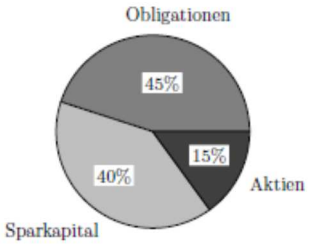
2. Beschreibung von eindimensionalen (univariaten) Stichproben

Bei eindimensionalen (univariaten) Daten wird nur ein Merkmal untersucht. Der Fall von zwei- oder mehrdimensionalen Daten wird im nächsten Kapitel untersucht.

2.1 Tabellarische und grafische Darstellungen

a)
Für die Darstellung von nominalen (qualitativen) Daten eignen sich **Kreisdiagramme** (englisch: Pie Charts).

Beispiel:
Zusammensetzung einer Kapitalanlage bzw. eines Portfolios (englisch: Asset Allocation) mit niedrigem Risiko: 45% Obligationen, 40% Liquidität, 15% Aktien



Übungsaufgabe:

Erstellen eines Kreisdiagramm, in dem das Verhältnis der Anzahl der weiblichen und männlichen Studierenden aufgezeigt wird.

Für metrische Daten ist die folgende Darstellung gut geeignet:

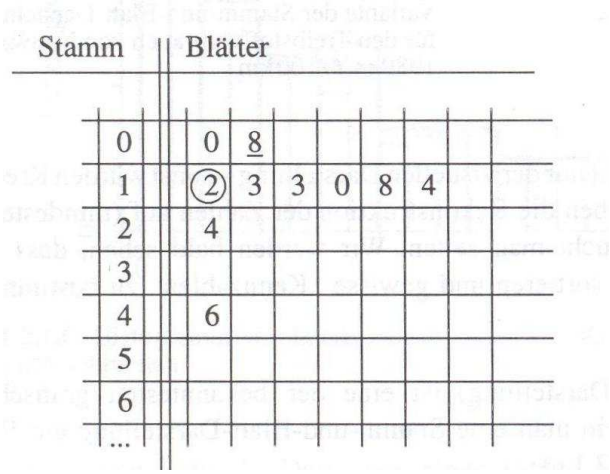
b) Die **Stamm-Blatt-Darstellung (engl. stem and leaf display)**

In einer Studie über die Wirksamkeit zweier Schlafmittel wurde bei 10 Personen die durchschnittliche Schlafverlängerung durch Medikament A gegenüber B gemessen. Für die beobachteten Werte ergab sich die folgende Urliste:

1.2 2.4 1.3 1.3 0.0 1.0 1.8 0.8 4.6 1.4 (Einheit: Stunden)

Geordnete Urliste
0.0 0.8 1.0 1.2 1.3 1.3 1.4 1.8 2.4 4.6

Die Stamm-Blatt-Darstellung erlaubt rasch einen Überblick über die Verteilung.
Beispielweise stellt die Ziffer 2 im Kreis den Wert 1.2 dar.



Bezeichnungen:
Die erhaltenen Daten werden mit $x_1, x_2, ….., x_n$ bezeichnet. n heisst Umfang der Stichprobe.
Im Beispiel ist $x_1 = 1.2$ und $x_{10} = 1.4$ und $n = 10$.
Ordnet man die Daten aufsteigend nach der Grösse, so erhält man die geordnete Stichprobe.
Im Beispiel:

	0.0	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6
Rang	1	2	3	4	5.5	5.5	7	8	9	10

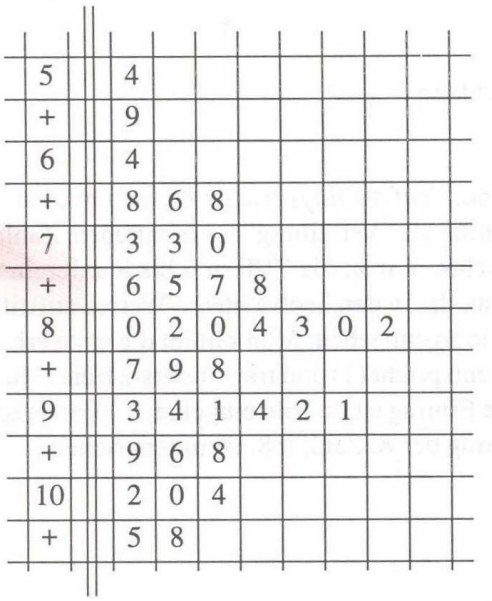
Damit erhält jeder Stichprobenwert x_i einen Rang $R(x_i)$. Bei gleichen Zahlen werden die entsprechenden Rangzahlen gemittelt.

Bemerkung:
Die Beispiele sind dem Buch von W. Stahel: Statistische Datenanalyse, Vieweg 1995 entnommen.

