Lösungen zu den Aufgaben

1. **Aufgabe**

   Die Varianzanalyse (ANOVA) ist ein inferenzstatistisches Verfahren des Frequentismus. Welches Skalenniveau passt zu diesem Verfahren?

   a. UV: nominal (mehrstufig), AV: metrisch
   b. UV: metrisch, AV: nominal (zweistufig)
   c. UV: nominal (mehrstufig), AV: nominal (mehrstufig)
   d. UV: metrisch, AV: nominal (zweistufig)
   e. UV: nominal (zweistufig), AV: metrisch

   Lösung

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

2. **Aufgabe**

   Welches der folgenden Beispiele ist *kein* Beispiel für eine Nullhypothese?

   a. $\beta_1 <= 0$
   b. $\mu_1 = \mu_2$
   c. $\mu_1 = \mu_2 = ... = \mu_k$
   d. $\rho = 0$
   e. $\pi_1 = \pi_2$

   Lösung

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

3. **Aufgabe**

   Der t-Test ist ein inferenzstatistisches Verfahren des Frequentismus. Welches Skalenniveau passt zu diesem Verfahren?

   a. UV: nominal (mehrstufig), AV: metrisch
   b. UV: metrisch, AV: nominal (zweistufig)
   c. UV: nominal (mehrstufig), AV: nominal (mehrstufig)
   d. UV: metrisch, AV: nominal (zweistufig)
   e. UV: nominal (zweistufig), AV: metrisch

   Lösung

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

4. **Aufgabe**

Für Statistiken (Stichprobe) verwendet man meist lateinische Buchstaben; für Parameter (Population) verwendet man entsprechend meist griechische Buchstaben.

Vervollständigen Sie folgende Tabelle entsprechend!

Kennwert	**Statistik**	**Parameter**
Mittelwert	$\overline{X}$	NA
Mittelwertsdifferenz	$\overline{X}_1 - \overline{X}_2$	NA
Streuung	sd	NA
Anteil	p	NA
Korrelation	r	NA
Regressionsgewicht	b	NA

Lösung

Kennwert	**Statistik**	**Parameter**
Mittelwert	$\overline{X}$	μ
Mittelwertsdifferenz	$\overline{X}_1 - \overline{X}_2$	μ_1 - μ_2
Streuung	sd	σ
Anteil	p	π
Korrelation	r	ρ
Regressionsgewicht	b	β

5. **Aufgabe**

Das Testen von Nullhypothesen wird u.a. deswegen kritisiert, weil die Nullhypothese zumeist apriori als falsch bekannt ist, weswegen es keinen Sinne mache, so die Kritiker, sie zu testen.

Nennen Sie ein Verfahren von John Kruschke, das einen Äquivalenzbereich testet und insofern eine Alternative zum Testen von Nullhypothesen anbietet.

