\infty} (x_i - \mu)^2 p_i$$

Wenn X stetig, so gilt:

$$var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx,$$

wobei f die Dichte von X ist.

$var(X)$: mittlere quadratische Abweichung von X und $\mathbf{E}X$.

Die Varianz

Eigenschaften der Varianz

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}(X - \mu)^2 = \\ &= \mathbf{E}(X^2 - 2\mu X + \mu^2) = \\ &= \mathbf{E}X^2 - \mu^2. \end{aligned}$$

$$var(aX + b) = a^2 var(X), \quad a, b \in \mathbb{R}.$$

$$var(X) = 0 \Longleftrightarrow \exists c : \quad P(X = c) = 1.$$

Die Varianz

Unabhängigkeit von Zufallsvariablen

Zwei Zufallsvariablen X und Y heißen unabhängig, falls

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$

für alle $x, y \in \mathbf{R}$.

Zwei Ereignisse A und B heißen unabhängig, falls

$$P(A, B) = P(A) \cdot P(B)$$

X und Y sind also unabhängig gdw. die Ereignisse $X \leq x$ und $Y \leq y$ unabhängig sind für alle $x, y \in \mathbf{R}$.

Seien X und Y unabhängig. Dann gilt

$$var(X + Y) = var(X) + var(Y).$$

Die Varianz

Poisson-Verteilung

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots$$

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 = \sum_{i=0}^{\infty} (i - \lambda)^2 p_i \\ &= \sum_{i=2}^{\infty} i \cdot (i-1) p_i + \sum_{i=0}^{\infty} i p_i - \\ &\quad 2\lambda \sum_{i=0}^{\infty} i p_i + \lambda^2 \sum_{i=0}^{\infty} p_i \\ &= e^{-\lambda} \lambda^2 \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} + \lambda - 2\lambda^2 + \lambda^2 = \lambda. \end{aligned}$$

Die Varianz

Binomialverteilung, $X \sim B(n,p)$

$$P(X=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

$$var(X) = np(1-p).$$

(ohne Beweis, ÜA)

Die Varianz

Gleichverteilung auf (a, b)

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a,b) \\ 0 & \text{sonst.} \end{cases} \qquad \mathbf{E}X = \frac{a+b}{2}.$$

$$\begin{aligned} \mathbf{E}X^2 &= \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{3}x^3\Big|_a^b \cdot \frac{1}{b-a} \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}. \\ var(X) &= \mathbf{E}X^2 - (\mathbf{E}X)^2 \\ &= \frac{1}{12}(4a^2+4ab+4b^2-3a^2 \\ &\quad -6ab-3b^2) \\ &= \frac{1}{12}(a^2-2ab+b^2) = \frac{(b-a)^2}{12}. \end{aligned}$$

Die Varianz

Exponentialverteilung

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases}$$

$$\begin{aligned} \mathbf{E}X &= \lambda. \\ \mathbf{E}X^2 &= \int_0^\infty x^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}}\, dx = 2 \cdot \lambda^2 \quad \text{(ÜA)}. \\ var(X) &= \lambda^2. \end{aligned}$$

Die Varianz

Normalverteilung $var(X) = \sigma^2$

$$
\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \\
\mathbf{E}(X-\mu)^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \\
&= \sigma^2 \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \\
&= \sigma^2 \int_{-\infty}^{\infty} (-t)(-t\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}})\, dt \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(-te^{-t^2/2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1) e^{-\frac{t^2}{2}}\, dt\right) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\, dt = \sigma^2.
\end{aligned}
$$

3.9 Normalverteilung (2)

Besondere Eigenschaften

(schwaches) Gesetz der Großen Zahlen

Seien X_i unabhängig, identisch verteilt, $EX_i = \mu$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow_p \mathbf{E}X$$

Zentraler Grenzwertsatz

Seien X_i unabhängig, identisch verteilt,
$EX_i = \mu, varX_i = \sigma^2$.

$$Z_n := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \rightarrow Z, \quad Z \sim N(0,1).$$

`Descr_Binomial_2.sas` `Descr_Exp.sas`

Normalverteilung

Fehlertheorie

Fehler sind unter folgenden Annahmen (asymptotisch) normalverteilt:

- Jeder Fehler ist Summe einer sehr großen Anzahl sehr kleiner, gleich großer Fehler, die verschiedene Ursachen haben.
- Die verschiedenen Fehlerkomponenten sind unabhängig.
- Jede Fehlerkomponente ist mit Wkt. 0.5 positiv und mit Wkt. 0.5 negativ.

Normalverteilung

Maximale Entropie

bei gegebenen

Mittelwert μ und Varianz σ^2.

f: Wkt.dichte auf $(-\infty, \infty)$.
$\int xf(x)\,dx = \mu, \quad \int (x-\mu)^2 f(x)\,dx = \sigma^2$
Entropie:

$$H(f) := -\int f(x) \log f(x)\,dx$$

ist zu maximieren unter den obigen Bedingungen.
$\Longrightarrow \quad f =$Normaldichte.

Literatur: Rao: Lineare Statistische Methoden, 3.a.1.

Normalverteilung

Die Summe normalverteilter Zufallsvariablen

Die Summe normalverteilter Zufallsvariablen ist normalverteilt.

Seien $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$. Dann

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2).$$

(ρ: Korrelationskoeffizient zwischen X_1 und X_2, s.u.)

Beweis: über charakteristische Funktionen (Fouriertransformationen der Dichte) oder über die Faltungsformel (Stochastik-Vorlesung).

- Statistische Maßzahlen für quantitative Merkmale
- Box-Plots
- Probability Plots
- Häufigkeitsdiagramme
- Häufigkeitstabellen
- Scatterplots, Zusammenhangsmaße
- Das Regressionsproblem

4. Beschreibende Statistik

4.1 Statistische Maßzahlen für quantitative Merkmale

4.1.1 Lagemaße
Mittelwert, Quantile, Median, Quartile, Modalwert

4.1.2 Eigenschaften von Schätzungen

4.1.3 Streuungsmaße
Varianz, Standardabweichung, Spannweite, Quartilsabstand, MAD, Variationskoeffizient

4.1.4 Formmaße
Schiefe, Exzess, Wölbung, Kurtosis

Lagemaße (Lokationsparameter)

Das arithmetische Mittel

Die angegebenen Maßzahlen sind empirisch, d.h. sie sind Schätzungen für die wahre (i.A. unbekannte) Lage.

Mittelwert (MEAN)

$$\overline{X} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\overline{X}_n \to_{n\to\infty} \mathbf{E}X$ Gesetz der Großen Zahlen.

Voraussetzungen:

a) X_i i.i.d., $\mathbf{E}X_i < \infty$ (Chintchin) oder

b) X_i beliebig, $\mathbf{E}X_i^2 < \infty$ (Tschebycheff)

Lagemaße (2)
Quantile

Die Beobachtungen $x_1, ..., x_n$ werden der Größe nach geordnet:
$x_{(1)} \leq ... \leq x_{(n)}$.
Sei $0 \leq \alpha \leq 1, \quad \alpha \cdot n = \lfloor \alpha \cdot n \rfloor + r =: j + r$.

Quantile (Perzentile)

$$x_\alpha = \begin{cases} x_{(j+1)} \text{ für r} > 0 \\ 1/2(x_{(j)} + x_{(j+1)}) \text{ für r} = 0 \end{cases}$$

(empirisches) α-Quantil bzw. $\alpha \cdot 100\%$ Perzentil

mindestens $\lfloor \alpha \cdot n \rfloor$ der Werte $(x_1, ..., x_n)$ sind $\leq x_\alpha$
mindestens $\lfloor (1 - \alpha) \cdot n \rfloor$ der Werte $(x_1, ..., x_n)$ sind $\geq x_\alpha$

Vereinbarung: $x_0 = x_{(1)} \quad x_1 = x_{(n)}$

Quantile
Beispiel

$$\begin{array}{ccccccccc} x_{(1)} & < & x_{(2)} & < & x_{(3)} & < & x_{(4)} & < & x_{(5)} \\ 1.5 & < & 2.7 & < & 2.8 & < & 3.0 & < & 3.1 \end{array}$$

$\alpha = 0.25:$

$\alpha \cdot n = 0.25 \cdot 5 = 1.25 = 1 + 0.25$

$\rightarrow x_\alpha = x_{0.25} = x_{(2)} = 2.7$

$\alpha = 0.75:$

$\alpha \cdot n = 0.75 \cdot 5 = 3.75 = 3 + 0.75$

$\rightarrow x_\alpha = x_{0.75} = x_{(4)} = 3.0$

$\alpha = 0.5:$

$\alpha \cdot n = 0.5 \cdot 5 = 2.5 = 2 + 0.5$

$\rightarrow x_\alpha = x_{0.5} = x_{(3)} = 2.8$

Lagemaße (3)

Median
ist das 0.5-Quantil $x_{0.5}$.

Quartile
heißen die 0.25- und 0.75-Quantile $x_{0.25}$ und $x_{0.75}$.

Modalwert
häufigster Wert

theoretischer Modalwert:
diskrete Merkmale: der wahrscheinlichste Wert
stetige Merkmale: Wert mit der größten Dichte

Lagemaße (4)

- Der Mittelwert ist in vielen Fällen eine ’gute’ Lageschätzung, aber nicht robust (gegen Ausreißer).
- Der Median ist robust, aber meist nicht so ’gut’.

getrimmte Mittel, (α-)getrimmtes Mittel

$$\overline{X}_\alpha := \frac{X_{(\lfloor n\cdot\alpha\rfloor+1)} + \ldots + X_{(n-\lfloor n\cdot\alpha\rfloor)}}{n - 2\lfloor n\cdot\alpha\rfloor}, \qquad \alpha \in [0, \frac{1}{2})$$

Die $\lfloor n\cdot\alpha\rfloor$ kleinsten und $\lfloor n\cdot\alpha\rfloor$ größten Werte werden weggelassen und dann das arithmetische Mittel gebildet.

$\overline{X}_\alpha$ ist robuster als $\overline{X}$ und effizienter als $x_{0.5}$.

Lagemaße (5)

winsorisiertes Mittel, (α-)winsorisiertes Mittel

Sei $\alpha \in [0, \frac{1}{2})$ und jetzt $n_1 := \lfloor n \cdot \alpha \rfloor + 1$.

$$\overline{X}_{\alpha,w} := \frac{n_1 X_{(n_1)} + X_{(n_1+1)} + \ldots + X_{(n-n_1)} + n_1 X_{(n-n_1+1)}}{n}$$

Die $\lfloor n \cdot \alpha \rfloor$ kleinsten und $\lfloor n \cdot \alpha \rfloor$ größten Werte werden "herangeschoben" und dann das arithmetische Mittel gebildet.

