Statistik I für Human- und SozialwissenschaftlerInnen

Vorlesung WS 2008/09

Prof. Dr. Michael Kohler
Fachbereich Mathematik
Technische Universität Darmstadt

kohler@mathematik.tu-darmstadt.de

"Those who ignore Statistics are condemned to reinvent it."

Brad Efron

Kapitel 1: Motivation

Statistik – 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.

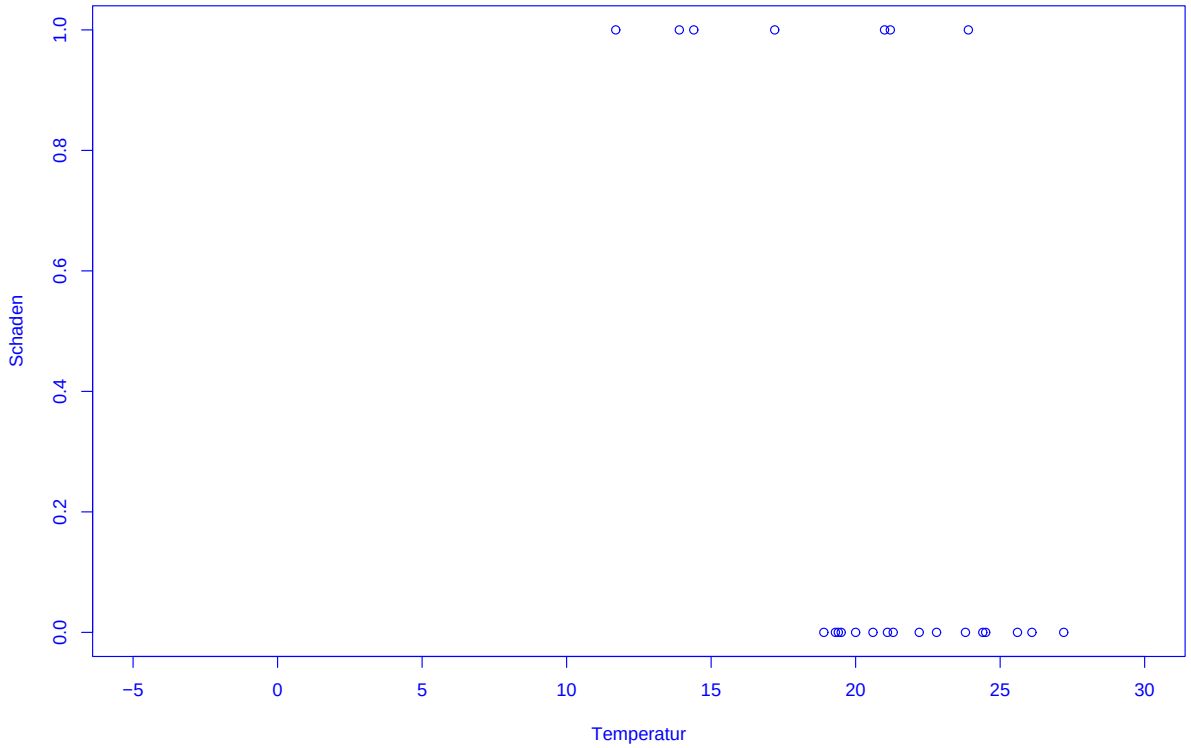
Zur Begründung verwendete Daten:

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

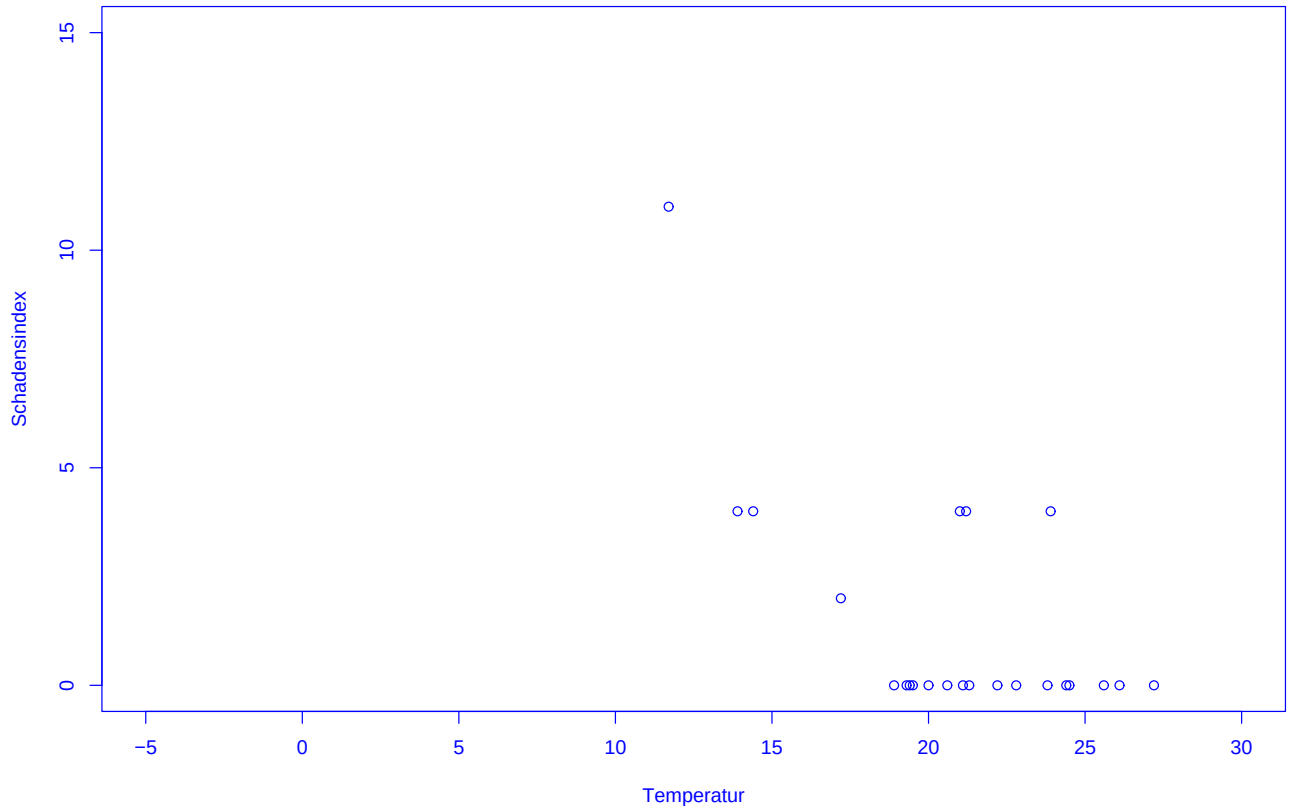
War für NASA leider nicht nachvollziehbar ...

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 Statistik ?

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.

Schön, aber:

Wozu braucht man Statistik in den **Human- und Sozialwissenschaften** ?

Um Theorien anhand von erhobenen Daten zu bilden bzw. zu überprüfen.

Z.B.:

- Wie entstehen Freundschaften - Ähnlichkeit oder Zufall ?

- Wie haben sich nach der Wiedervereinigung die Wohnverhältnisse im Osten verändert - z.B. hinsichtlich der Sozialstruktur in Plattenbausiedlungen ?

- Welches Bildungssystem ist besonders erfolgreich - und was folgt eigentlich aus der PISA-Studie ?

Schön, aber:

Braucht man den Stoff dieser Vorlesung wirklich im weiteren Studium der Psychologie, Soziologie oder Pädagogik in Darmstadt ?

JA, z.B.

- in der **Psychologie** als Grundlage der Vorlesung "Forschungsmethoden II" im 2. Semester sowie bei der selbständigen Durchführung empirischer Forschung.

- in der **Soziologie** als Grundlage der Vorlesung "Sozialwissenschaftliche Datenanalyse II" im 2. Semester sowie in allen empirischen Fächern.

- in der **Pädagogik** zur sicheren Interpretation empirischer Forschungsergebnisse.

FAZIT:
Statistik hat vielfältige Anwendungen in den Human- und Sozialwissenschaften und wird ihnen im Rahmen ihres Studiums immer wieder begegnen.

Die **Grundlagen** dazu lernen Sie in dieser Vorlesung.

Gliederung der Vorlesung (vorläufig):

- Kapitel 1: Einführung (heute)
- Kapitel 2: Erhebung von Daten im Rahmen von Studien und Umfragen (2V)
- Kapitel 3: Beschreibende Statistik (2V)
- Kapitel 4: Einführung in die W-Theorie (6V)
- Kapitel 5: Schließende Statistik (4V)

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 Statistik (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 W-Theorie und Statistik **lernen können**.

Spezielle “Tricks” dabei:

- Wiederholungsfolie zu Beginn
- Pause in der Mitte
- Umfrage am Schluss
- Intensiver Übungsbetrieb
- Skript

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

Zur Selbstkontrolle wird der Erwerb des Übungsscheines empfohlen.

TERMINE

1. Vorlesung: Montag, 16:25 Uhr - 17:55 Uhr, in S 206_030

2. Vortragsübungen: Dienstag, 8:00 Uhr - 9:40 Uhr, in S 103_226

   Die Vortragsübungen finden 14-täglich statt. Sie beginnen für

   - Pädagogen am **21.10.08**
   - Psychologen und Soziologen am **28.10.08**

3. Übungen:

   Die Übungen finden zu verschiedenen Terminen in Kleingruppen statt (Dauer 2 Stunden, wöchentlich).

Ergänzende Literatur:

Falls Sie sich über die Vorlesung hinaus in Statistik vertiefen möchten, empfehle ich die folgenden Bücher:

1. David Freedman, Robert Pisani, Roger Purves: *Statistics.* W. W. Norton & Company, New York, 1998.

Enthält viele sehr schöne Beispiele sowie keinerlei Mathematik, ca. 43 Euro.

2. L. Fahrmeir, R. Künstler, I. Pigeot und G. Tutz. *Statistik. Der Weg zur Datenanalyse.* Springer-Verlag, Berlin, 2001.

Anschauliche Erklärung des Stoffes unter weitgehender Vermeidung der mathematischen Hintergründe, deckt fast den gesamten Stoff der Vorlesung ab, ca. 30 Euro.

3. J. Bortz. *Statistik für Human- und Sozialwissenschaftler*. Springer, 2005.

Primär für Psychologen interessant, da sie dieses Buch im zweiten Semester verwenden werden (enthält aber auch Stoff aus dieser Vorlesung), ca. 50 Euro.

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.

Beispiele aus den Studienfächern der HörerInnen werden in den Übungen behandelt.

Bezug zum Studienfach:

- In der Psychologie führt man oft kontrollierte Studien durch, z.B.: Wie entstehen Freundschaften - Zufall oder Ähnlichkeit ?

- In der Soziologie analysiert man bei empirischen Arbeiten in der Regel Beobachtungsstudien oder Umfragen, z.B.: Wie wandeln sich die Werte bei Jugendlichen ?

- In der Pädagogik spielen Beobachtungsstudien und kontrollierte Studien eine wichtige Rolle, z.B.: PISA-Studie zum Vergleich der verschiedenen Schulformen.

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 Statistik-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.

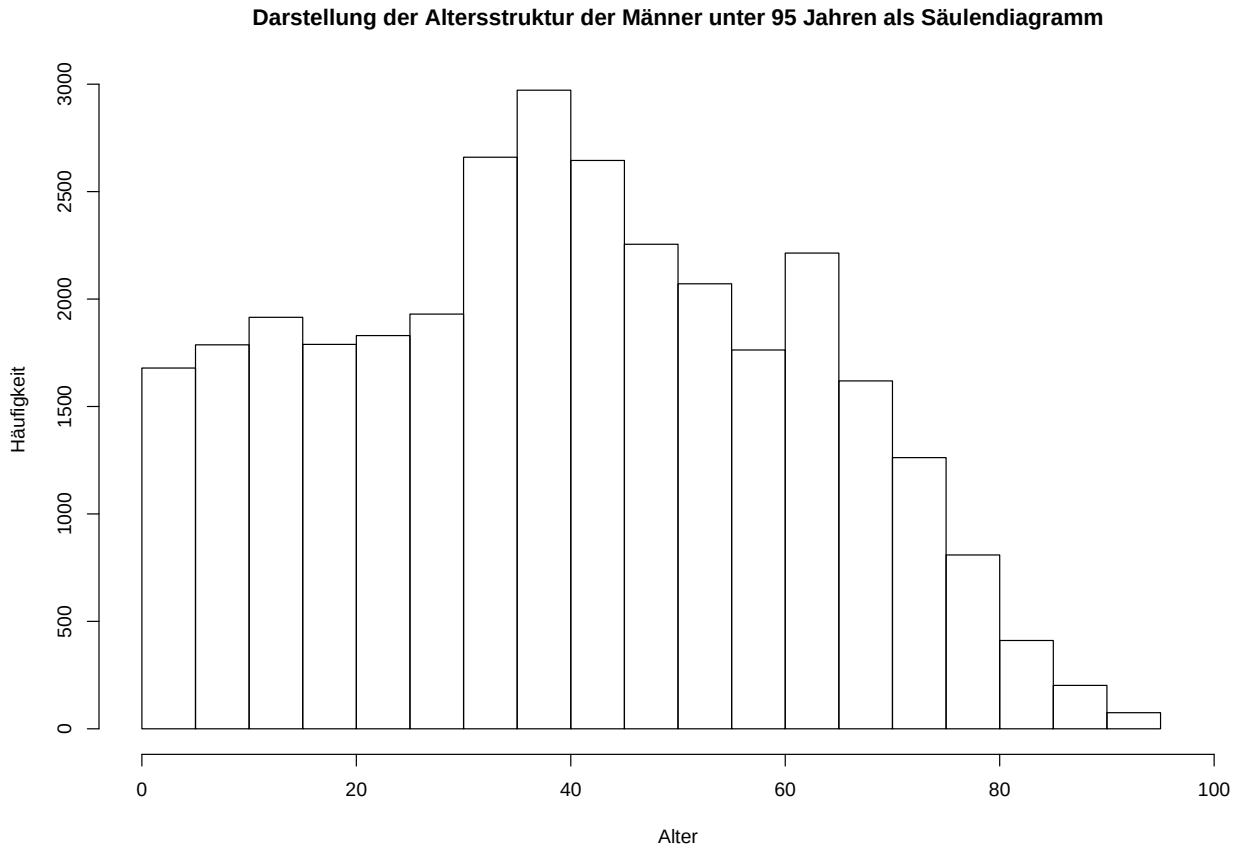
Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

