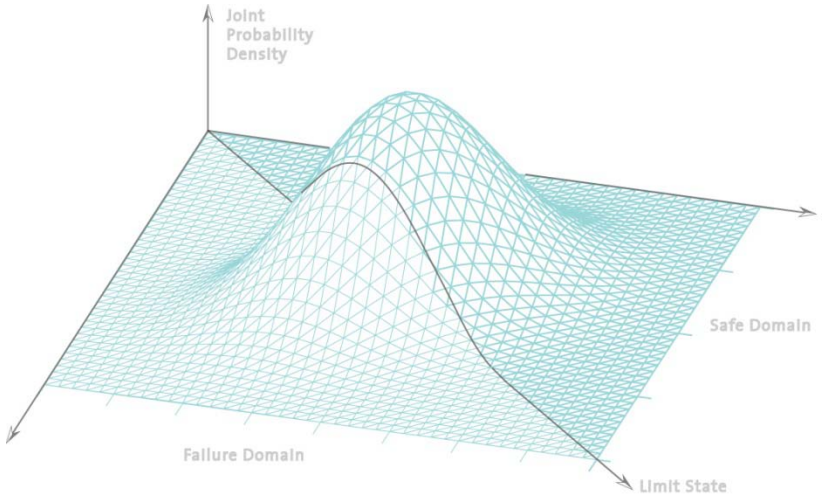


Statistik und Wahrscheinlichkeitsrechnung

Übung 3

Inhalt der heutigen Übung

- Vorrechnen der Hausübung B.7
- Beschreibende Statistik
- Gemeinsames Lösen der Übungsaufgaben
  - C.1: Häufigkeitsverteilung
  - C.2: Tukey Box Plot
  - C.5: Korrelation
- Vorstellen der Hausübung

Numerische Kennwerte der Stichprobe

Lageparameter:

Arithmetisches Mittel:	Schwerpunkt der Stichprobe
Median:	Mittlerer Wert einer Stichprobe
Modalwert:	Am häufigsten vorkommender Wert

Streuungsparameter:

Varianz / Standardabweichung:	Verteilung um den Mittelwert
Variationskoeffizient :	Variabilität relativ zum Mittelwert

Andere Parameter:

Schiefekoeffizient:	Schiefe relativ zum Mittelwert
Kurtosis:	Spitzigkeit/Gipfligkeit um den Mittelwert

Masse für die Korrelation:

Kovarianz:	Tendenz für paarweise beobachtete Eigenschaften
Korrelationskoeffizient :	Normalisierter Koeffizient zwischen -1 und +1

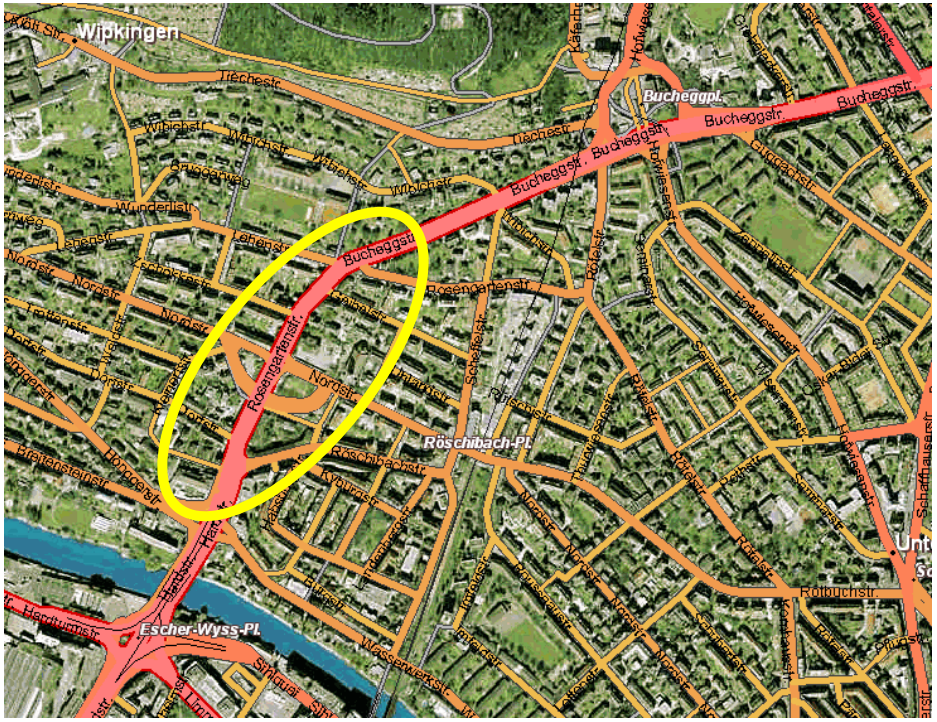
Grafische Darstellung

Ein-dimensionales Streudiagramm	Veranschaulicht den Bereich und die Verteilung von Datenreihen entlang einer Achse, und zeigt Symmetrie.
Zwei-dimensionales Streudiagramm	Veranschaulicht den paarweisen Zusammenhang von Daten.
Histogramm	Stellt die Verteilung von Daten über einem Bereich von Datenreihen dar, zeigt Modalwert und Symmetrie.
Quantil-Plot	Stellt Median, Verteilung und Symmetrie dar.
Tukey Box Plot	Stellt Median, obere/untere Quartile, Symmetrie und Verteilung dar.
Q-Q Plot	Vergleicht zwei Datenreihen, relatives Bild.
Mittelwert-Differenz Plot	Vergleicht zwei Datenreihen, relatives Bild.

Beschreibende Statistik

Uns wurden die unten dargestellten Verkehrsdaten der Rosengartenstrasse in Zürich aus dem Monat April 2001 zur Auswertung übergeben. Richtung 1 gibt die Verkehrsbelastung zum Bucheggplatz, Richtung 2 die Belastung zum Escher-Wyss-Platz an.

Datum	Richtung 1	Richtung 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877



Beschreibende Statistik

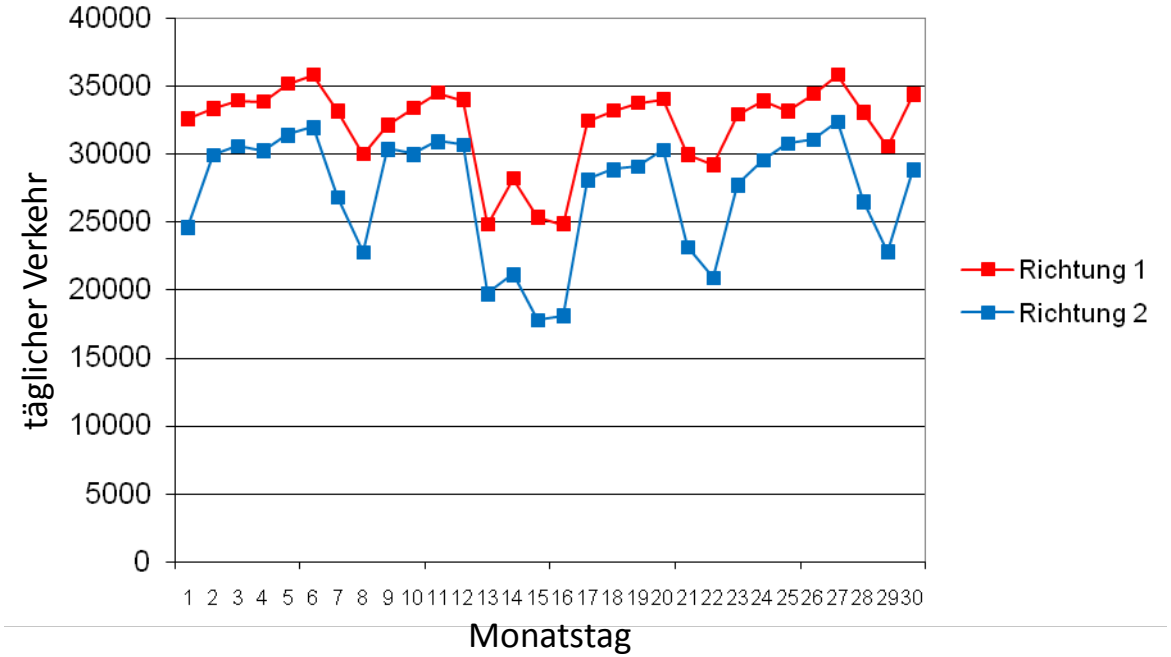
Was wollen wir wissen?

