rethinking-notes

Contents

Chapter 1 - The Golem of Prague

1.1 - Statistical Golems

"Golem" analogy to clay robots that just do what they're told without thinking, leading to carelessness. Common in statistics too, consider flowcharts with many different tests at the end. Need a more generalized approach (think engineering - don't start with building bridges, start with physics), rethinking inference as a set of strategies, not a set of tools.

1.2 - Statistical Rethinking

Many have approach that the objective of inference is to test null hypotheses, but science isn't described by falsification standard.

Hypotheses are not models

- Models can correspond to multiple hypotheses, multiple hypotheses to a model
- All models are false, so what does it mean to falsify a model?

Progression: - Hypotheses - Statements - Process Models - Causal structure model that formalize cause/effect relationships - Statistical Model - Don't embody causal relationships; express associations among variables

How to get from statistical model back to process model? Derive a expected frequency distribution of some quantity - "statistic" from the causal model. Histogram, for example. Unfortunately, other models can imply the same statistical model, reinforcing the many-to-many relationships:

- Any statistical model can correspond to many process models

- Any hypothesis can correspond to multiple process models
- Any statistical model can correspond to multiple hypotheses

So what to do? If you have multiple process models that all make similar predictions, then you search for a description where the processes look different.

Measurement matters Logic of falsification is have hypothesis, look for observation. If not found, then hypothesis is false. If we formalize H_0 to "All swans are white", no number of observations can prove it, one observation disproves - powerful but prone to observation errors, and quantitative hypotheses have degrees of existence

Observation Error - Doubted observations, measurement/instrumentation error

Continuous Hypotheses - Not trying to disprove, but understand a distribution

Falsification is consensual Communities argue toward consensus about the meaning of evidence, can be messy

1.3 - Tools for Golem Engineering

If falsification isn't the way, we can model. Models then can be made to testing procedures, as well as designs, forecasts, and arguments. This text focuses on several tools: - Bayesian Data Analysis - Model comparison - Multilevel models - Graphical models

Bayesian data analysis Takes question in form of a model and uses logic to produce an answer in the form of probability distributions. Literally just counting how the data may look according to our assumptions. Compare to frequentest, which is defined by the frequencies of events in large samples, based on imaginary data resampling.

Bayesian approaches treat randomness as a property of information, not of the world.

Model comparison and prediction If multiple models, how to choose? Cross validation and information criteria.

Help in 3 ways: provide expectations of accuracy, give an estimate of overfitting, help spot influential observations

Multilevel Models Parameters all the way down - parameters support inference. Multiple levels of uncertainty feed into the next, a multilevel (hierarchical, random effects, varying effects, mixed effects) model. Help us with overfitting, by exploiting partial pooling, which can be used to adjust estimates for repeat sampling, imbalance, variation, and to avoid averaging.

Diverse models turn out to be multilevel: models for missing data (imputation), measurement, factor analysis, some time series models, and types of spatial and network regression are all special applications of multilevel.

Fitting and interpreting multilevel models can be harder than traditional

Graphical Causal Models Statistical models are association engines, detect (not infer!) association between cause and effect. Due to overfitting though, causally incorrect models can make better predictions than causally correct ones, can't focus on just prediction.

So instead, causal model that can be used to design one or more statistical models for causal identification. "Graphical Causal Model" represents a causal hypothesis, the most simple being a Directed Acyclic Graph (DAG).

1.4 - Summary

4 parts to book

1. Bayesian inference (ch 2,3)
2. Multiple linear regression (ch 4-9)
3. Generalized linear models, with MCMC and maximum entropy (ch 9-12)
4. Multilevel models, specialized models (ch 13-17)

Chapter 2 - Small Worlds and Large Worlds

Christopher Columbus thought the world was 10,000 km smaller than it is, his prediction made him think he could have enough supplies to circle it - contrast between model and reality.

Small world - self-contained logical world of model

Large World - larger context where model is deployed

2.1 - The Garden of Forking Data

Bayesian inference is really just counting and comparing possibilities. Cannot guarantee a correct answer, but can guarantee the best possible answer given information provided.

Use example of blue/white marbles out of bag, examine each path of pulls to exclude different "paths" through the data. Introduce a "prior," if all routes are equally likely, can base prior on the number of counts to an outcome. Update prior after a pull, or when new information is added (e.g. the company says blue is rare). The example outlines the following terms:

1. Conjectured proportion of marbles (p in example) is a "parameter" value
2. Number of ways a value can produce data is a "likelihood" - derived by enumerating all data sequences that could have happened and eliminate those inconsistent with data
3. Prior plausibility of p is the "prior probability"
4. Updated plausibility of p is the "posterior probability"

2.2 - Building a model

Design loop with 3 steps:

1. Data story: motivate model by narrating how the data might arise
2. Update: educate model by feeding it data
3. Evaluate: All models require supervision, possibly leading to model revision

Example in chapter - calculate the amount of water on earth by throwing globe up in the air, see if right thumb is on water or land.

