Statistik - univariate Statistik
2.2 Maßzahlen der Verteilung

Dr. Wolfgang Böhme

Oktober 2010

2. Univariate Verteilungsanalyse - Gliederung

- 2.1 Häufigkeitsverteilung
- **2.2 Maßzahlen der Verteilung**
- 2.3 Konzentrationsanalyse

Intention / Gliederung

Zielstellung:

- Beschreibung von Datensätzen mit wenigen statistischen Kenngrößen
- Weitere Verdichtung der Verteilungsdarstellung

Gruppen von Maßzahlen

- 2.2.1 **Lagemaße**
- 2.2.2 Streuungsmaße
- 2.2.3 Formenmaße
- 2.2.4 lineare Transformationen

2.2.1 Lagemaße - Gliederung

- **2.2.1.1 Einführung**
- 2.2.1.2 Modus
- 2.2.1.3 Median
- 2.2.1.4 Quantile
- 2.2.1.5 Beispiel: Historische Simulation
- 2.2.1.6 Arithmetisches Mittel

Lagemaße Intention

Ausgangspunkt: **Urliste** für univariate Statistik

$$X_{(n,1)} = X_{(n)} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

mit

$$\begin{aligned} x_{i1} = x_i \quad &- \quad \textit{Auspraegung am } i-\textit{ten Merkmalstraeger} \\ i \quad &- \quad \textit{Index Merkmalstraeger } (i = 1, 2, ..., n) \end{aligned}$$

Zielstellung:

Beschreibung der Urliste durch wenige statistische Merkmale, die die Lage der Merkmalsausprägungen X_i auf der Achse der Merkmalsausprägungen (kurz $X - \textit{Achse}$) charakterisieren.

Lagemaße - grundsätzliches Vorgehen

Forderungen an den Wert des Lagemaßes:

- möglichst zentrale Lage
- Geringer Abstand zu den beobachteten Merkmalsausprägungen x_i

$\Rightarrow$
zentraler Begriff: **Abstand** bzw. **Abstandsfunktion**
Bezeichnung:

$$d(x_i, a) \quad - \quad \textit{Abstandsfunktion}$$

$$a \quad - \quad \textit{Lagemass}$$

$$x_i \quad - \quad \textit{Merkmalsauspraegung} \ (i = 1.2, ..., n)$$

Anforderung: $F(d(x_i, a)) \rightarrow min$

Abstandsfunktionen

Definition: Eine Abbildung d: $X\,x\,X \rightarrow \Re$ heißt **Metrik**, wenn für beliebige Elemente x, y und z von X die folgenden axiomatischen Bedingungen erfüllt sind:

$$\begin{aligned} d(x,x) &= 0 \\ d(x,y) &= 0 \Rightarrow x = y \\ d(x,y) &= d(y,x) \\ d(x,y) &\leq d(x,z) + d(z,y) \end{aligned}$$

typische Abstandsfunktionen (Metriken) sind:

$$\begin{aligned} d(x,y) &= |x-y| \; x,y \in \Re \\ d(x,y) &= \sqrt{\sum_{i=1}^{n}(x_i - a)^2}, \;\; x,a \in \Re \; x\Re x \ldots \Re \end{aligned}$$

Abstandsfunktionen -univarate Statistik

typische Abstandsfunktionen (Metriken) in der univariaten Statistik können sein:

$$
\begin{aligned}
d(x,y) &= |x-y| \; x,y \in \Re \\
d(x,y) &= (x-a)^2, \;\; x,y \; in \; \Re \\
d(x,y) &= \sqrt{(x-a)^2}, \;\; x,y \; in, \Re \\
d(x,a) &= \begin{cases} 0 & falls \; x = a \\ 1 & falls \; x \neq a \end{cases}
\end{aligned}
$$

Abstandsfunktionen - Beispiel

Beispiel: Vorlesungsbeispiel 1
Sheet: Abstandsfunktion_base

Aufgaben: Bestimmen von Abstaenden für metrische Variablen

2.2.1 Lagemaße - Gliederung

- 2.2.1.1 Einführung
- **2.2.1.2 Modus**
- 2.2.1.3 Median
- 2.2.1.4 Quantile
- 2.2.1.5 Beispiel: Historische Simulation
- 2.2.1.6 Arithmetisches Mittel

Modus - Definition

Definition: Jede disjunkte Merkmalsausprägung ξ_{j*}, deren absolute Häufigkeit

$$n_{j*} = max\{n_1, n_2, ..., n_n\}$$

oder deren relative Häufigkeit

$$p_{j*} = max\{p_1, p_2, ..., p_n\}$$

erfüllt wird als **Modus** (Modalwert) bezeichnet.

Schreibweise: X_M

Hinweis: wenn speziell nicht darauf hingewiesen wird, werden im weiteren die disjunkte Merkmalsausprägungen ξ_i kurz als Merkmalsausprägungen genannt und mit x_i bezeichnet.

Modus - Bemerkungen

1. für alle Skalierungen anwendbar
2. der Modalwert ist die Merkmalsausprägung in einem Datensatz beliebig skalierter Merkmale, die am häufigsten beobachtet wird,
3. der Modalwert entspricht im Säulendiagramm der Ausprägung mit der größten Säule,
4. der Modalwert ist nicht zwingend eindeutig,
5. verwendetes Abstandsmaß:

$$d(x_i, a) = \begin{cases} 0 & falls\, x_i = a \\ 1 & falls\, x_i \neq a \end{cases}$$

Vorlesungsbeispiel 1

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Summe		50	1

$$n_{j^*} = max\{n_1, n_2, ..., n_n\} = 23 \Rightarrow X_M = 4$$

oder deren relative Häufigkeit

$$p_{j^*} = max\{p_1, p_2, ..., p_n\} = 0,46 \Rightarrow X_M = 4$$

Modus mit Intervalbildung - Definition

Ausgangspunkt: Interval $I_l = (c_{l-1}, c_l)$

Problemstellung: Die am häufigsten auftretende Klasse läßt sich leicht bestimmen, aber welcher Wert ist als Modalwert zu verwenden?

2 Näherungsverfahren:

1. Klassenmitte des Intervalls $I_l \Rightarrow$

$$x_M \approx \frac{1}{2}(c_{l-1} + c_l)$$

2. Modalwert unter der Annahme der Gleichverteilung in den Intervallen I_{l-1}, I_l und $I_{l-1} \Rightarrow$

$$x_M \approx c_{l-1} + \frac{p_l^D - p_{l-1}^D}{(p_l^D - p_{l-1}^D) + (p_l^D - p_{l+1}^D)} \cdot \Delta_l$$

Vorlesungsbeispiel 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Die am häufigsten auftretende Klasse: $[0; 2)$

$$
\begin{aligned}
x_M &\approx \frac{1}{2}(c_{l-1} + c_l) = 1 \\
x_M &\approx c_{l-1} + \frac{p_l^D - p_{l-1}^D}{(p_l^D - p_{l-1}^D) + (p_l^D - p_{l+1}^D)} \cdot \Delta_l \\
x_M &\approx 0 + \frac{0,29 - 0}{(0,29 - 0) + (0,29 - 0,11)} \cdot 2 = 1,234
\end{aligned}
$$

2.2.1 Lagemaße - Gliederung

- 2.2.1.1 Einführung
- 2.2.1.2 Modus
- **2.2.1.3 Median**
- 2.2.1.4 Quantile
- 2.2.1.5 Beispiel: Historische Simulation
- 2.2.1.6 Arithmetisches Mittel

Median - Definition

gegeben sei die Rangwertreihe eines mindest ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ..., \ x_n$$

Definition: Der **Median** $\widetilde{x}$ des ordinalen Datensatzes ist bei einer ungeraden Anzahl von Beobachtungen:

$$\widetilde{x} = x_{\frac{n+1}{2}}$$

Interpretation: mindestens 50 % der Beobachtungen sind kleiner oder gleich als $\widetilde{x}$ und mindestens 50 % aller Beobachtungen sind größer oder gleich $\widetilde{x}$

Median - Definition

gegeben sei die Rangwertreihe eines mindest ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ..., \ x_n$$

Definition: Der **Median** $\widetilde{x}$ des ordinalen Datensatzes ist bei einer geraden Anzahl von Beobachtungen

$$\begin{aligned} \widetilde{x} &\in \left\{ x_{\frac{n}{2}}, x_{\frac{n}{2}+1} \right\} \\ \textit{hauefig} & \\ &= \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \end{aligned}$$

Interpretation: mindestens 50 % der Beobachtungen sind kleiner oder gleich als $\widetilde{x}$ und mindestens 50 % aller Beobachtungen sind größer oder gleich $\widetilde{x}$

Vorlesungsbeispiel 1

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Summe		50	1

Die Anzahl der Beobachtungen ist 50 (eine gerade Zahl)

$$\begin{aligned} \widetilde{x} &\in (x_{\frac{n}{2}}, x_{\frac{n}{2}+1}) \\ \widetilde{x} &\in (x_{25}, x_{26}) \\ \widetilde{x} &\in (4,4) = 4 \end{aligned}$$

hauefig

$$\widetilde{x} = \frac{1}{2}(x_{25} + x_{26}) = \frac{1}{2}(4+4) = 4$$

Median - Verallgemeinerung

gegeben sei die Rangwertreihe eines ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ..., \ x_n$$

p^* - vorgebener Wert (beim Median $p^* = 0,5$)

Interpretation: mindestens $p^* \cdot 100$ % der Beobachtungen sind kleiner oder gleich als $\widetilde{x}$ und mindestens $p^* \cdot 100$ % aller Beobachtungen sind größer oder gleich $\widetilde{x}$

Aufteilung der Rangwertreihe in 2 Teile:

$$\underbrace{x_1 \leq x_2 \leq \ ..., \ x_k}_{p^* \cdot 100 \ \%} \quad \underbrace{\leq x_{k+1} \leq x_{k+2} \leq \ ..., \ x_k}_{(1-p^*) \ \cdot 100 \ \%}$$

Diese Verallgemeinerung des Medians wird als **Quantil** bezeichnet

2.2.1 Lagemaße - Gliederung

- 2.2.1.1 Einführung
- 2.2.1.2 Modus
- 2.2.1.3 Median
- **2.2.1.4 Quantile**
- 2.2.1.5 Beispiel: Historische Simulation
- 2.2.1.6 Arithmetisches Mittel

Quantil - Definition

gegeben sei die Rangwertreihe eines mindest ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Definition: Es sei X ein mindestens ordinal skaliertes, zahlenmäßig erfaßtes und geordnetes Merkmal einer statistischen Gesamtheit und p, mit $0 < p < 1$, bezeichne einen vorgegebenen Wert der relativen Summenhäufigkeit (der empirischen Verteilungsfunktion), dann heißt der kleinste Wert x für den gilt $F(x) \geq p$, **Quantil der Ordnung p**

Bezeichnung: x_p

formal:

$$x_p = min\,(x \mid F(x) \geq p)$$

Interpretation: $p \cdot 100$ % aller Beobachtungen haben einen Wert, der nicht größer ist als (x_p)
oder der maximale Wert, den $p \cdot 100$ % aller Beobachtungen annehmen können, beträgt (x_p).

Quantil - Bestimmung aus Einzelwerten

gegeben sei die Rangwertreihe eines mindest ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ... \leq x_n$$

Bestimmen des Wertes $n \cdot p$

- wenn $g = n \cdot p$ - ganzahlig dann $x_p = x_g$
- sonst Wichtung der benachbarten ganzzahligen Werte x_g und x_{g+1} :

$$\begin{aligned} \textit{wenn } g &< n \cdot p \leq g + 1 \\ \textit{dann } x_p &= (1 - r) \cdot x_g + r \cdot x_{g+1} \\ \textit{mit } n \cdot p &= g + r \ (0 < r < 1 \ , r - \textit{nichtganzzahliger Bestandteil von } n \cdot p) \end{aligned}$$

Quantil - Bestimmung aus Einzelwerten

Aufgabe: Bestimmen des 0,2 - Quantils (p =0,2) bzw. des 20-Prozent-Quantils für Vorlesungsbeispiel 1

inhaltliche Fragestellung: Wie hoch ist die maximale Belegungszahl der Wohnung, die von 20 Prozent aller Wonungen erreicht wird?

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Summe		50	1

Die Anzahl der Beobachtungen ist 50 (n =50)

Quantil - Bestimmung aus Einzelwerten

Aufgabe: Bestimmen des 0,20 - Quantils (p =0,20) für Vorlesungsbeispiel 1

inhaltliche Fragestellung: Wie hoch ist die maximale Belegungszahl der Wohnung, die von 20 Prozent aller Wohnungen erreicht wird?

Bestimmen des Quantils:

$$\begin{aligned} g &= n \cdot p = 50 \cdot 0,20 = 10 \rightarrow \textit{ganzzahlige „Nummer der Beobachtung“} \\ x_{20} &= 3 \end{aligned}$$

Interpretation:

1. wenigstens 20 Prozent aller Wohnungen sind mit nicht mehr als 3 Personen belegt
2. Die maximale Belegungszahl, die von 20 Prozent aller Wohnungen erreicht wird, ist 3.
3. höchstens 80 Prozent aller Wohnungen ist mit nicht weniger als 3 Personen belegt.

Quantil - Bestimmung aus Einzelwerten -Wohnungsbelegung-

Bestimmen des 5-Prozent-Quantils:

$$
\begin{aligned}
g + r &= n \cdot p = 50 \cdot 0,05 = 2,5 \ \rightarrow \textit{nichtganzzahlig „Nummer der Beobachtu} \\
g &= 2 \\
r &= 0,5 \\
x_g &= x_2 = 2 \\
x_{g+1} &= x_3 = 2 \\
x_p &= 0,5x_2 + 0,5x_3 = 2
\end{aligned}
$$

Interpretation:

Die maximale Belegungszahl, die von 5 Prozent aller Wohnungen erreicht wird, ist 2.

