universität wien

Datenanalyse und Statistik
3. Deskriptive Maßzahlen

Marcus Hudec

Beschreibung quantitativer Daten

- Um die empirische Verteilung eines quantitativen Merkmals zu beschreiben, betrachten wir Parameter, die eine Verdichtung der Information des Datensatzes bzw. seiner Verteilung ermöglichen.
- Die wichtigsten Parameter sind die sog. **Lageparameter**, die das absolute Niveau (die Größenordnung) der Daten beschreiben sowie die **Streuungsparameter**, die messen, wie sehr die einzelnen Beobachtungen um das Zentrum konzentrieren
- Lagemaße

  Allgemeine Lagemaße:

  Minimum, Maximum, Quantile

  Zentrale Lagemaße:

  Arithmetisches Mittel, Median, (Modalwert)

- Streuungsmaße

  Spannweite, Quartilsabstand, Gini-Koeffizient
  Varianz, Standardabweichung, Variationskoeffizient

Lagemaßzahlen (1)

Statistische Maßzahlen, welche die absolute Lage der Verteilung beschreiben

Minimum

Der kleinste Wert eines quantitativen Merkmals

$\min(x_1,\ldots,x_n)$

Maximum

Der größte Wert eines quantitativen Merkmals

$\max(x_1,\ldots,x_n)$

Quantile

Wert einer quantitativen Variablen, welcher die geordneten Daten in Gruppen unterteilt, so dass ein bestimmter Prozentsatz darüber und ein bestimmter Prozentsatz darunter liegt (siehe Kapitel 2)

Zentrale Lagemaßzahlen (1)

Arithmetisches Mittel (mean)

Mittelwert; Summe der Werte aller Beobachtungen (Merkmalssumme) geteilt durch die Anzahl der Beobachtungen(Fälle)

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

Modalwert

Der am häufigsten auftretende Wert. Nur relevant bei uni-modalen Verteilungen bzw. bei diskreten Daten

Zentrale Lagemaßzahlen (2)

Median

Der Median oder Zentralwert eines (zumindest ordinal-skalierten) Merkmals ist der Wert jener Beobachtung, die in der nach diesem Merkmal geordneten Gesamtheit „in der Mitte“ zu liegen kommt.

$$\tilde{x}_{0,5} = \begin{cases} x_{((n+1)/2)} & \text{, n ungerade} \\ \frac{1}{2}\left(x_{(n/2)} + x_{((n+2)/2)}\right) & \text{, n gerade} \end{cases}$$

$x_{(i)}$ sind die geordneten Beobachtungen

Mittelwert oder Median?

- Das arithmetische Mittel reagiert sehr sensibel auf einzelne Extremwerte
- Der Median erweist sich gegenüber extremen Beobachtungen als relativ robust
- Eine statistische Methode ist robust, wenn sie auch bei Vorliegen fehlerbehafteter Daten vernünftige Ergebnisse liefert
- Robustheit muss man in der Regel durch Einbußen bei der Präzision erkaufen

Beispiel

- n=5 Beobachtungen
- 7, 5, 4, 8, 6
- Summe = 30
  ➔ arith. Mittel = 6
- Geordneten Beobachtungen:
- 4, 5, **6**, 7, 8
- Median=6

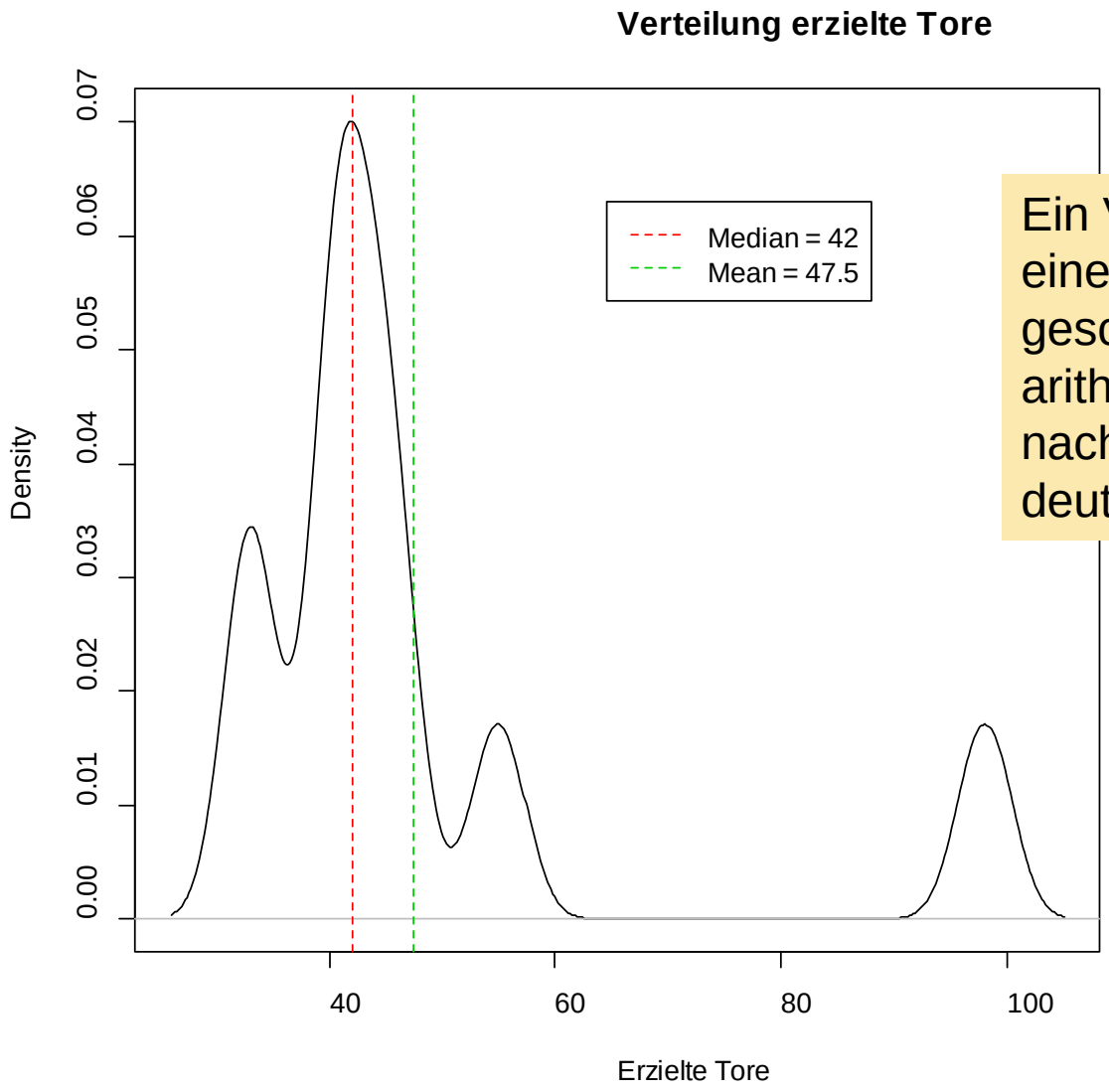
- n=5 Beobachtungen
- 7, 5, 4, **80**, 6
- Summe = 102
  ➔ arith. Mittel = 20,4
- Geordneten Beobachtungen:
- 4, 5, **6**, 7, **80**
- Median=6

Beispiel mit R: Bundesliga 2014

R Console

```
> setwd("c:\\work\\data")
> Daten <- read.csv(file="bundesliga.csv", header=T, sep=";", as.is=T)
> Daten
                  Verein Siege Remis Niederlagen Tore Gegentore
1      Red Bull Salzburg    22     4           2   98        23
2       SV Scholz Grödig    12     7           9   55        50
3             Rapid Wien    11     9           8   46        35
4           Austria Wien    11     9           8   45        35
5          SV Josko Ried     8    11           9   43        50
6  Puntigamer Sturm Graz     8     8          12   40        47
7         RZ Pellets WAC     8     8          12   41        54
8     SC Wiener Neustadt     8     8          12   33        65
9  Admira Wacker Mödling     9     6          13   41        55
10   FC Wacker Innsbruck     3    10          15   33        61
> attach(Daten)
> # Arithemetisches Mittel
> mean(Tore)
[1] 47.5
> sum(Tore)/length(Tore)
[1] 47.5
> median(Tore)
[1] 42
> plot(density(Tore), main="Verteilung erzielte Tore", xlab="Erzielte Tore")
> abline(v=median(Tore), col=2, lty=2)
> abline(v=mean(Tore), col=3, lty=2)
> legend(locator(1), legend=paste(c("Median =", "Mean ="),
+                                 c(median(Tore), mean(Tore))),
+        col=2:3, lty=c(2,2))
> |
```

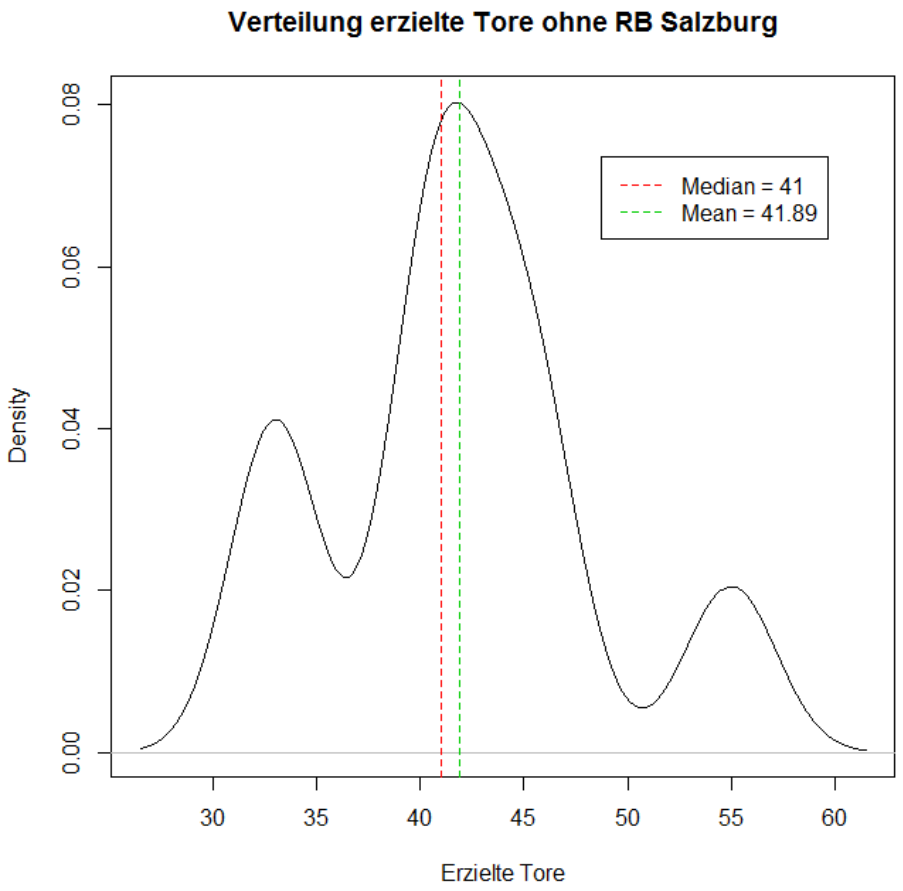
Visualisierung der Situation



Ein Verein (RB Salzburg) mit einer großen Anzahl geschossener Tore zieht das arithmetisches Mittel deutlich nach oben, während der Median deutlich geringer ist.

Visualisierung ohne RB Salzburg

```
R Console
> # Ohne Salzburg
> plot(density(Tore[-1]), main="Verteilung erzielte Tore ohne RB Salzburg",
+                         xlab="Erzielte Tore")
> abline(v=median(Tore[-1]), col=2, lty=2)
> abline(v=mean(Tore[-1]), col=3, lty=2)
> legend(locator(1), legend=paste(c("Median =", "Mean ="),
+                                 c(median(Tore[-1]), round(mean
+        col=2:3, lty=c(2,2))
> |
```



Durch die Elimination der extrem großen Anzahl geschossener Tore von RB Salzburg wird das Bild der Verteilung deutlich symmetrischer und das arithmetische Mittel und der Median divergieren nur mehr geringfügig.

