Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

Kapitel XI - Korrelationsrechnung

Deskriptive Statistik

Prof. Dr. W.-D. Heller
Hartwig Senska
Carlo Siebenschuh

Aufgabe der Korrelationsrechnung

Da es bei quantitativen Merkmalen nicht immer sinnvoll ist, mittels linearer Regression einen Zusammenhang zwischen beiden Merkmalen zu berechnen, erscheint es nützlich, eine Methode zu entwickeln, mit deren Hilfe man Hinweise erhalten kann, ob die Anwendung der linearen Regression berechtigt ist.

Ziel der Korrelationsrechnung

Messung der „Güte“ eines unterstellten linearen Zusammenhangs

Dazu vergleichen wir die Varianz der „Trendwerte“

$$\hat{y}_i = \hat{m}x_i + \hat{b}$$

mit der Varianz der y-Werte.

Aufgabe der Korrelationsrechnung

Da es bei quantitativen Merkmalen nicht immer sinnvoll ist, mittels linearer Regression einen Zusammenhang zwischen beiden Merkmalen zu berechnen, erscheint es nützlich, eine Methode zu entwickeln, mit deren Hilfe man Hinweise erhalten kann, ob die Anwendung der linearen Regression berechtigt ist.

Ziel der Korrelationsrechnung

Messung der „Güte“ eines unterstellten linearen Zusammenhangs

Dazu vergleichen wir die Varianz der „Trendwerte“

$$\hat{y}_i = \hat{m}x_i + \hat{b}$$

mit der Varianz der y-Werte.

Aufgabe der Korrelationsrechnung

Da es bei quantitativen Merkmalen nicht immer sinnvoll ist, mittels linearer Regression einen Zusammenhang zwischen beiden Merkmalen zu berechnen, erscheint es nützlich, eine Methode zu entwickeln, mit deren Hilfe man Hinweise erhalten kann, ob die Anwendung der linearen Regression berechtigt ist.

Ziel der Korrelationsrechnung

Messung der „Güte“ eines unterstellten linearen Zusammenhangs

Dazu vergleichen wir die Varianz der „Trendwerte“

$$\hat{y}_i = \hat{m}x_i + \hat{b}$$

mit der Varianz der y-Werte.

Bestimmtheitsmaß

Das Bestimmtheitsmaß stellt das Verhältnis aus der Varianz der $\hat{y}$-Werte und der Varianz der y-Werte dar.

Varianz der Trendwerte $\hat{y}$:

$$\begin{aligned} s_{\hat{y}}^2 &= Var(\hat{m}x_i + \hat{b}) \\ &= \hat{m}^2 s_x^2 \end{aligned}$$

Varianz der y-Werte

$$s_y^2.$$

Somit ist das Bestimmtheitsmaß definiert als

$$\frac{\hat{m}^2 \cdot s_x^2}{s_y^2}$$

Bestimmtheitsmaß

Das Bestimmtheitsmaß stellt das Verhältnis aus der Varianz der $\hat{y}$-Werte und der Varianz der y-Werte dar.
Varianz der Trendwerte $\hat{y}$:

$$\begin{aligned} s_{\hat{y}}^2 &= Var(\hat{m}x_i + \hat{b}) \\ &= \hat{m}^2 s_x^2 \end{aligned}$$

Varianz der y-Werte

$$s_y^2.$$

Somit ist das Bestimmtheitsmaß definiert als

$$\frac{\hat{m}^2 \cdot s_x^2}{s_y^2}$$

Bestimmtheitsmaß

Das Bestimmtheitsmaß stellt das Verhältnis aus der Varianz der $\hat{y}$-Werte und der Varianz der y-Werte dar.
Varianz der Trendwerte $\hat{y}$:

$$\begin{aligned} s_{\hat{y}}^2 &= Var(\hat{m}x_i + \hat{b}) \\ &= \hat{m}^2 s_x^2 \end{aligned}$$

Varianz der y-Werte

$$s_y^2.$$

Somit ist das Bestimmtheitsmaß definiert als

$$\frac{\hat{m}^2 \cdot s_x^2}{s_y^2}$$

Bestimmtheitsmaß

Das Bestimmtheitsmaß stellt das Verhältnis aus der Varianz der $\hat{y}$-Werte und der Varianz der y-Werte dar.
Varianz der Trendwerte $\hat{y}$:

$$\begin{aligned} s_{\hat{y}}^2 &= Var(\hat{m}x_i + \hat{b}) \\ &= \hat{m}^2 s_x^2 \end{aligned}$$

