Thema 3: Stichproben aus der Posteriori-Verteilung ziehen

QM2, ReThink, Kap. 3

Prof. Sauer

AWM, HS Ansbach

WiSe 21

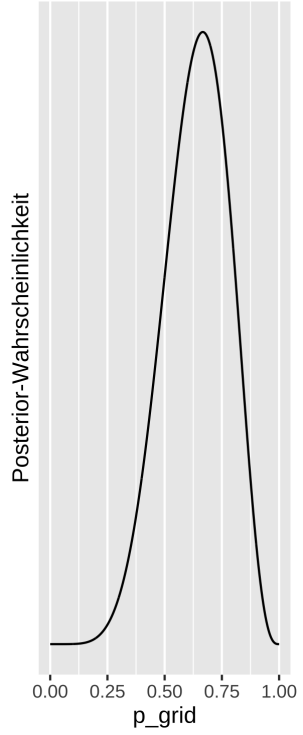
Gliederung

Mit Stichproben die Post-Verteilung zusammenfassen

Zur Erinnerung, die Gittermethode

Die Gittermethode ist ein Weg, die Posteriori-Verteilung zu berechnen. Die Posteriori-Verteilung birgt viele nützliche Informationen.

Modell: $W = 6$ Wasser, $N = 9$ Würfen und $k = 1000$ Gitterwerten.



Die ersten paar Zeilen aus der Tabelle d:

Tabelle *d* mit Daten zur Posteriori-Verteilung

p_grid	prior	likelihood	unstand_post	post
0	1	0	0	0
1×10^{-3}	1	8×10^{-17}	8×10^{-17}	8×10^{-19}
2×10^{-3}	1	5×10^{-15}	5×10^{-15}	5×10^{-17}
3×10^{-3}	1	6×10^{-14}	6×10^{-14}	6×10^{-16}
4×10^{-3}	1	3×10^{-13}	3×10^{-13}	3×10^{-15}
5×10^{-3}	1	1×10^{-12}	1×10^{-12}	1×10^{-14}

Befragen wir die Posteriori-Verteilung

Beispiele für Fragen an die Post-Verteilung*:

- Mit welcher Wahrscheinlichkeit liegt der Parameter unter einem bestimmten Wert?
- Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen zwei bestimmten Werten?
- Mit 5% Wahrscheinlichkeit liegt der Parameterwert nicht unter welchem Wert?
- Welcher Parameterwert hat die höchste Wahrscheinlichkeit?
- Wie ungewiss ist das Modell über die Parameterwerte?

Solche Fragen kann man in drei Gruppen aufteilen:

1. Fragen zu Bereichen von Parametern
2. Fragen zu Bereichen von Wahrscheinlichkeitsmassen
3. Fragen zu Punktschätzern von Parametern

*Post-Verteilung: Posteriori-Verteilung

Häufigkeiten sind einfacher als Wahrscheinlichkeiten

Tabelle mit Stichprobendaten aus der Posteriori-Verteilung (Tabelle `d`):

```
samples <-
  d %>%  # nimmt die Tabelle mit Posteriori-Daten,
  slice_sample(  # Ziehe daraus eine Stichprobe,
    n = 1e4,  # mit insgesamt n=10000 Elementen,
    weight_by = post,  # Gewichte nach Spalte mit Post-Wskt.,
    replace = T)  # Ziehe mit Zurücklegen
```

Die Wahrscheinlichkeit, einen Parameterwert aus Tabelle `d` zu ziehen, ist proportional zur Posteriori-Wahrscheinlichkeit (`post`) dieses Werts. Ziehen mit Zurücklegen hält die Wahrscheinlichkeiten während des Ziehens konstant.

