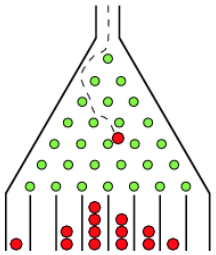
Deskriptive Statistik

KAD 2013.09.12



Die Statistik beschäftigt sich mit **Massenerscheinungen**, bei denen die dahinterstehenden Einzelereignisse meist zufällig sind.

Statistik benutzt die Methoden der Wahrscheinlichkeitsrechnung.

Fundamentalregeln:

Statistischen Aussagen beziehen sich nie auf ein Einzelereignis, sondern nur **auf Gesamtheiten vieler Ereignisse.**

Jede statistische Aussage ist mit einer **prinzipiell unvermeidlichen Unsicherheit** behaftet.

Wozu braucht eine Ärztin / ein Arzt Statistik?

- zum Verstehen der medizinischen Fachliteratur („How to Read a Paper“) insbesondere von Originalarbeiten in Fachzeitschriften über
  - experimentelle
  - klinische
  - epidemiologische
  - sonstige (z. B. gesundheitsökonomische) Studien
- „Evidence-based Medicine“ Bewertung und Kommunikation von Chancen und Risiken
- bei eigenen Untersuchungen
  - Doktorarbeit
  - Industrie
  - Gesundheitsbehörden

http://www.medizin.uni-koeln.de/kai/imsie/kursinfo/q1/Q1-01-Einfuehrung.pdf

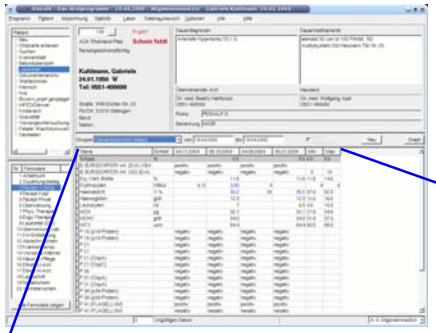
das erste Anwendungsgebiet der Statistik bestand in der **Staatsbeschreibung** (Völkszählung) *Status* = Zustand

Semmelweis (1818-1865) war der erste bekannte Arzt, der den Nutzen einer neuen Therapie **mit statistischen Methoden** belegte

Was messen Physiker, Arzt und Medizinstudent?

WER MISST WAS?		
PHYSIKER	ARZT	MEDIZINSTUDENT IM PHYSIKPRAKTIKUM
Länge	Körpergröße	Durchmesser von Erythrozyten (3)
Frequenz	Pulsfrequens	Impulshäufigkeit (9,20)
Temperatur	Körpertemperatur	—
Konzentration	Blutzuckerspiegel	Eiweißkonzentration im Blutplasma (5)
Spannung	EKG-Signal	EKG-Signal (24)
Leistungsdichte	Hörschwelle	Hörschwelle (22)
Druck	Blutdruck	—
Impedanz	Hautimpedanz (Hautwiderstand)	Hautimpedanz (21)

