Lösungen zu den Aufgaben

1. **Aufgabe**

   Welchen z-Wert hat eine Beobachtung mit

   $$x = 35$$

   bei einer Normalverteilung mit $\mu = 50$ und $\sigma = 5$?

   a. 0
   b. -2
   c. -3
   d. 3

   Lösung

   a. False
   b. False
   c. True
   d. False

2. **Aufgabe**

   Geben Sie den (am besten passenden) Wert der Verteilungsfunktion F für das im Folgenden dargestellten Quantil q an.

   Gehen Sie von dieser Verteilung aus: $\mathcal{N}\,(\mu = 0, \sigma = 1)$.

   Das Quantil lautet: -2.

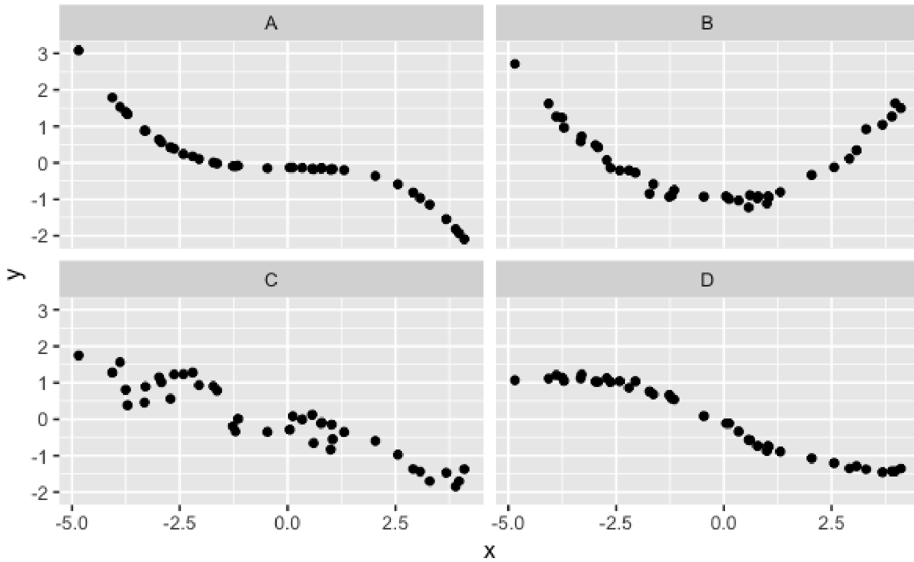
   a. 0.16
   b. 0.84
   c. 0.02
   d. 0.50
   e. 0.98

   Lösung

   Der Wert der Verteilungsfunktion von $z = -2$ ist 0.02.

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

3. **Aufgabe**



Bei welcher der Abbildungen ist eine Regression mit linearem Graph angemessen?

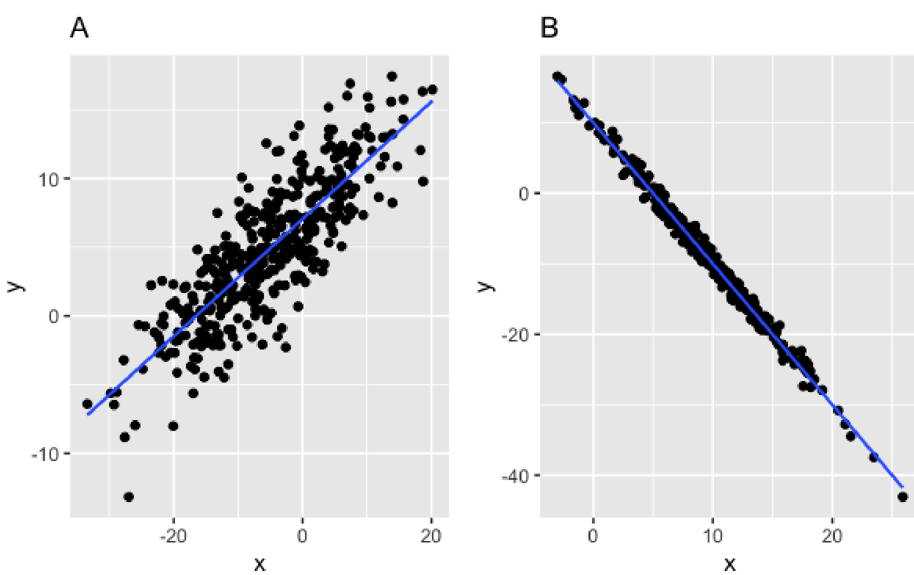
a. A
b. B
c. C
d. D

Lösung

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

4. **Aufgabe**

Die beiden folgenden Abbildungen zeigen zwei lineare Regressionen.



Welche Aussage stimmt?

a. $R_A^2 < R_B^2$

b. $R_A^2 \approx R_B^2$
c. $R_A^2 > R_B^2$

Lösung

Je enger die Punkte um die Gerade streuen, desto größer ist R^2.

a. Richtig
b. Falsch
c. Falsch

5. **Aufgabe**

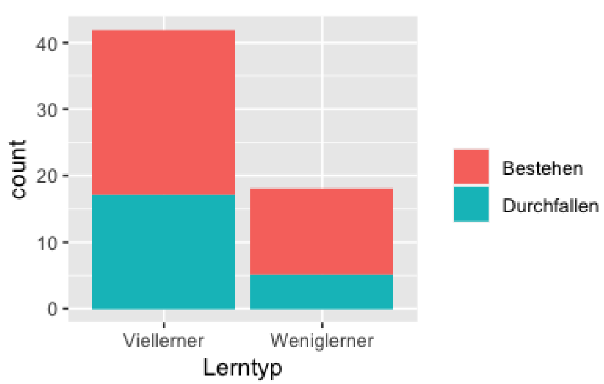
Prof. Salzig untersucht eine seiner Lieblingsfragen: Wie viel bringt das Lernen auf eine Klausur? Dabei konzentriert er sich auf das Fach Statistik (es gefällt ihm gut). In einer aktuellen Untersuchung hat er $n = 60$ Studierende untersucht (s. Tabelle und Diagramm) und jeweils erfasst, ob die Person die Klausur bestanden (b) hat oder durchgefallen (d) ist. Dabei hat er zwei Gruppen unterschieden: Die "Viel-Lerner" (VL) und die "Wenig-Lerner" (WL).

Berechnen Sie die folgende *bedingte Wahrscheinlichkeit*: p(Bestehen|Viellerner).

Beispiel: Wenn Sie ausrechnen, dass die Wahrscheinlichkeit bei 42 Prozentpunkten liegt, so geben Sie ein: `0,42` bzw. `0.42` (das Dezimalzeichen ist abhängig von Ihren Spracheinstellungen).

Hinweise:

- Geben Sie *nur eine Zahl* ein (ohne Prozentzeichen o.Ä.), z.B. `0,42`.
- Andere Angaben können u.U. nicht gewertet werden.
- Runden Sie auf zwei Dezimalstellen.
- Achten Sie darauf, das *korrekte Dezimaltrennzeichen* einzugeben; auf Geräten mit deutscher Spracheinstellung ist dies oft ein Komma.



Ergebnisse der Studie

	Viellerner	**Weniglerner**
Bestehen	25	13
Durchfallen	17	5

Lösung

Der gesuchte Wert liegt bei 0.6.

Lerntyp	Klausurergebnis	n	n_group	prop_conditional_group	N_gesamt
Viellerner	Bestehen	25	42	0.6	60

6. **Aufgabe**

Als Bildungsforscher(in) untersuchen Sie den Lernerfolg in einem Statistikkurs.

Eine Gruppe von Studierenden absolviert einen Statistikkurs. Ein Teil lernt gut mit (Ereignis A), ein Teil nicht (Ereignis A^C). Ein Teil besteht die Prüfung (Ereignis B); ein Teil nicht (B^C).

Hinweis: Das Gegenereignis zum Ereignis A wird oft das Komplementärereignis oder kurz Komplement von A genannt und mit A^C bezeichnet.

Wir ziehen zufällig eine/n Studierende/n: Siehe da – Die Person hat bestanden. Yeah!

Aufgabe: Gesucht ist die Wahrscheinlichkeit, dass *diese Person* gut mitgelernt hat, gegeben der Tatsache, dass dieser Person bestanden hat.

Die Anteile der Gruppen (bzw. Wahrscheinlichkeit des Ereignisses) lassen sich unten stehender Tabelle entnehmen.

row_ids	B	Bneg
A	0.59	0.10
Aneg	0.23	0.07

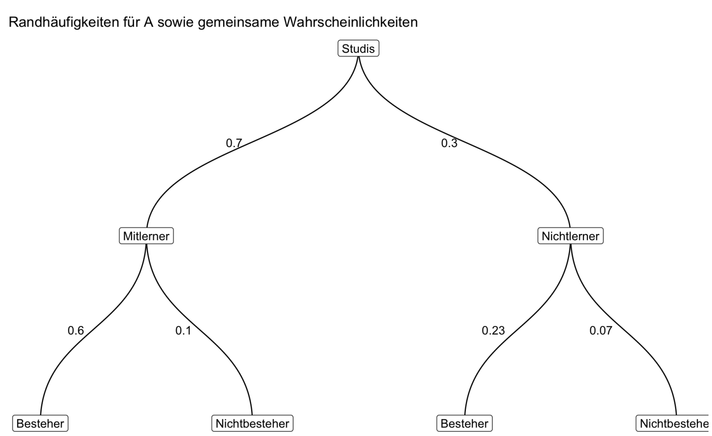
Hinweise:

- Runden Sie auf 2 Dezimalstellen.
- Geben Sie Anteile stets in der Form `0.42` an (mit führender Null und Dezimalzeichen).
- “Aneg” bezieht sich auf das Komplementärereignis zu A.

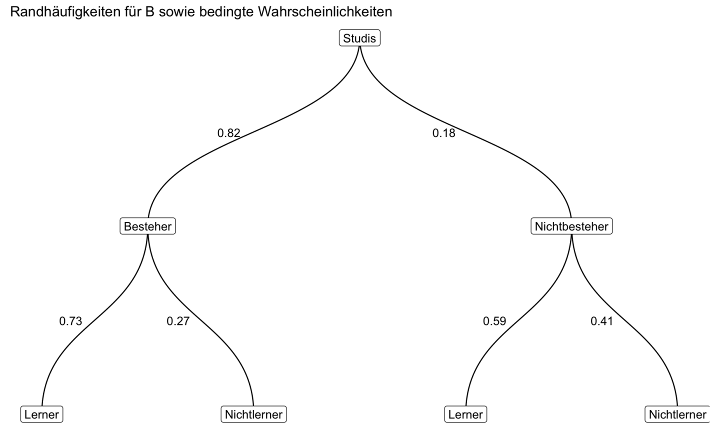
a. Zeichnen Sie (per Hand) ein Baumdiagramm, um die gemeinsamen Wahrscheinlichkeiten darzustellen. Weiterhin sollen die Randwahrscheinlichkeiten für A dargestellt sein.
b. Zeichnen Sie (per Hand) ein Baumdiagramm, um diesen Sachverhalt darzustellen.
c. Geben Sie die Wahrscheinlichkeit des gesuchten Ereignisses an.

