M:eta
Mathematik: Einführung, Theorie, Aufgaben
Statistik

Torsten Linnemann

Gymnasium Oberwil – Fachmittelschule

25. Februar 2023

9 Statistik

Dieses Kapitel verdanke ich zu grossen Teilen Martin Münch und Boris Girnat.

In der beschreibenden Statistik geht es vor allem darum, vorhandene Daten anschaulich darzustellen und auszuwerten. In der beurteilenden Statistik geht es darum, sich Zusammenhänge zwischen Daten zu erschliessen. Die beurteilende Statistik wird in Ansätzen in den Projekten dieses Kapitels bearbeitet.

Die Abschnitte dieses Kapitels:

- Mit Grafiken können wir sehr schnell erkennen, ob unsere Daten unerwartete Strukturen und Besonderheiten aufweisen. Dies ist der erste Schritt beim Auswerten von Statistiken.
- Lagemasse zeigen, wie die Daten liegen. Das bekannteste Lagemass ist der Mittelwert.
- Dann geht es um die Streuung der Daten: Auch wenn zwei Tests in zwei Klassen den gleichen Mittelwert haben, kann völlig verschiedene Auswertungen haben: ist die Streuung gross, gibt es sehr wahrscheinlich auch ungenügende Noten.
- Im ersten Exkurs werden verschiedene Datensätze an Hand von Lagemassen und Streuung verglichen.
- Im Projekt Sport werden Zusammenhänge von Datensätzen erforscht: es werden Daten aus der eigenen Schulklasse ermittelt.

9.1 Diagramme erstellen und auswerten

Die hier verwendeten Darstellungen kennen Sie bereits aus verschiedensten Zusammenhängen.

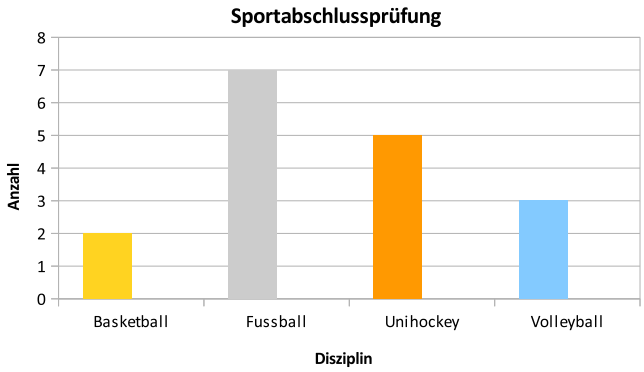
Um den Text zu verstehen braucht es einige Worte:

Das, was wir in der Statistik untersuchen, nennen wir «Merkmal». Beispielsweise die Körpergrösse oder die Anzahl Regentage in jedem Monat in Glasgow.

Die Merkmale treten dann in Ausprägungen auf: Das Merkmal Regentage in Glasgow hat im Juli die Ausprägung 17. Dann werden Häufigkeiten betrachtet: Unter den 12 Monaten gibt es drei Monate mit 17 Regentagen. Die absolute Häufigkeit von 17 Regentagen ist 3. Für die relative Häufigkeit teilen wir durch 12 (Monate). Die relative Häufigkeit von 17 Regentagen ist $\frac{3}{12} = 0.25 = 25$ Prozent. (Übrigens, der Mai ist mit durchschnittlich 14 Regentagen der sonnigste in Glasgow).

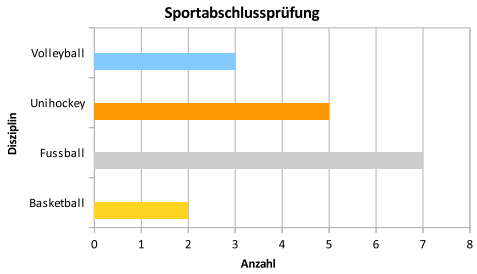
Säulendiagramm

Es werden auf der auf der y-Achse die absoluten oder relativen Häufigkeiten abgetragen. Die Rechtecke sollen nicht aneinander stossen.



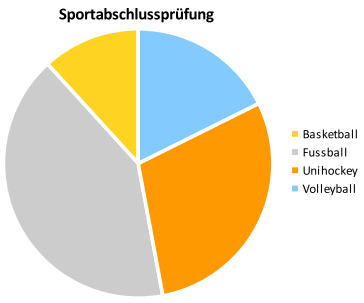
Balkendiagramm

Das Balkendiagramm ist eine Variante des Säulendiagramms. Die Ausprägungen werden nun aber auf der y-Achse und die absoluten oder relativen Häufigkeiten auf der x-Achse abgetragen, also genau umgekehrt zu einem Säulendiagramm.



Kreisdiagramm

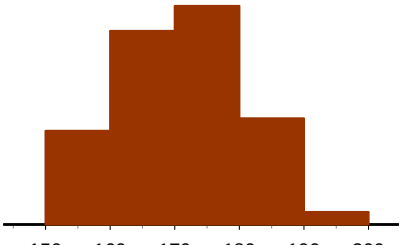
Ein weitere gebräuchliche Darstellungsform von Daten ist das sogenannte Kreisdiagramm. Der Winkel der einzelnen Sektoren ist proportional zu den absoluten oder relativen Häufigkeiten. Somit ist auch die Fläche der einzelnen Kreissektoren proportional zur Häufigkeit.



Klasseneinteilung und Histogrammm

Bei grossen Datenmengen ist es sinnvoll, sie in sogenannte «Klassen» einzuteilen: werden von allen Schülerinnen und Schülern einer Schule die Körpergrössen in cm erhoben, so gibt es viele Werte, zum Beispiel 155, 156, 157, 158, ...

Die erste Klasse enthält die Zahl der Schülerinnen und Schüler, die von 150 bis 159 cm gross sind, die nächste diejenigen von 160 bis 169 und so weiter. Üblich ist hier eine Darstellung in einem Histogrammm – das sieht wie ein Säulendiagrammm aus, aber ohne Abstände (das soll zeigen, dass die Klassen aneinander stossen)



Stamm-Blatt-Diagramm

Das Stamm-Blatt-Diagramm dient der Visualisierung von Häufigkeitsverteilungen. Im Gegensatz zum Histogramm bleiben die einzelnen Werte mit gewünschter Genauigkeit erhalten. Das Diagramm besteht aus zwei Spalten. In der ersten Spalte (Stamm) sind die Klassen (typisch ist eine Klassenbildung nach dem Dezimalsystem), in der rechten Spalte (Blätter) die einzelnen Werte.
Messreihe: 4.2, 3.5, 2.4, 5.2, 5.5, 6.0, 1.5, 4.6, 4.1, 4.0, 3.5, 3.6, 2.6, 5.7, 5.1 and 4.7

Tabelle 9.1: Stamm-Blatt-Diagramm am Beispiel von Noten

Stamm	Blätter
1	5
2	4 6
3	5 5 6
4	2 6 1 0 7
5	2 5 7 1
6	0

Beispiel 9.1: Darstellung absoluter Häufigkeiten

Aus einer Population von Frauen mit hohem Blutdruck (systolischer Wert $>$ 130) wurde eine Stichprobe vom Umfang 60 ausgewählt. Das relative Gewicht der Frauen (in % bezogen auf das sogenannte Normalgewicht für Frauen derselben Grösse) betrug:

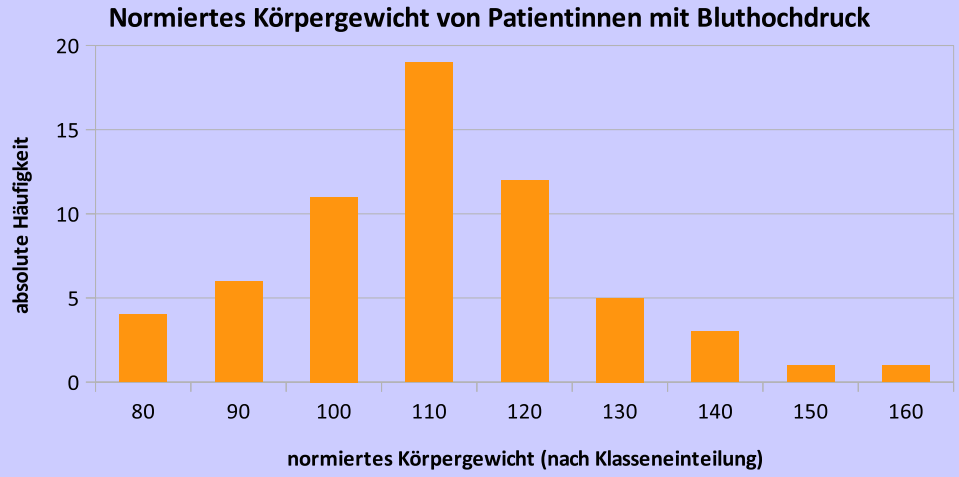
117, 89, 107, 102, 95, 132, 163, 103, 144, 108, 89, 118, 123, 92, 140, 102, 112, 115, 123, 130, 102, 119, 119, 114, 119, 100, 106, 115, 145, 115, 108, 122, 113, 153, 116, 81, 85, 122, 120, 113, 140, 117, 119, 130, 114, 107, 125, 94, 96, 118, 123, 91, 120, 125, 125, 103, 133, 102, 131 and 147

a) **Klasseneinteilung**:
Man teilt die Daten in Klassen (Abschnitte) ein. Relativ einfach geht das mit dem Stamm-Blatt-Diagramm (vgl. Tabelle 17.1), aber auch andere Methoden sind denkbar.

b) **Relative Häufigkeit**:
Wir berechnen die relative Häufigkeit als Prozentzahlen des Stichprobenumfangs.

Stamm	Blätter	absolute Häufigkeit	relative Häufigkeit in %
8	9 9 1 5	4	6.7
9	5 2 4 6 1	5	8.3
10	7 2 3 8 2 2 0 6 8 7 3 2	12	20
11	7 8 2 5 9 9 4 9 5 5 3 6 3 7 9 4 8	17	28.3
12	3 3 2 2 0 5 3 0 5 5	10	16.7
13	2 0 0 3 1	5	8.3
14	4 0 5 0 7	5	8.3
15	3	1	1.7
16	3	1	1.7
Summe		60	100

c) **Säulendiagramm**:
(Eigentlich sollte hier ein Histogramm stehen, ohne Abstände zwischen den Klassen. Sie müssen sich daran gewöhnen, dass gerade bei der statistischen Darstellung Flexibilität beim Lesen wichtig ist.)



9.1.1 Übungen

1. Auf dem Marktplatz vor dem Ratshaus in Basel soll ein Kunstwerk aufgestellt werden. Von 1550 befragten Passanten wollten 450 die Eisenplastik, 765 die Holzskulptur und die restlichen mögen beides nicht. Stellen Sie die Meinung der Passanten in einem Kreisdiagramm dar.

2. Es wird oft behauptet, im Strassenverkehr seien die Autos nicht ausgelastet. Für den Arbeitsverkehr gibt das Bundesamt für Statistik durchschnittlich 1.14 Personen pro Auto an, für den Einkauf 1.70 Personen und für die Freizeit 2.07 Personen. Zeichnen Sie dazu ein Balkendiagramm.

3. An einem Test beteiligten sich 35 Kandidatinnen und Kandidaten. Im ersten Anlauf erreichen nur 3 eine sehr gute Leistung (Note 6), 5 eine gute (5) und 16 eine befriedigende (4) Leistung; die restlichen Kandidatinnen und Kandidaten erzielen eine ungenügende Leistung (3). Deshalb wird der Test mit allen

Beteiligten wiederholt. Im zweiten Anlauf erreichen 8 das Prädikat „sehr gut“, 12 eine gute und 10 eine befriedigende Leistung. Zeichnen Sie zu diesen Angaben zwei Kreisdiagramme, sodass die Tendenz erkennbar wird.

