Statistik I für Human- und SozialwissenschaftlerInnen

Vorlesung WS 2009/10

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, Frühjahr 2009

Ergebnis der schriftlichen Prüfung zur Vorlesung "Statistik I für Human- und Sozialwissenschaftler" am 10.03.2009:

Anzahl Teilnehmer	:	341
Notendurchschnitt	:	2,71
Durchfallquote	:	6,74 %

StudentInnen hatten die Möglichkeit, freiwillig durch regelmäßige Mitarbeit bei den Übungen einen Bonus für die Klausur (ca. $0,3$ Notenpunkte) zu erwerben.

Anzahl Teilnehmer mit Bonus	:	287
Notendurchschnitt	:	2,59
Durchfallquote	:	5,23 %

Anzahl Teilnehmer ohne Bonus	:	54
Notendurchschnitt	:	3,35
Durchfallquote	:	15,4 %

Was folgt daraus hinsichtlich des Einflusses der regelmäßigen Teilnahme an den Übungen

- 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:
(vgl. Video, Quelle: Homepage des Kennedy Space Centers der NASA)

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.

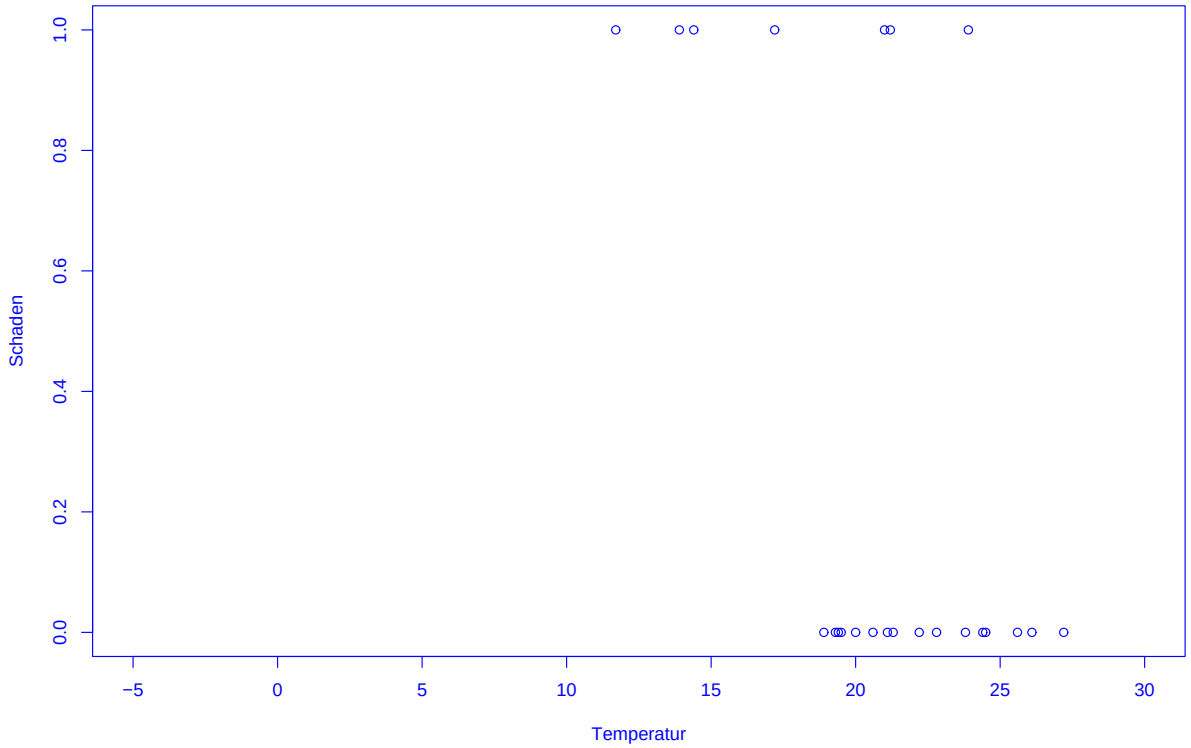
Zur Begründung verwendete Daten:

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

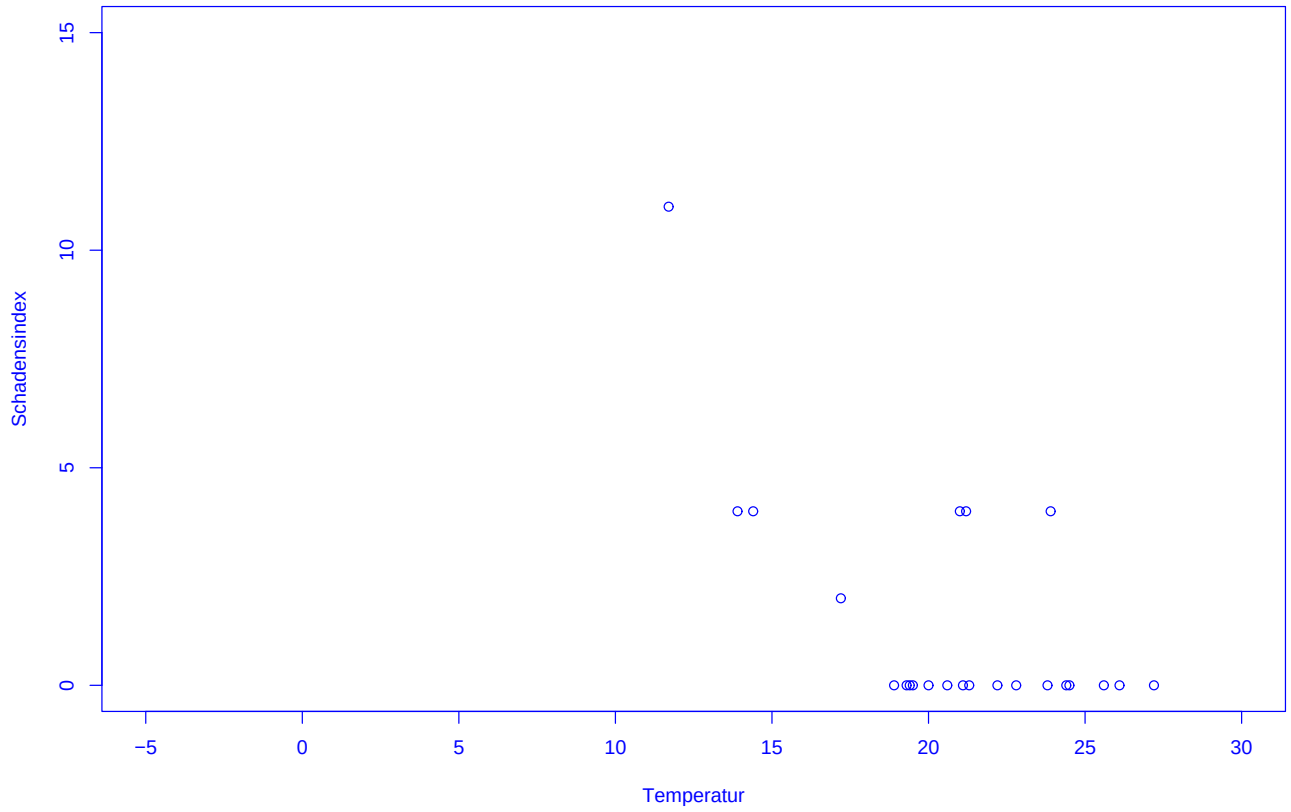
War für NASA leider nicht nachvollziehbar ...

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 . . .

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 ?

- 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 oder Pädagogik an der TU 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 **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 (1,5V)
- Kapitel 3: Beschreibende Statistik (2,5V)
- Kapitel 4: Einführung in die Wahrscheinlichkeitstheorie (5,5V)
- Kapitel 5: Schließende Statistik (3,5V)

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.

Das Erreichen der letzten Ebene ist in der Statistik wichtig, denn:

1. In der Statistik analysieren Sie Daten, die gewisse Unsicherheiten ($\approx$ Zufall) enthalten, mit Hilfe von mathematischen Modellen des Zufalls.

2. Das Anwenden eines statistischen Verfahrens entspricht dann dem Schluss innerhalb eines mathematischen Modells.

3. Damit Sie das Ergebnis auf die Realität übertragen können, muss aber das mathematische Modell zur Realität passen.

4. Das können Sie nur dann beurteilen, wenn Sie das mathematische Modell verstanden haben . . .

Dazu ist in Statistik ein gewisses abstraktes Verständnis der Verfahren 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 Wahrscheinlichkeitstheorie und Statistik **lernen können**.

Spezielle “Tricks” dabei:

- Formulieren von Lernzielen zu Beginn
- Minitest in der Mitte
- Zusammenfassung am Schluss
- Intensiver Übungsbetrieb
- Begleitendes Buch
- Vorlesungsaufzeichnung im Rahmen von E-Learning

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!

TERMINE

1. Vorlesung: Dienstag, 8:00 Uhr - 9:30 Uhr, in S 101_A01

2. Vortragsübungen: Montag, 16:15 Uhr - 17:55 Uhr, in S 103_221

   Die Vortragsübungen beginnen am **26.10.2009**.

3. Übungen:

   Die Übungen finden zu verschiedenen Terminen in Kleingruppen statt (Dauer 2 Stunden, wöchentlich ab 27.10.2009).

Begleitendes Buch zur Vorlesung:

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

In der Vorlesung wird der darin enthaltene Stoff unter Ausblendung der mathematischen Details behandelt.

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, die dieses Buch im zweiten Semester verwenden werden (enthält aber auch Stoff aus dieser Vorlesung), ca. 50 Euro.

Lernziele der Vorlesung am 27.10.2009

Nach dieser Vorlesung sollten Sie

1. verstanden haben, dass die Art, wie Daten entstehen, die möglichen Rückschlüsse aus den Daten beeinflusst,

2. wissen, dass nur bei prospektiv kontrollierten Studien mit Randomisierung ein Rückschluss auf kausale Zusammenhänge möglich ist,

3. erklären können, warum alle anderen Studien durch sogenannte konfundierende Faktoren verfälscht werden können.

Kapitel 2: Erhebung von Daten

Wie Daten entstehen bestimmt mit, welche Schlüsse man später daraus ziehen kann (bzgl. Verallgemeinerungen von Aussagen über den vorliegenden 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 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 abklingt.

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

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). Im Rahmen der Studie wird Einfluss auf die Versuchsobjekte genommen.

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

Aber:

Die Erkennung aller konfundierenden Faktoren ist meistens nicht möglich, weshalb **Beobachtungsstudien** (und ebenso retrospektiv kontrollierte Studien bzw. prospektiv kontrollierte Studien ohne Randomisierung) zwar zum **Aufstellen von Hypothesen** nützlich sind, aber **keine kausalen Zusammenhänge** nachweisen können.

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	10.268	10.268
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	$2.306\ (22,5\%)$	$2.312\ (22,5\%)$

Beispiel: Hat eine mediterrane Diät einen positiven Einfluss auf Herz-Kreislauf-Krankheiten ?

Eine Reihe von Beobachtungsstudien führte zu der Hypothese, dass eine mediterrane Diät einen positiven Einfluss auf Herz-Kreislauf-Krankheiten hat.

Im Rahmen einer prospektiv kontrollierten Studie mit Randomisierung wurden 1000 Hochrisikopatienten zufällig in Studien- und Kontrollgruppe unterteilt. Der Studiengruppe wurde eine mediterrane Diät empfohlen, die Kontrollgruppe erhielt die übliche Diätempfehlungen.

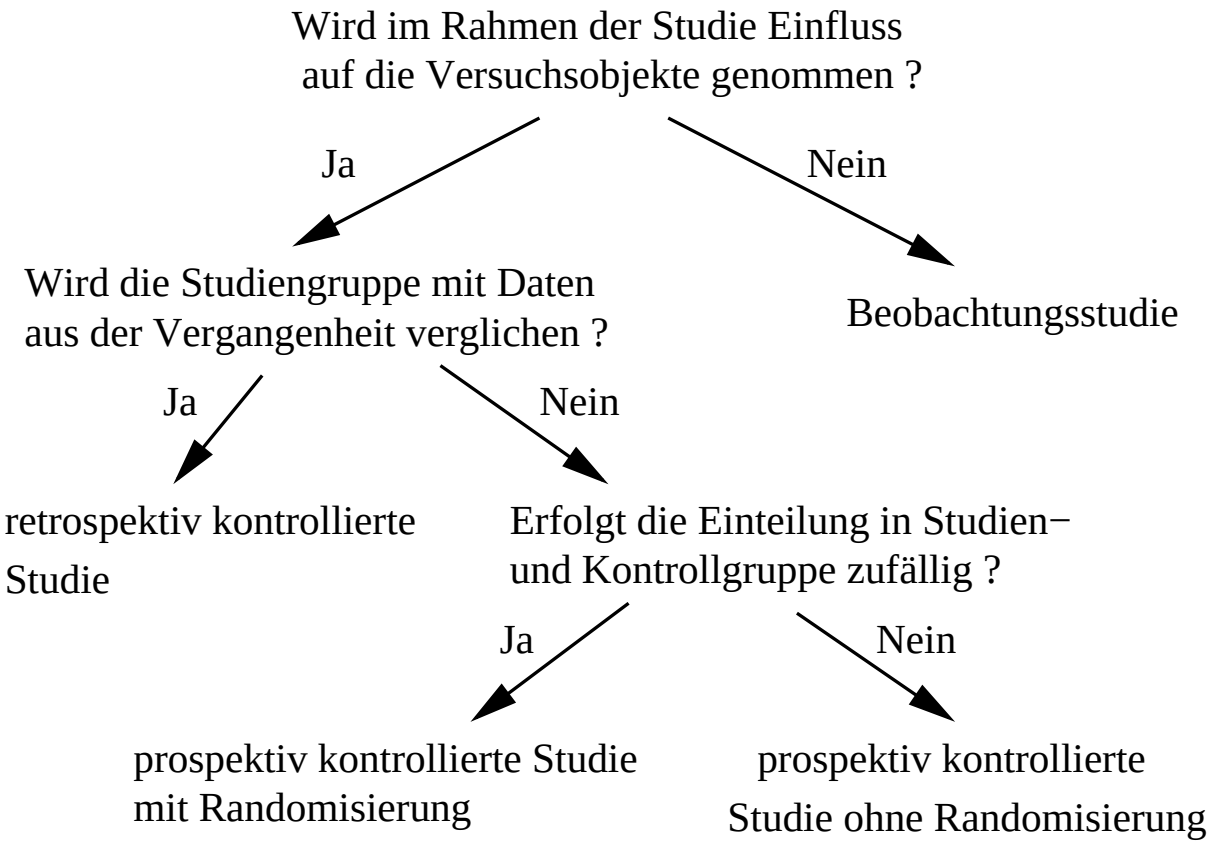
Nach zwei Jahren wurden beide Gruppen hinsichtlich neu aufgetretener Herz-Kreislauf-Krankheitsfälle verglichen.

Resultat:

	Gesamt	Studiengruppe	Kontrollgruppe
Alle	$1.000\ (100\%)$	$499\ (40,9\%)$	$501\ (50,1\%)$
Nicht tödlich verlaufende Myokardinfarkte	$63\ (6,3\%)$	$21\ (4,2\%)$	$43\ (8,6\%)$
Tödlich verlaufende Myokardinfarkte	$29\ (2,9\%)$	$12\ (2,4\%)$	$17\ (3,4\%)$
plötzlicher Herztod	$22\ (2,2\%)$	$6\ (1,2\%)$	$16\ (3,2\%)$

Da in der Studiengruppe weniger Herz-Kreislauf-Krankheitsfälle auftraten als in der Kontrollgruppe kann man davon ausgehen, dass die mediterrane Diät in der Tat einen positiven Einfluss auf Herz-Kreislauf-Krankheiten hat.

Übersicht über die verschiedenen Arten von Studien:



Zusammenfassung der Vorlesung am 27.10.2009

1. Bei einer Studie wird eine sogenannte Studiengruppe mit einer sogenannten Kontrollgruppe verglichen.

2. Im Rahmen von kontrollierten Studien wird Einfluss auf die Versuchsobjekte genommen, während diese bei Beobachtungsstudien nur beobachtet werden.

3. Nur mit Hilfe von prospektiv kontrollierten Studien mit Randomisierung kann auf kausale Zusammenhänge zurückgeschlossen werden.

4. Bei allen anderen Studien kann das Ergebnis durch sogenannte konfundierende Faktoren verfälscht werden, die gleichzeitig Einfluss auf die Einteilung der Versuchsobjekte in Studien- und Kontrollgruppe und auf das beobachtete Resultat haben.

Lernziele der Vorlesung am 03.11.2009

Nach dieser Vorlesung sollten Sie verstanden haben,

1. inwiefern Ergebnisse von Umfragen durch den sogenannten sampling bias und den sogenannten non-response bias verfälscht werden können,

2. wie Daten graphisch mit Hilfe eines Säulendiagramms und eines Histogramms dargestellt werden.

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 vor der Bundestagswahl am 27.09.2009:

	SPD	CDU/CSU	FDP	GRÜNE	DIE LINKE
Allensbach (22.09.09)	24,0	35,0	13,5	11,0	11,5
TNS Emnid (17.09.09)	25	35	13	11	12
Forsa (25.09.09)	25	33	14	10	12
Forschungsgruppe Wahlen (18.09.09)	25	36	13	10	11
Infratest-dimap (17.09.09)	26	35	14	10	11
amtliches Endergebnis	23,0	33,8	14,6	10,7	11,9

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 TNS 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.

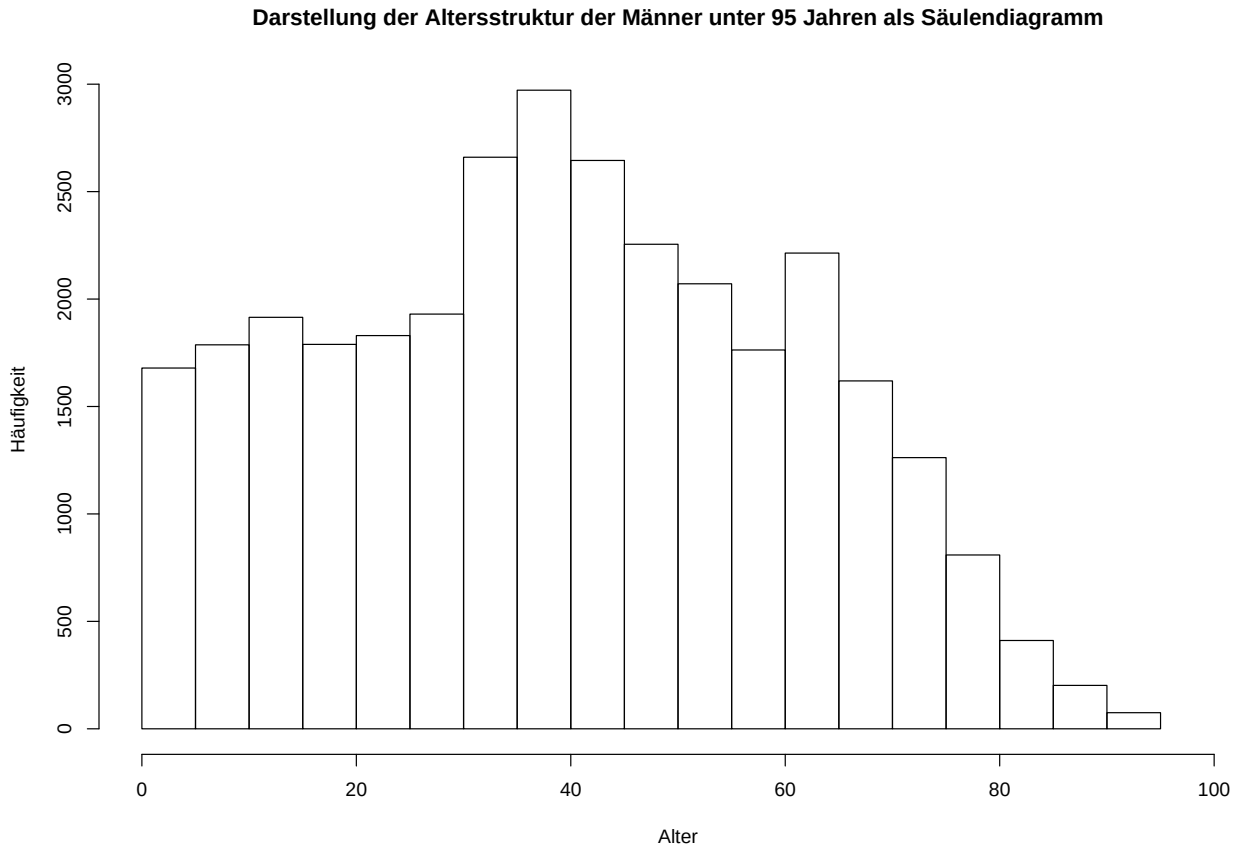
Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

