Statistik II für Betriebswirte
Vorlesung 6

Dr. Andreas Wünsche

TU Bergakademie Freiberg
Institut für Stochastik

18. November 2019

6. Korrelationsanalyse

6.1. Zwei normalverteilte Merkmale

- Mit Hilfe der Korrelationsanalyse sollen statistisch gesicherte Aussagen über bestimmte Aspekte des **Zusammenhangs** von zwei oder einer größeren Anzahl von Merkmalen getroffen werden.
- Im zugehörigen stochastischen Modell entsprechen den Merkmalen dann z.B. zwei Zufallsgrößen X und Y, die zu einem zweidimensionalen Zufallsvektor (X, Y) (auch zweidimensionale Zufallsvariable genannt) zusammengefasst werden können (analog für eine größere Anzahl von Merkmalen).
- Für derartige Zufallsvektoren interessieren Wahrscheinlichkeiten dafür, dass seine Realisierungen in bestimmten geeigneten Mengen liegen, diese Wahrscheinlichkeiten bilden die Verteilung (Wahrscheinlichkeitsverteilung) des Zufallsvektors.
- Im Folgenden soll zuerst kurz auf Zufallsvektoren eingegangen werden, bevor statistische Fragen behandelt werden.

Verteilungsfunktion eines Zufallsvektors

- Die Verteilung des Zufallsvektors (X, Y) kann durch die gemeinsame (oder Verbund-)Verteilungsfunktion bestimmt oder definiert werden. Für $x, y \in \mathbb{R}$ gilt

$$F_{(X,Y)}(x, y) = P(\{X < x\} \cap \{Y < y\}) = P(X < x, Y < y)\,.$$

- Diese Verbundverteilungsfunktionen haben ähnliche Eigenschaften wie die Verteilungsfunktionen reeller Zufallsgrößen, unter anderem
  - $0 \leq F_{(X,Y)}(x, y) \leq 1$;
  - $\lim\limits_{x \to -\infty} F_{(X,Y)}(x, y) = \lim\limits_{y \to -\infty} F_{(X,Y)}(x, y) = 0$;
  - $\lim\limits_{x,y \to \infty} F_{(X,Y)}(x, y) = 1$;
  - die Funktion $F_{(X,Y)}(x, y)$ ist bezüglich jeder Variable monoton nicht fallend.

Verteilungsdichte eines stetigen Zufallsvektors

- Für stetige Zufallsvektoren (Zufallsvektoren mit absolut stetiger Verteilung) kann die Verteilung auch durch die Verteilungsdichte $f_{(X,Y)}(s,t),\ (s,t) \in \mathbb{R}^2$, bestimmt werden:

$$F_{(X,Y)}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{(X,Y)}(s,t)\,\mathrm{d}t\mathrm{d}s\,.$$

- Dann gilt für geeignete Teilmengen $B \subset \mathbb{R}^2$:

$$P((X,Y) \in B) = \int \int_{B} f_{(X,Y)}(s,t)\,\mathrm{d}t\mathrm{d}s\,.$$

- Verteilungsdichten für Zufallsvektoren haben die bestimmenden Eigenschaften von Dichtefunktionen für reelle Zufallsgrößen:
  - $f_{(X,Y)}(s,t) \geq 0\,, \qquad (s,t) \in \mathbb{R}^2\,;$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(s,t)\,\mathrm{d}s\mathrm{d}t = 1\,.$

Verbundverteilung und Randverteilungen

- Die gemeinsame Verteilung des Zufallsvektors (X, Y), gegeben z.B. durch die Verbundverteilungsfunktion oder die gemeinsame Verteilungsdichte, bestimmt eindeutig die Verteilungen der Komponenten X und Y (die Randverteilungen), wenn diese als einzelne Zufallsgrößen betrachtet werden.
- So gelten:
  - $F_X(x) = P(X < x) = \lim_{y \to \infty} F_{(X,Y)}(x, y), \quad x \in \mathbb{R}$;
  - $F_Y(y) = P(Y < y) = \lim_{x \to \infty} F_{(X,Y)}(x, y), \quad y \in \mathbb{R}$;
  - falls die Verteilungsdichte für den Zufallsvektor (X, Y) existiert, existieren auch die Dichtefunktionen für X und Y und es gelten

$$f_X(s) = \int_{-\infty}^{\infty} f_{(X,Y)}(s, t)\,\mathrm{d}t, \ s \in \mathbb{R}, \qquad \text{sowie}$$

$$f_Y(t) = \int_{-\infty}^{\infty} f_{(X,Y)}(s, t)\,\mathrm{d}s, \ t \in \mathbb{R}.$$