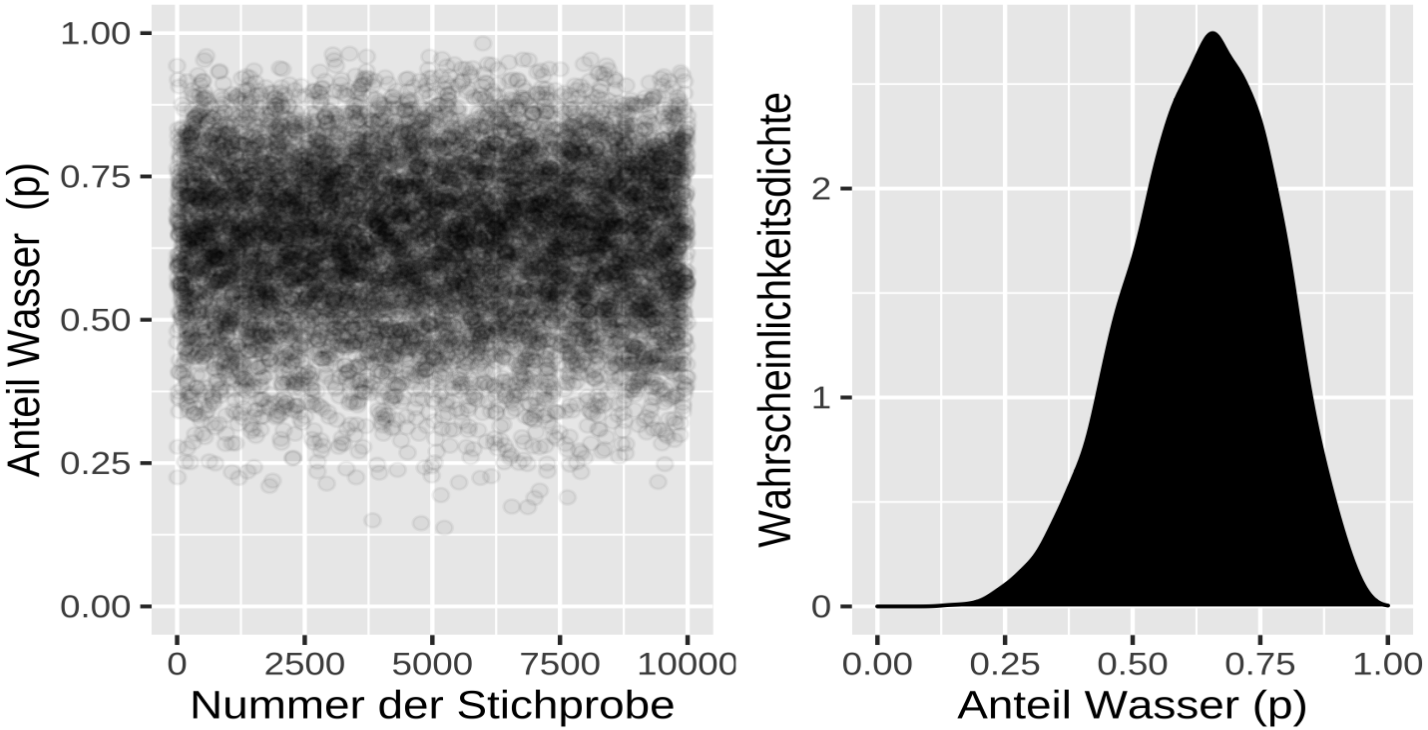
Stichprobendaten aus der Post-Verteilung

Nur die ersten Zeilen abgebildet

p_grid	prior	likelihood	unstand_post	post
0.683	1	0.272	3×10^{-1}	0.003
0.711	1	0.262	3×10^{-1}	0.003
0.515	1	0.178	2×10^{-1}	0.002

Visualisierung der Stichprobendaten

Datensatz `samples`, $n = 10^4$, basierend auf dem Modell oben.