Quantil - Bestimmung aus Einzelwerten -Telefongespräche-

Vorlesungsbeispiel 2 (Telefongespräche)
Bestimmen des 1,25 - Prozent - Quantils:

$$
\begin{aligned}
g &= n \cdot p = 100 \cdot 0,0125 = 1,25 \ \rightarrow \textit{nichtganzzahlig „Nummer der Beobachtung“} \\
g &= 1 \\
r &= 0,25 \ \textit{nichtganzzahligerBestandteil} \\
x_1 &= 0,04 \\
x_2 &= 0,09 \\
x_p &= (1-r) \cdot x_g + r \cdot x_{g+1} \\
x_{0,0125} &= (1-0,25) \cdot x_1 + 0,25 \cdot x_2 \\
x_{0,0125} &= 0,75 * 0,04 + 0,25 * 0,09 = 0,0525
\end{aligned}
$$

Interpretation:

Die maximale Dauer eines Telefongespräches, das von 1,25 Prozent aller Telefongespräche erzielt wird, ist 0,0525 min..

Vorlesungsbeispiel 1

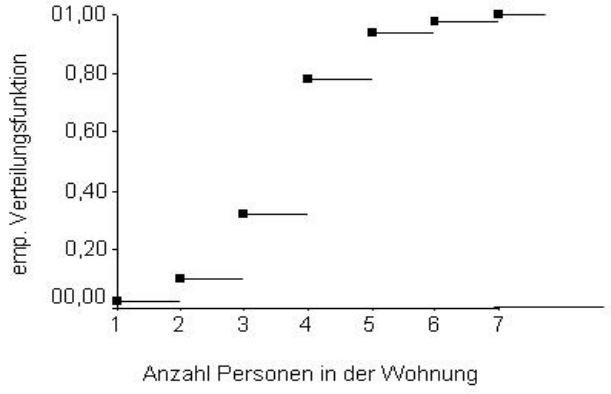
Urliste **aufsteigend sortierte Urliste**

$$\begin{pmatrix} x_1 \\ x_2 \\ . \\ x_i \\ . \\ x_n \end{pmatrix} \Rightarrow \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ . \\ x_{(i)} \\ . \\ x_{(n)} \end{pmatrix}$$

Häufigkeitsverteilung, tabellarisch

j	ξ_j	n_j	p_j	H_j	F_j
1	1	1	0,02	1	0,02
2	2	4	0,08	5	0,10
3	3	11	0,22	16	0,32
4	4	23	0,46	39	0,78
5	5	8	0,16	47	0,94
6	6	2	0,04	49	0,98
m=7	7	1	0,02	50	1,00
Σ		n= 50	1,00		

Graph der empirischen Verteilungsfunktion

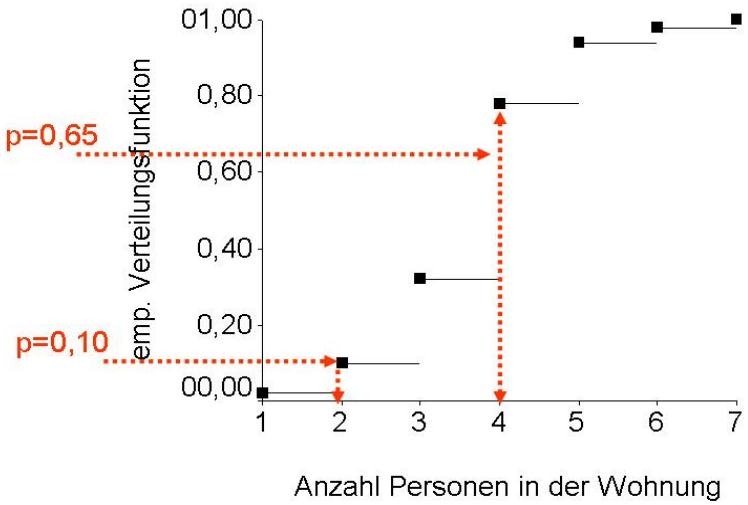


empirische Verteilungsfunktion im analytischen Ausdruck

$$F(x) = \begin{cases} 0 & \text{falls} \quad x < 1 \\ 0{,}02 & \text{falls} \quad 1 \le x < 2 \\ 0{,}10 & \text{falls} \quad 2 \le x < 3 \\ 0{,}32 & \text{falls} \quad 3 \le x < 4 \\ 0{,}78 & \text{falls} \quad 4 \le x < 5 \\ 0{,}94 & \text{falls} \quad 5 \le x < 6 \\ 0{,}98 & \text{falls} \quad 6 \le x < 7 \\ 1{,}00 & \text{falls} \quad x \ge 7 \end{cases}$$

Quantil - Bestimmen aus empirischer Verteilungsfunktion

Ermittlung Quantil unter Verwendung Treppenfunktion



Weitere Quantile: 0,20-Quantil

Quantil - Bestimmung für Intervalle

Ausgangspunkt: Interval $I_l = (c_{l-1}, c_l]$

Quantil : $x_p \in I_l \;\Rightarrow\; c_{l-1} < x_p \leq c_l$

$$\begin{aligned} x_p &\approx c_{l-1} + \frac{p - F_{l-1}}{p_l^D} \\ mit \quad & F_{l-1} < p < F_l \end{aligned}$$

Vorlesungsbeispiel 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Berechnen des 0,7-Quantils:

$$
\begin{aligned}
g &= n \cdot p = 100 \cdot 0,7 = 70 \\
x_{0,7} &\in (2, 4] = (c_1, c_2] \\
F_1 &= 0,58 \\
F_2 &= 0,80 \\
x_p &\approx c_{l-1} + \frac{p - F_{l-1}}{p_l^D} = 2 + \frac{0,7 - 0,58}{0,11} = 3,09
\end{aligned}
$$

Vorlesungsbeispiel 2

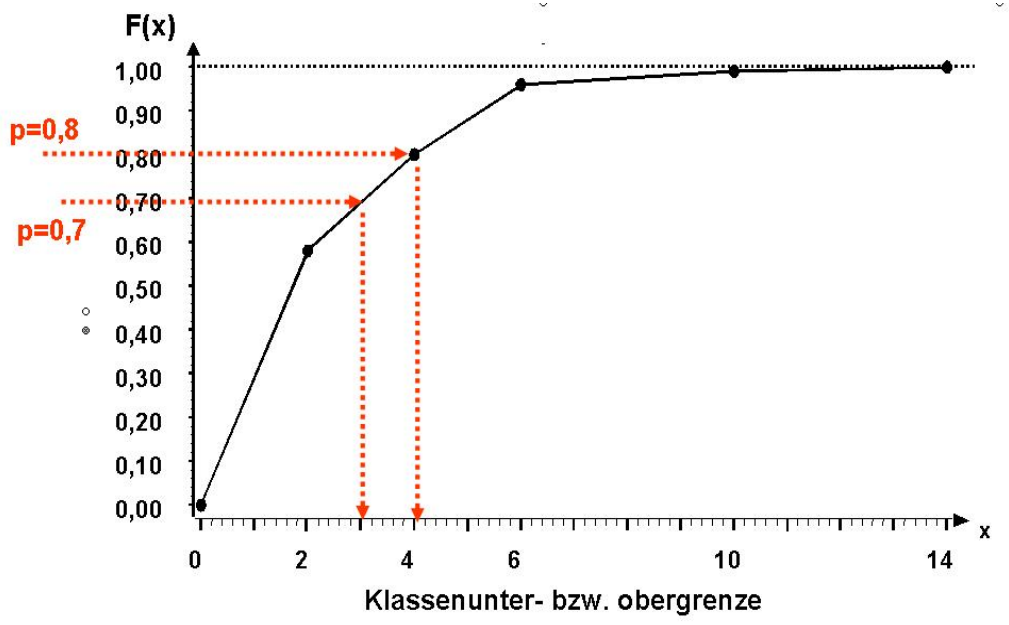
Interpretation:

$x_{0,7} = 3,09\,min = 3\,min + 5,4\,Sekunden$

1. wenigstens 70 Prozent aller Telefongespräche überschreitet nicht die Dauer von 3,09 min.
2. die maximale Dauer von 70 Prozent aller Telefongespräche beträgt 3,09 min.
3. höchstens 30 Prozent aller Telefongespräche ist länger als 3,09 min.

Hinweis: Vergleich mit dem 70-Prozent-Quantil aus Einzelwerten ($x_{0,7} = 3,11$)

Quantil für Intervalle - Bestimmen aus empirischer Verteilungsfunktion



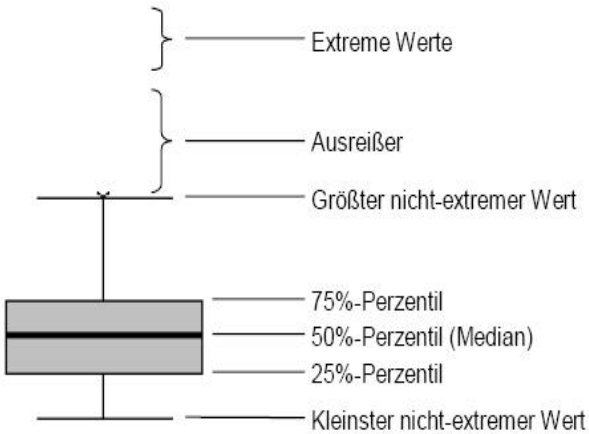
Weitere Quantile: 0,05-Quantil; 0,02-Quantil; 0,19-Quantil

Spezielle Quantile

p	*Benennung der Quantile*
$0,01$	*Perzentil*
$0,10$	*Dezil* (*unteres*)
$0,25$	*Quartil* (*unteres*)
$0,5$	*Median*, *Quartil* (*mittleres*)
$0,75$	*Quartil* (*oberes*)
$0,90$	*Dezil* (*oberes*)
$0,99$	*Perzentil*

Box-and-Whisker Plot

Boxplots: Ein Boxplot (Box-and-Whisker-Plot) beschreibt die Verteilung der Werte. Es stellt die Quartile, extreme Werte und Ausreißer sowie den größten und den kleinsten nicht extremen Wert dar. Extrem große Werte bzw. Ausreißer nach oben sind dadurch gekennzeichnet, daß ihr Abstand zum 75%-Perzentil größer ist als das 1,5fache des Abstands zwischen dem 75%- und dem 25%-Perzentil. Extrem kleine Werte bzw. Ausreißer werden entsprechend an ihrem Abstand zum 25%-Perzentil gemessen.



Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Häufigkeiten - Fortsetzung
- Diagramme (Boxplots) - Fortsetzung

Modus, Median, Quantile - Beispiel

mit Excel:

Beispiel 1 (aus Urliste): BeispielW95.xls
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:

Aufgaben: Bestimmen von Modus, Median, Quantile

mit SPSS:

Beispiel 1: World95.sav
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:

Aufgaben: Bestimmen von Modus, Median, Quantile

2.2.1 Lagemaße - Gliederung

- 2.2.1.1 Einführung
- 2.2.1.2 Modus
- 2.2.1.3 Median
- 2.2.1.4 Quantile
- **2.2.1.5 Beispiel: Historische Simulation**
- 2.2.1.6 Arithmetisches Mittel

Definitionen 1

Gegeben sei ein Portfolio, dessen Marktwert man in funktionaler Abhängigkeit von Marktdaten (wie z.B. Wechselkursen oder Zinssätzen) beschreiben kann.

$P(t) = P(t)$ - der Marktwert des Portfolios zum Zeitpunkt t
$P(t_0) = P(t_0)$ - der Marktwert des Portfolios zum Anfangseitpunkt t_0

Annahme:

die Zusammensetzung des Portfolios ändert sich nicht im Zeitraum $t - t_0 > 0$

$\Delta P(t) = P(t) - P(t_0)$ - Wertveränderung des Portfolios

(Profit and Loss - PnL)

Definitionen 2

Bezeichnungen:

$Y(t)$ - Vektor der Marktdaten zum Zeitpunkt t
$Y_i(t)$ - i-te Komponente des Vektors der Marktdaten
$P(Y(t))$ - Preisfunktion (Wert des Portfolios zum Zeitpunkt t)

Die Wertveränderung des Portfolios (PnL) lässt sich dann darstellen:

$$\Delta P(t) = P(Y(t)) - P(Y(t_0))$$

Value at Risk (VaR)

Das Ziel der Abschätzung des Marktrisikos ist eine Aussage über den potentiellen Wertverlust des Portfolios innerhalb einer fixierten Halteperiode t.

Der potentielle Wertverlust heißt Value at Risk (VaR).

Annahme: Wir verfügen über statistisches Material der Wertveränderung des Portfolios

Zunächst wird die Wertveränderung des Portfolios der Größe nach geordnet: $\Delta P(t) < \Delta P(t+1)$

Es wird das $\alpha \cdot 100\%$-Quantil berechnet:

$x_\alpha = min(\Delta P(t) \,|F(\Delta P(t)) \geq \alpha)$

Mathematisch gesehen ist $VaR(t, \alpha)$ das α-Quantil der Verteilung von $\Delta P(t)$.

Value at Risk (VaR)

Interpretation

1. der maximale Wert, der von $\alpha \cdot 100\%$ aller Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) erreicht wird, beträgt x_α.
2. in mindestens $\alpha \cdot 100\%$ aller Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) wird der Wert von x_α nicht überschritten.
3. von höchstens $(1 - \alpha) \cdot 100\%$ aller Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) wird der Wert von x_α überschritten.

Hinweis: in der Regel sind die ersten Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) negativ, deshalb wird von Wertverlust gesprochen.

