Lösungen zu den Aufgaben

1. **Aufgabe**

   Im Hinblick auf die lineare Regression: Welche der folgenden Aussage passt am besten?

   a. Die einfache Regression - $y = \alpha + \beta_1 x_1 + \epsilon$ - prüft, inwieweit zwei Variablen zusammenhängen (linear oder anderweitig).
   b. Obwohl statistische Zusammenhänge nicht ohne Weiteres Kausalschlüsse erlauben, kann man die Regression für Vorhersagen gut nutzen.
   c. Regressionskoeffizienten kann man so interpretieren: “Erhöht man X um eine 1 Einheit, so steigt daraufhin Y um β_1 Einheiten” (β_1 sei der entsprechende Regressionskoeffizient).
   d. “Lineare Regression” bedeutet, dass z.B. keine Polynome wie $y = \alpha + \beta_1 x_1^2 + \beta_2 x_1 + \epsilon$ berechnet werden dürfen, bzw. nicht zur *linearen* Regression zählen.
   e. Zentrieren der Prädiktoren ist bei der linearen Regression nicht zulässig.

   Lösung

   a. Falsch. Die lineare Regression $y = \alpha + \beta_1 x_1 + \epsilon$ untersucht, wie die Korrelation, den Grad des linearen Zusammenhangs. Allerdings sind auch nicht-lineare Zusammenhänge von y und den Prädiktoren erlaubt, etwa $y = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \epsilon$. *Linear* ist dabei so zu verstehen, dass y eine additive Funktion der Prädiktoren ist. Vielleicht wäre es daher besser, anstelle von “linearen” Modellen von “additiven” Modellen zu sprechen.
   b. Richtig. Für Vorhersagen ist Kenntnis einer Kausalstruktur nicht unbedingt nötig, kann aber sehr hilfreich sein.
   c. Falsch. Diese Interpretation suggeriert einen Kausaleffekt. Besser ist die Interpretation “Vergleicht man zwei Beobachtungen, die sich um 1 Einheit in X unterscheiden, so findet man im Durchschnitt einen Unterschied von β_1 in Y”.
   d. Falsch.Die Gleichung $y = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \epsilon$ ist linear in ihren Summanden.
   e. Falsch. Zentrieren der Prädiktoren ist bei der linearen Regression zulässig und oft sinnvoll.

2. **Aufgabe**

   Die folgende Frage bezieht sich auf dieses Ergebnis einer Regressionsanalyse:

```
Call:
lm(formula = y ~ x, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-1.667 -0.464  0.077  0.512  1.726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.207      0.114   -1.81    0.076 .
x             -0.693      0.108   -6.40  4.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.85 on 53 degrees of freedom
```

```
Multiple R-squared:  0.436, Adjusted R-squared:  0.425
F-statistic:   41 on 1 and 53 DF,  p-value: 4.13e-08
```

Welche der folgenden Aussagen passt am besten?

