Technische Hochschule Wildau [FH]
Technical University of Applied Sciences
WILDAU

Maschinelles Lernen – Theorie
Wahrscheinlichkeit & Statistik

Prof. Dr. Rainer Stollhoff

- Wahrscheinlichkeitsrechnung
    - Zufallsvariable
    - Wahrscheinlichkeitsrechnung
    - Wahrscheinlichkeitsverteilung
- Statistik
    - Zusammenhang zweier Zufallsvariablen
    - Korrelation und Kausalität
  - Lineare Regression
    - Trendlinie
    - Verlustminimierung
    - Erklärte Varianz

Zufallsvariable

- Für noch unbeobachtete Ereignisse z.B. eines Merkmals kennen wir keinen Wert.
- Mathematisch können wir noch unbeobachtete Ereignisse als Variable beschreiben z.B. X oder A.
- Da der Wert der Variable nicht bekannt ist, sondern zufällig bei der Beobachtung festgelegt wird, sprechen wir von einer Zufallsvariable.
- Entscheidend für die mathematische Analyse, sind die möglichen Werte oder Ausprägungen der Zufallsvariable und die Wahrscheinlichkeiten, mit denen die einzelnen Ausprägungen realisiert werden.

Beispiel 1

Tab. 11.4 Präferenzen bezüglich Softdrinks

Getränk	Material		
	Glas	Plastik	Summe
Coca-Cola	501	110	611
Pepsi	203	186	389
Summe	704	296	1000

- Annahme: $P = \frac{Anzahl\ der\ günstigen\ Fälle}{Anzahl\ der\ möglichen\ Fälle}$
- P(Coca-Cola) =
- P(Glas) =
- P(Coca-Cola und Glas) =

Wahrscheinlichkeitstheorie

1. **Wahrscheinlichkeiten als relative Häufigkeiten (frequentistischer Ansatz)**

$$P = \frac{Anzahl\ der\ günstigen\ Fälle}{Anzahl\ der\ möglichen\ Fälle}$$

2. **Wahrscheinlichkeiten als mathematische Objekte (axiomatischer Ansatz)**
   - a) **Wertebereich zwischen 0 und 1**

     $$0 \leq P \leq 1$$

     P=1: Sicheres Ereignis und P=0: Unmögliches Ereignis
   - b) **Additionssatz**

     $P(A\ oder\ B) = P(A) + P(B)$ wenn $(A\ und\ B)\ unmöglich$
   - c) **Unabhängigkeit**

     Gilt $P(A\ und\ B) = P(B) \cdot P(A)$ dann heißen A und B **unabhängig**.

Beispiel 1

Tab. 11.4 Präferenzen bezüglich Softdrinks

Getränk	Material		
	Glas	Plastik	Summe
Coca-Cola	501	110	611
Pepsi	203	186	389
Summe	704	296	1000

- P(Coca-Cola | Glas) =
- P(Glas | Coca-Cola) =

Wahrscheinlichkeitstheorie

1. **Wahrscheinlichkeiten als relative Häufigkeiten (frequentistischer Ansatz)**

$$P = \frac{Anzahl\ der\ günstigen\ Fälle}{Anzahl\ der\ möglichen\ Fälle}$$

2. **Wahrscheinlichkeiten als mathematische Objekte (axiomatischer Ansatz)**

   a) **Wertebereich zwischen 0 und 1**

   $$0 \leq P \leq 1$$

   P=1: Sicheres Ereignis und P=0: Unmögliches Ereignis

   b) **Additionssatz**

   $P(A\ oder\ B) = P(A) + P(B)$ wenn $(A\ und\ B)\ unmöglich$

   c) **Unabhängigkeit**

   Gilt $P(A\ und\ B) = P(B) \cdot P(A)$ dann heißen A und B **unabhängig**.

   d) **bedingte Wahrscheinlichkeit**

   $$P(A|B) := \frac{P(A\ und\ B)}{P(B)}$$

   e) **Multiplikationssatz**

   $$P(A\ und\ B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

   f) **Satz von Bayes**

   $$P(A|B) = P(B|A) \cdot P(A)/P(B)$$

Beispiel 1

Tab. 11.4 Präferenzen bezüglich Softdrinks

	Material		
Getränk	Glas	Plastik	Summe
Coca-Cola	501	110	611
Pepsi	203	186	389
Summe	704	296	1000

- P(Coca-Cola | Glas) =
- P(Glas | Coca-Cola) =

Intuitive Wahrscheinlichkeiten

- Matthias ist 27 Jahre alt. Er ist Single und hatte bisher nur sporadisch feste Beziehungen. Er ist im Allgemeinen eher zurückhaltend und sorgfältig. Er interessiert sich nicht für Politik oder soziale Fragen und verwendet den größten Teil seiner Freizeit auf eines seiner vielen Hobbys, wie z. B. Computer, Modellbau und mathematische Denksportaufgaben.
- Matthias ist Teilnehmer einer Marktforschungsstudie unter Ingenieuren und Juristen.

- Ist Matthias eher Ingenieur oder Jurist?

Intuitive Wahrscheinlichkeiten

- Matthias ist 27 Jahre alt. Er ist Single und hatte bisher nur sporadisch feste Beziehungen. Er ist im Allgemeinen eher zurückhaltend und sorgfältig. Er interessiert sich nicht für Politik oder soziale Fragen und verwendet den größten Teil seiner Freizeit auf eines seiner vielen Hobbys, wie z. B. Computer, Modellbau und mathematische Denksportaufgaben.
- Matthias ist Teilnehmer einer von **1000** Teilnehmern einer Marktforschungsstudie unter **500** Ingenieuren und **500** Juristen.

- Ist Matthias eher Ingenieur oder Jurist?

Intuitive Wahrscheinlichkeiten

- Matthias ist 27 Jahre alt. Er ist Single und hatte bisher nur sporadisch feste Beziehungen. Er ist im Allgemeinen eher zurückhaltend und sorgfältig. Er interessiert sich nicht für Politik oder soziale Fragen und verwendet den größten Teil seiner Freizeit auf eines seiner vielen Hobbys, wie z. B. Computer, Modellbau und mathematische Denksportaufgaben.
- Matthias ist Teilnehmer einer von **1000** Teilnehmern einer Marktforschungsstudie unter **100** Ingenieuren und **900** Juristen.

