

Gauss-Modelle

QM2, Thema 4

AWM, HS Ansbach

Gliederung

1. Teil 1: Verteilungen
2. Teil 2: Wie groß sind die !Kung San?
3. Hinweise
4. Literatur

Verteilungen

Häufigkeitsverteilung

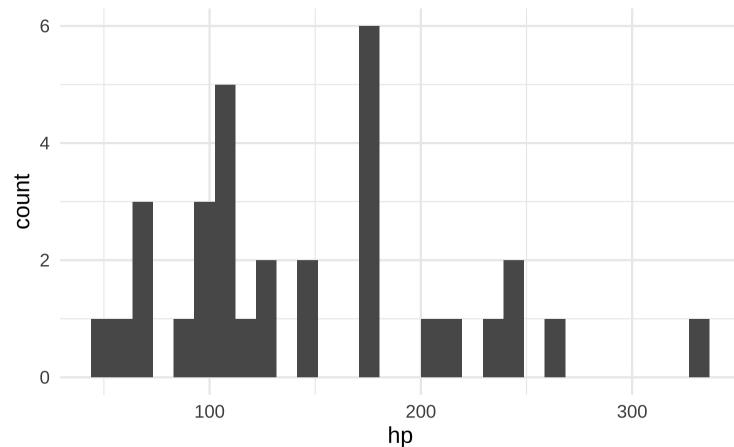
Die Verteilung eines *diskreten* Merkmals X mit k Ausprägungen zeigt, wie häufig die einzelnen Ausprägungen sind.

```
data(mtcars)
mtcars %>%
  count(cyl)
```



```
##   cyl  n
## 1   4 11
## 2   6  7
## 3   8 14
```

Ein *stetiges* Merkmal lässt sich durch Klassenbildung diskretisieren:

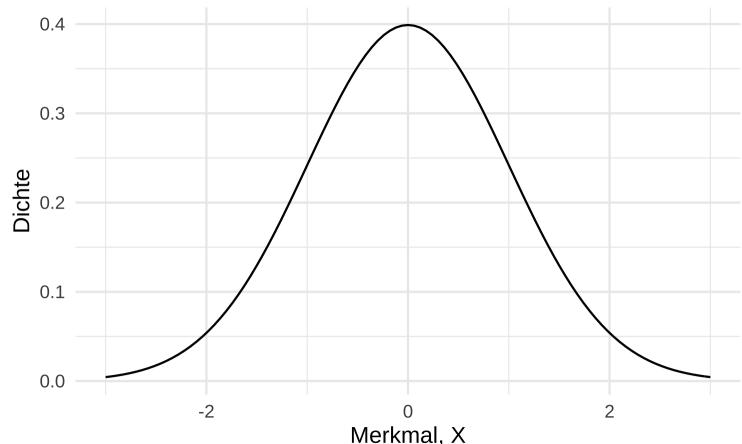


Wahrscheinlichkeitsverteilung

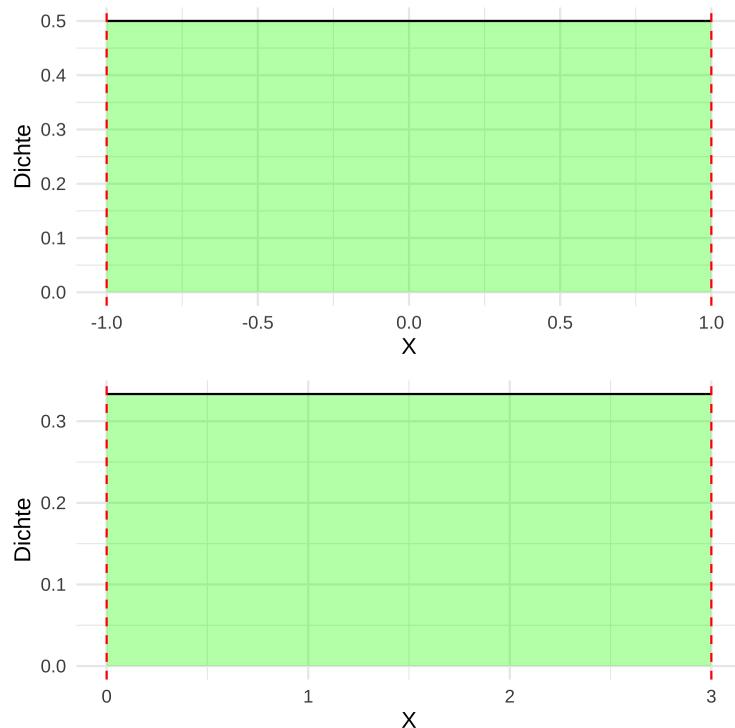
Eine *diskrete* Wahrscheinlichkeitsverteilung des Merkmals X ordnet jeder der k Ausprägungen $X = x$ eine Wahrscheinlichkeit p zu. So hat die Variable *Geschlecht eines Babies* die beiden Ausprägungen *Mädchen* und *Junge* mit den Wahrscheinlichkeiten $p_M = 51.2\%$ bzw. $p_J = 48.8\%$ ([Gelman, Hill, and Vehtari, 2021](#)).

Bei *stetigen* Merkmalen X geht man von unendlich vielen Ausprägungen aus; die Wahrscheinlichkeit einer bestimmten Ausprägung ist (praktisch) Null: $p(X = x_j) = 0$, $j = 1, \dots, k$. Daher gibt man stattdessen die *Dichte* der Wahrscheinlichkeit an: Das ist die Wahrscheinlichkeit(smasse) pro eine Einheit von X .

Beispiele für Wahrscheinlichkeitsdichte



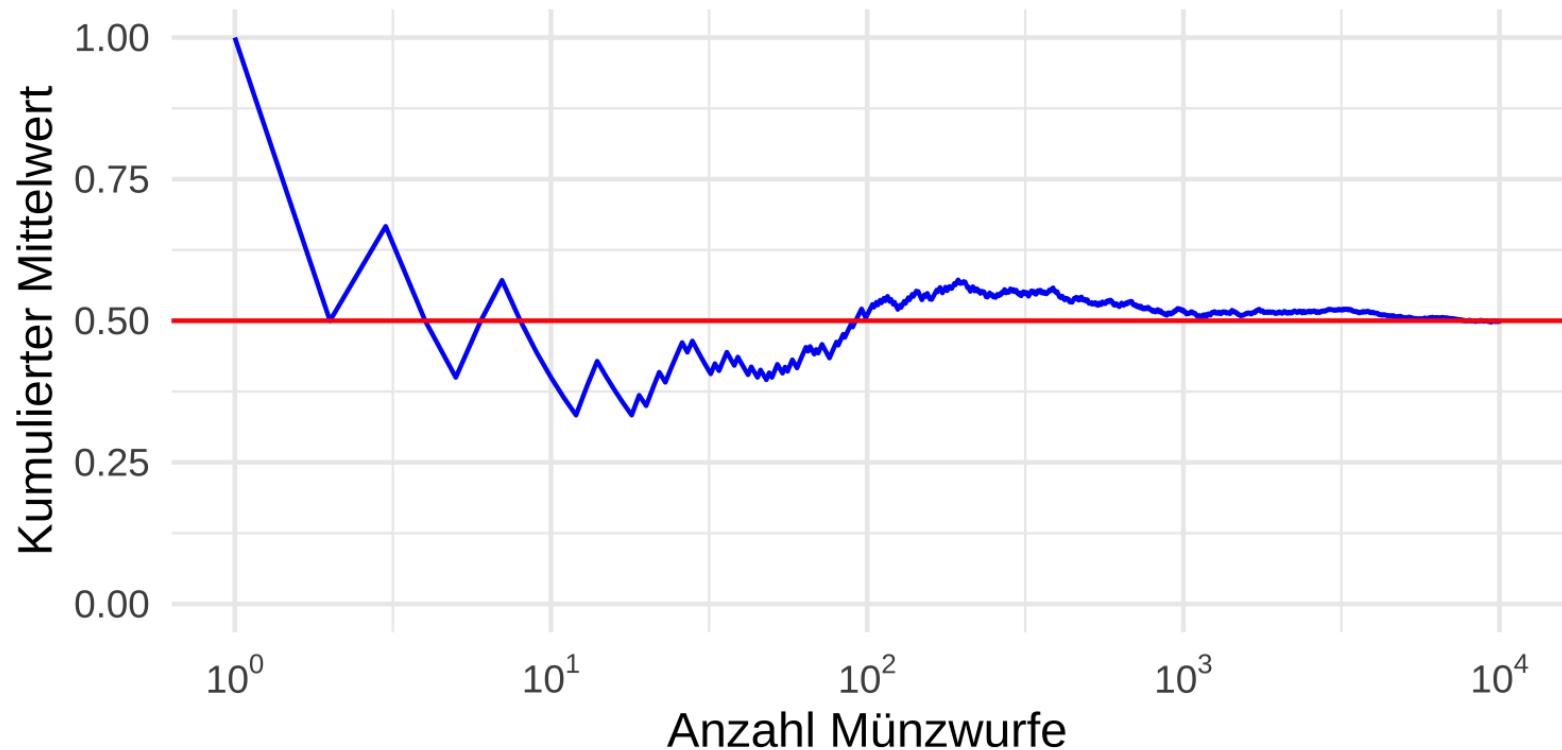
Bei $X = 0$ hat eine Einheit von X die Wahrscheinlichkeitsmasse von 40%.