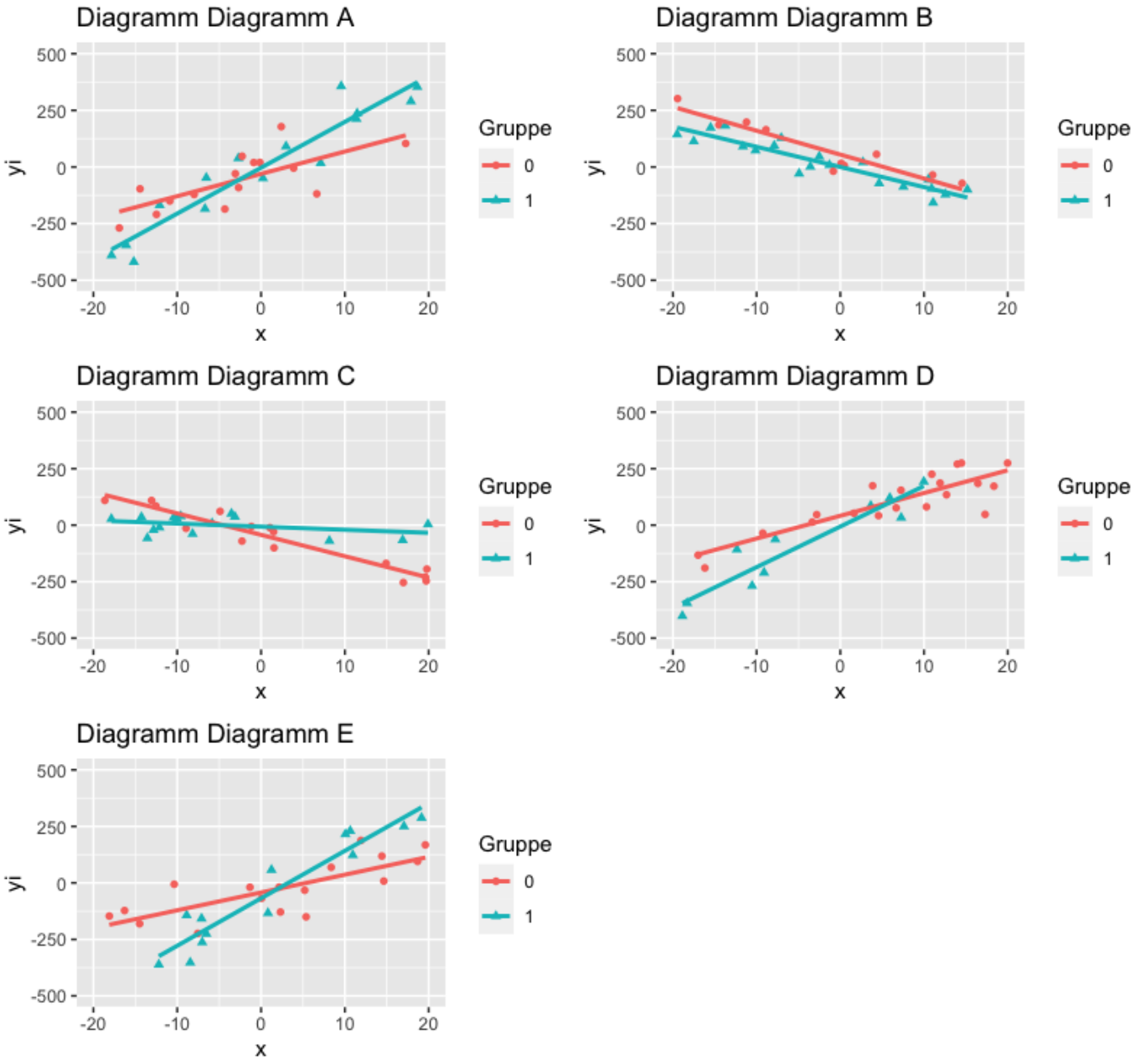
Hinweise:

- Geben Sie nur Kleinbuchstaben ein.
- Geben Sie nur ein einziges Wort ein.

