

Start:Bayes!

Die Post befragen

Sebastian Sauer



Die Post befragen

Kapitelthema: Wie wir die **Posteriori-Verteilung** auslesen und zusammenfassen.



Lernziele

Nach diesem Kapitel können Sie ...

- die Post-Verteilung anhand einer Stichprobenverteilung auslesen
- Fragen nach Wahrscheinlichkeitsanteilen der Post-Verteilung beantworten
- Fragen nach Quantilen der Post-Verteilung beantworten
- die Post-Verteilung mit R visualisieren
- Schätzbereiche (Konfidenzintervalle) bestimmen: Perzentilintervalle und Highest Density Intervals (HDI)

Begleitliteratur: McElreath ([2020](#)), Kap. 3.1 und 3.2



Zur Erinnerung: Bayesbox in R berechnen



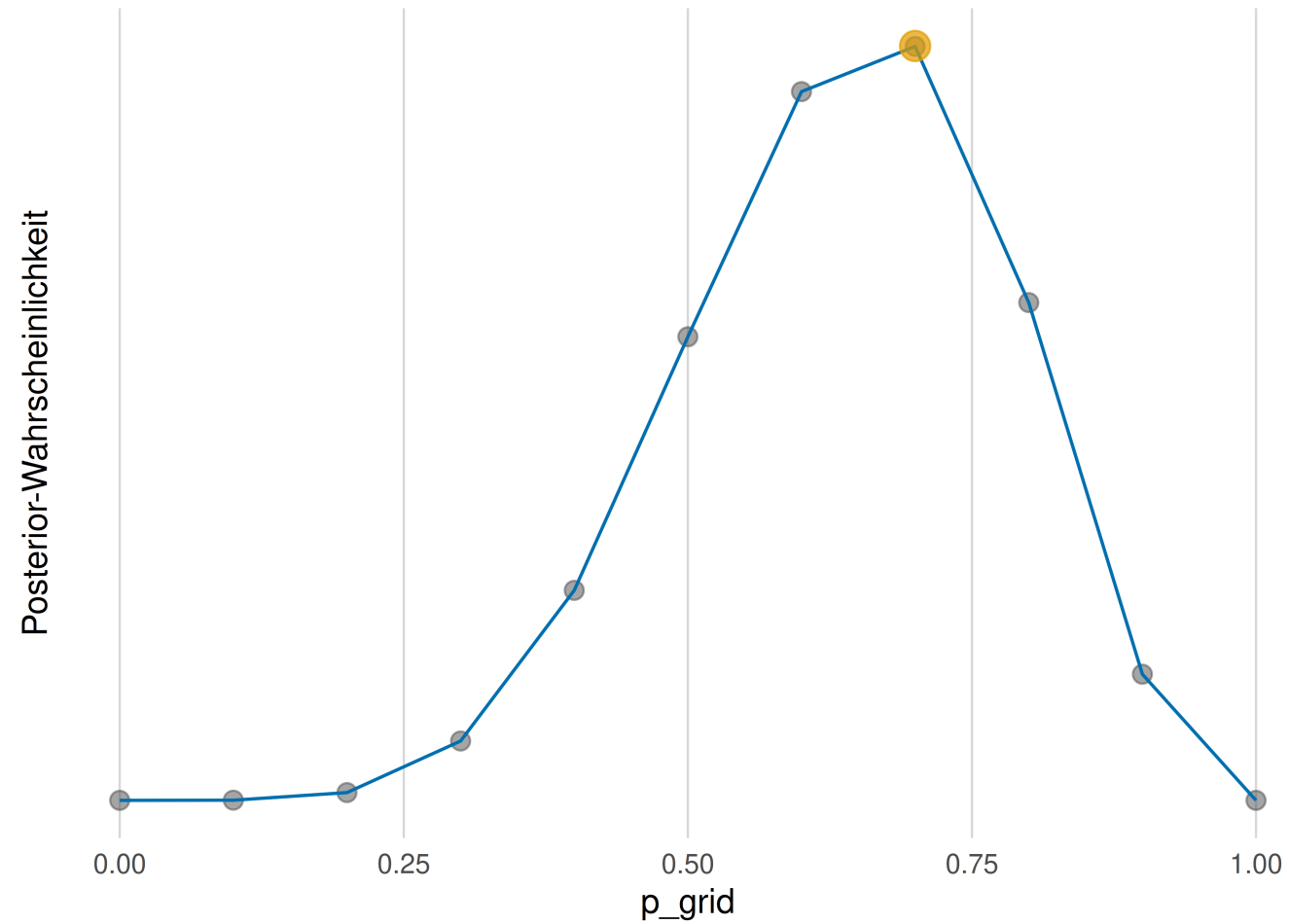
Zur Erinnerung: Die Bayesbox

- Die **Bayesbox** (“Gittermethode”) berechnet die Posteriori-Verteilung
- Die Post-Verteilung ist das **zentrale Ergebnis** einer Bayes-Analyse
- Beispiel: Globusversuch mit $W = 6$ Wasser bei $N = 9$ Würfeln, $g = 11$ Gitterwerte (0 bis 1, Schritte von 0.1)

Bei Apriori-Indifferenz entsprechen die Posteriori-Wahrscheinlichkeiten den Likelihoods — das Maximum liegt bei $6/9 = 2/3$.



Die Bayesbox als Grafik



Zur Erinnerung: Bayesbox in R berechnen



Bayesbox automatisiert

- Die Berechnung lässt sich als Funktion `bayesbox()` bündeln
- Verfügbar im R-Paket `prada`
(`pak::pak("sebastiansauer/prada")`)
- Aufruf: `bayesbox(hyps = p_grid, priors = 1, liks = L)`