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

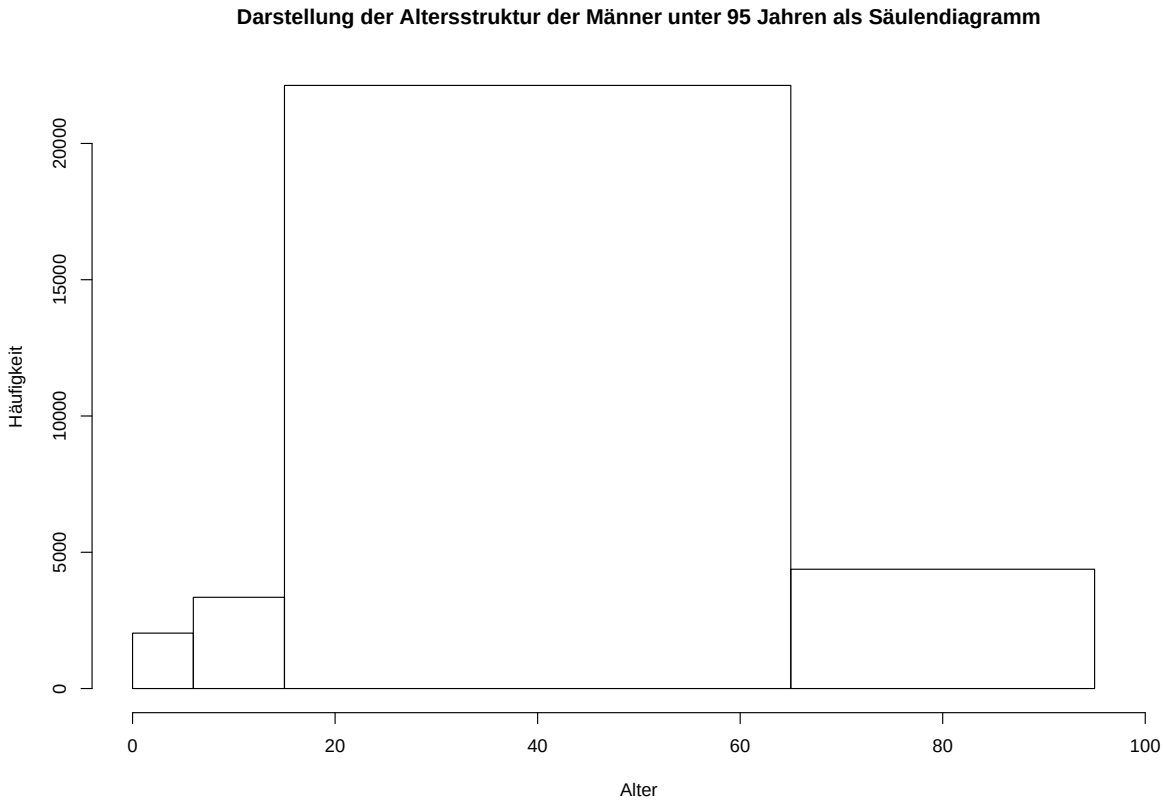
Unterteilung in 19 Klassen ergibt

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:

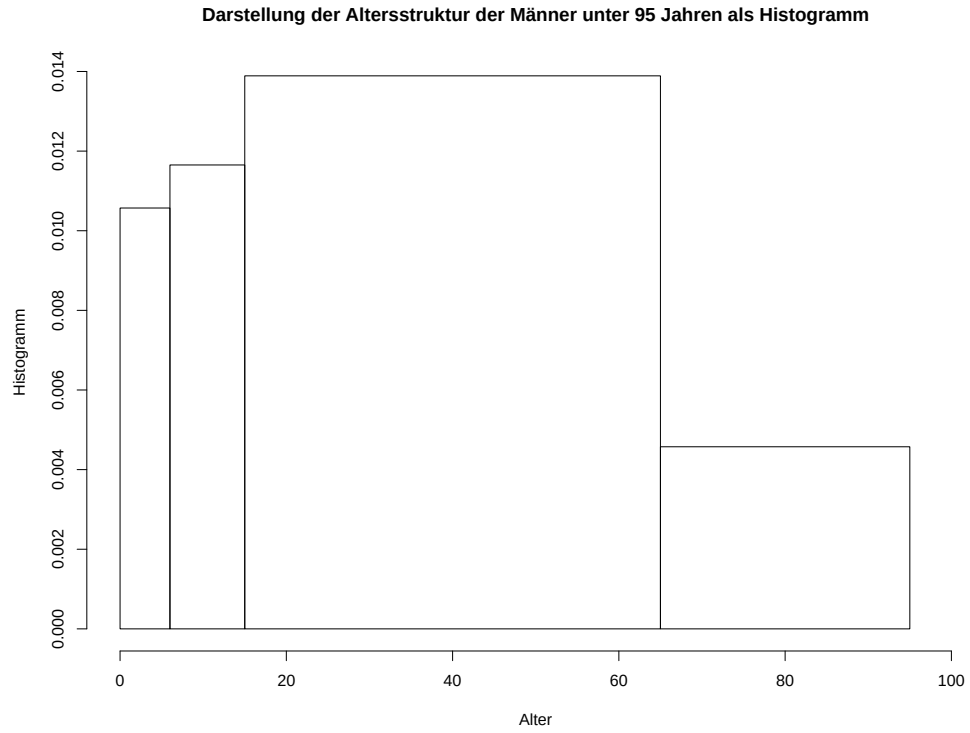
- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.
- Trage über I_j den Wert

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

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

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

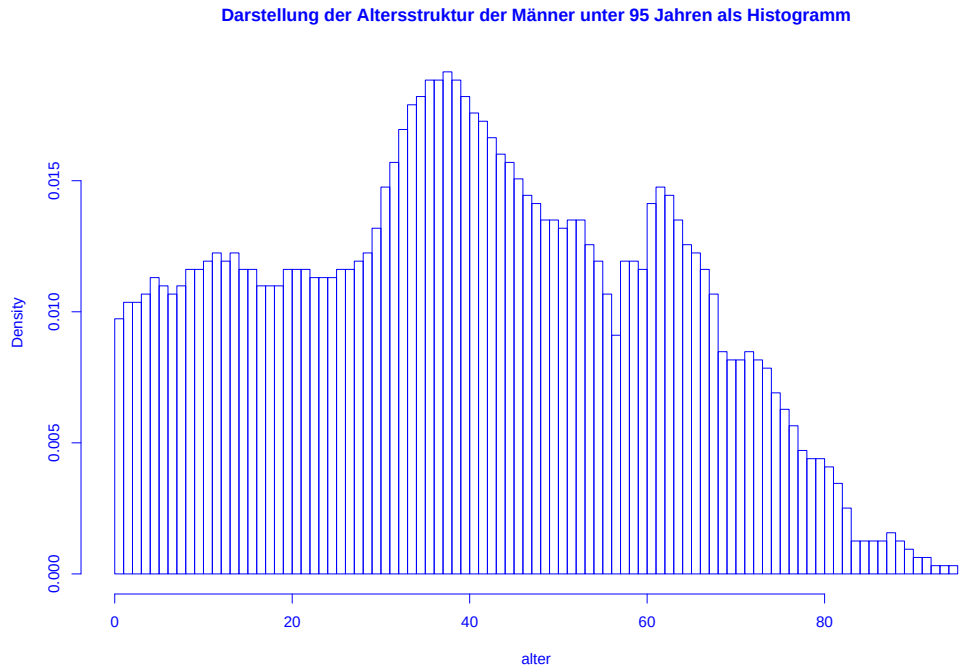
In Beispiel 3 oben erhält man



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.**

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

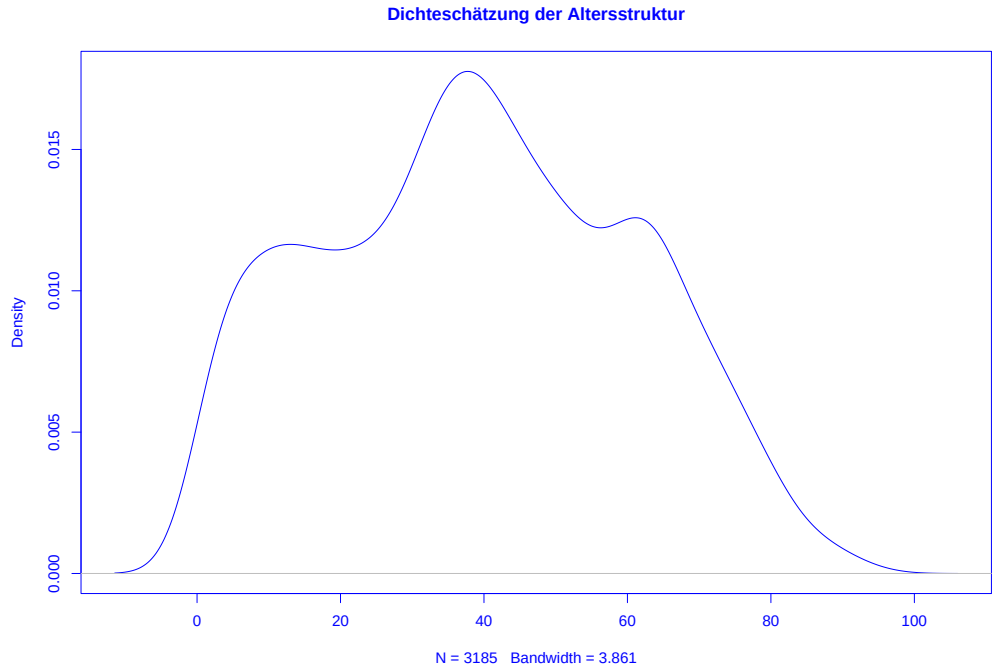
mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

Z.B. Epanechnikov-Kern:

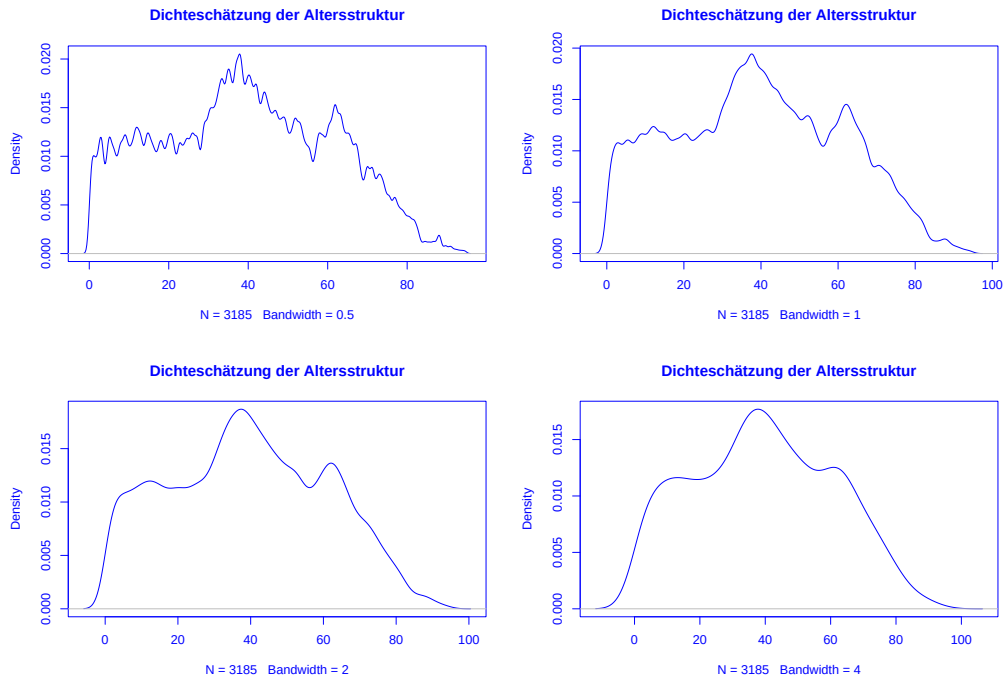
$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}$$

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:

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

Bei den Beschäftigungsquoten für Männer: $Md = 72.10$

(Wert bei den Frauen: $Md = 59.3$)

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

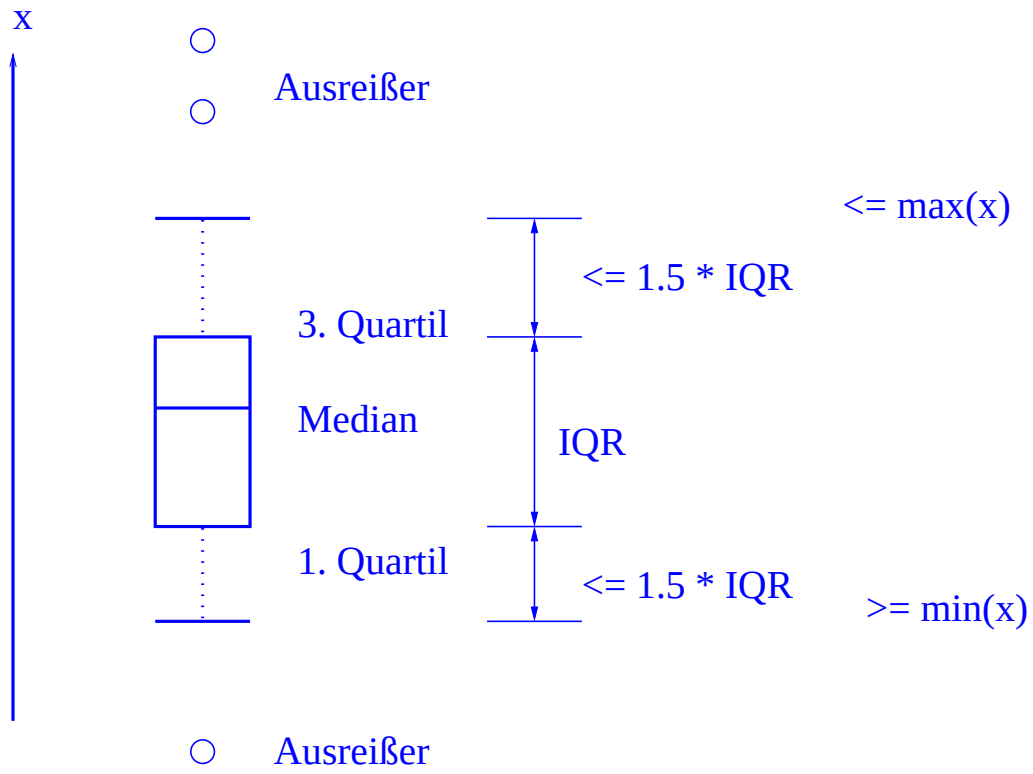
$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

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:

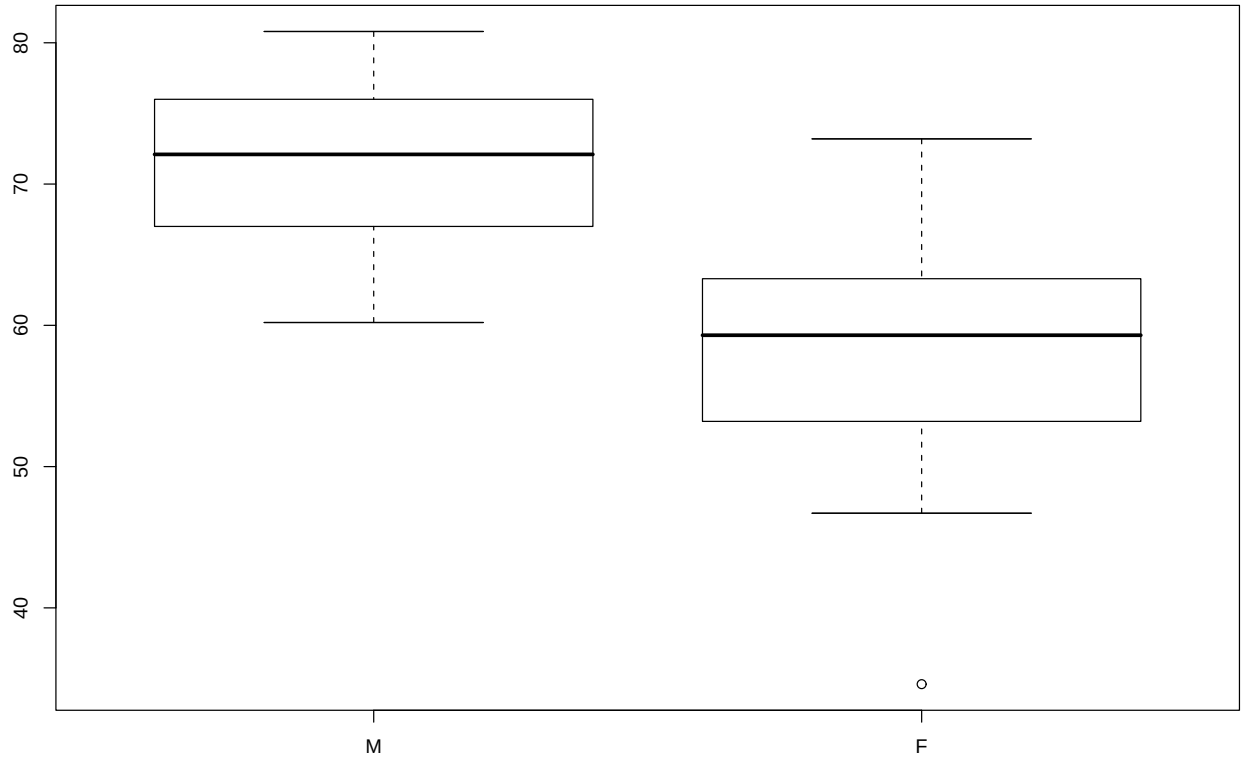
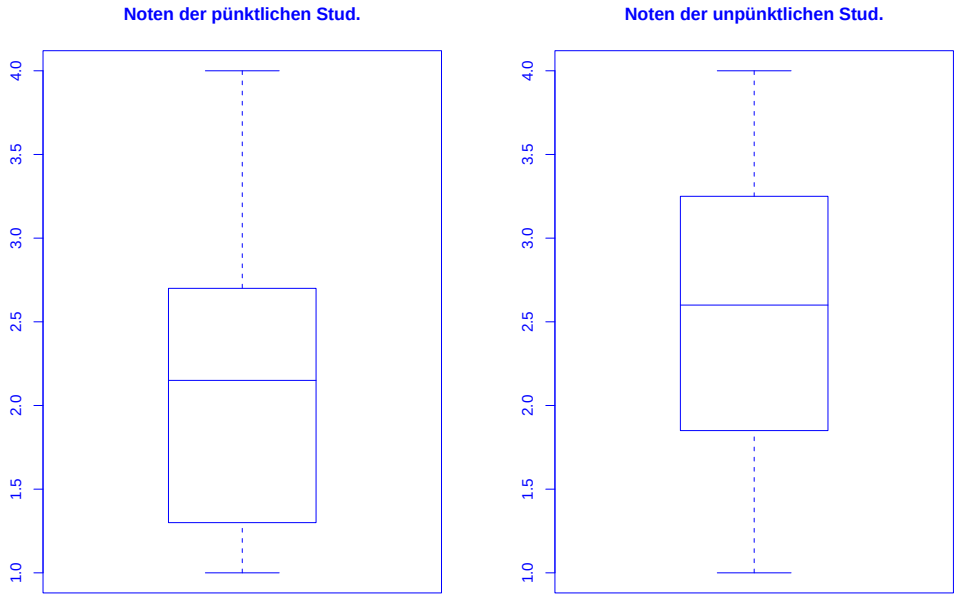
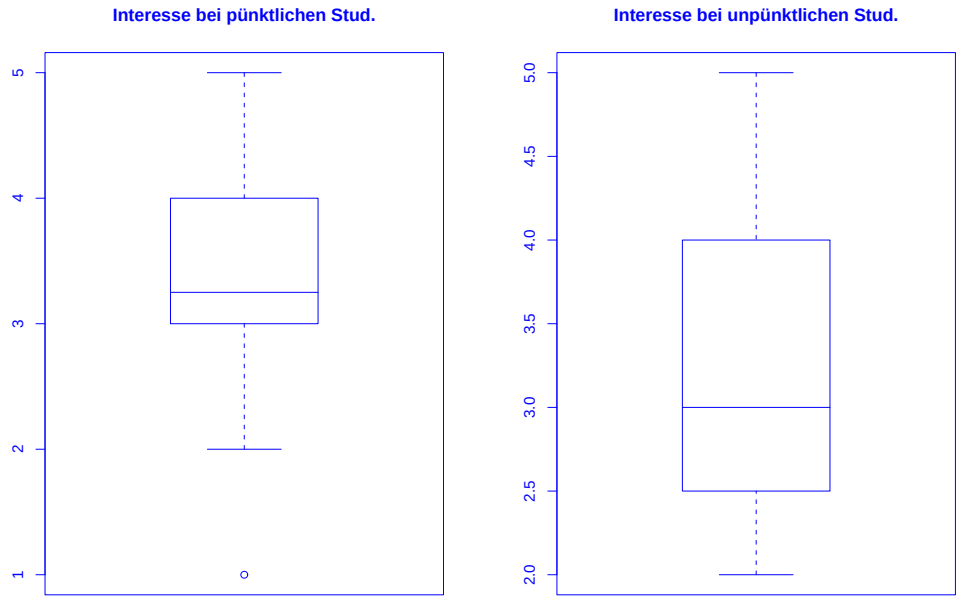


Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:




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

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

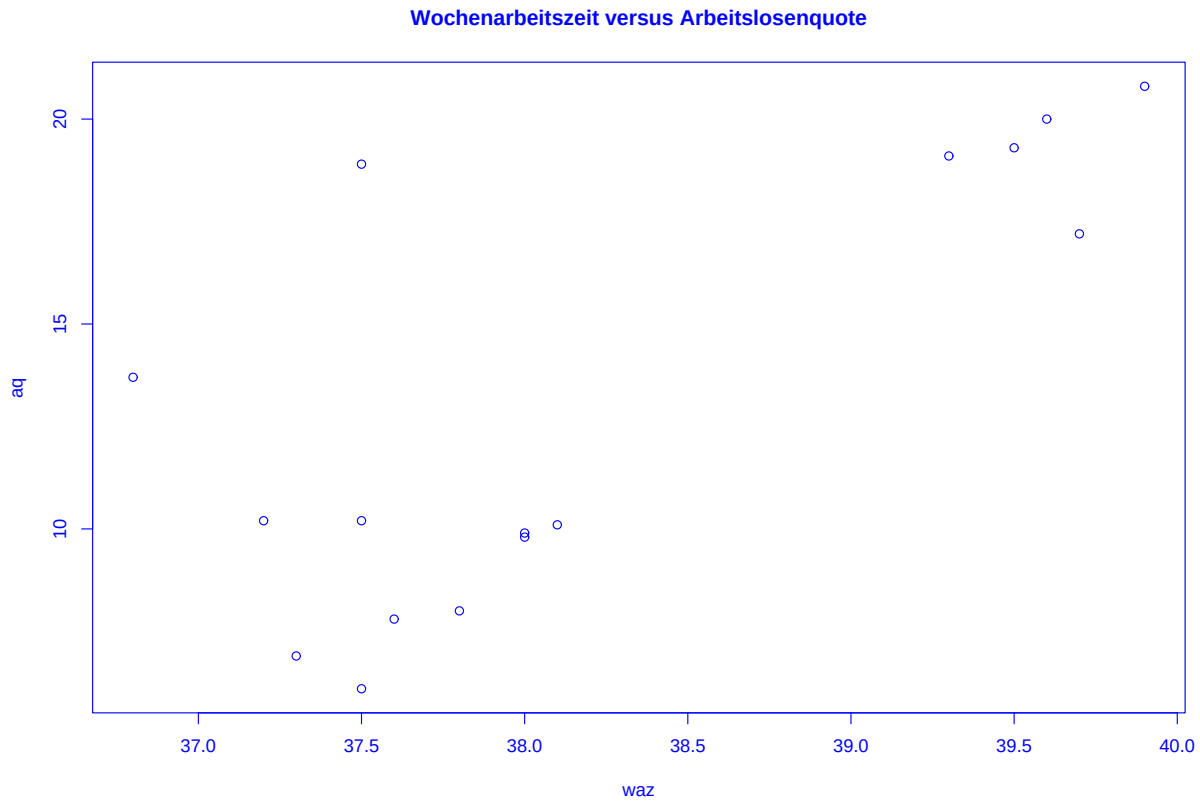
vom Umfang n.

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

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

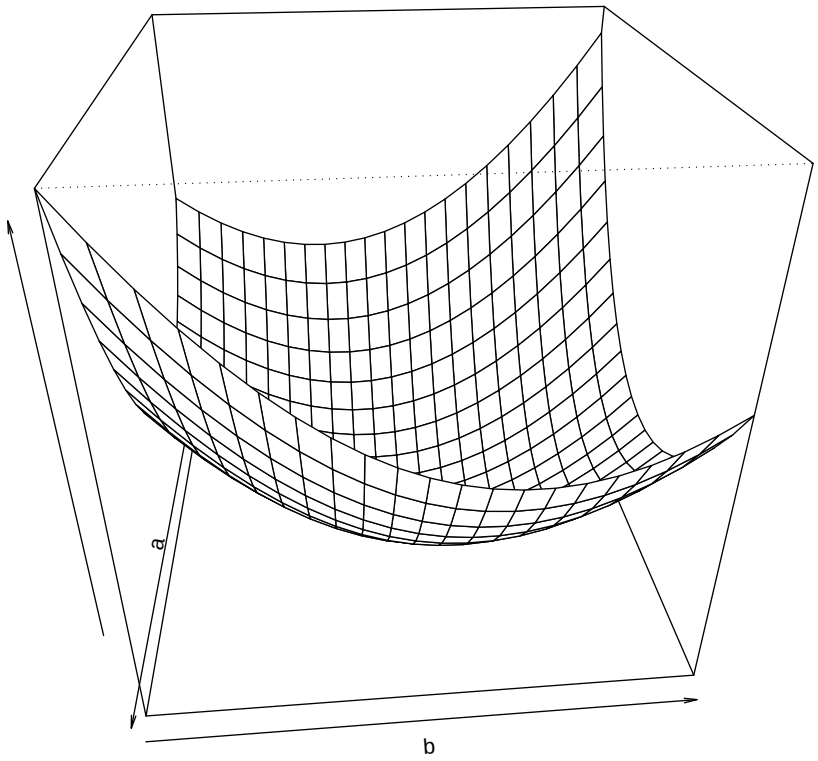
Gesucht ist dann eine Gerade

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

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

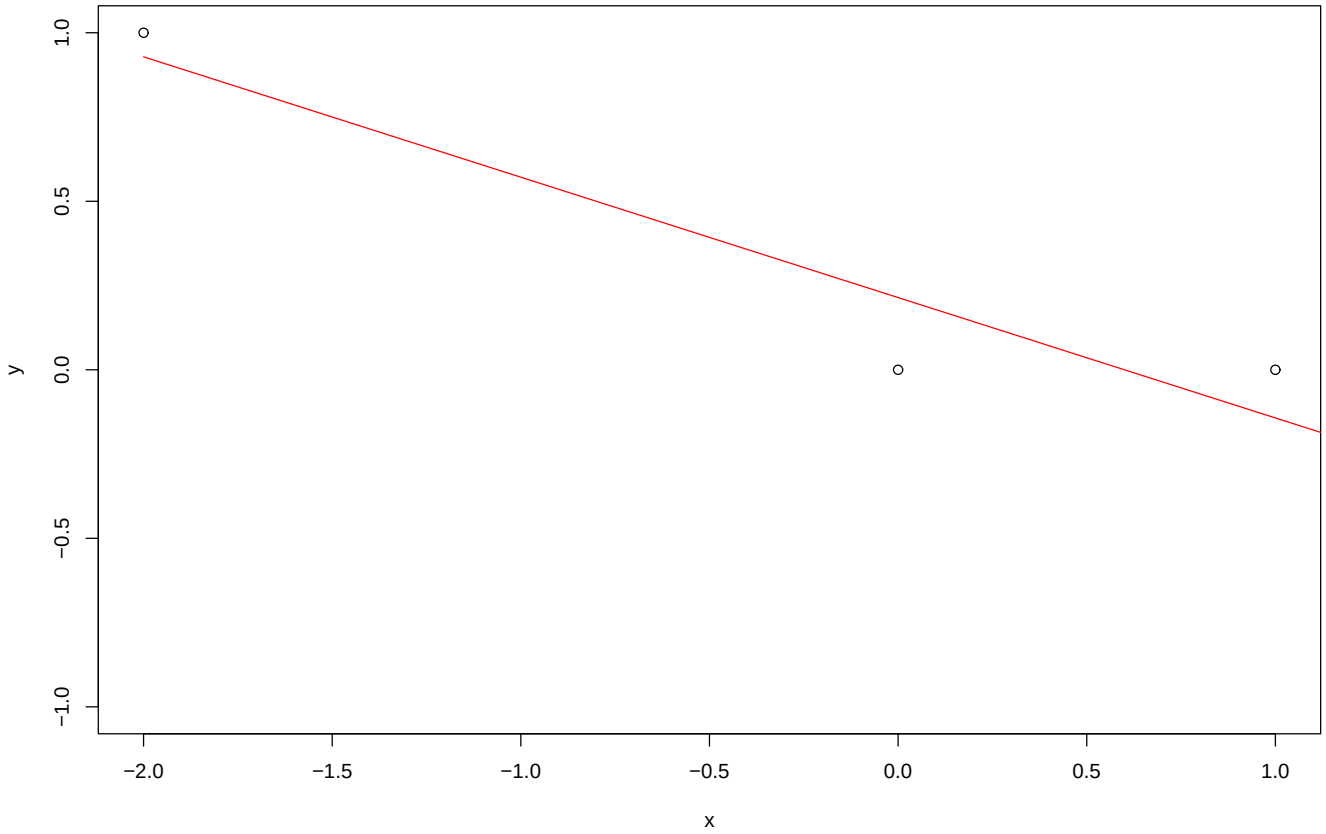
$$b^2 + (a+b)^2 + (1 + 2 \cdot a - b)^2$$