Unabhängigkeit von Zufallsgrößen

- **Definition:** Zwei Zufallsgrößen X und Y heißen stochastisch unabhängig, falls für beliebige reelle Zahlen x, y gilt:

$$P\left(\{X < x\} \cap \{Y < y\}\right) = P(X < x) \cdot P(Y < y)\,.$$

- D.h. die gemeinsame Verteilungsfunktion ist das Produkt der Randverteilungsfunktionen:

$$F_{(X,Y)}(x, y) = F_X(x) \cdot F_Y(y)\,, \qquad (x, y) \in \mathbb{R}^2\,.$$

- Damit ist auch die gemeinsame Dichtefunktion das Produkt der Randdichten:

$$f_{(X,Y)}(s, t) = f_X(s) \cdot f_Y(t)\,, \qquad (s, t) \in \mathbb{R}^2\,.$$

Momente von Zufallsvektoren

- Wichtige von der Verteilung eines Zufallsvektors abgeleitete Kenngrößen sind die Momente, für einen stetigen Zufallsvektor ist für nichtnegative ganze Zahlen k, l

$$\mathbf{E}\left[X^k Y^l\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s^k t^l f_{(X,Y)}(s,t)\,\mathrm{d}s\mathrm{d}t$$

  ein (im Allgemeinen gemischtes) Moment der Ordnung $k + l$ (falls es existiert).

- Momente erster Ordnung sind

$$\mathbf{E}X = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s f_{(X,Y)}(s,t)\,\mathrm{d}s\mathrm{d}t = \int_{-\infty}^{\infty} s f_X(s)\,\mathrm{d}s\,;$$

$$\mathbf{E}Y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t f_{(X,Y)}(s,t)\,\mathrm{d}s\mathrm{d}t = \int_{-\infty}^{\infty} t f_Y(t)\,\mathrm{d}t\,.$$

Zweite Momente von Zufallsvektoren

- Momente 2. Ordnung sind $\mathbf{E}X^2$ und $\mathbf{E}Y^2$ sowie das gemischte zweite Moment

$$\mathbf{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} stf_{(X,Y)}(s,t)\,\mathrm{d}s\mathrm{d}t\,.$$

- Die entsprechenden zentralen zweiten Momente sind

$$\begin{aligned}
\mathbf{Var}X &= \mathbf{E}(X-\mathbf{E}X)^2 = \mathbf{E}\left[X^2\right] - \mathbf{E}X\cdot\mathbf{E}X\,,\\
\mathbf{Var}Y &= \mathbf{E}(Y-\mathbf{E}Y)^2 = \mathbf{E}\left[Y^2\right] - \mathbf{E}Y\cdot\mathbf{E}Y\,,\\
\mathbf{Cov}\,(X,Y) &= \mathbf{E}[(X-\mathbf{E}X)(Y-\mathbf{E}Y)] = \mathbf{E}[XY] - \mathbf{E}X\cdot\mathbf{E}Y\,.
\end{aligned}$$

- Es gilt:

$$\mathbf{Var}(X+Y) = \mathbf{Var}X + \mathbf{Var}Y + 2\cdot\mathbf{Cov}\,(X,Y)\,.$$

Korrelationskoeffizient

- Gilt für X und Y jeweils $0 < \mathbf{Var}X < \infty\,, 0 < \mathbf{Var}\,Y < \infty$, dann definiert man den Korrelationskoeffizient von X und Y als

$$\mathbf{Corr}\,(X, Y) = \varrho_{X,Y} = \frac{\mathbf{Cov}\,(X, Y)}{\sqrt{\mathbf{Var}X}\sqrt{\mathbf{Var}\,Y}}\,.$$

- Der Korrelationskoeffizient liegt immer zwischen -1 und 1 :

$$-1 \leq \varrho_{X,Y} \leq 1.$$

- Im Fall $|\varrho_{X,Y}| = 1$ besteht ein vollständiger linearer Zusammenhang zwischen beiden Größen.
- Zwei Zufallsgrößen X und Y heißen unkorreliert, falls

$$\mathbf{Cov}\,(X, Y) = 0$$

und damit $\varrho_{X,Y} = 0$ gilt.

