TU Dresden, Institut für Soziologie 17.04.07
FS: Berufsverläufe und Berufserfolg von Hochschulabsolventen II
Dozent: Dipl.-Soz. Mike Kühne
Referentin: Marlen Gnerlich

# Korrelation und Kausalität

## 1. Korrelation und Kausalität – Begriffsdifferenzierung

- ***Korrelation*** ist ein quantitatives Maß zur Beschreibung linearer Zusammenhänge

→ Enge des Zusammenhangs wird durch den Korrelationskoeffizienten charakterisiert

- ***Kausalität*** impliziert ein Ursache-Wirkungs-Prinzip

→ eine Kausalhypothese muss drei Vorraussetzungen erfüllen:

1.) zwischen X und Y besteht ein statistischer Zusammenhang
2.) die Ursache X geht der Wirkung Y zeitlich voraus
3.) Zusammenhang zw. X und Y besteht auch nach Eliminierung von Drittvariablen

## 2. Formen bivariater Zusammenhänge

*Bsp. 1*: Zusammenhang Ökonomischer Status des Elternhauses und Studium*
n=100/ $n_{priv}$=50/ $n_{nichtpriv}$=50

| Ökonomischer Status X | Studium ja | nein |
|---|---|---|
| privilegiert | 39 | 11 |
| nicht privilegiert | 6 | 44 |

[*fiktiver Datensatz]

### 2.1 Stochastische Anhängigkeit

- Stochastik = Teilgebiet der Statistik, das sich mit der Untersuchung vom Zufall abhängiger Ereignisse und Prozesse befasst
- stochastisch [griech. stochastikós = mutmaßend] bedeutet vom Zufall abhängig

→ stochastische Abhängigkeit = zufallsbedingte Abhängigkeit

*Exkurs: Gegensatz zu stochastischer Abhängigkeit ist deterministische Abhängigkeit mit eindeutiger Funktion, wie z.B. y = f(x); deterministische Zusammenhänge sind in den empirischen Wissenschaften nur selten*

- stochastische Abhängigkeit herrscht vor, wenn sich die Wahrscheinlichkeit für das Auftreten von Variable Y durch das Auftreten von Variable X ändert
- Konzept der stochastischen Abhängigkeit ist die allgemeinste Form, um auszudrücken, dass es einen Zusammenhang zwischen Zufallsvariablen gibt, d.h. es zeigt an, dass es in Abhängigkeit der Werte der einen Variable zu ***irgendwelchen Änderungen*** der Verteilung der anderen Variable kommt

*Bsp. 1*: P(Y=Studium|X=priv) = 39/50 = **0.78** ≠ **0.45** P(Y=Studium)

→ Unterschied zwischen bedingter und unbedingter Wkt. = es liegt stochast. Abhängigkeit vor

### 2.2 Regressive Abhängigkeit

- beschreibt die Beziehung zwischen zwei Zufallsvariablen, indem aufgezeigt wird, in welchem Maß sich bedingte Erwartungswerte der einen Zufallsvariable in Abhängigkeit vom jeweiligen Wert der anderen Variable verändern → ***bedingte Erwartung (synonym: Regression)***

→ zeigt, ob ein linearer Zusammenhang besteht

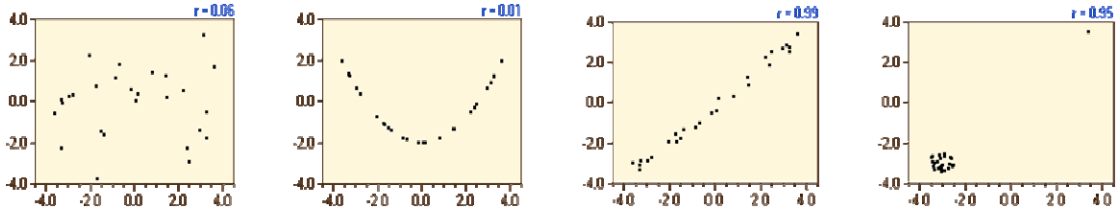
- Messung: Summierung aller Werte Y - gewichtet mit der Wkt. des Auftretens dieser Werte