Using R

```
> # Reihung der Vereine nur nach Tordifferenz
> td   <- Tore-Gegentore
> perm <- order(td, decreasing = T)
> res  <- cbind(Daten[perm, c(1,5,6)], td[perm], perm)
> res                                    # Im Prinzip ok aber nicht gut lesbar
                 Verein Tore Gegentore td[perm] perm
1      Red Bull Salzburg   98         23       75    1
3              Rapid Wien   46         35       11    3
4            Austria Wien   45         35       10    4
2       SV Scholz Grödig   55         50        5    2
5          SV Josko Ried   43         50       -7    5
6  Puntigamer Sturm Graz   40         47       -7    6
7         RZ Pellets WAC   41         54      -13    7
9  Admira Wacker Mödling   41         55      -14    9
10   FC Wacker Innsbruck   33         61      -28   10
8     SC Wiener Neustadt   33         65      -32    8
> rownames(res) <- 1:10
> colnames(res)[4] <-"Tordifferenz"
> colnames(res)[5] <-"Tabellenrang"
> res                                    # das wollten wir
                 Verein Tore Gegentore Tordifferenz Tabellenrang
1      Red Bull Salzburg   98         23          75           1
2              Rapid Wien   46         35          11           3
3            Austria Wien   45         35          10           4
4       SV Scholz Grödig   55         50           5           2
5          SV Josko Ried   43         50          -7           5
6  Puntigamer Sturm Graz   40         47          -7           6
7         RZ Pellets WAC   41         54         -13           7
8  Admira Wacker Mödling   41         55         -14           9
9    FC Wacker Innsbruck   33         61         -28          10
10    SC Wiener Neustadt   33         65         -32           8
> |
```

Mittelwert versus Median

- Das arithmetische Mittel reagiert sehr sensibel auf einzelne Extremwerte
- Der Median erweist sich gegenüber extremen Beobachtungen als relativ robust

Richter	Note
A	19
B	18,5
C	19,5
D	19
E	19

Arith. Mittel	19
Median	19

Trimmed Mean	19

Richter	Note
A	12
B	18,5
C	19,5
D	19
E	19

Arith. Mittel	17,6
Median	19

Trimmed Mean	18,83

Richter	Note
A	12
B	12
C	19,5
D	19
E	19

Arith. Mittel	16,3
Median	19

Trimmed Mean	16,67

Der kleinste und größte Wert wird von der Mittelung ausgeschlossen! (Trimmed Mean α=0,2)

α-Trimmed Mean

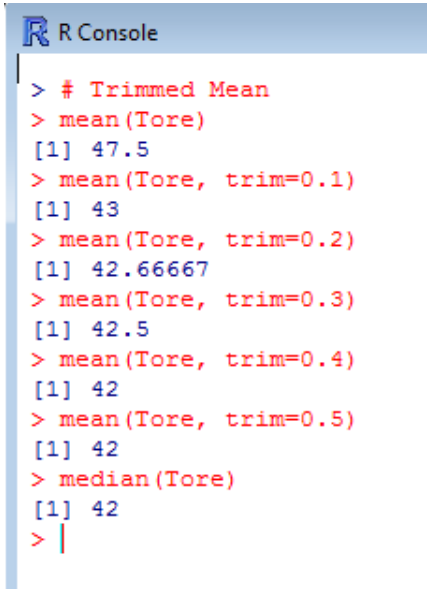
- Grundgedanke: entferne vor der Mittelung die extremen Beobachtungen, die das Ergebnis stark beeinflussen können
- Der Datensatz wird um 2*n*α Beobachtungen reduziert. Es werden also die n*α kleinsten und die n*α größten Beobachtungen von der Mittelwertsbildung ausgeschlossen
- n=20 und α=0,2 ➔ 8 Beobachtungen werden eliminiert (die 4 größten und die 4 kleinsten)

Beispiel zu α-Trimmed Mean

- Beispiel: n=40 α-Trimmed Mean mit α=5%
- Ausgeschlossen werden in dem Beispiel die vier extremen Beobachtungen: $x_{(1)}$, $x_{(2)}$,$x_{(39)}$, $x_{(40)}$ und

$$\overline{x}_{\alpha=0,05} = \frac{1}{n-4}\sum_{i=3}^{n-2} x_{(i)}$$

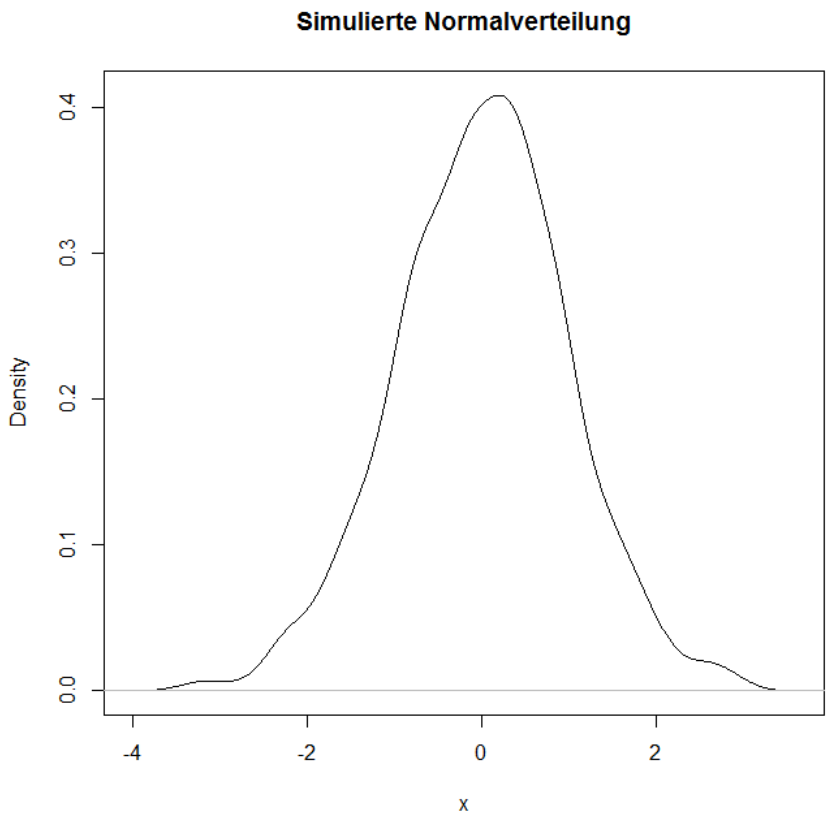
- In Excel Funktion: GESTUTZTMITTEL
- Allgemeines Prinzip umfasst eine Familie von Lagemaßzahlen:
  - α=0 ergibt das Arithmetischen Mittel
  - α➔50% ergibt den Median

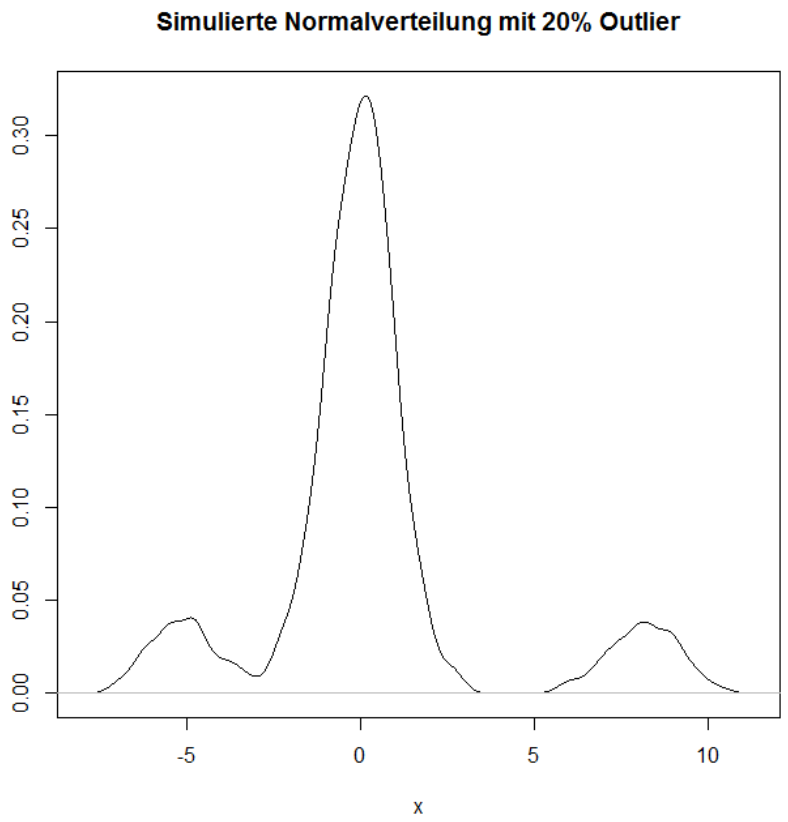


Simulation zum Effekt von Datenfehlern

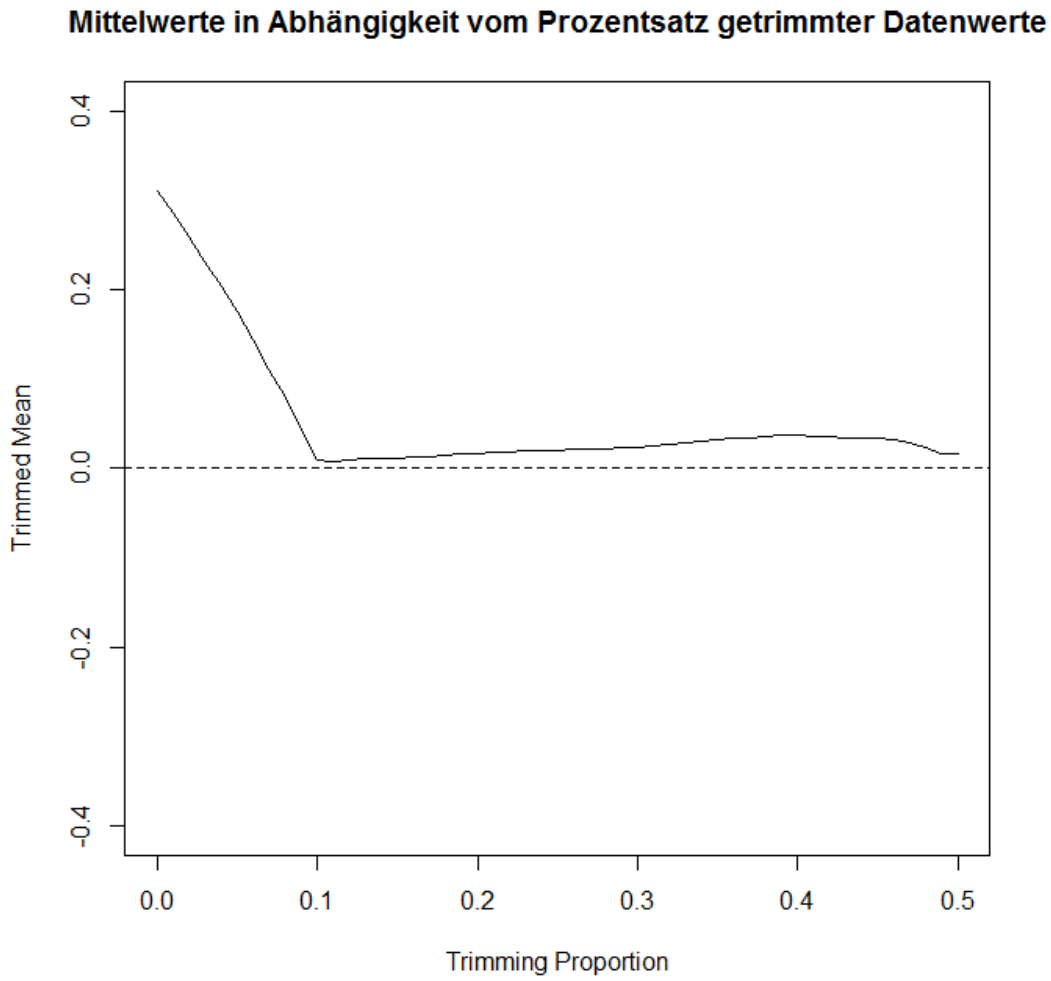
```
# ============================================================================
# Eine kleine Simulation
# ============================================================================
# Löschen des Workspace
rm(list=ls())

set.seed(111)                    # Zur Reproduzierbarkeit
x <- rnorm(800, 0, 1)
plot(density(x), main="Simulierte Normalverteilung", xlab="x")
mean(x)
x <- c(rnorm(100, -5, 1), x, rnorm(100, 8, 1))
win.graph()
plot(density(x), main="Simulierte Normalverteilung mit 20% Outlier", xlab="x")
mean(x)
median(x)
hi  <- (0:50)/100
tmi <- numeric(length(hi))
for (i in 1:length(hi)) tmi[i] <- mean(x, trim=hi[i])
win.graph()
plot(hi, tmi, type="l", xlab="Trimming Proportion",
     ylab="Trimmed Mean", ylim=c(-0.4, 0.4),
     main="Mittelwerte in Abhängigkeit vom Prozentsatz getrimmter Datenwerte")
abline(h=0, lty=2)
```

Szenario


Simulierte Normalverteilung
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
x


Simulierte Normalverteilung mit 20% Outlier
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-5
0
5
10
x

Effekt des Trimmens



Hinweis zur Definition: α-Trimmed Mean

- In manchen Büchern wird das α-Trimmed Mean in Bezug auf die Bedeutung des Parameters α anders definiert. Es wird dann der Datensatz um n*α Beobachtungen reduziert, wobei die n*α/2 kleinsten und die n*α/2 größten Beobachtungen von der Mittelung ausgeschlossen werden.
- Unsere Definition ist an die Logik des trim-Parameters der R-Funktion mean angelehnt. Die Excel-Funktion GESTUTZTMITTEL verwendet die alternative oben beschriebene Parametrisierung

Anwendungsbeispiel für Trimmed Mean

- **EURIBOR:**

  Euro Interbank Offered Rate; wird täglich um 11 Uhr Brüsseler Zeit als ungewichteter Durchschnitt aus Briefsätzen von Interbankeinlagen erstklassiger Institute auf Basis der Meldungen von rund 50 Banken berechnet. Dabei werden die jeweils 15% höchsten und tiefsten Werte eliminiert.

