technische universität
dortmund

Bachelorarbeit

**Bootstrapping Ansätze zur Bestimmung von Konfidenzbändern für Verteilungsfunktionen**

Dennis Richter
Monat der Abgabe

Gutachter:
Prof. Dr. Peter Buchholz
Name des Zweitgutachters

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für praktische Informatik (LS 4)
https://ls4-www.cs.tu-dortmund.de

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

## 1.1 Motivation

Bei der statistischen Analyse von Daten ist es oft üblich aufwendige theoretische Modelle vorauszusetzen zu müssen, um aussagekräftige Ergebnisse über eine Stichprobe zu erhalten.

Die Auswahl eines statistischen Modells welches das wahre Modell so gut wie möglich repräsentiert, ist oft eine Herausforderung aber gleichzeitig ausschlaggebend für den Erfolg der Analyse.

Nur bedingt erfüllte Annahmen führen zu falschen Aussagen, zu spezifische Modelle hingegen lassen sich nicht Computer gestützt umsetzen und müssen per Hand analysiert werden.

Die Simulation der wahren Population durch die gegeben Stichprobe liefern hier einen Weg, diese Schwierigkeit zu umgehen.

Die Idee solcher sogenannten Resampling-Verfahren ist aus einer kleinen Anzahl von Stichproben beliebig viele Stichproben zu generieren, indem die ursprünglichen Daten als Schätzer für die Grundgesamtheit dienen, von der nun beliebig oft gesampelt werden kann.

Anstelle eine Verteilung vorauszusetzen, kann diese so mittels Monte-Carlo-Schätzung unter sehr allgemeinen Voraussetzungen angenähert werden.

Efron ... zeigt das sogenannte Bootstrap Verfahren statistisch exakt ist und neben den zahlreichen Anwendungsgebieten überraschend gute Eigenschaften haben kann.

Einziger Nachteil ist der zu leistende Rechenaufwand, allerdings wird die Rechenleistung von Computern immer besser und günstiger.

Bootstrap Verfahren sind sehr Einfach zu implementieren und liefern somit eine wunderbare Alternative gegenüber analytischen Verfahren, um die Verteilung einer Stichprobe zu bestimmen

beliebtes anwendungsgebiet sind konfidenz intervalle für den schätzer einer zufallsvariable...

Bootstrap-Ansätze zur Bestimmung von Konfidenz-Intervallen sind bereits gut bekannt und weit verbreitet.

Implementierungen sowohl der analytischen Ansätze als auch Bootstrap-Methoden sind eigentlich in allen gängigen statistischen Analyse-Tools vorhanden.

hat man statt einer jedoch mehrere zufallsvariablen, ist eine punktweise schätzung nicht sehr aussagekräftig...

stattessen benötigt man eine simultane berechischätung welche die schätzer aller zufallsvariablen mit gegebener wahrscheinlichkeit enthält, man spricht dann von konfidenzband

Methoden zu Bestimmung von Konfindenz-Bändern wurden bisher wenn überhaupt, eher in der Literatur behandelt.

Gerade Bootstrap finden noch kaum Anwendung bei der Bestimmung von Konfidenz-Bändern.

## 1.2 Zielsetzung

Ziel dieser Arbeit ist es einige solcher Ansätze vorstellen und im Kontekt von OMNeT++ umsetzten.

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

## 1.3 Aufbau der Arbeit

In Kapitel 2 werden dazu zuerst ein paar Grundlagen besprochen.

Die darauf folgenden Abschnitte sind in die drei Hautschwerpunkte der Arbeit unterteilt, Vorstellung der Algorithmen, Implementierung und Auswertung.

Schließlich folgt in Kapitel 6 noch ein Fazit. -

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

# Kapitel 2

# Grundlagen

## 2.1 Simulationsstudien

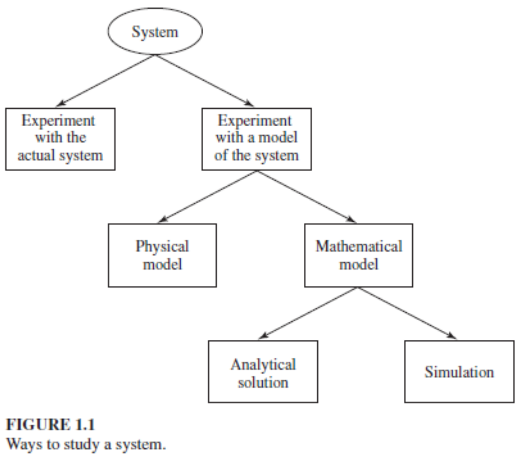


**Abbildung 2.1:** Arten von Systemstudien

### 2.1.1 M/M/1-Modell


**Abbildung 2.2:** M/M/1 Warteschlangenmodell

$$f_T(x) = \lambda e^{-\lambda t}, \quad t \geq 0 \tag{2.1}$$

$$f_S(x) = \mu e^{-\mu t}, \quad t \geq 0 \tag{2.2}$$

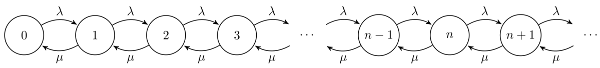


**Abbildung 2.3:** M/M/1 Zustandsübergangsdiagramm

$$P_n = (1-\rho)\rho^n \tag{2.3}$$

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \tag{2.4}$$

$$W_q = \frac{\rho}{\mu - \lambda} = \frac{\lambda}{\mu(\mu - \lambda)} \tag{2.5}$$

## 2.2 Parameter-Schätzer

$$y_i = \eta(x_i, \theta) + \epsilon_i, \quad i = 1, 2, ..., n \tag{2.6}$$

$$\epsilon \sim N(0, \sigma^2) \tag{2.7}$$

$$E(y|x) = \eta(x, \theta) \tag{2.8}$$

$$\hat{y} = \eta(x, \hat{\theta}) \tag{2.9}$$

$$\arg\min_{\theta \in \mathcal{R}} \sum_{i=1}^{n} [\eta(x_i, \theta) - y_i]^2 \tag{2.10}$$

$$\sum_{i=1}^{n} \psi(y_i, \theta) = 0 \tag{2.11}$$

### 2.2.1 Maximum-Likelihood-Schätzer

$$Lik(\theta, y) = \prod_{i=1}^{n} f(y_i, \theta) \tag{2.12}$$

$$L(\theta, y) = \log \prod_{i=1}^{n} f(y_i, \theta) = \sum_{i=1}^{n} \log f(y_i, \theta) \tag{2.13}$$

$$\frac{\partial}{\partial \theta} L(\theta, y) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(y_i, \theta) = 0 \tag{2.14}$$

$$\psi = \frac{-f}{f} \tag{2.15}$$

$$\mathrm{Var}(\theta) = [I(\theta)]^{-1} \tag{2.16}$$

$$I(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}L(\theta, y)\right) \tag{2.17}$$

$$\mathrm{Var}(\hat{\theta}) \approx \left[-\frac{\partial^2}{\partial\theta^2}L(\theta, y)\bigg|_{\theta=\hat{\theta}}\right]^{-1} \tag{2.18}$$

## 2.3 GoF-Tests

### 2.3.1 Kolmogorov-Smirnov Test

$$D = \sup_{y_i}\{F_n(y_i) - F(y_i, \theta)\} \tag{2.19}$$

### 2.3.2 Anderson-Darling Test

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\log F(y_i, \hat{\theta}) + \log(1 - F(y_{n+1-i}, \hat{\theta}))\right] \tag{2.20}$$

### 2.3.3 Chi-Quadrat Test

## 2.4 Konfidenzintervalle

$$W_q = \frac{\lambda\left[\mathrm{Var}(S) + (E(S))^2\right]}{2(1 - \lambda E(S))} \tag{2.21}$$

## 2.5 Konfidenzbänder

## 2.6 Coverage Error

## 2.7 Resampling Verfahren

### 2.7.1 Bootstrap

# Kapitel 3

# Vorstellung der Algorithmen

Banks, J. 1998. Handbook of simulation: - 7.2.3: Quantile Estimation

Was sind Voraussetzungen, die durch Bootstrap abgelöst werden?

Es werden im wesentlich zwei Ansätze vorgestellt, die Resampling einsetzen, um Konfidenzbänder zu bestimmen.

Der erste Ansatz Setzt Bootstrap ein, um

## 3.1 Standard Bootstrap

## 3.2 Bootstrapping der Residuen

## 3.3 Parametrsches Bootstrap

## 3.4 Wild Bootstrap

## 3.5 Bayes'sches Bootstrap

## 3.6 Resampling von $\partial R(\alpha)$

Banks, J. 1998. Handbook of simulation: - 5: Random Variate Generation (S. 143 ff)

# Kapitel 4

# Implementierung

## 4.1 OMNeT++

### 4.1.1 Überblick

### 4.1.2 Simulationen

## 4.2 Parameterstudien in OMNeT++

### 4.2.1 Datenerfassung

### 4.2.2 Auswertung

### 4.2.3 Darstellung

## 4.3

Kapitel 5

# Auswertung

## 5.1 Analytisches Verfahren

Eine Referenz [1].

## 5.2 parametrisches Bootstrapping

## 5.3 nicht-parametrisches Bootstrapping

## 5.4 Vergleich

# Kapitel 6

# Schlussteil

## 6.1 Fazit

## 6.2 Ausblick

## Anhang A

# Weitere Informationen

# Abbildungsverzeichnis

# Algorithmenverzeichnis

# Literaturverzeichnis

[1] Aggarwal, Alok und Jeffrey Scott Vitter: *The Input/Output Complexity of Sorting and Related Problems.* Communications of the ACM, 31(9):1116–1127, 1988.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 28. Februar 2021

Muster Mustermann