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

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

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

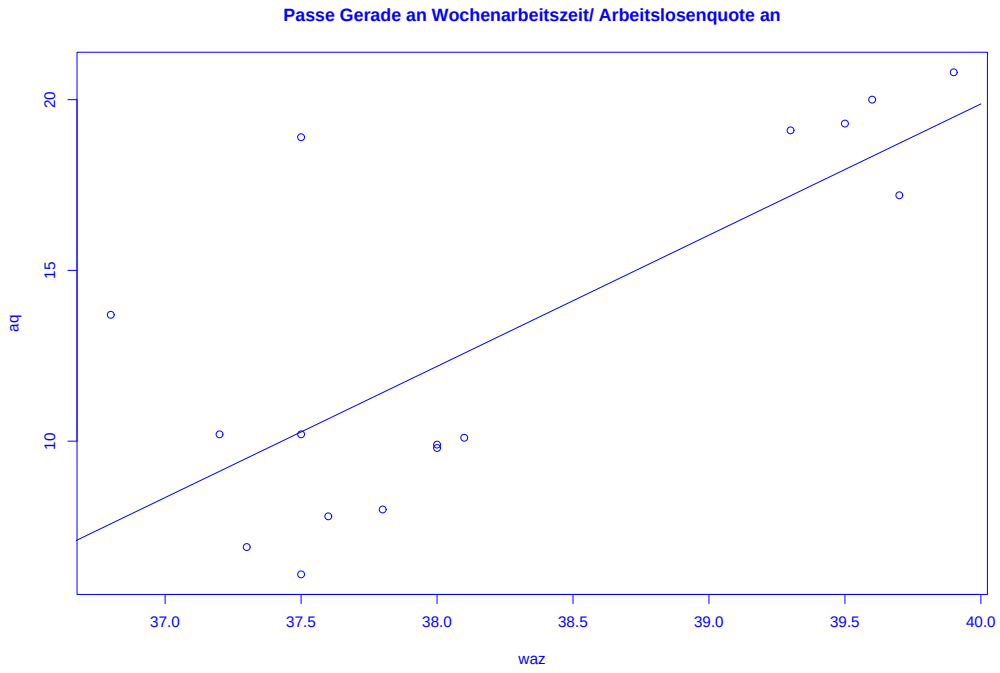
Hierbei wird

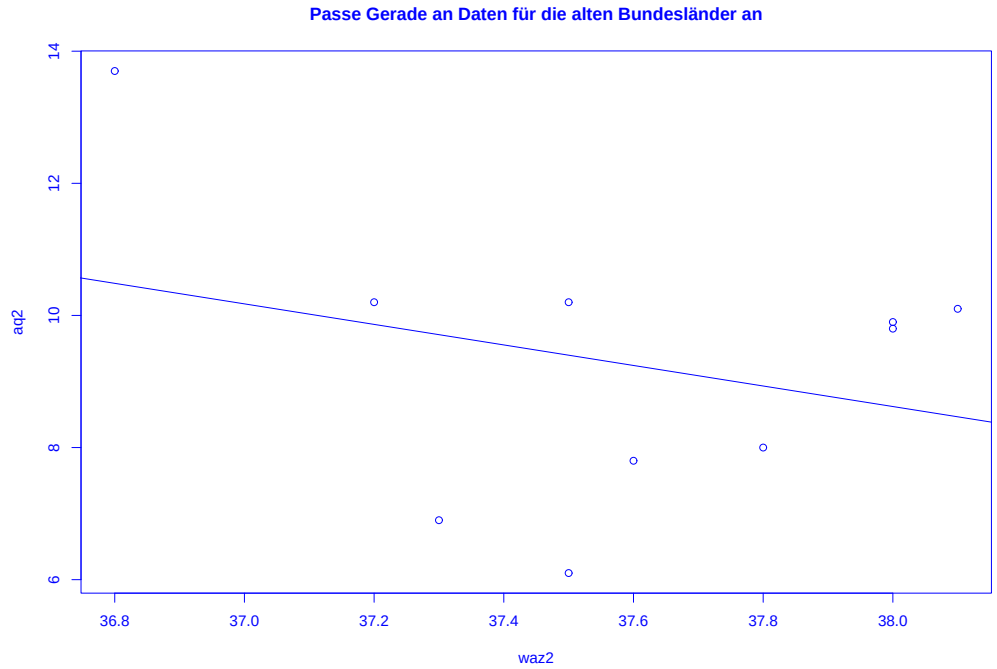
$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

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.

Kapitel 4: Wahrscheinlichkeitstheorie

4.1 Motivation

Die Statistik möchte Rückschlüsse aus Beobachtungen ziehen, die unter dem Einfluss des Zufalls enstanden sind.

Beispiel: Welche Rückschlüsse kann man aus den Ergebnissen beim Werfen eines Würfels

- über den Würfel ziehen ?
- über zukünftige Ergebnisse bei dem Würfel ziehen ?

Dazu hilfreich: **Mathematische Beschreibung des Zufalls!**

4.2 Mathematische Beschreibung des Zufalls

Ausgangspunkt der folgenden Betrachtungen ist ein sogenanntes *Zufallsexperiment*:

Definition. Ein **Zufallsexperiment** ist ein Experiment mit vorher unbestimmtem Ergebnis, das im Prinzip unbeeinflusst voneinander unter den gleichen Bedingungen beliebig oft wiederholt werden kann.

Die Menge Ω aller möglichen Ergebnisse heißt **Grundmenge**.

z.B. beim Werfen eines echten Würfels:

Ergebnis des Zufallsexperiments ist die Zahl, die auf der Seite des Würfels steht, die nach dem Wurf oben liegt.

$\Rightarrow \quad \Omega = \{1, 2, 3, 4, 5, 6\}$

Mehrfaches Durchführen eines Zufallsexperiments führe auf Ergebnisse x_1, ..., x_n.

z.B.: 10-maliges Werfen eines echten Würfels liefert die Ergebnisse

$x_1 = 5$, $x_2 = 1$, $x_3 = 5$, $x_4 = 2$, $x_5 = 4$, $x_6 = 6$, $x_7 = 3$, $x_8 = 5$, $x_9 = 3$, $x_{10} = 6$

Hier ist $n = 10$.

Absolute und relative Häufigkeit des Auftretens der einzelnen Zahlen:

	1	2	3	4	5	6
absolute Häufigkeit	1	1	2	1	3	2
relative Häufigkeit	0.1	0.1	0.2	0.1	0.3	0.2

Der Begriff des Ereignisses

Ein Ereignis ist eine Teilmenge der Grundmenge.

Ereignisse im Beispiel oben sind z.B. $A = \{1, 3, 5\}$ oder $B = \{1, 2, 3, 4, 5\}$.

Die einelementigen Teilmengen der Ergebnismenge heißen Elementarereignisse.

Die Elementarereignisse im Beispiel oben sind

$A_1 = \{1\}$, $A_2 = \{2\}$, $A_3 = \{3\}$, $A_4 = \{4\}$, $A_5 = \{5\}$ und $A_6 = \{6\}$

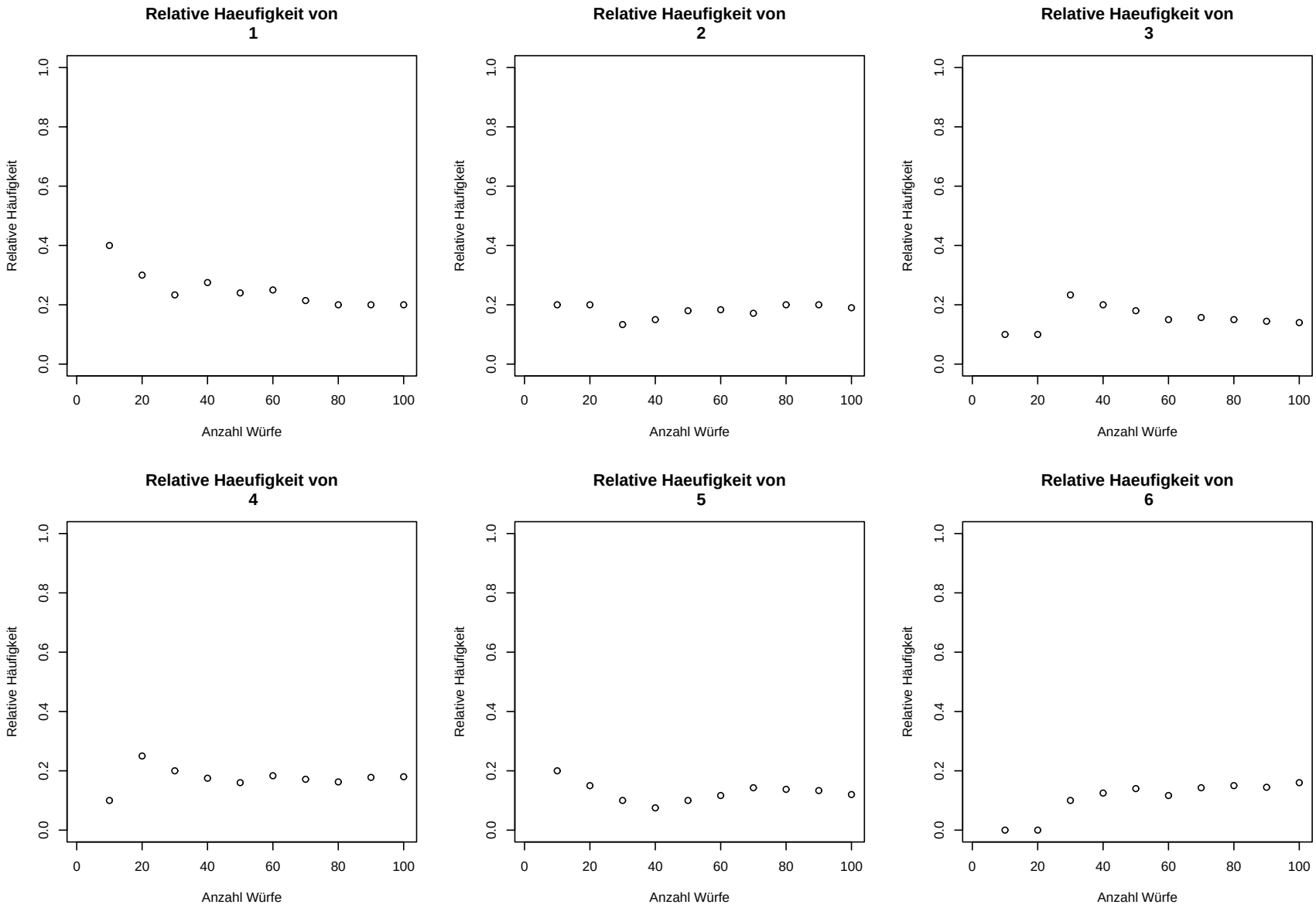
Ein Ereignis tritt ein, falls das Ergebnis des Zufallsexperiments im Ereignis liegt, andernfalls tritt es nicht ein.

Das empirische Gesetz der großen Zahlen:

Beobachtung aus der Praxis:

Führt man ein Zufallsexperiment **unbeeinflusst voneinander immer wieder** durch, so **nähert** sich die **relative Häufigkeit** des Auftretens eines festen Ereignisses A einer **festen Zahl** $\mathbf{P}(A) \in [0, 1]$ an.

Die Zahl $\mathbf{P}(A)$ nennen wir Wahrscheinlichkeit des Ereignisses A.


Relative Haeufigkeit von 1
Relative Haeufigkeit von 2
Relative Haeufigkeit von 3
Relative Haeufigkeit von 4
Relative Haeufigkeit von 5
Relative Haeufigkeit von 6
Relative Häufigkeit
Anzahl Würfe
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

Ziel im Folgenden: Bestimmung der Wahrscheinlichkeiten bei Zufallsexperimenten.

Möglichkeiten zur Bestimmung von Wahrscheinlichkeiten:

1. Zufallsexperiment sehr häufig durchführen, relative Häufigkeiten bestimmen.

2. Mit Symmetrieüberlegungen auf die Wahrscheinlichkeiten schließen.

3. Versuchen, durch allgemeine theoretische Überlegungen auf die Wahrscheinlichkeiten zu schließen.

Da 1. zu aufwendig ist, 2. nicht immer klappt, verfolgen wir primär Zugang 3.

Eigenschaften der Zuweisung von Wahrscheinlichkeiten zu Mengen:

(i) Für alle $A \subseteq \Omega$ gilt $0 \leq \mathbf{P}(A) \leq 1$.

(ii) $\mathbf{P}(\emptyset) = 0, \mathbf{P}(\Omega) = 1$.

(iii) Für alle $A \subseteq \Omega$ gilt: $\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A)$. (Hierbei $\bar{A} = \Omega \setminus A$).

(iv) Für alle $A, B \subseteq \Omega$ mit $A \cap B = \emptyset$ gilt: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

(v) Für alle $A_1, A_2, \cdots \subseteq \Omega$ mit $A_i \cap A_j = \emptyset$ für alle $i \neq j$ gilt:

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n) \quad \text{(sog. } \sigma\text{–Additivität)}.$$

Folgerungen aus (i)-(v):

Gelten die Bedingungen (i)-(v), so gilt z.B. auch:

- Für $A, B \subseteq \Omega$ mit $A \subseteq B$ gilt immer:

$$\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A).$$

- Für $A, B \subseteq \Omega$ mit $A \subseteq B$ gilt immer:

$$\mathbf{P}(A) \leq \mathbf{P}(B).$$

- Für beliebige $A, B \subseteq \Omega$ gilt immer:

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

Definition: Ein Paar $(\Omega, \mathbf{P})$ bestehend aus einer nichtleeren Menge Ω und einer Zuweisung $\mathbf{P}$ von Wahrscheinlichkeiten $\mathbf{P}(A)$ zu Ereignissen $A \subseteq \Omega$, die die Forderungen (i)-(v) von oben erfüllt, heißt **Wahrscheinlichkeitsraum**.

In diesem Falle heißt $\mathbf{P}$ **Wahrscheinlichkeitsmaß**.

Bemerkung: Aus technischen Gründen kann man meist nicht die Wahrscheinlichkeiten für **alle** Teilmengen von Ω sinnvoll festlegen, was hier aber im Folgenden vernachlässigt wird.

Im Beispiel oben führen Symmetrieüberlegungen auf

$$\mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = \mathbf{P}(\{3\}) = \mathbf{P}(\{4\}) = \mathbf{P}(\{5\}) = \mathbf{P}(\{6\}) = \frac{1}{6}.$$

Wegen (iv) folgt daraus sofort:

$$\mathbf{P}(A) = \frac{|A|}{6} = \frac{|A|}{|\Omega|}.$$

Damit ist der Wahrscheinlichkeitsraum in diesem Beispiel gegeben durch

$$(\Omega, \mathbf{P}) \quad \text{mit} \quad \Omega = \{1, \ldots, 6\} \text{ und } \mathbf{P}(A) = \frac{|A|}{6}.$$

4.3 Der Laplacesche Wahrscheinlichkeitsraum

Definition: Ein Wahrscheinlichkeitsraum $(\Omega, \mathbf{P})$ mit einer endlichen Grundmenge Ω und

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} \quad \text{für } A \subseteq \Omega$$

heißt **Laplacescher Wahrscheinlichkeitsraum**.