Varianz der y-Werte

$$s_y^2.$$

Somit ist das Bestimmtheitsmaß definiert als

$$\frac{\hat{m}^2 \cdot s_x^2}{s_y^2}$$

Bestimmtheitsmaß

Dies kann man umformen:

$$\begin{aligned}
\frac{\hat{m}^2 s_x^2}{s_y^2} &= \frac{s_x^2}{s_y^2} \cdot \left(\frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \cdot \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i^2)} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n y_i - \frac{\sum_{i=1}^n y_i}{n} \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - 2 \cdot \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i + \frac{(\sum_{i=1}^n x_i)^2}{n}} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \bar{x} \cdot \sum_{i=1}^n y_i - \bar{y} \cdot \sum_{i=1}^n x_i + n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - 2 \cdot \bar{x} \sum_{i=1}^n x_i + n \cdot \bar{x}^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 = \frac{1}{s_x^2 s_y^2} \cdot \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2 .
\end{aligned}$$

Der Ausdruck in der Klammer stellt die Kovarianz dar (später).

Bestimmtheitsmaß

Dies kann man umformen:

$$
\begin{aligned}
\frac{\hat{m}^2 s_x^2}{s_y^2} &= \frac{s_x^2}{s_y^2} \cdot \left(\frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \cdot \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i^2)} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n y_i - \frac{\sum_{i=1}^n y_i}{n} \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - 2 \cdot \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i + \frac{(\sum_{i=1}^n x_i)^2}{n}} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \bar{x} \cdot \sum_{i=1}^n y_i - \bar{y} \cdot \sum_{i=1}^n x_i + n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - 2 \cdot \bar{x} \sum_{i=1}^n x_i + n \cdot \bar{x}^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 = \frac{1}{s_x^2 s_y^2} \cdot \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2 .
\end{aligned}
$$

Der Ausdruck in der Klammer stellt die Kovarianz dar (später).

Bestimmtheitsmaß

Dies kann man umformen:

$$
\begin{aligned}
\frac{\hat{m}^2 s_x^2}{s_y^2} &= \frac{s_x^2}{s_y^2} \cdot \left(\frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \cdot \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i^2)} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n y_i - \frac{\sum_{i=1}^n y_i}{n} \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - 2 \cdot \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i + \frac{(\sum_{i=1}^n x_i)^2}{n}} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\sum_{i=1}^n x_i y_i - \bar{x} \cdot \sum_{i=1}^n y_i - \bar{y} \cdot \sum_{i=1}^n x_i + n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - 2 \cdot \bar{x} \sum_{i=1}^n x_i + n \cdot \bar{x}^2} \right)^2 \\
&= \frac{s_x^2}{s_y^2} \cdot \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 = \frac{1}{s_x^2 s_y^2} \cdot \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2 .
\end{aligned}
$$

Der Ausdruck in der Klammer stellt die Kovarianz dar (später).

Bestimmtheitsmaß

Das Bestimmtheitsmaß beschreibt den Anteil an der Varianz der y-Werte, der sich bei linearer Regression aus der Varianz der x-Werte begründen lässt, also den Teil der Varianz, den man auch erhalten würde, wenn der lineare Zusammenhang exakt eingehalten würde, die Störgrößen also alle gleich 0 wären.

Kovarianz

Den Klammerausdruck aus dem Bestimmtheitsmaß nennt man die Kovarianz der beiden Merkmale:

$$Cov(x,y) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Analog zur Varianz lässt sich dieser Ausdruck umformen:

$$\begin{aligned}
Cov(x,y) &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}) \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\bar{x}\left(\sum_{i=1}^{n} y_i\right) - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i\right)\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.
\end{aligned}$$

Kovarianz

Den Klammerausdruck aus dem Bestimmtheitsmaß nennt man die Kovarianz der beiden Merkmale:

$$Cov(x,y) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Analog zur Varianz lässt sich dieser Ausdruck umformen:

$$\begin{aligned}
Cov(x,y) &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}) \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\bar{x}(\sum_{i=1}^{n} y_i) - (\frac{1}{n} \cdot \sum_{i=1}^{n} x_i)\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.
\end{aligned}$$

Kovarianz

Es werden das Leergewicht und das maximale Zulandungsgewicht bei PKWs betrachtet. Die Summe dieser beiden bildet das zulässige Gesamtgewicht.

$i =$	1	2	3	4	5	6	7
Leergewicht x_i	873	967	932	1345	1138	996	1782
Zuladungsgewicht y_i	481	518	495	647	584	489	691
zul. Gesamtgewicht $z_i = x_i + y_i$	1354	1485	1427	1992	1722	1485	2473

$i =$	8	9	10	$\sum$
Leergewicht x_i	2020	1420	1382	12855
Zuladungsgewicht y_i	360	648	585	5498
zul. Gesamtgewicht $z_i = x_i + y_i$	2380	2068	1967	18353

Kovarianz

Es werden das Leergewicht und das maximale Zulandungsgewicht bei PKWs betrachtet. Die Summe dieser beiden bildet das zulässige Gesamtgewicht.