Beim folgenden Beispiel wird der Stamm in zwei Blattansätze unterteilt, je einen für tiefe (0 – 4) und für hohe (5 – 9) Werte.

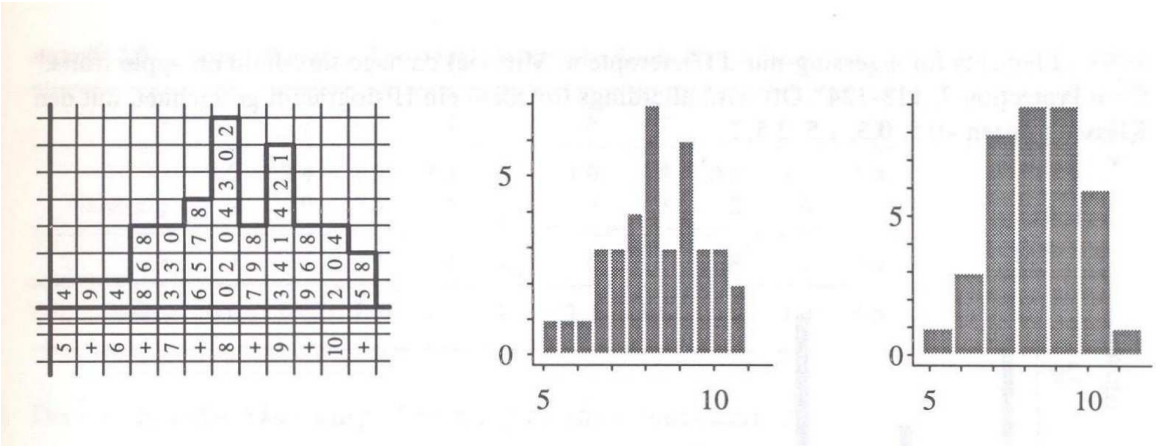
Treibstoffverbrauch von Neuwagen 1989 in ℓ/100 km

Die Verbrauchswerte liegen zwischen 5.4 ℓ/100 km und 10.8 ℓ/100 km.



c) Das **Histogramm**

Das Histogramm (Säulendiagramm, englisch Bar Chart) kann aus der Stamm-Blatt-Darstellung erhalten werden, indem man diese um 90° dreht und die Konturen nachzeichnet.



Die Abbildung zeigt, dass die Klasseneinteilung das Histogramm beeinflussen kann.
Die Fläche muss proportional zu Häufigkeit sein (und nicht zur Säulenhöhe).

2.2 Kennwerte und Masszahlen einer Stichprobe

Die Verteilung der Stichprobe kann mit Kennzahlen für die Lage und die Streuung charakterisiert werden.

2.2.1 **Lagemasse**

a) **Empirischer Mittelwert $\bar{x}$**

Die bekannteste Kennzahl für die Lage ist der empirische Mittelwert $\bar{x}$ einer Stichprobe vom Umfang n

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n) \qquad 1)$$

Im Beispiel der Schlafdaten ist

$$\bar{x} = \frac{1}{10} \cdot (0.0 + 0.8 + 1.0 + 1.2 + 1.3 + 1.3 + 1.4 + 1.8 + 2.4 + 4.6) = 1.58$$

Bemerkungen:

Für die Differenzen zum Mittelwert gilt:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Der Mittelwert entspricht also dem Schwerpunkt in der Physik:
Denkt man sich nämlich einen leichten Stab, der in den Abständen x_i von einem Ende mit gleichen Gewichten belastet ist, dann ist er bei $\bar{x}$ - dem Schwerpunkt - zu unterstützen, um ihn im Gleichgewicht zu halten.

Dieses Lagemass ist allerdings nicht robust, denn das Hinzufügen oder Weglassen einer (extremen) Beobachtung kann den Mittelwert stark beeinflussen. Wird im Beispiel der Wert 4.6 wegen eines Tippfehlers durch 46 ersetzt, so ergibt sich als arithmetisches Mittel 5.72 statt 1.58.

b) **Median** und **Quartile**

Der Median (oder Zentralwert) med teilt die der Grösse nach geordnete Stichprobe in zwei gleiche Hälften. Unterhalb und oberhalb des Medians liegen also gleich viele Beobachtungen. Ist die Stichprobenzahl ungerade so ist der Median gleich dem Glied in der Mitte der geordneten Stichprobe, bei gerader Stichprobenzahl wie im Beispiel ist der Median med gleich dem arithmetischen Mittel der beiden Mittelglieder.
Im Beispiel der Schlafdaten ist

$$med = \frac{1}{2}(1.3 + 1.3) = 1.3$$

Bemerkungen:

Der Median ist unempfindlich gegenüber statistischen Ausreissern, also im Gegensatz zum empirischen Mittelwert ein robustes Lagemass.

Zur Beschreibung von asymmetrischen Verteilungen z.B. beim Einkommen ist der Median als Lagemass besser geeignet als der empirische Mittelwert.

