# PDF2Text und Korrelationen

## Die Pandemie in Sprache und Text – Corona-Podcasts & co.

Jürgen Hermes, Lukas Mönch, Felix Rau, Nils Reiter

23. April 2021

Section 1

# PDF $\rightarrow$ Text

# PDF → Text

- PDF
  - File format for print and presentation
  - Contains everything needed (including fonts)

# PDF → Text

- PDF
  - File format for print and presentation
  - Contains everything needed (including fonts)
- Text
  - Not stored as a single stream of characters
  - Multiple text elements per page, located at specific positions
  - Reconstructing plain text is difficult

# PDF → Text

- PDF
  - File format for print and presentation
  - Contains everything needed (including fonts)
- Text
  - Not stored as a single stream of characters
  - Multiple text elements per page, located at specific positions
  - Reconstructing plain text is difficult
- Libraries
  - Java: Apache PDFBox `https://pdfbox.apache.org`
  - Python: PyPDF2 `https://pypi.org/project/PyPDF2/`

*demo*

# Einleitung

- Vom Ende her gedacht: Was wollen wir aussagen?
- Es besteht ein Zusammenhang zwischen Variablen $x$ und $y$
- Zur Quantifizierung des Zusammenhangs ist der Datentyp der Variablen entscheidend: Numerisch, ordinal oder nominal?

# Einleitung

- Vom Ende her gedacht: Was wollen wir aussagen?
- Es besteht ein Zusammenhang zwischen Variablen $x$ und $y$
- Zur Quantifizierung des Zusammenhangs ist der Datentyp der Variablen entscheidend: Numerisch, ordinal oder nominal?

## Beispiel (Numerisch – Numerisch)

- „Aus Sprechweise von Drosten lässt sich Inzidenz vorhersagen“
  - $x$: Anzahl der Seufzer in Drostens Redeanteil
  - $y$: 7-Tages-Inzidenz am Tag nach Aufnahme des Podcasts
- Hypothese: Wenn $x$ steigt, dann steigt auch $y$

Bortz/Schuster (2010, Kap. 10)

# Einleitung

- Vom Ende her gedacht: Was wollen wir aussagen?
- Es besteht ein Zusammenhang zwischen Variablen $x$ und $y$
- Zur Quantifizierung des Zusammenhangs ist der Datentyp der Variablen entscheidend: Numerisch, ordinal oder nominal?

## Beispiel (Numerisch – Ordinal)

- „Inzidenz schlägt Drosten aufs Gemüt“
  - $x$: 7-Tages-Inzidenz am Tag vor Aufnahme des Podcasts (numerisch)
  - $y$: Sentiment von Drostens Äußerungen: positiv/neutral/negativ (ordinal)
    - Ordinal: Kategorien, aber mit Reihenfolge
- Hypothese: $x$ ist höher, wenn $y =$ negativ

Bortz/Schuster (2010, Kap. 10)

# Einleitung

- Vom Ende her gedacht: Was wollen wir aussagen?
- Es besteht ein Zusammenhang zwischen Variablen $x$ und $y$
- Zur Quantifizierung des Zusammenhangs ist der Datentyp der Variablen entscheidend: Numerisch, ordinal oder nominal?

## Beispiel (Numerisch – Nominal)

- „Cisek benutzt mehr Verben!“
  - $x$: Anzahl Verben in einer Sendung (numerisch)
  - $y$: Wissenschaftler:in in einer Sendung (nominal)
- Hypothese: $x$ ist höher, wenn $y =$Cisek

# Einleitung

- Vom Ende her gedacht: Was wollen wir aussagen?
- Es besteht ein Zusammenhang zwischen Variablen $x$ und $y$
- Zur Quantifizierung des Zusammenhangs ist der Datentyp der Variablen entscheidend: Numerisch, ordinal oder nominal?