Bei $X = 0$ hat eine Einheit von X die Wahrscheinlichkeitsmasse von 50% bzw. 33%.

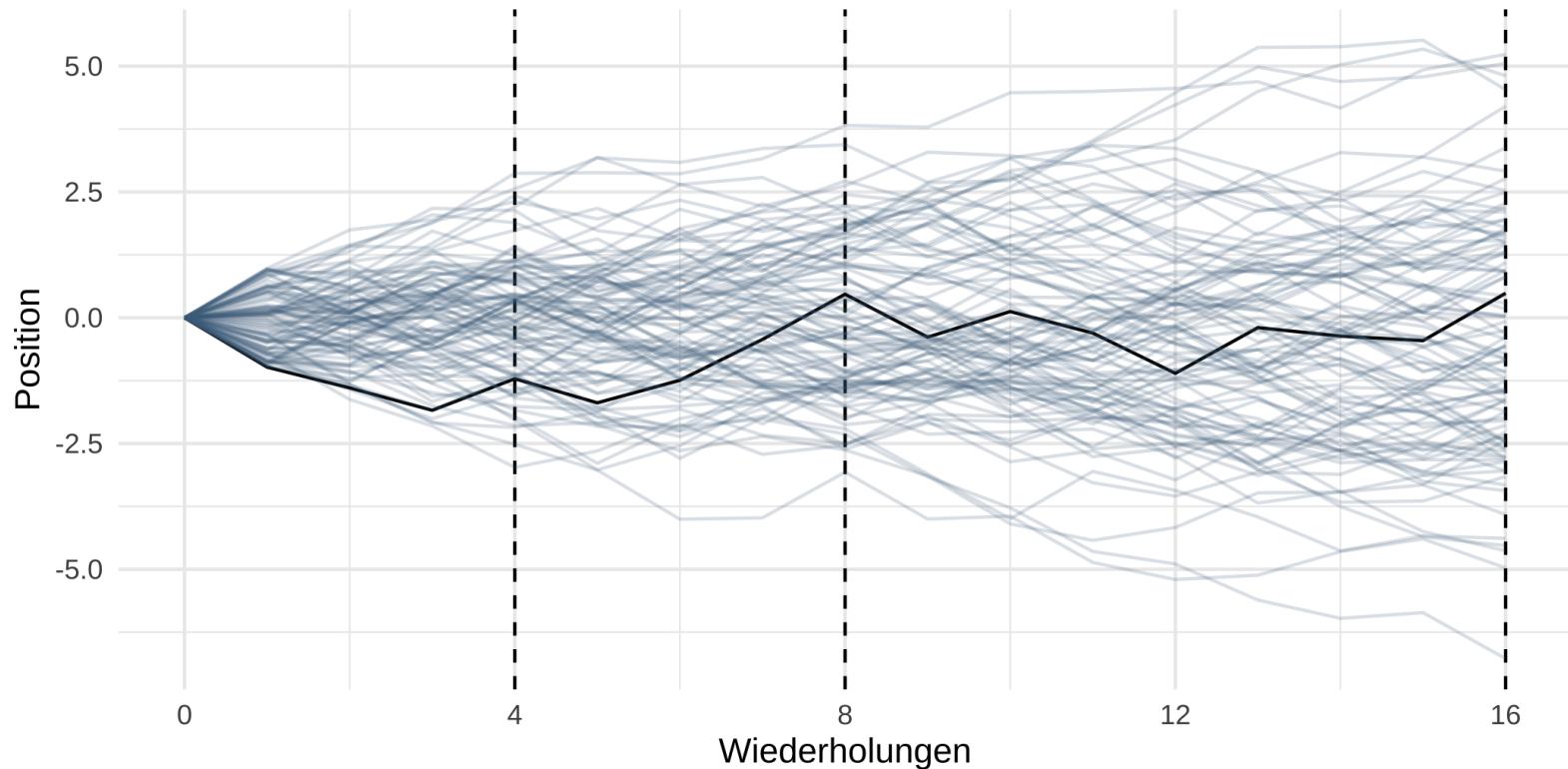
Gesetz der großen Zahl

Zieht man (zufällig) immer mehr Werte aus einer Verteilung (mit endlichem Mittelwert), nähert sich der Mittelwert der Stichprobe immer mehr mit dem Mittelwert (oft als Erwartungswert bezeichnet) der Verteilung an ([Taleb, 2019](#))



Normal auf dem Fußballfeld

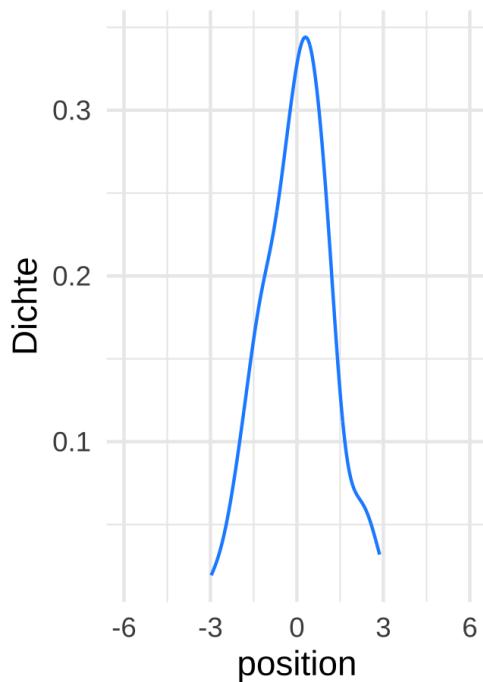
Sie und 1000 Ihrer besten Freunde stehen auf der Mittellinie eines Fußballfelds (eng). Auf Kommando werfen alle jeweils eine Münze; bei Kopf geht man einen Schritt nach links, bei Zahl nach rechts. Das wird 16 Mal wiederholt. Wie wird die Verteilung der Positionen wohl aussehen?



