

# Statistik

CH.3 - Maßzahlen

SS 2021 | | Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

Ziel der folgenden Maßzahlen ist die Reduktion der Daten auf Kennzahlen, die einen Großteil der *wesentlichen* Informationen der zugrundeliegenden statistischen Variablen enthalten.

- **Lagemaße:** Beschreiben das Zentrum / die Mitte einer Beobachtungsreihe
- **Streuungsmaße:** Beschreiben die Abweichung vom Zentrum einer Häufigkeitsverteilung.
- **Konzentrationsmaße:** beschreiben, wie sich die Summe der Merkmalswerte der Beobachtungsreihe auf die Untersuchungseinheiten verteilt.

**1** Lagemaße

**2** Streuungsmaße

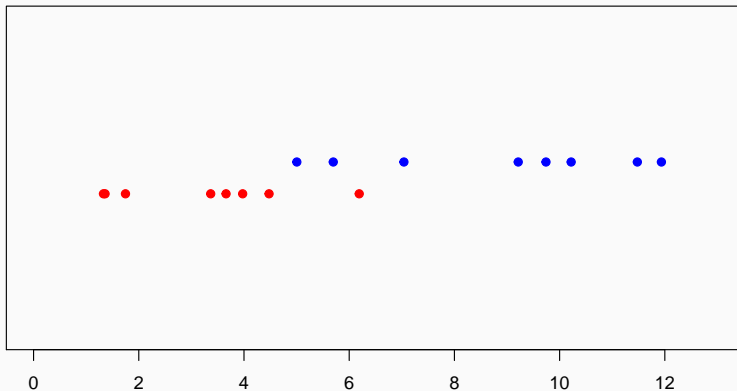
**3** Konzentrationsmaße

Die folgenden Daten bilden das Gewicht (in kg) von zufällig ausgewählten Kugeln aus der Produktion einer Fabrik für Bowlingkugeln ab.

- Handelt es sich bei den vorliegenden Daten um eine Stichprobe oder um eine Grundgesamtheit?
- Welches Skalenniveau weisen die gezeigten Daten auf?
- Wie könnte man die Daten beschreiben?

	i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8
red	1.7471	3.3673	1.3287	6.1906	3.6590	1.3591	3.9749	4.4766
blue	9.7409	11.9352	7.0403	10.2196	11.4776	5.6971	9.2134	5.0044

**Gewicht von 8 roten und 8 blauen Kugeln**



Lagemaß	Symbol	Berechnung
Modus	$\bar{x}_{Modus}$	$h_{Modus} \geq h_j$
Median	$\bar{x}_{Median}$	$x_{\frac{n+1}{2}}$ oder $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$
Quantil	$Q_\alpha$	Wert der Verteilungsfunktion
Arithmetisches Mittel	$\bar{x}$	$\frac{1}{n} \sum_{i=1}^n x_i$
Geometrisches Mittel	$\bar{x}_{geo}$	$\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$
Harmonisches Mittel	$\bar{x}_{harm}$	$\frac{n}{\sum_{i=1}^n 1/x_i}$

**Achtung:** Nicht jede Maßzahl ist für jede Art der *Skalierung* und damit nicht für jede Variable (sinnvoll) bestimmbar.

## Definition: Modus

Der **Modus** oder **Modalwert** ist die häufigste Ausprägung einer Verteilung.

- Der Modus *kann* für beliebig skalierte Variablen bestimmt werden.
- Bei klassierten Daten wird die am häufigsten auftretende Klasse als Modalklasse bezeichnet.
- Der Modus hat keine eigene **R-Funktion**, kann aber mittels absoluten oder relativen Häufigkeiten bestimmt werden.

## Definition: Median

Sind  $x_1 \leq x_2 \leq \dots \leq x_n$  die der Größe nach geordneten Beobachtungswerte eines metrisch skalierten Merkmals  $X$ , ergibt sich der **Median**  $\bar{x}_{Median}$  als

$$\bar{x}_{Median} = \begin{cases} x_{\frac{n+1}{2}} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{falls } n \text{ gerade} \end{cases}$$

- Der Median wird auch als Zentralwert bezeichnet.
- Der Median teilt die Daten in zwei gleich Große Hälften.
- Kann für metrische- und ordinalskalierte Merkmale verwendet werden.
- Ist *robust* gegenüber Ausreißern.
- **R-Funktion:** `median()`



