Bayesian Estimation

Please draw your own subjective distributions for the following events/items.

1. The probability that it will snow at Reed this winter.
2. The probability that, on a given night, the sun has gone supernova.
3. The total number of individual socks that you own.

Karl Broman's Socks

Classical H test

Assert a model

H_0: I have N_{pairs} pairs of socks and $N_{singles}$ singletons. The first 11 socks that I pull out of the machine are a random sample from this population.

Decide on a test statistic

The number of singletons in the sample: 11.

Construct the sampling distribution

Probability theory or simulation.

See where your observed stat lies in that distribution

Find the p-value if you like.

H_0

$$N_{pairs} = 9$$

H_0

$$N_{pairs} = 9; \quad N_{singles} = 5$$

Contructing the sampling dist.

We'll use simulation.

Create the population of socks:

```
sock_pairs <- c("A", "B", "C", "D", "E",
                "F", "G", "H", "I", "J", "K")
sock_singles <- c("l", "m", "n", "o", "p")
socks <- c(rep(sock_pairs,
               each = 2),
           sock_singles)
```

```
socks
```

```
##  [1] "A" "A" "B" "B" "C" "C" "D" "D" "E" "E" "F" "F" "
## [18] "I" "J" "J" "K" "K" "l" "m" "n" "o" "p"
```

One draw from the machine

```
picked_socks <- sample(socks, size = 11, replace = FALSE)
picked_socks
```

```
##  [1] "J" "A" "I" "D" "F" "m" "A" "H" "G" "o" "C"
```

```
sock_counts <- table(picked_socks)
sock_counts
```

```
## picked_socks
## A C D F G H I J m o
## 2 1 1 1 1 1 1 1 1 1
```

```
n_singles <- sum(sock_counts == 1)
n_singles
```

```
## [1] 9
```

Our simulator

Constructing the sampling dist.

```
pick_socks(N_pairs = 9,N_singles = 5,
           N_pick = 11)
```

```
## [1] 11
```

```
pick_socks(9, 5, 11)
```

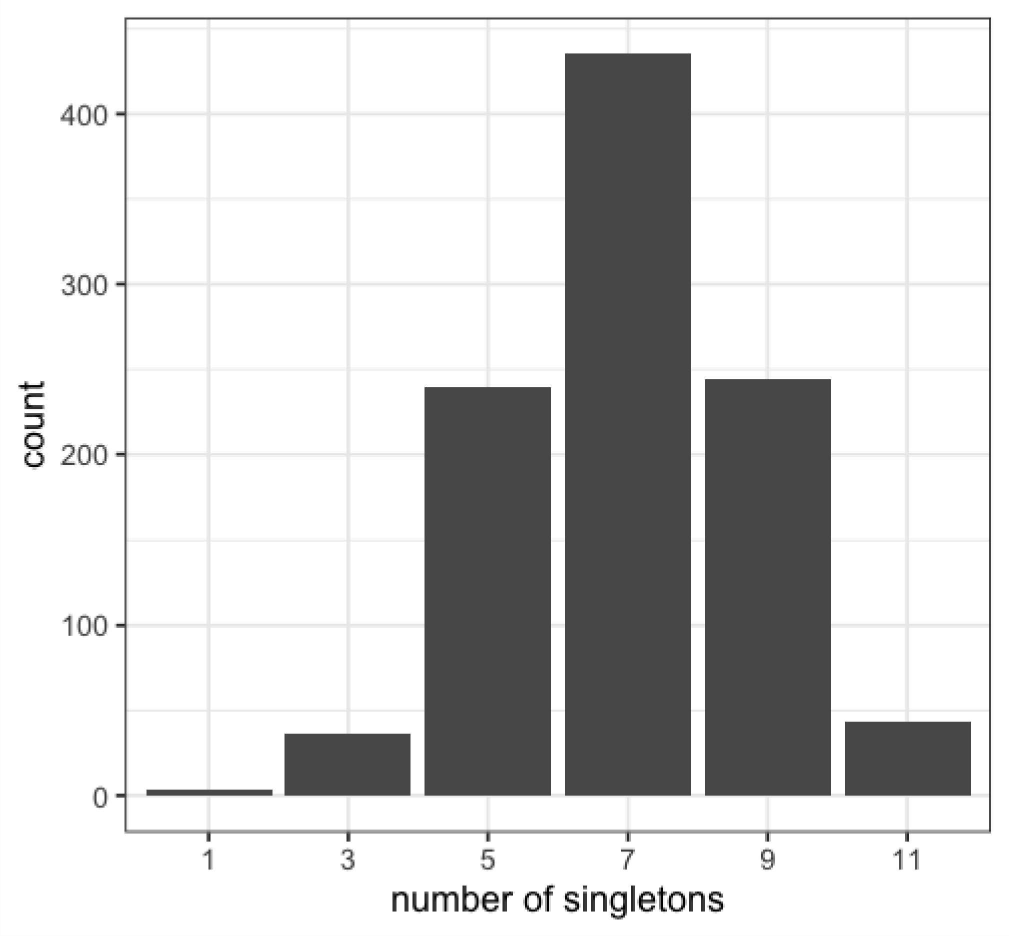
```
## [1] 7
```

```
pick_socks(9, 5, 11)
```

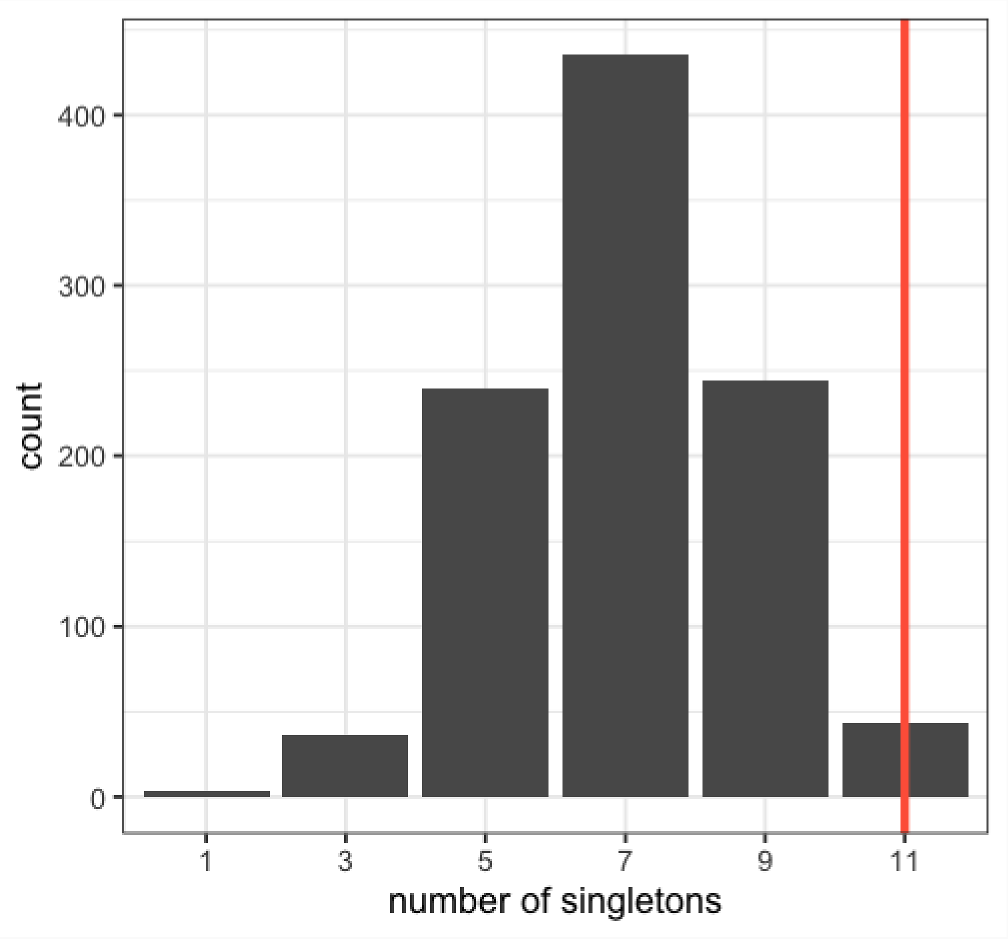
```
## [1] 7
```

Repeat many, many times...

