Lösungen zu den Aufgaben

1. **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.

2. **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

3. **Aufgabe**

   Gegeben dem folgenden Modell, schreiben Sie die passende Form des Bayes-Theorem auf.

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

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

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

   Lösung

   Die allgemeine Form des Bayes-Theorem hatten wir so kennen gelernt:

   $$Pr(Hyp|Daten) = \frac{Pr(Daten|Hyp) \cdot Pr(Hyp)}{Pr(Daten)}$$

   $Pr(\mu, \sigma|h)$ gibt die Posteriorie-Wahrscheinlichkeit für ein bestimmte Hypothese an, z.B. für die Hypothse $\mu = 0$.

   $Pr(Daten|Hyp)$ ist der Likelihood unserer Daten gegeben der gerade untersuchten Hypothese.

   $Pr(Hyp)$ ist die Apriori-Wahrscheinlichkeit (das “Apriori-Gewicht”) der gerade untersuchten Hypothese.

   Der Zähler gibt die *unstandardisierte* Posteriori-Wahrscheinlichkeit der gerade untersuchten Hypothese an.

   Der Nenner ist nur ein *Normalisierungsfaktor*, der dafür sorgt, dass der ganze Bruch die *standardisierte* Posteriori-Wahrscheinlichkeit angibt.

   In diesem konkreten Fall untersuchen wir Hypothesen zu einem “Parameter-Pärchen”, $\mu\sigma$. Wir fragen also, wie wahrscheinlich es ist, einen gewissen Mittelwert μ und (gleichzeitig) eine gewisse Streuung σ aufzufinden.

Zum Beispiel könnten wir fragen: “Wie wahrscheinlich ist es, dass $\mu = 194$ und $\sigma = 12$?”. Bayes’ Theorem gibt uns die Wahrscheinlichkeit für diese Hypothese:

$$Pr(\mu, \sigma | h) = \frac{Pr(h|\mu, \sigma) \cdot Pr(\mu) \cdot Pr(\sigma)}{Pr(h)}$$

Hier ist zu beachten, dass die Apriori-Wahrscheinlichkeit auf *zwei* Termen besteht, $Pr(\mu)$ und $Pr(\sigma)$. Sind diese unabhängig, so kann man ihre Wahrscheinlichkeiten multiplizieren, um die gemeinsame Wahrscheinlichkeit zu erhalten, also die Wahrscheinlichkeit für ein bestimmten “Mu-Sigma-Pärchen”, etwa $\mu = 194, \sigma = 12$.

4. **Aufgabe**

   Gegeben dem folgenden Modell, simulieren Sie Daten aus der Prior-Verteilung (Priori-Prädiktiv-Verteilung).

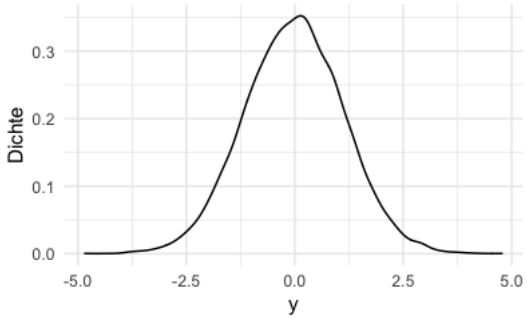
   Likelihood: $h_i \sim \mathcal{N}(\mu, \sigma)$

   Prior für μ: $\mu \sim \mathcal{N}(0, 1)$

   Prior für σ: $\sigma \sim \mathcal{U}(0, 10)$

   Lösung

```
library(tidyverse)

n <- 1e4

sim <- tibble(
  mu = rnorm(n = n),  # Default-Werte sind mean=0, sd = 1
  sigma = runif(n = n, 0, 1)) %>%
  mutate(
    y = rnorm(n = n, mean = mu, sd = sigma))

ggplot(sim, aes(x = y)) +
  geom_density() +
  labs(x = "y", y = "Dichte") +
  theme_minimal()
```



5. **Aufgabe**

   Gegeben dem folgenden Modell, geben Sie den Befehl mit `quap()` an, um die Posteriori-Verteilung zu berechnen.