$i =$	1	2	3	4	5	6	7
Leergewicht x_i	873	967	932	1345	1138	996	1782
Zuladungsgewicht y_i	481	518	495	647	584	489	691
zul. Gesamtgewicht $z_i = x_i + y_i$	1354	1485	1427	1992	1722	1485	2473

$i =$	8	9	10	$\sum$
Leergewicht x_i	2020	1420	1382	12855
Zuladungsgewicht y_i	360	648	585	5498
zul. Gesamtgewicht $z_i = x_i + y_i$	2380	2068	1967	18353

Kovarianz

Beispiel 11.1

Arithmetisches Mittel für Einzelwerte und zulässiges Gesamtgewicht:

$i =$	1	2	3	4	5	6
x_i	873	967	932	1345	1138	996
y_i	481	518	495	647	584	489
x_i^2	762129	935089	868624	1809025	1295044	992016
y_i^2	231361	268324	245025	418609	341056	239121
$x_i + y_i$	1354	1485	1427	1992	1722	1485
$(x_i + y_i)^2$	1833316	2205225	2036329	3968064	2965284	2205225

$i =$	7	8	9	10	$\sum$
x_i	1782	2020	1420	1382	12855
y_i	691	360	648	585	5498
x_i^2	3175524	4080400	2016400	1909924	17844175
y_i^2	477481	129600	419904	342225	3112706
$x_i + y_i$	2473	2380	2068	1967	18353
$(x_i + y_i)^2$	6115729	5664400	4276624	3869089	35139285

Kovarianz

Beispiel 11.1

Arithmetisches Mittel für Einzelwerte und zulässiges Gesamtgewicht:

$i =$	1	2	3	4	5	6
x_i	873	967	932	1345	1138	996
y_i	481	518	495	647	584	489
x_i^2	762129	935089	868624	1809025	1295044	992016
y_i^2	231361	268324	245025	418609	341056	239121
$x_i + y_i$	1354	1485	1427	1992	1722	1485
$(x_i + y_i)^2$	1833316	2205225	2036329	3968064	2965284	2205225

$i =$	7	8	9	10	$\sum$
x_i	1782	2020	1420	1382	12855
y_i	691	360	648	585	5498
x_i^2	3175524	4080400	2016400	1909924	17844175
y_i^2	477481	129600	419904	342225	3112706
$x_i + y_i$	2473	2380	2068	1967	18353
$(x_i + y_i)^2$	6115729	5664400	4276624	3869089	35139285

Kovarianz

Beispiel 11.1

Arithmetisches Mittel für Einzelwerte und zulässiges Gesamtgewicht:

$i =$	1	2	3	4	5	6
x_i	873	967	932	1345	1138	996
y_i	481	518	495	647	584	489
x_i^2	762129	935089	868624	1809025	1295044	992016
y_i^2	231361	268324	245025	418609	341056	239121
$x_i + y_i$	1354	1485	1427	1992	1722	1485
$(x_i + y_i)^2$	1833316	2205225	2036329	3968064	2965284	2205225

$i =$	7	8	9	10	$\sum$
x_i	1782	2020	1420	1382	12855
y_i	691	360	648	585	5498
x_i^2	3175524	4080400	2016400	1909924	17844175
y_i^2	477481	129600	419904	342225	3112706
$x_i + y_i$	2473	2380	2068	1967	18353
$(x_i + y_i)^2$	6115729	5664400	4276624	3869089	35139285

Kovarianz

Beispiel 11.1
Aus der Tabelle ermitteln wir:

$s_x^2 = 131907.25$; $s_y^2 = 8990.56$;

$Cov(x, y) = 2352.3$ und

$s_{x+y}^2 = 145602.41$,

also

$s_{x+y}^2 = s_x^2 + s_y^2 + 2Cov(x, y)$.

Kovarianz

Varianz der Summe

$$
\begin{aligned}
s^2_{x+y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - (\bar{x} + \bar{y}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + y_i - \bar{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= s_x^2 + s_y^2 + 2Cov(x, y).
\end{aligned}
$$

Bei der Varianz der Summe geht also wesentlich die Beziehung zwischen den Merkmalen ein, wie sie sich aus der Kovarianz ergibt.

Kovarianz

Varianz der Summe

$$
\begin{aligned}
s^2_{x+y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - (\bar{x} + \bar{y}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + y_i - \bar{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= s_x^2 + s_y^2 + 2Cov(x, y).
\end{aligned}
$$

Bei der Varianz der Summe geht also wesentlich die Beziehung zwischen den Merkmalen ein, wie sie sich aus der Kovarianz ergibt.