  Es handelt sich also in unserer Diktion um ein 15%-Trimmed Mean

Lagemaße

Quantil

Wert einer quantitativen Variablen, welcher die geordneten Daten in Gruppen unterteilt, so dass ein bestimmter Prozentsatz darüber und ein bestimmter Prozentsatz darunter liegt

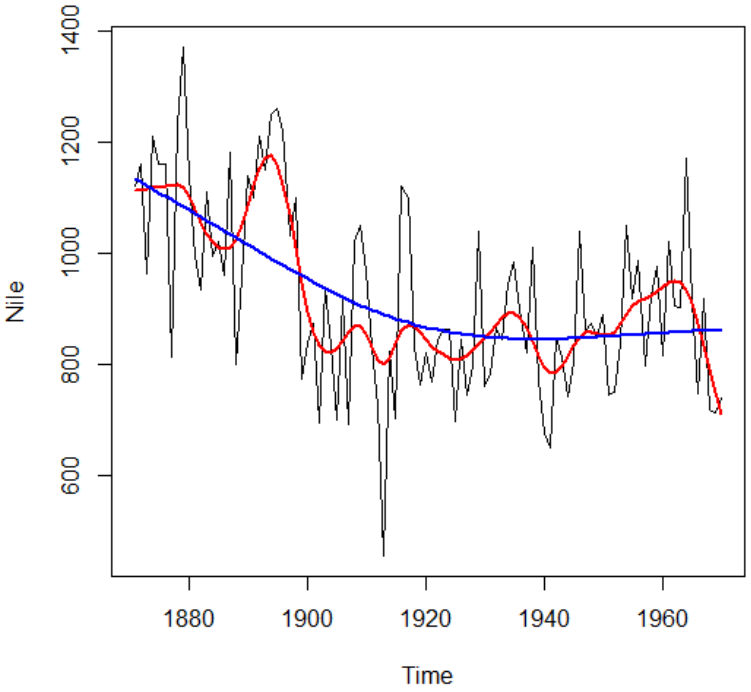
- Quartil (0%, 25%, 50%, 75%, 100%) $\tilde{x}$
- Dezil (0%, 10%, 20%, 30%, ... 90%,100%)

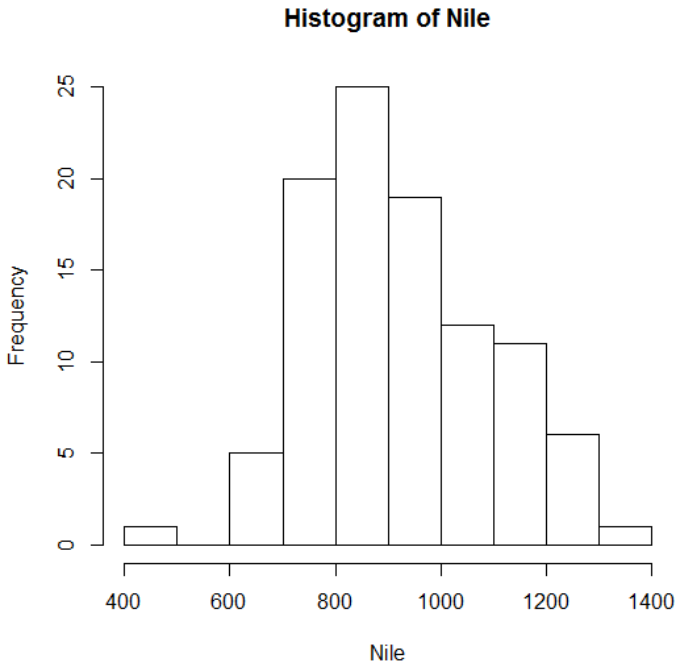
Zwei Konzepte (siehe Kapitel 2)

a) Empirisches Quantil zum Niveau α:
   $x(k)$ wobei $(k-1) < \alpha * n \leq k$
b) Alpha-Quantil: Interpolation

Nile Dataset

Measurements of the annual flow of the river Nile at Aswan (formerly Assuan), 1871–1970, in 10^8 m^3





Deskriptive Statistiken

```
> mean(Nile)
[1] 919.35
> median(Nile)
[1] 893.5
> summary(Nile)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  456.0   798.5   893.5   919.4  1032.0  1370.0
> sort(Nile)
  [1]  456  649  676  692  694  698  701  702  714  718
 [11]  726  740  742  744  744  746  749  759  764  768
 [21]  771  774  781  796  797  799  801  812  813  815
 [31]  821  822  824  831  832  833  838  840  845  845
 [41]  845  846  848  860  862  864  865  874  874  890
 [51]  897  901  906  912  916  918  919  923  935  940
 [61]  944  958  960  963  969  975  984  986  994  995
 [71] 1010 1020 1020 1020 1030 1040 1040 1050 1050 1100
 [81] 1100 1100 1110 1120 1120 1140 1140 1150 1160 1160
 [91] 1160 1170 1180 1210 1210 1220 1230 1250 1260 1370
> quantile(Nile, 1:10/10)
   10%    20%    30%    40%    50%    60%    70%    80%
 725.2  770.4  819.2  845.0  893.5  941.6  999.5 1100.0
   90%   100%
1160.0 1370.0
>
```

4 Eigenschaften des arithmetischen Mittels

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

Summe der Abweichungen ist gleich Null!

2) $\sum_{i=1}^{n}(x_i - c)^2 = \min! \quad \text{für } c = \bar{x}$

➔Prinzip der Kleinsten Quadrate

3) $y_i = a + bx_i \quad \Rightarrow \bar{y} = a + b\bar{x}$ Linearität

4) 2 Teilgesamtheiten A, B:

A $x_1, x_2, \ldots, x_{n_A}$ mit $\bar{x}$ B $y_1, y_2, \ldots, y_{n_B}$ mit $\bar{y}$

Arithmetische Mittel der Grundgesamtheit $A \cup B$:

$$\bar{z} = \frac{n_A \cdot \bar{x} + n_B \cdot \bar{y}}{n_A + n_B} \quad \text{bzw.} \quad \bar{z} = h_A \cdot \bar{x} + h_B \cdot \bar{y}$$

falls $n_A = n_B \Rightarrow \bar{z} = (\bar{x} + \bar{y})/2$

Beispiele:

- ad 3) Linearität
- Das Durchschnittseinkommen in einer bestimmten Gruppe von Arbeitern beträgt 1.000,- €
- Im Zuge einer Tarifverhandlung wird eine 4% Lohnerhöhung beschlossen. Gleichzeitig wird beschlossen, dass vom neuen Gehalt von jedem Arbeiter 10 € für einen Solidaritätsfond einbehalten werden. Wie hoch ist das neue Durchschnittseinkommen ?
- X ... Einkommen bisher
- Y ... Einkommen neu Y=1,04*X-10
- Neues Durchschnittseinkommen

$$\bar{y} = a + b\bar{x} = -10 + 1{,}04 \cdot \bar{x} = -10 + 1040 = 1030$$

Beispiele

- ad 4)
- In einem Unternehmen sind 120 Personen beschäftigt. Die Lohnsumme beträgt 240.000 €
- Das Durchschnittseinkommen der 40 männlichen Angestellten sei 2.400 €.
- Wie hoch ist das Durchschnittseinkommen der weiblichen Angestellten?
- Gesamtmittelwert: 2.000€
- 2000=(40*2.400+80*X)/120 ➔X=1.800€

Gewogenes Arithmetisches Mittel

- Die Merkmalsausprägungen werden bei der Summation gewichtet.
- Die gewichtete Summe wird dann nicht durch die Anzahl der Beobachtungen sondern durch die Summe der Gewichte dividiert:

$$\overline{x} = \frac{1}{W} \sum_{i=1}^{n} w_i x_i$$

- Falls alle $w_i = 1$ sind, ergibt sich das ungewichtete arithmetische Mittel.
- Falls die Summe der Gewichte gleich 1 ist, reduziert sich die obige Gleichung auf die Summenbildung (z.B. relative Häufigkeiten)

Arithmetisches Mittel bei diskreten Daten

- Aus vorigem Beispiel ergeben sich folgende Formeln:
- k ...Anzahl der verschiedenen Ausprägungen
- n_i … absolute Häufigkeit des Vorkommens von x_i
- h_i … relative Häufigkeit des Vorkommens von x_i

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

Arithmetisches Mittel

- Bei 12 Würfen mit dem Würfel wurde folgendes Ergebnis beobachtet:
- 5, 3, 4, 5, 5, 2, 6, 1, 4, 1, 3, 6
- Der Durchschnitt (das Arithmetische Mittel) dieser 12 Augenzahlen ist 3,75.

Beispiel Würfelwurf

- Das Ergebnis des Würfelwurfs lässt sich ohne Informationsverlust auch in einer Häufigkeitstabelle kompakt zusammenfassen:

x_i	n_i	h_i	$x_i h_i$	$x_i n_i$
1	2	2/12	2/12	2
2	1	1/12	2/12	2
3	2	2/12	6/12	6
4	2	2/12	8/12	8
5	3	3/12	15/12	15
6	2	2/12	12/12	12
Summe	12	1	45/12	45

45/12 = 3,75

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

Arithmetisches Mittel bei klassierten Daten

- Sind die einzelnen Merkmalswerte nicht mehr bekannt, so kann man das arithmetische Mittel nur mehr näherungsweise berechnen, indem man als Approximation die Klassenmitten m_i verwendet
- k ...Anzahl der Klassen

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i m_i = \sum_{i=1}^{k} h_i m_i$$

Arithmetisches Mittel bei klassierten Daten

- Bildet man die Summe der 100 Einzelbeobachtungen so ergibt sich 17440 und somit für das arithmetische Mittel 174,4 cm.
- Liegt nur noch die Tabelle mit Häufigkeiten auf der Basis der Klasseneinteilung vor so verwendet man die Klassenmitten (siehe Rechnung).
- Es ergibt sich eine kleine Abweichung: 174,3 cm

Klasse	m_i	n_i	m_i*n_i
(150 - 155]	153	3	459
(155 - 160]	158	4	632
(160 - 165]	163	10	1630
(165 - 170]	168	16	2688
(170 - 175]	173	23	3979
(175 - 180]	178	20	3560
(180 - 185]	183	11	2013
(185 - 190]	188	10	1880
(190 - 195]	193	1	193
(195 - 200]	198	2	396
	1755	100	17430
			174,3

Beachte Wahl der Klassenmitte, aufgrund der Tatsache, dass alle Messungen auf Basis cm vorliegen!

Handling von Missing Values in R

```
> head(airquality)                    # tail(airquality) gibt die letzten Records
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
> dim(airquality)
[1] 153   6
>
> attach(airquality)                  # Unterstützt die Namensauflösung
Die folgenden Objekte sind maskiert from airquality (position 3):

    Day, Month, Ozone, Solar.R, Temp, Wind
Die folgenden Objekte sind maskiert from airquality (position 4):

    Day, Month, Ozone, Solar.R, Temp, Wind
> mean(Ozone)                         # Mittelwert der Ozonwerte
[1] NA
> mean(Ozone, na.rm=T)                # erfordert explizite Behandlung der NAs
[1] 42.12931
>
> sum(Ozone, na.rm=T)/length(Ozone)         # wäre falsch
[1] 31.94118
> head(is.na(Ozone))                        # Prüfung, ob Ozonwerte fehlen
[1] FALSE FALSE FALSE FALSE  TRUE FALSE
> head(!is.na(Ozone))                       # Prüfung, ob Ozonwerte vorhanden sind
[1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
> sum(Ozone, na.rm=T)/sum(!is.na(Ozone))    # ist korrekt
[1] 42.12931
>
> # Welche Fälle haben für alle Attribute gültige Werte?
> head(complete.cases(airquality))
[1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE
>
> # Aufbau eines reduzierten Datensatzes bestehend aus kompletten Records
> newdata <- na.omit(airquality)
> head(newdata)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
7    23     299  8.6   65     5   7
8    19      99 13.8   59     5   8
```

R Console

```
> # Typische Fragstellungen in der Praxis
> # ------------------------------------
> mean(Temp)                            # Ausgangspunkt
[1] 77.88235
>
> #Durchschnittstemperatur der einzelnen Monate?
>
> # Quick and Dirty Solution - Nützt nicht die Möglichkeiten von R
> for (i in min(Month):max(Month)) print(paste(i, mean(Temp[Month==i])))
[1] "5 65.5483870967742"
[1] "6 79.1"
[1] "7 83.9032258064516"
[1] "8 83.9677419354839"
[1] "9 76.9"
>
> # R-like Solution
> tapply(Temp, Month, mean)
       5        6        7        8        9
65.54839 79.10000 83.90323 83.96774 76.90000
>
> # Speichern der Ergebnisse in einem Objekt
> res <- rbind(tapply(Temp, Month, mean),
+              tapply(Temp, Month, mean, trim=0.1),
+              tapply(Temp, Month, median))
> rownames(res) <- c("Mean", "10% Trimmed Mean", "Median")
> res
                      5        6        7        8        9
Mean             65.54839 79.10000 83.90323 83.96774 76.90000
10% Trimmed Mean 65.04000 78.91667 83.96000 83.72000 76.45833
Median           66.00000 78.00000 84.00000 82.00000 76.00000
> # Mittelwerte aller Variablen
> apply(airquality, 2, mean, na.rm=T)           # Letzten beiden Werte sinnlos
     Ozone    Solar.R       Wind       Temp      Month        Day
 42.129310 185.931507   9.957516  77.882353   6.993464  15.803922
> apply(airquality[1:4], 2, mean, na.rm=T)      # ist daher besser
     Ozone    Solar.R       Wind       Temp
 42.129310 185.931507   9.957516  77.882353
>
```

Exkurs über Mittelwerte

TESTAUFGABE

Startkapital 100.000
Verzinsung in 3 Jahren: +24%, +40%, -40%
Was ist die durchschnittliche jährliche Verzinsung?

