Aufgabe 1

1. absolutskaliert/stetig
2. absolutskaliert/diskret
3. nominal/diskret
4. ordinal/diskret
5. absolutskaliert/diskret
6. nominal/diskret

Aufgabe 2

Die Berechnung des arithmetischen Mittels ist für nominale Merkmale völlig unsinnig, da die Merkmalswerte nur Codes für Begrifflichkeiten darstellen, die keine inhärente numerische Bedeutung haben. Auch bei ordinalen Merkmalen ist die Berechnung des arithmetischen Mittels problematisch, da sie, um aussagefähig zu sein, eine Abstandstreue zwischen den einzelnen Merkmalsausprägungen voraussetzt, die hier nicht gegeben ist.

Aufgabe 3

Die absoluten Häufigkeiten errechnen sich durch Multiplikaton der relativen Häufigkeiten mit der Gesamtanzahl der Studierenden (=11) als:

Stunden (y_i)	absolute Häufigkeit (h_i)
1	2
2	1
3	3
4	5

Eine mögliche Urliste ist demnach:

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

Aufgabe 4

Bezeichnet n_1 die Anzahl der Befragten in der Altersklasse von unter 20 Jahren, n_2 die Anzahl der Befragten in der Altersklasse von 20 bis unter 50 Jahren, n_3 die Anzahl der Befragten in der Altersklasse von 50 Jahren oder darüber und h die gesuchte absolute Häufigkeit, so gilt:

$$\begin{aligned} n &= n_1 + n_2 + n_3 \\ n_1 &= 0.35n \\ n_2 &= 0.4n \\ n_3 &= 0.25n \\ \Rightarrow h &= 0.8n_1 + 0.7n_2 + 0.3n_3 \\ &= 0.8 \cdot 0.35n + 0.7 \cdot 0.4n + 0.3 \cdot 0.25n \\ &= 0.555n \end{aligned}$$

Die relative Häufigkeit der regelmäßigen Internetnutzer in allen Altersklassen zusammengefaßt beträgt demnach 55.5%.

Von diesem Ergebnis läßt sich auf die entsprechende Anzahl in der Gesamtbevölkerung nur dann zurückschließen, wenn die Stichprobe erstens ausreichend groß und zweitens repräsentativ war, wenn also insbesondere auch der Anteil aller Altersklassen an der Gesamtbevölkerung mit dem entsprechenden Anteil in der Stichprobe annähernd übereinstimmt.

Aufgabe 5

Das Merkmal hat insgesamt $k = 20$ Ausprägungen:

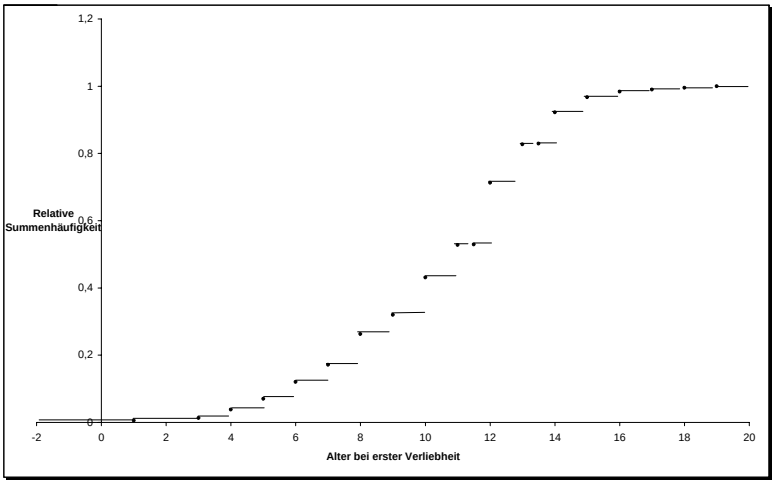
$$y_1 = 1,\ y_2 = 3,\ y_3 = 4,\ y_4 = 5, \ldots, y_{20} = 19$$

Häufigkeitstabelle

Index i	Ausprägung y_i	h_i	H_i	F_i
1	1.0	3	3	0.005
2	3.0	5	8	0.013
3	4.0	16	24	0.038
4	5.0	20	44	0.070
5	6.0	32	76	0.120
6	7.0	32	108	0.171
7	8.0	58	166	0.263
8	9.0	36	202	0.320
9	10.0	70	272	0.431
10	11.0	61	333	0.528
11	11.5	1	334	0.529
12	12.0	116	450	0.713
13	13.0	72	522	0.827
14	13.5	1	523	0.829
15	14.0	59	582	0.922
16	15.0	28	610	0.967
17	16.0	11	621	0.984
18	17.0	4	625	0.990
19	18.0	3	628	0.995
20	19.0	3	631	1

Hierbei bezeichne wie üblich h_i die absolute Einzelhäufigkeit, H_i die absolute Summenhäufigkeit und F_i die relative Summenhäufigkeit der Merkmalsausprägung y_i, $i = 1, 2, \ldots, 20$. Aus der Tabelle läßt sich unmittelbar ablesen, daß 76 bzw. 12% der Befragte jünger als 7 Jahre sowie 631-522 = 109 bzw. 100%-82.7%=17.3% der Befragten älter als 13 Jahre beim ersten Verliebtsein waren und daß im Alter von 12 Jahren 60% der Befragten schon mindestens einmal verliebt waren.