- winsorisiertes Mittel ist robuster als $\overline{X}$ und effizienter als $x_{0.5}$.

Empfehlung für $\overline{X}_\alpha, \overline{X}_{\alpha,w}$: $\alpha : 0.1 \quad \ldots \quad 0.2$.

Lageschätzungen mit SAS

Mittelwert:	PROC MEANS;
Median:	PROC MEANS MEDIAN;
	PROC UNIVARIATE;
getrimmte Mittel:	PROC UNIVARIATE
	TRIMMED=Zahl;
winsorisierte Mittel:	PROC UNIVARIATE
	WINSORIZED=Zahl;
Quartile:	PROC UNIVARIATE;
Modalwert:	PROC UNIVARIATE;
Quantile:	PROC UNIVARIATE;

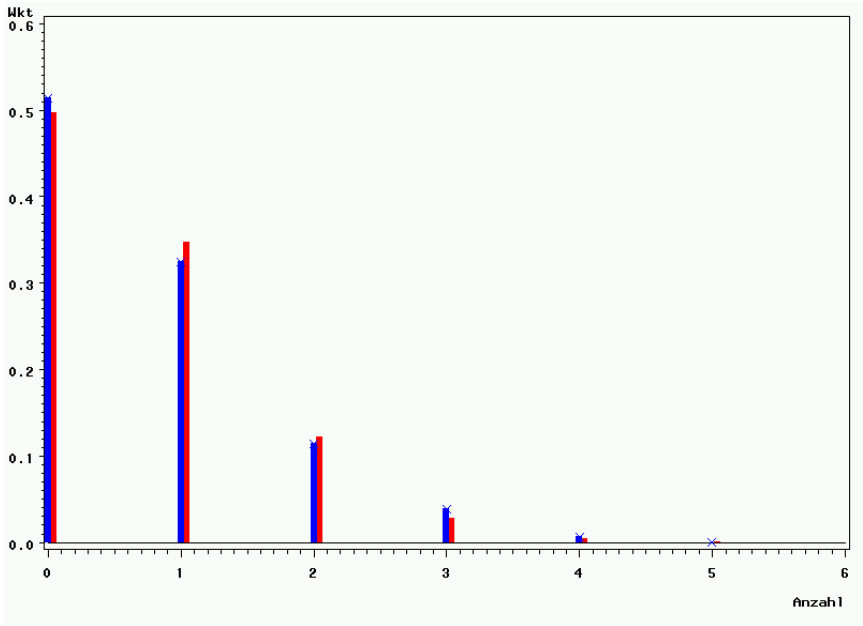
`Descr1.sas` `Mean.sas`

Beispiele (1)

Tödliche Unfälle durch Pferdetritte

14 Corps, 20 Jahre, insges. 280 Einheiten. Erfasst wurde für jede Einheit die Anzahl der tödlichen Unfälle durch Pferdetritte.

Anzahl	Häufigkeit
0	144
1	91
2	32
3	11
4	2
5	0



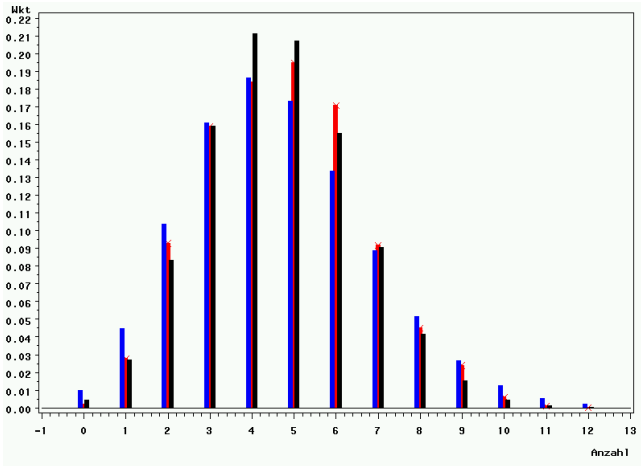
Poisson-Verteilung geeignet (?)

Beispiele (2)

Anzahl von schwarzen Feldern

Ein Zufallszahlengenerator soll zufällige Bildpunkte erzeugen, weiß mit Wkt. 0.71 und schwarz mit Wkt. 0.29.

Dazu wurde ein großes Quadrat in 1000 Teilquadrate mit je 16 Bildpunkten zerlegt. Gezählt wurde jeweils die Anzahl der schwarzen Bildpunkte.



n	0	1	2	3	4	5	6	7	8	9	10	11	12
h	2	28	93	159	184	195	171	92	45	24	6	1	0

Binomial-Verteilung (schwarz) geeignet (?)

Eigenschaften von Schätzungen (1)

Sei $\hat{\theta}_n$ eine Schätzung von θ, die auf n Beobachtungen beruht.

Konsistenz (Minimalforderung)

$\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$

Erwartungstreue, Asymptotische Erwartungstreue

$\mathbf{E}\hat{\theta}_n = \theta$
$\mathbf{E}\hat{\theta}_n \to_{n\to\infty} \theta$

“gute”, “effiziente” Schätzung

var $\hat{\theta}_n$ möglichst klein

Eigenschaften von Schätzungen (2)

optimale Schätzung

wenn var $\hat{\theta}_n$ den kleinstmöglichen Wert annimmt für alle e-treuen Schätzungen

Mean Square Error (MSE)

MSE = var $\hat{\theta}_n$ + bias^2 $\hat{\theta}_n$

= var $\hat{\theta}_n + (E\hat{\theta}_n - \theta)^2$

soll minimal oder möglichst klein sein.

robuste Schätzung

Eigenschaften sollten “möglichst” auch bei (kleinen) Abweichungen von der (Normal-) Verteilungsannahme gelten

Eigenschaften von Schätzungen (3)

Cramer-Rao Ungleichung

θ: zu schätzender Parameter einer Population (Dichte f).
$\hat{\theta} = \theta_n$: eine erwartungstreue Schätzung von θ.

Cramer-Rao-Ungleichung

$$var(\hat{\theta}) \geq \frac{1}{nI(f,\theta)},$$

Fisher-Information

$$\begin{aligned} I(f,\theta) &= \mathbf{E}\left(\frac{\partial \ln f(X,\theta)}{\partial\theta}\right)^2 \\ &= \int \left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)\, dx \end{aligned}$$

Die Varianz einer Schätzung kann, bei gegebenem Stochprobenumfang, nicht beliebig klein werden.

Eigenschaften von Schätzungen (4)

Beispiele

f normal

$$f(x,\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned}
\ln f(x,\mu) &= -\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \\
\frac{\partial \ln f(x,\mu)}{\partial \mu} &= \frac{x-\mu}{\sigma} \cdot \frac{1}{\sigma} \\
I(f,\mu) &= \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \cdot f(x,\mu)\, dx = \frac{1}{\sigma^2}.
\end{aligned}$$

Eigenschaften von Schätzungen (5)

Beispiele (2)

Nach der Cramer-Rao-Ungleichung gilt also für jede Lageschätzung

$$var(\hat{\theta}) \geq \frac{1}{nI(f,\theta)} = \frac{\sigma^2}{n},$$

insbesondere

$$var\overline{X} \geq \frac{\sigma^2}{n}.$$

Vergleichen Sie das mit:

$$var\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} varX_i = \frac{\sigma^2}{n}.$$

Bei Normalverteilung ist also $\overline{X}$ Lageschätzung mit minimaler Varianz.

Eigenschaften von Schätzungen (6)

Beispiele (3)

f exponential

$$f(x,\lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} & \text{falls} \quad x \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

$$I(f,\lambda) = \frac{1}{\lambda^2} \qquad \text{(ÜA, 2 P.)}$$

Die Cramer-Rao-Schranke ist also:

$$\frac{1}{nI(\lambda)} = \frac{\lambda^2}{n}.$$

Vergleichen Sie mit $\quad var\overline{X} = \frac{\lambda^2}{n}.$

Bei Exponentialverteilung ist also $\overline{X}$ Parameterschätzung mit minimaler Varianz.

Eigenschaften von Schätzungen (7)

Beispiele (4)

f Doppelexponential (=Laplace)

$$f(x,\lambda,\mu) = \frac{1}{2}\begin{cases} \frac{1}{\lambda}e^{-\frac{1}{\lambda}(x-\mu)} & \text{falls} \quad x \geq \mu \\ \frac{1}{\lambda}e^{\frac{1}{\lambda}(x-\mu)} & \text{falls} \quad x < \mu \end{cases}$$

Der hier interssierende (Lage-) Paramter ist μ.

$$I(f,\mu) = \frac{1}{\lambda^2}. \quad \text{(ÜA, 5 P.)} \qquad var(\overline{X}) = \frac{2\lambda^2}{n}. \quad \text{(ÜA, 2 P.)}$$

Für den Median $x_{0.5}$ gilt:

$$\text{var}(x_{0.5}) \sim \frac{\lambda^2}{n}. \quad \text{(ÜA, 10 P.)}$$

Streuungsmaße

Die angegebenen Maßzahlen sind empirisch, d.h. sie sind Schätzungen für die wahre Varianz

(empirische) Varianz (Streuung)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$s^2 \rightarrow_{n\rightarrow\infty} var(X)$$

Warum Division durch $(n-1)$: Erwartungstreue (ÜA)

Standardabweichung

$$s = \sqrt{s^2}$$

Streuungsmaße (2)

Spannweite (Range)

$$X_{(n)} - X_{(1)}$$

(Inter-)Quartilsabstand, IR

$$IR = x_{0.75} - x_{0.25}$$

Wenn $X \sim \mathcal{N}$ so $\mathbf{E}(IR/1.34898) = \sigma$.

Mittlere absolute Abweichung vom Median

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{0.5}|$$

Streuungsmaße (3)

Median absolute deviation, MAD

$$MAD = med(|X_i - x_{0.5}|)$$

Wenn $X \sim \mathcal{N}$ so $\mathbf{E}(1.4826 \cdot MAD) = \sigma$

Variationskoeffizient

$$CV = \frac{s \cdot 100}{\overline{X}}$$

Gini's Mean Difference

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j|$$

$X \sim \mathcal{N} \Rightarrow \mathbf{E}\left(\frac{\sqrt{\pi}}{2} G\right) = \sigma$

Streuungsmaße (4)

S_n und Q_n (Croux, Rousseuw 1992, 1993)

$$\begin{aligned} S_n &= 1.1926 \cdot med_i(med_j|x_i - x_j|) \\ Q_n &= 2.219 \cdot \{|x_i - x_j|, i < j\}_{(k)} \end{aligned}$$

$k = h^2, h = \lfloor \frac{n}{2} \rfloor + 1$

SAS verwendet einen modifizierten Schätzer (Korrekturfaktor) für kleine Umfänge.

