Technische Hochschule Wildau [FH]
Technical University of Applied Sciences

Wahrscheinlichkeit & Statistik Zusammenfassung

Prof. Dr. Rainer Stollhoff

- Wahrscheinlichkeitsrechnung
  - Wahrscheinlichkeitsrechnung
  - Zufallsvariable
    - Wahrscheinlichkeitsfunktion, - verteilung und -dichte
    - Beispiele: Binomialverteilung, Normalverteilung
  - Zentraler Grenzwertsatz
- Statistik
  - Deskriptive Statistik einer Zufallsvariablen
    - Lage- und Streuungsmaße
  - Zusammenhang zweier Zufallsvariablen
    - Korrelation und Kausalität

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.

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

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. $\mathrm{F}(X) = \sum_{z \leq X} P(z)$

Beispiel: 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

- Zufallsereignis mit zwei möglichen Ergebnissen (0,1)
- Wahrscheinlichkeit für 1:

$$p = \mathrm{P}(1)$$

- Wahrscheinlichkeit für 0:

$$\mathrm{P}(0) = 1 - \mathrm{P}(1) = 1 - p$$

- Wiederholungen sind unabhängig, z.B.

$$\mathrm{P}(1,0) = \mathrm{P}(1) \cdot \mathrm{P}(0) = p \cdot (1 - p)$$

- Wahrscheinlichkeit bei n Wiederholungen k mal 1 zu erhalten:

$$B_{n,p}(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

  sogenannte ***Wahrscheinlichkeitsfunktion***

- Wahrscheinlichkeit bei n Wiederholungen höchstens k mal 1 zu erhalten:

$$B_{n,p}(x \leq k) = \sum_{j=0}^{k} \binom{n}{j} \cdot p^j \cdot (1-p)^{n-j}$$

  sogenannte ***Verteilungsfunktion***

Erwartungswert und Varianz einer diskreten Zufallsvariablen

Rechenregel

Erwartungswert einer diskreten Zufallsvariablen

$$E(X) = \sum_{i=1}^{n} x_i \cdot f(x_i)$$

- Der Erwartungswert einer Zufallsvariablen entspricht dem Mittelwert
- Bsp.:
  - Erwartungswert eines fairen Münzwurfs
  - Erwartungswert einer Binomialverteilung mit p=0,1
  - Erwartungswert eines Würfelwurfs

Rechenregel

Varianz einer diskreten Zufallsvariablen:

$$Var(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 \cdot f(x_i)$$

- Die Varianz einer Zufallsvariablen entspricht der mittleren quadratischen Abweichung

Stetige Zufallsvariablen

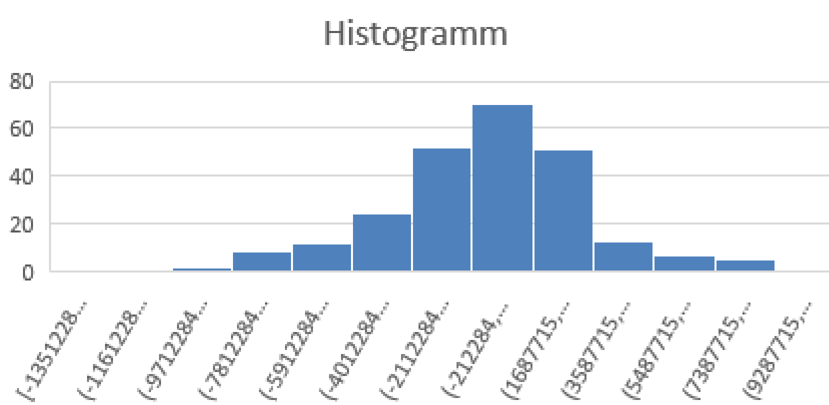
- Stetige Zufallsvariablen sind nicht auf eine diskrete Auswahl begrenzt, sie können beliebige Zahlenwerte in einem zusammenhängenden Intervall annehmen
  - *Bsp: Wie lange warte ich auf die nächste S-Bahn?*
- Was ist ein Ereignis x_i?
  - *Bsp:* x_i *= Wartezeit auf S-Bahn*
- Was ist die Wahrscheinlichkeitsfunktion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

