Chapter 3 – Sampling The Imaginary

3.1 Sampling from a grid-approximate posterior

- R Code 3.2:

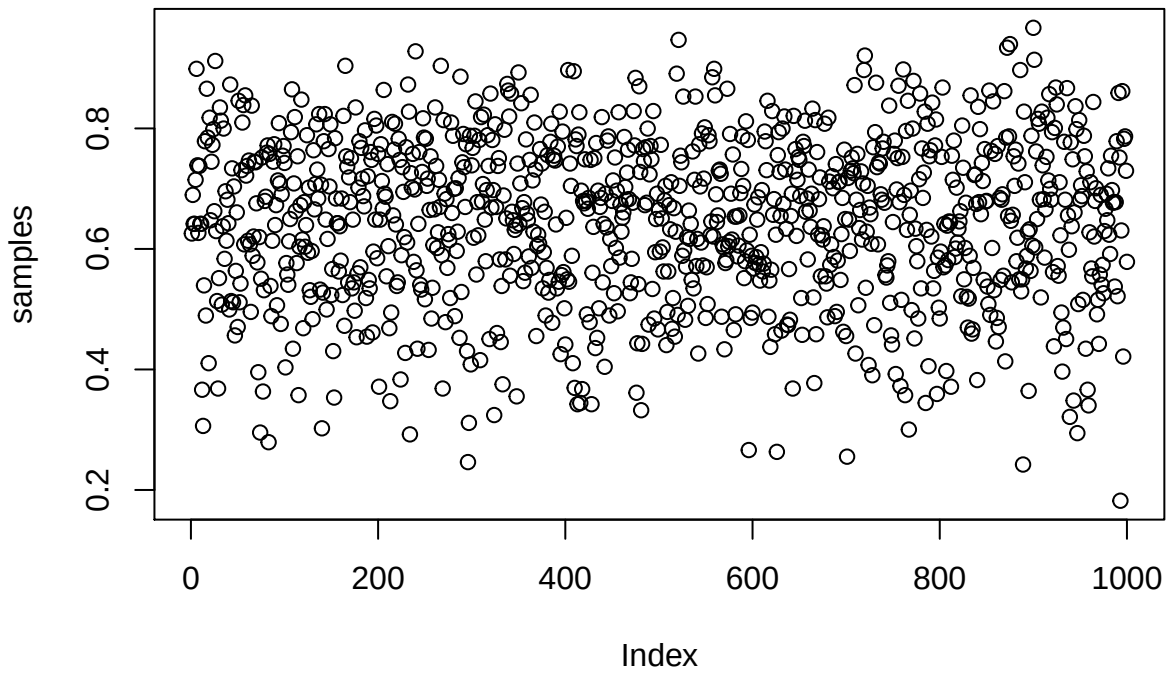
```
n = 1000
p_grid <- seq(from=0, to=1, length.out=n)
prior <- rep(1, n)
likelihood <- dbinom(x=6, size=9, prob=p_grid)
posterior_notnorm <- likelihood * prior
posterior <- posterior_notnorm / sum(posterior_notnorm)
```

Draw 10,000 samples: * R Code 3.3:

```
samples_orig <- sample(p_grid, prob=posterior, size=n, replace=T)
samples = samples_orig
```

