# Multivariate Analysemethoden


q-q-Plot Methode zur Prüfung
der Multivariaten Normalverteilung

***Günter Meinhardt***
***Johannes Gutenberg Universität Mainz***

**Prüfung der Verteilungs-annahme**

# Verteilungsanpassung/Prüfung

- Ausreißeranalyse:
  Vor der Schätzung der Parameter ($\mu$,$\Sigma$) für die multivariate NV- wird eine Analyse der Rohdaten auf Ausreißer vorgenommen.

- Effiziente Tests:
  Die NV- Annahme ist mit effektiven Methoden und trennscharfen Test zu prüfen, um ihre Gültigkeit sicherzustellen

- Korrekturen und Datentransformationen:
  Ist die NV- Annahme auf den originalen Skalen verletzt, können Skalentransformationen für die einzelnen Variablen des Variablen-verbundes gefunden werden, mit denen die multivariate Normalver- auf den transformierten Skalen gilt.

**Prüfung der Verteilungs-annahme & Outlier-Identifikation**

# Mahalanobisdistanz Δ

- Kernkonzept der Ausreißer-Identifikation
  Der Abstand einer Beobachtung vom Schwerpunkt der Verteilung wird über die **multivariate Distanz** Δ bestimmt. Dabei werden stets die quadrierten Distanzen $\Delta^2$ verwendet, da diese Chi-Quadrat verteilt sind, wenn die Variablen einer multivariaten Normalverteilung entstammen.

Dann definiert

$$\Delta^2 = (\vec{x} - \vec{\mu})^t \, \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu})$$

mit $\mathbf{\Sigma}^{-1}$ die Inverse der Varianz-Kovarianz Matrix **Σ,** die verallgemeinerte quadrierte Distanz im multivariaten Raum. Sie heißt quadrierte Mahalanobis-Distanz.

**Data Clearing**
p-dimensions

# Identifikation von Ausreißern

- Auch im multivariaten Fall sind Ausreißer in kleinen Stichproben nicht zuverlässig bestimmbar,
- Bei N > 30 legt man die Quantile der multivariaten Normalverteilung zugrunde ($\chi^2$) und eliminiert die Beobachtungen, dessen quadrierte Mahalanobis-Distanzen jenseits der äußeren Quantile liegen. Dies sollten nicht mehr als 7%-8% sein.

$$p_i = \frac{i-0.5}{N},\ i = 1,\ldots,N \Rightarrow \frac{N-0.5}{N} = 1-\frac{1}{2N} = p_{\max}$$

$$q_e = \chi^2_p(p_i) \Rightarrow \chi^2_p(p_{\max}) = \Delta^2_{\max}$$

| *p* | 0.02 | 0.05 | 0.08 | 0.12 | 0.15 | 0.18 | 0.22 | 0.25 | 0.28 | 0.32 | 0.35 | 0.38 | 0.42 | 0.45 | 0.48 | 0.52 | 0.55 | 0.58 | 0.62 | 0.65 | 0.68 | 0.72 | 0.75 | 0.78 | 0.82 | 0.85 | 0.88 | 0.92 | 0.95 | 0.98 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *χ2* | 0.39 | 0.71 | 0.95 | 1.17 | 1.37 | 1.56 | 1.74 | 1.92 | 2.1 | 2.29 | 2.47 | 2.66 | 2.85 | 3.05 | 3.25 | 3.46 | 3.69 | 3.92 | 4.17 | 4.44 | 4.73 | 5.04 | 5.39 | 5.77 | 6.22 | 6.74 | 7.39 | 8.24 | 9.49 | 12.09 |
| *Δ2* | 0.13 | 0.48 | 0.62 | 0.79 | 0.82 | 1.12 | 1.41 | 1.45 | 1.52 | 1.54 | 2 | 2.3 | 2.46 | 2.63 | 2.67 | 2.79 | 3.1 | 3.52 | 3.62 | 4.13 | 4.73 | 5.16 | 5.23 | 5.39 | 5.67 | 6.5 | 7.88 | 10.24 | 12.69 | 17.43 |

Ausreißer:

$$\Delta^2 > \chi^2_p(p_{\max})$$

[Excel-Beispiel q-q-Plot]

**Q-Q Plot**
**Methode**
**multivariat**

# Test über Quantilskorrelation

- Nach Ausreißerbereinigung werden den Meßvektoren empirische Quantile $q_o$ zugeordnet über die Reihe der Meßwerte sortiert nach Mahalanobisdistanz.
- Mit aus den Daten geschätzten Parametern ($\mu$,$\Sigma$) werden für die linearen Prozentränge erwartete Quantile $q_e$ aus der $\chi^2$ - Verteilung bestimmt.
- Man trägt $q_o$ (y-Achse) und $q_e$ (x-Achse) gegeneinander ab. Perfekte Passung liegt vor, wenn die Daten auf der Winkelhalbierenden liegen.
- Man bestimmt Anteil der aufgeklärten Varianz und Korrelation.

$$r_{qq} = \frac{\mathrm{cov}(q_o, q_e)}{\sqrt{\mathrm{var}(q_o) \cdot \mathrm{var}(q_e)}} \qquad \eta^2 = 1 - \frac{\sum_i (q_{oi} - q_{ei})^2}{\sum_i (q_{oi} - \bar{q}_o)^2} = 1 - \frac{\sum_i \left(\Delta_i^2 - \chi_p^2(p_i)\right)^2}{\sum_i \left(\Delta_i^2 - \overline{\Delta^2}\right)^2}$$

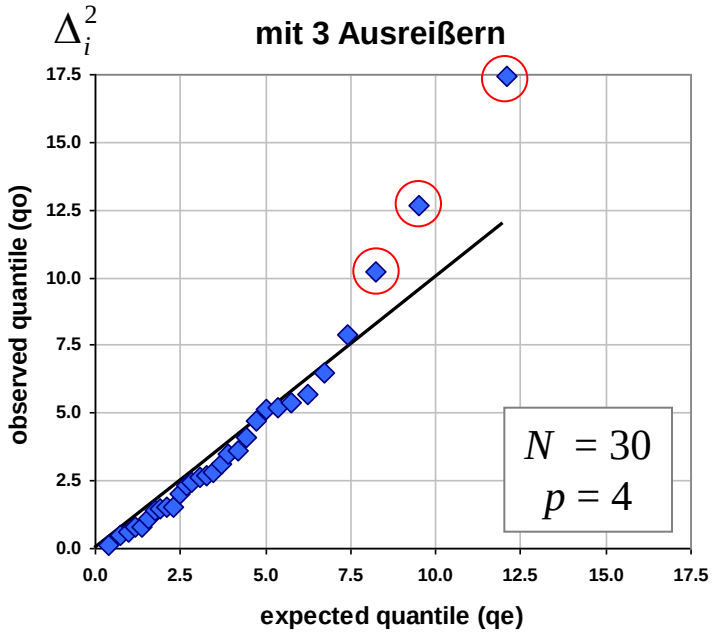
- Für den Test des Korrelationskoeffizienten verfährt man exakt wie im univariaten Fall.

**Q-Q Plot Methode**

# Datenbeispiel (p = 4 Variablen)

$$\eta^2 = 1 - \frac{\sum_i (q_{oi} - q_{ei})^2}{\sum_i (q_{oi} - \bar{q}_o)^2} = 0.889$$

$$r = \frac{\mathrm{cov}(q_o, q_e)}{\sqrt{\mathrm{var}(q_o) \cdot \mathrm{var}(q_e)}} = .979$$




$\chi_p^2(p_i)$

**Korrelations-Test**

$.979 > .9715 \Rightarrow r_{qq} > r_{crit(\alpha=.1)}$

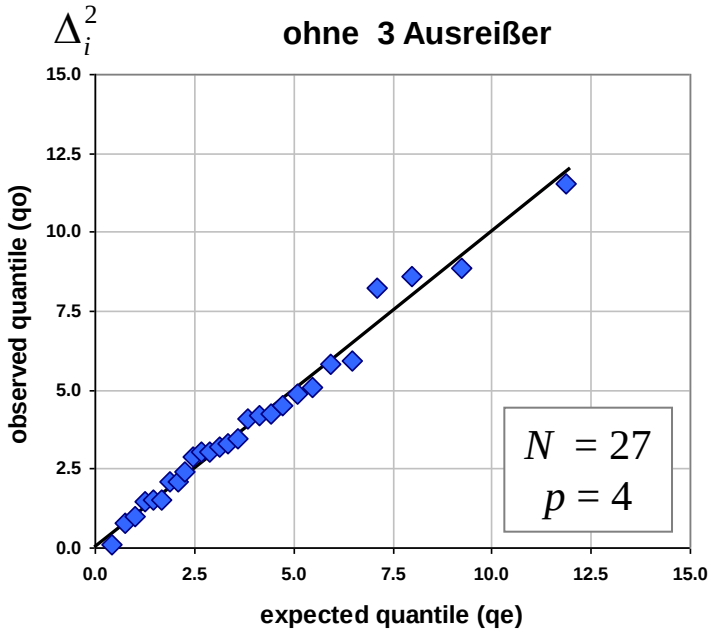
- NV Test knapp im Annahmebereich, aber 2 Ausreißer verschlechtern die Passung beträchtlich, auch in den unteren Quantilen
- Die beiden größten Ausreißer erfüllen das Kriterium, aber der 3. höchste Wert ist ebenfalls suspekt (hoher Intervallabstand)

**Q-Q Plot Methode**

# Datenbeispiel (p = 4 Variablen)

$$\eta^2 = 1 - \frac{\sum_i (q_{oi} - q_{ei})^2}{\sum_i (q_{oi} - \bar{q}_o)^2} = 0.984$$

$$r = \frac{\mathrm{cov}(q_o, q_e)}{\sqrt{\mathrm{var}(q_o) \cdot \mathrm{var}(q_e)}} = .992$$




$$\chi_p^2(p_i)$$

**Korrelations-Test**

$$.992 > .9715 \Rightarrow r_{qq} > r_{crit(\alpha=.1)}$$

NV Test zeigt nach Entfernung der höchsten 3 Werte (nicht nur 2) nun eine gute Passung der multivariaten NV

