Quantitative Methoden der Informatik

Elemente der Deskriptiven Statistik I

Prof. Dr. rer. nat.Thomas Wiebringhaus

Herzlich Willkommen!

Klausurdauer (coronabedingt) 60min., sonst 90min.
Klausur: mehr verstehen, wenig rechnen (Formelsammlung wird hochgeladen und mit der Klausur ebenfalls ausgeteilt;
TR (laut TR- Liste) wird benötigt)

Software: R und RStudio ab der zweiten Veranstaltung,
Installationshinweise folgen im OC.

Ich empfehle diese beiden FOM Springer Lehrbücher, die Skripte (und das Internet für schnelles Nachschlagen).

FOM Springer Lehrbücher

Sebastian Sauer (2019): Moderne Datenanalyse mit R – Daten einlesen, aufbereiten, visualisieren und kommunizieren.
https://link.springer.com/content/pdf/10.1007%2F978-3-658-21587-3.pdf

Angewandte Wirtschaftsstatistik (Lübke, Vogt) [Online Campus]

Weitere Quellen des Skripts: https://www.pearson-studium.de/statistik-ohne-angst-vor-formeln.html#tabs

Elemente der Deskriptiven Statistik I

Elemente der Deskriptiven Statistik I

- Merkmale + Skalenniveaus
- Absolute + relative Häufigkeiten
- Säulendiagramm , Empirische Verteilungsfunktion und Histogramm
- Lagemaße (MW, Median und Modus), Boxplots und Quantile
- Streumaße: Varianz und Standardabweichung (sd)

- Deskriptive („beschreibende“) Statistik
  - Reine Darstellung/ Ordnung von empirischen Daten durch Tabellen, Kennzahlen oder Grafiken
- Explorative („erkundende“) Statistik
  - Erste Datenanalyse, Hypothesen bilden etc.
- Inferentielle („schließende“) Statistik
  - Hypothesen verifizieren/ falsifizieren
  - Schluss von Stichprobe auf Grundgesamtheit

Merkmale + Skalenniveaus

1.1 Merkmalsausprägungen: Beispiele

Merkmal	Merkmalsausprägung
Haarfarbe	Blond, braun,…
Blutgruppe	A, B, 0,..
Geschlecht	m/w/d
Temperatur	25°C, -273,15°C (0 Kelvin),..
Schulnoten	Sehr gut, ungenügend,..

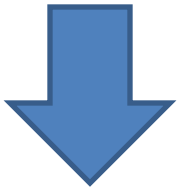
Skalenniveaus

metrisch

Rangfolge + gleiche Abstände
z.B. Körpergröße, Investitionen
Intervallskala: Nullpunkt festgelegt (z.B. Celsius)
Verhältnisskala: Nullpunkt existiert (z.B. Kelvin, Blutdruck)

ordinal

Rangfolge + ungleiche Abstände
z.B. Platzierungen, Schulnoten



nominal

Ohne Rangfolge + ungleiche Abstände
z.B. Kategorien wie Haarfarben

1.1 Merkmalsausprägungen: weitere Beispiele

Merkmal	Merkmals-ausprägungen	nominal/ordinal/ metrisch	diskret/stetig
Familienstand von Befragten	ledig (=1), verheiratet (=2), geschieden (=3), verwitwet (=4), ver-partnert (=5), ...	nominal	diskret
Zeiten der Teilnehmer an einem 100-m-Lauf	11,21 sec., 11,24 sec., ...	metrisch	stetig
Preis eines Sportartikels	29,90 €, 34,90 €, ...	metrisch	diskret
Platzierungen in einem 100-m-Lauf	1., 2., 3., ...	ordinal	diskret
Marke verkaufter LCD-Fernsehgeräte	SONY, Philips, ...	nominal	diskret
Einwohnerzahlen verschiedener Bundesländer	2,362.929, 4,746.014, ...	metrisch	diskret
Weitsprungleistung von Schülern (in ganzen cm)	516 cm, 392 cm, ...	metrisch	stetig
Beurteilung der Quali-tät einer TV-Show durch ausgewählte Konsumenten	1 = sehr gut, 2 = gut, 3 = teils-teils, 4 = schlecht, 5 = sehr schlecht	ordinal	diskret
Gewicht von TV-Gerä-ten im Lager eines Unternehmens	20,426 kg, 22,822 kg, ...	metrisch	stetig

gerundet:
quasi- stetig

Absolute + relative Häufigkeiten

1.3 Häufigkeitstabelle: Beispiel

Es wurden 64 Personen nach der Anzahl ihrer Geschwister befragt. Die Urliste sei gegeben durch:

1 2 3 8 1 4 2 4 1 3 2 2 4 2 2 5 2 2 1 0 3 4 1 1 3 4 3 2 1 2 2 4
6 2 2 3 0 2 4 5 3 7 1 2 2 5 4 1 1 3 3 2 2 2 1 3 3 1 0 1 1 1 5 2

(a) Erstellen Sie die Häufigkeitstabelle.

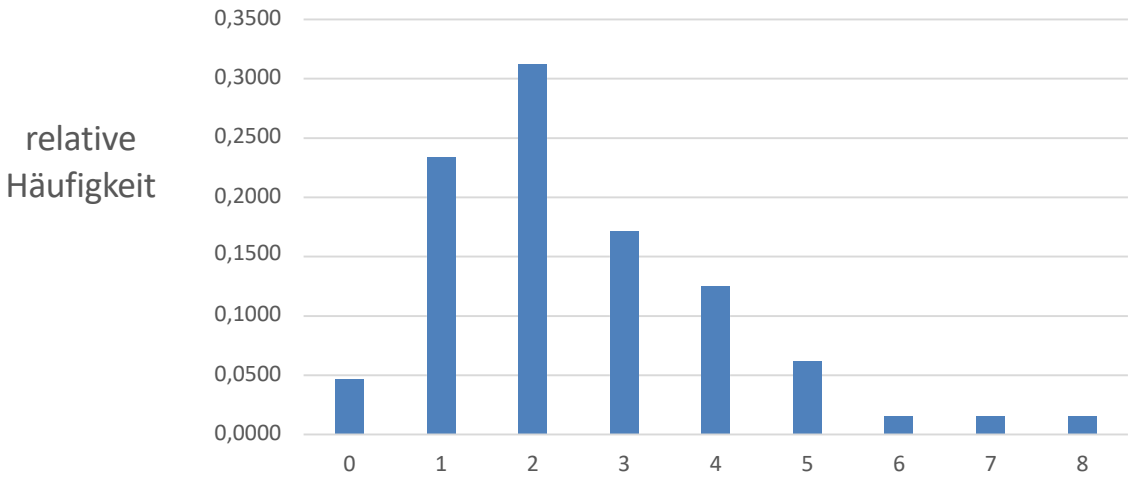
← Wie kann ich die Daten übersichtlicher darstellen?

