Algorithms for the Social Web

Simon Ginzinger / Markus Zanker

Arbeitsaufwand

- 5 ECTS = 125 Arbeitstunden
- 30 Stunden in LVs
- Bleiben 95 Arbeitsstunden
- 16 Termine, pro Termin zusätzlich **6h Arbeit**

Bewertung (Ginzinger)

- Tasks mit zugeordneten Punkten
- Kommuniziert in der LV (Slides)
- Punkteliste im Wiki

Die ersten 3 LVs

- Statistik (Stichprobe, Kennwerte 1D)
- Kennwerte für höherdimensionale Stichproben
- Grundlagen der Wahrscheinlichkeitsrechnung
- Satz von Bayes, Anwendungen

Statistik

Zusammenhang: Recommender Systems

- **Ziel:** Allgemeine Schlussfolgerungen durch Analyse einer Stichprobe
- Fachbegriffe:
  - *Statistische Einheiten:* Wohnungen, Menschen, Unternehmen, ...
  - *Grundgesamtheit:* Alle Wohnungen in Österreich, alle Menschen in Salzburg, ...
  - *(Zufalls-)Stichprobe:* Untersuchte Teilmenge der Grundgesamtheit
  - *Merkmal mit Ausprägungen:* Größe, Alter, Miete, ...

Merkmal

- **Diskret:** endlich viele oder abzählbar unendlich viele Ausprägungen
- **Stetig:** alle Werte in einem reellen Intervall können als Ausprägung angenommen werden

Univariate Statistik/ Multivariate Statistik

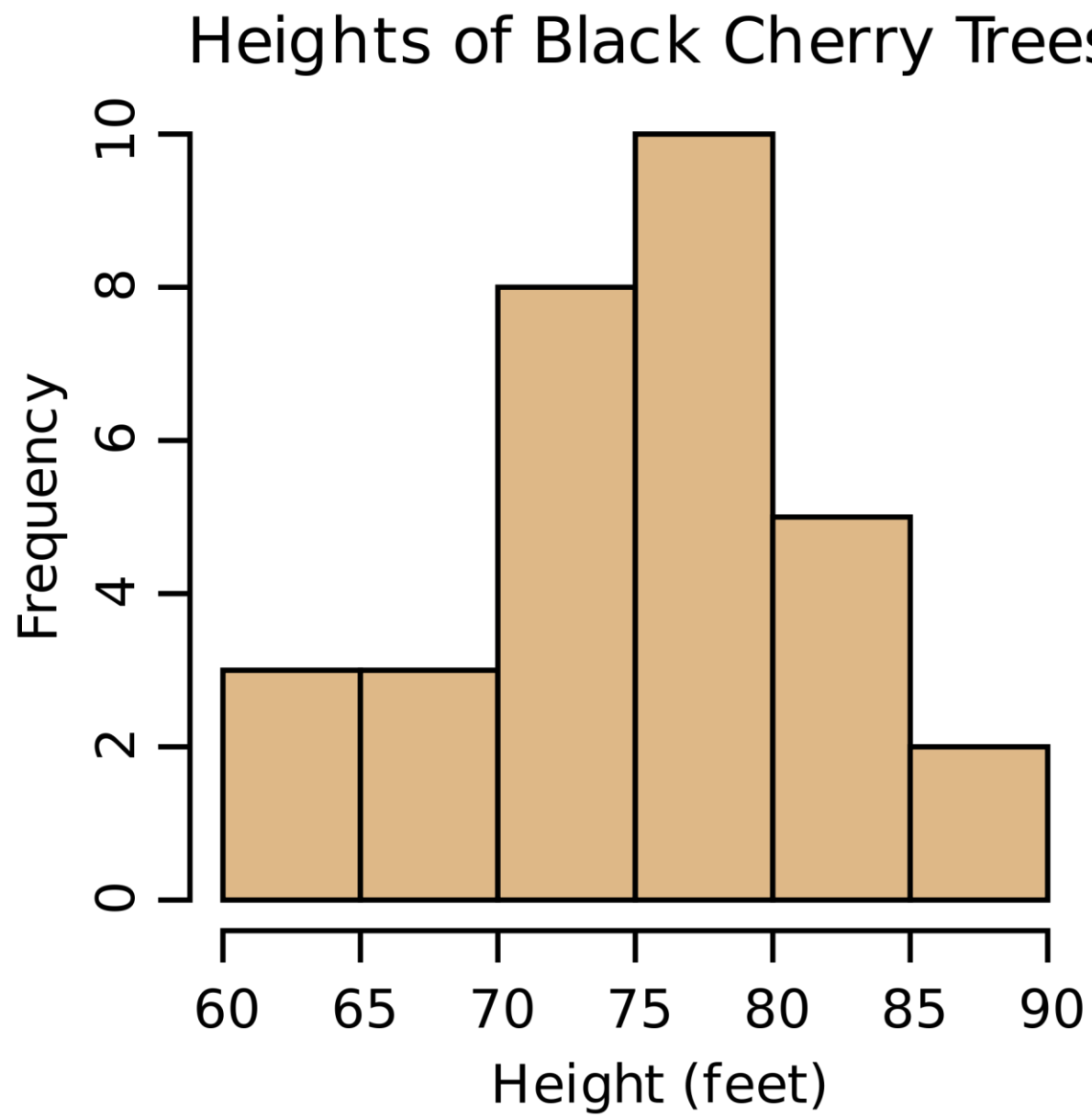
- **Univariate** statistische Fragestellungen untersuchen Daten, die aus der Beobachtung **eines** Merkmals resultieren.
- **Univariate** statistische Methoden bilden die Grundlage für **multivariate** statistische Fragestellungen.