Unabhängigkeit und Unkorreliertheit

- Sind zwei Zufallsgrößen X und Y mit endlichen Erwartungswerten stochastisch unabhängig, dann gilt $\mathbf{E}(X \cdot Y) = \mathbf{E}X \cdot \mathbf{E}Y$.
- Damit folgt aus der Unabhängigkeit zweier Zufallsgrößen X und Y deren Unkorreliertheit :

$$\mathbf{Cov}\,(X, Y) = \mathbf{E}(X \cdot Y) - \mathbf{E}X \cdot \mathbf{E}Y = 0\,.$$

- Sind zwei Zufallsgrößen X und Y stochastisch unabhängig (oder unkorreliert), dann gilt für deren Summe:

$$\mathbf{Var}(X + Y) = \mathbf{Var}X + \mathbf{Var}Y\,.$$

- Achtung:
  Aus der Unkorreliertheit folgt i. Allg. nicht die Unabhängigkeit.
- Bei einer zweidimensionalen Normalverteilung folgt aus der Unkorreliertheit der Komponenten auch deren Unabhängigkeit .

Beispiel: zweidimensionale Normalverteilung

- Ein stetiger Zufallsvektor (X, Y) besitzt eine zweidimensionale Normalverteilung, wenn seine Dichtefunktion lautet

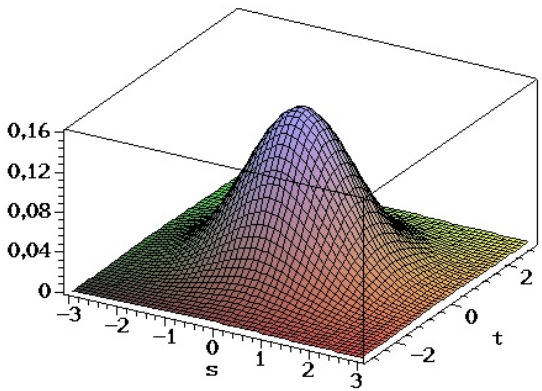
$$f_{(X,Y)}(s,t) = c \cdot \mathrm{e}^{-\frac{1}{2(1-\varrho^2)}\left[\frac{(s-\mu_X)^2}{\sigma_X^2} - 2\varrho\frac{(s-\mu_X)(t-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(t-\mu_Y)^2}{\sigma_Y^2}\right]}$$

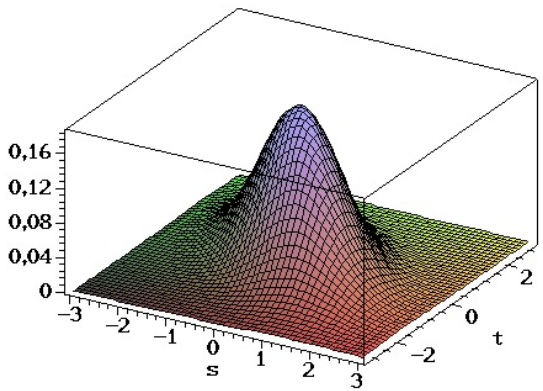
$$\text{mit} \quad c = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}}\,.$$

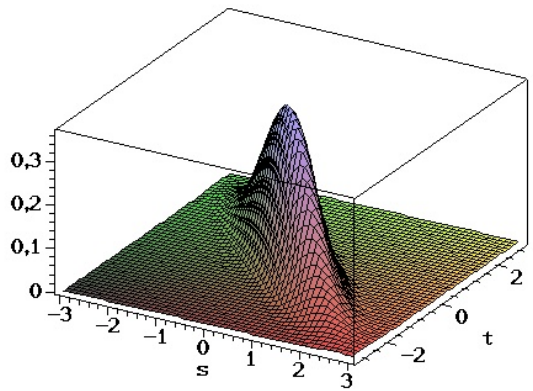
- Dann gelten: $\mathbf{E}X = \mu_X$, $\mathbf{E}Y = \mu_Y$, $\mathbf{Var}X = \sigma_X^2$, $\mathbf{Var}Y = \sigma_Y^2$, $\varrho_{X,Y} = \varrho \in (-1, 1)$.
- Die einzelnen Komponenten X und Y des Zufallsvektors sind hier normalverteilte Zufallsgrößen mit den oben angegebenen Parametern.
- In diesem Fall sind X und Y genau dann unabhängig, wenn sie unkorreliert sind, d.h. $\varrho_{X,Y} = \varrho = 0$ gilt.