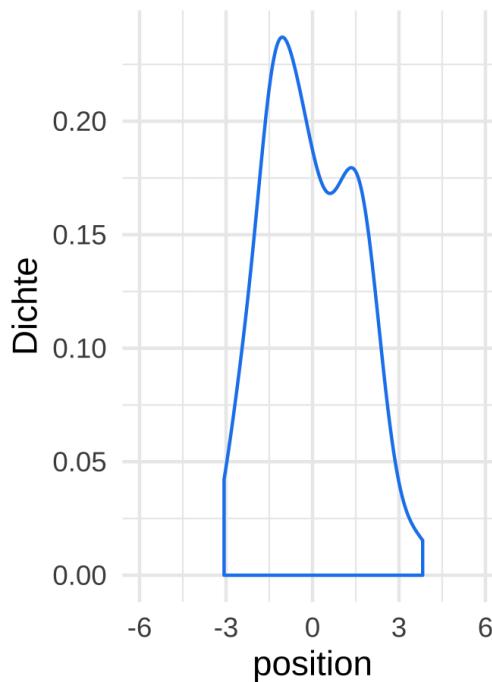
Normal durch Addieren

Die Summe vieler (gleich starker) Zufallswerte (aus der gleichen Verteilung) erzeugt eine Normalverteilung; egal aus welcher Verteilung die Zufallswerte kommen (Zentraler Grenzwertsatz).

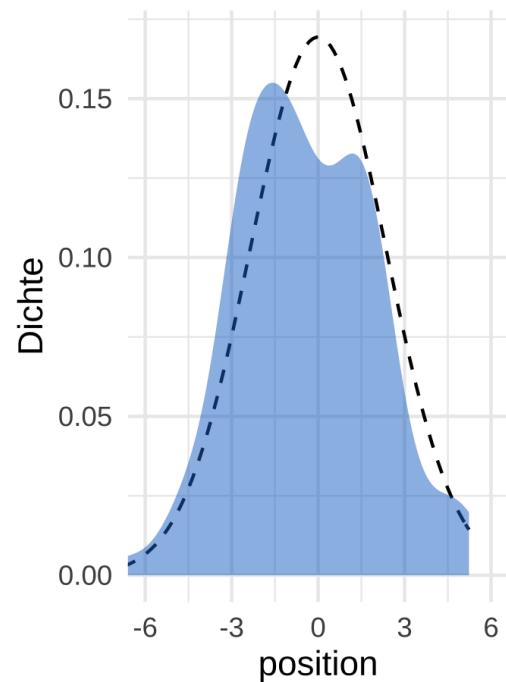
4 Wiederholungen



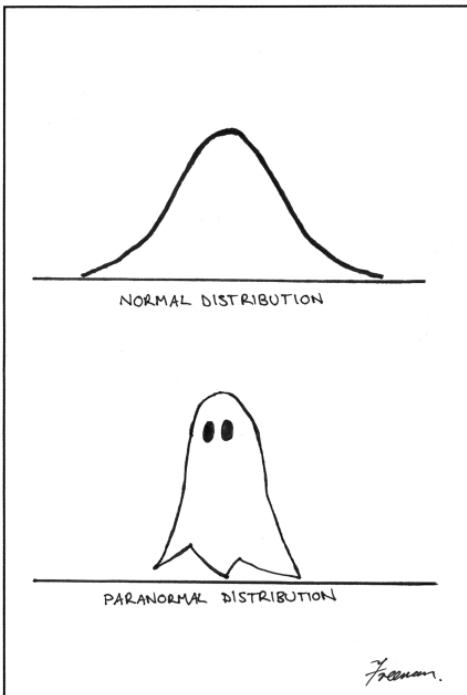
8 Wiederholungen



16 Wiederholungen

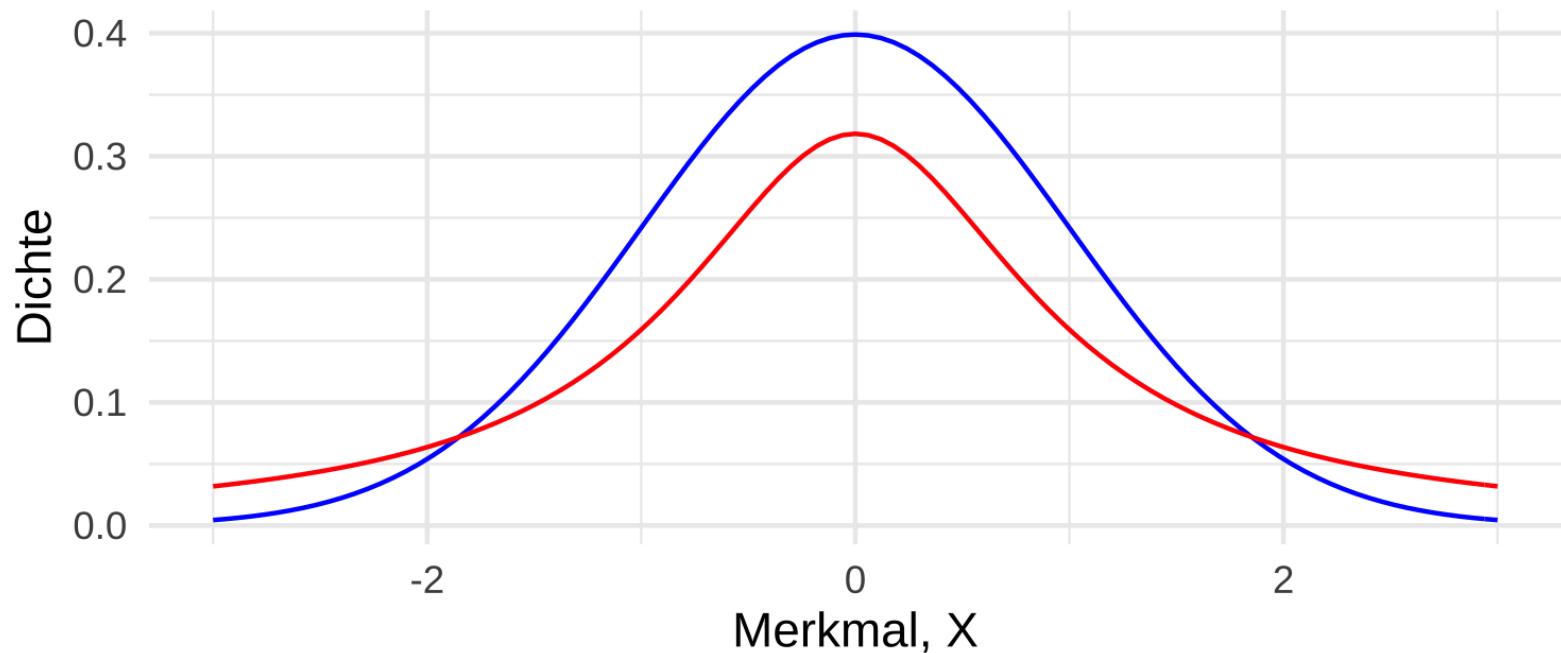


Nicht verwechseln



(Freeman, 2006)

Normalverteilung vs. randlastige Verteilungen



Blau: Normalverteilung
Rot: randlastige Verteilung (t-Verteilung mit $df=1$)

