KARLSRUHER INSTITUT FÜR TECHNOLOGIE

Lehrstuhl für Ökonometrie und Statistik

S T A T I S T I K I

Lösungsblätter zu den Übungen

Sommersemester 2014

Prof. Dr. Wolf-Dieter Heller
Frieder Conrad
Hartwig Senska

Institut für Volkswirtschaftslehre
76131 Karlsruhe

Aufgabe 1

Geben Sie für die folgenden Fragestellungen die statistischen Einheiten und Massen an, und grenzen Sie die Massen sachlich, räumlich und zeitlich ab.

(a) Es soll der Anteil der Pkw des Typs A an den im März 2008 in der Bundesrepublik neu zugelassenen Pkw ermittelt werden.

(b) Es soll die durchschnittliche Studiendauer an deutschen Universitäten ermittelt werden.

Lösung: (Deskriptive Statistik, S. 10ff)[1]

(a) **1. Vorschlag: Zwei statistische Massen**

„Neu zugelassene Pkw insgesamt“

i) Sachlich: neu zugelassener Pkw
ii) Räumlich: Zulassung in der Bundesrepublik Deutschland
iii) Zeitlich: Zulassung erfolgte im Monat März 2008

Statistische Einheiten sind hier die einzelnen Pkw, die die Forderungen der Abgrenzung einhalten.

„Neu zugelassene Pkw vom Typ A“

i) Sachlich: neu zugelassener Pkw vom Typ A
ii) Räumlich: Zulassung in der Bundesrepublik Deutschland
iii) Zeitlich: Zulassung erfolgte im Monat März 2008

Statistische Einheiten sind hier die einzelnen Pkw, die die Forderungen der Abgrenzung einhalten, also insbesondere vom Typ A sind.

Der gesuchte Anteil berechnet sich jetzt als Verhältnis der Anzahl der statistischen Einheiten der Masse „Neu zugelassene Pkw vom Typ A“ zur Anzahl der statistischen Einheiten der Masse „Neu zugelassene Pkw insgesamt“.

2. Vorschlag: eine statistische Masse mit Merkmal

„Neu zugelassene Pkw insgesamt“

i) Sachlich: neu zugelassener Pkw
ii) Räumlich: Zulassung in der Bundesrepublik Deutschland

[1] Die theoretischen Grundlagen zur Lösung der Aufgabe finden Sie jeweils an der hier angegebenen Stelle.

iii) Zeitlich: Zulassung erfolgte im Monat März 2008

Statistische Einheiten sind hier die einzelnen Pkw, die die Forderungen der Abgrenzung einhalten.

Betrachtet wird das Merkmal „Typ“. Gesucht ist die relative Häufigkeit der Merkmalsausprägung A.

(b) **Statistische Masse:**

„Absolvent(inn)en an deutschen Universitäten“

i) Sachlich: (erfolgreich?) abgeschlossenes Studium an einer Universität
Bemerkung: Entsprechend dem Untersuchungsgegenstand ist noch festzulegen, welche Studiengänge berücksichtigt werden sollen: z.B. Differenzierung nach Diplom, Lehramt, Magister, Promotion. Sollen auch Kurz- und Aufbaustudiengänge erfasst werden? Sollen auch die mit berücksichtigt werden, die beim Abschlussexamen nicht bestanden haben?

ii) Räumlich: Universität in der Bundesrepublik Deutschland

iii) Zeitlich: Es ist noch ein Zeitraum festzulegen, in dem das Studium abgeschlossen wurde.

Statistische Einheiten sind die einzelnen Absolvent(inn)en, die die Abgrenzungskriterien einhalten.

Betrachtet wird das Merkmal „Dauer des Studiums“. Gesucht ist das arithmetische Mittel. Soll entsprechend der Bemerkung eine Differenzierung durchgeführt werden, sind weitere Merkmale zu erfassen oder die statistische Masse ist entsprechend der Kriterien (führt zu unterschiedlichen sachlichen Abgrenzungen) in mehrere statistische Massen zu zerlegen.

Aufgabe 2

(a) Welche der folgenden Beispiele sind Bestands- und welche Ereignismassen?

- Studierende an einer Universität
- Todesfälle in einer Gemeinde
- Personenkraftwagen der Deutsche Post AG
- Maschinenausfälle in einer Fabrik
- Anmeldungen in einem Einwohnermeldeamt
- Wartende Postkunden vor einem Postschalter
- Leichte Verkehrsunfälle in der Bundesrepublik

(b) Zu den Bestandsmassen aus (a) gebe man jeweils die korrespondierenden Ereignismassen, zu den Ereignismassen die zugehörigen Bestandsmassen an.

(c) Geben Sie ein einfaches Beispiel an, bei dem *einer* Bestandsmasse *mehrere* korrespondierende Ereignismassen zugeordnet sind.

Lösung: (Deskriptive Statistik, S. 13ff)

Begriffe:

Bestandsmasse = statistische Masse, deren zeitliche Abgrenzung ein Zeitpunkt ist; Bestandseinheiten gehören der Bestandsmasse über einen Zeitraum an.

Ereignismasse = statistische Masse, deren zeitliche Abgrenzung ein Zeitraum ist; der Ereigniseinheit kann nur ein Zeitpunkt (des Eintritts des Ereignisses) zugeordnet werden.

Korrespondierende Massen = Bestands- und Ereignismassen, die dadurch zusammengehören, dass die Ereignismassen Zu- und Abgänge der (korrespondierenden) Bestandsmassen beschreiben.

(a)

• Studierende an einer Universität (1)	Bestandsmasse
• Todesfälle in einer Gemeinde (2)	Ereignissmasse
• Personenkraftwagen der Deutsche Post AG (3)	Bestandsmasse
• Maschinenausfälle in einer Fabrik (4)	Ereignismasse
• Anmeldungen in einem Einwohnermeldeamt (5)	Ereignismasse
• Wartende Postkunden vor einem Postschalter (6)	Bestandsmasse
• Leichte Verkehrsunfälle in der Bundesrepublik (7)	Ereignismasse

(b) Die korrespondierenden Ereignismassen bzw. Bestandsmassen zu (a) lauten:

- Im-/Exmatrikulationen
- Mitglieder der Gemeinde
- Zugänge/Abgänge an Pkw bei der Post

- Maschinenbestand der Fabrik
- Einwohner der Stadt
- Zugänge/Abfertigungen von Kunden
- Erfasste Unfälle seit Erfassungsbeginn (es gibt keine Abgangsmasse zu dieser Bestandsmasse, es sei denn nach einem bestimmten Kriterium werden „alte“ Unfälle aus den Bestand enfernt.)

(c) Beispiele:

- Einwohner einer Stadt (Bestandsmasse) $\leftrightarrow$ Geburten, Todesfälle, Zu- bzw. Wegzüge (Ereignismassen)
- Lagerbestand (Bestandsmasse) $\leftrightarrow$ Anlieferung, Verkauf, Diebstahl (Ereignissmassen)

Aufgabe 3

Interpretieren Sie die abbgebildete grafische Darstellung (Quelle: Deutsche Bank Research). Gehen Sie dabei insbesondere auf die folgenden Fragen ein.

(a) Welche statistische(n) Masse(n) ist(sind) involviert? Handelt es sich dabei um Bestands- oder Ereignismassen? Geben Sie jeweils die sachliche, räumliche und zeitliche Abgrenzung an.

(b) Welches Merkmal (welche Merkmale) wurden bei dieser Untersuchung beobachtet? Welcher Art ist dieses Merkmal (sind diese Merkmale)? Ist es (sind sie) häufbar?

(c) Um was für Zahlenangaben handelt es sich dabei? Wie werden sie in der Statistik bezeichnet?

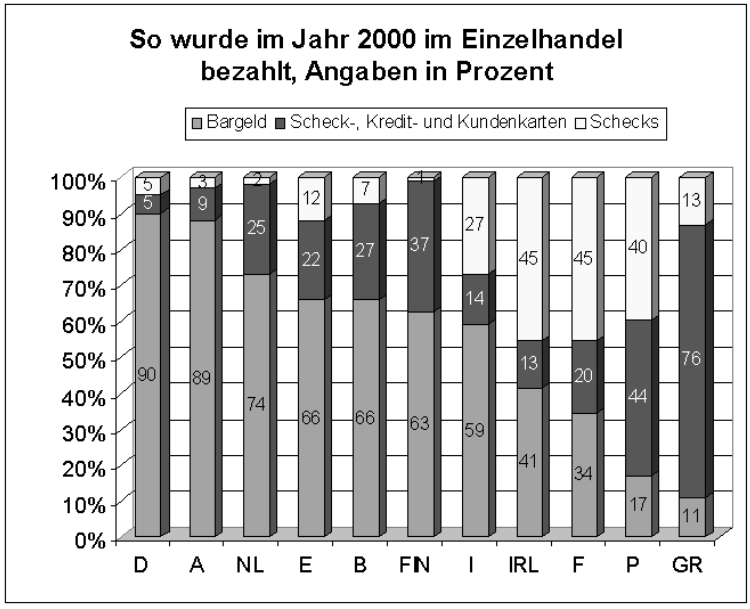


Abbildung 1: Einkäufe: Die Deutschen zahlen bar.

Lösung:

(a) Statistische Massen:
Für jedes der elf Länder werden die Einkäufe im Jahr 2000 betrachtet, bei denen eine der drei Zahlungsarten „Bargeld“, „Scheck-, Kredit- und Kundenkarten“, „Scheck“ vorlag, nicht also Einkäufe auf Rechnung, Ratenzahlung, Kredit,. . . .

Es ergeben sich elf statistische Massen mit den Abgrenzungen:

- **zeitlich:** im Laufe des Jahres 2000
- **sachlich:** Einkauf mit direkter Bezahlung (nicht auf Rechnung, ...)

- **räumlich:** jeweils im zugehörigen Land

Es handelt sich um Ereignismassen, da die zeitliche Abgrenzug jeweils ein Zeitraum ist.

(b) Das betrachtete Merkmal ist die Zahlungsart mit den Ausprägungen

- Bargeld
- Scheck-, Kredit- und Kundenkarten
- Scheck

Das Merkmal ist qualitativ und nicht häufbar[2].

(c) Mit den Zahlenangaben wird die Aufteilung der einzelnen statistischen Einheiten in den einzelnen Ländern auf die Zahlungsarten wiedergegeben. Es handelt sich um relative Häufigkeiten.

Bemerkung: Alternativ kann auch die statistische Masse aller Einkäufe mit den beiden Merkmalen „Land“ und „Zahlungsart“ betrachtet werden. In diesem Fall sind die Zahlenangaben bedingte relative Häufigkeiten (mit den einzelnen Ländern als Bedingung), siehe Deskriptive Statistik Kapitel 8.

[2]Unter der Annahme, dass die Rechnungssumme nicht auf zwei oder mehr Zahlungsarten aufgeteilt wird.

Aufgabe 4

(a) Welche der folgenden Merkmale sind *diskret*, welche sind *stetig*?

- Geschwindigkeit eines Pkws
- Hörerzahl der Vorlesung „Statistik I“
- Anzahl der Mitarbeiter eines Betriebes
- Einkommen
- Zeit für die Beschleunigung eines Pkws von 0 auf 100 km/h
- Klausurpunkte

(b) Welche Probleme ergeben sich bei der Messung *stetiger* Merkmale?

Lösung: (Deskriptive Statistik, S. 22)

(a)
- Geschwindigkeit eines Pkw — stetig
- Hörerzahl der Statistik I — diskret
- Anzahl der Mitarbeiter eines Betriebs — diskret
- Einkommen — stetig
- Zeit für die Beschleunigung eines Pkw von 0 auf 100 km/h — stetig
- Klausurpunkte — diskret

Messgrössen, die dem Raum, der Zeit, der Masse oder Funktionen dieser Grössen zugeordnet sind, können als stetig aufgefasst werden. Diskrete Merkmale werden oft wie stetige behandelt, wenn die Schrittgrösse zwischen den Ausprägungen in Relation zur Grösse sehr klein ist (z.B. monetäre Grössen: Einkommen, Umsatz, ...).

(b) Bedingt durch eine endliche Messgenauigkeit ist jedes theoretisch stetige Merkmal in der praktischen Anwendung letztlich diskret (entscheidend ist hier, dass jeder Punkt des entsprechenden Intervalls unabhängig von den technischen Möglichkeiten als Ausprägung gedacht werden kann). (s. Urlistenintervall)

Literatur: (Ferschl, 1978, S. 31f)

Aufgabe 5

(a) Erläutern Sie das Prinzip der mengen- und zahlenmäßigen Fortschreibung anhand der Pkw-Zulassungen in der Bundesrepublik im Monat April dieses Jahres.

(b) Überlegen Sie sich, wann eine Fortschreibung sinnvoll ist.

Lösung: (Deskriptive Statistik, S. 15)

(a) Begriffserläuterung: Fortschreibung = Korrektur einer Bestandsmasse in t_1 durch die korrespondierenden Ereignismassen für den Zeitraum $[t_1;t_2]$,$t_2>t_1$, um die Bestandsmasse zum Zeitpunkt t_2 zu bestimmen. Man unterscheidet zwischen:

- **Mengenmässige Fortschreibung :**

  $$\text{Endbestandsmasse} = \text{Anfangsbestandsmasse} \cup \text{Zugangsmasse} \setminus \text{Abgangsmasse}$$

- **Zahlenmässige Fortschreibung:**

  $$\text{Endbestand} = \text{Anfangsbestand} + \text{Zugänge} - \text{Abgänge}$$

  Beispiel[3]:

Pkw-Bestand zum Monatsanfang:	20365
+ Neuzulassungen im April 2000	5200
− Abmeldungen im April 2000	4735
= Pkw-Bestand zum Monatsende	20830

(b)
- mögliche Alternativen für die Aktualisierung einer Bestandsmasse:
  1. Neuerfassung der Bestandseinheiten
  2. Fortschreibung
- eine Fortschreibung ist sinnvoll, wenn
  1. die Bestandsmasse gegenüber den korrespondierenden Ereignismassen sehr gross ist bzw. die Ereignismassen ohnehin erhoben werden.
  2. die letzte Neuerfassung der Bestandsmasse noch nicht zu lange zurückliegt ($\rightarrow$ Gefahr der Fehlerakkumulation).

[3] Die Zahlen sind willkürlich gewählt und entsprechen nicht den tatsächlichen Werten.

Aufgabe 6

Für Merkmalsarten sind auch folgende Bezeichnungen üblich

(a) Nominalskala,

(b) Ordinalskala,

(c) Intervallskala

(d) Verhältnisskala und

(e) Absolutskala

Welche anderen Bezeichnungen sind hierfür üblich. Geben Sie jeweils ein Merkmal als Beispiel an und nennen Sie eine statistische Masse, bei der dieses Merkmal beobachtet werden kann. Welche Anforderungen müssen Skalentransformationen erfüllen, damit die wesentlichen Eigenschaften der jeweiligen Skala erhalten bleiben? Veranschaulichen Sie sich anhand von jeweils mindestens einem Beispiel, welche Folgen eine Missachtung des Skalentyps bei der Transformation haben kann.

Lösung: (Deskriptive Statistik, S. 23)

Begriffserläuterung:

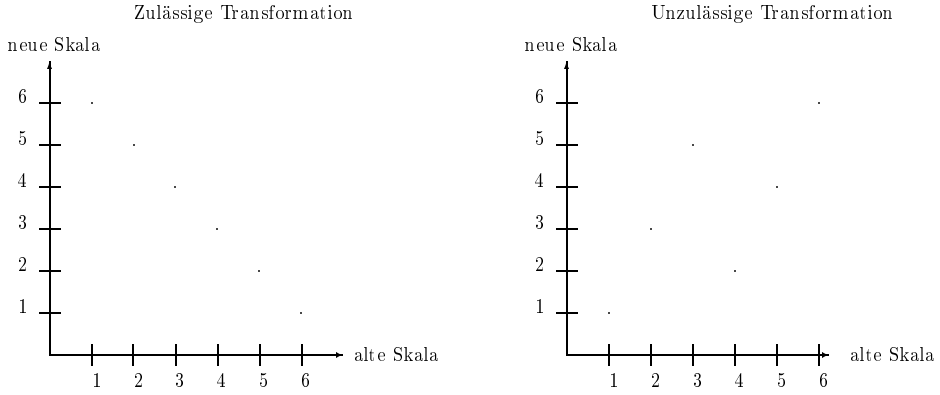
> Kodierung = Zuordnung von (in der Regel ganzer) Zahlen zu den Merkmalsausprägungen (z.B.: 1 zu weiblich, 0 zu männlich).
>
> Skalierung = relationstreue Abbildung der Merkmalsausprägungen auf reelle Zahlen (Scala italienisch: die Leiter, übertragen auch der Wertebereich bei Messungen). Skalierung steht auch für Verfahren, ein Rangmerkmal zu einem quantitativen Merkmal zu machen (z.B. Geben Sie auf einer Skala von -5 bis +5 die Zufriedenheit mit der aktuellen Regierung an).

(a) Nominalskala (andere Bezeichnung für qualitative Merkmale):

- Transformationen: alle bijektiven Abbildung möglich.
- Beispiel: Aufteilung der Bundesrepublik in Zustellbereiche durch die Post, Kodierung durch fünfstellige Zahlen.

(b) Ordinalskala (andere Bezeichnung für Rangmerkmale):

- Transformationen: zusätzlich zur Bijektivität muss die Ordnung erhalten bleiben, d.h. die Abbildung muss streng monoton steigend oder streng monoton fallend sein.
- Beispiel: Schulnoten „sehr gut“, „gut“, ..., „mangelhaft“
- übliche Skalierung: sehr gut $\to$ 1, gut $\to$ 2, ..., mangelhaft $\to$ 6.



Die nun folgenden Skalentypen c), d) und e) sind metrische Skalen = Kardinalskalen (andere Bezeichnung für quantitative Merkmale)

(c) Intervallskalen:

- Eigenschaft: Nullpunkt und Einheit der Skala sind willkürlich gewählt, d.h. Differenzen von Werten sind aussagekräftig, Verhältnisse von Ausprägungen nicht.
- zulässige Transformationen: lineare Abbildung vom Typ $y = ax + b, a \neq 0$ (in der Regel $a > 0$).
- Beispiel: Temperatur °Celsius $\leftrightarrow$ °Fahrenheit, wobei die Umrechnung gemäß der Transformation $y = \frac{9}{5}x + 32$ erfolgt; hierbei bezeichnet x die Temperatur in °C.

20°C	-	10°C	=	60°C	-	50°C	
↓		↓		↓		↓	Bemerkung: Verhältnisse von Differenzen
68°F	-	50°F	=	140°F	-	122°F	

können verglichen werden (sie sind zusätzlich invariant unter zulässigen Transformationen).

(d) Verhältnisskala:

- Nullpunkt ist durch natürliche Weise vorgegeben, d.h. neben Differenzen sind auch Verhältnisse aussagekräftig.
- zulässige Transformation: $y = ax,\ a \neq 0$ (in der Regel $a > 0$).
- Beispiele: Alter, Gewicht, Einkommen.

(e) Absolutskala:

- Nullpunkt und Einheit sind natürlich vorgegeben, d.h. es sind keine Transformationen möglich
- Beipiele: Stückzahl, Semesterzahl

Eine Missachtung der Skalenart kann zu folgenden Problemen führen:

- Überinterpretation bzw. Informationsverlust bei Wechsel zwischen den Skalen
- Verlust bzw. Verfälschung der Aussagekraft von Messwerten

Aufgabe 7

(a) Geben Sie bei den folgenden Merkmalen an, welcher Art sie sind. Welche der Merkmale sind häufbar?

- Unfallursache
- Datum des Semesterendes
- Hubraum eines Motors
- erlernter Beruf
- Windstärke
- Telefonnummer eines Faxgerätes
- Kinderzahl
- Körpergröße

(b) Im Studiengang Wirtschaftsingenieurwesen wird die Möglichkeit geschaffen, im Hauptstudium freiwillig eines oder mehrere der Fächer Statistik, Informatik und Operations Research zusätzlich zu den Pflichtfächern zu belegen. Erläutern Sie, wie das häufbare Merkmal „zusätzliche Fächer“ der statistischen Einheiten „Studierende des Wirtschaftsingenieurwesens“ auf ein nicht häufbares zurückgeführt werden kann.

(c) Machen Sie sich an mindestens zwei Beispielen den Unterschied in der Verwendung der Begriffe „Merkmalsausprägung“ und „Merkmalswert“ deutlich.

Lösung: (Deskriptive Statistik, S. 17)

* nicht häufbares Merkmal: Merkmal, bei dem jedem Merkmalsträger genau eine Merkmalsausprägung als Mermalswert zugeordnet ist.

* häufbares Merkmal: Merkmal, bei dem den Merkmalsträgern gleichzeitig mehrere Merkmalsausprägungen zugeordnet werden können.

(a)

Unfallursache	qualitatives Merkmal bzw. Nominalskala/häufbar
Datum des Semesterendes	quantitatives Merkmal; Intervallskala/nicht häufbar
Hubraum eines Motors	quantitatives Merkmal; Verhältnisskala/nicht häufbar
erlernter Beruf	qualitatives Merkmal bzw. Nominalskala/häufbar
Windstärke	quantitatives Merkmal; Verhältnisskala/nicht häufbar[4]
Telefonnummer eines Faxgerätes	qualitatives Merkmal bzw. Nominalskala/nicht häufbar
Kinderzahl	quantitatives Merkmal; Absolutskala/nicht häufbar
Körpergröße	quantitatives Merkmal; Verhältnisskala/nicht häufbar

(b) Auflösung durch Bildung eines neuen Merkmals, dessen Ausprägungen durch die möglichen Kombinationen der ursprünglichen Merkmalsausprägungen gebildet werden (Bildung der Potenzmenge).

häufbares Merkmal mit n Ausprägungen $\rightarrow$ nicht häufbares Merkmal mit 2^n Ausprägungen

Hier: Ausprägungen $\in \mathcal{P}(\{Statistik,\ Informatik,\ OR\})$

(c) Merkmalausprägung: möglicher Wert einer statistischen Einheit unabhängig davon, ob beobachtet oder nicht.

Merkmalswert: einer statistischen Einheit zugeordnete, d.h. tatsächlich bei ihr beobachtete Ausprägung.

Beispiele:

1. Bei einer statistischen Masse von 10 PKW wird das Merkmal Farbe beobachtet. Auch wenn die Farbe rot nicht vorkommt, ist rot eine Merkmalsausprägung. In diesem Fall aber kein Merkmalswert.
2. Bei der Messung des Gewichts von Paketen in Gramm kann jede natürliche Zahl von einem Minimalwert m bis zu einem Maximalwert M auftreten. Merkmalsausprägungen sind also alle natürlichen Zahlen zwischen m und M. Merkmalswerte sind aber nur die bei den betrachteten Paketen gemessenen Gewichte.

Eine Liste der Merkmalsausprägungen enthält sinnvollerweise keine Wiederholungen. Eine Liste der Merkmalswerte (die Urliste) wird meist Wiederholungen aufweisen, da statistische Einheiten übereinstimmende Merkmalswerte haben können.

Aufgabe 8

Bei der 2001. Vorführung von „Müllers Büro“ in der Karlsruher Schauburg wurde das Alter von 50 Studierenden ermittelt:

38, 23, 17, 34, 43, 22, 21, 26, 19, 25, 29, 18, 22, 24, 20, 16, 25, 24, 22, 21, 26, 28, 45, 22, 18, 23, 27, 21, 34, 42, 21, 62, 18, 17, 24, 36, 30, 20, 25, 21, 37, 29, 33, 19, 26, 28, 22, 41, 24, 26.

(a) Erstellen Sie für das Merkmal „Alter“ eine geordnete Urliste.

(b) Durch weiteres Nachfragen stellt man fest, dass 23 der 50 Zuschauer an der Universität studieren. Diese gaben als Studienfach an (C$\hat{=}$ Ciw, W$\hat{=}$ WiWi, M$\hat{=}$ Machbau, E$\hat{=}$ Etech, I$\hat{=}$ Info, Ch$\hat{=}$ Chemie, B$\hat{=}$ Biologie, A$\hat{=}$ Architektur, Ma$\hat{=}$ Mathematik):

M, C, B, W, A, Ma, M, Ch, B, M, I, Ma, I, E, A, M, I, E, C, B, C, Ch, I

Sortieren Sie die Urliste, und geben Sie für das Merkmal „Studienfach“ eine Häufigkeitsverteilung an.

(c) Die folgende Stiel-und-Blatt-Darstellung gibt die Werte des Merkmals „Alter“ für die 23 Studierenden wieder. Fassen Sie die Daten mittels einer geeigneten Klassierung zusammen (5 Klassen).

1|
1| 6 7 8 9 9
2| 0 1 2 3 3 4
2| 5 5 6
3| 0 3 4
3| 6 7 8
4| 1 2
4| 5

Lösung: (Deskriptive Statistik, S. 28, 31ff, 37ff)

(a) Geordnete Urliste: der Größe nach geordnete Liste der beobachteten Merkmalswerte. Schreibweise: $x_{(1)}, \ldots, x_{(n)}$.

hier: $n = 50$, Merkmal „Alter der Zuschauer“. Das Erstellen der geordneten Urliste erfolgt übersichtlich mit Hilfe des Stiel- u. Blatt-Diagramms:

1	
1	7 9 8 6 8 8 7 9
2	3 2 1 2 4 0 4 2 1 2 3 1 1 4 0 1 2 4
2	6 5 9 5 6 8 7 5 9 6 8 6
3	4 4 0 3
3	8 6 7
4	3 2 1
4	5
5	
5	
6	2
6	

1	
1	6 7 7 8 8 8 9 9
2	0 0 1 1 1 1 1 2 2 2 2 2 3 3 4 4 4 4
2	5 5 5 6 6 6 6 7 8 8 9 9
3	0 3 4 4
3	6 7 8
4	1 2 3
4	5
5	
5	
6	2
6	

$\Rightarrow$ Es ergibt sich die geordnete Urliste $x_{(1)}, \ldots, x_{(50)}$: 16, 17, 17, 18, 18,..., 43, 45, 62.

(b) $n = 23$, Merkmal: Studienfach

Nominalskala $\Rightarrow$ Stiel- und Blatt-Darstellung nicht möglich, nur Umsortieren möglich.

- M, M, M, M, C, C, C, B, B, B, W, A, A, Ma, Ma, Ch, Ch, I, I, I, I, E, E
  bzw.
- M(4), C(3), B(3), W(1), A(2), Ma(2), Ch(2), I(4), E(2)
  (Reihenfolge beliebig!)

Absolute/relative Häufigkeitsverteilung: Zusammenstellung der absoluten/relativen Häufigkeiten h(a) / p(a) für alle Merkmalsausprägungen a $\in$ M.

- $h : M \to \mathbb{N}_0$ mit $h(a) = \#\{i \in \{1, \ldots, n\} | x_i = a\}$
- $p : M \to [0, 1]$ mit $p(a) = \frac{h(a)}{n}$ (bzw. $p(a) = \frac{h(a)}{n} \cdot 100\%$)

Studienfach	M	C	B	W	A	Ma	Ch	I	E	$\sum$
abs.Häufigk.	4	3	3	1	2	2	2	4	2	23
rel.Häufigk.	$\frac{4}{23}$	$\frac{3}{23}$	$\frac{3}{23}$	$\frac{1}{23}$	$\frac{2}{23}$	$\frac{2}{23}$	$\frac{2}{23}$	$\frac{4}{23}$	$\frac{2}{23}$	1

Bemerkung: Bei Rangmerkmalen und quantitativen Merkmalen kann zusätzlich für jede Merkmalsausprägung $a \in M$ die Anzahl bzw. der Anteil der Merkmalswerte, die gemäß der Ordnung vor a kommen oder mit a übereinstimmen bzw. kleiner oder gleich a sind, bestimmt werden.Es ergeben sich die folgenden Funktionen:

absolute/relative Summenhäufigkeit: H(a)= $\sum_{a' \leq a} h(a')$, F(a)= $\sum_{a' \leq a} p(a')$

Bei quantitativen Merkmalen ($M \subseteq \mathbb{R}$) kann F(a) letztlich für alle $a \in \mathbb{R}$ bestimmt werden:

empirische Verteilungsfunktion $F(x) = \frac{1}{n}\#\{i \in \{1, \ldots, n\} | x_i \leq x\}$

(c) Klassierung: Zusammenfassung von Merkmalsausprägungen zu Klassen, jede Merkmalsausprägung gehört zu genau einer Klasse (Bildung eines neuen Merkmals mit einer kleineren Menge von Ausprägungen).
Bei quantitativen Merkmalen erfolgt die Klassierung durch eine Zerlegung von M in disjunkte Intervalle (Partition).

- Klassenbreite: $\Delta_I = \beta_I - \alpha_I$, i.d.R. für alle Klassen gleich.
- Klassenmitten: Es wird $z_I = \frac{1}{2}(\alpha_I + \beta_I)$, z.T. als Repräsentant / Ersatzwert für die Klasse verwendet. Bei diskreten Merkmalen sollte die Klassenmitte mit dem Mittel der Ausprägungen in der Klasse übereinstimmen. Bei stetigen Merkmalen ist die Messgenauigkeit beschränkt, d.h. der exakte Wert liegt in einem sogenannten Urlistenintervall; Klassengrenzen dürfen nicht im Inneren eines solchen Intervalls liegen.

- Anzahl der Klassen: beeinflusst den Informationsverlust. Wahl z.B. nach DIN 55302 oder „Wurzelregel“ ($\#Klassen \approx \sqrt{n}$)

Anzahl der Beobachtungswerte	Anzahl der Klassen
bis 100	mindestens 10
etwa 1.000	mindestens 13
etwa 10.000	mindestens 16
etwa 100.000	mindestens 20

- hier: n=23, $\sqrt{23} = 4.79$, in diesem Fall Wahl von 5 Klassen.
- Klassenbreite $\Delta I = \frac{1}{5}(45 - 16 + 1) = 6 \Rightarrow$ [16;22),[22;28), ..., [40;46).
- Das Alter wird in der Regel in vollendeten Lebensjahren angegeben, d.h. Urlistenintervalle sind z.B.: [16,17), [17,18), usw. Damit liegen die Klassengrenzen auf Grenzen von Urlistenintervallen.

	1	2	3	4	5	
Klasse	16 b.u. 22	22 b.u. 28	28 b.u. 34	34 b.u. 40	40 b.u. 46	$\sum$
abs.H	7	7	2	4	3	23
rel.H	$\frac{7}{23}$	$\frac{7}{23}$	$\frac{2}{23}$	$\frac{4}{23}$	$\frac{3}{23}$	1

- Bemerkung: absolute/relative Summenhäufigkeit bei klassierten Merkmalen für eine Klassengrenze a:

  $\mathrm{H(a)}= \sum_{I:\beta_I \leq a} h(I)$, $\mathrm{F(a)}= \sum_{I:\beta_I \leq a} p(I)$

Aufgabe 9

(a) Von 2000 im Rahmen einer Umfrage befragten Personen haben 50 das Auto X als Wagen des Jahres gewählt.

Benennen Sie den Merkmalswert und bestimmen Sie seine absolute und relative Häufigkeit.

(b) Geben Sie eine allgemeingültige Unter- und Obergrenze für relative Häufigkeiten an.

(c) Welchen Wert nimmt die Summe der relativen Häufigkeiten bei einer Erhebung normalerweise an? Gibt es Ausnahmen?

Lösung: (Deskriptive Statistik, S. 29ff)

(a) absolute Häufigkeit einer Merkmalsausprägung a: Anzahl der Merkmalswerte der Urliste, die mit a übereinstimmen (analog relative Häufigkeit).