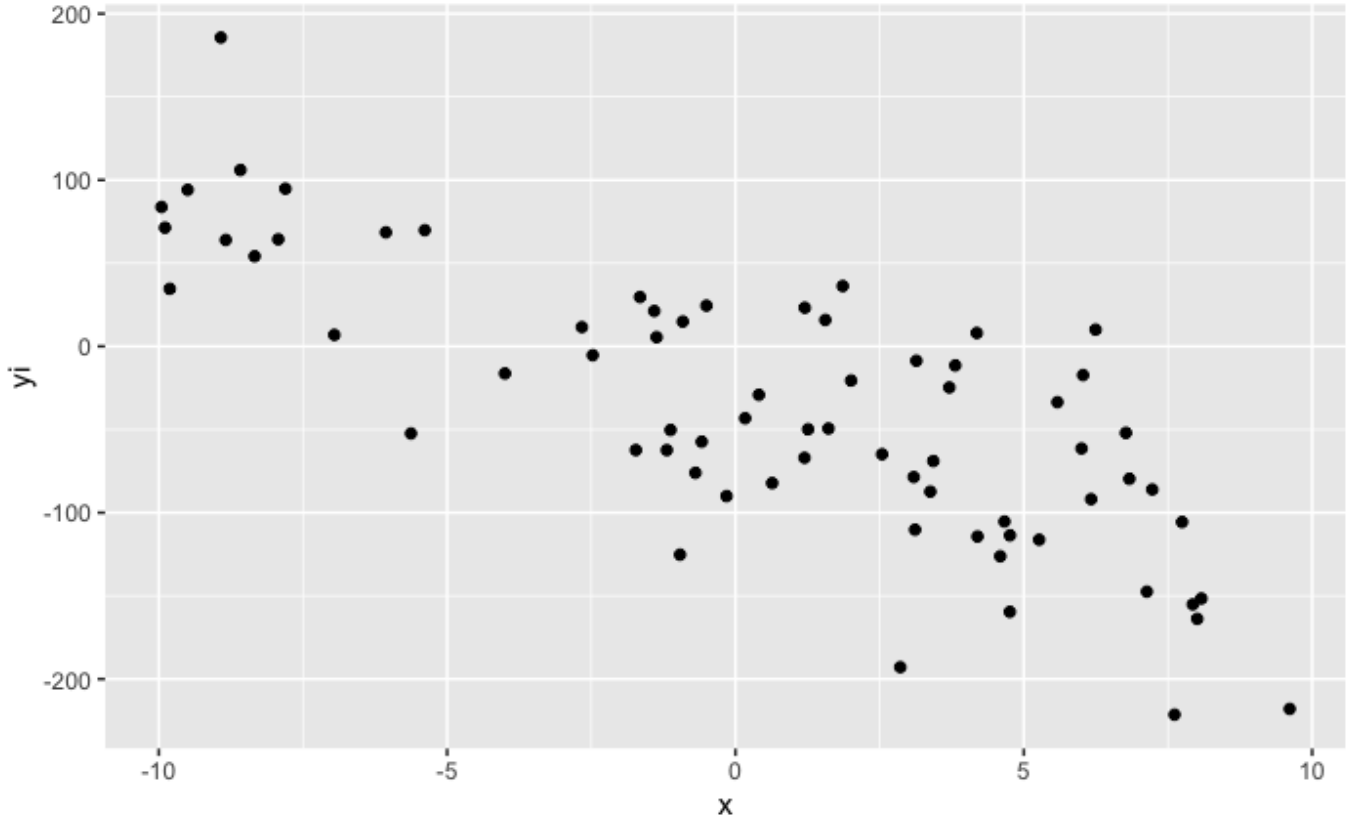
a. Wenn `x` um 1 Einheit steigt, dann kann eine Veränderung um etwa -0.69 Einheiten in `y` erwartet werden (nicht kausal zu verstehen).
b. Der Mittelwert der abhängigen Variaben `y` steigt mit zunehmenden `x`.
c. Wenn `x=0`, dann ist ein Mittelwert von `y` in Höhe von etwa -0.9 zu erwarten.
d. Wenn `x=1`, dann ist ein Mittelwert von `y` in Höhe von ca. -0.21 zu erwarten.
e. Wenn `x=2`, dann ist ein Mittelwert von `y` in Höhe von ca. -0.9 zu erwarten.

Lösung

a. Wahr
b. Falsch
c. Falsch
d. Falsch
e. Falsch Für diese Aufgabe ist es hilfreich, wenn Sie wissen, wie man $\widehat{y}$ berechnet: $\widehat{y} = \alpha + \beta x$. In Worten “Das vorhergesagte Y ist die Summe von Achsenabschnitt (alpha) und Steigung (beta) mal x”. Ein einfaches Rechenbeispiel: Wenn man nichts für die Klausur lernt, hat man 7 Punkte (Achsenabschnitt). Pro Stunde lernen kommt ein halber Klausurpunkte dazu. Wie viele Punkte hat man nach diesem Modell, wenn man 20 Stunden lernt? Antwort: $\widehat{y} = 7 + 0.5 * 20 = 7 + 10 = 17$

3. **Aufgabe**

Ein Streudiagramm von x und y ergibt folgende Abbildung:

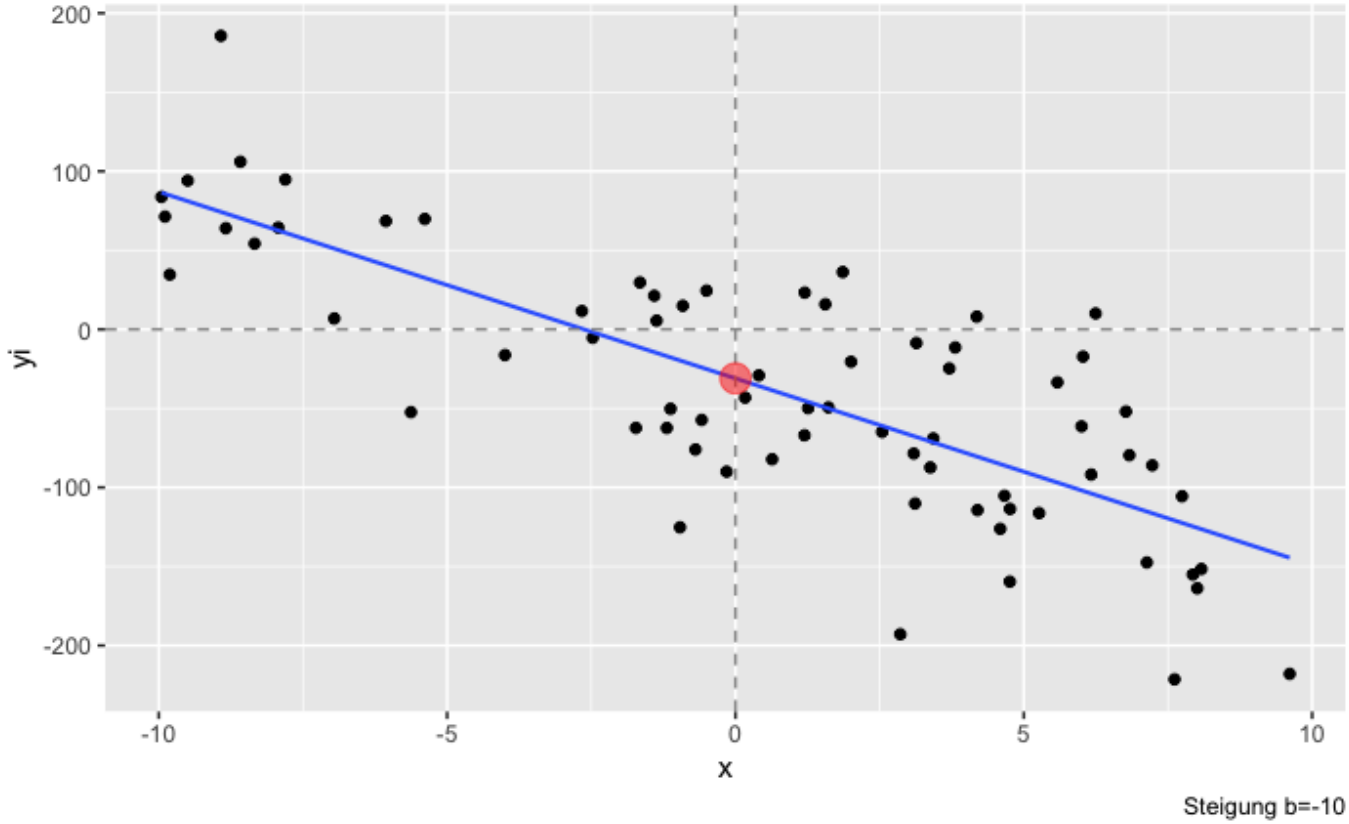


Wählen Sie das am besten passende Modell aus der Liste aus!

a. $y = 40 + -10 \cdot x + \epsilon$
b. $y = 40 + 10 \cdot x + \epsilon$
c. $y = -40 + -10 \cdot x + \epsilon$
d. $y = -40 + 10 \cdot x + \epsilon$
e. $y = 0 + -40 \cdot x + \epsilon$

Lösung

Das dargestellte Modell lautet $y = -40 + -10 \cdot x + \epsilon$.



a. Falsch
b. Falsch
c. Richtig
d. Falsch
e. Falsch

4. **Aufgabe**

Welcher R-Code passt am besten, um folgende Frage aus der Post-Verteilung herauszulesen:

- *Wie wahrscheinlich ist es, dass die mittlere Größe bei mind. 155 cm liegt?*

Hinweise:

- `a` ist der Achsenabschnitt, `b` ist das Regressionsgewicht.
- `post_tab_df` ist eine Tabelle (in Form eines R-Dataframe), die die Stichproben aus der Post-Verteilung enthält.
- Es handelt sich um Regressionsmodell, das mit der Bayes-Methode berechnet wurde.
- Der bzw. die Prädiktoren sind zentriert.