1. Der minimale Wertverlust, der von $\alpha \cdot 100\%$ aller Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) erreicht wird, beträgt x_α.
2. Von höchstens $(1 - \alpha) \cdot 100\%$ aller Fälle der Wertveränderungen des Portfolios ($\Delta P(t)$) wird ein Wertverlust von x_α unterschritten.

Inhalte

- Idee der Historischen Simulation
- Das Verfahren
- ein Beispiel

Idee der Historische Simulation 1

gegeben:

- Aktuelle Marktdaten - $Y_i(0),\ i = 1, 2, ..., n$
- Historische Marktdaten - $Y_i(t),\ t = -1, -2, ..., -T$
- aktueller Marktwert des Porfolios $P(0) = P(Y_1(0), Y_2(0), ..., Y_n(0))$

an Stelle der Marktdaten verwenden wir die sogenannten **logarithmischen Rendite** bzw. **logreturns**

$$R_i(0) = ln(\frac{Y_i(0)}{Y_i(-1)})$$

allgemein :

$$R_i(t+1) = ln(\frac{Y_i(t+1)}{Y_i(t)}),\ t = -1, -2, ..., -T$$

$$R_i(t) = ln(\frac{Y_i(t)}{Y_i(t-1)}),\ t = 0, -1, ..., -T+1$$

Bemerkung: häufig T = 250

Idee der Historische Simulation 1

gegeben:

- Aktuelle Marktdaten (logarithmische Rendite) - $R_i(0),\ i = 1, 2, ..., n$
- Historische Marktdaten (logarithmische Rendite) - $R_i(t),\ t = -1, -2, ..., -250$
- aktueller Marktwert des Porfolios $P(0) = P(R_1(0), R_2(0), ..., R_n(0))$

Bewerten des aktuellen Portfolios mit den historischen Marktdaten (logarithmische Rendite)

$$P(t) = P(R_1(t), R_2(t), ..., R_n(t)),\ t = -1, -2, ..., -250$$

Berechnen der Wertveränderung des Portfolios (aktuelle Marktdaten vs. historische Marktdaten)

$$\begin{aligned} \Delta P(t) &= P(t) - P(0) \\ &= P(R_1(t), R_2(t), ..., R_n(t)) - P(R_1(0), R_2(0), ..., R_n(0)) \end{aligned}$$

$\Rightarrow$ 250 Wertveränderungen des Portfolios:

$\Delta P(-1), \Delta P(-2), ..., \Delta P(-250) \Rightarrow$ *diskreteVerteilung*

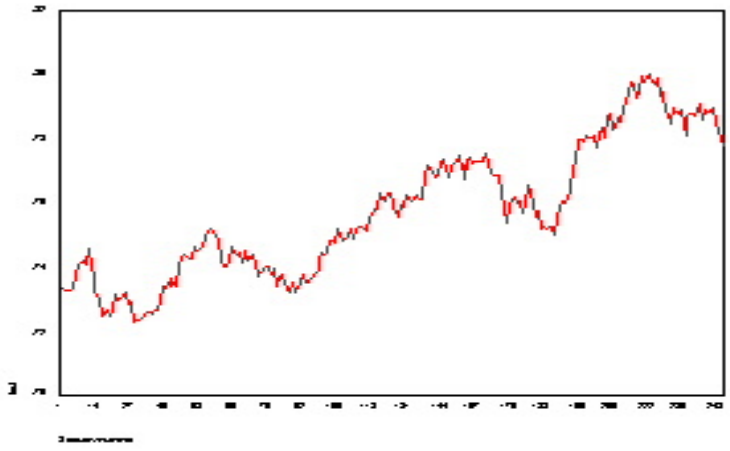
Idee der Historische Simulation 2

Quantilsbestimmung einer diskrete Verteilung

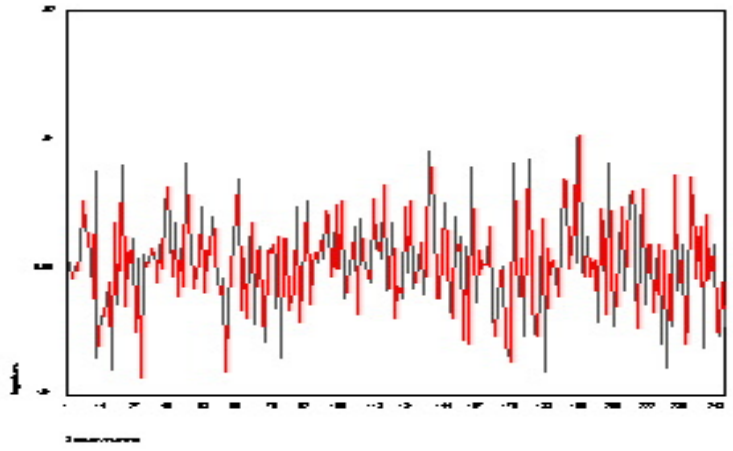
- Ordnen der Wertveränderungen der Größe nach $P(1) \leq P(2) \leq ... \leq P(250)$ Achtung (neue Numerierung!)
- Berechnen des Quantils der diskreten Verteilung (bei gegebenen Konfidenznivaus α)
- $\alpha = 0,01$ bei n =250 $\Rightarrow$ Mittelwert der zweit- und drittkleinsten Wertveränderung $\frac{1}{2}(P(2) + P(3))$ neue Numerierung
- $\alpha = 0,10$ bei n =250 $\Rightarrow$ die 25-kleinste Wertveränderung ($P(25)$) neue Numerierung
- $\alpha = 0,05$ bei n =250 $\Rightarrow$ Mittelwert der 12- und 13-kleinsten Wertveränderung $\frac{1}{2}(P(12) + P(13))$ neue Numerierung

Beispiel: VaR einer FX-Cash Position - historische Simulation

Position N = 1,000,000 USD Heimatwährung EUR
Haltedauer $t = 1$
Konfidenzniveau $\alpha = 0.01$
Wechselkurs am 31.8.2007 $Y(t_0) = 0.73338$



Y(t)



Logarithmische Rendite

Rohdaten

Wechselkurs Dollar - Euro (4.9.2006-31.8.2007)
$\Rightarrow$ 251(0) Beobachtungen

(Inverse Zeitachse)

Beispiel: VaR einer FX-Cash Position - hist. Sim.

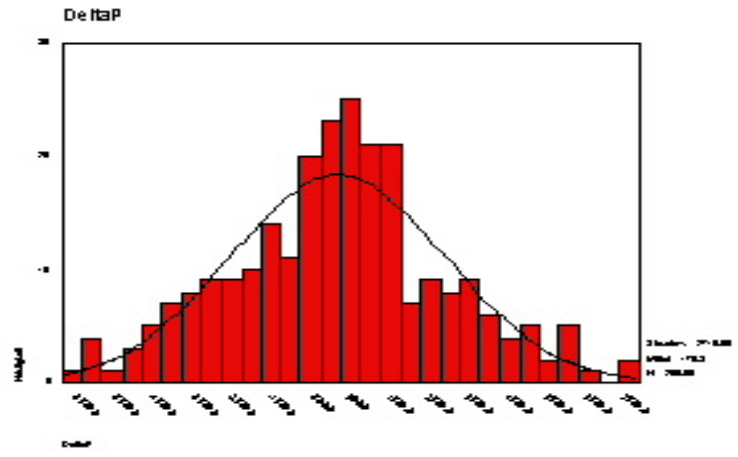
t	*Datum*	$Y_i(t+1)$	$R_{i+1}(t)$	$\widetilde{Y_i(t+1)}$	$\Delta P(t+1)$
-1	$2007-08-31$	0.7334	-0.000388	0.7331	-284.5
-2	$2007-08-30$	0.7337	0.000925	0.7341	679.0
-3	$2007-08-29$	0.7330	-0.000005	0.7334	-3.5
-4	$2007-08-28$	0.7330	0.000259	0.7336	190.2
...	...	...	...	...	...

$$
\begin{aligned}
R_i(t) &= ln(\frac{Y_i(t)}{Y_i(t-1)}),\ t=-1,-2,\ldots,-250 \\
\widetilde{Y_i(t)} &= e^{R_i(t+1)}Y_i(0),\ t=-1,-2,\ldots,-250 \\
\Delta P(t+1) &= N(\widetilde{Y_i(t)} - Y_i(0)),\ t=-1,-2,\ldots,-250
\end{aligned}
$$

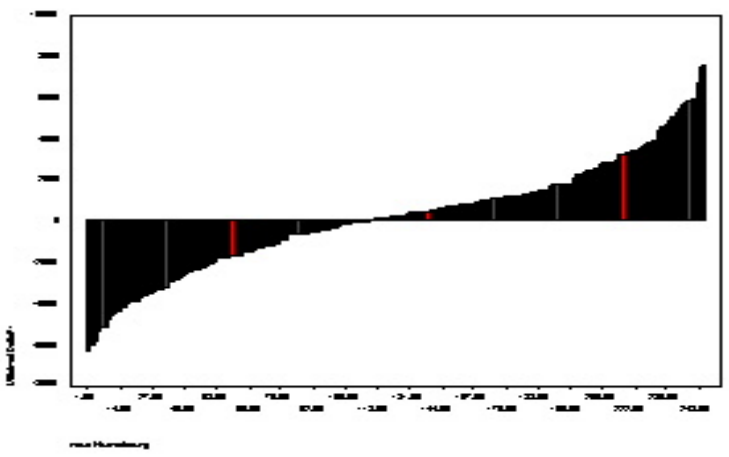
zum Beispiel für t = -1

$$
\begin{aligned}
R_i(0) &= ln(\frac{Y_i(0)}{Y_i(-1)}) = ln(\frac{0.7334}{0.7337}) = -0.000388 \\
\widetilde{Y_i(0)} &= e^{R_i(0)}Y_i(0) = e^{-0.000388}0.7334 = 0.7331 \\
\Delta P(0) &= N(\widetilde{Y_i(0)} - Y_i(0)) = 1.000.000(0.7331-0.7334) = -284.5
\end{aligned}
$$

VaR einer FX-Cash Position - hist. Sim.



geordnete $\Delta P(t)$



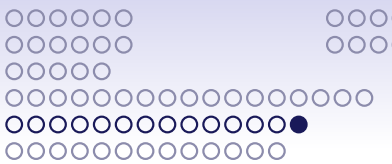
$\Delta P(t)$

Beispiel: VaR einer FX-Cash Position - hist. Sim.

Datum	$\Delta P(t)$	t	*neueNum.*
$2006-11-22$	-7535.5	-194	1
$2006-11-24$	-7393.4	-192	2
$2007-02-14$	-6575.9	-137	3
$2006-12-19$	-6072.0	-175	4
$2006-11-07$	-5924.5	-205	5
$2007-07-02$	-5896.3	-44	6
...	...	...	...

Achtung: neue Numerierung verwenden

$$VaR(\alpha = 0,01) = \frac{1}{2}(P(2) + P(3)) = \frac{1}{2}(-7393.4 - 6575.9) = -6984.653307$$



Bemerkung 1: Berücksichtigung Antithetic

Bei der Szenariobildung unter Berücksichtigung von Antithetik wird zu jeder in der historischen Zeitperiode aufgetretenden Marktbewegung auch die genau entgegengesetzte Marktbewegung in die Szenariomatrix übernommen. Auf diese Weise wird berücksichtigt, daß selbst für den Fall einer gleichlaufenden positiven Markttendenz oder einer entgegengesetzten Sensitivität des Portfolios über die Beobachtungsperiode ein angemessenes Risiko berechnet wird.

$$\widetilde{\Delta Y_i(t)} = e^{R_i(t)} Y_i(0),\ t = -1, -2, ..., -250$$

und

$$\widehat{\Delta Y_i(t)} = e^{-R_i(t)} Y_i(0),\ t = -1, -2, ..., -250$$

2.2.1 Lagemaße - Gliederung

- 2.2.1.1 Einführung
- 2.2.1.2 Modus
- 2.2.1.3 Median
- 2.2.1.4 Quantile
- 2.2.1.5 Beispiel: Historische Simulation
- **2.2.1.6 Arithmetisches Mittel**

Arithmetisches Mittel - Definitionen 1

Abstandsfunktion: $d(x_i, a) = (x_i - a)^2$

Forderung: $f(a) = \sum_{i=1}^{n} d(x_i, a) = \sum_{i=1}^{n} (x_i - a)^2 = \text{Minimum}$

Extremwertproblem für eine Funktion von 1 Variablen.

$$\begin{aligned} \frac{\partial f(a)}{\partial a} &= 0 \\ &= -2 \sum_{i=1}^{n} (x_i - a) = -2(\sum_{i=1}^{n} x_i - na) = 0 \quad \Rightarrow \\ a &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

Arithmetisches Mittel - Definitionen 2

Definition:

Ist X ein kardinal skaliertes Merkmal einer statistischen Gesamtheit vom Umfang n, dann heißt der Wert a, der sich ergibt, wenn man die Summe aller beobachteten Merkmalswerte x_i $(i = 1, 2, ..., n)$ gleichmäßig auf alle Merkmalsträger verteilt, **arithmetisches Mittel**.

