Fachhochschule Graubünden
University of Applied Sciences

Übungsblatt Sto 8

Computational and Data Science
BSc HS2024

Lösungen

Differentialgleichungen und Stochastik

Lernziele:

- Sie kennen die Begriffe Grundgesamtheit, Merkmal, Merkmalsträger, Merkmalsausprägung, Nominal-, Ordinal-, metrische Skala, kumulierte Häufigkeit, empirische Verteilungsfunktion, Modus, Median, arithmetisches Mittel, Quantil, mittlere absolute Abweichung, Spannweite, empirische Varianz und können diese auf konkrete Datensätze anwenden.
- Sie kennen verschiedene Methoden, um einen Datensatz grafisch darzustellen: Stabdiagramm, Rechteckdiagramm, Kreisdiagramm, Histogramm, Polygonzug, Box-Plot und können diese auch für konkrete Datensätze anwenden.

---

1. Baumbestand in D

Für das Jahr 1997 wurden in den deutschen Bundesländern (ausser Berlin) folgende Zahlen für den Anteil (in %) von Bäumen mit deutlichen Umweltschäden ausgewiesen:

Bundesland	HE	NS	NRW	SH	BB	MV	S
Anteil	16	15	20	20	10	10	19

Bundesland	SA	TH	BW	B	HH	RP	SL
Anteil	14	38	19	19	33	24	19

Erläutern Sie die Begriffe Grundgesamtheit, Untersuchungseinheit, Merkmal und Ausprägung anhand dieses Beispiels. Zeichnen Sie für die obigen Angaben einen Boxplot. Vergleichen Sie arithmetisches Mittel und Median der Angaben.

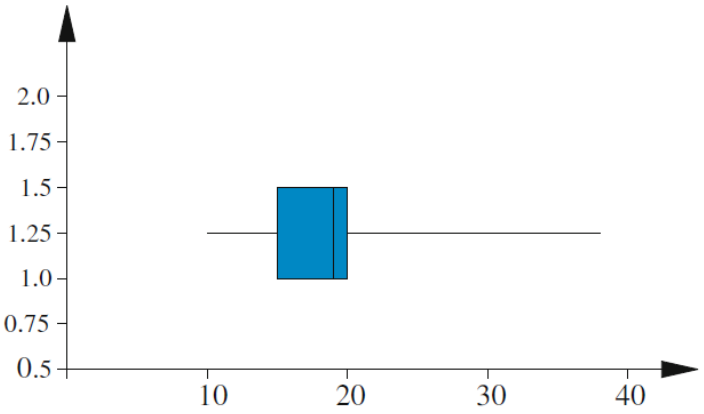
Grundgesamtheit: Bäume in den deutschen Bundesländern (ausser Berlin) im Jahr 1997.
Untersuchungseinheit: Baum.
Untersuchungsmerkmal: Umweltschaden.
Ausprägung: Umweltschaden Ja/Nein.

Die geordneten Daten sind

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	10	10	14	15	16	19	19	19	19	20	20	24	33	38

Die für den Boxplot notwendigen Größen sind: Minimum $x_{(1)} = 10$, unteres Quartil $x_{0.25} = x_{(4)} = 15$, Median $x_{\text{Med}} = x_{(7)} = x_{(8)} = 19$, oberes Quartil $x_{0.75} = x_{(11)} = 20$, Maximum $x_{(14)} = 38$.

Der Boxplot zeigt eine linkssteile Verteilung. Die wird durch die Relation: Modus = Median = $19 < \bar{x} = 19.71$ bekräftigt.



2. Fussball

Sieben ehemalige Profifussballer werden nach der Anzahl der Tore befragt, die sie im Laufe ihrer Karriere geschossen haben; sie machen folgende Angaben:

Spieler	1	2	3	4	5	6	7
Tore	1500	2300	1560	3000	1950	1560	2150

Berechnen Sie für diese Werte den Median, das untere und das obere Quartil, den Mittelwert, die mittlere absolute Abweichung und die empirische Varianz.

Median: 1950
Unteres Quartil: 1560
Oberes Quartil: 2300
Mittelwert: 2002,86
Mittlere absolute Abweichung: 404,29
Empirische Varianz: 292023,81

3. Beliebtheit in Schule

In einer Studie wurden 478 amerikanische Schüler in der 4. bis 6.Klassenstufe befragt, durch welche Eigenschaften Jugendliche beliebt werden. Die Schüler stammten sowohl aus städtischen, vorstädtischen und ländlichen Schulbezirken und wurden zusätzlich nach einigen demografischen Informationen gefragt. Die erhobenen Merkmale in der Studie waren u. a.

