Stochastik-Praktikum

Deskriptive Statistik

Peter Frentrup

Humboldt-Universität zu Berlin

7. November 2017

Übersicht

Kenngrößen univariater Verteilungen

Endlichdimensionale Kenngrößen:

- Mittelwert $\mathbb{E}[X]$
- Median $\frac{1}{2}(F^{-1}(0.5) + F^{-1}(0.5-))$
- Quantile $Q_\alpha = F^{-1}(a) = \inf\{x \mid F(x) \geq \alpha\}$ für $\alpha \in (0,1)$
- Varianz und Standardabweichung $\sigma^2 = \mathbb{V}\text{ar}(X)$ bzw. $\sigma = \sqrt{\mathbb{V}\text{ar}(X)}$
- allgemeiner: (un-)zentrierte Momente $\mathbb{E}[X^k]$ bzw. $\mathbb{E}[(X - \mathbb{E}X)^k]$ für $k \in \mathbb{N}$
- Schiefe (skewness) $\mathbb{E}[(X - \mathbb{E}X)^3]/\sigma^3$,
  Wölbung (kurtosis) $\mathbb{E}[(X - \mathbb{E}X)^4]/\sigma^4$,...

Kenngrößen univariater Verteilungen

Die unendlich-dimensionalen Kenngrößen

- (kumulative) Verteilungsfunktion $F(x) = P[X \leq x]$, $x \in \mathbb{R}$
- charakteristische Funktion $\varphi(t) = \mathbb{E}[\exp(itX)]$

beschreiben univariate Verteilungen vollständig.

Kenngrößen für Stichproben

Typischerweise auf eine von zwei Weisen berechnet:

- wie zuvor, bloß bezüglich der empirischen Verteilung $\frac{1}{n}\sum_1^n \delta_{x_i}$ der Stichprobe,
- ggf. modifiziert zu einem erwartungstreuem Schätzer der entsprechenden Kenngröße der zugrundeliegenden theoretischen Verteilung.

Kenngrößen für Stichproben

Beispiele

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

  $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (nicht erwartungstreu)
- Schiefe $\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$ (erwartungstreu)
- Median aus der Ordnungsstatistik...

Kenngrößen multivariater Verteilungen

- Mittelwert und Median (Vektoren)
- gemischte (zentrierte) Momente
- Kovarianzmatrix (alle zentrierten Momente der Ordnung 2)
- gemeinsame charakteristische Funktion oder Verteilungsfunktion
- Randverteilungen und *Copula* der gemeinsamen Verteilung

Letztere beschreiben Verteilung eindeutig.

Übersicht

Visualisierungen

... von uni- bzw. multivariaten Verteilungen/Stichproben in $\mathbb{R}^d$:

- Dichtefunktion bzw Histogramm (skaliert auf PDF-Form)
- Regressionsgeraden, -polynome, ...
- Quantil/Quantil-Plots (QQ-Plots)
- Scatterplot, Copula-Scatterplot

Für Beispielcode siehe auch *multinorm.py* (Folien aus Block 1)

Visualisierungen

... von uni- oder multivariaten kategorialen Datenstichproben:

- Kuchendiagramm
- Balkendiagramme...
- Kontingenztafeln

Beispiele: Studienfächer, Nationalitäten, Geschlecht, ...

Übersicht

Beispiel: Median vs. Mean

Untersuche Verteilung von Mean und Median für Stichprobe von je 50 Elementen von $X \sim \mathcal{N}(0,1)$, $Y = X^5$

```
import numpy as np

for k in range(0,5):
    print("Experiment %d:" % k)
    X = np.random.normal(size=50)
    Y = X**5
    print([[np.mean(X), np.median(X)], [np.mean(Y), np.median(Y)]])
```

Beobachtung? Erklärung??

Visualisieren Sie, z.B. mit

- `numpy.histogram` + `matplotlib.pyplot.bar`,
- `matplotlib.pyplot.boxplot`,
- `scipy.stats.probplot`, ...