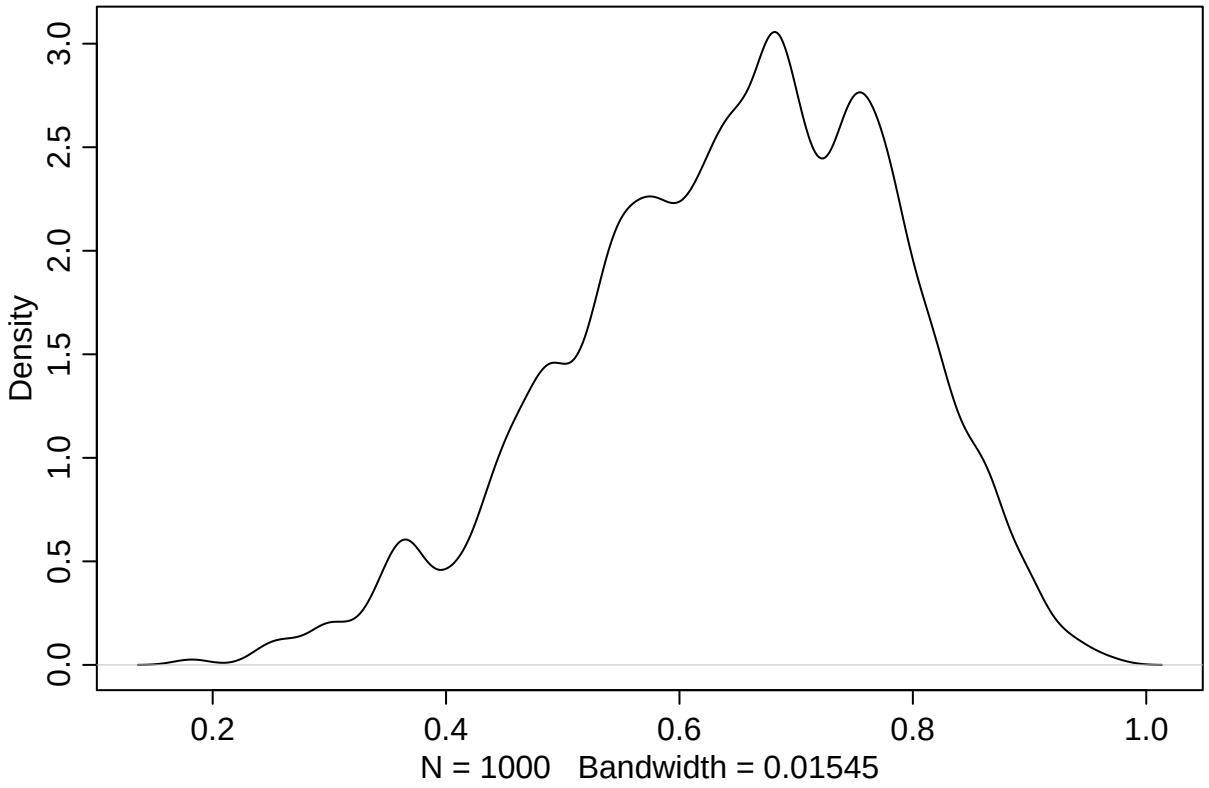
- 3.4:

```
plot(samples)
```



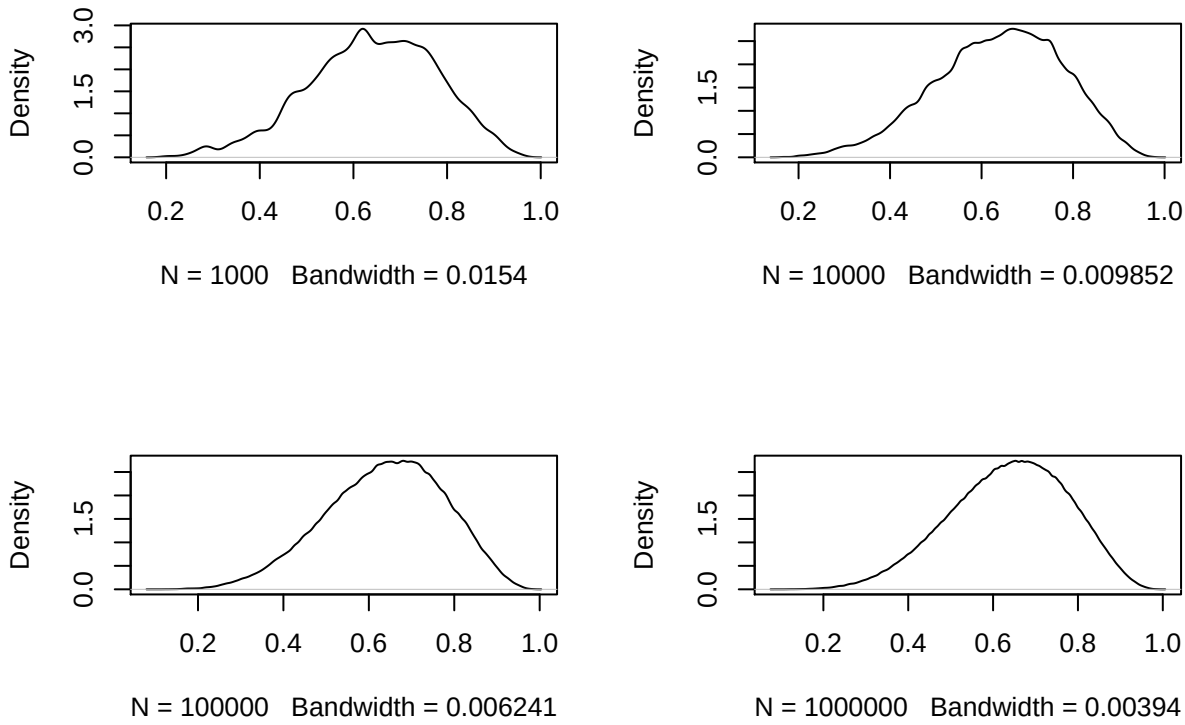
- 3.5:

```
dens(samples)
```



Let's try more samples:

```
par(mfrow=c(2, 2))
dens(sample(p_grid, prob=posterior, size=1e3, replace=T))
dens(sample(p_grid, prob=posterior, size=1e4, replace=T))
dens(sample(p_grid, prob=posterior, size=1e5, replace=T))
dens(sample(p_grid, prob=posterior, size=1e6, replace=T))
```



3.2 Sampling to Summarize

3.2.1. Intervals of defined boundaries.

The posterior probability that the proportion of water is less than 0.5:

- 3.6:

```
p_grid < 0.5
```

```
##    [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [12]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [23]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [78]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   [89]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [100]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [111]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [122]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [133]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [144]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [155]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [166]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
##  [177]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [188]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [199]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [210]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [221]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [232]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [243]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [254]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [265]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [276]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [287]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [298]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [309]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [320]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [331]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [342]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [353]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [364]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [375]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [386]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [397]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [408]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [419]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [430]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [441]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [452]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [463]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [474]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [485]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [496]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##  [507] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [518] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [540] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [551] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [562] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [573] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [584] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [595] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [606] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [617] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [628] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [639] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [650] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [661] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [672] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [683] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [694] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [705] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [716] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [727] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [738] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [749] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [760] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##  [771] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [782] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [804] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [815] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [826] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [837] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [848] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [859] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [870] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [881] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [892] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [903] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [914] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [936] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [947] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [958] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [969] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [980] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [991] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
sum(posterior[p_grid < 0.5])
```

```
## [1] 0.1718746
```

Samples array:

```
head(samples, 100)
```

```
##   [1] 0.6256256 0.6896897 0.6416416 0.6416416 0.7147147 0.8988989 0.7387387
##   [8] 0.6266266 0.7377377 0.6416416 0.6416416 0.3663664 0.3063063 0.5395395
##  [15] 0.7787788 0.4894895 0.8658659 0.7847848 0.4104104 0.8178178 0.6476476
##  [22] 0.7447447 0.7727728 0.7987988 0.6626627 0.9119119 0.6296296 0.5135135
##  [29] 0.3683684 0.5515516 0.8348348 0.8128128 0.5075075 0.6376376 0.7997998
##  [36] 0.5835836 0.6956957 0.6136136 0.6816817 0.6426426 0.4994995 0.8728729
##  [43] 0.5135135 0.7337337 0.5115115 0.7037037 0.4564565 0.5635636 0.6606607
##  [50] 0.4704705 0.8458458 0.5115115 0.5425425 0.7307307 0.8098098 0.8388388
##  [57] 0.6076076 0.8548549 0.7217217 0.7377377 0.6136136 0.7467467 0.6076076
##  [64] 0.4954955 0.8378378 0.5885886 0.6206206 0.7447447 0.7437437 0.6756757
##  [71] 0.5775776 0.3953954 0.6196196 0.2952953 0.5505506 0.7517518 0.3633634
##  [78] 0.5315315 0.6796797 0.6866867 0.7597598 0.7717718 0.2792793 0.7577578
##  [85] 0.5385385 0.4884885 0.6136136 0.7447447 0.7737738 0.6396396 0.6726727
##  [92] 0.5065065 0.7137137 0.8088088 0.7087087 0.4754755 0.6906907 0.7547548
##  [99] 0.7707708 0.7427427
```