The sampling distribution



The sampling distribution



The p-value

Quantifying how far into the tails our observed count was.

```r
table(sim_singles)
```

```
## sim_singles
##   1   3   5   7   9  11
##   3  36 239 435 244  43
```

```r
table(sim_singles)[6]/1000
```

```
##    11
## 0.043
```

Our two-tailed p-value is 0.086.

Question

What is the best definition for our one-tailed p-value in probability notation?

1. P(H_0 is true | data) = 0.043
2. P(H_0 is false | data) = 0.043
3. P(H_0 is true) = 0.043
4. P(data | H_0 is true) = 0.043
5. P(data) = 0.043

Question

What is the best definition for our one-tailed p-value in probability notation?

1. P(H_0 is true | data) = 0.043
2. P(H_0 is false | data) = 0.043
3. P(H_0 is true) = 0.043
4. **P(data | H_0 is true) = 0.043**
5. P(data) = 0.043

The challenge with the classical method

The result of a hypothesis test is a probability of the form:

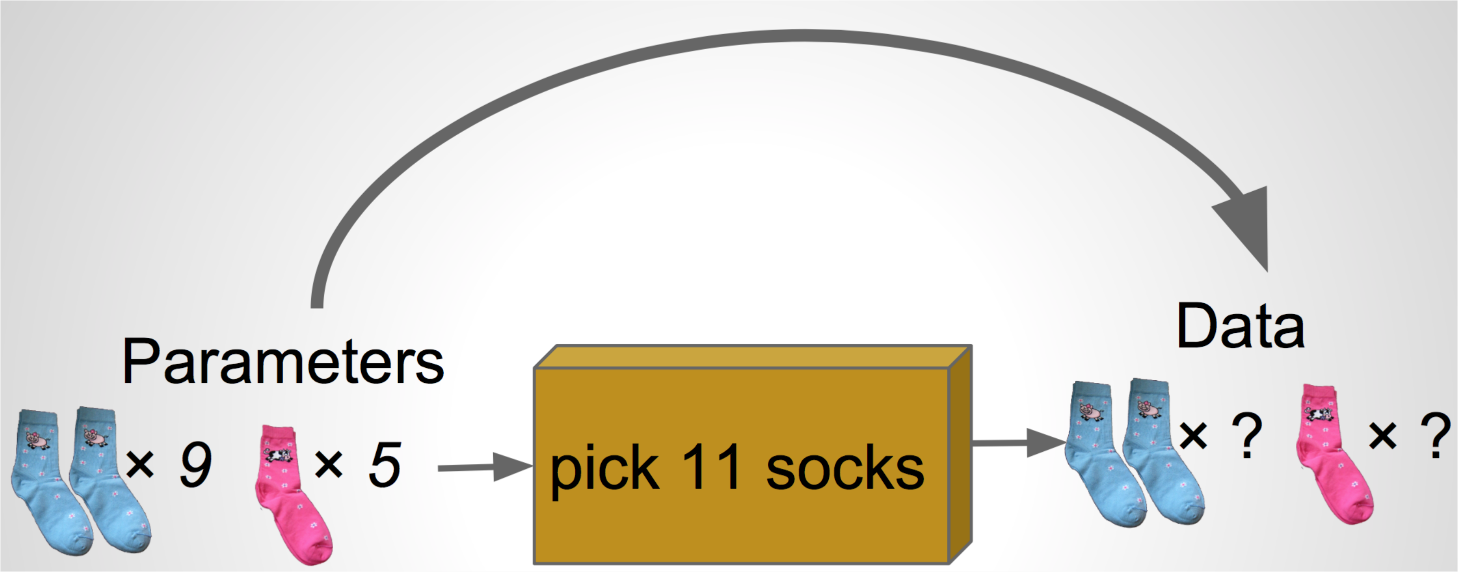
$$P(\text{ data or more extreme} \mid H_0 \text{ true })$$

while most people *think* they're getting

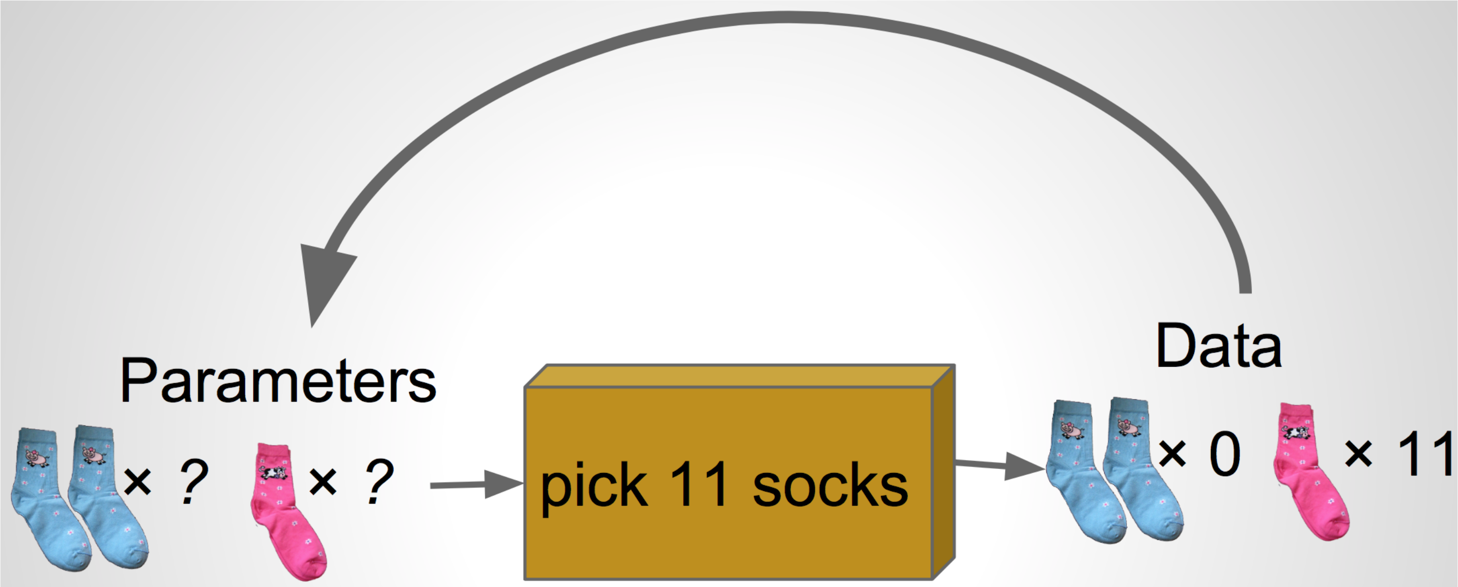
$$P(\, H_0 \text{ true} \mid \text{data })$$

How can we go from the former to the latter?

