Einführung in die Forschungsmethoden der Psychologie

BSc Philosophie-Neurowissenschaften-Kognition WiSe 23/24

BSc Psychologie WiSe 23/24

Prof. Dr. Dirk Ostwald

(7) Korrelative Designs

Korrelative Designs

Definition

Nicht-experimentelle Datenanordnungen, die typischerweise nur den wechselseitigen Zusammenhang (Korrelation) zwischen zwei oder mehreren Variablen betreffen.

Reiß and Sarris (2012)

Bemerkungen

- "Correlation is not causation!"

⇒ Korrelationen können immer durch Drittvariablen kausal bedingt sein.

Ja, aber ...

- Der Begriff der "Kausalität" ist nicht eindeutig definiert.
- Experimentelle Designs werden im Normalfall mit Korrelationen (Regression, ALM) untersucht.
- Kausale Inferenz nutzt auch "nur" probabilistische Modelle, wird wenig angewendet/gelehrt.
- Im Sinne der zeitlichen Präzedenz benutzt kausale Inferenz zum Teil Zeitserienmodelle.
- Das Psychologiestudium sieht keine Auseinandersetzung mit Graphical Models und Differentialgleichungsmodellen vor, die für ein Verständnis zeitgenössischer kausaler Inferenz nötig wäre.

Kausalzusammenhänge sind ein latentes Konstrukt das nur datenanalytisch erschlossen werden kann!

⇒ Kausalzusammenhänge sind sowohl in experimentellen als auch korrelativen Designs latent!

Anwendungsszenario

Psychotherapie

Mehr Therapiestunden

⇒ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

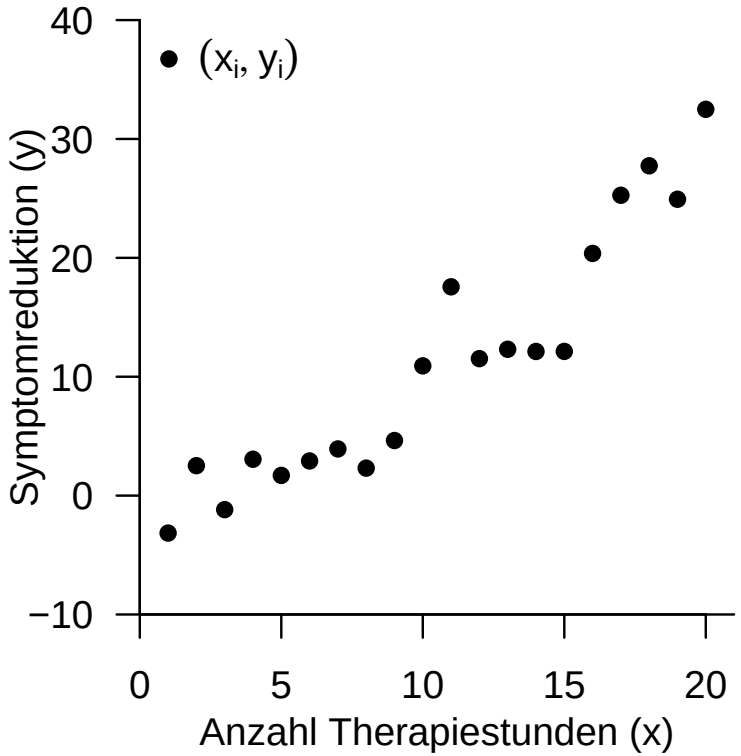
- Symptomreduktion

Simulierter Datensatz aus den Fundus einer psychotherapeutischen Hochschulambulanz

$i = 1, \ldots, 20$ Patient:innen, y_i Symptomreduktion bei Patient:in i, x_i Anzahl Therapiestunden von Patient:in i

y_i	x_i
-3.2	1
2.5	2
-1.2	3
3.1	4
1.7	5
2.9	6
3.9	7
2.3	8
4.6	9
10.9	10
17.6	11
11.5	12
12.3	13
12.1	14
12.1	15
20.4	16
25.3	17
27.7	18
24.9	19
32.5	20

Beispieldatensatz



Wie stark hängen Anzahl Therapiestunden und Symptomreduktion zusammen?

