Quantitative Methoden

Inferenzstatistik – Korrelation
13.06.2008

Prof. Dr. Walter Hussy und David Tobinski
UDE.EDUcation College
im Rahmen des dokFORUMs
Universität Duisburg-Essen

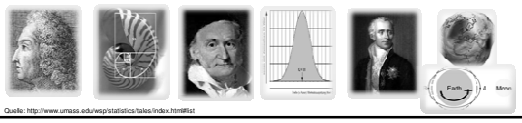
Merkmalszusammenhänge
Eine kurze Geschichte der Statistik

Systematische Datensammlungen zu Bevölkerung und Wirtschaft wurden erstmals in der Renaissance in den italienischen Stadtstaaten Venedig und Florenz zusammengestellt.

Der Begriff „Statistik“, abgeleitet von dem lateinischen *status* „Stand, Verfassung“, beschrieb damals eine Datensammlung, die für den Staat von Interesse war. Die Idee, solche Daten zu sammeln, bereitet sich von Italien über die anderen Länder Westeuropas aus.

Merkmalszusammenhänge
Eine kurze Geschichte der Statistik

Obwohl die Wahrscheinlichkeitsrechnung von solch berühmten Mathematikern wie Jacob Bernoulli, Carl Friedrich Gauß und Pierre-Simon Laplace entwickelt wurde, spielte sie im ganzen 19.Jahrhundert für die Untersuchung statistischer Ergebnisse praktisch keine Rolle, da die meisten sozialen Statistiker zu dieser Zeit vorzogen, ihre Daten für sich selbst sprechen zu lassen.



Quelle: http://www.umass.edu/wsp/statistics/tales/index.html#list

Merkmalszusammenhänge
Eine kurze Geschichte der Statistik

Erst im späte 19. Jahrhundert befasste sich die Statistik damit auch Schlüsse aus ihren numerischen Daten zu ziehen. Dies begann mit den Arbeiten von Francis Galton (1822-1911) zur Vererbung, in denen er Verfahren einsetzte, die wir heute als Regressions- und Korrelationsanalyse bezeichnen würden.

Weitere Impulse stammten aus den Arbeiten von Karl Pearson (1857-1936). Er war der erste Direktor des Galton Laboratory, das Francis Galton 1904 eingerichtet hatte. Einer seiner ersten Forschungsgäste war der Chemiker William Sealey Gosset (1876-1937), der seine Arbeiten unter dem Pseudonym „Student“ veröffentlichte.



Merkmalszusammenhänge
Einführung

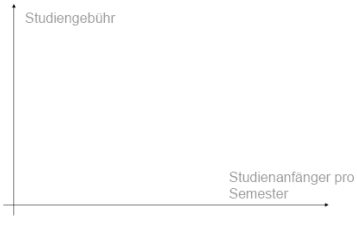
Die meisten Hypothesen über einen empirischen Sachverhalt beinhalten offen oder verdeckt formulierte Annahmen über Kausalbeziehungen. Das Aufdecken solcher Kausalzusammenhänge erlaubt uns, über das bloße Beschreiben der phänomenologischen Umwelt hinauszugehen und Erklärungen für empirische Sacherverhalte anzubieten.

Die Kenntnis von Zusammenhängen ermöglicht überdies Vorhersagen über künftige Ereignisse.

„Wenn – Dann“ „Je – Desto“

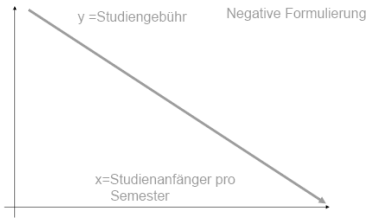
Merkmalszusammenhänge
Einführung

Hypothesen: Je-desto:



Merkmalszusammenhänge
Einführung

Hypothesen: Je-desto:



Merkmalszusammenhänge
Kovarianz und Korrelation

Der Grad des Zusammenhangs zwischen zwei intervallskalierten Variablen lässt sich mathematisch durch die Kovarianz und die auf ihr aufbauende Produkt-Moment-Korrelation beschreiben.

Merkmalszusammenhänge
Der Begriff des Zusammenhangs

Ein Zusammenhang kann in zwei „Richtungen“ vorliegen: positiv oder negativ.
Wenn hohe Werte auf der einen Variable hohen Werten auf der anderen entsprechen und niedrige Werte auf der einen Variable niedrigen auf der anderen, so ist der Zusammenhang positiv.

