Mathematik und Statistik für BiologInnen

Vorlesung WS 2007/08

Prof. Dr. Michael Kohler
Fachbereich Mathematik
Technische Universität Darmstadt

kohler@mathematik.tu-darmstadt.de

Ein zutreffender Titel für diese Veranstaltung ist:

Mathematik und Statistik für BiologInnen

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

Brad Efron

Kapitel 1: Motivation

Statistik – wozu braucht man das ?

1.1 Statistik-Prüfung, Sommer 2002

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

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

StudentInnenen hatten die Möglichkeit, freiwillig einen Übungsschein zu erwerben.

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

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

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

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

1.2 Sex und Herzinfarkt

Studie von Prof. Shah Ebrahim, Universität Bristol:

2400 gesunde Männer wurden unter anderem zu ihrem Sexualleben befragt und über einen Zeitraum von 10 Jahren beobachtet.

Resultat:

Bei drei bis vier Orgasmen pro Woche sinkt das Infarktrisiko um mehr als die Hälfte.

Was folgt daraus ?

1.3 Die Challenger-Katastrophe

Start der Raumfähre Challenger am 28. Januar 1986:

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

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

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

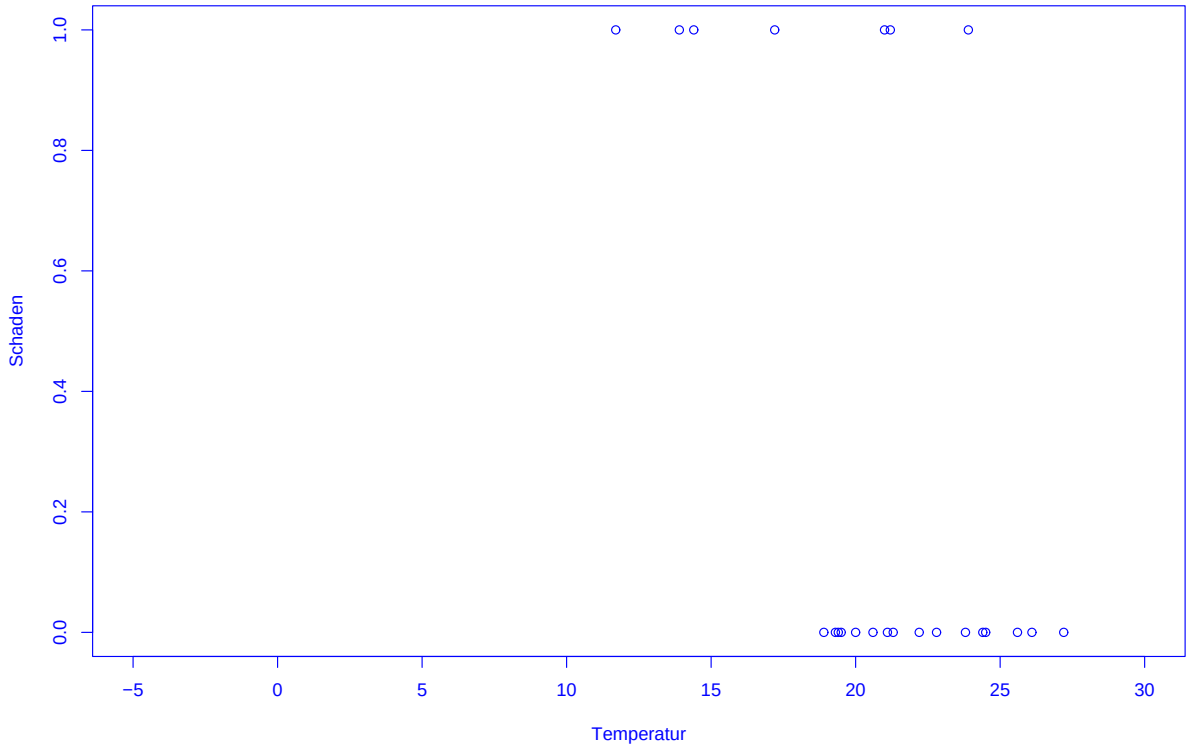
Zur Begründung verwendete Daten:

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

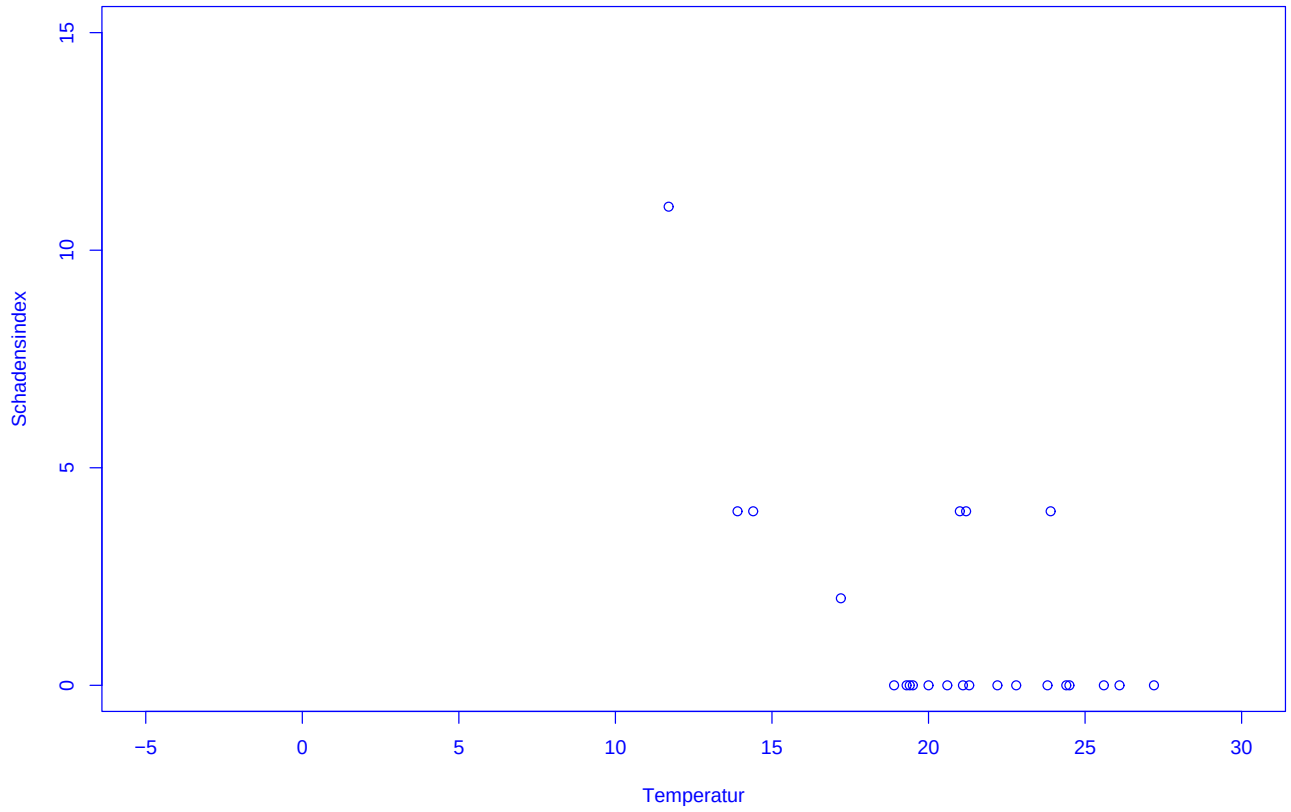
War für NASA leider nicht nachvollziehbar ...

Probleme bei der Analyse dieser Daten:

1. Flüge ohne Schädigungen nicht berücksichtigt.



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



1.4 Genetischer Fingerabdruck

Sehr erfolgreiches Hilfsmittel bei der Aufklärung von Kapitalverbrechen.

Am Tatort gefundenes DNA Material wird mit dem eines Verdächtigen verglichen.

- Falls Übereinstimmung: Verdächtiger = Täter
- Falls keine Übereinstimmung: Verdächtiger $\neq$ Täter

Wozu benötigt man da Statistik ?

DNA $\approx$ lange Kette bestehend aus $\geq 1.000.000$ Mononukleotiden.

Falls alle Mononukleotide übereinstimmen, stammt das Material vom gleichen Menschen.

Aber: Man kann nicht 1.000.000 Mononukleotide auf Übereinstimmung vergleichen.

Ausweg: Vergleiche nur (kurze) Sequenz von Mustern in der DNA

Falls keine Übereinstimmung: Verdächtiger $\neq$ Täter

Aber was tun bei Übereinstimmung ?

Falls Übereinstimmung:

Schätze, wie häufig eine Übereinstimmung auftritt, wenn man Menschen zufällig auswählt und ihre DNA mit dem Material des Täters vergleicht.

Falls dies nur selten der Fall ist, schließe, dass Verdächtiger der Täter ist.

Problem:
Aus welcher Menge von Menschen wird ausgewählt:

- gesamte Menschheit ?
- Großfamilie ?
- abgeschiedenes Dorf ?

1.5 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 der **Biologie** ?

Zu den (klassischen) Arbeitsmethoden der Biologie gehören **strukturiertes Beobachten, Dokumentation** und das **Überprüfen formulierter Hypothesen und Theorien durch Experimente**.

Hier ist die **beschreibende Statistik** bei der übersichtlichen und kompakten Darstellung von erhobenen Daten hilfreich, während die **schließende Statistik** Verfahren bereit stellt, um aus den erhobenen Daten Rückschlüsse auf die Grundgesamtheit zu ziehen.

Beispiel: Analyse des Wachstumsverhaltens von Krebsen.

In den modernen Teilgebieten der Biologie muss eine Fülle von molekularbiologischen Daten analysiert werden.

Z.B. in der **Systembiologie:**

Erforschung der **Wechselwirkung zwischen Proteinen und regulatorischen Faktoren** mit Hilfe von DNA-Microarray-Daten.

Dies ist ohne moderne statistische Verfahren undenkbar !

Schön, aber:

Braucht man den Stoff dieser Vorlesung wirklich im weiteren Studium der Biologie in Darmstadt ?

JA, z.B.

- **"Beschreibende Statistik"** in Vorlesungen zur **Zoo-Ökologie**,
- **"Grundbegriffe über W-Verteilungen"** in der Vorlesung **Quantitative Molekularbiologie**,
- **"Schließende Statistik"** im Rahmen von **Forschungspraktika**,

⋮

FAZIT:
Statistik hat vielfältige Anwendungen in der Biologie 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
- Kapitel 3: Beschreibende Statistik
- Kapitel 4: Einführung in die W-Theorie
- Kapitel 5: Schließende Statistik

Zum Niveau dieser Vorlesung:

Verschiedene Ebenen des **“Lernens”**:

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

Ziel der Ausbildung an der Universität ist die letzte Ebene.

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

Zum didaktischen Konzept dieser Vorlesung:

Lehr-Lern-Kurzschluss:

Gelernt wird nicht, was gelehrt wird!

Was ich hier mache:

Bereitsstellung einer "Umgebung", in der **Sie** möglichst einfach möglichst viel über W-Theorie und Statistik **lernen können**.

Spezielle Tricks"dabei:

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

und ganz wichtig:

Motivierung der StudentInnen !

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

AKTIV AN DIESER VERANSTALTUNG TEILNEHMEN, d.h.

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

Zur Selbstkontrolle wird der Erwerb des Übungsscheines empfohlen.