- Argumente: `hyps` (Parameterwerte), `priors` (Priori-Werte), `liks` (Likelihoods)



Grenzen der Bayesbox

- Für einfache Modelle funktioniert die Bayesbox gut
- Bei einer Regression mit Parametern β_0, β_1, σ : je 10 Ausprägungen $\rightarrow 10 \cdot 10 \cdot 10 = 10^3$ Zeilen
- Bei 100 Ausprägungen: $100^3 = 10^6$ Zeilen
- Bei 10 Parametern mit je 100 Ausprägungen: 10^{100} Zeilen — unmöglich zu berechnen

Wir brauchen eine andere Lösung.



Mit Stichproben die Post-Verteilung zusammenfassen



Der Trick: Häufigkeiten statt Wahrscheinlichkeiten

- Komplexere Bayes-Modelle lassen sich nicht mehr exakt ausrechnen
- Lösung: Wir arbeiten mit **Stichproben (Häufigkeiten)** statt mit exakten Wahrscheinlichkeiten
- In der Praxis übernehmen das Monte-Carlo-Markov-Ketten-Verfahren (MCMC), z. B. der R-Golem **Stan** (Paket [rstanarm](#))
- Wie MCMC im Detail funktioniert, ist nicht Gegenstand dieses Kurses — wir wenden es einfach an



Stichproben aus der Post-Verteilung ziehen

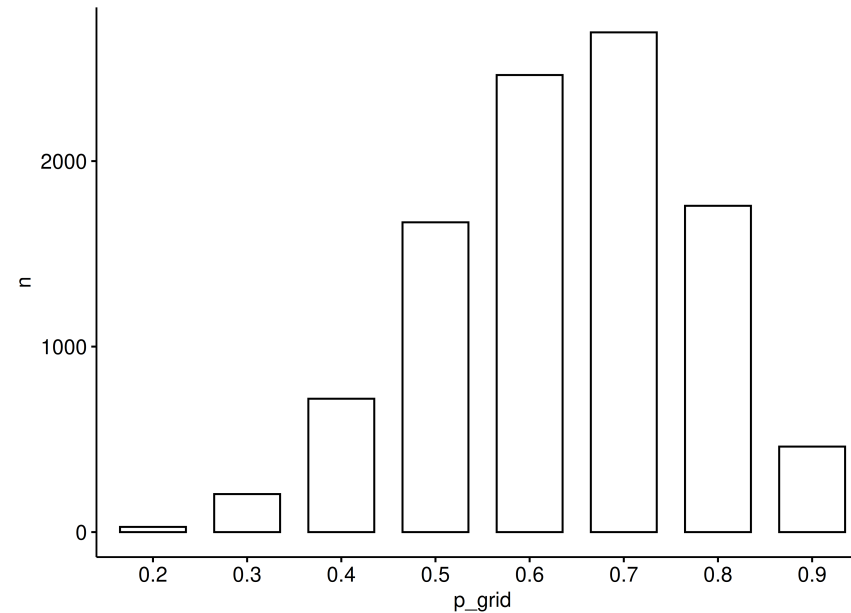
- Aus der Bayesbox `bayesbox_6w_9v` ziehen wir $n = 10\,000$ Stichproben
- Gewichtet nach der Posteriori-Wahrscheinlichkeit (`weight_by = post`)
- Mit Zurücklegen (`replace = TRUE`), damit die Wahrscheinlichkeiten konstant bleiben

```
1 samples_6w_9v <- bayesbox_6w_9v %>%  
2   slice_sample(n = 1e4, weight_by = post, replace = TRUE) %>%  
3   select(p_grid)
```