Die beiden Lagemasse haben folgende Extremaleigenschaften:

Illustration am folgenden
Beispiel:

i	1	2	3	4	5
x_i	3	3	4	6	9

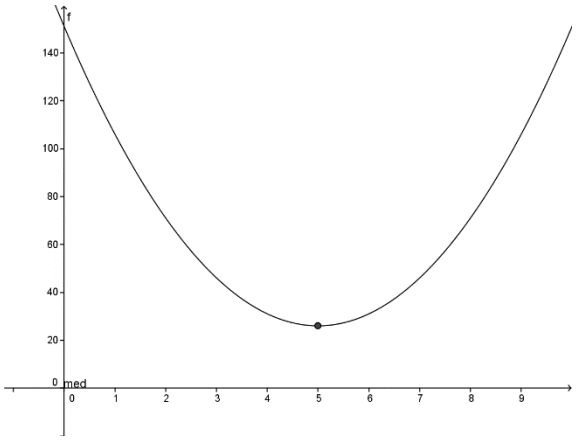
Für das arithmetische Mittel ist

$$\sum_{i=1}^{n}(x_i - x)^2$$

minimal. Das arithmetische Mittel ist also im Sinne der (im Kapitel „Lineare Regression“ erwähnten) Methode der kleinsten Quadrate minimal.
In der Abbildung ist die Funktion

$$f(x) = \sum_{i=1}^{5}(x_i - x)^2$$

dargestellt. Ihr Minimum liegt beim empirischen Mittelwert
$\bar{x} = 5$.



Für den Median ist

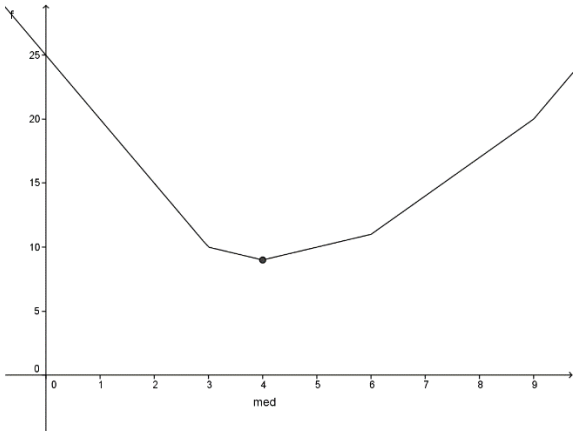
$$\sum_{i=1}^{n}|x_i - x|$$

minimal.

Beispiel:

i	1	2	3	4	5
x_i	3	3	4	6	9

Die Stichprobe hat den Median med = 4.



In der Abbildung ist die Funktion

$$f(x) = \sum_{i=1}^{5}|x_i - x|$$

dargestellt. Ihr Minimum liegt beim Median med = 4.

Quartile
Unter dem 1.Quartil Q_1 versteht man den Median der Werte unterhalb vom Median
Unter dem 3. Quartil Q_3 versteht man den Median der Werte oberhalb vom Median

Im Beispiel der Schlafdaten ist das 1. Quartil 1.0 und das 3. Quartil 1.8.

Übungsaufgabe:
Welche der folgenden Aussagen über die Prüfungsnoten einer Klasse sind immer richtig?
a)
Ist meine Prüfungsnote grösser als das arithmetische Mittel, dann gehöre ich zur besseren Hälfte der Klasse.
b)
Mit der Note 3.5 bin ich bei einem arithmetischen Mittel von 4 sicher nicht der schlechteste der Klasse.
c)
Liegt meine Prüfungsnote unter dem Median, dann gehöre ich zur schlechteren Hälfte der Klasse.
d)
Habe ich eine 6 und der Median ist 4, so gibt es mindestens eine Note 2 bzw. zwei Noten 3, resp. 4 Noten 3.5,..).
e) Entspricht meine Note genau dem Median, dann hat die Klasse eine ungerade Anzahl von Studierenden.
f) Haben drei Studierende eine ungenügende Note und das arithmetische Mittel ist 4, dann habe auch drei Studierende Noten über 4.

Lösungen:
a) falsch, Gegenbeispiel: 6, 4.5, 4, 1
b) falsch, Gegenbeispiel: 4.5, 4, 4, 4, 4, 4, 4, …, 3.5
c) richtig
d) falsch, die Aussage würde für das empirische Mittel gelten. Gegenbeispiel: 6, 4, 4
e) falsch, die Aussage gilt nur, wenn alle Noten verschieden sind. Gegenbeispiel: 2, 4, 4, 5
f) falsch, die Aussage würde für den Median gelten, Gegenbeispiel: 3.5, 3.5, 3.5, 5.5.

