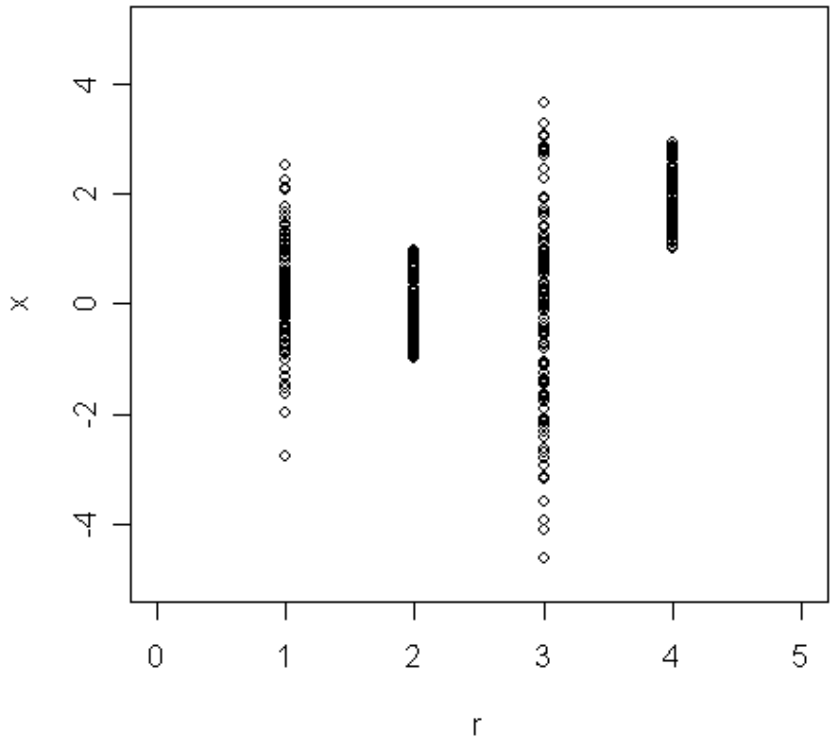
Wirtschaftsstatistik

Claudia Redenbach

SS 2010

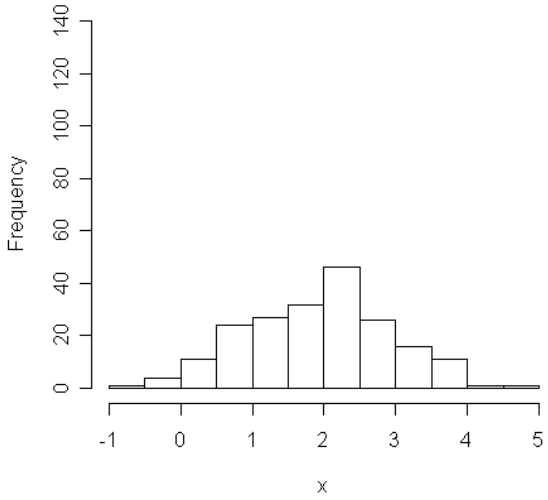
Lagemaßzahlen



	x_n	x_{med}
1	0.15	0.19
2	0.00	0.00
3	-0.08	0.10
4	1.98	1.99

Verteilungstypen

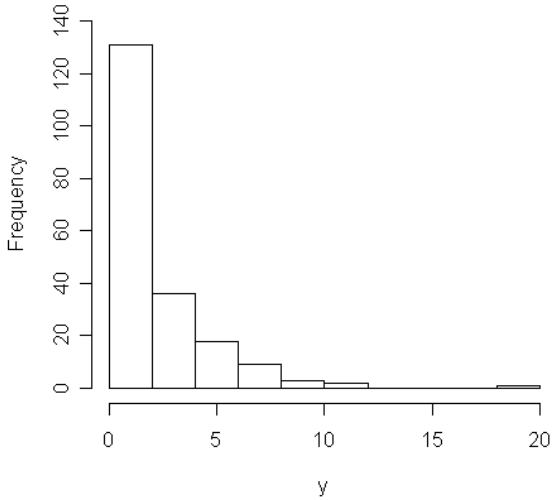
Symmetrisch



x_{med}=2.01

x_n = 1.91

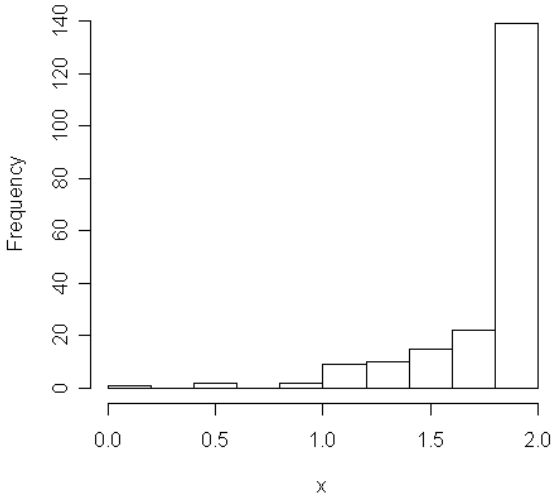
Linkssteil



x_{med}=1.07

x_n = 1.99

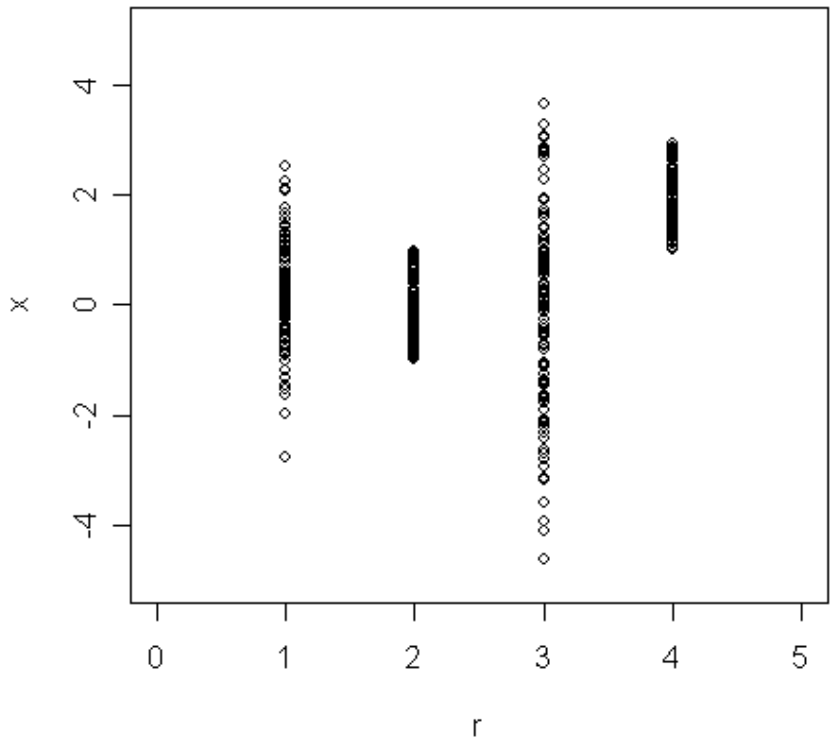
Rechtssteil



x_{med}=1.91

x_n = 1.78

Streuung



	x_n	x_{med}
1	0.15	0.19
2	0.00	0.00
3	-0.08	0.10
4	1.98	1.99

	s^2_n	$(s^2_n)^{0.5}$
1	0.86	0.93
2	0.38	0.61
3	3.51	1.87
4	0.32	0.57

Zur Berechnung von Quantilen

- Beispiel: Autounfälle in einer Stadt pro Tag
- n=14

  20,17,22,21,26,45,25,18,23,25,15,24,30,26
- $z_{0.25}$: 14*0.25 = 3.5, also $z_{0.25} = x_{(4)} = 20$
- $z_{0.75}$: 14*0.75 = 10.5, also $z_{0.75} = x_{(11)} = 26$