Die konstanten Faktoren sichern Erwartungstreue bei Normalverteilung, $X \sim \mathcal{N} \Rightarrow \mathbf{E}(S_n) = \mathbf{E}(Q_n) = \sigma$

Streuungsmaße (5)

Eigenschaften:

- Varianz und Standardabweichung und Spannweite sind nicht “robust”.
- IR und MAD sind robust.
  (MAD etwas besser da höherer “Bruchpunkt”)
- G ist bedingt robust, effizient bei F normal.
- IR und MAD sind wenig effizient.
  (0.37 bei Normal)
- S_n oder Q_n sind geeignetste Schätzungen.

Streuungsmaße (6)

Nicht-Robuste Skalenschätzungen

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \\
\text{Range} &= X_{(n)} - X_{(1)} \\
CV &= \frac{s \cdot 100}{\overline{X}}
\end{aligned}
$$

Streuungsmaße (7)

Robuste Skalenschätzungen

$$
\begin{aligned}
IR &= x_{0.75} - x_{0.25} \\
MAD &= med\left(\left|x_i - x_{0.5}\right|\right) \\
G &= \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j| \\
S_n &= 1.1926 \cdot med_i\left(med_j |x_i - x_j|\right) \\
Q_n &= 2.219 \cdot \left\{|x_i - x_j|, i < j\right\}_{(k)} \\
& \quad k = h^2, h = \lfloor \frac{n}{2} \rfloor + 1
\end{aligned}
$$

Lage- und Streuungsmaße in SAS (1)

```
PROC MEANS;
   VAR Zeit;
RUN;
```

Standardausgabe:
N, Mean, Std Dev, Minimum, Maximum

Vorteil: übersichtliche Darstellung

Nachteil: nur wenige Statistiken

Es können aber zusätzliche Statistiken durch Optionen angefordert werden, z.B.

```
PROC MEANS Median Sum CL;
```

`Descr1.sas`

Lage- und Streuungsmaße in SAS (2)

Die Prozedur Univariate

```
PROC UNIVARIATE;
        VAR Zeit;
RUN;
```

N, Mean, Std Deviation, Variance
Sum Observations, Median, Mode
Range, Interquartile Range
Lokationstests (später)
Quantile
Extreme Beobachtungen

Lage- und Streuungsmaße in SAS (3)

Getrimmte Mittel und robuste Skalenschätzer können einfach berechnet werden durch:

```
PROC UNIVARIATE ROBUSTSCALE TRIMMED=10
WINSORISED=10;
VAR ...;
RUN;
```

TRIMMED: getrimmte Mittel

TRIMMED=10: die je 10 kleinsten und größten Beobachtungen werden weggelassen.

WINSORIZED: winsorisierte Mittel

ROBUSTSCALE: robuste Skalenschätzer

Lage- und Streuungsmaße in SAS (4)

Abkürzung

```
PROC CAPABILITY ROBUSTSCALE TRIMMED=10
  WINSORISED=10;
ODS SELECT BASICMEASURES
  TRIMMEDMEANS ROBUSTSCALE
VAR ...;
RUN;
```

Formmaße

(Theoretische) Schiefe

$$\beta_1 = E\left(\frac{X - EX}{\sqrt{var(X)}}\right)^3$$

(Empirische) Schiefe

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s}\right)^3 \\ \hat{\beta}_{1,SAS} &= \hat{\beta}_1 \frac{n^2}{(n-1)(n-2)} \end{aligned}$$

$\beta_1 = 0$ falls F symmetrisch
$\beta_1 < 0$ falls F linksschief
$\beta_1 > 0$ falls F rechtsschief

ÜA: Berechnen Sie die (theor.) Schiefe von

$$X : \begin{pmatrix} \frac{1}{2}(-4-\sqrt{6}) & -1 & \frac{1}{2}(-4+\sqrt{6}) & 2 & 3 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

und von

$$Y : \begin{pmatrix} -9 & -7 & 2 & 4 & 10 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

PROC MEANS skewness;
PROC MEANS skewness vardef=n; (ohne Faktor)

Formmaße (2)

(Theoretische) Wölbung, Kurtosis

$$\beta_2 = E\left(\frac{X - EX}{\sqrt{var(X)}}\right)^4 - 3$$

(Empirische) Wölbung, Kurtosis

$$\begin{aligned}
\hat{\beta}_2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s}\right)^4 - 3 \\
\hat{\beta}_{2,SAS} &= \hat{\beta}_2 \frac{n^2(n+1)}{(n-1)(n-2)(n-3)} \\
&\quad -3\frac{4n^2 - 3n + 1}{(n-1)(n-2)(n-3)}
\end{aligned}$$

Exzeß

$\beta_2 + 3$ $\hat{\beta}_2 + 3$

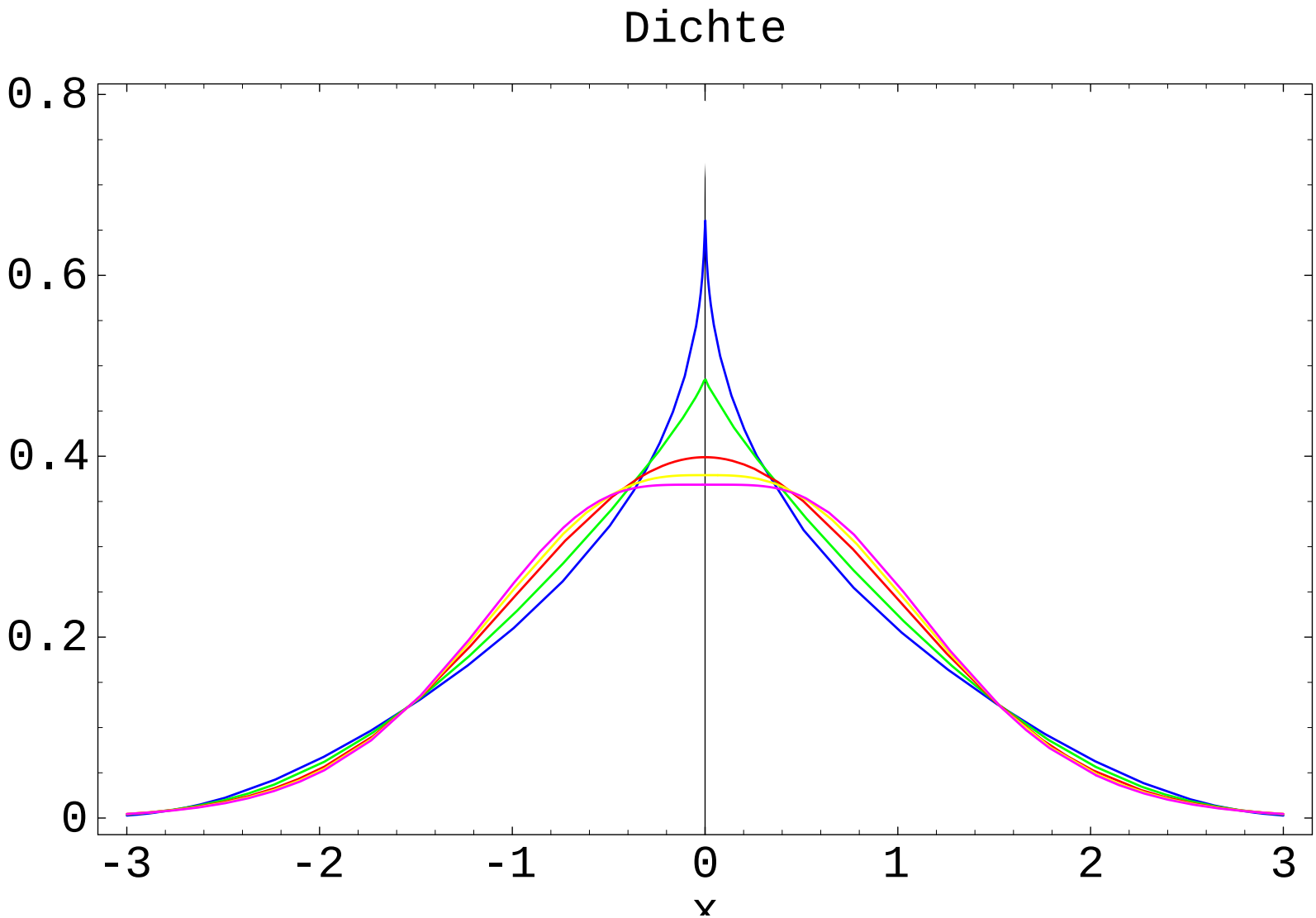
$\beta_2 = 0$	bei Normalverteilung
$\beta_2 > 0$	Tails “dicker, länger, stärker” als bei NV
$\beta_2 < 0$	Tails “dünner, kürzer, schwächer” als bei NV

PROC MEANS kurtosis;
PROC MEANS kurtosis vardef=n; (ohne Faktor)

$\beta_2 = 0$ heißt nicht notwendig: $F \sim$ Normal.

Kurtosis

Dichten mit $\mathbf{E}(X) = 0, var(X) = 1, \beta_1 = 0, \beta_2 = 0$



Formmaße (3)

Theoretische Schiefe und Kurtosis verschiedener Verteilungen

Verteilung	Schiefe	Kurtosis
normal	0	0
gleich	0	-1.2
Doppelexp	0	3
Exponential	2	6
Bi(n,p)	$\frac{1-2p}{\sqrt{np(1-p)}}$	$-\frac{6}{n}+\frac{1}{np(1-p)}$
Poi(λ)	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
Geo(p)	$\frac{2-p}{\sqrt{1-p}}$	$6+\frac{p^2}{1-p}$

Einschub: GPLOT (vgl. ÜA 9)

Darstellung zweidimensionaler Zusammenhänge

```
SYMBOL1 i=spline c=green v=point;
SYMBOL2 i=needle c=blue v=plus;
PROC GPLOT;
   PLOT y1*x=1 y2*x=2 /overlay;
RUN;
```

Die darzustellenden Paare (x,y) sind vorher in einem DATA-Step zu erzeugen oder einzulesen.

Nach dem Gleichheitszeichen im Plot-Kommando steht die Nummer der zugehörigen SYMBOL-Anweisung.

