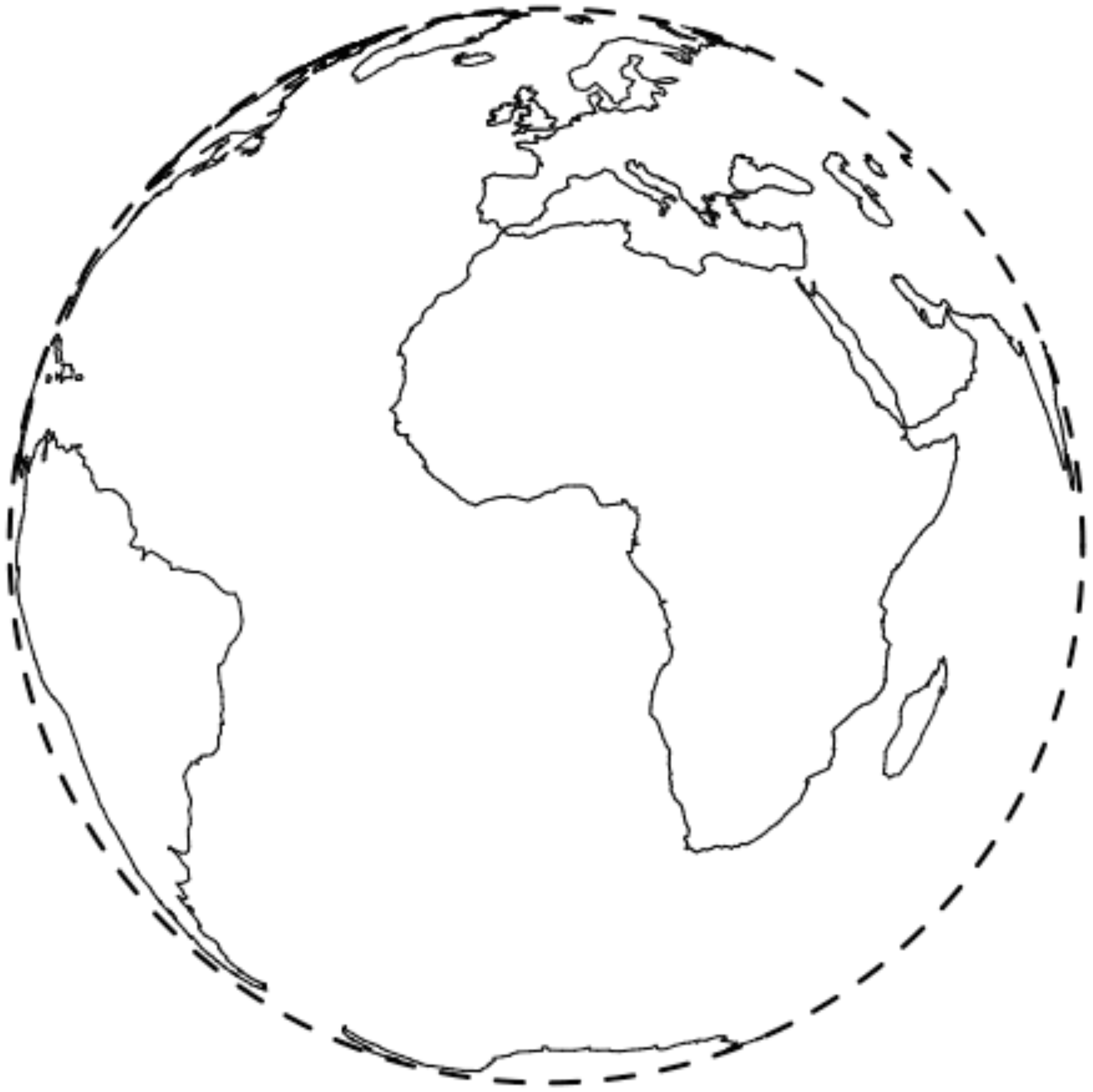
Statistical Rethinking

2023

2. The Garden of Forking Data

What proportion of the surface is covered with water?



How should we use the sample?

How to produce a summary?

How to represent uncertainty?

Workflow

(1) Define generative model of the sample

(2) Define a specific estimand

(3) Design a statistical way to produce estimate

(4) Test (3) using (1)

(5) Analyze sample, summarize

Generative model of the globe

Begin conceptually: How do the variables influence one another?