4. 685 Schülerinnen und Schüler antworten auf die Frage: „Sollen Noten in der Schule abgeschafft werden?“ folgendermassen:

	unbedingt abschaffen	fände ich toll	ist mir egal	fände ich nicht so toll	ich bin dagegen
Stimmen	96	247	123	164	55
Anteil in %					

a) Berechnen Sie die Anteile in %.

b) Stellen Sie die oben berechneten relativen Häufigkeiten in einem Kreisdiagramm dar.

5. Die 32 Schülerinnen und Schüler einer Klasse werden nach ihren Hobbys befragt.

	Sport	Computer	Lesen	Musik	Basteln	Sonstiges
Antworten	22	6	11	16	8	9
Anteil in %						

a) Berechnen Sie die Anteile.

b) Stellen Sie die Anteile in einem Säulendiagramm dar.

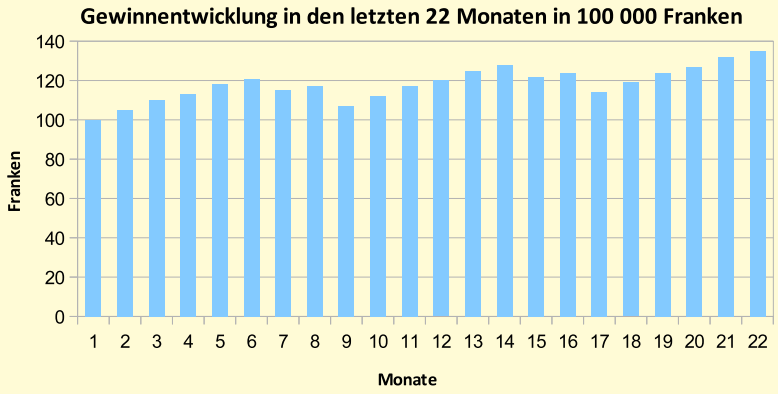
Auftrag 9.1

Die folgenden Zahlen geben den Gewinn der Firma M&M in den letzten 22 Monaten (in CHF 10 000.-) an:

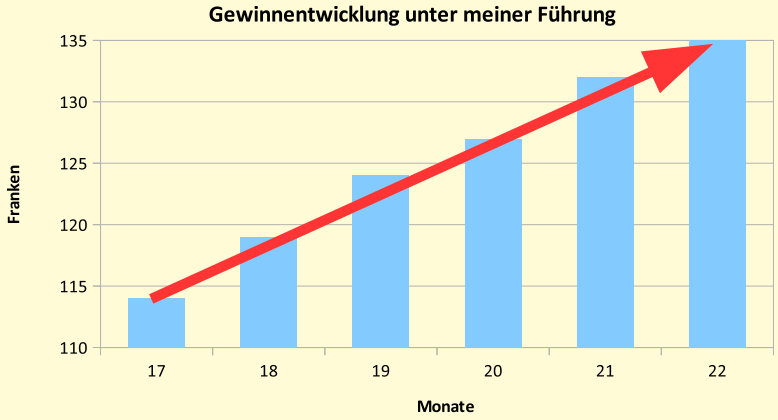
100, 105, 110, 113, 118, 121, 115, 117, 107, 112, 117, 120, 125, 128, 122, 124, 114, 119, 124, 127, 132 and 135

Etwas übersichtlicher könnte man die Daten in einer Tabelle darstellen. Erstellen Sie eine solche.

Der Trend ist viel leichter zu erkennen, wenn man die Daten in einem Balken- oder Säulendiagramm darstellt.
Es ist klar eine Art Zyklus zu erkennen.



Ein ehrgeiziger Manager möchte, dass die Gewinnentwicklung besonders positiv aussieht. Er wendet drei Tricks an, welche?



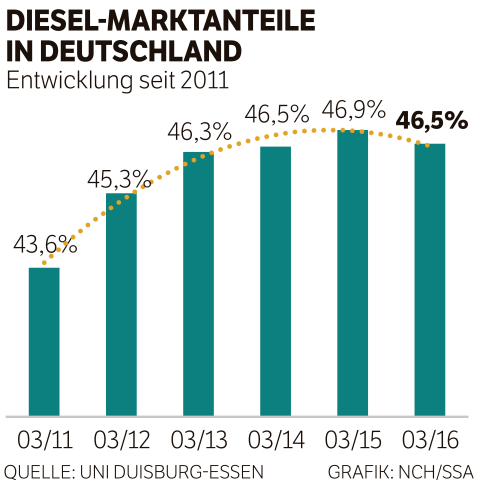
Auftrag 9.2

Autohersteller haben den Abgasausstoss von Dieselautos manipuliert. Das wird als Abgasskandal bezeichnet. In der Basellandschaftlichen Zeitung vom 4.5.2016 erschien ein Artikel mit der Überschrift: Der Abgas-Skandal verunsichert – Erstmals nach dem Abgas-Skandal ist ein Trendwechsel in Deutschland messbar: Verkäufe von Diesel-Fahrzeugen nehmen ab.

Zur Veranschaulichung wurde die rechts stehende Graphik veröffentlicht.

a) Schildern Sie, mit welchen Mitteln der Trend deutlich gemacht wird.

b) Erstellen Sie mit Zahlen aus der Graphik eine eigene Graphik mit der Aussage: der Verkauf von Diesel- Fahrzeugen bleibt stabil. Verwenden Sie also graphische Mittel, um diese Aussage zu verdeutlichen. Teilen Sie mit, welche Mittel Sie dazu verwendet haben.



Auftrag 9.3: Struktur der Schweizer Wirtschaft

Wichtig ist es, Graphiken interpretieren zu können. Der Auftrag auf den nächsten drei Seiten stammt von Pascale Herrmann, und wurde für das Fach "Wirtschaft und Recht» entwickelt.

Wählen Sie drei der «20 Fakten zum Unternehmerland Schweiz» aus und interpretieren Sie diese, wie es im Beispiel vorgegeben wird.

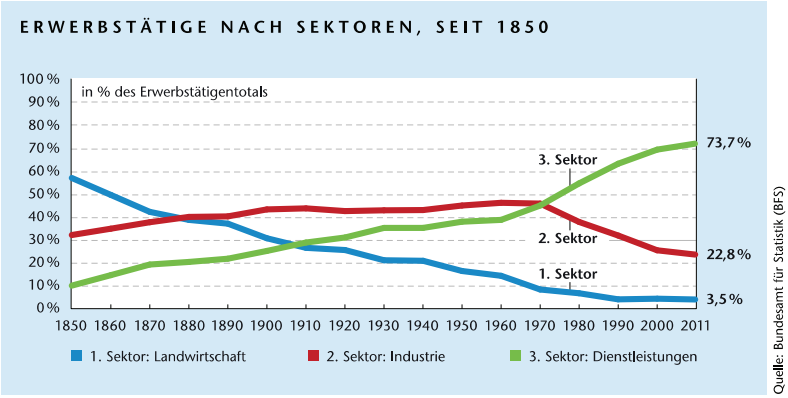
Struktur der Schweizer Wirtschaft

Im Dossier finden Sie Grafiken und Tabellen, die Informationen zur Unternehmenslandschaft Schweiz beinhalten. Wählen Sie eine Grafiken aus und analysieren Sie diese gemäss folgender Struktur:

Drei Grundfragen zum Lesen von statistischen Grafiken und Tabellen

Grundfrage	Detail	Ergebnis
1. Was ist genau das Thema?	Titel/Untertitel lesen Was ist die Fragestellung dieser Statistik? Auf welchen Zeitraum bezieht sie sich? (Jahr/Reihe von Jahren)	In eigenen Worten formulieren
2. Was wird gemessen (Messgrössen)?	Was sind die Masseinheiten? (Franken/pro Kopf/Prozent von was/etc.) Entwicklung? Veränderung? Zeit von bis Quelle?	Stichwörter
3. Was kann festgestellt werden (Kernaussagen)?	1 bis 2 sachliche und präzise Feststellungen pro Messgrösse	1-3 Sätze

Beispiel



Titel der Graphik	Erwerbstätige nach Sektoren, seit 1850
1. Thema	• Wie hat sich die Beschäftigung im 1., 2. und 3. Sektor von 1850 bis 2011 entwickelt?
2. Messgrössen	• Prozentualer Anteil der Erwerbstätigen in den 3 Wirtschaftssektoren am Total aller Erwerbstätigen. • Entwicklung von 1850 bis 2011 • Quelle: Bundesamt für Statistik
3. Kernaussagen	• Der Anteil der Beschäftigten in der Landwirtschaft hat sich von knapp 60% im Jahr 1850 kontinuierlich auf 3.5% im Jahr 2011 abgeschwächt. • Der Anteil der Beschäftigten in der Industrie ist von rund 30% im Jahr 1850 bis 1900 auf rund 45% angestiegen und dort bis 1970 verharrt. Danach ist er bis auf 22.8% im Jahr 2011 gefallen. • Der Anteil der Beschäftigten im Dienstleistungssektor ist von 1850 bis in die 1960er Jahre stetig von 10% auf rund 40% angewachsen. Danach legte der Beschäftigungsanteil markant auf 73.7% im Jahr 2011 zu.

20 Fakten zum Unternehmerland Schweiz

Internationalen Rankings zufolge zählt die Schweiz zu den konkurrenzfähigsten Ländern der Welt. Doch der Wettbewerb ist hart. Gefragt sind Unternehmergeist, Innovationsbereitschaft und Qualitätsprodukte. In der Übersicht präsentieren wir interessante Zahlen und Grafiken zum Wirtschaftsstandort Schweiz.

1. Wirtschaftsstandorte im Vergleich

2. Platz

Der Wirtschaftsstandort Schweiz bietet Unternehmen im internationalen Vergleich hervorragende Rahmenbedingungen: Im Global Entrepreneurship and Development Index (GEDI) 2018 belegt die Schweiz den zweiten Platz.

Rang	Land	GEDI
1.	USA	83.6
2.	Schweiz	80.4
3.	Kanada	79.2
4.	Grossbritannien	77.8
5.	Australien	75.5
6.	Dänemark	74.3
7.	Island	74.2
8.	Irland	73.7
9.	Schweden	73.1
10.	Frankreich	68.5

Quelle: Global Entrepreneurship and Development Index (GEDI) 2018

5. Firmengründungen

43 174

2018 wurden 43 174 Unternehmen neu ins Handelsregister eingetragen, am meisten in den Branchen Unternehmensdienstleistungen (B2B), Unternehmens- und Steuerberatung, Handwerk, Einzelhandel und Gastgewerbe.

Quelle: Schweizerisches Handelsamtsblatt SHAB; IFJ Institut für Jungunternehmen AG

35,3%

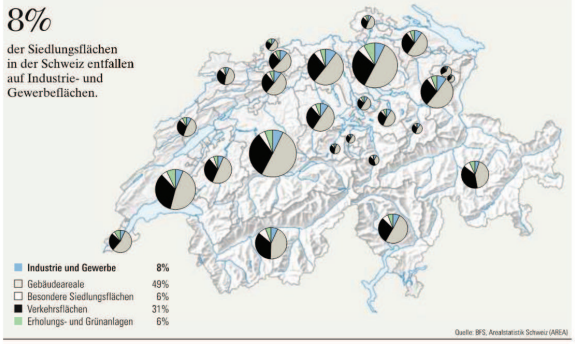
Mehr als ein Drittel der Firmen werden von Frauen gegründet, 54,9% von Männern und 9,7% von Männern und Frauen gemeinsam.

Quelle: BFS, Statistik der Unternehmensdemografie (2016)

6. Industrie- und Gewerbeflächen

8%

der Siedlungsflächen in der Schweiz entfallen auf Industrie- und Gewerbeflächen.

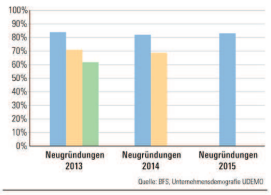


Quelle: BFS, Arealstatistik Schweiz (AREA)

7. «Überlebensrate» neuer Unternehmen

83%

Die durchschnittliche Überlebensrate der neuen Firmen in der Schweiz liegt ein Jahr nach der Gründung bei 83%. So waren 32 820 Unternehmen, die 2015 «ex nihilo» gegründet wurden, auch 2016 noch aktiv. Besonders hoch ist die Überlebensrate in der Branche «Gesundheits- und Sozialwesen».