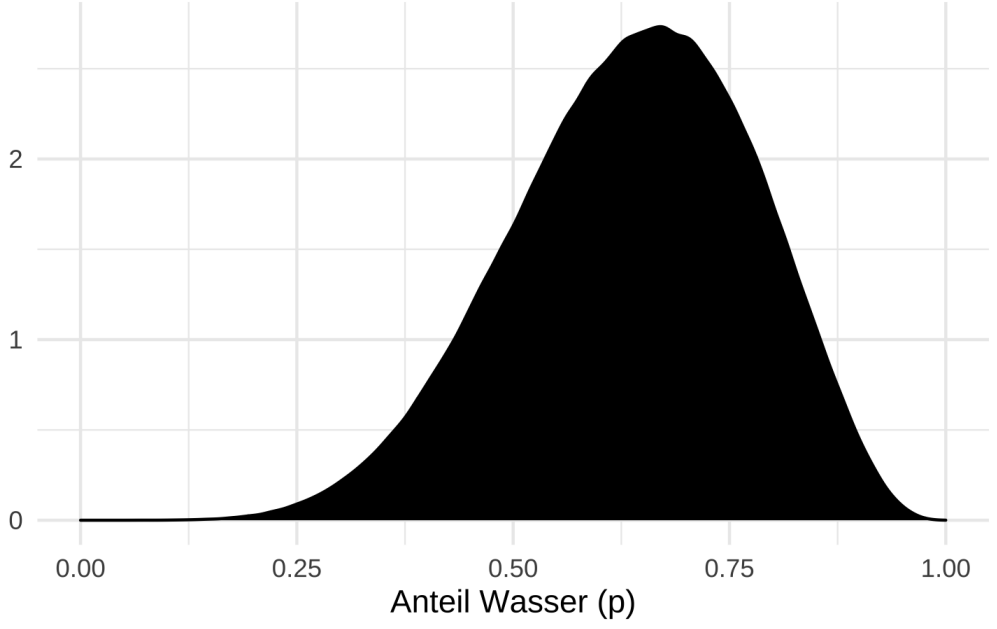


Die Stichprobendaten nähern sich der Posteriori-Verteilung an.

Mehr Stichproben glätten die Verteilung

$n = 10^6$ Stichproben aus der Posteriori-Verteilung

```
d %>%
  slice_sample(n = 1e6, weight_by = post, replace = T) %>%
  ggplot(aes(x = p_grid)) +
  geom_density(fill = "black") +
  scale_x_continuous("Anteil Wasser (p)", limits = c(0, 1)) +
  labs(y = "")
```



Fragen zu Bereichen von Parametern 1

Wie groß ist die Wahrscheinlichkeit, dass der Wasseranteil unter 50% liegt?

Aus der Posteriori-Verteilung mit der Gridmethode:

```
d %>%
  filter(p_grid < .5) %>%
  summarise(sum = sum(post))
```

```
## # A tibble: 1 × 1
##     sum
##   <dbl>
## 1 0.172
```

Aus den Stichproben der Posteriori-Verteilung:

```
samples %>%
  filter(p_grid < .5) %>%
  summarise(sum = n() / 1e4)
```

```
## # A tibble: 1 × 1
##     sum
##   <dbl>
## 1 0.169
```

Einfach wie 🍰 essen.

Die Gridmethode funktioniert bei großen Modellen nicht gut (im Gegensatz zur quadratischen Approximation `quap`). Daher werden wir ab jetzt mit den Stichproben arbeiten, weil das für `quap` auch funktioniert. Das ist außerdem einfacher.

Fragen zu Bereichen von Parametern 2

Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen 0.5 und 0.75?

```
samples %>%
  filter(p_grid > .5 & p_grid < .75) %>%
  summarise(sum     =       n() / 1e4,
            percent = 100 * n() / 1e4)  # In Prozent
```

```
## # A tibble: 1 × 2
##     sum percent
##   <dbl>   <dbl>
## 1 0.608    60.8
```

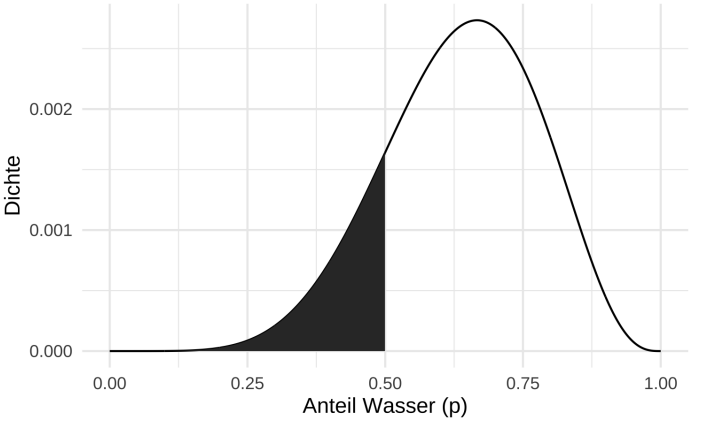
Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen 0.9 und 1?

```
samples %>%
  filter(p_grid >= .9 & p_grid <= 1) %>%
  summarise(sum     =       n() / 1e4,
            percent = 100 * n() / 1e4)  # In Prozent
```

