

# Werkzeuge der Empirischen Forschung

Sommersemester 2019

Wolfgang Kössler

Humboldt-Universität zu Berlin, Institut für Informatik

3. Juli 2019

# Inhalt (1)

- 1 Einleitung
- 2 Dateneingabe und Transformation
- 3 Wahrscheinlichkeitsrechnung
- 4 Statistische Maßzahlen für quantitative Merkmale
- 5 Datenvisualisierung

## Inhalt (2)

- 6 Beschreibung von Zusammenhängen
- 7 Statistische Tests
- 8 Varianzanalyse
- 9 Anpassungstests
- 10 Nichtparametrische Tests

# Inhalt (3)

11 Korrelation und Regression

12 Zufallszahlen

13 Clusteranalyse

14 Hauptkomponentenanalyse

15 Faktorenanalyse

# Inhalt (4)

## 16 Zusammenfassung

# Statistik und Wahrscheinlichkeitsrechnung

## Stochastik

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

# Statistik und Wahrscheinlichkeitsrechnung

## Stochastik

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

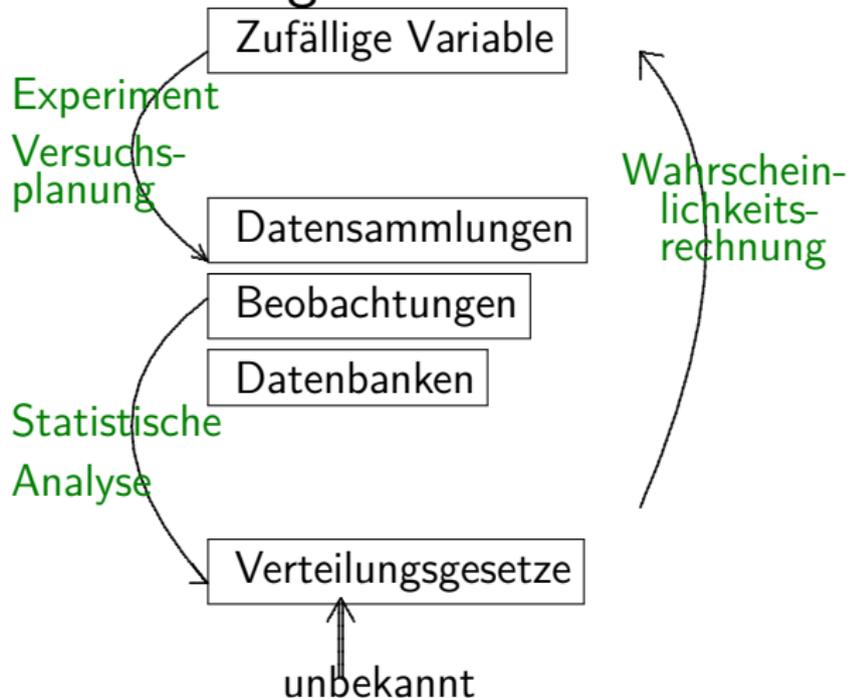
## Wahrscheinlichkeitsrechnung

gegebene Grundgesamtheit (Verteilung) → Aussagen über Realisierungen einer Zufallsvariablen treffen.

## 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
  - Explorative oder Hypothesen-generierende Statistik (data mining)

# Überblick: Statistik und Wahrscheinlichkeitsrechnung

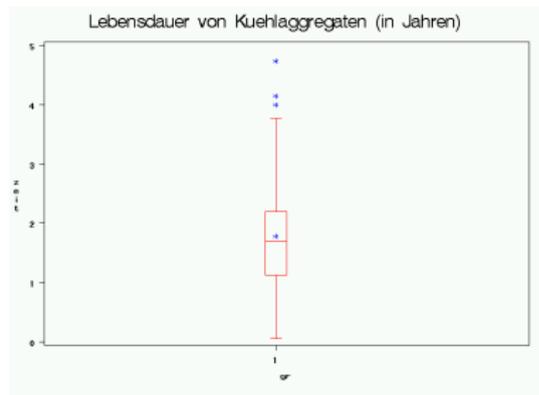


## Beschreibende Statistik

- statistische Maßzahlen: Mittelwerte, Streuungen, Quantile, ...
- Box-Blots
- Q-Q Plots
- Balkendiagramme
- Zusammenhangsmaße
- Punktediagramme (Scatterplots)

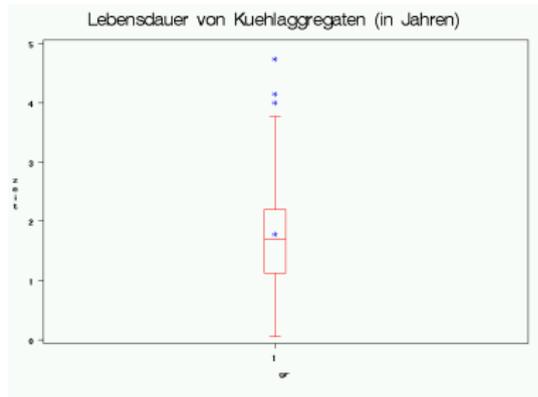
# Boxplots - Beispiele

Lebensdauern von 100  
Kühlaggregate

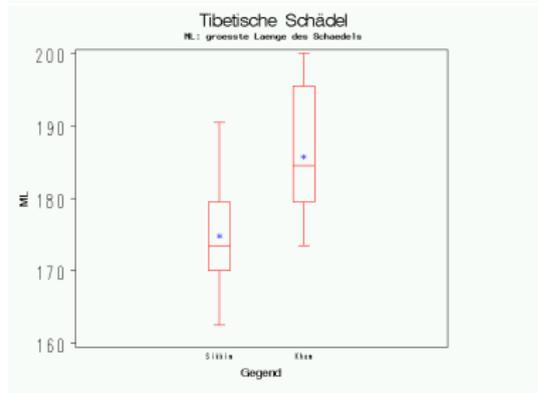


# Boxplots - Beispiele

Lebensdauern von 100  
Kühlaggregaten



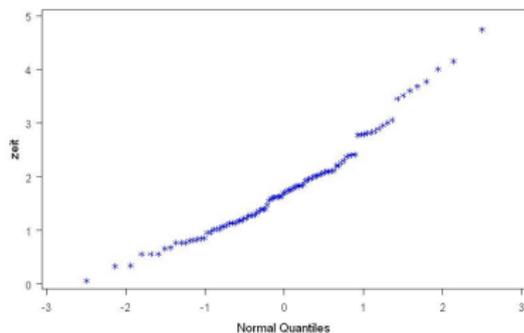
Schädelmaße in zwei  
Regionen Tibets



# Q-Q Plots - Beispiele (1/2)

Lebensdauern von 100  
Kühlaggregaten

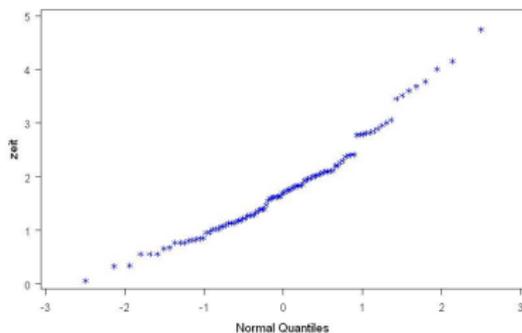
**Lebensdauer von Kühlaggregaten**  
(in Jahren)



# Q-Q Plots - Beispiele (1/2)

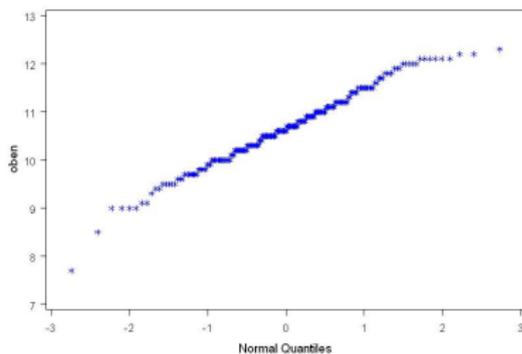
Lebensdauern von 100  
Kühlaggregaten

**Lebensdauer von Kühlaggregaten**  
(in Jahren)



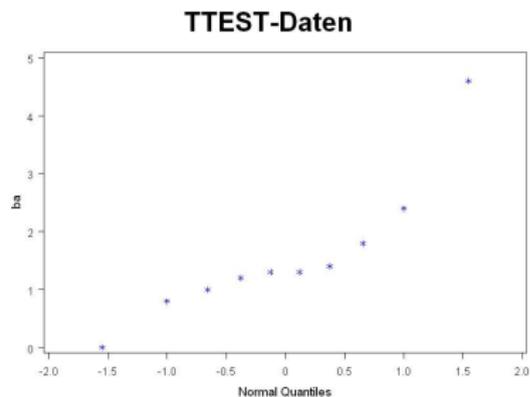
Abmessungen von  
Banknoten

**Banknoten, Variable oben**



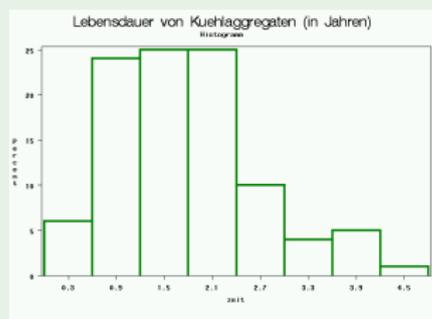
# Q-Q Plots - Beispiele (2/2)

Verlängerung der  
Schlafdauer



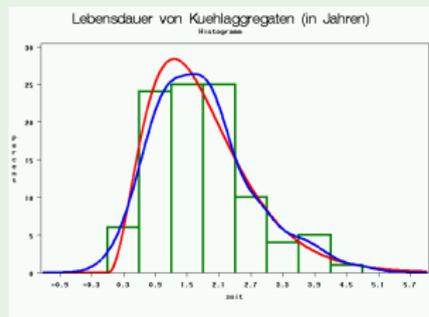
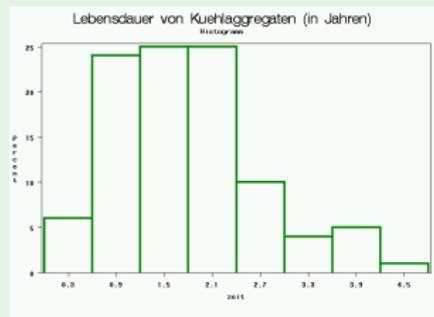
# Dichteschätzung, Beispiel

## Kühlaggregate



# Dichteschätzung, Beispiel

## Kühlaggregate

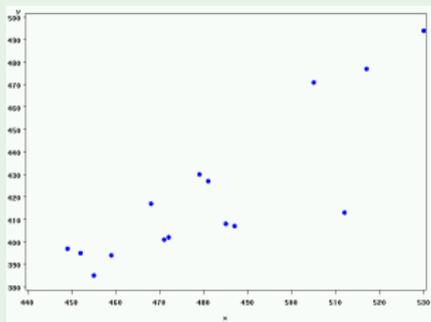


## Histogramm

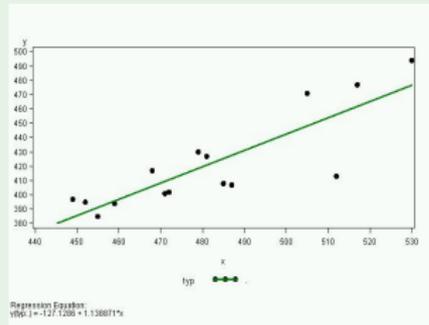
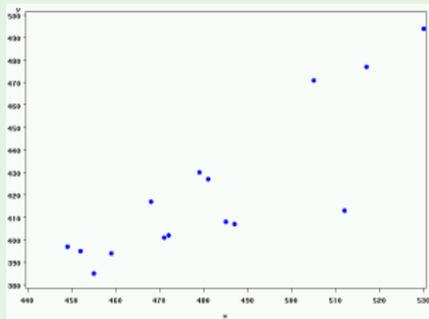
Parametrische Dichteschätzung (Gamma)

Nichtparametrische Dichteschätzung

## Länge und Breite von Venusmuscheln



## Länge und Breite von Venusmuscheln



## Schließende Statistik (1)

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

## Schließende Statistik (2)

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

## 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

# Ein- und Zweistichprobenproblem

## Eine Stichprobe

- Banknoten: vorgegebene Länge eingehalten?

→ Einstichproben t-Test, Signed-Wilcoxon-Test

# 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

# Mehrstichprobenproblem

## Abhängige und Unabhängige Stichproben

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

## Ursache - Wirkungsanalysen

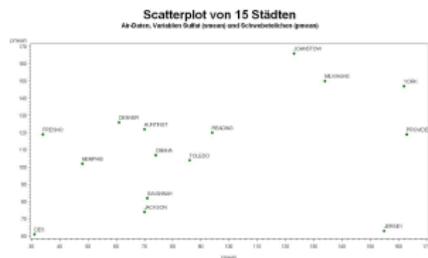
- Ursache - Wirkungsanalysen
    - Zusammenhangsanalyse
    - Bestimmen funktionaler Beziehungen
    - Trends, Vorhersagen
  - Beispiele:
    - Bluthochdruck - Rauchgewohnheiten
    - Blutdruck - Proteinuria
    - Größe - Gewicht
    - Sterblichkeit - Wasserhärte
- Lineare, Nichtlineare und  
Nichtparametrische Regression
- Korrelation

## Klassifikation

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

# Hierarchische Clusteranalyse

Beispiel: Luftverschmutzung in USA-Städten

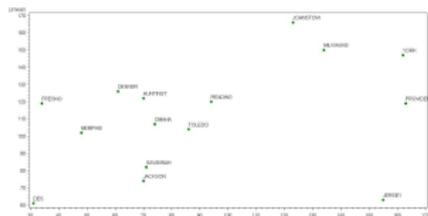




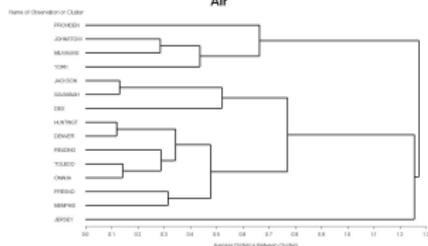
# Hierarchische Clusteranalyse

Beispiel: Luftverschmutzung in USA-Städten

Scatterplot von 15 Städten  
Air-Qualität, Variablen Sulfat (sulfate) und Schwefeldioxid (sulfur)



Air

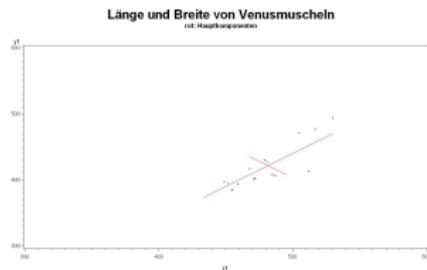


Complete Linkage Cluster Analyse



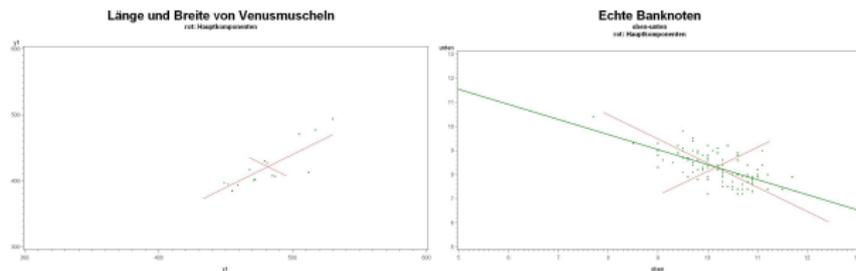
# Hauptkomponentenanalyse

## Beispiele



# Hauptkomponentenanalyse

## Beispiele



**Frage:** Wie kann man diese ausgezeichnete Richtung erfassen?

## 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-Online Dokumentation, SAS-Handbücher

## Literatur (2)

Steland A. (2013). Basiswissen Statistik, Springer.

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

Sachs (1999). Angewandte Statistik, Springer.

Handl, A. (2002). Multivariate Analysemethoden, Springer.

Schlittgen, R. (2008). Einführung in die Statistik, Oldenbourg.

Backhaus, Erichsen, Plinke, Weiber (2010). Multivariate Analysemethoden, Springer.

Büning, Trenkler (1994). Nichtparametrische Statistische Methoden, DeGruyter Berlin.

Bortz, J. (1999). Statistik für Sozialwissenschaftler, Springer.

## Statistik-Software

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

STATA, STATGRAPHICS, XPLORE, MATHEMATICA, MATLAB ..

	SAS	R	
Umfang	+	+	
Verfügbarkeit	+	++	
Preis	(-)	++	
Validierung	+	-	
Dokumentation	+	-	
Große Datensätze	+	-	
User Community	+	+	
Graphik		+	
Kontinuität	+	Kern gut,	Zusatzpakete ?
Haftung	?	?	
Erlernbarkeit	+	+	

## Starten und Beenden von SAS

- Starten von SAS

1. beim Terminal-Server orkan oder tornado einloggen:

`xfreerdp -v orkan -u Nutzername` Passwort angeben;

2. Start von SAS: Icon anklicken oder

`Programme > Mathematik > SAS > SAS 9.4 (English)`

- Starten von SAS von außerhalb: hier ist die Anleitung

<https://www2.informatik.hu-berlin.de/~tmstern/werkzeuge/ts-zugang/>

- Beenden der Sitzung

`Datei beenden > Logoff`

# Allgemeine Struktur von SAS (1)

## 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

# Allgemeine Struktur von SAS (2)

## Hilfen

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

# Allgemeine Struktur eines SAS-Programms

## Aufbau einer SAS-Datei

DATA

PROC

DATA

PROC

PROC

...

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

## Daten (1)

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

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

$x_{ij}$ : Wert oder  
Ausprägung des  
Merkmals j am  
Objekt i

## Daten (2)

$p$ : Anzahl der Merkmale

$N$ : Gesamtanzahl der einbezogenen Objekte (Individuen)

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

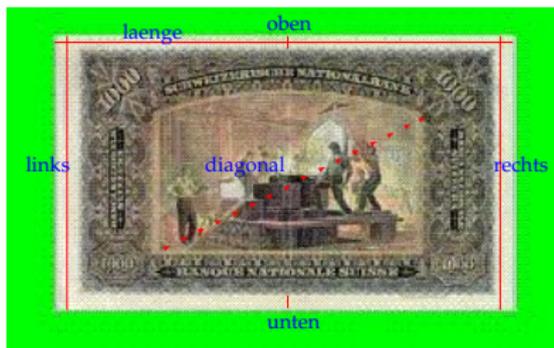
Qualität des Datenmaterials wird im Wesentlichen durch die Auswahl der Objekte aus einer größeren Grundgesamtheit bestimmt.

## 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

# Ein 1000-Franken Schein

1



## Daten (4), Datenmatrix

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

Banknote	Merkmale						
	laenge	oben	unten	..	j	..	gr
1							
2							
3							
..							
i							
..							

 $x_{ij}$

## Daten (5), 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							
..							

# Inhalt

- 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

**DATA** <dateiname <(dateioptionen)»;

...

**RUN**;

<... > kennzeichnet optionale Parameter

**DATA** <dateiname <(dateioptionen)»;

...

**RUN**;

<... > kennzeichnet optionale Parameter

Externes File

```
INFILE ' ... ' ;
```

```
INPUT ... ;
```

**DATA** <dateiname <(dateioptionen)>>;

...

**RUN**;

<... > kennzeichnet optionale Parameter

Externes File

```
INFILE ' ... ' ;
```

```
INPUT ... ;
```

SAS-System-File

```
SET SAS-dateiname;
```

**DATA** <dateiname <(dateioptionen)>>;

...

**RUN**;

<... > kennzeichnet optionale Parameter

Externes File

```
INFILE ' ... ' ;
```

```
INPUT ... ;
```

SAS-System-File

```
SET SAS-dateiname;
```

Tastatur

```
INPUT ... ;
```

```
CARDS;
```

Datenmatrix

```
;
```

+ zusätzliche Anweisungen

Programmbeispiele: Eingabe... .sas

# Inhalt

- 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

# 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.

## Aktivierung des Programms

- klicken auf MännchenLogoGrafik oder
- klicken auf 'run' → 'submit' oder
- F3-Taste oder F8-Taste

## Die Datei Eingabe1 hat

3 Beobachtungen (Individuen, Wertesätze)

5 Merkmale (Variablen) a, x, y, z und s.

# Dateneingabe und Transformation

Wo werden die Daten abgelegt?

	Bibliothek	Dateiname
--	------------	-----------

---

DATA Eingabe1;	WORK	Eingabe1
----------------	------	----------

DATA sasuser.Eing1;	SASUSER	Eing1
---------------------	---------	-------

DATA MyBib.Eing1;	MyBib	Eing1
-------------------	-------	-------

DATA;	WORK	DATA1
-------	------	-------

		DATA2 ...
--	--	-----------

# Dateneingabe und Transformation

Wo werden die Daten abgelegt?

	Bibliothek	Dateiname
DATA Eingabe1;	WORK	Eingabe1
DATA sasuser.Eing1;	SASUSER	Eing1
DATA MyBib.Eing1;	MyBib	Eing1
DATA;	WORK	DATA1 DATA2 ...

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

# Dateneingabe und Transformation

Wo werden die Daten abgelegt?

	Bibliothek	Dateiname
DATA Eingabe1;	WORK	Eingabe1
DATA sasuser.Eing1;	SASUSER	Eing1
DATA MyBib.Eing1;	MyBib	Eing1
DATA;	WORK	DATA1 DATA2 ...

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

Eigene Bibliotheken erstellen

LIBNAME MyBib Pfadname

# Dateneingabe und Transformation

Automatisch generierte Variablen

*\_N\_*

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.

# Inhalt

- 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

Transformationen immer nach der INPUT-Anweisung angeben!

IF THEN ELSE und logische Operationen

vgl. Programm [Eingabe2](#)

Funktionen

vgl. Programm [Eingabe3](#)

Transformationen immer nach der INPUT-Anweisung angeben!

IF THEN ELSE und logische Operationen

vgl. Programm Eingabe2

Funktionen

vgl. Programm Eingabe3

Arithmetische Operationen

$+, -, *, /, **$

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)

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;
```

# 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;

# Inhalt

- 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

**DATA** Eingabe4;

**INFILE** 'Pfadname';

**INPUT** Variablen;

evtl. Transformationen;

**RUN**;

**DATA** Eingabe4;

**INFILE** 'Pfadname';

**INPUT** Variablen;

evtl. Transformationen;

**RUN**;

**DATA** Eingabe4url;

**FILENAME** fname URL 'http:// ...!';

**INFILE** fname;

**INPUT** Variablen;

**RUN**;

**DATA** Eingabe4;

**INFILE** 'Pfadname';

**INPUT** Variablen;

evtl. Transformationen;

**RUN**;

**DATA** Eingabe4url;

**FILENAME** fname URL 'http:// ...';

**INFILE** fname;

**INPUT** Variablen;

**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"  
dbms=excel  
out=Dateiname; /*SAS-Datei*/  
getnames=no; /*Variablennamen werden nicht übernommen*/  
sheet=spreadsheetname;  
RUN;
```

# Inhalt

- 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

- 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.sas](#)

[Eingabe6.sas \(komplexere Formate\)](#)

# 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

# Inhalt

- 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

## 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;**

## 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.

# 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** Maxlength \$12;

# Inhalt

- 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

## Files 'untereinander'

```
SASfile_1
```

```
...
```

```
SASfile_n
```

```
DATA; /* Eingabe_Banknote13.sas */
```

```
SET SASfile_1 <(options)>... SASfile_n<(options)>;
```

```
RUN;
```

## 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.

# 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.

Eingabe\_Merge.sas

# Inhalt

- 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

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

Eingabe12.sas

# Inhalt

- 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

## Allgemeine Syntax

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

Eingabe13.sas

Eingabe14.sas

Eingabe15.sas

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

## Eine Grundgesamtheit (oder Population)

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

### Beispiele

- 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

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, wobei jede Stichprobe gleichen Umfangs gleichwahrscheinlich ist.

(oder: bei der jedes Element mit 'der gleichen Wahrscheinlichkeit' ausgewählt wird).

**Bemerkung:** Ein (auszuwertender) Datensatz ist (i.d.R.) eine Stichprobe.

# Klassifikation von Merkmalen

## Nominale Merkmale

Die Ausprägungen sind lediglich Bezeichnungen für Zustände oder Sachverhalte.

Sie können auch durch Zahlen kodiert sein!

## Beispiele

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

## Ordinale Merkmale (Rangskala)

Die Menge der Merkmalsausprägungen besitzt eine Rangordnung!

### Beispiele

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

## Metrische Merkmale (kardinale/quantitative M.)

Werte können auf der Zahlengeraden aufgetragen werden (metrische Skala)

### Beispiele

Messwerte, Längen, Größen, Gewichte, Alter

Metrische Merkmale werden unterschieden nach:

### Diskrete Merkmale

nehmen höchstens abzählbar viele Werte an.

### Beispiele

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.

## Der Stichprobenraum $\Omega$ eines zufälligen Experiments

ist die Menge aller möglichen Versuchsausgänge

Die Elemente  $\omega$  des Stichprobenraums  $\Omega$  heißen Elementarereignisse.

### Beispiele

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

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

### Beispiele

Lebensdauer  $\leq 10$  min.

Augensumme gerade.

Warteschlange hat Länge von  $\leq 10$  Personen.

Realisierungen sind die Ergebnisse des Experiments

(die realisierten Elemente von  $\Omega$ )

Verknüpfungen von Ereignissen werden durch entsprechende Mengenverknüpfungen beschrieben

$A \cup B$       A oder B tritt ein

$A \cap B$       A und B tritt ein

$\bar{A} = \Omega \setminus A$     A tritt nicht ein.

Forderung (damit die Verknüpfungen auch immer ausgeführt werden können):

Die Ereignisse liegen in einem Ereignisfeld ( $\sigma$ -Algebra)  $\mathfrak{E}$ .

### Ereignisfeld

Das Mengensystem  $\mathfrak{E} \subseteq \mathfrak{P}(\Omega)$  heißt Ereignisfeld, falls gilt:

1.  $\Omega \in \mathfrak{E}$
2.  $A \in \mathfrak{E} \implies \bar{A} \in \mathfrak{E}$
3.  $A_i \in \mathfrak{E}, i = 1, 2, \dots \implies \bigcup_{i=1}^{\infty} A_i \in \mathfrak{E}$ .

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

# Das Axiomensystem 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,  $A_i \in \mathfrak{E}$ ,

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

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

# Wahrscheinlichkeit

## Eigenschaften (1)

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

Beweis:

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

# Wahrscheinlichkeit

## Eigenschaften (2)

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

Beweis:

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

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

Eine (messbare) Abbildung heißt Zufallsvariable.

$$X : \Omega \longrightarrow \mathbb{R}$$

$$\omega \longmapsto r$$

Eine (messbare) Abbildung heißt Zufallsvariable.

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

### 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.  
( $p_i > 0$ )

### Beispiele

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

## Stetige Zufallsvariable

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

## Beispiele

- 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  $X$

Manchmal wird die Verteilungsfunktion auch durch  $P(X < x)$  definiert.

## Stetige Zufallsvariable

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

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

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

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

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

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

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

## 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.

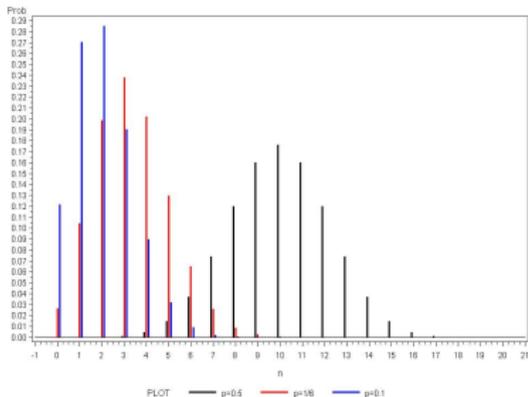
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Binomialwahrscheinlichkeit}$$

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

# Wahrscheinlichkeitsfunktionen

## Binomial

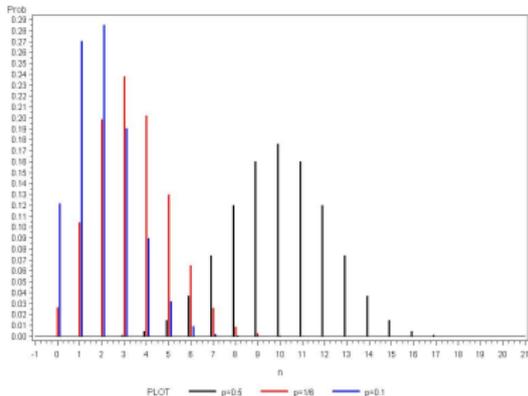
Binomial-Verteilung mit  $n=20$  und  $p=0.5, 1/6, 0.1$



# Wahrscheinlichkeitsfunktionen

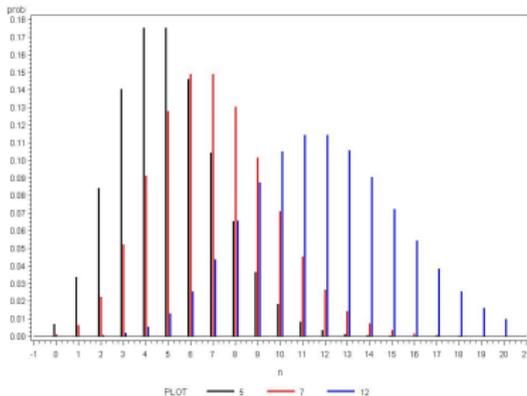
## Binomial

Binomial-Verteilung mit  $n=20$  und  $p=0.5, 1/6, 0.1$



## Poisson

Poisson-Verteilung mit  $\lambda=5, 7, 12$





Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

## Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

$X$ : Anzahl der Sechsen.

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

## Würfel 20 mal. Wkt. für mindestens 4 Sechsen?

$X$ : Anzahl der Sechsen.

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

$$= 1 - \left(\frac{5}{6}\right)^{20} - 20\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{19} - \frac{20 \cdot 19}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{18} - \\ - \frac{20 \cdot 19 \cdot 18}{6}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{17}$$

$$= 1 - \text{CDF}(\text{'Binomial'}, 3, 1/6, 20)$$

$$= \text{SDF}(\text{'Binomial'}, 3, 1/6, 20)$$

$$\approx 0.43.$$

# 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  $\lambda$  ein noch unbestimmter Parameter ist. Er kann als mittlere Rate aufgefasst werden.

## 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 Zufallsvariablen

$$N_{t+h} - N_t \text{ und } N_t - N_{t-h} \text{ sind unabhängig}$$

V2: es ist egal wo wir das Zeitintervall betrachten, dh.

$$N_{t+h} - N_t \text{ und } N_t - N_{t-h} \text{ haben dieselbe Verteilung}$$

V3: Wahrscheinlichkeit, dass mindestens ein Ereignis in der Zeit  $h$  eintritt, z.B. ein Kunde ankommt.

$$p(h) = a \cdot h + o(h), \quad a > 0, h \rightarrow 0$$

V4: Wahrscheinlichkeit für  $k \geq 2$  Ereignisse in der Zeit  $h$ :  $o(h)$

## Poisson (3)

Frage: Wahrscheinlichkeit, dass bis zum Zeitpunkt  $t$  genau  $i$  Ereignisse? (eingetroffene Kunden, zerfallene Teilchen) eintreten?

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

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

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

Beweis: Stochastik-Vorlesung.

# Poisson (4)

## Binomial und Poisson

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

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

Beweis:

# Poisson (4)

## Binomial und Poisson

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

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

Beweis:

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

# Poisson (4)

## Binomial und Poisson

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

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

Beweis:

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

# Poisson (4)

## Binomial und Poisson

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

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

Beweis:

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

# Geometrische Verteilung

Münzwurf solange bis B(Blatt) kommt

$$\Omega = \{B, ZB, ZZB, \dots\}$$

$X :=$  Anzahl der Fehl-Würfe (Misserfolge) vor dem ersten Blatt.

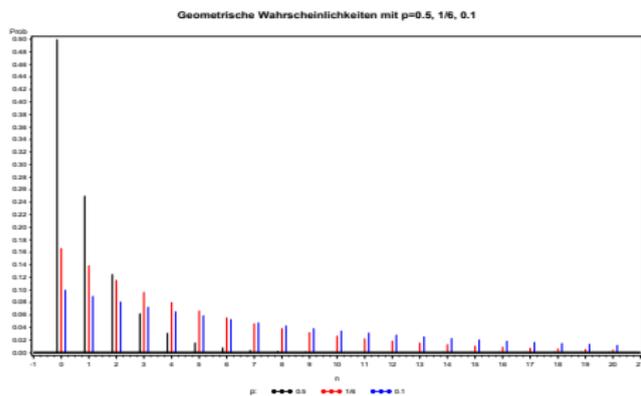
$$X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & \dots & n & \dots \\ 1/2 & (1/2)^2 & (1/2)^3 & (1/2)^4 & (1/2)^5 & \dots & (1/2)^{n+1} & \dots \end{pmatrix}$$

$$\sum_{i=0}^{\infty} p_i = \sum_{i=1}^{\infty} (1/2)^i = \frac{1}{1 - \frac{1}{2}} - 1 = 1 \quad \text{geometrische Reihe}$$

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

allgemeiner:  $P(X = i) = p_i = p(1 - p)^i$ .

# Geometrische Verteilung



## Qualitätskontrolle

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

## Qualitätskontrolle

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

$X$ : Anzahl der schlechten Stücke in der Stichprobe.

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

	schlecht	gut	total
Stichprobe	$k$		$m$
Rest			$N - m$
total	$n$	$N - n$	$N$

$\binom{N}{m}$ : # möglichen Stichproben

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

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

## Hypergeometrische Verteilung (2)

Offenbar:  $0 \leq x \leq \min(n, m), \quad m - x \leq N - n.$

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.

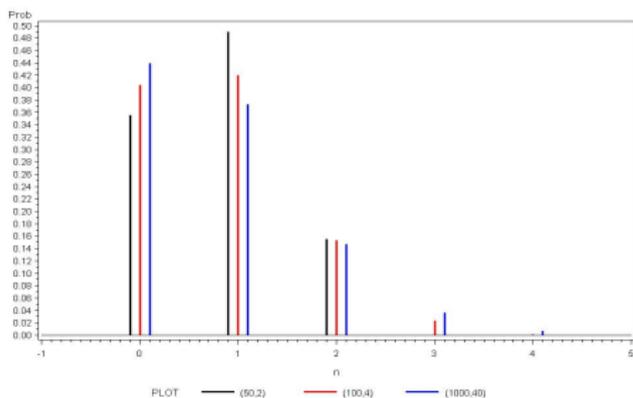
Bemerkung: Für  $N \rightarrow \infty, n \rightarrow \infty, \frac{n}{N} \rightarrow p$  gilt:

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

# Hypergeometrische Verteilung

## Hypergeometrische Verteilung mit $m=20$

und  $(N,n)=(1000,40), (100,4), (50,2)$



# Multinomialverteilung

Wir betrachten ein zufälliges Experiment mit den Ausgängen  $A_1, A_2, \dots, A_l$ . Wir setzen  $p_i = P(A_i)$ ,  $\sum_{i=1}^l p_i = 1$ .

Es sei ein Behälter mit  $k$  Kugeln in  $l$  verschiedenen Farben gegeben, wobei  $k_i$  Kugeln die Farbe  $i$  ( $i = 1, \dots, l$ ) besitzen,  $\sum_{i=1}^l k_i = k$ .  
Wahrscheinlichkeit, mit der eine Kugel einer bestimmten Farbe aus dem Behälter entnommen wird:

# Multinomialverteilung

Wir betrachten ein zufälliges Experiment mit den Ausgängen  $A_1, A_2, \dots, A_l$ . Wir setzen  $p_i = P(A_i)$ ,  $\sum_{i=1}^l p_i = 1$ .

Es sei ein Behälter mit  $k$  Kugeln in  $l$  verschiedenen Farben gegeben, wobei  $k_i$  Kugeln die Farbe  $i$  ( $i = 1, \dots, l$ ) besitzen,  $\sum_{i=1}^l k_i = k$ .  
Wahrscheinlichkeit, mit der eine Kugel einer bestimmten Farbe aus dem Behälter entnommen wird:

$$P(\text{Kugel der Farbe } i) = p_i = \frac{k_i}{k}.$$

## Multinomiale Wahrscheinlichkeiten (2)

Das Experiment soll nun  $n$ -mal wiederholt werden.

$B_{n_1, n_2, \dots, n_l}$ : das Ereignis, daß die Ereignisse  $A_1$   $n_1$ -mal,  $A_2$   $n_2$ -mal,  $\dots$ , und  $A_l$   $n_l$ -mal eintreten.

## Multinomiale Wahrscheinlichkeiten (2)

Das Experiment soll nun  $n$ -mal wiederholt werden.

$B_{n_1, n_2, \dots, n_l}$ : das Ereignis, daß die Ereignisse  $A_1$   $n_1$ -mal,  $A_2$   $n_2$ -mal,  $\dots$ , und  $A_l$   $n_l$ -mal eintreten.

$$P(B_{n_1, n_2, \dots, n_l}) = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_l!} \cdot p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_l^{n_l}.$$

Derartige Wahrscheinlichkeiten bezeichnen wir auch als multinomiale Wahrscheinlichkeiten (polynomiale Wktn.)

# Potenzen von Summen

Vergleichen Sie:

$$(a_1 + \dots + a_l)^n = \sum \frac{n!}{n_1! \dots n_l!} a_1^{n_1} \dots a_l^{n_l}$$

wobei die Summe über alle Tupel  $(n_1, \dots, n_l)$  gebildet wird mit

$$\sum_{i=1}^l n_i = n.$$

# Multinomiale Wahrscheinlichkeiten (3)

## Fragebogen

Bei einem Fragebogen wird (u.a.) nach dem Alter der befragten Personen gefragt. Das Alter sei in Klassen eingeteilt, 10-20, 21-40, 41-60, über 60 Jahre. Der Bevölkerungsanteil beträgt jeweils  $p_i$  für die  $i$ -te Altersklasse,  $i = 1, \dots, 4$ ,  $\sum_i p_i = 1$ .

Es werden  $n=1000$  Personen befragt.

## SAS-Anweisungen

$X \sim Bi(p, n)$	CDF('Binomial',m,p,n)	PDF('Binomial',m,p,n)
$X \sim Poi(\lambda)$	CDF('Poisson',m, $\lambda$ )	PDF('Poisson',m, $\lambda$ )
$X \sim Geo(p)$	CDF('Geometric',m-1,p)	PDF('Geometric',m-1,p)
$X \sim H(N, n, m)$	CDF('Hyper',k,N,n,m)	PDF('Hyper',k,N,n,m)

**Bem.:** Bei SAS weicht bei der geometrischen Verteilung die Parametrisierung von der üblichen Definition ab.

[Descr\\_Binomial\\_neu.sas](#)

[Descr\\_Poisson.sas](#)

[Descr\\_Geometr2015.sas](#)

[Descr\\_Hypergeom.sas](#)

## SAS-Anweisungen

$X \sim Bi(p, n)$	CDF('Binomial',m,p,n)	PDF('Binomial',m,p,n)
$X \sim Poi(\lambda)$	CDF('Poisson',m, $\lambda$ )	PDF('Poisson',m, $\lambda$ )
$X \sim Geo(p)$	CDF('Geometric',m-1,p)	PDF('Geometric',m-1,p)
$X \sim H(N, n, m)$	CDF('Hyper',k,N,n,m)	PDF('Hyper',k,N,n,m)

**Bem.:** Bei SAS weicht bei der geometrischen Verteilung die Parametrisierung von der üblichen Definition ab.

[Descr\\_Binomial\\_neu.sas](#)      [Descr\\_Poisson.sas](#)

[Descr\\_Geometr2015.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! → Aufgabe der Statistik

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

Sei  $X$  stetig auf  $(a,b)$ , wobei  $a, b$  unendlich sein können,

$$a \leq x_0 < x_1 \leq b$$

$$P(X = x_0) = 0, \quad P(x_0 < X < x_1) > 0 \text{ (wenn } f > 0 \text{)}.$$

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 unbekanntem Parametern ab, die geschätzt werden müssen.

# Beispiele

Gleich- und Exponentialverteilung

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

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

- Referenzverteilung
- Zufallszahlen

# Beispiele

Gleich- und Exponentialverteilung

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

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

- Referenzverteilung      - Zufallszahlen

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

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

- 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-Funktionalgleichung

Eine Verteilung ist also gedächtnislos, gdw.

$$\frac{P(X \geq s + t)}{P(X \geq t)} = P(X \geq s) \quad \text{gdw.} \quad \frac{1 - F(s + t)}{1 - F(t)} = 1 - F(s).$$

# Gedächtnislosigkeit

Cauchy-Funktionalgleichung

Eine Verteilung ist also gedächtnislos, gdw.

$$\frac{P(X \geq s+t)}{P(X \geq t)} = P(X \geq s) \quad \text{gdw.} \quad \frac{1 - F(s+t)}{1 - F(t)} = 1 - F(s).$$

Überlebensfunktion (oder Zuverlässigkeitsfunktion)

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

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

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

# Cauchy-Funktionalgleichung

Eine Lösung

**Satz:** Die Exponentialverteilung ist gedächtnislos.

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

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

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

$$F(0) = 0 \quad \text{und} \quad G(t) = 1 - F(t).$$

Es gelte die Cauchy-Funktionalgleichung

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

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

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

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

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

Bez.:  $X \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\mu$ : Lageparameter,  $\sigma$ : Skalenparameter

Normalverteilung: wichtigste Verteilung in der Statistik

warum?  $\rightarrow$  später.

# SAS-Anweisungen

PDF('Exponential',x, $\lambda$ )

Dichtefunktion

CDF('Exponential',x, $\lambda$ )

Verteilungsfunktion

PDF('Normal',x, $\mu, \sigma$ )

Dichtefunktion

CDF('Normal',x, $\mu, \sigma$ )

Verteilungsfunktion

PROBNORM(x, $\mu, \sigma$ )

Quantile('Normal',u, $\mu, \sigma$ )

Quantilfunktion

PROBIT(u, $\mu, \sigma$ )

## Weitere wichtige Verteilungen

Weibull-Verteilung     $\text{CDF}(\text{'Weibull'}, x, a, \lambda)$

Gamma-Verteilung     $\text{CDF}(\text{'Gamma'}, x, a, \lambda)$

$\chi^2$ -Verteilung     $\text{CDF}(\text{'Chisq'}, x, \nu, \lambda)$

t-Verteilung     $\text{CDF}(\text{'t'}, x, \nu, \delta)$

F-Verteilung     $\text{CDF}(\text{'F'}, x, \nu_1, \nu_2, \delta)$

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

[Descr\\_Weibull](#)

[Descr\\_Gamma](#)

# Wahrscheinlichkeitsverteilungen in SAS (1)

↪ 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 (2)

CDF('Verteilung',x,Parameterliste)      Verteilungsfunktion

PDF('Verteilung',x,Parameterliste)      Dichtefunktion  
(Wahrscheinlichkeitsfunktion)

SDF ('Verteilung',x,Parameterliste)      = 1-CDF  
Überlebensfunktion       $(1 - F(x))$

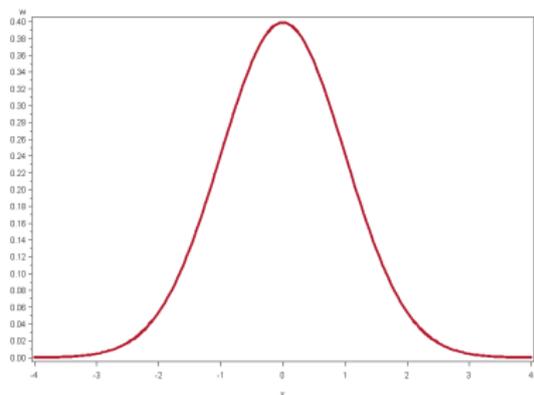
Quantile('Verteilung',u,Parameterliste)      Quantilfunktion

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

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

Dichtefunktion der Standard-Normalverteilung



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

Gauß

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

Eine Zufallsvariable mit dieser Dichte  $f(x)$  heißt normalverteilt mit Parametern  $\mu$  und  $\sigma^2$ .

## Normalverteilung (2)

**Satz:**  $f$  auf der letzten Folie ist Dichte.

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

2. bleibt z.z.

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

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 (3)

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 (4)

Substitution:

$$s := \frac{x - \mu}{\sigma} \quad t := \frac{y - \mu}{\sigma}.$$
$$dx = \sigma ds \quad dy = \sigma dt.$$

Wir erhalten damit:

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

Weitere Substitution (Polarkoordinaten):

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

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 (6)

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

## Normalverteilung (6)

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

# Normalverteilung

## Standard-Normalverteilung

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

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

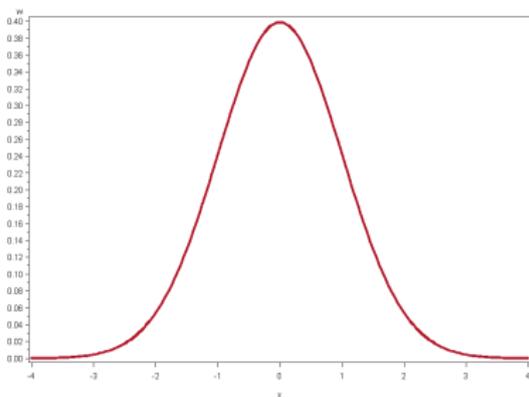
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad \text{Verteilungsfunktion}$$

$\varphi(x)$ ,  $\Phi(x)$  sind tabelliert.

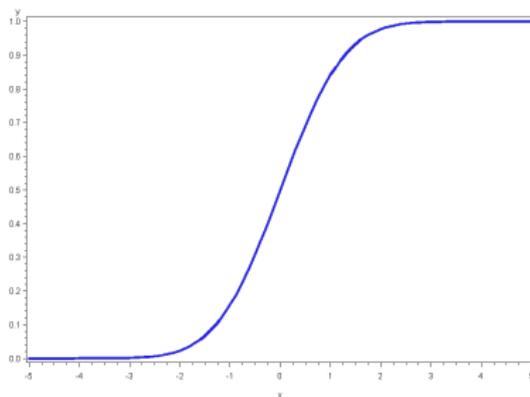
Es geht auch einfacher mit CDF und PDF.

# Standardnormalverteilung (1)

Dichtefunktion der Standard-Normalverteilung



Verteilungsfunktion der Standard-Normalverteilung



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

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

$$P(a < X < b) = \Phi(b) - \Phi(a)$$

Descr\_normal.sas

# Standardnormalverteilung (2)

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

# Standardnormalverteilung (2)

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

$x = \Phi^{-1}(\alpha)$   $\alpha$ -Quantil.

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

SAS: QUANTILE('normal', $\alpha$ ,0,1)

# Normalverteilung

Beziehung zur Standard-Normalverteilung

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

# Normalverteilung

Beziehung zur Standard-Normalverteilung

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

**Satz.** Es gilt:

$$X \sim \mathcal{N}(0, 1) \iff \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff \alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

# Normalverteilung

Beziehung zur Standard-Normalverteilung

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

**Satz.** Es gilt:

$$X \sim \mathcal{N}(0, 1) \iff \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff \alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

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

$$\begin{aligned} P(\sigma X + \mu \leq x) &= P\left(X \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \\ &= \int_{-\infty}^{\frac{x - \mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)} du \end{aligned}$$

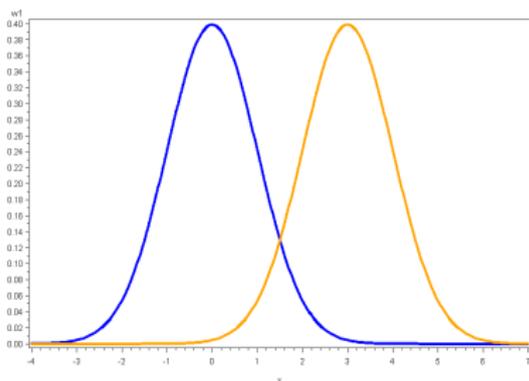
# Normalverteilung

## Unterschiedliche Parameter (1)

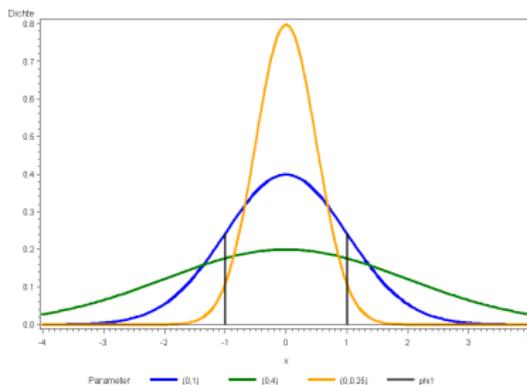
### Vergleichen Sie

- $\sigma^2$  fest,  $\mu$  verschieden
- $\mu$  fest,  $\sigma^2$  verschieden

Dichtefunktion verschiedener Normalverteilungen  
Lageunterschied



Dichtefunktion verschiedener Normalverteilungen  
Skalenunterschied



Descr\_Normal\_1.sas

# Normalverteilung

Unterschiedliche Parameter (2)

Satz: Seien  $X_1 \sim \mathcal{N}(\mu, \sigma_1^2)$ ,  $X_2 \sim \mathcal{N}(\mu, \sigma_2^2)$ ,

$\sigma_1^2 < \sigma_2^2$  und  $a > 0$ . Dann gilt:

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

# Normalverteilung

Unterschiedliche Parameter (2)

Satz: Seien  $X_1 \sim \mathcal{N}(\mu, \sigma_1^2)$ ,  $X_2 \sim \mathcal{N}(\mu, \sigma_2^2)$ ,

$\sigma_1^2 < \sigma_2^2$  und  $a > 0$ . Dann gilt:

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

Beweis:

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

# Normalverteilung

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

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

# Normalverteilung

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

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

$$\begin{aligned}P(9 < X_2 < 11) &= \Phi\left(\frac{11 - 10}{3}\right) - \Phi\left(\frac{9 - 10}{3}\right) \\&= \Phi\left(\frac{1}{3}\right) - \Phi\left(-\frac{1}{3}\right) = 2 \cdot \Phi\left(\frac{1}{3}\right) - 1 \\&= 2 \cdot 0.6306 - 1 = 0.2612.\end{aligned}$$

# 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  $X \sim \text{Exp}(\lambda)$

“gedächtnislose” stetige Verteilung.

