Korrelationsanalyse

Inhaltsverzeichnis

Korrelationsanalyse

Lernhinweise

Benötigte Vorkenntnisse
Für diesen Lernschritt sollten Sie wissen, was eine Zufallsauswahl ist, wie Stichproben zu Stande kommen, welche Bedeutung Messniveaus haben, was abhängige und unabhängige Variablen und was es mit Freiheitsgraden auf sich hat.

Lernziele
Sie lernen anhand von Übungen und Beispielen, wie Sie bestimmen können, ob bestimmte Variablen korrelieren. Dazu sind folgende Schritte notwendig:

- Verständnis linearer Beziehungen;
- Verständnis unterschiedlicher Datenverteilungen;
- Berechnung des Korrelationskoeffizienten;
- Bestimmung des sog. kritischen Wertes von r auf einem gewählten Signifikanzniveau.

Einführung

Die Korrelationsanalyse untersucht, ob ein Zusammenhang zwischen zwei metrischen Variablen besteht, wie stark dieser Zusammenhang ist und welche "Richtung" er hat.

Beispiel:
Besteht ein statistisch signifikanter Zusammenhang zwischen dem Wohlstand eines Staates und der Immigrationsrate?
Zur Bestimmung des Zusammenhangs wird der Pearson’sche Korrelationskoeffizient r berechnet.

Theorie (1-8)

Inhaltsübersicht:

1. Produktmoment-Korrelationskoeffizient

Anhand des Korrelationskoeffizienten lässt sich bestimmen:

- ob eine Beziehung zwischen zwei metrischen Variablen (Messniveau Intervall- oder Ratio-Skala) besteht;
- wie eng diese Beziehung ist;
- welche Richtung diese Beziehung hat.

Der Einfachheit halber gehen wir von einer linearen Beziehung aus.
Der Korrelationskoeffizient (genau: der Pearson'sche Produktmoment-Korrelationskoeffizient) wird mit r bezeichnet. Die Statistik r kann Werte zwischen +1 und -1 annehmen.

- Werte in der Nähe von +1 deuten auf einen engen "positiven" Zusammenhang (je grösser x, desto grösser y);
- Werte in der Nähe von -1 deuten auf einen engen "negativen" Zusammenhang (je grösser x, desto kleiner y);
- Kleine Werte von r deuten auf eine geringe oder keine Beziehung zwischen den Variablen.

Im nachfolgenden Streudiagramm werden verschiedene mehr oder weniger eng korrelierte Daten aus den Mitgliedsstaaten der EU - sowie einige Musterdaten - dargestellt.
Schauen Sie sich auf dem nachfolgenden Diagramm die unterschiedlichen Streuungen an und achten Sie auf die Veränderung von r.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

2. Verteilung von Daten

Der Korrelationskoeffizient r ist absolut umso grösser (liegt also umso näher bei +1 oder bei -1), je dichter die Datenpunkte an einer Diagonalen liegen.
Auf der nachfolgenden Tabelle können Sie die Datenpunkte aus dem vorherigen Beispiel manipulieren:

- Verschieben Sie einzelne Datenpunkte und beobachten Sie, wie sich der Wert von r verändert.
- Schieben Sie alle Punkte auf eine Diagonale und sehen Sie, wie r gegen 1 bzw. -1 konvergiert.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

3. Berechnung des Produktmoment-Korrelationskoeffizienten r

Nachdem Sie sich mit dem Verhalten mehr oder weniger korrelierender (also mehr oder weniger zusammenhängender) Variablen angefreundet haben, sollten Sie nun lernen, wie der Produktmoment-Korrelationskoeffizient berechnet wird.
Der Einfachheit halber stellen wir Ihnen hier allerdings nur 2 von vielen Berechnungsmöglichkeiten vor. Dabei wird der Korrelations-Koeffizient r formell wie folgt definiert:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \overline{X})^2)(\sum_{i=1}^{n}(Y_i - \overline{Y})^2)}}$$

Wie liest sich diese Formel?