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit		kum. Häufigkeit	
1	0	3	$\frac{3}{64}$	0,0469	$\frac{3}{64}$	0,0469
2	1	15	$\frac{15}{64}$	0,2344	$\frac{18}{64}$	0,2813
3	2	20	$\frac{20}{64}$	0,3125	$\frac{38}{64}$	0,5938
4	3	11	$\frac{11}{64}$	0,1719	$\frac{49}{64}$	0,7656
5	4	8	$\frac{8}{64}$	0,1250	$\frac{57}{64}$	0,8906
6	5	4	$\frac{4}{64}$	0,0625	$\frac{61}{64}$	0,9531
7	6	1	$\frac{1}{64}$	0,0156	$\frac{62}{64}$	0,9688
8	7	1	$\frac{1}{64}$	0,0156	$\frac{63}{64}$	0,9844
9	8	1	$\frac{1}{64}$	0,0156	1	1,0000

$$rel.H. = \frac{abs.H.}{Anzahl}$$

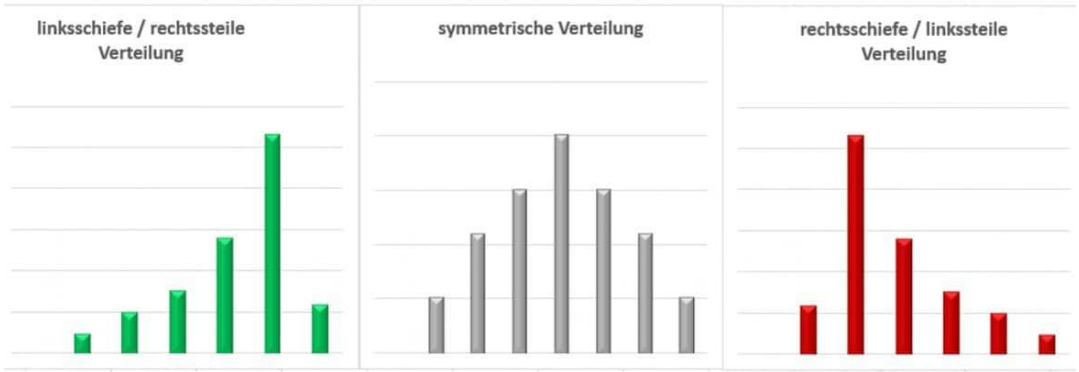
Säulendiagramm, Empirische Verteilungsfunktion und Histogramm

1.4 Stabdiagramm



← linkssteil = rechtsschief

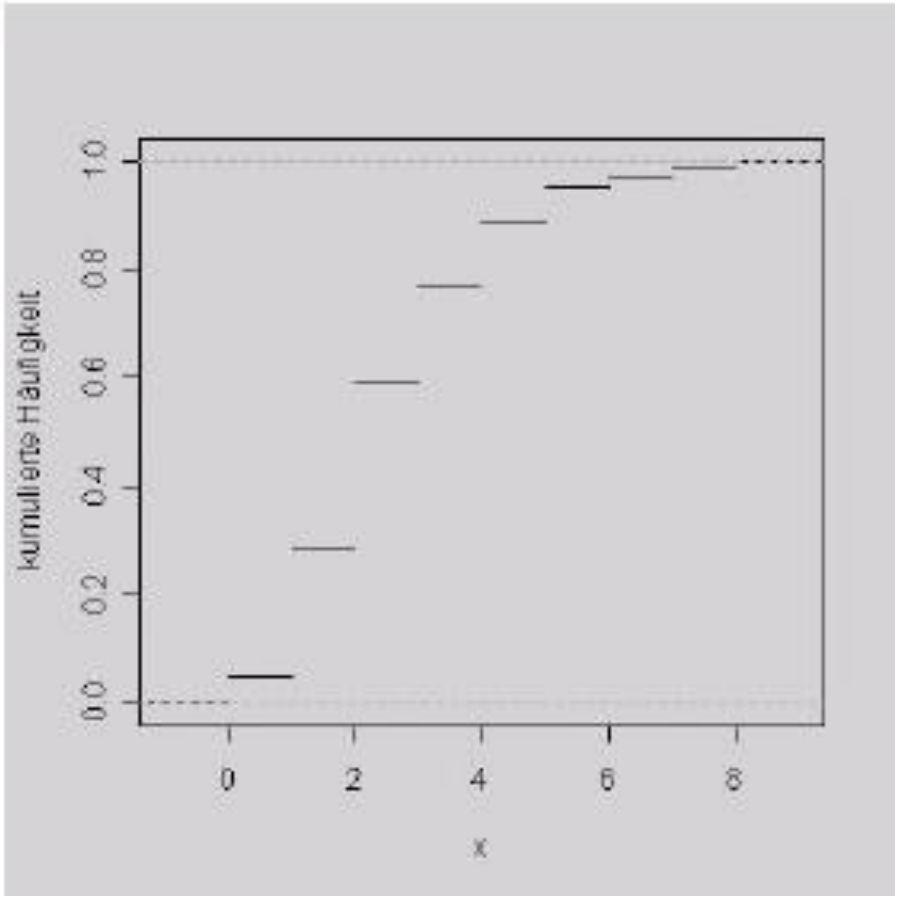
Allgemein gilt:



https://www.statistik-nachhilfe.de/ratgeber/statistik/deskriptive-statistik/masszahlen/parameter-der-form/schiefe

1.5 Empirische Verteilungsfunktion

Merkmals-Ausprägung x_i	***kumulative Häufigkeit*** $\sum_{i=1}^{n} h_i$
0	0,0469
1	0,2813
2	0,5938
3	0,7656
4	0,8906
5	0,9531
6	0,9688
7	0,9844
8	1,0000



(d) Wie groß ist der Anteil der Personen, die genau 3 Geschwister haben?

