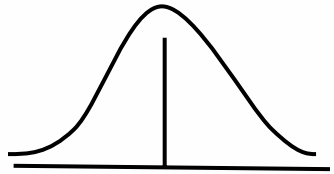
Lageregeln

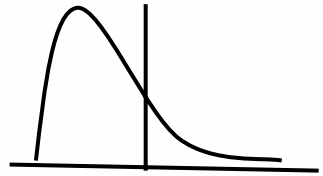
Bei symmetrischen Verteilungen:

$$\bar{x} \approx \tilde{x} \approx x_{\text{mod}}$$



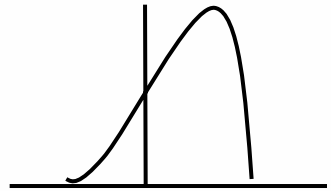
Linkssteile Verteilungen:
(rechtsschiefe Verteilungen)

$$\bar{x} > \tilde{x} > x_{\text{mod}}$$



Rechtssteile Verteilungen:
(linksschiefe Verteilungen)

$$\bar{x} < \tilde{x} < x_{\text{mod}}$$



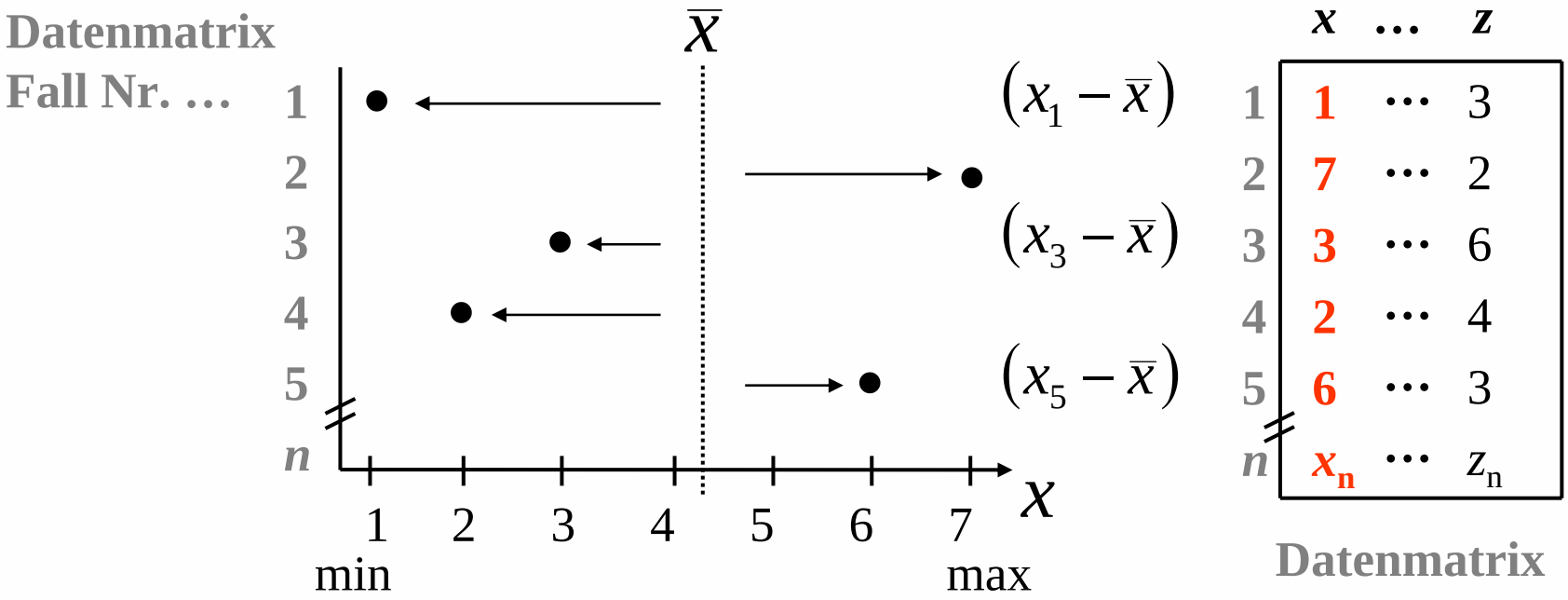
Im Beispiel:

Zeitauto: $\bar{x} = 28{,}8$ $\tilde{x} = 20{,}0$ $x_{\text{mod}} = 15$

Die **Varianz** einer Messwertreihe $x_1, \ldots, x_n$ ergibt sich als ..

$$s_x^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right]$$

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


Datenmatrix
Fall Nr. …
$\bar{x}$
$(x_1 - \bar{x})$
$(x_3 - \bar{x})$
$(x_5 - \bar{x})$
1
2
3
4
5
n
1 2 3 4 5 6 7
min
max
x
x … z
1 1 … 3
2 7 … 2
3 3 … 6
4 2 … 4
5 6 … 3
n x_n … z_n
Datenmatrix

Variation
Sum of Squares

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

Varianz

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

Standardabweichung

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

Beispiel: Variable „Zeitauto“

Varianz

$$s_x^2 = 512{,}83$$

Durchschnittliche Streuung
im
Quadrat der Skaleneinheit

Standardabweichung

$$s_x = \sqrt{512{,}82} = 22{,}65$$

Durchschnittliche Streuung,
in der Skaleneinheit
der Variable [hier: Minuten]

Wenn x in etwa normalverteilt ist, gilt:

$\bar{x} \pm s_x$ umfaßt ca. 68% aller Daten

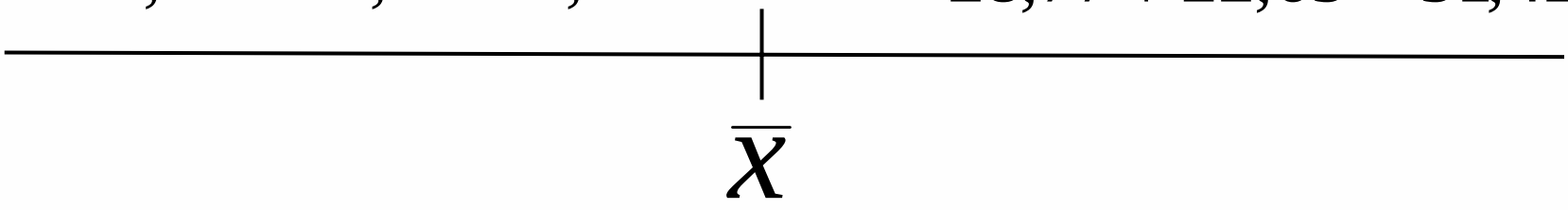
$\bar{x} \pm 2s_x$ umfaßt ca. 95% aller Daten

$\bar{x} \pm 3s_x$ umfaßt ca. 99% aller Daten

Im Beispiel:

$\bar{x} - s_x$ $\quad$ $\bar{x} + s_x$

$= 28{,}77 - 22{,}65 = 6{,}12$ $\quad$ $= 28{,}77 + 22{,}65 = 51{,}42$



Nur heuristisch, da die Variable „Zeitauto“ nicht normalverteilt ist

Variationskoeffizient

$$V_x = \frac{s_x}{\bar{x}} \qquad \bar{x} > 0$$

Maßstabsunabhängiges Streuungsmaß für Merkmale
mit nichtnegativen Ausprägungen
und arithmetischem Mittel größer Null,
das zum Vergleich unterschiedlicher Streuungen geeignet ist

Momente

Erwartungswert (Mittelwert) von X^k = k^{tes} Moment um den Ursprung

1. Moment um den Ursprung = Mittelwert

Momente um den Mittelwert

Erwartungswert (Mittelwert) von $\left(x_i - \bar{x}\right)^k$ = k^{tes} Moment um den Mittelwert

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

Vgl. z.B. H.M. Blalock (1981) Social Statistics. Auckland, p. 139

Momente **um den Mittelwert**

(Durchschnitte von Potenzen von Differenzen vom Mittelwert)

1. Moment um den Mittelwert gleich Null

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

2. Moment um den Mittelwert= Varianz

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

3. Moment um den Mittelwert

wobei zusätzlich durch die 3. Potenz der Standardabweichung geteilt wird

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

Positive Schiefewerte -> rechtsschiefe
Negative Schiefewerte = linksschiefe
Null -> symmetrische Verteilung

Positive Werte -> steile
Negative Werte = flache Verteilung

$$Steilheit=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{s_x^4}-3$$

Wie lange benötigen Sie bzw. würden Sie für Ihren täglichen Weg (einfache Strecke) benötigen, ...

... wenn Sie ausschließlich **das Auto** benutzen?

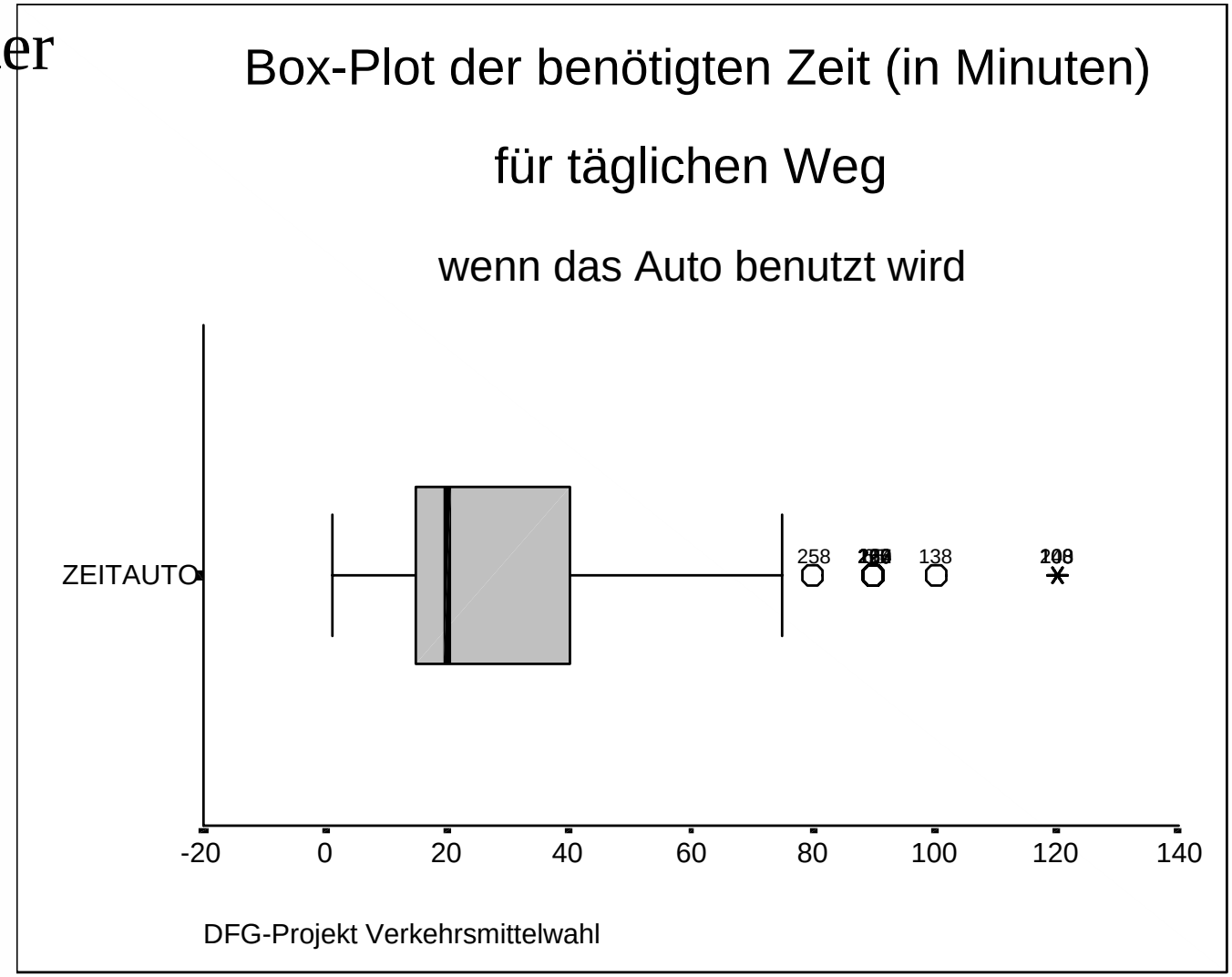
(In Minuten)

Häufigkeitstabelle für ZEITAUTO

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente	
Gültig	1,00	2	,7	,7	,7	
	3,00	1	,3	,3	1,0	
	4,00	2	,7	,7	1,6	
	5,00	16	5,3	5,3	6,9	
	6,00	2	,7	,7	7,6	
	7,00	6	2,0	2,0	9,5	
	8,00	5	1,6	1,6	11,2	
	10,00	38	12,5	12,5	23,7	
	12,00	2	,7	,7	24,3	
	14,00	1	,3	,3	24,7	
	15,00	46	15,1	15,1	39,8	Q_1
	17,00	1	,3	,3	40,1	
	20,00	41	13,5	13,5	53,6	Q_2
	25,00	18	5,9	5,9	59,5	
	27,00	1	,3	,3	59,9	
	28,00	1	,3	,3	60,2	
	30,00	30	9,9	9,9	70,1	
	35,00	10	3,3	3,3	73,4	
	40,00	11	3,6	3,6	77,0	Q_3
	45,00	20	6,6	6,6	83,6	
	50,00	6	2,0	2,0	85,5	
	55,00	1	,3	,3	85,9	
	60,00	22	7,2	7,2	93,1	
	65,00	1	,3	,3	93,4	
	70,00	1	,3	,3	93,8	
	75,00	6	2,0	2,0	95,7	
	80,00	1	,3	,3	96,1	
	90,00	8	2,6	2,6	98,7	
	100,00	1	,3	,3	99,0	
	120,00	3	1,0	1,0	100,0	
	Gesamt	304	100,0	100,0		
Gesamt		304	100,0			

Quelle: dto.

Modifizierter **Box-Plot**



Box-Plot:

» Anfang der Box: 1. Quartil (15 Min.)

» Median: Vertikaler Strich innerhalb der Box (20 Min).

» Ende der Box: 3. Quartil (40 Min.)

» Länge der Box = Quartilabstand (40-15 = 25)

» Zwei Linien außerhalb der Box gehen zum kleinsten und größten *x*-Wert

$$x_{\min}, Q_{0.25}, \tilde{x}, Q_{0.75}, x_{\max}$$

Modifizierter Box-Plot:
Die Linien außerhalb der Box werden nur bis zum kleinsten und größten *x*-Wert gezogen, falls diese innerhalb der Zäune liegen. Ansonsten gehen die Linien bis zum kleinsten bzw. größten Wert innerhalb der Zäune. Außerhalb der Zäune liegen die **Ausreißer** und **Extremfälle**.

Ausreißer: Fälle mit Werten, die zw. 1,5 und 3 Boxlängen vom oberen oder unteren Rand der Box entfernt sind.

Extremfälle: Fälle mit Werten, die mehr als 3 Boxlängen vom oberen oder unteren Rand des Balkens entfernt sind.

Devianz

$$D = -2\sum_{i=1}^{I} \ln\left(\frac{n_i}{n}\right) \cdot n_i$$

» **Maximal**wert bei Gleichverteilung

» **Null**, wenn nur eine Kategorie der Variable besetzt ist

Relative Devianz

$$d = -2\sum_{i=1}^{I} \ln\left(\frac{n_i}{n}\right) \cdot \left(\frac{n_i}{n}\right) = -2\sum_{i=1}^{I} \ln(p_i) \cdot p_i$$

Entropie (Informationsgehalt)

$$H = -\sum_{i=1}^{I} \frac{n_i}{n} \ln \frac{n_i}{n}$$

$H_{min} = 0$ Alle Fälle befinden sich in derselben Kategorie der Variablen; Einpunktverteilung; minimale Unsicherheit in der Vorhersage der Beobachtungswerte

$H_{max} = \ln I$ Fälle verteilen sich gleich über die Kategorien; alle Kategorien sind gleich häufig besetzt;

(H variiert mit der Anzahl I der Kategorien; das erschwert den Vergleich der Streuungen von Verteilungen von Variablen mit unterschiedlicher Anzahl von Kategorien)

Relative Entropie

$$RH = \frac{H}{\ln I}$$

$H_{min} = 0$ Alle Fälle befinden sich in derselben Kategorie der Variablen; Einpunktverteilung

$H_{max} = 1$ Fälle verteilen sich gleich über die Kategorien;

Vgl. z.B.: Jann (2002) Einführung in die Statistik. München, S. 50-52 oder Clauß/Finze/Partsch (1999) Statistik. Thun, S. 48f.

Entropie

Zahlen-beispiele ..

	p_i	p_i	p_i	p_i
1	0,5	0,4	0,3	0,2
2	0,5	0,6	0,7	0,8
	0,69	**0,67**	**0,61**	**0,50**

ln *I* = ln 2

.. unter Verwendung des natürlichen Logarithmus *ln*

z.B.: $-1\times[(0{,}5\times\log_e(0{,}5))+(0{,}5\times\log_e(0{,}5))]=0{,}69$
z.B.: $-1\times[(0{,}4\times\log_e(0{,}4))+(0{,}6\times\log_e(0{,}6))]=0{,}67$

Proportionale Abweichung der beobachteten Verteilung von max. Streuung (Gleichverteilung)

$$\frac{\ln I - H}{\ln I}$$

	Gruppe 1	Gruppe 2	p_i
1	0,75	0,17	0,4
2	0,25	0,83	0,6
Entropie	**0,56**	**0,46**	**0,67**
Entropie (gewichtet)	0,56*0,40 = 0,22	0,46*0,60 = 0,28	
n	40	60	100

$$Pseudo - R^2 = \frac{S_y^2 - S_e^2}{S_y^2} = \frac{0{,}67 - (0{,}22 + 0{,}28)}{0{,}67} = \frac{0{,}17}{0{,}67} = 0{,}25$$