Karlsruher Institut für Technologie

Geodatenanalyse I:

Schließende und Bivariate Statistik

Kathrin Menberg

Stundenplan

Karlsruher Institut für Technologie

Vorläufiger Stundenplan		
Datum	**Thema**	**Dozent**
20.10.2021	Einführung in die Programmierung mit *Python*	Gabriel Rau
25.10.2021	Univariate Statistik und statistisches Testen	Kathrin Menberg
01.11.2021	*Feiertag*	
08.11.2021	Variablen, Datentypen und Logik eines Programms	Gabriel Rau
15.11.2021	Bivariate und schließende Statistik	Kathrin Menberg
22.11.2021	Umgang und Berechnung von Datensätzen	Gabriel Rau
29.11.2021	Multivariate Statistik	Kathrin Menberg
06.12.2021	Datenvisualisierung mit *matplotlib*	Gabriel Rau
13.12.2021	Monte-Carlo Methoden	Kathrin Menberg
20.12.2021	Datenformate, Datenspeicherung und Datenbanken	Gabriel Rau
27.12.2021	*Weihnachtsferien*	
03.01.2022	*Weihnachtsferien*	
10.01.2022	Sensitivitätsanalyse	Kathrin Menberg
17.01.2022	Analyse und Visualisierung von Geodaten	Gabriel Rau
24.01.2022	Räumliche Interpolation	Kathrin Menberg
31.01.2022	Datenethik, Lizensierung und Entwicklungstools	Gabriel Rau
07.02.2022	Regressionsanalyse	Kathrin Menberg

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:20	**Schließende Statistik**
10:20 – 11:00	Übung
11:00 – 11:10	Diskussion und Reflexion
11:10 – 11:25	Pause
11:25 – 11:45	Bivariate Statistik
11:45 – 12:20	Übung
12:20 – 12:30	Diskussion und Reflexion

Lernziele Block 2.3

Am Ende der Stunde werden die Teilnehmer:

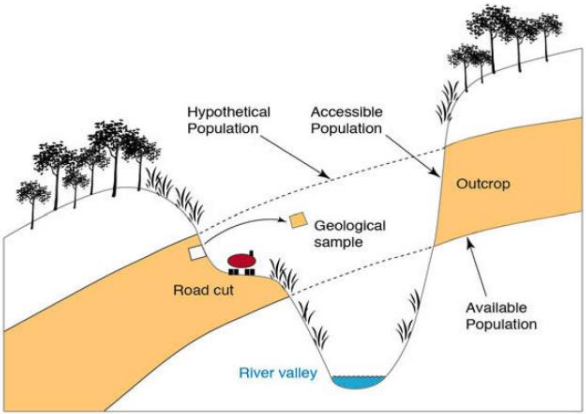
- … verschiedene theoretische Verteilungen und deren statistische Momente kennen.
- … Verteilungen an Datensätze anpassen und die Übereinstimmung bewerten und diskutieren können.
- … mit den Grundlagen der Wahrscheinlichkeitsrechnung vertraut sein.

Anknüpfung

- Übung 1: Charakterisierung von Stichproben anhand von statistischen Parametern

- … nun schauen wir uns die Verteilung der Grundgesamtheit an
- Annahme: $n \rightarrow \infty$

- Schließende Statistik
- Wahrscheinlichkeit



Trauth (2015) (Fig. 1.1)

Was ist Wahrscheinlichkeit?

- Relative Häufigkeit in Zufallsexperimenten
- Zufallsexperiment = Vorgang, der beliebig oft unter den gleichen Bedingungen wiederholbar ist
- … und dessen Ausgang nicht mit Sicherheit vorhergesagt werden kann

<u>Beispiel:</u>

- Medikament, dass bei 80% der Patienten wirkt
- Wahrscheinlichkeit der Wirkung bei zufällig herausgegriffenem Patienten $p = 0.8$

Rechnen mit Wahrscheinlichkeiten

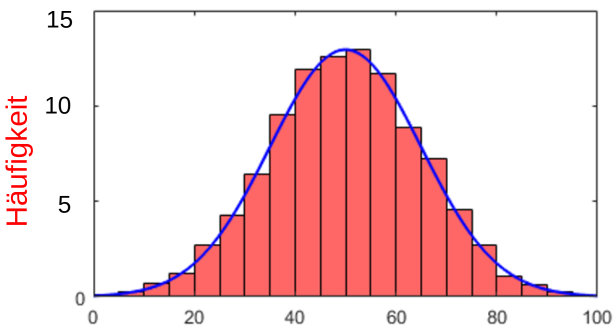
- Drei Grundregeln (nach A. Kolmogorov):
  - Wahrscheinlichkeit als reelle, nichtnegative Zahl: $1 \geq p(A) \geq 0$
  - Sicheres Ereignis hat Wahrscheinlichkeit 1: $p(S) = 1$
  - Wenn sich *A* und *B* ausschließen gilt: $p(A + B) = p(A) + p(B)$
- Bedingte Wahrscheinlichkeit (conditional probability)
  - Wahrscheinlichkeit für Ereignis *A*, unter der Bedingung ein Ereignis *B* sei eingetreten: $p(A|B)$
- Totale Wahrscheinlichkeit
  - Wahrscheinlichkeit für Ereignis *A* ergibt sich aus den bedingten Wahrscheinlichkeiten und den Wahrscheinlichkeiten dafür dass die Bedingungen eintreten: $p(\mathrm{A}) = \sum_i p(A|B_i)\, p(B_i)$

