Aufgabenblatt 2.1: Statistische Grundlagen (Wahrscheinlichkeiten)

TU Dortmund, Lehrstuhl für Mikroökonomie, Wintersemester 2026/27

Kausalanalyse: Economics meets Data Science

Marc Justin Schmidt

22. Oktober 2026

Aufgabe 1: Vom Zufallsexperiment zur Verteilung

Wir betrachten einen fairen sechsseitigen Würfel. Warum ist der Befehl `set.seed()` hilfreich?

a. Simulieren Sie zunächst 20 Würfe. Verwenden Sie dazu `sample()` und speichern Sie die Ergebnisse in einem Vektor `wuerfe_20`.

b. Bestimmen Sie mit `table()` die absoluten Häufigkeiten der sechs möglichen Ergebnisse. Berechnen Sie anschließend mit `prop.table()` die relativen Häufigkeiten.

c. Wiederholen Sie die Simulation mit 10.000 Würfen. Stellen Sie die relativen Häufigkeiten in einem Balkendiagramm dar und ergänzen Sie eine horizontale Linie bei der theoretischen Wahrscheinlichkeit. Vergleichen Sie die Ergebnisse für 20 und 10.000 Würfe und erklären Sie in eigenen Worten den Unterschied zwischen einer theoretischen Wahrscheinlichkeit und einer empirischen relativen Häufigkeit.

d. Berechnen Sie anhand der 10.000 simulierten Würfe die folgenden empirischen Wahrscheinlichkeiten

$$\mathbb{P}(X = 4), \qquad \mathbb{P}(X \leq 4), \qquad \mathbb{P}(X > 4).$$

e. Berechnen Sie für $x = 1, \ldots, 6$ die kumulierten relativen Häufigkeiten mit `cumsum()`. Stellen Sie damit die empirische Verteilungsfunktion $F_X(x) = \mathbb{P}(X \leq x)$ grafisch dar.

Aufgabe 2: Gemeinsame und marginale Wahrscheinlichkeiten

Verwenden Sie den bereitgestellten Datensatz `students_performance.csv`. Er enthält Ergebnisse von 1.000 Schüler:innen in Mathematik, Lesen und Schreiben sowie verschiedene demografische Merkmale. Behandeln Sie den Datensatz in dieser Aufgabe als Population.

a. Laden Sie den Datensatz mit `read.csv()` und verschaffen Sie sich mit `head()` (oder `tail()`)und `names()` einen Überblick. Verwenden Sie auch den `summary()`-Befehl.

b. Erstellen Sie die Variable `pass_math`, die den Wert 1 annimmt, wenn der Mathematikscore mindestens 60 beträgt, und andernfalls den Wert 0.

c. Erstellen Sie eine Kreuztabelle von `pass_math` und Geschlecht. Wandeln Sie diese mit `prop.table()` in eine Tabelle gemeinsamer Wahrscheinlichkeiten um.

d. Berechnen Sie aus dem Datensatz die folgenden Wahrscheinlichkeiten:

- $\mathbb{P}(\text{bestanden})$,
- $\mathbb{P}(\text{weiblich})$,
- $\mathbb{P}(\text{bestanden}, \text{weiblich})$,
- $\mathbb{P}(\text{nicht bestanden}, \text{männlich})$.

e. Erstellen Sie ein gestapeltes Balkendiagramm der gemeinsamen relativen Häufigkeiten aus c: je ein Balken pro Geschlecht, unterteilt nach *bestanden* und *nicht bestanden*.

Aufgabe 3: Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Verwenden Sie weiterhin den Datensatz aus Aufgabe 2.

a. Berechnen Sie die folgenden bedingten Wahrscheinlichkeiten durch geeignetes Filtern und anschließende Bildung relativer Häufigkeiten:

$$\mathbb{P}(\text{bestanden} \mid \text{weiblich}) \quad \text{und} \quad \mathbb{P}(\text{weiblich} \mid \text{bestanden})$$

b. Berechnen Sie beide bedingten Wahrscheinlichkeiten zusätzlich direkt aus gemeinsamen und marginalen Wahrscheinlichkeiten. Überprüfen Sie damit die folgende Formel

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}$$

c. Erklären Sie, weshalb $\mathbb{P}(A \mid B)$ im Allgemeinen nicht dasselbe ist wie $\mathbb{P}(B \mid A)$.

d. Prüfen Sie anhand der beobachteten relativen Häufigkeiten die beiden Unabhängigkeitskriterien (siehe unten). Wir erwarten in empirischen Daten in der Regel keine perfekte Unabhängigkeit. Beurteilen Sie anhand der Größenordnung der Unterschiede vorsichtig, ob die Variablen im Datensatz unabhängig erscheinen.

$$\mathbb{P}(\text{bestanden} \mid \text{weiblich}) \stackrel{?}{=} \mathbb{P}(\text{bestanden})$$

$$\mathbb{P}(\text{bestanden}, \text{weiblich}) \stackrel{?}{=} \mathbb{P}(\text{bestanden})\mathbb{P}(\text{weiblich})$$

e. Angenommen, die Bestehenswahrscheinlichkeiten unterscheiden sich zwischen den Gruppen. Begründen Sie, warum dieser statistische Zusammenhang allein noch nicht zeigt, dass das Geschlecht den Prüfungserfolg kausal beeinflusst.