Lösung

A.



B.



C.

0.73

```
A_cond_B <- AandB / B_marg %>% round(2)
Aneg_cond_B <- AnegandB / B_marg %>% round(2)
A_cond_Bneg <- AandBneg / Bneg_marg %>% round(2)
Aneg_cond_Bneg <- AnegandBneg / Bneg_marg %>% round(2)
```

$Pr(A) = 0.7$.

$Pr(B) = 0.82$.

$Pr(AB) = 0.6$.

$Pr(A|B) = 0.73$.

$Pr(\neg A|B) = 0.27$.

$Pr(A|\neg B) = 0.59$.

$Pr(\neg A|\neg B) = 0.41$.

7. **Aufgabe**

Als Bildungsforscher(in) untersuchen Sie den Lernerfolg in einem Statistikkurs.

Eine Gruppe von Studierenden absolviert einen Statistikkurs. Ein Teil lernt gut mit (Ereignis A), ein Teil nicht (Ereignis A^C). Ein Teil besteht die Prüfung (Ereignis B); ein Teil nicht (B^C).

Wir ziehen zufällig eine/n Studierende/n: Siehe da – Die Person hat bestanden. Yeah!

Die Anteile der Gruppen (bzw. Wahrscheinlichkeit des Ereignisses) lassen sich unten stehender Tabelle entnehmen.

row_ids	B	Bneg	Summe
A	0.78	0.13	0.91
A_neg	0.07	0.02	0.09
Summe	0.86	0.15	1.01

Aufgabe: Gesucht ist die (bedingte) Wahrscheinlichkeit, dass *diese Person* gut mitgelernt hat, gegeben der Tatsache, dass sie bestanden hat. Geben Sie die Wahrscheinlichkeit des gesuchten Ereignisses an!

Hinweise:

- Runden Sie auf 2 Dezimalstellen.
- Geben Sie Anteile stets in der Form `0.42` an (mit führender Null und Dezimalpunkt).
- “A_neg” bezieht sich auf das Komplementärereignis zu A.

Lösung

Der gesuchte Wert lautet: 0.92.

8. **Aufgabe**

Ob wohl die *PS-Zahl* (Ereignis A) und der *Spritverbrauch* (Ereignis B) voneinander abhängig sind? Was meinen Sie? Was ist Ihre Einschätzung dazu? Vermutlich haben Sie ein (wenn vielleicht auch implizites) Vorab-Wissen zu dieser Frage. Lassen wir dieses Vorab-Wissen aber einmal außen vor und schauen uns rein Daten dazu an. Vereinfachen wir die Frage etwas, indem wir fragen, ob die Ereignisse "hoher Spritverbrauch" (A) und "hohe PS-Zahl" voneinander abhängig sind.

Um es konkret zu machen, nutzen wir den Datensatz `mtcars`:

```
library(tidyverse)
data(mtcars)
glimpse(mtcars)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21, 21, 23, 21, 19,…
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8…
## $ disp <dbl> 160, 160, 108, 258,…
## $ hp   <dbl> 110, 110, 93, 110, …
## $ drat <dbl> 3.9, 3.9, 3.9, 3.1,…
## $ wt   <dbl> 2.6, 2.9, 2.3, 3.2,…
## $ qsec <dbl> 16, 17, 19, 19, 17,…
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0…
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0…
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3…
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4…
```

Weitere Infos zum Datensatz bekommen Sie mit `help(mtcars)` in R.

Definieren wir uns das Ereignis "hohe PS-Zahl" (und nennen wir es `hp_high`, klingt cooler). Sagen wir, wenn die PS-Zahl größer ist als der Median, dann trifft `hp_high` zu, ansonsten nicht:

```
mtcars %>%
  summarise(median(hp))
```

median(hp)
123

Mit dieser "Wenn-Dann-Abfrage" können wir die Variable `hp_high` mit den Stufen `TRUE` und `FALSE` definieren:

```
mtcars <-
  mtcars %>%
  mutate(hp_high = case_when(
    hp > 123 ~ TRUE,
    hp <= 123 ~ FALSE
  ))
```

Genauso gehen wir mit dem Spritverbrauch vor (`mpg_high`):

```
mtcars <-
  mtcars %>%
  mutate(mpg_high = case_when(
    mpg > median(mpg) ~ TRUE,
    mpg <= median(mpg) ~ FALSE
  ))
```

Berechnen Sie $Pr(\text{mpg\_high}|\neg\text{hp\_high})$*!*

Lösung

Schauen wir zuerst mal in den Datensatz:

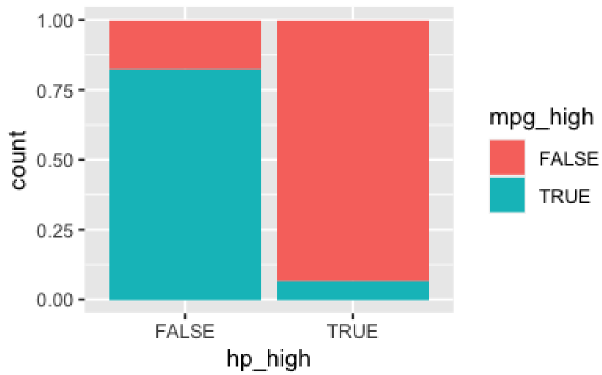
```
mtcars %>%
  select(hp, hp_high, mpg, mpg_high) %>%
  slice_head(n = 5)
```

hp	hp_high	mpg	mpg_high

	hp	hp_high	mpg	mpg_high
Mazda RX4	110	FALSE	21	TRUE
Mazda RX4 Wag	110	FALSE	21	TRUE
Datsun 710	93	FALSE	23	TRUE
Hornet 4 Drive	110	FALSE	21	TRUE
Hornet Sportabout	175	TRUE	19	FALSE

Dann visualisieren wir die bedingnte Wahrscheinlichkeiten:

```
mtcars %>%
  #select(hp_high, mpg_high) %>%
  ggplot() +
  aes(x = hp_high, fill = mpg_high) +
  geom_bar(position = "fill")
```



Hey, sowas von abhängig voneinander, die zwei Variablen, `mpg_high` und `hp_high`!

Der *rechte* Balken zeigt $Pr(\text{mpg\_high}|\ \text{hp\_high})$ und $Pr(\neg\text{mpg\_high}|\text{hp\_high})$. Der *linke* Balken zeigt $Pr(\text{mpg\_high}|\neg\text{hp\_high})$ und $Pr(\neg\text{mpg\_high}|\neg\text{hp\_high})$.

Berechnen wir die relevanten Anteile:

```
mtcars %>%
  #select(hp_high, mpg_high) %>%
  count(mpg_high, hp_high) %>%  # Anzahl pro Zelle der Kontingenztabelle
  group_by(hp_high) %>%  # die Anteile pro "Balken" s. Diagramm
  mutate(prop = n / sum(n))
```

```
## # A tibble: 4 × 4
## # Groups:   hp_high [2]
##   mpg_high hp_high     n   prop
##   <lgl>    <lgl>   <int>  <dbl>
## 1 FALSE    FALSE       3 0.176
## 2 FALSE    TRUE       14 0.933
## 3 TRUE     FALSE      14 0.824
## 4 TRUE     TRUE        1 0.0667
```

Die richtige Antwort lautet: 0.82.

Am besten, Sie führen den letzten Code Schritt für Schritt aus und schauen sich jeweils das Ergebnis an, das hilft beim Verstehen.

Alternativ kann man sich die Häufigkeiten auch schön bequem ausgeben lassen:

```
library(mosaic)
tally(mpg_high ~ hp_high,
      data = mtcars,
      format = "proportion")
```

	TRUE	FALSE
TRUE	0.07	0.82
FALSE	0.93	0.18

9. **Aufgabe**

Betrachten wir das Ereignis “Schwerer Coronaverlauf” (S); ferner betrachten wir das Ereignis “Blutgruppe ist A” (A) und das Gegenereignis von A: “Blutgruppe ist nicht A”. Ein Gegenereignis wird auch als *Komplementärereignis* oder *Komplement* (complement) mit dem Term A^C bezeichnet.

Sei $Pr(S|A) = 0.01$ und sei $Pr\left(S|A^C\right) = 0.01$.

Was kann man auf dieser Basis zur Abhängigkeit der Ereignisse S und A sagen?

Geben Sie ein *Adjektiv* an, dass diesen Sachverhalt kennzeichnet!

Lösung

Die Lösung lautet: unabhängig.

S und A sind unabhängig: Offenbar ist die Wahrscheinlichkeit eines schweren Verlaufs gleich groß unabhängig davon, ob die Blutgruppe A ist oder nicht. In diesem Fall spricht man von *stochastischer Unabhängigkeit*.

$$Pr(S|A) = Pr\left(S|A^C\right) = Pr(S)$$

10. **Aufgabe**

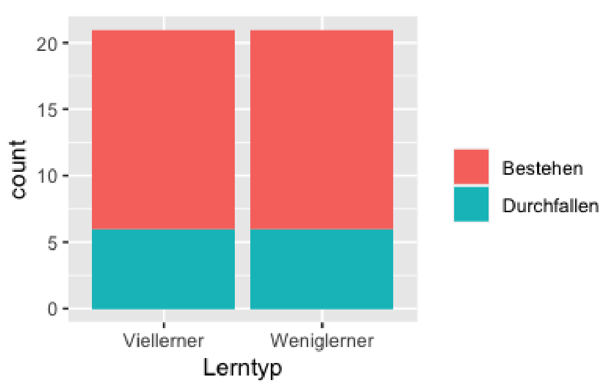
Prof. Bitter untersucht eine seiner Lieblingsfragen: Wie viel bringt das Lernen auf eine Klausur? Dabei konzentriert er sich auf das Fach Statistik (es gefällt ihm gut). In einer aktuellen Untersuchung hat er $n = 42$ Studierende untersucht (s. Tabelle und Diagramm) und jeweils erfasst, ob die Person die Klausur bestanden (b) hat oder durchgefallen (d) ist. Dabei hat er zwei Gruppen unterschieden: Die “Viel-Lerner” (VL) und die “Wenig-Lerner” (WL).