Bezeichnung: $\overline{x}$

$$\overline{x} = a$$

Arithmetisches Mittel - Berechnungen

- aus der Urliste:

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

- aus der Häufigkeitsverteilung (disjunkte Merkmale):

$$\overline{x} = \frac{\sum_{l=1}^{m} \xi_l \cdot n_l}{n} = \sum_{l=1}^{m} \xi_l \cdot p_l$$

- aus der Häufigkeitsverteilung (Intervalle):

Arithmetisches Mittel - Berechnungen

aus der Häufigkeitsverteilung (Intervalle):

$$\overline{x} \approx \frac{\sum_{l=1}^{r} \widetilde{x}_l \cdot n_l}{n} = \sum_{l=1}^{r} \widetilde{x}_l \cdot p_l$$

$$mit\ \widetilde{x}_l = \frac{1}{2}(c_{l-1} + c_l)$$

Arithmetisches Mittel - Beispiele

1. aus der Urliste:
   Vorlesungsbeispiel 1 (Excel-Sheet)
2. aus der Häufigkeitsverteilung (disjunkte Merkmale):
   Vorlesungsbeispiel 1 (Excel-Sheet)
3. aus der Häufigkeitsverteilung (Intervalle):
   Vorlesungsbeispiel 2 (Excel-Sheet)

Arithmetisches Mittel - Eigenschaften - I

- Nulleigenschaft / Schwerpunkteigenschaft

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

$$\sum_{l=1}^{m}(x_l - \overline{x})n_l = 0$$

- Hochrechnungseigenschaft:

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

- Arithmetisches Mittel für linear transformierten datensatz::

$$y_i = a + bx_i \Rightarrow \overline{y} = a + b\overline{x}$$

- Sensitivität bezüglich Ausreißer

gepoolter Datensatz

Definition: Der Datensatz ist in disjunkte Gruppen $D_1, D_2, ..., D_r$ untergliedert. Für jede der disjunkten Gruppen sind die statistischen Kenngrößen berechnet und damit bekannt. Einen solchen Datensatz nennt man **gepoolter Datensatz** bzw. **Datenpool**.

Aufgabe: Berechnen von statistischen Kenngrößen für den zusammengefaßten (gepoolten) Datensatz $(D_1 \cup D_2 \cup ... \cup D_r)$ mit Hilfe der Kenngrößen der disjunkten Gruppen $D_1, D_2, ..., D_r$

arithmetisches Mittel für gepoolten Datensatz

Sind die Beobachtungswerte in disjunkten Gruppen $D_1, D_2, ..., D_r$ gegeben und ist für jede Gruppe das arithmetische Mittel $\overline{x_p}$ bekannt, kann das arithmetische Mittel für den aus allen r Gruppen zusammengefaßten (gepoolten) Datensatz $(D_1 \cup D_2 \cup ... \cup D_r)$ mit der folgenden Formel berechnet werden, wobei n_p die absoluten Häufigkeiten Beobachtungswerte der p-ten Gruppe ist:

$$\overline{x} = \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n}$$

$$mit\ n = \sum_{p=1}^{r} n_p$$

Arithmetisches Mittel (gepoolter Datensatz) - Beispiele

Datensatz: Vorlesungsbeispiel 2

j	c_u	c_o	n_j	p_j	arithmetisches Mittel	empirische Varianz
1	0	2	58	0.58	0.92	0.3508
2	2	4	22	0.22	2.9	0.3694
3	4	6	16	0.16	4.63	0.2578
4	6	10	3	0.03	7.98	0.038
5	10	14	1	0.01	13.13	0
			100	1		

Bemerkung:
Mittelwerte und empirische Varianzen werden aus der Urliste für die entsprechenden Gruppen berechnet.

$$\overline{x} = \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n} = 2,2821$$

Mittelwert - Spezialfälle

- Mittelwert für Zeitreihen (durchschnittliches jährliches Wachstumstempo) - geometrisches Mittel
- Harmonisches Mittel (Statisches Merkmal ist eine Verhältniszahl)
- gewogenes arithmetisches Mittel

Einführung in SPSS - Themen

Formaler Umgang mit SPSS

- Umgang mit Dateien und Dateitransformationen - Fortsetzung (Gruppenbildung unter SPSS)

Statistische Themen: Deskriptive Statistik

- Häufigkeiten - Fortsetzung

Mittelwerte - Beispiel

mit Excel:

Beispiel 1 (aus Urliste): BeispielW95.xls
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:
Beispiel 3 (gepoolter Datensatz):Zinsätze 2007_4.xls

Aufgaben: Bestimmen der Mittelwerte

mit SPSS:

Beispiel 1 (mit Gruppenbilung unter SPSS): World95.sav.xls
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:

Aufgaben: Bestimmen der Mittelwerte

Maßzahlen der Verteilung

- 2.2.1 Lagemaße (Maße der zentralen Tendenz)
- **2.2.2 Streuungsmaße**
- 2.2.3 Formenmaße
- 2.2.4 lineare Transformationen

2.2.2 Streuungsmaße - Gliederung

- **2.2.2.1 Disparitätsmaße**
- 2.2.2.2 metrische Streungsmaße

Streuungsmaße Intention

Zielstellung:
Informationen bereitstellen, welche die Konzentration der Beobachtungen um das Lagemaß quantifizieren. In Abhängigkeit von den zugrundeliegenden Skalen werden folgende Maßzahlen verwendet:

- Nominalskala $\Rightarrow$ nominales Disparitätsmaß
- Ordinalskala $\Rightarrow$ ordinales Dispritätsmaß
- Kardinalskala (metrische Skala)
  - Spannweite
  - zentraler Quantilsabstand
  - (empirische) Varianz
  - (empirische) Standardabweichung
  - Variationskoeffizient

nominales Disparitätsmaß

Ausgangspunkt: relative Häufigkeitsverteilung eines **nominalen** Merkmales:

$$f : \xi_l \rightarrow p_l = f(\xi_l) = p(\xi_l),\ l = 1, 2, ..., m$$

Definition: das **nominale Disparitätsmaß** d_N ist die Größe:

$$d_N = \sqrt{1 - \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2)}\ \textit{mit}\ 0 \leq d_N < 1$$

Häugig auch: normierte Streuungsmaß nach HERFINDAHL

$$d_H = \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2)\ \textit{mit}\ 0 \leq d_N < 1$$

Interpretation:

$d_N \approx 0$ *oder* $d_H \approx 1$ $\Rightarrow$ Gleichverteilung des Merkmals.
$d_N \approx 1$ *oder* $d_H \approx 0$ $\Rightarrow$ fast Einpunktverteilung.

nominales Disparitätsmaß - Interpretation

Annahme: Gleichverteilung $\Rightarrow \quad p_l = \frac{n_l}{n} = \frac{1}{m}$

$$\begin{aligned} \sum_{l=1}^{m} p_l^2 &= \sum_{l=1}^{m} \frac{m}{m^2} = \frac{1}{m} \\ d_N &= \sqrt{1 - \frac{m}{m-1} \cdot (1 - \frac{1}{m})} \\ &= \sqrt{1 - \frac{m}{m-1} \cdot \frac{m-1}{m}} = 0 \end{aligned}$$

Annahme: Einpunktverteilung $\Rightarrow \quad p_1 = 1; \; p_2 = \ldots p_m = 0$

$$\begin{aligned} \sum_{l=1}^{m} p_l^2 &= 1 \\ d_N &= \sqrt{1 - \frac{m}{m-1} \cdot (1 - 1 = 1} \end{aligned}$$

nominales Disparitätsmaß - Vorlesungsbeispiel 4

Region or economic group (1995)

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	*OECD*	21	19.3
2	*EastEurope*	14	12.8
3	*Pacific/Asia*	17	15.6
4	*Africa*	19	17.4
5	*MiddleEast*	17	15.6
6	*LatnAmerica*	21	19.3
Summe	*Gesamt*	109	100.0

$$d_N = \sqrt{1 - \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2)} = \sqrt{1 - \frac{6}{6-1}(1 - 0.169766855)} = 0.060993652$$

Interpretation:

$d_N \approx 0 \Rightarrow$ annähernde Gleichverteilung des Merkmals.

nominales Disparitätsmaß - Vorlesungsbeispiel 4

Region or economic group (1995)

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	*OECD*	21	19.3
2	*EastEurope*	14	12.8
3	*Pacific/Asia*	17	15.6
4	*Africa*	19	17.4
5	*MiddleEast*	17	15.6
6	*LatnAmerica*	21	19.3
Summe	*Gesamt*	109	100.0

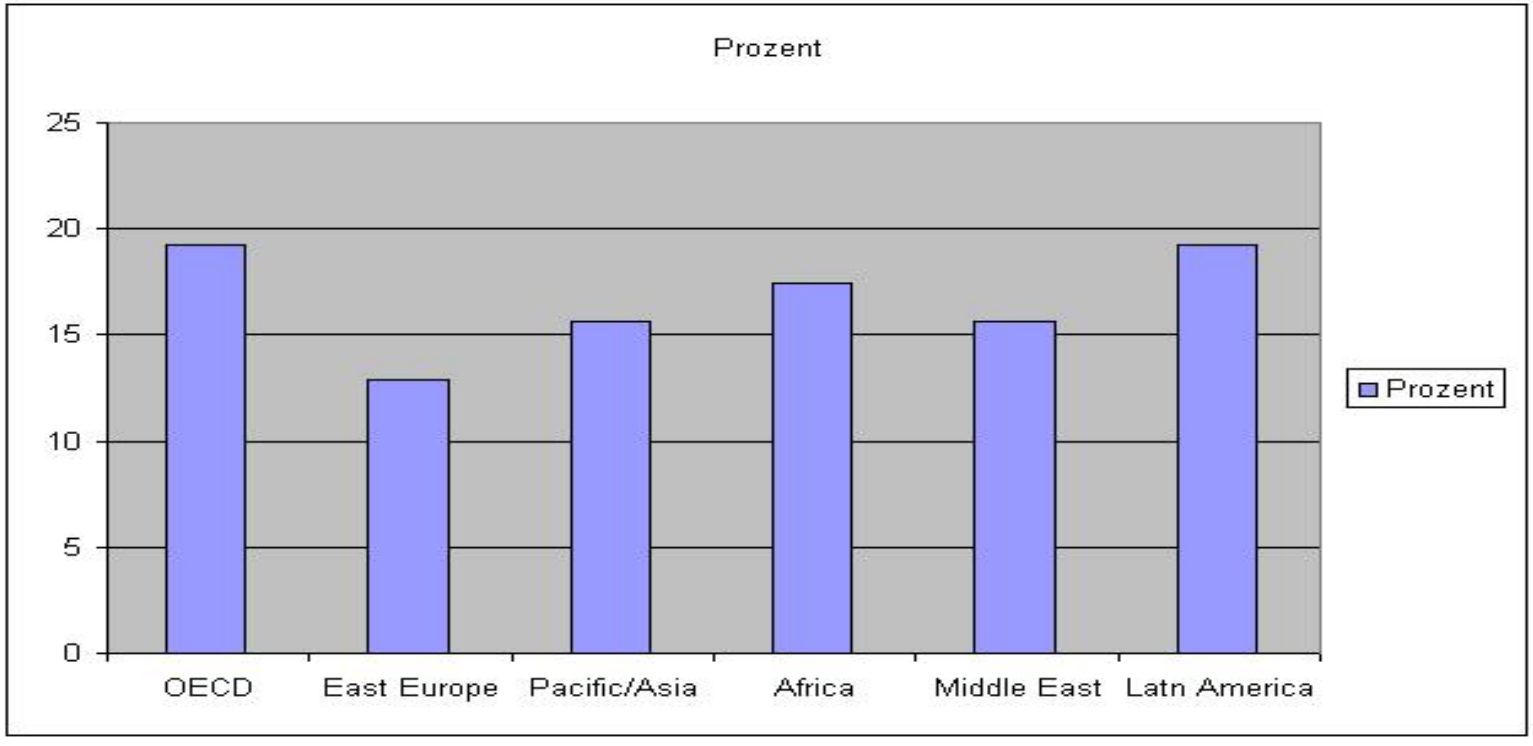
$$d_H = \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2) = \frac{6}{6-1}(1 - 0.169766855) = 0,996279774$$

Interpretation:

$d_H \approx 1 \Rightarrow$ annähernde Gleichverteilung des Merkmals.

nominales Disparitätsmaß - Vorlesungsbeispiel 4

Datensatz: Vorlesungsbeispiel 4



Bemerkung:

$d_N \approx 0 \Rightarrow$ annähernde Gleichverteilung des Merkmals.

ordinales Disparitätsmaß

Ausgangspunkt: absolute Summenhäufigkeitsverteilung eines **ordinalen** Merkmales:

$$f : \xi_l \rightarrow F_l = f(\xi_l) = F(X \leq \xi_l)$$

Definition: das **ordinale Disparitätsmaß** d_O ist die Größe:

$$d_O = 1 - \frac{6m}{m^2 - 1} \sum_{l=1}^{m-1} F_l(1 - F_l) \; mit \; 0 \leq d_N < 1$$

Interpretation:

$d_O = 0 \;\Rightarrow$ Gleichverteilung des Merkmals.
$d_O \approx 1 \;\Rightarrow$ fast Einpunktverteilung.

ordinales Disparitätsmaß - Vorlesungsbeispiel 5

Klausurnoten (1995)

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*	*relative Summenhaeufigkeit*
j	ξ_j	n_j	p_j	F_j
1	*sehrgut*	5	11.4	11.4
2	*gut*	20	45.5	56.9
3	*befriedigend*	10	22.7	79.6
4	*genuegend*	6	13.6	93.2
5	*ungenuegend*	4	6.8	100.0
Summe	*Gesamt*	45	100.0	