Definition (Korrelation)

Die *Korrelation* zweier Zufallsvariablen ξ und υ ist definiert als

$$\rho(\xi, \upsilon) := \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} \tag{1}$$

wobei $\mathbb{C}(\xi, \upsilon)$ die Kovarianz von ξ und υ und $\mathbb{S}(\xi)$ und $\mathbb{S}(\upsilon)$ die Standardabweichungen von ξ und υ, respektive, bezeichnen.

Bemerkungen

- $\rho(\xi, \upsilon)$ wird auch *Korrelationskoeffizient* von ξ und υ genannt.
- Wir haben bereits gesehen, dass $-1 \leq \rho(\xi, \upsilon) \leq 1$ gilt.
- Wenn $\rho(\xi, \upsilon) = 0$ ist, werden ξ und υ *unkorreliert* genannt.
- Wir haben bereits gesehen, dass aus der Unabhängigkeit von ξ und υ, folgt dass $\rho(\xi, \upsilon) = 0$.
- Aus $\rho(\xi, \upsilon) = 0$ folgt aber wie bereits gesehen die Unabhängigkeit von ξ und υ im Allgemeinen nicht.

Definition (Stichprobenkorrelation)

$\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ sei ein Datensatz. Weiterhin seien:

- Die Stichprobenmittel der x_i und y_i definiert als

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \text{ und } \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{2}$$

- Die Stichprobenstandardabweichungen x_i und y_i definiert als

$$s_x := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ und } s_y := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{3}$$

- Die Stichprobenkovarianz der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

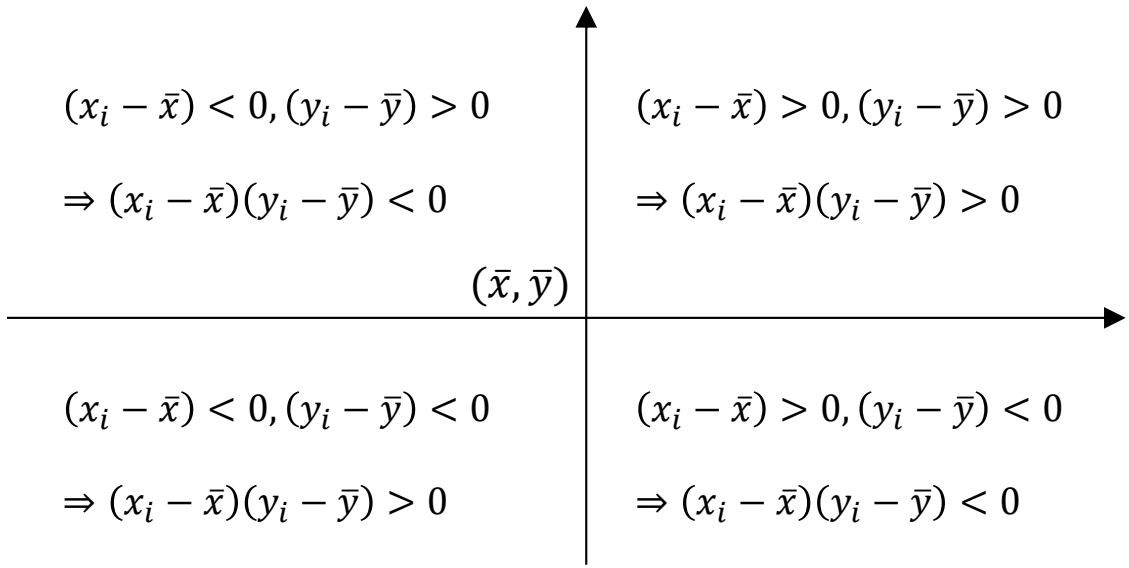
$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{4}$$

Dann ist die *Stichprobenkorrelation* der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

$$r_{xy} := \frac{c_{xy}}{s_x s_y} \tag{5}$$

und wird auch *Stichprobenkorrelationskoeffizient* genannt.