Absolute und relative Häufigkeiten

- Stichprobenwerte (Messwerte)
  $x_1, \dots, x_n \Longrightarrow$Urliste
- Verschiedene Werte $a_1, \dots, a_k$mit $\mathrm{k} \leq n$
- Absolute Häufigkeit h_i Anzahl der Vorkommen des Wertes $a_i$in der Urliste
- **Relative Häufigkeit:** $\boldsymbol{f_i = \frac{h_i}{n}}$
- Beispiel

Stichproben mit vielen verschiedenen Messwerten

- Man betrachtet Intervalle (*Klassen*)
- h_i entspricht nun der Anzahl der Messwerte, die in die i-te Klasse fallen
- Histogramm:
  - Balkendiagramm
  - 1 Balken pro Klasse
  - Fläche entspricht Anzahl der Messwerte in der Klasse


Heights of Black Cherry Trees
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (feet)

Kennwerte einer Stichprobe

Arithmetisches Mittel:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} h_j a_j$$

$$\bar{x} = \sum_{j=1}^{k} f_j \, a_j$$

Weiters:

- Geometrisches Mittel
- Harmonisches Mittel

Median einer **geordneten** Stichprobe

$$\tilde{x} = \begin{cases} x_{m+1} & , \quad n = 2m+1 \\ \frac{1}{2}(x_m + x_{m+1}) & , \quad n = 2m \end{cases}$$

- Unempfindlicher gegenüber Ausreißern als Mittelwert
- Beispiel: -100,1,1,1,2,3,4,4,4,4

p-Quantil einer **geordneten** Stichprobe

$$0 < p < 1$$

$$\tilde{x}_p = \begin{cases} x_{\lfloor np \rfloor + 1} & , \quad np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{np} + x_{np+1}\right) & , \quad np \in \mathbb{N} \end{cases}$$

Beispiel:

Stichprobe: 1, 2, 5, 6, 8, 8, 8, 10

Berechnen $\tilde{x}_{0.25}$

Wert der größer als 25% der Stichprobe ist.

Empirische Varianz und Standardabweichung

Varianz:

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

Standardabweichung:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Erinnerung

- Statistik **schätzt** Kennzahlen von **Grundgesamtheiten**
- **„Gedankenexperiment“:**
  Wir nehmen an, es gäbe **exakte** Daten, die aber nicht erhoben / ausgewertet werden können. Durch Statistik sollen die Kennzahlen einer Stichprobe **möglichst genau** die Kennzahlen der **exakten Daten wiederspiegeln**.