Prozedur GPLOT (2)

Die Symbol-Anweisung beschreibt die Art, den Stil des Plot

i=needle: Nadelplot (für diskrete Wahrscheinlichkeiten praktisch)

i=join: (nach x) aufeinander folgende Punkte werden verbunden

i=spline: Punkte werden durch einen Spline verbunden

c=<Farbe>

v=<Zeichen>

overlay: alles in ein Plot.

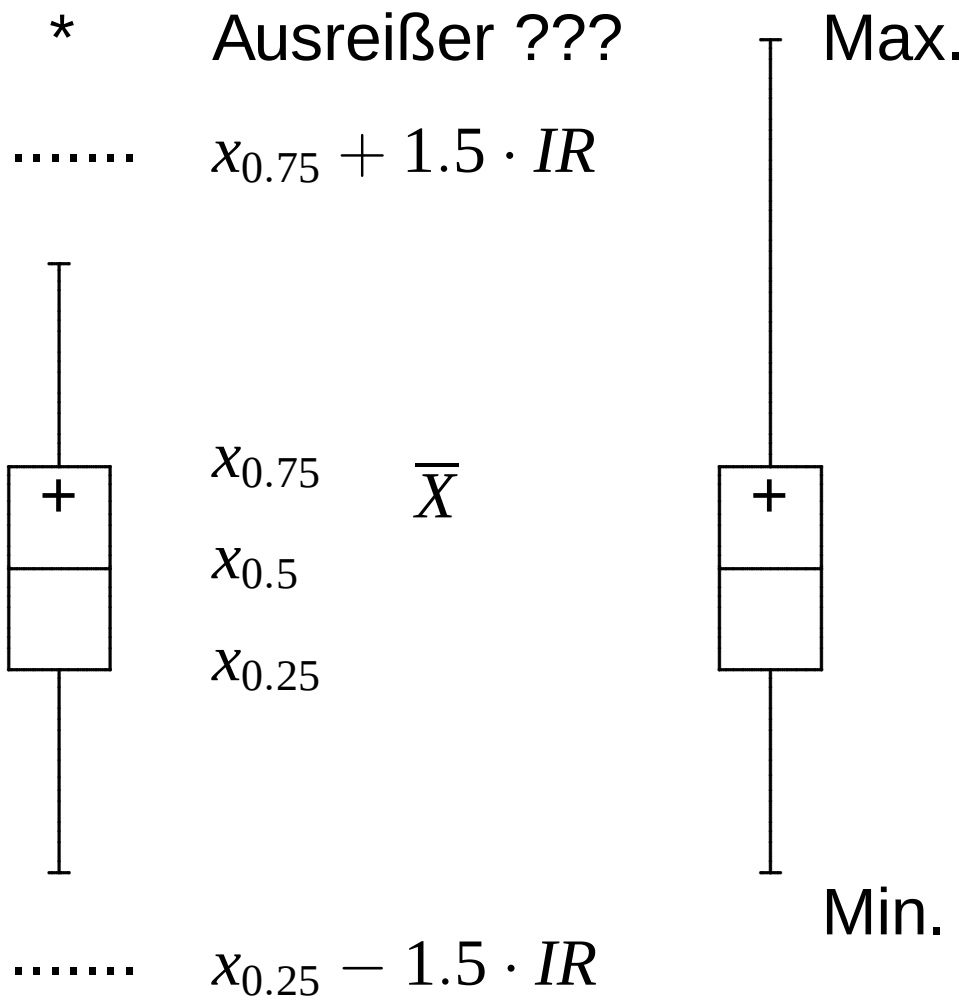
4.2 Box-Plots

Ziel: übersichtliche Darstellung der Daten.
Boxplot zu dem Eingangsbeispiel mit n=5:

`Descr_Boxplot0.sas`

Prozeduren: UNIVARIATE, GPLOT, BOXPLOT
PROC UNIVARIATE PLOT;
SYMBOL1 INTERPOL=BOXT10; **PROC GPLOT**; PLOT y*x=1; **PROC BOXPLOT**; PLOT y*x /BOXSTYLE=SCHEMATIC; /BOXSTYLE=SKELETAL;

Prozedur BOXPLOT



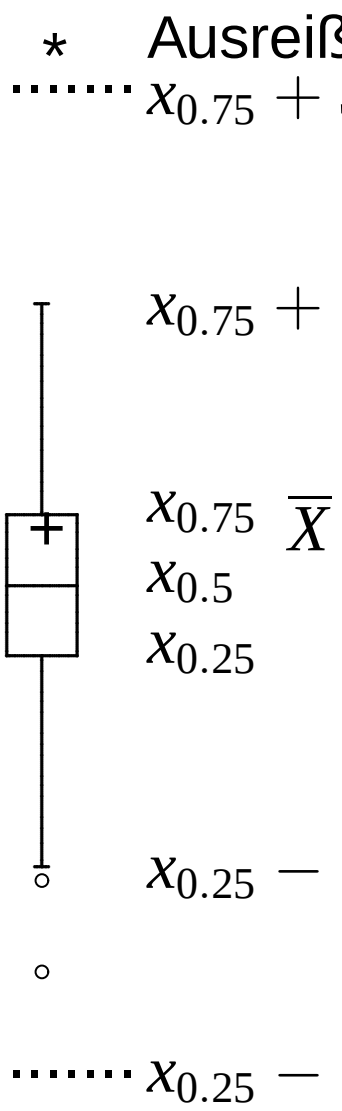
Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$

etwa 99% der Daten liegen zwischen den “fences”.

$$
\begin{aligned}
0.99 &= 0.995 - 0.005 \\
&= \Phi(2.575) - \Phi(-2.575) \\
&= P(\mu - 2.575\sigma < X < \mu + 2.575\sigma) \\
&\approx P(x_{0.5} - 2.575 \cdot \underbrace{0.7434 \cdot IR}_{} < X < \\
&\qquad x_{0.5} + 2.575 \cdot \underbrace{0.7434 \cdot IR}_{}) \\
&= P(x_{0.5} - 1.914 \cdot IR < X < x_{0.5} + 1.914 \cdot IR) \\
&\approx P(x_{0.5} - 2 \cdot IR < X < x_{0.5} + 2 \cdot IR) \\
&= P(x_{0.25} - 1.5 \cdot IR < X < x_{0.75} + 1.5 \cdot IR)
\end{aligned}
$$

Prozedur UNIVARIATE, Option PLOT



* Ausreißer ??

$x_{0.75} + 3 \cdot IR$

$x_{0.75} + 1.5 \cdot IR$ oder Max.

$x_{0.75}$ $\overline{X}$

$x_{0.5}$

$x_{0.25}$

$x_{0.25} - 1.5 \cdot IR$ oder Min.

$x_{0.25} - 3 \cdot IR$

Box-Plots in SAS

Ein Merkmal, eine Gruppe (Merkmal gr)

```
gr = 1;
PROC BOXPLOT;
   PLOT zeit*gr; RUN;
```

Ein Merkmal (zeit), mehrere Gruppen (gr)

```
PROC BOXPLOT;
   PLOT zeit*gr; RUN;
```

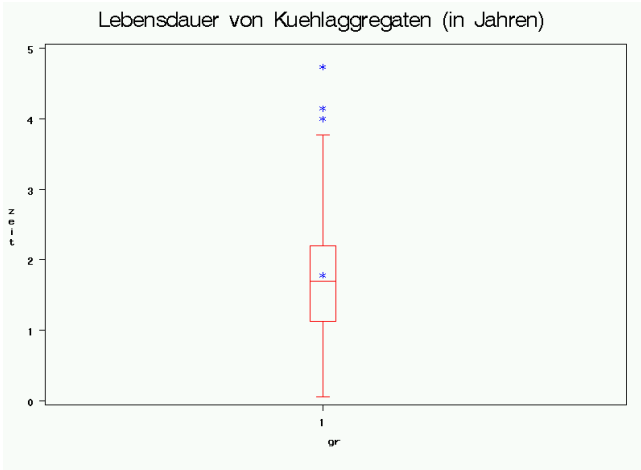
Ein Merkmal (X), mehrere Gruppen (gr)

```
SYMBOL INTERPOL=BOXT10;
PROC GPLOT; PLOT X*gr; RUN;
```

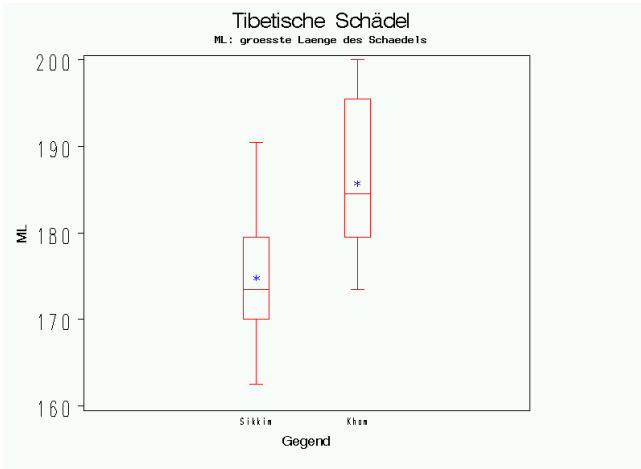
`Descr_Boxplot.sas` `Descr_Boxplot1.sas`

Boxplots - Beispiele

Lebensdauern von 100 Kühlaggregaten



Schädelmaße in zwei Regionen Tibets



Box-Plots in SAS (2)

Box-Plots von mehreren Variablen

`Descr_Boxplot2.sas`

1. Data-Step:
   Definition von neuen Variablen, die konstant gesetzt werden.
2. Symbol-Anweisungen für die einzelnen darzustellenden Variablen definieren.
3. Achsenbeschriftung entsprechend den Variablen definieren.
4. Prozedur GPLOT;

4.3 Probability Plots

Erinnerung: Normalverteilung

(i) Dichte der Standard-Normalverteilung

$$\phi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty$$

(ii) Verteilungsfunktion der Standard-Normal

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{t^2}{2}} dt, \qquad -\infty < x < \infty$$

(iii) Dichte der Normalverteilung

$$\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{\sigma^2}},$$

mit Erwartungswert μ und Varianz σ^2.

Probability Plots

Erinnerung: Normalverteilung, Quantile

Der Wert $\Phi^{-1}(u)$ heißt u-Quantil
der Standard-Normalverteilung.

Die Funktion $\Phi^{-1}(u), u \in (0,1)$, heißt Quantilfunktion
der Standard-Normalverteilung.