Die empirische Verteilungsfunktion stellt sich wie folgt dar:



Eine Einbeziehung der 17 Jugendlichen, die noch zum Zeitpunkt der Befragung noch nie verliebt waren, verändert die Ergebnisse wie folgt: Da alle befragten Jugendlichen zwischen 12 und 19 Jahren waren, steht fest, daß diejenigen, die bisher noch nicht verliebt waren, im Alter von 7 Jahren auf jeden Fall auch noch nicht verliebt gewesen sein können. Deswegen bleibt die absolute Häufigkeit konstant bei 76 und der entsprechende Prozentsatz verringert sich von 12 auf $\frac{76}{648} \cdot 100 = 11.7$ Prozent. Der Prozentsatz der Befragten, die älter als 13 Jahre beim ersten Verliebtsein waren, läßt sich nur über zwei Extremfälle abschätzen: Entweder alle 17 Jugendlichen sind jünger als 13 Jahre und verlieben sich noch vor dem 13. Lebensjahr oder alle verlieben sich erst später. Demnach liegt die gesuchte absolute Häufigkeit zwischen 109 und 109+17 = 126, so daß sich relative Häufigkeit f wie folgt abschätzen läßt:

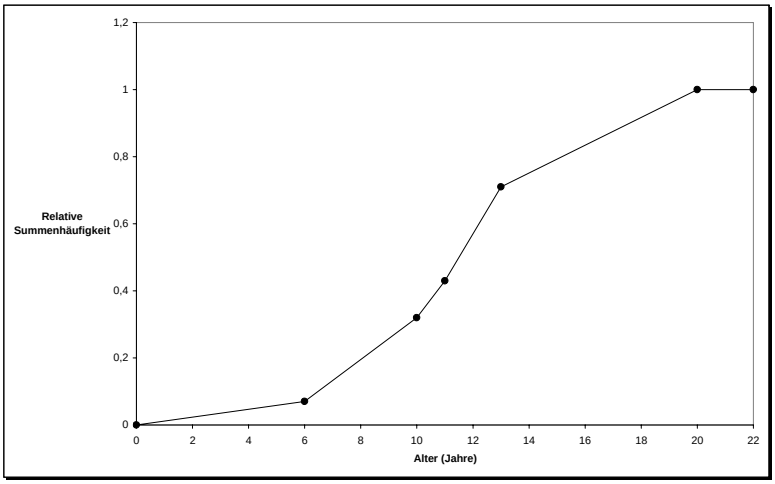
$$\begin{aligned} & \frac{109}{648} \leq \quad f \quad \leq \frac{109+17}{648} \\ \Leftrightarrow \quad & 0.168 \leq \quad f \quad \leq 0.194 \end{aligned}$$

Aufgabe 6

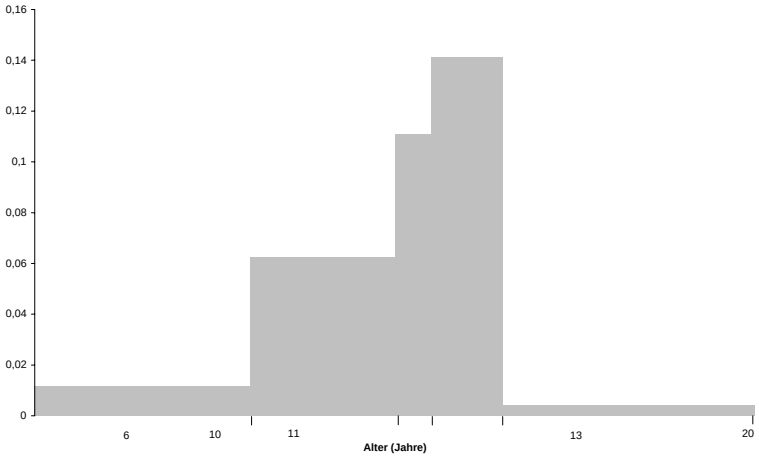
Die Häufigkeitstabelle für die gewählte Klasseneinteilung stellt sich wie folgt dar:

Index i	Klasse K_i (Alter in Jahren)	h_i	f_i	F_i	Klassenbreite Δ_i	Histogrammhöhe d_i
1	0 bis unter 6	44	0.07	0.07	6	0.0117
2	6 bis unter 10	158	0.25	0.32	4	0.0625
3	10 bis unter 11	70	0.11	0.43	1	0.1109
4	11 bis unter 13	178	0.28	0.71	2	0.1410
5	13 bis unter 20	181	0.29	1	7	0.0041