    $x_1 = ?$

    (x_i *in Minuten, Sekunden, ms?)*

Diskretisierung stetiger Zufallsvariablen

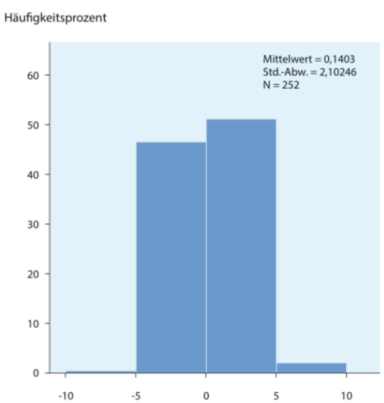
- Bei einer Messung (spätestens bei der Digitalisierung) diskretisiert man stetige Variablen, d.h.
  - Man teilt den Wertebereich in eine endliche Anzahl von Intervallen auf
  - Anstelle der Messwerts speichert man das Intervall, in dem der Messwert liegt
- Ein Histogramm stellt die beobachteten Häufigkeiten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

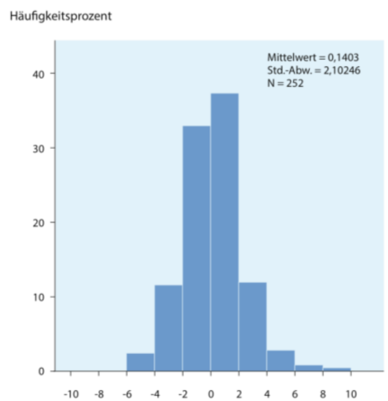


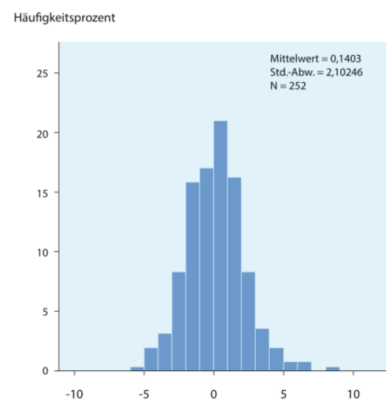
- Aber: Jede Diskretisierung ist mit Verlust behaftet (vgl. Schallplatte vs. CD)

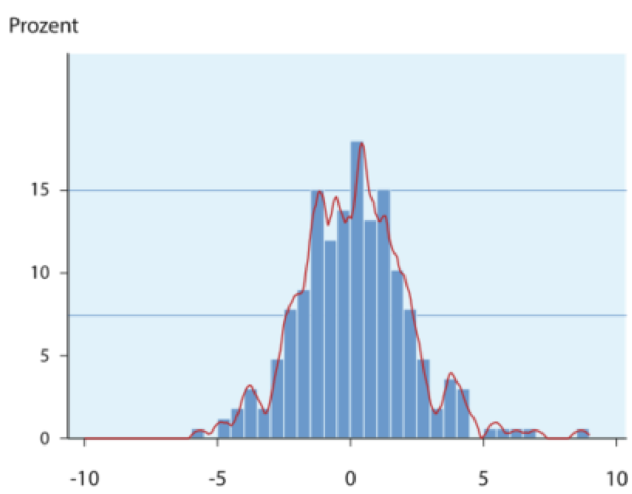
Histogramm und Dichtefunktion

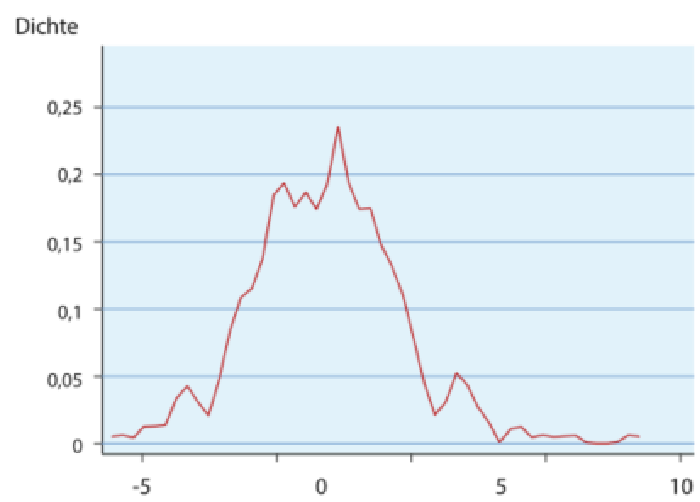
- Wenn man für die Diskretisierung immer kleinere Intervalle wählt, wird aus dem Histogramm eine Dichtefunktion