Mechanik der Kovariationsterme



Häufige richtungsgleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Positive Korrelation

Häufige richtungsungleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Negative Korrelation

Keine häufigen richtungsgleichen oder -entgegengesetzten Abweichungen $\Rightarrow$ Keine Korrelation

Beispiel

```r
# Laden des Beispieldatensatzes
fname = file.path(getwd(), "7_Daten", "7_Korrelation_Beispieldatensatz.csv")     # Dateipfad
D     = read.table(fname, sep = ",", header = TRUE)                              # Laden als Dataframe
x_i   = D$x_i                                                                    # x_i Werte
y_i   = D$y_i                                                                    # y_i Werte
n     = length(x_i)                                                              # n

# "Manuelle" Berechnung der Stichprobenkorrelation
x_bar = (1/n)*sum(x_i)                                                           # \bar{x}
y_bar = (1/n)*sum(y_i)                                                           # \bar{y}
s_x   = sqrt(1/(n-1)*sum((x_i - x_bar)^2))                                       # s_x
s_y   = sqrt(1/(n-1)*sum((y_i - y_bar)^2))                                       # s_y
c_xy  = 1/(n-1) * sum((x_i - x_bar) * (y_i - y_bar))                             # c_{xy}
r_xy  = c_xy/(s_x * s_y)                                                         # r_{xy}
print(r_xy)                                                                      # Ausgabe
```

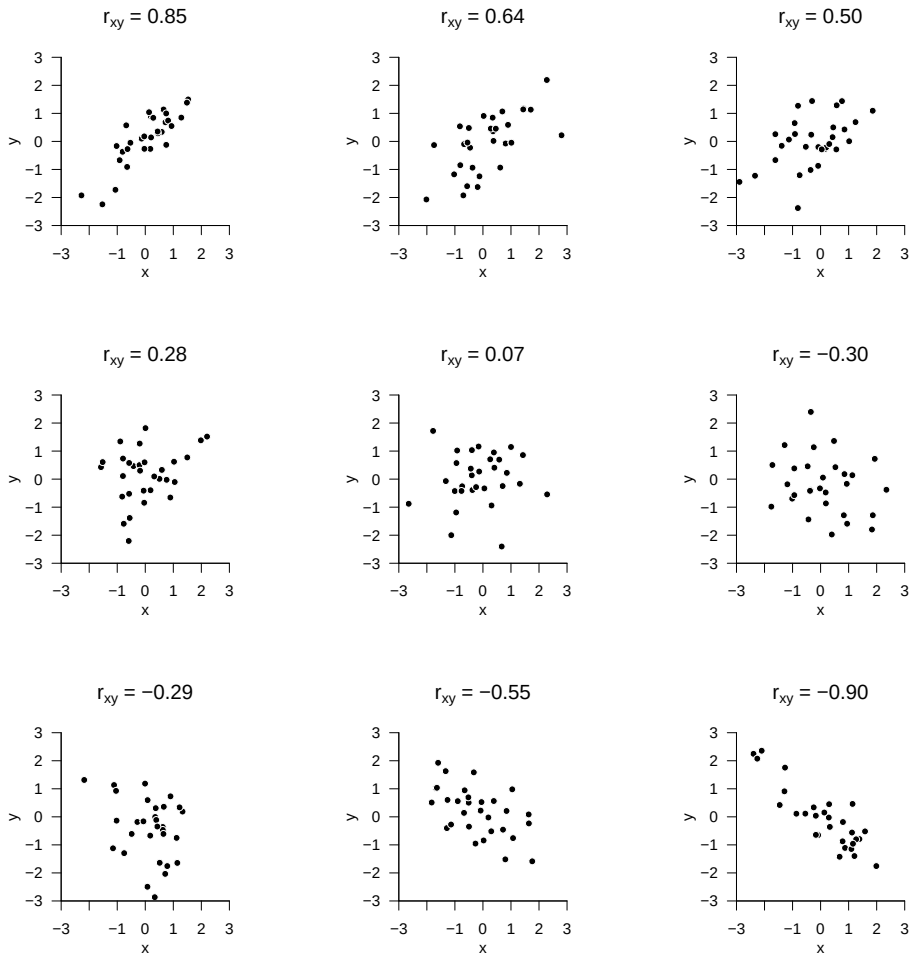
```
[1] 0.9378162
```

```r
# Automatische Berechnung mit cor()
r_xy  = cor(x_i,y_i)                                                             # r_{xy}
print(r_xy)                                                                      # Ausgabe
```

```
[1] 0.9378162
```