- X und Y sind die Werte der beiden Variablen.
- i=1... n indiziert die Anzahl der Fälle für X und Y.
- $\overline{X}$
  (X-quer) und
  $\overline{Y}$
  (Y-quer) sind die Mittelwerte (d.h. das arithmetische Mittel) der beiden Variablen.
- Im **Zähler** der Formel werden:
  1.) die Abweichungen der Werte von X bzw. Y von ihrem jeweiligen Mittel für jeden Fall gebildet,
  2.) miteinander multipliziert und
  3.) dann über alle Fälle aufsummiert.
- Im **Nenner** der Formel werden die schon genannten Abweichungen der Variablenwerte von ihren Mittelwerten zuerst
  1.) quadriert, dann
  2.) über alle Fälle aufsummiert,
  3.) die Resultate miteinander multipliziert und
  4.) daraus die Wurzel gezogen.

4. Vereinfachung der Formel

Auf den ersten Blick scheint diese Formel kompliziert. Sie lässt sich jedoch stark vereinfachen. Nochmals zur Erinnerung die ausführliche Schreibweise:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \overline{X})^2)(\sum_{i=1}^{n}(Y_i - \overline{Y})^2)}}$$

Wenn wir davon ausgehen, dass immer über alle Fälle (i=1... n) summiert wird, können die Indizes aus der Berechnung weggelassen werden. Wir berechnen dann mit

- $x = (X - \overline{X})$
  die Abweichung aller X-Werte vom Mittel.
- $y = (Y - \overline{Y})$
  die Abweichung aller Y-Werte vom Mittel.

Damit kann die Formel folgendermassen vereinfacht werden:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Oder in Worten: Die Summe der Abweichungs-Produkte (**SP**, sum of products in der engl. Terminologie) wird dividiert durch die Wurzel aus dem Produkt der aufsummierten Abweichungsquadrate (**SQx** bzw.**SQy**, sum of squares).
Die Formel lässt sich also noch einmal vereinfachen:

$$r = \frac{SP}{\sqrt{SQ_x SQ_y}}$$

Abgekürzt: r ist das Verhältnis der Produktsummen zum geometrischen Mittel der Quaudratsummen.
Aufgabe: Versuchen Sie, ausgehend von

$$r = SP / \sqrt{SQ_x SQ_y}$$

die ursprüngliche, ausführliche Formel mit Papier und Bleistift zu rekonstruieren. Auf diese Weise wird es Ihnen in Zukunft leichter fallen, sich an das Berechnungsprinzip zu erinnern.
Hinweis: Eine ausführliche Definition des Korrelationskoeffizienten finden Sie in Bohley (2000), S. 232-254.

5. Berechnungsbeispiel

Nachdem Sie sich mit der Formel vertraut gemacht haben, werden Sie sie nun anwenden. Sie werden mit Hilfe des Korrelationskoeffizienten die Frage beantworten: Besteht ein Zusammenhang zwischen Immigration und Wohlstand?
Hypothese: Je höher der Wohlstand in einem Land, umso mehr Menschen werden einwandern. Erwartet wird eine positive Korrelation (r > 0).
Operationalisiert werden die beiden Variablen wie folgt:

- X: BSP pro Kopf 2000 (in 1000 USD)
- Y: geschätzte Netto-Zuwanderungsrate 2001 (Anz. Migranten pro 1000 Einwohner)

Folgende Datenpaare (X/Y) sind gegeben:*

Österreich	(AUT) 25.0/2.45	Griechenland	(GRE) 17.2/1.96
Belgien	(BEL) 25.3/0.97	Irland	(IRL) 21.6/4.69
Dänemark	(DAN) 25.5/1.98	Italien	(ITA) 22.1/1.73
Spanien	(ESP) 18.0/0.87	Luxemburg	(LUX) 36.4/9.26
Finnland	(FIN) 22.9/0.61	Niederlande	(NDL) 24.4/2.34
Frankreich	(FRA) 24.4/0.64	Portugal	(POR) 15.8/0.50
Grossbritannien	(GBR) 22.8/1.07	Schweden	(SWE) 22.2/0.91
Deutschland	(GER) 23.4/4.00		