Berechnen Sie die folgende: *gemeinsame Wahrscheinlichkeit*: p(Durchfallen UND Weniglerner).

Beispiel: Wenn Sie ausrechnen, dass die Wahrscheinlichkeit bei 42 Prozentpunkten liegt, so geben Sie ein: `0,42` bzw. `0.42` (das Dezimalzeichen ist abhängig von Ihren Spracheinstellungen).

- Geben Sie *nur eine Zahl* ein (ohne Prozentzeichen o.Ä.), z.B. `0,42`.
- Andere Angaben können u.U. nicht gewertet werden.
- Runden Sie auf zwei Dezimalstellen.
- Achten Sie darauf, das *korrekte Dezimaltrennzeichen* einzugeben; auf Geräten mit deutscher Spracheinstellung ist dies oft ein Komma.

Das folgende Diagramm zeigt die Häufigkeiten pro Gruppe:



Hier ist die Kontingenztabelle mit den Häufigkeiten pro Gruppe:

Lerntyp	Bestehen	Durchfallen
Viellerner	15	6
Weniglerner	15	6

Lösung

Die gemeinsame Wahrscheinlichkeit beträgt 0.14.

Lerntyp	Klausurergebnis	n	n_group	prop_conditional_group	joint_prob
Weniglerner	Durchfallen	6	21	0.29	0.14

Die gemeinsame Wahrscheinlichkeit berechnet sich hier als der Quotient der Zellenhäufigkeit und der Gesamthäufigkeit.

11. **Aufgabe**

Wir betrachten einen Datensatz, der Kredite analysiert:

```
library(openintro)
data("loans_full_schema")
```

Quelle

Hier ist ein Überblick über den Datensatz:

```
## tibble [10,000 × 3] (S3: tbl_df/tbl/data.frame)
##  $ interest_rate   : num [1:10000] 14.07 12.61 17.09 6.72 14.07 ...
##  $ annual_income   : num [1:10000] 90000 40000 40000 30000 35000 34000 35000 110000 65000 30000 ...
##  $ application_type: Factor w/ 2 levels "individual","joint": 1 1 1 1 2 1 2 1 1 1 ...
```

Eine Analystin möchte den Zinssatz (`interest_rate`) auf Basis dieses Datensatzes vorhersagen.

Sie berechnet folgendes Regressionsmodell:

```
lm1 <- lm(interest_rate ~ annual_income + application_type, data = loans_full_schema)
```

Folgende Ergebnisse bekommt Sie zurück geliefert:

term	estimate	std_error
intercept	12.90	0.083
annual_income	0.00	0.000
application_typejoint	0.71	0.140

Welche Aussage ist korrekt?

a. `Estimate` liefert eine Schätzung zur Modellgüte.
b. Das Verhältnis von Signal zu Rauschen für `application_typejoint` ist kleiner als 1.
c. Es liegt ein Fehler vor, denn `application_typejoint` hat neben `joint` noch eine weitere Stufe (`individual`), diese ist aber nicht aufgeführt.
d. Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `individual` bei `application_type` und ohne Jahreseinkommen.
e. Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `joint` bei `application_type` und ohne Jahreseinkommen.

Lösung

Der Wert bei `Intercept` gibt den Wert der abhängigen Variable an, bei Fällen mit dem Wert `individual` bei `application_type` und ohne Jahreseinkommen.

```
predict(lm1, newdata = data.frame(annual_income = 0,
                                  application_type = "individual"))
```

```
##  1
## 13
```

a. Falsch
b. Falsch
c. Falsch

d. Wahr
e. Falsch

12. **Aufgabe**

Welche der folgenden Zeilen zeigt den Likelihood?

a. $\mu \sim \mathcal{N}(0, 10)$
b. $\sigma \sim \mathcal{U}(0, 1)$
c. $y_i = \beta_0 + \beta_1 \cdot x$
d. $y_i \sim \mathcal{N}(\mu, \sigma)$

Lösung

a. Falsch. Priori-Verteilung.
b. Falsch. Priori-Verteilung.
c. Falsch. Regressionsformel.
d. Wahr. Likelihood.

13. **Aufgabe**

Wie viele Parameter hat das folgende Modell?

Likelihood: $h_i \sim \mathcal{N}(\mu, \sigma)$

Prior für μ: $\mu \sim \mathcal{N}(178, 20)$

Prior für σ: $\sigma \sim \mathcal{U}(0, 50)$

a. 0
b. 1
c. 2
d. 3
e. mehr

Lösung

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

14. **Aufgabe**

Eine statistische Analyse, wie eine Regression, ist mit mehreren Arten an Ungewissheit konfrontiert. Zum einen gibt es die *Ungewissheit in den Modellparametern*. Für die Regression bedeutet das: "Liegt die Regressionsgerade in"Wahrheit" (in der Population) genauso wie in der Stichprobe, sind Achsenabschnitt und Steigung in der Stichprobe also identisch zur Popuation?". Zum anderen die *Ungewissheit innerhalb des Modells*. Auch wenn wir die"wahre" Regressionsgleichung kennen würden, wären (in aller Regel) die Vorhersagen trotzdem nicht perfekt. Auch wenn wir etwa wüssten, wieviel Klausurpunkte "in Wahrheit" pro Stunde Lernen herausspringen (und wenn wir den wahren Achsenabschnitt kennen würden), so würde das Modell trotzdem keine perfekten Vorhersagen zum Klausurerfolg liefern. Vermutlich fehlen dem Modell wichtige Informationen etwa zur Motivation der Studentis.

Vor diesem Hintergrund, betrachten Sie folgendes statistisches Regressionsmodell, das mit Methoden der Bayes-Statistik berechnet werden soll:

```
## mpg ~ gear
## <environment: 0x11f7390f0>
```

```
## stan_glm
##  family:       gaussian [identity]
```

```
##  formula:      mpg ~ gear
##  observations: 32
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 5.6    5.1
## gear        3.9    1.4
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 5.4    0.7
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Beantworten Sie vor diesem Hintergrund folgende Frage:

Welche Zahl(en) kennzeichnet die Ungewissheit des Modells des Modells gegeben der Modellparameter (die Ungewissheit innerhalb des Modells)? (Gesucht ist nicht die Ungewissheit für die Ungewissheit des Modells).

Hinweise:

- Geben Sie nur Zahlen ein.
- Runden Sie auf 1 Dezimale.

Lösung

Die Antwort lautet: 5.45.

15. **Aufgabe**

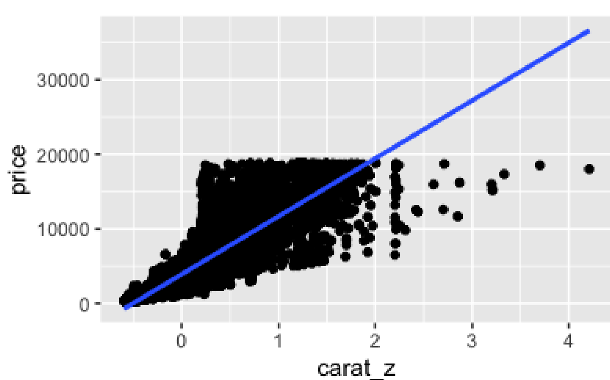
Betrachten Sie folgendes Modell, das den Zusammenhang des Preises (`price`) und dem Gewicht (`carat`) von Diamanten untersucht (Datensatz `diamonds`).

Aber zuerst zentrieren wir den metrischen Prädiktor `carat`, um den Achsenabschnitt besser interpretieren zu können.

Dann berechnen wir ein (bayesianisches) Regressionsmodell, wobei wir auf die Standardwerte der Prior zurückgreifen.

```
Estimates:
              mean   sd     10%    50%    90%
(Intercept) 3932.5    6.8 3923.7 3932.5 3941.1
carat_z     7756.3   14.2 7737.8 7756.2 7774.7
sigma       1548.6    4.8 1542.5 1548.6 1554.7
```

Zur Verdeutlichung ein Diagramm zum Modell:



Aufgabe:

Geben Sie eine Regressionsformel an, die `lm1` ergänzt, so dass die Schliffart (`cut`) des Diamanten kontrolliert (adjustiert) wird. Anders gesagt: Das Modell soll die mittleren Preise für jede der fünf Schliffarten angeben.

Hinweis:

- Geben Sie nur die Regressionsformel an.
- Lassen Sie zwischen Termen der Regressionsformel jeweils ein Leerzeichen Abstand.

- Beziehen Sie sich auf das Modell bzw. die Angaben oben.
- Es gibt (laut Datensatz) folgende Schliffarten (und zwar in der folgenden Reihenfolge):

cut
Ideal
Premium
Good
Very Good
Fair