Data Story May be *descriptive*, specifying associations to predict outcomes given observations; may be *causal*, a theory of how some events produce others. Generally causal stories are easily descriptive, descriptive stories may be hard to be causal. Can motivate by explaining how each piece of data is born. The value of the story is to more strongly define hypotheses and resolve ambiguities.

Example: true proportion of water is p, single toss has p chance of producing water W and $1 - p$ of land L, each toss is independent.

Bayesian Updating Model begins with one set of plausibilities assigned to each possibility. As data is collected, posteriors are produced.

Example: Give uniform prior. First case lands on water, $p = 0$ is excluded, since there is no longer a possibility of no water. Continue tossing and distribution shifts and closes more and more.

One benefit of Bayesian approach is that estimates are valid for any sample size. Of course better with more data.

Evaluate Because of differences between the model and real world, no guarantee of large world performance. Keep in mind two principles. Model's certainty is no guarantee the model is good - this is telling you that, given this model, plausible values are in some range. Second, it's important to supervise and critique the work - in the example, order of tosses shouldn't change final curve, but may indirectly affect it because data depends on order, so check on data it does not know about.

2.3 - Components of the Model

Variables Symbols that can take different values - for globe example: target p (proportion of water) cannot be observed. Unobserved are called "parameters." Other variables are count of water W and count of land L, these are observed.

Definitions In defining, we build a model relating variables to one another. For each value of unobserved, need to define the number of ways/probability that the values of each observed could arise. For each unobserved also need a prior.

Using example:

Observed Variables - For specific p, need to define how plausible combinations of W and L would be, using a mathematical function called a likelihood. In this case, if tosses are independent and probability are the same, we use binomial distribution.

$$\Pr(W, L|p) = \frac{(W+L)!}{W!L!} p^W (1-p)^L$$

In R:

```
dbinom( 6 , size=9 , prob=0.5)
```

```
## [1] 0.1640625
```

This gives the relative number of ways to get 6 water results for $p = 0.5$ after 9 total tosses ($N = W + L = 9$).

Unobserved Variables - Distributions for observed variables typically have own variables; p not observed, so a parameter. Many common data questions are answered directly by parameters, e.g. average difference between groups, association strength, covariate dependence, variation. For every parameter, you must define a prior.

Some schools of thought that emphasize choosing priors on personal belief, known as "subjective Bayesian." If you don't have a strong argument for any prior, try different ones.

Model is born Can summarize as the following:

$$W \sim \text{Binomial}(N, p) p \sim \text{Uniform}(0, 1)$$

Telling us W is binomial, and p is flat over the range 0 to 1.

2.4 - Making the Model Go

Once you have named all the variables, definitions, update prior to posterior - the relative plausibility of parameter values conditional on fdata and model, for our example $\Pr(p|W, L)$.

Bayes Theorem Mathematical definition of posterior arises from Bayes. Joint probability $\Pr(W, L, p) = \Pr(W, L|p)\Pr(p)$. The right side can also be reversed $\Pr(p|W, L)\Pr(W, L)$, which can be solved to