Lösung

rope

6. **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

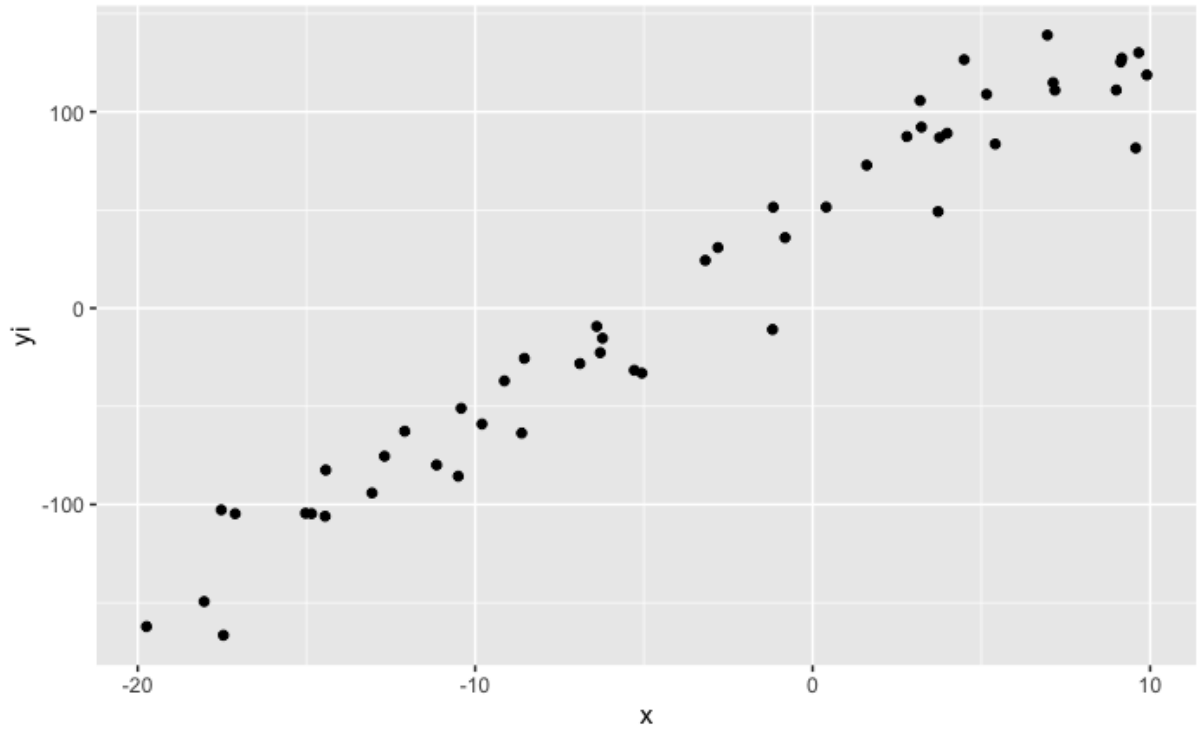
Lösung

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

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

7. **Aufgabe**

Ein Streudiagramm von x und y ergibt folgende Abbildung:

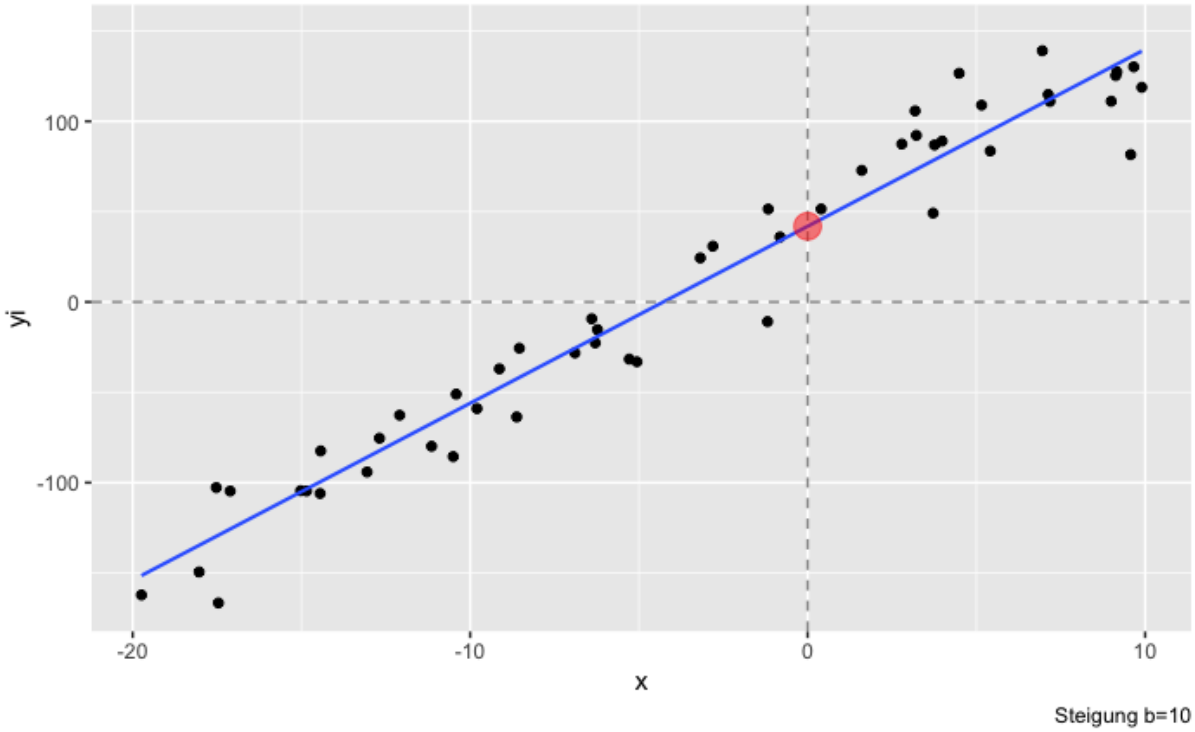


Wählen Sie das am besten passende Modell aus der Liste aus!

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

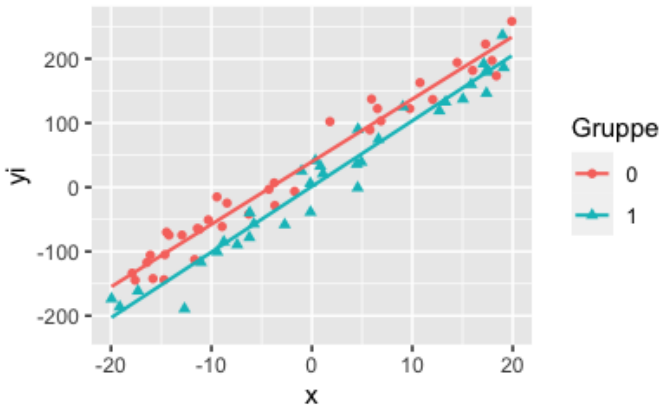
Das dargestellte Modell lautet $y = 40 + 10 \cdot x + \epsilon$.