Wahrscheinlichkeitsverteilungen

- Deskriptive Statistik: Stichprobe → Messwert
- Schließende Statistik: Zufallsgröße → Wahrscheinlichkeit

Deskriptive Statistik

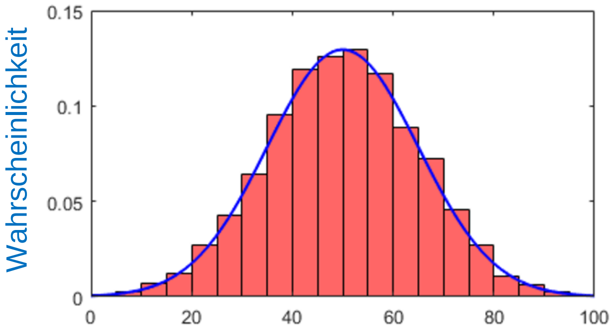
Häufigkeitsverteilung



Gesamtheit der Verteilung ergibt Anzahl der Stichproben

Schließende Statistik

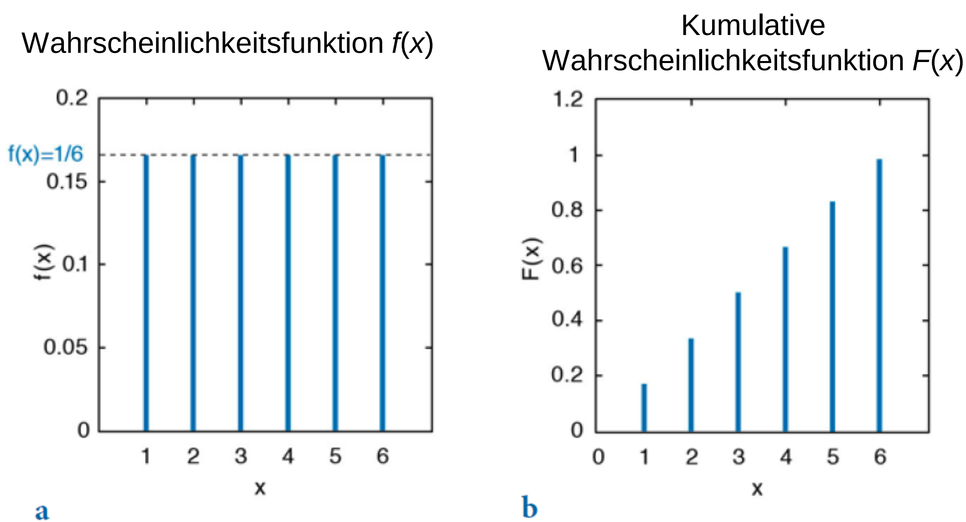
Wahrscheinlichkeitsverteilung



Gesamtheit der Verteilung ergibt Wahrscheinlichkeit $p = 1$

Theoretische Verteilungen

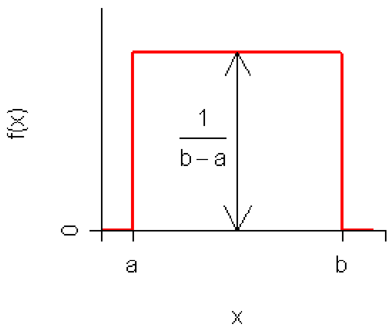
- Diskrete Werte: Wahrscheinlichkeitsfunktionen (probability mass function)
- Uniformverteilung, Gleichverteilung (Minimum, Maximum)



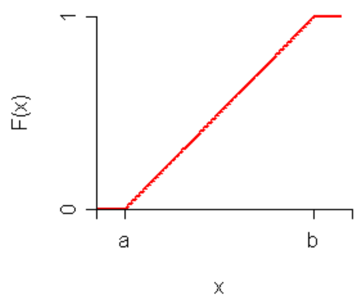
Theoretische Verteilungen

- Stetige Werte: Wahrscheinlichkeits<u>dichte</u>funktionen (probability density function)
- Uniformverteilung, Gleichverteilung (Minimum, Maximum)
  - *Uniform* (min, max), *U* (min, max)

Dichtefunktion *f*(*x*)



