Mathematik 2 - Statistik

Beschreibende Statistik

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

A - Statistische Variablen

Untersucht man eine gewisse Eigenschaft (auch Merkmal oder Variable) von Personen oder Objekten, dann erhält man **Daten**. Diese Daten werden **Werte**, oder **Ausprägungen** der entsprechenden **statistischen Variable** genannt.

Beispiel 1 Das Alter der Studenten einer gewissen Klasse der BFH sei durch die folgende Liste gegeben :

$$\{22, 22, 24, 20, 23, 25, 23, 21, 20, 22, 24, 27, 25, 23, 22\} \ .$$

Diese Zahlen sind die Werte der Variable "'Alter der Stundenten in dieser Klasse"'.

Wir gehen von einer solchen Erhebung vom Umfang n aus, und bezeichnen mit $x_1, \ldots x_n$ die Ausprägungen eines Merkmals X. Die so entstehende Liste wird **Urliste** genannt.

Variablentypen

Es gibt verschiedene Merkmaltypen : wir werden uns vor allem mit **quantitativen Variablen** befassen, d.h. Variablen, die ein Ausmass widerspiegeln, wie *Messungen* oder *Abzählungen*. Solche Variable können *diskret* oder *stetig* sein :

- Eine **diskrete Variable** kann nur endlich viele oder abzählbar unendlich viele verschiedene Werte annehmen. Sie beschreibt Daten, die im allgemeinen durch *Abzählungen* gewonnen werden.
  Beispiel : die Anzahl Personnen in jedem Auto, das während einer Stunde an eine Stelle passieren.
- Eine **stetige Variable** kann theoretisch jeden beliebigen Wert innerhalb eines vogegebenen Intervalls annehmen. Sie beschreibt Daten, die im allgemeinen durch *Messungen* gewonnen werden.
  Beispiel : Die Körpergrösse.

Variablentypen

Der Unterschied zwischen diskreten und stetigen Variablen ist manchmal willkürlich.

Tatsächlich könnte eine diskrete Variable mit einer grossen Anzahl möglichen Werten als eine stetige Variable betrachtet werden.
Beispiel : die Anzahl Einwohnern in Städten einer Region.

Im Gegenteil, für eine stetige Variable wird manchmal keine hohe Messgenauigkeit verlangt. Man wird sie dann wie eine diskrete Variable behandeln.
Beispiel : die Wohnfläche aller Wohnungen einer Stadt kann durch Abrundung als eine diskrete Variable mit ganzzahligen Werten betrachtet werden.

In solchen Zwischenfällen kann man von **quasi-stetigen** Variablen sprechen.

Daten ohne Klasseneinteilung

Sei $\{x_1, x_2, ..., x_n\}$ eine Urliste. Wir bezeichnen die verschiedenen vorkommenden Ausprängungen dieser Urliste mit $a_1, a_2, \ldots a_m$ wobei $a_1 < a_2 < \ldots < a_m$.

Die absolute bzw. relative Häufigkeiten werden folgenderweise definiert :

- Die **absolute Häufigkeit** h_k der Ausprägung a_k ist die Anzahl von Werten in der Urliste, die mit a_k übereinstimmen.
- Die **relative Häufigkeit** f_k der Ausprägung a_k ist der Anteil von Werten in der Urliste, die mit a_k übereinstimmen. Es gilt

$$f_k = \frac{h_k}{n}$$

Die Häufigkeiten $f_1, \ldots f_m$ bzw. $h_1, \ldots h_m$ fasst man in einer **Häufigkeitstabelle** zusammen.

Beispiel 2 Wir betrachten wieder das Alter der Studenten einer gewissen Klasse der HTI-Biel. Geben Sie die Häufigkeitstabelle an :

$$\{22, 22, 24, 20, 23, 25, 23, 21, 20, 22, 24, 27, 25, 23, 22\} \ .$$

Häufigkeitstabelle :

Alter (a_k)	absolute Häufigkeit h_k	relative Häufigkeit f_k
20	2	$2/15 = 13.33 \cdot 10^{-2}$
21	1	$1/15 = 6.67 \cdot 10^{-2}$
22	4	$4/15 = 26.67 \cdot 10^{-2}$
23	3	$1/15 = 6.67 \cdot 10^{-2}$
24	2	$2/15 = 13.33 \cdot 10^{-2}$
25	2	$2/15 = 13.33 \cdot 10^{-2}$
26	0	0
27	1	$1/15 = 6.67 \cdot 10^{-2}$

Daten mit Klasseneinteilung (*data binning*)

Die Komprimierung der Urliste zu einer deutlich kleineren Menge $a_1, a_2, \ldots a_n$ von Werten ist oft nicht möglich, insbesondere bei quasi-stetigen Variablen. In diesem Fall kommen viele (denkbare) Werte überhaupt nicht vor, und haben also die Häufigkeit 0. Die Idee besteht darin, die Daten in **Klassen** zu gruppieren, und eine Häufigkeitstabelle für die gruppierten Daten zu erstellen.

Die Häufigkeiten werden hier nicht einzelnen Ausprägungen, sondern Intervallen zugeordnet. Es gibt immer mehrere mögliche Einteilungen ; als Faustregel wird die Anzahl k der Klassen gemäss der empirischen Formel

$$k \approx \sqrt{n}$$

gewählt, wobei n den Stichprobenumfang bezeichnet.

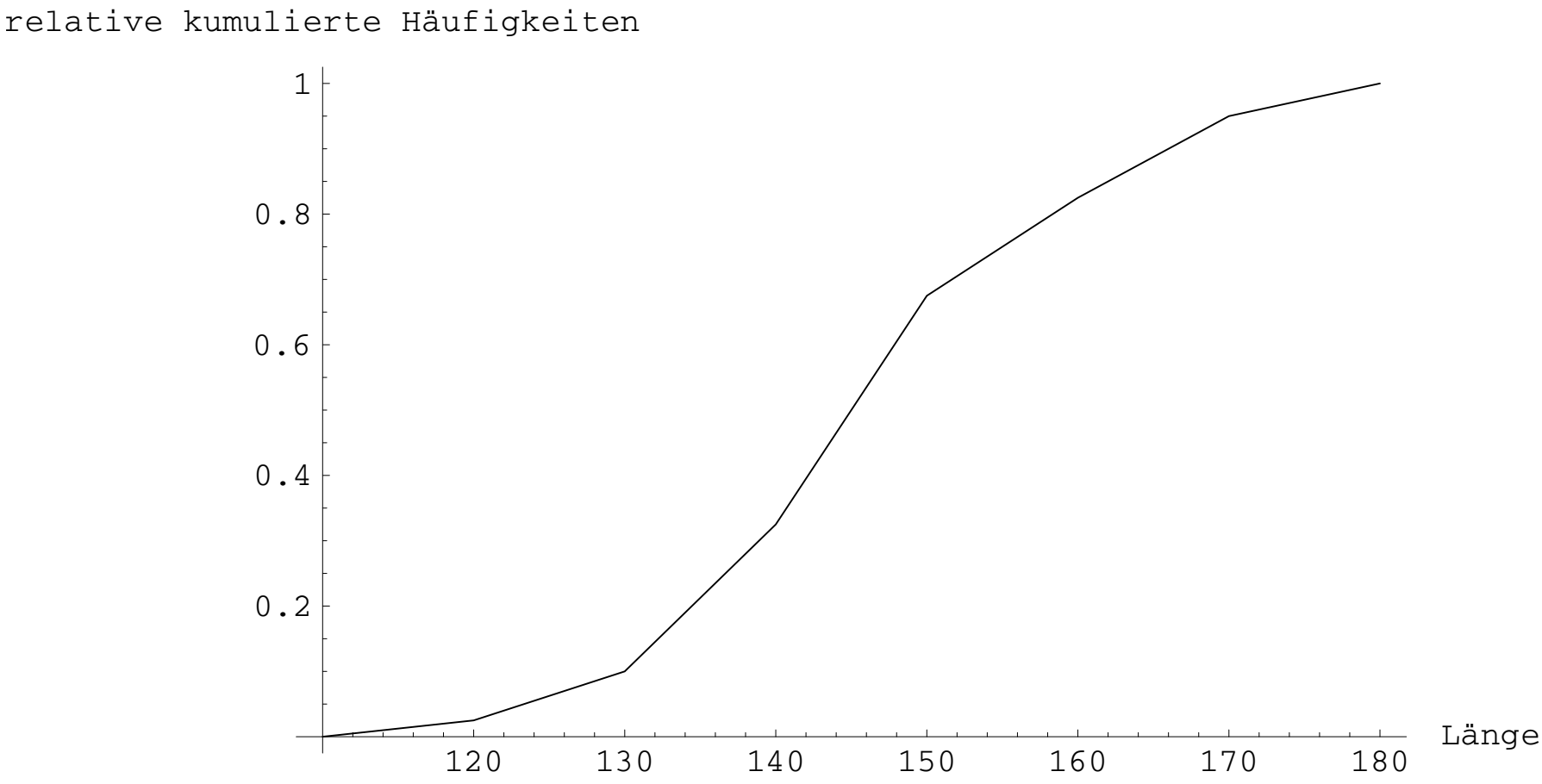
Beispiel 3 Die Länge (in mm) von vierzig Schrauben wurde gemessen. Geben Sie eine Häufigkeitstabelle mit Klasseneinteilung an.