Quelle: BFS, Unternehmensdemografie UDEMO

8. Familienbetriebe

75%

Von den rund 585 000 Schweizer KMU sind drei Viertel Familienunternehmen, also vollständig im Besitz der Gründerfamilie.

Quelle: Credit Suisse Succession Survey 2016

8%

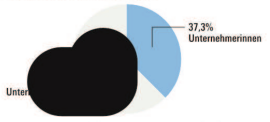
Die Aktien von Schweizer Familienunternehmen schneiden besser ab als jene von Firmen, die nicht in Familienbesitz sind. Eine Studie der Credit Suisse zeigt, dass Unternehmen, die mehrheitlich in Familienbesitz sind, in den vergangenen zehn Jahren rund 8 Prozent mehr Rendite im Jahr erzielten.

Quelle: Credit Suisse

9. Unternehmerinnen

37,3%

Frauen sind in der Kategorie der Selbstständigen (einschliesslich der im eigenen Unternehmen angestellten Personen) weiterhin unterrepräsentiert, aber ihr Anteil steigt. Im zweiten Quartal 2018 belief sich der Frauenanteil im Kreis der Unternehmenden auf 37,3%.

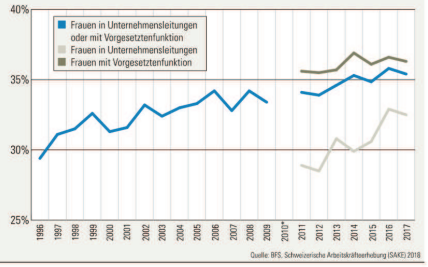


Quelle: BFS, Schweizerische Arbeitskräfteerhebung (SAKE), 2. Quartal 2018

10. Frauen in Führungspositionen

35,4%

Der Anteil der Frauen in Führungspositionen lag 2017 bei 35,4%. Zehn Jahre zuvor waren es 32,8%. 1996 betrug der Anteil erst 29,4%.



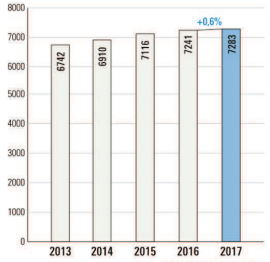
* Die Daten 2010 werden nicht veröffentlicht, weil die Umformulierung der Frage nach der Stellung im Beruf erst nach dem ersten Quartal 2011 gültig war.

Quelle: BFS, Schweizerische Arbeitskräfteerhebung (SAKE) 2018

16. Patentanmeldungen

7283

Das Europäische Patentamt verzeichnete 2017 insgesamt 7283 Patentanmeldungen aus der Schweiz – eine neue Höchstmarke. Mit 884 Anmeldungen pro eine Million Einwohner liegt die Schweiz im Pro-Kopf-Ranking europaweit vorne.



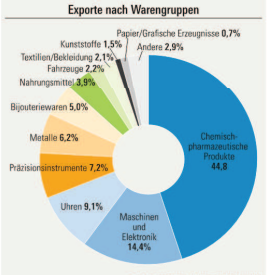
Quelle: IGE, Europäisches Patentamt (2018)

17. Export

233 146 000 000

Die Schweizer Wirtschaft exportierte im Jahr 2018 Waren im Wert von 233,146 Milliarden Franken (nominal) – ein neuer Rekord. Die Importe zogen ebenfalls an, auf 201,8 Milliarden Franken.

Exporte nach Warengruppen



Quelle: Eidgenössische Zollverwaltung (EZV), 2019

18. Pendler

4 000 000

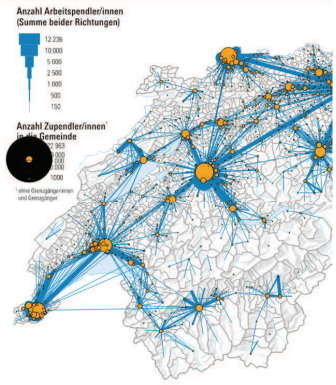
9 von 10 Erwerbstätigen in der Schweiz sind Pendlerinnen bzw. Pendler, was rund 4 Millionen Menschen entspricht. (Stand 2017)

52%

Etwas mehr als die Hälfte der Pendelnden benutzte 2017 das Auto für den Arbeitsweg, 31% benutzten den öffentlichen Verkehr und 15% gingen zu Fuss zur Arbeit oder fuhren mit dem Velo.

15

Pro Arbeitsweg (ein Hinweg) legten die Pendlerinnen und Pendler 2017 rund 15 Kilometer zurück und benötigten dafür 31 Minuten.

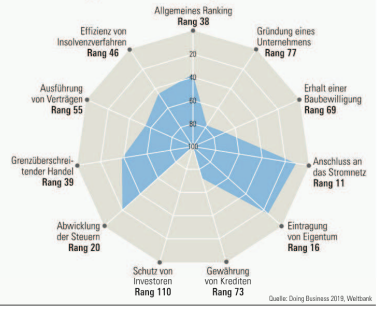


2. Geschäftsklima

38. Rang

Im Weltbank-Index «Ease of Doing Business» 2019 zum Geschäftsklima für Unternehmen belegt die Schweiz im internationalen Vergleich Rang 38. Sie hat damit gegenüber dem Vorjahr fünf Plätze eingebüsst. Das Ranking wird angeführt von Neuseeland und Singapur.

Platzierung der Schweiz im Vergleich von 190 Ländern

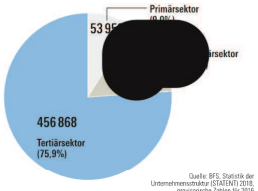


Quelle: Doing Business 2019, Weltbank

3. Gesamtanzahl Unternehmen

601 755

Unternehmen gab es 2016 in der Schweiz, über zwei Drittel davon stammten aus dem tertiären Sektor (Dienstleistungen), etwa 15% aus dem Sekundär- (Industrie, Kunst und Gewerbe) und nur 9% entfielen auf den Primärsektor (Land-, Forstwirtschaft und Fischerei).

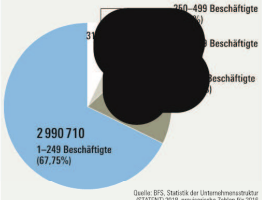


Quelle: BFS, Statistik der Unternehmensstruktur (STATENT) 2018, provisorische Zahlen für 2016

4. Anzahl Arbeitsplätze

4 400 000

99,7% aller marktwirtschaftlichen Unternehmen in der Schweiz zählen weniger als 250 Beschäftigte (KMU). Sie stellen dabei rund zwei Drittel der insgesamt rund 4,4 Millionen Arbeitsplätze in der Schweiz.

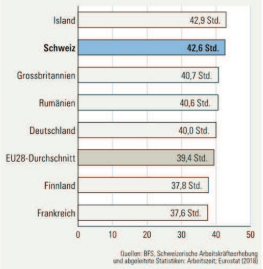


Quelle: BFS, Statistik der Unternehmensstruktur (STATENT) 2018, provisorische Zahlen für 2016

11. Arbeitszeit pro Woche

42,6

2017 wurde gemäss EU28/EFTA-Ländervergleich in Island mit 42,9 Stunden am längsten gearbeitet pro Woche, gefolgt von der Schweiz (42,6), Grossbritannien (40,7) und Rumänien (40,6). In Deutschland lag die wöchentliche Arbeitszeit bei 40 Stunden. Am Schluss der Rangliste lagen Finnland (378) und Frankreich (376).



Quellen: BFS, Schweizerische Arbeitskräfteerhebung und abgeleitete Statistiken: Arbeitszeit; Eurostat (2018)

7 861 000 000

2017 wurden in der Schweiz insgesamt 7,861 Milliarden Arbeitsstunden geleistet.

Quelle: BFS, Schweizerische Arbeitskräfteerhebung und abgeleitete Statistiken: Arbeitszeit

12. Lehrstellen

218 539

2017 wurden in der Statistik der beruflichen Grundbildung 218 539 Lehrverhältnisse registriert. Damit ist etwa jeder 23. Beschäftigte in der Schweiz ein Lernender bzw. eine Lernende. Den höchsten Anteil erreichen mit über 90% die Berufl ichen Grundbildungen mit EFZ (Eidgenössisches Fähigkeitszeugnis).

Ausbildungstyp	2005	2010	2015	2017
Total	200 166	225 530	221 435	218 539
Männer	119 970	131 086	128 918	128 343
Frauen	80 196	94 444	92 517	90 196
> Berufliche Grundbildung mit EFZ	178 333	200 162	208 159	205 013
> Handelsmittelschulen	10 068	12 611	–	–
> Berufliche Grundbildung mit EBA	1 604	8 457	12 946	13 466
> Anlehre	4 686	2 020	234	–
> Nicht BBG-reglementierte BGB	5 475	2 280	96	60

Stand: Mai 2018 Quelle: BFS, Statistik der beruflichen Grundbildung (SBG-SFP)

13. Neue Stellen pro Tag

+100

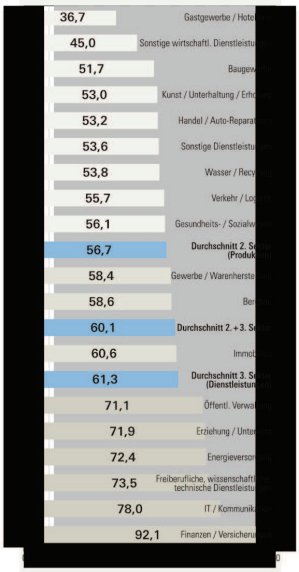
Der Schweizer Arbeitsmarkt ist von einer hohen Dynamik geprägt: Jeden Tag werden hierzulande durchschnittlich etwa 1350 Stellen geschaffen, davon 37% durch neue Unternehmen. Gleichzeitig gehen rund 1250 Jobs verloren. Im Schnitt kommen somit täglich rund 100 Stellen (netto) hinzu.

Quelle: Economiesuisse (2017)

14. Kosten pro Arbeitsstunde

60,05

Eine Arbeitsstunde in der Schweiz kostete 2016 durchschnittlich 60,05 Franken in Unternehmen des sekundären und tertiären Sektors.



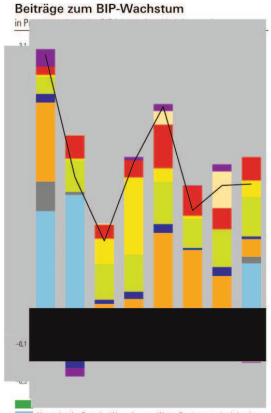
Quelle: BFS, Strukturelle Arbeitskostenerhebung (2018)

15. Bruttoinlandprodukt

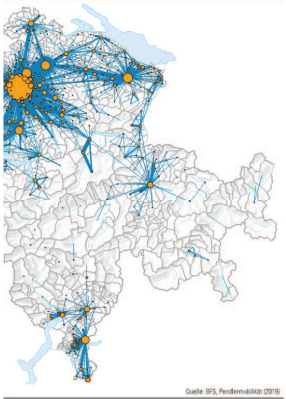
668 572 000 000

Das Bruttoinlandprodukt der Schweiz betrug 2017 insgesamt 668,572 Milliarden Franken. Die Agrarwirtschaft trug rund 0,7% zur Bruttowertschöpfung bei, die Industrie 25,5% und der Dienstleistungssektor 73,8%.

Beiträge zum BIP-Wachstum



Quelle: BFS, Volkswirtschaftliche Gesamtrechnung (2018)

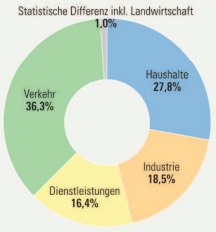


Quelle: BFS, Pendlermobilität (2016)

19. Energieverbrauch

296 310

Der Gesamtenergieverbrauch des Dienstleistungssektors lag 2017 bei 139 230 Terajoule (TJ), jener des Industriesektors bei 157 080 TJ. Zusammen ergibt dies einen Verbrauch von 296 310 TJ. 1 Terajoule entspricht 277 778 Kilowattstunden.