- Merkmal: „Auto des Jahres“, genauer Wunschauto (der befragten Person) für die Auszeichnung „Auto des Jahres“.
- Merkmalswert a: Auto X
- absolute Häufigkeit: $h(a) = 50$
- relative Häufigkeit: $p(a) = \frac{50}{2000} = \frac{1}{40}$

(b) Für nicht häufbare Merkmale gilt: $0 \leq p(a) \leq 1$ (bzw. $0 \leq p(a)$ in $\% \leq 100$), wobei $p(a) := \frac{h(a)}{n} = \frac{h(a)}{\sum_{a \in M} h(a)}$.

(c) Aus (b) folgt:

$$\sum_{a \in M} p(a) = \sum_{a \in M} \frac{h(a)}{\sum_{a \in M} h(a)} = 1$$

Ausnahmen sind bei häufbaren Merkmalen möglich, z. B. erlernter Beruf:

Person	Beruf
A	Kaufmann
B	Maurer
C	Fahrer, Maurer
D	Kaufmann
E	Kaufmann, Maurer

$\Rightarrow$

Beruf	
Kaufmann	$\frac{3}{5} \mathrel{\hat{=}} 60\%$
Maurer	$\frac{3}{5} \mathrel{\hat{=}} 60\%$
Fahrer	$\frac{1}{5} \mathrel{\hat{=}} 20\%$
$\sum$	$\frac{7}{5} \mathrel{\hat{=}} 140\%$

Aufgabe 10

Die folgenden Daten bzgl. der Entwicklung der Studierendenzahlen in Westdeutschland können Sie dem Kompendium „Zahlen 1996“ des Instituts der Deutschen Wirtschaft Köln entnehmen:

Studenten in Westdeutschland nach Hochschulart in 1000			
	1960	1975	1994
Insgesamt	291.1	840.8	1676.1
Uni und PH	238.4	680.2	1253.4
Kunsthochschulen	8.5	15.4	24.5
Fachhochschulen	44.2	145.2	398.2

(a) Stellen Sie für das Jahr 1994 die Aufteilung der Studierenden auf die verschiedenen Hochschularten mit Hilfe eines Stabdiagramms dar. Was müssen Sie bei Verwendung eines Flächen- bzw. Volumendiagramms beachten?

(b) Veranschaulichen Sie sich die Daten der Tabelle mit Hilfe eines kombinierten Flächen-/Kreissektorendiagramms. Kennen Sie Alternativen zu diesem Diagramm?

Lösung: (Deskriptive Statistik, S. 42ff)

(a) Stab-/ Linien bzw. Säulen-/Balkendiagramm: Höhe der Stäbe/...ist proportional zur darzustellenden absoluten/relativen Häufigkeit; Balken-/Säulenbreite bzw. Grundfläche ist ohne Bedeutung (sollte übereinstimmend sein!)

Flächen-/Volumendiagramm: Flächen bzw. Volumina sind proportional zu den darzustellenden absoluten/relativen Häufigkeiten (Darstellung flächen-/volumenproportional)

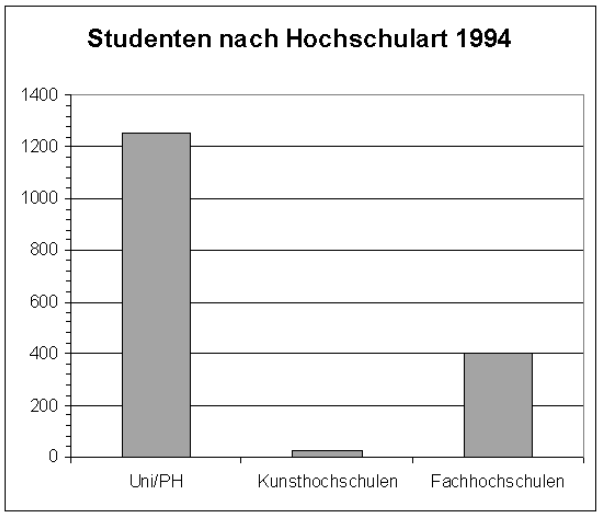


Abbildung 2: Balkendiagramm

(b) Kreissektorendiagramm = Flächendiagramm; relative Häufigkeiten sind proportional zur Größe der Sektoren bzw. der Sektorenwinkel

Kombiniertes Flächen-/Kreissektorendiagramm: Kreissektorendiagramm, bei dem die Gesamtfläche zusätzlich proportional zur Anzahl der statistischen Einheiten der betrachteten Masse ist. Sinnvoll, wenn mehrere Massen verglichen werden sollen.

Studenten nach Hochschulart in 1000				relative Häufigkeiten		
	West					
	1960	1975	1994	1960	1975	1994
Insgesamt	291.1	840.8	1676.1	1	1	1
Uni und PH	238.4	680.2	1253.4	0.819	0.809	0.748
Kunsthochschulen	8.5	15.4	24.5	0.029	0.018	0.015
Fachhochschulen	44.2	145.2	398.2	0.152	0.173	0.238
Radius des Kreises in mm	$\sqrt{291.1} = 17.06$	$\sqrt{840.8} = 29$	$\sqrt{1676.1} = 40.9$			

Exemplarisch ist hier ein Kreissektorendiagramm für die Studentenzahlen 1994 abgebildet. Für die anderen Jahre kann analog verfahren werden, wobei darauf zu achten ist, die Radien proportional zu den in der Tabelle angegebenen zu wählen.

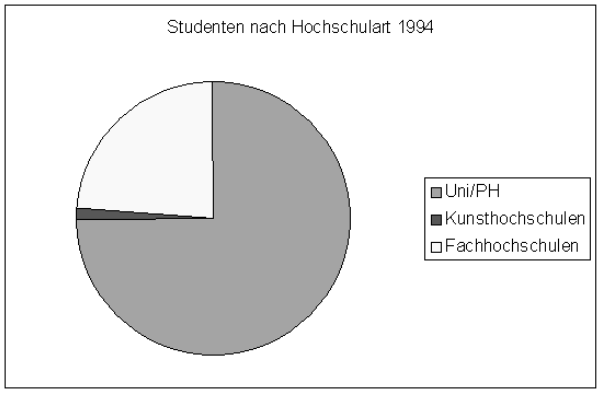


Abbildung 3: Kreissektorendiagramm

Eine alternative Darstellungsmöglichkeit bietet ein gestapeltes Balkendiagramm.

Aufgabe 11

(a) Eine Erhebung des Merkmals „Körpergröße“ in einer Gemeinde ergab folgende Messwerte:

$[0.30, 0.60)$	$[0.60, 0.90)$	$[0.90, 1.20)$	$[1.20, 1.50)$	$[1.50, 1.80)$	$[1.80, 2.10)$
500	3580	4760	3800	10128	2780

Zeichnen Sie das Histogramm und die Summenhäufigkeitsfunktion.

(b) Zu einer detaillierteren Analyse der Daten werden diese wie folgt umgruppiert:

$[0.30, 0.60)$	$[0.60, 0.75)$	$[0.75, 0.90)$	$[0.90, 1.05)$	$[1.05, 1.20)$	$[1.20, 1.50)$
500	2500	1080	1000	3760	3800

$[1.50, 1.65)$	$[1.65, 1.80)$	$[1.80, 2.00)$
2000	8128	2780

Zeichnen Sie das Histogramm für diese Häufigkeitstabelle.

(c) Zeichnen Sie in die beiden Histogramme aus (a) und (b) jeweils ein Häufigkeitspolygon ein. Welche Schwierigkeiten ergeben sich beim Erstellen und Interpretieren der Diagramme?

(d) Berechnen Sie mit Hilfe der beiden Häufigkeitstabellen den Anteil der Merkmalswerte in der Urliste,

- die kleiner oder gleich 0.7 sind,
- im abgeschlossenen Intervall $[1.4, 1.6]$ liegen bzw.
- größer oder gleich 1.65 sind,

und vergleichen Sie die Ergebnisse.

Lösung: (Deskriptive Statistik, S. 52ff)

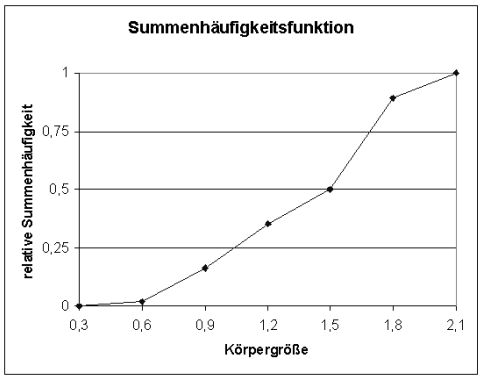
(a) Histogramm: Flächendiagramm für klassierte quantitative Merkmale; Höhe des Rechtecks über einem Intervall I: $\frac{h(I)}{\Delta_I}$ bzw. $\frac{p(I)}{\Delta_I}$; Problem offener Randklassen: entweder fehlende Klassengrenze „sinnvoll“ ergänzen oder Höhe des Rechtecks beträgt 0 und Vermerk in der Grafik oder Legende zur Grafik. $\frac{h(I)}{\Delta_I}$ bzw. $\frac{p(I)}{\Delta_I}$ heißt absolute/relative Häufigkeitsdichte.

Summenhäufigkeitsfunktion (relative bzw. absolute) für klassierte stetige quantitative Merkmale:

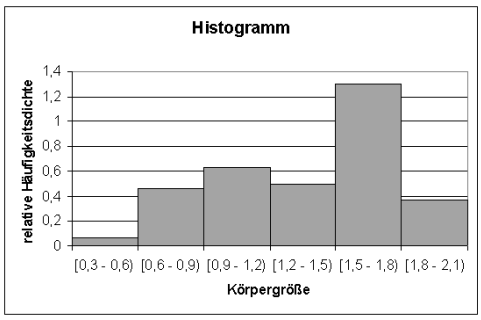
$$SF(z) = F(\alpha_I) + \frac{z - \alpha_I}{\beta_I - \alpha_I} p(I), \text{ für } z \in I \ = \ [\alpha_I, \beta_I) \text{ bzw. } z \in I \ = \ (\alpha_I, \beta_I]$$

$SF(z)$ entspricht der Histogrammfläche links von z und basiert auf der Annahme der Gleichverteilung der Beobachtungen in den Klassen.

I	ΔI	h(I)	p(I)	$\frac{p(I)}{\Delta I}$	$F(\beta_I)$
$[0.3; 0.6)$	0.3	500	0.02	0.067	0.02
$[0.6; 0.9)$	0.3	3580	0.14	0.467	0.16
$[0.9; 1.2)$	0.3	4760	0.19	0.63	0.35
$[1.2; 1.5)$	0.3	3800	0.15	0.5	0.5
$[1.5; 1.8)$	0.3	10128	0.39	1.3	0.89
$[1.8; 2.1)$	0.3	2780	0.11	0.367	1
		25548	1		



(a) Summenhäufigkeitsfunktion



(b) Histogramm

Abbildung 4:

(b)

I	ΔI	h(I)	p(I)	$\frac{p(I)}{\Delta I}$	$F(\beta_I)$
$[0.3; 0.6)$	0.3	500	0.02	0.067	0.02
$[0.6; 0.75)$	0.15	2500	0.10	0.67	0.12
$[0.75; 0.9)$	0.15	1080	0.04	0.267	0.16
$[0.9; 1.05)$	0.15	1000	0.04	0.267	0.20
$[1.05; 1.2)$	0.15	3760	0.14	0.93	0.34
$[1.2; 1.5)$	0.3	3800	0.15	0.5	0.49
$[1.5; 1.65)$	0.15	2000	0.08	0.53	0.57
$[1.65; 1.8)$	0.15	8128	0.32	2.13	0.89
$[1.8; 2.0)$	0.2	2780	0.11	0.55	1
		25548	1		

Die Grafiken können nun analog zu Aufgabenteil (a) angefertigt werden.

(c) Häufigkeitspolygon: Ein Polygonzug, der die Mitten aller oberen Rechteckseiten eines Histogramms verbindet. Problem bilden hierbei die Randklassen. Mögliche Konventionen:

1. Polygonzug endet an den Randklassen $\rightarrow$ wirkt unvollständig.
2. Einführung zweier neuer Randklassen mit $h(I) = 0 \rightarrow$ Gefahr der Fehlinterpretation.

Haben alle Klassen die gleiche Breite und wird nach 2. vorgegangen, so entspricht die Fläche unter dem Polygonzug der des Histogramms. Andernfalls ist keine einheitliche Interpretation der Fläche unter dem Polygonzug möglich. Häufigkeitspolygone werden wegen der angegebenen Problematik selten angewendet.

(d) Wichtig: Im folgenden wird unterstellt, dass die Merkmalswerte in jeder Klasse gleichverteit sind.

Erste Häufigkeitstabelle ($n = 25548$):

Relative Häufigkeit der Merkmalswerte ≤ 0.7 :

$$SF(0.7) = \frac{500}{25548} + \frac{1}{3} \cdot \frac{3580}{25548} \approx 0.06628$$

Relative Häufigkeit der Merkmalswerte in [1.4, 1.6]:

$$\begin{aligned} SF(1.6) - SF(1.4) &= \frac{1}{25548} \cdot (500 + 3580 + 4760 + 3800 + \frac{1}{3} \cdot 10128 - \\ &\quad (500 + 3580 + 4760 + \frac{2}{3} \cdot 3800)) \approx 0.1817 \end{aligned}$$

Relative Häufigkeit der Merkmalswerte ≥ 1.65 :

$$1 - SF(1.65) = \frac{1}{25548} \cdot (\frac{10128}{2} + 2780) \approx 0.307$$

Zweite Häufigkeitstabelle:

$$\begin{aligned} SF(0.7) &= \frac{1}{25548} \cdot (500 + \frac{2}{3} \cdot 2500) \approx 0.0848 \\ SF(1.6) - SF(1.4) &= \frac{1}{25548} \cdot (H(1.5) + \frac{2}{3} \cdot 2000 - H(1.2) - \frac{2}{3} \cdot 3800) \approx 0.1018 \\ 1 - SF(1.65) &= p([1.65, 1.80)) + p([1.80, 2.00)) = \frac{8128 + 2780}{25548} \approx 0.427 \end{aligned}$$

Aufgabe 12

Im Rahmen einer Umfrage unter Gästen der Fast-Food-Kette „Crazy Cow“ wurde bei 50 Personen das Körpergewicht (in [kg]) gemessen:

$62, 72, 83, 98, 52, 84, 94, 95, 68, 65, 85, 50, 72, 59, 76, 61, 60, 72, 93, 47$

$61, 58, 75, 65, 79, 81, 55, 63, 53, 50, 62, 86, 50, 74, 94, 70, 88, 48, 79, 76$

$70, 47, 72, 58, 83, 74, 85, 79, 75, 76$

(a) Berechnen Sie aus den Daten der Urliste alle Ihnen aus der Vorlesung bekannten Lage- und Streuungsparameter.

(b) Klassieren Sie die Daten sinnvoll, und zeichnen Sie ein Histogramm.

(c) Zeichnen Sie die empirische Verteilungsfunktion für den Datensatz sowie die Summenhäufigkeitsfunktion für die klassierten Daten. Wie läßt sich die Summenhäufigkeitsfunktion interpretieren?

(d) Berechnen Sie nun aus den klassierten Daten analog zu (b) die Lageparameter und die Varianz. Welche Annahmen treffen Sie dabei?

(e) Vergleichen Sie die Ergebnisse aus (d) mit denen aus (a). Was läßt sich über die Verteilung innerhalb der Klassen aussagen?

Lösung: (Deskriptive Statistik, S. 63ff)

(a) Um eine bessere Übersicht über die Daten zu erhalten wird eine Stiel- und Blatt-Darstellung erstellt.

4	7	7	8					
5	0	0	0	2	3			
5	5	8	8	9				
6	0	1	1	2	2	3		
6	5	5	8					
7	0	0	2	2	2	2	4	4
7	5	5	6	6	6	9	9	9
8	1	3	3	4				
8	5	5	6	8				
9	3	4	4					
9	5	8						

Lageparameter geben Information über die Lage einer Verteilung:

- Unabhängig von der Skalierung der vorliegenden Daten existiert der Modus oder Modalwert= a_M =arg $\max_{a \in M}(h(a))$
  $x_M = 72 = x_{(24)} = \cdots = x_{(27)}$
- Sind die Daten mindestens ordinalskaliert, so existiert der Median oder Zentralwert:
  $x_Z = x_{(25)} = x_{(26)} = 72 \qquad x_Z = \begin{cases} x_{(\frac{n+1}{2})} & \text{für n ungerade} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] & \text{für n gerade} \end{cases}$
- Für kardinalskalierte Daten existiert weiterhin das arithmetische Mittel:
  $\bar{x} = 70.68 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Streuungsparameter machen eine Aussage über die Gestalt der Verteilung, unabhängig von der Lage und sind nur für quantitative Merkmale berechenbar:

- Spannweite: R = $\max_i x_i - \min_i x_i$ (weniger anfällig gegen Ausreißer ist der Quartilsabstand QA, siehe Übung 14)
- mittlere absolute Abweichung (vom Zentralwert): $d = \frac{1}{n}\sum_i |x_i - x_z|$
- Varianz: $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ (korr. Varianz $s^{*2} = \frac{n}{n-1}s^2$ bei Stichproben)
- Standardabweichung: $s = \sqrt{s^2}$ (bzw. $s^* = \sqrt{s^{*2}}$)

Mit den vorliegenden Daten ergibt sich: $R = 51$; $d = 11.6$; $s^2 = 193.7$; $s^{*2} = 197.61$, $s = 13.9$; $s^* = 14.1$ (Zum Zusammenhang zwischen Lage- und Streuungsparametern vgl. Übung 13)

(b) Es bietet sich die Wahl von 6 Klassen an:

$[\alpha_I; \beta_I)$	z_I	Δ_I	$p(I)$	$\frac{p(I)}{\Delta_I}$	$SF(\beta_I)$
$[39.5, 49.5)$	44.5	10	$\frac{3}{50}$	0.006	0.06
$[49.5, 59.5)$	54.5	10	$\frac{9}{50}$	0.018	0.24
$[59.5, 69.5)$	64.5	10	$\frac{9}{50}$	0.018	0.42
$[69.5, 79.5)$	74.5	10	$\frac{16}{50}$	0.032	0.74
$[79.5, 89.5)$	84.5	10	$\frac{8}{50}$	0.016	0.9
$[89.5, 99.5)$	94.5	10	$\frac{5}{50}$	0.01	$\underbrace{1.0}_{c)}$

Wahl der Klassengrenzen: vgl. Übung 8. Urlistenintervall ist hier z.B. für die Gewichtsangabe 62: [61.5 , 62.5).

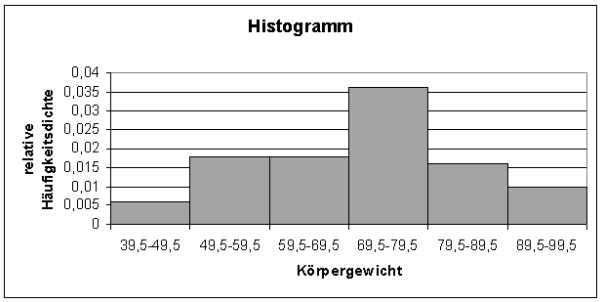
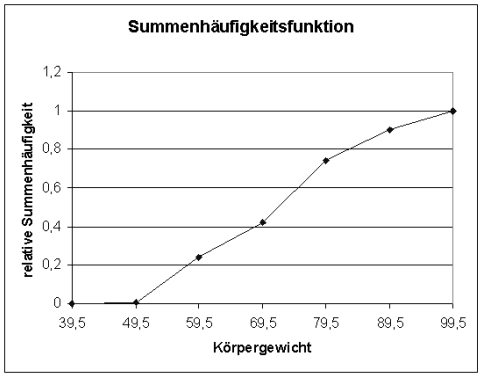
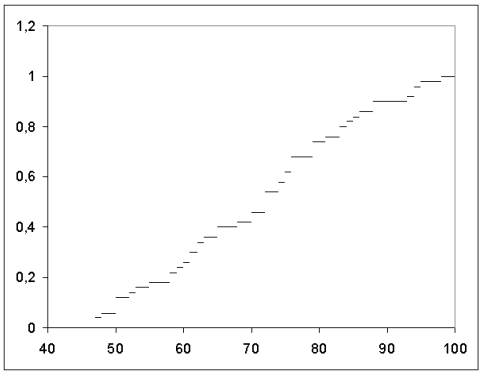


Abbildung 5: Histogramm

(c)



(a) Summenhäufigkeitsfunktion



(b) Verteilungsfunktion

Die Summenhäufigkeitsfunktion kann als Näherung der empirischen Verteilungsfunktion interpretiert werden.

(d) Die Berechnung der Lage- und Streungsparameter beruht auf der Annahme, dass die Klassenmitten als Repräsentant für die Klassen geeignet sind.

- Modale Klasse I_M mit maximaler Häufigkeitsdichte, d.h. mit $\frac{h(I_M)}{\Delta I_M} = \max_I \frac{h(I)}{\Delta I}$ bzw. $\frac{p(I_M)}{\Delta I_M} = \max_I \frac{p(I)}{\Delta I}$. Es ergibt sich $I_M = [69.5, 79.5)$
- (Feinberechneter) Zentralwert a_Z mit $\mathrm{SF}(a_Z) = 0.5$ oder in anderen Worten: eine senkrechte Linie durch a_Z halbiert die Histogrammfläche. Berechnung:
  1. Man bestimmt die Einfallsklasse, d.h. diejenige Klasse $[\alpha_E; \beta_E)$ mit $\mathrm{SF}(\alpha_E) < \frac{1}{2} < SF(\beta_E)$ (bei '=' ist a_Z die entsprechende Klassengrenze). Es ergibt sich:

$$I_E = [69.5, 79.5)$$

  2. a_Z entspricht nun der Abszisse desjenigen Punktes auf der Geraden durch $SF(\alpha_E), SF(\beta_E))$ Ordinate 0,5 ist. Man erhält:

$$SF(a_z) \stackrel{!}{=} \frac{1}{2} \quad \Rightarrow \quad a_z = \alpha_E + \frac{\frac{1}{2} - SF(\alpha_E)}{SF(\beta_E) - SF(\alpha_E)}(\beta_E - \alpha_E) = 69.5 + \frac{\frac{1}{2} - 0.42}{0.32} \cdot 10 = 72.$$

Im weiteren bezeichnet z_I die Klassenmitte der Klasse I, d.h. $(z_I = \frac{1}{2}(\beta_I + \alpha_I))$.

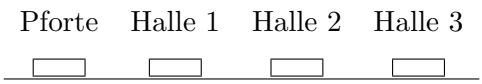
- Arithmetisches Mittel: $\bar{x} = \frac{1}{n} \sum_I z_I h(I) = \sum_I z_I p(I) \;=\; 70.9$
- Spannweite: $R = \max_{I:p(I)\neq 0} \beta_I - \min_{I:p(I)\neq 0} \alpha_I \;=\; 60$ (geringe Aussagekraft, da über die Wahl der Klassengrenzen manipulierbar!)
- Mittlere absolute Abweichung (vom feinberechneten Zentralwert x_Z): $d = \frac{1}{n} \sum_I |z_I - x_z| h(I) = \sum_I |z_I - a_z| p(I) \;=\; 11.5$
- Varianz: $s^2 = \frac{1}{n} \sum_I (z_I - \bar{x})^2 h(I) = \sum_I (z_I - \bar{x})^2 p(I) \;=\; 187.04$

(e) Zentralwert bei den Ausgangsdaten und feinberechneter Zentralwert bei den klassierten sind gleich. Es gibt Abweichungen bei den übrigen Lageparametern, da die Verteilung in den Klassen nicht gleichmäßig ist, die Klassenmitten z_I also keine perfekten Repräsentanten sind. Die Varianz wird durch die Klassierung geringer, da durch die Klassierung auch Information (z.B. bzgl. der Streuung) verloren geht. Die aus den klassierten Daten berechnete Varianz ist tendenziell niedriger, da die Streuung innerhalb der Klassen vernachlässigt wird.

Aufgabe 13

Im folgenden sollen Sie sich die Bedeutung der verschiedenen in der Vorlesung betrachteten Lageparameter noch einmal anhand eines Beispiels verdeutlichen:

Situation: *Ein Werk mit 1000 Beschäftigten, das in Form einer Werkstraße angelegt ist (vgl. Skizze), soll eine neue Kantine erhalten. Ihr Standort soll möglichst „optimal“ gewählt werden, wobei die Geschäftsleitung sich bzgl. der zu verwendenden Zielfunktion noch nicht ganz einig werden konnte.*



Anzahl Beschäftigte	3	200	300	497
Abstand von der Pforte	0	100	600	900

Im folgenden soll m den noch zu bestimmenden Abstand der Kantine und x_i den Abstand des Arbeitsplatzes des i'ten Mitarbeiters von der Pforte bezeichnen.

(a) Die Geschäftsleitung beschließt, die Kantine in einer der drei Hallen einzurichten. Zeigen Sie, dass der Modalwert x_M der Entfernungen x_i der Arbeitsplätze von der Pforte den optimalen Standort der Kantine angibt, wenn dieser die „Gesamtrüstzeiten des Kantinengangs“ aller Mitarbeiter minimieren soll (muss ein Mitarbeiter für den Besuch der Kantine die Halle, in der er arbeitet, verlassen, so entsteht durch Umziehen etc. ein Verlust von w Zeiteinheiten).

(b) Zeigen Sie, dass das arithmetische Mittel der Wegstrecken x_i der Mitarbeiter zur Pforte die Summe der quadrierten Wegstrecken $\sum(x_i - m)^2$ minimiert (Sie können sich das Quadrieren als eine Form der Gewichtung der Wegstrecken vorstellen, d.h. lange Wege werden im Vergleich zu relativ kurzen überproportional bestraft).

(c) Zeigen Sie, dass der Median x_z der Wegstrecken zur Kantine die Summe der Wegstrecken aller Beschäftigten, d.h. $\sum |x_i - m|$, minimiert.

(d) Erklären Sie basierend auf den Ergebnissen der Aufgabenteile (b) und (c), warum die mittlere absolute Abweichung mit Hilfe des Medians und die Varianz mit Hilfe des arithmetischen Mittels berechnet wird.

Lösung: (Deskriptive Statistik, S. 63ff)

Bezeichnungen: Der Standort der Kantine sei m (= Entfernung von der Pforte), x_i sei der Abstand des Arbeitsplatzes des i-ten Mitarbeiters von der Pforte.

(a) Minimierung der Gesamtrüstzeit:

$$\min_m \sum_{i=1}^{1000} f(x_i, m) \quad \text{mit } f(x_i, m) \;=\; \begin{cases} 0 & \text{, falls } x_i = m \\ 1 & \text{, falls } x_i \neq m \end{cases}$$

Es ergibt sich als optimaler Standort: $\boxed{x_{mod} = 900}$

(b) Minimierung der Summe aller quadrierten Wegstrecken der Mitarbeiter:

$$\min_{m \in \mathbb{R}} \sum_{i=1}^{n} (x_i - m)^2$$

Sei $WQ(m) := \sum_{i=1}^{n} (x_i - m)^2$. Es ergibt sich:

$$\begin{aligned} \frac{dWQ(m)}{dm} &= \sum_{i=1}^{n} 2 \cdot (x_i - m) \cdot (-1) = -2 \cdot \sum_{i=1}^{n} x_i + 2 \cdot nm \stackrel{!}{=} 0 \\ \Leftrightarrow m^* &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \bar{x} \qquad \text{(Zweite Ableitung prüfen!)} \end{aligned}$$

Es ergibt sich als optimaler Standort: $\boxed{\bar{x} = 647.3}$

Folgerung: Das arithmetische Mittel minimiert $s^2(m) = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2$:

$$s^2 = s^2(\bar{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \min_{m \in \mathbb{R}} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - m)^2 = \min_{m \in \mathbb{R}} s^2(m)$$

(c) Minimierung der Summe der Wegstrecken aller Mitarbeiter:

$$\min_{m \in \mathbb{R}} \sum_{i=1}^{1000} |x_i - m| \quad \text{bzw.} \quad \min_{m \in \mathbb{R}} \sum_{i=1}^{n} |x_{(i)} - m|$$

$x_{(i)},\ i = 1, \ldots, n$ bezeichnet die geordnete Urliste der Entfernungen der Mitarbeiter von der Pforte.

Behauptung: $\boxed{m^{opt.} = x_Z \text{ (Median)}}$

Beweis:

1.) verbal:

Liegt die Kantine im Zentralwert, so kommt bei geradzahligem n die Hälfte der Mitarbeiter von der einen, die andere von der anderen Seite[5]. Eine Verschiebung der Kantine wirkt sich in der Gesamtwegstrecke nicht aus, solange die Zugangsrichtung für alle gleich bleibt, da die Wegstrecke, die die eine Hälfte weniger geht, von der anderen mehr geleistet werden muss. Ist die Verschiebung so groß, dass bei einem oder mehreren Mitarbeitern sich die Richtung zur Kantine ändert, wächst die Gesamtstrecke, da die Erhöhung bei der einen Hälfte nicht mehr vollständig kompensiert wird.

[5] sofern sie überhaupt einen Weg zur Kantine zurückzulegen haben

2.) formal: siehe Deskriptive Statistik, Seite 182ff

Es ergibt sich somit als optimaler Standort: $\boxed{x_z = 600}$

Folgerung: Der Median minimiert $d(m) = \frac{1}{n}\sum_{i=1}^{n} |x_i - m|$:

$$d = d(x_z) = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_z| = \min_{m \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |x_i - m| = \min_{m \in \mathbb{R}} d(m)$$

(d) Median bzw. arithmetisches Mittel können als Lageparameter Aufschluss über die Lage einer Verteilung geben. Mit Hilfe der mittleren absoluten Abweichung und der Varianz soll *unabhängig von der Lage* die Verteilung weiter beschrieben werden. Dazu werden die Abweichungen der Beobachtungen vom jeweiligen Lageparameter betrachtet. (b) bzw. (c) zeigen, dass die mittlere absolute Abweichung bzw. die Varianz, als Funktion der Abweichungen der Beobachtungen von einem Referenzwert m ihr Minimum für den Median ($m = x_z$) bzw. das arithmetische Mittel ($m = \bar{x}$) annehmen, und damit eine Zusammengehörigkeit von Median und mittlerer absoluter Abweichung bzw. von arithmetischem Mittel und Varianz (Standardabweichung) gegeben ist.

Aufgabe 14

Berechnen Sie zu den Gewichtsangaben der 50 Personen aus Übung 12 die α-Quantile für $\alpha = 20\%$ und $\alpha = 60\%$. Stellen Sie die Verteilung der Gewichtsangaben mit Hilfe eines Boxplots dar.