⇒ Anzahl Therapiestunden und Symptomreduktion sind hochkorreliert.

Beispiele


r_{xy} = 0.85
r_{xy} = 0.64
r_{xy} = 0.50
r_{xy} = 0.28
r_{xy} = 0.07
r_{xy} = −0.30
r_{xy} = −0.29
r_{xy} = −0.55
r_{xy} = −0.90
x
y

Theorem (Stichprobenkorrelation bei linear-affinen Transformationen)

Für einen Datensatz $\{(x_i, y_i)\}_{i=1,\ldots n} \subset \mathbb{R}^2$ sei $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1,\ldots n} \subset \mathbb{R}^2$ eine linear-affin transformierte Wertemenge mit

$$(\tilde{x}_i, \tilde{y}_i) = (a_x x_i + b_x, a_y y_i + b_y), a_x, a_y \neq 0. \tag{6}$$

Dann gilt

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \tag{7}$$

Bemerkungen

- Der Betrag der Stichprobenkorrelation ändert sich bei linear-affiner Datentransformation nicht.
- Man sagt, dass die Stichprobenkorrelation im Gegensatz zur Stichprobenkovarianz *maßstabsunabhängig* ist.
- Rechnet man z.B. die x_i durch Multiplikation mit 1.000 von *kg* in *g* um, bleibt die Korrelation gleich.

Beweis

Es gilt

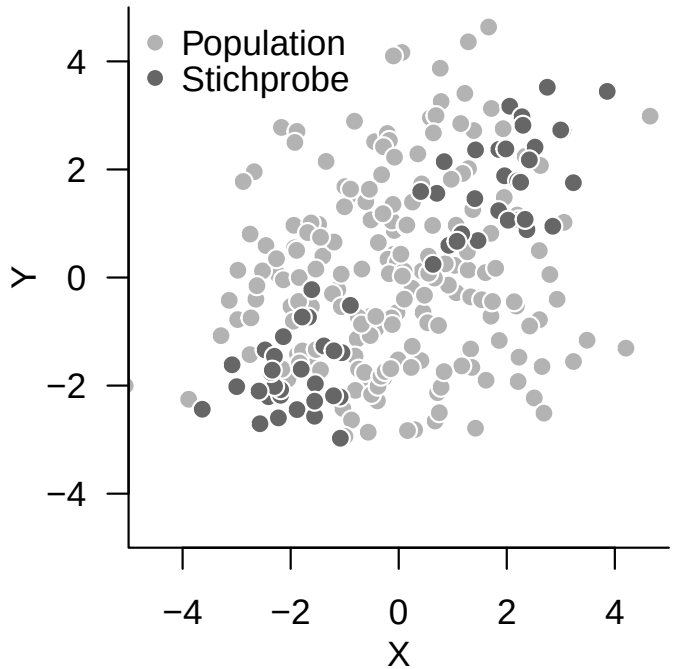
$$
\begin{aligned}
r_{\tilde{x}\tilde{y}} &:= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{x}_i - \bar{\tilde{x}}\right)^2}\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{y}_i - \bar{\tilde{y}}\right)^2}} \\
&= \frac{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x} + b_x))(a_y y_i + b_y - (a_y\bar{y} + b_y))}{\sqrt{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x} + b_x))^2}\sqrt{\sum_{i=1}^{n}(a_y y_i + b_y - (a_y\bar{y} + b_y))^2}} \\
&= \frac{a_x a_y \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{a_x^2\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{a_y^2\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (8) \\
&= \frac{a_x a_y}{|a_x||a_y|}\,\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{a_x a_y}{|a_x||a_y|}\,\frac{c_{xy}}{s_x s_y} \\
&= \frac{a_x a_y}{|a_x||a_y|}\,r_{xy}.
\end{aligned}
$$