Wie können wir das mithilfe von Daten beschreiben?

- Grafik, Histogramm
- numerische Kennwerte etc.

Beispiel:

Man will etwas über die Änderung des Verkehrs in Richtung 1 im Monat April erfahren.



Beschreibende Statistik

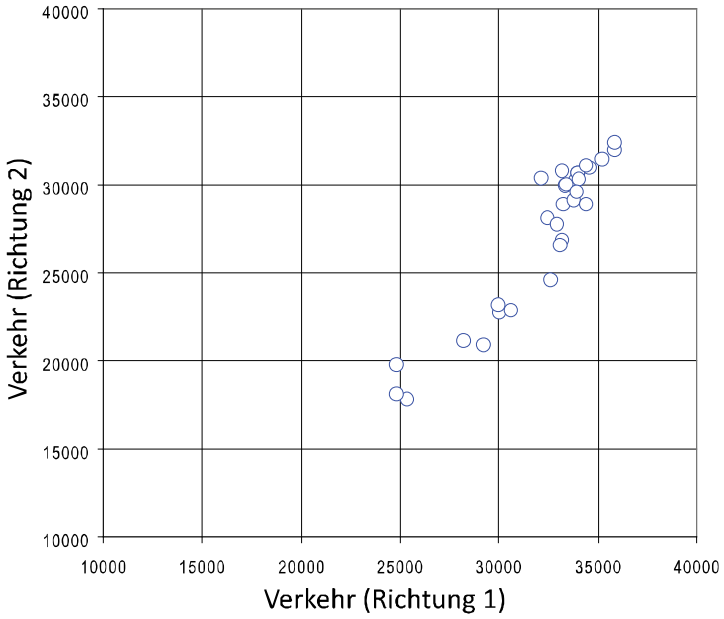
Was wollen wir wissen?

Wie können wir das mithilfe von Daten beschreiben?

- Grafik, Histogramm
- numerische Kennwerte etc.

Beispiel:

Man will etwas über die Änderung des Verkehrs in Richtung 1 im Monat April erfahren.



Beschreibende Statistik

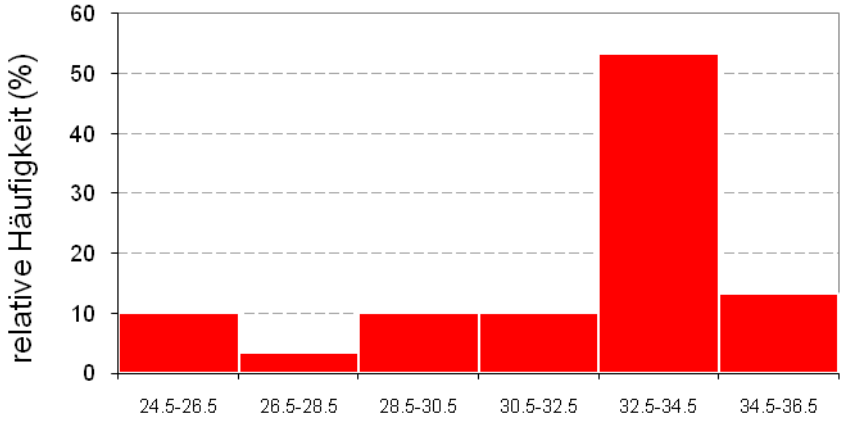
Was wollen wir wissen?

Wie können wir das mithilfe von Daten beschreiben?

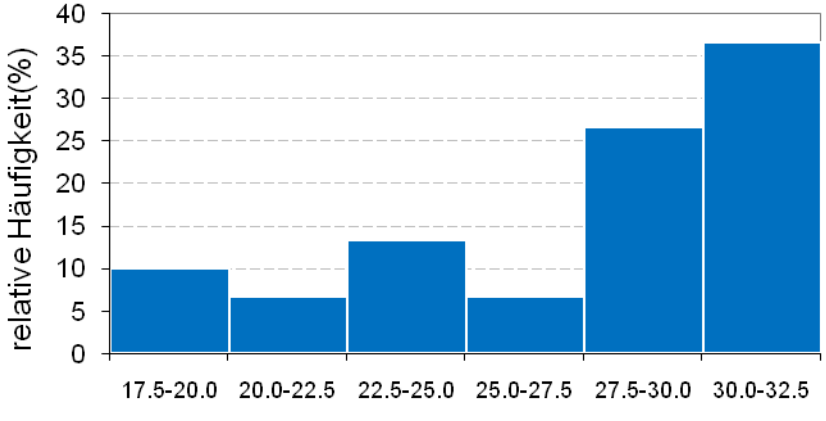
- Grafik, Histogramm
- numerische Kennwerte etc.

Beispiel:

Man will etwas über die Änderung des Verkehrs in Richtung 1 im Monat April erfahren.



Richtung 1



Richtung 2

…was ist bei diesem Vergleich **nicht** gut gelöst?

Beschreibende Statistik

Wir üben heute…

wie man die Eigenschaften von gegebenen Daten darstellen kann, und zwar:

- Grafisch

  Häufigkeitsdiagramm (Histogramm)
  kumulative Häufigkeitsverteilung

- Numerisch

  Stichprobenmittelwert
  Standardabweichung der Stichproben

- Zusammenfassungen

  Tukey Box Plot

- Korrelation von Datenreihen

Anmerkung: Du kannst Excel, Matlab und/oder andere Statistikprogramme verwenden.

ABER !!!!!

Stelle IMMER sicher, Funktionen selbst einzusetzen oder zu prüfen, ob die Funktionen, die du vom verwendeten Programm bereitgestellt bekommst, mit denen des Skripts übereinstimmen!

Aufgabe C.1

Erstelle aus den erhobenen Verkehrsdaten nach deren Einteilung in Intervalle eine Häufigkeitsdiagramm sowie eine kumulierte Häufigkeitsdiagramm und stelle deren Verläufe in den geeigneten Graphen dar.
Wie würdest Du die Daten einem ersten Eindruck nach charakterisieren?
Fertige einen Vergleich der Verkehrsflüsse beider Richtungen an.

Aufgabe C.1

Datum	Richtung 1	Richtung 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

Die einzelnen Schritte:

1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. Kumulative Häufigkeitsdiagramm zeichnen

Aufgabe C.1

Schritt 1

Schritte:
1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. Kumulative Häufigkeitsdiagramm zeichnen

Datum	Richtung 1	Richtung 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

sortieren

Richtung 1 sortiert	Richtung 2 sortiert
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

Schritte:
1. Daten sortieren
2. ⇨ Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. Kumulative Häufigkeitsdiagramm zeichnen

Aufgabe C.1

Schritt 2

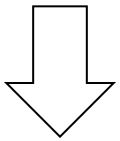
Es gibt keine allgemeingültige Regel, aber eine Faustregel:

$$k = 1 + 3.3 \log_{10} n$$

dabei ist k die Anzahl an Intervallen und n die Anzahl an Daten.