138	164	150	132	144	125	149	157	146	158
140	147	136	148	152	144	168	126	138	176
163	119	154	165	146	173	142	147	135	153
140	135	161	145	135	142	150	156	145	128

Alle Werte liegen im Intervall zwischen 119 und 176mm, aber von den 58 möglichen Ausprägungen werden nur 30 beobachtet. Gruppiert man die Urliste in 7 Klassen, dann erhält man die Häufigkeitstabelle.

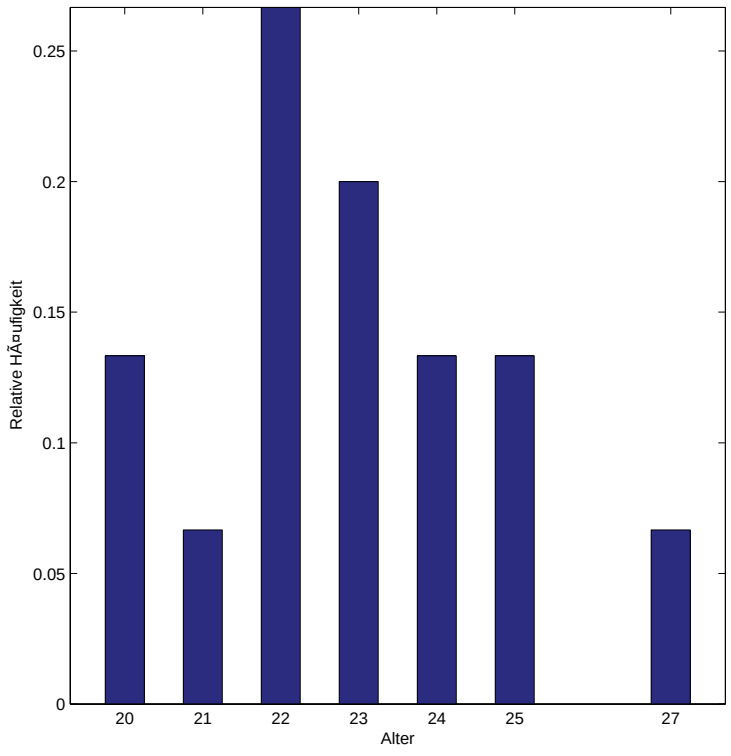
Klasse Nr	Klassen	Klassen-mitte	absolute Häufigk. h_j	relative Häufigk. f_j	kumulierte rel. Häufigkeit
1	]110,120]	115	1	1/40=0.025	1/40
2	]120,130]	125	3	3/40=0.075	4/40
3	]130,140]	135	9	9/40=0.225	13/40
4	]140,150]	145	14	14/40=0.350	27/40
5	]150,160]	155	6	6/40=0.150	33/40
6	]160,170]	165	5	5/40=0.125	38/40
7	]170,180]	175	2	2/40=0.050	1

Relative kumulierte Häufigkeiten



Stabdiagramm

Bei einem **Stabdiagramm** (auch **Säulendiagramm**) werden auf der horizontalen Achse die Ausprägungen des Merkmals abgetragen und auf der vertikalen Achse die absolute (oder die relative) Häufigkeiten der jeweiligen Ausprägungen in Form eines Stabes.
Beispiel 4 Dem Beispiel mit dem Alter entspricht das folgende Stabdiagramm :