Gehen dagegen hohe Werte auf der einen Variable mit niedrigen Werten auf der anderen einher und umgekehrt, so liegt ein negativer Zusammenhang vor.

Merkmalszusammenhänge
Die Kovarianz

Die Kovarianz ist das durchschnittliche Produkt aller korrespondierenden Abweichungen der Messwerte von den Mittelwerten der beiden Merkmale x und y.

Die folgende Formel zeigt, dass die Kovarianz im Gegensatz zur Varianz Aussagen über die gemeinsame Variation zweier Merkmale macht:

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

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Merkmalszusammenhänge
Die Kovarianz

Für jedes Wertepaar wird die Abweichung des x-Werts vom Mittelwert der x-Werte mit der Abweichung des y-Werts vom Mittelwert der y-Werte multipliziert.

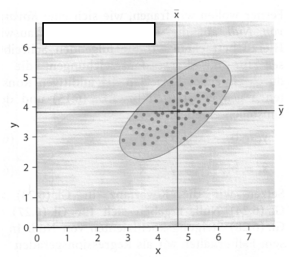
Die Summe der einzelnen Abweichungsprodukte wird als Kreuzproduktsumme zweier Variablen bezeichnet.

Diese Kreuzproduktsumme wird über alle Beobachtungen gemittelt.

Allerdings wird analog zur Varianz im Nenner durch n-1 geteilt, um einen erwartungstreuen Schätzer der Populationskovariation zu erhalten.

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

Merkmalszusammenhänge
Die Kovarianz



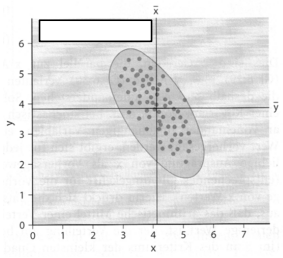
$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Merkmalszusammenhänge
Die Kovarianz

Eine positive Kovarianz resultiert, wenn die beiden Variablen weitgehend gemeinsam in die gleiche Richtung von ihrem Mittelwert abweichen, d.h. positive Abweichungen der einen Variable werden mit positiven Abweichungen der anderen multipliziert, bzw. negative mit negativen. Der Zusammenhang ist positiv.

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

Merkmalszusammenhänge
Die Kovarianz



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

Merkmalszusammenhänge
Die Kovarianz

Dagegen ergibt sich eine negative Kovarianz, wenn viele entgegengesetzt gerichtete Abweichungen vom jeweiligen Mittelwert auftreten, d.h. eine positive Abweichung auf der einen Variable korrespondiert mit einer negativen Abweichung auf der anderen und umgekehrt.
Die Kreuzproduktsumme und somit auch die Kovarianz werden negativ. Die Merkmale weisen einen negativen Zusammenhang auf.

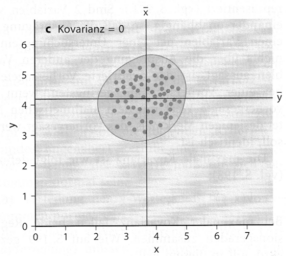
$$\mathrm{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n-1}$$

Merkmalszusammenhänge
Die Kovarianz

Sind die Abweichungen mal gleich, mal entgegengesetzt gerichtet, so heben sich die Abweichungsprodukte gegenseitig auf und es resultiert eine Kovarianz nahe Null.
In diesem Fall besteht kein systematischer Zusammenhang zwischen den Variablen x und y. Die Ausprägung des Merkmals x sagt also nichts über die Ausprägung des Merkmals y aus.

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

Merkmalszusammenhänge
Die Kovarianz



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

Merkmalszusammenhänge
Die Kovarianz

Der Betrag der maximalen Kovarianz ist für positive wie auch negative Zusammenhänge identisch. Er ist definiert als das Produkt der beiden Merkmalsstreuungen:

$$\left|\text{cov}(\max) = \hat{\sigma}_x \cdot \hat{\sigma}_y\right|$$

Die Kovarianz ist also kein standardisiertes Maß und folglich zur quantitativen Kennzeichnung des Zusammenhangs zweier Merkmale nur bedingt geeignet. Sie kann allerdings in ein standardisiertes Maß überführt werden: den Korrelationskoeffizienten.