Kovarianz

Kovarianz bei Verteilungen mit **absoluten** bzw. **relativen Häufigkeiten**
$h(a,b)$ bzw. $p(a,b)$ für $a \in M_1, b \in M_2$:

$$\begin{aligned} Cov(x,y) &= \frac{1}{n} \sum_{a \in M_1} \sum_{b \in M_2} (a-\bar{x})(b-\bar{y}) h(a,b) \\ &= \sum_{a \in M_1} \sum_{b \in M_2} (a-\bar{x})(b-\bar{y}) p(a,b). \end{aligned}$$

Entsprechend gilt:

$$\begin{aligned} Cov(x,y) &= \frac{1}{n} \sum_{a \in M_1} \sum_{b \in M_2} a \cdot b \cdot h(a,b) - \bar{x}\bar{y} \\ &= \sum_{a \in M_1} \sum_{b \in M_2} a \cdot b \cdot p(a,b) - \bar{x}\bar{y}. \end{aligned}$$

Kovarianz

Kovarianz bei Verteilungen mit **absoluten** bzw. **relativen Häufigkeiten**
$h(a,b)$ bzw. $p(a,b)$ für $a \in M_1, b \in M_2$:

$$\begin{aligned} Cov(x,y) &= \frac{1}{n} \sum_{a \in M_1} \sum_{b \in M_2} (a-\bar{x})(b-\bar{y}) h(a,b) \\ &= \sum_{a \in M_1} \sum_{b \in M_2} (a-\bar{x})(b-\bar{y}) p(a,b). \end{aligned}$$

Entsprechend gilt:

$$\begin{aligned} Cov(x,y) &= \frac{1}{n} \sum_{a \in M_1} \sum_{b \in M_2} a \cdot b \cdot h(a,b) - \bar{x}\bar{y} \\ &= \sum_{a \in M_1} \sum_{b \in M_2} a \cdot b \cdot p(a,b) - \bar{x}\bar{y}. \end{aligned}$$

Kovarianz

Sind die beiden Merkmale unabhängig, so gilt nach Definition:

$$p(a,b) = p(a) \cdot p(b) \quad \text{für } a \in M_1, b \in M_2.$$

Damit ergibt sich in diesem Fall

$$\begin{aligned} Cov(x,y) &= \sum_{a \in M_1} \sum_{b \in M_2} (a - \bar{x})(b - \bar{y}) p(a,b) \\ &= \sum_{b \in M_2} \left(\sum_{a \in M_1} (a - \bar{x}) p(a) \right) (b - \bar{y}) p(b). \end{aligned}$$

Kovarianz

Sind die beiden Merkmale unabhängig, so gilt nach Definition:

$$p(a,b) = p(a) \cdot p(b) \quad \text{für } a \in M_1, b \in M_2.$$

Damit ergibt sich in diesem Fall

$$\begin{aligned} Cov(x,y) &= \sum_{a \in M_1} \sum_{b \in M_2} (a - \bar{x})(b - \bar{y}) p(a,b) \\ &= \sum_{b \in M_2} \left(\sum_{a \in M_1} (a - \bar{x}) p(a) \right) (b - \bar{y}) p(b). \end{aligned}$$

Kovarianz

Für den Klammerausdruck gilt:

$$\begin{aligned}\sum_{a\in M_1}(a-\bar{x})p(a) &= \sum_{a\in M_1} a\cdot p(a) - \sum_{a\in M_1}\bar{x}p(a)\\ &= \bar{x}-\bar{x}\sum_{a\in M_1}p(a) = \bar{x}-\bar{x}=0.\end{aligned}$$

Damit gilt für **unabhängige Merkmale**

$$Cov(x,y)=0.$$

Kovarianz

Für den Klammerausdruck gilt:

$$\begin{aligned}\sum_{a\in M_1}(a-\bar{x})p(a) &= \sum_{a\in M_1} a\cdot p(a) - \sum_{a\in M_1}\bar{x}p(a)\\ &= \bar{x}-\bar{x}\sum_{a\in M_1}p(a) = \bar{x}-\bar{x}=0.\end{aligned}$$

Damit gilt für **unabhängige Merkmale**

$$Cov(x,y)=0.$$

Kovarianz

Beispiel 11.2

Bei der Messung von Körpergröße und -gewicht von 10 Personen ergab sich folgende Urliste:

(186,85), (155,70), (165,70), (186,75), (160,75), (155,50), (165,60), (175,60), (175,70), (160,65).

