Einführung in die Stochastik

Vorlesung SS 2009

Prof. Dr. Michael Kohler
Fachbereich Mathematik
Technische Universität Darmstadt

kohler@mathematik.tu-darmstadt.de

Kapitel 1: Motivation

Stochastik – wozu braucht man das ?

1.1 Statistik-Prüfung, Sommer 2002

Ergebnis der Vordiplomsprüfung "Statistik II für WirtschaftswissenschaftlerInnen" am 31.07.2002:

Anzahl Teilnehmer	:	295
Notendurchschnitt	:	2,68
Durchfallquote	:	5,4 %

StudentInnenen hatten die Möglichkeit, freiwillig einen Übungsschein zu erwerben.

Anzahl Teilnehmer mit Statistik-Schein	:	190
Notendurchschnitt	:	2,46
Durchfallquote	:	3,16 %

Anzahl Teilnehmer ohne Statistik–Schein	:	105
Notendurchschnitt	:	3,07
Durchfallquote	:	9,52 %

Was folgt daraus hinsichtlich des Einflusses des Erwerbs des Statistik–Übungsscheines

- auf die Note ?
- auf das Bestehen der Prüfung ?

1.2 Sex und Herzinfarkt

Studie in Caerphilly (Wales), 1979-2003:

914 gesunde Männer im Alter von 45 bis 95 Jahren wurden zufällig ausgewählt, unter anderem zu ihrem Sexualleben befragt und über einen Zeitraum von 10 Jahren beobachtet.

Resultat:

	Gesamt	≥ 2 Orgasmen / W.	< 1 Orgasmus / M.
Alle	914 (100%)	231 (25,3%)	197 (21,5%)
Herzinfarkte	105 (11,5%)	19 (8,2%)	33 (16,8%)

Was folgt daraus ?

1.3 Die Challenger-Katastrophe

Start der Raumfähre Challenger am 28. Januar 1986:

Raumfähre explodiert genau 73 Sekunden nach dem Start, alle 7 Astronauten sterben.

Grund: Dichtungsringe, die aufgrund der geringen Außentemperatur von unter 0 Grad beim Start undicht geworden waren.

Am Tag vor dem Start:
Experten von Morton Thiokol, dem Hersteller der Triebwerke, hatten angesichts der geringen vorhergesagten Außentemperatur Bedenken hinsichtlich der Dichtungsringe und empfahlen, den Start zu verschieben.

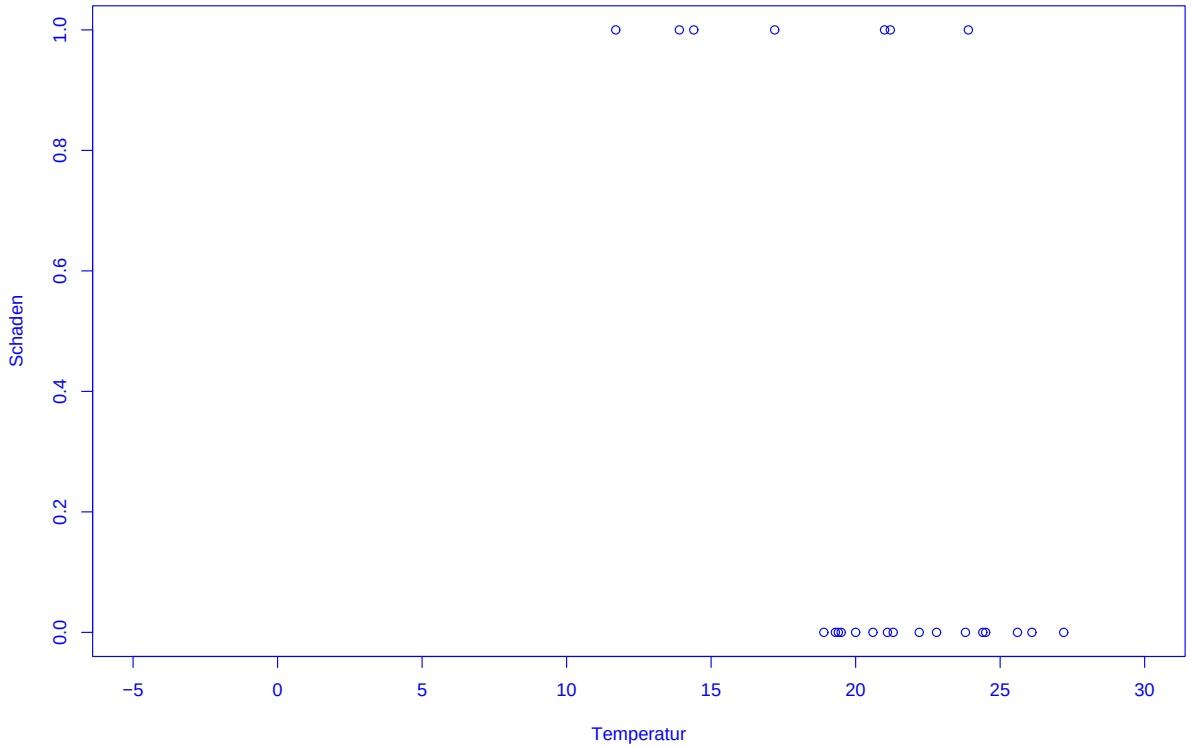
Zur Begründung verwendete Daten:

Flugnummer	Datum	Temperatur (in Grad Celsius)
STS-2	12.11.81	21,1
41-B	03.02.84	13,9
41-C	06.04.84	17,2
41-D	30.08.84	21,1
51-C	24.01.85	11,7
61-A	30.10.85	23,9
61-C	12.01.86	14,4

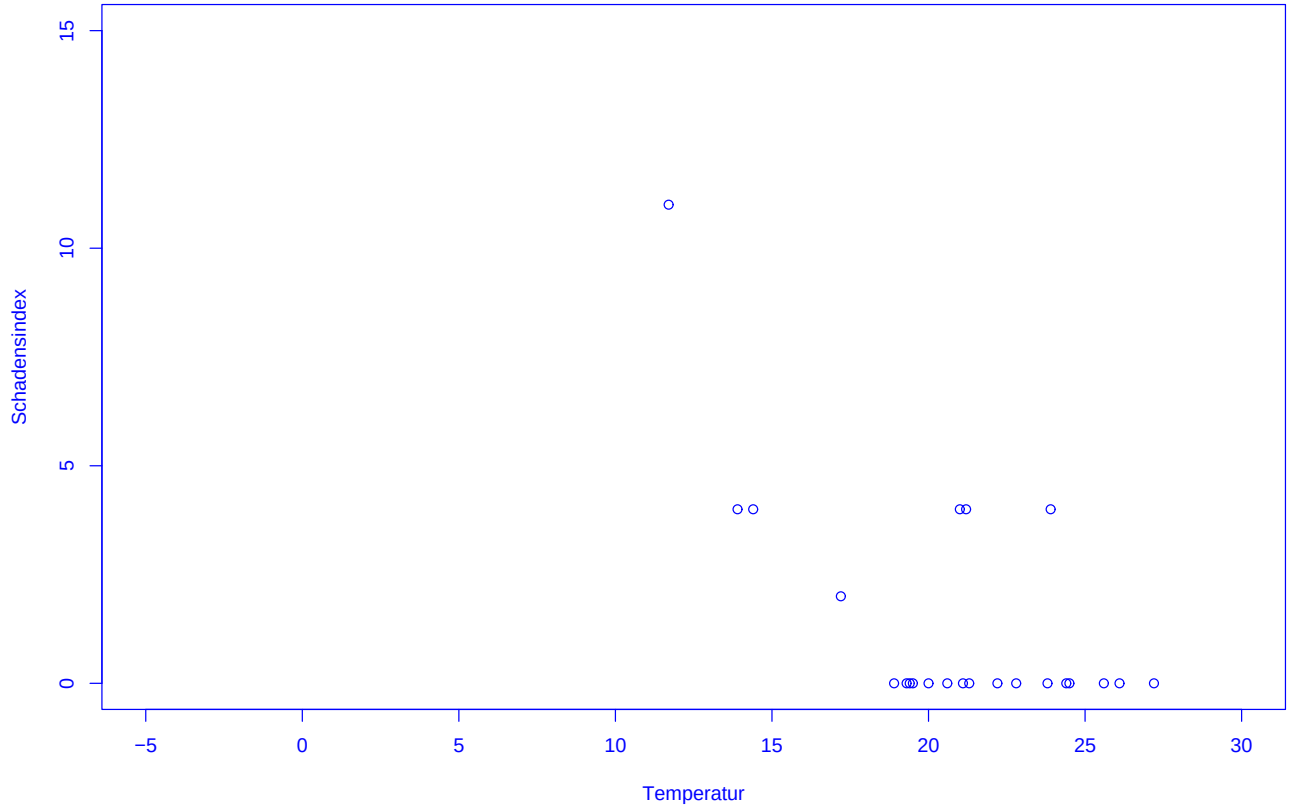
War für NASA leider nicht nachvollziehbar ...

Probleme bei der Analyse dieser Daten:

1. Flüge ohne Schädigungen nicht berücksichtigt.



2. Stärke der Schädigungen nicht in Abhängigkeit von der Temperatur dargestellt.



1.4 Präsidentschaftswahl in den USA, Herbst 2000

Auszählung der Präsidentschaftswahl in den USA:

Pro Bundesstaat werden die gültigen abgegebenen Stimmen pro Kandidat ermittelt. Wer die meisten Stimmen erhält, bekommt die Wahlmänner/-frauen zugesprochen, die für diesen Bundesstaat zu vergeben sind.

Wozu braucht man da Stochastik ?

Problem im Herbst 2000:

In Florida gewann George Bush die 25 Wahlmänner/-frauen mit einem Vorsprung von nur 537 Stimmen.

Al Gore versuchte danach, in einer Reihe von Prozessen eine (teilweise) manuelle Nachzählung der Stimmen zu erreichen.

Zentraler Streitpunkt:

Stimmabgabe erfolgte durch Lochung von Lochkarten.

Soll man auch unvollständig gelochte Lochkarten (ca. 2 % der Stimmen) berücksichtigen ?

Im Prozess vor dem Supreme Court in Florida hat Statistik Professor Nicholas Hengartner aus Yale für Al Gore ausgesagt.

Sein Argument:

Unabsichtliche unvollständige Lochung tritt bei Kandidaten, die wie Al Gore auf der linken Seite der Lochkarte stehen, besonders häufig auf.