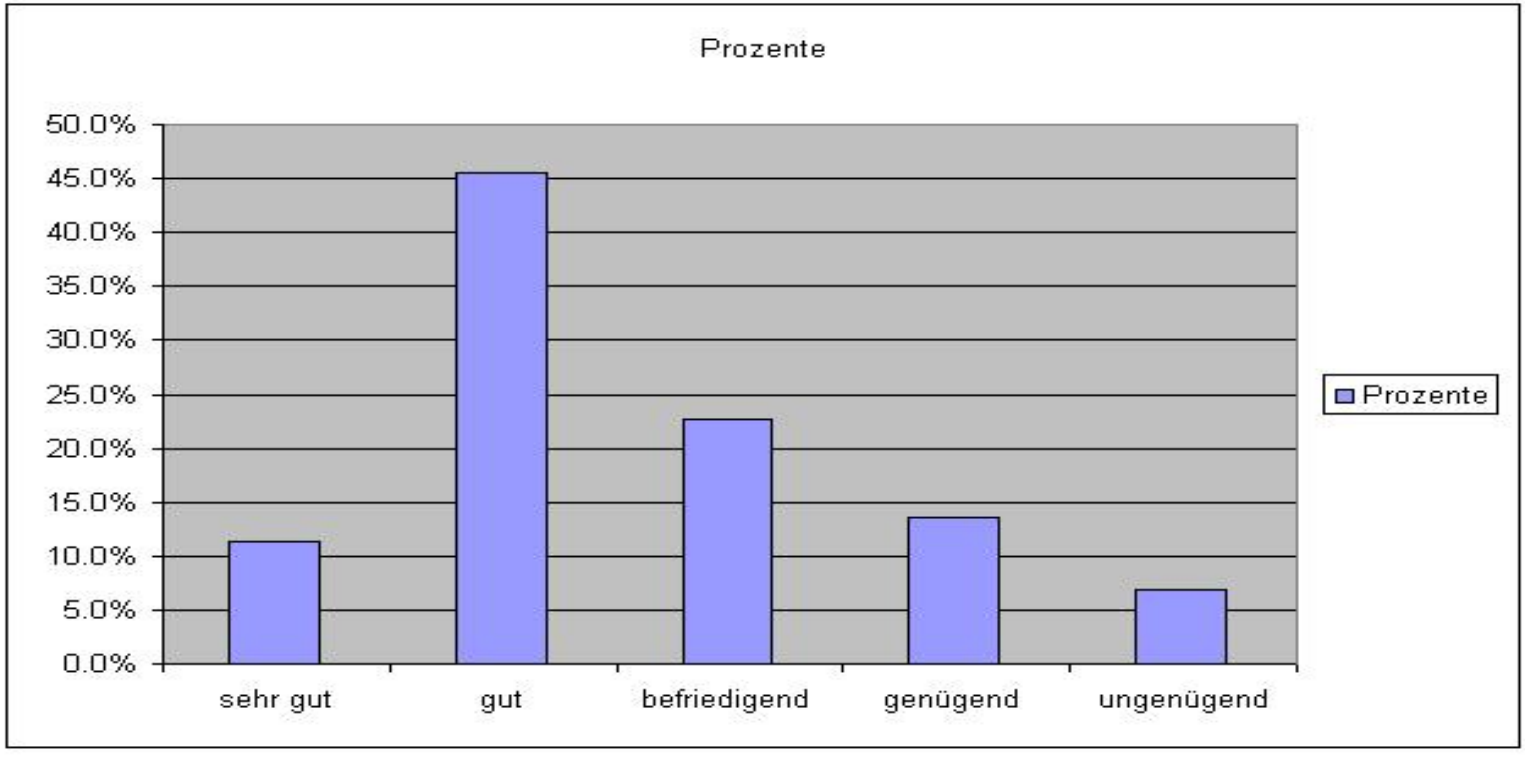
$$d_O = 1 - \frac{6m}{m^2 - 1} \sum_{l=1}^{m-1} F_l(1 - F_l) = 1 - \frac{6 \cdot 5}{5^2 - 1} \sum_{l=1}^{5-1} F_l(1 - F_l) = 0,2846$$

Interpretation:

$d_O > 0 \Rightarrow$ kaum Gleichverteilung des Merkmals.

ordinales Disparitätsmaß - Vorlesungsbeispiel 5

Datensatz: Vorlesungsbeispiel 4



Bemerkung:

$d_O > 0 \Rightarrow$ kaum Gleichverteilung des Merkmals.

2.2.2 Streuungsmaße - Gliederung

- 2.2.2.1 Disparitätsmaße
- **2.2.2.2 metrische Streungsmaße**

Spannweite

gegeben sei die Rangwertreihe eines ordinal skalierter Datensatzes:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Definition: **Spannweite R**:

$$R = X_{MAX} - X_{MIN} = x_n - x_1$$

Definition: **Spannweite R** mit Intervallbildung:

$$R = c_r - c_0$$

zentraler Quantilsabstand

Ist X ein in seinen Merkmalswerten aufsteigend geordnetes, kardinal skaliertes Merkmal einer statistischen Gesamtheit:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

dann heißt die Differenz aus den (1+p)/2 – Quantil und dem (1-p)/2 – Quantil für ein vorgegebenes p, wobei $0 < p < 1$, zentraler Quantilsabstand (Interquantilsabstand) der Ordnung p, und wird mit Q_p bezeichnet.

ein spezieller zentraler Quantilsabstand:

$$Q_p = x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}}$$

zentraler Quantilsabstand der Ordnung p = 0,5 (Interquartilsabstand):

$$Q_{0,5} = x_{\frac{1+0,5}{2}} - x_{\frac{1-0,5}{2}} = x_{0,75} - x_{0,25}$$

empirische Varianz - Berechnungsvorschriften

Ist X ein kardinal skaliertes Merkmal einer statistischen Gesamtheit, dann heißt das arithmetische Mittel der quadrierten Abweichungen der Merkmalswerte von ihrem arithmetischen Mittel empirische Varianz d_x^2:

- aus der Urlisten (Beobachtungen x_i $i = 1, 2, ..., n$)

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

- aus der Häufigkeitstabelle (disjunkte Merkmale ξ_l, relative Häufigkeiten p_l, $l = 1, 2, ..., m$)

$$d_x^2 = \sum_{l=1}^{m} (\xi_l - \overline{x})^2 \cdot p_l$$

empirische Varianz - Berechnungsvorschriften:

- aus der Häufigkeitstabelle mit Intervallbildung (Intervalmitte x_l^*, relative Häufigkeiten p_l, $l = 1, 2, ..., r$)

$$d_x^2 = \sum_{l=1}^{r} (x_l^* - \overline{x})^2 \cdot p_l$$

- für gepoolte Daten (arithmetisches Mittel $\overline{x_l}$, empirische Varianz d_l^2, relative Häufigkeiten p_l, der Gruppe l, $l = 1, 2, ..., r$)

$$d_x^2 = \sum_{l=1}^{r} d_l^2 \cdot p_l + \sum_{l=1}^{r} (\overline{x_l} - \overline{x})^2 \cdot p_l = d_a^2 + d_b^2$$

d_a^2 - Varianz innerhalb der Gruppen
d_a^2 - Varianz zwischen den Gruppen

empirische Varianz für gepoolte Daten Herleitung

$$
\begin{aligned}
d_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 == \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x})^2 = \\
&= \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x_l}+\overline{x_l}-\overline{x})^2 == \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}[(x_{il}-\overline{x_l})+(\overline{x_l}-\overline{x})]^2 = \\
&= \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x_l})^2+\frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_l-\overline{x})^2+2\frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x_l})(x_l-\overline{x}) = \\
&= \frac{1}{n}\sum_{l=1}^{m}n_l(x_{il}-\overline{x_l})^2+\frac{1}{n}\sum_{l=1}^{m}n_l(x_l-\overline{x})^2 \\
&= \sum_{l=1}^{m}d_l^2\cdot p_l+\sum_{l=1}^{m}(\overline{x_l}-\overline{x})^2\cdot p_l = d_a^2+d_b^2
\end{aligned}
$$

$$
da \quad 2\frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x_l})(x_l-\overline{x}) = \sum_{i=1}^{n_l}(x_{il}-\overline{x_l}) = 0
$$

empirische Standardabweichung - Definition

Die (positive) Wurzel aus der empirische Varianz eines kardinalskalierten Merkmals X heißt empirische Standardabweichung(d_x).

$$d_x = \sqrt[2]{d_x^2}$$

empirische Standardabweichung - Interpretation

- Im Intervall$[\overline{x} - d_x, \overline{x} + d_x]$ liegen mindestens 50 % aller Beobachtungen
- bei normalverteilten Beobachtungen liegen im Intervall $[\overline{x} - d_x, \overline{x} + d_x]$ ca.68 % aller Beobachtungen.
- bei normalverteilten Beobachtungen liegen im Intervall $[\overline{x} - 2d_x, \overline{x} + 2d_x]$ ca.95 % aller Beobachtungen.
- bei normalverteilten Beobachtungen liegen im Intervall $[\overline{x} - 3d_x, \overline{x} + 3d_x]$ ca.99 % aller Beobachtungen.

Variationskoeffizient

Der Koeffizient aus empirischer Standardabweichung d_x und arithmetischen Mittel $\overline{x}$ heißt **Variationskoeffizient** (V_x)

$$V_x = \frac{d_x}{\overline{x}}$$

$$\textit{Voraussetzung}: \quad \overline{x} > 0 \ \textit{und} \ x_i > 0 \ \forall i = 1, 2, ..., n$$

Interpretation; Ein Variationskoeefizient größer als 0,5 ist ein Indiz dafür, dass das arithmetische Mittel $\overline{x}$ wegen der sehr großen Streung d_x kein geeigneter Repräsentant der Einzelwerte x_i ist.

metrische Streuungsmaße - Beispiel 1

gegeben geordnete Urliste: (Telefongespräche)

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
R &= X_{MAX} - X_{MIN} = X_n - X_1 \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Streungsmasse Vorlesungsbeispiel 1

aufsteigend sortierte Urliste:

0,04	0,09	0,10	0,10	0,11
0,12	0,14	0,16	0,22	0,23
0,26	0,27	0,27	0,34	0,38
0,39	0,44	0,46	0,48	0,51
0,51	0,55	0,63	0,66	0,76
0,77	0,84	0,87	0,87	0,97
0,97	0,97	1,03	1,05	1,06
1,07	1,15	1,19	1,20	1,26
1,29	1,33	1,33	1,42	1,54
1,57	1,57	1,62	1,68	1,70
1,72	1,74	1,74	1,78	1,79
1,86	1,93	1,99	2,01	2,02
2,13	2,14	2,15	2,28	2,38
2,58	2,59	2,82	2,92	3,11
3,11	3,13	3,15	3,23	3,47
3,51	3,60	3,64	3,86	3,90
4,11	4,14	4,23	4,28	4,32
4,32	4,41	4,41	4,52	4,59
4,60	4,67	4,68	5,39	5,42
5,98	7,71	8,05	8,17	13,13

metrische Streuungsmaße - Beispiel 2

gegeben Häufigkeitstabelle mit disjunkten Merkmalsausprägungen: (z.B. 50 Wohnungen)

$$\xi_1 \leq \xi_2 \leq \ \ldots, \leq \xi_m$$

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{m} \xi_i * n_i}{n} = \sum_{i=1}^{m} \xi_i * p_i \\
R &= \xi_{MAX} - \xi_{MIN} = \xi_m - \xi_1 \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \frac{1}{m} \sum_{i=1}^{m} (\xi_i - \overline{x})^2 * n_i = \sum_{i=1}^{m} (\xi_i - \overline{x})^2 * p_i \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Streungsmasse Vorlesungsbeispiel 1

Aufgabe; Bestimmen des 0,05 - Quantils (p =0,05)

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Summe		50	1

metrische Streuungsmaße - Beispiel 3

gegeben Klassifizierung (Intervalle, Telefongespräche)

$$I_1, I_2, ..., I_3 = [c_0, c_1], [c_1, c_2], ..., (c_{r-1}, c_r]$$

$$\begin{aligned}
\overline{x} &\approx \frac{\sum_{l=1}^{r} \widetilde{x}_l \cdot n_l}{n} = \sum_{l=1}^{r} \widetilde{x}_l \cdot p_l \\
R &= c_r - c_0 \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \sum_{l=1}^{r} (\widetilde{x}_l - \overline{x})^2 \cdot p_l = \frac{1}{n} \sum_{l=1}^{r} (\widetilde{x}_l - \overline{x})^2 * n_l \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Streuungsmasse Vorlesungsbeispiel 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00

Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

metrische Streuungsmaße - Beispiel 4

gegeben geopoolter Datensatz (Telefongespräche)

disjunkten Gruppen $D_1, D_2, ..., D_r$ mit $(D_1 \cup D_2 \cup ... \cup D_r)$

$$\begin{aligned}
\overline{x} &= \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n} \; mit \; n = \sum_{p=1}^{r} n_p \\
R &= c_r - c_0 \\
d_x^2 &= \sum_{l=1}^{r} d_l^2 \cdot p_l + \sum_{l=1}^{r} (\overline{x_l} - \overline{x})^2 \cdot p_l = d_a^2 + d_b^2 \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Streuungsmasse (gepoolter Datensatz) - Vorlesungsbeispiel 2

Datensatz: Vorlesungsbeispiel 2

j	c_u	c_o	n_j	p_j	arithmetisches Mittel	empirische Varianz
1	0	2	58	0.58	0.92	0.3508
2	2	4	22	0.22	2.9	0.3694
3	4	6	16	0.16	4.63	0.2578
4	6	10	3	0.03	7.98	0.038
5	10	14	1	0.01	13.13	0
			100	1		

Bemerkung:
Mittelwerte und empirische Varianzen werden aus der Urliste für die entsprechenden Gruppen berechnet.