→ Unterscheiden sich die Erwartungswerte von Y in Abhängigkeit von X, dann ist Y ***regressiv abhängig***; unterscheiden sich die Erwartungswerte nicht, dann ist Y regressiv unabhängig von X

### 2.3 Korrelation

- Kennwert, der Stärke sowie Richtung eines linearen Zusammenhangs zweier Variablen angibt
- Bedeutung der Korrelation manchmal überbewertet → hoher Korrelationskoeffizient bedeutet nicht immer hohe Korrelation; Korrelationskoeffizient von null bedeutet nicht notwendigerweise, dass keinerlei Beziehung zwischen zwei Variablen vorherrscht
  - → Abb. 1 unkorrelierter Datensatz, während Abb. 2 eine perfekte parabolische Beziehung zeigt, obwohl der Korrelationskoeffizient in beiden Fällen nahe null ist
  - → umgekehrt muss ein hoher Korrelationskoeffizient nicht unbedingt einer hohen Korrelation zw. Daten zuzuschreiben sein (wie Abb.3), sondern kann auch auf einen einzelnen Ausreißer, der abseits des unkorrelierten Rests der Datenpunkte liegt, zurückzuführen sein (s. Abb. 4)

Abb. 1 Abb. 2 Abb. 3 Abb. 4




## 3. Korrelationsanalyse

Ziel der Korrelationsanalyse: Beziehungen zwischen Variablen zu entdecken und Zusammenhänge aufzuklären → Maß für die Korrelation ist Korrelationskoeffizient

### 3.1 Spearman-Rangkorrelation $r_s$

Für **ordinalskalierte** Variablen, eignet sich der Rangkorrelationskoeffizient nach Spearman. Hier werden die einzelnen Beobachtungen von x bzw. y der Größe nach geordnet. Jedem Wert wird seine Rangzahl zugewiesen. Es entstehen so n Paare mit Rangzahlen $rg(x_i)$ und $rg(y_i)$. Aus diesen Rängen wird der Korrelationskoeffizent nach Bravais-Pearson errechnet. Man erhält so den Korrelationskoeffizenten nach Spearman-Pearson:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

*Bsp. 2.:* Kenntnisse von Studenten in Methoden und Statistik an Hand von Prüfungsergebnissen*

| Student (n=10) | Rangplatz Methoden | Rangplatz Statistik | $D_i$ | $D_i^2$ |
|---|---|---|---|---|
| A | 8 | 9 | 1 | 1 |
| B | 2 | 1 | 1 | 1 |
| C | 6 | 10 | 4 | 16 |
| D | 3 | 3 | 0 | 0 |
| E | 7 | 5 | 2 | 4 |
| F | 1 | 2 | 1 | 1 |
| G | 4 | 6 | 2 | 4 |
| H | 5 | 4 | 1 | 1 |
| I | 10 | 7 | 3 | 9 |
| J | 9 | 8 | 1 | 1 |

$D_i$ – Differenz zwischen beiden Rängen des jeweiligen Studenten
$r_s$ kann Werte zwischen -1 und 1 annehmen, wobei $r_s = 1$: völlige Übereinstimmung der Rangreihen; $r_s = -1$: völlig gegenläufige Rangreihen; $r_s = 0$: Unabhängigkeit zwischen beiden Rangreihen

$$r_s = 1 - \frac{6 \cdot 38}{10\ (100-1)} = 1 - \frac{228}{990} = 1 - 0{,}230 = \underline{\mathbf{0{,}77}}$$

→ relativ starker Zusammenhang zwischen den Kenntnissen in Methoden und Statistik