proportion of water p

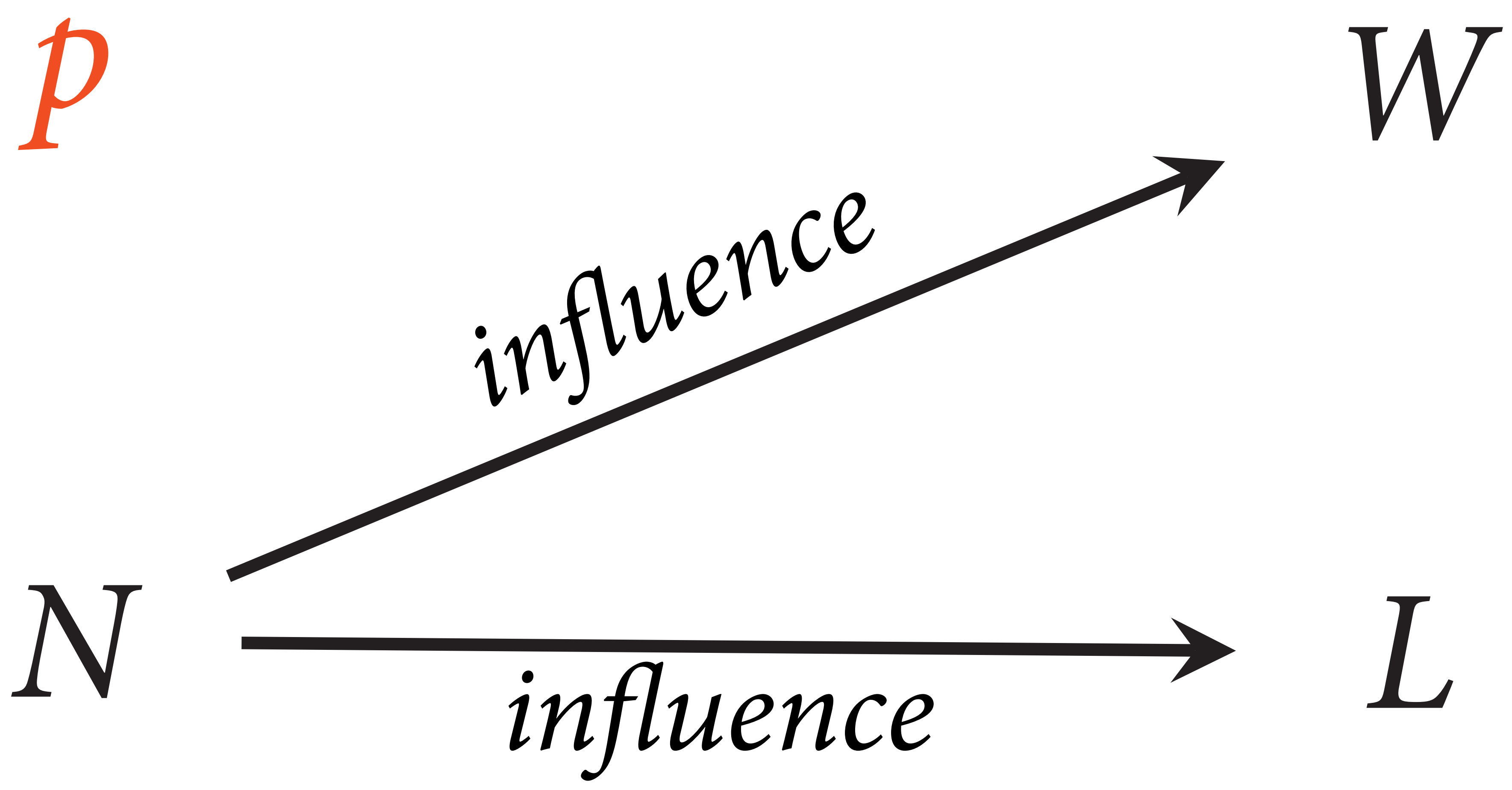
W water observations

number of tosses N

L land observations

Generative model of the globe

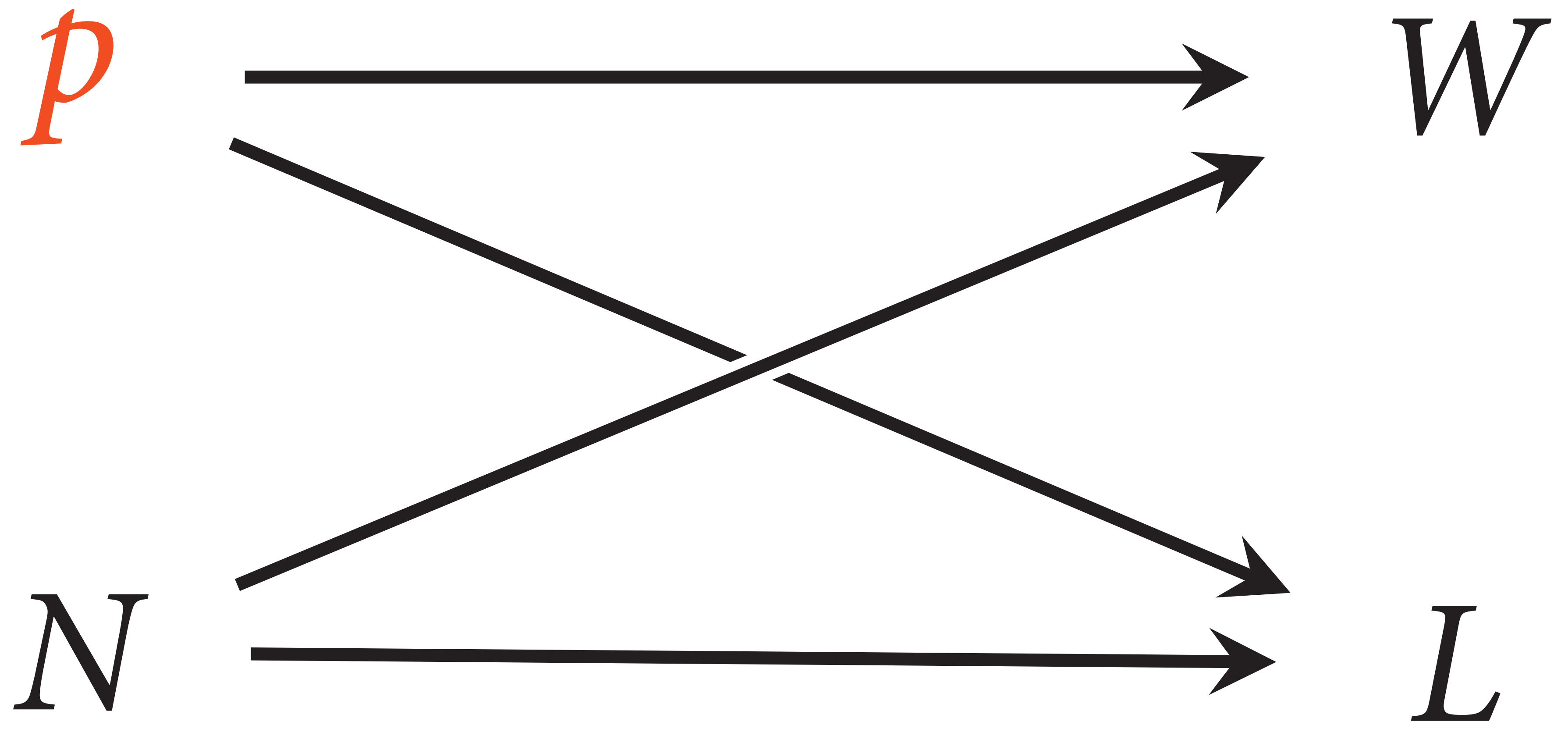
Begin conceptually: How do the variables influence one another?



N influences W and L

Generative model of the globe

Begin conceptually: How do the variables influence one another?



Generative model of the globe

Generative assumptions: What do the arrows mean exactly?

$$\begin{tikzcd}
p \arrow[r] \arrow[dr] & W \\
N \arrow[r] \arrow[ur] & L
\end{tikzcd}$$

$$W, L = f(p, N)$$

Workflow

(1) Define generative model of the sample

(2) Define a specific estimand

(3) Design a statistical way to produce estimate

(4) Test (3) using (1)

(5) Analyze sample, summarize

Bayesian data analysis

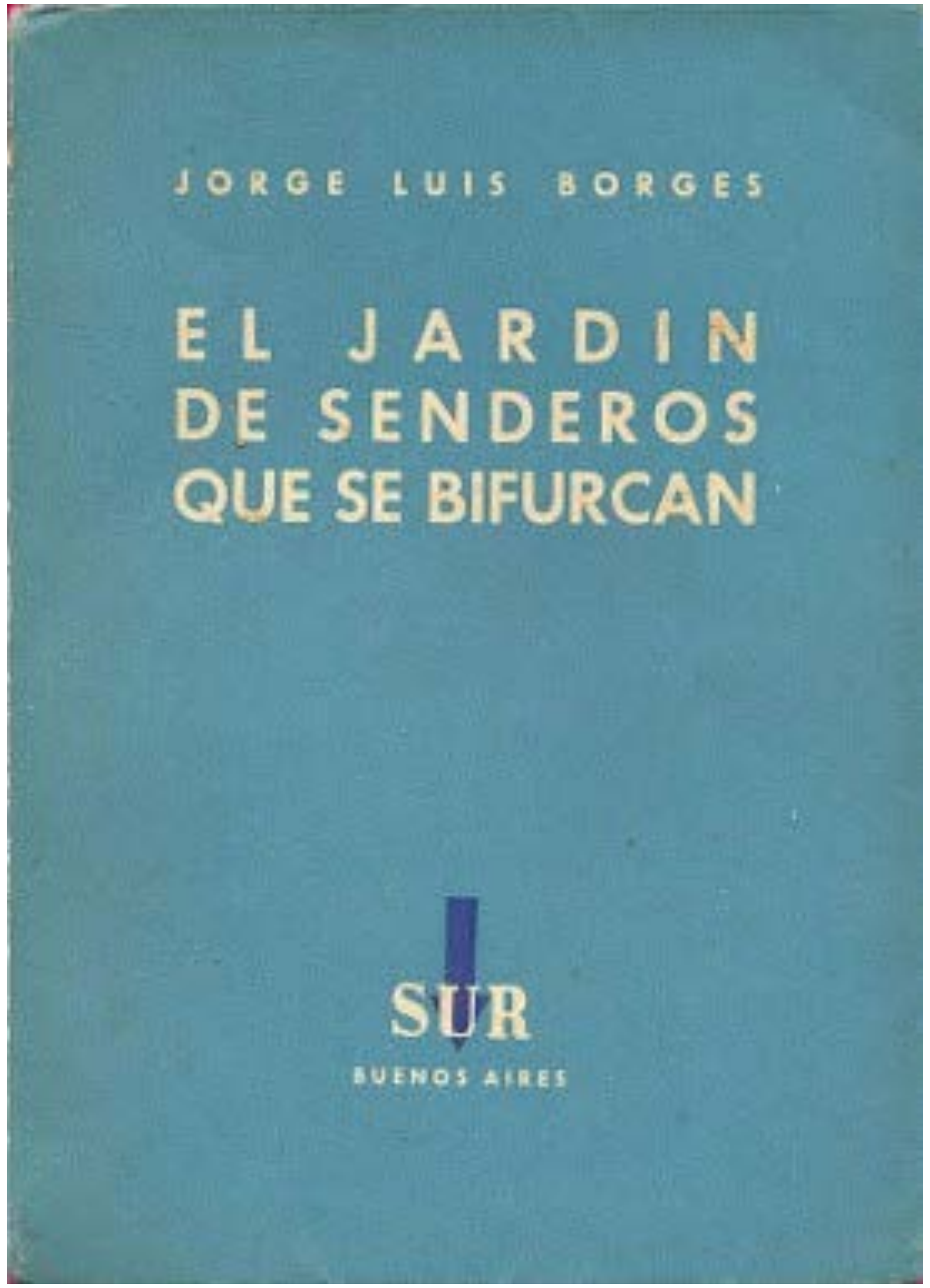
For each possible explanation of the sample,

Count all the ways the sample could happen.

Explanations with more ways to produce the sample are more plausible.

The Garden of Forking Data

El jardín de los datos que se bifurcan



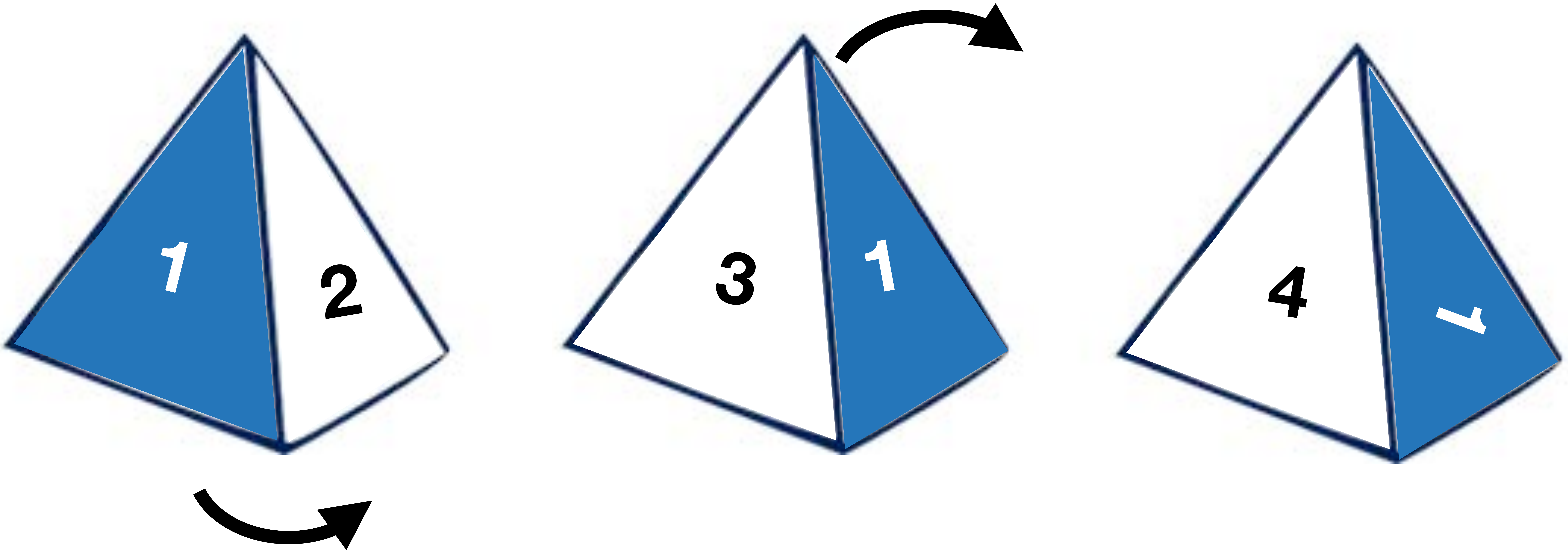
For each possible proportion of water on the globe,

Count all the ways the sample of tosses could happen.