Geometrisches Mittel

- Anwendungsbeispiele: Wachstumsprozesse
- Beobachtungen $x_1, x_2, ..., x_{n+1}$
- Wachstumsfaktoren:

$$w_i = \frac{x_{i+1}}{x_i} \quad \text{bzw.} \quad x_{i+1} = x_i \cdot w_i$$

  Verhältnis von zwei aufeinanderfolgenden Beobachtungen

- Wachstumsraten:

$$r_i = \frac{x_{i+1} - x_i}{x_i} \quad \text{bzw.} \quad r_i = w_i - 1 \quad \text{bzw.} \quad w_i = r_i + 1$$

  Relative Änderung von zwei aufeinanderfolgenden Beobachtungen

Mittleres Wachstum

- Der Wert in der nächsten Periode ergibt sich durch den aktuellen Wert multipliziert mit dem Wachstumsfaktor

$$x_{n+1} = x_n \cdot w_n$$

- Wendet man diese Überlegung wiederholt an, so ergibt sich für das Wachstum von x_1 bis x_{n+1}

$$x_{n+1} = x_n \cdot w_n = x_{n-1} \cdot w_{n-1} \cdot w_n = \ldots = x_1 \cdot w_1 \cdot w_2 \cdot \ldots \cdot w_n$$

- Das durchschnittliche Wachstum ist durch jenen konstanten Wachstumsfaktor charakterisiert, der ausgehend von x_1 zum selben Endwert x_{n+1} führt

Geometrisches Mittel

Beispiel zum geometrischen Mittel

Startkapital 100.000,00

Verzinsung in 3 Jahren: +25%, +40%, -40%

Was ist die durchschnittliche Verzinsung?

$$X_{n+1} = X_n \cdot W_n = X_{n-1} \cdot W_{n-1} \cdot W_n = \ldots$$

$$\ldots = X_1 \cdot W_1 \cdot W_2 \cdot \ldots \cdot W_n$$

$$X_{n+1} = X_1 \cdot \overline{W}_{geom} \cdot \overline{W}_{geom} \cdot \ldots \cdot \overline{W}_{geom}$$

$$\overline{W}_{geom}^{\,n} = W_1 \cdot W_2 \cdot \ldots \cdot W_n = \prod_{i=1}^{n} W_i$$

$$\overline{W}_{geom} = \sqrt[n]{\prod_{i=1}^{n} W_i}$$

X_t … Bestandsgröße zum Zeitpunkt t

W_t … Wachstumsfaktor von t auf t+1

Forderung an den Mittelwert

n-te Wurzel aus dem Produkt der n Wachstumsfaktoren

Korrekte Mittelung

Beispiel zum geometrischen Mittel

Startkapital 100.000

Verzinsung in 3 Jahren: +24%, +40%, -40%

Was ist die durchschnittliche jährliche Verzinsung?

	Rendite (Wachstumsrate)	Wachstums-faktor	Endkapital	Probe
	24,00%	1,24	124.000,00	101.367,87
	40,00%	1,40	173.600,00	102.754,45
	-40,00%	0,60	104.160,00	104.160,00
FALSCH	**8,00%**	1,013679	1,0137	
RICHTIG	**1,37%**			

$$\sqrt[3]{1.24 \times 1.4 \times 0.6} = \sqrt[3]{1.464} = 1.0137$$

Wachstumsprozess über mehrere Perioden

- Bestimmung des Durchschnitts auf der Basis von Anfangs- und Endwertes

$$x_{n+1} = x_1 \cdot w_1 \cdot w_2 \cdot \ldots \cdot w_n$$

$$\prod_{i=1}^{n} w_i = \frac{x_{n+1}}{x_1}$$

$$\bar{w}_{geom} = \sqrt[n]{\prod_{i=1}^{n} w_i} = \sqrt[n]{\frac{x_{n+1}}{x_1}}$$

Logarithmieren

- Durch Logarithmieren können Wachstumsprozesse linearisiert werden.
- Anwendung des arithmetischen Mittels auf der logarithmischen Skala und danach Rücktransformation

$$\bar{w}_{geom} = \sqrt[n]{\prod_{i=1}^{n} w_i}$$

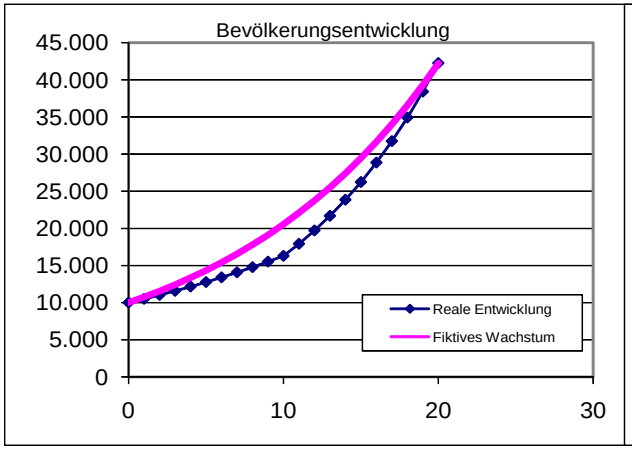
$$\log(\bar{w}_{geom}) = \log(\sqrt[n]{\prod_{i=1}^{n} w_i}) = \frac{1}{n}\log(\prod_{i=1}^{n} w_i) = \frac{1}{n}\sum_{i=1}^{n}\log(w_i)$$

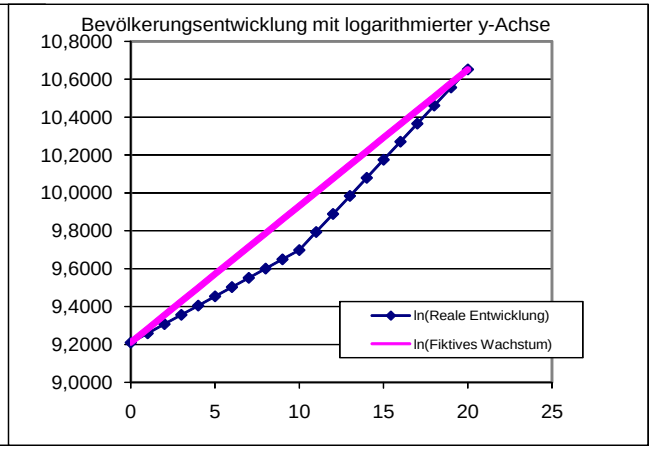
$$\bar{w}_{geom} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(w_i)\right)$$

Siehe Beispiel „Populationsdynamik“ mit Excel

Anfangspopulation: 10.000

Jahr	Wachstums-rate	Wachstums-faktor	Reale Entwicklung	Fiktives Wachstum	ln(Reale Entwicklung)	Differenz	ln(Wachstums-faktor)	ln(Fiktives Wachstum)
0			10.000	10.000	9,2103			9,2103
1	0,05	1,05	10.500	10.747	9,2591	0,04879	0,04879	9,2824
2	0,05	1,05	11.025	11.550	9,3079	0,04879	0,04879	9,3544
3	0,05	1,05	11.576	12.413	9,3567	0,04879	0,04879	9,4265
4	0,05	1,05	12.155	13.340	9,4055	0,04879	0,04879	9,4985
5	0,05	1,05	12.763	14.337	9,4543	0,04879	0,04879	9,5706
6	0,05	1,05	13.401	15.408	9,5031	0,04879	0,04879	9,6426
7	0,05	1,05	14.071	16.559	9,5519	0,04879	0,04879	9,7147
8	0,05	1,05	14.775	17.796	9,6007	0,04879	0,04879	9,7867
9	0,05	1,05	15.513	19.126	9,6495	0,04879	0,04879	9,8588
10	0,05	1,05	16.289	20.555	9,6982	0,04879	0,04879	9,9308
11	0,10	1,10	17.918	22.090	9,7936	0,09531	0,09531	10,0029
12	0,10	1,10	19.710	23.741	9,8889	0,09531	0,09531	10,0749
13	0,10	1,10	21.681	25.514	9,9842	0,09531	0,09531	10,1470
14	0,10	1,10	23.849	27.420	10,0795	0,09531	0,09531	10,2190
15	0,10	1,10	26.234	29.469	10,1748	0,09531	0,09531	10,2911
16	0,10	1,10	28.857	31.671	10,2701	0,09531	0,09531	10,3631
17	0,10	1,10	31.743	34.037	10,3654	0,09531	0,09531	10,4352
18	0,10	1,10	34.917	36.580	10,4607	0,09531	0,09531	10,5072
19	0,10	1,10	38.408	39.312	10,5560	0,09531	0,09531	10,5793
20	0,10	1,10	42.249	42.249	10,6513	0,09531	0,09531	10,6513
Durchschnittliches Wachstum			1,0747093	7,47%		0,0720502		
			1,0747093			1,0747093		





Beispiel zum Geometrischen Mittel

Jahr	Umsatz in Mio EURO	Wachstums-faktor	Wachstums-rate	ln(W.-Faktor)
2009	2			
2010	2,4	1,200	20,0%	0,1823
2011	2,9	1,208	20,8%	0,1892
2012	2,7	0,931	-6,9%	-0,0715
2013	3,1	1,148	14,8%	0,1382
Produkt		1,550		0,1096
4.Wurzel		1,116	11,6%	1,116

Arithmetisches Mittel der logarithmierten Wachstumsfaktoren

exp(0,1096)

$$\bar{w}_{geom} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(w_i)\right)$$

Testaufgabe 1

- Sie fahren staubedingt eine Strecke von 50km mit einer durchschnittlichen Reisegeschwindigkeit von 20km/h
- Die nächsten 50 km erzielen Sie eine durch-schnittliche Reisegeschwindigkeit von 100km/h.
- Wie hoch ist die Durchschnittsgeschwindigkeit für die gesamte Fahrt?

Berechnung

50 km 20 km/h
50 km 100 km/h
Ungewogenes arithmetisches Mittel
= (20 + 100) / 2 = 60 ==> FALSCH

50 km 20 km/h
50 km 100 km/h
Gewogenes arithmetisches Mittel Weg als Gewicht
= (20*50 + 100*50) / (100) = 60 ==> FALSCH

50 km 20 km/h 2,5 h
50 km 100 km/h 0,5 h
Gewogenes arithmetisches Mittel Zeit als Gewicht!!!
= (20*2,5 + 100*0,5) / (2,5 + 0,5) = 33,3 ==> KORREKT

Testaufgabe 2

- Sie planen eine Autofahrt auf der Autobahn von Wien nach Salzburg und rechnen mit einer durchschnittlichen Reisegeschwindigkeit von 100km/h
- Auf der ersten Hälfte der Strecke kommen Sie in einen Stau, so dass Sie für den halben Weg nur eine Durchschnittsgeschwindigkeit von 50km/h erreichen.
- Wie schnell müssen Sie die zweite Hälfte der Strecke zurücklegen, damit Sie noch die geplante Geschwindigkeit von 100km/h erreichen?

Harmonisches Mittel

$$\bar{x} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \quad \text{bzw.} \quad \frac{1}{\frac{1}{W}\sum_{i=1}^{n}\frac{w_i}{x_i}} \qquad W = \sum_{i=1}^{n} w_i$$

- Das untersuchte Merkmal sei ein Quotient:
  Dimension-A/Dimension-B
- Anzahl (Gewichte) der Merkmalsträger:
  - Dimension-B: Mittelung durch arithmetisches Mittel
  - Dimension-A: Mittelung durch harmonisches Mittel

Beispiel zum Harmonischen Mittel

- Geschw. Strecke
  30 km/h 60km
  90 km/h 60km

$$\frac{1}{1/2*(\frac{1}{30}+\frac{1}{90})}=\frac{1}{1/2*4/90}=\frac{180}{4}=45$$

Ungewichtetes Harmonisches Mittel

(nur zulässig, da beide Strecken gleich lang sind)

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

$$\frac{1}{1/120*(\frac{60}{30}+\frac{60}{90})}=\frac{1}{1/120*240/90}=45$$

Gewichtetes Harmonisches Mittel (Gewichte sind Strecken)

$$\overline{x}=\frac{1}{\frac{1}{W}\sum_{i=1}^{n}\frac{w_i}{x_i}} \qquad W=\sum_{i=1}^{n}w_i$$

Beispiel zum Harmonischen Mittel

- Berechnung der Durchschnittsgeschwindigkeit:

Geschw.	Dauer
30 km/h	2h
90 km/h	2/3h

- Berechnung mittels gewogenem arithmetischen Mittels:
- Gewichte sind Zeitdauern
- (30*2 +90*2/3)/(2+2/3)=3*120/8=45

$$\overline{x} = \frac{1}{W}\sum_{i=1}^{n} w_i x_i \qquad W = \sum_{i=1}^{n} w_i$$

Periodischer Kauf von Wertpapieren

Vier Käufe von Wertpapieren mit konstantem Budget von 100.000 €

Ankauf	Kurse
1	5.000,00
2	6.000,00
3	10.000,00
4	4.000,00

??? Durchschnittskurs des Käufers ???