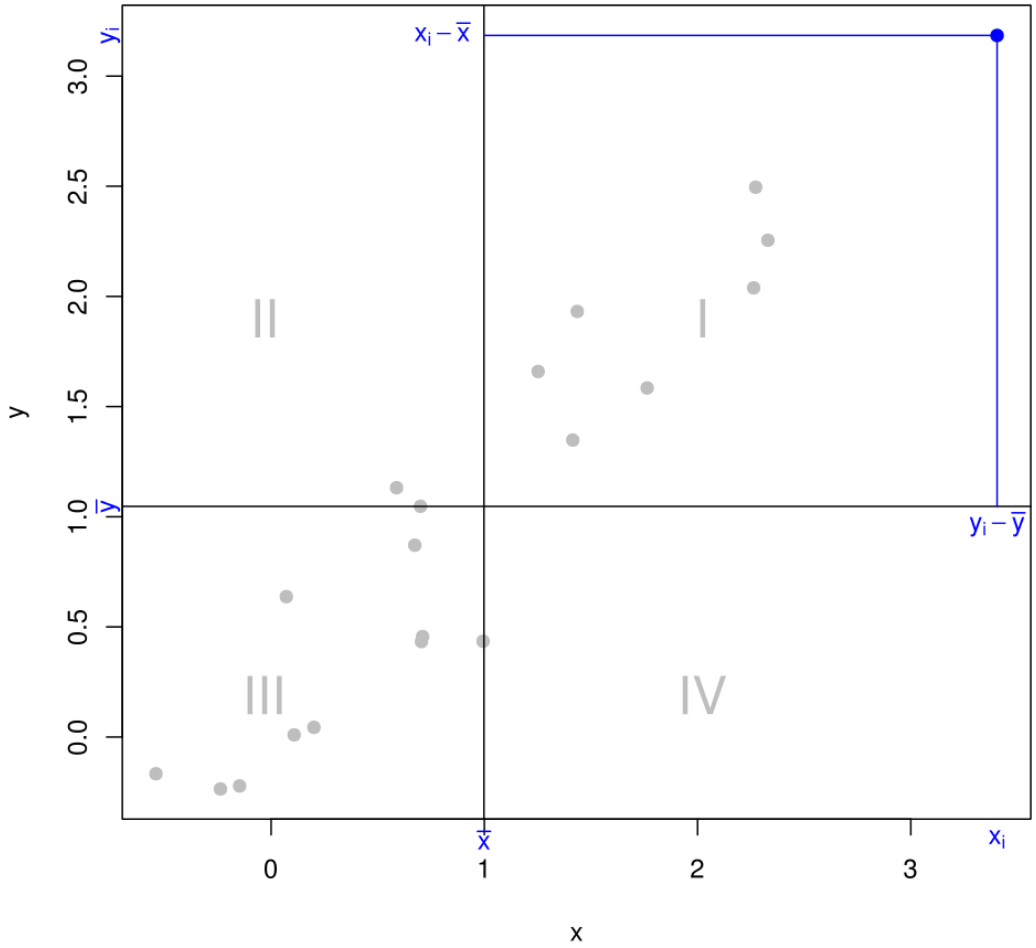
English

Empirical Covariance

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

The empirical covariance is positive if the values x_i and y_i are linearly and directionally related.

Example: Empirical Covariance



Erstellt von Sigbert
(http://de.wikipedia.org/w/index.php?title=Datei:Covariance.svg&page=1&filetimestamp=20110820142905)

