Aufgabenblatt 2.2: Statistische Grundlagen

TU Dortmund, Lehrstuhl für Mikroökonomie, Wintersemester 2026/27

Kausalanalyse: Economics meets Data Science

Marc Justin Schmidt

05. November 2026

Aufgabe 1: Monte Carlo Simulation

Angenommen, das monatliche Einkommen einer Population sei normalverteilt mit einem wahren Mittelwert von 3.000€ und einer Standardabweichung von 500€. Wir untersuchen, wie gut Stichprobenmittelwerte den wahren zentralen Lageparameter schätzen.

a. Simulieren Sie die Population mit der Funktion `rnorm()`.
b. Ziehen Sie 1.000 Stichproben mit jeweils 10 Beobachtungen. Berechnen Sie für jede Stichprobe den Mittelwert und visualisieren Sie die Verteilung dieser 1.000 Mittelwerte in einem Histogramm. Intepretieren Sie, ob der Schätzer erwartungstreu ist.
c. Wiederholen Sie Schritt b mit Stichprobengrößen von 50 und 100. Vergleichen Sie die Streuung der Mittelwerte. Interpretieren Sie bzgl. der Konsistenz des Mittelwert-Schätzers.
d. Definieren Sie zusätzlich einen neuen Schätzer: den Mittelwert nur der ersten Hälfte jeder Stichprobe. Welcher der Schätzer ist effizienter?

Aufgabe 2: Zentraler Grenzwertsatz (CLT)

Der Zentrale Grenzwertsatz (CLT) besagt: Auch wenn Daten nicht normalverteilt sind, folgt die Verteilung der Stichprobenmittel bei großen Stichproben näherungsweise einer Normalverteilung. Nehmen Sie an, Sie analysieren die Akkulaufzeiten von Smartphones, die exponentialverteilt mit einem durchschnittlichen (erwarteten) Wert von 7 Stunden sind. Für eine Exponentialverteiltung gilt $\mathbb{E}[X] = 1/\lambda$.

a. Simulieren Sie 10.000 Akkulaufzeiten, die exponentialverteilt mit Erwartungswert 7 sind (`rexp()`).
b. Ziehen Sie 5.000 Stichproben mit jeweils 30 Nutzer*innen und berechnen Sie die Mittelwerte.
c. Erstellen Sie ein Histogramm der Mittelwerte.
d. Interpretieren Sie: Warum ergibt sich eine Normalverteilung, obwohl die Einzeldaten exponentialverteilt sind?

Aufgabe 3: Hypothesentests und p-Wert

Stellen Sie sich vor, der Basketball-Europa- und Weltmeister Dennis Schröder trainiert Freiwürfe. Er gibt an, im Durchschnitt 70 % seiner Würfe zu treffen.

a. Sie schauen sich die nächsten 100 Würfe von ihm an und notieren jedes Mal, ob er trifft (1) oder daneben wirft (0). Daraus entsteht Ihre Stichprobe. Simulieren Sie diese Stichprobe mit `rbinom()`.
b. Testen Sie die Nullhypothese $H_0 : p = 0.7$ mit einem Ein-Stichproben-t-Test auf den Mittelwert (d. h. Schröder trifft im Durchschnitt 70 % seiner Freiwürfe). Interpretieren Sie die Ausgabe: Wann lehnen Sie die Nullhypothese ab?
c. Der Trainer behauptet jedoch, Schröder treffe noch besser als die von ihm angegebenen 70%. Wiederholen Sie den Test für $H_0 : p = 0.8$. Was verändert sich?
d. Notieren Sie den p-Wert aus Ihrem Test und erklären Sie in eigenen Worten, was dieser p-Wert im Kontext bedeutet.
e. Variieren Sie die Stichprobengröße und diskutieren Sie, wie sich der p-Wert verändert. Was sagt dies über die Aussagekraft von Tests aus?