Der Anteil der Personen, die drei Geschwister haben, kann aus der Häufigkeitstabelle abgelesen werden:

$$h(X = 3) = \frac{11}{64} = 0,172$$

17,2 % der Personen haben drei Geschwister.

← Erkenntnisse gewinnen

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit	kum. Häufigkeit
1	0	3	$\frac{3}{64}$	$\frac{3}{64}$
2	1	15	$\frac{15}{64}$	$\frac{18}{64}$
3	2	20	$\frac{20}{64}$	$\frac{38}{64}$
4	3	11	$\frac{11}{64}$ ←	$\frac{49}{64}$
5	4	8	$\frac{8}{64}$	$\frac{57}{64}$
6	5	4	$\frac{4}{64}$	$\frac{61}{64}$
7	6	1	$\frac{1}{64}$	$\frac{62}{64}$
8	7	1	$\frac{1}{64}$	$\frac{63}{64}$
9	8	1	$\frac{1}{64}$	1

$h(X = 3)$

(e) Wie groß ist der Anteil der Personen, die mindestens 1 Geschwisterteil haben?

Das Gegenteil von dem Anteil derjenigen, die mindestens einen Geschwisterteil haben, ist der Anteil derer, die gar keine Geschwister haben. Der gesuchte Anteil lässt sich demnach einfach über das Gegenereignis bestimmen:

← Erkenntnisse gewinnen

$$h(X \geq 1) = 1 - h(X < 1) = 1 - h(X = 0) = 1 - \frac{3}{64} = \frac{61}{64} = 0,953$$

Der Anteil der Personen, die mindestens einen Geschwisterteil haben, beträgt 95,3 %.

$X =$

Nummer	Realisations-möglichkeit	abs. Häufigkeit	rel. Häufigkeit	kum. Häufigkeit
1	0	3	$\frac{3}{64}$	$\frac{3}{64}$
2	1	15	$\frac{15}{64}$	$\frac{18}{64}$
3	2	20	$\frac{20}{64}$	$\frac{38}{64}$
4	3	11	$\frac{11}{64}$	$\frac{49}{64}$
5	4	8	$\frac{8}{64}$	$\frac{57}{64}$
6	5	4	$\frac{4}{64}$	$\frac{61}{64}$
7	6	1	$\frac{1}{64}$	$\frac{62}{64}$
8	7	1	$\frac{1}{64}$	$\frac{63}{64}$
9	8	1	$\frac{1}{64}$	1

$1 - h(X{=}0)$

Beispiel

Aus den Fragebögen wurden 50 Studenten ausgewählt und die Schulabschlussnoten notiert. Daraus ergibt sich folgende Urliste:

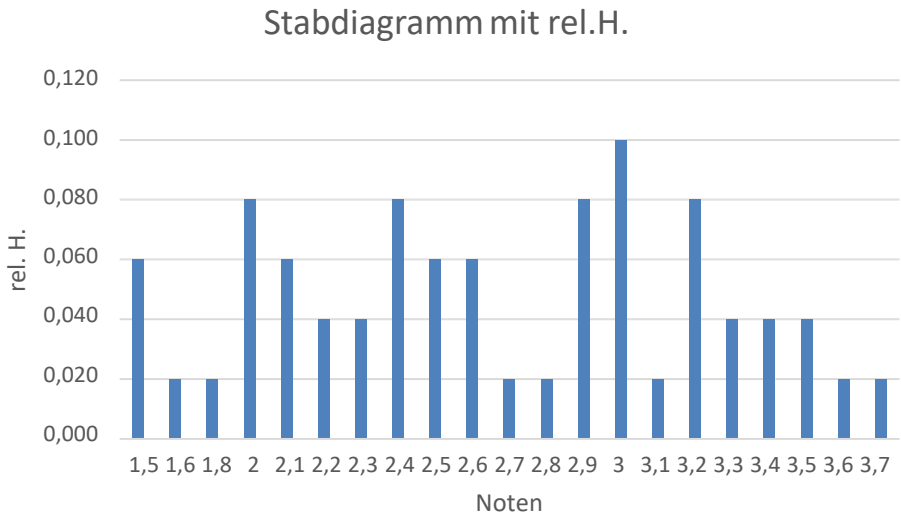
2.1	2.1	2.2	2.5	3.0	2.0	2.2	2.4	2.9	2.4
3.6	3.1	3.2	2.3	3.3	1.8	2.7	3.7	2.9	2.9
3.2	3.4	2.3	2.4	3.0	2.0	2.6	3.5	3.2	3.0
3.4	2.6	2.5	1.5	1.5	3.0	2.0	2.5	2.9	2.8
1.6	2.0	2.6	2.1	3.2	3.0	3.5	2.4	1.5	3.3