Kovarianz

Die Kovarianz kann mit Hilfe von folgendem Arbeitsschema berechnet werden:

$i =$	1	2	3	4	5
x_i	186	155	165	186	160
y_i	85	70	70	75	75
$x_i y_i$	15810	10850	11550	13950	12000

$i =$	6	7	8	9	10	$\sum$
x_i	155	165	175	175	160	1682
y_i	50	60	60	70	65	680
$x_i y_i$	7750	9900	10500	12250	10400	114960

Damit erhält man:

$$Cov(x, y) = 0.1 \cdot 114960 - 0.01 \cdot 1682 \cdot 680 = 58.4.$$

Kovarianz

Die Kovarianz kann mit Hilfe von folgendem Arbeitsschema berechnet werden:

$i =$	1	2	3	4	5
x_i	186	155	165	186	160
y_i	85	70	70	75	75
$x_i y_i$	15810	10850	11550	13950	12000

$i =$	6	7	8	9	10	$\sum$
x_i	155	165	175	175	160	1682
y_i	50	60	60	70	65	680
$x_i y_i$	7750	9900	10500	12250	10400	114960

Damit erhält man:

$$Cov(x, y) = 0.1 \cdot 114960 - 0.01 \cdot 1682 \cdot 680 = 58.4.$$

Korrelationskoeffizient

Der **(Bravais-Pearson-)Korrelationskoeffizient** entspricht der Kovarianz zweier Merkmale über einer statistischen Masse dividiert durch das Produkt der beiden Standardabweichungen :

$$r = \frac{Cov(x, y)}{s_x \cdot s_y}$$

Er ist also ein Maß für den linearen Zusammenhang zweier Merkmale. Das **Bestimmtheitsmaß** ist wegen

$$r^2 = \frac{\hat{m}^2 s_x^2}{s_y^2} = \frac{(Cov(x, y))^2}{s_x^2 \cdot s_y^2}.$$

das Quadrat des Korrelationskoeffizienten.

Korrelationskoeffizient

Der **(Bravais-Pearson-)Korrelationskoeffizient** entspricht der Kovarianz zweier Merkmale über einer statistischen Masse dividiert durch das Produkt der beiden Standardabweichungen :

$$r = \frac{Cov(x, y)}{s_x \cdot s_y}$$

Er ist also ein Maß für den linearen Zusammenhang zweier Merkmale.
Das **Bestimmtheitsmaß** ist wegen

$$r^2 = \frac{\hat{m}^2 s_x^2}{s_y^2} = \frac{(Cov(x, y))^2}{s_x^2 \cdot s_y^2}.$$

das Quadrat des Korrelationskoeffizienten.

Korrelationskoeffizient

Der **(Bravais-Pearson-)Korrelationskoeffizient** entspricht der Kovarianz zweier Merkmale über einer statistischen Masse dividiert durch das Produkt der beiden Standardabweichungen :

$$r = \frac{Cov(x, y)}{s_x \cdot s_y}$$

Er ist also ein Maß für den linearen Zusammenhang zweier Merkmale. Das **Bestimmtheitsmaß** ist wegen

$$r^2 = \frac{\hat{m}^2 s_x^2}{s_y^2} = \frac{(Cov(x, y))^2}{s_x^2 \cdot s_y^2}.$$

das Quadrat des Korrelationskoeffizienten.

Korrelationskoeffizient

Beispiel 11.3
Mit den Standardabweichungen $s_x = 11.09$ und $s_y = 9.27$ erhält man den Korrelationskoeffizienten zu Beispiel 11.2

$$r = \frac{58.4}{11.09 \cdot 9.27} = 0.568.$$

Korrelationskoeffizient

Bemerkung: Für den Korrelationskoeffizienten gilt

$$r = \frac{\hat{m} \cdot s_x}{s_y},$$

Damit erhält man folgende Beziehungen:

(a) Liegen alle Beobachtungswerte auf einer *steigenden Geraden*, dann ist $r > 0$ und $r^2 = 1$, also $r = 1$.

(b) Liegen alle Beobachtungswerte auf einer *fallenden Geraden*, dann ist $r < 0$ und $r^2 = 1$, also $r = -1$.

Es gilt für r:

$$-1 \leq r \leq 1.$$

wobei $\hat{m}$ der Anstieg der Regressionsgeraden ist. Ist also $r < 0$, so ist der Anstieg der Regressionsgeraden negativ, bei $r > 0$ ist er positiv.

Korrelationskoeffizient

Bemerkung: Für den Korrelationskoeffizienten gilt

$$r = \frac{\hat{m} \cdot s_x}{s_y},$$

Damit erhält man folgende Beziehungen:

(a) Liegen alle Beobachtungswerte auf einer *steigenden Geraden*, dann ist $r > 0$ und $r^2 = 1$, also $r = 1$.

(b) Liegen alle Beobachtungswerte auf einer *fallenden Geraden*, dann ist $r < 0$ und $r^2 = 1$, also $r = -1$.