Proportions with more ways to produce the sample are more plausible.

A Four-sided Globe

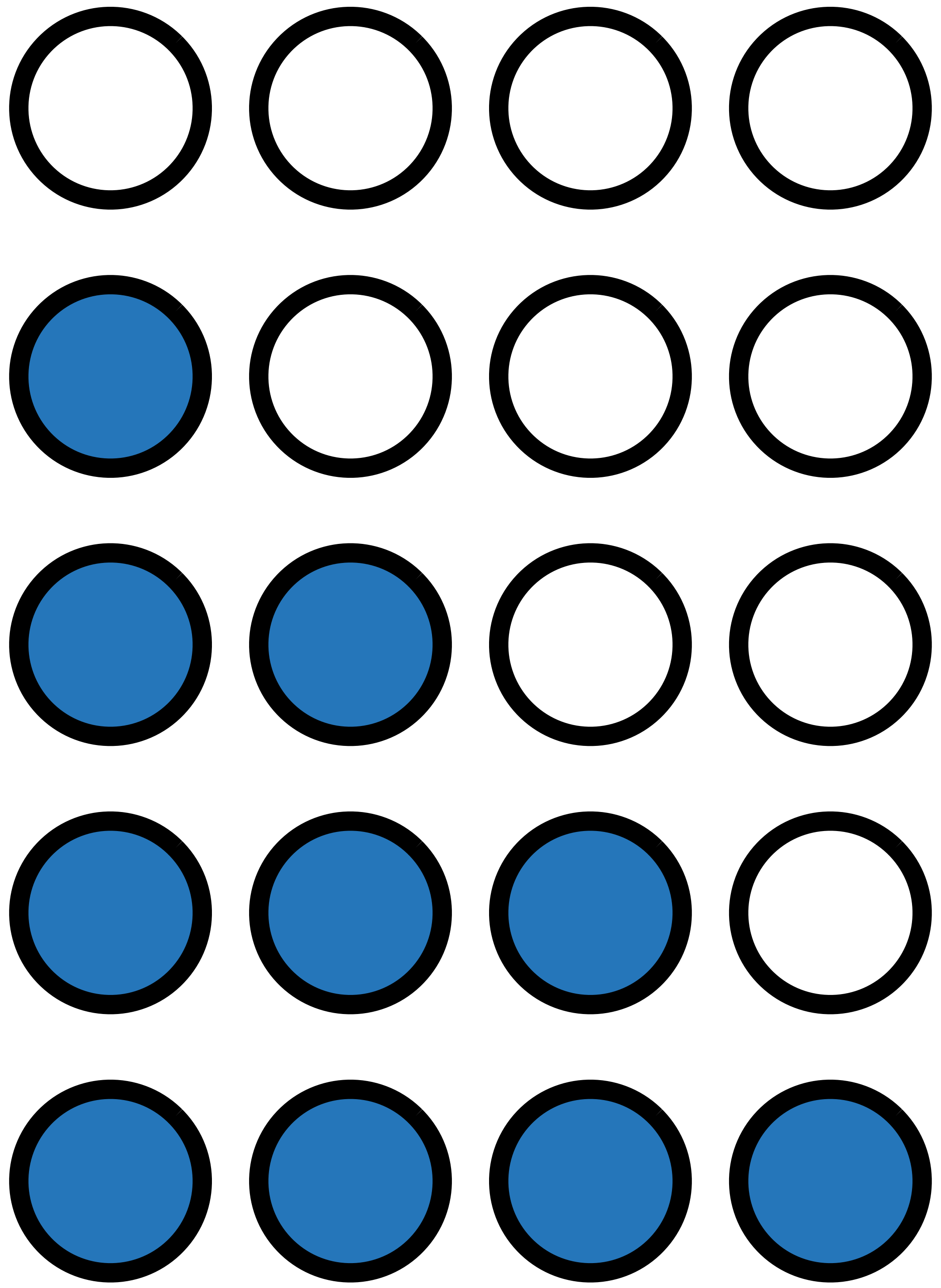
covered 25% by water



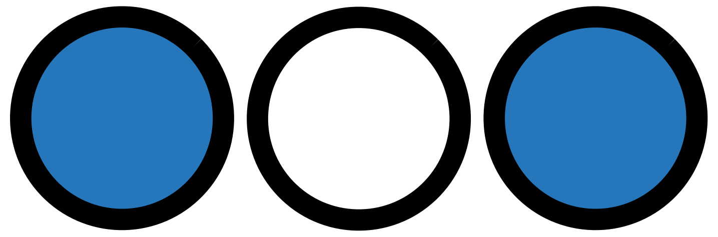
Garden of Forking Data

Possible d4 globes:

(1)
(2)
(3)
(4)
(5)

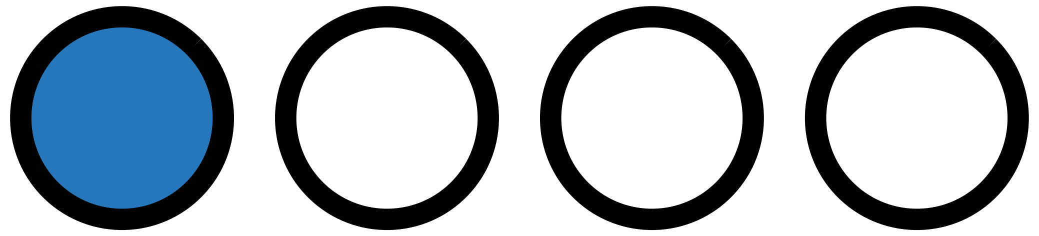


Observe:

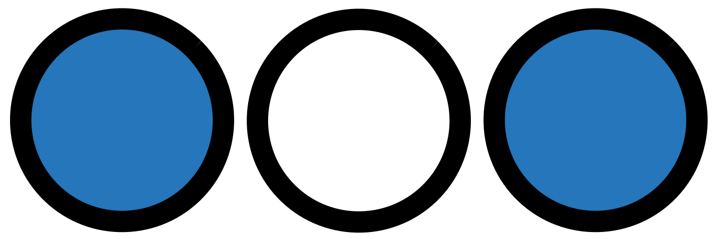


Garden of Forking Data

Possible d4 globes:

(2)  25%

Observe:



First Possibility

Figure 2.2

Second Possibility

Figure 2.2

Third Possibility

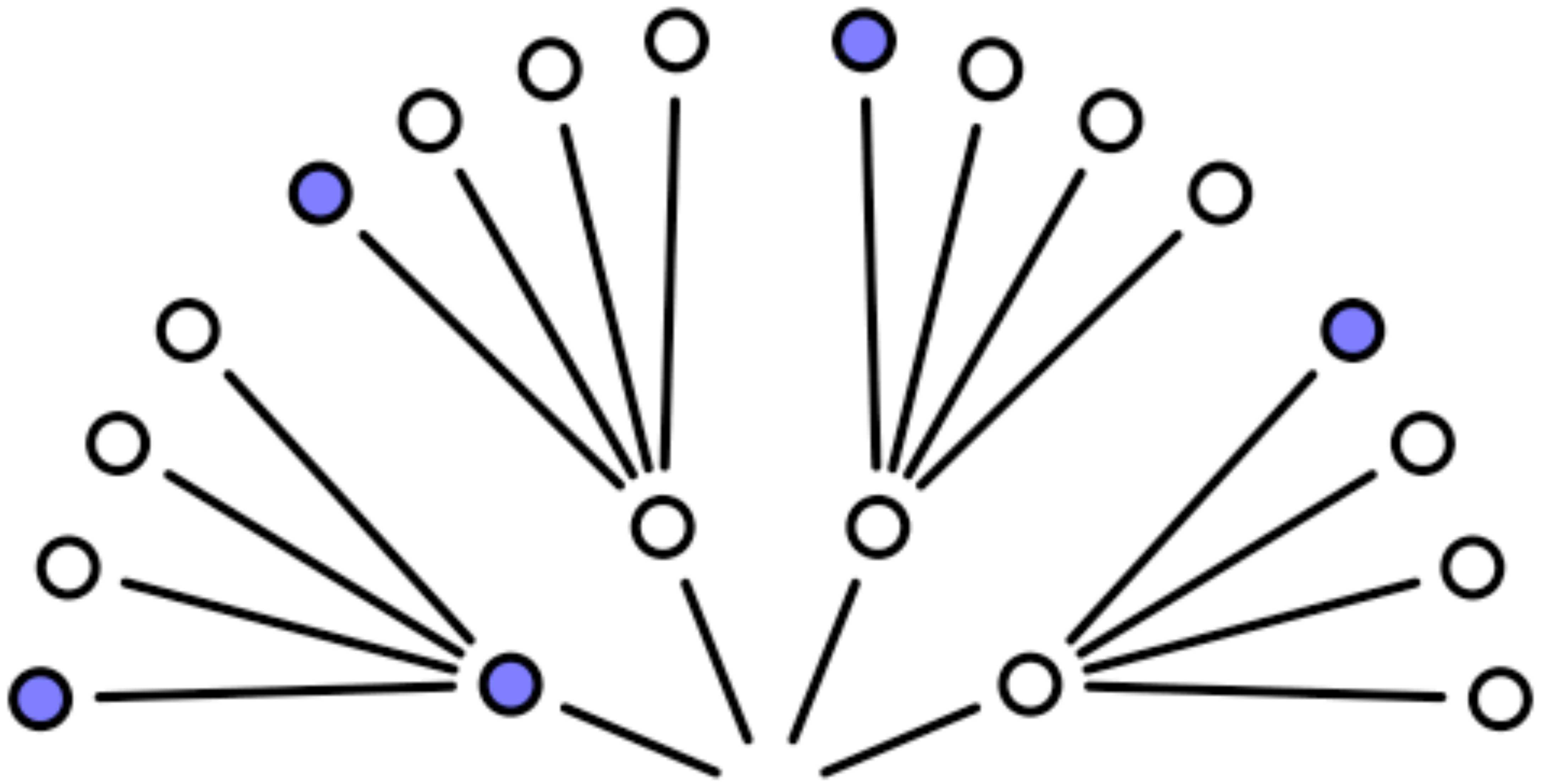


Figure 2.2

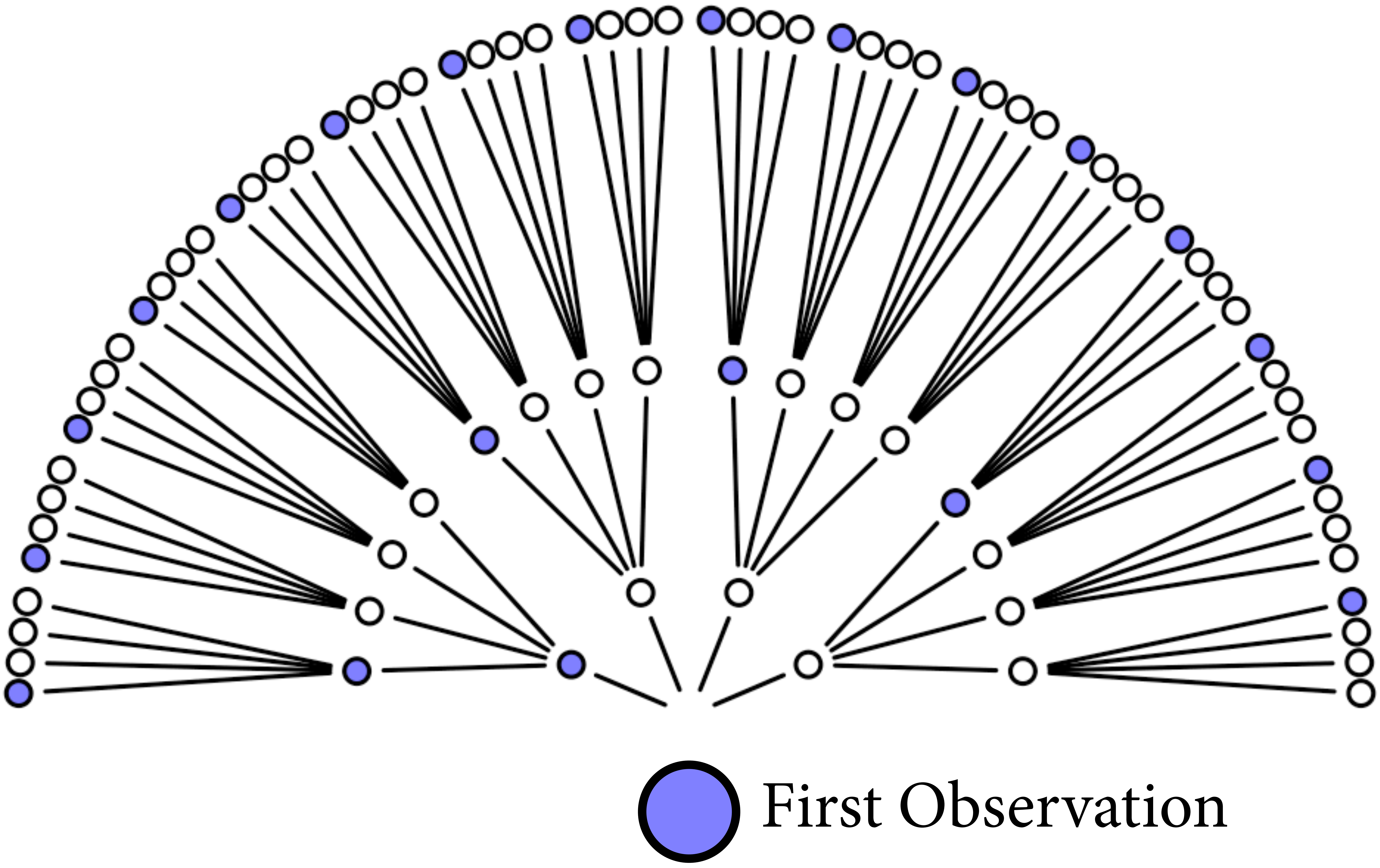


Figure 2.2

