Aufgabenblatt 2.1: Statistische Grundlagen (Lösungsvorschlag)

TU Dortmund, Lehrstuhl für Mikroökonomie, Wintersemester 2026/27

Kausalanalyse: Economics meets Data Science

Marc Justin Schmidt

26. 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`.

```
set.seed(123)
wuerfe_20 <- sample(1:6, size = 20, replace = TRUE)
wuerfe_20
```

```
##  [1] 3 6 3 2 2 6 3 5 4 6 6 1 2 3 5 3 3 1 4 1
```

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

```
# factor erlaubt auch nicht gewürfelte Augenzahlen
haeufig_20 <- table(factor(wuerfe_20, levels = 1:6))
haeufig_20
```

```
##
## 1 2 3 4 5 6
## 3 3 6 2 2 4
```

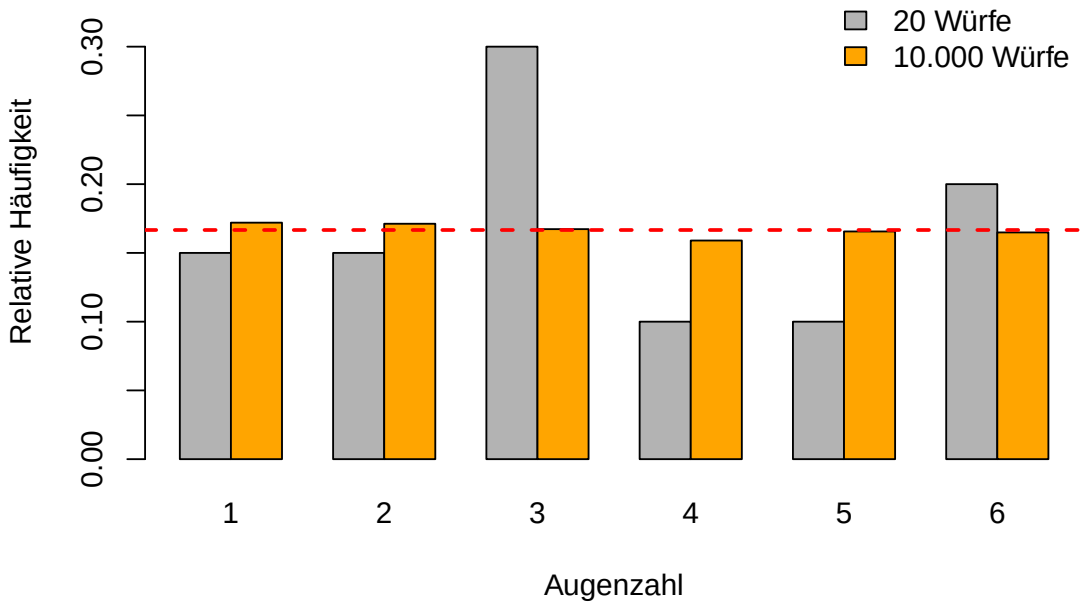
```
rel_20 <- prop.table(haeufig_20)
rel_20
```

```
##
##    1    2    3    4    5    6
## 0.15 0.15 0.30 0.10 0.10 0.20
```

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.

```r
set.seed(123)
wuerfe_10000 <- sample(1:6, size = 10000, replace = TRUE)
rel_10000 <- prop.table(table(factor(wuerfe_10000, levels = 1:6)))
rel_vergleich <- rbind("20 Würfe" = rel_20, "10.000 Würfe" = rel_10000)

barplot(
  rel_vergleich,
  beside = TRUE,
  col = c("grey70", "orange"),
  ylim = c(0, max(rel_vergleich, 1 / 6) * 1.15),
  xlab = "Augenzahl",
  ylab = "Relative Häufigkeit",
  legend.text = rownames(rel_vergleich),
  args.legend = list(x = "topright", bty = "n")
)
abline(h = 1 / 6, col = "red", lwd = 2, lty = 2)
```



Die theoretische Wahrscheinlichkeit beträgt für jede Augenzahl 1/6. Die empirische relative Häufigkeit schwankt, liegt bei 10.000 Würfen aber typischerweise näher an 1/6 als bei 20 Würfen.

