Gegeben ist ein Datensatz:

Student Nummer	Rastas	Haarlänge	Haarfarbe	Ausgaben für Haare	Anzahl Frisör-Besuche
1	0	3	1	31	2
2	0	3	3	63	2
3	1	1	2	32	1
4	0	2	4	45	3
5	0	1	0	11	0
6	1	1	2	34	2
7	0	3	1	23	1
8	1	2	0	30	3
9	1	3	2	73	4
10	0	2	1	8	0
11	0	1	2	23	2
12	0	2	3	52	4
13	1	3	1	70	4
14	1	2	1	58	3
15	0	2	3	85	4
16	0	3	2	14	0
17	0	1	1	35	3
18	1	3	0	24	1
19	1	3	1	61	3
20	0	3	0	45	3

Bezeichnungen:

N = Umfang der statistischen Gesamtheit. Hier ist N die Gesamtzahl aller Studenten, die man befragen könnte.
n = Stichprobenumfang. Hier $n = 20$, die Anzahl der befragten Studenten.

$a_1, a_2, a_3, \ldots, a_i, \ldots, a_n$ ist eine **Urliste** zu einem **Merkmal X**, wobei $\boxed{i = 1, \ldots, n}$ gilt.

Oben sind die 5 Urlisten zu den 5 Merkmalen als Spalten geschrieben.

$x_1, x_2, x_3, \ldots, x_j, \ldots, x_m$ sind die m **verschiedenen Ausprägungen** eines **Merkmals X**, wobei $\boxed{j = 1, \ldots, m}$ mit $m \le n$ gilt. Bei Rastas ist $m = 2$, bei der Haarlänge ist m = 3, bei der Haarfarbe ist m = 5, usw.

Die **absolute Häufigkeit h_j einer Ausprägung x_j eines Merkmals X** gibt an, wie oft die Ausprägung x_j in der Urliste des Merkmals X auftritt.
Beispiel: Bei der Haarlänge tritt die Ausprägung $x_1 = 1$ genau $h_1 = 5$ mal, die Ausprägung $x_2 = 2$ genau $h_2 = 6$ mal und die Ausprägung $x_3 = 3$ genau $h_3 = 9$ mal auf. Die Probe stimmt: $h_1 + h_2 + h_3 = 20$. Allgemein ist für jede Ausprägung die Summe der absoluten Häufigkeiten gleich dem Stichprobenumfang:

$$\boxed{\sum_{j=1}^{m} h_j = n}$$

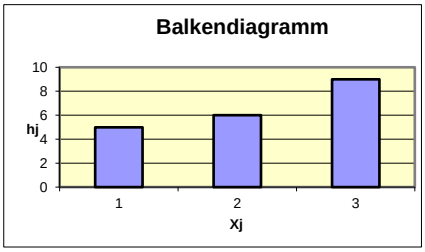
Die **relative Häufigkeit f_j einer Ausprägung x_j eines Merkmals X** gibt an, zu welchem Bruchteil die Ausprägung x_j in der Urliste des Merkmals X auftritt: $\boxed{f_j = \frac{h_j}{n}, \quad j = 1, \ldots, m}$

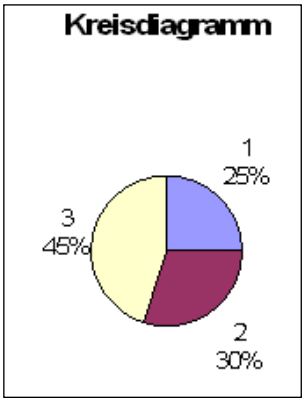
Beispiel: Bei der Haarlänge gilt für die Ausprägung $x_1 = 1$: $f_1 = \frac{5}{20} = 25\%$, für die Ausprägung $x_2 = 2$: $f_2 = \frac{6}{20} = 30\%$ und für die Ausprägung $x_3 = 3$: $f_3 = \frac{9}{20} = 45\%$. Die Probe stimmt: $f_1 + f_2 + f_3 = 1 = 100\%$.

Allgemein gilt $\boxed{\sum_{j=1}^{m} f_j = 1}$, d.h. für jede Ausprägung ist die Summe der relativen Häufigkeiten gleich 1.

Verschiedene Darstellungen der Haarlänge

Tabelle					
j	x_j	h_j	$f_j = \frac{h_j}{n}$	f_j in %	Winkel $= f_j \cdot 360°$
1	1	5	0,25	25	90°
2	2	6	0,30	30	108°
m = 3	3	9	0,45	45	162°
	$\sum_{j=1}^{3}$	n = 20	1	100	360°





Unter den **kumulierten absoluten Häufigkeiten H_j** einer Ausprägung x_j eines Merkmals X versteht man die

Summen $H_1 = \sum_{k=1}^{1} h_k = h_1$, $H_2 = \sum_{k=1}^{2} h_k = h_1 + h_2$, ... , $H_m = \sum_{k=1}^{m} h_k = n$.

Beispiel: Bei den Haarlängen

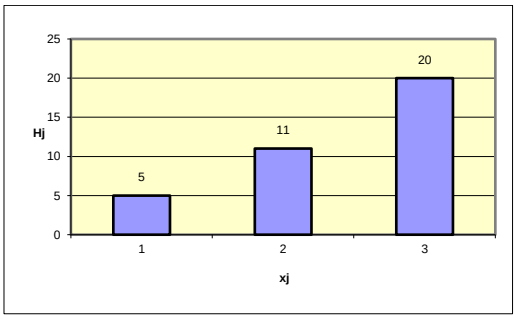
Haarlänge = 1: $H_1 = h_1 = 5$, Haarlänge ≤ 2: $H_2 = h_1 + h_2 = 11$, Haarlänge ≤ 3: $H_3 = h_1 + h_2 + h_3 = 20$.

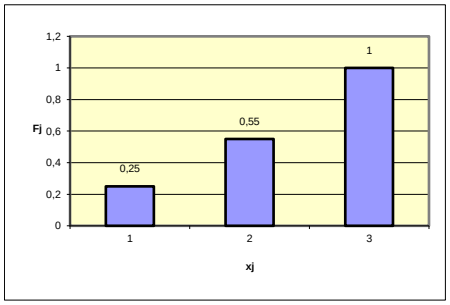
Unter den **kumulierten relativen Häufigkeiten F_j** einer Ausprägung x_j eines Merkmals X versteht man die

Summen $F_1 = \sum_{k=1}^{1} f_k$, $F_2 = \sum_{k=1}^{2} f_k$, ... , $F_m = \sum_{k=1}^{m} f_k = 1$.

Beispiel: Bei den Haarlängen

Haarlänge = 1: $F_1 = f_1 = \frac{5}{20}$, Haarlänge ≤ 2: $F_2 = f_1 + f_2 = \frac{11}{20}$, Haarlänge ≤ 3: $F_3 = f_1 + f_2 + f_3 = \frac{20}{20} = 1$.