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

8. **Aufgabe**

Ein Streudiagramm von x und y ergibt folgende Abbildung; dabei wird noch die Gruppierungsvariable g (mit den Stufen `0` und `1`) berücksichtigt (vgl. Farbe und Form der Punkte). Zur besseren Orientierung ist die Regressionsgerade pro Gruppe eingezeichnet.



Wählen Sie das (für die Population) am besten passende Modell aus der Liste aus!

Hinweis: Ein Interaktionseffekt der Variablen x und g ist mit xg gekennzeichnet.

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

Lösung

Das dargestellte Modell lautet $y = 40 + 10 \cdot x + -40 \cdot g + 0 \cdot xg + \epsilon$. Der Modellfehler ϵ hat den Anteil 0.2 im Vergleich zur Streuung von y.

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

9. **Aufgabe**

Gegeben sei ein Datensatz mit folgenden Prädiktoren, wobei Studierende die Beobachtungseinheit darstellen:

- X_1: Quereinsteiger (0: nein, 1: ja)
- X_2: Intelligenz-Testwert (z-Wert)
- X_3: Matheanteil im Studium (z-Wert)
- X_4: Interaktion von $X1$ und $X2$

Die vorherzusagende Variable (Y; Kriterium) ist *Gehalt nach Studienabschluss*.

Folgende Modellparameter einer Regression (Least Squares) seien gegeben:

- $\beta_0 : 50$
- $\beta_1 : 40$
- $\beta_2 : 3$
- $\beta_3 : 10$
- $\beta_4 : 7$

Welche der Aussagen ist korrekt?

a. Für einen bestimmten (festen) Wert von $X_2 =$ `Intelligenz-Testwert (z-Wert)` und $X_3 =$ `Matheanteil im Studium (z-Wert)` gilt, dass das Gehalt im Mittel höher ist bei $X_1 = 1$ im Vergleich zu $X_1 = 0$, laut dem Modell.
b. Für einen bestimmten (festen) Wert von $X_2 =$ `Intelligenz-Testwert (z-Wert)` und $X_3 =$ `Matheanteil im Studium (z-Wert)` gilt, dass das Gehalt im Mittel höher ist bei $X_1 = 0$ im Vergleich zu $X_1 = 1$, laut dem Modell.
c. Der mittlere Gehaltsunterschied Y zweier Personen a und b, wobei bei Person a gilt $X_1 = 0$ und bei Person b gilt $X_1 = 1$, beträgt stets 50, laut dem Modell.
d. Der mittlere Gehaltsunterschied Y zweier Personen a und b, wobei bei Person a gilt $X_1 = 0$ und bei Person b gilt $X_1 = 1$, kann *nicht* ohne weitere Angaben auf eine Zahl fixiert werden, laut dem Modell.
e. Der mittlere Gehaltsunterschied von Menschen ist eine Wirkung von genau drei Ursachen: `Quereinsteiger (0: nein, 1: ja)`, `Intelligenz-Testwert (z-Wert)`, `Matheanteil im Studium (z-Wert)`, laut dem Modell.

Lösung

- Wahr
- Falsch
- Falsch
- Falsch
- Falsch

10. **Aufgabe**

Welches Ergebnis hat der R-Befehl `posterior_interval()` (R-Paket `rstanarm`)?

Wählen Sie die (am besten) passende Antwort aus.

Hinweis:

- Soweit nicht anders benannt, ist immer die Voreinstellung der betreffenden Funktion gemeint.

a. Er liefert einen Vorhersagewert aus der Posteriori-Verteilung.
b. Er liefert ein Vorhersageintervall aus der Posteriori-Verteilung.
c. Er liefert ein 90%-Vorhersageintervall aus der Posteriori-Verteilung.
d. Er liefert ein 95%-Vorhersageintervall aus der Posteriori-Verteilung.
e. Er liefert ein HDI-Vorhersageinterval aus der Posteriori-Verteilung.

