Ein Leben mit Zahlen, vom Datenjournalist zum Kryptographen

Niclas Richter

1. September 2023

Handout zum Vortrag *Ein Leben mit Zahlen, vom Datenjournalist zum Kryptographen* am 01. September 2023 an der Ernst-Haeckel-Oberschule, Berlin. Hier stehen weitere Informationen zu Studien- und Berufsmöglichkeiten und Verweise auf die im Vortrag erwähnten Beispiele.

Warum Daten wichtig sind

Betrachte das Fallbeispiel: Eine Ärztin wirbt für ihre neue Methode für Lebertransplantationen. Bei ihrer Methode hatten 3 von 62 Patient:innen Komplikationen, während durchschnittlich 10% alle Patient:innen in den USA Komplikationen nach einer Lebertransplantation haben. Ist die neue Methodik besser?[1]

Die Antwort ist nein. In der Statistik gibt es sogenannte *Hypothesentests* die erlauben zu beurteilen, ob eine Beobachtung nur zufällig ist, oder die Beobachtung eine signifikante Änderung darstellt. Diese helfen Unterschiede zu detektieren und können auch unterschiedliche Ausgangslagen herausrechnen.

[1] Beispiel genommen aus: . (Ein sehr schönes und frei verfügbares Buch zu Statistik!)

Mine Çetinkaya Rundel and Johanna Hardin. *Introduction to Modern Statistics*. First edition, 2021. ISBN 9781943450145. URL `https://openintro-ims.netlify.app/`

Daten sind in der Gesetzgebung oftmals wichtig, wenn es um das festlegen von Schadstoffgrenzwerten geht. Der Grenzwert für Stickstoffdioxid liegt bei 40 $\mu g/m^3$[2] Zur Einhaltung der Grenzwerte werden Messstationen benötigt und diese müssen ausgelesen werden, idealerweise vollautomatisch.

[2] Mehr Hintergründe hat das Umweltbundesamt auf seiner Website `https://www.umweltbundesamt.de/themen/stickstoffdioxid-belastung-hintergrund-zu-eu`.

Warum es sich lohnt

Berufe mit Zahlenbezug sind oftmals sehr gut bezahlt, wie man im Boxplot in der Abbildung 1 sieht. Allerdings gibt es oftmals einen Unterschied zwischen akademischen Berufen und nicht-akademischen Berufen. Eine Faustregel ist, je höher der Bildungsabschluss, desto höher ist das Einkommen nach der Ausbildung ist. Nach einem Studium arbeitet man häufig an einem Computer und hat fast nur sitzende Tätigkeiten. Dies kann negative gesundheitliche Auswirkungen haben.

Wenn man allerdings Mathematik oder Informatik studiert, ist die Tätigkeit eher abstrakt und man sollte einen gewissen Hang zur

Ein Boxplot ist ein Diagramm, das uns hilft, Daten zu verstehen und Muster zu erkennen. In einem *Boxplot* siehst du eine rechteckige Box und manchmal auch Linien darin. Die Box zeigt, wo die meisten Daten liegen, also den Bereich, in dem sich die meisten Werte befinden. Die Linien, die aus der Box herausragen, zeigen die Ausreißer, also Werte, die weit von den anderen entfernt sind. Ein Boxplot hilft uns zu sehen, wie die Daten verteilt sind, ob es viele oder wenige Werte gibt und ob es ungewöhnliche Werte gibt, die anders sind als die anderen.

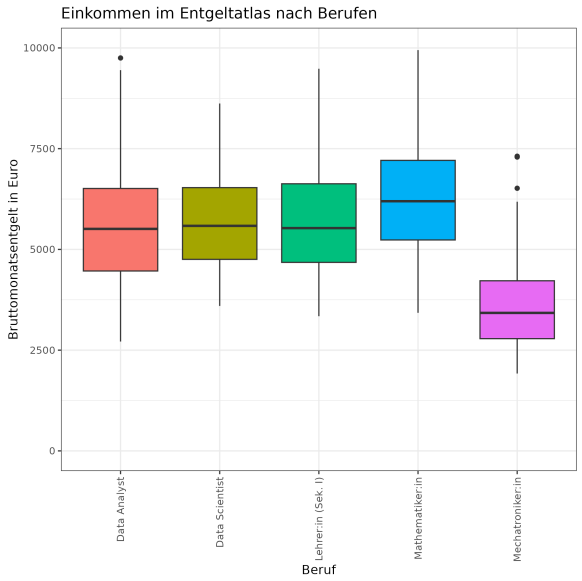

Abb. 1: Simulierte Einkommensverteilung von Lehrer:innen, Mechatroniker:innen, Data Scientists / Analysts, sowie Mathematiker:innen.