2.2.2 Streuungsmasse

Der Mittelwert oder der Median allein kann eine Verteilung noch nicht genügend kennzeichnen, denn die Daten können mehr oder weniger um die mittlere Lage streuen. Gesucht ist eine geeignete Zahl als Mass dafür, wie stark die Beobachtungen im Mittel vom Mittelwert bzw. Median abweichen.

a) Die **Quartilsdifferenz**
Als Streuungsmass kann man etwa die Quartilsdifferenz $Q_3 - Q_1$verwenden. Sie misst die Länge eines Intervalls, in dem die Hälfte aller Werte liegt.

Im Beispiel Schlafdaten hat sie den Wert 1.8 – 1 = 0.8

b) **Die mittlere absolute Abweichung**

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Im Beispiel der Schlafdaten ergibt sich

$$\bar{x} = \frac{1}{10}(|0 - 1.58| + |0.8 - 1.58| + \cdots + |4.6 - 1.58|) = 0.812$$

Aus mathematischen Gründen werden aber nach Gauss die im Folgenden definierten Streuungsmasse benützt. Statt der Beträge werden die Quadrate der Abweichungen vom Mittelwert betrachtet.

c) Varianz s^2 und **Standardabweichung s**
Die Quadratwurzel aus der Varianz heisst Standardabweichung s (englisch: standard deviation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad 2)$$

Im Beispiel der Schlafdaten ergibt sich

$$s^2 = \frac{1}{9}((0 - 1.58)^2 + (0.8 - 1.58)^2 + + \cdots + (4.6 - 1.58)^2) = 1.51$$

$s = \sqrt{1.51} = 1.23$

Die Varianz kann auch in der folgenden Form geschrieben werden:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right) \qquad 2')$$

Diese Form eignet sich allerdings nicht für das Programmieren.

Auch hier ergibt sich eine physikalische Analogie: Die Varianz misst das Trägheitsmoment bezüglich des Schwerpunkts.

Bemerkung:
Sowohl Varianz als auch Standardabweichung sind die gebräuchlichsten Streuungsmasse. Da sie aber von Ausreissern stark beeinflusst werden, sind sie nicht robust.

Hin und wieder ist es sinnvoll, die Streuung mit dem Mittelwert zu vergleichen.

Der Variationskoeffizient cv (englisch coefficient of variation) ist folgendermassen definiert:

$$cv = \frac{s}{\bar{x}}$$

Im Beispiel der Schlafdaten ist
$cv \approx \frac{1.23}{1.58} \approx 0.78$

Bemerkung:
Sind die Standardabweichungen zu den entsprechenden Mittelwerten ungefähr proportional, so bleibt der Variationskoeffizient ungefähr gleich.

Mit den Formeln 1) und 2‘) können die Kennzahlen aus geeigneten Summen berechnet werden. Im Beispiel der Schlafdaten erhält man die folgende Tabelle:

Schlafdaten

i	x_i	x_i^2	
1	1.2	1.44	
2	2.4	5.76	
3	1.3	1.69	
4	1.3	1.69	
5	0	0	
6	1	1	
7	1.8	3.24	
8	0.8	0.64	
9	4.6	21.16	
10	1.4	1.96	
	15.8	38.58	Summe
		1.58	Mittelwert
		1.51	Varianz
		1.23	Standardabweichung
		0.78	Variationskoeffizient

Eine Anwendung aus der Finanzwirtschaft

Die Renditen einer Anlage oder eines Portfolios streuen mehr oder weniger stark. Als Mass für die Volatilität (das Risiko) der Renditen wird häufig die (annualisierte) Standardabweichung verwendet.

Beispiel:
Für das folgende Portfolio ergibt sich in den 8 Jahren eine mittlere Rendite von 2.7% und eine Volatilität von 11.5%.

Wie im Abschnitt 2.3 gezeigt wird, ermöglicht die Volatilität eine Vorhersage, in welchem Intervall sich die Renditen in ungefähr 95% aller Jahre bewegen.

Jahr t	Rendite r	r^2	
1	0.1398	0.01954	
2	-0.1200	0.01440	
3	-0.0711	0.00506	
4	-0.0321	0.00103	
5	0.1292	0.01669	
6	0.2007	0.04028	
7	0.0000	0.00000	
8	-0.0336	0.00113	
Summe	0.2129	0.0981	
	s^2	0.0132	2.7% mittlere Rendite
	s	0.1149	11.5% Volatilität

2.3. Klassierte Werte

Ist die Stichprobenzahl grösser, dann werden die Daten nicht einzeln angegeben, sondern sie werden in Klassen gleicher Breite mit den Klassenmitten z_j eingeteilt. Für die Anzahl k der Klassen wird etwa die Faustregel $k \approx \sqrt{n}$ (Variante: $k \leq 5 \cdot log_{10} n$) verwendet.