Dieser beschreibt ein Zufallsexperiment, bei dem

1. nur endlich viele verschiedene Werte auftreten,

2. jeder dieser Werte mit der gleichen Wahrscheinlichkeit $\frac{1}{|\Omega|}$ auftritt.

Im Laplaceschen Wahrscheinlichkeitsraum gilt:

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{\#\text{günstige Fälle}}{\#\text{mögliche Fälle}}.$$

Beispiel: Dezember 2007:
Höchster Jackpot aller Zeiten (43 Millionen Euro) beim Lotto "6 aus 49"

Spekulation der Medien: Was sind vielversprechende Zahlen beim Lotto ?

Häufigste Zahlen in den 4599 Ziehungen seit Oktober 1955:

1. 38 (614-mal gezogen)

2. 26 (606-mal gezogen)

3. 25 (600-mal gezogen)

Zum Vergleich: $4599 \cdot 6/49 \approx 563$

Frage: Ist es sinnvoll, speziell auf solche Zahlen zu tippen ?

Im Folgenden wollen wir entscheiden, ob diese Zahlen bei der Maschine, die die Lottozahlen erzeugt, vermutlich besonders häufig in der Zukunft auftreten werden.

Idee des Statistikers zur Entscheidung dieser Frage:

1. Gehe hypothetisch davon aus, dass die Zahlen "rein zufällig" gezogen werden, d.h. dass jede der endlich vielen möglichen Zahlenkombinationen mit der gleichen Wahrscheinlichkeit auftritt ($\Rightarrow$ Laplacescher W-Raum kann verwendet werden).

2. Berechne unter dieser Annahme die Wahrscheinlichkeit, dass bei 4599 Ziehungen ein Resultat auftritt, dass mindestens so stark gegen die obige Hypothese spricht wie das beobachtete Resultat (bei dem 614-mal die Zahl 38 gezogen wurde).

3. Falls die Wahrscheinlichkeit oben klein ist (z.B. kleiner als 0.05), so verwerfe die Hypothese oben, andernfalls verwerfe sie nicht.

Sei N die Anzahl der Möglichkeiten, 6 Zahlen aus 49 Zahlen *ohne Zurücklegen* und *ohne Beachtung der Reihenfolge* zu ziehen.

Dann gilt:

$$N \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44,$$

also ist

$$N = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \binom{49}{6} = 13983816$$

Hierbei

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n \cdot (n-1) \cdots 1}{k \cdot (k-1) \cdots 1 \cdot (n-k) \cdot (n-k-1) \cdots 1}.$$

Soll dabei aber einmal die 38 auftreten, so ist eine der Zahlen fest, und die übrigen 5 können noch aus 48 verschiedenen Zahlen ausgewählt werden, so dass dabei

$$\binom{48}{5}$$

verschiedene Möglichkeiten auftreten.

Daher tritt bei einer einzigen Ziehung die 38 mit Wahrscheinlichkeit

$$p = \frac{\binom{48}{5}}{\binom{49}{6}} = \frac{\frac{48!}{5!\cdot(48-5)!}}{\frac{49!}{6!\cdot(49-6)!}} = \frac{6}{49}$$

auf.

Zieht man nun n-mal unbeeinflusst voneinander rein zufällig 6 Zahlen aus 49, so ist die Wahrscheinlichkeit dass bei den ersten k Ziehungen die 38 auftritt, und bei den anschließenden $n-k$ Ziehungen die 38 nicht auftritt, gerade

$$\frac{\#\text{günstige Fälle}}{\#\text{mögliche Fälle}} = \frac{\binom{48}{5}^k \cdot \left(\binom{49}{6} - \binom{48}{5}\right)^{n-k}}{\binom{49}{6}^n} = p^k \cdot (1-p)^{n-k}.$$

Beachtet man, dass es nun $\binom{n}{k}$ viele verschiedene Möglichkeiten für die Anordnung der k Ziehungen gibt, bei denen die 38 jeweils auftritt, so sieht man, dass die Wahrscheinlichkeit für das k-malige Auftreten der 38 gegeben ist durch

$$\frac{\binom{n}{k} \cdot \binom{48}{5}^k \cdot \left(\binom{49}{6} - \binom{48}{5}\right)^{n-k}}{\binom{49}{6}^n} = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Damit erhalten wir für die Wahrscheinlichkeit, dass die 38 bei den $n = 4599$ Ziehungen mindestens 614-mal auftritt

$$\sum_{k=614}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \sum_{k=614}^{4599} \binom{4599}{k} \cdot \left(\frac{6}{49}\right)^k \cdot \left(1 - \frac{6}{49}\right)^{4599-k} \approx 0.01$$

Problem: Hypothese kann noch nicht abgelehnt werden, da nicht nur ein Ergebnis, bei dem die 38 mindestens 614-mal gezogen wird, sondern ebenso jedes andere Ergebnis, bei dem irgendeine der Zahlen zwischen 1 und 49 mindestens 614-mal gezogen wird, gegen die Hypothese spricht.

Also nötig: Berechnung der Wahrscheinlichkeit, dass mindestens eine der 49 Zahlen bei 4599 Ziehungen mindestens 614-mal gezogen wird.

Statt Berechnung: **Computersimulation.**

Wir simulieren mit einem Zufallszahlengenerator am Rechner $n = 4599$ Lottoziehungen, und bestimmen, ob dabei eine Zahl mindestens 614-mal auftritt. Anschließend wiederholen wir das Experiment sehr oft, bestimmen die relative Häufigkeit des Auftretens des obigen Ereignisses bei diesen Wiederholungen, und verwenden diese Zahl als Approximation für die gesuchte Wahrscheinlichkeit.

100000-malige Durchführung dieses Zufallsexperiments ergab als Schätzwert für die gesuchte Wahrscheinlichkeit ungefähr

$$\mathbf{0.47},$$

also bei fast jeder zweiten simulierten Abfolge der Lottoziehungen trat eine der Zahlen mindestens so häufig auf wie in der Realität beobachtet.

Folgerung: Auch beim rein zufälligen und unbeeinflussten Ziehen der Lottozahlen tritt ein solches Ergebnis keineswegs selten auf, so dass wir aufgrund der beobachteten Lotto-Zahlen nicht auf irgendwelche Defekte der Apparatur zur Ziehung der Lotto-Zahlen schließen können.

Also besser nicht auf eine der in der Vergangenheit häufig gezogenen Zahlen tippen, da dass vermutlich viele (mathematisch nicht ganz so gebildeten) Personen machen und daher bei diesen Zahlen der ausgezahlte Gewinn besonders klein ist.

4.4 Zufallsvariablen und Verteilungen

Oft interessieren nur Teilaspekte des Ergebnisses eines Zufallsexperimentes.

Idee: Wähle Abbildung

$$X : \Omega \to \Omega'$$

und betrachte anstelle des Ergebnisses ω des Zufallsexperimentes nur $X(\omega)$.

Beispiel: Werfen zweier echter Würfel

Kann modelliert werden durch Laplaceschen W-Raum $(\Omega, \mathbf{P})$ mit

$$\begin{aligned} \Omega &= \{(1,1),(1,2),\ldots,(1,6),\ldots,(6,1),(6,2),\ldots,(6,6)\}, \\ \mathbf{P}(\{\omega\}) &= \frac{1}{|\Omega|} = \frac{1}{36} \quad \text{für } \omega \in \Omega \quad \text{bzw.} \\ \mathbf{P}(A) &= \frac{|A|}{|\Omega|} = \frac{|A|}{36} \quad \text{für } A \subseteq \Omega. \end{aligned}$$

Falls nur die **Summe** der Augenzahlen interessiert:

Wähle

$$\Omega' = \{2, 3, \ldots, 12\}$$

und definiere $X : \Omega \to \Omega'$ durch

$$X((k, l)) = k + l.$$

Definition: Ist $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum, Ω' eine beliebige Menge und $X : \Omega \to \Omega'$ eine Abbildung, so heißt X **Zufallsvariable**.

Frage: Wie sieht ein Wahrscheinlichkeitsmaß $\mathbf{P}_X$ aus, dass das Zufallsexperiment mit unbestimmten Ergebnis $X(\omega)$ beschreibt ?

Idee: Für $A' \subseteq \Omega'$ setzen wir

$$\mathbf{P}_X(A') = \mathbf{P}\left(\{\omega \in \Omega \quad : \quad X(\omega) \in A'\}\right).$$

Im Beispiel oben: Hier war $\Omega' = \{2, 3, \ldots, 12\}$ und $X((k,l)) = k + l$. Dann ist

$$\begin{aligned}
\mathbf{P}_X\left(\{10, 11, 12\}\right) &= \mathbf{P}\left(\{\omega \in \Omega \quad : \quad X(\omega) \in \{10, 11, 12\}\}\right) \\
&= \mathbf{P}\left(\{(k,l) \in \Omega \quad : \quad k + l \in \{10, 11, 12\}\}\right) \\
&= \mathbf{P}\left(\{(4,6), (5,5), (5,6), (6,4), (6,5), (6,6)\}\right) = \frac{6}{36}.
\end{aligned}$$

Satz: Ist $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum, Ω' eine beliebige Menge und $X : \Omega \to \Omega'$ eine Abbildung, so wird durch

$$\mathbf{P}[X \in A] := \mathbf{P}_X(A') = \mathbf{P}\left(\{\omega \in \Omega \quad : \quad X(\omega) \in A'\}\right)$$

ein Wahrscheinlichkeitsmaß auf Ω' definiert (und damit ist auch $(\Omega', \mathbf{P}_X)$ ein Wahrscheinlichkeitsraum).

Definition: Das Wahrscheinlichkeitsmaß $\mathbf{P}_X$ heißt **Verteilung** der Zufallsvariablen X.

Bemerkungen:
a) Häufig verwendet man die Begriffe Wahrscheinlichkeitsmaß und Verteilung synonym.
b) Der große Vorteil von Zufallsvariablen ist, dass damit Operationen wie Aufsummieren der Ergebnisse von Zufallsexperimenten leicht beschreibbar sind.

4.5 Beispiele für Wahrscheinlichkeitsmaße und Verteilungen

Definition. Eine Folge $(p_n)_{n\in\mathbb{N}_0}$ reeller Zahlen mit

$$p_n \geq 0 \quad \text{für alle } n \in \mathbb{N}_0 \quad \text{und} \quad \sum_{n=0}^{\infty} p_n = 1$$

heißt **Zähldichte**.

Für sogenannte **diskrete Verteilungen** wählen wir $\Omega = \mathbb{N}_0$ und eine Zähldichte $(p_n)_{n\in\mathbb{N}_0}$ und setzen

$$\mathbf{P}(A) = \sum_{k\in A} p_k.$$

Hierbei gibt p_k die Wahrscheinlichkeit für das Eintreten des Elementarereignisses $\{k\}$ an.

Beispiele für diskrete Verteilungen:

1. Sei $n \in \mathbb{N}$ und $p \in [0,1]$. Die zur Zähldichte

$$p_k = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{für} \quad 0 \leq k \leq n, \\ 0 & \text{für} \quad k > n, \end{cases}$$

gehörende Verteilung heißt **Binomialverteilung** mit Parametern n und p.

Eine Zufallsvariable X heißt **binomialverteilt** mit Parametern n und p, falls ihre Verteilung eine **Binomialverteilung** mit Parametern n und p ist.

Einsatz in der Modellierung:

Wird ein Zufallsexperiment n-mal unbeeinflusst voneinander durchgeführt, wobei jedesmal mit Wahrscheinlichkeit p Erfolg und mit Wahrscheinlichkeit $1-p$ Misserfolg eintritt, so ist die **Anzahl der Erfolge binomialverteilt mit Parametern n und p.**

2. Sei $\lambda \in \mathbb{R}_+ \setminus \{0\}$. Die zur Zähldichte

$$p_k = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

gehörende Verteilung heißt **Poisson-Verteilung** mit Parameter λ.

Eine Zufallsvariable X heißt **Poisson-verteilt** mit Parameter λ, falls ihre Verteilung eine **Poisson-Verteilung** mit Parameter λ ist.

Einsatz in der Modellierung:
Eine binomialverteilte Zufallsvariable mit Parametern n und p kann für n groß und p klein durch eine **Poisson-verteilte** Zufallsvariable mit Parameter $\lambda = n \cdot p$ approximiert werden.

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.

Für sogenannte **stetige Verteilungen** wählen wir $\Omega = \mathbb{R}$ und eine Dichte $f : \mathbb{R} \to \mathbb{R}$ und setzen

$$\mathbf{P}(A) = \int_A f(x)\, dx.$$

Hierbei sind die Wahrscheinlichkeiten für das Eintreten eines Elementarereignisses immer Null.

Beispiele für stetige Verteilungen:

1. Die *Gleichverteilung* $U(a,b)$ mit Parametern $-\infty < a < b < \infty$ ist das durch die Dichte

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für} \quad a \leq x \leq b, \\ 0 & \text{für} \quad x < a \text{ oder } x > b \end{cases}$$

festgelegte W–Maß.

Eine Zufallsvariable X heißt **gleichverteilt** auf dem Intervall $[a,b]$, falls ihre Verteilung eine **Gleichverteilung** mit Parametern a und b ist.

Einsatz in der Modellierung:

"Rein zufälliges Ziehen" einer Zahl aus einem Intervall.

2. Die *Exponentialverteilung* $\exp(\lambda)$ mit Parameter $\lambda > 0$ ist das durch die Dichte

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{für} \quad x \geq 0, \\ 0 & \text{für} \quad x < 0 \end{cases}$$

festgelegte W–Maß.

Eine Zufallsvariable X heißt **exponentialverteilt** mit Parameter λ, falls ihre Verteilung eine **Exponentialverteilung** mit Parameter λ ist.

Einsatz in der Modellierung:
Lebensdauern oder Wartevorgänge werden häufig durch Exponentialverteilungen modelliert.

3. Die *Normalverteilung* $N(\mu, \sigma^2)$ mit Parametern $\mu \in \mathbb{R}, \sigma > 0$ ist das durch die Dichte

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

festgelegte W–Maß.

Eine Zufallsvariable X heißt **normalverteilt** mit Parametern μ und σ^2, falls ihre Verteilung eine **Normalverteilung** mit Parametern μ und σ^2 ist.

Einsatz in der Modellierung:
Summen von Zufallsvariablen der gleichen Art, die sich gegenseitig nicht beeinflussen, werden häufig durch Normalverteilungen approximiert.

4.6 Erwartungswert und Varianz

Sei $(\Omega, \mathbf{P})$ Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit Werten in $\mathbb{R}$ (sog. *reelle Zufallsvariable*).

Gesucht: Definieren wollen wir einen *mittleren Wert* des Zufallsexperiments mit Ergebnis $X(\omega)$, den wir als **Erwartungswert EX** bezeichnen werden.

Vor Definition des Erwartungswertes beschreiben wir zuerst drei allgemeine Eigenschaften des Erwartungswertes, die sich anschaulich mit der Vorstellung als "mittlerer Wert" begründen lassen.