$$\Pr(p|W, L) = \frac{\Pr(W, L|p)\Pr(p)}{\Pr(W, L)}$$

This is Bayes theorem but can really just be said

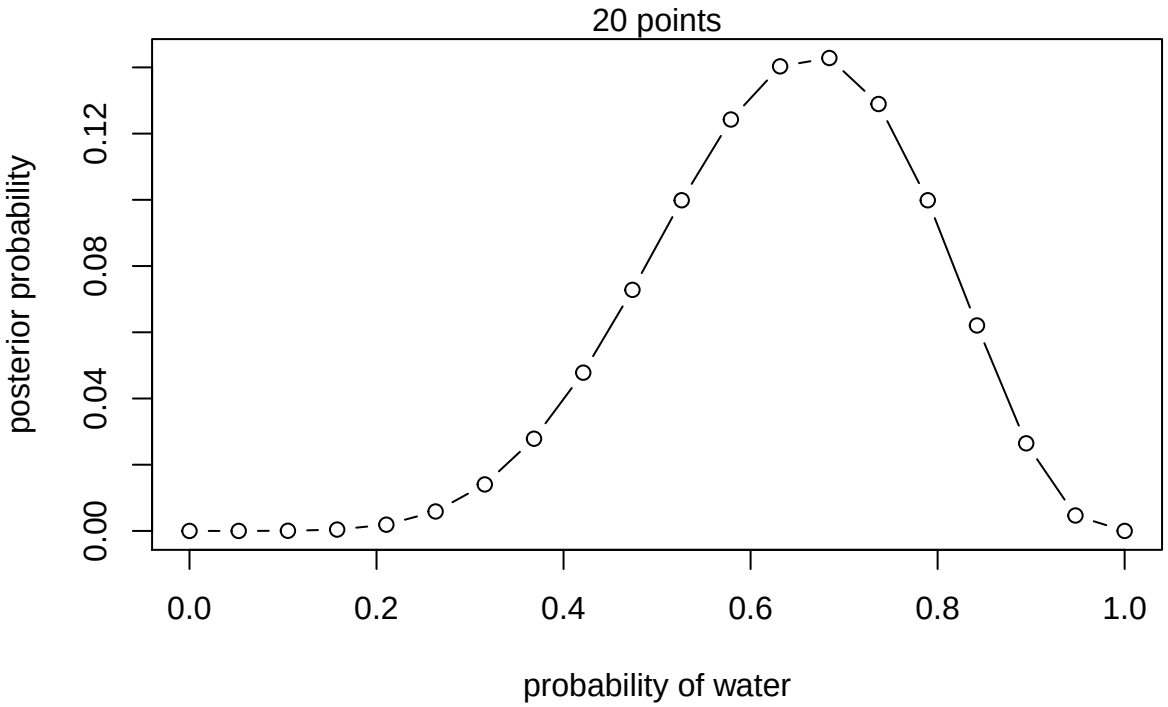
$$\text{Posterior} = \frac{\text{Probability of data} \times \text{Prior}}{\text{Average probability of the data}}$$

Denominator is averaged over the prior - meant to standardize the posterior to make teh sum one.

Three different numerical techniques for computing posterior: grid approximation, quadratic approximation, MCMC.

Grid Approximation - Consider a finite number of values, compute posterior by multiplying prior by likelihood, repeate until getting an apprximate picture of the posterior. Mostly a pedagogical tool, since not typically practical.

```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=20 )
# define prior
prior <- rep( 1 , 20 )
# compute likelihood at each value in grid
likelihood <- dbinom( 6 , size=9 , prob=p_grid )
# compute product of likelihood and prior
unstd.posterior <- likelihood * prior
# standardize the posterior, so it sums to 1
posterior <- unstd.posterior / sum(unstd.posterior)
plot(p_grid , posterior , type="b" ,
     xlab="probability of water" , ylab="posterior probability")
mtext( "20 points" )
```



Quadratic Approximation - More parameters make grid apprixmations tough (N^p for p parameters and N data points). Use quadratic approximation when the region near the peak of the posterior will be gaussian in shape, easy because it can be described by just mean and variance.