Gewichte von n = 100 zweiwöchigen Küken in g

107	117	105	106	114	105	113	88	119	116
108	98	104	126	102	100	120	121	87	110
111	114	121	114	104	94	101	94	95	114
101	82	111	108	100	109	92	96	108	108
97	92	112	105	112	100	108	105	97	119
113	102	103	100	94	102	104	110	127	102
109	100	76	101	95	96	118	91	118	107
105	112	92	99	118	100	130	112	110	103
116	115	96	125	97	114	111	101	101	90
122	106	109	116	103	134	86	124	107	107

Die Daten dieser Urliste werden in k = 12 Klassen der Breite 5 eingeteilt. Die absoluten Häufigkeiten n_j und die relativen Häufigkeiten $\frac{n_j}{n}$ können aus dem Stamm-Blatt-Diagramm herausgelesen werden. Die einzelnen Werte in einer Klasse werden durch die Klassenmitte ersetzt.
Beispielsweise werden die Daten 88, 87, 86 der Klasse 85 – 89 durch die Klassenmitte 87 ersetzt. Bei gerundeten Werten würde die Klasse aus dem Intervall [84.5, 89.5[bestehen.

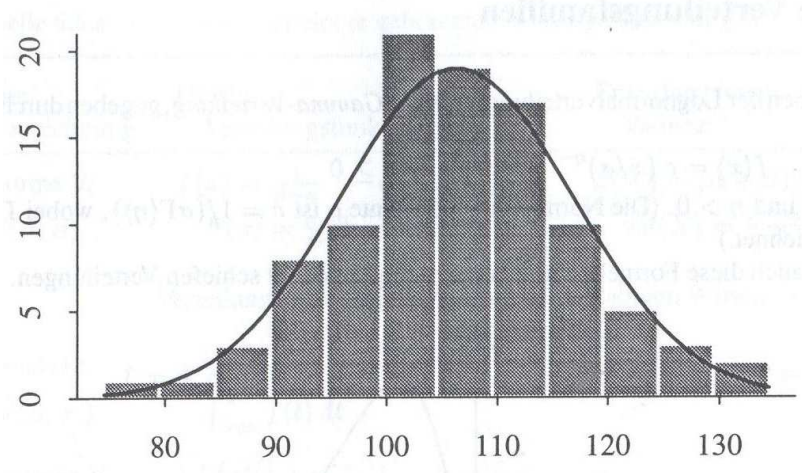
Klasse	Stamm	Blatt	Häufigkeiten abs.	rel.
75-79	7+	6	1	0.01
80-84	8	2	1	0.01
85-89	8+	8 7 6	3	0.03
90-94	9	4 4 2 2 4 1 2 0	8	0.08
95-99	9+	8 5 6 7 7 5 6 9 6 7	10	0.10
100-104	10	4 2 0 4 1 1 0 0 2 3 0 2 4 2 0 1 0 3 1 1 3	21	0.21
105-109	10+	7 5 6 5 8 8 9 8 8 5 8 5 9 7 5 6 9 7 7	19	0.19
110-114	11	4 3 0 1 4 4 4 1 2 2 3 0 2 2 0 4 1	17	0.17
115-119	11+	7 9 6 9 8 8 8 6 5 6	10	0.10
120-124	12	0 1 1 2 4	5	0.05
125-129	12+	6 7 5	3	0.03
130-134	13	0 4	2	0.02

Aus dem Diagramm kann der Median als Mittelwert der 50. und 51. Zahlen in der geordneten Stichprobe bestimmt werden. Ordnet man die Klasse 105-109 der Grösse nach, so erhält man den Median
$\frac{1}{2}(106 + 106) = 106$. Das 1.Quartil ergibt sich als 25. Wert zu 100 und das 3. Quartil als 75.Wert zu 114. Der Quartilsabstand beträgt damit 14.

Dreht man das Stamm-Blattdiagramm um 90°, so ergibt sich das folgende Histogramm.

Klasse	Stamm	Blatt
75-79	7+	6
80-84	8	2
85-89	8+	8 7 6
90-94	9	4 4 2 2 4 1 2 0
95-99	9+	8 5 6 7 7 5 6 9 6 7
100-104	10	4 2 0 4 1 1 0 0 2 3 0 2 4 2 0 1 0 3 1 1 3
105-109	10+	7 5 6 5 8 8 9 8 8 5 8 5 9 7 5 6 9 7 7
110-114	11	4 3 0 1 4 4 4 1 2 2 3 0 2 2 0 4 1
115-119	11+	7 9 6 9 8 8 8 6 5 6
120-124	12	0 1 1 2 4
125-129	12+	6 7 5
130-134	13	0 4

Zu erkennen ist die Glockenform des Histogramms. Im Kapitel Schliessende Statistik wird gezeigt, dass in diesem Fall die Normalverteilung ein gutes Modell darstellt. Dies tritt häufig dann auf, wenn viele voneinander unabhängige kleine Faktoren die Messgrösse positiv oder negativ beeinflussen und sich damit ungefähr aufheben. Im Beispiel Küken ist es plausibel, dass sich wachstumshemmende und -fördernde Einflüsse ungefähr kompensieren. In Bild ist bereits eine angepasste Normalverteilung dargestellt.