The same calculation using samples. Add up all samples that lie in the grid < 0.5, and divide by the total number of samples to get the frequency $\sim$ probability:

- 3.7:

```
n = 1e4
samples = sample(p_grid, prob=posterior, size=n, replace=T)
sum(samples < 0.5) / n
```

```
## [1] 0.1728
```

How much probability lies between 0.5 and 0.75: * 3.8:

```
sample_points = sum(samples > 0.5 & samples < 0.75)
sample_points
```

```
## [1] 6070
```

```
sample_points / n
```

```
## [1] 0.607
```

3.2.2. Intervals of defined mass.

Boundaries of the lower 80% posterior probability lies:

- 3.9:

```
quantile(samples, probs = .8)
```

```
##       80%
## 0.7587588
```

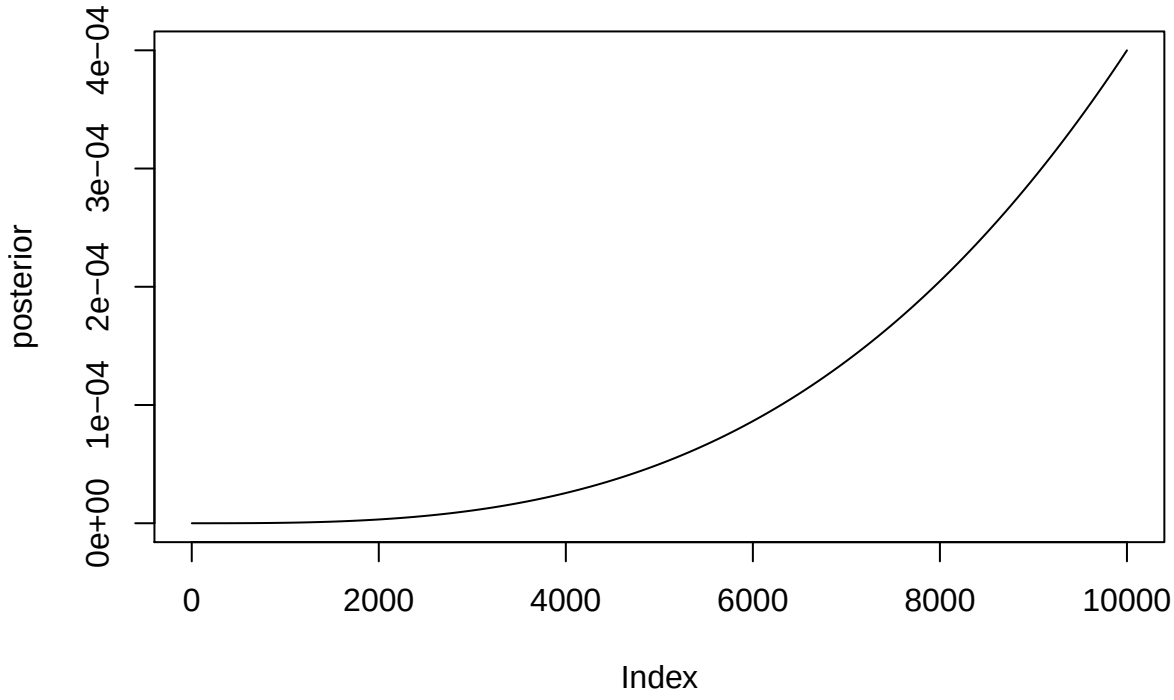
Middle 80%, i.e. lying between 10% and 90%:

```
# 3.10
quantile(samples, probs = c(0.1, 0.9))
```

```
##       10%       90%
## 0.4494494 0.8108108
```

The above are PERCENTILE INTERVALS. Percentiles can be misleading if the distribution is highly skewed.

```
# 3.11
n <- 10000
p_grid <- seq(0, 1, length.out = n)
prior <- rep(1, n)
likelihood <- dbinom(3, size=3, prob=p_grid)
posterior_notnorm <- likelihood * prior
posterior <- posterior_notnorm / sum(posterior_notnorm)
samples <- sample(p_grid, size=1e5, replace=T, prob=posterior)
plot(posterior, type='l')
```




```
# 3.12
PI(samples, prob=0.5)
```

```
##       25%       75%
## 0.7063706 0.9306931
```