- Beachte:

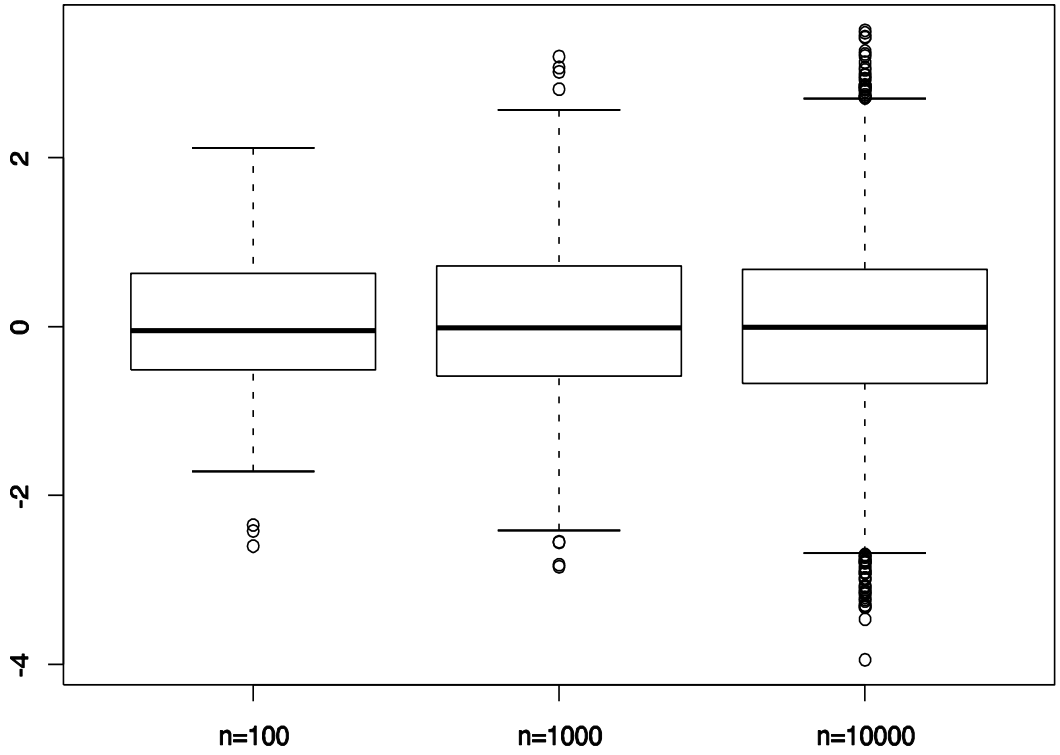
  Die R-Funktion quantile() implementiert eine gegenüber der Vorlesung veränderte Definition der Quantile.

  Sie liefert hier $z_{0.25} = 20.25$ und $z_{0.75} = 25.75$

Boxplots

- Strich: Median
- Box: Quartile
- Whisker: 5% - und 95%-Quantile der Daten
- Kreise: Ausreißer

- Ablesbar: Wertebereich, Symmetrie

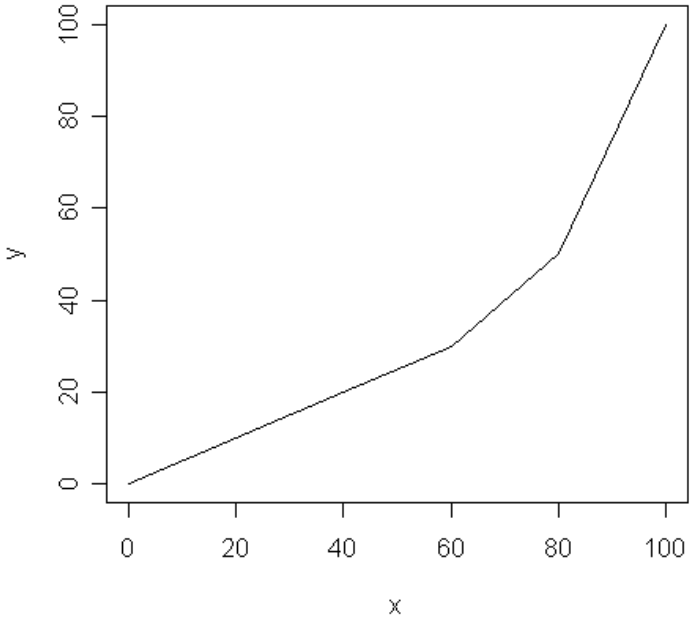


Berechnung der Whisker in R:
Extremster Datenpunkt, der nicht weiter als der 1.5-fache Abstand zwischen den Quartilen von der Box entfernt ist.

Lorenzkurve

Fünf Unternehmen einer Branche

Marktanteile: (10%, 10%, 10%, 20%, 50%)



Lorenzkurve

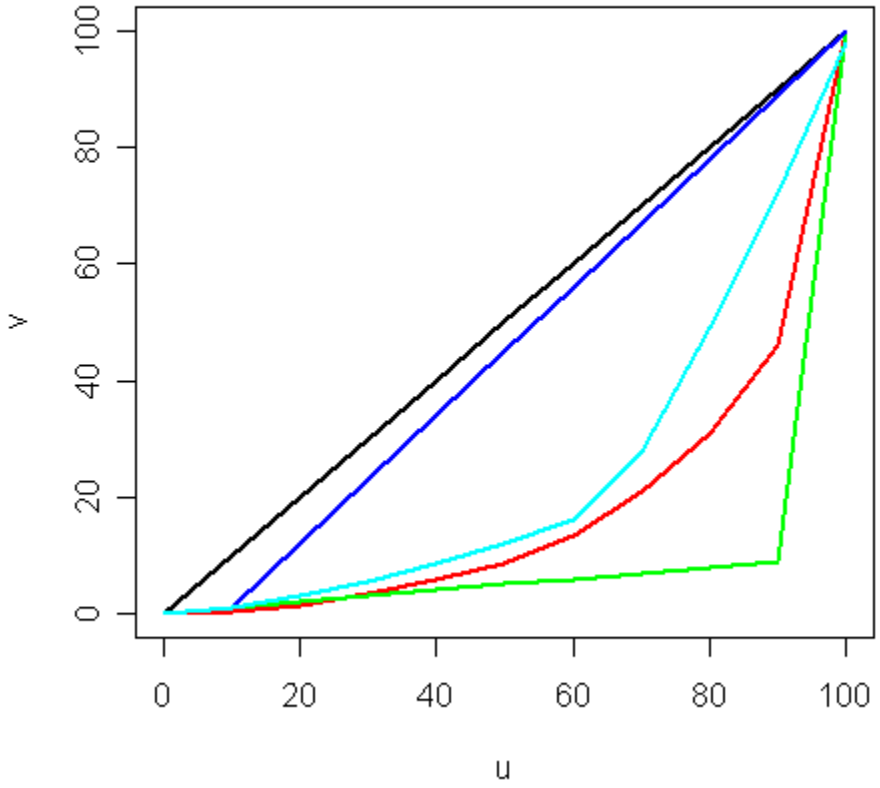
- n=10, Gesamtmerkmalssumme = 10

9 x 0.1, 1 x 9.1

1 x 0.1, 9 x 1.1

(0.05, 0.1, 0.2, 0.25, 0.25,
0.5, 0.75, 1, 1.5, 5.4)