```
## [1] "Fair"      "Good"
## [3] "Very Good" "Premium"
## [5] "Ideal"
```

Lösung

Die richtige Antwort lautet: `price ~ carat_z + cut`

Das Modell könnten wir so berechnen:

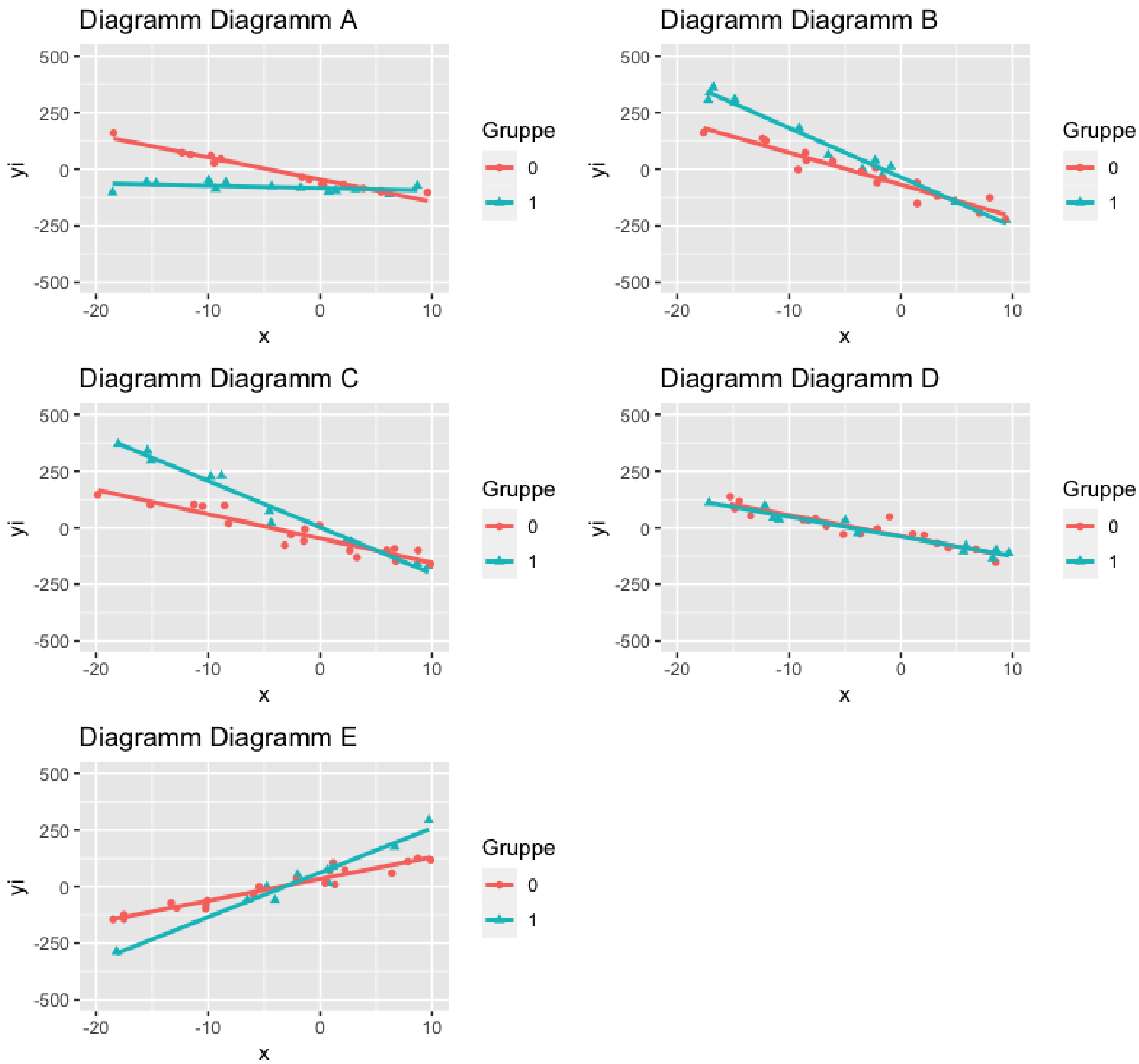
Oder auch so, mit der klassischen Regression:

```
Estimates:
              mean   sd     10%    50%    90%
(Intercept) 3579.4    9.7 3566.9 3579.5 3591.9
carat_z     7871.5   14.2 7853.1 7871.4 7890.3
cut.L       1239.4   26.3 1205.7 1239.6 1272.6
cut.Q       -527.9   23.4 -557.9 -528.3 -497.7
cut.C        367.7   20.4  341.8  367.7  393.5
cut^4         74.9   16.5   53.6   75.0   95.5
sigma       1511.5    4.6 1505.6 1511.5 1517.4
```

```
##
## Call:
## lm(formula = price ~ carat_z + cut, data = diamonds)
##
## Coefficients:
## (Intercept)      carat_z
##      3579.3       7871.1
##       cut.L        cut.Q
##      1239.8       -528.6
##       cut.C        cut^4
##       367.9         74.6
```

Man könnte hier noch einen Interaktionseffekt ergänzen.

16. **Aufgabe**



Wählen Sie das Diagramm, in dem *kein* Interaktionseffekt (in der Population) vorhanden ist (bzw. wählen Sie Diagramm, dass dies am ehesten darstellt).

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

Lösung

Das Streudiagramm `Diagramm D` zeigt *keinen* Interaktionseffekt.

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

17. **Aufgabe**

Der *t-Test* kann als Spezialfall der Regressionsanalyse gedeutet werden.

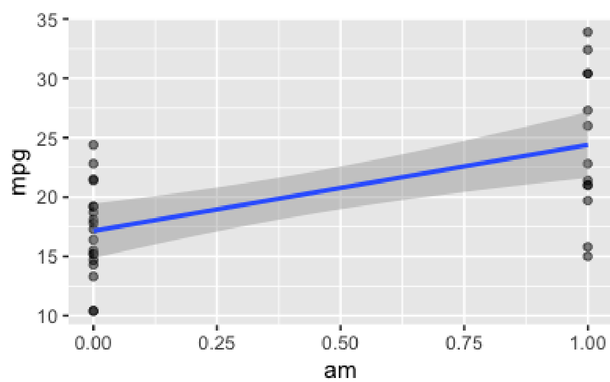
Hierbei ist es wichtig, sich das Skalenniveau der Variablen, die ein t-Test verarbeitet, vor Augen zu führen.

Benennen Sie die Skalenniveaus der AV eines t-Tests! Geben Sie nur *ein* Wort ein. Verwenden Sie nur Kleinbuchstaben (z.B. `regression`).

Lösung

Die richtige Antwort lautet: *metrisch*.

```
data(mtcars)
mtcars %>%
  ggplot() +
  aes(x = am, y = mpg) +
  geom_point(alpha = .5) +
  geom_smooth(method = "lm")
```



18. **Aufgabe**

Betrachten Sie folgende Ausgabe eines Bayesmodell, das mit `rstanarm` “gefittet” wurde:

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      price ~ cut
##  observations: 1000
##  predictors:   5
## ------
##              Median  MAD_SD
## (Intercept)   4571.7   675.1
## cutGood       -570.2   777.2
## cutIdeal     -1288.3   688.1
## cutPremium     362.5   709.8
## cutVery Good  -807.4   706.3
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 3795.0   82.4
```

Welche Aussage passt (am besten)?

Hinweise:

- Mit “Nullhypothese” ist im Folgenden dieser Ausdruck gemeint: $\mu_1 = \mu_2 = \ldots = \mu_k$.

a. Die Nullhypothese muss verworfen werden.
b. Die Nullhypothese muss beibehalten werden.
c. Unter der Annahmen, dass die Posteriori-Verteilung für jeden Regressionsparameter normal verteilt ist, kann man schließen, dass bei allen Gruppenmittelwerte 95% der Posteriori-Verteilung ungleich Null sind.
d. Unter der Annahmen, dass die Posteriori-Verteilung für jeden Regressionsparameter normal verteilt ist, kann man schließen, dass *nicht* bei allen Gruppenmittelwerte 95% der Posteriori-Verteilung gleich Null sind.
e. Unter der Annahmen, dass die Posteriori-Verteilung für jeden Regressionsparameter normal verteilt ist, kann man schließen, dass bei allen Gruppenmittelwerte 95% der Posteriori-Verteilung ungleich Null sind. Daher muss die Nullhypothese verworfen werden.

Lösung

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

19. **Aufgabe**

Berechnet man eine Posteriori-Verteilung mit `stan_glm()`, so kann man entweder die schwach informativen Prioriwerte der Standardeinstellung verwenden, oder selber Prioriwerte definieren.

Betrachten Sie dazu dieses Modell:

```
stan_glm(price ~ cut, data = diamonds,
                   prior = normal(location = c(100, 100, 100, 100),
                                  scale = c(100, 100, 100, 100)),
                   prior_intercept = normal(3000, 500))
```

Wie viele Parameter gibt es in diesem Modell?

Hinweise:

- Geben Sie nur eine (ganze) Zahl ein.

Lösung

Die Anzahl der Parameter in diesem Modell ist: 11

- Achsenabschnitt: 2 Parameter (MW, SD)
- 4 Regressionsparameter: je 2 Parameter (MW, SD)
- Sigma (Streuung der y-Werte): 1 Parameter (Rate lambda)

20. **Aufgabe**

Sei $r_{xy} = 0.7, s_x = 1, s_y = 1$.

Berechnen Sie b, den empirischen Regressionskoeffizienten!

Hinweise:

- Geben Sie nur Zahlen ein (ggf. inkl. Dezimaltrennzeichen).
- Runden Sie auf zwei Dezimalstellen.
- Führende Nullen sind (ggf.) anzugeben.

Lösung

Die richtige Antwort lautet: 0.7.

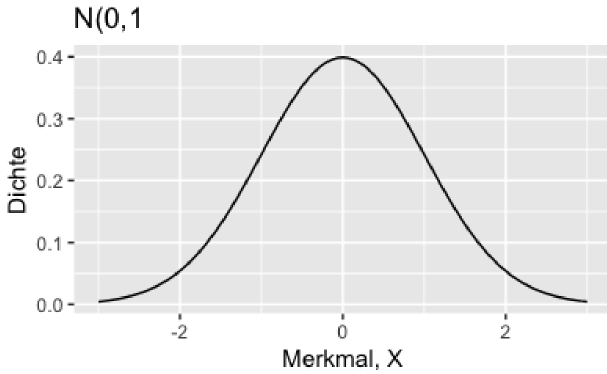
21. **Aufgabe**

Welche Verteilung ist (am besten) geeignet, um Streuung (σ) zu modellieren?

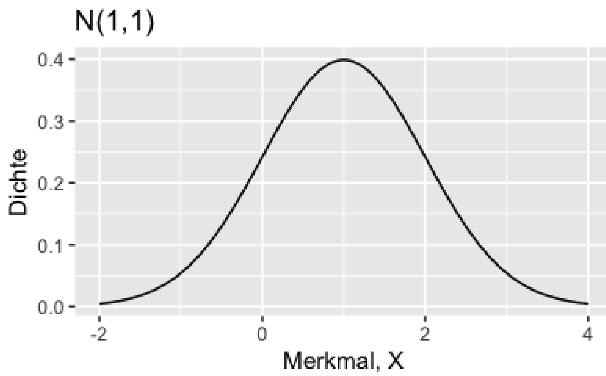
a. N(0,1)
b. N(1,1)
c. Exp(1)
d. Exp(0)
e. Exp(-1)

Lösung

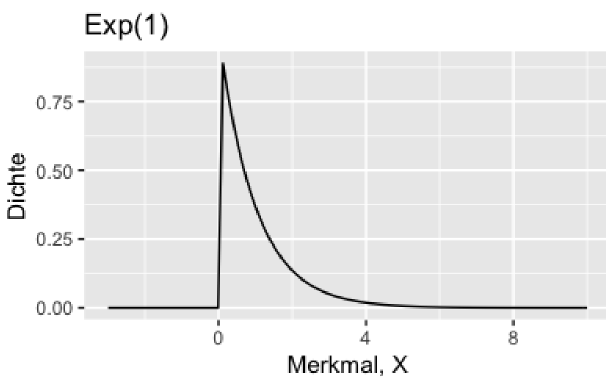
a. Falsch
b. Falsch
c. Wahr
d. Falsch
e. Falsch Da Streuung σ per Definition positiv ist, kommt eine Verteilung, die negative Werte erlaubt, nicht in Frage. Die Normalverteilung scheidet also aus. Die Rate der Exponentialverteilung regelt gleichzeitig Streuung und Mittelwert. Allerdings hat $Exp\,(0)$ eine unendliche Streuung, was nicht wünschenswert ist. Eine negative Rate ist für die Exponentialverteilung nicht definiert. Normalverteilungen: A)