Highest Posterior Density Interval described the distribution better. It's the *narrowest* interval containing the specified probability mass, e.g. 50%.

```
# 3.13
HPDI(samples, prob=0.5)
```

```
##      |0.5      0.5|
## 0.8405841 1.0000000
```

3.2.3. Point Estimates

A parameter with the highest posterior probability is called *a maximum a posteriori* estimate, or *MAP*.

```
# 3.14
which.max(posterior)
```

```
## [1] 10000
```

```
p_grid[which.max(posterior)]
```

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

Use samples to get the same (or similar) result:

```
# 3.15
chainmode(samples, adj=0.01)
```

```
## [1] 0.996744
```

```
# 3.16
mean(samples)
```

```
## [1] 0.7999124
```

```
median(samples)
```

```
## [1] 0.8405841
```

If the loss function is the absolute difference, then the posterior loss for p = 0.5 is

```
# 3.17
sum(posterior * abs(0.5 - p_grid))
```

```
## [1] 0.3125375
```

```
# 3.18
loss <- sapply(p_grid, function(d) sum(posterior * abs(d - p_grid)))
```

```
# 3.19
which.min(loss)
```

```
## [1] 8410
```

```
p_grid[which.min(loss)]
```

```
## [1] 0.8409841
```

The posterior median minimizes the abs loss function. Let's test the quadratic loss function:

```
loss2 <- sapply(p_grid, function(d) sum(posterior * (d - p_grid)^2))
which.min(loss2)
```

```
## [1] 8001
```

```
p_grid[which.min(loss2)]
```

```
## [1] 0.80008
```

This is a mean.

3.3. Sampling to Simulate Prediction

3.3.1. Dummy Data

```
# 3.20
dbinom(0:2, size=2, prob=0.7)
```

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

We can sample from this distribution:

```
# 3.22
rbinom(10, size=2, prob=0.7)
```

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

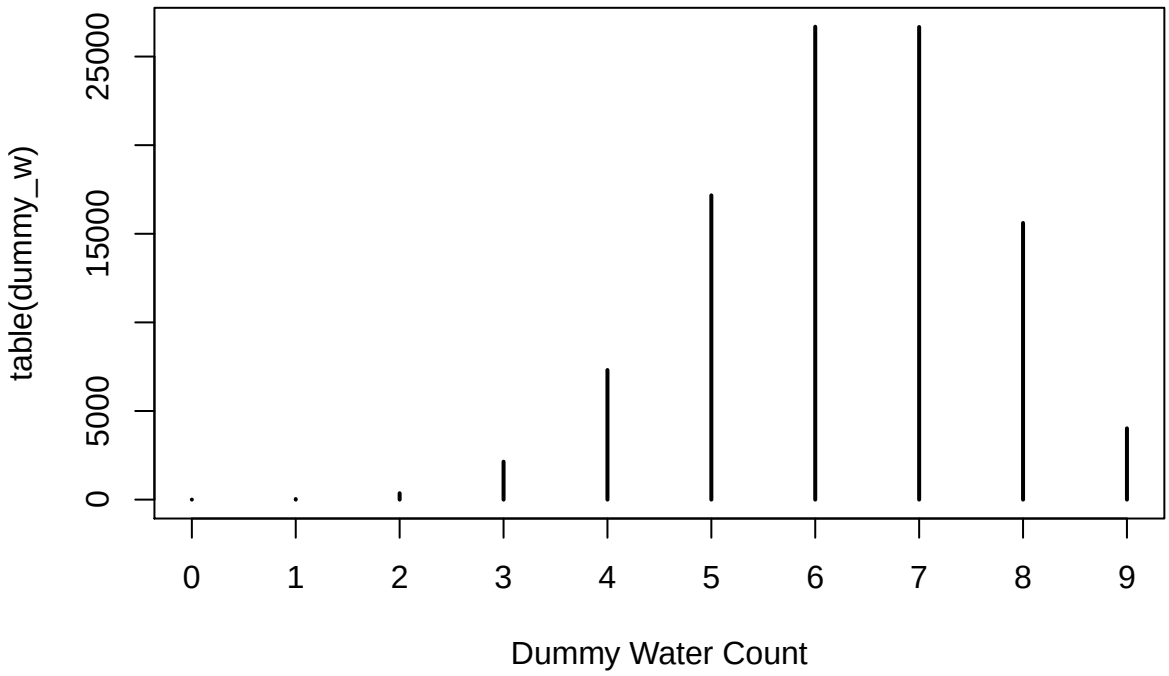
Let's generate 100,000 dummy observations to verify that each values 0, 1, and 2 appear in proportion to its likelihood:

```
# 3.23
dummy_w <- rbinom(1e5, size=2, prob=0.7)
table(dummy_w) / 1e5
```

```
## dummy_w
##       0       1       2
## 0.08909 0.41902 0.49189
```