In unserem Fall: $n = 30$ $\quad k = 1 + 3.3 \log_{10} 30 = 5.87 \approx 6$ Intervalle

Für Richtung 1: Minimum = 24846 und Maximum= 35852



Wir könnten folgende Intervalle wählen [Anzahl Fahrzeuge $\cdot 10^3$]:
(24.0,26.0] ,(26.0,28.0], (28.0,30.0],(30.0,32.0],(32.0,34.0],(34.0,36.0]

Aufgabe C.1

Schritt 3

Schritte:
1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. ⇨ Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. Kumulative Häufigkeitsdiagramm zeichnen

Richtung 1 sortiert

24846
24862
25365
28252
29224
29976
30035
30613
32158
32472
32618
32962
33091
33197
33198
33245
33380
33406
33788
33888
33937
34007
34013
34076
34425
34455
34576
35237
35843
35852

zählen ⇨

Richtung 1	Intervall (Anzahl der Autos *10^3)	Intervall-Mittelpunkt (Anzahl der Autos *10^3)	Abs. Häufigkeit in dem Intervall
	24.0-26.0	25.0	3
	26.0-28.0	27.0	0
	28.0-30.0	29.0	3
	30.0-32.0	31.0	2
	32.0-34.0	33.0	13
	34.0-36.0	35.0	9

Schritte:
1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. ⇨ Häufigkeitsdiagramm zeichnen
5. Kumulative Häufigkeitsdiagramm zeichnen

Aufgabe C.1

Schritt 4

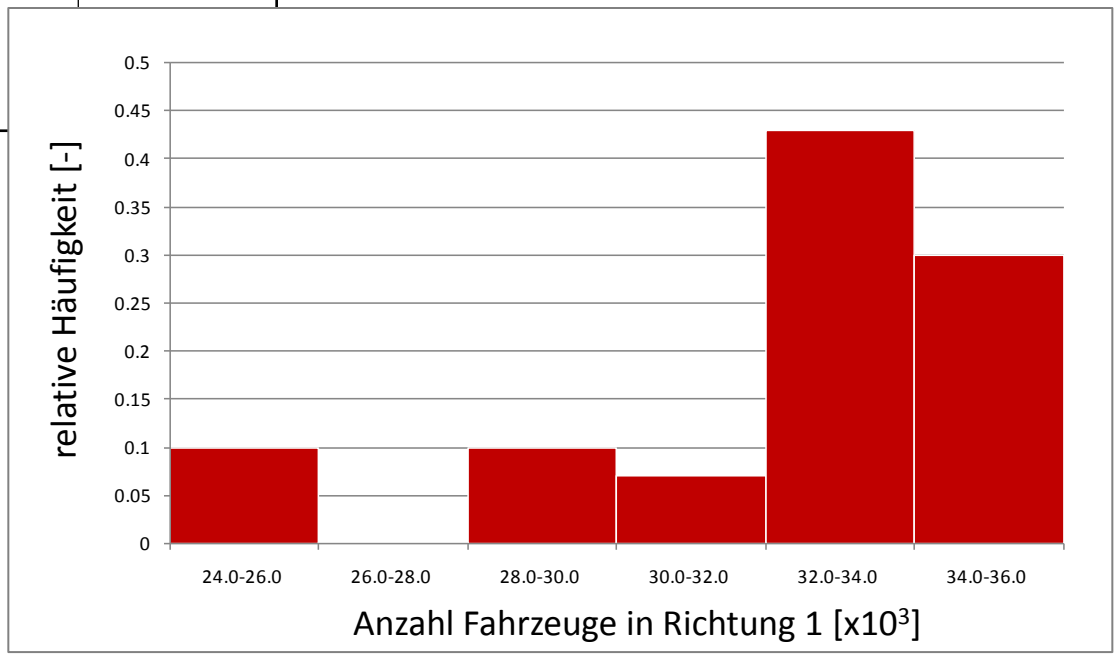
Zuerst einige Berechnungen ….

	Intervall [Anzahl Autos x10³]	Intervall-Mittelpunkt [Anzahl Autos x10³]	absolute Häufigkeit [-]	relative Häufigkeit [-]
Richtung 1	24.0-26.0	25.0	3	0.10
	26.0-28.0	27.0	0	0.00
	28.0-30.0	29.0	3	0.10
	30.0-32.0	31.0	2	0.07
	32.0-34.0	33.0	13	
	34.0-36.0	35.0	9	

$$rel.\ Häufigkeit = \frac{n_k}{n_{ges}} = \frac{3}{30} = 0.10$$

… und dann zeichnen.

Histogramm



Schritte:
1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. ⇨ Kumulative Häufigkeitsdiagramm zeichnen

Aufgabe C.1

Schritt 5

Berechnung der kumulierten relativen Häufigkeit:

	Intervall [Anzahl Autos x10³]	**Intervall-Mittelpunkt** [Anzahl Autos x10³]	**absolute Häufigkeit** [-]	**relative Häufigkeit** [-]	**kumulierte, rel. Häufigkeit** [-]
Richtung 1	24.0-26.0	25.0	3	0.10	0.10
	26.0-28.0	27.0	0	0.00	**0.10**
	28.0-30.0	29.0	3	0.10	0.20
	30.0-32.0	31.0	2	0.07	0.27
	32.0-34.0	33.0	13	0.43	0.70
	34.0-36.0	35.0	9	0.30	1.00

schrittweise addieren (kumulieren)

Schritte:
1. Daten sortieren
2. Geeignete Anzahl von Intervallen wählen
3. Die Anzahl der Daten für jedes Intervall zählen
4. Häufigkeitsdiagramm zeichnen
5. ⇨ Kumulative Häufigkeitsdiagramm zeichnen

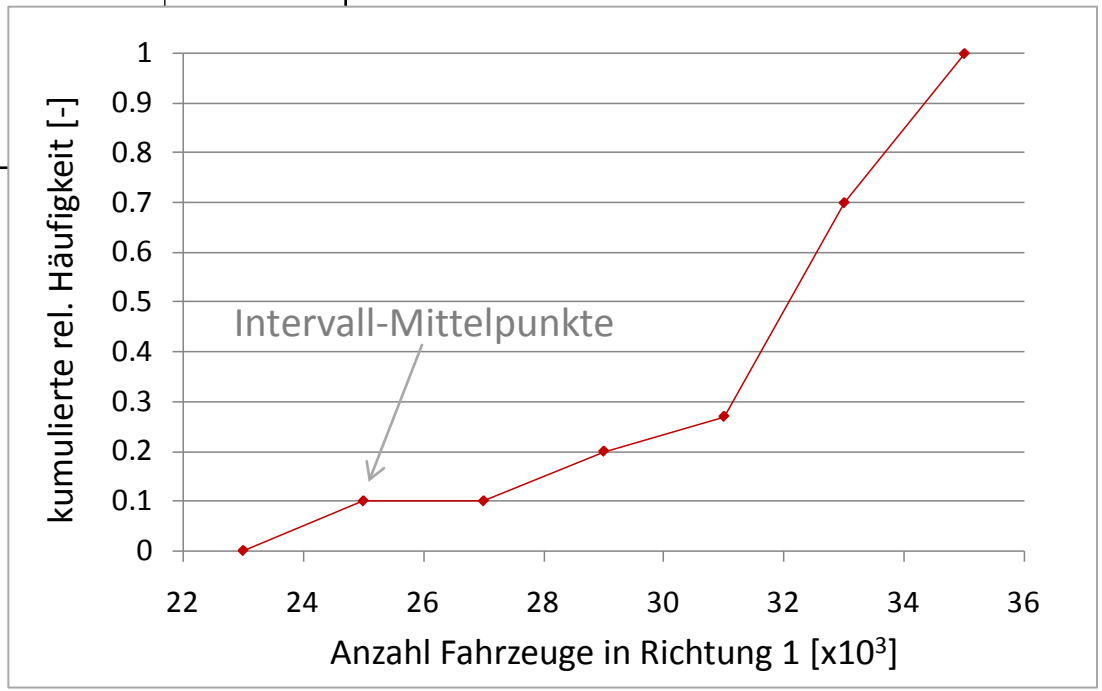
Aufgabe C.1

Schritt 5

Berechnung der kumulierten relativen Häufigkeit:

	Intervall [Anzahl Autos x10³]	Intervall-Mittelpunkt [Anzahl Autos x10³]	absolute Häufigkeit [-]	relative Häufigkeit [-]	kumulierte, rel. Häufigkeit [-]
Richtung 1	24.0-26.0	25.0	3	0.10	0.10
	26.0-28.0	27.0	0	0.00	0.10
	28.0-30.0	29.0	3	0.10	0.20
	30.0-32.0	31.0	2		
	32.0-34.0	33.0	13		
	34.0-36.0	35.0	9		