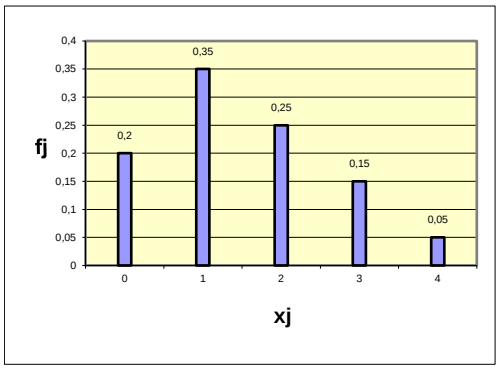




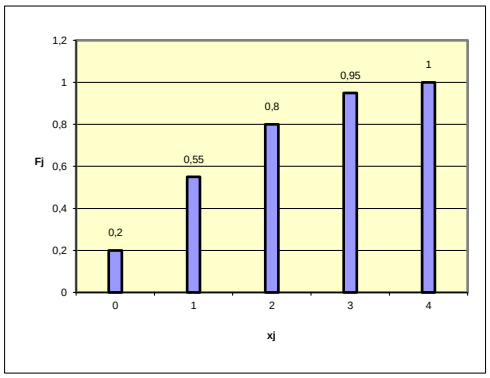
Und nun zur Haarfarbe:

j	x_j	h_j	f_j	H_j	F_j
1	0	4	0,20	4	0,20
2	1	7	0,35	11	0,55
3	2	5	0,25	16	0,80
4	3	3	0,15	19	0,95
m = 5	4	1	0,05	20	1,00
	$\sum_{j=1}^{5}$	20	1	–	–

Relative Häufigkeitsverteilung Haarfarbe

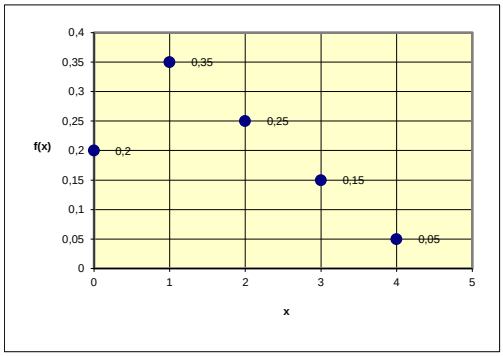


Kumulierte relative Häufigkeitsverteilung Haarfarbe

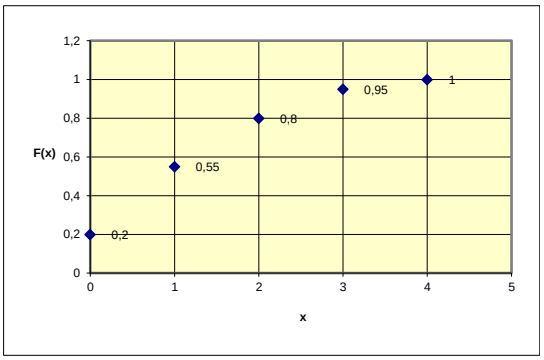


Die Häufigkeitsfunktionen am Beispiel der Haarfarbe

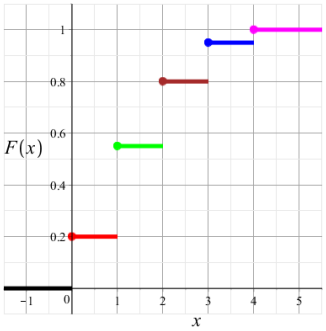
Relative Häufigkeitsfunktion $f(x)$



Relative summierte Häufigkeitsfunktion $F(x)$



Empirische Verteilungsfunktion $F(x)$



j	1	2	3	4	5
x_j	0	1	2	3	4
f_j	0,20	0,35	0,25	0,15	0,05
F_j	0,20	0,55	0,80	0,95	1,00

$$F(x) = \sum_{x_j \le x} f_j$$

Z.B. $F(1,3) = \sum_{x_j \le 1,3} f_j = f_1 + f_2 = 0,55$, $F(0) = \sum_{x_j \le 0} f_j = f_1 = 0,20$

Für $x < x_1$ ist $F(x) = 0$ und für $x \ge x_m$ ist $F(x) = 1$

Der Begriff der Konzentration

Bei einer Gruppe von $n = 20$ Studenten wird untersucht, wie viele Statistik-Bücher sie jeweils besitzen. Dabei bedeute $x_1 = 0$ kein, $x_2 = 1$ ein, ... , $x_5 = 4$ vier Statistik-Bücher, d.h. $m = 5$ Ausprägungen.

j	1	2	3	4	5
x_j	0	1	2	3	4
h_j	4	7	5	3	1
$h_j \cdot x_j$	0	7	10	9	4

Die Summe der Studenten ist $n = \sum_{j=1}^{m} h_j = \sum_{j=1}^{5} h_j = 20$ Studenten.

Die Summe aller vorhandenen Statistik-Bücher ist $\sum_{k=1}^{m} h_k \cdot x_k = \sum_{k=1}^{5} h_k \cdot x_k = 4 \cdot 0 + 7 \cdot 1 + 5 \cdot 2 + 3 \cdot 3 + 1 \cdot 4 = 30$.

Unter der **relativen Merkmalssumme von Merkmal j** versteht man $\boxed{\ell_j = \frac{h_j \cdot x_j}{\sum_{k=1}^{m} h_k \cdot x_k}}$, $j = 1, 2, ..., m$.

Z.B. gibt ℓ_3 den Anteil der 30 Bücher an, den die 5 Studenten mit je 2 Büchern besitzen:

$\ell_3 = \frac{5 \cdot 2}{30} = \frac{1}{3} \approx 33,3\%$. D,h, etwa $33,3\%$ aller Bücher gehören Studenten, die jeweils zwei Bücher besitzen.

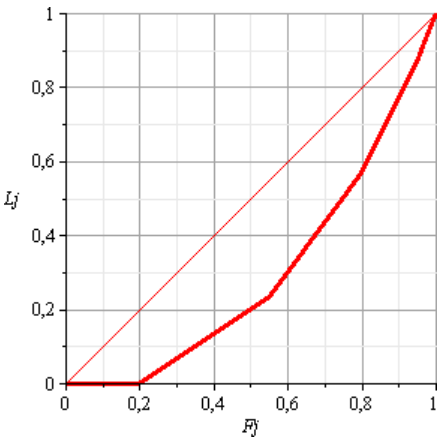
Unter der **kumulierten relativen Merkmalssumme von Merkmal j** versteht man $\boxed{L_j = \sum_{k=1}^{j} \ell_k = \frac{\sum_{k=1}^{j} h_k \cdot x_k}{\sum_{k=1}^{m} h_k \cdot x_k}}$.

Z.B. gibt die kumulierte relative Merkmalssumme $L_3 = \sum_{k=1}^{3} \ell_k$ den Anteil der 30 Bücher an, den die 16 Studenten mit höchstens je 2 Büchern besitzen: $L_3 = \sum_{k=1}^{3} \ell_k = \frac{\sum_{k=1}^{3} h_k \cdot x_k}{\sum_{k=1}^{5} h_k \cdot x_k} = \frac{17}{30} \approx 56,7\%$. D,h, etwa $56,7\%$ aller Bücher gehören Studenten, die höchstens zwei Bücher besitzen.

Zur Erinnerung: $F_j = \sum_{k=1}^{j} f_k = \sum_{k=1}^{j} \frac{h_k}{n}$, $j = 1, 2, ..., m$.

j	Merkmale x_j	absolute Häufigkeit h_j	relative Häufigkeit f_j	kumulierte relative Häufigk. F_j	relative Merkmalssumme ℓ_j	kumulierte rel. MS. $L_j = \sum_{k=1}^{j} \ell_k$
1	0	4	0,20	0,20	$0/30 = 0$	$0/30 = 0$
2	1	7	0,35	0,55	$7/30$	$7/30 \approx 0,23$
3	2	5	0,25	0,80	$10/30$	$17/30 \approx 0,57$
4	3	3	0,15	0,95	$9/30$	$26/30 \approx 0,87$
5	4	1	0,05	1,00	$4/30$	$30/30 = 1$

Wir tragen nun L_j über F_j auf und erhalten die sogenannte **Lorenzkurve** nach Max Otto Lorenz (1876-1959, US-amerikanischer Mathematiker).



Zur Lorenz-Kurve

Die Lorenzkurve ist der dicke Streckenzug.
Die Merkmale müssen der Größe nach geordnet werden.
Die Lorenzkurve beginnt in $(0/0)$ und endet in $(1/1)$.

Die Lorenzkurve verläuft nie oberhalb der Diagonale.
Bei einer **Gleichverteilung** (s.u.) fällt die Lorenzkurve mit der Diagonale zusammen. Je mehr sich die Lorenzkurve von der Diagonale entfernt, desto mehr haben wir es mit einer **Konzentration** zu tun.
Die Stärke der Konzentration lässt sich mit dem **Gini-Koeffizienten G** angeben: G ist der Quotient aus den Inhalten der Fläche zwischen Diagonale und Lorenzkurve und der Dreiecksfläche zwischen Diagonale und Abszisse (waagerechte Achse) , so dass $0 \le G \le 1$ gilt.
(Corrado Gini, italienischer Statistiker, 1884 – 1965)

Interpretation: Die nicht so lesefreudigen 40% aller Studenten besitzen etwa 14% der 30 Bücher, die nicht so lesefreudigen 50% der Studenten besitzen etwa 20% der 30 Bücher, usw.
Die Fläche zwischen der Diagonalen und der Lorenzkurve ist ein Maß für die **Konzentration**. Je kleiner diese Fläche ausfällt, desto geringer ist die Konzentration.