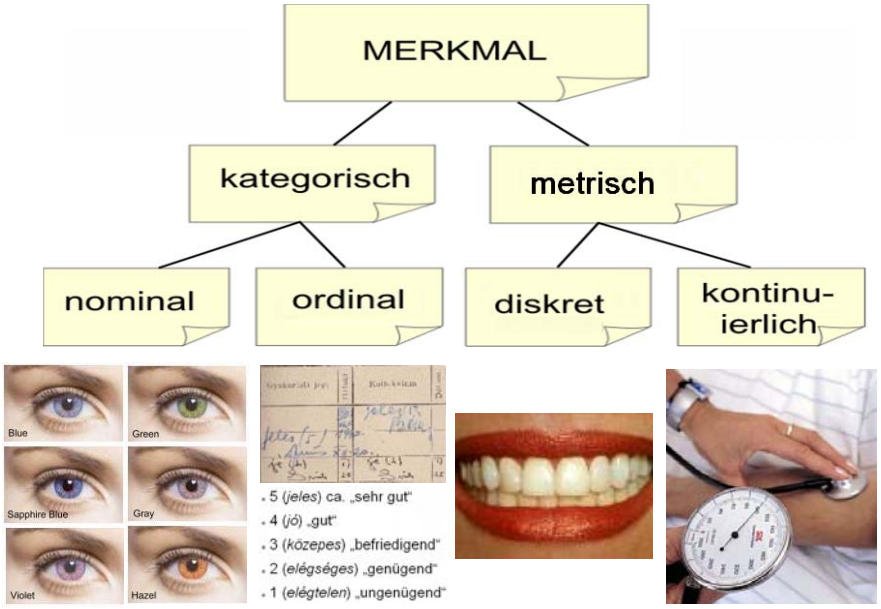
Pr.Buch Tabelle 3



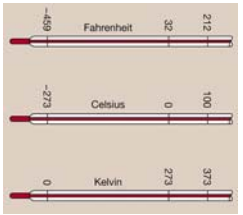
Labormessergebnisse

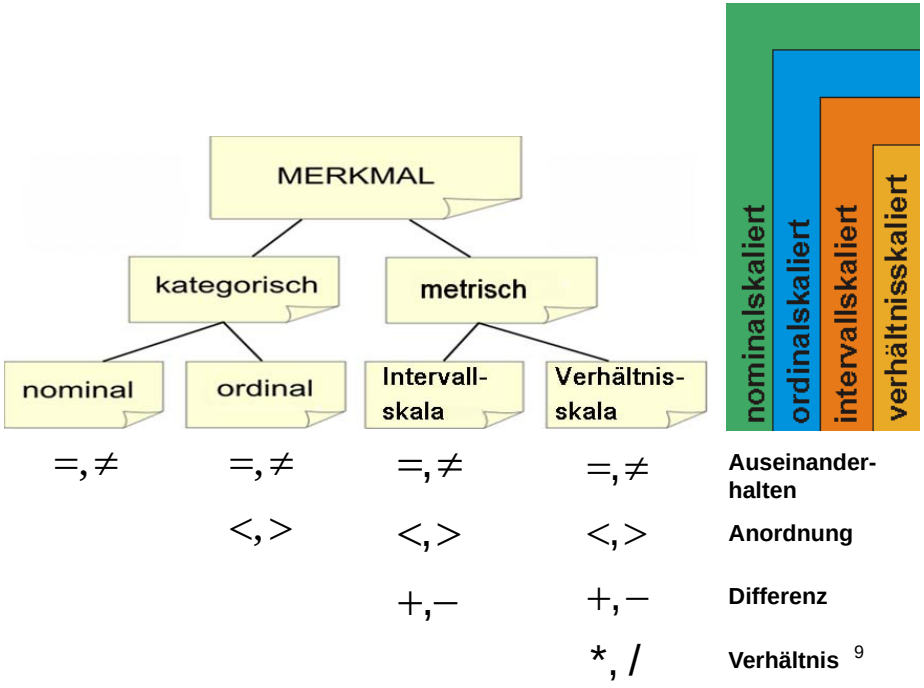
Name	Einheit	04.11.2004	05.10.2004	04.08.2004	05.07.2004	Min	Max
%Hypo	%		0.5		0.5	0.0	5.0
B. BURGDORFERI-AK (EIA) IGM		positiv	positiv	positiv	positiv		
B. BURGDORFERI-AK IGG (EIA)		negativ	negativ	negativ	negativ	5	10
Ery.-Vert.-Breite	%		11.6		11.6	11.5	14.5
Erythrozyten	Mill/ul	4,12	3,95	4		4	6
Haematokrit	V %		36.2	36	36.2	37.0	52.0
Haemoglobin	g/dl		12.3		12.3	12.0	16.0
Leukozyten	/ul		7		6.5	4.0	10.0
MCH	pg		32.1		32.1	27.0	34.0
MCHC	g/dl		34.0		34.0	31.0	37.0
MCV	ucm		94.4		94.4	80.0	99.0
P 18 (p18-Protein)		negativ	negativ	negativ	negativ		