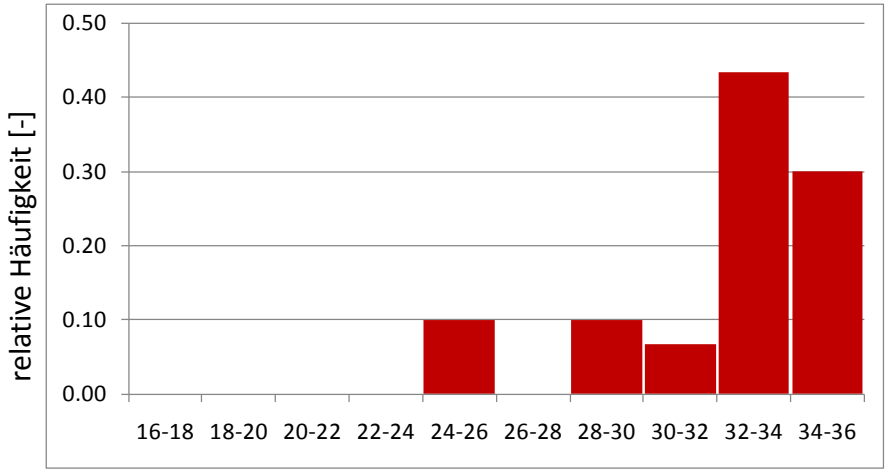
Diagramm der kumulierten relativen Häufigkeiten



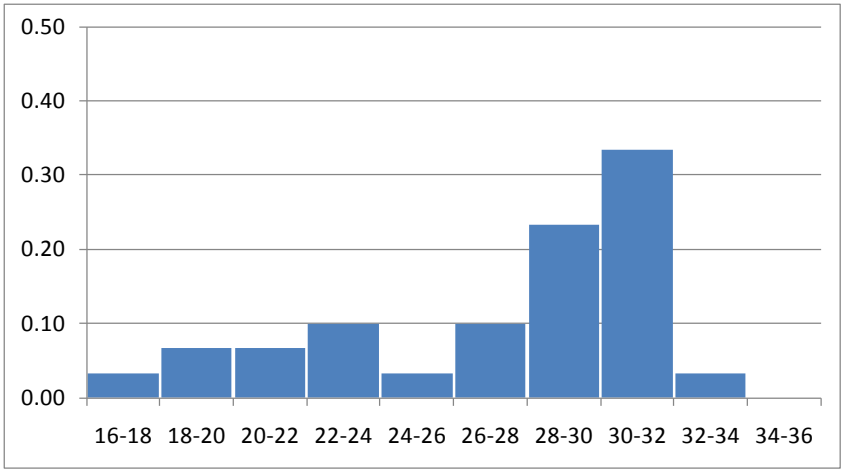
Aufgabe C.1

Die gleichen Schritte können auch für Richtung 2 durchgeführt werden. Was kann aus diesen Diagrammen erkannt werden?

Histogramme



Anzahl Fahrzeuge in Richtung 1 [x10^3]



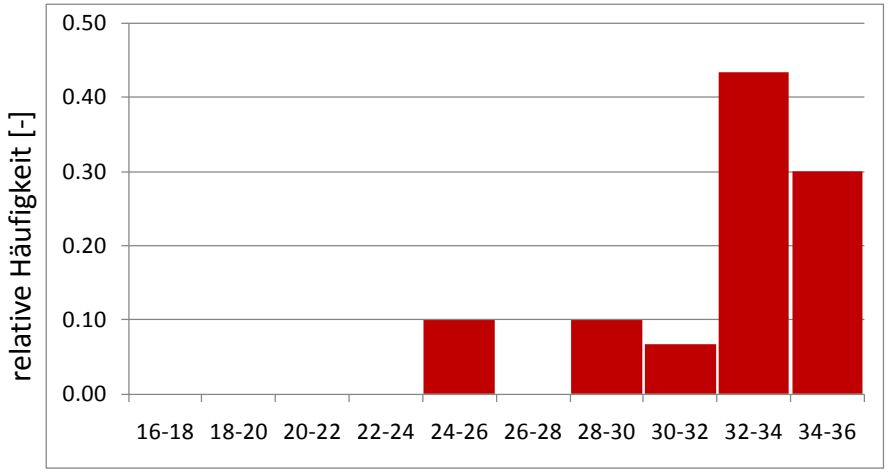
Anzahl Fahrzeuge in Richtung 2 [x10^3]

Damit die Histogramme vergleichbar sind, werden für das zweite Histogramm (Richtung 2) die gleichen Intervalle wie in dem Histogramm für Richtung 1 gewählt.

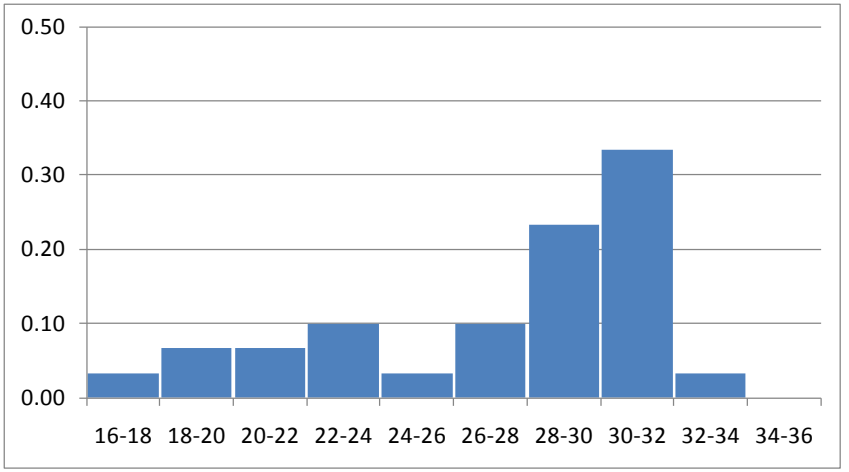
Aufgabe C.1

Die gleichen Schritte können auch für Richtung 2 durchgeführt werden. Was kann aus diesen Diagrammen erkannt werden?

Histogramme

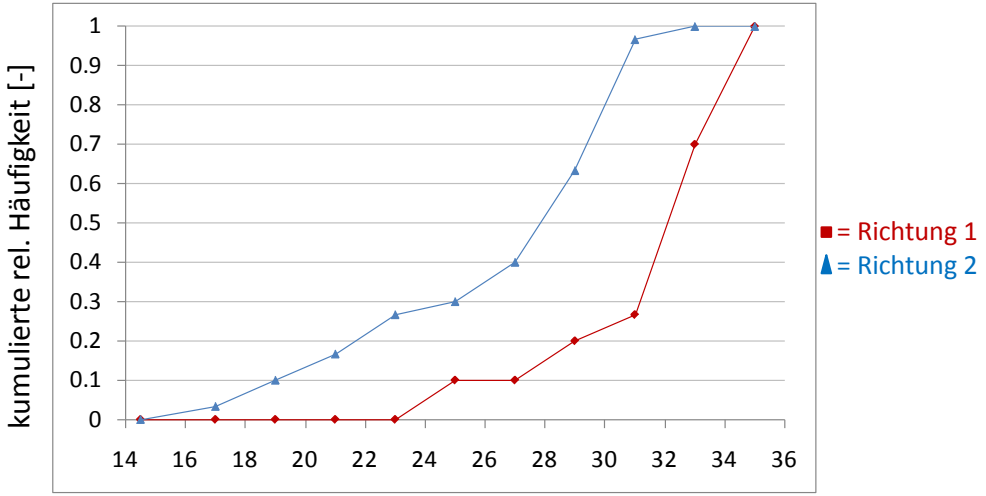


Anzahl Fahrzeuge in Richtung 1 [x10^3]



Anzahl Fahrzeuge in Richtung 2 [x10^3]

Diagramm der kumulierten relativen Häufigkeiten



Anzahl Fahrzeuge [x10^3]

Quantil

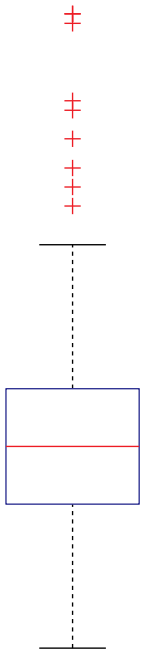
Das Quantil ist für eine gegebene Anzahl an Beobachtungen wie folgt definiert:

- Das υ-Quantil ist der Wert, der die unteren $\upsilon \cdot 100\%$ der Messwerte von den oberen $100\% - \upsilon \cdot 100\%$ trennt.
- Beispiel: Das 0.75-Quantil (das obere Quartil) trennt die unteren 75% von den oberen 25% der Daten.
- Die Quantile werden von der geordneten (sortierten) Stichprobe berechnet: $x_1^0 \le x_2^0 \le \ldots \le x_n^0$
- Der Quantilindex wird wie folgt berechnet:

$$\upsilon = \frac{i}{n+1}; \qquad n: \text{ Gesamtanzahl der Beobachtungen, } i=1,2\ldots,n$$

Aufgabe C.2

Verwende für beide Datenreihen der Verkehrsdaten den Tukey Box Plot, um eine zusammenfassende Übersicht über die Eigenschaften der jeweiligen Verteilung zu bekommen. Trage beide Darstellungen in die gleiche Grafik auf, um eine verbesserte Vergleichbarkeit zu erzielen und beurteile die Datenreihen hinsichtlich ihrer Symmetrie.