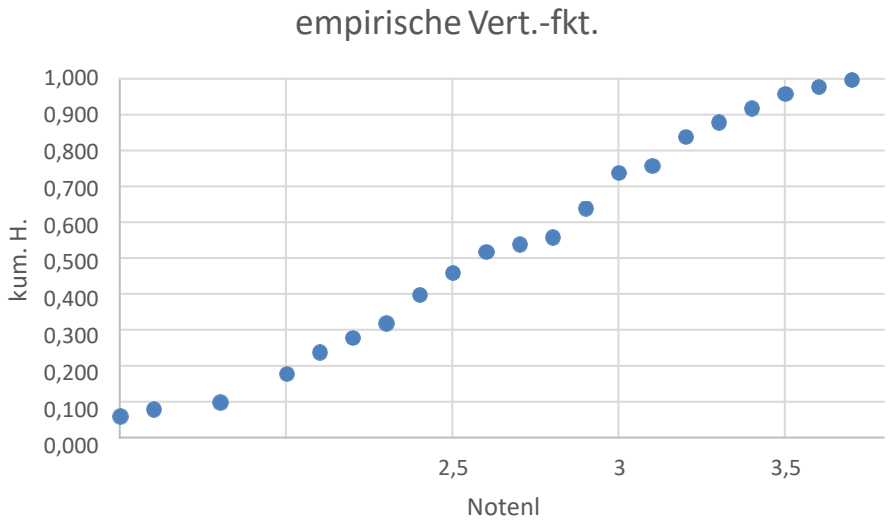
Beispiel

Häufigkeitstabelle

Note	abs.H.	rel.H.	kum.H.
1,5	3	0,060	0,060
1,6	1	0,020	0,080
1,8	1	0,020	0,100
2	4	0,080	0,180
2,1	3	0,060	0,240
2,2	2	0,040	0,280
2,3	2	0,040	0,320
2,4	4	0,080	0,400
2,5	3	0,060	0,460
2,6	3	0,060	0,520
2,7	1	0,020	0,540
2,8	1	0,020	0,560
2,9	4	0,080	0,640
3	5	0,100	0,740
3,1	1	0,020	0,760
3,2	4	0,080	0,840
3,3	2	0,040	0,880
3,4	2	0,040	0,920
3,5	2	0,040	0,960
3,6	1	0,020	0,980
3,7	1	0,020	1,000

n=50

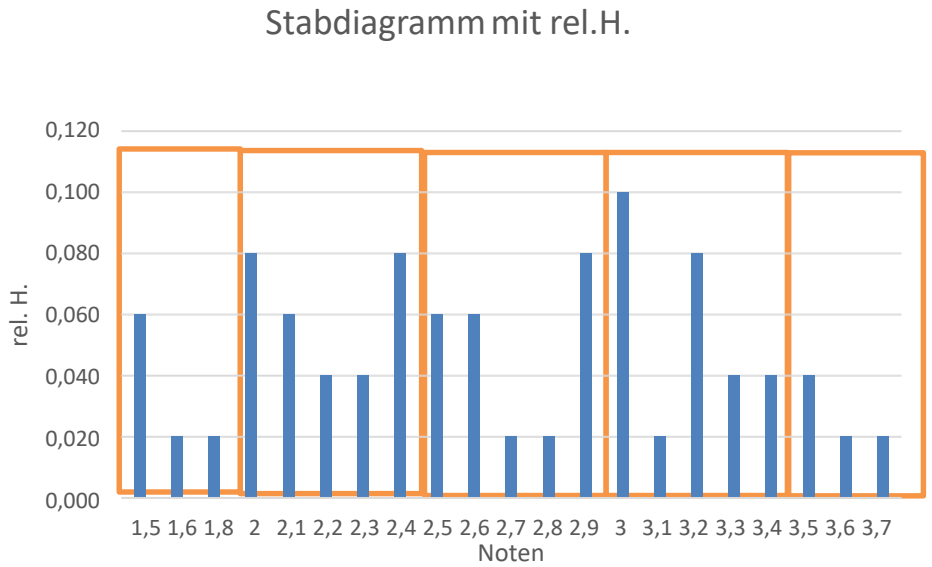




1.6 Klassenbildung und Histogramm

Es werden folgende Klassen gebildet:

1.Klasse : von 1.0 bis unter 1.5
2.Klasse : von 1.5 bis unter 2.0
3.Klasse : von 2.0 bis unter 2.5
4.Klasse : von 2.5 bis unter 3.0
5.Klasse : von 3.0 bis unter 3.5
6.Klasse : von 3.5 bis unter 4.0



Zusammenfassung von Datenpunkten zu Gruppen: Klassenbildung/ Klassierung.
Hier: Klassenbreite jeweils 0,5
Histogramm: Fläche eines Rechtecks entspricht der rel. Häufigkeit
Gesamtfläche = 1

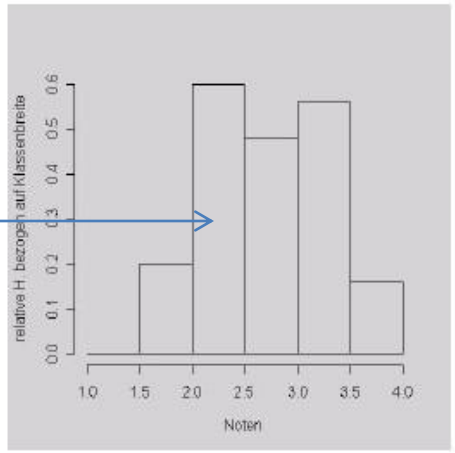
1.6 Klassenbildung und Histogramm

(a) Erstellen Sie die Häufigkeitstabelle.

Klasse	Intervall	abs. Häufigkeit	rel. Häufigkeit
1	$[1; 1,5)$	0	0
2	$[1,5; 2)$	5	0,1
3	$[2; 2,5)$	15	0,3
4	$[2,5; 3)$	12	0,24
5	$[3; 3,5)$	14	0,28
6	$[3,5; 4)$	4	0,08

0,5 * 0,6 = 0,3

Halboffenes Intervall: die 3,5 ist mit drin, die 4 nicht mehr



(b) Zeichnen und interpretieren Sie das Histogramm.