Code A

```
post_tab_df %>%
  count(gross = a == 155) %>%
  mutate(prop = n / sum(n))
```

Code B

```
post_tab_df %>%
```

```
  count(gross = a > 155) %>%
  mutate(prop = n / sum(n))
```

Code C

```
post_tab_df %>%
  count(gross = a <= 155) %>%
  mutate(prop = n / sum(n))
```

Code D

```
post_tab_df %>%
  count(gross = a >= 155) %>%
  mutate(prop = n / sum(n))
```

Code E

```
post_tab_df %>%
  count(gross = a < 155) %>%
  mutate(prop = n / sum(n))
```

a. Code A
b. Code B
c. Code C
d. Code D
e. Code E

Lösung

Vgl. Skript 5.

a. Falsch
b. Falsch
c. Falsch
d. Wahr
e. Falsch

5. **Aufgabe**

Betrachten wir den biologisch fundierten Zusammenhang von Gewicht (UV) und Körpergröße (AV).

Welche der folgenden Priori-Verteilungen passt am besten für β?

Gehen Sie von z-standardisierten Variablen aus.

a. $N(0, 1)$
b. $N(0, 100)$
c. $N(1, 0)$
d. $N(0, 0)$
e. $N(-1, 1)$

Lösung

a. Wahr. Plausibler Prior. Bei z-standardisierten Werten sind die Koeffizienten meist kleiner 1. Noch sinnvoller wäre vermutlich, wenn $\mu > 0$ und nicht $\mu = 0$.
b. Falsch. Zu weit.
c. Falsch. Keine Streuung.
d. Falsch. Keine Streuung.
e. Falsch. Negativer Mittelwert ist nicht sehr plausibel. Eine weitere, sinnvolle Überlegung ist, eine Priorverteilung zu wählen, die nur positive Werte zulässt wie die Exponentialverteilung, m it der Begründung, dass dies biologisch fundiert ist. Allerdings lässt `stan_glm()` nur normalverteilte Prior in diesem Fall zu.

6. **Aufgabe**

Ei Forschi wählt für ein Regressionsmodell $\beta \sim \mathcal{N}(0, 500)$ (Priori), wobei die empirischen Variablen z-standardisiert sind. Beziehen Sie Stellung zu diesem Prior.

Lösung

Die Priori-Verteilung ist nicht sinnvoll spezifiziert. Die Streuung der Normalverteilung ist so groß, dass sie fast schon uniform verteilt ist. Dieser Priori-Verteilung nimmt z.B. an, $Pr(|\beta| < 250) < Pr(|\beta| > 250)$, was eine sehr wilde Vorstellung ist. Man könnte sagen: Die Verteilung nimmt an, dass es wahrscheinlicher ist, dass ihr bester Freund 100 Millionen Lichtjahre entfernt lebt, als dass er näher als diese Distanz bei Ihnen lebt.

Weitere Hinweise hier

Zur Verdeutlichung: Wie wahrscheinlich ist $q = 1, 2, ..., 10$ bei einer Normalverteilung zu betrachten?