Quelle: BFE, Schweizerische Gesamtenergiestatistik 2017

20. Firmennachfolge

73 786

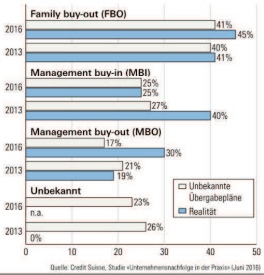
Gemäss einer Analyse von Bisnode D&B müssen 73 786 Schweizer Unternehmen in den nächsten Jahren ihre Nachfolge regeln. Betroffen sind rund eine halbe Million Arbeitsplätze. 13,4% der Firmen haben ein Nachfolgeproblem, insbesonders kleine Betriebe bis neun Mitarbeitende.

Mitarbeitende	Analysierte Unternehmen	Unternehmen mit offener Nachfolge	Prozentualer Anteil
1 bis 9	501 800	68 026	13,6
10 bis 49	41 011	5 306	12,9
50 bis 249	6 591	454	6,9
Total	549 402	73 786	13,4

Quelle: Bisnode D&B, Nachfolge-Studie KMU Schweiz 2018

45%

Rund 45% der Unternehmen werden im Kreis der Familie übertragen («Family buy-out»), 30% an Mitarbeitende, die nicht zur Familie gehören («Management buy-out») und 25% an eine firmenexterne Person («Management buy-in»).



Quelle: Credit Suisse, Studie «Unternehmensnachfolge in der Praxis» (Juni 2018)

9.2 Beschreibung von Verteilung

Mit Hilfe der grafischen Darstellung lässt sich schon sehr viel über die Verteilung von Daten aussagen. Dennoch werden die meisten Statistiken mit sogenannten Masszahlen oder Parametern belegt, mit denen sich die Eigenschaften der Verteilungen in komprimierter Form darstellen lassen. Diese Masszahlen lassen sich auch in einer geeigneten Form, einem Box-Plot, visualisieren.

9.2.1 Lagemasse

Arithmetisches Mittel (Mittelwert)

Eines der bekanntesten Lagemasse ist das arithmetische Mittel. Man erhält dieses, indem man alle Werte aufsummiert und diese Summe durch die Anzahl der Werte dividiert.

Beispiel 9.2

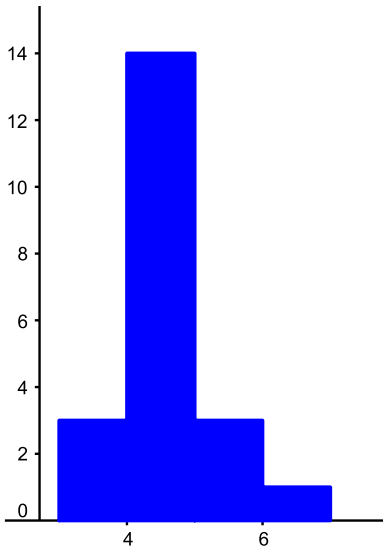
Folgende Noten wurden im Fach Sport erreicht: 5, 4, 4.5, 6, 3, 4 and 5.5

$$\overline{x} = \frac{5 + 4 + 4.5 + 6 + 3 + 4 + 5.5}{7} = 4.57$$

$\Rightarrow$ Mittelwert $\overline{x} = 4.57$

Mittelwert bei Klassenbildung

Gegeben sind hier einige Daten, zum Beispiel eines Weitsprungvergleichs:
3.2; 3.7; 3.5; 4.1; 4.1; 4.2; 4.3; 4.3; 4.3; 4.3; 4.5; 4.6; 4.6; 4.6; 4.8; 4.8; 4.8; 5.1; 5.1; 5.1; 6.2
Diese Daten haben den Mittelwert 4.49. Eine sinnvolle Darstellung ist ein Histogramm mit Klassenbreite 1, also so wie rechts. Ist nur die Graphik vorhanden, aber nicht alle Daten, lässt sich damit auch ein näherungsweiser Mittelwert berechnen, der *Mittelwert nach Klasseneinteilung*: Es wird davon ausgegangen, dass alle Einträge in der Klasse genau in der Mitte liegen. Hier ergibt sich damit: $\frac{3 \cdot 3.5 + 14 \cdot 4.5 + 3 \cdot 5.5 + 1 \cdot 6.5}{21} = 4.59$ Der Mittelwert wird also gar nicht so schlecht abgeschätzt.



Median

in Nachteil des arithmetischen Mittels ist die Aussreisserempfindlichkeit. Wenn ein Wert extrem nach oben oder nach unten ausreisst, wird der Mittelwert sehr stark beeinflusst. Beispiel: in einem kleinen Dorf wohnen 4 Familien: Drei Bauernfamilien mit einem Jahreseinkommen von 37000, 66000 und 73000 Franken. Ausserdem wohnt dort ein Millinär mit einem Jahresverdienst von 40 000 000 Franken. Im Mittelwert sind die Dorfbewohner sehr reich. Aber ...

Der Median ist ein Lagemass, welches nicht stark auf Ausreisser reagierrt. Der Median wird so in der Datenmitte platziert, dass eine Hälfte der Daten oberhalb und die andere Hälfte unterhalb des Medians liegt. Dazu muss man aber zunächst alle Werte x_1, x_2, ..., x_n der Grösse nach ordnen. Dies ergibt eine *geordnete Liste*. Wenn die geordnete Liste aus einer geraden Anzahl Werten besteht, wird der Median aus dem Mittelwert der beiden mittleren Werte gebildet.

Beispiel 9.3

Folgende Noten wurden im Fach Sport erreicht: 5, 4, 4.5, 6, 3, 4 and 5.5

geordnete Liste: 3, 4, 4, 4.5, 5, 5.5 and 6

$\Rightarrow$ Median $\widetilde{x} = 4.5$

Beispiel 9.4

Die vorige Liste wird um eine 5 ergänzt. 5, 4, 4.5, 6, 3, 4, 5.5 and 5

geordnete Liste: 3, 4, 4, 4.5, 5, 5, 5.5 and 6

$\Rightarrow$ Median $\widetilde{x} = (4.5 + 5)/2 = 4.75$

Ein Vorteil, neben der bereits erwähnten Robustheit, ist seine einfache Interpretierbarkeit. Zum Beispiel sagt mir auf der Suche nach einer Wohnung der Median, dass 50 % aller Wohnungen günstiger, 50 % aber teurer sind, der Mittelwert hingegen kann von extrem billigen bzw. teuren Wohnungen beeinflusst werden.

Modus

Ein letztes wichtiges Lagemass ist der Modus. In der Darstellung mit einem Stamm-Blatt-Diagramm ist der Modus die Ausprägung mit dem längsten Balken.

Beispiel 9.5

Folgende Noten wurden im Fach Sport erreicht: 5, 4, 4.5, 6, 3, 4 and 5.5

Die Note 4 kommt zweimal vor

$\Rightarrow$ Der Modus beträgt $x_{mod} = 4$

Der Modus bei den Regentagen in Glasgow ist übrigens 17.

Das arithmetische Mittel und der Median stimmen bei den meisten Statistiken mit keiner der möglichen Ausprägungen überein. So besitzt kein Schweizer Bürger 8.45 Bücher, auch wenn dies als arithmetisches Mittel einer Stichprobe resultieren mag.

9.2.2 Quantile und Box-Plot

Quantile und Boxplots sind geeignete Mittel, Daten und ihre Streuung grafisch darzustellen. Der Boxplot ist eine erste Möglichkeit, die Streuung der Daten festzuhalten.

Quantile

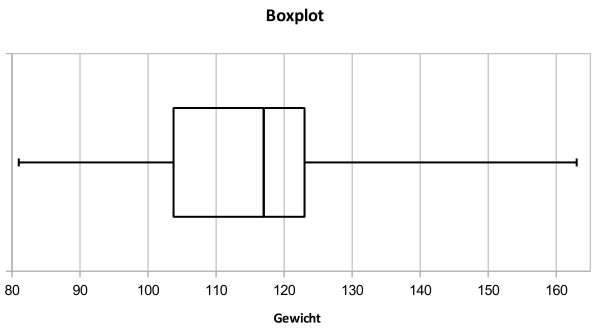
Definition 9.1

Das erste Quartil q_1, auch bekannt als 25 %-Quantil, ist der Median der Ergebnisse links vom Median $\widetilde{x}$. Das dritte Quartil q_3, auch bekannt als 75 %-Quantil, ist der Median der Ergebnisse rechts vom Median $\widetilde{x}$.

Etwas salopper gesagt: die Quartile teilen die Daten in vier Abschnitte mit jeweils gleich vielen Einträgen ein. Der Median ist also das 50 %-Quantil.

Box-Plot

Das Minimum, das untere Quartil, der Median, das obere Quartil und das Maximum bilden zusammen fünf Werte. Diese Werte führen zur komprimierten Visualisierung einer Verteilung in einem Boxplot. Man erhält damit eine grafische Darstellung der Daten, die sehr gut zum Vergleich verschiedener Verteilungen geeignet ist. Es lässt sich so sehr schnell erkennen, ob die Beobachtungen annähernd symmetrisch verteilt sind oder Ausreisser in dem Datensatz auftreten.



Wir erkennen im Boxplot, basierend auf den Daten aus Kapitel 17.1, dass 50 % aller Frauen mit hohem Blutdruck ein Gewicht zwischen 102 und 122 kg haben. Ausserdem zeigt der Boxplot die enorme Streuung in den obersten 25 % der Probandinnen.

Definition 9.2

a) $x_{0.25}$ = Anfang der Schachtel (Box);
$x_{0.75}$ = Ende der Schachtel;
IQR = Länge der Schachtel (siehe Auftrag **??**)

b) Der Median wird durch einen Punkt oder einen vertikalen Strich in der Box markiert.

c) Zwei Linien ausserhalb der Box gehen bis zu x_{min} und x_{max}.

Der Boxplot gibt eine erste Abschätzung der Streuung der Daten: wie weit sind die Quartile voneinander entfernt?

Die Differenz von grösstem und kleinsten Wert heisst übrigens «Spannweite»

9.2.3 Streuung

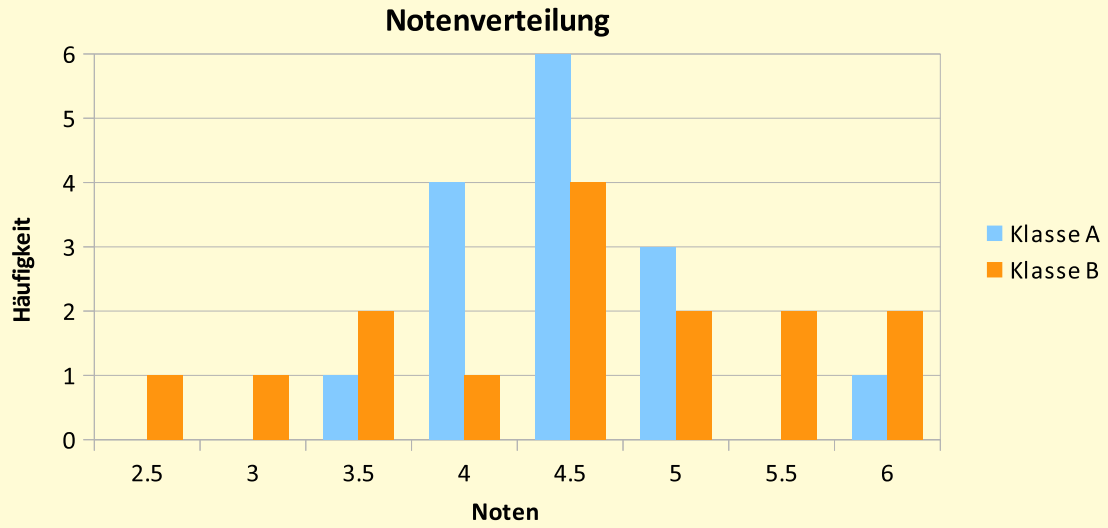
Auftrag 9.4

Lesen Sie das folgende Beispiel gut durch, diskutieren Sie es mit Ihrem Tischnachbarn. Die Lagemasse können Sie bereits berechnen, die Streuung sollten Sie kreativ diskutieren.