- Ist Matthias eher Ingenieur oder Jurist?

Intuitive Wahrscheinlichkeiten

- Matthias ist 27 Jahre alt. Er ist Single und hatte bisher nur sporadisch feste Beziehungen. Er ist im Allgemeinen eher zurückhaltend und sorgfältig. Er interessiert sich nicht für Politik oder soziale Fragen und verwendet den größten Teil seiner Freizeit auf eines seiner vielen Hobbys, wie z. B. Computer, Modellbau und mathematische Denksportaufgaben.
- Matthias ist Teilnehmer einer von **1000** Teilnehmern einer Marktforschungsstudie unter **5** Ingenieuren und **995** Juristen.

- Ist Matthias eher Ingenieur oder Jurist?

Intuitive Wahrscheinlichkeiten

- Matthias ist 27 Jahre alt. Er ist Single und hatte bisher nur sporadisch feste Beziehungen. Er ist im Allgemeinen eher zurückhaltend und sorgfältig. Er interessiert sich nicht für Politik oder soziale Fragen und verwendet den größten Teil seiner Freizeit auf eines seiner vielen Hobbys, wie z. B. Computer, Modellbau und mathematische Denksportaufgaben.
- Matthias ist
  - einer von **1000** Teilnehmern einer Marktforschungsstudie unter **500** Ingenieuren und **500** Juristen.
  - einer von **1000** Teilnehmern einer Marktforschungsstudie unter **100** Ingenieuren und **900** Juristen.
  - einer von **1000** Teilnehmern einer Marktforschungsstudie unter **5** Ingenieuren und **995** Juristen.
- Ist Matthias eher Ingenieur oder Jurist?
- Diskutieren Sie in Kleingruppen, wie sich die Aufgabenstellung mit Hilfe der Bayes Formel mathematisch darstellen lässt.

Wahrscheinlichkeitstheorie

1. **Wahrscheinlichkeiten als relative Häufigkeiten (frequentistischer Ansatz)**
2. **Wahrscheinlichkeiten als mathematische Objekte (axiomatischer Ansatz)**
3. **Wahrscheinlichkeiten als Grad persönlicher Überzeugung (bayesianischer Ansatz)**

$$P(\text{Modell}|\text{Daten}) \propto P(\text{Daten}|\text{Modell}) \cdot P(\text{Modell})$$

Posterior **Likelihood** **Prior bzw. A-priori-Wahrscheinlichkeit**

Wahrscheinlichkeitsfunktion

- Eine Wahrscheinlichkeitsfunktion ordnet Ereignissen x_i einer diskreten Zufallsvariable eine Wahrscheinlichkeit $f(x_i)$ zu
  - Beispiel 1: Wahrscheinlichkeiten als relative Häufigkeiten

$$P = \frac{Anzahl\ der\ günstigen\ Fälle}{Anzahl\ der\ möglichen\ Fälle}$$

  - Beispiel 2: Zufallsexperimente - Wir werfen eine faire Münze
    - In 50% der Fälle erhalten wir Kopf: P(Kopf)=0,5
    - In 50% der Fälle erhalten wir Zahl: P(Zahl)=0,5
- Eine Verteilungsfunktion gibt für einen Wert X die Wahrscheinlichkeit an, einen Wert kleiner oder gleich X zu beobachten. $\mathrm{F}(X) = \sum_{z \leq X} P(z)$

Münzwurf

- Wir werfen eine faire Münze
  - mit p=50% erhalten wir Kopf
  - mit (1-p)=50% erhalten wir Zahl
- Die Münzwürfe sind voneinander unabhängig
  P(a,b) = P(a)*P(b)
- Wie hoch ist bei n Wiederholungen die Wahrscheinlichkeit $f(x_i)$ für
  - X_0 = keinmal Kopf
  - X_1 = einmal Kopf
  - X_2 = zweimal Kopf

Binomialverteilung

- Wir beobachten wiederholt ein Zufallsereignis mit zwei möglichen Ergebnissen (0,1)
  - mit der Wahrscheinlichkeit p erhalten wir 1
  - mit der Wahrscheinlichkeit (1-p) erhalten wir 0
- Die Wiederholungen sind voneinander unabhängig $P(a,b) = P(a)*P(b)$
- Wie hoch ist bei n Wiederholungen die Wahrscheinlichkeit $f(x_i)$ für
  - X_0 = keinmal 1
  - X_1 = einmal 1
  - X_2 = zweimal 1

Stetige Zufallsvariablen

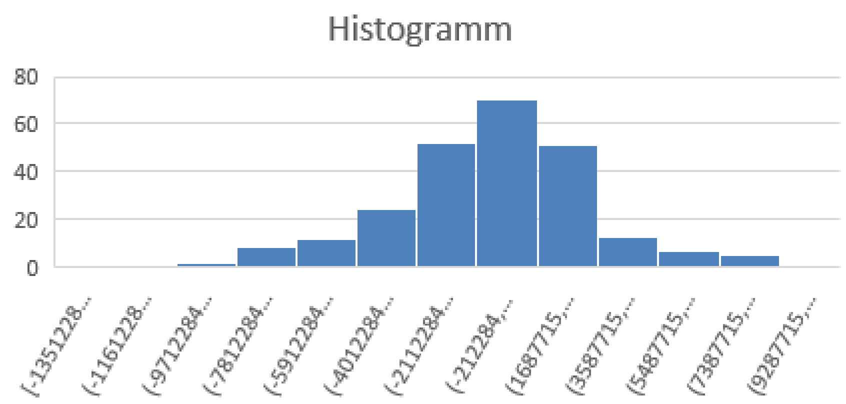
- Stetige Zufallsvariablen sind nicht auf eine diskrete Auswahl begrenzt, sie können beliebige Zahlenwerte in einem zusammenhängenden Intervall annehmen
  - *Bsp: Wie lange warte ich auf die nächste S-Bahn?*
- Was ist ein Ereignis x_i?
  - *Bsp: x_i = Wartezeit auf S-Bahn*
- Was ist die Wahrscheinlichkeitsfunktion $f(x_i)$?
  - *Bsp: $x_0 = 0$*

    $x_1 = ?$

    (x_i in Minuten, Sekunden, ms?)