Lösung: (Deskriptive Statistik, S. 66, 87ff)

α-Quantil:

- Definition nicht einheitlich (vgl. Gessler, 1993, Seite 100 für verschiedene Definitionen des α-Quantils).

hier:

$$
\begin{aligned}
q_\alpha &= \begin{cases} \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}), & \text{für } n\alpha \in \mathbb{N}_0 \\ x_{\lceil n\cdot\alpha \rceil}, & \textit{sonst} \end{cases} \\
&= \begin{cases} \frac{1}{2}(\min\{x|F(x) = \alpha\} + \min\{x|F(x) > \alpha\}), & \text{falls } x \textit{ mit } F(x) = \alpha \textit{ existiert} \\ \min\{x|F(x) > \alpha\}, & \text{sonst} \end{cases}
\end{aligned}
$$

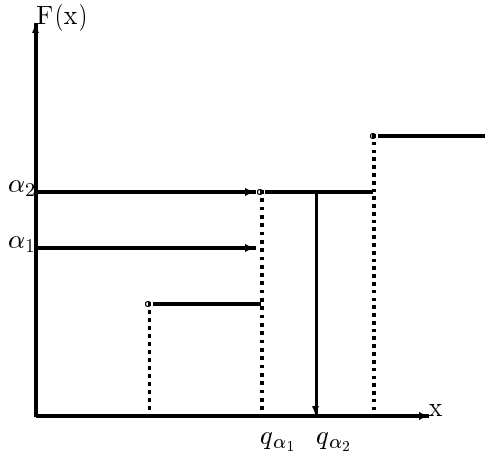


Abbildung 6: Veranschaulichung der Quantilsbestimmung anhand der empirischen Verteilungsfunktion

Spezielle Quantile:

- α=0,25; 0.75 : unteres/oberes Quartil
- α=0.5 : Median

Mit Hilfe des oberen und unteren Quartils wird als weiteres Streuungsmaß der Quartilsabstand berechnet (weniger anfällig gegen Ausreißer als R): QA=$q_{0,75} - q_{0,25}$ (vgl. auch Übung 12 (a))

$n = 50$:

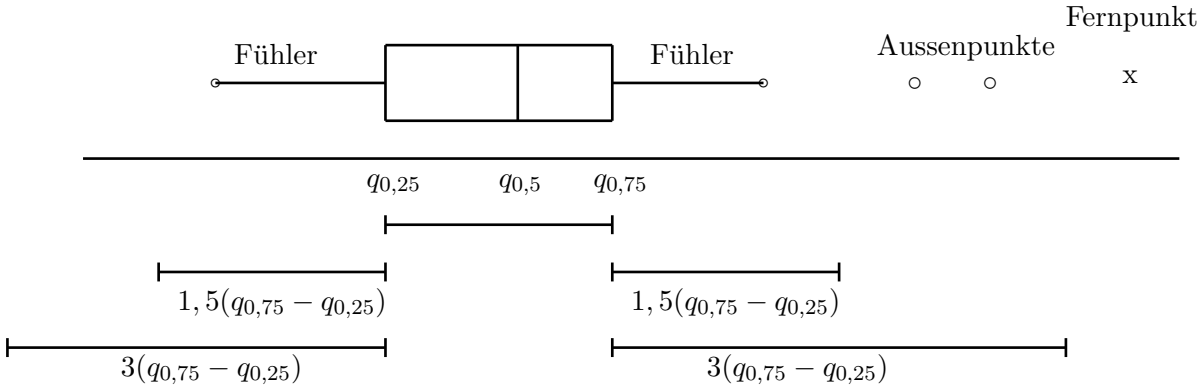
- Für $\alpha = 0,2$ ist $n\alpha = 50 \cdot 0,2 = 10$ und damit:

  $q_{0,2} = \frac{1}{2}(x_{(10)} + x_{(11)}) = 58.$

- Für $\alpha = 0,6$ ist $n\alpha = 50 \cdot 0,6 = 30$ und damit:

  $q_{0,6} = \frac{1}{2}(x_{(30)} + x_{(31)}) = 75.$

Boxplot: zur Darstellung einer Verteilung



hier:

$$\left.\begin{array}{l} q_{0,25} = x_{(13)} = 60 \\ q_{0,75} = x_{(38)} = 81 \end{array}\right\} \Rightarrow QA = 21$$

$x_z = q_{0,5} = 72$

1,5 QA = 31,5

$\Rightarrow$

unterer Anrainer: $x_{(1)} = 47$,
oberer Anrainer: $x_{(50)} = 98$,
es gibt keine Aussen- und Fernpunkte.

Aufgabe 15

An 21 Personen – anhand ihres Ernährungsverhaltens in drei Gruppen unterteilt – wurden folgende Messwerte des Cholesterolgehaltes im Blut (in [mg/100ml]) festgestellt:

Gruppe	Cholesterolwerte									
I	403	311	269	336	259					
II	312	222	302	420	540	386	353	210	286	290
III	403	244	353	235	319	260				

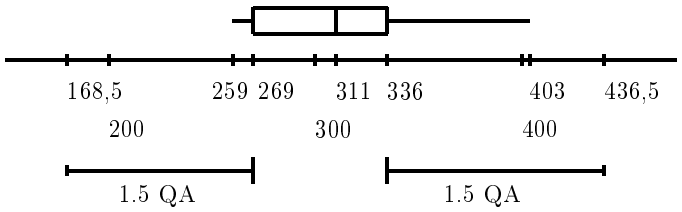
Vergleichen Sie die Cholesterol-Verteilungen der drei Gruppen mit Hilfe von Boxplots.

Lösung: (Deskriptive Statistik, S. 87ff)

I (n=5) : $x_z = x_{(3)} = 311$

$$\left.\begin{array}{l} q_{0,25} = x_{(2)} = 269 \\ q_{0,75} = x_{(4)} = 336 \end{array}\right\} \Rightarrow QA = 67 \Rightarrow QA \cdot 1.5 = 100,5$$

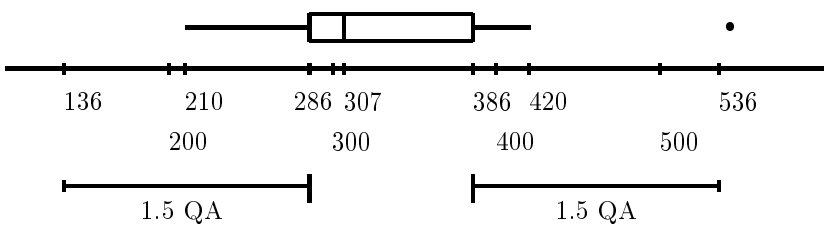
Boxplot:



II (n=10): $x_z = \frac{1}{2}(x_{(5)} + x_{(6)}) = 307$

$$\left.\begin{array}{l} q_{0,25} = x_{(3)} = 286 \\ q_{0,75} = x_{(8)} = 386 \end{array}\right\} \Rightarrow QA = 100 \Rightarrow QA \cdot 1.5 = 150$$

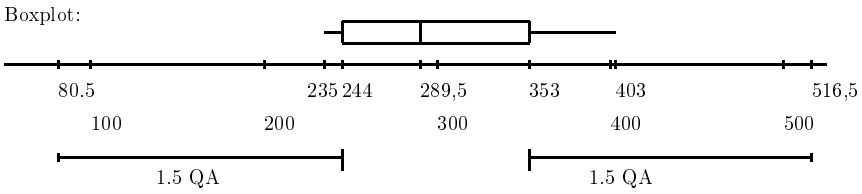
Boxplot:



Bemerkung: II hat einen Ausreißer (• ist ein Aussenpunkt).

III (n=6): $x_z = \frac{1}{2}(x_{(3)} + x_{(4)}) \;=\; 289,5$

$$\left.\begin{array}{rcl} q_{0,25} & = & x_{(2)} \;=\; 244 \\ q_{0,75} & = & x_{(5)} \;=\; 353 \end{array}\right\} \Rightarrow \; QA \;=\; 109 \;\Rightarrow\; QA \cdot 1.5 \;=\; 163.5$$



Aufgabe 16

Die Quartile für den Alkoholkonsum von 28 Männern lauten:

25%-Quantil:	5	g/Tag
Median	12	g/Tag
75%-Quantil:	12	g/Tag

Die angegebenen Quantile sind gleichzeitig auch gemessene Werte des zugrunde liegenden Datensatzes.

(a) Welche der Quantile $q_{0.15},\ q_{0.25},\ q_{0.40},\ q_{0.50},\ q_{0.70},\ q_{0.75},\ q_{0.90}$ sind für $n = 28$ notwendigerweise identisch mit tatsächlich gemessenen Größen, welche nur zufälligerweise im hier angegebenen Datensatz? Begründung!

(b) Schätzen Sie mit Hilfe der obigen Angaben ab, wie viele der 28 Männer 5 g bzw. 12 g pro Tag zu sich nehmen. Begründung!

g Alkohol/Tag	minimal mögliche Anzahl	maximal mögliche Anzahl
5		
12		

(c) Zeichnen Sie unter Verwendung der zusätzlichen Angabe, dass im Datensatz die Messwerte 0, 13, 17, 18, 19, 17, 28 vorkommen, einen Boxplot.

Lösung: (Deskriptive Statistik, S. 66, 87ff)

(a) Ein α-Quantil entspricht einer gemessenen Größe, wenn $\alpha \cdot n \notin \mathbb{N}$, wobei n die Anzahl der statistischen Einheiten bezeichnet. Hieraus ergibt sich, dass alle diejenigen α-Quantile zwingender Weise mit gemessenen Größen übereinstimmen, für die α kein ganzzahliges Vielfaches von 1/28 ist. Bei uns sind dies: $q_{0.15},\ q_{0.4},\ q_{0.7},\ q_{0.9}$

(b)

$$\left.\begin{array}{l} q_{0,25} = \frac{1}{2}(x_{(7)} + x_{(8)}) \\ \text{gemessener Wert} \end{array}\right\} \Rightarrow x_{(7)} = x_{(8)} = 5\frac{g}{Tag}$$

- Damit nehmen mindestens 2 Personen 5 $\frac{g}{Tag}$ zu sich.

$$\left.\begin{array}{l} q_{0,5} = \frac{1}{2}(x_{(14)} + x_{(15)}) \\ \text{gemessener Wert} \end{array}\right\} \Rightarrow\ x_{(14)} = x_{(15)} = 12\frac{g}{Tag} \neq 5\frac{g}{Tag}$$

- Damit nehmen höchstens $x_{(1)}, \ldots x_{(13)}$ den Wert $5\frac{g}{Tag}$ an.

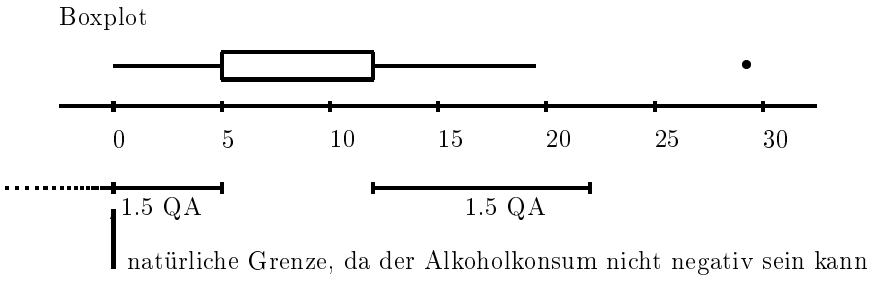
$$\left.\begin{array}{l} q_{0,5} = \frac{1}{2}(x_{(14)} + x_{(15)}) = q_{0,75} = \frac{1}{2}(x_{(21)} + x_{(22)}) \\ \text{gemessene Werte} \end{array}\right\} \Rightarrow x_{(14)} = x_{(15)} = x_{(21)} = x_{(22)} = 12\frac{g}{Tag}$$

- Damit nehmen mindestens $x_{(14)}, \ldots, x_{(22)}$, den Wert $12\frac{g}{Tag}$ an.
- $x_{(8)} = 5\frac{g}{Tag}$: Damit nehmen höchstens $x_{(9)}, \ldots, x_{(28)}$ den Wert $12\frac{g}{Tag}$ an.

$\frac{gAlkohol}{Tag}$	minimal mögliche Anzahl	maximal mögliche Anzahl
5	2	13
12	9	20

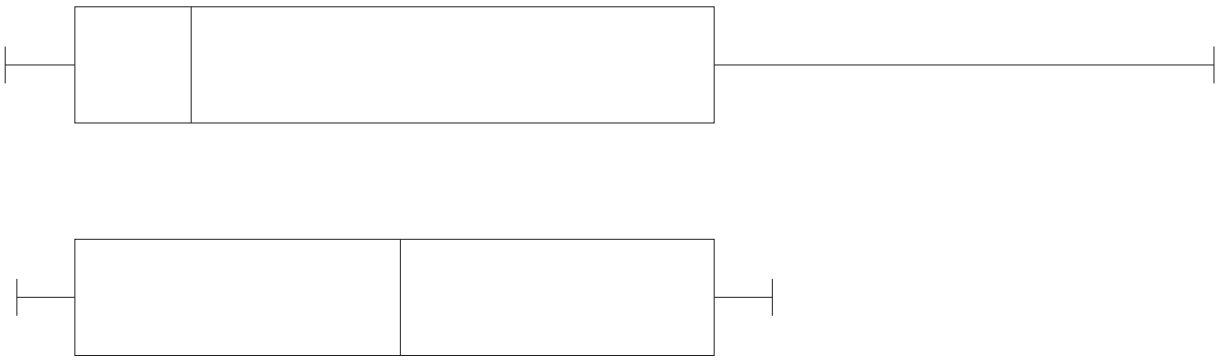
(c)

- $x_{(14)}, \ldots, x_{(22)} = 12\frac{g}{Tag}$, n=28 $\Rightarrow x_{(23)} = 13$, $x_{(24)} = x_{(25)} = 17$, $x_{(26)} = 18$, $x_{(27)} = 19$ und $x_{(28)} = 28$
- $x = 0$ kleinster möglicher Wert $\Rightarrow x_{(1)}$=0
- $q_{0,25} = 5$, $q_{0,5} = q_{0,75} = 12 \Rightarrow$ QA=7; 1,5QA=10,5
- $x_{(28)}$=28 ist Aussenpunkt.



Aufgabe 17

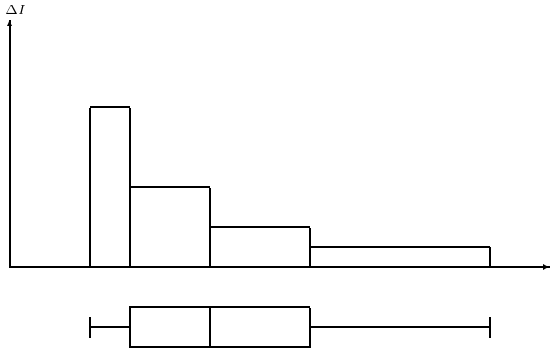
Gegeben sind die folgenden beiden Boxplots:



Was kann man über die zugehörigen Datenverteilungen sagen? Stellen Sie diese in groben Zügen durch Punktediagramme und Histogramme dar.

Lösung: (Deskriptive Statistik, S. 87ff)

Keine Ausreißer: jeweils mindestens 25% der Werte in $[x_{(1)}; q_{0,25}]$, $[q_{0,25};\ q_{0,5}]$, $[q_{0,5};\ q_{0,75}]$, $[q_{0,75}; x_{(n)}]$ Histogramm zu den vier Klassen $[x_{(1)}; q_{0,25})$, $[q_{0,25};\ q_{0,5})$, $[q_{0,5};\ q_{0,75})$, $[q_{0,75}; x_{(n)}]$ unter der Annahme, dass die Quartile selbst nicht beobachtet werden.



Analoges Vorgehen für Boxplot 2.

- Verteilung in den Klassen unbekannt, daher keine Aussage über die Anzahl der Gipfel möglich (analog keine Aussage über die Schiefe möglich).
- Verteilung sicher nicht symmetrisch (vgl. Anrainer und Median)

⇒ Boxplot enthält relativ wenig Informationen über die Verteilung

Aufgabe 18

In der Physik gibt das Potential eines Punktes x die Spannung zwischen x und einem vorab gewählten Referenzpunkt x_0 an.

(a) Um was für Merkmale (Skalentypen) handelt es sich bei der elektrischen Spannung bzw. dem elektrischen Potential.

(b) Sind die Variationskoeffizienten der Verteilungen der Merkmale Spannung und elektrisches Potential invariant gegenüber Skalentransformationen?

Lösung:

(a)

- Bei dem elektrischen Potential sind weder Einheit noch Nullpunkt fest vorgegeben, d.h. es handelt sich um eine Intervallskala.
- Bei der elektrischen Spannung ist der Nullpunkt natürlich vorgegeben, d.h. es handelt sich um eine Verhältnisskala.

(b)

- Der Variationskoeffizient ist nur invariant gegenüber Transformationen der Form $y = a \cdot x$. Dies entspricht gerade den Skalentransformationen einer Verhältnisskala.

Aufgabe 19

Betrachtet werden 5 Straßenhändler, die an einem bestimmten Tag auf der Kaiserstraße zwischen 16 und 17 Uhr Modeschmuck verkaufen.

(a) Nehmen Sie an, dass nur ein Händler in dieser Zeit überhaupt etwas verkauft. Wie sehen die Lorenzkurve und der Gini-Koeffizient aus?

(b) Nehmen Sie nun an, der Gesamtumsatz der Händler betrage 300 DM. Betrachten Sie die folgenden 5 Verteilungen:

Händler i / Verteilung	1	2	3	4	5
I	60	60	60	60	60
II	40	30	60	70	100
III	100	100	100	0	0
IV	200	25	25	25	25
V	10	10	260	10	10

Geben Sie zu jeder Verteilung die Lorenzkurve und den Gini-Koeffizienten an. Diskutieren Sie anhand der Ergebnisse die Aussagekraft des Gini-Koeffizienten.

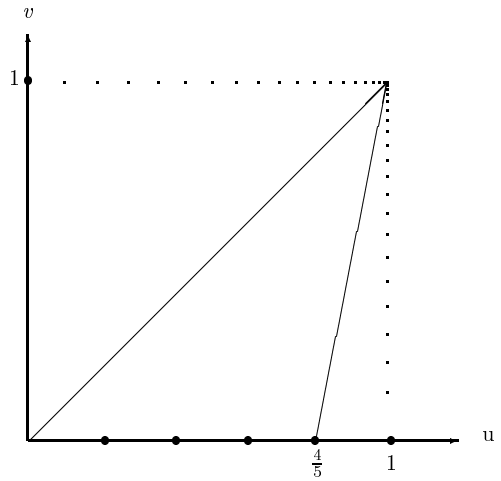
(c) Verdeutlichen Sie sich die Aussagekraft des Herfindahl-Indexes. Interpretieren Sie dazu den Anteil des Merkmals x_i an der Merkmalssumme, d.h. $q_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$, als neuen Merkmalswert, und zeigen Sie anschließend, dass $H = \sum_{i=1}^{n} q_i^2$ sich aus der Varianz der q_i berechnen lässt. Was ist der Minimal- und Maximalwert von H?

(d) Bestimmen Sie für jede der Verteilungen den Herfindahl-Index und den Konzentrationskoeffizienten CR_3.

Lösung: (Deskriptive Statistik, S. 99ff, 104f)

(a)

$$G = \tfrac{4}{5}\,(\ldots = G_{max}) \;=\; \sum_{i=0}^{n-1} (u_{i+1} - u_i)(u_i - v_i + u_{i+1} - v_{i+1})$$



(b) Die statistische Masse der schmuckverkaufenden Strassenhändler in diesem Beispiel umfasst fünf Elemente.

$\rightarrow u_0 = 0; u_i = \frac{i}{5}, i = 1, \ldots, 5$

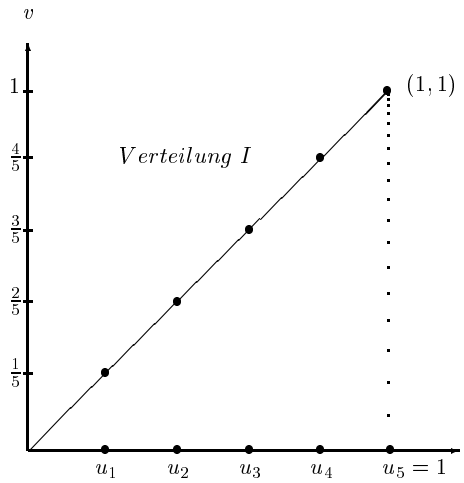
Die Gesamtsumme - hier der Gesamtumsatz - ergibt sich zu 300. Die fünf angenommenen Umsatzverteilungen ergeben folgende v_i:

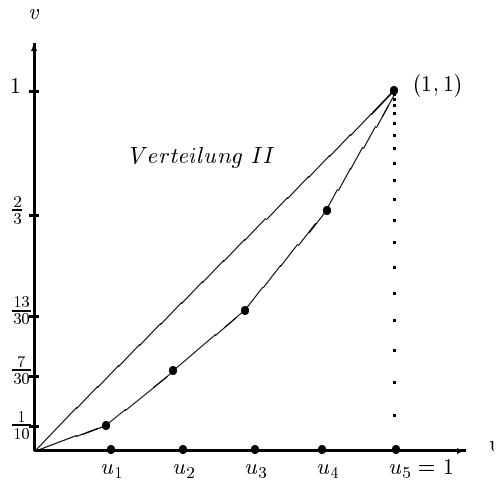
$$v_i = \frac{\sum_{k=1}^{i} x_k}{\sum_{k=1}^{n} x_k}$$

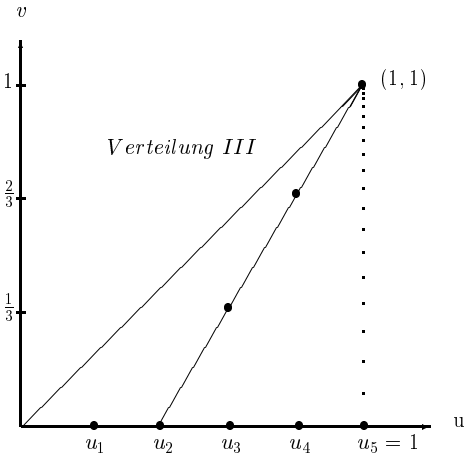
	v_0	v_1	v_2	v_3	v_4	v_5
Umsatzverteilung						
I	0	1/5	2/5	3/5	4/5	1
II	0	3/30	7/30	13/30	20/30	1
III	0	0	0	1/3	2/3	1
IV	0	1/12	2/12	3/12	4/12	1
V	0	1/30	2/30	3/30	4/30	1

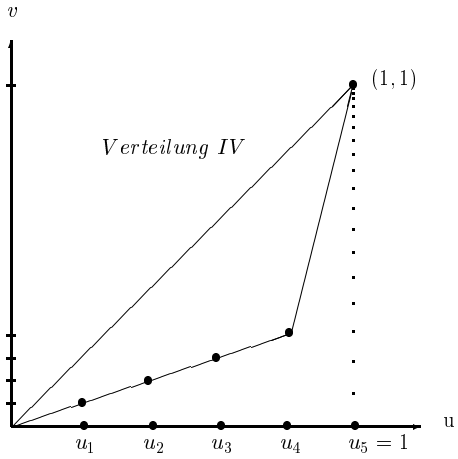
Achtung: Bei mehreren gleichen Merkmalsausprägungen reicht die Berechnung an den „Grenzen“ (sowohl für die Lorenzkurve als auch für den Gini-Koeffizienten).

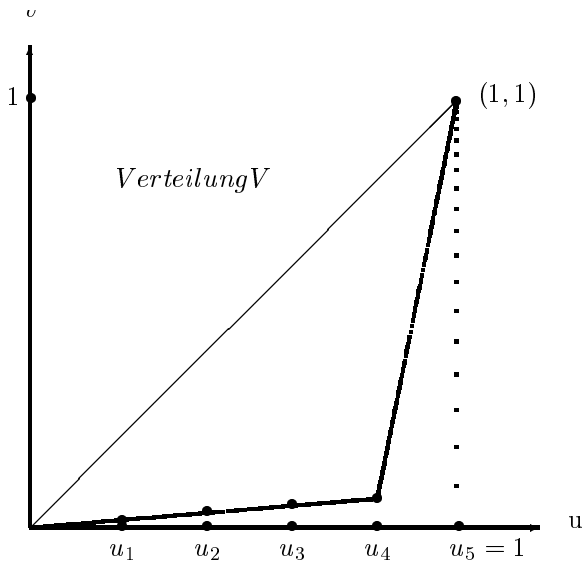
Lorenzkurven:


v
1
$\frac{4}{5}$
$\frac{3}{5}$
$\frac{2}{5}$
$\frac{1}{5}$
(1, 1)
Verteilung I
u
u_1
u_2
u_3
u_4
$u_5 = 1$


v
1
$\frac{2}{3}$
$\frac{13}{30}$
$\frac{7}{30}$
$\frac{1}{10}$
(1, 1)
Verteilung II
u
u_1
u_2
u_3
u_4
$u_5 = 1$


v
1
$\frac{2}{3}$
$\frac{1}{3}$
(1, 1)
Verteilung III
u
u_1
u_2
u_3
u_4
$u_5 = 1$


v
1
(1, 1)
Verteilung IV
u
u_1
u_2
u_3
u_4
$u_5 = 1$



Die Gini-Konzentrationsmaße G ergeben sich zu:

Umsatzverteilung	I	II	III	IV	V
	0	$17/75 \approx 0.23$	$2/5 = 0.4$	$7/15 = 0.47$	$2/3 = 0.67$

(c)

* Herfindahl-Index: $H := \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i} \right)^2 = \sum_{i=1}^{n} q_i^2$ mit $q_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$

* Formel für die Berechnung der Varianz: $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, mit $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

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

$\Rightarrow$ Varianz der q_i: Mit $\bar{q} = \frac{1}{n} \sum_{i=1}^{n} q_i = \frac{1}{n}$

$$\begin{aligned} s_q^2 &= \frac{1}{n} \sum_{i=1}^{n} (q_i - \bar{q})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} q_i^2 - \bar{q}^2 \\ &= \frac{1}{n} H - \bar{q}^2 \\ &= \frac{1}{n}(H - \frac{1}{n}) \\ H &= n s_q^2 + \frac{1}{n} \end{aligned}$$

Mit $s_q^2 \geq 0$ folgt direkt $H \geq \frac{1}{n}$. Da $\sum_{i=1}^{n} q_i = 1$ sowie $q_i \geq 0$ ($x_i \geq 0$ nach Voraussetzung) und damit auch $q_i \leq 1$ für alle $i = 1, \cdots, n$ gilt, nimmt H sein Maximum an, falls $q_i = 1$ für ein i gilt ($q_i < 1 \Rightarrow q_i^2 < q_i < 1$), d.h. größter Wert für H ist $H_{max} = 1$.

(d)

* Herfindahl-Index

$$\mathrm{H} = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i} \right)^2$$

$H_{min} = \frac{1}{n} \leq H \leq 1 = H_{max}$ (je grösser H, desto stärker die Konzentration)

* Konzentrationskoeffizient CR_g:

$$CR_g = \frac{\sum_{i=n-g+1}^{n} x_{(i)}}{\sum_{i=1}^{n} x_{(i)}} \quad \text{für } g = 1, 2, 3, \ldots$$

Gibt an, welchen Anteil der Merkmalssumme die g letzten Merkmalswerte der geordneten statistischen Reihe in sich vereinen.

Umsatzverteilung	I	II	III	IV	V
H	$1/5 = 0.2$	0.233	$1/3 = 0.\bar{3}$	0.472	0.756
CR_3	3/5	23/30	1	5/6	14/15

Bemerkung: Man vergleiche die Berechnung der Konzentrationskoeffizienten mit der Konstruktion des Lorenzkurve.

Aufgabe 20

In einem Unternehmen wurde im Jahr 2001 die folgende Verteilung der Monatsgehälter von 20 Mitarbeitern festgestellt:

Klasse	Monatsgehalt in DM von	bis	unter	Anzahl Mitarbeiter
1	0	-	2000	4
2	2000	-	4000	6
3	4000	-	6000	4
4	6000	-	8000	3
5	8000	-	10000	2
6	über		10000	1

(a) Zeichnen Sie die Lorenzkurve, und berechnen Sie für die klassierten Daten den Ginikoeffizienten der Gehaltskonzentration auf die Mitarbeiter.

(b) Wie ändern sich Lorenzkurve und Ginikoeffizient, wenn der Gesetzgeber die Einkommen folgendermaßen besteuert:

(i) Kopfsteuer von 500,– DM pro Gehaltsempfänger und Monat

(ii) Proportionalsteuer von 30% auf das monatliche Bruttogehalt

(iii) progressive Besteuerung mit den entsprechenden Steuersätzen der Einkommensklassen:

Klasse	1	2	3	4	5	6
Durchschnittssteuersatz in %	6	13	20	27	34	41

(c) Welche Umverteilungseffekte bewirken demnach die unterschiedlichen Besteuerungsarten im Hinblick auf die Verteilung des gesamten Volkseinkommens?

Bemerkung: Der Duchschnittssteuersatz für die Klasse $i, i = 1, \ldots, 6$ gibt denjenigen Steuersatz an, mit dem ein Mitarbeiter der entsprechenden Lohnklasse sein *gesamtes* Einkommen versteuern muss.

- **Lorenz-Kurve aus klassierten Daten**

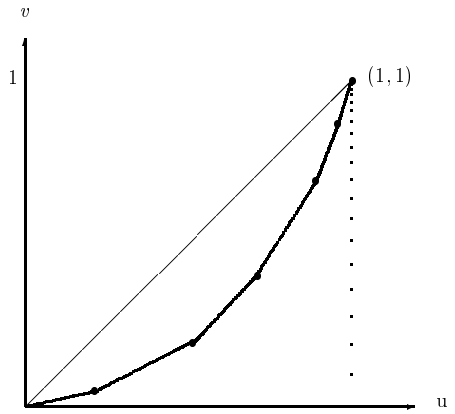
(a)

k	0	1	2	3	4	5	6
$h(I_k)$	-	4	6	4	3	2	1
u_k	0	4/20	1/2	7/10	17/20	19/20	1
z_{I_k}	-	1000	3000	5000	7000	9000	11000 [1)]
v_k	0	4/92	22/92	42/92	63/92	81/92	1

1) offene Randklasse: hier wird die Annahme getroffen, dass das höchste Gehalt in [10000; 12000) liegt

$x^* = \sum h(I_k) z_{I_k} = 92000.$

Lorenzkurve (zu a):



(b) analog

k	0	1	2	3	4	5	6	
I z_{I_k}		500	2500	4500	6500	8500	10500	$x^* = 82000$
v_k	0	2/82	17/82	35/82	545/820	715/820	1	
II z_{I_k}		700	2100	3500	4900	6300	7700	$x^* = 64400$
v_k	0	28/644	154/644	294/644	441/644	567/644	1	
III z_{I_k}		940	2610	4000	5110	5940	6490	$x^* = 69120$
v_k	0	376/6912	1942/6912	3520/6912	5075/6912	6263/6912	1	

Ginikoeffizienten : Aufgabe a)

G= 0.346

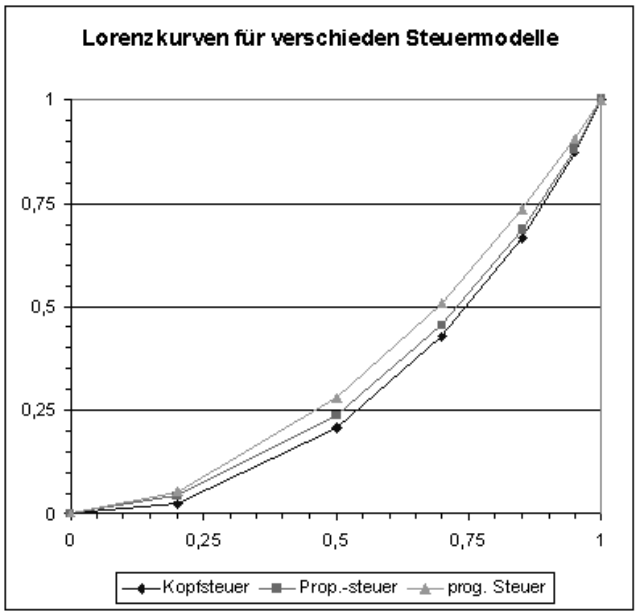
Aufgabe b)

I G= 0.388

II G= 0.346

III G= 0.285

Proportionalsteuer hat keinen Einfluss auf G; Kopfsteuer vergrössert die Konzentration, progressive Steuer vermindert die Einkommensdivergenz eher(gemessen jeweils durch den Gini-Koeffizient).