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

Zentraler Grenzwertsatz

Fehlergesetz (viele kleine unabhängige Fehler)

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

# 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?

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

# 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

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

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

heißt Erwartungswert von  $X$ .

# Erwartungswert

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

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

$$\mathbf{E}X = \sum_{i=0}^{\infty} p_i i$$

# Erwartungswert

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

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

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

# Erwartungswert

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

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

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

# Erwartungswert

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

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

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

Interpretation: z.B. mittlere Ankunftsrate.

# Erwartungswert

$$X \sim Bi(n, p)$$

$$\mathbf{EX} = \sum_{k=0}^n k \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

# Erwartungswert

$$X \sim Bi(n, p)$$

$$\begin{aligned} \mathbf{EX} &= \sum_{k=0}^n k \binom{n}{k} p^k \cdot (1-p)^{n-k} \\ &= p \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \end{aligned}$$

# Erwartungswert

$$X \sim \text{Bi}(n, p)$$

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

# Erwartungswert

$$X \sim Bi(n, p)$$

$$\begin{aligned}
 \mathbf{EX} &= \sum_{k=0}^n k \binom{n}{k} p^k \cdot (1-p)^{n-k} \\
 &= p \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
 &= p \cdot n \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
 &= p \cdot n \underbrace{\sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-1-i}}_{=1}, \quad k = i + 1
 \end{aligned}$$

# Erwartungswert

$$X \sim \text{Bi}(n, p)$$

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

# 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  $X$ .

# 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  $X$ .

$X \sim \text{Exp}(\lambda)$ ,  $\lambda > 0$

$$\mathbf{E}X = \int_0^{\infty} x \cdot \frac{1}{\lambda} \cdot e^{-\frac{x}{\lambda}} dx = \lambda$$

# Erwartungswert

Normalverteilung

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

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx$$

# Erwartungswert

Normalverteilung

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

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

# Erwartungswert

Normalverteilung

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

$$\begin{aligned} \mathbf{E}X &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-(\frac{x-\mu}{\sigma})^2/2} dx \\ &= \int_{-\infty}^{\infty} (\sigma t + \mu) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad \frac{x - \mu}{\sigma} = t, \quad dx = \sigma dt \\ &= \mu + \frac{1}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} \sigma \cdot t \cdot e^{-\frac{t^2}{2}} dt}_{=0} = \mu. \end{aligned}$$

# Erwartungswert

Gleichverteilung

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

$$\mathbf{EX} = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b$$

# Erwartungswert

Gleichverteilung

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

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

# Eigenschaften des Erwartungswertes

**E** ist Linearer Operator

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

**Beweis:**

# Eigenschaften des Erwartungswertes

**E** ist Linearer Operator

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

**Beweis:** folgt aus Eigenschaften von Reihen und Integralen. ■

# Eigenschaften des Erwartungswertes

## E ist Linearer Operator

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

**Beweis:** folgt aus Eigenschaften von Reihen und Integralen. ■

## Regel des Faulen Statistikers

Sei  $X$  Zufallsvariable,  $g: \mathbb{R} \rightarrow \mathbb{R}$  (rechtsseitig) stetig  $\Rightarrow$

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

vorausgesetzt die Erwartungswerte existieren.

**Beweis:** Transformationsformel (s. Stochastik)

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

Ang., die betrachteten Erwartungswerte existieren.

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

heißt Varianz der Zufallsvariable  $X$ .

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

heißt Standardabweichung der Zufallsvariablen  $X$ .

**Bez.:**  $\text{var}(X)$ ,  $\text{Var}(X)$ ,  $\text{var}X$ ,  $\sigma^2$ ,  $\sigma_X^2$ ,  $\sigma$ ,  $\sigma_X$ .

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

# Die Varianz

Stetige und diskrete Zufallsvariablen

Wenn  $X$  diskret, so gilt:

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

# Die Varianz

Stetige und diskrete Zufallsvariablen

Wenn  $X$  diskret, so gilt:

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

Wenn  $X$  stetig, so gilt:

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

wobei  $f$  die Dichte von  $X$  ist.

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

# Eigenschaften der Varianz

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

# Eigenschaften der Varianz

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

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

# Eigenschaften der Varianz

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

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

$$\text{var}(X) = 0 \iff \exists c : P(X = c) = 1.$$

# Unabhängigkeit

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  $x, y \in \mathbb{R}$ .

# Unabhängigkeit

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  $x, y \in \mathbb{R}$ .

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  $x, y \in \mathbb{R}$ .

# Erwartungswert und Varianz

## Eigenschaften

Seien  $X$  und  $Y$  stochastisch unabhängig. Dann

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

**Beweis:** Übung

# Erwartungswert und Varianz

## Eigenschaften

Seien  $X$  und  $Y$  stochastisch unabhängig. Dann

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

**Beweis:** Übung

Seien  $X$  und  $Y$  unabhängig. Dann gilt

$$\mathit{var}(X + Y) = \mathit{var}(X) + \mathit{var}(Y).$$

**Beweis:** Übung

# Die Varianz

Poisson-Verteilung

Wahrscheinlichkeitsfunktion

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

# Die Varianz

Poisson-Verteilung

## Wahrscheinlichkeitsfunktion

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

$$\mathit{var}(X) = \mathbf{E}(X - \mathbf{E}X)^2 = \sum_{i=0}^{\infty} (i - \lambda)^2 p_i$$

# Die Varianz

## Poisson-Verteilung

### Wahrscheinlichkeitsfunktion

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

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

# Die Varianz

## Poisson-Verteilung

### Wahrscheinlichkeitsfunktion

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

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

# Die Varianz

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

## Wahrscheinlichkeitsfunktion

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

$$\text{var}(X) = np(1 - p).$$

(ohne Beweis, ÜA)

# Die Varianz bei Gleichverteilung auf $(a, b)$

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

# Die Varianz bei Gleichverteilung auf $(a, b)$

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

$$\mathbf{E}X^2 = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3} x^3 \Big|_a^b \cdot \frac{1}{b-a}$$

# Die Varianz bei Gleichverteilung auf $(a, b)$

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

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

# Die Varianz bei Gleichverteilung auf $(a, b)$

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

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

$$\text{var}(X) = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \frac{1}{12}(4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2)$$

# Die Varianz bei Gleichverteilung auf $(a, b)$

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

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

$$\begin{aligned} \text{var}(X) &= \mathbf{E}X^2 - (\mathbf{E}X)^2 = \frac{1}{12}(4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2) \\ &= \frac{1}{12}(a^2 - 2ab + b^2) = \frac{(b-a)^2}{12}. \end{aligned}$$

# Die Varianz

## Exponentialverteilung

### Dichte

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

$$\mathbf{EX} = \lambda.$$

$$\mathbf{EX}^2 = \int_0^{\infty} x^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx = 2 \cdot \lambda^2 \quad (\text{ÜA}).$$

$$\mathit{var}(X) = \lambda^2.$$

# Die Varianz

Normalverteilung:  $\text{var}(X) = \sigma^2$

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

# Die Varianz

Normalverteilung:  $\text{var}(X) = \sigma^2$

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

$$\mathbf{E}(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

# Die Varianz

Normalverteilung:  $\text{var}(X) = \sigma^2$

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

# Die Varianz

Normalverteilung:  $\text{var}(X) = \sigma^2$

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

# Die Varianz

Normalverteilung:  $\text{var}(X) = \sigma^2$

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

Bei Normalverteilung sind also die Parameter  $\mu$  und  $\sigma^2$  Erwartungswert und Varianz.

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

## (Theoretische) Schiefe

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

## (Theoretische) Schiefe

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

$\beta_1 = 0$  falls  $F$  symmetrisch

$\beta_1 < 0$  falls  $F$  linksschief

$\beta_1 > 0$  falls  $F$  rechtsschief

## (Theoretische) Schiefe

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

$\beta_1 = 0$  falls  $F$  symmetrisch

$\beta_1 < 0$  falls  $F$  linksschief

$\beta_1 > 0$  falls  $F$  rechtsschief

ÜA: Berechnen Sie die (theoretische) Schiefe von

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

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

# Formmaße (2)

(Theoretische) Wölbung, Kurtosis

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

## Formmaße (2)

(Theoretische) Wölbung, Kurtosis

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

$\beta_2 = 0$  bei Normalverteilung

$\beta_2 > 0$  Tails “dicker, länger, stärker” als bei NV (?)

$\beta_2 < 0$  Tails “dünner, kürzer, schwächer” als  
bei NV (?)

## Formmaße (2)

(Theoretische) Wölbung, Kurtosis

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

$\beta_2 = 0$  bei Normalverteilung

$\beta_2 > 0$  Tails “dicker, länger, stärker” als bei NV (?)

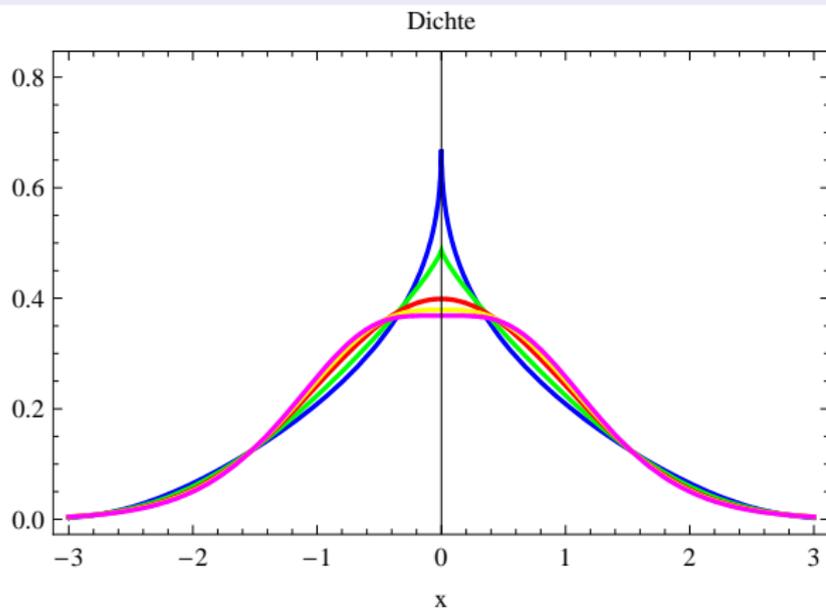
$\beta_2 < 0$  Tails “dünner, kürzer, schwächer” als  
bei NV (?)

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

# Formmaße (3)

## Kurtosis

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



# Formmaße (4)

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( $\lambda$ )	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
Geo(p)	$\frac{2-p}{\sqrt{1-p}}$	$6 + \frac{p^2}{1-p}$

# Inhalt

- Grundgesamtheit, Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz
- Formmaße
- Besondere Eigenschaften der Normalverteilung

## (schwaches) Gesetz der Großen Zahlen

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

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

## Zentraler Grenzwertsatz

Seien  $X_i$  unabhängig, identisch verteilt,

$$EX_i = \mu, \text{var}X_i = \sigma^2.$$

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

**Bem.:** Die  $X_i$  selbst müssen nicht normalverteilt sein.

[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 (zur Information)

gegeben: Erwartungswert  $\mu$  und Varianz  $\sigma^2$

gesucht: Wahrscheinlichkeitsdichte  $f$  auf  $(-\infty, \infty)$  mit

$$\int xf(x) dx = \mu, \quad \int (x - \mu)^2 f(x) dx = \sigma^2$$

und maximaler Entropie:

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

$\implies f = \text{Normaldichte.}$

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

# Normalverteilung

Die Summe normalverteilter Zufallsvariablen

Die Summe normalverteilter Zufallsvariablen ist normalverteilt.

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

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

( $\rho$ : Korrelationskoeffizient zwischen  $X_1$  und  $X_2$ , s.u.)

Beweis: über charakteristische Funktionen (Fouriertransformationen der Dichte) oder

über die Faltungsformel (Stochastik-Vorlesung) oder

über eine Verallgemeinerung des Satzes der Totalen

Wahrscheinlichkeit.

## 4. Statistische Maßzahlen für quantitative Merkmale

### 4.1 Lagemaße

Mittelwert, Quantile, Median, Quartile, Modalwert

### 4.2 Eigenschaften von Schätzungen

### 4.3 Schätzmethoden

### 4.4 Streuungsmaße

Varianz, Standardabweichung, Spannweite, Quartilsabstand, MAD, Variationskoeffizient

### 4.5 Formmaße

Schiefe, Exzess, Wölbung, Kurtosis

# Inhalt

- Lagemaße
- Eigenschaften von Schätzungen
- Schätzmethoden
- Streuungsmaße
- Formmaße

# Lagemaße (Lokationsparameter)

Das arithmetische Mittel

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

## Mittelwert (MEAN)

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

$\bar{X}_n \xrightarrow{n \rightarrow \infty} \mathbf{EX}$       Gesetz der Großen Zahlen.

Voraussetzungen:

- a)  $X_i$  i.i.d.,       $\mathbf{EX}_i < \infty$  (Chintchin)      oder
- b)  $X_i$  beliebig,       $\mathbf{EX}_i^2 < \infty$  (Tschebychev)

## Lagemaße (2)

### Quantile

Die Beobachtungen  $x_1, \dots, x_n$  werden der Größe nach geordnet:

$$x_{(1)} \leq \dots \leq x_{(n)}.$$

Sei  $0 \leq \alpha \leq 1$ ,  $\alpha \cdot n = \lfloor \alpha \cdot n \rfloor + r =: j + r$ ,  $r \in [0, 1)$

### (empirische) 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)  $\alpha$ -Quantil bzw.  $\alpha \cdot 100\%$  Perzentil

mindestens  $\lfloor \alpha \cdot n \rfloor$  der Werte  $(x_1, \dots, x_n)$  sind  $\leq x_\alpha$

mindestens  $\lfloor (1 - \alpha)n \rfloor$  sind  $\geq x_\alpha$

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

# Quantile

Beispiel

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

$\alpha = 0.25$  :

# Quantile

## Beispiel

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

$$\alpha = 0.25 :$$

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

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

$$\alpha = 0.75 :$$

# Quantile

## Beispiel

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

$$\alpha = 0.25 :$$

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

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

$$\alpha = 0.75 :$$

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

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

$$\alpha = 0.5 :$$

# Quantile

## Beispiel

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

$$\alpha = 0.25 :$$

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

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

$$\alpha = 0.75 :$$

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

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

$$\alpha = 0.5 :$$

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

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

# Lagemaße (3)

## Median

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

# Lagemaße (3)

## Median

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

## Quartile

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

# 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).

## 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'.

## 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, ( $\alpha$ -)getrimmtes Mittel

$$\bar{X}_\alpha := \frac{x_{(\lfloor n \cdot \alpha \rfloor + 1)} + \dots + x_{(n - \lfloor n \cdot \alpha \rfloor)}}{n - 2 \lfloor n \cdot \alpha \rfloor}, \quad \alpha \in [0, \frac{1}{2})$$

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

## 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, ( $\alpha$ -)getrimmtes Mittel

$$\bar{X}_\alpha := \frac{x_{(\lfloor n \cdot \alpha \rfloor + 1)} + \dots + x_{(n - \lfloor n \cdot \alpha \rfloor)}}{n - 2 \lfloor n \cdot \alpha \rfloor}, \quad \alpha \in [0, \frac{1}{2})$$

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

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

# Lagemaße (5)

winsorisiertes Mittel, ( $\alpha$ -)winsorisiertes Mittel

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

$$\bar{X}_{\alpha, w} := \frac{n_1 x_{(n_1)} + x_{(n_1+1)} + \dots + x_{(n-n_1)} + n_1 x_{(n-n_1+1)}}{n}$$

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

# Lagemaße (5)

winsorisiertes Mittel, ( $\alpha$ -)winsorisiertes Mittel

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

$$\bar{X}_{\alpha,w} := \frac{n_1 x_{(n_1)} + x_{(n_1+1)} + \dots + x_{(n-n_1)} + n_1 x_{(n-n_1+1)}}{n}$$

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  $\bar{X}$  und effizienter als  $x_{0.5}$ .

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

# Lageschätzungen mit SAS

Mittelwert:	PROC MEANS; PROC SUMMARY;
Median:	PROC MEANS MEDIAN; PROC UNIVARIATE;
getrimmte Mittel:	PROC UNIVARIATE TRIMMED=Zahl;
winsorisierte Mittel:	PROC UNIVARIATE WINSORIZED=Zahl;
Modalwert:	PROC UNIVARIATE;
Quartile:	PROC UNIVARIATE;
Quantile:	PROC UNIVARIATE; PROC MEANS p1 p5 p10 p25 p75 p99; (etwa)

## Beispiel: 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

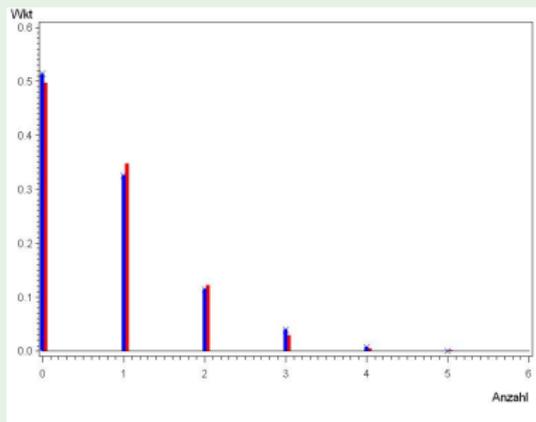
5            0

## Beispiel: 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

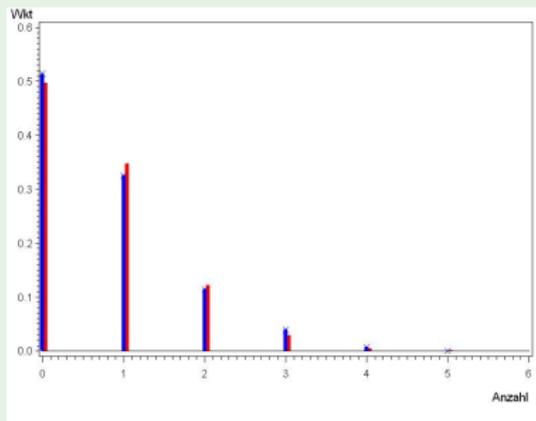


## Beispiel: Tödliche Unfälle durch Pferdetritte

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



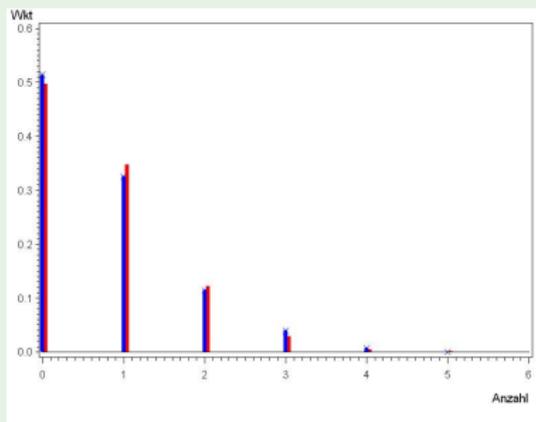
Poisson-Verteilung geeignet (?)

## Beispiel: 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 (?) Schätzung von  $\lambda$  durch  $\bar{X}$ .

## Beispiel: Anzahl von schwarzen Feldern

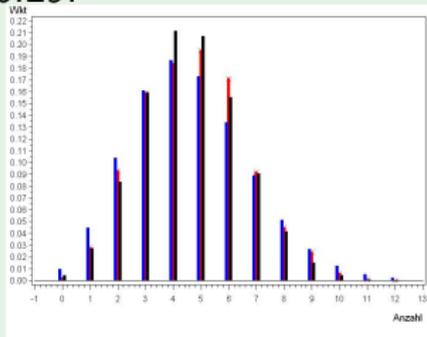
Ein Zufallszahlengenerator soll zufällige Bildpunkte erzeugen, weiß mit Wkt. 0.71 und schwarz mit Wkt. 0.29. Dazu wurde ein großes Rechteck in 1000 Teilquadrate mit je 16 Bildpunkten zerlegt. Gezählt wurde jeweils die Anzahl der schwarzen Bildpunkte.

## Beispiel: 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 Rechteck in 1000 Teilquadrate mit je 16 Bildpunkten zerlegt. Gezählt wurde jeweils die Anzahl der schwarzen Bildpunkte.



## Beispiel: Anzahl von schwarzen Feldern

Ein Zufallszahlengenerator soll zufällige Bildpunkte erzeugen, weiß mit

Wkt. 0.71 und schwarz mit Wkt. 0.29.

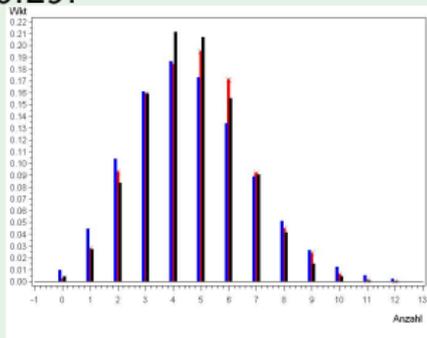
Dazu wurde ein großes Recht-

eck in 1000 Teilquadrate mit

je 16 Bildpunkten zerlegt. Ge-

zählt wurde jeweils die Anzahl

der schwarzen Bildpunkte.



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

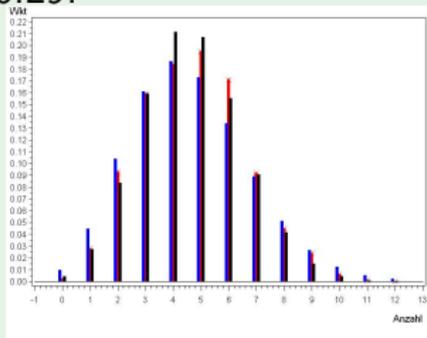
Binomial-Verteilung (schwarz) geeignet (?)

## Beispiel: 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 Rechteck 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
<b>h</b>	2	28	93	159	184	195	171	92	45	24	6	1	0

Binomial-Verteilung (schwarz) geeignet (?)

Ang.  $p$  unbekannt. Schätzung von  $np = 16p$  durch

## Beispiel: Anzahl von schwarzen Feldern

Ein Zufallszahlengenerator soll zufällige Bildpunkte erzeugen, weiß mit

Wkt. 0.71 und schwarz mit Wkt. 0.29.

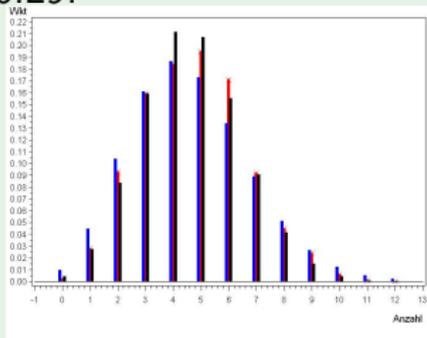
Dazu wurde ein großes Recht-

eck in 1000 Teilquadrate mit

je 16 Bildpunkten zerlegt. Ge-

zählt wurde jeweils die Anzahl

der schwarzen Bildpunkte.



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

Binomial-Verteilung (schwarz) geeignet (?)

Ang.  $p$  unbekannt. Schätzung von  $np = 16p$  durch  $\bar{X}$ .

# Inhalt

- Lagemaße
- Eigenschaften von Schätzungen
- Schätzmethoden
- Streuungsmaße
- Formmaße

# Eigenschaften von Schätzungen (1)

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

Konsistenz (Minimalforderung)

$$\hat{\theta}_n \xrightarrow{n \rightarrow \infty} \theta$$

# Eigenschaften von Schätzungen (1)

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

## Konsistenz (Minimalforderung)

$$\hat{\theta}_n \xrightarrow{n \rightarrow \infty} \theta$$

## Erwartungstreue, Asymptotische Erwartungstreue

$$\mathbf{E}\hat{\theta}_n = \theta$$

$$\mathbf{E}\hat{\theta}_n \rightarrow_{n \rightarrow \infty} \theta$$

# Eigenschaften von Schätzungen (1)

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

Konsistenz (Minimalforderung)

$$\hat{\theta}_n \xrightarrow{n \rightarrow \infty} \theta$$

Erwartungstreue, Asymptotische Erwartungstreue

$$\mathbf{E}\hat{\theta}_n = \theta$$

$$\mathbf{E}\hat{\theta}_n \rightarrow_{n \rightarrow \infty} \theta$$

“gute”, “effiziente” Schätzung

$\text{var } \hat{\theta}_n$  möglichst klein

## Eigenschaften von Schätzungen (2)

### optimale Schätzung

wenn  $\text{var } \hat{\theta}_n$  den kleinstmöglichen Wert annimmt für alle erwartungstreuen (e-treuen) Schätzungen.

## Eigenschaften von Schätzungen (2)

### optimale Schätzung

wenn  $\text{var } \hat{\theta}_n$  den kleinstmöglichen Wert annimmt für alle erwartungstreuen (e-treuen) Schätzungen.

### Mean Square Error (MSE)

$$\begin{aligned}\text{MSE} &= \mathbf{E}(\hat{\theta}_n - \theta)^2 = \mathbf{E}(\hat{\theta}_n - \mathbf{E}\hat{\theta}_n + \mathbf{E}\hat{\theta}_n - \theta)^2 \\ &= \text{var } \hat{\theta}_n + (\mathbf{E}\hat{\theta}_n - \theta)^2 = \text{var } \hat{\theta}_n + \text{bias}^2 \hat{\theta}_n\end{aligned}$$

soll minimal oder möglichst klein sein.

## Eigenschaften von Schätzungen (2)

### optimale Schätzung

wenn  $\text{var } \hat{\theta}_n$  den kleinstmöglichen Wert annimmt für alle erwartungstreuen (e-treuen) Schätzungen.

### Mean Square Error (MSE)

$$\begin{aligned}\text{MSE} &= \mathbf{E}(\hat{\theta}_n - \theta)^2 = \mathbf{E}(\hat{\theta}_n - \mathbf{E}\hat{\theta}_n + \mathbf{E}\hat{\theta}_n - \theta)^2 \\ &= \text{var } \hat{\theta}_n + (\mathbf{E}\hat{\theta}_n - \theta)^2 = \text{var } \hat{\theta}_n + \text{bias}^2 \hat{\theta}_n\end{aligned}$$

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

$\theta$ : zu schätzender Parameter einer Population (Dichte  $f$ ).

$\hat{\theta} = \hat{\theta}_n$ : eine erwartungstreue Schätzung von  $\theta$ .

### Cramer-Rao-Ungleichung

$$\text{var}(\hat{\theta}) \geq \frac{1}{n \cdot I(f, \theta)},$$

# Eigenschaften von Schätzungen (3)

## Cramer-Rao Ungleichung

$\theta$ : zu schätzender Parameter einer Population (Dichte  $f$ ).

$\hat{\theta} = \hat{\theta}_n$ : eine erwartungstreue Schätzung von  $\theta$ .

### Cramer-Rao-Ungleichung

$$\text{var}(\hat{\theta}) \geq \frac{1}{n \cdot I(f, \theta)},$$

### Fisher-Information

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

# Eigenschaften von Schätzungen (3)

## Cramer-Rao Ungleichung

$\theta$ : zu schätzender Parameter einer Population (Dichte  $f$ ).

$\hat{\theta} = \hat{\theta}_n$ : eine erwartungstreue Schätzung von  $\theta$ .

### Cramer-Rao-Ungleichung

$$\text{var}(\hat{\theta}) \geq \frac{1}{n \cdot I(f, \theta)},$$

### Fisher-Information

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

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

# Eigenschaften von Schätzungen (4)

Beispiele

$f$  normal

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

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

$$\frac{\partial \ln f(x, \mu)}{\partial \mu} = \frac{x-\mu}{\sigma} \cdot \frac{1}{\sigma}$$

$$I(f, \mu) = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \cdot f(x, \mu) dx = \frac{1}{\sigma^2}.$$

# Eigenschaften von Schätzungen (5)

Beispiele (2)

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

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

# Eigenschaften von Schätzungen (5)

Beispiele (2)

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

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

insbesondere

$$\text{var}(\bar{X}) \geq \frac{\sigma^2}{n}.$$

# Eigenschaften von Schätzungen (5)

Beispiele (2)

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

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

insbesondere

$$\text{var}(\bar{X}) \geq \frac{\sigma^2}{n}.$$

Vergleichen Sie das mit:

$$\text{var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{var}X_i = \frac{\sigma^2}{n}.$$

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

# Eigenschaften von Schätzungen (6)

Beispiele (3)

$f$  exponential

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

# Eigenschaften von Schätzungen (6)

Beispiele (3)

$f$  exponential

$$f(x, \lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} & \text{falls } x \geq 0 \\ 0 & \text{sonst.} \end{cases} \quad I(f, \lambda) = \frac{1}{\lambda^2} \quad (\text{ÜA})$$

# Eigenschaften von Schätzungen (6)

Beispiele (3)

$f$  exponential

$$f(x, \lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} & \text{falls } x \geq 0 \\ 0 & \text{sonst.} \end{cases} \quad I(f, \lambda) = \frac{1}{\lambda^2} \quad (\text{ÜA})$$

Die Cramer-Rao-Schranke ist also:  $\frac{1}{nI(\lambda)} = \frac{\lambda^2}{n}$ .

Vergleichen Sie mit:  $\text{var}(\bar{X}) = \frac{\lambda^2}{n}$ .

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

# Eigenschaften von Schätzungen (7)

Beispiele (4)

$f$  Doppelsexponential (=Laplace)

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

Der hier interessierende (Lage-) Parameter ist  $\mu$ .

$$I(f, \mu) = \frac{1}{\lambda^2}. \quad (\ddot{U}A) \quad \text{var}(\bar{X}) = \frac{2\lambda^2}{n}. \quad (\ddot{U}A)$$

# Eigenschaften von Schätzungen (7)

Beispiele (4)

$f$  Doppelsexponential (=Laplace)

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

Der hier interessierende (Lage-) Parameter ist  $\mu$ .

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

Für den Median  $x_{0.5}$  gilt:

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

# Inhalt

- Lagemaße
- Eigenschaften von Schätzungen
- Schätzmethoden
- Streuungsmaße
- Formmaße

## Momentenmethode

Man drückt den zu schätzenden Parameter durch die Momente, z.B.  $\mathbf{E}(X)$ , aus.

Dann werden die Momente durch die entsprechenden *empirischen* Momente, z.B. der Erwartungswert durch  $\bar{X}$ , ersetzt.

## Maximum-Likelihood-Schätzung (ML-Schätzung)

Es wird der Schätzwert für den unbekannt Parameter ermittelt, der anhand der vorliegenden Daten, am meisten für diesen Parameter spricht (most likely).

## Kleinste-Quadrat-Schätzung (KQS)

Sei  $\theta$  der zu schätzende Parameter. Man geht aus von einem Modell, z.B.

$$Y_i = g(\theta, X_i) + \epsilon_i$$

Dann versucht man die Summe der Fehlerquadrate

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - g(\theta, X_i))^2.$$

zu minimieren (Kleinste Quadrate).

# Momentenschätzung

## Momentenschätzung bei Normalverteilung

Seien  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ .

$$\mu = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{\mu} = \bar{X}$$

$$\sigma^2 = \mathbf{E}(X - \mathbf{E}X)^2 \quad \Rightarrow \quad \hat{\sigma}^2 = \overline{(X_i - \bar{X})^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

# Momentenschätzung

## Momentenschätzung bei Normalverteilung

Seien  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ .

$$\mu = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{\mu} = \bar{X}$$

$$\sigma^2 = \mathbf{E}(X - \mathbf{E}X)^2 \quad \Rightarrow \quad \hat{\sigma}^2 = \overline{(X_i - \bar{X})^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

## Momentenschätzung bei Exponentialverteilung

Seien  $X_1, \dots, X_n \sim \text{Exp}(\lambda)$ .

$$\lambda = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{\lambda} = \bar{X}$$

# Momentenschätzung

## Momentenschätzung bei Binomialverteilung

Seien  $X_1, \dots, X_n \sim Bi(p, 1)$ .

$$p = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{p} = \bar{X}$$

der relative Anteil der Realisierungen  $x_i = 1$ .

# Maximum-Likelihood-Schätzung

## ML-Schätzung bei Binomialverteilung

Beobachten  $n=1000$  Jugendliche. Stichprobe  $(X_1, \dots, X_n)$

$X_i = 1$  falls Übergewicht festgestellt

$X_i = 0$  sonst.

Die Wahrscheinlichkeit, dass die beobachtete Stichprobe auftritt, wenn der Parameter  $p$  vorliegt ist (die Beobachtungen werden als unabhängig angenommen)

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i}$$

# Maximum-Likelihood-Schätzung

## ML-Schätzung bei Binomialverteilung

Beobachten  $n=1000$  Jugendliche. Stichprobe  $(X_1, \dots, X_n)$

$X_i = 1$  falls Übergewicht festgestellt

$X_i = 0$  sonst.

Die Wahrscheinlichkeit, dass die beobachtete Stichprobe auftritt, wenn der Parameter  $p$  vorliegt ist (die Beobachtungen werden als unabhängig angenommen)

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n) &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\ &= p^k (1-p)^{n-k}, \quad \text{wobei} \quad k = \sum_{i=1}^n x_i. \end{aligned}$$

# Maximum-Likelihood-Schätzung

Binomialverteilung

Der ML-Schätzer ist der Wert, der diese Funktion,  $L_n(p)$ , Likelihood-Funktion genannt, bzgl.  $p$  maximiert.

Maximieren statt  $L_n(p)$ :  $\ln L_n(p)$  (Arg.Max. ist dasselbe).

# Maximum-Likelihood-Schätzung

## Binomialverteilung

Der ML-Schätzer ist der Wert, der diese Funktion,  $L_n(p)$ , Likelihood-Funktion genannt, bzgl.  $p$  maximiert.

Maximieren statt  $L_n(p)$ :  $\ln L_n(p)$  (Arg.Max. ist dasselbe).

$$\begin{aligned}\ln L_n(p) &= \ln(p^k(1-p)^{n-k}) \\ &= k \ln p + (n-k) \ln(1-p).\end{aligned}$$

# Maximum-Likelihood-Schätzung

Binomialverteilung

Der ML-Schätzer ist der Wert, der diese Funktion,  $L_n(p)$ , Likelihood-Funktion genannt, bzgl.  $p$  maximiert.

Maximieren statt  $L_n(p)$ :  $\ln L_n(p)$  (Arg.Max. ist dasselbe).

$$\begin{aligned}\ln L_n(p) &= \ln(p^k(1-p)^{n-k}) \\ &= k \ln p + (n-k) \ln(1-p).\end{aligned}$$

Ableiten nach  $p$  und Nullsetzen liefert:

$$\frac{k}{p} - \frac{n-k}{1-p} = 0$$

# Maximum-Likelihood-Schätzung

Binomialverteilung

Die einzige Lösung ist:

$$\hat{p} = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Für ein relatives Extremum in  $(0,1)$  kommt nur dieser Wert in Betracht.

# Maximum-Likelihood-Schätzung

Binomialverteilung

Die einzige Lösung ist:

$$\hat{p} = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Für ein relatives Extremum in  $(0,1)$  kommt nur dieser Wert in Betracht.

Müssen aber noch die Likelihood-Funktion an den Rändern betrachten:

Für  $p = 0$  und  $p = 1$  wird  $\ln L(p) = -\infty$ . Also:

$$\hat{p}_{ML} = \frac{k}{n}.$$

# Maximum-Likelihood-Schätzung

Normalverteilung,  $\mu$  unbekannt,  $\sigma^2$  bekannt

## ML-Schätzung bei Normalverteilung

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

Seien  $X_1, \dots, X_n$  unabhängig,  $X_i \sim \mathcal{N}(\mu, 1)$ .

Likelihood:

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

# Maximum-Likelihood-Schätzung

Normalverteilung, 2

$$\ln L_n(\mu) = -n \ln(\sqrt{2\pi}) + \sum_{i=1}^n \left( -\frac{(x_i - \mu)^2}{2} \right)$$
$$\frac{\partial \ln L_n(\mu)}{\partial \mu} = \sum_{i=1}^n (x_i - \mu)$$

Nullsetzen liefert die Maximum-Likelihood-Schätzung

$$\hat{\mu} = \bar{X}.$$

# Maximum-Likelihood-Schätzung

Normalverteilung,  $\mu$  und  $\sigma^2$  unbekannt

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ ,  $x_1, \dots, x_n$ : Beobachtungen

$$L_n(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$

# Maximum-Likelihood-Schätzung

Normalverteilung,  $\mu$  und  $\sigma^2$  unbekannt

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ ,  $x_1, \dots, x_n$ : Beobachtungen

$$\begin{aligned} L_n(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

# Maximum-Likelihood-Schätzung

Normalverteilung,  $\mu$  und  $\sigma^2$  unbekannt

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ ,  $x_1, \dots, x_n$ : Beobachtungen

$$\begin{aligned} L_n(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

wobei  $S^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

Die letzte Gleichung folgt aus:

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$$

# Maximum-Likelihood-Schätzung

Normalverteilung, Fortsetzung

Log-Likelihood:

$$\ln L(\mu, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

# Maximum-Likelihood-Schätzung

Normalverteilung, Fortsetzung

Log-Likelihood:

$$\ln L(\mu, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

Lösen des Gleichungssystems

$$0 = \frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \frac{\bar{X} - \mu}{\sigma^2}$$

$$0 = \frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{nS^2}{\sigma^3} + \frac{n(\bar{X} - \mu)^2}{\sigma^3}$$

# Maximum-Likelihood-Schätzung

Normalverteilung, Fortsetzung

Log-Likelihood:

$$\ln L(\mu, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

Lösen des Gleichungssystems

$$0 = \frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \frac{\bar{X} - \mu}{\sigma^2}$$

$$0 = \frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{nS^2}{\sigma^3} + \frac{n(\bar{X} - \mu)^2}{\sigma^3}$$

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = S^2$$

# Maximum-Likelihood-Schätzung

Gleichverteilung

## ML-Schätzung bei Gleichverteilung auf $(0, \theta)$

Likelihood:  $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$ ,

die gemeinsame Dichtefunktion der  $X_i$ .

Seien  $X_1, \dots, X_n$  unabhängig,  $X_i \sim R(0, \theta)$ , d.h.

$$f_{X_i}(x_i) = \begin{cases} \frac{1}{\theta} & \text{falls } 0 \leq x_i \leq \theta \\ 0 & \text{sonst} \end{cases}$$

# Maximum-Likelihood-Schätzung

Gleichverteilung, 2

Likelihood:

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n f_{X_i}(x_i) && \text{(Unabhängigkeit)} \\ &= \begin{cases} \frac{1}{\theta^n} & \text{falls } 0 \leq x_i \leq \theta \quad \forall x_i \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Maximal, wenn  $\theta \geq x_1, \dots, x_n$ , und wenn  $\theta$  möglichst klein, also

# Maximum-Likelihood-Schätzung

Gleichverteilung, 2

Likelihood:

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n f_{X_i}(x_i) && \text{(Unabhängigkeit)} \\ &= \begin{cases} \frac{1}{\theta^n} & \text{falls } 0 \leq x_i \leq \theta \quad \forall x_i \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Maximal, wenn  $\theta \geq x_1, \dots, x_n$ , und wenn  $\theta$  möglichst klein, also

$$\hat{\theta} = \max(x_1, \dots, x_n).$$

# Maximum-Likelihood-Schätzung

## Gemischte Normalverteilung

Dichte ( $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$ ):

$$f(x; \theta) = (1 - p)\phi\left(\frac{x - \mu_1}{\sigma_1}\right) + p\phi\left(\frac{x - \mu_2}{\sigma_2}\right)$$

$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$  mit Wkt.  $(1 - p)$  und  $X_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$  mit Wkt.  $(1 - p)$ , aber  $p$  ist nicht bekannt.

# Maximum-Likelihood-Schätzung

## Gemischte Normalverteilung

Dichte ( $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$ ):

$$f(x; \theta) = (1 - p)\phi\left(\frac{x - \mu_1}{\sigma_1}\right) + p\phi\left(\frac{x - \mu_2}{\sigma_2}\right)$$

$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$  mit Wkt.  $(1 - p)$  und  $X_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$  mit Wkt.  $p$ , aber  $p$  ist nicht bekannt.

Likelihood:

$$L(\theta) = \prod_{i=1}^n \left( (1 - p)\phi\left(\frac{x_i - \mu_1}{\sigma_1}\right) + p\phi\left(\frac{x_i - \mu_2}{\sigma_2}\right) \right)$$

Maximieren des (log-)Likelihood  $\rightarrow$  Newton-Raphson o.

EM-Algorithmus (Stochastik-Vorlesung)

# Eigenschaften von ML-Schätzern

Unter Regularitätsannahmen gilt

- ML-Schätzungen sind konsistent.

# Eigenschaften von ML-Schätzern

## Unter Regularitätsannahmen gilt

- ML-Schätzungen sind konsistent.
- Wenn sie erwartungstreu ist: sie sind (asymptotisch) effizient, d.h. sie haben minimale Varianz.

Die Varianz ist durch die Cramér-Rao Ungleichung gegeben.

# Eigenschaften von ML-Schätzern

## Unter Regularitätsannahmen gilt

- ML-Schätzungen sind konsistent.
- Wenn sie erwartungstreu ist: sie sind (asymptotisch) effizient, d.h. sie haben minimale Varianz.  
Die Varianz ist durch die Cramér-Rao Ungleichung gegeben.
- sie sind asymptotisch normal verteilt (wichtig für die Konstruktion von Konfidenzintervallen, s.u.)

# Eigenschaften von ML-Schätzern

## Unter Regularitätsannahmen gilt

- ML-Schätzungen sind konsistent.
- Wenn sie erwartungstreu ist: sie sind (asymptotisch) effizient, d.h. sie haben minimale Varianz.  
Die Varianz ist durch die Cramér-Rao Ungleichung gegeben.
- sie sind asymptotisch normal verteilt (wichtig für die Konstruktion von Konfidenzintervallen, s.u.)
- Nachteil: ML-Schätzungen beruhen auf Verteilungsannahmen.

# Kleinste Quadrat Schätzung

## KQS des Lageparameters

Modell:

$$Y_i = \mu + \epsilon_i$$

Die Summe der Fehlerquadrate

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \mu)^2.$$

minimieren:

# Kleinste Quadrat Schätzung

## KQS des Lageparameters

Modell:

$$Y_i = \mu + \epsilon_i$$

Die Summe der Fehlerquadrate

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \mu)^2.$$

minimieren: Differenzieren und Nullsetzen liefert:

$$\hat{\mu}_{KQS} = \bar{Y}.$$

# Kleinste Quadrat-Schätzung

## KQS im einfachen linearen Regressionsmodell

$$Y_i = \theta_2 + \theta_1 X_i + \epsilon_i$$

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

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

# Kleinste Quadrat-Schätzung

## KQS im einfachen linearen Regressionsmodell

$$Y_i = \theta_2 + \theta_1 X_i + \epsilon_i$$

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

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

Minimieren von  $\sum (Y_i - f(X_i, \theta_1, \theta_2))^2$  liefert:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - (\theta_1 X_i + \theta_2)) \cdot X_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n (Y_i - (\theta_1 X_i + \theta_2)) \cdot 1 = 0$$

# Kleinste Quadrat-Schätzung

⇒

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_2 \sum_i X_i = 0$$

$$\sum_i Y_i - \theta_1 \sum_i X_i - \theta_2 \cdot n = 0$$

Die zweite Gleichung nach  $\theta_2$  auflösen:

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

und in die erste einsetzen:

# Kleinste Quadrat-Schätzung

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

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

⇒

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

$$\hat{\theta}_2 = \frac{1}{n} \left( \sum_i Y_i - \hat{\theta}_1 \sum_i X_i \right)$$

# Einschub: Die Prozedur GPLOT

(vgl. ÜA 9)

Darstellung von Dichten und Wahrscheinlichkeitsfunktionen

```
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.

## Die 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.

# Inhalt

- Lagemaße
- Eigenschaften von Schätzungen
- Schätzmethoden
- Streuungsmaße
- Formmaße

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

(empirische) Varianz (Streuung)

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

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

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

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

(empirische) Varianz (Streuung)

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

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

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

Standardabweichung

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

## Spannweite (Range)

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

## Spannweite (Range)

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

## (Inter-)Quartilsabstand, IR

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

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

## 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

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

## Median absolute deviation, MAD

$$MAD = \text{med}(|x_i - x_{0.5}|)$$

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

## Variationskoeffizient

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

kein Skalenmaß, weder lageinvariant noch skalenequivariant

## Gini's Mean Difference

$$G = \frac{1}{\binom{n}{2}} \sum_{i < j} |x_i - x_j| \quad X \sim \mathcal{N} \Rightarrow \mathbf{E}\left(\frac{\sqrt{\pi}}{2} G\right) = \sigma$$

$S_n$  und  $Q_n$  (Croux, Rousseeuw 1992, 1993)

$$S_n = 1.1926 \cdot \text{lomed}_i(\text{highmed}_j | x_i - x_j |)$$

$$Q_n = 2.2219 \cdot \{ |x_i - x_j|, i < j \}_{(k)}$$

Bei  $n = 2m + 1$  ist  $\text{lomed} = \text{highmed} = \text{med}$ ,

bei  $n = 2m$  ist  $\text{lomed} = X_{(\frac{n}{2})}$ ,  $\text{highmed} = X_{(\frac{n}{2}+1)}$ .

$$k = \binom{h}{2}, h = \lfloor \frac{n}{2} \rfloor + 1 \quad \left( k \approx \binom{n}{2} / 4 \right)$$

$\{ \dots \}_{(k)}$  bezeichnet das  $k$ -te Element der geordneten (Multi-)Menge.

- SAS verwendet einen modifizierten Schätzer (Korrekturfaktor) für kleine Umfänge.
- Die konstanten Faktoren sichern Konsistenz und 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

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

$$\text{Range} = x_{(n)} - x_{(1)}$$

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

# Streuungsmaße (7)

## Robuste Skalenschätzungen

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

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

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

$$S_n = 1.1926 \cdot med_i(med_j |x_i - x_j|)$$

$$Q_n = 2.219 \cdot \{|x_i - x_j|, i < j\}_{(k)}$$

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

SAS (Option ROBUSTSCALE) gibt neben diesen Werten auch die (im Fall der Normalverteilung) erwartungstreuen Schätzungen an.

# 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  
WINSORIZED=10; VAR ...; RUN;
```

TRIMMED: getrimmte Mittel

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

WINSORIZED: winsorisierte Mittel

ROBUSTSCALE: robuste Skalenschätzer

Descr\_MAD

# Lage- und Streuungsmaße in SAS (4)

Abkürzung

```
PROC CAPABILITY ROBUSTSCALE TRIMMED=10
```

```
WINSORISED=10;
```

```
ODS SELECT BASICMEASURES
```

```
TRIMMEDMEANS ROBUSTSCALE
```

```
VAR ...;
```

```
RUN;
```

# Inhalt

- Lagemaße
- Eigenschaften von Schätzungen
- Schätzmethoden
- Streuungsmaße
- Formmaße

# Formmaße (1)

(Theoretische) Schiefe

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

# Formmaße (1)

(Theoretische) Schiefe

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

(Empirische) Schiefe

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{X}}{s} \right)^3$$

$$\hat{\beta}_{1,SAS} = \hat{\beta}_1 \frac{n^2}{(n-1)(n-2)}$$

**PROC MEANS** skewness;

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

# Formmaße (2)

(Theoretische) Wölbung, Kurtosis

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

# Formmaße (2)

(Theoretische) Wölbung, Kurtosis

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

(Empirische) Wölbung, Kurtosis

$$\hat{\beta}_2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{X}}{s} \right)^4 - 3$$
$$\hat{\beta}_{2,\text{SAS}} = \hat{\beta}_2 \frac{n^2(n+1)}{(n-1)(n-2)(n-3)} - 3 \frac{(n-1)^2}{(n-2)(n-3)}$$

## Formmaße (3)

### Exzeß

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

$\beta_2 = 0$  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)

Erinnerung:

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

# Inhalt

- Box-Plots
- Probability Plots
- Histogramme
- Dichteschätzung

Ziel: übersichtliche Darstellung der Daten.

Boxplot zu dem Eingangsbeispiel mit  $n=5$ :

[Descr\\_Boxplot0.sas](#)

Prozeduren: UNIVARIATE, GPLOT, BOXPLOT

Ziel: übersichtliche Darstellung der Daten.

Boxplot zu dem Eingangsbeispiel mit  $n=5$ :

`Descr_Boxplot0.sas`

Prozeduren: UNIVARIATE, GPLOT, BOXPLOT

**PROC UNIVARIATE PLOT; RUN;**

Ziel: übersichtliche Darstellung der Daten.

Boxplot zu dem Eingangsbeispiel mit  $n=5$ :

Descr\_Boxplot0.sas

Prozeduren: UNIVARIATE, GPLOT, BOXPLOT

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

```
SYMBOL1 INTERPOL=BOXT10;
```

```
PROC GPLOT;
```

```
    PLOT y*x=1;
```

```
RUN;
```

Ziel: übersichtliche Darstellung der Daten.

Boxplot zu dem Eingangsbeispiel mit  $n=5$ :

Descr\_Boxplot0.sas

Prozeduren: UNIVARIATE, GPLOT, BOXPLOT

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

```
SYMBOL1 INTERPOL=BOXT10;
```

```
PROC GPLOT;
```

```
    PLOT y*x=1;
```

```
RUN;
```

```
PROC BOXPLOT;
```

```
    PLOT y*x /BOXSTYLE=SCHEMATIC;
```

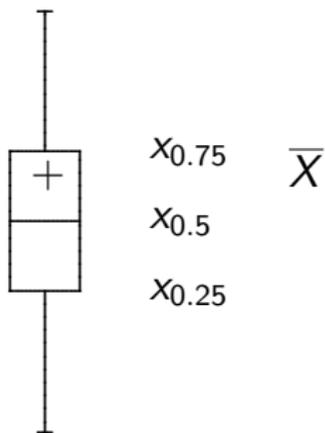
```
        /BOXSTYLE=SKELETAL;
```

```
RUN;
```

## Prozedur BOXPLOT

\* Ausreißer ???