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

   Prior für μ: $\mu \sim \mathcal{N}(0, 1)$

   Prior für σ: $\sigma \sim \mathcal{U}(0, 10)$

   Lösung

```
library(rethinking)

model_def <-
  alist(
    y ~ dnorm(mu, sigma),
    mu ~ dnorm(0, 1),
    sigma ~ dunif(0, 1)
  )

model <-
  quap(
    alist = model_def,
```

```
    daten = meine_Daten
  )
```

6. **Aufgabe**

Betrachten Sie den Datensatz zur Größe der !Kung:

```
library(tidyverse)
url_kung <- "https://raw.githubusercontent.com/rmcelreath/rethinking/master/data/Howell1.csv"
d <-
  read_delim(url_kung, delim = ";")  # Strichpunkt als Trennzeichen in der CSV-Datei
```

a. Untersuchen Sie mit Hilfe eines Diagramms, ob bzw. inwieweit sich die Größe der erwachsenen Personen normalverteilt.

b. Kennzahlen, die angegeben, inwieweit sich eine Größe normalverteilt, sind *Schiefe* und *Kurtosis*. Die Schiefe gibt an, wie symmetrische eine Verteilung ist.

Normalverteilungen sind symmetrisch und haben daher einen Wert von 0 für *Schiefe*. *Kurtosis* gibt die “Wölbung”, also wie “spitz” oder “plattgedrückt” eine Verteilung ist. Eine Normalverteilung hat eine Wert von 3 für Kurtosis.

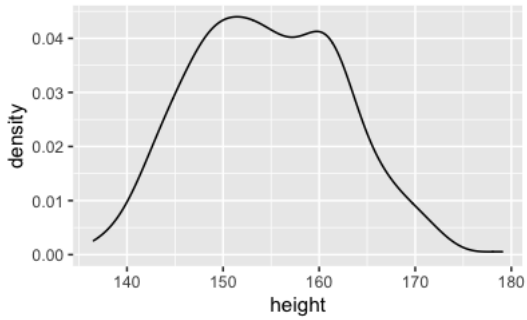
Entsprechende R-Funktionen finden Sie z.B. im Paket `moments`. iBerechnen Sie die beiden Kennzahlen für die Gruppe der Erwachsenen sowie aufgeteilt nach dem Geschlecht. Interpretieren Sie das Ergebnis.

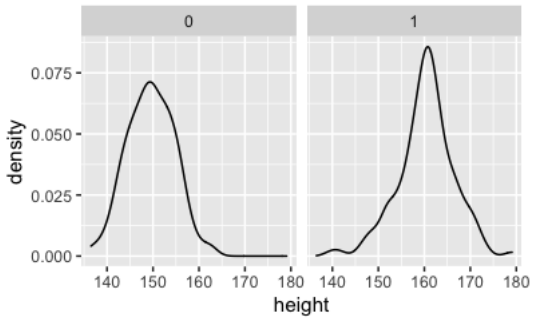
c. Diskutieren Sie, inwieweit man aus biologisch fundierten Sachverhalten (also *ontologisch*) eine Normalverteilung der Körpergröße annehmen kann.

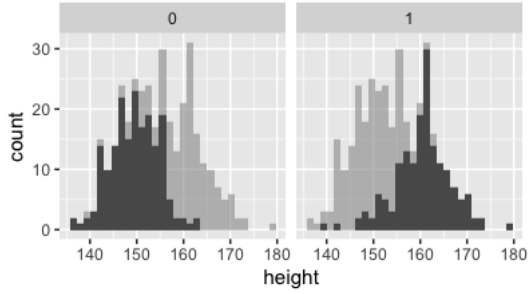
Lösung

a. Visuelle Prüfung der Normalverteilung

```
d2 <- d %>%
  filter(age >= 18)

d3 <- d2 %>%
  select(-male)

ggplot(d2, aes(x = height)) +
  geom_density()

ggplot(d2, aes(x = height )) +
  facet_wrap(~ male) +
    geom_density()

ggplot(d2, aes(x = height)) +
  facet_wrap(~ male) +
  geom_histogram(data = d3, fill = "grey60", alpha = .6) +
    geom_histogram() +
  labs(caption = "Grau hinterlegt ist das Histogramm für die Daten über beide Geschlechter")
```







rau hinterlegt ist das Histogramm für die Daten über beide Geschlechter

b. Schiefe und Kurtosis

```
library(moments)
d2 %>%
  summarise(skew = skewness(height),
            kurtosis = kurtosis(height))
```

skew	kurtosis
0.15	2.5

```
d2 %>%
  group_by(male) %>%
    summarise(skew = skewness(height),
              kurtosis = kurtosis(height))
```

male	skew	kurtosis
0	0.00	2.7
1	-0.33	4.0

c. Normalverteilung, Begründung

Es ist plausibel anzunehmen, dass der Phänotyp *Körpergröße* das Resultat des (kausalen) Einflusses vieler Gene ist, vieler Gene, die über einen vergleichbar starken Einfluss verfügen.

