

Statistik für Digital Humanities

Kovarianz & Korrelation

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

16. November 2020

[Letzte Aktualisierung: 15/11/2020, 18:00]

Beziehungen zwischen Variablen

Mögliche Beziehung zwischen Variablen

- positiv: Je höher x , desto höher y
Übungszeit \rightarrow Sprachverständnis
- nicht vorhanden: Kein Zusammenhang zwischen x und y
Übungszeit \rightarrow Anzahl Sonneneruptionen
- negativ: Je höher x desto niedriger y
Übungszeit \rightarrow Freizeit

2 wesentliche Beziehungsmaße

- Kovarianz
- Korrelation Wir konzentrieren uns erstmal nur auf bivariate Korrelation, also zwischen 2 Variablen

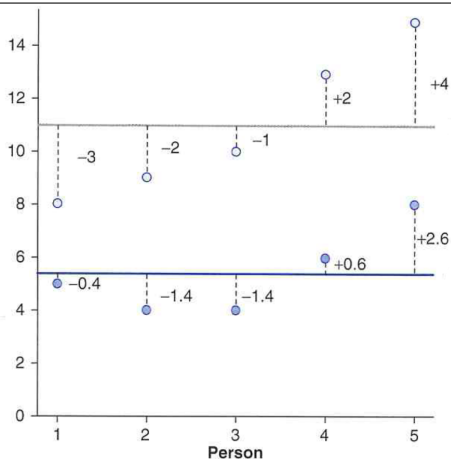
Anknüpfungspunkt Varianz

- Abweichung (deviance) = $x_i - \bar{x}$
- Naiv: Abweichungen addieren = $\sum(x_i - \bar{x})$
 - $X = \{22, 40, 53, 57\}$
 - $\bar{x} = 43$
 - Totaler Fehler = $-21 + -3 + 10 + 14 = 0$
- Halbgut: Quadratabweichungen addieren $SS = \sum(x_i - \bar{x})^2$
 - Sum of Squares steigt mit Stichprobengröße
- Gut: SS mit Stichprobengröße normalisieren
Varianz $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

→ Kovarianz bestimmt, wie sehr zwei Variablen ko-variiieren

Kovarianz

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92



Kovarianz

- Abweichung (deviance) = $x_i - \bar{x}$
- Halbgut: Kreuzprodukt der Abweichung (cross-product deviance) = $(x_i - \bar{x}) * (y_i - \bar{y})$
 - positiv wenn x und y positiv oder negativ abweichen
 - negativ wenn x und y in verschiedene Richtungen abweichen
 - Summe der Kreuzprodukte der Abweichung steigt mit Stichprobengröße
- Gut: mit Stichprobengröße normalisieren
Kovarianz $cov(X, Y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Kovarianz

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92

$$\begin{aligned}\text{Kovarianz } \text{cov}(X, Y) &= \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1} \\ &= \frac{(-0.4)*(-3) + (-1.4)*(-2) + (-1.4)*(-1) + (0.6)*(2) + (2.6)*(4)}{4} \\ &= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\ &= \frac{17}{4} = 4.25\end{aligned}$$

```
werbung<-c(5,4,4,6,8)
gekauft<-c(8,9,10,13,15)
advertData<-data.frame(werbung, gekauft)
cov(advertData$werbung, advertData$gekauft)
```

Kovarianz

Kovarianz wird durch Maßskalierung verzerrt

Proband	1	2	3	4	5	\bar{x}	s
Arbeitsweg (Meilen)	5	4	4	6	8	5.4	1.67
Wanderlust (Meilen)	8	9	10	13	15	11.0	2.92

$$\text{cov}(x, y) = 4.25$$

Nach Umrechnung in Kilometer (*1.6)

$$\text{cov}(x, y) = 11$$

→ Vergleiche zwischen Datensätzen mit Kovarianz problematisch, deshalb...

Pearsons Korrelationskoeffizient

→ Vergleiche zwischen Datensätzen mit Kovarianz problematisch, deshalb...