Arithmetisches Mittel	6.250,00	ist natürlich nicht gleich dem durchschnitlichen Ankaufskurs
Harmonisches Mittel	5.581,40	ist gleich dem durchschnitlichen Ankaufskurs

Ankauf	Kurse	1/x
1	5.000,00	0,00020
2	6.000,00	0,00017
3	10.000,00	0,00010
4	4.000,00	0,00025
		0,00072

0,000179
5.581,40

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

Ankauf	Kurse	Menge	Budget
1	5.000,00	20,00	100.000,00
2	6.000,00	16,67	100.000,00
3	10.000,00	10,00	100.000,00
4	4.000,00	25,00	100.000,00
SUMME		**71,67**	**400.000,00**

Gewogenes Arithm. Mittel 5.581,40

Cost Average Effect

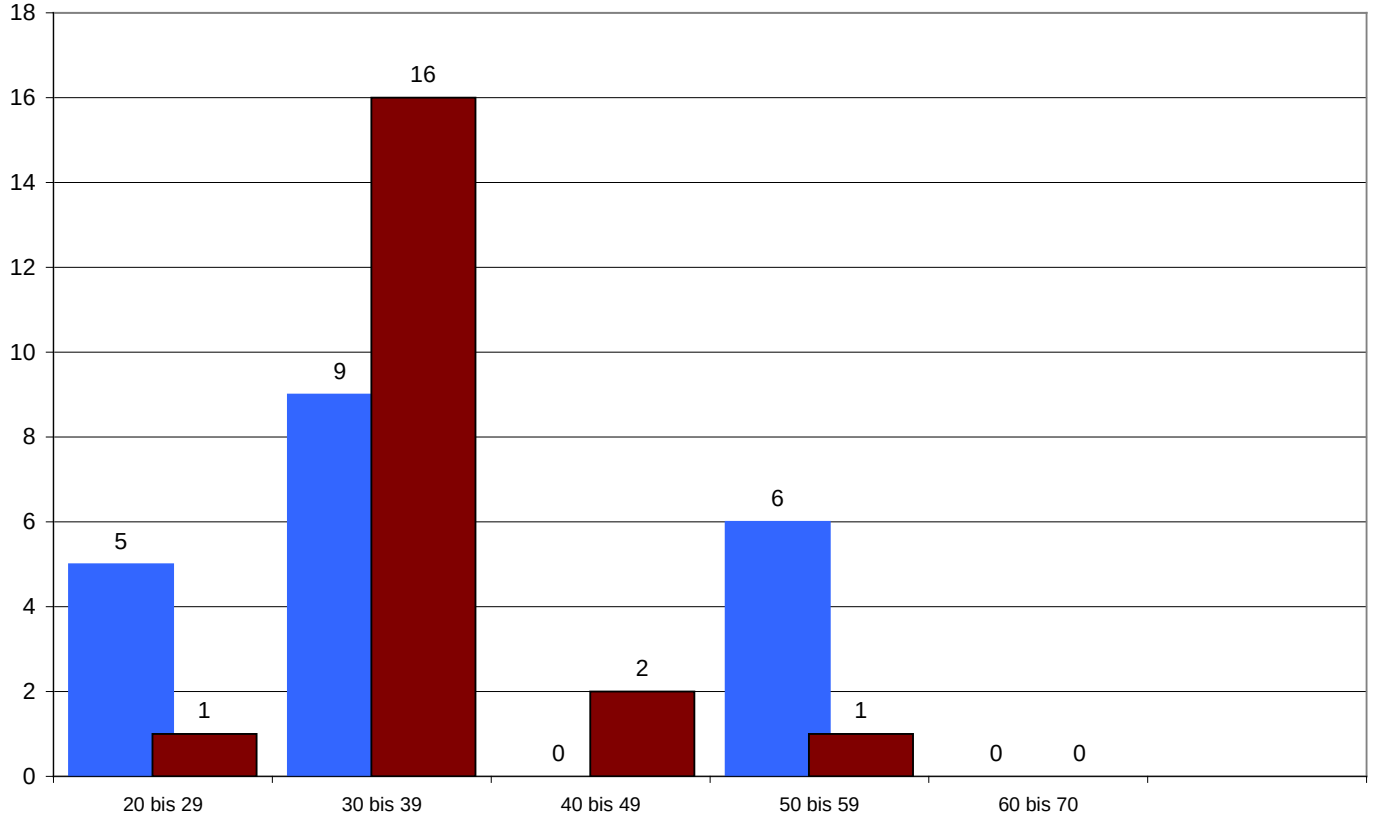
Streuungsmaße

- Statistische Maßzahlen, welche die Variabilität oder die Streubreite in den Daten messen.
- Sie beschreiben die Abweichung vom Zentrum einer Häufigkeitsverteilung
- Wie eng liegen die Merkmalsausprägungen eines quantitativen Merkmals beieinander?
- Maßzahlen:
  - Differenz von Quantilen
  - Summe der Abstände aller Merkmalsausprägungen von einem Lagemaß

Beispiel: 2 Altersverteilungen mit gleichem Mittel

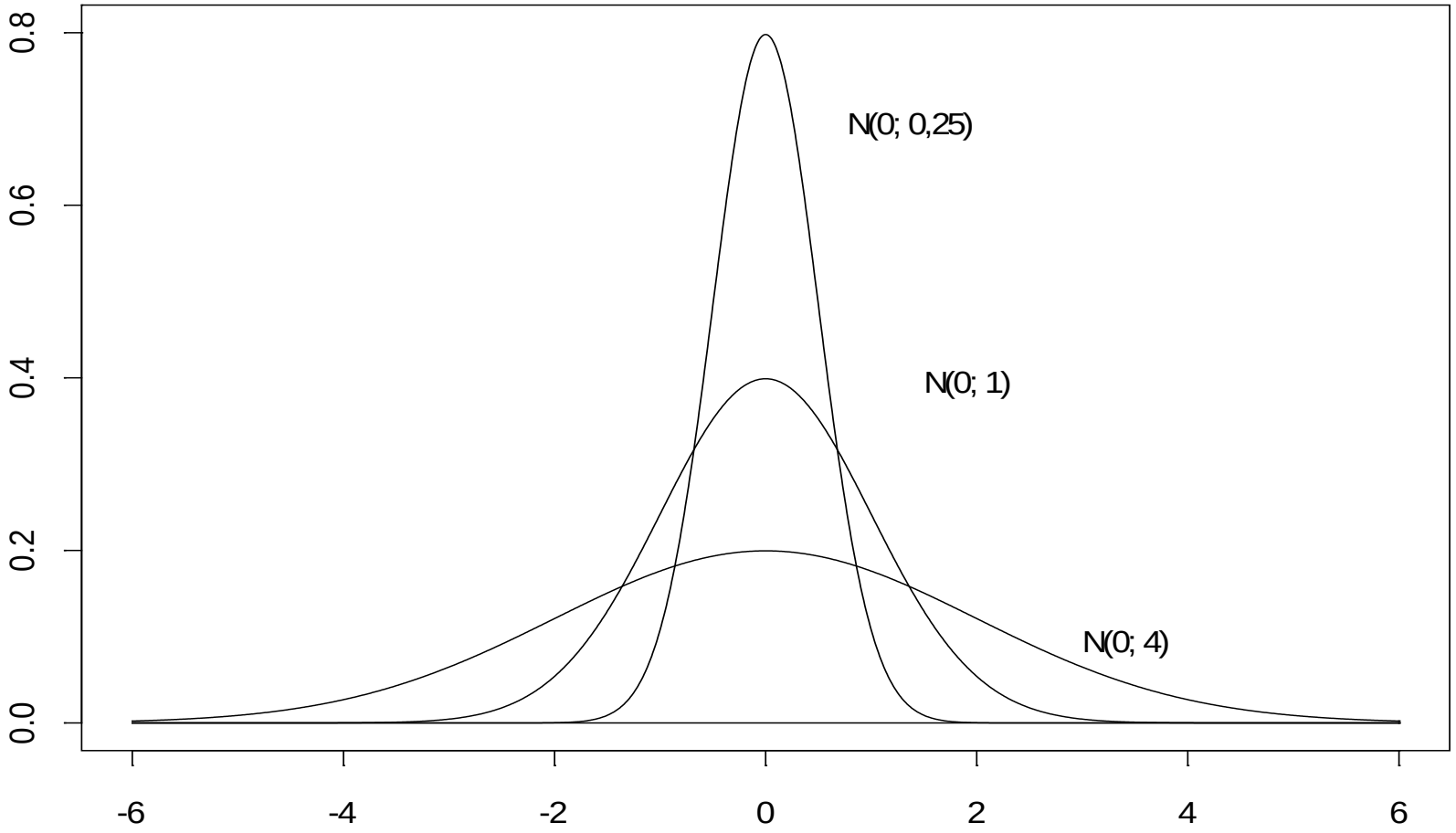
Gruppe A	Gruppe B
24	21
34	35
34	34
35	37
23	35
32	36
21	37
36	34
30	35
32	39
34	57
38	37
22	36
53	38
54	37
51	34
23	38
57	39
54	40
54	42

	Gruppe A	Gruppe B
Durchschnitt	37,1	37,1
Minimum	21,0	21,0
Maximum	57,0	57,0
Spannweite	36,0	36,0



Verschiedene Verteilungen

Alle 3 Verteilungen sind unimodal, symmetrisch und weisen den selben Mittelwert auf; sie unterscheiden sich aber in Ihrer Streuung um die Mitte



Streuungsmaße (1)

Spannweite (range)

Differenz zwischen größtem und kleinstem Wert einer numerischen Variablen;

Wertebereich, in dem alle Merkmalswerte liegen

$$R = x_{(n)} - x_{(1)}$$

Für die Praxis der Streuungsmessung oft nur wenig aussagekräftig, da stark von einzelnen Beobachtungen abhängig

Streuungsmaße (2)

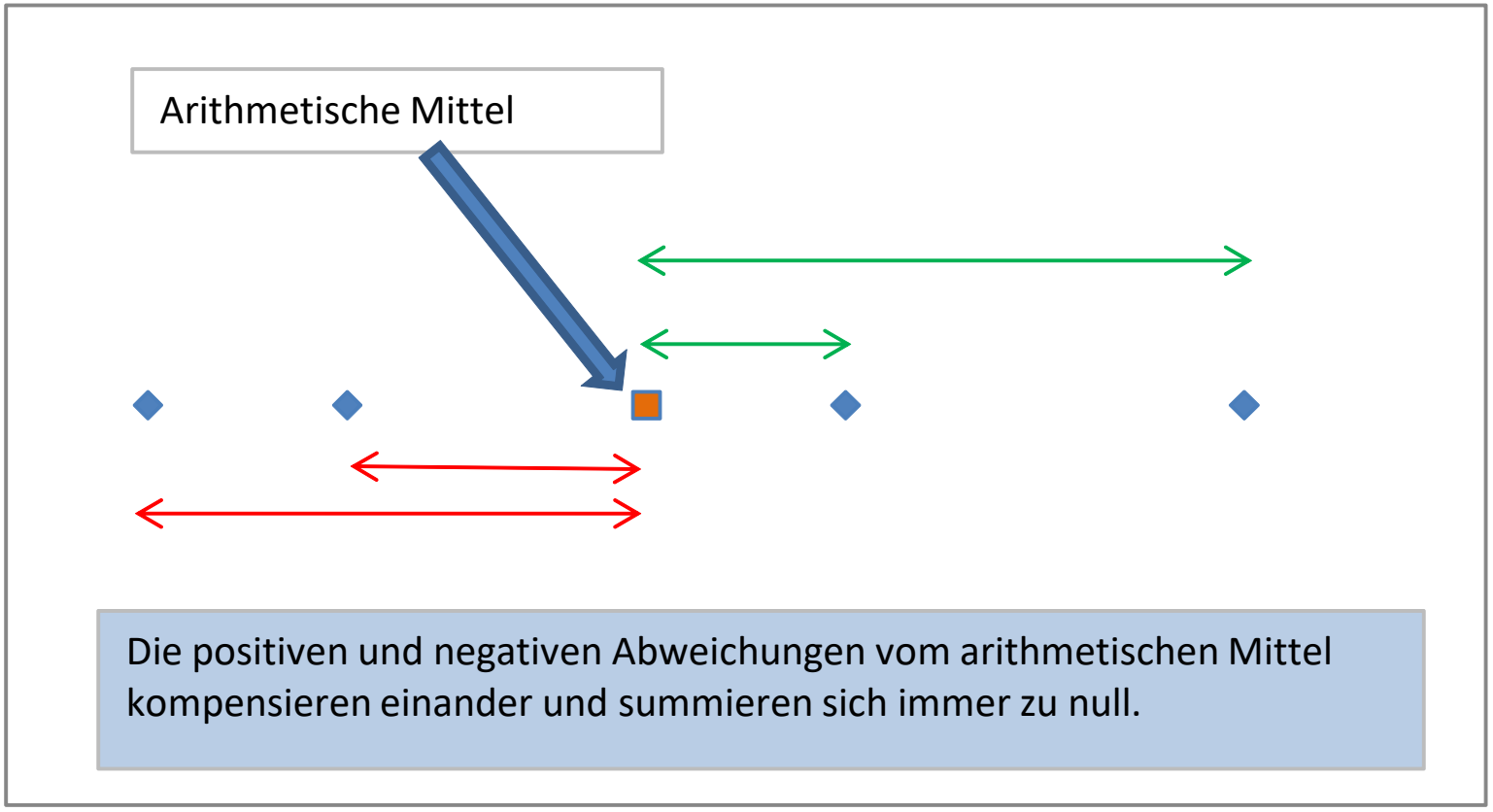
Quartilsabstand (IQ-range)

Differenz zwischen drittem und erstem Quartil;
Innerhalb des Quartilsabstandes liegen 50% der Werte;
unempfindlich gegenüber Extremwerten

$$\tilde{x}_{0,75} - \tilde{x}_{0,25}$$

Entspricht der Boxlänge im Boxplot

Abstände vom arithmetischen Mittel



Die positiven und negativen Abweichungen vom arithmetischen Mittel kompensieren einander und summieren sich immer zu null.