Empirische Verteilungsfunktion $F(x)$:



Histogramm:



Wegen

$$F(7) = \frac{44}{631} + \frac{\frac{202}{631} - \frac{44}{631}}{10-6}(7-6) = \frac{167}{1262} \approx 0.132$$

bzw.

$$F(13) = \frac{450}{631} \approx 0.713$$

sind (approximativ) 13.2% bzw. 100-71.3=28.7% der Befragten beim ersten Verliebtsein jünger als 7 bzw. älter als 13 Jahre. Da sich die Merkmalswerte in der Klasse, die den Merkmalswert

7 Jahre enthält, relativ gleichmäßig verteilen, fällt der Fehler in dem durch die Klassenbildung berechnetem approximativen Wert relativ gering aus. Der entsprechende Fehler fällt beim Wert 13 deutlich größer aus, da die 72 bzw. 11.4% der Befragten, die beim ersten Verliebtsein genau 13 Jahre alt waren, bei der genauen Berechnung nicht gezählt werden, bei der approximativen Berechnung über die Klassen hingegen miteinberechnet werden.

Wegen

$$\begin{aligned} 0.6 &= \frac{272}{631} + \frac{\frac{450}{631} - \frac{272}{631}}{13-11}(x-11) \\ \Rightarrow \quad x &= \frac{5428}{445} \approx 12.2 \end{aligned}$$

waren (approximativ) 60% der Befragten spätestens mit 12.2 Jahren das erste Mal verliebt. Da in der Altersklasse zwischen 11 und unter 13 Jahren das Gros der Merkmalswerte bei 12 Jahren liegt, unterschätzt die approximative Berechnung, die von einer gleichmäßigen Verteilung der Merkmalswerte in den Klassen ausgeht, die zugehörige relative Häufigkeit (laut Einzeldaten sind mit 12 Jahren schon 71.3% der Befragten verliebt gewesen).

Aufgabe 7

Gipfel: Die Verteilung hat einen Gipfel, ist also unimodal. Der Gipfel befindet sich bei Angstwerten von 55-60, also ziemlich in der Mitte der möglichen Werte von 0 bis 100.

Symmetrisch/schiefe Verteilung: Wenn man einmal von den kleinen "Störungen" am Anfang der Skala absieht, dann ist die Verteilung der Werte relativ regelmäßig und symmetrisch um den Gipfel verteilt.

Normalverteilung: Aufgrund der Symmetrie der Verteilung und da Verteilung der Daten weder besonders flach noch besonders steil erscheint, scheint eine Normalverteilung plausibel zu sein. Dies legt auch die Natur der Daten als subjektive Bewertung eines Gefühlszustandes nahe, die häufig normal verteilt sind.

Zentrum: Das Zentrum der Daten befindet sich um den Gipfel herum, also bei Angstwerten von 50 bis 65.

Streuung: Als groben Streubereich läßt sich die Spanne zwischen 0 bis 95 angeben, in der alle Angstwerte zu finden sind. Der Großteil der Daten (über 80%) liegen im Bereich von Werten zwischen 20 und 85.

Ausreißerwerte: Es sind keine Ausreißerwerte zu erkennen.

Einbrüche/Spitzen: Es liegt ein Einbruch bei Angstwerten von 10 bis 20 vor. Es steht aber vermuten, daß dieser Einbruch nicht auf eine besondere Eigentümlichkeit in der Verteilung von Angstwerten zurückzuführen ist, sondern es sich um eine "normale" Schwankung innerhalb der erhobenen Daten handelt.

Aufgabe 8

Das Merkmal X sei in der Studie S_i an n_i Teilnehmern mit Ausprägungen $x_{i1}, x_{i2}, \ldots, x_{in_i}$ erfaßt worden. Dann gilt mit $n := \sum_{i=1}^{k} n_i$:

$$\bar{x} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \cdot \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{x_i}}{n} = \sum_{i=1}^{k} \frac{n_i}{n}\bar{x_i}$$

Deswegen wird zur Berechnung des arithmetischen Mittels der Meta-Analyse neben dem arithmetischen Mittel der einzelnen Studien entweder noch die Kenntnis der Anzahl der Teilnehmer n_i in jeder der Studien benötigt oder aber mindestens der Prozentsatz der Teilnehmer in jeder der einzelnen Studien im Verhältnis zur Gesamtzahl der Teilnehmer (nämlich $\frac{n_i}{n}$).