What we have



What we want



Bayesian modeling via Bayes' rule

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$P(model \mid data) = \frac{P(data \mid model)\, P(model)}{P(data)}$$

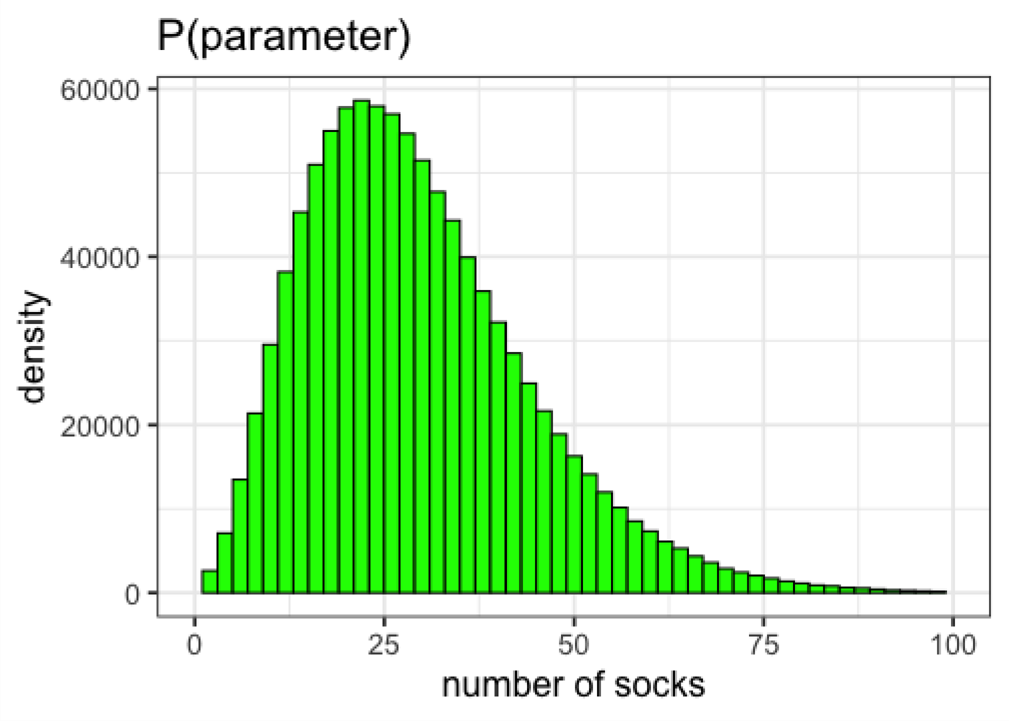
What does it mean to think about $P(model)$?

Please draw your own subjective distributions for the following events/items.

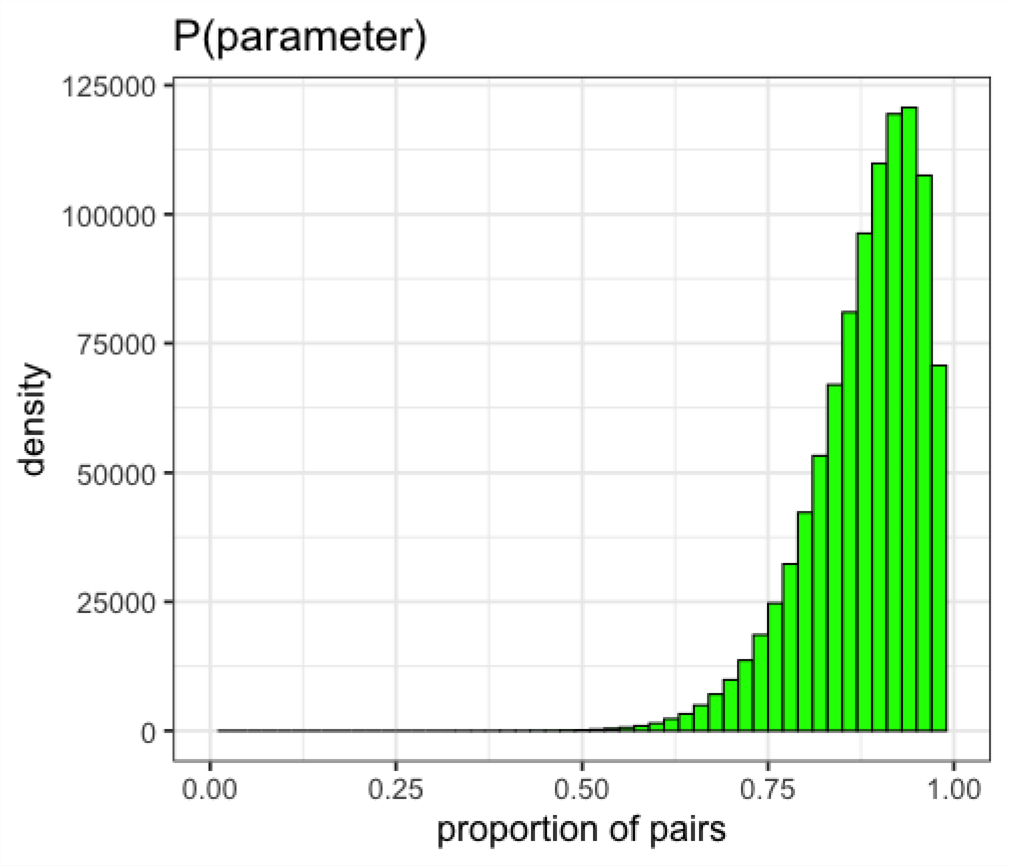
1. The probability that it will snow at Reed this winter.
2. The probability that, on a given night, the sun has gone supernova.
3. The total number of individual socks that you own.

Prior distribution

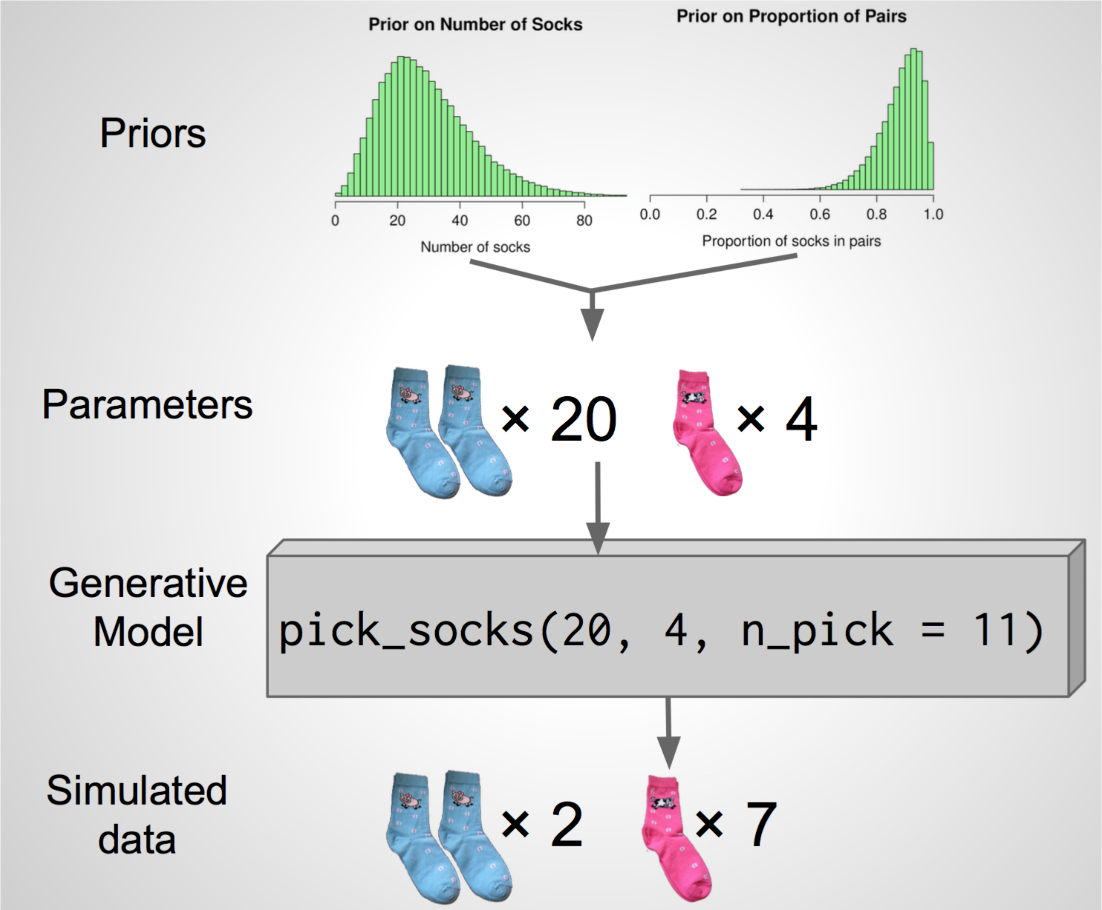
A *prior distribution* is a probability distribution for a *parameter* that summarizes the information that you have before seeing the data. Prior on N:



Prior on proportion pairs

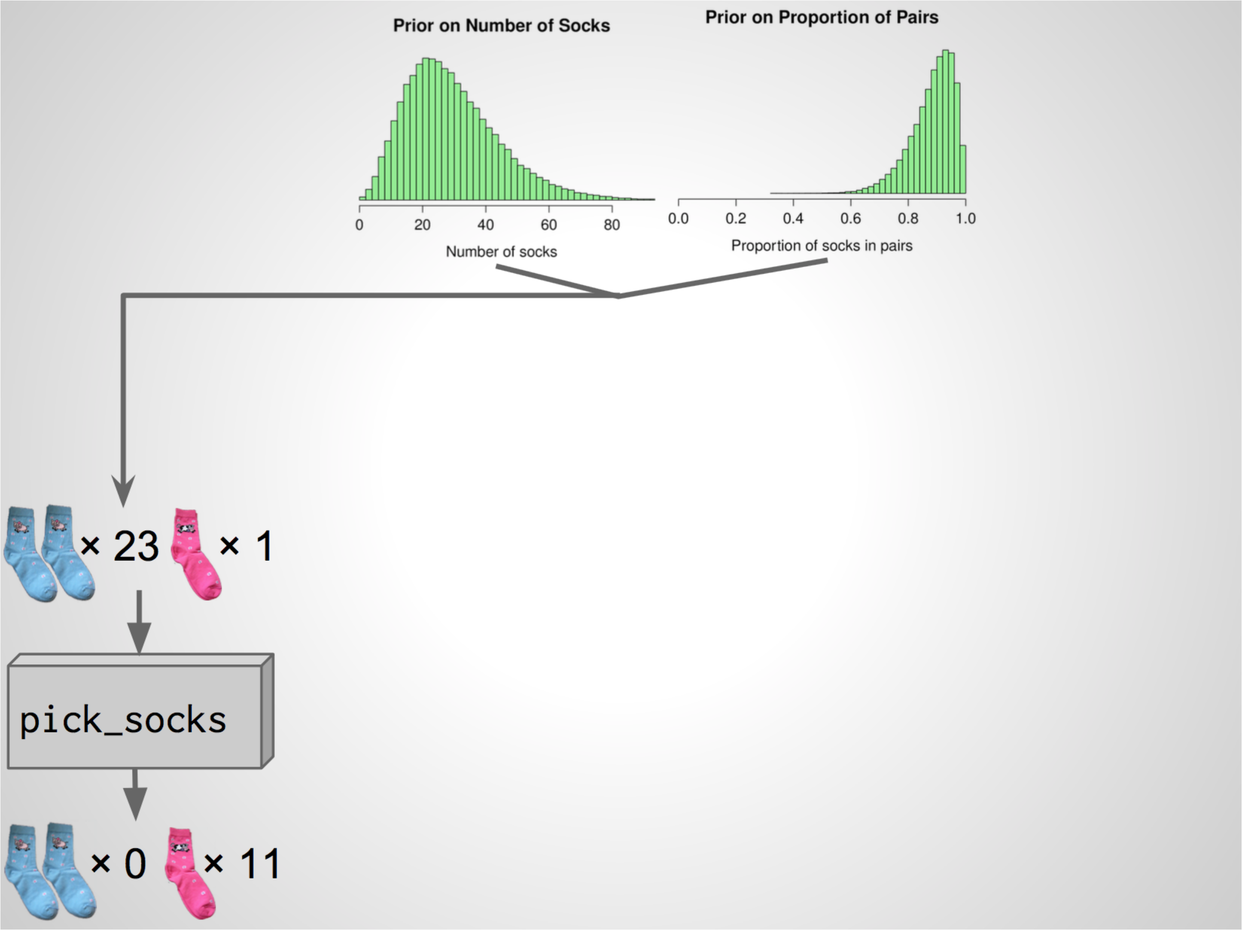


Our scheme

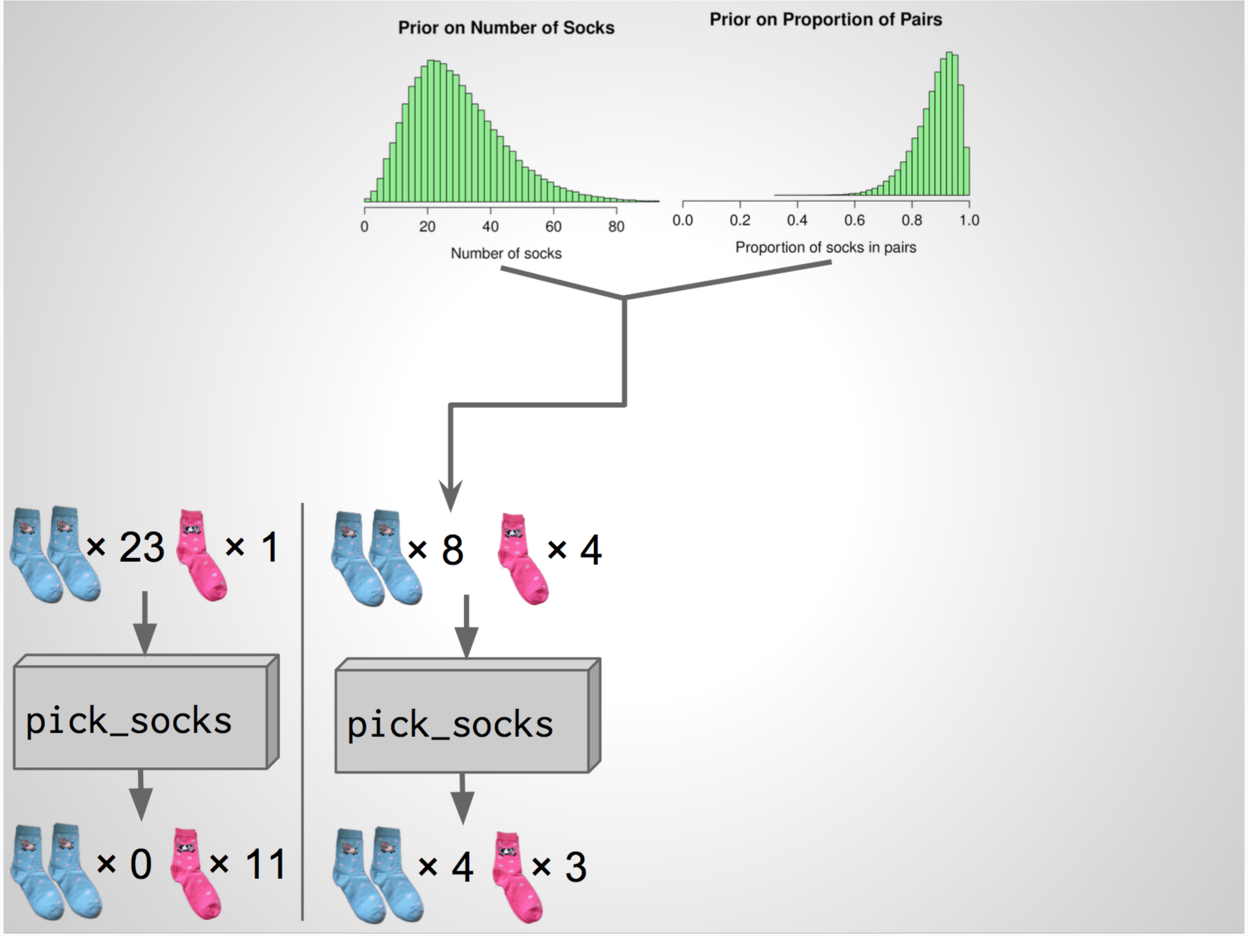


Prior on Number of Socks
Prior on Proportion of Pairs
Priors
0
20
40
60
80
Number of socks
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of socks in pairs
Parameters
× 20
× 4
Generative Model
pick_socks(20, 4, n_pick = 11)
Simulated data
× 2
× 7