## Beispiel (Nominal – Nominal)

- „Die häufigste Wortart von Drosten sind Nomen“
  - $x$: Meistverwendete Wortart in einer Sendung (nominal)
  - $y$: Wissenschaftler:in in einer Sendung (nominal)
- Hypothese: Wenn $y =$Drosten, dann steigt die Wahrscheinlichkeit dass $x =$Nomen

Section 3

# Numerisch – Numerisch

# Kovarianz

- Ausgangspunkt: Zwei Stichproben (= Datenreihen)
- Wir berechnen zunächst die Kovarianz der Stichproben mit $n$ Einträgen

# Kovarianz

- Ausgangspunkt: Zwei Stichproben (= Datenreihen)
- Wir berechnen zunächst die Kovarianz der Stichproben mit $n$ Einträgen

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

- $x_i$, $y_i$: Erhobene Werte an Stelle $i$
- $\bar{x}$, $\bar{y}$: Durchschnitt der Stichprobe

# Kovarianz

- Ausgangspunkt: Zwei Stichproben (= Datenreihen)
- Wir berechnen zunächst die Kovarianz der Stichproben mit $n$ Einträgen

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

- $x_i$, $y_i$: Erhobene Werte an Stelle $i$
- $\bar{x}$, $\bar{y}$: Durchschnitt der Stichprobe
- Intuition: Grad des „miteinander Variierens“ Bortz/Schuster (2010, 155)
- Sind beide Werte stark über- oder unterdurchschnittlich, wird $(x_i - \bar{x})(y_i - \bar{y})$ sehr groß

# Beispiel

Hohe Kovarianz

$$\begin{aligned} x &= (1, 3, 7, 9) \\ y &= (2, 4, 8, 10) \end{aligned}$$

# Beispiel

Hohe Kovarianz

$$
\begin{aligned}
x &= (1, 3, 7, 9) \\
y &= (2, 4, 8, 10) \\
\bar{x} = 5 \qquad & \qquad \bar{y} = 6
\end{aligned}
$$

# Beispiel

Hohe Kovarianz

$$
\begin{aligned}
x &= (1, 3, 7, 9) \\
y &= (2, 4, 8, 10) \\
\bar{x} = 5 \qquad & \qquad \bar{y} = 6 \\
\operatorname{cov}(x, y) &= \frac{40}{3} \\
&= 13.334
\end{aligned}
$$

# Beispiel

Hohe Kovarianz

$$
\begin{aligned}
x &= (1,3,7,9) \\
y &= (2,4,8,10) \\
\bar{x} = 5 \qquad & \qquad \bar{y} = 6 \\
\operatorname{cov}(x,y) &= \frac{40}{3} \\
&= 13.334 \\
y' &= (10,8,4,2) \\
\operatorname{cov}(x,y') &= \frac{-40}{3} \\
&= -13.334
\end{aligned}
$$

# Beispiel

Hohe Kovarianz

$$
\begin{aligned}
x &= (1,3,7,9) \\
y &= (2,4,8,10) \\
\bar{x} = 5 \qquad & \bar{y} = 6 \\
\operatorname{cov}(x,y) &= \frac{40}{3} \\
&= 13.334 \\
y' &= (10,8,4,2) \\
\operatorname{cov}(x,y') &= \frac{-40}{3} \\
&= -13.334
\end{aligned}
$$

Niedrige Kovarianz

$$
\begin{aligned}
x &= (1,3,7,9) \\
y &= (10,11,9,10)
\end{aligned}
$$

# Beispiel

Hohe Kovarianz

$$
\begin{aligned}
x &= (1,3,7,9) \\
y &= (2,4,8,10) \\
\bar{x} = 5 \quad & \quad \bar{y} = 6 \\
\operatorname{cov}(x,y) &= \frac{40}{3} \\
&= 13.334 \\
y' &= (10,8,4,2) \\
\operatorname{cov}(x,y') &= \frac{-40}{3} \\
&= -13.334
\end{aligned}
$$

Niedrige Kovarianz

$$
\begin{aligned}
x &= (1,3,7,9) \\
y &= (10,11,9,10) \\
\bar{x} = 5 \quad & \quad \bar{y} = 10 \\
\operatorname{cov}(x,y) &= \frac{4}{3} \\
&= 1.334
\end{aligned}
$$