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

Mittlere Absolutabstände von einem Lagemaß

- Mittlere absolute Abweichung vom Median
  (mean absolute deviation (from the median) - MD)

$$MD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}\right|$$

- Mittlere absolute Abweichung vom arithmetischen Mittel
  (mean absolute deviation (from the mean) - MAA)

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

- Median der absoluten Abweichungen vom Median (median absolute deviation - MAD)

$$MAD = median(\left|x_i - \tilde{x}\right|)$$

Mittlerer quadrierter Abstand vom Mittelwert

- Varianz (variance)

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

- Die mittlere quadrierte Abweichung vom arithmetischen Mittel
- Häufig wird in der Praxis auch statt durch n durch n-1 dividiert. Dies ist v.a. dann sinnvoll, wenn man auf der Basis einer Stichprobe Aussagen für die Grundgesamtheit treffen möchte (Erklärung folgt später)
- Excel-Funktionen: *Varianz* bzw. *Varianzen*
- In R immer Division durch n-1

Berechnung von Streuungsmaßen

Nr.	Gruppe A	Abweichung vom Mittel	Absolute Abweichung vom Median	Quadratierte Abweichung vom Mittel	Nr.	Gruppe B	Abweichung vom Mittel	Absolute Abweichung vom Median	Quadratierte Abweichung vom Mittel
1	24	-13,1	10,0	170,3	1	21	-16,1	16,0	257,6
2	34	-3,1	0,0	9,3	2	35	-2,1	2,0	4,2
3	34	-3,1	0,0	9,3	3	34	-3,1	3,0	9,3
4	35	-2,1	1,0	4,2	4	37	0,0	0,0	0,0
5	23	-14,1	11,0	197,4	5	35	-2,1	2,0	4,2
6	32	-5,1	2,0	25,5	6	36	-1,1	1,0	1,1
7	21	-16,1	13,0	257,6	7	37	0,0	0,0	0,0
8	36	-1,1	2,0	1,1	8	34	-3,1	3,0	9,3
9	30	-7,1	4,0	49,7	9	35	-2,1	2,0	4,2
10	32	-5,1	2,0	25,5	10	39	2,0	2,0	3,8
11	34	-3,1	0,0	9,3	11	57	20,0	20,0	398,0
12	38	1,0	4,0	0,9	12	37	0,0	0,0	0,0
13	22	-15,1	12,0	226,5	13	36	-1,1	1,0	1,1
14	53	16,0	19,0	254,4	14	38	1,0	1,0	0,9
15	54	17,0	20,0	287,3	15	37	0,0	0,0	0,0
16	51	14,0	17,0	194,6	16	34	-3,1	3,0	9,3
17	23	-14,1	11,0	197,4	17	38	1,0	1,0	0,9
18	57	20,0	23,0	398,0	18	39	2,0	2,0	3,8
19	54	17,0	20,0	287,3	19	40	3,0	3,0	8,7
20	54	17,0	20,0	287,3	20	42	5,0	5,0	24,5
	Summe:	0,0	191,0	2893,0		Summe:	0,0	67,0	741,0
			9,55	144,6				3,35	37,0
			MD	Varianz				MD	Varianz

Abgeleitete Streuungsparameter

- Standardabweichung – (standard deviation)

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

Quadratwurzel aus der Varianz
Ist wieder in derselben Dimension wie die Beobachtungen und ist somit anschaulicher als die Varianz

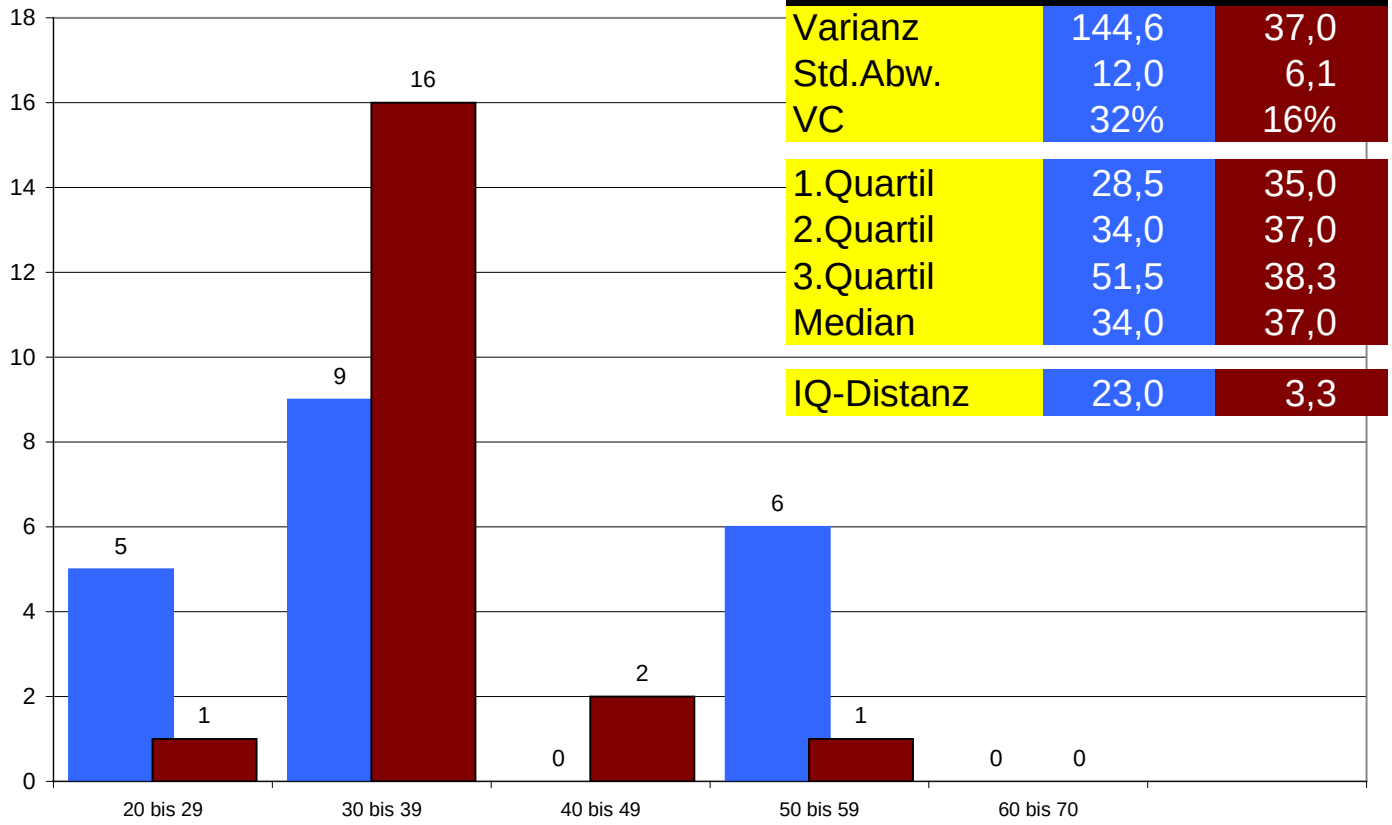
- Variationskoeffizient (coefficient of variance)

$$v = \frac{s}{\bar{x}} \quad \text{bzw.} \quad v = \frac{s}{\bar{x}} \cdot 100$$

Standardabweichung ausgedrückt in Einheiten des Mittelwerts; dimensionslose Größe gut geeignet für Vergleiche

Beispiel: 2 Altersverteilungen mit gleichem Mittel

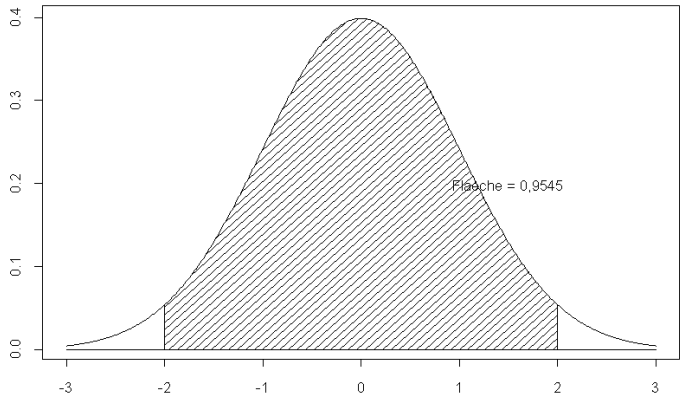
Gruppe A	Gruppe B
24	21
34	35
34	34
35	37
23	35
32	36
21	37
36	34
30	35
32	39
34	57
38	37
22	36
53	38
54	37
51	34
23	38
57	39
54	40
54	42



	Gruppe A	Gruppe B
Durchschnitt	37,1	37,1
Minimum	21,0	21,0
Maximum	57,0	57,0
Spannweite	36,0	36,0
Varianz	144,6	37,0
Std.Abw.	12,0	6,1
VC	32%	16%
1.Quartil	28,5	35,0
2.Quartil	34,0	37,0
3.Quartil	51,5	38,3
Median	34,0	37,0
IQ-Distanz	23,0	3,3

Standardabweichung & Intervalle

- Die Standardabweichung bildet den Ausgangspunkt für Konfidenzaussagen
- Intervalle sind oft reliabler als punktuelle Aussagen
- Falls die Daten einer Normalverteilung folgen (Glockenkurve) liegen etwa 95% der Daten in dem Intervall zwischen Mittelwert minus bzw. Mittelwert plus zweifacher Standardabweichung; jedenfalls aber 75%.



Bedeutung des Variationskoeffizienten

- Eine Standardabweichung von 300,- € beim monatlichen Einkommen ist in einer Gesellschaftsschicht mit einem Durchschnittseinkommen von 1.500,- € von wesentlich größerer Bedeutung als in einer Gruppe von Einkommensbeziehern, die im Monatsdurchschnitt 2.400,- € verdienen.
- In der ersten Gruppe ist der Variationskoeffizient 20%, während er in der zweiten Gruppe nur 12,5% beträgt, obwohl die Standardabweichung in beiden Gruppen gleich groß ist.

Volatilität

- Die Volatilität gilt als Einschätzung des künftigen Risikos einer Aktie. Als Maß für die Volatilität verwendet man häufig den Variationskoeffizienten
- Beispiel: Im Durchschnitt über 250 Handelstage betrug der mittlere Kurs einer Aktie A 50,59 € bei einer Standardabweichung von 36,18€. Über den selben Vergleichszeitraum betrug der mittlere Kurs einer Aktie B 396,10 € bei einer Standardabweichung von 182,96 €.
- Obwohl die Standardabweichung der Aktie A deutlich geringer ist, muss ein Investor hier mit einem größeren Risiko rechnen als bei der Aktie B, dies wird durch den Variationskoeffizienten quantifiziert:
- VK_A=36,18/50,59*100=72% VK_B=182,96/396,10*100=46%

Rechenbeispiel: Reaktionszeiten

i	x_i	x_i-xq	abs(x_i-xq)	abs(x_i-med)	(x_i-xq)²
1	0,30	0,04	0,042	0,035	0,00176
2	0,21	-0,05	0,048	0,055	0,00230
3	0,19	-0,07	0,068	0,075	0,00462
4	0,27	0,01	0,012	0,005	0,00014
5	0,32	0,06	0,062	0,055	0,00384
6	0,30	0,04	0,042	0,035	0,00176
7	0,26	0,00	0,002	0,005	0,00000
8	0,22	-0,04	0,038	0,045	0,00144
9	0,31	0,05	0,052	0,045	0,00270
10	0,20	-0,06	0,058	0,065	0,00336
Summe	**2,58**	**0,00**	**0,424**	**0,420**	**0,02196**

Arithmetisches Mittel (xq)	0,258
Median (med):	0,265

MD:	0,0420	Mittlere absolute Abweichung vom Median
MAA:	0,0424	Mittlere absolute Abweichung vom arithmetischen Mittel
MAD:	0,0450	Median der absoluten Abweichungen vom Median

Varianz:	0,002196
Standardabweichung:	0,046861
Variationskoeffizient:	18,2%

Interpretation

- Die Angabe der Standardabweichung erfolgt oft in der Form

$$\bar{x} \pm \sigma$$

- Im Beispiel: 0,258±0,047
- Unter der Annahme einer „Normalverteilung“ (Form der Häufigkeitsdichte entspricht einer Glockenkurve) liegen ca. 95% der Datenpunkte in einem Bereich von

$$\bar{x} \pm 2\sigma$$

Alternative Berechnungsformeln

- (1) $$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}\right) - \overline{x}^2$$
- (2) $$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$
- (3) $$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - n\overline{x}^2\right)$$

Mittelwert der quadrierten Werte minus dem Quadrat des Mittelwertes

Alternative Berechnungsformeln im Beispiel

i	x_i	x_i^2
1	0,30	0,09
2	0,21	0,04
3	0,19	0,04
4	0,27	0,07
5	0,32	0,10
6	0,30	0,09
7	0,26	0,07
8	0,22	0,05
9	0,31	0,10
10	0,20	0,04
Summe	**2,58**	**0,69**