Aufgabe 9

Bezeichne mit $x_{i(j)}$, $i \in \{1,2\}$, $j = 1,2,\ldots,n_i$, den j-ten Wert in der geordneten Liste der Merkmalswerte der Studie S_i sowie mit $x_{(j)}$, $j = 1,2,\ldots,n_1+n_2$, den j-tenWert in der geordneten Liste der Merkmalswerte beider Studien zusammengenommen. Also:

$$x_{1(1)} \leq x_{1(2)} \leq \ldots \leq x_{1(n_1)} \text{ und } x_{2(1)} \leq x_{2(2)} \leq \ldots \leq x_{2(n_2)}$$
$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n_1+n_2)} \text{wobei } x_{(l)} = x_{i(j)} \text{ mit } i,j \text{ geeignet gewählt}$$

Sei zunächst n_1 gerade, n_2 ungerade. Nach Definition des Medians folgt dann:

$$\begin{aligned} & x_{1(1)}, x_{1(2)}, ..., x_{1\left(\frac{n_1}{2}\right)} \leq \tilde{x_1} \text{ und } x_{2(1)}, x_{2(2)}, ..., x_{2\left(\frac{n_2+1}{2}\right)} \leq \tilde{x_2} \\ \Rightarrow \quad & x_{1(1)}, x_{1(2)}, ..., x_{1\left(\frac{n_1}{2}\right)}, x_{2(1)}, x_{2(2)}, ..., x_{2\left(\frac{n_2+1}{2}\right)} \leq \max\{\tilde{x_1}, \tilde{x_2}\} \\ \Rightarrow \quad & \tilde{x} = x_{\left(\frac{n_1+n_2+1}{2}\right)} \leq \max\{\tilde{x_1}, \tilde{x_2}\} \end{aligned}$$

Die letzte Ungleichung folgt einfach daraus, daß mindestens $\frac{n_1}{2} + \frac{n_2+1}{2} = \frac{n_1+n_2+1}{2}$ Werte kleiner als $\max\{\tilde{x_1}, \tilde{x_2}\}$ sind, also mit Sicherheit also auch der $\frac{n_1+n_2+1}{2}$-kleinste Wert von allen Werten zusammengenommen, also $x_{\left(\frac{n_1+n_2+1}{2}\right)}$.

Der Fall n_1 ungerade, n_2 gerade folgt genauso durch Vertauschung der Rollen von n_1 und n_2.

Seien jetzt n_1 und n_2 gerade. Analog zu oben erhalten wir:

$$x_{\left(\frac{n_1+n_2}{2}\right)} \leq \max\{x_{1\left(\frac{n_1}{2}\right)}, x_{2\left(\frac{n_2}{2}\right)}\}$$

O.B.d.A. gelte $\max\{x_{1\left(\frac{n_1}{2}\right)}, x_{2\left(\frac{n_2}{2}\right)}\} = x_{1\left(\frac{n_1}{2}\right)}$. Wegen

$$x_{\left(\frac{n_1+n_2}{2}\right)} \leq x_{1\left(\frac{n_1}{2}\right)} \leq x_{1\left(\frac{n_1}{2}+1\right)}$$

folgt

$$x_{\left(\frac{n_1+n_2}{2}+1\right)} \leq x_{1\left(\frac{n_1}{2}+1\right)}$$

und damit

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n_1+n_2}{2}\right)} + x_{\left(\frac{n_1+n_2+}{2}+1\right)}\right) \leq \frac{1}{2}\left(x_{1\left(\frac{n_1}{2}\right)} + x_{1\left(\frac{n_1}{2}+1\right)}\right) \leq \tilde{x_1} = \max\{\tilde{x_1}, \tilde{x_2}\}$$

Seien im letzten Fall n_1 und n_2 ungerade. Wieder folgt daraus analog zum ersten Fall:

$$x_{\left(\frac{n_1+n_2}{2}+1\right)} \leq \max\{\tilde{x_1}, \tilde{x_2}\}$$

Wegen $x_{\left(\frac{n_1+n_2}{2}\right)} \leq x_{\left(\frac{n_1+n_2}{2}+1\right)}$ folgt

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n_1+n_2}{2}\right)} + x_{\left(\frac{n_1+n_2}{2}\right)+1}\right) \leq x_{\left(\frac{n_1+n_2}{2}+1\right)} \leq \max\{\tilde{x_1}, \tilde{x_2}\}$$

Die Behauptung für das Minimum ergibt sich, indem man das "gespiegelte" Merkmal $-X$ mit den Werten $-x_{i1}, -x_{i2}, \ldots, -x_{in_i}$, $i \in \{1,2\}$, betrachtet. Da für beliebige Merkmalswerte $x_1, x_2, \ldots, x_n$ offensichtlich gilt:

$$\text{Median}(-x_1, -x_2, \ldots, -x_n) = -\text{Median}(x_1, x_2, \ldots, x_n)$$

folgt mit den Ergebnissen von oben

$$\begin{aligned} & -\tilde{x} \leq \max\{-\tilde{x_1}, -\tilde{x_2}\} = -\min\{\tilde{x_1}, \tilde{x_2}\} \\ \Rightarrow \quad & \tilde{x} \geq \min\{\tilde{x_1}, \tilde{x_2}\} \end{aligned}$$