# Beispiele, visuell

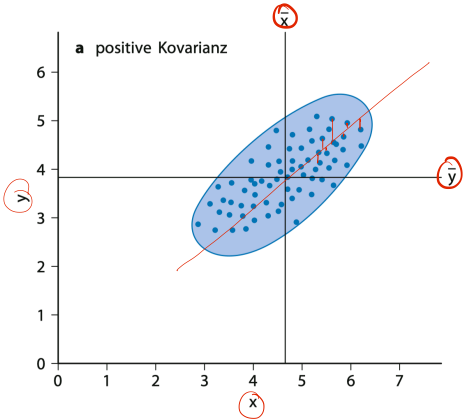



Abbildung: Positive Kovarianz (Bortz/Schuster, 2010)

# Beispiele, visuell

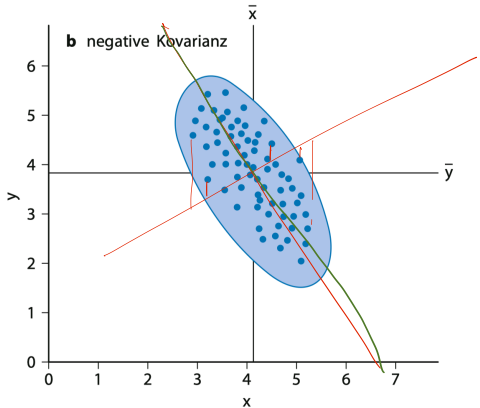



Abbildung: Negative Kovarianz (Bortz/Schuster, 2010)

# Beispiele, visuell

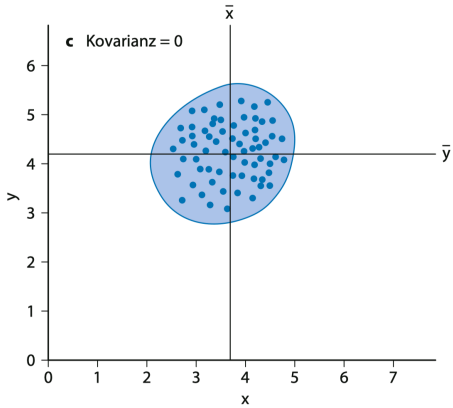



Abbildung: Keine Kovarianz (Bortz/Schuster, 2010)

# Skalierung

- Kovarianz ist sensitiv gegenüber der Skalierung
  - $\operatorname{cov}(x, 2y) = 2\operatorname{cov}(x, y)$
- Viele Skalen in Anwendungen sind unterschiedlich skaliert

# Skalierung

- Kovarianz ist sensitiv gegenüber der Skalierung
  - $\operatorname{cov}(x, 2y) = 2\operatorname{cov}(x, y)$
- Viele Skalen in Anwendungen sind unterschiedlich skaliert
- Lösung: Division durch Produkt der Standardabweichungen der Stichproben

$$\frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)} = \rho$$

# Skalierung

- Kovarianz ist sensitiv gegenüber der Skalierung
  - $\operatorname{cov}(x, 2y) = 2\operatorname{cov}(x, y)$
- Viele Skalen in Anwendungen sind unterschiedlich skaliert
- Lösung: Division durch Produkt der Standardabweichungen der Stichproben

$$\frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)}$$

Das ist die Korrelation!

# Korrelation

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)}$$

- Liegt im Intervall $[-1; 1]$
  - $1$ $(-1)$: Proportional (umgekehrt proportional)
    - $x \to 2x \Rightarrow y \to 2cy$ $(x \to 2x \Rightarrow y \to \frac{y}{2c})$
  - $0$: Kein Zusammenhang zwischen $x$ und $y$
    - $x \to 2x \Rightarrow y \to cy$

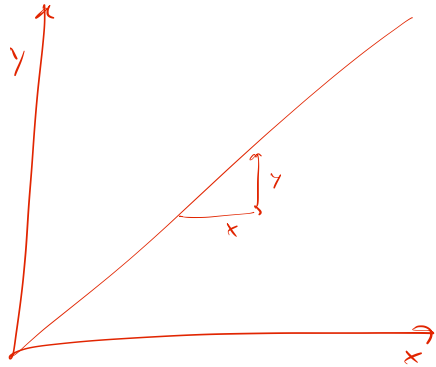
# Korrelation

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)}$$

- Liegt im Intervall $[-1; 1]$
  - $1$ ($-1$): Proportional (umgekehrt proportional)
    - $x \to 2x \Rightarrow y \to 2cy$ ($x \to 2x \Rightarrow y \to \frac{y}{2c}$)
  - $0$: Kein Zusammenhang zwischen $x$ und $y$
    - $x \to 2x \Rightarrow y \to cy$
- „Pearson Correlation Coefficient“ Pearson (1895)