Bei randlastigen Verteilungen ("fat tails") kommen Extremereignisse viel häufiger vor als bei Normalverteilungen. Deshalb ist es wichtig sein, zu wissen, ob eine Normalverteilung oder eine randlastige Verteilung vorliegt. Viele statistische Methoden sind nicht zuverlässig bei (stark) randlastigen Methoden ([Taleb, 2019](#))

Beispiele für Normal- und randlastige Verteilungen

Normal verteilt

- Größe
- Münzwürfe
- Gewicht
- IQ
- Blutdruck
- Ausschuss einer Maschine

Randlastig verteilt

- Vermögen
- Verkaufte Bücher
- Ruhm
- Aktienkurse
- Erdbeben
- Pandemien
- Kriege
- Erfolg auf Tinder
- Meteoritengröße
- Stadtgrößen

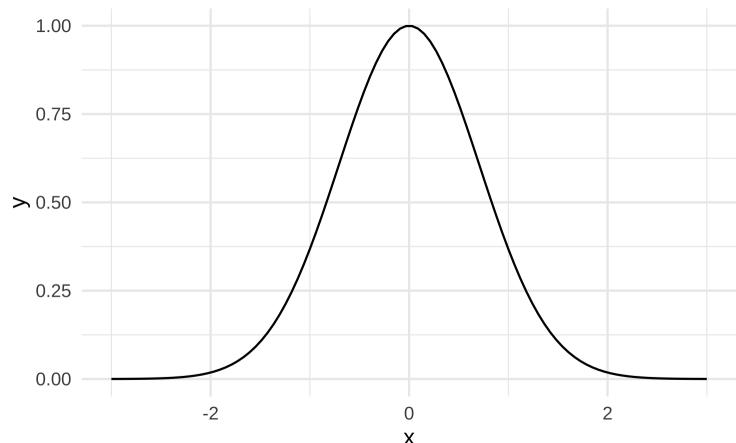
Formel der Normalverteilung

Vereinfacht ausgedrückt lässt die Normalverteilung \mathcal{N} durch Exponenzieren einer Quadratfunktion beschreiben:

$$\mathcal{N} \propto e^{-x^2}$$

mit $e = 2.71\dots$, der Eulerschen Zahl.

```
d <-  
  tibble(  
    x = seq(-3, 3,  
            length.out = 100),  
    y = exp(-x^2)  
)  
  
d %>%  
  ggplot() +  
  aes(x = x, y = y) +  
  geom_line()
```



Die Normalverteilung wird auch *Gauss*-Verteilung oder *Glockenkurve* genannt.



Normalverteilung als konservative Wahl



Uni Greifswald, Public domain, via
Wikimedia Commons

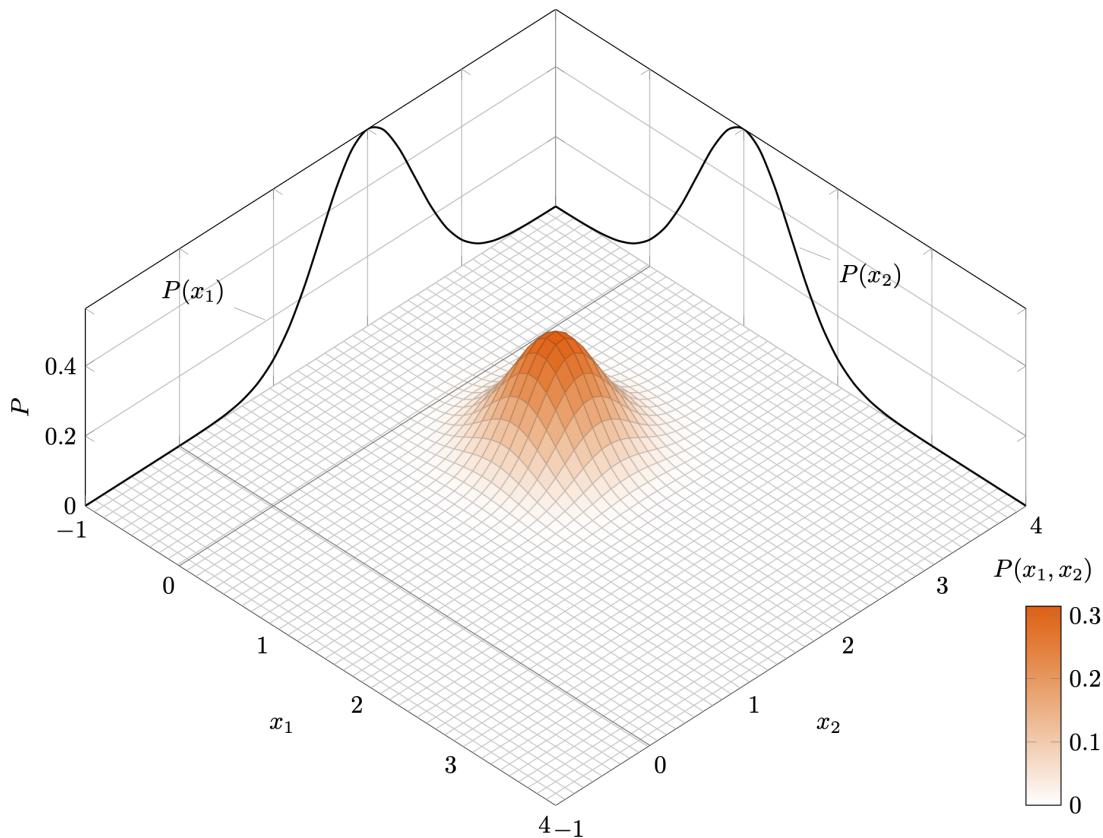
Ontologische Begründung

- Wirken viele, gleichstarke Einflüsse additiv zusammen, entsteht eine Normalverteilung ([McElreath, 2020](#)), Kap. 4.1.4.

Epistemologische Begründung

- Wenn wir nur wissen, dass eine Variable über einen endlichen Mittelwert und eine endliche Varianz verfügt und wir keine weiteren Annahmen treffen bzw. über kein weiteres Vorwissen verfügen, dann ist die Normalverteilung die plausibelste Verteilung (maximale Entropie) ([McElreath, 2020](#)), Kap. 7 und 10.

Zweidimensionale Normalverteilung, unkorreliert



Quelle

Vgl. auch dieses Diagramm

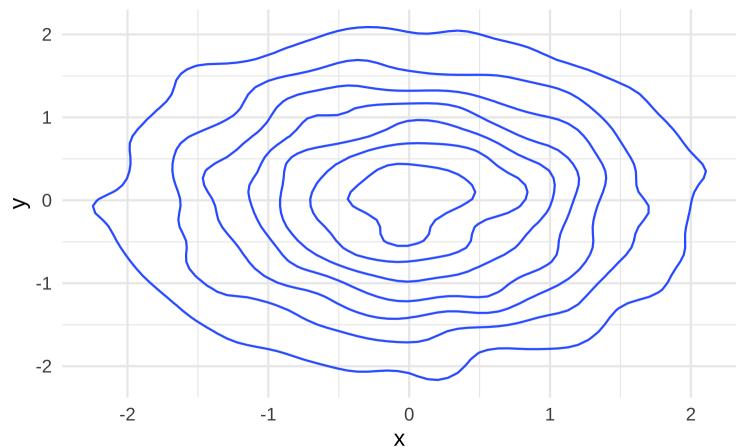
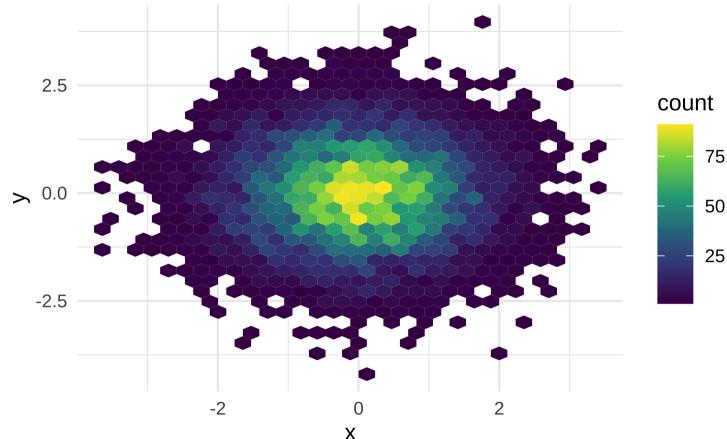
2D-Normalverteilung mit R, unkorreliert

$$r(X, Y) = 0$$

```
d1 <-  
  tibble(  
    x=rnorm(1e4),  
    y=rnorm(1e4)  
)  
  
ggplot(d1) +  
  aes(x, y) +  
  geom_hex()  
  
ggplot(d1) +  
  aes(x, y) +  
  geom_density2d()
```