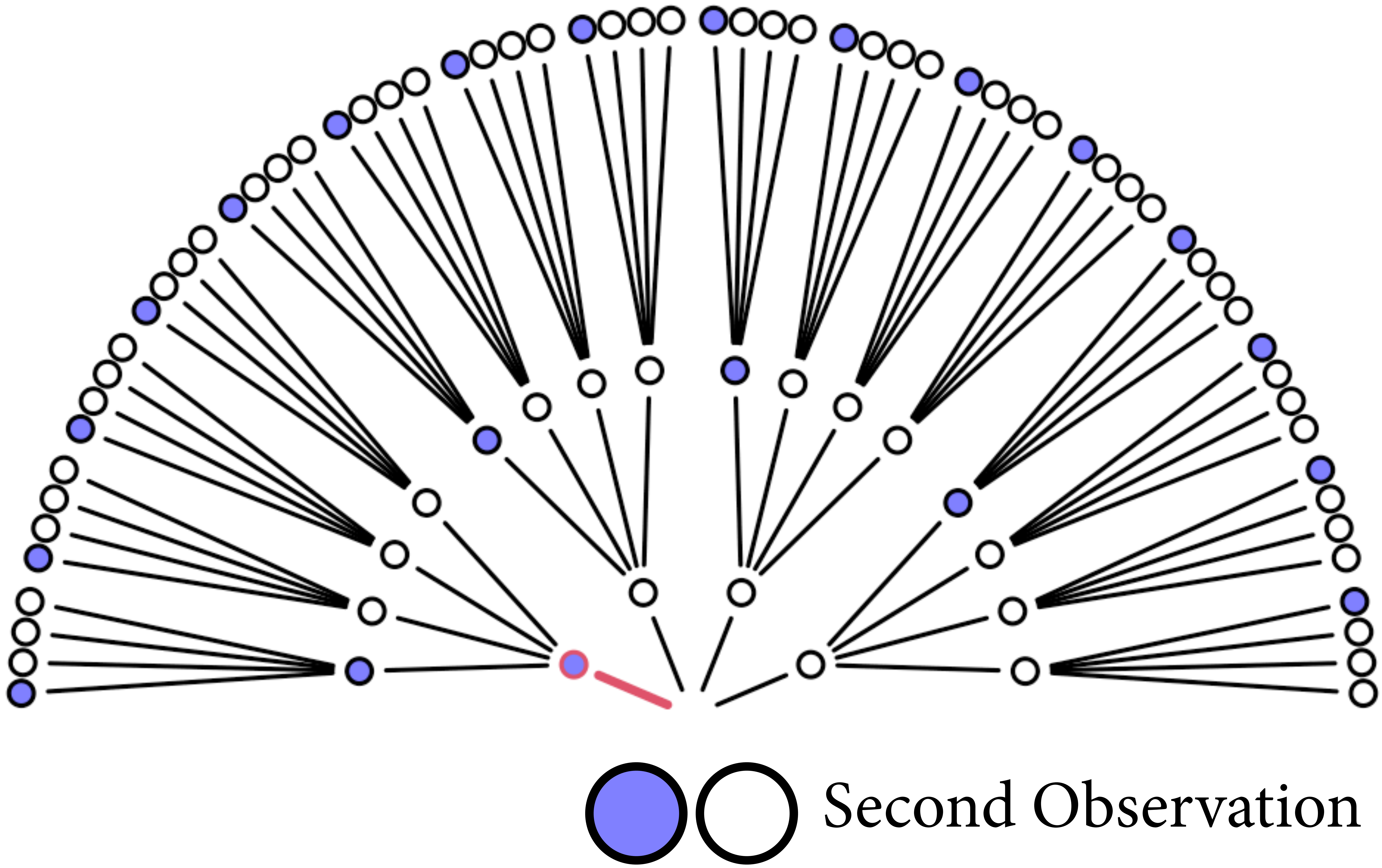


Figure 2.2

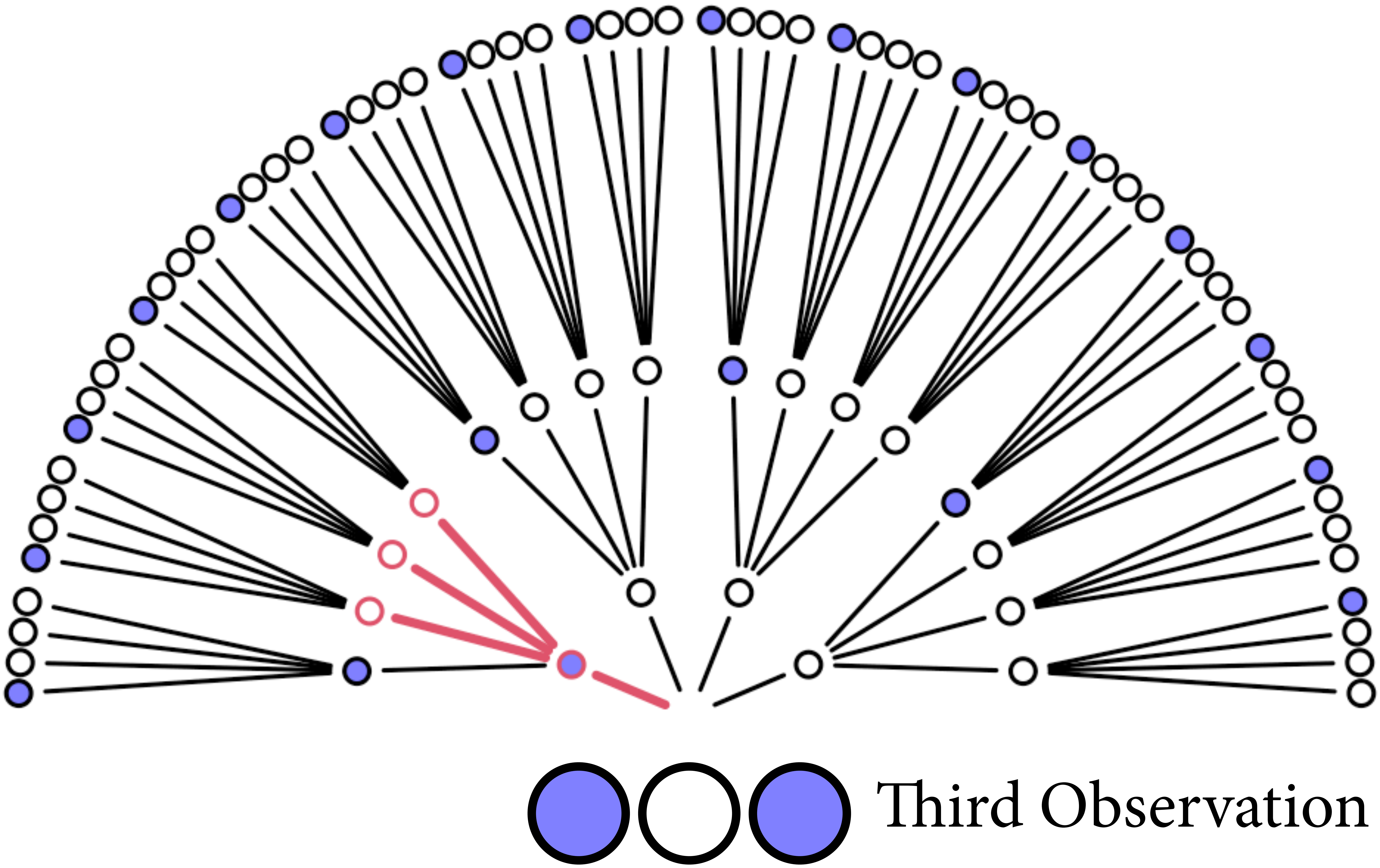


Figure 2.2

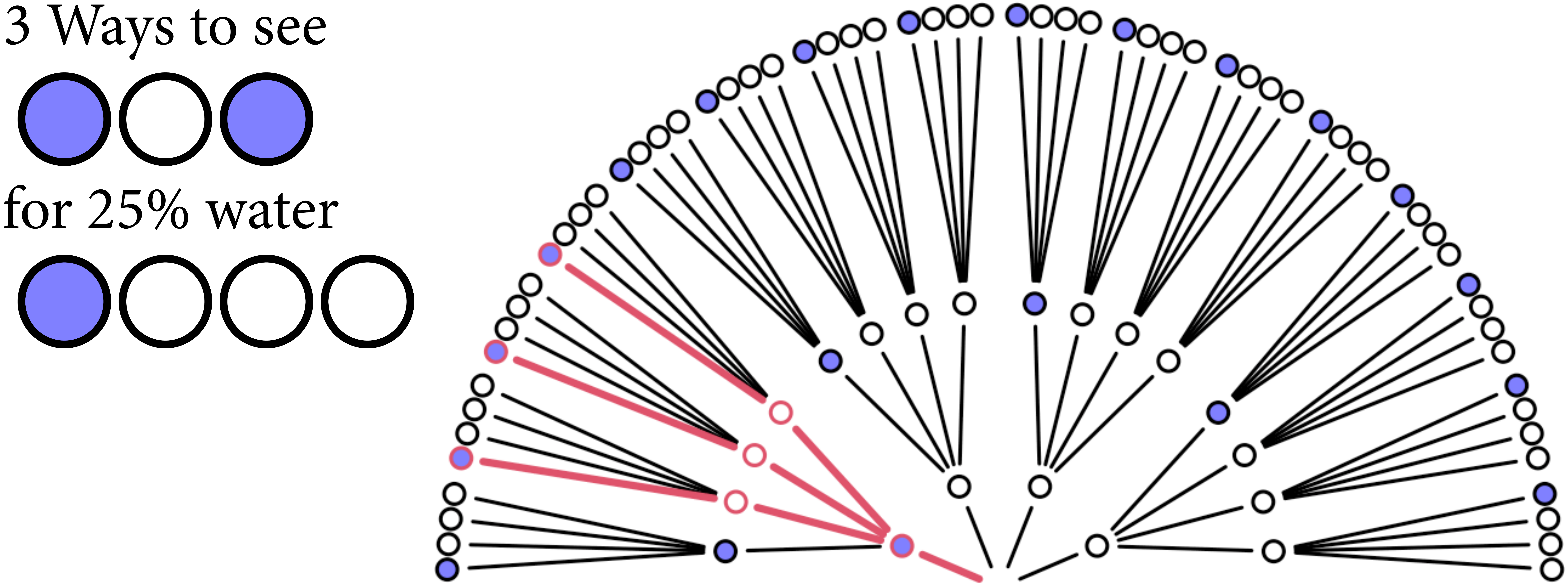
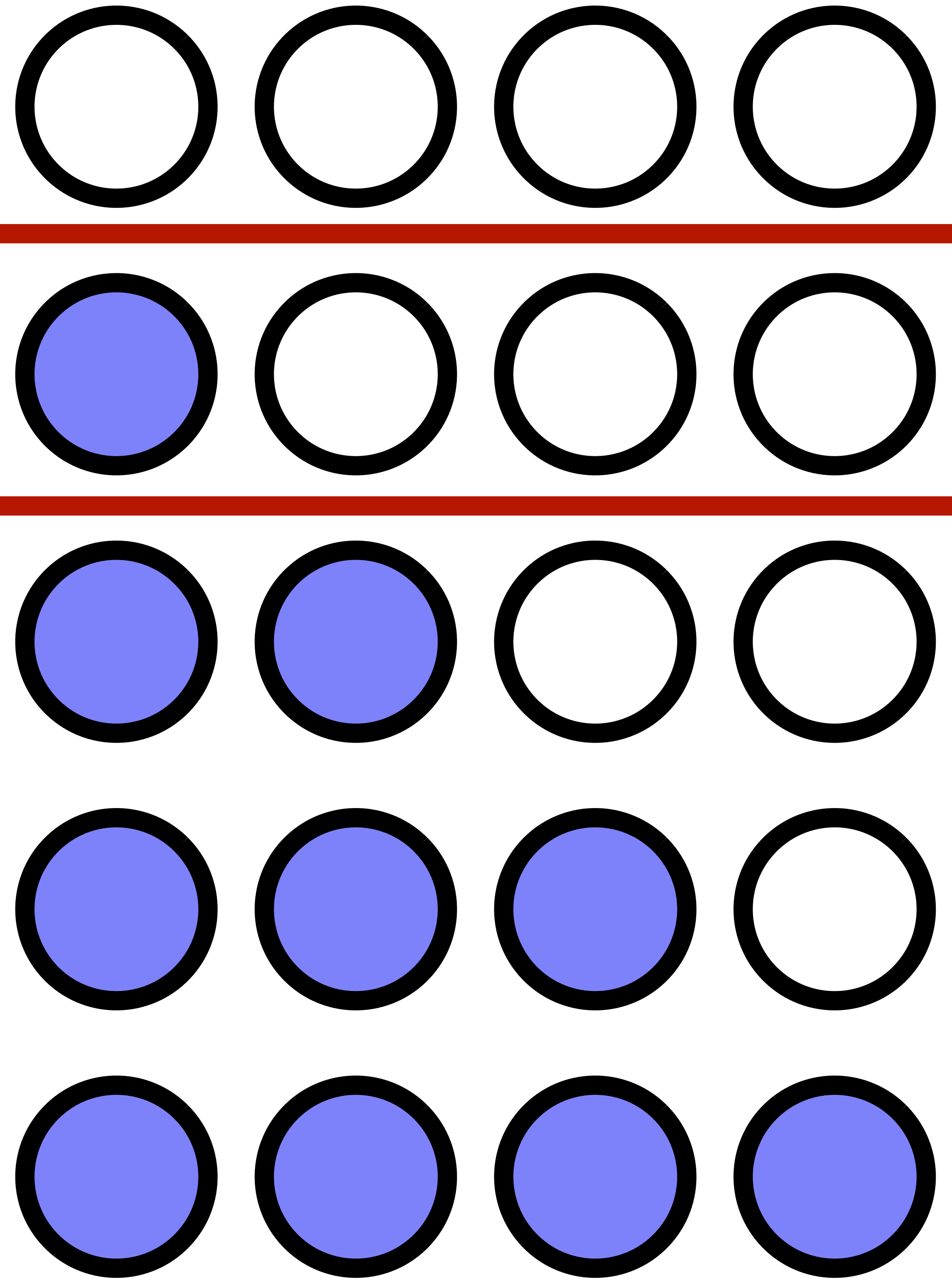
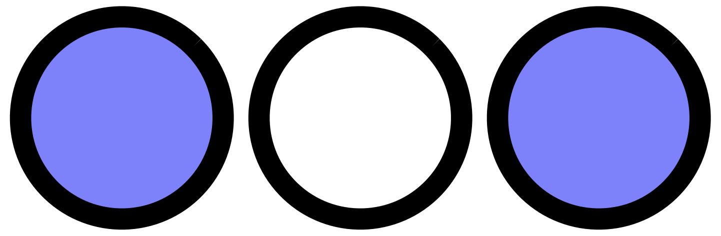