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

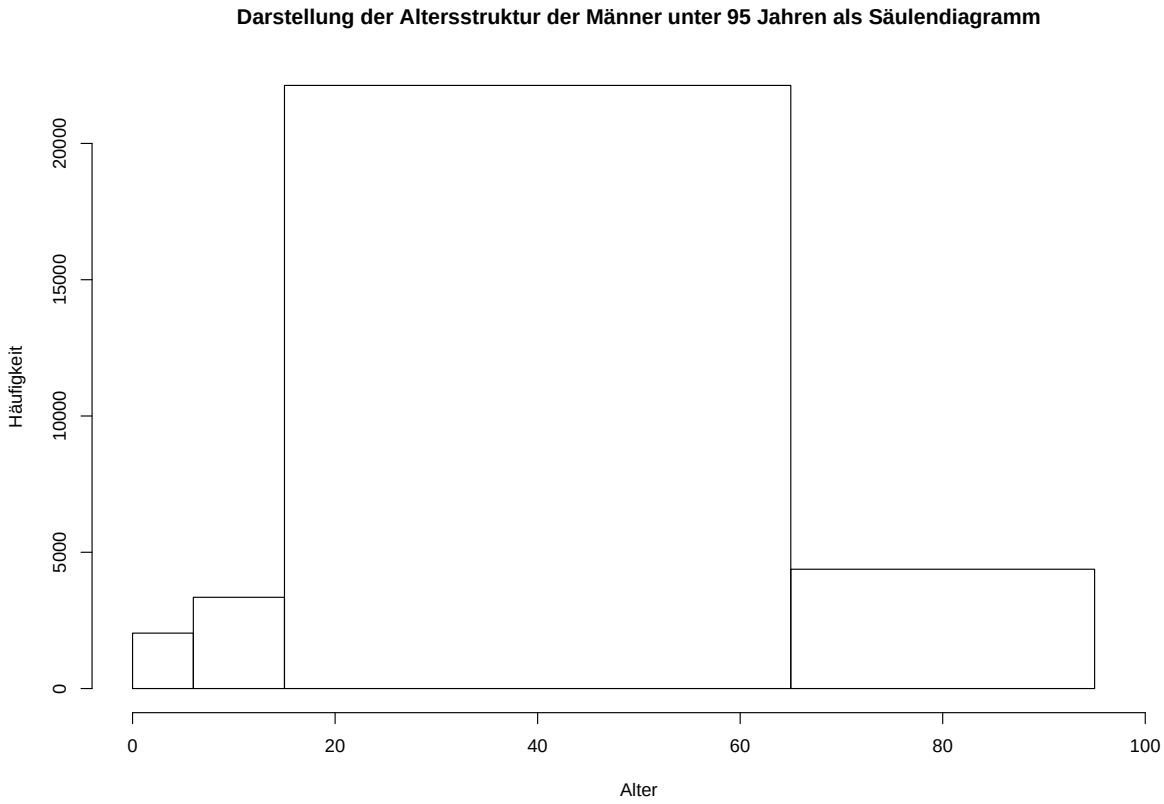
Unterteilung in 19 Klassen ergibt

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:

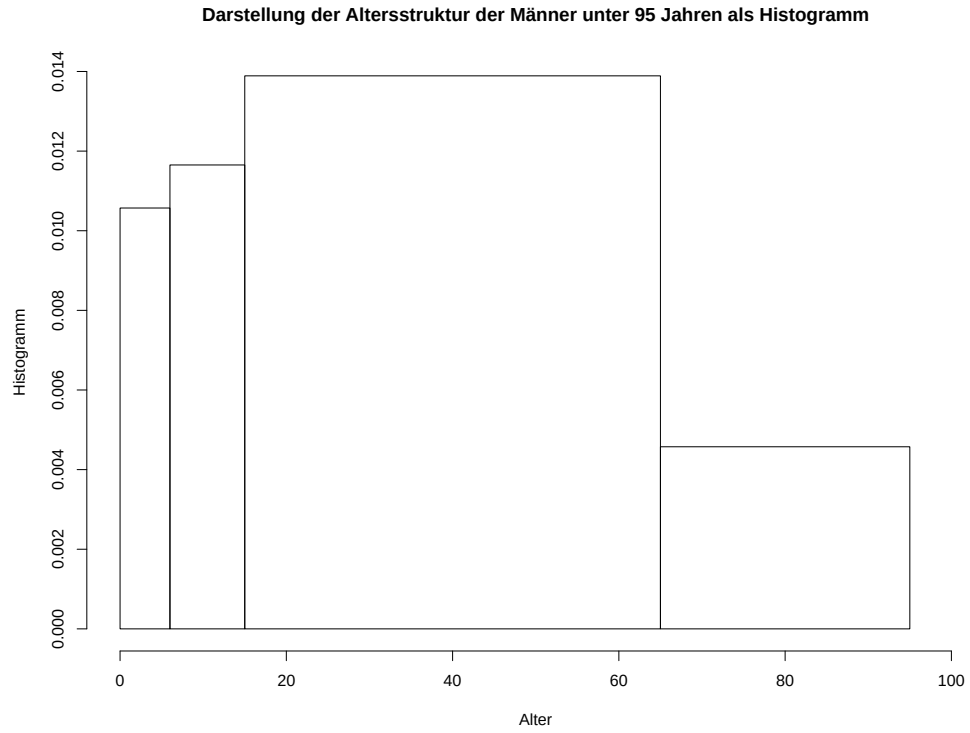
- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.
- Trage über I_j den Wert

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

  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



Zusammenfassung der Vorlesung am 03.11.2009

1. Bei einer Umfrage versteht man unter dem sogenannten sampling bias, dass gewisse Untergruppen, deren Antwortverhalten von der Allgemeinheit abweicht, in der Stichprobe zu häufig vorkommen und daher die Resultate verzerrt werden. Ein sogenannter non-response bias führt zu einer Verfälschung der Ergebnisse, indem Teile der Befragten, deren Antwortverhalten vom Rest abweicht, die Teilnahme an der Umfrage verweigern.

2. Bei der graphischen Darstellung eines Datensatzes in einem Säulendiagramm (bzw. Histogramm) wird über jedem zugrundeliegenden Intervall ein Balken gezeichnet, dessen Höhe (bzw. Flächeninhalt) gleich dem Anzahl der Datenpunkte (bzw. dem prozentualen Anteil der Datenpunkte) in diesem Intervall ist.

Lernziele der Vorlesung am 10.11.2009

Nach dieser Vorlesung sollten Sie

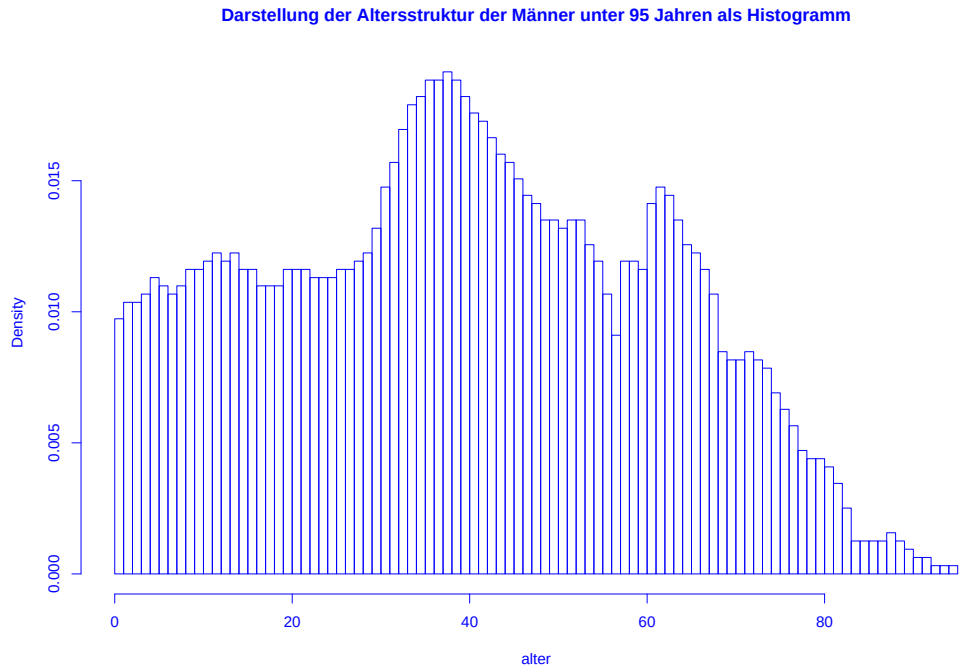
1. verstanden haben, was man unter einer Dichte versteht und was es anschaulich bedeutet, dass diese eine Datenmenge beschreibt,

2. die wichtigsten statistischen Maßzahlen sowie Boxplots kennen.

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.

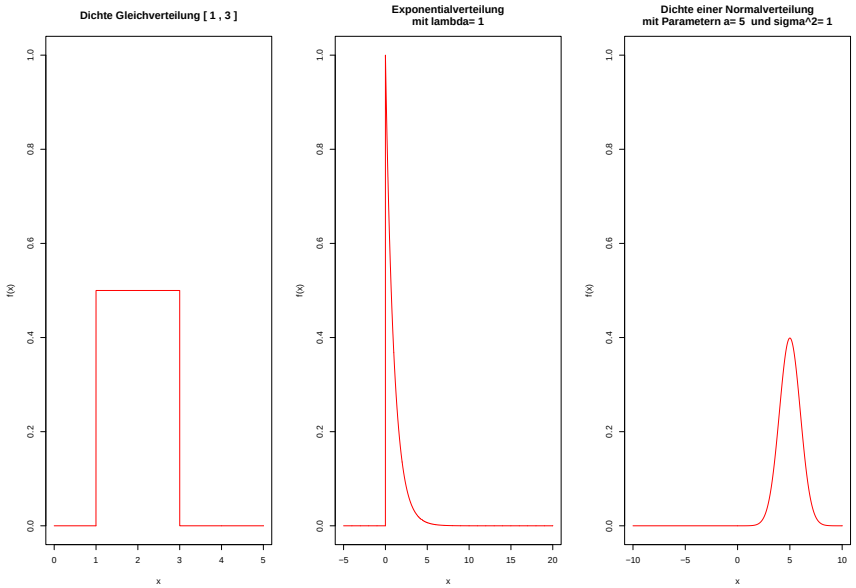
Definition: *Eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *mit*

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

und

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

heißt Dichte.



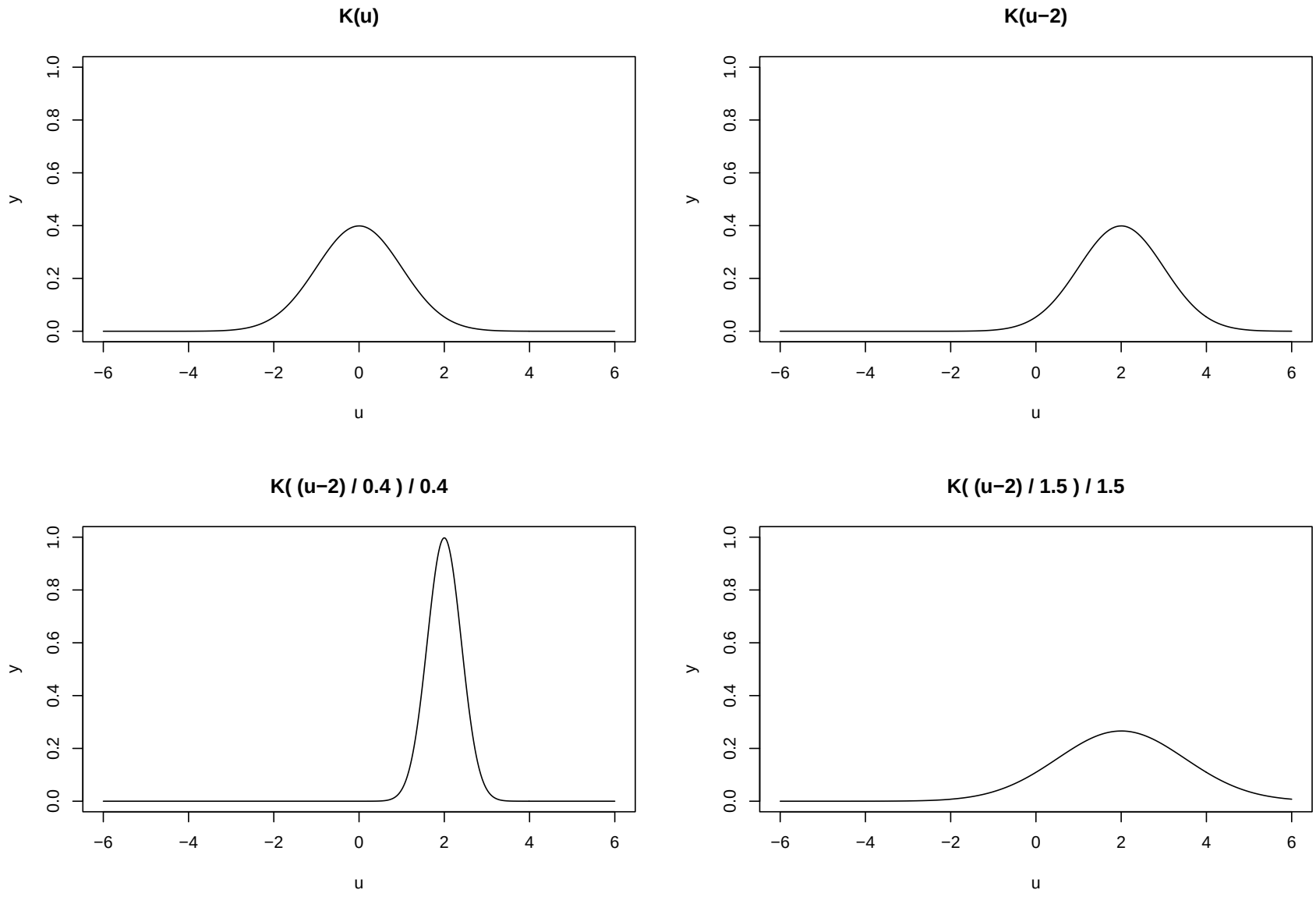
Anpassung einer glatten Dichtefunktion an Daten mit Hilfe des sogenannten Kerndichteschätzers:

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

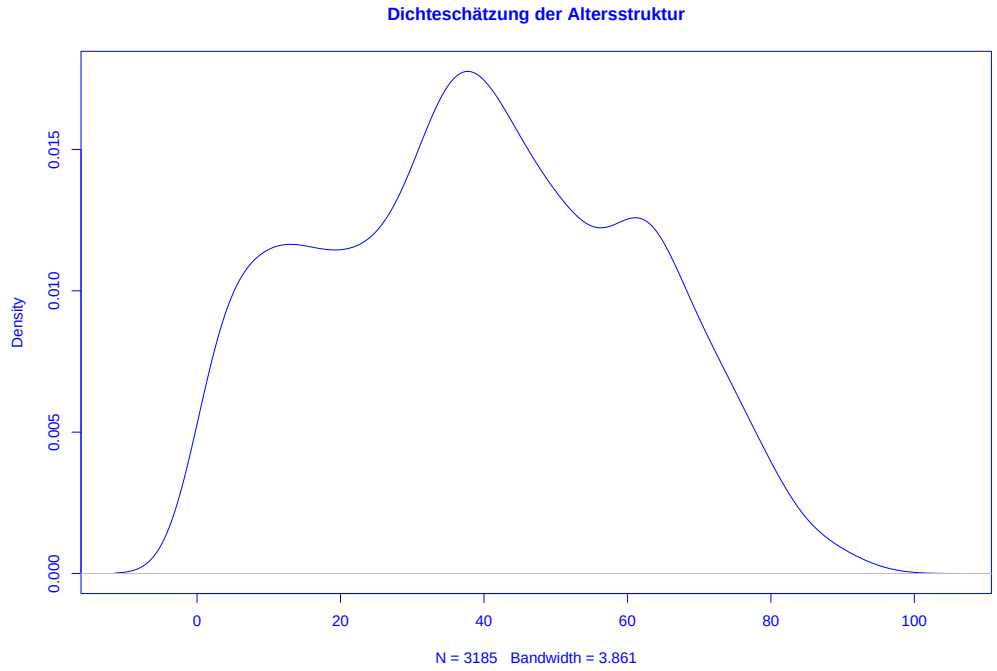
mit Parameter $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**), z.B.

$$K(v) = \frac{1}{\sqrt{2\pi}}\exp\left(-v^2/2\right) \quad \text{(sog. Gauß-Kern).}$$

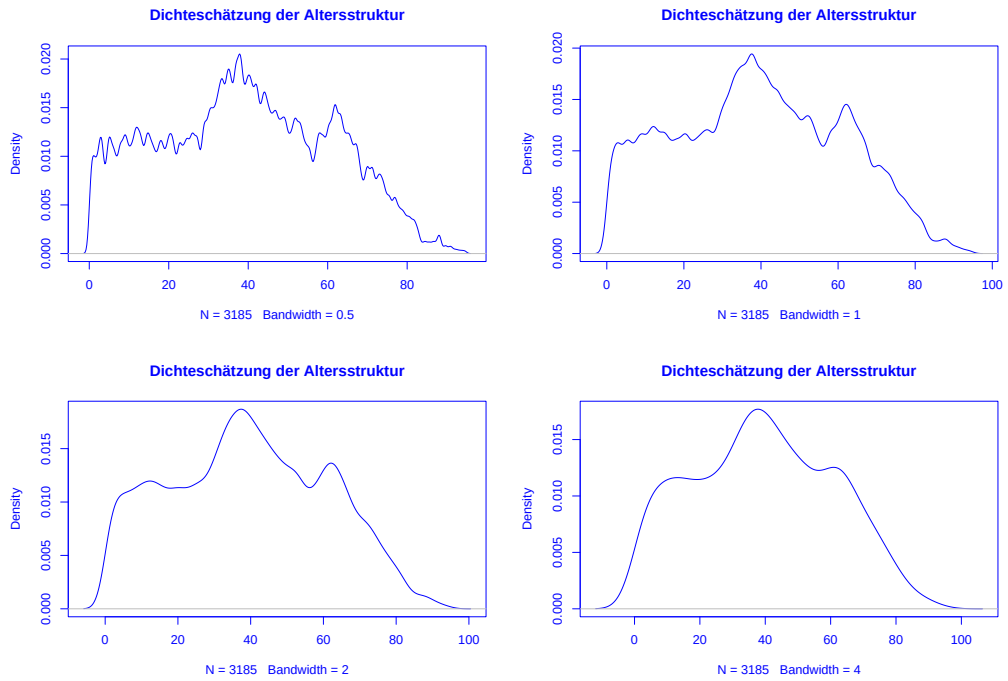
Deutung: Mittelung von Dichtefunktionen, die um die einzelnen Datenpunkte konzentriert sind.


K(u)
K(u−2)
K((u−2) / 0.4) / 0.4
K((u−2) / 1.5) / 1.5
y
u
0.0
0.2
0.4
0.6
0.8
1.0
−6
−4
−2
0
2
4
6

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:

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

Vorfaktor $1/(n-1)$ statt $1/n$, da $x_1 - \bar{x}, \ldots, x_n - \bar{x}$ nur $n-1$ Freiheitsgrade hat.