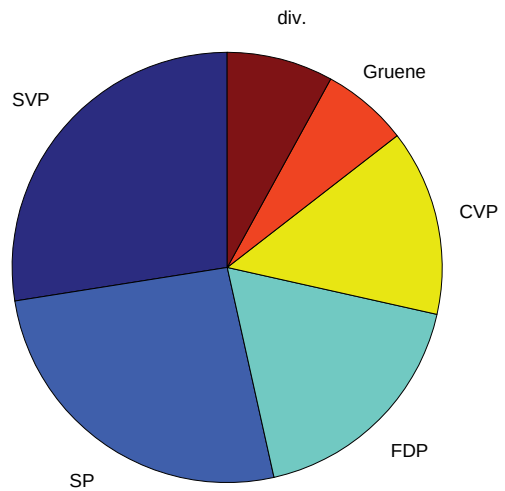


Kreisdiagramm

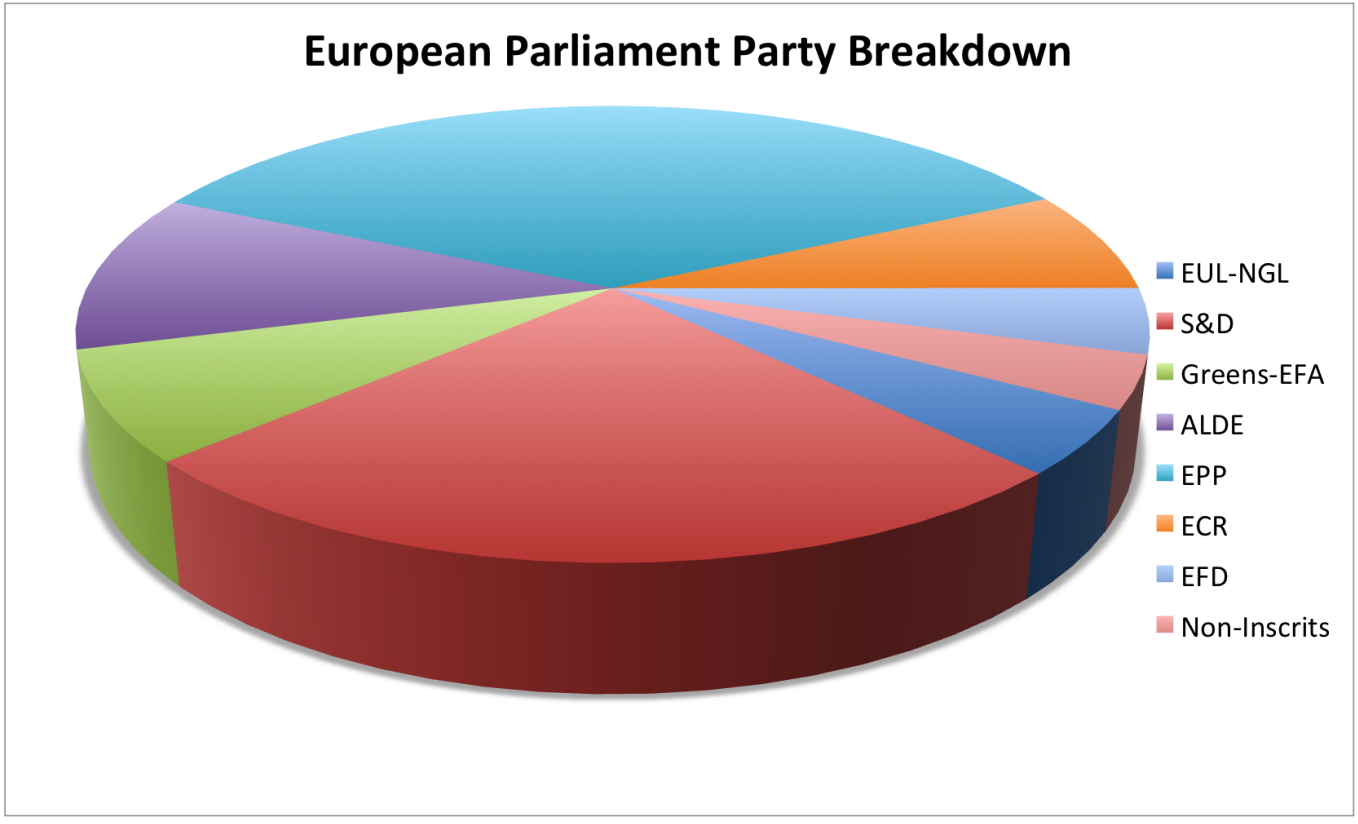
Eine weitere Darstellungsform ist das **Kreisdiagramm**, bei dem der Winkel, der den Kreisausschnitt einer Kategorie oder Ausprägung festlegt, proportional zur absoluten (oder relativen) Häufigkeit ist. Damit ist natürlich auch die Fläche des Kreissektors proportional zur Häufigkeit.
Beispiel 5 Die Parteistärke im Nationalrat sind in der folgenden Tabelle zusammengefasst (Jahr 2009) :

SVP	SP	FDP	CVP	Grüne	div.
55	52	36	28	13	16

Um diese Stärke zu vergleichen ist ein Kreisdiagram geeignet :



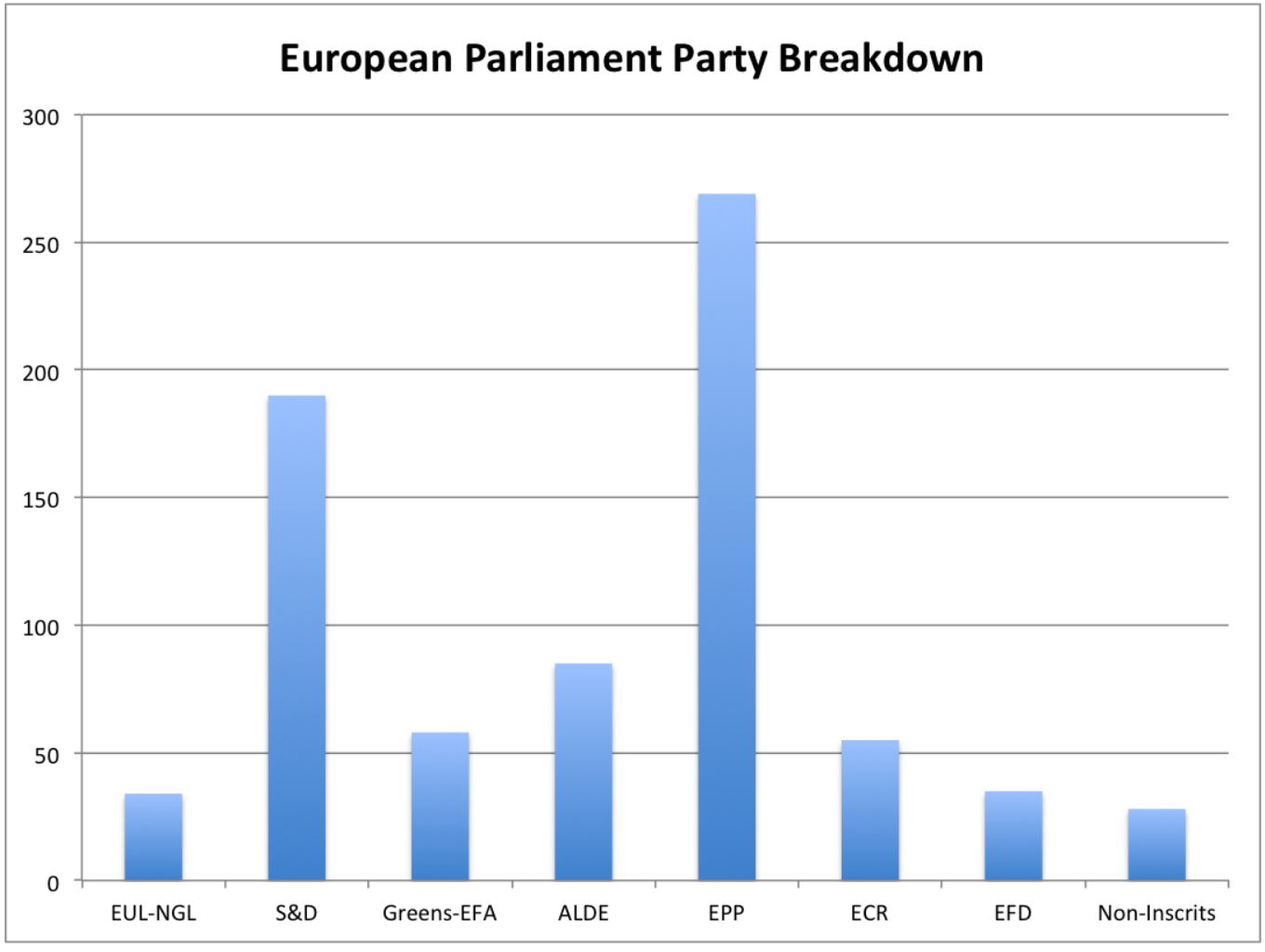
Kreisdiagramm : bitte kein 3 !



S&D $\sim$ EPP ?

Quelle businessinsider.com.au

Kreisdiagramm : bitte kein 3 !