Dichtefunktionsgrafiken zweidimensionaler Normalverteilungen

Dichtefunktionen von normalverteilten Zufallsvektoren (X, Y) mit $\mathbf{E}X = \mathbf{E}Y = 0$, $\mathbf{Var}X = \mathbf{Var}Y = 1$ sowie $\varrho = 0$ (links), $\varrho = -0.5$ (Mitte) und $\varrho = -0.9$ (rechts).





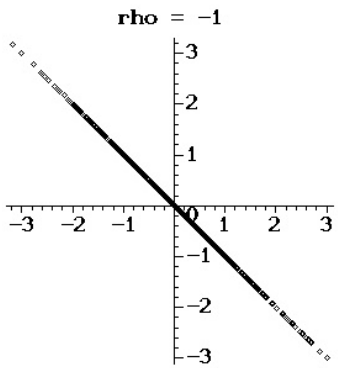


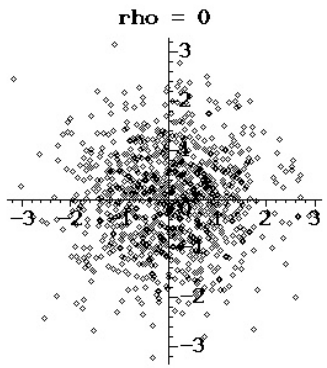
Streudiagramme für simulierte Werte

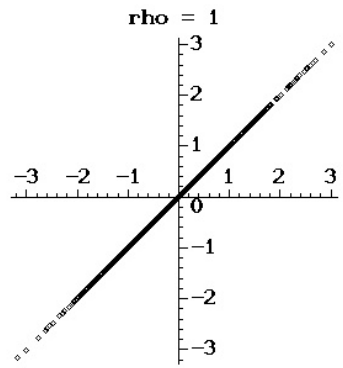
Streudiagramme (Scatterplots) von 1000 simulierten Realisierungen von normalverteilten Zufallsvektoren (X, Y) mit

$$\mathbf{E}X = \mathbf{E}Y = 0\,, \quad \mathbf{Var}X = \mathbf{Var}Y = 1 \quad \text{sowie}$$

$\varrho = -1$ (links), $\varrho = 0$ (Mitte) und $\varrho = 1$ (rechts).

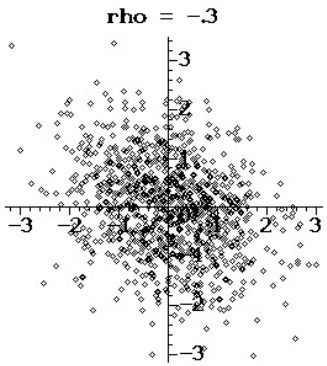


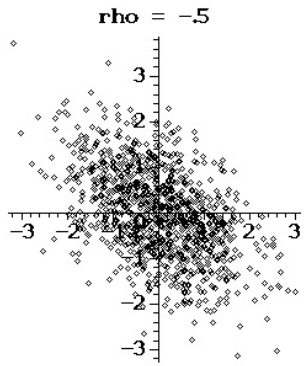


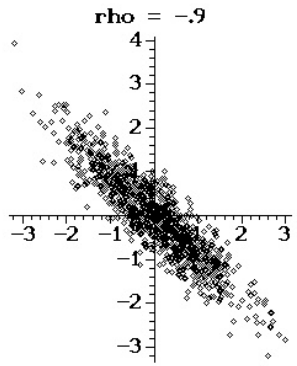


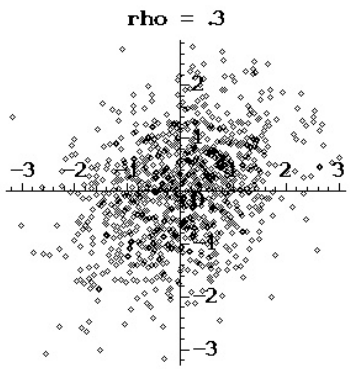
Streudiagramme für simulierte Werte – Fortsetzung