Kapitel 2: Erhebung von Daten

2.1 Kontrollierte Studien

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

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

Vorgehen in drei Phasen üblich:

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

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

Vorgehen 1: Retrospektiv kontrollierte Studie

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

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

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

Vorgehen 2: Prospektiv kontrollierte Studie ohne Randomisierung

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

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

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

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

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

Nachteil: Fehleranfällig !

Vorgehen 3: Prospektiv kontrollierte Studie mit Randomisierung

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

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

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

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

a) und b): doppelte Blindstudie

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

Ergebnis der Studie:

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

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

Lohnt sich der Aufwand einer
prospektiv kontrollierten Studie mit Randomisierung ?

Beispiel: Abnahme der Pinguinpopulation am Südpol durch die Erderwärmung ?

Um die Veränderung der Pinguinpopulation am Südpol zu analysieren, wurden über Jahre hinweg Pinguine mit Gummiringen markiert und dann fortlaufend die Zahl der markierten Tiere gezählt. Dabei ergab sich, dass die Anzahl der gezählten Pinguine im Laufe der Zeit abnahm, was auf die Erderwärmung zurückgeführt wurde.

Aber: Als im Rahmen einer kontrollierten Studie mit Randomisierung 100 Pinguine mit einem Chip markiert wurden und dann von diesen die Hälfte auch noch mit einem Gummiring gekennzeichnet wurden stellte sich heraus:

Von den mit Gummiringen markierten Tieren überlebten weniger den Winter als von denen ohne Gummiringe ...

Beispiel: Studie zur Überprüfung der Wirksamkeit eines Impfstoffes gegen Polio (USA, 1954).

Prospektiv kontrollierte Studie mit Randomisierung ergab:

	Größe	# Fälle	Infektionsrate
SG	200.000	56	28
KG	200.000	142	71
KZdE	350.000	161	46

- SG = Studiengruppe, KG = Kontrollgruppe
- KZdE = Gruppe bestehend aus allen Kindern, bei denen die Eltern der Impfung nicht zugestimmt haben
- Infektionsrate = Anzahl Polio–Fälle pro 100.000 Kinder.

Beispiel: Studien über Bypass–Operationen am Herzen

Resultate von 8 prospektiv kontrollierten Studien mit Randomisierung:

positiv: 1

negativ: 7

Resultate von 21 retrospektiv kontrollierten Studien:

positiv: 16

negativ: 5

Woher kommt der Unterschied ?

Bei 6 der prospektiv kontrollierten Studien und bei 9 der retrospektiv kontrollierten Studien wurde die Überlebensrate nach 3 Jahren angegeben. Resultat:

	prospektiv, randomisiert	retrospektiv
Operation	87.6 %	90.9 %
keine Operation	83.2 %	71.1 %

Überlebensraten bei den operierten Patienten ungefähr gleich, bei den nicht operierten Patienten aber bei den retrospektiven Studien viel geringer als bei den prospektiven Studien.

Grund: Für die Operation (und die KG bei prospektiven Studien !) kommen nur nicht zu kranke Patienten in Frage.

2.2 Beobachtungsstudien

Unterschied zu kontrollierten Studien:

Kontrollierte Studie (auch: geplanter Versuch):

Untersucht wird Einfluss einer Einwirkung (z.B. Impfung) auf Objekte (z.B. Kinder). Statistiker kann entscheiden, auf welche Objekte eingewirkt wird und teilt die Objekte entsprechend in SG und KG ein.

Beobachtungsstudie:

Die Objekte entscheiden selbst, ob sie zur SG oder KG gehören.

Hauptproblem bei Beobachtungsstudien:

Ist die KG wirklich ähnlich zur SG ?

Beispiel 1: 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 konfundierter Faktor.

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

Vergleiche

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

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

Nötig daher:

- Erkennung aller konfundierten Faktoren
- Bildung von vielen Untergruppen

Beispiel 2: Beeinflusst Ultraschall das Geburtsgewicht von Kindern ?

Beobachtungsstudie am John Hopkins Krankenhaus, Baltimore:

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

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

Aber: Kontrollierte Studie mit Randomisierung ergab:

Geburtsgewicht nach Ultraschalluntersuchung sogar etwas höher als ohne Ultraschalluntersuchung.

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

Beispiel 3: Diskreminierung von Frauen bei der Zulassung zum Studium

Zulassungsdaten Universität Berkeley, Herbst 1973:

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

Folgt daraus, dass die Uni Berkely Frauen diskreminiert ?

Zulassungsdaten nach Fächern getrennt:

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

Folgerung:

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

Problem bei Studien:

Die Mehrzahl obiger Studien weist Assoziation aber nicht Kausalität nach.

Grund:

Existenz konfundierter Faktoren.

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

2.3 Umfragen

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

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

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

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

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

Problem bei Wahlumfragen: Befragung aller Wahlberechtigten zu aufwendig.

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

Fragen:

1. Wie wählt man die Stichprobe ?

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

Mögliche Antwort im Beispiel oben:

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

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

Wir werden in Kapitel 5 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: Durchführung von Wahlumfragen (USA):

1. USA wird gemäß Zeitzone und Bevölkerungsdichte unterteilt.

2. Für jeden Teil wird eine Umfrage mit Hilfe von zufälliger Wahl von Telefonnummern durchgeführt.

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

4. Gewichte berücksichtigen demographische Aspekte.

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: Größen (in mm) von 362 im Jahr 1982 in Kalifornien gefangenen Weibchen des Kalifornischen Taschenkrebs (Cancer Magister):

143.9, 153.8, 144.0, 163.2, 149.3, 140.2, 155.3, 138.9, 153.7, 163.0, 157.1, 132.8, 157.5, 139.1, 155.7, 115.5, 133.3, 144.2, 148.7, 137.7, 144.8, 161.1, 119.6, 139.5, 153.3, 153.4, 139.5, 143.2, 126.7, . . .

Beispiel 2: Alter der ca. 82 Millionen Einwohner Deutschlands im Jahr 2001 (Angabe in Jahren):

79, 4, 34, 60, . . .

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.

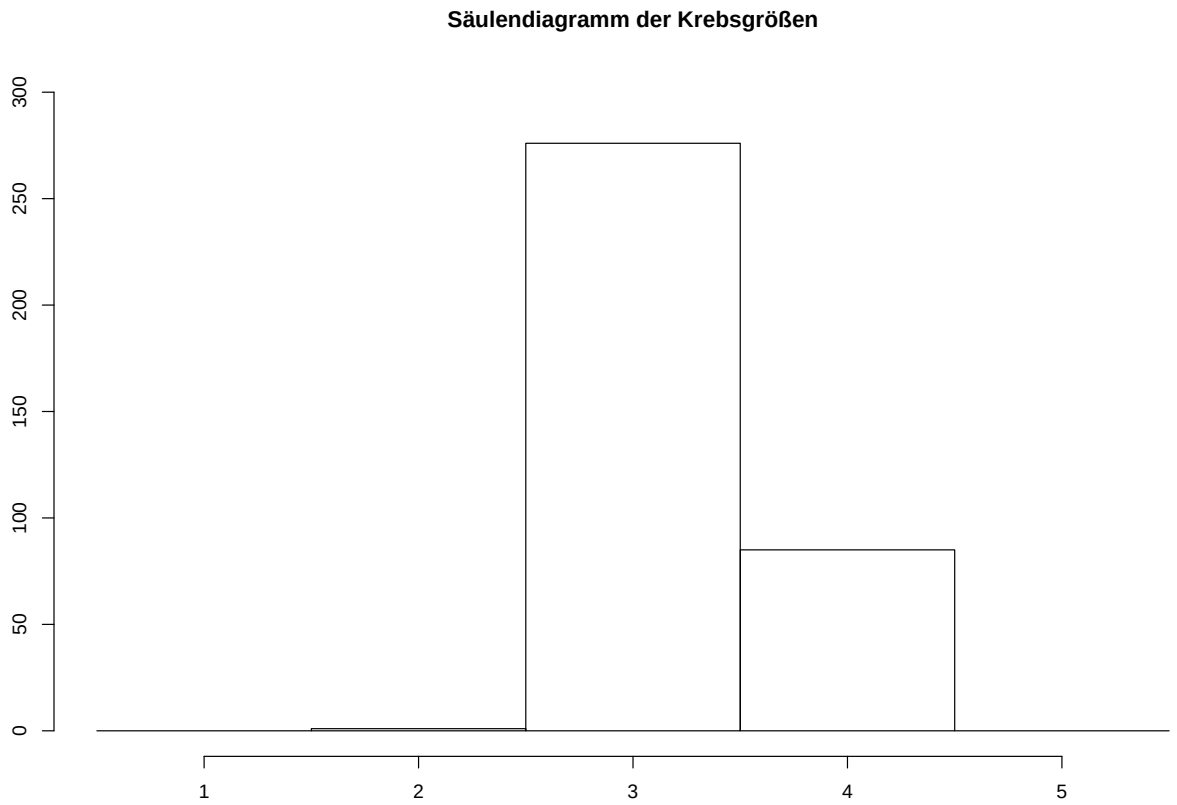
Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

In Beispiel 1 oben (Größen von gefangenen Weibchen des Kalifornischen Taschenkrebs):

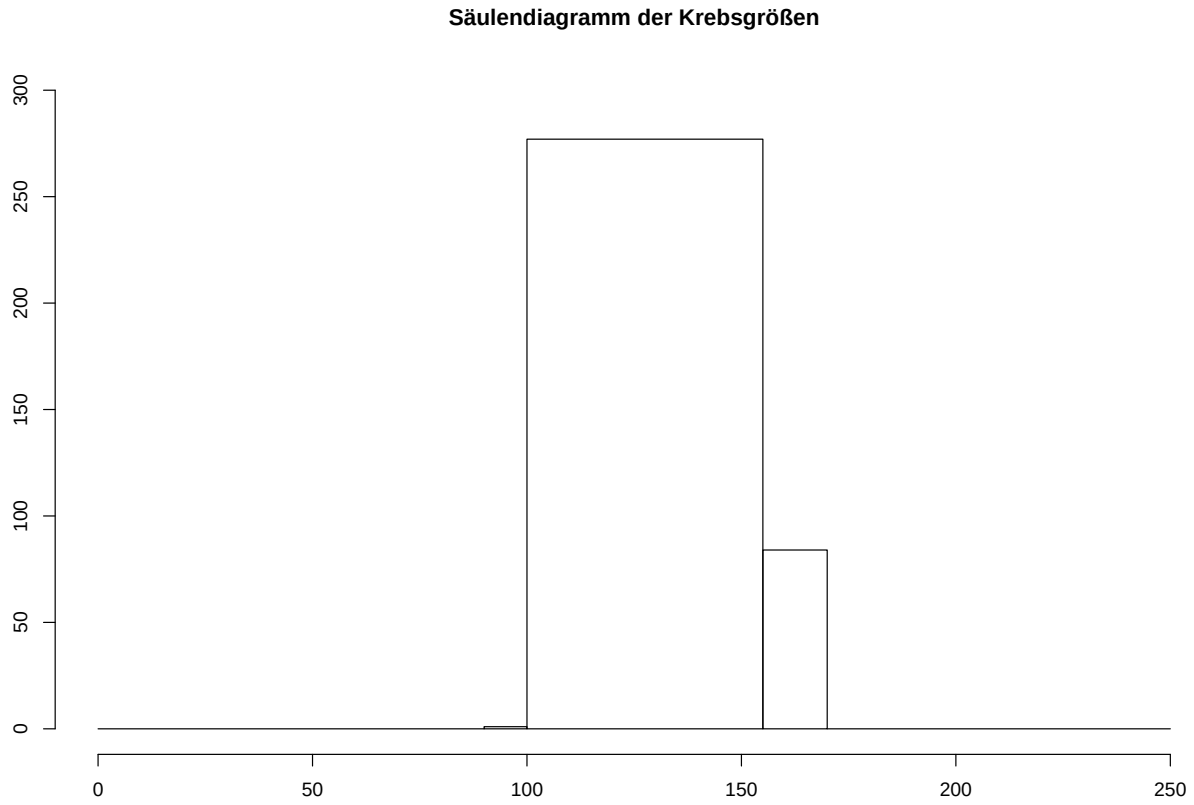
Unterteilung in 5 Klassen ergibt

Klasse	Altersgruppe	Häufigkeit (in 1000)
1	$[0, 90)$	0
2	$[90, 100)$	1
3	$[100, 155)$	276
4	$[155, 170)$	85
5	$[170, 250)$	0