Es gilt für r:

$$-1 \leq r \leq 1.$$

wobei $\hat{m}$ der Anstieg der Regressionsgeraden ist. Ist also $r < 0$, so ist der Anstieg der Regressionsgeraden negativ, bei $r > 0$ ist er positiv.

Korrelationskoeffizient

Bemerkung: Für den Korrelationskoeffizienten gilt

$$r = \frac{\hat{m} \cdot s_x}{s_y},$$

Damit erhält man folgende Beziehungen:

(a) Liegen alle Beobachtungswerte auf einer *steigenden Geraden*, dann ist $r > 0$ und $r^2 = 1$, also $r = 1$.

(b) Liegen alle Beobachtungswerte auf einer *fallenden Geraden*, dann ist $r < 0$ und $r^2 = 1$, also $r = -1$.

Es gilt für r:

$$-1 \leq r \leq 1.$$

wobei $\hat{m}$ der Anstieg der Regressionsgeraden ist. Ist also $r < 0$, so ist der Anstieg der Regressionsgeraden negativ, bei $r > 0$ ist er positiv.

Korrelationskoeffizient

Bemerkung: Für den Korrelationskoeffizienten gilt

$$r = \frac{\hat{m} \cdot s_x}{s_y},$$

Damit erhält man folgende Beziehungen:

(a) Liegen alle Beobachtungswerte auf einer *steigenden Geraden*, dann ist $r > 0$ und $r^2 = 1$, also $r = 1$.

(b) Liegen alle Beobachtungswerte auf einer *fallenden Geraden*, dann ist $r < 0$ und $r^2 = 1$, also $r = -1$.

Es gilt für r:

$$-1 \leq r \leq 1.$$

wobei $\hat{m}$ der Anstieg der Regressionsgeraden ist. Ist also $r < 0$, so ist der Anstieg der Regressionsgeraden negativ, bei $r > 0$ ist er positiv.

Korrelationskoeffizient

Bezeichnungen

Gilt $r > 0$, *so heißen die Merkmale* ***positiv korreliert***,

$r = 0$, *so heißen die Merkmale* ***unkorreliert***,

$r < 0$, *so heißen die Merkmale* ***negativ korreliert***.

Wichtig: Sind zwei Merkmale unabhängig, so sind sie also auch unkorreliert, aber die Umkehrung gilt nicht. Unkorrelierte Merkmale können abhängig sein.

Korrelationskoeffizient

Bezeichnungen

Gilt $r > 0$, *so heißen die Merkmale* ***positiv korreliert***,

$r = 0$, *so heißen die Merkmale* ***unkorreliert***,

$r < 0$, *so heißen die Merkmale* ***negativ korreliert***.

Wichtig: Sind zwei Merkmale unabhängig, so sind sie also auch unkorreliert, aber die Umkehrung gilt nicht. Unkorrelierte Merkmale können abhängig sein.

Korrelationskoeffizient

Bezeichnungen

Gilt $r > 0$, *so heißen die Merkmale* ***positiv korreliert***,
$r = 0$, *so heißen die Merkmale* ***unkorreliert***,
$r < 0$, *so heißen die Merkmale* ***negativ korreliert***.

Wichtig: Sind zwei Merkmale unabhängig, so sind sie also auch unkorreliert, aber die Umkehrung gilt nicht. Unkorrelierte Merkmale können abhängig sein.

Korrelationskoeffizient

Bezeichnungen

Gilt $r > 0$, *so heißen die Merkmale* ***positiv korreliert***,
$r = 0$, *so heißen die Merkmale* ***unkorreliert***,
$r < 0$, *so heißen die Merkmale* ***negativ korreliert***.

Wichtig: Sind zwei Merkmale unabhängig, so sind sie also auch unkorreliert, aber die Umkehrung gilt nicht. Unkorrelierte Merkmale können abhängig sein.

Spearmannscher Rangkorrelationskoeffizient

Bei Rangmerkmalen hat die Differenz zweier Merkmalswerte keine Aussagekraft. Damit ist auch der Korrelationskoeffizient, wenn er gebildet werden kann (etwa aus skalierten Werten), ohne direkte Bedeutung.

Die natürliche Reihenfolge lässt sich dennoch ausnutzen. Dazu ordnet man die Merkmalswerte der statistischen Reihe in ihrer natürlichen Reihenfolge und ersetzt den Merkmalswert durch die entsprechende „Rangziffer“.

Spearmannscher Rangkorrelationskoeffizient

Bei Rangmerkmalen hat die Differenz zweier Merkmalswerte keine Aussagekraft. Damit ist auch der Korrelationskoeffizient, wenn er gebildet werden kann (etwa aus skalierten Werten), ohne direkte Bedeutung.