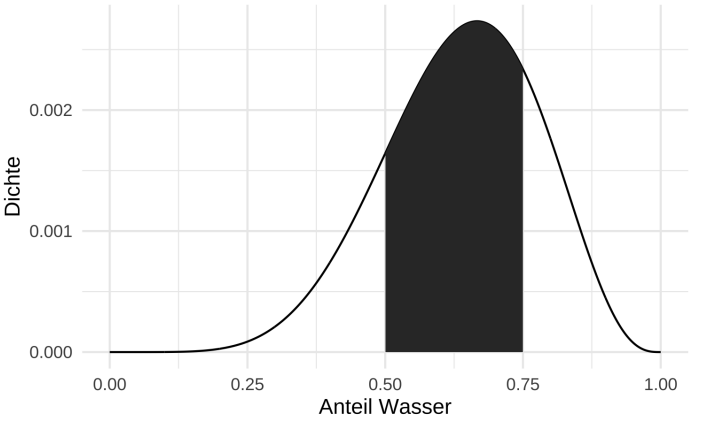
```
## # A tibble: 1 × 2
##      sum percent
##    <dbl>   <dbl>
## 1 0.0128    1.28
```

Visualisierung von Parameterwert-Intervalle

```
d %>%
  ggplot(aes(x = p_grid,
             y = post)) +
  geom_line() +
  geom_area(data = d %>%
              filter(p_grid < .5)) +
  labs(x = "Anteil Wasser (p)",
       y = "Dichte")
```




```
d %>%
  ggplot(aes(x = p_grid,
             y = post)) +
  geom_line() +
  geom_area(data = d %>%
            filter(p_grid < .75 &
                   p_grid > .5)) +
  labs(x = "Anteil Wasser",
       y = "Dichte")
```



Fragen zu Bereichen von Wahrscheinlichkeitsmassen

- Nennt man auch *Konfidenz- oder Vertrauensintervall*.
- Synonym: *Kompatibilitätsintervall* oder *Passungsbereich*.
- 💥 Traue niemals einem Golem, er lebt in der kleinen Welt.

Welcher Parameterwert wird mit 80% Wahrscheinlichkeit nicht überschritten, laut unserem Modell? (Gesucht sind also die unteren 80% Posteriori-Wahrscheinlichkeit)

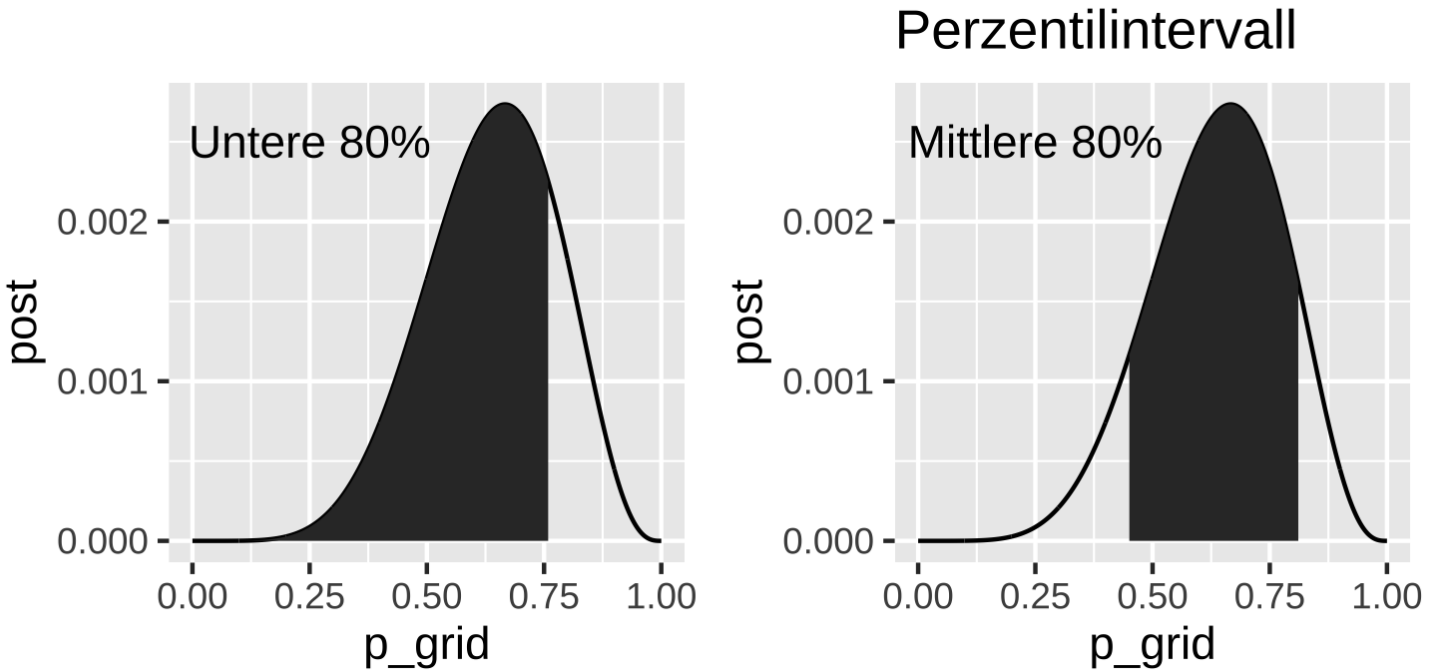
```
## # A tibble: 1 × 1
##   quantil80
##       <dbl>
## 1     0.760
```