Abstraktion haben. Ein konkretes Produkt erstellt man nur selten, sondern arbeitet mit abstrakten Modellen und Prototypen.

Berufsfelder

Ein Data Scientist ist eine Art Daten-Detektiv. Er oder sie findet Muster in großen Datensätzen, oder auch sehr kleinen. Dabei werden modernste Methoden verwendet und man forscht zu neuen Methoden, die bessere Ergebnisse für bekannte Probleme liefern. Die Probleme umfassen oftmals Klassifikationen, Schätzungen von numerischen Ergebnissen oder Vorhersagen von bestimmten Ereignissen.

Ein Data Analyst ist ein Übersetzer von komplexen Daten zu einfach verständlichen Informationen. Man hilft Unternehmen oder Organisationen herauszufinden, welche Produkte sich wann am besten Verkaufen. Beispielsweise will ein Unternehmen wissen, wie viel Eis mehr verkauft wird, wenn es im Juli 3°C wärmer ist.

Oftmals arbeiten diese 3 Berufsfelder Hand in Hand innerhalb eines Teams. Manchmal ist es schwer feste Grenzen zu setzen und die Übergänge sind fließend.

Ein Data Engineer ist ein Baumeister oder eine Baumeisterin für Daten. Er oder sie baut eine Infrastruktur auf um Daten automatisiert zu verarbeiten oder andere diese einfach nutzen können. Data Engineers sorgen für einen reibungslosen Ablauf bei der Arbeit mit Daten, von der Gewinnung der Daten bis zur Speicherung und Löschung.

Der Berufsalltag dieser Berufe findet vor einem Computer statt. Man nutzt Skriptsprachen with Python[3]. Auch wird die Programmiersprache R genutzt, die speziell für Statistik und Datenverarbeitung genutzt werden kann.[4] Man kann sehr einfach Grafiken erstellen mit R, da Datenvisualisierung täglich vorkommt.

Der Code für die Grafik auf der rechten Seite ist sehr kurz mit:

```
library(tidyverse)
tibble("Male" = rnorm(200, 1.75, 0.1),
 "Female" = rnorm(200, 1.65, 0.12),
 "Diverse"=rnorm(200, 1.7, 0.14)) %>%
  pivot_longer(cols = everything(),
     names_to = "Sex", values_to = "Height") %>%
  ggplot(aes(x=Sex, y=Height, fill=Sex)) +
  geom_violin() +
  theme_bw() +
  labs(title = "Randomly generated Heights")
```

[3] Python ist eine allgemeine Programmiersprache, die besonders einfach zu erlernen ist. Siehe auch `https://www.python.org/about/gettingstarted/`

[4] Hier ein kleines Tutorial zu R: `https://www.statmethods.net/r-tutorial/index.html`

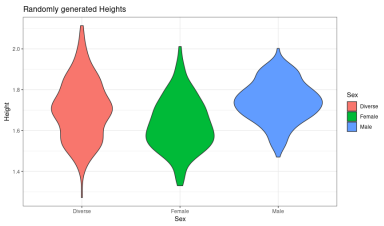

Abb. 2: Minibeispiel für eine R-Grafik

Datenjournalismus ist eine moderne Form des Journalismus, bei der Daten analysiert und visuell aufbereitet werden. Das Ziel dabei ist es, komplexe Sachverhalte verständlich zu vermitteln. Die Deutsche Welle bietet eine Vielzahl kostenfreier Beiträge an, in denen Daten genutzt werden, um Informationen zu veranschaulichen[5]. Die Daten reichen von Themen von Umweltthemen bis hin zu Kriminalität. Allerdings ist dies nie rein objektiv, sondern man trifft immer ein Wahl, egal ob es die Farbgebung ist oder der Beginn einer Zeitreihe. Zu der Objektivität von Daten ist das Buch von Tin Fischer [6] ein guter Einstieg.

[5] Zu finden unter `https://www.dw.com/en/data/t-43091100`

[6] Tin Fischer. *Linke Daten, Rechte Daten*. HOFFMANN UND CAMPE VERLAG GmbH, 2022. URL `https://hoffmann-und-campe.de/products/53007-linke-daten-rechte-daten`

Ohne Kryptographie haben wir keine Privatssphäre. Kryptographie erlaubt es unsere Daten sicher zu versenden und zu wissen, dass sie vom richtigen Sender kommen. Sei es bei einer Banküberweisung, dem Schließen der Fenster per App, oder dem Aufschließen des Autos. Einen verständlichen Einblick liefert [7]. Berufe, die dies benötigen, sind Cypersecruity-Expert:innen, IT-Sicherheitsanalyst:innen oder Datenschutzbeauftragte.

[7] Simon Rubinstein-Salzedo. *Cryptography*. Springer. ISBN 9783319948171

Anwendungsbeispiele

Life Science

Die Sicherheit von Medikamenten beruht auf Statistik, indem wir mit statistischen Methoden nachweisen, dass es nur akzeptierbare Nebenwirkungen gibt und die Medikamente auch

wirken. Eine weitere Punkt ist die Entwicklung von Medikamenten, so hilft *Computer Aided Drug Design (CADD)* bei der Entwicklung neuer Medikamente[8].

[8] Eine Übersicht gibt es unter `https://encyclopedia.pub/entry/24806`

Pflanzenschutz profitiert zunehmend von mathematischen Modellen, so werden Plfanzenmedikamente über mathematische Verfahren gefunden (s.o.), oder die Epidemologie von Pflanzenkrankheiten modeliert um rechtzeitig Gegenmaßnahmen zu ergreifen [9]. Auch kann mit Methoden der Bioinformatik Gene identifiziert werden, die Pflanzen besonders resistent gegenüber dem Klimaschutz machen[10].

[9] Noelia Bazarra, Michele Colturato, José R. Fernández, Maria Grazia Naso, Anna Simonetto, and Gianni Gilioli. Analysis of a mathematical model arising in plant disease epidemiology. *Applied Mathematics and Optimization*, 85 (2), apr 2022. DOI: 10.1007/s00245-022-09858-z

[10] Siehe: `https://www.bbc.com/news/science-environment-28789716`

Die mathematische Modellierung des Herzens hilft bei der Erkennung von Herzinfarkten und anderen Herz-Keislauferkrankungen. Bei der Modellierung kommen verschiedene physikalische Phänomene zusammen wie elektrische Impulse und Strömungsmechanik

Klimavorhersagen / Meteorologie

Wettervorhersagen sind eine Neuerung, die vor 200 Jahren noch undenkbar war. Heute ist es möglich, dass Wetter einige Tage im voraus zu berechnen, in dem man komplizierte Gleichungen[11] nährungsweise löst. Warum nur nährungsweise? Naja, es sind solche:

[11] Jean Coiffier. *Fundamentals of Numerical Weather Prediction*. Cambridge University Press, dec 2011. DOI: 10.1017/cbo9780511734458

$$
\begin{aligned}
\frac{dV_3}{dt} &= -2\Omega \times V_3 - \frac{1}{\rho}\nabla p - \nabla\Phi + F \\
\frac{dT}{dt} &= \frac{R}{C_p}\frac{T}{p}\frac{dp}{dt} + \frac{Q}{C_p} \\
\frac{d\rho}{dt} &= -\rho \operatorname{div} V_3 \\
\frac{dq}{dt} &= M \\
p &= \rho R T
\end{aligned}
$$

Diese Gleichungen muss man nicht verstehen, insbesondere weil die Lösung(en) keine Zahlen sind, sondern mathematische Funktionen, die man noch nicht einmal aufschreiben kann. Auch muss man die Lösung in einem unendlichdimensionalen Raum suchen. Das ist überhaupt nicht einfach. Hinzu kommt, dass die Erde kein 3-dimensionaler Euklidischer Raum ist, sondern eher einer Kartoffel ähnelt, das macht die Sache nochmal komplizierter.

Einen Überlick gibt der Artikel

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, sep 2015. DOI: 10.1038/nature14956

Vorhersagen über das langfristige Klima sind nochmal schwieriger. Doch warum setzt man nicht KI ein? Weil KI oft nur so gut ist, wie ihre Trainingsdaten und durch den Klimawandel wandelt

sich die Grundgesamtheit der Wetterereignisse. Extremwetter-ereignisse werden häufiger, aber die KI hat in ihren Trainingsdaten nur sehr wenige Extremwetterereignisse, daher ist es wie so häufig nur eine Ergänzung für die physikalischen Modelle und der Arbeit von Meteorolog:innen[12]

[12] Der SWR hat auch einen Beitrag dazu gebracht, der hier veröffentlicht wurde.

Die digitale Welt

Zahlen alle den gleichen Preis? Als Oma noch in den Laden ging und ein handschriftliches Preisschild las, war es noch so. Heute ist es nicht mehr so. Unternehmen verwenden im Bereich des E-Commerce dynamische Preisalgorithmen, um ihre Preise nach verschiedenen Parametern, wie der Tageszeit, der Verfügbarkeit der Waren oder auch dem/der Kund:in. Die Preisstrategie wird nahezu nie offengelegt, so dass es schwer ist eine konkrete Diskriminierung nachzuweisen. Trotzdem ist bekannt, dass Frauen bereit sind höhere Preise für bestimmte Produkte zu zahlen [13] und Frauen für bestimmte Güter und Dienstleistungen bereits mehr zahlen[14]. Auch bei Platformen wie *Uber* oder *AirBnB* kann es zu Diskriminierung kommen, wenn schwarze Uberfahrer schlechter bewertet werden als weiße Uberfahrer[15] Weitere Beispiele findet man in dem sehr zugänglichen Buch von Hanna Fry[16].

engl. dynamic pricing, gutes Schlagwort zum eintauchen in die Thematik

[13] Iris an der Halden and Maria Wersig. *Preisdifferenzierung nach Geschlecht in Deutschland*. Nomos, 2018

[14] ebd.
Ein Beisipiel wie soetwas passiert, gibt der einseitige Blogartikel `https://mypermissions.com/blog/2017/07/30/the-ugliness-of-dynamic-pricing/`.

[15] Siehe: `https://hbr.org/2016/12/fixing-discrimination-in-online-marketplaces`, Abrufdatum 27.08.23

[16] Hannah Fry. *Hello World*. C.H. Beck Verlag, München, 2019. ISBN 9783406732195

Recommendation Systems sind das Kernstück sozialer Medien, den sie entscheiden, was die Nutzer sehen. Dies erstreckt sich von TikTok Videos, Vorschläge bei Amazon oder Tinder Matches. Der englische Wikipedia Artikel [17] ist sehr gut zu diesem Thema.

[17] `https://en.wikipedia.org/wiki/Recommender_system`

Ausbildungsmöglichkeiten

Berufsausbildungen sind eher weniger geeignet, wenn man mit Daten und Zahlen arbeiten will. Es gibt Ausbildungen in diesem Bereich, wie die zum Mathematisch-Softwaretechnologischen Assistent:in oder Fachinformatiker:in, aber oftmals ist der Beruf danach in den Aufgaben konkreter als das Abstraktionslevel.

Ein Studium im MINT-Bereich ist häufig der Einstieg. Dabei steht MINT für **M**athematik, **I**nformatik, **N**aturwissenschaften und **T**echnik. In allen diesen Bereichen gibt es Vertiefungen, die sich mit der Verwendung von Daten beschäftigen, sei es Data Assimiliation, Computational Chemistry, Signal Verarbeitung oder Computational Engineering. Letzteres gibt es sogar als eigenständigen als eigenständigen

Studiengangs. Allerdings sind in allen Bereichen Programmierkenntnisse notwendig. Oftmals ist es nachrangig, wo man einen Bachelor macht und beim Master kann man immer noch die Uni wechseln. Oftmals hat man vor dem Bachelor keine Ahnung, wo es genau hingehen soll, während man zu Beginn des Masterstudiums bessere Vorstellungen von den eigenen Zielen hat. (Bei bis zu 70% der Studienanfänger:innen ist die Nähe zum bisherigen Wohnort der ausschlaggebene Punkt bei der Wahl der Uni.)

EIN QUEREINSTIEG IST MANCHMAL MÖGLICH, so können quantitative Sozialwissenschaftler:innen über Datenauswertungsmethoden in den Datenbereich kommen. Diese haben den Vorteil, dass sie gut mit Daten umgehen können, die einen Bias haben. Ein Bias ist eine systematische Abweichung vom tatsächlichen Wert. Allerdings sollte man dies auch belegen können, in dem man spezielle Kurse im Studium dazu belegt.

Literatur

Iris an der Halden and Maria Wersig. *Preisdifferenzierung nach Geschlecht in Deutschland*. Nomos, 2018.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, sep 2015. DOI: 10.1038/nature14956.

Noelia Bazarra, Michele Colturato, José R. Fernández, Maria Grazia Naso, Anna Simonetto, and Gianni Gilioli. Analysis of a mathematical model arising in plant disease epidemiology. *Applied Mathematics and Optimization*, 85(2), apr 2022. DOI: 10.1007/s00245-022-09858-z.

Jean Coiffier. *Fundamentals of Numerical Weather Prediction*. Cambridge University Press, dec 2011. DOI: 10.1017/cbo9780511734458.

Tin Fischer. *Linke Daten, Rechte Daten*. HOFFMANN UND CAMPE VERLAG GmbH, 2022. URL `https://hoffmann-und-campe.de/products/53007-linke-daten-rechte-daten`.

Hannah Fry. *Hello World*. C.H. Beck Verlag, München, 2019. ISBN 9783406732195.

Simon Rubinstein-Salzedo. *Cryptography*. Springer. ISBN 9783319948171.

Mine Çetinkaya Rundel and Johanna Hardin. *Introduction to Modern Statistics*. First edition, 2021. ISBN 9781943450145. URL `https://openintro-ims.netlify.app/`.