Für die vereinfachte Berechnung von Mittelwert und Standardabweichung werden die Beobachtungen in einer bestimmten Klasse durch die Klassenmitten ersetzt.

Im Beispiel erhält man für

$$\bar{x} \approx \frac{1}{100}(1 \cdot 77 + 1 \cdot 82 + 3 \cdot 87 + 8 \cdot 92 + \cdots + 3 \cdot 127 + 2 \cdot 132) = 106.3$$

Bemerkung:
Bei der Berechnung mit den einzelnen Werten ergibt sich $\bar{x} = 106.2$.

Für die Varianz gilt dann näherungsweise

$$s^2 \approx \frac{1}{100-1}(1\cdot(77-106.3)^2+1\cdot(82-106.3)^2+8\cdot(92-106.3)^2+\cdots \\ +3\cdot(127-106.3)^2+2\cdot(132-106.3)^2)=111.12$$

und damit für die Standardabweichung s

$s=\sqrt{111.12}\approx 10.5$

Bemerkung:
Die Berechnung mit den einzelnen Werten ergibt für s den Wert 10.6.

Allgemein:
Die n Einzelwerte werden in k Klassen mit den Klassenmitten z_j eingeteilt. Ist n_j die absolute Häufigkeit der Werte in der Klasse j dann gelten für den Mittelwert $\bar{x}$ und die Varianz bzw. Standardabweichung s näherungsweise die folgenden Formeln:

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} n_j z_j$$

Mittelwert 3)

$$s^2 \approx \frac{1}{n-1}\sum_{j=1}^{k} n_j \left(z_j-\bar{x}\right)^2$$

Varianz bzw. Standardabweichung 4)

Bequemer ist die folgende Formel 4‘)

$$s^2=\frac{1}{n-1}\cdot\left(\sum_{j=1}^{k} n_j z_j^2 - n\cdot\bar{x}^2\right)$$

4´)

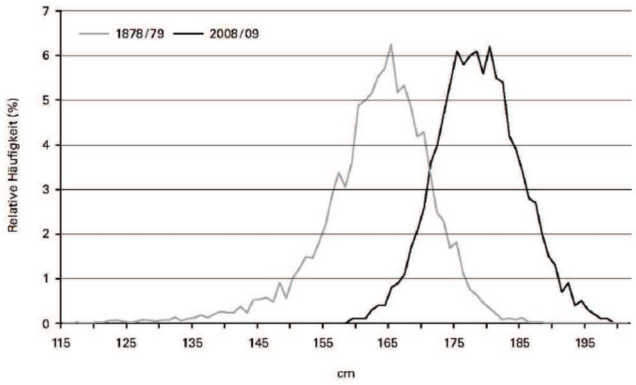
Sind die Daten ungefähr normalverteilt, dann ermöglicht die Standardabweichung die folgende Aussage:

Das Intervall $[\bar{x}-s,\bar{x}+s]$ enthält etwa 68% und
das Intervall $[\bar{x}-2s,\bar{x}+2s]$ etwa 95% aller Werte.
Ausserhalb des Intervalls $[\bar{x}-3s,\bar{x}+3s]$ liegen fast keine Werte.

Diese Aussage ist im Kükenbeispiel tatsächlich ungefähr erfüllt.

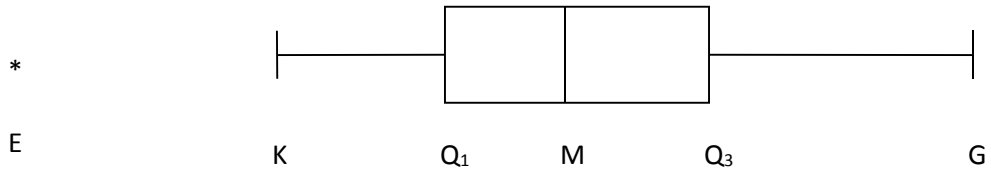
Auch im folgenden Beispiel (Quelle: *Swiss Medical* Weekly 141 2011) ist die Normalverteilung gut zu erkennen:

Verteilung der Körperhöhe von 19-jährigen Stellungspflichtigen im Kanton Bern 1878/79 und 2008/09



2.4 Box- und Whiskerplot

Einen Gesamtüberblick über die Quartilwerte gibt das Boxplot.
Die Kiste (Box) wird vom oberen 3. Quartil Q_3 und vom unteren ersten Quartil Q_1 begrenzt und beim Median M unterteilt. Für die Whiskers (engl. Schnurrhaar einer Katze, vgl. „Katzen würden Whiskas kaufen!“) trägt man nach beiden Seiten das Eineinhalb-fache des Quartilsabstands ab. Extrem grosse Beobachtungen wie E werden mit Sternchen bezeichnet. Alle kleineren bzw. grösseren Werte werden als Extremwerte bezeichnet und mit einem Stern markiert.
Die Regelungen für die Whiskers sind allerdings nicht einheitlich.