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

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

```r
c(
  P_X_4 = mean(wuerfe_10000 == 4),
  P_X_bis_4 = mean(wuerfe_10000 <= 4),
  P_X_ueber_4 = mean(wuerfe_10000 > 4)
)
```

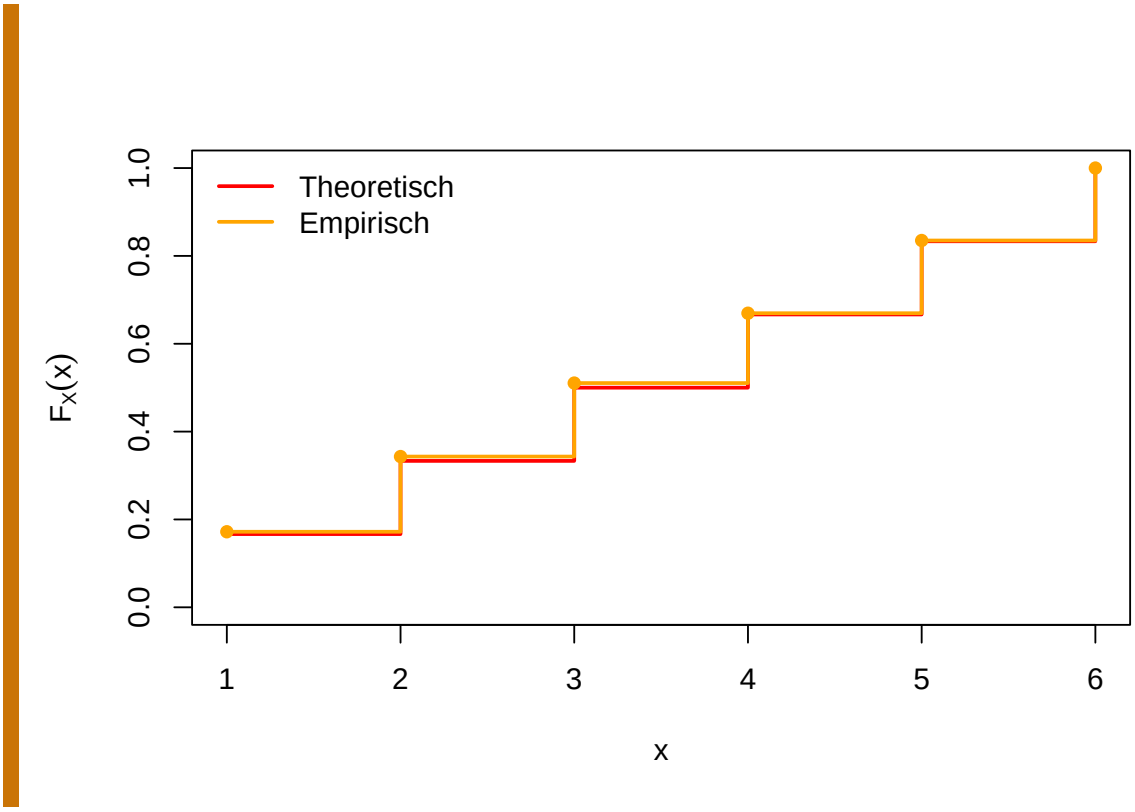
```
##       P_X_4   P_X_bis_4 P_X_ueber_4
##      0.1590      0.6695      0.3305
```

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.

```r
F_emp <- cumsum(rel_10000)
F_theo <- cumsum(rep(1 / 6, 6))

# Empirisch
plot(
  1:6,
  F_theo,
  type = "s",
  lwd = 2,
  col = "red",
  ylim = c(0, 1),
  xlab = "x",
  ylab = expression(F[X](x))
)
# Theoretisch
lines(1:6, F_emp, type = "s", lwd = 2, col = "orange") # theoretischer Graph
points(1:6, F_emp, pch = 16, col = "orange")

legend(
  "topleft",
  legend = c("Theoretisch", "Empirisch"),
  col = c("red", "orange"),
  lwd = 2,
  bty = "n"
)
```


Theoretisch
Empirisch
$F_X(x)$
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
x

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.

```r
# Datensatz ist im selben Ordner wie das Skript gespeichert
students <- read.csv("students_performance.csv")
head(students)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

```r
names(students)
```

```
## [1] "gender"                       "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"     "math.score"
## [7] "reading.score"               "writing.score"
```