Spezialfall: Falls alle Merkmale x_j gleich sind, dann gibt es keine Konzentration und die Lorenzkurve fällt mit der Diagonalen zusammen. Diese Diagonale heißt deshalb auch **Gleichverteilungsgerade**.
Für alle j gilt $L_j = F_j$.

j	Merkmale x_j	absolute Häufigkeit h_j	relative Häufigkeit f_j	kumulierte relative Häufigk. F_j	relative Merkmalssumme ℓ_j	kumulierte rel. MS. $L_j = \sum_{k=1}^{j} \ell_k$
1	2	4	0,20	0,20	$8/40 = 0,2$	0,20
2	2	7	0,35	0,55	$14/40 = 0,35$	0,55
3	2	5	0,25	0,80	$10/40 = 0,25$	0,80
4	2	3	0,15	0,95	$6/40 = 0,15$	0,95
5	2	1	0,05	1,00	$2/40 = 0,05$	1,00

Klassierte Verteilungen am Beispiel der Ausgaben für die Haare

Die Ausgaben für die Haare schwanken zwischen 8 und 85 Euro. In diesem Fall ist es nicht sinnvoll, $85 - 7 = 78$ Rechtecke zu zeichnen. Es bieten sich verschiedene Klassen an.
Die **Klassenbreiten** w_j können unterschiedlich groß sein: Zum Beispiel $w_1 = 20$, $w_2 = 10$, ..., $w_5 = 30$ wie in der folgenden Tabelle.
Die absolute Häufigkeit h_j und die relative Häufigkeit $f_j = h_j / n$, n = Stichprobenumfang, einer Klasse hängen natürlich von deren Klassenbreite w_j ab. Mit steigender Breite w_j werden h_j und f_j dieser Klasse zunehmen. Aus Gründen der Vergleichbarkeit führt man deshalb den Begriff der Dichte einer Klasse ein:

Die **absolute Dichte** h_j^* einer Klasse ist definiert durch $\boxed{h_j^* = \frac{h_j}{w_j}}$ für j = 1, ... m.

Die **relative Dichte** f_j^* einer Klasse ist definiert durch $\boxed{f_j^* = \frac{f_j}{w_j}}$ für j = 1, ... m.

Wir führen m = 5 Klassen ein. F_j ist wieder die kumulierte relative Häufigkeit.

Ausgaben von ... bis unter ...	w_j	h_j	h_j^*	f_j	f_j^*	F_j
0 – 20	20	3	0,15	0,15	0,0075	0,15
20 – 30	10	3	0,3	0,15	0,015	0,30
30 – 40	10	5	0,5	0,25	0,025	0,55
40 – 60	20	4	0,2	0,2	0,01	0,75
60 – 90	30	5	0,167	0,25	0,0083	1
$\sum_{j=1}^{m=5}$	90	n = 20	–	1	–	–

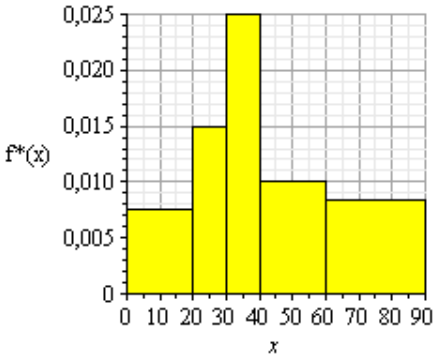
Wir zeichnen nun ein **Histogramm**.
Beim Histogramm ist nicht die Höhe der Rechtecke proportional zu den relativen Häufigkeiten f_j, sondern der Flächeninhalt der Rechtecke ist proportional zu den relativen Häufigkeiten f_j.

Somit ist die Höhe der Rechtecke gleich der relativen Dichte $\boxed{f_j^* = \frac{f_j}{w_j}}$.

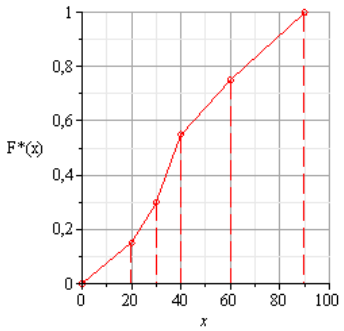
Die Summe aller Rechteckflächen ist damit immer 1.

Bei der klassierten Verteilungsfunktion werden die relativen Häufigkeiten f_j aufsummiert.

Histogramm



Klassierte Verteilungsfunktion



Manchmal sieht man Histogramme, bei denen h_j^* statt f_j^* über x aufgetragen wird. Dann ist die Summe der Rechteckflächen gleich n.

Maßzahlen der zentralen Tendenz

Die Maßzahlen der zentralen Tendenz beschreiben das Zentrum einer Verteilung. Dazu gehören der Modus, der Median, das arithmetische und das geometrische Mittel. Sie geben durch eine einzige Zahl eine Information über die Verteilung.

Def.: Unter dem **Modus (Modalwert) D** versteht man die am häufigsten vorkommende Ausprägung einer Variablen.
Beispiel 1: In der Zahlenreihe 1,2,3,3,3,3,4,4,5,5,5,6,6,6 ist 3 der Modus, da die 3 häufiger auftritt als die anderen Zahlen.

Vereinbarung: Haben zwei oder mehrere Werte die größte Häufigkeit, so liegt eine bimodale oder multimodale Verteilung vor.
Beispiel 2: In der Zahlenreihe 1,2,3,3,3,3,4,4,5,5,5,5,6,6,6 ist bimodal mit den Modalwerten 3 und 5..

Vereinbarung: Bei klassierten Daten ist der Modus diejenige Klasse mit der größten Zahl an Einträgen.

Def.: Unter dem **Median (Zentralwert Z, 2. Quartil Q_2)** versteht man denjenigen Wert, der in einer der Größe nach geordneten Reihe in der Mitte steht.

Beispiel 1: Die Zahlenreihe 3, 5, 2, 3, 2, 4, 2 wird zunächst geordnet: 2, 2, 2, $\boxed{3}$, 3, 4, 5. Der Median liegt bei den sieben Zahlen an vierter Stelle, also ist $Z = 3$.

Bei einer **ungeraden Anzahl n** von Daten $a_1, a_2, \ldots, a_n$, die durch das Kleiner-oder-gleich-Zeichen ($\leq$) geordnet sind, ist also der $\boxed{\text{Median } Z = a_k}$ mit $\boxed{k = \frac{1}{2}(n+1)}$.

Bei einer **geraden Anzahl n** von Daten ergibt sich nach dieser Formel für k keine ganze Zahl. In diesem Fall verwendet man den Mittelwert $\boxed{\text{Median } Z = \frac{a_k + a_{k+1}}{2}}$ mit $\boxed{k = \frac{1}{2}n}$.

Beispiel 2: Die Zahlenreihe 3, 2, 3, 2, 4, 2 wird zunächst geordnet: 2, 2, $\boxed{2}$, $\boxed{3}$, 3, 4. Also $Z = 2{,}5$.

Der Median (Zentralwert Z, 2. Quartil Q_2) teilt die Datenreihe in zwei gleich große Hälften.

Def.: Das **untere Quartil** Q_1 (**1. Quartil** oder **25%−Quantil**) ist derjenige Wert a_k der geordneten Datenreihe, für den mindestens 25% der Werte kleiner oder gleich a_k sind **und** mindestens 75% der Werte größer oder gleich a_k sind.