Lorenzkurven für verschieden Steuermodelle
1
0,75
0,5
0,25
0
0
0,25
0,5
0,75
1
Kopfsteuer
Prop.-steuer
prog. Steuer

Aufgabe 21

Aus dem statistischen Jahrbuch entnimmt man für 1975 die folgenden Zahlen für die Verteilung des Umsatzes auf die Umsatzgrößenklassen der Betriebe der verarbeitenden Industrie:

Klasse	Umsatz von ... bis unter ... DM	Anzahl der Unternehmen	Umsatz in Mill. DM	Beschäftigte
1	unter 1 Mill. DM	7.047	4.577	119.746
2	1 Mill. - 2 Mill.	8 167	11 887	209 079
3	2 Mill. - 5 Mill.	10 594	34 163	485 871
4	5 Mill. - 10 Mill.	6 237	44 102	552 409
5	10 Mill. - 25 Mill.	5 238	81 856	928 103
6	25 Mill. - 50 Mill.	2 167	82 669	776 854
7	50 Mill. - 100 Mill.	1 194	82 669	813 254
8	100 Mill. und mehr	948	484 975	3.578.868

(a) Geben Sie die Lorenzkurve und den Gini-Koeffizienten für die Umsatzkonzentration auf die Unternehmensanzahl an. Wie müßten Sie vorgehen, wenn der Umsatz der einzelnen Klassen nicht bekannt wäre?

(b) Wie würden sich die Lorenzkurve und der Gini-Koeffizient ändern, wenn die Klassen paarweise (Klasse 1 und 2, Klasse 3 und 4 usw.) zusammengefasst würden?

Lösung:

Lorenzkurve aus klassierten Daten:

I_j seien die Klassen ($j = 1, \ldots, r$) (sortiert nach Klassengrenzen) $x(I_j)$ Merkmalssumme in Klasse I_j; n_j die Anzahl der statistischen Einheiten in I_j und p_j die relative Häufigkeit von Klasse I_j. Da $x(I_j)$ in der Regel nicht bekannt ist, dient $n_j \cdot \tilde{x}_j$ als Approximation für die Merkmalssumme in Klasse j. Der kumulierte Anteil an statistischer Masse berechnet sich gemäß:

$$u_0 = 0$$

$$u_k = \frac{\sum_{j=1}^{k} n_j}{n} = \sum_{j=1}^{k} p_j, \quad \text{für } k = 1, \ldots, r$$

Der kumulierte Anteil an der Merkmalssumme berechnet sich gemäß:

$$v_0 = 0$$

$$v_k = \frac{\sum_{j=1}^{k} x(I_j)}{\sum_{j=1}^{r} x(I_j)} = \frac{\sum_{j=1}^{k} n_j \cdot x_j}{\sum_{j=1}^{r} n_j \cdot x_j} = \frac{\tilde{X}_k}{\tilde{X}} \quad \text{für } k = 1, \ldots, r$$

(a) Umsatzkonzentration auf Unternehmenszahl:

Die Merkmalssummen innerhalb der Klassen $x(I_j)$ sind hier bekannt, müssen also nicht unter Zuhilfenahme der Klassenmitten approximiert werden.

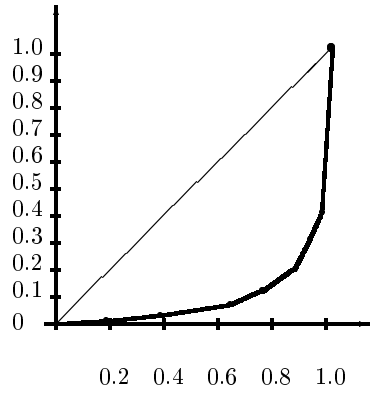
$$\Rightarrow \quad u_1 = \frac{n_1}{n} = \frac{7.047}{41.592} \approx 0.170$$
$$u_2 = \frac{n_1+n_2}{n} = \frac{7.047+8.167}{41.592} \approx 0.366$$
$$u_3 \approx 0.621 \qquad u_6 \approx 0.949$$
$$u_4 \approx 0.771 \qquad u_7 \approx 0.977$$
$$u_5 \approx 0.896 \qquad u_8 \approx 1.000$$

Gesamtumsatz: $X = \sum_{j=1}^{r} x(I_j) = 826.898$

$\Rightarrow$

$$v_1 = \frac{4.577}{826.898} \approx 0.006$$
$$v_2 = \frac{4577+11887}{826.898} \approx 0.02$$
$$v_3 \approx 0.06 \qquad v_5 \approx 0.215 \qquad v_6 \approx 0.314$$
$$v_4 \approx 0.116 \qquad v_7 \approx 0.414 \qquad v_8 \approx 1.0$$

Lorenzkurve der Umsatzkonzentration auf die Unternehmenszahl



Gini-Koeffizient:

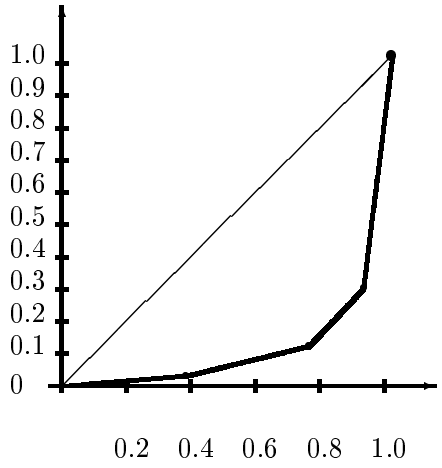
$$\begin{aligned} G &= \sum_{j=0}^{r-1} (u_{j+1} - u_j) \cdot (u_j - v_j + u_{j+1} - v_{j+1}) \\ &= 0.170(0.170 - 0.006) + \ldots + (1 - 0.977) \cdot (0.977 - 0.408 + 1 - 1) \\ &= 0.827 \end{aligned}$$

Sind die Umsätze $x(I_j)$ nicht bekannt, müssen diese wie erwähnt mittels der Klassenmitten $\bar{x_j}$ geschätzt werden. Für die 1. Klasse würde sich dann ergeben:

$$x(I_j) = n_1 \cdot \bar{x_1} = 7.047 \cdot 500.000 = 3523,5 \text{ Mio.}$$

(b) Sind zwei verschiedene Klasseneinteilungen hierarchisch angeordnet, d.h. ist die eine eine Verfeinerung der anderen (wie im vorliegenden Fall) so interpoliert die Lorenzkurve der gröberen Klasseneinteilung diejenige der feineren an den Stellen der übereinstimmenden kumulierten Anteile an statistischen Massen. In unserem Fall gilt nämlich:

$$
\begin{aligned}
\tilde{u_1} :&= \tfrac{\tilde{h}_1}{u} &&= \tfrac{h_1+h_2}{u} &&= u_2\\
\tilde{u_2} :&= \tfrac{\tilde{h}_1+\tilde{h}_2}{u} &&= \tfrac{h_1+h_2+h_3+h_4}{u} &&= u_4\\
\tilde{u_3} :&= \tfrac{\tilde{h}_1+\tilde{h}_2+\tilde{h_3}}{u} &&= \tfrac{h_1+\ldots+h_6}{u} &&= u_6\\
\tilde{u_4} :&= 1
\end{aligned}
$$



Die Lorenzkurve hängt weniger durch und folglich wird der Gini-Koeffizient kleiner.

$$
\begin{aligned}
G &= \sum_{j=0}^{\tilde{r}-1}(\tilde{u}_{j+1}-\tilde{u}_j)\cdot(\tilde{u}_j-\tilde{v}_j+\tilde{u}_{j+1}-\tilde{v}_{j+1})\\
&= 0.366\cdot(0-0+0.346)+0.405\cdot(0.346+0.655)+\\
&\quad +0.178\cdot(0.655+0.641)+0.51\cdot(0.641+0)\\
&= 0.80
\end{aligned}
$$

Aufgabe 22

In einem 24-Stunden-Dauertest wurde die Leuchtdauer von 24 Glühbirnen getestet. Hierbei ergaben sich folgende Messwerte (in [h]):

1.0	1.2	1.4	1.7	2.2	2.5	3.1	3.5	4.4
5.1	6.1	7.5	8.9	10.7	12.6	15.4	18.3	22.1

6 Birnen brannten am Ende des 24 Stunden-Tests noch immer.

Erläutern Sie, inwiefern bei der Erstellung eines Boxplots bzw. eines Histogramms auf Basis der gegebenen Messwerte Probleme entstehen. Welche Lösungsansätze fallen Ihnen ein?

Lösung:

Problem: 6 Glühbirnen haben eine unbekannte Lebensdauer (> 24 Stunden)

Lösungsmöglichkeiten:

1) alle 6 Werte auf 24 Stunden setzen (da Werte > 24 Stunden beim Test nicht auftreten können)

2) alle 6 Lebensdauern auf einen Wert x > 24 Stunden setzen („im Durchschnitt werden die Glühbirnen wohl x Stunden gebrannt haben“)

3) maximal mögliche Lebensdauer festlegen (willkürlich, z.B. 29 Stunden) und Leuchtdauer äquidistant in [24;29] annehmen $\rightarrow\ x_{(19)} = 24, \ldots, x_{(24)} = 29$.

4) geordnete Urliste so fortsetzen, dass die fehlenden Werte $x_{(19)}, \ldots, x_{(24)}$ von $x_{(18)} = 22,1$ Stunden den gleichen Abstand wie $x_{(17)}, x_{(16)}, \ldots$, haben, d.h.

   $x_{(18+i)} = 22,1 \text{ Std} + (22,1 \text{ Std } - x_{(18-i)}), i = 1, \ldots, 6$

Bemerkung: Jede der vier Möglichkeiten ist willkürlich und damit problematisch.

Aufgabe 23

Eine Befragung von 20 Personen, welcher Partei sie bei der nächsten Bundestagswahl ihre Stimme geben würden, ergab folgendes Ergebnis:

Person	Geschlecht	Partei
1	m	SPD
2	m	CDU
3	w	Grüne
4	m	SPD
5	w	CDU
6	w	CDU
7	w	FDP
8	m	SPD
9	m	CDU
10	w	SPD

Person	Geschlecht	Partei
11	m	Grüne
12	m	CDU
13	m	SPD
14	w	CDU
15	w	CDU
16	m	Grüne
17	m	SPD
18	w	CDU
19	w	SPD
20	m	FDP

(a) Erstellen Sie eine Kontingenztabelle der absoluten Häufigkeiten, und bestimmen Sie die Randhäufigkeiten.

(b) Stellen Sie die Kontingenztabelle der relativen Häufigkeiten auf und geben Sie die Randverteilungen an.

(c) Bestimmen Sie die bedingten Häufigkeitsverteilungen des Wahlverhaltens für Frauen und Männer.

(d) Stellen Sie die Daten aus (b) geeignet graphisch dar.

Lösung:

Mehrdimensionale Merkmale:

Messung mehrerer (k) Merkmale an einer statistischen Einheit $b_j : S \to M_j, \quad j = 1, \ldots, k$. Die Häufigkeitsverteilung gibt die Häufigkeit jeder Merkmalskombination an. Im weiteren werden zwei Merkmale gemessen. Wir bezeichnen die möglichen Ausprägungen bei Merkmal 1 resp. 2 mit $a_1, \ldots, a_r$ resp. $b_1, \ldots, b_s$. Insgesamt sind somit $r * s$ Kombinationen von Merkmalsausprägungen möglich, die zusammen mit ihren Häufigkeiten in einer 2-dimensionale Kontingenztabelle (bzw. Korrelationstabelle, falls beide Merkmale mindestens ordinales Skalenniveau aufweisen) dargestellt werden können.

(a)

Partei / Geschlecht	SPD	CDU	FDP	Grüne	
m	5	3	1	2	11
w	2	5	1	1	9
	7	8	2	3	20

(b)

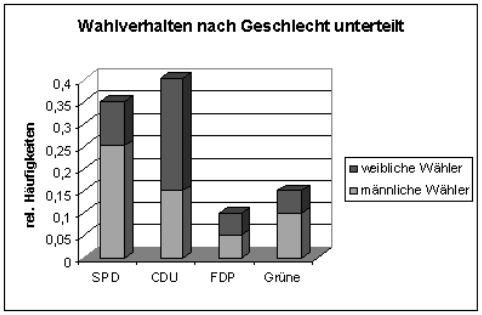
Partei / Geschlecht	SPD	CDU	FDP	Grüne	
m	$\frac{1}{4}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{11}{20}$
w	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{9}{20}$
	$\frac{7}{20}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{20}$	1

(c)

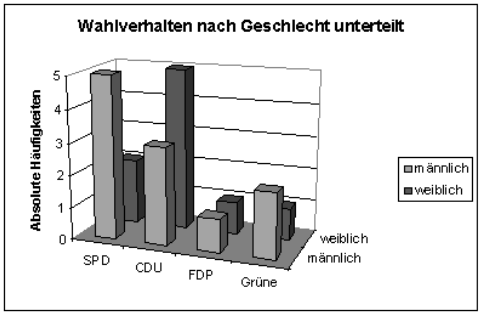
Bedingte Verteilungen (allg. $p(a|b) = \frac{p(a,b)}{p(b)}$)

Partei	SPD	CDU	FDP	Grüne	
$p(\text{Partei}\|m)$	$\frac{\frac{1}{4}}{\frac{11}{20}} = \frac{5}{11}$	$\frac{3}{11}$	$\frac{1}{11}$	$\frac{2}{11}$	1
$p(\text{Partei}\|w)$	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	1
$p(\text{Partei})$	$\frac{7}{20}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{20}$	1

(d) Eine Darstellung wäre z.B. durch ein 3-dimensionales Stab-/ Balkendiagramm möglich. Eine 2-dimensionale Darstellung ist möglich, wenn die absolute/ relative Häufigkeit durch Grautöne, Farben oder Schraffur dargestellt werden. Bei klassierten Merkmalen und $k = 2$ bietet sich ein 3-dimensionales Histogramm (volumenproportional) an.



(a) Relative Häufigkeiten



(b) Absolute Häufigkeiten

Aufgabe 24

Der Karlsruher Wiwi-Student John Highway fährt freitags grundsätzlich mit dem Auto von Karlsruhe in seine hessische Heimat, wo er das (verlängerte) Wochenende verbringt. Je nach Motivation fährt er dann montags, dienstags oder mittwochs nach Karlsruhe zurück, um sich dem Studienstress zu unterziehen. Da er häufig in den Stau kommt, hat er sich bei seinen letzten 60 Fahrten nach Karlsruhe notiert, an welchem Wochentag er gefahren ist, und ob er in einen Stau gekommen ist oder nicht. Hier sind seine Ergebnisse:

- Ein Drittel der Fahrten wurde montags gemacht, ein weiteres Drittel mittwochs.
- Insgesamt gab es genauso viele Fahrten mit Stau wie ohne Stau.
- Montags war in 60% der Fälle Stau.
- Mittwochs war sechsmal kein Stau.

(a) Erstellen Sie die Kontingenztabellen der absoluten und relativen Häufigkeiten.

(b) Berechnen Sie die bedingten Häufigkeiten für das (Nicht-)Auftreten eines Staus an den drei genannten Wochentagen.

(c) Was meinen Sie: Hat der Wochentag, an dem die Fahrt unternommen wird, einen Einfluss auf das Stauaufkommen?

Lösung: (Deskriptive Statistik, S. 117ff)

(a) Angegebene Werte:

	Mo	Di	Mi	Summe
Stau				30
kein Stau			6	30
Summe	20		20	60

Durch berechnen der fehlenden Werte erhält man die Kontingenztabelle der absoluten Häufigkeiten:

	Mo	Di	Mi	Summe
Stau	12	4	14	30
kein Stau	8	16	6	30
Summe	20	20	20	60

dabei berechnet sich z.B. $h(\text{Stau, Mo})$ gemäß:

$$0.6 \;=\; p(\text{Stau} \mid \text{Mo}) \;=\; \frac{h(\text{Stau,Mo})}{h(\text{Mo})} \;=\; \frac{h(\text{Stau,Mo})}{20} \text{ also } h(\text{Stau,Mo}) = 12.$$

Weiterhin die Kontingenztabelle für die relativen Häufigkeiten:

	Mo	Di	Mi	Summe
Stau	$\frac{1}{5}$	$\frac{1}{15}$	$\frac{7}{30}$	$\frac{1}{2}$
kein Stau	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{1}{10}$	$\frac{1}{2}$
Summe	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

(b) Die bedingte Häufigkeiten lauten:

	Mo	Di	Mi
p (Stau| Wochentag)	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{7}{10}$
p (kein Stau| Wochentag)	$\frac{2}{5}$	$\frac{4}{5}$	$\frac{3}{10}$

(c) Die Merkmale sind abhängig, da ansonsten die bedingten Häufigkeiten in (b) zeilenweise gleich groß wären. Bei Unabhängigkeit müsste bei den selben Randverteilungen gelten:

	Mo	Di	Mi	Summe
p (Stau, Wochentag)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
p (kein Stau, Wochentag)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
Summe	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

bzw.

	Mo	Di	Mi	Summe
h (Stau, Wochentag)	10	10	10	30
h (kein Stau, Wochentag)	10	10	10	30
Summe	20	20	20	60

Aufgabe 25

Bei einer Untersuchung von Familienstand und Geschlecht bei 50 Personen hat sich herausgestellt, dass die Merkmale unabhängig sind. Für die Randverteilungen gilt:

$$\begin{array}{lll} h(\text{weiblich}) = 30, & h(\text{männlich}) = 20, & h(\text{ledig}) = 5, \\ h(\text{verheiratet}) = 30, & h(\text{geschieden}) = 10, & h(\text{verwitwet}) = 5. \end{array}$$

Wie lautet die gemeinsame Häufigkeitsverteilung?

Lösung: (Deskriptive Statistik, S. 120f)

Definition: Zwei Merkmale a und b heißen unabhängig, wenn eine der folgenden drei äquivalenten Bedingungen erfüllt ist:

(i) für alle $a \in M_1$, für alle $b \in M_2$ gilt: $p(a,b) = p(a) \cdot p(b)$ bzw. $h(a,b) = \frac{h(a) \cdot h(b)}{n}$

(ii) für alle $a,\ a' \in M_1$ mit $p(a),\ p(a') \neq 0$ gilt: $p(b|a) = p(b|a')$ für alle $b \in M_2$.

(iii) für alle $b,\ b' \in M_2$ mit $p(b),\ p(b') \neq 0$ gilt: $p(a|b) = p(a|b')$ für alle $a \in M_1$.

Damit ergibt sich aus den Randverteilungen die gemeinsame Häufigkeitsverteilung:

	verheiratet	geschieden	ledig	verwitwet	
w	$= \frac{30 \cdot 30}{50} = 18$	6	3	3	30
m	12	4	2	2	20
	30	10	5	5	50

Aufgabe 26

Gegeben seien die Daten aus Aufgabe 23.

(a) Prüfen Sie

- mittels der Definition der Unabhängigkeit,
- mit Hilfe der bedingten Verteilungen und
- mittels des (korrigierten) Kontingenzkoeffizienten,

ob das Merkmal Geschlecht einen Einfluss auf die Parteipräferenz hat.

(b) Können Sie die Unabhängigkeit auch mit Hilfe eines gestapelten Balkendiagramms überprüfen?

(c) Erstellen Sie ein kombiniertes Flächen-/Kreissektorendiagramm. Eignet sich dieses Diagramm ebenfalls zur Überprüfung der Unabhängigkeit?

Lösung: (Deskriptive Statistik, S. 120f, 125ff)

(a)

- mit Hilfe der Definition:

  Bei Unabhängigkeit müsste gelten: $p(a,b) = p(a) \cdot p(b)$ bzw. $h(a,b) = \frac{h(a)h(b)}{n}$ für alle $a,\ b$. Es gilt aber z.B. $p\,(\text{m,SPD}) = \frac{1}{4} \neq \frac{7}{20} \cdot \frac{11}{20} = \frac{77}{400} = p(\text{m}) \cdot p\,(\text{SPD})$, folglich sind die Merkmale nicht unabhängig.

- mit Hilfe bedingter Verteilungen:

  Die nach dem Geschlecht bedingten Häufigkeitsverteilungen der Parteipräferenz müssten bei Unabhängigkeit übereinstimmen. Die bedingten Verteilungen lauten:

Partei	SPD	CDU	FDP	Grüne	
p(Partei| m)	$\frac{\frac{1}{4}}{\frac{11}{20}} = \frac{5}{11}$	$\frac{3}{11}$	$\frac{1}{11}$	$\frac{2}{11}$	1
p(Partei| w)	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	1
p(Partei)	$\frac{7}{20}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{20}$	1

  Die bedingten Verteilungen stimmen nicht überein, folglich sind die Merkmale nicht unabhängig.

  Bemerkung: analog kann p (Geschlecht|Partei) verglichen werden.

- über den (korrigierten) Kontingenzkoeffizienten: Der Kontingenzkoeffizient von Pearson berechnet sich gemäß:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} \text{ mit } \chi^2 = \sum_{\substack{a\in M_1\\ h(a)\neq 0}} \sum_{\substack{b\in M_2\\ h(b)\neq 0}} \frac{\left(h(a,b) - \frac{h(a)\cdot h(b)}{n}\right)^2}{\frac{h(a)\cdot h(b)}{n}}$$

$$= n \sum_{\substack{a\in M_1\\ p(a)\neq 0}} \sum_{\substack{b\in M_2\\ p(b)\neq 0}} \frac{\left(p(a,b) - p(a)\cdot p(b)\right)^2}{p(a)\cdot p(b)}$$

Der Kontingenzkoeffizient C ist ein Maß für die Abhängigkeit der betrachtenen Mermale. Es gilt:

$$0 \leq C \leq \sqrt{\frac{k-1}{k}} \quad \text{mit } k = \min\{r, s\},$$

wobei r bzw. s die Anzahl der Merkmalsausprägungen bei Merkmal 1 bzw. 2 bezeichnet. Um den Einfluss, der Anzahl an Merkmalsausprägungen zu eliminieren, wird analog zum Gini-Koeffizienten der korrigierte Kontingenzkoeffizient verwendet:

$$C_{corr} = \sqrt{\frac{k}{k-1}} \cdot C \quad \text{mit } 0 \leq C_{corr} \leq 1$$

Es gilt: $C_{corr} = 0 \iff$ Merkmale unabhängig. Es ergibt sich:

	SPD	CDU	FDP	Grüne	p (Geschlecht)
m	0.25 0.0033 0.0575 0.1925	0.15	0.05	0.1	0.55
w	0.1	0.25	0.05	0.05	0.45
p(Partei)	0.35	0.4	0.1	0.15	1

Man verwendet zur Berechnung folgendes Schema (vgl. Deskriptive Statistik, S.127):

$p(a,b)$	$\left(p(a,b) - p(a)p(b)\right)^2$
$p(a,b) - p(a)\cdot p(b)$	$p(a)\cdot p(b)$

Man erhält:

$$\begin{aligned}\chi^2 &= 20 \cdot \left(\frac{(0.25 - 0.1925)^2}{0.1925} + \frac{(0.15 - 0.55 \cdot 0.4)^2}{0.55 \cdot 0.4} + \ldots + \frac{(0.05 - 0.45 \cdot 0.15)^2}{0.45 \cdot 0.15} \right) \\ &= 1.938\end{aligned}$$

$$C = \sqrt{\frac{1.938}{20 + 1.938}} = 0.2972$$

$$C_{corr} = \sqrt{\frac{k}{k-1}} \cdot C = \sqrt{\frac{2}{1}}\, 0.2972 = 0.4203 \neq 0 \Longrightarrow \text{ nicht unabhängig.}$$

(b) Mit gestapeltem Balkendiagramm: Jeder Balken gibt die Häufigkeitsverteilung aller Kombinationen (a_i, b_j) mit festem $b_j = b^*$ an, d.h. bei Unabhängigkeit ist die Struktur der Balken (Grössenverhältnisse der Balkenabschnitte) gleich; Unabhängigkeit lässt sich z.B. über Strahlensatz überprüfen.

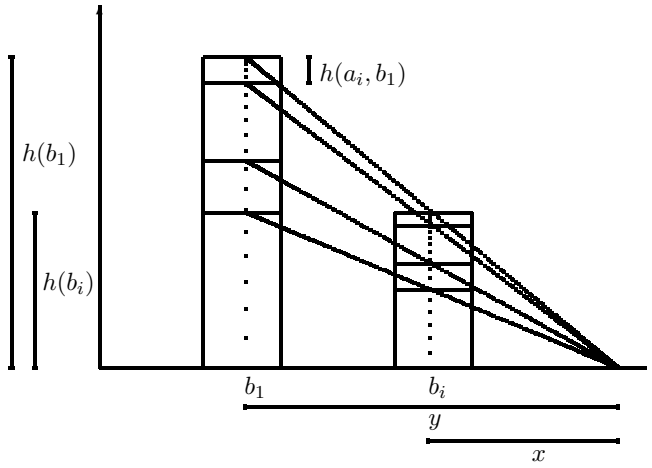


Abbildung 7: Überprüfen der Unabhängigkeit über den Strahlensatz (Balken geben die Randverteilung für b an.)

Bei Unabhängigkeit gilt:

$$\underbrace{\frac{h(a_i, b_1)}{h(b_1)}}_{p(a_i|b_1)} = \underbrace{\frac{h(a_i, b_2)}{h(b_2)}}_{p(a_i|b_2)}$$

(c) Flächendiagramm, d.h. Fläche proportional zur dargestellten statistischen Masse. In unserem Fall heißt dies, dass die Fläche proportional zur Anzahl der befragten Männer bzw. Frauen sein soll , d.h. $\sqrt{\text{Anzahl Personen}} \sim r$, r bezeichnet den Kreisradius.

Seien hierzu A_j die Fläche des Kreises für Merkmalsausprägung b_j, d.h. $A_j = \pi {r_j}^2 = c \cdot h(b_j)$, c beliebig gewählte Konstante. A^s_{ij} die Fläche des Kreissektors, der $h(a_i, b_j)$ darstellt. $A^s_{ij} = A_j \frac{\alpha_{ij}}{360^\circ} = A_j \cdot \frac{h(a_i,b_j)}{h(b_j)}$. Aus $\frac{h(a_i,b_j)}{h(b_j)} = p(a_i|b_j)$, folgt $\frac{\alpha_{ij}}{360^\circ} = p(a_i|b_j)$, und somit $\alpha_{ij} = 360^\circ \cdot p(a_i|b_j)$.

Bei Unabhängigkeit sind die Winkel α_{ij} der Kreissektoren zu $h(a_i, b_j)$ für alle $b_j \in M_2$ gleich, d.h. Unabhängigkeit lässt sich z.B. durch „Übereinandersschieben“ der Kreise prüfen. Man erhält:

$b_j = m : A_j = 11$ und $r_m = \sqrt{11} \cdot c$
$b_j = w : A_j = 9$ und $r_b = \sqrt{9} \cdot c$
$a_i =$SPD, damit ist $\alpha_{ij} = 360° p(\text{SPD}|\text{m}) = 360° \cdot \frac{5}{11} = 163,64°$

Analog lassen sich die Winkel für die übrigen Parteien bestimmen. Man erhält für $b_j = m$:

a_i	SPD	CDU	FDP	Grüne
α_{ij}	163,64°	98,18°	32,72°	65,45°

und für $b_j = w$:

a_i	SPD	CDU	FDP	Grüne
α_{ij}	80°	200°	40°	40°

Resultat: Die Winkel stimmen nicht überein folglich sind die Merkmale nicht unabhängig.

Aufgabe 27

Bestimmen Sie für die beiden folgenden Häufigkeitstabellen den Kontingenzkoeffizienten.

b \ a	a_1	a_2	a_3	a_4
b_1	20	12	8	4
b_2	10	6	4	2
b_3	5	3	2	1

b \ a	a_1	a_2	a_3
b_1	0	0	20
b_2	10	0	0
b_3	0	15	0

Lösung: (Deskriptive Statistik, S. 120f, 125ff)

Für den korrigierten Kontingenzkoeffizienten gilt: $0 \leq C_{corr} \leq 1,\ C_{corr} = 0 \Longleftrightarrow$ Merkmale sind unabhängig.

Häufigkeitstabelle 1:

$h(a,b)$

b \ a	a_1	a_2	a_3	a_4	
b_1	20	12	8	4	44
b_2	10	6	4	2	22
b_3	5	3	2	1	11
	35	21	14	7	77

$p(a|b)$

b \ a	a_1	a_2	a_3	a_4	
b_1	$\frac{20}{44}$	$\frac{12}{44}$	$\frac{8}{44}$	$\frac{4}{44}$	1
b_2	$\frac{10}{22}$	$\frac{6}{22}$	$\frac{4}{22}$	$\frac{2}{22}$	1
b_3	$\frac{5}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	1

nach Kürzen ergibt sich die bedingte Häufigkeitsverteilung:

b \ a	a_1	a_2	a_3	a_4	
b_1	$\frac{5}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	1
b_2	$\frac{5}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	1
b_3	$\frac{5}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	1

Die Merkmale sind unabhängig, also $C_{corr} = 0$.

Häufigkeitstabelle 2:

$h(a|b)$

b \ a	a_1	a_2	a_3	
b_1	0	0	20	20
b_2	10	0	0	10
b_3	0	15	0	15
	10	15	20	45

$p(a|b)$

b \ a	a_1	a_2	a_3	
b_1	0	0	1	1
b_2	1	0	0	1
b_3	0	1	0	1

Die Merkmale sind perfekt voneinander abhängig: Bedingte Verteilungen sind jeweils auf eine Merkmalsausprägung konzentriert. Bei Kenntnis einer Merkmalsausprägung lässt sich mit Sicherheit die Merkmalsausprägung des anderen Merkmals angeben (bzw. umgekehrt). Damit gilt:

$$C = \sqrt{\frac{k-1}{k}} \quad ; \qquad C_{corr} = \sqrt{\frac{k}{k-1}} \cdot \sqrt{\frac{k-1}{k}} = 1$$

Nachteil des unkorrigierten Kontingenzmaßes: Der größte mögliche Wert ist abhängig von der Anzahl der Merkmalsausprägungen der beiden Merkmale.

$\Rightarrow$ kein direkter Vergleich verschiedener unkorrigierter Kontingenzkoeffizienten möglich.

Aufgabe 28

Im folgenden sind die Ergebnisse einer Umfrage unter 28 Personen wiedergegeben.

Körpergröße	180	179	162	170	180	184	170	160	169	172	182	166	158	164
Gewicht	72	98	52	84	94	95	85	50	59	60	93	47	75	65

Körpergröße	170	170	160	168	179	173	182	178	166	181	176	173	170	161
Gewicht	79	55	53	86	94	70	88	79	76	70	72	79	75	76

(a) Klassieren Sie die Personen anhand des Merkmals „Körpergröße“. Verwenden Sie dazu die Klasseneinteilung $[150, 160), [160, 170), \ldots, [180, 190)$. Berechnen Sie anschließend für jede der Klassen die Quartile, das arithmetische Mittel und die Varianz der zugehörigen Gewichtsangaben.

(b) Stellen Sie die Ergebnisse aus (a) grafisch dar, indem Sie für die Körpergewichte der Personen in den einzelnen Klassen Boxplots erstellen. Zeichnen Sie die Boxplots in ein Diagramm (Körpergröße auf der Abszissen-, Körpergewicht auf der Ordinatenachse) jeweils senkrecht über der entsprechenden Klassenmitte des Merkmals „Körpergröße“ ein. Welche Informationen entnehmen Sie der Grafik?

(c) Klassieren Sie die Personen nun zusätzlich nach dem Merkmal „Gewicht“ anhand der Klassen $[45, 65), [65, 85)$ und $[85, 105)$. Stellen Sie eine Korrelationstabelle der absoluten Häufigkeiten auf, und berechnen Sie anschließend den Kontingenzkoeffizienten.

(d) Erstellen Sie ein dreidimensionales Histogramm für die klassierten Daten.