(0.1,0.2,0.25, 0.3, 0.35,
0.4, 1.2, 2.1, 2.3, 2.6)



Gini-Koeffizient

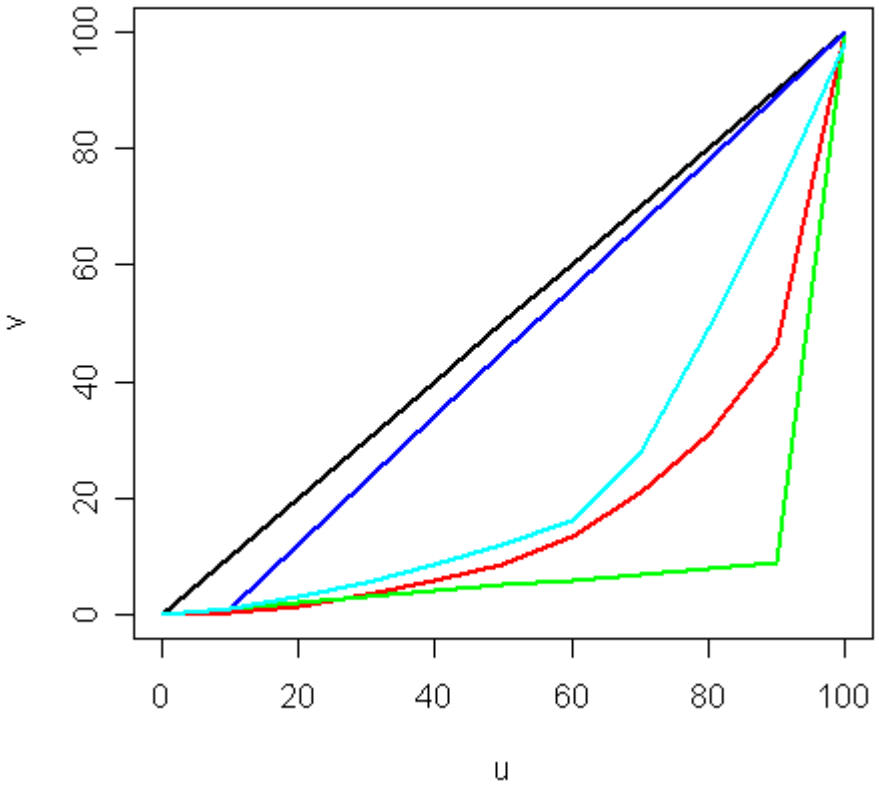
- n=10, Gesamtmerkmalssumme = 10

$\gamma = 0.81$, $\gamma^* = 0.9$

$\gamma = 0.09$, $\gamma^* = 0.1$

$\gamma = 0.64$, $\gamma^* = 0.71$

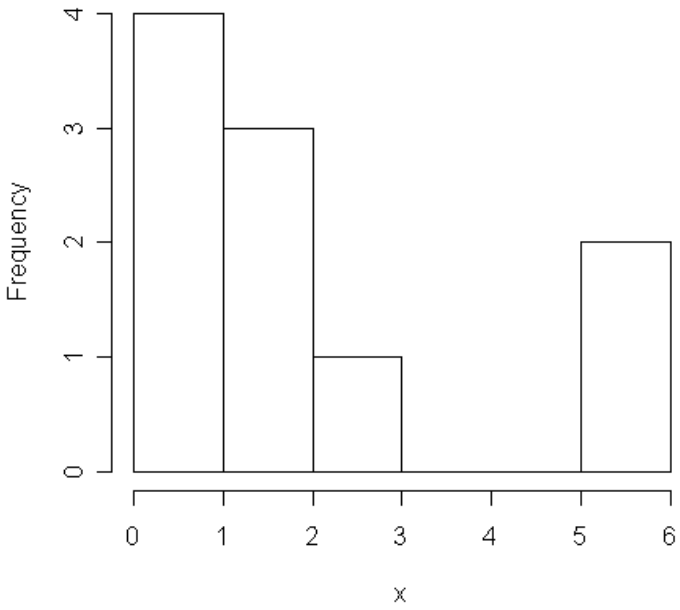
$\gamma = 0.47$, $\gamma^* = 0.52$

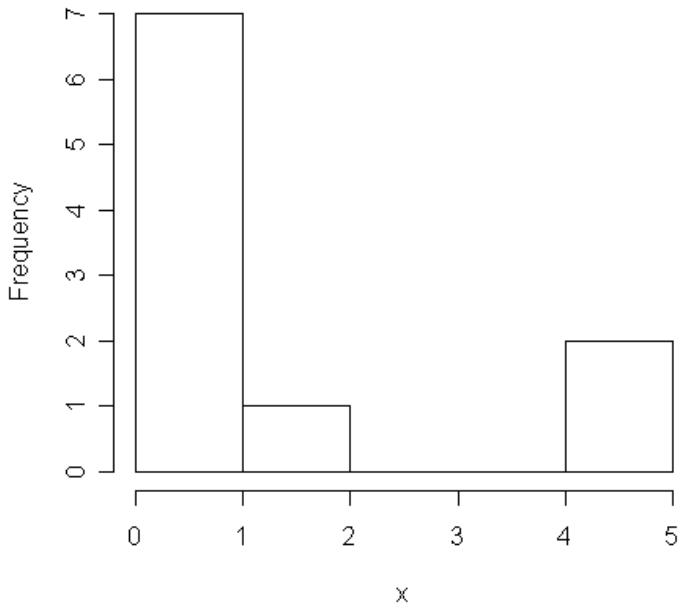


Histogramm

- Zähne

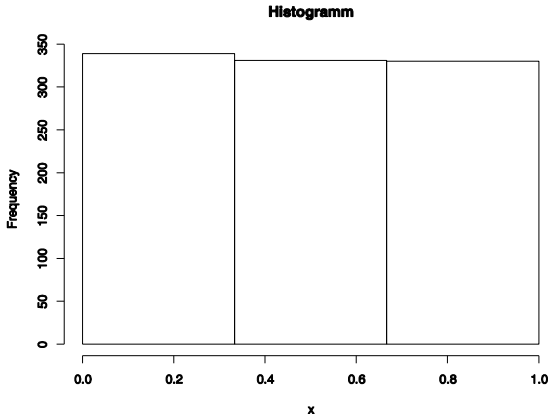
Beachte: R wählt standardmäßig linksoffene Intervalle (Plot rechts) . Die Option right = F liefert den Plot links.