Def.: Das **obere Quartil** Q_3 (**3. Quartil** oder **75%−Quantil**) ist derjenige Wert a_k der geordneten Datenreihe, für den mindestens 75% der Werte kleiner oder gleich a_k sind **und** mindestens 25% der Werte größer oder gleich a_k sind.

Mit der folgenden Formel erhält man das 1. Quartil Q_1 für $p = 1/4$, den Median $Z = Q_2$ für $p = 1/2$ und das 3. Quartil Q_3 für $p = 3/4$.

$$Q = \begin{cases} a_{[n\cdot p]+1} & \text{falls } n\cdot p \text{ nicht ganzzahlig} \\ \dfrac{a_{n\cdot p} + a_{n\cdot p+1}}{2} & \text{falls } n\cdot p \text{ ganzzahlig} \end{cases}$$

mit der Gauß-Klammer [...]. Z.B. $[2,3] = 2$, $[2,9] = 2$.

Allgemein: $[x] = \max\{k \in \mathbb{Z} / k \le x\}$.

Beispiel 1: $n = 8$ Daten geordnet: 1, 3, 5, 6, 7, 9, 10, 11. Dann ist $Q_1 = 4$, $Z = Q_2 = 6{,}5$ und $Q_3 = 9{,}5$.
2 der 8 Daten, d.h. 25% sind $\le Q_1$. Ebenso: 6 der 8 Daten, d.h. 75% sind $\ge Q_1$.
4 der 8 Daten, d.h. 50% sind $\le Z$. Ebenso: 4 der 8 Daten, d.h. 50% sind $\ge Z$.
6 der 8 Daten, d.h. 75% sind $\le Q_3$. Ebenso: 2 der 8 Daten, d.h. 25% sind $\ge Q_3$.

Beispiel 2: $n = 7$ Daten geordnet: 1, 3, 5, 6, 7, 9, 10. Dann ist $Q_1 = 3$, $Z = 6$ und $Q_3 = 9$.
2 der 7 Daten, d.h. $2/7 \approx 28{,}6\%$ sind $\le Q_1$. Ebenso: 6 der 7 Daten, d.h. $6/7 \approx 85{,}7\%$ sind $\ge Q_1$.
4 der 7 Daten, d.h. $4/7 \approx 57{,}1\%$ sind $\le Z$. Ebenso: 4 der 7 Daten, d.h. $4/7 \approx 57{,}1\%$ sind $\ge Z$.
6 der 7 Daten, d.h. $6/7 \approx 85{,}7\%$ sind $\le Q_3$. Ebenso: 2 der 7 Daten, d.h. $2/7 \approx 28{,}6\%$ sind $\ge Q_3$.

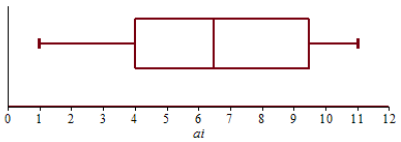
Beispiel 3: $n = 18$ Daten geordnet: $a_1, a_2, a_3, ..., a_{18}$. Wir suchen die vier Quintile, d.h. das 20%-Quantil, das 40%-Quantil, das 60%-Quantil und das 80%-Quantil.
Zum 20%-Quantil: $n \cdot p = 18 \cdot 0{,}2 = 3{,}6$ nicht ganzzahlig. Also $Q_{0,2} = a_{[3,6]+1} = a_4$.
4 der 18 Daten, d.h. $4/18 \approx 22{,}2\%$ sind $\le Q_{0,2}$ und 15 der 18 Daten, d.h. $15/18 \approx 83{,}3\%$ sind $\ge Q_{0,2}$.
Analog folgen $Q_{0,4} = a_{[18\cdot 0,4]+1} = a_9$, $Q_{0,6} = a_{[18\cdot 0,6]+1} = a_{11}$ und $Q_{0,8} = a_{[18\cdot 0,8]+1} = a_{15}$.

Der Box – and – Whisker Plot (kurz: Der Box-Plot)

Box = Schachtel, Whisker = Antenne, Barthaar, Schnurrbart

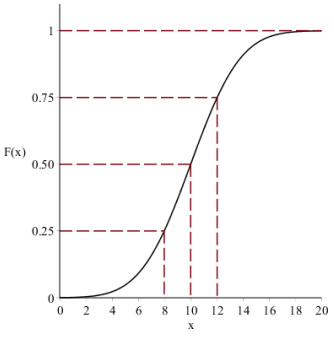
Die Box besteht aus einem Rechteck (Box) beliebiger Höhe, das die mittleren 50% der Verteilung umfasst, und vom 1. Quartil (25%-Grenze) bis zum 3. Quartil (75%-Grenze) reicht. Der Median wird durch einen senkrechten Strich in der Box gekennzeichnet. Die Länge der Whisker wird durch a_{min} und a_{max} bestimmt.

Box-Plot zum Beispiel 1:



Spannweite $a_{max} - a_{min} = 10$.

Zusatz: Gegeben sei eine Verteilungsfunktion $F(x)$, siehe Schaubild.
Z.B. ist $F(7) = 0{,}16$, d.h. die relative Häufigkeit für $x \le 7$ beträgt 0,16.
Umgekehrt kann man dem Schaubild die drei Quartile entnehmen.
Das 1. Quartil bei $F(x) = 0{,}25$ ergibt $Q_1 = 8$, so dass 25% der Daten kleiner oder gleich 8 und 75% der Daten größer oder gleich 8 sind.
Analog folgt $Q_2 = 10$ und $Q_3 = 12$.



Def.: Es sei $a_1, a_2, a_3, ..., a_i, ..., a_n$ die Urliste zu den Ausprägungen $x_1, x_2, x_3, ..., x_j, ..., x_m$. Dann versteht man unter dem **arithmetischen Mittel** $\bar{x}$:

1. $\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i}$. Folgerung: $\boxed{n \cdot \bar{x} = \sum_{i=1}^{n} a_i}$.

   Ein Scherz: „Ein Fußballer schießt zweimal aufs Tor, zuerst einen Meter rechts vorbei, danach einen Meter links vorbei. Statistisch gesehen landete der Ball mitten im Tor."

2. $\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot x_j}$ mit den absoluten Häufigkeiten h_j der Ausprägung x_j.

3. Wegen $\bar{x} = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot x_j = \sum_{j=1}^{m} \frac{h_j}{n} \cdot x_j = \sum_{j=1}^{m} f_j \cdot x_j$ folgt

   $\boxed{\bar{x} = \sum_{j=1}^{m} f_j \cdot x_j}$ mit den relativen Häufigkeiten f_j der Ausprägung x_j.

4. Den Mittelwert bei klassierten Daten kann man auf folgende Weise bekommen:
   Fall 1: Wenn die Mittelwerte $\bar{x}_j$ der einzelnen Klassen bekannt sind, dann ist $h_j \cdot \bar{x}_j$ die Summe der Elemente der j-ten Klasse. Und damit folgt $\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot \bar{x}_j}$.

   Fall 2: Wenn die Mittelwerte $\bar{x}_j$ der einzelnen Klassen nicht bekannt sind, dann ersetzt man in obiger Formel den Klassenmittelwert $\bar{x}_j$ durch die Klassenmitte x_j, um einen Schätzwert für $\bar{x}$ zu erhalten: $\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot x_j}$

5. Unter dem **geometrischen Mittel** $\bar{x}_{geom}$ versteht man die n-te Wurzel $\boxed{\bar{x}_{geom} = \sqrt[n]{a_1 \cdot a_2 \cdot ... \cdot a_n} = \sqrt[n]{\prod_{i=1}^{n} a_i}}$.

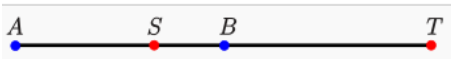
   Es ist nur sinnvoll für positive Zahlen.
   $\sqrt{a \cdot b}$ ist die Seitenlänge eines Quadrats, inhaltsgleich einem Rechteck mit den Seitenlängen a, b.
   $\sqrt[3]{a \cdot b \cdot c}$ ist die Kantenlänge eines Würfels, volumengleich einem Quader der Kantenlängen a, b, c.