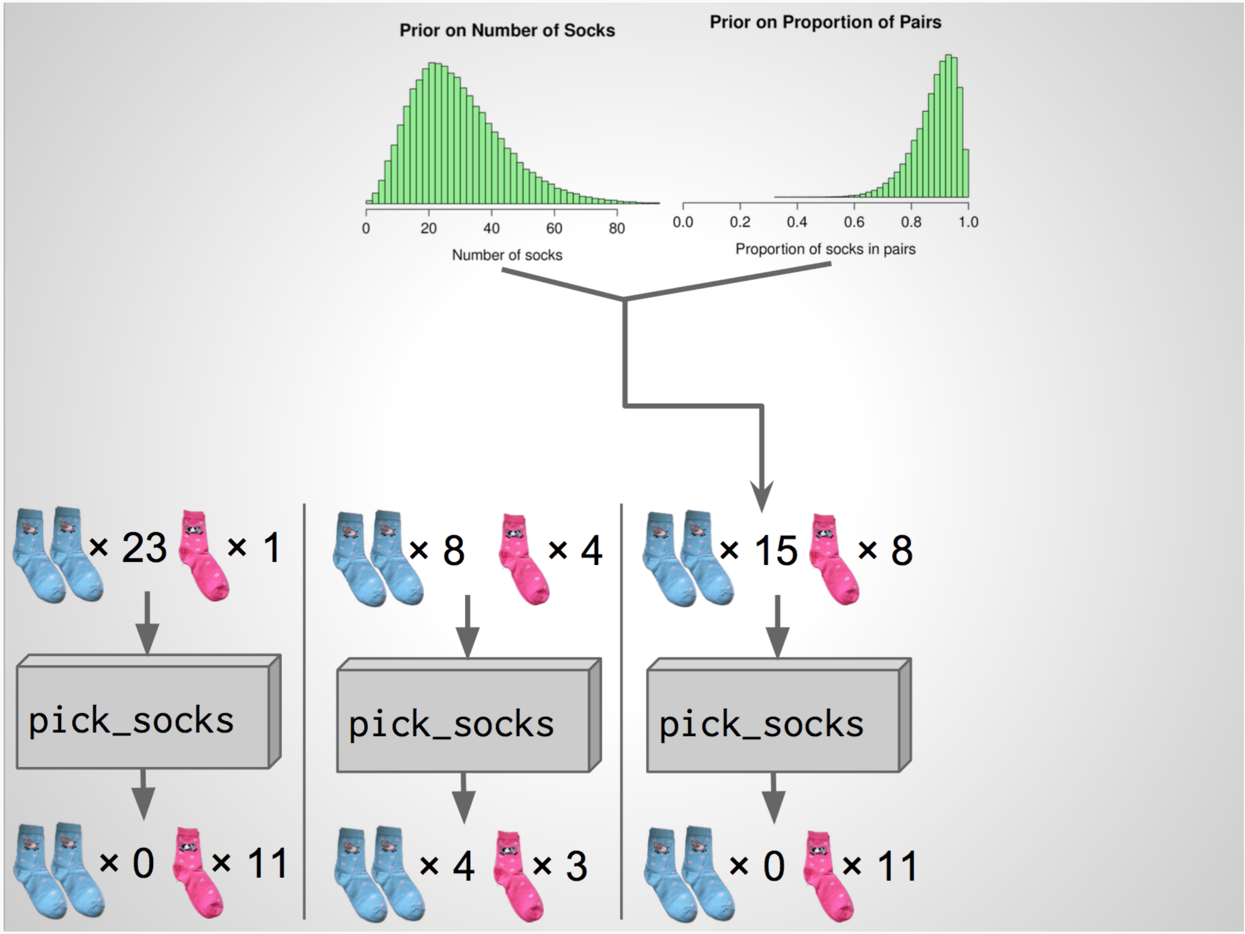
One simulation



A second simulation

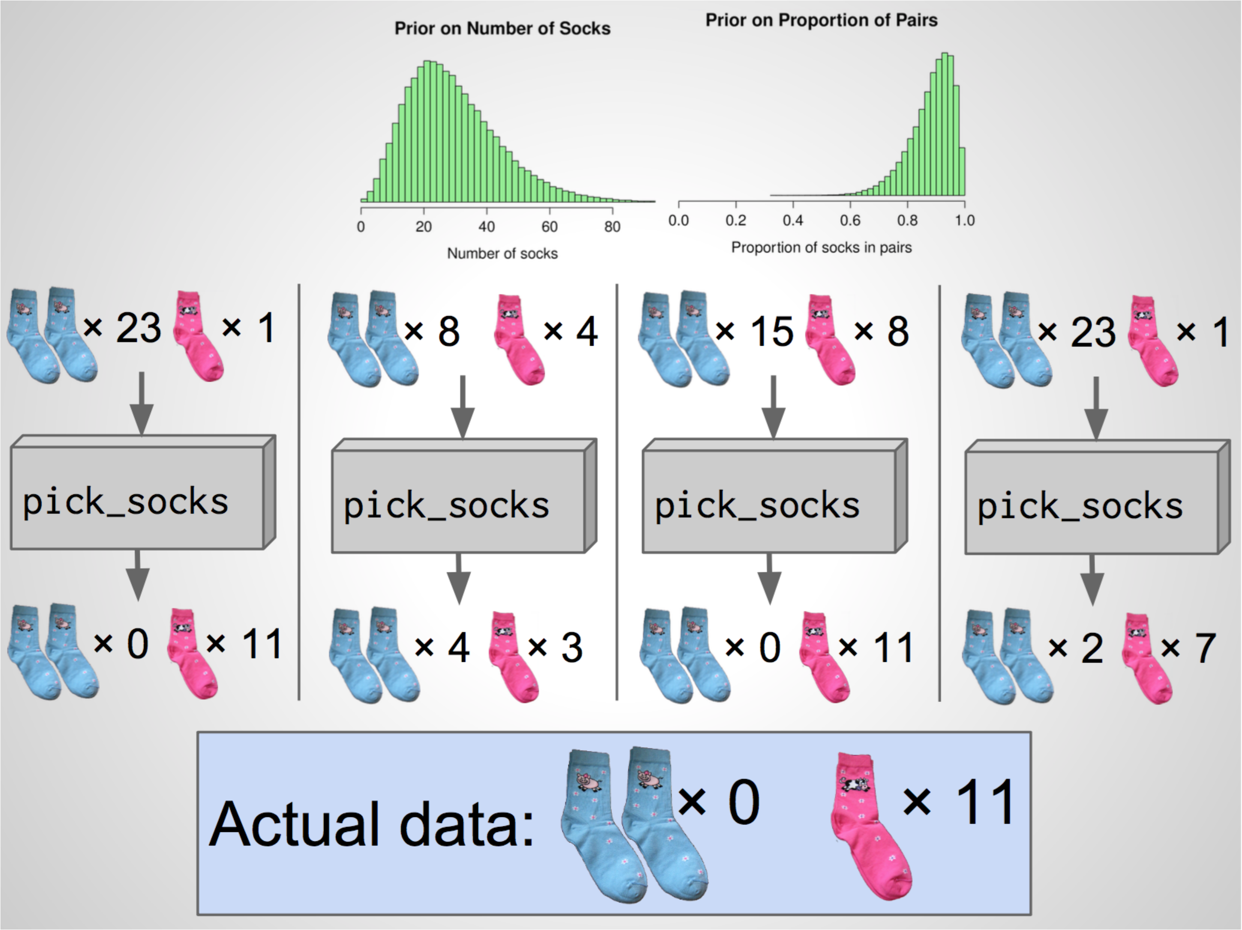


A third simulation

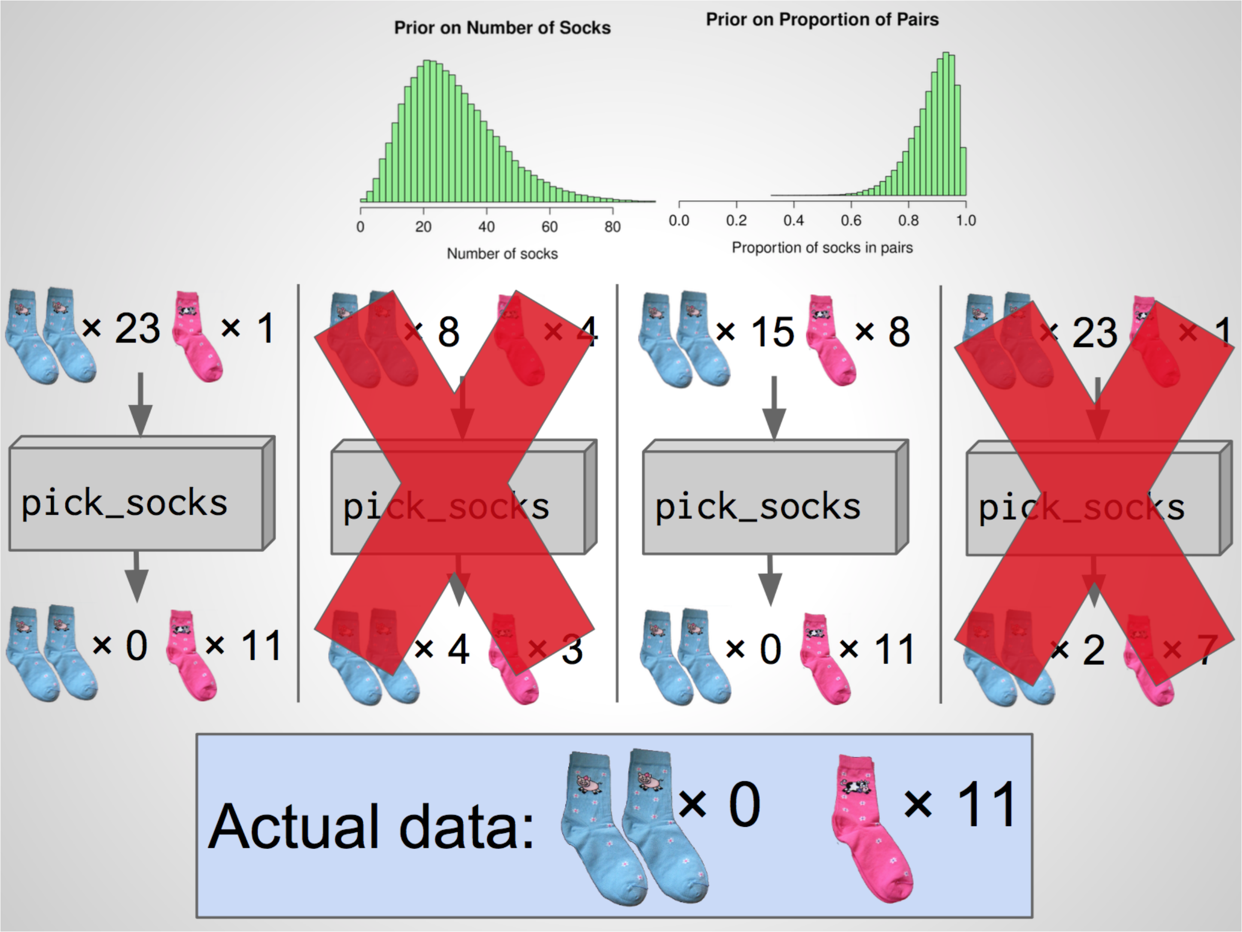


A fourth simulation

The actual data



The actual data



Full simulation

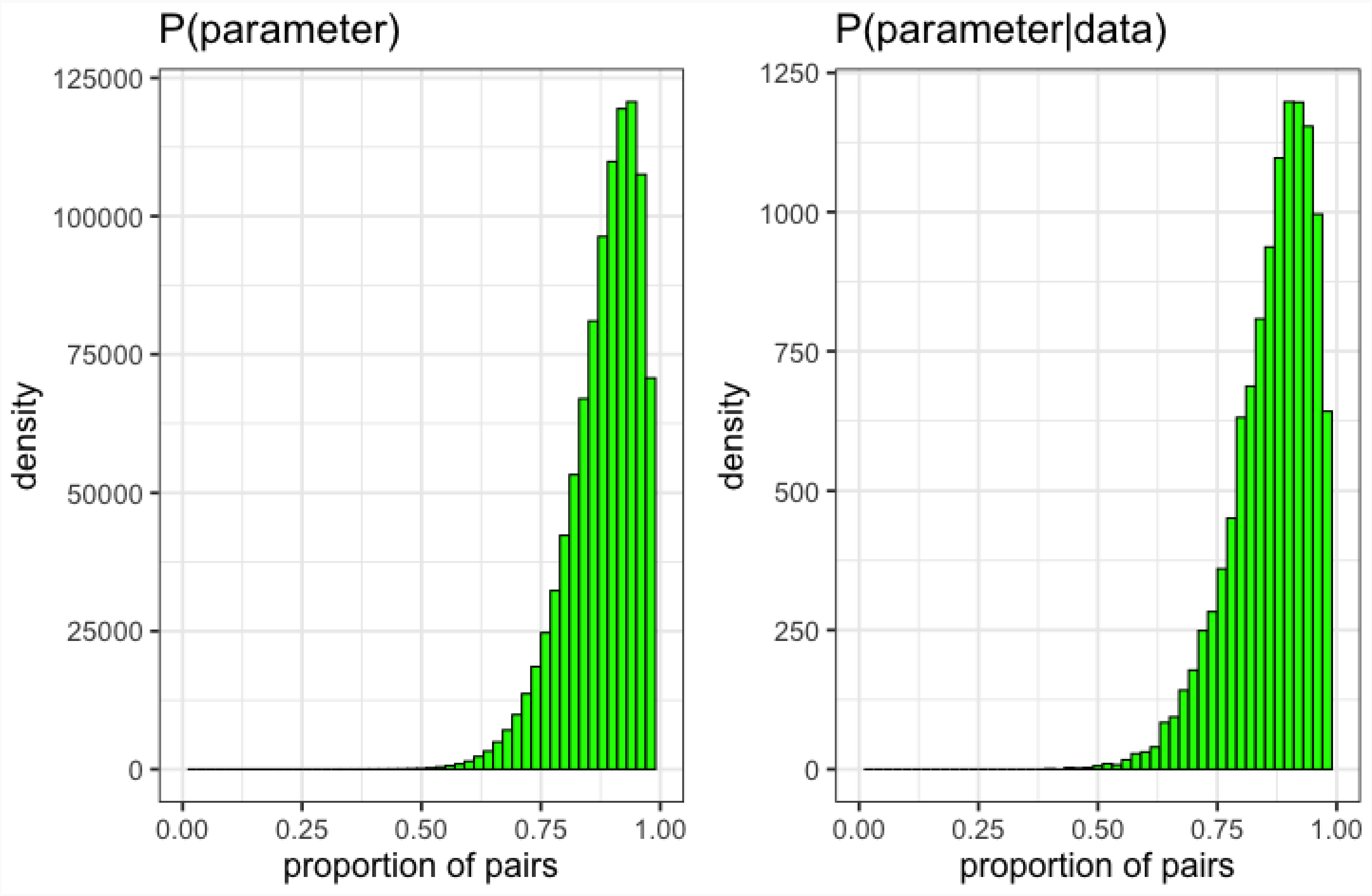
```
head(sock_sim)
```

```
##   singletons pairs n_socks prop_pairs
## 1          5     3      18      0.826
## 