Arithmetisches Mittel (xq)	0,258

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \overline{x}^2$$

=(1/10)*0,69-0,258²

0,002196

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

=(1/10)*(0,69-2,58²/10)

0,002196

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - n\overline{x}^2\right)$$

=(1/10)*(0,69-10*0,258²)

0,002196

Eigenschaften der Varianz

- "Steiner´scher Verschiebungssatz"

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

- Auswirkung linearer Transformationen

$$y = a + bx \quad \Rightarrow \quad s^2(y) = b^2 s^2(x)$$

Verschiebungs-Invarianz

Varianzermittlung aus 2 Teilpopulationen

2 Teilgesamtheiten A, B:

$A \quad x_1, x_2, \ldots, x_{n_A}$ mit $\bar{x}_A, s_A^2$ $\qquad B \quad x_1, x_2, \ldots, x_{n_B}$ mit $\bar{x}_B, s_B^2$

$$s^2 = \frac{n_A s_A^{\,2} + n_B s_B^{\,2}}{n_A + n_B} + \frac{n_A (\bar{x}_A - \bar{x})^2 + n_B (\bar{x}_B - \bar{x})^2}{n_A + n_B}$$

$$p_A = \frac{n_A}{n_A + n_B} \qquad p_B = \frac{n_B}{n_A + n_B}$$

$\bar{x} \ldots$ Gesamtmittelwert

$n = n_A + n_B$

$$s^2 = p_A \cdot s_A^{\,2} + p_B \cdot s_B^{\,2} + p_A (\bar{x}_A - \bar{x})^2 + p_B (\bar{x}_B - \bar{x})^2$$

Gesamte Varianz ist das gewichtete Mittel der Teilvarianzen plus dem gewichteten Mittel der quadratischen Abweichungen der Gruppenmittel vom Gesamtmittel ($V_{total} = V_{within} + V_{between}$)

Beispiel: Ermittlung des Mittelwertes aus 2 Teilgesamtheiten

- Es liegen Daten aus 2 Betrieben A und B vor:
- 400 Angestellte aus Betrieb A verdienen monatlich im Mittel 1.920,84 €
- 300 Angestellte aus Betrieb B verdienen monatlich im Mittel 2.012,17 €
- Dann beträgt der Gesamtmittelwert nach dem Prinzip des gewogenen arithmetischen Mittels: 1.959,98 €

$$\overline{x}_{Gesamt} = \frac{400 \cdot 1920,84 + 300 \cdot 2012,17}{400 + 300} = 1.959,98$$

Beispiel: Ermittlung der Varianz aus 2 Teilgesamtheiten

- Das Einkommen der 400 Angestellten aus Betrieb A hat eine Standardabweichung von 220,32 €
- Das Einkommen der 300 Angestellten aus Betrieb B hat eine Standardabweichung von 411,98 €
- Dann beträgt die Standardabweichung der Angestellten beider Betriebe zusammen: 320,19 €

$$s^2 = p_A \cdot s_A^{\,2} + p_B \cdot s_B^{\,2} + p_A\left(\bar{x}_A - \bar{x}\right)^2 + p_B\left(\bar{x}_B - \bar{x}\right)^2$$

$$s^2 = \frac{400 \cdot 220,32^2 + 300 \cdot 411,98^2}{400+300} +$$

$$+ \frac{400 \cdot (1920,84 - 1959,98)^2 + 300 \cdot (2012,17 - 1959,98)^2}{400+300} =$$

$$= 102520,76$$

$$s = \sqrt{102520,76} = 320,19$$

Standardisierung (z-Transformation)

- Gegeben seien Beobachtungen $x_1, ..., x_n$ mit
  Mittelwert $\bar{x}$ und Varianz s_x^2
- Gesucht ist eine lineare Transformation z=a+bx, so dass für die transformierten Daten das arithmetische Mittel 0 und die Varianz 1 wird.

$$z_i = \frac{x_i - \bar{x}}{s_x} = \underbrace{-\frac{\bar{x}}{s_x}}_{a} + \underbrace{\frac{1}{s_x}}_{b} x_i$$

$$\bar{z} = -\frac{\bar{x}}{s_x} + \frac{1}{s_x}\bar{x} = 0 \qquad s_z^2 = \frac{1}{s_x^2} s_x^2 = 1$$

- Unterscheide: x_i-$x_{(1)}$/($x_{(n)}$-$x_{(1)}$) bildet x_i in [0,1] ab

Beispiel zur Standardisierung

i	x_i	x_i-xq	$z_i=(x_i-xq)/s$	z_i^2
1	0,30	0,04	0,8963	0,803278689
2	0,21	-0,05	-1,0243	1,049180328
3	0,19	-0,07	-1,4511	2,10564663
4	0,27	0,01	0,2561	0,06557377
5	0,32	0,06	1,3230	1,750455373
6	0,30	0,04	0,8963	0,803278689
7	0,26	0,00	0,0427	0,001821494
8	0,22	-0,04	-0,8109	0,657559199
9	0,31	0,05	1,1097	1,23132969
10	0,20	-0,06	-1,2377	1,531876138
Summe	**2,58**	**0,00**	**0,00**	**10,00**

Arithmetisches Mittel (xq):	0,258	Varianz(z):	1
Standardabweichung (x):	0,046861	Standardabweichung (z):	1

Standardabweichung vs. Standardfehler

Scorewerte zwischen 0 und 100 bei n=100 Personen gemessen

12	93	17	93	17	57	93	10	14	81
8	86	82	42	91	0	65	13	68	4
10	73	28	57	53	56	30	8	68	25
50	94	67	43	18	86	94	78	73	46
66	61	98	70	38	97	94	62	11	7
2	80	25	9	76	79	85	1	12	55
6	15	42	64	8	50	26	40	6	36
89	69	38	95	100	32	3	73	53	11
21	79	25	79	96	36	86	7	23	86
7	37	30	60	85	23	96	29	93	73

Arithmetisches Mittel 49,6
Standardabweichung 31,8

Variabilität der Einzelwerte σ

Wir ziehen 10-mal eine zufällige Stichprobe von 9 Beobachtungen

	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	Sample-9	Sample-10
	38	11	8	42	93	9	38	2	61	30
	93	95	0	64	79	21	36	79	85	57
	29	40	21	89	67	62	12	80	15	21
	18	50	89	95	10	86	0	79	57	67
	50	38	18	7	7	93	9	86	12	89
	93	64	38	10	73	46	23	73	40	86
	8	3	60	42	30	32	25	60	53	43
	15	61	69	30	73	93	2	93	80	50
	6	64	79	25	68	43	67	23	14	30
arithm. Mittel	38,9	47,3	42,4	44,9	55,6	53,9	23,6	63,9	46,3	52,6

Standardfehler 10,6
Std.Abw. der 10 Stichprobenmittelwerte 10,4

Variabilität des Mittelwertes $\sigma/\sqrt{n}$

Varianz bei diskreten Daten

Treten nur k unterschiedliche Werte

$x_1, ..., x_k$ mit zugehörigen absoluten Häufigkeiten

$n_1, ..., n_k$ bzw. relativen Häufigkeiten $h_1, ..., h_k$

auf, so ergibt sich die Varianz als:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 n_i = \sum_{i=1}^{k}(x_i - \overline{x})^2 h_i$$

$$\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i^2\right) - \overline{x}^2 = \left(\sum_{i=1}^{k} h_i \cdot x_i^2\right) - \overline{x}^2$$

Berechnung Varianz – diskrete Daten

- Bei 12 Würfen mit dem Würfel wurde folgendes Ergebnis beobachtet:
- 5, 3, 4, 5, 5, 2, 6, 1, 4, 1, 3, 6
- Die Summe der 12 Augenzahlen ist 45. Der Durchschnitt (das Arithmetische Mittel) dieser 12 Augenzahlen ist 3,75. Die Summe der quadrierten Augenzahlen beträgt 25+9+…+9+36=203
- Demnach ist die Varianz:

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

$$s^2 = 1/12(203 - 12 * 3,75^2) = 2,85$$

Berechnung mittels Häufigkeiten

x_i	n_i	h_i	$x_i h_i$	$x_i n_i$	$x_i^2 n_i$	x_i-xq	$(x_i$-xq$)^2$	$(x_i$-xq$)^2 h_i$	$(x_i$-xq$)^2 n_i$
1	2	0,17	0,17	2	2	-2,75	7,56	1,26	15,13
2	1	0,08	0,17	2	4	-1,75	3,06	0,26	3,06
3	2	0,17	0,50	6	18	-0,75	0,56	0,09	1,13
4	2	0,17	0,67	8	32	0,25	0,06	0,01	0,13
5	3	0,25	1,25	15	75	1,25	1,56	0,39	4,69
6	2	0,17	1,00	12	72	2,25	5,06	0,84	10,13
	12	**1,00**	**3,75**	**45**	**203**			**2,85**	**34,25**

Arithmetisches Mittel → **3,75**

Varianz → **2,85**

Alternativ: $203/12-3{,}75^2=$**2,85**

$$s^2 = \left(\frac{1}{n} \sum_{i=1}^{k} n_i \cdot x_i^2 \right) - \overline{x}^2$$

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} \left(x_i - \overline{x} \right)^2 n_i = \sum_{i=1}^{k} (x_i - \overline{x})^2 h_i$$

Berechnung bei klassifizierten Daten

Sind die Daten in k Klassen eingeteilt, kann man auch die Varianz nur näherungsweise berechnen, indem man mit den Klassenmittelwerten $m_1, ..., m_k$ arbeitet.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}(m_i - \overline{x})^2 n_i = \sum_{i=1}^{k}(m_i - \overline{x})^2 h_i$$

bzw.

$$\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{k} m_i^2 n_i\right) - \overline{x}^2 = \left(\sum_{i=1}^{k} m_i^2 h_i\right) - \overline{x}^2$$

Beispiel Körpergröße von 100 Studenten

Klasse	n_i	h_i	m_i	$m_i h_i$	$m_i^2 h_i$
bis 150	0	0,00			
(150 - 155]	3	0,03	153	4,59	702,27
(155 - 160]	4	0,04	158	6,32	998,56
(160 - 165]	10	0,10	163	16,30	2656,90
(165 - 170]	16	0,16	168	26,88	4515,84
(170 - 175]	23	0,23	173	39,79	6883,67
(175 - 180]	20	0,20	178	35,60	6336,80
(180 - 185]	11	0,11	183	20,13	3683,79
(185 - 190]	10	0,10	188	18,80	3534,40
(190 - 195]	1	0,01	193	1,93	372,49
(195 - 200]	2	0,02	198	3,96	784,08
	100	**1**		**174,30**	**30468,80**

Mittelwert= **174,30**

Varianz= 30.468,8-174,3²=**88,31** Korrektur **86,23**

Standardabweichung= **9,40** **9,29**

Exakte Berechnungen auf Basis der Urliste

n= **100**

Summe x = **17.440**

Summe x² = **3.049.914**

Arithmetisches Mittel = **174,4**

Varianz= **83,78**

Standardabweichung= **9,15**

Sheppard-Korrektur

- Es lässt sich theoretisch zeigen, dass bei einer unimodalen Verteilung die auf der Basis der klassifizierten Daten berechnete Varianz größer ist als die aus den Originaldaten.
- Bei einer Klasseneinteilung mit konstanter Breite Δ kann der Fehler mit der sog. Sheppard-Korrektur annähernd ausgeglichen werden:

$$\sigma^2_{korr.} = \sigma^2 - \frac{\Delta^2}{12}$$

$$\sigma^2_{korr.} = 88,31 - \frac{25}{12} = 86,23$$

Höhere Verteilungsmaßzahlen

- Ein stetiges Merkmal wurde in 3 Gruppen beobachtet und in Form der folgenden Häufigkeitstabelle berichtet:

Klasse	m_i	Gruppe A	Gruppe B	Gruppe C
0-2	1	0	4	0
2-4	3	12	4	4
4-6	5	24	20	40
6-8	7	28	44	24
8-10	9	24	20	20
10-12	11	12	4	8
12-14	13	0	4	4
		100	100	100