   Beispiel: Ein Kapital K_0 wird im ersten Jahr mit 1%, im zweiten Jahr mit 2% und in dritten Jahr mit 3% verzinst. Bestimmen Sie die mittlere Verzinsung p.
   $K_3 = K_0 \cdot \left(1 + \frac{1}{100}\right) \cdot \left(1 + \frac{2}{100}\right) \cdot \left(1 + \frac{3}{100}\right) = K_0 \cdot \left(1 + \frac{p}{100}\right)^3$. Es folgt
   $\left(1 + \frac{p}{100}\right) = \sqrt[3]{\left(1 + \frac{1}{100}\right) \cdot \left(1 + \frac{2}{100}\right) \cdot \left(1 + \frac{3}{100}\right)} = \sqrt[3]{1,061106} \approx 1,01997$, so dass $p \approx 1,997$ beträgt.

6. Daneben gibt es noch das **harmonische Mittel** $\bar{x}_{harm} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + ... \frac{1}{a_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}$, bzw. $\frac{1}{\bar{x}_{harm}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i}$,

   also: **Der Kehrwert des harmonischen Mittels ist das arithmetische Mittel der Kehrwerte.**

In der Geometrie heißt eine Strecke $\overline{AB}$ durch die beiden Punkte S und T harmonisch geteilt, wenn $\frac{|\overline{SA}|}{|\overline{SB}|} = \frac{|\overline{TA}|}{|\overline{TB}|}$ gilt.



S ist dabei ein innerer, T ein äußerer Teilpunkt. Wenn man $|\overline{AS}| = s$, $|\overline{AB}| = e$ und $|\overline{AT}| = t$ wählt, dann gilt $\frac{s}{e-s} = \frac{t}{t-e}$, d.h. $s \cdot t - e \cdot s = e \cdot t - s \cdot t$ bzw. $2s \cdot t = e \cdot s + e \cdot t$. Nach Division durch $2s \cdot t \cdot r$ folgt $\frac{1}{e} = \frac{1}{2}\left(\frac{1}{t} + \frac{1}{s}\right)$ bzw. $\frac{1}{|\overline{AB}|} = \frac{1}{2}\left(\frac{1}{|\overline{AS}|} + \frac{1}{|\overline{AT}|}\right)$, so dass in diesem Fall das harmonische Mittel von $|\overline{AS}|$ und $|\overline{AT}|$ gleich $|\overline{AB}|$ ist.

Beispiel: Ein Fahrzeug fährt von A nach B mit der Geschwindigkeit v_1 und zurück von B nach A mit der Geschwindigkeit v_2. Wie groß ist die Durchschnittgeschwindigkeit v?

Mit $s = |AB|$ folgt die Zeit für die Hinfahrt $t_1 = \frac{s}{v_1}$ und für die Rückfahrt $t_2 = \frac{s}{v_2}$. Daraus folgt die Durchschnittsgeschwindigkeit $v = \frac{2s}{t_1 + t_2}$. Es folgt $\frac{1}{v} = \frac{t_1 + t_2}{2s} = \frac{1}{2} \cdot \left(\frac{t_1}{s} + \frac{t_2}{s}\right) = \frac{1}{2} \cdot \left(\frac{1}{v_1} + \frac{1}{v_2}\right)$, so dass v das harmonische Mittel von v_1 und v_2 ist.

Falls alle $a_i > 0$ sind, dann gilt: $\boxed{\bar{x}_{harm} \le \bar{x}_{geom} \le \bar{x}_{arith}}$. Und $\bar{x}_{harm} = \bar{x}_{geom} = \bar{x}_{arith} \Leftrightarrow$ alle a_i sind gleich. Siehe Hausaufgaben im Spezialfall $n = 2$.

7. Quadratische Mittel $\bar{x}_{quadr} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^{\,2}}$, Kubische Mittel $\bar{x}_{kub} = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} a_i^{\,3}}$. …

Beispiel 1: Es geht um die Anzahl der Frisörbesuche der Tabelle auf Seite 1.

a. Wir verwenden die einzelnen Beobachtungswerte.
Ihre Anzahlen werden der Größe nach geordnet: 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4.
Der Modus D ist $D = 3$. Ein dreifacher Frisörgang kommt am häufigsten vor.

Der Median Z (Zentralwert) ist der Mittelwert $Z = \frac{a_{10} + a_{11}}{2} = \frac{2+3}{2} = 2{,}5$. D.h. 50% aller Studenten gingen höchstens zweimal zum Frisör.

Das arithmetische Mittel $\bar{x}$ ist $\bar{x} = \frac{1}{n}\sum_{j=1}^{5} h_j \cdot x_j = \frac{1}{20}(3 \cdot 0 + 3 \cdot 1 + 4 \cdot 2 + 6 \cdot 3 + 4 \cdot 4) = \frac{45}{20} = 2{,}25$.

Außerdem ist das 1. Quartil $Q_1 = \frac{a_5 + a_6}{2} = 1$ und das 3. Quartil $Q_3 = \frac{a_{15} + a_{16}}{2} = 3$.

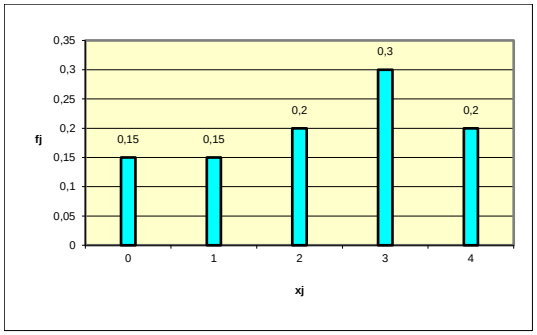
b. Wir verwenden die Häufigkeitstabelle für die Anzahl der Frisörbesuche:

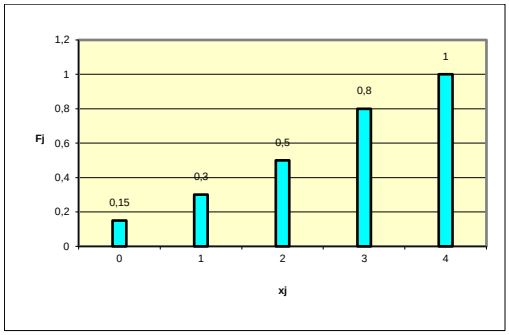
j	x_j	h_j	$h_j \cdot x_j$	f_j	$f_j \cdot x_j$	F_j
1	0	3	0	0,15	0	0,15
2	1	3	3	0,15	0,15	0,3
3	2	4	8	0,2	0,4	0,5
4	3	6	18	0,3	0,9	0,8
m = 5	4	4	16	0,2	0,8	1
$\sum$	–	20	45	1	**$\bar{x}$ = 2,25**	

Der Modus D ist $D = 3$, da h_4 am größten ist.
Der Median Z (Zentralwert) ist $Z = 2$, da für $j = 3$ zum ersten Mal $F_j \ge 0{,}5$ ist.