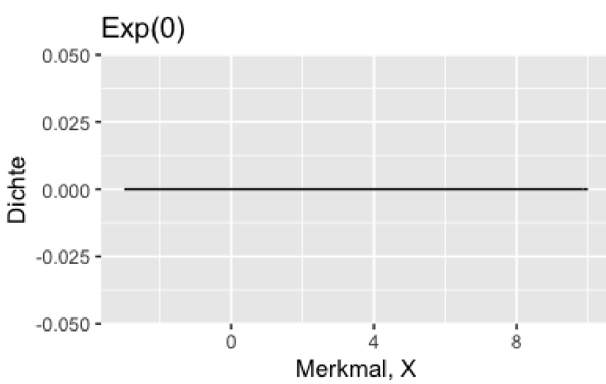
$N\,(0, 1)$:



B) $N\,(1, 1)$: Exponentialverteilungen: C) $Exp\,(1)$:

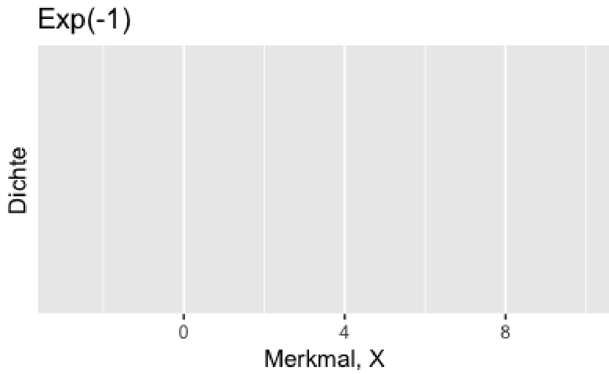


D) $Exp\,(0)$:



E) $Exp\,(-1)$: `## Warning in (function (x, rate = 1, ## log`

```
= FALSE) : NaNs produced ## Warning: Removed 101 row(s) ## containing missing values ## (geom_path).
```

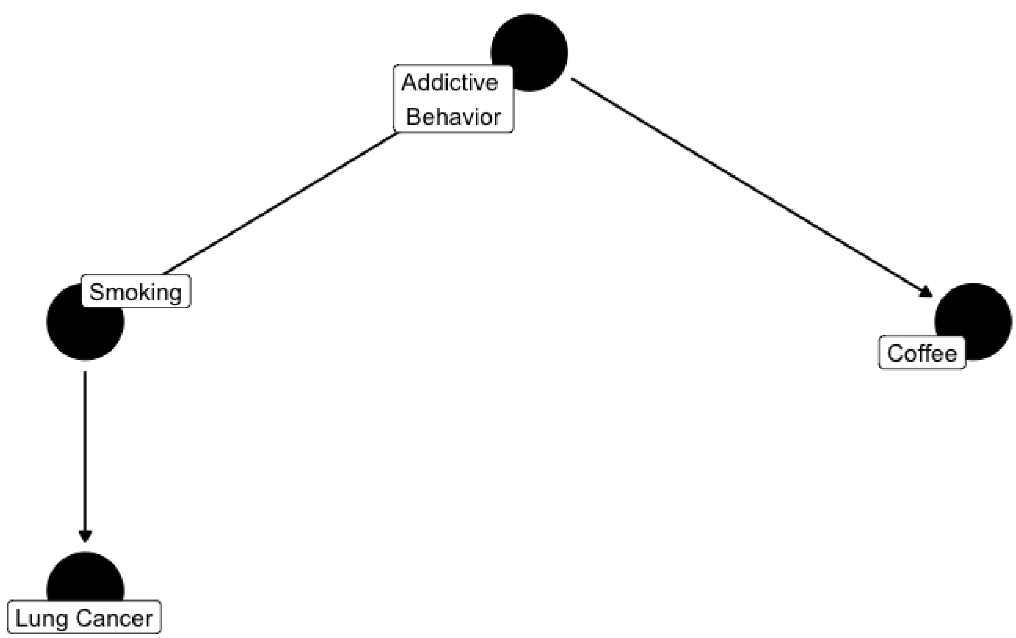


22. **Aufgabe**

    Gegeben sei die Theorie (oder schlichter: das Modell), demzufolge eine Anlage zu Suchtverhalten die Ursache von sowohl Rauchen als auch Kaffeegewohnheit darstellt. Lungenkrebs wiederum hat als (alleinige) Ursache Rauchen (laut diesem Modell).

    Daten zeigen, dass Kaffeegenuss und Lungenkrebs assoziiert sind: Bei Kaffeetrinkern ist die Lungenkrebsrate höher als bei Nichttrinkern (von Kaffee). Ob Kaffeegebrauch Lungenkrebs erzeugt?

    Eine alternative Erklärung bietet folgender DAG.

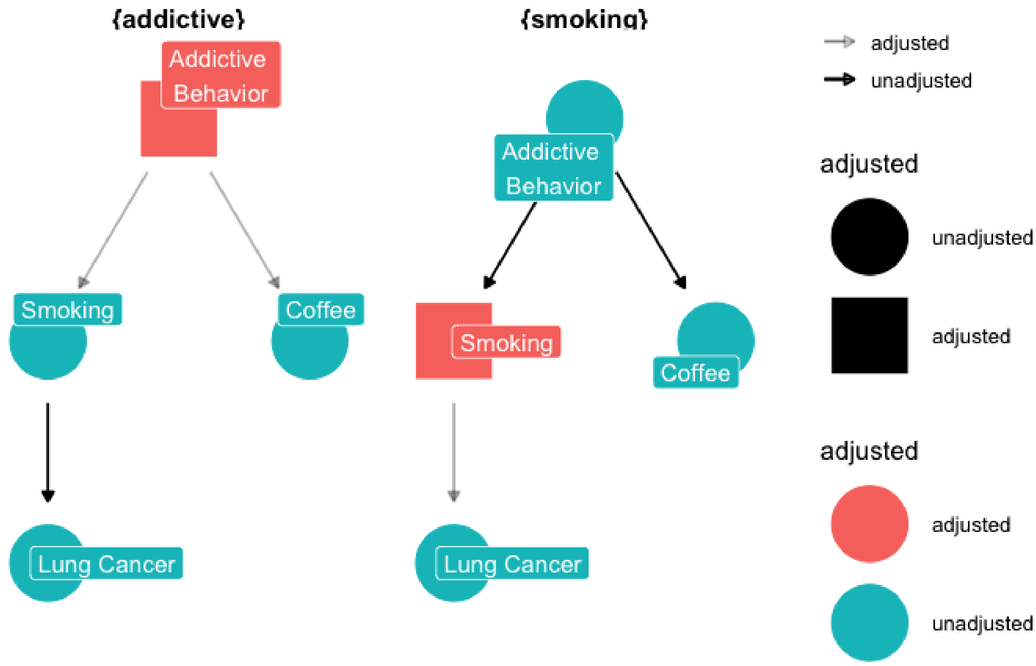


Welche Variablenmenge muss *mindestens* kontrolliert werden, um Konfundierung auszuschließen und damit den kausalen Effekt von Kaffee auf Lungenkrebs zu identifizieren?

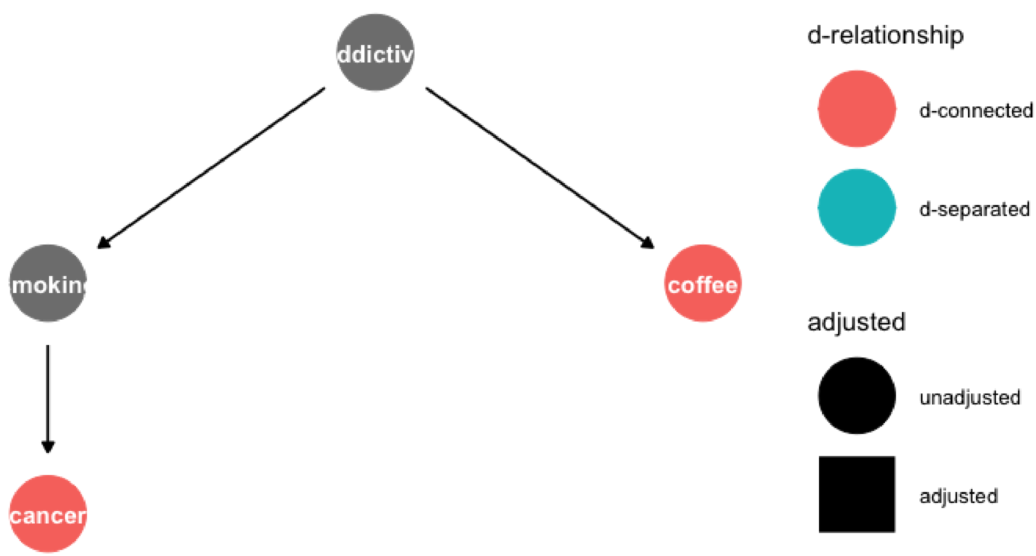
a. {Addictive Behavior *oder aber* Rauchen}
b. {Addictive Behavior *und* Rauchen}
c. {Rauchen}
d. {Addictive Behavior}
e. {Addictive Behavior *und* Lungenkrebs}

Lösung

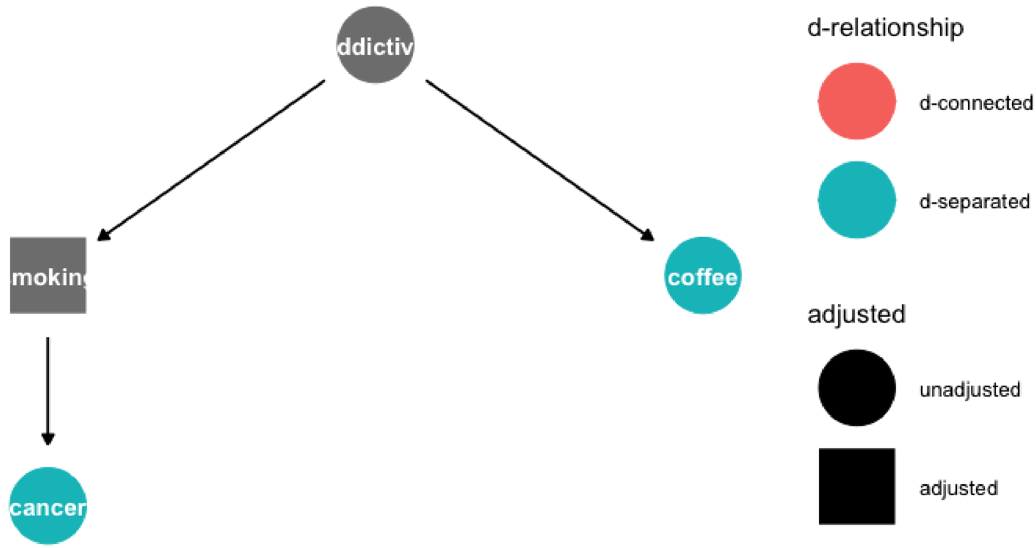
Durch Kontrolle von `Addictive Behavior` *oder aber* `Rauchen` wird der kausale Effekt von `Coffee` auf `Lung Cancer` identifiziert.