Diskretisierung stetiger Zufallsvariablen

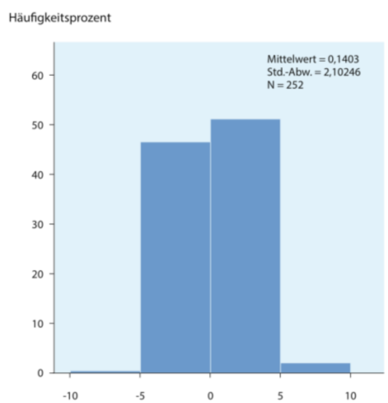
- Bei einer Messung (spätestens bei der Digitalisierung) diskretisiert man stetige Variablen, d.h.
  - Man teilt den Wertebereich in eine endliche Anzahl von Intervallen auf
  - Anstelle der Messwerts speichert man das Intervall, in dem der Messwert liegt
- Ein Histogramm stellt die beobachteten Häufigkeiten als Säulendiagram dar

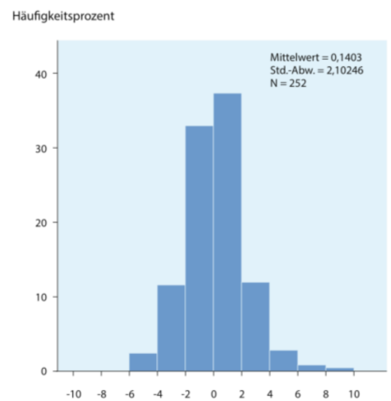


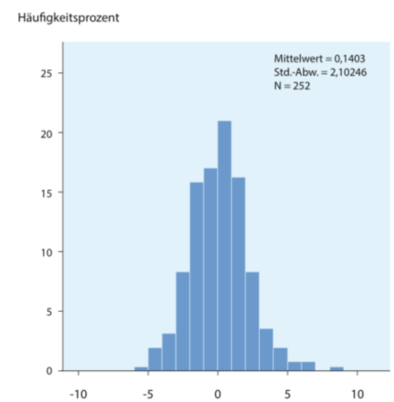
- Aber: Jede Diskretisierung ist mit Verlust behaftet (vgl. Schallplatte vs. CD)

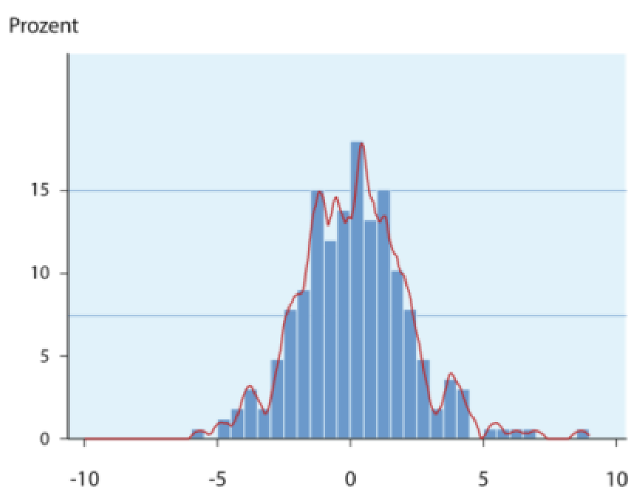
Histogramm und Dichtefunktion

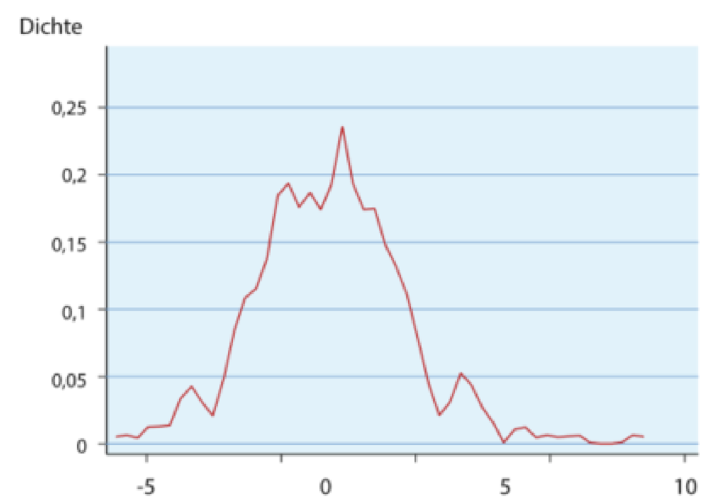
- Wenn man für die Diskretisierung immer kleinere Intervalle wählt, wird aus dem Histogramm eine Dichtefunktion