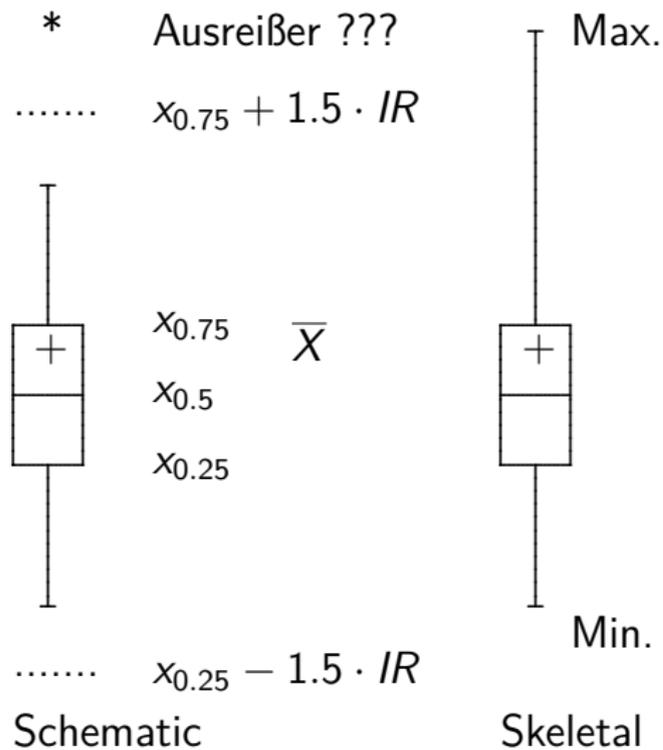
.....  $x_{0.75} + 1.5 \cdot IR$



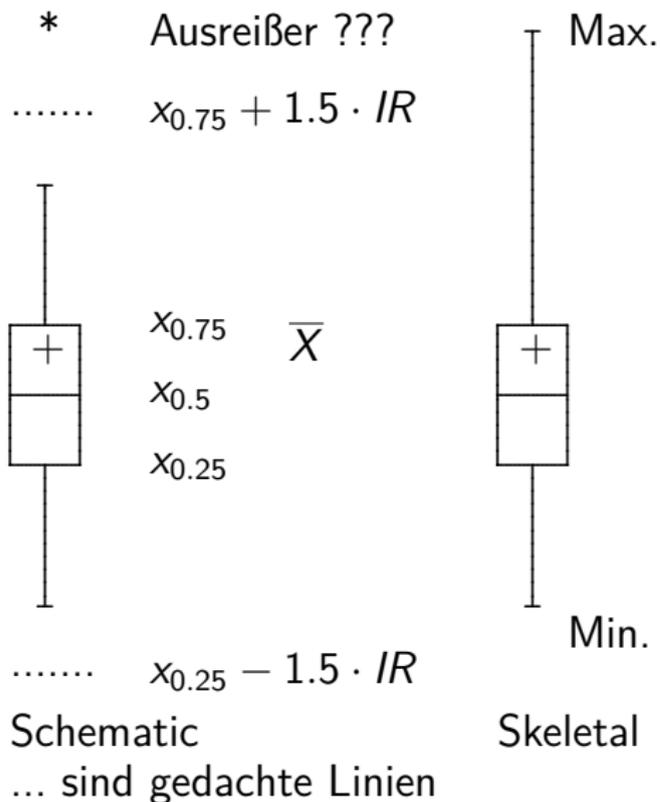
.....  $x_{0.25} - 1.5 \cdot IR$

Schematic

## Prozedur BOXPLOT



## Prozedur BOXPLOT



## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den “fences” (den ...).

$$0.99 = 0.995 - 0.005$$

## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den “fences” (den ...).

$$\begin{aligned} 0.99 &= 0.995 - 0.005 \\ &= \Phi(2.575) - \Phi(-2.575) \end{aligned}$$

## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den “fences” (den ...).

$$\begin{aligned} 0.99 &= 0.995 - 0.005 \\ &= \Phi(2.575) - \Phi(-2.575) \\ &= P(\mu - 2.575\sigma < X < \mu + 2.575\sigma) \end{aligned}$$

## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den "fences"  
(den ...).

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

## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den "fences"  
(den ...).

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

## Erläuterung zu BOXSTYLE=Schematic

$X \sim \mathcal{N}(\mu, \sigma^2)$  etwa 99% der Daten liegen zwischen den "fences"  
(den ...).

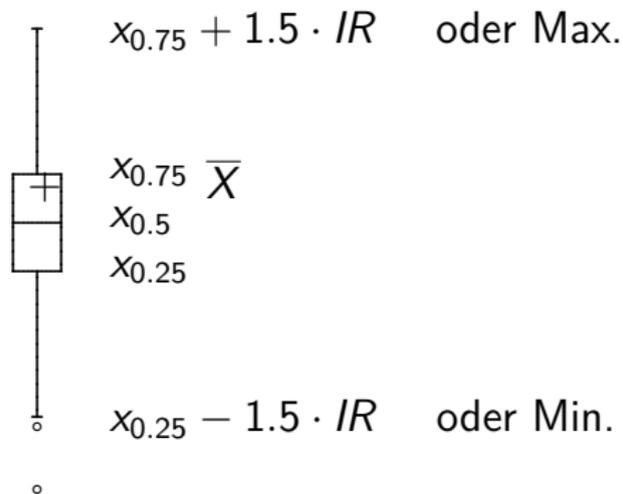
$$\begin{aligned}0.99 &= 0.995 - 0.005 \\&= \Phi(2.575) - \Phi(-2.575) \\&= P(\mu - 2.575\sigma < X < \mu + 2.575\sigma) \\&\approx P(x_{0.5} - 2.575 \cdot \underbrace{0.7434 \cdot IR}_{\text{}} < X < \\&\quad x_{0.5} + 2.575 \cdot \underbrace{0.7434 \cdot IR}_{\text{}}) \\&= P(x_{0.5} - 1.914 \cdot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

# Prozedur UNIVARIATE, Option PLOT

Zum Vergleich: es gibt auch andere Boxplotdefinitionen

\* Ausreißer ??

.....  $x_{0.75} + 3 \cdot IR$



# Box-Plots in SAS

Ein Merkmal, eine Gruppe (Merkmal gr)

```
gr = 1;
```

```
PROC BOXPLOT;
```

```
  PLOT zeit*gr; RUN;
```

# Box-Plots in SAS

Ein Merkmal, eine Gruppe (Merkmal gr)

```
gr = 1;
```

```
PROC BOXPLOT;
```

```
  PLOT zeit*gr; RUN;
```

Ein Merkmal (zeit), mehrere Gruppen (z.B. gr=1,2,3)

```
PROC BOXPLOT;
```

```
  PLOT zeit*gr; RUN;
```

# Box-Plots in SAS

Ein Merkmal, eine Gruppe (Merkmal gr)

```
gr = 1;
```

```
PROC BOXPLOT;
```

```
  PLOT zeit*gr; RUN;
```

Ein Merkmal (zeit), mehrere Gruppen (z.B. gr=1,2,3)

```
PROC BOXPLOT;
```

```
  PLOT zeit*gr; RUN;
```

Ein Merkmal (X), mehrere Gruppen (gr)

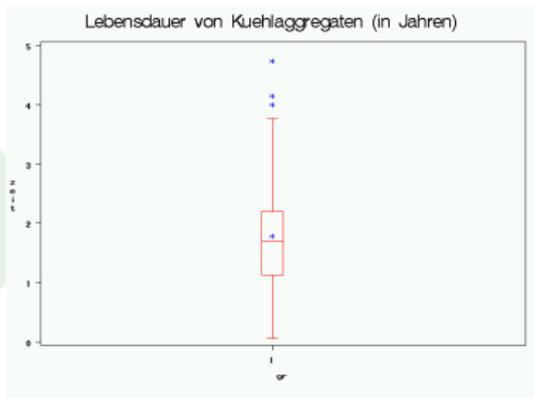
```
SYMBOL INTERPOL=BOXT10;
```

```
PROC GPLOT; PLOT X*gr; RUN;
```

Descr\_Boxplot.sas    Descr\_Boxplot1.sas

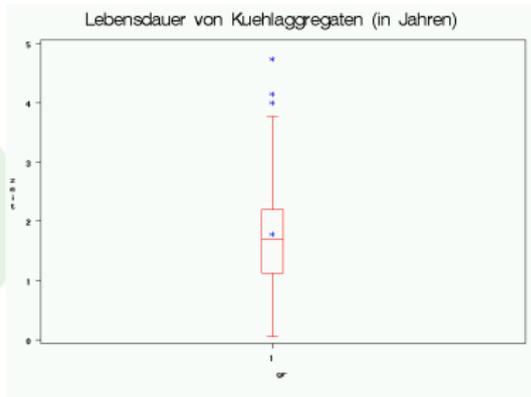
# Boxplots - Beispiele

Lebensdauern von 100  
Kühlaggregate

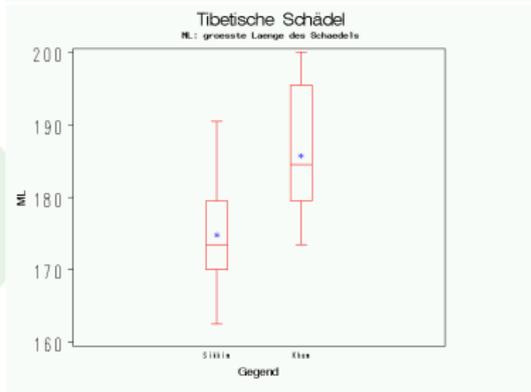


# Boxplots - Beispiele

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;

# Inhalt

- Box-Plots
- Probability Plots
- Histogramme
- Dichteschätzung

## Erinnerung: Normalverteilung

### (i) Dichte der Standard-Normalverteilung

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

### (ii) Verteilungsfunktion der Standard-Normal

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

### (iii) Dichte der Normalverteilung

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

mit Erwartungswert  $\mu$  und Varianz  $\sigma^2$ .

# 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.

$\alpha = 0.05$

$$u_{1-\alpha} = \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.95) = 1.645$$

$$u_{1-\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

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

$\hat{x}_\alpha$   $\alpha$ -Quantil, empirisch, z.B.  $= x_{(\lfloor \alpha n \rfloor + 1)}$ : falls  $\alpha n$  nicht ganzz

# Q-Q-Plot, Variante 1

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$\frac{x_\alpha - \mu}{\sigma} = u_\alpha = \Phi^{-1}(\alpha) \quad \text{gdw.} \quad x_\alpha = \sigma \Phi^{-1}(\alpha) + \mu$$

## Q-Q-Plot, Variante 1

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$\frac{x_\alpha - \mu}{\sigma} = u_\alpha = \Phi^{-1}(\alpha) \quad \text{gdw.} \quad x_\alpha = \sigma\Phi^{-1}(\alpha) + \mu$$

Wenn Normalverteilung zutrifft, so müssen die Punkte  $(\Phi^{-1}(\alpha), \hat{x}_\alpha)$  etwa auf einer Geraden liegen,

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

**ODS GRAPHICS ON; PROC UNIVARIATE PLOT; RUN;**

## Q-Q-Plot, Variante 1

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$\frac{x_\alpha - \mu}{\sigma} = u_\alpha = \Phi^{-1}(\alpha) \quad \text{gdw.} \quad x_\alpha = \sigma\Phi^{-1}(\alpha) + \mu$$

Wenn Normalverteilung zutrifft, so müssen die Punkte  $(\Phi^{-1}(\alpha), \hat{x}_\alpha)$  etwa auf einer Geraden liegen,

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

**ODS GRAPHICS ON; PROC UNIVARIATE PLOT; RUN;**

Die theoretischen Werte ( $\circ$ ) und die theoretische Gerade werden eingezeichnet.

Je näher die  $\circ$  an der Geraden desto mehr spricht es für Normalverteilung und umgekehrt. [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

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

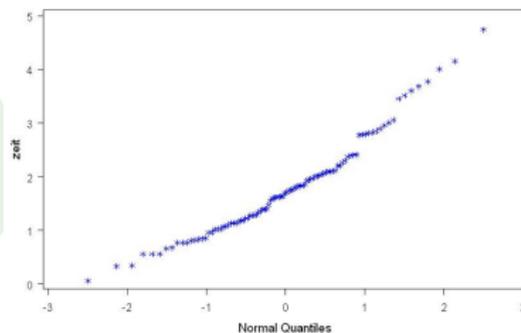
geplottet.  $i = 1, \dots, n$ .

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 (1/2)

Lebensdauern von 100  
Kühlaggregaten

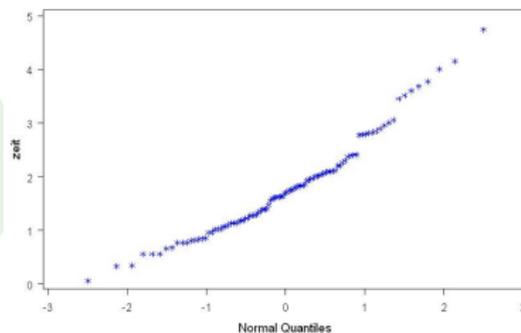
Lebensdauer von Kühlaggregaten  
(in Jahren)



# Q-Q Plots - Beispiele (1/2)

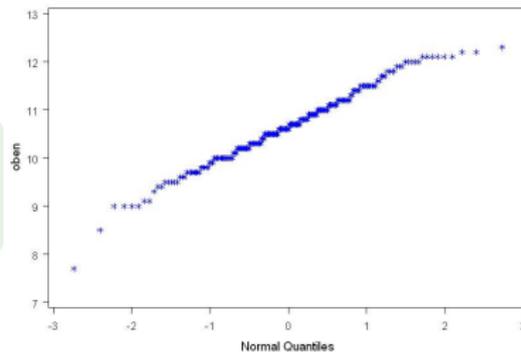
Lebensdauern von 100  
Kühlaggregaten

**Lebensdauer von Kühlaggregaten**  
(in Jahren)



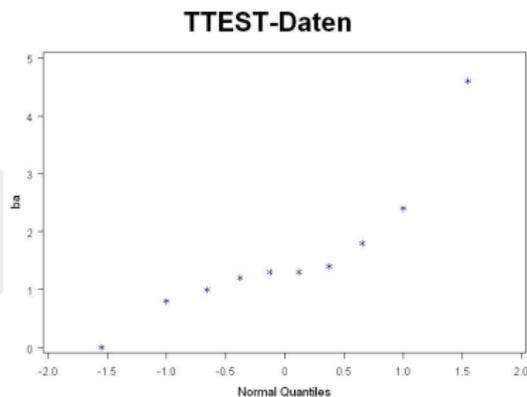
Abmessungen von  
Banknoten

**Banknoten, Variable oben**



# Q-Q Plots - Beispiele (2/2)

Verlängerung der  
Schlafdauer



# Probability Plot

**PROC UNIVARIATE;**

**PROBPLOT** var /Optionen;

**RUN;**

wie oben, x-Achse hat die selbe Skala, aber eine andere Beschriftung, statt  $\Phi^{-1}(u)$  steht  $u$ , 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

Eigenschaften der QQ-Kurve	Interpretation
<p>wenige Punkte weg von der Geraden</p> <p>linkes Ende unter der Linie</p> <p>rechtes Ende über der Linie</p> <p>linkes Ende über der Linie</p> <p>rechtes Ende unter der Linie</p> <p>gebogene Kurve, steigender Anstieg</p> <p>gebogene Kurve, fallender Anstieg</p> <p>Plateaus und Sprünge</p>	<p>Ausreißer</p> <p>lange Tails</p> <p>kurze Tails</p> <p>rechtsschief</p> <p>linksschief</p> <p>diskrete Daten</p> <p>gerundete Daten</p>
<p>Gerade <math>y = x</math></p> <p>Gerade <math>y = ax + b</math></p>	<p>empirische <math>\approx</math> theoretische Verteilung</p> <p>nur Lage- oder Skalenunterschied</p>

# Inhalt

- Box-Plots
- Probability Plots
- Histogramme
- Dichteschätzung

**PROC GCHART DATA**=sasdatei;

<b>VBAR</b>	Variablenliste /Optionen;	/*vertikales Histogramm*/
<b>HBAR</b>	Variablenliste /Optionen;	/*horizontales Histogramm*/
<b>PIE</b>	Variablenliste /Optionen;	/*Kreisdiagramm*/
<b>STAR</b>	Variablenliste /Optionen;	/*Sterndiagramm*/
<b>BLOCK</b>	Variablenliste /Optionen;	/*3 dim. Balkendiagramm*/
<b>VBAR3D</b>	Variablenliste /Optionen;	
<b>HBAR3D</b>	Variablenliste /Optionen;	
<b>PIE3D</b>	Variablenliste /Optionen;	
<b>RUN</b> ;		

# Histogramme, 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 kumulative Häufigkeiten

= CPERCENT kumulative Prozente

= SUM Summen (nur mit SUMVAR)

SUMVAR = anzahl Anzahl ist bereits aufsummierte Häufigkeit

## Histogramme, 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](#)

# Histogramme (Design)

PATTERN<sub>xn</sub>    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.

# Inhalt

- Box-Plots
- Probability Plots
- Histogramme
- Dichteschätzung

# 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	Nichtparametrische 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;**

# 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:  $\bar{X}$  und  $s^2$  sind optimale Schätzungen für  $\mu$  und  $\sigma^2$ .

# 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:  $\bar{X}$  und  $s^2$  sind optimale Schätzungen für  $\mu$  und  $\sigma^2$ .

Wie findet man (gute) Schätzungen bei anderen Verteilungen? → Abschnitt Schätzmethoden.

SAS berechnet in der Regel Maximum-Likelihood-Schätzungen.

# Nichtparametrische Dichteschätzung

Überlagerung der Daten mit einer (Dichte-) Funktion

$K(t)$  eine Kernfunktion,

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

# Nichtparametrische Dichteschätzung

Überlagerung der Daten mit einer (Dichte-) Funktion

$K(t)$  eine Kernfunktion,

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

Dichteschätzung oder Dichtefunktionsschätzung.

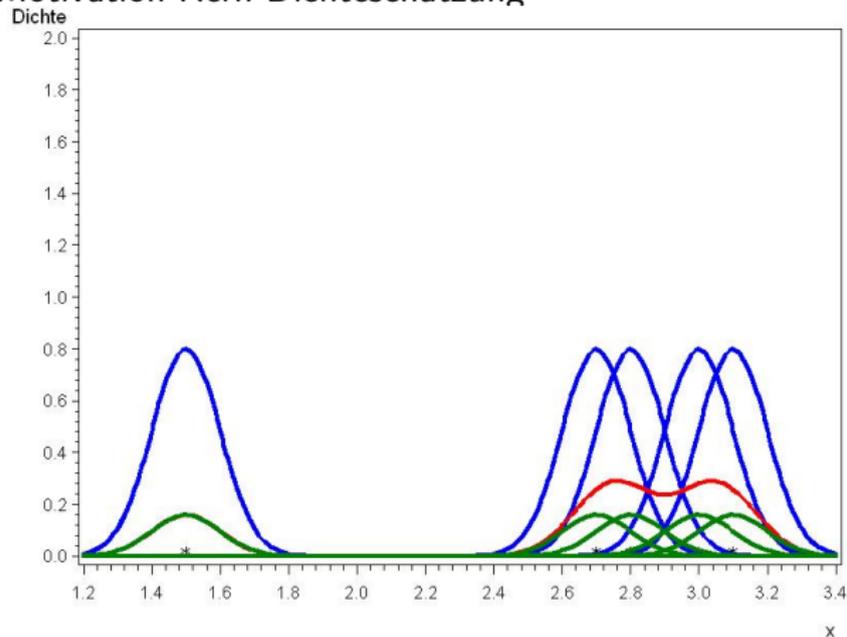
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

$x_1, \dots, x_n$ : die Beobachtungen.

$h$ : ein sogenannter Glättungsparameter.

# Dichteschätzung

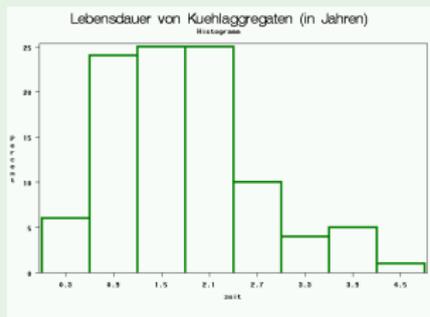
## Motivation Kern-Dichteschätzung



Descr\_Dichteschätzung.sas

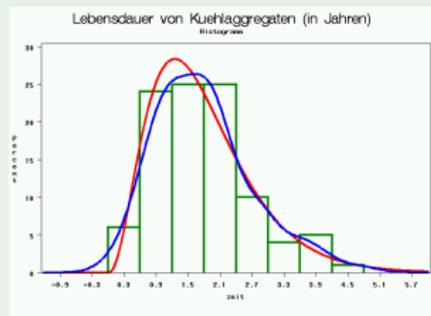
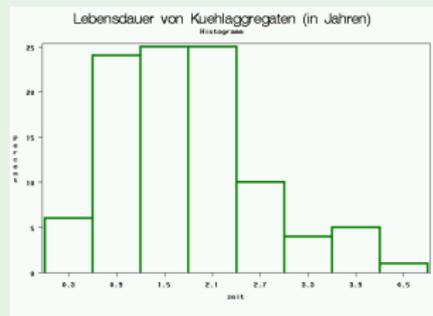
# Dichteschätzung, Beispiel

## Kühlaggregate



# Dichteschätzung, Beispiel

## Kühlaggregate



## Histogramm

Parametrische Dichteschätzung (Gamma)

Nichtparametrische Dichteschätzung

# Dichteschätzung

## Wahl des Kernes $K$

- Normaldichte
- Epanechnikov-Kern (minimiert, bei gegebenem  $h$ , den IMSE)

# Dichteschätzung

## Wahl des Kernes $K$

- Normaldichte
- Epanechnikov-Kern (minimiert, bei gegebenem  $h$ , den IMSE)

## Wahl des Glättungsparameters $h$

Minimiere den Integrated Mean Square Error bzgl.  $h$

$$\begin{aligned} \text{IMSE} &= \int (\mathbf{E}\hat{f}_h(t) - f(t))^2 dt + \int \text{var}(\hat{f}_h(t)) dt \\ &\approx \frac{h^4}{4} \int (f''(t))^2 dt + \frac{1}{nh} \int K^2(t) dt \end{aligned}$$

# Dichteschätzung

## Wahl des Kernes $K$

- Normaldichte
- Epanechnikov-Kern (minimiert, bei gegebenem  $h$ , den IMSE)

## Wahl des Glättungsparameters $h$

Minimiere den Integrated Mean Square Error bzgl.  $h$

$$\begin{aligned} \text{IMSE} &= \int (\mathbf{E}\hat{f}_h(t) - f(t))^2 dt + \int \text{var}(\hat{f}_h(t)) dt \\ &\approx \frac{h^4}{4} \int (f''(t))^2 dt + \frac{1}{nh} \int K^2(t) dt \end{aligned}$$

Frage: Was ist hier  $f$ ? Das ist doch zu schätzen!

# Dichteschätzung

## Wahl des Kernes $K$

- Normaldichte
- Epanechnikov-Kern (minimiert, bei gegebenem  $h$ , den IMSE)

## Wahl des Glättungsparameters $h$

Minimiere den Integrated Mean Square Error bzgl.  $h$

$$\begin{aligned} \text{IMSE} &= \int (\mathbf{E}\hat{f}_h(t) - f(t))^2 dt + \int \text{var}(\hat{f}_h(t)) dt \\ &\approx \frac{h^4}{4} \int (f''(t))^2 dt + \frac{1}{nh} \int K^2(t) dt \end{aligned}$$

Frage: Was ist hier  $f$ ? Das ist doch zu schätzen!

Standard-Annahme hier:  $f$  ist normal-Dichte.

# Inhalt

- Häufigkeitstabellen
- Scatterplots, Zusammenhangsmaße
- Das Regressionsproblem

# Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

## Eindimensionale Zufallsvariablen

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

Die  $p_i$  sind zu schätzen:

# Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

## Eindimensionale Zufallsvariablen

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

Die  $p_i$  sind zu schätzen:

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

$N$ : Stichprobenumfang  $n_i$ : relative Häufigkeiten

# Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

## Eindimensionale Zufallsvariablen

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

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**;

# Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel

## 3maliges Werfen einer Münze

$X$ : Anzahl von Blatt nach 3 Würfeln

$Y$ : Anzahl von Blatt nach 2 Würfeln

Element von $\Omega$	$X$	$Y$
BBB	3	2
BBZ	2	2
BZB	2	1
BZZ	1	1
ZBB	2	1

## Besetzungswahrscheinlichkeiten

$X Y$	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

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \quad Y : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

# Tabelle der zweidimensionalen Wahrscheinlichkeiten

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

## Zweidimensionale Zufallsvariable

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

## 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  $(X, Y)$ ,  $i = 1, \dots, M, j = 1, \dots, N$ .

gemeinsame Wahrscheinlichkeitsfunktion von  $(X, Y)$

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

## 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  $(X, Y)$ ,  $i = 1, \dots, M, j = 1, \dots, N$ .

## gemeinsame Wahrscheinlichkeitsfunktion von $(X, Y)$

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

$$\begin{array}{lll} p_{ij} \geq 0 & p_{i.} := \sum_{j=1}^N p_{ij} & p_{.j} := \sum_{i=1}^M p_{ij} \\ \sum_{i,j} p_{ij} = 1 & & \end{array}$$

$X$  und  $Y$  heißen unabhängig, wenn

$$p_{ij} = P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) = p_{i.} \cdot p_{.j}$$

$X$  und  $Y$  heißen unabhängig, wenn

$$p_{ij} = P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) = p_{i.} \cdot p_{.j}$$

$p_{i.}$  und  $p_{.j}$  heißen Randwahrscheinlichkeiten.

## 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?

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	$n_{00}$	$n_{01}$	$n_{0\cdot}$ Nichtsportler
1	$n_{10}$	$n_{11}$	$n_{1\cdot}$ Sportler
	$n_{\cdot 0}$	$n_{\cdot 1}$	
	# 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.

# 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.

# Optionen der TABLES-Anweisung

nur für mehrdimensionale Tabellen

CHISQ:  $\chi^2$ -Unabhängigkeitstest

CMH: u.a. Odds Ratio

MEASURES: Assoziationsmaße,  
Korrelationskoeffizient

NO... keine Ausgabe von:

NOFREQ: absoluten Häufigkeiten

NOPERCENT: relativen Häufigkeiten

NOROW: Zeilenhäufigkeiten

NOCOL: Spaltenhäufigkeiten

# Assoziationsmaße

nur für mehrdimensionale Tabellen

 $\chi^2$ 

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

# Assoziationsmaße

nur für mehrdimensionale Tabellen

 $\chi^2$ 

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

$\Phi$ -Koeffizient für 2x2 Tafeln

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

# Assoziationsmaße

nur für mehrdimensionale Tabellen

 $\chi^2$ 

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

$\Phi$ -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:

# Assoziationsmaße

nur für mehrdimensionale Tabellen

 $\chi^2$ 

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

$\Phi$ -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.

## Beispiel: Mendelsche Kreuzungsversuche

**DATA** Erbsen;

**INPUT** rund gruen Anzahl;

**CARDS**;

0 0 101

0 1 32

1 0 315

1 1 108

;

**RUN**;

## 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;
```

## 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;
```

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

# Inhalt

- Häufigkeitstabellen
- Scatterplots, Zusammenhangsmaße
- Das Regressionsproblem

Erinnerung: Varianz der Zufallsvariablen  $X$ 

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

Erinnerung: Varianz der Zufallsvariablen  $X$ 

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

Kovarianz der Zufallsvariablen  $X$  und  $Y$ 

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

Erinnerung: Varianz der Zufallsvariablen  $X$ 

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

Kovarianz der Zufallsvariablen  $X$  und  $Y$ 

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

Korrelation der Zufallsvariablen  $X$  und  $Y$ 

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

# Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

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

## Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

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

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$$

## Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

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

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$$

empirische Korrelation, Pearson-Korrelationskoeffizient

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

# Pearson-Korrelationskoeffizient, Eigenschaften

- Es gilt stets:

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

# Pearson-Korrelationskoeffizient, Eigenschaften

- Es gilt stets:

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

- Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen

$$x \longrightarrow a + bx$$

# 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,

$$y = mx + b, m \neq 0$$

$$r_{XY} = 1 \rightarrow \text{Anstieg} > 0$$

$$r_{XY} = -1 \rightarrow \text{Anstieg} < 0$$

# Pearson-Korrelationskoeffizient

- Der Pearson-Korrelationskoeffizient ist also ein Maß für die lineare Abhängigkeit von  $X$  und  $Y$ .
- $r_{XY} \approx 0 \rightarrow$  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:

# Pearson-Korrelationskoeffizient

- Der Pearson-Korrelationskoeffizient ist also ein Maß für die lineare Abhängigkeit von  $X$  und  $Y$ .
- $r_{XY} \approx 0 \rightarrow$  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

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

$R_i$ : Rang von  $X_i$  in der geordneten Stichprobe  $X_{(1)} \leq \dots \leq X_{(n)}$

$S_i$ : Rang von  $Y_i$  in der geordneten Stichprobe  $Y_{(1)} \leq \dots \leq Y_{(n)}$

# Spearman-Korrelationskoeffizient

## Spearman-Rangkorrelationskoeffizient

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

$R_i$ : Rang von  $X_i$  in der geordneten Stichprobe  $X_{(1)} \leq \dots \leq X_{(n)}$

$S_i$ : Rang von  $Y_i$  in der geordneten Stichprobe  $Y_{(1)} \leq \dots \leq Y_{(n)}$

```
PROC CORR SPEARMAN;
```

```
VAR X Y;
```

```
RUN;
```

# Spearman-Korrelationskoeffizient

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

# Spearman-Korrelationskoeffizient

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

# Spearman-Korrelationskoeffizient

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

# Spearman-Korrelationskoeffizient

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

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

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

# Spearman-Korrelationskoeffizient

Beweis der letzten Formel (1)

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

Nenner:

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

# Spearman-Korrelationskoeffizient

Beweis der letzten Formel (1)

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

Nenner:

$$\begin{aligned} \sum_{i=1}^n (R_i - \bar{R})^2 &= \sum_{i=1}^n (S_i - \bar{S})^2 = \sum_{i=1}^n \left(i - \frac{n+1}{2}\right)^2 \\ &= \sum i^2 - 2 \cdot \frac{n+1}{2} \sum i + n \cdot \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n \cdot (n+1) \cdot (2n+1)}{6} - \frac{n \cdot (n+1)^2}{2} + \frac{n \cdot (n+1)^2}{4} \\ &= \frac{n \cdot (n+1)}{12} \cdot [2 \cdot (2n+1) - 3 \cdot (n+1)] \\ &= \frac{(n-1) \cdot n \cdot (n+1)}{12} = \frac{n \cdot (n^2 - 1)}{12} \end{aligned}$$

# Spearman-Korrelationskoeffizient

Beweis der letzten Formel (2)

Zähler:

$$\begin{aligned}\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S}) &= \sum_{i=1}^n \left(R_i - \frac{n+1}{2}\right) \left(S_i - \frac{n+1}{2}\right) \\ &= \sum_{i=1}^n R_i S_i - 2 \cdot \frac{n+1}{2} \sum_{i=1}^n R_i + n \cdot \left(\frac{n+1}{2}\right)^2 \\ &= \sum_{i=1}^n R_i S_i - \frac{n \cdot (n+1)^2}{4}\end{aligned}$$

# Spearman-Korrelationskoeffizient

Beweis der letzten Formel (2)

Zähler:

$$\begin{aligned}\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S}) &= \sum_{i=1}^n \left(R_i - \frac{n+1}{2}\right) \left(S_i - \frac{n+1}{2}\right) \\ &= \sum_{i=1}^n R_i S_i - 2 \cdot \frac{n+1}{2} \sum_{i=1}^n R_i + n \cdot \left(\frac{n+1}{2}\right)^2 \\ &= \sum_{i=1}^n R_i S_i - \frac{n \cdot (n+1)^2}{4}\end{aligned}$$

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

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

Bem.: Es gilt:

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

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

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

# 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 Normalverteilung.

# Vergleich der Korrelationskoeffizienten

Pearson - Spearman

## Nachteile Spearman

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

# Kendalls $\tau$ (Konkordanzkoeffizient)

$(X_i, Y_i), i = 1, \dots, n$

$$a_{ij} = \begin{cases} 1, & \text{falls } x_i < x_j \wedge y_i < y_j \text{ oder} \\ & x_i > x_j \wedge y_i > y_j \\ -1, & \text{falls } x_i < x_j \wedge y_i > y_j \text{ oder} \\ & x_i > x_j \wedge y_i < y_j \\ 0, & \text{sonst} \end{cases}$$
$$= \operatorname{sgn}[(x_i - x_j)(y_i - y_j)]$$

Falls  $a_{ij} = 1$  so heißen die Punktepaare  $(x_i, y_i), (x_j, y_j)$  konkordant

Falls  $a_{ij} = -1$  so heißen sie diskordant

Falls  $a_{ij} = 0$  so heißen sie gebunden

# Kendalls $\tau$ (Konkordanzkoeffizient)

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

# Kendalls $\tau$ (Konkordanzkoeffizient)

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

Bem.: einfache Berechnung, wenn neue Paare hinzukommen

Bem.: Es gilt, falls  $X, Y$  stetig:  $-1 \leq 3\tau - 2r_S \leq 1$ .

$$\text{Approximation: } \tau_{\text{appr.}} = \frac{2}{3} \frac{N+1}{N} r_S$$

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

# Inhalt

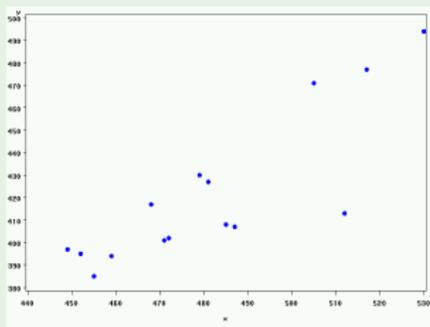
- Häufigkeitstabellen
- Scatterplots, Zusammenhangsmaße
- Das Regressionsproblem

## 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:

$$Y = f\left(X, \underbrace{\theta_1, \dots, \theta_p}_{\text{Parameter}}\right) + \underbrace{\epsilon}_{\text{zufälliger Fehler}}, \quad \epsilon \sim (0, \sigma^2).$$

# Das Regressionsproblem

$X, Y$ : Zufallsvariablen (auch mehrdimensional)

Modell:

$$Y = f\left(X, \underbrace{\theta_1, \dots, \theta_p}_{\text{Parameter}}\right) + \underbrace{\epsilon}_{\text{zufälliger Fehler}}, \quad \epsilon \sim (0, \sigma^2).$$

$f$  linear, bekannt bis auf Parameter:

lineare Regression

# Das Regressionsproblem

$X, Y$ : Zufallsvariablen (auch mehrdimensional)

Modell:

$$Y = f\left(X, \underbrace{\theta_1, \dots, \theta_p}_{\text{Parameter}}\right) + \underbrace{\epsilon}_{\text{zufälliger Fehler}}, \quad \epsilon \sim (0, \sigma^2).$$

$f$  linear, bekannt bis auf Parameter:

lineare Regression

$f$  nichtlinear, bekannt bis auf Parameter:

nichtlineare Regression

# Das Regressionsproblem

$X, Y$ : Zufallsvariablen (auch mehrdimensional)

Modell:

$$Y = f\left(X, \underbrace{\theta_1, \dots, \theta_p}_{\text{Parameter}}\right) + \underbrace{\epsilon}_{\text{zufälliger Fehler}}, \quad \epsilon \sim (0, \sigma^2).$$

$f$  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:

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

Parameter  $\theta_1, \dots, \theta_p$ : unbekannt.

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

# Regression

$f$  bekannt (bis auf Parameter)

Aufgabe:

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

Parameter  $\theta_1, \dots, \theta_p$ : unbekannt.

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

Erwartungswert durch arithmetisches Mittel ersetzen

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

Kleinste Quadrat-Schätzung für  $\theta_1, \dots, \theta_p$  (KQS)

Least-Squares-Estimation (LSE)

# Regression

$f$  bekannt (bis auf Parameter)

## Lösung des Minimum-Problems

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

# Regression

$f$  bekannt (bis auf Parameter)

## Lösung des Minimum-Problems

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

zu minimierende Funktion nach den Parametern differenzieren und Null setzen:

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

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

# Regression

$f$  linear: lineares Gleichungssystem (1)

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

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

$$\frac{1}{n} \sum_{i=1}^n (Y_i - (\theta_1 X_i + \theta_2)) \cdot X_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n (Y_i - (\theta_1 X_i + \theta_2)) \cdot 1 = 0$$

---

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_2 \sum_i X_i = 0$$

$$\sum_i Y_i - \theta_1 \sum_i X_i - \theta_2 \cdot n = 0$$

# Regression

$f$  linear: lineares Gleichungssystem (2)

Die zweite Gleichung nach  $\theta_2$  auflösen:

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

# Regression

$f$  linear: lineares Gleichungssystem (2)

Die zweite Gleichung nach  $\theta_2$  auflösen:

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

und in die erste einsetzen:

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

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

# Regression

$f$  linear: lineares Gleichungssystem (2)

Die zweite Gleichung nach  $\theta_2$  auflösen:

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

und in die erste einsetzen:

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

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

$\Rightarrow$

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

# Regression

Zähler und Nenner in  $\hat{\theta}_1$

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

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

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

Minimierungsaufgabe:

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

Lösung:

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

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

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

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

Schätzung des Schätzfehlers

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

Schätzfehler:

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

# Lineare und Nichtlineare Regression

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

$\theta_1$  und  $\theta_2$  werden geschätzt.

[Descr\\_Scatter\\_1.sas](#)    [Descr\\_Scatter\\_Heroin.sas](#)

# Lineare und Nichtlineare Regression

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

$\theta_1$  und  $\theta_2$  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, z.B.

$$Y = \ln(\theta_1 X + \theta_2) + \epsilon$$

$$e^Y = \theta_1 X + \theta_2 + \tilde{\epsilon}$$

Modelle sind aber i.A. nicht äquivalent!

# Weitere nichtlineare Regressionsfunktionen

Auswahl

$$f(t) = a + bt + ct^2 \quad \text{Parabel}$$

$$f(t) = at^b \quad \text{Potenzfunktion}$$

$$f(t) = ae^t \quad \text{Exponentialfunktion}$$

$$f(t) = k - ae^{-t}$$

$$f(t) = \frac{k}{1 + be^{-ct}} \quad \text{logistische Funktion}$$

$$\ln f(t) = k - \frac{a}{b + t} \quad \text{Johnson-Funktion}$$

$$\ln f(t) = k - \lambda e^{-t} \quad \text{Gompertz-Funktion}$$

# Nichtparametrische Regression

$f$  unbekannt, aber "glatt"

Sei  $f$  2x stetig differenzierbar,  $f \in C_2$ ,  $\lambda \geq 0$

Ziel: 
$$\min_{f \in C_2} \left( \sum_{i=1}^n (Y_i - f(X_i))^2 \right)$$

# Nichtparametrische Regression

$f$  unbekannt, aber "glatt"

Sei  $f$  2x stetig differenzierbar,  $f \in C_2$ ,  $\lambda \geq 0$

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

# Nichtparametrische Regression

$f$  unbekannt, aber "glatt"

Sei  $f$  2x stetig differenzierbar,  $f \in C_2$ ,  $\lambda \geq 0$

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

Lösung: Glättender Kubischer Spline.

**PROC GPLOT**      Descr\_Scatter.sas

# Nichtparametrische Regression

$f$  unbekannt, aber "glatt"

Sei  $f$  2x stetig differenzierbar,  $f \in C_2$ ,  $\lambda \geq 0$

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

Lösung: Glättender Kubischer Spline.

**PROC GPLOT**      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, **wird in SAS nicht mehr angeboten, s.u.**

$K$ : Kernfunktion, standardisierte Dichte (z.B. Normal, Epanechnikov-Kern).

Regressionsmodell:

$$Y = f(X) + \epsilon, \quad \epsilon \sim (0, \sigma^2) \quad \text{also}$$
$$\mathbf{E}(Y|X = x) = f(x)$$

# Nichtparametrische Regression

Kernschätzung, Motivation, **wird in SAS nicht mehr angeboten, s.u.**

$K$ : Kernfunktion, standardisierte Dichte (z.B. Normal, 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

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

$g(x, y)$ : gemeinsame Dichte von  $(X, Y)$

$f_0(x)$ : Randdichte von  $X$

$f_{Y|X}$ : bedingte Dichte von  $Y$

# Regression, Kernschätzung

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

$g(x, y)$ : gemeinsame Dichte von  $(X, Y)$

$f_0(x)$ : Randdichte von  $X$

$f_{Y|X}$ : bedingte Dichte von  $Y$

Der Nenner wird geschätzt durch

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

# Regression, Kernschätzung

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

$g(x, y)$ : gemeinsame Dichte von  $(X, Y)$

$f_0(x)$ : Randdichte von  $X$

$f_{Y|X}$ : bedingte Dichte von  $Y$

Der Nenner wird geschätzt durch

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

und der Zähler durch

$$\frac{1}{n} \sum_{i=1}^n y_i \hat{g}(x_i, y_i) = \frac{1}{n} \sum_{i=1}^n y_i \cdot \frac{1}{h} \cdot K\left(\frac{x - x_i}{h}\right)$$

# 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

**Bem.** Diese (feste) Kernschätzung hat einen Bias, insbesondere an den Rändern. Deshalb nimmt man *variable* Kernschätzungen, oder was sehr ähnlich ist, sogen. LOESS (local weighted scatterplot smoothing): in SAS: **proc loess**

# 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 S: Daten müssen vorher nach dem x-Merkmal sortiert sein	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

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

## stetige Verteilung

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

Bsp: diskrete Verteilung: Binomial, Poisson

stetige Verteilung: Normal, Gleich, Exponential

# 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

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

### Normalverteilung, Dichte

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

# Beschreibende Statistik

## Zusammenfassung (3)

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

$$\bar{X} \longrightarrow \mathbf{E}X, \quad \bar{X} = \frac{1}{n} \sum X_i$$

# Beschreibende Statistik

## Zusammenfassung (3)

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

$$\bar{X} \longrightarrow \mathbf{E}X, \quad \bar{X} = \frac{1}{n} \sum X_i$$

### Zentraler Grenzwertsatz ( $X_i$ iid)

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

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

$$s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 \rightarrow \sigma^2$$

# Beschreibende Statistik

Zusammenfassung (4)

## Statistische Maßzahlen

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

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

Formmaße:  $\beta_1$ ,  $\beta_2$

**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**

HISTOGRAM

Häufigkeitstabellen:

**PROC FREQ**

Zusammenhangsmaße:

**PROC CORR**

Pearson, Spearman, Kendall-Korrelationskoeffizient

Scatterplots, Regression

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

Sei  $X$  ein Merkmal (eine Zufallsvariable),

$$F_X(x) = P(X \leq x) = P_\theta(X \leq x) = F_{X,\theta}(x) \quad \theta: \text{Parametervektor}$$

Beispiel:  $\theta = (\mu, \sigma^2)$

$\mu$ : Erwartungswert von  $X$

$\sigma^2$ : Varianz von  $X$

Sei  $X$  ein Merkmal (eine Zufallsvariable),

$$F_X(x) = P(X \leq x) = P_\theta(X \leq x) = F_{X,\theta}(x) \quad \theta: \text{Parametervektor}$$

Beispiel:  $\theta = (\mu, \sigma^2)$

$\mu$ : Erwartungswert von  $X$

$\sigma^2$ : Varianz von  $X$

$X_1, X_2, \dots, X_n$  Beobachtungen von  $X$

$$\mu \approx \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

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

D.h. die unbekannt Parameter werden geschätzt.

# Statistische Tests: Einführung

## Problem

Schätzungen können sehr schlecht ausfallen!

I.a. vertritt der Fachexperte gewisse Hypothesen bzgl. der (unbekannten) Parameterwerte!

## Hypothesenpaar: Nullhypothese-Alternativhypothese

Die Nullypothesen werden verworfen, wenn die erhaltenen Schätzwerte (z.B.  $\bar{X}$ ,  $s^2$ ) mit ihnen nicht in Einklang stehen.

# Statistische Tests: Einführungsbeispiele

- Einstichprobenproblem, einfache Alternative

$$H_0 : \mu = \mu_0 \quad H_1 : \mu = \mu_1, \quad (\mu_1 \neq \mu_0)$$

- Einstichprobenproblem, zusammengesetzte (zweiseitige)

Alternative  $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$

- Einstichprobenproblem, zusammengesetzte (einseitige)

Alternative  $H_0 : \mu = \mu_0 \quad H_1 : \mu > \mu_0$

- Zweistichprobenproblem, einfache Alternative

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 = \mu_2 + \theta, \quad (\theta > 0, \text{ fest})$$

- Zweistichprobenproblem, zusammengesetzte (zweiseitige)

Alternative  $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2,$

# Statistische Tests: Eine verwandte Problemstellung

## Elektronischer Großhandel: TV-Geräte

Händler sagt: Ausschußquote  $p \leq 1\%$  ( $p = 0.01$ )

Käufer wäre einverstanden, prüft aber  $N$  Geräte!

# Statistische Tests: Eine verwandte Problemstellung

## Elektronischer Großhandel: TV-Geräte

Händler sagt: Ausschußquote  $p \leq 1\%$  ( $p = 0.01$ )

Käufer wäre einverstanden, prüft aber  $N$  Geräte!

Davon:  $N_f$  fehlerhaft,  $N_f$ : Teststatistik

$$\frac{N_f}{N} \cdot 100\% \gg 1\% \Rightarrow \text{Ablehnung}$$

# Statistische Tests: Eine verwandte Problemstellung

## Elektronischer Großhandel: TV-Geräte

Händler sagt: Ausschußquote  $p \leq 1\%$  ( $p = 0.01$ )

Käufer wäre einverstanden, prüft aber  $N$  Geräte!

Davon:  $N_f$  fehlerhaft,  $N_f$ : Teststatistik

$$\frac{N_f}{N} \cdot 100\% \gg 1\% \Rightarrow \text{Ablehnung}$$

Zwei Fehler möglich

a) Zufällig  $N_f$  zu groß!  $p < 0.01$

$\Rightarrow$  Käufer lehnt ab

b) Zufällig  $N_f$  zu klein!  $p$  groß,  $p \gg 0.01$

$\Rightarrow$  Käufer kauft

# Statistische Tests: Risiken - Fehler

## Risiko des Händlers

Käufer lehnt gute Ware ab (weil  $N_f$  zufällig zu groß)

# Statistische Tests: Risiken - Fehler

## Risiko des Händlers

Käufer lehnt gute Ware ab (weil  $N_f$  zufällig zu groß)

## Risiko des Käufers

Käufer kauft schlechte Ware (weil  $N_f$  zufällig zu klein)

# Statistische Tests: Risiken - Fehler

## Risiko des Händlers

Käufer lehnt gute Ware ab (weil  $N_f$  zufällig zu groß)

## Risiko des Käufers

Käufer kauft schlechte Ware (weil  $N_f$  zufällig zu klein)

## Risiken sollen quantifiziert werden:

- a)  $P(\text{Nicht kaufen} \mid p \leq 1\%)$
- b)  $P(\text{Kaufen} \mid p > 1\%)$

Beide Risiken nicht gleichzeitig zu minimieren.

# Statistische Tests: Risiken - Fehler

## Risiko des Händlers

Käufer lehnt gute Ware ab (weil  $N_f$  zufällig zu groß)

## Risiko des Käufers

Käufer kauft schlechte Ware (weil  $N_f$  zufällig zu klein)

## Risiken sollen quantifiziert werden:

- a)  $P(\text{Nicht kaufen} \mid p \leq 1\%)$
- b)  $P(\text{Kaufen} \mid p > 1\%)$

Beide Risiken nicht gleichzeitig zu minimieren.

## Lösung:

$P(\text{Nicht kaufen} \mid p \leq 1\%) = \alpha$  vorgeben

$P(\text{Kaufen} \mid p > 1\%)$  minimieren (oder es versuchen)

# Hypothesentest

Beispiel: Einstichproben-Lagetest

Sei  $\mu$  ein Lageparameter, z.B. der Erwartungswert.

Sei  $\mu_0$  ein vorgegebener Wert.

## Nullhypothese und Alternativhypothese

$$\text{a) } H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$$

$$\text{b) } H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$$

$$\text{c) } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

# Hypothesentest

Beispiel: Einstichproben-Lagetest

Sei  $\mu$  ein Lageparameter, z.B. der Erwartungswert.

Sei  $\mu_0$  ein vorgegebener Wert.

## Nullhypothese und Alternativhypothese

$$\text{a) } H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$$

$$\text{b) } H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$$

$$\text{c) } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

## Teststatistik, Testgröße, Prüfgröße, Stichprobenfunktion

$$T = T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

# Hypothesentest, allgemein

Die Entscheidung für  $H_A$  oder für  $H_0$  wird anhand einer Teststatistik

$$T = T(x_1, \dots, x_n)$$

gefällt.

Liegt der Wert von  $T$  in einem vorher bestimmten Bereich  $K$ , dem sogen. Ablehnungsbereich oder kritischen Bereich, dann wird  $H_0$  abgelehnt, anderenfalls wird  $H_0$  nicht abgelehnt.

# Hypothesentest, allgemein

Die Entscheidung für  $H_A$  oder für  $H_0$  wird anhand einer Teststatistik

$$T = T(x_1, \dots, x_n)$$

gefällt.

Liegt der Wert von  $T$  in einem vorher bestimmten Bereich  $K$ , dem sogen. Ablehnungsbereich oder kritischen Bereich, dann wird  $H_0$  abgelehnt, anderenfalls wird  $H_0$  nicht abgelehnt.