Eine Lehrerin führt in zwei Parallelklassen die gleiche Prüfung durch. Beide Klasse umfassen je 15 Schülerinnen und Schüler. Sie erzielen folgende Noten:

Klasse A	4.0	5.0	4.5	4.5	4.0	4.5	4.0	5.0	6.0	4.5	4.5	5.0	4.0	4.5	3.5
Klasse B	4.5	3.5	5.5	3.0	4.0	4.5	3.5	4.5	5.5	6.0	4.5	6.0	5.0	2.5	5.0

Wir stellen die beiden Notenverteilungen in einem Säulendiagramm dar.



Nun sortieren wir die Notenlisten, die tiefste Note links, die höchste rechts.

Klasse A	3.5	4.0	4.0	4.0	4.0	4.5	4.5	4.5	4.5	4.5	4.5	5.0	5.0	5.0	6.0
Klasse B	2.5	3.0	3.5	3.5	4.0	4.5	4.5	4.5	4.5	5.0	5.0	5.5	5.5	6.0	6.0

1.Quartil Median = 2.Quartil 3.Quartil

Beide Notenverteilungen haben ungefähr die gleiche Lage. Das Zentrum liegt jeweils bei ca. 4.5. Die beiden Notenverteilungen haben deutlich unterschiedlich grosse Streuungen. Die Streuung in der Klasse A ist deutlich kleiner als in der Klasse B.

In beiden Klassen beträgt der Median je 4.5: Die Hälfte der Noten liegt jeweils bei 4.5 oder darüber, die andere Hälfte bei 4.5 oder darunter.
Auch der Mittelwert beträgt in beiden Klassen je 4.5: in der Klasse A ergibt sich für den Mittelwert

$$\frac{1 \cdot 3.5 + 4 \cdot 4.0 + 6 \cdot 4.5 + 3 \cdot 5.0 + 1 \cdot 6.0}{15} = 4.5$$

in der Klasse B

$$\frac{1 \cdot 2.5 + 1 \cdot 3.0 + 2 \cdot 3.5 + 1 \cdot 4.0 + 4 \cdot 4.5 + 2 \cdot 5.0 + 2 \cdot 5.5 + 2 \cdot 6.0}{15} = 4.5$$

Das wichtigste Streuungsmass ist die Standardabweichung, manchmal auch einfach Streuung genannt. Zunächst werden die Abstände vom Mittelwert von jedem Eintrag gebildet. Dann werden wie beim Satz von Pythagoras deren Quadrate summiert, und daraus eine Wurzel gezogen. Die daraus folgende Formel ist zunächst einmal kompliziert, hat aber wichtige Eigenschaften in der fortgeschritteneren Statistik, die dieses Mass wichtig werden lassen.

Definition 9.3

Die Standardabweichung s des Datensatzes, der aus den Daten x_1 bis x_n besteht, wird folgendermassen errechnet:

a) Der Mittelwert $\overline{x}$ der Daten wird gebildet.

b) Die Abstände der Daten vom Mittelwert werden berechnet, quadriert und aufsummiert:
$(x_1 - \overline{x})^2 + \ldots (x_n - \overline{x})^2$

c) Alles wird durch die Zahl der Daten geteilt und die Wurzel wird gezogen:

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + \ldots (x_n - \overline{x})^2}{n}}$$

Beispiel 9.6

In der Klasse A beträgt

$$s = \sqrt{\frac{(3.5 - 4.5)^2 + \ldots + (6.0 - 4.5)^2}{14}} = \sqrt{\frac{1 \cdot 1.5^2 + 1 \cdot 1.0^2 + 7 \cdot 0.5^2 + 6 \cdot 0.0^2}{14}} \approx 0.6$$

In der Klasse B

$$s = \sqrt{\frac{(2.5 - 4.5)^2 + \ldots + (6.0 - 4.5)^2}{14}} = \sqrt{\frac{1 \cdot 2.0^2 + 3 \cdot 1.5^2 + 4 \cdot 1.0^2 + 3 \cdot 0.5^2 + 4 \cdot 0.0^2}{14}} \approx 1.0$$

Das Quadrat der Standardabweichung heisst übrigens Varianz.

Wichtig zu wissen ist noch, dass bei der Untersuchung von Stichproben (statt aller SchülerInnen einer Schule werden zum Beispiel nur 20 untersucht) oft durch $n-1$ statt durch n geteilt wird. Das heisst dann empirische Standardabweichung.

$$\sqrt{\frac{(x_1-\overline{x})^2+...+(x_n-\overline{x})^2}{n-1}}$$

9.3 Übungen

6. Berechnen Sie den Mittelwert und den Median.

a) 165 cm, 165 cm, 167 cm, 169 cm, 170 cm, 172 cm, 174 cm, 175 cm, 179 cm and 180 cm

b) 51.2 kg, 53.5 kg, 54.6 kg, 59.8 kg, 60.4 kg, 62.3 kg, 65.7 kg and 70.3 kg

c) 2.8 m, 3.2 m, 3.6 m, 3.7 m, 3.9 m, 4.5 m, 4.6 m, 4.7 m, 4.8 m, 5.1 m and 6.3 m

7. Bei einem 100-m-Lauf werden in der Klasse F2a folgende Zeiten (in Sekunden) gestoppt:

11.8, 12.0, 12.2, 12.2, 12.5, 12.6, 12.7, 12.8, 12.8, 12.9, 13.1, 13.1, 13.3, 13.4, 13.5, 13.6, 13.6, 13.8, 13.9, 14.2, 14.3, 14.5, 14.5, 14.5, 14.7, 15.0, 15.1 and 15.1

a) Nehmen Sie eine Klasseneinteilung vor.

b) Berechnen Sie den Mittelwert nach Klasseneinteilung.

c) Berechnen Sie den Mittelwert ohne Klasseneinteilung.

Klasse (Sekunden)	$11.5 \leq x < 12$	$12 \leq x < 12.5$	
Anzahl			
Anteil in %			

8. Teilen Sie mit, wenn eine der Aufgaben nicht lösbar ist. Lösen Sie die Aufgabe andernfalls.

Geben Sie jeweils Merkmalsausprägungen einer beliebigen Stichprobe an, so dass

a) der Mittelwert 10 und der Median 1 ist.

b) der Mittelwert keinem der Ergebnisse entspricht. Welches ist der Mittelwert der Ergebnisse?

c) der Median keinem der Ergebnisse entspricht. Welches ist der Median der Ergebnisse?

d) der Modus keinem der Ergebnisse entspricht. Welches ist der Modus der Ergebnisse?

9. Drei Parallelklassen führen einen Test in Mathematik durch. Dabei werden von den 64 Schülerinnen und Schülern die folgenden Punkte erreicht (das Punktemaximum ist 40):

26, 22, 15, 17, 16, 28, 33, 22, 23, 20, 27, 16, 10, 26, 25, 26, 33, 35, 28, 13, 23, 21, 12, 24, 18, 28, 30, 19, 26, 22, 15, 17, 31, 23, 25, 11, 8, 20, 16, 32, 22, 19, 38, 6, 34, 21, 26, 37, 27, 18, 29, 16, 19, 15, 23, 13, 32, 34, 20, 18, 18, 10, 27 and 26

a) Ordnen Sie die erreichten Punktezahlen der Grösse nach, berechnen Sie den Mittelwert und den Median.

b) Stellen Sie die Punkteverteilung dreimal grafisch dar. Wählen Sie zuerst Klassenbreite 2, dann 4, dann 8. Welche Klassenbreite gibt die aussagekräftigste Grafik?

10. Der Body-Mass-Index $\text{BMI}=\frac{\text{Gewicht in kg}}{(\text{Grösse in m})^2}$ dient der Bewertung des Körpergewichts in Relation zur Körpergrösse. Für Männer und Frauen gelten folgende Werte: Untergewicht für $\text{BMI} < 20$, Normalgewicht BMI = 20 to 25, Übergewicht BMI = 25 to 30, Adipositas $\text{BMI} > 30$.
Im Folgenden sind die BMI-Werte von 60 Männern aufgelistet, alle zwischen 20 und 60 Jahre alt, der Grösse nach geordnet.

18.8, 20.1, 20.3, 20.5, 20.6, 21.0, 21.1, 21.3, 21.3, 21.4, 21.5, 21.9, 22.3, 22.8, 22.9, 23.0, 23.2, 23.2, 23.2, 23.3, 23.4, 23.5, 23.5, 23.6, 23.6, 23.7, 24.5, 24.5, 24.6, 24.8, 24.9, 25.1, 25.4, 25.5, 25.8, 25.8, 26.1, 26.1, 26.2, 26.3, 26.4, 26.5, 26.6, 26.8, 27.2, 27.3, 27.5, 27.6, 28.1, 28.3, 28.4, 28.9, 29.1, 29.2, 29.2, 29.5, 30.4, 31.1, 31.9 and 35

a) Bestimmen Sie Median und Mittelwert sowie den Prozentsatz der Männer mit Übergewicht.

b) Suchen Sie eine geeignete Klasseneinteilung, ermitteln Sie die zugehörige Häufigkeit, zeichnen Sie ein geeignetes Diagramm.

c) Erstellen Sie zu den gegebenen Werten einen Boxplot.

11. Die Polizei hat bei Autokontrollen 30 Autoreifen auf ihre Profiltiefe (in mm) überprüft. Folgende Liste ergab sich:
3.5, 4.6, 3.6, 2.3, 4.9, 2.3, 4.6, 7.5, 4.1, 2.3, 3.4, 5.6, 4.3, 2.3, 3.4, 2.3, 4.3, 4.6, 2.3, 3.5, 6.2, 3.2, 1.3, 2.3, 2.4, 5.1, 2.3, 4.5, 6.3 and 2.3
Berechnen Sie die mittlere Profiltiefe und die Standardabweichung dieser 30 Autoreifen. Bestimmen Sie auch den Median und den Interquartilabstand.

12. Die Leute in einer Disco wurden nach ihrem Alter gefragt.

Alter	14	15	16	17	18	19	20	21	22	23	24	25
Anzahl	6	10	12	15	17	14	10	7	4	2	1	1

Berechnen Sie das durchschnittliche Alter und die Standardabweichung. Interpretieren Sie die Werte.

13. Kira hat 100-mal zwei Würfel gleichzeitig geworfen und dabei die Augensumme festgehalten. Werten Sie ihre Versuchsergebnisse aus: Mittelwert, Median, Spannweite und Standardabweichung.

Summe	2	3	4	5	6	7	8	9	10	11	12
Anzahl	1	2	5	10	22	21	19	12	7	1	0

9.4 Exkurse: Vergleich von Datensätzen

Auftrag 9.5: Muskelzuwachs

Eine Forschungsanstalt testete zum Aufbau von Muskelmasse zwei unterschiedliche Trainingsprogramme im Kraftraum. Untersucht wurde dabei die Muskelzunahme am Oberschenkel nach einer vorgängig während 4 Wochen medizinisch zwangsbedingten Ruhestellung (z.B. durch Gips, Schiene, etc.) des rechten Beines.
Die beiden Trainingsprogramme I und II wurden jeweils mit 80 Probanden (Sportler im Alter zwischen 25-30 Jahren, ungefähr gleicher Statur) während 6 Wochen durchgeführt. Anschliessend wurde die Zunahme des Oberschenkelumfanges gemessen.
Die Auswertung hat nun folgendes ergeben:

Versuch Muskelaufbau		
Muskelzunahme (Umfangzunahme am Oberschenkel in cm) Klassenmitten	Trainingsprogramm I Absolute Häufigkeit (Anzahl Personen)	Trainingsprogramm II Absolute Häufigkeit (Anzahl Personen)
1.5	25	15
3.0	30	35
4.5	10	10
6.0	5	7
7.5	10	13

a) Zeichnen Sie jeweils ein Histogramm mit den relativen Häufigkeiten für das Trainingsprogramm I und das Trainingsprogramm II.

b) Bestimmen Sie für das Trainingsprogramm I: Mittelwert, Median, Modus, Spannweite und Standardabweichung.
Verwenden Sie anstelle der Einzelwerte die vorgegebenen Klassenmitten aus der Tabelle.

c) Das Trainingsprogramm II hat folgende Werte ergeben:
Mittelwert $\overline{x}$ = 3.9 cm, Median $\widetilde{x}$ = 3 cm, Modus= 3 cm,
Spannweite R = 6 cm und Standardabweichung s = 2.0054 cm.