Um die Frage nach dem Zusammenhang zwischen Wohlstand und Migration zu beantworten, müssen Sie die Daten zuerst in einer Tabelle erfassen. Nur so können Sie sich ein Bild von der Streung der Daten machen und verzerrende Elemente wie z.B. Ausreisser vorzeitig erkennen.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Ermitteln Sie anschliessend Schritt für Schritt die Werte, die für die Berechnung von r benötigt werden, und tragen Sie diese in die Tabelle ein:

1. Berechnen Sie die Summen und die Mittel;
2. Berechnen Sie x, d.h. die Abweichungen der X-Werte von ihrem Mittel;
3. Berechnen Sie y, d.h. die Abweichungen der Y-Werte von ihrem Mittel;
4. Berechnen Sie xy, d.h. die Produkte der Abweichungen x und y.

Zur Erinnerung: Die Formel von r

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \overline{X})^2)(\sum_{i=1}^{n}(Y_i - \overline{Y})^2)}}$$

Wie liest sich diese Formel?

- X und Y sind die Werte der beiden Variablen.
- i=1... n indiziert die Anzahl der Fälle für X und Y.
- $\overline{X}$
  (X-quer) und
  $\overline{Y}$

(Y-quer) sind die Mittelwerte (d.h. das arithmetische Mittel) der beiden Variablen.

- Im **Zähler** der Formel werden:
  1.) die Abweichungen der Werte von X bzw. Y von ihrem jeweiligen Mittel für jeden Fall gebildet,
  2.) miteinander multipliziert und
  3.) dann über alle Fälle aufsummiert.
- Im **Nenner** der Formel werden die schon genannten Abweichungen der Variablenwerte von ihren Mittelwerten zuerst
  1.) quadriert, dann
  2.) über alle Fälle aufsummiert,
  3.) die Resultate miteinander multipliziert und
  4.) daraus die Wurzel gezogen.

Wenn wir davon ausgehen, dass immer über alle Fälle (i=1... n) summiert wird, können die Indizes aus der Berechnung weggelassen werden. Wir berechnen dann mit

- $x = (X - \overline{X})$
  die Abweichung aller X-Werte vom Mittel.
- $y = (Y - \overline{Y})$
  die Abweichung aller Y-Werte vom Mittel.

Damit kann die Formel folgendermassen vereinfacht werden:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Oder in Worten: Die Summe der Abweichungs-Produkte (**SP**, sum of products in der engl. Terminologie) wird dividiert durch die Wurzel aus dem Produkt der aufsummierten Abweichungsquadrate (**SQx** bzw.**SQy**, sum of squares).
Die Formel lässt sich also noch einmal vereinfachen:

$$r = \frac{SP}{\sqrt{SQ_x SQ_y}}$$

Abgekürzt: r ist das Verhältnis der Produktsummen zum geometrischen Mittel der Quaudratsummen.

* Quelle der Daten ist das World Factbook 2001 der CIA.

6. Perspektivenwechsel

Schauen Sie sich die Berechnung des Korrelationskoeffizienten nochmals aus umgekehrter Perspektive an. Beachten Sie dabei, wie schnell sich die einfachen Berechnungen hinter den "beispiellosen" Zahlen verändern, wenn sie die Datenpunkte in der Ansicht "Diagramm" verschieben.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

7. Aufgabe

Suchen Sie im Internet oder in der (aktuellen) Literatur Eckdaten zu den Staaten der EU-15 und untersuchen Sie diese **theoriegeleitet** auf Zusammenhänge die in den letzten Monaten in der Tagespresse diskutiert wurden (Halten Sie sich dabei an die Aufgabenstellung zum Lernschritt).