# Korrelation

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)}$$




- Liegt im Intervall $[-1; 1]$
  - $1$ ($-1$): Proportional (umgekehrt proportional)
    - $x \to 2x \Rightarrow y \to 2cy$ ($x \to 2x \Rightarrow y \to \frac{y}{2c}$)
  - $0$: Kein Zusammenhang zwischen $x$ und $y$
    - $x \to 2x \Rightarrow y \to cy$
- „Pearson Correlation Coefficient“ Pearson (1895)
- `http://guessthecorrelation.com`

# Implementations

- Python
  - NumPy: `np.corrcoef(x, y)`
    - Returns a correlation matrix of x-x, y-y, x-y and y-x
  - Pandas: `x.corr(y, method="pearson")`
    - Returns the decimal value

# Implementations

- Python
  - NumPy: `np.corrcoef(x, y)`
    - Returns a correlation matrix of x-x, y-y, x-y and y-x
  - Pandas: `x.corr(y, method="pearson")`
    - Returns the decimal value
- R
  - `cor(x, y)`

# Implementations

- Python
  - NumPy: `np.corrcoef(x, y)`
    - Returns a correlation matrix of x-x, y-y, x-y and y-x
  - Pandas: `x.corr(y, method="pearson")`
    - Returns the decimal value
- R
  - `cor(x, y)`
- Java
  - Apache Commons Math:

```
new PearsonsCorrelation().correlation(double[] xArray, double[] yArray)
```

# Korrelation und Kausalität

- Korrelation != Kausalität
- Mögliche Situationen, wenn $x$ und $y$ korrelieren
  - $x$ beeinflusst $y$ kausal
  - $y$ beeinflusst $x$ kausal
  - $x$ und $y$ werden von einer (oder mehr) dritten Variablen beeinflusst
  - $x$ und $y$ beeinfluss sich wechselseitig kausal

# Korrelation und Kausalität

- Korrelation != Kausalität
- Mögliche Situationen, wenn $x$ und $y$ korrelieren
  - $x$ beeinflusst $y$ kausal
  - $y$ beeinflusst $x$ kausal
  - $x$ und $y$ werden von einer (oder mehr) dritten Variablen beeinflusst
  - $x$ und $y$ beeinfluss sich wechselseitig kausal
- Korrelationskoeffizient gibt keine Auskunft, welche Situation vorliegt!

# Korrelation und Kausalität

- Korrelation != Kausalität
- Mögliche Situationen, wenn $x$ und $y$ korrelieren
  - $x$ beeinflusst $y$ kausal
  - $y$ beeinflusst $x$ kausal
  - $x$ und $y$ werden von einer (oder mehr) dritten Variablen beeinflusst
  - $x$ und $y$ beeinfluss sich wechselseitig kausal
- Korrelationskoeffizient gibt keine Auskunft, welche Situation vorliegt!
- „Die meisten korrelativen Zusammenhänge dürften vom Typus 3 sein, d. h. der Zusammenhang der beiden Variablen ist ursächlich auf andere Variablen zurückzuführen, die auf beide Variablen Einfluss nehmen.“ (Bortz/Schuster, 2010, 159)
  - Das dürfte auch bei uns gelten!

# References I


- Bortz, Jürgen/Christof Schuster (2010). *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin, Heidelberg: Springer [1977].
- Cramér, Harald (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Manning, Christopher D./Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts und London, England: MIT Press.
- Pearson, Karl (1895). „Notes on regression and inheritance in the case of two parents“. In: *Proceedings of the Royal Society of London* 58, S. 240–242.
- — (1900). „On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling“. In: *Philosophical Magazine* 50.302, S. 157–175.
- Spearman, Charles (1904). „The proof and measurement of association between two things“. In: *American Journal of Psychology* 15.1, S. 72–101.