$T \in K \Rightarrow H_0$  ablehnen, Entscheidung für  $H_A$

$T \notin K \Rightarrow H_0$  nicht ablehnen, Entscheidung für  $H_0$ .

$$\phi(x_1, \dots, x_n) = \begin{cases} 1 & \text{falls } T \in K \\ 0 & \text{sonst} \end{cases}$$

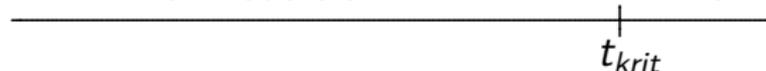
# Hypothesentest, Annahme- und Ablehnungsbereich

• a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmebereich

Krit. Bereich



# Hypothesentest, Annahme- und Ablehnungsbereich

- a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmebereich

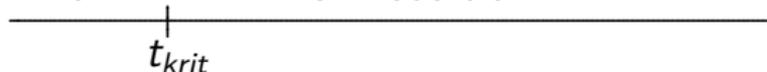
Krit. Bereich



- b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

kleine Werte von  $T$  sprechen für  $H_A$ .

Krit. B.      Annahmebereich



# Hypothesentest, Annahme- und Ablehnungsbereich

- a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmebereich

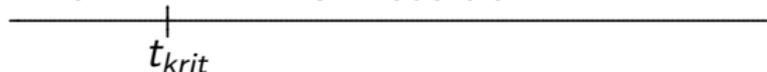
Krit. Bereich



- b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

kleine Werte von  $T$  sprechen für  $H_A$ .

Krit. B.      Annahmebereich



- c)  $H_0 : \mu = \mu_0$        $H_A : \mu \neq \mu_0$

große Werte von  $|T|$  sprechen für  $H_A$ .

Annahmebereich



# Hypothesentest, Fehler 1. Art, Fehler 2. Art

## Fehler 1. Art

Entscheidung für  $H_A$  obwohl  $H_0$  richtig ist.

# Hypothesentest, Fehler 1. Art, Fehler 2. Art

## Fehler 1.Art

Entscheidung für  $H_A$  obwohl  $H_0$  richtig ist.

## Fehler 2.Art

Entscheidung für  $H_0$  obwohl  $H_A$  richtig ist

# Hypothesentest, Fehler 1. Art, Fehler 2. Art

## Fehler 1.Art

Entscheidung für  $H_A$  obwohl  $H_0$  richtig ist.

## Fehler 2.Art

Entscheidung für  $H_0$  obwohl  $H_A$  richtig ist

	Entscheidung für $H_0$	Entscheidung für $H_A$
$H_0$ richtig	richtig, Sicher- heitswkt. $1 - \alpha$	Fehler 1. Art Fehlerwkt. $\alpha$ .
$H_A$ richtig	Fehler 2.Art Fehlerwkt. $1-\beta$	richtig, Güte $\beta$

# Hypothesentest, Fehler 1. Art, Fehler 2. Art

## Fehler 1.Art

Entscheidung für  $H_A$  obwohl  $H_0$  richtig ist.

## Fehler 2.Art

Entscheidung für  $H_0$  obwohl  $H_A$  richtig ist

	Entscheidung für $H_0$	Entscheidung für $H_A$
$H_0$ richtig	richtig, Sicher- heitswkt. $1 - \alpha$	Fehler 1. Art Fehlerwkt. $\alpha$ .
$H_A$ richtig	Fehler 2.Art Fehlerwkt. $1-\beta$	richtig, Güte $\beta$

Entscheidung für  $H_0$  heißt nicht notwendig, dass  $H_0$  richtig ist.

# Hypothesentest, Fehler 1. Art, Fehler 2. Art

$\alpha$  und  $(1 - \beta)$  können nicht gleichzeitig minimiert werden.

⇒ Man gibt  $\alpha$  vor (z.B.  $\alpha = 0.05$ ), d.h. man behält  $\alpha$  unter Kontrolle und versucht die Teststatistik so zu definieren, daß  $\beta$  maximal wird.

$\beta$  (und manchmal auch  $\alpha$ ) hängen von wahren (i.A. unbekanntem) Parametern ab.

## Signifikanzniveau

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

$\Theta_0$ : Nullhypothesenraum, also z.B. die Menge

$$\{\mu : \mu \geq \mu_0\} \quad \text{oder} \quad \{\mu : \mu = \mu_0\}.$$

## Gütefunktion

$$\beta = \beta(\theta) = \beta(\mu) = P_{\mu}(T \in K)$$

$K$  heißt Ablehnungsbereich oder Kritischer Bereich.

### Beispiel: $t$ -Test

$$\begin{aligned}\beta(\mu) &= P(T \in K) \quad K: \text{kritischer Bereich} \\ &= P(T > t_{1-\alpha, n-1} | \mu, \sigma^2) \\ &= 1 - CDF('T', t_{1-\alpha, n-1}, n - 1, nc)\end{aligned}$$

$$nc = \sqrt{n} \frac{\mu - \mu_0}{\sigma}: \quad \text{Nichtzentralitätsparameter}$$

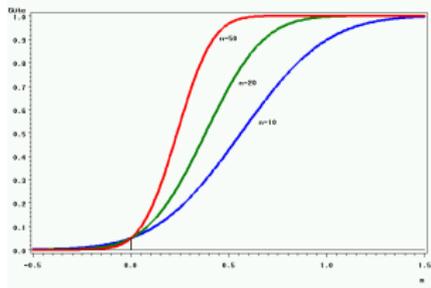
$$t_{1-\alpha, n-1}: \quad \text{kritischer Wert}$$

$$K = [t_{1-\alpha, n-1}, \infty): \quad \text{kritischer Bereich.}$$

# Gütefunktion

## Einseitiger Test

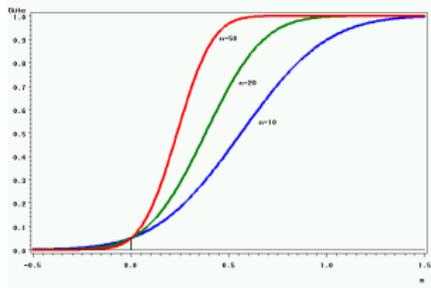
Gütefunktion des einseitigen Einstichproben t-Tests



# Gütefunktion

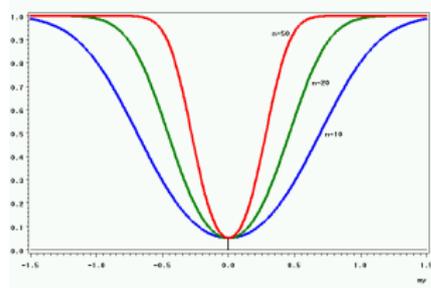
## Einseitiger Test

Gütefunktion des einseitigen Einstichproben t-Tests



## Zweiseitiger Test

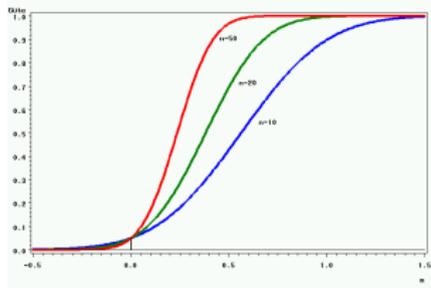
Gütefunktion des zweiseitigen Einstichproben t-Tests



# Gütefunktion

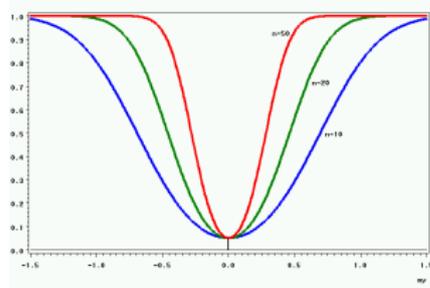
## Einseitiger Test

Gütefunktion des einseitigen Einstichproben t-Tests



## Zweiseitiger Test

Gütefunktion des zweiseitigen Einstichproben t-Tests



Test\_Guete\_t.sas

Test\_Guete\_t2.sas

# Gütefunktion

Ideal:

Unter  $H_0$ : Güte 0 (d.h. Fehler 1. Art =0)

Unter  $H_A$ : Güte 1 (d.h. Fehler 2. Art =0)

Das ist aber nicht möglich!

# Gütefunktion

## Ideal:

Unter  $H_0$ : Güte 0 (d.h. Fehler 1. Art = 0)

Unter  $H_A$ : Güte 1 (d.h. Fehler 2. Art = 0)

Das ist aber nicht möglich!

## Ziel:

Test mit möglichst großer Gütefunktion (unter  $H_A$ ).

Wir schlagen natürlich nur solche "sinnvollen" Tests vor.

# Lagetests (bei Normalverteilungsannahme, 1)

## Einstichprobenproblem

$$H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$$

$$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

# Lagetests (bei Normalverteilungsannahme, 1)

## Einstichprobenproblem

$H_0 : \mu \leq \mu_0$      $H_A : \mu > \mu_0$     Einstichproben t-Test

$H_0 : \mu \geq \mu_0$      $H_A : \mu < \mu_0$     **PROC UNIVARIATE**

$H_0 : \mu = \mu_0$      $H_A : \mu \neq \mu_0$     **PROC TTEST**

# Lagetests (bei Normalverteilungsannahme, 1)

## Einstichprobenproblem

$H_0 : \mu \leq \mu_0$     $H_A : \mu > \mu_0$    Einstichproben t-Test

$H_0 : \mu \geq \mu_0$     $H_A : \mu < \mu_0$    **PROC UNIVARIATE**

$H_0 : \mu = \mu_0$     $H_A : \mu \neq \mu_0$    **PROC TTEST**

## Zweistichprobenproblem

$H_0 : \mu_1 \leq \mu_2$     $H_A : \mu_1 > \mu_2$

$H_0 : \mu_1 \geq \mu_2$     $H_A : \mu_1 < \mu_2$

$H_0 : \mu_1 = \mu_2$     $H_A : \mu_1 \neq \mu_2$

# Lagetests (bei Normalverteilungsannahme, 1)

## Einstichprobenproblem

$H_0 : \mu \leq \mu_0$     $H_A : \mu > \mu_0$    Einstichproben t-Test

$H_0 : \mu \geq \mu_0$     $H_A : \mu < \mu_0$    **PROC UNIVARIATE**

$H_0 : \mu = \mu_0$     $H_A : \mu \neq \mu_0$    **PROC TTEST**

## Zweistichprobenproblem

$H_0 : \mu_1 \leq \mu_2$     $H_A : \mu_1 > \mu_2$    Einstichproben  $t$ -Test (verbundene  
Stichproben)

$H_0 : \mu_1 \geq \mu_2$     $H_A : \mu_1 < \mu_2$     $t$ -Test (unverbundene Stichproben)

$H_0 : \mu_1 = \mu_2$     $H_A : \mu_1 \neq \mu_2$    **PROC UNIVARIATE**  
**PROC TTEST**

# Lage- und Skalentests (bei Normalverteilungsannahme)

## c-Stichprobenproblem

$$H_0 : \mu_1 = \dots = \mu_c \quad H_A : \exists(i, j) : \mu_i \neq \mu_j$$

einfache Varianzanalyse

**PROC ANOVA, PROC GLM**

Andere Alternativen sind z.B.:  $\mu_1 \leq \dots \leq \mu_c$      $\mu_1 \geq \dots \geq \mu_c$

# Lage- und Skalentests (bei Normalverteilungsannahme)

## c-Stichprobenproblem

$$H_0 : \mu_1 = \dots = \mu_c \quad H_A : \exists(i, j) : \mu_i \neq \mu_j$$

einfache Varianzanalyse

**PROC ANOVA, PROC GLM**

Andere Alternativen sind z.B.:  $\mu_1 \leq \dots \leq \mu_c$      $\mu_1 \geq \dots \geq \mu_c$

## Skalentest

Zwei unverbundene Stichproben

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_A : \sigma_1^2 \neq \sigma_2^2$$

**PROC TTEST** (nur wenn wirklich Normalverteilung)

**PROC ANOVA, PROC GLM**

# p-Werte

bisher: “ $H_0$  abgelehnt” oder “ $H_0$  beibehalten”

⇒ wenig informativ.

Wir könnten uns auch bei jedem  $\alpha$  fragen, ob  $H_0$  abgelehnt wird oder nicht.

Wenn der Test bei Signifikanzniveau  $\alpha$  ablehnt, wird er das auch für  $\alpha' > \alpha$  tun.

Es gibt also ein kleinstes  $\alpha$ , bei dem der Test  $H_0$  ablehnt.

## Der p-Wert

ist das kleinste  $\alpha$ , bei dem wir  $H_0$  ablehnen können.

# p-Wert

$T$ : (zufällige) Teststatistik,  $t$ : beobachtete Teststatistik

## Nullhypothese:

$$H_0 : \mu = \mu_0$$

## Zweiseitige Alternative: $\mu \neq \mu_0$

$$\text{p-Wert} = P_0(|T| > |t|)$$

## Einseitige Alternative: $\mu < \mu_0$

$$\text{p-Wert} = P_0(T < t)$$

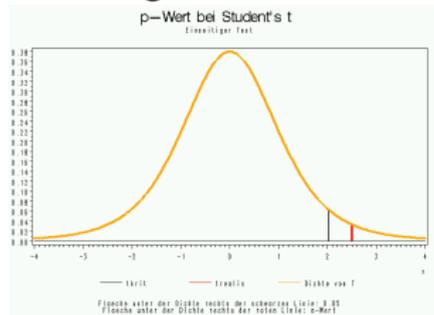
## Einseitige Alternative: $\mu > \mu_0$

$$\text{p-Wert} = P_0(T > t)$$

Der p-Wert heißt auch Überschreitungswahrscheinlichkeit.

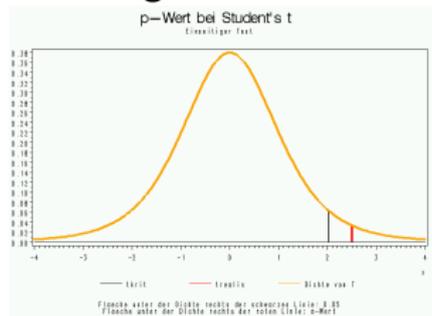
# p-Wert, Illustration

## Einseitiger Test

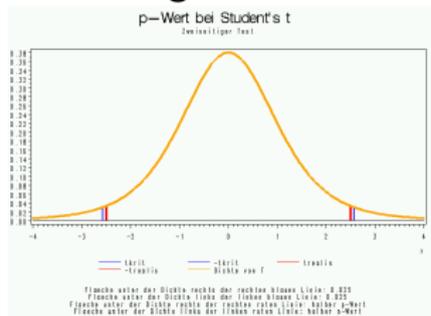


# p-Wert, Illustration

## Einseitiger Test

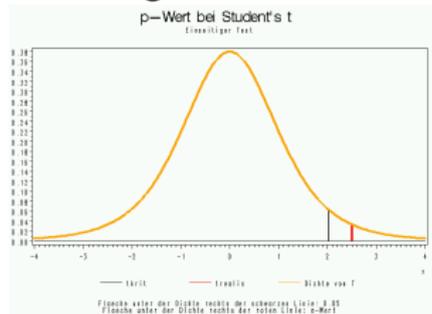


## Zweiseitiger Test

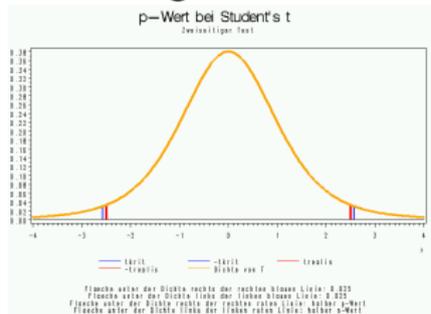


# p-Wert, Illustration

## Einseitiger Test



## Zweiseitiger Test



Fäche unter der Dichte rechts der schwarzen Linie:

0.05

0.025

Fäche unter der Dichte rechts der roten Linie:

p-Wert

halber p-Wert

links entsprechend.

# Bewertung von p-Werten

Der p-Wert ist also, grob, ein Maß für den Grad dafür, dass die Nullhypothese nicht zutrifft.

## (vorsichtige) Interpretation

p-Wert	Grad des Nicht-Zutreffens von $H_0$
$< 0.01$	sehr streng gegen $H_0$
$0.01 \dots 0.05$	streng gegen $H_0$
$0.05 \dots 0.1$	schwach gegen $H_0$
$> 0.1$	wenig oder gar nichts gegen $H_0$

# Bewertung von p-Werten

Der p-Wert ist also, grob, ein Maß für den Grad dafür, dass die Nullhypothese nicht zutrifft.

## (vorsichtige) Interpretation

p-Wert	Grad des Nicht-Zutreffens von $H_0$
$< 0.01$	sehr streng gegen $H_0$
$0.01 \dots 0.05$	streng gegen $H_0$
$0.05 \dots 0.1$	schwach gegen $H_0$
$> 0.1$	wenig oder gar nichts gegen $H_0$

## Warnung:

Ein großer p-Wert heißt noch lange nicht, dass  $H_0$  zutrifft.  $H_0$  kann zutreffen, Der große p-Wert kann aber auch daran liegen, dass der Test niedrige Güte hat!

# Bewertung von p-Werten

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

aus <https://xkcd.com/1478/> (Dank an Herrn Rieger)

# p-Wert und kritischer Wert

Einseitige Alternative,  $t_{krit} = t_{1-\alpha}$

$t \leq t_{krit} \Leftrightarrow \text{p-Wert} \geq \alpha \implies H_0$  angenommen,

$t > t_{krit} \Leftrightarrow \text{p-Wert} < \alpha \implies H_0$  abgelehnt.

# p-Wert und kritischer Wert

Einseitige Alternative,  $t_{krit} = t_{1-\alpha}$

$t \leq t_{krit} \Leftrightarrow \text{p-Wert} \geq \alpha \implies H_0$  angenommen,

$t > t_{krit} \Leftrightarrow \text{p-Wert} < \alpha \implies H_0$  abgelehnt.

Zweiseitige Alternative,  $t_{krit} = t_{1-\alpha/2}$

$|t| \leq t_{krit} \Leftrightarrow \text{p-Wert} \geq \alpha \implies H_0$  angenommen,

$|t| > t_{krit} \Leftrightarrow \text{p-Wert} < \alpha \implies H_0$  abgelehnt.

## Ausgabe bei SAS

Wenn nicht anders vermerkt: zweiseitige p-Werte.

# p-Wert und kritischer Wert

Einseitige Alternative,  $t_{krit} = t_{1-\alpha}$

$t \leq t_{krit} \Leftrightarrow \text{p-Wert} \geq \alpha \implies H_0$  angenommen,

$t > t_{krit} \Leftrightarrow \text{p-Wert} < \alpha \implies H_0$  abgelehnt.

Zweiseitige Alternative,  $t_{krit} = t_{1-\alpha/2}$

$|t| \leq t_{krit} \Leftrightarrow \text{p-Wert} \geq \alpha \implies H_0$  angenommen,

$|t| > t_{krit} \Leftrightarrow \text{p-Wert} < \alpha \implies H_0$  abgelehnt.

## Ausgabe bei SAS

Wenn nicht anders vermerkt: zweiseitige p-Werte.

Der p-Wert ist **nicht** die Wahrscheinlichkeit, dass  $H_0$  zutrifft,

$P(H_0|\text{Daten}) \neq \text{p-Wert}$

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

Nulhypothese      Alternative

a)  $H_0 : \mu \leq \mu_0$      $H_A : \mu > \mu_0$

b)  $H_0 : \mu \geq \mu_0$      $H_A : \mu < \mu_0$

c)  $H_0 : \mu = \mu_0$      $H_A : \mu \neq \mu_0$

## Teststatistik

$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

Nulhypothese      Alternative

a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

c)  $H_0 : \mu = \mu_0$        $H_A : \mu \neq \mu_0$

### Teststatistik

$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$



'Student' in 1908

'Student'

Nulhypothese      Alternative

a)  $H_0 : \mu \leq \mu_0$      $H_A : \mu > \mu_0$

b)  $H_0 : \mu \geq \mu_0$      $H_A : \mu < \mu_0$

c)  $H_0 : \mu = \mu_0$      $H_A : \mu \neq \mu_0$

### Teststatistik

$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

Durchführung des Tests mit

**PROC UNIVARIATE MU0= $\mu_0$**     oder

**PROC TTEST H0= $\mu_0$**



'Student' in 1908

'Student'

## Beispiel: Test\_t1\_Banknote.sas

$\mu_0$	gr	p-Wert		
		Pr>  t		
215	1	0.4258	$> \alpha = 0.05$	nosign
	2	$< 0.0001$	$< \alpha = 0.05$	sign.
214.9	1	0.0784	$> \alpha = 0.05$	nosign.
	2	0.03	$< \alpha = 0.05$	sign.

Das sind also zweiseitige p-Werte (Alternative c)).

Was machen wir bei Alternative a) oder b)?  $\rightarrow$  s.u.

vorgegeben: Fehler 1. Art  $\alpha$  (Signifikanzniveau)

(üblich ist  $\alpha = 0.05$  oder  $\alpha = 0.01$ )

d.h.  $P_{\mu_0}(|T| > t_{krit}) = \alpha$ .

# Verteilung der Teststatistik $T$

Nehmen wir in unserem Beispiel an, die Beobachtungen

$$X_i \sim \mathcal{N}(\mu_0, \sigma^2), \quad , i = 1, \dots, n$$

sind normal und unabhängig, dann hat die (zufällige) Teststatistik  $T$  eine  $t$ -Verteilung (Student's  $t$ ),

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{1}{n-1}\chi_{n-1}^2}} =: t_{n-1}$$

( $t$ -Verteilung mit  $n - 1$  Freiheitsgraden) und

$$t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$$

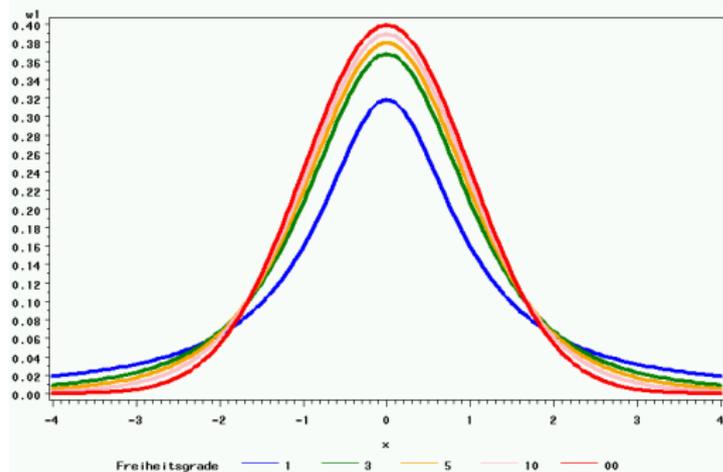
ist  $(1 - \frac{\alpha}{2})$  - Quantil einer  $t$ -Verteilung mit  $n - 1$  Freiheitsgraden.

# Dichtefunktion einer $t$ -Verteilung

mit  $\nu(= n - 1)$  Freiheitsgraden (FG)

$$f_{t_\nu}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \cdot \pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad PDF('T', x, \nu)$$

Dichtefunktion von Student's  $t$



Test\_t\_Dichte.sas

# Einstichprobenproblem

*t*-Test

$$\text{a) } H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$$

⇒ große Werte von

$$T = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

indizieren Gültigkeit von  $H_A$ .

# Einstichprobenproblem

*t*-Test

$$\text{a) } H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$$

⇒ große Werte von

$$T = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

indizieren Gültigkeit von  $H_A$ .

$$\text{b) } H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$$

⇒ kleine Werte von  $T$  indizieren  $H_A$

# Einstichprobenproblem

*t*-Test

a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

⇒ große Werte von

$$T = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

indizieren Gültigkeit von  $H_A$ .

b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

⇒ kleine Werte von  $T$  indizieren  $H_A$

c)  $H_0 : \mu = \mu_0$        $H_A : \mu \neq \mu_0$

⇒  $|T|$  groß indiziert Gültigkeit von  $H_A$ .

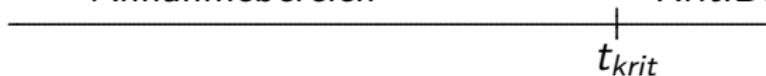
# Hypothesentest, Annahme- und Ablehnungsbereich

- a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmebereich

Krit. Bereich



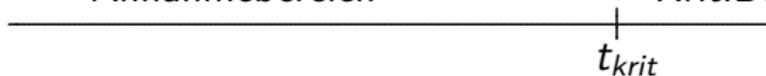
# Hypothesentest, Annahme- und Ablehnungsbereich

- a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmehereich

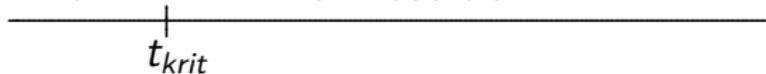
Krit.Bereich



- b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

kleine Werte von  $T$  sprechen für  $H_A$ .

Krit.B.      Annahmehereich



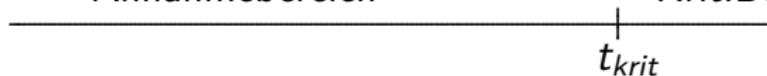
# Hypothesentest, Annahme- und Ablehnungsbereich

- a)  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

große Werte von  $T$  sprechen für  $H_A$ .

Annahmehereich

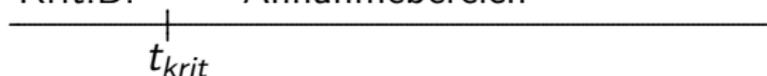
Krit.Bereich



- b)  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

kleine Werte von  $T$  sprechen für  $H_A$ .

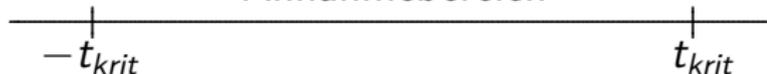
Krit.B.      Annahmehereich



- c)  $H_0 : \mu = \mu_0$        $H_A : \mu \neq \mu_0$

große Werte von  $|T|$  sprechen für  $H_A$ .

Annahmehereich

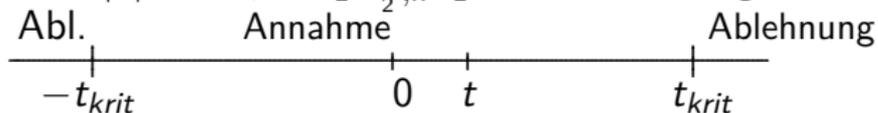


Hypothesentest, sei jetzt  $t$  eine Realisierung von  $T$ .

Zweiseitige Alternative  $H_A : \mu \neq \mu_0$

Wenn  $|t| > t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  abgelehnt.

Wenn  $|t| \leq t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  nicht abgel.

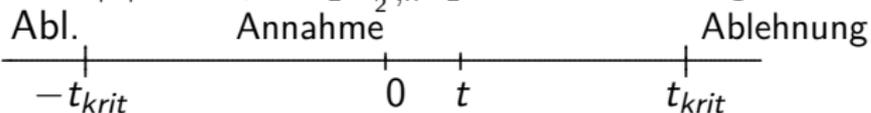


Hypothesentest, sei jetzt  $t$  eine Realisierung von  $T$ .

Zweiseitige Alternative  $H_A : \mu \neq \mu_0$

Wenn  $|t| > t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  abgelehnt.

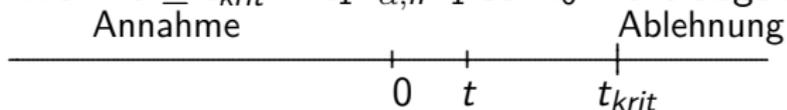
Wenn  $|t| \leq t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  nicht abgel.



Einseitige Alternative  $H_A : \mu > \mu_0$

Wenn  $t > t_{krit} = t_{1-\alpha, n-1}$  so  $H_0$  abgelehnt.

Wenn  $t \leq t_{krit} = t_{1-\alpha, n-1}$  so  $H_0$  nicht abgel.

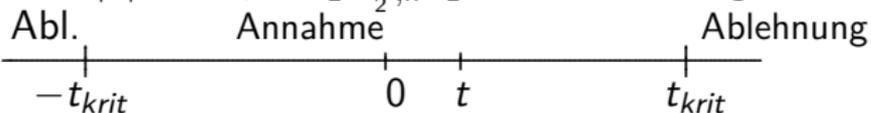


Hypothesentest, sei jetzt  $t$  eine Realisierung von  $T$ .

Zweiseitige Alternative  $H_A : \mu \neq \mu_0$

Wenn  $|t| > t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  abgelehnt.

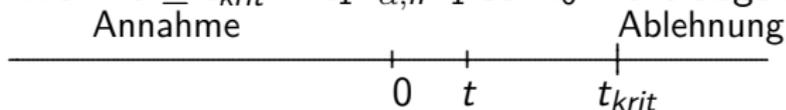
Wenn  $|t| \leq t_{krit} = t_{1-\frac{\alpha}{2}, n-1}$  so  $H_0$  nicht abgel.



Einseitige Alternative  $H_A : \mu > \mu_0$

Wenn  $t > t_{krit} = t_{1-\alpha, n-1}$  so  $H_0$  abgelehnt.

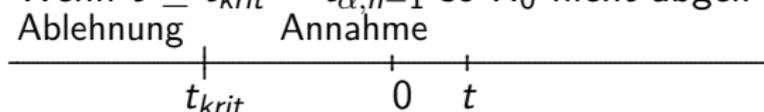
Wenn  $t \leq t_{krit} = t_{1-\alpha, n-1}$  so  $H_0$  nicht abgel.



Einseitige Alternative:  $H_A : \mu < \mu_0$

Wenn  $t < t_{krit} = t_{\alpha, n-1}$  so  $H_0$  abgelehnt.

Wenn  $t \geq t_{krit} = t_{\alpha, n-1}$  so  $H_0$  nicht abgel.



# p-Werte bei einseitigen Alternativen (1)

Erinnerung: Der zweiseitige p-Wert ist:  $P(|T| > t)$ .

$$\begin{aligned}P(|T| > t) &= P((T > t) \vee (-T > t)) \\ &= P((T > t) \vee (T < -t)) \\ &= 2 \cdot P(T > t), \quad t > 0\end{aligned}$$

# p-Werte bei einseitigen Alternativen (1)

Erinnerung: Der zweiseitige p-Wert ist:  $P(|T| > t)$ .

$$\begin{aligned}P(|T| > t) &= P((T > t) \vee (-T > t)) \\ &= P((T > t) \vee (T < -t)) \\ &= 2 \cdot P(T > t), \quad t > 0\end{aligned}$$

$$\begin{aligned}P(T > t) &= P(T < -t) \\ &= 1 - P(T \geq -t) \\ &= 1 - \frac{1}{2}P(|T| > -t), \quad t \leq 0\end{aligned}$$

(Die Verteilung von  $T$  ist stetig und symmetrisch.)

## p-Werte bei einseitigen Alternativen (2)

Fall a)  $H_0 : \mu \leq \mu_0$        $H_a : \mu > \mu_0$

$$\text{p-Wert} = P(T > t) = \begin{cases} \frac{1}{2}P(|T| > t), & \text{falls } t > 0 \\ 1 - \frac{1}{2}P(|T| > -t), & \text{falls } t \leq 0 \end{cases}$$

Ablehnung von  $H_0$  falls  $P(T > t) < \alpha$ .

Die p-Werte von SAS sind zweiseitig, sie sind also (wenn  $t > 0$ )  
durch 2 zu dividieren

(wenn  $t \leq 0$  wird  $H_0$  ohnehin nicht abgelehnt)

**PROC TTEST H0= $\mu_0$**     p-Wert Modifikation nötig

**PROC TTEST H0= $\mu_0$  sides=u**    (u: upper)

keine p-Wert Modifikation nötig

## p-Werte bei einseitigen Alternativen (3)

Fall b)  $H_0 : \mu \geq \mu_0$        $H_a : \mu < \mu_0$

$$\text{p-Wert} = P(T < t) = \begin{cases} \frac{1}{2}P(|T| > |t|), & \text{falls } t \leq 0 \\ 1 - \frac{1}{2}P(|T| > -t), & \text{falls } t > 0 \end{cases}$$

Ablehnung von  $H_0$  falls  $P(T < t) < \alpha$

also wenn  $t < 0$  so SAS-p-Wert durch 2 teilen!

**PROC TTEST**  $H_0=\mu_0$     p-Wert Modifikation nötig

**PROC TTEST**  $H_0=\mu_0$  sides=l    (l: lower)

keine p-Wert Modifikation nötig

Im Fall der zweiseitigen Alternative (c) ist der p-Wert  $P(|T| > t)$

genau das was SAS ausgibt, wir brauchen also nichts zu ändern.

# Zusammenfassung Einstichprobenproblem (1)

## Teststatistik

$$T = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{S} \quad \text{Realisierung } t$$

$$\bar{X} = \frac{1}{n} \sum_i X_i, \quad S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

Zweiseitige Alternative,  $H_0 : \mu = \mu_0$        $H_A : \mu \neq \mu_0$

$$|t| > t_{krit} \quad \Leftrightarrow \quad H_0 \text{ ablehnen}$$

$$\text{p-value} < \alpha \quad \Leftrightarrow \quad H_0 \text{ ablehnen}$$

$$\text{"Pr} > |t| \text{"} < \alpha \quad \Rightarrow \quad H_0 \text{ ablehnen}$$

# Zusammenfassung Einstichprobenproblem (2)

Einseitige Alternative,  $H_0 : \mu \leq \mu_0$        $H_A : \mu > \mu_0$

$t > 0$  und  $\frac{\text{p-value}}{2} < \alpha \Leftrightarrow H_0$  ablehnen

Einseitige Alternative,  $H_0 : \mu \geq \mu_0$        $H_A : \mu < \mu_0$

$t < 0$  und  $\frac{\text{p-value}}{2} < \alpha \Leftrightarrow H_0$  ablehnen

# Konfidenzbereiche (1)

am Beispiel des  $t$ -Tests

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \sqrt{n} \cdot \frac{\bar{X} - \mu}{S} \sim t_{n-1} \text{ wenn } \mu \text{ der wahre (Lokations-)}$$

Parameter ist.  $\Rightarrow$

$$P\left(\underbrace{-t_{1-\frac{\alpha}{2}, n-1} \leq \sqrt{n} \cdot \frac{\bar{X} - \mu}{S} \leq t_{1-\frac{\alpha}{2}, n-1}}_{(*)}\right) = 1 - \alpha$$

# Konfidenzbereiche (1)

am Beispiel des  $t$ -Tests

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \sqrt{n} \cdot \frac{\bar{X} - \mu}{s} \sim t_{n-1} \text{ wenn } \mu \text{ der wahre (Lokations-)}$$

Parameter ist.  $\Rightarrow$

$$P\left(\underbrace{-t_{1-\frac{\alpha}{2}, n-1} \leq \sqrt{n} \cdot \frac{\bar{X} - \mu}{s} \leq t_{1-\frac{\alpha}{2}, n-1}}_{(*)}\right) = 1 - \alpha$$

Die Ungleichungen sind äquivalent zu

$$\begin{aligned} (*) &\Leftrightarrow -\frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \leq \bar{X} - \mu \leq \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \\ &\Leftrightarrow -\bar{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \leq -\mu \leq -\bar{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \\ &\Leftrightarrow \bar{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \geq \mu \geq \bar{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \\ &\Leftrightarrow \bar{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \leq \mu \leq \bar{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \end{aligned}$$

## Konfidenzbereiche (2)

$(1 - \alpha)$  Konfidenzintervall für den (unbekannten) Parameter  $\mu$

$$\left[ \bar{X} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}, n-1}, \bar{X} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}, n-1} \right]$$

**PROC TTEST** ALPHA=Wert

**PROC UNIVARIATE** ALPHA=Wert CIBASIC

ALPHA: Konfidenzniveau (=Signifikanzniveau)

CIBASIC: Konfidenzintervalle für  $\mu, \sigma^2, \sigma$  basierend auf Normalverteilung

CIPCTLDF: verteilungsfreie Konfidenzintervalle (basierend auf empirischen Quantilen)

# Konfidenzbereiche (3)

Beispiel

## Test\_t1\_Banknote

$(1 - \alpha)$ -Konfidenzintervalle für den Lageparameter  $\mu = \mathbf{E}'\text{laenge}'$ :

	echt		gefälscht	
$\alpha = 0.01$	214.87	215.07	214.73	214.92
$\alpha = 0.05$	214.89	215.05	214.75	214.89
$\alpha = 0.05$	214.9	215.1	214.7	214.9
nichtparam. KI (für Median)				

**PROC TTEST ALPHA=**Wert

# Konfidenzintervalle für $\sigma^2$

bei Normalverteilung

$$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2), \text{ unabhängig} \quad \Rightarrow \quad (n-1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Dichte einer  $\chi_{\nu}^2$ -Verteilung

$$f_{\chi_{\nu}^2}(x) = \begin{cases} \frac{1}{2^{\nu/2} \Gamma(\frac{\nu}{2})} e^{-x/2} x^{\nu/2-1} & \text{falls } x \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

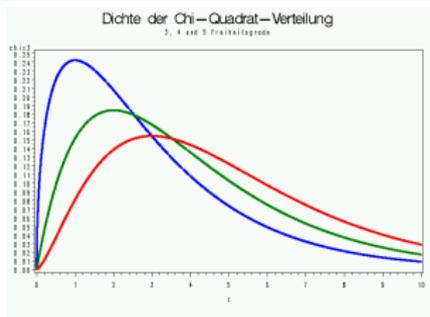
# Konfidenzintervalle für $\sigma^2$

bei Normalverteilung

$$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2), \text{ unabhängig} \Rightarrow (n-1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Dichte einer  $\chi_\nu^2$ -Verteilung

$$f_{\chi_\nu^2}(x) = \begin{cases} \frac{1}{2^{\nu/2} \Gamma(\frac{\nu}{2})} e^{-x/2} x^{\nu/2-1} & \text{falls } x \geq 0 \\ 0 & \text{sonst.} \end{cases}$$



Test\_Chi2\_Dichte

# Konfidenzintervall für $\sigma^2$ (2)

bei Normalverteilung

$$P\left(\chi_{\alpha/2, n-1}^2 \leq (n-1) \frac{s^2}{\sigma^2} \leq \chi_{1-\alpha/2, n-1}^2\right) = 1 - \alpha$$

auflösen nach  $\sigma^2$ :

$$\begin{aligned} 1 - \alpha &= P\left(\chi_{\alpha/2, n-1}^2 \leq (n-1) \frac{s^2}{\sigma^2} \leq \chi_{1-\alpha/2, n-1}^2\right) \\ &= P\left(\frac{1}{\chi_{1-\alpha/2, n-1}^2} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{\chi_{\alpha/2, n-1}^2}\right) \\ &= P\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2}\right) \end{aligned}$$

# Konfidenzintervall für $\sigma^2$ (3)

nur bei Normalverteilung!

## Konfidenzintervall

(Vertrauensintervall) für den (unbekannten) Parameter  $\sigma^2$

$$\left[ \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \right]$$

**PROC TTEST**

**PROC UNIVARIATE ALPHA CIBASIC CIPCTLDF**

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

# verbundene Stichproben

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$$

$$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$$

- Gewicht einer Person zu den Zeitpunkten  $t_1, t_2$ .
- Banknoten (oben- unten, links - rechts)
- Patient nimmt Medikament 1 und 2
- Kreuz- und selbstbefruchtete Pflanzen

Test\_t2\_Banknote    Test\_t2\_Darwin

# Vergleich zweier abhängiger Gruppen

Folgende Möglichkeiten:

a) Transformation  $Z := X_1 - X_2$  und testen auf  $\mu = 0$

PROC UNIVARIATE; VAR Z; RUN;   oder

PROC TTEST H0=0; VAR Z; RUN;

b) Mit der Prozedur TTEST:

**PROC TTEST;**

**PAIRED** X1\*X2;

**RUN;**

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

# unverbundene Stichproben

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 < \mu_2 \quad H_1 : \mu_1 \geq \mu_2$$

$$H_0 : \mu_1 > \mu_2 \quad H_1 : \mu_1 \leq \mu_2$$

- Tibetische Schädel (Sikkim - Kham)
- Wasserhärte (Nord - Süd)
- Klinikaufenthalt (Klinik1 - Klinik2)
- Banknoten (echt - gefälscht)

Test\_t2\_Tibetan

Test\_t2\_Heroin

Test\_t2\_Banknote

## Vergleich zweier unabhängiger Gruppen (2)

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Fall 1: Varianzen  $\sigma_1^2, \sigma_2^2$  sind gleich

Fall 2: Varianzen  $\sigma_1^2, \sigma_2^2$  sind verschieden

## Vergleich zweier unabhängiger Gruppen (2)

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Fall 1: Varianzen  $\sigma_1^2, \sigma_2^2$  sind gleich

Fall 2: Varianzen  $\sigma_1^2, \sigma_2^2$  sind verschieden

Fall 1:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma^2}$$

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

Fall 1: Varianzen  $\sigma_1^2, \sigma_2^2$  sind gleich

Fall 2: Varianzen  $\sigma_1^2, \sigma_2^2$  sind verschieden

Fall 1. Seien  $n, m$ : Umfänge von Stichprobe 1 und 2.

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}}$$

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2, \quad S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (X_{2i} - \bar{X}_2)^2$$

Bem.: Bisher haben wir die erwartungstreue empirische Varianz mit  $s^2$  (Kleinbuchstabe) bezeichnet. Da wir hier betonen, dass es Zufallsvariablen sind, verwenden wir jetzt Großbuchstaben, also  $S_1^2$  und  $S_2^2$ .

# Erläuterung des Quotienten $T$

$$X_1 \sim \mathcal{N}(\mu_1, \sigma^2), X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$$

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \sigma^2 \cdot \frac{1}{n}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \sigma^2 \cdot \frac{1}{m}\right)$$

$$\frac{(n-1)}{\sigma^2} \cdot S_1^2 \sim \chi_{n-1}^2, \quad \frac{(m-1)}{\sigma^2} \cdot S_2^2 \sim \chi_{m-1}^2$$

# Erläuterung des Quotienten $T$

$$X_1 \sim \mathcal{N}(\mu_1, \sigma^2), X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$$

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \sigma^2 \cdot \frac{1}{n}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \sigma^2 \cdot \frac{1}{m}\right)$$

$$\frac{(n-1)}{\sigma^2} \cdot S_1^2 \sim \chi_{n-1}^2, \quad \frac{(m-1)}{\sigma^2} \cdot S_2^2 \sim \chi_{m-1}^2$$

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

$$\frac{1}{\sigma^2} \left( (n-1) \cdot S_1^2 + (m-1) \cdot S_2^2 \right) \sim \chi_{n+m-2}^2$$

$$T \sim t_{n+m-2} \quad \text{unter } H_0 (\mu_1 = \mu_2)$$

**Bem.:** Bei Normalverteilung sind Zähler und Nenner stochastisch unabhängig!

## Vergleich zweier unabhängiger Gruppen (4)

$T$  ist eine Zufallsgröße!

Werte von  $T$  werden mit gewissen Wahrscheinlichkeiten angenommen!

Die Wahrscheinlichkeit dafür, daß  $T$  sehr große Werte annimmt (wenn  $H_0$  richtig ist) ist also sehr klein.

Sei jetzt  $t$  eine Realisierung von  $T$  (also der Wert, der bei Ausrechnen anhand der gegebenen Daten entsteht).

Wenn jetzt  $t$  sehr groß,  $|t| \in K$  (kritischer Bereich)  
(aber die Wahrscheinlichkeit dafür ist sehr klein, wenn  $H_0$  richtig ist)  
 $\Rightarrow H_0$  ablehnen.

# Vergleich zweier unabhängiger Gruppen

Fall 2: Varianzen ungleich

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

# Vergleich zweier unabhängiger Gruppen

## Fall 2: Varianzen ungleich

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

$T \sim t_\nu$  approximativ. Die Zahl  $\nu$  der Freiheitsgrade wird auch approximativ berechnet. (Welch-Test, 1937)

SAS bietet Tests für beide Fälle (gleiche, ungleiche Varianzen) an. Satterthwaite-Approximation (1946).

**PROC TTEST;**

**CLASS** Klassifikationsvariable;

**VAR** auszuwertende Variable(n); **RUN;**

# Vergleich zweier unabhängiger Gruppen

## Welchen Test soll man nehmen?

- Aus Vorinformation ist vielleicht bekannt, ob man gleiche Varianzen annehmen kann.
- Man könnte einen Test auf gleiche Varianzen vorschalten

# Vergleich zweier unabhängiger Gruppen

## Welchen Test soll man nehmen?

- Aus Vorinformation ist vielleicht bekannt, ob man gleiche Varianzen annehmen kann.
- Man könnte einen Test auf gleiche Varianzen vorschalten

## Problem: 2 stufiger Test

Wird das Signifikanzniveau eingehalten??

# Vergleich zweier unabhängiger Gruppen

## Welchen Test soll man nehmen?

- Aus Vorinformation ist vielleicht bekannt, ob man gleiche Varianzen annehmen kann.
- Man könnte einen Test auf gleiche Varianzen vorschalten

## Problem: 2 stufiger Test

Wird das Signifikanzniveau eingehalten??

## Vorschlag

**gleich den t-Test für ungleiche Varianzen nehmen**

ist einigermaßen robust gegen Abweichungen von der Normalverteilung, aber nicht gegen Ausreißer

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

Voraussetzung: Normalverteilung!!

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{S_1^2}{S_2^2} \sim F_{n-1, m-1}$$

(Fisher-)  $F$ - Verteilung mit  $(n - 1, m - 1)$  Freiheitsgraden.

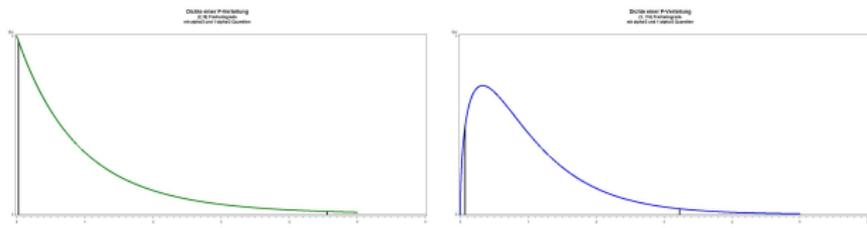
$F$  ist Quotient zweier unabhängiger  $\chi^2$ -verteilter Zufallsgrößen.

$H_0$  ablehnen, falls

$$\frac{s_1^2}{s_2^2} < F_{\frac{\alpha}{2}, n-1, m-1} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2}, n-1, m-1}$$

# Test auf Gleichheit der Varianzen (bei Normalverteilung)

## F-Test



$$F_{\frac{\alpha}{2}, n-1, m-1} = \frac{1}{F_{1-\frac{\alpha}{2}, m-1, n-1}}$$

(beachten: Freiheitsgrade vertauschen sich)

⇒  $H_0$  ablehnen, falls

$$\frac{s_1^2}{s_2^2} < \frac{1}{F_{1-\frac{\alpha}{2}, m-1, n-1}} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2}, n-1, m-1} \quad \Leftrightarrow$$

$$\frac{s_2^2}{s_1^2} > F_{1-\frac{\alpha}{2}, m-1, n-1} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2}, n-1, m-1}$$

# Test auf Gleichheit der Varianzen

F-Test, prakt. Durchführung

$$s_M^2 := \max(s_1^2, s_2^2) \quad s_m^2 := \min(s_1^2, s_2^2)$$

$n_M, n_m$ : die entsprechenden Stichprobenumfänge

$$\Rightarrow H_0 \text{ ablehnen, falls } \frac{s_M^2}{s_m^2} > F_{1-\frac{\alpha}{2}, n_M-1, n_m-1}.$$

Formulierung mit  $p$ -Werten

$$\Rightarrow H_0 \text{ ablehnen, falls } p\text{-Wert} = P\left(F > \frac{s_M^2}{s_m^2}\right) < \frac{\alpha}{2}$$

$$F \sim F_{n_M-1, n_m-1}$$

Test\_F\_Dichte

# Zweistichprobenproblem

## Output der Prozedur TTEST

- Konfidenzintervalle für  $\mu_1, \mu_2$  und für  $\mu_1 - \mu_2$   
Für die ersten beiden siehe Abschnitt 5.2  
Konfidenzintervalle für  $\mu = \mu_1 - \mu_2$  bekommt man analog zum Einstichprobenfall
- Tabelle der durchgeführten  $t$ -Tests
  - für gleiche Varianzen (pooled)
  - für ungleiche Varianzen (Satterthwaite)
- F-Test zum Vergleich der Varianzen (bitte ignorieren)

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

# Ein- und Zweistichprobenproblem

## Anmerkungen (1)

- Der  $F$ -Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme  
⇒ mit größter Vorsicht genießen.

# Ein- und Zweistichprobenproblem

## Anmerkungen (1)

- Der  $F$ -Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme  
⇒ mit größter Vorsicht genießen.
- Der Einstichproben-  $t$ -Test ist nicht robust!

# Ein- und Zweistichprobenproblem

## Anmerkungen (1)

- Der  $F$ -Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme  
⇒ mit größter Vorsicht genießen.
- Der Einstichproben-  $t$ -Test ist nicht robust!
- Der Zweistichproben  $t$ -Test ist etwas robuster als der  $t$ -Test im Einstichprobenproblem

# Ein- und Zweistichprobenproblem

## Anmerkungen (1)

- Der  $F$ -Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme  
⇒ mit größter Vorsicht genießen.
- Der Einstichproben-  $t$ -Test ist nicht robust!
- Der Zweistichproben  $t$ -Test ist etwas robuster als der  $t$ -Test im Einstichprobenproblem
- Ausreißer können extremen Einfluss haben (ÜA).

# Ein- und Zweistichprobenproblem

## Anmerkungen (1)

- Der  $F$ -Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme  
⇒ mit größter Vorsicht genießen.
- Der Einstichproben-  $t$ -Test ist nicht robust!
- Der Zweistichproben  $t$ -Test ist etwas robuster als der  $t$ -Test im Einstichprobenproblem
- Ausreißer können extremen Einfluss haben (ÜA).
- Wenn Gleichheit der Varianzen unklar ⇒  
 **$t$ -Test mit ungleichen Varianzen nehmen.**  
(ist bei gleichen Varianzen nur ganz wenig weniger effizient)

# Ein- und Zweistichprobenproblem

## Anmerkungen (2)

- Besser nicht auf das Ergebnis des  $F$ -Tests verlassen.  
(Problematik: 2-Stufentest, Nicht-Robustheit).
- Es gibt robustere Skalentests  
⇒ Levene Test und Brown-Forsythe Test.