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:

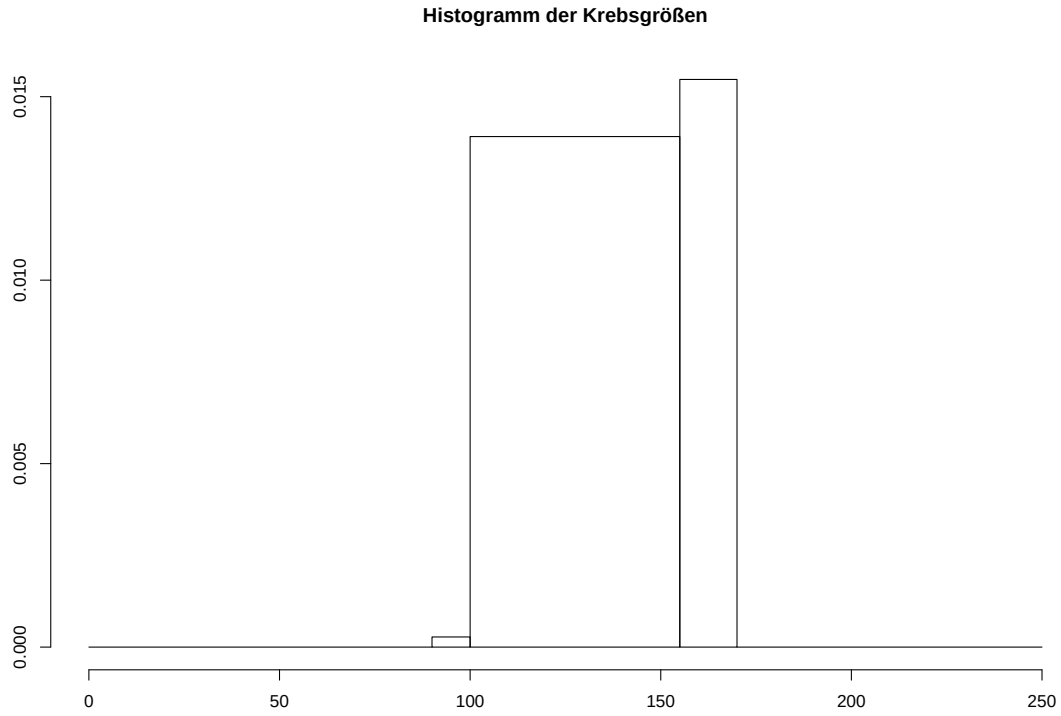
- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.
- Trage über I_j den Wert

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

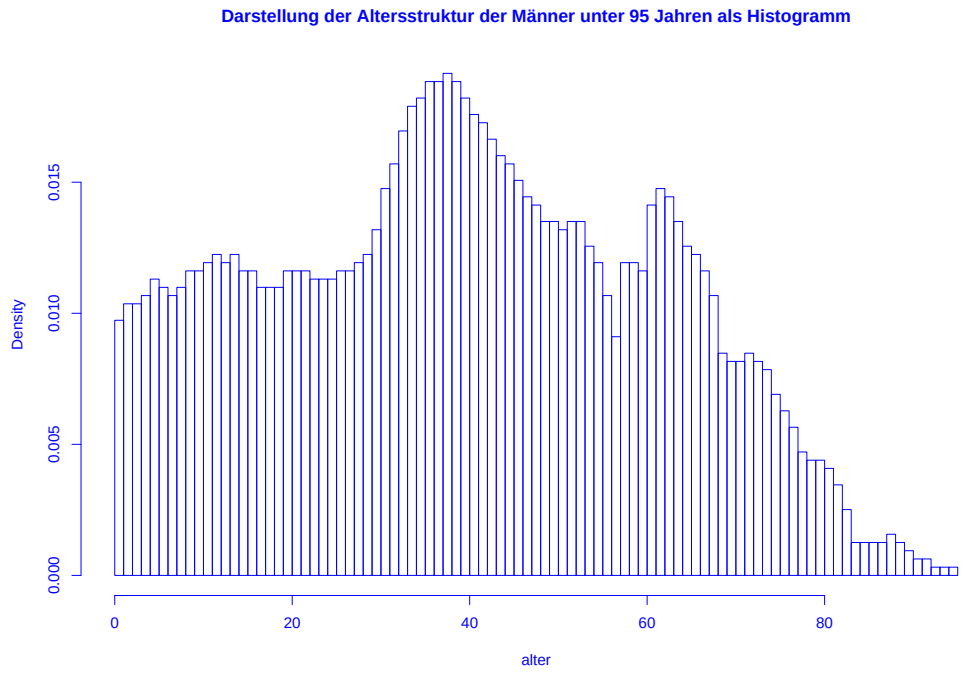
  auf, wobei $\lambda(I_j) =$ Länge von I_j.

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

In Beispiel 1 oben erhält man



Beispiel 3: Altersverteilung der männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD (ohne Berlin-West) im Jahr 2001:



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 proportional zur Anzahl Datenpunkte in dem Intervall.

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

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

und

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

heißt Dichte.

Ziel: Beschreibe Lage der Daten durch glatte Dichtefunktion.

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

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

Mit

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

erhält man

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

mit Dichte

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

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

2. Schritt: **Verallgemeinerung.**

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

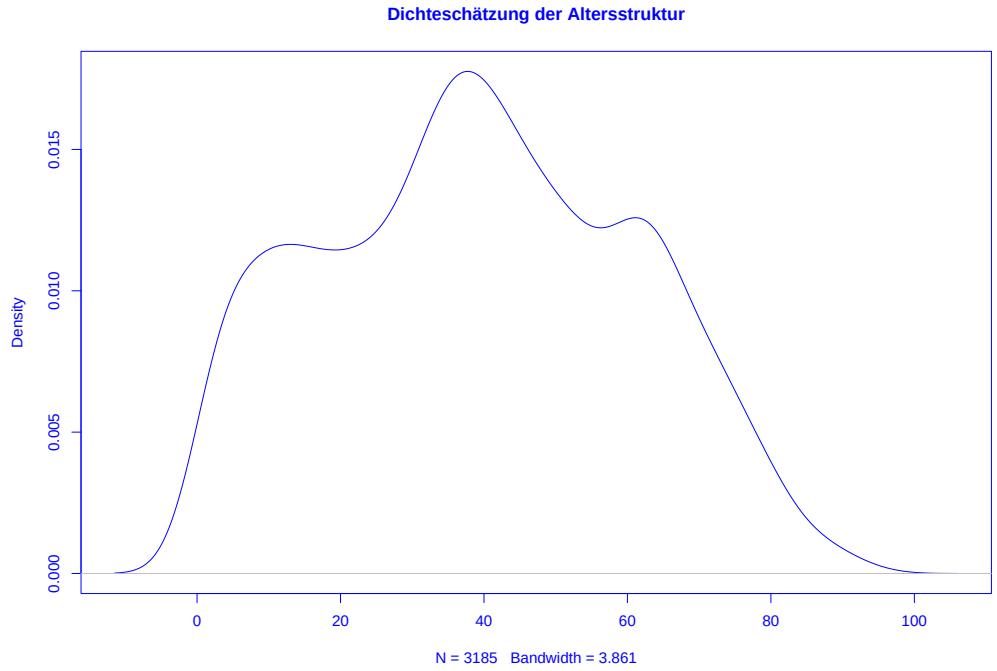
mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

Z.B. Epanechnikov-Kern:

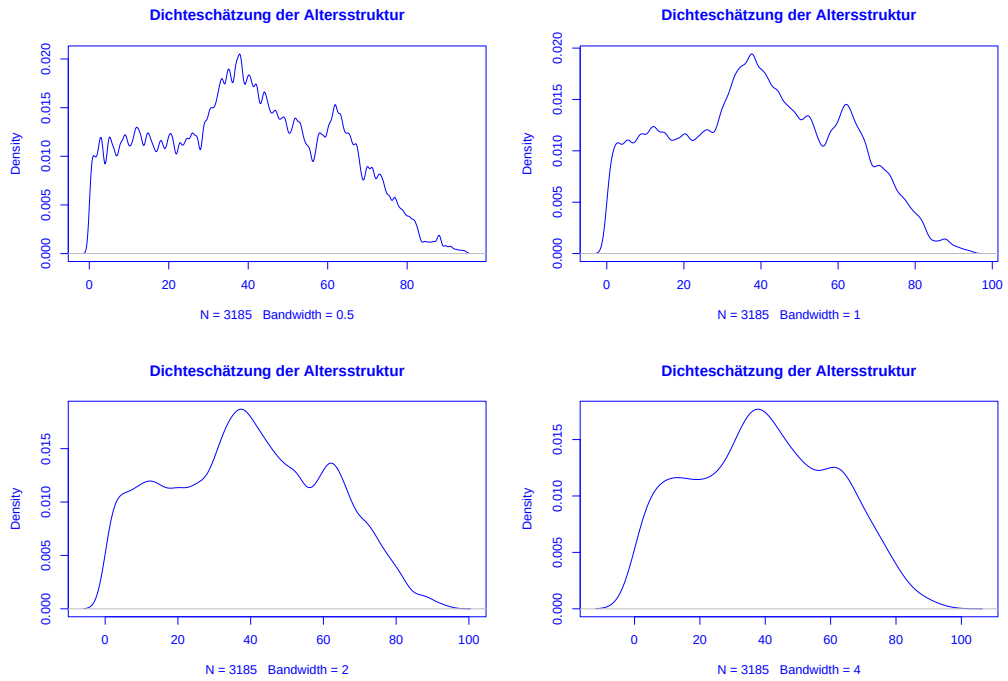
$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}$$