Eine besondere Situation stellt das X- bzw. Y-Chromosom dar, das Gene zum Geschlecht bereitstellt. Das Geschlecht ist ein einzelner Faktor, der (erfahrungsgemäß) einen relativ großen Einfluss auf die Körpergröße hat (in Anbetracht, dass vielleicht Tausende Gene additiv die Größe bestimmen). Insofern ist eine klarere Annäherung an die Normalverteilung zu erwarten, wenn man die Geschlechter einzeln betrachtet.

7. **Aufgabe**

Pupillendaten sind ein verbreiteter Analysegegenstand in Bereichen wie Psychologie, Marktforschung und Marketing.

Betrachten wir dazu ein R-Paket (zum Vorverbarbeitung, preprocessing) und einen Datensatz der Uni Münster.

```
library(PupilPre)
data("Pupildat")
d <-
  Pupildat %>%
  select(size = RIGHT_PUPIL_SIZE,
         time = TIMESTAMP) %>%
  mutate(size = size / 100) # in millimeter
```

Mit dem R-Paket `rstatix` kann man sich bequem typische Statistiken ausgeben lassen. Aber natürlich können Sie auch mit `summarise(mw = mean(size))` arbeiten.

```
library(rstatix)
d %>%
  get_summary_stats()
```

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
size	45343	1	25	8.2e+00	6.6e+00	11	3.9e+00	2.6e+00	1e+01	5.1e+00	0.02	0.05
time	46950	1443974	4062110	3.6e+06	1.9e+06	3821381	1.9e+06	5.1e+05	3e+06	9.9e+05	4557.21	8932.20

Wir verzichten hier auf eine Aufbereitung der Daten (was eigentlich nötig wäre, aber nicht Gegenstand dieser Übung ist). Stattdessen konzentrieren wir uns auf die Posteriori-Verteilung zur Pupillengröße.

Wir sind also interessiert an einem Modell zur Schätzung der (Verteilung der) Pupillengröße; die Posteriori-Verteilung bildet das ab.

a. Formulieren Sie ein passendes Modell.

b. Verteidigen Sie Ihre Modellspezifikation.

c. Simulieren Sie Daten aus der Priori-Verteilung. Kritisieren Sie die Wahl der Priori-Werte.

d. Berechnen Sie die Posteriori-Verteilung mit den Pupillendaten `d`. Geben Sie zentrale Statistiken an.

e. Geben Sie ein 90%-Intervall für die mittlere Pupillengröße an auf Basis der Posteriori-Verteilung.

Lösung

a. Modelldefinition

$$
\begin{aligned}
s_i &\sim \mathcal{N}(\mu, \sigma) && \mid \text{s wie size} \\
\mu &\sim \mathcal{N}(10, 5) \\
\sigma &\sim \mathcal{U}(0, 20)
\end{aligned}
$$

b. Begründung der Modellspezifikation

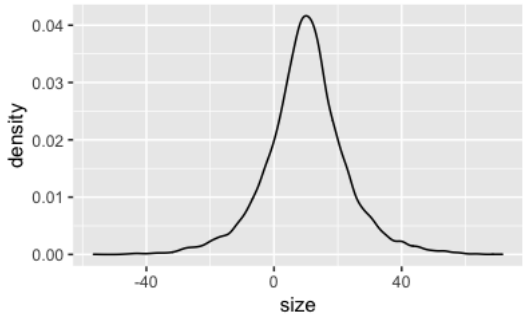
s_i: Pupillengrößen sind normalverteilt, da viele Gene additiv auf die Größe hin zusammenwirken

μ: Da wir nicht viel wissen über die mittlere Pupillengröße, entscheiden wir uns für Normalverteilung für diesen Parameter, da dies keine weiteren Annahmen (außer dass Mittelwert und Streuung endlich sind) hinzufügt. Ein Modell mit wenig Annahmen nennt man “sparsam” oder konservativ. Es ist wünschenswert, dass Modelle mit so wenig wie möglich Annahmen auskommt (aber so vielen wie nötig).

σ: Die Streuung muss positiv sein, daher kommt keine Normalverteilung in Frage. Da wir (noch) keine passenden Verteilungen kennen außer der Gleichverteilung, entscheiden wir uns für eine vage Gleichverteilung. Die große Stichprobe wird den Priori-Wert vermutlich überstimmen.

c. Priori-Prädiktiv-Verteilung

```
n <- 1e4
sim_prior_pred <-
  tibble(
    mu = rnorm(n, mean = 10, sd = 5),
    sigma = runif(n, min = 0, max = 20),
    size = rnorm(n, mu, sigma)
  )

sim_prior_pred %>%
  ggplot(aes(x = size)) +
  geom_density()
```



Da es viele negative Pupillengröße-Werte gibt, sieht man deutlich, dass das Modell nicht gut spezifiziert ist. So könnte kleinere Streuungswerte zu einem realistischeren Modell führen. Oder man verwendet Verteilungen, die rein positiv sind (hier nicht weiter ausgeführt).

d. Berechnen Sie die Posteriori-Verteilung.

Die Modelle wie `quap()` tun sich leichter, wenn man nur die relevanten Daten, ohne fehlende Werte und schon schön fertig vorverarbeitet, zur Analyse in die Modellberechnung gibt:

```
d3 <-
  d %>%
  select(size) %>%
  drop_na()
```

Die Posteriori-Verteilung kann man mit `rstanarm` d.h. mit `stan_glm()` berechnen:

```
library(rstanarm)
m_pupil <- stan_glm(size ~ 1,
                    data = d3,
                    refresh = 0)

summary(m_pupil)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      size ~ 1
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 45343
##  predictors:   1
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept) 10.0    0.0 10.0  10.0  10.0
## sigma        5.1    0.0  5.1   5.1   5.1
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 10.0    0.0 10.0  10.0  10.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for details see help('summary.stanreg'))
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  2000
## sigma         0.0  1.0  3015
## mean_PPD      0.0  1.0  2625
## log-posterior 0.0  1.0  1587
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample size, and Rhat is the potential s
```

Oder man kann alternativ `quap()` aus `rethinking` verwenden:

```
library(rethinking)
m_pupil2 <-
  quap(
    alist(
      size ~ dnorm(mu, sigma),
      mu ~ dnorm(10, 5),
      sigma ~ dunif(0, 20)
    ),
```

```
    data = d3
  )

precis(m_pupil2) %>%
  gt()
```

mean	sd	5.5%	94.5%
10.0	0.024	10.0	10.0
5.1	0.017	5.1	5.1

Ich werde mich künftig auf `stan_glm()` konzentrieren, da die Syntax einfacher ist.

Sie müssen die Syntax aus `rethinking` NICHT können, es reicht die Syntax von `rstanarm`.

Hier die ersten paar Zeilen des Post-Verteilung:

(Intercept)	sigma
10.03	5.08
10.00	5.07
9.97	5.10
10.00	5.12
9.99	5.13

e. Geben Sie ein 90%-Intervall für die mittlere Pupillengröße an auf Basis der Posteriori-Verteilung.

```
posterior_interval(m_pupil)
```

	5%	95%
(Intercept)	10.0	10.1
sigma	5.1	5.1

Alternativ mit `rethinking`:

Ziehen wir Stichproben aus der Posteriori-Verteilung:

```
post_m_pupil <-
  extract.samples(m_pupil2, n = 1e4)
```

Und dann erstellen wir ein 89%-PI:

```
library(rethinking)

post_m_pupil %>%
  summarise(PI(mu))
```

PI(mu)
10
10

8. **Aufgabe**

Intelligenz von Studentis

Eine Psychologin möchte die Intelligenz von Studentis bestimmen: Was ist wohl der Mittelwert? Wie schlau sind die schlausten 10%? Von wo bis wo geht das mittlere 90%-Intervall von IQ-Werten? Natürlich ist ihr klar, dass es nicht reicht, einen Mittelwert zu schätzen. Nein, sie will alles, sprich: die Posteriori-Verteilung.