# Inhalt

- Statistische Tests: Einführung und Übersicht
- Einstichprobenproblem
- Vergleich zweier abhängiger Gruppen
- Vergleich zweier unabhängiger Gruppen
- Test auf Gleichheit der Varianzen (1)
- Anmerkungen
- Test auf Gleichheit der Varianzen (2)

Seien  $X_1, \dots, X_n$  und  $Y_1, \dots, Y_m$  unabhängige Beobachtungen.

### Levene-Test

Bilden die Werte

$$X_j^* := |X_j - \bar{X}|$$

$$Y_j^* := |Y_j - \bar{Y}|$$

Skalenunterschiede in  $(X, Y)$  spiegeln sich jetzt in Lageunterschieden in  $(X^*, Y^*)$  wieder.

Mit den “neuen Beobachtungen” wird jetzt ein  $t$ -Test durchgeführt.

Die  $t$ -Verteilung der entsprechenden Teststatistik gilt nur approximativ.

# Test auf Gleichheit der Varianzen

## Brown-Forsythe Test

Analog zum Levene-Test, nur hier bilden wir die Werte

$$X_j^* := |X_j - \text{med}(X_1, \dots, X_n)|$$

$$Y_j^* := |Y_j - \text{med}(Y_1, \dots, Y_m)|$$

Beide Tests, Levene und Brown-Forsythe, sind (einigermaßen) robust gegen Abweichungen von der Normalverteilung.

# Test auf Gleichheit der Varianzen

Syntax

```
PROC ANOVA;
```

```
  CLASS Klasse;
```

```
  MODEL var=Klasse;
```

```
  MEANS Klasse / HOVTEST=Levene (TYPE=ABS);
```

```
  MEANS Klasse / HOVTEST=BF;
```

```
RUN;
```

Test\_t2\_Banknote

# Inhalt

- Vergleich von  $k$  unabhängigen Gruppen
- Multiple Vergleiche
- Vergleich von  $k$  abhängigen Gruppen
- Weitere Varianzanalyse-Modelle

# Einfaktorielle, einfache Varianzanalyse

A: Faktor (Gruppenvariable) mit  $k$  Stufen (Faktorstufen)

## Modell

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1 \dots k, j = 1 \dots n_i$$

- $\mu$ : Gesamterwartungswert
- $\alpha_i$ : Effekt der  $i$ -ten Stufe von A
- $\epsilon_{ij}$ : Fehler,  $\epsilon_{ij} \sim (0, \sigma^2)$ , unabhängig
- $Y_{ij}$ :  $j$ -te Beobachtung der  $i$ -ten Faktorstufe
- $\sum_{i=1}^k \alpha_i = 0$  Parametrisierungsbedingung

Zum Testen brauchen wir (vorläufig) die Annahme  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

# Einfache Varianzanalyse

## Testproblem

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k$$

$$H_1 : \alpha_i \neq \alpha_l \text{ (für ein } i \neq l)$$

Im Fall  $k = 2$  führt dieses Testproblem auf das Zweistichprobenproblem ( $\rightarrow$  t-Test).

Output der Maschinen gleich?

Klausurergebnisse unterschiedlich?

Mageninhalt der Eidechsen gleich?

Cortisolgehalt unterschiedlich?

ANOVA\_Maschinen

PI12erg

GLM\_Eidechsen

GLM\_Cortisol

# Varianzanalyse

Varianzanalyse macht eine Streuungserlegung:

Quadratsumme = Quadratsumme + Quadratsumme

zwischen                      innerhalb

Gesamtfehler                den Faktorstufen                der Faktorstufen

---

SST = SSB + SSW | (SSE)

(Total)                      (Between)                      (Within)      (Error)

$$N = \sum_{i=1}^k n_i$$

$$\bar{Y}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{i,j}$$

# Einfache Varianzanalyse

**Satz:** Es gilt

$$SSB + SSW = SST$$

wobei

$$SSB = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 \quad (\underline{\text{B}}\text{etween})$$

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad (\underline{\text{W}}\text{ithin})$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2. \quad (\underline{\text{T}}\text{otal})$$

# Einfache Varianzanalyse

**Satz:**  $SSB + SSW = SST$

Beweis:

$$SSB = \sum_i n_i \bar{Y}_i^2 - 2 \cdot N \cdot \bar{Y}^2 + \bar{Y}^2 \cdot N$$

$$SSW = \sum_{i,j} Y_{ij}^2 - 2 \cdot \sum_i n_i \bar{Y}_i^2 + \sum_i n_i \bar{Y}_i^2$$

# Einfache Varianzanalyse

**Satz:**  $SSB + SSW = SST$

Beweis:

$$SSB = \sum_i n_i \bar{Y}_i^2 - 2 \cdot N \cdot \bar{Y}^2 + \bar{Y}^2 \cdot N$$

$$SSW = \sum_{i,j} Y_{ij}^2 - 2 \cdot \sum_i n_i \bar{Y}_i^2 + \sum_i n_i \bar{Y}_i^2$$

$$SSB + SSW =$$

$$= \sum_{i,j} Y_{ij}^2 + \sum_i n_i \bar{Y}_i^2 - N \cdot \bar{Y}^2 - \sum_i n_i \bar{Y}_i^2$$

$$= \sum_{i,j} Y_{ij}^2 - N \cdot \bar{Y}^2 = \sum_j \sum_i (Y_{ij} - \bar{Y})^2 = SST \quad \blacksquare$$

# Varianzanalyse

Programm

```
PROC ANOVA;
```

```
  CLASS A; /*A: Faktor */
```

```
  MODEL var=A;
```

```
  MEANS A /
```

```
    HOVTEST=Levene (TYPE=ABS);
```

```
    HOVTEST=BF;
```

```
  MEANS OUT=SAS-Ausgabedatei;
```

```
RUN;
```

# Varianzanalyse, Programm

```
PROC ANOVA;           PROC GLM;  
  CLASS A; /*A: Faktor */  
  MODEL var=A;  
  MEANS A /  
    HOVTEST=Levene (TYPE=ABS);  
    HOVTEST=BF;  
  MEANS OUT=SAS-Ausgabedatei;  
RUN;
```

- ANOVA: schneller
- GLM: zusätzliche Auswertungen möglich, z.B. Ausgabe der Residuen ( $\hat{Y}_i - Y_i$ )
- HOVTEST: Test auf Varianzhomogenität (nur bei ANOVA!)

# Varianzanalyse-Tabelle

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Squares	F-value	$Pr > F$
MODEL	$k-1$	SSB(M)	MSB	$\frac{MSB}{MSE}$	p-Wert
ERROR	$N-k$	SSW(E)	MSE		
Total	$N-1$	SST			

$$MSB = \frac{SSB}{k-1}, \quad MSE = \frac{SSW}{N-k}, \quad F = \frac{MSB}{MSE} \sim F_{k-1, N-k}$$

$$H_0 : \alpha_1 = \dots = \alpha_k \quad H_1 : \exists(i, j) : \alpha_i \neq \alpha_j$$

# Varianzanalyse-Tabelle

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Squares	F-value	$Pr > F$
MODEL	$k-1$	SSB(M)	MSB	$\frac{MSB}{MSE}$	p-Wert
ERROR	$N-k$	SSW(E)	MSE		
Total	$N-1$	SST			

$$MSB = \frac{SSB}{k-1}, \quad MSE = \frac{SSW}{N-k}, \quad F = \frac{MSB}{MSE} \sim F_{k-1, N-k}$$

$$H_0 : \alpha_1 = \dots = \alpha_k \quad H_1 : \exists(i, j) : \alpha_i \neq \alpha_j$$

Wenn  $H_0$  richtig ist und die Beobachtungen normalverteilt sind, so hat  $F = \frac{MSB}{MSE}$  eine  $F$ -Verteilung mit  $(k-1, N-k)$  Freiheitsgraden.

## Einfache Varianzanalyse (2)

$H_0$  wird getestet mit

$$F = \frac{MSB}{MSE} = \frac{\text{geschätzte Varianz zwischen den Gruppen}}{\text{geschätzte Varianz innerhalb der Gruppen}}$$
$$= \frac{N - k}{k - 1} \cdot \frac{SSB}{SSW} = \frac{N - k}{k - 1} \cdot \frac{SST - SSW}{SSW}$$

$F$  groß,  $F > F_{1-\alpha, k-1, N-k} \Leftrightarrow H_0$  abgelehnt

## Einfache Varianzanalyse (2)

$H_0$  wird getestet mit

$$F = \frac{MSB}{MSE} = \frac{\text{geschätzte Varianz zwischen den Gruppen}}{\text{geschätzte Varianz innerhalb der Gruppen}}$$
$$= \frac{N - k}{k - 1} \cdot \frac{SSB}{SSW} = \frac{N - k}{k - 1} \cdot \frac{SST - SSW}{SSW}$$

$F$  groß,  $F > F_{1-\alpha, k-1, N-k} \Leftrightarrow H_0$  abgelehnt

### Bestimmtheitsmaß

$$R^2 := \frac{SSB}{SST} = \frac{SST - SSW}{SST} = 1 - \frac{SSW}{SST}$$

Der Anteil der Fehlerquadratsummen, der durch das Modell bestimmt wird, heißt Bestimmtheitsmaß

## Einfache Varianzanalyse (3)

Offenbar:  $0 \leq R^2 \leq 1$ .

$$F = \frac{MSB}{MSE} = \frac{N - k}{k - 1} \cdot \frac{SSB}{SST} \cdot \frac{SST}{SSW} = \frac{N - k}{k - 1} \cdot \frac{R^2}{1 - R^2}$$

$$R^2 \rightarrow 0 \implies F \rightarrow 0$$

$$R^2 \rightarrow 1 \implies F \rightarrow \infty.$$

## Einfache Varianzanalyse (3)

Offenbar:  $0 \leq R^2 \leq 1$ .

$$F = \frac{MSB}{MSE} = \frac{N - k}{k - 1} \cdot \frac{SSB}{SST} \cdot \frac{SST}{SSW} = \frac{N - k}{k - 1} \cdot \frac{R^2}{1 - R^2}$$

$$R^2 \rightarrow 0 \implies F \rightarrow 0$$

$$R^2 \rightarrow 1 \implies F \rightarrow \infty.$$

Schätzung der Modellstandardabweichung  $\sigma$

$$\text{RootMSE} = \sqrt{MSE} = \sqrt{\frac{1}{N-k} SSE}$$

Variationskoeffizient

$$CV = \frac{100 \cdot \text{RootMSE}}{\bar{Y}}$$

# Einfache Varianzanalyse

## Anmerkungen (1)

- Der  $F$ -Test in der Varianzanalyse ist (einigermaßen) robust gegenüber Abweichungen von der Normalverteilungsannahme

# Einfache Varianzanalyse

## Anmerkungen (1)

- Der  $F$ -Test in der Varianzanalyse ist (einigermaßen) robust gegenüber Abweichungen von der Normalverteilungsannahme
- Wenn man die Prozedur GLM verwendet, dann kann man die sogenannten Residuen

$$\hat{\epsilon}_{ij} = Y_{ij} - \hat{\alpha}_i - \hat{\mu}$$

abspeichern

(Option **RESIDUAL** im **OUTPUT**-Statement)

und später auf Normalität testen.

(**PROC UNIVARIATE NORMAL**)

# Varianzanalyse

## Anmerkungen (2)

- $F$ -Test verlangt auch Varianzhomogenität  
Daten balanziert (gleiche Stichprobenumfänge)  
→ Abweichungen nicht so schwerwiegend.
- Wenn die Varianzen verschieden sind, kann die Welch-Modifikation verwendet werden:  
`MEANS Var/ WELCH;`

# Einfache Varianzanalyse

Test auf Varianzhomogenität

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \exists(i, l) : \sigma_i^2 \neq \sigma_l^2$$

## Levene Test (1960)

HOVTEST = LEVENE im MEANS-Statement

$$Z_{ij}^* = |Y_{ij} - \bar{Y}_{i.}|$$

## Brown-Forsythe-Test (1974)

HOVTEST = BF

$$Z_{ij}^* = |Y_{ij} - med Y_i|$$

# Einfache Varianzanalyse

Test auf Varianzhomogenität (2)

Mit diesen neuen ZV wird eine Varianzanalyse durchgeführt.

$$W = \frac{\frac{1}{k-1} \sum n_i (\bar{Z}_{i.}^* - \bar{Z}^*)^2}{\frac{1}{N-k} \sum_{i,j} (Z_{ij}^* - \bar{Z}_{i.}^*)^2} \sim F_{k-1, N-k}.$$

GLM\_Cortisol

# Geometrische Veranschaulichung

$$\mathbf{Y} = (Y_{11}, \dots, Y_{kn_k}) \quad \text{Dimension } N$$

$$\hat{\mathbf{Y}} = (\underbrace{\bar{Y}_1, \dots, \bar{Y}_1}_{n_1 \text{ mal}}, \dots, \underbrace{\bar{Y}_k, \dots, \bar{Y}_k}_{n_k \text{ mal}})$$

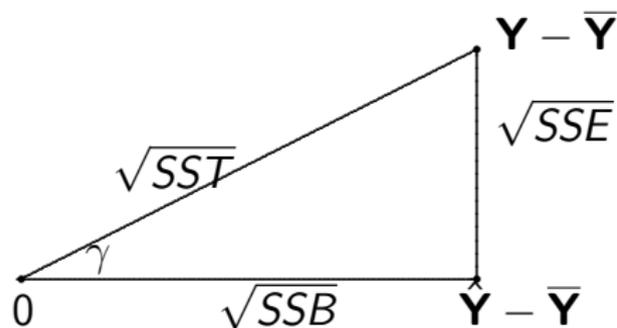
$$\bar{\mathbf{Y}} = (\underbrace{\bar{Y}, \dots, \bar{Y}}_{N \text{ mal}}), \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$$

# Geometrische Veranschaulichung

$$\mathbf{Y} = (Y_{11}, \dots, Y_{kn_k}) \quad \text{Dimension } N$$

$$\hat{\mathbf{Y}} = (\underbrace{\bar{Y}_1, \dots, \bar{Y}_1}_{n_1 \text{ mal}}, \dots, \underbrace{\bar{Y}_k, \dots, \bar{Y}_k}_{n_k \text{ mal}})$$

$$\bar{\mathbf{Y}} = (\underbrace{\bar{Y}, \dots, \bar{Y}}_{N \text{ mal}}), \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$$



$$SSB + SSW = SST$$

$$R^2 = \cos^2 \gamma$$

$$\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$$

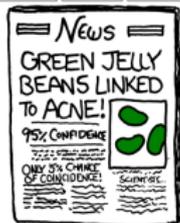
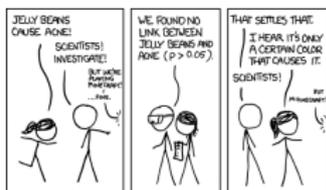
# Inhalt

- Vergleich von  $k$  unabhängigen Gruppen
- Multiple Vergleiche
- Vergleich von  $k$  abhängigen Gruppen
- Weitere Varianzanalyse-Modelle

Problemstellung:  $H_0$  abgelehnt, aber zwischen welchen Faktorstufen liegt der Unterschied?

- Idee: Alle Paarvergleiche machen.
- Problem: Wenn wir das Signifikanzniveau  $\alpha (= 0.05)$  so lassen, wird das Testniveau nicht eingehalten!
- Veranschaulichung: Bei 20 gleichzeitigen Tests können wir  $20 \cdot \alpha = 1$  Ablehnung erwarten, auch wenn  $H_0$  richtig ist.

# Illustration zum Problem vieler gleichzeitiger Tests



aus: <https://xkcd.com/882/>  
(Dank an Herrn Tobias Rieger)

# Multiple Vergleiche

Lösungsmöglichkeiten (1)

## Option BON im MEANS Statement

Signifikanzniveau für die gleichzeitigen Tests wird herabgesetzt auf

$$\frac{\alpha_{nom}}{\binom{k}{2}},$$

bei  $k = 4$  und  $\alpha_{nom} = 0.05$  wäre das  $\frac{\alpha_{nom}}{\binom{4}{2}} = \frac{0.05}{6}$ .

Begründung: Bonferroni-Ungleichung.

# Multiple Vergleiche

Lösungsmöglichkeiten (1)

## Option BON im MEANS Statement

Signifikanzniveau für die gleichzeitigen Tests wird herabgesetzt auf

$$\frac{\alpha_{nom}}{\binom{k}{2}},$$

bei  $k = 4$  und  $\alpha_{nom} = 0.05$  wäre das  $\frac{\alpha_{nom}}{\binom{4}{2}} = \frac{0.05}{6}$ .

Begründung: Bonferroni-Ungleichung.

$A_i$ : Ereignis,  $H_{0i}$  ( $i$ -ter Paarvergleich) abgelehnt.

$$\underbrace{P_0\left(\bigcup A_i\right)}_{\text{Wkt, } H_{0i} \geq 1 \text{ mal abgelehnt}} \leq \sum_{i=1}^M P(A_i) \leq M \cdot \frac{\alpha}{M} = \alpha$$

M: Anzahl der Paarvergleiche.

## Multiple Vergleiche, Lösungsmöglichkeiten (2)

### Option TUKEY im MEANS Statement

Bilden die  $\bar{Y}_j$  und die Spannweite dazu  $w = \max_{i,j} |\bar{Y}_i - \bar{Y}_j|$ .

Dazu kommt noch die empirische Standardabweichung  $s$ .

$$t_{\max} = \frac{w}{s}$$

die sogenannte *studentisierte Spannweite*.

Diese hat (wenn die  $Y_i \sim \mathcal{N}$ ) eine (dem SAS-Programmierer) wohlbekannte Verteilung, und entsprechende Quantile und kritische Werte. Damit erhalten wir simultane Konfidenzintervalle für alle Paardifferenzen  $\mu_i - \mu_j$ . Liegt 0 nicht darin, so wird  $H_{0,ij} : \mu_i = \mu_j$  abgelehnt zugunsten von  $H_{A,ij} : \mu_i \neq \mu_j$ .

**Bem.:** Es gibt eine Fülle weiterer Varianten.

# Inhalt

- Vergleich von  $k$  unabhängigen Gruppen
- Multiple Vergleiche
- Vergleich von  $k$  abhängigen Gruppen
- Weitere Varianzanalyse-Modelle

## Zwei-faktorielle Varianzanalyse

### Modell:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \text{unabhängig}$$

$i = 1, \dots, a, j = 1, \dots, b.$  (eine Beobachtung je Zelle)

Das Modell ist überparametrisiert, deswegen Bedingung:

$$\sum_{i=1}^a \alpha_i = 0, \quad \sum_{j=1}^b \beta_j = 0.$$

## Zwei-faktorielle Varianzanalyse

### Modell:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \text{unabhängig}$$

$i = 1, \dots, a, j = 1, \dots, b.$  (eine Beobachtung je Zelle)

Das Modell ist überparametrisiert, deswegen Bedingung:

$$\sum_{i=1}^a \alpha_i = 0, \quad \sum_{j=1}^b \beta_j = 0.$$

Folgende Hypothesen sind zu testen:

$$H_{0a} : \alpha_1 = \dots = \alpha_a = 0 \quad \text{gegen} \quad H_{1a} : \exists(i_1, i_2) : \alpha_{i_1} \neq \alpha_{i_2}$$

$$H_{0b} : \beta_1 = \dots = \beta_b = 0 \quad \text{gegen} \quad H_{1b} : \exists(j_1, j_2) : \beta_{j_1} \neq \beta_{j_2}$$

GLM\_Synchro    GLM\_Cache

Zum Testen brauchen wir wieder (vorläufig) die Annahme

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

## 2-faktorielle Varianzanalyse

$$\bar{Y}_{..} = \frac{1}{a \cdot b} \sum_{i=1}^a \sum_{j=1}^b Y_{ij} \quad \text{arithmet. Mittel aller Beobachtungen}$$

$$\bar{Y}_{i.} = \frac{1}{b} \sum_{j=1}^b Y_{ij} \quad \text{Mittel aller Beobachtungen der } i\text{-ten Stufe von A}$$

$$\bar{Y}_{.j} = \frac{1}{a} \sum_{i=1}^a Y_{ij} \quad \text{Mittel aller Beobachtungen der } j\text{-ten Stufe von B}$$

$$SSA := b \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad SSB := a \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$SSE := \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

$$SST := \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2$$

## Varianzanalyse-Tabelle

Source	DF	Sum Squ.	Mean Squ.	F-value	$Pr > F$ p-value
A	a-1	SSA	MSA	$\frac{MSA}{MSE}$	$H_{1a}$
B	b-1	SSB	MSB	$\frac{MSB}{MSE}$	$H_{1b}$
Model	a+b-2	SSM	MSM	$\frac{MSM}{MSE}$	$H_1$
Error	(a-1)(b-1)	SSE	MSE		
Total	a b - 1	SST			

$$SSM = SSA + SSB$$

$$MSA = \frac{SSA}{(a-1)}$$

$$MSM = \frac{SSA + SSB}{a + b - 2}$$

$$SST = SSA + SSB + SSE$$

$$MSB = \frac{SSB}{(b-1)}$$

$$MSE = \frac{SSE}{(a-1)(b-1)}$$

## 2-faktorielle Varianzanalyse

Tests (1), Teilhypothesen

$H_{0a}$  gegen  $H_{1a}$ :

$$F_1 = \frac{MSA}{MSE} = \frac{\text{geschätzte Varianz zwischen Stufen von A}}{\text{geschätzte Varianz innerhalb der Gruppen}}$$
$$F_1 \sim F_{a-1, (a-1)(b-1)}$$

$H_{0b}$  gegen  $H_{1b}$ :

$$F_2 = \frac{MSB}{MSE} = \frac{\text{geschätzte Varianz zwischen Stufen von B}}{\text{geschätzte Varianz innerhalb der Gruppen}}$$
$$F_2 \sim F_{b-1, (a-1)(b-1)}$$

große Werte von  $F$  führen zur Ablehnung!

$$F_1 > F_{1-\alpha, a-1, (a-1)(b-1)} \rightarrow \text{Ablehnung von } H_{0a}$$

$$F_2 > F_{1-\alpha, b-1, (a-1)(b-1)} \rightarrow \text{Ablehnung von } H_{0b}$$

## 2-faktorielle Varianzanalyse

Tests (2), Globale Hypothese

$H_0: \alpha_1 = \dots = \alpha_a = 0$  und  $\beta_1 = \dots = \beta_a = 0$  gegen

$H_1: \exists(i_1, i_2): \alpha_{i_1} \neq \alpha_{i_2} \vee \exists(j_1, j_2): \beta_{j_1} \neq \beta_{j_2}$ .

$$F = \frac{MS_{Modell}}{MSE} = \frac{SSA + SSB}{SSE} \cdot \frac{(a-1)(b-1)}{a+b-2}$$

$$MS_{Modell} = \frac{SS_{Modell}}{a+b-2}$$

$$SS_{Modell} = SSA + SSB.$$

$H_0$  ablehnen, falls

$$F > F_{1-\alpha, a+b-2, (a-1)(b-1)}.$$

# Zweifaktorielle Varianzanalyse

Programm

```
PROC GLM;
```

```
  CLASS A B; /*die beiden Faktoren*/
```

```
  MODEL Y = A B;
```

```
RUN;
```

Output

- Balanzierter Fall: Variante I und III identisch
- Unbalanzierter Fall: Typ III-Summen sind vorzuziehen, da der entsprechende Test unabhängig von den Stichprobenumfängen ist.

# Inhalt

- Vergleich von  $k$  unabhängigen Gruppen
- Multiple Vergleiche
- Vergleich von  $k$  abhängigen Gruppen
- Weitere Varianzanalyse-Modelle
  - Mehrere Beobachtungen pro Kombination der Faktoren A und B
  - Wechselwirkungen ins Modell mit aufnehmen
  - Faktoren (Effekte, Faktorstufen) sind zufällig
  - Mehr als 2 Faktoren
  - Hierarchische Modelle

# Mehrere Beobachtungen pro Kombination der Faktoren A und B

SAS-Prozedur ändert sich nicht!

Output ändert sich gegebenenfalls

a) balanzierter Fall → eindeutig

b) unbalanzierter Fall → Es gibt verschiedene Möglichkeiten die Fehlerquadratsummen zu zerlegen.

→ SAS bietet die Varianten an

3 Forscher graben eine Reihe von Schädeln in 3 verschiedenen Schichten aus.

Gemessen wird die Nasenlänge.

? Forschereffekt, Schichteneffekt

## Klinische Untersuchung in mehreren Zentren

Ein Medikament zur Gewichtsreduktion soll getestet werden.

1: Medikament

0: Placebo

1-6: Zentren

## Klinische Untersuchung in mehreren Zentren

Ein Medikament zur Gewichtsreduktion soll getestet werden.

1: Medikament

0: Placebo

1-6: Zentren

### Modell:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) \quad \text{unabhängig}$$

Es interessiert nur das Medikament, nicht das Zentrum:

$$H_0 : \alpha_0 = \alpha_1 \quad H_1 : \alpha_0 < \alpha_1$$

# Weitere Varianzanalyse-Modelle

**PROC GLM;**

```
CLASS Medik Zentrum; /*die beiden Faktoren*/
```

```
MODEL Y = Medik Zentrum;
```

**RUN;** (dieselbe Prozedur wie oben)

GLM\_Drugeffect

Zum Output: wie bisher.

Balanzierter Fall: Variante I und III identisch.

Unbalanzierter Fall: Typ III-Summen zu bevorzugen, da der entsprechende Test unabhängig von den Stichprobenumfängen ist.

## Wechselwirkungen ins Modell mit aufnehmen

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) \quad \text{unabhängig}$$

(+Reparametrisierungsbedingungen)

geht nur, wenn für jede Faktorstufenkombination mehrere Beobachtungen vorliegen.

**PROC GLM;**

**CLASS** A B; /\*die beiden Faktoren\*/

**MODEL** Y = A B A\*B;

**RUN;**

GLM\_Insekten

# Weitere Varianzanalyse-Modelle

Modell mit Wechselwirkungen

Folgende Hypothesen sind zu testen:

$$H_{0a} : \alpha_1 = \dots = \alpha_a = 0 \quad \text{gegen}$$

$$H_{1a} : \exists(i_1, i_2) : \alpha_{i_1} \neq \alpha_{i_2}$$

$$H_{0b} : \beta_1 = \dots = \beta_b = 0 \quad \text{gegen}$$

$$H_{1b} : \exists(j_1, j_2) : \beta_{j_1} \neq \beta_{j_2}$$

$$H_{0c} : \gamma_{11} = \dots = \gamma_{a*b} = 0 \quad \text{gegen}$$

$$H_{1c} : \exists(j_1, j_2) : \gamma_{j_1, j_2} \neq 0$$

## Faktoren (Effekte, Faktorstufen) zufällig

hier ist Schätzung der Varianzkomponenten interessant und evtl. ein Hypothesentest

Preisrichter seien zufällig ausgewählt.

Die Frage ist, ob die Variabilität in den Scores an den Preisrichtern liegt?

$$Y_{ij} = \mu + \underbrace{A_i}_{\text{zufällig}} + b_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \text{unabhängig}$$

$$A_i \sim (0, \sigma_P^2)$$

$$\epsilon_{ij} \sim (0, \sigma^2)$$

# Varianzkomponentenschätzung

```
PROC VARCOMP METHOD=Type1;
```

```
  CLASS Preisrichter Wettkaempfer;
```

```
  MODEL Score = Preisrichter;
```

```
RUN;
```

```
  GLM_syncro_zufaelligeEffekte
```

METHOD=Type1: Auf den Quadratsummen beruhende  
Varianzschätzungen

Annahme:  $A_i$  und  $\epsilon_{ij}$  unabhängig.

$$\text{var}(Y_{ij}) = \text{var}(A_i) + \text{var}(\epsilon_{ij})$$

Output: Schätzungen für die Varianzkomponenten.

# Mehr als 2 Faktoren- höherfaktorielle Varianzanalyse

## Frequenzspektren

Gemessen wird die Amplitude bei 35 verschiedenen Frequenzen, 4 Füllungen, 3 Richtungen, jede Messung wird 5 mal wiederholt.

? Füllungs-, Richtungseffekt, Wiederholungseffekt? Frequenzeffekt?

→ 4 Faktoren.

**PROC GLM;**

**CLASS** A B C D;

**MODEL** Y = A B C D; **RUN;**

# Hierarchische Modelle

Die Faktoren liegen in hierarchischer Ordnung vor.

A											
A1			A2			A3			A4		
B11	B12	B13	B21	B22	B23	B31	B32	B33	B41	B42	B43

## Kalzium-Gehalt verschiedener Pflanzen und von verschiedenen Blättern

4 Pflanzen werden zufällig ausgewählt (zufällige Effekte)

3 Blätter davon

2 Stichproben zu 100mg von jedem Blatt

Frage: Gibt es zwischen Pflanzen oder zwischen Blättern unterschiedliche CA-Konzentrationen?

# Weitere Varianzanalyse-Modelle

## Hierarchische Modelle (2)

### Modell

$$Y_{ijk} = \mu + A_i + B_{ij} + \epsilon_{ijk}$$

$$A_i \sim \mathcal{N}(0, \sigma_a^2), \quad B_{ij} \sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

alle unabhängig

$$\text{hier: } n = 2 \quad a = 4 \quad b = 3$$

$$\begin{aligned} \text{var} Y_{ijk} &= \text{var} A_i + \text{var} B_{ij} + \text{var} \epsilon_{ijk} \\ &= \sigma_a^2 + \sigma_b^2 + \sigma^2 \end{aligned}$$

$$H_{0a} : \sigma_a^2 = 0 \quad H_{0b} : \sigma_b^2 = 0$$

# Weitere Varianzanalyse-Modelle

Hierarchische Modelle (3)

**PROC GLM;**

CLASS A B;

MODEL Y = A B(A);      (hierarchische Struktur)

RANDOM A B(A);      (Faktoren sind zufällig)

**RUN;**

**PROC VARCOMP;**

CLASS A B;

MODEL Y=A B(A);

**RUN;**

# Inhalt

- Einführung
- EDF-Tests
- Shapiro-Wilk-Test
- Anpassungstests auf weitere Verteilungen

- 9.1 Einführung
  - empirische Verteilungsfunktion
- 9.2 EDF-Anpassungstests
  - Kolmogorov-Smirnov-Test
  - Anderson-Darling-Test
  - Cramer-von Mises-Test
- 9.3 Anpassungstest auf Normalverteilung -
  - Shapiro-Wilk-Test
- 9.4. Anpassungstests auf weitere Verteilungen

## Problem

Klassische Test- und Schätzverfahren sind oft konzipiert unter der Normalverteilungsannahme.

## Frage

Gilt sie überhaupt?

# Gilt die Normalverteilung? (1)

## Hampel, 1980, Biometrisches Journal

Eine Zeitlang glaubte (fast) jeder an das 'normale Fehlergesetz',

die Mathematiker, weil sie es für ein empirisches Faktum hielten,

und die Anwender, weil sie es für ein mathematisches Gesetz hielten.

## Gilt die Normalverteilung? (2)

**Geary 1947, Biometrika**

Normality is a myth;  
there never was,  
and never will be,  
a normal distribution.

# Anpassungstests

Seien  $X_1, \dots, X_n$  unabhängige identisch verteilte Beobachtungen,  
 $X_i \sim F$ ,  $F$  unbekannt.

Anpassungstest auf eine spezifizierte Verteilung:

$$H_0 : F = F_0 \quad \text{gegen} \quad H_1 : F \neq F_0.$$

I.A. hängt  $F$  von unbekanntem Parametern ab.

Anpassungstest auf eine Normalverteilung:

$$H_0 : F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (\mu, \sigma \text{ unbekannt})$$

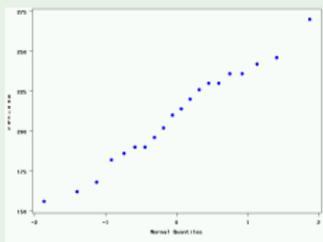
$$H_1 : F(x) \neq \Phi\left(\frac{x - \mu}{\sigma}\right) \quad \forall \mu, \sigma, \sigma > 0$$

( $\Phi$ : Verteilungsfunktion der Standardnormal.).

# Anpassungstests

## Gewicht von Hühnern

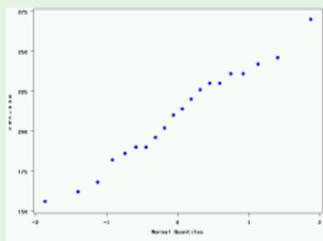
156	162	168	182	186
190	190	196	202	210
214	220	226	230	230
236	236	242	246	270



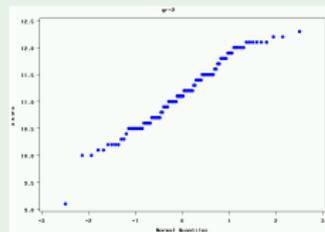
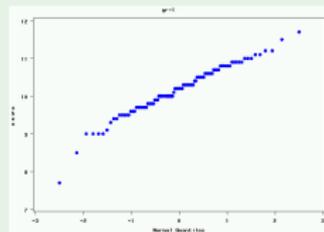
# Anpassungstests

## Gewicht von Hühnern

156	162	168	182	186
190	190	196	202	210
214	220	226	230	230
236	236	242	246	270



## Abmessungen von Banknoten, oben (echt, falsch)



# Inhalt

- Einführung
- EDF-Tests
- Shapiro-Wilk-Test
- Anpassungstests auf weitere Verteilungen

## Empirische Verteilungsfunktion

Seien  $X_1, \dots, X_n$  unabhängige Beobachtungen, und

$X_{(1)} \leq \dots \leq X_{(n)}$  die geordneten Beobachtungen

Die Funktion

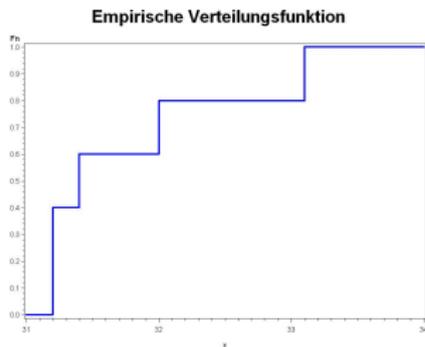
$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x < X_{(i+1)} \\ 1 & X_{(n)} \leq x \end{cases} \quad i = 1 \dots n$$

heißt empirische Verteilungsfunktion.

Satz v. Glivento-Cantelli:  $F_n(x) \rightarrow F(x)$ .

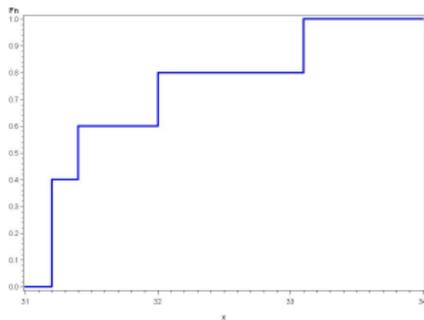
(Hauptsatz der math. Statistik genannt)

# Die empirische Verteilungsfunktion

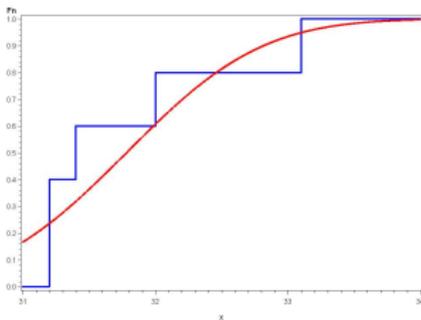


# Die empirische Verteilungsfunktion

Empirische Verteilungsfunktion



Empirische Verteilungsfunktion



# Anpassungstests

Auf der empirischen Verteilungsfunktion beruhende Tests

## Kolmogorov-Smirnov-Test

$$D = \sqrt{n} \sup_x |F_n(x) - F_0(x)|$$

## Cramér-von Mises-Test

$$W\text{-sq} = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x)$$

## Anderson-Darling-Test

$$A\text{-sq} = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

# Anpassungstests auf Normalverteilung

Auf der empirischen Verteilungsfunktion beruhende Tests

hier:

$$F_0(x) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right),$$
$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$
$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$D \sim D_n$  (Kolmogorov-Verteilung) approx.

$$\lim_{n \rightarrow \infty} P_0(D < \frac{x}{\sqrt{n}}) = 1 - 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2}$$

(Kolmogorov, 1933).

# Anpassungstests auf Normalverteilung

Auf der empirischen Verteilungsfunktion beruhende Tests (2)

Modifikationen für endliche Stichproben (zur Info.)

$$\mathbf{D:} \quad D \cdot (\sqrt{n} - 0.01 + 0.85/\sqrt{n})/\sqrt{n}$$

$$\mathbf{A - sq:} \quad A\text{-sq} \cdot (1.0 + 0.75/n + 2.25/n^2)$$

$$\mathbf{W\text{-sq:}} \quad W\text{-sq} \cdot (1.0 + 0.5/n)$$

Große Werte von  $D$ ,  $A\text{-sq}$  und  $W\text{-sq}$  führen jeweils zur Ablehnung von  $H_0$ .

$p$ -Werte werden vom Programm berechnet.

[Test\\_GoF\\_Banknote.sas](#)

[Test\\_GoFDarwin.sas](#)

# Inhalt

- Einführung
- EDF-Tests
- Shapiro-Wilk-Test
- Anpassungstests auf weitere Verteilungen

# Anpassungstests, Shapiro-Wilk-Test (1)

Vorbemerkungen:

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad Y_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$i = 1, \dots, n.$$

Geordnete Beobachtungen:

$$X_{(1)} \leq \dots \leq X_{(n)} \quad Y_{(1)} \leq \dots \leq Y_{(n)}.$$

Die Erwartungswerte

$$\begin{aligned} m_i &:= E(Y_{(i)}) \\ &= \frac{n!}{(i-1)!(n-i)!} \cdot \int_{-\infty}^{\infty} t \Phi^{i-1}(t) (1 - \Phi(t))^{n-i} \phi(t) dt \end{aligned}$$

sind bekannt (und vertafelt).

## Shapiro-Wilk-Test (2)

Approximation (Blom, 1958)

$$m_i \approx \tilde{m}_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

$$\mathbf{E}X_{(i)} = \mu + \sigma m_i$$

$$X_{(i)} = \mu + \sigma m_i + \epsilon_i$$

einfaches lineares Regressionsmodell mit Parametern  $\mu, \sigma$ .

$\mathbf{E}\epsilon_j = 0$ , aber die  $\epsilon_j$  sind nicht unabhängig.

$$\mathbf{V} := \text{cov}(Y_{(i)}, Y_{(j)}), \quad \mathbf{m}' := (m_1, \dots, m_n)$$

$$\mathbf{X}' := (X_{(1)}, \dots, X_{(n)}).$$

## Shapiro-Wilk-Test (3)

Verallgemeinerter Kleinstes Quadrat-Schätzer von  $\sigma$ :

$$\hat{\sigma} = \frac{\mathbf{m}'\mathbf{V}^{-1}\mathbf{X}}{\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}}$$

## Shapiro-Wilk-Test (3)

Verallgemeinerter Kleinsten Quadrat-Schätzer von  $\sigma$ :

$$\hat{\sigma} = \frac{\mathbf{m}'\mathbf{V}^{-1}\mathbf{X}}{\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}}$$

wird verglichen mit der gewöhnlichen empirischen Standardabweichung  $s$

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

## Shapiro-Wilk-Test (3)

Verallgemeinerter Kleinsten Quadrat-Schätzer von  $\sigma$ :

$$\hat{\sigma} = \frac{\mathbf{m}'\mathbf{V}^{-1}\mathbf{X}}{\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}}$$

wird verglichen mit der gewöhnlichen empirischen Standardabweichung  $s$

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

**Bem.:** Der gewöhnliche kleinste Quadrat-Schätzer von  $\sigma$  ist:

$$\hat{\sigma}_{KQS} = \frac{\mathbf{m}'\mathbf{X}}{\mathbf{m}'\mathbf{m}}.$$

# Shapiro-Wilk Test (4)

## Shapiro-Wilk-Statistik

$$W = \frac{\hat{\sigma}^2}{s^2(n-1)} .$$

# Shapiro-Wilk Test (4)

## Shapiro-Wilk-Statistik

$$W = \frac{\hat{\sigma}^2}{s^2(n-1)} \cdot \frac{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{m})^2}{\mathbf{m}'\mathbf{V}^{-2}\mathbf{m}} = \frac{(\mathbf{h}'\mathbf{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \mathbf{h}'\mathbf{h}}$$

wobei  $\mathbf{h}' = \mathbf{m}'\mathbf{V}^{-1}$  (bekannt, vertafelt).

# Shapiro-Wilk Test (4)

## Shapiro-Wilk-Statistik

$$W = \frac{\hat{\sigma}^2}{s^2(n-1)} \cdot \frac{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{m})^2}{\mathbf{m}'\mathbf{V}^{-2}\mathbf{m}} = \frac{(\mathbf{h}'\mathbf{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \mathbf{h}'\mathbf{h}}$$

wobei  $\mathbf{h}' = \mathbf{m}'\mathbf{V}^{-1}$  (bekannt, vertafelt).

Wegen  $\sum h_i = 0$  folgt:

$W$  ist Quadrat des (empirischen) Korrelationskoeffizienten von  $\mathbf{h}$  und  $\mathbf{X}$ :

$$W = \frac{\left(\sum_{i=1}^n (X_i - \bar{X})(h_i - \bar{h})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (h_i - \bar{h})^2},$$

## Shapiro-Wilk Test (5)

$$W = \frac{\left(\sum_{i=1}^n (X_i - \bar{X})(h_i - \bar{h})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (h_i - \bar{h})^2},$$

- Offenbar:  $0 \leq W \leq 1$ .
- $W \approx 1$  indiziert,  $\mathbf{h}' = \mathbf{m}'\mathbf{V}^{-1} (\approx 2\mathbf{m}')$  ist Vielfaches von  $\mathbf{X}$ .  
D.h. die Punkte  $(m_i, X_{(i)})$  liegen etwa auf einer Geraden, was Normalverteilung indiziert.
- $H_0$  wird ablehnt, falls  $W < W_\alpha(n)$ . SAS verwendet dabei noch eine (Normalisierungs-)Transformation von  $W$ .

# Shapiro-Wilk Test (6)

zur groben Illustration:

## Scores der 1. Wettkämpferinnen (5 Preisrichter)

31.2, 31.2, 31.4, 32.0, 33.1

Mit der Prozedur UNIVARIATE erhalten wir  $s = 0.80747$  und mit der Prozedur GPLOT (Option REGEQN)  $\sigma_{\hat{KQS}} = 0.805$  im Regressionsmodell  $Y_i = \mu + \sigma m_i + \epsilon_i$

Es wird allerdings die verallgemeinerte KQS  $\hat{\sigma}$  verwendet. Für die Shapiro-Wilk Statistik bekommen wir

$$W = \frac{\hat{\sigma}^2}{s^2} \cdot c = 0.966 \quad (c: \text{Normierungsfaktor})$$

Nach der Transformation wird daraus:  $W = 0.8125$ .

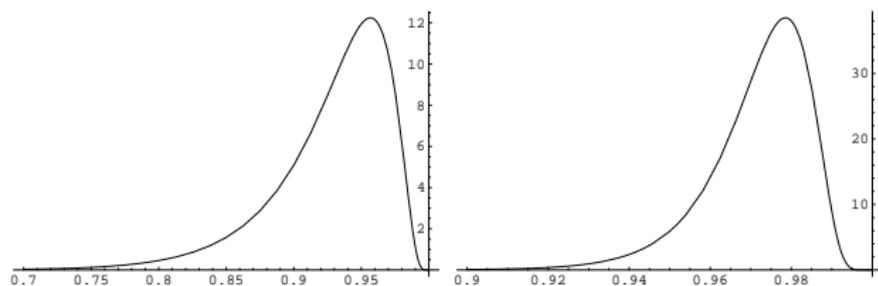
# Shapiro-Wilk Test (7)

Approximative Dichtefunktion von  $W$  (unter  $H_0$ )

$$W = \frac{\left(\sum_{i=1}^n (X_i - \bar{X})(h_i - \bar{h})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (h_i - \bar{h})^2},$$

$n = 10$

$n = 50$



# Anpassungstests

- SAS verwendet eine Approximation von  $W$ .
- Der Shapiro-Wilk-Test erweist sich für kleinere, mittlere und größere Stichprobenumfänge als geeignetster Test (er hat für die meisten Alternativen die höchste Güte).
- Früher wurde meist der sogenannte  $\chi^2$ -Anpassungstest verwendet. Dieser hat jedoch geringe Güte.
- $W$  ist etwas besser als A-sq, besser als W-sq, und viel besser als  $D$  und  $\chi^2$ .
- $D$  ist nur für sehr große Stichprobenumfänge zu empfehlen ( $n \geq 2000$ ).

# Anpassungstests

- Man sollte beim Test auf Normalverteilung das Signifikanzniveau auf  $\alpha = 0.1$  hochsetzen, insbesondere wenn wenig robuste Tests (die Normalverteilung verlangen) angewendet werden sollen.
- Die Verwendung eines Test auf Normalverteilung als Vortest für weitere Tests ist etwas umstritten.
- Robuste Tests haben meist geringen Effizienzverlust bei Vorliegen von Normalverteilung.

# Durchführung des Tests auf Normalverteilung

Unter Verwendung von  $\hat{\mu}, \hat{\sigma}$ :

```
PROC UNIVARIATE NORMAL; RUN;
```

```
PROC UNIVARIATE;
```

```
HISTOGRAM variable / NORMAL; RUN;
```

mit vorgebenen  $\mu, \sigma$ :

```
PROC UNIVARIATE;
```

```
HISTOGRAM variable / NORMAL(mu=0, sigma=1);
```

```
RUN;
```

**Bem.:** Mit der Prozedur UNIVARIATE (Kommando HISTOGRAM) können Sie auch auf andere Verteilungen testen.

# Inhalt

- Einführung
- EDF-Tests
- Shapiro-Wilk-Test
- Anpassungstests auf weitere Verteilungen

## $\chi^2$ -Anpassungstest (Pearson, 1900)

Prinzip: Daten werden in  $p$  Klassen eingeteilt.

Klassenhäufigkeiten:  $N_i$

theoretische Klassenhäufigkeiten:  $np_i$

$$\chi^2 = \sum_{i=1}^p \frac{(N_i - np_i)^2}{np_i}$$

$\chi^2 \sim \chi_{p-1}^2$  asymptotisch (bei bekannten  $\mu, \sigma^2$ ) (Fisher, 1922)

$\chi^2 \sim \chi_{p-3}^2$  approx. (bei 2 zu schätzenden Parametern, ML-Schätzung mit gruppierten Daten oder Minimum- $\chi^2$ -Schätzung).

# Anpassungstests

## $\chi^2$ -Anpassungstest

### Nachteile des $\chi^2$ -Anpassungstests

- Wert von  $\chi^2$  abhängig von Klasseneinteilung.
- $\chi^2$ - Anpassungstest auf Normalverteilung hat geringe Güte.

### Diskrete Verteilungen

Hier kann der  $\chi^2$ -Anpassungstest genommen werden (natürliche Klasseneinteilung)

Prozedur FREQ, Option CHISQ

# $\chi^2$ -Anpassungstest

## Diskrete Gleichverteilung

```
PROC FREQ;
```

```
    TABLES var1 /CHISQ; RUN;
```

## Sonstige diskrete Verteilungen

wie oben, zusätzlich sind die Einzelwktn. explizit zu formulieren,

```
/CHISQ TESTP=( $p_1, \dots, p_k$ );
```

```
Test_GoF_Poisson    PoissonHorsekicks
```

## Anzahlen schon gegeben

Die Variablen, die Anzahlen bezeichnen, werden durch ein WEIGHT-Kommando angegeben.

# Anpassungstests

## EDF-Tests

### Stetige Verteilungen

zugelassen sind:

Normal, Gamma, Weibull, Lognormal, Exponential

```
HISTOGRAM var1 / Gamma;
```

```
Descr_Plot_Kuehl.sas
```

```
Test_GoF_Darwin_1.sas
```

# Inhalt

- Einführung
- Einstichprobenproblem
- Zweistichprobenproblem
- Mehrere unverbundene Stichproben
- Mehrere verbundene Stichproben

Es werden die wichtigsten Rang-Analoga zu den Tests in 7.2.-7.4., 8.1,8.3 behandelt.

10.1 Einführung

10.2 Einstichprobenproblem (vgl 7.2), 2 verbundene Stichproben (vgl. 7.3)

Vorzeichentest, Vorzeichen-Wilcoxon-Test

10.3 Zwei unverbundene Stichproben (vgl. 7.4)

Wilcoxon-Test

10.4 Mehrere unabhängige Stichproben (vgl. 8.1)

Kruskal-Wallis-Test

10.5 Mehrere verbundene Stichproben (vgl. 8.3)

Friedman-Test

Was tun wenn Normalverteilung nicht vorliegt?

Was tun wenn Normalverteilung nicht vorliegt?

## Nichtparametrische Tests

- sie verwenden keine Parameterschätzung (wie  $\bar{X}$ ,  $s$ )
- sie halten das Signifikanzniveau ( $\alpha$ ) für jede stetige Verteilung (approximativ) ein.  $\alpha$  hängt also nicht von der zugrundeliegenden Verteilungsfunktion ab.
- sie sind relativ effizient. Der Effizienzverlust bei Normalverteilung ist in vielen Fällen gering!