Denn:

$$x_1 - \bar{x} + \cdots + x_n - \bar{x} = x_1 + \cdots + x_n - n \cdot \bar{x} = 0.$$

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 Interquartilsabstand

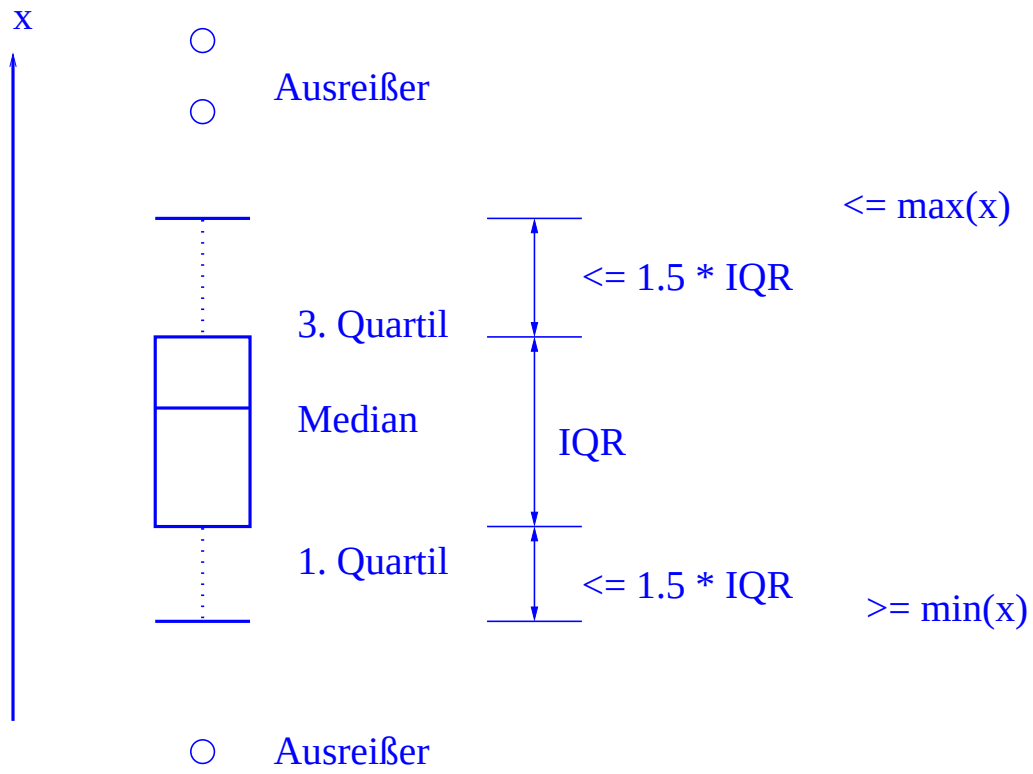
$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

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:

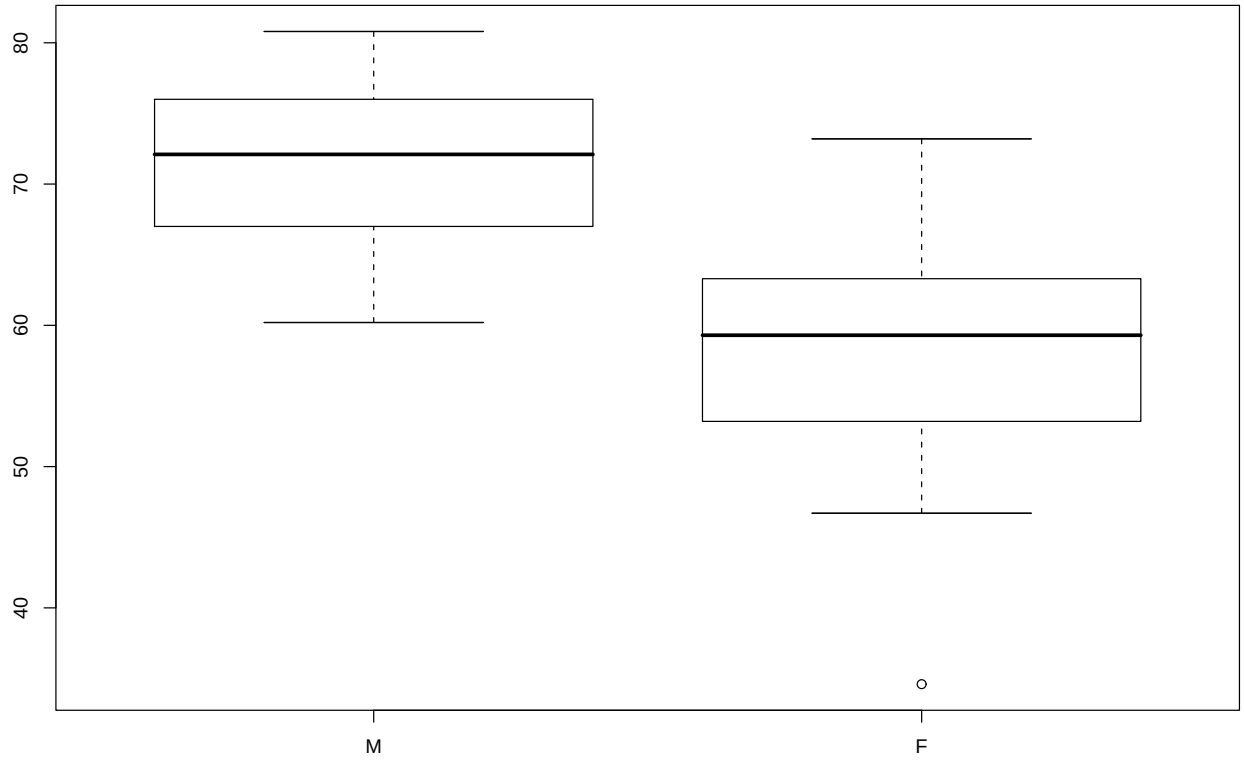
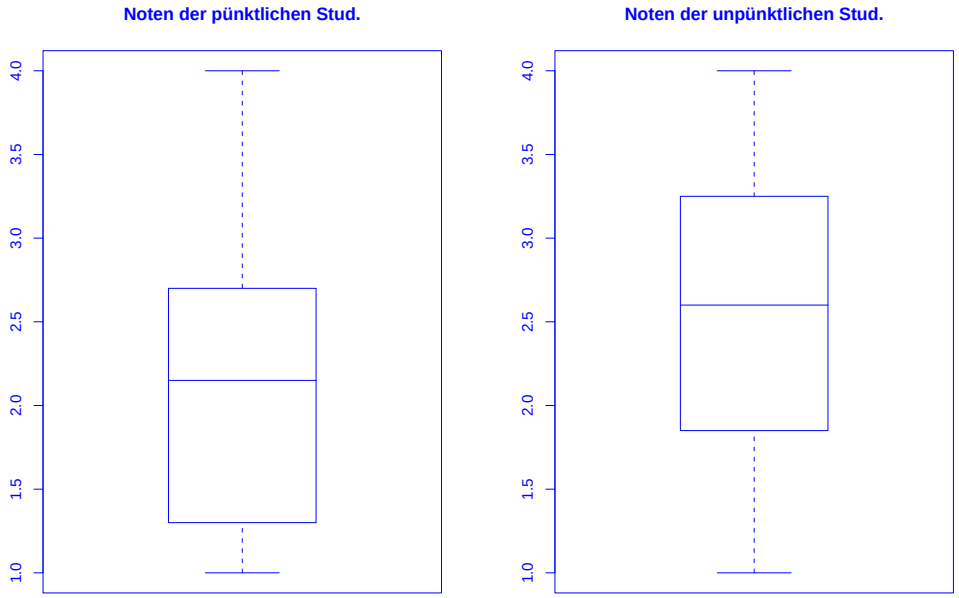
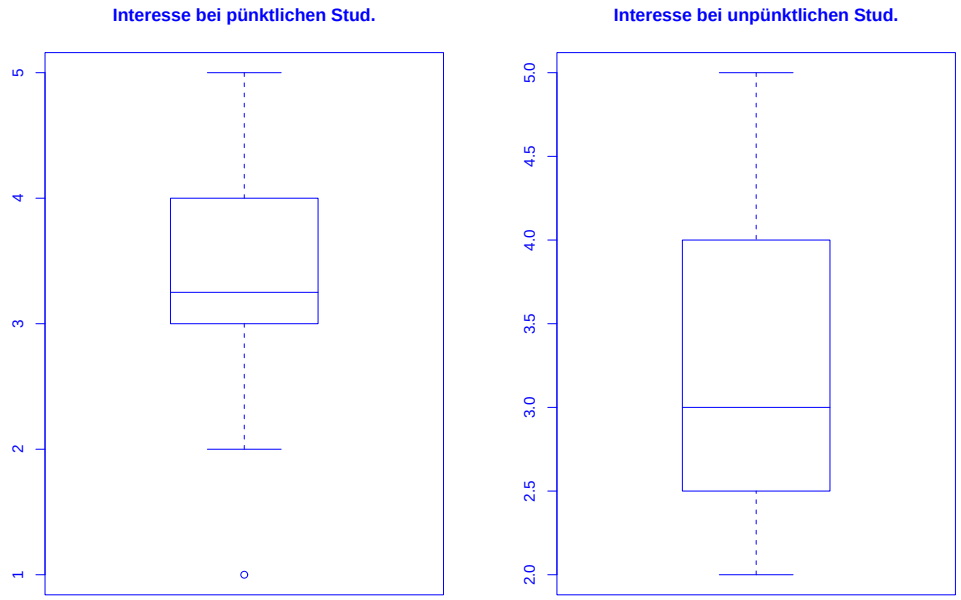


Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:




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

Zusammenfassung der Vorlesung am 10.11.2009

1. Eine Dichte ist eine nichtnegative reellwertige Funktion mit der Eigenschaft, dass der Flächeninhalt zwischen der x-Achse und der Funktion gleich Eins ist. Sie beschreibt eine Datenmenge, wenn die prozentualen Anteile der Datenpunkte in jedem Intervall ungefähr gleich dem Flächeninhalt zwischen x-Achse und Funktion über diesem Intervall sind.

2. Die "Mitte" der Daten wird durch Lagemaßzahlen wie (empirisches) arithmetisches Mittel und Median beschrieben, die "Streuung" der Daten um den mittleren Wert geben Streuungsmaßzahlen wie (empirische) Varianz und Interquartilsabstand an.

3. Ein Boxplot beschreibt eine Datenmenge durch Angabe von Median (mittlere Linie), 1. und 3. Quartil (Enden der Box, Länge ist Interquartilsabstand) sowie dem von Ausreißern bereinigten Maximum und Minimum der Daten.

Lernziele der Vorlesung am 17.11.2009

Nach dieser Vorlesung sollten Sie

1. verstanden haben, nach welchem Prinzip bei der linearen Regression eine Gerade an Daten angepasst wird, und den qualitativen Verlauf einer solchen Gerade in einfachen Fällen angeben können,

2. die Begriffe Kovarianz und Korrelation kennen und ihren Zusammenhang mit der linearen Regression erläutern können,

3. das Prinzip der Regressionsschätzung durch lokale Mittelung erklären können.

3.4 Regressionsrechnung

Geg.: 2–dimensionale Messreihe

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

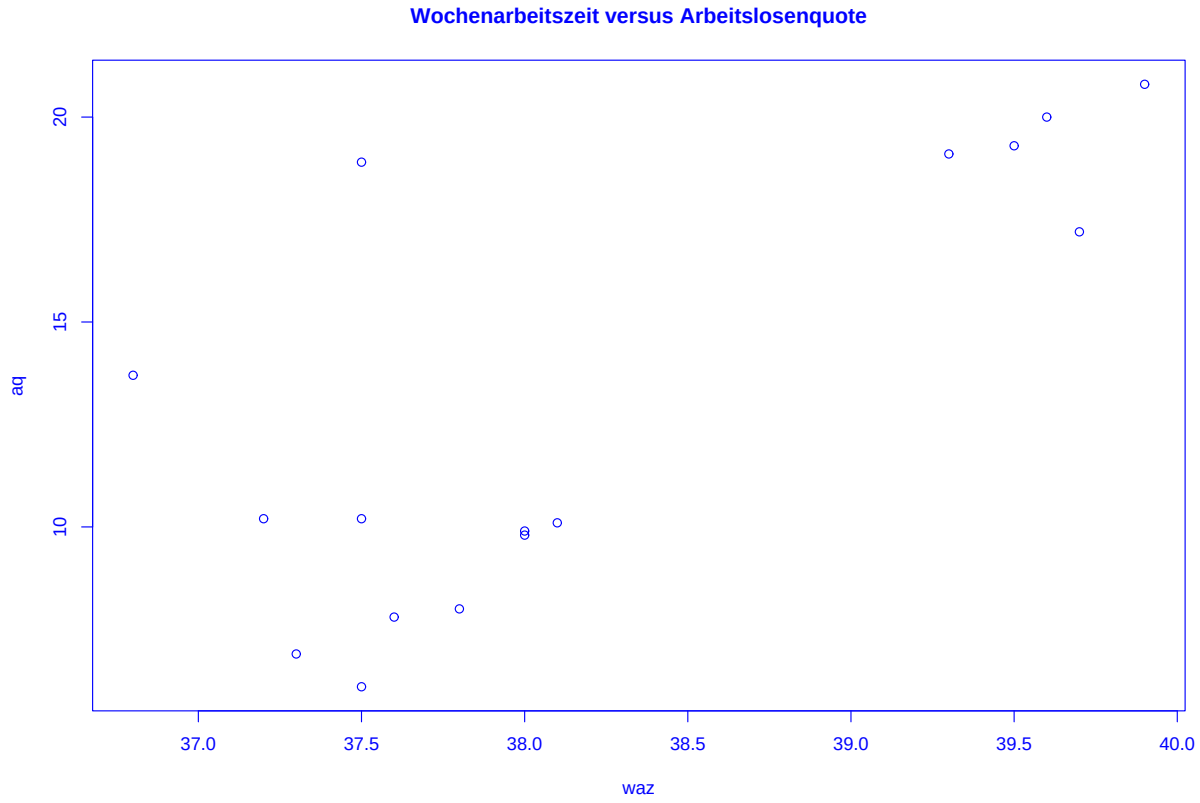
vom Umfang n.

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.

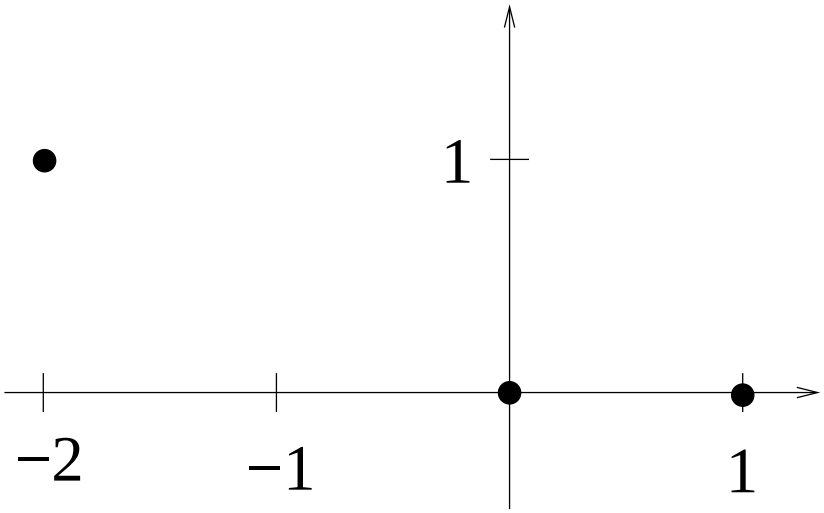
Eine Möglichkeit dafür:

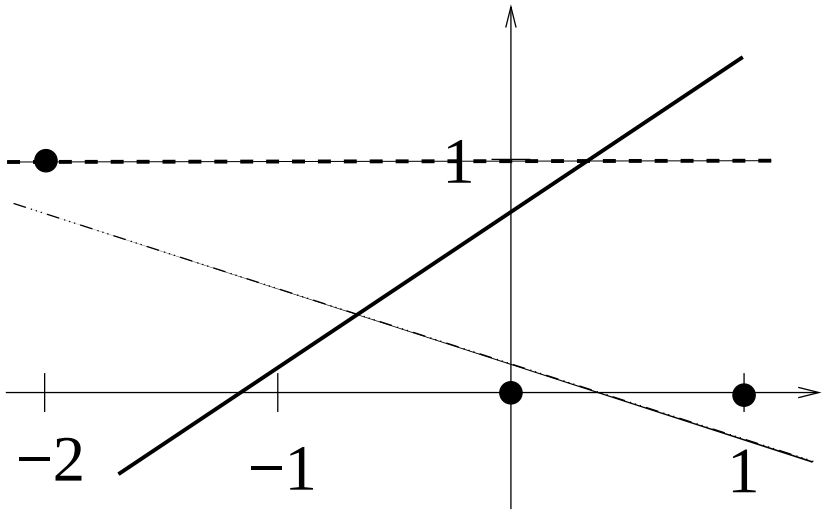
Wähle **a, b** $\in \mathbb{R}$ durch Minimierung von

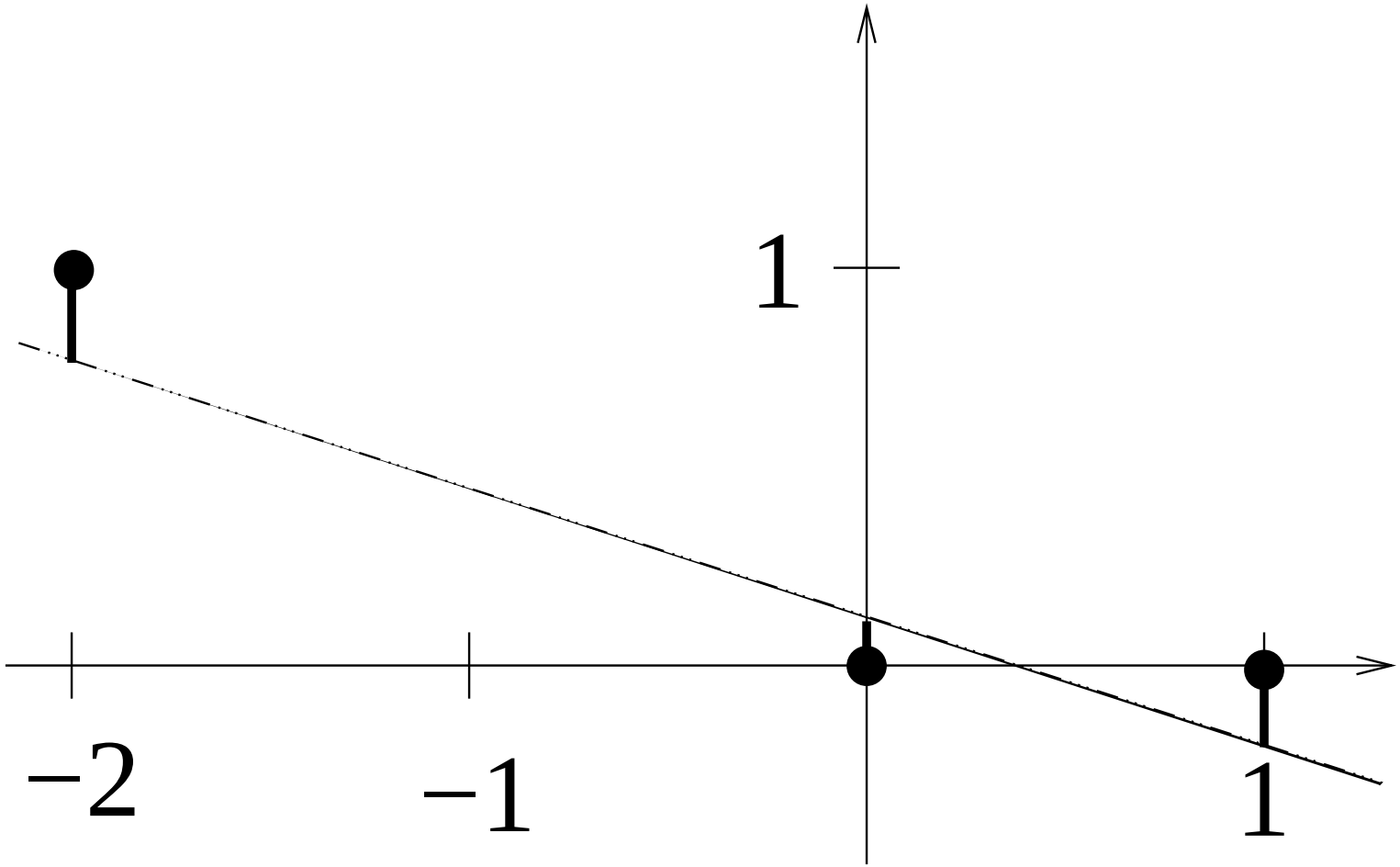
$$\sum_{i=1}^{n} \left(y_i - (a \cdot x_i + b)\right)^2 .$$

Beispiel: Es sei $n = 3$ und

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






1
−2
−1
1

Es ist $n = 3$ und

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

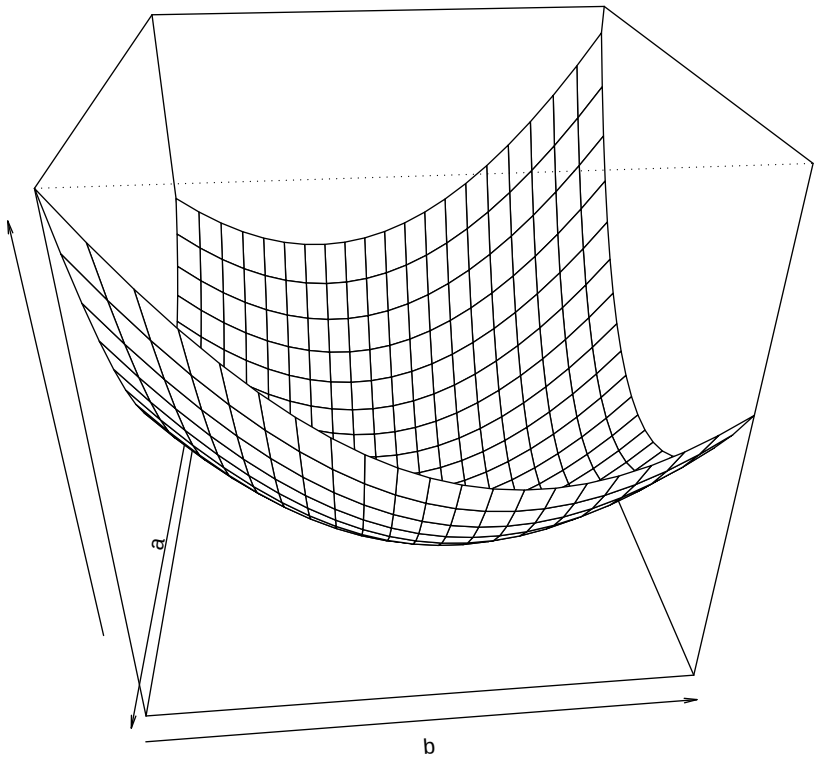
Gesucht ist dann eine Gerade

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

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

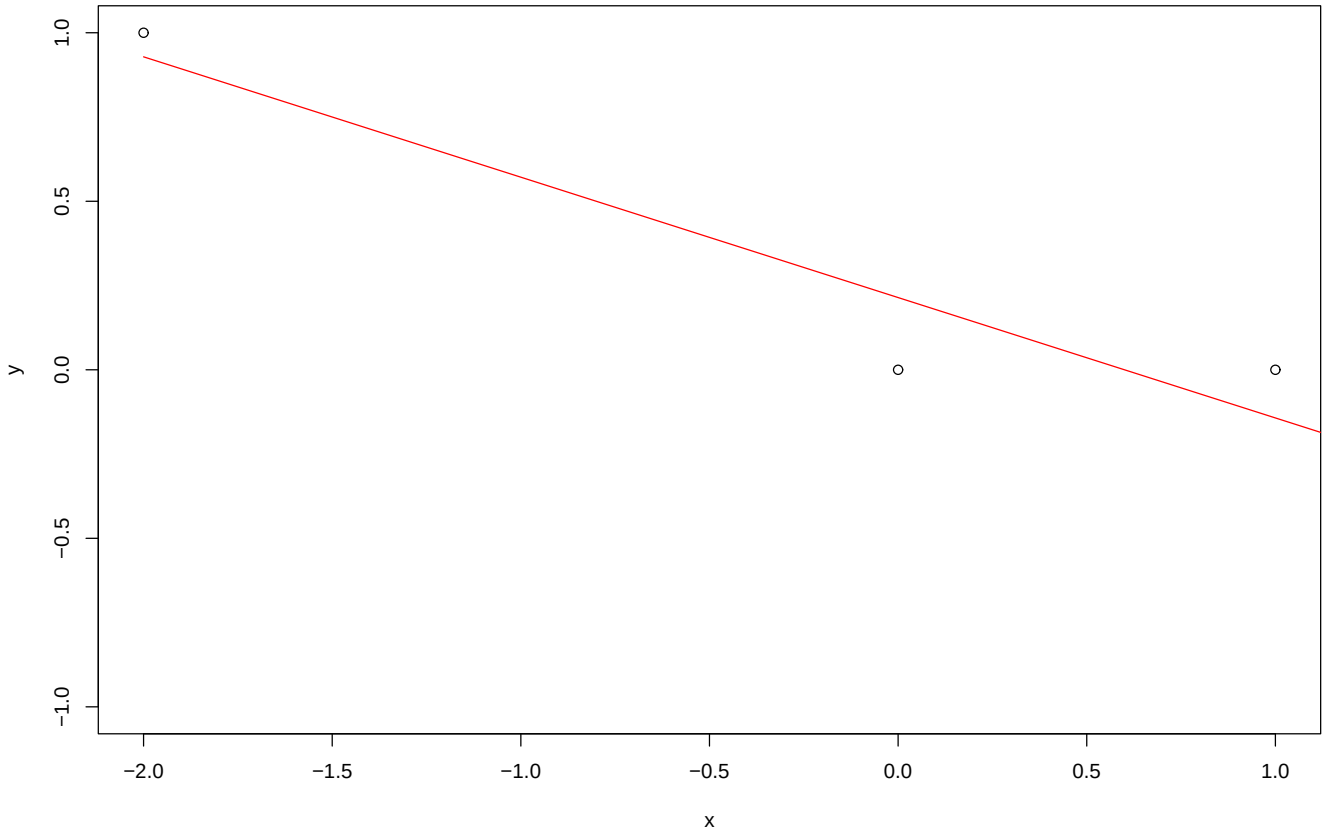
$$b^2 + (a+b)^2 + (1 + 2 \cdot a - b)^2$$

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{a} = \frac{s_{x,y}}{s_x^2} \quad (\text{mit } \tfrac{0}{0} := 0),$$

wobei

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

die (empirische) Varianz der x-Koordinaten ist und

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

die sogenannte empirische Kovarianz der zweidimensionalen Messreihe ist.