Aufgabe 4: Satz von Bayes und Basisraten

Ein diagnostischer Test wird in einer Population von 10.000 Personen eingesetzt. Ein Prozent der Personen ist tatsächlich erkrankt. Der Test ist bei 90 Prozent der erkrankten Personen positiv. Bei fünf Prozent der gesunden Personen fällt der Test fälschlicherweise ebenfalls positiv aus.

a. Wie viele der 10.000 Personen sind erkrankt und wie viele gesund?

b. Vervollständigen Sie die folgende Vierfeldertafel mit absoluten Häufigkeiten:

	Test positiv	Test negativ	Gesamt
erkrankt			
gesund			
Gesamt			10.000

c. Eine zufällig ausgewählte Person erhält ein negatives Testergebnis. Berechnen Sie anhand der Tabelle die untenstehende Wahrscheinlichkeit. Interpretieren Sie das Ergebnis.

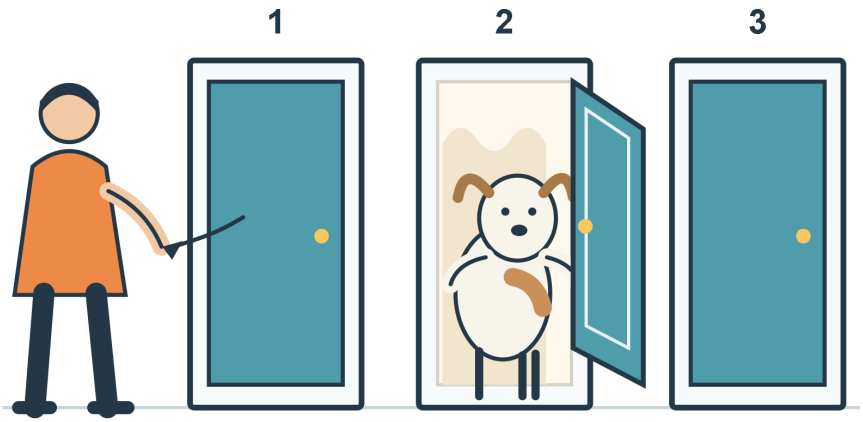
$$\mathbb{P}(\text{gesund} \mid \text{Test negativ})$$

d. Stellen Sie sich nun vor, dass keine Vierfeldertafel für eine konkrete Population von 10.000 Personen vorliegt. Aus getrennten Erhebungen sind jedoch die zu Beginn genannten Raten bekannt. Da die gemeinsame Wahrscheinlichkeit $\mathbb{P}(v \cap \text{Test negativ})$ nicht direkt beobachtet wird, nutzen Sie den Satz von Bayes, um aus $\mathbb{P}(\text{Test negativ} \mid \text{gesund})$ die Wahrscheinlichkeit $\mathbb{P}(\text{gesund} \mid \text{Test negativ})$ zu bestimmen.

$$\mathbb{P}(\text{gesund} \mid \text{Test negativ}) = \frac{\mathbb{P}(\text{Test negativ} \mid \text{gesund})\mathbb{P}(\text{gesund})}{\mathbb{P}(\text{Test negativ})}$$

Aufgabe A1: Monty Hall

In einer Spielshow stehen drei Türen zur Auswahl. Hinter einer Tür befindet sich ein Auto, hinter den beiden anderen jeweils eine Ziege. Sie wählen zunächst eine Tür. Der Moderator kennt den Standort des Autos, öffnet immer eine andere Tür mit einer Ziege und bietet Ihnen anschließend an, zur verbleibenden geschlossenen Tür zu wechseln.



a. Intuitiv: sollten Sie bei Ihrer ersten Wahl bleiben, wechseln oder ist beides gleich gut? Begründen Sie Ihre Vermutung.

b. Bestimmen Sie die Gewinnwahrscheinlichkeit der Strategie *Bleiben*.

c. Schreiben Sie eine Funktion `monty_hall(wechseln)`, die eine Spielrunde simuliert und `TRUE` zurückgibt, wenn Sie das Auto gewinnen. Beachten Sie dabei die Informationsregel des Moderators: Er öffnet weder Ihre Tür noch die Tür mit dem Auto.

d. Simulieren Sie mit `replicate()` jeweils 10.000 Spiele für die Strategien *Bleiben* und *Wechseln*. Vergleichen Sie die relativen Gewinnhäufigkeiten mit Ihren theoretischen Ergebnissen.

e. Ändern Sie die Regel: Der Moderator kennt den Standort des Autos nicht und öffnet zufällig eine der beiden nicht gewählten Türen. Spiele, in denen er das Auto zeigt, enden sofort. Diskutieren Sie, weshalb diese neue Datenerzeugung die Interpretation der beobachteten Information verändert.