2         11     0      53      0.715
## 3          9     1      27      0.973
## 4          7     2      35      0.724
## 5          9     1      31      0.869
## 6          9     1      33      0.758
```

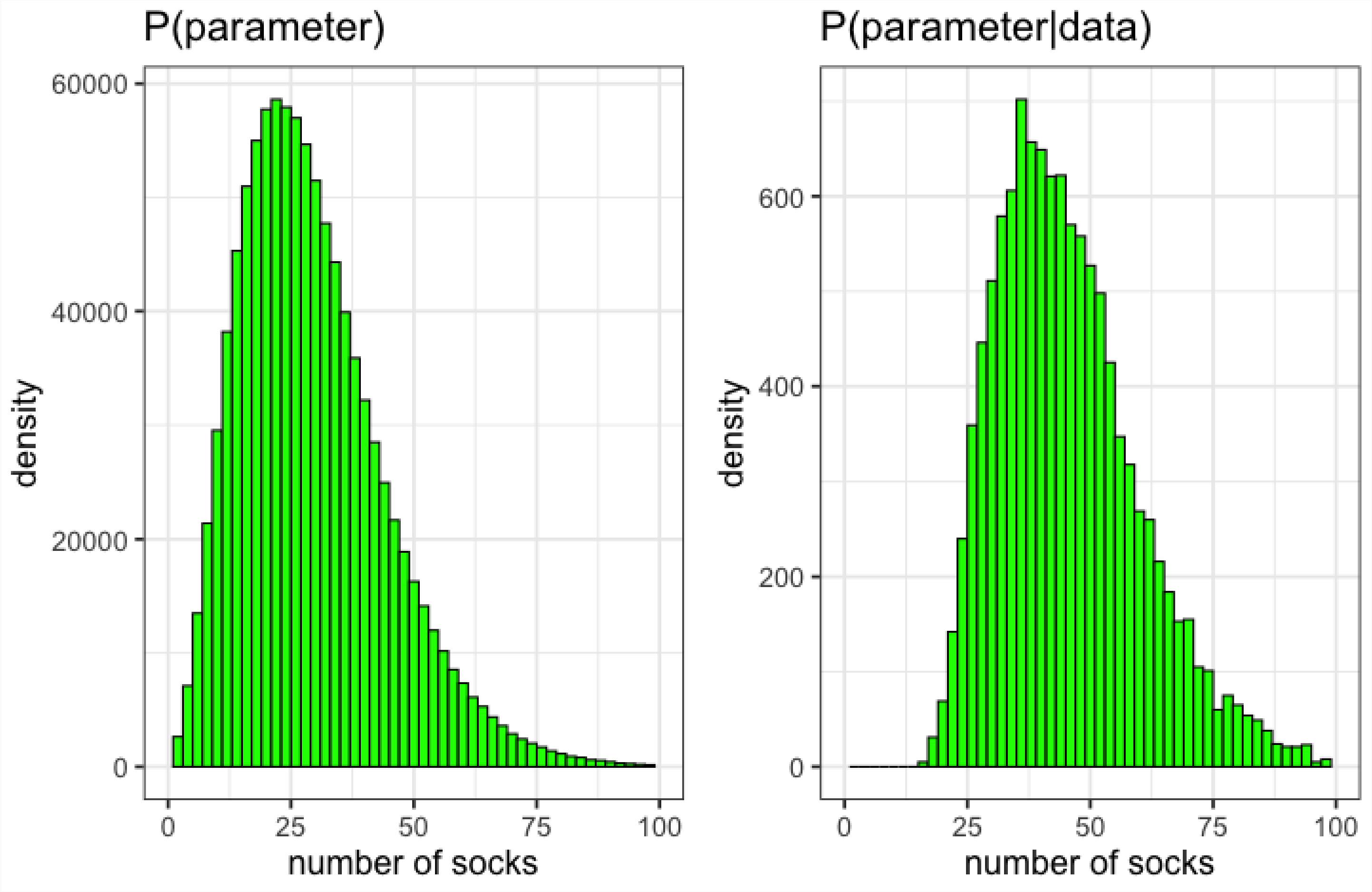
```
sock_sim %>%
  filter(singletons == 11, pairs == 0) %>%
  head()
```

```
##   singletons pairs n_socks prop_pairs
## 1         11     0      53      0.715
## 2         11     0      41      0.885
## 3         11     0      53      0.957
## 4         11     0      37      0.773
## 5         11     0      45      0.880
## 6         11     0      51      0.754
```

Proportion of pairs



Number of socks



Karl Broman's Socks

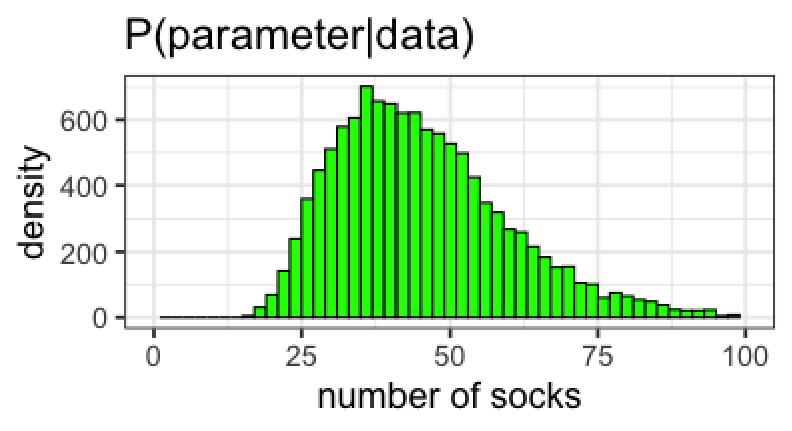
Karl Broman
@kwbroman

Follow

That the 1st 11 socks in the laundry are each distinct suggests there are a lot more socks.

RETWEETS 4 LIKES 11

The posterior distribution

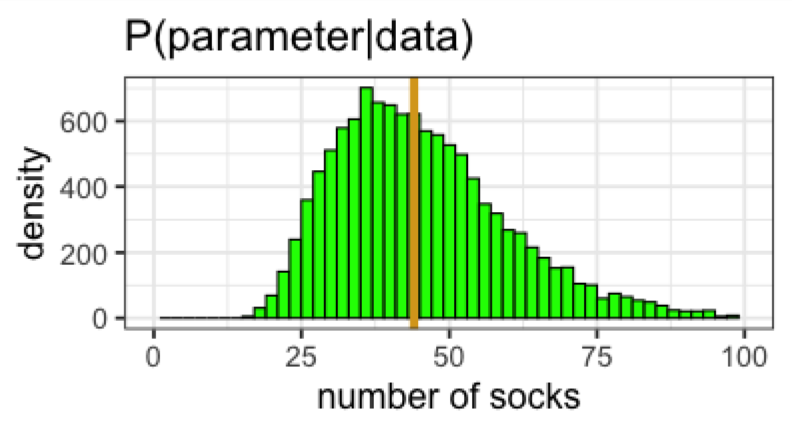


- Distribution of a parameter after conditioning on the data
- Synthesis of prior knowledge and observations (data)

Question

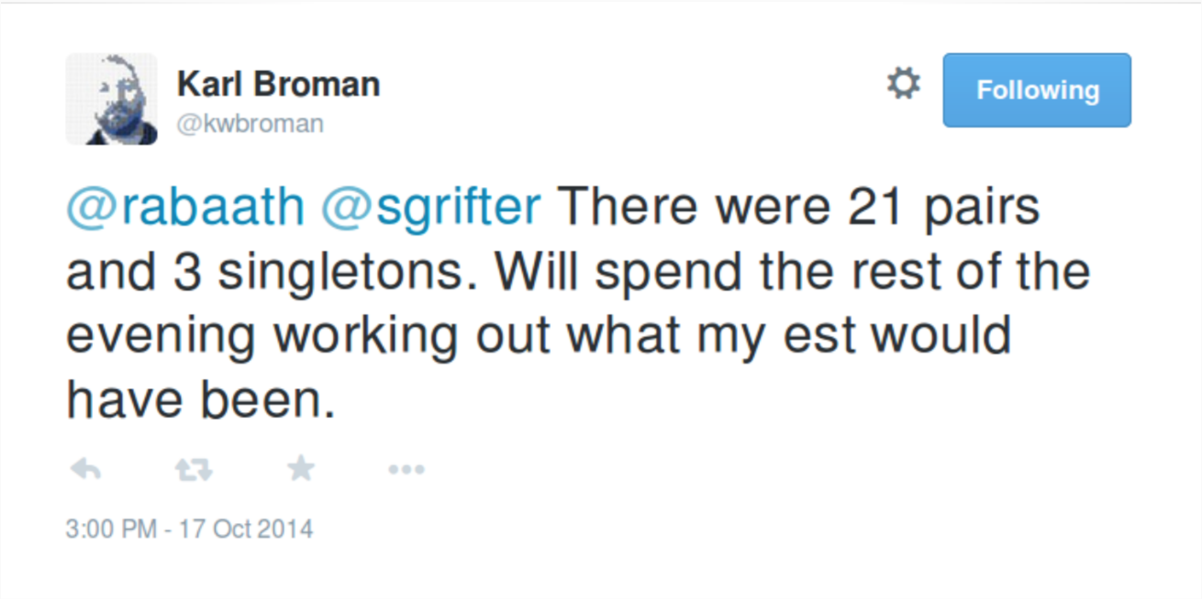
What is your best guess for the number of socks that Karl has?

Our best guess



- The posterior median is 44 socks.

Karl Broman's Socks

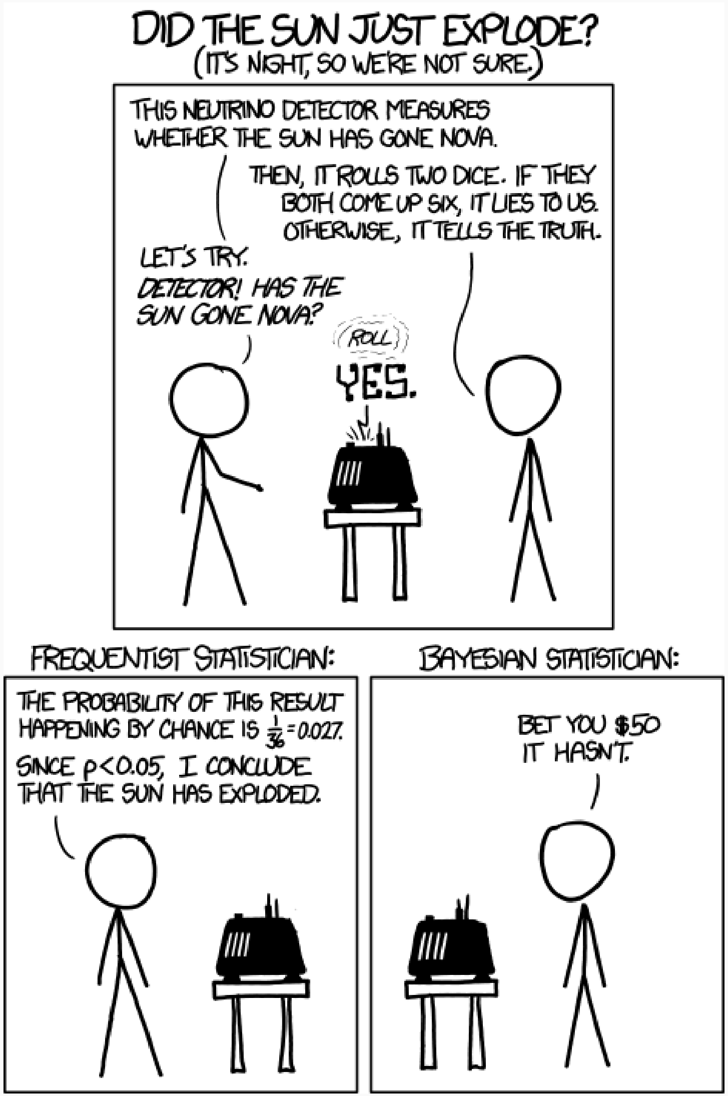


$$21 \times 2 + 3 = 45 \text{ socks}$$

Summary

Bayesian methods . . .

- Require the subjective specification of your prior knowledge
- Provide a posterior distribution on the parameters
- Are usually computationally intensive
- Have strong intuition


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
BAYESIAN STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BET YOU $50 IT HASN'T.