Zuerst überlegt sie sich die Prioris: “Was ist meine Einschätzung zur Intelligenz von Studentis?”. Dazu liest sie alle verfügbare Literatur, beurteilt die methodische Qualität jeder einzelnen Studie und spricht mit den Expertis. Auf dieser Basis kommt sie zu folgenden Prioris:

$$\mu \sim \mathcal{N}(115, 5)$$

Ein paar Überlegungen, die unsere Psychologin dazu hatte: Die Studentis sind im Mittel schlauer als die Normalbevölkerung. Um ein Gefühl für die Verteilungsfunktion vom IQ zu bekommen, nutzt sie folgenden R-Befehl:

```
pnorm(q = 115, mean = 100, sd = 15)
```

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

Dieser Befehl gibt ihr an, welcher Prozentsatz der allgemeinen Bevölkerung (die Wahrscheinlichkeitsmasse) nicht schlauer ist als 115.

Dann versucht sie ein Gefühl für die Streuung zu bekommen, folgender R-Befehl hilft ihr:

```
q_iq <- 50
rate_lambda <- 0.1
pexp(q = q_iq, rate = rate_lambda)
```

```
## [1] 0.99
```

Ah! Nimmt man an, dass Sigma exponentialverteilt ist mit einer Rate von 0.1, dass sind etwa 99 Prozent der Leute nicht mehr als `q_iq` IQ-Punkte vom Mittelwert μ entfernt. Das deckt sich mit ihren Informationen aus der Literatur.

Damit sind die Priors spezifiziert.

a. Geben Sie die Priors an.
b. Simulieren Sie die Prior-Prädiktiv-Verteilung dazu.
c. Befragen Sie die Prior-Prädiktiv-Verteilung mit geeigneten Fragen Ihrer Wahl.

Lösung

a. Geben Sie die Priors an.

$$\mu \sim \mathcal{N}(115, 5)$$

$$\sigma \sim \mathcal{E}(0.1)$$

b. Simulieren Sie die Prior-Prädiktiv-Verteilung dazu.

Ziehen wir Zufallszahlen entsprechend der Priori-Werte:

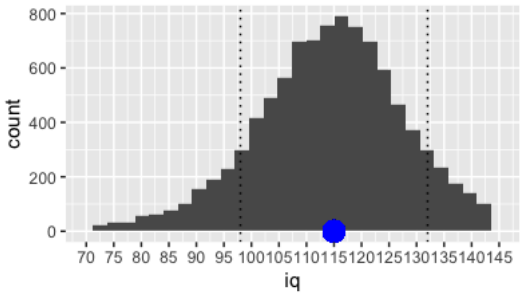
```
library(tidyverse)
n <- 1e4

sim <-
  tibble(
  sample_mu  =
    rnorm(n,
          mean = 115,
          sd   = 10),
  sample_sigma =
    rexp(n,
         rate = 0.1)) %>%
  mutate(iq  =
           rnorm(n,
            mean = sample_mu,
            sd   = sample_sigma))
```

Was ist wohl der Mittelwert und die SD dieser Priori-Prädiktiv-Verteilung?

```
height_sim_sd <-
  sd(sim$iq) %>% round()
height_sim_mean <-
  mean(sim$iq) %>% round()
```

Und jetzt plotten wir diese Verteilung:

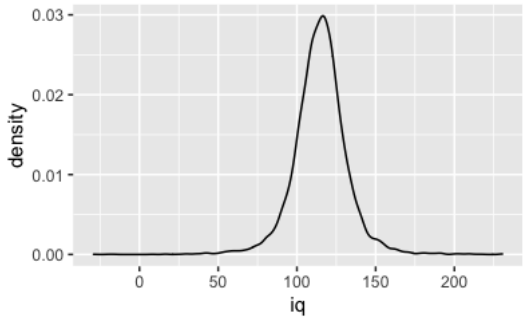
```
sim %>%
  ggplot() +
  aes(x = iq) +
  geom_histogram() +
  geom_point(y = 0, x = height_sim_mean, size = 5,
             color = "blue", alpha = .5) +
  geom_vline(xintercept = c(height_sim_mean + height_sim_sd,
                            height_sim_mean - height_sim_sd),
             linetype = "dotted") +
  labs(caption = "Der blaue Punkt zeigt den Mittelwert; die gepunkteten Linien MD±SD") +
  scale_x_continuous(limits = c(70, 145),
                     breaks = seq(70, 145, by = 5))
```