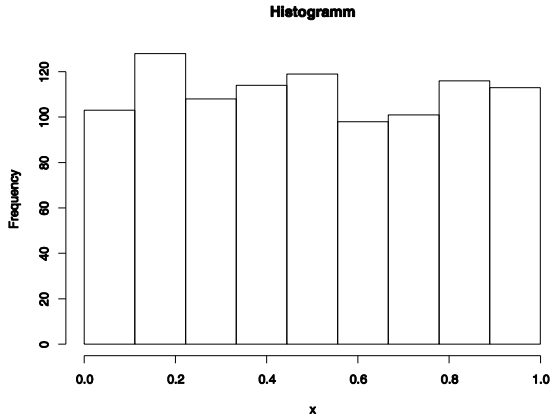


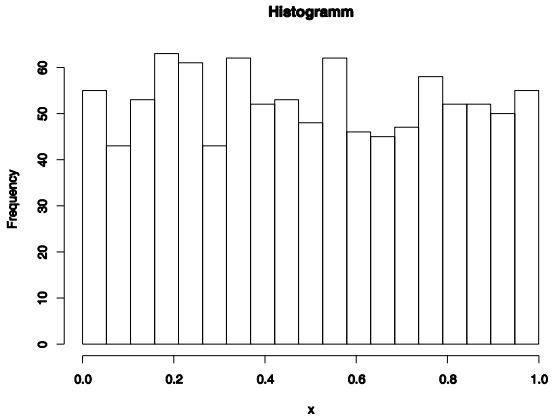


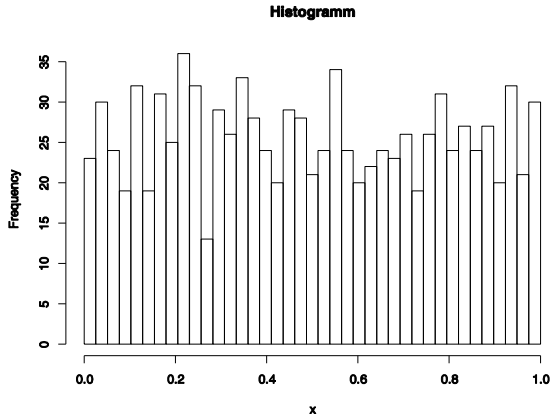
Wahl der Grenzen

- $X_i \sim U(0,1)$
- n=1000





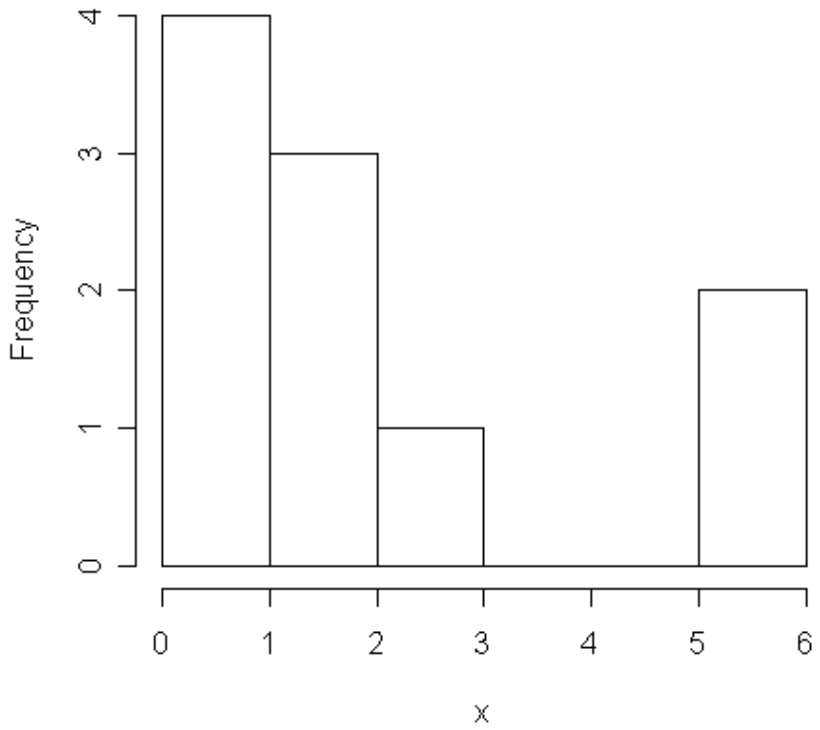


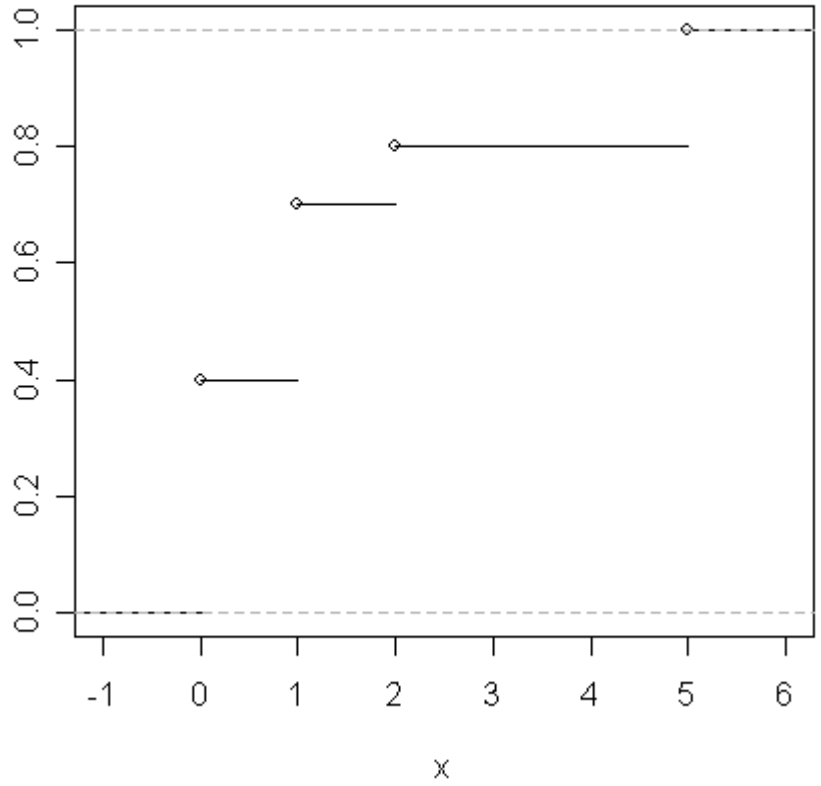


- Gleiche Realisierung in allen 4 Plots

Empirische Verteilungsfunktion

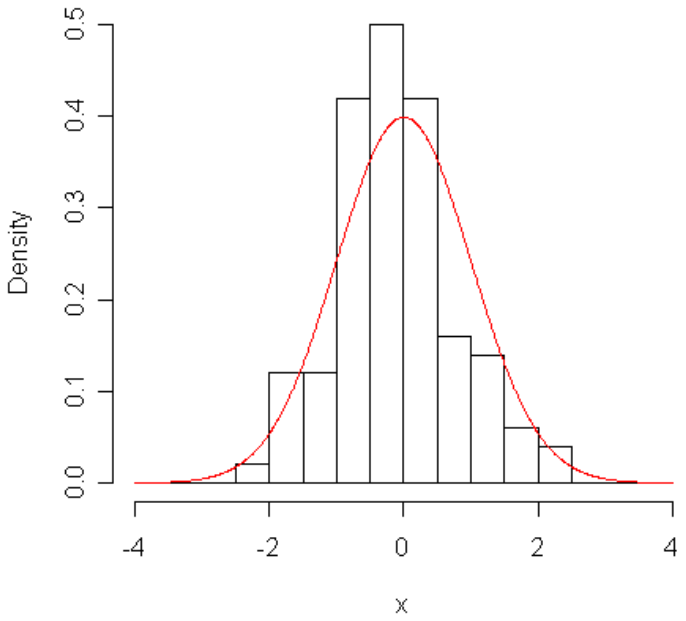
- Zähne

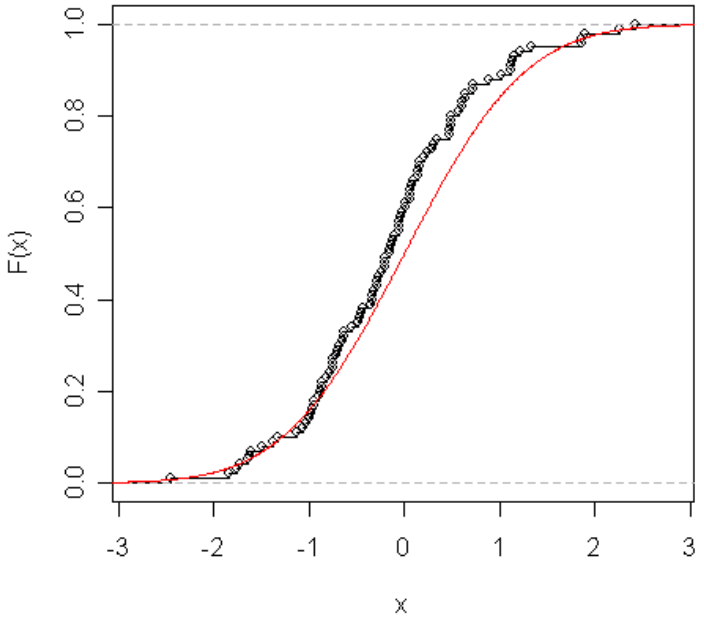




Normalverteilte Daten

- $X_i \sim N(0,1)$, n =100





Kontingenztafel (absolut)

- n = 447 männliche deutsche Arbeitslose
- X: Ausbildungsniveau: keine Ausbildung, Lehre, fachspezifische Ausbildung, Hochschulabschluss
- Y: Dauer der Arbeitslosigkeit: $\leq$ 6 Monate, 7-12 Monate, > 12 Monate

	$\leq$ 6	7-12	>12	
keine Ausbildung	86	19	18	123
Lehre	170	43	20	233
fachspezifische Ausbildung	40	11	5	56
Hochschulabschluss	28	4	3	35
	324	77	46	447

Kontingenztafel (relativ)

- n = 447 männliche deutsche Arbeitslose
- X: Ausbildungsniveau: keine Ausbildung, Lehre, fachspezifische Ausbildung, Hochschulabschluss
- Y: Dauer der Arbeitslosigkeit: $\leq$ 6 Monate, 7-12 Monate, > 12 Monate

	$\leq$ 6	7-12	>12	
keine Ausbildung	0.192	0.043	0.040	0.275
Lehre	0.380	0.096	0.045	0.521
fachspezifische Ausbildung	0.089	0.025	0.011	0.125
Hochschulabschluss	0.063	0.009	0.007	0.078