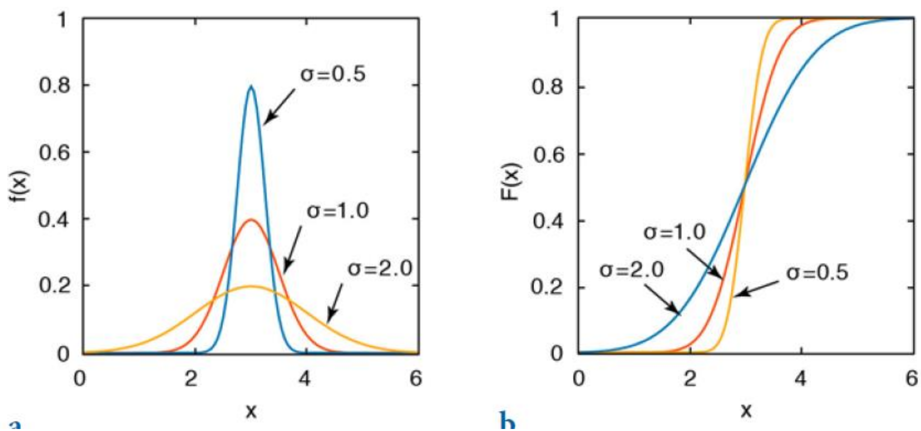
Verteilungsfunktion *F*(*x*)



Häufig verwendete Verteilungen

- Normal-, Gaußverteilung (Mittelwert, Varianz)
  - *Normal* (μ, σ^2), *N* (μ, σ^2)
  - Mean = Median = Mode
  - Skewness = 0
  - Kurtosis = 3
  - x ϵ (-∞, +∞)

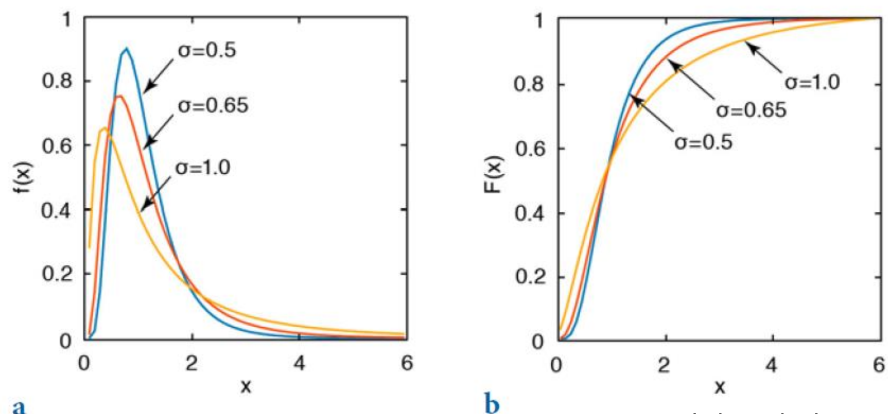
$\mu = 3$



Trauth (2015) Fig. 3.7

Häufig verwendete Verteilungen

- Log-Normalverteilung (Mittelwert μ_{logn}, Standardabweichung σ_{logn})
  - Mean ≠ Median ≠ Mode
  - Skewness > 0
  - x > 0

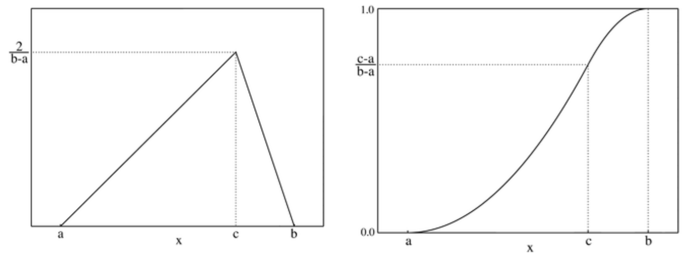


Trauth (2015) Fig. 3.7

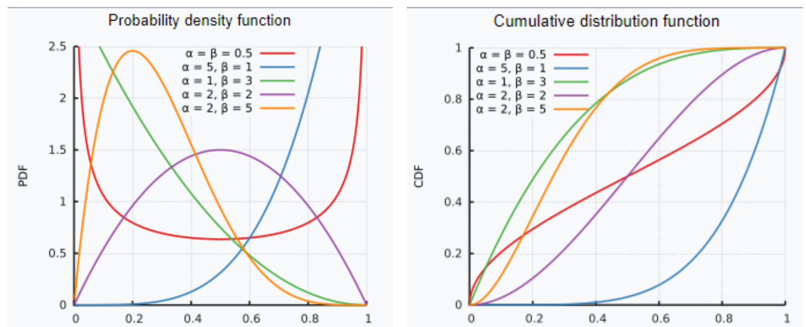
Häufig verwendete Verteilungen

Karlsruher Institut für Technologie

- Triangularverteilung (min, mode, max)
  - Mean ≠ Median ≠ Mode
  - x ϵ (min, max)



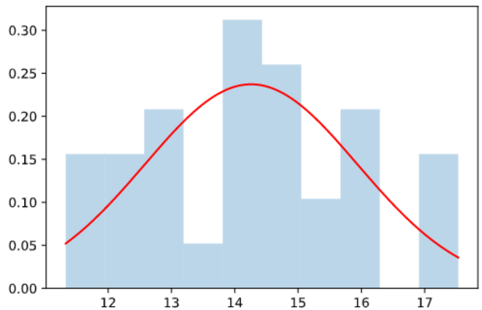
- Betaverteilung (α, β)
  - x ϵ (0, 1)