3D ist irreführend

Quelle businessinsider.com.au

Histogramm

Für eine Variable mit vielen verschiedenen Werten kann man annehmen, dass die Daten in Klassen eingeteilt sind. Die Klassen sind benachbarte halboffene Intervalle

$$]c_0, c_1], \;]c_1, c_2], \; \ldots]c_{k-1}, c_k] \; .$$

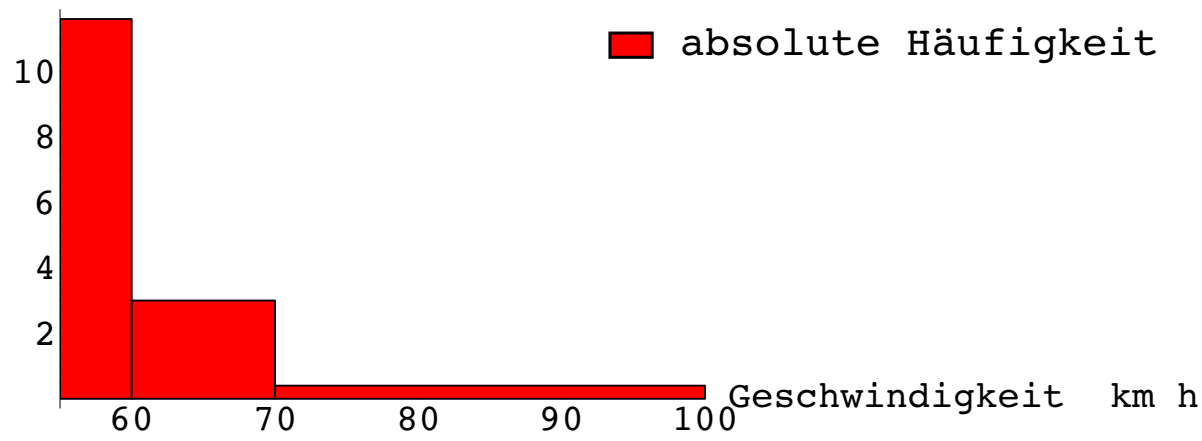
Über jedem von diesen Intervallen wird ein Rechteck so konstruiert, dass seine **Fläche** proportional zu der absoluten bzw. relativen Häufigkeit der entsprechenden Klasse ist. Die abzutragene Höhe des Rechtecks Nr j muss also gleich oder proportional zu h_j/d_j bzw. f_j/d_j gewählt werden, wobei $d_j = c_j - c_{j-1}$ die Klassenbreite bezeichnet.

Falls möglich und sinnvoll, sollten die Klassenbreiten d_j gleich gross sein. Dann kann als Höhe der Rechtecke auch die absoluten oder die relativen Häufigkeiten gewählt werden. Weiter sollten offene Randklassen vermieden werden.

Beispiel 6 Bei einer Radarkontrolle in einer Stadt wurden diejenigen Fahrzeuge registriert, welche die Geschwindigkeit von 55km/h überschritten haben. Bezüglich der Höhe des Verwarnungsgeldes wurden drei Klassen gebildet :

Klasse	Geschwindigkeit	absolute Häufigkeit
1	55 bis 60 km/h	58
2	60 bis 70 km/h	30
3	70 bis 100 km/h	12

Geben Sie das Histogramm an.
Die Klassenbreiten sind hier gegeben durch $d_1 = 5$, $d_2 = 10$ und $d_3 = 30$.
Die Höhe der Rechtecke können also folgenderweise definiert werden : $58/5 = 11.6$ (Klasse 1), $30/10 = 3$ (Klasse 2) und $12/30 = 0.4$ (Klasse 3).

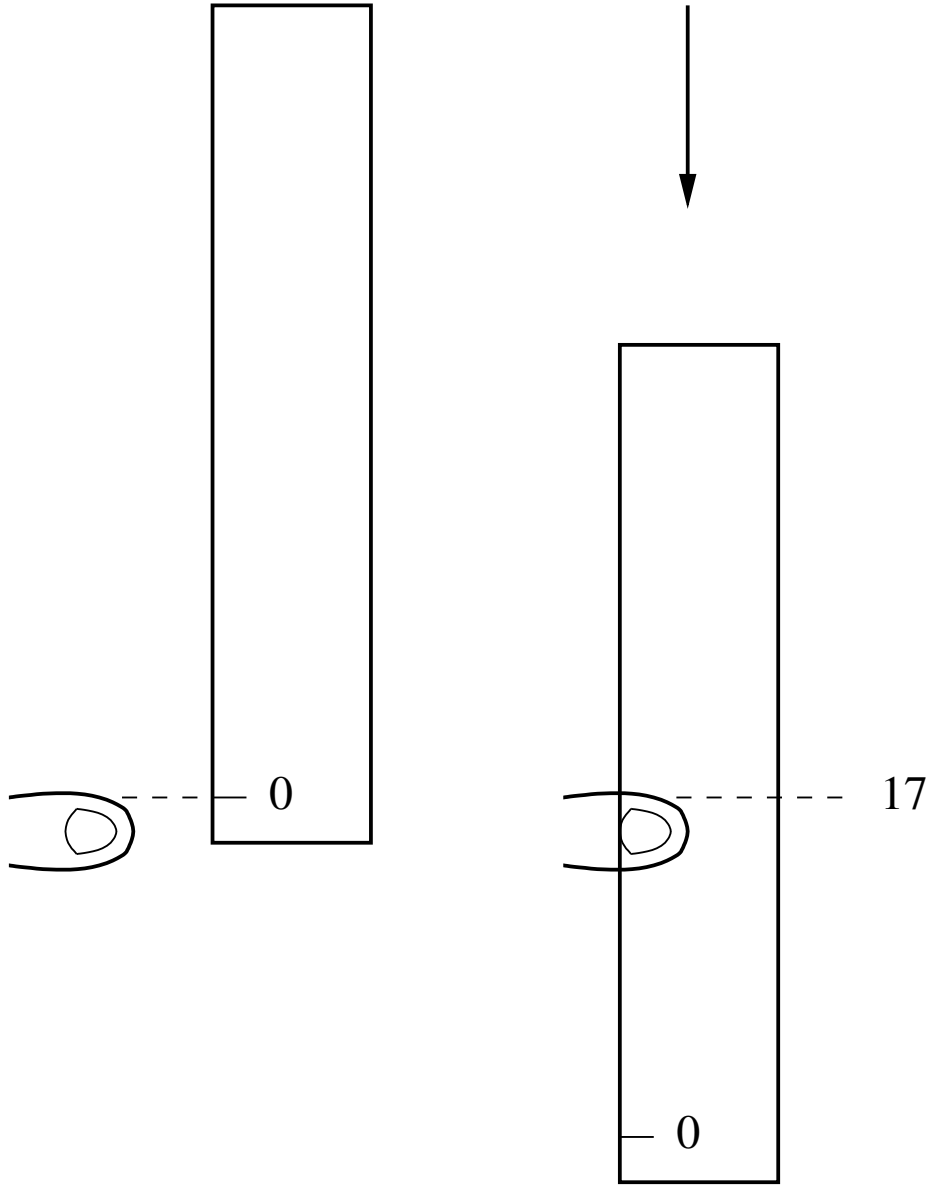


Ein wenig Übung !

Frage :

Die typische Greifzeit eines Objektes bestimmen.