### 3.2 Pearson-Produkt-Moment-Korrelation r

Die Produkt-Moment-Korrelation r ist ein statistischer Kennwert für die Enge des linearen Zusammenhangs. Neben der Voraussetzung **intervallskalierter** Variablen, zwischen denen ein linearer Zusammenhang besteht, gilt, dass beide Variablen normalverteilt sein müssen.

$$r(x,y) = \frac{1}{n} \cdot \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{cov(x,y)}{s_x \cdot s_y}$$

*Bsp. 3.:* Bildungsgrad (Bildungsjahre) und Einkommen (entnommen aus: Diekmann, S. 581)

| Bildungsjahre $x_i$ | Einkommen $y_i$ |
|---|---|
| 9 | 3500 |
| 8 | 2400 |
| 18 | 5200 |
| 9 | 3200 |
| 9 | 2300 |
| 10 | 4500 |
| 18 | 12000 |
| 10 | 6500 |
| 9 | 2300 |
| 13 | 4600 |
| 10 | 1600 |
| 9 | 2900 |

| | |
|---|---|
| ohne Abschluss | = 8 |
| Hauptschule | = 9 |
| Realschule | = 10 |
| Abitur | = 13 |
| Hochschule | = 18 |

$\bar{x} = 11$
$\bar{y} = 4250$
$s_x = 3{,}49$
$s_y = 2824$
$\underline{\mathbf{r_{xy} = 0{,}75}}$

→ relativ starker Zusammenhang zwischen Bildungsgrad und Einkommen

### 3.3 Fehler und Fallen bei der Interpretation von Korrelationen

Eine Korrelation zwischen zwei Variablen zu beobachten, kann dazu verleiten, eine kausale Beziehung zwischen diesen Variablen zu sehen.
→ *Storchenbsp.*: Abnahme der Anzahl von Störchen korrelierte mit der Abnahme der Geburtenzahlen, weil Industrialisierung sowohl zum Rückgang von Geburten als auch der Storchenpopulation führte
**!** Im strengen wissenschaftstheoretischen Sinn ist Kausalität ***nicht beweisbar***, Kausalität lässt sich korrelationsstatistisch nur widerlegen.

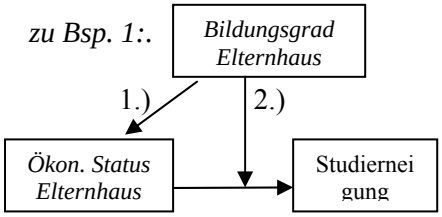
## 4. Konfundierung

### 4.1 Was ist Konfundierung?

- Zusammenhänge zwischen Zufallsvariablen können durch dritte Variablen beeinflusst werden, was die kausale Interpretation von Effekten verfälscht → z.B. durch *Moderation* (Einfluss einer sog. *Moderatorvariable* auf Zusammenhang zweier Merkmale) oder *Konfundierung*
- von Konfundierung spricht man dann, wenn es eine Störvariable W gibt, die sowohl mit der unabhängigen Variable X stochastisch zusammenhängt (1.Bedingung für Konfundierung), als auch den regressiven Zusammenhang zwischen der abhängigen Variable Y der unabhängigen Variable X verändert (2. Bedingung für Konfundierung)

**!** Eine Regression E(Y|X) ist folglich konfundiert, wenn gilt:

*zu Bsp. 1:.*




1.) die Ereignisse X=x und W=w sind stochastisch abhängig sind ***und***
2.) $E(Y|X=x) \neq E(Y|X=x, W=w)$

**!** potenzielle Störvariablen sind Variablen, die Eigenschaften von Personen beschreiben (z.B. Geschlecht, Bildungsgrad, Motivation) → sie teilen Populationen in Subpopulationen (männlich-weiblich, Hauptschule-Realschule-Abitur)

### 4.2 Testen von Konfundierung und Problematik bei quasi-experimentellen Studien

- bei „echten“, d.h. kontrollierten Experimenten ermöglicht die Randomisierung (zufällige Zuteilung der Probanden zu einer Kategorie der unabhängigen Variable) die Kontrolle von Drittvariablen, so dass eine Variation der unabhäng. Variable allein auf Variation der abhäng. Variable zurückgeführt werden kann → Konfundierung wird eliminiert bzw. kontrolliert
- bei der Arbeit mit Daten aus Bevölkerungsumfragen ist eine Randomisierung nicht möglich