$\alpha = 0.05$

$$\begin{aligned} \Phi^{-1}(1-\alpha) &= \Phi^{-1}(0.95) = 1.645 \\ \Phi^{-1}(1-\frac{\alpha}{2}) &= \Phi^{-1}(0.975) = 1.96 \end{aligned}$$

$\Phi^{-1}(\alpha)$: α-Quantil, theoretisch
$x_\alpha = x_{(\lfloor \alpha n \rfloor)}$: α-Quantil, empirisch

Q-Q-Plot

Variante 1

Wenn Normalverteilung zutrifft, so müssen die Punkte

$$(\Phi^{-1}(\alpha), x_\alpha)$$

etwa auf einer Geraden liegen,

$$\Phi^{-1}(\alpha) \approx \frac{x_\alpha - \mu}{\sigma} = \frac{x_{(\lfloor \alpha n \rfloor)} - \mu}{\sigma}$$

```
PROC UNIVARIATE PLOT; RUN;
```

Die theoretischen Werte (+) werden durch die empirischen Werte (*) überschrieben.
Je weniger “+”-Zeichen zu sehen sind, desto näher sind wir an der NV.

`Descr_QQPlot.sas`

Q-Q-Plot

Variante 2

```
PROC UNIVARIATE;
   QQPLOT var /Optionen;
RUN;
```

wie oben, bessere Grafik, aber keine Linie.
Es werden die Punkte

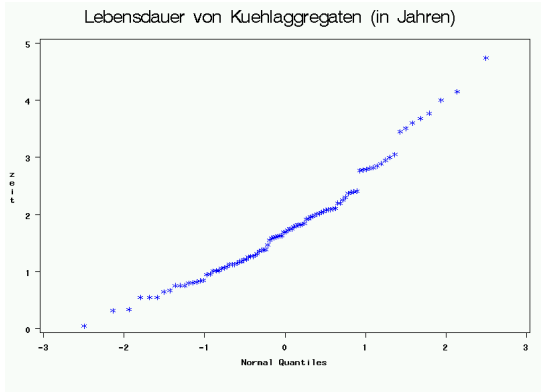
$$\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right), x_{(i)}\right)$$

geplottet. $i = 1, \ldots, n$.

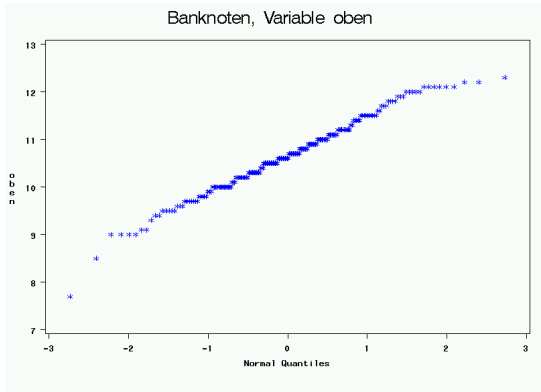
Bem.: $\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$ ist eine Approximation von $\mathbf{E}X_{(i)}$ bei Standard-Normalverteilung.

Q-Q Plots - Beispiele

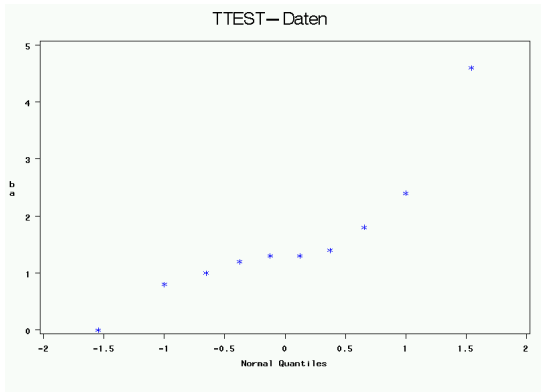
Lebensdauern von 100 Kühlaggregaten



Abmessungen von Banknoten



Verlängerung der Schlafdauer



Probability Plot

```
PROC UNIVARIATE;
   PROBPLOT var /Optionen;
RUN;
```

wie oben, x-Achse hat die selbe Skala, aber eine andere Beschriftung, statt x_α steht α, also

$$\left(\alpha, x_{(i)}\right) = \left(\frac{i - 0.375}{n + 0.25}, x_{(i)}\right)$$

Bem.: Es können auch einige andere Verteilungen verwendet werden.

Q-Q Plot
Übersicht

wenige Punkte weg von der Geraden	Ausreißer
linkes Ende unter der Linie rechtes Ende über der Linie	lange Tails
linkes Ende über der Linie rechtes Ende unter der Linie	kurze Tails
gebogene Kurve, steigender Anstieg	rechtsschief
gebogene Kurve, fallender Anstieg	linksschief
Plateaus und Sprünge	diskrete Daten gerundete Dat.
Gerade $y = x$	empirische $\approx$ theoretische Verteil.
Gerade $y = ax + b$	nur Lage- oder Skalenunterschied

4.4 Häufigkeitsdiagramme

PROC UNIVARIATE
PROC GCHART

```
PROC GCHART <DATA=sasdatei>;
  VBAR variablenliste </Optionen>;
       /* vertikales Histogramm */
  HBAR var.list </Optionen>;
       /* horizontales Histogramm */
  PIE var.list </Optionen>; /* Kreisdiagr. */
  STAR var.list </Optionen>; /* Sterndiagr. */
  BLOCK var.list </Optionen>;
       /* 3 dim. Balkendiagramm */
RUN;
```

Häufigkeitsdiagramme

Optionen (1)

VBAR3D, HBAR3D, PIE3D anstelle von
VBAR, HBAR, PIE liefern schönere Bilder.

DISCRETE Zusammenfassung von Ausprägungen wird unterdrückt, d.h. für jeden Wert wird eine Säule erzeugt.

LEVELS = anzahl	gewünschte Anzahl Säulen
TYPE = FREQ	Häufigkeiten (Standard)
= PERCENT	Prozente
= CFREQ	kum. Häufigkeiten
= CPERCENT	kum. Prozente
= SUM	Summen (nur mit SUMVAR)
SUMVAR = anzahl	Anzahl ist bereits aufsummierte Häufigkeit

Häufigkeitsdiagramme
Optionen (2)

MIDPOINTS = Mittelpunkte der Balken.
Balken haben alle die gleiche Breite!

GROUP= Gruppierungsvariable

SUBGROUP= Gruppierungsvariable, gemeinsame Auswertung

PATTERNID=Musterzuordnung
Vergleiche die PATTERN-Anweisung

```
Descr_Gchart_1a.sas
Descr_Gchart_1b.sas
Descr_Gchart_3.sas    3a,3b
Descr_Gchart_1.sas
```

Häufigkeitsdiagramme

Design der Diagramme

PATTERNxn C= V=

C, COLOR	Farbe: blue,cyan,red,black... black ist Voreinstellung
V, VALUE	Wert: star,plus point,...
x	Muster: X_n: schraffiert S_n: Solid R_n: /// L_n: \\\
n	1-5: Dichte des Musters.

Histogramme und Dichteschätzung

Auch Prozedur UNIVARIATE liefert Histogramme

```
PROC UNIVARIATE;
  HISTOGRAM varname </Optionen>;
RUN;
```

Sie liefert auch Tabellen von Histogrammen;

```
PROC UNIVARIATE;
  CLASS Klassenvariablen;
  HISTOGRAM varname </Optionen>;
RUN;
```

```
Descr_Plot_Kuehl.sas
Desc_ZweidimHisto_Heroin.sas
```

Histogramme und Dichteschätzung

Optionen

CBARLINE=	Farbe des Histogramms
WBARLINE=	Dicke der Histogrammlinien
L=	Linientyp (Standard: 1, solid)
MIDPOINTS=	wie bei GPLOT
KERNEL	Nichtparametr. Dichteschätzung
COLOR=	Farbe der Dichtekurve
NORMAL	Parametrische Dichteschätzung (Normalverteilung)
GAMMA	Parametrische Dichteschätzung (Gammaverteilung)

Parametrische Dichteschätzung

Vorgabe: Modell, z.B. Normalverteilung oder Gammaverteilung
Lediglich die Parameter werden geschätzt.

```
PROC UNIVARIATE;
  HISTOGRAM varn/normal gamma;/*Parametrisch*/
  HISTOGRAM varn/kernel; /*Nichtparametrisch*/
RUN;
```

Frage: Wie wird geschätzt?

bei Normalverteilung ist das klar: $\overline{X}$ und s^2 sind optimale Schätzungen für μ und σ^2.

Wie findet man (gute) Schätzungen bei anderen Verteilungen?

Schätzmethoden

Momentenmethode

Man drückt den zu schätzenden Parameter durch die Momente, z.B. $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, aus.
Dann werden die Momente durch die empirischen Momente, hier $\overline{X}$, $\frac{1}{n}\sum X_i^2$ ersetzt.

Maximum-Likelihood-Schätzung

Es wird der Schätzwert für den unbekannten Parameter ermittelt, bei dem die Beobachtungen am meisten für diesen Parameter sprechen (most likely).

Normalverteilung $\mathcal{N}(\mu, \sigma^2)$

$\overline{X}$ und $\hat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2$ sind Momentenschätzungen für μ und σ^2. Sie sind auch ML-schätzungen für μ und σ^2.

SAS berechnet in der Regel Maximum-Likelihood-Schätzungen.

Maximum-Likelihood-Schätzung

$X_i \sim \mathcal{N}(\mu, 1)$ unabhängig

Likelihood: $L_n := f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$, die gemeinsame Dichtefunktion der X_i.

$$
\begin{aligned}
L_n(\mu) &= \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \text{(Unabhängigkeit)} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} \\
\ln L_n(\mu) &= -n \ln(\sqrt{2\pi}) + \sum_{i=1}^{n} (-\frac{(x_i-\mu)^2}{2}) \\
\frac{\partial L_n(\mu)}{\partial \mu} &= \sum_{i=1}^{n} (x_i - \mu)
\end{aligned}
$$

Nullsetzen liefert: $\hat{\mu} = \overline{X}$.