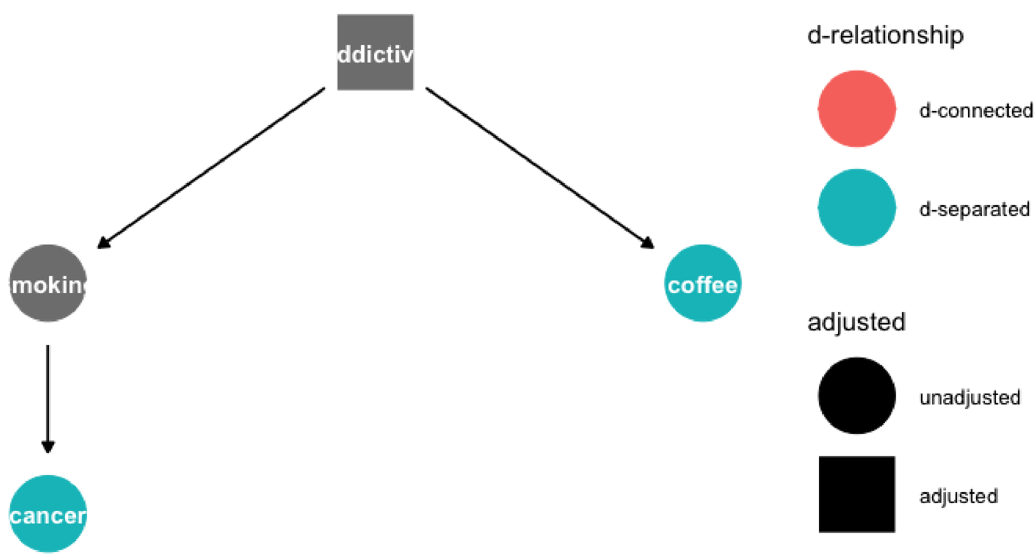
OHNE Kontrolle:



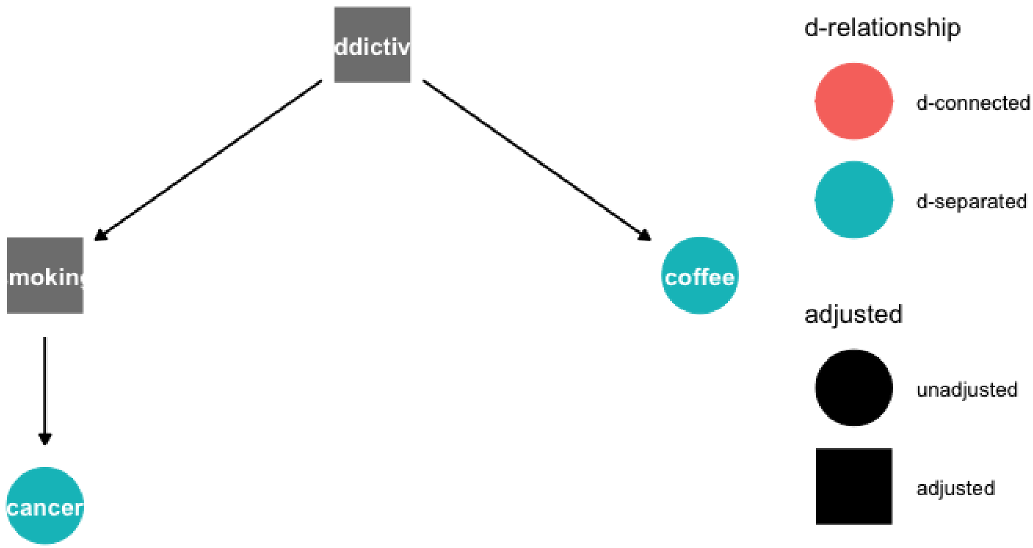
MIT Kontrolle von `smoking` :



MIT Kontrolle von `addictive` :



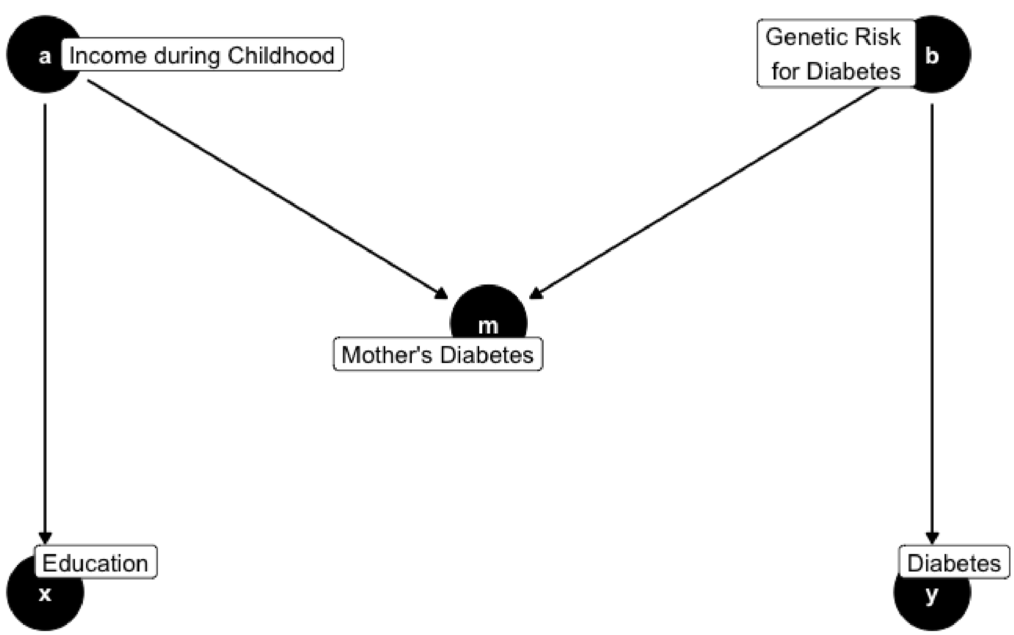
MIT Kontrolle von `addictive` und `smoking`:



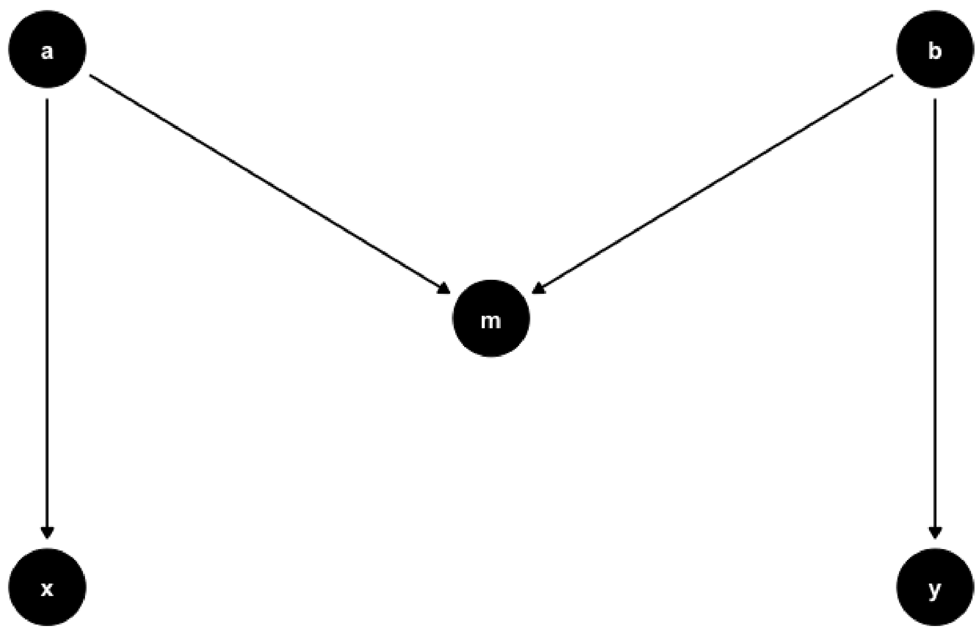
a. Wahr
b. Falsch
c. Falsch
d. Falsch
e. Falsch

23. **Aufgabe**

Ein Forschungsteam aus Epidemiologen untersucht den (möglicherweise kausalen) Zusammenhang von Erziehung (`education`) und Diabetes (`diabetes`). Das Team schlägt folgendes Modell zur Erklärung des Zusammenhangs vor (s. DAG).