Die natürliche Reihenfolge lässt sich dennoch ausnutzen. Dazu ordnet man die Merkmalswerte der statistischen Reihe in ihrer natürlichen Reihenfolge und ersetzt den Merkmalswert durch die entsprechende „Rangziffer“.

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.4

Seien als Prüfungsergebnisse von 6 Kandidaten befriedigend, ungenügend, sehr gut, gut, ausreichend, sehr gut bis gut erzielt worden, so erhält man die geordnete Reihe

sehr gut, sehr gut bis gut, gut, befriedigend, ausreichend, ungenügend.

Die Rangziffern lauten dann:

1 für sehr gut
2 für sehr gut bis gut
3 für gut
4 für befriedigend
5 für ausreichend
6 für ungenügend

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.4

Seien als Prüfungsergebnisse von 6 Kandidaten befriedigend, ungenügend, sehr gut, gut, ausreichend, sehr gut bis gut erzielt worden, so erhält man die geordnete Reihe

sehr gut, sehr gut bis gut, gut, befriedigend, ausreichend, ungenügend.

Die Rangziffern lauten dann:

1 für sehr gut
2 für sehr gut bis gut
3 für gut
4 für befriedigend
5 für ausreichend
6 für ungenügend

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.4

Seien als Prüfungsergebnisse von 6 Kandidaten befriedigend, ungenügend, sehr gut, gut, ausreichend, sehr gut bis gut erzielt worden, so erhält man die geordnete Reihe

sehr gut, sehr gut bis gut, gut, befriedigend, ausreichend, ungenügend.

Die Rangziffern lauten dann:

1 für sehr gut
2 für sehr gut bis gut
3 für gut
4 für befriedigend
5 für ausreichend
6 für ungenügend

Spearmannscher Rangkorrelationskoeffizient

Zusammenfassung:
Die Merkmalspaare gehen damit bei einem zweidimensionalen Merkmal (gebildet aus Rangmerkmalen) in Rangzifferpaare über.
Aus den Rangzifferpaaren kann dann der Korrelationskoeffizient bestimmt und damit festgestellt werden, ob eine Beziehung zwischen den Rangfolgen der statistischen Einheiten bzgl. der beiden Merkmale besteht.

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.5
Die sechs oben angeführten Kandidaten haben (in derselben Reihenfolge wie in der geordneten Urliste) in einem weiteren Fach die Bewertungen

$$\underbrace{\textit{gut}, \textit{gut}}_{2;3}, \underbrace{\textit{sehr gut bis gut}}_{1}, \underbrace{\textit{befriedigend}, \textit{befriedigend}}_{4;5}, \underbrace{\textit{ausreichend}}_{6}$$

erhalten.
Damit erhalten sie hier die Rangziffern[1]

$$2.5,\ 2.5,\ 1,\ 4.5,\ 4.5,\ 6.$$

[1] Sind zwei oder mehr Merkmalswerte gleich, so ordnet man ihnen das arithmetische Mittel der zur Verfügung stehenden Rangziffern zu.
Beispiel: Liegen $x_3 = x_7$ auf den Plätzen 3 und 4, so ist $r_3 = r_7 = 3.5$.

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.5
Die sechs oben angeführten Kandidaten haben (in derselben Reihenfolge wie in der geordneten Urliste) in einem weiteren Fach die Bewertungen

$$\underbrace{\textit{gut}, \textit{gut}}_{2;3}, \underbrace{\textit{sehr gut bis gut}}_{1}, \underbrace{\textit{befriedigend}, \textit{befriedigend}}_{4;5}, \underbrace{\textit{ausreichend}}_{6}$$

erhalten.
Damit erhalten sie hier die Rangziffern[1]

2.5, 2.5, 1, 4.5, 4.5, 6.

[1] Sind zwei oder mehr Merkmalswerte gleich, so ordnet man ihnen das arithmetische Mittel der zur Verfügung stehenden Rangziffern zu.
Beispiel: Liegen $x_3 = x_7$ auf den Plätzen 3 und 4, so ist $r_3 = r_7 = 3.5$.

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.5

Rangzifferpaare sind also

(1, 2.5) (2, 2.5) (3, 1) (4, 4.5) (5, 4.5) (6, 6).

Spearmannscher Rangkorrelationskoeffizient

Allgemein

Seien

$$(x_1, y_1), \ldots, (x_n, y_n)$$

die Merkmalspaare und

$$(r_1, s_1), \ldots, (r_n, s_n)$$

die zugehörigen Rangzifferpaare, so erhält man als Korrelationskoeffizienten

$$r_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})^2}}.$$

Spearmannscher Rangkorrelationskoeffizient

Allgemein

Seien

$$(x_1, y_1), \ldots, (x_n, y_n)$$

die Merkmalspaare und

$$(r_1, s_1), \ldots, (r_n, s_n)$$

die zugehörigen Rangzifferpaare, so erhält man als Korrelationskoeffizienten

$$r_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})^2}}.$$

Spearmannscher Rangkorrelationskoeffizient

Falls kein Merkmalswert doppelt auftritt und damit die Rangziffern jeweils die Werte $1, \ldots, n$ durchlaufen, lässt sich r_S umformen zu

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (r_i - s_i)^2}{n(n^2 - 1)}.$$

r_S heißt Spearmanscher Rangkorrelationskoeffizient.