Aufgabe 10

1. $0, 0, 0, 1, 1, 2, 10$ mit $x_{mod} = 0$, $\tilde{x} = 1$, $\bar{x} = 2$.
2. $0, 1, 9, 9, 10, 10, 10$ mit $x_{mod} = 10$, $\tilde{x} = 9$ und $\bar{x} = 7$.
3. $0, 0, 1, 1, 1, 2, 2$ mit $x_{mod} = 1$, $\tilde{x} = 1$, $\bar{x} = 1$.

Die unterschiedliche Lagen der drei Parameter zueinander sind eine Folge davon, daß die erste Verteilung positiv schief, die zweite Verteilung negativ schief und die dritte Verteilung symmetrisch ist.

Aufgabe 11

Die Lageparameter der unklassierten Daten errechnen sich als:

Modus: $x_{mod} = 12$

Mittelwert: $\bar{x} = \sum_{i=1}^{20} f_i y_i = \frac{1}{631}(1 \cdot 3 + 3 \cdot 5 + 4 \cdot 16 + \ldots + 19 \cdot 3) = \frac{6711}{631} \approx 10.64$

Median: $\tilde{x} = 11$

Die Lageparameter der klassierten Daten ergeben sich dagegen wie folgt:

Modus x_{mod}:

Die Altersklasse 11-13 Jahre verfügt über den höchsten Balken im Histogramm.

$$\begin{aligned} {\Delta x_u}^i &= \frac{178}{2 \cdot 631} - \frac{70}{631} = \frac{19}{631} \\ {\Delta x_o}^i &= \frac{178}{2 \cdot 631} - \frac{181}{7 \cdot 631} = \frac{442}{4417} \\ \Rightarrow \quad x_{mod} &= 11 + \frac{\frac{19}{631}}{\frac{19}{631} + \frac{442}{4417}} \cdot (13 - 11) = \frac{6591}{575} \approx 11.46 \end{aligned}$$

Mittelwert $\bar{x}$:

$$\begin{aligned} \bar{x} &= \frac{1}{631}(44 \cdot 3 + 158 \cdot 8 + 70 \cdot 10.5 + 178 \cdot 12 + 181 \cdot 16.5) \\ &= \frac{7253.5}{631} \approx 11.49 \end{aligned}$$

Median $\tilde{x}$:

$$\begin{aligned} 0.5 &= \frac{272}{631} + \frac{\frac{450}{631} - \frac{272}{631}}{13 - 11}(\tilde{x} - 11) \\ \Leftrightarrow \quad \tilde{x} &= 11 + \frac{87}{178} \approx 11.49 \end{aligned}$$

Aufgrund der ziemlich symmetrischen Verteilung der Daten ohne Ausreißer liegen die drei Lageparameter in allen Fällen dicht beeinander. Dies führt auch dazu, daß das arithmetische Mittel der klassierten und der unklassierten Daten nicht stark voneinander abweicht. Für Median und Modus gilt dies ebenso, und zwar insbesondere auch weil die Klasse, in der diese beide Parameter fallen, nur eine geringe Breite hat, so daß der Fehler, der durch den Mittelungsprozess entsteht, nicht sehr hoch ausfallen kann.

Aufgabe 12

Die Wachstumsrate der Anzahl der Informatik-Studenten beträgt in den letzten drei Jahre zusammengenommen $1.21 \cdot 1.26 \cdot 1.09 = 1.661814$. Somit ist die Anzahl der Studenten in den letzten drei Jahren um insgesamt ca. 66.2% gestiegen. Der durchschnittliche Anstieg pro Jahr betrug demnach $\sqrt[3]{1.701381} \approx 1.184$ bzw. 18.4 Prozent.Ein Abfall auf das Ausgangsniveau nach Ablauf der nächsten zwei Jahre bedeutet eine durchschnittliche Reduzierung x pro Jahr mit der Eigenschaft:

$$\begin{aligned} 1 &= 1.21 \cdot 1.26 \cdot 1.09 \cdot x^2 \\ \Rightarrow \quad x &= \sqrt{\frac{1}{1.21 \cdot 1.26 \cdot 1.09}} \approx 0.776 \end{aligned}$$

Demnach darf sich die Anzahl der Studenten in den nächsten zwei Jahren im Schnitt maximal um jeweils 22.4 Prozent verringern, um nicht unter das Ausgangsniveau von vor drei Jahren zu fallen.