Problem: Konnte nicht bewiesen werden . . .

1.5 Positionsbestimmung mittels GPS

Anwendung:

- Navigation von Flugzeugen, Schiffen und Autos
- Erdbebenfrühwarnsysteme

Idee:

Kennt man den Abstand seiner Position zu drei Punkten im Raum, so kann man diese durch Schnitt dreier Kugeloberflächen bestimmen.

Grundlage:

ca. 30 Satellitten, die die Erde in ca. 20200 km Höhe umkreisen und im Sekundentakt Position und Signalaussendezeit zur Erde senden. Bestimme daraus Abstand zu den Satellitten durch Vergleich der Empfangszeit mit der Aussendezeit.

Probleme:

- Uhrenfehler
- Signalgeschwindigkeit schwankt aufgrund von Veränderungen in der Ionosphäre.

Lösung:

Verwende Signale von 4 bis 5 Satellitten und wende statistische Verfahren an, um Fehler bei der Abstandsbestimmung auszugleichen.

1.6 Anwendung der Stochastik in der Finanzmathematik

In der modernen Finanzmathematik modelliert man den zukünftigen **unbestimmten Wert einer Finanzinvestition** (z.B. in eine Aktie) mit Hilfe der Stochastik als **zufälligen Wert**.

Fragestellungen der modernen Finanzmathematik:

1. Bewertung von Optionen

   Was ist das Recht Wert, eine (konkrete) Aktie in der Zukunft zu einem bereits jetzt festgelegten Preis verkaufen zu dürfen ?

2. Beurteilung des Risikos von Kapitalanlagen

   Wieviel Geld wird eine Bank, die Geld in verschiedene Aktien und andere Anlagen investiert hat, voraussichtlich verlieren, falls es zu Kurseinbrüchen an der Aktienbörse kommt ?

3. Portfoliooptimierung

Wie verteilt man einen festen Geldbetrag optimal auf verschiedene Anlageprodukte (z.B. Festgeld und verschiedene Aktien) ?

1.7 Anwendung der Stochastik in der Versicherungsmathematik

Bei einer Versicherung bietet das Versicherungsunternehmen an, gegen Erhalt eines im voraus fälligen Geldbetrages (Prämie) bei Eintritt von näher definierten ungewissen Ereignissen (Schäden) gewisse meist vom betreffenden Ereignis abhängende Zahlungen an den Versicherungsnehmer zu leisten.

In der Versicherungsmathematik werden diese **ungewissen Schäden** mit Hilfe der Stochastik **als zufällig modelliert**.

Zentrale Fragen sind dann:

- Wie groß sind die Schäden im Mittel ?
- Wieviel Geld rechnet man in die Prämie ein für die Schwankungen der Schäden um den Mittelwert ?

- Wie berücksichtigt man späte Schadenmanifestation ?

- Wie berechnet man die Prämie bei Übernahme nur eines Teils der Schadenhöhe (z.B. Selbstbeteiligung, Deckungssumme, Rückversicherung) ?

Schön, aber:

Braucht man Stochastik als **MathematikstudentIn** wirklich?

z.B.:

- um das Fach später selber in einer Schule unterrichten zu können . . .
- um im Rahmen von darauf aufbauenden Vorlesungen (wie z.B. Finanz- und Versicherungsmathematik) nützliches Wissen für den späteren Beruf erwerben zu können . . .

Anmerkung: Zum Jahr der Mathematik haben in dem Buch

Mathematik - Motor der Wirtschaft

20 große Unternehmen erläutert, wo in ihrem Unternehmen Mathematik zum Einsatz kommt. Bei **13** der Unternehmen stammten die Anwendungsgebiete aus der **Stochastik**.

Sollten Sie sich in Stochastik vertiefen wollen, ist dazu der Besuch der Vorlesung

Probability theory bzw. Wahrscheinlichkeitstheorie

Voraussetzung. Diese werde ich im WS 2009/10 halten, und darauf aufbauend desweiteren folgende Vorlesungen anbieten:

- Einführung in die Finanzmathematik (SS 10),
- Mathematische Statistik (WS 10/11),
- Schadenversicherungsmathematik (SS 11).

Desweiteren werde ich im SS 10 ein *Bachelor-Seminar* zur Stochastik anbieten, in dessen Anschluss direkt eine Bachelor-Arbeit verfasst werden kann. Für das SS 11 ist ein *Master-Seminar* geplant.

Ziel der Vorlesung "Einführung in die Stochastik":

Erlernen der wichtigsten Grundprinzipien der Wahrscheinlichkeitstheorie und der Statistik, so dass man die Frage beantworten kann:

> Wie modelliert man zufällige Phänomene mathematisch, und was fängt man damit an ?

Ein tiefes Verständnis des behandelten Stoffes wird aber erst in der Vorlesung "Wahrscheinlichkeitstheorie" und darauf aufbauenden Veranstaltungen vermittelt.

Gliederung der Vorlesung "Einführung in die Stochastik":

- Kapitel 1: Einführung (heute)
- Kapitel 2: Erhebung von Daten im Rahmen von Studien und Umfragen (2V)
- Kapitel 3: Beschreibende Statistik (3V)
- Kapitel 4: Einführung in die W-Theorie (14V)
- Kapitel 5: Schließende Statistik (7V)

Die schriftliche Prüfung zur Vorlesung "Einführung in die Stochastik" findet am

Montag, 17.08.2009

statt.

Für StudentInnen, die ihre Prüfungsnote nachweislich schon sehr früh benötigen, wird am

Freitag, 17.07.2009

ein separater Prüfungstermin angeboten. Alle anderen StudentInnen bekommen diese Klausur als $Probeklausur$ zur Verfügung gestellt.

Zum Niveau dieser Vorlesung:

Verschiedene Ebenen des **“Lernens”**:

1. Wissen, was es gibt.
2. Verstehen, wie es funktioniert.
3. Anwenden können.
4. Analysieren können.
5. Synthetisieren können.
6. Bewerten können.

Ziel der Ausbildung an der Universität ist die letzte Ebene.

Dazu ist in Stochastik (wie in jeder Vorlesung aus der Mathematik) ein gewisses Abstraktionsniveau unabdingbar !!!

Zum didaktischen Konzept dieser Vorlesung:

Lehr-Lern-Kurzschluss:

Gelernt wird nicht, was gelehrt wird!

Was ich hier mache:

Bereitsstellung einer "Umgebung", in der **Sie** möglichst einfach möglichst viel über Stochastik **lernen können**.

Spezielle “Tricks” dabei:

- Wiederholungsfolie zu Beginn
- Pause bzw. Minitest in der Mitte
- Umfrage am Schluss
- Intensiver Übungsbetrieb
- Begleitendes Buch (s.u.)
- Recording der Vorlesung

und ganz wichtig:

Motivierung der StudentInnen !

Was können bzw. sollten Sie tun, um in dieser Vorlesung erfolgreich zu sein ?

AKTIV AN DIESER VERANSTALTUNG TEILNEHMEN, d.h.

- anwesend sein (bei Vorlesung und Gruppenübung).
- Vorlesung nach jedem Termin kurz nacharbeiten (ca. 5-10 Minuten genügen dazu).
- Übungsaufgaben in Gruppen aktiv bearbeiten.
- Bei Unklarheiten: FRAGEN!

TERMINE

1. Vorlesung:

   - Mittwoch, 14:25 Uhr - 15:55 Uhr, in S 311/08
   - Freitag, 9:50 Uhr - 11:30 Uhr, in S 311/0012

2. Tutorium (für das erste Semester) und Gruppenübungen:

   Siehe Homepage der Vorlesung:

   https://www3.mathematik.tu-darmstadt.de/fb/mathe/lehre-und-studium/elektronisches-veranstaltungssystem.html?evsid=23&evsver=102

Begleitendes Buch zur Vorlesung:

Judith Eckle-Kohler und Michael Kohler:
Eine Einführung in die Statistik und ihre Anwendungen.
Springer 2009. Ca. EUR 25.

Kapitel 2: Erhebung von Daten

Wie Daten entstehen bestimmt mit, welche Schlüsse man später daraus ziehen kann (bzgl. Verallgemeinerungen von Aussagen über den vorligenden Datensatz hinaus).

Im Folgenden betrachten wir die Erhebung von Daten im Zusammenhang mit Studien und Umfragen.

2.1 Kontrollierte Studien

Beispiel: Überprüfung der Wirksamkeit der Anti-Grippe-Pille Tamiflu (1997/98)

Wie stellt man fest, ob eine im Labor erfolgreich getestete Anti-Grippe-Pille auch in der realen Welt hilft ?

Vorgehen in drei Phasen üblich:

- Phase 1: Test auf Nebenwirkung an kleiner Gruppe gesunder Menschen.
- Phase 2: Überprüfung der Wirksamkeit an kleiner Gruppe Grippekranker.
- Phase 3: Überprüfung der Wirksamkeit unter realistischen Bedingungen an Hunderten von Menschen.

Grundidee bei Phasen II / III: Vergleiche Studiengruppe (SG) bestehend aus mit neuem Medikament behandelten Grippekranken mit Kontrollgruppe (KG) bestehend aus traditionell behandelten Grippekranken.

Vorgehen 1: Retrospektiv kontrollierte Studie

Größere Anzahl Grippekranker mit neuem Medikament behandeln (SG). Nach einiger Zeit durchschnittliche Krankheitsdauer bestimmen. Vergleichen mit durchschnittlicher Krankheitsdauer von in der Vergangenheit an Grippe erkrankten Personen (KG).

Vergleich von durchschnittlicher Behandlungsdauer ermöglicht Vernachlässigung von Unterschieden bei den Gruppengrößen.

Problem: Grippe tritt in Epidemien auf und Grippe-Virus verändert sich Jahr für Jahr stark.

Vorgehen 2: Prospektiv kontrollierte Studie ohne Randomisierung

Größere Zahl von Grippekranken auswählen. Diejenigen, die einverstanden sind, mit neuem Medikament behandeln (SG). Rest bildet die KG. Nach einiger Zeit durchschnittliche Krankheitsdauern vergleichen.

Hier entscheiden die Grippekranken, ob sie zur SG oder zur KG gehören.

Problem: KG unterscheidet sich nicht nur durch Behandlung von SG. Z.B. denkbar: Besonders viele ältere Grippekranke, bei denen es oft zu Komplikationen wie z.B. Lungenentzündung kommt, stimmen neuer Behandlungsmethode zu.