## Definition: Quantil

Das  $\alpha$ -Quantil eines Merkmals ist der Wert, unterhalb dessen ein vorgegebener Anteil  $\alpha$  aller Beobachtungswerte der Verteilung liegt. Dieser Wert ergibt sich aus der (empirischen) Verteilungsfunktion  $S()$ .

$$S(Q_\alpha) = \alpha$$

- Quantile sind Verallgemeinerungen des Medians, dieser ist  $Q_{0.5}$ .
- Einige Gruppen von Quantilen haben spezielle Namen
  - Quartile:  $Q_{0.25}$ ,  $Q_{0.5}$ ,  $Q_{0.75}$
  - Percentile:  $Q_{0.01}$ ,  $Q_{0.02}$ ,  $Q_{0.03}$ ,  $Q_{0.04}$ , . . .
- **R-Funktion:** `quantile()`

## Definition: Arithmetisches Mittel

Sind  $x_1, \dots, x_n$  die Beobachtungswerte eines metrisch skalierten Merkmals  $X$ , so errechnet sich das **arithmetische Mittel** durch

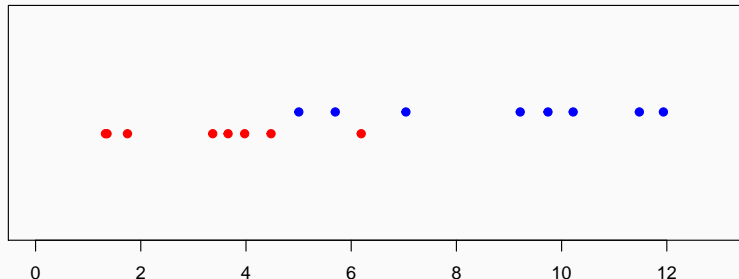
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Das arithmetische Mittel ist nur für metrisch skalierte Daten definiert!
- Ist eine Maßzahl die empfindlich gegenüber Ausreißern ist.
- Das gewichtete arithmetische Mittel erlaubt die Bestimmung des arithmetischen Mittels für klassierte Daten.
- **R-Funktion:** `mean()`

## Beispiel: Median und arithmetisches Mittel für die roten Kugeln

	i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8
red	1.7471	3.3673	1.3287	6.1906	3.6590	1.3591	3.9749	4.4766
blue	9.7409	11.9352	7.0403	10.2196	11.4776	5.6971	9.2134	5.0044

**Gewicht von 8 roten und 8 blauen Kugeln**



# Beispiel: Median und arithmetisches Mittel für die roten Kugeln

```
# Ausgabe der Daten
```

```
red
```

```
## [1] 1.747092 3.367287 1.328743 6.190562 3.659016 1.359063 3.974858 4.476649
```

```
# Arithmetisches Mittel
```

```
mean(red)
```

```
## [1] 3.262909
```

```
## Median
```

```
median(red)
```

```
## [1] 3.513151
```

```
# Zusammenfassung wesentlicher Lagemaße
```

```
summary(red)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.329   1.650   3.513   3.263   4.100   6.191
```

- Es gibt zahlreiche spezialisierte Mittelwerte, wie das **geometrische Mittel**  $\bar{x}_{geo}$  und das **harmonische Mittel**  $\bar{x}_{harm}$ . Welcher Mittelwert genutzt werden muss, hängt von den zugrundeliegenden Daten ab.
- Ziel der Mittelwertbildung ist, die durchschnittliche *Gesamtwirkung* von  $n$  meist unterschiedlichen Werten mit einem einzigen Wert zu beschreiben.

**Geometrisches Mittel:**  $\bar{x}_{geo} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$

**Harmonisches Mittel:**  $\bar{x}_{harm} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$

- **Anwendung:** Geometrische Mittelwerte eignen sich für Wachstumsraten, harmonische Mittelwerte für Geschwindigkeiten.