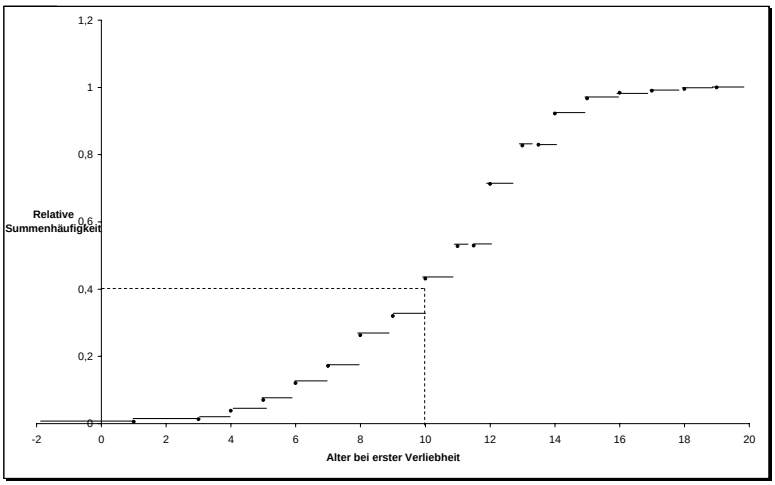
Aufgabe 13

Seien $x_1, x_2 \in \mathbb{R}_+$, dann gilt:

$$\begin{aligned} 0 &\leq \left(\frac{1}{2}(x_1 + x_2)\right)^2 = \frac{1}{4}x_1^2 - \frac{1}{2}x_1x_2 + \frac{1}{4}x_2^2 \\ \Leftrightarrow \quad x_1x_2 &\leq \frac{1}{4}x_1^2 + \frac{1}{2}x_1x_2 + \frac{1}{4}x_2^2 \\ \Leftrightarrow \quad \left(\sqrt{x_1x_2}\right)^2 &\leq \left(\frac{1}{2}(x_1 + x_2)\right)^2 \\ \Leftrightarrow \quad \sqrt{x_1x_2} &\leq \frac{1}{2}(x_1 + x_2) \text{ wegen der Monotonie der Wurzelfunktion} \\ \Leftrightarrow \quad x_g &\leq \bar{x} \text{ für } n = 2 \end{aligned}$$

Aufgabe 14

Zeichnerische Bestimmung des 0.4-Quantils:



Rechnerische Bestimmung des $\frac{108}{631}$-Quantils:

Da $n\alpha = 631 \cdot \frac{108}{631}$ eine ganze Zahl ist, gilt:

$$x_{\frac{108}{631}} = \frac{1}{2}\left(x_{(108)} + x_{(109)}\right) = \frac{1}{2}(7 + 8) = 7.5$$

Das 0.4-Quantil beträgt demnach 10 Jahre und das $\frac{108}{631}$-Quantil 7.5 Jahre. Dies bedeutet, daß ca. 17.1% der Befragten zum Zeitpunkt der ersten Verliebtheit nicht älter als siebeneinhalb Jahre bzw. 40% der Befragten nicht älter als 10 Jahre waren.

Aufgabe 15

$Y := X - \bar{X}$ und $\bar{X}$ stehen senkrecht aufeinander, da:

$$\langle X - \bar{X}, \bar{X} \rangle = \sum_{i=1}^{n} (x_i - \bar{x})\bar{x} = \bar{x} \left(\sum_{i=1}^{n} x_i - n\bar{x} \right) = \bar{x} \cdot 0 = 0$$

Im zweidimensionalen Raum stellt sich die Situation graphisch wie folgt dar:

Das bedeutet, daß nach dem Satz des Pythargoras für die Längen a, b, und c der Vektoren (und für beliebiges n) $\bar{X}$, $X - \bar{X}$ und X gilt:

$$\begin{aligned} & a^2 + b^2 &=& c^2 \\ \Leftrightarrow \quad & n\bar{x}^2 + \sum_{i=1}^{n} (x_i - \bar{x})^2 &=& \sum_{i=1}^{n} x_i^2 \\ \Leftrightarrow \quad & n\bar{x}^2 + n s^2(X) &=& \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Das heißt: Der Merkmalsvektor $(x_1, x_2, \ldots, x_n)$ läßt sich aufspalten in einen Vektor in Richtung der Diagonalen, der der "gleichmäßigsten Messung" $(\bar{x}, \ldots, \bar{x})$ entspricht, und einen Vektor, der die Länge der Entfernung von der Diagonalen mißt, die bis auf den konstanten Faktor n durch die Varianz $s^2(X)$ der Merkmalswerte bestimmt wird.