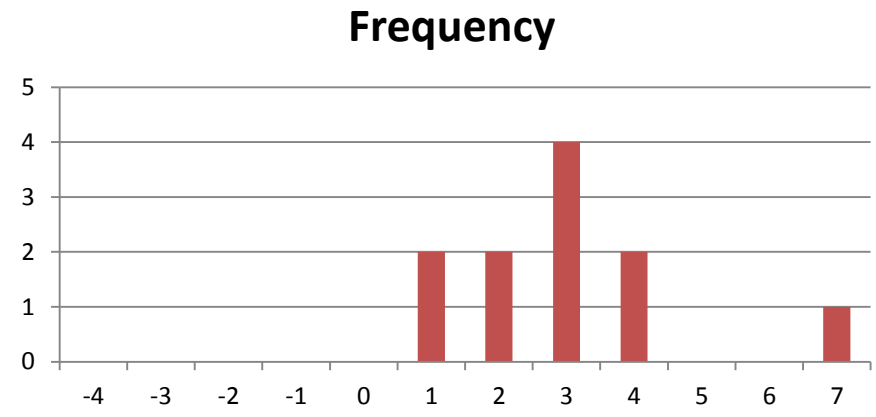
Example: Standard Deviation

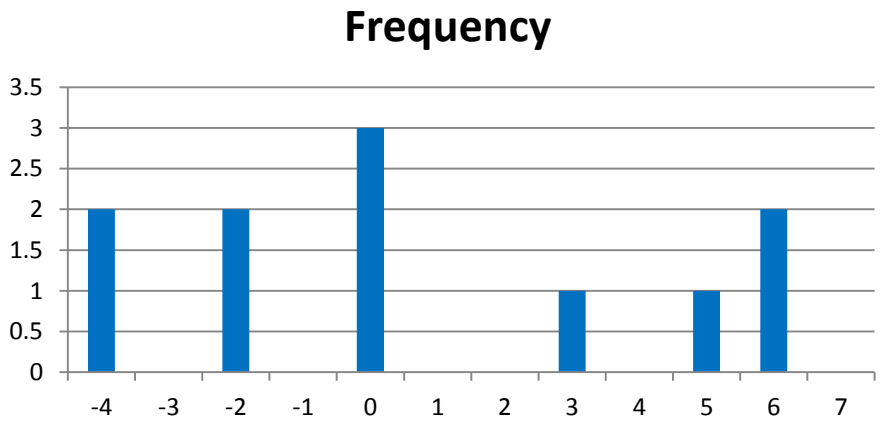
Stichprobe 1	Stichprobe 2
1	-5
1	-5
2	-2
2	-2
3	0
3	0
3	0
3	3
4	5
4	6
7	6
3.00	**0.55**
2.80	**16.07**
1.67	**4.01**

$\bar{x}$

s^2

s





Pearson Correlation

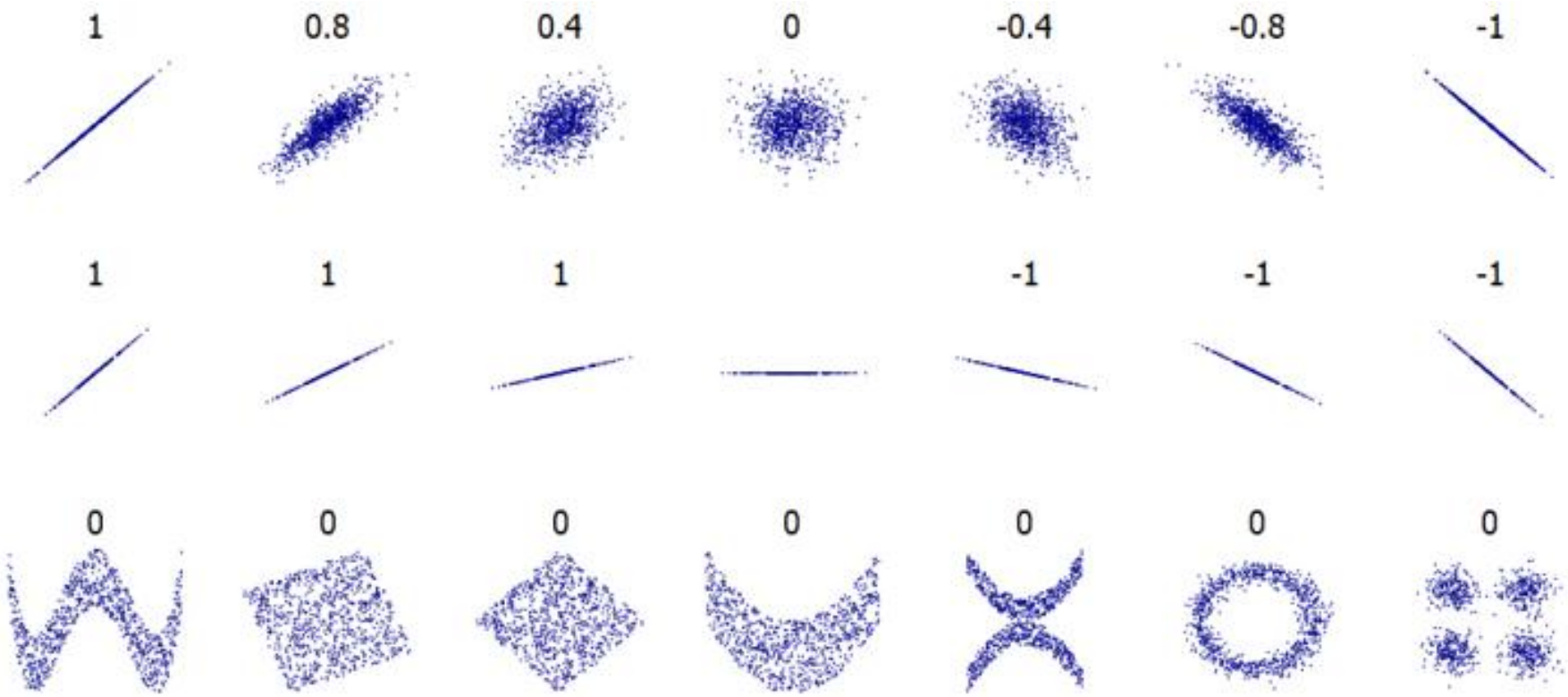
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson Correlation