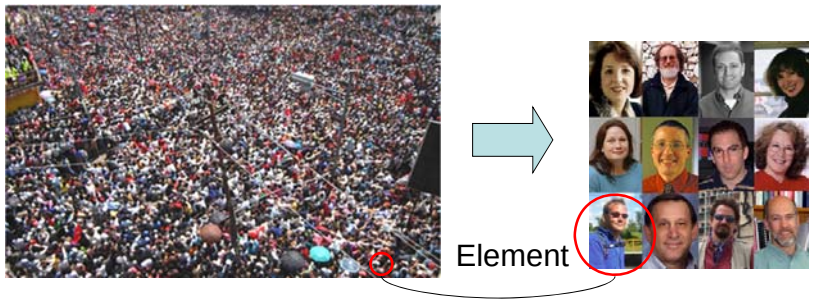
Klassifizierung der Merkmale



Skalentypen der metrischen Merkmale

	diskret	kontinuierlich
Intervall-skala definierte Differenz, „kein" 0 Punkt	Tage in einem Kalender 	Tempe-ratur in °C 
Verhältnis-skala definiertes Verhältnis, 0 Punkt	Anzahl der Zähne 	Tempe-ratur in K 





Grundgesamtheit (Population):

Gesamtheit der Individuen (Elemente), deren Eigenschaften bei der Studie untersucht werden sollen. Die gesamte Menge der interessierenden Daten.

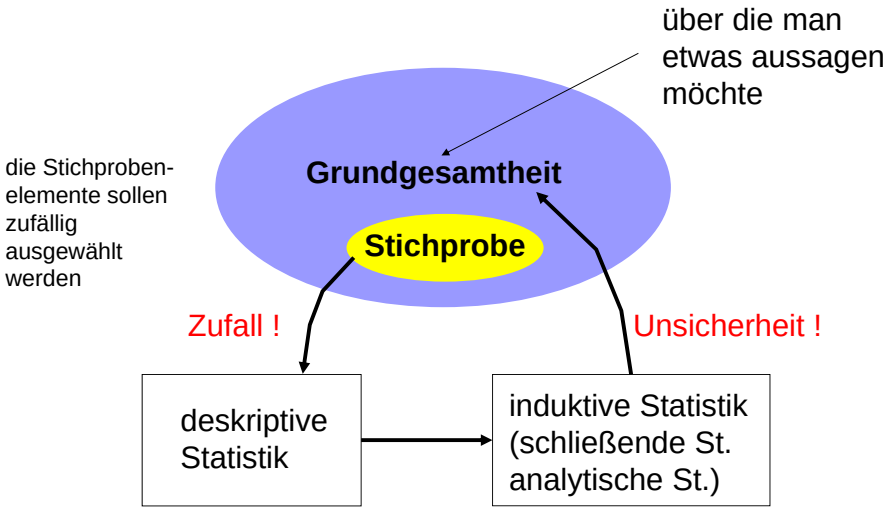
N = „unendlich"

Stichprobe:

Der für die Studie ausgewählte Teil der Population.

n = endlich

N >>*n* (Umfang)



Die deskriptive Statistik ist die Vorstufe zur induktiven Statistik

Wie hoch ist die normale Pulsfrequenz (einer Population)?

Merkmal: Pulsfrequenz

zufällige Erhebung einiger Elementen der Population: **Stichprobe**

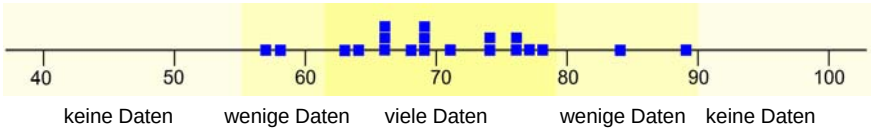
Daten der Stichprobe liegen in Form einer Urliste vor:

66, 56, 89, 63, 66, 69, 71, 68, 58, 69, 78, 66, 64, 84, 74, 76, 69, 77, 74, 76 (Einheit: 1/Min), oder:

66	56	89	63	66	69	71	68	58	69
78	66	64	84	74	76	69	77	74	76

„Die Werte sollen **geordnet** und **verdichtet** werden." !?

Stellen wir die Daten entlang einer Zahlengeraden dar!



Pr.Buch Abb. 4

Verfeinern wir die Klassen noch weiter!

Unterteilen wir die Zahlengerade in gleich breite Klassen (Intervalle) und zählen wir ab, wie viele Daten sich in den so erhaltenen **Klassen** befinden!