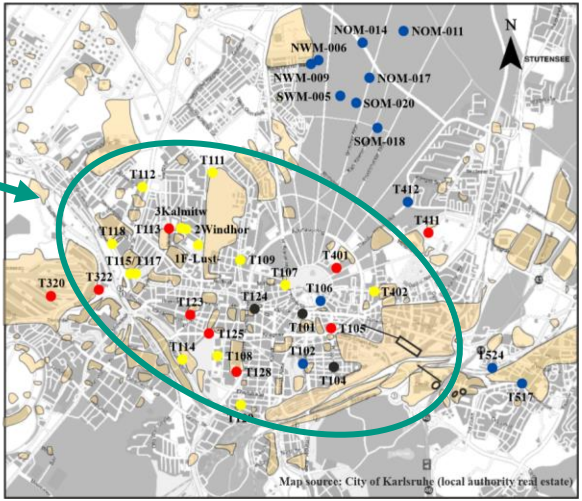


wikipedia.org

Übung 2.3: Schließende Statistik

- Grundwasserdatensatz Karlsruhe
  - Anpassung theoretische Verteilung an gemessene Stichproben





- Aufgaben in Jupyter Notebook: 03_Schliessende_Statistik_uebung

Aufgabenbesprechung

- Anpassung Normalverteilung an Grundwassertemperaturen
  - mean_fit = 14.19, variance_fit = 1.7
- Zufallswerte mit n = 50
  - mean_sample = 14.16, variance_sample = 2.9
- Zufallswerte mit n = 500,000
  - mean_sample2 = 14.19, variance_sample2 = 2.8
  - min = 6.5, max = 23.0
- Gestutzte Normalverteilung
  - lower_bound = 12, upper_bound = 18
  - Min = 11.33°C, max = 19.11°C

Karlsruher Institut für Technologie

Pause

… bis 11:25 Uhr

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:20	Schließende Statistik
10:20 – 11:00	Übung
11:00 – 11:10	Diskussion und Reflexion
11:10 – 11:25	Pause
11:25 – 11:45	**Bivariate Statistik**
11:45 – 12:20	Übung
12:20 – 12:30	Diskussion und Reflexion

Lernziele Block 2.4

Am Ende der Stunde werden die Teilnehmer:

- … mit den Konzepten von Kovarianz, Randverteilungen und Copulas vertraut sein.
- … verschiedene Korrelationskoeffizienten kennen und diese differenziert auf Geodaten anwenden können.

Karlsruher Institut für Technologie

Bivariate Statistik

- Meistens wird an Standorten mehr als nur eine Messgröße aufgenommen

- Bivariate Statistik untersucht die Art und Stärke der Beziehung zwischen **zwei** Messgrößen

- Nomenklatur:
  - x: unabhängige Variable
  - y: abhängige Variable

O_2
Temperatur

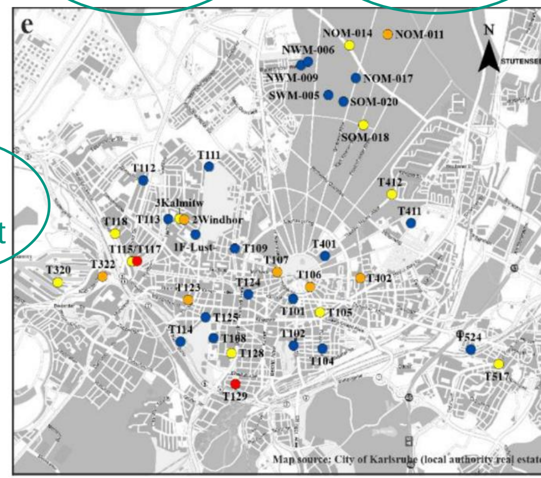
Eisen
Mangan

pH

Nitrat
Phosphat

Tiefe

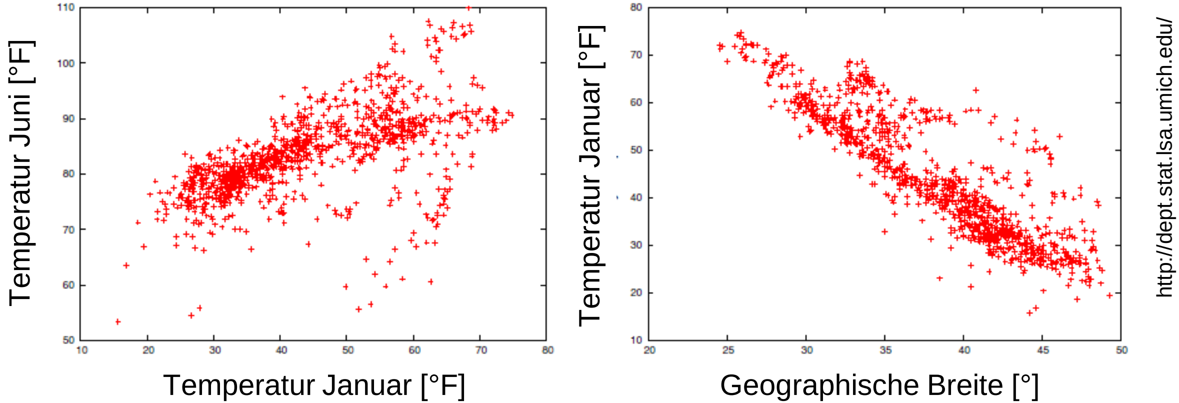
usw.