1. Geben Sie die Werte von X und Y ein und berechnen Sie (auf Papier) x, x2, y2 und xy.
2. Kontrollieren Sie Ihre Berechnungen (klicken sie auf "Kontrolliere Berechnung";
3. Füllen Sie die Tabelle vollständig aus (Beschriftung der Achsen, Quelle, ihren Namen und ihre Matrikelnummer) und drucken Sie sie aus.
4. Berechnen Sie r anhand der Formel;

$$r = \frac{SP}{\sqrt{SQ_x SQ_y}}$$

5. **Überprüfen** und notieren Sie Ihr Resultat und gehen Sie weiter zur nächsten Seite.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

8. Signifikanz

Sie haben Ihre Daten nach Zusammenhängen untersucht und einen bestimmten Wert für r berechnet. Was bedeutet dieser Wert nun aber? Ist 0.82 ein "guter" Wert? Oder wäre 0.62 besser? Oder -0.42?
Um das zu entscheiden, müssen Sie zuerst in der Lage sein, die Signifikanz von r zu bestimmen.

Grundüberlegung:
Korrelationskoeffizienten werden üblicherweise nicht für eine Grundgesamtheit von Fällen berechnet, sondern nur für eine Stichprobe. Je grösser die Anzahl Fälle n der Stichprobe, umso näher kommt der berechnete Korrelationskoeffizient r dem "echten" Korrelationskoeffizienten (rho), der aber natürlich nicht bekannt ist.
Da eine Stichprobe nur begrenzt aussagekräftig ist stellt sich die Frage nach der Verlässlichkeit des von ihnen berechneten Korrelationskoeffizienten: beschreibt er die tatsächliche Korrelation zwischen den untersuchten Variablen in der Grundgesamtheit, oder muss man davon ausgehen, dass er zufällige Zusammenhänge einer Stichprobe aufzeigt?
Zur Beantwortung dieser Frage wird überprüft, ob der berechnete Korrelationskoeffizient signifikant ist, d.h. ob man die Null-Hypothese: „es besteht keine Korrelation“ im vorliegenden Fall verwerfen kann oder nicht.

Anmerkung:
Wir gehen von einem sog. einseitigen Test aus, d.h. uns interessiert nur das Risiko in einer Richtung (zu geringe Korrelation).

Vorgehen:

- Zunächst wird das sog. Signifikanzniveau festgelegt. Dabei handelt es sich um die Irrtumswahrscheinlichkeit, die man bei der Rückweisung der Null-Hypothese „keine Korrelation“ zu akzeptieren gewillt ist. Üblich sind 5%, 1%, 0.1% oder 0.01%. Bei einer kleinen Anzahl von Fällen n wird man z.B. 1% wählen. Die Prozentwerte werden üblicherweise als Wahrscheinlichkeiten p (Teile von 1) angegeben. 1% entspricht z.B. p=0.01.
- Danach wird mit der Teststatistik F die Varianz der Stichproben berechnet. Der F-Test ist definiert als:

$$F_{1,n-2} = \frac{r^2}{(1-r^2)}(n-2)$$

Im Index von F stehen neben den Freiheitsgraden (degrees of freedom, df) der Variablen df1 (in unserem Fall 1) die Freiheitsgrade df2 der n-2 Fälle. Anhand der Freiheitsgrade kann in einer entsprechenden statistischen Tafel überprüft werden, welcher Wert von F erreicht werden muss, damit der zugehörige r-Wert bei der gewählten Irrtumswahrscheinlichkeit als Signifikant gelten kann.

Beispiel:
Bei 13 Freiheitsgraden (entspricht n=15) muss der F-Wert auf dem Signifikanzniveau von 1% (entspricht p=0.01) F=9.07 erreichen, damit die Null-Hypothese verworfen werden kann. Bei F-Werten, die kleiner als 9.07 sind, kann die Null-Hypothese nicht mit ausreichender Sicherheit verworfen werden. Man muss dann also davon ausgehen, dass zwischen den Variablen kein (signifikanter) Zusammenhang besteht.
Klicken Sie hier für einen Auszug aus der Tabelle mit F-Werten.

SPSS-Kochbuch

Korrelationsanalyse in SPSS *.pdf, 36 KB

Übungsdaten:

- im SPSS-eigenen *.sav Format, ready to use, 1 KB
- als Excel Tabelle, die umkodiert werden muss, 14 KB