KLASSENGRENZEN	HÄUFIGKEIT
$55 \le x_i < 60$	2
$60 \le x_i < 65$	2
$65 \le x_i < 70$	7
$70 \le x_i < 75$	3
$75 \le x_i < 80$	4
$80 \le x_i < 85$	1
$85 \le x_i < 90$	1
insgesamt:	$n = 20$

in Excel:

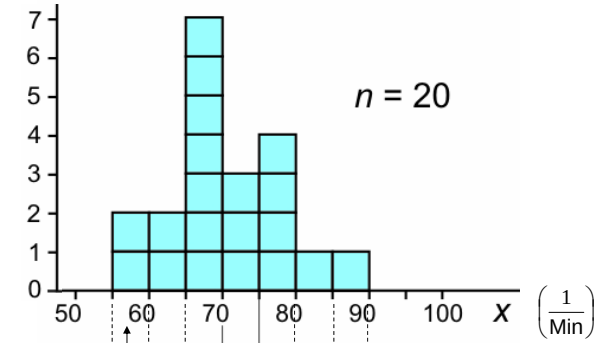
=frequency(...)
=Häufigkeit(...)

Die Grenzwerte und die Breiten der Klassen sind willkürlich. Stellen wir diese Treppenfunktion dar!

Pr.Buch Tabelle 5

Häufigkeitsdichte

$$\frac{\Delta n}{\Delta x} \quad \left(\frac{1}{5\frac{1}{\text{Min}}} = \left(\frac{\text{Min}}{5}\right)\right)$$



Die Fläche unter der Treppenfunktion zwischen 55 und 60:

$$5\frac{1}{\text{Min}} \cdot 2\frac{\text{Min}}{5} = 2$$

Die Gesamtfläche unter der Treppenfunktion: 20 = *n*,

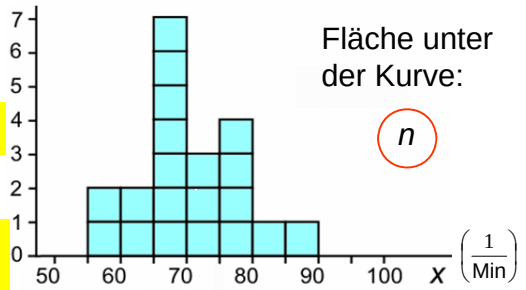
Anzahl der Messdaten in der Stichprobe

KLASSENGRENZEN	HÄUFIGKEIT
$55 \le x_i < 60$	2
$60 \le x_i < 65$	2
$65 \le x_i < 70$	7
$70 \le x_i < 75$	3
$75 \le x_i < 80$	4
$80 \le x_i < 85$	1
$85 \le x_i < 90$	1
insgesamt:	$n = 20$

Häufigkeitsdichte-verteilung

absolute

$$\frac{\Delta n}{\Delta x} \quad \left(\frac{\text{Min}}{5}\right)$$

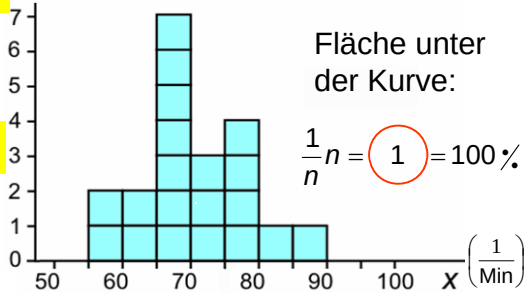


Fläche unter der Kurve:

n

relative

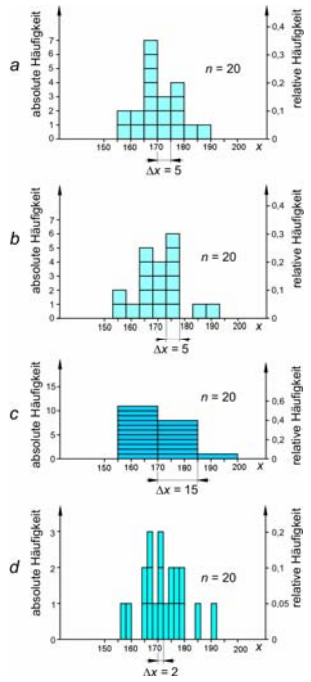
$$\frac{1}{n}\frac{\Delta n}{\Delta x} \quad \left(\frac{1}{20}\frac{\text{Min}}{5}\right)$$