Vergleichen Sie die Wirkung der zwei Trainingsprogramme: Welches Programm würden Sie einzig unter der Berücksichtigung der unter a) bis c) aufgearbeiteten bzw. ermittelten Daten in Zukunft einsetzen?
Begründen Sie Ihre Auswahl!

d) Bei einer anderen Untersuchung wurden die beiden Trainingsprogramme mit je 220 Probanden mit unterschiedlichen Körpergrössen (155 cm bis 195 cm), unterschiedlichen Gewichten (50 kb bis 110 kg) und unterschiedlichen Alter (22 bis 75 Jahre) durchgeführt.
Die statistische Auswertung erfolgte ohne die Berücksichtigung der unterschiedlichen Körpermerkmale (Körpergrösse, Gewicht, Alter, etc.). Wie beurteilen Sie bei dieser Untersuchung das Vorgehen bei der Datenermittlung und die Aussagekraft der Resultate?
Begründen Sie Ihre Antwort mit wenigen prägnanten und aussagekräftigen Sätzen.

Auftrag 9.6: Düngungsmittel

Bei einer Forschungsanstalt wurden zwei neue Düngemittel zur Wachstumsförderung von Weizen eingesetzt. Anschliessend wurde nach einer bestimmten Zeit das Wachstum der einzelnen Pflanzen in cm gemessen und in Klassen eingeteilt:

Versuch Düngemittel		
Wachstum in cm (Klassenmitten) Klassenmitten	Düngemittel DM I Absolute Häufigkeit (Anzahl Pflanzen)	Düngemittel DM II Absolute Häufigkeit (Anzahl Pflanzen)
50	5	3
75	35	12
100	32	45
125	18	20
150	10	20

a) Zeichnen Sie ein Histogramm mit den relativen Häufigkeiten für das Düngemittel DM I.

b) Bestimmen Sie für das Düngemittel DM I: Mittelwert, Median, Modus, Spannweite und Standardabweichung.
Verwenden Sie anstelle der Einzelwerte die vorgegebenen Klassenmitten aus obenstehender Tabelle.

c) Für das Düngemittel DM II haben sich folgende Werte ergeben:
Mittelwert $\overline{x}$ = 110.5 cm, Median $\widetilde{x}$ = 100 cm, Modus= 100 cm
Spannweite R = 100 cm und Standardabweichung s = 25.91 cm.
Vergleichen Sie die Wirkung der zwei Düngemittel: welches Mittel würden Sie in Zukunft einsetzen? Begründen Sie Ihre Auswahl!

d) Bei einem weiteren Düngemittel (DM III) wurde bei der statistischen Erfassung „gemogelt“. Bei 172 Werten wurde ein Mittelwert von 92.5 cm und eine Standardabweichung von 22.0 cm ermittelt. Eine Kommission, welche die erfassten Daten überprüft, stellt fest, dass in der geordneten Liste die 5 kleinsten Werte 44.0 cm, 44.5 cm, 45.0 cm, 45.5 cm und 46.5 cm (absichtlich) nicht berücksichtigt wurden.

Wie gross wäre der Mittelwert bei korrekter Berechnung mit allen 177 Werten?

Wie verändert sich die Standardabweichung bei korrekter Berechnung mit allen 177 Werten: wird sie kleiner, grösser oder bleibt sie gleich? Begründen Sie Ihre Antwort in Worten (ohne Berechnung).

Auftrag 9.7: Abfüllmaschine

Ein Getränkehersteller evaluiert den Kauf einer neuen vollautomatischen Abfüllmaschine. Dabei werden bei zwei Herstellern Testserien für das Abfüllen eines Energiedrinks ausgewertet. Massgebend für die Auswahl der Abfüllmaschine ist die Volumenhaltigkeit für eine Getränkedose von 0.33 Liter.
Die Testserien haben nun folgende Resultate geliefert:

Abfüllmaschine CNC0815	
effektives Volumen in dl	Absolute Häufigkeit
3.2	50
3.25	120
3.30	210
3.35	100
3.40	120

Abfüllmaschine CNC2010	
effektives Volumen in dl	Absolute Häufigkeit
3.2	190
3.25	100
3.30	50
3.35	50
3.40	210

a) Bestimmen Sie Mittelwert, Median, Modus und Spannweite für die Abfüllmaschine CNC0815.

b) Bestimmen Sie die Standardabweichung s für die Abfüllmaschine CNC0815.

c) Vergleichen Sie die Resultate aus den beiden Aufgaben a) und b) mit den nachfolgenden Werten für CNC2010:
Mittelwert $\overline{x}$ = 3.299 dl, Median $\widetilde{x}$ = 3.3 dl, Modus= 3.4 dl, Standardabweichung s = 0.085458 dl.

Welche Abfüllmaschine würden Sie in Zukunft einsetzen? Begründen Sie Ihre Auswahl.

d) Erstellen Sie ein Histogramm mit den relativen Häufigkeiten für die Abfüllmaschine CNC0815.

e) Die Abfüllmaschine einer dritten Firma wurde aus dem Evaluationsverfahren ausgeschlossen, weil der zuständige Sachbearbeiter aufgrund einer Bestechung die Testserie manipulierte.

Folgende manipulierten Daten, ermittelt aus 600 Werten, wurden abgeliefert:

Mittelwert $\overline{x}$ = 3.3 dl

Standardabweichung s = 0.08 dl

Median $\widetilde{x}$ = 3.3 dl

Modus= 3.4 dl

Die Auswahlkommission stellte fest, dass 150mal an Stelle des Wertes 3.2 dl der Wert 3.3 dl für die Statistik verwendet wurde.

i) Welche der nachfolgenden Werte Mittelwert, Median, Modus und Standardabweichung können Sie anhand der unter e) vorgegebenen Informationen überprüfen und wie gross wären diese bei der korrekten Berechnung?

ii) Über welche Werte, die Sie nicht berechnen können, können Sie eine quantitative Aussage machen und welche?

9.5 Projekt: Lineare Regression und Korrelation

Im einleitenden theoretischen Teil werden Werkzeuge zur Verfügung gestellt, die für die Untersuchung zweier Datensätze wichtig sind. Damit sollen dann in einem Projekt Zusammmenhänge untersucht werden. Dies soll mit selbst erhobenen Daten geschehen. Beispiele für Fragestellungen, denen nachgegangen werden kann:

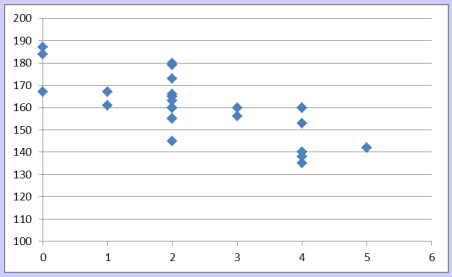
- Schneiden diejenigen, die in der Woche häufiger Sport treiben, beimm 30Min-Lauf besser ab?
- Haben diejenigen, die beim 30Minuten-Lauf weiter kommen, einen niedrigeren Ruhepuls?

9.6 Punktwolken

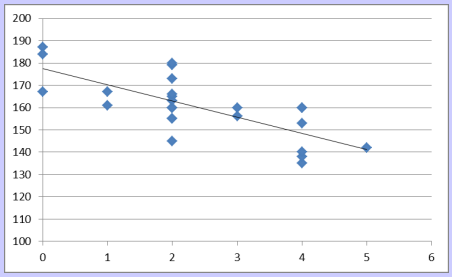
Abhängigkeiten zwischen Merkmalen lassen sich oft aus dem Sachkontext heraus vermuten: Das Gewicht einer Person hängt vermutlich (auch) von ihrer Grösse ab; wer mehr verdient, hat eher eine grössere Wohnung als einer, der wenig verdient; je stärker man ein Gewebe radioaktiv bestrahlt, desto stärker ist der Schaden. In all diesen Fällen scheint ein Zusammenhang zu bestehen. Ob der aber tatsächlich besteht und – wenn ja – ob er dann auch linear ist, soll nun mit statistischen Methoden untersucht werden. Ein erster Schritt besteht darin, sich einen grafischen Überblick über Daten zu verschaffen, die zu zwei Merkmalen gehören.

Beispiel 9.7

Hier handelt es um Daten von FachmittelschülerInnen aus Oberwil aus dem Jahr 2013. Die SchülerInnen haben angegeben, wie oft sie Sport treiben in der Woche. Ausserdem sind sie schnell drei Etagen die Treppen hoch gelaufen, und haben ihren Puls gemessen. In der Graphik repräsentiert jeder Punkt einE SchülerIn. Auf der x-Achse ist die Anzahl Sport pro Woche ersichtlich, auf der y-Achse der Puls.



Aus physiologischen Gründen lässt sich vermuten, dass es einen Zusammenhang gibt: wer oft Sport treibt, bei dem wird der Puls nicht so stark nach oben getrieben beim Treppenlauf. Der Zusammenhang ist allerdings nicht eindeutig. Es gibt Leute mit niedrigem Puls und wenig Sport in der Woche. Einen Trend gibt es aber schon, wie die hier von der Tabellenkalkulation erstellte Gerade in der Graphik zeigt.



Wie sich diese Trendlinien zeichnen lassen, wird nun herausgearbeitet, mit Hilfe eines kleineren Datensatzes.

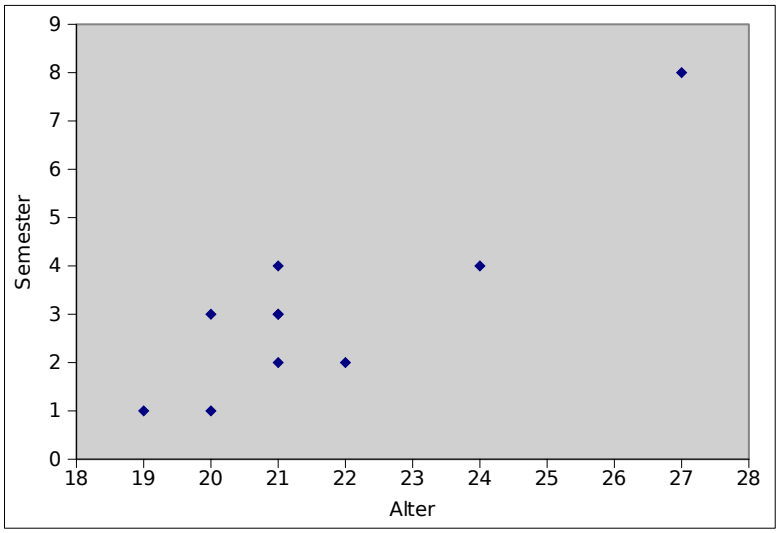
Beispiel 9.8

Wir beginnen mit einem kleinen Datensatz, in dem das Alter und die Semesterzahl einer Studentengruppe erhoben ist. Man vermutet: Je älter ein Student ist, desto höher ist sein Semester. Auch dass dieser Zusammenhang linear sein könnte, ist nicht unplausibel: Sieht man von Studienwechsel und Auslands-, Urlaubs- und Freisemestern ab, so erhöhen sich Alter und Semesterzahl proportional zueinander, sogar mit dem Proportionalitätsfaktor 1.