Annahme: Verteilungsfunktion ist stetig (wenn nicht anders vermerkt)

# Inhalt

- Einführung
- Einstichprobenproblem
- Zweistichprobenproblem
- Mehrere unverbundene Stichproben
- Mehrere verbundene Stichproben

Nulhypothese

Alternative

a)  $H_0 : \mu \leq \mu_0$

$H_A : \mu > \mu_0$

b)  $H_0 : \mu \geq \mu_0$

$H_A : \mu < \mu_0$

c)  $H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

### Vorzeichentest

Wie bisher werden die Differenzen  $X_i - \mu_0$  gebildet.

$$V_i := \begin{cases} 1 & \text{falls } X_i - \mu_0 > 0 \\ 0 & \text{falls } X_i - \mu_0 < 0 \end{cases}$$

Nulhypothese

Alternative

a)  $H_0 : \mu \leq \mu_0$

$H_A : \mu > \mu_0$

b)  $H_0 : \mu \geq \mu_0$

$H_A : \mu < \mu_0$

c)  $H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

### Vorzeichentest

Wie bisher werden die Differenzen  $X_i - \mu_0$  gebildet.

$$V_i := \begin{cases} 1 & \text{falls } X_i - \mu_0 > 0 \\ 0 & \text{falls } X_i - \mu_0 < 0 \end{cases}$$

$$V^+ = \sum_{i=1}^n V_i$$

$V^+ = \#$  Differenzen mit positivem Vorzeichen

## Vorzeichentest (2)

Der Fall  $X_i - \mu_0 = 0$  tritt wegen der Stetigkeit

der Verteilungsfunktion nur mit Wahrscheinlichkeit Null auf. Sollte der Fall trotzdem eintreten (Messungenauigkeit) so wird die entsprechende Beobachtung weggelassen und der Stichprobenumfang entsprechend verringert.

(Nachteil: Es werden gerade Beobachtungen weggelassen, die für die Nullhypothese sprechen!)

## Vorzeichentest (2)

Der Fall  $X_i - \mu_0 = 0$  tritt wegen der Stetigkeit

der Verteilungsfunktion nur mit Wahrscheinlichkeit Null auf. Sollte der Fall trotzdem eintreten (Messungenauigkeit) so wird die entsprechende Beobachtung weggelassen und der Stichprobenumfang entsprechend verringert.

(Nachteil: Es werden gerade Beobachtungen weggelassen, die für die Nullhypothese sprechen!)

Es gilt:  $V^+ \sim$

## Vorzeichentest (2)

Der Fall  $X_i - \mu_0 = 0$  tritt wegen der Stetigkeit

der Verteilungsfunktion nur mit Wahrscheinlichkeit Null auf. Sollte der Fall trotzdem eintreten (Messungenauigkeit) so wird die entsprechende Beobachtung weggelassen und der Stichprobenumfang entsprechend verringert.

(Nachteil: Es werden gerade Beobachtungen weggelassen, die für die Nullhypothese sprechen!)

Es gilt:  $V^+ \sim Bi(n, \frac{1}{2})$

( $V^+ = \#$  "Erfolge" bei  $n$  Versuchen mit Wkt. je  $\frac{1}{2}$ ).

⇒ kritische Werte könnten leicht selbst bestimmt werden:

$BINV(1 - \alpha, n, \frac{1}{2})$  oder  $QUANTILE('Binomial', 1 - \alpha, n, \frac{1}{2})$

# Vorzeichentest (3)

## Teststatistik

$$\underline{M} = V^+ - \frac{n}{2} \quad \left( = \frac{V^+ - V^-}{2} \right) \quad (\text{zentrierte Statistik})$$

$n^+$ : Realisierung von  $V^+$

$n^-$ : Realisierung von  $V^-$

Zweiseitiger p-Wert:

$$P(|\underline{M}| \geq |n^+ - \frac{n}{2}|) = P(|\underline{M}| \geq \max(n^+, n^-) - \frac{n}{2}) = (*)$$

$$\text{denn } |n^+ - \frac{n}{2}| = \begin{cases} n^+ - \frac{n}{2} & n^+ > \frac{n}{2} \\ \frac{n}{2} - n^+ & n^+ < \frac{n}{2} \\ = n^- - \frac{n}{2} & \end{cases}$$

# Vorzeichentest (4)

Der p-Wert ist gleich

$$(*) = P\left(V^+ - \frac{n}{2} \geq \max(n^+, n^-) - \frac{n}{2}\right) + P\left(\frac{n}{2} - V^+ \geq \max(n^+, n^-) - \frac{n}{2}\right)$$

# Vorzeichentest (4)

Der p-Wert ist gleich

$$\begin{aligned} (*) &= P\left(V^+ - \frac{n}{2} \geq \max(n^+, n^-) - \frac{n}{2}\right) + P\left(\frac{n}{2} - V^+ \geq \max(n^+, n^-) - \frac{n}{2}\right) \\ &= P\left(V^+ \geq \max(n^+, n^-)\right) + P\left(n - V^+ \geq \max(n^+, n^-)\right) \end{aligned}$$

# Vorzeichentest (4)

Der p-Wert ist gleich

$$\begin{aligned}
 (*) &= P\left(V^+ - \frac{n}{2} \geq \max(n^+, n^-) - \frac{n}{2}\right) + P\left(\frac{n}{2} - V^+ \geq \max(n^+, n^-) - \frac{n}{2}\right) \\
 &= P(V^+ \geq \max(n^+, n^-)) + P(n - V^+ \geq \max(n^+, n^-)) \\
 &= 2 \sum_{j=\max(n^+, n^-)}^n \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} = \left(\frac{1}{2}\right)^{n-1} \sum_{j=\max(n^+, n^-)}^n \binom{n}{j} \\
 &= \left(\frac{1}{2}\right)^{n-1} \sum_{j=0}^{\min(n^+, n^-)} \binom{n}{j}.
 \end{aligned}$$

## Vorzeichentest (5)

Die Verteilung von  $V^+$  ist diskret, d.h. es gibt nicht zu jedem  $\alpha$  einen entsprechenden kritischen Wert.

Aber: p-Werte gibt es immer, d.h.:

$$p < \alpha \quad \Rightarrow \quad H_0 \text{ (c) ablehnen}$$

$$M > 0 \wedge \frac{p}{2} < \alpha \quad \Rightarrow \quad H_0 \text{ (b) ablehnen}$$

$$M < 0 \wedge \frac{p}{2} < \alpha \quad \Rightarrow \quad H_0 \text{ (a) ablehnen}$$

Der Vorzeichentest ist meist nicht sehr effizient

(Ausnahme: Verteilung=Doppelexponential)

besser ist der Wilcoxon-Vorzeichen-Rangtest

## Wilcoxon-Vorzeichen-Rangtest

Bilden zu den "Beobachtungen"  $D_i = |X_i - \mu_0|$  die Rangzahlen, d.h. den Rang (den Platz) in der geordneten Stichprobe

$$\underbrace{D_{(1)}}_{\text{Rang 1}} \leq \dots \leq \dots \leq \underbrace{D_{(n)}}_{\text{Rang n}}$$

Sei  $R_i^+$  der Rang von  $D_i$ .

## Wilcoxon-Vorzeichen-Rangtest

Bilden zu den "Beobachtungen"  $D_i = |X_i - \mu_0|$  die Rangzahlen, d.h. den Rang (den Platz) in der geordneten Stichprobe

$$\underbrace{D_{(1)}}_{\text{Rang 1}} \leq \dots \leq \dots \leq \underbrace{D_{(n)}}_{\text{Rang n}}$$

Sei  $R_i^+$  der Rang von  $D_i$ .

$$W_n^+ = \sum_{i=1}^n R_i^+ \cdot V_i$$

Summe der Ränge von  $D_i$  für die  $X_i - \mu_0 > 0$ .

# Wilcoxon-Vorzeichen-Rangtest (2)

Erwartungswert und Varianz von  $W_n^+$

$$\mathbf{E}_0 W_n^+ = \frac{1}{2} \sum_{i=1}^n R_i^+ = \frac{1}{2} \sum_{i=1}^n i = \frac{n \cdot (n+1)}{4} \quad \mathbf{E} V_i = \frac{1}{2}$$

$$\text{var } W_n^+ = \mathbf{E}(W_n^+ - \mathbf{E}W_n^+)^2 = \frac{n \cdot (n+1)(2n+1)}{24} \quad (\text{ÜA})$$

# Wilcoxon-Vorzeichen-Rangtest (2)

## Erwartungswert und Varianz von $W_n^+$

$$\mathbf{E}_0 W_n^+ = \frac{1}{2} \sum_{i=1}^n R_i^+ = \frac{1}{2} \sum_{i=1}^n i = \frac{n \cdot (n+1)}{4} \quad \mathbf{E} V_i = \frac{1}{2}$$

$$\text{var } W_n^+ = \mathbf{E}(W_n^+ - \mathbf{E}W_n^+)^2 = \frac{n \cdot (n+1)(2n+1)}{24} \quad (\text{ÜA})$$

Die Berechnung der exakten Verteilung von  $W_n^+$  kann durch Auszählen aller Permutationen erfolgen

( $\rightarrow$  schon für kleinere  $n$  größere Rechenzeit!)

Deshalb verwendet man (für mittlere und große  $n$ ) die asymptotische Verteilung.

# Wilcoxon-Vorzeichen-Rangtest (3)

## Asymptotische Verteilung

$$W_n^+ \sim \mathcal{N}(EW_n^+, \text{var} W_n^+) \quad \text{asymptotisch}$$

Große Werte von

$$\frac{|W_n^+ - EW_n^+|}{\sqrt{\text{var} W_n^+}}$$

führen zur Ablehnung von  $H_0$  (gegen die zweiseitige Alternative).

# Wilcoxon-Vorzeichen-Rangtest (4)

## SAS-Implementation (Wilcoxon-Vorzeichen-Test)

$$S = W_n^+ - EW_n^+ = \sum_{X_i - \mu_0 > 0} R_i^+ V_i - \frac{n(n+1)}{4}$$

$R_i^+$  Rang von  $|X_i - \mu_0|$ ,

Summe nur über positive  $X_i - \mu_0$

$n \leq 20$ : p-Werte aus der exakten Verteilung von  $S$ .

$n > 20$ : Es wird eine  $t$ -Approximation verwendet:

$$t = \frac{S \cdot \sqrt{n-1}}{\sqrt{n \operatorname{Var}(S) - S^2}} \sim t_{n-1}$$

# Wilcoxon-Vorzeichen-Rangtest (5)

Bindungen (= Messwertwiederholungen): Ränge werden gemittelt.

Sei  $t_i$ : # Bindungen in der  $i$ -ten Gruppe.

Korrektur in  $\text{Var}(S)$ :

$$\text{var}(S) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{2} \sum t_i(t_i+1)(t_i-1)$$

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110 \quad H_1 : \mu > \mu_0$$

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110$$

$$H_1 : \mu > \mu_0$$

$$d_i = x_i - 110$$

$x_i = \text{IQ}$	$d_i$	$ d_i $	$r_i^+$	$V_i$
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110$$

$$H_1 : \mu > \mu_0$$

$$d_i = x_i - 110$$

$x_i = \text{IQ}$	$d_i$	$ d_i $	$r_i^+$	$V_i$
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

**Vorzeichentest:**

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110$$

$$H_1 : \mu > \mu_0$$

$$d_i = x_i - 110$$

$x_i = \text{IQ}$	$d_i$	$ d_i $	$r_i^+$	$V_i$
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

**Vorzeichentest:**

$$M = 8 - \frac{10}{2}$$

$$\text{p-Wert(exakt)} = 0.1094$$

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110$$

$$H_1 : \mu > \mu_0$$

$$d_i = x_i - 110$$

$x_i = \text{IQ}$	$d_i$	$ d_i $	$r_i^+$	$V_i$
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

**Vorzeichentest:**

$$M = 8 - \frac{10}{2}$$

$$\text{p-Wert(exakt)} = 0.1094$$

**Wilcoxon-signed**

$$W^+ - \mathbf{E}(W^+) =$$

# Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi),

$$H_0 : \mu = \mu_0 = 110$$

$$H_1 : \mu > \mu_0$$

$$d_i = x_i - 110$$

$x_i = \text{IQ}$	$d_i$	$ d_i $	$r_i^+$	$V_i$
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

**Vorzeichentest:**

$$M = 8 - \frac{10}{2}$$

$$\text{p-Wert(exakt)} = 0.1094$$

**Wilcoxon-signed**

$$W^+ - \mathbf{E}(W^+) =$$

$$48 - \frac{10 \cdot 11}{4} = 20.5.$$

$$\text{p-Wert} = 0.0371.$$

Test\_IQ\_Daten

# Wilcoxon-Vorzeichen-Rangtest (7)

- Im Gegensatz zum Vorzeichentest ist der Vorzeichen-Wilcoxon-Test (= signed rank test) sehr effizient, bei Normalverteilung nur wenig schlechter, bei den meisten Verteilungen besser als der  $t$ -Test.  
⇒ Wenn Normalverteilung nicht gesichert ist Vorzeichen-Wilcoxon-Rang-Test nehmen!

# Wilcoxon-Vorzeichen-Rangtest (7)

- Im Gegensatz zum Vorzeichentest ist der Vorzeichen-Wilcoxon-Test (= signed rank test) sehr effizient, bei Normalverteilung nur wenig schlechter, bei den meisten Verteilungen besser als der  $t$ -Test.  
⇒ Wenn Normalverteilung nicht gesichert ist Vorzeichen-Wilcoxon-Rang-Test nehmen!
- Der Vorzeichentest und der (Vorzeichen-)Wilcoxon-Test sind sogenannte Rangtests, da sie nur auf den Rangzahlen der Beobachtungen beruhen. Es gibt weitere Rangtests.

# Wilcoxon-Vorzeichen-Rangtest (7)

- Im Gegensatz zum Vorzeichentest ist der Vorzeichen-Wilcoxon-Test (= signed rank test) sehr effizient, bei Normalverteilung nur wenig schlechter, bei den meisten Verteilungen besser als der  $t$ -Test.  
⇒ Wenn Normalverteilung nicht gesichert ist Vorzeichen-Wilcoxon-Rang-Test nehmen!
- Der Vorzeichentest und der (Vorzeichen-)Wilcoxon-Test sind sogenannte Rangtests, da sie nur auf den Rangzahlen der Beobachtungen beruhen. Es gibt weitere Rangtests.
- Durchführung der Tests:

**PROC UNIVARIATE MU0=**Wert;

## Zwei verbundene Stichproben

Bilden  $Z := X - Y$  und testen wie beim Einstichprobenproblem,  
z.B.

$$H_0 : \mu_Z = 0$$

$$H_1 : \mu_Z \neq 0$$

Banknoten: oben-unten, links-rechts

Darwin: kreuz-selbstbefruchtete Pflanzen

(zur Illustration mit Prozedur RANK)

```
PROC UNIVARIATE;
```

```
VAR Z;
```

```
RUN;
```

```
Npar_1_Banknote  Npar_1_Darwin
```

# Weitere Fragestellungen im Einstichprobenfall (1)

## Binärvariablen

Sei  $X$  eine 0-1 Variable, d.h.

$$P(X = 0) = p, \quad P(X = 1) = 1 - p$$

$H_0 : p = p_0$        $T$ : Anzahl der Beobachtungen in Klasse 0.

$H_{1a} \quad p < p_0$  :      p-Wert =  $P(T \leq t) = \text{CDF}(\text{'Binomial'}, t, n, p_0)$

$H_{1b} \quad p > p_0$  :      p-Wert =  $P(T \geq t)$

$H_{1c} \quad p \neq p_0$  :      p-Wert =  $P(T \leq t \text{ oder } T \geq n - t + 1)$

## Weitere Fragestellungen im Einstichprobenfall (2)

### Binomialtest

Prozedur FREQ, Option Binomial im Tables-Kommando.

$$\hat{p} = \frac{T}{n} = \frac{\# \text{ Beobachtungen in Klasse 0}}{n}$$

$$se(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n} = ASE$$

$$Z = \frac{\hat{p} - p_0}{se(\hat{p})}$$

Einseitige p-Werte bei SAS sind

$$\begin{cases} P(Z > z) & \text{falls } z > 0 \\ P(Z < z) & \text{falls } z \leq 0 \end{cases}$$

# Binomialtest

```
PROC FREQ;
```

```
TABLES var / BINOMIAL(P=0.8);
```

```
RUN;
```

Binomialtest\_toxaemia.sas

## Warenlieferung, ÜA

Der Hersteller behauptet, höchstens 5% sind schlecht.

Sie haben  $n = 20$  Stücke geprüft, und  $X = 3$  schlechte Stücke gefunden.

Hat der Hersteller recht?

Betrachten Sie sowohl die exakte als auch die asymptotische Version.

## Konfidenzintervalle:

a) Normalapproximation:  $\hat{p} \pm u_{\alpha/2} se(\hat{p})$

b) exakt: Binomialverteilung (CDF('Binomial',....))

# Weitere Fragestellungen im Einstichprobenfall (4)

Zum Vergleich, zur Erinnerung und Ergänzung

## $\chi^2$ -Anpassungstest

Anpassungstest auf diskrete Gleichverteilung:

```
PROC FREQ;  
    TABLES var /CHISQ;  
RUN;
```

## Anpassungstest auf vorgegebene diskrete Verteilung

```
PROC FREQ;  
    TABLES var /CHISQ TESTP=( $p_1, \dots, p_k$ );  
RUN;
```

# Nichtparametrische Konfidenzintervalle

Option CIPCTLDF in der PROC UNIVARIATE

$(1 - \alpha)$ -Konfidenzintervall für  $p$ -Quantil, d.h. für  $x_p$

Die Verteilung der  $j$ -ten Ordnungsstatistik  $X_{(j)}$ :

$$P(X_{(j)} < x) = \sum_{i=j}^n \binom{n}{i} F(x)^i (1 - F(x))^{n-i}$$

“Erfolg” gdw.  $X_i < x$ , “Erfolgswahrscheinlichkeit”  $F(x)$ .

Insbesondere, für  $x = x_p$  (das wahre  $p$ -Quantil)

$$\begin{aligned} P(X_{(j)} < x_p) &= \sum_{i=j}^n \binom{n}{i} F(x_p)^i (1 - F(x_p))^{n-i} \\ &= \sum_{i=j}^n \binom{n}{i} p^i (1 - p)^{n-i} \end{aligned}$$

# Nichtparametrische Konfidenzintervalle

Option CIPCTLDF in der PROC UNIVARIATE (2)

$$P(X_{(j)} < x_p) = \sum_{i=j+1}^n \binom{n}{i} p^i (1-p)^{n-i}$$

Untere und obere Konfidengrenzen  $X_{(l)}$  und  $X_{(u)}$  für  $x_p$  werden so bestimmt, dass  $l$  und  $u$  (möglichst) symmetrisch um  $\lfloor np \rfloor + 1$  und so dass

$$P(X_{(l)} \leq x_p < X_{(u)}) = \sum_{i=l}^{u-1} \binom{n}{i} p^i (1-p)^{n-i} \geq 1 - \alpha$$

( $X_{(\lfloor np \rfloor)}$  ist Schätzung für  $x_p$ .)

**PROC UNIVARIATE CIPCTLDF;**

# Inhalt

- Einführung
- Einstichprobenproblem
- Zweistichprobenproblem
- Mehrere unverbundene Stichproben
- Mehrere verbundene Stichproben

# Zwei unverbundene Stichproben: Wilcoxon Test

Wir setzen keine Normalverteilung voraus, aber den gleichen Verteilungstyp, insbesondere gleiche Varianzen

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$$

$$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$$

Wir fassen die Beobachtungen  $X_{11}, \dots, X_{1n}, X_{21}, \dots, X_{2m}$  zu einer Stichprobe zusammen und bilden die Rangzahlen  $R_{ij}$ ,  $i = 1, 2, j = 1 \dots n, m$

$$\underbrace{Z_{(1)}}_{\text{Rang 1}} \leq \dots \leq \underbrace{Z_{(n+m)}}_{\text{Rang } n+m}$$

# Nichtparametrische Tests

## Wilcoxon-Test

Summe der Ränge zur 1. bzw. 2. Stichprobe

$$S_1 = \sum_{j=1}^n R_{1j}$$

$$S_2 = \sum_{j=1}^m R_{2j}$$

Die Erwartungswerte (unter  $H_0$ ) sind

$$\mathbf{E}_0 S_1 = \frac{n(n+m+1)}{2} \quad \text{und} \quad \mathbf{E}_0 S_2 = \frac{m(n+m+1)}{2}$$

und die Varianzen

$$\text{var} S_1 = \text{var} S_2 = \frac{n \cdot m(n+m+1)}{12}.$$

## Wilcoxon-Test (2)

Sei  $S$  die Statistik  $S_1$  oder  $S_2$ , die zur kleineren Stichprobe gehört.

Die Teststatistik des Wilcoxon-Tests ist

$$Z = \frac{S - E(S)}{\sqrt{\text{var}S}} \quad \text{SAS: } Z = \frac{S - E(S) + 0.5}{\sqrt{\text{var}S}}$$
$$Z \sim \mathcal{N}(0, 1) \quad \text{approximativ}$$

(0.5 = Stetigkeitskorrektur)

bei Bindungen: korrigierte (kleinere) Varianz

[Npar1way\\_Carnitinfraktion.sas](#)

[Npar1way\\_Banknote.sas](#)

[Npar1way\\_Heroin.sas](#)

[Npar1way\\_Tibetan.sas](#)

## Wilcoxon-Test (3)

- SAS gibt die Teststatistik ( $Z$ ) und die ein- und zweiseitigen  $p$ -Werte an.

a)  $H_0 : \mu_1 = \mu_2$                        $H_1 : \mu_1 \neq \mu_2$   
 $\Rightarrow$  two-sided  $Pr > |Z| = P(|Z| > Z)$

b)  $H_0 : \mu_1 \leq \mu_2$                        $H_1 : \mu_1 > \mu_2$   
 $\Rightarrow$  one-sided  $z > 0$   
 $\rightarrow P(Z > z) = Pr > Z$

c)  $H_0 : \mu_1 \geq \mu_2$                        $H_1 : \mu_1 < \mu_2$   
 $\Rightarrow$  one-sided  $z < 0$   
 $\rightarrow P(Z < z) = Pr < Z$

- SAS bietet die Normalapproximation und die t-Approximation an.

## Wilcoxon-Test (4)

- Manche Programmpakete, z.B. R, verwenden die Mann-Whitney-Version des Wilcoxon Tests mit folgender Prüfgröße

$$U = \sum_{i,j} I(X_i < Y_j),$$

wobei  $I$  Indikatorfunktion ist.

$U$  zählt also alle Paare von  $X$  und  $Y$ -Beobachtungen, für die eine  $X$ -Beobachtung kleiner ist als eine  $Y$ -Beobachtung.

Beide Versionen sind äquivalent, d.h. die  $p$ -Werte sind dieselben. Es gilt:

$$U = n(n + m) - \underbrace{\sum_{i=1}^n R_i}_W - \frac{(n - 1)n}{2}$$

# Fligner-Policello Test (1)

Verteilungsannahme: keine, außer Symmetrie.

Seien  $\theta_1, \theta_2$  die Mediane von  $X$  bzw.  $Y$ .

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$

$$H_0 : \theta_1 \leq \theta_2 \quad H_1 : \theta_1 > \theta_2$$

$$H_0 : \theta_1 \geq \theta_2 \quad H_1 : \theta_1 < \theta_2$$

Placements:

$$PI(X_i) = \sum_{j=1}^{n_2} I(Y_j < X_i) + \frac{1}{2} I(Y_j = X_i),$$

$$\overline{PI}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} PI(X_i)$$

$$PI(Y_j) = \sum_{i=1}^{n_1} I(X_i < Y_j) + \frac{1}{2} I(X_i = Y_j),$$

$$\overline{PI}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} PI(Y_j)$$

# Fligner-Policello Test (2)

## Fligner-Policello Test

$$FP = \frac{\sum_{j=1}^{n_2} PI(Y_j) - \sum_{i=1}^{n_1} PI(X_i)}{2\sqrt{V_1 + V_2 + \overline{PI}_1 \overline{PI}_2}}$$

$$V_1 = \sum_{i=1}^{n_1} (PI(X_i) - \overline{PI}_1)^2, \quad V_2 = \sum_{j=1}^{n_2} (PI(Y_j) - \overline{PI}_2)^2$$

$$FP \sim \mathcal{N}(0, 1) \quad \text{unter } \theta_1 = \theta_2$$

**PROC NPAR1WAY FP**

Wir machen gar keine Verteilungsannahmen.

$$H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$$

$$H_0 : F_1 \leq F_2 \quad H_1 : F_1 > F_2$$

$$H_0 : F_1 \geq F_2 \quad H_1 : F_1 < F_2$$

### Kolmogorov-Smirnov Test

$$D = \max_i |F_{1,n}(x) - F_{2,m}(x)| \quad (\text{zweiseitig, EDF})$$

$$D^+ = \max_i (F_{1,n}(x) - F_{2,m}(x)) \quad (\text{einseitig, D})$$

$$D^- = \max_i (F_{2,m}(x) - F_{1,n}(x)) \quad (\text{einseitig, D})$$

**PROC NPAR1WAY EDF D;**

# Cramér-von Mises Test

Wir machen gar keine Verteilungsannahmen.

$$H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$$

$$H_0 : F_1 \leq F_2 \quad H_1 : F_1 > F_2$$

$$H_0 : F_1 \geq F_2 \quad H_1 : F_1 < F_2$$

## Cramér-von Mises Test

$$CM = \frac{1}{n^2} \sum_{i=1}^2 n_i \sum_{j=1}^{n_i} t_j (F_{1,n}(x_j) - F_n(x_j))^2$$

$$F_n(x_j) = \frac{1}{n} \sum_{i=1}^2 n_i F_{i,n}$$

**PROC NPAR1WAY EDF ;**

# Allgemeine Empfehlungen (1)

- Wenn Normalverteilung, gleiche Varianzen und keine Ausreißer:

# Allgemeine Empfehlungen (1)

- Wenn Normalverteilung, gleiche Varianzen und keine Ausreißer:  
*t*-Test

# Allgemeine Empfehlungen (1)

- Wenn Normalverteilung, gleiche Varianzen und keine Ausreißer:  
*t*-Test
- Wenn Normalverteilung, ungleiche oder unbekannte Varianzen und keine Ausreißer:

# Allgemeine Empfehlungen (1)

- Wenn Normalverteilung, gleiche Varianzen und keine Ausreißer: *t*-Test
- Wenn Normalverteilung, ungleiche oder unbekannte Varianzen und keine Ausreißer: **Welch-Test** (*t*-Test, unpooled, Satterthwaite)
- Wenn “sehr nahe” an Normalverteilung und keine Ausreißer: wie bei Normalverteilung
- keine Normalverteilung oder unbekannte Verteilung, gleiche Varianzen, und etwa gleicher Verteilungstyp (Ausreißer in begrenztem Maße erlaubt): **Wilcoxon Test**  
oder: Adaptiver Test (von SAS nicht angeboten)

## Allgemeine Empfehlungen (2)

- keine Normalverteilung oder unbekannte Verteilung, ungleiche Varianzen, und etwa gleicher Verteilungstyp (Ausreißer in begrenztem Maße erlaubt)  
 $n_1 \approx n_2$  oder ( $n_1 > n_2, \sigma_1 < \sigma_2$ ): [Wilcoxon Test](#)
- keine Normalverteilung oder unbekannte symmetrische Verteilung, ungleiche Varianzen, nicht zu kleine Stichprobenumfänge: [Fligner-Policello Test](#)
- keine Normalverteilung, Verteilungstypen verschieden, ungleiche Varianzen (kleine Varianz zu kleinem Stichprobenumfang): [K-S Test](#)  
oder: Brunner-Munzel Test (von SAS selbst nicht angeboten, kann aber heruntergeladen werden (TSP.sas):

# Inhalt

- Einführung
- Einstichprobenproblem
- Zweistichprobenproblem
- Mehrere unverbundene Stichproben
- Mehrere verbundene Stichproben

## Modell:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2) \quad \text{unabhängig, } i = 1, \dots, k, j = 1, \dots, n_i.$$

$$H_0 : \mu_1 = \dots = \mu_k \quad H_1 : \exists (i_1, i_2) \mu_{i_1} \neq \mu_{i_2}$$

Wir fassen alle Beobachtungen  $X_{11}, \dots, X_{1n_1}, \dots, X_{k1}, \dots, X_{kn_k}$  zusammen und bilden die Rangzahlen  $R_{ij}$ ,  $i = 1 \dots k, j = 1 \dots n_i$ .

Mit den Rangzahlen führen wir eine  
einfaktorielle Varianzanalyse durch  
= Kruskal-Wallis Test

## Kruskal-Wallis Test

$$KW = \frac{\sum_{i=1}^k (T_i - E_0(T_i))^2 \cdot n_i}{S^2 / (N - 1)} = \frac{12 \sum_{i=1}^k (T_i - E_0(T_i))^2 \cdot n_i}{N(N + 1)},$$

wobei  $T_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$  mittlere Rangsumme der i-ten Gruppe

## Kruskal-Wallis Test

$$KW = \frac{\sum_{i=1}^k (T_i - E_0(T_i))^2 \cdot n_i}{S^2 / (N - 1)} = \frac{12 \sum_{i=1}^k (T_i - E_0(T_i))^2 \cdot n_i}{N(N + 1)},$$

wobei  $T_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$  mittlere Rangsumme der i-ten Gruppe

Kruskal-Wallis

Varianzanalyse

 $T_i$  $\bar{Y}_i$  $E_0 T_i = \frac{N+1}{2}$  $\bar{Y}_{..} = \bar{Y}$ 

Zähler

SSB

 $S^2 = \frac{(N-1)N(N+1)}{12} = \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2$  SST

 $N = \sum_{i=1}^k n_i$  Gesamtstichprobenumfang

# Kruskal-Wallis-Test (2)

$$S^2 = \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2 =$$

# Kruskal-Wallis-Test (2)

$$S^2 = \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2 = \sum_{k=1}^N (k - \frac{N+1}{2})^2$$

# Kruskal-Wallis-Test (2)

$$\begin{aligned} S^2 &= \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2 = \sum_{k=1}^N (k - \frac{N+1}{2})^2 \\ &= \sum_k k^2 - (N+1) \sum_k k + \frac{(N+1)^2}{4} \cdot N \end{aligned}$$

## Kruskal-Wallis-Test (2)

$$\begin{aligned} S^2 &= \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2 = \sum_{k=1}^N (k - \frac{N+1}{2})^2 \\ &= \sum_k k^2 - (N+1) \sum_k k + \frac{(N+1)^2}{4} \cdot N \\ &= \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{(N+1)^2}{4} \cdot N \\ &= \frac{(N+1) \cdot N}{12} (4N+2 - 6N - 6 + 3N + 3) \\ &= \frac{N(N+1)}{12} \cdot (N-1) = \frac{(N-1) \cdot N \cdot (N+1)}{12}. \end{aligned}$$

## Kruskal-Wallis-Test (3)

Vorteil:  $S^2$  ist nicht zufällig, hängt nur vom Stichprobenumfang ab.

$KW \sim \chi_{k-1}^2$  (asymptotisch)

$H_0$  ablehnen, falls p-value = "Pr > Chi Square" <  $\alpha$

### SAS-Output

Mean Score:  $T_i$

Chi-Square: realisierte  $KW$

DF= $k - 1$ : Freiheitsgrade.

[Npar1way\\_Maschinen.sas](#)

[PI12erg.sas](#)

## Kruskal-Wallis-Test (4)

- Bei Bindungen erfolgt eine Korrektur der Statistik
- KW-Test ist relativ effizient bei Normalverteilung. Bei Nicht-Normalverteilung meist besser als der Varianzanalyse-F-Test.
- KW-Test hält (wie alle nichtparametrischen Tests) asymptotisch das Signifikanzniveau ein.
- kleine Stichproben ( $N \leq 20$ ): Option EXACT möglich

```
PROC NPAR1WAY WILCOXON;
```

```
  CLASS Faktor;
```

```
  VAR var;
```

```
  EXACT Wilcoxon; RUN;
```

# Inhalt

- Einführung
- Einstichprobenproblem
- Zweistichprobenproblem
- Mehrere unverbundene Stichproben
- Mehrere verbundene Stichproben

# Friedman Test

Modell, wie bei der 2-faktoriellen Varianzanalyse

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2) \quad \text{unabhängig}$$

$$j = 1, \dots, k, i = 1, \dots, n$$

$$H_0 : \beta_1 = \dots = \beta_k (= 0) \quad H_1 : \exists (j_1, j_2) : \beta_{j_1} \neq \beta_{j_2}$$

Ränge werden zeilenweise gebildet,  $Y_{1(1)} \leq \dots \leq Y_{1(k)}$

$R_{ij}$  der Rang von  $Y_{ij}$  in der  $i$ -ten Zeile.

# Nichtparametrische Tests

## Friedman Test (2)

Block	Behandlung				Zeilensumme
	1	2	...	k	
1	$R_{11}$	$R_{12}$	...	$R_{1k}$	$\frac{k(k+1)}{2}$
.					
.					
$n$	$R_{n1}$	$R_{n2}$	...	$R_{nk}$	$\frac{k(k+1)}{2}$
	$R_{.1}$	$R_{.2}$	...	$R_{.k}$	$\frac{nk(k+1)}{2}$
	$n\bar{R}_{.1}$	$n\bar{R}_{.2}$	...	$n\bar{R}_{.k}$	

# Nichtparametrische Tests

## Friedman Test (3)

$$F_k = \frac{n^2 \sum_{j=1}^k (\bar{R}_{.j} - E(\bar{R}_{.j}))^2}{n \cdot k(k+1)/12}$$

$\bar{R}_{.j} = \frac{1}{n} \sum_{i=1}^n R_{ij}$  Spaltenmittel der j-ten Spalte (Vergleiche mit  $\bar{Y}_{.j}$ )

$E\bar{R}_{.j} = \frac{1}{n} \cdot \frac{n(k+1)}{2} = \frac{k+1}{2}$  (Vergleiche mit  $\bar{Y}_{..}$ )

Unter  $H_0$  :  $F_k \sim \chi_{k-1}^2$  (asymptotisch)

$H_0$  ablehnen, falls  $F_k > \chi_{1-\alpha, k-1}^2$

oder falls p-value  $< \alpha$  .

# Nichtparametrische Tests

## Friedman-Test (4)

- Bei Bindungen Korrektur des Nenners.
- Für kleinere  $n$  ist Friedman-Test (asymptotisch) meist etwas konservativ (d.h. der wahre Fehler 1. Art ist kleiner als z.B. 0.05).
- Für größere  $k$  (etwa  $k \geq 5$ ) ist der Friedman-Test (bei Normalverteilung) einigermaßen effizient.
- Für  $k = 2$  ist der Friedman-Test zum Vorzeichentest äquivalent (also nicht besonders effizient).

# Durchführung des Friedman-Tests

**PROC FREQ;**

**TABLES** Faktor A \* Faktor B \* Y

**/CMH2 SCORES=RANK NOPRINT; RUN;**

NOPRINT: unterdrückt den Druck von  
Kontingenztafeln

SCORES=RANK: Ränge werden (zeilenweise)  
gebildet.

CMH2: Cochran-Mantel-Haenszel

[Test\\_Friedman\\_Hypnose.sas](#) [Test\\_Friedman\\_Synchro.sas](#)

Hier ist nur die folgende Zeile interessant:

*Row Mean Scores Differ*

# Übersicht

- 11.1 Korrelation und Unabhängigkeit
- 11.2 Lineare Regression
- 11.3 Nichtlineare Regression
- 11.4 Nichtparametrische Regression
- 11.5 Logistische Regression

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

## Unabhängigkeit und Unkorreliertheit, Wdh.

Die Zufallsvariablen  $X_1, \dots, X_N$  heißen unabhängig, falls für alle  $x_1, \dots, x_N \in \mathbb{R}$

$$P(X_1 < x_1, \dots, X_N < x_N) = P(X_1 < x_1) \cdots P(X_N < x_N)$$

Die Zufallsvariablen  $X_1, \dots, X_N$  heißen unkorreliert, falls

$$\mathbf{E}(X_1 \cdots X_N) = \mathbf{E}(X_1) \cdots \mathbf{E}(X_N).$$

Unabhängigkeit  $\Rightarrow$  Unkorreliertheit

$\nLeftarrow$

Unabhängigkeit  $\Leftrightarrow$  Unkorreliertheit falls  $X_i \sim \mathcal{N}$

# Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale

Seien  $(X_i, Y_i)$ ,  $i = 1, \dots, N$  unabhängige bivariate Zufallsvariablen.

Wir testen

$H_0$  :  $X$  und  $Y$  sind unabhängig (unkorreliert) gegen

$H_1$  :  $X$  und  $Y$  sind linear abhängig (korreliert)

## Pearson-Korrelation

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

$$T = \sqrt{N-2} \cdot \frac{r_{XY}}{\sqrt{1-r_{XY}^2}} \sim t_{N-2}$$

wird in SAS zur Berechnung der  $p$ -Werte verwendet.

# Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale (3)

$H_0$  :  $X$  und  $Y$  sind unabhängig (unkorreliert) gegen

$H_1$  :  $X$  und  $Y$  sind monoton abhängig

## Spearman-Rangkorrelationskoeffizient

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

Weitere Korrelationskoeffizienten: Kendall.

Wenn keine Normalverteilung vorliegt, so Spearman oder Kendall nehmen!

# Korrelation und Unabhängigkeit

a) Metrisch skalierte Merkmale

```
PROC CORR PEARSON SPEARMAN KENDALL;
```

```
VAR vars;
```

```
RUN;
```

# Korrelation und Unabhängigkeit

a) Metrisch skalierte Merkmale

```
PROC CORR PEARSON SPEARMAN KENDALL;  
    VAR vars;  
RUN;
```

b) Ordinal oder nominal skalierte Merkmale

```
PROC FREQ;  
    TABLES var1*var2 / CHISQ;  
RUN;
```

Descr\_Scatter.sas    Descr\_Scatter\_Heroin.sas

# Ordinal oder nominal skalierte Merkmale

Frage: Bestehen Abhängigkeiten?

Geschlecht - Studienfach      Studiengang - Note

Geburtsmonat - IQ

Antwort:  $\chi^2$  - Unabhängigkeitstest (Pearson, 1908)

Annahme:

$X$  hat Ausprägungen  $a_1, \dots, a_m$

$Y$  hat Ausprägungen  $b_1, \dots, b_l$

(sind die Daten metrisch, so wird automatisch eine Klasseneinteilung vorgenommen.)

$$P(X = a_i) = p_i, \quad P(Y = b_j) = p_j, \quad P(X = a_i, Y = b_j) = p_{ij}$$

# Häufigkeitstabelle (= Kontingenztafel)

$X Y$	$b_1$	$b_2$	$\dots$	$b_j$	$\dots$	$b_l$	
$a_1$	$h_{11}$	$h_{12}$	$\dots$	$h_{1j}$	$\dots$	$h_{1l}$	$h_{1.}$
$a_2$	$h_{21}$	$h_{22}$	$\dots$	$h_{2j}$	$\dots$	$h_{2l}$	$h_{2.}$
$\dots$							
$a_i$	$h_{i1}$	$h_{i2}$	$\dots$	$h_{ij}$	$\dots$	$h_{il}$	$h_{i.}$
$\dots$							
$a_m$	$h_{m1}$	$h_{m2}$	$\dots$	$h_{mj}$	$\dots$	$h_{ml}$	$h_{m.}$
	$h_{.1}$	$h_{.2}$	$\dots$	$h_{.j}$	$\dots$	$h_{.l}$	$h_{..} = n$

$h_{ij}$ : Häufigkeiten

# Unabhängigkeitstests

Die Häufigkeiten  $h_{ij}$  werden verglichen mit den theoretischen Häufigkeiten  $np_{ij}$ .

$$H_0: p_{ij} = p_{i.} \cdot p_{.j}, \quad i = 1, \dots, m, j = 1, \dots, l$$

$$H_1: p_{ij} \neq p_{i.} \cdot p_{.j}, \quad \text{für ein Paar}(i, j)$$

$H_0$ :  $X$  und  $Y$  sind unabhängig.

$H_1$ :  $X$  und  $Y$  sind abhängig.

Betrachten zunächst die Stichprobenfunktion

$$\tilde{T} = \sum_i \sum_j \frac{(h_{ij} - np_{ij})^2}{np_{ij}}$$

# Unabhängigkeitstests

## Konstruktion der Teststatistik

Unter  $H_0$  ( $X$  und  $Y$  unabhängig) gilt:  $p_{ij} = p_{i.} \cdot p_{.j}$ .

Problem:  $p_{i.}$  und  $p_{.j}$  sind unbekannt. Sie müssen also geschätzt werden,

das sind  $m + l - 2$  Parameter ( $\sum p_{i.} = \sum p_{.j} = 1$ )

$$\hat{p}_{i.} = \frac{h_{i.}}{n} \quad \hat{p}_{.j} = \frac{h_{.j}}{n}$$

$$h_{i.} = \sum_{j=1}^l h_{ij} \quad h_{.j} = \sum_{i=1}^m h_{ij}$$

# Unabhängigkeitstests

Einsetzen der Schätzungen in  $\tilde{T}$

$$Q_P = \sum_i \sum_j \frac{(h_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}}$$

# Unabhängigkeitstests

Einsetzen der Schätzungen in  $\tilde{T}$

$$\begin{aligned} Q_P &= \sum_i \sum_j \frac{(h_{ij} - n\hat{p}_i \cdot \hat{p}_j)^2}{n\hat{p}_i \cdot \hat{p}_j} \\ &= n \sum_i \sum_j \frac{(h_{ij} - \frac{h_i \cdot h_j}{n})^2}{h_i \cdot h_j} \\ &\sim \chi_{(m-1)(l-1)}^2 \quad \text{approximativ, unter } H_0 \end{aligned}$$

# Unabhängigkeitstests

Einsetzen der Schätzungen in  $\tilde{T}$

$$\begin{aligned}
 Q_P &= \sum_i \sum_j \frac{(h_{ij} - n\hat{p}_i \cdot \hat{p}_j)^2}{n\hat{p}_i \cdot \hat{p}_j} \\
 &= n \sum_i \sum_j \frac{(h_{ij} - \frac{h_i \cdot h_j}{n})^2}{h_i \cdot h_j} \\
 &\sim \chi_{(m-1)(l-1)}^2 \quad \text{approximativ, unter } H_0
 \end{aligned}$$

Die Anzahl der Freiheitsgrade ergibt sich aus:

$$m \cdot l - 1 - \underbrace{(m + l - 2)}_{\text{\#geschätzte Werte}}$$

$H_0$  ablehnen, falls

$$Q_P > \chi_{(m-1)(l-1)}^2, \quad \text{bzw. falls p-Wert} < \alpha$$

# Korrelation und Unabhängigkeit

Faustregel für die Anwendung des  $\chi^2$ -Unabhängigkeitstests:

- alle  $h_{ij} > 0$ .
- $h_{ij} \geq 5$  für mindestens 80% der Zellen, sonst Klassen zusammenfassen.

[Descr\\_Freq\\_Heroin\\_Unabhaengigkeitstest](#)

# Weitere Unabhängigkeitstests (1)

- LQ- $\chi^2$ - Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{nh_{ij}}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

# Weitere Unabhängigkeitstests (1)

- LQ- $\chi^2$ - Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{nh_{ij}}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Continuity Adjusted  $\chi^2$  (bei SAS nur: 2x2-Tafel)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.}h_{.j}}{N}| - 0.5)^2}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

## Weitere Unabhängigkeitstests (1)

- LQ- $\chi^2$ - Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{nh_{ij}}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Continuity Adjusted  $\chi^2$  (bei SAS nur: 2x2-Tafel)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.}h_{.j}}{N}| - 0.5)^2}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Mantel-Haenszel ( $r_{XY}$ : Pearson-Korrelation)

$$Q_{MH} = (N - 1)r_{XY}^2 \sim \chi^2_1$$

# Weitere Unabhängigkeitstests (1)

- LQ- $\chi^2$ - Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{nh_{ij}}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Continuity Adjusted  $\chi^2$  (bei SAS nur: 2x2-Tafel)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.}h_{.j}}{N}| - 0.5)^2}{h_{i.}h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Mantel-Haenszel ( $r_{XY}$ : Pearson-Korrelation)

$$Q_{MH} = (N - 1)r_{XY}^2 \sim \chi^2_1$$

- Phi-Koeffizient

$$\Phi = \begin{cases} \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1.}h_{2.}h_{.1}h_{.2}}} & m = l = 2 \\ \sqrt{Q_p/n} & \text{sonst} \end{cases}$$

## Weitere Unabhängigkeitstests (2)

- Kontingenzkoeffizient

$$P = \sqrt{\frac{Q_P}{Q_P + n}}$$

- Fishers Exact Test (bei 2x2-Tafeln)

durch Auszählen aller Tafel-Möglichkeiten bei gegebenen Rändern.  
(gilt als etwas konservativ.)

- Cramers  $V$

$$V = \begin{cases} \Phi & \text{falls } 2 \times 2 \text{ Tafel} \\ \sqrt{\frac{Q_P/n}{\min(m-1, l-1)}} & \text{sonst} \end{cases}$$

# Weitere Unabhängigkeitstests (3)

## Anmerkungen

- Mantel- Haenszel Test verlangt ordinale Skalierung, vgl.  $(n - 1)r_{XY}^2$  'gut' gegen lineare Abhängigkeit.
- Der  $\chi^2$  Unabhängigkeitstest testet gegen allgemeine Abhängigkeit.
- Der LQ-Test  $G^2$  ist plausibel und geeignet.
- Der LQ-Test  $G^2$  und der  $\chi^2$  Unabhängigkeitstest sind asymptotisch äquivalent.

# Unabhängigkeitstests

## $\Phi$ -Koeffizient (2x2 Tafel)

Y \ X	Sportler	Nichtsportler	Summe
w	$p_{11}$	$p_{12}$	$p_{1.}$
m	$p_{21}$	$p_{22}$	$p_{2.}$
Summe	$p_{.1}$	$p_{.2}$	1

$$X \sim \text{Bi}(1, p_{.2}) \quad Y \sim \text{Bi}(1, p_{2.})$$

$$\mathbf{E}(X) = p_{.2} \quad \text{var}(X) = p_{.2}(1 - p_{.2}) = p_{.2}p_{.1}$$

$$\mathbf{E}(Y) = p_{2.} \quad \text{var}(Y) = p_{2.}(1 - p_{2.}) = p_{2.}p_{1.}$$

# Unabhängigkeitstests

## Korrelationskoeffizient in einer 2x2 Tafel

$$\rho = \frac{p_{22} - p_{.2}p_{2.}}{\sqrt{p_{.2}p_{1.}p_{2.}p_{.1}}} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{.2}p_{2.}p_{1.}p_{.1}}}$$

$$\begin{aligned} p_{22} - p_{.2}p_{2.} &= p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) \\ &= p_{22} - (p_{21}p_{12} + p_{22}p_{12} + p_{21}p_{22} + p_{22}^2) \\ &= p_{22}(1 - p_{12} - p_{21} - p_{22}) - p_{21}p_{12} \\ &= p_{22}p_{11} - p_{21}p_{12} \end{aligned}$$

Für  $m = l = 2$  ist der Phi-Koeffizient eine Schätzung des Korrelationskoeffizienten.

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

## Einfache lineare Regression (vgl. Kap. 6.3)

$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i \quad \epsilon_i \sim (0, \sigma^2)$$

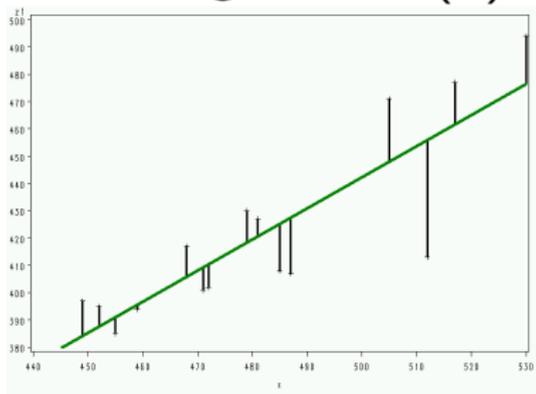
$$\hat{\theta}_1 = \frac{S_{XY}}{S_X^2}$$

$$\hat{\theta}_0 = \frac{1}{n} \left( \sum Y_i - \hat{\theta}_1 \sum X_i \right) = \bar{Y} - \hat{\theta}_1 \bar{X}$$

als Lösung der Minimumaufgabe

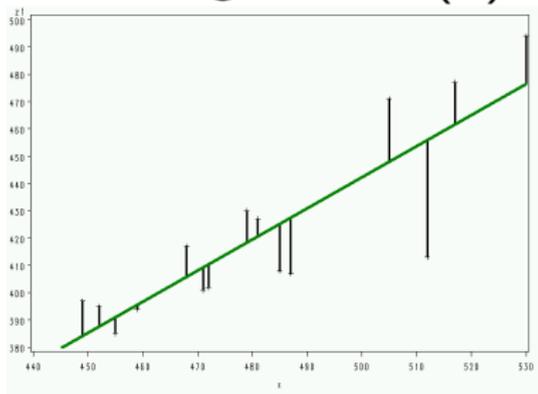
$$\sum_{i=1}^n (Y_i - \theta_1 X_i - \theta_0)^2 \rightarrow \min.$$

## Lineare Regression (2)



Die Summe der Quadrate  
der Länge der  
Streckenabschnitte soll  
minimal werden.

## Lineare Regression (2)



Die Summe der Quadrate  
der Länge der  
Streckenabschnitte soll  
minimal werden.

$$S_{XY} = \frac{1}{n-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_X^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

Regression\_Venusmuscheln

Regression\_Plot

# Lineare Regression (3)

```
PROC REG;
```

```
    MODEL y = x1 / Optionen;
```

```
RUN;
```

## Multiple lineare Regression

### Modell

$$Y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \dots + \theta_m x_{mi} + \epsilon_i$$

$$Y_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \dots + \theta_m X_{mi} + \epsilon_i$$

$Y_i, \epsilon_i$  Zufallsvariablen, unabhängig,  $\epsilon_i \sim (0, \sigma^2)$ ,  $i = 1 \dots n$

$\theta_0 \dots \theta_m, \sigma$  : Modellparameter  $\Rightarrow$  zu schätzen

Man unterscheidet Fälle:

$x_i = (x_{1i}, \dots, x_{mi})$  fest, und  $X_i = (X_{1i}, \dots, X_{mi})$  zufällig

oder auch gemischt.