Figure 2.2

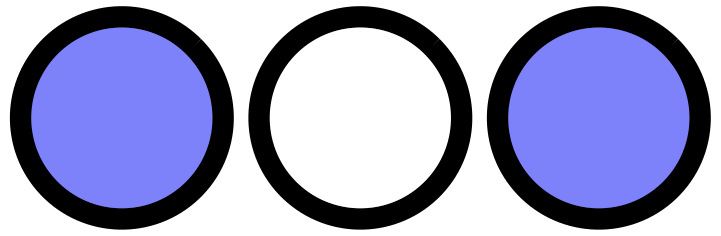
Garden of Forking Data

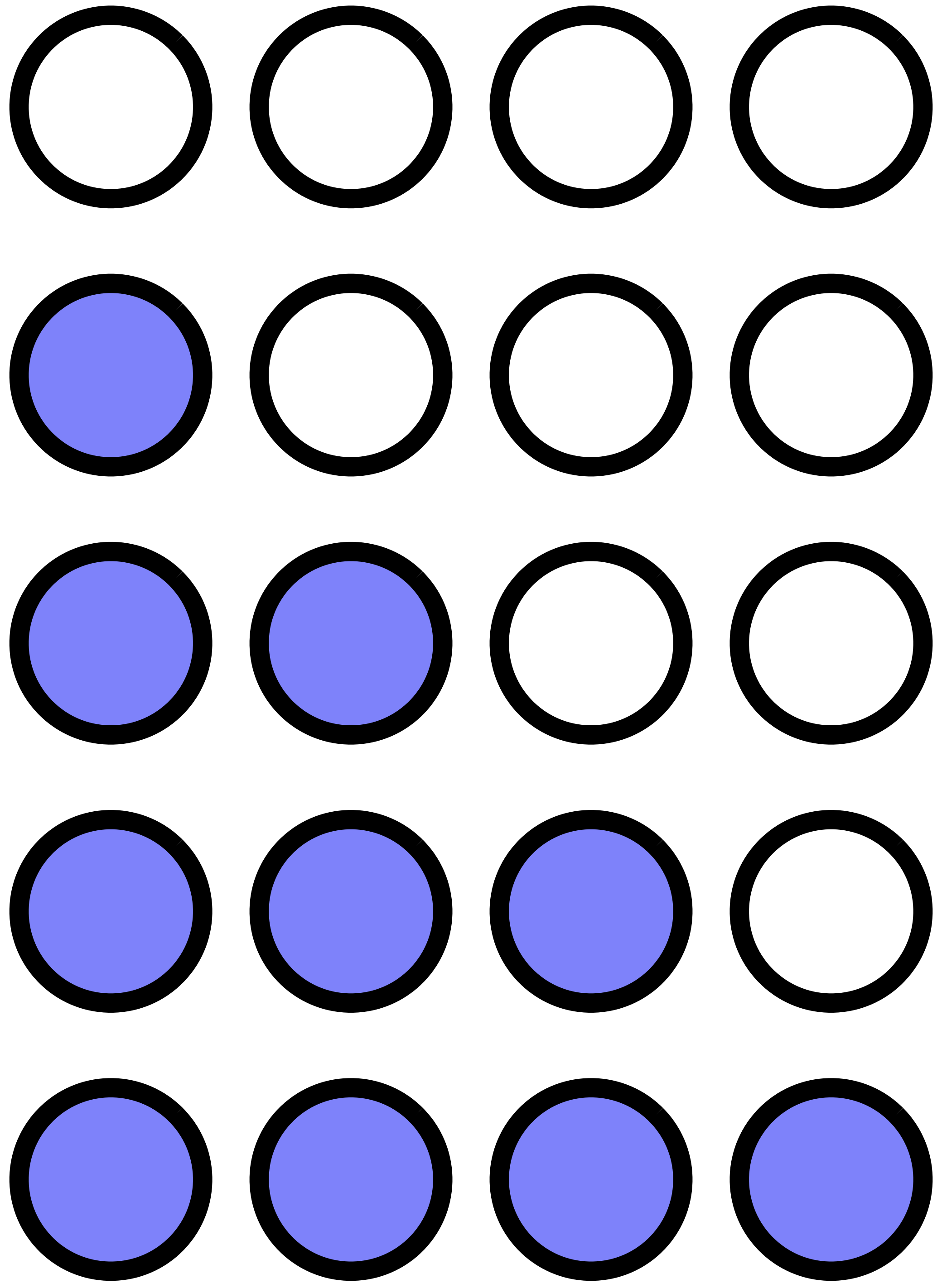


	Possible globes:	Ways to produce 
(1)		?
(2)		3
(3)		?
(4)		?
(5)		?

Garden of Forking Data

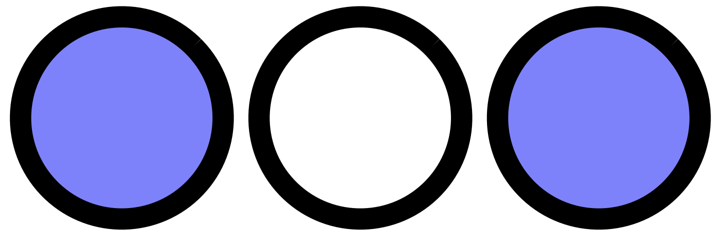
Possible globes:

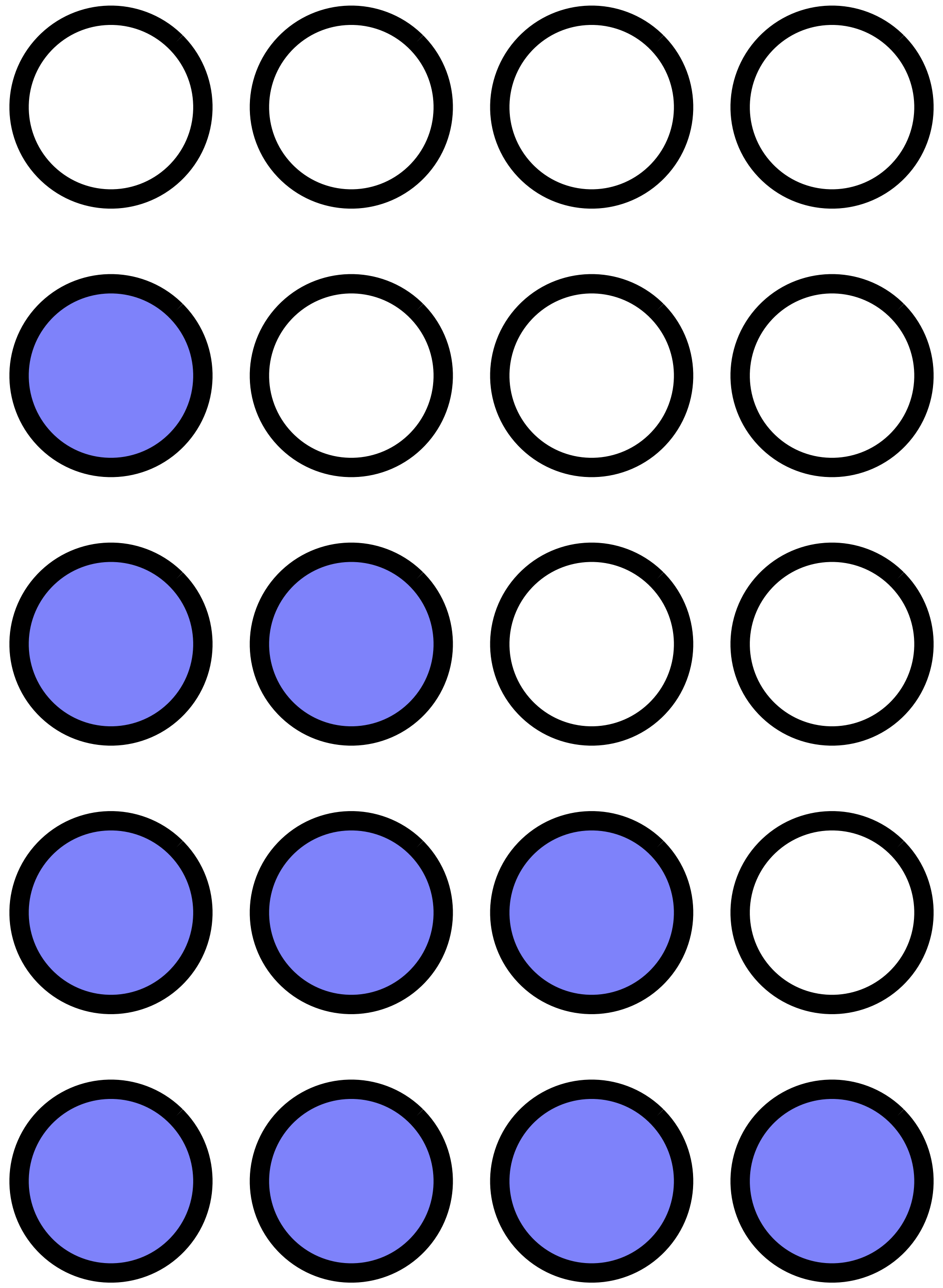
Ways to produce 

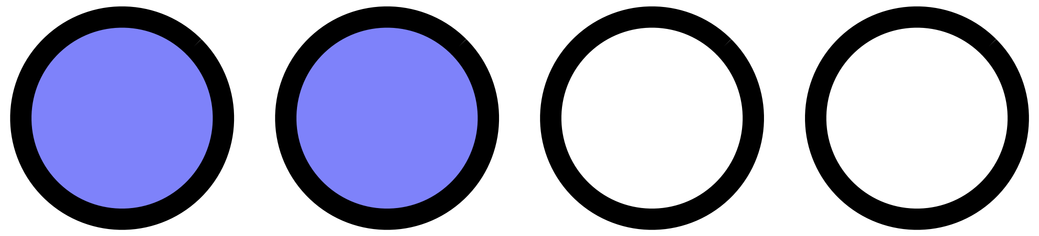


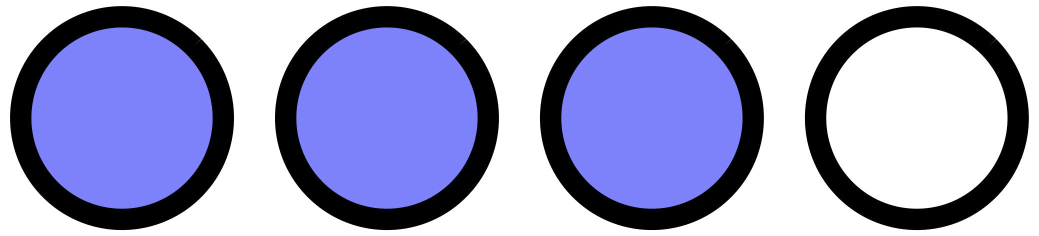
Globe	Ways to produce
(1)	0
(2)	3
(3)	?
(4)	?
(5)	?

Garden of Forking Data

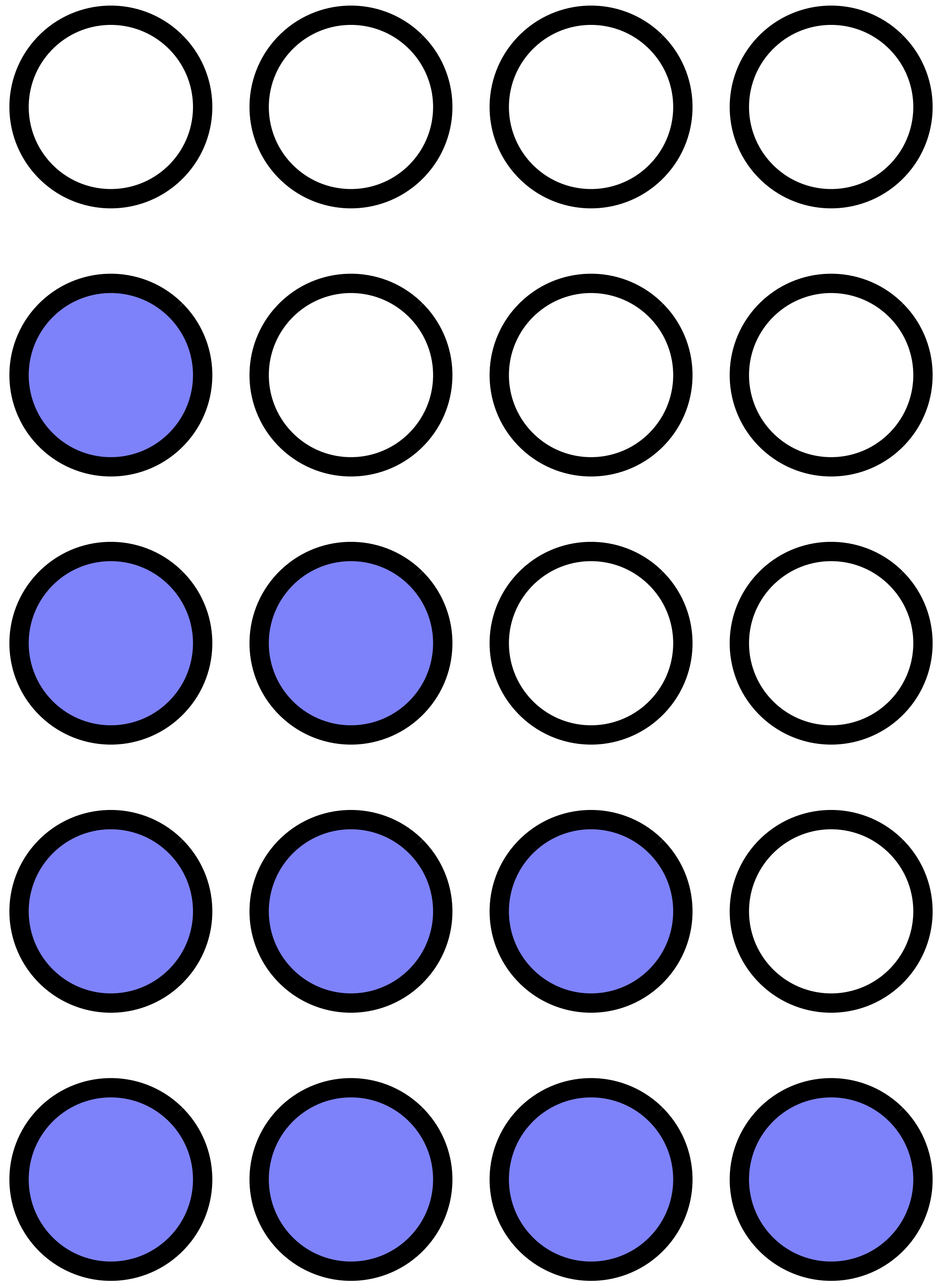
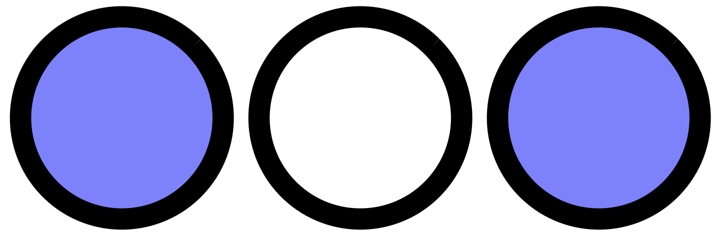
Possible globes:	Ways to produce 
(1)	0
(2)	3
(3)	?
(4)	?
(5)	0



(3) 

(4) 

Garden of Forking Data

Possible globes: 	Ways to produce 
(1)	0
(2)	3
(3)	8
(4)	9
(5)	0

Counts to plausibility

Unglamorous basis of applied probability:
Things that can happen more ways are more plausible.

Possibility	
[○○○○]	0%
[●○○○]	25%
[●●○○]	50%
[●●●○]	75%
[●●●●]	100%

Counts to plausibility

Unglamorous basis of applied probability:
Things that can happen more ways are more plausible.