Was ist das *mittlere* Intervall, das mit 80% Wahrscheinlichkeit den Parameterwert enthält, laut dem Modell?

```
## # A tibble: 1 × 2
##   quant_10 quant_90
##      <dbl>    <dbl>
## 1    0.449    0.812
```

Visualisierung der Massen-Intervalle

Intervalle (Bereiche), die die Wahrscheinlichkeitsmasse hälftig auf die beiden Ränder aufteilen, nennen wir *Perzentilintervalle*.



Schiefe Posteriori-Verteilungen sind möglich

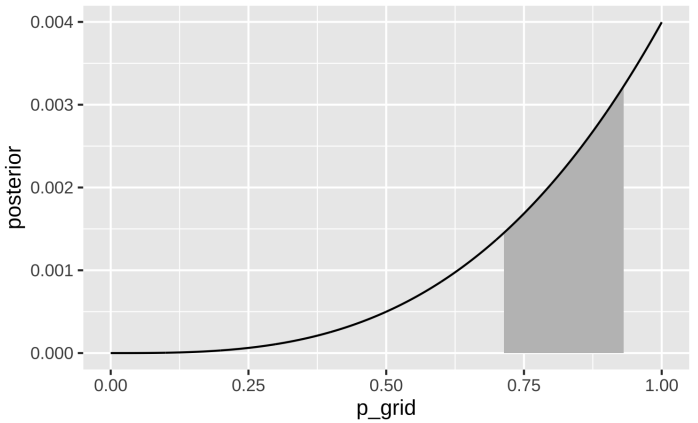
```
d <- d %>%
  mutate(likelihood = dbinom(3, size = 3, prob = p_grid),
         unstand_post = likelihood * prior) %>%
  mutate(posterior  = unstand_post / sum(unstand_post))

samples <- d %>%
    slice_sample(n = 1e4,
                 weight_by = posterior,
                 replace = T)
```

p_grid	prior	likelihood	unstand_post	posterior
0.649	1	0.273	0.273	0.001
0.658	1	0.284	0.284	0.001
0.829	1	0.569	0.569	0.002
0.962	1	0.890	0.890	0.004
0.811	1	0.533	0.533	0.002
0.869	1	0.656	0.656	0.003

Intervalle höchster Dichte

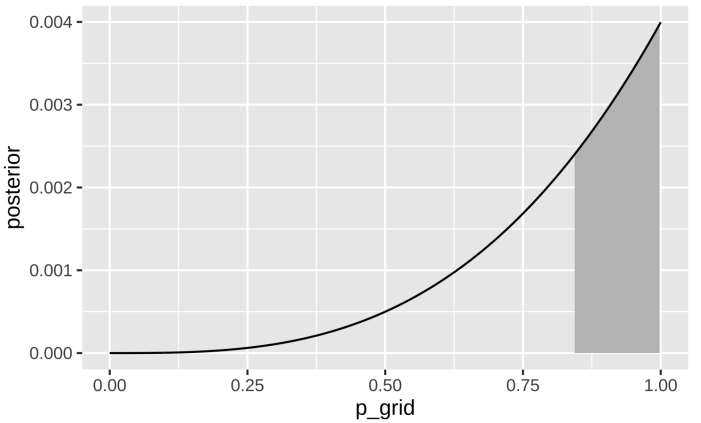
Daten: 3 Würfe mit 3 Treffern

50%-Perzentil-Intervall



Der wahrscheinlichste Paramterwert (1) ist *nicht* im Intervall enthalten!