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: $x_0 = 0$*

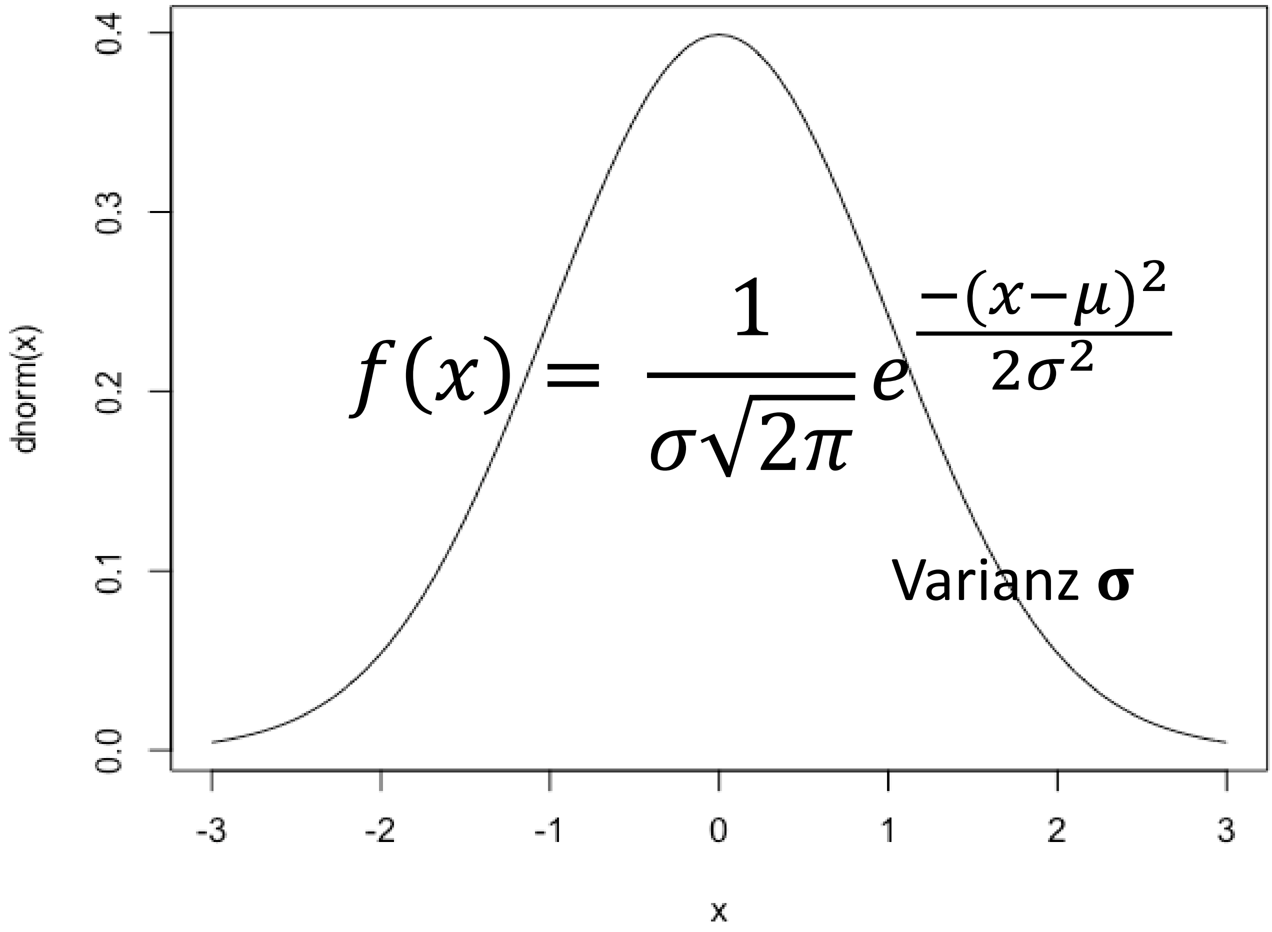
    $x_1 = ?$

    (x_i in Minuten, Sekunden, ms?)
- Es gibt nur unendlich kleine (infinitesimal) Ereignisse.
- Eine Dichtefunktion $f(x)$oder $p(x)$ ordnet infinitesimalen Ereignissen einer metrischen Variable eine Wahrscheinlichkeit zu
- Eine Verteilungsfunktion $\mathrm{F}(x)$oder $P(x)$ gibt für einen Wert X die Wahrscheinlichkeit an, einen Wert kleiner oder gleich X zu beobachten: $P(x) = \int_{-\infty}^{x} p(z)\, dz$

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

- Dichtefunktion

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

  - Erwartungswert μ
  - Varianz σ^2
- Heuristiken:
  - 68,27% der Werte liegen im Intervall [μ-σ, μ+σ]
  - 95,45% der Werte liegen im Intervall [μ-2σ, μ+2σ]
- Standardisierung