Lösung

So können Sie sich Hilfe zu diesem Befehl ausgeben lassen:

```
help(posterior_interval)
```

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

11. **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 ungleich 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

12. **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))
```

Welche Aussage dazu passt (am besten)?

a. Es wird für (genau) einen Parameter eine Priori-Verteilung definiert.
b. Für keinen Parameter liegt apriori die Wahrscheinlichkeit für einen negativen Wert bei mehr als 5%.
c. Mit `prior = normal()` werden Gruppenmittelwerte definiert.
d. Alle Parameter des Modells sind normalverteilt.
e. mit `prior_intercept = normal(3000, 500)` wird praktisch eine Gleichverteilung definiert (da die Streuung sehr hoch ist).

Lösung

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

13. **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)

14. **Aufgabe**

    Sei $X \sim \mathcal{N}(42, 7)$ und $x_1 = 28$.

    Berechnen Sie den z-Wert für x_1!

    Hinweis:

    - Runden Sie ggf. auf die nächste ganze Zahl.

    Lösung

    ```
    x1_z = (x1 - x_mw) / x_sd
    ```

    -2

15. **Aufgabe**

    John Kruschke hat einen (Absolut-)Wert vorschlagen, als Grenze für Regressionskoeffizienten "vernachlässigbarer" Größe.

    Nennen Sie diesen Wert!

    Hinweise:

    - Geben Sie nur Zahlen ein (und ggf. Dezimaltrennzeichen).
    - Führende Nullen dürfen auch bei Zahlen kleiner als 1 nicht weggelassen werden.

    Lösung

    0.05

16. **Aufgabe**

    Im Datensatz `mtcars`: Ist der (mittlere) Unterschied im Spritverbrauch zwischen den beiden Gruppen *Automatik* vs. *Schaltgetriebe* vernachlässigbar?

    Definieren Sie selber, was "vernachlässigbar klein" bedeutet. Oder greifen Sie auf die Definition "höchstens eine Meile" zurück.

    Prüfen Sie rechnerisch, anhand des angegebenen Datensatzes, folgende Behauptung:

    Behauptung: "Der Unterschied ist vernachlässigbar klein!"

    Wählen Sie die Antwortoption, die am besten zu der obigen Behauptung passt!

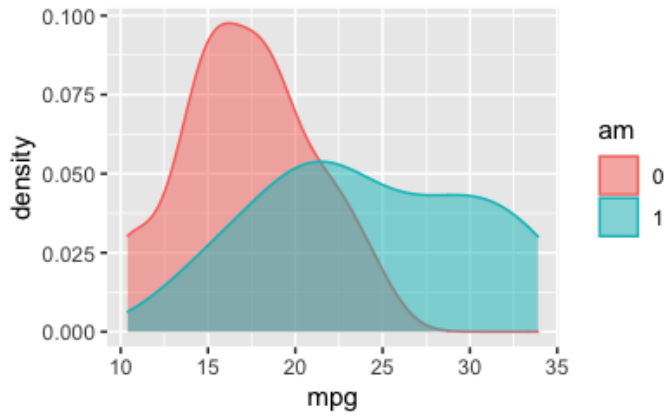
    Hinweise:

    - Sie benötigen einen Computer, um diese Aufgabe zu lösen.
    - Verwenden Sie die statistischen Methoden, die im Unterricht behandelt wurden.
    - Verwenden Sie Ansätze aus der Bayes-Statistik zur Lösung dieser Aufgabe.

a. Ja, die Behauptung ist korrekt.
b. Nein, die Behauptung ist falsch.
c. Die Daten sind bzw. das Modell nicht konkludent; es ist keine Entscheidung über die Behauptung möglich.
d. Auf Basis der bereitgestellten Informationen ist keine Entscheidung möglich über die Behauptung.

Lösung

```
mtcars %>%
  mutate(am = factor(am)) %>%
  ggplot() +
  aes(x = mpg, color = am, fill = am) +
  geom_density(alpha = .5)
```



Modell berechnen:

```
library(rstanarm)
library(tidyverse)
data(mtcars)

m1_mtcars <- stan_glm(mpg ~ am, data = mtcars, refresh = 0)
```

Posteriori-Verteilung betrachten:

```
m1_mtcars
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      mpg ~ am
##  observations: 32
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 17.2    1.1
## am           7.1    1.8
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 4.9    0.6
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
coef(m1_mtcars)
```