50%-Intervall höchster Dichte



Der wahrscheinlichste Paramterwert (1) *ist* im Intervall enthalten!

Intervalle höchster Dichte vs. Perzentilintervalle

- Bei symmetrischer Posteriori-Verteilung sind beide Intervalle ähnlich
- Perzentilintervalle sind verbreiteter
- *Intervalle höchster Dichte* (Highest Posteriori Density Interval, HPDI) sind bei schiefen Post-Verteilungen zu bevorzugen
- Intervalle höchster Dichte sind die *schmalsten* Intervalle für eine gegebene Wahrscheinlichkeitsmasse

Intervallbreite HDPI: 0.16

```
rethinking::HPDI(samples$p_grid, prob = .5)
```

```
##       |0.5       0.5| 
## 0.8428428 0.9989990
```

Intervallbreite PI: 0.23

```
rethinking::PI(samples$p_grid, prob = .5)
```

```
##       25%       75% 
## 0.7134635 0.9319319
```

Punktschätzungen

Daten: 3 Treffer bei 3 Würfen.

Lageparameter

```
library(tidybayes)
samples %>%
  summarise(
    mean   = mean(p_grid),
    median = median(p_grid),
    # Mode() ist aus tidybayes:
    modus = Mode(p_grid))
```

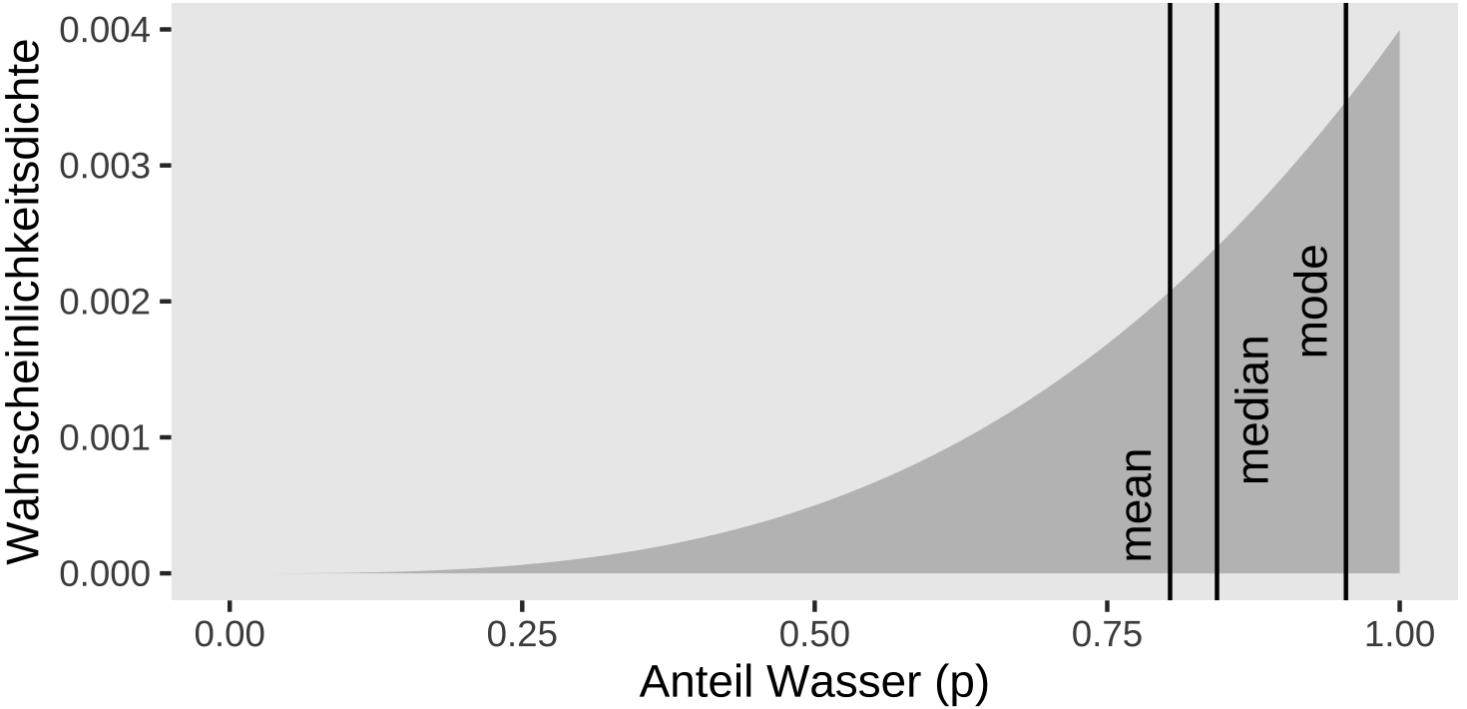
```
## # A tibble: 1 × 3
##    mean median modus
##   <dbl>  <dbl> <dbl>
## 1 0.804  0.844 0.954
```