Schritte

1. Berechne den Median
2. Berechne das 0.75- und 0.25-Quantil.
3. Berechne die Nachbarschaftswerte.
4. Bestimmung der Ausreisser
5. Zeichne den Tukey Box Plot

Aufgabe C.2

Schritt 1

Schritte:
1. Berechne den Median
2. Berechne das 0.75- und 0.25- Quantil.
3. Berechne die Nachbarschaftswerte.
4. Ausreisser
5. Zeichne den Tukey Box Plot

Der Median ist der 0.50-Quantil.

Richtung 1

24846
24862
25365
28252
29224
29976
30035
30613
32158
32472
32618
32962
33091
33197
33198
33245
33380
33406
33788
33888
33937
34007
34013
34076
34425
34455
34576
35237
35843
35852

Aber …. wenn die Anzahl der Daten gerade ist, ist das nicht möglich!

In diesem Fall müssen wir linear interpolieren.

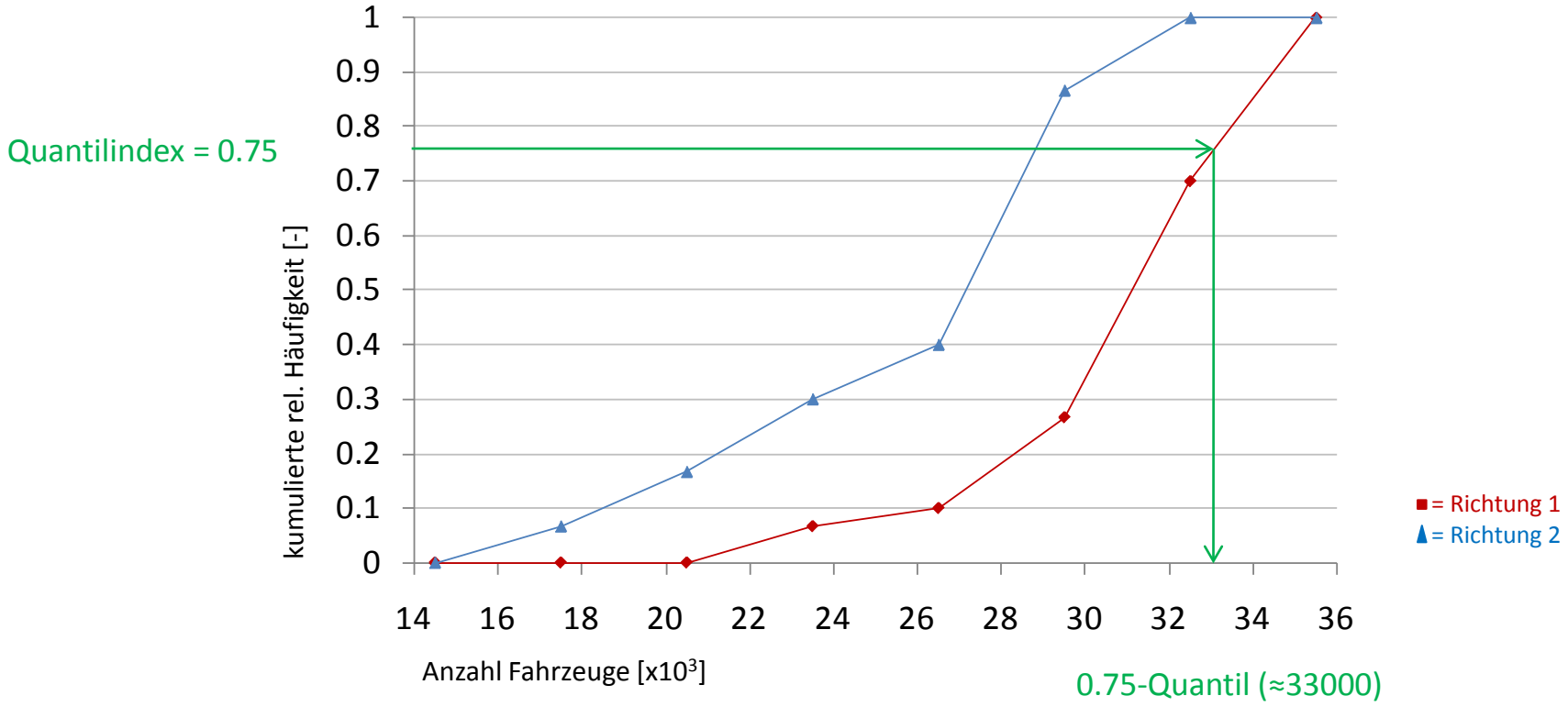
Somit beträgt der Median: $\frac{33198 + 33245}{2} = 33221.5$

Aufgabe C.2

Schritt 2

Schritte
1. Berechne den Median
2. ⇨ Berechne das 0.75- und 0.25- Quantil.
3. Berechne die Nachbarschaftswerte.
4. Ausreisser
5. Zeichne den Tukey Box Plot

Ermittlung der Quartil-Werte - veranschaulicht am 0.75-Quantil (oberes Quartil):



Aufgabe C.2

Schritt 2

Berechnung des Quantilindex und Ermittlung der Quantile:

$$v = \frac{i}{n+1}$$

n : Gesamtanzahl der Beobachtungen, $i = 1, 2, \ldots, n$

Richtung 1	i	i/31	
24846	1	0.03	
24862	2	0.06	
25365	3	0.10	
28252	4	0.13	
29224	5	0.16	
29976	6	0.19	
30035	7	0.23	
30613	8	0.26	
32158	9	0.29	
32472	10	0.32	
32618	11	0.35	
32962	12	0.39	
33091	13	0.42	
33197	14	0.45	
33198	15	0.48	
33245	16	0.52	
33380	17	0.55	
33406	18	0.58	
33788	19	0.61	
33888	20	0.65	
33937	21	0.68	
34007	22	0.71	
34013	23	0.74	← 75%
34076	24	0.77	
34425	25	0.81	
34455	26	0.84	
34576	27	0.87	
35237	28	0.90	
35843	29	0.94	
35852	30	0.97	