Der blaue Punkt zeigt den Mittelwert; die gepunkteten Linien MD±SD

Oder vielleicht besser als Dichte-Diagramm, das zeigt das “Big Picture” vielleicht besser:

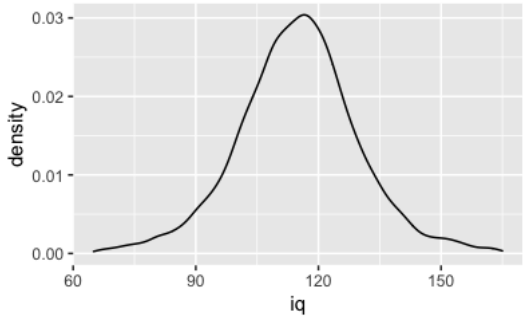
```
sim %>%
  ggplot() +
  aes(x = iq) +
    geom_density()
```



Hm, etwas randlastig die Verteilung.

Zoomen wir etwas mehr rein:

```
sim %>%
  ggplot() +
  aes(x = iq) +
  geom_density() +
  scale_x_continuous(limits = c(65, 165))
```



c. Befragen Sie die Prior-Prädiktiv-Verteilung mit geeigneten Fragen Ihrer Wahl.

Was ist der Mittelwert und die SD und die üblichen deskriptiven Kennwerte?

```
library(rstatix)

sim %>%
  select(iq) %>%
  get_summary_stats()
```

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
iq	10000	-29	230	115	106	124	18	13	115	17	0.17	0.34

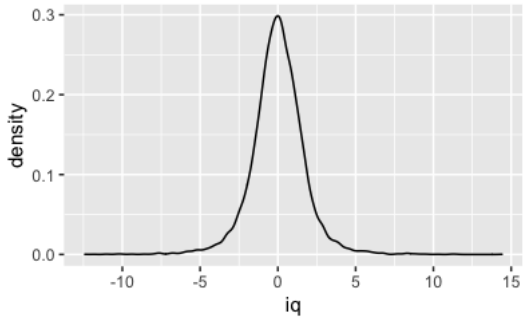
In welchem Bereich liegen die mittleren 95% der IQ-Werte?

```
sim %>%
  summarise(pi_95 = quantile(iq, probs = c(0.025,
                                           0.975)))
```

pi_95
81
149

Alternativ könnten wir in z-transformierten Daten denken:

```
sim2 <-
  tibble(
  sample_mu  =
    rnorm(n,
          mean = 0,
          sd   = 1),
  sample_sigma =
    rexp(n,
         rate = 1)) %>%
  mutate(iq  =
           rnorm(n,
            mean = sample_mu,
            sd   = sample_sigma))

sim2 %>%
  ggplot() +
  aes(x = iq) +
  geom_density()
```



9. **Aufgabe**

In diesem Diagramm sehen Sie etwas Nomenklatur für eine Verteilung: Gipfel (Peak), Schultern (shoulders) und Ränder (tails). Bitte klicken Sie den Link, um sich das Diagramm anzuschauen.

Taleb, N. N. (2019). *The statistical consequences of fat tails, papers and commentaries*. https://nassimtaleb.org/2020/01/final-version-fat-tails/

Zwar sind viele Daten in der Welt normalverteilt, aber längst nicht alle. In jüngerer Zeit sind sog. “Fat Tails” in die Aufmerksamkeit gerückt. Das sind Variablen, bei denen Werte in den Rändern (tails) wahrscheinlicher sind als bei einer Normalverteilung; ein Beispiel für eine Fat-Tail-Verteilung ist die t-Verteilung mit 1 Freiheitsgrad. Sie müssen diese Verteilung nicht weiter kennen, teilung handelt.

Recherchieren Sie (Fach-)Artikel, die argumentieren, dass ein bestimmtes Phänomen Fat-Tails zeigt!

Lösung

- Kriege
- Pandemien
- Erfolg auf der Singlebörse Tinder
- Kapitelmarkt