1. Find posterior mode
2. Estimate curvature, either analytically or computationally.

For this book use `quap` from rethinking programming package

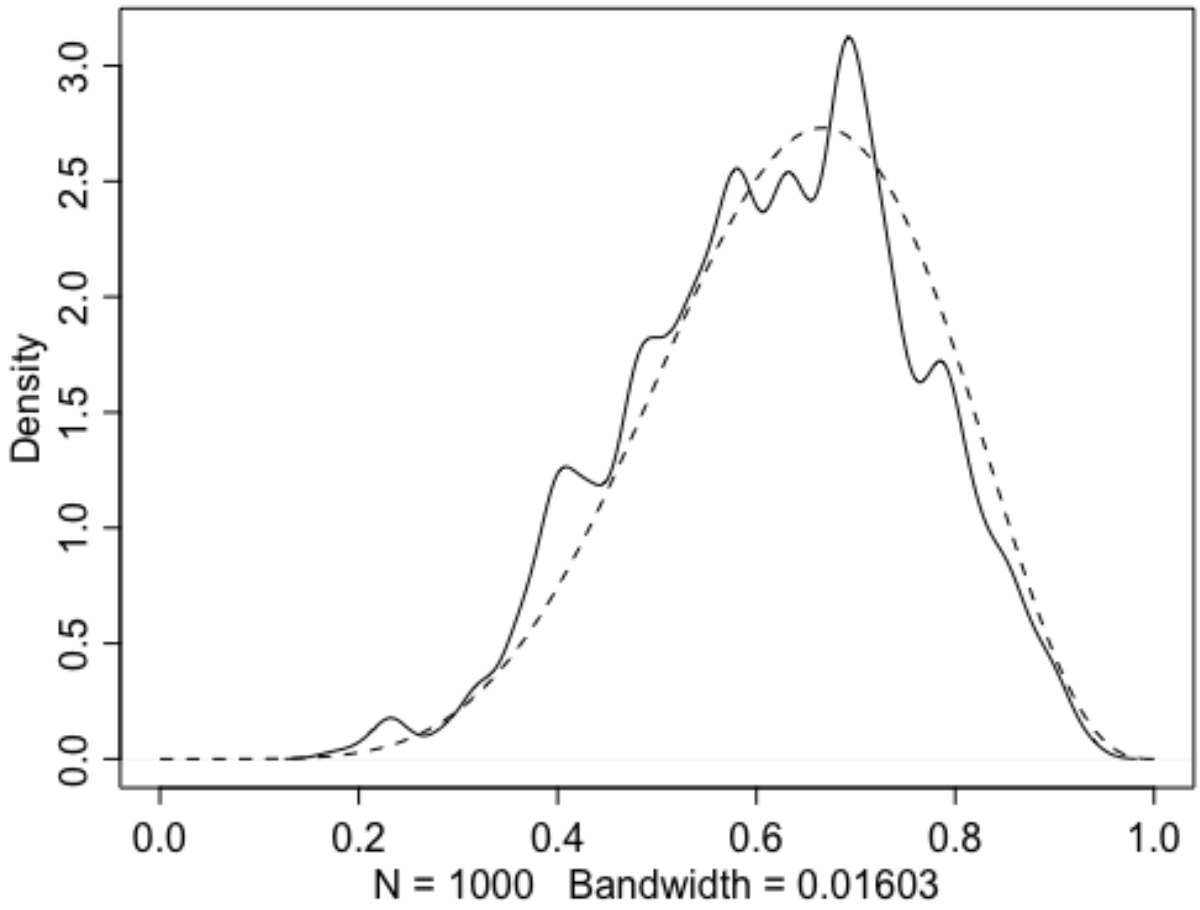
```
library(rethinking)
globe.qa <- quap(
  alist(
    W ~ dbinom (W+L, p), # Binomial
    p ~ dunif(0,1) # Uniform
  ), data = list(W=6, L=3)
)
precis(globe.qa)
```

```
##        mean        sd      5.5%     94.5%
## p 0.6666666 0.1571338 0.4155364 0.9177967
```

The quadratic approximation is often equivalent to a Maximum Likelihood Estimate and its standard error.

note - quadratic is solved by computing the Hessian, a square matrix of second derivatives of the log of posterior probability wrt parameters. Derivatives sufficient to describe a gaussian. Std is typically computed from Hessian, which can occassionally cause problems in computation.

Markov chain Monte Carlo (MCMC) - Many models, like multilevel/mixed-effects don't work for grid approximation (many parameters) or quadratic (non-gaussian posterior). Function to maximize isn't known, computed in pieces via MCMC. Rather than computing or approximating posterior, MCMC draws samples, a collection of parameter values.



2.5 - Summary

Looked at conceptual ideas in Bayesian data analysis. Models are composite of variables and distributional definitions, fit to data using numerical techniques

Problem sets

Chapter 2 problems

2E1. Which of the expressions below correspond to the statement:the probability of rain on Monday?

1. Pr(rain)
2. **Pr(rain|Monday)**
3. Pr(Monday|rain)
4. Pr(rain,Monday)/Pr(Monday)

2E2. Which of the following statements corresponds to the expression: Pr(Monday|rain)?

1. The probability of rain on Monday.

2. The probability of rain, given that it is Monday.
3. **The probability that it is Monday, given that it is raining.**
4. The probability that it is Monday and that it is raining.

2E3. Which of the expressions below correspond to the statement: the probability that it is Monday, given that it is raining?

1. **Pr(Monday|rain)**
2. Pr(rain|Monday)
3. Pr(rain|Monday)Pr(Monday)
4. Pr(rain|Monday)Pr(Monday)/Pr(rain)
5. Pr(Monday|rain)Pr(rain)/Pr(Monday)

2E4. The Bayesian statistician Bruno de Finetti (1906–1985) began his book on probability theory with the declaration: "PROBABILITY DOES NOT EXIST." The capitals appeared in the original, so I imagine de Finetti wanted us to shout this statement. What he meant is that probability is a device for describing uncertainty from the perspective of an observer with limited knowledge; it has no objective reality. Discuss the globe tossing example from the chapter, in light of this statement. What does it mean to say "the probability of water is 0.7"?