Experimentelles Protokoll



Ergebnisse

Messungen

22 16 13 na 23 ... 20

Stichprobe

x_1 x_2 x_3 x_4 x_5 ... x_n

Fragen

1. Wie kann man diese Daten grafisch darstellen ?
2. Wie bestimmt man die "typische" Reaktionszeit ?

Elemente

Mittelwert :

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

Median : mit $x_{(i)}$ für die Daten in aufsteigender Reihenfolge :

$$x_{1/2} = \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & n \text{ gerade} \end{cases}$$

Zeit Umrechnung :

$$h = \frac{1}{2}gt^2 \qquad\qquad t = \sqrt{\frac{2h}{g}}$$

t	Reaktionszeit [s]
h	Messung [m]
g	Erdbeschleunigung [m/s^2]

Relative kumulierte Häufigkeiten

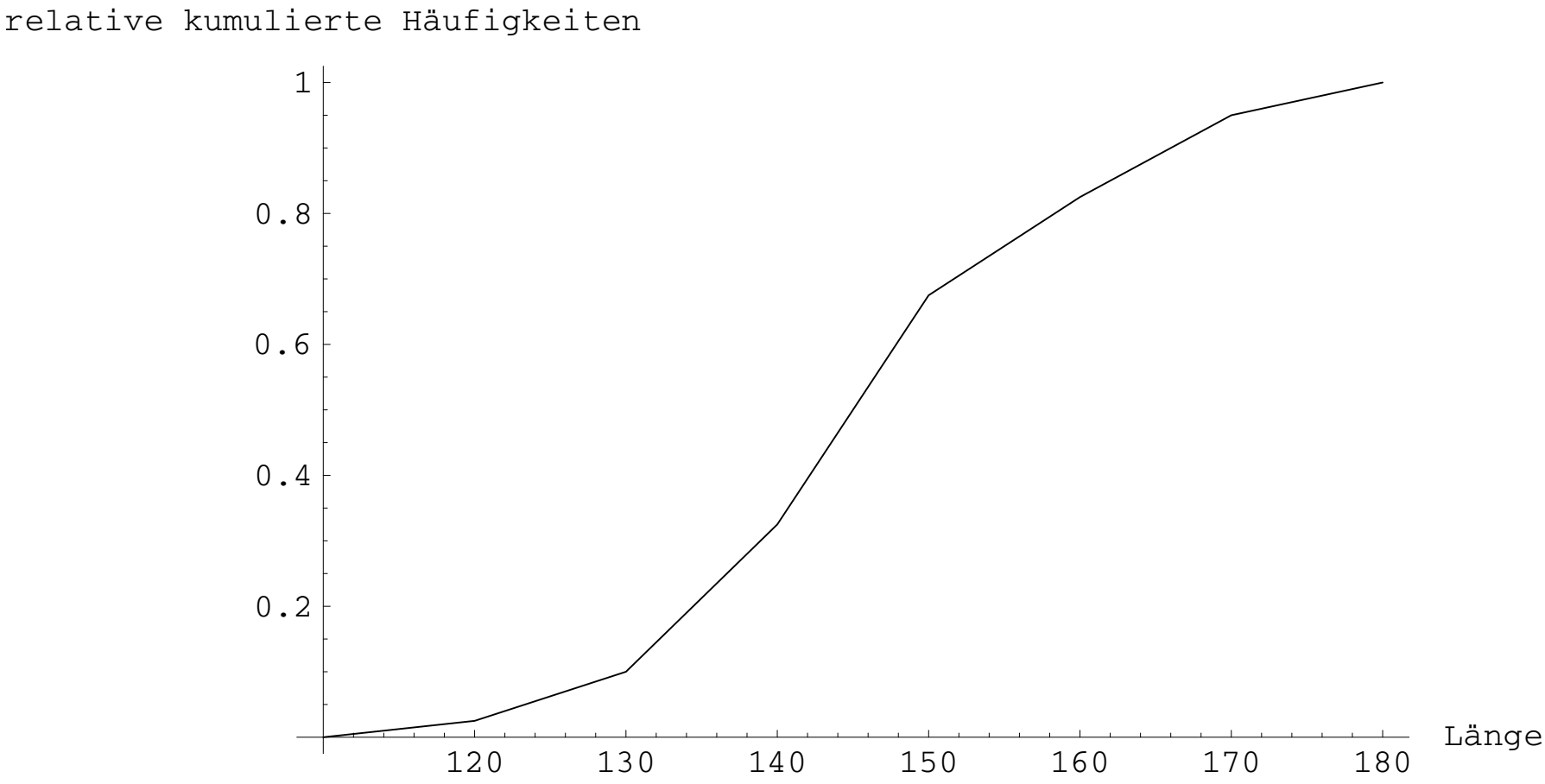
für Daten, die in Klassen gruppiert sind

138	164	150	132	144	125	149	157	146	158
140	147	136	148	152	144	168	126	138	176
163	119	154	165	146	173	142	147	135	153
140	135	161	145	135	142	150	156	145	128

Klasse Nr	Klassen	Klassen-mitte	absolute Häufigk. h_j	relative Häufigk. f_j	kumulierte rel. Häufigkeit
1	]110,120]	115	1	1/40=0.025	1/40
2	]120,130]	125	3	3/40=0.075	4/40
3	]130,140]	135	9	9/40=0.225	13/40
4	]140,150]	145	14	14/40=0.350	27/40
5	]150,160]	155	6	6/40=0.150	33/40
6	]160,170]	165	5	5/40=0.125	38/40
7	]170,180]	175	2	2/40=0.050	1