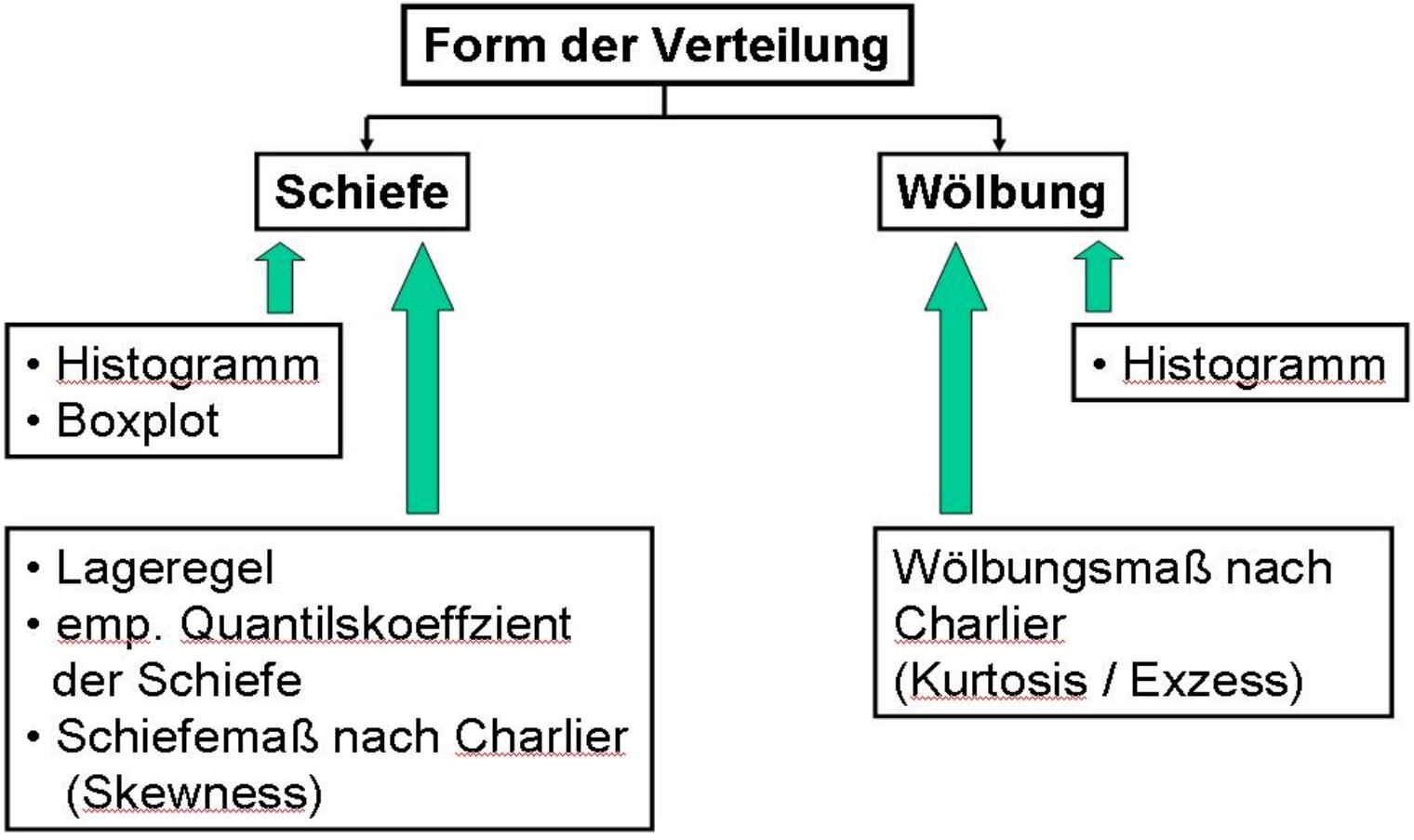
Maßzahlen der Verteilung

- 2.2.1 Lagemaße (Maße der zentralen Tendenz)
- 2.2.2 Streuungsmaße
- **2.2.3 Formenmaße**
- 2.2.4 lineare Transformationen

2.2.3 Formenmaße - Gliederung

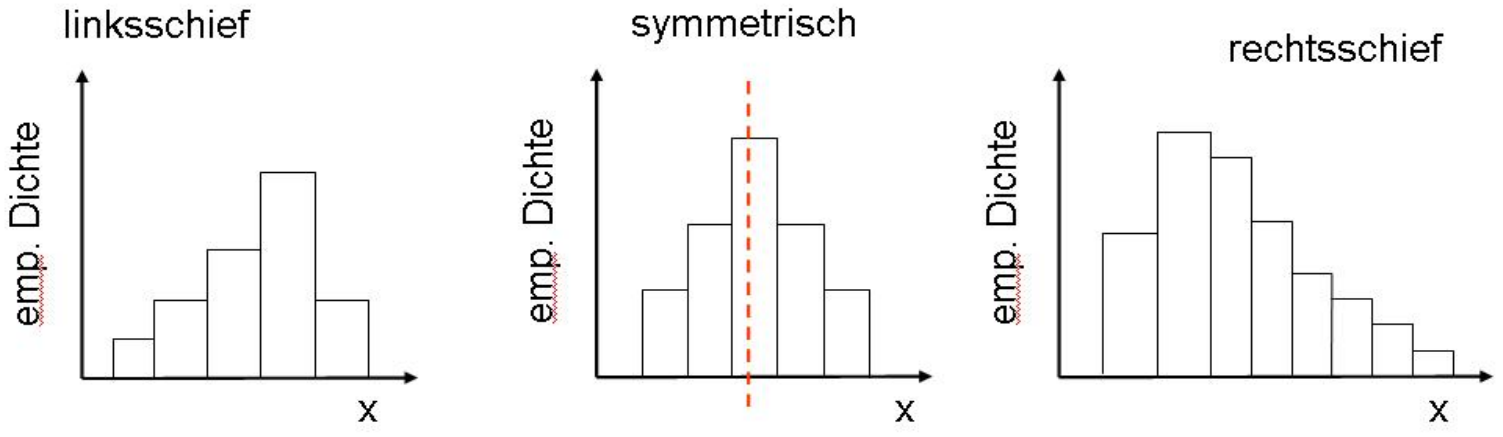
- **2.2.3.1 Schiefe**
- 2.2.3.2 Wölbung
- 2.2.3.3 empirische Momente

Formenmaße Einführung



Symmetrie / Schiefe - Histogramm

Symmetrie / Asymmetrie einer empirischen Verteilung (Histogramm)



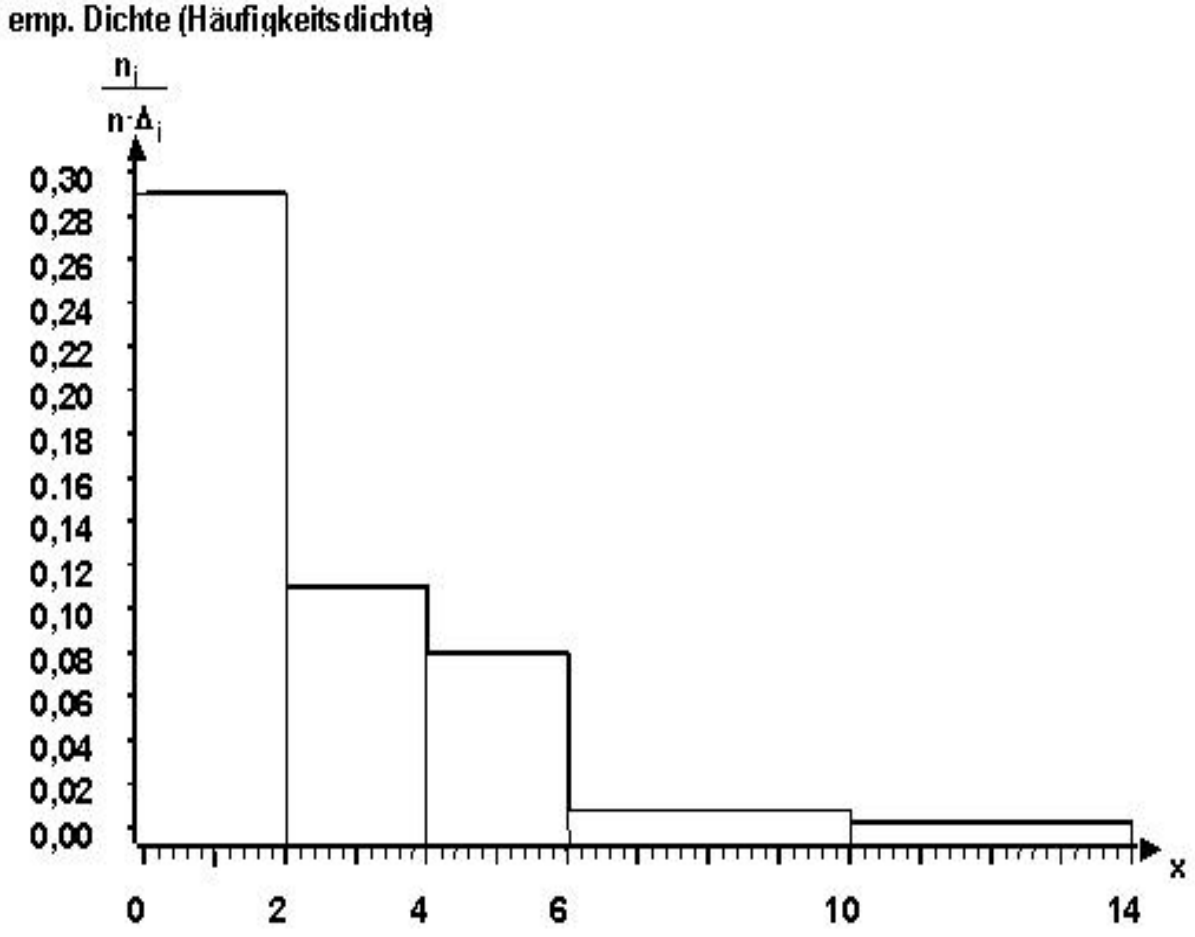
Lageregel:

(bezüglich der Modalwerte x_M und des arithmetischen Mittels $\overline{x}$)

$\overline{x} < x_{0,5} < x_M$ $\quad x_M \cong \overline{x} \cong x_{0,5}$ $\quad x_M < x_{0,5} < \overline{x}$

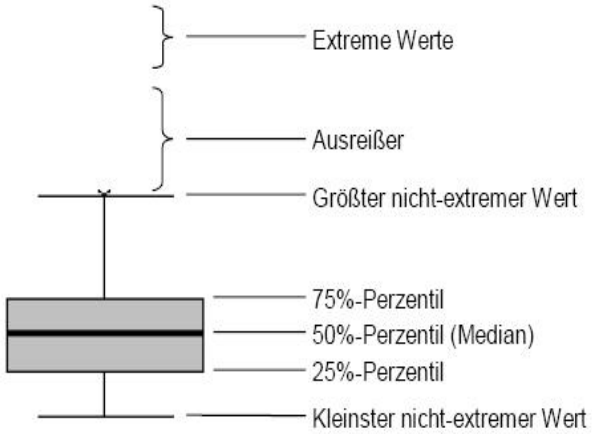
Symmetrie / Schiefe - Beispiel

Beispiel: Telefongespräche:
Analyse der Symmetrie / Asymmetrie einer empirischen Verteilung (Histogramm)



Symmetrie / Schiefe - Box-and-Whisker-Plot

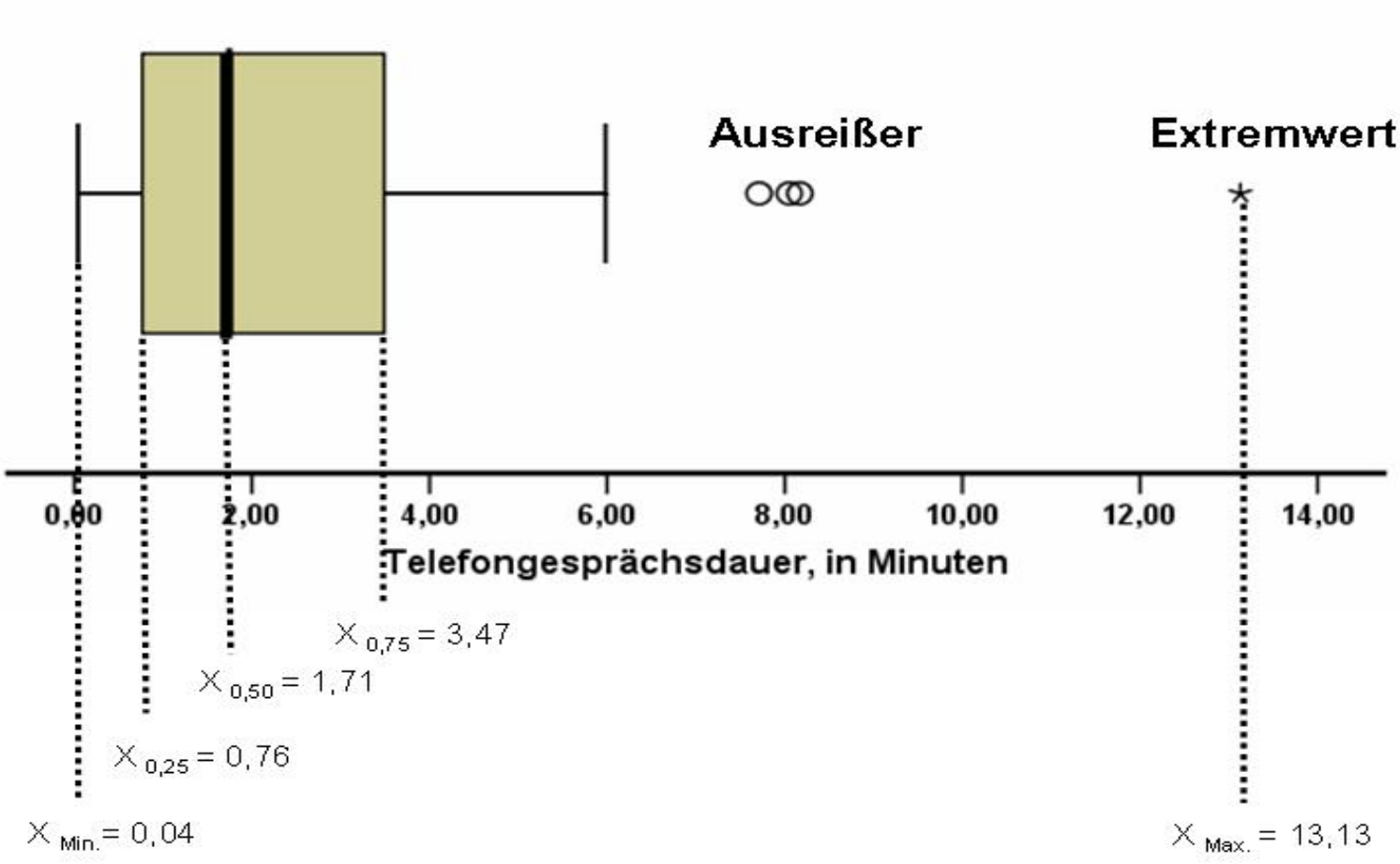
Symmetrie / Asymmetrie einer empirischen Verteilung (Box-and-Whisker-Plot)



Die Form und die Lage der Box kann zur Analyse der Symmetrie / Asymmetrie genutzt werden.

