2 Analyse statistischer Daten zu einem Merkmal – Lösungshinweise

Aufgabe 2.1: In der folgenden Tabelle ist eine Teilstichprobe zu den Studierenden in Münster gegeben, die hinsichtlich ihres Studienfachs ausgewählt wurde. Vergleichen Sie die beiden Studierendengruppen. Verwenden Sie grafische Darstellungen, Lage- und Streuparameter, um Gemeinsamkeiten oder Unterschiede aufzudecken.

	Studierende des Lehramts				Studierende der Jura			
	geordnet nach Alter		geordnet nach Abi		geordnet nach Alter		geordnet nach Abi	
Nr.	Alter	Abi	Alter	Abi	Alter	Abi	Alter	Abi
1	19	2,4	25	1,2	18	1,5	19	1
2	19	2,4	25	1,2	19	2,3	21	1
3	20	1,9	26	1,3	19	1,3	20	1
4	20	2,6	22	1,7	19	1,3	21	1,1
5	20	2,5	20	1,9	19	1	21	1,2
6	20	2,7	23	1,9	19	2,2	27	1,2
7	20	2,6	23	1,9	19	2,5	19	1,3
8	20	3	29	1,9	20	1,8	19	1,3
9	20	2,7	22	2	20	2,6	18	1,5
10	20	2	25	2	20	2	27	1,6
11	20	2	22	2	20	1,9	21	1,7
12	21	2,6	20	2	20	2,7	23	1,7
13	21	2,4	20	2	20	1	25	1,7
14	21	2,2	28	2,1	21	1,7	20	1,8
15	22	2	25	2,2	21	2,2	28	1,8
16	22	2,4	21	2,2	21	2	20	1,9
17	22	2,6	22	2,4	21	1,1	29	1,9
18	22	2	24	2,4	21	1,2	23	1,9
19	22	3	21	2,4	21	2,1	21	2
20	22	1,7	19	2,4	21	1	20	2
21	23	2,8	19	2,4	21	2,3	24	2
22	23	2,7	20	2,5	23	1,7	25	2
23	23	1,9	20	2,6	23	2,6	23	2
24	23	1,9	21	2,6	23	1,9	21	2,1
25	24	2,8	22	2,6	23	2	26	2,1
26	24	3,4	20	2,6	24	2,5	21	2,2
27	24	2,4	20	2,7	24	2	19	2,2
28	25	1,2	20	2,7	25	1,7	19	2,3
29	25	2	23	2,7	25	2	21	2,3
30	25	2,2	24	2,8	26	2,1	28	2,4
31	25	1,2	23	2,8	27	1,6	24	2,5
32	26	2,9	26	2,9	27	1,2	19	2,5
33	26	1,3	20	3	28	1,8	20	2,6
34	28	2,1	22	3	28	2,4	23	2,6
35	28	3	28	3	29	1,9	20	2,7
36	29	1,9	24	3,4				

Lösungsskizze Aufgabe 2.1

Lehramtsstudierende:

- Lagemaße:
  - Merkmal Alter (Angaben in Jahren): Modalwert 20; Minimum 19; unteres Quartil 20; Median 22; oberes Quartil 24,5; Maximum 29; arithmetisches Mittel 22,6
  - Merkmal Abiturnote: Modalwert 2,0 und 2,4 (bimodale Häufigkeitsverteilung); Minimum 1,2; unteres Quartil 2; Median 2,4; oberes Quartil 2,7; Maximum 3,4; arithmetisches Mittel 2,32
- Streumaße:
  - Merkmal Alter (Angaben in Jahren): Spannweite 10; Quartilsabstand 4,5; empirische Varianz 6,9; empirische Standardabweichung 2,6; mittlere absolute Abweichung 2,1; Variationskoeffizient 0,12
  - Merkmal Abiturnote: Spannweite 2,2; Quartilsabstand 0,7; empirische Varianz 0,26; empirische Standardabweichung 0,51; mittlere absolute Abweichung 0,41; Variationskoeffizient 0,22
- Form der Verteilung:
  - Merkmal Alter: Schiefe $g = 0,68$, Quartilskoeffizient $QS_{0,25} = 0,11$
  - Merkmal Abiturnote: Schiefe $g = -0,34$, Quartilskoeffizient $QS_{0,25} = -0,14$