– Pearson, K. (1920): *Notes on the History of Correlation*

– mit Standardabweichung normieren

Korrelationskoeffizient $r(X, Y) = \frac{\text{cov}(X, Y)}{s_X * s_Y} = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(n-1) * s_X * s_Y}$

– auch Pearsons Produkt-Moment Korrelationskoeffizient

– r liegt zwischen -1 und 1

→ +1 : perfekt positive Korrelation, x steigt proportional zu y

→ 0 : kein linearer Zusammenhang, während x steigt, bleibt y gleich

→ -1 : perfekt negative Korrelation, x steigt indirekt proportional zu y

– Indikator für Effektstärke Kein Beweis → Kontext und vergleichbare Ergebnisse beachten

→ ± 0.1 : schwacher Effekt

→ ± 0.3 : moderater Effekt

→ ± 0.5 : starker Effekt

Pearsons Korrelationskoeffizient

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92

$$\text{cov}(X, Y) = \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

$$= \frac{17}{4} = 4.25$$

$$s_x * s_y = 4.88$$

$$r = 4.25 / 4.88 = 0.87$$

```
werbung<-c(5,4,4,6,8)
```

```
gekauft<-c(8,9,10,13,15)
```

```
advertData<-data.frame(werbung, gekauft)
```

```
cor(advertData$werbung, advertData$gekauft)
```

r als Teststatistik

Hypothesentest für

- Ist die Korrelation ungleich 0?
- Ist r wahrscheinlich wenn es keinen messbaren Effekt gäbe?

r ist nicht normalverteilt, deswegen z-Transformation

Fisher R.A.(1921): *On the probable Error of a coefficient of correlation deduced from a small sample*

- $z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$
- $SE_{zr} = \frac{1}{\sqrt{n-3}}$

Dann normal mit z-Score arbeiten

- $z = \frac{z_r}{SE_{zr}}$
- $p(r)$ aus der z-Tabelle ablesen ("Smaller portion")
- $p(r)$ verdoppeln weil two-tailed
- $p < 0.05 \rightarrow$ Korrelation signifikant

oder bald folgenden t-Test verwenden mit $df = 2$

- $t_r = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$

r als Teststatistik

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92

$$r = 4.25/4.88 = 0.87$$

$$z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) = 1.33$$

$$SE_{z_r} = \frac{1}{\sqrt{n-3}} = 0.71$$

$$z = \frac{z_r}{SE_{z_r}} = \frac{1.33}{0.71} = 1.87$$

$$p(r) = 0.0307 \text{ (Tabelle)}$$

→ Korrelation signifikant für one-tailed, nicht signifikant für two-tailed

```
werbung<-c(5,4,4,6,8)
```

```
gekauft<-c(8,9,10,13,15)
```

```
advertData<-data.frame(werbung, gekauft)
```

```
cor.test(advertData$werbung, advertData$gekauft)
```

```
#Liefert auch Konfidenzintervalle
```

```
#Ergebnisse leicht anders als hier -> Rundungsfehler
```

Konfidenzintervalle für r

r ist nicht normalverteilt, deswegen z-Transformation

Fisher R.A.(1921): *On the propable Error of a coefficient of correlation deduced from a small sample*

$$- z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

$$- SE_{zr} = \frac{1}{\sqrt{n-3}}$$

Dann KI für z_r berechnen (für 95%)

$$- \text{Untergrenze} = z_r - (1.96 * SE_{zr})$$

$$- \text{Obergrenze} = z_r + (1.96 * SE_{zr})$$

Dann auf r zurücktransformieren

$$- r = \frac{e^{2*z_r} - 1}{e^{2*z_r} + 1}$$

Konfidenzintervalle für r

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92

$$r = 4.25/4.88 = 0.87$$

$$z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) = 1.33$$

$$SE_{z_r} = \frac{1}{\sqrt{n-3}} = 0.71$$

Zwischenergebnis:

$$\text{Untergrenze} = 1.33 - (1.96 * 0.71) = -0.062$$

$$\text{Obergrenze} = 1.33 + (1.96 * 0.71) = 2.72$$

$$KI_U = \frac{e^{2*z_r} - 1}{e^{2*z_r} + 1} = \frac{e^{2*-0.062} - 1}{e^{2*-0.062} + 1} = -0.062 \quad KI_O = \frac{e^{2*z_r} - 1}{e^{2*z_r} + 1} = \frac{e^{2*2.72} - 1}{e^{2*2.72} + 1} = 0.991$$