Streuungsparameter

```
samples %>%
  summarise(
    mean   = sd(p_grid),
    median = IQR(p_grid),
    modus = mad(p_grid))
```

```
## # A tibble: 1 × 3
##    mean median modus
##   <dbl>  <dbl> <dbl>
## 1 0.161  0.218 0.150
```

Visualisierungen der Punktschätzer



Je symmetrischer die Verteilung, desto näher liegen die Punktschätzer aneinander (und umgekehrt).

Mit Stichproben neue Beobachtungen simulieren

Wir simulieren die Wasserzahl bei Globuswürfen

Wir berechnen den Likelihood (L) für $w = 0, 1, 2$ bei einem Globusversuch mit $N = 2$ und $p = 0.7$:

```
## [1] 0.09 0.42 0.49
```

Wir simulieren $n = 1$ neuen Globusversuch mit $N = 2, p = 0.7$ auf Basis dieser Wahrscheinlichkeiten:

```
rbinom(n = 1, size = 2, prob = 0.7)
```

```
## [1] 0
```

Dieser Versuch ergab $W = 0$. Warum nicht $n = 10$ neue Globusversuche simulieren:

```
rbinom(n = 10, size = 2, prob = 0.7)
```

```
##  [1] 0 2 1 1 1 1 2 1 1 2
```

Diese Versuche geben Aufschluss, welche Daten (wie oft Wasser) man bei einem bestimmten Modell, p, N, erwarten kann. Damit können wir prüfen, ob der Golem uns nicht in die Irre führt.

Mit guten Simulationen kommt man den wahren Werten nahe

Warum nicht $n = 10^6$ neue Globusversuche simulieren:

```
d <-
  tibble(
    draws =
      rbinom(1e6,
             size = 2,
             prob = .7))
d %>%
  count(draws) %>%
  mutate(proportion =
           n / nrow(d))
```

```
## # A tibble: 3 × 3
##   draws      n proportion
##   <int>  <int>      <dbl>
## 1     0  89770     0.0898
## 2     1 420629     0.421
## 3     2 489601     0.490
```

Diese simulierten Häufigkeiten sind sehr ähnlich zu den theoretisch bestimmten Häufigkeiten mit `dbinom`: Unser Modell liefert plausible Vorhersagen.

```
dbinom(0:2, size = 2, prob = .7)
```

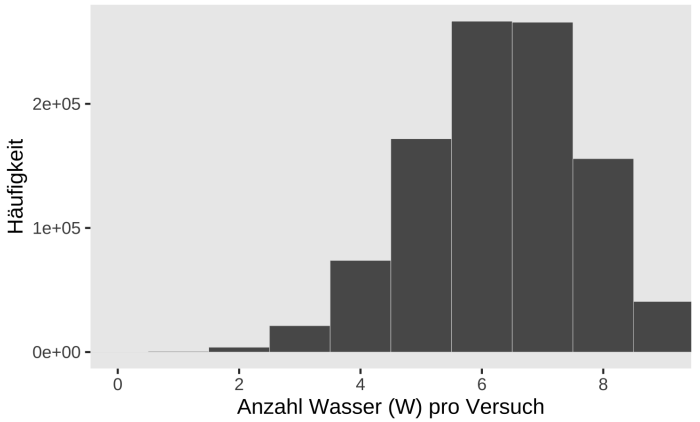
```
## [1] 0.09 0.42 0.49
```

Stichprobenverteilung

Wir ziehen viele Stichproben für den Versuch $N = 9$ Globuswürfe mit $p = 0.7$.