Matrix-Schreibweise:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

# Multiple lineare Regression (2)

## Modell

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1m} \\ \cdot & \cdot & \dots & \cdot \\ 1 & X_{n1} & \dots & X_{nm} \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \dots \\ \theta_m \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

Methode der kleinsten Quadrate: Bestimme  $\hat{\boldsymbol{\theta}}$  so daß

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$$

## Multiple lineare Regression (2a)

Ableiten von  $(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$  nach  $\boldsymbol{\theta}$  und Nullsetzen liefert:

$$2(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'\mathbf{X} = 0$$

$$\mathbf{Y}'\mathbf{X} - \boldsymbol{\theta}'\mathbf{X}'\mathbf{X} = 0$$

$$\mathbf{Y}'\mathbf{X} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\theta}$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Kleinste Quadrat Schätzung.

# Multiple lineare Regression (2a)

Ableiten von  $(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$  nach  $\boldsymbol{\theta}$  und Nullsetzen liefert:

$$2(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'\mathbf{X} = 0$$

$$\mathbf{Y}'\mathbf{X} - \boldsymbol{\theta}'\mathbf{X}'\mathbf{X} = 0$$

$$\mathbf{Y}'\mathbf{X} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\theta}$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$2\underbrace{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})'}_{\hat{\boldsymbol{\epsilon}}}\mathbf{X} = \mathbf{0}$$

$$\hat{\boldsymbol{\epsilon}}'\mathbf{X} = \mathbf{0} \quad \text{insbesondere}$$

$$\sum_{i=1}^n \hat{\epsilon}_i = 0$$

$$\sum_{i=1}^n \hat{\epsilon}_i X_{ij} = 0 \quad \forall j$$

Kleinste Quadrat Schätzung.

# Multiple lineare Regression (3)

## Kleinste Quadrat-Schätzung

Voraussetzung:  $\text{rg}(\mathbf{X}'\mathbf{X}) = m$  (voll)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

wenn  $(\mathbf{X}'\mathbf{X})$  nicht regulär: verallgemeinerte Inverse  
(Moore-Penrose)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

# Multiple lineare Regression (4)

## Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (1)

$$(\mathbf{X}'\mathbf{X})^{-1} = \left[ \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{11} & \cdot & \dots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ \dots & \dots \\ 1 & X_{n1} \end{pmatrix} \right]^{-1}$$

# Multiple lineare Regression (4)

## Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (1)

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1} &= \left[ \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{11} & \cdot & \dots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ \dots & \dots \\ 1 & X_{n1} \end{pmatrix} \right]^{-1} \\ &= \begin{pmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{pmatrix}^{-1} \quad (X_i = X_{1i})\end{aligned}$$

# Multiple lineare Regression (4)

## Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (1)

$$\begin{aligned}
 (\mathbf{X}'\mathbf{X})^{-1} &= \left[ \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{11} & \cdot & \dots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ \dots & \dots \\ 1 & X_{n1} \end{pmatrix} \right]^{-1} \\
 &= \begin{pmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{pmatrix}^{-1} \quad (X_i = X_{1i}) \\
 &= \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}
 \end{aligned}$$

# Multiple lineare Regression (5)

Kleinste Quadrat-Schätzung, Spezialfall  $m = 1$  (2)

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & . & \dots & X_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

# Multiple lineare Regression (5)

Kleinste Quadrat-Schätzung, Spezialfall  $m = 1$  (2)

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & . & \dots & X_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

$$\begin{aligned} \hat{\theta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i \\ -\sum X_i \sum Y_i + n \sum X_i Y_i \end{pmatrix} \end{aligned}$$

## Multiple lineare Regression (6)

Schätzung für  $\mathbf{Y}$ :  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$

Vergleiche mit  $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$

Einsetzen von  $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ :

$$\begin{aligned}\hat{\mathbf{Y}} &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y} \\ &= \mathbf{H}'\mathbf{Y}\end{aligned}$$

**H**: Hat-Matrix

Aus dem Beobachtungsvektor  $\mathbf{Y}$  wird der geschätzte Beobachtungsvektor  $\hat{\mathbf{Y}}$ .

# Multiple Lineare Regression (7)

Quadratsummenaufspaltung:

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{SSM} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SSE}$$

$MST = \frac{1}{n-1} SST$ : Schätzung für die Gesamtvarianz.

$MSE = \frac{1}{n-m-1} SSE = \hat{\sigma}^2$ . (erwartungstreu)

$MSM = \frac{1}{m} SSM$  ( $m + 1$  Einflussvariablen)

Bestimmtheitsmaß (wie bei der Varianzanalyse)

$$R^2 = \frac{SSM}{SST}$$

# Multiple Lineare Regression (7a)

Quadratsummenaufspaltung:  $\sum_i (Y_i - \bar{Y})^2 =$

$$= \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (Y_i - \bar{Y})^2 + 2 \sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

Der letzte Summand  $\sum_i \underbrace{(Y_i - \hat{Y}_i)}_{\hat{\epsilon}_i} (\hat{Y}_i - \bar{Y})$  ist

$$= \sum_i \hat{\epsilon}_i (\mathbf{x}\hat{\theta})_i - \bar{Y} \underbrace{\sum_i \hat{\epsilon}_i}_{=0} = \sum_i \hat{\epsilon}_i \sum_j x_{ij} \hat{\theta}_j$$

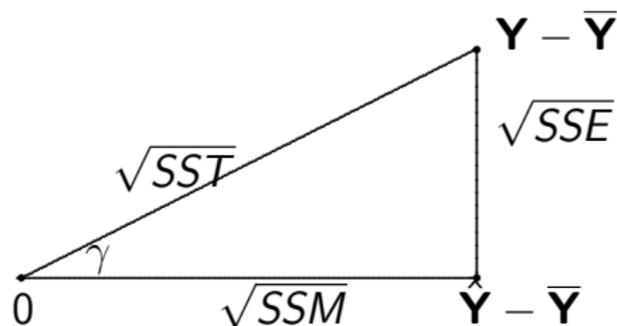
$$= \sum_j \hat{\theta}_j \underbrace{\sum_i \hat{\epsilon}_i x_{ij}}_{=0 \quad \forall j} = 0$$

# Geometrische Veranschaulichung

$$\mathbf{Y} = (Y_{11}, \dots, Y_{kn_k}) \quad \text{Dimension } N$$

$$\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_{kn_k})$$

$$\bar{\mathbf{Y}} = (\underbrace{\bar{Y}, \dots, \bar{Y}}_{n \text{ mal}}), \quad \bar{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$$



$$SSM + SSE = SST$$

$$R^2 = \cos^2 \gamma$$

$$\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$$

## Multiple Lineare Regression (8)

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_m = 0 \quad H_1 : \sim H_0$$

Unter der Annahme  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  gilt:

$$F = \frac{SSM}{SSE} \cdot \frac{n - m - 1}{m} \sim F_{m, n-m-1}$$

**PROC REG;**

**MODEL** y = x1 x2 x3 / Optionen;

**TEST** x2=0 x3=0; /\*zusaeztl. Hypothesen\*/

**RUN;**

Regression\_Tibetan    Regression\_Phosphor

# Multiple Lineare Regression (9)

Zusätzliche Hypothesen, z.B.

$$H_{0a} : \theta_1 = 0 , \quad H_{1a} : \theta_1 \neq 0$$

$$H_{0b} : \theta_k = 0 , \quad H_{1b} : \theta_k \neq 0$$

$$H_{0c} : \theta_1 = \theta_2 = 0 , \quad H_{1c} : \theta_1 \neq 0 \vee \theta_2 \neq 0$$

# Multiple Lineare Regression (10)

$R^2$ -adjustiert für Anzahl  $p$  der Parameter im Modell

$$Adj\_R^2 = 1 - \frac{n - i}{n - p}(1 - R^2)$$

$$\begin{cases} i = 0 & \text{ohne intercept} \\ i = 1 & \text{mit intercept} \end{cases}$$

Dependent Mean: Mittelwert der abhängigen Variable ( $Y$ )

StdError MeanPredict: Standardfehler für vorhergesagten Erwartungswert

# Multiple Lineare Regression (11)

## Optionen (Auswahl)

XPX:	Ausgabe der Matrizen <b><math>\mathbf{X}'\mathbf{X}</math>, <math>\mathbf{X}'\mathbf{Y}</math>, <math>\mathbf{Y}'\mathbf{Y}</math></b>
I:	Ausgabe der Inversen von <b><math>\mathbf{X}'\mathbf{X}</math></b>
COVB:	Schätzung der Kovarianzmatrix der Schätzung = $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$
CLM, CLI:	Konfidenzbereiche (s.u.)
CLB:	Konfidenzintervall für Parameter $\theta$
R:	studentisierte Residuen (s.u.)
DW:	Durbin-Watson "Test" auf Autokorrelation (s.u.)

## Output Statistics (Optionen CLI, CLM, R)

Dependent Variable	$Y_i$
Predicted Value	$\hat{Y}_i = (\hat{\theta}\mathbf{X})_i$
StdErrorMeanPredict	$\hat{\sigma}_{\hat{Y}_i}$
95% CL Mean (s.u.)	nur Variabilität in Parameter-schätzung berücksichtigt
95% CL Predict (s.u.)	Variabilität im Fehlerterm mit berücksichtigt
Residual	$e_i = Y_i - \hat{Y}_i$
StdErrorResidual	$s\sqrt{1 - h_{ii}}$ , s.u.
Student Residual	$r_i$ s.u.
Cook's $D_i$	s.u.
Predicted Residual SS	s.u.

# Lineare Regression

## Multiple Linear Regression (13)

Konfidenzintervalle für allgemeinen Parameter  $\vartheta_j$ :

$$\frac{\hat{\vartheta}_j - \vartheta_j}{s_{\hat{\vartheta}_j}} \sim t_{n-1} \quad \text{Vor. } \epsilon_j \sim \mathcal{N}(0, \sigma^2)$$

$$\text{KI: } [\hat{\vartheta}_j - t_{1-\frac{\alpha}{2}, n-1} \cdot s_{\hat{\vartheta}_j}, \hat{\vartheta}_j + t_{1-\frac{\alpha}{2}, n-1} \cdot s_{\hat{\vartheta}_j}]$$

# Lineare Regression

## Multiple Lineare Regression (13)

Konfidenzintervalle für allgemeinen Parameter  $\vartheta_i$ :

$$\frac{\hat{\vartheta}_i - \vartheta_i}{s_{\hat{\vartheta}_i}} \sim t_{n-1} \quad \text{Vor. } \epsilon_j \sim \mathcal{N}(0, \sigma^2)$$

$$\text{KI: } [\hat{\vartheta}_i - t_{1-\frac{\alpha}{2}, n-1} \cdot s_{\hat{\vartheta}_i}, \hat{\vartheta}_i + t_{1-\frac{\alpha}{2}, n-1} \cdot s_{\hat{\vartheta}_i}]$$

95% Konfidenzintervall für  $E(Y_i)$

( $\vartheta_i = \mathbf{E}(Y_i)$ , Option CLM)

Nur die Variabilität in der Parameterschätzung wird berücksichtigt.

# Lineare Regression

Multiple Lineare Regression (14)

95% Konfidenzintervall für Vorhersagen  $\underline{Y}_i$

( $\vartheta_i = Y_i$ , Option CLI)

Die Variabilität im Fehlerterm wird mit berücksichtigt.

# Lineare Regression

Multiple Lineare Regression (14)

95% Konfidenzintervall für Vorhersagen  $\underline{Y}_i$

( $\vartheta_i = Y_i$ , Option CLI)

Die Variabilität im Fehlerterm wird mit berücksichtigt.

95% Konfidenzintervall für  $\theta$

( $\vartheta_i = \theta_j$ , Option CLB)

# Lineare Regression

Multiple Lineare Regression (14)

95% Konfidenzintervall für Vorhersagen  $\underline{Y}_i$

( $\vartheta_i = Y_i$ , Option CLI)

Die Variabilität im Fehlerterm wird mit berücksichtigt.

95% Konfidenzintervall für  $\theta$

( $\vartheta_i = \theta_j$ , Option CLB)

Darstellung von Konfidenzbereichen bei der einfachen  
Regressionsanalyse

```
SYMBOL I=RLCLI95;
```

```
PROC GPLOT;
```

# Residualanalyse (1)

## Studentisierte Residuen (Option R)

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

$e_i = y_i - \hat{y}_i$  (Residuen) sind korreliert,

$\text{var } e_i = \sigma^2(1 - h_{ii})$        $s = \hat{\sigma}$

# Residualanalyse (1)

## Studentisierte Residuen (Option R)

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

$e_i = y_i - \hat{y}_i$  (Residuen) sind korreliert,  
 $\text{var } e_i = \sigma^2(1-h_{ii})$        $s = \hat{\sigma}$

## Cook's $D_i$

$$D_i = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})}{(m+1)S^2}, \quad i = 1 \dots n$$

beschreibt den Einfluß der  $i$ -ten Beobachtung auf die Parameterschätzung

$\hat{\boldsymbol{\theta}}_{(i)}$ : KQS von  $\boldsymbol{\theta}$  ohne Beobachtung  $i$ .

Faustregel:  $D_i > 1 \rightarrow$  'starker' Einfluß

# Residualanalyse (2)

## Predicted Residual SS (PRESS)

$$\sum (y_i - \hat{y}_{i(i)})^2$$

$\hat{y}_{i(i)}$ :  $i$ -te Beobachtung weggelassen.

## Residualanalyse (2)

### Predicted Residual SS (PRESS)

$$\sum (y_i - \hat{y}_{i(i)})^2$$

$\hat{y}_{i(i)}$ :  $i$ -te Beobachtung weggelassen.

### “Test” auf Autokorrelation: Durbin-Watson-Test (Option DW)

$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

DW=2: Unkorreliertheit der Residuen

## Weitere Bewertung der Residuen

Kommando PLOT in der Prozedur REG

```
PLOT rstudent.*obs.;
```

```
PLOT residual.*y    residual.*predicted.;
```

```
OUTPUT OUT=dateiname RESIDUAL=;
```

und evtl. Test auf Normalverteilung.

rstudent. : studentisierte Residuen

residual. : Residuen

obs : Beobachtungsnummer

y : beobachteter Wert von  $Y$

predicted. : geschätzter Wert von  $Y$ :  $\hat{Y}$

Anmerkung: wenn Sie ODS graphics on gesetzt haben, kommen eine Reihe der o.g. Plots automatisch.

# Multiple Lineare Regression (Fortsetzung)

## Modellwahl in der linearen Regression

SELECTION=

BACKWARD: Alle Variablen, die mit größten p-Wert werden nacheinander herausgenommen (min. p-Wert: SLSTAY [=0.1])

FORWARD: Start ohne Variablen, die Var. mit kleinstem p-Wert kommt hinzu (max. p-Wert: SLENTY[= 0.5])

STEPWISE: Start ohne Variable, 1.Schritt wie bei FORWARD (Standard: SLENTY = 0.15), Variablen können wieder eliminiert werden (Standard: SLSTAY=0.1)

MAXR: Für jeweils eine feste Anzahl von Variablen wird das Modell mit max.  $R^2$  ausgegeben.

# Multiple Lineare Regression (Fortsetzung)

a) Wenn  $rg(\mathbf{X}'\mathbf{X})$  nicht voll ( $< m + 1$ )

$\Rightarrow (\mathbf{X}'\mathbf{X})^{-}$  und Anmerkung im Output

b) Condition number

$\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$   $\lambda_{max}, \lambda_{min}$  größter und kleinster Eigenwert von  $\mathbf{X}'\mathbf{X}$  (ohne 1-Spalte).

große Konditionszahl (etwa  $> 30$ ): schlechte Kondition ( $\approx$  lineare Abhängigkeit)

c)  $C(p)$ : Mallows (1973) Kriterium für die Modellwahl

$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

$SSE_p$ : SSE im Modell mit  $p$  Parametern

# Multiple Lineare Regression (Fortsetzung)

Modellwahl in der linearen Regression

$$R^2 = \frac{SSM}{SST}$$
$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

$SSE_p$ : SSE im Modell mit  $p$  Parametern

Ziel:  $R^2$  groß,  $C(p)$  nahe  $p$

Idee von  $C(p)$ : Wenn die Wahl von  $p$  Parametern gut, dann

$$MSE \approx MSE_p = \frac{SSE_p}{n-p} \Rightarrow C(p) \approx n - p - n + 2p = p$$

Regression\_Tibetan\_Modellwahl

Einfache Varianzanalyse:  $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ 

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ \dots \\ Y_{n_1 1} \\ Y_{12} \\ \dots \\ Y_{n_2 2} \\ \dots \\ Y_{1k} \\ \dots \\ Y_{n_k k} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & 1 & \dots & 0 \\ \dots & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \dots \\ \epsilon_{n_k k} \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

# Multiple Linear Regression (Fortsetzung)

$$\begin{pmatrix} Y_1 \\ \dots \\ \dots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ \cdot & \dots & & \dots \\ \cdot & \dots & & \dots \\ 1 & X_{N1} & \dots & X_{Np} \end{pmatrix} \begin{pmatrix} \mu \\ \theta_1 \\ \dots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \dots \\ \dots \\ \epsilon_N \end{pmatrix} \Leftrightarrow$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

## Mögliche Probleme bei der linearen Regression

### Probleme

- Ausreißer
- keine Normalverteilung
- kein linearer Zusammenhang
- Zielvariable nicht stetig

## Mögliche Probleme bei der linearen Regression

### Probleme

- Ausreißer
- keine Normalverteilung
- kein linearer Zusammenhang
- Zielvariable nicht stetig

### Lösungsansätze

Robuste Lineare Regression

## Mögliche Probleme bei der linearen Regression

### Probleme

- Ausreißer
- keine Normalverteilung
- kein linearer Zusammenhang
- Zielvariable nicht stetig

### Lösungsansätze

Robuste Lineare Regression  
Datentransformation,  
( $L_1$ -Regression)

## Mögliche Probleme bei der linearen Regression

### Probleme

- Ausreißer
- keine Normalverteilung
- kein linearer Zusammenhang
- Zielvariable nicht stetig

### Lösungsansätze

- Robuste Lineare Regression
- Datentransformation,  
( $L_1$ -Regression)
- Nichtlineare Regression

## Mögliche Probleme bei der linearen Regression

### Probleme

- Ausreißer
- keine Normalverteilung
- kein linearer Zusammenhang
- Zielvariable nicht stetig

### Lösungsansätze

- Robuste Lineare Regression
- Datentransformation, ( $L_1$ -Regression)
- Nichtlineare Regression
- Nichtparametrische Regression

## Mögliche Probleme bei der linearen Regression

Probleme	Lösungsansätze
• Ausreißer	Robuste Lineare Regression
• keine Normalverteilung	Datentransformation, ( $L_1$ -Regression)
• kein linearer Zusammenhang	Nichtlineare Regression Nichtparametrische Regression
• Zielvariable nicht stetig	Logistische Regression

# Robuste Lineare Regression (Skizze)

Ausreißer können auftreten in

- Y-Richtung
- X-Richtung(en) (Leverage points)
- Y- und X- Richtungen

Fall: Ausreißer(verdacht) in Y-Richtung:

es werden nicht die Abstandskquadrate minimiert, sondern (z.B.) die Gewichtsfunktion (Bisquare Biweight, Huber,  $c=4.685$ , Voreinstellung bei SAS)

$$W(x, c) = \begin{cases} 1 - \left(\frac{x}{c}\right)^2 & \text{falls } |x| < c \\ 0 & \text{sonst.} \end{cases}$$

verwendet.

## Robuste Lineare Regression (2)

Außerdem wird der Skalenparameter  $\sigma$  nicht durch  $s$  sondern durch den *MAD* geschätzt.

```
PROC ROBUSTREG;
```

```
MODEL y=x1 x2 x3/DIAGNOSTICS LEVERAGE;
```

```
RUN;
```

```
Regression_Phosphor
```

# Robuste Lineare Regression (3)

## Diagnosestatistiken

Ausreißer: standardis. robust residual  $>$  cutoff (outlier)

Leverage Point: robuste MCD-Distanz  $>$  cutoff (Leverage)

### Mahalanobis-Distanz

$\approx$  mit Kovarianzmatrix gewichteter mittlerer quadratischer Abstand von  $\bar{X}$ .

### Robust MCD Distance:

anstelle von  $\bar{X}$ : robuste multivariate Lokationsschätzung (MCD)

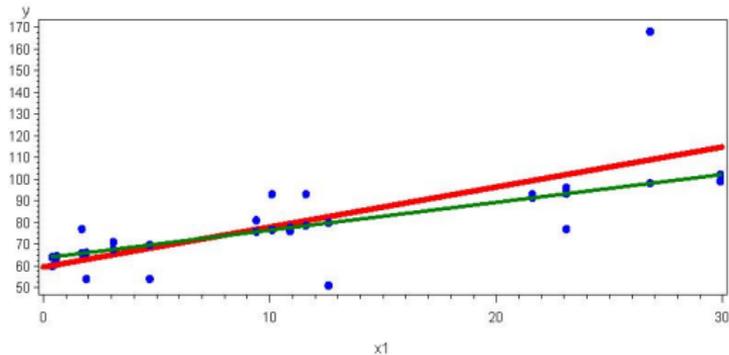
### Goodness of fit: zum Modellvergleich

je größer  $R^2$ , je kleiner AICR, BICR desto besser.

# Robuste Lineare Regression (4)

Beispiel: Phosphorfractionen

## Lineare (rot) vs. Robuste lineare Regression



# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

## Quasilineare Regression

z.B. Polynomregression

$$Y_i = a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + \epsilon_i$$

wird auf lineare Regression zurückgeführt

$$x_{ij} := x_i^j$$

## Quasilineare Regression

z.B. Polynomregression

$$Y_i = a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + \epsilon_i$$

wird auf lineare Regression zurückgeführt

$$x_{ij} := x_i^j$$

## Echt nichtlineare Regression

- Wachstumskurven

$$y = \alpha + \frac{\gamma}{1 + \exp(-\beta(x - \mu))} \quad \text{logistische Fkt.}$$

$$y = \alpha + \gamma \exp(-\exp(-\beta(x - \mu))) \quad \text{Gompertzfkt.}$$

- Wettervohersagemodell ( $Y(t)$ : mittlere Temperatur im Monat  $t$ )

$$Y(t) = aY(t - 6) + bY(t - 12) + \epsilon_t$$

Modell,  $f$  wird als bekannt angenommen

$$Y = f(x, \theta) + \epsilon \quad \epsilon \sim (0, \sigma^2)$$

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \theta) + \epsilon$$

$$L(\theta) = \epsilon' \epsilon = \sum_i (Y_i - \mathbf{F}(\mathbf{X}_i, \theta))^2 \longrightarrow \min_{\theta}$$

Dazu werden Iterationsverfahren verwendet.

Modell,  $f$  wird als bekannt angenommen

$$Y = f(x, \theta) + \epsilon \quad \epsilon \sim (0, \sigma^2)$$

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \theta) + \epsilon$$

$$L(\theta) = \epsilon' \epsilon = \sum_i (Y_i - \mathbf{F}(\mathbf{X}_i, \theta))^2 \longrightarrow \min_{\theta}$$

Dazu werden Iterationsverfahren verwendet.

**PROC NLIN METHOD** = NEWTON;

**MODEL** abhaengige Variable = Ausdruck;

**PARMS** Anfangswerte;

**RUN**;

Andere Methoden: Gauss-Newton, Marquardt, steepest descent

## Nichtlineare Regression (2)

### Ausgabe

R, PPC, RPC: Konvergenzmaßzahlen

Object: Zielfunktionswertänderung (letzte Iteration)

Objective: Zielfunktionswert  $L(\theta)$

Details zu den Iterationsverfahren siehe OnlineDoc.

Parameterschätzungen (mit Iterationsverlauf) und  
Konfidenzintervallen

[Nlin1\\_usapop.sas](#)    [Nlin1\\_usapop\\_est.sas](#)

[Nlin2\\_wind.sas](#)

Anmerkung: Es gibt noch andere Prozeduren, die nichtlineare  
Regressionen durchführen, z.B. **PROC OPTMODEL**

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

Modell:  $f$  unbekannt, aber "glatt"

$$Y_i = f(x_i) + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$  ( $\mathbf{x}_i$  fest oder zufällig)

$$\min_{f \in \mathcal{C}^2} \sum_{i=1}^n (Y_i - f(x_i))^2$$

Modell:  $f$  unbekannt, aber "glatt"

$$Y_i = f(x_i) + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$  ( $\mathbf{x}_i$  fest oder zufällig)

$$\min_{f \in C^2} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

- $\int (f'')^2$ : Strafterm
- $\lambda$ : Glättungsparameter
  - $\lambda \rightarrow 0$ : Interpolierender Spline
  - $\lambda \rightarrow \infty$ : lineare Regression

Modell:  $f$  unbekannt, aber "glatt"

$$Y_i = f(x_i) + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$  ( $\mathbf{x}_i$  fest oder zufällig)

$$\min_{f \in C^2} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

- $\int (f'')^2$ : Strafterm
- $\lambda$ : Glättungsparameter
  - $\lambda \rightarrow 0$ : Interpolierender Spline
  - $\lambda \rightarrow \infty$ : lineare Regression

Lösung der Minimumaufgabe: natürlicher kubischer Spline

## Nichtparametrische Regression (2)

**PROC TPSPLINE;**

**MODEL** abh.Var = (unabhaengige Variablen);

**OUTPUT OUT=**Datei1 **PRED RESID;**

**RUN;**

### Wahl der Glättungsparameter

Kreuzvalidierung (Standard)

vorgeben: **LAMBDA0=Wert**

Es kann eine ganze Liste abgearbeitet werden mit der Option **LOGNLAMBDA** in der **MODEL**-Anweisung, z.B.

**MODEL** y = (x) / **LOGNLAMBDA=-4 to -2 by 0.1;**

# Nichtparametrische Regression (3)

## Ausgabe

$\text{Log}_{10}(n * \hat{\lambda})$

Strafterm  $\int (f'')^2(t) dt$

Residual Sum of Squares

Schätzung für  $\sigma$ ,  $\hat{\sigma}^2 = \frac{RSS}{sp(I-A)}$ ,  $A$ : entspricht der Hat-Matrix.

`Npar_USApop.sas`

Anwendung in der 3D-Darstellung:

`Npar_Banknote.sas`

## Visualisierung

```
PROC GPLOT DATA=Datei1;
```

```
    PLOT pred*x; RUN;
```

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

$Y$ : Binäre Zielgröße,  $P(Y = 1) = p, P(Y = 0) = 1 - p,$   
 $Y \sim B(1, p)$

Wenn wir lineare Regression machen würden:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

$$\mathbf{E}Y_i = \alpha + \beta x_i, \quad \mathbf{E}\epsilon_i = 0$$

$$p_i = \alpha + \beta x_i$$

Problem: Wahrscheinlichkeiten sind beschränkt, lineare Funktionen aber nicht.

$Y$ : Binäre Zielgröße,  $P(Y = 1) = p, P(Y = 0) = 1 - p,$   
 $Y \sim B(1, p)$

Wenn wir lineare Regression machen würden:

$$\begin{aligned}Y_i &= \alpha + \beta x_i + \epsilon_i \\ \mathbf{E}Y_i &= \alpha + \beta x_i, \quad \mathbf{E}\epsilon_i = 0 \\ p_i &= \alpha + \beta x_i\end{aligned}$$

Problem: Wahrscheinlichkeiten sind beschränkt, lineare Funktionen aber nicht.

Ausweg: Odds ratio  $OR := \frac{p}{1-p}$

nach oben unbeschränkt, aber nicht nach unten

# Logistische Regression (2)

Logit

$$\text{Logit}(p) := \ln\left(\frac{p}{1-p}\right)$$

ist auch nach unten unbeschränkt.

## Modell

$$\begin{aligned}\text{Logit}(p_i) &= \ln\left(\frac{p_i}{1-p_i}\right) \\ &= \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \boldsymbol{\beta}' \mathbf{x}_i,\end{aligned}$$

$i = 1, \dots, n, p_i = P(Y_i = 1)$ .

$\mathbf{x}'_i = (1, x_{i1}, \dots, x_{ik}), \boldsymbol{\beta}' = (\alpha, \beta_1, \dots, \beta_k)$ .

Umstellen der letzten Gleichung liefert

## Logistische Regression (3)

$$p_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} = 1 - \frac{1}{1 + e^{\beta' \mathbf{x}_i}}.$$

Gegeben sind Beobachtungen:  $(y_i, \mathbf{x}_i)$ .

Unbekannt sind  $p_i$ .

Frage: Wie schätzen wir  $\beta$  ?

## Logistische Regression (3)

$$p_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} = 1 - \frac{1}{1 + e^{\beta' \mathbf{x}_i}}.$$

Gegeben sind Beobachtungen:  $(y_i, \mathbf{x}_i)$ .

Unbekannt sind  $p_i$ .

Frage: Wie schätzen wir  $\beta$  ?

Methode: Maximum-Likelihood

**PROC LOGISTIC;**

**MODEL** Y=X1 X2 /Optionen;

**RUN;**

Logistic\_banknote   Logistic\_tibetan   Logistic\_water

# Maximum-Likelihood Schätzung der Parameter

Idee: Eine Schätzung ist “gut”, wenn sie für die beobachteten Daten die “plausibelste” ist, d.h. wenn sie eine hohe Wahrscheinlichkeit produziert.

Ziel: maximiere (die Beobachtungen sind unabhängig)

$$L = P(y_1) \cdot P(y_2) \cdot \dots \cdot P(y_n) = \prod_{i=1}^n P(y_i)$$

$$y_i = \begin{cases} 1 & \text{mit Wkt. } p_i \\ 0 & \text{mit Wkt. } 1 - p_i \end{cases} \quad P(y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$P(0) = p_i^0 (1 - p_i)^{1-0} = 1 - p_i$$

$$P(1) = p_i^1 (1 - p_i)^{1-1} = p_i$$

hier:  $y_i$  bekannt (Beobachtungen),  $p_i$  zu schätzen.

# Logistische Regression (5)

Maximum-Likelihood Schätzung der Parameter (2)

Einsetzen

$$\begin{aligned}L &= \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \\&= \prod_{i=1}^n \left( \frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) \\\ln L &= \sum_{i=1}^n y_i \ln \left( \frac{p_i}{1 - p_i} \right) + \sum_{i=1}^n \ln(1 - p_i) \\&= \sum_{i=1}^n \beta' \mathbf{x}_i y_i - \sum_{i=1}^n \ln(1 + e^{\beta' \mathbf{x}_i})\end{aligned}$$

Da der Logarithmus monoton wachsend ist, genügt es  $\ln L$  zu maximieren.

# Logistische Regression (6)

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i \left(1 + e^{\beta' \mathbf{x}_i}\right)^{-1} e^{\beta' \mathbf{x}_i} \\ &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i \left(1 + e^{-\beta' \mathbf{x}_i}\right)^{-1} \\ &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i \tilde{y}_i,\end{aligned}$$

wobei

$$\tilde{y}_i = \frac{1}{1 + e^{-\beta' \mathbf{x}_i}}$$

die Vorhersagewahrscheinlichkeit für  $y_i = 1$  bei gegebenen  $\mathbf{x}_i$ .

# Logistische Regression (7)

$$\frac{\partial \ln L}{\partial \beta} = 0$$

ist Nichtlineares Gleichungssystem

→ numerische Lösung, z.B. Newton-Raphson Methode

hier: = Fisher Scoring (nimmt statt  $I(\beta)$ :  $\mathbf{E}(I(\beta))$ )

$\mathbf{U}(\beta)$ : Vektor der ersten Ableitungen von  $\ln L$

$\mathbf{I}(\beta)$ : Matrix der zweiten Ableitungen von  $\ln L$

Iteration

$$\beta_{j+1} = \beta_j - \mathbf{I}^{-1}(\beta_j) \mathbf{U}(\beta_j)$$

Konvergenz?

# Logistische Regression (7)

$$\frac{\partial \ln L}{\partial \beta} = 0$$

ist Nichtlineares Gleichungssystem

→ numerische Lösung, z.B. Newton-Raphson Methode

hier: = Fisher Scoring (nimmt statt  $I(\beta)$ :  $\mathbf{E}(I(\beta))$ )

$\mathbf{U}(\beta)$ : Vektor der ersten Ableitungen von  $\ln L$

$\mathbf{I}(\beta)$ : Matrix der zweiten Ableitungen von  $\ln L$

Iteration

$$\beta_{j+1} = \beta_j - \mathbf{I}^{-1}(\beta_j) \mathbf{U}(\beta_j)$$

Konvergenz? hoffentlich.

Vergleiche: Newton-Verfahren ( $k = 1$ ) zur Lösung von  $g(x) = 0$ .

# Logistische Regression (8)

## Output

- Modellinformationen
- Konvergenzstatus
- Modellanpassungsstatistiken
- Test der globalen Nullhypothese  $\beta = 0$
- ML-Schätzungen mit  $\chi^2$ -Statistiken und  $p$ -Werten
- Schätzungen der Odds Ratios  $e^{\hat{\beta}_j}$

# Logistische Regression (9)

## Modellanpassungsstatistiken

zum Vergleich verschiedener Modelle

- je größer AIC, SC, desto besser
- je kleiner  $-2 \ln L$  desto besser
  - $-2 \ln L$ : Abweichung vom saturierten Modell, d.h. vom anhand der Daten (bei perfekter Anpassung) möglichen Modell (perfekte Anpassung bei  $L = 1$ )
- Hosmer-Lemeshov Anpassungstest (Option LACKFIT)

# Logistische Regression (10)

Vorhersagefähigkeit des Modells

- Association of Predicted Probabilities and Observed Responses
- alle möglichen Paare  $(y_i, y_j)$  werden verglichen bzgl. ihres Vorhersagewertes, d.h. mit  $(\hat{y}_i, \hat{y}_j)$
- Anteil der konkordanten Paare C
- Kendall-Konkordanzkoeffizient Tau-a  
(kann als Bestimmtheitsmaß interpretiert werden)
- Somer's D, Gamma, c hängen mit C zusammen.

Modellwahl durch Selektion möglich (Option SELECTION= in Model-Anweisung)

# Hosmer-Lemeshov Test

Alle Beobachtungen haben nach der Schätzung der Parameter eine Vorhersagewahrscheinlichkeit in die Referenzgruppe (z.B. 'echte Banknoten')

$$\hat{p}_i = \frac{1}{1 + e^{-\hat{\beta}'\mathbf{x}_i}}$$

Das Intervall  $[0, 1]$  wird in etwa 10 Teile  $G_j$  eingeteilt, und ausgezählt, wieviele der  $\hat{p}_i$  in die einzelnen Teile fallen, total, und für beide Gruppen (z.B. echte und gefälschte Banknoten) getrennt.

Diese Anzahlen werden jeweils verglichen mit der erwarteten vorhergesagten Häufigkeit  $\sum_{i \in G_j} \hat{p}_i$  mit Hilfe eines  $\chi^2$  Anpassungstests.

# Inhalt

- Korrelation und Unabhängigkeit
- Lineare Regression
- Robuste lineare Regression
- Nichtlineare Regression
- Nichtparametrische Regression
- Logistische Regression
- Übersicht Regressionsverfahren

# Regressionsverfahren, kurze Übersicht (1)

## a) Lineare Regression

Modell:

$$Y_i = \theta_0 + \sum_{j=1}^m \theta_j X_{ij} + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$ ,  $i = 1, \dots, n$

$Y_i, \epsilon_i$  zufällig

$X_i$  zufällig oder fest

$\theta_0 \dots \theta_m; \sigma$ : Modellparameter

**PROC REG;**

**MODEL** abh.Variable = unabh.Variable(n)

    / **R DW;**

**RUN;**

# Regressionsverfahren, kurze Übersicht (2)

## b) Robuste Lineare Regression

Modell wie bei der linearen Regression

$$Y_i = \theta_0 + \sum_{j=1}^m \theta_j X_{ij} + \epsilon_i$$

robuste Abstandsfunktion

*MAD* statt *s* als Skalenschätzung.

**PROC ROBUSTREG;**

**MODEL** abh.Variable = unabh.Variable(n)

    / diagnostics leverage;

**RUN;**

# Regressionsverfahren

## Kurze Übersicht (3)

### c) Nichtlineare Regression

Modell:

$$Y_i = f(X_{1i}, \dots, X_{mi}, \theta_1, \dots, \theta_p) + \epsilon_i$$

$f$ : bekannt (i.A. nichtlinear)

**PROC NLIN;**

**MODEL** abh.Variable = Ausdruck;

**PARMS** Parameter = Anfangswert(e);

**RUN;**

# Regressionsverfahren

Kurze Übersicht (4)

## d) Nichtparametrische Regression

Modell:

$$Y_i = f(X_{1i}, \dots, X_{mi}) + \epsilon_i$$

$f$  unbekannt, aber "glatt", z.B.  $f \in C^2$ .

**PROC TPSPLINE;**

**MODEL** abh.Var. = (unabh. Var);

**RUN;**

Regression\_Phosphor\_Uebersicht.sas

# Regressionsverfahren

Kurze Übersicht (5)

## e) Logistische Regression

Y: binäre Zielgröße

$$p_i = P(Y_i = 1) = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}.$$

Parameter:  $\beta$ .

Odds ratio:  $\frac{p_i}{1-p_i}$

**PROC LOGISTIC;**

**MODEL** binaere Variable = abhaengige Variablen;

**RUN;**

Zufallszahlen werden nach einem deterministischen Algorithmus erzeugt  $\Rightarrow$  Pseudozufallszahlen

- wirken wie zufällige Zahlen (sollen sie jedenfalls)

Algorithmus:

Startwert  $x_0$ ,  $x_{n+1} = f(x_n)$  (z.B. Kongruenzen)

Ein Generator von SAS

CALL RANUNI Routine, RANUNI Funktion

$$x_{n+1} = \underbrace{397204094}_{2 \cdot 7 \cdot 7 \cdot 4053103} x_n \bmod (2^{31} - 1) \quad u_n = \frac{x_n}{2^{31} - 1}$$

liefert gleichverteilte Zufallszahlen  $u_n \in (0, 1)$ .

## Zufallszahlen (2)

### Ein weiterer Generator von SAS: RAND Funktion

Mersenne Twister hat eine Periode von  $2^{19937} - 1$  und liefert gleichverteilte Zufallszahlen  $u_n \in (0, 1)^{623}$ .

### zufälliger Startwert

seed = -1;

Der interne Startwert wird dann durch  $x_1$  ersetzt, der folgende Aufruf von ranuni/rannor(seed) liefert eine neue Zufallszahl.

### auf (0,1) gleichverteilte Zufallszahlen

x=ranuni(seed)

### Standardnormalverteilte Zufallszahlen

x=rannor(seed)

## Zufallszahlen (3)

### vorgegebene stetige Verteilung

wird z.B. aus gleichverteilter Zufallsvariable  $U_i$  mittels Quantilfunktion ( $F^{-1}(U_i)$ ) gewonnen.

### diskrete Verteilungen

werden erzeugt durch Klasseneinteilung des Intervalls  $(0, 1)$  entsprechend der vorgegebenen Wahrscheinlichkeiten  $p_i$ , also

$$(0, p_1], (p_1, p_1 + p_2], (p_1 + p_2, p_1 + p_2 + p_3], \dots, (p_1 + \dots + p_{k-1}, 1)$$

Call `rantbl(seed, p1, ..., pk-1, x)`

Zweidimensionale Normalverteilung,  $\rho = \text{corr}(X, Y)$

`probbnrm(x, y,  $\rho$ )`

## Zufallszahlen (4), wünschenswerte Eigenschaften

- Einfacher Algorithmus, wenig Rechenzeit.
- möglichst viele verschiedene Zufallszahlen sollen erzeugbar sein  
⇒ lange Periode.

- $k$ -Tupel  $(U_1, \dots, U_k) \sim R(0, 1)^k$ ,  $k \leq 10$   
⇒ Test auf Gleichverteilung.

- “Unabhängigkeit”

Test auf Autokorrelation (z.B. Durbin-Watson Test, vgl. Regression)

Plot der Punkte  $(U_i, U_{i+k})$ ,  $k = 1, 2, \dots$

es sollten keine Muster zu erkennen sein.

- Ziel der Clusteranalyse: Zusammenfassung von
- “ähnlichen” Objekten zu Gruppen (Clustern),
  - unähnliche Objekte in verschiedene Cluster.
- Cluster sind vorher nicht bekannt.

### 20 Patienten, Blutanalyse

Merkmale: Eisengehalt  $X_1$ , alkalische Phosphate  $X_2$

### Umweltverschmutzung in verschiedenen Städten

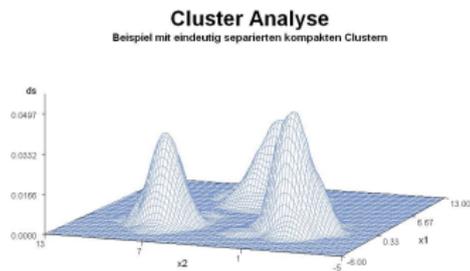
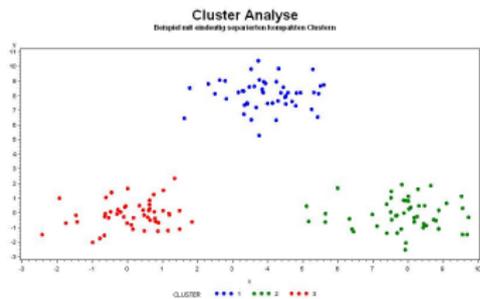
Merkmale: Schwebeteilchen, Schwefeldioxid

### Byzantinische Münzen

Lassen sich gesammelte Münzen verschiedenen Epochen zuordnen?

# Clusteranalyse

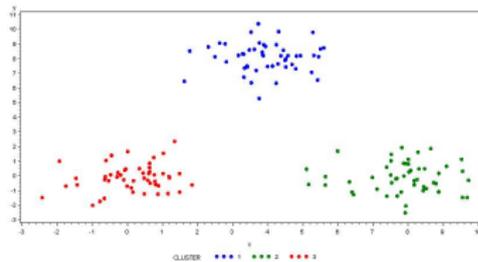
## Beispiel



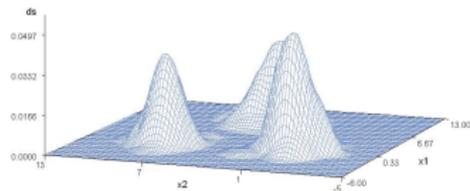
# Clusteranalyse

## Beispiel

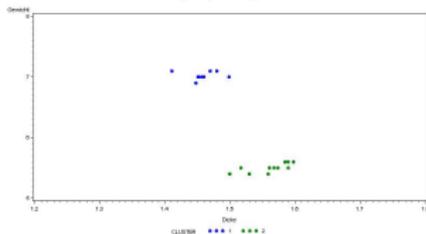
Cluster Analyse  
Beispiel mit eindeutig separierten kompakten Clustern



Cluster Analyse  
Beispiel mit eindeutig separierten kompakten Clustern



Byzantinische Münzen  
Single-Linkage Cluster Analyse



# Clusteranalyse

Wir unterscheiden:

partitionierende Clusteranalyse

Zahl der Cluster ist vorgegeben (MAXCLUSTERS=)

**PROC FASTCLUS** (k-means),

**PROC MODECLUS** (nichtparametrische Dichteschätzung)

# Clusteranalyse

Wir unterscheiden:

## partitionierende Clusteranalyse

Zahl der Cluster ist vorgegeben (MAXCLUSTERS=)

**PROC FASTCLUS** (k-means),

**PROC MODECLUS** (nichtparametrische Dichteschätzung)

## hierarchische Clusteranalyse

**PROC CLUSTER**, gefolgt von

**PROC TREE** und evtl.

**PROC GPLOT**

# Clusteranalyse

Wir unterscheiden:

## partitionierende Clusteranalyse

Zahl der Cluster ist vorgegeben (MAXCLUSTERS=)

**PROC FASTCLUS** (k-means),

**PROC MODECLUS** (nichtparametrische Dichteschätzung)

## hierarchische Clusteranalyse

**PROC CLUSTER**, gefolgt von

**PROC TREE** und evtl.

**PROC GPLOT**

## Fuzzy Clusteranalyse

# Clusteranalyse

Abstandsdefinitionen ( $p$ : # Merkmale)

Euklidischer Abstand (das ist Standard)

$$d_E^2(x, y) = \sum_{i=1}^p (x_i - y_i)^2$$

# Clusteranalyse

Abstandsdefinitionen ( $p$ : # Merkmale)

Euklidischer Abstand (das ist Standard)

$$d_E^2(x, y) = \sum_{i=1}^p (x_i - y_i)^2$$

City-Block Abstand (Manhattan-Abstand)

$$d_C(x, y) = \sum_{i=1}^p |x_i - y_i|$$

# Clusteranalyse

Abstandsdefinitionen ( $p$ : # Merkmale)

Euklidischer Abstand (das ist Standard)

$$d_E^2(x, y) = \sum_{i=1}^p (x_i - y_i)^2$$

City-Block Abstand (Manhattan-Abstand)

$$d_C(x, y) = \sum_{i=1}^p |x_i - y_i|$$

Tschebyschev-Abstand

$$d_T(x, y) = \max_i |x_i - y_i|$$

# Clusteranalyse

## Anmerkungen zu den Abständen

- Nichteuklidische Abstände müssen selbst berechnet werden.  
Macro %DISTANCE
- Abstandsmatrix kann in der DATA-Anweisung angegeben werden. **DATA**=name (**TYPE**=DISTANCE)
- Die Variablen sollten i.A. vor der Analyse standardisiert werden, da Variablen mit großer Varianz sonst großen Einfluß haben (Option **STANDARD** oder die Prozedur **ACECLUS** zuvor laufen lassen).  
davor: Ausreißer beseitigen.

# Hierarchische Clusteranalyse

## Methoden (1)

Die Methoden unterscheiden sich durch die Definition der Abstände  $D(C_i, C_j)$  zwischen Clustern  $C_i$  und  $C_j$ .

### Single Linkage

$$D_S(C_i, C_j) = \min \{d(k, l), k \in C_i, l \in C_j\}$$

# Hierarchische Clusteranalyse

## Methoden (1)

Die Methoden unterscheiden sich durch die Definition der Abstände  $D(C_i, C_j)$  zwischen Clustern  $C_i$  und  $C_j$ .

### Single Linkage

$$D_S(C_i, C_j) = \min \{d(k, l), k \in C_i, l \in C_j\}$$

### Complete Linkage

$$D_C(C_i, C_j) = \max \{d(k, l), k \in C_i, l \in C_j\}$$

# Hierarchische Clusteranalyse

## Methoden (1)

Die Methoden unterscheiden sich durch die Definition der Abstände  $D(C_i, C_j)$  zwischen Clustern  $C_i$  und  $C_j$ .

### Single Linkage

$$D_S(C_i, C_j) = \min \{d(k, l), k \in C_i, l \in C_j\}$$

### Complete Linkage

$$D_C(C_i, C_j) = \max \{d(k, l), k \in C_i, l \in C_j\}$$

### Centroid

$$D_{CE}(C_i, C_j) = d(\bar{X}_i, \bar{X}_j) \quad \text{Abstände der Schwerpunkte}$$

# Hierarchische Clusteranalyse

Methoden (2)

## Average Linkage

$$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, l \in C_j} d(k, l)$$

# Hierarchische Clusteranalyse

Methoden (2)

## Average Linkage

$$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, l \in C_j} d(k, l)$$

## Ward

ANOVA-Abstände innerhalb der Cluster minimieren, außerhalb maximieren. Nach Umrechnen erhält man

$$D_W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} D_{CE}(C_i, C_j).$$

# Hierarchische Clusteranalyse

## Methoden (2)

### Average Linkage

$$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, l \in C_j} d(k, l)$$

### Ward

ANOVA-Abstände innerhalb der Cluster minimieren, außerhalb maximieren. Nach Umrechnen erhält man

$$D_W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} D_{CE}(C_i, C_j).$$

### Density Linkage

beruht auf nichtparametrischer Dichteschätzung (DENSITY, TWOSTAGE)

# Hierarchische Clusteranalyse

## Tendenzen

**WARD:** Cluster mit etwa gleicher Anzahl von  
Objekten

**AVERAGE:** ballförmige Cluster

**SINGLE:** große Cluster, “Ketteneffekt”,  
langgestreckte Cluster

**COMPLETE:** kompakte, kleine Cluster

Im Mittel erweisen sich **Average Linkage** und **Ward** sowie die nichtparametrischen Methoden als die geeignetsten Methoden.

# Hierarchische Clusteranalyse

## Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.

$$Z = \{C_1, \dots, C_n\}, C_i \cap C_j = \emptyset \quad C_i = \{O_i\}$$

# Hierarchische Clusteranalyse

## Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.

$$Z = \{C_1, \dots, C_n\}, C_i \cap C_j = \emptyset \quad C_i = \{O_i\}$$

2. Suche  $C_r, C_l$  :  $d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$

# Hierarchische Clusteranalyse

## Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.

$$Z = \{C_1, \dots, C_n\}, C_i \cap C_j = \emptyset \quad C_i = \{O_i\}$$

2. Suche  $C_r, C_l$ :  $d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$

3. Fusioniere  $C_r, C_l$  zu einem neuen Cluster:

$$C_r^{new} = C_r \cup C_l$$

# Hierarchische Clusteranalyse

## Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.

$$Z = \{C_1, \dots, C_n\}, C_i \cap C_j = \emptyset \quad C_i = \{O_i\}$$

2. Suche  $C_r, C_l$ :  $d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$

3. Fusioniere  $C_r, C_l$  zu einem neuen Cluster:

$$C_r^{new} = C_r \cup C_l$$

4. Ändere die  $r$ -te Zeile und Spalte der Distanzmatrix durch Berechnung der Abstände von  $C_r^{new}$  zu den anderen Clustern!  
Streiche die  $l$ -te Zeile und Spalte!

# Hierarchische Clusteranalyse

## Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.

$$Z = \{C_1, \dots, C_n\}, C_i \cap C_j = \emptyset \quad C_i = \{O_i\}$$

2. Suche  $C_r, C_l$ :  $d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$

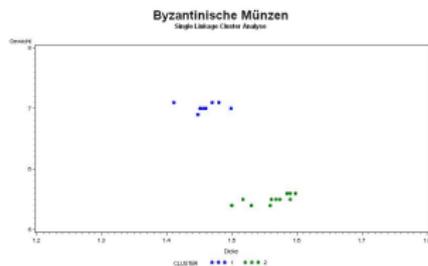
3. Fusioniere  $C_r, C_l$  zu einem neuen Cluster:

$$C_r^{new} = C_r \cup C_l$$

4. Ändere die  $r$ -te Zeile und Spalte der Distanzmatrix durch Berechnung der Abstände von  $C_r^{new}$  zu den anderen Clustern!  
Streiche die  $l$ -te Zeile und Spalte!
5. Beende nach  $n-1$  Schritten, ansonsten fahre bei 2. mit geänderter Distanzmatrix fort!