→ Überprüfung von Konfundierung ist daher bei quasi-experimentellen Designs sehr bedeutend

1.) Suchen nach potentieller Störvariable, um sie zu erheben und berücksichtigen zu können:
2.) Prüfen der 1.Bedingung für Konfundierung (X und W stochastisch abhängig?)
3.) Prüfen der 2. Bedingung (Gleichheit vs. Verschiedenheit der bedingten Erwartungswerte)

*zu Bsp. 1*: Zusammenhang Ökonomischer Status des Elternhauses und Studium

→ Überlegung: Theorie Bourdieus, dass kulturelles Kapital wesentlichen Einfluss auf Studienentscheidung hat, daher Berücksichtigung des Bildungsgrades der Eltern als mögliche Konfundierung - potenzielle Störvariable W=Hochschulabschluss

*Bildung von Subpopulationen aus Datensatz von Bsp. 1:*

Hochschulabschluss (n=46)

| Ökonomischer Status X | Studium Y ja | nein |
|---|---|---|
| privilegiert | 37 | 1 |
| nicht privilegiert | 5 | 2 |

Kein Hochschulabschluss (n= 54)

| Ökonomischer Status X | Studium Y ja | nein |
|---|---|---|
| privilegiert | 2 | 10 |
| nicht privilegiert | 1 | 42 |

*1. X und W stochastisch abhängig?*

P(X=priv) = 50/100 = 0.50
P(W=Hochschul|X=priv) = 38/46 = 0.82
→ da sich unbedingte und bedingte Wkt. unterscheiden, liegt stochastische Abhängigkeit vor

*2. Verschiedenheit der Erwartungswerte?*

$E(Y|X=x) \neq E(Y|X=x, W=w)$
→ $E(Y=1|X=priv) = P(Y=1|X=priv) = 39/50 = 0.78$
→ $E(Y=1|X=priv, W=Hochschul) = P(Y=1|X=priv, W=Hochschul) = 37/38 = 0.97$
→ Erwartungswerte von Y unterscheiden sich in Abhängigkeit von X und W, folglich wird die regressive Abhängigkeit zwischen X und Y durch W konfundiert.

**!** Einfluss von Drittvariablen ist von großer Bedeutung für die Interpretierbarkeit von Effekten.
**!** Man wird nie jede potenzielle Störvariable erfassen, zudem können aus praktischen Gründen nicht beliebig viele Variablen berücksichtigt werden.

---


**Literatur:**
**Bortz**, Jürgen: Statistik für Sozialwissenschaftler, 5. Aufl., 1999.
**Bortz**, Jürgen, **Döhring**, Nicola: Forschungsmethoden und Evaluation, 2. Aufl., 1995.
**Diekmann**, Andreas: Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 10. Aufl., 2003.
**Ghanbari**, Azizi S.: Einführung in die Statistik für Sozial- und Erziehungswissenschaftler, 2002.
**Küpper**, Claudia: Korrelation und Kausalität. Arbeitspapier, 2000.
**Nachtigall**, Christof, **Suhl**, Ute**, Steyer**, Rolf: Einführung in die Konfundierungsanalyse, in: methevalreport 2(1) 2000.
**Steyer**, Rolf: Was wollen und was können wir durch empirische Kausalforschung erfahren?, 1997.

**Internet:**
http://wirtschaft.fh-duesseldorf.de/fileadmin/dekanat/Schmeink/025_rangkorrelation.pdf (10.04.2007)
http://user.uni-frankfurt.de/~reiss/stat_b.pdf (14.04.2007)
http://www.phil.uni-sb.de/~jakobs/seminar/vpl/expost/kausal/htm (14.04.2007)