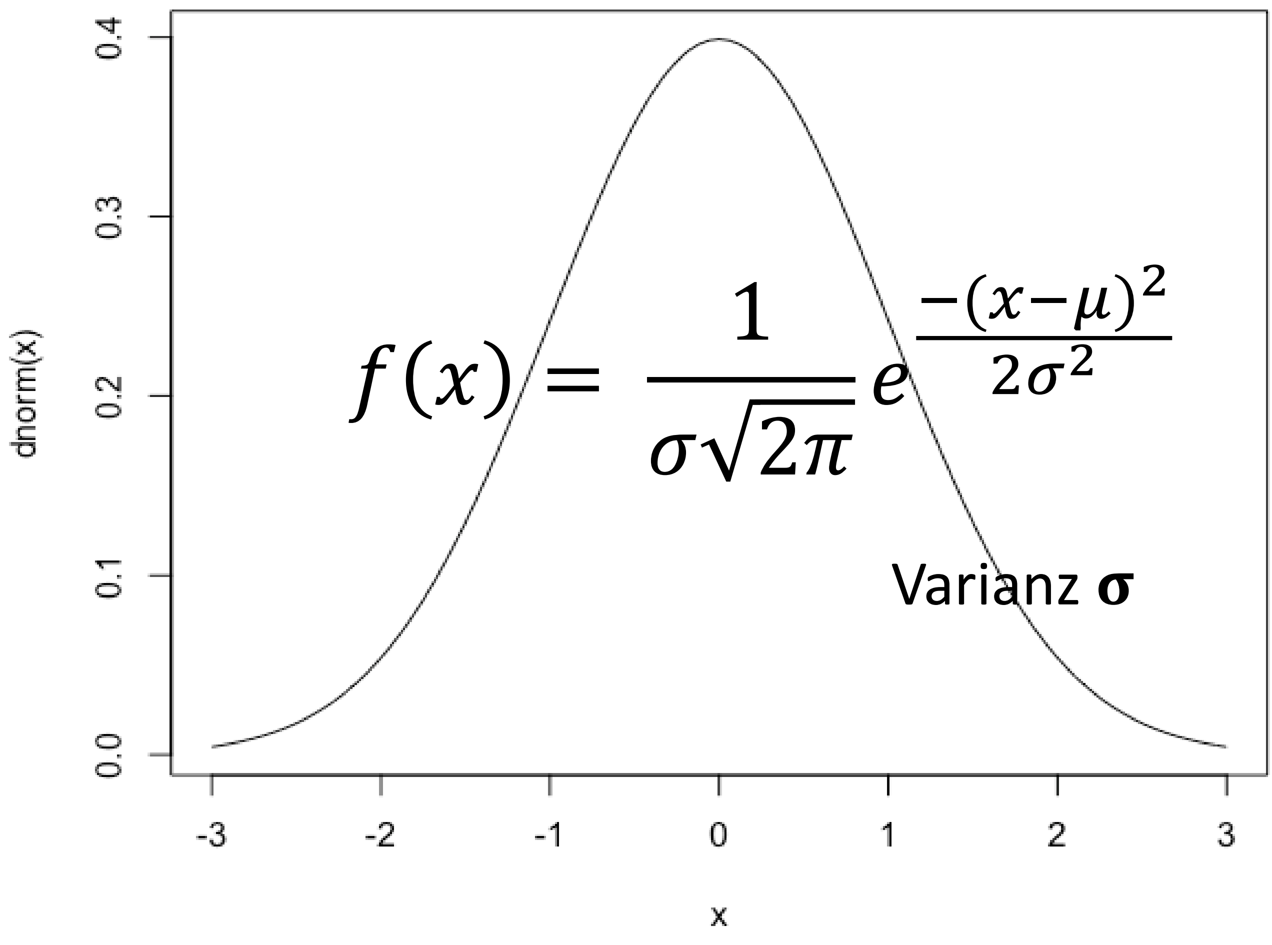




Normalverteilung

Gauß‘sche Glockenkurve

Erwartungswert µ



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Varianz **σ**

Normalverteilung

Gauß'sche Glockenkurve

- Wahrscheinlichkeitsrechnung
    - Zufallsvariable
    - Wahrscheinlichkeitsrechnung
    - Wahrscheinlichkeitsverteilung
- Statistik
    - Zusammenhang zweier Zufallsvariablen
    - Korrelation und Kausalität
  - Lineare Regression
    - Trendlinie
    - Verlustminimierung
    - Erklärte Varianz

Wiederholung

- Was sind die Unterschiede zwischen Supervised und Unsupervised Learning?

Supervised Learning

- Erfahrung E
  - Input X
  - Outcome Y
- Aufgabe A
  - Vorhersage $\hat{Y} = A(X)$
- Qualität Q
  - Verlustfunktion $\mathrm{L}\left(\hat{Y}, \mathrm{Y}\right)$

Unsupervised Learning

- Erfahrung E
  - Input X
- Aufgabe A
  - Repräsentation $\hat{X} = A(X)$
- Qualität Q
  - von Repräsentationen $\mathrm{Q}(\widehat{X_1}, \widehat{X_2})$

Statistisches Lernen

- Wie lassen sich die Fragestellungen statistisch formulieren?

Supervised Learning

- Erfahrung E
  - Input X
  - Outcome Y
- Aufgabe A
  - Vorhersage $\hat{Y} = A(X)$
- Qualität Q
  - Verlustfunktion $\mathrm{L}(\hat{Y}, \mathrm{Y})$

Unsupervised Learning

- Erfahrung E
  - Input X
- Aufgabe A
  - Repräsentation $\hat{X} = A(X)$
- Qualität Q
  - von Repräsentationen $\mathrm{Q}(\widehat{X_1}, \widehat{X_2})$

- Wahre Wahrscheinlichkeitsverteilungen
  P(X) und P(Y): Randverteilungen
  P(X,Y): Gemeinsame Verteilung
  P(Y|X): Bedingte Verteilung
- Subjektive Schätzungen der Wahrschienlichkeitsverteilungen
  $\hat{P}(X)$ als Repräsentation von P(X)
  $\hat{P}(Y|X)$ zur Vorhersage von Y als wahrscheinlichstem Outcome
- Lernen als Anpassen der Modelle an beobachtete Daten
  P(Modell|Daten) $\propto$ P(Daten|Modell) · P(Modell)

Zusammenhangsmaße

- Ein Zusammenhangsmaß gibt das Ausmaß des Zusammenhangs als Zahl an
- Ein einfaches Zusammenhangsmaß ist die Kovarianz:
  $Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  – Sind x_i und y_i größer als der jeweilige Mittelwert $\bar{x}$ bzw. $\bar{y}$ dann steigt die Kovarianz
  – Ist nur ein Wert größer und der andere kleiner, dann sinkt die Kovarianz

- Positiver Zusammenhang

X	-1	0	1
Y	-1	0	1

- Negativer Zusammenhang

X	-1	0	1
Y	1	0	-1

- Kein Zusammenhang

X	-1	0	1
Y	1	0	1

Standardisiertes Zusammenhangsmaß

- Die Kovarianz wächst mit der Varianz der Variablen. Das erschwert die Vergleichbarkeit.
- Meist ist daher eine standardisierte Variante, der sog. Bravais-Pearson-Koorelationskoeffizient passender:

$$\rho_{x,y} = \frac{Cov(x,y)}{\sqrt{s^2(x) \cdot s^2(y)}} = \frac{Cov(x,y)}{s(x) \cdot s(y)}$$

- Die Werte des Korrelationskoeffizienten liegen dann bei

  +1 für einen perfekt positiven Zusammenhang, z.B. $\rho_{x,x}$

  -1 für einen perfekt negativen Zusammenhang, z.B. $\rho_{x,-x}$

  0 für keinen Zusammenhang, z.B. x und y unabhängig

Zusammenhang und Unabhängigkeit

- Sind X und Y unabhängig, dann gilt

  $Cov(x, y) = 0$ und damit auch $\rho_{x,y} = 0$

- Gilt das umgekehrt auch, also wenn $\rho_{x,y} = 0$ dann sind X und Y unabhängig?