[ggplot-Referenz](#), [Quellcode](#)

Mit `scale_fill_continuous(type = "viridis")` kann man die Farbpalette der Füllfarbe ändern.



2D-Normalverteilung mit R, korreliert, $r=0.7$

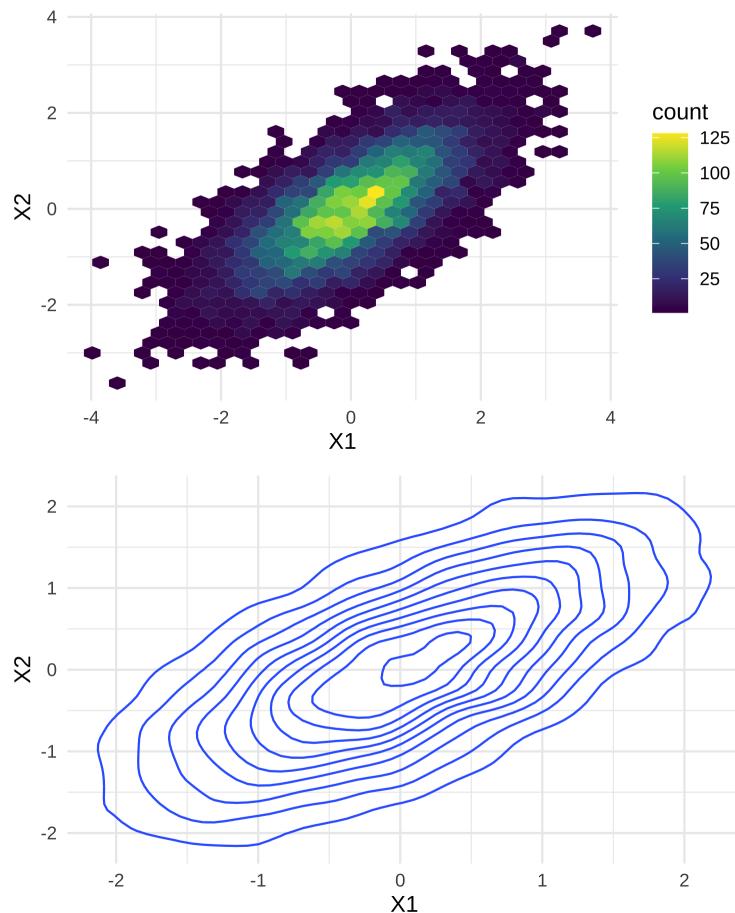
Die ersten paar Zeilen der Daten:

| X1 | X2 |
|-------|-------|
| -0.10 | -0.26 |
| -0.19 | 0.81 |
| -1.45 | -1.33 |

Berechnen wir die Korrelation r :

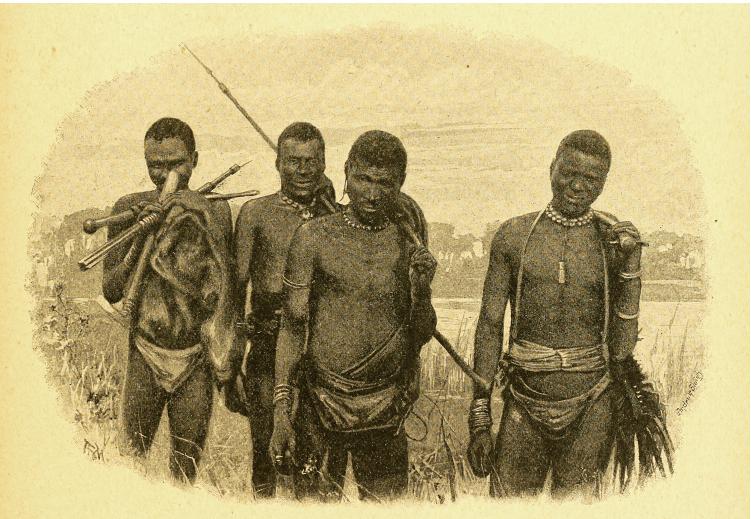
```
d2 %>%
  summarise(
    r = cor(X1,X2),
    n = n()
  )
```

| r | n |
|------|-----------|
| 0.70 | 10,000.00 |

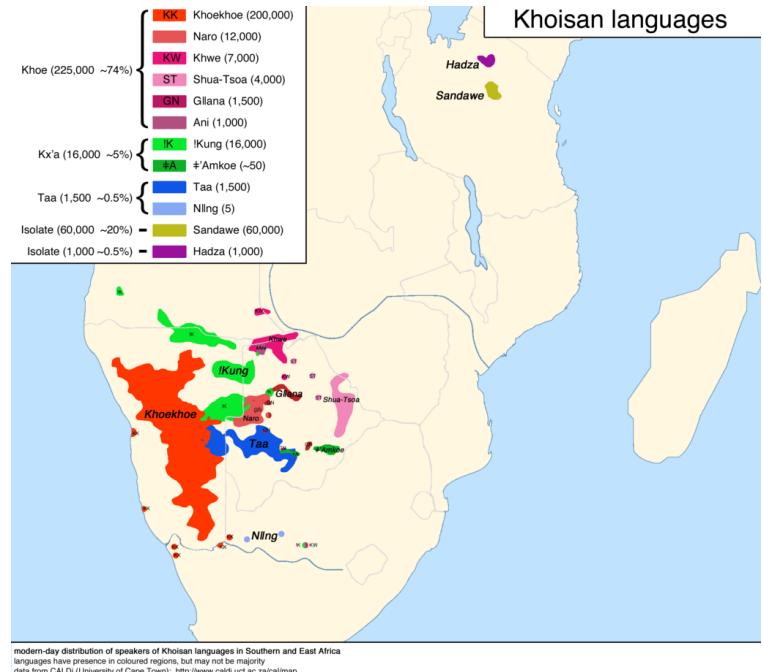


Wie groß sind die !Kung San?

!Kung San



Quelle Internet Archive Book Images, No restrictions, via Wikimedia Commons



By Andrewwik.0 - Own work, CC BY-SA 4.0, Quelle



Winfried Bruenken (Amrum), CC BY-SA 2.5 <https://creativecommons.org/licenses/by-sa/2.5>, via Wikimedia Commons

!Kung Data

```
library(rethinking)
data(Howell1)
d <- Howell1
```

```
d %>%
  head(n=3) %>%
  gt()
```

Alternativ kann man die Daten [hier](#) herunterladen.

Wir interessieren uns für die Größe der erwachsenen !Kung, $N = 352$:

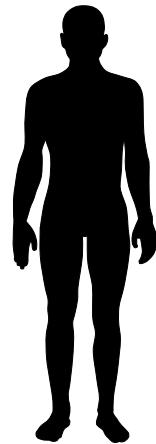
```
d2 <-
  d %>%
  filter(age >= 18)
```

| height | weight | age | male |
|---------|----------|-----|------|
| 151.765 | 47.82561 | 63 | 1 |
| 139.700 | 36.48581 | 63 | 0 |
| 136.525 | 31.86484 | 65 | 0 |

```
##               mean           sd      5.5%     94.5%
## height  138.2635963 27.6024476 81.108550 165.73500
## weight   35.6106176 14.7191782  9.360721 54.50289
## age     29.3443934 20.7468882  1.000000 66.13500
## male    0.4724265  0.4996986  0.000000  1.00000
```

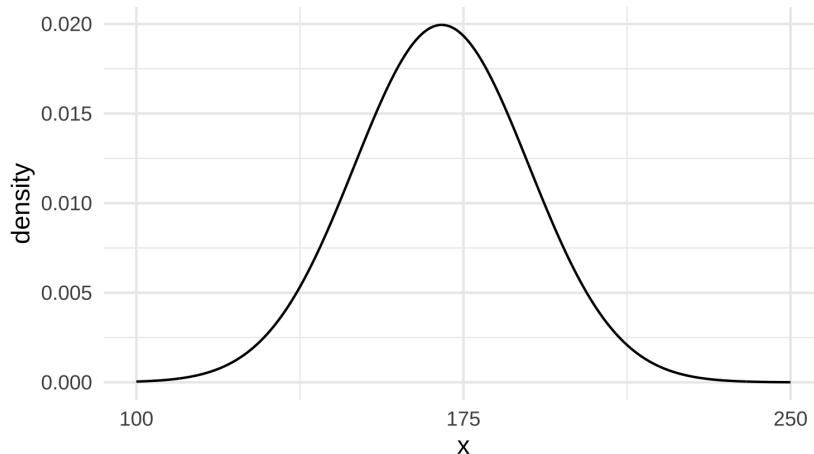


Wir gehen apriori von normalverteilter Größe aus



$$\mu \sim \mathcal{N}(178, 20)$$

$$\text{mu} \sim \text{dnorm}(178, 20)$$



Unser Gauss-Modell der !Kung

Wir nehmen an, dass μ und h_i normalverteilt, σ gleichverteilt sind:

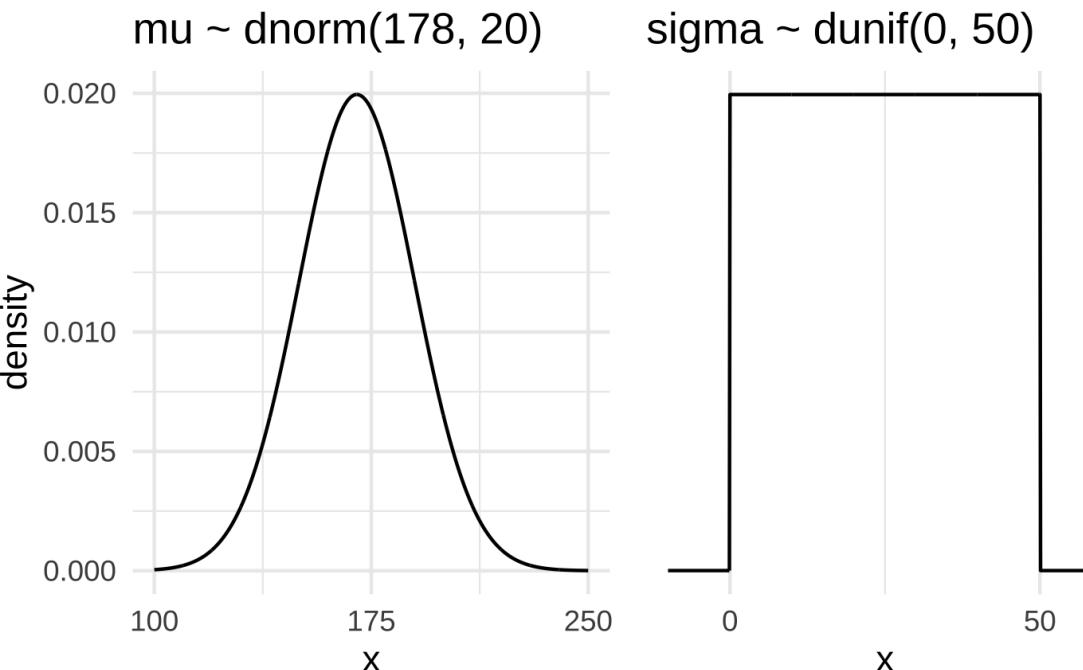
$$h_i \sim \mathcal{N}(\mu, \sigma)$$

$$\mu \sim \mathcal{N}(178, 20)$$

$$\sigma \sim \mathcal{U}(0, 50)$$

$$95\%KI(\mu) :$$

$$178 \pm 40$$



Welche Beobachtungen sind auf Basis unseres Modells zu erwarten?

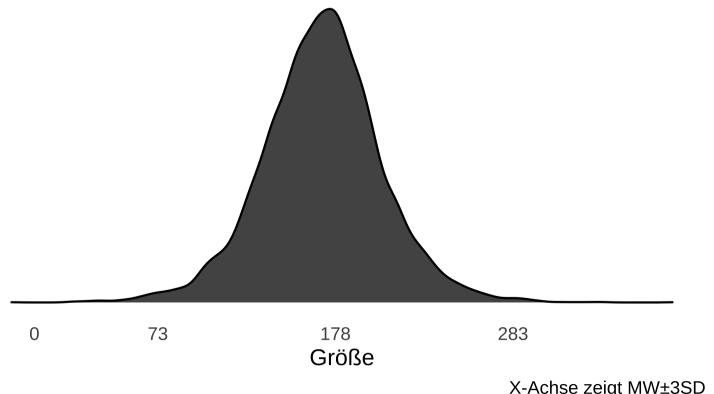
```
n <- 1e4

sim <-
  tibble(
    sample_mu      =
      rnorm(n,
            mean = 170,
            sd   = 20),
    sample_sigma =
      runif(n,
            min = 0,
            max = 50)) %>%
  mutate(
    height =
      rnorm(n,
            mean = sample_mu,
            sd = sample_sigma))
```

💡 Was denkt der Golem apriori von der Größe der !Kung?