Fläche unter der Kurve:

$$\frac{1}{n}n = 1 = 100\,\%$$

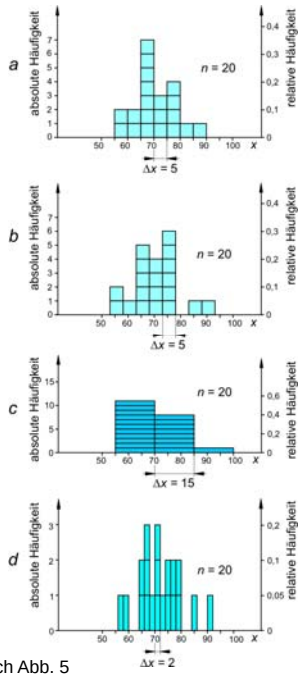
absolute Häufigkeits-dichte (Histogramm)



relative Häufigkeits-dichte (Histogramm)

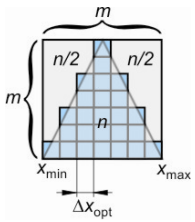
„Jedes Rechteck entspricht einem Messwert. „

Pr.Buch Abb. 5



Pr.Buch Abb. 5

Bestimmung der optimalen Klasseneinteilung



optimale Klassenanzahl m:

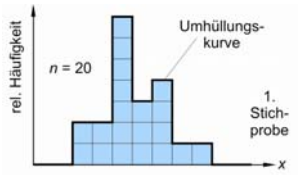
$$m^2 = 2n$$

$$m = \sqrt{2n} \qquad m = \sqrt{40} = 6.3$$

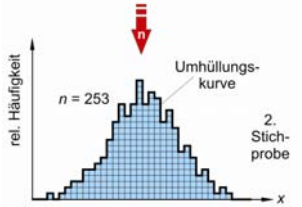
optimale Klassenbreite Δx:

$$\Delta x = \frac{x_{max} - x_{min}}{m} \qquad \Delta x = \frac{89 - 56}{6.3} = 5.2$$

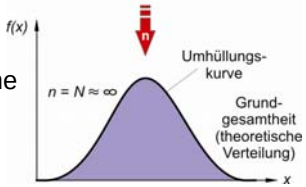
empirische Funktion



empirische Funktion



theoretische Funktion



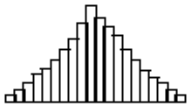
Pr.Buch Abb. 6

n vergrößert sich, die Klassenbreite Δx kann verkleinert werden

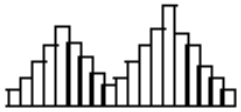
Bei großen Stichproben ergibt die empirische Verteilungsfunktion **eine sehr gute Näherung** der theoretischen Verteilungsfunktion. (Die Stichprobe ist „gleich" der Grundgesamtheit.)

Analyse von Häufigkeitsverteilungen

homogene symmetrische Stichprobe:

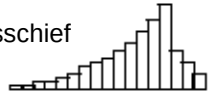


heterogene Stichprobe:

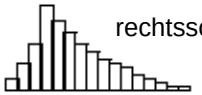


homogene nichtsymmetrische Stichproben:

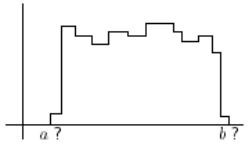
linksschief



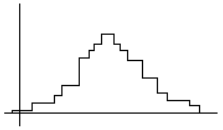
rechtsschief



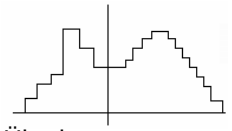
Vermutung:



Gleichverteilung?



Normalverteilung?



Überlagerung von zwei Normalverteilungen?

