Small Worlds and Large Worlds & Sampling the Imaginary

Chapter 2 & 3 of Statistical Rethinking

Small worlds and large worlds

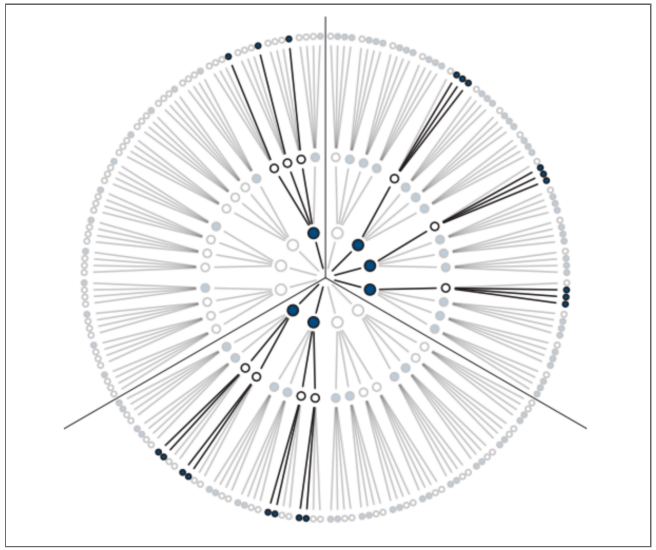
Small World

Large World

- Small World: Self-contained logical world of the **model**.
- Bayesian models is a logic machine optimally uses the data to make inferences about the small world.
- Large World: The real world.
- There may be events not imagined in the model.
- The logical consistency of the model is no guarantee to be optimal in the large world.
- Large world performance has to be demonstrated rather than logically deduced.

The garden of forking data

The garden of forking paths



The garden of forking data

- A bayesian analysis is a garden of forking data, in which alternative sequences of events are cultivated.
- It generates a quantitative ranking of hypotheses in the small world.

Building a model

- You want to infer the ratio of the earth surface covered by water.
- Data: W L W W L W L W
- Data Story

1. *underlying reality*: true proportion of water is p
2. *sampling process*:

- A single toss has a probability p of being water.
- Each toss is independent.
- Bayesian updating

$$\text{Posterior} = \frac{\text{Prob of the data} \times \text{Prior}}{\text{Average prob of the data}}$$