Aufgabe 16

Bezeichne $\bar{x}$, $\overline{ax}$ bzw. $\overline{x+b}$ das arithmetische Mittel der Merkmalswerte von X, aX bzw. $X + b$ so gilt:

$$\overline{ax} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} ax_i = a \cdot \tfrac{1}{n} \sum_{i=1}^{n} x_i = a\bar{x}$$

sowie

$$\overline{x+b} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (x_i + b) = \tfrac{1}{n} \sum_{i=1}^{n} x_i + \tfrac{1}{n} \cdot nb = \bar{x} + b$$

Damit ergeben sich folgende Formeln für die Varianzen:

$$\begin{aligned} s^2(aX) &= \frac{1}{n} \sum_{i=1}^{n} (ax_i - \overline{ax})^2 = \frac{1}{n} \sum_{i=1}^{n} (ax_i - a\bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} a^2 (x_i - \bar{x})^2 = a^2 \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= a^2 s^2(X) \end{aligned}$$

und

$$\begin{aligned} s^2(X+b) &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i + b - \overline{x+b} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i + b - (\bar{x} + b))^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i + b - \bar{x} - b)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= s^2(X) \end{aligned}$$

Wie auch zu erwarten war, verändert damit eine Translation, d.h. eine reine Verschiebung der Werte, die Varianz nicht, wohingegen die Multiplikation der Merkmalswerte mit einem skalaren

Faktor die Varianz (quadratisch) staucht oder streckt, je nachdem ob der Faktor betragsmäßig größer oder kleiner als eins ist.

Da eine Translation die relative Lage der Merkmalswerte zueinander nicht beeinflußt, bleiben dabei auch Spannweite und Quartilsabstand unverändert. Für die geordnete Liste der Merkmalswerte ax_i, $i = 1, \ldots, n$, die im folgenden mit $(ax)_{(1)} \leq (ax)_{(2)} \leq \ldots \leq (ax)_{(n)}$ bezeichnet sei, gilt offensichtlich:

$$(ax)_{(i)} = ax_{(i)}, \text{ für } a \geq 0 \text{ und } i = 1, 2, \ldots, n$$

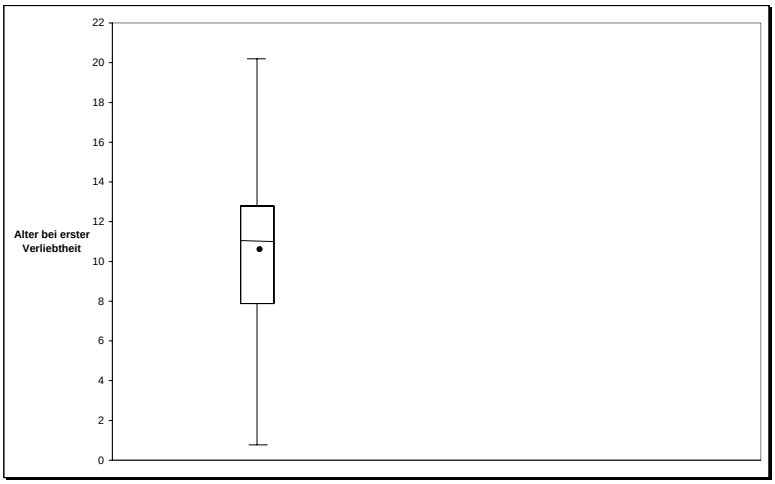
sowie

$$(ax)_{(i)} = ax_{(n-i+1)}, \text{ für } a < 0 \text{ und } i = 1, 2, \ldots, n$$

Der größte und kleinste Wert bzw. das 0.25- bzw. 0.75-Quantil werden damit um den Faktor a gestaucht bzw. gestreckt und falls a kleiner Null ist, werden zusätzlich größter und kleinster Wert bzw. 0.25- bzw. 0.75-Quantil vertauscht. Deswegen verändern sich sowohl Spannweite als auch Quartilsabstand um den Faktor $|a|$.

Aufgabe 17

Median und Mittelwert waren Aufgabe 12 bestimmt worden als $\bar{x} \approx 10.64$ und $\tilde{x} = 11.$. Wegen $x_{([0.25*631]+1)} = x_{(158)} = 8$ und $x_{([0.75*631]+1)} = x_{(474)} = 13$ beträgt der Quartilsabstand q $13 - 8 = 5$ Jahre. Ausreißer wären demnach Merkmalswerte größer als $13 + 7.5 = 20.5$ bzw. kleiner als $8 - 7.5 = 0.5$ Jahre und sind somit nicht vorhanden. Damit hat der zugehörige Boxplot folgende Gestalt:



Aufgabe 18

Da bei im Schnitt höherer Körpergröße auch die Streuung naturgemäß größer ausfallen wird als bei im Schnitt niedrigerer Körpergröße, ist es hier nicht angemessen, die Standardabweichung direkt als Streuungsmaß für den Vergleich beider Altersgruppen zu betrachten. Besser ist der Variationskoeffizient v, der die Standardabweichung ins Verhältnis zum Mittelwert setzt. Er errechnet sich hier als:

Altersklasse	arithmetisches Mittel	Standardabweichung	Variationskoeffizient
4 Jahre	101.3 cm	4.35 cm	4.29%
18 Jahre	176.2 cm	6.26 cm	3.55%

Wir sehen also, daß im Gegensatz zur absoluten Streuung der Standardabweichung, die bei den 18-jährigen größer ist als bei den 4-jährigen, die relative Streuung, gemessen durch den Variationskoeffizient, mit zunehmenden Alter abnimmt. Dies war auch zu erwarten, da manche Kinder schneller als andere wachsen, was - zusätzlich zur Veranlagung zur Größe - eine weitere Variabilität in diese Variable bringt.

