GMF-Tagung „Zukunft der Lehre in Statistik und Stochastik"

Zürich, 21. Oktober 2017

Datenanalyse für Ingenieure im betriebswirtschaftlichen Umfeld: Gedanken zum Curriculum

Andreas Ruckstuhl

Dozent für Statistische Datenanalyse

Institut für Datenanalyse und Prozessdesign (idp), ZHAW

Inhalt

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

- Einige einleitende Gedanken
- Kurs I (Wahrscheinlichkeit & Statistik)
- Kurs II (Datenanalyse und Prognose)
- Schlussbemerkungen

Einige einleitende Gedanken (1/2)

- **Was hat Statistik mit Datenanalyse zu tun?**

- Meine Antwort dazu:
  Datenanalyse ist angewandte Statistik

  Denn die Statistik befasst sich heute neben dem Problem
  - *Wie sollen welche Daten gewonnen werden?*

  vor allem mit den Fragen
  - *Wie soll man Daten beschreiben?* und
  - *Welche Schlüsse kann man aus Daten ziehen?*

- **Jedoch „Datenanalyse ist angewandte Statistik" wird so vielfach nicht akzeptiert**

Einige einleitende Gedanken (2/2)

- **Ingenieure im betriebswirtschaftlichen Umfeld**
- Sie werden sich oft auch mit Betriebsabläufen befassen
- Um diese qualitativ oder quantitativ beschreiben zu können,
  - greift man auf geeignete Vorstellungen und Theorien zurück und
  - **kalibriert diese mit Daten**, die aus dem eigenen Betrieb stammen, und
  - macht **Prognosen**

- Das „Kalibrieren mit Daten / Prognose" ist mehr als das, was in einer Einführung in Wahrscheinlichkeit und Statistik abgedeckt wird!
  → also nennt ***man*** es Datenanalyse

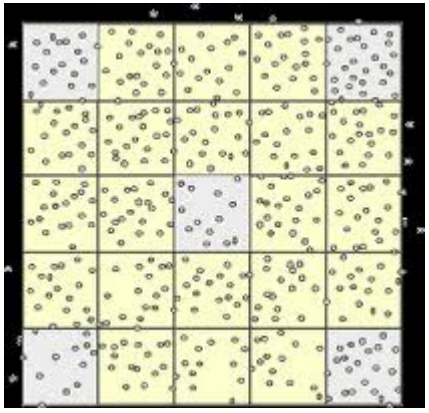
- Meine Antwort dazu: Wir sollten einen zweisemestrigen Kurs anbieten!!
  - Kurs I: „Einführung in Wahrscheinlichkeit und Statistik"
  - Kurs II: „Datenanalyse und Prognose"
  - Unter Verwendung von R (oder einer ähnlichen Software)

Kurs I (W'keit & Stat)