- 1.Schritt: Berechnung von Mittelwert und Streuung

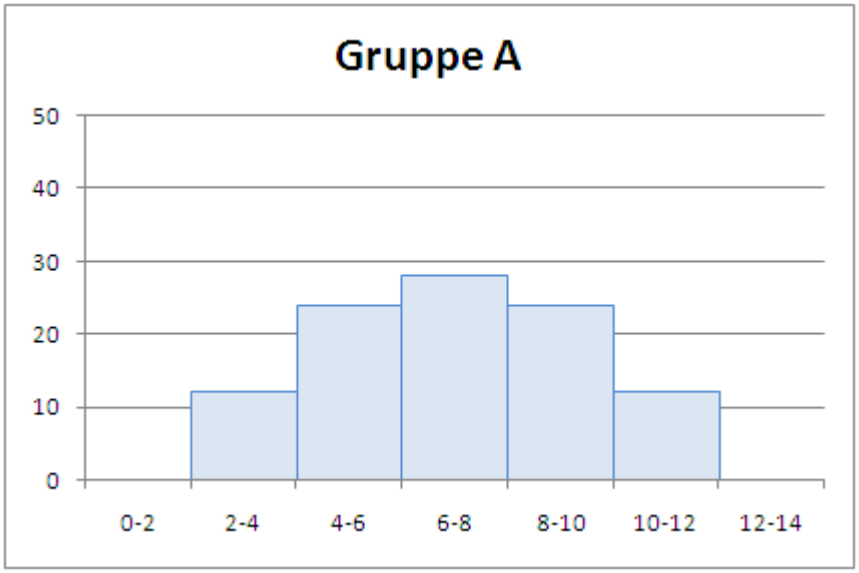
Berechnung von Mittelwert und Streuung

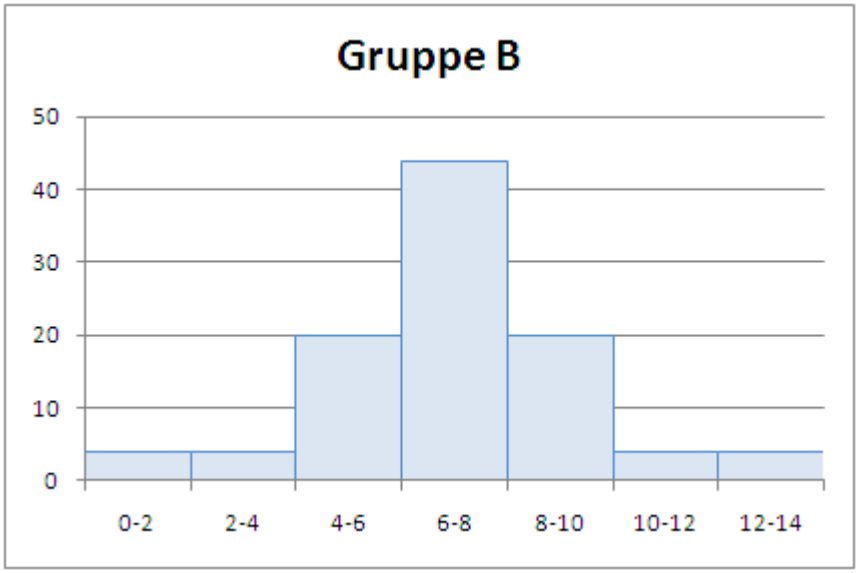
Klasse	m_i	Gruppe A	Gruppe B	Gruppe C	Gruppe A m_i*n_i	Gruppe B m_i*n_i	Gruppe C m_i*n_i	Gruppe A $m_i^2*n_i$	Gruppe B $m_i^2*n_i$	Gruppe C $m_i^2*n_i$
0-2	1	0	4	0	0	4	0	0	4	0
2-4	3	12	4	4	36	12	12	108	36	36
4-6	5	24	20	40	120	100	200	600	500	1000
6-8	7	28	44	24	196	308	168	1372	2156	1176
8-10	9	24	20	20	216	180	180	1944	1620	1620
10-12	11	12	4	8	132	44	88	1452	484	968
12-14	13	0	4	4	0	52	52	0	676	676
		100	100	100	700	700	700	5476	5476	5476
					7	**7**	**7**	**5,76**	**5,76**	**5,76**
								2,4	**2,4**	**2,4**

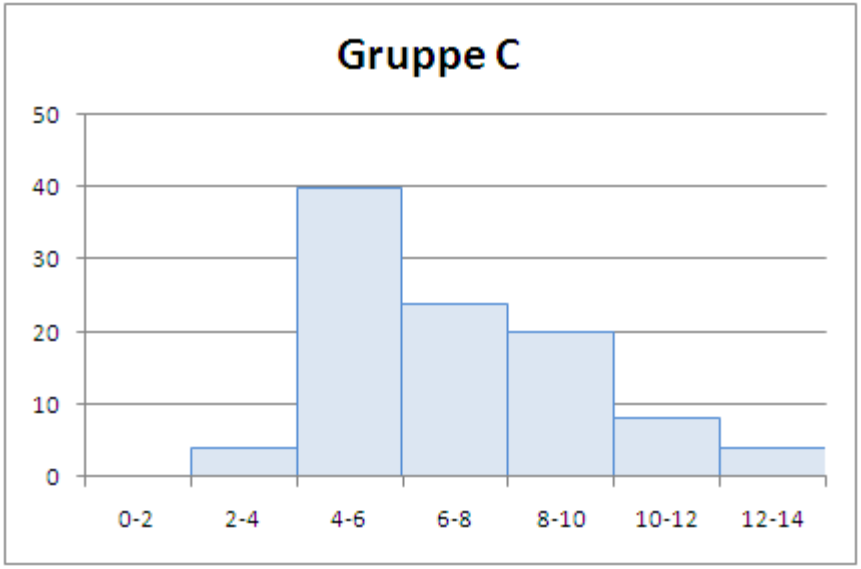
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i m_i \qquad \sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{k} m_i^2 n_i\right) - \overline{x}^2$$

Das arithmetische Mittel ist an allen 3 Gruppen gleich 7.
Die Varianz ist an allen 3 Gruppen gleich 5,76 bzw. ist die Standardabweichung in allen 3 Gruppen 2,4.

Visualisierung der 3 Verteilungen

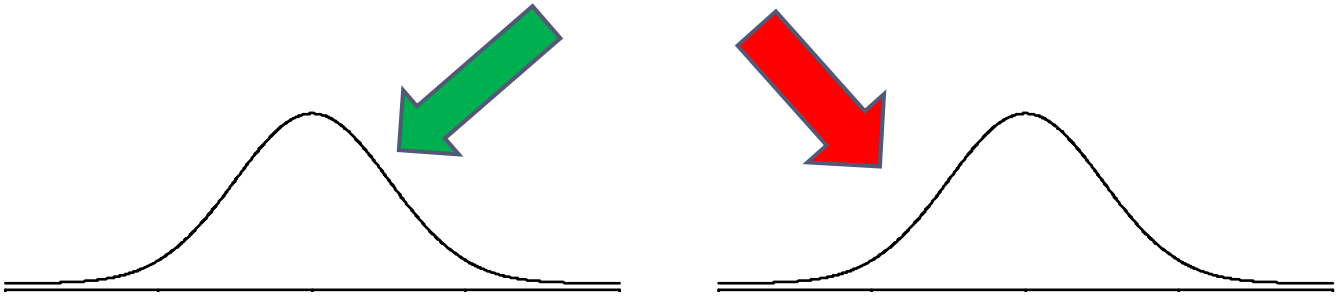






Maßzahlen der Schiefe

- Konvention:
- Positiv ➔ rechtsschief bzw. linkssteil
- Negativ ➔ linksschief bzw. rechtssteil



Unimodale symmetrische Verteilung

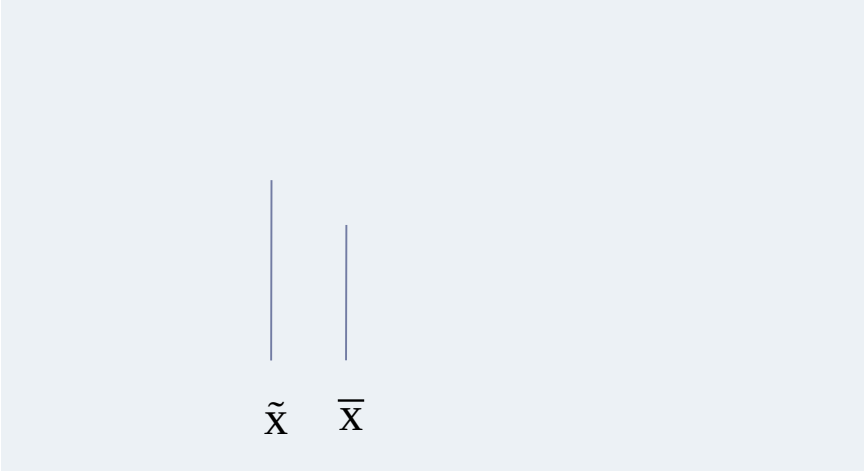
rechtsschief

linksschief

Typische Maße

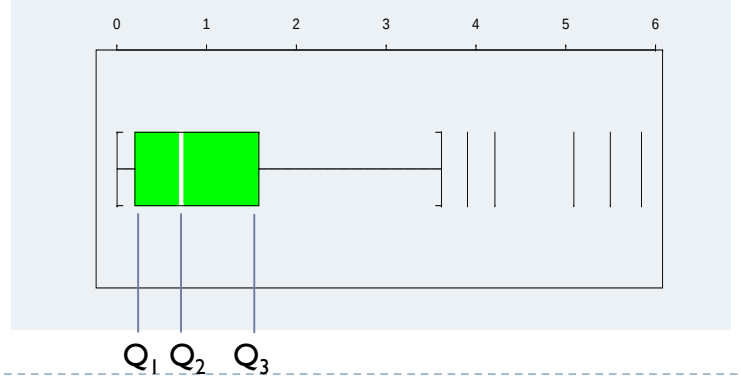
- Pearson‘scher Schiefekoeffizient:

$$S_k = \frac{3 \cdot \left(\overline{x} - \tilde{x} \right)}{\sigma}$$



- Quartilskoeffizient der Schiefe

$$S_q = \frac{\left(Q_3 - Q_2 \right) - \left(Q_2 - Q_1 \right)}{\left(Q_3 - Q_1 \right)}$$



Momentenkoeffizient der Schiefe (Fisher)

- Im Fall von Einzeldaten:

$$S_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}$$

← 3. Zentrales Moment

- Bei klassierten Daten

$$S_m = \frac{\frac{1}{n}\sum_{i=1}^{k}(m_i - \bar{x})^3 \cdot n_i}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{x})^2 \cdot n_i}\right)^3}$$

Wölbung (Kurtosis) einer Verteilung

- Die **Wölbung** oder **Kurtosis** beschreibt die Steilheit bzw. „Spitzigkeit“ einer (eingipfeligen) Häufigkeitsverteilung.
- Verteilungen mit geringer Wölbung streuen relativ gleichmäßig; bei Verteilungen mit hoher Wölbung resultiert die Streuung mehr aus extremen, aber seltenen Ereignissen.
- Um das Ausmaß der Wölbung besser einschätzen zu können, wird sie mit der Wölbung einer Gauß‘schen Glockenkurve (Normalverteilung) verglichen, deren Wölbung konstant 3 ist.
- Der **Exzess** gibt die Differenz der Wölbung einer empirischen Verteilung zur Wölbung einer Gauß‘schen Glockenkurve an.
- Der Exzess ist daher definiert als:
- Exzess = Wölbung -3

Momentenkoeffizient der Wölbung (Kurtosis)

Ermittlung des Exzesses

4. Zentrales Moment

- Im Fall von Einzeldaten:

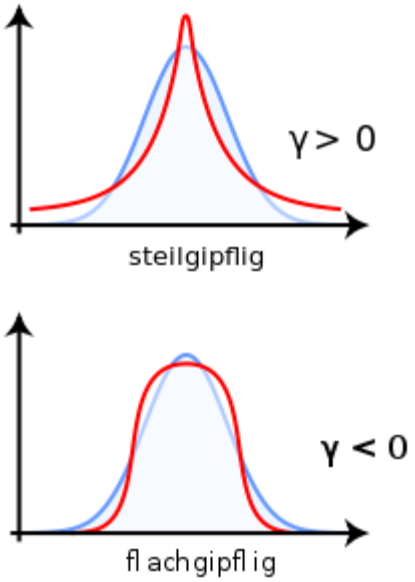
$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3$$

- Bei klassierten Daten

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{k}(m_i - \bar{x})^4 \cdot n_i}{\left[\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{x})^2 \cdot n_i\right]^2} - 3$$

Wölbung (Kurtosis) einer Verteilung

- Verteilungen werden entsprechend ihres Exzesses eingeteilt in:
- $\gamma = 0$: *normalgipflig* oder *mesokurtisch*
  Empirische Wölbung entspricht der Wölbung einer Gauß'schen Glockenkurve (blaue Linie)
- $\gamma > 0$: *steilgipflig* oder *leptokurtisch*.
  Im Vergleich zur Normalverteilung eine spitzere Verteilung, d.h. mit stark ausgeprägten Peak.
- $\gamma < 0$: *flachgipflig* oder *platykurtisch*.
  Im Vergleich zur Normalverteilung abgeflachte Verteilung.



Beispiel (siehe auch XLS)

Klasse	m_i	Gruppe A	Gruppe B	Gruppe C	Schiefe			Wölbung		
0-2	1	0	4	0	0	-864	0	0	5184	0
2-4	3	12	4	4	-768	-256	-256	3072	1024	1024
4-6	5	24	20	40	-192	-160	-320	384	320	640
6-8	7	28	44	24	0	0	0	0	0	0
8-10	9	24	20	20	192	160	160	384	320	320
10-12	11	12	4	8	768	256	512	3072	1024	2048
12-14	13	0	4	4	0	864	864	0	5184	5184
		100	100	100	0	0	960	6912	13056	9216
					0	**0**	**0,694**	**-0,9167**	**0,9352**	**-0,2222**

$$S_m = \frac{\frac{1}{n}\sum_{i=1}^{k}(m_i - \bar{x})^3 \cdot n_i}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{x})^2 \cdot n_i}\right)^3} \qquad \gamma = \frac{\frac{1}{n}\sum_{i=1}^{k}(m_i - \bar{x})^4 \cdot n_i}{\left[\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{x})^2 \cdot n_i\right]^2} - 3$$