- Evaluation: the model's certainty $\neq$ good model

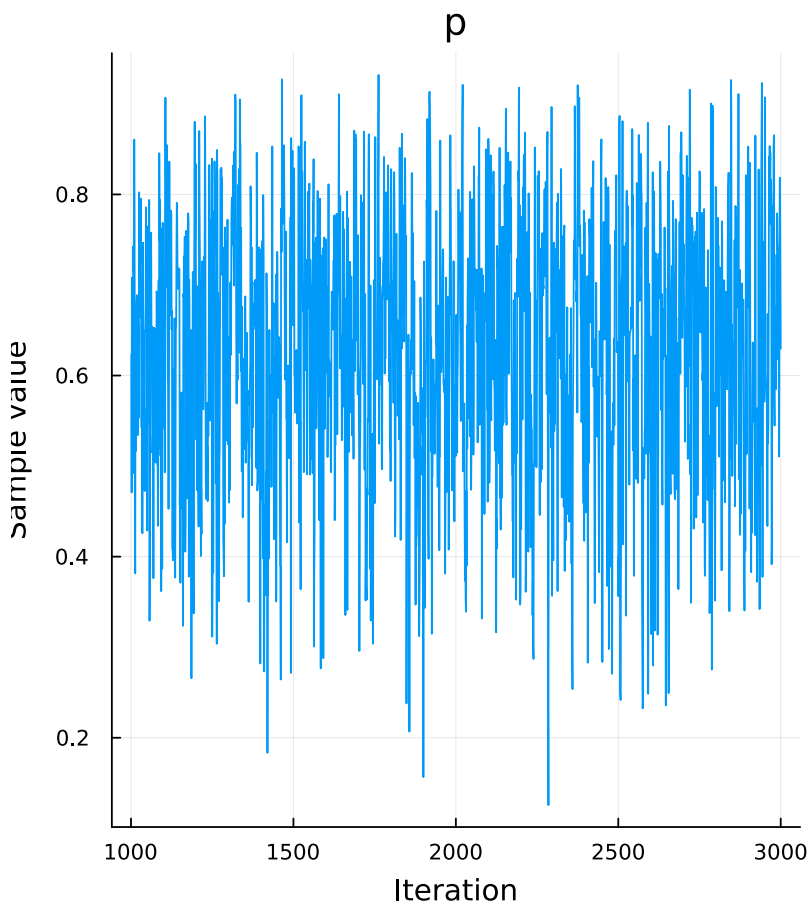
Turing Code

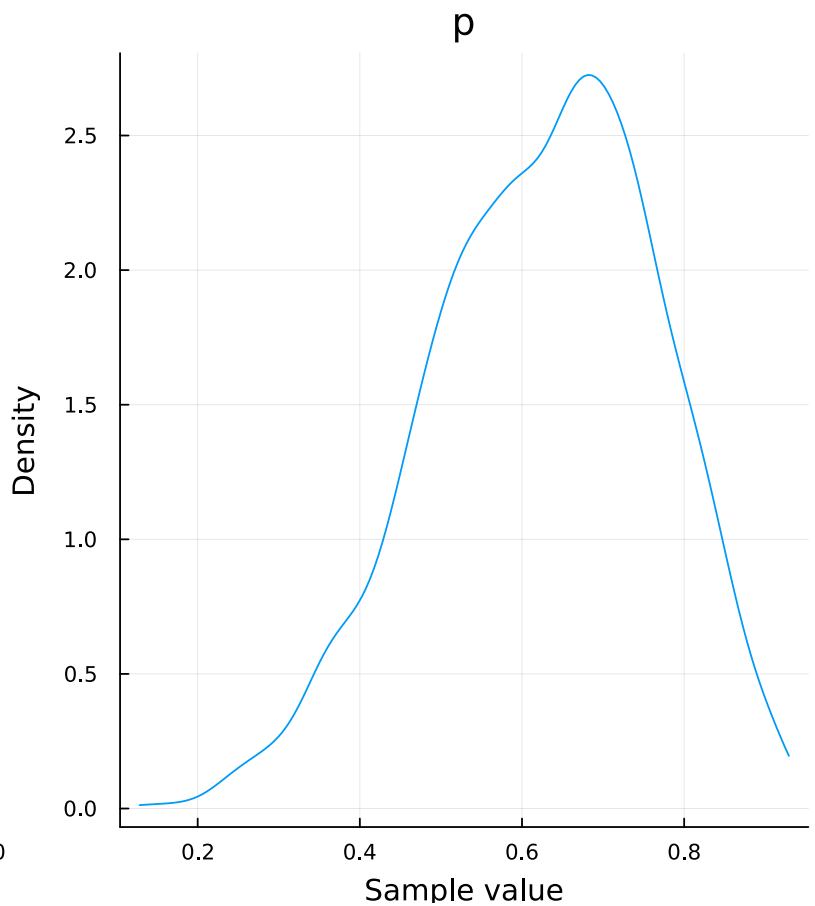
```
using StatsPlots, Turing, Logging;
default(labels=false)

@model function water_land(W, L)
    p ~ Uniform(0, 1)
    W ~ Binomial(W + L, p)
end

Logging.disable_logging(Logging.Warn)
chain = sample(water_land(6, 3), NUTS(), 2000)

plot(chain, size=(900, 500))
```