Aufgabe 19

Die Häufigkeitstabelle läßt sich unter Benutzung der Randhäufigkeiten wie folgt vervollständigen:

	Alter (Jahre)					
Testergebnis	0-	10-	20-	30-	40-	Total
positiv	14 (4.9%)	16 (5.6%)	14 (4.9%)	7 (2.4%)	6 (2.1%)	57 (19.8%)
negativ	87 (30.2%)	33 (11.5%)	66 (22.9%)	34 (11.8%)	11 (3.8%)	231 (80.2%)
Total	101 (35.1%)	49 (17.0%)	80 (27.8%)	41 (14.2%)	17 (5.9%)	288(100%)

$f_{3.} \approx 0.278$, d.h. ca. 27.8 Prozent der untersuchten Personen waren zwischen 20 und 30 Jahren alt.

$h_{.2} = 231$, d.h. bei insgesamt 231 der untersuchten Personen konnten keine Eier von Schistosoma mansoni im Stuhl nachgewiesen werden.

Aufgabe 20

Es gilt mit den üblichen Bezeichnungen:

Bedingte Verteilung der Altersgruppen bei positivem Testergebnis:

$$f_{1|y_1} = \frac{14}{57}, f_{2|y_1} = \frac{16}{57}, f_{3|y_1} = \frac{14}{57}, f_{4|y_1} = \frac{7}{57}, f_{5|y_1} = \frac{6}{57}$$

Bedingte Verteilung der Altersgruppen bei negativem Testergebnis

$$f_{1|y_2} = \frac{87}{231}, f_{2|y_2} = \frac{33}{231}, f_{3|y_2} = \frac{66}{231}, f_{4|y_2} = \frac{34}{231}, f_{5|y_2} = \frac{11}{231}$$

Stellt man die beiden bedingten Verteilungen (in Prozent) in einer Tabelle gegenüber, so ergibt sich folgendes Bild:

	Relative Häufigkeiten der bedingten Verteilungen					
Testergebnis	0-	10-	20-	30-	40-	Total
positiv	24.6%	28.1%	24.6%	12.3%	10.5%	100%
negativ	37.7%	14.3%	28.6%	14.7%	4.8%	100%

Demnach sind zwar Unterschiede zwischen den beiden Altersverteilungen zu sehen, diese sind aber nicht besonders stark ausgeprägt: In den Altersgruppen 20 bis 30 und 30 bis 40 Jahre sind die Unterschiede marginal, in der Altersgruppe über 40 verfügen prozentual zwar mehr als doppelt so viele Probanden über ein positives wie über ein negatives Testergebnis, aber diese Altersgruppe ist dennoch in beiden bedingten Verteilungen die Gruppe mit der geringsten relative Häufigkeit. In die Altersgruppen von 0 bis 10 und 10 bis 20 Jahren fallen insgesamt in beiden Verteilungen ca. 50% der Probanden, allerdings ist der Anteil beider Gruppen nur bei positivem Testergebnis ungefähr gleich groß, wohingegen bei negativem Testergebnis die Gruppe der jüngeren Probanden (0 bis 10 Jahre) stärker gewichtet ist.

Aufgabe 21

Tabelle der bei statistischer Unabhängigkeit erwarteten relativen Häufigkeiten $\tilde{f_{ij}} = f_{i.}f_{.j}$

	Alter				
Testergebnis	0-	10-	20-	30-	40-
positiv	$\frac{5757}{82944}(6.9\%)$	$\frac{2793}{82944}(3.4\%)$	$\frac{4560}{82944}(5.5\%)$	$\frac{2337}{82944}(2.8\%)$	$\frac{969}{82944}(1.2\%)$
negativ	$\frac{23331}{82944}(28.1\%)$	$\frac{11319}{82944}(13.6\%)$	$\frac{18480}{82944}(22.3\%)$	$\frac{9471}{82944}(11.4\%)$	$\frac{3927}{82944}(4.7\%)$

Die Unterschiede zwischen den tatsächlichen und den bei statistischer Unabhängigkeit erwarteten relativen Häufigkeiten sind mit einem maximalen Unterschied von unter 3 Prozentpunkten ziemlich klein, was darauf hindeutet, daß keine besonders starke Abhängigkeit zwischen den zwei Merkmalen besteht.