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

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

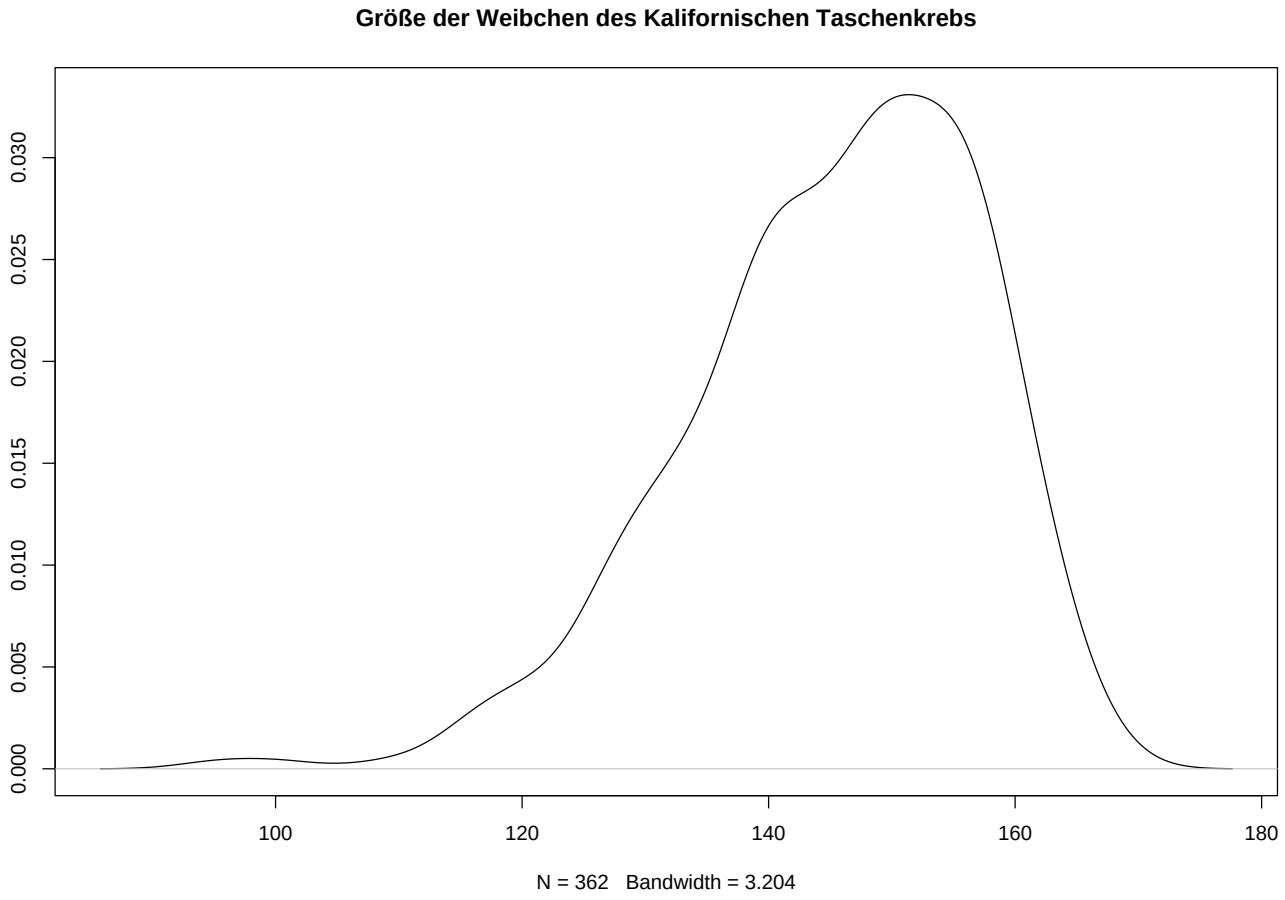


Mittels h lässt sich die "Glattheit" des Kern-Dichteschätzers $f_h(x)$ kontrollieren:



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

In Beispiel 1 (Kalifornischer Taschenkrebs) erhält man



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: Größe (in mm) der Weibchen des Kalifornischen Taschenkrebs:

$x_1, \ldots, x_{29}$:
143.9, 153.8, 144.0, 163.2, 149.3, 140.2, 155.3, 138.9, 153.7, 163.0, 157.1, 132.8, 157.5, 139.1, 155.7, 115.5, 133.3, 144.2, 148.7, 137.7, 144.8, 161.1, 119.6, 139.5, 153.3, 153.4, 139.5, 143.2, 126.7

$x_{(1)}, \ldots, x_{(29)}$:
115.5, 119.6, 126.7, 132.8, 133.3, 137.7, 138.9, 139.1, 139.5, 139.5, 140.2, 143.2, 143.9, 144.0, 144.2, 144.8, 148.7, 149.3, 153.3, 153.4, 153.7, 153.8, 155.3, 155.7, 157.1, 157.5, 161.1, 163.0, 163.2

Beispiele für Lageparameter:

(empirisches arithmetisches) Mittel:

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

Im Beispiel oben: $\bar{x} = 145.1$.

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

In diesen Fällen besser geeignet:

(empirischer) Median:

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

Im Beispiel oben: $\tilde{x} = 144.2$.

Beispiele für Streuungsparameter:

(empirische) Spannweite oder Variationsbreite:

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

Im Beispiel oben: $r = 163.2 - 115.5 = 47.7$.

(empirische) Varianz:

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

Im Beispiel oben: $s^2 \approx 146.9$.

(empirische) Standardabweichung oder Streuung:

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

Im Beispiel oben: $s \approx 12.12$.

Variationskoeffizient:

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

Im Beispiel oben: $V \approx 0.084$.

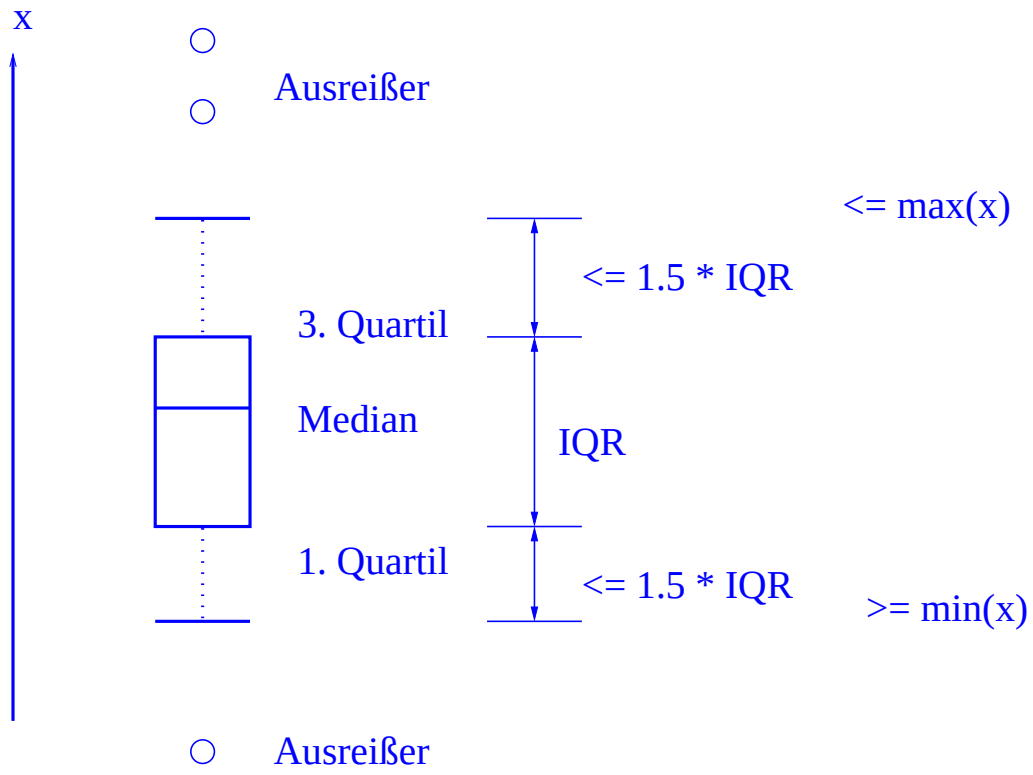
Bei nicht reellen Messgrößen oder Vorhandensein von Ausreißern ist der sogenannte Interquartilabstand

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

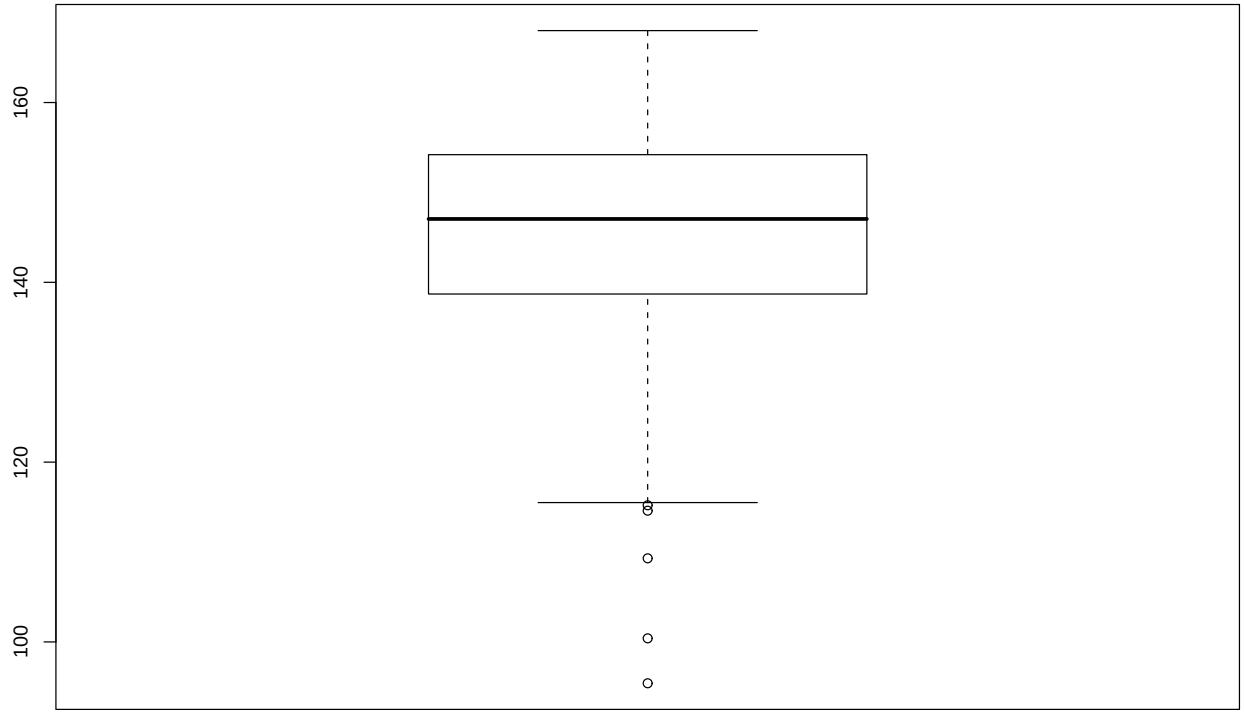
günstiger.

Im Beispiel oben:$IQR = 153.8 - 139.1 = 14.7$.