Für das Histogramm braucht man die Höhen der einzelnen Rechtecke. Man teilt also die relativen Häufigkeiten der einzelnen Klassen durch die jeweilige Klassenbreite und erhält die Höhen. Für dieses Beispiel ist die Klassenbreite für jede Klasse identisch i.H.v. $\Delta_i = 0,5$. Die einzelnen Höhen für die Klassen lauten:

1. Klasse: $\frac{h_1}{\Delta_1} = \frac{0}{0,5} = 0$
2. Klasse: $\frac{h_2}{\Delta_2} = \frac{0,1}{0,5} = 0,2$
3. Klasse: $\frac{h_3}{\Delta_3} = \frac{0,3}{0,5} = 0,6$
4. Klasse: $\frac{h_4}{\Delta_4} = \frac{0,24}{0,5} = 0,48$
5. Klasse: $\frac{h_5}{\Delta_5} = \frac{0,28}{0,5} = 0,56$
6. Klasse: $\frac{h_6}{\Delta_6} = \frac{0,08}{0,5} = 0,16$

Die meisten Studenten haben eine Schulabschlussnote zwischen 2.0 und 3.5 erreicht. Sehr wenige erreichten eine sehr gute oder sehr schlechte Note.

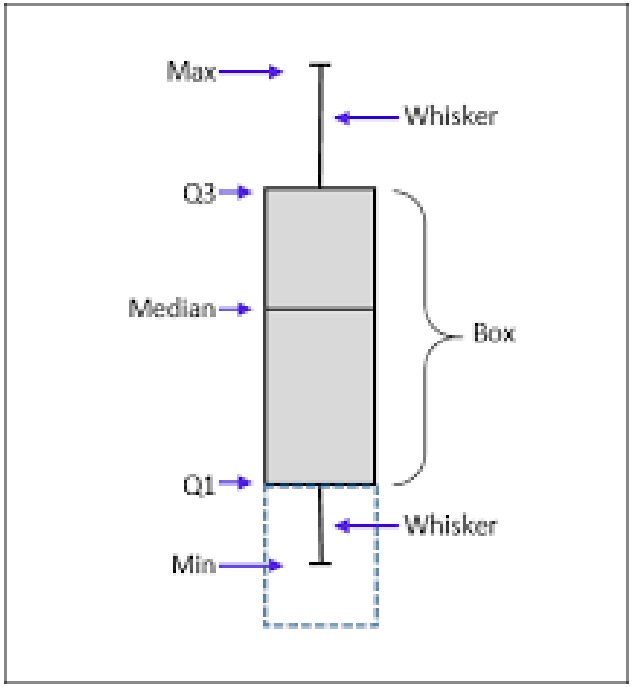
A=a*b (Fläche eines Rechtecks)

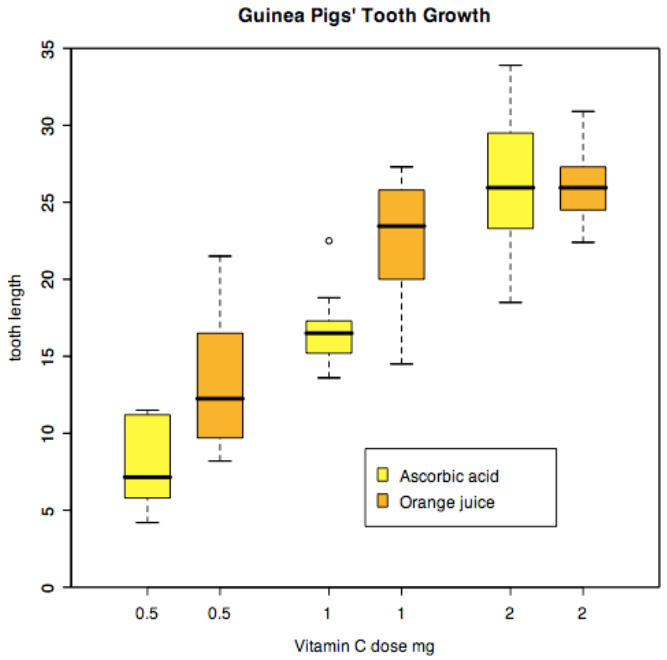
rel.H. = Klassenbreite * Höhe
Höhe = rel.H./ Klassenbreite
0,6 = 0,3 / 0,5

Lagemaße (MW, Median und Modus),
Boxplots und Quantile

2. Lagemaße

1. Arithmetischer Mittelwert (MW) und Median (Zentralwert)
2. Boxplot (Median, Quantile, Min./Max.)





Olympics Boxplot

JUL 9 Posted by statsinthewild

[7/23/2012 Addition: I've updated these plots using ggplot2 to look nicer. They can be found here.]

