Übungsblatt 1 - mit Lösungen

Beschreibende Statistik

Stochastik@AIN2
Prof. Dr. Barbara Staehle
Wintersemester 2021/2022
HTWG Konstanz

|Einfache und mittelschwere Aufgaben|

Aufgabe 1.1 Eigenschaften von Merkmalen, 3 Punkte

Geben Sie für jedes Merkmal an, ob es quantitativ diskret, quantitativ stetig, qualitativ ordinal oder qualitativ nominal ist.

a) Anzahl der Insassen in einem PKW bei der Verkehrszählung

b) Reisegeschwindigkeit eines Flugzeugen

c) Schultypen

d) Temperaturangaben in °C

e) Zugriffszeiten auf die Daten einer Festplatte

f) Fassungsvermögen eines Binnenschiffs

g) Energieeffizienz von Elektrogeräten

h) Ölverbrauch eines Einfamilienhauses pro Jahr

i) Anzahl von Likes eines Instagram-Posts

j) Nationalität von Studierenden

k) Intelligenzquotient

l) Wohnort von Studierenden

Lösung

a) Anzahl der Insassen in einem PKW bei der Verkehrszählung: quantitativ, diskret

b) Reisegeschwindigkeit eines Flugzeugen: quantitativ, stetig

c) Schultypen: qualitativ, nominal (Schultypen lassen sich nicht anordnen)

d) Temperaturangaben in ° C: quantitativ, stetig

e) Zugriffszeiten auf die Daten einer Festplatte: quantitativ, stetig

f) Fassungsvermögen eines Binnenschiffs: quantitativ, stetig

g) Energieeffizienz von Elektrogeräten: qualitativ, ordinal (A ist besser als E)

h) Ölverbrauch eines Einfamilienhauses pro Jahr: quantitativ, stetig

i) Beurteilung eines Instagram-Posts (Like): qualitativ, ordinal (Like ist besser als kein Like)

j) Nationalität von Studierenden: qualitativ, nominal (Nationalitäten lassen sich nicht anordnen)

k) Intelligenzquotient: qualitativ, ordinal (Zahlen werden als Bewertung verwendet, Streitfall; auch quantitativ, diskret)

l) Wohnort von Studierenden qualitativ, nominal (Wohnorte lassen sich nicht anordnen)

Aufgabe 1.2 TikTok-Videos

Sie interessieren sich für die Frage, wie ein typisches TikTok-Video aussieht und führen eine statistische Stichprobenerhebung durch.

Teilaufgabe 1.2.1 3 Punkte

Geben Sie **Beispiele** an, was für diese Untersuchung die folgenden Kenngrößen sein könnten:

- Statistische Einheiten
- Grundgesamtheit
- Stichprobe
- Merkmale (mindestens 3)
- **Beispielhafte** Merkmalsausprägungen (jeweils für Ihre Beispiele)
- Eigenschaften der Merkmale (qualitativ ordinal/nominal, quantitativ diskret/kontinuierlich) (jeweils für Ihre Beispiele)

Lösung

Beispielhafte Lösung, andere Ideen gut vorstellbar

- Statistische Einheiten: TikTok-Videos
- Grundgesamtheit: alle auf TikTok existierenden Videos
- Stichprobe: 100 Videos, die Sie durch eine Zufallssuche finden
- Merkmale:
  - Länge [sec]
  - Größe [MB]
  - live [j/n]
  - Kategorie
  - Account
- Merkmalsausprägungen:
  - Länge im Intervall 3-60 sec
  - Größe im Intervall 0.1-10 MB (?)
  - ja oder nein

  - Performance, Entertainment, Sports & Outdoors, ...
  - Charli D'Amelio, Addison Rae, ...

- Eigenschaften der Merkmale
  - Länge: quantitativ, kontinuierlich
  - Größe: quantitativ, kontinuierlich
  - live: qualitativ, ordinal
  - Kategorie: qualitativ, nominal
  - Account: qualitativ, nominal

TEILAUFGABE 1.2.2 1 PUNKT

Welche Eigenschaften von TikTok-Videos finden Sie besonders spannend und einer näheren Untersuchung wert? Begründen Sie Ihre Meinung.

LÖSUNG

Keine richtige oder falsche Lösung. Vorstellbar sind sicher die genauere Analyse der Längen der Videos, der Codierung, der Größe in Bytes, Zusammenhänge zwischen Art des Videos und dessen Länge, oder Downloadanzahlen, ...
Begründung ist dann logischerweise abhängig vom Merkmal.

AUFGABE 1.3 WOHNUNGSGRÖSSEN

10 Personen aus Konstanz wurden nach der Anzahl der Zimmer in ihrer Wohnung gefragt. Es ergab sich folgende Urliste:

$$3, 4, 5, 1, 5, 2, 1, 3, 1, 3$$

TEILAUFGABE 1.3.1 2 PUNKTE

Geben Sie die relative und absolute Häufigkeiten der genannten Zimmeranzahlen an.

LÖSUNG

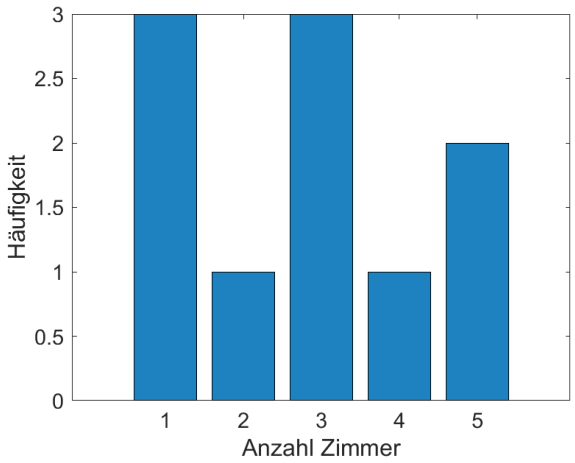
$n = 10$ (Größe der Stichprobe)

Zimmeranzahl z_i	1	2	3	4	5
$h(z_i)$	3	1	3	1	2
$f(z_i)$	0.3	0.1	0.3	0.1	0.2

TEILAUFGABE 1.3.2 2 PUNKTE

Zeichnen Sie ein Histogramm für die absoluten Häufigkeiten.

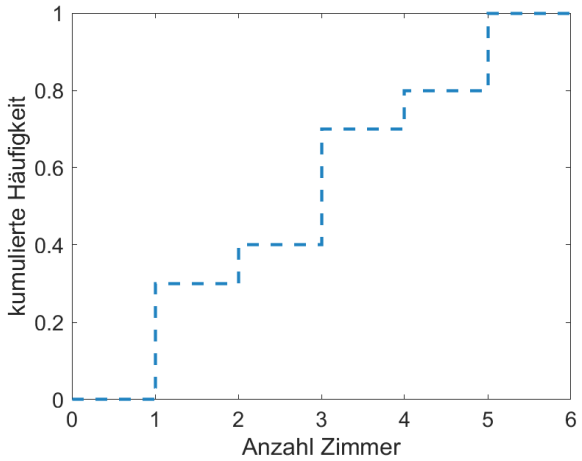
LÖSUNG



TEILAUFGABE 1.3.3 2 PUNKTE

Zeichnen Sie ein Diagramm für die empirische Verteilungsfunktion.

LÖSUNG



TEILAUFGABE 1.3.4 1 PUNKT

Geben Sie an, wie viele Prozent der Befragten

a) maximal 4 Zimmer zur Verfügung haben.

b) mindestens 2 und maximal 4 Zimmer zur Verfügung haben.

LÖSUNG

a) $f_1 + f_2 + f_3 + f_4 = 0.8 = 80\%$ der befragten Personen.

b) $f_2 + f_3 + f_4 = 0.5 = 50\%$ der befragten Personen.

TEILAUFGABE 1.3.5 5 PUNKTE

Berechnen Sie für die Stichprobe (der Zimmergrößen):

a) das arithmetische Mittel

b) den Median

c) den Modalwert

d) das 25%-Quantil

e) das 30%-Quantil

f) das 66%-Quantil

g) die Varianz

h) die Standardabweichung

i) den Interquartilabstand

j) die Spannweite

LÖSUNG

Sortierte Liste der Stichprobe:

i	1	2	3	4	5	6	7	8	9	10
x_i	1	1	1	2	3	3	3	4	5	5

a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{28}{10} = 2.8$

b) $\tilde{x} = \frac{1}{2}(x_6 + x_5) = \frac{1}{2}(3+3) = 3$

c) Es gibt zwei Modalwerte: $x_m = \{3, 1\}$ (da $h(3) = h(1)$ am größten ist)

d) $0.25 \cdot 10 = 2.5 \Rightarrow \tilde{x}_{0.25} = x_{\lceil 2.5 \rceil} = x_3 = 1$

e) $0.3 \cdot 10 = 3 \Rightarrow \tilde{x}_{0.3} = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(1+2) = 1.5$

f) $0.66 \cdot 10 = 6.6 \Rightarrow \tilde{x}_{0.66} = x_{\lceil 6.6 \rceil} = x_7 = 3$

g) $s^2 = \frac{1}{n-1}(\sum_{i=1}^{k} h(z_i) z_i^2 - n\bar{x}^2) = \frac{1}{9}(3 \cdot 1 + 1 \cdot 4 + 3 \cdot 9 + 1 \cdot 16 + 2 \cdot 25 - 10 \cdot 7.84) = \frac{12}{5} = 2.4$

h) $s = \sqrt{s^2} = 1.549$

i) $I = \tilde{x}_{0.75} - \tilde{x}_{0.25}$
$0.75 \cdot 10 = 7.5 \Rightarrow \tilde{x}_{0.75} = x_{\lfloor 7.56 \rfloor + 1} = x_8 = 4$
$I = \tilde{x}_{0.75} - \tilde{x}_{0.25} = 4 - 1 = 3$

j) $R = max(x) - min(x) = 5 - 1 = 4$

Teilaufgabe 1.3.6 1 Punkt

In München werden 20 anderen Personen ebenfalls nach der Anzahl der Zimmer in ihrer Wohnung gefragt. Die Werte in dieser Stichprobe ergeben den gleichen Mittelwert wie die Ergebnisse aus Konstanz, allerdings ergeben diese eine **höhere Varianz**.

Was schließen Sie hieraus hinsichtlich des Unterschieds zwischen den beiden Stichproben?

Lösung

Bei gleichem Mittelwert und größerer Varianz weichen die Wohnungsgrößen in München stärker vom Mittelwert ab als in Konstanz, es gibt also mehr Wohnungen mit sehr vielen (5,6, ...) oder nur 1-2 Zimmern.

Aufgabe 1.4 Geldscheine, 2 Punkte

Angenommen, es wären von allen vorhandenen Werten (5 €, 10 €, 20 €, 50 €, 100 €, 200 €, 500 €) gleich viele Geldscheine im Umlauf.

a) Bestimmen Sie den Median des Werts der Geldscheine im Umlauf.

b) Berechnen Sie den Median, der sich nach der Einführung eines 1000 €-Scheins (und davon ebenfalls so viele in den Umlauf gebracht wie alle anderen) ergeben würde.

Lösung

a) $n = 7 = 2 \cdot 3 + 1 \Rightarrow \tilde{x} = x_{3+1} = x_4 = 50\,€$

b) $n = 8 = 2 \cdot 4 \Rightarrow \tilde{x} = \frac{1}{2}(x_4 + x_{4+1}) = \frac{1}{2}(50 + 100) = 75\,€$

AUFGABE 1.5 AN DER FAHRRADBRÜCKE (KLAUSURAUFGABE WS 16/17), 4 PUNKTE

Alice steht an der Konstanzer Fahrradbrücke und zählt die Anzahl der Räder der Fahrzeuge (Einräder, Fahrräder, Dreiräder, Kinderwagen, ...) welche an ihr vorbei rollen.
Ihre Beobachtungen resultieren in folgender Urliste:

$$2, 4, 3, 1, 2, 4, 2, 2, 2, 3.$$

Berechnen Sie für die Beobachtungen von Alice:

a) das arithmetische Mittel

b) den Median

c) den Modalwert

d) das 10%-Quantil

e) das 25%-Quantil

f) das 75%-Quantil

g) den Interquartilabstand

h) die Spannweite

Geben Sie jeweils an, wie Sie die geforderten Charakteristika berechnet haben!

LÖSUNG

Zähle: $n = 10$. Sortierte Liste der Stichprobe:

i	1	2	3	4	5	6	7	8	9	10
x_i	1	2	2	2	2	2	3	3	4	4

a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{25}{10} = 2.5$

b) $\tilde{x} = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(2+2) = 2$

c) $x_m = 2$ (da 2 am meisten vorkommt)

d) $0.1 \cdot 10 = 1 \Rightarrow \tilde{x}_{0.1} = \frac{1}{2}(x_1 + x_2) = \frac{1}{2}(1+2) = 1.5$

e) $0.25 \cdot 10 = 2.5 \Rightarrow \tilde{x}_{0.25} = x_{\lceil 2.5 \rceil} = x_3 = 2$

f) $0.75 \cdot 10 = 7.5 \Rightarrow \tilde{x}_{0.75} = x_{\lceil 7.5 \rceil} = x_8 = 3$

g) $R = max(x) - min(x) = 4 - 1 = 3$

h) $IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25} = 3 - 2 = 1$

AUFGABE 1.6 TASCHENGELD (KLAUSURAUFGABE SS 18)

Frank schreibt für seine Schülerzeitung einen Artikel zum Thema „Taschengeld eines Achtklässlers“. Hierfür befragt er 6 zufällig ausgewählte Schüler der Klasse 8a nach ihrem monatlichen Taschengeld in Euro. Seine Umfrage ergibt die folgende Urliste:

$$25, 10, 30, 25, 35, 25$$

TEILAUFGABE 1.6.1 2 PUNKTE

Berechnen Sie für Frank's Urliste folgende Größen. Stellen Sie Ihren Rechenweg nachvollziehbar dar!

a) Arithmetisches Mittel

b) Median

c) 75%-Quantil

d) empirische Standardabweichung

LÖSUNG

$n = 6$. Sortierte Liste der Stichprobe:

i	1	2	3	4	5	6
x_i	10	25	25	25	30	35

a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{150}{6} = 25$

b) $\tilde{x} = \frac{x_3 + x_4}{2} = 25$

c) $0.75 \cdot 6 = 4.5 \Rightarrow \tilde{x}_{0.75} = x_{\lceil 4.5 \rceil} = x_5 = 30$

d) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{5}((10-25)^2 + 3\cdot(25-25)^2 + (30-25)^2 + (35-25)^2) = \frac{1}{5}(225+0+25+100) = \frac{350}{5} = 70$
$s = \sqrt{s^2} = \sqrt{70} = 8.3666$

TEILAUFGABE 1.6.2 3 PUNKTE

Beantworten Sie folgende Fragen möglichst allgemeinverständlich!

a) Frank berechnet außerdem noch die Spannweite seiner Stichprobe zu $R = 5$ und den Interquartilabstand zu $I = 25$. Machen diese Berechnungen Sinn? Begründen Sie Ihre Meinung!

b) Mittelwert und Median von Franks Daten sind gleich. Ist das immer so, oder ist das Zufall?

c) Was hat der Wert des 75%-Quantils zu bedeuten? Fassen Sie Ihr Ergebnis in Worte.

d) Frank führt seine Umfragen nach dem monatlichen Taschengeld noch in zwei anderen Klassen, den Klassen 8b und 8c, durch. In beiden Fällen erhält er das selbe arithmetische Mittel wie bei Klasse 8a. Im Fall von Klasse 8b erhält er allerdings eine Standardabweichung von 0, im Fall von Klasse 8c erhält er eine Standardabweichung, die deutlich größer ist als die der Klasse 8a. Was schließen Sie hieraus hinsichtlich der Stichproben des monatlichen Taschengelds aus den anderen beiden Klassen?

e) Frank interessiert sich weiterhin für die Schuhgröße der Befragungsteilnehmer, sowie für ihre Mathenote. Er berechnet hierfür die Korrelationskoeffizienten von Schuhgröße und Taschengeld zu $r_{S,T} = 0.875$ und von Mathenote und Taschengeld zu $r_{M,T} = -0.038$. Was schließen Sie hieraus über den Zusammenhang der betrachteten Größen?

f) Grace befragt 3 Schüler nach ihrem Taschengeld und berechnet den arithmetische Mittelwert dieser Stichprobe zu 300 €. Beschreibt diese Zahl die Stichprobe ausreichend gut, oder sollte sie besser noch andere Werte berechnen?

LÖSUNG

a) Quatsch, IQR muss immer kleiner gleich Range sein

b) Es ist Zufalls, dass Mittelwert und Median gleich sind, die weichen oft voneinander ab, vor allem für asymmetrische Werteverteilungen. Siehe Beispiel in letzter Teilaufgabe.

c) 75% der Schüler kriegen 30 € oder weniger Taschengeld.
Andere mögliche Lösung: 25% der Schüler bekommen 30 € oder mehr Taschengeld.

d) 8b: alle kriegen gleich viel Taschengeld (25), 8c: Werte schwanken stärker, es gibt also mehr Kinder die deutlich unter bzw. über 25 € kriegen.

e)
- Schuhgröße und Taschengeld sind (deutlich) positiv korreliert, heißt dass ein Schüler mit größeren Füßen tendenziell mehr Taschengeld bekommt.
- Mathenote und Taschengeld sind (sehr schwach) negativ korreliert bis eher unabhängig, heißt dass man von der Mathenote nicht aufs Taschengeld rückschließen kann.

f) Wenn Mittelwert 300 ist, ein Wert 890, könnte die Stichprobe z.B. 5,5,890 sein. Daher hilft es, sich z.B. Median und Modalwert, beide 5, anzusehen.

AUFGABE 1.7 SPORTWAGEN (KLAUSURAUFGABE WS 17/18)

Charlie besucht die IAA (Internationale Automobilausstellung). Er interessiert sich für Sportwagen und sucht daher Informationen zu den Höchstgeschwindigkeiten der vorgestellten Autos.

Am Stand von Hersteller 1 analysiert er die Datenblätter von 10 verschiedenen Fahrzeugen, denen er folgende Höchstgeschwindigkeiten entnimmt:

$$250, 330, 250, 350, 270, 300, 260, 330, 200, 400$$

TEILAUFGABE 1.7.1 2 PUNKTE

Berechnen Sie für die Stichprobe der von Charlie ermittelten Höchstgeschwindigkeiten folgende Größen. Stellen Sie Ihren Rechenweg nachvollziehbar dar!

a) Arithmetisches Mittel

b) Median

c) 90%-Quantil

d) Standardabweichung

LÖSUNG

$n = 10$. Sortierte Liste der Stichprobe:

i	1	2	3	4	5	6	7	8	9	10
x_i	200	250	250	260	270	300	330	330	350	400

a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{2940}{10} = 294$

b) $\tilde{x} = \frac{x_5 + x_6}{2} = \frac{270+300}{2} = 285$

c) $0.9 \cdot 10 = 9 \Rightarrow \tilde{x}_{0.9} = \frac{x_9 + x_10}{2} = \frac{350+400}{2} = 375$

d) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}(\ldots) = 349.333$
$s = \sqrt{s^2} = 59.104$

TEILAUFGABE 1.7.2 3 PUNKTE

Gefragt sind Beschreibungen in Ihren eigenen Worten (Zielgruppe: nicht-technisches Publikum).

a) Nennen Sie **zwei** Fakten, wie die von Ihnen berechneten Größen einzeln oder gemeinsam betrachtet die von Charlie erhobene Stichprobe der Höchstgeschwindigkeiten charakterisieren.

b) Charlie entnimmt aus einem Datenblatt von Hersteller 2 ebenfalls die Höchstgeschwindigkeiten von 10 verschiedenen Fahrzeugen und berechnet den Mittelwert dieser Höchstgeschwindigekeiten zu 294 sowie die Standardabweichung zu 20. Was schließen Sie hieraus hinsichtlich des Unterschieds zwischen den Fahrzeugen der beiden Hersteller?

c) Charlie betrachtet weiterhin die Preise der Reifen welche für die Fahrzeuge von Hersteller 1 zugelassen sind, sowie die Umweltfreundlichkeit des Fahrzeugs auf einer Skala von 0 (gar nicht) bis 10 (sehr).
Er berechnet hierfür die Korrelationskoeffizienten von Höchstgeschwindigkeit und Reifenpreis zu $r_{G,R} = 0.853$ und von Höchstgeschwindigkeit und Umweltfreundlichkeit zu $r_{G,U} = -1.137$. Was schließen Sie hieraus über den Zusammenhang der betrachteten Größen?

LÖSUNG

a)
- Mittelwert ist deutlich größer als der Median, es gibt also unverhältnismäßige Ausreiser die den Wert nach oben ziehen.
- Vor allem Vergleich von Mittelwert / Median und 90%-Quantil zeigt, dass es sehr viel schnellere Werte als die im Mittel gibt.
- Standardabweichung $\approx 60 \Rightarrow$ im Mittel weicht die Geschwindigkeit der Autos ungefähr 60 kmh vom Mittelwert ab.

b) Mittelwert und Stichprobegröße sind gleich, die Standardabweichung ist kleiner, daher weichen die Höchstgeschwindigkeiten weniger stark vom Mittelwert ab, die Fahrzeuge von Hersteller 2 sind also in sich ähnlicher (schnell) wie die von Hersteller 1.

c)
- Geschwindigkeit und Reifenpreise sind positiv korreliert, heißt dass ein schnelleres Auto mit einer hohen Wahrscheinlichkeit auch teurere Reifen braucht.
- Der Korrelationskoeffizient von Geschwindigkeit und Umweltfreundlichkeit wäre rund um -1 zu erwarten, kann aber nicht kleiner als -1 werden, daher ist der Wert Quatsch.

|Mittelschwere und schwere Aufgaben|

Aufgabe 1.8 Stichprobe mit Klassen (elektronisch lösen!), 3 Punkte

Gegeben ist folgende Urliste einer Stichprobe vom Umfang $n = 20$:

$$3, 7, 12, 18, 19, 20, 25, 25, 27, 28, 29, 31, 32, 34, 37, 38, 40, 41, 45, 47$$

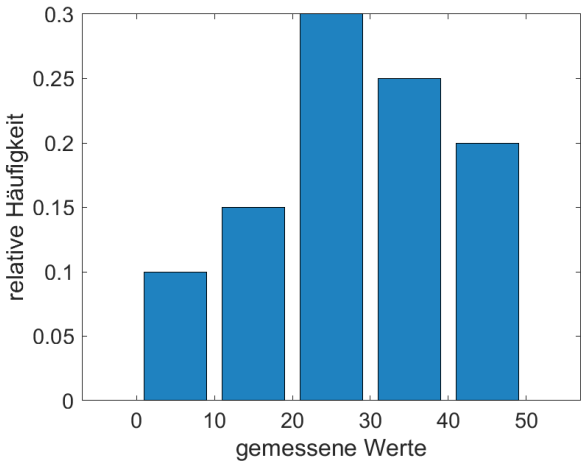
a) Gruppieren Sie die Stichprobenwerte in geeignete Klassen und bestimmen Sie die absoluten und die relativen Häufigkeiten der Klassen.

b) Stellen Sie die für die Klassen erhaltenen absoluten Häufigkeiten als Histogramm dar.

Lösung

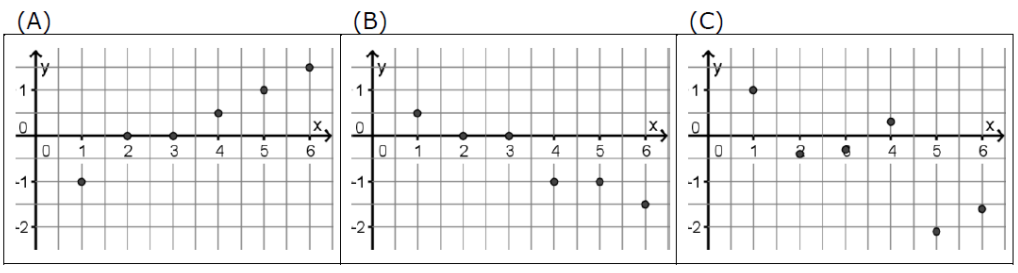
- Die Klassen müssen alle vorkommenden Stichprobenwerte (liegen zwischen 3 und 47) überdecken. Wähle z.B. die Intervalle $[0;10), [10;20), [20;30), [30;40), [40;50)$ als Klassen.
- Da nur $3, 7 \leq 10$ gilt $h_1 = 2$, analog $h_2 = 3, \ldots$
- Alle absoluten und relativen Häufigkeiten zusammengefasst:

Klasse	$[0, 10)$	$[10, 20)$	$[20, 30)$	$[30, 40)$	$[40, 50)$
Strichliste	||	|||	𝍸|	𝍸	||||
absolute Häufigkeit h_i	2	3	6	5	4
relative Häufigkeit f_i	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{1}{4}$	$\frac{1}{5}$



Aufgabe 1.9 Punktwolken

Teilaufgabe 1.9.1 3 Punkte



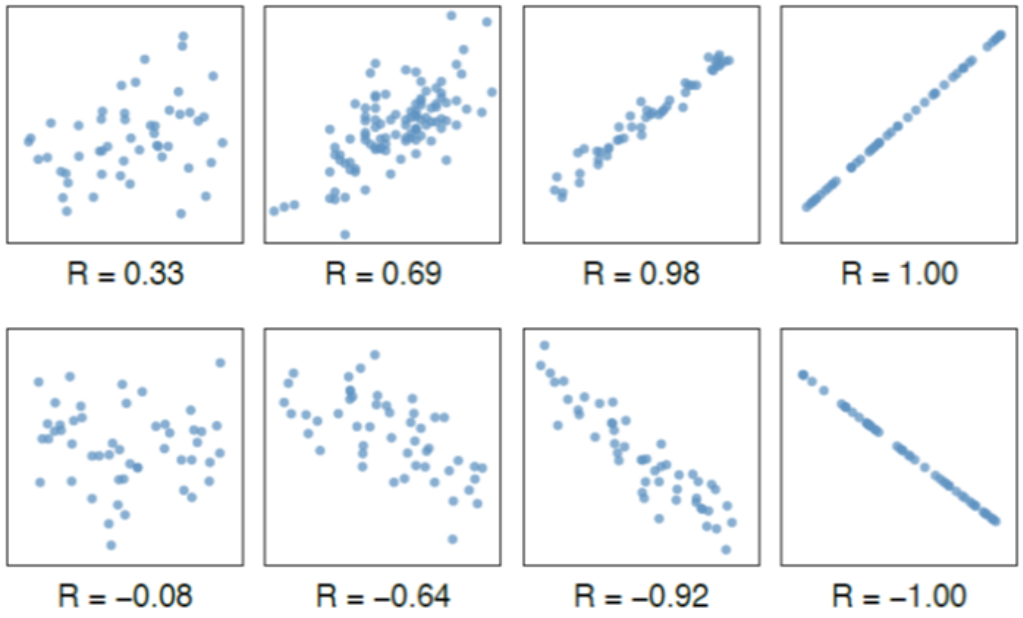
Bei welcher der dargestellten Punktwolken (A), (B) oder (C) erhält man als Korrelationskoeffizienten $r_{x,y} = -0,966$? Begründen Sie Ihre Meinung und nennen Sie für die beiden anderen Punktwolken jeweils einen Ablehnungsgrund.