Also folgt, durch Durchspielen aller möglichen Vorzeichenfälle, dass

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \qquad (9)$$

□

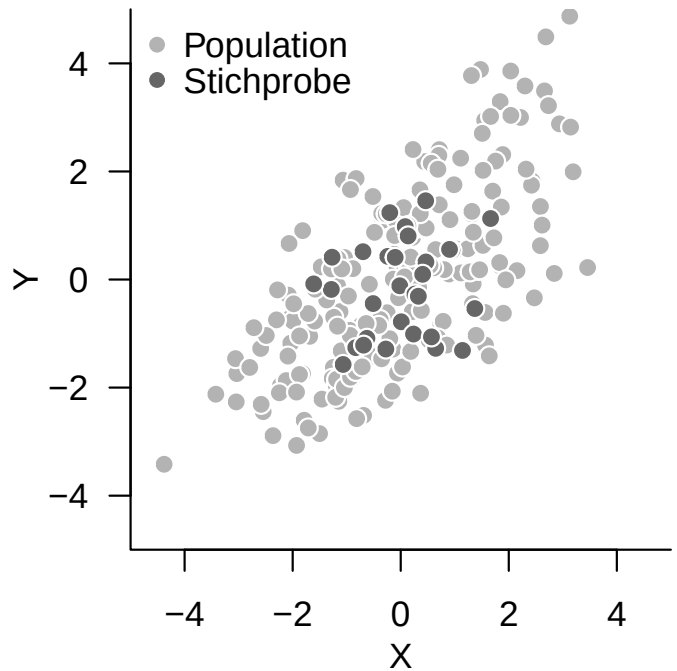
Stichprobenauswahleffekte | Stichprobe aus Extremgruppen




```
Korrelation basierend auf Gesamtpopulationsdaten = 0.3567178

Korrelation basierend auf Stichprobendaten       = 0.9188452
```

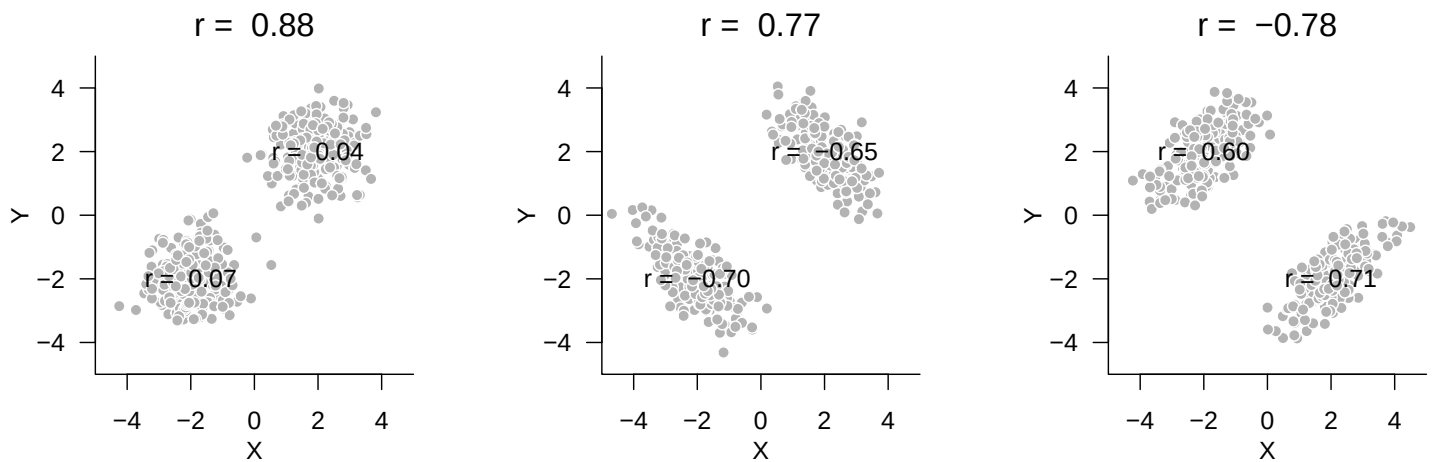
Stichprobenauwahleffekte | Stichprobe mit zu kleiner Streubreite