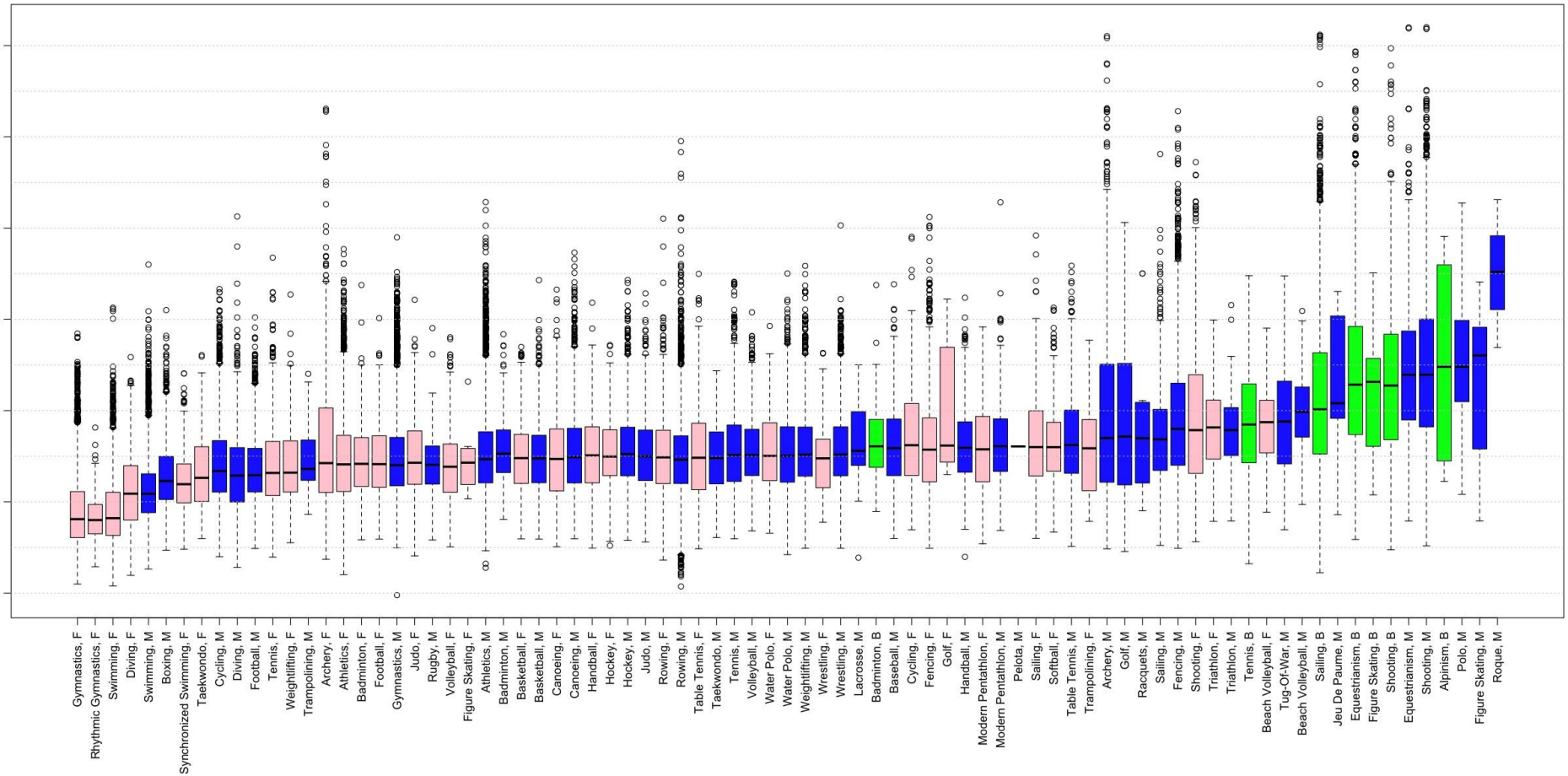
Recently, I saw this pretty cool chart at the Washington Post (I originally saw the chart at this wonderful blog here) about the ages of olympians from the past three olympics. I commented to myself that I thought it would be more interesting with boxplots of the data, rather than simple ranges, and I also wondered what it would look like if we used data from all of the past olympics.

So, I wrote some R code and began scraping sports-reference.com/olympics to get a data set with all of the olympic athletes from all of the games. This took me quite some time (and work kept getting in the way), but I eventually got it right and collected the data.

Here are some of the resulting graphs:

Below is a graph of side-by-size boxplots of age for each sport by gender with blue for male, pink for female, and green for mixed competition. And no the 11 year old female swimmer is not a typo like I originally thought.

Age distribution of Olympic Athletes by Sport and Gender: All-time
Female = Pink, Male = Blue, Both = Green



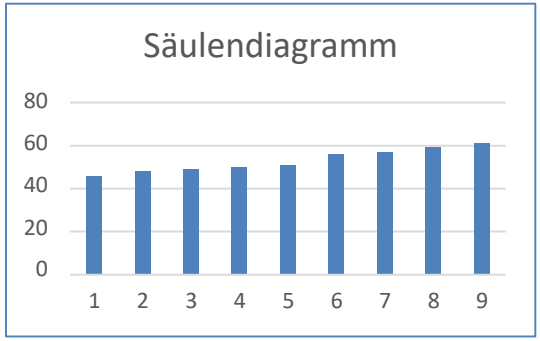
http://statsinthewild.com/2012/07/09/olympics-boxplot/

2.1 Mittelwert und Median

Aus einem Fragebogen wurde als interessierendes Merkmal das Gewicht ausgewählt. Hierzu wurde das Gewicht von 9 Personen aufgelistet:

51 56 57 48 49 61 46 50 59

(a) Bestimmen Sie den Mittelwert und den Median des Merkmals Gewicht.



Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \quad \bar{x} = \frac{1}{9}(46 + 48 + 49 + \ldots + 59 + 61) = 53$$

51 (Median)

Ungerade Anzahl an n

i	1	2	3	4	5	6	7	8	9
x_i	46	48	49	50	51	56	57	59	61

Median

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

Modus = häufigster Wert

Gerade Anzahl an n

i	1	2	3	4	5	6	7	8	9	10
x_i	46	48	49	50	51	56	57	59	61	61

$$\bar{x} = \frac{51 + 56}{2} = 53{,}5 \quad \text{(Median)}$$

2.2 Boxplot

(c) Erstellen und interpretieren Sie den Boxplot der Daten.

5-Zahlen-Zusammenfassung:

$$x_{(1)} = 46 \quad x_{0.25} = 49 \quad x_{0.5} = 51 \quad x_{0.75} = 57 \quad x_{(n)} = 61$$

Ungerade Anzahl an n:

46(Min) Q_{25} (49) 51 (Median) Q_{75} (57) 61 (Max)

i	1	2	3	4	5	6	7	8	9
x_i	46	48	49	50	51	56	57	59	61

→ Es gibt unterschiedliche Methoden für die Berechnung von Quantilen, hier verwenden wir die Folgende:

$$n * p = 9 * 0{,}25 = 2{,}25; \textit{Aufrunden auf die nächste} \text{ Position} = 3$$
$$n * p = 9 * 0{,}75 = 6{,}75; \textit{Aufrunden auf die nächste} \text{ Position} = 7$$