	0.725	0.172	0.103	1

Kontingenztafel (bedingt)

- n = 447 männliche deutsche Arbeitslose
- X: Ausbildungsniveau: keine Ausbildung, Lehre, fachspezifische Ausbildung, Hochschulabschluss
- Y: Dauer der Arbeitslosigkeit: $\leq$ 6 Monate, 7-12 Monate, > 12 Monate
- Relative bedingte Häufigkeiten $f_Y(\cdot \mid i)$

	$\leq$ 6	7-12	>12	
keine Ausbildung	0.699	0.154	0.147	1
Lehre	0.730	0.184	0.086	1
fachspezifische Ausbildung	0.714	0.197	0.089	1
Hochschulabschluss	0.800	0.114	0.086	1

Kontingenzkoeffizient

n = 100 männliche deutsche Arbeitslose

X: Ausbildungsniveau:
keine Ausbildung, Lehre

Y: Dauer der Arbeitslosigkeit:
7-12 Monate, > 12 Monate

T = 2.826
T‘ = 0.165
T*= 0.234

	d_1	d_2	
c_1	19	18	37
c_2	43	20	63
	62	38	100

Streudiagramm

- Weinanbau im Jahr i:
- x_i: mittlere Anzahl von Beeren je Traube im Juli
- y_i: Ertrag in Tonnen je 100 m²

Jahr	1971	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
x_i	116.37	82.77	110.68	97.50	115.88	80.19	125.24	116.15	117.36	93.31	107.46	122.30
y_i	5.6	3.2	4.5	4.2	5.2	2.7	4.8	4.9	4.7	4.1	4.4	5.4

- 1972: keine Ernte wegen Sturm

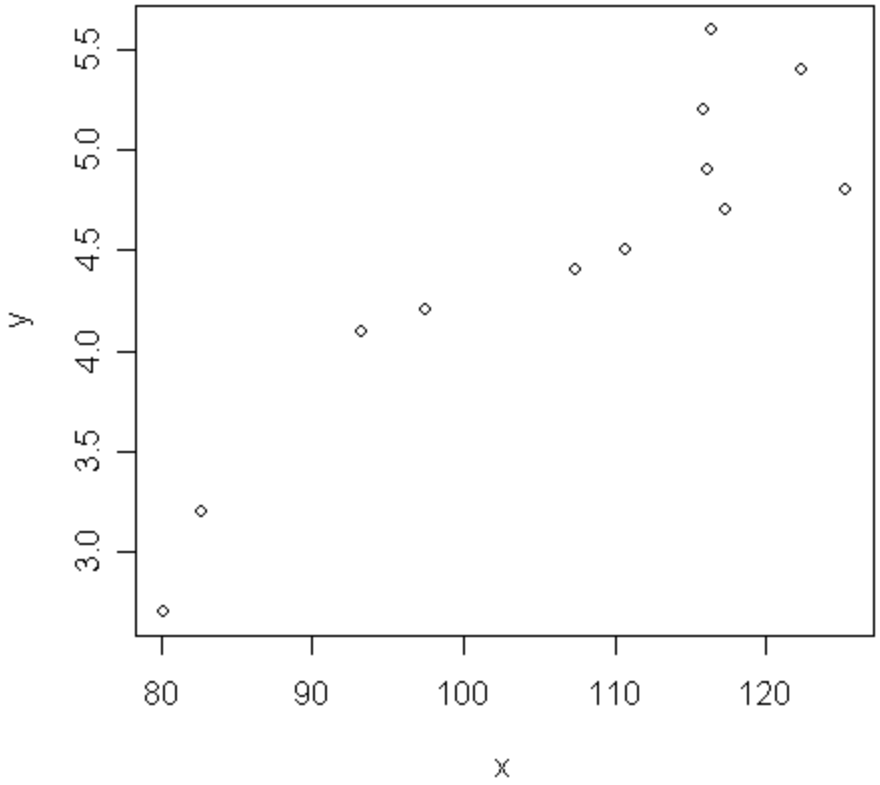
Streudiagramm

- Weinanbau im Jahr i:
- x_i: mittlere Anzahl von Beeren je Traube im Juli
- y_i: Ertrag in Tonnen je 100 m²

- Offenbar besteht ein Zusammenhang zwischen den Variablen X und Y:

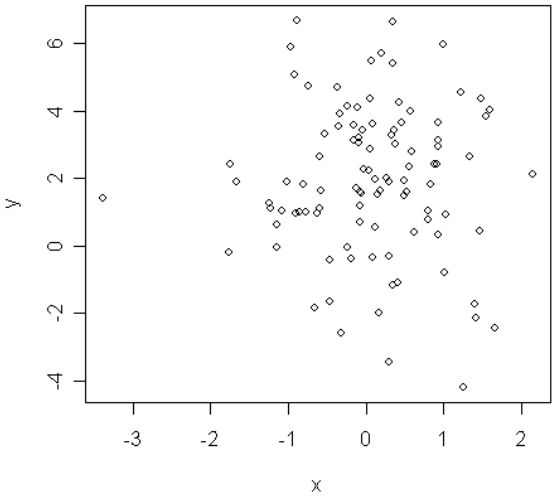
  x groß <-> y groß

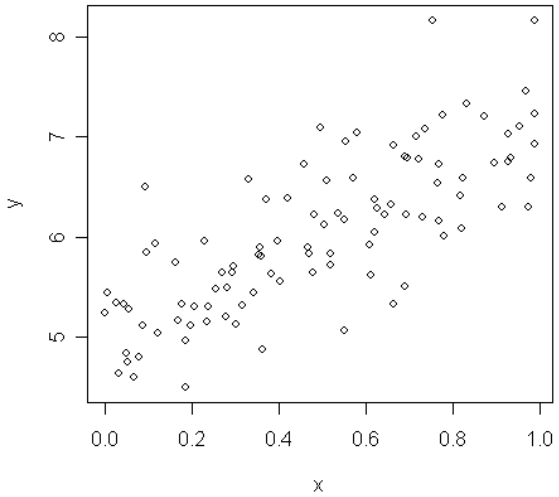
-> Korrelationskoeffizient: 0.91

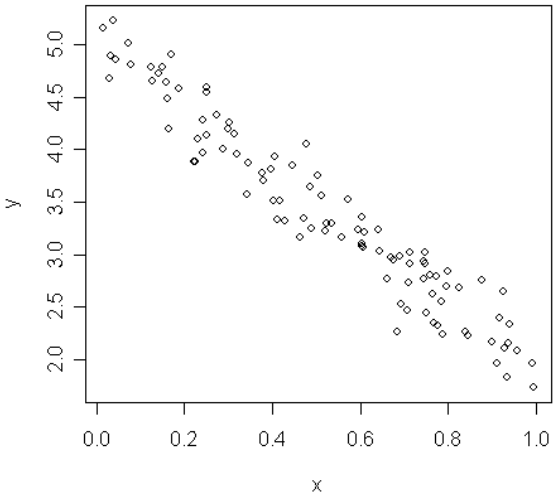


Korrelationskoeffizient

$\rho_{xy} = -0.025$ $\rho_{xy} = 0.773$ $\rho_{xy} = -0.965$



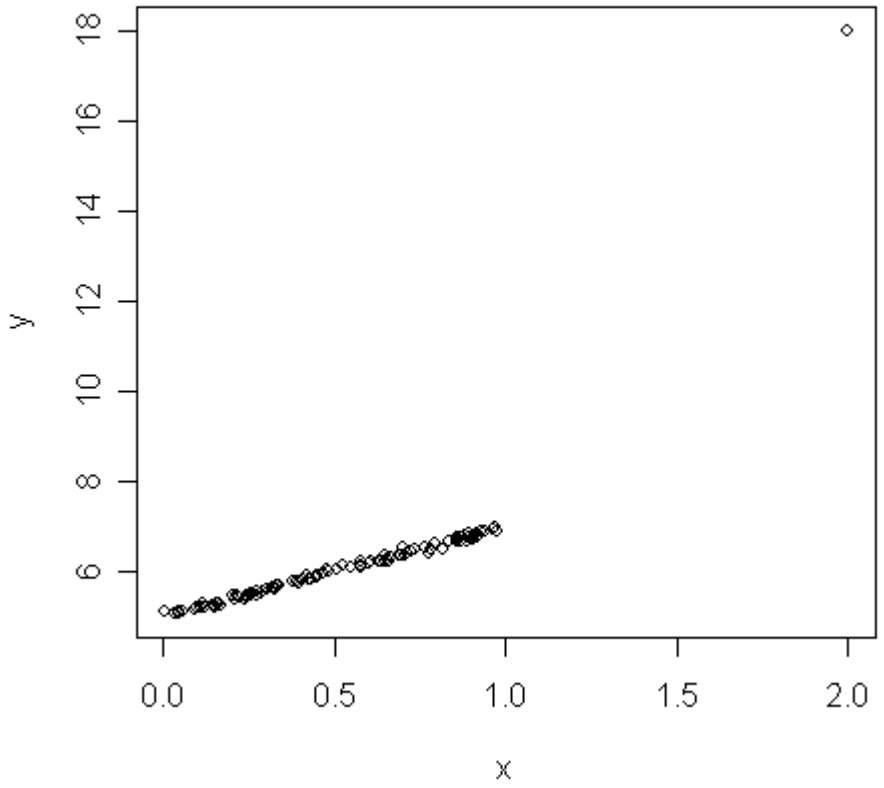




Korrelationskoeffizient

$\rho_{xy} = 0.798$

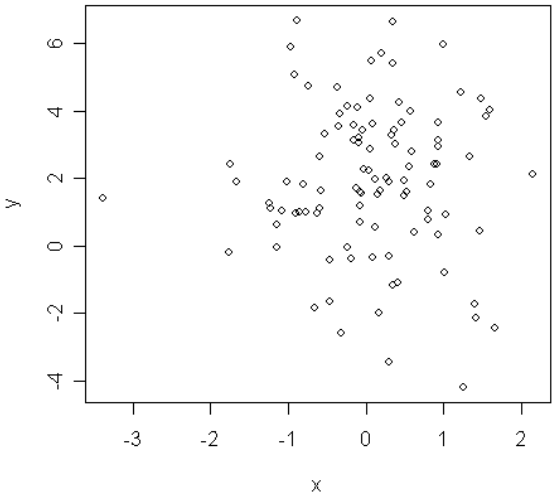
$\rho'_{xy} = 0.996$

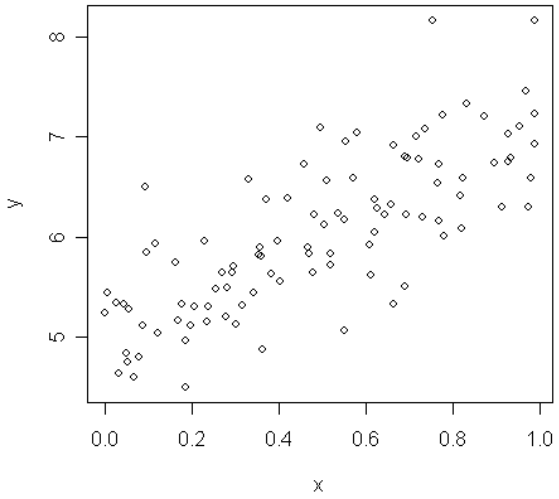


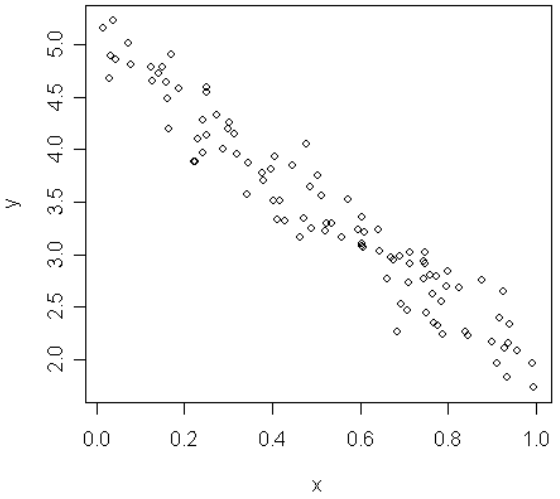
Korrelationskoeffizient

$\rho_{xy} = -0.025$ $\rho_{xy} = 0.773$ $\rho_{xy} = -0.965$

$\rho'_{xy} = 0.033$ $\rho'_{xy} = 0.788$ $\rho_{xy}' = -0.967$







Beispiel: Zusammenhangsanalyse

Daten:

Geburtsgewicht in g von 20 Säuglingen und Gewichtszunahme in g zwischen dem 70. und 100. Tag

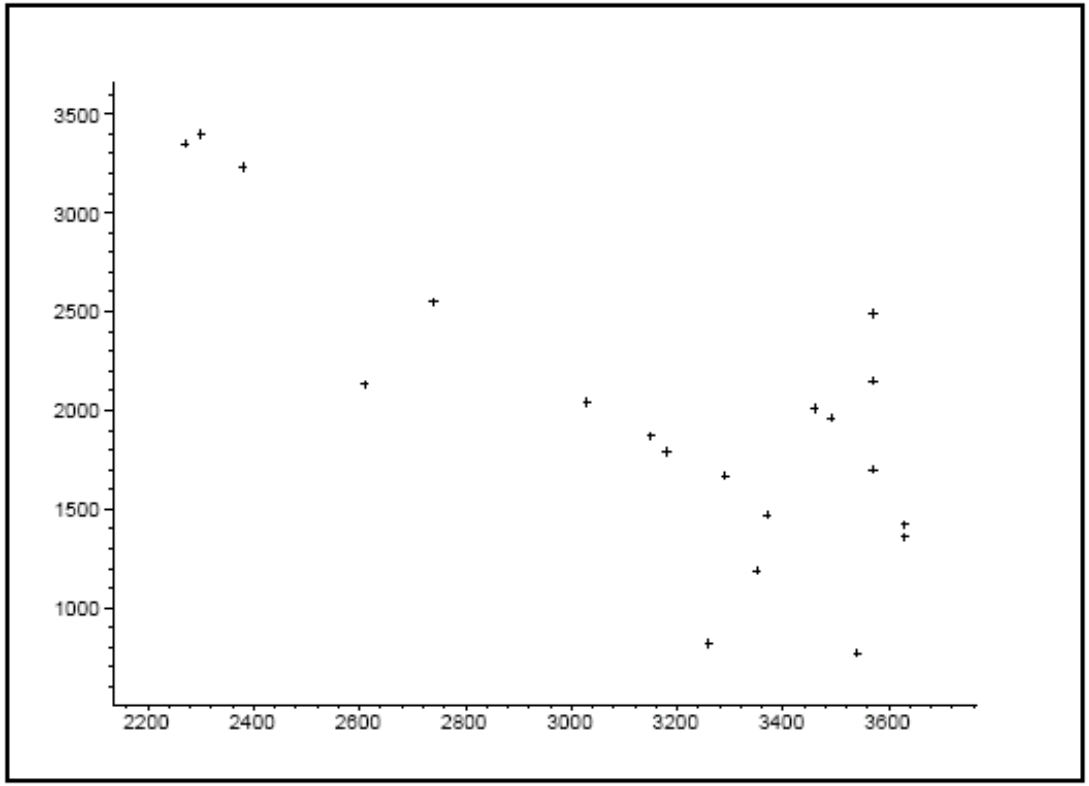
Säugling	Geburts-gewicht x_i	Gewichts-zunahme y_i	Säugling	Geburts-gewicht x_i	Gewichts-zunahme y_i
1	2740	2550	11	3260	820
2	3180	1790	12	3350	1190
3	3150	1870	13	3630	1360
4	3030	2040	14	3630	1420
5	3370	1470	15	3490	1960
6	2610	2130	16	3290	1670
7	3570	2150	17	3540	770
8	2270	3350	18	3570	1700
9	2300	3400	19	3460	2010
10	2380	3230	20	3570	2490

Beispiel (vgl. J. Hüsler, H. Zimmermann (2001) *Statistische Prinzipien für medizinische Projekte.* Huber-Verlag, Bern, S. 183 ff.)

Beispiel: Zusammenhangsanalyse

Streudiagramm

$\rho_{xy} = -0.762$



Beispiel: Zusammenhangsanalyse

Folgerungen:

- Die beiden Merkmale sind negativ korreliert:

  kleines Geburtsgewicht -> große Zunahme
- Die Wertepaare liegen aber nicht sehr nahe an einer Geraden.
- Der lineare Eindruck entsteht vor allem durch Datensätze mit sehr geringem Geburtsgewicht.

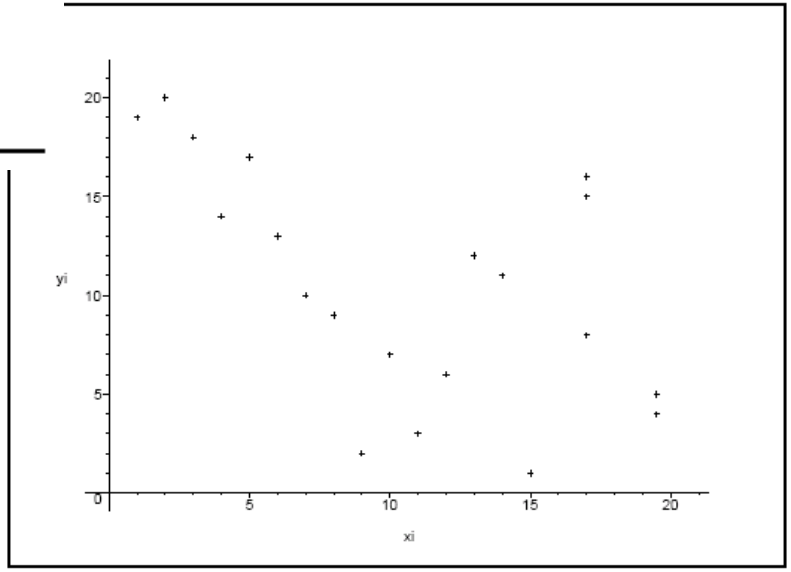
  -> Betrachte die Ränge in dieser Stichprobe

Beispiel: Zusammenhangsanalyse

Säugling	Rang Geb.-gewicht $\mathrm{rg}(x_i)$	Rang Gew.-zunahme $\mathrm{rg}(y_i)$	Säugling	Rang Geb.-gewicht $\mathrm{rg}(x_i)$	Rang Gew.-zunahme $\mathrm{rg}(y_i)$
1	5	17	11	9	2
2	8	9	12	11	3
3	7	10	13	19.5	4
4	6	13	14	19.5	5
5	12	6	15	14	11
6	4	14	16	10	7
7	17	15	17	15	1
8	1	19	18	17	8
9	2	20	19	13	12
10	3	18	20	17	16