Ab jetzt arbeiten wir überwiegend mit Post-Verteilungen **auf Stichprobenbasis**, weil das für komplexere Modelle nötig wird.



Stichproben-Bayesbox als Balkendiagramm

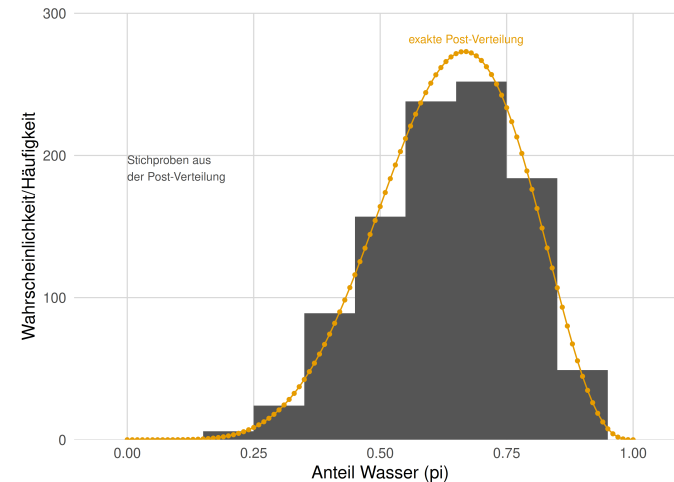


- Wasseranteile zwischen ca. 50 % und 70 % sind am wahrscheinlichsten
- Kleine Anteile (z. B. 25 %) sind nicht ausgeschlossen
- Die Stichprobenverteilung nähert sich gut an die exakte Bayesbox-Verteilung an

Mit Stichproben die Post-Verteilung zusammenfassen



Feineres Gitter: $g = 100$



- Exakte Post-Verteilung (Linie) und Stichproben-Post-Verteilung (Histogramm) stimmen gut überein
- Zentrale Tendenz: ca. 70 % Wasseranteil ist der “typische” Wert
- Das Modell zeigt deutliche Streuung — es ist sich nicht sehr sicher

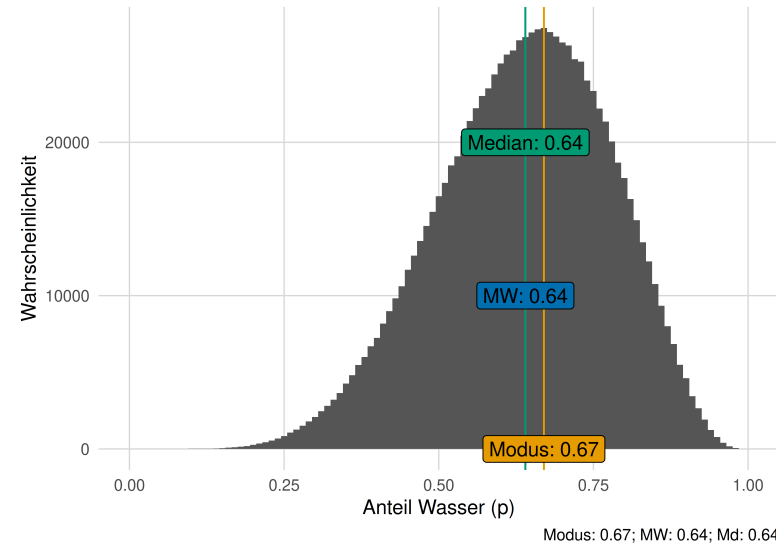
Hinweis

Mehr Stichproben und mehr Gitterwerte glätten die Stichproben-Post-Verteilung.