Interquartilsabstand $iqr = Q_{75} - Q_{25}$

Gerade Anzahl an n

i	1	2	3	4	5	6	7	8	9	10
x_i	46	48	49	50	51	56	57	59	61	61

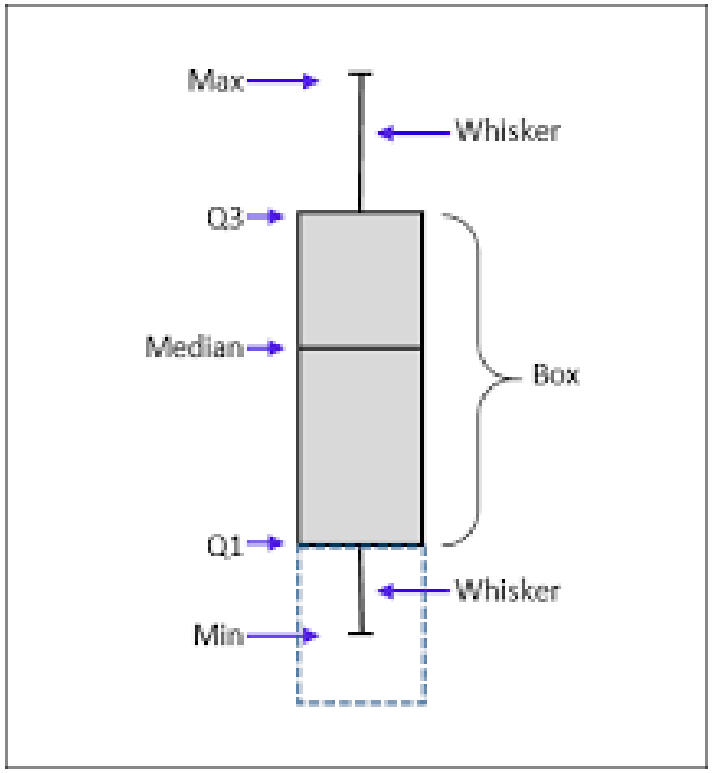
$$\bar{x} = \frac{51 + 56}{2} = 53{,}5 \text{ (Median)}$$

2.2 Boxplot

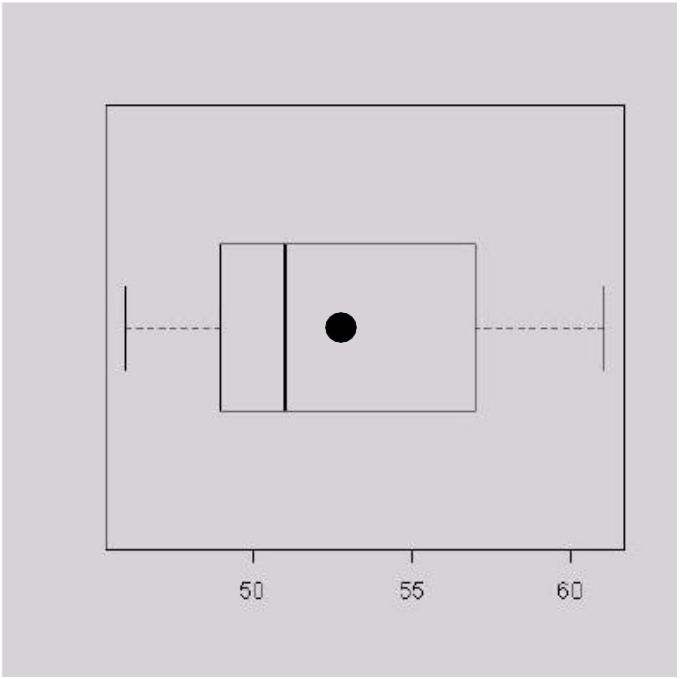
(c) Erstellen und interpretieren Sie den Boxplot der Daten.

5-Zahlen-Zusammenfassung:

$$x_{(1)} = 46 \quad x_{0.25} = 49 \quad x_{0.5} = 51 \quad x_{0.75} = 57 \quad x_{(n)} = 61$$



Aus diesen fünf Zahlen lässt sich der Boxplot konstruieren.



Boxplot Beispiel

Aufgabe 5

Zwei der häufigsten Todesursachen unter jungen Amerikanern sind Trauma und Krebs. Trauma bezeichnet dabei eine Verletzung des Körpers durch Gewalteinwirkung von außen. Trunkey (1983) gibt für von 20 bzw. 25 an diesen beiden Ursachen Gestorbenen das Alter der Gestorbenen an.

Todesfälle durch Krebs: 2, 3, 5, 9, 13, 16, 17, 19, 20, 22, 23, 26, 27, 27, 28, 29, 30, 31, 32, 34

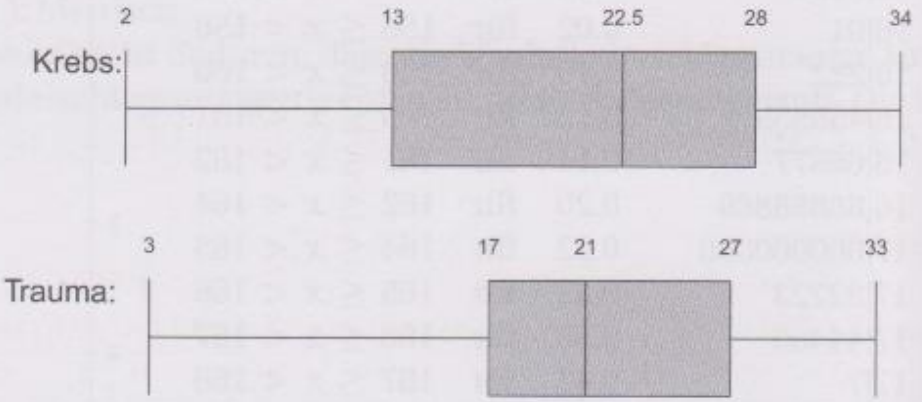
Todesfälle durch Trauma: 3, 6, 9, 14, 15, 16, 17, 17, 18, 19, 20, 20, 21, 22, 22, 23, 24, 26, 27, 28, 30, 30, 31, 32, 33

Vergleichen Sie die beiden Altersverteilungen anhand

1. der 5-Zahlen-Zusammenfassungen und der zugehörigen Box-Plots.