LÖSUNG

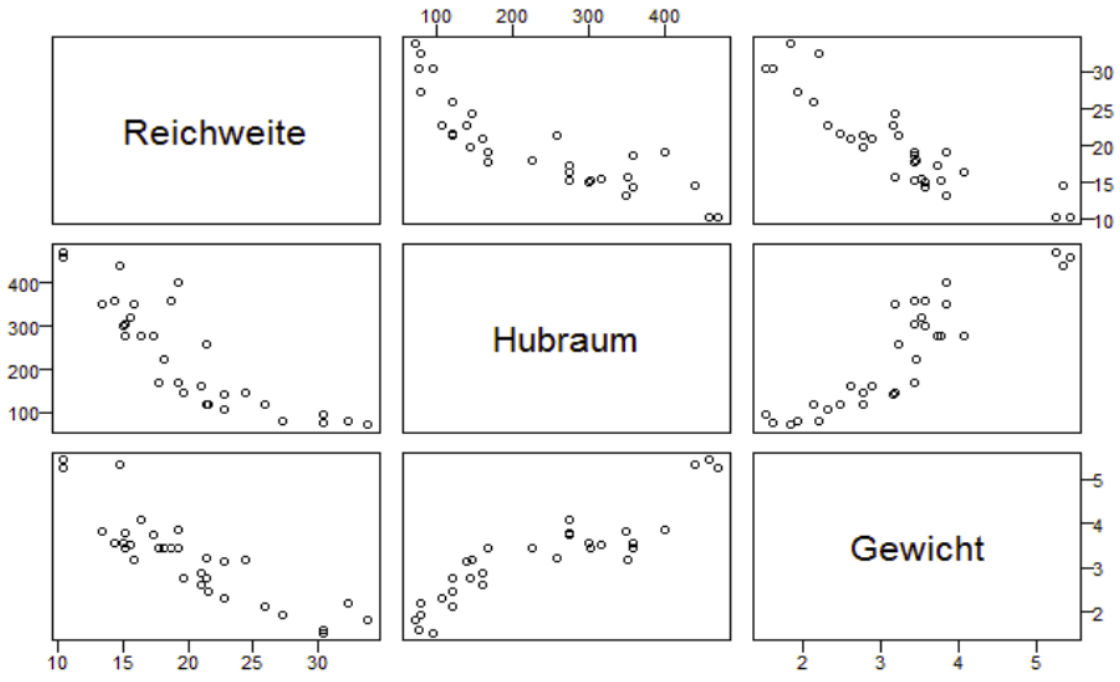
Bei Punktwolke B erhält man $r_{x,y} = -0,966$, da der Trend erkennbar linear negativ ist. Punktwolke A hat eine aufsteigende Trendlinie; deshalb wäre $r_{x,y} > 0$. Bei Punktwolke C ist die Streuung der Punkte größer; deshalb wäre $|r_{x,y}| < 0,966$, also weniger stark korreliert.

TEILAUFGABE 1.9.2 3 PUNKTE

LÖSUNG



TEILAUFGABE 1.9.3 3 PUNKTE



Die abgebildete Matrix stellt Punktwolken für einige Eigenschaften einer Auswahl verschiedener Autotypen des gleichen Herstellers dar. Der Plot oben rechts z.B. zeigt auf der x-Achse den Faktor Gewicht, auf der y-Achse den Faktor Reichweite. Ein Punkt entspricht einem Auto, geeignete Einheiten fehlen, diese dürfen Sie sich selbst ausdenken.

Welche Aussagen über die Zusammenhänge zwischen diesen drei Faktoren lassen sich aus der Grafik ableiten?

Lösung

- Reichweite und Hubraum sind negativ korreliert, ein Auto mit großem Hubraum hat mit einer großen Wahrscheinlichkeit eine kleine Reichweite, je größer der Hubraum desto kleiner ist tendenziell die Reichweite.
- Reichweite und Gewicht sind negativ korreliert, ein schweres Auto hat mit einer großen Wahrscheinlichkeit eine kleine Reichweite, je schwerer das Auto desto kleiner ist tendenziell die Reichweite.
- Gewicht und Hubraum sind positiv korreliert, ein Auto mit großem Hubraum hat mit einer großen Wahrscheinlichkeit ein hohes Gewicht, je größer der Hubraum desto schwerer ist tendenziell das Auto.
- Korrelationen sind symmetrisch, diese 3 Zusammenhänge sind die einzig dargestellten.
- Welche der Korrelationen die stärkste ist, ist schwierig zu beurteilen, da die Darstellungen nicht normiert sind.

AUFGABE 1.10 KLAUSURNOTEN (KLAUSURAUFGABE SS19)

Prof. Schmidt liest für alle Studierenden des Studiengangs „Bachelor Super-Informatik“ die Vorlesungen „Mathematik 1“ und „Mathematik 2“.
Um die Noten der beiden Vorlesungen vergleichen, analysiert sie die Noten, welche 6 zufällig ausgewählte Studierende $(s_1, s_2, \ldots, s_6)$ in den beiden Klausuren jeweils erreichten. Sie erhält die folgende Tabelle:

Studierende(r)	s_1	s_2	s_3	s_4	s_5	s_6
Note Mathematik 1	1.3	5.0	1.3	2.7	4.0	3.7
Note Mathematik 2	2.0	3.0	1.0	2.0	1.0	2.7

TEILAUFGABE 1.10.1 2 PUNKTE

Berechnen Sie **nur für die Noten der Vorlesung Mathematik 1** folgende Größen. Stellen Sie Ihren Rechenweg nachvollziehbar dar, bzw. geben Sie die Verwendung des Taschenrechners oder eines anderen elektronischen Tools an!