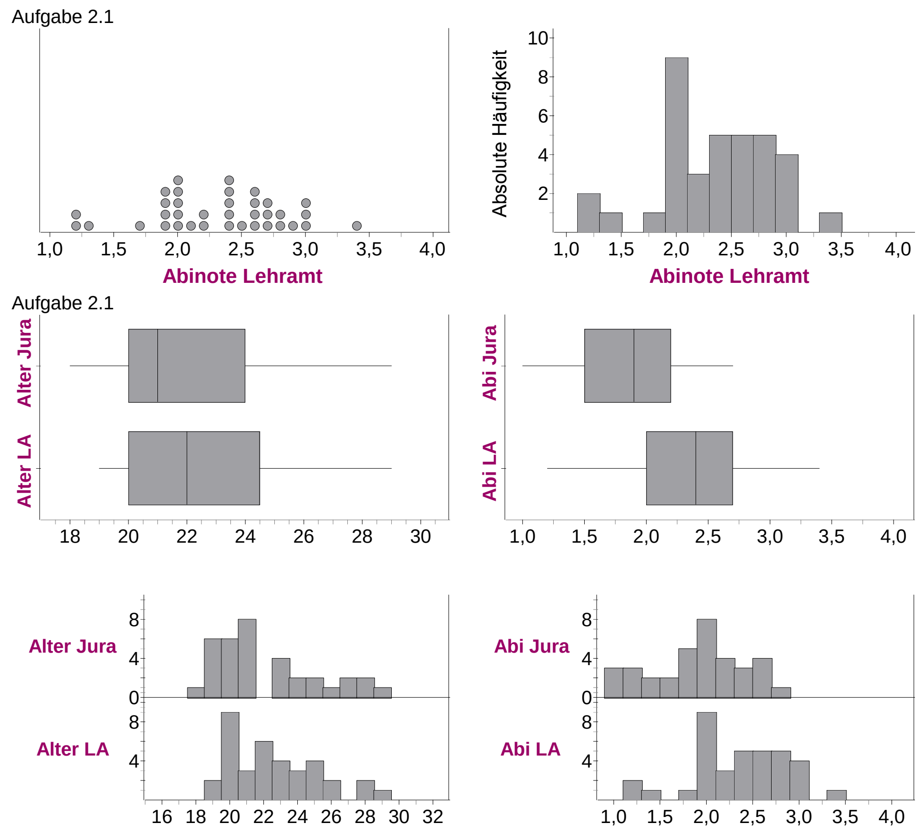
Jurastierende:

- Lagemaße:
  - Merkmal Alter (Angaben in Jahren): Modalwert 21; Minimum 18; unteres Quartil 20; Median 21; oberes Quartil 24; Maximum 29; arithmetisches Mittel 22,1
  - Merkmal Abiturnote: Modalwert 2,0 (unimodale Häufigkeitsverteilung); Minimum 1; unteres Quartil 1,5; Median 1,9; oberes Quartil 2,2; Maximum 2,7; arithmetisches Mittel 1,86
- Streumaße:
  - Merkmal Alter (Angaben in Jahren): Spannweite 11; Quartilsabstand 4,0; empirische Varianz 9,1; empirische Standardabweichung 3,0; mittlere absolute Abweichung 2,3; Variationskoeffizient 0,14
  - Merkmal Abiturnote: Spannweite 2,2; Quartilsabstand 0,7; empirische Varianz 0,26; empirische Standardabweichung 0,51; mittlere absolute Abweichung 0,41; Variationskoeffizient 0,26
- Form der Verteilung:
  - Merkmal Alter: Schiefe $g = 0,78$, Quartilskoeffizient $QS_{0,25} = 0,5$
  - Merkmal Abiturnote: Schiefe $g = -0,24$, Quartilskoeffizient $QS_{0,25} = -0,14$

Die errechneten Schiefemaße lassen für die Form der Verteilungen – bei den Lehramtsstudierenden wie bei den Jurastudierenden – im Fall des Merkmals *Alter* auf Linkssteilheit und im Fall des Merkmals *Abiturnote* (jeweils etwas schwächer) auf Rechtssteilheit schließen. Dies passt mit den Kriterien, die sich aus dem jeweiligen Vergleich von Median und arithmetischem Mittel ergeben (vgl. S. 35 dieses Buches), zusammen.