✍️ Ziehen wir mal ein paar Stichproben auf Basis des Modells. voilà:

$\text{height} \sim \text{dnorm}(\mu, \sigma)$

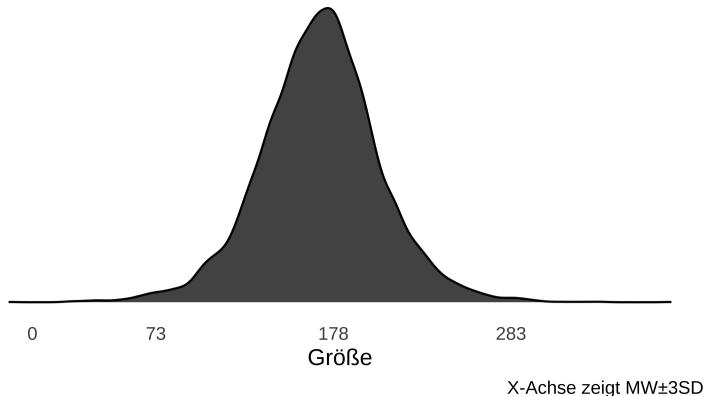


Priori-Werte prüfen mit der Die Priori-Prädiktiv-Verteilung

- Die Priori-Prädiktiv-Verteilung simuliert Beobachtungen (nur) auf Basis der Priori-Annahmen: $h_i \sim \mathcal{N}(\mu, \sigma)$, $\mu \sim \mathcal{N}(178, 20)$, $\sigma \sim \mathcal{U}(0, 50)$
- Priori-Prädiktiv-Verteilungen sind praktisch, um zu prüfen, ob die Priori-Werte vernünftig sind

Die Priori-Prädiktiv-Verteilung zeigt, dass unsere Priori-Werte ziemlich vage sind, also einen zu breiten Bereich an Größenwerten zulassen:

height ~ dnorm(mu, sigma)



Anteil $h_i < 100$:

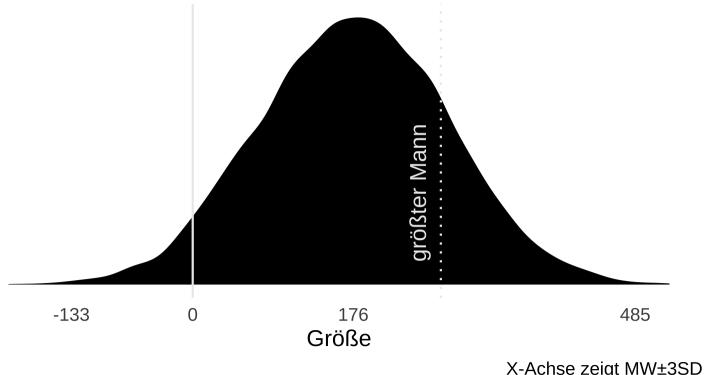
```
sim %>%
  count(height < 100) %>%
  mutate(prop = n()/n)
```

```
## # A tibble: 2 × 3
##   `height < 100`     n      prop
##   <lgl>        <int>    <dbl>
## 1 FALSE          9731  0.000206
## 2 TRUE           269   0.00743
```

🤔 Sehr kleine Buschleute?

Extrem vage Priori-Verteilung, $\sigma = 100$

height ~ dnorm(mu, sigma)
mu ~ dnorm(178, 100)



Anteil negativer Größe:

```
sim %>%
  count(height < 0) %>%
  mutate(prop = n()/n)
```

```
## # A tibble: 2 × 3
##   `height < 0`     n      prop
##   <lgl>       <int>    <dbl>
## 1 FALSE         9571 0.000209
## 2 TRUE          429  0.00466
```

🧐 Das Modell geht apriori von ein paar Prozent Menschen mit *negativer* Größe aus. Ein Haufen Riesen 🤨 werden auch erwartet.

🧐 Vage (flache, informationsarme, "neutrale", "objektive") Priori-Werte machen oft keinen Sinn.

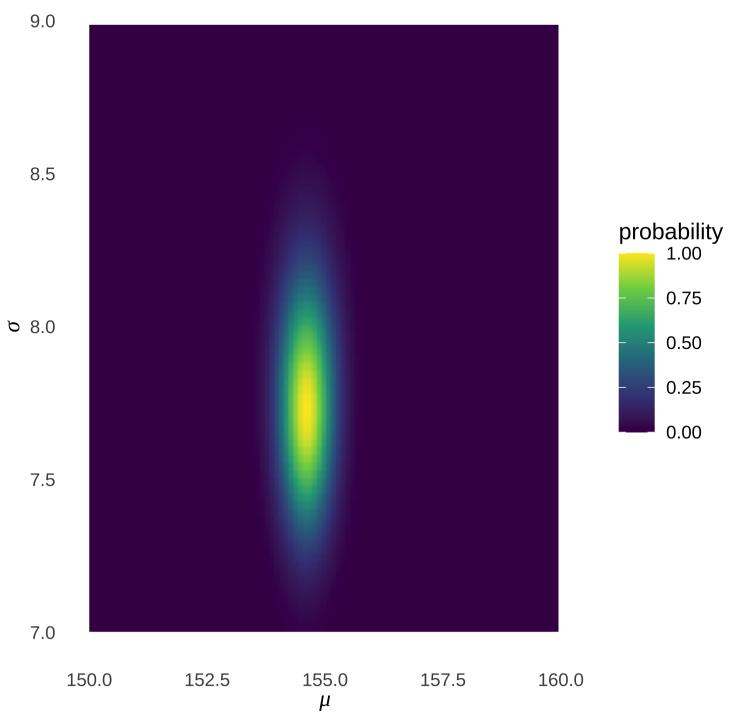
Zufällige Motivationsseite



**PRETTY, PRETTY, PRETTY,
PRETTY GOOD.**

HBO

Posteriori-Verteilung des Größen-Modells

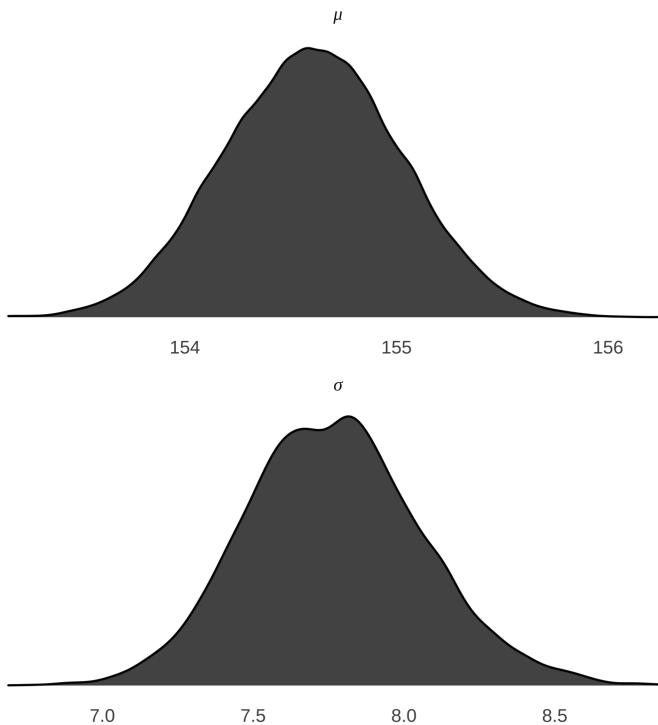


- Wir bekommen eine Wahrscheinlichkeitsverteilung für μ und eine für σ (bzw. eine zweidimensionale Verteilung, für die μ, σ -Paare).
- Trotz des vagen Priors sind die Posteriori-Werte für μ und σ klein: Die große Stichprobe hat die Priori-Werte überstimmt.
- Ziehen wir Stichproben aus der Posteriori-Verteilung, so können wir interessante Fragen stellen.

Hallo, Posteriori-Verteilung

... wir hätten da mal ein paar Fragen an Sie. 🧑

- Mit welcher Wahrscheinlichkeit ist die mittlere !Kung-Person größer als 1,55m?
- Welche mittlere Körpergröße wird mit 95% Wahrscheinlichkeit nicht überschritten, laut dem Modell?
- In welchem 90%-PI liegt μ vermutlich?
- Mit welcher Unsicherheit ist die Schätzung der mittleren Körpergröße behaftet (σ)?
- Welcher Wert der mittleren Körpergröße hat die höchste Wahrscheinlichkeit?