1. Geschlecht: Mädchen oder Junge
2. Klassenstufe: 4, 5 oder 6
3. Alter (in Jahren)
4. Hautfarbe: Weiß, Andere
5. Region: ländlich, vorstädtisch, städtisch
6. Schule: Brentwood Elementary, Brentwood Middle, usw.
7. Ziele: die Antwortalternativen waren 1 = gute Noten, 2 = beliebt sein, 3 = gut im Sport
8. Noten: Wie wichtig sind Noten für die Beliebtheit (1 = am wichtigsten bis 4 = am unwichtigsten)

Geben Sie für die acht Merkmale jeweils den Typ und die Skalierung sowie geeignete Parameter und grafische Darstellungen an.

Merkmal	Typ	Skalierung
Geschlecht	diskret	nominal
Stufe	diskret	kardinal/ordinal
Alter	diskret	kardinal
Rasse	diskret	nominal
Region	diskret	nominal
Schule	diskret	nominal/ordinal
Ziele	diskret	nominal
Noten	diskret	ordinal

Merkmal	Lageparameter	Streuung	graf. Darstellung
Geschlecht	Modus	–	Kreis
Stufe	Modus	–	Kreis, Balken
Alter	Mittelwert	Varianz	Histogramm, Box-Plot
Rasse	Modus	–	Kreis, Balken
Region	Modus	–	Kreis, Balken
Schule	Modus	–	Kreis, Balken
Ziele	Modus	–	Kreis, Balken
Noten	Modus	–	Kreis, Balken

4. Fehlzeiten

In der nachstehenden Tabelle sind die Fehlzeiten (in Tagen) von 50 Arbeitnehmern (AN) eines Unternehmens für das vergangene Jahr angegeben.

Fehlzeit (Tage)	0	3	5	9	12	18	21
Anzahl der AN	5	9	13	9	8	4	2

a) Berechnen Sie das arithmetische Mittel, den Modus und den Median.
b) Berechnen Sie das 3. Quartil und das 9. Dezil.
c) Berechnen Sie die mittlere absolute Abweichung, die Varianz und die Standardabweichung!

Die nachstehende Arbeitstabelle dient der Durchführung und zugleich der übersichtlichen Darstellung erforderlicher Rechenoperationen:

(1) x_i	(2) h_i	(3) H_i	(4) F_i	(5) $x_i \cdot h_i$	(6) H_i^*	(7) F_i^*	(8) $\lvert x_i - \bar{x} \rvert$	(9) $\lvert x_i - \bar{x} \rvert \cdot h_i$	(10) $(x_i - \bar{x})^2 \cdot h_i$
0	5	5	0,10	0	0	0,00	7,66	38,30	293,38
3	9	14	0,28	27	27	0,07	4,66	41,94	195,44
5	13	27	0,54	65	92	0,24	2,66	34,58	91,98
9	9	36	0,72	81	173	0,45	1,34	12,06	16,16
12	8	44	0,88	96	269	0,70	4,34	34,72	150,68
18	4	48	0,96	72	341	0,89	10,34	41,36	427,66
21	2	50	1,00	42	383	1,00	13,34	26,68	355,91
	50			383				229,64	1.531,21

x_i = Merkmalswert
h_i = absolute einfache Häufigkeit
H_i = absolute kumulierte Häufigkeit
F_i = relative kumulierte Häufigkeit
H_i^* = absolute kumulierte Häufigkeit (wobei $h_i^* = x_i \cdot h_i$; Spalte 5)
$\bar{x}$ = arithmetisches Mittel
n = Anzahl der Merkmalsträger (hier: n = 50)
v = Anzahl der verschiedenen Merkmalswerte (hier: v = 7)

a) Mittelwerte

i) Arithmetisches Mittel:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{v} x_i \cdot h_i = \frac{1}{50} \cdot 383 = 7{,}66 \qquad \text{(Berechnung s. Spalte 5)}$$

Die durchschnittliche Fehlzeit der Arbeitnehmer beträgt 7,66 Tage.