```
## (Intercept)          am
##        17.2         7.1
```

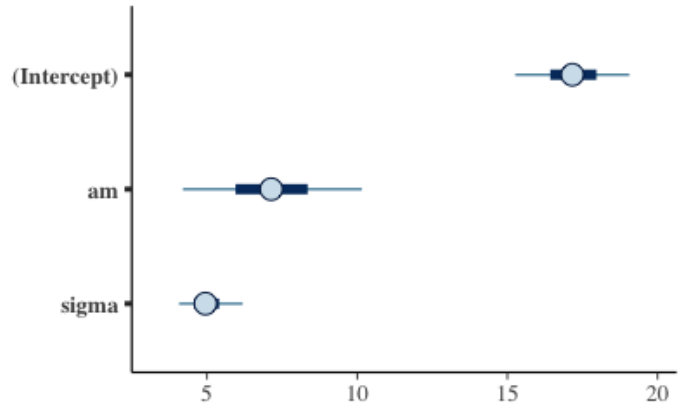
```
posterior_interval(m1_mtcars, prob = .95)
```

```
##             2.5% 97.5%
## (Intercept) 14.9  19.5
## am           3.6  10.6
## sigma        3.9   6.5
```

Spuckt ein PI aus, kein HDI (HDI noch nicht implementiert in `rstanarm`).

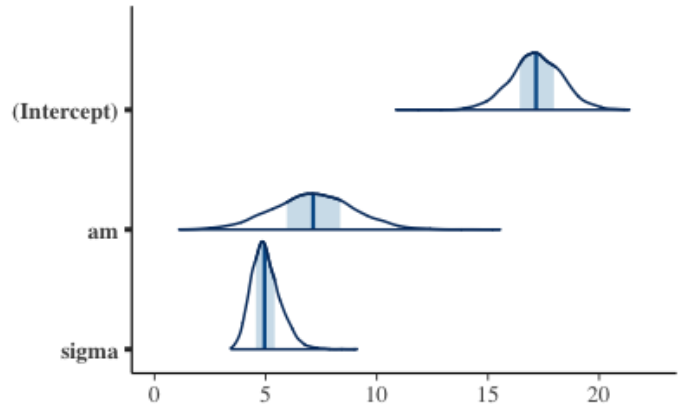
Visualisieren der Posteriori-Verteilung:

```
plot(m1_mtcars)
```



Oder als Histogramm:

```
library(bayesplot)
mcmc_areas(m1_mtcars)
```



Man sieht direkt, dass der Unterschied komplett außerhalb des Rope liegt.

Rope berechnen:

```
library(bayestestR)
rope(m1_mtcars)
```

```
## # Proportion of samples inside the ROPE [-0.60, 0.60]:
##
## Parameter   | inside ROPE
## ------------------------
## (Intercept) |      0.00 %
## am          |      0.00 %
```

Rope visualisieren:

```
plot(rope(m1_mtcars))
```

```
## Error: Failed at retrieving data :( Please provide original model or data through the `data` argument
```

Man sieht, dass der “Berg” - die Posteriori-Verteilung bzw. der Bereich plausibler Werte - außerhalb des Rope-Bereichs liegt.

Wir können also die Hypothese, dass der Unterschied zwischen beiden Gruppen praktisch Null ist, verwerfen.

Natürlich ist das nur ein deskriptiver Befund, wir können nichts dazu sagen, ob der Unterschied auch ein kausaler Effekt ist.

Alternative Rope-Definition: Z-Standardisieren.

Ein kleiner Effekt ist, laut Kruschke 2018, ein Unterschied der nicht größer ist als ±0.1 SD.

```
m2_mtcars <-
  mtcars %>%
  mutate(mpg_z = scale(mpg)) %>%
  stan_glm(mpg_z ~ am, data = ., refresh = 0)

rope(m2_mtcars)
```

```
## # Proportion of samples inside the ROPE [-0.10, 0.10]:
##
## Parameter   | inside ROPE
## ------------------------
## (Intercept) |      0.00 %
## am          |      0.00 %
```

```
plot(rope(m2_mtcars))
```

```
## Error: Failed at retrieving data :( Please provide original model or data through the `data` argument
```

a. Richtig
b. Falsch
c. keine Entscheidung zur Behauptung möglich
d. Die Frage ist mit den gegebenen Daten nicht zu beantworten.

17. **Aufgabe**

Einer der (bisher) größten Studien der Untersuchung psychologischer Konsequenzen (oder Korrelate) der Covid-Zeit ist die Studie COVIDiStress.

Im Folgenden sollen Sie folgende Forschungsfrage untersuchen:

Ist der Zusammenhang von Stress (`PSS10_avg`, AV) und Neurotizismus (`neu`, UV) vernachlässigbar klein?

Den Datensatz können Sie so herunterladen (Achtung, groß):