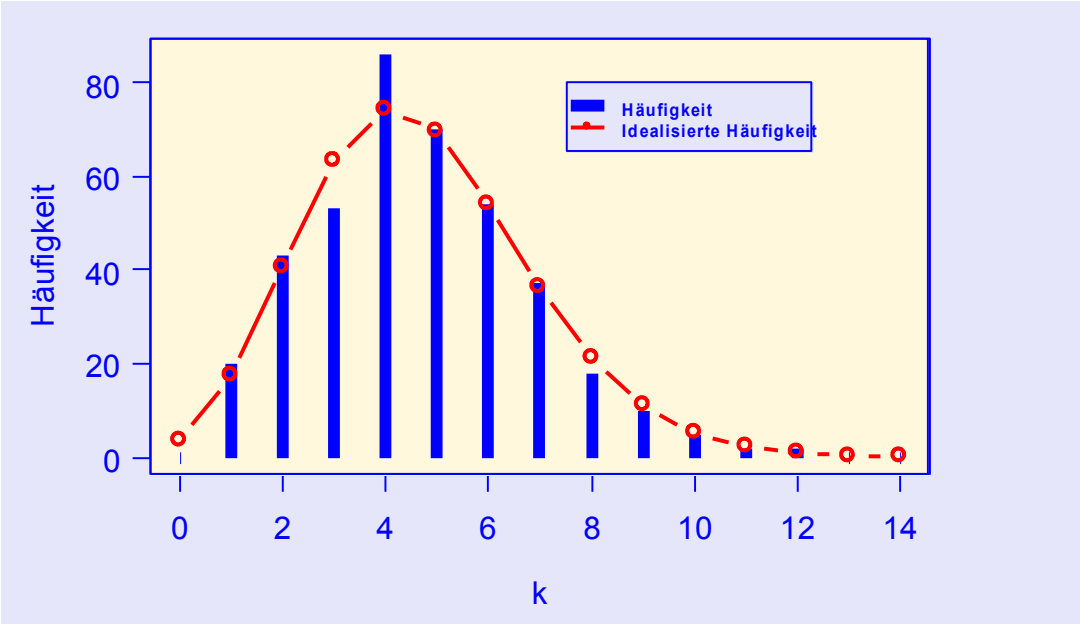
- **Inhalt ist „klar":**
  - Deskriptive Statistik (Variablentypen, Kennzahlen, Korrelation, Grafiken, Simpson, ...)
  - Wahrscheinlichkeitsrechnen
    - Zufall, Wahrscheinlichkeiten, bedingte Wahrscheinlichkeiten, Unabhängigkeit
    - Diskrete und stetige Wahrscheinlichkeitsverteilungen
    - Erwartungswert und Varianz
    - Zentraler Grenzwertsatz
  - Inferenz: Schätzen, Statistische Tests, Vertrauensintervalle
    - für Binomial-, Poisson- und Normalverteilungs-Modell
    - (Vergleich von zwei Stichproben, Chi-Quadrat-Anpassungstest, Bootstrap, Bayes)
  - Weitere wichtige Aspekte
    - Interpretation von Wahrscheinlichkeiten (Laplace, frequentistisch, subjektivistisch, ...)
    - Interpretation von Testentscheidungen
    - Signifikanz vs. Relevanz
    - Welche Variabilität berücksichtigt das Vertrauensintervall?
      (≠ Prognoseintervall, Problem von nicht repräsentative Stichproben, Unabhängigkeit verletzt)
  - Auch **Beispiele mit echten Daten und mit R** durcharbeiten, **Simulationen**

Skizze zu „Bootstrap“ (1/2)

- William Sealy Gosset (1876 -1937) studierte die Variation, die sich beim Zählen von Hefezellen in Bier mit einem Hämozytometer ergab. …
  Anschliessend wurde die Scheibe in 400 Quadrate unterteilt und die darin enthaltenen Hefezellen gezählt:
  2, 6, 5, 5, 3, 4, 5, 3, 8, 3, 4, 7, 1, 5, 3, 9, 8, 5, …



- Vergleich zwischen den beobachteten Häufigkeiten und den idealisierten Häufigkeiten (= angepasstes Poisson-Modell $n \cdot \widehat{p}_k$)



Skizze zu „Bootstrap“ (2/2)

- Bei der Poissonverteilung ist die **Dispersion** $\frac{\mathrm{var}\langle X\rangle}{E\langle X\rangle}$ ist gleich 1.
- Folglich muss bei Poisson verteilten Daten die empirische Varianz ungefähr gleich dem arithmetischen Mittel sein: $s^2/\bar{x} \approx 1$
  - Für das Hefezellen-Beispiel ist $s^2/\bar{x} = 0.9548658$
  - Ist diese Abweichung von 1 plausibel bei einer Poisson-Verteilung?
- Zur Beantwortung kann die **Bootstrap-Simulation verwendet** werden
  - nichtparametrischen Bootstrap erklären
  - In R: `library(boot); ...; HZ.boot <- boot(...)`

```
> boot.ci(HZ.boot, conf=0.95, type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 4999 bootstrap replicates
... snip ...
Intervals :
Level   Percentile
95%   (  0.8229,  1.0915 )
Calculations and Intervals on Original Scale
```

  - **Weil das 95%-Vertrauensintervall von [0.823, 1.091] den Wert λ_0=1 (Nullhypothese) enthält, habe wir keine Evidenz gegen die Nullhypothese auf dem 5% Niveau.**

Kurs II (Datenanalyse und Prognose)

- Regressionsanalyse ist die meist verwendete Methode in der statistischen Datenanalyse
  - Wird zur Inferenz der Parameter benutzt
  - Wird zur Prognose (nicht nur zeitlich) verwendet

  - Also Kurs II, 1. Teil: Einführung in die Regressionsanalyse

- Daten gestützte zeitliche Prognosen basieren auf Zeitreihen
  - Also Kurs II, 2. Teil: Einführung in die Zeitreihenanalyse

Kurs II (Datenanalyse und Prognose)

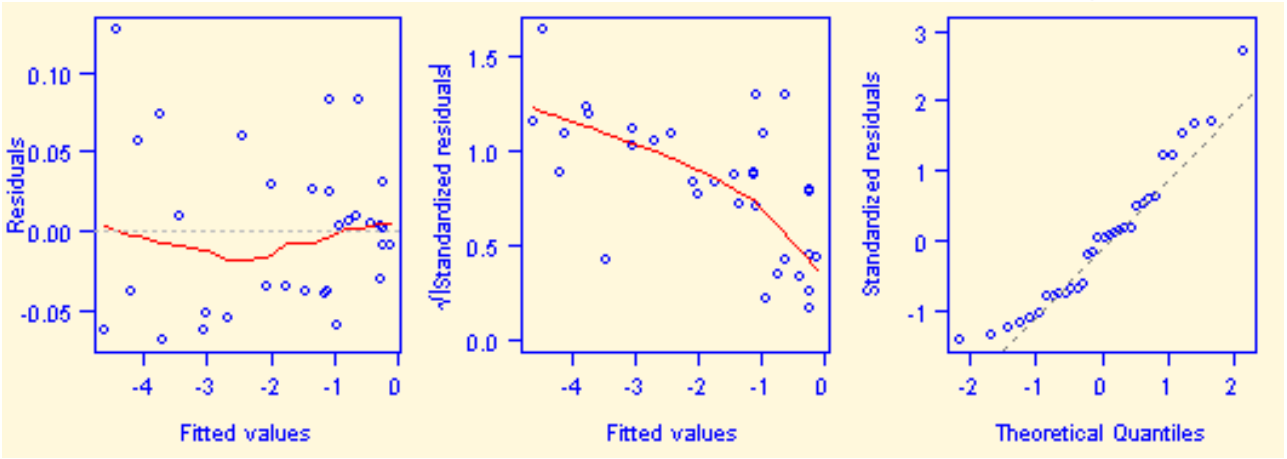
- Inhalt des **Regressionsteils** (zuerst einfache, dann multiple)
  - Regressionsmodell für eine oder mehrere erklärende Variablen
  - Inferenz (Schätzen, Testen, Vertrauensintervalle)
  - Prognose und Prognoseintervalle
  - Residuenanalyse, Einflussdiagnostik, Kreuzvalidierung (Diskussion der Ziele Prognose vs Inferenz vs Kausalität)
  - Transformationen und kategorielle erklärende Variablen (machen Regression zu einem sehr potenten Werkzeug)
  - Variablenselektion (mit AIC), Modellierungsstrategien
  - Kollinearität und die Interpretation der geschätzten Parameterwerte

- … **und immer wieder Beispiele mit (echten) Daten**

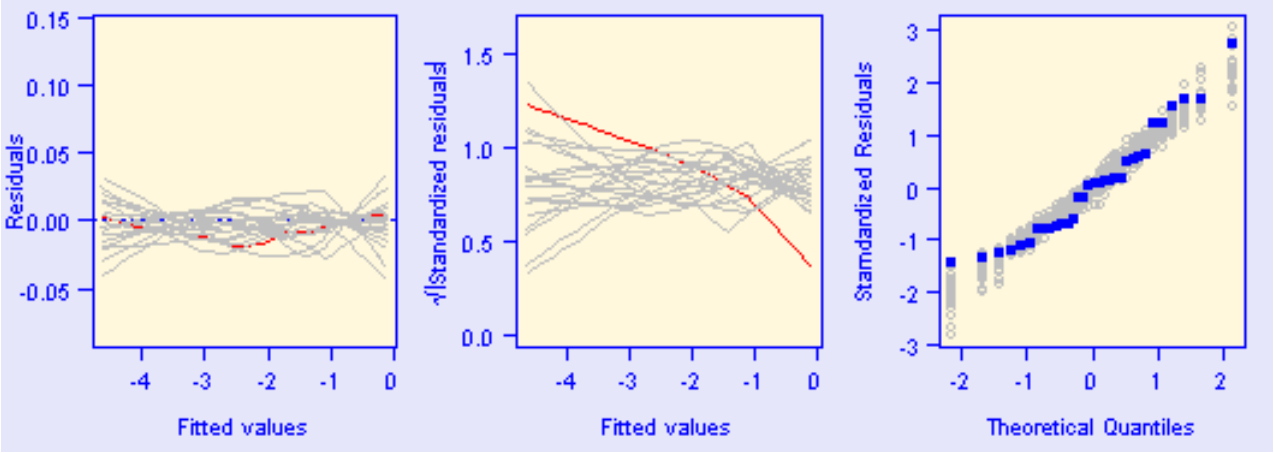
Bsp Residuen-Analyse

- Eine Residuen-Analyse sollte mit den drei Diagrammen Residuals vs Fitted (A), Scale-location (B) and Normal QQ (C)

A B C



und den entsprechenden Bootstrap-Simulationen durchgeführt werden.

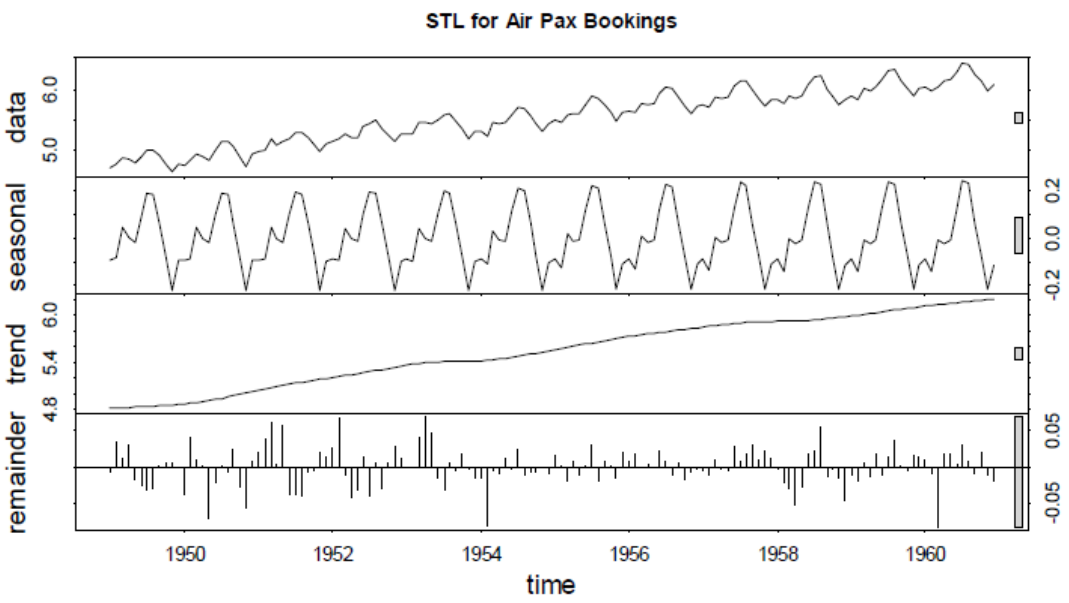


Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

- Inhalt des **Zeitreihen-Teils**
  - Beispiele aus …
  - Zeitreihen-Prozess (Beobachtungen sind eine einzige (!) Realisation davon).
  - Stationarität, Erwartungswert, Varianz, Autokorrelation
  - Zeitreihen in R
  - Deskriptive Zerlegung einer Zeitreihe (Trend, Saisoneffekt, Restterm) Also

STL mit zeitabhängigem Saisoneffekt

```
> lap.stl <- stl(lap, s.window=13)
> plot(lap.stl, main="STL for Air Pax Bookings")
```



Kurs II (Datenanalyse und Prognose)

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

- Inhalt des **Zeitreihen-Teils (Fortsetzung)**
  - Autokorrelation, lagged Scatterplot, Korrelogramm (ACF), Partielle Autokorrelation
  - Auswirkungen von Ausreissern
  - Modelle für stationäre Zeitreihen: Weisses Rauschen, Autoregressive (AR) Modelle
  - (Zeitliche) Vorhersage
    - mit AR(p): Ein-Schritt- und Mehr-Schritt-Vorhersage
    - Vorhersage von Saison-Trend-Zeitreihen
    - Exponentieller Glätter: einfacher und für Saison/Trend (Holt-Winters-Methode)
    - Warnung: zeitliche Vorhersage ist eine Extrapolation und ist wie Autofahren nur mit Blick in den Rückspiegell

Schlussbemerkung

- Statistik ist mehr als Mittelwert und Standardabweichung oder eine Tabelle
- **Angewandte Statistik ist Datenanalyse**

- Also darf die Einführung in angewandte Statistik
  - nicht beim 1x1 (übliche Einführung in die Statistik) stecken bleiben,
  - sondern sollte zumindest bei betrieblichen Ingenieuren auch **Regression und Prognose enthalten**

- Es sollten also (6 bis) 8 Credits investiert werden