Nichtparametrsiche Dichteschätzung

Überlagerung der Daten mit einer (Dichte-) Funktion

$K(t)$ eine Kernfunktion,

$$\begin{aligned} \int K(t)\,dt &= 1, \quad \int tK(t)\,dt = 0, \\ \int t^2K(t)\,dt &= 1, \quad \int K^2(t)\,dt < \infty \end{aligned}$$

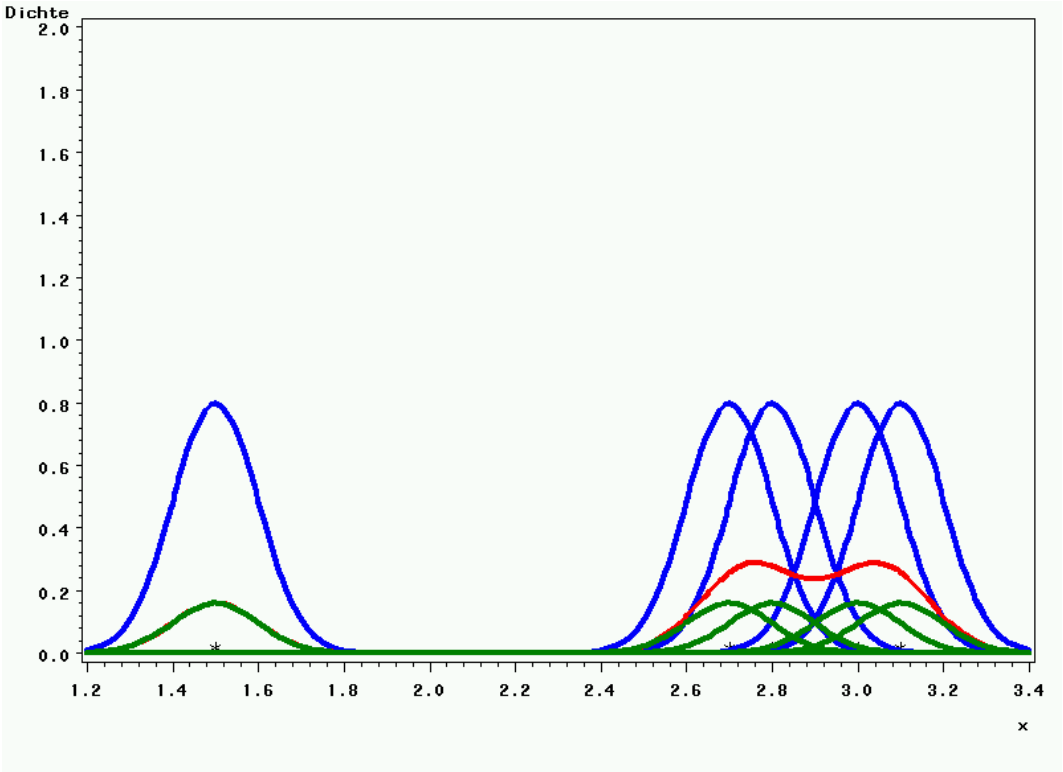
Dichteschätzung oder Dichtefunktionsschätzung.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x - x_i}{h})$$

$x_1, \ldots, x_n$: die Beobachtungen.
h: ein sogenannter Glättungsparameter.

Dichteschätzung

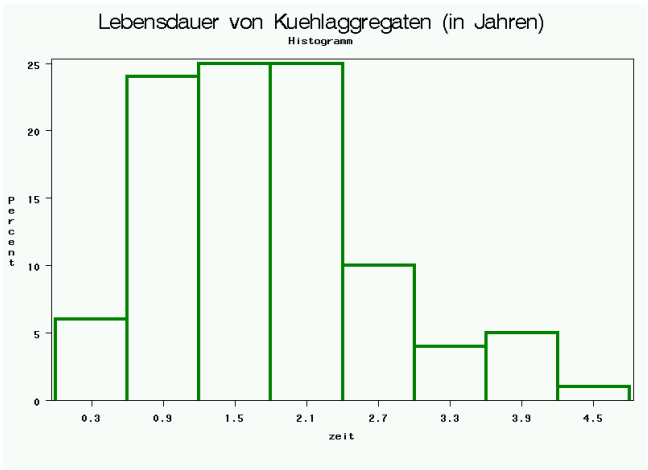
Motivation Kern-Dichteschätzung

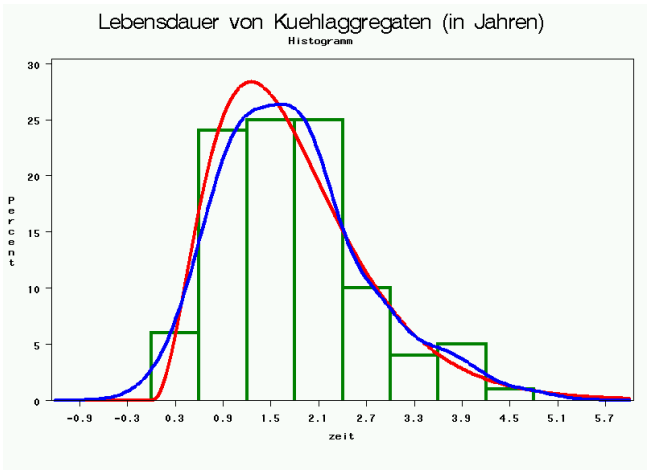


`Descr_Dichteschaetzung.sas`

Dichteschätzung, Beispiel

Kühlaggregate





Histogramm

Parametrische Dichteschätzung (Gamma)

Nichtparametrische Dichteschätzung

4.5 Häufigkeitstabellen

Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

Eindimensionale Zufallsvariablen

$$X: \quad \begin{pmatrix} x_0 & x_1 & \cdots & x_n & \cdots \\ p_0 & p_1 & \cdots & p_n & \cdots \end{pmatrix}$$

Die p_i sind zu schätzen:

$$\hat{p}_i = \frac{n_i}{N}$$

N: Stichprobenumfang n_i: relative Häufigkeiten

PROC FREQ Optionen;
    TABLES variablenliste /Optionen;
RUN;

`DescrFreqBanknote.sas  DescrFreq`

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel

3maliges Werfen einer Münze

X: Anzahl von Blatt nach 3 Würfen

Y: Anzahl von Blatt nach 2 Würfen

Element von Ω	X	Y
BBB	3	2
BBZ	2	2
BZB	2	1
BZZ	1	1
ZBB	2	1
ZBZ	1	1
ZZB	1	0
ZZZ	0	0

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel (Fortsetzung)

Besetzungswahrscheinlichkeiten

$X \mid Y$	0	1	2	
0	$\frac{1}{8}$	0	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{4}$	0	$\frac{3}{8}$
2	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{8}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

$$Y : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Tabelle der zweidimensionalen Wahrscheinlichkeiten

$X \mid Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_N	
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1N}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2N}	$p_{2.}$
$\cdots$							
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iN}	$p_{i.}$
$\cdots$							
x_M	p_{M1}	p_{M2}	$\cdots$	p_{Mj}	$\cdots$	p_{MN}	$p_{M.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.N}$	1

Zweidimensionale diskrete Zufallsgrößen

Zweidimensionale Zufallsvariable

Seien X, Y Zufallsgrößen. Das Paar (X, Y) heißt zweidimensionale Zufallsvariable.

Seien X und Y diskret und (x_i, y_j) die möglichen Ergebnisse von (X, Y), $\quad i = 1, \ldots, M, j = 1, \ldots, N$.

gemeinsame Wahrscheinlichkeitsfunktion von (X, Y)

$$p_{ij} = P(X = x_i, Y = y_j),$$

$$\begin{aligned} p_{ij} &\geq 0 \\ \sum_{i,j} p_{ij} &= 1 \end{aligned}$$

$$p_{i.} := \sum_{j=1}^{N} p_{ij} \qquad p_{.j} := \sum_{i=1}^{M} p_{ij}$$

Zweidimensionale diskrete Zufallsgrößen

Beispiel

Treiben Sie Sport?

X: 0 - nein 1 - ja

Y: 0 - weiblich 1 - männlich

X | Y	0	1	
0	p_{00}	p_{01}	$p_{0.}$
1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	

p_{ij}: unbekannt!

Frage: Ist das Sportverhalten von Männern und Frauen unterschiedlich? Hängt das Sportverhalten vom Geschlecht ab?

Zweidimensionale diskrete Zufallsgrößen

Kontingenztafel

Befragung liefert Häufigkeiten für die einzelnen Felder. Anhand dieser Häufigkeiten werden die Wahrscheinlichkeiten geschätzt!

Die Tabelle der Häufigkeiten heißt Kontingenztafel

X | Y	0	1	# der beobachteten	
0	n_{00}	n_{01}	$n_{0.}$	Nichtsportler
1	n_{10}	n_{11}	$n_{1.}$	Sportler
	$n_{.0}$	$n_{.1}$		
	# der befragten			
	Frauen	Männer		

$$p_{ij} \approx \frac{n_{ij}}{n} = \hat{p}_{ij}$$

Zweidimensionale diskrete Zufallsgrößen

Häufigkeitstabellen in SAS

```
PROC FREQ Optionen;
TABLES variablenliste /Optionen;
TABLES vliste1*vliste2 /Optionen;
TABLES vliste1*vliste2*varliste3;RUN;
```

Option im Prozedur-Step

ORDER=schlüsselwort, z.B. ORDER=FREQ
wenn die Ausgabe nach Häufigkeiten geordnet.

Optionen der TABLES-Anweisung

MISSING: fehlende Werte werden bei der Berechnung relativer Häufigkeiten mit einbezogen.

OUT=sasfile: Ausgabe der Tabelle in ein SAS-File

Optionen der TABLES-Anweisung

nur für mehrdim. Tabellen

CHISQ:	χ^2-Unabhängigkeitstest
CMH:	u.a. Odds Ratio
MEASURES:	Assozitionsmaße, Korrelationskoeffizient
NO...	keine Ausgabe von:
NOFREQ:	absoluten Häufigkeiten
NOPERCENT:	relativen Häufigkeiten
NOROW:	Zeilenhäufigkeiten
NOCOL:	Spaltenhäufigkeiten

Assoziationsmaße

nur für mehrdim. Tabellen

χ^2-Statistik

$$\sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

Φ-Koeffizient für 2x2 Tafeln

$$\Phi^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1.}p_{2.}p_{.1}p_{.2}}$$

Odds Ratio für 2x2 Tafeln

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Schätzung: Ersetzen der Wahrscheinlichkeiten durch die jeweiligen relativen Häufigkeiten.

Assoziationsmaße, Beispiel

Mendelsche Kreuzungsversuche

```
DATA Erbsen;
INPUT rund gruen Anzahl;
CARDS;

0 0 101
0 1  32
1 0 315
1 1 108

;
RUN;
```

```
PROC FREQ;
WEIGHT Anzahl;
TABLES rund*gruen \
  chisq cmh;
RUN;
```

$\chi^2 = 0.1163$ Φ-Koeffizient=0.0145.