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



Boxplot für Beispiel 1 (mit allen 362 Daten):



Vergleich der Größen von Weibchen des Kalifornischen Taschenkrebs, die kürzlich ihren Panzer abgestoßen haben, mit denen, die ihren Panzer schon länger haben:

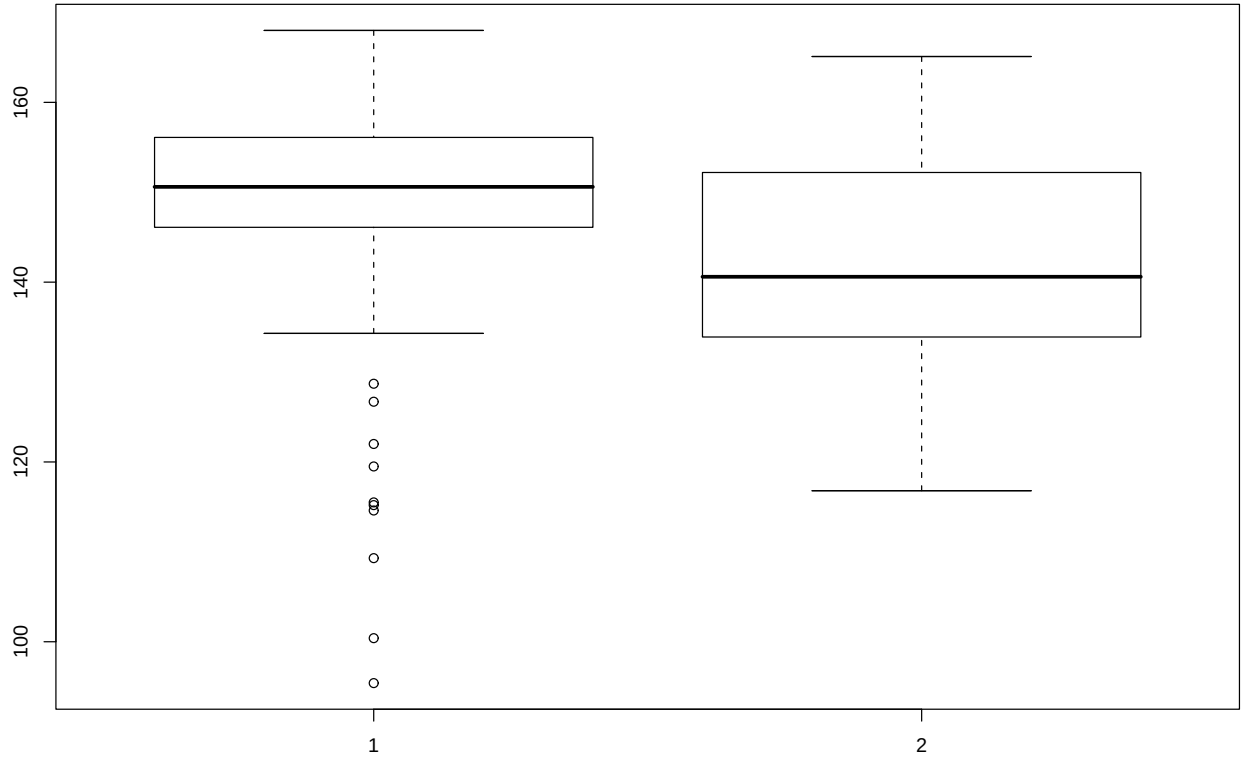
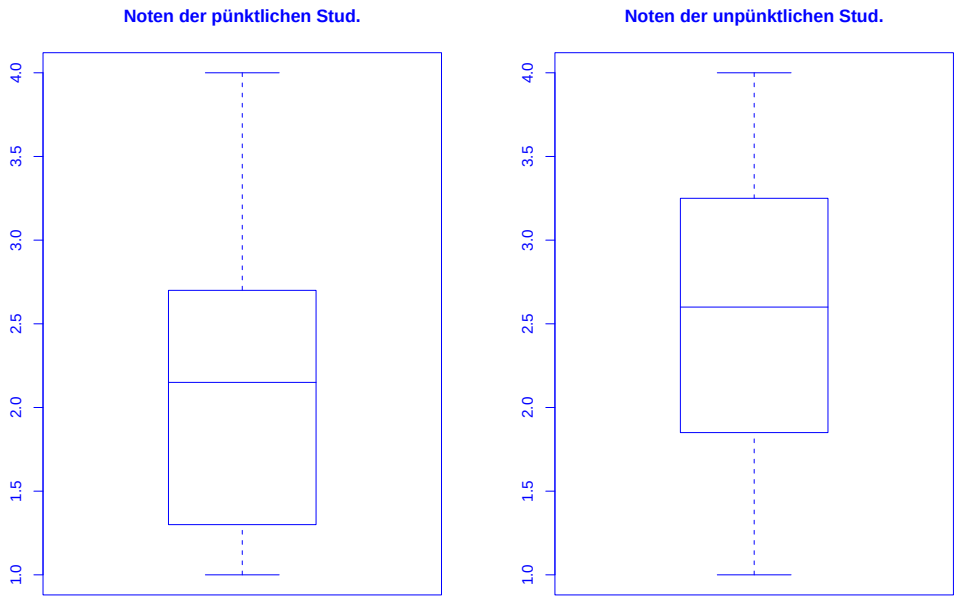
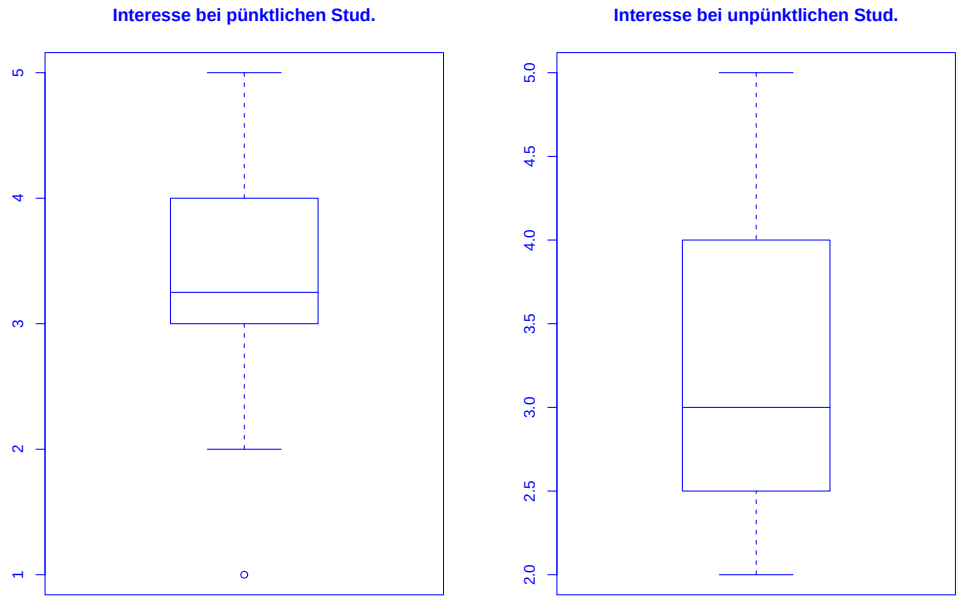


Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:




Interesse bei pünktlichen Stud.
1
2
3
4
5

Interesse bei unpünktlichen Stud.
2.0
2.5
3.0
3.5
4.0
4.5
5.0

3.4 Regressionsrechnung

Geg.: 2–dimensionale Messreihe

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

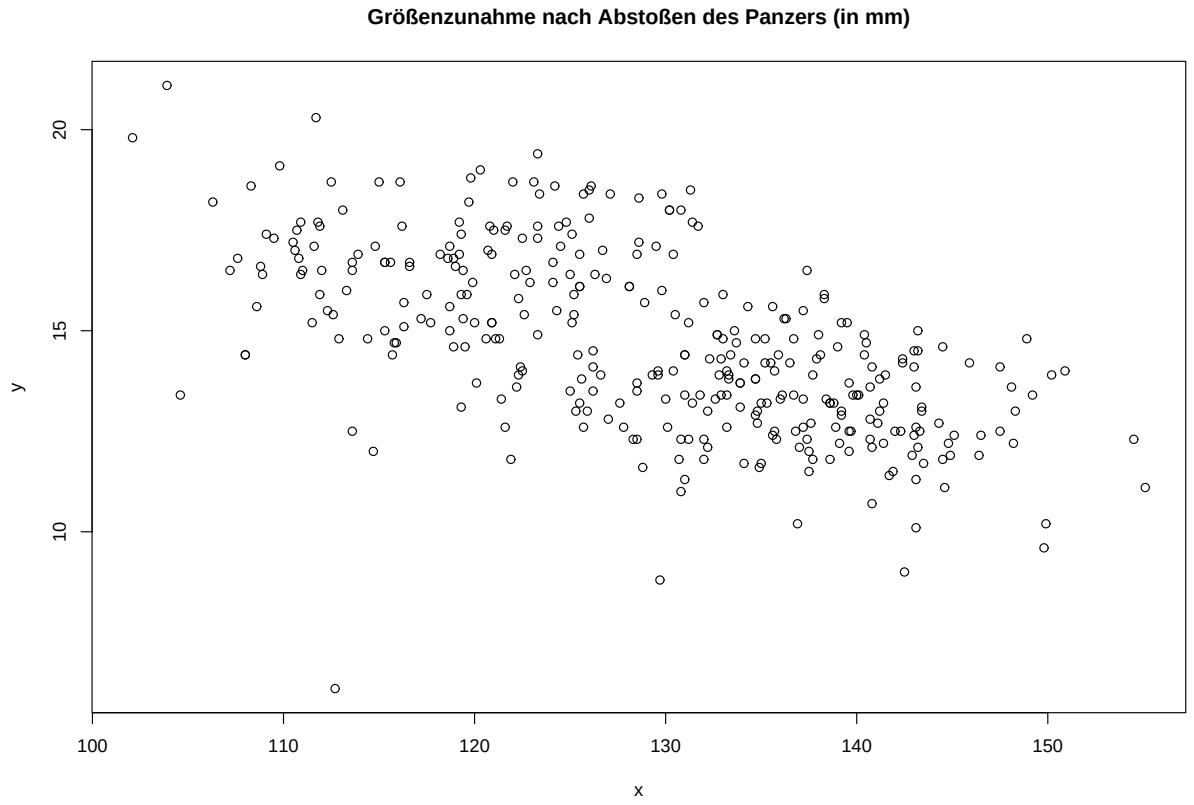
vom Umfang n.

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

Beispiel: Besteht ein Zusammenhang zwischen

- Größe des Weibchens des Kalifornischen Taschenkrebs und Wachstum nach Abstoßung des Panzers ?

Darstellung der Messreihe im Scatterplot (Streudiagramm), wobei 342 Weibchen größer als 100 mm berücksichtigt werden (capture / recapture data von 1981, 1982 und 1992):



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

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

Nach kleinerer Rechnung ergibt sich die sogenannte **Regressionsgerade**

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

mit

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

und

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

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

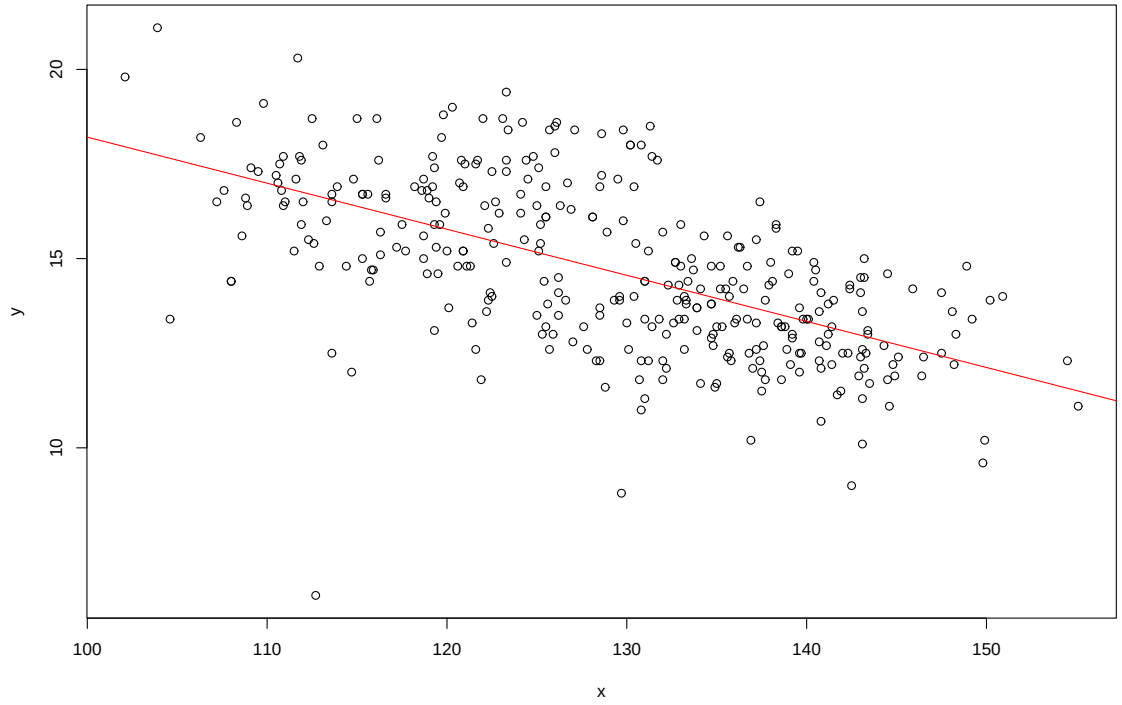
Hierbei wird

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

als empirische Kovarianz der zweidimensionalen Messreihe bezeichnet.