a) Arithmetisches Mittel

b) Alle Quartile

c) 90%-Quantil

d) empirische Standardabweichung

LÖSUNG

$n = 6$. Sortierte Liste der Stichprobe:

Studierende(r)	s_1	s_2	s_3	s_4	s_5	s_6
Note Mathematik 1	1.3	1.3	2.7	3.7	4.0	5.0

a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{18}{6} = 3$

b) via Taschenrechner oder sortierte Liste

- $Q1 : 1.3$ (25%-Quantil: $\tilde{x}_{0.25} = x_{\lceil 6*0.25 \rceil} = x_2 = 1.3$)
- $Q2 : 3.2$ (Median: $\tilde{x} = (x_3 + x_4)/2 = 3.2$)
- $Q3 : 4.0$ (75%-Quantil: $\tilde{x}_{0.75} = x_{\lceil 6*0.75 \rceil} = x_5 = 4.0$)

c) $0.9 \cdot 6 = 5.4 \Rightarrow \tilde{x}_{0.9} = x_{\lceil 5.4 \rceil} = x_6 = 5.0$

d) $s = 1.5073$ (TR) Falscher Ansatz: $\sigma = 1.37598$

TEILAUFGABE 1.10.2 3 PUNKTE

Beantworten Sie folgende Fragen in ganzen Sätzen und möglichst allgemeinverständlich!

a) Was bedeutet der Wert des 90%-Quantils (Ihr Ergebnis aus Aufgabe 1.10.1c) für die Noten von Mathematik 1?

b) Prof. Schmidt berechnet für die Noten der Vorlesung Mathematik 2 folgende Größen:

$$\bar{x} = 2.0, \quad (Q1,\ Q2,\ Q3) = (1.0,\ 2.0,\ 2.7), \quad \tilde{x}_{0.9} = 3.0, \quad s = 0.8462$$

Verwenden Sie diese Zahlen, sowie Ihre eigenen Berechnungen aus Aufgabe 1.10.1 um **zwei** Gemeinsamkeiten oder Unterschiede für die Noten in Mathematik 1 und 2 abzuleiten. Begründen Sie Ihre Meinung!

c) Welcher der folgenden Werte ist der korrekte empirische Korrelationskoeffizienten für die Noten in Mathematik 1 und Mathematik 2 $r_{1,2}$ (für die obige Stichprobe)? Begründen Sie Ihre Meinung.

$$r_{1,2} = \quad 0 \quad -0.571 \quad 0.423 \quad 1.1 \quad 0.952$$

LÖSUNG

a) $\tilde{x}_{0.9} = 5.0$: 90% aller Studierenden erhalten eine 5.0 oder besser, bzw. 10% aller Studierenden erhalten eine 5.0 oder schlechter (nicht sehr sinnvoll).

b)
- Bei Mathe 2 (im Vergleich zu Mathe 1) sind die Noten im Durchschnitt deutlich besser, sieht man am Vergleich von Mittelwert, allen Quantilen, Min, Max.
- Bei Mathe 2 (im Vergleich zu Mathe 1) sind Standardabweichung ist kleiner, IQR, Range auch, Noten sind homogener.
- Bei beiden Vorlesungen ist der Median größer als der Mittelwert, heißt, es gibt im Verhältnis mehr gute Noten, die die wenigen schlechten Noten im Mittel wieder ausgleichen.

c) $r_{1,2} = 0.952$, da die Daten einen deutlich positiven Zusammenhang (je größer M1 desto größer M2) zeigen
alternative Begründung über Ausschlussverfahren: Wert 1.1 ist nicht möglich für eine Korrelation, keine Korrelation (0) so sehen die Daten nicht aus, genauso wenig negativ korreliert (-0.571), 0.423 wäre möglich, aber die Korrelation ist tatsächlich stärker, damit 0.952.

AUFGABE 1.11 JOBSUCHE (KLAUSURAUFGABE WS20/21)

Sie suchen nach Ihrem Studium einen neuen Job, der bestimmte Anforderungen erfüllen soll.

- Der neue Job soll in einer attraktiven Stadt sein: Sie bewerten die Stadt in dieser Hinsicht mit Punkten von 1 (furchtbar) bis 10 (total toll).
- Natürlich ist Ihnen auch das Gehalt wichtig: Sie notieren das (geschätzte) Brutto-Gehalt in Tausend Euro ($55 \triangleq 55\,000$).
- Und die Tätigkeit soll interessant sein: Auch hier vergeben Sie Punkte von 1 (langweilig) bis 10 (genau Ihr Ding).

Nach zwei Wochen Suche haben Sie sich auf folgende 8 Jobs eingeschränkt und diese wie folgt bewertet:

Job	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8
Attraktivität der Stadt	7	4	10	10	9	3	9	6
Gehalt	65	60	40	70	45	40	65	55
interessant	3	10	9	1	5	10	2	3

TEILAUFGABE 1.11.1 3 PUNKTE

Berechnen Sie **nur für Ihre Punktevergabe der Attraktivität der Städte** (Zeile 1 der Tabelle) folgende Größen. Stellen Sie Ihren Rechenweg nachvollziehbar dar, bzw. geben Sie die Verwendung des Taschenrechners (TR), bzw. des genutzten Tools an!:

a) Modalwert

b) Arithmetisches Mittel

c) empirische Standardabweichung

d) Alle Quartile

e) 60%-Quantil

LÖSUNG

a) Modalwert: 10

b) $\bar{a} = 7.25$

c) $a_s = 2.7124$ falscher Wert: $\sigma_a = 2.5372$

d) $Q1 = 4, Q2 = 7, Q3 = 9$ (Matlab)

e) Sortierte Liste: $3, 4, 5, 6, 7, 9, 9, 10$
$0.6 \cdot 8 = 4.2 \Rightarrow a_{0.6} = a_5 = 7$ Matlab: 7

TEILAUFGABE 1.11.2 6 PUNKTE

a) Für welchen der 8 Jobs würden Sie sich auf Basis dieser Daten entscheiden? Begründen Sie Ihre Meinung (es gibt keine richtige oder falsche Antwort).

b) Ihre Freunde Alice und Bob haben sich jeweils auch 8 Jobs ausgesucht und die Attraktivität der Städte ebenso mit 1-10 bewertet. Alice berechnet den Mittelwert der Städte als $\overline{a_B} = 6.875$ und die entsprechende Standardabweichung zu $s_{a_B} = 5$. Bob berechnet den Mittelwert der Städte als $\overline{a_C} = 5$ und die entsprechende Standardabweichung zu $s_{a_C} = 0$. Was können Sie aus diesen Werten über die Attraktivität der Städte von Alice und Bob im Vergleich zu „ihren“ Städten sagen?

c) **Schätzen** Sie die empirischen Korrelationskoeffizienten $r_{A,G}, r_{A,I}, r_{G,I}$ zwischen „Attraktivität der Stadt“ und „Gehalt“, „Attraktivität der Stadt“ und „interessant“ sowie „Gehalt“ und „interessant“ für die ausgewählten Jobs.
Rechnen Sie diese nicht aus, sondern wählen Sie unter den folgenden Werten. Begründen Sie Ihre Meinung!