1. *Monotonie*: Für zwei beliebige reelle ZVen X und Y gilt immer:

$$X(\omega) \leq Y(\omega) \quad \text{für alle } \omega \in \Omega \quad \Rightarrow \quad \mathbf{E}X \leq \mathbf{E}Y$$

2. *Linearität:* Für zwei beliebige reelle ZVen X und Y und beliebige reelle Zahlen $\alpha, \beta \in \mathbb{R}$ gilt immer:

$$\mathbf{E}\left(\alpha \cdot X + \beta \cdot Y\right) = \alpha \cdot \mathbf{E}X + \beta \cdot \mathbf{E}Y.$$

3. *Erwartungswert des Produktes unabhängiger Zufallsvariablen:*

   Beeinflussen sich die Werte der reellen Zufallsvariablen X und Y gegenseitig nicht, so gilt immer:

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

Unabhängigkeit von Ereignissen

Sei $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum, und seien $A, B \subseteq \Omega$ zwei Ereignisse. Bei n-maligen Durchführen des zugrundeliegenden Zufallsexperiments seien A bzw. B bzw. $A \cap B$ jeweils n_A bzw. n_B bzw. $n_{A\cap B}$ mal eingetreten.

Falls sich die Ereignisse A und B gegenseitig nicht beeinflussen, sollte für großes n approximativ gelten:

$$\frac{n_{A\cap B}}{n_B} \approx \frac{n_A}{n} \quad \text{und} \quad \frac{n_{A\cap B}}{n_A} \approx \frac{n_B}{n} \quad \Leftrightarrow \quad \frac{n_{A\cap B}}{n} \approx \frac{n_A}{n} \cdot \frac{n_B}{n}.$$

Definition. A und B heißen **unabhängig**, falls gilt:

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

Die folgende Definition beschreibt formal, wann sich zwei Zufallsvariablen gegenseitig nicht beeinflussen:

Definition. Sei $(\Omega, \mathbf{P})$ Wahrscheinlichkeitsraum und $X, Y : \Omega \to \mathbb{R}$ reelle Zufallsvariablen. Dann heißen X und Y **unabhängig**, falls für alle $A, B \subseteq \mathbb{R}$ gilt:

$$\mathbf{P}[X \in A, Y \in B] = \mathbf{P}[X \in A] \cdot \mathbf{P}[Y \in B].$$

Die obige Regel besagt also, dass für unabhängige reelle Zufallsvariablen immer gilt:

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

4.6.1 Erwartungswert von diskreten Zufallsvariablen

Sei X eine diskrete Zufallsvariable, die mit Wahrscheinlichkeit Eins nur einen der Werte $x_1, x_2, \ldots, x_K \in \mathbb{R}$ annimmt.

n-maliges Durchführen des Zufallsexperiment mit Ergebnis $X(\omega)$ liefere die Werte $z_1, \ldots, z_n$.

Idee:

$$\begin{aligned} \mathbf{E}X &\approx \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n} \cdot \left(\sum_{k=1}^{K} x_k \cdot \#\{1 \le i \le n \,:\, z_i = x_k\} \right) \\ &= \sum_{k=1}^{K} x_k \cdot \frac{\#\{1 \le i \le n \,:\, z_i = x_k\}}{n} \approx \sum_{k=1}^{K} x_k \cdot \mathbf{P}[X = x_k]. \end{aligned}$$

Definition: Sei X eine diskrete Zufallsvariable, die mit Wahrscheinlichkeit Eins nur einen der Werte $x_1, x_2, \ldots, x_K \in \mathbb{R}$ bzw. $x_1, x_2, \cdots \in \mathbb{R}$ annimmt. Dann heißt

$$\mathbf{E}X = \sum_{k=1}^{K} x_k \cdot \mathbf{P}[X = x_k]$$

bzw. (sofern existent)

$$\mathbf{E}X = \sum_{k=1}^{\infty} x_k \cdot \mathbf{P}[X = x_k]$$

der **Erwartungswert** von X.

Hierbei: $\mathbf{P}[X = x_k] := \mathbf{P}_X(\{x_k\}) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = x_k\})$.

Beispiel. Betrachtet wird das (zufällige) Werfen zweier echter Würfel. Die Zufallsvariable X gebe die Summe der beiden Augenzahlen an, die oben landen.

X ist diskret verteilt, nimmt mit Wahrscheinlichkeit Eins nur einen der Werte in $\{2, 3, \ldots, 12\}$ an und es gilt:

k	2	3	4	5	6	7	8	9	10	11	12
$\mathbf{P}[X = k]$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Damit

$$\begin{aligned}\mathbf{E}X &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} \\ &\quad + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} \\ &= \frac{252}{36} = 7.\end{aligned}$$

Einfacher: Es gilt $X = X_1 + X_2$ wobei X_1 bzw. X_2 die Augenzahlen des ersten bzw. zweiten Würfels ist.

Dabei ist

$$\mathbf{E}X_1 = \mathbf{E}X_2 = \sum_{k=1}^{6} k \cdot \frac{1}{6} = 3.5$$

und damit

$$\mathbf{E}(X_1 + X_2) = \mathbf{E}X_1 + \mathbf{E}X_2 = 3.5 + 3.5 = 7.$$

Allgemeiner gilt:

Ist X eine diskrete Zufallsvariable die mit Wahrscheinlichkeit Eins nur einen der Werte $x_1, x_2, \ldots, x_K \in \mathbb{R}$ bzw. $x_1, x_2, \cdots \in \mathbb{R}$ annimmt, und ist $h : \mathbb{R} \to \mathbb{R}$ eine beliebige reelle Funktion.

Dann ist $h(X)$ eine diskrete Zufallsvariable, deren Erwartungswert gegeben ist durch

$$\mathbf{E}h(X) = \sum_{k=1}^{K} h(x_k) \cdot \mathbf{P}[X = x_k]$$

bzw. (sofern existent)

$$\mathbf{E}h(X) = \sum_{k=1}^{\infty} h(x_k) \cdot \mathbf{P}[X = x_k].$$

4.6.2 Erwartungswert von Zufallsvariablen mit Dichten

Im Falle einer stetig verteilten Zufallsvariablen X mit Dichte f ersetzt man die Summe in den vorigen Definitionen durch das entsprechende Integral:

Definition: Sei X eine stetig verteilte Zufallsvariable mit Dichte f. Dann heißt

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

– sofern existent – der **Erwartungswert** von X.

Allgemeiner setzt man wieder:

Ist X eine stetig verteilte Zufallsvariable mit Dichte f, und ist $h : \mathbb{R} \to \mathbb{R}$ eine beliebige reelle Funktion.

Dann definieren wir den **Erwartungswert von** $h(X)$ als

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

(sofern existent).

Beispiel: Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ^2, d.h. X ist eine stetig-verteilte Zufallsvariable mit Dichte

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

Dann gilt:

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \overset{(!)}{=} \mu.$$

4.6.3 Varianz

Der Ewartungswert beschreibt den Wert, den man "im Mittel" bei Durchführung eines Zufallsexperiments erhält. Ein Kriterium zur Beurteilung der zufälligen Schwankung des Resultats eines Zufallsexperiments um diesen Mittelwert ist die sogenannte Varianz, die die mittlere quadratische Abweichung zwischen einem zufälligen Wert und seinem Mittelwert beschreibt:

Definition: Sei X eine reelle ZV für die $\mathbf{E}X$ existiert. Dann heißt

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

die **Varianz** von X.

Beispiel: Für eine normalverteilte Zufallsvariable X mit Parametern μ und σ^2 gilt

$$\begin{aligned} V(X) &= \mathbf{E}(|X - \mathbf{E}X|^2) \\ &= \mathbf{E}(|X - \mu|^2) \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \, dx \\ &\overset{(!)}{=} \sigma^2. \end{aligned}$$

Nützliche Rechenregeln für die Berechnung von Varianzen:

Lemma: Sei X eine reelle ZV für die $\mathbf{E}X$ existiert. Dann gilt:

a)

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

b) Für alle $\alpha \in \mathbb{R}$:

$$V(\alpha \cdot X) = \alpha^2 \cdot V(X).$$

c) Für alle $\beta \in \mathbb{R}$:

$$V(X + \beta) = V(X).$$

Für **unabhängige** Zufallsvariablen ist darüberhinaus die **Varianz der Summe gleich der Summe der Varianzen**:

Satz:

Sind X und Y zwei unabhängige reelle Zufallsvariablen definiert auf dem gleichen Wahrscheinlichkeitsraum, so gilt:

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

Entsprechendes gilt für beliebige endliche Summen unabhängiger Zufallsvariablen.

Kapitel 5: Schließende Statistik

Wir gehen in der schließenden Statistik davon aus, dass die Daten gemäß einem stochastischen Modell erzeugt wurden. Eigenschaften dieses Modells beschreiben dann die zugrunde liegende Grundgesamtheit.

Ziel:
Herleitung von Aussagen über **Eigenschaften dieses Modells**, wie z.B.:
Wie groß sind Erwartungswert und Varianz im stochastischen Modell ?

Dies wird es uns ermöglichen, von dem vorliegenden Datensatz auf die Grundgesamtheit zu schließen!

Beispiel: Im Rahmen der Shell Jugendstudie 2006 wurden 2532 Jugendliche aus ganz Deutschland befragt. Dabei gaben 39% der Befragten an, dass sie politisch interessiert sind.

Wir fassen diese Daten als $n = 2532$ unbeeinflusst voneinander entstandene Realisierungen $x_1, \ldots, x_n$ einer $\{0, 1\}$-wertigen ZV X auf, die den Wert 1 genau dann annimmt, wenn der befragte Jugendliche politisch interessiert ist. Wir nehmen an, dass

$$p = \mathbf{P}[X = 1]$$

mit dem Prozentsatz der politisch interessierten Jugendlichen in Deutschland im Jahr 2006 übereinstimmt.

Ausgehend von den Beobachtungen $x_1, \ldots, x_n$ mit

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \frac{\#\{1 \leq i \leq n : x_i = 1\}}{n} = 0.39$$

wollen wir Rückschlüsse auf p ziehen.

Annahme an die Erzeugung der Daten:

Informal: Wir gehen davon aus, dass alle Datenpunkte unbeeinflusst voneinander und nach dem gleichen Prinzip erzeugt werden.

Formal: Unsere Stichprobe $x_1, \ldots, x_n$ ist Realisierung der ersten n-Glieder $X_1, \ldots, X_n$ einer Folge $(X_k)_{k \in \mathbb{N}}$ von reellen Zufallsvariablen definiert auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathbf{P})$, die **unabhängig** und **identisch verteilt** sind in dem Sinne, dass:

1.
$$\mathbf{P}\left[X_1 \in A_1, \ldots, X_n \in A_n\right] = \mathbf{P}\left[X_1 \in A_1\right] \cdots \mathbf{P}\left[X_n \in A_n\right]$$
für alle $A_1, \ldots, A_n \subseteq \mathbb{R}$ und alle $n \in \mathbb{N}$.

2.
$$\mathbf{P}_{X_1} = \mathbf{P}_{X_2} = \mathbf{P}_{X_3} = \ldots$$

5.1 Punktschätzverfahren

geg.: Realisierungen $x_1, \ldots, x_n$ von reellen Zufallsvariablen $X_1, \ldots, X_n$, wobei $X_1, X_2, \ldots$ unabhängig identisch verteilt sind.

ges.: Schätzung $T_n(x_1, \ldots, x_n)$ von einem "Parameter" θ der Verteilung von X_1, z.B. vom Erwartungswert oder von der Varianz von X_1.

Beispiele:

1. $T_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$ ist Schätzung von $\mathbf{E}X_1$.

2. $T_n(x_1, \ldots, x_n) = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2$ ist Schätzung von $V(X_1)$.

Sinnvolle Eigenschaften von Schätzungen:

a) Asymptotisch (d.h. sofern der Stichprobenumfang n gegen Unendlich strebt) ergibt sich der richtige Wert.

b) Im Mittel (d.h. bei wiederholter Erzeugung der Stichproben und Mittelung der Ergebnisse) ergibt sich (asymptotisch mit wachsender Zahl der Wiederholungen) der richtige Wert.

Formal:

Definition:

a) Eine Schätzung $T_n(x_1, \ldots, x_n)$ von θ heißt stark konsistente Schätzung für θ, falls gilt

$$\mathbf{P}\left(\{\omega \in \Omega \,:\, T_n(X_1(\omega), \ldots, X_n(\omega)) \to \theta \quad (n \to \infty)\}\right) = 1.$$

a) Eine Schätzung $T_n(x_1, \ldots, x_n)$ von θ heißt erwartungstreue Schätzung für θ, falls gilt

$$\mathbf{E}\big(T_n(X_1, \ldots, X_n)\big) = \theta.$$

Bemerkung: Bei a) handelt es sich um sogenannte fast sichere (f.s.) Konvergenz einer Folge von Zufallsvariablen:

Sind $Z, Z_1, Z_2, \ldots$ reelle Zufallsvariablen definiert auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathbf{P})$, so sagt man: Z_n konvergiert gegen Z fast sicher (Schreibweise: $Z_n \to Z \ f.s.$), falls gilt:

$$\mathbf{P}\left(\{\omega \in \Omega : Z_n(\omega) \to Z(\omega) \quad (n \to \infty)\}\right) = 1.$$

Mit der fast sicheren Konvergenz kann man rechnen wie mit reellen Zahlenfolgen, d.h. es gilt z.B. für beliebige reelle Zahlen $\alpha, \beta \in \mathbb{R}$:

$$X_n \to X \quad f.s., Y_n \to Y \quad f.s. \quad \Rightarrow \quad \alpha \cdot X_n + \beta \cdot Y_n \to \alpha \cdot X + \beta \cdot Y \quad f.s.$$

$$X_n \to X \quad f.s. \quad \Rightarrow \quad X_n^2 \to X^2 \quad f.s.$$

Die Schätzung $T_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$ ist erwartungstreue Schätzung für $\mathbf{E}X_1$, denn es gilt:

$$\mathbf{E}\left(T_n(X_1, \ldots, X_n)\right) = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(X_1\right) = \mathbf{E}(X_1).$$

Die Schätzung $T_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$ ist auch stark konsistente Schätzung für $\mathbf{E}X_1$, denn es gilt:

Satz (Starkes Gesetz der großen Zahlen):

Sind die auf dem selben Wahrscheinlichkeitsraum definierten reellen Zufallsvariablen $X_1, X_2, \ldots$ unabhängig und identisch verteilt, und existiert $\mathbf{E}X_1$, so gilt:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbf{E}X_1 \quad f.s.$$

Beispiel zum starken Gesetz der großen Zahlen:

Beim wiederholten unbeeinflussten Werfen eines echten Würfels nähert sich das arithmetische Mittel der bisher geworfenen Augenzahlen für große Anzahl von Würfen (mit Wahrscheinlichkeit Eins) immer mehr dem Erwartungswert 3.5 an.

