2 Zusammenhänge untersuchen

2.1 Kovarianz und Korrelation

Wir kommen nun zu den bivariaten statistischen Kenngrößen. Die Bezeichnung „bivariat“ bezieht sich auf die Verwendung von zwei und „multivariat“ allgemein auf die Verwendung von mehreren Variablen. Wir wollen empirische Maßzahlen berechnen, die uns einen Hinweis darauf geben, in wie weit zwei Variablen korrelieren, das heißt, ob es einen linearen Zusammenhang zwischen ihnen gibt. Entsprechend der Varianz einer einzelnen Variablen als Maß für ihre Streuung, gibt es die Kovarianz zwischen zwei Variablen, mit der man eine Aussage darüber machen kann, wie stark die Abhängigkeit der beiden Variablen ist. Bei normalverteilten Daten gilt nämlich, dass bei verschwindender (theoretischer, nicht empirischer bzw. geschätzter) Kovarianz auf die Unabhängigkeit der Variablen geschlossen werden kann. Den Betrag der Kovarianz kann man allerdings schlecht interpretieren, da dieser noch von der Streuung der beiden Variablen abhängt. In diesem Fall verwendet man den Korrelationskoeffizienten. Man kann sagen, dass bei einer positiven theoretischen Kovarianz ein positiver Zusammenhang zwischen den beiden Variablen in dem Sinn besteht, dass mit dem Anstieg der Werte der einen Variable, auch ein Anstieg der Werte der anderen Variablen „zu erwarten“ ist (hier ist dann der Steigungsparameter der Regressionsgerade - die den linearen Zusammenhang beschreibt - positiv). Analoges gilt für eine negative theoretische Kovarianz.

Da wir immer nur die empirische Kovarianz bzw. später den empirischen Korrelationskoeffizient aus den Daten erhalten, sollte man eine vermutete Korrelation mit einem Test absichern. Aussagen anhand empirischer Größen werden umso sicherer, je größer der Stichprobenumfang ist (was allgemein für erwartungstreue und konsistente Schätzer von theoretischen Kenngrößen gilt). Ansonsten

kann man im Rahmen der schließenden Statistik einen Test auf Korrelation durchführen, wie wir ihn im zweiten Abschnitt dieses Kapitels zeigen.

Wir möchten im Folgenden die Kovarianz und die Korrelationen schätzen (bzw. die empirischen Entsprechungen berechnen) und danach mit diesem einen Test auf Korrelation durchführen.

Hier sind unsere Daten:

v1	v2
175	79
178	81
177	80
181	84
185	83
183	90

Den Output erhalten Sie, wenn Sie auf den Button **→Zusammenhänge untersuchen** klicken**,** wobei Sie zuvor die Variablen v1 und v2 unter diesem Menüpunkt auswählen müssen. Danach können Sie **→Pearsonscher Korrelationskoeffizient** wählen.

Stichprobenumfang	6
arithmetisches Mittel Spalte 1	179.83333333333
arithmetisches Mittel Spalte 2	82.833333333333
empirische Varianz Spalte 1	14.566666666669
empirische Varianz Spalte 2	15.766666666667
empirische Kovarianz	10.966666666666
Pearsonscher Korrelationskoeffizient	0.72364340603714
Prüfgröße ('t-Value') nach R. A. Fisher für Test (H0: corr=0; gegen H1: corr<>0)	2.0969881577483
p-Wert	0.104

Im Folgenden verwenden wir $x_1, x_2, \ldots, x_n$ für die Beobachtungen der ersten Stichprobe und $y_1, y_2, \ldots, y_n$ für die Beobachtungen der zweiten Stichprobe.

Es werden nun berechnet: Der Stichprobenumfang n, der Mittelwert der ersten und zweiten Spalte $\bar{x}$ und $\bar{y}$ und die empirischen Varianzen s_x^2 und s_y^2.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 179{,}833\ldots$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 82{,}833\ldots$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 14{,}566\ldots$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 15{,}766..$$

Es folgt die Berechnung der empirischen Kovarianz s_{xy} und des empirischen Korrelationskoeffizienten r_{xy} (bzw. des Pearsonschen Korrelationskoeffizienten).

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 10{,}966\ldots$$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = 0{,}7236\ldots$$

Test auf Korrelation bei Normalverteilung:
Unter der Voraussetzung, dass die beiden Stichproben $x_1, x_2, \ldots, x_n$ und $y_1, y_2, \ldots, y_n$ jeweils aus einer normalverteilten Grundgesamtheit stammen (wir gehen davon aus, dass die $x_1, x_2, \ldots, x_n$ Realisierungen von unabhängig und identisch normalverteilten Zufallvariablen $X_1, X_2, \ldots, X_n$ sind und dies analog für die zweite Stichprobe gilt) berechnen wir die Prüfgröße als Realisierung (unter H_0) der mit n - 2 Freiheitsgraden t-verteilten Prüfgröße und führen einen Test durch mit:

H_0: Die Korrelation zwischen zwei Variablen ist gleich Null (d.h. der Korrelationskoeffizient $\rho = 0$)

gegen

H_1: Die Korrelation zwischen zwei Variablen ist von Null verschieden (d.h. $\rho \neq 0$)

Nun berechnen wir die Prüfgröße t als Realisierung einer (unter H_0), wie beschrieben, mit n - 2 Freiheitsgraden t-verteilten Zufallsvariablen. Danach berechnen wir mit dieser den p-Wert.

$$t = r_{xy} \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}} = 2{,}096988\ldots$$

H_0 kann auf einem Signifikanzniveau von α verworfen werden, wenn

$$|t| \geq F_{t_{n-2}}^{-1}(1-\alpha/2)\,,$$

bzw., falls

$$F_{t_{n-2}}(|t|) \geq 1-\alpha/2 \Leftrightarrow \underbrace{2(1-F_{t_{n-2}}(|t|))}_{:=\text{p-Wert}} \leq \alpha \quad .$$

Dabei ist $F_{t_{n-2}}$ die Verteilungsfunktion der t-Verteilung mit n - 2 Freiheitsgraden.

Im Beispiel hat der p-Wert einen Wert von 0,1040. Auf einem Signifikanzniveau von 5% kann man die Nullhypothese, dass die Korrelation gleich Null ist, nicht verwerfen (da p-Wert > 0,05). Man kann somit keinen Zusammenhang zwischen den beiden Variablen nachweisen. Wir weisen an dieser Stelle nochmals darauf hin, dass die Normalverteilungsvoraussetzungen erfüllt sein müssen. Falls diese nicht erfüllt sind, so kann ein nichtparametrisches Verfahren angewandt werden (Rangkorrelation nach Spearman, siehe nächstes Kapitel).

Umsetzung mit SAS:

```
data dat1;
input x y;
datalines;
175      79
178      81
177      80
181      84
185      83
183      90
run;

proc print data=dat1;

proc corr data = dat1 cov;
   var x y;
run;
```

SAS-Output zur Prozedur CORR:

Die Prozedure CORR

2 Variablen:	x y

Kovarianzmatrix, DF = 5		
	x	**y**
x	14.56666667	10.96666667
y	10.96666667	15.76666667

Einfache Statistiken						
Variable	**N**	**Mittelwert**	**Std.abweichung**	**Summe**	**Minimum**	**Maximum**
x	6	179.83333	3.81663	1079	175.00000	185.00000
y	6	82.83333	3.97073	497.00000	79.00000	90.00000

Pearsonsche Korrelationskoeffizienten, N = 6 Prob > |r| unter H0: Rho=0		
	x	**y**
x	1.00000	0.72364 0.1040
y	0.72364 0.1040	1.00000