Das arithmetische Mittel $\bar{x}$ ist $\bar{x} = \frac{1}{n}\sum_{j=1}^{5} h_j \cdot x_j = \frac{1}{20}(3\cdot 0 + 3\cdot 1 + 4\cdot 2 + 6\cdot 3 + 4\cdot 4) = \frac{45}{20} = 2{,}25$.

c. Wir verwenden die grafischen Häufigkeitsverteilungen für die relative Häufigkeit und die kumulierte relative Häufigkeit





d. Wir bilden zwei Gruppen, mit und ohne Rastas.
Die Frisörbesuchszahlen der Rastas sind: 1, 2, 3, 4, 4, 3, 1, 3.
Die Frisörbesuchszahlen der Nicht-Rastas sind: 2, 2, 3, 0, 1, 0, 2, 4, 4, 0, 3, 3.
Das arithmetische Mittel der Frisörbesuchszahlen für die Rastas ist

$$\bar{x}_R = \frac{1}{8}(1+2+3+4+4+3+1+3) = \frac{21}{8} = 2{,}625$$

Das arithmetische Mittel der Frisörbesuchszahlen für die Nicht-Rastas ist

$$\bar{x}_{NR} = \frac{1}{12}(2+2+3+0+1+0+2+4+4+0+3+3) = \frac{24}{12} = 2.$$

Das arithmetische Mittel aller Frisörbesuchszahlen beträgt folglich

$$\frac{1}{20}\cdot\left(8\cdot\bar{x}_R + 12\cdot\bar{x}_{NR}\right) = \frac{1}{20}\cdot(8\cdot 2{,}625 + 12\cdot 2) = \frac{45}{20} = 2{,}25.$$

Beispiel 2: Es geht um die Ausgaben für die Haare nach der Tabelle auf Seite 1.

a. Wir verwenden die einzelnen Beobachtungswerte.
Sie werden der Größe nach geordnet:
8, 11, 14, 23, 23, 24, 30, 31, 32, 34, 35, 45, 45, 52, 58, 61, 63, 70, 73, 85.
Der Modus D ist nicht sinnvoll zu definieren.
Der Median Z ist $\frac{a_{10} + a_{11}}{2} = \frac{34+35}{2} = 34{,}5$, d.h. mindestens 50% der Studenten geben 34,50€ oder weniger für ihre Haare aus. Ebenso: mindestens 50% der Studenten geben 34,50€ oder mehr für ihre Haare aus.
Der Mittelwert ist $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} a_i = \frac{1}{20}\cdot 817 = 40{,}85$.

b. Wir verwenden jetzt die klassierten Daten von Seite 4.

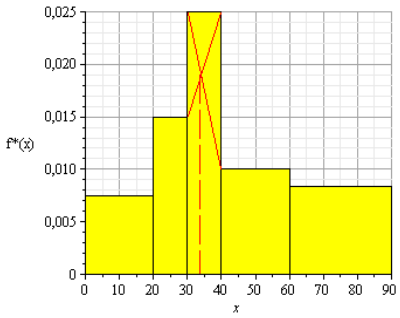
Ausgaben von ... bis unter ...	w_j	h_j	h_j^*	x_j	$h_j \cdot x_j$	f_j	$f_j^* = \frac{f_j}{w_j}$	F_j
0 – 20	20	3	0,15	10	30	0,15	0,0075	0,15
20 – 30	10	3	0,3	25	75	0,15	0,015	0,30
30 – 40	10	5	0,5	35	175	0,25	0,025	0,55
40 – 60	20	4	0,2	50	200	0,2	0,01	0,75
60 – 90	30	5	0,167	75	375	0,25	0,00833	1
$\sum_{j=1}^{m=5}$	90	20	–	–	855	1	–	–

In der Klasse von 30€ bis unter 40€ liegt die größte absolute Dichte h_j^* mit 0,5 Studenten pro €.

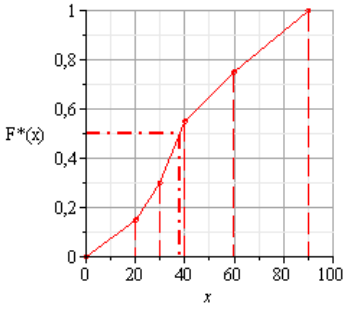
Für den Mittelwert folgt in unserem Beispiel $\bar{x} = \frac{855€}{20} = 42,75€$.

c. Wir verwenden jetzt die grafischen Häufigkeitsverteilungen.

Histogramm



Klassierte Verteilungsfunktion



Aus der linken Grafik folgt der Modus zu etwa D = 34.
Aus der rechten Grafik folgt der Zentralwert zu Z = 38.

Noch drei Eigenschaften des arithmetischen Mittels

1. Wenn die Messwerte $a_1, a_2, a_3, ..., a_n$ das arithmetische Mittel $\bar{x}$, haben, dann haben die Werte $y_i = c + d \cdot a_i$ das arithmetische Mittel $\bar{y} = c + d \cdot \bar{x}$.

   Denn $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(c + d \cdot a_i) = \frac{1}{n}\sum_{i=1}^{n} c + d \cdot \frac{1}{n}\sum_{i=1}^{n} a_i = c + d \cdot \bar{x}$.

2. Die Summe aller Abweichungen vom arithmetischen Mittel ist Null.

   Denn $\sum_{i=1}^{n}(a_i - \bar{x}) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \bar{x} = n \cdot \bar{x} - n \cdot \bar{x} = 0$.

3. Minimumseigenschaft: $\sum (a_i - A)^2 \geq \sum (a_i - \bar{x})^2$ für beliebige Zahlen A.

   Denn $\sum (a_i - A)^2 = \sum \left[(a_i - \bar{x}) + (\bar{x} - A) \right]^2$

   $= \sum (a_i - \bar{x})^2 + 2 \cdot (\bar{x} - A) \cdot \sum (a_i - \bar{x}) + \sum (\bar{x} - A)^2$

   $= \sum (a_i - \bar{x})^2 + 2 \cdot (\bar{x} - A) \cdot 0 + n \cdot (\bar{x} - A)^2$

   $\geq \sum (a_i - \bar{x})^2$.

   Oder: $\frac{d}{dA}\sum (a_i - A)^2 = -2\sum (a_i - A) = 0 \Rightarrow \sum (a_i - A) = 0 \Rightarrow \sum a_i - n \cdot A = 0$, also $A = \frac{1}{n}\sum_{i=1}^{n} a_i$.

   Wegen $\frac{d^2}{A^2}\sum (a_i - A)^2 = 2 > 0$ handelt es sich um ein Minimum.

<u>Maßzahlen für die Streuung: Die Varianz</u> (lateinisch: variantia = Verschiedenheit)

Gegeben sind die beiden Messreihen, die beide den Mittelwert $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i = 1$ besitzen.

①	a_i	0,5	1,2	1,4	0,8	1,1
	$a_i - \bar{x}$	–0,5	0,2	0,4	–0,2	0,1
②	a_i	1,0	1,1	1,0	0,9	1,0
	$a_i - \bar{x}$	0,0	0,1	0,0	–0,1	0,0

Der Messung ② vertraut man eher, da die Messwerte nicht so sehr streuen. Wir suchen nach einem Maß für die Streuung der einzelnen Messwerte um den Mittelwert $\bar{x}$. Die einzelnen Abweichungen betragen $a_i - \bar{x}$.

1. Versuch: Mittlere Abweichung $\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x}) = \frac{1}{n}\left(\sum_{i=1}^{n} a_i - n \cdot \bar{x}\right) = \frac{1}{n}\sum_{i=1}^{n} a_i - \bar{x} = \bar{x} - \bar{x} = 0$.

Der Wert 0 war zu erwarten, da der Mittelwert so liegt, dass die Abweichungen nach oben und unten im Mittel gleich groß sind.

2. Versuch: Mittlerer Abweichungsbetrag $\frac{1}{n}\sum_{i=1}^{n}\left|a_i - \bar{x}\right|$. Leider ist das Rechnen mit Beträgen nicht komfortabel.

3. Versuch: **Standardabweichung** $\boxed{\mathbf{s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2}}}$.

Für die Messreihe ① ergibt sich $s = \sqrt{\frac{1}{5}\left((-0,5)^2 + 0,2^2 + (-0,4)^2 + (-0,2)^2 + 0,1^2\right)} = \sqrt{0,1} \approx 0,316$

Für die Messreihe ② ergibt sich $s = \sqrt{\frac{1}{5}\left(0^2 + 0,1^2 + 0^2 + (-0,1)^2 + 0^2\right)} = \sqrt{0,004} \approx 0,063$ deutlich weniger als für die Messreihe ①.

Def.: Varianz $\boxed{\mathbf{s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2}}$ **.** Die Varianz ist das mittlere Abweichungsquadrat.