Koch et al. (2020)

Scatterplots

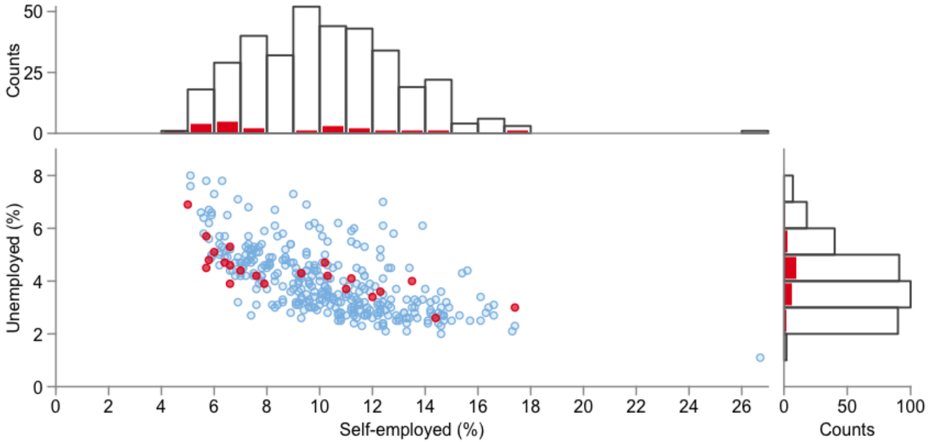
- Wichtigste graphische Zusammenfassung von bivariaten Daten



- Positive oder negativer Zusammenhang → Trend
- Keine Aussage über kausalen Zusammenhang möglich!

Randverteilungen

- Bivariate Datensätze haben gemeinsame (joint) und Randverteilungen (marginal distribution)

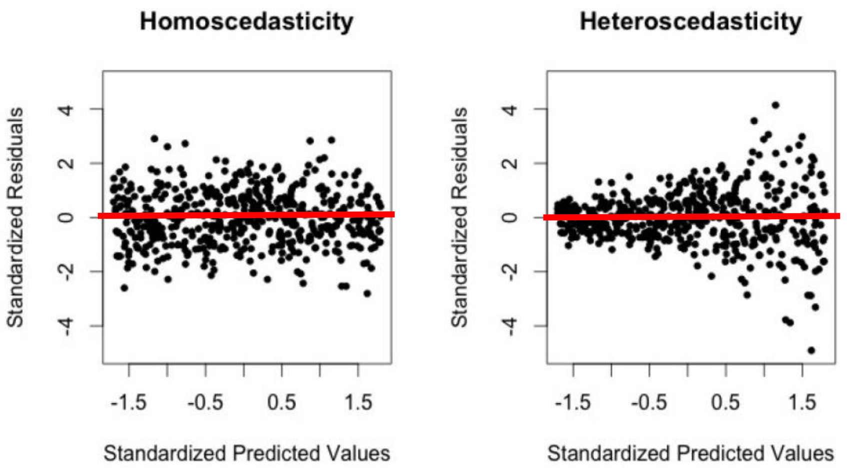


http://dept.stat.lsa.umich.edu/

- Das gleiche gilt für statistische Parameter, usw.

Trend ist nicht alles

▶ Trend: Zusammenhang der Mittelwerte



http://dept.stat.lsa.umich.edu/

▶ Heterogenität der Varianz: Heteroskedastizität (heteroscedasticity)

Varianz und Kovarianz

- Varianz (σ^2): Streuung eines Parameters

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

- Maß für die Assoziation zwischen zwei Zufallsvariablen
  - $\mathrm{cov}_{xy} > 0$: hohe (niedrige) Werte von *x*, gehen mit hohen (niedrigen) Werten von *y* einher
  - $\mathrm{cov}_{xy} < 0$: hohe (niedrige) Werte von *x*, gehen mit niedrigen (hohen) Werten von *y* einher
  - $\mathrm{cov}_{xy} = 0$: kein Zusammenhang zwischen *x* und *y*