Symmetrie / Schiefe - Beispiel

Beispiel: Telefongespräche:
Analyse der Symmetrie / Asymmetrie der empirischen Verteilung
(Box-and-Whisker-Plot)



Symmetrie / Schiefe Maßzahlen

Maßzahlen zur Messung der Schiefe:
Quantilskoeffzient der Schiefe:

$$\frac{(x_{(1+p)/2} - x_{0,5}) - (x_{0,5} - x_{(1-p)/2})}{x_{(1+p)/2} - x_{(1-p)/2}}$$

$p = 0,5 \Rightarrow$ *Quartilskoeffizient*
Der Quartilskoeffizient der Schiefe erlaubt eine Aussage über 50 % der mittleren Merkmalsausprägungen:

- wenn Quartilskoeffizient < 0, dann linksschief
- wenn Quartilskoeffizient $= 0$, dann symmetrisch
- wenn Quartilskoeffizient > 0, dann rechtsschief

Symmetrie / Schiefe Vorlesungsbeispiel 1

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*	*absolute Summenhaeufigkeit*
j	ξ_j	n_j	p_j	F_j
1	1	1	0.02	1
2	2	4	0.08	5
3	3	11	0.22	16
4	4	23	0.46	39
5	5	8	0.16	47
6	6	2	0.04	49
7	7	1	0.02	50
Summe		50	1	

Quantilskoeffizient:
Berechnen der Quantile:

$$
\begin{aligned}
p = 0.5 &\quad\Rightarrow\quad g = 25 \Rightarrow x_p = 4 \\
p = 0.25 &\quad\Rightarrow\quad g = 12.5 \Rightarrow x_p = 3 \\
p = 0.75 &\quad\Rightarrow\quad g = 37.5 \Rightarrow x_p = 4 \\
p = 0.1 &\quad\Rightarrow\quad g = 5 \Rightarrow x_p = 2 \\
p = 0.9 &\quad\Rightarrow\quad g = 45 \Rightarrow x_p = 6
\end{aligned}
$$

Symmetrie / Schiefe Vorlesungsbeispiel 1

$$
\begin{aligned}
p = 0.5 &\Rightarrow g = 25 \Rightarrow x_p = 4 \\
p = 0.25 &\Rightarrow g = 12.5 \Rightarrow x_p = 3 \\
p = 0.75 &\Rightarrow g = 37.5 \Rightarrow x_p = 4 \\
p = 0.1 &\Rightarrow g = 5 \Rightarrow x_p = 2 \\
p = 0.9 &\Rightarrow g = 45 \Rightarrow x_p = 6
\end{aligned}
$$

Quantilskoeffzient der Schiefe:

$$
\begin{aligned}
p \quad &: \quad \frac{(x_{(1+p)/2} - x_{0,5}) - (x_{0,5} - x_{(1-p)/2})}{x_{(1+p)/2} - x_{(1-p)/2}} \\
p = 0.5 \quad &: \quad \frac{(x_{(1.5)/2} - x_{0,5}) - (x_{0,5} - x_{(0.5)/2})}{x_{(1.5)/2} - x_{(0.5)/2}} \\
&: \quad \frac{(x_{0.75} - x_{0,5}) - (x_{0,5} - x_{0,25})}{x_{0.75} - x_{0.25}} = \frac{(4-4) - (4-3)}{4-3} = -1 \\
p = 0.8 \quad &: \quad \frac{(x_{(1.8)/2} - x_{0,5}) - (x_{0,5} - x_{(0.2)/2})}{x_{(1.8)/2} - x_{(0.8)/2}} \\
&: \quad \frac{(x_{0.9} - x_{0,5}) - (x_{0,5} - x_{0,1})}{x_{0.9} - x_{0.1}} = \frac{(6-4) - (4-2)}{6-2} = 0
\end{aligned}
$$

Symmetrie / Schiefe Vorlesungsbeispiel 1

$$p = 0.5 \quad : \quad \textit{Quartilskoeffizient} = -1$$
$$p = 0.8 \quad : \quad \textit{Quantilskoeffizient} = 0$$

Interpretation:

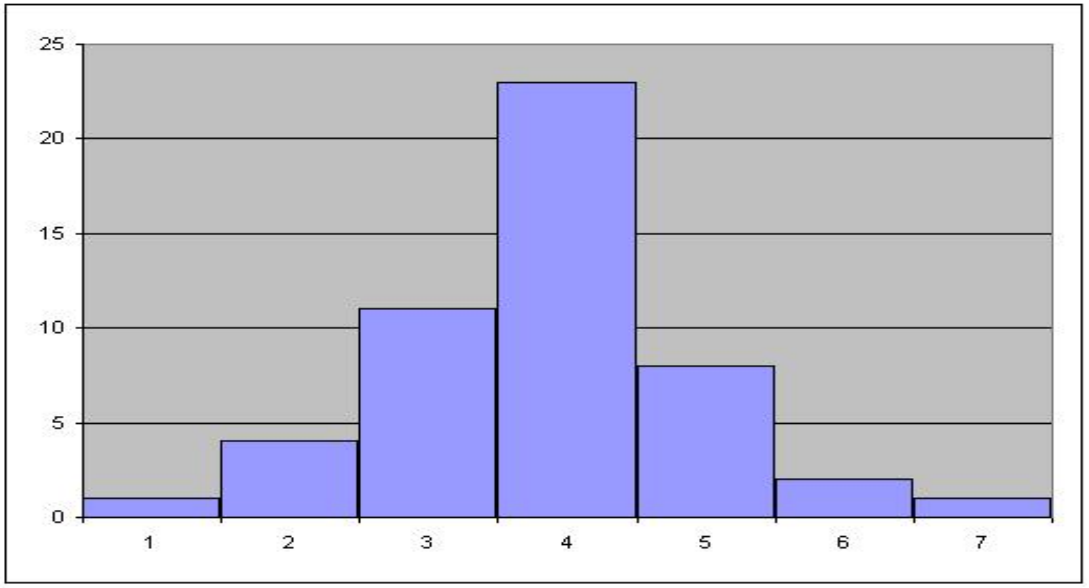
- Quartilskoeffizient = -1 $\Rightarrow$ linksschief
- Quantilskoeffizient = 0 $\Rightarrow$ symmetrisch

$\Rightarrow$ widersprüchliches und grobes Ergebnis

siehe Histogramm

Symmetrie / Schiefe Vorlesungsbeispiel 1

Histogramm:



Interpretation: annähernd symmetrisch, leicht rechtsschief

Symmetrie / Schiefe Maßzahlen

Maßzahlen zur Messung der Schiefe (skewness) aus der Urliste:
Schiefemaß nach CHARLIER S_x:

$$
\begin{aligned}
S_x &= \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} \\
mit\ m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
m_{\overline{x}}^3 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3
\end{aligned}
$$

Interpretation:

- wenn $S_X < 0$, dann linksschief
- wenn $S_X = 0$, dann symmetrisch
- wenn $S_X > 0$, dann rechtsschief

Symmetrie / Schiefe Maßzahlen

Maßzahlen zur Messung der Schiefe (Skewness) aus Häufigkeitsverteilung (disjunkte Merkmalsausprägungen):
Schiefemaß nach CHARLIER S_x:

$$
\begin{aligned}
S_x &= \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} \\
mit\ m_{\overline{x}}^2 &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot p_i \\
m_{\overline{x}}^3 &= \sum_{i=1}^{n} (x_i - \overline{x})^3 \cdot p_i
\end{aligned}
$$

Interpretation:

- wenn $S_X < 0$, dann linksschief
- wenn $S_X = 0$, dann symmetrisch
- wenn $S_X > 0$, dann rechtsschief

Symmetrie / Schiefe - Beispiel

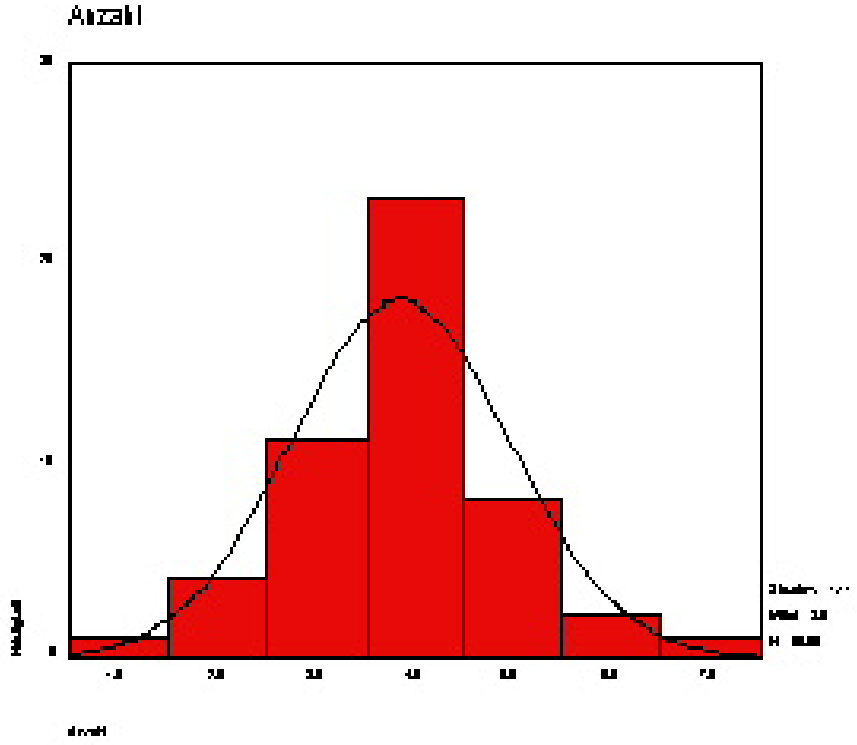
Schiefemaß nach CHARLIER S_x aus der Urliste:

$$
\begin{aligned}
S_x &= \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} \\
\ldots &\quad \frac{0.126912}{\sqrt{1.2004^3}} = 0.096496977 \\
mit\ m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1.2004 \\
m_{\overline{x}}^3 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 = 0.126912
\end{aligned}
$$

$S_X \approx 0$, deshalb annähernd symmetrisch
$S_X > 0$, deshalb leicht rechtsschief

Symmetrie / Schiefe Vorlesungsbeispiel 1

Histogramm mit Normalverteilungskurve:

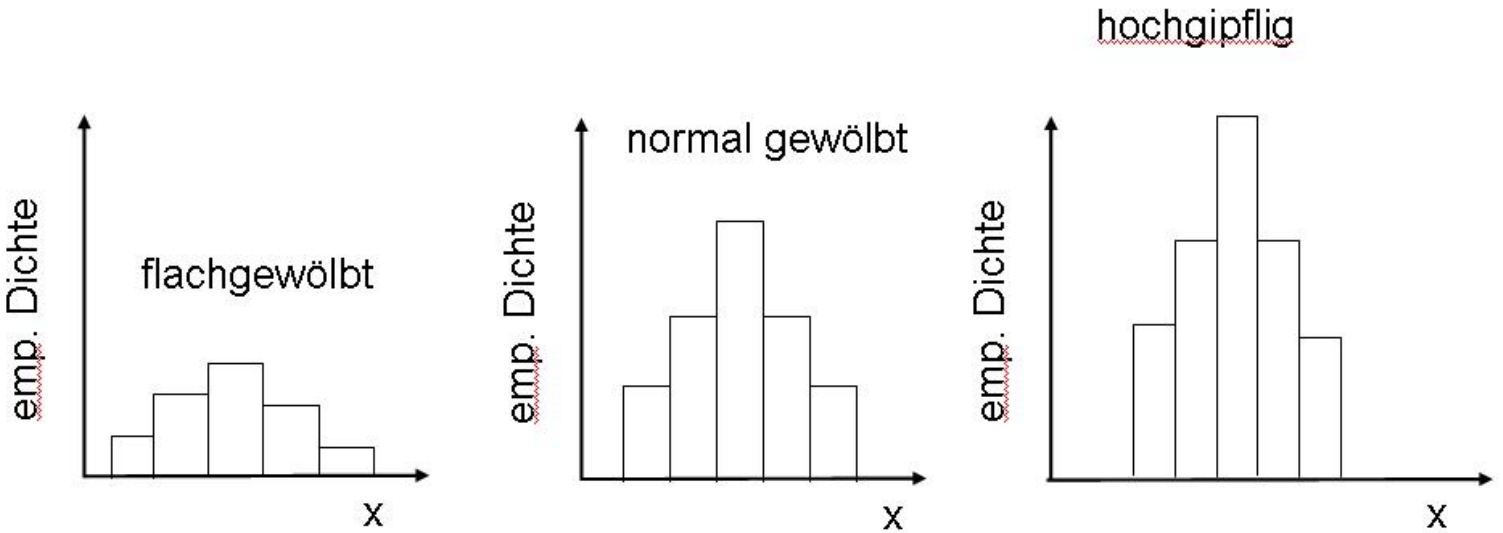


2.2.3 Formenmaße - Gliederung

- 2.2.3.1 Schiefe
- **2.2.3.2 Wölbung**
- 2.2.3.3 empirische Momente

Wölbung grafische Darstellung

Wölbung einer Verteilung einer empirischen Verteilung (Histogramm)