$r_{A,G}$	a) 0.9308	b) 9.308	c) -0.9308	d) 0.0093
$r_{A,I}$	a) 8.152	b) -0.8152	c) 0.8152	d) 0.0815
$r_{G,I}$	a) 0.6659	b) 0.0066	c) -0.6659	d) -6.659

LÖSUNG

a) keine Musterlösung, begründen Sie Ihre Meinung nur sinnvoll.

b) Die Städte von mir Alice sind im Mittel gleich interessant, allerdings streut die Stichprobe stärker, die Attraktivitäten sind verschiedener da Standardabweichung größer.
Stichprobe von Bob : Standardabweichung $= 0$, heißt alle Werte haben die gleiche Attraktivität 5.

c)
- $r_{A,G} = 0.9308$, weil Werte eindeutig zusammenhängen
positiv korreliert, je attraktiver die Stadt, desto höher das Gehalt (vielleicht weil Stadt teuer)
- $r_{G,I} = -0.6659$, weil tendenziell je besser bezahlt der Job desto uninteressanter, aber nicht immer
negativ korreliert, je interessanter die Stadt, desto weniger gut bezahlt der Job (aber nicht so stark wie Gehalt)
- $r_{A,I} = -0.8152$, weil tendenziell je attraktiver die Stadt desto uninteressanter der Job, aber nicht immer
negativ korreliert, je attraktiver die Stadt desto weniger interessant der Job, allerdings schwächerer Zusammenhang als die anderen.

Alternativ: Auch Ausschlussverfahren möglich: Korrelationskoeffizient muss kleiner als 1 sein ...

|Digitalaufgaben|

Lösen Sie alle Aufgaben in diesem Abschnitt digital mit einem Werkzeug Ihrer Wahl (MATLAB, Python, Excel, ...) und laden Sie den verwendeten Code bzw. Ihre Bilder und Ergebnisse nach Moodle hoch.

AUFGABE 1.12 AUSBILDUNG UND GEHÄLTER, 3 PUNKTE

Fünf Mitglieder eines Sportclubs gehen der Frage nach, ob zwischen ihrer Schul- und Ausbildungszeit und ihrem Jahreseinkommen ein statistischer Zusammenhang besteht. Das Ergebnis der anonymen Befragung ist in folgender Tabelle wiedergegeben.

Ausbildungsdauer X	9	13	15	18	20
Jahresgehalt Y [1000 €]	18	37	61	125	59

a) Berechnen Sie den empirischen Korrelationskoeffizient der beiden untersuchten Merkmale. Interpretieren Sie diesen Wert!

b) Bestimmen Sie eine lineare Regressionsfunktion und zeichnen Sie diese. Bewerten Sie die Passgenauigkeit dieser Regressionsgerade optisch und durch das Bestimmtheitsmaß.

LÖSUNG

a) Der empirische Korrelationskoeffizient berechnet sich als

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

MATLAB: `corrcoef` und die Einträge der Nebendiagonalen verwenden.

TR (TI): Menü `stat-reg`→`2-Var Stats` (wenn die Daten für x und y schon in den Spalten L1 und L2 eingegeben waren)

Per Hand sind folgende Berechnungsschritte notwendig:

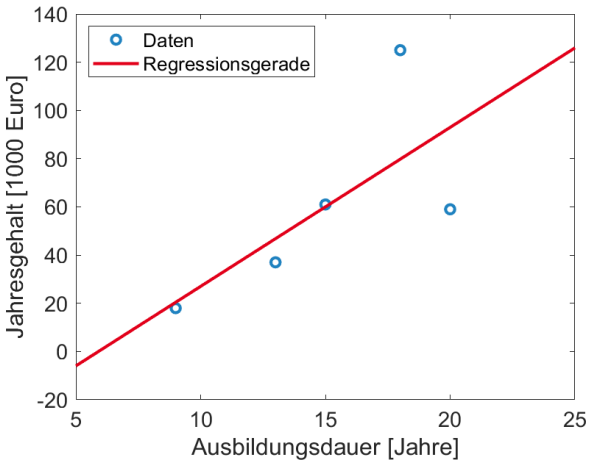
- $\bar{x} = \frac{1}{5}(9+13+15+18+20) = \frac{75}{5} = 15$
- $\bar{y} = \frac{1}{5}(18+37+61+125+59) = \frac{300}{5} = 60$
- $s_{x,y} = \frac{1}{5-1}\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})$
  $= \frac{1}{4}[(9-15)(18-60)+(13-15)(37-60)+(15-15)(61-60)+(18-15)(125-60)+(20-15)(59-6$
  $= \frac{1}{4}(6\cdot 42+2\cdot 23+0\cdot 1+3\cdot 25+5\cdot 1) = \frac{1}{4}(252+46+0+195+(-5)) = \frac{488}{4} = 122$
- $s_x = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(x_i-\bar{x})^2} = \sqrt{\frac{1}{4}((9-15)^2+(13-15)^2+(15-15)^2+(18-15)^2+(20-15)^2)}$
  $= \sqrt{\frac{1}{4}(36+4+0+9+25)} = \sqrt{\frac{74}{4}} = \sqrt{\frac{37}{2}} = 4.3012$
- $s_y = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(y_i-\bar{y})^2} = \sqrt{\frac{1}{4}((18-60)^2+(37-60)^2+(61-60)^2+(125-60)^2+(59-60)^2)}$
  $= \sqrt{\frac{1}{4}(42^1+23^2+1^2+65^2+1^2)} = \sqrt{1630} = 40.3733$
- $r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{122}{4.3012\cdot 40.3733} = 0.70256$

Ein Korrelationskoeffizient von mehr als 0.7 zeigt, dass Ausbildungsdauer und Jahresgehalt stark korreliert sind, eine Person mit einer längeren Ausbildung hat also mit einer höheren Wahrscheinlichkeit auch ein höheres Jahresgehalt.

b) **MATLAB:** `regression` und oder `plotregression`.

Per Hand sind folgende Berechnungsschritte notwendig::

- $k = r_{x,y} \cdot \frac{s_y}{s_x} = 0.70256 \cdot \frac{40.3733}{4.3012} = 6.5946$
- $d = \bar{y} - k\bar{x} = 60 - 6.5946 \cdot 15 = -38.9189$
- Regressionsgerade: $f(x) = kx + d = 6.5946 - 38.9189x$
- Bestimmtheitsmaß: $R^2 = r_{x,y}^2 = 0.70256^2 = 0.4936$



Sowohl R^2 als auch der optische Vergleich zeigen, dass die Regressionsgerade „ganz gut“ passt. Allerdings ist die Passung für längere Ausbildungszeiten deutlich schlechter. Das wird erst durch die optische Analyse sichtbar!