Im Beispiel oben: $n = 3$, $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 0)$, $(x_3, y_3) = (-2, 1)$ ist

$$\bar{x} = \frac{1}{3} \cdot (0 + 1 + (-2)) = -\frac{1}{3}, \quad \bar{y} = \frac{1}{3} \cdot (0 + 0 + 1) = \frac{1}{3},$$

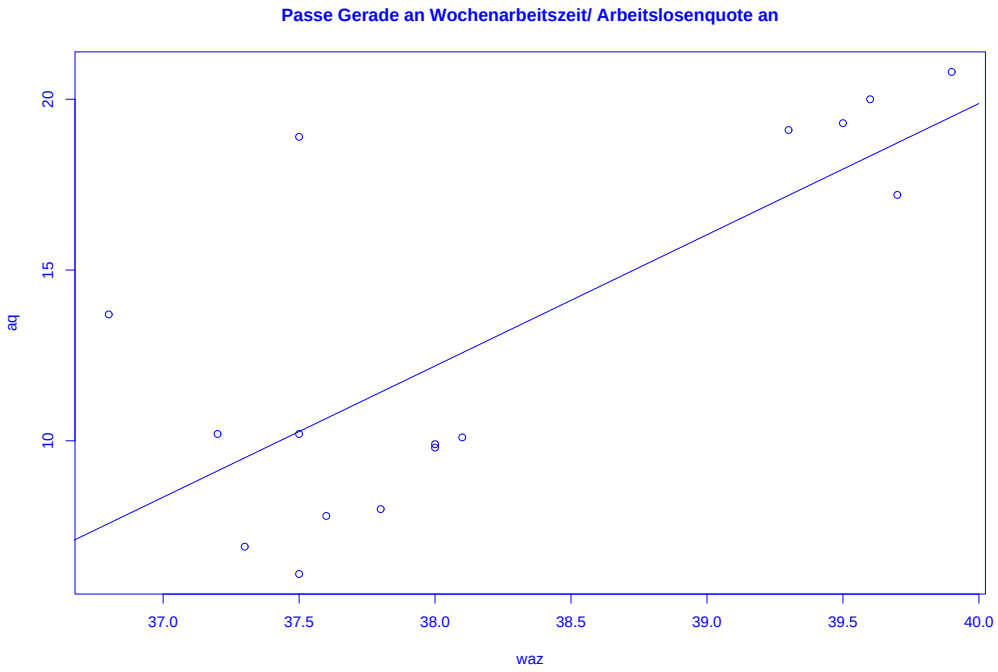
$$s_x^2 = \frac{1}{3-1} \left((0 - (-1/3))^2 + (1 - (-1/3))^2 + (-2 - (-1/3))^2 \right) = \frac{21}{9}$$

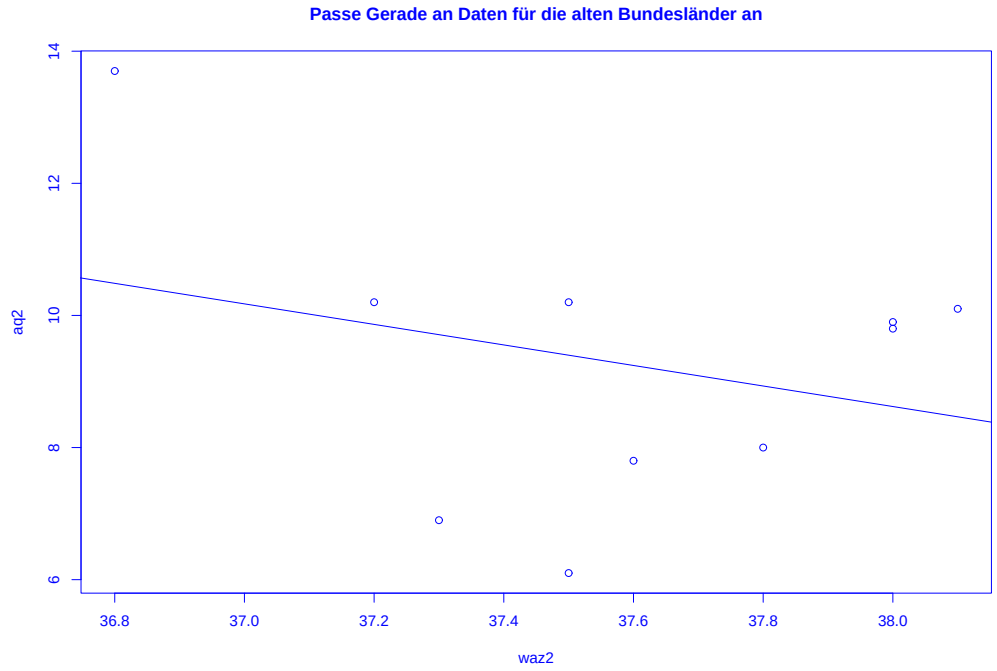
und

$$\begin{aligned} s_{x,y} &= \frac{1}{3-1} \Big((0 - (-1/3)) \cdot (0 - 1/3) + (1 - (-1/3)) \cdot (0 - 1/3) \\ &\qquad + (-2 - (-1/3)) \cdot (1 - 1/3) \Big) = -\frac{15}{18}. \end{aligned}$$

$$\Rightarrow \hat{a} = \frac{s_{x,y}}{s_x^2} = \frac{-15/18}{21/9} = -\frac{5}{14} \quad \text{und} \quad y = -\frac{5}{14} \cdot (x + 1/3) + 1/3 = -\frac{5}{14} \cdot x + \frac{3}{14}$$

Beispiel:


Passe Gerade an Wochenarbeitszeit/ Arbeitslosenquote an
aq
waz
10
15
20
37.0
37.5
38.0
38.5
39.0
39.5
40.0


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

Eine maßstabsunabhängige Variante der (empirischen) Kovarianz

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

ist die sogenannte empirische Korrelation:

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

Sowohl die (empirische) Kovarianz $s_{x,y}$ als auch die empirische Korrelation

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

haben das gleiche Vorzeichen wie die Steigung

$$\hat{a} = \frac{s_{x,y}}{s_x^2}$$

der Regressionsgeraden und machen daher eine Aussage über einen linearen Zusammenhang zwischen den x- und den y-Koordinaten einer Datenmenge.

Daher gilt:

- Die empirische Kovarianz oder Korrelation ist genau dann positiv (bzw. negativ), wenn auch die Steigung der Regressionsgeraden positiv (bzw. negativ) ist.

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

Darüberhinaus kann man zeigen:

Die empirische Korrelation nimmt nur Werte in $[-1, 1]$ an und ist sie gleich -1 oder $+1$, so liegen die Punkte alle auf einer Geraden.

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**.

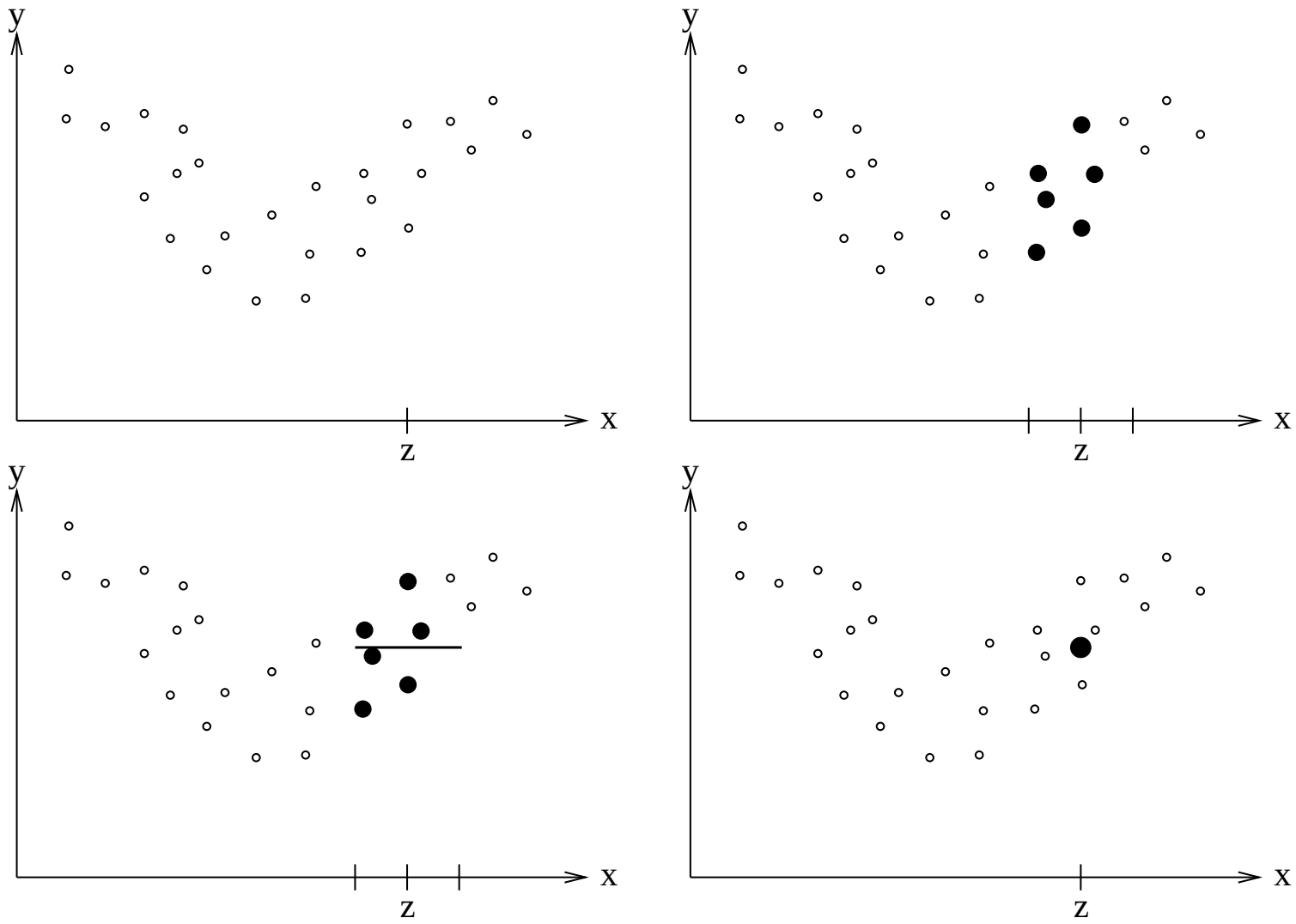
Anderer Ansatz:

Nichtparametrische Regressionsschätzung.

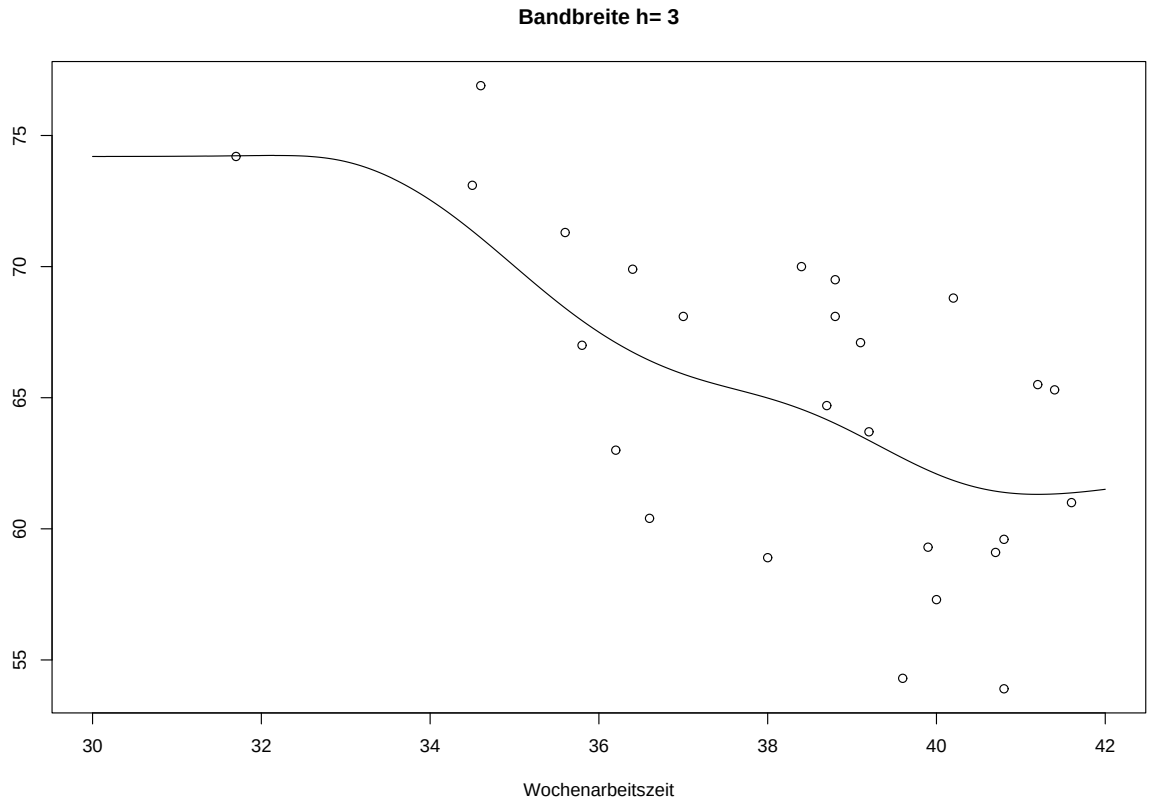
Keine Annahme über die Bauart der anzupassenden Funktion.

Einfachstes Beispiel: lokale Mittelung

Der Wert des Schätzers an einer Stelle z ist das arithmetische Mittel der y-Werte aller der Datenpunkte, bei denen der Abstand vom x-Wert zu z kleiner als eine vorgegebene Schranke ist.


y
x
z
y
x
z
y
x
z
y
x
z

Beispiel: Zusammenhang zwischen Wochenarbeitszeit und Beschäftigungsquote in 26 europäischen Staaten im Jahr 2006:



Zusammenfassung der Vorlesung am 17.11.2009

1. Bei der linearen Regression passt man eine Gerade so an gegebene Punkte an, dass die Summe der Quadrate der Abstände zwischen den y-Werten der Punkte und den y-Werten auf der Gerade minimal ist.

2. Kovarianz und Korrelation haben das gleiche Vorzeichen wie die Steigung der Regressionsgeraden und können daher zur Beurteilung eines linearen Zusammenhangs zwischen den x- und den y-Werten einer gegebenen Menge von Punkten verwendet werden.

3. Die Korrelation ist maßstabsunabhängig und liegt im Intervall $[-1, 1]$.

4. Bei der Regressionsschätzung durch lokale Mittelung wird die Wert an einer Stelle als arithmetisches Mittel der y-Werte derjenigen Datenpunkte berechnet, deren x-Wert in der Nähe der Stelle liegt.

Lernziele der Vorlesung am 24.11.2009

Nach dieser Vorlesung sollten Sie

1. die Begriffe Zufallsexperiment, Grundmenge, Ereignis und absolute bzw. relative Häufigkeit des Eintretens eines Ereignisses erkäutern können,

2. erklären können, was wir anschaulich in dieser Vorlesung unter einer Wahrscheinlichkeit verstehen,

3. den Begriff des Wahrscheinlichkeitsraums kennen.

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 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.

$\Rightarrow \quad \Omega =$

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						
relative Häufigkeit						

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.

Im Beispiel oben:

10-maliges Werfen eines echten Würfels liefert die Ergebnisse

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

Absolute und relative Häufigkeit des Eintretens von Ereignissen:

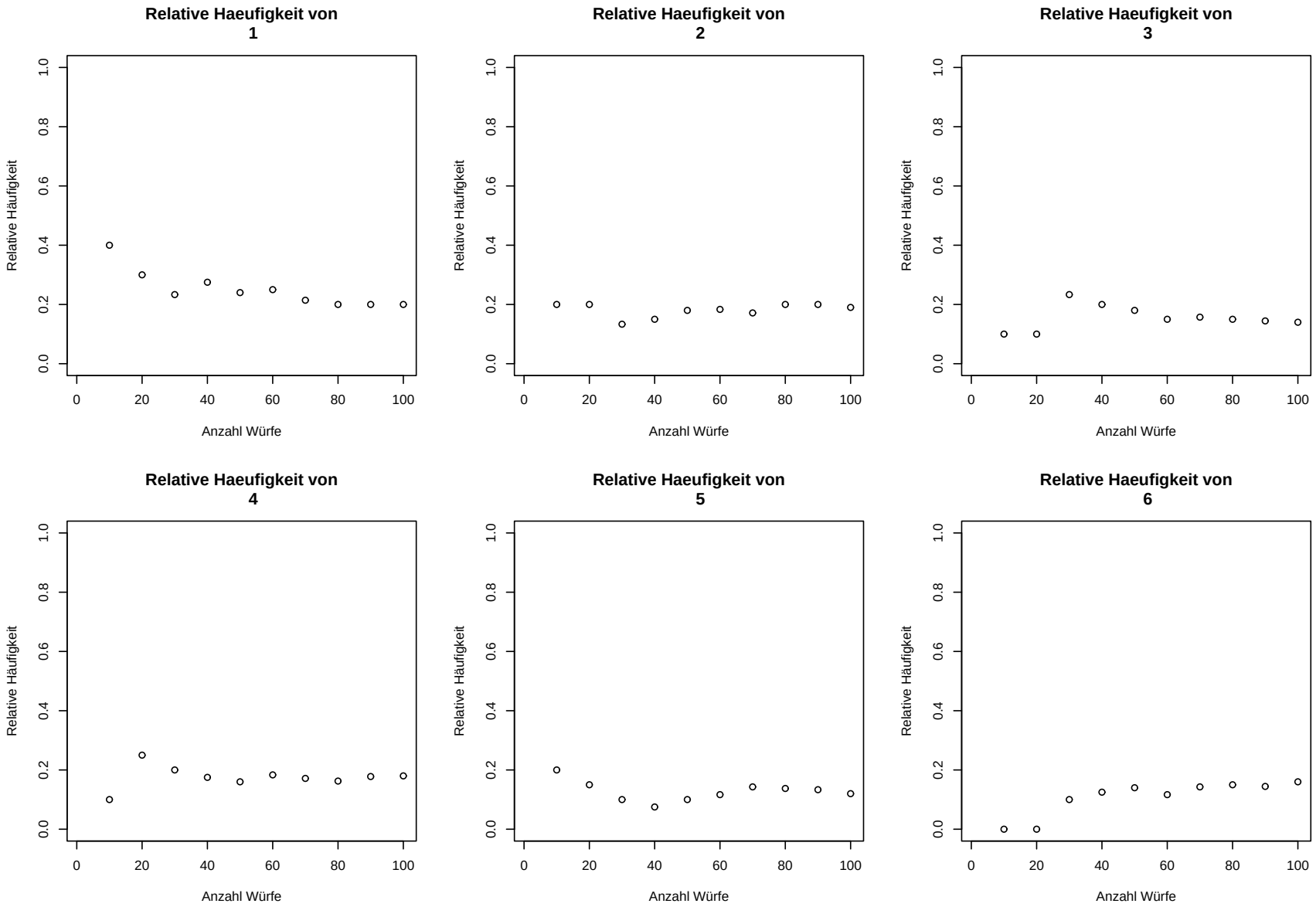
	$A = \{1, 3, 5\}$	$B = \{1, 2, 3, 4, 5\}$
absolute Häufigkeit des Eintretens		
relative Häufigkeit des Eintretens		