Ausreisser-Kontrolle

## Allgemeines zur Verteilungskorrektur

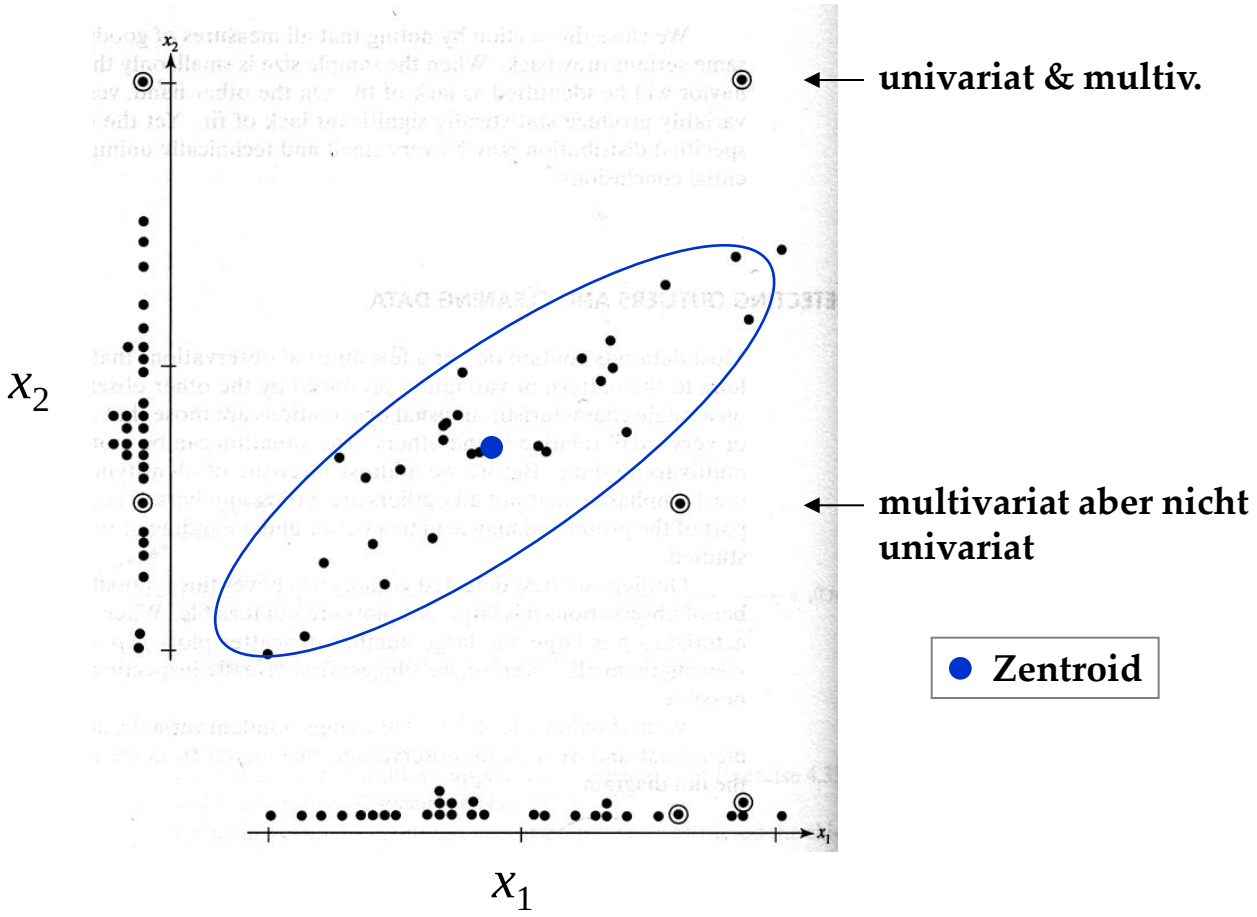
- Ausreißerbereinigung sollte immer multivariat erfolgen, da ein Ausreisser in einer einzelnen Variable noch nicht einen Ausreißer im Variablenverbund definiert.
- Das Entfernen extremer Beobachtungen ändert die Korrelationsmatrix, daher können iterative Bereinigungen nötig werden.

Transformationen

## Skalentransformationen

- Skalentransformationen können nur univariat erfolgen. (Keine Methode definiert eine Transformation für den Variablenverbund)
- Es ist ratsam eine univariate Untersuchung systematischer Verteilungsabweichungen nach der multivariaten Ausreißerkontrolle durchzuführen, und die einzelnen Variablen mit geeigneten Potenz-transformationen zu korrigieren.
- Sind die Randverteilungen (univariate) alle normal, so wird auch die multivariate Verteilung normalverteilt sein.