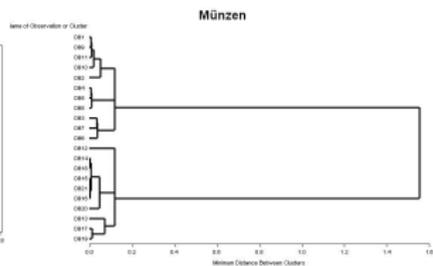
# Clusteranalyse

## Beispiel



# Clusteranalyse

## Beispiel



# Hierarchische Clusteranalyse

## Anmerkungen

- Alle von SAS angebotenen hierarchischen Methoden sind agglomerativ.
- Es gibt auch divisive Methoden.
- Fall großer Datensätze:

PROC FASTCLUS: Vorclusteranalyse mit großer Anzahl von Clustern

PROC CLUSTER: Clusteranalyse mit diesen Clustern.

# Hierarchische Clusteranalyse

zu WARD:

ANOVA Abstände innerhalb eines Clusters  $i$

$$D_i = \frac{1}{n_i} \sum_{l \in C_i} d^2(O_l, \bar{X}_i)$$

Fusioniere die Cluster  $C_i$  und  $C_j$ , wenn

$$D_{CE}(C_i, C_j) - D_i - D_j \longrightarrow \min_{i,j}$$

# Clusteranalyse, Durchführung

```
PROC CLUSTER /*hierarchische Clusteranalyse*/  
  METHOD=methode  
  STANDARD /*Standardisierung*/  
  OUTTREE=datei; /*Eingabedatei für Proc Tree*/  
RUN;
```

# Clusteranalyse, Durchführung

```
PROC CLUSTER /*hierarchische Clusteranalyse*/  
  METHOD=methode  
  STANDARD /*Standardisierung*/  
  OUTTREE=datei; /*Eingabedatei für Proc Tree*/  
  
RUN;  
  
PROC TREE DATA=datei  
  OUT=out /*Ausgabedatei z.B.für PROC GPLOT*/  
  NCLUSTERS=nc /*Anz. Cluster*/  
  COPY vars /*vars in die Ausgabedatei*/  
  
RUN;
```

# Clusteranalyse, Durchführung

```
PROC CLUSTER /*hierarchische Clusteranalyse*/  
  METHOD=methode  
  STANDARD /*Standardisierung*/  
  OUTTREE=datei; /*Eingabedatei für Proc Tree*/  
  
RUN;  
  
PROC TREE DATA=datei  
  OUT=out /*Ausgabedatei z.B.für PROC GPLOT*/  
  NCLUSTERS=nc /*Anz. Cluster*/  
  COPY vars /*vars in die Ausgabedatei*/  
  
RUN;  
  
PROC GPLOT;  
  PLOT variablen=cluster; /*Symbol-Anweisungen  
  vorher definieren*/  
  
RUN;
```

# Hierarchische Clusteranalyse

Die Ausgabedatei OUTTREE=

- `_NAME_` Bezeichnung der Cluster
  - $\geq 2$  Beobachtungen: CLn
  - 1 Beobachtung: OBn
- `_NCL_` Anzahl der Cluster
- `_FREQ_` Anzahl der Beobachtungen  
im jeweiligen Cluster

n: Clusternummer (CLn) oder

Beobachtungsnummer (OBn = `_N_`)

`Cluster_Air.sas`   `Cluster.sas`   `Cluster_Banknoten.sas`

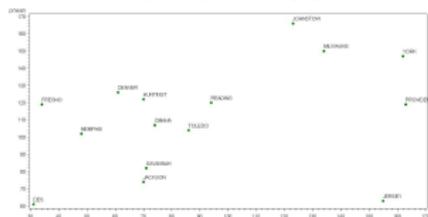
`Cluster_Muenzen.sas`

# Hierarchische Clusteranalyse

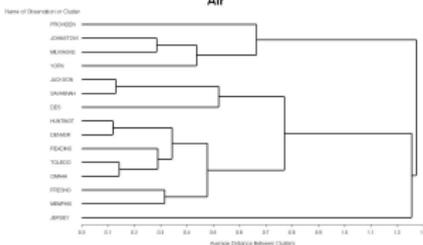
Beispiel: Luftverschmutzung in USA-Städten

Scatterplot von 15 Städten

Air-Qualität, Variablen Ozone (ozon) und Schwefeldioxid (sulfur)

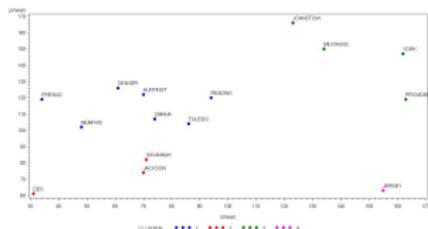


Air



Complete Linkage Cluster Analyse

Ozone and Sulfur in USA-Cities



## 3D-Darstellung von Datenpunkten

```
PROC G3D;  
SCATTER y*x = z;  
RUN;
```

Wertetabelle erstellen,  
vgl. z.B. [Texashut.sas](#)

```
PROC G3D;  
PLOT y*x = z;  
RUN;
```

# Glatte 3D-Darstellung, Kontur-Plot

## Glatte 3D-Darstellung

```
PROC G3GRID;
```

```
    GRID var1*var2=y/SPLINE SMOOTH=Wert;
```

```
    AXIS1=von TO bis BY Schrittweite;
```

```
    AXIS2=von TO bis BY Schrittweite;
```

```
RUN;
```

# Glatte 3D-Darstellung, Kontur-Plot

## Glatte 3D-Darstellung

```
PROC G3GRID;
```

```
    GRID var1*var2=y/SPLINE SMOOTH=Wert;
```

```
    AXIS1=von TO bis BY Schrittweite;
```

```
    AXIS2=von TO bis BY Schrittweite;
```

```
RUN;
```

## Kontur-Plot

```
PROC GCONTOUR;
```

```
    PLOT var1*var2 = y /LLEVEL=1;
```

```
RUN;
```

Erläuterung dazu siehe Programm [Npar\\_Banknote.sas](#)

# Inhalt

- Problemstellung und Übersicht
- Berechnung der Hauptkomponenten
- Anzahl der Hauptkomponenten

# Problemstellung

- viele (hoch) korrelierte Variablen  
→ diese sollen ersetzt werden, durch neue, unkorrelierte Variablen,  
durch eine lineare Transformation

# Problemstellung

- viele (hoch) korrelierte Variablen  
→ diese sollen ersetzt werden, durch neue, unkorrelierte Variablen,  
durch eine lineare Transformation
- **Ziel:** wenig neue Variablen,  
die aber möglichst viel Information aus den Daten erhalten.

# Problemstellung

- viele (hoch) korrelierte Variablen  
→ diese sollen ersetzt werden, durch neue, unkorrelierte Variablen, durch eine lineare Transformation
- **Ziel:** wenig neue Variablen, die aber möglichst viel Information aus den Daten erhalten.

Daten: Punkte im  $p$ -dimensionalen Raum

Ziel: Projektion in einen  $p'$ -dimensionalen

( $p' \leq p$ ) Teilraum mit möglichst viel erhaltener Information.

[Hauptkomponenten\\_Venusmuscheln.sas](#) ( $p = 2$ )

# Hauptkomponentenanalyse (2)

## Annahmen

Daten sind Realisierungen eines  $p$ -variaten zufälligen Vektors

$$\mathbf{X} := (X_1, \dots, X_p) \text{ mit } \mathbf{E}(\mathbf{X}) = \mathbf{0} \quad \text{und} \quad \text{cov}\mathbf{X} = \mathbf{\Sigma} > 0$$

(Kovarianzmatrix, positiv definit)

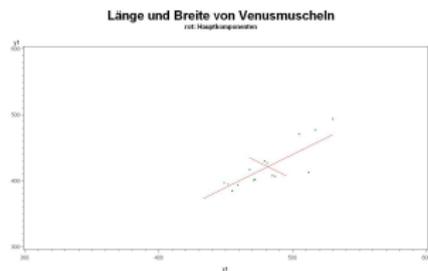
Bem: Die erste Bedingung erreicht man durch zentrieren um die Mittelwerte  $\bar{X}_j, j = 1, \dots, p$

Wenn zwischen einzelnen Komponenten des zufälligen Vektors ein gewisser (etwa ein linearer) Zusammenhang besteht, so ist eine Dimensionsreduzierung möglich.

Der Zusammenhang wird durch Gerade dargestellt (ausgezeichnete Richtung in der Ebene).

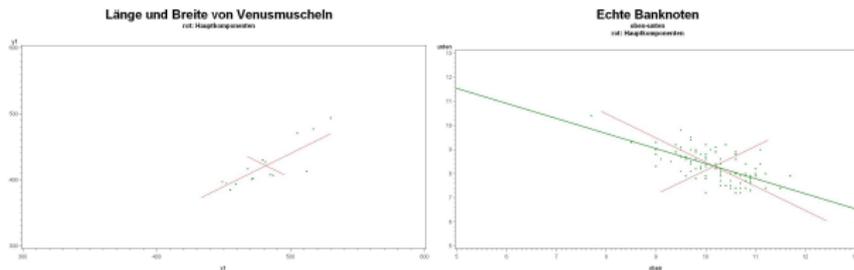
# Hauptkomponentenanalyse

## Beispiele



# Hauptkomponentenanalyse

## Beispiele



**Frage:** Wie kann man diese ausgezeichnete Richtung erfassen?

# Hauptkomponentenanalyse (3)

## 1. Hauptkomponente. Die Linearkombination

$$Y_1 = \sum_{j=1}^p b_{1j} X_j$$

ist so zu bestimmen, dass  $\text{var } Y_1 \rightarrow \max.$

unter der Normierungsbedingung  $(\sum_j b_{1j}^2 = 1)$

Die Variablen werden vorher zentriert,  $X_j := X_j - \bar{X}_j$

# Hauptkomponentenanalyse (3)

## 1. Hauptkomponente. Die Linearkombination

$$Y_1 = \sum_{j=1}^p b_{1j} X_j$$

ist so zu bestimmen, dass  $\text{var } Y_1 \rightarrow \max.$

unter der Normierungsbedingung  $(\sum_j b_{1j}^2 = 1)$

Die Variablen werden vorher zentriert,  $X_j := X_j - \bar{X}_j$

## 2. Hauptkomponente. Die Linearkombination

$$Y_2 = \sum_{j=1}^p b_{2j} X_j$$

ist so zu bestimmen, dass  $\text{var } Y_2 \rightarrow \max,$

unter Normierungsbedingung  $(\sum_j b_{2j}^2 = 1)$

und unter der Bedingung  $\text{cov}(Y_1, Y_2) = 0$

Die Bedingung  $\text{cov}(Y_1, Y_2) = 0$  sichert Unkorreliertheit der Hauptkomponenten.

Hauptkomponenten sind durch die Korrelationsmatrix eindeutig bestimmt.

**Hauptachsentransformation:  $\Sigma = \mathbf{B}\mathbf{\Lambda}\mathbf{B}'$**

**$\Sigma$**  : (empirische) Korrelationsmatrix (bekannt)

**$\mathbf{B}$** : Orthogonalmatrix

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}$$

$\lambda_i$  : Eigenwerte, sei  $\lambda_1 \geq \dots \geq \lambda_p \geq 0$

# Hauptkomponentenanalyse (5)

## Hauptkomponenten

$$\mathbf{Y} = \mathbf{B}' \cdot \mathbf{X}$$

**Mahalanobis-Distanz** eines Datenpunktes  $\mathbf{X} = (X_1, \dots, X_p)$  zum Ursprung:

$$\begin{aligned} \mathbf{X}'\Sigma^{-1}\mathbf{X} &= \mathbf{X}'\mathbf{B}\Lambda^{-1}\mathbf{B}'\mathbf{X} = \mathbf{Y}'\Lambda^{-1}\mathbf{Y} \\ &= \sum_{i=1}^p \frac{Y_i^2}{\lambda_i}. \end{aligned}$$

Die Konturen sind Ellipsoide.

# Inhalt

- Problemstellung und Übersicht
- Berechnung der Hauptkomponenten
- Anzahl der Hauptkomponenten

## Aufgabe

gesucht Linearkombination  $Y_1$  der Komponenten von  $\mathbf{X}$  derart, dass die Varianz  $\text{var } Y_1$  maximal.

Ansatz:  $Y_1 := \mathbf{b}'_1 \mathbf{X}$

$\mathbf{b}_1$ : -  $p$ -variater Gewichtsvektor,  $Y_1$  eindimensional

$$\text{var } Y_1 = \text{var}(\mathbf{b}'_1 \mathbf{X}) = \mathbf{b}'_1 \boldsymbol{\Sigma} \mathbf{b}_1.$$

Sinnvolle Lösung der Aufgabe bei Normierung von  $\mathbf{b}_1$  :  $\mathbf{b}'_1 \mathbf{b}_1 = 1$ .

Maximiere

$$L(\mathbf{b}_1) = \mathbf{b}'_1 \boldsymbol{\Sigma} \mathbf{b}_1 - \lambda(\mathbf{b}'_1 \mathbf{b}_1 - 1) \quad \text{bzgl. } \mathbf{b}_1$$

Maximiere

$$\begin{aligned}
 L(\mathbf{b}_1) &= \mathbf{b}'_1 \boldsymbol{\Sigma} \mathbf{b}_1 - \lambda(\mathbf{b}'_1 \mathbf{b}_1 - 1) \quad \text{bzgl. } b_1 \\
 \frac{\partial L(\mathbf{b}_1)}{\partial \mathbf{b}_1} &= 2\boldsymbol{\Sigma} \mathbf{b}_1 - 2\lambda \mathbf{b}_1 \\
 &= 2(\boldsymbol{\Sigma} - \lambda \mathbf{I}) \mathbf{b}_1 = 0 \quad (*) \\
 \frac{\partial L(\mathbf{b}_1)}{\partial \lambda} &= \mathbf{b}'_1 \mathbf{b}_1 - 1 = 0 \Rightarrow \mathbf{b}'_1 \mathbf{b}_1 = 1
 \end{aligned}$$

Lösungen von (\*) sind die Eigenwerte von  $\boldsymbol{\Sigma}$  :

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q > 0.$$

$$\text{var} \mathbf{Y}_1 = \mathbf{b}'_1 \boldsymbol{\Sigma} \mathbf{b}_1 \stackrel{(*)}{=} \mathbf{b}'_1 \lambda \mathbf{I} \mathbf{b}_1 = \mathbf{b}'_1 \lambda \mathbf{b}_1 = \lambda$$

maximal für  $\lambda = \lambda_1$  (maximaler Eigenwert von  $\boldsymbol{\Sigma}$ .)

$\mathbf{b}_1$ : der zu  $\lambda_1$  gehörende Eigenvektor.

Def.:  $\mathbf{Y}_1$  heisst erste Hauptkomponente von  $\mathbf{X}$ .

## Berechnung der Hauptkomponenten (2)

Aufgabe: gesucht Linearkombination  $\mathbf{Y}_2$  der Komponenten von  $\mathbf{X}$  derart, dass  $\text{var } Y_2$  maximal und  $Y_1$  und  $Y_2$  unkorreliert sind.

Ansatz:  $Y_2 := \mathbf{b}'_2 \mathbf{X}$

$b_2$ :  $p$ -variater Gewichtsvektor,

$Y_2$  eindimensional

$$\text{var } Y_2 = \text{var}(\mathbf{b}'_2 \mathbf{X}) = \mathbf{b}'_2 \boldsymbol{\Sigma} \mathbf{b}_2$$

Sinnvolle Lösung der Aufgabe bei Normierung von  $\mathbf{b}_2$  :

$$\mathbf{b}'_2 \mathbf{b}_2 = 1.$$

## Berechnung der Hauptkomponenten (3)

Jetzt kommt noch die Bedingung  $cov(Y_1, Y_2) = 0$  hinzu:

$$\begin{aligned} 0 &= cov(Y_1, Y_2) = cov(\mathbf{b}'_1 \mathbf{X}, \mathbf{b}'_2 \mathbf{X}) = \\ &= E((\mathbf{b}'_1 \mathbf{X})(\mathbf{b}'_2 \mathbf{X})') \quad \text{da } E\mathbf{X} = 0 \\ &= \mathbf{b}'_1 \underbrace{E(\mathbf{X} \cdot \mathbf{X}')}_{\Sigma} \mathbf{b}_2 \\ &= \mathbf{b}'_1 \Sigma \mathbf{b}_2 \\ &= \lambda_1 \mathbf{b}'_1 \mathbf{b}_2 \quad \text{da } \mathbf{b}_1 \text{ Eigenvektor} \\ &= \lambda_1 \mathbf{b}'_1 \mathbf{b}_2 \end{aligned}$$

$\Rightarrow \mathbf{b}_1$  und  $\mathbf{b}_2$  sollen orthogonal sein.

# Berechnung der Hauptkomponenten (4)

Maximiere bzgl.  $\mathbf{b}_2$ :

$$\begin{aligned}
 L(\mathbf{b}_2) &= \mathbf{b}_2' \boldsymbol{\Sigma} \mathbf{b}_2 - \lambda(\mathbf{b}_2' \mathbf{b}_2 - 1) + \theta(\mathbf{b}_1' \mathbf{b}_2) \\
 \frac{\partial L(\mathbf{b}_2)}{\partial \mathbf{b}_2} &= 2\boldsymbol{\Sigma} \mathbf{b}_2 - 2\lambda \mathbf{b}_2 + \theta \mathbf{b}_1 = 0 & (**) \\
 \frac{\partial L(\mathbf{b}_2)}{\partial \lambda} &= \mathbf{b}_2' \mathbf{b}_2 - 1 = 0 \\
 \frac{\partial L(\mathbf{b}_2)}{\partial \theta} &= \mathbf{b}_1' \mathbf{b}_2 = 0.
 \end{aligned}$$

Multiplizieren (\*\*) von links mit  $\mathbf{b}_1'$ :

$$\underbrace{2\mathbf{b}_1' \boldsymbol{\Sigma} \mathbf{b}_2}_{=0} - \underbrace{2\lambda \mathbf{b}_1' \mathbf{b}_2}_{=0} + \theta \mathbf{b}_1' \mathbf{b}_1 = 0$$

$\Rightarrow \theta = 0 \quad \Rightarrow (\boldsymbol{\Sigma} - \lambda \mathbf{I}) \mathbf{b}_2 = 0 \quad \Rightarrow |\boldsymbol{\Sigma} - \lambda \mathbf{I}| = 0$  (wie im ersten Schritt).

$$\text{var } Y_2 = \mathbf{b}'_2 \boldsymbol{\Sigma} \mathbf{b}_2 = \lambda \mathbf{b}'_2 \mathbf{b}_2 = \lambda$$

var  $Y_2$  maximal unter der Voraussetzung

$Y_2$  und  $Y_1$  unkorreliert für:  $\lambda = \lambda_2$

$\mathbf{b}_2$ : der zu  $\lambda_2$  gehörige Eigenvektor

Def.:  $Y_2 := \mathbf{b}'_2 \mathbf{X}$  heißt zweite Hauptkomponente von  $\mathbf{X}$ .

Allgemein: Die  $k$ -te Hauptkomponente von  $\mathbf{X}$  wird definiert durch die  
Linearkombination

$$Y_k = \mathbf{b}'_k \mathbf{X}$$

unter der Voraussetzung  $Y_k$  ist unkorreliert zu  $Y_1, \dots, Y_{k-1}$ ,

und var  $Y_k$  ist maximal ( $=\lambda_k$ ) ( $k = 2, \dots, p$ )

Bem: Wenn der Eigenwert  $\lambda$  mehrmals auftritt, so ist der zugehörige  
Eigenvektor nicht eindeutig.

# Inhalt

- Problemstellung und Übersicht
- Berechnung der Hauptkomponenten
- Anzahl der Hauptkomponenten

## Ziel: Dimensionen verkleinern.

Dazu brauchen wir ein Maß für Übereinstimmung an Information.

Betrachten als skalares Maß für die Gesamtvariabilität des Vektors  $X$  die Spur von  $\Sigma$ :

$$sp(\Sigma) = \sum_{i=1}^p \sigma_{ii}^2 = \sum_{i=1}^p \sigma_i^2$$

$\Sigma \mathbf{b}_i = \lambda_i \mathbf{b}_i, \quad i = 1, \dots, p, \quad \mathbf{b}_i$ : Eigenvektoren von  $\Sigma$ .

$\mathbf{B} := (\mathbf{b}_1, \dots, \mathbf{b}_p)$  (orthogonale) Matrix  $(p, p)$

$$\mathbf{B}'\mathbf{B} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ 0 & \cdot & \dots & 1 \end{pmatrix} = \mathbf{I} = \mathbf{B}\mathbf{B}'$$

Hauptkomponenten:

$$\mathbf{Y} := \begin{pmatrix} Y_1 \\ \dots \\ \dots \\ Y_p \end{pmatrix} = \mathbf{B}'\mathbf{X}$$

$$\text{cov}\mathbf{Y} = \mathbf{B}'\mathbf{\Sigma}\mathbf{B} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ & \dots & \\ 0 & \dots & \lambda_p \end{pmatrix}$$

Hauptachsentransformation

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}' = \sum_{i=1}^p \lambda_i \mathbf{b}_i \mathbf{b}_i'$$

Spektralzerlegung von  $\mathbf{\Sigma}$ .

## Anzahl der Hauptkomponenten (3)

$$sp(\boldsymbol{\Sigma}) = sp(\mathbf{B}\boldsymbol{\Lambda}\mathbf{B}') = sp(\underbrace{\boldsymbol{\Lambda}\mathbf{B}\mathbf{B}'}_{=I}) = \sum_{i=1}^p \lambda_i$$

Aufgabe:  $X$  soll durch einen  $r$ -dimensionalen Vektor so ersetzt werden, dass sich die Gesamtvariation  $sp(\boldsymbol{\Sigma})$  möglichst wenig ändert.

Lösung: Man nehme die ersten  $r$  Hauptkomponenten.

$$\mathbf{B}^* := (\mathbf{b}_1, \dots, \mathbf{b}_r)$$

$$\mathbf{Y}^* := (\mathbf{B}^*)'\mathbf{X}$$

Dann gilt:

$$sp(\text{var}\mathbf{Y}^*) = sp(\mathbf{B}^*)'\boldsymbol{\Sigma}\mathbf{B}^* = sp\boldsymbol{\Lambda}_r = \sum_{i=1}^r \lambda_i$$

Anmerkung: Es gilt  $sp(\mathbf{A}\mathbf{C}) = sp(\mathbf{C}\mathbf{A})$  (ÜA).

Zum Vergleich betrachtet man den Quotienten

$$Q := \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i}$$

und fordert z.B.  $Q > 0.85$ .

Bemerkung: Es gibt auch andere Kriterien, z.B.

$r =$  Anzahl der Eigenwerte  $> 1$

$r =$  Anzahl der Eigenwerte bis zu einem evtl. Knick in der Eigenwertkurve (vgl. Scree-Plot bei der Prozedur **FACTOR**)

# Bestimmung der Hauptkomponenten, wenn $\Sigma$ unbekannt

- 1. Schätzung für  $\Sigma$  durch das Beobachtungsmaterial durchführen  $\Rightarrow \hat{\Sigma}$
- 2. Berechnung der Hauptkomponenten auf Basis von  $\hat{\Sigma}$ .  
Diese Hauptkomponenten werden dann die empirischen Hauptkomponenten genannt.
- 3. Um Erwartungswert = 0 zu erreichen  $\bar{X}_{.j}$  berechnen, dann Verschiebung  $X_{ij} - \bar{X}_{.j}$

Die Höhenlinien der Dichten beschreiben Ellipsen mit Hauptachsen in Hauptkomponentenrichtung.

(falls Normalverteilung vorliegt)

Beispiel **X**: 2-dimensional normalverteilte Zufallsvariable,  $\Sigma > 0$

gesucht: Hauptkomponenten von **X**:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Dichte:

$$f_{N(0, \Sigma)}(x) = \frac{1}{2\pi \sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2} x' \Sigma^{-1} x\right\}$$

Betrachten die Menge der Punkte  $x$  mit

$$c = f_{N(0, \Sigma)}(x) = \frac{1}{2\pi \sqrt{|\Sigma|}} e^{-\frac{1}{2} c_1} \Rightarrow c_1 = x' \Sigma^{-1} x$$

d.h. die Konturlinien sind Ellipsen mit Ursprung in  $\mathbf{0}$ .

$\mathbf{Y}$  sei Vektor der Hauptkomponenten

$$\mathbf{Y} = \mathbf{B}'\mathbf{X} \quad \mathbf{B} = (b_1, b_2)$$

$b_1$  Eigenvektor von  $\Sigma$ , der zum größeren Eigenwert gehört

$b_2$  Eigenvektor von  $\Sigma$ , der zum kleineren Eigenwert gehört

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

$$\begin{aligned} c_1 &= \mathbf{x}'\Sigma^{-1}\mathbf{x} = \mathbf{x}'/\Sigma^{-1}/\mathbf{x} \stackrel{I=\mathbf{B}\mathbf{B}'}{=} \mathbf{x}'\mathbf{B}\mathbf{B}'\Sigma^{-1}\mathbf{B}\mathbf{B}'\mathbf{x} \\ &= (\mathbf{B}'\mathbf{x})'\mathbf{B}'\Sigma^{-1}\mathbf{B}(\mathbf{B}'\mathbf{x}) = \mathbf{Y}'\mathbf{B}'\Sigma^{-1}\mathbf{B}\mathbf{Y} \\ &= \mathbf{Y}'\Lambda^{-1}\mathbf{Y} = \sum_{i=1}^2 \frac{1}{\lambda_i} Y_i^2 = \mathbf{B}\Lambda^{-1}\mathbf{B}' \end{aligned}$$

Nebenrechnung:

$$\Sigma = \mathbf{B}\Lambda\mathbf{B}' \quad \Rightarrow \quad \Sigma^{-1} = (\mathbf{B}')^{-1}\Lambda^{-1}\mathbf{B}^{-1}$$

Darstellung einer Ellipse in Hauptachsen  $Y_1, Y_2$ :

$$\frac{Y_1^2}{a^2} + \frac{Y_2^2}{b^2} = 1$$

$$a = \lambda_1 c_1 \quad b = \lambda_2 c_1$$

## PROC PRINCOMP

OUTSTAT=Statistiken;

VAR varnamen;

**RUN;**

## PROC PRINCOMP

OUTSTAT=Statistiken;

VAR varnamen;

RUN;

## PROC FACTOR;

VAR varnamen;

RUN;

## PROC PRINCOMP

OUTSTAT=Statistiken;

VAR varnamen;

RUN;

Ausgabe PRINCOMP: Eigenwerte und normierte Eigenvektoren von

$\Sigma$ .

OUTSTAT: Ausgabestatistiken

[Hauptkomponenten\\_Banknote.sas](#)

Zwei Variablen (oben, unten), nur echte: Interpretation:

1. Hauptkomponente: unten-oben
2. Hauptkomponente: unten+oben

Ausgabe FACTOR: siehe Faktoranalyse

PROC FACTOR;

VAR varnamen;

RUN;

# Inhalt

- Modell

# Faktorenanalyse

- Analyse der Beziehungen zwischen verschiedenen Merkmalen (Variablen).
- Finden gemeinsamer Ursachenkomplexe, sogenannter Faktoren.
- Geschichte: Spearman (1904)  
Thurstone (1931,1947)  
Psychologie: Finden von 'Faktoren' der Intelligenz
- hängt eng mit der Hauptkomponentenanalyse zusammen, ist jedoch viel mehr.

## Modell:

$$X_j = d_{j1}F_1 + \dots + d_{jr}F_r + \epsilon_j, \quad j = 1 \dots p, \quad r \leq p$$

in Matrixschreibweise

$$\mathbf{X} = \mathbf{D} \cdot \mathbf{F} + \boldsymbol{\epsilon} \quad \mathbf{X} = (X_1, \dots, X_p)'$$

$$\mathbf{D} = \begin{pmatrix} d_{11} & \dots & d_{1r} \\ \dots & \dots & \dots \\ d_{p1} & \dots & d_{pr} \end{pmatrix}$$

$\mathbf{F} = (F_1 \dots F_r)'$  zufälliger Vektor der Faktoren (unbekannt)

$\boldsymbol{\epsilon} = (\epsilon_1 \dots \epsilon_p)'$  zufälliger Vektor

$\text{var}(\boldsymbol{\epsilon}) = V$  Diagonalmatrix, d.h.  $\epsilon_j$  sind unkorreliert

$\epsilon_j$  wirkt nur auf  $X_j$  (nicht auf die anderen Komponenten)

$$\mathbf{E}(\mathbf{X}) = \mathbf{E}(\mathbf{F}) = \mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$$

$\mathbf{F}$  und  $\boldsymbol{\epsilon}$  seien unabhängig, insbesondere

$$\text{cov}(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{cov}\mathbf{F} = \mathbf{I}$$

d.h. die Faktoren seien auch standardisiert und unabhängig voneinander.

Annahme: Die Variablen sind standardisiert;  
gegebenenfalls sind sie zu standardisieren  
(SAS: Option STANDARD)

$$X_{ij} := \frac{X_{ij} - \overline{X_{.j}}}{s_j}$$

$\boldsymbol{\Sigma} = \text{cov}\mathbf{X}$  Korrelationsmatrix

# Faktorenanalyse

Einsetzen in das Modell liefert:

$$\begin{aligned}\Sigma &= \text{cov}\mathbf{X} = \text{cov}(\mathbf{D} * \mathbf{F} + \epsilon) \\ &= \mathbf{D} \underbrace{(\text{cov}\mathbf{F})}_{=I} \mathbf{D}' + \text{cov}(\epsilon) \\ &= \mathbf{D}\mathbf{D}' + \mathbf{V}\end{aligned}$$

$\Sigma$ : bekannt = (empir.) Kovarianzmatrix

$\mathbf{D}$ ,  $\mathbf{V}$  sind gesucht

Lösung nicht eindeutig.

# Faktorenanalyse

Nehmen fürs erste an  $r = p$ , d.h.  $\epsilon = \mathbf{0}$

und  $\mathbf{X}$  ist eine Lineartransformation der (unbekannten) Faktoren,

$$\mathbf{X} = \mathbf{D} \cdot \mathbf{F}$$

Nach dem vorigen Kapitel (Hauptkomponentenanalyse) wissen wir:

$\mathbf{Y} = \mathbf{B}'\mathbf{X}$  ist Vektor der Hauptkomponenten

Wegen  $\mathbf{B}'\mathbf{B} = \mathbf{I}$  ( $\mathbf{B}$  ist Orthogonalmatrix) folgt:

$$\mathbf{X} = \mathbf{B}\mathbf{Y}$$

Vergleichen wir die Modelle miteinander, so sehen wir, daß

$$\mathbf{F} = \mathbf{Y} \quad (\text{mit } \mathbf{D} = \mathbf{B})$$

eine (vollständige Faktoren-) Lösung wäre.

# Faktorenanalyse

Aber: die Faktoren sind noch nicht standardisiert.

Zur Lösung des Problems berechnen wir die Korrelationsmatrix

$$\begin{aligned}\Sigma &= \text{corr}\mathbf{X} = \mathbf{B} \underbrace{\text{cov}(\mathbf{Y})}_{=\Lambda} \mathbf{B}' = \mathbf{B}\Lambda\mathbf{B}' \\ &= \mathbf{B}\Lambda^{1/2} \cdot \Lambda^{1/2}\mathbf{B}' = (\mathbf{B}\Lambda^{1/2}) \cdot (\mathbf{B}\Lambda^{1/2})' \\ &= \mathbf{D}\mathbf{D}'\end{aligned}$$

mit  $\mathbf{D} = \mathbf{B}\Lambda^{1/2}$ .

$\Lambda$ : Diagonalmatrix, Hauptdiagonale: Eigenwerte von  $\Sigma$

$\mathbf{D}$ : Faktorenmatrix, Faktorladungsmatrix

$d_{jk}$ : Elemente von  $\mathbf{D}$ : Faktorladungen,  $j = 1, \dots, p$ ,  $k = 1, \dots, p$

# Faktorenanalyse

Die Matrix  $\mathbf{D}$  beschreibt die Beziehungen zwischen Faktoren  $\mathbf{F}$  und den Variablen  $\mathbf{X}$

$$\text{cov}(\mathbf{X}, \mathbf{F}) = \text{cov}(\mathbf{D} \cdot \mathbf{F}, \mathbf{F}) = \mathbf{D} \text{cov} \mathbf{F} = \mathbf{D}$$

und, da die  $X$  standardisiert sind, ist  $d_{jk}$  die Korrelation zwischen Variable  $X_j$  und Faktor  $F_k$ .

Erinnerung:

$$\mathbf{D}\mathbf{D}' = \mathbf{\Sigma}$$

$\mathbf{\Sigma}$ : Korrelationsmatrix, also insbesondere

$$1 = \sigma_{jj} = \sum_{k=1}^p d_{jk} d_{kj} = \sum_{k=1}^p d_{jk}^2 \quad \forall j.$$

# Faktorenanalyse

Wir haben also gesehen, daß die vollständige Faktorenlösung mit der Hauptkomponentenmethode bestimmt werden kann.

Ziel: Möglichst wenig Faktoren  $F_1 \dots F_r$ ,  $r < p$

mit möglichst wenig Informationsverlust.

Erinnerung: vollständige Faktorenlösung

$$X_j = d_{j1}F_1 + \dots + d_{jr}F_r + \underbrace{d_{jr+1}F_{r+1} + \dots + d_{jq}F_q}_{=:\epsilon_j}$$

$$X_j = d_{j1}F_1 + \dots + d_{jr}F_r + \epsilon_j \quad j = 1 \dots p$$

Faktoranalyse Modell.

# Faktorenanalyse

Zerlegen für jedes Merkmal  $X_j$  die Quadratsummen

$$\sum_{k=1}^q d_{jk}^2 = 1 = \underbrace{\sum_{k=1}^r d_{jk}^2}_{\text{Var}(d_{j1}F_1 + \dots + d_{jr}F_r)} + \underbrace{\sum_{k=r+1}^q d_{jk}^2}_{\text{Var}(d_{jr+1}F_{r+1} + \dots + d_{jp}F_p)}$$

Erinnerung: Die Faktoren sind unabhängig.

Wenn wir  $r$  Faktoren auswählen, dann ist der erste Summand durch die  $r$  Faktoren bedingt, die sogenannte Kommunalität.

Der zweite Summand ist die sogenannte Restvarianz - diese ist variablenspezifisch.

Die Kommunalität beschreibt also wie gut die Variable durch die  $r$ -Faktoren beschrieben werden kann.

# Faktorenanalyse

Ziel: hohe Kommunalität (und kleine Restvarianzen)

Ziel:  $r$  klein

Beide Ziele sind gegenläufig.

Vorgehen:

1. Vollständige Faktorlösung bestimmen
2. Anzahl der Faktoren anhand der Eigenwerte von  $\Sigma$  bestimmen (als ersten Ansatz).

Erinnerung: Die höchsten Eigenwerte liefern die ersten

Hauptkomponenten, den größten Anteil an Gesamtvarianz.

Als Kriterium für die Wahl der Anzahl der Faktoren kann der

Scree-Plot (Option SCREE) dienen: “Knick” in der “Eigenwertkurve”.

# Faktorenanalyse

Evtl. ist eine Interpretation der Faktoren möglich. Meistens jedoch nicht.

Deshalb werden die Faktoren rotiert.

- Die Unabhängigkeit der Faktoren bleibt erhalten
- Die Varianzanteile ändern sich.

# Faktorenanalyse

## Varimax-Rotation (Standard)

Die Rotation der Faktoren geschieht so dass die neuen Faktoren möglichst dicht an den Koordinatenachsen liegen (jede Variable wird möglichst nur einem Faktor zugeordnet),  
eine sogenannte “Einfachstruktur”

heißt: Korrelationen  $\rho_{jk}$  in der Faktorenmatrix sind -1, 0, 1 möglichst nahe. ( $j = 1 \dots p, k = 1 \dots r$ )

$$v_k^2 := \sum_{j=1}^p (\rho_{jk}^2)^2 - \frac{1}{p} \left( \sum_{j=1}^p \rho_{jk}^2 \right)^2$$

ist bis auf Faktor  $\frac{1}{p-1}$  die Varianz der quadrierten Elemente der  $k$ -ten Spalte der Faktormatrix.

# Faktorenanalyse

rohes Varimax-Kriterium:

$$v_{roh}^2 = \sum_{k=1}^q v_k^2$$

ist zu maximieren (deshalb Varimax)

Praxis: zusätzl. Berücksichtigung der Kommunalitäten

ersetzen  $\rho_{jk}^2$  durch  $\frac{\rho_{jk}^2}{c_j^2}$ .,

wobei  $c_j^2$  die Kommunalität der (standardisierten) Variable  $X_j$  ist.

# Faktorenanalyse

## Varimax-Kriterium

Maximiere:

$$v^2 = \sum_{k=1}^r \left[ \sum_{j=1}^q \left( \frac{\rho_{jk}^2}{c_j^2} \right)^2 - \frac{1}{q} \left( \sum_{j=1}^q \frac{\rho_{jk}^2}{c_j^2} \right)^2 \right].$$

Praktische Durchführung der Rotation:

- $r = 2$  analytisch.
- $r > 2$  nehmen jeweils 2 Faktoren, halten die anderen fest und drehen nach dem Varimax-Kriterium. Das geschieht mit allen Paaren. (Nacheinanderausführung von Drehungen ist wieder eine Drehung.)
- Iteration bis sich das Varimax-Kriterium nicht mehr wesentlich

# Faktorenanalyse

Durchführung mit SAS:

```
PROC FACTOR ROTATE=VARIMAX NFACTORS=n;
```

```
VAR Variablen; RUN;
```

```
FaktoranayseBanknoten.sas
```

## **Ausgabe:**

Korrelationskoeffizienten

Eigenwerte von  $\Sigma$

Faktorladungsmatrizen

Kommunalitäten

rotierte Faktorladungsmatrizen

Rotation ist optional.

# Zusammenfassung (1)

## Basiswissen

- Klassifikation von Merkmalen
- Wahrscheinlichkeit
- Zufallsvariable
- Diskrete Zufallsvariablen (insbes. Binomial)
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert, Varianz
- Gesetz der großen Zahlen,  
Zentraler Grenzwertsatz

# Zusammenfassung (2)

Beschreibende Statistik

(Robuste) Lage- und Skalenschätzungen

```
PROC UNIVARIATE TRIMMED=Zahl ROBUSTSCALE; RUN;
```

Boxplots

```
PROC BOXPLOT; PLOT Variable*Faktor  
/BOXSTYLE=SCHEMATIC; RUN;
```

Häufigkeitsdiagramme:

```
PATTERN1 ...; PROC GCHART; VBAR Variable; RUN;
```

# Zusammenfassung (2)

## Beschreibende Statistik

(Robuste) Lage- und Skalenschätzungen

```
PROC UNIVARIATE TRIMMED=Zahl ROBUSTSCALE; RUN;
```

Boxplots

```
PROC BOXPLOT; PLOT Variable*Faktor  
/BOXSTYLE=SCHEMATIC; RUN;
```

Häufigkeitsdiagramme:

```
PATTERN1 ...; PROC GCHART; VBAR Variable; RUN;
```

Scatterplots, Regressionsgerade:

```
SYMBOL1 ...;  
PROC GPLOT; PLOT y*x=1 / REGEQN; RUN;
```

# Zusammenfassung (3)

Statistische Tests

Testproblem: Nullhypothese - Alternative, z.B.

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Entscheidung für  $H_0$ /gegen  $H_0$ : anhand einer

# Zusammenfassung (3)

## Statistische Tests

Testproblem: Nullhypothese - Alternative, z.B.

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Entscheidung für  $H_0$ /gegen  $H_0$ : anhand einer

Teststatistik, z.B.

$$T = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

# Zusammenfassung (3)

## Statistische Tests

Testproblem: Nullhypothese - Alternative, z.B.

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Entscheidung für  $H_0$ /gegen  $H_0$ : anhand einer

Teststatistik, z.B.

$$T = \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n}$$

Entscheidung

$$|t| > t_{krit} \Rightarrow H_0 \text{ ablehnen, } P(|T| > t_{krit}) = \alpha$$

$\alpha$  : Fehler 1. Art, Signifikanzniveau (in der Regel vorgegeben)

# Zusammenfassung (4)

## Statistische Tests (2)

### p-Wert (zweiseitig)

$P(|T| > t)$ , wobei  $t$ : Realisierung von  $T$

p-Wert  $< \alpha \Rightarrow H_0$  ablehnen

p-Wert  $\geq \alpha \Rightarrow H_0$  nicht ablehnen

# Zusammenfassung (4)

## Statistische Tests (2)

### p-Wert (zweiseitig)

$P(|T| > t)$ , wobei  $t$ : Realisierung von  $T$

p-Wert  $< \alpha \Rightarrow H_0$  ablehnen

p-Wert  $\geq \alpha \Rightarrow H_0$  nicht ablehnen

### Gütefunktion

$P(H_0 \text{ abgelehnt} | \mu \text{ richtig}) = \beta(\mu)$

Fehler 2.Art:  $1 - \beta(\mu)$

Wir betrachten Tests mit einer vergleichsweise hohen Gütefunktion.

# Zusammenfassung (5)

## Einseitige Tests

Alternative geht in eine Richtung, (aus sachlichen Gründen kann es nur eine Richtung geben)

$$\text{z.B. } \mu > \mu_0$$

## Zweiseitige Tests

Alternative geht in alle Richtungen,

$$\text{z.B. } \mu \neq \mu_0$$

k	unverbunden	verbunden
1	Einstichproben t-Test, Vorzeichen-Wilcoxon-Test	
	PROC UNIVARIATE; o. PROC TTEST H0=Wert; VAR Variable; RUN	
2	t-Test	t-Test
	PROC TTEST; CLASS=Faktor; VAR Variable; RUN;	PROC TTEST; PAIRED Var1*Var2; RUN;
	Wilcoxon-Test	Vorzeichen-Wilcoxon-Test
	PROC NPAR1WAY WILCOXON CLASS=Faktor;	diff=a-b; PROC UNIVARIATE.

einfache Varianzanalyse = einfaktorielle VA	einfaches Blockexperiment = zweifaktorielle VA
PROC ANOVA; (GLM) CLASS Faktor; MODEL Y=Faktor; RUN;	PROC GLM; CLASS FaktorA FaktorB; MODEL Y=FaktorA FaktorB; RUN;
Kruskal-Wallis-Test	Friedman-Test
PROC NPAR1WAY Wilcoxon; CLASS Faktor; VAR var; RUN;	PROC FREQ; TABLES FaktorA*FaktorB*Y / CMH2 SCORES=RANK NOPRINT; RUN;

# Zusammenfassung (8)

Anpassungstest auf Normalverteilung:

```
PROC UNIVARIATE NORMAL; VAR var; RUN;
```

Shapiro-Wilk-Test oder Anderson-Darling-Test

## Zusammenfassung (8)

Anpassungstest auf Normalverteilung:

```
PROC UNIVARIATE NORMAL; VAR var; RUN;
```

Shapiro-Wilk-Test oder Anderson-Darling-Test

Anpassungstest auf Verteilung mit begrenzter Anzahl von Ausprägungen

```
PROC FREQ; TABLES Var1 /CHISQ NOPRINT
```

```
TESTP=(p1,p2,...pk);
```

```
RUN;
```

( $p_1, \dots, p_k$  vorher ausrechnen)

# Zusammenfassung (9)

Test auf Korrelation (metrisch oder ordinal skalierte Merkmale)

```
PROC CORR PEARSON SPEARMAN KENDALL;
```

```
RUN;
```

## Zusammenfassung (9)

Test auf Korrelation (metrisch oder ordinal skalierte Merkmale)

```
PROC CORR PEARSON SPEARMAN KENDALL;  
RUN;
```

Test auf Unabhängigkeit (beliebig skalierte Merkmale):

```
PROC FREQ;  
TABLES Var1*Var2 /CHISQ NOPRINT; RUN;
```

# Zusammenfassung (10)

## Lineare Regression (1)

### Parameterschätzung und Test

```
PROC REG;
```

```
MODEL Y=Var1 Var2 ... Varn / CLI CLM R; TEST Var1=0
```

```
Var2=0; /*Zusaetzl.Hypothesen */ RUN;
```

# Zusammenfassung (10)

## Lineare Regression (1)

### Parameterschätzung und Test

```
PROC REG;
```

```
MODEL Y=Var1 Var2 ... Varn / CLI CLM R; TEST Var1=0  
Var2=0; /*Zusaetzl.Hypothesen */ RUN;
```

### Modellwahl

```
PROC REG;
```

```
MODEL Y=Var1 Var2 ... Varn / SELECTION=backward;  
RUN;
```

# Zusammenfassung (11)

## Lineare Regression (2)

### Residualanalyse

**PROC REG;**

**MODEL** Y=Var1 Var2 ... Varn / **R**;

**PLOT** rstudent.\*obs.; /\*und/oder\*/

**PLOT** residual.\*y; residual.\*predicted.;

**RUN;**

und evtl. Test auf Normalverteilung.

# Zusammenfassung (12)

Sonstige Regressionsverfahren, nur Übersicht

Robuste Lineare Regression

Nichtlineare Regression

Nichtparametrische Regression

Logistische Regression

## Zusammenfassung (13)

```
PROC CLUSTER /*hierarchische Clusteranalyse*/  
  METHOD=methode  
  STANDARD /*Standardisierung*/  
  OUTREE=datei; /*Eingabedatei für Proc Tree*/  
RUN;
```

## Zusammenfassung (13)

**PROC CLUSTER** /\*hierarchische Clusteranalyse\*/

**METHOD**=methode

**STANDARD** /\*Standardisierung\*/

**OUTTREE**=datei; /\*Eingabedatei für Proc Tree\*/

**RUN**;

**PROC TREE DATA**=datei

**OUT**=out /\*Ausgabedatei z.B.für PROC GPLOT\*/

**NCLUSTERS**=nc /\*Anz. Cluster\*/

**COPY** vars /\*vars in die Ausgabedatei\*/

**RUN**;

## Zusammenfassung (13)

**PROC CLUSTER** /\*hierarchische Clusteranalyse\*/

METHOD=methode

STANDARD /\*Standardisierung\*/

OUTREE=datei; /\*Eingabedatei für Proc Tree\*/

**RUN;**

**PROC TREE** DATA=datei

OUT=out /\*Ausgabedatei z.B.für PROC GPLOT\*/

NCLUSTERS=nc /\*Anz. Cluster\*/

COPY vars /\*vars in die Ausgabedatei\*/

**RUN;**

**PROC GPLOT;**

PLOT variablen=cluster; /\*Symbol-Anweis. vorher definieren\*/

**RUN;**

# Zusammenfassung (14)

## Konfidenzbereiche

für Parameter im Regressionsmodell

```
PROC REG;  
MODEL Y=var1...varn/ CLI CLM;  
RUN;
```

Grafische Darstellung von Konfidenzbereichen bei der Regression

```
SYMBOL1 I=RLCLI95;  
PROC GPLOT; PLOT y*x=1; RUN;
```

# Zusammenfassung (15)

Wichtige Sprachelemente

## Normalverteilte Zufallsvariable

mit zufälligem Startwert: `seed=-1; RANNOR(seed);`

## Gleichverteilte Zufallsvariable

mit zufälligem Startwert: `seed=-1; RANUNI(seed);`

# Zusammenfassung (16)

Wahrscheinlichkeitsverteilungen:

Verteilungsfunktion (Parameter)

$CDF('Verteilung', z, \text{Parameterliste})$

Dichte oder Wahrscheinlichkeitsfunktion (Parameter)

$PDF('Verteilung', z, \text{Parameterliste})$

z.B.: ('normal', z, 0, 1)

('binomial', z, n, p)

Quantile

Standardnormal:  $PROBIT(u)$ ,  $u \in (0, 1)$ .

$Quantile('Verteilung', z, \text{Parameterliste})$

# Zusammenfassung (17)

Hauptkomponentenanalyse, nur Übersicht

**PROC PRINCOMP**

# Übungen (1)

1. Folgen und Reihen, Potenzreihen
2. Differential- und Integralrechnung, Normalverteilung
3. Integralrechnung, Rechnen mit Erwartungswerten, Kovarianz, Korrelation, Berechnen von Korrelationen
4. Normalverteilung und Binomialverteilung, Berechnen von robusten Lage- und Skalenschätzungen
5. Cauchy- und Paretoverteilung, Dichtefunktionen, Voraussetzungen des Zentralen Grenzwertsatzes
6. Berechnen von Korrelationen, Einfluss von Ausreißern
7. Motivation der  $\chi^2$ - und  $t$ -Verteilung, Hypothesentests
8.  $t$ -Test und Varianzanalyse

# Übungen (2)

9. Aufgabenblatt 9, Untersuchung der Robustheit des  $t$ -Tests  
Produkt von Matrizen, Eigenwerte, Eigenvektoren
10. Lineare Algebra, Matrizenrechnung, Idempotenz,  $\chi^2$ -Verteilung
11. Erwartungswert und Varianz der Wilcoxon-Vorzeichen  
Rangstatistik,  
Transformation von normalverteilten Zufallsvariablen,  
 $\chi^2$ -Verteilung
12. Zusammenfassung

# Übungsaufgaben

7,8,9 Wahrscheinlichkeitsverteilungen

10,11 Statist. Maßzahlen, Boxplots

11 Histogramme, Dichteschätzung

14,15,26,30,33,34,35 Korrelation, Unabhängigkeit, Lineare  
Regression

16-18,20-22,23-25 Lagetests, Anpassungstests

20,23 Varianzanalyse

27-29,31-32 Nichtparametrische Tests

36,37 Zufallszahlen

37 Clusteranalyse