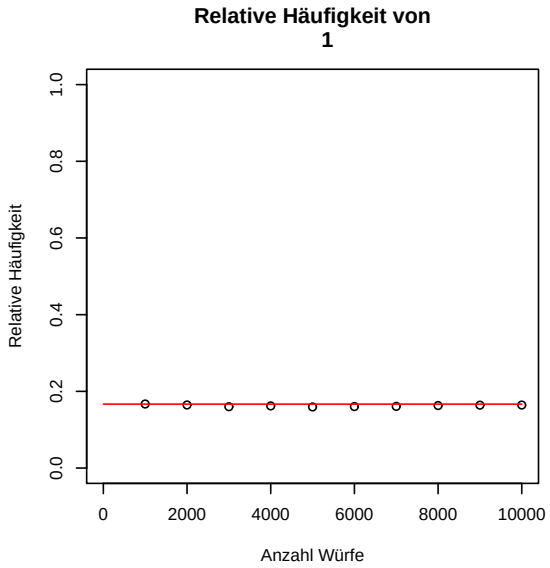
Das empirische Gesetz der großen Zahlen:

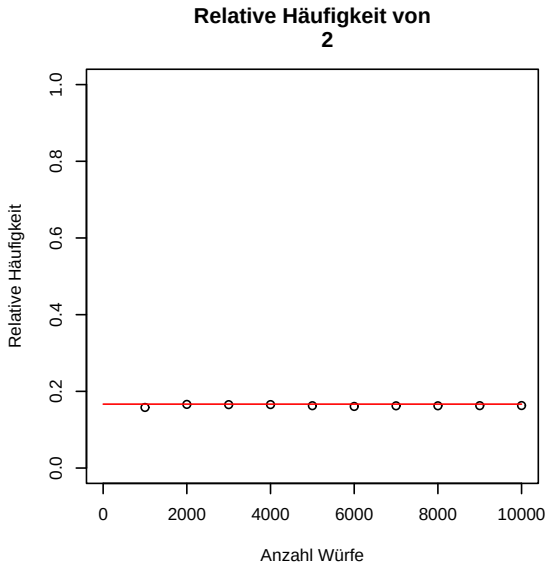
Beobachtung aus der Praxis:

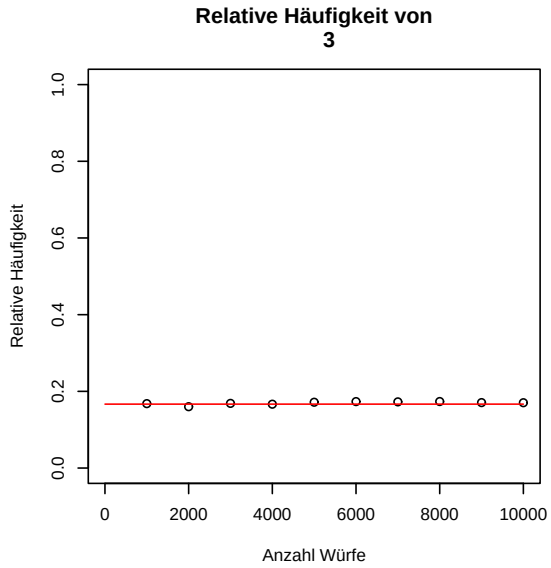
Führt man ein Zufallsexperiment **unbeeinflusst voneinander immer wieder** durch, so **nähert** sich die **relative Häufigkeit** des Auftretens eines beliebigen Ereignisses A einer (von A abhängenden) **festen Zahl** $\mathbf{P}(A) \in [0, 1]$ an.

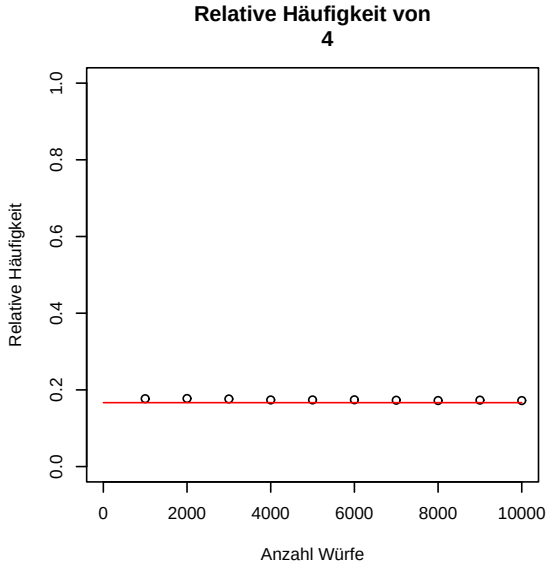
Die Zahl $\mathbf{P}(A)$ nennen wir Wahrscheinlichkeit des Ereignisses A.

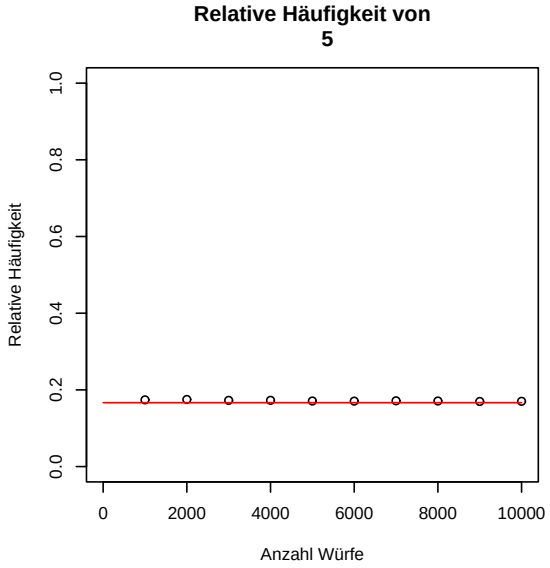

Relative Haeufigkeit von
1
Relative Haeufigkeit von
2
Relative Haeufigkeit von
3
Relative Haeufigkeit von
4
Relative Haeufigkeit von
5
Relative Haeufigkeit von
6
Relative Häufigkeit
Anzahl Würfe
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

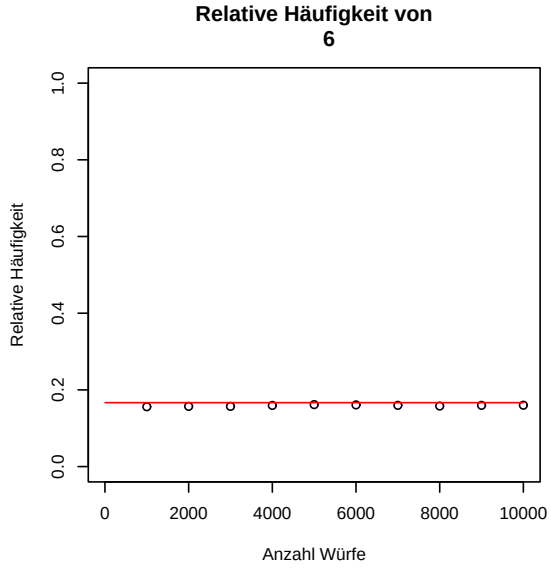

Relative Häufigkeit von
1
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe


Relative Häufigkeit von
2
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe


Relative Häufigkeit von
3
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe


Relative Häufigkeit von
4
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe


Relative Häufigkeit von
5
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe


Relative Häufigkeit von
6
Relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Anzahl Würfe

Im Folgenden überlegen wir uns einige Gesetzmäßigkeiten, die für Wahrscheinlichkeiten immer gelten:

(I)
$$0 \leq \mathbf{P}(A) \leq 1 \quad \text{für alle } A \subseteq \Omega$$

(denn dies haben wir schon in der Definition gefordert, da es aus der Tatsache folgt, dass relative Häufigkeiten immer zwischen 0 und 1 liegen).

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

(denn die relativen Häufigkeiten des Eintretens von $\emptyset$ und Ω sind immer 0 bzw. 1, also muss dies auch für die Wahrscheinlichkeiten als Grenzwerte dieser relativen Häufigkeiten gelten).

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

(denn die relative Häufigkeit des Eintretens des Komplements eines Ereignisses A ist immer gleich 1 minus der relativen Häufigkeit des Eintretens von A).

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

(denn haben A und B keine Elemente gemeinsam, so ist die relative Häufigkeit des Eintretens von A oder B gleich die Summe der relativen Häufigkeit des Eintretens von A und der relativen Häufigkeit des Eintretens von B).

(V) Für alle $n \in \mathbb{N}$ and alle $A_1, A_2, \ldots, A_n \subseteq \Omega$ mit $A_i \cap A_j = \emptyset$ für alle $1 \leq i, j \leq n$ mit $i \neq j$ gilt:

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n).$$

(analog zu (IV)).

(VI) Für den Aufbau einer mathematischen Theorie sinnvoll:

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)-(VI):

Gelten die Bedingungen (I)-(VI), 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:

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

Bemerkung: Das intuitive Verständnis von Wahrscheinlichkeiten ist oft schwierig.

Beispiel: *Linda ist 31 Jahre alt. Sie ist Single, verbal versiert und sehr intelligent. Sie hat auf einem College Philosophie studiert. Als Studentin war sie sehr engagiert in Fragen sozialer Diskriminierung und anderen sozialen Problemen: sie nahm auch an Anti-Kernkraft-Demonstrationen teil.*

Was ist wahrscheinlicher:

1) Linda ist Bankangestellte.

2) Linda ist Bankangestellte und aktiv in der Frauenbewegung.

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)-(VI) 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 (Werfen eines echten Würfels) führen Symmetrieüberlegungen auf

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

Wegen (V) folgt daraus sofort:

$$\mathbf{P}(A) =$$

Damit ist der Wahrscheinlichkeitsraum in diesem Beispiel gegeben durch

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

Zusammenfassung der Vorlesung am 24.11.2009

1. Ein Zufallsexperiment ist eine Experiment mit vorher unbestimmten Ausgang, das unbeeinflusst voneinander beliebig oft wiederholt werden kann.

2. Nach dem empirischen Gesetz der großen Zahlen nähert sich die relative Häufigkeit eines Ereignisses (für große Anzahlen von unbeeinflussten Wiederholungen des Zufallsexperiments) immer mehr einer (von dem Ereignis abhängenden) Zahl an, die wir als Wahrscheinlichkeit dieses Ereignisses bezeichnen.

3. Ein Wahrscheinlichkeitsraum ist ein Paar $(\Omega, \mathbf{P})$, wobei Ω eine nichtleere Menge ist und $\mathbf{P}$ jeder Teilmenge A von Ω eine Wahrscheinlichkeit $\mathbf{P}(A) \in [0, 1]$ so zuweist, dass gewisse Gesetzmäßigkeiten gelten.

Lernziele der Vorlesung am 01.12.2009

Nach dieser Vorlesung sollten Sie

1. den Begriff des Laplaceschen Wahrscheinlichkeitsraumes kennen und erläutern können, wann man diesen zur Modellierung eines Zufallsexperimentes einsetzen kann,

2. wissen, was ein diskreter Wahrscheinlichkeitsraum ist und wie man in diesem Wahrscheinlichkeiten von Ereignissen berechnet.

Modelle für Wahrscheinlichkeiten

4.3.1 Der Laplacesche Wahrscheinlichkeitsraum

Laplacesche Wahrscheinlichkeitsräume werden zur Beschreibung von Zufallsexperimenten verwendet, bei denen

1. nur endlich viele verschiedene Werte als Ergebnis vorkommen können,

2. jeder dieser Werte mit der gleichen Wahrscheinlichkeit auftritt.

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**.

Im Laplaceschen Wahrscheinlichkeitsraum gilt:

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

Beispiel: Bei einem Glücksspiel werden nach einem Einsatz von 1 Euro vier Münzen geworfen, und zwar zwei 50 Cent Münzen, eine 1 Euro Münze und eine 2 Euro Münze, und der Spieler bekommt als Gewinn alle die Münzen, die mit $Zahl$ oben landen.

Wie groß ist die Wahrscheinlichkeit, dass der Gewinn mindestens so groß ist wie der Einsatz ?

Als Ergebnis des Zufallsexperiments betrachten wir die $Lage\ der\ Münzen$. Dazu denken wir uns die Münzen durchnummeriert mit den Zahlen 1 bis 4, wobei die Münzen 1 und 2 den Wert 50 Cent haben, die Münze 3 den Wert 1 Euro und die Münze 4 den Wert 2 Euro hat.

Da jede der 16 möglichen Kombinationen mit der gleichen Wahrscheinlichkeit auftritt, können wir das Zufallsexperiment durch einen Laplaceschen Wahrscheinlichkeitsraum beschreiben mit Grundmenge

$$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \quad : \quad \omega_i \in \{Z, W\}\},$$

wobei $\omega_i = Z$ bedeutet, dass die i-te Münze mit Zahl oben gelandet ist.

Gesucht: $\mathbf{P}(A)$ mit

$$A = \{(\omega_1, \omega_2, \omega_3, \omega_4) \in \Omega \quad : \quad \text{Wert der Münzen mit Zahl oben} \geq 1 \text{ Euro}\}$$

Wegen

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{16}$$

müssen wir hierzu die Anzahl der Elemente in A bestimmen.

ω_1 50 Cent	ω_2 50 Cent	ω_3 1 Euro	ω_4 2 Euro	Gewinn	Gewinn $\geq$ 1 Euro ?
W	W	W	W		
W	W	W	Z		
W	W	Z	W		
W	W	Z	Z		
W	Z	W	W		
W	Z	W	Z		
W	Z	Z	W		
W	Z	Z	Z		
Z	W	W	W		
Z	W	W	Z		
Z	W	Z	W		
Z	W	Z	Z		
Z	Z	W	W		
Z	Z	W	Z		
Z	Z	Z	W		
Z	Z	Z	Z		

Damit gilt $|A| =$ und

$$\mathbf{P}(A) =$$

Einfacher: Es gilt

$$\bar{A} =$$

was

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

impliziert.

4.3.2 Diskrete Wahrscheinlichkeitsräume

Diskrete Wahrscheinlichkeitsräume verwenden wir zur Beschreibung aller der Zufallsexperimente, bei denen nur endlich viele oder abzählbar unendlich viele verschiedene Werte für das Ergebnis möglich sind.

In diesem Fall berechnen wir die Wahrscheinlichkeit eines Ereignisses als Summe der Wahrscheinlichkeiten aller darin enthaltener Elementarereignisse.

Beispiel: Mit einem echten Würfel wird solange gewürfelt, bis der Würfel zum ersten Mal mit 6 oben landet.

Wie groß ist die Wahrscheinlichkeit, dass die Anzahl Würfe kleiner als vier ist?

Wir bestimmen zunächst für $k \in \mathbb{N}$ fest die Wahrscheinlichkeit, dass der Würfel genau beim k-ten Wurf zum ersten Mal mit 6 oben landet.

Werfen wir einen echten Würfel k-mal hintereinander, so können bei dieser Sequenz von k Würfen

$$6 \cdot 6 \cdot \cdots \cdot 6 = 6^k$$

viele verschiedene Ergebnisse auftreten.

Soll dabei der letzte Wurf eine 6 ergeben und alle anderen nicht, so gibt es davon

viele verschiedene Sequenzen.

Da bei k-maligen Werfen jede einzelne Sequenz der Ergebnisse mit der gleichen Wahrscheinlichkeit $1/6^k$ auftritt, gilt für die auf dieser Folie gesuchte Wahrscheinlichkeit

$$\mathbf{P}(\{k\}) =$$

Damit ist die Wahrscheinlichkeit, dass die Anzahl Würfe kleiner als vier ist, gegeben durch

$$\begin{aligned}\mathbf{P}(\{1,2,3\}) &= \mathbf{P}(\{1\}) + \mathbf{P}(\{2\}) + \mathbf{P}(\{3\}) \\ &= \frac{5^0}{6^1} + \frac{5^1}{6^2} + \frac{5^2}{6^3} = \frac{1}{6} + \frac{5}{36} + \frac{25}{216} = \frac{36+30+25}{216} = \frac{91}{216}.\end{aligned}$$

Im Folgenden formulieren wir den zugrunde liegenden Wahrscheinlichkeitsraum allgemein. Dazu nehmen wir ohne Beschränkung der Allgemeinheit an, dass beim Zufallsexperiment als Ergebnis eine der Zahlen 0, 1, 2, . . . auftritt (manche davon evt. nur mit Wahrscheinlichkeit Null).

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 einen sogenannten **diskreten Wahrscheinlichkeitsraum** 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.

In diesem Falle bezeichnen wir $(\Omega, \mathbf{P})$ als **diskreten Wahrscheinlichkeitsraum** und $\mathbf{P}$ als **diskretes Wahrscheinlichkeitsmaß**.

Im Beispiel oben: $\Omega = \mathbb{N}_0$, $p_0 = 0$ und $p_n = \frac{5^{n-1}}{6^n}$ für $n \in \mathbb{N}$.

$(p_n)_{n\in\mathbb{N}_0}$ ist Zähldichte, da $p_n \geq 0$ für alle $n \in \mathbb{N}_0$ und

$$\begin{aligned}
\sum_{n=0}^{\infty} p_n &= 0 + \frac{5^0}{6^1} + \frac{5^1}{6^2} + \frac{5^2}{6^3} + \dots \\
&= \frac{1}{6} \cdot \left(\left(\frac{5}{6}\right)^0 + \left(\frac{5}{6}\right)^1 + \left(\frac{5}{6}\right)^2 + \dots \right) \\
&= \frac{1}{6} \cdot \frac{1}{1 - \frac{5}{6}} = 1.
\end{aligned}$$

Wir haben dann berechnet:

$$\mathbf{P}(\{1,2,3\}) = p_1 + p_2 + p_3 = \sum_{k\in\{1,2,3\}} p_k.$$

Beispiele für diskrete Wahrscheinlichkeitsmaße:

1. Sei $n \in \mathbb{N}$ und $p \in [0,1]$. Das 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 diskrete Wahrscheinlichkeitsmaß heißt **Binomialverteilung** mit Parametern n und p.

Hierbei

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k \cdot (k-1) \cdot \cdots \cdot 1} \quad (\text{sog. } \textit{Binomialkoeffizient}).$$

Einsatz in der Modellierung: Siehe nächste Vorlesung.

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

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

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

Hierbei:

$$k! = k \cdot (k-1) \cdot \cdots \cdot 1 \quad (\text{sog. } \textit{Fakultät}).$$

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

Zusammenfassung der Vorlesung am 01.12.2009

1. Ein Laplacescher Wahrscheinlichkeitsraum ist ein Paar $(\Omega, \mathbf{P})$ mit endlicher Grundmenge Ω und

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

   Er wird zur Modellierung von Zufallsexperimenten eingesetzt, bei denen jedes der endlich vielen möglichen Ergebnisse mit der gleichen Wahrscheinlichkeit auftritt.

2. In einem diskreten Wahrscheinlichkeitsraum $(\mathbb{N}_0, \mathbf{P})$ mit Zähldichte $(p_n)_{n\in\mathbb{N}_0}$ (d.h., $p_n \geq 0$ $(n \in \mathbb{N}_0)$ und $\sum_{n=0}^{\infty} p_n = 1$) gilt

$$\mathbf{P}(A) = \sum_{k\in A} p_k \quad (A \subseteq \mathbb{N}_0).$$

   D.h. hier ist die Wahrscheinlichkeit eines Ereignisses A gleich der Summe der Wahrscheinlichkeiten p_k aller in A enthaltenen Elementarereignisse $\{k\}$.

Lernziele der Vorlesung am 08.12.2009

Nach dieser Vorlesung sollten Sie

1. eine wichtige statistische Schlussweise und eine Anwendung der Binomialverteilung kennengelernt haben,

2. wissen, was eine Wahrscheinlichkeitsraum mit Dichte ist und wie man darin Wahrscheinlichkeiten berechnet.

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 von Oktober 1955 bis Dezember 2007:

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.

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.

Aufgabe

Bestimmen Sie die Wahrscheinlichkeit, dass bei einer Ziehung von 6 Zahlen aus der Menge der Zahlen

$$1, 2, 3, \ldots, 49$$

die Zahl 38 gezogen wird.

Hinweis: Betrachten Sie das Ziehen ohne Zurücklegen und ohne Beachtung der Reihenfolge und verwenden Sie die Formel

$$\frac{\text{Anzahl der “günstigen” Fälle}}{\text{Anzahl der “möglichen” Fälle}}$$

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

verschiedene Möglichkeiten auftreten.

Daher tritt bei einer einzigen Ziehung die 38 mit Wahrscheinlichkeit

$$p =$$

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 ist bisher gezeigt:

Bei einer einzelnen Lottoziehung tritt die Zahl 38 mit Wahrscheinlichkeit

$$p = \frac{6}{49}$$

auf.

Und führen wir unbeeinflusst voneinander n solche Lottoziehungen hintereinander durch, so wird die (zufällige) Anzahl $\in \{0, 1, \ldots, n\}$ der Ziehungen, bei denen die Zahl 38 auftritt, durch eine Binomialverteilung mit Parametern n und p beschrieben.

Also 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.

4.3.3 Wahrscheinlichkeitsräume mit Dichten

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.