$\Rightarrow$ Einfluss der Behandlung konfundiert (vermengt sich) mit Einfluss des Alters der Grippekranken.

Möglicher Ausweg: KG so wählen, dass möglichst ähnlich (z.B. bzgl. Alter, ...) zu SG.

Nachteil: Fehleranfällig !

Vorgehen 3: Prospektiv kontrollierte Studie mit Randomisierung

Nur Grippekranke betrachten, die mit der neuen Behandlungsmethode einverstanden sind. Diese zufällig (z.B. durch Münzwürfe) in SG und KG aufteilen. SG mit neuem Medikament behandeln, KG nicht. Nach einiger Zeit durchschnittliche Krankheitsdauern vergleichen.

Studie wurde gemäß Vorgehen 3 in den Jahren 1997/98 durchgeführt. Weitere Aspekte dabei:

a) Um Einfluss des neuen Medikaments vom Einfluss der Einnahme einer Tablette zu unterscheiden, wurden den Personen in der KG eine gleich aussehende Tablette ohne Wirkstoff (sog. Placebo) verabreicht.

b) Um Beeinflussung der (manchmal schwierigen) Beurteilung der Symptome von Grippe zu vermeiden, wurde den behandelnden Ärzten nicht mitgeteilt, ob ein Grippekranker zur SG oder zur KG gehört.

a) und b): doppelte Blindstudie

c) Um sicherzustellen, dass SG (und KG) einen hohen Anteil an Grippekranken enthält, wurden nur dort Personen in die Studie aufgenommen, wo in der Woche davor durch Halsabstriche mindestens zwei Grippefälle nachgewiesen wurden.

Ergebnis der Studie:

Einnahme des neuen Medikaments innerhalb von 36 Stunden nach Auftreten der ersten Symptome führt dazu, dass die Grippe etwa eineinhalb Tage früher abgeklingt.

Medikament ist seit Mitte 2002 unter dem Namen Tamiflu in Apotheken erhältlich.

Lohnt sich der Aufwand einer
prospektiv kontrollierten Studie mit Randomisierung ?

Beispiel: Wirkt sich die Einnahme von Vitamin E positiv auf das Auftreten von Gefäßerkrankung am Herzen (die z.B. zu Herzinfarkten) führen aus ?

Beobachtungsstudie in den USA (Nurses Health Study)

Ab dem Jahr 1980 wurden mehr als 87000 Krankenschwestern zu ihrer Ernährung befragt und anschließend über 8 Jahre hinweg beobachtet.

Resultat: 34% weniger Gefäßerkrankungen bei denen, die viel Vitamin E zu sich nahmen.

Effekt trat auch noch nach Kontrolle von konfundierenden Faktoren auf.

Überprüfung des Resultats in einer kontrollierten Studie mit Randomisierung.

Zwischen 1994 und 2001 wurden 20536 Erwachsene mit Vorerkrankungen zufällig in Studien- und Kontrollgruppe unterteilt.

SG bekam täglich Tablette mit 600mg Vitamin E, 250mg Vitamin C und 20mg Beta-Karotin als Nahrungsmittelergänzung.

Resultat:

	Studiengruppe	Kontrollgruppe
Alle	10.288	10.288
Todesfälle	$1.446\ (14,1\%)$	$1.389\ (13,5\%)$
Todesfälle in Zusammenhang mit Gefäßerkrankungen	$878\ (8,6\%)$	$840\ (8,2\%)$
Herzinfarkt	$1.063\ (10,4\%)$	$1.047\ (10,2\%)$
Schlaganfall	$511\ (5,0\%)$	$518\ (5,0\%)$
Erstauftritt schwere Herzerkrankung	$2.306\ (22,5\%)$	$2.312\ (22,5\%)$

2.2 Beobachtungsstudien

Unterschied zu kontrollierten Studien:

Kontrollierte Studie (auch: geplanter Versuch):

Untersucht wird Einfluss einer Einwirkung (z.B. Impfung) auf Objekte (z.B. Kinder). Statistiker entscheidet, auf welche Objekte wie eingewirkt wird.

Beobachtungsstudie:

Die Objekte werden nur beobachtet, und während der Studie keinerlei Intervention ausgesetzt. Die Aufteilung der Objekte in SG und KG erfolgt hier immer anhand gewisser vorgegebener Merkmale der Objekte.

Hauptproblem bei Beobachtungsstudien:

Ist die KG wirklich ähnlich zur SG ?

Beispiel: Verursacht Rauchen Krankheiten ?

Vergleich Todesraten Raucher (SG) mit Todesraten Nichtraucher (KG).

Problem: Besonders viele Männer rauchen. Herzerkrankungen häufiger bei Männern als bei Frauen.

$\Rightarrow$ Geschlecht ist konfundierender Faktor.

Ausweg: Nur Gruppen vergleichen, bei denen dieser konfundierende Faktor übereinstimmt.

Vergleiche

- männliche Raucher (SG1) mit männlichen Nichtrauchern (KG1)
- weibliche Raucher (SG2) mit weiblichen Nichtrauchern (KG2)

Neues Problem: Es gibt weitere konfundierende Faktoren, z.B. Alter.

Nötig daher:

- Erkennung aller konfundierenden Faktoren
- Bildung von vielen Untergruppen

Beispiel: Beeinflusst Ultraschall das Geburtsgewicht von Kindern ?

Beobachtungsstudie am John Hopkins Krankenhaus, Baltimore:

Geburtsgewicht von Kindern, deren Mütter während der Schwangerschaft eine Ultraschalluntersuchung durchführen haben lassen, ist geringer als das von Kindern, bei denen bei der Mutter keine Ultraschalluntersuchung durchgeführt wurde.

Effekt besteht selbst dann, wenn eine Vielzahl von konfundierenden Faktoren (z.B. Rauchen, Alkoholgenuss, Ausbildung der Mutter, etc.) berücksichtigt wird.

Aber: Kontrollierte Studie mit Randomisierung ergab:

Geburtsgewicht nach Ultraschalluntersuchung sogar etwas höher als ohne Ultraschalluntersuchung.

Erklärung: In SG gaben überproportional viele Mütter das Rauchen auf.

Beispiel: Diskriminierung von Frauen bei der Zulassung zum Studium

Zulassungsdaten Universität Berkeley, Herbst 1973:

Für das Master-/PhD-Programm hatten sich 8442 Männer und 4321 Frauen beworden. Zugelassen wurden 44% der Männer und 35% der Frauen.

Folgt daraus, dass die Uni Berkely Frauen diskriminiert ?

Zulassungsdaten nach Fächern getrennt:

Fach	#Männer	Zugel.	#Frauen	Zugel.
A	825	62%	108	82%
B	560	63%	25	68 %
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Folgerung:

Wahl des Faches konfundiert mit Geschlecht, Frauen haben sich vor allem für Fächer beworben, in denen nur wenige zugelassen wurden.

Problem bei Studien:

Die Mehrzahl obiger Studien weist Assoziation aber nicht Kausalität nach.

Grund:

Existenz konfundierender Faktoren.

Diese haben Einfluss auf die Aufteilung in SG und KG und auf das beobachtete Resultat.

2.3 Umfragen

geg.: Menge von Objekten (Grundgesamtheit) mit Eigenschaften.

Ziel: Stelle fest, wie viele Objekte der Grundgesamtheit eine gewisse Eigenschaft haben.

Beispiel: Wie viele der Wahlberechtigten in der BRD würden für die einzelnen Parteien stimmen, wenn nächsten Sonntag Bundestagswahl wäre ?

Ergebnisse von Wahlumfragen ca. drei Wochen vor der Bundestagswahl am 22.09.2002:

	SPD	CDU/CSU	FDP	GRÜNE	PDS
Allensbach	35,2	38,2	11,2	7,2	4,9
Emnid	37	39	8	6	5
Forsa	39	39	9	7	4
Forschungsgruppe Wahlen	38	38	8	7	4
Infratest-dimap	38	39,5	8,5	7,5	4
amtliches Endergebnis	38,5	38,5	7,4	8,6	4,0

Problem bei Wahlumfragen: Befragung aller Wahlberechtigten zu aufwendig.

Ausweg: Befrage nur "kleine" Teilmenge (Stichprobe) der Grundgesamtheit und "schätze" mit Hilfe des Resultats die gesuchte Größe.

Fragen:

1. Wie wählt man die Stichprobe ?

2. Wie schätzt man ausgehend von der Stichprobe die gesuchte Größe ?

Mögliche Antwort im Beispiel oben:

1. Bestimme Stichprobe durch "rein zufällige" Auswahl von n Personen aus der Menge der Wahlberechtigten (z.B. $n = 2000$).

2. Schätze die prozentualen Anteile der Stimmen für die einzelnen Parteien in der Menge aller Wahlberechtigten durch die entsprechenden prozentualen Anteile in der Stichprobe.

Wir werden später sehen: 2. ist eine gute Idee.

Durchführung von 1. ???

Vorgehen 1: Befrage die Studenten einer Stochastik-Vorlesung.

Vorgehen 2: Befrage die ersten n Personen, die Montag morgens ab 10 Uhr einen festen Punkt der Fußgängerzone in Darmstadt passieren.

Vorgehen 3: Erstelle eine Liste aller Wahlberechtigten (mit Adresse). Wähle aus dieser "zufällig" n Personen aus und befrage diese.

Vorgehen 4: Wähle aus einem Telefonbuch für Deutschland rein zufällig Nummern aus und befrage die ersten n Personen, die man erreicht.

Vorgehen 5: Wähle zufällig Nummern am Telefon, und befrage die ersten n Privatpersonen, die sich melden.

Probleme:

- Vorgehen 3 ist zu aufwendig.

- Verzerrung durch Auswahl (sampling bias)

  Stichprobe ist nicht repräsentativ: Bestimmte Gruppen der Wahlberechtigten, deren Wahlverhalten vom Durchschnitt abweicht, sind überrepräsentiert, z.B.:

  - Studenten,
  - Einwohner von Darmstadt,
  - Personen, die dem Interviewer sympathisch sind,
  - Personen mit Eintrag im Telefonbuch,
  - Personen, die telefonisch leicht erreichbar sind,
  - Personen, die in einem kleinem Haushalt leben.

- Verzerrung durch Nicht–Antworten (non–response bias)

  Ein Teil der Befragten wird die Antwort verweigern. Deren Wahlverhalten kann vom Rest abweichen.