Lageparameter. Charakterisierung das Zentrum der Daten

Durchschnittswert (der arithmetische Mittelwert)

=average(...)
=Mittelwert(...)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Modus (Modalwert, Dichtemittel): der Wert mit der größten Wahrscheinlichkeit;
der häufigste Wert einer Häufigkeitsverteilung

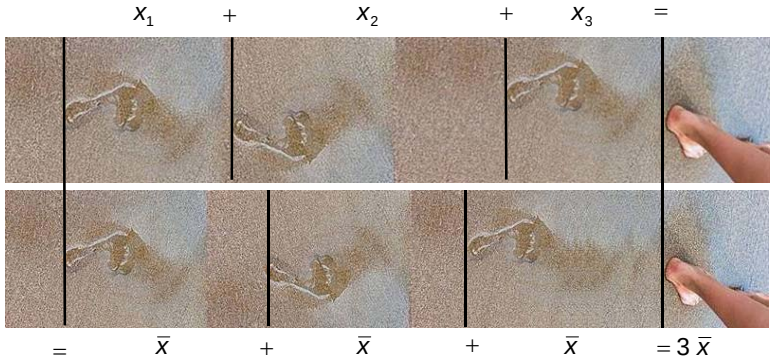
=mode(...)
=Modalwert(...)

Median (Zentralwert): halbiert eine Stichprobe.
Anzahl der Daten der Stichprobe kleiner als Median =
= Anzahl der Daten der Stichprobe größer als Median

$$x_{med} = \begin{cases} x_{(n+1)/2} & \text{falls } n \text{ ungerade} \\ (x_{n/2} + x_{(n/2+1)})/2 & \text{falls } n \text{ gerade} \end{cases}$$

=median(...)
=Median(...)

Durchschnittswert (der arithmetische Mittelwert)



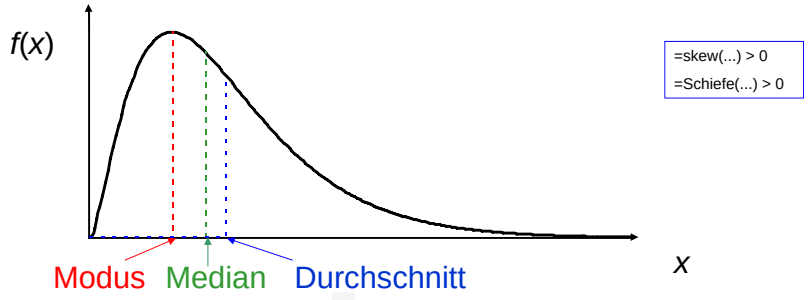
$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = 0$$

Die Summe der Abweichungen der Daten von diesem Wert ist gleich Null.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

=average(...)
=Mittelwert(...)

Linkssteile bzw. rechtschiefe Verteilung



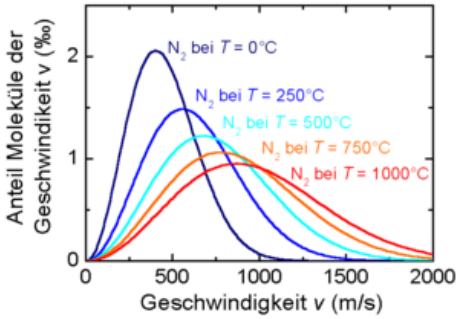
=skew(...) > 0
=Schiefe(...) > 0

z.B. Einkommensverteilungen in einem Land:

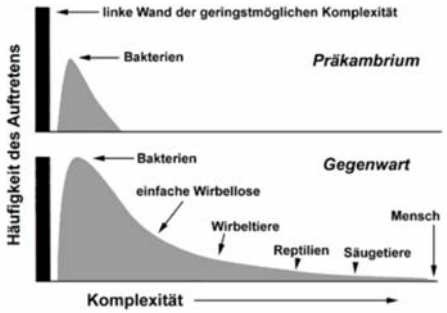
Der Großteil der Bevölkerung verdient relativ wenig, während es nur wenig Leute gibt, die sehr viel verdienen.