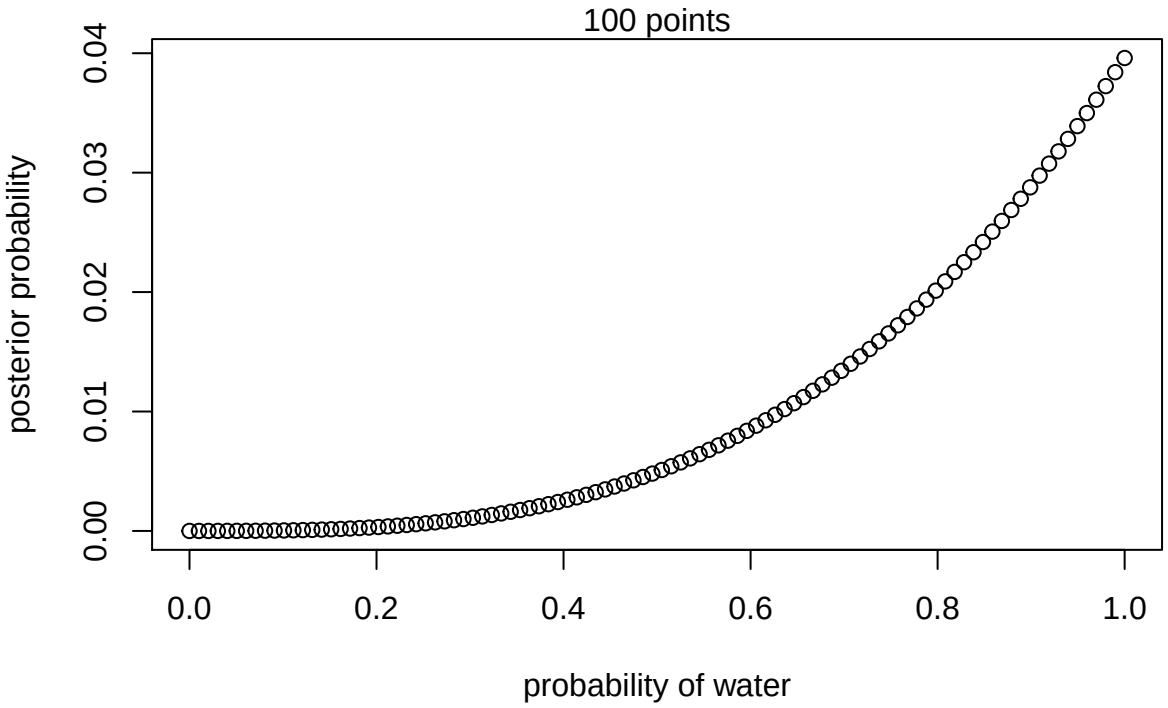
Based on the tosses that we've performed, on the globe we hold, we can expect 70% of future tosses to also land on water. The source of the uncertainty and limited knowledge is that the point we're landing on is random.

2M1. Recall the globe tossing model from the chapter. Compute and plot the grid approximate posterior distribution for each of the following sets of observations. In each case, assume a uniform prior for p.

```
globe_grid <- function(trial_list, grid_size){
  # define grid
  p_grid <- seq( from=0 , to=1 , length.out=grid_size )
  # define prior
  prior <- rep( 1 , grid_size )
  # compute likelihood at each value in grid
  likelihood <- dbinom( sum(trial_list) , size=length(trial_list) , prob=p_grid )
  # compute product of likelihood and prior
  unstd.posterior <- likelihood * prior
  # standardize the posterior, so it sums to 1
  posterior <- unstd.posterior / sum(unstd.posterior)
  plot(p_grid , posterior , type="b" ,
       xlab="probability of water" , ylab="posterior probability")
  mtext( sprintf("%i points", grid_size ))
}
```