Lösung: (Deskriptive Statistik, S. 87f, 118, 125f)

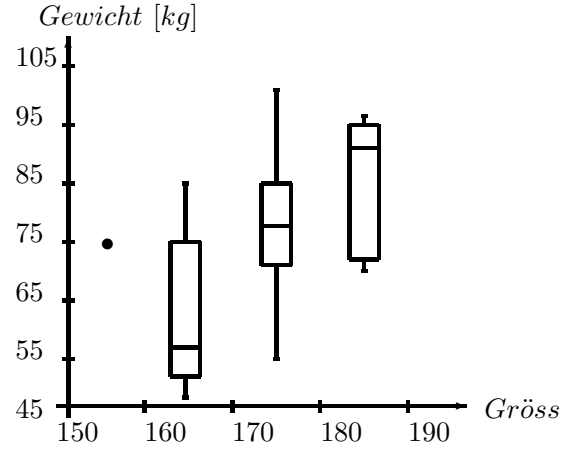
(a) bedingte Lage- und Streuungsparameter

$$\bar{x}_{a|b_j} = \sum_{a \in M_1} a p(a|b_j), \quad \text{bedingtes arithmetisches Mittel von } a \text{ unter Bedingung } b_j$$

$$s^2_{a|b_j} = \sum_{a \in M_1} (a - \bar{x}_{a|b_j})^2 p(a|b_j), \quad s^2_{a|b_j} \text{ bedingte Varianz von } a \text{ unter der Bedingung } b_j$$

	[150;160)	Gewicht	[160;170)	Gewicht	[170:180)	Gewicht	[180:190)	Gewicht
	158	75	160	50	170	55	180	72
			160	53	170	75	180	94
			161	76	170	79	181	70
			162	52	170	84	182	88
			164	65	170	85	182	93
			166	47	172	60	184	95
			166	76	173	70		
			168	86	173	79		
			169	59	176	72		
					178	79		
					179	94		
					179	98		
Mittelwert		75.00		62.67		77.50		85.33
Varianz		0.00		170.22		141.92		107.89
$q_{0.25}$				52.00		71.00		72.00
$q_{0.5}$				59.00		79.00		90.50
$q_{0.75}$				76.00		84.50		94.00

(b) Boxplot: nicht sinnvoll für $n < 5$



- Gewicht und Grösse sind nicht unabhängig
- höhere Gewichte sind tendentiell mit höheren Körpergrössen verbunden (Ursache-Wirkung-Aussagen nicht möglich)

(c)

	[150;160)	Gewicht	[160;170)	Gewicht	[170:180)	Gewicht	[180:190)	Gewicht
[45;65)			166	47	170	55		
			160	50	172	60		
			162	52				
			160	53				
			169	59				
[65;85)	158	75	164	65	173	70	181	70
			161	76	176	72	180	72
			166	76	170	75		
					170	79		
					173	79		
					178	79		
					170	84		
[85;105)			168	86	170	85	182	88
					179	94	182	93
					179	98	180	94
							184	95

	$[150;160)$	$[160;170)$	$[170;180)$	$[180;190)$	
$[45;65)$	0	5	2	0	7
$[65;85)$	1	3	7	2	13
$[85;105)$	0	1	3	4	8
	1	9	12	6	28

$\frac{(h(a,b)-\frac{h(a)h(b)}{n})^2}{\frac{h(a)h(b)}{n}}$	$[150;160)$	$[160;170)$	$[170;180)$	$[180;190)$
$[45;65)$	0.25	3.36111111	0.33333333	1.5
$[65;85)$	0.61813187	0.33241758	0.36630037	0.22161172
$[85;105)$	0.28571429	0.96031746	0.05357143	3.01761905

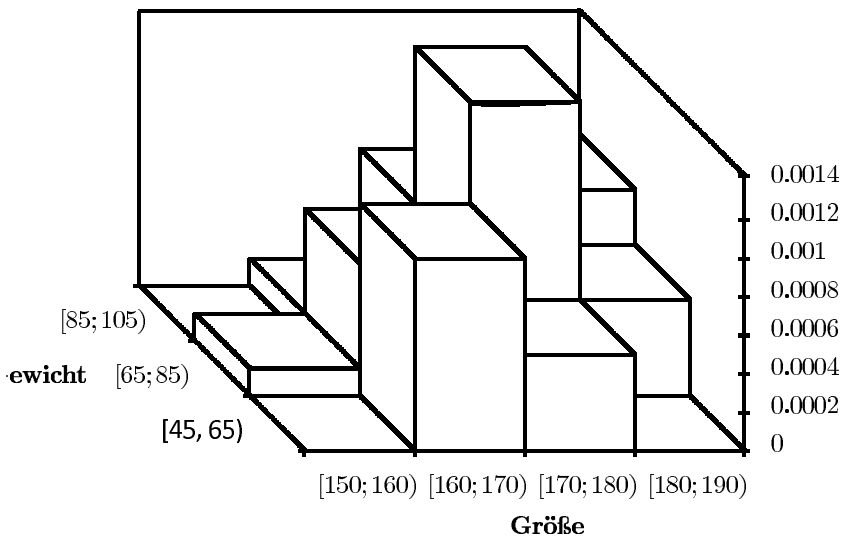
- $\chi^2 = 11.3301282$
- $C = 0.53672859$ und $C_{corr} = 0.65735559$

(d)

	[150; 160)	[160; 170)	[170; 180)	[180; 190)
[45; 65)	–	0.00089286	0.00035714	-
[65; 85)	0.00017857	0.00053571	0.00125	0.00035714
[85; 105)	–	0.00017857	0.00053571	0.00071429

A: Grundfläche der Balken, die Tabelle gibt die Häufigkeitsdichte $= \frac{relative\ Häufigkeit}{A}$ an.

3-dim. Histogramm: volumenproportional



Aufgabe 29

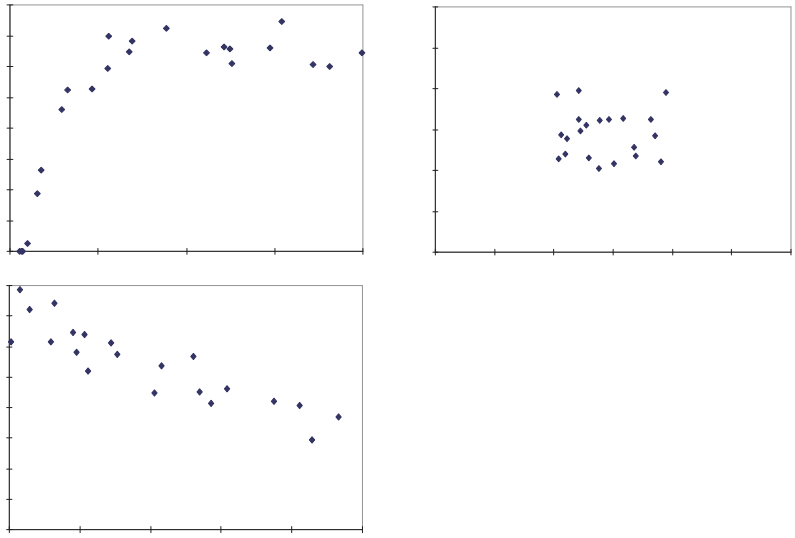
Bei einer Mathematik- und einer Statistik-Klausur wurden von 11 Wiwi-Studenten folgende Punktzahlen erreicht:

Student	1	2	3	4	5	6	7	8	9	10	11
Mathematik	38	47	44	51	35	29	22	14	12	19	9
Statistik	39	34	31	48	46	23	17	12	16	28	10

(a) Zeichnen Sie ein Streudiagramm.

(b) Nehmen Sie an, dass zwischen den Ergebnissen in Mathematik (x) und denen in Statistik (y) der funktionale Zusammenhang $y = f(x)+\epsilon$ besteht, wobei $f(x)$ eine beliebige Funktion von x und ϵ ein Störterm ist.
Welche Form könnte f hier haben?
Wie groß wird der Einfluss des Störterms sein?

(c) Beantworten Sie die Fragen aus (b) für die folgenden Streudiagramme.



Lösung: (Deskriptive Statistik, S. 131ff)

(a)+(b)

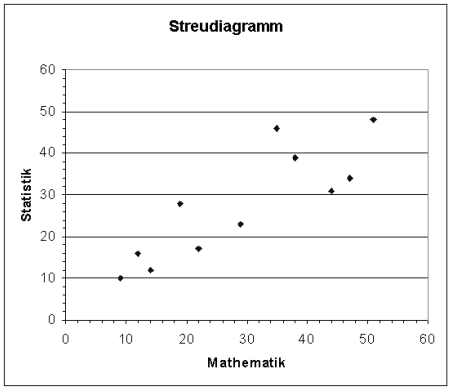


Abbildung 8: **Streudiagramm**

- linearer Zusammenhang ($f(x) = ax + b$) plausibel auf Grund des Streuungsdiagramms
- Ursache – Wirkung – Interpretation nicht zulässig
- Interpretation von ϵ: Abweichung der Punkte vom linearen Zusammenhang wegen nicht beobachteter Einflussfaktoren und Zufall („Tagesform“).

(c)

links oben: es könnte ein logarithmischer bzw. Quadratwurzel Zusammenhang vermutet werden.

rechts oben: kein funktionaler Zusammenhang erkennbar.

links unten: ein linear fallender Zusammenhang erscheint plausibel.

Aufgabe 30

Die folgende Tabelle gibt für die Jahre 1975 - 1994 die Aufwendungen der deutschen Wirtschaft für FuE (Merkmal x) und die Anzahl der Patentanmeldungen (Merkmal y) an.

Jahr	FuE-Aufwendungen in Mrd. DM x_i	Patentanmeldungen in Tausend y_i
1975	14.54	64.595
1985	39.55	45.213
1989	50.81	41.244
1991	55.12	41.799
1992	56.93	43.663
1993	56.76	45.380
1994	56.25	49.011

(a) Zeichnen Sie das Streudiagramm.

(b) Bestimmen Sie die Regressionsgerade $y = a \cdot x + b$.

(c) Berechnen Sie den Korrelationskoeffizienten zwischen x und y.

(d) Welchen Wert für die Anzahl der Patentanmeldungen würden Sie voraussagen, wenn Sie wüssten, dass im Jahr 1998 50 bzw. 60 Mrd. DM für Forschung und Entwicklung aufgewendet werden sollen?

(e) Wiederholen Sie die Aufgabenteile (b)–(d) unter Weglassung des ersten Wertepaares. Wie sind die veränderten Ergebnisse zu interpretieren?

Lösung: (Deskriptive Statistik, S. 131ff, 136ff)

(a)

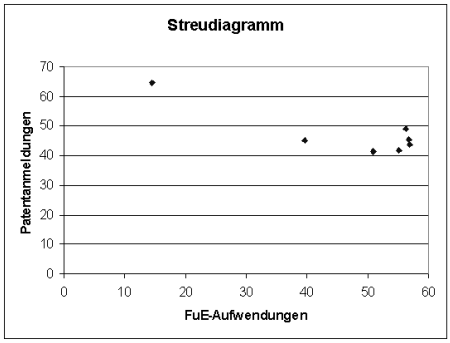


Abbildung 9: **Streudiagramm**

(b) Lineare Regression:

$$\text{Steigung: } \hat{m} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{cov(x,y)}{s_x^2}$$

$$y\text{-Achsenabschnitt: } \hat{b} = \quad \bar{y} - \hat{m}\bar{x} \quad \left(= \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \right)$$

Aus den Werten ergibt sich:

$$\bar{x} = 47,137,\ s_x^2 = 209,843,\ \bar{y} = 47,272,\ s_y^2 = 55,749,\ cov(x,y) = -93,227$$

Somit erhält man für die Steigung, den y-Achsenabschnitt und die Regressionsgerade:

$$\hat{m} = -0,444; \quad \hat{b} = 68,214; \quad \hat{y} = -0,444x + 68,214$$

(c) Der Korrelationskoeffizient (Bravais–Pearson–Korrelationskoeffizient) ist ein Maß für den linearen Zusammenhang der Ausprägungen bzw. Werte zweier Merkmale und besitzt die folgenden Eigenschaften:

- $r = \frac{cov(x,y)}{s_x\, s_y}$
- $-1 \leq r \leq 1$
- $|r| = 1$, falls alle beobachteten Punkte (x, y) auf einer Geraden liegen, die weder waagerecht noch senkrecht ist.

Folgende Bezeichnungen sind üblich:

- Für $r > 0$ heißen die Merkmale positiv korreliert.
- Für $r = 0$ heißen die Merkmale unkorreliert.
- Für $r < 0$ heißen die Merkmale negativ korreliert.

Aus den berechneten Werten ergibt sich $r = -0,862$.

(d) Anhand der berechneten Regressionsgeraden können Prognosen erstellt werden. Konkret ergibt sich:

$$x = 50 \quad \Rightarrow \quad \hat{y} = \hat{m}x + \hat{b} = 46,01$$
$$x = 60 \quad \Rightarrow \quad \hat{y} = \hat{m}x + \hat{b} = 41,57$$

Allerdings muss man sich hierbei folgende Fragen stellen:

- Ist der lineare Trend fortsetzbar? ($x = 50$ Interpolation, d.h. innerhalb des Bereiches der Beobachtungswerte, $x = 60$ Extrapolation, d.h. ausserhalb)
- Ist die Folgerung, dass steigende FuE - Aufwendungen zu einer Abnahme der Patentanmeldungen führen, plausibel? (erste Beobachtung evtl. weglassen, vgl. (e))

(e) Analog zu Aufgabenteil (b) bzw. (c) ergibt sich:

- Regressionsgerade: $y = 0,027x + 42,964$
- Korrelationskoeffizient: $r = 0,0646$

Der Korrelationskoeffizient zeigt, dass ein linearer Zusammenhang nicht als fundiert angesehen werden kann. Eine mögliche Interpretation für diesen stark veränderten Sachverhalt könnte sein:

Die Datenreihe besteht aus relativ wenigen Werten, die noch dazu fast alle sehr dicht beieinander liegen. Dies führt dazu, dass der Wert für das Jahr 1975 das Verhalten der Regressionsgeraden wesentlich beeinflusst bzw. eine Regression nicht mehr gerechtfertigt erscheint, wenn man ihn weglässt.

Aufgabe 31

Von der Controlling-Abteilung eines Konsumgüterherstellers wurde der Zusammenhang zwischen dem Werbebudget w und dem Umsatz u eines Produktes X im Jahr 2001 untersucht. Man erhielt für die 12 Monate folgende Größen:

- Werbeausgaben: insgesamt 120 000 DM im untersuchten Jahr.
- Der durchschnittliche Umsatz des Produktes X in DM beträgt das zehnfache des durchschnittlichen Werbeetats pro Monat.
- Standardabweichung der monatlichen Werbeausgaben im Untersuchungszeitraum: 9 000 DM.
- Standardabweichung der Umsatzdaten: 50 000 DM.
- Als Maß für den Zusammenhang zwischen Werbeausgaben und Umsatz wurde ein Korrelationskoeffizient von 0.90 ermittelt.

(a) Berechnen Sie die Regressionsgerade für den Fall eines einfachen linearen Zusammenhangs der beiden Merkmale auf Monatsebene in der Form $u = m \cdot w + b$.

(b) Welche Maßnahmen würden Sie als Werbemanager ergreifen, falls Sie diesem ermittelten Zusammenhang ohne Einschränkungen vertrauen könnten und das Produkt mit zunehmendem Umsatz den Gewinn des Unternehmens steigert? Warum wäre ein solches Vertrauen unrealistisch?

(c) Skizzieren Sie eine realistischere Funktionsbeziehung, die nicht die Nachteile des Modells in (a) besitzt, und begründen Sie Ihre Modellierung.

Lösung: (Deskriptive Statistik, S. 131ff, 136ff)

(a) Aus den angegebenen Werten entnimmt man bzw. ergibt sich:

- Die durchschnittlichen Werbeausgaben $\bar{w}$ pro Monat von $\bar{w} = \sum_{i=1}^{12} w_i/12 = 120.000\text{DM}/12 = 10\,000$ DM
- Ein durchschnittlicher Umsatz $\bar{u}$ von $\bar{u} = 10 \cdot \bar{w} = 100\,000$ DM
- Die Standardabweichung der Werbeausgaben $s_w = 9\,000$ DM
- Die Standardabweichung des Umsatzes $s_u = 50\,000$ DM
- Der Korrelationskoeffizient $r = 0.9$

Für den Korrelationskoeffizienten gilt : $r = \frac{cov(u,w)}{s_u s_w}$ und daraus ergibt sich: $cov(u,w) = r\, s_u s_w$. Somit erhält man für die Kovarianz:

$$cov(u,w) \;=\; 9\,000 \cdot 50\,000 \cdot 0.9 \;=\; 405\,000\,000.$$

Nun lassen sich Steigung und y-Achsenabschnitt der Regressionsgerade berechnen:

$$\begin{aligned} \hat{m} &= \frac{cov(u,w)}{s_w^2} = \frac{405\,000\,000}{(9\,000)^2} = 5 \quad \left(\text{bzw. } \hat{m} = r\frac{s_u}{s_w} = 0.9 \cdot \frac{50\,000}{9\,000}\right) \\ \bar{u} &= \hat{m}\bar{w} + \hat{b} \Rightarrow \hat{b} = \bar{u} - \bar{m}\bar{w} = 100\,000 - 5 \cdot 10\,000 = 50\,000 \end{aligned}$$

Wir erhalten somit die Regressionsgerade $u = 5w + 50.000$.

(b) **Zur Frage 1:** Diese Modellbeziehung würde den Werbemanager veranlassen, die Werbeausgaben immer weiter zu erhöhen, da damit auch der Gewinn erhöht werden kann.

Zur Frage 2: Die Modellbeziehung geht von einer unbegrenzten Marktnachfrage aus (realistisch: Marktsättigung).

(c) Lineare Funktionen dürften kaum geeignet sein, um einen Zusammenhang zwischen Umsatz und Werbung auszudrücken. Solche Beziehungen werden im allgemeinen durch degressiv wachsende Funktionen beschrieben.

Aufgabe 32

Für 80 Gemeinden verschiedener Größe werden das monatliche Müllaufkommen (= Merkmal y) und die Anzahl der zu Monatsanfang gemeldeten Einwohner (= Merkmal x) einer linearen Regressionsanalyse unterzogen.

(a) Der Analytiker berechnet eine Regressionsgerade, die durch 79 der beobachteten Werte verläuft. Lediglich der Beobachtungswert (x_{80}, y_{80}) liegt nicht auf ihr. Muss zwangsläufig ein Rechenfehler vorliegen, oder ist ein derartiger Befund bei empirischen Daten möglich?

(b) Wären (genau) zwei von Null verschiedene Residuen möglich? Wenn ja, wie?

Lösung: (Deskriptive Statistik, S. 131ff)

(a) Die Residuen geben die Abweichung zwischen den theoretischen (auf der Regressionsgerade) und den tatsächlichen (aus den Daten) Werten an, d.h. die Residuen u_i berechnen sich gemäß: $u_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n$. Für die Summe der Residuen u_i gilt allgemein:

$$\begin{aligned}
\sum_{i=1}^{n} u_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} \left(y_i - \hat{m}\, x_i - \hat{b} \right) \\
&= \sum_{i=1}^{n} y_i - \hat{m} \sum_{i=1}^{n} x_i - n\, \hat{b} \\
&= n\, \bar{y} - \hat{m}\, n\, \bar{x} - n\, \hat{b} \\
&= n \underbrace{\left(\bar{y} - \hat{m}\, \bar{x} - \hat{b} \right)}_{=0} \\
&= 0
\end{aligned}$$

Nach den Berechnungen des Analytikers ist $u_1 = \ldots = u_{79} = 0$, und $u_{80} \neq 0$. Damit ist $\sum\limits_{i=1}^{80} u_i \neq 0$, d.h. es muss ein Rechenfehler vorliegen.

(b) Nach Teilaufgabe (a) gilt: $\sum\limits_{i=1}^{80} u_i = 0$. Seien $u_{i_1},\ u_{i_2} \neq 0$ und $u_i = 0$ für alle $i \neq i_1,\ i_2$, dann folgt aus $\sum\limits_{i=1}^{80} u_i = \sum\limits_{i \neq i_1, i_2} u_i + (u_{i_1} + u_{i_2}) = 0$, dass $u_{i_1} = -u_{i_2}$, da $\sum\limits_{i \neq i_1, i_2} u_i = 0$ ist. Es gilt somit für $u_{i_1} = -u_{i_2} =: \varepsilon$:

$$\begin{aligned}
y_i &= \hat{y}_i &&= \hat{m}\, x_i + \hat{b} \quad \text{für } i \neq i_1,\ i_2 \\
y_{i_1} &= \hat{y}_{i1} &&= \hat{m}\, x_{i_1} + \hat{b} + \varepsilon \\
y_{i_2} &= \hat{y}_{i2} &&= \hat{m}\, x_{i_2} + \hat{b} - \varepsilon
\end{aligned}$$

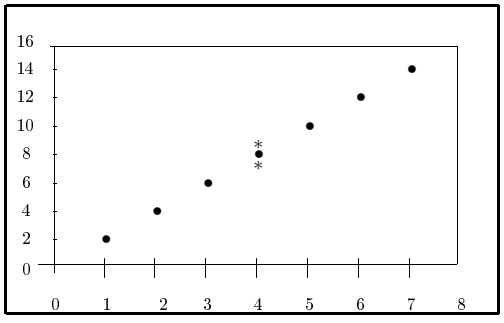
Setzt man dies in die Formel $\hat{m} = \frac{cov(x,y)}{s_x^2}$ ein, so erhält man:

$$\begin{aligned}
\hat{m} &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\
&= \frac{\frac{1}{80}(\sum_{i=1}^{80} x_i\,(\hat{m}\,x_i + \hat{b}) + x_{i_1}\,\varepsilon - x_{i_2}\,\varepsilon) - \bar{x}\,\bar{y}}{\frac{1}{80}\sum_{i=1}^{80} x_i^2 - (\frac{1}{80}\sum_{i=1}^{80} x_i)^2} \\
&= \frac{\frac{1}{80}\sum \hat{m}\,x_i^2 + \bar{x}\,\hat{b} + \frac{\varepsilon}{80}(x_{i_1} - x_{i_2}) - \hat{m}\,\bar{x}_i^2 - \bar{x}\,\hat{b}}{\frac{1}{80}\sum x_i^2 - (\frac{1}{80}\sum x_i)^2} \\
&= \frac{\hat{m}[\frac{1}{80}\sum x_i^2 - (\frac{1}{80}\sum x_i)^2] + \frac{\varepsilon}{80}(x_{i_1} - x_{i_2})}{\frac{1}{80}\sum x_i^2 - (\frac{1}{80}\sum x_i)^2} \\
&= \hat{m} + \frac{\frac{\varepsilon}{80}\left(x_{i_1} - x_{i_2}\right)}{\frac{1}{80}\sum x_i^2 - \left(\frac{1}{80}\sum x_i\right)^2}
\end{aligned}$$

Um diese Gleichung zu erfüllen muss gelten: $x_{i_1} = x_{i_2}$, da $\varepsilon > 0$. Somit wären zwei von Null verschiedene Residuen unter der Bedingung möglich, dass sich die Residuen nur im Vorzeichen unterscheiden und die zugehörigen x-Werte übereinstimmen. Folgendes Zahlenbeispiel verdeutlicht den Sachverhalt:

i	1	2	3	4	5	6	7	8
x_i	1	2	3	4	4	5	6	7
y_i	2	4	6	7	9	10	12	14
$\hat{y}_i$	2	4	6	8	8	10	12	14
u_i	0	0	0	−1	1	0	0	0

Man erhält für die Parameter der Regressionsgeraden $\hat{m} = 2$ und $\hat{b} = 0$.



Aufgabe 33

Aus der Produktionswirtschaft ist das Erfahrungskurvenkonzept bekannt, das den zeitlichen Zusammenhang zwischen Produktionsmenge und Stückkosten beschreibt. Ein Wirtschaftswissenschaftler hat dies folgendermaßen formuliert: „*Mit jeder Verdoppelung der im Zeitablauf kumulierten Produktionsmenge x fallen die realen Stückkosten K potentiell um 20-30%*“. Es wird also der folgende funktionale Zusammenhang unterstellt:

$$K_t = a x_t^{-b} \qquad (t = 1, \ldots, n)$$

(a) Berechnen Sie Schätzwerte für die Parameter a und b des Erfahrungskurvenkonzeptes mit Hilfe des Kleinste-Quadrate-Ansatzes (KQ-Ansatz), indem Sie eine Linearisierung vornehmen.

(b) Entsprechen die über den Umweg einer Linearisierung gewonnenen Schätzer für a und b denen, die man bei einer direkten KQ-Lösung des Ausgangsproblems erhalten würde? Versuchen Sie, Ihre Antwort zu begründen!

Lösung: (Deskriptive Statistik, S. 131ff)

(a) Gegeben sind Beobachtungen $(x_t, K_t),\ t = 1, \ldots, n$, wobei ein funktionaler Zusammenhang des Typs $K_t = a\, x_t^{-b}$ vermutet wird. Durch Logarithmieren der Werte für die Stückkosten K_t ergibt sich der korrespondierende linearisierte Zusammenhang:

$$\underbrace{\ln K_t}_{=:v_t} = \underbrace{\ln a}_{=:\alpha} + \underbrace{(-b)}_{=:\beta} \cdot \underbrace{\ln x_t}_{=:u_t}$$

Für die auf diese Weise transformierten Beobachtungen kann eine lineare Regression durchgeführt werden. Als KQ-Werte für α und β ergeben sich:

$$\hat{\beta} = \frac{\sum_{t=1}^{n}(u_t - \bar{u})(v_t - \bar{v})}{\sum_{t=1}^{n}(u_t - \bar{u})^2}, \qquad \hat{\alpha} = \bar{v} - \hat{\beta}\bar{u}$$

Für a und b folgt dann:

$$\begin{aligned} \hat{b} &= -\hat{\beta} = -\frac{\sum(u_t - \bar{u})(v_t - \bar{v})}{\sum(u_t - \bar{u})^2} \\ \hat{a} &= e^{\hat{\alpha}} = e^{(\bar{v} - \hat{\beta}\bar{u})} \end{aligned}$$

(b) Der KQ-Ansatz für K_t lautet:

$$\min_{a,b} \sum_{t=1}^{n} (K_t - a x_t^{-b})^2$$

Es ergeben sich die Normalgleichungen:

$$\begin{aligned}
\frac{\partial}{\partial a}\sum_{t=1}^{n}(K_t - ax_t^{-b})^2 &= 2\sum_{t=1}^{n}(K_t - ax_t^{-b})(-x_t^{-b}) \stackrel{!}{=} 0 \\
\frac{\partial}{\partial b}\sum_{t=1}^{n}(K_t - ax_t^{-b})^2 &= 2\sum_{t=1}^{n}(K_t - ax_t^{-b})(-a\cdot(-\ln(x_t))\,x_t^{-b}) \stackrel{!}{=} 0
\end{aligned}$$

Dieses Optimierungsproblem lässt sich nur numerisch lösen. Die auf diesem Wege erhaltenen Schätzer für a und b entsprechen im allgemeinen Fall sicherlich nicht denjenigen, die man über das linearisierte Problem erhält. Beispielsweise besitzt $\hat{K}_t = e^{\hat{\alpha}}x^{\hat{\beta}}$ im allgemeinen nicht mehr die Eigenschaft, die Summe der Abstandsquadrate der Beobachtungen (x_t, K_t) von der Kurve zu minimieren ($\hat{K}_t$ ist nicht die Lösung eines KQ-Ansatzes für K_t).

Aufgabe 34

Passen Sie an die folgenden Daten mit Hilfe des KQ-Ansatzes das Modell $y = mx^2 + b$ an, und berechnen Sie den Wert des Bestimmtheitsmaßes.

x	0.497	2.648	0.247	3.3	3.936	3.505	0.881	2.835	0.025	1.383
y	3.435	15.052	-0.810	24.110	34.402	23.956	5.0323	16.211	0.595	2.735

Lösung: (Deskriptive Statistik, S. 131ff)

Der KQ - Ansatz für $y = mx^2 + b$ lautet:

$$\min_{m,b} \sum_{i=1}^{n} \left(y_i - mx_i^2 - b\right)^2$$

Als notwendige Bedingungen für ein Minimum erhalten wir die Normalengleichungen:

$$\begin{aligned} \frac{\partial \sum(\ldots)^2}{\partial\, m} &= 2\sum_{i=1}^{n} \left(y_i - m\, x_i^2 - b\right)\left(-x_i^2\right) \stackrel{!}{=} 0 \\ \frac{\sum(\ldots)^2}{\partial\, b} &= 2\sum_{i=1}^{n} \left(y_i - m\, x_i^2 - b\right)(-1) \stackrel{!}{=} 0 \end{aligned}$$

Das Ergebnis lautet [6]:

$$\hat{m} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2\, y_i - \bar{x^2}\,\bar{y}}{\frac{1}{n}\sum\limits_{i=1}^{n} {x_i}^4 - \overline{x^2}^2}\,, \qquad \hat{b} = \bar{y} - \hat{m}\,\overline{x^2}$$

Bemerkung: Das gleiche Ergebnis erhält man durch die Transformation $z_i = x_i^2$ und eine lineare Regression mit dem Ansatz $y_i = m\, z_i + b$.

Für das Bestimmtheitsmaß ergibt sich:

$$r^2 = \frac{\hat{m}^2 s_{x^2}^2}{s_y^2} = \frac{cov(x^2, y)^2}{s_{x^2}^2 s_y^2} = 0.979$$

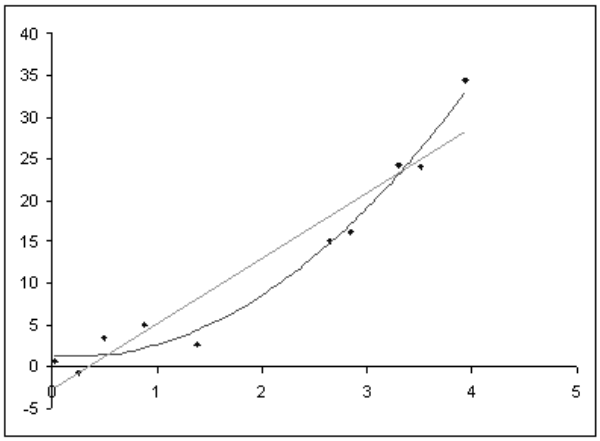
x	0.025	0.247	0.497	0.881	1.383	2.648	2.835	3.300	3.505	3.936
x^2	0.001	0.061	0.247	0.776	1.913	7.012	8.037	10.890	12.285	15.492
y	0.595	-0.810	3.435	5.032	2.735	15.052	16.211	24.110	23.956	34.402

Man erhält die Regressionsparameter $\hat{m} = 2.053$ und $\hat{b} = 0.826$. Somit ergibt sich:

$\hat{y}$	0.827	0.951	1.333	2.420	4.754	15.224	17.330	23.188	26.052	32.638

[6] Eigentlich muss noch eine hinreichende Bedingung für ein Minimum nachgewiesen werden, dies wird aber hier nicht durchgeführt.

Zur grafischen Illustration folgt ein Schaubild in der zusätzlich die Lösung der linearen Regression mit dem Ansatz $y_i = \alpha\, x_i + \beta$ eingezeichnet ist:



Aufgabe 35

Im folgenden sind noch einmal die Klausurergebnisse aus Übung 29 gegeben:

Student	1	2	3	4	5	6	7	8	9	10	11
Mathematik	38	47	44	51	35	29	22	14	12	19	9
Statistik	39	34	31	48	46	23	17	12	16	28	10

(a) Berechnen Sie den Rangkorrelationskoeffizienten von Spearman.

(b) Bei Student 5 erhöht sich die Anzahl der Punkte in der Mathematikklausur nachträglich um 3 auf 38 Punkte. Welche Auswirkung hat diese Änderung auf den Wert des Rangkorrelationskoeffizienten?