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

Beispiel:



Man rechnet leicht nach:

$$\sum_{i=1}^{n} \left(y_i - (\hat{a}(x_i - \bar{x}) + \bar{y})\right)^2$$

$$= (n-1) \cdot s_y^2 \cdot \left(1 - \frac{s_{x,y}^2}{s_x^2 \cdot s_y^2}\right).$$

Der obige Wert liegt (nach Konstruktion) zwischen Null und $(n-1) \cdot s_y^2$!

Daraus folgt, dass die sogenannte empirische Korrelation

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

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

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

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

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

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

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

Nachteil der linearen Regression:

Nicht sinnvoll, sofern der Zusammenhang zwischen x und y nicht durch eine lineare Funktion gut approximiert werden kann.

Ob dies der Fall ist, ist insbesondere für hochdimensionale Messsreihen (Dimension von $x > 1$) nur schlecht feststellbar.

3.5 Nichtparametrische Regressionsschätzung

Verallgemeinerung der linearen Regression:

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

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

Anderer Ansatz:

Nichtparametrische Regressionsschätzung.

Keine Annahme über die Bauart der anzupassenden Funktion.

Einfachstes Beispiel: lokale Mittelung

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

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

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

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

z.B. naiver Kern

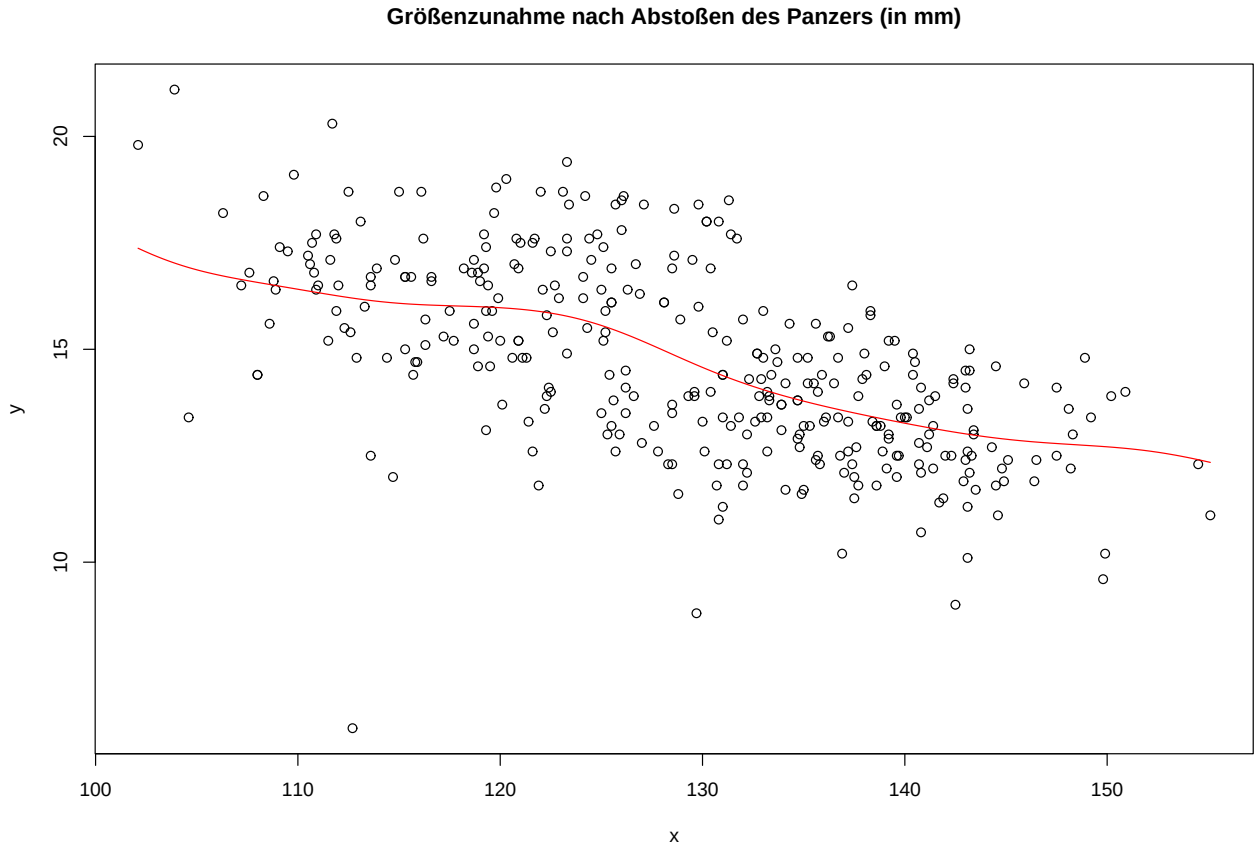
$$K(u) = \frac{1}{2} 1_{[-1,1]}(u)$$

oder Gauss-Kern

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

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

Im Beispiel oben ergibt Anwendung des Kernschätzers mit Gauss-Kern und Bandbreite $h = 10$:



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 *Zufallsexperiment mit unbestimmten Ergebnis* $\omega \in \Omega$.

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

z.B. beim Werfen eines echten Würfels:

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

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

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

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

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

Hier ist $n = 10$.

Absolute und relative Häufigkeit des Auftretens der einzelnen Zahlen:

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

Der Begriff des Ereignisses

Ein Ereignis ist eine Teilmenge der Grundmenge.

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

Die einelementigen Teilmengen der Ergebnismenge heißen Elementarereignisse.

Die Elementarereignisse im Beispiel oben sind

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

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

Das empirische Gesetz der großen Zahlen:

Beobachtung aus der Praxis:

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

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

Ziel im Folgenden: Bestimmung der Wahrscheinlichkeiten bei Zufallsexperimenten.

Möglichkeiten zur Bestimmung von Wahrscheinlichkeiten:

1. Zufallsexperiment sehr häufig durchführen, relative Häufigkeiten bestimmen.

2. Mit Symmetrieüberlegungen auf die Wahrscheinlichkeiten schließen.

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

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

Eigenschaften der Zuweisung von Wahrscheinlichkeiten zu Mengen:

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

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

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

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

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

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

Folgerungen aus (i)-(v):

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

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

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

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

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

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

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

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

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

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

Lemma. Die Eigenschaften (i)-(v) sind genau dann erfüllt, wenn gilt:

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

2. Es gilt $\mathbf{P}(\Omega) = 1$.

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

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

Im Beispiel oben führen Symmetrieüberlegungen auf

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

Wegen (iv) folgt daraus sofort:

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

4.3 Grundformeln der Kombinatorik

Motivation: Im Rahmen einer Beobachtungsstudie hat Prof. P. K. Gadhia von der Universität Süd-Gujurat in Indien festgestellt, dass das Telefonieren mit dem Handy zu Genschäden an menschlichen Blutzellen führt.

Wie bei Beobachtungsstudien üblich wurden dabei verschiedene ausgewählte Personengruppen (bestehend aus Handy-Benutzern und Nicht-Handy-Benutzern unterteilt nach potenziellen konfundierenden Faktoren wie Rauch- oder Trinkgewohnheiten) miteinander verglichen. Da dies extrem fehleranfällig ist, können daraus aber eigentlich keine Rückschlüsse auf die Gefahren der Handy-Strahlung gezogen werden.

Das Bundesamt für Strahlenschutz hat daraufhin eine Reihe von kontrollierten Studien mit Randomisierung in Auftrag gegeben, bei denen untersucht werden sollte, ob niederfrequente Magnetfelder **Chromosomenschäden in Zellkulturen** auslösen.

Dazu sollten (in verschiedenen Altersgruppen und mit je zehn Versuchspersonen) je zwei Zellkulturen bestehend aus je 100 Zellen pro Versuchsperson gewonnen werden. Anschließend wurde eine davon einem niederfrequentem Magnetfeld ausgesetzt und danach beide auf Chromosomenschäden hin untersucht.

Dies war relativ aufwendig, da

- Chromosomenschäden nicht leicht zu erkennen sind,
- Zellkulturen sich nur durch das Magnetfeld unterscheiden sollten,
- das Magnetfeld homogen sein sollte,

$\vdots$

Anschließend sollte z.B. durch die Anzahlen der Zellen mit Chromosomenschäden pro Zellkultur auf die Wirkung der Strahlung zurückgeschlossen werden.

Problem: Selbst wenn die Strahlung keine Auswirkung auf die Zellen hat, wird sich die Anzahl der geschädigten Zellen pro Zellkultur unterscheiden (da Chromosomenschäden bei den Zellen auch zufällig auftreten).

Frage: Wie unterscheidet man den Einfluss des Zufalls von dem Einfluss der Strahlung ?

Zahlenbeispiel (hypothetisch): In den beiden Zellkulturen mit je 100 Zellen seien in der bestrahlten Zellkultur 3 Zellen mit Chromosomenschäden, in der nichtbestrahlten Zellkultur aber nur 1 Zelle mit Chromosomenschaden.

Ist das Zufall oder nicht ?

Idee beim statistischen Test:

Wir wollen uns zwischen zwei Hypothesen

- H_0: Strahlung hat keinen Einfluss auf Chromosomenschäden
- H_1: Strahlung hat Einfluss auf Chromosomenschäden

entscheiden.

Wir gehen zuerst hypothetisch davon aus, dass H_0 stimmt. Unter dieser Annahme bestimmen wir dann, mit welcher Wahrscheinlichkeit ein Ergebnis auftritt, dass mindesten so stark für H_1 spricht wie das beobachtete Ergebnis. Ist dann diese Wahrscheinlichkeit sehr klein (z.B. nicht größer als 0.05), so verwerfen wir die Annahme.

Im Zahlenbeispiel oben:

1. Wir betrachten ein Zufallsexperiment, bei dem $3 + 1 = 4$ geschädigte und $200 - 4 = 196$ nicht-geschädigte Zellen zufällig in zwei Gruppen zu je 100 Zellen unterteilt werden.

2. Wir bestimmen dann die Wahrscheinlichkeit, dass mindestens 3 der geschädigten Zellen in der ersten Gruppe sind, d.h., dass die erste Gruppe 3 oder 4 geschädigte Zellen enthält.