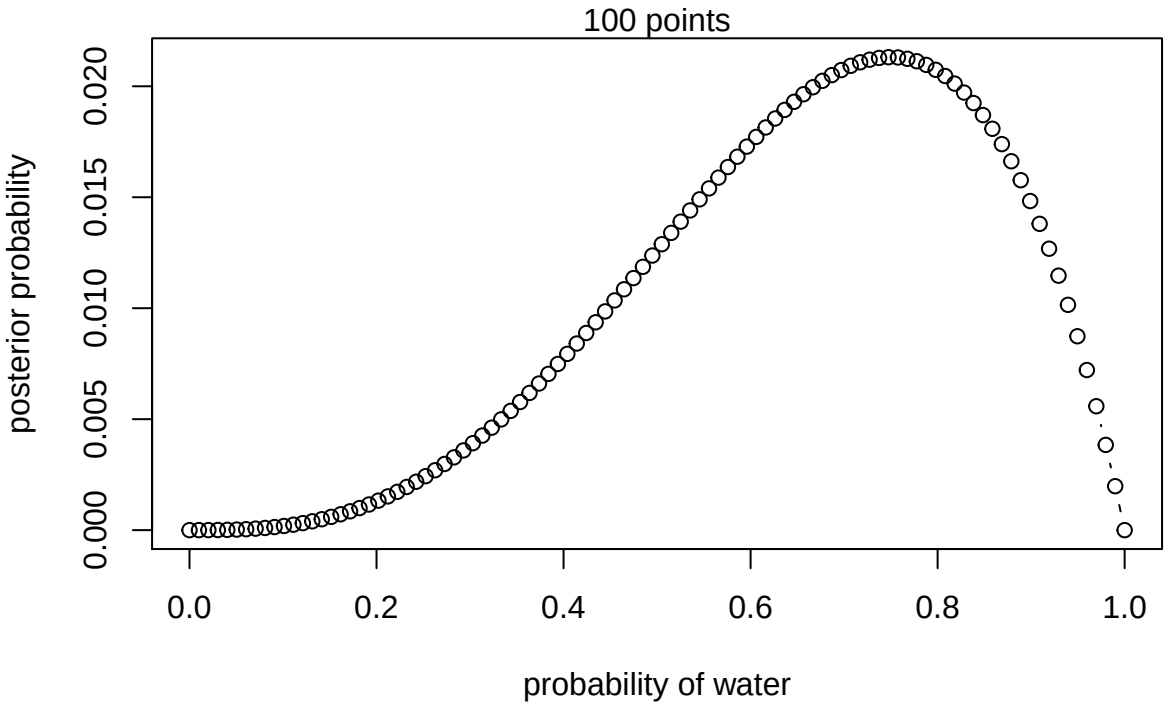
1. W, W, W

```
globe_grid(c(1,1,1),100)
```



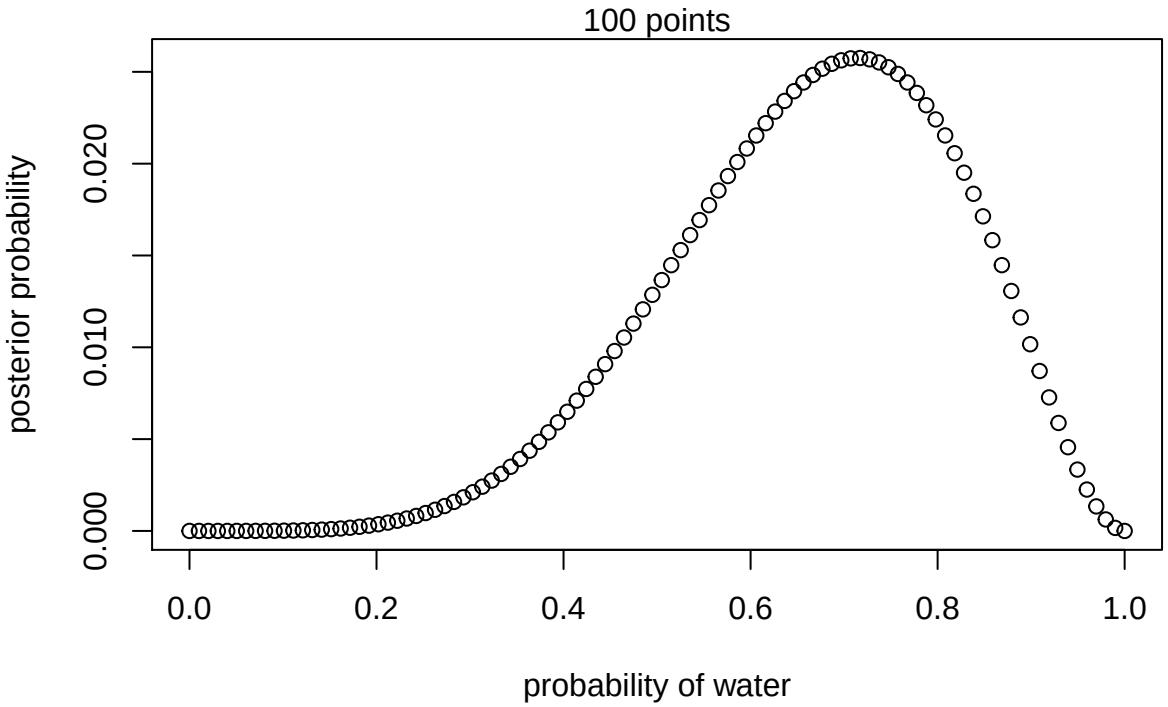
2. W, W, W, L

```
globe_grid(c(1,1,1,0),100)
```