Wie viele Wasser (W) erhalten wir wohl typischerweise?

```
n_draws <- 1e6

d <-
  tibble(draws =
           rbinom(
             n_draws,
             size = 9,
             prob = .7
           ))

plot1 <-
d %>%
  ggplot(aes(x = draws)) +
  geom_histogram()
```

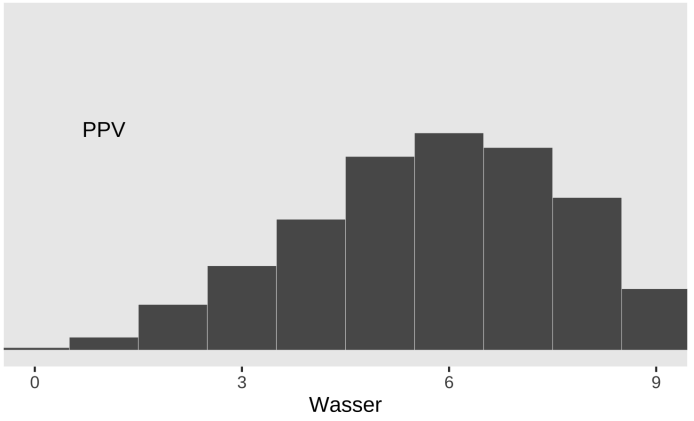


Die *Stichprobenverteilung* zeigt, welche Stichrpobendaten laut unserem Modell zu erwarten sind. Wir können jetzt prüfen, ob die echten Daten zu den Vorhersagen des Modells passen.

Visualisierung der PPV

Vorhersagen sind schwierig

... gerade wenn sie die Zukunft betreffen. Das zeigt uns die PPV: Der PPV unseres Modells gelingt es zwar, der theoretisch wahrscheinlichste Parameterwert ist auch der häufigste in unseren Stichproben, aber die Vorhersagen haben eine große Streuung, birgt also hohe Ungewissheit.



Würde man die Vorhersagen nur anhand eines bestimmten Parameterwertes (z.B $p = 0.6$) vornehmen, hätten die Vorhersagen zu wenig Streuung, würden also die Ungewissheit nicht ausreichend abbilden (Übergewissheit, Overconfidence).

Zwei Arten von Ungewissheit in Vorhersagen von Modellen

1. *Ungewissheit innerhalb des Modells*: Auch wenn der (oder die) Modellparameter eines Modells mit Sicherheit bekannt sind, so bleibt Unsicherheit, welche Beobachtung eintreten wird: Auch wenn man sicher weiß, dass $p = 1/4$ Murmeln blau sind, so kann man nicht sicher sagen, welche Farbe die nächste Murmel haben wird (Ausnahme: $p = 1$ oder $p = 0$).

2. *Ungewissheit in den Modellparametern*: Wir sind uns nicht sicher, welchen Wert p (bzw. die Modellparameter) haben. Diese Unsicherheit ist in der Post-Verteilung dargestellt.

Um zu realistischen Vorhersagen zu kommen, möchte man beide Arten von Ungewissheit berücksichtigen: Das macht die *Posteriori-Prädiktiv-Verteilung (PPV)*.

Die PPV zeigt, welche Daten das Modell vorhersagt (prädiktiv) und mit welcher Häufigkeit, basierend auf der Post-Verteilung.

Hinweise

Zu diesem Skript

Dieses Skript bezieht sich auf folgende Lehrbücher:

- Kapitel 3 aus McElreath (2020)
- R-Code stammt aus Kurz (2021).

Dieses Skript wurde erstellt am 2021-10-13 22:47:49.

Literatur

[1] A. S. Kurz. *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition*. 2021. URL: https://bookdown.org/content/4857/ (visited on 09/08/2021).

[2] R. McElreath. *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, 2020. ISBN: 978-0-367-13991-9.