Lösung: (Deskriptive Statistik, S. 142ff)

(a) Die Rangziffern r_i (für Mathematik) bzw. s_i (für Statistik) des Studenten i ergeben sich aus der Rangfolge der erreichten Punktzahlen in der jeweiligen Klausur, d.h. dem Platz in der geordneten Urliste. Man erhält somit die folgende Tabelle:

i	1	2	3	4	5	6	7	8	9	10	11
r_i	8	10	9	11	7	6	5	3	2	4	1
s_i	9	8	7	11	10	5	4	2	3	6	1

Laut Definition berechnet sich nun der Rangkorrelationskoeffizient r_S von Spearman gemäß:

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n(n^2-1)} = 1 - \frac{6\left((8-9)^2 + (10-8)^2 + \cdots + (1-1)^2\right)}{11(11^2-1)} = 0.882$$

Bemerkung: Da in diesem Fall die Rangziffern r_i und s_i jeweils die Werte $1, \ldots, n$ annehmen, stimmt r_S mit dem Korrelationskoeffizient nach Bravais-Pearson der Rangziffernpaare überein.

(b) Die Änderung führt zu einer Bindung beim Merkmal Mathematik (Student 1 und 5). Die Studenten 1 und 5 konkurrieren folglich um die Plätze 7 und 8 und erhalten beide die Rangziffer 7.5. Es ergeben sich also folgende Änderungen zu (a): $(r_1, s_1) = (7.5; 9)$ und $(r_5, s_5) = (7.5; 10)$. Man erhält: $r_S = 0.884$.

Bemerkung: Der Korrelationskoeffizient nach Bravais-Pearson der Rangzifferpaare ist 0.888, d.h. die doppelt auftretende Rangziffer 7.5 wirkt sich kaum auf das Ergebnis aus.

Aufgabe 36

Gegeben sei die folgende Aussage: *In einem x, y-Koordinatensystem befinde sich eine Punktewolke mit Schwerpunkt im Ursprung. Dann gilt: Die Gerade g aus einer linearen Regression von y auf x und die Gerade g' aus einer linearen Regression von x auf y sind nicht identisch, außer x und y sind vollständig miteinander korreliert ($|r| = 1$).*
Führen Sie einen rechnerischen Nachweis für diese Aussage, und überlegen Sie sich die Ursache für den angesprochenen Unterschied.

Lösung: (Deskriptive Statistik, S. 131ff, 136ff)

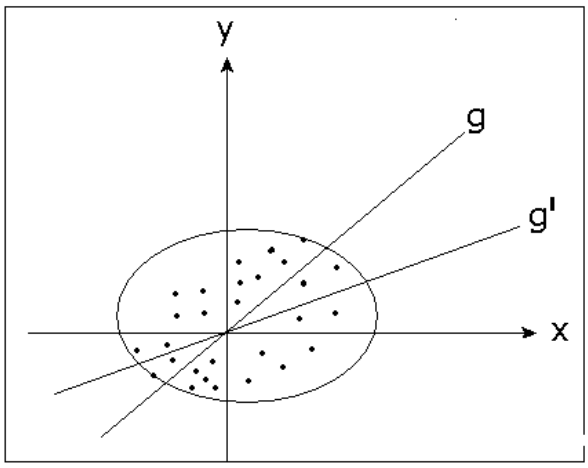


Abbildung 10: Streudiagramm mit hypothetischen Regressionsgeraden

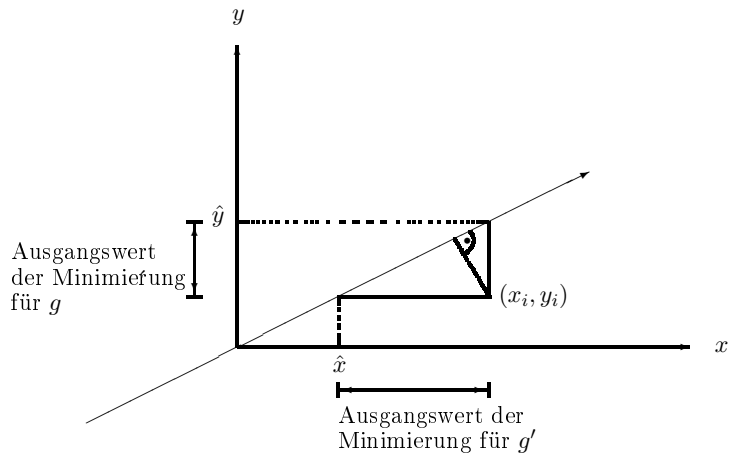
Bemerkung: Es ist durch ein Verschieben des Ursprungs auf den Schnittpunkt (falls es mehrere gibt, einen beliebigen) der Geraden g und g' immer möglich, dass beide Geraden einen y-Achsenabschnitt von 0 haben. Folglich erhalten wir als Bedingung für die Gleichheit der beiden Geraden $g : y = m_1\, x$ bzw. $g' : y = m_2\, x$ folgende Bedingung:

$$g \equiv g' \iff m_1 = m_2.$$

Aus $m_1 = \frac{cov(x,y)}{s_x^2}$ bzw. $m_2 = \frac{s_y^2}{cov(x,y)}$ ergibt sich durch Umformung folgende Bedingung für die Gleichheit der beiden Geraden:

$$r^2 = \frac{cov(x,y)^2}{s_x^2 s_y^2} = 1$$

Die beiden Geraden g und g' stimmen genau dann überein, wenn die angegeben Daten eine Gerade bilden. Im allgemeinen gilt also $m_1 \neq m_2$ bzw. $g \neq g'$ und der Grund liegt darin, dass bei einer Regression von y auf x ($y = m\, x + b$) die senkrechten Abstände der Punkte von der Geraden in die Summe der quadrierten Abweichungen, bei einer Regression von x auf y ($x = m'\, y + b'$) die waagerechten Abstände eingehen.



Bemerkung: Bei Verwendung des Orthogonalabstandes ergibt sich für g und g' die gleiche Lösung, aber i.a. nicht die Lösung der linearen Regression.

Aufgabe 37

(a) Welche der folgenden Aussagen sind richtig, welche falsch? (Begründung!)

(1) Je größer der Anstieg der Regressionsgeraden, desto größer ist der Korrelationskoeffizient.

(2) Ist der Anstieg der Regressionsgeraden positiv, so ist auch der Korrelationskoeffizient positiv.

(3) Das Bestimmtheitsmaß wird mit wachsender Anzahl der Beobachtungen größer.

(4) Ist das Bestimmtheitsmaß gleich 1, so sind alle Residuen gleich 0.

(5) Ist der Korrelationskoeffizient gleich 0, so liegen alle Beobachtungswerte auf einer waagerechten Geraden.

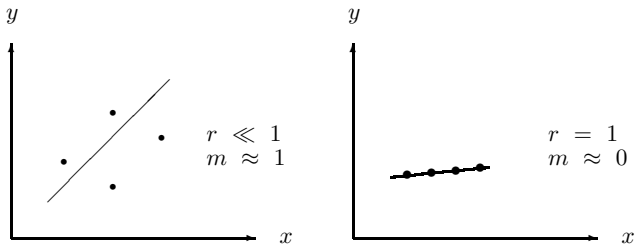
(6) Ist der Korrelationskoeffizient gleich 0, so sind die Merkmale X und Y unabhängig.

(b) Gegeben sei das Modell $y = b + \frac{m}{x}$. Wie ist x zu transformieren, damit eine lineare Regression durchgeführt werden kann?

(c) Überlegen Sie sich ein Beispiel für zwei Merkmale, die unkorreliert aber nicht unabhängig sind.

Lösung: (Deskriptive Statistik, S. 131ff, 136ff)

(a) (1) Falsch: Es gilt $m = \frac{cov(x,y)}{s_x^2}$, $r = \frac{cov(x,y)}{s_x s_y}$, also $r = m\frac{s_x}{s_y}$, r ist folglich nicht alleine von m abhängig und insbesondere nicht monoton in m. Die folgende Skizze illustriert ein Gegenbeispiel:



(2) Richtig: $r = m\frac{s_x}{s_y}, \quad s_x, s_y > 0 \Rightarrow sgn(r) = sgn(m)$

(3) Falsch, Gegenbeispiel: Für $n = 2$ gilt stets $r^2 = 1$, für $n > 2$ i.a. nicht (falls keine perfekte Korrelation vorliegt)

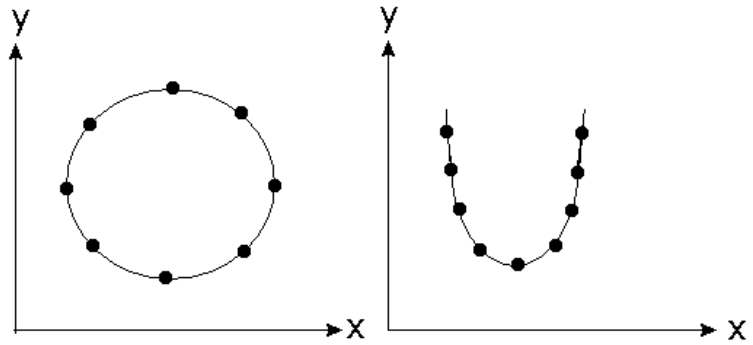
(4) Richtig: Wenn $r^2 = 1$ gilt, liegt perfekte Korrelation vor (also alle beobachteten Werte liegen auf der Regressionsgeraden) und damit $u_i = y_i - \hat{y}_i = 0$ für alle $i = 1, \ldots, n$.

(5), (6) Falsch: $r = 0$ bedeutet, dass die betrachteten Merkmale unkorreliert sind. Der Korrelationskoeffizient ist ein Maß für den linearen Zusammenhang zwischen den Merkmalen, d.h. selbst wenn dieser nicht vorliegt ($r = 0$), kann sehr wohl ein anderer

Zusammenhang bestehen (s. Skizze). Unabhängigkeit bedeutet, dass kein funktionaler Zusammenhang zwischen den Merkmalen besteht, insbesondere also auch kein linearer. Wir erhalten also Aussage:

Unabhängigkeit $\Longrightarrow$ Unkorreliertheit, Unkorreliertheit $\not\Longrightarrow$ Unabhängigkeit

Gegenbeispiele für unkorrelierte abhängige Merkmale:



(b) y ist linear in $\frac{1}{x}$, d.h. mit $z = \frac{1}{x}$ gilt: $y = mz + b$. Man wählt folglich die Transformation $x \mapsto z = \frac{1}{x}$.

(c) Man definiere die folgenden Beobachtungspaare zweier Merkmale A und B (bsp.: $1 = x^2 + y^2$):

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right); (1,0); \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right); (0,-1); \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right); (-1,0); \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right); (0,1)$$

Mit $\bar{x} = \bar{y} = 0$ folgt: $cov(x,y) = \frac{1}{n}\sum_{i=1}^{8} x_i y_i - \bar{x}\bar{y} = 0$, d.h. die Merkmale sind unkorreliert (da $s_A, s_B > 0$). Die Merkmale sind jedoch nicht unabhängig, denn die bedingte relative Häufigkeit der Merkmalsausprägung $\frac{\sqrt{2}}{2}$ bei Merkmal A unter der Bedingung 0 bei Merkmal B

$$p(\frac{\sqrt{2}}{2}|0) \;=\; \frac{p(\frac{\sqrt{2}}{2}, 0)}{p(0)} \;=\; 0$$

ist verschieden von der bedingten relativen Häufigkeiten von $\frac{\sqrt{2}}{2}$ bei Merkmal A unter der Bedingung $\frac{\sqrt{2}}{2}$ bei Merkmal B

$$p(\frac{\sqrt{2}}{2}|\frac{\sqrt{2}}{2}) \;=\; \frac{p(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})}{p(\frac{\sqrt{2}}{2})} \;=\; \frac{\frac{1}{8}}{\frac{2}{8}} \;=\; \frac{1}{2}.$$

Die Gleichheit der bedingten relativen Häufigkeitsverteilungen ist aber eine zur Unabhängigkeit äquivalente Eigenschaft.

Aufgabe 38 (Bamberg / Bauer, S. 67ff)

Die folgende Zeitreihe beschreibt monatliche Einfuhrmengen eines Industriezweiges in die Bundesrepublik Deutschland:

Monat / Jahr	J	F	M	A	M	J	J	A	S	O	N	D
1954	21	29	38	32	31	24	34	54	63	52	46	38
1955	26	34	40	36	24	28	43	69	63	60	46	42
1956	34	41	46	37	35	37	42	61	47	42	35	37

(a) Berechnen Sie „sinnvolle“ gleitende Durchschnitte für diese Zeitreihenwerte, und vergleichen Sie diese mit den gleitenden Durchschnitten der Ordnung 7.

(b) Zeichnen Sie die zugehörigen Zeitreihen in ein Koordinatensystem.

(c) Bestimmen Sie die saisonbereinigten Werte für einen additiven Ansatz mit konstanter Saisonfigur.

Lösung: (Deskriptive Statistik, S. 146ff)

Zur Berechnung gleitender Durchchnitte:

ungerade Ordnung (2k+1): $x_t^* = \frac{1}{2k+1} \sum_{\tau=t-k}^{t+k} x_\tau \quad (t = k+1, \ldots, n-k)$

gerade Ordnung (2k): $x_t^* = \frac{1}{2k}(\frac{1}{2}x_{t-k} + \frac{1}{2}x_{t+k} + \sum_{\tau=t-(k-1)}^{t+(k-1)} x_\tau) \quad (t = k+1, \ldots, n-k)$

(a) „Sinnvolle“ gleitende Durchschnitte sind hier die gleitenden 12-Monats-Durchschnitte, da es sich hier um Daten (Einfuhrmengen) handelt, die eine über die Jahre gleichförmige Entwicklung vermuten lassen (zur Bestätigung der Vermutung kann die Graphik aus (b) herangezogen werden).

	1954			1955			1956		
		Ordnung 7	Ordnung 12		Ordnung 7	Ordnung 12		Ordnung 7	Ordnung 12
Monat	y_t	y_t^*	y_t^*	y_t	y_t^*	y_t^*	y_t	y_t^*	y_t^*
Januar	21	-	-	26	39	40	34	44	46
Februar	29	-	-	34	35	41	41	40	46
März	38	-	-	40	32	42	46	39	45
April	32	30	-	36	33	42	37	39	43
Mai	31	35	-	24	39	42	35	43	42
Juni	24	39	-	28	43	42	37	44	41
Juli	34	41	39	43	46	43	42	43	-
August	54	43	39	69	48	44	61	43	-
September	63	44	39	63	50	44	47	43	-
Oktober	52	45	40	60	51	44	42	-	-
November	46	45	40	46	51	45	35	-	-
Dezember	38	43	39	42	47	46	37	-	-

(b)

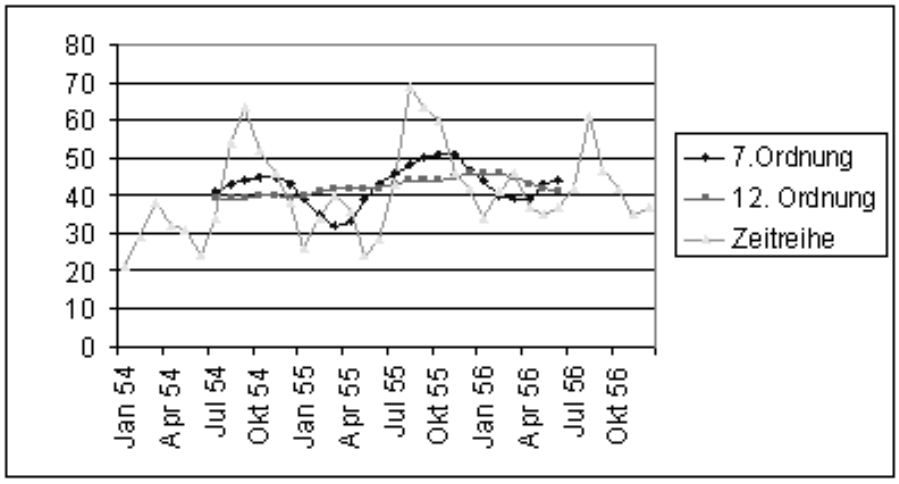


Abbildung 11: Zeitreihe mit unterschiedlichen gleitenden Durchschnitten

(c)

Monat	J	F	M	A	M	J	J	A	S	O	N	D
$y_t - y_t^*$							-5	15	24	12	6	-1
	-14	-7	-2	-6	-18	-14	0	25	19	16	1	-4
	-12	-5	1	-6	-7	-4						
$\bar{S}_j$	-13	-6	-0.5	-6	-12.5	-9	-2.5	20	21.5	14	3.5	-2.5
$\hat{S}_j$	-13,6	-6.6	-1.1	-6.6	-13.1	-9.6	-3.1	19.4	20.9	13.4	2.9	-3.1

Korrekturglied $\frac{1}{12}\sum_{i=1}^{n}\bar{S}_j = \frac{7}{12} = 0.583 \approx 0.6$. Als saisonbereinigte Zahlen ergeben sich:

Einfuhrmengen	1954	1955	1956
Januar	34.6	39.6	47.6
Februar	35.6	40.6	47.6
März	39.1	41.1	47.1
April	38.6	42.6	43.6
Mai	44.1	37.1	48.1
Juni	33.6	37.6	46.6
Juli	37.1	46.1	45.1
August	34.6	49.6	41.6
September	42.1	42.1	26.1
Oktober	38.6	46.6	28.6
November	43.1	43.1	32.1
Dezember	41.1	45.1	40.1

Aufgabe 39

Ermitteln Sie die 12 Saisonindexziffern für folgende Zeitreihe der Arbeitslosenzahl eines Landes (die Werte sind bereits um die glatte Komponente bereinigt) unter der Voraussetzung einer variablen Saisonfigur.

Monat / Jahr	J	F	M	A	M	J	J	A	S	O	N	D
1976	1.7	1.7	1.1	1.2	0.8	0.7	0.8	0.8	0.9	1.0	1.1	1.4
1977	1.8	1.6	1.5	0.9	0.9	0.6	0.6	0.8	1.0	0.9	1.2	1.3
1978	1.6	1.5	1.3	0.9	0.7	0.8	0.7	0.8	0.8	1.1	1.0	1.2

Lösung: (Deskriptive Statistik, S. 157ff)

Gegeben sind die Werte $\frac{y_t}{y_t^*}$ und die Periodenlänge $s = 12$:

j	1	2	3	4	5	6	7	8	9	10	11	12
$\bar{I}_j$	1.7	1.6	1.3	1.0	0.8	0.7	0.7	0.8	0.9	1.0	1.1	1.3
$\hat{I}_j$	1.58	1.49	1.21	0.93	0.74	0.65	0.65	0.74	0.84	0.93	1.02	1.21

- $\hat{I}_j$: Saisonindexziffern
- Der Korrekturfaktor beträgt $\frac{12}{\sum_{j=1}^{12} \bar{I}_j} = 0.93$
- $\hat{I}_j = 0.93 \cdot \bar{I}_j$
- Verfahren enthält einen systematischen Fehler, d.h. es liefert auch ohne die Störkomponente i.a. keine exakten Werte.

Aufgabe 40 (Heller et. al., S. 188)

(a) Man stelle den Laspeyres- und Paasche-Mengenindex als gewichtetes arithmetisches Mittel von Mengenmeßzahlen dar und interpretiere die Gewichte.

(b) Es gilt für ein Gut i: Umsatz = Preis x Menge.
Läßt sich diese Gleichung auch auf die entsprechenden Indices übertragen?

(c) Ein Unternehmer stelle 3 Produkte A, B und C her; in den Jahren 2010 und 2011 werden folgende Mengen zu folgenden Preisen abgesetzt:

	Preise je Stück		**abgesetzte Mengen**	
	2010	2011	2010	2011
Produkt A	20	21	100	110
Produkt B	45	44	95	115
Produkt C	53	58	80	85

Zeigen Sie, dass die folgende Indentitäten gelten:

$$W_{0,n} = P^L_{0,n} \cdot Q^P_{0,n} \qquad\qquad W_{0,n} = P^P_{0,n} \cdot Q^L_{0,n}.$$

Lösung: (Deskriptive Statistik, S. 170ff)

(a)

$$Q^L_{0,n} = \frac{\sum\limits_{i\in G} q_n^{(i)} \cdot p_0^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_0^{(i)}} = \frac{\sum\limits_{i\in G} \frac{q_n^{(i)}}{q_0^{(i)}} \cdot q_0^{(i)} \cdot p_0^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_0^{(i)}} = \sum_{i\in G} \frac{q_n^{(i)}}{q_0^{(i)}} \cdot \frac{q_0^{(i)} \cdot p_0^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_0^{(i)}} = \sum_{i\in G} \frac{q_n^{(i)}}{q_0^{(i)}} \cdot f_i^{(0)}$$

$$f_i^{(0)} = \left\{ \begin{array}{l} \text{Wertanteil des Gutes } i \text{ am Gesamtwert des} \\ \text{Warenkorbs zum Basiszeitpunkt} \end{array} \right.$$

$$Q^P_{0,n} = \frac{\sum\limits_{i\in G} q_n^{(i)} \cdot p_n^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_n^{(i)}} = \sum_{i\in G} \frac{q_n^{(i)}}{q_0^{(i)}} \cdot \frac{q_0^{(i)} \cdot p_n^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_n^{(i)}} = \sum_{i\in G} \frac{q_n^{(i)}}{q_0^{(i)}} \cdot f_i^{(0,n)}$$

$$f_i^{(0,n)} = \left\{ \begin{array}{l} \text{Wertanteil des Gutes } i \text{ am Warenkorb des} \\ \text{Basiszeitpunktes, falls dieser zu Preisen} \\ \text{des Berichtszeitpunktes bewertet wird.} \end{array} \right.$$

Die Gewichte $f_i^{(0)}$ und $f_i^{(0,n)}$ sind bei den Quantitätsindices dieselben wie bei den entsprechenden Preisindices.

(b)

Die Gleichung Umsatz = Preis x Menge auf die entsprechenden Indices übertragen, würde lauten:

$$W_{0,n} = P^P_{0,n} \cdot Q^P_{0,n}$$

bzw.

$$W_{0,n} = P^L_{0,n} \cdot Q^L_{0,n}.$$

Aus Aufgabenteil a) ist aber ersichtlich, dass diese Gleichungen nicht gelten.

(c)

$$P^L_{0,n} \cdot Q^P_{0,n} = \frac{\sum\limits_{i\in G} p_n^{(i)} \cdot q_0^{(i)}}{\sum\limits_{i\in G} p_0^{(i)} \cdot q_0^{(i)}} \cdot \frac{\sum\limits_{i\in G} q_n^{(i)} \cdot p_n^{(i)}}{\sum\limits_{i\in G} q_0^{(i)} \cdot p_n^{(i)}} = \frac{\sum\limits_{i\in G} q_n^{(i)} \cdot p_n^{(i)}}{\sum\limits_{i\in G} p_0^{(i)} \cdot q_0^{(i)}} = W_{0,n}$$

$$P^P_{0,n} \cdot Q^L_{0,n} = \frac{\sum\limits_{i\in G} p_n^{(i)} \cdot q_n^{(i)}}{\sum\limits_{i\in G} p_0^{(i)} \cdot q_n^{(i)}} \cdot \frac{\sum\limits_{i\in G} q_n^{(i)} \cdot p_0^{(i)}}{\sum\limits_{i\in G} p_0^{(i)} \cdot q_0^{(i)}} = \frac{\sum\limits_{i\in G} p_n^{(i)} \cdot q_n^{(i)}}{\sum\limits_{i\in G} p_0^{(i)} \cdot q_0^{(i)}} = W_{0,n}$$

$$P^L_{10,11} = \frac{21 \cdot 100 + 44 \cdot 95 + 58 \cdot 80}{20 \cdot 100 + 45 \cdot 95 + 53 \cdot 80} = \frac{10920}{10515} \approx 1.039$$

$$P^P_{10,11} = \frac{21 \cdot 110 + 44 \cdot 115 + 58 \cdot 85}{20 \cdot 110 + 45 \cdot 115 + 53 \cdot 85} = \frac{12300}{11880} \approx 1.035$$

$$Q^L_{10,11} = \frac{110 \cdot 20 + 115 \cdot 45 + 85 \cdot 53}{100 \cdot 20 + 95 \cdot 45 + 80 \cdot 53} = \frac{11880}{10515} \approx 1.130$$

$$Q^P_{10,11} = \frac{110 \cdot 21 + 115 \cdot 44 + 85 \cdot 58}{100 \cdot 21 + 95 \cdot 44 + 80 \cdot 58} = \frac{12300}{10920} \approx 1.126$$

$$P^L_{10,11} \cdot Q^P_{10,11} = 1.039 \cdot 1.126 = \frac{10920}{10515} \cdot \frac{12300}{10920} = \frac{12300}{10515} \approx 1.170$$

$$P^P_{10,11} \cdot Q^L_{10,11} = 1.035 \cdot 1.030 = \frac{12300}{11880} \cdot \frac{11880}{10515} = \frac{12300}{10515} \approx 1.170$$

Aufgabe 41

(a) Erläutern Sie die folgenden Begriffe, und geben Sie jeweils Beispiele an:

- Grundraum Ω
- σ-Algebra $A(\Omega)$
- Wahrscheinlichkeitsmaß P

Welche Bedeutung haben die einzelnen Komponenten für den *Wahrscheinlichkeitsraum*?

(b) Bilden Sie alle σ-Algebren für den Grundraum $\Omega := \{a, b, c\}$.

(c) Gegeben sei der Grundraum $\Omega := \{a, b, c, d\}$. Zeigen Sie, dass $A(\Omega) := \{\emptyset, \{a, d\}, \{b, c\}, \Omega\}$ eine σ-Algebra auf Ω ist.

Lösung: (Wahrscheinlichkeitstheorie, S. 17ff)

(a) Im Kontext der Wahrscheinlichkeit handelt es sich bei

- einem Grundraum Ω um eine beliebige nichtleere Menge. Die Elemente $\omega \in \Omega$ repräsentieren dann mögliche Realisationen eines Zufallsexperiments und heißen Elementarereignisse. Beispielsweise könnte der Grundraum $\Omega = \{0, \ldots, 36\}$ die möglichen Realisationen bei einem Roulette-Spiel beschreiben.
- einer σ-Algebra $A(\Omega)$ um ein Mengensystem von Teilmengen von Ω mit den Eigenschaften:

  1.) $\emptyset \in A(\Omega)$

  2.) $A \in A(\Omega) \Rightarrow A^c \in A(\Omega)$

  3.) $A_1, A_2, \ldots \in A(\Omega) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in A(\Omega)$

  Die σ-Algebra dient der Modellierung von Ereignissen, die mehrere Elementarereignisse umfassen. Beispielsweise könnte man beim Roulette-Spiel sich für das Ereignis „Pair“ interessieren, welches bedeutet, dass das Ergebnis eine gerade Zahl ist und aus der Menge $\{2, 4, 6, \ldots, 36\} \subseteq \Omega$ besteht. Ein Beispiel für eine σ-Algebra ist im Falle eines höchstens abzählbar unendlichen Grundraums Ω immer $A(\Omega) := \mathcal{P}(\Omega)$, d.h die Potenzmenge von Ω.
- einem Wahrscheinlichkeitsmaß P, um eine Abbildung von einer σ-Algebra $A(\Omega)$ in das reelle Einheitsintervall [0,1] mit den Eigenschaften :

  1.) $P(\Omega) = 1$

  2.) $P(A) \geq 0$ für alle $A \in A(\Omega)$

  3.) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ für beliebige paarweise disjunkte $A_i \in A(\Omega), i = 1, 2, \ldots$

  Die erste Eigenschaft normiert das Wahrscheinlichkeitsmaß auf einen größtmöglichen Wert von 1. Die zweite Eigenschaft verhindert negative Wahrscheinlichkeiten, während die dritte Eigenschaft gewährleistet, dass sich das Wahrscheinlichkeitsmaß auf einer abzählbaren disjunkten Zerlegung einer Menge additiv verhält (σ-Additivität). Beispielsweise könnte man für das Roulette-Spiel ein Wahrscheinlichkeitsmaß P auf der Potenzmenge von $\Omega = \{0, 1, 2, \ldots, 36\}$ definieren, indem man einer Menge $A \subseteq \Omega$ die Wahrscheinlichkeit $P(A) := \frac{\#A}{37}$ zuordnet.

(b) Gegeben ist der Grundraum $\Omega = \{a, b, c\}$ und gesucht werden alle σ-Algebren über Ω.

Behauptung: Jede σ-Algebra $A(\Omega)$ über Ω ist von einem der folgenden drei Typen:

a) $A(\Omega) = \{\emptyset,\ \Omega\}$
b) $A(\Omega) = \{\emptyset,\ A,\ A^c,\ \Omega\}$ mit $\emptyset \subsetneq A \subsetneq \Omega$
c) $A(\Omega) = \mathcal{P}(\Omega)$

Offensichtlicherweise erfüllen alle der genannten Mengensystemen die Anforderungen an eine σ-Algebra. Zu zeigen ist also, dass es keine anderen gibt.

Beweis: Sei $A(\Omega)$ eine σ-Algebra über Ω, dann gilt:

1.) $\Omega \,\epsilon\, A(\Omega)$
2.) $\Omega^c = \emptyset \,\epsilon\, A(\Omega)$

Ist $A(\Omega) \neq \{\emptyset, \Omega\}$ gibt es eine Teilmenge A ($\emptyset \neq A \neq \Omega$) mit $A \in A(\Omega)$. Dann ist wegen Eigenschaft 2 einer σ-Algebra $A^c \in A(\Omega)$. Damit ist $A(\Omega)$ vom Typ b) oder es gibt eine weitere Teilmenge B in $A(\Omega)$ mit $\emptyset \neq B \neq \Omega$ und $A \neq B \neq A^c$. A oder A^c ist einelementig. Sei also o.B.d.A. $A = \{a\}$, dann ist B einelementig (also $B = \{b\}$ oder $B = \{c\}$) oder $B \cap A^c$ einelementig. Damit ist $(A \cup B)^c$ oder $(A \cup (B \cap A^c))^c$ in $A(\Omega)$ die verbleibende einelementige Teilmenge. Ist also $A(\Omega)$ nicht vom Typ a) oder b), so enthält $A(\Omega)$ jede einelementige Teilmenge und damit wegen Eigenschaft 3 jede Teilmenge. $A(\Omega)$ ist also die Potenzmenge (Typ c).

(c) Wir beweisen die allgemeinere Aussage, dass für einen beliebigen Grundraum Ω und eine beliebige nicht triviale Teilmenge $\emptyset \subsetneq A \subsetneq \Omega$ das Mengensystem $A(\Omega) := \{\emptyset,\ A,\ A^C,\ \Omega\}$ eine σ-Algebra über Ω definiert.

Beweis durch Überprüfen der Eigenschaften einer σ-Algebra:

1.) Es gilt $\emptyset \in A(\Omega)$, d.h. die erste Eigenschaft ist erfüllt.
2.) $\emptyset^c = \Omega \in A(\Omega),\ A^c \in A(\Omega),\ (A^c)^c = A \in A(\Omega),\ \Omega^c = \emptyset \in A(\Omega)$, d.h. die zweite Eigenschaft ist erfüllt.
3.) Da $A(\Omega)$ lediglich vier Elemente enthält, können wir uns auf den Nachweis der endlichen Additivität beschränken.
Es gilt für beliebige $C_1,\ C_2\ \epsilon\ A(\Omega)$, dass $C := C_1 \cup C_2\ \epsilon\ A(\Omega)$, folglich ist ebenfalls die dritte Eigenschaft erfüllt.

Fazit: $A(\Omega) = \{\emptyset,\ A,\ A^C,\ \Omega\}$ ist eine σ-Algebra.

alternativer Erklärungsansatz:

Ist $A(\Omega) \neq \{\emptyset,\ \Omega\}$, so enthält $A(\Omega)$ eine einelementige Teilmenge, o.B.d.A. $\{a\}$ und damit auch das Komplement $\{b, c\}$.