Fehlerquelle: Division der gesamten Fehlzeit 383 Tage mit v = 7 (Anzahl der verschiedenen Merkmalswerte) anstatt mit n = 50 (Anzahl der Arbeitnehmer).

ii) Modus:
Am häufigsten, nämlich 13-mal, wurde die Fehlzeit 5 Tage beobachtet.

iii) Median:
Die relative kumulierte Häufigkeit F = 0,50 (Spalte 4) wird bei dem Merkmalswert 5 erreicht. (Mindestens) 50 % der Arbeitnehmer haben höchstens 5 Tage gefehlt, (mindestens) 50 % der Arbeitnehmer haben mindestens 5 Tage gefehlt.

b) Quantile

i) 3. Quartil (75 % / 25 %):
Die relative kumulierte Häufigkeit F = 0,75 (Spalte 4) wird bei dem Merkmalswert 12 erreicht.
(Mindestens) 75 % der Arbeitnehmer haben höchstens 12 Tage gefehlt.

ii) 9. Dezil (90 % / 10 %):
Die relative kumulierte Häufigkeit F = 0,90 (Spalte 4) wird bei dem Merkmalswert 18 erreicht.
(Mindestens) 90 % der Arbeitnehmer haben höchstens 18 Tage gefehlt.

c) Streuungsmaße

i) Mittlere absolute Abweichung:

$$\delta = \frac{1}{n} \cdot \sum_{i=1}^{v} \left| x_i - \bar{x} \right| \cdot h_i = \frac{1}{50} \cdot 229,64 = 4,59 \qquad \text{(Berechnung s. Sp. 9)}$$

Die Fehlzeit der Arbeitnehmer weicht durchschnittlich um 4,59 Tage von der durchschnittlichen Fehlzeit 7,66 Tage ab.

Fehlerquelle: Division der gesamten Abweichung 229,64 mit v = 7 (Anzahl der verschiedenen Merkmalswerte) anstatt mit n = 50 (Anzahl der Arbeitnehmer).

ii) Varianz (mittlere quadratische Abweichung) und Standardabweichung:

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{v} (x_i - \bar{x})^2 \cdot h_i = \frac{1}{50} \cdot 1.531,21 = 30,62 \text{ Tage}^2 \qquad \text{(s. Sp. 10)}$$

Fehlerquelle: Division der gesamten quadrierten Abweichung 1.531,21 mit v = 7 (Anzahl der Merkmalswerte) anstatt mit n = 50 (Anzahl der Arbeitnehmer).

$$\sigma = \sqrt{\sigma^2} = \sqrt{30,62} = 5,53 \text{ Tage}$$

Eine inhaltliche Interpretation der Varianz (Dimension: Tage^2!) und damit auch der Standardabweichung als Wurzel aus der Varianz ist nicht möglich.

5. Überstunden

Die nachstehende Häufigkeitsverteilung zeigt auf, wie viele Überstunden die 30 Arbeitnehmer (AN) einer Firma in der letzten Woche geleistet haben.

Überstunden	0	1	2	3	4	5	8
Anzahl der AN	7	3	4	9	4	2	1

a) Wie viele Überstunden haben die Arbeitnehmer insgesamt geleistet?
b) Berechnen und interpretieren Sie das arithmetische Mittel, den Median und den Modus!
c) Ermitteln und interpretieren Sie das 3. Quartil und das 9. Dezil!
d) Berechnen Sie die mittlere absolute Abweichung, die Varianz, die Standardabweichung und den Variationskoeffizienten!

e) Welchen Anteil an den Gesamtüberstunden haben die unteren 25 % der Arbeitnehmer, welchen die oberen 25 % der Arbeitnehmer und welchen die unteren 25 Arbeitnehmer?

f) Auf welchen Anteil der Arbeitnehmer entfallen die unteren 80 % der Überstunden, auf welchen die unteren 50 % der Überstunden?

a) $\sum_{i=1}^{7} x_i \cdot h_i = 72$

b) $\bar{x} = \frac{1}{30} \cdot 72 = 2{,}4$; Modus: 3; Median: 3

c) 3. Quartil: 3; 9. Dezentil: 4

d) $\delta = \frac{1}{30} \cdot 45{,}2 = 1{,}51$; $\sigma^2 = \frac{1}{30} \cdot 105{,}2 = 3{,}51$; $\sigma = \sqrt{3{,}51} = 1{,}87$

$VK = \frac{1{,}87}{2{,}4} \cdot 100 = 77{,}9\ \%$

e) i) $F^* = 0{,}00 + \frac{0{,}25 - 0{,}23}{0{,}33 - 0{,}23} \cdot (0{,}04 - 0{,}00) = 0{,}008$

ii) $1 - F^* = 1 - [0{,}15 + \frac{0{,}75 - 0{,}47}{0{,}77 - 0{,}47} \cdot (0{,}53 - 0{,}15)] = 0{,}495$