E Extremwert

K kleinster Wert, der nicht zu den Extremwerten gehört

Q_1 erster Quartilwert

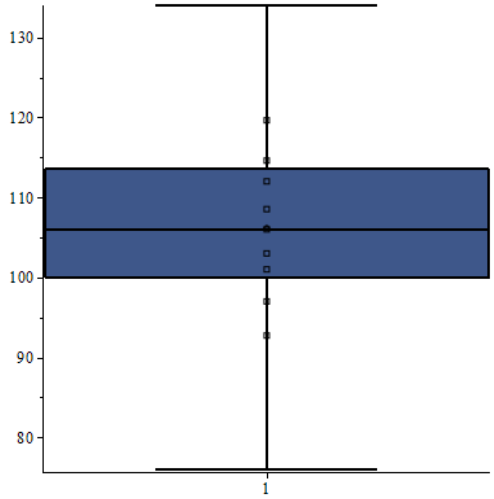
M Median

Q_3 dritter Quartilwert

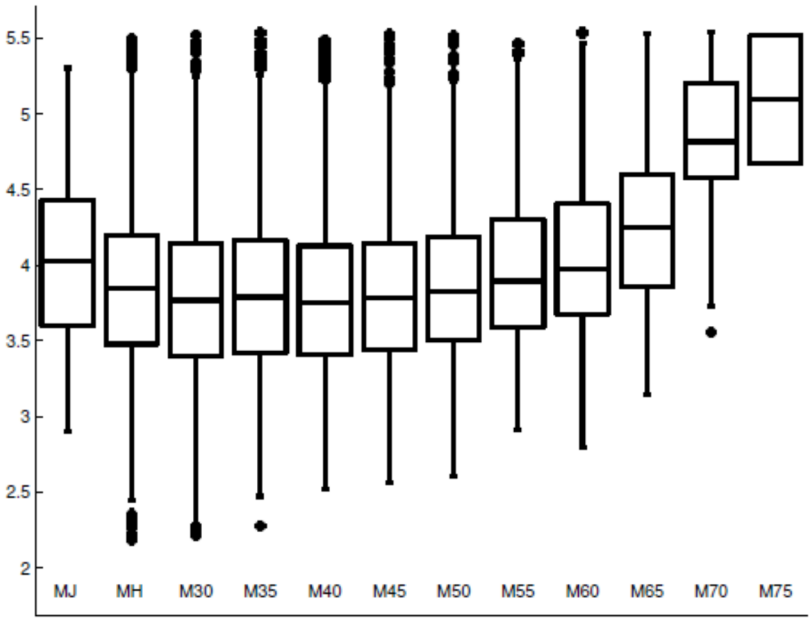
G grösster Wert, der nicht zu den Extremwerten gezählt wird

Wenn das Histogramm nicht glockenförmig ist (etwa zum Beispiel schief), reichen die Kennzahlen Mittelwert und Standardabweichung nicht aus. Dann gibt ein Boxplot besser über die Verteilung der Daten Auskunft.

Im Kükenbeispiel ergibt sich der folgende Boxplot



Boxplots sind gut geeignet, wenn viele Gruppen verglichen werden, wie etwa im folgenden Beispiel zum Hamburg-Marathon 2000.
Dargestellt sind die Netto-Laufzeiten [h] der Männer in Abhängigkeit von ihrer Altersklasse.
Der schnellste Läufer erreichte das Ziel nach 2 Stunden, 11 Minuten und 6 Sekunden, der langsamste Läufer kam nach 5 Stunden, 32 Minuten und 2 Sekunden an.
Der Median der Laufzeiten ist 3 Stunden, 52 Minuten 10 Sekunden. Interessanterweise steigt der Median mit dem Alter nicht monoton an und ist in den Gruppen M30 bis M45 nahezu konstant.
Abkürzungen:
MJ: Alter 18, 19, MH: Alter 20 – 29, M30: Alter 30 – 34, …
Quelle: Lutz Dümbgen: Einführung in die Statistik 2009



Für den Vergleich von zwei Gruppen eignen sich auch das Rücken-an-Rückendiagramm wie etwa in der abgebildeten Bevölkerungspyramide