Alter	20	21	24	21	22	21	27	19	20	21
Semester	3	3	4	2	2	3	8	1	1	4

Wenn ein linearer Zusammenhang zwischen diesen beiden Merkmalen besteht, so könnte man die zehn Paare von Messwerten in ein Koordinatensystem einzeichnen, und die Datenpunkte müssten dann ungefähr auf einer Geraden liegen. Den Graphen, der aus den Datenpaaren besteht, nennt man *Punktwolke*.

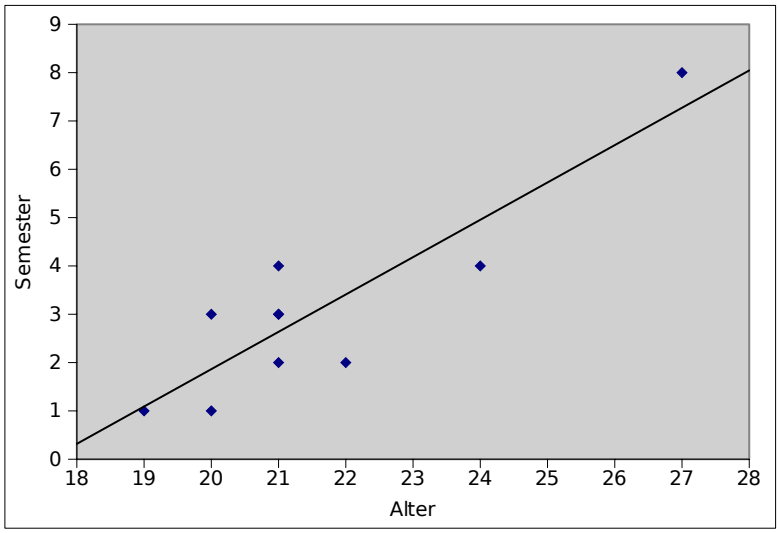
Punktwolken kann man wie jede Menge reeller Zahlenpaare grafisch in einem Koordinatensystem darstellen. In der Abbildung ist die Punktwolke der beiden Merkmale aus der letzten Tabelle in einem kartesischen Koordinatensystem dargestellt.



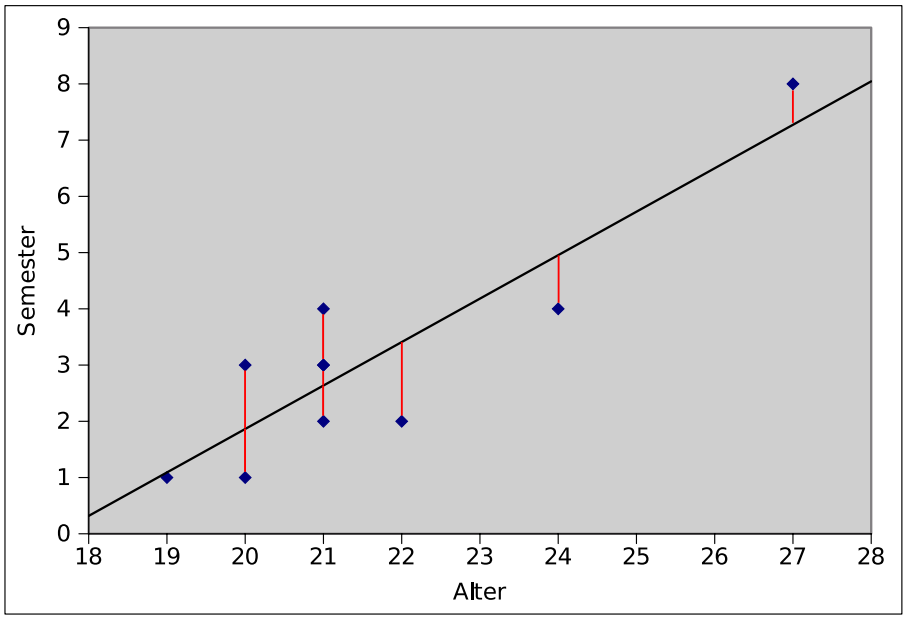
9.6.1 Lineare Regression nach Augenmass

Die Punktwolke in der obigen Abbildung scheint linear anzusteigen. Eine Gerade, auf der alle Punkte liegen, wird man nicht finden – allein schon, weil dem Alter 23 drei verschiedene Semesterzahlen zugeordnet sind. Aufgabe der linearen Regression ist es, dennoch eine Gerade zu finden, welche die Punktwolke möglichst gut annähert, also die lineare Grundtendenz der Punktwolke möglichst gut darzustellen, so dass die Abweichung von dieser Tendenz möglichst gering sind. Das Ziel ist also eine lineare Funktion y_{fit} zu finden, die zu jedem x-Wert x_i einen y-Wert $y_{\text{fit}}(x_i)$ liefert, sodass die Abweichung vom tatsächlich gemessenen y-Wert y_i für alle Messwerte x_i möglichst gering ist. Diese Abweichung $r_i = y_i - y_{\text{fit}}(x_i)$ bezeichnet man als *Residuum*.

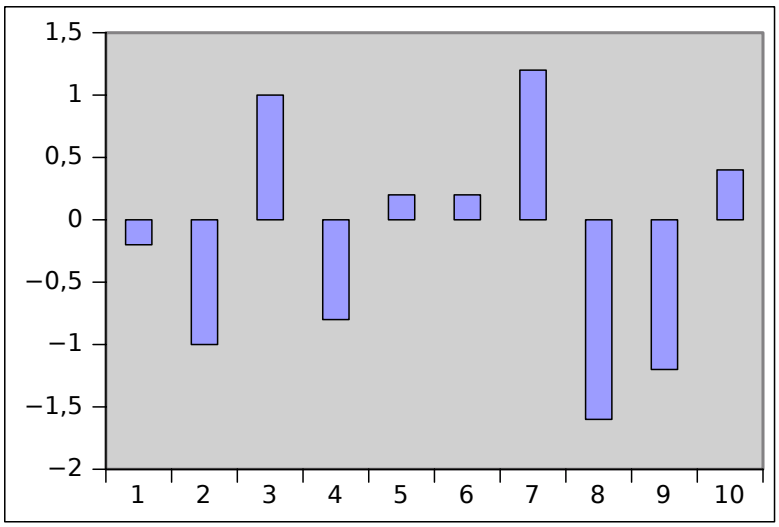
Eine einfache und für Schulzwecke oft ausreichende Methode, eine passende Regressionsfunktion y_{fit} zu finden – insbesondere wenn es sich (vermutlich) um einen linearen Zusammenhang handelt und die Regressionsfunktion eine Gerade ist – besteht darin, den Graphen von y_{fit} nach Augenmass möglichst passend in die Punktwolke einzuzeichnen und aus der grafischen Darstellung die Funktionsgleichung von y_{fit} abzulesen. In der Abbildung **??** ist per Augenmass eine Regressionsgerade eingezeichnet worden, so dass die Gerade möglichst dicht an den Punkten verläuft und möglichst gleichmässig Punkte ober- und unterhalb der Geraden liegen. Als Funktionsgleichung der Regressionsgeraden kann mit den üblichen Verfahren näherungsweise die Funktionsgleichung $y_{\text{fit}} = 0.8x - 14$ aus dem Schaubild ermitteln.



In der nächsten Grafik sind die Residuen $r_i = y_i - y_{\text{fit}}(x_i)$ rot eingezeichnet. Eine Möglichkeit, die Einpassung der Geraden per Augenmass zu verbessern, ist es, die Residuen als Säulendiagramm darzustellen und nach systematischen Fehlern zu suchen.

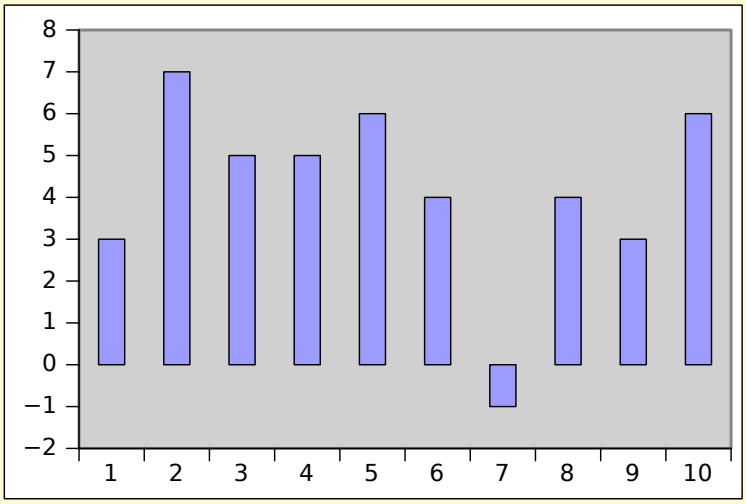


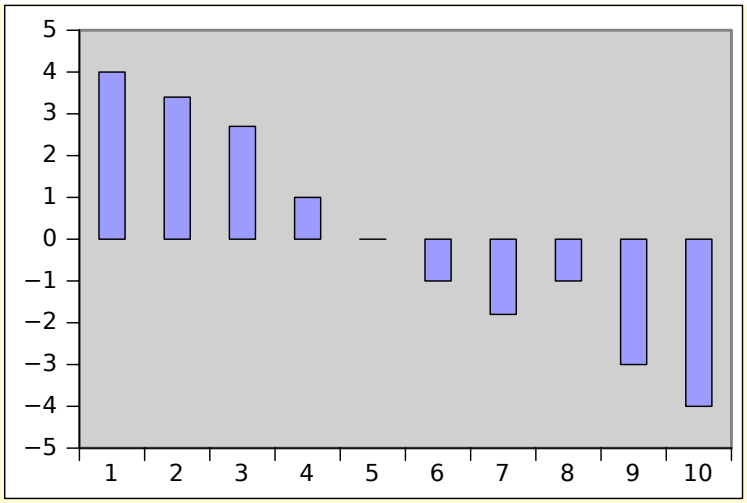
In der nächsten Abbildung sind die Residuen zur Regressionsgeraden aus der Grafik 17.6.1 aufgetragen. Man kann keinen systematischen Fehler erkennen: Die Residuen sind nicht allzu gross und verteilen sich einigermassen gleichmässig im positiven wie im negativen Bereich.



Auftrag 9.8

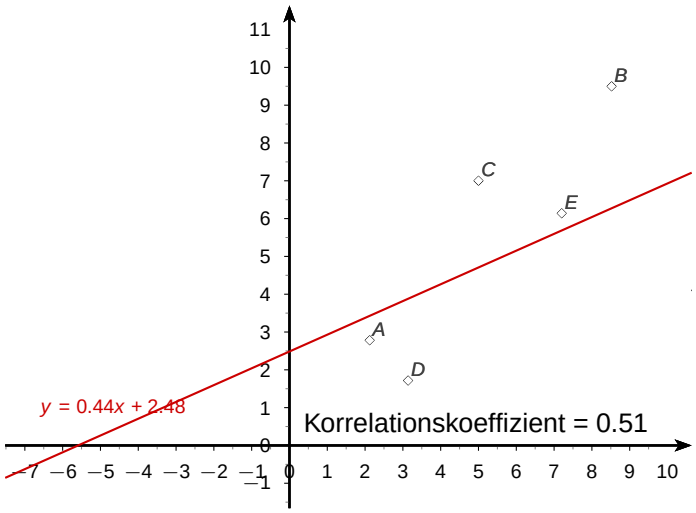
In den Abbildungen 17.8 und 17.8 sehen Sie Residuenplots, die auf einen systematischen Fehler beim Einzeichnen der Regressionsgeraden hindeuten. a) Beschreiben Sie, worin der Fehler besteht; b) erläutern Sie, wie sich dieser Fehler auf die Lage der Regressionsgeraden bezüglich der Punktwolke auswirkt, und c) geben Sie begründet an, wie man die Regressiongeraden verändern sollte, um den Fehler zu vermeiden, und wie sich diese Änderung in der Funktionsgleichung der Regressionsgeraden bemerkbar macht.