Interpretation von r

Proband	1	2	3	4	5	\bar{x}	s
Werbung gesehen	5	4	4	6	8	5.4	1.67
Bambinas gekauft	8	9	10	13	15	11.0	2.92

Determinationskoeffizient oder Bestimmtheitsmaß R^2

- Maß für Effektstärke
- $R^2 = r * r$
- Beispiel $r = 0.87$
 - $R^2 = 0,7569$
 - *Werbung gesehen* ist für 75,69% der Variation bei *Bambinas gekauft* verantwortlich
 - 24,31 % der Variation bei *Bambinas gekauft* durch andere Variablen

Achtung: Kein Nachweis für Kausalität, auch wenn es oft so fehlinterpretiert wird

Statistische Annahmen von r

- Linearer Zusammenhang (ja/nein): Intervallskalierung Sortiert, Abstände zwischen den Werten der Skala aussagekräftig
- Darüber hinaus: Beide Variablen normalverteilt oder eine normal und die andere binärskaliert

→ Sonst: nicht parametrisches Korrelationsmaß oder Bootstrapping

Nicht-Parametrische Korrelationsmaße

- Spearman's Korrelationskoeffizient ρ
- Kendalls tau τ

Spearman's Korrelationskoeffizient ρ

- Spearman C. (1904): *The proof and measurement of association between two things*
- wie Pearsons r aber statt x und y wird $Rang(x)$ und $Rang(y)$ verwendet
- $Rang(x)$ = Position in sortierter Liste
- Also: X und Y sortieren, x_i und y_i mit dem jeweiligen Rang in X und Y ersetzen und r berechnen.
- Interpretation von ρ analog zu r

```
cor(advertData$werbung, advertData$gekauft, method="spearman")  
cor.test(advertData$werbung, advertData$gekauft,  
method="spearman", alternative="less") #alternative -> one/two-sided
```

Kendalls tau τ

- Kendall M. G. (1970): *Rank correlation methods*
- scheinbar besser als Spearman (Howell, D.C. (1997): *Statistical Methods for Psychology*)
- definitiv besser bei kleinen Datensätzen und vielen gleichrangigen Werten
- Interpretation von τ analog zu r

Berechnung (Laut Wikipedia):

- Sortiere Paare $\{x_i, y_i\}$ nach x
- vergleiche alle Paare $\{x_i, y_i\}$ mit allen Paaren $\{x_j, y_j\}$ mit $i < j$
 - C = Anzahl der Paare : $x_i < x_j, y_i < y_j$ Konkordanz
 - D = Anzahl der Paare : $x_i < x_j, y_i > y_j$ Diskordanz
 - T_Y = Anzahl der Paare : $x_i \neq x_j, y_i = y_j$ Bindung in Y
 - T_X = Anzahl der Paare : $x_i = x_j, y_i \neq y_j$ Bindung in X
- **Kendalls tau** $\tau = \frac{C-D}{\sqrt{(C+D+T_Y)*(C+D+T_X)}}$

```
cor(advertData$werbung, advertData$gekauft, method="kendall")
```

Vergleiche zwischen Korrelationen

- Vergleiche zwischen unabhängigen Korrelationen
- Vergleiche zwischen abhängigen Korrelationen

Vergleiche zwischen unabhängigen Korrelationen

Werbung - Bambina Studie mit 51 *male* und 52 *female* wiederholt

- $n_{male} = 51, n_{female} = 52$
- $r_{male} = -0.506, r_{female} = -0.381$

$$Z_{Differenz} = \frac{z_{male} - z_{female}}{\sqrt{\frac{1}{n_{male}-3} + \frac{1}{n_{female}-3}}}$$