  - $Z = \frac{X-\mu}{\sigma}$

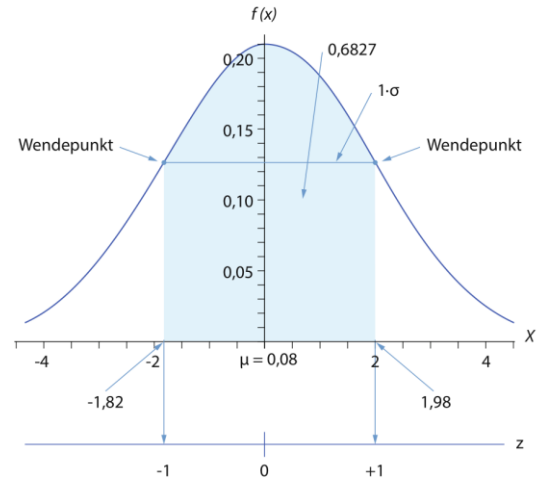


Abb. 11.47 Dichtefunktion der Tagesrendite einer Daimler-Aktie, $-1{,}82 < x < 1{,}98$

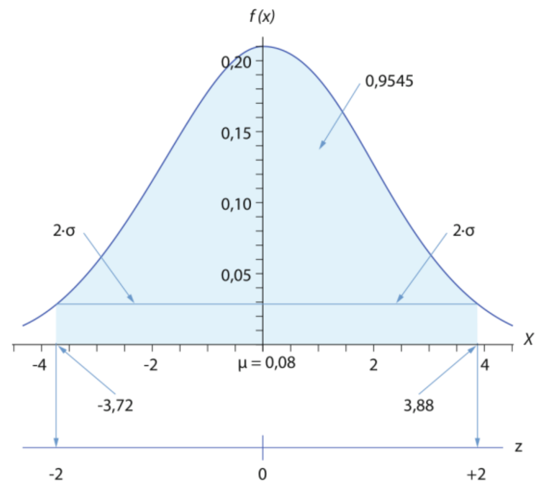


Abb. 11.48 Dichtefunktion der Tagesrendite einer Daimler-Aktie, $-3{,}72 < x < 3{,}88$

Erwartungswert und Varianz einer diskreten Zufallsvariablen

Rechenregel

Erwartungswert für eine stetige Variable: Der Erwartungswert $E(X)$ wird folgendermaßen ermittelt:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx.$$

Rechenregel

Für die **Varianz einer stetigen Zufallsvariablen** gilt:

$$Var(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x)dx.$$

- Der Erwartungswert einer Zufallsvariablen wird mit dem Mittelwert geschätzt

- Die Varianz einer Zufallsvariablen wird mit der mittleren quadratischen Abweichung geschätzt