(Anmerkung: Da die Schulnoten keine metrisch skalierten Merkmalsausprägungen haben, sind streng genommen diejenigen Maße nicht sinnvoll zu interpretieren, die auf der Berechnung eines arithmetischen Mittels basieren.)

Ausgewählte grafische Darstellungen: Vergleicht man Punktdiagramm und Histogramm zur Verteilung der Abiturnoten der Lehramtsstudierenden, erkennt man, dass die Eigenschaft der Bimodalität im Histogramm bei der gewählten Klasseneinteilung nicht zu sehen ist. Im Unterschied zum Punktediagramm, das die einzelnen Merkmalsausprägungen in einzelnen Punkten grafisch abbildet, werden bei Histogramm die Merkmalsausprägungen in Klassen zusammengefasst. Je nachdem wie die Klassenbreite und Startwert gewählt werden, ergeben sich unterschiedliche Histogramme zum selben Datensatz.



Im Boxplotvergleich zeigt sich, dass sich Lehramtsstudierende und Jurastudierende hinsichtlich ihrer Altersstruktur vergleichsweise gering unterscheiden: Die Boxen, die die (mindestens) mittleren 50 % der jeweiligen Studierendengruppe repräsentieren, kommen annähernd gleich zu liegen und unterscheiden sich in der Länge um ein halbes Jahr. Die Antennen, die die unteren 25 % resp. die oberen 25 % repräsentieren, unterscheiden sich im unteren Fall um ein, im oberen Fall um ein halbes Jahr. Im Unterschied dazu fällt der Boxplotvergleich beim Merkmal Abiturnote deutlicher aus: Die Box der Lahramtsstudierenden ist deutlich nach rechts in den weniger guten Notenbereich verschoben, so weit, dass das obere Quartil der Jurastudierenden noch unter dem Median der Lehramtsstudierenden zu liegen kommt.

Der Blick auf die Histogramme zeigt die vergleichsweise geringen Schiefen: Während die Verteilung der Merkmalsausprägungen der Abiturnote im Fall der Jurastudierenden vergleichsweise symmetrisch aussieht, lässt sich bei den Lehramtsstudierenden eine Tendenz zur Linkssteilheit bzw. Rechtsschiefe erahnen. Im Unterschied dazu sind – beide Studiengänge betreffend – die Verteilungen der jeweiligen Alterstruktur augenscheinlich deutlich linksssssteiler bzw. rechtsschiefer.

Eine Excel-Datei mit den Berechnungen steht in den online-Materialien zum Download zur Verfügung.

Aufgabe 2.2: Verwenden Sie den Gesamtdatensatz (Datenquelle: siehe Vorwort).

a) Untersuchen Sie anhand dieses Datensatzes Eigenschaften der Studierenden verschiedener Hochschulen mithilfe grafischer Darstellungen, Lage- und Streuparametern. Versuchen Sie, Ihre Interpretationen der Ergebnisse einerseits sachlich und andererseits plakativ zu gestalten.

b) Untersuchen Sie die Streuung verschiedener Merkmale sowie die normierte Streuung mit dem Variationskoeffizienten.

c) Untersuchen Sie, welche der Merkmale eine symmetrische bzw. asymmetrische Verteilung aufweisen. Überlegen Sie sich, ob sich die Form der Verteilung plausibel erklären lässt.

Eine Excel-Datei mit den Berechnungen steht in den online-Materialien zum Download zur Verfügung.

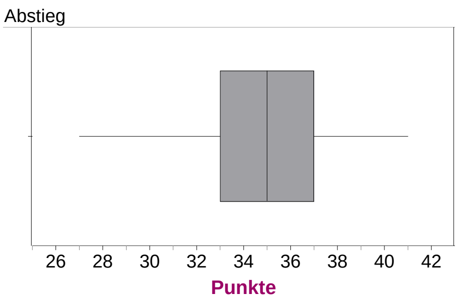
Lösungsskizze Aufgabe 2.2

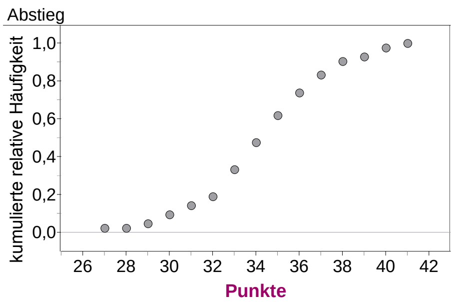
a) und b) Entsprechende rechnergestützte Vorgehensweise wie in der Lösung zur Aufgabe 2.1.