Auch unsere Schätzung für die Varianz ist stark konsistent, denn es gilt:

$$
\begin{aligned}
T_n(X_1,\ldots,X_n) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2 \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2 \\
&\overset{(!)}{=} \frac{n}{n-1}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right) \\
&\to 1\cdot(\mathbf{E}(X_1^2)-(\mathbf{E}X_1)^2)=V(X_1) \quad f.s.
\end{aligned}
$$

Darüberhinaus ist sie wegen

$$\mathbf{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right)\overset{(!)}{=}V(X_1)$$

auch erwartungstreu.

5.2 Bereichsschätzverfahren

geg.: Realisierungen $x_1, \ldots, x_n$ von reellen Zufallsvariablen $X_1, \ldots, X_n$, wobei $X_1, X_2, \ldots$ unabhängig identisch verteilt sind.

ges.: Sogenannter **Konfidenzbereich** $C_n(x_1, \ldots, x_n) \subseteq \mathbb{R}$, in dem ein "Parameter" $\theta \in \mathbb{R}$ der Verteilung von X_1, z.B. der Erwartungswert oder die Varianz von X_1, mit "möglichst großer" Wahrscheinlichkeit liegt.

Im Folgenden sind wir in erster Linie an Intervallen der Form $[a, b]$ bzw. $(-\infty, b]$ bzw. $[a, \infty)$ interessiert, in denen der gesuchte Parameter mit möglichst großer Wahrscheinlichkeit liegt.

Beispiel: Wie kann man aus den Ergebnissen der Shell Jugendstudie 2006 (bei der 39% der 2532 befragten Jugendlichen politisch interessiert waren) ein "möglichst kleines" Intervall konstruieren, in dem der Anteil p der politisch interessierten Jugendlichen mit "möglichst großer" Wahrscheinlichkeit liegt ?

Modellieren wir die Antwort der befragten Jugendlichen als Realisierungen einer $b(1,p)$-verteilten Zufallsvariablen X, so sieht man, dass wir wegen

$$\mathbf{E}X = 0 \cdot \mathbf{P}[X=0] + 1 \cdot \mathbf{P}[X=1] = \mathbf{P}[X=1] = p$$

eigentlich ein Konfidenzintervall für den Erwartungswert unserer Stichprobe suchen.

Def.: Sei $\alpha \in [0,1]$. Dann heißt $C_n(x_1,\ldots,x_n) = [a(x_1,\ldots,x_n), b(x_1,\ldots,x_n)]$ **zweiseitiges Konfidenzintervall zum Niveau** α für den Erwartungswert, falls für **alle** (in dem Kontext zugelassenen) **Verteilungen** von X_1 gilt:

$$\mathbf{P}\left[\mathbf{E}X_1 \in C_n(X_1,\ldots,X_n)\right] \geq \alpha.$$

Entsprechend werden einseitige Konfidenzintervalle zum Niveau α als Konfidenzintervalle der Form $(-\infty, b(x_1,\ldots,x_n)]$ bzw. $[a(x_1,\ldots,x_n), \infty)$ definiert.

Beispiel: Naheliegender Ansatz für ein zweiseitiges Konfidenzintervall zum Niveau α für den Erwartungswert ist mit $c > 0$ geeignet:

$$C_n(x_1, \ldots, x_n) = \left[\frac{1}{n} \sum_{i=1}^{n} x_i - c, \frac{1}{n} \sum_{i=1}^{n} x_i + c \right]$$

Frage: Wie wählt man c in Abhängigkeit von α und der Stichprobe $x_1, \ldots, x_n$?

Der zentrale Grenzwertsatz:

Sind $X_1, X_2, \ldots$ unabhängige und identisch verteilte reelle Zufallsvariablen mit $\mathbf{E}X_1^2 < \infty$, so ist für n groß

$$\frac{\sum_{i=1}^n X_i - \mathbf{E}\left(\sum_{i=1}^n X_i\right)}{\sqrt{V(\sum_{i=1}^n X_i)}} = \frac{\sqrt{n}}{\sqrt{V(X_1)}} \cdot \left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right)$$

annähernd $N(0,1)$- verteilt.

Genauer gilt dann für jedes $x \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbf{P}\left[\frac{\sqrt{n}}{\sqrt{V(X_1)}} \cdot \left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right) \leq x\right] = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

Beispiel: X_i sei die Augenzahl die man beim i-ten unbeeinflussten Werfen eines echten Würfel erhält. Dann gilt

$$\begin{aligned} \mathbf{E}X_1 &= \sum_{i=1}^{6} i \cdot \mathbf{P}[X_1 = i] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = 3.5, \\ V(X_1) &= \mathbf{E}(X_1^2) - (\mathbf{E}X_1)^2 = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} - (3.5)^2 = \frac{35}{12}. \end{aligned}$$

Nach dem zentralen Grenzwertsatz verhält sich also

$$\frac{\sqrt{n}}{\sqrt{35/12}} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i - 3.5 \right)$$

für große n annähernd wie eine $N(0,1)$-verteilte Zufallsvariable.

Aufgabe: Werfen Sie einen echten Würfel $n = 15$-mal und notieren Sie sich die Summe $(x_1 + \cdots + x_{15})$ der Augenzahlen $x_1, \ldots, x_{15}$ die oben landen.

Folgerung: Wählen wir $\delta \in \mathbb{R}$ so, dass für eine $N(0,1)$-verteilte Zufallsvariable Z gilt

$$\mathbf{P}[|Z| \leq \delta] \geq \alpha,$$

so gilt für n groß approximativ:

$$\mathbf{P}\left[\left|\frac{\sqrt{n}}{\sqrt{V(X_1)}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X_1\right)\right| \leq \delta\right] \geq \alpha,$$

Wegen

$$\left|\frac{\sqrt{n}}{\sqrt{V(X_1)}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X_1\right)\right| \leq \delta$$

$$\Leftrightarrow \quad \mathbf{E}X_1 \in \left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot \delta, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot \delta\right]$$

gilt auch in diesem Fall für n groß approximativ:

$$\mathbf{P}\left[\mathbf{E}X_1 \in \left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot \delta, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot \delta\right]\right] \geq \alpha.$$

Damit das so konstruierte Konfidenzintervall möglichst klein wird, wählen wir δ so klein wie möglich, was auf die Bedingung

$$\mathbf{P}[|Z| \leq \delta] = \alpha$$

führt (für eine $N(0,1)$-verteilte ZV Z).

Problem: Konfidenzintervall hängt noch von der (in aller Regel unbekannten) Varianz von X_1 ab.

Ausweg: Varianz durch empirische Varianz schätzen.

Für das gesuchte Konfidenzintervall erhalten wir dann

$$C(x_1, \ldots, x_n) = \left[\bar{x} - \frac{\sqrt{s_n^2}}{\sqrt{n}} \cdot \delta, \bar{x} + \frac{\sqrt{s_n^2}}{\sqrt{n}} \cdot \delta \right].$$

Beispiel: Zweiseitiges Konfidenzintervall für den Anteil der politisch interessierten Jugendlichen in Deutschland im Jahr 2006 zum Niveau $\alpha = 0.95$:

- Hier ist $n = 2532$, $\bar{x} = 0.39$ und

$$s_n^2 = \frac{n}{n-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\frac{1}{n} \sum_{j=1}^{n} x_j)^2 \right) = \frac{n}{n-1} \cdot (\bar{x} - (\bar{x})^2) \approx 0.238$$

- Für eine $N(0,1)$-verteilte Zufallsvariable Z gilt: $\mathbf{P}[|Z| \leq 1.96] = 0.95$.
  Also wählen wir $\delta = 1.96$.

- Das gesuchte (approximative) Konfidenzintervall ist dann

$$C(x_1, \ldots, x_n) = \left[\bar{x} - \frac{\sqrt{s_n^2}}{\sqrt{n}} \cdot \delta, \bar{x} + \frac{\sqrt{s_n^2}}{\sqrt{n}} \cdot \delta \right] \approx [0.38, 0.40].$$

5.3 Statistische Testverfahren

5.3.1. Beispiele:

1. Sind Examenskandidaten in der Lage, ihre eigene Leistungsfähigkeit einzuschätzen ?

   $n = 15$ Kandidaten wurde eine Klausur mit 70 Aufgaben gestellt. Nach Bearbeitung der Klausur wurden Sie gebeten, die Anzahl der richtig gelösten Aufgaben zu schätzen. Nach der Korrektur der Klausur wurden die Differenzen

   $$X_i = \text{Gesch. Anz. gelöster Aufgaben } - \text{Tatsächliche Anz. gelöster Aufgaben}$$

   gebildet.

   Beschreibung der gemessenen Daten: $n = 15$, $\bar{x} = -6.4$, $s^2 = 61.7$

2. Sprechen Frauen mehr als Männer ?

Vorhandene Daten:

Im Rahmen einer Studie an der Universität Arizona wurden bei 210 Studentinnen und 186 Studenten approximativ die Anzahl der gesprochenen Worte über einen Zeitraum von mehreren Tagen bestimmt. Für die empirischen arithmetischen Mittel der Anzahlen der gesprochenen Wörter pro Tag ergab sich:

- $n = 210$ Studentinnen: $\bar{x} = 16215$, $s_x = 7301$
- $m = 186$ Studenten: $\bar{y} = 15669$, $s_y = 8633$

Frage: Wie kann man ausgehend von den Daten in der Stichprobe Rückschlüsse auf die zugrunde liegende Grundgesamtheit so ziehen, dass man die dabei zwangsläufig auftretenden Fehler quantitativ kontrollieren kann ?

5.3.2. Mathematische Modellbildung:

1. Wir gehen davon aus, dass die Daten unter Einfluss des Zufalls (wie im mathematischen Modell des Zufalls in dieser Vorlesung beschrieben) entstanden sind.

2. Wir fassen die Daten als Stichprobe einer uns unbekannten (stochastischen) Verteilung auf:

   - In Beipiel 1 fassen wir unsere Daten als Realisierungen $x_1, \ldots, x_{15}$ von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots X_{15}$ auf.
   - In Beipiel 2 fassen wir unsere Daten als Realisierungen $x_1, \ldots, x_{210}$ von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots X_{210}$ bzw. $y_1, \ldots, y_{186}$ von unabhängigen identisch verteilten Zufallsvariablen $Y_1, \ldots Y_{186}$ auf.

3. Wir formulieren unsere Frage so um, dass sie nur von den zugrunde liegenden Verteilungen abhängt:

   - In Beispiel 1 wollen wir wissen, welche von den beiden Hypothesen

$$H_0: \quad \mathbf{E}X_1 = 0$$
$$H_1: \quad \mathbf{E}X_1 \neq 0$$

     zutrifft.
   - In Beispiel 2 wollen wir wissen, welche von den beiden Hypothesen

$$H_0: \quad \mathbf{E}X_1 = \mathbf{E}Y_1$$
$$H_1: \quad \mathbf{E}X_1 \neq \mathbf{E}Y_1$$

     zutrifft.

Prinzipieller Unterschied zwischen den beiden Fragestellungen:

- In Beispiel 1 haben wir **eine** Stichprobe $x_1, \ldots x_{15}$ der Verteilung von X_1 gegeben, und wollen wissen, ob $\mathbf{E}X_1 = 0$ gilt (**Einstichprobenproblem**).

- In Beipiel 2 haben wir **zwei** Stichproben $x_1, \ldots, x_{210}$ bzw. $y_1, \ldots, y_{186}$ der Verteilungen von X_1 bzw. Y_1 gegeben, und wollen wissen, ob $\mathbf{E}X_1 = \mathbf{E}Y_1$ gilt (**Zweistichprobenproblem**).

Anmerkung:

Wir haben die auftretenden Fragestellungen als **zweiseitige Testprobleme** formuliert. Alternativ könnte man auch sogenannte **einseitige Testprobleme** betrachten, wie z.B.

- Gilt in Beispiel 1 $H_0 : \mathbf{E}X_1 \leq 0$ oder $H_1 : \mathbf{E}X_1 > 0$?
- Gilt in Beispiel 2 $H_0 : \mathbf{E}X_1 \leq \mathbf{E}Y_1$ oder $H_1 : \mathbf{E}X_1 > \mathbf{E}Y_1$?

4. Um die Fragestellung zu vereinfachen, machen wir Annahmen über die Art der in den Beispielen auftretenden Verteilungen:

   Wir gehen im folgenden davon aus, dass alle auftretenden Verteilungen **Normalverteilungen** mit **unbekanntem Erwartungswert** und **bekannter oder unbekannter Varianz** sind.

5. Unter diesen Annahmen ermitteln wir geeignete Verfahren, die es uns (mit kontrollierter Fehlerwahrscheinlichkeit) ermöglichen, zwischen den beiden Hypothesen zu entscheiden.

5.3.3 Grundbegriffe der Testtheorie

geg.: Realisierungen $x_1, \ldots, x_n$ von unabhängigen identisch verteilten reellen Zufallsvariablen $X_1, \ldots, X_n$.

ges.: Entscheidungsvorschrift zur Entscheidung zwischen zwei Hypothesen über die zugrunde liegenden Verteilung, z.B. Hypothesen wie

$$H_0: \quad \mathbf{E}X_1 = 0$$
$$H_1: \quad \mathbf{E}X_1 \neq 0$$

Definition. Ein **statistischer Test** ist eine Abbildung

$$\phi : \mathbb{R}^n \to \{0, 1\}.$$

Deutung des Tests: Im Falle von $\phi(x_1, \ldots, x_n) = 0$ entscheiden wir uns für H_0, im Falle $\phi(x_1, \ldots, x_n) = 1$ entscheiden wir uns für H_1.

Bezeichnung für die auftretenden Fehler:

- Gilt H_0 (die sogenannte Nullhypothese), liefert unser Test aber fälschlicherweise $\phi(x_1, \ldots, x_n) = 1$ und entscheiden wir uns daher für H_1 (die sogenannte Alternativhypothese), so sprechen wir von einem **Fehler erster Art**.

- Gilt H_1 (die Alternativhypothese), liefert unser Test aber fälschlicherweise $\phi(x_1, \ldots, x_n) = 0$ und entscheiden wir uns daher für H_0 (die Nullhypothese), so sprechen wir von einem **Fehler zweiter Art**.

Die entsprechenden Wahrscheinlichkeiten für das Auftreten eines Fehlers erster bzw. zweiter Art bezeichnen wir als **Fehlerwahrscheinlichkeiten erster** bzw. **zweiter Art**.

Genauer: Im Beispiel oben (teste $H_0 : \mathbf{E}X_1 = 0$ versus $H_1 : \mathbf{E}X_1 \neq 0$) ist die **Fehlerwahrscheinlichkeit erster Art** eines Tests ϕ gegeben durch

$$\mathbf{P}_{\mathbf{E}X_1=0}\left[\phi(X_1, \ldots, X_n) = 1\right],$$

während die **Fehlerwahrscheinlichkeiten zweiter Art** gegeben sind durch

$$\mathbf{P}_{\mathbf{E}X_1=\mu}\left[\phi(X_1, \ldots, X_n) = 0\right] \qquad \textbf{mit } \mu \in \mathbb{R} \setminus \{0\} \ .$$

Wünschenswert: Konstruiere einen statistischen Tests, bei dem sowohl die Fehlerwahrscheinlichkeiten erster Art als auch die Fehlerwahrscheinlichkeiten zweiter Art kleiner als bei allen anderen Tests sind.