3. Ist diese Wahrscheinlichkeit nicht größer als 0.05, so gehen wir davon aus, dass H_0 nicht stimmt und dass die Strahlung Chromosomenschäden auslößt. Andernfalls gehen wir davon aus, dass H_0 stimmt und dass die Strahlung keinen Einfluss auf Chromosomenschäden hat.

Achtung: Da man hier nur die Wahrscheinlichkeit einer falschen Entscheidung für H_1 bei Gültigkeit von H_0 kontrolliert, spricht man asymmetrisch davon, dass entweder H_0 nicht abgelehnt wird oder dass H_0 abgelehnt und gleichzeitig H_1 angenommen wird.

Nötig: Kombinatorische Hilfsmittel zur Ermittlung der obigen Wahrscheinlichkeit!

Einführung in die Kombinatorik:

Betrachtet wird das Ziehen von k Elementen aus einer Grundmenge Ω vom Umfang $|\Omega| = n$.

Die Anzahl aller möglichen Stichproben sei N.

Dabei kann man vier verschiedene Vorgehensweisen unterscheiden, und zwar je nachdem, ob man die Elemente unmittelbar nach dem Ziehen wieder zurücklegt oder nicht, und je nachdem, ob man die Reihenfolge, in der die Elemente gezogen werden, beachtet oder nicht.

Es gilt:

Anzahl Möglichkeiten	Ziehen mit Zurücklegen	Ziehen ohne Zurücklegen
Ziehen mit Berücksichtigung der Reihenfolge	n^k	$\frac{n!}{(n-k)!}$
Ziehen ohne Berücksichtigung der Reihenfolge	$\binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!\cdot k!}$	$\binom{n}{k} = \frac{n!}{(n-k)!\cdot k!}$

Anwendung im Zahlenbeispiel oben:

Wir betrachten das Ziehen von 100 Zellen aus der Gesamtmenge von 200 Zellen **ohne Zurücklegen** und **ohne Beachtung der Reihenfolge**.

Anzahl Möglichkeiten: $\binom{200}{100}$

Anzahl Möglichkeiten mit drei geschädigten Zellen: $\binom{196}{97} \cdot \binom{4}{3}$

Anzahl Möglichkeiten mit vier geschädigten Zellen: $\binom{196}{96} \cdot \binom{4}{4}$

Damit ist die gesuchte Wahrscheinlichkeit gleich

$$\frac{\binom{196}{97} \cdot \binom{4}{3} + \binom{196}{96} \cdot \binom{4}{4}}{\binom{200}{100}} \approx 0.31 \quad \Rightarrow \quad H_0 \text{ kann nicht abgelehnt werden !}$$

4.4 Zufallsvariablen und Verteilungen

Oft interessieren nur Teilaspekte des Ergebnisses eines Zufallsexperimentes.

Idee: Wähle Abbildung

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

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

Beispiel: Werfen zweier echter Würfel

Kann modelliert werden durch

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

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

Wähle

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.5 Beispiele für Wahrscheinlichkeitsmaße und Verteilungen

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

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

heißt **Zähldichte**.

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

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

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

Beispiele für diskrete Verteilungen:

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

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

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

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

Einsatz in der Modellierung:

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

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

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

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

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

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

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

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

und

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

heißt Dichte.

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

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

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

Beispiele für stetige Verteilungen:

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

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

festgelegte W–Maß.

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

Einsatz in der Modellierung:

"Rein zufälliges Ziehen" einer Zahl aus einem Intervall.

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

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

festgelegte W–Maß.

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

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

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

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

festgelegte W–Maß.

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

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

4.6 Erwartungswert und Varianz

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

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

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

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

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

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

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

3. *Erwartungswert des Produktes unabhängiger Zufallsvariablen:*

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

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

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

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

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

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

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

4.6.1 Erwartungswert von diskreten Zufallsvariablen

Sei X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \ldots, x_K \in \mathbb{R}$.

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 mit Werten $x_1, x_2, \ldots, x_K \in \mathbb{R}$ bzw. $x_1, x_2, \cdots \in \mathbb{R}$. 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.

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

X ist diskret verteilt mit Werten in $\{2, 3, \ldots, 12\}$ und es gilt:

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

Damit

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

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

Dabei ist

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

und damit

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

Allgemeiner gilt:

Ist X eine diskrete Zufallsvariable mit Werten $x_1, x_2, \ldots, x_K \in \mathbb{R}$ bzw. $x_1, x_2, \cdots \in \mathbb{R}$, und ist $h : \mathbb{R} \to \mathbb{R}$ eine beliebige reelle Funktion.

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

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

bzw. (sofern existent)

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

4.6.2 Erwartungswert von Zufallsvariablen mit Dichten

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

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

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

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

Allgemeiner setzt man wieder:

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

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

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

(sofern existent).

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

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

Dann gilt:

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

4.6.3 Varianz

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

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

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

die **Varianz** von X.

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

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

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

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

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

Satz:

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

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

Entsprechendes gilt für beliebige endliche Summen unabhängiger Zufallsvariablen.

Kapitel 5: Schließende Statistik

Bisher behandelt: **beschreibende Statistik**

geg.: Messreihe (Stichprobe, Datensatz): $x_1, \ldots, x_n$ (n=Stichprobenumfang)

ges.: *Übersichtliche Darstellung von Eigenschaften dieser Messreihe.*

z.B.: Beschreibe "Mitte" der Werte durch *arithmetisches Mittel*: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

z.B.: Beschreibe "Streuung" der Werte um den Mittelwert durch *empirische Varianz*:

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

Beispiel: Größen (in mm) von 362 im Jahr 1982 in Kalifornien gefangenen Weibchen des Kalifornischen Taschenkrebs (Cancer Magister):

143.9, 153.8, 144.0, 163.2, 149.3, 140.2, 155.3, 138.9, 153.7, 163.0, 157.1, 132.8, 157.5, 139.1, 155.7, 115.5, 133.3, 144.2, 148.7, 137.7, 144.8, 161.1, 119.6, 139.5, 153.3, 153.4, 139.5, 143.2, 126.7, . . .

Arithmetisches Mittel:

$$\bar{x} = \frac{1}{362} \cdot (143.9 + 153.8 + \dots) = 145.1$$

Empirische Varianz:

$$s^2 = \frac{1}{362 - 1} \cdot \Big((143.9 - 145.1)^2 + (153.8 - 145.1)^2 + \dots \Big) \approx 146.9$$

Neu jetzt:

Wir gehen in der schließenden Statistik davon aus, dass die Daten gemäß einem stochastischen Modell erzeugt wurden und wollen:

- Aussagen über **Eigenschaften dieses Modells** machen, z.B.: Wie groß sind Erwartungswert und Varianz im stochastischen Modell ?

- Aussagen über **zukünftige Werte** machen, die bei diesem Modell entstehen, z.B.: Wie sieht ein möglichst kleines Intervall aus, in dem zukünftige Werte mit möglichst großer Wahrscheinlichkeit liegen ?

Dies wird es uns ermöglichen, von dem vorliegenden Datensatz auf die Grundgesamtheit bzw. zukünftige neue (!) Werte zu schließen.

Annahme an die Erzeugung der Daten:

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

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

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

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

5.1 Punktschätzverfahren

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

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

Beispiele:

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

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

Sinnvolle Eigenschaften von Schätzungen:

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

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

Formal:

Definition:

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

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

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

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

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

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

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

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

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

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

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

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

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

Satz (Starkes Gesetz der großen Zahlen):

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

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

Beispiel zum starken Gesetz der großen Zahlen:

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

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

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

Darüberhinaus ist sie wegen

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

auch erwartungstreu.

5.2 Bereichsschätzverfahren

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

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

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

Beispiel: Wie sieht ein "möglichst kleines" Intervall aus, in dem die mittlere Größe des Weibchens des Kalifornischen Taschenkrebs mit "möglichst großer" Wahrscheinlichkeit liegt ?

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

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

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

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

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

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

Der zentrale Grenzwertsatz:

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

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

annähernd $N(0,1)$- verteilt.

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

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

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

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

Nach dem zentralen Grenzwertsatz verhält sich also

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

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

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

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

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

so gilt für n groß approximativ:

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

Wegen

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

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

gilt in diesem Fall auch:

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

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

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

führt.

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

Ausweg: Varianz durch empirische Varianz schätzen.

Man kann zeigen:

Sind $X_1, \ldots, X_n$ unabhängige und identisch **normalverteilte** reelle Zufallsvariablen, so ist

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

t- verteilt mit $n-1$ Freiheitsgraden. Hierbei sind die Werte der Verteilungsfunktion der t_{n-1}-Verteilung tabelliert.

Konstruktion eines zweiseitigen Konfidenzintervalls zum Niveau α für den Erwartungswert:

1. Gegeben sind $\alpha \in (0,1)$, $n \in \mathbb{N}$ und $x_1, \ldots x_n \in \mathbb{R}$.

2. Bestimme

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

und

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

3. Bestimme $\delta \in \mathbb{R}$ so, dass für eine t_{n-1}-verteilte Zufallsvariable Z gilt:

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

4. Das gesuchte Konfidenzintervall ist

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

Beispiel: Zweiseitiges Konfidenzintervall für die mittlere Größe des Weibchens des Kalifornischen Taschenkrebs zum Niveau $\alpha = 0.95$:

- Hier ist $n = 362$, $\hat{\mu} = 145.1$ und $\hat{\sigma}^2 = 146.9$.
- Für eine $t_{n-1} = t_{361}$-verteilte Zufallsvariable Z gilt:

$$\mathbf{P}[|Z| \leq 1.967] = 0.95.$$

  Also wählen wir $\delta = 1.967$.

- Das gesuchte Konfidenzintervall ist dann

$$C(x_1, \ldots, x_n) = \left[\hat{\mu} - \frac{\sqrt{\hat{\sigma}^2}}{\sqrt{n}} \cdot \delta, \hat{\mu} + \frac{\sqrt{\hat{\sigma}^2}}{\sqrt{n}} \cdot \delta\right] = [143.8, 146.4].$$

5.3 Statistische Testverfahren

5.3.1. Beispiele:

1. Ändern sich Spinnenweben mit den Lichtverhältnissen ?

   Vorhandene Daten:

   17 Spinnen wurden zwei verschiedenen Lichtstufen (Dämmerung und Tageslicht) ausgesetzt und für jede Spinne wurde die Differenzen x bzw. y zwischen vertikalen bzw. horizontalen Durchmesser der im jeweiligen Licht gewebten Netze ermittelt.

   Beschreibung der gemessenen Daten (in cm) : $n = 17$ und

   - $\bar{x} = 46.18,\ s_x = 21.49$
   - $\bar{y} = 20.59,\ s_y = 21.32$

2. Unterscheidet sich der Energiehaushalt bei brütenden Eissturmvögeln (fulmar glacialis) zwischen männlichen (x) und weiblichen (y) Vögeln ?

Beschreibung der gemessenen Daten:

- $n_x = 8,\ \bar{x} = 1563.78,\ s_x = 894.37$
- $n_y = 6,\ \bar{y} = 1285.52,\ s_y = 420.96$

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 er im mathematischen Modell dieser Vorlesung beschrieben wurde) entstanden sind.