Posteriori-Verteilung mit quap() berechnen

- `quap()` (Quadratische Approximation) erlaubt uns, komfortabel die Posteriori-Verteilung zu bestimmen.
- Rechnerisch wird die quadratische Form der Normalverteilung ausgenutzt.
- Die Gittermethode wird nicht verwendet, aber die Ergebnisse sind - in bestimmten Situationen - ähnlich.
- `quap()` hat auch den Vorteil, dass es wenig rechenintensiv ist als die Gittermethode, gerade bei komplexeren Modellen.

```
library(rethinking)
# berechnet Post.-Vert.:
quap(
  alist(
    # modelldefinition
  ),
  data = meine_daten
)
```

Modelldefinition:

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

$$\mu \sim \mathcal{N}(178, 20)$$

$$\sigma \sim \mathcal{U}(0, 50)$$

Posteriori-Stichproben mit quap() berechnen

Modelldefinition:

```
modell_definition <-
  alist(
    height ~ dnorm(mu, sigma),
    mu ~ dnorm(178, 20),
    sigma ~ dunif(0, 50)
  )

m41 <- quap(
  modell_definition,
  data = d2
)
```

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

$$\mu \sim \mathcal{N}(178, 20)$$

$$\sigma \sim \mathcal{U}(0, 50)$$

Posteriori-Verteilung für Modell 4.1 m41:

```
precis(m41) # precis, engl. "Kurzfassung"

##           mean        sd      5.5%     94.5%
## mu     154.607022 0.4119945 153.948575 155.26547
## sigma  7.731329 0.2913857   7.265638   8.19702
```

Stichproben aus der Posteriori-Verteilung ziehen

```
post_m41 <- extract.samples(m41, n = 1e4)
precis(post_m41)
```

```
##           mean      sd    5.5%    94.5%
## mu     154.611253 0.4101497 153.950117 155.263315
## sigma  7.732909 0.2919296  7.269278  8.201933
```



Hier die ersten paar Zeilen von post_m41:

| | mu | sigma |
|---|----------|----------|
| 1 | 155.0387 | 7.660676 |
| 2 | 154.4266 | 7.532136 |
| 3 | 154.9872 | 8.263670 |
| 4 | 154.1836 | 7.989497 |
| 5 | 154.4975 | 7.482399 |
| 6 | 154.5213 | 8.020057 |

Mit welcher Wahrscheinlichkeit ist $\mu > 155$?

```
post_m41 %>%
  count(mu > 155) %>%
  mutate(prop = n/sum(n))
```

```
##   mu > 155     n   prop
## 1 FALSE 8282 0.8282
## 2 TRUE 1718 0.1718
```

Antworten von der Posteriori-Verteilung

Welche mittlere Körpergröße wird mit 95% Wahrscheinlichkeit nicht überschritten, laut dem Modell?

```
post_m41 %>%
  summarise(
    q95 =
      quantile(mu, .95))

##           q95
## 1 155.2825
```

In welchem 90%-PI liegt μ vermutlich?

```
post_m41 %>%
  summarise(
    pi_90 =
      quantile(mu, c(0.05,
                    0.95)))

##          pi_90
## 1 153.9298
## 2 155.2825
```

⚠ Ähnliche Fragen bleiben als Übung für die Leseris 😊.

Priori-Werte mit mehr Information, m4.2

- Die Priori-Verteilung von μ in m4.1 ist sehr vage (fast flach, indifferent, informationsarm, fast gleichverteilt, fast "objektiv").
- Aufgrund der großen Stichprobe ist der Schäzbereich für μ trotzdem schmal. Die vage Priori-Verteilung kommt daher nicht zum tragen.
- Bei kleineren Stichproben oder komplexeren Modellen kann die Priori-Verteilung deutlich mehr Einfluss haben.

Untersuchen wir den Effekt von Prior-Werten für μ mit mehr Information:

```
m4.2 <-  
  quap(  
    alist(  
      height ~ dnorm(mu, sigma),  
      mu ~ dnorm(170, 0.1),  
      sigma ~ dunif(0, 50)  
    ),  
    data = d2  
  )
```

```
##           mean        sd      5.5%     94.5%  
## mu     169.81621 0.1003582 169.65582 169.97661  
## sigma  17.07505 0.6500034 16.03622 18.11388
```

Man beachte, dass σ sich im Vergleich zu m4.1 geändert hat.

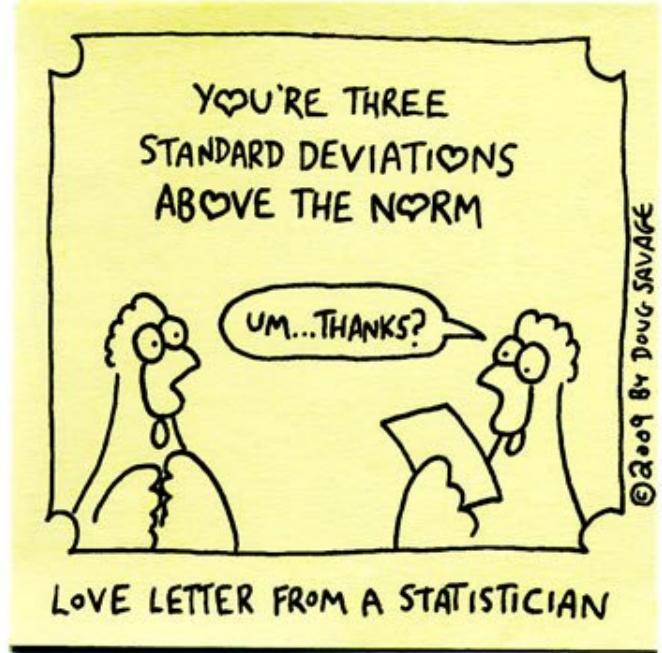
Fazit

- Wir haben die Posteriori-Verteilung für ein Gauss-Modell berechnet.
- Dabei hatten wir ein einfaches Modell mit metrischer Zielvariablen, ohne Prädiktoren, betrachtet.
- Die Zielvariablen, Körpergröße (`height`) haben wir als normalverteilt mit den Parametern μ und σ angenommen.
- Für μ und σ haben wir jeweils keinen einzelnen (fixen) Wert angenommen, sondern eine Wahrscheinlichkeitsverteilung, der mit der Priori-Verteilung für μ bzw. σ festgelegt ist.

♥ Bleiben Sie dran!

Savage Chickens

by Doug Savage



Hinweise

Zu diesem Skript

- Dieses Skript bezieht sich auf folgende Lehrbücher:
 - *Statistical Rethinking* (2. Auflage), Kapitel 4.1 - 4.3, [McElreath \(2020\)](#)
 - Der R-Code stammt aus [Kurz \(2021\)](#).
- Dieses Skript wurde erstellt am 2021-10-20 16:35:11 (WiSe 21).
- Lizenz: [CC-BY](#)
- Autor ist Sebastian Sauer.
- Um diese HTMLM-Folien korrekt darzustellen, ist eine Internet-Verbindung nötig.
- Eine PDF-Version kann erzeugt werden, indem man im Chrome-Browser druckt (Drucken als PDF).
- Mit der Taste ? bekommt man eine Hilfe über Shortcuts.

Literatur

Freeman, M. (2006). "A visual comparison of normal and paranormal distributions". In: *Journal of Epidemiology and Community Health* 60.1, p. 6. ISSN: 0143-005X. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2465539/> (visited on Sep. 09, 2021).

Gelman, A., J. Hill, and A. Vehtari (2021). *Regression and other stories*. Analytical methods for social research. Cambridge: Cambridge University Press. 534 pp. ISBN: 978-1-107-67651-0 978-1-107-02398-7.

Kurz, A. S. (2021). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition*. URL: <https://bookdown.org/content/4857/> (visited on Sep. 08, 2021).

McElreath, R. (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.

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