Mit Stichproben die Post-Verteilung zusammenfassen



Eine Million Stichproben



- Bei symmetrischer Verteilung liegen Modus, Mittelwert (MW) und Median (Md) eng beieinander
- Je mehr Stichproben, desto glatter die resultierende Verteilung

Die Post-Verteilung befragen

Zwei Arten von Fragen an die Post

! Wichtig

Die Post-Verteilung ist das zentrale Ergebnis einer Bayes-Analyse. Wir können viele nützliche Fragen an sie stellen.

Zwei Arten von Fragen:

A. Fragen nach **Wahrscheinlichkeiten** (p) B. Fragen nach **Parameterwerten** (Quantilen, q)

Beispiele für Fragen an die Post-Verteilung

A. Nach Wahrscheinlichkeiten (p):

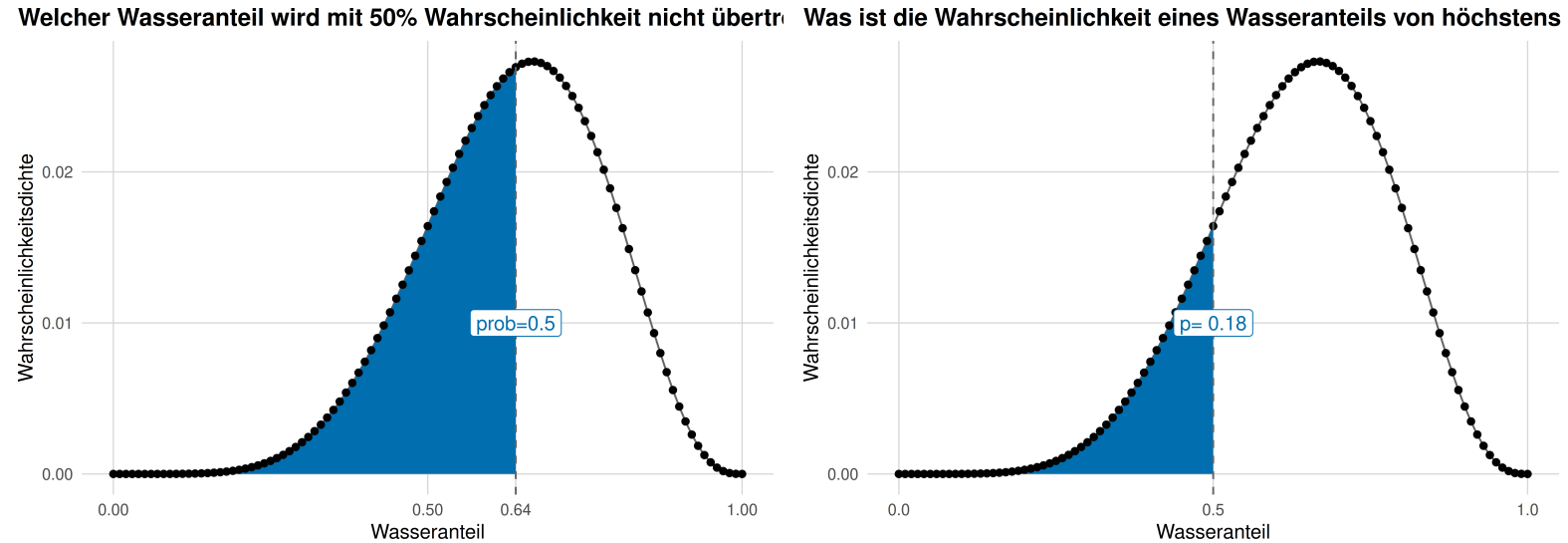
- Mit welcher Wahrscheinlichkeit liegt der Parameter unter/über einem bestimmten Wert?
- Mit welcher Wahrscheinlichkeit liegt der Parameter zwischen zwei Werten?

B. Nach Parameterwerten (Quantilen, q):

- Welcher Wert wird mit 95 % Wahrscheinlichkeit nicht überschritten?
- Welcher Parameterwert hat die höchste Wahrscheinlichkeit?
- Wie ungewiss ist das Modell über die Parameterwerte?



p vs. q — der Unterschied



- **q-Frage:** Welcher Wasseranteil wird mit gegebener Wahrscheinlichkeit p nicht überschritten? (gesucht: Parameterwert)
- **p-Frage:** Wie wahrscheinlich ist ein Wasseranteil von höchstens q ? (gesucht: Wahrscheinlichkeit)

Fragen nach Wahrscheinlichkeiten (p)

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

- Wir zählen, wie viele Stichproben die Bedingung erfüllen
- Und teilen durch die Gesamtzahl der Stichproben

```
1 samples_6w_9v %>% count(p_grid < .5)
```

Ergebnis: ca. **10 %** Wahrscheinlichkeit für einen Wasseranteil unter 50 %.