9.6.2 Lineare Regression mit der Methode der kleinsten Quadrate

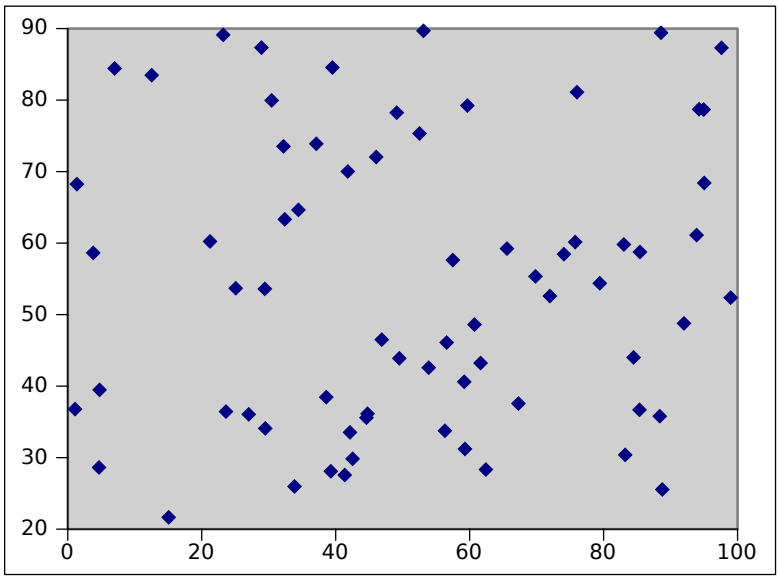
Bei der Einpassung einer guten Regressionsgerade geht es darum, die Residuen möglichst klein zu halten. Um hier die optimale Gerade zu finden, gibt es exakte Formeln (Methode der kleinsten Quadrate), die beispielsweise in Excel, OpenOffice oder Geogebra einprogrammiert sind. Diese Programme können also die beste Regressionsgerade legen, und geben dann auch die Geradengleichung an.



(Gleich ausprobieren: https://tube.geogebra.org/material/simple/id/3260657 Die Zahl Korrelationskoeffizient in der Graphik ist ein Mass dafür, wie gut die Gerade zu den Datenpunkten passt. Das wird im nächsten Abschnitt erklärt.)

9.6.3 Korrelationskoeffizienten

Durch jede Punktwolke lässt sich so eine Regressionsgerade legen – auch wenn das gar nicht sinnvoll ist. In der Abbildung 17.6.3 ist eine Punktwolke eingezeichnet, in der kein linearer Zusammenhang (und auch nicht irgendein anderer) zwischen den x- und y-Werten erkennbar ist. Auch für diese Punktwolke würde die Methode der kleinsten Quadrate eine Regressionsgerade ermitteln.



Wir brauchen ein Mass dafür, wie gut die Regressionsgerade passt. In der Statistik wird dieses Mass durch den *Korrelationskoeffizienten* definiert. Es gibt verschiedene mögliche Definitionen. Gebräuchlich ist die folgende Definition (die hier nie per Hand ausgerechnet wird, das wird dem Computer überlassen).

Definition 9.4

Gegeben sind die Datenpunkte $(x_1, y, 1)$ bis (x_n, y_n). Der Mittelwert des ersten Merkmals ist $\overline{x}$, der des zweiten $\overline{y}$. Der Korrelationskoeffizient nach Pearson ist dann:

$$r = \frac{(x_1 - \overline{x}) \cdot (y_1 - \overline{y}) + ... + (x_n - \overline{x}) \cdot (y_n - \overline{y})}{\sqrt{(x_1 - \overline{x})^2 + ... + (x_n - \overline{x})^2} \cdot \sqrt{(y_1 - \overline{y})^2 + ... + (y_n - \overline{y})^2}}$$

Liegen die Punkte perfekt auf einer Geraden, so ist der Korrelationskoeffizient 1, bzw -1 bei einer Geraden mit negativer Steigung. Es ist üblich, den Korrelationskoeffizienten wie in der folgenden Tabelle abzuschätzen, wobei es sich nur um Richtlinien handelt:

Wertebereich von r	Interpretation
$0.8 \leq r \leq 1$	sehr starke Korrelation
$0.6 < r < 0.8$	starke Korrelation
$0.4 \leq r \leq 0.6$	mittlere Korrelation
$0.2 < r < 0.4$	schwache Korrelation

Die gleichen Bezeichnungen gelten für negative Korrelationen.

9.6.4 Übungen

14. Schwache Korrelation

a) Erstellen Sie ein Streudiagramm mit sechs Datenpaaren einer zweidimensionalen Verteilung. Wählen Sie die Datenpaare so, dass eine schwach negative Korrelation zwischen den Variablen besteht.

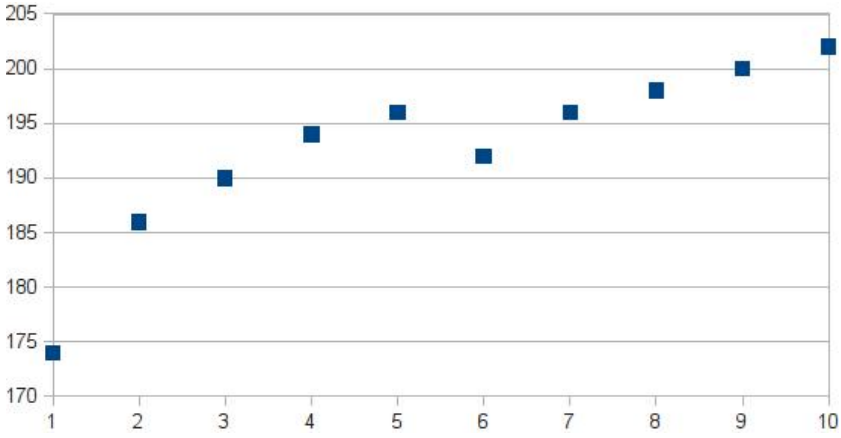
b) Zeichnen Sie per Augenmass eine Regressionsgerade ein.

c) Bestimmen Sie die Residuen und stellen Sie diese in einem Säulendiagramm dar.

d) Beurteilen Sie mit Hilfe der Residuen: Ist ihre Regressionsgerade gut getroffen, wie könnten Sie diese allenfalls verbessern?

15. Fliff ist ein exzellenter Mittelstreckenläufer. Während eines 2500m Laufs wurde nach jeder der zehn Runden sein Puls gemessen. Die Daten sind in der folgenden Tabelle festgehalten.

Runde	1	2	3	4	5	6	7	8	9	10
Puls	174	186	190	194	196	192	196	198	200	202



a) Zeichnen Sie per Augenmass eine Regressionsgerade ein.

b) Bestimmen Sie die Residuen und stellen Sie diese in einem Säulendiagramm dar.

c) Beurteilen Sie mit Hilfe der Residuen: Ist ihre Regressionsgerade gut getroffen, wie könnten Sie diese allenfalls verbessern?

d) Wie gut ist die Korrelation zwischen Rundenzahlen und Pulswerten?

e) Ist es sinnvoll, eine Gerade durch die Punkte zu legen? Welche Kurvenform wäre allenfalls besser geeignet? Warum?

f) Berechnen Sie den Mittelwert und den Median von Fliffs Pulsdaten.

g) Berechnen Sie die Spannweite.

h) Wenn Sie eine der Messungen weglassen würden, würde sich eine viel kleinere Standardabweichun ergeben. Welche Messung ist das? Warum hat diese Messung einen besonders grossen Einfluss auf die Streuung?

i) Berechnen Sie ohne den Taschenrechner zu nutzen möglichst effektiv den Mittelwert der Pulsdaten.

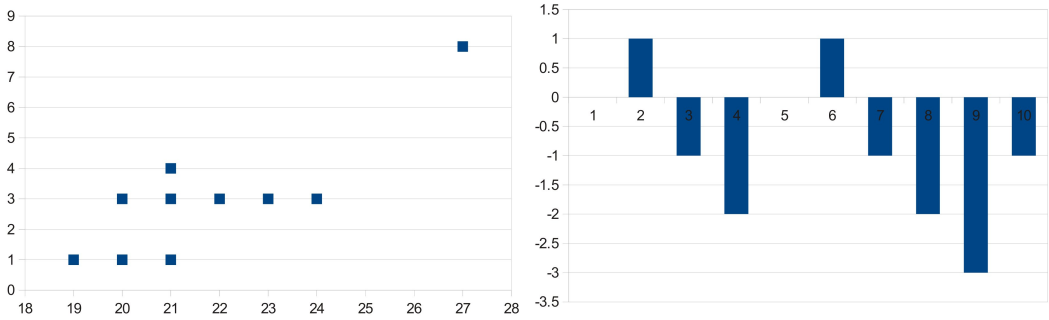
j) Berechnen Sie den Quartilsabstand (zwischen erstem und drittem Quartil).

16. Die folgende Tabelle zeigt das Alter einer Gruppe von Studierenden und die Zahl ihrer Studienjahre.

19	20	20	21	21	21	22	23	24	27
1	3	1	1	3	4	3	3	3	8

a) Bestimmen Sie den Median und die Spannweite der Studienjahre der Studierenden.

b) Die beiden folgenden Graphiken zeigen die Punktwolke zur Tabelle und die Residuen zur Gerade $y = x - 18$. Bei dieser Gerade wird angenommen, das 19-jährige im ersten Studienjahr sind und dann jedes Jahr ein Studienjahr dazukommt.



Die Residuen im rechten Bild zeigen, dass es einen systematischen Fehler gibt bei der Trendgeraden $y = x - 18$ gibt.

Wie muss die Trendgerade verändert werden? (Zeichnen Sie die Trendgerade ein, dann sehen Sie es.)

c) Legen Sie eine gute Trendgerade durch die Punktwolke.

d) Ist die Korrelation zwischen Alter und Semesterzahl eher nahe bei Null, eher negativ oder eher positiv?

9.7 Exkurs: Ausdauersport

In diesem Abschnitt geht es darum, statistische Kenntnisse zu verwenden, um Zusammenhänge zwischen Datenreihen zu bestätigen oder zu widerlegen. Dies wird beispielhaft am Beispiel vom Ausdauersport durchgeführt.

Zwei Fachmittelschulklassen haben 30 Minuten-Läufe durchgeführt. Die Daten liegen vor. Die SchülerInnen wurden gebeten, zu überlegen, wie Ausdauerleistung mit anderen Eigenschaften zusammenhängt. Hier sind, sinngemäss, einige Hypothesen.

Überlegen Sie, welche Hypothesen sich sinnvoll untersuchen lassen, und welche Daten noch erhoben werden müssen. Welche Hypothesen fallen Ihnen noch ein?

- Geschlecht und Ausdauerleistung korrelieren.
- Alter und Ausdauerleistung korrelieren.
- Ausdauerleistung und das sonstige Sporttreiben korrelieren.
- Gewicht und Ausdauerleistung korrelieren.
- Grösse und Ausdauerleistung korrelieren.
- Hochsprung/ Weitsprungleistung hängt nicht mit der Sprintleistung/Ausdauerleistung zusammen.
- Ausdauerleistung verändert den Charakter.
- wer Ausdauerleistung treibt ist sozialer.
- Sporttreiben und Depression korrelieren.
- Ausdauerleistung und Depression korrelieren.
- IQ und Ausdauerleistung korrelieren.
- Wie viel Schlaf man hatte ist korreliert mit der Ausdauerleistung.
- Schuhe und Kleidung korrelieren mit der Ausdauerleistung.
- Das Wetter korreliert mit der Ausdauerleistung.
- Die Anfahrt zur Schule (Velo, Bus) korreliert mit der Ausdauerleistung.
- der Ruhepuls hat nichts mit der Ausdauerleistung zu tun.
- der Puls während der Ausdauerleistung hängt mit der Leistung zusammen.
-

10 Lösungsverzeichnis