Übersicht

Univariate Daten

Michelsons Lichtgeschwindigkeits-Daten

Lichtgeschwindigkeit $c \approx s + 299\,000$ km/s. ↪ A. A. Michelson

Nr.	Wert s	Durchlauf	Experiment
1	850	1	1
2	740	2	1
3	900	3	1
4	1070	4	1
5	930	5	1
6	850	6	1
7	950	7	1
⋮	⋮	⋮	⋮
98	800	18	5
99	810	19	5
100	870	20	5

Interpretation der Daten

Nr.	Wert s	Durchlauf	Experiment
1	850	1	1
2	740	2	1
3	900	3	1
4	1070	4	1
⋮	⋮	⋮	⋮

Erste Spalte: Fortlaufende Nummer der Messungen (1-100)
Zweite Spalte: (Gemessene Geschwindigkeit - 299 000) in km/s
Dritte Spalte: Fortlaufende Nummer in der Messreihe (1-20)
Vierte Spalte: Nummer der Messreihe (1-5)

Einlesen der Daten

```
michelson = np.loadtxt("michelson.dat", skiprows=1)
(number, speed, run, experiment) = michelson.T

# oder alternativ:
number = michelson[:,0]
speed = michelson[:,1]
run = michelson[:,2]
experiment = michelson[:,3]
```

Auswahl der ersten Messreihe

```
indices = experiment == 1
s = speed[indices]
```

Statistische Kenngrößen

(Arithmetischer) Mittelwert $\bar{x} = \sum_{i=1}^{n} x_i$:

```
>>> np.mean(s)
909.0
```

Standardabweichung (erwartungstreu) $\sqrt{(1/(n-1))\sum_i (x_i - \bar{x})^2}$:

```
>>> np.std(s, ddof=1)
104.926039114
```

Median $med(\mathbf{x}) = x_{((n+1)/2)}$:

```
>>> np.median(s)
940
```

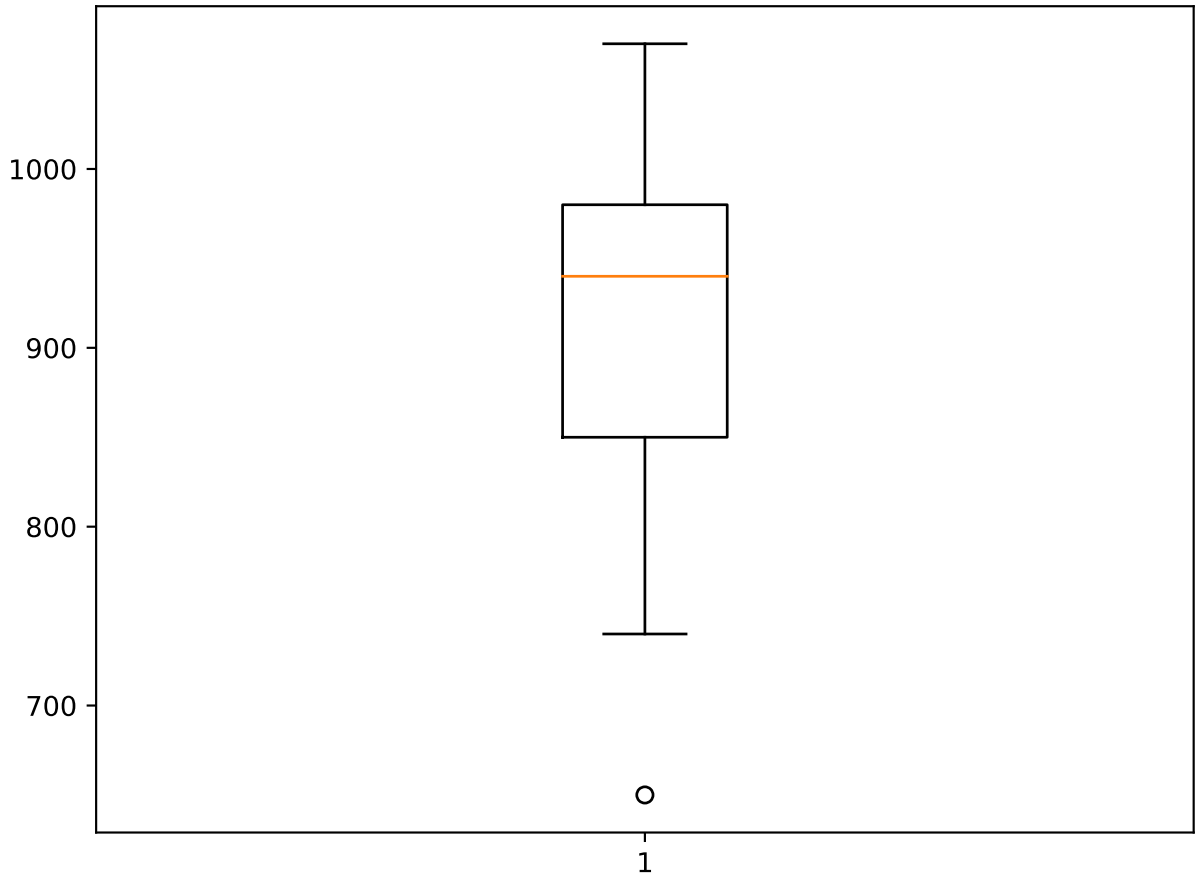
Mittel absoluter Abweichungen zum Median $MAD = n^{-1}\sum_i |x_i - med(\mathbf{x})|$:

```
>>> np.mean(np.absolute(s - np.median(s)))
79.0
```

Achtung: MAD auch "median absolute deviation": $med\big((|x_i - med(\mathbf{x})|)_i\big)$
sowie andere "mean absolute deviation"s: $n^{-1}\sum_i |x_i - \bar{x}|$

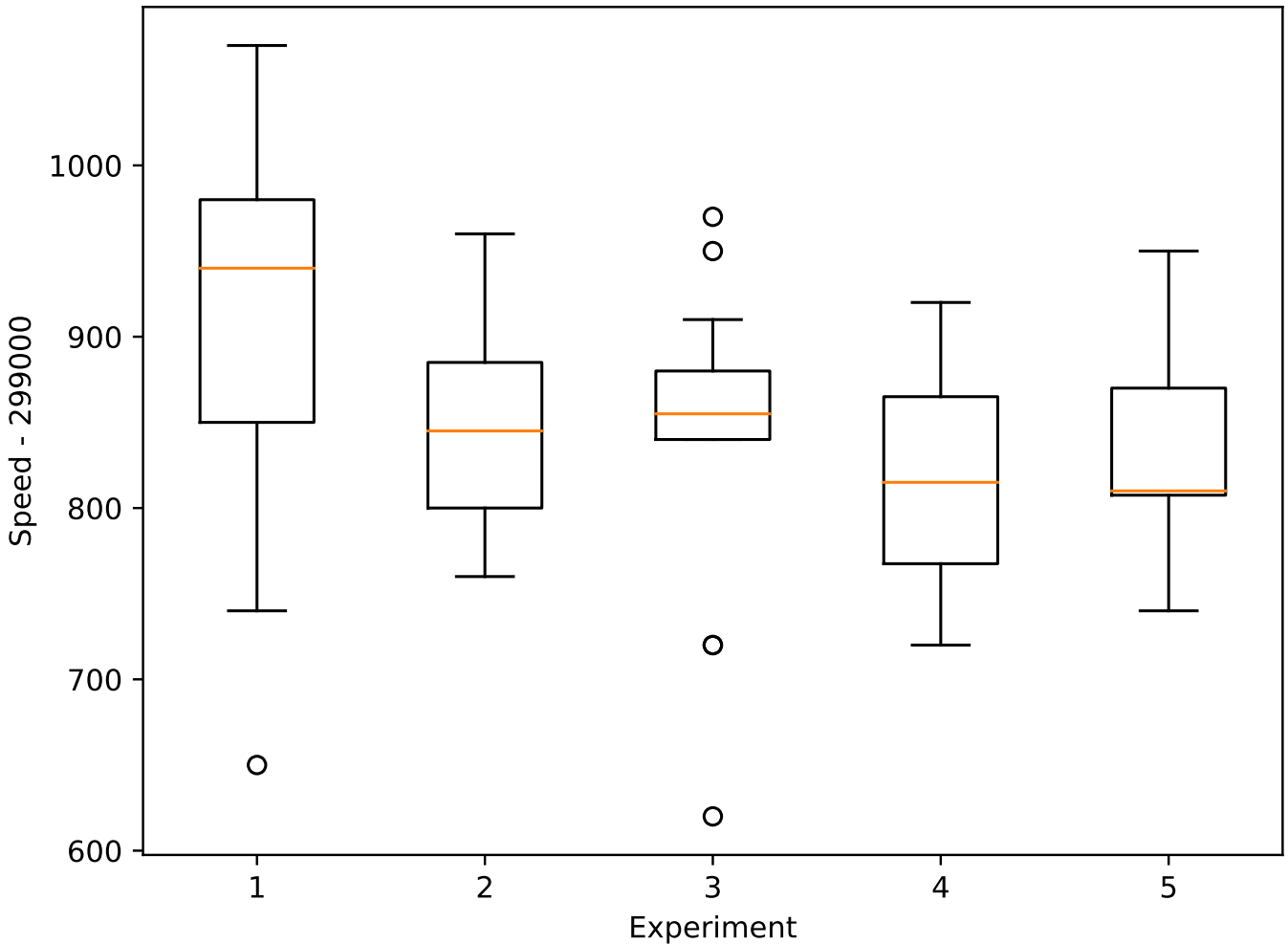
Der Box–Whisker–Plot

```
>>> matplotlib.pyplot.boxplot(s)
```



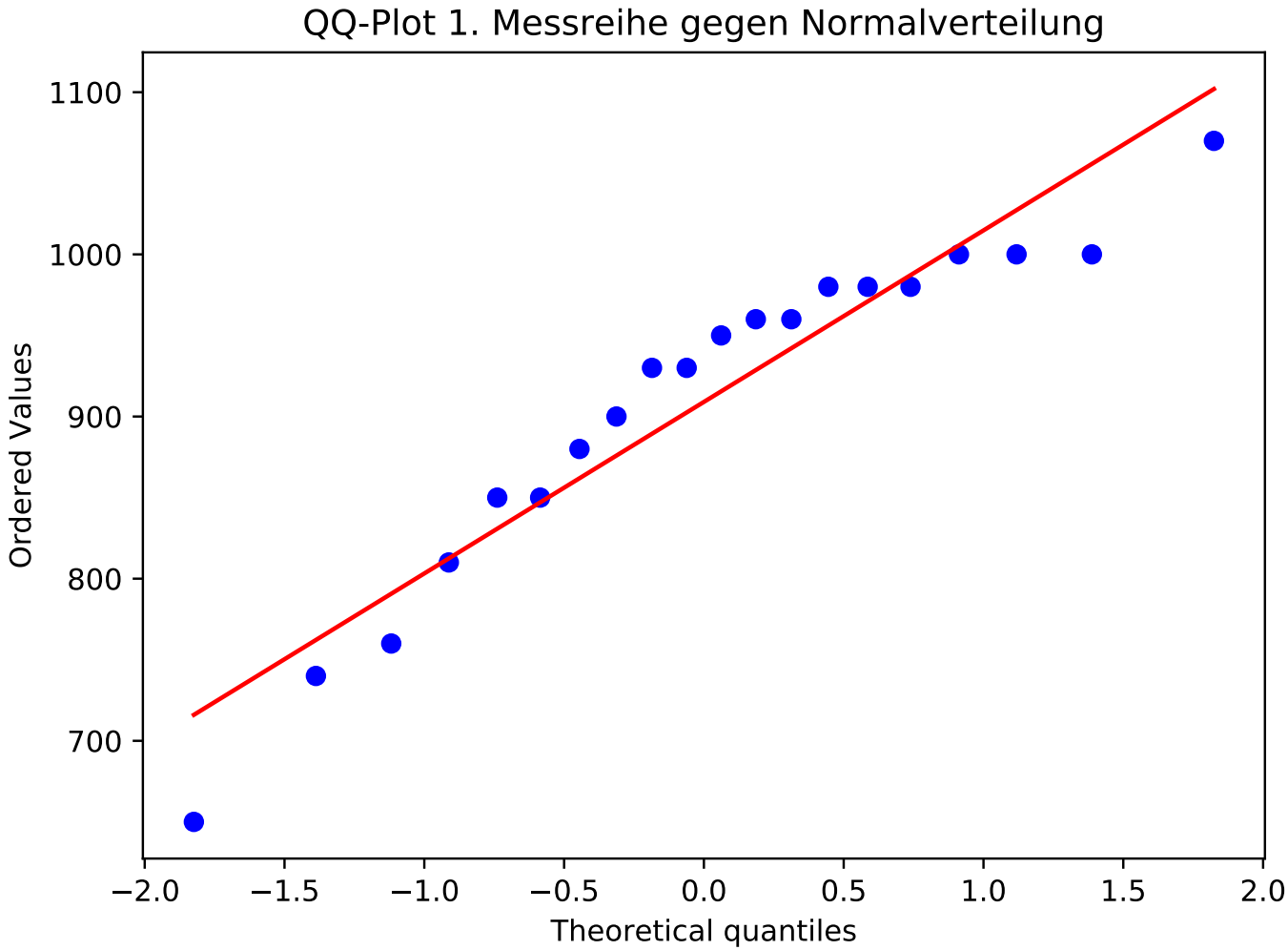
Box-Plots der verschiedenen Messreihen

```
plt.figure()
plt.boxplot([speed[experiment==i] for i in [1,2,3,4,5]])
plt.xlabel("Experiment")
plt.ylabel("Speed - 299000")
```