Weitere p-Fragen

Wie wahrscheinlich liegt der Wasseranteil zwischen 0.5 und 0.75?

```
1 samples_6w_9v %>%  
2   count(p_grid > .5 & p_grid < .75) %>%  
3   mutate(Anteil = n / 1e4, Prozent = 100 * n / 1e4)
```

Wie wahrscheinlich liegt der Wasseranteil zwischen 0.9 und 1?

- Ergebnis: sehr unwahrscheinlich, dass der Wasseranteil mindestens 90 % beträgt



Modus, Mittelwert, Median der Post

- Modus (häufigster/wahrscheinlichster Wert):
`map_estimate()` — **MAP** = Maximum A posteriori
- Mittelwert: `mean(p_grid)`
- Median: `median(p_grid)`

Alle drei sind **Punktschätzer** der Post-Verteilung und liegen bei symmetrischen Verteilungen eng beieinander.



Fragen nach Parameterwerten (Quantile)

! Wichtig

Schätzbereiche von Parameterwerten nennt man auch *Konfidenz-* oder *Vertrauensintervalle* (auch: Kompatibilitätsintervall, Ungewissheitsintervall).

Frage: *Welcher Parameterwert wird mit 90 % Wahrscheinlichkeit nicht überschritten?*

```
1 samples_6w_9v %>% summarise(quantil90 = quantile(p_grid, p = .9))
```

Antwort: Wir können zu 90 % sicher sein, dass der Wasseranteil kleiner als ca. 78 % ist.

Das mittlere 90 %-Intervall

Frage: *Welches Intervall enthält mit 90 % Wahrscheinlichkeit den Parameterwert?*

- Links und rechts je 5 % der extremsten Stichproben abschneiden

```
1 samples_6w_9v %>%  
2   summarise(quant_05 = quantile(p_grid, 0.05),  
3             quant_95 = quantile(p_grid, 0.95))
```

Solche Fragen lassen sich mit **Quantilen** beantworten.

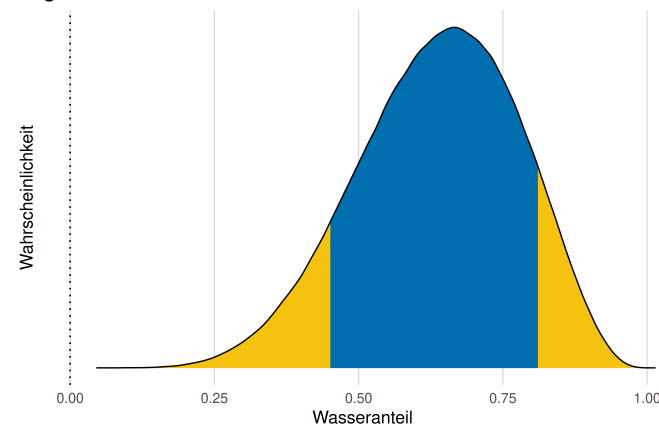
Visualisierung der Verteilungen

Definition: Perzentilintervall (PI)

Definition: Perzentilintervall (PI)

Intervalle, die die “abzuschneidende” Wahrscheinlichkeitsmasse hälftig auf beide Ränder aufteilen, nennt man **Perzentilintervalle (PI)** oder (synonym) **Equal-Tails-Intervalle (ETI)**.