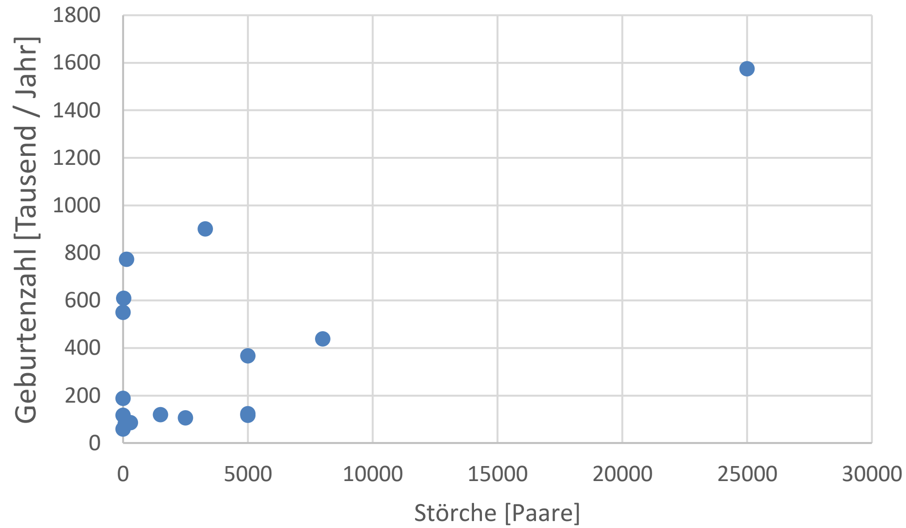
  Nein! Zum Beispiel sind X und Y=X2 unkorreliert, aber nicht unabhängig

- Die logisch korrekte Umkehrung gilt aber, d.h. wenn $\rho_{x,y} \neq 0$ dann sind X und Y auch nicht unabhängig.
- Gilt dann auch, dass bei einer Korrelation, d.h. einem statistischen Zusammenhang auch immer ein kausaler Zusammenhang, d.h. eine Ursache-Wirkung Beziehung gilt?

Gibt es einen Zusammenhang zwischen der Anzahl der Störche und der Anzahl der Geburten?

R. Matthews, Storks Deliver Babies (p=0.008)

Land	Störche [Paare]	Geburtenzahl [Tausend / Jahr]
Albania	100	83
Austria	300	87
Belgium	1	118
Bulgaria	5000	117
Denmark	9	59
France	140	774
Germany	3300	901
Greece	2500	106
Holland	4	188
Hungary	5000	124
Italy	5	551
Poland	30	610
Portugal	1500	120
Romania	5000	367
Spain	8000	439
Switzerland	150	82
Turkey	25000	1576
Korrelationen	Störche und Geburtenzahl	0,72



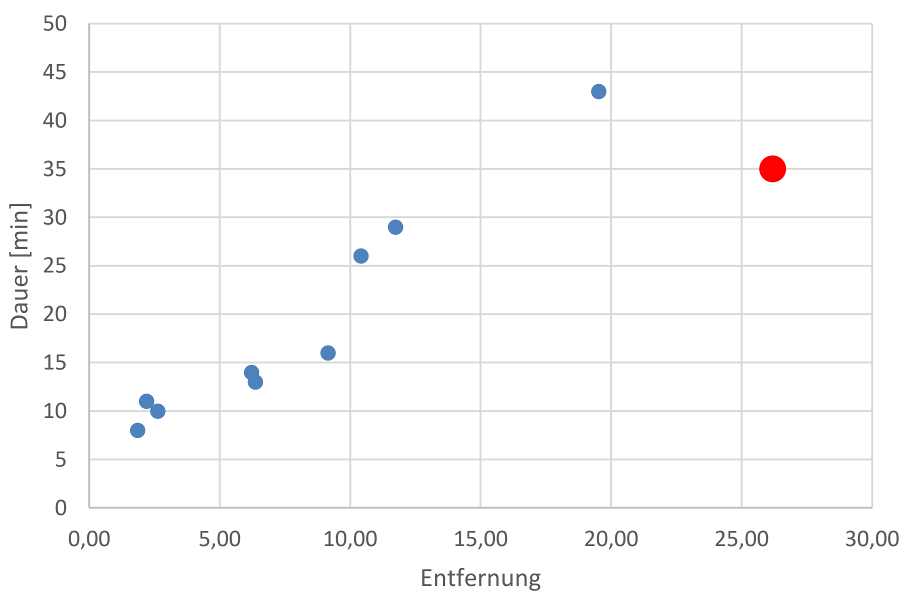
Gibt es einen Zusammenhang zwischen der Anzahl der Störche und der Anzahl der Geburten?

Ja, aber …

Statistische und kausale Zusammenhänge

1. Ein statistischer Zusammenhange hat eine kausale Ursache.
   Bsp: Je länger die zurückgelegte Entfernung, desto länger die dafür benötigte Zeit (bei gleicher Fortbewegungsart).

Entfernung = Geschwindigkeit * Dauer
Dauer = 1/Geschwindigkeit * Entfernung



Statistische und kausale Zusammenhänge

1. Ein statistischer Zusammenhange hat eine kausale Ursache.
   Bsp: Je länger die zurückgelegte Entfernung, desto länger die dafür benötigte Zeit (bei gleicher Fortbewegungsart).
2. Ein statistischer Zusammenhang hat keine kausale Ursache, sog. Scheinkorrelation

Statistische und kausale Zusammenhänge

1. Ein statistischer Zusammenhange hat eine kausale Ursache.
   Bsp: Je länger die zurückgelegte Entfernung, desto länger die dafür benötigte Zeit (bei gleicher Fortbewegungsart).
2. Ein statistischer Zusammenhang hat keine kausale Ursache, sog. Scheinkorrelation
   a) Konfundierende Variable, d.h. es gibt eine zugrundeliegende Ursache die einen statistischen Zusammenhang zwischen beiden Merkmalen herstellt.
   Bsp.: Zahl der Störche und Zahl der Geburten