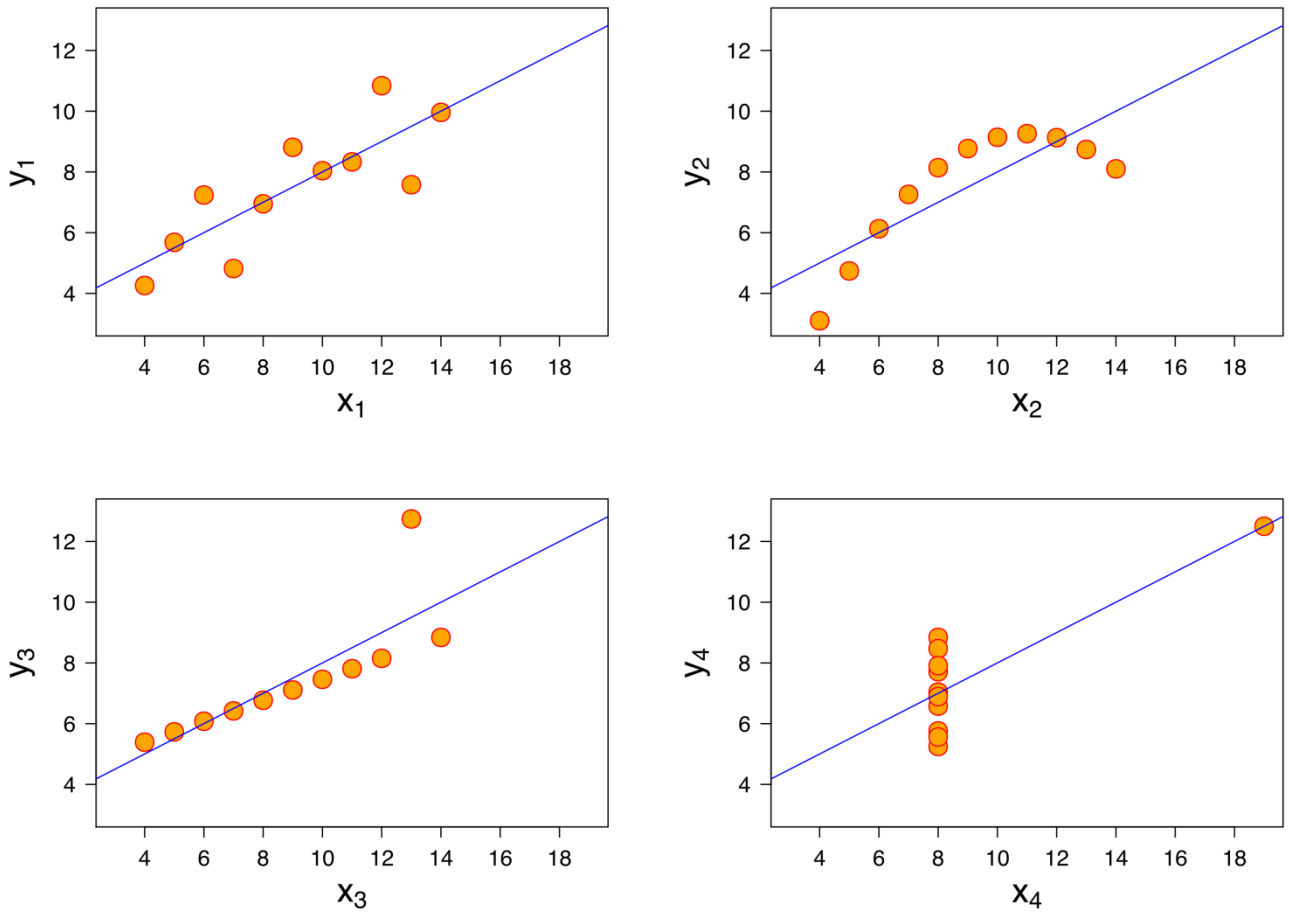
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