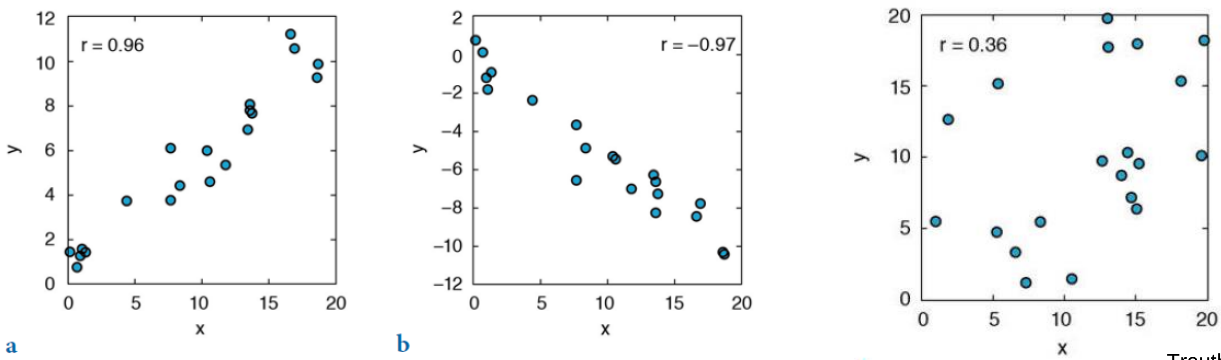
Korrelationskoeffizienten

Karlsruher Institut für Technologie

- Pearsons Korrelationskoeffizient

$$\rho = \frac{Kovarianz}{std\,(x)\ \cdot std\,(y)}$$

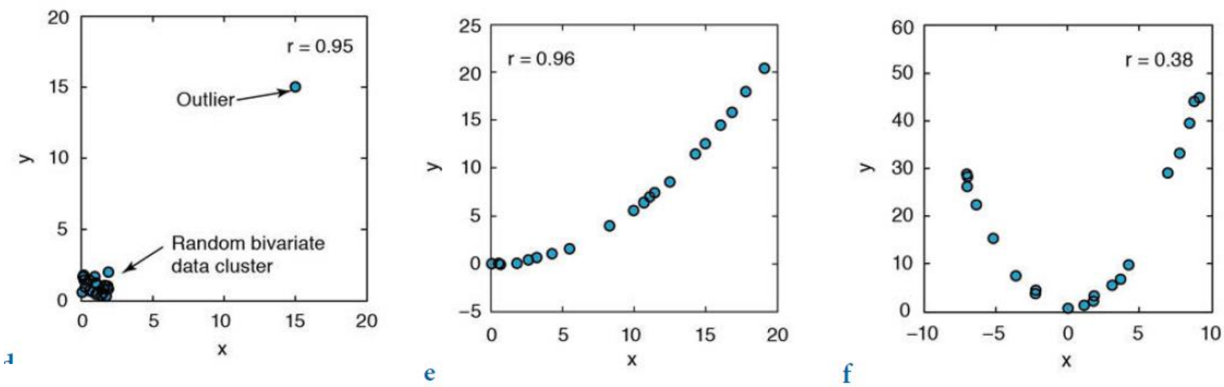
- Maß für die Stärke des linearen Trends von (*x*, *y*)



Trauth (2015)

Korrelationskoeffizienten

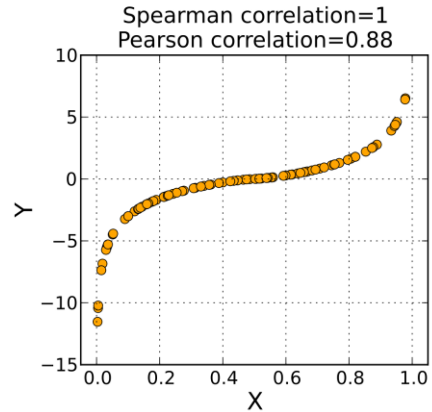
- Ausreißer und nicht-lineare Verbundenheit



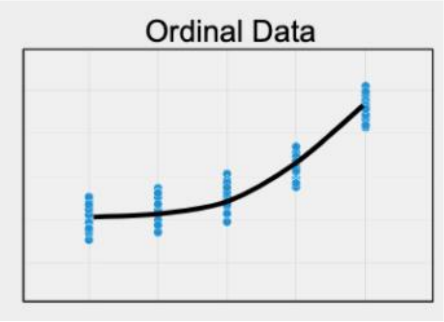
Trauth (2015)

Rang-Korrelationskoeffizienten

- Spearmans rho (ρ)
  - monotone Funktion
  - Statt Werte, Rang (ranking) der Daten



- Kendalls tau (τ)
  - Ähnlichkeit der Ränge von x und y
  - Robust gegenüber Ausreißern
  - Auch für ordinale Daten geeignet



www.wikipedia.org

Korrelationskoeffizienten

- Interpretation von Korrelationskoeffizienten:
  - Beschreibung der Stärke der Verbundenheit (association)

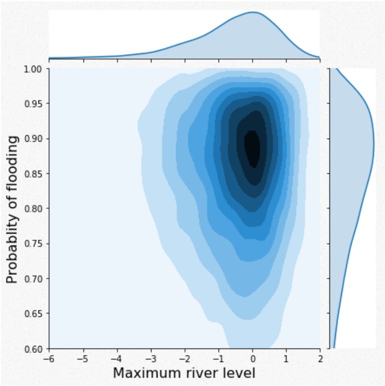
Perfekte Korrelation	ρ = ± 1
Hohe Korrelation	± 0.5 < ρ < ± 0.99
Moderate Korrelation	± 0.3 < ρ < ± 0.49
Niedrige Korrelation	0 < ρ < ± 0.29
Keine Korrelation	ρ = 0

  - Abhängige Formulation der Messgrößen wenn $-\frac{2}{\sqrt{n}} > \rho > \frac{2}{\sqrt{n}}$
  - Aussagekraft von Korrelationskoeffizienten mit *p*-Wert angeben.