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

   - In Beipiel 1 fassen wir unsere Daten als Realisierungen $x_1, \ldots, x_{17}$ von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots X_{17}$ auf.
   - In Beipiel 2 fassen wir unsere Daten als Realisierungen $x_1, \ldots, x_8$ von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots X_8$ bzw. $y_1, \ldots, y_6$ von unabhängigen identisch verteilten Zufallsvariablen $Y_1, \ldots Y_6$

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

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

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

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

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

     zutrifft.

Prinzipieller Unterschied zwischen den beiden Fragestellungen:

- In Beispiel 1 haben wir **eine** Stichprobe $x_, \ldots x_{17}$ der Verteilung von X_1 gegeben, und wollen wissen, ob $\mathbf{E}X_1 = 0$ gilt (**Einstichprobenproblem**).

- In Beipiel 2 haben wir **zwei** Stichproben $x_1, \ldots, x_8$ bzw. $y_1, \ldots, y_6$ der Verteilungen von X_1 bzw. Y_1 gegeben, und wollen wissen, ob $\mathbf{E}X_1 = \mathbf{E}Y_1$ gilt (**Zweistichprobenproblem**).

Anmerkung:

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

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

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

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

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

5.3.3 Grundbegriffe der Testtheorie

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

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

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

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

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

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

Bezeichnung für die auftretenden Fehler:

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

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

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

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

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

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

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

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

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

Ausweg: Asymmetrische Betrachtungsweise der Fehlerwahrscheinlichkeiten erster und zweiter Art:

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

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

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

Achtung:

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

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

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

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

5.3.4 Der Gauß-Test

1. Fragestellungen beim Gauß-Test für eine Stichprobe

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

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

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

   testen.

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

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

   testen.

2. Fragestellungen beim Gauß-Test für zwei Stichproben

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

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

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

testen.

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

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

testen.

3. Grundidee beim Gauß-Test für eine Stichprobe

(a) Wir betrachten

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

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

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

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

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

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

4. Einseitiger Gauß-Test für eine Stichprobe

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

5. Zweiseitiger Gauß-Test für eine Stichprobe

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

6. Grundidee beim Gauß-Test für zwei Stichproben

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

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

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

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

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

7. Einseitiger Gauß-Test für zwei Stichproben

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

8. Zweiseitiger Gauß-Test für zwei Stichproben

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

5.3.5 Der t-Test von Student

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

Ausweg: Wir schätzen die Varianz aus unseren Daten.

Einfach, bei Test für eine Stichprobe:

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

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

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

Zur Einhaltung des Niveaus beachten wir:

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

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

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

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

Beispiel: Zweiseitiger t-Test für eine Stichprobe

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

Anwendung im Spinnenbeispiel:

17 Spinnen wurden zwei verschiedenen Lichtstufen (Dämmerung und Tageslicht) ausgesetzt und für jede Spinne wurde die Differenzen x bzw. y zwischen vertikalen bzw. horizontalen Durchmesser der im jeweiligen Licht gewebten Netze ermittelt.

Beschreibung der gemessenen Daten (in cm) : $n = 17$ und

- $\bar{x} = 46.18$, $s_x = 21.49$

- $\bar{y} = 20.59$, $s_y = 21.32$

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

Hierbei gilt: $t_{n-1,\alpha} = t_{16,0.05/2} \approx 2.12$

Für die Differenzen der vertikalen Durchmesser erhalten wir

$$\frac{\sqrt{n}}{s_x} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) = \frac{\sqrt{17}}{21.49} \cdot (46.18 - 0) \approx 8.86 > t_{16,0.025},$$

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

Für die Differenzen der horizontalen Durchmesser erhalten wir

$$\frac{\sqrt{n}}{s_y} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} y_i - \mu_0 \right) = \frac{\sqrt{17}}{21.32} \cdot (20.59 - 0) \approx 3.98 > t_{16,0.025},$$

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

Resultat: Die vertikalen und die horizontalen Durchmesser unterscheiden sich jeweils je nach Lichstufe (Tageslicht oder Dämmerung).

t-Test für zwei Stichproben

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

Zu testen sei

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

bzw.

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

Problem: Wie schätzen wir diesmal die Varianz ?

Schätzung der Varianz:

Wir verwenden die sogenannte gepoolte Stichprobenvarianz

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

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

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

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

Beispiel: Zweiseitiger t-Test für zwei Stichproben

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

Zu testen sei

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

H_0 wird abgelehnt, falls

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

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

Anwendung bei den brütenden Eissturmvögeln:

Unterscheidet sich der Energiehaushalt bei brütenden Eissturmvögeln (fulmar glacialis) zwischen männlichen (x) und weiblichen (y) Vögeln ?

Beschreibung der gemessenen Daten:

- $n_x = 8,\ \bar{x} = 1563.78,\ s_x = 894.37$

- $n_y = 6,\ \bar{y} = 1285.52,\ s_y = 420.96$

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: $t_{n_x+n_y-2,\alpha} = t_{8+6-2,0.05/2} = t_{12,0.05/2} \approx 2.179$

Für die beobachteten Daten erhalten wir

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

$$= \frac{|1563.78 - 1285.52|}{\sqrt{\frac{1}{8}+\frac{1}{6}}\cdot\sqrt{\frac{1}{8+6-2}\cdot((8-1)\cdot 894.37 + (6-1)\cdot 420.96)}}$$

$$\approx 19.51 > t_{12,0.05/2},$$

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

Resultat: Der Energiehaushalt bei brütenden Eissturmvögeln (fulmar glacialis) unterscheidet sich zwischen männlichen (x) und weiblichen (y) Vögeln.

5.4 Einfaktorielle Varianzanalyse

5.4.1 Einführung

Wir interessieren uns für den Einfluß eines potentiellen **Einflussfaktors** auf eine zufällige Größe, insbesondere wollen wir wissen, ob der Wert der zufälligen Größe durch den Wert des Einflussfaktors überhaupt beeinflusst wird.

Anwendungsbeispiel: Nach der Eiszeit waren in Nordamerika so gut wie keine Regenwürmer mehr vorhanden. Erst in jüngster Zeit nimmt deren Zahl durch menschliche Einflüsse (von Holzfällern und Anglern) in den Wäldern wieder zu.

Was passiert bei der Invasion von Regenwürmern in einen Laubwald ?

Fragestellungen:

1. Verändert sich die Zahl der vorkommenden Regenwürmern mit der Entfernung zum Waldrand ?

2. Beeinflusst die Zahl der vorhandenen Regenwürmer die Beschaffenheit des Waldbodens, z.B. die Dicke der Humusschicht ?

Vereinfachung:

Wir unterteilen den Wertebereich des Einflussfaktors in endlich viele **Faktorstufen** (die nicht unbedingt geordnet sein müssen), z.B.

- Ermittlung der Zahl der Regenwürmer pro Quadratmeter entlang drei verschiedener Strecken zwischen Waldrand und Waldmitte, wobei jeweils an zehn gleichweit voneinander entfernten Stellen die Regenwürmer gezählt werden (1. Datensatz mit 10 Faktorstufen und 3 Beobachtungen pro Faktorstufe).

- Ermittlung der Dicken der Humusschicht pro untersuchtem Waldstück, und Angabe, ob in diesem Waldstück Regenwürmer vorhanden bzw. nicht vorhanden waren (2. Datensatz mit 2 Faktorstufen).

Der zweite Datensatz kann mit Hilfe des t-Tests analysiert werden, für den ersten Datensatz benötigen wir eine Verallgemeinerung des t-Tests auf mehr als zwei Variablen.

5.4.2 Das Testproblem

geg.: k unabhängige Stichproben von k verschiedenen Normalverteilungen mit gleicher Varianz, wobei alle auftretenden Zufallsvariablen unabhängig sind:

- Realisierungen $x_1(1), \ldots, x_{n_1}(1)$ von unabhängig $N(\mu_1, \sigma^2)$-verteilten Zufallsvariablen $X_1(1), \ldots, X_{n_1}(1)$.

$$\vdots$$

- Realisierungen $x_1(k), \ldots, x_{n_k}(k)$ von unabhängig $N(\mu_k, \sigma^2)$-verteilten Zufallsvariablen $X_1(k), \ldots, X_{n_k}(k)$.

Zu testen:

$$\mathbf{H_0}: \mu_1 = \mu_2 = \cdots = \mu_\mathbf{k} \quad \text{versus} \quad \mathbf{H_1}: \text{Es existieren } \mathbf{1} \leq \mathbf{i} < \mathbf{j} \leq \mathbf{k} \text{ mit } \mu_\mathbf{i} \neq \mu_\mathbf{j}$$

5.4.3 Grundidee der einfaktoriellen Varianzanalyse

Wir vergleichen zwei verschiedene Varianzschätzer und ziehen damit Rückschlüsse auf die Erwartungswerte.

1. Varianzschätzer:

Basiert auf Abweichungen zwischen den Mittelwerten der einzelnen Stichproben und dem Gesamtmittel, also auf

$$\bar{X}(l) = \frac{1}{n_l} \sum_{i=1}^{n_l} X_i(l) \text{ und } \bar{X} = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} X_i(l),$$

wobei $n = n_1 + \cdots + n_k$, und ist gegeben durch

$$SS_1^2 = \frac{1}{k-1} \sum_{l=1}^{k} \sum_{i=1}^{n_l} (\bar{X}(l) - \bar{X})^2 = \frac{1}{k-1} \sum_{l=1}^{k} n_l (\bar{X}(l) - \bar{X})^2.$$

2. Varianzschätzer:

Wird mittelbar gebildet unter Verwendung der Schätzer für die Varianzen der einzelnen Stichproben:

$$S_X^2(l) = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (X_i(l) - \bar{X}(l))^2 \quad \text{mit } \bar{X}(l) = \frac{1}{n_l} \sum_{i=1}^{n_l} X_i(l).$$

Damit schätzen wir die Gesamtvarianz durch

$$SS_2^2 = \frac{1}{n-k} \sum_{l=1}^{k} \sum_{i=1}^{n_l} (X_i(l) - \bar{X}(l))^2 = \frac{1}{n-k} \sum_{l=1}^{k} (n_l - 1) \cdot S_X^2(l).$$

Man kann nun zeigen:

Sind alle Erwartungswerte identisch (also gilt H_0), so hängt die Verteilung von

$$\frac{SS_1^2}{SS_2^2}$$

nur von k und n ab, und zwar handelt es sich um eine sogenannte $F_{k-1,n-k}$-Verteilung, deren Verteilungsfunktionen bekannt und tabelliert sind.

Idee des Tests: Lehne H_0 ab, falls

$$\frac{SS_1^2}{SS_2^2}$$

größer als das α-Fraktil der $F_{k-1,n-k}$-Verteilung ist (d.h. größer ist als der Wert, an dem die Verteilungsfunktion der $F_{k-1,n-k}$-Verteilung den Wert $1-\alpha$ annimmt).

5.4.4 Der Test

geg. $x_1(1), \ldots, x_{n_1}(1)$, $x_1(2), \ldots, x_{n_2}(2)$, $\ldots$, $x_1(k), \ldots, x_{n_k}(k)$

zu testen:

H_0 : alle Erwartungswerte gleich versus H_1 : nicht alle Erwartungswerte gleich

Vorgehen: Berechnen mit diesen Werten die beiden Varianzschätzer SS_1^2 und SS_2^2 wie oben.

Ergebnis: Lehne H_0 ab, falls

$$\frac{SS_1^2}{SS_2^2}$$

größer als das α-Fraktil der $F_{k-1,n-k}$-Verteilung ist.