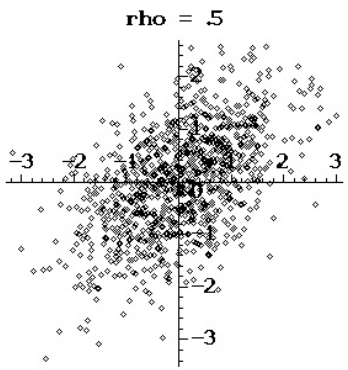
$\varrho = \pm 0.3$ (links), $\varrho = \pm 0.5$ (Mitte), $\varrho = \pm 0.9$ (rechts).

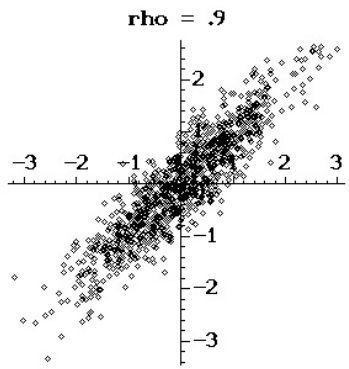












Schätzung des Korrelationskoeffizienten

- Für eine geeignete Stichprobe $(X_1, Y_1), \ldots, (X_n, Y_n)$ ist der Stichprobenkorrelationskoeffizient eine gute Schätzfunktion für den Korrelationskoeffizienten,

$$\hat{\varrho}_{X,Y} = R_{X,Y} := \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

- Für eine konkrete Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ erhält man so den empirischen Korrelationskoeffizienten

$$r_{X,Y} := \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}.$$

Schätzung des Korrelationskoeffizienten – Fortsetzung

- Diese Formeln basieren auf der Schätzung der Kovarianz zwischen X und Y durch die empirische Kovarianz

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

und der Beziehung $r_{X,Y} = \dfrac{s_{X,Y}}{s_X s_Y}$.

- Möglich ist die Berechnung von $r_{X,Y}$ auch durch

$$r_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}.$$

- Da auch andere Korrelationskoeffizienten in der Statistik eine Rolle spielen, wird $r_{X,Y}$ auch gewöhnlicher oder Bravais-Pearsonscher Korrelationskoeffizient genannt.

Eigenschaften des gewöhnlichen Korrelationskoeffizienten

- Es gelten $r_{X,Y} = r_{Y,X}$ und $-1 \leq r_{X,Y} \leq 1$.
- Der gewöhnliche Korrelationskoeffizient $r_{X,Y}$ ist ein Maß für die Stärke und Richtung des linearen Zusammenhanges zwischen den $x-$ und $y-$Werten der Stichprobenwerte $(x_i, y_i), i = 1, \ldots, n$.
- $r_{X,Y} > 0$ bedeutet, dass großen $x-$Werten vorwiegend große $y-$Werte entsprechen und umgekehrt. Man spricht dann von positiver oder gleichsinniger Korrelation.
  $r_{X,Y} < 0$ bedeutet, dass großen $x-$Werten vorwiegend kleine $y-$Werte entsprechen und umgekehrt. Man spricht dann von negativer oder ungleichsinniger Korrelation.
- Das Quadrat des gewöhnlichen Korrelationskoeffizienten $B_{X,Y} = r_{X,Y}^2$ heißt empirisches Bestimmtheitsmaß.
- Es gilt $0 \leq B_{X,Y} \leq 1$.