Aufgabe 1.13 Fernsehen und Agressionen, 3 Punkte

In einer Studie zur Auswirkung von YouTube-Filmen mit gewalttätigen Szenen auf das Sozialverhalten von Kindern wurden ein Aggressivitätsmaß Y (auf einer Skala von 0-10), die Zeitdauer X (in Minuten), während der das Kind gewöhnlich solche Filme sieht, und das Geschlecht Z des Kindes (0 männlich, 1 weiblich) erfasst.

Die Ergebnisse in einer zufällig gewählten Testgruppe waren wie folgt:

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	10	50	30	70	80	60	90	40	10	20	30	50	60
y_i	4	5	2	6	6	8	7	2	7	3	5	1	3
z_i	0	0	0	0	0	0	0	1	1	1	1	1	1

Verwenden Sie diese Testergebnisse, für eine (empirisch begründete) Aussage zum Thema "macht YouTube agressiv"?

Empfohlene Vorgehensweise:

a) Zeichnen Sie ein Streudiagramm für die 13 Kinder, und berechnen Sie den Korrelationskoeffizienten zwischen Dauer des Medienkonsums und Aggressionslevel ohne Berücksichtigung des Geschlechts.

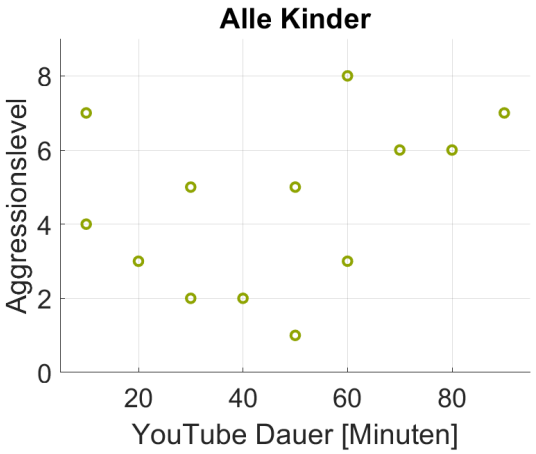
b) Zeichnen Sie nun für Jungen und Mädchen getrennt jeweils ein Streudiagramm und berechnen Sie für beide Geschlechter den Korrelationskoeffizienten.

c) Vergleichen Sie alle Ergebnisse und ziehen Sie Ihre Folgerungen.

Lösung

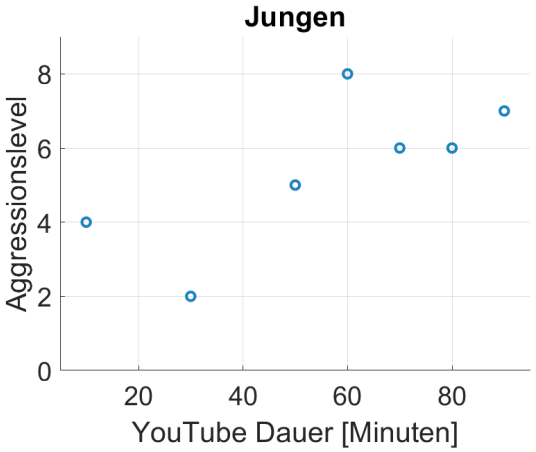
a) Alle Kinder:

- $\bar{x} = 46.154$
- $\bar{y} = 4.538$
- $r_{x,y} = 0.332$



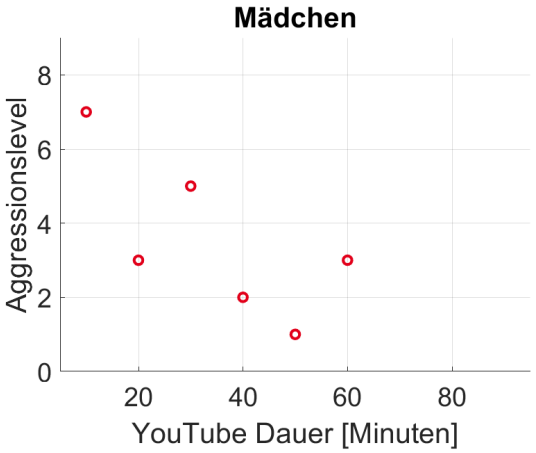
b) Jungen:

- $\bar{x} = 55.714$
- $\bar{y} = 5.429$
- $r_{x,y} = 0.722$



Mädchen:

- $\bar{x} = 35$
- $\bar{y} = 3.5$
- $r_{x,y} = -0.715$



c) Betrachtet man alle Kinder auf einmal, so lässt sich eine schwache positive, aber nicht sehr deutliche (siehe Bild) Korrelation feststellen.

Nach Geschlechtern getrennt zeigen die Ergebnisse jedoch, dass die Aggressivität bei Jungen mit der Dauer des Fernsehkonsums zunimmt, während sie bei Mädchen erkennbar abnimmt. weiterhin sind sowohl Dauer des Medienkonsums als auch Aggressionslevel im Mittel bei den Jungen deutlich höher als bei den Mädchen. Die beiden entgegengesetzten Wirkungen mitteln sich bei der Untersuchung beider Geschlechter zusammen heraus, sodass kaum eine Wirkung erkennbar ist.

AUFGABE 1.14 3 KEY FACTS, 6 PUNKTE

Suchen Sie sich einen der in Moodle verlinkten Datensätze, oder einen beliebigen Datensatz Ihrer Wahl aus und finden Sie 3 Key Facts (also Sachverhalte, die Sie für spannend halten). Verwenden Sie hierfür Methoden, und Kenngrößen, die Sie in der Vorlesung kennengelernt haben und MATLAB, Python, Excel, Java, ... Stellen Sie jeden Sachverhalt mit einem Bild, einer Zahl, einer Tabelle, ... dar. Beschreiben Sie die Aussage, die Sie daraus ableiten, sowie wie Sie das Bild, die Zahl, ... erhalten haben, kurz. Bündeln Sie alles (Bild, Code, Zahl, Erklärung) in einer Datei. Keine Punkte ohne (kurze) Vorstellung in der Übung.

Hinweise:

- Teamgröße: 3-5 (ein Team darf identische Lösungen abgeben, Namen der Teammitglieder bitte in der Lösung angeben)
- Bewertung: 1-2 Punkte pro gut erklärtem Fakt guten Code und gute Präsentation
- Geben Sie in Ihrer Lösung an, aus welcher Quelle Sie Ihre Daten haben.
- Beispielhafte Lösungen: siehe Moodle