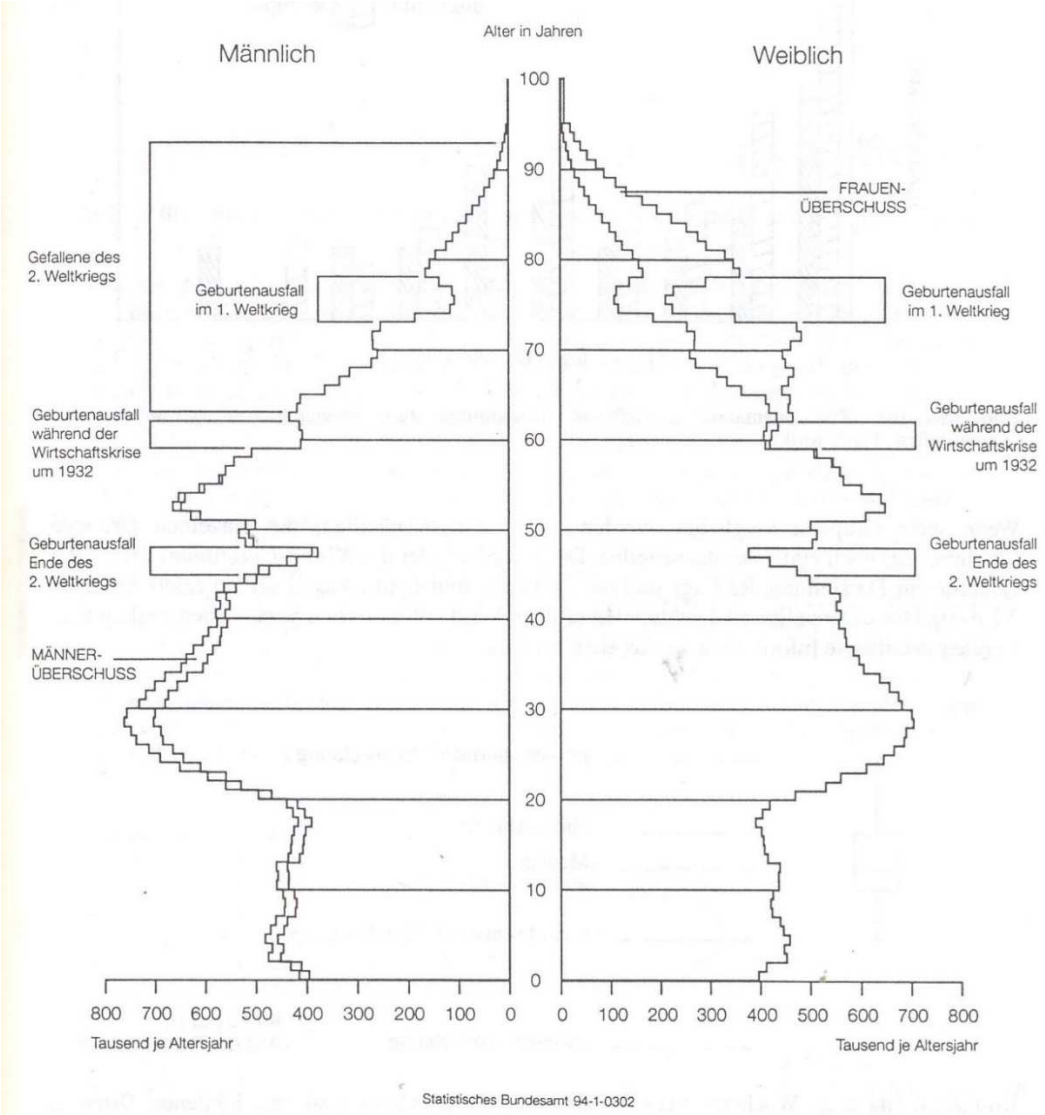


Bild 2.5.c (i) Rücken-an-Rücken-Histogramm für die Altersverteilung der deutschen Bevölkerung am 1. 1. 1993

Oder zwei ineinander gezeichnet Histogramme zum Vergleich von Hagelwolken

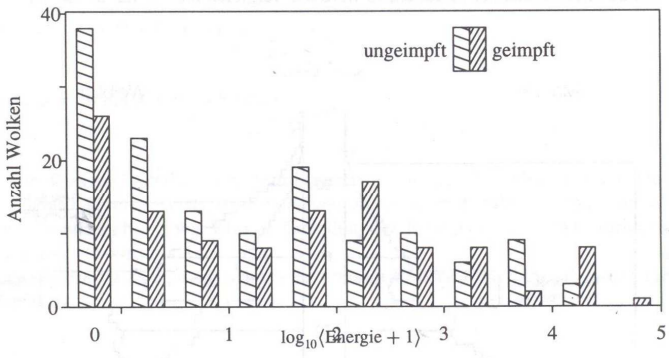


Bild 2.5.c (ii) Zwei ineinander gezeichnete Histogramme zum Vergleich von „geimpften" und „ungeimpften" Hagelwolken