80%-ETI: Die mittleren 80% der Wahrscheinlichkeit liegen zwischen 0.45 und 0.81



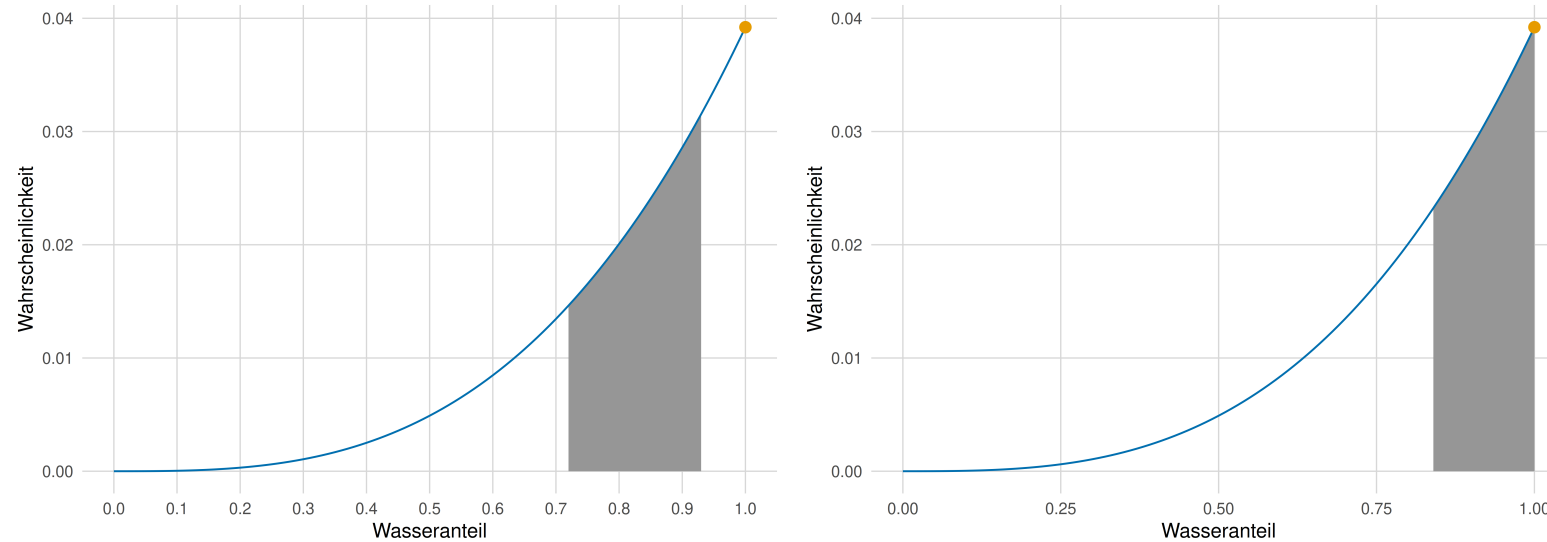
Schiefe Post-Verteilungen

Beispiel: Globusversuch mit 3 Würfeln, 3 Treffern (Wasser)

- Extremfall: alle Würfe zeigen Wasser
- Die resultierende Post-Verteilung ist stark **schief** (rechtslastig, Maximum am Rand bei $p = 1$)
- Gut geeignet, um PI und HDI zu vergleichen



PI vs. HDI bei schiefer Verteilung



- Ein **Perzentilintervall** kann bei schiefen Verteilungen den wahrscheinlichsten Parameterwert ausschließen — das ergibt wenig Sinn
- Ein **Highest-Density-Intervall (HDI)** ist schmaler und enthält immer den wahrscheinlichsten Wert