Let's simulate the sample with 9 tosses:

```
# 3.24
dummy_w <- rbinom(1e5, size=9, prob=0.7)
plot(table(dummy_w), xlab="Dummy Water Count")
```




```
table(dummy_w)
```

```
## dummy_w
##     0     1     2     3     4     5     6     7     8     9
##     3    32   361  2140  7313 17169 26684 26668 15609  4021
```

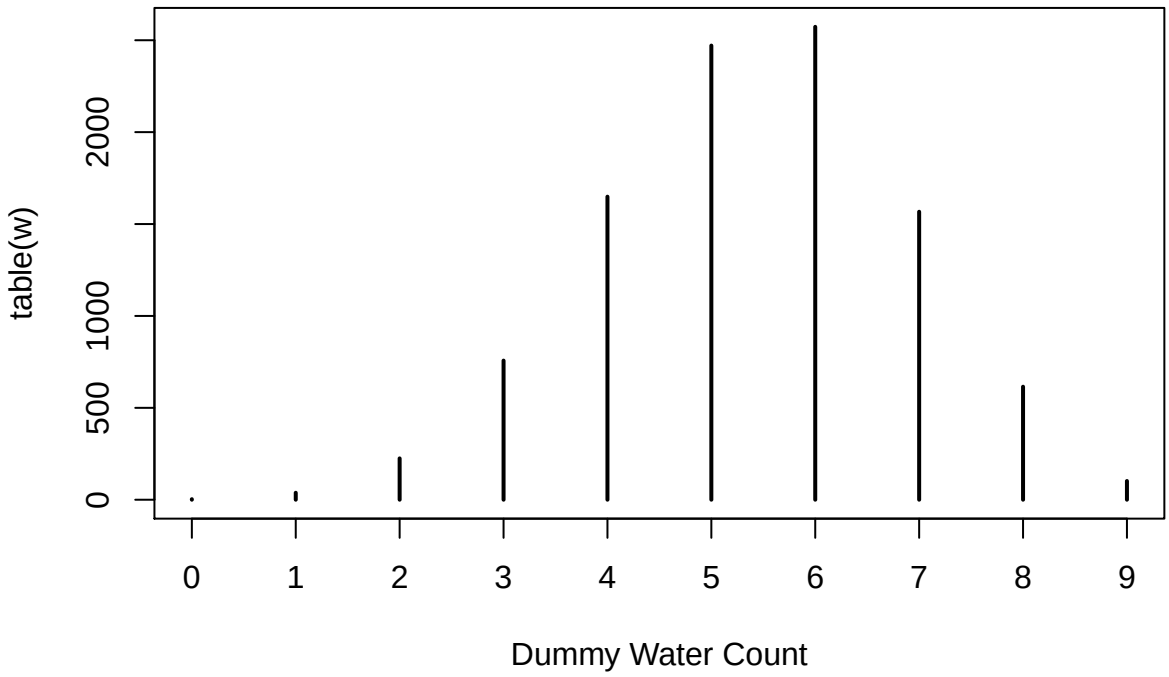
```
dummy_w[1:100]
```

```
##   [1] 6 7 5 8 9 8 7 5 5 6 6 6 7 8 6 7 7 6 8 7 6 6 7 5 6 8 5 5 7 5 4 6 6 8 6
##  [36] 4 5 6 6 6 7 7 5 7 4 7 8 6 5 4 8 5 4 3 7 5 7 8 6 5 9 3 7 6 9 6 8 5 6 7
##  [71] 3 9 7 4 8 7 6 7 5 7 7 7 9 6 6 8 5 6 5 7 5 5 5 6 6 7 6 6 5 6
```