Zur Definition von **Wahrscheinlichkeitsräumen mit Dichten** wählen wir $\Omega = \mathbb{R}$ und eine Dichte $f : \mathbb{R} \to \mathbb{R}$ und setzen

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

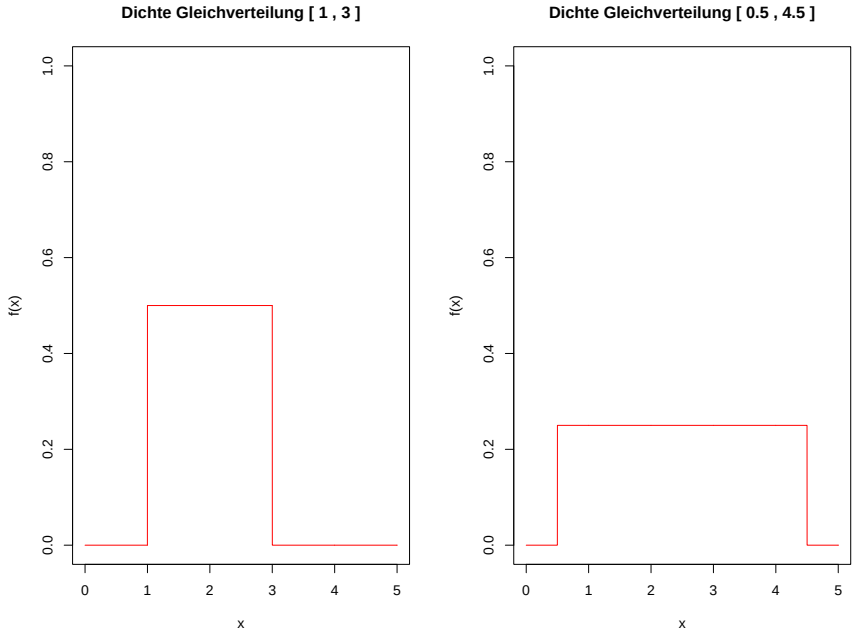
Hierbei sind die Wahrscheinlichkeiten für das Eintreten eines Elementarereignisses immer Null.

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 Wahrscheinlichkeitsmaß.



Beispiel: Eine Zahl wird rein zufällig aus dem Intervall $[0, 10]$ gezogen.

Wie groß ist die Wahrscheinlichkeit, dass die Zahl kleiner oder gleich 4 ist?

Die zufällige Zahl wird durch eine *Gleichverteilung* $U(0, 10)$ beschrieben, d.h. durch ein Wahrscheinlichkeitsmaß $\mathbf{P}$ mit Dichte

$$f(x) = \begin{cases} \frac{1}{10} & \text{für} \quad 0 \leq x \leq 10, \\ 0 & \text{für} \quad x < 0 \text{ oder } x > 10 \end{cases}$$

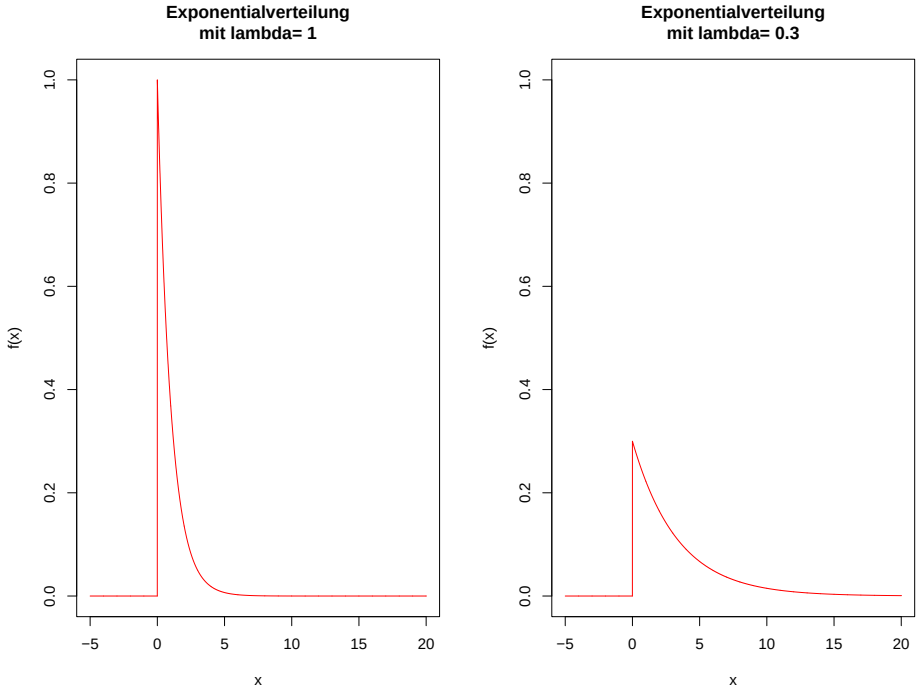
Daher ist die gesuchte Wahrscheinlichkeit gegeben durch

$$\mathbf{P}((-\infty, 4]) \quad = \quad \int_{(-\infty,4]} f(x)\, dx =$$

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 Wahrscheinlichkeitsmaß.



Beispiel: Die Zeit (in Monaten), in der sich ein Universitätsabsolvent nach dem Studium bis zum Antritt der ersten Arbeitsstelle bewirbt, sei Exponentialverteilt mit Parameter $\lambda = 0,3$. Wie groß ist dann die Wahrscheinlichkeit, dass er länger als ein halbes Jahr auf den ersten Arbeitsplatz wartet?

Für eine Exponentialverteilung $\mathbf{P}$ mit Parameter $\lambda = 0,3$, d.h. für ein Wahrscheinlichkeitsmaß mit Dichte

$$f(x) = \begin{cases} 0,3 \cdot e^{-0,3 \cdot x} & \text{für} \quad x \geq 0, \\ 0 & \text{für} \quad x < 0 \end{cases}$$

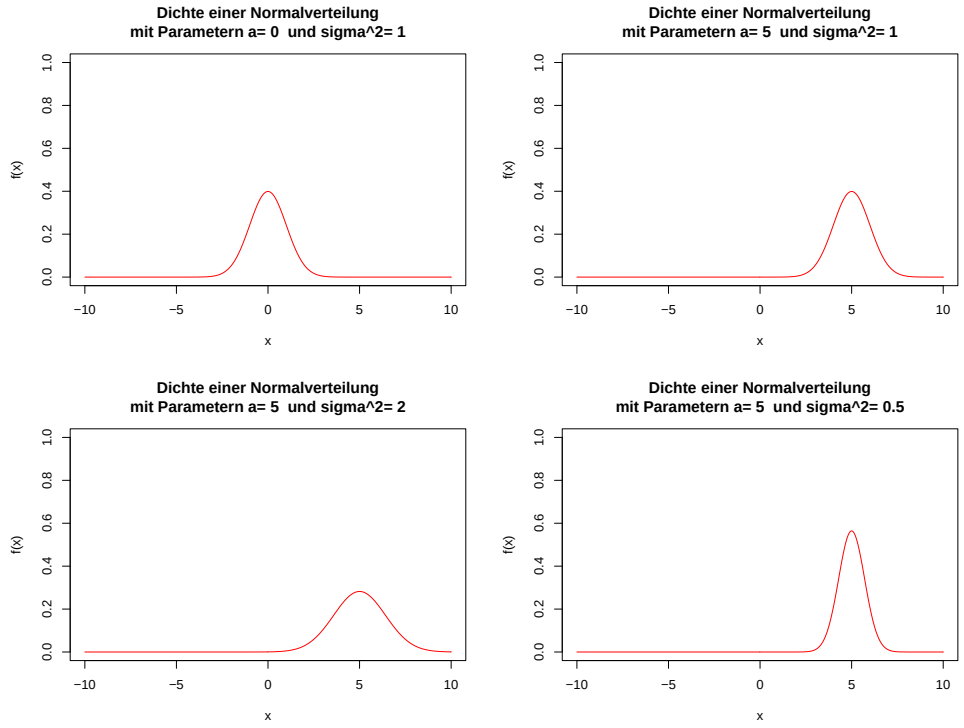
gilt:

$$\begin{aligned} \mathbf{P}((6,\infty)) &= \int_{(6,\infty)} f(x)\,dx = \int_6^\infty 0,3 \cdot e^{-0,3 \cdot x}\,dx \\ &= -e^{-0,3 \cdot x}\Big|_{x=6}^{\infty} = e^{-0,3 \cdot 6} = e^{-1,8} \approx 0,165. \end{aligned}$$

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

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

festgelegte Wahrscheinlichkeitsmaß.



Zusammenfassung der Vorlesung am 08.12.2009

1. Eine Binomialverteilung tritt dann auf, wenn man ein Zufallsexperiment mit möglichem Ergebnis "Erfolg" oder "Misserfolg" wiederholt durchführt und die Anzahl der Erfolge zählt.

2. Eine Dichte ist eine nichtnegative Funktion, für die der Flächeninhalt zwischen Dichte und x-Achse gleich Eins ist.

3. In einem Wahrscheinlichkeitsraum mit Dichte $f : \mathbb{R} \to \mathbb{R}_+$ ist die Wahrscheinlichkeit einer Menge $A \subseteq \mathbb{R}$ gleich dem Flächeninhalt

$$\int_A f(x)\,dx$$

zwischen Dichte und x-Achse im Bereich dieser Menge.

Lernziele der Vorlesung am 15.12.2009

Nach dieser Vorlesung sollten Sie die folgenden Begriffe verstanden haben:

1. Zufallsvariable,
2. Verteilung einer Zufallsvariablen,
3. Unabhängigkeit.

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: Wie groß ist die Wahrscheinlichkeit, dass beim unbeeinflussten Werfen zweier echter Würfel die Summe der Augenzahlen größer oder gleich 10 ist ?

Wir modellieren zuerst (weil das einfach ist!) das unbeeinflusste Werfen zweier echter Würfel.

Dazu definieren wir einen Laplaceschen W-Raum $(\Omega, \mathbf{P})$ durch:

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

Im Beispiel oben interessiert nur die **Summe** der Augenzahlen:

Dazu wählen wir

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

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 die Wahrscheinlichkeit, dass die Summe der Augenzahlen mindestens 10 ist, gegeben durch:

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

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.

Sprechweisen für Zufallsvariablen:

Eine diskrete Zufallsvariable ist eine Zufallsvariablen, die mit Wahrscheinlichkeit Eins nur Werte aus einer endlichen oder höchsten abzählbaren Menge annimmt.

Eine Zufallsvariable X heißt **binomialverteilt** mit Parametern n und p, falls ihre Verteilung eine **Binomialverteilung** mit Parametern n und p ist (kurz: X $b(n,p)$-**verteilt**).

Eine Zufallsvariable X heißt **Poisson-verteilt** mit Parameter λ, falls ihre Verteilung eine **Poisson-Verteilung** mit Parameter λ ist (kurz: X $\pi(\lambda)$-**verteilt**).

Ist $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum, so heißt jede Abbildung

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

eine **reelle Zufallsvariable**.

Eine **stetig verteilte Zufallsvariable mit Dichte** ist eine reelle Zufallsvariable, deren Verteilung eine Dichte hat.

Eine Zufallsvariable X heißt **gleichverteilt** auf dem Intervall $[a, b]$, falls ihre Verteilung eine **Gleichverteilung** mit Parametern a und b ist (kurz: X $U(a, b)$-**verteilt**).

Eine Zufallsvariable X heißt **exponentialverteilt** mit Parameter λ, falls ihre Verteilung eine **Exponentialverteilung** mit Parameter λ ist (kurz: X $\exp(\lambda)$-**verteilt**).

Eine Zufallsvariable X heißt **normalverteilt** mit Parametern a und σ^2, falls ihre Verteilung eine **Normalverteilung** mit Parametern a und σ^2 ist (kurz: X $N(a, \sigma^2)$-**verteilt**).

4.5 Unabhängigkeit

Sei $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum.

Im folgenden interessiert uns die Frage, wann sich zwei Ereignisse gegenseitig nicht beeinflussen.

Beispiel: Wir betrachten das Werfen zweier echter Würfel und definieren Ereignisse A, B und C durch

$$\begin{aligned} A &= \text{“Augenzahl beim } 1. \text{ Würfel ist } 6\text{”} \\ B &= \text{“Augenzahl beim } 2. \text{ Würfel ist } 3\text{”} \\ C &= \text{“Summe der Augenzahlen ist größer als 10”} \end{aligned}$$

Beeinflussen sich A und B bzw. A und C gegenseitig, d.h., hat das Eintretens eines der Ereignisse Auswirkung auf die Wahrscheinlichkeit des Eintretens des anderen Ereignisses?

Formal:

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:

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

Im Beispiel oben:

Wir beschreiben des Werfen der beiden Würfel durch einen Laplaceschen Wahrscheinlichkeitsraum mit Grundmenge

$$\Omega = \{(i,j) \,:\, i,j \in \{1,2,\ldots,6\}\}\,.$$

Dann gilt

$$\begin{aligned} A &= \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\,, \\ B &= \{(1,3),(2,3),(3,3),(4,3),(5,3),(6,3)\}\,, \\ C &= \{(5,6),(6,5),(6,6)\}\,. \end{aligned}$$

$$\Rightarrow \quad \mathbf{P}(A) = \mathbf{P}(B) = \quad , \quad \mathbf{P}(C) =$$

Bereits gesehen:

$$\begin{aligned} A &= \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}, \mathbf{P}(A) = \frac{1}{6} \\ B &= \{(1,3),(2,3),(3,3),(4,3),(5,3),(6,3)\}, \mathbf{P}(B) = \frac{1}{6} \\ C &= \{(5,6),(6,5),(6,6)\}, \mathbf{P}(C) = \frac{3}{36}. \end{aligned}$$

Damit

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

und

$$\mathbf{P}(A \cap C) =$$

Also sind A, B unabhängig, aber A und C nicht.

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

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

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

Bei zwei unabhängigen Zufallsvariablen sind also alle Paare von Ereignissen unabhängig, die mit diesen beiden Zufallsvariablen gebildet werden können.

Anschaulich sind zwei Zufallsvariablen X und Y unabhängig, falls sich ihre Werte gegenseitig nicht beeinflussen.

Zusammenfassung der Vorlesung am 15.12.2009

1. Ein **Zufallsvariable** $X : \Omega \to \Omega'$ weist dem Ergebnis ω eines Zufallsexperiments einen neuen Wert $X(\omega)$ zu.

2. Auf diese Art entsteht ein neues Zufallsexperiment mit Ergebnis $X(\omega)$, welches durch das als **Verteilung** von X bezeichnete Wahrscheinlichkeitsmaß

$$\mathbf{P}_X(A') = \mathbf{P}(\{\omega \in \Omega \,:\, X(\omega) \in A'\}) \quad (A' \subseteq \Omega')$$

   beschrieben wird.

3. Ist $(\Omega, \mathbf{P})$ ein Wahrscheinlichkeitsraum, und sind $A, B \subseteq \Omega$ zwei Ereignisse, so sind diese **unabhängig** (anschaulich: beeinflussen sich diese gegenseitig nicht), falls gilt:

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

Lernziele der Vorlesung am 12.01.2010

Nach dieser Vorlesung sollten Sie

1. erklären können, was man anschaulich unter dem Erwartungswert einer Zufallsvariablen versteht,

2. Erwartungswerte in den verschiedenen Spezialfällen berechnen können.

4.6 Der Erwartungswert einer Zufallsvariablen

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.

Beispiel: Bei einem Glücksrad, das rein zufällig auf einem von 36 Feldern stehen bleibt, wird bei dem einzigem roten Feld ein Gewinn von 3 Euro, bei jedem der 12 blauen Felder ein Gewinn von 1 Euro, und bei den 23 übrigen Feldern kein Gewinn ausgezahlt.

Wie groß ist der ausgezahlte Gewinn im Mittel ?

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

Also:

Der Erwartungswert einer diskreten Zufallsvariablen wird berechnet, indem man

a) die auftretenden Werte mit den zugehörigen Wahrscheinlichkeiten multipliziert,

b) die entstehenden Produkte aufaddiert.

Dies lässt sich als Mittelwert der auftretenden Werte interpretieren, wobei die einzelnen Werte mit den Wahrscheinlichkeiten gewichtet werden.

Im Beispiel oben:

Sei X der beim Glücksrad ausgezahlte Gewinn.

Dann ist X diskrete Zufallsvariable, die die Werte 3, 1 und 0 mit den Wahrscheinlichkeiten $1/36$, $12/36$ und $23/36$ annimmt.

Daher gilt:

$$\begin{aligned} \mathbf{E}X &= \\ &= \end{aligned}$$

Beispiel: Ist X $b(1,p)$-verteilt, d.h. gilt

$$\mathbf{P}[X=1]=p \quad \text{und} \quad \mathbf{P}[X=0]=1-p,$$

so erhalten wir

$$\mathbf{E}X =$$

Beispiel: Ist X $\pi(\lambda)$-verteilt, d.h. gilt

$$\mathbf{P}[X=k]=\frac{\lambda^k}{k!}\cdot e^{-\lambda},$$

so erhalten wir

$$\mathbf{E}X=\sum_{k=0}^{\infty} k\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}=\lambda\cdot\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\cdot e^{-\lambda}=\lambda.$$

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.

Also:

Der Erwartungswert einer Zufallsvariablen mit Dichte wird berechnet, indem man

a) den Wert der Dichte an der Stelle x mit x multipliziert,

b) das entstandene Produkt über ganz $\mathbb{R}$ integriert.

Statt $Wert * Wahrscheinlichkeit$ aufzuaddieren, wird jetzt also $Wert * Dichte$ integriert.

Beispiel: Sei X eine auf $U(a,b)$-verteilte Zufallsvariable (mit $a < b$), d.h. X sei eine stetig-verteilte Zufallsvariable mit Dichte

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

Dann gilt

$$\mathbf{E}X \quad =$$

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

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

Dann gilt:

$$\begin{aligned}
&\mathbf{E}X \\
&= \int_{-\infty}^{\infty} x\cdot f(x)\,dx = \int_{-\infty}^{\infty} x\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\,dx \\
&= \int_{-\infty}^{\infty}\frac{x-a}{\sqrt{2\pi}\cdot\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\,dx + a\cdot\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\,dx \\
&\overset{(!)}{=} 0 + a\cdot 1 = a.
\end{aligned}$$

4.6.3 Eigenschaften von Erwartungswerten

Erwartungswerte haben die folgenden drei Eigenschaften, die anschaulich aus der Motivation des Erwartungswertes als "mittlerer Wert" folgen:

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:*

   Sind die reellen Zufallsvariablen X und Y unabhängig (d.h. anschaulich: beeinflussen sich die Werte von X und Y gegenseitig nicht) so gilt immer:

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

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

Wie groß ist $\mathbf{E}X$?

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

Dabei ist

$$\mathbf{E}X_1 = \mathbf{E}X_2 =$$

und damit

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

4.6.4 Weitere Formeln zur Berechnung von Erwartungswerten

Sei X eine reelle Zufallsvariable und $h : \mathbb{R} \to \mathbb{R}$ eine beliebige reelle Funktion.

Dann definieren wir

$$\mathbf{E}h(X) = \begin{cases} \sum_{k=0}^{\infty} h(k) \cdot \mathbf{P}[X = k] & \text{falls} \quad \mathbf{P}_X(\mathbb{N}_0) = 1, \\ \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx & \text{falls} \quad X \text{ die Dichte } f : \mathbb{R} \to \mathbb{R} \text{ hat.} \end{cases}$$

Beispiel: Ist X $\pi(\lambda)$-verteilt und ist Y $U(0,2)$-verteilt, so gilt

$$\mathbf{E}(X^2) = \sum_{k=0}^{\infty} k^2 \cdot \mathbf{P}[X = k] = \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = \dots$$

und

$$\mathbf{E}(e^Y) = \int_{-\infty}^{\infty} e^x \cdot f_Y(x)\, dx = \int_0^2 e^x \cdot \frac{1}{2}\, dx = \dots$$

Zusammenfassung der Vorlesung am 12.01.2010:

1. Zur Berechnung des Erwartungswertes einer diskreten Zufallsvariable X bilden wir ein mit den Wahrscheinlichkeiten gewichtetes Mittel der Werte von X:

$$\mathbf{E}X = \sum_{k=0}^{\infty} k \cdot \mathbf{P}[X = k] \quad \text{falls } \mathbf{P}_X(\mathbb{N}_0) = 1.$$

2. Zur Berechnung des Erwartungswertes einer stetig verteilten Zufallsvariablen Y mit Dichte f integrieren wir $x \cdot f(x)$ über $\mathbb{R}$:

$$\mathbf{E}Y = \int_{\mathbb{R}} x \cdot f(x)\, dx \quad \text{falls } f \text{ Dichte von } Y$$

3. Allgemeiner: $\mathbf{E}(X^2) = \sum_{k=0}^{\infty} k^2 \cdot \mathbf{P}[X = k]$ und $\mathbf{E}\left(e^Y\right) = \int_{\mathbb{R}} e^x \cdot f(x)\, dx$.

Lernziele der Vorlesung am 19.01.2010

Nach dieser Vorlesung sollten Sie

1. erklären können, was man anschaulich unter der Varianz einer Zufallsvariablen versteht,

2. die wichtigsten Rechenregeln für Varianzen kennen,

3. und Varianzen in den verschiedenen Spezialfällen berechnen können.

4.7 Die Varianz einer Zufallsvariablen

Beispiel: Beim Glückspiel Roulette bleibt eine Kugel rein zufällig in einem von insgesamt 37 gleich großen Fächern liegen. Die Fächer sind mit den Zahlen 0 bis 36 durchnummeriert, wobei 18 dieser Zahlen rot sind, nämlich:

$$1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36.$$

Die restlichen Zahlen (außer der Null) sind schwarz. Man kann nun beim Roulette vor dem Werfen der Kugel sein Geld z.B. darauf setzen, ob die Kugel in einem Feld mit einer geraden, einer ungeraden, einer roten oder einer schwarzen Zahl landet. Sofern dann der Fall eintritt, auf den man gesetzt hat, und die Kugel außerdem nicht auf der Null landet, bekommt man den doppelten Einsatz ausgezahlt. Andernfalls verliert man seinen Einsatz.