Wölbung Maßzahlen

Maßzahlen zur Messung der Wölbung (kurtosis) aus der Urliste:
Wölbungsmaß nach CHARLIER W_x:

$$
\begin{aligned}
W_x &= \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 \\
mit\ m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
m_{\overline{x}}^4 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4
\end{aligned}
$$

Interpretation:

- wenn $W_X < 0$, dann flach gewölbt
- wenn $W_X = 0$, dann normal gewölbt
- wenn $W_X > 0$, dann stark gewölbt

Wölbung Maßzahlen

Maßzahlen zur Messung der Wölbung (kurtosis) aus Häufigkeitsverteilung (disjunkte Merkmalsausprägungen):
Wölbungsmaß nach CHARLIER W_x:

$$
\begin{aligned}
W_x &= \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 \\
mit\ m_{\overline{x}}^2 &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot p_i \\
m_{\overline{x}}^4 &= \sum_{i=1}^{n} (x_i - \overline{x})^4 \cdot p_i
\end{aligned}
$$

Interpretation:

- wenn $W_X < 0$, dann flach gewölbt
- wenn $W_X = 0$, dann normal gewölbt
- wenn $W_X > 0$, dann stark gewölbt

Wölbung - Vorlesungsbeispiel 1

Wölbung nach CHARLIER W_x aus der Urliste:

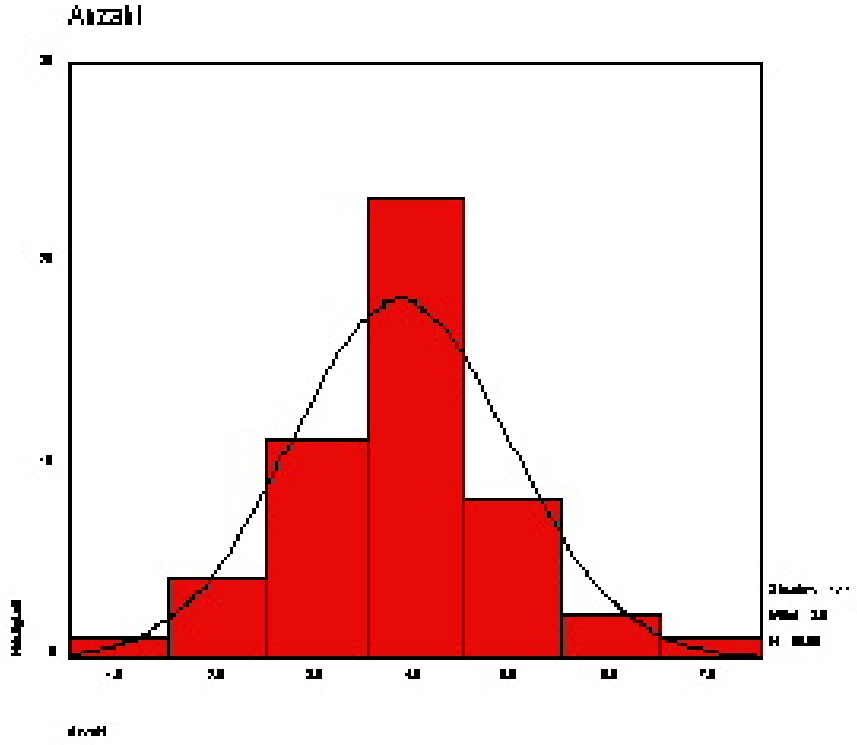
$$
\begin{aligned}
W_x &= \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 = \\
\ldots &\quad \frac{5.46951952}{1.2004^2} - 3 = 0.795746525 \\
mit\ m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1.2004 \\
m_{\overline{x}}^4 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 = 5.46951952
\end{aligned}
$$

Interpretation:

$W_x = 0.795746525 > 0 \Rightarrow$ stärker als normal gewölbt

Wölbung - Vorlesungsbeispiel 1

Histogramm mit Normalverteilungskurve:



2.2.3 Formenmaße - Gliederung

- 2.2.3.1 Schiefe
- 2.2.3.2 Wölbung
- **2.2.3.3 empirische Momente**

empirische Momente

Empirisches Moment um c der Ordnung r

1. aus der Urliste

$$m_c^r = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^r$$

2. aus einer Häufigkeitsliste (ohne Intervallbildung)

$$m_c^r = \sum_{l=1}^{m}(\xi_l - c)^r \cdot p_l$$

3. aus einer Häufigkeitsliste (mit Intervallbildung)

$$m_c^r \approx \sum_{l=1}^{m}(\xi_l^* - c)^r \cdot p_l \text{ mit } \xi_l^* - \textit{Klassenmitte}$$

empirische Nullmomente

Empirisches Moment um 0 der Ordnung r aus der Urliste

$$m_0^r = \frac{1}{n}\sum_{i=1}^{n}(x_i - 0)^r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

r =1: erstes Nullmoment = arithmetisches Mittel:

$$m_0^1 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

empirische Zentralmomente

Empirisches Moment um $\overline{x}$ der Ordnung r aus der Urliste

$$m_{\overline{x}}^{r} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$$

r =2: zweites Zentralmomentt = empirische Varianz:

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

weitere wichtige Zentralmomente für Berechnung der Schiefe und Wölbung nach CHARLIER

$$r = 3: \quad m_{\overline{x}}^{3} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

$$r = 3: \quad m_{\overline{x}}^{4} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$$

Null- und Zentralmomente - Vorlesungsbeispiel 1

Urliste / Häufigkeitstabelle:

$$
\begin{aligned}
m_0^1 &= \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3.86 \\
m_{\overline{x}}^2 &= d_X = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = 1.2004 \\
m_{\overline{x}}^3 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^3 = 0.126912 \\
m_{\overline{x}}^4 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^4 = 5.46951952
\end{aligned}
$$

Null- und Zentralmomente - Vorlesungsbeispiel 2

Datensatz mit Intervallbildung:

$$
\begin{aligned}
m_0^1 &= \overline{x} = \sum_{i=1}^{n} \xi_i^* \cdot p_i = 2.4 \\
m_{\overline{x}}^2 &= d_X = \sum_{i=1}^{n} (\xi_i^* - \overline{x})^2 \cdot p_i = 4.16 \\
m_{\overline{x}}^3 &= \sum_{i=1}^{n} (\xi_i^* - \overline{x})^3 \cdot p_i = 15.384 \\
m_{\overline{x}}^4 &= \sum_{i=1}^{n} (\xi_i^* - \overline{x})^4 \cdot p_i = 124.0064 \\
\text{mit } \xi_i^* &- \textit{Klassenmitte}
\end{aligned}
$$

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Häufigkeiten - Fortsetzung
- Deskriptive Statistik

Formenmaße - Beispiel

mit Excel:

Beispiel 1 (aus Urliste): BeispielW95.xls
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:
Beispiel 3 (gepoolter Datensatz):Zinsätze 2007_4.xls

Aufgaben: Bestimmen Schiefe, Kurtosis, Momente

mit SPSS:

Beispiel 1 (mit Gruppenbilung unter SPSS): World95.sav.xls
Beispiel 2 (aus Klassifizierten Daten): Zinsätze 2007_1.xls:

Aufgaben: Bestimmen Schiefe, Kurtosis, Momente

Maßzahlen der Verteilung

- 2.2.1 Lagemaße (Maße der zentralen Tendenz)
- 2.2.2 Streuungsmaße
- 2.2.3 Formenmaße
- **2.2.4 lineare Transformationen**

2.2.4 lineare Transformationen - Gliederung

- **2.2.4.1 Einführung**
- 2.2.4.2 Standardisierung
- 2.2.4.3 standardisierte Maßzahlen

lineare Transformation

- x_i - beobachtete Merkmalswerte
- y_i - transformierte Merkmalswerte
- $y_i = a + b \cdot x_i$ mit $b \neq 0$ - Transformationsforschrift

transformierte Verteilungdsparameter

arithmetisches Mittel	–	$\overline{y} = a + b\overline{x}$
empirische Varianz	–	$d_y^2 = b^2 \cdot d_x^2$
empirische Standardabweichung	–	$d_y = \lvert b \rvert \cdot d_x$
Schiefe	–	$S_y = S_x$
Woelbung	–	$W_y = W_x$

transformierte Verteilungdsparameter - Herleitung

$y_i = a + b \cdot x_i$ mit $b \neq 0$ - Transformationsforschrift

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \cdot \sum_{i=1}^{n} (a + b \cdot x_i) = a + b \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = a + b \cdot \bar{x}$$

$$d_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} (a + b \cdot x_i - a - b \cdot \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} b^2 \cdot (x_i - \bar{x})^2 = b^2 \cdot$$

$$d_y = \sqrt{d_y^2} = |b| \cdot d_x$$

transformierte Verteilungdsparameter - Herleitung

$y_i = a + b \cdot x_i$ mit $b \neq 0$ - Transformationsforschrift
Schiefe - $S_y = S_x$

$$
\begin{aligned}
m_{\overline{y}}^3 &= b^3 \cdot m_{\overline{x}}^3 \\
\sqrt{(m_{\overline{y}}^2)^3} &= b^3 \cdot \sqrt{(m_{\overline{x}}^2)^3} \\
S_y &= \frac{m_{\overline{y}}^3}{\sqrt{(m_{\overline{y}}^2)^3}} = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} = S_x
\end{aligned}
$$

transformierte Verteilungdsparameter - Herleitung

$y_i = a + b \cdot x_i$ mit $b \neq 0$ - Transformationsforschrift
Woelbung - $W_y = W_x$

$$
\begin{aligned}
m_{\overline{y}}^4 &= b^4 \cdot m_{\overline{x}}^4 \\
(m_{\overline{y}}^2)^2 &= b^4 \cdot (m_{\overline{x}}^2)^2 \\
W_y &= \frac{m_{\overline{y}}^4}{(m_{\overline{y}}^2)^2} - 3 = \frac{b^4 \cdot m_{\overline{x}}^4}{b^4 \cdot (m_{\overline{x}}^2)^2} - 3 = \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 = W_x
\end{aligned}
$$

2.2.4 lineare Transformationen - Gliederung

- 2.2.4.1 Einführung
- **2.2.4.2 Standardisierung**
- 2.2.4.3 standardisierte Maßzahlen

Standardisierung

- x_i - beobachtete Merkmalswerte
- z_i - transformierte Merkmalswerte
- $z_i = \frac{x_i - \overline{x}}{d_x}$ - Transformationsforschrift

wegen

$a = -\frac{\overline{x}}{d_x}$ und $b = \frac{1}{d_x}$

$z_i = \frac{x_i - \overline{x}}{d_x}$ - Spezielle lineare Transformetion

2.2.4 lineare Transformationen - Gliederung

- 2.2.4.1 Einführung
- 2.2.4.2 Standardisierung
- **2.2.4.3 standardisierte Maßzahlen**

standardisierte Maßzahlen

$z_i = \frac{x_i - \overline{x}}{d_x}$ - Transformationsforschrift

- $\overline{z} = 0$ - arithmetisches Mittel
- $d_z = 1$ - transformierte Merkmalswerte
- dimensionslos

$z_i = \frac{x_i - \overline{x}}{d_x}$ - Transformationsforschrift

$$
\begin{aligned}
\overline{z} &= \frac{1}{n} \cdot \sum_{i=1}^{n} z_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{d_x}\right) = \frac{1}{n} \cdot \frac{0}{d_x} = 0 \\
d_z^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (z_i - \overline{z})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} z_i^2 = \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{d_x}\right)^2 = \frac{1}{d_x^2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = 1 \\
d_z &= \sqrt{d_z^2} = \sqrt{1} = 1
\end{aligned}
$$

Interpretation Schiefe - Wölbung

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

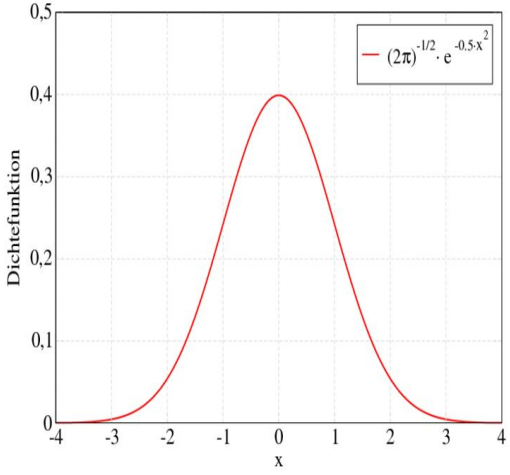
$$\frac{1}{n} \cdot \frac{1}{d_x^3} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^3 = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} = S_x$$

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{d_x}\right)^4 = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(x_i - \overline{x})^4}{d_x^4} =$$

$$\frac{1}{n} \cdot \frac{1}{d_x^4} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^4 = \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} = W_x + 3$$

Exkurs: Normalverteilung und Maßzahlen der Verteilung

Normalverteilung = idealtypische symmetrische Verteilung mit normaler Wölbung
Standard-Normalverteilung: zusätzlich (Mittelwert = 0, Standardabweichung =1)



Paramater der Verteilung:

Mittelwert = 0
Standardabweichung = 1
Schiefe = 0
Wölbung = 0