3.2. Model Checking

Below is a misleading distribution plot. While p = 0.6 is the likeliest estimate, if we simply use it as a point estimate we will obtain a much more narrow, "overly confident" predictions:

```
# 3.25
w <- rbinom(1e4, size=9, prob=0.6)
plot(table(w), xlab="Dummy Water Count")
```



The correct way to generate predictions is to incorporate our uncertainty about p. We can do it by using sampled values of p, and averaging over all of them. The sampled values will appear with the right frequency, described by our posterior. Samples from the *posterior* distribution of p:

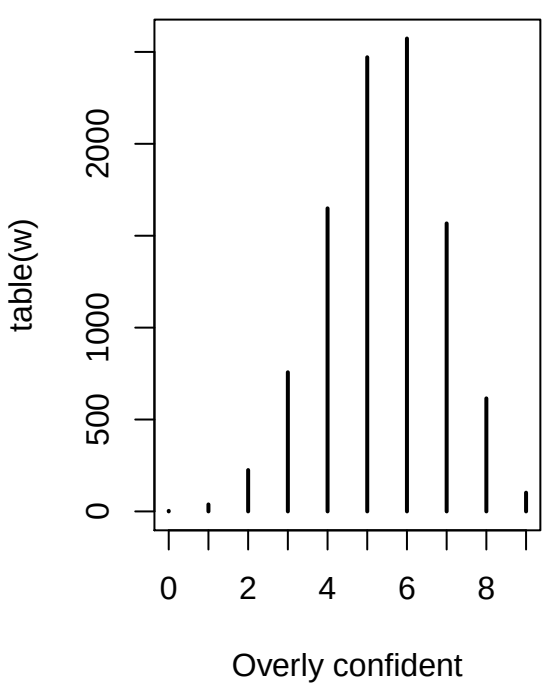
```
samples[1:20]
```

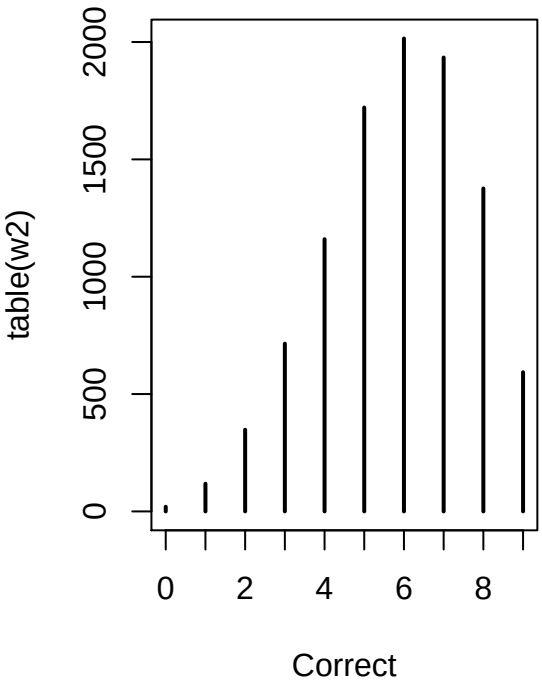
```
##  [1] 0.9985999 0.9157916 0.9017902 0.8860886 0.9622962 0.2513251 0.3799380
##  [8] 0.8337834 0.5165517 0.9456946 0.7757776 0.9045905 0.4982498 0.9819982
## [15] 0.9124912 0.8585859 0.8111811 0.8829883 0.9723972 0.5428543
```

```
# 3.26
w2 <- rbinom(1e4, size=9, prob=samples_orig)
table(w2)
```

```
## w2
##    0    1    2    3    4    5    6    7    8    9
##   20  118  348  715 1160 1721 2015 1934 1376  593
```

```
par(mfrow=c(1, 2))
plot(table(w), xlab ="Overly confident")
plot(table(w2), xlab="Correct")
```


table(w)
0
500
1000
2000
0
2
4
6
8
Overly confident


table(w2)
0
500
1000
1500
2000
0
2
4
6
8
Correct