c) Die Verteilungen der Merkmale Alter, Semester, Entfernung zur Hochschule, Zeit für den Hochschulweg, Stundenanzahlen für sportliche und musikalische Freizeitaktivitäten, für Arbeit, für Computeraktivitäten sind (unterschiedlich stark, aber in der Tendenz alle) linkssteil. Dies lässt sich in den meisten Fällen plausibel erklären: So ist z. B. bei der Altersverteilung aufgrund der relativ großen Altershomogenität beim Übertritt von Schule in die Hochschule kaum eine symmetrische oder gar rechtssteile Verteilung zu erwarten. Auch die Verteilungen der genannten Stundenanzahlen lassen durch die Begrenztheit der zur Verfügung stehenden Zeit eine linkssteile Verteilung plausibel erscheinen. Demgegenüber ist beispielsweise das Merkmal der geäußerten Zufriedenheit vergleichsweise symmetrisch verteilt. Hier ist die Verteilung das Ergebnis einer freien subjektiven Einschätzung. Die Abiturnote weist sogar eine rechtssteile Tendenz auf.

Aufgabe 2.3: In den Zusatzmaterialien ist ein Datensatz zu den Ergebnissen der Fußballbundesliga gegeben. Stellen Sie eine Regel auf, wie viele Punkte eine Mannschaft in einer Saison erhalten muss, um nicht abzusteigen bzw. um Meister zu werden. (Datenquelle: siehe Vorwort)

Lösungsskizze Aufgabe 2.3 Nicht-Abstieg: Die Frage, welche Punktzahl vor dem Abstieg rettet, ist mit der Angabe eines Durchschnitts nicht befriedigend zu beantworten. So nützt etwa die Information, dass das arithmetische Mittel 34,7 beträgt, also dass die Absteiger im Durchschnitt ca. 35 Punkte erzielt hatten, wenig. Gleiches gilt für die alleinige Angabe des Medians von 35 Punkten. Argumentiert man hingegen auf der Datenlage mit dem Maximum von 41 Punkten, so hat die Aussage „Mit mehr als 41 Punkten ist noch niemand abgestiegen" durchaus eine empirische Aussagekraft (gleichwohl kann sie keine Garantie für die Zukunft geben). Nachteil von der Orientierung am Maximum ist, dass die Aussage recht sicher erscheint, aber mit dem Nachteil erkauft wird, dass sie je nach Punkteniveau der laufenden Saison dennoch unrealistisch hoch erscheinen kann und sich kein Spielraum für die Schätzung einer Prognose ergibt.





Informativer ist es im vorliegenden Fall, die Streuung zu betrachten und Intervalle anzugeben, in denen ein bestimmter Anteil der zur Verfügung stehenden Daten liegt. Im Boxplot ergibt sich, dass die mittleren (mindestens) 50 % der Absteiger zwischen 33 (unteres Quartil) und 37 Punkten (oberes Quartil) erreicht haben (Quartilsabstand 4). Daraus lässt sich vorsichtig abschätzen: Erreicht einen Mannschaft eine Punktzahl in diesem Intervall, ist recht sicher davon auszugehen, dass sie absteigt, diese Sicherheit ist noch als höher anzusehen, wenn sie eine Punktzahl darunter erzielt. Im Punktebereich der oberen Boxplotantenne ist durchaus noch die Möglichkeit zum Klassenerhalt zu sehen (ist aber keinesfalls als sicher anzusehen). Eine Alternative zur feineren entsprechenden Abschätzung ergibt sich mit dem bereits in einem anderen Kontext verwendeten Diagramm (Buch, S. 19/20), das die kumulierten relativen Häufigkeiten über den Punktzahlen abträgt, die zum Abstieg geführt haben. Aus dieser Darstellung lässt sich z. B. als vorsichtige Abschätzung ablesen, dass eine Punktzahl im Intervall $[38;41]$ nur in etwa 10 % aller bisherigen Fälle zum Abstieg geführt hat.