Aufgabe 4: Konfidenzintervalle

Harry Kane möchte wissen, wie zuverlässig er Elfmeter trifft. In den letzten 30 Elfmetern hat er 25 Treffer erzielt. Das Team möchte ein Konfidenzintervall für seine wahre Trefferwahrscheinlichkeit berechnen.

a. Berechnen Sie das 95%-Konfidenzintervall für die wahre Trefferwahrscheinlichkeit p. Nutzen Sie `rbinom()` und `prop.test()`.
b. Was bedeutet der Bereich, den das KI angibt? Wenn ein Trainer behauptet, Kane treffe *immer*, liegt seine Aussage im Intervall oder außerhalb? Wie sicher können wir auf Basis dieser 30 Schüsse sein?
c. Simulieren Sie, wie sich das Konfidenzintervall verändert, wenn Kane in Zukunft mehr Elfmeter schießt (300) und davon 250 trifft. Wie verändert sich die Breite des Intervalls? Warum wird das Intervall bei größerer Stichprobe enger?

Aufgabe A1: Schätzer auf echten Daten

Verwenden Sie den bereitgestellten Datensatz `students_performance.csv` (von Kaggle unter https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?resource=download). Er enthält Noten in Mathematik, Lesen und Schreiben für 1.000 Schüler*innen sowie demografische Merkmale.

a. Laden Sie den Dateneinsatz ein (z.B. mit `read.csv()`).
b. Berechnen Sie den Populationsmittelwert der Mathematiknoten und betrachten Sie ihn als "wahren Wert".
c. Ziehen Sie 500 Stichproben mit jeweils 20 Schüler*innen und berechnen Sie den Mittelwert jeder Stichprobe. Stellen Sie die Verteilung der Mittelwerte grafisch dar.
d. Wiederholen Sie die Simulation mit Stichprobengrößen 50 und 200. Interpretieren Sie wie sich die Streuung der Mittelwerte ändert.
e. Erstellen Sie eine neue Variable `pass_math`, die den Wert 1 annimmt, wenn der Mathematikscore ≥ 60 beträgt, und 0 sonst. Berechnen Sie anschließend die folgenden Wahrscheinlichkeiten mithilfe relativer Häufigkeiten:
   - $\mathbb{P}(X = 1)$: Wahrscheinlichkeit, dass ein'e Schüler'in die Mathematikprüfung besteht.
   - $\mathbb{P}(Y = "female")$: Wahrscheinlichkeit, dass ein'e zufällig gezogene Schüler'in weiblich ist.
   - $\mathbb{P}(X = 1, Y = "female")$: Gemeinsame Wahrscheinlichkeit, dass ein'e zufällig gezogene Schüler'in weiblich ist und die Mathematikprüfung besteht.
f. Berechnen und vergleichen Sie die bedingten Wahrscheinlichkeiten:
   - $\mathbb{P}(X = 1|Y = "female")$: Wahrscheinlichkeit, dass ein'e Schüler'in die Mathematikprüfung besteht, gegeben, dass sie weiblich ist.
   - $\mathbb{P}(Y = "female"|X = 1)$: Wahrscheinlichkeit, dass ein'e Schüler'in weiblich ist, gegeben, dass sie die Mathematikprüfung besteht.

Aufgabe A2: Weiteres praktisches Monte Carlo Beispiel

Stellen Sie sich vor, Sie testen die ausgegebene Kaffeemenge von Starbucks-Maschinen. Im Durchschnitt sollen 200 ml ausgegeben werden, tatsächlich schwankt es aber leicht (Standardabweichung ca. 10 ml).

a. Simulieren Sie 10.000-mal Stichproben von jeweils 5 Bechern aus Maschinen, deren Füllmengen $N(200, 10^2)$ (d.h. Mittelwert 200 ml, Standardabweichung 10 ml) verteilt sind. Berechnen Sie jedes Mal den Stichprobenmittelwert.
b. In der Realität würden Sie keine 10.000 Starbucks besuchen können. Wiederholen Sie daher die Simulation, aber ziehen Sie nur fünf Stichproben mit jeweils 5 Bechern. Vergleichen Sie diesen Mittelwert mit dem wahren Wert von 200 ml.
c. Erhöhen Sie die Stichprobengröße auf 100 Becher. Wie verändert sich die Stabilität des Mittelwerts im Vergleich zu den beiden anderen Fällen?
d. Stellen Sie sich vor, Sie hätten keine Zeit, mehrere Becher pro Starbucks zu testen. Verwenden Sie in jeder Stichprobe nur die erste Beobachtung als Schätzer. Vergleichen Sie dessen Varianz mit dem Mittelwertschätzer bei $n = 5$.