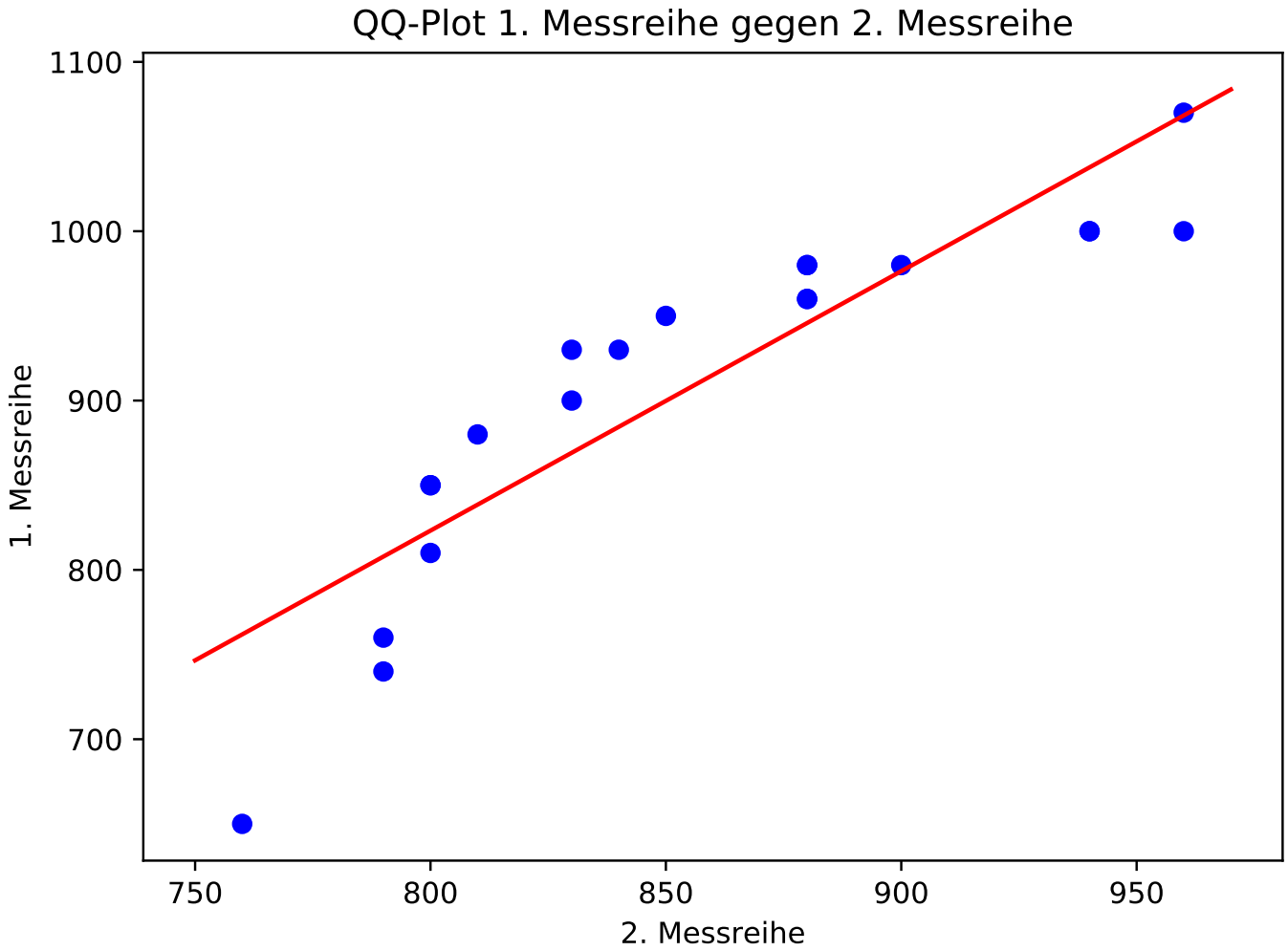
Vergleich verschiedener Messreihen

QQ-Plot von Quantilen 1-ter Messreihe gegen Quantile einer $\mathcal{N}(0,1)$-Verteilung



Vergleich verschiedener Messreihen

QQ-Plot von Quantilen 1-ter Messreihe gegen Quantile 2-ter Messreihe



Multivariate Daten

Mietspiegel–Daten

Einlesen der Daten

```
miete = np.genfromtxt('miete03p.csv',delimiter='\t',skip_header=1)

# TODO: bessere Namen
(GKM, QMKM, QM, Zi, BJ, B, L, best, WW, ZH, BK, BA, KUE) = miete.T
```

Abgeleitete Variablen

Hier: Klassfizierung von Baujahr und Quadratmeterzahl

```
# Einsortieren in Baujahr-Klasse 1-6
BJKL = 1 + (BJ > 1918) + (BJ > 1948) + (BJ > 1965) \
     + (BJ > 1977) + (BJ > 1983)

# Einsortieren in Groessenklasse 1-3
QMKL = 1 + (QM > 50) + (QM > 80)
```