4.6 Zusammenhangsmaße

zwischen Zufallsvariablen X, Y

Erinnerung: Varianz der Zufallsvariablen X

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 \\ &= \mathbf{E}\big[(X - \mathbf{E}X)(X - \mathbf{E}X)\big] \end{aligned}$$

Kovarianz der Zufallsvariablen X und Y

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y) \\ &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \end{aligned}$$

Korrelation der Zufallsvariablen X und Y

$$Corr(X, Y) = \frac{\mathbf{E}\big[(X - \mathbf{E}X)(Y - \mathbf{E}Y)\big]}{\sqrt{var(X) \cdot var(Y)}}$$

Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})$$

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

empirische Korrelation, Pearson-Korrelationskoeffizient

$$r_{XY} := \frac{s_{XY}}{s_X s_Y}$$

Pearson-Korrelationskoeffizient

Eigenschaften

- Es gilt stets:

$$-1 \leq r_{XY} \leq 1.$$

- Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen

$$x \longrightarrow a + bx$$

- $|r_{XY}| = 1$ gdw. alle Punkte auf einer Geraden liegen,
  $y = mx + b, m \neq 0$
  $r_{XY} = \quad 1 \rightarrow$ Anstieg > 0
  $r_{XY} = -1 \rightarrow$ Anstieg < 0

Pearson-Korrelationskoeffizient

- Der Korrelationskoeffizient ist also ein Maß für die lineare Abhängigkeit von X und Y.
- $r_{XY} \approx 0 \longrightarrow$ keine lineare Beziehung zwischen X und Y erkennbar, aber es sind durchaus andere Abhängigkeiten möglich!
- Der Pearson-Korrelationskoeffizient ist nicht robust gegen Ausreißer (siehe Übung)

Realisierung in SAS:

```
PROC CORR PEARSON;
VAR X Y;
RUN;
```

Spearman-Korrelationskoeffizient

Spearman-Rangkorrelationskoeffizient

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

R_i: Rang von X_i in der geordneten Stichprobe $X_{(1)} \leq \ldots \leq X_{(n)}$
S_i: Rang von Y_i in der geordneten Stichprobe $Y_{(1)} \leq \ldots \leq Y_{(n)}$

```
PROC CORR SPEARMAN;
VAR X Y;
RUN;
```

Spearman-Korrelationskoeffizient

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_i - S_i)^2}{n \cdot (n^2 - 1)}
\end{aligned}
$$

$$-1 \leq r_S \leq +1$$

$|r_S| = 1$ gdw. X_i, Y_i in gleicher oder entgegengesetzter Weise geordnet sind!

Spearman-Korrelationskoeffizient

Beweis der letzten Formel (1)

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})^2}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})}}$$

Nenner:

$$\sum_{i=1}^{n}(R_i - \overline{R})^2 = \sum_{i=1}^{n}(S_i - \overline{S})^2 = \sum_{i=1}^{n}(i - \frac{n+1}{2})^2$$

$$= \sum i^2 - 2 \cdot \frac{n+1}{2} \sum i + n \cdot (\frac{n+1}{2})^2$$

$$= \frac{n \cdot (n+1) \cdot (2n+1)}{6} - \frac{n \cdot (n+1)^2}{2} + \frac{n \cdot (n+1)^2}{4}$$

$$= \frac{n \cdot (n+1)}{12} \cdot [2 \cdot (2n+1) - 3 \cdot (n+1)]$$

$$= \frac{(n-1) \cdot n \cdot (n+1)}{12} = \frac{n \cdot (n^2-1)}{12}$$

Spearman-Korrelationskoeffizient

Beweis der letzten Formel (2)

Zähler:

$$\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}) = \sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})$$

$$= \sum_{i=1}^{n} R_i S_i - 2 \cdot \frac{n+1}{2} \sum_{i=1}^{n} R_i + n \cdot (\frac{n+1}{2})^2$$

$$= \sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}$$

Damit erhalten wir eine weitere Darstellung für r_S:

$$r_S = 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}}{(n-1) \cdot n \cdot (n+1)}$$

Spearman-Korrelationskoeffizient

Andere Darstellung für den Zähler

Setzen: $d_i := R_i - S_i = (R_i - \frac{n+1}{2}) + (\frac{n+1}{2} - S_i)$

$$
\begin{aligned}
\sum d_i^2 &= \sum (R_i - \frac{n+1}{2})^2 + \sum (S_i - \frac{n+1}{2})^2 \\
&\quad -2 \sum (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2}) \\
&= \frac{(n-1)n(n+1)}{12} + \frac{(n-1)n(n+1)}{12} \\
&\quad -2 \cdot r_S \cdot \frac{(n-1)n(n+1)}{12} \\
&= \frac{(n-1)n(n+1)}{6}(1 - r_S) \\
r_S &= 1 - \frac{6 \sum d_i^2}{(n-1)n(n+1)}
\end{aligned}
$$

Spearman-Korrelationskoeffizient

Drei Darstellungen

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}} \\
&= 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n\cdot(n+1)^2}{4}}{(n-1)n(n+1)} \\
&= 1 - \frac{6\sum (R_i - S_i)^2}{(n-1)n(n+1)}
\end{aligned}
$$

Bem.: Es gilt:

a) $-1 \leq r_S \leq 1$

b) $r_S = 1 \Leftrightarrow R_i = S_i \quad \forall i = 1, \ldots, n$

c) $r_S = -1 \Leftrightarrow R_i = n + 1 - S_i \quad \forall i = 1, \ldots, n$

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Vorteile Spearman

- es genügt ordinales Meßniveau
- leicht zu berechnen
- r_S ist invariant gegenüber monotonen Transformationen
- gute Interpretation, wenn $r_S \approx -1, 0, 1$ (wie bei Pearson)
- eignet sich als Teststatistik für einen Test auf Unabhängigkeit
- ist robust gegen Abweichungen von der NV.

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Nachteile Spearman

- wenn kardinales (stetiges) Meßniveau $\longrightarrow$ Informationsverlust
- schwierige Interpretation,
wenn r_S nicht nahe 0, 1, oder -1
(gilt eingeschränkt auch für Pearson)

Kendalls τ (Konkordanzkoeffizient)

$(X_i, Y_i), i = 1, ..., n$

$$a_{ij} = \begin{cases} 1, \text{ falls } x_i < x_j \wedge y_i < y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i > y_j \\ -1, \text{ falls } x_i < x_j \wedge y_i > y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i < y_j \\ 0, \text{ sonst} \end{cases}$$

$$= sgn[(X_i - X_j)(Y_i - Y_j)]$$

Falls $a_{ij} = 1$ so heißen die Paare konkordant
Falls $a_{ij} = -1$ " diskordant
Falls $a_{ij} = 0$ " gebunden

Kendalls τ (Konkordanzkoeffizient)

$$\begin{aligned}\tau &= \frac{2 \cdot \sum_{i<j} a_{ij}}{N \cdot (N-1)} = \frac{1}{\binom{N}{2}} \cdot \sum_{i<j} a_{ij} \\ &= \frac{\sharp \text{ konkordanter Paare - } \sharp \text{ diskordanter Paare}}{\binom{N}{2}}\end{aligned}$$

Bem.: einfache Berechnung, wenn neue Paare hinzukommen

Bem.: meist gilt: $|\tau| < |r_S|$. Approximation von τ:

$$\hat{\tau} = \frac{2}{3}\frac{N+1}{N} r_S$$

PROC CORR KENDALL; VAR X Y; **RUN**;

4.7 Das Regressionsproblem

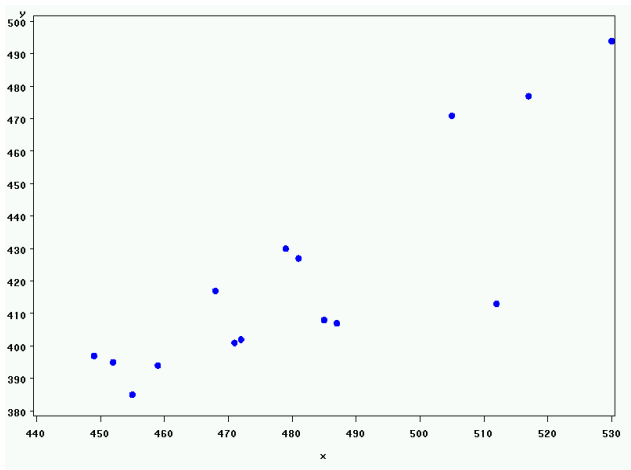
Scatterplots

Scatterplot

Zweidimensionale Stichproben können als Punkte in der Ebene dargestellt werden

Länge und Breite von Venusmuscheln

```
PROC GPLOT;
PLOT y*x;
RUN;
```



`Descr_Scatter.sas` `Descr_Scatter1.sas`

Das Regressionsproblem

X, Y: Zufallsvariablen (auch mehrdimensional)

Modell:

$$Y = f(X, \underbrace{\theta_1, ..., \theta_p}) + \epsilon, \qquad \epsilon \sim (0, \sigma^2).$$

Parameter zuf. Fehler

f linear, bekannt bis auf Parameter:
lineare Regression
f nichtlinear, bekannt bis auf Parameter:
nichtlineare Regression
f unbekannt: nichtparametrische Regression

Regression

f bekannt (bis auf Parameter)

Aufgabe:

$$min_{\theta_1,\ldots,\theta_p}\mathbf{E}(Y - f(\mathbf{X}, \theta_1, \ldots, \theta_p))^2$$

$\theta_1, \ldots, \theta_p$ unbekannt.

Beobachtungen: $(Y_i, \mathbf{X}_i)$.

Erwartungswert durch arithmetisches Mittel ersetzen

$$min_{\theta_1,\ldots,\theta_p}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i, \theta_1, \ldots, \theta_p))^2$$

Kleinste Quadrat-Schätzung für $\theta_1, \ldots, \theta_p$ (KQS)
Least-Squares-Estimation (LSE)

Regression

f bekannt (bis auf Parameter)

Lösung des Minimum-Problems

$$min_{\theta_1,...,\theta_p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i, \theta_1, ..., \theta_p))^2$$

zu minimierende Funktion differenzieren und Null setzen:

$$\frac{2}{n} \cdot \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i, \theta_1, ..., \theta_p)) \cdot \frac{\partial f(\mathbf{X}_i, \theta_1, ..., \theta_p)}{\partial \theta_j} = 0$$

$j = 1, ..., p, \Rightarrow$ Gleichungssystem mit p Gleichungen.