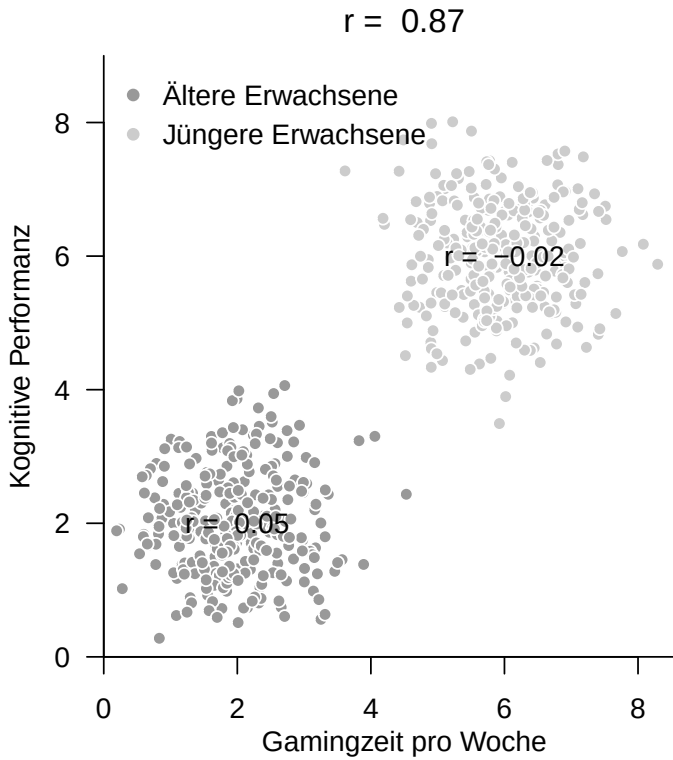

```
Korrelation basierend auf Gesamtpopulationsdaten = 0.6850208

Korrelation basierend auf Stichprobendaten       = 0.1859573
```

Stichprobenauswahleffekte | Bedingte vs. unbedingte Korrelationen (Simpson's Paradox)



Stichprobenauswahleffekte | Bedingte vs. unbedingte Korrelationen (Simpson's Paradox)



⇒ Durch Drittvariable Alter (jung, alt) induzierte Korrelation!

Selbstkontrollfragen

1. Definieren Sie den Begriff des Korrelativen Designs nach Reiß und Sarris (2012).

2. Geben Sie die Definition der Stichprobenkorrelation wieder.

3. Erläutern Sie die Mechanik der Kovariationsterme.

4. Geben Sie das Theorem zur Stichprobenkorrelation bei linear-affiner Transformation wieder.

5. Erläutern Sie die Bedeutung des Theorems zur Stichprobenkorrelation bei linear-affiner Transformation.

Referenzen

Reiß, Siegbert, and Viktor Sarris. 2012. *Experimentelle Psychologie: von der Theorie zur Praxis*. Pearson Studium Psychologie. München: Pearson.