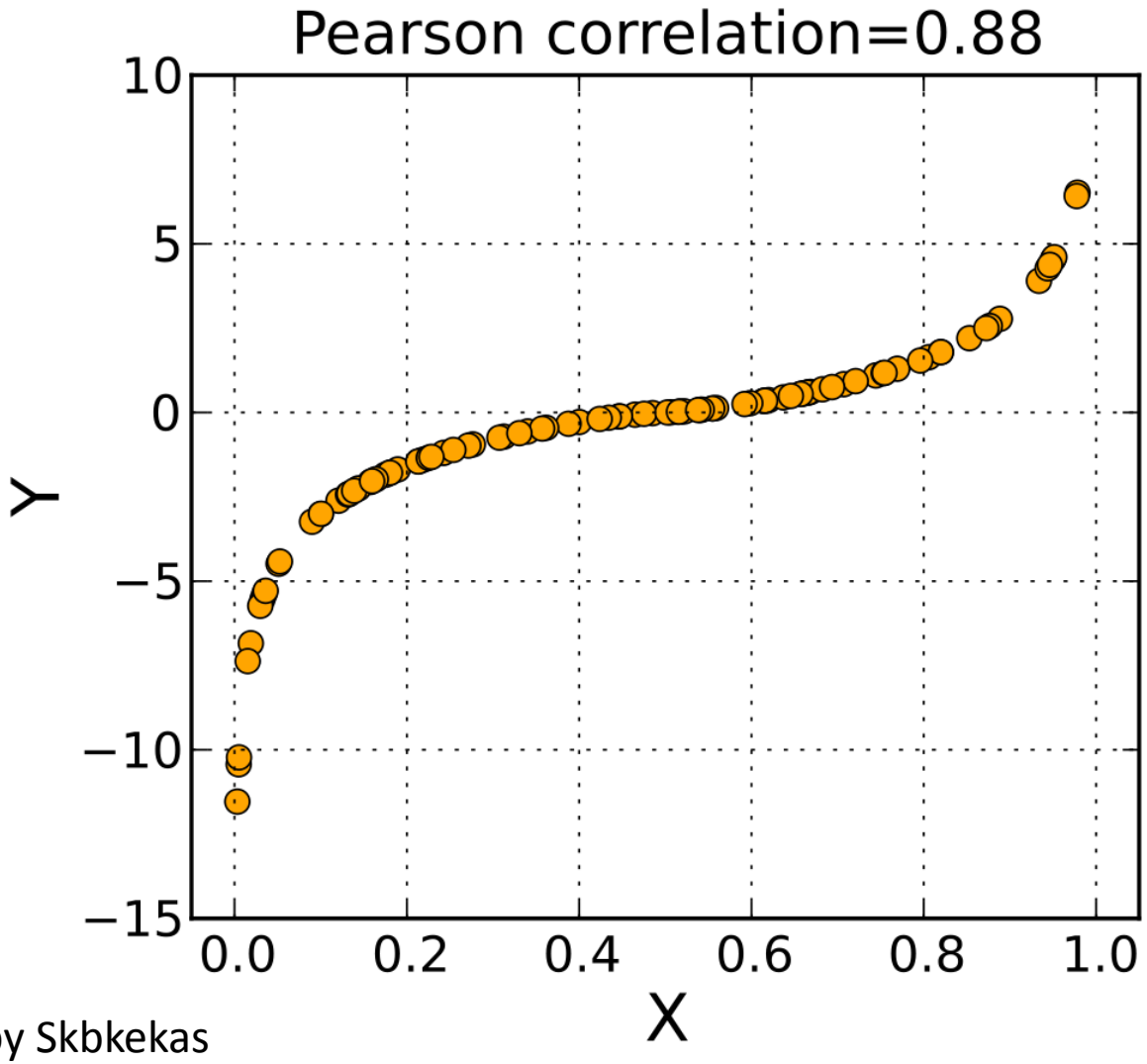
Created by DenisBoigelot
(http://en.wikipedia.org/w/index.php?title=File:Correlation_examples2.svg&page=1)