Der Term für die Varianz lässt sich umformen:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i^2 - 2a_i \cdot \bar{x} + \bar{x}^2) = \frac{1}{n}\sum_{i=1}^{n}a_i^2 - 2\bar{x} \cdot \frac{1}{n}\sum_{i=1}^{n}a_i + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}a_i^2 - 2\bar{x} \cdot \bar{x} + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}a_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

also **Varianz** $\boxed{\mathbf{s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 \quad mit \quad \overline{x^2} = \left(\frac{1}{n}\sum_{i=1}^{n}a_i^2\right) \quad und \quad \bar{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{n}a_i\right)^2}}$.

Die Varianz ist der **Mittelwert der Quadrate a_i minus dem Quadrat des Mittelwertes der a_i.**

①	a_i	0,5	1,2	1,4	0,8	1,1
	a_i^2	0,25	1,44	1,96	0.64	1,21
②	a_i	1,0	1,1	1,0	0,9	1,0
	a_i^2	1,00	1,21	1,00	0,81	1,00

Für die Messreihe ① ergibt sich $s = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{\frac{1}{5}\left(0,25 + 1,44 + 1,96 + 0,64 + 1,21\right) - 1^2} = \sqrt{0,1} \approx 0,316$.

Für die Messreihe ② ergibt sich $s = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{\frac{1}{5}\left(1,00 + 1,21 + 1,00 + 0,81 + 1,00\right) - 1^2} = \sqrt{0,004} \approx 0,063$.

Obige Formel verändert sich leicht, wenn der Messwert x_j mit der absoluten Häufigkeit h_j, $j = 1,...,m$ auftritt:

$$\mathbf{s^2 = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot (x_j - \bar{x})^2 = \overline{x^2} - \bar{x}^2 \quad bzw. \quad s^2 = \sum_{j=1}^{m} f_j \cdot (x_j - \bar{x})^2 = \overline{x^2} - \bar{x}^2}$$

$$\mathbf{mit \quad \overline{x^2} = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot x_j^2 = \sum_{j=1}^{m} f_j \cdot x_j^2 \quad und \quad \bar{x}^2 = \left(\frac{1}{n}\sum_{j=1}^{m} h_j \cdot x_j\right)^2 = \left(\sum_{j=1}^{m} f_j \cdot x_j\right)^2}$$

Beispiel:

j	x_j	h_j	$h_j \cdot x_j$	$h_j \cdot x_j^2$	f_j	$f_j \cdot x_j$	$f_j \cdot x_j^2$
1	0,80	3	2,4	1,92	0,06	0,048	0,0384
2	0,90	8	7,2	6,48	0,16	0,144	0,1296
3	1,00	20	20	20	0,4	0,4	0,4
4	1,10	14	15,4	16,94	0,28	0,308	0,3388
5	1,20	5	6	7,2	0,1	0,12	0,144
$m = 5$	$\sum_{j=1}^{5}$	$n = 50$	51,0	52,54	1	1,02	1,0508

Es folgt $\bar{x} = \frac{1}{50}\sum_{j=1}^{5} h_j \cdot x_j = \frac{1}{50} \cdot 51,0 = 1,02$ oder $\bar{x} = \sum_{j=1}^{5} f_j \cdot x_j = 1,02$.

Außerdem gilt $\overline{x^2} = \frac{1}{50}\sum_{j=1}^{5} h_j \cdot x_j^2 = \frac{1}{50} \cdot 52,54 = 1,0508$ oder $\overline{x^2} = \sum_{j=1}^{5} f_j \cdot x_j^2 = 1,0508$, so dass

$s^2 = \overline{x^2} - \bar{x}^2 = 0,0104$, also $s \approx 0,10$,

Satz zur linearen Transformation:

Die Beobachtungswerte a_i werden linear transformiert zu $\boxed{z_i = c + d \cdot a_i}$.

Dann gilt für die Werte z_i: $\boxed{\bar{z} = c + d \cdot \bar{x}}$ und $\boxed{s^2_Z = d^2 \cdot s^2_X}$ bzw. $\boxed{s_Z = |d| \cdot s_X}$.

Denn es gilt $\bar{z} = \frac{1}{n}\sum_{i=1}^{n}(c + d \cdot a_i) = \frac{1}{n}\left(n \cdot c + d \cdot \sum_{i=1}^{n} a_i\right) = c + d \cdot \frac{1}{n}\sum_{i=1}^{n} a_i = c + d \cdot \bar{x}$.

Außerdem ist $s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}\left(c + d \cdot a_i - (c + d \cdot \bar{x})\right)^2 = d^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\left(a_i - \bar{x}\right)^2 = d^2 \cdot s_X^2$.

Beispiel: Es sei $z_i = \frac{a_i - \bar{x}}{s_X} = -\frac{\bar{x}}{s_X} + \frac{1}{s_X} \cdot a_i$. Dann gilt: $\bar{z} = -\frac{\bar{x}}{s_X} + \frac{1}{s_X} \cdot \bar{x} = 0$ und $s_Z = \frac{1}{s_X} \cdot s_X = 1$.

$$\mathbf{Standardisierung: \ Für \ z_i = \frac{a_i - \bar{x}}{s_X} \ gilt \ \bar{z} = 0 \ und \ s_Z = 1}.$$

Zusammensetzung von mehreren Messreihen

Gegeben sind die Messreihen: 1. Messreihe: n_1 Messwerte, Mittelwert $\bar{x}_1$ und Varianz s_1^2

...

m. Messreihe: n_m Messwerte, Mittelwert $\bar{x}_m$ und Varianz s_m^2

Dann gilt für den gemeinsamen Mittelwert $\boxed{\bar{x} = \frac{1}{n_1 + n_2 + \ldots + n_m} \cdot \left(n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2 + \ldots n_m \cdot \bar{x}_m\right) = \frac{1}{n} \cdot \sum_{j=1}^{m} n_j \cdot \bar{x}_j}$

und für die gemeinsame Varianz

$$s^2 = s_{int}^2 + s_{ext}^2 \quad \text{mit} \quad s_{int}^2 = \frac{1}{n_1+n_2+\ldots+n_m}\cdot\left(n_1\cdot s_1^2 + n_2\cdot s_2^2 + \ldots n_m\cdot s_m^2\right) = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j\cdot s_j^2$$

$$\text{und} \quad s_{ext}^2 = \frac{1}{n_1+n_2+\ldots+n_m}\cdot\left(n_1\cdot\left(\bar{x}_1-\bar{\bar{x}}\right)^2 + n_2\cdot\left(\bar{x}_2-\bar{\bar{x}}\right)^2 + \ldots + n_m\cdot\left(\bar{x}_m-\bar{\bar{x}}\right)^2\right) = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j\cdot\left(\bar{x}_j-\bar{\bar{x}}\right)^2$$

Bemerkung: $s_{int}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j\cdot s_j^2$ lässt sich interpretieren als Mittelwert der einzelnen Varianzen s_j^2 und

$s_{ext}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j\cdot\left(\bar{x}_j-\bar{\bar{x}}\right)^2$ bedeutet die Varianz der einzelnen Mittelwerte $\bar{x}_j$.

Beweis: Die Messwerte der 1. Messreihe seien $x_{11}, x_{21}, x_{31}, \ldots, x_{n_1 1}$.

Die Messwerte der 2. Messreihe seien $x_{12}, x_{22}, x_{32}, \ldots, x_{n_2 2}$.

. . .

Die Messwerte der m-ten Messreihe seien $x_{1m}, x_{2m}, x_{3m}, \ldots, x_{n_m m}$ mit $n_1+n_2+\ldots+n_m = n$.

Die einzelnen Mittelwerte sind $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$, also $\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j) = 0$. Die einzelnen Varianzen sind

$s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$, so dass $\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 = n_j\cdot s_j^2$ gilt.