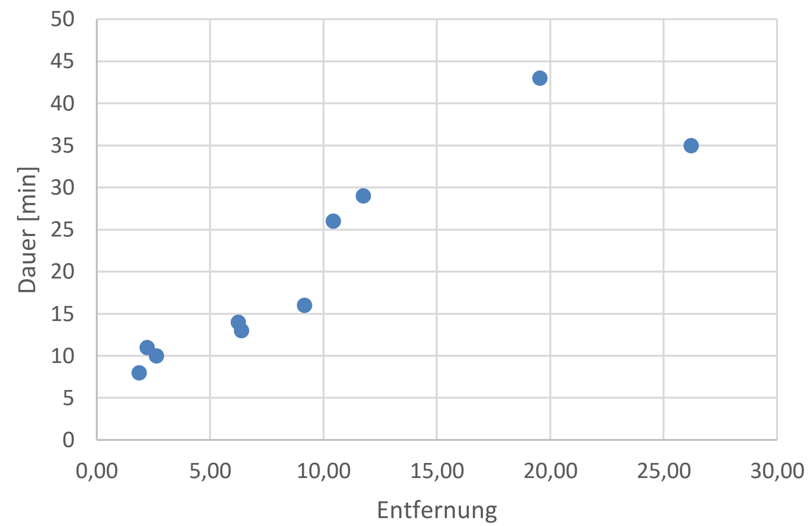
Statistische und kausale Zusammenhänge

1. Ein statistischer Zusammenhange hat eine kausale Ursache.
   Bsp: Je länger die zurückgelegte Entfernung, desto länger die dafür benötigte Zeit (bei gleicher Fortbewegungsart).
2. Ein statistischer Zusammenhang hat keine kausale Ursache, sog. Scheinkorrelation
   a) Konfundierende Variable, d.h. es gibt eine zugrundeliegende Ursache die einen statistischen Zusammenhang zwischen beiden Merkmalen herstellt.
      Bsp.: Zahl der Störche und Zahl der Geburten
   b) Explizite (bewusste) oder implizite (zufällige) Datenselektion, d.h. der Zusammenhang ist nur auf einer speziell ausgewählten Teilmenge gültig

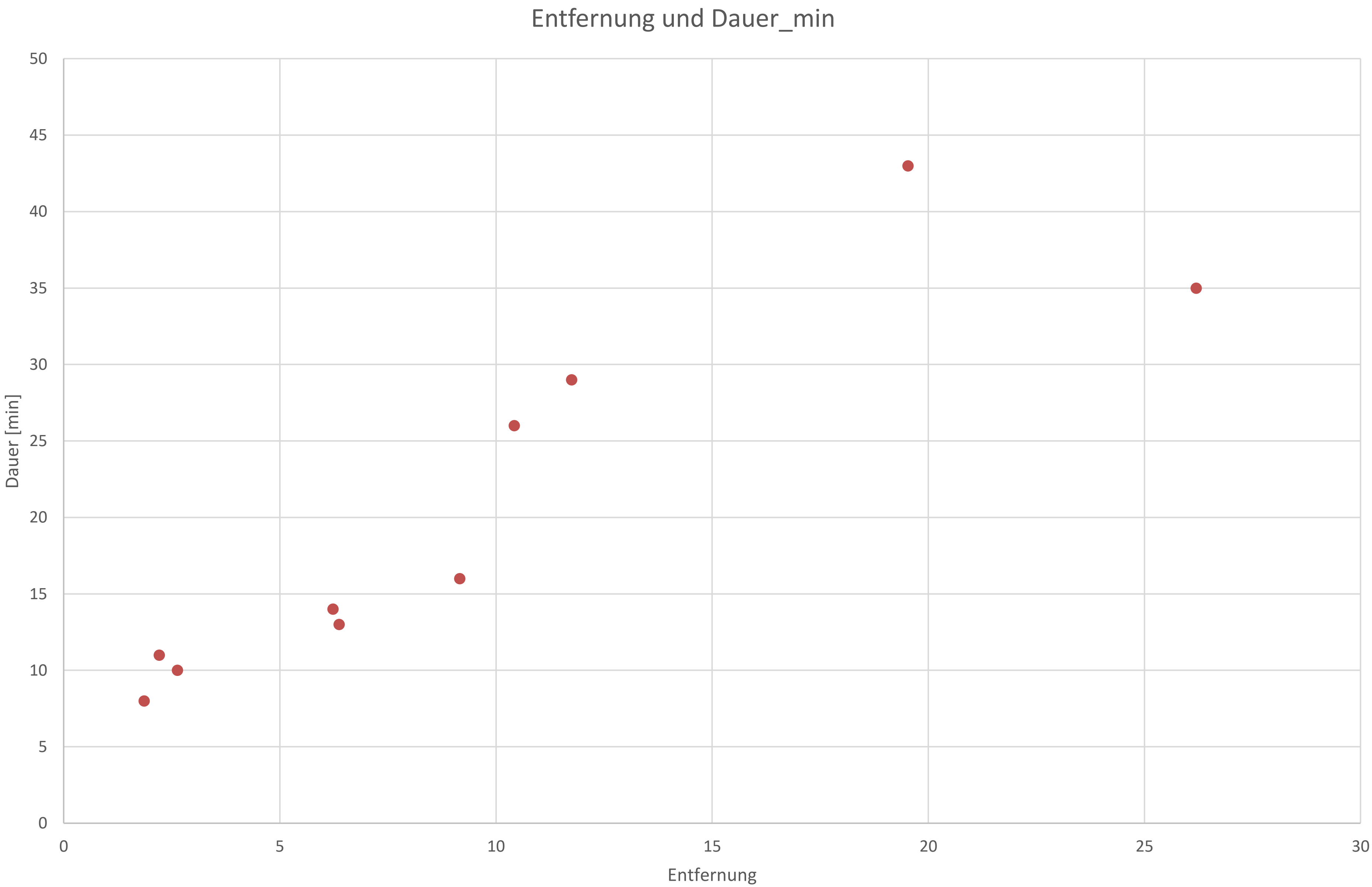
Datenpaare – Beispiel Car-Sharing

	Entfernung	Dauer_min
M =MITTELWERT	9,64	20,5
S =STABW.N	7,51	11,38
Cov =KOVARIANZ.P	77,52	
Cor =KORREL	0,91	
Rang-Cor	0,96	