```
## Warning: One or more parsing issues,
## see `problems()` for details
```

Hinweise:

- Sie benötigen einen Computer, um diese Aufgabe zu lösen.
- Verwenden Sie die statistischen Methoden, die im Unterricht behandelt wurden.
- Verwenden Sie Ansätze aus der Bayes-Statistik zur Lösung dieser Aufgabe.

a. Ja
b. Nein
c. Die Daten sind nicht konkludent; es ist keine Entscheidung möglich.
d. Auf Basis der bereitgestellten Informationen ist keine Entscheidung möglich.

Lösung

Pakete laden:

Wie groß ist der Datensatz (im Speicher) eigentlich, in Megabyte?

```
## 156.8 bytes
```

Relevante Spalten auswählen:

Datensatz aufbereiten:

Modell berechnen:

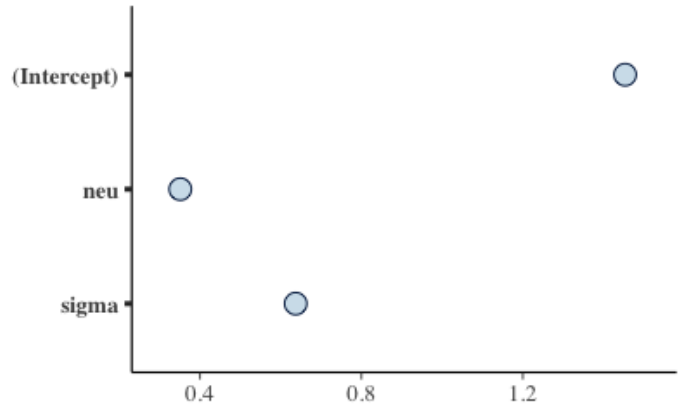
Modellkoeffizienten auslesen:

```
## (Intercept)         neu
##        1.45        0.35
```

Posteriori-Verteilung auslesen:

```
##             5.5% 94.5%
## (Intercept) 1.44  1.46
## neu         0.35  0.35
## sigma       0.63  0.64
```

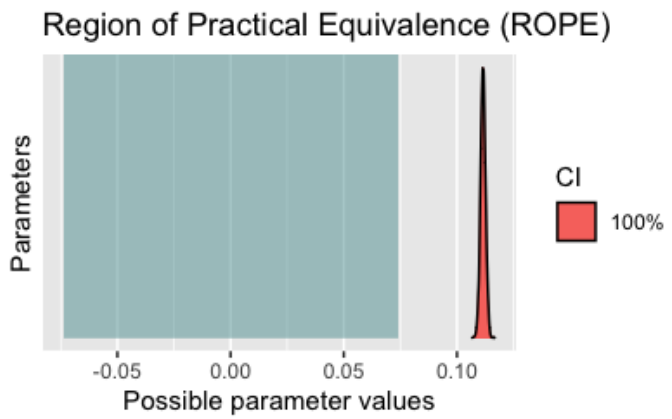
Posteriori-Verteilung plotten:



Rope berechnen:

```
## # Proportion of samples inside the ROPE [-0.07, 0.07]:
##
## Parameter   | inside ROPE
## ------------------------
## (Intercept) |      0.00 %
## neu         |      0.00 %
```

Rope visulasieren:



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

18. **Aufgabe**

Einer der (bisher) größten Studien der Untersuchung psychologischer Konsequenzen (oder Korrelate) der Covid-Zeit ist die Studie COVIDiStress.

Im Folgenden sollen Sie folgende Forschungsfrage untersuchen:

Forschungsfrage:

Ist der Unterschied zwischen Männern und Frauen (`Dem_gender`) im Hinblick zum Zusammenhang von Stress (`PSS10_avg`, AV) und Neurotizismus (`neu`, UV) vernachlässigbar klein?

Den Datensatz können Sie so herunterladen (Achtung, groß):

```
## Warning: One or more parsing issues,
## see `problems()` for details
```

Hinweise:

- Sie benötigen einen Computer, um diese Aufgabe zu lösen.
- Verwenden Sie die statistischen Methoden, die im Unterricht behandelt wurden.
- Verwenden Sie Ansätze aus der Bayes-Statistik zur Lösung dieser Aufgabe.
- Bei der Variable für Geschlecht können Sie sich auf Fälle begrenzen, die Männer und Frauen umfassen.
- Wandeln Sie die die Variable für Geschlecht in eine binäre Variable - also Werte mit 0 und 1 - um.

a. Ja
b. Nein
c. Die Daten sind nicht konkludent; es ist keine Entscheidung möglich.
d. Auf Basis der bereitgestellten Informationen ist keine Entscheidung möglich.

Lösung

Pakete laden:

Relevante Spalten auswählen:

Das sind die Variablen:

- Stress
- Neurotizismus
- Geschlecht

Deskriptive Statistiken zum Datensatz:

```
## # A tibble: 2 × 4
##   variable       n median   iqr
##   <chr>      <dbl>  <dbl> <dbl>
## 1 neu       108367   3.33  1.33
## 2 PSS10_avg 116097   2.6   1
```

```
## # A tibble: 4 × 2
##   Dem_gender                     n
##   <chr>                      <int>
## 1 Female                     90400
## 2 Male                       33126
## 3 Other/would rather not say  1474
## 4 <NA>                         306
```

Datensatz aufbereiten:

Check:

```
## # A tibble: 2 × 2
##   Female     n
##    <dbl> <int>
```

```
## 1      0 28371
## 2      1 78472
```

Check:

```
## # A tibble: 3 × 13
##   variable       n   min   max median
##   <chr>      <dbl> <dbl> <dbl>  <dbl>
## 1 Female    106843     0     1   1
## 2 neu       106843     1     6   3.33
## 3 PSS10_avg 106843     1     5   2.6
## # … with 8 more variables: q1 <dbl>,
## #   q3 <dbl>, iqr <dbl>, mad <dbl>,
## #   mean <dbl>, sd <dbl>, se <dbl>,
## #   ci <dbl>
```

Modell berechnen:

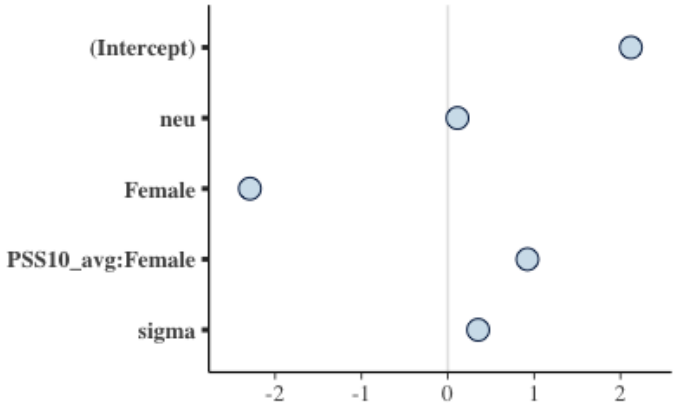
Modellkoeffizienten auslesen:

```
##      (Intercept)              neu
##             2.12             0.11
##           Female PSS10_avg:Female
##            -2.29             0.92
```

Posteriori-Verteilung auslesen:

```
##                    5.5% 94.5%
## (Intercept)        2.11  2.13
## neu                0.11  0.11
## Female            -2.30 -2.28
## PSS10_avg:Female   0.92  0.93
## sigma              0.35  0.35
```

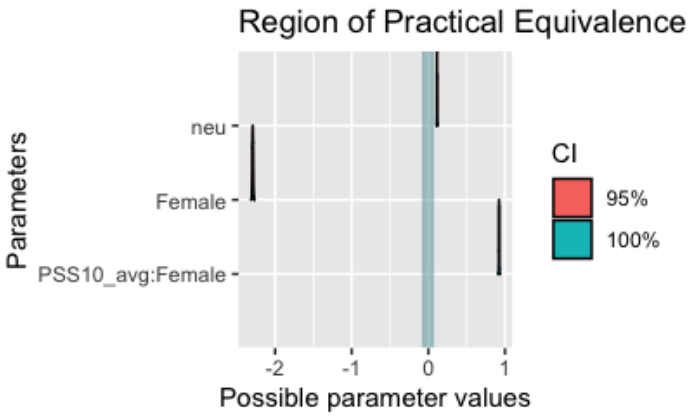
Posteriori-Verteilung plotten:



Rope berechnen:

```
## # Proportion of samples inside the ROPE [-0.07, 0.07]:
##
## Parameter        | inside ROPE
## -----------------------------
## (Intercept)      |      0.00 %
## neu              |      0.00 %
## Female           |      0.00 %
## PSS10_avg:Female |      0.00 %
```

Rope visulasieren:



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