```r
summary(students)
```

```
##     gender          race.ethnicity     parental.level.of.education
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     lunch           test.preparation.course   math.score     reading.score
##  Length:1000        Length:1000             Min.   :  0.00   Min.   : 17.00
##  Class :character   Class :character        1st Qu.: 57.00   1st Qu.: 59.00
##  Mode  :character   Mode  :character        Median : 66.00   Median : 70.00
##                                             Mean   : 66.09   Mean   : 69.17
##                                             3rd Qu.: 77.00   3rd Qu.: 79.00
##                                             Max.   :100.00   Max.   :100.00
```

```
##  writing.score
##  Min.   : 10.00
##  1st Qu.: 57.75
##  Median : 69.00
##  Mean   : 68.05
##  3rd Qu.: 79.00
##  Max.   :100.00
```

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

```
students$pass_math <- ifelse(students$math.score >= 60, 1, 0)
table(students$pass_math)
```

```
##
##   0   1
## 323 677
```

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

```
tab_abs <- table(pass_math = students$pass_math, Geschlecht = students$gender)
tab_abs
```

```
##          Geschlecht
## pass_math female male
##         0    197  126
##         1    321  356
```

```
tab_rel <- prop.table(tab_abs)
tab_rel
```

```
##          Geschlecht
## pass_math female  male
##         0  0.197 0.126
##         1  0.321 0.356
```

d. Berechnen Sie aus dem Datensatz die folgenden Wahrscheinlichkeiten:

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

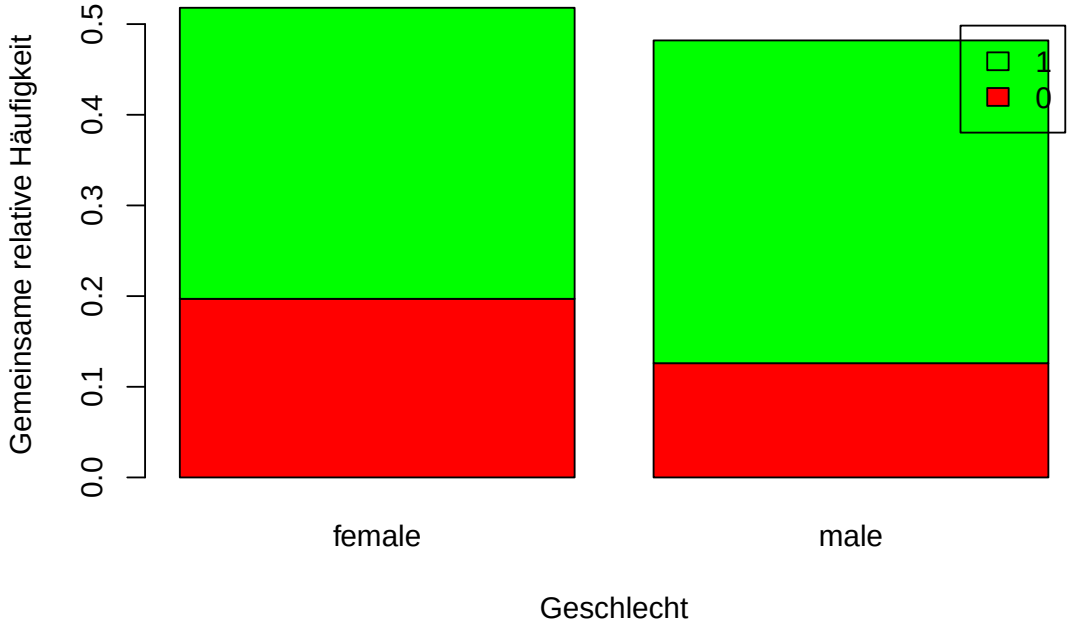
```
c(
  P_bestanden = sum(tab_rel["1", ]),
  P_weiblich = sum(tab_rel[, "female"]),
  P_bestanden_und_weiblich = tab_rel["1", "female"],
  P_nichtbestanden_und_maennlich = tab_rel["0", "male"]
)
```

```
##                    P_bestanden                      P_weiblich
##                          0.677                           0.518
```

```
##       P_bestanden_und_weiblich P_nichtbestanden_und_maennlich
##                          0.321                          0.126
```

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

```r
barplot(
  tab_rel,
  beside = FALSE,
  col = c("red", "green"),
  legend.text = rownames(tab_rel),
  xlab = "Geschlecht",
  ylab = "Gemeinsame relative Häufigkeit"
)
```



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:

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

```
p_bestanden_gegeben_weiblich <- mean(students$pass_math[students$gender == "female"]
== 1)
p_bestanden_gegeben_weiblich
```

```
## [1] 0.6196911
```

```
p_weiblich_gegeben_bestanden <- mean(students$gender[students$pass_math == 1]
== "female")
p_weiblich_gegeben_bestanden
```

```
## [1] 0.4741507
```

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

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

```
P_bestanden_gegeben_weiblich <- tab_rel["1", "female"] /
  sum(tab_rel[, "female"])
P_bestanden_gegeben_weiblich
```

```
## [1] 0.6196911
```

```
P_weiblich_gegeben_bestanden <- tab_rel["1", "female"] /
  sum(tab_rel["1", ])
P_weiblich_gegeben_bestanden
```

```
## [1] 0.4741507
```

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

Die Bedingungen legen unterschiedliche Grundgesamtheiten fest: einmal alle weiblichen Schülerinnen, einmal alle Schüler:innen, die bestanden haben.

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.

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

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

```r
c(
  P_bestanden = sum(tab_rel["1", ]),
  P_bestanden_gegeben_weiblich = p_bestanden_gegeben_weiblich,

  P_gemeinsam = tab_rel["1", "female"],
  Produkt_der_Raender = sum(tab_rel["1", ]) * sum(tab_rel[, "female"])
)
```

```
##                  P_bestanden P_bestanden_gegeben_weiblich
##                    0.6770000                   0.6196911
##                  P_gemeinsam          Produkt_der_Raender
##                    0.3210000                   0.3506860
```

Die beobachteten Werte weichen voneinander ab. Streng genommen sind die Variablen in dieser Population daher nicht unabhängig.

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.

Weitere Merkmale wie Vorbildung, Kurswahl oder soziale Faktoren können sowohl mit Geschlecht als auch mit dem Ergebnis zusammenhängen. Diese Themen werden hier im Kurs vertieft und sind zentral.

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?

100 Personen sind erkrankt, 9.900 Personen sind gesund.

b. Vervollständigen Sie die folgende Vierfeldertafel mit absoluten Häufigkeiten:

	Test positiv	Test negativ	Gesamt
erkrankt	90	10	100
gesund	495	9.405	9.900
Gesamt	**585**	**9.415**	**10.000**

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

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

$$\mathbb{P}(\text{gesund} \mid \text{Test negativ}) = \frac{9.405}{9.415} \approx 99{,}9\%$$

Unter den Personen mit negativem Test sind etwa 99,9 Prozent tatsächlich gesund.

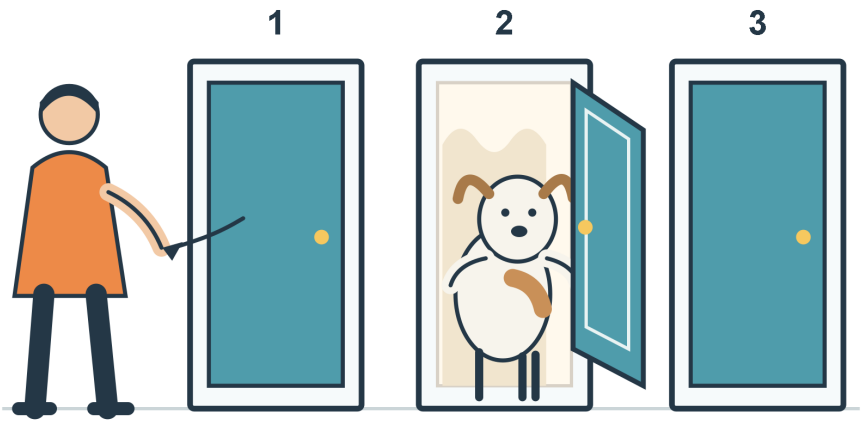
d. Stellen Sie sich nun vor, dass keine Vierfeldertafel für eine konkrete Population von 10.000 Personen vorliegt. Aus getrennten Erhebungen sind jedoch die zu Beginn genannten Raten bekannt. Da die gemeinsame Wahrscheinlichkeit $\mathbb{P}(\textit{gesund} \cap \text{Test negativ})$ nicht direkt beobachtet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

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

$$\mathbb{P}(\text{gesund} \mid \text{Test negativ}) = \frac{\left(\frac{9.405}{9.900}\right)\left(\frac{9.900}{10.000}\right)}{\frac{9.415}{10.000}} = \frac{9.405}{9.415} \approx 99{,}9\%$$

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.

Zunächst könnte man vermuten: zwei Türen bleiben geschlossen, also beträgt die Gewinnwahrscheinlichkeit jeweils 1/2, und Wechseln wäre egal.

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

Betrachten wir die drei gleich wahrscheinlichen Möglichkeiten für die erste Wahl:

Erste Wahl	Was passiert beim Bleiben?	Was passiert beim Wechseln?
Auto gewählt	Gewinn	Verlust
Ziege gewählt	Verlust	Gewinn
Ziege gewählt	Verlust	Gewinn

Nur in einem von drei Fällen wurde das Auto gleich zu Beginn gewählt. Daher gilt

$$\mathbb{P}(\text{Gewinn} \mid \text{Bleiben}) = \frac{1}{3} \quad \text{und} \quad \mathbb{P}(\text{Gewinn} \mid \text{Wechseln}) = \frac{2}{3}$$

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.