Ist es günstiger, zwei Euro auf Gerade oder je einen Euro auf Gerade und auf Schwarz zu setzen ?

Wir überlegen uns zuerst, wie groß der Gewinn (also der ausgezahlte Betrag minus dem Einsatz von zwei Euro) im Mittel ist.

Sei X bzw. Y der Gewinn, sofern man zwei Euro auf Gerade bzw. je einen Euro auf Gerade und auf Schwarz setzt. Dann gilt

sowie

Also erhalten wir bei beiden Strategien im Mittel den gleichen (negativen) Gewinn.

Es stellt sich aber die Frage, ob beidesmal der zufällige Wert gleich stark um den mittleren Wert schankt.

Ein Kriterium zur Beurteilung der zufälligen Schwankung des Resultats eines Zufallsexperiments um den 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.

Im Beispiel oben:

$$\mathbf{P}[X=2]=\frac{18}{37},\quad \mathbf{P}[X=-2]=\frac{19}{37}\quad \text{und}\quad \mathbf{E}X=-\frac{2}{37},$$

woraus folgt

$$\begin{aligned}
V(X) &= \mathbf{E}\left(|X-\mathbf{E}X|^2\right)\\
&= \mathbf{E}\left(|X+\frac{2}{37}|^2\right)\\
&= \left(2+\frac{2}{37}\right)^2\cdot\mathbf{P}[X=2]+\left(-2+\frac{2}{37}\right)^2\cdot\mathbf{P}[X=-2]\\
&= \left(2+\frac{2}{37}\right)^2\cdot\frac{18}{37}+\left(-2+\frac{2}{37}\right)^2\cdot\frac{19}{37}\\
&\approx 3.997.
\end{aligned}$$

Weiter gilt

$$\mathbf{P}[Y=2]=\frac{10}{37},\quad \mathbf{P}[Y=0]=\frac{16}{37},\quad \mathbf{P}[Y=-2]=\frac{11}{37}\quad \text{und}\quad \mathbf{E}Y=-\frac{2}{37},$$

woraus folgt

$$\begin{aligned} V(Y) &= \mathbf{E}\left(|Y+\frac{2}{37}|^2\right) \\ &= \left(2+\frac{2}{37}\right)^2\cdot\frac{10}{37}+\left(0+\frac{2}{37}\right)^2\cdot\frac{16}{37}+\left(-2+\frac{2}{37}\right)^2\cdot\frac{11}{37} \\ &\approx 2.267. \end{aligned}$$

Da der Gewinn im Mittel negativ ist, ist es naheliegend, die erste Strategie mit der stärkeren Schwankung des Gewinnes zu bevorzugen . . .

Beispiel: Sei X eine $b(1,p)$-verteilte Zufallsvariable, d.h. es gilt $\mathbf{P}[X=1]=p$ und $\mathbf{P}[X=0]=1-p$.

Dann gilt $\mathbf{E}X=p$ und damit

$$\begin{aligned} V(X) &= \mathbf{E}\left((X-\mathbf{E}X)^2\right) \\ &= \mathbf{E}\left((X-p)^2\right) \\ &= (1-p)^2\cdot\mathbf{P}[X=1]+(0-p)^2\cdot\mathbf{P}[X=0] \\ &= (1-p)^2\cdot p+p^2\cdot(1-p) \\ &= p-2p^2+p^3+p^2-p^3=p-p^2=p\cdot(1-p). \end{aligned}$$

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

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

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 $\beta \in \mathbb{R}$:

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

Beispiel: Sei X eine $U(0,1)$-verteilte Zufallsvariable, d.h. X sei stetig verteilt mit Dichte

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

Dann gilt

Beispiel: Sei X eine $\exp(\lambda)$-verteilte Zufallsvariable, d.h. X sei stetig verteilt mit Dichte

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

Dann gilt $\mathbf{E}X = \frac{1}{\lambda}$ (vgl. Übungen) und

$$\mathbf{E}X^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \int_0^{\infty} x^2 \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx$$

$$\stackrel{(!)}{=} \frac{2}{\lambda^2},$$

also

$$V(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

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.

Beispiel: Berechnung des Erwartungswertes und der Varianz einer $b(n,p)$-verteilten Zufallsvariablen X.

Bereits gesehen:

Wird ein Zufallsexperiment, bei dem mit Wahrscheinlichkeit p Erfolg und mit Wahrscheinlichkeit $1-p$ Misserfolg eintritt, n-mal unbeeinflusst voneinander durchgeführt, so ist die Anzahl der Erfolge $b(n,p)$-verteilt.

Daher:

$$\mathbf{E}(X) = \mathbf{E}(X_1 + X_2 + \cdots + X_n) \quad \text{und} \quad V(X) = V(X_1 + X_2 + \cdots + X_n)$$

wobei $X_1, \ldots, X_n$ unabhängig $b(1,p)$-verteilt sind.

Mit $\mathbf{E}X_1 = p$ und $V(X_1) = p \cdot (1-p)$ (s.o.) folgt

$$\begin{aligned}
\mathbf{E}(X) &= \mathbf{E}(X_1 + X_2 + \cdots + X_n) \\
&= \mathbf{E}(X_1) + \mathbf{E}(X_2) + \cdots + \mathbf{E}(X_n) \\
&= n \cdot \mathbf{E}(X_1) = n \cdot p
\end{aligned}$$

und

$$\begin{aligned}
V(X) &= V(X_1 + X_2 + \cdots + X_n) \\
&\overset{\text{Unabhängigkeit}}{=} V(X_1) + V(X_2) + \cdots + V(X_n) \\
&= n \cdot V(X_1) = n \cdot p \cdot (1-p).
\end{aligned}$$

Zusammenfassung der Vorlesung am 19.01.2010:

1. Ist X eine relle Zufallsvariable, so beschreibt die sogenannte Varianz von X

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

die mittlere quadratische Schwankung von X um seinen Erwartungswert.

2. Gemäß den bisher bereits gelernten Rechenregeln für Erwartungswerte gilt

$$V(X) = \begin{cases} \sum_{k=0}^{\infty}(k - \mathbf{E}X)^2 \cdot \mathbf{P}[X = k] & \text{falls} \quad \mathbf{P}_X(\mathbb{N}_0) = 1, \\ \int_{-\infty}^{\infty}(x - \mathbf{E}X)^2 \cdot f(x)\,dx & \text{falls} \quad X \text{ die Dichte } f : \mathbb{R} \to \mathbb{R} \text{ hat.} \end{cases}$$

3.

	$b(n,p)$	$\pi(\lambda)$	$U(a,b)$	$\exp(\lambda)$	$N(a,\sigma^2)$
Erwartungswert	$n \cdot p$	λ	$(a+b)/2$	$1/\lambda$	a
Varianz	$n \cdot p \cdot (1-p)$	λ	$(b-a)^2/12$	$1/\lambda^2$	σ^2

Lernziele der Vorlesung am 26.01.2010

Nach dieser Vorlesung sollten Sie

1. wissen, wie man Erwartungswert und Varianz einer unbekannten Verteilung schätzt,

2. wissen, was man unter Erwartungstreue bzw. Konsistenz eines Punktschätzverfahren versteht und wie man diese nachweist,

3. die Aussage des zentralen Grenzwertsatzes kennen.

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!

Beispiele:

1. Um festzustellen, inwiefern Examenskandidaten in der Lage sind, ihre eigene Leistungsfähigkeit einzuschätzen, wurde 15 Kandidaten eine Klausur mit 70 Aufgaben gestellt. Nach Bearbeitung der Klausur wurden die Kandidaten gebeten, die Anzahl der richtig gelösten Aufgaben zu schätzen, und nach der Korrektur der Klausur wurden die geschätzen Anzahlen der gelösten Aufgaben mit den tatsächlichen Anzahlen verglichen.

2. Im Rahmen einer Studie wurden bei 396 Studenten approximativ die Anzahl der gesprochenen Worte über einen Zeitraum von mehreren Tagen bestimmt.

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 ?

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

Ziel der Analyse der Daten:

Informal: Aussagen über das Prinzip, nach dem die Daten erzeugt werden, z.B.

- Wie groß sind die Werte "im Mittel" ?
- Wie stark schwanken die Werte um ihren "mittleren Wert" ?

Formal: Aussagen über die Verteilung $\mathbf{P}_{X_1}$ der Zufallsvariablen, z.B.

- Wie groß ist der Erwartungswert $\mathbf{E}X_1$?
- Wie groß ist die Varianz $V(X_1)$?

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" 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 $T_n(x_1, \ldots, x_n)$ heißt konsistente Schätzung für $\mathbf{E}X_1$, falls gilt

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

a) Eine Schätzung $T_n(x_1, \ldots, x_n)$ heißt erwartungstreue Schätzung für $\mathbf{E}X_1$, falls gilt

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

Analog: Konsistente bzw. erwartungstreue Schätzung für $V(X_1)$ . . .

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:

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

Anschaulich: Mit Wahrscheinlichkeit Eins nähern sich die Werte von Z_n mit wachsendem n immer mehr dem Wert von Z an.

Man kann zeigen:

Mit der fast sicheren Konvergenz kann man rechnen wie mit reellen Zahlenfolgen . . .

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 konsistente Schätzung für $\mathbf{E}X_1$, denn es gilt:

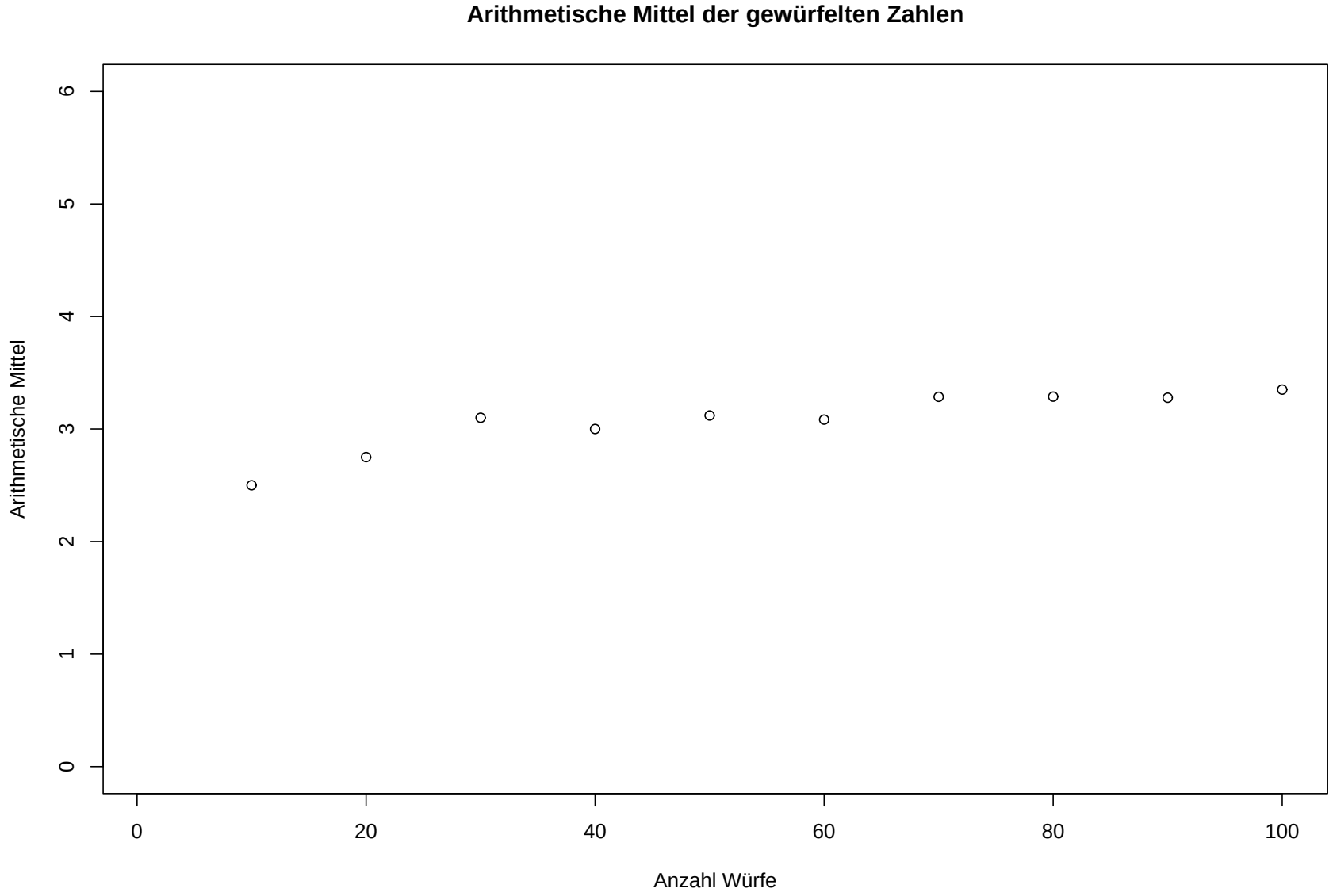
Satz (Starkes Gesetz der großen Zahlen):

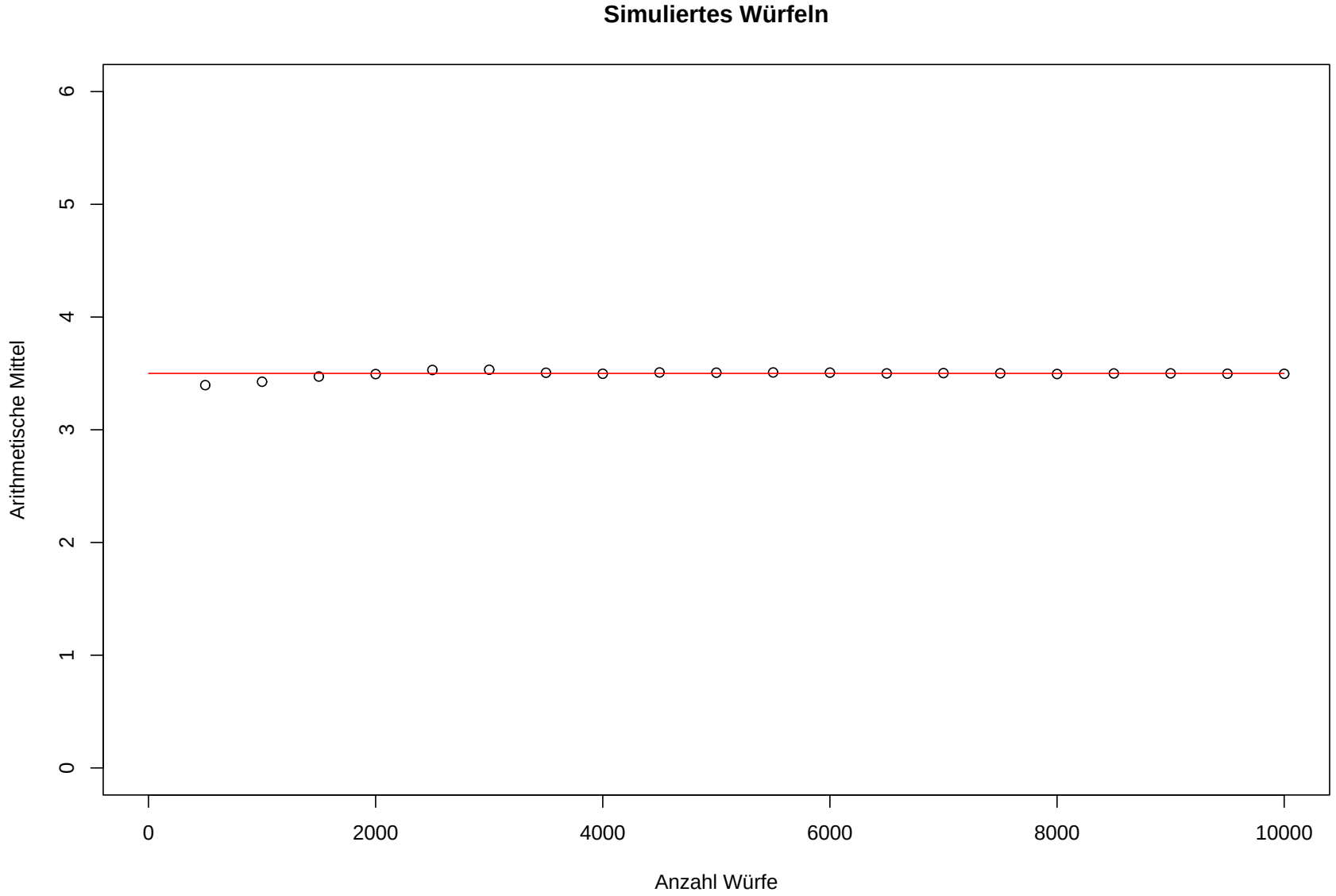
Sind die auf dem selben Wahrscheinlichkeitsraum definierten reellen Zufallsvariablen $X_1, X_2, \ldots$ unabhängig und identisch verteilt, und existiert $\mathbf{E}X_1$, so gilt:

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

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.


Arithmetische Mittel der gewürfelten Zahlen
Arithmetische Mittel
0
1
2
3
4
5
6
0
20
40
60
80
100
Anzahl Würfe


Simuliertes Würfeln
Arithmetische Mittel
0
1
2
3
4
5
6
0
2000
4000
6000
8000
10000
Anzahl Würfe

Auch unsere Schätzung für die Varianz ist 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 Die Bedeutung der Normalverteilung

Häufig verwendet man die sogenannte Normalverteilung bei der Modellierung von Daten in der Praxis.

Grund ist der sogenannte **zentrale Grenzwertsatz**:

Dieser besagt anschaulich, dass Summen bestehend aus vielen zufälligen Werten, die unbeeinflusst voneinander nach dem gleichen Prinzip erzeugt wurden, sich approximativ wie eine normalverteilte Zufallsvariable verhalten.

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

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

annähernd $N(0,1)$- verteilt.

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

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

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

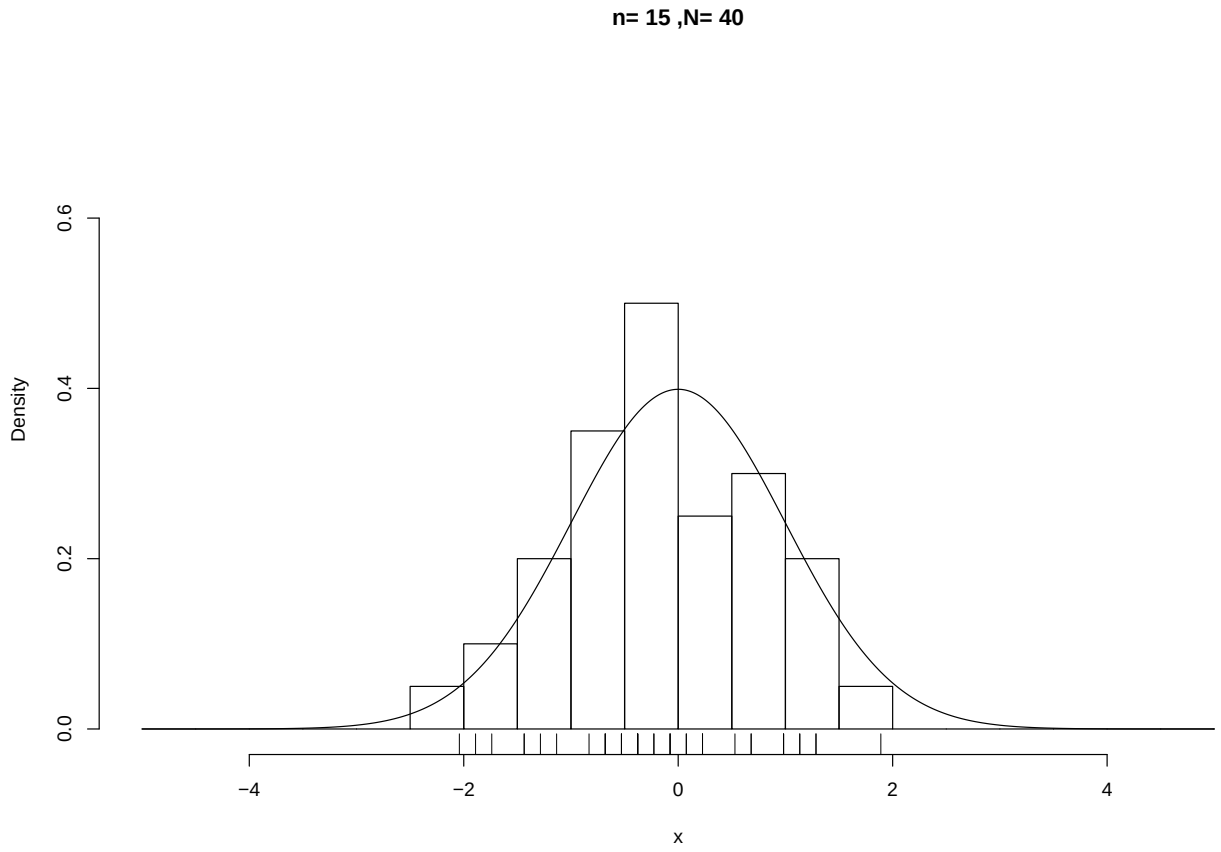
$$\begin{aligned} \mathbf{E}X_1 &= \sum_{i=1}^{6} i \cdot \mathbf{P}[X_1 = i] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = 3.5, \\ V(X_1) &= \mathbf{E}(X_1^2) - (\mathbf{E}X_1)^2 = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} - (3.5)^2 = \frac{35}{12}. \end{aligned}$$