Relative kumulierte Häufigkeiten

für Daten, die in Klassen gruppiert sind



Matlab/Octave : `plot`

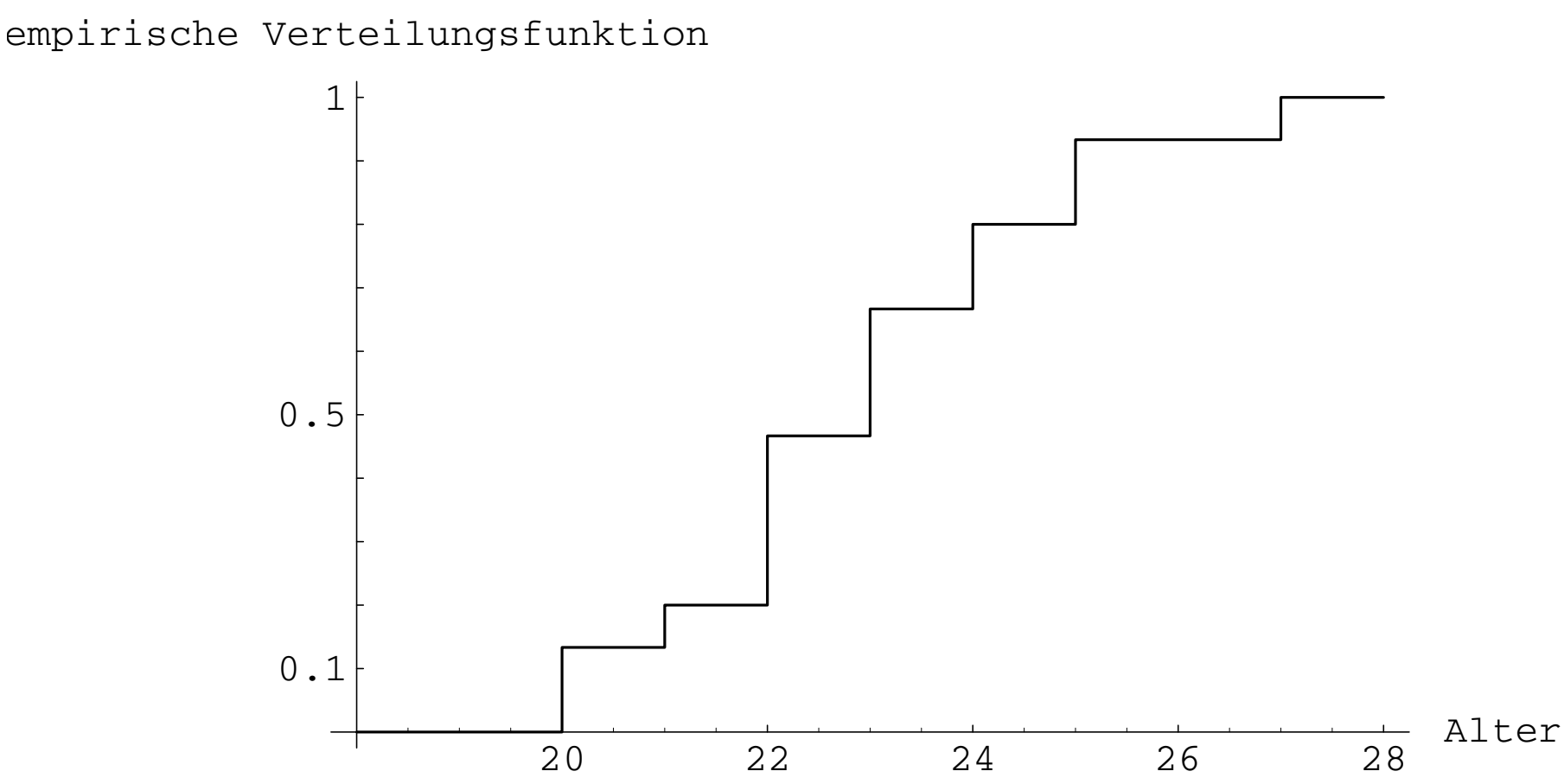
Relative kumulierte Häufigkeiten

für Daten ohne Gruppierung

Alter a_i	h_i	f_i	$H(a_i)$	$F(a_i)$
20	2	$2/15$=0.133	2	$2/15$=0.133
21	1	$1/15$=0.067	3	$3/15$=0.200
22	4	$4/15$=0.267	7	$7/15$=0.467
23	3	$3/15$=0.200	10	$10/15$=0.667
24	2	$2/15$=0.133	12	$12/15$=0.800
25	2	$2/15$=0.133	14	$14/15$=0.933
26	0	$0/15$=0.000	14	$14/15$=0.933
27	1	$1/15$=0.067	15	$15/15$=1.000

Relative kumulierte Häufigkeiten

für Daten ohne Gruppierung

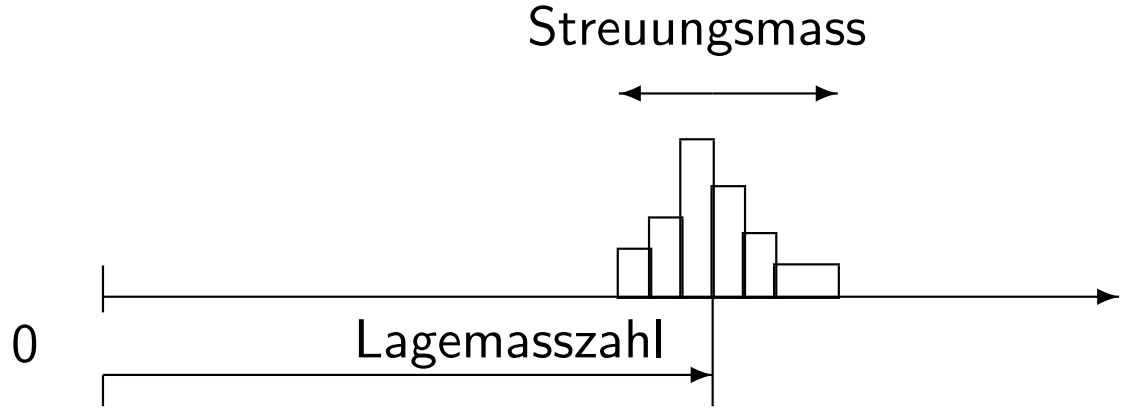


Matlab/Octave : `ecdf` / `empirical_cdf` , `stairs`

Masszahlen

Die graphische Darstellung einer Häufigkeitsverteilung ist zwar nützlich, sie bringt aber keine Antwort zu den folgenden Fragen : wo liegt das Zentrum der Daten ? Wie breit sind die Daten um das Zentrum verteilt ?

Zu diesem Zweck müssen die Daten auf *Masszahlen*, oder *Parameter*, reduziert werden. Man unterscheidet zwischen **Lagemasszahlen**, die angeben, wie gross die Stichprobenwerte etwa sind, und **Streuungsmasse**, die besagen wie breit die Daten gestreut sind.



Der arithmetische Mittelwert

Der **arithmetische Mittelwert** von n Zahlen $x_1, x_2, \ldots, x_n$ ist

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

Für Häufigkeitsdaten mit Ausprägungen $a_1, \ldots a_k$ und relativen Häufigkeiten $f_1, \ldots f_k$ gilt

$$\bar{x} := \sum_{i=1}^{k} a_i \cdot f_i$$

Man spricht dann von einem **gewogenen** oder **gewichteten** arithmetischen Mittelwert.

Sind die Daten bereits in Klassen gruppiert, so werden die Ausprägungen durch die Klassenmitten c_i ersetzt.

$$\bar{x} \approx \sum_{i=1}^{k} c_i \cdot f_i \,.$$

Beispiel 7 Berechnen Sie den arithmetischen Mittelwert im Beispiel der Radarkontrolle :

Klasse	Geschwindigkeit	absolute Häufigkeit
1	55 bis 60 km/h	58
2	60 bis 70 km/h	30
3	70 bis 100 km/h	12

$$\bar{v} \approx 0.58 \cdot 57.5 + 0.30 \cdot 65 + 0.12 \cdot 85 = 63.05 \text{km/h}\,.$$

Der Median oder Zentralwert

In gewissen Fällen ist der arithmetische Mittelwert nicht die beste Masszahl, um das Zentrum der Daten zu beschreiben. Dies wird anhand des folgenden Beispiels veranschaulicht.

Beispiel 8 Die Löhne einer Stichprobe von Mitarbeitern seien in der folgenden Tabelle gegeben :

Lohn (CHF)	4300	5100	6200	6750	9300	12800	18600

Als mittlerer Lohn erhalten wir 9007 CHF. Dieser Wert ist offensichtlich wegen der einigen grossen Ausprägungen nach oben gezogen. Wird der grösste Lohn weiter auf CHF 21000 ändert, dann verändert sich der mittlere Lohn zu 9350 CHF.

Der arithmetsiche Mittelwert reagiert also empfindlich auf extreme Werte (Ausreisser). Dies kann unerwünscht sein, beispielsweise wenn solche Ausreisser durch Fehler bei der Datenerhebung oder Messung verursacht wurden.