## Ausreißer als Distanz vom Zentroid




**Ausreisser in 2D: einer univariat & multivariat und einer multivariat**

# Ausreißer als Distanz vom Zentroid

**Ausreisser in 4D: einer uni+multi und zwei multivariat**

| Z1 | Z2 | Z3 | Z4 | D2 | |
|---|---|---|---|---|---|
| -0.05 | -0.31 | 0.17 | 0.17 | 0.62 | |
| 1.56 | 0.95 | 1.94 | 1.49 | 5.67 | |
| 0.67 | -0.16 | 1.03 | 1.57 | 7.88 | |
| -0.82 | -0.39 | -1.34 | -0.60 | 5.39 | |
| 0.22 | 0.53 | 0.35 | 0.50 | 1.45 | |
| -0.61 | -0.12 | -0.24 | -0.56 | 2.30 | |
| 0.12 | -0.21 | -0.80 | -0.17 | 5.16 | |
| 0.62 | 0.22 | 0.70 | 0.47 | 1.54 | |
| 3.37 | 3.33 | 3.03 | 2.70 | 12.69 | ← **uni+multivariat** |
| -0.50 | -0.48 | -0.42 | -0.68 | 0.79 | |
| -0.61 | -0.51 | 0.03 | -0.18 | 2.00 | |
| 0.44 | 0.50 | 0.40 | 0.55 | 0.48 | |
| -0.21 | 0.29 | 0.29 | 0.05 | 2.79 | |
| -0.12 | -0.21 | -0.05 | -0.15 | 0.13 | |
| -0.15 | -0.32 | -0.40 | -0.03 | 1.12 | |
| 0.15 | 1.28 | -1.10 | -1.40 | 17.43 | ← **multivariat** |
| -1.82 | -1.85 | -1.70 | -1.73 | 3.62 | |
| -1.52 | -1.21 | -0.86 | -1.31 | 4.13 | |
| -0.24 | -0.37 | 0.31 | 0.09 | 1.41 | |
| -0.57 | -0.50 | -0.66 | -0.25 | 1.52 | |
| 1.16 | 1.40 | 0.13 | 1.22 | 10.24 | ← **multivariat** |
| -0.02 | -0.43 | -0.29 | -0.78 | 5.23 | |
| -0.85 | -0.76 | -0.74 | -0.66 | 0.82 | |
| 0.48 | 0.37 | 0.46 | 0.98 | 2.63 | |
| -0.16 | -0.82 | -0.51 | -0.60 | 4.73 | |
| -0.56 | -1.08 | -0.91 | -0.81 | 3.52 | |
| 0.82 | 0.47 | 0.64 | 0.34 | 2.46 | |
| -0.79 | -0.24 | -0.32 | -0.40 | 3.10 | |
| 1.31 | 1.76 | 1.86 | 1.60 | 6.50 | |
| -1.30 | -1.17 | -0.99 | -1.39 | 2.67 | |

**Q-Q Plot Methode**

**Korrelations-Test**

$$r_{qq} < r_{crit(\alpha)}$$

# Kritische Q-Q- Korrelationen

| Smple Size N | Significance level α | | |
|---|---|---|---|
| | 0.01 | 0.05 | 0.10 |
| 5 | 0.8299 | 0.8788 | 0.9032 |
| 10 | 0.8801 | 0.9198 | 0.9351 |
| 15 | 0.9126 | 0.9389 | 0.9503 |
| 20 | 0.9269 | 0.9508 | 0.9604 |
| 25 | 0.9410 | 0.9591 | 0.9665 |
| 30 | 0.9479 | 0.9652 | 0.9715 |
| 35 | 0.9538 | 0.9682 | 0.9740 |
| 40 | 0.9599 | 0.9726 | 0.9771 |
| 45 | 0.9632 | 0.9749 | 0.9792 |
| 50 | 0.9671 | 0.9768 | 0.9809 |
| 55 | 0.9695 | 0.9787 | 0.9822 |
| 60 | 0.9720 | 0.9801 | 0.9836 |
| 75 | 0.9771 | 0.9838 | 0.9866 |
| 100 | 0.9822 | 0.9873 | 0.9895 |
| 150 | 0.9879 | 0.9913 | 0.9928 |
| 200 | 0.9905 | 0.9931 | 0.9942 |
| 300 | 0.9935 | 0.9953 | 0.9960 |

Ist $r_{qq} < r_{crit(\alpha)}$, wird die Annahme der NV auf dem gewählten α Level verworfen. α sollte progressiv gewählt sein (10%), da man eine Sicherheit für die Beibehaltung wünscht.