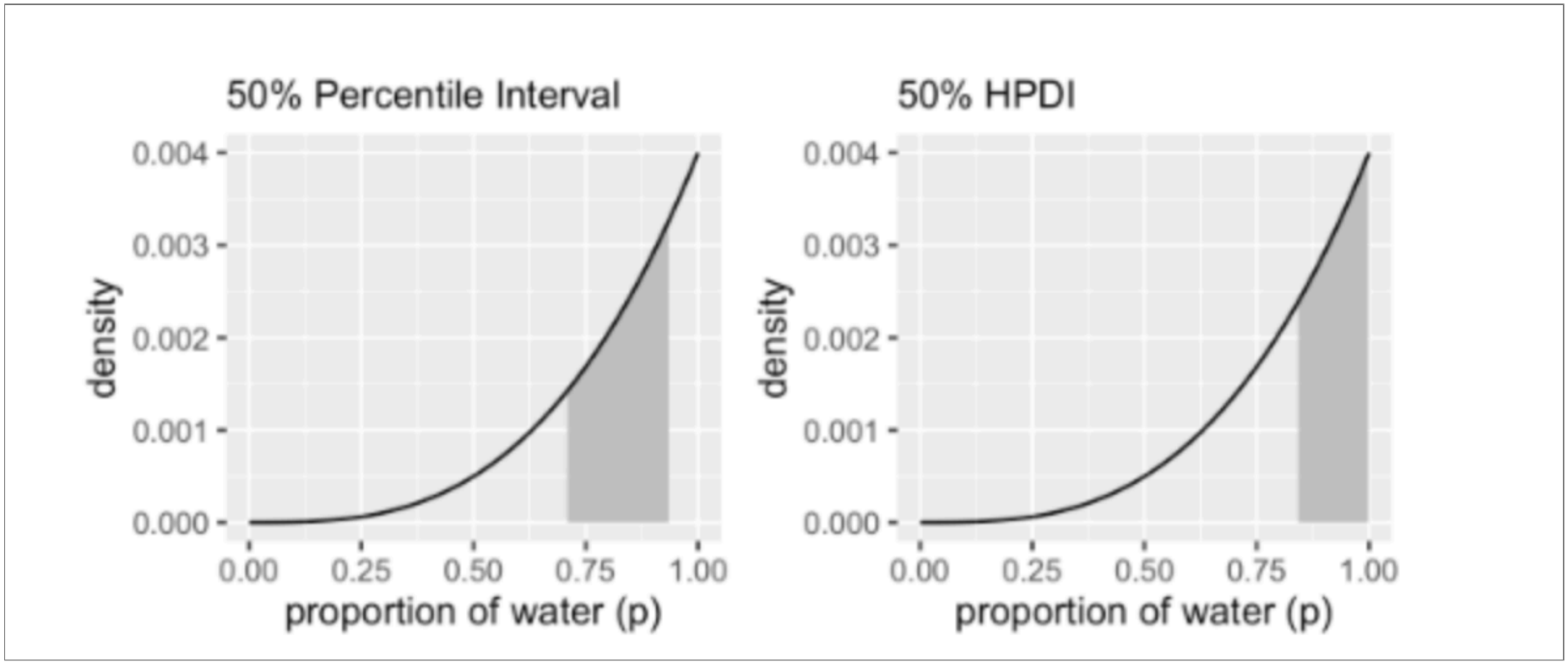

p
Sample value
0.8
0.6
0.4
0.2
1000
1500
2000
2500
3000
Iteration


p
Density
2.5
2.0
1.5
1.0
0.5
0.0
0.2
0.4
0.6
0.8
Sample value

Sampling the Imaginary

- Most people think about the world in terms of counts rather than probability.
- Why statistics cannot save bad science?
- You want to test a hypothesis, science history shows $Pr(true) = 0.01$.
- You have a positive finding (the statistical test has power 0.95 and size 0.05)
- $Pr(true \mid positive) = \frac{Pr(positive|true) \cdot Pr(true)}{Pr(positive)} \approx 0.161$
- Even if you choose significant threshold of 0.01, $Pr(true \mid positive) \approx 0.5$.
- Thinking (increasing $Pr(true)$) is more important than testing.