Merkmalszusammenhänge
Die Produkt-Moment-Korrelation

Die Produkt-Moment-Korrelation nach Pearson ist das gebräuchlichste Maß für die Stärke des Zusammenhangs zweier Variablen. Sie drückt sich aus im Korrelationskoeffizienten r.

Es stellt die Standardisierung der im vorherigen Abschnitt behandelten Kovarianz dar. Dabei wird die empirisch ermittelte Kovarianz an der maximalen Kovarianz relativiert.

$$r_{xy} = \frac{\mathrm{cov}_{emp}}{\mathrm{cov}_{\max}} = \frac{\mathrm{cov}(x, y)}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

Merkmalszusammenhänge
Die Produkt-Moment-Korrelation

Die Formel gibt zu erkennen, dass der Korrelationskoeffizient niemals größer als 1 oder kleiner als -1 werden kann, denn die empirisch gefundene Kovarianz kann die maximal mögliche Kovarianz zwischen den beiden Variablen in ihrem Wert nicht übersteigen. Der Wertebereich der Korrelation ist somit im Gegensatz zu dem der Kovarianz begrenzt zwischen -1 und +1.

$$r_{xy} = \frac{\mathrm{cov}_{emp}}{\mathrm{cov}_{\max}} = \frac{\mathrm{cov}(x, y)}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

Merkmalszusammenhänge
Die Produkt-Moment-Korrelation

Eine Umwandlung der Formel der Korrelation ist sehr aufschlussreich:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot \hat{\sigma}_x \cdot \hat{\sigma}_y} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\hat{\sigma}_x} \cdot \frac{y_i - \bar{y}}{\hat{\sigma}_y} \right)$$

Die Quotienten in der Klammer entsprechen der Formel für z-Standardisierung. $z_i = \frac{x_i - \mu}{\sigma}$

Die z-Standardisierung übernimmt dabei die Funktion, die unterschiedlichen Streuungen der beiden Verteilungen aus der Kovarianz heraus zu rechnen. Die Korrelation ist also im Grunde genommen nichts anderes als die Kovarianz zweier z-standardisierter Variablen mit den Mittelwerten 0 und der Streuung 1:

Merkmalszusammenhänge
Signifikanz von Korrelationen

Die Nullhypothese des Signifikanztests für Korrelationen besagt, dass eine empirisch ermittelte Korrelation r zweier Variablen aus einer Grundgesamtheit stammt, in der eine Korrelation ρ von Null besteht.

	In Wirklichkeit gilt die H_0	In Wirklichkeit gilt die H_1
Entscheidung zugunsten der H_0	Richtige Entscheidung	**β-Fehler**
Entscheidung zugunsten der H_1	**α-Fehler**	Richtige Entscheidung

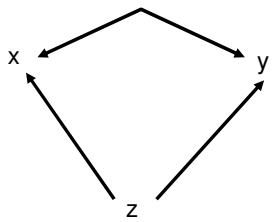
Merkmalszusammenhänge
Korrelation und Kausalität

Ist eine Korrelation vorhanden so sagt diese noch nichts über zugrunde liegende Ursache-Wirkungs-Beziehungen zwischen den beteiligten Merkmalen aus. Nicht immer ist klar, in welche Richtung die Kausalität verläuft.

Darüber hinaus ist eine hohe Korrelation kein Garant dafür, dass überhaupt ein direkter ursächlicher Zusammenhang zwischen den untersuchten Merkmalen besteht. Beide Variablen x und y können von einer dritten gemeinsamen Ursache abhängen. Dieses Phänomen wird als Scheinkorrelation bezeichnet.

„Man vermutet oft dort einen Kausalnexus, wo lediglich eine temporäre Koinzidenz oder eine zufällige Korrelation zugrunde liegen„ (Jürgen von der Lippe)

Merkmalszusammenhänge
Korrelation und Kausalität



Merkmalszusammenhänge
Korrelation

	Intervallskala	Rangskala	Nominalskala (dichotom)
Intervallskala	Produkt-Moment-Korrelation	Rangkorrelation	Punktbiseriale Korrelation
Rangskala		Rangkorrelation	Punktbiseriale Korrelation
Nominalskala (dichotom)			Phi-Koeffizient

Aussicht nächster Termin
Workshop August

Vielen Dank für Ihre Aufmerksamkeit