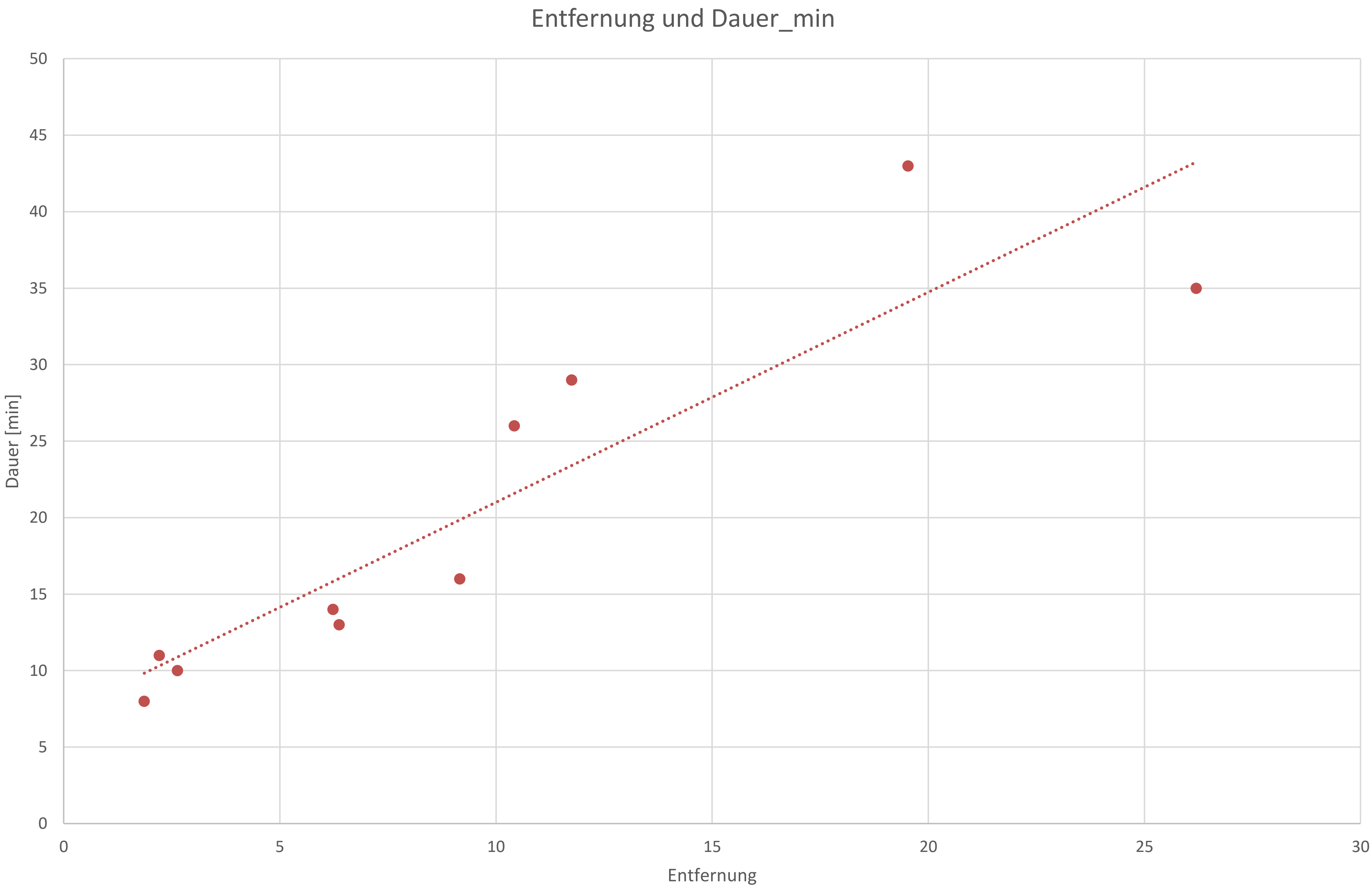
Entfernung	Dauer_min
1,86	8
2,21	11
19,53	43
26,20	35
2,63	10
6,23	14
6,37	13
9,16	16
10,42	26
15,00	**????**



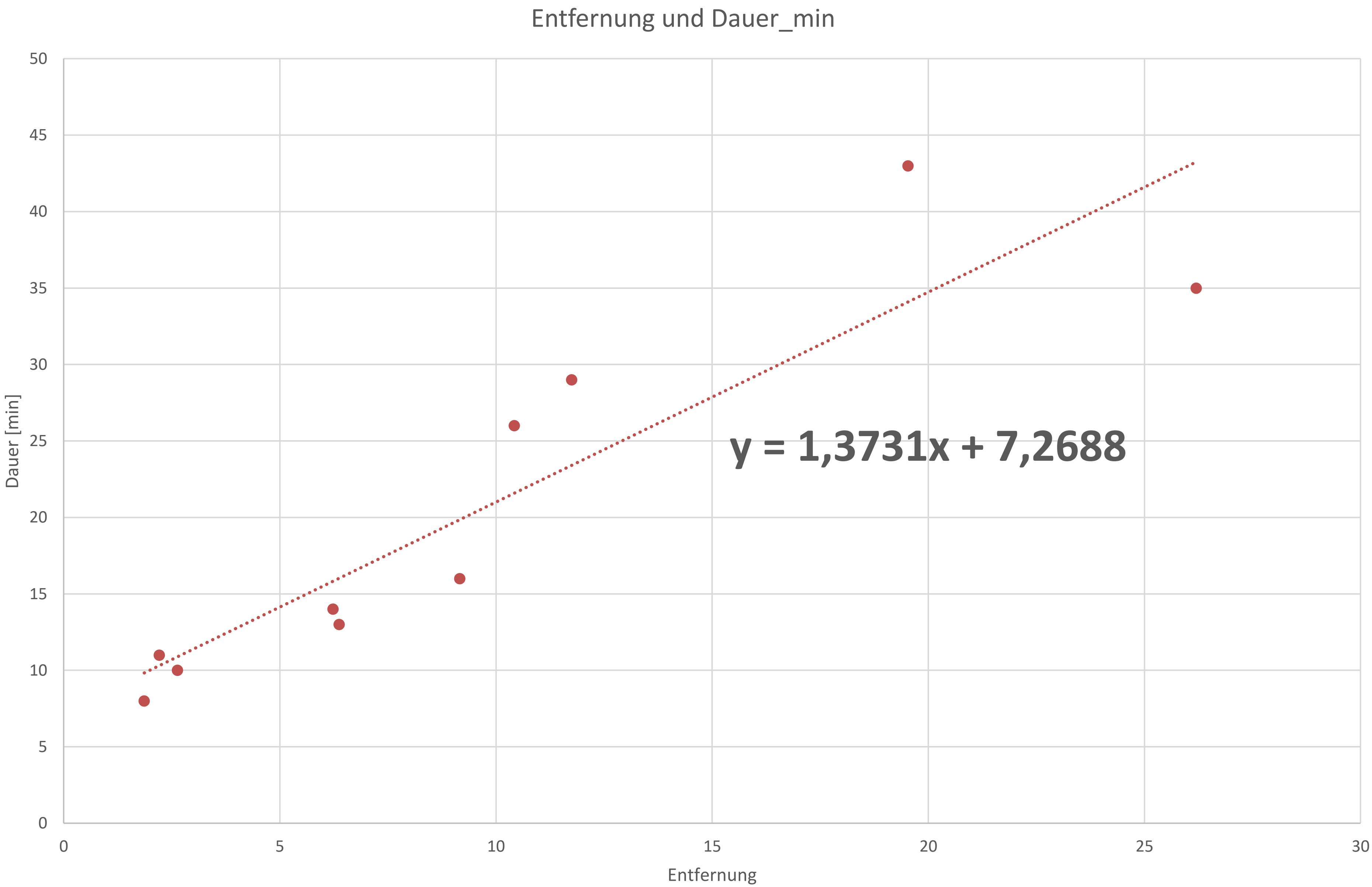
Streudiagramm (XY-Plot)