Spearmannscher Rangkorrelationskoeffizient

Analog zum Korrelationskoeffizienten gilt:

Rangfolgen bei beiden Merkmalen	r_S
gegenläufig	-1
unkorreliert	0
identisch	$+1$

Entsprechend wird man Zwischenwerte interpretieren:

$r_S \approx -1$ gegenläufiger Trend in den Merkmalen,
$r_S \approx 0$ kein Trend erkennbar,
$r_S \approx 1$ gleichlaufender Trend.

Spearmannscher Rangkorrelationskoeffizient

Analog zum Korrelationskoeffizienten gilt:

Rangfolgen bei beiden Merkmalen	r_S
gegenläufig	-1
unkorreliert	0
identisch	$+1$

Entsprechend wird man Zwischenwerte interpretieren:

$r_S \approx -1$ gegenläufiger Trend in den Merkmalen,
$r_S \approx 0$ kein Trend erkennbar,
$r_S \approx 1$ gleichlaufender Trend.

Spearmannscher Rangkorrelationskoeffizient

Rangziffern bei quantitativen Merkmalen
Dies ist dann sinnvoll, wenn zwar ein Trend zwischen den Werten, aber kein funktionaler Zusammenhang vermutet wird und daher insbesondere eine lineare Regression nicht angebracht ist.

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.6

8 Studenten der Statistik-Vorlesung wurden befragt, wieviele Stunden sie für die Nacharbeitung der Vorlesung im Durchschnitt wöchentlich aufgewandt haben und welche Punktzahlen sie in der Klausur erreichten:

	Student							
	1	2	3	4	5	6	7	8
Stunden	0	1.5	3	3	4	4.5	5	2
Punkte	25	15	30	35	50	45	55	30

Daraus ergeben sich die Rangzifferpaare (r_i, s_i):

(1,2), (2,1), (3,3.5), (4.5,3.5), (4.5,5), (6,7), (7,6), (8,8)

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.6

8 Studenten der Statistik-Vorlesung wurden befragt, wieviele Stunden sie für die Nacharbeitung der Vorlesung im Durchschnitt wöchentlich aufgewandt haben und welche Punktzahlen sie in der Klausur erreichten:

	Student							
	1	2	3	4	5	6	7	8
Stunden	0	1.5	3	3	4	4.5	5	2
Punkte	25	15	30	35	50	45	55	30

Daraus ergeben sich die Rangzifferpaare (r_i, s_i):

(1,2), (2,1), (3,3.5), (4.5,3.5), (4.5,5), (6,7), (7,6), (8,8)

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.6

Nach der ersten Formel für r_S ergibt sich

$$r_S = \frac{Cov(r,s)}{\sqrt{s_r^2 \cdot s_s^2}}$$

$$= \frac{\frac{200.75}{8} - 20.25}{\sqrt{5.1875 \cdot 5.1875}} = \frac{4.84375}{5.1875} = 0.934$$

Die zweite Formel darf hier wegen bestehender Bindungen nicht verwendet werden und liefert ein abweichendes Ergebnis

$$r_S = 1 - \frac{6}{8 \cdot 63}(1^2 + 1^2 + 0.5^2 + 1^2 + 0.5^2 + 1^2 + 1^2 + 0^2)$$

$$= 1 - \frac{6}{8 \cdot 63} \cdot 5.5 = 0.935.$$

Spearmannscher Rangkorrelationskoeffizient

Beispiel 11.6

Nach der ersten Formel für r_S ergibt sich

$$\begin{aligned} r_S &= \frac{Cov(r,s)}{\sqrt{s_r^2 \cdot s_s^2}} \\ &= \frac{\frac{200.75}{8} - 20.25}{\sqrt{5.1875 \cdot 5.1875}} = \frac{4.84375}{5.1875} = 0.934 \end{aligned}$$

Die zweite Formel darf hier wegen bestehender Bindungen nicht verwendet werden und liefert ein abweichendes Ergebnis

$$\begin{aligned} r_S &= 1 - \frac{6}{8 \cdot 63}(1^2 + 1^2 + 0.5^2 + 1^2 + 0.5^2 + 1^2 + 1^2 + 0^2) \\ &= 1 - \frac{6}{8 \cdot 63} \cdot 5.5 = 0.935. \end{aligned}$$