Aufgabe A2: Simpson-Paradoxon bei Hochschulzulassungen

Der in R enthaltene Datensatz `UCBAdmissions` dokumentiert Bewerbungen und Zulassungen zu sechs großen Graduate Departments der University of California, Berkeley, aus dem Jahr 1973. Die aggregierten Daten wurden bekannt, weil die Zulassungsquoten nach Geschlecht zunächst ein anderes Bild vermitteln als die nach Department aufgeteilten Daten.

a. Laden Sie den Datensatz und wandeln Sie ihn in einen `data.frame` um (mithilfe von `as.data.frame()`). Welche Variablen enthält der resultierende Datensatz? Wofür steht `Freq`?

```r
data("UCBAdmissions")
```

b. Berechnen Sie zunächst ohne Berücksichtigung des Departments die Zulassungsquote getrennt nach Geschlecht und dann anschließend für jedes Department und Geschlecht separat. Stellen Sie diese bedingten Wahrscheinlichkeiten grafisch gegenüber.

$$\mathbb{P}(\text{zugelassen} \mid \text{Geschlecht})$$

$$\mathbb{P}(\text{zugelassen} \mid \text{Geschlecht}, \text{Department})$$

c. Untersuchen Sie, wie sich die Bewerbungen der Geschlechter auf die sechs Departments verteilen.

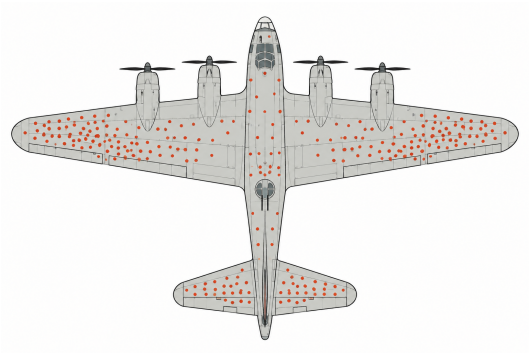
d. Erklären Sie in eigenen Worten, wie die aggregierte und die nach Department bedingte Betrachtung zu unterschiedlichen Schlussfolgerungen führen können.

Quelle: Bickel, Hammel und O'Connell (1975), „Sex Bias in Graduate Admissions: Data from Berkeley", *Science*, 187(4175), 398–404.

Aufgabe A3: Abraham Wald und die fehlenden Flugzeuge

Während des zweiten Weltkriegs untersuchte Abraham Wald die Verwundbarkeit von Flugzeugen anhand der Schäden zurückgekehrter Maschinen. Die folgende Tabelle enthält illustrative Messwerte:

Bereich des Flugzeugs	Beobachtete Treffer an zurückgekehrten Flugzeugen
Tragflächen	180
Rumpf	150
Heck	90
Motor	20
Cockpit	10



a. Stellen Sie die beobachteten Treffer in einem Balkendiagramm dar. Welche Bereiche würde man verstärken, wenn man ausschließlich die größten beobachteten Häufigkeiten betrachtet?

b. Definieren Sie die Ereignisse T_j: "Ein Treffer liegt im Bereich j" und R: "Das Flugzeug kehrt zurück". Welche bedingte Verteilung beschreibt die Tabelle: $\mathbb{P}(T_j \mid R)$ oder $\mathbb{P}(R \mid T_j)$? Welche Größe wäre für die Entscheidung über zusätzliche Panzerung eigentlich relevant?

c. Erklären Sie, weshalb wenige Treffer an Motor und Cockpit sowohl (i) geringe Trefferwahrscheinlichkeiten als auch (ii) eine geringe Rückkehrwahrscheinlichkeit nach einem Treffer bedeuten könnten. Warum lassen sich diese beiden Erklärungen mit den vorliegenden Daten nicht unterscheiden?

d. Übertragen Sie das Problem auf eine heutige Anwendung, beispielsweise Unternehmen, erfolgreiche Start-ups, Krankenhausdaten oder Studienabbrüche. Welche nicht beobachtete Gruppe könnte dort zu einem ähnlichen Fehlschluss führen?

Quelle: Wald, „A Method of Estimating Plane Vulnerability Based on Damage ofSurvivors", wiederabgedruckt durch das Center for Naval Analyses (1980): https://www.cna.org/analyses/1980/a-method-of-estimating-plane-vulnerability