iii) $F^* = 0{,}53 + \frac{25 - 23}{27 - 23} \cdot (0{,}75 - 0{,}53) = 0{,}64$

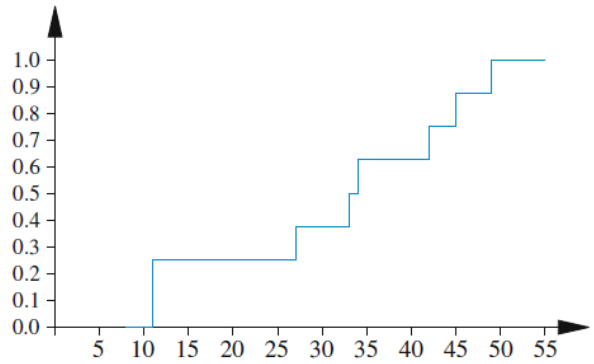
f) i) $F = 0{,}90 + \frac{0{,}80 - 0{,}75}{0{,}89 - 0{,}75} \cdot (0{,}97 - 0{,}90) = 0{,}925$

ii) $F = 0{,}47 + \frac{0{,}50 - 0{,}15}{0{,}53 - 0{,}15} \cdot (0{,}77 - 0{,}47) = 0{,}746$

6.
Sie erhalten die folgenden 8 Datenwerte: 34, 45, 11, 42, 49, 33, 27, 11.

a) Bestimmen Sie die empirische Verteilungsfunktion F.
b) Geben Sie arithmetisches Mittel, Median und Modus an.
c) Berechnen Sie die Varianz i) aus den ungruppierten Originalwerten, ii) aus den geordneten Werten und iii) mit dem Verschiebungssatz.
d) Berechnen Sie die Standardabweichung, die mittlere absolute Abweichung vom Median und die Spannweite.

a)



b) Arithmetisches Mittel:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(34 + 45 + 11 + 42 + 49 + 33 + 27 + 11)$$
$$= 31.5$$

Bestimmung des Medians: Die geordneten Daten sind: $x_{(1)} = 11$, $x_{(2)} = 11$, $x_{(3)} = 27$, $x_{(4)} = 33$, $x_{(5)} = 34$, $x_{(6)} = 42$, $x_{(7)} = 45$, $x_{(8)} = 49$,

$$x_{\text{med}} = \frac{1}{2}(x_{(4)} + x_{(5)}) = \frac{1}{2}(33 + 34) = 33.5.$$

Bestimmung des Modus: Der Modus liegt bei 11, da dieser Wert zweimal auftritt und alle anderen nur einmal.

c)

Bestimmung der Varianz aus der Urliste:

$$\begin{aligned}\text{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{8}\big((34 - 31.5)^2 + (45 - 31.5)^2 \\ &\quad + (11 - 31.5)^2 + (42 - 31.5)^2 \\ &\quad + (49 - 31.5)^2 + (33 - 31.5)^2 \\ &\quad + (27 - 31.5)^2 + (11 - 31.5)^2\big) \\ &= \frac{1}{8} \cdot 1468 = 183.5\end{aligned}$$

Bestimmung der Varianz aus geordneten Liste:

i	x_i	n_i	$x_i - \overline{x}$	$n_i(x_i - \overline{x})^2$	$n_i x_i^2$
	11	2	−20.5	840.5	242
	27	1	−4.5	20.25	729
	33	1	1.5	2.25	1089
	34	1	2.5	6.25	1156
	42	1	10.5	110.25	1764
	45	1	13.5	182.25	2025
	49	1	17.5	306.25	2401
$\sum$		8		1468	9406

Dann ist $\text{var}(\boldsymbol{x}) = \frac{1}{8}\sum n_i\,(x_i - \overline{x})^2 = \frac{1468}{8} = 183.5$. Der Verschiebungssatz liefert

$$\text{var}(\boldsymbol{x}) = \frac{1}{8}\sum n_i x_i^2 - \overline{x}^2 = \frac{9406}{8} - 31.5^2 = 183.5$$

d)

Die Standardabweichung ist:

$$\sqrt{\text{var}(x)} = \sqrt{183.5} = 13.546$$

Bestimmung der absoluten Abweichung:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|x_i - x_{\text{med}}| &= \frac{1}{8}\big(|34-33.5| + |45-33.5| \\ &\quad + |11-33.5| + |42-33.5| \\ &\quad + |49-33.5| + |33-33.5| \\ &\quad + |27-33.5| + |11-33.5|\big) \\ &= \frac{1}{8}\big(0.5 + 11.5 + 22.5 + 8.5 \\ &\quad + 15.5 + 0.5 + 6.5 + 22.5\big) \\ &= \frac{1}{8}\cdot 88 = 11\end{aligned}$$

Die Spannweite ist $x_{(n)} - x_{(1)} = 49 - 11 = 38$