Dann gilt $s^2 = \frac{1}{n}\sum_{i=0}^{n}(x_i-\bar{\bar{x}})^2 = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\bar{\bar{x}})^2 = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left[(x_{ij}-\bar{x}_j)+(\bar{x}_j-\bar{\bar{x}})\right]^2 =$

$$\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left[(x_{ij}-\bar{x}_j)^2 + 2(x_{ij}-\bar{x}_j)(\bar{x}_j-\bar{\bar{x}}) + (\bar{x}_j-\bar{\bar{x}})^2\right] =$$

$$\frac{1}{n}\sum_{j=1}^{m}\left[\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 + 2\cdot(\bar{x}_j-\bar{\bar{x}})\cdot\underbrace{\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)}_{=0} + n_j\cdot(\bar{x}_j-\bar{\bar{x}})^2\right] =$$

$$\frac{1}{n}\sum_{j=1}^{m}\underbrace{\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2}_{=n_j\cdot s_j^2} + \frac{1}{n}\cdot\sum_{j=1}^{m} n_j\cdot(\bar{x}_j-\bar{\bar{x}})^2 =$$

$$\underbrace{\frac{1}{n}\sum_{j=1}^{m} n_j\cdot s_j^2}_{=s_{int}^2} + \underbrace{\frac{1}{n}\sum_{j=1}^{m} n_j\cdot(\bar{x}_j-\bar{\bar{x}})^2}_{=s_{ext}^2} = s_{int}^2 + s_{ext}^2 .$$

Beispiel für n = 2:

1. Messreihe	1,1	0,9	1,2	1,0	–
2. Messreihe	1,2	1,0	0,9	1,1	0,9

Es ist: $n_1 = 4$, $\bar{x}_1 = 1{,}05$, $s_1^2 = 0{,}0125$, $n_2 = 5$, $\bar{x}_2 = 1{,}02$, $s_2^2 = 0{,}0136$.

Es folgt $\bar{\bar{x}} = \frac{1}{n_1+n_2}\cdot\left(n_1\cdot\bar{x}_1 + n_2\cdot\bar{x}_2\right) = \frac{1}{4+5}\cdot(4\cdot 1{,}05 + 5\cdot 1{,}02) = \frac{31}{30} \approx 1{,}03$.

Außerdem folgt $s_{int}^2 = \frac{1}{n_1+n_2}\cdot\left(n_1\cdot s_1^2 + n_2\cdot s_2^2\right) = \frac{1}{4+5}\cdot(4\cdot 0{,}0125 + 5\cdot 0{,}0136) = \frac{59}{4500} \approx 0{,}0131$

und $s_{ext}^2 = \frac{1}{n_1+n_2}\cdot\left(n_1\cdot(\bar{x}_1-\bar{\bar{x}})^2 + n_2\cdot(\bar{x}_2-\bar{\bar{x}})^2\right) = \frac{1}{4+5}\cdot\left(4\cdot\left(1{,}05-\frac{31}{30}\right)^2 + 5\cdot\left(1{,}02-\frac{31}{30}\right)^2\right) = \frac{1}{4500}$, so dass

$s^2 = s_{int}^2 + s_{ext}^2 = \frac{60}{4500} = \frac{1}{75} \approx 0{,}0133$ und $s \approx 0{,}115$.

Die Varianz bei klassierten Daten

Wir verwenden obige Formeln $\boxed{\bar{x} = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j \cdot \bar{x}_j = \sum_{j=1}^{m} f_j \cdot \bar{x}_j}$ und

$$\boxed{s^2 = s_{int}^2 + s_{ext}^2 \quad \text{mit} \quad s_{int}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j \cdot s_j^2 = \sum_{j=1}^{m} f_j \cdot s_j^2 \quad \text{und} \quad s_{ext}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j \cdot \left(\bar{x}_j - \bar{x}\right)^2 = \sum_{j=1}^{m} f_j \cdot \left(\bar{x}_j - \bar{x}\right)^2}.$$

Für die Mittelwerte $\bar{x}_j$ der einzelnen Klassen nehmen wir einfach **näherungsweise die Klassenmitten** x_j.

Aber wie können wir die Varianzen s_j^2 der einzelnen Klassen bekommen? Die Varianz hängt stark von der Verteilung der einzelnen Daten innerhalb der Klasse ab.

→ Falls alle Daten innerhalb einer Klasse identisch sind, dann ist die Varianz $s^2 = 0$.

→ Der andere Extremfall wäre gegeben, wenn bei der Klasse $a \le x < b$ der Wert a $\frac{n}{2}$-mal und ein Wert dicht bei b (also b) ebenfalls $\frac{n}{2}$-mal vorkommt. Dann ist das arithmetische Mittel $\bar{x} = \frac{\frac{n}{2}\cdot a + \frac{n}{2}\cdot b}{n} = \frac{a+b}{2}$

und die Varianz $s^2 = \frac{1}{n}\left(\frac{n}{2}\cdot\left(a - \frac{a+b}{2}\right)^2 + \frac{n}{2}\cdot\left(b - \frac{a+b}{2}\right)^2\right) = \frac{1}{2}\left(\left(\frac{a-b}{2}\right)^2 + \left(\frac{b-a}{2}\right)^2\right) = \frac{(b-a)^2}{4}$.

→ Oder wenn in der Klasse $a \le x < b$ die drei Werte a, $\frac{a+b}{2}$ und b gleich oft (jeweils $\frac{n}{3}$-mal) vorkommen, dann ist der Mittelwert $\bar{x} = \frac{a+b}{2}$ und die Varianz

$$s^2 = \frac{1}{n}\left(\frac{n}{3}\cdot\left(a - \frac{a+b}{2}\right)^2 + \frac{n}{3}\cdot\left(\frac{a+b}{2} - \frac{a+b}{2}\right)^2 + \frac{n}{3}\cdot\left(b - \frac{a+b}{2}\right)^2\right) = \frac{1}{3}\left(\left(\frac{a-b}{2}\right)^2 + 0 + \left(\frac{b-a}{2}\right)^2\right) = \frac{(b-a)^2}{6}.$$

→ Wenn viele Werte gleichmäßig auf die Klasse $a \le x < b$ verteilt sind, dann ist $\bar{x} = \frac{a+b}{2}$ und

$$s^2 = \frac{1}{b-a}\cdot\int_a^b \left(x - \tfrac{a+b}{2}\right)^2 dx = \frac{1}{3(b-a)}\cdot\left[\left(x - \tfrac{a+b}{2}\right)^3\right]_a^b = \frac{1}{3(b-a)}\cdot\left(\left(\tfrac{b-a}{2}\right)^3 - \left(\tfrac{a-b}{2}\right)^3\right) = \frac{(b-a)^2}{12}.$$

Diese Verteilung wollen wir im Folgendem verwenden. Dann gilt

$$\boxed{s^2 = s_{int}^2 + s_{ext}^2 \quad \text{mit} \quad s_{int}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j \cdot s_j^2 = \sum_{j=1}^{m} f_j \cdot s_j^2 = \sum_{j=1}^{m} f_j \cdot \frac{w_j^2}{12} \quad \text{und} \quad s_{ext}^2 = \frac{1}{n}\cdot\sum_{j=1}^{m} n_j \cdot \left(\bar{x}_j - \bar{x}\right)^2 = \sum_{j=1}^{m} f_j \cdot \left(\bar{x}_j - \bar{x}\right)^2}$$

Als **Beispiel** verwenden wir wieder die klassierten Daten von Seite 5.

Ausgaben von ... bis unter ...	w_j	h_j	f_j	$f_j \cdot w_j^2$	x_j	$f_j \cdot x_j$	$(x_j - \bar{x})^2$	$f_j \cdot (x_j - \bar{x})^2$
0 – 20	20	3	0,15	60	10	1,5	1072,5625	160,884375
20 – 30	10	3	0,15	15	25	3,75	315,0625	47,259375
30 – 40	10	5	0,25	25	35	8,75	60,0625	15,015625
40 – 60	20	4	0,2	80	50	10	52,5625	10,5125
60 – 90	30	5	0,25	225	75	18,75	1040,0625	260,015625
$\sum_{j=1}^{m=5}$	90	20	1	405	–	$\bar{x} = 42,75$	–	493,6875

Und damit folgt $s_{int}^2 = \frac{405}{12} = 33,75$, $s_{ext}^2 = 493,6875$, also $s^2 = s_{int}^2 + s_{ext}^2 \approx 527,4375$ und $s \approx 22,97$.