Nochmal den gleich DAG ohne "Schilder", damit man die Pfeilspitzen besser sieht:



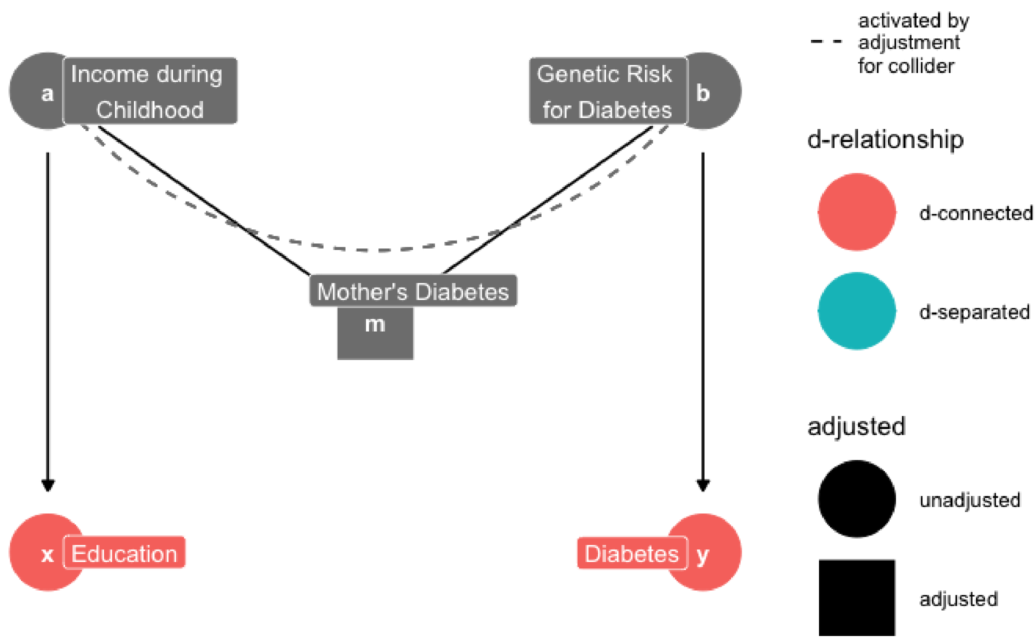
Sollte die Krankengeschichte der Mutter hinsichtlich Diabetes kontrolliert werden, um den kausalen Effekt von Erziehung auf Diabetes zu identifizieren?

a. Ja, `Mother's Diabetes` sollte kontrolliert werden, da so eine Konfundierung vermieden wird.
b. Ja, `Mother's Diabetes` sollte kontrolliert werden, da so ein Collider Bias (Kollisionsverzerrung) vermieden wird.
c. Nein, `Mother's Diabetes` sollte *nicht* kontrolliert werden, da zwar keine Verzerrung entsteht, es aber auch nicht nötig ist.
d. Nein, `Mother's Diabetes` sollte *nicht* kontrolliert werden, da so eine Konfundierung resultiert.
e. Nein, `Mother's Diabetes` sollte *nicht* kontrolliert werden, da so ein Collider Bias (Kollisionsverzerrung) resultiert.

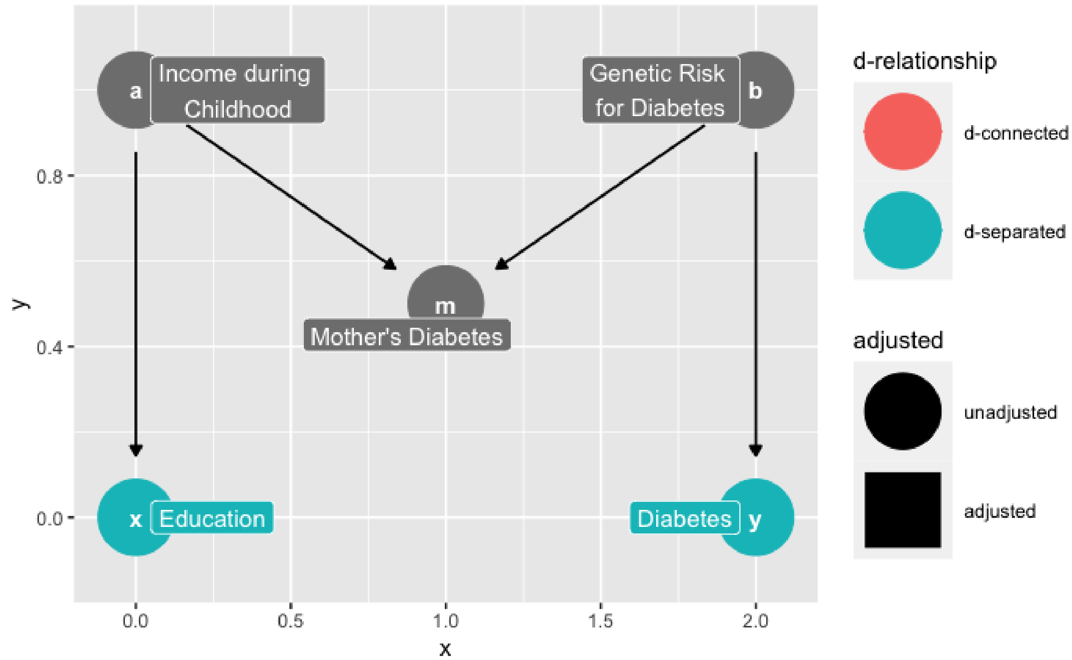
Lösung

Durch Kontrolle von `Mother's Diabetes` wird eine Scheinkorrelation erzeugt, wo es vorher keine gab. Das nennt man eine Kollisionsverzerrung (collider bias). Daher sollte `Mother's Diabetes` *nicht* kontrolliert werden.

Im foglenden Diagramm ist der durch Kontrolle einer Kollisionsvariable geöffntete Pfad von `a` nach `b` im DAG eingezeichnet:



OHNE Kontrolle gibt es keine Verbindung zwischen `x` und `y` (sie sind d-separiert).

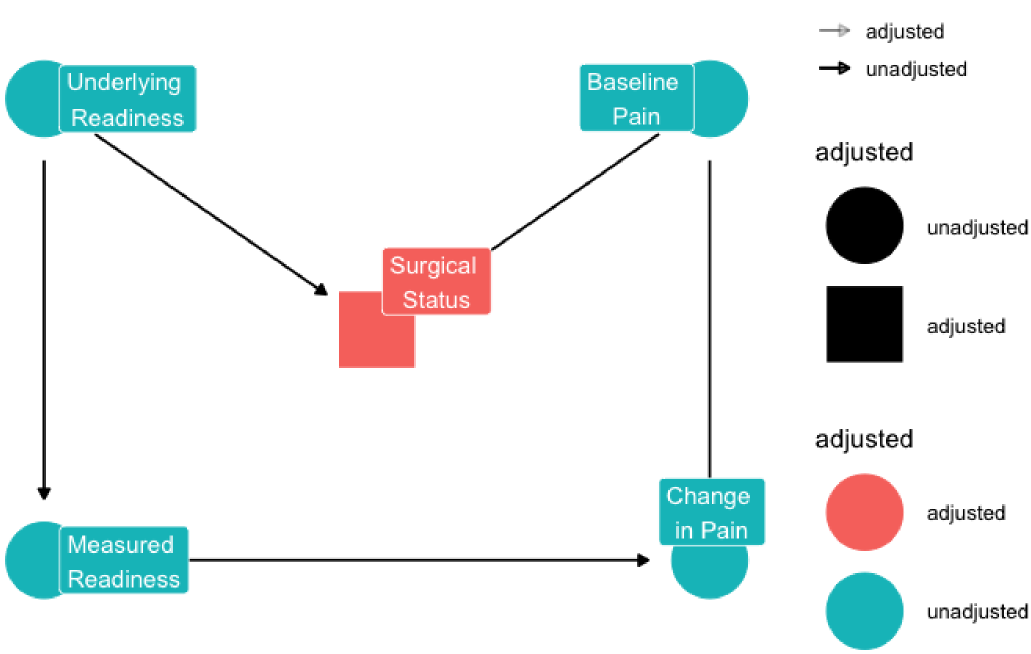


a. Falsch
b. Falsch
c. Falsch
d. Falsch
e. Wahr

24. **Aufgabe**

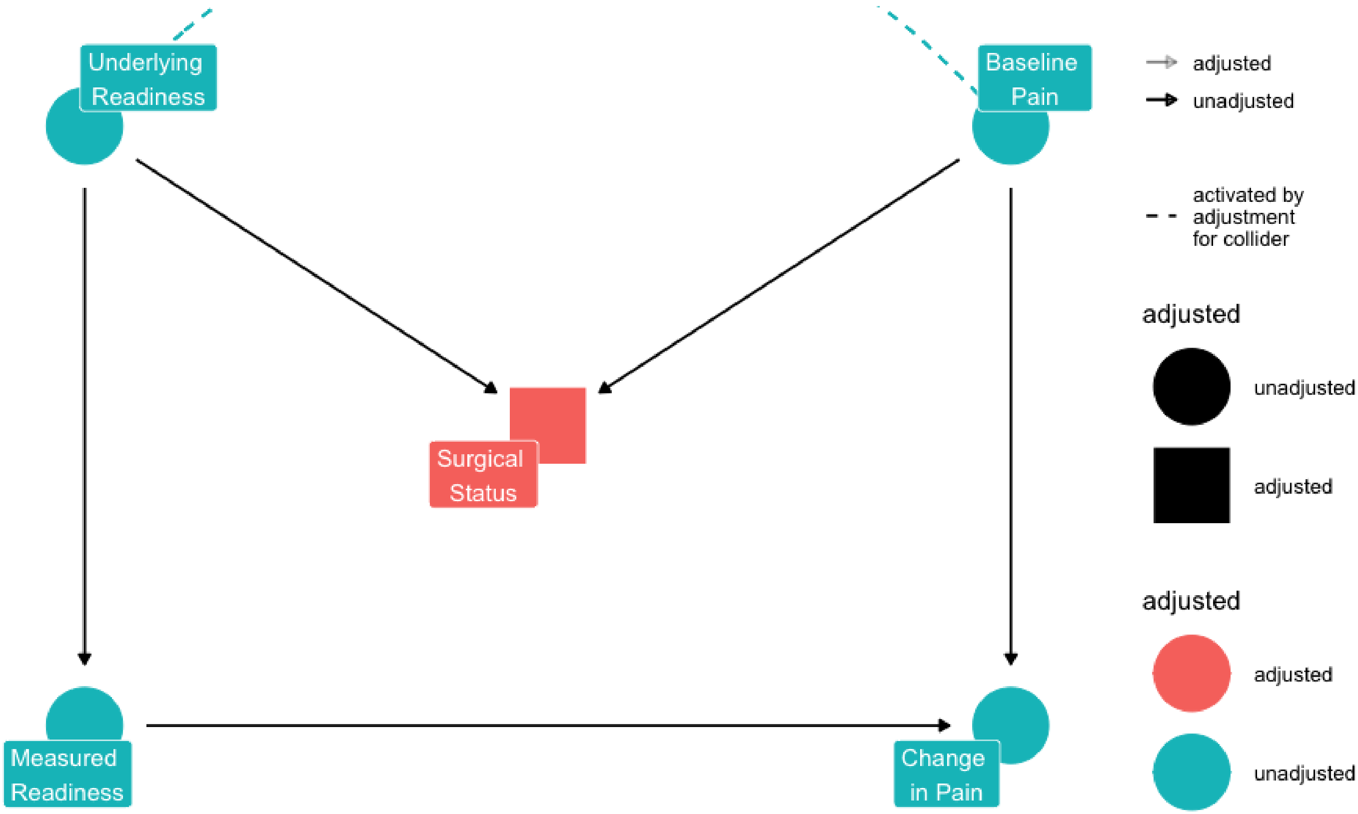
Ein Forschungsteam aus Psychologen und Medizinern untersucht die Frage, ob Menschen, die eine hohe Bereitschaft für eine OP und zu Veränderung in ihrer Lebensführung aufweisen (mittels Fragebogen gemessen), nach einem Jahr über einen höheren Schmerzrückgang verfügen als Patienten geringerer Bereitschaft. Die Studie umfasst ausschließlich Patienten, die eine OP wegen Rückenschmerzen durchlaufen sind (s. DAG).