$\rho'_{xy} = -0.56$ liegt näher bei 0 als $\rho_{xy} = -0.762$

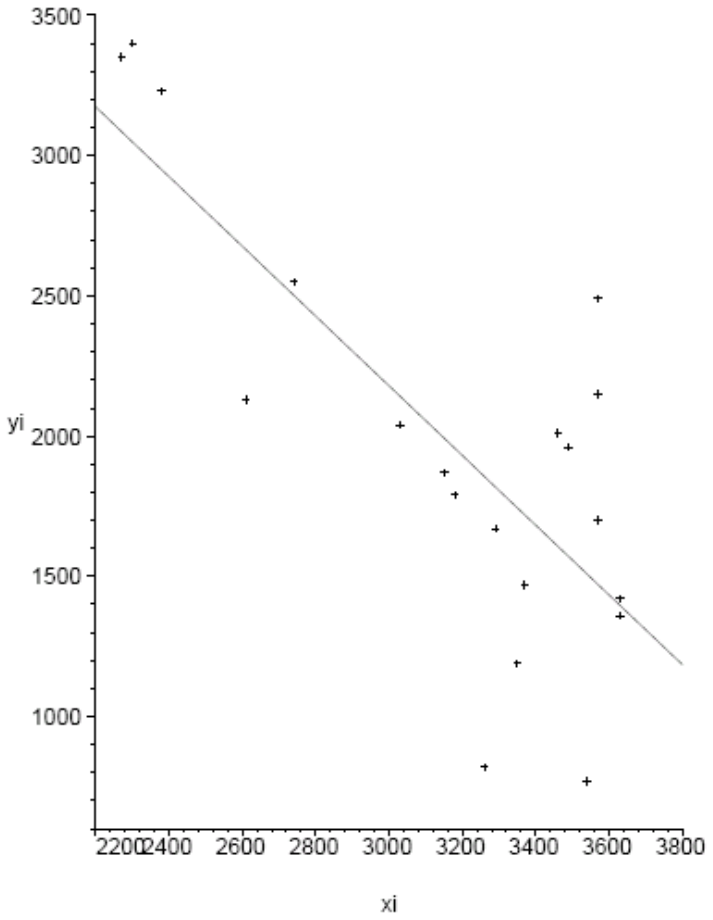
Der lineare Zusammenhang ist also nicht besonders ausgeprägt.



Beispiel: Lineare Regression

- Für die Merkmale Geburtsgewicht und Gewichtszuwachs erhält man

  α = 5905 und β = -1.242



Beispiel: Regressionsanalyse

- 10 LKW-Lieferungen
- Besteht ein Zusammenhang zwischen Weglänge und Lieferzeit?

Nummer der Lieferung	1	2	3	4	5	6	7	8	9	10
Weglänge (in km)	825	215	1070	550	480	920	1350	325	670	1215
Lieferzeit (in Tagen)	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

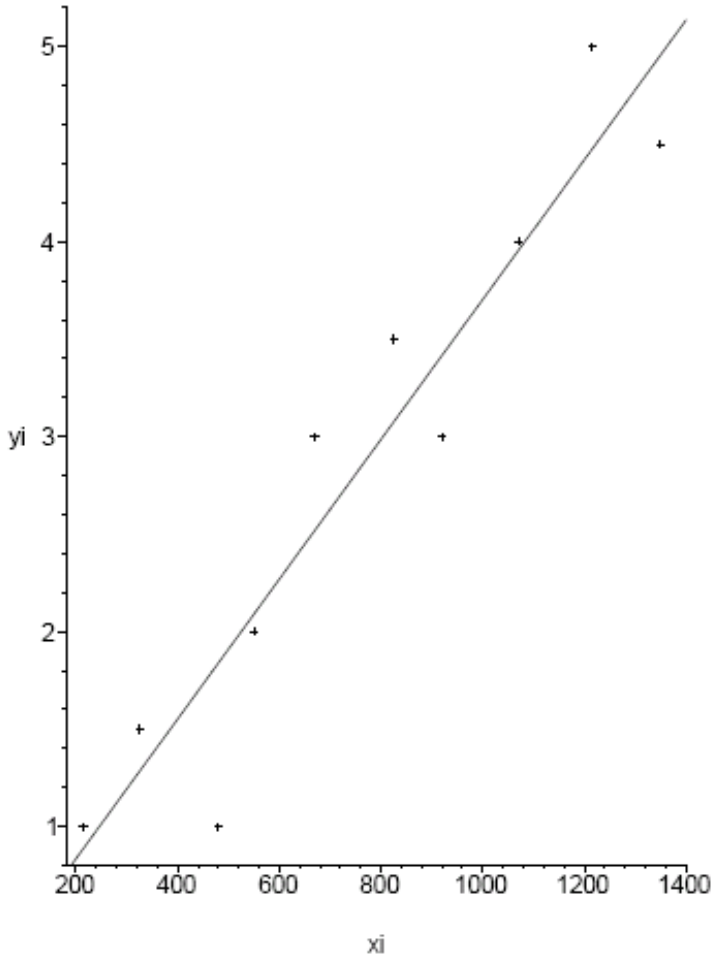
L.J. Kazmier (1999) *Wirtschaftsstatistik.* McGraw-Hill, S. 256 ff

Beispiel: Regressionsanalyse

- Streudiagramm mit Regressionsgerade:

$$\widehat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} = 0.0036\,, \qquad \widehat{\alpha} = \overline{y}_{10} - \widehat{\beta}\overline{x}_{10} = 0.12$$

- $R^2 = 0.90$
- $S^2 = 0.23 = 0.48^2$



Beispiel: Regressionsanalyse

- Teste H_0: $\beta = 0$ gegen H_1: $\beta \neq 0$ zum Niveau $1-\gamma = 0.05$

$$\overline{x}_{10} = 762\,, \quad \sum_{i=1}^{10} x_i^2 = 7104300\,, \quad \sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2} = 1139.24$$

$$\frac{|\,\widehat{\beta}\,|}{S/\sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2}} = \frac{0.0036}{0.48/1139.24} = \frac{0.0036}{0.0004} = 9.00$$

$$\mathrm{t}_{8,0.975} = 2.306$$

- Lehne H_0 ab. Es besteht also ein signifikanter Zusammenhang.

Beispiel: Regressionsanalyse

- Konfidenzintervall für β zum Niveau γ = 0.95
- Linke Grenze $\widehat{\beta} - \mathrm{t}_{n-2,1-(1-\gamma)/2} S \Big/ \sqrt{(n-1)s_{xx}^2}$
- Rechte Grenze $\widehat{\beta} + \mathrm{t}_{n-2,1-(1-\gamma)/2} S \Big/ \sqrt{(n-1)s_{xx}^2}$
- Hier:

$$\frac{S}{\sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2}} = \frac{0.48}{1139.24} = 0.0004$$

$$(0.0036 \mp 2.306 \cdot 0.0004) = (0.0036 \mp 0.0009) = (0.0027,\, 0.0045)$$

Beispiel: Regressionsanalyse

- Sei $x_0 = 1000$. Dann ist

$$\widehat{\alpha} + \widehat{\beta} x_0 = 3.70$$

$$S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}} = 0.182$$

- Konfidenzintervall für $\alpha+\beta\ x_0$ zum Niveau $\gamma = 0.95$

  (3.70± 2.306*0.182) = (3.28, 4.12)

Beispiel: Regressionsanalyse

- Sei $x_0 = 0$ (also außerhalb des zur Modellanpassung verwendeten Wertebereichs).
- Dann ist die geschätzte Lieferzeit gegeben durch $y_0 = 0.12$.
- Als reine Fahrzeit nicht sinnvoll.
- Mögliche Interpretation: Verladezeit.

Beispiel: Prognoseintervall

- $x_0 = 1000$, $\gamma = 0.95$

  $S = 0.48$, $\tau' = 2.306*0.513$

  Prognoseintervall für $Y_0 = \alpha+\beta+\varepsilon_0$

  $(3.70 \pm 2.306*0.513) = (2.52, 4.89)$

Beispiel: Konfidenzband

- Konfidenzband für die Regresionsgerade der Lieferzeiten zum Niveau $\gamma = 0.95$