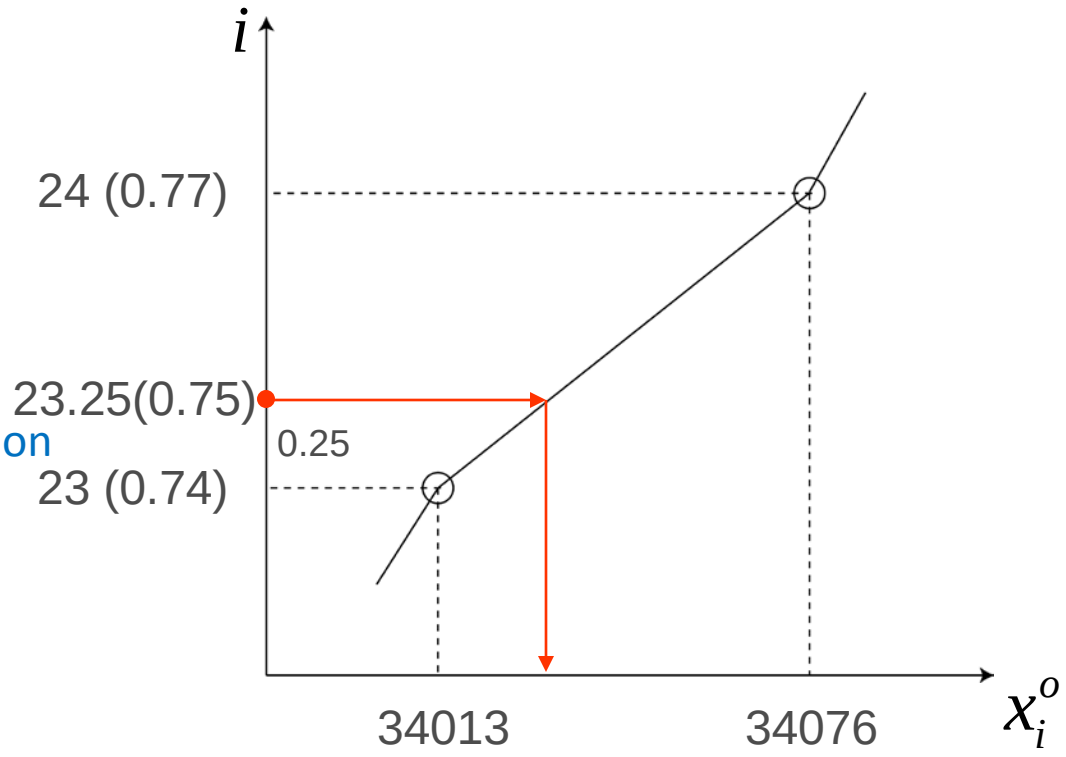
Aufgabe C.2

Schritt 2

Schritte
1. Berechne den Median
2. ⇨ Berechne das 0.75- und 0.25- Quantil.
3. Berechne die Nachbarschaftswerte.
4. Ausreisser
5. Zeichne den Tukey Box Plot

Beobachtungen (geordnet)	"Rang"	Quantilindex
x_i^o	i	υ
33788	19	0.61
33888	20	0.65
33937	21	0.68
34007	22	0.71
34013	23	0.74
34076	24	0.77
34425	25	0.81

⇨ Lineare Interpolation



$$\upsilon = \frac{i}{n+1} \quad \Rightarrow \quad i = \upsilon(n+1)$$

$$i = 0.75(30+1) = 23.25$$

$$x^o_{23.25} = (1-p)x^o_{23} + px^o_{23+1} = x^o_{23} + p(x^o_{24} - x^o_{23}) =$$
$$= 34013 + 0.25 \cdot (34076 - 34013) = 34028.75 \approx 34029 \text{ Autos}$$

Schritte
1. Berechne den Median
2. Berechne das 0.75- und 0.25- Quantil.
3. ⇨ Berechne die Nachbarschaftswerte.
4. Ausreisser
5. Zeichne den Tukey Box Plot

Aufgabe C.2

Schritt 3

Berechnung des oberen Nachbarschaftswertes

$$\left.\begin{aligned} Q_{0.75} &= 34029 \\ Q_{0.25} &= 30469 \end{aligned}\right\} \quad r \equiv Q_{0.75} - Q_{0.25} = 34029 - 30469 = 3560$$

interquartile Differenz

oberer Nachbarschaftswert: $\textit{grösster Wert} \leq (\textit{0.75-Quantil}) + 1.5 \cdot r$

In diesem Fall, oberer Grenzwert: $34029 + 1.5 \cdot 3560 = 39363$

33198
33245
33380
33406
33788
33888
33937
34007
34013
34076
34425
34455
34576
35237
35843
35852

Der grösste Wert der Datenreihe kleiner/gleich dem berechneten Grenzwert ist der *obere Nachbarschaftswert*.

oberer Nachbarschaftswert = 35852

Schritte
1. Berechne den Median
2. Berechne das 0.75- und 0.25- Quantil.
3. ⇨ Berechne die Nachbarschaftswerte.
4. Ausreisser
5. Zeichne den Tukey Box Plot

Aufgabe C.2

Schritt 3

Berechnung des unteren Nachbarschaftswertes

$$\left.\begin{aligned} Q_{0.75} &= 34029 \\ Q_{0.25} &= 30469 \end{aligned}\right\} \quad r \equiv Q_{0.75} - Q_{0.25} = 34029 - 30469 = 3560$$

unterer Nachbarschaftswert: *kleinster Wert* $\geq$ (*0.25-Quantil*) - $1.5 \cdot r$

In diesem Fall, unterer Grenzwert: $30469 - 1.5 \cdot 3560 = 25129$

Direction 1
24846
24862
25129 25365
28252
29224
29976
30035
30613
32158
32472
32618
32962
33091
33197
33198

Der kleinste Wert der Datenreihe grösser/gleich dem berechneten Grenzwert ist der *untere Nachbarschaftswert*.

unterer Nachbarschaftswert = 25365

Aufgabe C.2

Schritt 4

Schritte
1. Berechne den Median
2. Berechne das 0.75- und 0.25- Quantil.
3. Berechne die Nachbarschaftswerte.
4. ⇨ Ausreisser
5. Zeichne den Tukey Box Plot

Richtung 1	i	i/31
24846	1	0.03
24862	2	0.06
25365	3	0.10
28252	4	0.13
29224	5	0.16
29976	6	0.19
30035	7	0.23
30613	8	0.26
32158	9	0.29
32472	10	0.32
32618	11	0.35
32962	12	0.39
33091	13	0.42
33197	14	0.45
33198	15	0.48
33245	16	0.52
33380	17	0.55
33406	18	0.58
33788	19	0.61
33888	20	0.65
33937	21	0.68
34007	22	0.71
34013	23	0.74
34076	24	0.77
34425	25	0.81
34455	26	0.84
34576	27	0.87
35237	28	0.90
35843	29	0.94
35852	30	0.97

Ausreisser:

Ausserhalb der oberen und unteren Nachbarschaftswerte

24846

24862

Zusammenfassung:

oberer Nachbarschaftswert	35852
0.75-Quantil	34029
Median	33222
0.25- Quantil	30469
unterer Nachbarschaftswert	25365