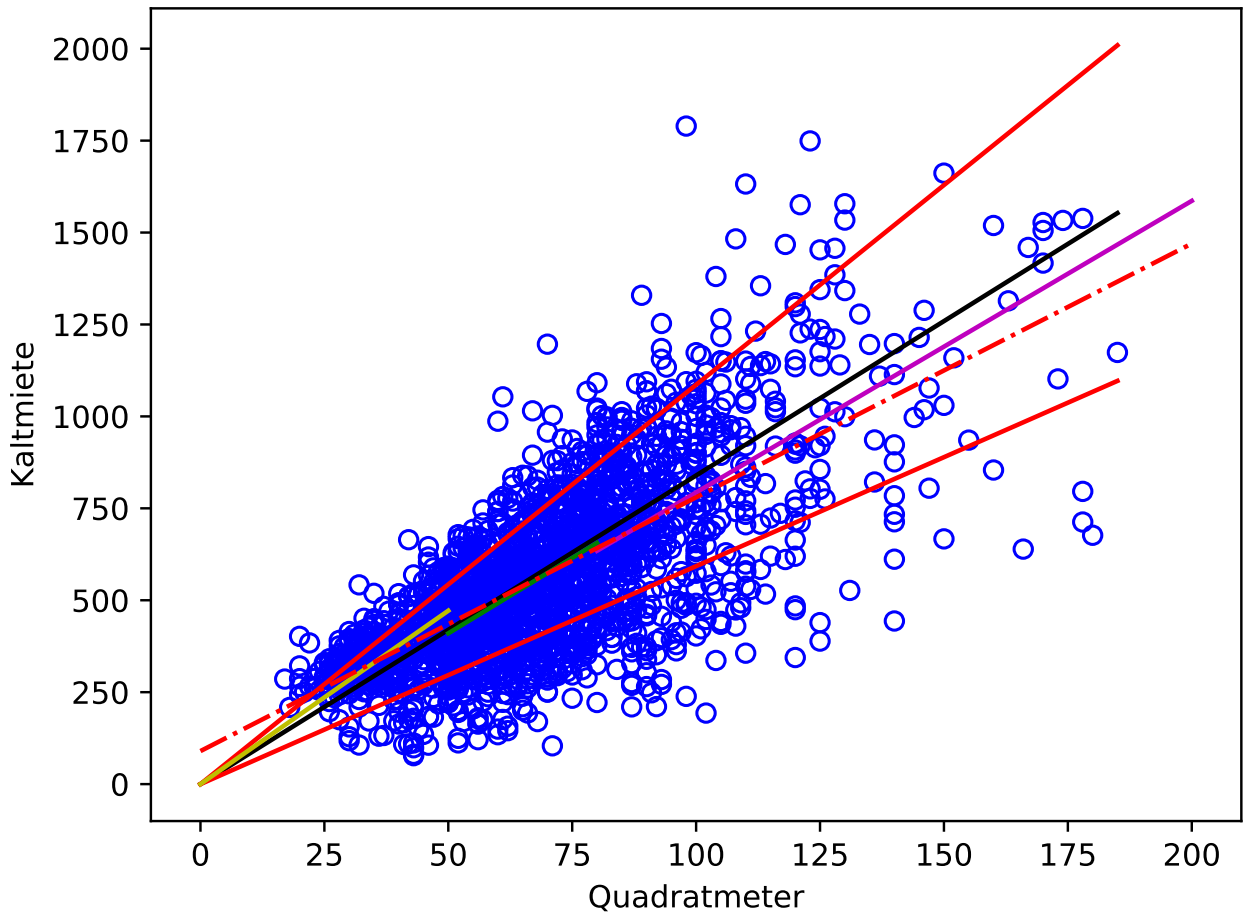
Zusammenhänge zwischen Variablen

Darstellung linearer Zusammengänge : Regressionsgraden

```
## Regressionsgerade
p = np.poly1d(np.polyfit(QM, GKM, 1))
x = [0, 200]
y = p(x)
plt.plot(x, y, '-.r')
plt.show()
```

lineare Zusammenhänge

Regressionsgrade, (mittlere QM-Miete±1Std)*QM, QMKL-weise...



Visualisierung von Kontingenztafeln

Miete $\leftrightarrow$ Wohnlage

Kontingenztafel der QM- u BJ-Klassen bestimmen und mit matplotlib . pyplot . bar() aufbauen...