Beispiel: Wöchentliche Wahlumfrage von EMNID im Auftrag von n-tv:

1. Telefonisch werden pro Woche ca. 1000 Wahlberechtigte befragt.

2. Gewählte Telefonnummern werden zufällig aus Telefonbüchern und CD-ROMs ausgewählt. Dabei wird die letzte Ziffer zufällig modifiziert.

3. Innerhalb des so ausgewählten Haushalts wird die Zielperson durch Zufallsschlüssel ermittelt.

4. Schätzung wird durch gewichtete Mittelung der Angaben der Personen in der Stichprobe gebildet.

5. Gewichte berücksichtigen z.B. Haushaltsgröße, demographische Zusammensetzung der Menge der Wahlberechtigten, evt. auch angegebenes Abstimmungsverhalten bei letzter Bundestagswahl.

Kapitel 3: Deskriptive und explorative Statistik

geg.: Messreihe (Stichprobe, Datensatz):

$x_1, \ldots, x_n$ (n=Stichprobenumfang)

Aufgabe der deskriptiven (beschreibenden) Statistik:

Übersichtliche Darstellung von Eigenschaften dieser Messreihe.

Aufgabe der explorativen (erforschenden) Statistik:

Finden von (unbekannten) Strukturen.

Beispiel 1: Beschäftigungsquote der Männer zwischen 15 und 64 Jahren in 26 Ländern der europäischen Union im Jahr 2006 (Quelle: Eurostat):

67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2,
66.4, 63.9, 73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

Beispiel 2: Beschäftigungsquote der Frauen zwischen 15 und 64 Jahren in 26 Ländern der europäischen Union im Jahr 2006 (Quelle: Eurostat):

53.2, 55, 56.8, 73.2, 61.4, 66.4, 58.8, 47.5, 53.2, 57.7, 46.7, 59.8, 62.9,
61.1, 51.1, 34.6, 67.5, 63, 47.8, 62.4, 54.1, 63.3, 51.6, 68.1, 70.6, 65.8

Beispiel 3: Alter der ca. 32 Millionen männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD im Jahr 2001 (Quelle: Statistisches Bundesamt, Angabe in Jahren):

79, 2, 34, . . .

Typen von Messgrößen (Merkmalen, Variablen):

1. mögliche Unterteilung:

- diskret: endlich oder abzählbar unendlich viele Ausprägungen
- stetig: alle Werte eines Intervalls sind Ausprägungen

2. mögliche Unterteilung:

	Abstandbegriff vorhanden ?	Ordnungsrelation vorhanden ?
reell	ja	ja
ordinal	nein	ja
zirkulär	ja	nein
nominal	nein	nein

3.1 Histogramme

Häufigkeitstabelle:

- Einteilung der Daten in k Klassen (z.B. $k \approx \sqrt{n}$ oder $k \approx 10 \cdot \log_{10} n$),
- Ermittlung der Klassenhäufigkeiten n_i $(i = 1, \ldots, k)$,
- Darstellung des Resultats in einer Tabelle.

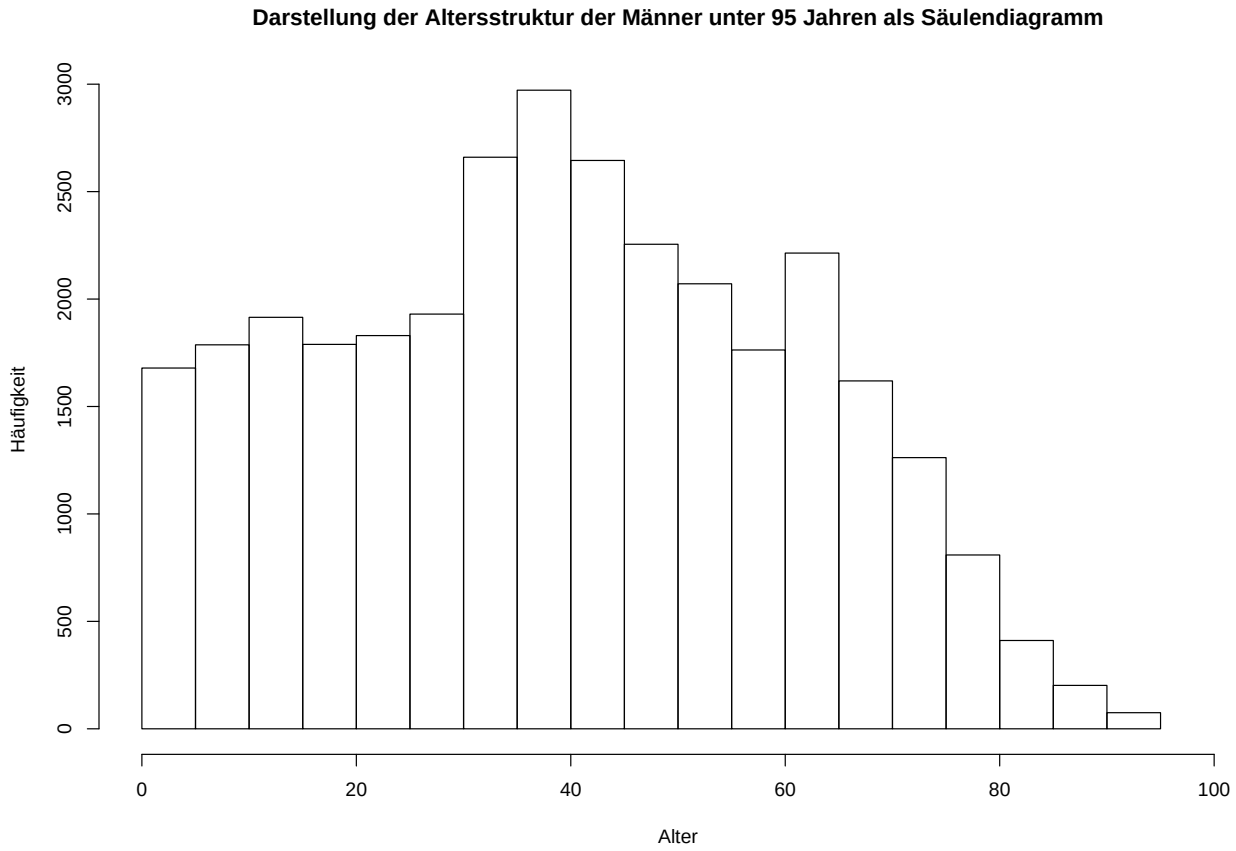
Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

In Beispiel 3 oben (Alter der ca. 32 Millionen männlichen Einwohner unter 95 Jahren im Jahr 2001, Quelle: Statistisches Bundesamt):

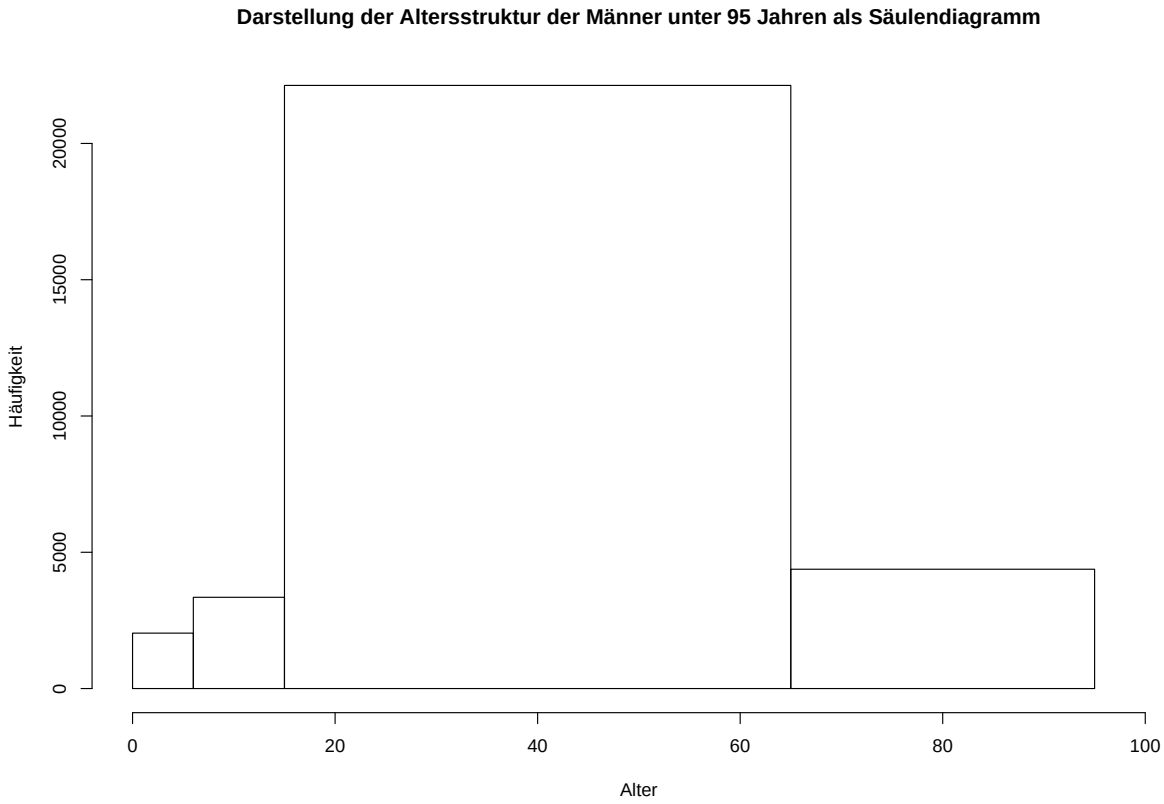
Unterteilung in 19 Klassen ergibt

Alter	Anzahl (in Tausenden)
$[0, 5)$	1679.3
$[5, 10)$	1787.2
$[10, 15)$	1913.2
$[15, 20)$	1788.7
$\vdots$	$\vdots$
$[65, 70)$	1618.4
$[70, 75)$	1262.2
$[75, 80)$	808.4
$[80, 85)$	411.9
$[85, 90)$	202.4
$[90, 95)$	73.9

Graphische Darstellung als Säulendiagramm:



Irreführend, falls die Klassen nicht alle gleich lang sind und die Klassenbreiten mit dargestellt werden:



Histogramm:

Im Gegensatz zum Säulendiagramm wird hier auch die Breite der Klassen mit berücksichtigt.