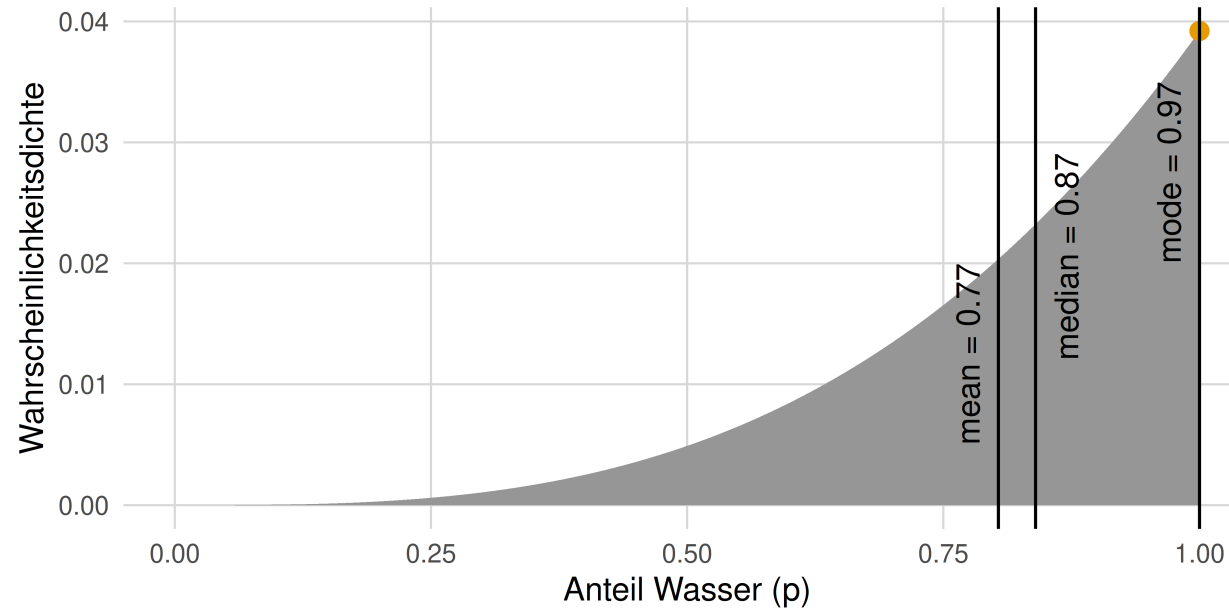
Definition: Intervalle höchster Dichte (HDI)

Definition: Intervalle höchster Dichte (HDI)

Intervalle höchster Dichte (Highest Density Intervals, HDI oder HDPI) sind definiert als das **schmälste** Intervall, das den gesuchten Parameter enthält (bezogen auf ein gegebenes Modell).

- Beim HDI sind die abgeschnittenen Ränder nicht zwangsläufig gleich groß
- Beim PI ist die Wahrscheinlichkeitsmasse in beiden Rändern gleich groß

Punktschätzer bei schiefer Verteilung



Je unsymmetrischer die Verteilung, desto weiter liegen die Punktschätzer auseinander.

HDI berechnen

```
1 samples_6w_9v %>%  
2   select(p_grid) %>%  
3   hdi(ci = .5) # aus dem Paket easystats
```

Ergebnis: Das Modell ist sich zu 50 % sicher, dass der Wasseranteil im Bereich von ca. 0.67 bis 0.78 liegt (HDI-basiert).

Fazit

Fazit: PI vs. HDI

- Bei **symmetrischer** Posteriori-Verteilung sind PI und HDI ähnlich oder identisch
- Perzentilintervalle sind in der Praxis verbreiteter
- **Intervalle höchster Dichte (HDI)** sind bei **schiefen** Post-Verteilungen zu bevorzugen
- HDI sind die **schmalsten** Intervalle für eine gegebene Wahrscheinlichkeitsmasse



Zusammenfassung am Beispiel

Globusversuch: 6 Treffer bei 9 Würfeln (`samples_6w_9v`)

Lageparameter — welcher mittlere Wasseranteil ist zu erwarten?

```
1 samples_6w_9v %>% summarise(mean = mean(p_grid), median = median(p_grid))
```

Streuungsparameter — wie unsicher ist die Schätzung?

```
1 samples_6w_9v %>% summarise(p_sd = sd(p_grid), p_iqr = IQR(p_grid), p_mad =
```

Üblicher als reine Streuungsmaße: ein HDI oder PI für den Parameter angeben.



Merksatz zum Abschluss

! Wichtig

Der Wasseranteil ist ein Beispiel für einen **Parameter** — eine unbekannte Größe eines Modells. Die Post-Verteilung sagt uns, wie plausibel welche Parameterwerte sind, und erlaubt uns, diese Unsicherheit präzise zu beziffern.

Kernpunkte dieses Kapitels

- Die Bayesbox liefert die exakte Post-Verteilung, stößt aber bei komplexen Modellen an Grenzen
- **Stichproben** aus der Post-Verteilung sind der praktikable Ersatz (Basis für MCMC/Stan)
- Man kann die Post-Verteilung nach **Wahrscheinlichkeiten (p)** und nach **Parameterwerten (q)** befragen
- **PI** (Perzentilintervall) und **HDI** (Intervall höchster Dichte) fassen die Unsicherheit als Bereich zusammen
- Bei schiefen Verteilungen ist das HDI vorzuziehen



Vielen Dank!

Fragen?

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2. Aufl.). Taylor and Francis, CRC Press.