Zentraler Grenzwertsatz

- Für eine beliebige Folge von unabhängigen, identisch verteilten metrischen Zufallsvariablen $X_1, \dots, X_n$ (z.B. Messdaten,...) mit
  - Erwartungswert μ
  - Standardabweichung σ

  berechne den empirischen Mittelwert

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

- Für diesen empirischen Mittelwert gilt
  - Erwartungs-/Mittelwert $\mathrm{E}(\bar{X}_n) = \mu$
  - Standardabweichung $\mathrm{Std}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$
- Der Mittelwert folgt für große n annähernd einer Normalverteilung $N(\mu, \sigma^2/n)$

Zentraler Grenzwertsatz

(Lindeberg, Levy)

- Wir führen ein Zufallsexperiment insgesamt n mal durch
- Die Wiederholungen $(X_1, \cdots, X_n)$ sind unabhängig und identisch verteilt
  - Erwartungswert μ
  - Varianz σ^2

  und bilden die Summe S_n der n Wiederholungen

$$S_n = X_1 + \cdots + X_n$$

- Für die Summe S_n gilt
  - Erwartungswert $\mathrm{E}[S_n] = n \cdot \mu$
  - Varianz $\mathrm{Var}[S_n] = n \cdot \sigma^2$
- Die Summe S_n ist für große n annähernd normalverteilt

$$S_n \sim N(n \cdot \mu, n \cdot \sigma^2)$$

- Wahrscheinlichkeitsrechnung
  - Wahrscheinlichkeitsrechnung
  - Zufallsvariable
    - Wahrscheinlichkeitsfunktion, - verteilung und -dichte
    - Beispiele: Binomialverteilung, Normalverteilung
  - Zentraler Grenzwertsatz
- Statistik
  - Deskriptive Statistik einer Zufallsvariablen
    - Lage- und Streuungsmaße
  - Zusammenhang zweier Zufallsvariablen
    - Korrelation und Kausalität

Merkmale und Skalen

Man erhebt Daten von Merkmalen anhand von Skalen

Diskrete Merkmale

- Nominales Skalenniveau - (lat. nomen = Namen)
  - Bsp.: {rot, gelb, grün}, {Mann, Frau}, {BWL, Mathe, Jura}
  - Kategorien für verschiedene Objekte
  - Übliche Fragen: X = Y?, Wieviele X haben auch Y?
- Ordinales Skalenniveau – (lat. ordo = Reihe)
  - Bsp.: (Grundschule, Gymnasium, Hochschule), (Gewährleistungszeit, Garantiezeit, außerhalb der Garantie)
  - Kategorien mit Rangordnung
  - Übliche Fragen: X > Y?, Wenn X1 > Y1 dann auch X2 > Y2?

Stetige Merkmale

- Metrisches Skalenniveau – (lat. metor - abmessen)
  - Bsp.: reelle Zahlen, Intervall [0,1]
  - Einzelne Messwerte mit Anordnung und Abstandsmaß
  - Übliche Fragen: (X-Y) > Z? Mittlere Wert von X?

nominal/ordinal: Häufigkeiten

- Für nominale (und z.T. ordinale) Skalen kann man die Häufigkeit mit der Kombinationen beobachtet werden in einer (mehrdimensionalen/Pivot-) Tabelle eintragen.

	Elektrik	Sauberkeit	Bedienung	Summe
innerhalb der Gewährleistung	40	40	20	100
innerhalb der Garantiezeit	4	176	320	500
außerhalb der Garantiezeit	216	144	40	400
Summe	260	360	380	1000

- Man spricht von
  - Absoluten Häufigkeiten z.B. 40 Gewährleistungen wegen Elektrik
  - Relativen Häufigkeiten z.B. 40/1000=4% Gewährleistungen wegen Elektrik
  - Kumulierten Häufigkeiten (beim schrittweisen Zusammenfassen einer ordinalen Skala) z.B. 44 Erstattungsfälle (Gewährleistungen und Garantien) bei Elektrik