```r
monty_hall <- function(wechseln = TRUE) {
  auto <- sample(1:3, 1) # Autotür
  erste_wahl <- sample(1:3, 1) # erste Wahl

  # Mögliche Ziegentüren
  moegliche_tueren <- setdiff(1:3, c(erste_wahl, auto))
  geoeffnete_tuer <- moegliche_tueren[sample.int(length(moegliche_tueren), 1)]

  # Bei Wechsel: andere geschlossene Tür wählen
  endgueltige_wahl <- if (wechseln) {
    setdiff(1:3, c(erste_wahl, geoeffnete_tuer))
  } else {
    erste_wahl # bei erster Wahl bleiben
  }

  endgueltige_wahl == auto # Gewinnüberprüfung
}
monty_hall()
```

```
## [1] TRUE
```

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

```r
set.seed(123)

c(
  Bleiben = mean(replicate(10000, monty_hall(wechseln = FALSE))),
  Wechseln = mean(replicate(10000, monty_hall(wechseln = TRUE)))
)
```

```
##  Bleiben Wechseln
##   0.3306   0.6643
```

Die relativen Gewinnhäufigkeiten liegen (bis auf zufällige Abweichungen) bei 1/3 für Bleiben und 2/3 für Wechseln. Die Simulation bestätigt damit die Rechnung aus Teil b: Bei vielen Wiederholungen gewinnt die Wechselstrategie ungefähr doppelt so häufig.

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.

Der Moderator öffnet nun zufällig eine Tür und kann dabei das Auto zeigen. Ist die erste Wahl richtig, öffnet er sicher eine Ziege. Ist sie falsch, geschieht dies nur mit Wahrscheinlichkeit 1/2.

$$
\begin{aligned}
\mathbb{P}(\text{richtig} \mid \text{Ziege}) &= \frac{\mathbb{P}(\text{richtig}, \text{Ziege})}{\mathbb{P}(\text{Ziege})} \\
&= \frac{\mathbb{P}(\text{Ziege} \mid \text{richtig})\mathbb{P}(\text{richtig})}{\mathbb{P}(\text{Ziege} \mid \text{richtig})\mathbb{P}(\text{richtig}) + \mathbb{P}(\text{Ziege} \mid \text{falsch})\mathbb{P}(\text{falsch})} \\
&= \frac{1 \cdot (1/3)}{1 \cdot (1/3) + (1/2) \cdot (2/3)} \\
&= \frac{1/3}{2/3} \\
&= \frac{1}{2},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(\text{falsch} \mid \text{Ziege}) &= \frac{\mathbb{P}(\text{falsch}, \text{Ziege})}{\mathbb{P}(\text{Ziege})} \\
&= \frac{\mathbb{P}(\text{Ziege} \mid \text{falsch})\mathbb{P}(\text{falsch})}{\mathbb{P}(\text{Ziege} \mid \text{richtig})\mathbb{P}(\text{richtig}) + \mathbb{P}(\text{Ziege} \mid \text{falsch})\mathbb{P}(\text{falsch})} \\
&= \frac{(1/2) \cdot (2/3)}{1 \cdot (1/3) + (1/2) \cdot (2/3)} \\
&= \frac{1/3}{2/3} \\
&= \frac{1}{2}.
\end{aligned}
$$

Beim Bleiben gewinnen Sie, wenn die erste Wahl richtig war; beim Wechseln, wenn sie falsch war. Bedingt auf eine geöffnete Ziege sind daher beide Strategien gleich gut.