3. L,W,W,L,W,W,W

```
globe_grid(c(0,1,1,0,1,1,1),100)
```



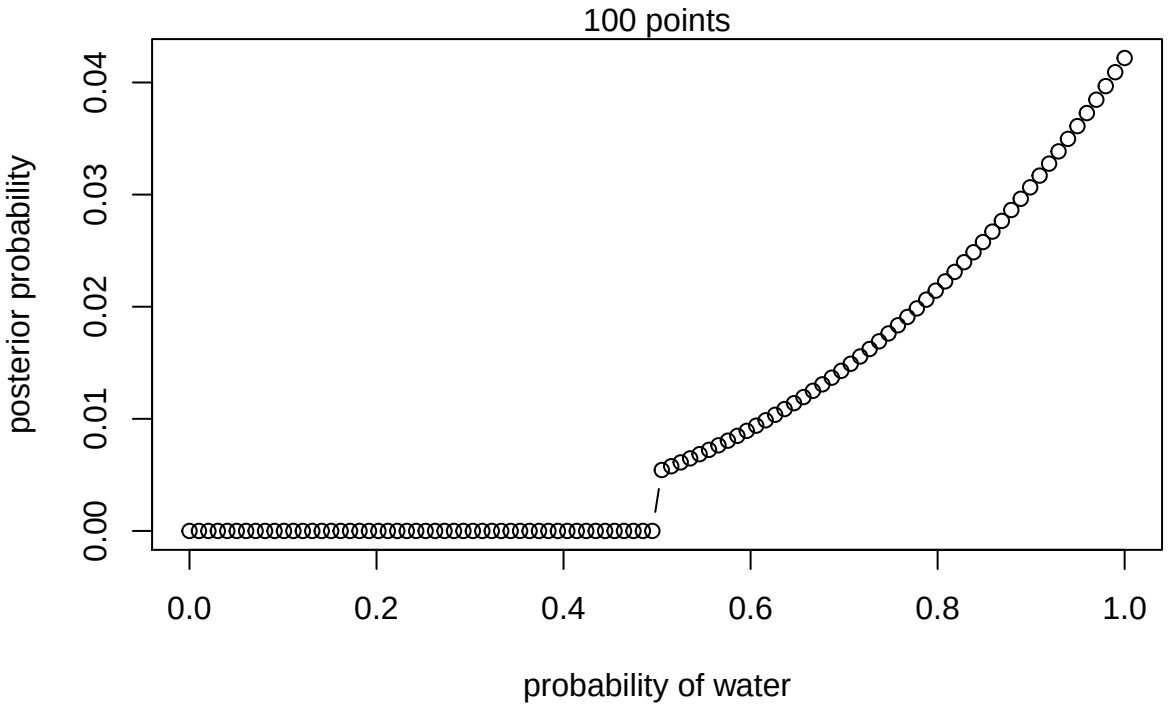
2M2. Now assume a prior for p that is equal to zero when $p < 0.5$ and is a positive constant when $p \geq 0.5$. Again compute and plot the grid approximate posterior distribution for each of the sets of observations in the problem just above.

Change function prior definition to:

```
# define prior
prior <- (p_grid >= 0.5) * prior_const
```