Das Studiendesign impliziert, dass nur Patienten, die eine OP durchlaufen haben, in die Studie aufgenommen wurde. Damit wird per Design diese Variable stratifiziert (kontrolliert).



Durch die Stratifizierung wird ein Hintertürpfad geöffnet; dieser muss geschlossen werden. Wie sollte dies geschehen (in diesem Modell)?

Im folgenden Diagramm ist der Kollisionsbias kenntlich gemacht, der durch die Stratifizierung von `Surgical Status` entsteht:



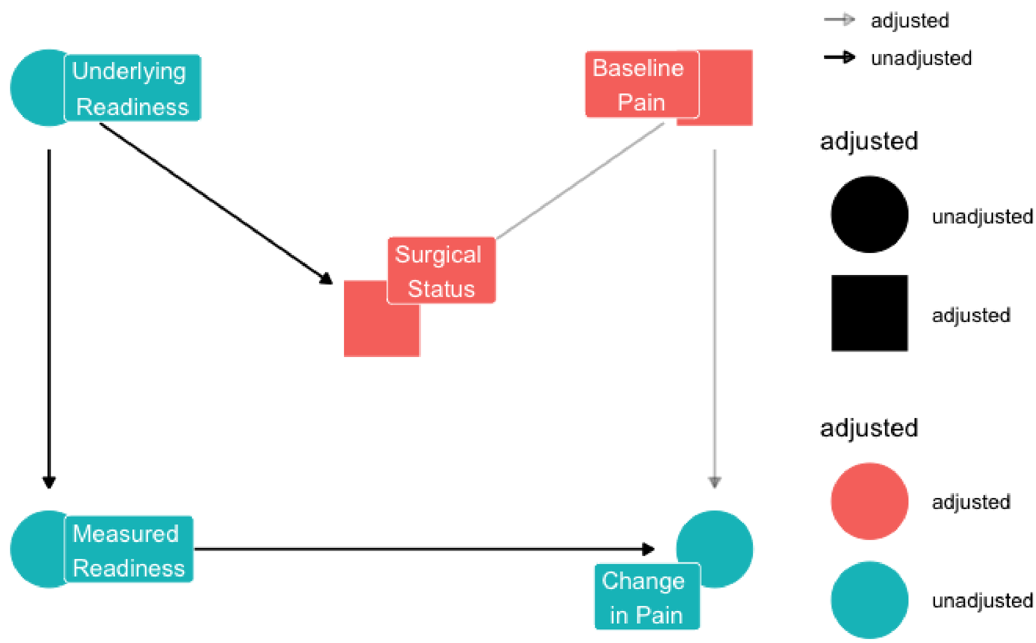
Hinweis:

- Wenn von “kausaler Effekt” gesprochen wird, ist stets der kausale Effekt wie oben definiert gemeint (von Erziehung auf Diabetes).

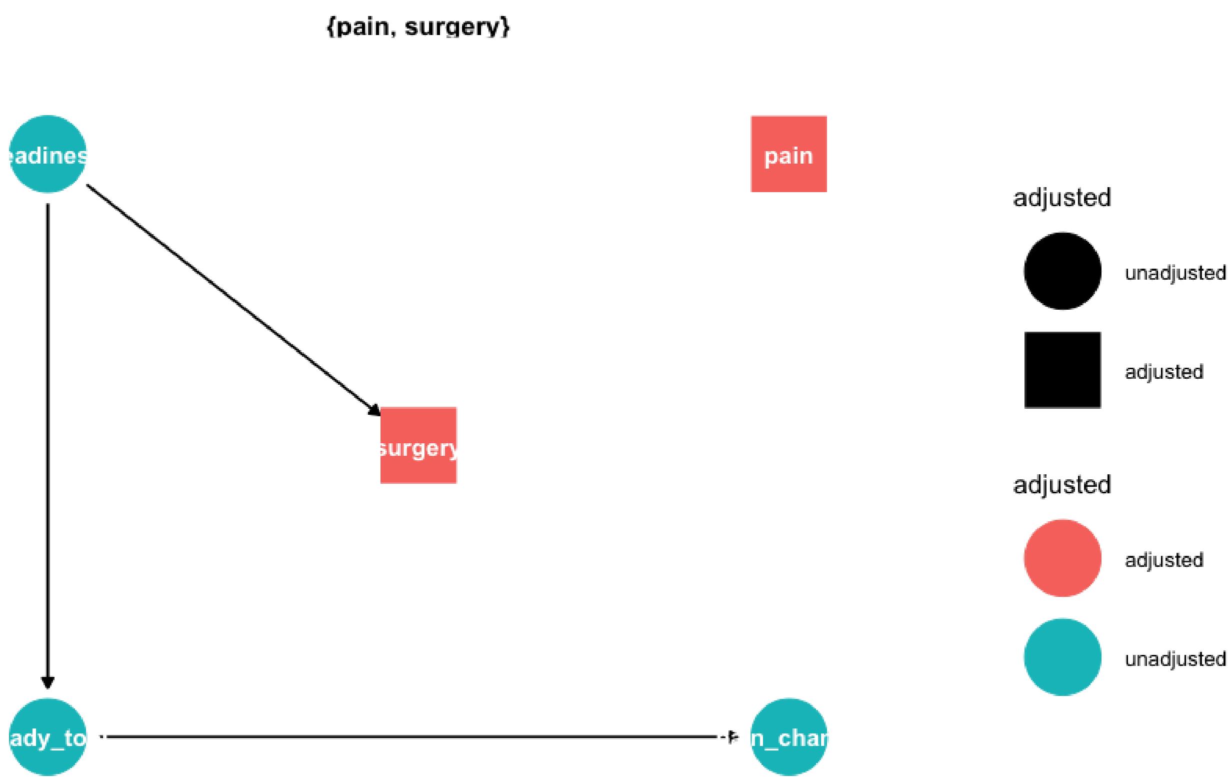
a. Es sollte vom Forschungsteam auf `underlying readiness` kontrolliert werden, um den kausalen Effekt zu identifizieren.
b. Es sollte vom Forschungsteam auf `Change in Pain` kontrolliert werden, um den kausalen Effekt zu identifizieren.
c. Es sollte vom Forschungsteam auf `Measured Readiness` kontrolliert werden, um den kausalen Effekt zu identifizieren.
d. Es sollte vom Forschungsteam auf `underlying readiness` kontrolliert werden, um den kausalen Effekt zu identifizieren.
e. Es sollte vom Forschungsteam auf `Baseline Pane` kontrolliert werden, um den kausalen Effekt zu identifizieren.

Lösung

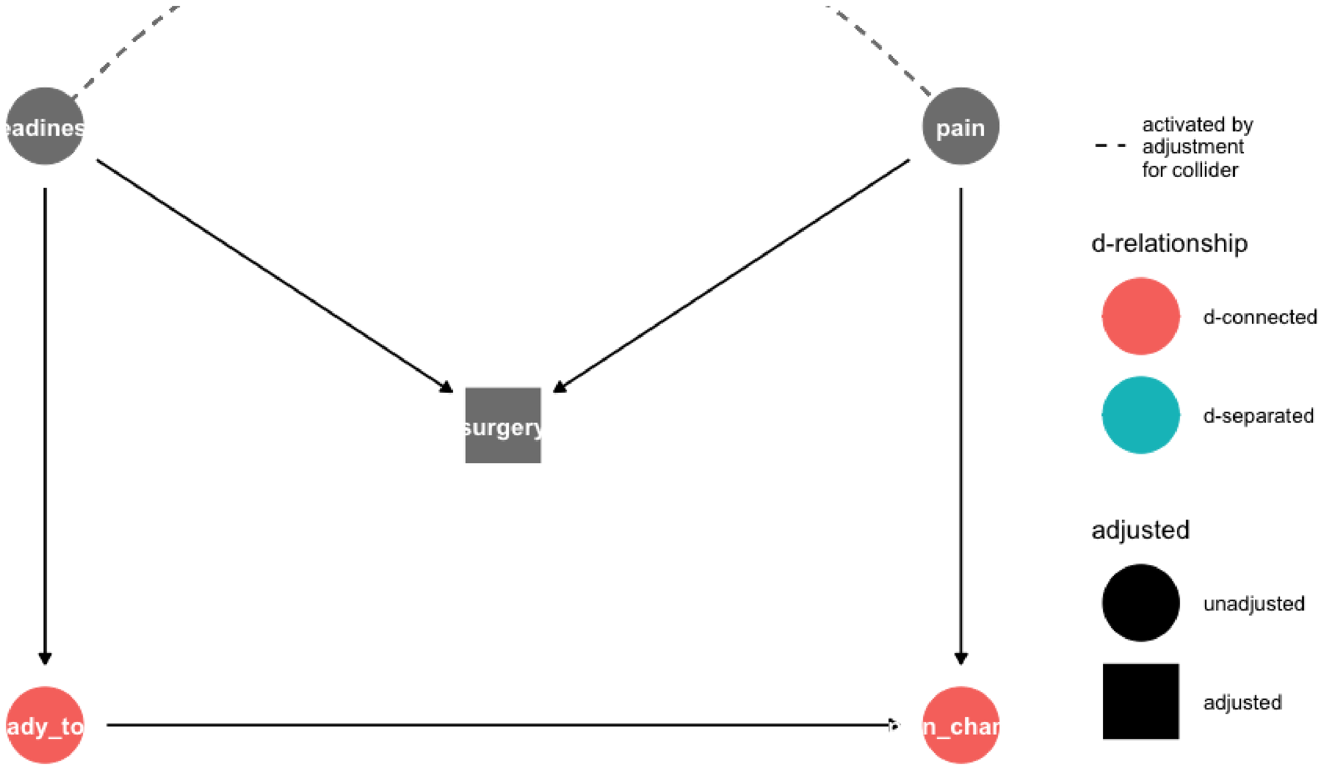
Es sollte auf `Baseline Pain` kontrolliert werden, um den kausalen Effekt zu identifizieren.



Mit Kontrolle von `Baseline Pain` (und `surgery`) ist der kausale Effekt identifiziert:



Ohne Kontrolle von `Baseline Pain` ist der Effekt *nicht* identifiziert; es gibt einen Hintertürpfad:



[Quelle]()

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