Für $q = 1$ beträgt die Wahrscheinlichkeit für einen Wert nicht höher als $q = 1$ etwa 84%:

```
pnorm(q = 1)
```

```
## [1] 0.84
```

Allgemeiner:

```
options(digits = 20)  # Mehr Nachkommastellen
pnorm(q = 1:10)
```

```
##  [1] 0.84134474606854292578 0.97724986805182079141 0.99865010196836989653
##  [4] 0.99996832875816688002 0.99999971334842807646 0.99999999901341229958
##  [7] 0.99999999999872013490 0.99999999999999933387 1.00000000000000000000
## [10] 1.00000000000000000000
```

Die Wahrscheinlichkeiten für Sigma-Ereignisse bis zu ±7 finden sich z.B. hier.

```
options(digits = 2)
```

Vertiefung:

Nassim Taleb hat dieses Argument in seinem Buch “Statistical Consequences of Fat Tails” aufgegriffen (ein anspruchsvolles Buch). Hier finden Sie eine interessante Darstellung eines Arguments daraus.

7. **Aufgabe**

Beziehen Sie sich auf das Regressionsmodell, für das die Ausgabe mit `stan_glm()` hier dargestellt ist:

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      height ~ weight_c
##  observations: 346
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 154.6    0.3
## weight_c      0.9    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 5.1    0.2
```

Betrachten Sie wieder folgende Beziehung (Gleichung bzw. Ungleichung):

$$Pr\left(\text{height}_i = 155 | \text{weight\_c}_i = 0, \alpha, \beta, \sigma\right) \quad \square \quad Pr\left(\text{height}_i = 156 | \text{weight\_c}_i = 0, \alpha, \beta, \sigma\right)$$

Die in der obigen Beziehung angebenen Parameter beziehen sich auf das oben dargestellt Modell.

Ergänzen Sie das korrekte Zeichen in das Rechteck $\square$!

a. $<$
b. $\leq$
c. $>$
d. $\geq$
e. $=$

Lösung

Als Prädiktorwert wurde der Achsenabschnitt spezifiziert, also $x = 0$. Der Achsenabschnitt wird mit 154.6 angegeben. Je weiter ein y_i von 154.6 entfernt ist, desto unwahrscheinlicher ist es, gegeben $x = 0$.

a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch

8. **Aufgabe**

Was ist *nicht* Ziel oder Gegenstand einer Bayes-Analyse?

a. updating beliefs
b. quantifying uncertainty
c. including prior knowledge of the domain, possibly of subjective nature

d. drawing inferential conclusions solely based on the likelihood

Lösung

Bei der Bayes-Analyse werden die Schlussfolgerungen nicht nur auf Basis des Likelihoods gezogen (im Gegensatz zum Frequentistischen Ansatz).

a. Falsch
b. Falsch
c. Falsch
d. Wahr

9. **Aufgabe**

Der Likelihood eines Datensatzes ist definiert als das Produkt der Likelihoods aller Beobachtungen:

$$\mathscr{L} = \prod_{i=1}^{n} \mathscr{L}_i$$

wobei die Beobachtungen bzw. ihre Likelihood als unabhängig angenommen werden:
$\mathscr{L}_i \perp \mathscr{L}_j, \quad i \neq j.$

Je größer n, desto …….. $\mathscr{L}$!

Füllen Sie die Lücke!

a. größer
b. kleiner
c. unabhängig voneinander
d. keine Aussage möglich
e. kommt auf weitere, hier nicht benannte Bedingungen an

Lösung

Multipliziert man zwei (oder mehr) Anteile p_i (Wahrscheinlichkeiten), $p \in [0, 1]$, so ist das resultierende Produkt nicht größer als p_i. Je mehr Anteile p_i man multipliziert, desto kleiner (näher an Null, aber positiv) das resultierende Produkt.

Beispiel: Die Wahrscheinlichkeit, dass eine zufällig bestimmte (“gezogene”) Person eine Frau ist, sei $p = 1/2$. Die Wahrscheinlichkeit, dass unter Personen zwei Frauen sind, beträgt $p_2 = p \cdot p = 1/4$ (unter der Annahme, dass die Ziehungen unabhängig sind). Wir sehen: Je mehr Wahrscheinlichkeiten (“Anteile”) man multipliziert, desto kleiner (näher an Null) das resultierende Produkt.