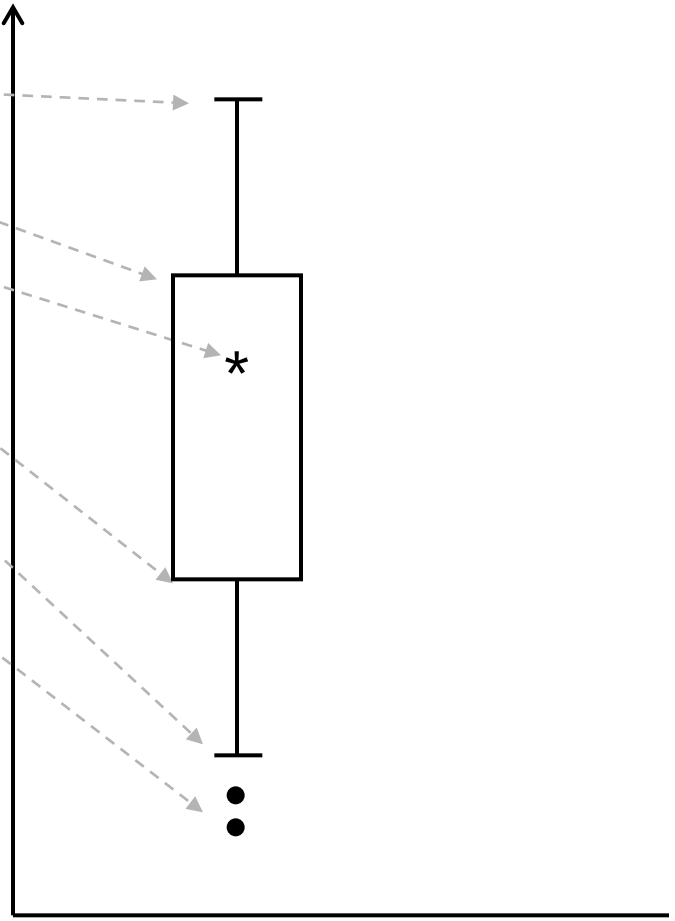
Aufgabe C.2

Schritt 5

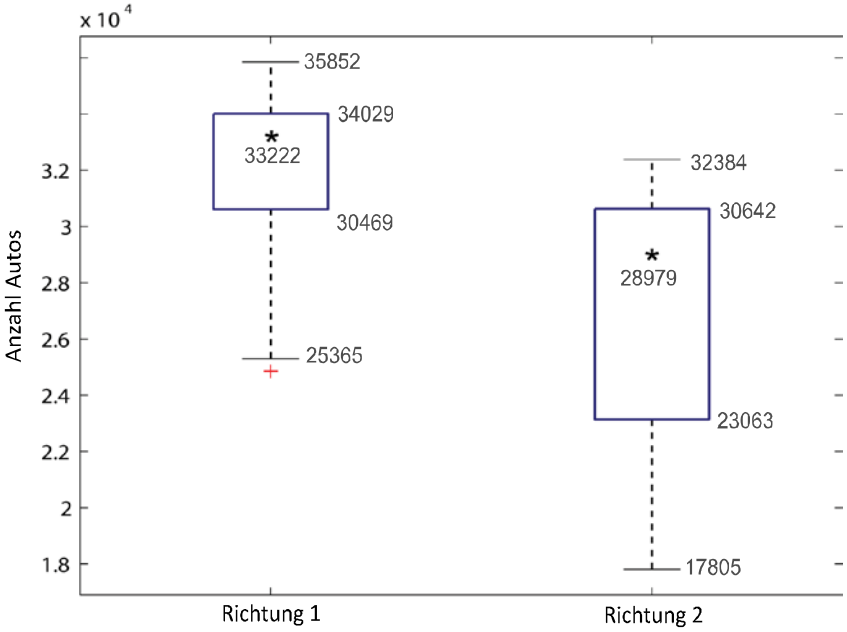
Schritte
1. Berechne den Median
2. Berechne das 0.75- und 0.25- Quantil.
3. Berechne die Nachbarschaftswerte.
4. Ausreisser
5. ⇨ Zeichne den Tukey Box Plot

erforderliche Kennwerte:

oberer Nachbarschaftswert	35852
0.75- Quantil	34029
Median	33222
0.25- Quantil	30469
unterer Nachbarschaftswert	25365
Ausreisser	24846
	24862



Aufgabe C.2 - Lösung



Kennwerte der jeweiligen Verteilung:

- Median
- Nachbarschaftswerte
- oberes und unteres Quartil (0,75- und 0.25-Quantil)
- Ausreisser

Vergleich der Datenreihen:

- Alle Kennwerte sind in Richtung1 grösser.
- Grösseres Verkehrsvolumen in Richtung 1.
- Grössere Interquartile Differenz in Richtung 2: Beobachtungen sind weiter gestreut um den Median.

Symmetrie der Datenreihen:

- keine Symmetrie beobachtet.
- Der Median ist bei beiden Datenreihen näher am oberen Nachbarschaftswert.
- linksschief.

Aufgabe C.5

Aus den in Tabelle C.5.1 gegebenen Daten ist die Korrelation zwischen der Anzahl an Studienanfängern X und der Gesamtzahl Studierender Y an einer Universität zu bestimmen. Benutze das angefügte Berechnungsblatt.

	Uni A	Uni B	Uni C	Uni D	Uni E	Uni F
Studienanfänger	3970	732	499	1300	3463	2643
Studentenzahl	24273	5883	2847	5358	23442	17076

Tabelle C.5.1 Anzahl Studienanfänger und Studentenzahl (gesamt)

Aufgabe C.5

Die Korrelation dieser Beobachtungen ist mit Hilfe des Berechnungsblattes zu bestimmen.

	Uni A	Uni B	Uni C	Uni D	Uni E	Uni F
Studienanfänger	3970	732	499	1300	3463	2643
Studentenzahl	24273	5883	2847	5358	23442	17076

Tabelle C.5.1 Anzahl Studienanfänger und Studentenzahl (gesamt)

Was ist bekannt?

Beobachtungen/Universitäten: $n = 6$
Studienanfänger: $x_i, i = 1,\ldots,6$
Studentenzahlen: $y_i, i = 1,\ldots,6$

Was wird gesucht?

Korrelationskoeffizient :

$$r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{x}_i - \overline{x})(\hat{y}_i - \overline{y})}{s_X s_Y}$$

Mittelwert: $\overline{x}$ $\overline{y}$

Standardabweichung: s_X s_Y

Aufgabe C.5

Bestimmen Sie die Korrelation dieser Zahlen mit Hilfe des Berechnungsblattes.

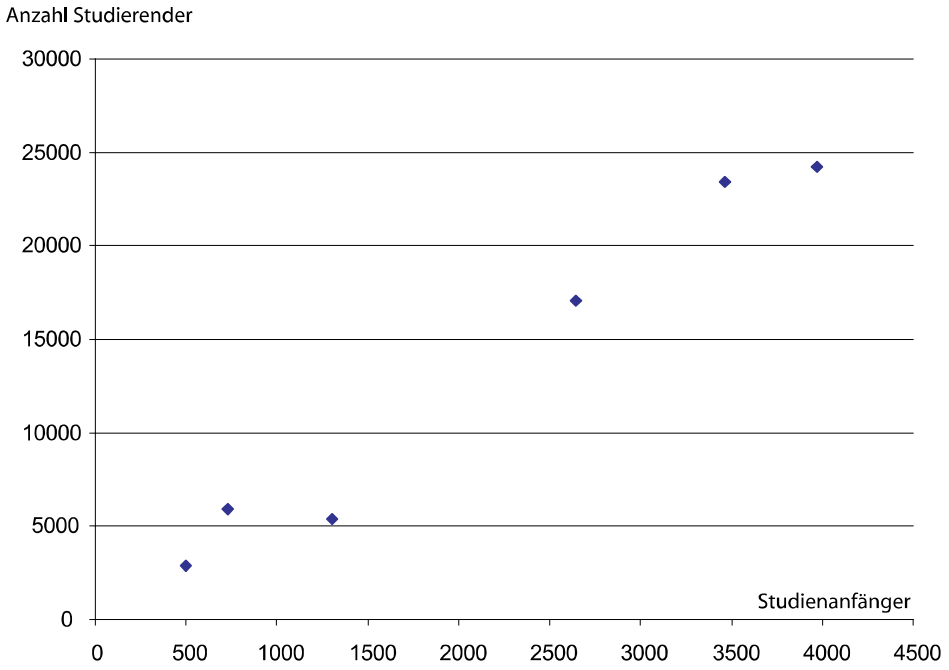
	Uni A	Uni B	Uni C	Uni D	Uni E	Uni F
Studienanfänger	3970	732	499	1300	3463	2643
Studentenzahl	24273	5883	2847	5358	23442	17076

Tabelle C.5.1 Anzahl Studienanfänger und Studentenzahl (gesamt)

Sind die Beobachtungen auf dem ersten Blick korreliert?

Gebe eine grobe Schätzung des Korrelationskoeffizienten.

$$-1 \le r_{XY} \le 1$$



Aufgabe C.5

	Uni A	Uni B	Uni C	Uni D	Uni E	Uni F
Studienanfänger	3970	732	499	1300	3463	2643
Studentenzahl	24273	5883	2847	5358	23442	17076

Tabelle C.5.1 Anzahl Studienanfänger und Studentenzahl (gesamt)

Was ist gesucht?