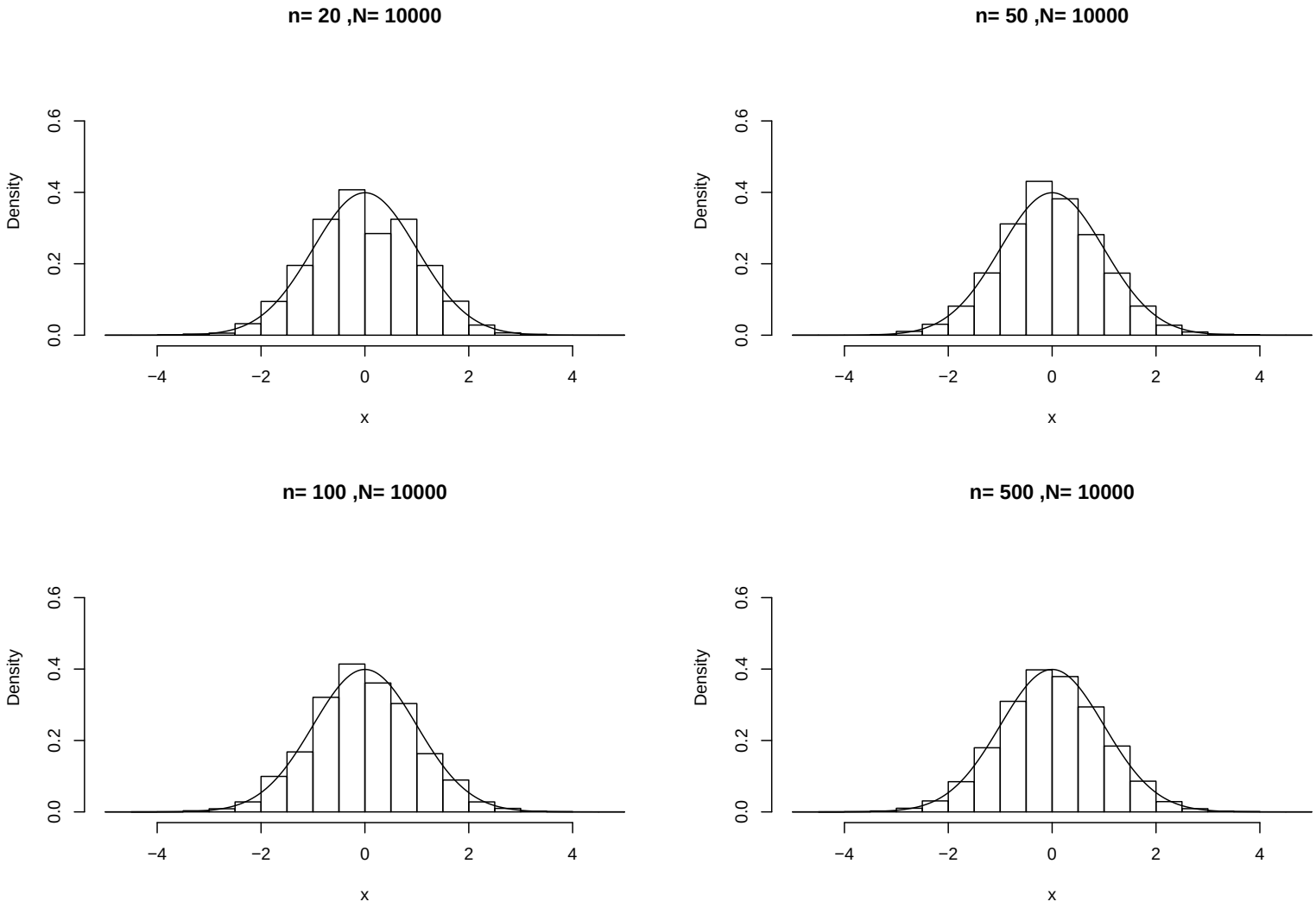
Nach dem zentralen Grenzwertsatz verhält sich also

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

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.




n= 20 ,N= 10000
n= 50 ,N= 10000
n= 100 ,N= 10000
n= 500 ,N= 10000
Density
x
0.0
0.2
0.4
0.6
−4
−2
0
2
4

Zusammenfassung der Vorlesung am 26.01.2010:

1. Sinnvolle Schätzer für Erwartungswert bzw. Varianz einer unbekannten Verteilung sind das empirische arithmetische Mittel bzw. die empirische Varianz.

2. Diese sind konsistent (d.h. mit wachsendem Stichprobenumfang nähert sich der Schätzer immer mehr dem zu schätzenden Wert an) und erwartungstreu (d.h. bei festem Stichprobenumfang ergibt sich im Mittel der zu schätzende Wert).

3. In der Praxis modelliert man Daten häufig mit Hilfe der sogenannten Normalverteilung. Nach dem zentralen Grenzwertsatz verhalten sich Summen bestehend aus vielen zufälligen Werten, die unbeeinflusst voneinander nach dem gleichen Prinzip erzeugt wurden, approximativ normalverteilt.

Lernziele der Vorlesung am 02.02.2010

Nach dieser Vorlesung sollten Sie

1. erläutern können, was ein Test zum Niveau α ist, und wann wir einen solchen Test als optimal ansehen,

2. das Resultat eines statistischen Tests interpretieren können,

3. den einseitigen Gauß-Test kennen und anwenden können.

5.3 Statistische Testverfahren, Teil I

5.3.1. Beispiel: Schätzen Examenskandidaten ihre eigene Leistungsfähigkeit eher zu gut oder eher zu schlecht ein ?

$n = 15$ 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

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

gebildet.

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

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: Wir fassen wir unsere Daten als Realisierungen $x_1, \ldots, x_{15}$ von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots X_{15}$ auf.

3. Wir formulieren unsere Frage so um, dass sie nur von der zugrunde liegenden Verteilung abhängt: Im Beispiel oben wollen wir wissen, welche von den beiden Hypothesen H_0 bzw. H_1 zutrifft, wobei:

$$H_0: \quad \mathbf{E}X_1 \leq 0 \quad \text{(sog. \textbf{Nullhypothese})},$$
$$H_1: \quad \mathbf{E}X_1 > 0 \quad \text{(sog. \textbf{Alternativhypothese})}.$$

4. Um die Fragestellung zu vereinfachen, machen wir Annahmen über die Art der in dem Beispiel auftretenden Verteilung:

   Wir gehen im Folgenden davon aus, dass die auftretende Verteilung eine **Normalverteilungen** mit **unbekanntem Erwartungswert** und **bekannter oder unbekannter Varianz** ist.

5. Unter diesen Annahmen ermitteln wir geeignete Verfahren, mit Hilfe derer wir uns (mit kontrollierter Fehlerwahrscheinlichkeit) zwischen den beiden Hypothesen entscheiden können.

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 liegende Verteilung, z.B. Hypothesen wie

$$H_0: \quad \mathbf{E}X_1 \leq 0,$$
$$H_1: \quad \mathbf{E}X_1 > 0.$$

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

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

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

Bezeichnung für die auftretenden Fehler:

- Gilt H_0 (die sogenannte Nullhypothese), liefert unser Test aber fälschlicherweise $\varphi(x_1, \ldots, x_n) = 1$ 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 $\varphi(x_1, \ldots, x_n) = 0$ 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 $H_0 : \mathbf{E}X_1 \leq 0$ versus $H_1 : \mathbf{E}X_1 > 0$) sind die **Fehlerwahrscheinlichkeiten erster Art** eines Tests φ gegeben durch

$$\mathbf{P}_{\mathbf{E}X_1=\mu}\left[\varphi(X_1, \ldots, X_n) = 1\right] \qquad \textbf{mit } \mu \leq 0,$$

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

$$\mathbf{P}_{\mathbf{E}X_1=\mu}\left[\varphi(X_1, \ldots, X_n) = 0\right] \qquad \textbf{mit } \mu > 0.$$

Wünschenswert: Konstruiere einen statistischen Test, 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 $\alpha > 0$ 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 einseitige Gauß-Test für eine Stichprobe

1. Fragestellungen

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 $\sigma_0^2 > 0$.

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.

2. Grundidee

(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, $H_0 : \mu \leq \mu_0$ abzulehnen, falls $\frac{1}{n}\sum_{i=1}^{n} X_i$ "sehr viel größer" als 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.

3. 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: $\mathbf{P}[Z > u_\alpha] = \alpha$.

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

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

gebildet.

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

Wir gehen vereinfachend davon aus, dass die Varianz durch $\sigma_0^2 = s^2 = 61.7$ gegeben ist, und führen einen einseitigen Gauß-Tests für $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ zum Niveau $\alpha = 0.05$ durch.

Hierbei gilt: $u_\alpha = u_{0.05} \approx 1.64$

Wir erhalten

$$\frac{\sqrt{n}}{\sigma_0} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) = \frac{\sqrt{15}}{\sqrt{61.7}} \cdot (6.4 - 0) \approx 3.16 > u_{0.05},$$

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

Resultat: Examenskandidaten schätzen die eigene Leistungsfähigkeit eher zu schlecht ein.

Zusammenfassung der Vorlesung am 02.02.2010

1. Ein statistischer Test ist eine Entscheidungsvorschrift zwischen zwei Hypothesen (Nullhypothese und Alternativhypothese).

2. Bei einem Test zum Niveau α ist die Wahrscheinlichkeit, dass sich der Test bei Vorliegen der Nullhypothese für die Alternativhypothese entscheidet, immer kleiner oder gleich α. Bei einem optimalen Test zum Niveau α ist gleichzeitig die Wahrscheinlichkeit für die fälschliche Entscheidung für die Nullhypothese so klein wie möglich.

3. Beim einseitigen Gauß-Test liegt eine Stichprobe einer Normalverteilung mit unbekanntem Erwatungswert μ und bekannter Varianz vor. Die Hypothese $H_0 : \mu \leq \mu_0$ wird abgelehnt, sofern das arithmetische Mittel der Beobachtungen groß ist. Dabei wird die Schranke für die Ablehnung so gewählt, dass für $\mu = \mu_0$ ein Überschreiten der Schranke genau mit Wahrscheinlichkeit α auftritt.

Lernziele der Vorlesung am 09.02.2010

Nach dieser Vorlesung sollten Sie

1. sollten Sie wissen, was ein ein- bzw. zweiseitiges Testproblem und was ein Einstichprobenproblem bzw. ein Zweistichprobenproblem ist,

2. den Unterschied zwischen einem Gauß-Test und einem t-Test kennen und beide in den verschiedenen Situationen anwenden können.

5.4 Statistische Testverfahren, Teil II

5.4.1. **Beispiel:** Sprechen Frauen mehr als Männer ?

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 ?

5.4.2. 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. Im Beispiel oben fassen wir die 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: Wir wollen wissen, welche von den beiden Hypothesen

$$H_0: \quad \mathbf{E}X_1 = \mathbf{E}Y_1 \quad \text{(sog. \textbf{Nullhypothese})}$$
$$H_1: \quad \mathbf{E}X_1 \neq \mathbf{E}Y_1 \quad \text{(sog. \textbf{Alternativhypothese})}$$

   zutrifft.

Prinzipieller Unterschied zur letzten Vorlesung:

- Im Beispiel der letzten Vorlesung hatten wir **eine** Stichprobe $x_1, \ldots x_{15}$ der Verteilung von X_1 gegeben, und wollten wissen, ob $H_0 : \mathbf{E}X_1 \leq 0$ oder $H_1 : \mathbf{E}X_1 > 0$ gilt. Da nur eine Stichprobe vorliegt, sprechen wir hierbei von einem sog. **Einstichprobenproblem**, und da bei H_1 Abweichungen von der Null nur in eine Richtung betrachtet werden, handelt es sich hierbei um ein sog. **einseitiges Testproblem**.

- Im obigen Beipiel 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 $H_0 : \mathbf{E}X_1 = \mathbf{E}Y_1$ oder $H_1 : \mathbf{E}X_1 \neq \mathbf{E}Y_1$ gilt. Da Stichproben von zwei Verteilungen vorliegen, sprechen wir jetzt von einem sog. **Zweistichprobenproblem**, und da bei H_1 Abweichungen der Erwartungswerte in beide Richtungen betrachtet werden, handelt es sich jetzt um ein sog. **zweiseitiges Testproblem**.

4. Um die Fragestellung zu vereinfachen, machen wir Annahmen über die Art der in dem Beispiel auftretenden Verteilung:

   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, mit Hilfe derer wir uns (mit kontrollierter Fehlerwahrscheinlichkeit) zwischen den beiden Hypothesen entscheiden können.

5.4.3 Der zweiseitige Gauß-Test für zwei Stichproben

1. Fragestellung
   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 $\sigma_0^2 > 0$.

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

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

   entscheiden.

2. Grundidee beim zweiseitigen 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) Wir lehnen $H_0 : \mu_X = \mu_Y$ ab, falls $\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j$ "weit entfernt" von 0 ist.

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

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

$N(0,1)$-verteilt ist, da diese Zufallsvariable analog zur letzten Vorlesung normalverteilt ist und Erwartungswert Null und Varianz Eins hat.

3. 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.

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 beobachteten Daten:

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

Wir führen einen zweiseitigen Gauß-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, wobei wir

$$u_{\alpha/2} = u_{0.025} \approx 1.97$$

verwenden und die Varianz als bekannt voraussetzen müssen.

Dazu schätzen wir die Varianz durch die die sogenannte gepoolte Stichprobenvarianz

$$\begin{aligned}\hat{\sigma}^2_{X,Y} &= \frac{\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2 + \sum_{i=1}^{m}(Y_i - \frac{1}{m}\sum_{j=1}^{m} Y_j)^2}{n+m-2} \\ &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},\end{aligned}$$

was eine erwartungstreue Schätzung der Varianz ist, da:

$$\mathbf{E}\hat{\sigma}^2_{X,Y} = \frac{(n-1)\mathbf{E}(s_X^2) + (m-1)\mathbf{E}(s_Y^2)}{n+m-2} = \frac{(n-1)\sigma_0^2 + (m-1)\sigma_0^2}{n+m-2} = \sigma_0^2.$$

Also verwenden wir im Folgenden

$$\sigma_0 = \sqrt{\frac{(n_x-1)\cdot s_x^2 + (n_y-1)\cdot s_y^2}{n_x+n_y-2}} \approx 7970.$$

Für die beobachteten Daten erhalten wir

$$\frac{\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{j=1}^{m} y_j\right|}{\sqrt{\frac{1}{n}+\frac{1}{m}}\cdot\sigma_0}$$

$$= \frac{|16215 - 15669|}{\sqrt{\frac{1}{210}+\frac{1}{186}}\cdot 7970}$$

$$\approx 0.68 < u_{\alpha/2},$$

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

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

5.4.4 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 $V(X_1)$.

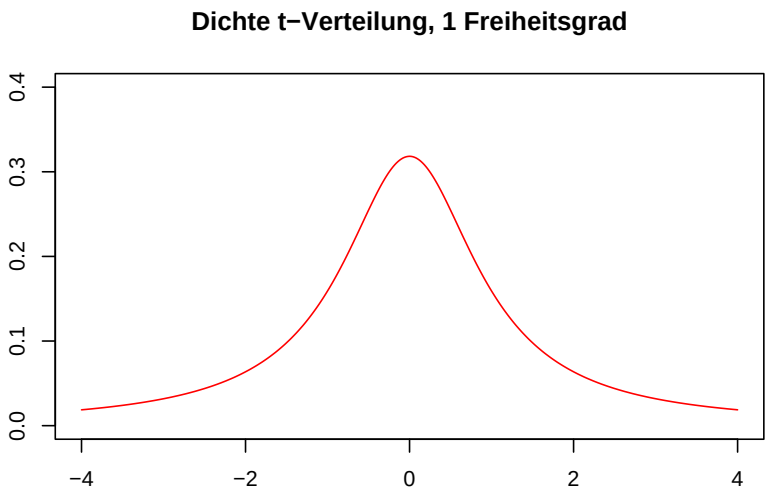
Zur Einhaltung des Niveaus beachten wir:

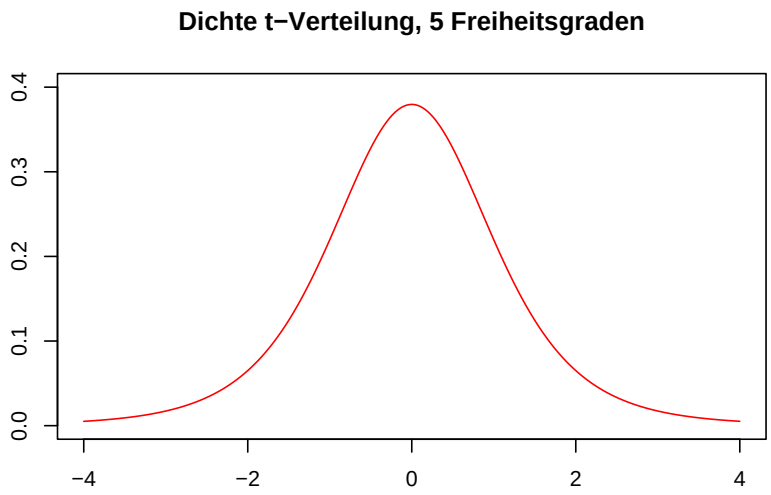
Sind $X_1, \ldots, X_n$ unabhängig $N(\mu_0, \sigma^2)$-verteilt, so ist

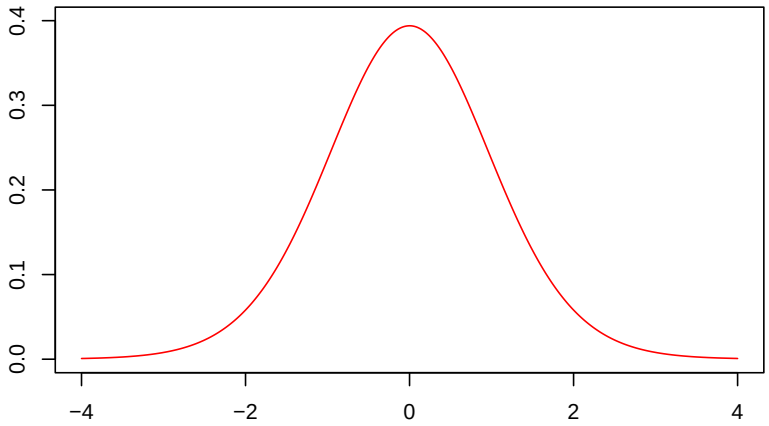
$$\frac{\sqrt{n}}{S_X} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu_0 \right)$$

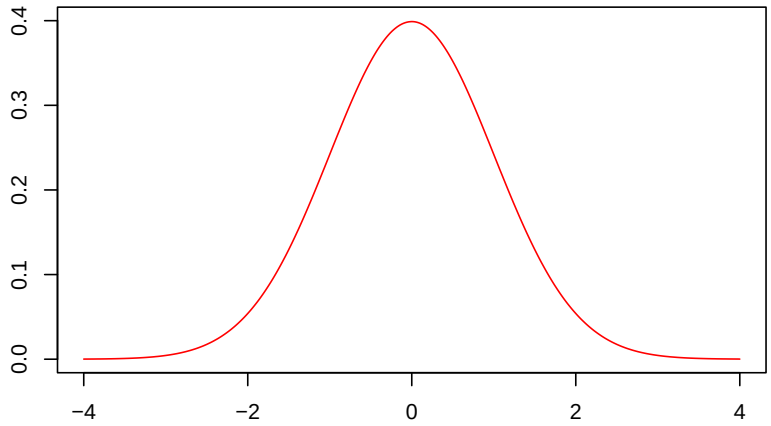
nicht länger $N(0,1)$-verteilt, sondern **t-verteilt mit $n-1$-Freiheitsgraden.**

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


Dichte t−Verteilung, 1 Freiheitsgrad
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4


Dichte t−Verteilung, 5 Freiheitsgraden
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4


Dichte t−Verteilung, 20 Freiheitsgraden
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4


Dichte von N(0,1)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

Beispiel: Einseitiger 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

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

H_0 wird abgelehnt, falls

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

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

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

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

gebildet.

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

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

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

Wir erhalten

$$\frac{\sqrt{n}}{s_x} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) = \frac{\sqrt{15}}{\sqrt{61.7}} \cdot (6.4 - 0) \approx 3.16 > t_{14,0.05},$$

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

Resultat: Examenskadidaten schätzen ihre eigene Leistungsfähigkeit eher zu schlecht ein.

Der zweiseitige 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

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

Problem: Wie schätzen wir diesmal die Varianz ?

Schätzung der Varianz:

Wir verwenden wieder die sogenannte gepoolte Stichprobenvarianz

$$\begin{aligned}\hat{\sigma}^2_{X,Y} &= \frac{\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)^2 + \sum_{i=1}^{m}(Y_i - \frac{1}{m}\sum_{j=1}^{m} Y_j)^2}{n+m-2} \\ &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},\end{aligned}$$

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

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

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

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

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 beobachteten 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

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

$$= \frac{|16215 - 15669|}{\sqrt{\frac{1}{210}+\frac{1}{186}}\cdot 7970}$$

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

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 von der bei Studenten unterscheidet.

Zusammenfassung der Vorlesung am 09.02.2010

1. In Anwendungen ist meist die Varianz der Daten unbekannt, daher wird statt dem sogenannten Gauß-Test der t-Test angewendet, bei dem statt der Varianz eine Schätzung derselben verwendet wird, und die Testgröße statt mit Fraktilen der Normalverteilung mit Fraktilen der sogenannten t-Verteilung verglichen wird.

2. Bei dem behandelten Zweistichprobenproblem werden die Erwartungswerte zweier normalverteilter Stichproben mit gleicher Varianz verglichen. Der Test hängt dabei von der Differenz der arithmetischen Mittel der beiden Stichproben ab, die geeignet normalisiert wird.