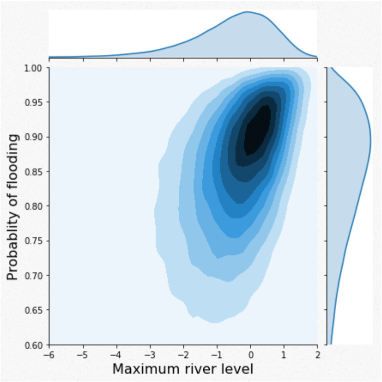
Copulas

- Randverteilungen von Zufallsvariablen die korrelieren
- Copula = „coupling function“ zwischen gemeinsamer Wahrscheinlichkeitsverteilung und Randverteilungen
- Erzeugen von zufälligen Wertepaaren mit Korrelation

ohne Korrelation



mit Korrelation



www.twiecke.io

Übung 2.4: Bivariate Statistik

Karlsruher Institut für Technologie

- Grundwasserdatensatz Karlsruhe
  - Graphische Darstellung
  - Quantifizierung der Beziehung zwischen einzelnen Parameter-Paaren
  - Kovarianzen
  - Korrelationskoeffizienten

- Aufgaben in Jupyter Notebook: geodatenanalyse_1-2-4

O_2
Temperatur

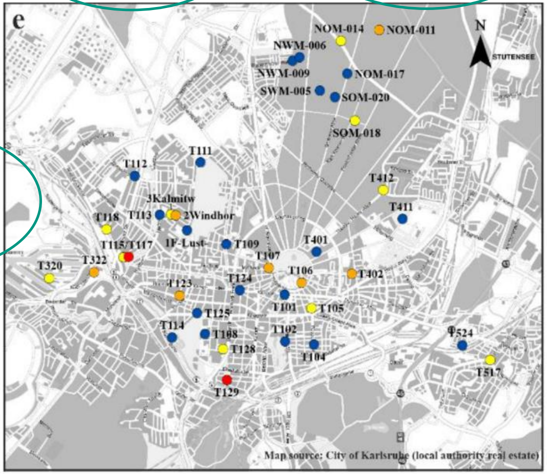
Eisen
Mangan

pH

Nitrat
Phosphat

Tiefe

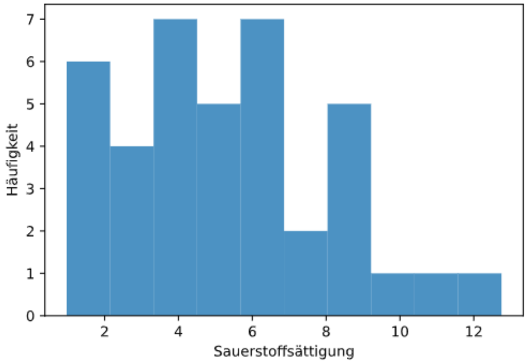
usw.

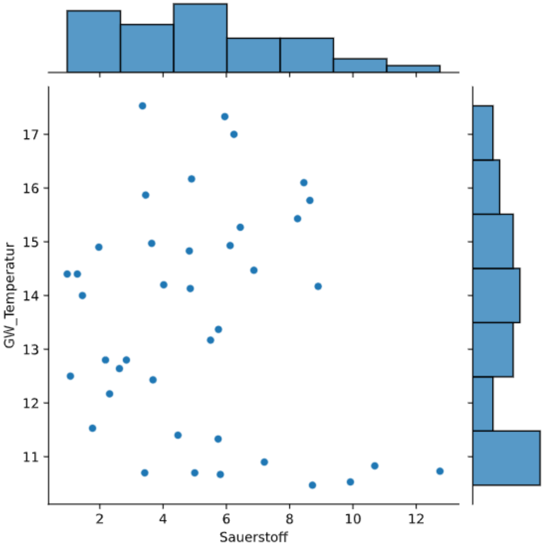


Koch et al. (2020)