Ist weiter $(\{a\}\epsilon A(\Omega))$ $A(\Omega) \neq \{\emptyset, \{a\}, \{b, c\}, \Omega = \{a, b, c\}\}$, so enthält $A(\Omega)$ mindestens eine der einelementigen Mengen $\{b\}$ oder $\{c\}$ oder eine weitere zweielementige Teilmenge, in der dann aber a als Element enthalten ist, also $\{a, b\}$ oder $\{a, c\}$. Im zweiten Fall ist aber $\{a, b\} \cap \{b, c\} = \{b\}$ oder $\{a, c\} \cap \{b, c\} = \{c\}$, also ist auch eine der einelementigen Teilmengen $\{b\}$ oder $\{c\}$ Element von $A(\Omega)$. O.B.d.A. sei also $\{b\}\epsilon A(\Omega)$, dann ist aber $\{a\} \cup \{b\} = \{a, b\}\ \epsilon\ A(\Omega)$ und damit auch $\{a, b\}^c = \{c\}\ \epsilon\ A(\Omega)$. $A(\Omega)$ enthält also alle einelementigen Teilmengen und damit jede Teilmenge: $A(\Omega) = \mathcal{P}(\Omega)$.

Aufgabe 42

Betrachten Sie eine vereinfachte Variante eines Roulette-Spiels für einen Spieler mit den folgenden Spielregeln: In jeder Runde kann der Spieler auf eine der Zahlen $1, \ldots, 36$ (nicht aber auf die 0!) setzen. Er gewinnt, falls die von ihm gesetzte Zahl mit der durch das Roulette-Rad ausgespielten übereinstimmt.

(a) Beschreiben Sie den Wahrscheinlichkeitsraum.

(b) Sei X die erste und Y die zweite Ziffer der gespielten Zahl (für die Zahlen $0, \ldots, 9$ wird X auf 0 gesetzt). Berechnen Sie die Wahrscheinlichkeit dafür, dass eine Zahl

- aus $\{1, \ldots, 36\}$,
- mit $X = 3$,
- mit $X = 2$ oder $X = 3$,
- aus $\{25, \ldots, 28\}$,
- mit $X \geq 2$ und $Y \leq 4$ bzw.
- mit $X + Y = 4$

gewinnt.

Lösung:(Wahrscheinlichkeitstheorie, S. 17ff)

1.) Wahrscheinlichkeitsraum: $(\Omega,\ A(\Omega), P)$

mit Ω : Grundgesamtheit

$A(\Omega)$: $\sigma-$Algebra (Menge der Teilmengen von Ω (Ereignisse), denen über P eine Wahrscheinlichkeit zugeordnet wird)

P : Wahrscheinlichkeitsmaß

Wir wählen in diesem Fall:

- Ω= {0,1,2,...,36}
- $A(\Omega)= \mathcal{P}(\Omega)$ (Potenzmenge)
- $P : \mathcal{P}(\Omega) \to [0,1]$ mit $A \in \mathcal{P}(\Omega) : P(A) = \frac{\#A}{\#\Omega}$ (Laplace'scher Wahrscheinlichkeitsraum)

2.) Es gelten folgende Rechenregeln:

- Für alle $A, B \in A(\Omega)$ mit $A \subset B$ gilt: $P(B \backslash A) = P(B) - P(A)$ und $P(A) \leq P(B)$
- Insbesondere gilt: $P(\Omega \backslash A) = 1 - P(A)$
- $P(\emptyset) \leq P(A) \leq P(\Omega)\ \forall A \in A(\Omega)$, wobei $P(\emptyset) = 0, P(\Omega) = 1$
- für $A_i \in A(\Omega)$, $i = 1, 2, \ldots,$ mit $A_i \cap A_j = \emptyset$, $i \neq j$, gilt:
  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

In unserem Fall gilt also:

- Menge der Zahlen ungleich null:

$$\begin{aligned} P(\{1,\ldots,36\}) &= P(\Omega\backslash\{0\}) \quad \text{oder } P(\{1,\ldots,36\}) = P(\{1\}\cup\ldots\cup\{36\}) \\ &= \tfrac{36}{37} \qquad\qquad\qquad\qquad\quad = P(\{1\})+\ldots+P(\{36\}) \\ &\qquad\qquad\qquad\qquad\qquad\quad = \tfrac{36}{37}) \end{aligned}$$

- Menge der Zahlen mit $X = 3 : \{30, 31, \ldots, 36\}$

$$\begin{aligned} P(\{30,31,\ldots,36\}) &= P(\{30\})+P(\{31\})+\ldots+P(\{36\}) \\ &= \tfrac{7}{37} \end{aligned}$$

- Menge der Zahlen mit $X = 2$ oder $X = 3$: {20, 21,22,. . . ,36}

$$\begin{aligned} P(\{20,21,\ldots,36\}) &= P(\{20,\ldots,29\}\cup\{30,\ldots,36\}) \\ &= P(\{20\})+\ldots+P(\{29\})+P(\{30,\ldots,36\}) \\ &= \tfrac{17}{37} \end{aligned}$$

- Menge der Zahlen von 25 bis 28:

$$\begin{aligned} P(\{25,\ldots,28\}) &= P(\{25\})+\ldots+P(\{28\}) \\ &= \tfrac{4}{37} \end{aligned}$$

- Menge der Zahlen mit $X \geq 2$ und $Y \leq 4 : \{20, 21, 22, 23, 24, 30, 31, 32, 33, 34\}$

$$P(\{20,\ldots,24,30,\ldots,34\}) = \tfrac{10}{37}$$

- Menge der Zahlen mit $X + Y = 4 : \{4, 13, 22, 31\}$

$$P(\{4,13,22,31\}) = \tfrac{4}{37}$$

Aufgabe 43

Bei zahlreichen Brettspielen (u.a. „Mensch ärgere dich nicht“) gilt folgende Vereinbarung für das Würfeln:

Würfelt ein Spieler eine der Zahlen 1 bis 5, so kommt anschließend der nächste Spieler dran. Würfelt er hingegen eine 6, so darf er noch mal würfeln (so lange, bis er keine 6 mehr würfelt).

Formulieren Sie den zugehörigen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}(\Omega), P)$.

Lösung:(Wahrscheinlichkeitstheorie, S. 17ff)

Wir werden den Wahrscheinlichkeitsraum Ω als eine abzählbare disjunkte Vereinigung von Mengen Ω_i für $i = 0, 1, 2, \ldots$ modellieren, wobei Ω_i diejenigen Elementarereignisse enthält, in denen der Spieler i-mal eine “6” würfelt bevor er eine Zahl ungleich 6 würfelt und somit sein Spielzug beendet ist. Sei also

$$\Omega := \bigcup_{i=0}^{\infty} \Omega_i \text{ mit } \Omega_i = \{\underbrace{66\ldots6}_{i-\text{mal}}1,\ \underbrace{66\ldots6}_{i-\text{mal}}2, \ldots,\ \underbrace{66\ldots6}_{i-\text{mal}}5\}$$

Da Ω eine abzählbare Menge ist, wählen wir die größtmögliche σ-Algebra nämlich die Potenzmenge von Ω:

$$A(\Omega) := \mathcal{P}(\Omega)$$

Im Falle eines endlichen oder abzählbar unendlichen Grundraums mit der Potenzmenge als σ-Algebra ist es üblich, das Wahrscheinlichkeitsmaß auf jedem beliebigen Elementarereignis $\omega \in \Omega$ zu definieren. Für eine beliebige Menge $A \in A(\Omega)$ ergibt sich dann $P(A)$ als Summe über die Wahrscheinlichkeiten, der in A enthaltenen Elementarereignisse:

$$P(A) := \sum_{\omega \in A} P(\omega)$$

Sei also $\omega \in \Omega$ beliebig. Dann existiert genau ein $i \in \mathbb{N}_0$ mit $\omega \in \Omega_i$. Wir setzen:

$$P(\omega) := (\frac{1}{6})^i \cdot \frac{1}{6}$$

Offensichtlich gilt:

1.) $P(\omega) \geq 0$ für alle $\omega \in \Omega$

2.)
$$\sum_{\omega \in \Omega} P(\omega) = \sum_{i=0}^{\infty} \left(\sum_{\omega \in \Omega_i} P(\omega) \right) = \sum_{i=0}^{\infty} |\Omega_i| \cdot \left(\frac{1}{6}\right)^i \cdot \frac{1}{6}$$
$$= \sum_{i=0}^{\infty} 5 \cdot \left(\frac{1}{6}\right)^i \cdot \frac{1}{6} = \frac{5}{6} \cdot \sum_{i=0}^{\infty} \left(\frac{1}{6}\right)^i$$
$$= \frac{5}{6} \cdot \frac{1}{1 - \frac{1}{6}} = 1.$$

Fazit: P definiert ein Wahrscheinlichkeitsmaß auf Ω und $(\Omega,\ A(\Omega),\ P)$ ist ein Wahrscheinlichkeitsraum.

Aufgabe 44

Gegeben ist die folgende Summenhäufigkeitstabelle für die Einkommensverteilung einer Gruppe von Berufstätigen (Einkommen in [k€]):

rechte Klassengrenzen	10	20	25	30	40	60	85
Summenhäufigkeit (in [%])	5	20	40	65	85	90	100

Die rechte Klassengrenze gehöre dabei jeweils zur Klasse.

(a) Bestimmen und zeichnen Sie die Summenhäufigkeitsfunktion. Erstellen Sie ein Histogramm.

(b) Definieren Sie auf Grundlage der Summenhäufigkeitsfunktion ein Wahrscheinlichkeitsmaß P für die Höhe des Einkommens einer zufällig aus der Gruppe der Berufstätigen herausgegriffenen Person.

(c) Bestimmen Sie mit Hilfe von P die Wahrscheinlichkeiten für folgende Ereignisse: Das Einkommen einer zufällig herausgegriffenen Person ist:

- in dem Intervall $(20; 30]$.
- in dem Intervall $[20; 30)$.
- gleich 55.
- mindestens 75.
- geringer als 25 oder größer als 54.
- in sämtlichen Intervallen A_i der Form $A_i := [20 + \frac{5}{i}; 30 + \frac{15}{i}]$, $i \in \mathbb{N}$.

Veranschaulichen Sie sich Ihre Lösungen anhand des Histogramms sowie des Graphen der Summenhäufigkeitsfunktion.

Lösung: (Wahrscheinlichkeitstheorie, S. 11ff, 17ff)

(a)

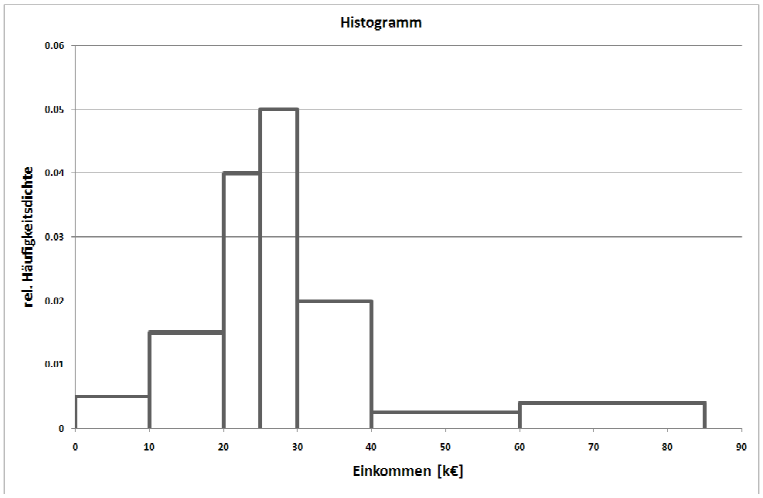


Abbildung 12: Histogramm

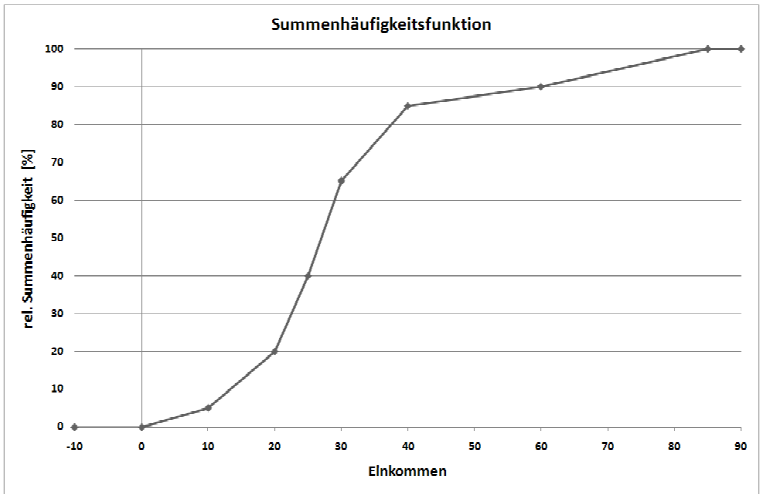


Abbildung 13: Summenhäufigkeitsfunktion

(b) Im Rahmen der Interpretation einer relativen Häufigkeit als Wahrscheinlichkeit ist $SF(\alpha)$ somit Näherung für $P((-\infty;\alpha])$, d.h. für die Wahrscheinlichkeit, einen Merkmalswert aus $(-\infty;\alpha]$ zu erhalten. Es gilt:

Eine Summenhäufigkeitsfunktion erfüllt die folgenden Eigenschaften:

- SF wächst monoton
- SF ist rechtsstetig
- $\lim\limits_{z\to-\infty} SF(z) = 0, \quad \lim\limits_{z\to+\infty} SF(z) = 1$

Dies sind genau diejenigen Eigenschaften, die eine Verteilungsfunktion charakterisieren, somit kann die Summenhäufigkeitsfunktion der empirischen Einkommensverteilung als Verteilungsfunktion der zufälligen Einkommensverteilung eines zufällig ausgewählten Berufstätigen betrachtet werden. Als Wahrscheinlichkeitsraum kann man nun wählen:

1) $\Omega = \mathbb{R}$
2) $A(\Omega) = \mathcal{B}(\mathbb{R})$
3) $P((-\infty,\alpha]) := SF(\alpha)$

Da die Mengen I der Form $I = (-\infty,\alpha]$ mit $\alpha \in \mathbb{R}$ einen Erzeuger der Borelschen σ-Algebra bilden, ist das Wahrscheinlichkeitsmaß P durch 3) eindeutig auf $\mathcal{B}(\mathbb{R})$ definiert.

(c)
- $P((20;30]) = P([20;30)) = SF(30) - SF(20) = 0,45$
- $P\{55\} = 0$
- $P([75,\infty)) = 1 - P((-\infty,75)) = 1 - P((-\infty,75]) = 1 - 0,96 = 0,04$
- $P((-\infty,25] \cup [54,\infty)) = P((0;25]) + P((54;\infty))$
  $= SF(25) + 1 - SF(54) = 0,52$
- $\bigcap\limits_{i=1}^{\infty} A_i = [25;30]: \quad P((25;30)) = SF(30) - SF(25) = 0,25$

Bemerkung: A_{i+1} ist nicht in A_i enthalten, folglich gilt auch i.A.

$$P(\bigcap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i) \quad \text{nicht,}$$

wie man bei der Aufgabe leicht sieht.

Aufgabe 45

Ein Händler will zu Silvester 25 Feuerwerkskörper, die ihm aus früheren Jahren übriggeblieben sind, loswerden. Er verspricht einem daran Interessierten, daß mindestens 60 % davon noch funktionsfähig sind. Dieser verlangt, 5 der 25 Feuerwerkskörper sofort ausprobieren zu dürfen, und ist bereit, die restlichen 20 zu kaufen, wenn mindestens 3 der 5 geprüften funktionieren. Der Händler ist mit einem Test einverstanden, will jedoch nur 3 Feuerwerkskörper dafür zur Verfügung stellen.

(a) Stellen Sie den Wahrscheinlichkeitsraum für den Test von 3 Feuerwerkskörpern auf.

(b) Sei X eine Zufallsvariable, die die Anzahl der funktionsfähigen Feuerwerkskörper in der Stichprobe angibt. Bestimmen Sie die Wahrscheinlichkeitsverteilung von X.

(c) Wie groß ist die Wahrscheinlichkeit, daß das Geschäft zustande kommt, wenn der Händler sich auf den Test von 5 Feuerwerkskörpern einläßt und tatsächlich

- 60 %
- 80 %
- 20 %

der 25 Feuerwerkskörper noch funktionsfähig sind?

Lösung: (Wahrscheinlichkeitstheorie, S. 17ff, 33ff)

(a)

- Wahrscheinlichkeitsraum $(\Omega,\ A(\Omega),\ P)$
- Jeder Feuerwerkskörper ist entweder defekt (d) oder funktionsfähig (f).

$$\Rightarrow \Omega = \{ddd, ddf, dfd, fdd, dff, fdf, ffd, fff\}$$
$$= \{\omega_1, \omega_2, \ldots, \omega_8\}$$

- $A(\Omega) = \mathcal{P}(\Omega)$
- $P:\ A(\Omega) \to [0;1]$ mit Ziehen ohne Zurücklegen [7]

$$P(\omega_1) = \tfrac{D}{N} \cdot \tfrac{D-1}{N-1} \cdot \tfrac{D-2}{N-2}$$
$$P(\omega_2) = P(\omega_3) = P(\omega_4) = \tfrac{D}{N} \cdot \tfrac{D-1}{N-1} \cdot \tfrac{F}{N-2}$$
$$P(\omega_5) = P(\omega_6) = P(\omega_7) = \tfrac{D}{N} \cdot \tfrac{F}{N-1} \cdot \tfrac{F-1}{N-2}$$
$$P(\omega_8) = \tfrac{F}{N} \cdot \tfrac{F-1}{N-1} \cdot \tfrac{F-2}{N-2}$$

[7]
D : Anzahl defekter Feuerwerkskörper
F : Anzahl funktionsfähiger Feuerwerkskörper
$N = D + F$: Anzahl Feuerwerkskörper insgesamt
Die exakte Schreibweise für $P(\omega_1)$ lautet $P(\{\omega_1\})$ usw.

(b)

$X : (\Omega,\ A(\Omega),\ P) \to \{0,1,2,3\}$
X gibt die Anzahl funktionsfähiger Feuerwerkskörper in der Stichprobe an, d.h.
$X(\omega_1) = 0,\ X(\omega_2) = X(\omega_3) = X(\omega_4) = 1,\ X(\omega_5) = X(\omega_6) = X(\omega_7) = 2,\ X(\omega_8) = 3.$

$$\begin{aligned} P_X(\{0\}) &= P(X^{-1}(0)) \\ &= P(\{\omega_i | X(\omega_i) = 0\}) \\ &= P(\omega_1) \end{aligned}$$

$$\begin{aligned} P_X(\{1\}) &= P(\{\omega_2, \omega_3, \omega_4\}) \\ &= P(\omega_2) + P(\omega_3) + P(\omega_4) \\ &= 3 \cdot \tfrac{D(D-1)\cdot F}{N(N-1)(N-2)} \\ &= \tfrac{\binom{D}{2}\binom{F}{1}}{\binom{N}{3}} \end{aligned}$$

$$P_X(\{2\}) = P(\{\omega_5, \omega_6, \omega_7\}) = P(\omega_5) + P(\omega_6) + P(\omega_7) = 3 \cdot \tfrac{D}{N} \cdot \tfrac{F}{N-1} \cdot \tfrac{F-1}{N-2} \cdots = \tfrac{\binom{D}{1}\binom{F}{2}}{\binom{N}{3}}$$

$$P_X(\{3\}) = P(\omega_8) = \tfrac{\binom{D}{0}\binom{F}{3}}{\binom{N}{3}}$$

(vgl. hypergeometrische Verteilung)

(c)

- Das Geschäft kommt zustande, falls $X \geq 3$, d.h.
  $P(\text{Geschäft}) = P(X = 3) + P(X = 4) + P(X = 5) = P(X \geq 3)$
  Für $N = 25,\ F = p \cdot 25,\ D = (1-p) \cdot 25$ ($p = 0,6;\ 0,8$ bzw. $0,2$) gilt:
  $P(\text{Geschäft}) = \tfrac{\binom{F}{3}\binom{D}{2}}{\binom{N}{5}} + \tfrac{\binom{F}{4}\binom{D}{1}}{\binom{N}{5}} + \tfrac{\binom{F}{5}\binom{D}{0}}{\binom{N}{5}}$

- $p = 0.6$:

$$P(X \geq 3) = \frac{\left[\binom{15}{3}\binom{10}{2} + \binom{15}{4}\binom{10}{1} + \binom{15}{5}\binom{10}{0}\right]}{\binom{25}{5}} = 0.6988$$

- $p = 0.8$:

$$P(X \geq 3) = \frac{\left[\binom{20}{3}\binom{5}{2} + \binom{20}{4}\binom{5}{1} + \binom{20}{5}\binom{5}{0}\right]}{\binom{25}{5}} = 0.9623$$

- $p = 0.2$:

$$P(X \geq 3) = \frac{\left[\binom{5}{3}\binom{20}{2} + \binom{5}{4}\binom{20}{1} + \binom{5}{5}\binom{20}{0}\right]}{\binom{25}{5}} = 0.0377$$

Aufgabe 46

Im Wareneingang einer Unternehmung werden Transistoren auf ihre Funktionsfähigkeit hin untersucht. Bei einer Warenpartie von $N = 100$ Teilen wird eine Stichprobe vom Umfang $n = 10$ gezogen. Aus langjähriger Erfahrung weiss man, dass im Mittel 3% der Transistoren fehlerhaft sind. Die Warenpartie wird abgelehnt, wenn mindestens 1 Transistor in der Stichprobe defekt ist.

(a) Berechnen Sie die Wahrscheinlichkeit, dass die Warenpartie abgelehnt wird, wenn die Stichprobe ohne Zurücklegen gezogen wird.

(b) Ein Mitarbeiter schlägt vor, die Stichprobe mit Zurücklegen zu ziehen, weil dies den Rechenaufwand vermindere. Berechnen Sie für diesen Fall die Wahrscheinlichkeit, die Warenpartie abzulehnen, und vergleichen Sie das Ergebnis mit dem aus (a).

(c) Ein weiterer Mitarbeiter schlägt vor, den Umfang der Warenpartie auf $N = 1000$ zu erhöhen. Überprüfen Sie, ob für diese Warenpartie der Unterschied zwischen Ziehen ohne Zurücklegen und mit Zurücklegen ins Gewicht fällt.

Lösung: (Wahrscheinlichkeitstheorie, S. 47ff)

(a) $N = 100, \quad n = 10, \quad M = p \cdot N = 3.$

Die Zufallsvariable Z gebe die Anzahl defekter Teile in der Stichprobe an. Bei Stichproben ohne Zurücklegen ist die Zufallsvariable Z hypergeometrisch verteilt. Die Wahrscheinlichkeit für das Ablehnen der Warenpartie beträgt

$$P(Z \geq 1) = 1 - P(Z = 0) = 1 - \frac{\binom{3}{0} \cdot \binom{97}{10}}{\binom{100}{10}} = 0.27347.$$

(b) Bei Ziehen mit Zurücklegen ist die Zufallsvariable Z binomialverteilt. Die Wahrscheinlichkeit für das Ablehnen der Warenpartie beträgt dann

$$P(Z \geq 1) = 1 - P(Z = 0) = 1 - \binom{10}{0} \cdot 0.03^0 \cdot 0.97^{10} = 0.26257,$$

d.h. bei Ziehen mit Zurücklegen wird die Warenpartie mit einer um rund 1%-Punkt niedrigeren Wahrscheinlichkeit abgelehnt als bei Ziehen ohne Zurücklegen.

(c) Die Erhöhung des Umfangs der Warenpartie hat bei Ziehen mit Zurücklegen keinen Einfluss auf die Ablehnungswahrscheinlichkeit. Bei Ziehen ohne Zurücklegen ergibt sich mit dem erhöhten Umfang folgende Ablehnungswahrscheinlichkeit:

$$M' = p \cdot N' = 30.$$

$$P(Z \geq 1) = 1 - P(Z = 0) = 1 - \frac{\binom{30}{0}\binom{970}{10}}{\binom{1000}{10}} = 0.26361.$$

Der Unterschied der Ablehnungswahrscheinlichkeiten zwischen Ziehen mit und ohne Zurücklegen reduziert sich auf rund 0.1 Prozentpunkte.

Aufgabe 47

(a) Für welche Parameter $a, b \in \mathbb{R}$ sind auf $\Omega = \{0, 1, \ldots, k, \ldots, n\}$ die Werte

$$\alpha_k = \binom{n}{k} a^k b^{n-k}$$

für ein Wahrscheinlichkeitsmaß brauchbar? Interpretieren Sie dieses Wahrscheinlichkeitsmaß im Hinblick auf praktische Anwendungen. Skizzieren Sie die zugehörige Verteilungsfunktion.

(b) Geben Sie Bedingungen für die Parameter $p_1, \ldots, p_m \in \mathbb{R}$ an, so dass auf der Menge Ω mit

$$\Omega \;=\; \{(k_1, \ldots, k_m) \,|\, k_i \in \{0, \ldots, n\}, \sum_{i=1}^{m} k_i = n\}$$

durch die Werte

$$\alpha_{k_1,\ldots,k_m} \;=\; \binom{n}{k_1 \ldots k_m} p_1^{k_1} \cdots p_m^{k_m}$$

ein Wahrscheinlichkeitsmaß definiert wird. Interpretieren Sie dieses Wahrscheinlichkeitsmaß im Hinblick auf praktische Anwendungen.

(c) Leiten Sie aus der geometrischen Reihe mit den Gliedern $(\frac{1}{2})^n$ eine Wahrscheinlichkeitsverteilung auf $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ ab. Verallgemeinern Sie Ihr Ergebnis für die Glieder a^n.

Lösung: (Wahrscheinlichkeitstheorie S.17ff und S.47ff)

(a)

Gegeben: $\alpha_k \;=\; \binom{n}{k}\; a^k\; b^{n-k}$ mit $a, b \,\epsilon\, \mathbb{R}$ und $k \;<\; n \,\epsilon\, \mathbb{N}$

Bedingungen an α_k, damit die $\alpha_k,\; k \;=\; 0,\; 1,\; \ldots\;,\; n$ eine Wahrscheinlichkeitsverteilung auf Ω erzeugen, sind:

1.) $\alpha_k \;\geq\; 0$ für alle $k \,\epsilon\, \{0,\; 1,\; \ldots\;,\; n\}$

2.) $\sum\limits_{k=0}^{n} \alpha_k \;=\; 1$

Es gilt:

1.) $\alpha_k \;=\; \underbrace{\binom{n}{k}}_{>\,0}\; \underbrace{a^k}_{\epsilon\,\mathbb{R}}\; \underbrace{b^{n-k}}_{\epsilon\,\mathbb{R}} \;\geq\; 0 \iff \begin{cases} a \;\geq\; 0,\; b \;\geq\; 0 \text{ für } n \text{ ungerade} \\ a \cdot b \;\geq\; 0 \text{ für } n \text{ gerade} \end{cases}$ [8]

[8] n ungerade: k gerade $\Leftrightarrow$ $n-k$ ungerade, also $a^k b^{n-k} \geq 0$ für alle k genau dann, wenn $a \geq 0$ und $b \geq 0$ ist.
n gerade: k gerade $\Leftrightarrow$ $n-k$ gerade, also $a^k b^{n-k} \geq 0$ für alle k genau dann, wenn $a \geq 0$ und $b \geq 0$ oder $a < 0$ und $b < 0$ ist.

2.) $\sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k} \;=\; (a \,+\, b)^n \;=\; 1 \iff \begin{cases} a \,+\, b \;=\; 1 \text{ für } n \text{ ungerade} \\ a \,+\, b \;=\; \pm 1 \text{ für } n \text{ gerade} \end{cases}$

Fazit: Die α_k erzeugen eine Wahrscheinlichkeitsverteilung auf Ω falls gilt:

a) $a \;\geq\; 0,\; b \;\geq\; 0,\; a \,+\, b \;=\; 1$ für n ungerade

b) $a \cdot b \;\geq\; 0,\; a \,+\, b \;=\; \pm 1$ für n gerade

Anmerkung: Da im Fall b) die α_k unabhängig vom gemeinsamen Vorzeichen von a, b sind, wählt man im allgemeinen $a, b \;\geq\; 0$ und somit $a \,+\, b \;=\; 1$. Die Wahrscheinlichkeitsverteilung entspricht der Binomialverteilung $B(n, p)$ mit der Trefferwahrscheinlichkeit $p \;=\; a$.

(b) Gegeben: $\alpha_{k_1,\ldots,k_m} \;=\; \binom{n}{k_1 \ldots k_m}\; p_1^{k_1} \cdots p_m^{k_m}$ mit $p_1, \ldots, p_m \;\epsilon\; \mathbb{R}$ und $(k_1, \ldots, k_m) \;\epsilon\; \{0,\; 1, \ldots, n\}^m$

Die Bedingungen an $\alpha_{k_1,\ldots,k_m}$ damit die $\alpha_{k_1,\ldots,k_m}$ eine Wahrscheinlichkeitsverteilung auf $\{0,\; 1, \ldots, n\}^m$ erzeugen, lauten:

1.) $\alpha_{k_1,\ldots,k_m} \;\geq\; 0$ für alle $(k_1, \ldots, k_m) \;\epsilon\; \{0,\; 1,\; \ldots\; ,\; n\}^m$

2.) $\sum\limits_{\substack{(k_1,\ldots,k_m) \;\epsilon\; \{0,\; 1,\; \ldots\; ,\; n\}^m \\ mit \sum k_j \;=\; n}} \alpha_k \;=\; 1$

Es gilt:

1.) $\alpha_{k_1,\ldots,k_m} \;=\; \underbrace{\binom{n}{k_1 \ldots k_m}}_{>\,0}\; \underbrace{p_1^{k_1}}_{\epsilon\; \mathbb{R}} \cdots \underbrace{p_m^{k_m}}_{\epsilon\; \mathbb{R}} \;\geq\; 0,$ falls $p_1, \ldots, p_m \;\geq\; 0$ gilt.

2.) $\sum\limits_{\substack{(k_1,\ldots,k_m) \;\epsilon\; \{0,1,\ldots,n\}^m \\ mit \sum k_j \;=\; n}} \alpha_k \;=\; (p_1 \;+\ldots+\; p_m)^n \;=\; 1$, falls $\sum\limits_{i=1}^{m} p_i \;=\; 1$ gilt.

Fazit: Die $\alpha_{k_1,\ldots,k_m}$ erzeugen eine Wahrscheinlichkeitsverteilung auf $\{0,\; 1, \ldots, n\}^m$ falls $p_1, \ldots, p_m \;\geq\; 0$ und $\sum\limits_{i=1}^{m} p_i \;=\; 1$.

Die Wahrscheinlichkeitsverteilung entspricht der Multinomialverteilung. Die Multinomialverteilung beschreibt die Wahrscheinlichkeitsverteilung die entsteht, wenn man n Kugeln auf m Urnen aufteilt, wobei die Urne i $(i \;=\; 1, \ldots, m)$ mit der Wahrscheinlichkeit p_i ausgewählt wird. $\alpha_{k_1,\ldots,k_m}$ entspricht der Wahrscheinlichkeit in Urne i $(i \;=\; 1, \ldots, m)$ k_i Kugeln zu beobachten.

(c)

1.) Sei $\Omega \;=\; \mathbb{N}_0,\;\; \alpha_n \;=\; c \;\cdot\; (\frac{1}{2})^n$.

Die α_n erzeugen eine Wahrscheinlichkeitsverteilung auf Ω falls gilt:

1.) $\alpha_n \;\geq\; 0$ für alle $n \;\epsilon\; \mathbb{N}_0,$ also genau dann, wenn $c \;\geq\; 0$ ist.

2.) $\sum\limits_{n \;\epsilon\; \mathbb{N}_0} \alpha_n \;=\; 1,$ also wenn $c \;\cdot\; \frac{1}{1-\frac{1}{2}} \;=\; 1,$ d.h. $c \;=\; \frac{1}{2}$ ist.