Beispiel 6.1: Alter und Blutdruck

Alter X und Blutdruck Y von 15 zufällig ausgewählten Frauen

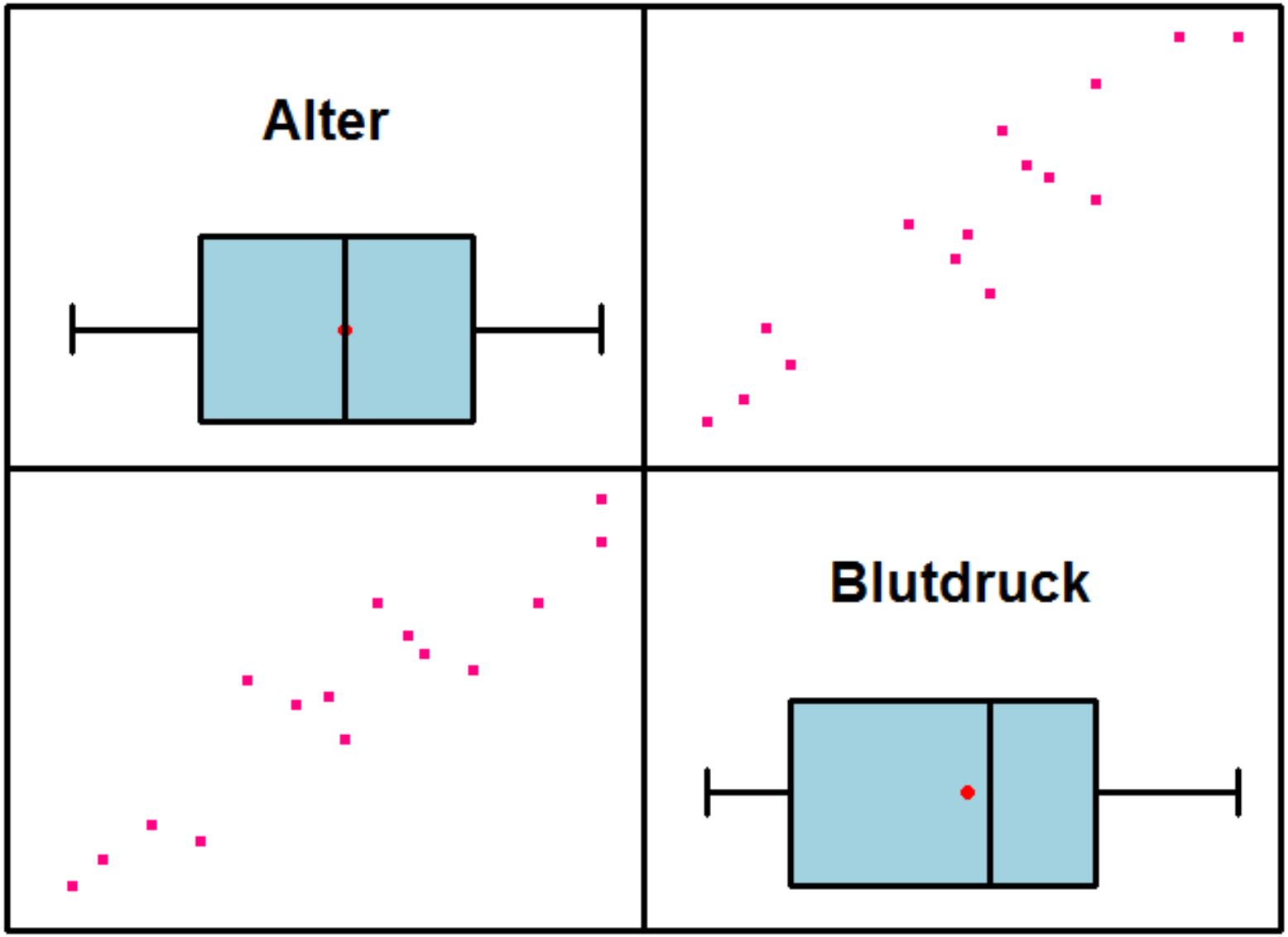
Quelle: J. Hartung: Statistik, Oldenbourg Verlag 2009, Kap. IX, Abschnitt 1

in Statgraphics:
Describe → Numeric Data →
Multiple-Variable Analysis
Beschreiben → Numerische Daten →
Analyse mehrerer Variablen

$n = 15$;
$\overline{x} = 47.0$, $\overline{y} = 134.067$;
$s_X = 10.5221$, $s_Y = 13.5408$;
$r_{X,Y} = 0.9375$.

i	x_i	y_i
1	47	129
2	52	139
3	30	112
4	35	119
5	59	145
6	44	133
7	63	152
8	38	117
9	49	145
10	41	136
11	32	115
12	55	137
13	46	134
14	51	141
15	63	157

Streudiagramm Beispiel 6.1 (Statgraphics)