Weitere Beispiele

Maxwell-Boltzmann-Verteilung

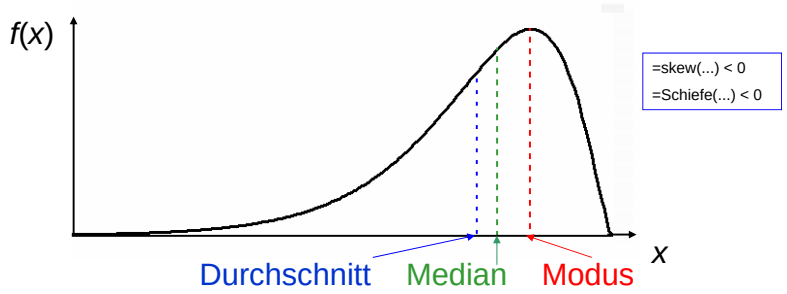


Komplexität der Tiere



www.vordenker.de/if_gould/images/verteilung.gif

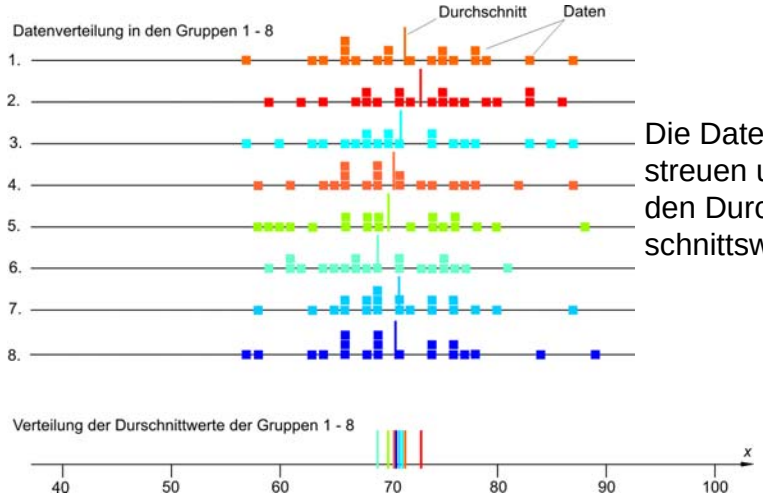
Linksschiefe bzw. rechtssteile Verteilung



=skew(...) < 0
=Schiefe(...) < 0

z.B. Dauer einer Schwangerschaft

Daten und ihre Durchschnittswerte



Die Daten streuen um den Durch-schnittswert.

Pulsfrequenzen (1/Min)

Pr.Buch Abb. 10

Streuungsmasse (Variabilitätsmaße, Variationsmaße)
Mass für die Streubreite von Daten

Streuungsparameter. Charakterisierung der Variation der Daten

Standardabweichung
(Streuung der Messdaten, *s*):
die mittlere Abweichung vom Durchschnitt:

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

=stdev(...)
=Stabw(...)

das Quadrat der Streuung, die mittlere quadratische Abweichung, auch als **Varianz** bezeichnet:

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

=var(...)
=Varianz(...)

Spannweite: x_{max}-x_{min}

=max(...)-min(...)

α-Quantil

$0 < \alpha < 1$

(seien dazu die x_i aufsteigend sortiert):

$$x_\alpha = \begin{cases} x_{[n\alpha]+1} & \text{falls } n\alpha \text{ keine ganze Zahl ist} \\ (x_{n\alpha} + x_{n\alpha+1})/2 & \text{falls } n\alpha \text{ ganzzahlig ist} \end{cases}$$

$x_{1/4}$ – unteres Quartil $x_{3/4}$ – oberes Quartil

$x_{1/10}$ – unteres Dezil $x_{9/10}$ – oberes Dezil

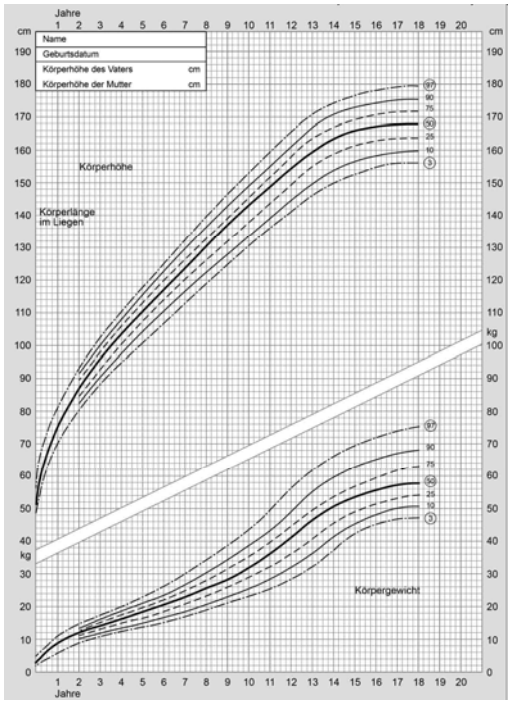
halber Quartilabstand : $(x_{3/4} - x_{1/4})/2$

=Quartil(...)