- $Z_{Differenz} = \frac{-0.557 - (-0.401)}{\sqrt{\frac{1}{49} + \frac{1}{48}}} = -0.768$
- z-Score Tabelle liefert 0.221 (one-tailed / gerichtet größer oder kleiner)
- verdoppeln liefert 0.442 (two-tailed / ungerichtet unterschiedlich)
- → Kein signifikanter Korrelationsunterschied Werte statistisch wahrscheinlich

Vergleiche zwischen unabhängigen Korrelationen

```
zdifference<-function(r1, r2, n1, n2){  
  zd<-(atanh(r1)-atanh(r2))/sqrt(1/(n1-3)+1/(n2-3))  
  p <-1 - pnorm(abs(zd))  
  print(paste("Z Difference: ", zd))  
  print(paste("One-Tailed P-Value: ", p))  
  print(paste("Two-Tailed P-Value: ", 2*p))  
}
```

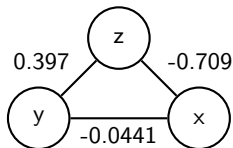
Vergleiche zwischen abhängigen Korrelationen

Studie zu Prüfungsstress

x = Prüfungsangst, y = Performanz, z = Abgabezeit

Ist die Korrelation zwischen x und y stärker als die zwischen z und y ?

H_0 → Kein signifikanter Unterschied.



– $n = 103, r_{xy} = -0.0441, r_{zy} = 0.397, r_{zx} = -0.709$

– $t_{\text{Differenz}} = (r_{xy} - r_{zy}) * \sqrt{\frac{(n-1)*(1+r_{zx})}{2*(1-r_{xy}^2 - r_{zx}^2 - r_{zy}^2 + 2*r_{xy} + 2*r_{zy} + 2*r_{zx})}}$

– $= (-0.838) * \sqrt{\frac{29.1}{2*(1-0.94-0.503-1.58+0.248)}} = -5.09$

– T-Tabelle ($df = n - 3$, two tailed) liefert 1.96 (95%) und 2.63 (99%) als Kritische Werte

– → Wert signifikant höher als kritischer Wert → H_0 widerlegt

Vergleiche zwischen abhängigen Korrelationen

```
tdifference<-function(rxy, rxz, rzy, n) {  
  df<-n-3  
  td<-(rxy-rzy)*sqrt((df*(1 + rxz))/(2*(1-rxy^2-rxz^2-rzy^2+(2*rxy*rxz*rzy))))  
  p <-pt(td, df)  
  print(paste("t Difference: ", td))  
  print(paste("One-Tailed P-Value: ", p))  
  print(paste("Two-Tailed P-Value: ", 2*p))  
}
```

Zusammenfassung

- Kovarianz als grobes Maß für Beziehung zwischen Variablen anfällig für Messskalierung
- Pearsons r als normalisiertes Maß unabhängig von Messskalierung aber parametrisch
- nicht parametrische Verfahren
 - Spearmans ρ
 - Kendalls τ
- Korrelationen liegen zwischen -1 und 1
 - -1 : negativ, indirekt proportional
 - 1 : positiv, direkt proportional
- Korrelationen sind Indikatoren für Effektstärke
 - ± 0.5 : starker Einfluss
 - ± 0.3 : moderater Einfluss
 - ± 0.1 : schwacher Einfluss

Korrelationen, die übersprungen wurden:

- Partiiell und Semi-Partiiell
- Biseriell und Point-Biseriell

Aktuelle Beispiele

- Paul Caruana-Galizia (2015): *Politics and the German language: Testing Orwell's hypothesis using the Google N-Gram corpus*
- "shows that six non-technical non-Naziwords – Demokratie(democracy), Freiheit(freedom), Frieden(peace), Herrlichkeit(glory), Gerechtigkeit(justice), and Heldentumd(heroism) – are (1) highly correlated with explicitly Nazi words; (2) negatively correlated with Germany's level of democracy; and (3) negatively correlated with the count of riots, anti-government protests, and government crises"

- kritischer Kommentar: vermutliche Drittvariable *Kriegsbereitschaft* und Korrelation wird fälschlicherweise als Hinweis auf Zusammenhang verwendet