# Beispiel: Geometrisches Mittel

```
# Das Wertpapier-Beispiel (Bitcoin) aus der Einführung liefert die folgenden  
# Wertveränderungen (returns) für ersten 3 Jahre des Assets.  
ret <- c(.12, .07, .01)
```

```
# Geometrisches Mittel  
mean_gm <- prod(1 + ret)^(1/length(ret))  
mean_gm
```

```
## [1] 1.065715
```

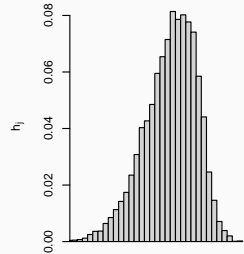
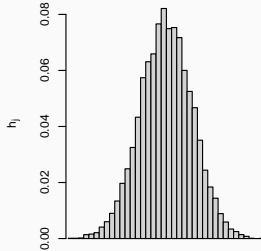
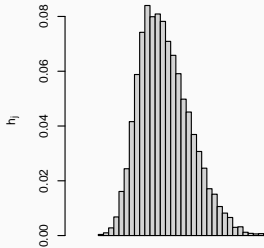
```
# Probe  
(100 * prod(1 + ret)) # Wert nach 3 Perioden bei 100 Euro Startwert
```

```
## [1] 121.0384
```

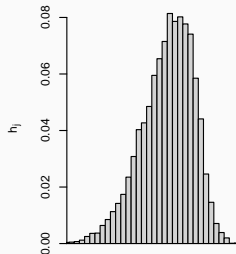
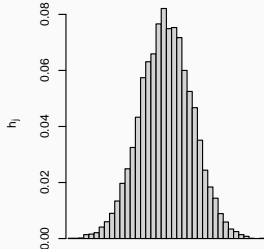
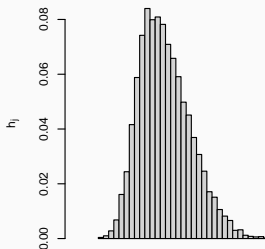
```
(100 * mean_gm^3) # Wert nach 3 Perioden berechnet mit x_{geom}
```

```
## [1] 121.0384
```

- Wo liegen  $\bar{x}$ ,  $\bar{x}_{Median}$  und  $\bar{x}_{Modus}$  bei den nachfolgend gezeigten Häufigkeitsverteilungen?



- Wo liegen  $\bar{x}$ ,  $\bar{x}_{Median}$  und  $\bar{x}_{Modus}$  bei den nachfolgend gezeigten Häufigkeitsverteilungen?



- Rechtsschiefe Häufigkeitsverteilung:  $\bar{x} > \bar{x}_{Median} > \bar{x}_{Modus}$
- Linksschiefe Häufigkeitsverteilung:  $\bar{x} < \bar{x}_{Median} < \bar{x}_{Modus}$
- Symmetrische Häufigkeitsverteilung:  $\bar{x} = \bar{x}_{Median} = \bar{x}_{Modus}$

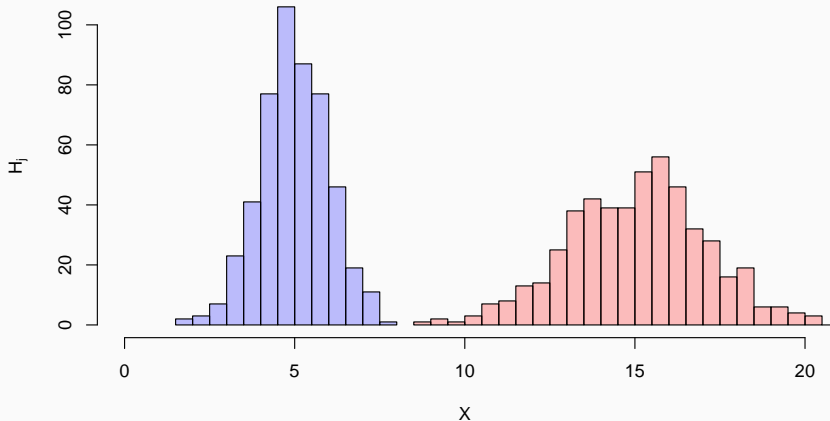


**1** Lagemaße

**2** Streuungsmaße

**3** Konzentrationsmaße

## Häufigkeitsverteilungen



Lagemaß	Symbol	Berechnung
Spannweite	$R$	$x_{max} - x_{min}$
Interquartilsabstand	$IQR$	$Q_{0.75} - Q_{0.25}$
(empirische) Varianz	$s^2$	$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Standardabweichung	$s$	$\sqrt{s^2}$
Variationskoeffizient	$V$	$s/\bar{x}$

**Achtung:** Nicht jede Maßzahl ist für jede Art der *Skalierung* und damit nicht für jede Variable (sinnvoll) bestimmbar.