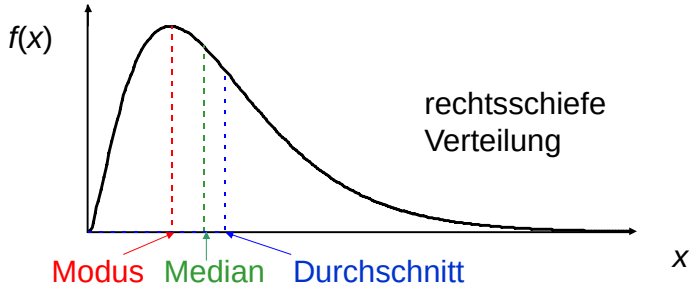
mit Wörter: z.B. **Dezile**
Durch Dezile (lat. „Zehntelwerte“) wird die Verteilung in 10 gleich große Teile zerlegt. Unterhalb des dritten Dezils liegen 30 % der Verteilung.

Perzentilenkurven sind ein Werkzeug für den Arzt.

Wachstums- und Gewichtskurven für Mädchen

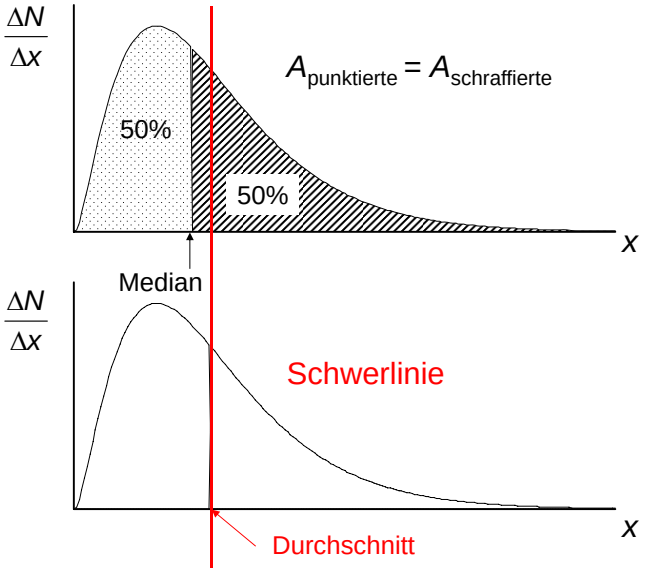


=percentile(...)
=Quantil(...)



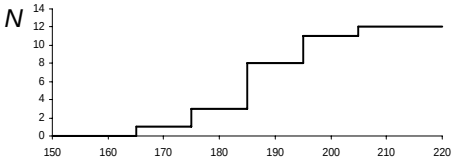
Skalentypen	zulässige Lage-Parameter	zulässige Streuungs-Parameter
Nominalskala	Modus	–
Ordinalskala	Modus, Median	–
numerische Skalen	Modus, Median, Durchschnittswert	Spannweite, Quartilabstand, Standardabweichung

Position des Medians und des Durchschnitts einer Verteilung



Summen- (kumulierte/kumulative) Häufigkeitsverteilung

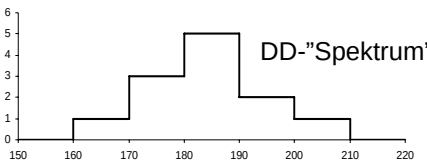
Summen-Häufigkeits-verteilung



Wieviele Werte sind kleiner als *h*?

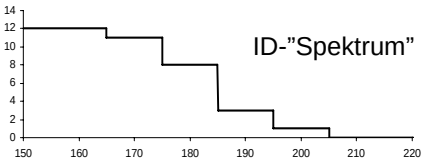
$\frac{\Delta N}{\Delta h}\left(\frac{1}{10\,\mathrm{cm}}\right)$

Häufigkeitsdichte-Verteilung



$M=N_0-N$

„Summen-Häufigkeits-verteilung"



Wieviele Werte sind grösser als *h*?

Coulter-Zähler