Skalen

- Nominales Skalenniveau - (lat. nomen = Namen)
  - Bsp.: {rot, gelb, grün}, {Mann, Frau}, {BWL, Mathe, Jura}
  - Kategorien für verschiedene Objekte
  - Übliche Fragen: X = Y?, Wieviele X haben auch Y?
- Ordinales Skalenniveau – (lat. ordo = Reihe)
  - Bsp.: (Grundschule, Gymnasium, Hochschule), (Gewährleistungszeit, Garantiezeit, außerhalb der Garantie)
  - Kategorien mit Rangordnung
  - Übliche Fragen: X > Y?, Wenn X1 > Y1 dann auch X2 > Y2?
- Metrisches Skalenniveau – (lat. metor - abmessen)
  - Bsp.: reelle Zahlen, Intervall [0,1]
  - Einzelne Messwerte mit Anordnung und Abstandsmaß
  - Übliche Fragen: (X-Y) > Z? Mittlere Wert von X?

metrisch: Lageparameter

- Für metrische Variablen $(x_i)_{i=1}^{n}$ geben Lageparameter Auskunft über den „Mittelpunkt“ der Daten an

Fahrt	1	2	3	4	5	6
Dauer_min	52	11	13	17	14	14

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

  Der Durchschnitt der Datenpunkte

- Median (): $\tilde{x} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} \text{ für n ungerade} \\ \left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n+2}{2}\right]}\right)/2 \text{ für n gerade} \end{cases}$

  Der durchschnittliche Datenpunkt: 50% größer, 50% kleiner

- Wie lange dauerte eine Fahrt im Mittel?
  - (Arithmetischer) Mittelwert
  - Median
  - Modalwert?

metrisch: Streuungsparameter

- Für metrische Variablen $(x_i)_{i=1}^n$ geben Streuungsparameter Auskunft über die „Streuung“ der Daten
  - Minimum, Maximum, Spannweite
  - X% - Quantile: z.B. 50%-Quantil=Median

    X% der Datenpunkte sind kleiner oder gleich groß
  - Abweichung vom Mittelwert $x_i - \bar{x}$
    - Mittlere Abweichung: $\frac{1}{n}\sum_{i=1}^n |x_i - \bar{x}|$
    - Mittlere quadrierte Abweichung einer Grundgesamtheit: $\frac{1}{\boldsymbol{n}}\sum_{i=1}^n (x_i - \bar{x})^2$
    - (Stichproben-)Varianz: $s^2 = \frac{1}{\boldsymbol{n-1}}\sum_{i=1}^n (x_i - \bar{x})^2$
    - Standardabweichung: $s = \sqrt{s^2}$

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

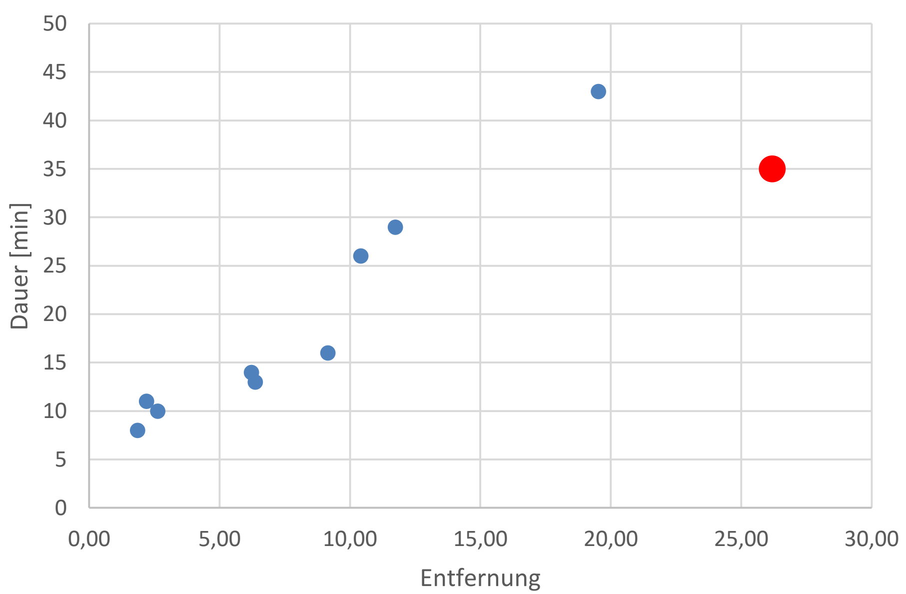
Zusammenhang und Unabhängigkeit

- 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?

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
   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