Problem: So ein Test existiert im Allgemeinen nicht ...

Ausweg: Asymmetrische Betrachtungsweise der Fehlerwahrscheinlichkeiten erster und zweiter Art:

Gebe Schranke für die Fehlerwahrscheinlichkeiten erster Art vor und verwende dann einen Test, der diese Schranke erfüllt und der bzgl. allen anderen Tests, die diese Schranke erfüllen, hinsichtlich der Fehlerwahrscheinlichkeiten zweiter Art optimal ist.

Die Optimalität der Tests werde wir in dieser Vorlesung nicht beweisen, aber die Schranke für die Fehlerwahrscheinlichkeiten erster Art formalisieren wir in

Definition. Ein Test ϕ heißt **Test zum Niveau** α (mit $\alpha \in [0,1]$ vorgegeben), wenn alle Fehlerwahrscheinlichkeiten erster Art von ϕ kleiner oder gleich α sind.

Achtung:

- Bei einem Test zum Niveau α kontrollieren wir nur die Wahrscheinlichkeit des Auftretens von Fehlern erster Art.

- Wie groß die Wahrscheinlichkeit des Auftretens von Fehlern zweiter Art ist, hängt beim optimalen Test von der Stichprobengröße ab (und wird meist nicht kontrolliert).

- Eine wiederholte Durchführung eines Tests zum Niveau α mit unabhängig erzeugten Daten für die gleiche Fragestellung wird zwangsläufig irgendwann zur Ablehnung von H_0 führen und ist daher nicht zulässig (**Problem des iterierten Testens**).

- In der Praxis gibt man häufig das minimale Niveau an, dass beim vorliegenden Datensatz und einem festen Test zur Ablehung von H_0 führt (sog. p-**Wert**). **Das ist aber nicht die Wahrscheinlichkeit für die Gültigkeit von** H_0.

5.3.4 Der Gauß-Test

1. Fragestellungen beim Gauß-Test für eine Stichprobe
   Geg.: Realisierungen $x_1, \ldots, x_n$ von unabhängigen identisch $N(\mu, \sigma_0^2)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$.

   (a) Beim einseitigen Gauß-Test für eine Stichprobe ist ein $\mu_0 \in \mathbb{R}$ gegeben und wir möchten zu gegebenem Niveau $\alpha \in (0, 1)$ die Hypothesen

   $$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

   testen.

   (b) Beim zweiseitigen Gauß-Test für eine Stichprobe ist ein $\mu_0 \in \mathbb{R}$ gegeben und wir möchten zu gegebenem Niveau $\alpha \in (0, 1)$ die Hypothesen

   $$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

   testen.

2. Fragestellungen beim Gauß-Test für zwei Stichproben

Geg.: Realisierungen $x_1, \ldots, x_n, y_1, \ldots, y_m$ von unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, wobei $X_1, \ldots, X_n$ identisch $N(\mu_X, \sigma_0^2)$-verteilt und $Y_1, \ldots, Y_m$ identisch $N(\mu_Y, \sigma_0^2)$-verteilt sind, mit unbekannten $\mu_X, \mu_Y \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$.

(a) Beim einseitigen Gauß-Test für zwei Stichproben möchten wir zu gegebenem Niveau $\alpha \in (0,1)$ die Hypothesen

$$H_0 : \mu_X \leq \mu_Y \quad \text{versus} \quad H_1 : \mu_X > \mu_Y$$

testen.

(b) Beim zweiseitigen Gauß-Test für zwei Stichproben möchten wir zu gegebenem Niveau $\alpha \in (0,1)$ die Hypothesen

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

testen.

3. Grundidee beim Gauß-Test für eine Stichprobe

(a) Wir betrachten

$$\frac{1}{n}\sum_{i=1}^{n} X_i,$$

was ein Schätzer von $\mathbf{E}X_1 = \mu$ ist.

(b) Also ist es naheliegend, $H_0 : \mu \leq \mu_0$ (bzw. $H_0 : \mu = \mu_0$) abzulehnen, falls $\frac{1}{n}\sum_{i=1}^{n} X_i$ "sehr viel größer" als μ_0 (bzw. "weit entfernt" von μ_0) ist.

(c) Um das Niveau einzuhalten, verwenden wir, dass Linearkombinationen unabhängiger normalverteilter Zufallsvariablen selbst normalverteilt sind, und dass daher für $\mu = \mu_0$

$$\frac{\sqrt{n}}{\sigma_0} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_0 \right)$$

$N(0,1)$-verteilt ist.

4. Einseitiger Gauß-Test für eine Stichprobe

geg.: Realisierungen $x_1, \ldots, x_n$ von unabhängigen identisch $N(\mu, \sigma_0^2)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

H_0 wird abgelehnt, falls

$$\frac{\sqrt{n}}{\sigma_0} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) > u_\alpha$$

ist, wobei u_α das sogenannte $\alpha - Fraktil$ von $N(0, 1)$ ist, d.h. u_α wird so bestimmt, dass für eine $N(0, 1)$-verteilte Zufallsvariable Z gilt: $\mathbf{P}[Z > u_\alpha] = \alpha$.

5. Zweiseitiger Gauß-Test für eine Stichprobe

geg.: Realisierungen $x_1, \ldots, x_n$ von unabhängigen identisch $N(\mu, \sigma_0^2)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

H_0 wird abgelehnt, falls

$$\left| \frac{\sqrt{n}}{\sigma_0} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) \right| > u_{\alpha/2}$$

ist, wobei $u_{\alpha/2}$ das sogenannte $\alpha/2 - Fraktil$ von $N(0, 1)$ ist, d.h. $u_{\alpha/2}$ wird so bestimmt, dass für eine $N(0, 1)$-verteilte Zufallsvariable Z gilt:
$\mathbf{P}[Z > u_{\alpha/2}] = \alpha/2$ (was $\mathbf{P}[|Z| > u_{\alpha/2}] = \alpha$ impliziert).

6. Grundidee beim Gauß-Test für zwei Stichproben

(a) Wir betrachten $\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j$, was ein Schätzer von $\mathbf{E}X_1 - \mathbf{E}Y_1 = \mu_X - \mu_Y$ ist.

(b) Also ist es naheliegend, $H_0 : \mu_X \leq \mu_Y$ (bzw. $H_0 : \mu_X = \mu_Y$) abzulehnen, falls $\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j$ "sehr viel größer" als 0 (bzw. "weit entfernt" von 0) ist.

(c) Um das Niveau einzuhalten, beachten wir, dass für $\mu_X = \mu_Y$ analog oben

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}\cdot\sigma_0}\cdot\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_i\right)$$

$N(0,1)$-verteilt ist.

7. Einseitiger Gauß-Test für zwei Stichproben

geg.: Realisierungen $x_1, \ldots, x_n, y_1, \ldots, y_m$ von unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, wobei $X_1, \ldots, X_n$ identisch $N(\mu_X, \sigma_0^2)$-verteilt und $Y_1, \ldots, Y_m$ identisch $N(\mu_Y, \sigma_0^2)$-verteilt sind, mit unbekannten $\mu_X, \mu_Y \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu_X \leq \mu_Y \quad \text{versus} \quad H_1 : \mu_X > \mu_Y.$$

H_0 wird abgelehnt, falls

$$\frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}} \cdot \sigma_0} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{m} \sum_{j=1}^{m} y_j \right) > u_\alpha$$

ist, wobei u_α das $\alpha - Fraktil$ von $N(0, 1)$ ist.

8. Zweiseitiger Gauß-Test für zwei Stichproben

geg.: Realisierungen $x_1, \ldots, x_n, y_1, \ldots, y_m$ von unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, wobei $X_1, \ldots, X_n$ identisch $N(\mu_X, \sigma_0^2)$-verteilt und $Y_1, \ldots, Y_m$ identisch $N(\mu_Y, \sigma_0^2)$-verteilt sind, mit unbekannten $\mu_X, \mu_Y \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y.$$

H_0 wird abgelehnt, falls

$$\left| \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}} \cdot \sigma_0} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{m} \sum_{j=1}^{m} y_j \right) \right| > u_{\alpha/2}$$

ist, wobei $u_{\alpha/2}$ das $\alpha/2 - Fraktil$ von $N(0, 1)$ ist.

5.3.5 Der t-Test von Student

Problem beim Gauß-Test: Varianz σ_0^2 wird in Anwendungen nie bekannt sein.

Ausweg: Wir schätzen die Varianz aus unseren Daten.

Einfach, bei Test für eine Stichprobe:

Sind $X_1, \ldots, X_n$ unabhängig identisch verteilt, so ist

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^2$$

eine erwartungstreue und stark konsistente Schätzung von $V(X_1)$.

Zur Einhaltung des Niveaus beachten wir:

Sind $X_1, \ldots, X_n$ unabhängig $N(\mu_0, \sigma^2)$-verteilt, so ist

$$\frac{\sqrt{n}}{S_X} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu_0 \right)$$

nicht länger $N(0,1)$-verteilt, sondern **t-verteilt mit $n-1$-Freiheitsgraden.**

Daher verwenden wir bei den Tests jetzt Fraktile der sogenannten t-Verteilung!

Beispiel: Zweiseitiger t-Test für eine Stichprobe

geg.: Realisierungen $x_1, \ldots, x_n$ von unabhängigen identisch $N(\mu, \sigma^2)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit unbekanntem $\mu \in \mathbb{R}$ und **unbekanntem** $\sigma^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

H_0 wird abgelehnt, falls

$$\left| \frac{\sqrt{n}}{s_x} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) \right| > t_{n-1,\alpha/2}$$

ist, wobei $t_{n-1,\alpha/2}$ das sogenannte $\alpha/2 - Fraktil$ der t_{n-1}-Verteilung ist, d.h. $t_{n-1,\alpha/2}$ wird so bestimmt, dass für eine t_{n-1}-verteilte Zufallsvariable Z gilt: $\mathbf{P}[Z > t_{n-1,\alpha/2}] = \alpha/2$.

Anwendung im Beispiel zu Einschätzung der Leistungsfähigkeit:

$n = 15$ Kandidaten wurde eine Klausur mit 70 Aufgaben gestellt. Nach Bearbeitung der Klausur wurden Sie gebeten, die Anzahl der richtig gelösten Aufgaben zu schätzen. Nach der Korrektur der Klausur wurden die Differenzen

$$X_i = \text{Gesch. Anz. gelöster Aufgaben } - \text{Tatsächliche Anz. gelöster Aufgaben}$$

gebildet.

Beschreibung der gemessenen Daten: $n = 15$, $\bar{x} = -6.4$, $s^2 = 61.7$

Wir führen einen zweiseitigen t-Tests für $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ zum Niveau $\alpha = 0.05$ durch.

Hierbei gilt: $t_{n-1,\alpha} = t_{14,0.05/2} \approx 2.14$

Wir erhalten

$$\frac{\sqrt{n}}{s_x} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) = \frac{\sqrt{15}}{\sqrt{61.7}} \cdot |-6.4 - 0| \approx 3.16 > t_{14,0.025},$$

so dass H_0 zum Niveau $\alpha = 0.05$ abgelehnt werden kann.

Resultat: Examenskadidaten können ihre eigene Leistungsfähigkeit nicht richtig einschätzen.

t-Test für zwei Stichproben

geg.: Realisierungen $x_1, \ldots, x_n, y_1, \ldots, y_m$ von unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, wobei $X_1, \ldots, X_n$ identisch $N(\mu_X, \sigma^2)$-verteilt und $Y_1, \ldots, Y_m$ identisch $N(\mu_Y, \sigma^2)$-verteilt sind, mit unbekannten $\mu_X, \mu_Y \in \mathbb{R}$, unbekanntem $\sigma^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu_X \leq \mu_Y \quad \text{versus} \quad H_1 : \mu_X > \mu_Y$$

bzw.

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

Problem: Wie schätzen wir diesmal die Varianz ?

Schätzung der Varianz:

Wir verwenden die sogenannte gepoolte Stichprobenvarianz

$$\begin{aligned} S^2_{X,Y} &= \frac{\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2 + \sum_{i=1}^{m}(Y_i - \frac{1}{n}\sum_{j=1}^{m} Y_j)^2}{n+m-2} \\ &= \frac{(n-1)s^2_X + (m-1)s^2_Y}{n+m-2}, \end{aligned}$$

Unter den obigen Voraussetzungen und bei Gültigkeit von $\mu_X = \mu_Y$ ist jetzt

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}} \cdot S_{X,Y}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_i \right)$$

t-verteilt mit $n+m-2$-Freiheitsgraden.

Beispiel: Zweiseitiger t-Test für zwei Stichproben

geg.: Realisierungen $x_1, \ldots, x_n, y_1, \ldots, y_m$ von unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, wobei $X_1, \ldots, X_n$ identisch $N(\mu_X, \sigma_0^2)$-verteilt und $Y_1, \ldots, Y_m$ identisch $N(\mu_Y, \sigma_0^2)$-verteilt sind, mit unbekannten $\mu_X, \mu_Y \in \mathbb{R}$ und bekanntem $\sigma_0^2 > 0$ und $\alpha \in (0, 1)$.

Zu testen sei

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y.$$

H_0 wird abgelehnt, falls

$$\left| \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}} \cdot s_{x,y}} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{m} \sum_{j=1}^{m} y_j \right) \right| > t_{n+m-2,\alpha/2}$$

ist, wobei $t_{n+m-2,\alpha/2}$ das $\alpha/2 - Fraktil$ von t_{n+m-2} ist.

Anwendung bei den Anzahlen gesprochener Wörter pro Tag:

Unterscheidet sich die Anzahl der gesprochenen Wörter pro Tag bei Frauen (x) von der bei Männern (y) ?

Beschreibung der gemessenen Daten:

- $n_x = 210$, $\bar{x} = 16215$, $s_x = 7301$
- $n_y = 186$, $\bar{y} = 15669$, $s_y = 8663$

Wir führen einen zweiseitigen t-Tests für zwei Stichproben für $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$ zum Niveau $\alpha = 0.05$ durch.

Hierbei gilt: $t_{n_x+n_y-2,\alpha} = t_{210+186-2,0.05/2} = t_{394,0.05/2} \approx 1.97$

Für die beobachteten Daten erhalten wir

$$\frac{\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{j=1}^{m} y_j\right|}{\sqrt{\frac{1}{n}+\frac{1}{m}}\cdot s_{x,y}}$$

$$= \frac{|16215 - 15669|}{\sqrt{\frac{1}{210}+\frac{1}{186}}\cdot\sqrt{\frac{1}{210+186-2}\cdot((210-1)\cdot 7301^2 + (186-1)\cdot 8663^2)}}$$

$$\approx 0.68 < t_{394,0.05/2},$$

so dass H_0 zum Niveau $\alpha = 0.05$ nicht abgelehnt werden kann.

Resultat: Der t-Test zum Niveau $\alpha = 0.05$ führt nicht darauf, dass sich die Anzahl der gesprochenen Wörter pro Tag bei Studentinnen und bei Studenten unterscheiden.