Vorgehen:

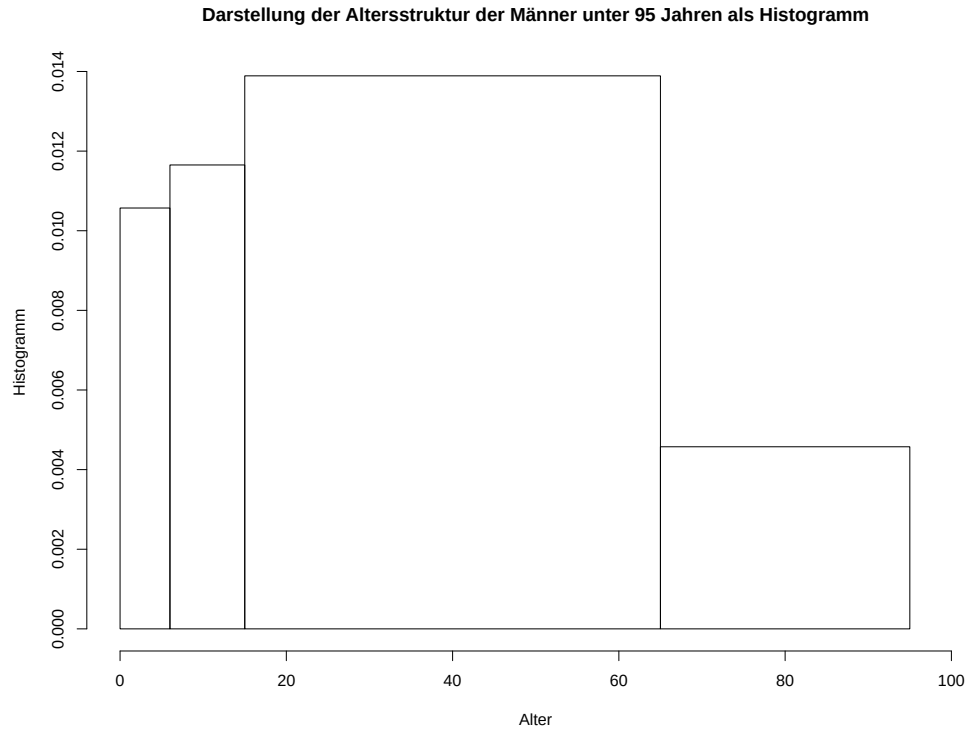
- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.
- Trage über I_j den Wert

$$\frac{n_j}{n \cdot \lambda(I_j)}$$

  auf, wobei $\lambda(I_j) =$ Länge von I_j.

Bemerkung: Flächeninhalt eines Rechtecks ist gleich dem prozentualen Anteil der Datenpunkte im zugrunde liegenden Intervall.

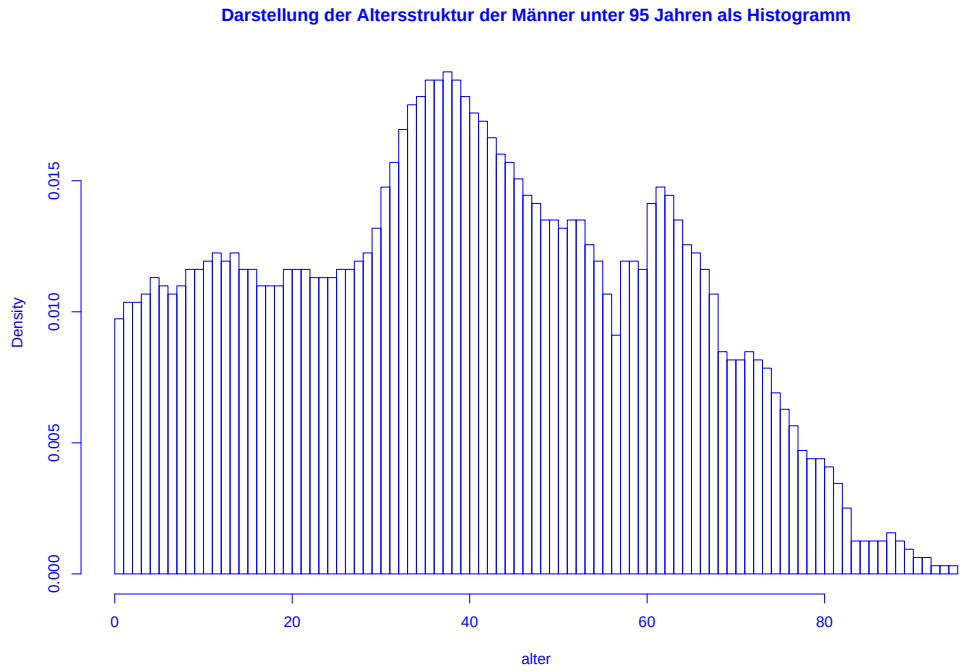
In Beispiel 3 oben erhält man



3.2 Dichteschätzung

Nachteil des Histogramms:

Unstetigkeit erschwert Interpretation zugrunde liegender Strukturen.



Ausweg:

Beschreibe Lage der Daten durch "glatte" Funktion.

Wie bisher soll gelten:

- Funktionswerte nichtnegativ.
- Flächeninhalt Eins.
- Fläche über Intervall ungefähr proportional zur Anzahl Datenpunkte in dem Intervall.

Definition: *Eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *mit*

$$f(x) \geq 0 \quad \textit{für alle } x \in \mathbb{R}$$

und

$$\int_{\mathbb{R}} f(x)\, dx = 1$$

heißt Dichte.

Ziel: Beschreibe Lage der Daten durch glatte Dichtefunktion.

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

$$\begin{aligned} f_h(x) &= \frac{\frac{1}{n} \cdot \text{Anzahl Datenpunkte } x_i \text{ in } [x-h, x+h]}{2h} \\ &= \frac{1}{n \cdot h} \sum_{i=1}^{n} \frac{1}{2} \cdot 1_{[x-h,x+h]}(x_i). \end{aligned}$$

Mit

$$1_{[x-h,x+h]}(x_i) = 1 \Leftrightarrow x - h \leq x_i \leq x + h \Leftrightarrow -1 \leq \frac{x - x_i}{h} \leq 1$$

erhält man

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

mit Dichte

$$K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u).$$

Deutung: Mittelung von Dichtefunktionen, die um die einzelnen Datenpunkte konzentriert sind.

2. Schritt: **Verallgemeinerung.**

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

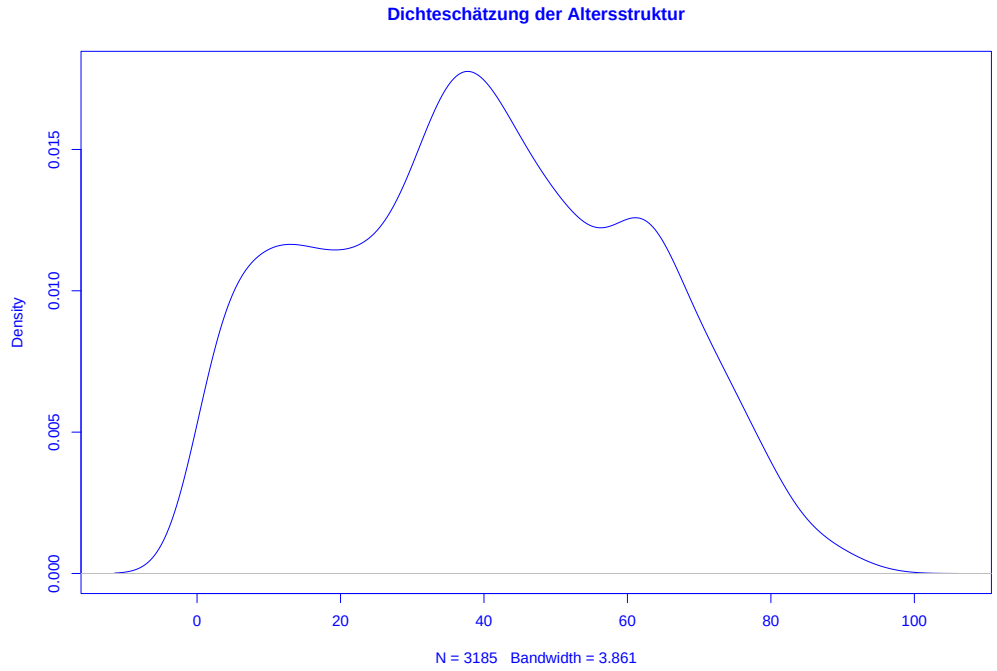
mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

Z.B. Epanechnikov-Kern:

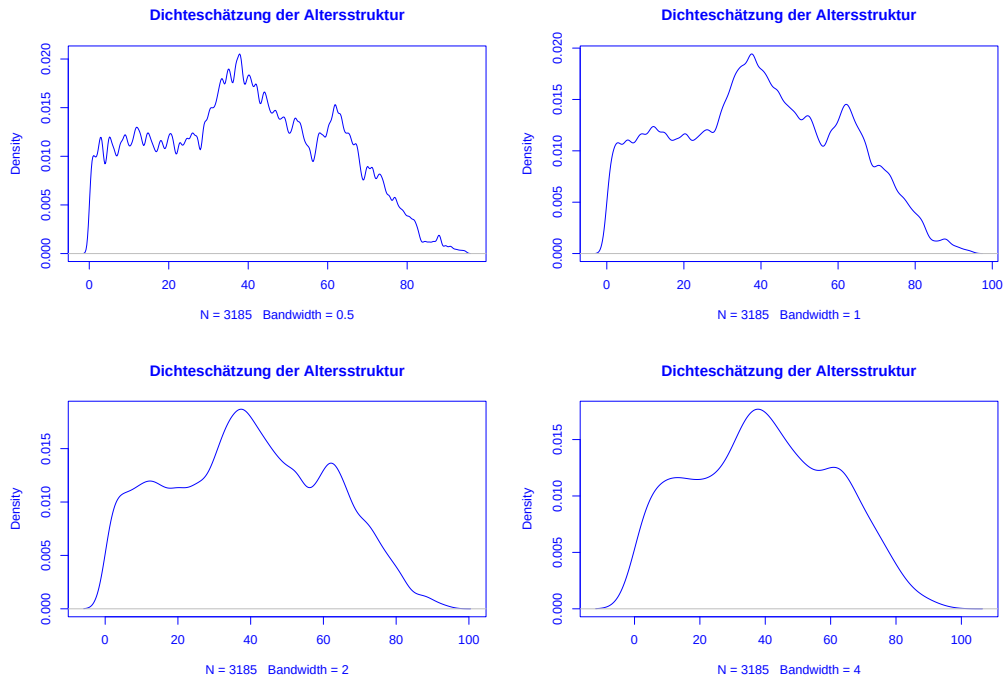
$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}$$

oder Gauss-Kern: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$.