Fazit: $\alpha_n \;= (\frac{1}{2})^{n+1}$ mit $n \;\epsilon\; \mathbb{N}_0$ ist eine Wahrscheinlichkeitsverteilung auf $\Omega \;=\; \mathbb{N}_0$.

2.) $\Omega \;=\; \mathbb{N}_0,\;\; \alpha_n \;=\; c \;\cdot\; a^n.$

Die α_n erzeugen eine Wahrscheinlichkeitsverteilung auf Ω falls gilt:

1.) $\alpha_n \geq 0$ für alle $n \epsilon \mathbb{N}_0$, also genau dann, wenn $a,\ c \geq 0$

2.) $\sum\limits_{n \epsilon \mathbb{N}_0} \alpha_n = 1$ also wenn $c \cdot \frac{1}{1-a} = 1$, d.h. $c = 1 - a$ ist.

Fazit: $\alpha_n = (1 - a) \cdot a^n$ mit $n \epsilon \mathbb{N}_0$ definieren eine Wahrscheinlichkeitsverteilung auf $\Omega = \mathbb{N}_0$, diese Verteilung heißt geometrische Verteilung mit Parameter $a \in (0, 1)$.

Aufgabe 48

$F(x_i) = 1 - 2^{-x_i}$ für $x_i = 1, 2, \ldots$ gebe die Werte der Verteilungsfunktion einer diskreten Zufallsvariable an sämtlichen Sprungstellen an. Geben Sie die Verteilungsfunktion auf $\mathbb{R}$ an. Stellen Sie fest, welche Werte die Zufallsvariable mit positiver Wahrscheinlichkeit annimmt und berechnen Sie die Wahrscheinlichkeitsverteilung der Zufallsvariable.

Lösung:

Da die Verteilungsfunktion einer diskreten Zufallsvariablen zwischen ihren Sprungstellen konstant ist, gilt[9]:

$$F(x) = \begin{cases} 1 - 2^{-\lfloor x \rfloor} & x \geq 1 \\ 0 & x < 1 \end{cases} = \begin{cases} 1 - 2^{-n} & n \leq x < n+1, n \in \mathbb{N} \\ 0 & x < 1 \end{cases}$$

Für die Wahrscheinlichkeitsverteilung einer Zufallsvariable X gilt:

$$P(X = x_0) = F_X(x_0) - \lim_{x \uparrow x_0} F_X(x)$$

Folglich nimmt die Zufallsvariable nur Werte mit positiver Wahrscheinlichkeit an, an denen die Verteilungsfunktion einen Sprung aufweist. Bei uns sind dies gerade die natürlichen Zahlen $n = 1, 2, 3, \ldots$. Konkret gilt für $n \in \mathbb{N}$

$$\begin{aligned} P(X = n) &= F_X(n) - \lim_{x \uparrow n} F_X(x) \\ &= 1 - 2^{-n} - (1 - 2^{-(n-1)}) \\ &= 2^{-n}. \end{aligned}$$

Die Wahrscheinlichkeitsverteilung lautet somit:

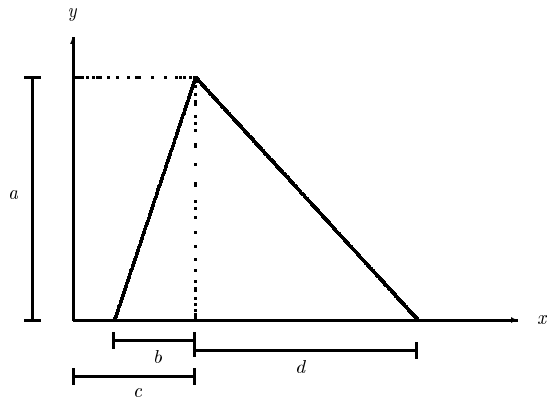
$$P(X = n) = (\frac{1}{2})^n.$$

Die Zufallszahlen $Y := X - 1$ folgt in diesem Fall einer geometrischen Verteilung mit Parameter $a = \frac{1}{2}$ (vergleiche Aufgabe 47c).

[9] [x] sei die größte ganze Zahl $\leq x$.

Aufgabe 49

(a) Bestimmen Sie für die folgende geometrische Figur zu den Parameter a, b und c den Parameter d derart, dass es sich um eine Dichtefunktion handelt. Berechnen Sie eine zugehörige Verteilungsfunktion. Ist diese eindeutig?



(b) Die Zufallsvariable Y habe die folgende Dichtefunktion:

$$f_Y(y) = \begin{cases} \alpha \cdot (1-(y-1)^2) & , y \in [0;1) \\ \alpha \cdot \sqrt{2-y} & , y \in [1;2] \\ \beta & , \text{sonst} \end{cases}$$

Bestimmen Sie die Konstanten α und β, und berechnen Sie die Verteilungsfunktion.

(c) Bestimmen Sie für die Zufallsvariable Y aus Aufgabenteil (b) den Median sowie den Erwartungswert.

Lösung: (Wahrscheinlichkeitstheorie, S. 57ff)

(a) **Satz:** Sei $f : \mathbb{R} \to \mathbb{R}$ eine bis auf endlich viele Stellen stetige Funktion mit:

1.) $f(x) \geq 0$ für alle $x \in \mathbb{R}$

2.) $\int_{-\infty}^{+\infty} f(x)dx = 1$

3.) Für alle $\alpha_0 \in \mathbb{R}$, für die $\lim_{\alpha \to \alpha_0} f(\alpha)$ existiert, folgt $f(\alpha_0) = \lim_{\alpha \to \alpha_0} f(\alpha)$.

Dann ist f Dichte einer Zufallsvariablen.

Bedingung 3 sagt, dass f keine unnötigen Unstetigkeitsstellen haben soll.

- hier
$$f(x) = \begin{cases} \frac{a}{b}(x-(c-b)) & x \in [c-b;c) \\ a - \frac{a}{d}(x-c) & x \in [c;c+d) \\ 0 & \text{sonst} \end{cases}$$
- erfüllt 1) und 3) des Satzes
$$F(z) = \int_{-\infty}^{z} f(x)dx = \begin{cases} 0 & x \in (-\infty; c-b] \\ \int_{c-b}^{z} \frac{a}{b}(x-c+b)dx & x \in (c-b;c] \\ F(c) + \int_{c}^{z} a - \frac{a}{d}(x-c)dx & x \in (c;c+d] \\ F(c+d) & x \in (c+d;\infty) \end{cases}$$
- Forderung 2:

$F(c+d) \stackrel{!}{=} 1$

$\int_{c-b}^{z} \frac{a}{b}(x-c+b)dx = \frac{a}{b}\left(\frac{1}{2}x^2 - cx + bx\right)\Big|_{c-b}^{z}$

$\Rightarrow F(c) = \frac{ab}{2}$ (s. Zeichnung)

$\int_{c}^{z} a - \frac{a}{d}(x-c)dx = \left[ax - \frac{a}{d}(\frac{1}{2}x^2 - cx)\right]\Big|_{c}^{z}$

$$\begin{aligned} \Rightarrow \quad F(c+d) &= F(c) + \left[ax - \tfrac{a}{d}\left(\tfrac{1}{2}x^2 - cx\right)\right]\Big|_{c}^{c+d} \\ (s.Zeichnung) \quad &= \tfrac{ab}{2} + \tfrac{ad}{2} \\ &\stackrel{!}{=} 1 \\ \Leftrightarrow \quad d &= \tfrac{2}{a} - b, \end{aligned}$$

d.h. f ist Dichtefunktion, wenn $d = \frac{2}{a} - b$ ist.[10]

(b) Aus Bedingung 1) im Satz der Teilaufgabe (a) folgt direkt: $\beta \geq 0,\ \alpha \geq 0$

Weiterhin muss gelten:

$$\begin{aligned} \int_{\mathbb{R}} f(y)dy &= 1 \\ &\Leftrightarrow \int_{-\infty}^{0} \beta\, dy + \int_{0}^{1} \alpha(1-(y-1)^2)dy + \int_{1}^{2} \alpha\sqrt{2-y}\, dy + \int_{2}^{\infty} \beta\, dy \\ &\stackrel{!}{=} 1 \\ &\Rightarrow \beta = 0, \text{ da sonst die Integrale } \int_{-\infty}^{0} \beta\, dy \text{ und } \int_{2}^{\infty} \beta\, dy \text{ divergieren.} \\ &\Rightarrow \int_{0}^{1} \alpha(1-(y-1)^2)dy + \int_{1}^{2} \alpha\sqrt{2-y}\, dy = 1 \\ &\Leftrightarrow \alpha\left[y^2 - \tfrac{1}{3}y^3\right]\Big|_{0}^{1} + \alpha\left[-\tfrac{2}{3}(2-y)^{\frac{3}{2}}\right]\Big|_{1}^{2} = \tfrac{2}{3}\alpha + \tfrac{2}{3}\alpha = \tfrac{4}{3}\alpha \stackrel{!}{=} 1 \end{aligned}$$

d.h. notwendige Bedingung ist: $\beta = 0, \alpha = \frac{3}{4}$.

[10] Für $d = b$ (Symmetrie) also $b = \frac{1}{a}$.

Für diese Werte ist die Funktion f_y stetig auf ganz $\mathbb{R}$ und folglich nach dem Satz aus Teilaufgabe a) eine Dichtefunktion.

$$f_Y(y) = \begin{cases} \frac{3}{4}(1-(y-1)^2) & y \in [0;1) \\ \frac{3}{4}\sqrt{2-y} & y \in [1;2) \\ 0 & \text{sonst} \end{cases}$$

Die Verteilungsfunktion lautet:

$$F_Y(y) = \begin{cases} 0 & y \in (-\infty;0) \\ \frac{3}{4}\left(y^2 - \frac{1}{3}y^3\right) & y \in [0;1) \\ 1 - \frac{1}{2}(2-y)^{\frac{3}{2}} & y \in [1;2) \\ 1 & y \in [2;\infty) \end{cases}$$

(c) x_z ist genau dann Median von X, falls

1) $F_X(x) \leq \frac{1}{2}$ für $x < x_z$
2) $F_X(x_z) \geq \frac{1}{2}$

ist.

Der Median ist eventuell nicht eindeutig definiert (falls es ein x^* mit $F_X(x^*) = \frac{1}{2}$ gibt, dann sind alle $x \in \mathbb{R}$ mit $F_X(x) = \frac{1}{2}$ sowie $\min\{x|F_X(x) > \frac{1}{2}\}$ Median).

Für stetige Zufallsvariablen X existiert immer mindestens ein $x \in \mathbb{R}$ mit $F_X(x) = \frac{1}{2}$ (F_X stetig mit Wertebereich [0;1]).

hier: $F_Y(y) = \frac{1}{2}$ für $y = 1$, d.h $y_z = 1$

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= \int_0^1 \frac{3}{4} y(1-(y-1)^2) dy + \int_1^2 \frac{3}{4} y\sqrt{2-y}\, dy \\ &= \frac{3}{4}\left[\frac{2}{3}y^3 - \frac{1}{4}y^4\right]\Big|_0^1 - \frac{1}{2}\left[y(2-y)^{\frac{3}{2}}\right]\Big|_1^2 - \frac{1}{5}(2-y)^{\frac{5}{2}}\Big|_1^2 \\ &= 1,0125 \end{aligned}$$

Die Berechnung des 2. Integrals erfolgt mit Hilfe partieller Integration.

Aufgabe 50

(a) Berechnen Sie für die folgende diskrete Verteilung den Modalwert, den Median, den Erwartungswert, die Varianz, das 2. Moment sowie das 0.85-Quantil.

x	-2	-1	0	1	2	3	4
$P(X=x)$	0.1	0.1	0.4	0.1	0.05	0.1	0.15

(b) Bestimmen Sie für die Pareto-Verteilung, die durch die Dichtefunktion

$$f_X(x) = \begin{cases} k\,x^{-n-1} & \text{für } x \geq c \\ 0 & \text{sonst} \end{cases}$$

charakterisiert wird $(c, n, k > 0)$ den Parameter k in Abhängigkeit von c und n, die Verteilungsfunktion sowie den Modalwert, den Erwartungswert, die Varianz und das 0.85-Quantil.

Lösung: (Wahrscheinlichkeitstheorie, S. 69ff)

(a) Modalwert einer diskreten Zufallsvariable: Wert der Zufallsvariable, der mit maximaler Wahrscheinlichkeit angenommen wird.

hier: $x_{\text{mod}} = 0$, da $P(X=0) = 0.4 \geq P(X=i) \quad i = -2, \cdots, 4$

- Median: x_z Median $\Leftrightarrow P(X \geq x_z) \geq \frac{1}{2}$ und $P(X \leq x_z) \geq \frac{1}{2}$

  Hilfssatz: x_z Median der ZV $X \Leftrightarrow$ 1) $F_X(x) \leq \frac{1}{2}$ für alle $x < x_z$
  2) $F_X(x_z) \geq \frac{1}{2}$

  Hier gilt $x_z = 0$, da $F_X(x) < 0.5$ für $x < 0$ und $F_X(0) = 0.6$.

- Erwartungswert einer diskreten Zufallsvariable:

  $E(X) = \sum_{i=1}^{\infty} \alpha_i \cdot P(X = \alpha_i)$, falls diese Reihe absolut konvergiert.

  hier: X hat endlichen Wertevorrat: $E(X)$ existiert

  $$\begin{aligned} E(X) &= (-2) \cdot 0,1 + (-1) \cdot 0,1 + 0 \cdot 0.4 + 1 \cdot 0.1 + 2 \cdot 0.05 + 3 \cdot 0.1 + 4 \cdot 0,15 \\ &= 0,8 \end{aligned}$$

- k-tes Moment:

  $E(X^k) = \sum_{i=1}^{\infty} \alpha_i^k P(X = \alpha_i)$ heißt k-tes Moment von X, falls $\sum_{i=1}^{\infty} |\alpha_i|^k P(X = \alpha_i)$ konvergiert.

  (Momente höherer Ordnung können zur genaueren Beschreibung von Verteilungen verwendet werden)

  hier : $E(X^2)$ existiert, da X einen endlichen Wertevorrat besitzt

  $$\begin{aligned} E(X^2) &= 4 \cdot 0,1 + 1 \cdot 0,1 + 0 \cdot 0.4 + 1 \cdot 0.1 + 4 \cdot 0.05 + 9 \cdot 0.1 + 16 \cdot 0,15 \\ &= 4,1 \end{aligned}$$

- Varianz: $Var(X) = \sum_{i=1}^{\infty} (\alpha_i - E(X))^2 \cdot P(X = \alpha_i)$, falls diese Reihe konvergiert.

hier:

$$\begin{aligned} Var(X) &= (-2-0,8)^2 \cdot 0,1 + (-1-0.8)^2 \cdot 0.1 + (0-0.8)^2 \cdot 0.4 + (1-0.8)^2 \cdot 0.1 \\ &\quad + (2-0.8)^2 \cdot (0.05+3-0.8)^2 \cdot 0.1 + (4-0,8)^2 \cdot 0,15 \\ &= 3,46 \end{aligned}$$

- α-Quantil:

  x_α ist α-Quantil von X, wenn 1.) $P(X \leq x_\alpha) = F(x_\alpha) \geq \alpha$

  2.) $P(X \geq x_\alpha) \geq 1 - \alpha$

  Analog zur Definition des Medians besitzt also die Menge der α-Quantile die Form: $x_\alpha = \{x \in \mathbb{R} \mid F(x) = \alpha \text{ oder } x = \min\{x \mid F(x) > \alpha\}\}$.

  Statt α-Quantil wird auch der Begriff $(1-\alpha)$-Fraktil verwendet.

  Hier: $\min\{x|F_X(x) \geq 0,85\} = 3$ und $F_X(3) = 0,85$: alle $x \in [3;4]$ sind 0,85-Quantile.

  (Analog zur Deskriptiven Statistik verwendet man oft das Intervallmittel als Quantilswert, d.h. hier $q_{0,85} = 3.5$.)

(b) $f_X(x)$ soll Dichtefunktion sein, d.h. u.a. muss $\int_{-\infty}^{\infty} f_X(x)dx = 1$ gelten.

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)dx &= \int_{c}^{\infty} kx^{-n-1}dx \\ &= k\left(-\frac{1}{n}\right)x^{-n}\Big|_c^\infty \\ &= -\frac{k}{n}(0-c^{-n}) \\ &= \frac{k}{n}c^{-n} \\ &\stackrel{!}{=} 1 \quad \Rightarrow k = nc^n \end{aligned}$$

$$\begin{aligned} F_X(z) &= \int_{-\infty}^{z} f_X(x)dx = 0 \quad für \quad z < c \\ &= \int_{c}^{z} nc^n x^{-n-1}dx \quad für \quad c \leq z \\ &= nc^n(-\frac{1}{n}x^{-n})\,|_c^z \\ &= -c^n z^{-n} + 1 \end{aligned}$$

- Modalwert: x_{mod} heißt Modalwert der Zufallsvariable X, wenn $f(x_{\text{mod}}) = \max_{x \in R} f(x)$

  hier: $\left.\begin{array}{l} f_X(x) = 0 \text{ auf } (-\infty; c) \\ f_X(x) \geq 0 \text{ und monoton fallend auf } [c, \infty) \end{array}\right\} \Rightarrow x_{\text{mod}} = c$

- Erwartungswert:

Sei X stetige Zufallsvariable mit Dichte f. Existiert $\int_{-\infty}^{\infty} |x| f(x)dx$, so heißt $E(X) = \int_{-\infty}^{\infty} x f(x)dx$ Erwartungswert von X.

$$\begin{aligned}
\text{hier:} \quad E(X) &= \int_{c}^{\infty} x\, n\, c^n x^{-n-1} dx \\
&= \int_{c}^{\infty} n\, c^n x^{-n} dx \\
&= nc^n \left(-\frac{1}{n-1}\, x^{-n+1} \right) \Big|_{c}^{\infty} \quad \text{, falls } n > 1 \\
&= \frac{nc^n}{n-1} \left(-\lim_{x\to\infty} x^{-n+1} + c^{-n+1} \right) \\
&= \frac{n}{n-1}\, c \qquad \text{(für } n \leq 1 \text{ existiert kein Erwartungswert)}
\end{aligned}$$

Bemerkung: Analog zu (1) können Momente höherer Ordnung durch

$E(X^k) = \int_{-\infty}^{\infty} x^k f(x)dx$ berechnet werden, falls das Integral absolut konvergiert.

Für $E(X^2)$ ergibt sich hier

$E(X^2) = \frac{n}{n-2}\, c^2$, falls $n > 2$ ist. Für $n \leq 2$ existiert das 2. Moment nicht.

- Varianz:
  X stetige Zufallsvariable mit Dichte f und Erwartungswert $E(X)$.
  Dann heißt $Var(X) = \int_R (x - E(X))^2 f(x)dx$ Varianz von X, falls das Integral existiert ($\sqrt{Var(X)}$ heißt Standardabweichung).
  hier: durch Ausnutzen von $Var(X) = E(X^2) - E^2(X)$ ergibt sich

  $Var(X) = c^2 \left(\frac{n}{n-2} - \left(\frac{n}{n-1} \right)^2 \right)$, falls $n > 2$ ist. Für $n \leq 2$ existiert die Varianz nicht.

- α-Quantil: X stetige ZV mit Verteilungsfunktion $F_X(x), 0 < \alpha < 1$. Dann heißen alle $x_\alpha \in \mathbb{R}$ mit $F_X(x_\alpha) = \alpha$ α-Quantil von X.

  hier: $F_X(x) = \begin{cases} 0, & x < c \\ 1 - c^n x^{-n}, & x \geq c \end{cases}$

  $\Rightarrow 0,85 \stackrel{!}{=} 1 - c^n x_{0,85}^{-n} \Leftrightarrow x_{0,85} = \frac{c}{|\sqrt[n]{0,15}|}$

  Bemerkung: Der Median von X ist analog $x_{0,5} = c\sqrt[n]{2}$

Aufgabe 51

Gegeben ist folgende Funktionenfamilie ($a > 0$):

$$f_k(x) = \begin{cases} k\,(a - |x|) & \text{für } -a < x < a \\ 0 & \text{sonst} \end{cases}$$

(a) Bestimmen Sie k in Abhängigkeit von a, so dass $f_k(x)$ eine Dichtefunktion wird. Skizzieren Sie diese.

(b) Bestimmen Sie die zugehörige Verteilungsfunktion.

Lösung: (Wahrscheinlichkeitstheorie, S. 57ff)

a) f^k stetig auf $\mathbb{R}$ für $a > 0, f^k(x) \geq 0$ für alle $x \in \mathbb{R}$.

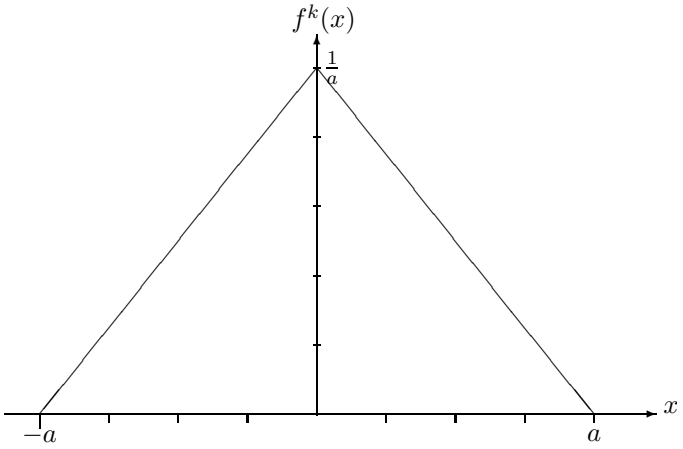
$$\begin{aligned} \int_{-\infty}^{\infty} f^k(x)dx &= \int_{-a}^{a} k(a - |x|)dx \\ &= \int_{-a}^{0} k(a + x)dx + \int_{0}^{a} k(a - x)dx \\ &= k(ax + \tfrac{1}{2}x^2)\Big|_{-a}^{0} + k(ax - \tfrac{1}{2}x^2)\Big|_{0}^{a} \\ &= ka^2 \\ &\stackrel{!}{=} 1 \qquad \Leftrightarrow k = \tfrac{1}{a^2} \end{aligned}$$

d.h.

$$f^k(x) = \begin{cases} \frac{1}{a^2}(a - |x|) & \text{für } |x| < a \\ 0 & \text{sonst} \end{cases}$$

ist Dichtefunktion (offensichtlich gilt $f(x) \geq 0$ und f ist stetig (s. Zeichnung)).

Somit ist auch die 3. Bedingung und natürlich auch die Bedingung, dass die Dichte nur endlich viele Unstetigkeitsstellen haben soll, erfüllt.



Dreiecksverteilung (vgl. Wahrscheinlichkeitstheorie, S. 62ff)

b) $F_X(x) = 0$ für $x \leq -a$ und $F_X(x) = 1$ für $x > a$.

Für $x \in (-a; a]$ gilt:

$x \leq 0$:

$$\begin{aligned} F_X(x) &= \int_{-a}^{x} \tfrac{1}{a^2}(a - |t|)dt \\ &= \int_{-a}^{x} \tfrac{1}{a^2}(a + t)dt \\ &= \tfrac{1}{a^2}\,(at + \tfrac{1}{2}t^2)\Big|_{-a}^{x} \\ &= \tfrac{x}{a}\left(1 + \tfrac{1}{2}\tfrac{x}{a}\right) + \tfrac{1}{2} \end{aligned}$$

$x > 0$:

$$\begin{aligned} F_X(x) &= \tfrac{1}{2} + \int_{0}^{x} \tfrac{1}{a^2}(a - t)dt \\ &= \tfrac{1}{2} + \tfrac{x}{a} - \tfrac{1}{2}\left(\tfrac{x}{a}\right)^2 \end{aligned}$$

Die Verteilungsfunktion lautet:

$$F_X(x) = \begin{cases} 0 & x \leq -a \\ \tfrac{1}{2} + \tfrac{x}{a}\left(1 + \tfrac{1}{2}\tfrac{x}{a}\right) & x \in (-a; 0] \\ \tfrac{1}{2} + \tfrac{x}{a}\left(1 - \tfrac{1}{2}\tfrac{x}{a}\right) & x \in (0; a] \\ 1 & x > a \end{cases}$$

Aufgabe 52

Gegeben ist die Dichtefunktion der Cauchy-Verteilung

$$f_X(x) \;=\; \frac{1}{\pi\,(1\;+\;x^2)},\; x\;\epsilon\;\mathbb{R}$$

(a) Untersuchen Sie, ob die Cauchy-Verteilung einen Erwartungswert besitzt. Was folgt aus Ihrem Ergebnis für die Varianz der Verteilung?

(b) Skizzieren Sie den Graphen der Dichtefunktion.

Lösung:Wahrscheinlichkeitstheorie, S. 69ff

(a) Es gilt: Der Erwartungswert $E(X)$ einer stetigen Verteilung mit Dichtefunktion $f(x)$ existiert genau dann, wenn gilt: $\int_{-\infty}^{\infty} |x|\; f(x)\,dx \;<\; \infty$. Überprüfen der Bedingung ergibt:

$$\begin{aligned}
\int_{-\infty}^{\infty} |x|\; f(x)\,dx \;&=\; \int_{-\infty}^{\infty} \frac{|x|}{\pi\,(1\;+\;x^2)}\,dx \\
&=\; \frac{1}{\pi}\int_{0}^{\infty} \frac{2x}{1\;+\;x^2}\,dx \;=\; \frac{1}{\pi}\lim_{N\to\infty}\int_{0}^{N} \frac{2x}{1\;+\;x^2}\,dx \\
&=\; \frac{1}{\pi}\lim_{N\to\infty}\left[\log\,(1\;+\;x^2)\right]_0^N \;=\; \frac{1}{\pi}\lim_{N\to\infty}\log\,(1\;+\;N^2) \\
&=\; \infty
\end{aligned}$$

Folglich existiert der Erwartungswert nicht und somit auch nicht die Varianz.

b)

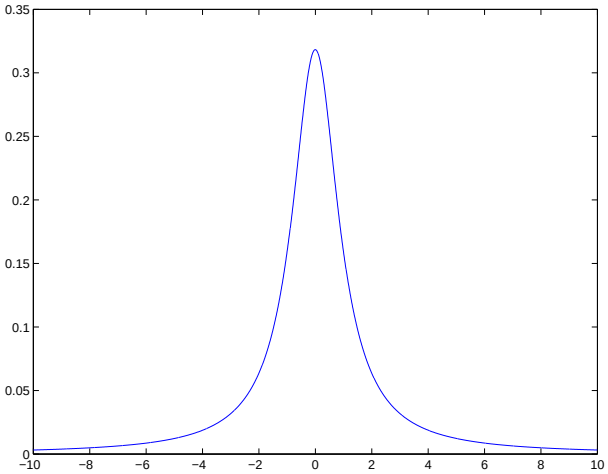


Abbildung 14: Dichtefunktion der Cauchyverteilung im Intervall $[-10, 10]$.

Aufgabe 53

(a) Geben Sie mit Hilfe einer Tabelle für die Standardnormalverteilung $N(0,1)$ an:

(i) $P((-\infty, 1.50))$, $P((-\infty, -2.05))$, $P((1.65, \infty))$, $P((-0.78, 1.75))$.

(ii) c mit $P([c, \infty)) = 0.85$.

(iii) c mit $P([-c, c]) = 0.95$.

(b) Berechnen Sie mit Hilfe der Funktionen "NORMVERT" und "NORMINV" aus Excel für die Normalverteilung $N(5,9)$:

(i) $P((-\infty, 2))$, $P((-\infty, 6))$, $P((2, 7))$, $P((4.5, \infty))$

(ii) c mit $P((-\infty, c]) = 0.84$.

(iii) c mit $P([5-c, 5+c]) = 0.57$.

Lösung: (Wahrscheinlichkeitstheorie, S. 66f, 107, 217f)

(a) Die Verteilungsfunktion Φ der Standardnormalverteilung (bzw. allg. der Normalverteilung) lässt sich nicht geschlossen darstellen. Die Werte der Verteilungsfunktion müssen mit Hilfe numerischer Methoden berechnet werden. Daher sind sie tabelliert für $z \geq 0$ und für $z \leq 0$ über $\Phi(-z) = 1 - \Phi(z)$ daraus berechenbar.

(i)
$$\begin{aligned} P((-\infty; 1,5)) &= \Phi(1,5) \\ &\approx 0{,}9332 \end{aligned}$$

$$\begin{aligned} P((-\infty; -2,05)) &= \Phi(-2,05) \\ &= 1 - \Phi(2,05) \\ &\approx 0{,}0202 \end{aligned}$$

$$\begin{aligned} P((1,65; \infty)) &= 1 - P((-\infty; 1,65]) \\ &= 1 - \Phi(1,65) \\ &\approx 0{,}0495 \end{aligned}$$

$$\begin{aligned} P((-0,78; 1,75)) &= \Phi(1,75) - \Phi(-0,78) \\ &= \Phi(1,75) + \Phi(0,78) - 1 \\ &\approx 0{,}7422 \end{aligned}$$

(ii)
$$\begin{aligned} 0{,}85 &\overset{!}{=} P([c; \infty)) \\ &= 1 - P((-\infty; c)) \\ &= 1 - \Phi(c) = \Phi(-c) \\ \Phi(-c) &= 0{,}85, \text{ d.h. } c \approx \text{-}1{,}04 \end{aligned}$$

(iii)
$$\begin{aligned} 0{,}95 &\overset{!}{=} P([-c; c]) \\ &= \Phi(c) - \Phi(-c) = \Phi(c) - (1 - \Phi(c)) \\ &= 2\Phi(c) - 1, \text{ d.h. } \Phi(c) = 1,95/2 \\ \Phi(c) &= 0{,}975, \text{ d.h. } c \approx 1{,}96 \end{aligned}$$

(b) Bemerkung: Die analytische Lösung lässt sich erst nach Einführung der Transformationsformel für Verteilungen verstehen.

Wenn X normalverteilt ist mit Mittelwert μ und Varianz σ^2, so gilt $P(X \leq \alpha) = \Phi(\frac{\alpha-\mu}{\sigma})$.

(i)
$$\begin{aligned} P_X((-\infty;2)) &= P(X \in (-\infty;2)) \\ &= P(Z \in (-\infty; \tfrac{2-5}{3})) \text{ mit } Z = \tfrac{X-\mu}{\sigma} \sim N(0,1) \\ &= \Phi(-1) \\ &\approx 0{,}1587 \end{aligned}$$

$$\begin{aligned} P((-\infty;6)) &= \Phi(\tfrac{6-5}{3}) \\ &\approx 0{,}6293 \end{aligned}$$

$$\begin{aligned} P((2;7)) &= \Phi(\tfrac{7-5}{3}) - \Phi(\tfrac{2-5}{3}) \\ &= \Phi(\tfrac{2}{3}) + \Phi(1) - 1 \\ &\approx 0{,}5899 \end{aligned}$$

$$\begin{aligned} P((4,5;\infty)) &= 1 - P(X \in (-\infty;4,5)) \\ &= 1 - \Phi(\tfrac{4{,}5-5}{3}) \\ &\approx 0{,}5675 \end{aligned}$$

(ii)
$$\begin{aligned} 0{,}84 &\overset{!}{=} P(X \in (\infty;c)) \\ &= \Phi(\tfrac{c-5}{3}) \\ \tfrac{c-5}{3} &\approx 0{,}99, \text{ d.h. } c \approx 7{,}97 \end{aligned}$$

(iii)
$$\begin{aligned} 0{,}57 &\overset{!}{=} P(X \in [5-c;5+c]) \\ &= \Phi(\tfrac{c}{3}) - \Phi(-\tfrac{c}{3}) \\ &= \Phi(\tfrac{c}{3}) + \Phi(\tfrac{c}{3}) - 1 \\ \Phi(\tfrac{c}{3}) &= 0{,}785, \text{ d.h. } c \approx 2{,}37 \end{aligned}$$