## Definition: Spannweite

Die Breite eines Streubereichs nennt man Spannweite  $R$ . Sie ergibt sich aus dem Maximum und Minimum der Daten.

$$R = x_{\max} - x_{\min}$$

- Nachteil: Nur zwei *extreme* Werte gehen in die Berechnung ein, der Großteil der Daten bleibt ungenutzt.
- Die Spannweite hat keine eigene **R-Funktion**, kann aber einfach mittels `max()` und `min()` berechnet werden.

## Definition: Interquartilsabstand

Der **Quartilsabstand** gibt die Größe des Bereiches zwischen dem oberen und dem unteren Quartil einer Verteilung an, in dem die mittleren 50% der Beobachtungen fallen.

$$IQR = Q_{0.75} - Q_{0.25}$$

- Zwischen dem oberen und dem unteren Quartil liegen 50% der Beobachtungen.
- Kann auch sinnvoll für ordinalskalierte Merkmale bestimmt werden.
- Ist *robust* in dem Sinne, dass die *IQR* weitgehend unempfindlich gegenüber Ausreißern ist.
- **R-Funktion:** `IQR()`

## Definition: Varianz

Der **Quartilsabstand** gibt die Größe des Bereiches zwischen dem oberen und dem unteren Quartil einer Verteilung an, in dem die mittleren 50% der Beobachtungen fallen.

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

- Es gilt immer  $s^2 \geq 0$
- Wird unterschiedlich für die Population und für die Grundgesamtheit berechnet.
- Grundidee: Einbezug aller Abweichungen vom Mittelwert
- Beobachtungen die weit von  $\bar{x}$  entfernt liegen werden überproportional stark gewichtet.
- **R-Funktion:** `var()`

## Definition: Standardabweichung

Die **Standardabweichung** ist die Wurzel aus der Varianz.

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

- Weißt die gleiche Maßeinheit wie die Daten auf
- Ist i.d.R. einfacher zu interpretieren als die Varianz.
- **R-Funktion:** `sd()`

Berechnung der Varianz der blauen Kugeln aus dem Eingangsbeispiel.

$i$	$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1.7471	-1.5158	2.2977
2	3.3673	0.1044	0.0109
3	1.3287	-1.9342	3.7410
4	6.1906	2.9277	8.5712
5	3.6590	0.3961	0.1569
6	1.3591	-1.9038	3.6246
7	3.9749	0.7119	0.5069
8	4.4766	1.2137	1.4732

$$n = 8 \quad \bar{x} = 3.2629 \quad \sum(x_i - \bar{x}) = 0 \quad \sum(x_i - \bar{x})^2 = 20.3823 \quad s^2 = 2.9118$$



## Definition: Variationskoeffizient

Der **Variationskoeffizient** ist der Quotient aus Standardabweichung und arithmetischem Mittel.

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

- Ist dimensionslos und vergleichbar
- Ist i.d.R. einfacher zu interpretieren als die Varianz.
- Der Variationskoeffizient hat keine eigene **R-Funktion**, kann aber einfach mittels `sd()` und `mean()` berechnet werden.

# Beispiel: Streuungsmaße

```
# Ausgabe der Daten
```

```
blue
```

```
## [1]  9.740948 11.935249  7.040281 10.219562 11.477642  5.697140  9.213390
```

```
## [8]  5.004441
```

```
# Spannweite
```

```
max(blue) - min(blue)
```

```
## [1] 6.930808
```

```
## Varianz
```

```
var(blue)
```

```
## [1] 6.742995
```

```
# Interquartilsabstand
```

```
IQR(blue)
```

```
## [1] 3.829586
```

```
# Variationskoeffizient
```

```
sd(blue) / sd(blue)
```

**1** Lagemaße

**2** Streuungsmaße

**3** Konzentrationsmaße

## Definition: Konzentration

Man spricht von Konzentration oder Ungleichheit, falls zu einem bestimmten Zeitpunkt ein relativ kleiner Anteil der Merkmalsträger einen hohen Anteil an der Summe der Merkmalswerte besitzt.

- Konzentration bzw. Ungleichheitsdiskussionen findet man häufig im Kontext von Einkommen oder Vermögen.
- Beispiele: In Deutschland besitzen 10% der Bevölkerung des Vermögens.