Aufgabenbesprechung

- Visualisierung mit seaborn





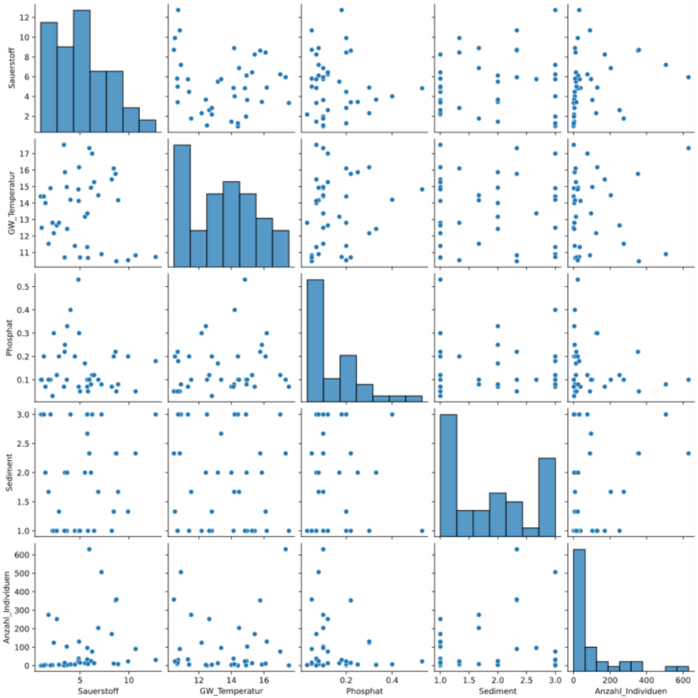
Aufgabenbesprechung

- Kovarianz (P_GWT) = 0.03
- Alle Kovarianzen:

```
cov_matrix = data.cov()
print(cov_matrix)

                  Sauerstoff  GW_Temperatur  Phosphat   Sediment  \
Sauerstoff          8.209471      -0.862375 -0.029748   0.141498
GW_Temperatur      -0.862375       4.440565  0.029929  -0.013659
Phosphat           -0.029748       0.029929  0.011623  -0.006913
Sediment            0.141498      -0.013659 -0.006913   0.648204
Anzahl_Individuen  97.744211      16.880526 -2.753158  17.569737

                  Anzahl_Individuen
Sauerstoff                97.744211
GW_Temperatur             16.880526
Phosphat                  -2.753158
Sediment                  17.569737
Anzahl_Individuen      22199.736842
```



Aufgabenbesprechung

- Korrelation O2_Individuen = 0.22
- Alle Korrelationen:

```
corr_matrix = data.corr()
print(corr_matrix)
```

```
                  Sauerstoff  GW_Temperatur  Phosphat  Sediment  \
Sauerstoff          1.000000      -0.142830 -0.096302  0.061339
GW_Temperatur      -0.142830       1.000000  0.131738 -0.008051
Phosphat           -0.096302       0.131738  1.000000 -0.079638
Sediment            0.061339      -0.008051 -0.079638  1.000000
Anzahl_Individuen   0.228960       0.053764 -0.171394  0.146466

                  Anzahl_Individuen
Sauerstoff                 0.228960
GW_Temperatur              0.053764
Phosphat                  -0.171394
Sediment                   0.146466
Anzahl_Individuen          1.000000
```

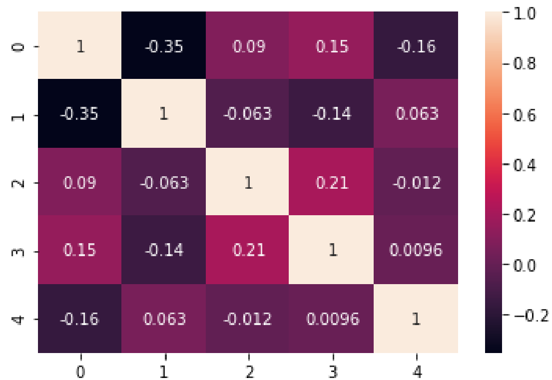
- Korrelation Pearson O2_GWT: r = - 0.14, p-Wert = 0.39
- Korrelation Spearman O2_GWT: r = - 0.06, p-Wert = 0.7

Aufgabenbesprechung

- Vergleich von drei Korrelationskoeffizienten

```
r_pear, p_pear = stats.pearsonr(data['Sediment'],data['Anzahl_Individuen'])
r_spear, p_spear = stats.spearmanr(data['Sediment'],data['Anzahl_Individuen'])
r_tau, p_tau = stats.kendalltau(data['Sediment'],data['Anzahl_Individuen'])
print (r_pear, p_pear, r_spear, r_tau, p_tau)
```

```
0.14646557616592315 0.3736111785965318 -0.07098288249384602 -0.05154355799732965 0.6701212825532348
```



Literatur

- Trauth (2015) MATLAB Recipes for Earth Sciences (4th Ed.), Springer
- Tschirk (2014) Statistik: Klassisch oder Bayes, Springer
- Koch et al. (2020) Groundwater fauna in an urban area: natural or affected?, Hydrology and Earth System Sciences Discussions

Nützliche Weblinks:

- Copluas: https://twiecki.io/blog/2018/05/03/copulas/


thank you
danke
gracias
merci
obrigado
spasibo in Cyrillic script: спасибо
teşekkür ederim
go raibh maith agat
mochchakkeram
dziękuję
bedankt
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
dank je
tapadh leat
takk
hvala
mauruuru
sagolun
мерси
ありがとう
감사합니다
謝謝
ευχαριστώ