Pearson Correlation: 0.816



Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Trouble?



Created by Skbkekas
(http://en.wikipedia.org/w/index.php?title=File:Spearman_fig2.svg&page=1)

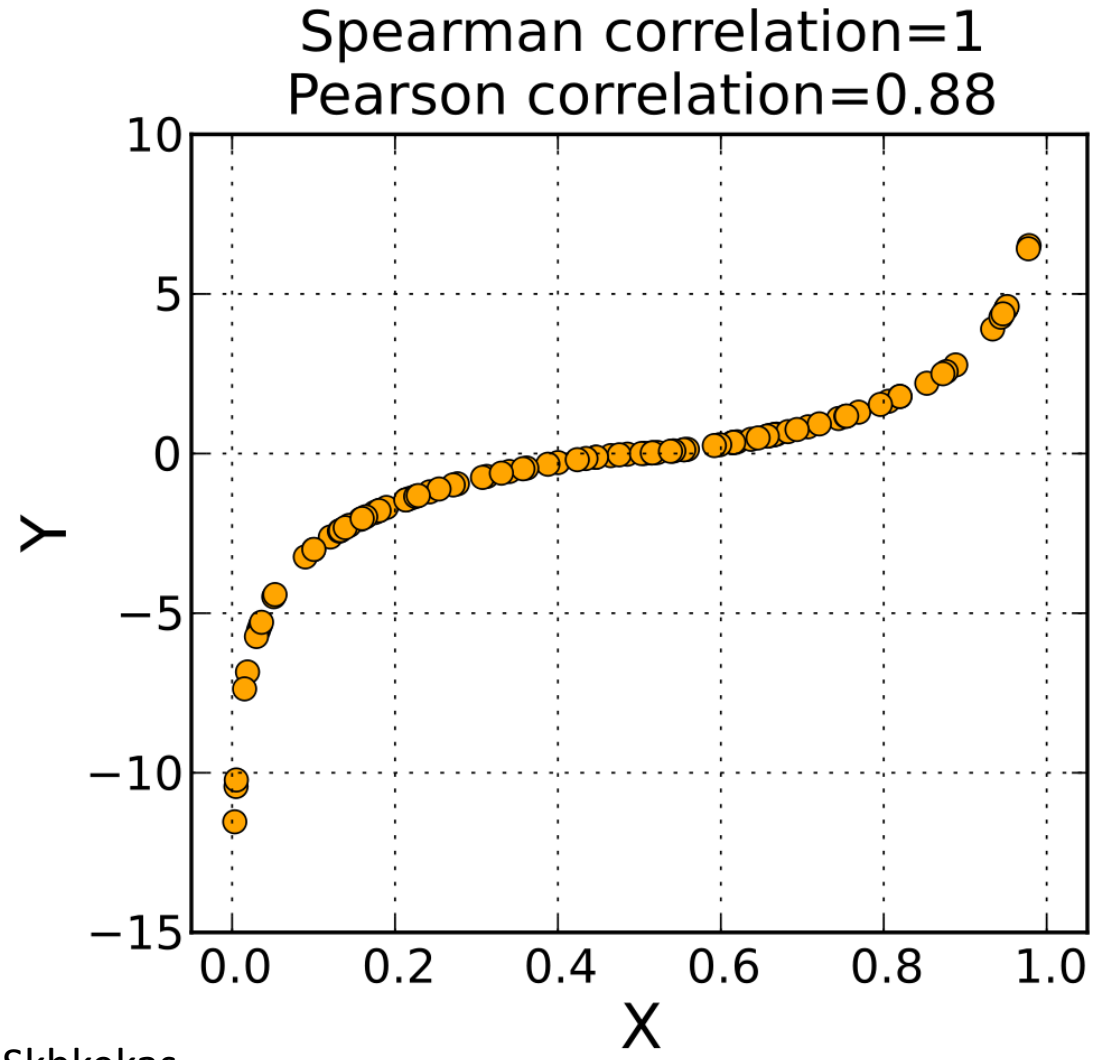
Spearman Rank Correlation

- Order Values x_i and y_i
- Replace the values by their ranks
- **Attention:** If there are multiple x_i und y_i with the same value their rank is set to their **average** rank.

Example:

x_i	y_i	Rank x_i	Rank y_i
0.2	1.0	1	1
0.5	11.34	3	3
0.5	12.4	3	4.5
0.5	8.1	3	2
1.73	12.4	5	4.5

Better!



Created by Skbkekas
(http://en.wikipedia.org/w/index.php?title=File:Spearman_fig2.svg&page=1)

Cosine Similarity

Cosine of the angle (γ) between the two data vectors:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}, \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}$$

$$\cos(\gamma) = \frac{\vec{x} \bullet \vec{y}}{\left|\vec{x}\right|\left|\vec{y}\right|}$$

Cosine Similarity & Pearson Correlation

- The correlation coefficient also corresponds to the cosine of the angle (γ) between the two (mean-subtracted) data vectors:

$$\vec{x}_{norm} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{pmatrix}, \vec{y}_{norm} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_{n-1} - \bar{y} \\ y_n - \bar{y} \end{pmatrix}$$