Streudiagramm mit Trendlinie



Streudiagramm mit Trendlinie



- AB hier nicht mehr gezeigt!!!

Lineare Regression

- Lineare Regression liefert Prognose
$$\hat{y} = \hat{y}(x) = a\,x + b$$
- Parameter
  - a: Steigung
  - b: Achsenabschnitt
- Wie bestimmt man a und b?
  - Quadratische Abweichung = Fehler der Prognose
    $\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{n}(y_i - (a\,x_i + b))^2$
    (mean squared error - mse)
  - Ziel: Minimierung des Prognosefehlers
    (kleinste-Quadrate-Regression, least-squares-regression)

Lineare Regression

$$L(a, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a\, x_i + b))^2$$

1. $\frac{\partial L(a,b)}{\partial b} = 0$

b = $\bar{y} - a \cdot \bar{x}$

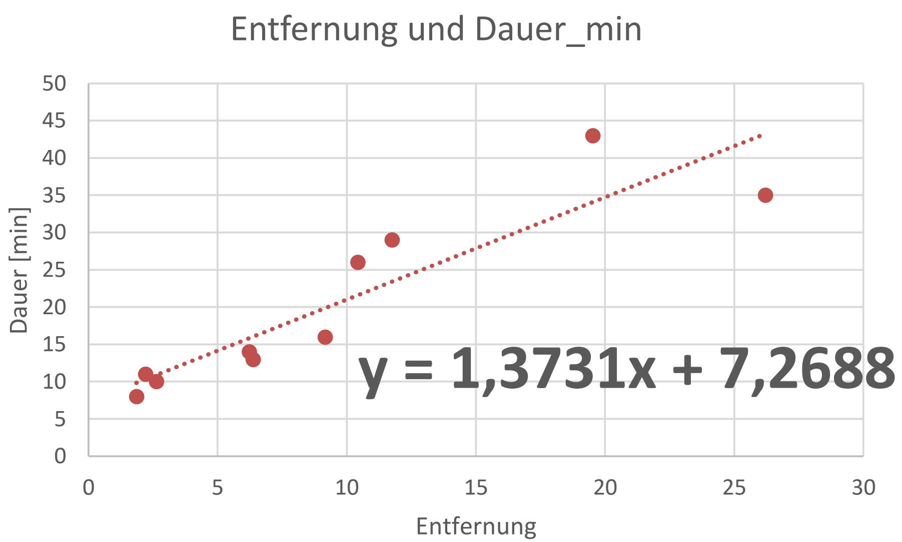
2. $\frac{\partial L(a,b)}{\partial a} = 0$

a = $\frac{Cov(x,y)}{Var(x)}$

Datenpaare – Beispiel Car-Sharing

	Entfernung	Dauer_min
M =MITTELWERT	9,64	20,5
S =STABW.N	7,51	11,38
Cov =KOVARIANZ.P	77,52	
Cor =KORREL	0,91	
Rang-Cor	0,96	

Entfernung	Dauer_min
1,86	8
2,21	11
19,53	43
26,20	35
2,63	10
6,23	14
6,37	13
9,16	16
10,42	26
11,75	29



Lineare Regression

- Lineare Regression liefert Prognose

$$\hat{y} = \hat{y}(x) = a\,x + b$$

- Wie gut ist die Prognose?
  - Mittlere Quadratische Abweichung = Fehler der Prognose

$$mse(y, \hat{y}\,) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2$$

  - Fehler der Prognose mit Mittelwert = Varianz

$$mse(y, \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = s_G^2(y)$$

  - Bestimmtheitsmaß R^2

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}\left(y_i - \hat{y}(x_i)\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{mse(y, \hat{y}\,)}{mse(y, \bar{y})}$$

Relativer Rückgang des Fehlers

$$= \frac{s_G^2(y) \; - s_G^2(y - \hat{y})}{s_G^2(y)} = 1 - \frac{s_G^2(y - \hat{y})}{s_G^2(y)}$$

Erklärte Varianz = Relativer Rückgang der Varianz

$$= \rho_{x,y}^2$$

Quadrierte Korrelation

Datenpaare – Beispiel Car-Sharing

	Entfernung	Dauer_min
M =MITTELWERT	9,64	20,5
S =STABW.N	7,51	11,38
Cov =KOVARIANZ.P	77,52	
Cor =KORREL	0,91	
Rang-Cor	0,96	

Entfernung	Dauer_min
1,86	8
2,21	11
19,53	43
26,20	35
2,63	10
6,23	14
6,37	13
9,16	16
10,42	26
11,75	29