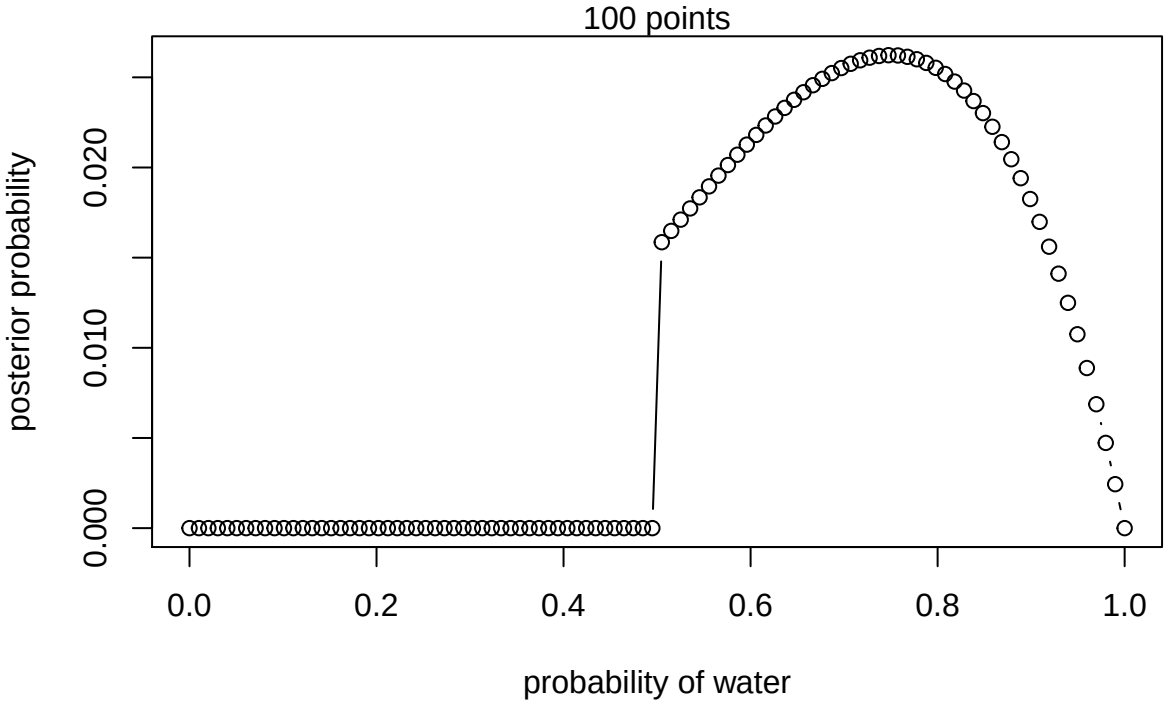
1. W, W, W

```
globe_grid_const(c(1,1,1),100, 1.0)
```



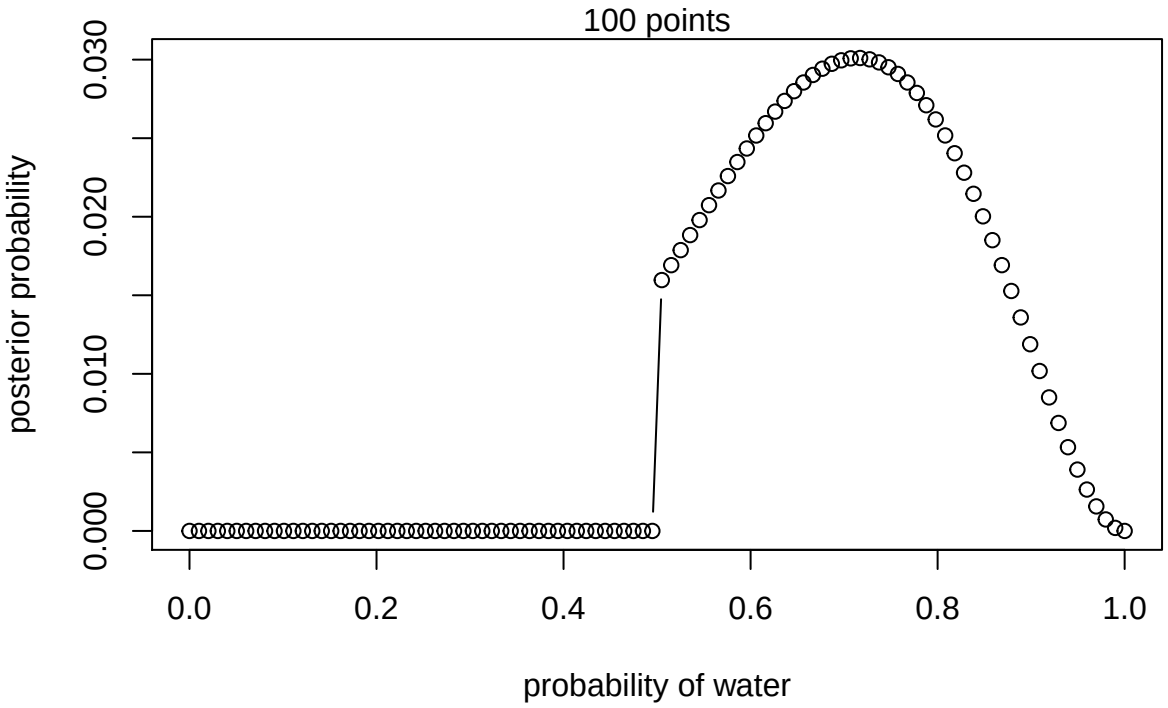
2. W, W, W, L

```
globe_grid_const(c(1,1,1,0),100, 1.0)
```



3. L,W,W,L,W,W,W

```
globe_grid_const(c(0,1,1,0,1,1,1),100,1)
```


100 points
posterior probability
0.000
0.010
0.020
0.030
0.0
0.2
0.4
0.6
0.8
1.0
probability of water