1. Die 5-Zahlen Zusammenfassungen sind:

	Krebs ($n=20$)	Trauma ($n=25$)
$x_{(1)}$	2	3
$x_{0.25}$	13	17
$\tilde{x}$	22.5	21
$x_{0.75}$	28	27
$x_{(n)}$	34	33



Quelle: Schlittgen, „Einführung in die Statistik“ S. 460

Streumaße: Varianz und Standardabweichung (sd)

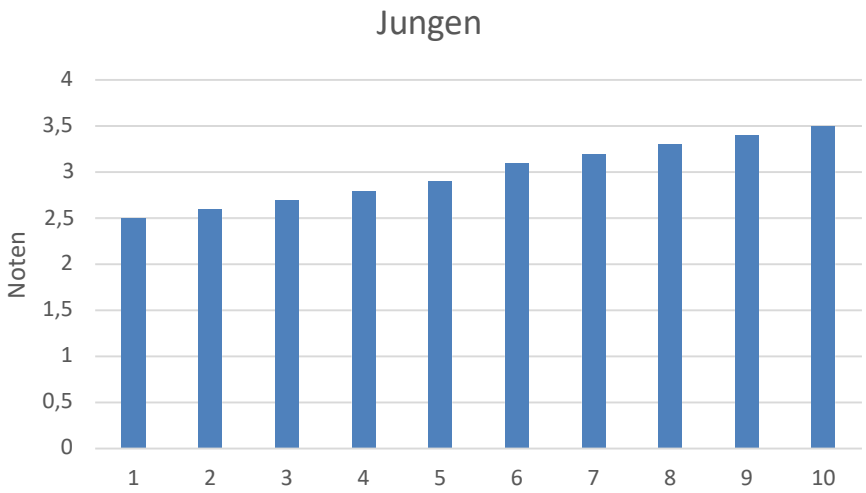
Beispiel Schulnoten.. Grenzen des Mittelwerts

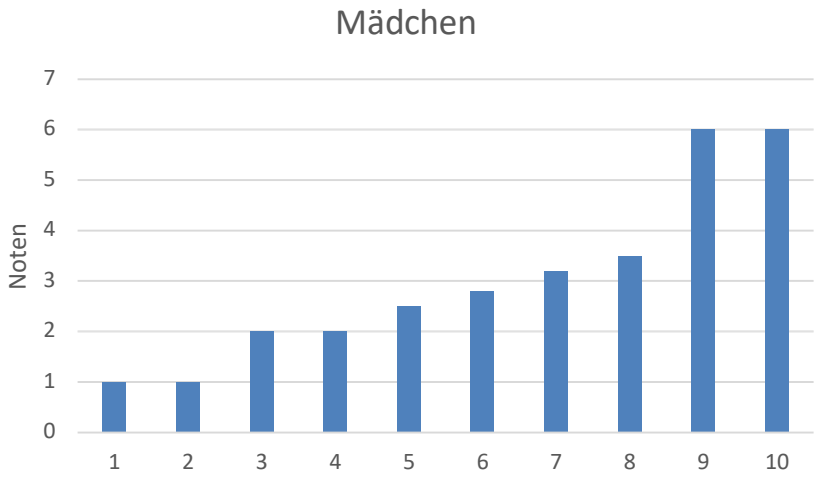
Nummer	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Jungen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6
Mädchen	1	1	2	2,5	3,2	2,8	3,5	2	6	6

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

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 3{,}0$$

Der Arithmetische MW beträgt jeweils 3,0.





Varianz = Streuung **um** den arithmetischen Mittelwert

Jungen	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	3,2	0,2	0,04
	3,5	0,5	0,25
	2,9	-0,1	0,01
	3,3	0,3	0,09
	3,4	0,4	0,16
	2,5	-0,5	0,25
	2,7	-0,3	0,09
	2,8	-0,2	0,04
	3,1	0,1	0,01
	2,6	-0,4	0,16
Summe	***30***	***0***	***1,1***

$1,1/\ 10 = 0,11$

Mädchen	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	1	-2	4
	1	-2	4
	2	-1	1
	2,5	-0,5	0,25
	3,2	0,2	0,04
	2,8	-0,2	0,04
	3,5	0,5	0,25
	2	-1	1
	6	3	9
	6	3	9
Summe	***30***	***0***	***28,58***

$28,58/\ 10 = 2,858$

Varianz

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

$$Standardabweichung\ \ s = \sqrt{s^2}$$

Beispiel: Zwei Filialen haben im letzten Jahr folgende monatliche Umsätze erzielt:

Nr.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
A	1,1	2,6	2,9	3,3	3,6	3,7	4,5	2,1	3,8	2,5	1,5	6,9
B	1,3	2,9	3,0	4,8	1,4	2,4	3,9	1,9	1,1	4,0	2,8	5,8

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1,1	-2,11	4,45
1,5	-1,71	2,92
2,1	-1,11	1,23
2,5	-0,71	0,50
2,6	-0,61	0,37
2,9	-0,31	0,10
3,3	0,09	0,01
3,6	0,39	0,15
3,7	0,49	0,24
3,8	0,59	0,35
4,5	1,29	1,67
6,9	3,69	13,63
		Summe 25,61

MW= 3,21

Varianz: 25,61/ 12 = 2,13

sd: 1,46

Varianz

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

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1,1	-1,84	3,39
1,3	-1,64	2,70
1,4	-1,54	2,38
1,9	-1,04	1,09
2,4	-0,54	0,29
2,8	-0,14	0,02
2,9	-0,04	0,00
3,0	0,06	0,00
3,9	0,96	0,92
4,0	1,06	1,12
4,8	1,86	3,45
5,8	2,86	8,17
		Summe 23,53

MW= 2,94

Varianz: 23,53/ 12 = 1,96

sd: 1,4