PI and HPDI



Point Estimates

- choose a point estimator $\approx$ choose a loss function
- negative posterior density $\Rightarrow$ MAP
- L1 loss $\Rightarrow$ posterior median

- L2 loss $\Rightarrow$ posterior mean

Simulation

- Why simulate observations from the model?
- Model design: understanding the implication of prior
- Model checking: check the model fit
  - Is tossing the globe i.i.d.?

- Forecasting: used for model revision

Excercise

2H1.

Suppose there are two species of panda bear. Both are **equally common** in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They **differ however in their family sizes**. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research. Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

2H1 Solution

$$Pr(A|D) \propto Pr(D|A) \cdot Pr(A) = 0.1 \cdot 0.5 = 0.05$$
$$Pr(B|D) \propto Pr(D|B) \cdot Pr(B) = 0.2 \cdot 0.5 = 0.1$$
$$\Rightarrow Pr(A|D) = \frac{0.05}{0.05 + 0.1} = \frac{1}{3}$$
$$Pr(B|D) = \frac{0.1}{0.05 + 0.1} = \frac{2}{3}$$
$$\Rightarrow Pr(Twins|D)$$
$$= Pr(Twins|A, D) \cdot Pr(A|D)$$
$$+ Pr(Twins|B, D) \cdot Pr(B|D)$$
$$= (1/3) \cdot 0.1 + (2/3) \cdot 0.2 = \frac{1}{6}$$

2H2

Recall all the facts from the problem above. Now compute the probability that the panda we have is from species A, assuming we have observed only the first birth and that it was twins.

2H2 Solution

$Pr(A|D) = \frac{1}{3}$

2H3

Continuing on from the previous problem, suppose the same panda mother has a second birth and that it is not twins, but a singleton infant. Compute the posterior probability that this panda is species A.

2H3 Solution

$$Pr(A|D_1, D_2) \propto Pr(D_2|A, D_1) \cdot Pr(A|D_1) = 0.9 \cdot (1/3)$$
$$Pr(B|D_1, D_2) \propto Pr(D_2|B, D_1) \cdot Pr(B|D_1) = 0.8 \cdot (2/3)$$
$$Pr(A|D_1, D_2) = \frac{0.9 \cdot (1/3)}{0.9 \cdot (1/3) + 0.8 \cdot (2/3)} = \frac{9}{25} = 0.36$$

2H4

A common boast of Bayesian statisticians is that Bayesian inference makes it easy to use all of the data, even if the data are of different types. So suppose now that a veterinarian comes along who has a new genetic test that she claims can identify the species of our mother panda.

- The probability it correctly identifies a species A panda is 0.8.
- The probability it correctly identifies a species B panda is 0.65.

  The vet administers the test to your panda and tells you that the test is positive for species A. First ignore your previous information from the births and compute the posterior probability that your panda is species A. Then redo your calculation, now using the birth data as well.

2H4 Solution

- Without birth data

$$Pr(A|pos) = \frac{Pr(pos|A) \cdot Pr(A)}{Pr(pos|A) \cdot Pr(A) + Pr(pos|B) \cdot Pr(B)}$$
$$= \frac{0.8 \cdot 0.5}{0.8 \cdot 0.5 + 0.35 \cdot 0.5} \approx 0.7$$

- With birth data

$$Pr(A|pos, D)$$
$$= \frac{Pr(pos|A, D) \cdot Pr(A|D)}{Pr(pos|A, D) \cdot Pr(A|D) + Pr(pos|B, D) \cdot Pr(B|D)}$$
$$= \frac{0.8 \cdot 0.36}{0.8 \cdot 0.36 + 0.35 \cdot 0.64}$$
$$= 0.5625$$

Speaker notes