Korrelationskoeffizient der Stichprobe:

$$r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{x}_i - \overline{x})(\hat{y}_i - \overline{y})}{s_X s_Y}$$

Mittelwert der Stichprobe:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$$

Standardabweichung der Stichprobe:

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - \overline{x})^2} \qquad s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}$$

Aufgabe C.5 - Lösung

	Uni A	Uni B	Uni C	Uni D	Uni E	Uni F
Studienanfänger	3970	732	499	1300	3463	2643
Studentenzahl	24273	5883	2847	5358	23442	17076

Tabelle C.5.1 Anzahl Studienanfänger und Studentenzahl (gesamt)

	$\hat{x}_i$	$\hat{y}_i$	$\hat{x}_i - \overline{x}$	$\hat{y}_i - \overline{y}$	$(\hat{x}_i - \overline{x})^2$	$(\hat{y} - \overline{y})^2$	$(\hat{x}_i - \overline{x})(\hat{y}_i - \overline{y})$
A	3970	24273	1868	11126	3493161	123787876	20793574
B	732	5883	-1369	-7264	1874161	52765696	9944942
C	499	2847	-1602	-10300	2566404	106090000	16501516
D	1300	5358	-801	-7789	641601	60668521	6239887
E	3463	23442	1362	10295	1855044	105987025	14020755
F	2643	17076	542	3929	293764	15437041	2129134
$\sum$	12607	78879	-	-	10724135	464736159	69629807
$\sum / n$	2101	13147	-	-	1787356	77456026.5	11604968
$\sqrt{\sum / n}$	-	-	-	-	1337	8801	-

Aufgabe C.5 - Lösung

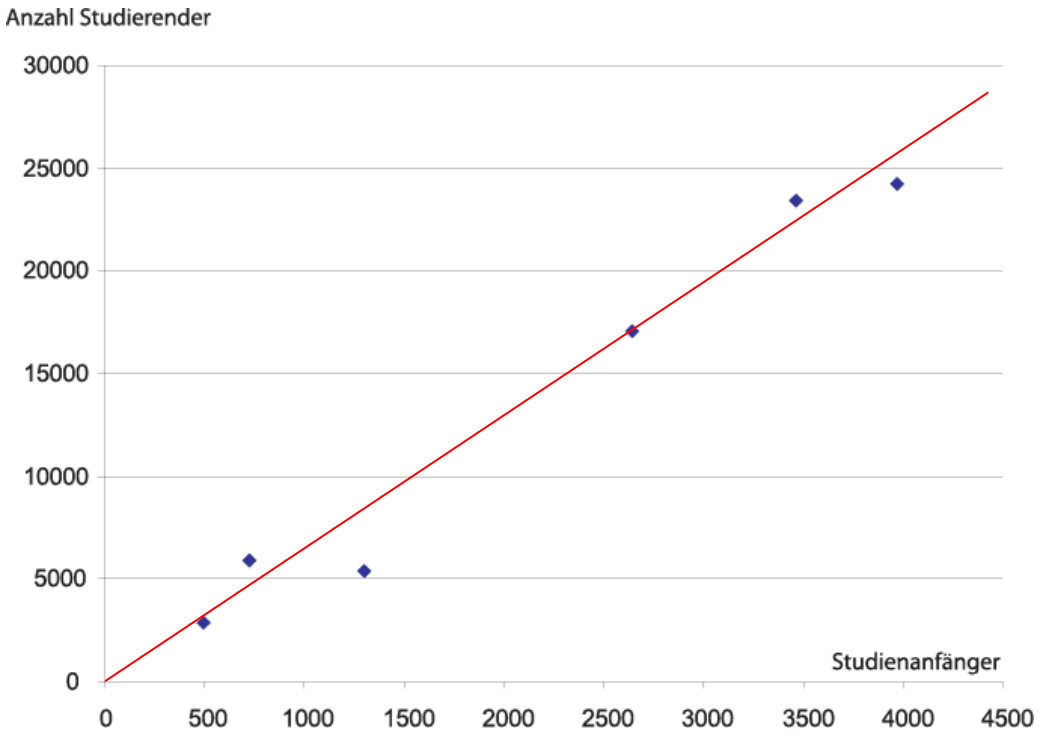
$$r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y})}{s_X s_Y} = \frac{11604968}{1337 \cdot 8801} = 0.99$$

	$\hat{x}_i$	$\hat{y}_i$	$\hat{x}_i - \bar{x}$	$\hat{y}_i - \bar{y}$	$(\hat{x}_i - \bar{x})^2$	$(\hat{y} - \bar{y})^2$	$(\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y})$
A	3970	24273	1868	11126	3493161	123787876	20793574
B	732	5883	-1369	-7264	1874161	52765696	9944942
C	499	2847	-1602	-10300	2566404	106090000	16501516
D	1300	5358	-801	-7789	641601	60668521	6239887
E	3463	23442	1362	10295	1855044	105987025	14020755
F	2643	17076	542	3929	293764	15437041	2129134
$\sum$	12607	78879	-	-	10724135	464736159	69629807
$\sum / n$	2101	13147	-	-	1787356	77456026.5	11604968
$\sqrt{\sum / n}$	-	-	-	-	1337	8801	-

Aufgabe C.5 - Lösung

$$r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y})}{s_X s_Y} = \frac{11604968}{1337 \cdot 8801} = 0.99$$

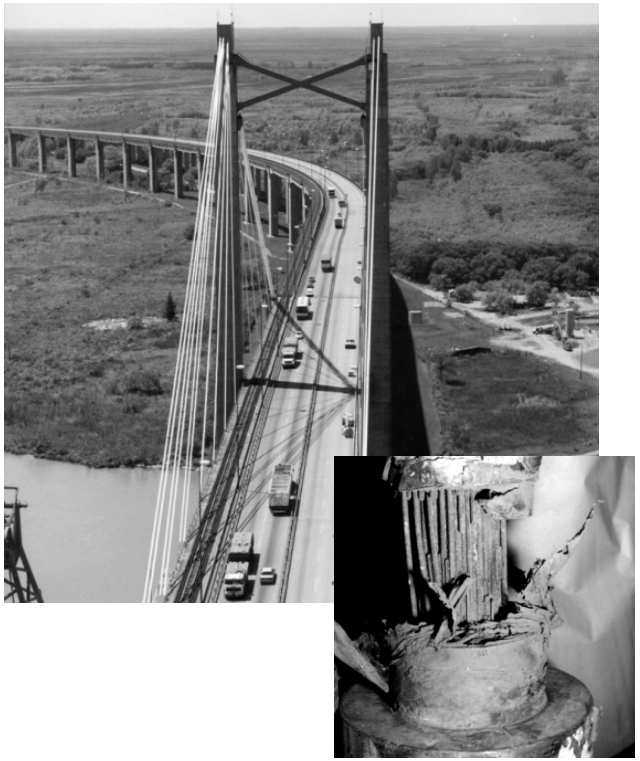
Wie erwartet ist der Korrelationskoeffizient positiv.



Aufgabe C.4 (Hausübung)

Potentialfeldmessungen helfen dabei, die mögliche Korrosion in Brückentragwerken vorherzusagen. Während einer routinemässigen Untersuchung an einer Brücke wurden die Daten in folgender Tabelle durch Potentialfeldmessungen entlang der beiden Fahrspuren (Richtung 1 und 2) erhoben:

Messung Nr. (i)	Richtung 1 Widerstand (kOhm)	Richtung 2 Widerstand (kOhm)
1	20.2	3.8
2	20.4	5.6
3	22.1	6.5
4	23.8	7.1
5	24.3	7.9
6	24.7	8.2
7	25.3	9.1
8	25.6	9.3
9	25.7	9.6
10	25.9	9.8
11	26.2	10.3
12	26.7	10.9
13	26.9	11.1
14	27.3	11.7
15	27.6	12.2
16	27.6	12.6
17	27.8	12.9
18	27.9	13.8
19	28.3	13.9
20	28.7	14.5
21	28.9	15
22	28.9	15.4
23	29.3	17.1
24	29.4	17.8
25	29.9	23.4



Aufgabe C.4 (Hausübung)

a) Nutze die beiden Datenreihen aus der Tabelle und fertige zwei Tukey Box Plots an (Richtung 1 und 2). Zeige die Hauptmerkmale der Tukey Box Plots und schreibe deren Werte neben die korrespondierenden Punkte auf das Diagramm. Zeichne auch vorhandene Werte, die ausserhalb liegen ein.

b) Der Tukey Box Plot ist ein hilfreiches Werkzeug zur Bewertung der Symmetrie von Datenreihen. Diskutiere Symmetrie/Schiefe der Potentialfeldmessdaten der beiden Fahrspuren.

c) Wähle eine geeignete Anzahl von Intervallen und zeichne ein Histogramm für die Potentialfeldmessdaten von Richtung 1.

d) Viel Erfolg ;-)