Test auf Unkorreliertheit für normalverteilte Merkmale

- **Voraussetzung:** Die Zufallsvektoren $(X_1, Y_1), ..., (X_n, Y_n)$ sind unabhängig und identisch normalverteilt mit Parametern $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho$.
- **Hypothesen:** $H_0 : \varrho = 0$, $H_A : \varrho \neq 0$.
- **Bemerkung:** Da hier eine Normalverteilung vorausgesetzt wird, ist die Nullhypothese gleichbedeutend mit der Hypothese über die stochastische Unabhängigkeit der beiden Zufallsgrößen X und Y.
- **Testgröße:** $$T = \frac{R_{X,Y}\sqrt{n-2}}{\sqrt{1-R_{X,Y}^2}} \overset{H_0}{\sim} t_{n-2}.$$
- **Kritischer Bereich:** $K = \{t \in \mathbb{R} : |t| > t_{n-2;1-\alpha/2}\}$.
- **Einseitige Tests:**
  Für $H_A : \varrho > 0$ gilt $K = \{t \in \mathbb{R} : t > t_{n-2;1-\alpha}\}$ und
  für $H_A : \varrho < 0$ gilt $K = \{t \in \mathbb{R} : t < -t_{n-2;1-\alpha}\}$.

Fortsetzung Beispiel 6.1 Alter und Blutdruck

- $H_0 : \varrho = 0 \quad$ gegen $\quad H_A : \varrho \neq 0\,,$
- $r_{X,Y} = 0.9375, n = 15 \quad \Longrightarrow \quad t = \frac{0.9375}{\sqrt{1-0.9375^2}}\sqrt{13} = 9.71\,,$
- $\alpha = 0.05 \quad \Longrightarrow \quad t_{n-2;1-\frac{\alpha}{2}} = t_{13;0.975} = 2.16$
- $|t| = 9.71 > 2.16 = t_{13;0.975} \quad \Longrightarrow \quad H_0$ wird abgelehnt.
- Die Korrelation zwischen Alter und Blutdruck ist signifikant von Null verschieden. Bei Frauen gibt es eine signifikante Abhängigkeit zwischen Alter und Blutdruck.
- **Statgraphics:**

Correlations

	Alter	Blutdruck
Alter		0,9375
		(15)
		0,0000
Blutdruck	0,9375	
	(15)	
	0,0000	

Correlation
(Sample Size)
P-Value

Test auf festen Wert $\varrho_0 \neq 0$

- **Voraussetzung:** Die Zufallsvektoren $(X_1, Y_1), ..., (X_n, Y_n)$ sind unabhängig und identisch normalverteilt mit Parametern $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho$.
- **Hypothesen:** $H_0 : \varrho = \varrho_0 (\neq 0), \quad H_A : \varrho \neq \varrho_0$.
- Für einen exakten Test auf Basis der Testgröße $R_{X,Y}$ existieren Tafeln.
- Man kann aber auch schon für kleine Werte n einen Test nutzen, der die Fishersche $Z-$Transformation verwendet:

$$Z = \operatorname{artanh}(R_{X,Y}) = \frac{1}{2} \ln\left(\frac{1 + R_{X,Y}}{1 - R_{X,Y}}\right),$$

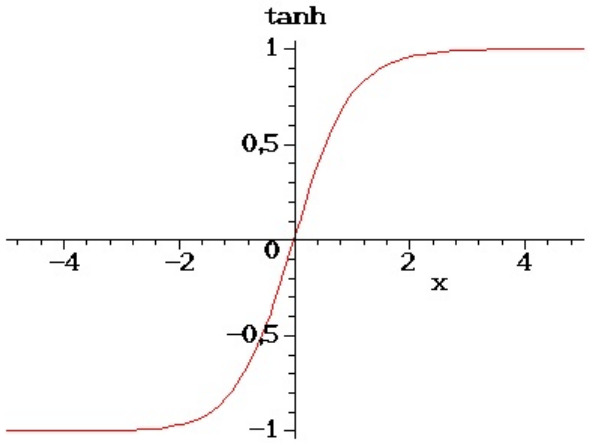
$$z = \operatorname{artanh}(r_{X,Y}) = \frac{1}{2} \ln\left(\frac{1 + r_{X,Y}}{1 - r_{X,Y}}\right).$$

Hyperbeltangens und seine Umkehrfunktion

Hyperbeltangens, Tangens hyperbolicus

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

für $x \in \mathbb{R}$.