Meisterschaft: Entsprechende Vorgehensweise zu den vorausgehenden Betrachtungen. Man beachte, dass sich hier das Phänomen zeigt, dass mit höheren Mittelwerten (oder Medianen) häufig eine größere Streuung einhergeht: Die Streuung der Punkteverteilung auf Platz 1 ist deutlich größer als auf Platz 16.

Aufgabe 2.4: Gegeben ist das Gehaltsgefüge einer fiktiven Firma:

		Gehalt in Euro	
Bezeichnung	**Anzahl**	**Durchschnitt**	**Gruppe**
Arbeiter	1.000	1.000	1.000.000
Arbeiter, gehobene Position	500	2.000	1.000.000
Leiter von Arbeitsgruppen	50	5.000	250.000
Management	10	10.000	100.000
Firmenbesitzer	1	1.000.000	1.000.000
Summe	1.561	—	3.350.000

a) Ein Arbeiter bekommt genau das Durchschnittsgehalt seiner Berufsgruppe, möchte aber mehr Geld bekommen. Der Firmenbesitzer möchte dagegen das Gehalt des Arbeiters nicht erhöhen. Argumentieren Sie aus der Sicht des Arbeiters und aus der Sicht des Firmenbesitzers, warum das Gehalt nicht ausreichend bzw. angemessen ist!

b) Das Gehaltsgefüge der Firma wird von außen, z. B. von einer Gewerkschaft einerseits und vom Arbeitgeberverband andererseits, begutachtet. Argumentieren Sie aus der Sicht der Gewerkschaft und aus der Sicht des Arbeitgeberverbandes, warum das Gehalt der Arbeiter angehoben werden muss bzw. beibehalten werden kann!

Hinweis: Verwenden Sie für Ihre Argumentationen insbesondere den Median und das arithmetische Mittel. Untersuchen Sie die Schiefe der Verteilung.

Lösungsskizze Aufgabe 2.4 Das arithmetische Mittel der Einkommen aller Firmenbeschäftigten errechnet sich zu: $3350000 : 1561 \approx 2146$ Euro. Der Median ergibt sich daraus, dass alle Einkommen der Größe nach geordnet werden und die Höhe des $n_{\frac{1561+1}{2}}$-ten Gehalts ermittelt wird. Im vorliegenden Fall ist dies das Gehalt eines einfachen Arbeiters, nämlich 1000 Euro. Aus dem Vergleich von arithmetischem Mittel und Median 2146 Euro > 1000 Euro wie aus der offensichtlichen Tatsache, dass viele wenig und wenige viel verdienen, ergibt sich eine linkssteile Verteilung der Einkommen.

a) Mögliche Argumentation des Arbeiters mit dem arithmetischen Mittel: Er könnte geltend machen, dass in der Firma durchschnittlich das Doppelte seines Gehalts verdient wird, und er aus diesem Grund eine Gehaltserhöhung als gerechtfertigt ansieht.

Mögliche Argumentation des Firmenbesitzers mit dem Median: Er könnte erwidern, dass über die Hälfte seiner Angestellten ebenso viel verdient wie der Arbeiter, und daher eine Erhöhung nicht gerechtfertigt erscheint.

b) Mögliche Argumentation eines Gewerkschaftsvertreters mit dem Median: Er könnte argumentieren, dass über die Hälfte der Angestellten mit lediglich 1000 Euro auskommen müssen, und die finanzielle Situation dieser Angestellten dringend verbessert werden muss.

Mögliche Argumentation eines Vertreters des Arbeitgeberverbands mit dem arithmetischen Mittel: Er könnte geltend machen, dass die Situation der Angestellten mit einem durchschnittlichen Gehalt von ca. 2150 Euro recht gut sei, da selbst Facharbeiter mit Spezialausbildung, die üblicherweise durchschnittlich ein Gehalt von 2100 Euro beziehen, nicht so viel verdienen. (Hinweis: Der Vergleich mit einem Facharbeiter mit Spezialausbildung, der durchschnittlich 2100 Euro (fiktives Gehalt, das höher liegt als das höchste Arbeitereinkommen in der Firma) verdient, ist in der Aufgabe nicht enthalten. Der Vergleich spitzt die Argumentation über das arithmetische Mittel zu.)