## Definition: Lorenzkurve

Der Polygonzug durch die Punkte  $P_0 = (0, 0)$  und  $P_j = (k_j, l_j)$  mit  $j = 1, \dots, q$  heißt **Lorenzkurve**.

$$k_j = \sum_{i=1}^j \frac{H_j}{n} = \sum_{i=1}^j h_j \quad l_j = \frac{\sum_{i=1}^j a_i H_i}{\sum_{i=1}^q a_i H_i}$$

- Die Lorenzkurve verläuft durch die Punkte  $(0, 0)$  und  $(1, 1)$
- Die Lorenzkurve verläuft immer **unterhalb** der Winkelhalbierenden.
- Die Lorenzkurve ist winkelhalbierend, wenn alle Merkmalsausprägungen gleich häufig vorkommen. Dann liegt keine Konzentration vor. Je weiter die Lorenzkurve sich von der Winkelhalbierenden entfernt, desto größer ist die Ungleichheit.
- **R-Funktion:**  $L_c()$  aus dem Zusatzpaket `ineq`

Wir betrachten vereinfachend die Einkommensverteilungen der folgenden drei sehr kleinen Länder.

```
A <- c(1000, 3000, 4000, 4000, 8000)
B <- c(2000, 2000, 4000, 8000)
C <- c(1000, 2000, 5000, 8000)
```

# Beispiel

```
# Lorenzkurve für A [ohne Punkt (0,0) ]  
A <- sort(A)  
H <- table(A)  
n <- length(A)  
h <- H/n  
k <- cumsum(h)  
l <- cumsum(H * unique(A)) / sum(A)
```

j	$a_j$	$k_j$	$l_j$
1	1000	0.2	0.05
2	3000	0.4	0.20
3	4000	0.8	0.60
4	8000	1.0	1.00

## Definition: Gini Koeffizient

Das Doppelte der Fläche zwischen der Lorenzkurve und der Winkelhalbierenden heißt **Gini-Koeffizient**  $G$  und wird als Konzentrationsmaß einer Häufigkeitsverteilung verwendet.

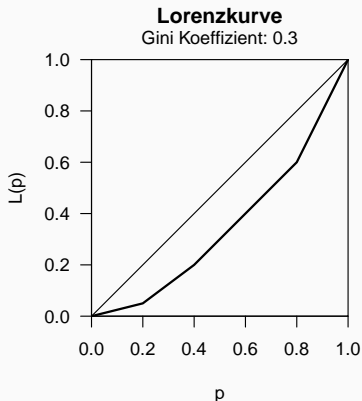
$$G = \sum_{i=1}^n (k_i + k_{i-1})(l_i + l_{i-1}) - 1$$

- Die Lorenzkurve verläuft durch die Punkte (0, 0) und (1, 1)
- Die Lorenzkurve verläuft immer **unterhalb** der Winkelhalbierenden.
- Die Lorenzkurve ist winkelhalbierend, wenn alle Merkmalsausprägungen gleich häufig vorkommen. Dann liegt keine Konzentration vor. Je weiter die Lorenzkurve sich von der Winkelhalbierenden entfernt, desto größer ist die Ungleichheit.
- **R-Funktion:** `Gini()` aus dem Zusatzpaket `ineq`



# Lorenzkurve mit Gini-Koeffizient

```
library(ineq); par(pty="s") # Laden des Zusatzpaketes ineq + Plot quadratisch
plot(Lc(A), main="Lorenzkurve")
subtitle <- paste("Gini Koeffizient:", Gini(A))
mtext(subtitle, line=0.5)
```



- Welche Lage- und Streuungsparameter eignen sich für ordinalskalierte Merkmale? Welche sind für nominalskalierte Merkmale geeignet?
- Welche Streuungsmaße berücksichtigen nur einzelne Beobachtungswerte der Häufigkeitsverteilung?
- Wie macht sich eine vollkommene Gleichheit in der Einkommensverteilung eines Landes in der Lorenzkurve bemerkbar? Wie groß ist dann der GINI-Koeffizient?

- Welche Lage- und Streuungsparameter eignen sich für ordinalskalierte Merkmale? Welche sind für nominalskalierte Merkmale geeignet?
  - **Ordinal:** Median, Quantile, Modus, Quartilsabstände.
  - **Nominal:** Modus.
- Welche Streuungsmaße berücksichtigen nur einzelne Beobachtungswerte der Häufigkeitsverteilung?
  - **Einzelne Beobachtungswerte:** Spannweite
  - **Alle Beobachtungswerte:** alle weiteren die vorgestellt wurden.
- Wie macht sich eine vollkommene Gleichheit in der Einkommensverteilung eines Landes in der Lorenzkurve bemerkbar? Wie groß ist dann der GINI-Koeffizient?
  - Winkelhalbierende,  $G = 0$