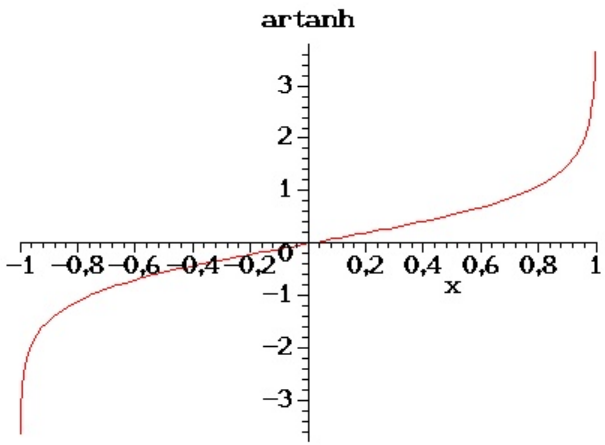


Dazu Umkehrfunktion (inverse Funktion):

Area Hyperbeltangens, Area Tangens hyperbolicus

$$\operatorname{artanh}(x) = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right),$$

für $-1 < x < 1$.



Approximativer Test auf festen Wert $\varrho_0 \neq 0$

- **Voraussetzung:** Die Zufallsvektoren $(X_1, Y_1), \ldots, (X_n, Y_n)$ sind unabhängig und identisch normalverteilt mit Parametern $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho$.
- **Hypothesen:** $H_0 : \varrho = \varrho_0 \, (\neq 0)$, $H_A : \varrho \neq \varrho_0$.
- **Testgröße:** $T = (Z - z_0)\sqrt{n-3} \overset{H_0}{\sim} \mathrm{N}(0,1)$ mit

$$Z = \operatorname{artanh}\left(R_{X,Y}\right) = \frac{1}{2} \ln\left(\frac{1 + R_{X,Y}}{1 - R_{X,Y}}\right),$$

$$z_0 = \mathbf{E}_{H_0}[Z] = \frac{1}{2} \ln\left(\frac{1 + \varrho_0}{1 - \varrho_0}\right) + \frac{\varrho_0}{2(n-1)}.$$

- **Kritischer Bereich:** $K = \{t \in \mathbb{R} : |t| > z_{1-\alpha/2}\}$.
- **Einseitige Tests:**
  Für $H_A : \varrho > \varrho_0$ gilt $K = \{t \in \mathbb{R} : t > z_{1-\alpha}\}$ und
  für $H_A : \varrho < \varrho_0$ gilt $K = \{t \in \mathbb{R} : t < -z_{1-\alpha}\}$.

Fortsetzung Beispiel 6.1 Alter und Blutdruck

- Wir betrachten nun zum Niveau $\alpha = 0.05$ den Test
  $H_0 : \varrho \leq 0.90$ gegen $H_A : \varrho > 0.90$.
- Dann erhalten wir aus

$$r_{X,Y} = 0.9375 \qquad \text{die Werte}$$

$$z = \frac{1}{2} \ln \left(\frac{1 + r_{X,Y}}{1 - r_{X,Y}} \right) = 1.717$$

$$z_0 = \frac{1}{2} \ln \left(\frac{1.90}{0.10} \right) + \frac{0.90}{2 \cdot 14} = 1.504$$

$$t = (z - z_0) \cdot \sqrt{n - 3} = 0.738 < 1.645 = z_{0.95} ,$$

- Folglich können wir die Hypothese H_0 nicht ablehnen.
- Zum Niveau 0.05 ist also nicht signifikant gesichert, dass die Korrelation zwischen Alter und Blutdruck bei Frauen größer als 0.90 ist.

Approximatives Konfidenzintervall im Beispiel 6.1 mit R

- Im Beispiel 6.1 erhält man ein approximatives Konfidenzintervall für ϱ zum Niveau $1-\alpha = 0.95$ z.B. mit Hilfe des Statistikprogrammes **R**:

```
> Alter<-c(47,52,30,35,59,44,63,38,49,41,32,55,46,51,63)
> Blutdruck<-c(129,139,112,119,145,133,152,117,145,136,115,137,134,141,157)
> cor.test(Alter,Blutdruck)

        Pearson's product-moment correlation

data:  Alter and Blutdruck
t = 9.7131, df = 13, p-value = 2.519e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8181349 0.9794044
sample estimates:
      cor
0.9374939
```