Regression

f linear: lineares Gleichungssystem (1)

$$f(X, \theta_1, \theta_2) = \theta_1 X + \theta_2$$

$$\frac{\partial f}{\partial \theta_1} = X \qquad \frac{\partial f}{\partial \theta_2} = 1$$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(Y_i - (\theta_1 X_i + \theta_2)) \cdot X_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n}(Y_i - (\theta_1 X_i + \theta_2)) \cdot 1 &= 0 \\
\hline
\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_2 \sum_i X_i &= 0 \\
\sum_i Y_i - \theta_1 \sum_i X_i - \theta_2 \cdot n &= 0
\end{aligned}$$

Regression

f linear: lineares Gleichungssystem (2)

Die zweite Gleichung nach θ_2 auflösen:

$$\theta_2 = \frac{1}{n}\sum_i Y_i - \theta_1 \frac{1}{n}\sum_i X_i$$

und in die erste einsetzen:

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \frac{1}{n}\sum_i Y_i \sum_i X_i + \theta_1 \frac{1}{n}\sum_i X_i \sum_i X_i = 0$$

$$\sum_i X_i Y_i - \frac{1}{n}\sum_i Y_i \sum_i X_i - \theta_1 \big((\sum_i X_i^2 - \frac{1}{n}\sum_i X_i \sum_i X_i\big) = 0$$

$\Rightarrow$

$$\hat{\theta}_1 = \frac{\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i}{\sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2} = \frac{S_{XY}}{S_X^2}, \hat{\theta}_2 = \frac{1}{n}\big(\sum_i Y_i - \hat{\theta}_1 \sum_i X_i\big)$$

Regression

Zähler und Nenner in $\hat{\theta}_1$

$$
\begin{aligned}
S_{XY} &= \frac{1}{n-1}\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - \overline{X}\sum_i Y_i - \overline{Y}\sum_i X_i + n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - n\overline{XY} - n\overline{XY} + n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i\Big) \\
S_{X^2} &= \frac{1}{n-1}\Big(\sum_i X_i X_i - \frac{1}{n}\sum_i X_i \sum_i X_i\Big)
\end{aligned}
$$

Spezialfall $f(X, \theta) = \theta$ (konstant)

$$Y_i = \theta + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2)$$

Minimierungsaufgabe:

$$\min_{\theta} \Big(\sum_{i=1}^{n} (Y_i - \theta)^2 \Big)$$

Lösung:

$$2 \sum_{i=1}^{n} (Y_i - \theta) = 0 \qquad \sum_{i=1}^{n} Y_i - n\theta = 0$$

$$\hat{\theta} = \frac{1}{n} \sum Y_i = \overline{Y}$$

D.h. $\overline{Y}$ ist auch KQS.

Spezialfall $f(X, \theta) = \theta$

Schätzung des Schätzfehlers

$$\sigma^2_{Y_i} = \sigma^2_{\theta+\epsilon_i} = \sigma^2_{\epsilon_i} = \sigma^2.$$

Schätzfehler:

$$\begin{aligned}
\sigma^2_{\hat{\theta}} &= var(\hat{\theta}) = var(\frac{1}{n} \cdot \sum Y_i) = \frac{1}{n^2} \cdot n \cdot varY_i \\
&= \frac{1}{n} \cdot \sigma^2 \qquad \rightarrow_{n\rightarrow\infty} \quad 0 \\
\hat{\sigma}^2_{\hat{\theta}} &= \frac{\hat{\sigma}^2}{n}
\end{aligned}$$

Lineare und Nichtlineare Regression

f: linear, $f(X, \theta_1, \theta_2) = \theta_1 X + \theta_2$

θ_1 und θ_2 werden geschätzt.

`Descr_Scatter_1.sas`
`Descr_Scatter_Heroin.sas`

f: nichtlinear, z.B. $f(X, \theta_1, \theta_2) = ln(\theta_1 X + \theta_2)$

a) Lösung des nichtlinearen Gleichungssystems

b) wird auf den linearen Fall zurückgeführt

$$\begin{aligned} Y &= ln(\theta_1 X + \theta_2) + \epsilon \\ e^Y &= \theta_1 X + \theta_2 + \tilde{\epsilon} \end{aligned}$$

Modelle sind aber i.A. nicht äquivalent!

Weitere nichtlineare Regressionsfunktionen

$$
\begin{aligned}
f(t) &= a + bt + ct^2 && \text{Parabel} \\
f(t) &= at^b && \text{Potenzfunktion} \\
f(t) &= ae^t && \text{Exponentialfunktion} \\
f(t) &= k - ae^{-t} && \\
f(t) &= \frac{k}{1 + be^{-ct}} && \text{logistische Funktion} \\
\ln f(t) &= k - \frac{a}{b + t} && \text{Johnson-Funktion} \\
\ln f(t) &= k - \lambda e^{-t} && \text{Gompertz-Funktion}
\end{aligned}
$$

Nichtparametrische Regression

f unbekannt, aber “glatt”

z.B. f 2x stetig differenzierbar, $f \in C_2$, $\lambda \geq 0$

Glättender Kubischer Spline ist Lösung von

$$min_{f \in C_2} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \cdot \int \left(f''(x)\right)^2 dx$$

```
Descr_Scatter.sas
SYMBOL I=SMnnS;
```

SM: Smoothing Spline

nn: Glättungsparameter
nn=00: Interpolierender Spline
nn=99: Gerade

S: Punktepaare werden vor der Auswertung nach dem Argument sortiert.

Nichtparametrische Regression

Kernschätzung, Motivation

geg.: Kernfunktion K, standardisierte Dichte, z.B. Normaldichte, Epanechnikov-Kern.

Regressionsmodell:

$$
\begin{aligned}
Y &= f(X) + \epsilon, \quad \epsilon \sim (0, \sigma^2) \quad \text{also} \\
\mathbf{E}(Y|X = x) &= f(x) \\
f(x) &= \mathbf{E}(Y|X = x) \\
&= \int y f_{Y|X}(y|x)\, dy \\
&= \int y \frac{g(x,y)}{f_0(x)}\, dy \\
&= \frac{\int y g(x,y) dy}{f_0(x)}
\end{aligned}
$$

Regression
Kernschätzung

$$f(x) = \frac{\int y g(x,y) dy}{f_0(x)}$$

$g(x,y)$: gemeinsame Dichte von (X,Y)
$f_0(x)$: Randdichte von X
$f_{Y|X}$: bedingte Dichte von Y

Der Nenner wird geschätzt durch

$$\hat{f}_0(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \cdot K(\frac{x - X_i}{h})$$

und der Zähler durch

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \hat{g}(X_i, Y_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot \frac{1}{h} \cdot K(\frac{x - X_i}{h})$$

Regression
Kernschätzung

Beide zusammen ergeben die

Kernschätzung

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i \cdot \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}$$

K: Kernfunktion
h: Glättungsparameter

Nichtparametrische Kurvenschätzung

Spline- und Kernschätzung

llustration: SAS-INSIGHT.

Analyse

Fit(Y X)

Output

Nonparametric Curves

Smoothing Spline

Normal kernel smoother

- Venus-Muschel Daten (WORK/Descr_Scatter)

- Heroin-Daten (SASUSER/heroin) (TIME-DOSE)

Glättende Splines können auch mit Hilfe der Prozedur GPLOT erzeugt werden.

Zeichnen von Funktionen mit der Prozedur GPLOT, die SYMBOL-Anweisung

SYMBOLnr I= (I steht für INTERPOL)

I=needle	Nadelplot	diskrete Wktn.
I=spline	interpolierender Spline	glatte Kurven
I=SMnnS	glättender Spline nn: Glättungsparameter	glatte Kurven
I=RL	Regressionsgerade	
I=RQ	quadratische Regressionskurve	
I=RC	kubische Regressionskurve	
I=RLCLI	Konfidenzbereiche für Beobachtungen	
I=RLCLM	Konfidenzbereiche für Regressionsgerade	

Beschreibende Statistik

Zusammenfassung (1)

Verteilungsfunktion

$$F(x) = P(X \leq x)$$

diskrete Verteilung

$$F(x) = \sum_{i:i \leq x} p_i \qquad p_i = P(X = x_i)$$

stetige Verteilung

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad f(t): \text{ Dichte.}$$

Bsp: diskrete Verteilung: Binomial, Poisson
stetige Verteilung: Normal,Gleich,Exp

Beschreibende Statistik

Zusammenfassung (2)

Erwartungswert

$$\mathbf{E}(X) = \begin{cases} \sum x_i p_i & X \text{ diskret} \\ \int x f(x) dx & X \text{ stetig} \end{cases}$$

Varianz

$$var(X) = \mathbf{E}(X - \mathbf{E}X)^2$$

Normalverteilung, Dichte

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{2}} \quad \text{Standard}$$

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Beschreibende Statistik

Zusammenfassung (3)

Gesetz der Großen Zahlen ($\mathbf{E}(X) < \infty$)

$$\overline{X} \longrightarrow \mathbf{E}X.$$

Zentraler Grenzwertsatz (X_i iid)

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \longrightarrow Z \sim \mathcal{N}(0,1)$$

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{s} \longrightarrow Z \sim \mathcal{N}(0,1)$$

$$\begin{aligned} \overline{X} &= \frac{1}{n} \sum X_i \\ s^2 &= \frac{1}{n-1} \sum (X_i - \overline{X})^2 \rightarrow \sigma^2 \end{aligned}$$

Beschreibende Statistik

Zusammenfassung (4)

Statistische Maßzahlen

Lagemaße: $\overline{X}, x_{0.5}, x_{\alpha}, x_{0.25}, x_{0.75}, \overline{x}_{\alpha}, \overline{x}_{\alpha,w}$

Skalenmaße: $s^2, s, R, IR, MAD, S_n, Q_n$

Formmaße: β_1, β_2

PROC UNIVARIATE
PROC UNIVARIATE ROBUSTSCALE
PROC UNIVARIATE TRIMMED=
PROC UNIVARIATE WINSORIZED=
PROC MEANS MEDIAN STD

Beschreibende Statistik

Zusammenfassung (5)

Boxplots	**PROC BOXPLOT**
	PROC GPLOT
Häufigkeitsdiagramme	**PROC GCHART**
	PROC UNIVARIATE Pearson, HISTOGRAM
Häufigkeitstabellen:	**PROC FREQ**
Zusammenhangsmaße:	**PROC CORR**

Spearman, Kendall-Korrelationskoeff.

Scatterplots, Regression, Schätzung der Regressionskoeffizienten: **PROC GPLOT**