Vortreffen am 15.10.2024

Grundlagen der Datenanalyse und Statistik mit R | WS 2024/25

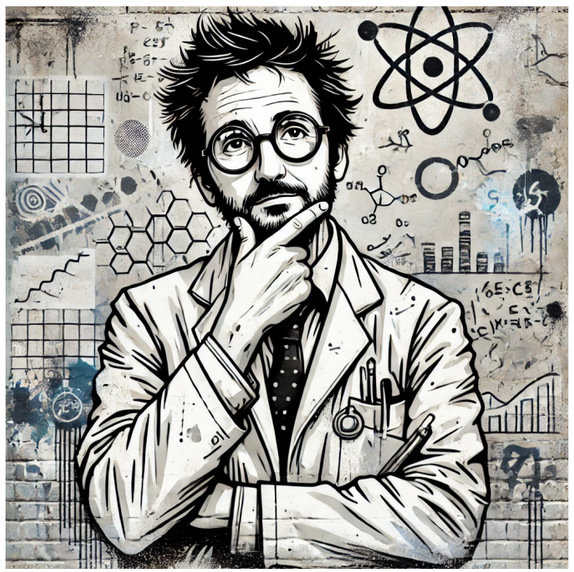
Prof. Dr. Daniel Schnitzlein

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen

Mit wem haben Sie es zu tun?

Daniel D. Schnitzlein

- Gründer und Geschäftsführer von Innside Statistics
  https://innside-statistics.de
- außerplanmäßiger Professor für Volkswirtschaftslehre
  an der Leibniz Universität Hannover
- Principle Investigator im DFG-Projekt InterVerm



Bisherige berufliche Stationen

Leibniz Universität Hannover | DIW Berlin | Humboldt Universität zu Berlin | Universität Hamburg | Friedrich-Alexander-Universität Erlangen-Nürnberg | Institut für Arbeitsmarkt- und Berufsforschung

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen

Seminarablauf

Datum	Uhrzeit	Thema
15. Oktober 2024	09:45 - 11:15 Uhr	Vorbesprechung
18. Oktober 2024	08:00 - 12:15 Uhr	Termin 1
22. November 2024	08:00 - 12:15 Uhr	Termin 2
06. Dezember 2024	08:00 - 12:15 Uhr	Termin 3
10. Januar 2025	08:00 - 12:15 Uhr	Termin 4
17. Januar 2025	08:00 - 10:00 Uhr (?)	Termin 5

Alle Termine finden **online** via **Zoom** statt. Der Link wird rechtzeitig vor dem jeweiligen Termin bereitgestellt. Nach jedem Termin wird es ein Problem Set (Hausaufgabe) geben, das bis zum nächsten Termin bearbeitet werden muss. Die Abgabe der Hausaufgaben ist Voraussetzung für die Bescheinigung der Kursteilnahme.

Ihre Umgebung auf Ihrem eigenen Laptop

- R und R-Studio werden beide unter Open-Source-Lizenzen vertrieben und sind für alle gängigen Betriebssysteme verfügbar. Das heißt, Sie können beides herunterladen und auf Ihrem eigenen Computer installieren.
- Bitte beachten Sie: Sie müssen für R und R-Studio **nicht bezahlen**. Kostenpflichtig ist nur die Enterprise Version, die wir im Kurs nicht! benötigen.
- Auch alle Datensätze, die wir in dieser Veranstaltung verwenden werden, können Sie auf Ihrem eigenen Computer nutzen und ich werde Ihnen entweder die Download-Links zur Verfügung stellen oder die Daten direkt über **Github** (oder StudOn) bereitstellen. Den Link zum Repository erhalten Sie jeweils vor dem relevanten Kurstermin.
- Das bedeutet, Sie können alle Übungen und Anwendungen mit Ihrem eigenen Gerät bearbeiten und wir müssen nicht auf die Computerräume der Universität zurückgreifen.
- Bitte bringen Sie Ihren Computer mit einer laufenden Installation von R und R-Studio zu den Veranstaltungsterminen mit.
- Idealerweise können Sie entweder **zwei Monitore** nutzen oder können mit einem **(mobilen) Gerät** am Zoommeeting teilnehmen, sodass Sie parallel an Ihrem Rechner arbeiten können.

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen

Was ist R und warum R?

- R ist eine Programiersprache und Umgebung für statistische Berechnungen.
- Die Arbeit an R begann 1992, die erste Version erschien 1993 und die erste stabile Betaversion wurde im Jahr 2000 veröffentlicht.
- R wird unter der GNU GPL v2-Lizenz vertrieben, oder anders gesagt, es ist Open Source.
- Während R in den ersten Jahren vor allem ein kleines Projekt für den akademischen Bereich war, ist es inzwischen eine der weltweit am häufigsten verwendeten Sprachen für statistische Analysen.
- R ist inzwischen in eine Vielzahl professioneller Lösungen für die Datenanalyse und Big Data-Verarbeitung integriert.
- Posit https://posit.co/ (früher R-Studio) bietet Enterprise-ready Lösungen für die Arbeit mit R (und Python) an.

Was sind R-Pakete?

- R besteht aus einer Basiskomponente, die durch sogenannte R-Pakete erweitert werden kann um bestimmte Aufgaben zu lösen.
- Das wichtigste Repository für R-Pakete, das Comprehensive R Archive Network (CRAN), listet zur Zeit 21.396 (September 2024) verfügbare Pakete auf (zuzüglich eine Unmenge von Paketen auf Github).
- Für alle (die meisten) verfügbaren Pakete ist eine umfangreiche Dokumentation verfügbar.
- Es gibt zahlreiche Tutorials (eine Google-Suche zum Thema "Introduction to R" ergibt über 6 Mio. Treffer), freie und kommerzielle Lehrbücher und eine sehr aktive Community rund um R.
- Praktisch auf jede Frage, die man am Anfang hat, kann man mit einer einfachen Google-Suche die Antwort finden.
- Mit R-Studio steht eine sehr benutzerfreundliche und ebenfalls unter einer Open-Source Lizenz stehende IDE zur Verfügung.

Wo kann man R und R-Studio herunterladen?

Sie können **R** hier herunterladen: https://cran.r-project.org/

- Auf der CRAN Seite finden Sie R-Versionen für alle gängigen (und nicht so gängigen) Betriebssysteme.

Sie können **R-Studio** hier herunterladen: https://posit.co/download/rstudio-desktop/

- Bitte beachten Sie, wir nutzen die **kostenlose Version** von R-Studio. Wenn Sie aufgefordert werden etwas zu bezahlen, oder ein Abo abzuschließen, sind Sie auf die Seite der Enterprise Version geraten. Diese benötigen wir nicht!
- Auch R-Studio ist für alle gängigen (und nicht so gängigen) Betriebssysteme verfügbar.

> **Achtung!**
>
> Sie müssen **zuerst** R und erst **danach** R-Studio installieren. Nur auf diese Weise erkennt R-Studio Ihre R-Version automatisch.

Liste der verwendeten Pakete

```r
# Notwendige Pakete

install.packages("tidyverse")
install.packages("psych")
install.packages("easystats")
install.packages("palmerpenguins")
install.packages("knitr")
install.packages("stargazer")
install.packages("gt")
install.packages("gtsummary")
install.packages("MASS")
install.packages("plotly")
install.packages("ggthemes")
```

```r
# Optionale Pakete

install.packages("cleaner")
install.packages("ineq")
install.packages("gglorenz")
install.packages("ggTimeSeries")
install.packages("ggpubr")
install.packages("readr")
install.packages("ggwordcloud")
install.packages("CGPfunctions")
install.packages("caret")
install.packages("boot")
install.packages("lubridate")
```

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen

(Vorläufige) Themenliste

1. **Einführung in R und R-Studio:** Überblick über die Installations- und Einrichtungsprozesse | Grundlegende Funktionen und Bedienung von R und R-Studio
2. **Grundlagen der Statistiksprache R:** Syntax und Datenstrukturen in R | Einführung in Funktionen und Pakete
3. **Datenmanagement in R:** Methoden der Datenorganisation und -vorbereitung | Importieren, Bereinigen und Transformieren von Datensätzen
4. **Einführung in die Pakete des tidyverse:** Überblick über die wichtigsten tidyverse-Pakete wie z.B. dplyr und ggplot2 | Anwendung dieser Pakete zur effizienten Datenanalyse und -visualisierung
5. **Deskriptive Statistik in R:** Berechnung und Interpretation grundlegender statistischer Kennzahlen | Anwendung von deskriptiven Methoden zur Datenexploration | Einführung in die statistische Modellierung am Beispiel linearer Modelle
6. **Datenvisualisierung in R:** Erstellen von publikationsreifen Grafiken und Diagrammen mit ggplot2 | Gestaltung und Interpretation von Datenvisualisierungen zur Unterstützung der Datenanalyse

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen

Bis Freitag!

GitHub-Repo: https://github.com/dschnitzlein/GDSR_0b_Vortreffen