a. Falsch
b. Richtig
c. Falsch
d. Falsch

e. Falsch

10. **Aufgabe**

Welche Zeile der folgenden Modellspezifikation zeigt den Likelihood?

$$
\begin{aligned}
\text{height}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta \cdot \text{weight}_i \\
\alpha &\sim \text{Normal}(178, 20) \\
\beta &\sim \text{Normal}(5, 3) \\
\sigma &\sim \text{Exp}(0.1)
\end{aligned}
$$

Zeile …

a. 1
b. 2
c. 3
d. 4
e. 5

Lösung

a. Richtig
b. Falsch. Lineares Modell.
c. Falsch. Prior Achsenabschnitt.
d. Falsch. Prior Regressiongewicht.
e. Falsch. Prior Streuung der AV.

11. **Aufgabe**

Sie möchten, im Rahmen einer Studie, ein einfaches lineare Modell spezifizieren, d.h. den Likelihood und die Priori-Verteilungen benennen.

Folgende Informationen sind gegeben:

- AV: `einnahmen`
- UV: `werbebudget`
- Alle empirischen Variablen sind z-standardisiert.
- Alle Variablen sollen als normalverteilt angegeben werden mit Ausnahme der Streuung der AV, diese ist exponenzialverteilt mit Rate 1 zu modellieren.
- Streuungen der Normalverteilung sind mit 2.5 SD anzugeben.

Schreiben Sie in mathematischer Notation folgende Notation auf:

Die Priori-Verteilung des Regressionsgewichts

Hinweise:

- Verzichten Sie auf Leerstellen in Ihrer Antwort.

- Benennen Sie β mit `b`, α mit `a` und σ mit `s`.
- Nutzen Sie die Tilde `~` um stochastische Relationen (Verteilungen) anzuzeigen.
- Geben Sie Normalverteilungen als `Normal(x;y)` und Exponentialverteilung als `Exp(x)` an (jeweils mit den korrekten Argumenten in der allgemein üblichen Form).

Lösung

b~Normal(0, 2.5)

12. **Aufgabe**

Sie möchten, im Rahmen einer Studie, ein einfaches lineare Modell spezifizieren, d.h. den Likelihood und die Priori-Verteilungen benennen.

Folgende Informationen sind gegeben:

- AV: `einnahmen`
- UV: `werbebudget`
- Alle empirischen Variablen sind z-standardisiert.
- Alle Variablen sollen als normalverteilt angegeben werden mit Ausnahme der Streuung der AV, diese ist exponenzialverteilt mit Rate 1 zu modellieren.
- Streuungen der Normalverteilung sind mit 2.5 SD anzugeben.

Schreiben Sie in mathematischer Notation folgende Notation auf:

Priori-Verteilung der Streuung der AV

Hinweise:

- Verzichten Sie auf Leerstellen in Ihrer Antwort.
- Benennen Sie β mit `b`, α mit `a` und σ mit `s`.
- Nutzen Sie die Tilde `~` um stochastische Relationen (Verteilungen) anzuzeigen.
- Geben Sie Normalverteilungen als `Normal(x;y)` und Exponentialverteilung als `Exp(x)` an (jeweils mit den korrekten Argumenten in der allgemein üblichen Form).

Lösung

s~Exp(1)

13. **Aufgabe**

Nach der Berechnung bzw. Schätzung der Modellparameter ein)es Regressionsmodells (mit Methoden der Bayes-Inferenz) erhält man u.a. auf die Prädiktorwerte x_i ($i = 1, 2, ..., n$) bedingte Wahrscheinlichkeiten für die AV, y_i, oder genauer $y_i|x_i, \theta$ (mit θ für die Modellparameter).

Betrachten Sie dazu folgende Aussage:

$Pr(y_i|x_i, \alpha, \beta, \sigma) = c$ für $i = 1, 2, ..., n$

Welche der Aussagen ist in diesem Zusammenhang *falsch*?

a. Das Regresssionsmodell hat 3 Parameter.
b. Das Regresssionsmodell hat 1 Prädiktor (im Sinne von 1 Inputvariablen).
c. $Pr(y_i|x_i, \alpha, \beta, \sigma) = c$ für $i = 1, 2, ..., n$
d. $\sum_{y_i=-\infty}^{+\infty} Pr(y_i|x_i, \alpha, \beta, \sigma) = 1$
e. $Pr(y_i|x_i, \alpha, \beta, \sigma) = p_i, \qquad p_i \in [0, 1]$

Lösung

a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch