4 Modellierung von Eingabedaten

- Repräsentation der Parameter spielt zentrale Rolle in jedem Simulationsprogramm!
- Oft werden stochastische Größen zur Abbildung der Realität verwendet

Wie kommt man an Verteilungen und Parameter?

Wird in diesem Abschnitt untersucht

Zwei wesentliche Quellen:

1. a priori Wissen
   - aus vorherigen Modellierungen
   - aus ähnlichen Modellen
   - aus Erfahrung
   - aus der Theorie
   - ...
2. Messungen (d.h. Stichproben)
   - am realen System
   - an ähnlichen Systemen
   - an anderen Modellen
   - ...

Schritte bei der Modellierung von Eingabedaten

1. Datensammlung
2. Entscheidung über die Darstellung der gemessenen Daten
    a. Deterministische Größe
    b. Diskrete empirische Verteilung
    c. Kontinuierliche empirische Verteilung
    d. Stochastische Verteilung

Falls 2d. gewählt

3. Auswahl eines Verteilungstyps
4. Schätzung der Verteilungsparameter
5. Überprüfung der Passgüte durch Anpassungstest

Sammlung/Messung von Daten

Datenerhebung ist aufwändig und oft frustrierend, aber eminent wichtig für alle nachfolgenden Schritte!

GIGO-Prinzip (garbage-in garbage-out)

Probleme bei der Datenerhebung

- Zu wenige Daten
  - geringer Stichprobenumfang
  - nur summarische Statistiken
  - lediglich qualitative Informationen
- Zu viele Daten
  - Daten aus automatischen Messungen
  - vollständige Traces
- Falsche Daten
  - falscher Aggregationszustand (z.B. Monat statt Tag)
  - korrelierte Daten
  - falscher zeitlicher Bezug (z.B. aus der Vergangenheit)
  - falscher räumlicher Bezug (z.B. vom falschen System)
  - ungenauer sachlicher Bezug (z.B. Umsatz statt Nachfrage)

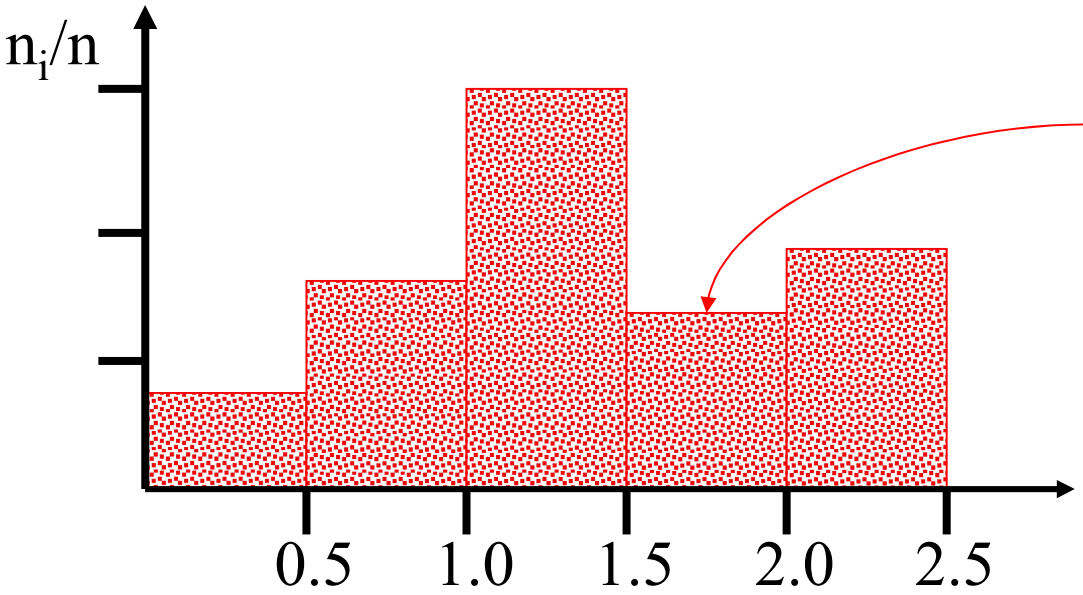
Regeln zur Datenerhebung:

- Vorläufe zur Bestimmung des Erhebungsintervalls, der Auflösung, des Stichprobenumfangs
- Falls möglich Daten schon während der Erhebung analysieren
- Bei mehreren Stichproben erst Homogenität und Herkunft aus einer Verteilung prüfen (z.B. mit Testverfahren)
- Auf zensierte oder verfälschte Daten achten
- Daten auf Korrelation untersuchen
- Zwischen Ein- und Ausgabedaten bei der Messung unterscheiden

Aufbereitung und Repräsentation von Daten

- Daten seien über den Beobachtungszeitraum identisch verteilt (ansonsten Darstellung des Verlauf über der Zeit)
- Zur Interpretation von Daten eignen sich graphische Repräsentationen und Maßzahlen
- Annahme: n Daten x_i nach Erhebungszeit geordnet
  $y_1,\ldots, y_n$ Stichprobe nach Größe geordnet (d.h. $y_i \leq y_{i+1}$)

Histogramm (als Approximation der Dfkt.)



Höhe proportional zur Anzahl Werte im Intervall dividiert durch die Anzahl der Werte der Stichprobe (Schätzer für den Wert der Dichtefunktion)

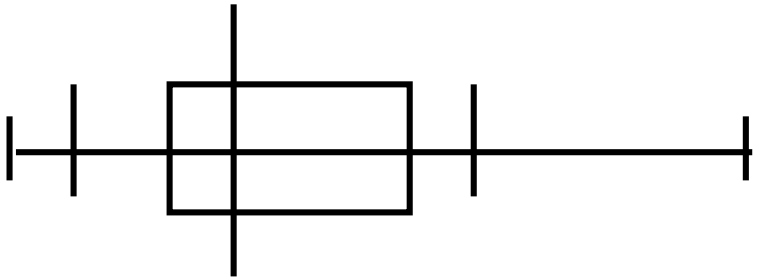
Intervallbreite frei wählbar, mehrere probieren!

Maßzahlen

- Momente, Varianz
- Quantilschätzer aus der geordneten Stichprobe $(y_1,\ldots,y_n)$:
  - Median y_i mit $i = (n+1)/2$
  - Quartile y_j und y_{n-j+1} mit $j = (\lfloor i \rfloor +1) / 2$
  - Octile y_k und y_{n-k+1} mit $k = (\lfloor j \rfloor +1) / 2$
  - Extremwerte y_1 und y_n

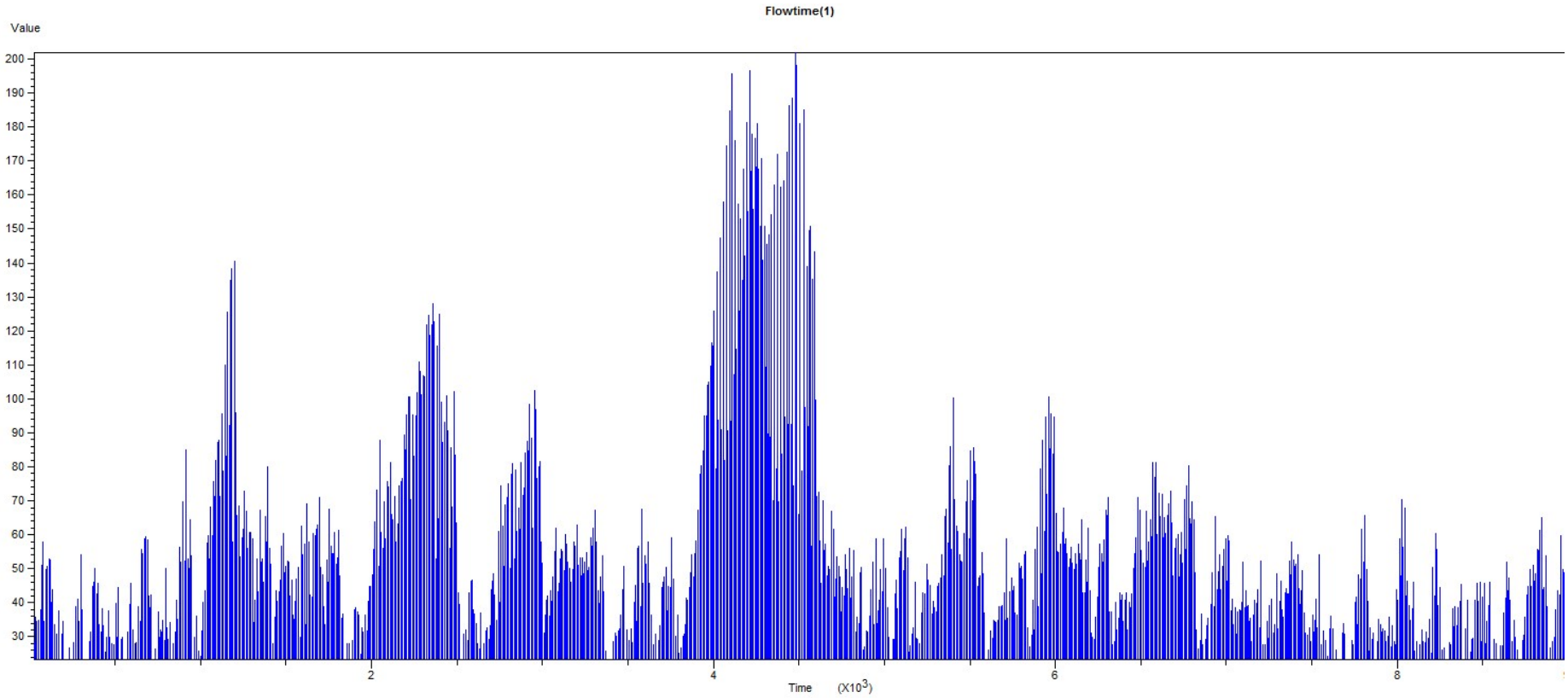
wobei $y_{m+0.5}$ mit $m \in \mathbb{N}$ $0.5(y_m+y_{m+1})$ entspricht

Graphische Darstellung Box-Plots



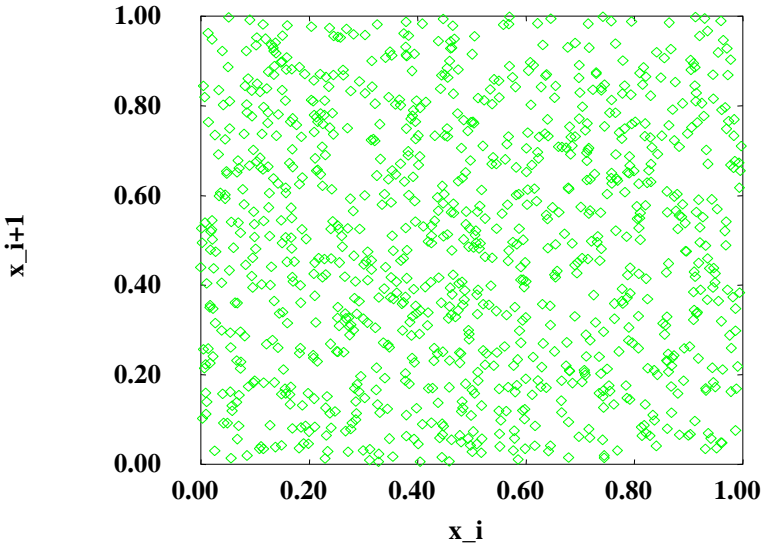
Beobachtete Werte über die Zeit:

Im Beispiel Durchlaufzeiten durch eine Fertigung

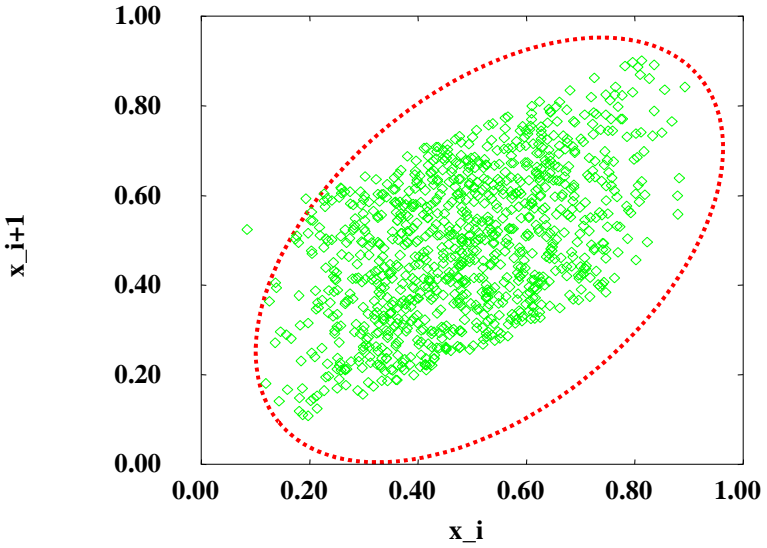


Graphische Darstellung von Abhängigkeiten durch Tupel (x_i, x_{i+1})

keine Korrelation erkennbar



positive Korrelation



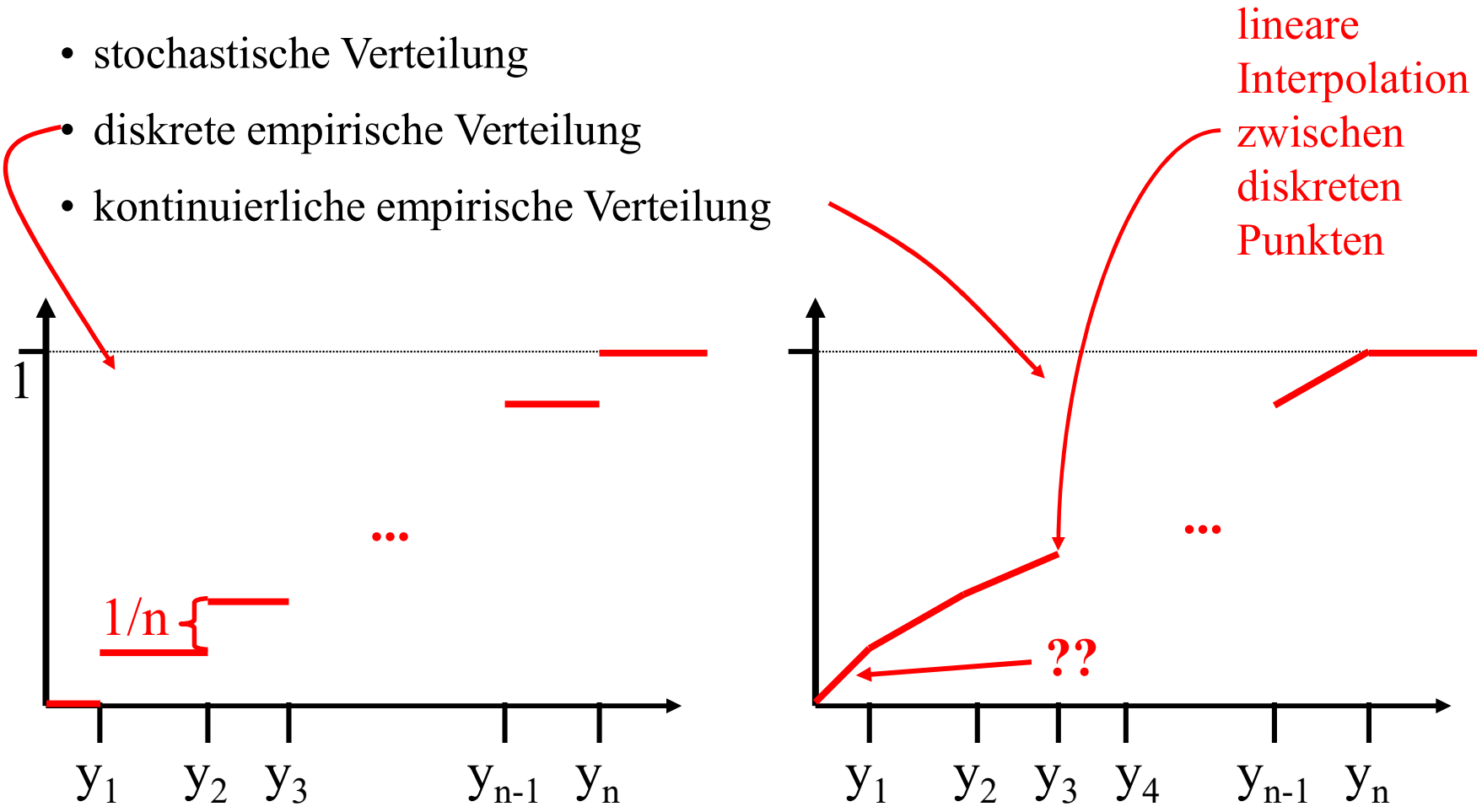
- Visueller Eindruck zeigt Korrelationen
- Auch zur Darstellung der Abhängigkeiten zwischen X_i und X_{i+k} (k>1) nutzbar
- Im Prinzip auch 3D-Darstellung möglich

Darstellungsalternativen

Konstante, wenn die Daten alle (fast) identisch sind

Ansonsten:

- stochastische Verteilung
- diskrete empirische Verteilung
- kontinuierliche empirische Verteilung



Empirische Verteilung:

- Nutzung der gesamten Information
- Aber kein mathematisches Modell

Theoretische Verteilung:

- Wohldefiniertes mathematisches Modell
- Struktur durch Verteilungstyp vorgegeben

Insgesamt kontrovers in der Literatur behandelt

Simulationsmodelle benutzen oft (aus Effizienzgründen) theoretische Verteilungen

Beide Darstellungsformen berücksichtigen keine Korrelationen!

4.5 Anpassung theoretischer Verteilungen

Auszuführende Schritte:

1. Bestimmung des Verteilungstyps
2. Schätzung der Parameter
3. Bestimmung der Anpassungsgüte

Typische Szenarien

a) Verteilungstyp aus der Theorie, Parameterschätzung aus der Stichprobe

b) Verteilungstyp aus der Stichprobe, Parameterschätzung aus der Stichprobe

c) Weder theoretische Erkenntnisse, noch Stichprobe vorhanden

a) und b) erfordern jeweils Parameterschätzung und Test der Anpassungsgüte

Fall a): Theoretische Erkenntnisse über den Verteilungstyp:

- ZV X, welche aus einer größeren Anzahl zufälliger Ereignisse resultiert, könnte normalverteilt sein (zentraler Grenzwertsatz)
- ZV X, welche das Minimum einer größeren Anzahl zufälliger Ereignisse ist, könnte Weibull-verteilt sein
- ZV X, welche zeitliche Abstände aufeinanderfolgender Ereignisse darstellt könnte exponentiell-verteilt sein, wenn anzunehmen ist, dass Ereignisse
  - einzeln auftreten
  - mit konstanter Rate λ auftreten
- ZV X, welche das Produkt einer größeren Anzahl zufälliger Einflüsse ist, könnte log-normal-verteilt sein
- ...

Zusätzlich oft Erkenntnisse aus dem Anwendungsgebiet

Fall b): Bestimmung des Verteilungstyps aus der Stichprobe

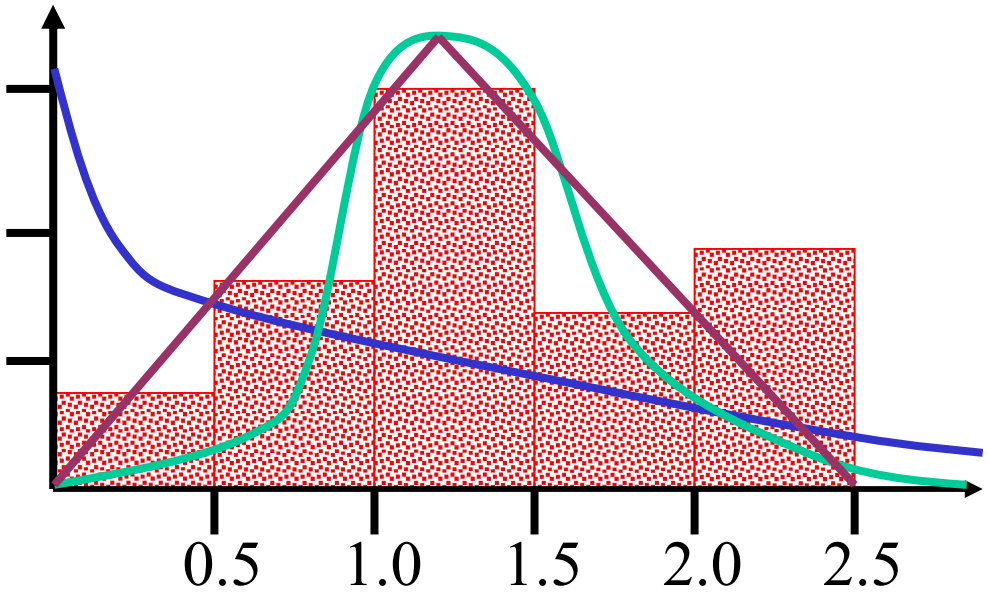
Erste Hinweise auf Ausschluss bestimmter Verteilungstypen auf Basis von Verteilungscharakteristika

- Variationskoeffizient VK(Y)= σ(Y) / E(Y)
  Ausschluss bestimmter Verteilungstypen z.B. Exponentialverteilung hat VK 1 ⇒ VK(Y) deutlich von 1 abweichend, keine Exponentialverteilung wählen (Schätzung von σ(Y) und E(Y) später)
- andere Charakteristika
  - Verhältnis Erwartungswert zu Median
  - höhere Momente (z.B. Schiefe $\upsilon = \frac{E((Y-E(Y))^3)}{\left(\sigma^2\right)^{3/2}}$)
  - ...

Ausschluss von Verteilungstypen, i.a. aber **keine** Festlegung auf einen Verteilungstyp!

Bestimmung des Verteilungstyps aus dem Histogramm der Stichprobe

Histogramm ist erwartungstreuer Schätzer der Dichtefunktion!



- Visueller Vergleich Dfkt. – Histogramm auf Basis der Form (also erst einmal ohne Kenntnis der Parameter)

Basis: Erfahrung/Wissen

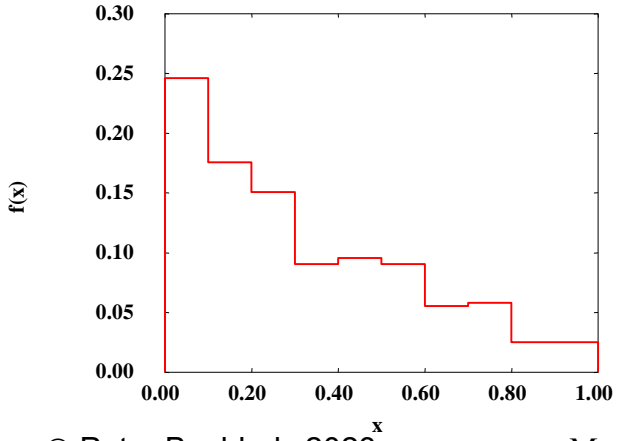
Heute Unterstützung durch Softwaretools (ExpertFit, Arena Input-Analyzer, ..) mit teilweiser automatischer Verteilungswahl und Parameterschätzung

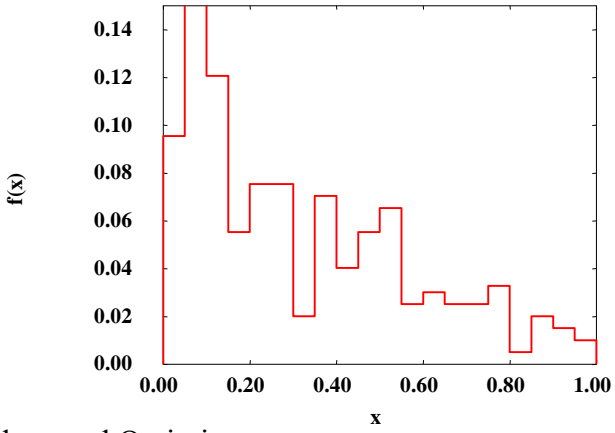
Vorsicht: Ergebnis kann durch (frei wählbare) Histogrammparameter beeinflusst werden!

Freiheitsgrade bei der Histogrammerstellung

- Anzahl und Breite der Zellen
  - i.d.R. Zellen gleicher Breite, ansonsten Höhe anpassen
  - Zahl der Zellen so wählen, dass
    - Form der Dichte erkennbar
    - jede Zelle mehrere Werte enthält (≥ 10)
- überdeckter Bereich
  - Start der ersten Zelle, Ende der letzten Zelle
  - Behandlung von Werten außerhalb des Histogramms (Ausreißern)

Quantitative Bewertung der Anpassung erst nach Parameterschätzung





4.6 Parameterschätzung

Jede Verteilungsfamilie weist gewisse Parameter auf:

- Exponentialverteilung: Rate λ
- Normalverteilung: Mittelwert μ, Standardabweichung σ
- Dreiecksverteilung: linke Grenze a, rechte Grenze b, Modalwert c
- ...

Allgemeine Form der Dichtefunktion $f(x,\Theta)$ mit Parametervektor $\Theta=(\Theta_1, ..., \Theta_p)$

Ziel: Bestimme die Werte von Θ so, dass die Dichtefunktion der Verteilung und die Stichprobe „möglichst gut korrespondieren“.

Zahlreiche Methoden existieren, wir betrachten

- Momentenmethode
- Maximum-Likelihood-Methode

Momentenmethode

Zur Notation:

- $(x_1, ..., x_n)$ ist die konkrete Stichprobe
- jeder Wert x_i ist eine Realisierung der ZV X_i (alle X_i sind identisch verteilt!)

Sei $\tilde{X}^i$ Schätzer für $E(X^i)$ und $\hat{X}^i$ der konkrete Schätzwert

Entsprechend definieren wir

- $\tilde{S}^2$ und $\hat{S}^2$ als Schätzer und Schätzwert für die Varianz
- $\widetilde{CV}$ und $\widehat{CV}$ als Schätzer und Schätzwert für den Variationskoeffizienten

Erwartungstreue Schätzer:

$$\tilde{X}^i = \frac{1}{n}\sum_{j=1}^{n}(X_j)^i \text{ und } \tilde{S}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \tilde{X}^1)^2$$

Parameter Θ_j lassen sich oft als Funktion der Momente darstellen

$$\Theta_j = \phi_j\left(E(X^1), \ldots, E(X^p)\right)$$

Momentenmethode substituiert die Momentenschätzer für die Momente und liefert damit einen Parameterschätzer

$$\tilde{\Theta}_j = \phi_j\left(\tilde{X}^1, \ldots, \tilde{X}^p\right)$$

Schätzer oft nicht erwartungstreu, aber asymptotisch erwartungstreu und konsistent

Trotzdem oft keine guten Schätzer, da die „Form“ der Verteilung nicht berücksichtigt wird

Beispiele:

Exponentialverteilung $E(X^1)=\lambda^{-1}$

$$\Rightarrow \quad \tilde{\lambda} = \frac{1}{\tilde{X}^1} \left(= \frac{n}{\sum_{j=1}^{n} \tilde{X}_j}\right)$$

Normalverteilung:

$E(X^1)=\mu$, $\sigma^2 = E(X^2)-E(X^1)^2$

$$\Rightarrow \ \mu = \tilde{X}^1 \text{ und } \sigma = \tilde{S}$$

Maximum-Likelihood-Methode (ML-Methode)

Suche nach den plausibelsten Parametern

Vorstellung der ML-Methode am Beispiel:

- Diskrete Verteilung mit Parameter θ, so dass $p_\theta(x)$ W. für Wert x bei Parameter θ
- Stichprobe $(x_1, ..., x_n)$
- Likelihoodfunktion $L(\theta) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n)$
- Ziel der ML-Methode: Wähle θ so, dass $L(\theta)$ maximal
- also $\max_\theta(L(\theta))$ Punkt der maximalen Beobachtungswahrscheinlichkeit
- analoges Vorgehen bei Parametervektor (mehrdimensionales Optimierungsproblem)

Kontinuierlicher Fall nicht ganz so intuitiv, da Wahrscheinlichkeit in jedem Punkt 0 sein kann

Verwendung der Dichtefunktion $f_\theta(x)$ mit Parameter θ

$L(\theta) = \prod_{j=1}^{n} f_\theta(x_j)$ finde θ_{max}, so dass $L(\theta_{max}) \geq L(\theta)$ für alle θ.

Beispiel Exponentialverteilung:

$$L(\lambda) = \prod_{j=1}^{n} \lambda \cdot e^{-\lambda \cdot x_j} = \lambda^n \cdot e^{-\lambda \cdot \sum_{j=1}^{n} x_j}$$

Optimierung des natürlichen Logarithmus ist einfacher
(natürlicher Logarithmus ist eine Transformation, welche die Lage des Optimums nicht verändert)

$$l(\lambda) = \ln(L(\lambda)) = n \cdot \ln \lambda - \lambda \cdot \sum_{j=1}^{n} x_j$$

$$l'(\lambda) = n / \lambda - \sum_{j=1}^{n} x_j$$

Nullstelle der Ableitung

$$\hat{\lambda} = n / \left(\sum_{j=1}^{n} x_j \right)$$

Schätzer in diesem Fall identisch zur Momentenmethode

Verfahren auch für Parametervektoren anwendbar

Allgemein Maximierung von l(θ) statt L(θ)

$$l(\theta) = \ln(L(\theta)) = \sum_{j=1}^{n} \ln(f_\theta(x_j))$$

Im Allgemeinen Optimierungsproblem ohne geschlossene Lösung

$\Rightarrow$ Anwendung von (nichtlinearen) Optimierungsverfahren

Eigenschaften der ML-Schätzer:

1. Asymptotisch erwartungstreu
2. Konsistent
3. Asymptotisch normalverteilt (Berechnung von Konfidenzintervallen)
4. I.d.R. nicht schlechter bzw. besser als Momentenschätzer

ML-Schätzer für die Normalverteilung: $\mu = \tilde{X}^1$ und $\sigma^2 = \frac{n-1}{n}\tilde{S}^2$

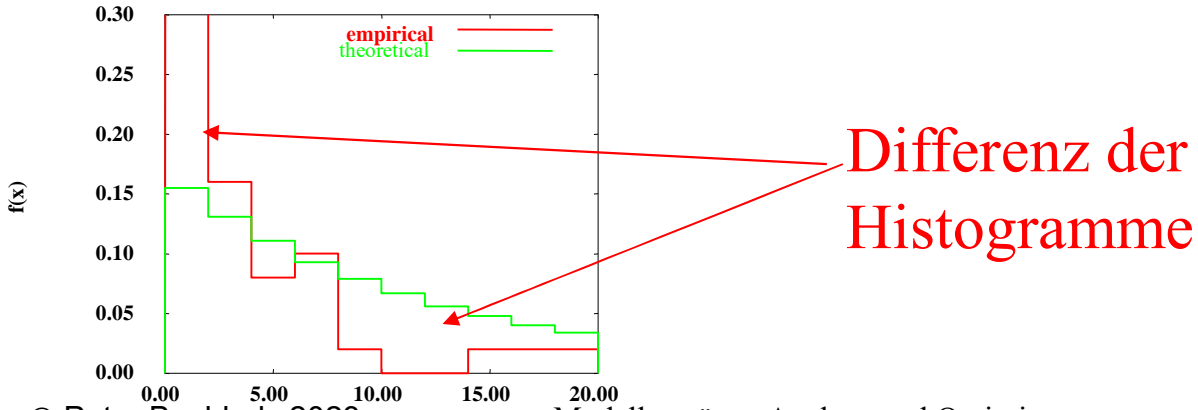
4.7 Bestimmung der Anpassungsgüte

Nach Verteilungsbestimmung und Parameterschätzung kann die Anpassungsgüte der theoretischen Verteilung quantifiziert oder getestet werden

Methoden zur Quantifizierung

Vergleich der Histogramme

- n_i Anzahl Werte der Stichprobe im i-ten Intervall und $h_i = n_i/n$
- $p_i = \int_{\Delta_i} f(x)\,dx$ wobei Δ_i das i-te Intervall ist
- Abstandsmaß $D = \Sigma_i |h_i - p_i|$ oder $D' = \Sigma_i (h_i - p_i)^2$



Differenz der Histogramme

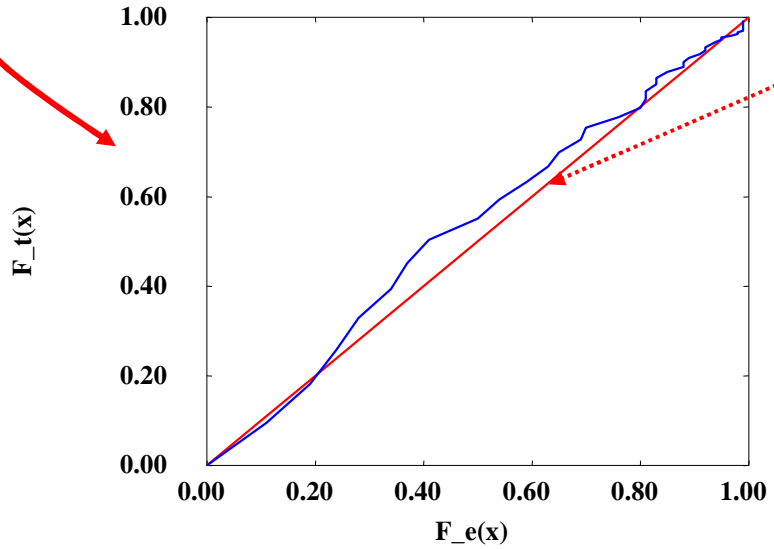
Wahrscheinlichkeitsplots

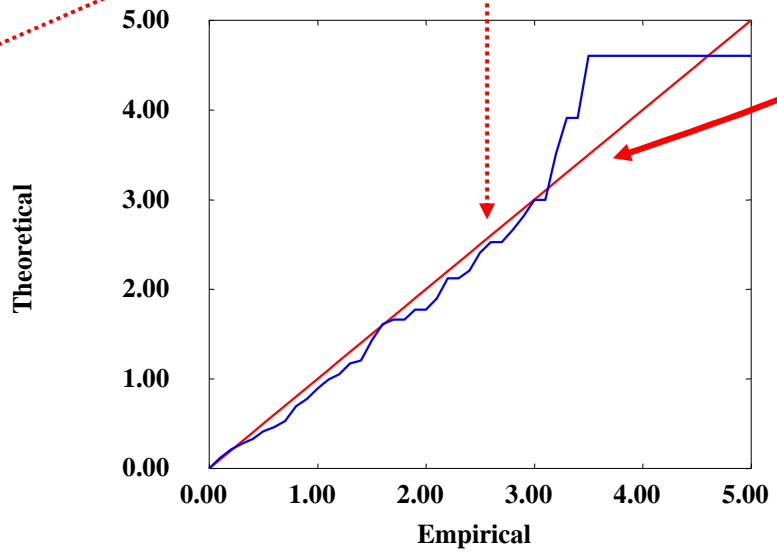
Grundidee: Vergleich der Werte der Verteilungsfunktion der empirischen und theoretischen Verteilung

- Für die empirische Verteilung gilt $F_e(y) = \max_{y_i \le y}(i/n)$
- $F_t(y)$ sei der Wert der theoretischen Verteilungsfunktion
- Q-Q-Plot Darstellung der Quantile (y_i,z_i) mit $F_e(y_i) = F_t(z_i)$
- P-P-Plot Darstellung $(F_e(y_i), F_t(y_i))$

y_i geordnet

Abweichung von einer Geraden beschreibt Abweichungen der Vfkts.





Anpassungstests

Zentrale Frage: Wann ist die Modellierung der Stichprobe durch eine theoretische Verteilung als adäquat anzusehen?

- Bisherige Ansätze erlauben Bewertung der Anpassung auf Basis des visuellen Vergleichs oder ausgesuchter Maßzahlen
- Auf Grund stochastischer Schwankungen ist zu erwarten, dass empirische und theoretische Verteilung immer Abweichungen aufweisen (müssen)
- Bleibt die Frage, welche Abweichungen (noch) tolerierbar sind

Alternative/Ergänzung zu den bisherigen Maßzahlen sind Tests:

Hypothese H_0: $(x_1,..., x_n)$ wurde aus Verteilung F(x) gezogen

Test liefert Antwort, ob H_0 verworfen oder angenommen werden soll

Chi-Quadrat Test

Sehr altes Testverfahren (ca. 1900) erlaubt einen formalen Vergleich der Histogramme empirischer und theoretischer Verteilungen

Definiere $b_0 < b_1 < ... < b_k$ als Intervallgrenzen

- $n_i = |\{y_j \mid b_{i-1} \leq y_j < b_i\}|$
- $p_i = \int_{b_{i-1}}^{b_i} f(y)dy$

Differenz $\Sigma_{i=1,..,k}|n_i - p_i \cdot n|$ liefert ein Maß für die Abweichung der beiden Histogramme: Je kleiner der Wert, desto wahrscheinlicher ist es, dass die Stichprobe aus der theoretischen Verteilung gezogen wurde.

Um ein Testverfahren anzuwenden, muss eine Teststatistik mit bekannter Verteilung definiert werden!

Teststatistik $d = \sum_{i=1}^{k} \frac{\left(n_i - n \cdot p_i\right)^2}{n \cdot p_i}$ Welche Verteilung hat d?

χ^2-Verteilung:

Seien Y_1, ..., Y_k unabhängig, identisch N(0,1)-verteilte ZVs, dann ist

$$Y = \sum_{i=1}^{k} Y_i^2$$

χ^2-verteilt mit k Freiheitsgraden.

Familie der –Verteilungen liegt in vertafelter Form vor
(keine explizite funktionale Form)

Wert d ist Realisierung einer ZV D

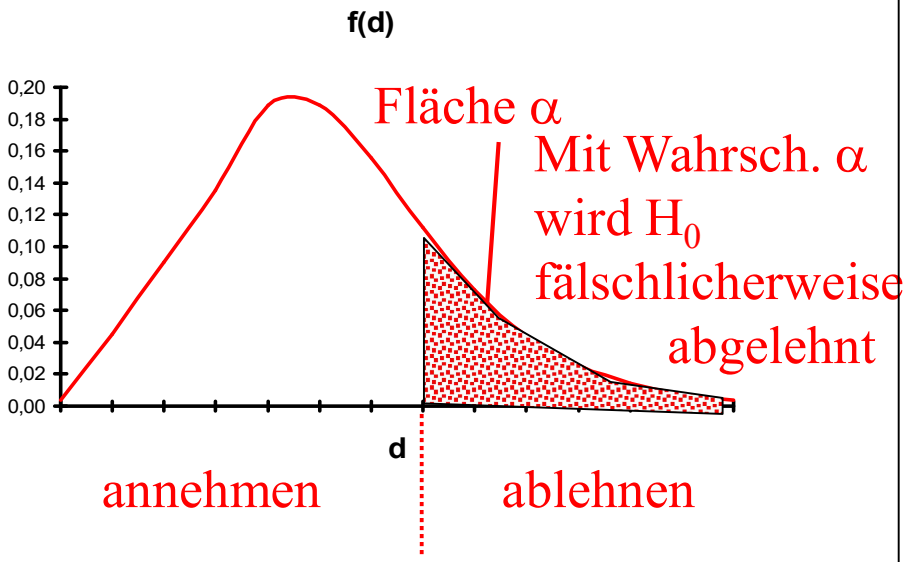
Falls Hypothese H_0 gilt:

- dann ist D asymptotisch χ^2-verteilt (d.h. für „genügend große“ n ist D approximativ χ^2-verteilt)
- Fallunterscheidung zur Bestimmung der Freiheitsgrade
  - falls keine Verteilungsparameter aus der Stichprobe geschätzt wurde mit k-1 Freiheitsgraden
  - falls p Verteilungsparameter aus der Stichprobe geschätzt wurden mit k-p-1 Freiheitsgraden

Vorgehen

- d berechnen und
- mit kritischen Werten zum gewählten Signifikanzniveau vergleichen (vertafelt)
- Hypothese annehmen/ablehnen

Skizze des Vorgehens:



Subjektive Komponenten: Lage, Größe und Anzahl der Intervalle (d.h. Festlegung der b_i)

Hinweise

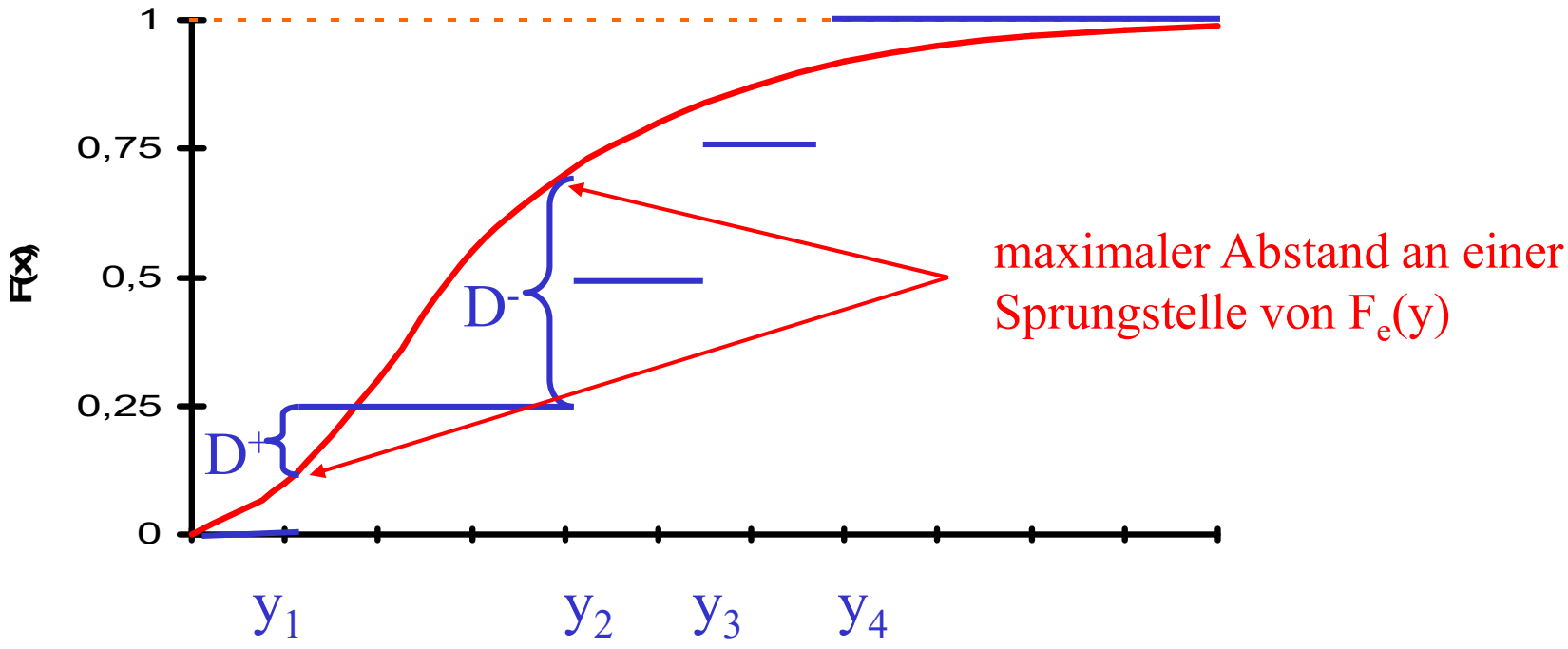
- Wähle Intervalle so, dass Werte p_j identisch/ähnlich sind (also unterschiedliche Intervallbreiten)
- Wähle Intervalle so, dass $n \cdot p_i \geq 5$

Kolmogorov-Smirnov Test

Grundidee: Vergleich der empirischen und theoretischen Verteilungsfunktion:

$F_e(y) = |\{y_i \leq y\}|/n$ (Treppenfunktion)

Teststatistik $D_n = \max_y(|F_e(y)-F_t(y)|)$ (falls nötig sup statt max)



Formale Definition von $D_n = \max \{D_n^-, D_n^+\}$ mit

$D_n^+ = \max_{1 \le i \le n} \{i/n - F_t(y_i)\}$ und $D_n^- = \max_{1 \le i \le n} \{F_t(y_i) - (i-1)/n\}$

Großer Wert von D_n deutet auf eine schlechte Anpassung hin

Fallunterscheidung bei der Anwendung des Tests:

- Falls keine Verteilungsparameter aus der Stichprobe geschätzt wurden, ist H_0 zu verwerfen, wenn

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) \cdot D_n \ge c_{1-\alpha}$$ Werte $c_{1-\alpha}$ sind vertafelt

- Falls Verteilungsparameter aus der Stichprobe geschätzt wurden, so sind Teststatistiken nur für spezielle Verteilungen bekannt
  z.B. Normalverteilung, Exponentialverteilung, Weibull-Verteilung

Aussagekraft von Testverfahren

Verfahren zur Parameterschätzung und Testverfahren sind in vielen Softwarepaketen implementiert:

- ExpertFit
- Arena Input Analyzer
- ..

Resultate der Software:

- Automatische Anpassung der Parameter
- Ausgabe der α-Werte für die Stichprobe
  (je größer α, desto besser)
- u.U. automatische Auswahl der „besten“ Verteilung

Vorsicht: die Software kennt nur die Daten eine manuelle Überprüfung ist notwendig!

Fall c): Keine Information und keine Stichprobe:

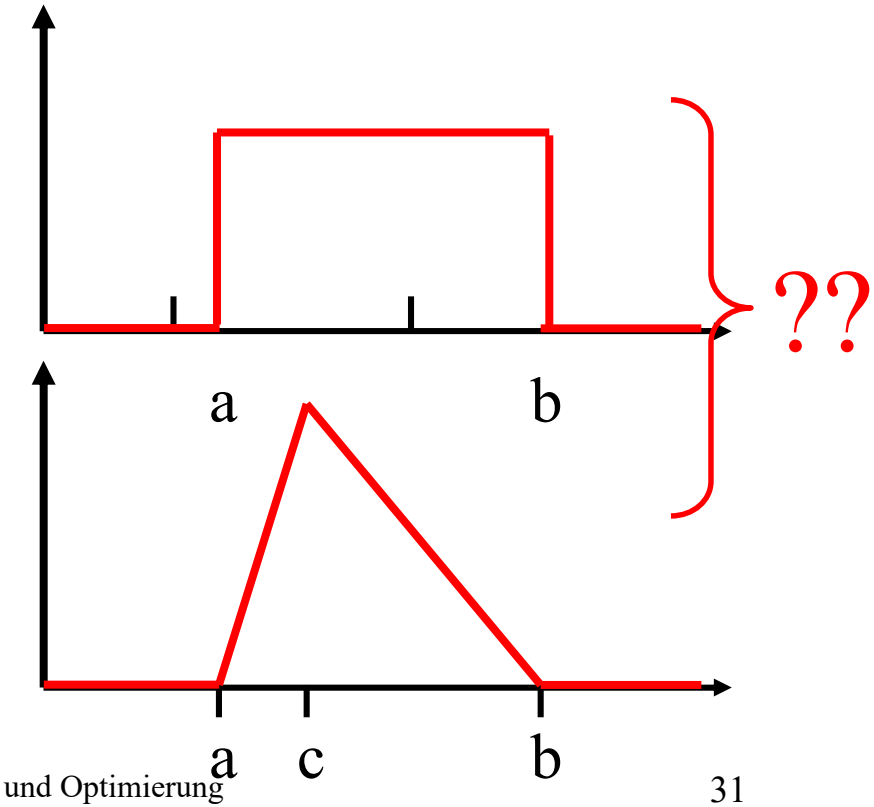
Äußerst „ungemütliche“ Situation, da eigentlich nicht genügend Information vorhanden

Heuristische Ansätze aus der Praxis

1. Raten des minimalen Wertes a und des maximalen Wertes b, so dass $P[x<a] \approx P[x>b] \approx 0$
2. Evtl. zusätzliches Raten des Mittelwertes c

Falls 1. vorliegt wähle [a,b]-Gleichverteilung

Falls 1. und 2. vorliegt wähle Dreiecksverteilung



Weitere Aspekte bei der Modellierung von Eingabedaten

Hier behandelt: unabhängige, identisch verteilte Daten

Reale Daten sind oft

- korreliert bzgl. der Zeitintervalle
  (z.B. Ankunftsprozess im Rechnernetz)
- korreliert bzgl. verschiedener Messgrößen
  (z.B. Größe und Gewicht von Menschen)
- über einen längeren Zeitraum nicht identisch verteilt
  (Ankunftsprozess in einem Restaurant)

In diesen Fällen sind andere Verteilungen zu verwenden

z.B. Zufallsvektoren, Markovsche Ankunftsprozesse,
nichtstationäre Poisson-Prozesse, autoregressive Modelle,
bivariate Normalverteilungen, ...

Beispiel Schätzung der Korrelation einer bivariaten Normalverteilung

$$\begin{aligned} COV(X_1, X_2) &= \frac{1}{n-1} \sum_{j=1}^{n} \left(X_{1j} - \tilde{X}_1 \right) \left(X_{2j} - \tilde{X}_2 \right) \\ &= \frac{1}{n-1} \left(\sum_{j=1}^{n} X_{1j} X_{2j} - n \tilde{X}_1 \tilde{X}_2 \right) \end{aligned}$$

$$\tilde{\rho} = \frac{COV(X_1, X_2)}{\tilde{S}_1 \tilde{S}_2}$$

Verwendung der Werte zur ZZ-Generierung (siehe Kap. 3)

Prinzip auf den n-dimensionalen Fall übertragbar!

Zeitreihenmodelle:

Einfachste Varianten zur Anpassung des Autokorrelationskoeffizienten ρ(1)
(es gilt für diese Modelle $\rho(k) = \rho^k$)

AR(1)-Prozess

$$X_h = \mu + \phi(X_{h-1} - \mu) + \epsilon_h$$

mit $\epsilon_h \sim N(0, \sigma)$ und $-1 \leq \phi \leq 1$

Es gilt $X_h \sim N\left(\mu, \sqrt{\frac{\sigma^2}{1-\phi^2}}\right)$ und $\rho(k) = \phi^k$

MA(1)-Prozess

$$X_h = \mu + \epsilon_h + \theta\epsilon_{h-1}$$

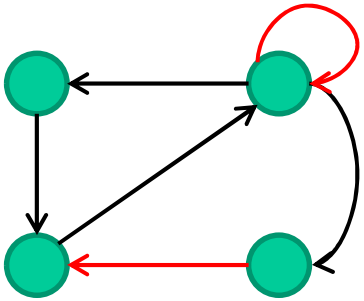
mit $\epsilon_h \sim N(0, \sigma)$ und Konstante θ

Es gilt $X_h \sim N\left(\mu, \sqrt{\sigma^2(1+\theta^2)}\right)$ und $\rho(1) = \frac{\theta}{1+\theta^2}$ sowie $\rho(k) = 0$ falls $k \geq 2$

Markovsche Ankunftsprozesse (MAPs)

(Erweiterung von Phasenverteilungen Hyperexp.- oder Erlang-Vert.)

Markov-Prozess mit markierten Transitionen



- Rote Transition erzeugen Ereignisse, schwarze sind intern
- Simulation einfach
- Theoretisch sehr große Klasse von Prozessen approximierbar
- Parameterbestimmung aufwändig