Der **Median** wird so definiert, dass er die Stichprobe *in zwei gleiche Hälfte teilt*. Sind die Daten einer Stichprobe der Grösse nach geordnet, so ist der Median, geschrieben $\tilde{x}$ oder $x_{1/2}$ gegeben durch :

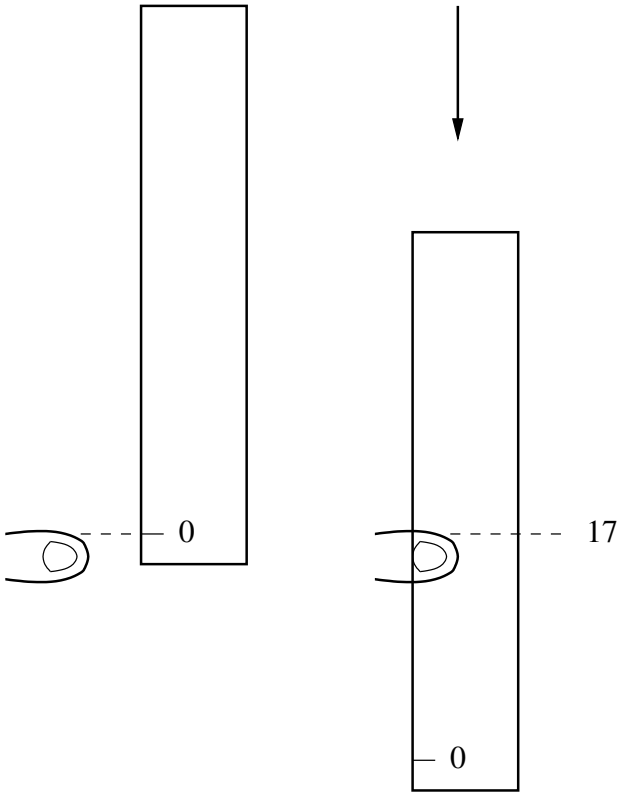
$$x_{1/2} := \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ impair} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & n \text{ pair} \end{cases}$$

Beispiel 9 Geben Sie den Median für die folgenden Daten an :

Lohn (CHF)	4300	5100	6200	6750	9300	12800	18600

Falls die Daten in Klassen gruppiert sind, muss man zuerst die *Einfallsklasse* bestimmen, d.h. die Klasse in welcher die Folge der relativen kumulierten Häufigkeiten $F(c_i)$ erstmals den Wert 0.5 überschreitet. Da die Urliste nicht mehr zu Verfügung steht, nimmt man an, dass die Daten in der Einfallsklasse gleichmässig verteilt sind.

Mittelwert vs. Median

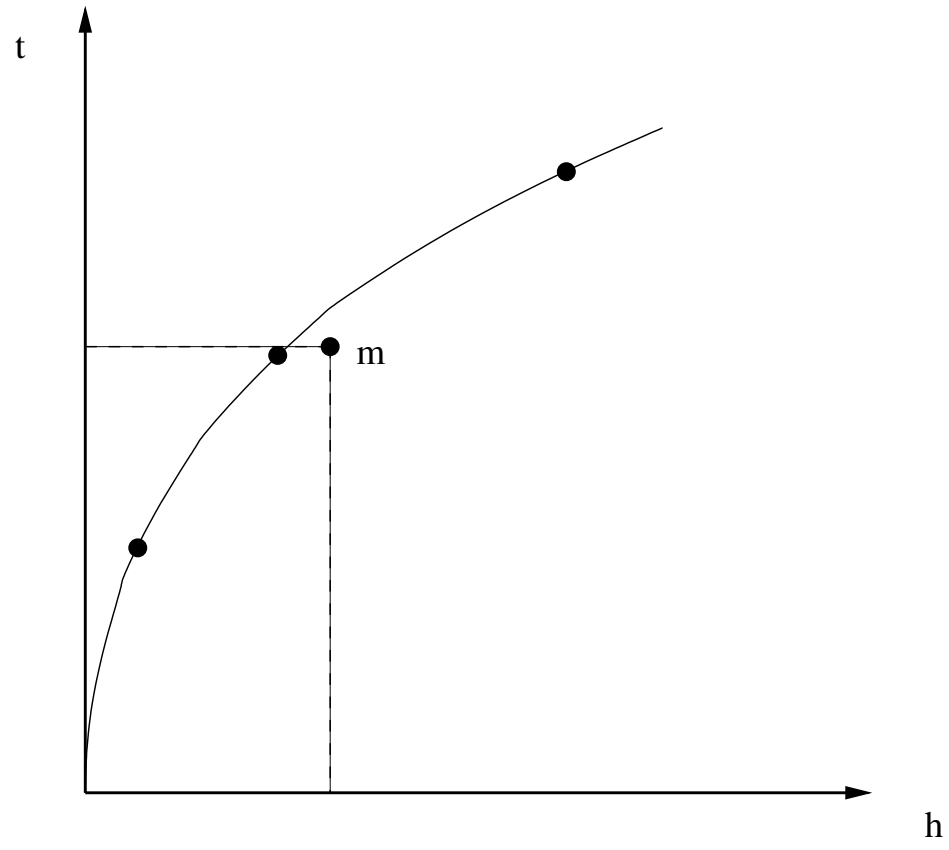


Zeit Umrechnung :

$$h = \frac{1}{2} g t^2 \qquad\qquad t = \sqrt{\frac{2h}{g}}$$

t	Reaktionszeit	[s]
h	Messung	[m]
g	Erdbeschleunigung	[m/s^2]

Ordnung der Verknüpfungen



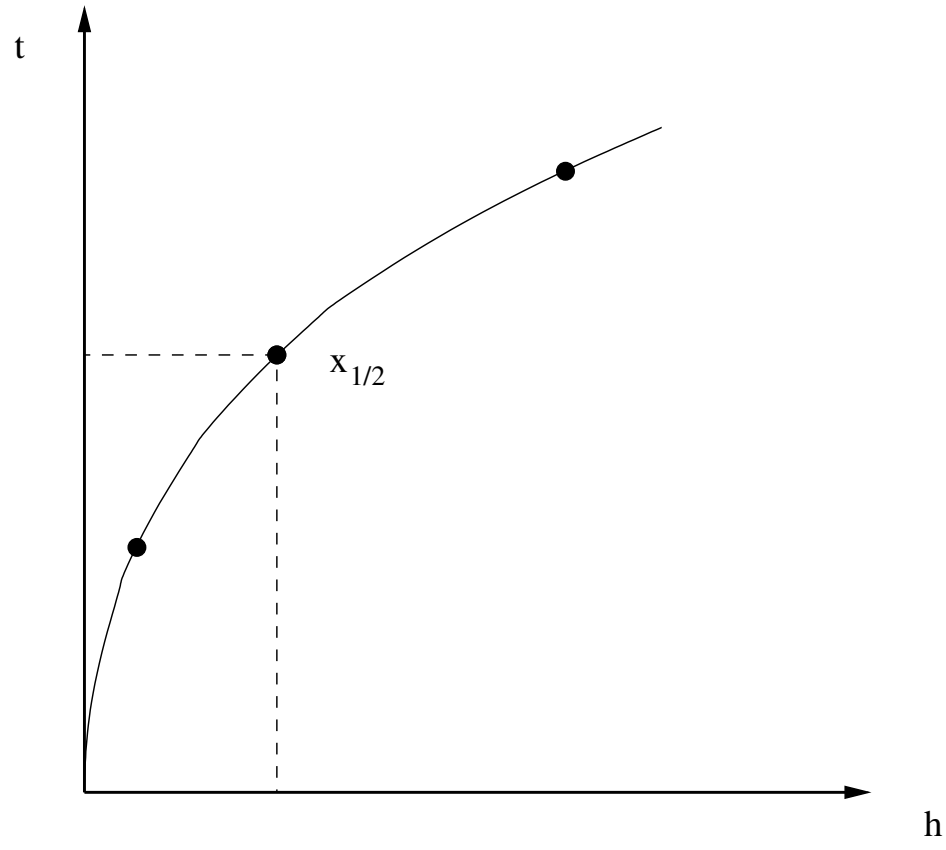
$$h = \frac{1}{2} g t^2 \qquad\qquad t = \sqrt{\frac{2h}{g}}$$

Der Mittelwert von t entspricht nicht dem Mittelwert von h !

Vorteile des Medians

1. "Robust" : nicht sehr empfindlich gegenüber Extremwerten (na...)

2. Der Median von t entspricht (*fast*) dem Median von h.



Quantil und Box-Plot

Sei $0 < p < 1$ eine Zahl. Das $p-$*Quantil* einer Verteilung wird so definiert, dass es die Daten in zwei Teile trennt und zwar so, dass etwa $p \cdot 100\%$ der Daten darunter und $(1-p) \cdot 100\%$ Prozent darüber liegen.

Wir bezeichnen die der Grösse nach geordnete Liste der Daten mit Klammern

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Das $p-$**Quantil** x_p der Stichpobe ist definiert durch

$$x_p := \begin{cases} x_{(\lfloor n \cdot p + 1 \rfloor)} & n \cdot p \notin \mathbb{Z} \\ \frac{1}{2}\left[x_{(n \cdot p)} + x_{(n \cdot p + 1)}\right] & n \cdot p \in \mathbb{Z} \end{cases}$$

wobei $\lfloor n \rfloor$ die grösste natürliche Zahl bezeichnet, welche kleiner oder gleich n ist :

$$\lfloor n \rfloor = \max\{k \in \mathbb{Z} : k \leq n\} .$$

Es folgt unmittelbar aus dieser Definition, dass mindestens $p \cdot 100\%$ (bzw. $(1-p) \cdot 100\%$) aller Stichprobenwerte kleiner oder gleich (bzw. grösser oder gleich) x_p sind. Neben dem Median als $0.5-$Quantil besitzen auch weitere häufig verwendete Quantile eigene Namen. So heissen $x_{0.25}$ und $x_{0.75}$ das **untere** bzw. **obere Quartil**, und $x_{j \cdot 0.1}$ das $j-$te Dezil $(j = 1, ..., 9)$.

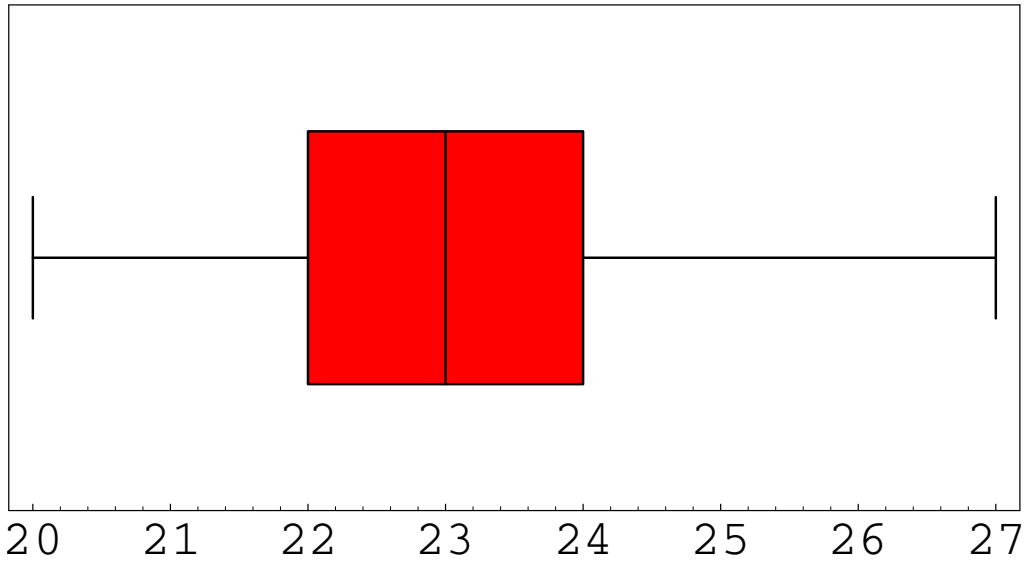
Zusammen mit dem kleinsten und grössten Wert geben die Quartile und der Median Informationen über die Verteilung der Daten. Diese Beschreibung, bestehend aus den Werten

$$x_{(1)},\ x_{0.25},\ x_{0.5},\ x_{0.75},\ x_{(n)},$$

heisst **Fünf-Punkte-Zusammenfassung** der Verteilung.

Diese Zusammenfassung führt zur komprimierten Visualisierung einer Verteilung durch den sogenannten **Box-Plot**. Der Box-Plot besteht aus einem Schachtel (*"'Box"'*), mit Anfang bzw. Ende bei dem unteren bzw. oberen Quartil. In der Box wird der Median durch einen vertikalen Strich dargestellt. Weiter gehen zwei Linien (*"'Whiskers"'*) aus der Box bis zu $x_{(1)}$ und $x_{(n)}$.

Beispiel 10 Die Daten im Beispiel des Alters werden folgenderweise durch den Box-Plot zusammengefasst :



Verschiedene Variante des Box-Plots stehen zu Verfügung !

Streumasse

Die Lagemasszahlen sagen nur, wie gross die Stichprobenwerte im Durchschnitt sind ; sie geben aber keine Information über die Streuung dieser Daten um das Zentrum.
Zum Beispiel besitzen die Stichproben $\{-1, 0, 1\}$ und $\{-10, 0, 10\}$ den gleichen arithmetischen Mittelwert ; die Werte der zweiten Stichprobe sind aber offensichtlich breiter verteilt als die Werte der ersten Stichprobe.
Ein **Streuungsmass** ist ein Mass, das beschreibt, wie breit die Stichprobenwerte um den Mittelwert verteilt sind.

Die **Varianz** von n Zahlen $x_1, x_2, \ldots, x_n$ ist der "Mittelwert" der quadratischen Abweichungen :

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

und die **Standardabweichnung** ist die Quadratwurzel aus der Varianz, nämlich

$$s := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \, .$$

Beispiel 11 Berechnen Sie die Varianz und die Standardabweichnung für das Beispiel des Alters.

Für Häufigkeitsdaten mit Ausprägungen $a_1, \ldots a_k$ und relativen Häufigkeiten $f_1, \ldots f_k$ wird die Varianz s^2 durch den folgenden gewichteten Mittelwert erhalten :

$$s^2 = \frac{n}{n-1} \sum_{i=1}^{n} (a_i - \bar{x})^2 f_i \ .$$

Sind die Daten bereits in Klassen gruppiert, so werden die Ausprägungen durch die Klassenmitten c_i ersetzt, was uns zu einem *Näherungswert* führt :

$$s^2 \approx \frac{n}{n-1} \sum_{i=1}^{k} (c_i - \bar{x})^2 f_i$$

Beispiel 12 Berechnen Sie die Standardabweichnung im Beispiel der Radarkontrolle :

Klasse	Geschwindigkeit	absolute Häufigkeit
1	55 bis 60 km/h	58
2	60 bis 70 km/h	30
3	70 bis 100 km/h	12