In Beispiel 3 (Altersverteilung der männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD (ohne Berlin-West) im Jahr 2001) erhält man als Schätzung der Dichte:



Mittels h lässt sich die "Glattheit" des Kern-Dichteschätzers $f_h(x)$ kontrollieren:



Ist h sehr klein, so wird $f_h(x)$ als Funktion von x sehr stark schwanken, ist dagegen h groß, so variiert $f_h(x)$ als Funktion von x kaum noch.

3.3 Statistische Maßzahlen

Zwei Arten von statistischen Maßzahlen:

Lagemaßzahlen:

In welchem Bereich der Zahlengeraden liegen die Werte (oder die "Mitte" der Werte) ?

Streuungsmaßzahlen:

Wie groß ist der "Bereich", über den sich die Werte im wesentlichen erstrecken ?

Im Folgenden sei

$$x_1, \ldots, x_n$$

die Messreihe. Die der Größe nach aufsteigend sortierten Werte seien

$$x_{(1)}, \ldots, x_{(n)}.$$

In Beispiel 1 oben: Beschäftigungsquoten der Männer im Jahr 2006:

$x_1, \ldots, x_{26}$:
67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2, 66.4, 63.9,
73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

$x_{(1)}, \ldots, x_{(26)}$:
60.2, 63.3, 63.9, 65.2, 66.4, 66.9, 67.0, 68.2, 68.5, 70.8, 71.1, 71.3, 71.7, 72.5,
73.6, 73.8, 74.0, 74.6, 75.5, 76.0, 77.0, 77.0, 77.3, 79.6, 80.6, 80.8

Beispiele für Lageparameter:

(empirisches arithmetisches) Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n)$$

Bei den Beschäftigungsquoten für Männer: $\bar{x} = 71.8$

(Wert bei den Frauen: $\bar{x} = 58.2$)

Problematisch bei nicht reellen Messgrößen oder falls Ausreißer in Stichprobe vorhanden.

In diesen Fällen besser geeignet:

(empirischer) Median:

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls} \quad n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls} \quad n \text{ gerade.} \end{cases}$$

Bei den Beschäftigungsquoten für Männer: $\tilde{x} = 72.10$

(Wert bei den Frauen: $\tilde{x} = 59.3$)

Beispiele für Streuungsparameter:

(empirische) Spannweite oder Variationsbreite:

$$r := x_{max} - x_{min} := x_{(n)} - x_{(1)}.$$

Bei den Beschäftigungsquoten für Männer: $r = 80.8 - 60.2 = 20.6$

(Wert bei den Frauen: $r = 73.2 - 34.6 = 29.6$)

(empirische) Varianz:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right).$$

Bei den Beschäftigungsquoten für Männer: $s^2 \approx 30.8$

(Wert bei den Frauen: $s^2 \approx 75.3$)

(empirische) Standardabweichung oder Streuung:

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

Bei den Beschäftigungsquoten für Männer: $s \approx 5.55$

(Wert bei den Frauen: $s \approx 8.68$)

Variationskoeffizient:

$$V = \frac{s}{\bar{x}}$$

Bei den Beschäftigungsquoten für Männer: $V \approx 0.077$

(Wert bei den Frauen: $V \approx 0.149$)

Bei nicht reellen Messgrößen oder Vorhandensein von Ausreißern ist der sogenannte Interquartilabstand

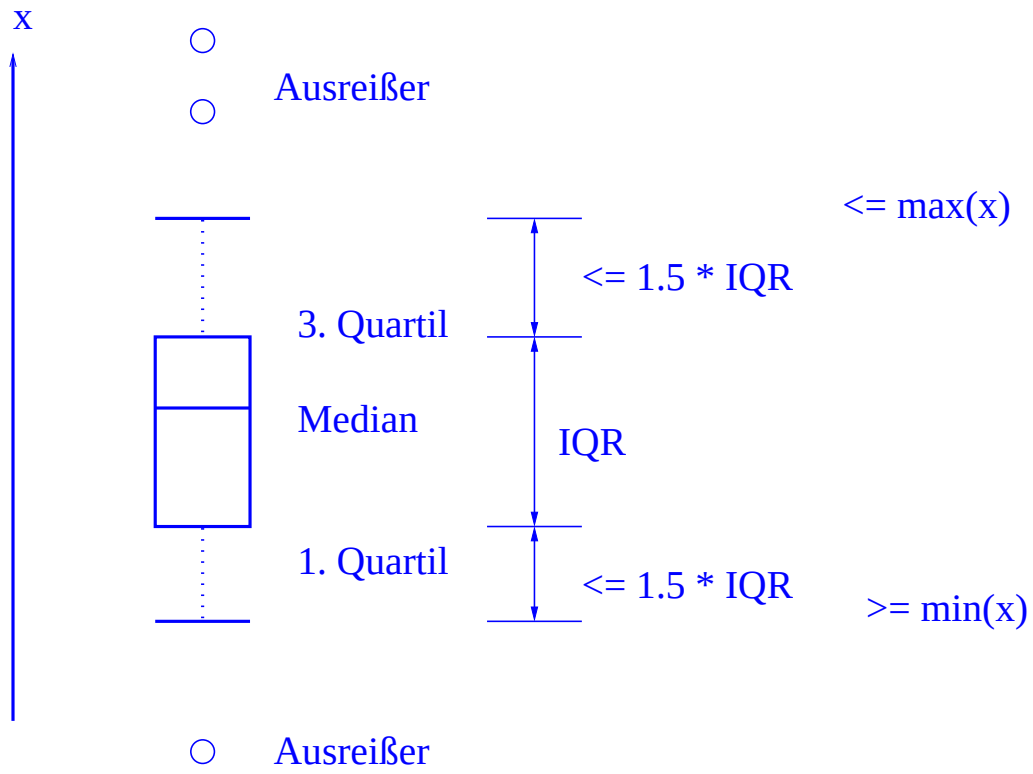
$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

günstiger.

Bei den Beschäftigungsquoten für Männer:$IQR = 76 - 67 = 9$

(Wert bei den Frauen: $IQR = 63.3 - 53.2 = 10.1$)

Graphische Darstellung einiger dieser Lage- und Streuungsparameter im sogenannten Boxplot:



Boxplot zum Vergleich der Beschäftigungsquoten von Männern und Frauen:

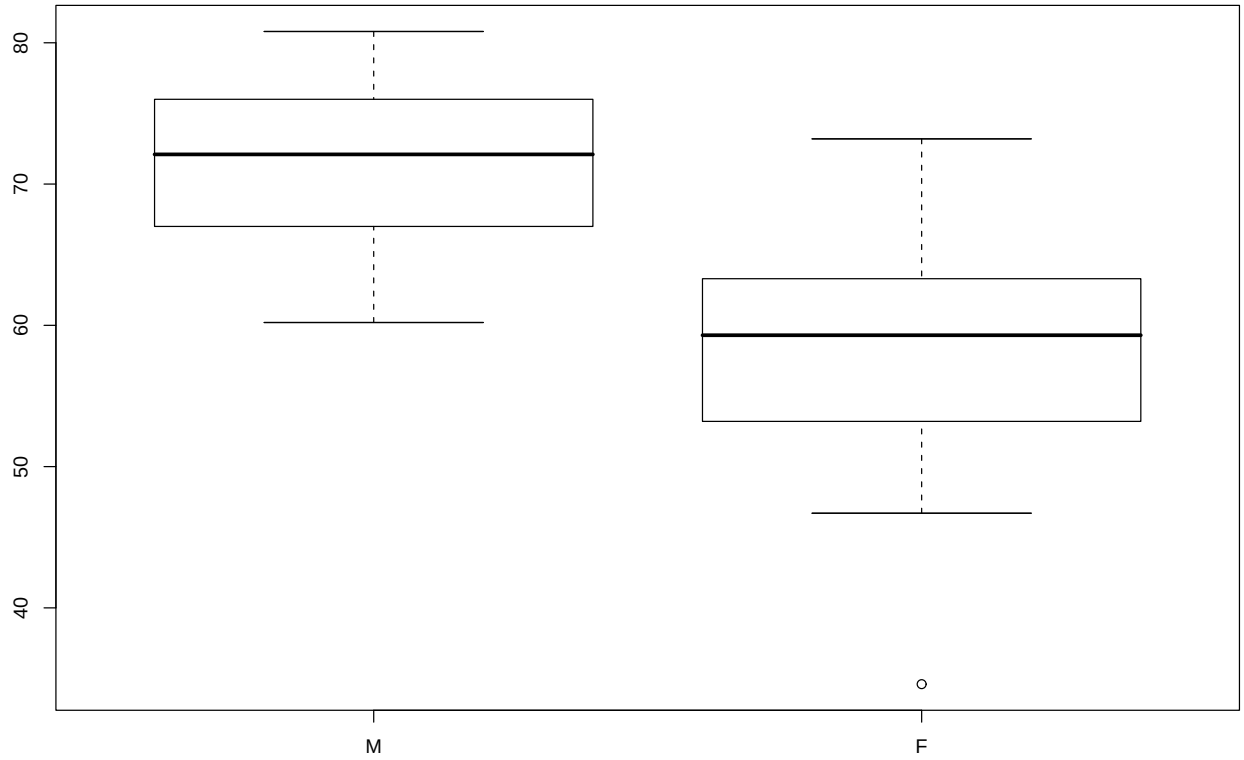
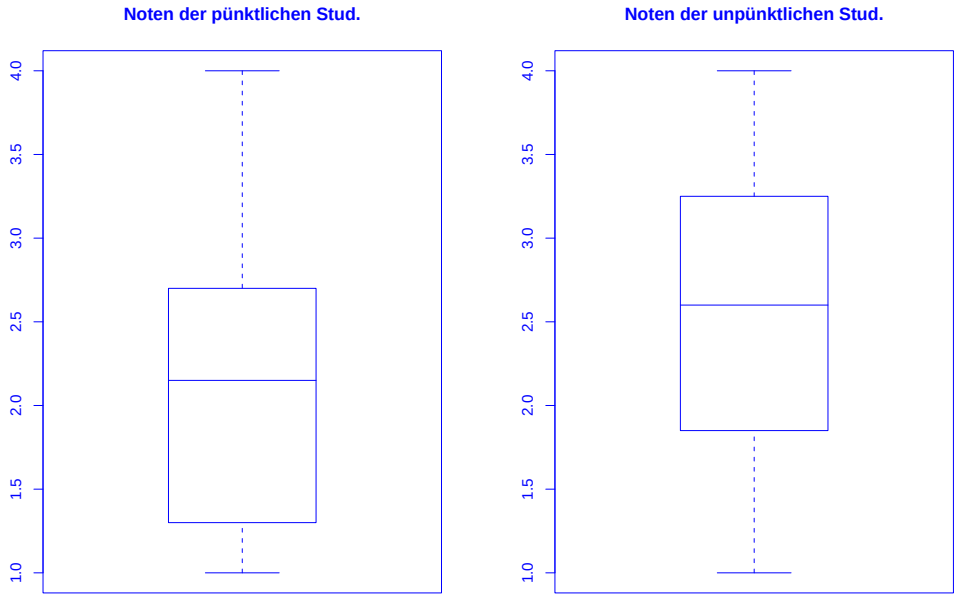
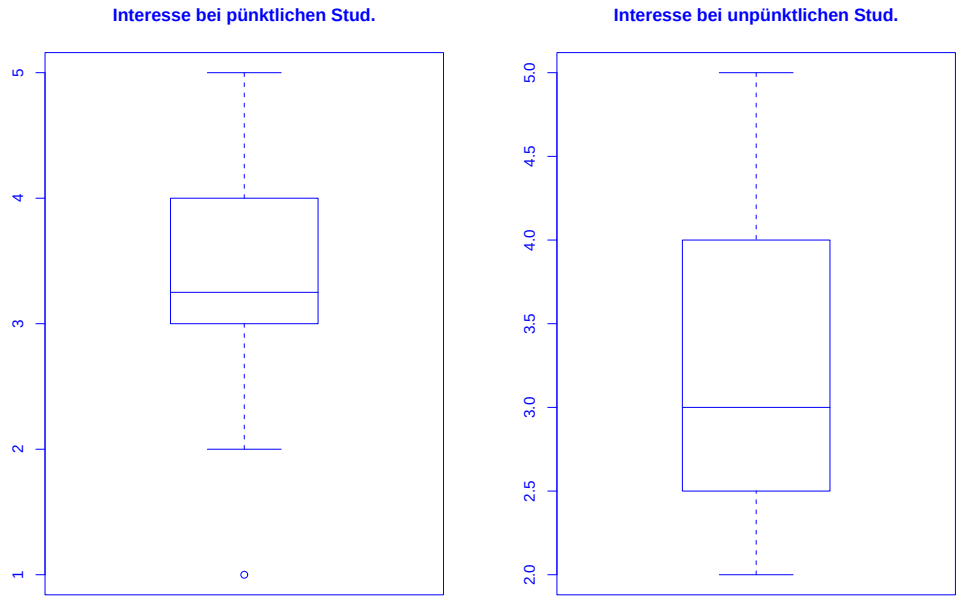


Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:




Interesse bei pünktlichen Stud.
1
2
3
4
5
Interesse bei unpünktlichen Stud.
2.0
2.5
3.0
3.5
4.0
4.5
5.0

3.4 Regressionsrechnung

Geg.: 2–dimensionale Messreihe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

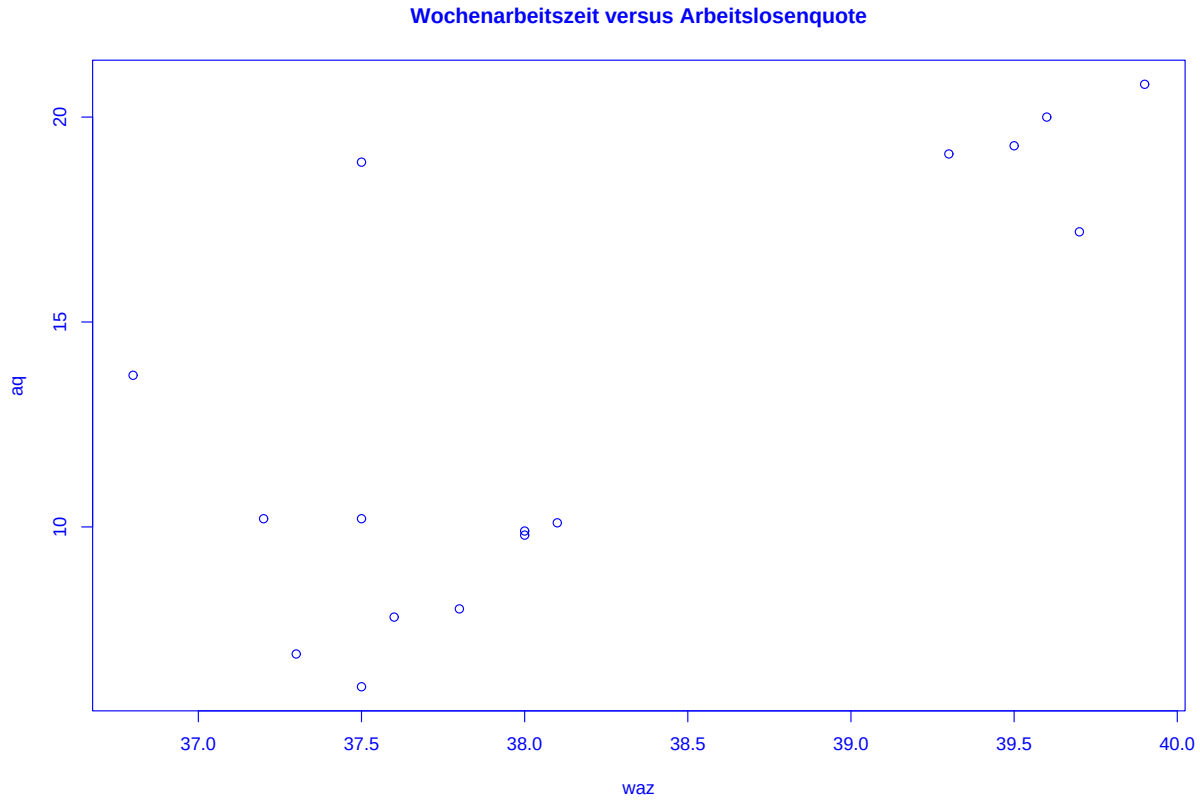
vom Umfang n.

Frage: Zusammenhang zwischen den x– und den y–Koordinaten ?

Beispiel: Besteht ein Zusammenhang zwischen

- der Wochenarbeitszeit im produzierenden Gewerbe und der Arbeitslosenquote in den 16 Bundesländern der BRD im Jahr 2002 ?

Darstellung der Messreihe (Quelle: Statistisches Bundesamt) im Scatterplot (Streudiagramm):



Bei der sogenannten **linearen Regression** passt man eine Gerade

$$y = \mathbf{a} \cdot x + \mathbf{b}$$

an die Daten an.

Eine Möglichkeit dafür:

Wähle **a, b** $\in \mathbb{R}$ durch Minimierung von

$$\sum_{i=1}^{n} \left(y_i - (a \cdot x_i + b)\right)^2 .$$

Beispiel: Es sei $n = 3$ und

$$(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (-2, 1).$$

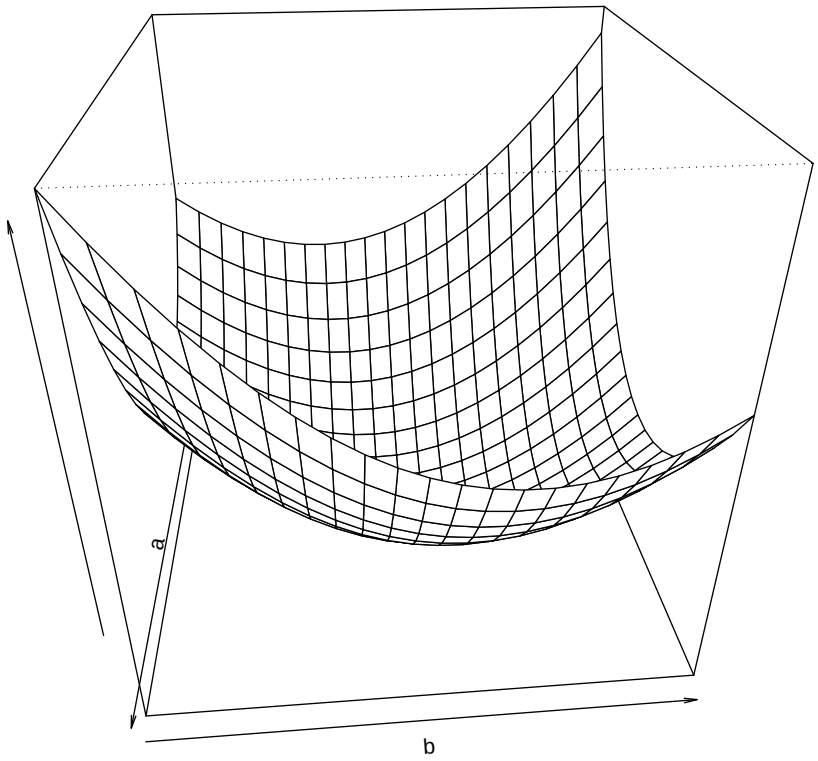
Gesucht ist dann eine Gerade

$$y = a \cdot x + b,$$

für die der folgende Ausdruck möglichst klein ist:

$$\begin{aligned}
&(y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + (y_3 - (a \cdot x_3 + b))^2 \\
&= (0 - (a \cdot 0 + b))^2 + (0 - (a \cdot 1 + b))^2 + (1 - (a \cdot (-2) + b))^2 \\
&= b^2 + (a + b)^2 + (1 + 2 \cdot a - b)^2.
\end{aligned}$$

In Abhängigkeit von a und b lässt sich der zu minimierende Ausdruck graphisch wie folgt darstellen:



Man kann zeigen: Der Ausdruck

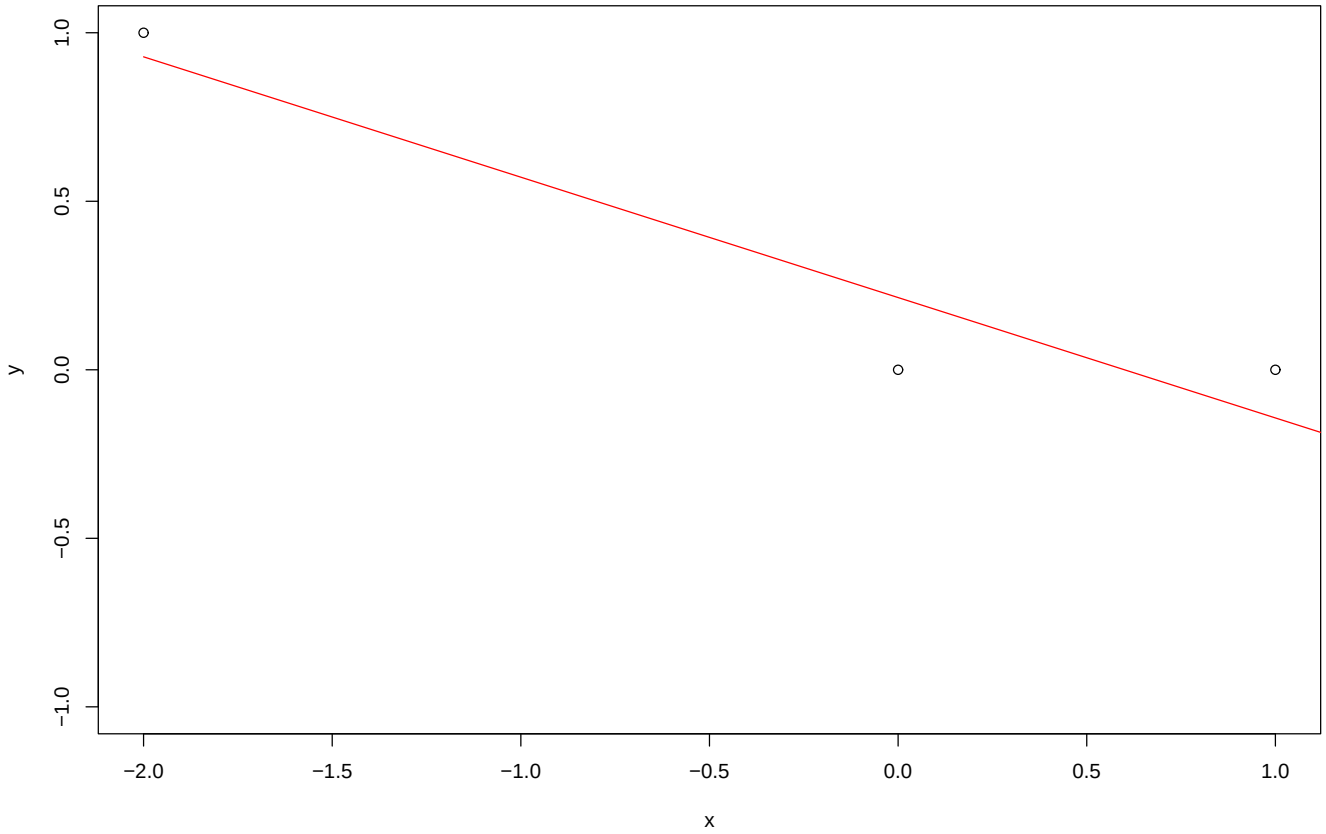
$$b^2 + (a+b)^2 + (1 + 2 \cdot a - b)^2$$

wird minimal für

$$a = -\frac{5}{14} \quad \text{und} \quad b = \frac{3}{14}.$$

Also ist die gesuchte Gerade hier gegeben durch

$$y = -\frac{5}{14} \cdot x + \frac{3}{14}.$$


1.0
0.5
0.0
−0.5
−1.0
y
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
x

Allgemein führt obige Minimierungsaufgabe auf die sogenannte **Regressionsgerade** gegeben durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$$

mit

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

und

$$\hat{a} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

$(\frac{0}{0} := 0)$.

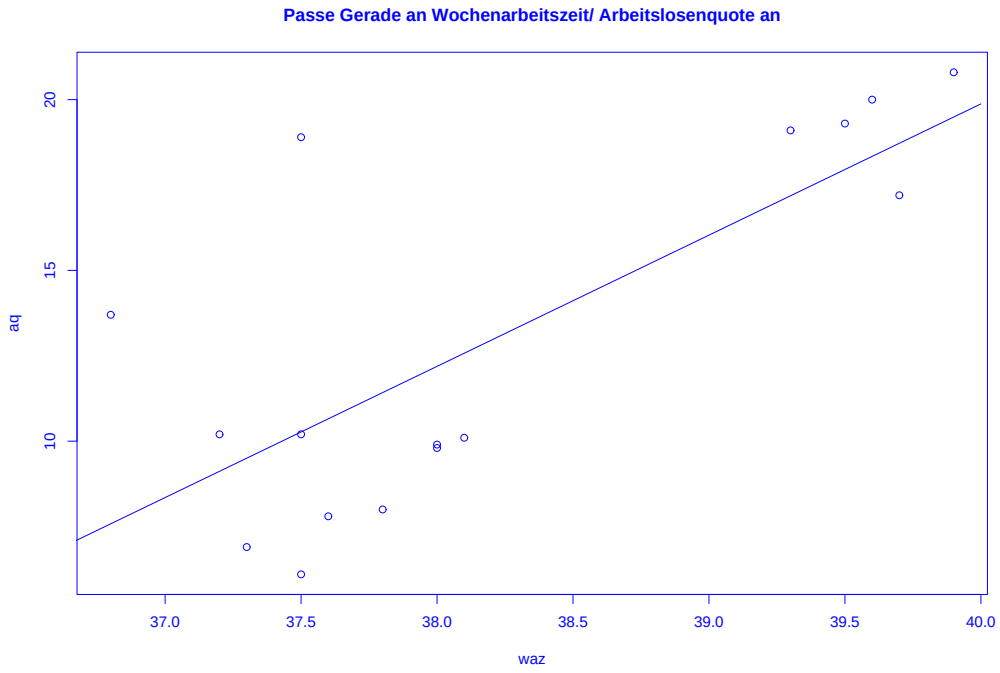
Hierbei wird

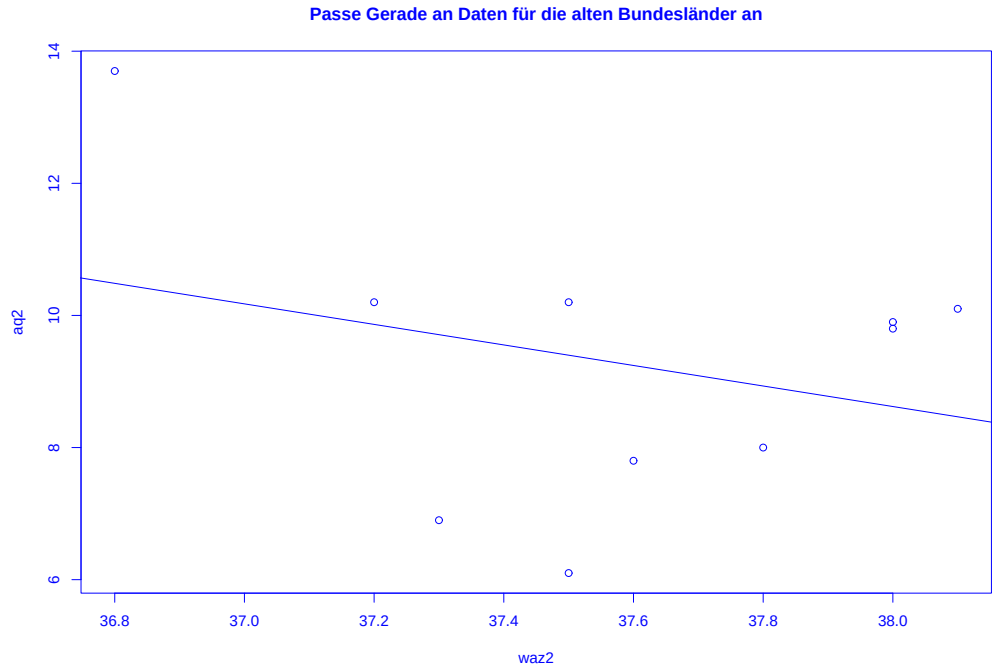
$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

als empirische Kovarianz der zweidimensionalen Messreihe bezeichnet.

Ist die empirische Kovarianz positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

Beispiel:




Passe Gerade an Daten für die alten Bundesländer an
aq2
14
12
10
8
6
36.8
37.0
37.2
37.4
37.6
37.8
38.0
waz2

Man kann weiter zeigen, dass die sogenannte empirische Korrelation

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

im Intervall $[-1, 1]$ liegt.

Die empirische Korrelation dient zur Beurteilung der Abhängigkeit der x– und der y–Koordinaten.

Sie macht Aussagen über die Regressionsgerade und die Lage der Punktwolke im Scatterplot:

- Ist die empirische Korrelation $+1$ oder -1, so liegen die Punkte (x_i, y_i) alle auf der Regressionsgeraden.

- Ist die empirische Korrelation positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

- Ist die empirische Korrelation Null, so verläuft die Regressionsgerade waagrecht.

3.5 Nichtparametrische Regressionsschätzung

Verallgemeinerung der linearen Regression:

Passe Funktionen allgemeinerer Bauart (z.B. Polynome) an Daten an. Z.B. wie bei linearer Regression durch Minimierung der Summe der quadratischen Fehler (Prinzip der Kleinsten-Quadrate).

Falls Bauart vorgegeben ist und diese nur von endlich vielen Parametern abhängt: **parametrische Regressionsschätzung**.

Anderer Ansatz:

Nichtparametrische Regressionsschätzung.

Keine Annahme über die Bauart der anzupassenden Funktion.

Einfachstes Beispiel: lokale Mittelung

Versucht wird, den durchschnittlichen Verlauf der y-Koordinaten der Datenpunkte in Abhängigkeit der zugehörigen x-Koordinaten zu beschreiben.

z.B. durch sogenannten **Kernschätzer**:

$$m_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \cdot y_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}.$$

Hierbei ist $K : \mathbb{R} \to \mathbb{R}_+$ die sogenannte **Kernfunktion** und $h > 0$ die sogenannte **Bandbreite**.

z.B. naiver Kern

$$K(u) = \frac{1}{2} 1_{[-1,1]}(u)$$

oder Gauss-Kern

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2).$$

Wie beim Kern-Dichteschätzer bestimmt die Bandbreite die Glattheit bzw. Rauheit der Schätzung.