$$r_{xy} = \cos(\gamma_{norm}) = \frac{\vec{x}_{norm} \bullet \vec{y}_{norm}}{\left|\vec{x}_{norm}\right|\left|\vec{y}_{norm}\right|}$$

Group Work based on the BX_Books Dataset

Ziegler, Cai-Nicolas, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. "Improving recommendation lists through topic diversification." In Proceedings of the 14th international conference on World Wide Web, pp. 22-32. ACM, 2005.

http://www.informatik.uni-freiburg.de/~cziegler/BX/

Use the CSV (tested by me) or SQL Files

Task 1 (40 Points)

Deadline 31.10.2012, 11:00AM

- (10 Points) Calculate $\bar{x}, \tilde{x}$ and $\tilde{x}_{0.25}$ of the ratings of books "0316095648", "0971880107", and "0446610038". What is the message of those values?
- (10 Points) Calculate $\bar{x}$, s_x of Users "1903", "2033", and "2766". Compare the values? What do these values tell us?
- (20 Points) Find the most similar user to user "276688" with respect to his/her ratings calculated using Pearson correlation, Spearman correlation, Cosine similarity (calculate the values only based on mutually rated values). Evaluate only users that have at least **7** ratings with user **"276688"** in common.

Result

The result of your work is a document (one page, A4, PDF) which describes the algorithm for solution, the results, your interpretation and the responsabilities (who-did-what). This document has to be uploaded to the Mediacube Wiki before the Deadline.