Possibility		Ways to produce ●○●
[○○○○]	0%	$0 \times 4 \times 0 = 0$
[●○○○]	25%	
[●●○○]	50%	
[●●●○]	75%	
[●●●●]	100%	

Counts to plausibility

Unglamorous basis of applied probability:
Things that can happen more ways are more plausible.

Possibility		Ways to produce ● ○ ●
[○○○○]	0%	$0 \times 4 \times 0 = 0$
[●○○○]	25%	$1 \times 3 \times 1 = 3$
[●●○○]	50%	
[●●●○]	75%	
[●●●●]	100%	

Counts to plausibility

Unglamorous basis of applied probability:
Things that can happen more ways are more plausible.

Possibility		Ways to produce ●○●
[○○○○]	0%	$0 \times 4 \times 0 = 0$
[●○○○]	25%	$1 \times 3 \times 1 = 3$
[●●○○]	50%	$2 \times 2 \times 2 = 8$
[●●●○]	75%	
[●●●●]	100%	

Counts to plausibility

Unglamorous basis of applied probability:
Things that can happen more ways are more plausible.

Possibility		Ways to produce ●○●
[○○○○]	0%	$0 \times 4 \times 0 = 0$
[●○○○]	25%	$1 \times 3 \times 1 = 3$
[●●○○]	50%	$2 \times 2 \times 2 = 8$
[●●●○]	75%	$3 \times 1 \times 3 = 9$
[●●●●]	100%	$4 \times 0 \times 4 = 0$

Updating

Another draw from the bag: ●

Possibility
[○○○○]
[●○○○]
[●●○○]
[●●●○]
[●●●●]

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●	Ways to produce ●
[○○○○]	$0 \times 4 \times 0 = 0$	0
[●○○○]	$1 \times 3 \times 1 = 3$	1
[●●○○]	$2 \times 2 \times 2 = 8$	2
[●●●○]	$3 \times 1 \times 3 = 9$	3
[●●●●]	$4 \times 0 \times 4 = 0$	4

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●	Ways to produce ●	Ways to produce ●○●●
[○○○○]	$0 \times 4 \times 0 = 0$	0	$0 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$	1	
[●●○○]	$2 \times 2 \times 2 = 8$	2	
[●●●○]	$3 \times 1 \times 3 = 9$	3	
[●●●●]	$4 \times 0 \times 4 = 0$	4	

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●	Ways to produce ●	Ways to produce ●○●●
[○○○○]	$0 \times 4 \times 0 = 0$	0	$0 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$	1	$3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$	2	
[●●●○]	$3 \times 1 \times 3 = 9$	3	
[●●●●]	$4 \times 0 \times 4 = 0$	4	

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●	Ways to produce ●	Ways to produce ●○●●
[○○○○]	$0 \times 4 \times 0 = 0$	0	$0 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$	1	$3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$	2	$8 \times 2 = 16$
[●●●○]	$3 \times 1 \times 3 = 9$	3	
[●●●●]	$4 \times 0 \times 4 = 0$	4	

Updating

Another draw from the bag: ●

Possibility	Ways to produce ●○●	Ways to produce ●	Ways to produce ●○●●
[○○○○]	$0 \times 4 \times 0 = 0$	0	$0 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$	1	$3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$	2	$8 \times 2 = 16$
[●●●○]	$3 \times 1 \times 3 = 9$	3	$9 \times 3 = 27$
[●●●●]	$4 \times 0 \times 4 = 0$	4	$0 \times 4 = 0$

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]										
[●●○○]										
[●●●○]										
[●●●●]										

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]		1	3	3	3	3	9	9	27	$27 = 1^6 \times 3^3$
[●●○○]										
[●●●○]										
[●●●●]										

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]		1	3	3	3	3	9	9	27	$27 = 1^6 \times 3^3$
[●●○○]		2	4	8	16	32	64	128	256	$512 = 2^6 \times 2^3$
[●●●○]										
[●●●●]										

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]		1	3	3	3	3	9	9	27	$27 = 1^6 \times 3^3$
[●●○○]		2	4	8	16	32	64	128	256	$512 = 2^6 \times 2^3$
[●●●○]		3	3	9	27	81	81	243	243	$729 = 3^6 \times 1^3$
[●●●●]		4	0	0	0	0	0	0	0	$0 = 4^6 \times 0^3$

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]		1	3	3	3	3	9	9	27	$27 = 1^6 \times 3^3$
[●●○○]		2	4	8	16	32	64	128	256	$512 = 2^6 \times 2^3$
[●●●○]		3	3	9	27	81	81	243	243	$729 = 3^6 \times 1^3$
[●●●●]		4	0	0	0	0	0	0	0	$0 = 4^6 \times 0^3$

The whole sample

Possibility	Observations:	●	○	●	●	●	○	●	○	●
[○○○○]		0	0	0	0	0	0	0	0	$0 = 0^6 \times 4^3$
[●○○○]		1	3	3	3	3	9	9	27	$27 = 1^6 \times 3^3$
[●●○○]		2	4	8	16	32	64	128	256	$512 = 2^6 \times 2^3$
[●●●○]		3	3	9	27	81	81	243	243	$729 = 3^6 \times 1^3$
[●●●●]		4	0	0	0	0	0	0	0	$0 = 4^6 \times 0^3$

Ways for p to produce $W,L = (4p)^W \times (4-4p)^L$

Probability

Probability: Non-negative values that sum to one

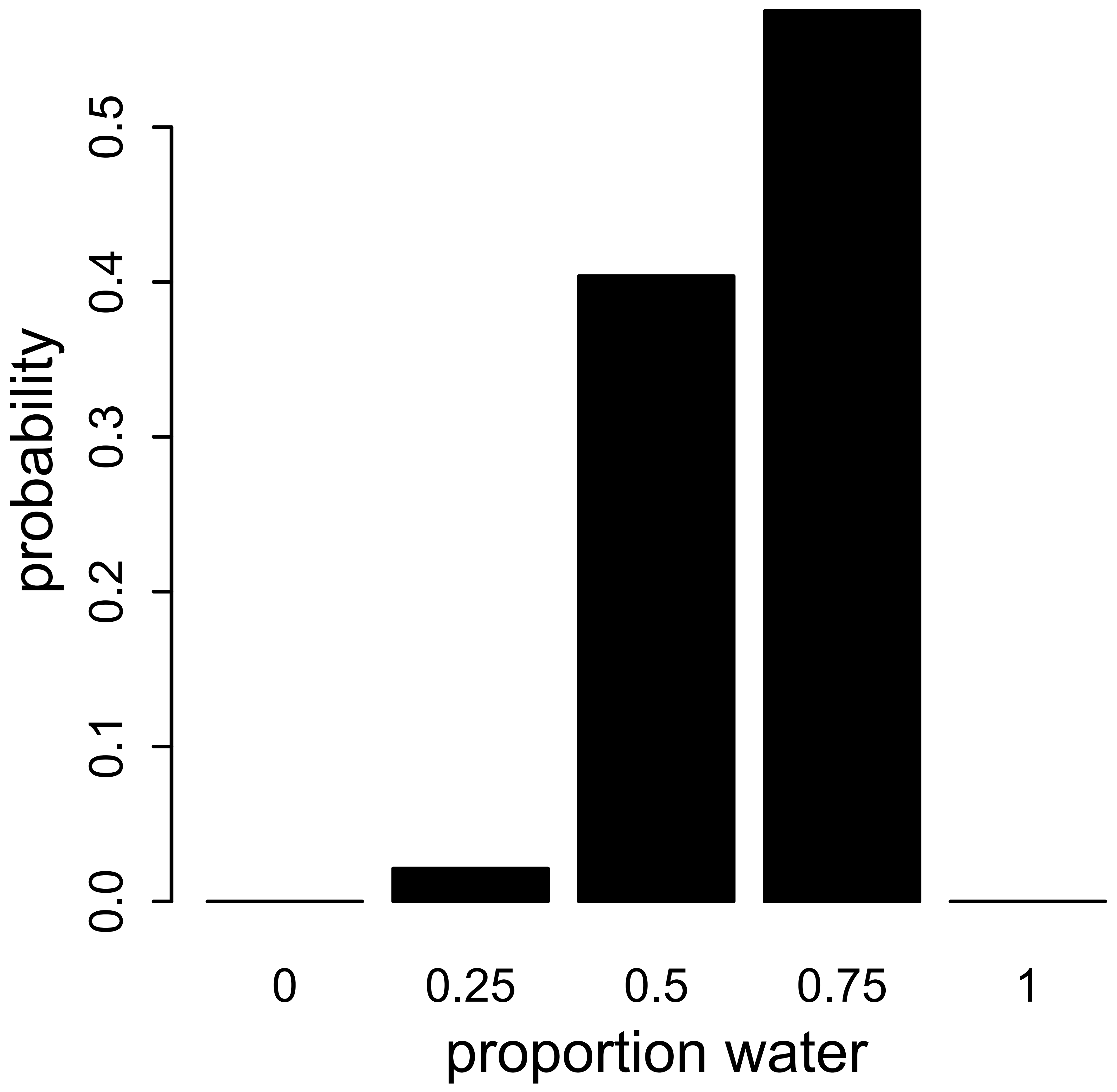
Suppose W=20, L=10. Then p=0.5 has

$$2^W \times 2^L = 1{,}073{,}741{,}824$$

ways to produce sample. Better to convert to probability.

Probability

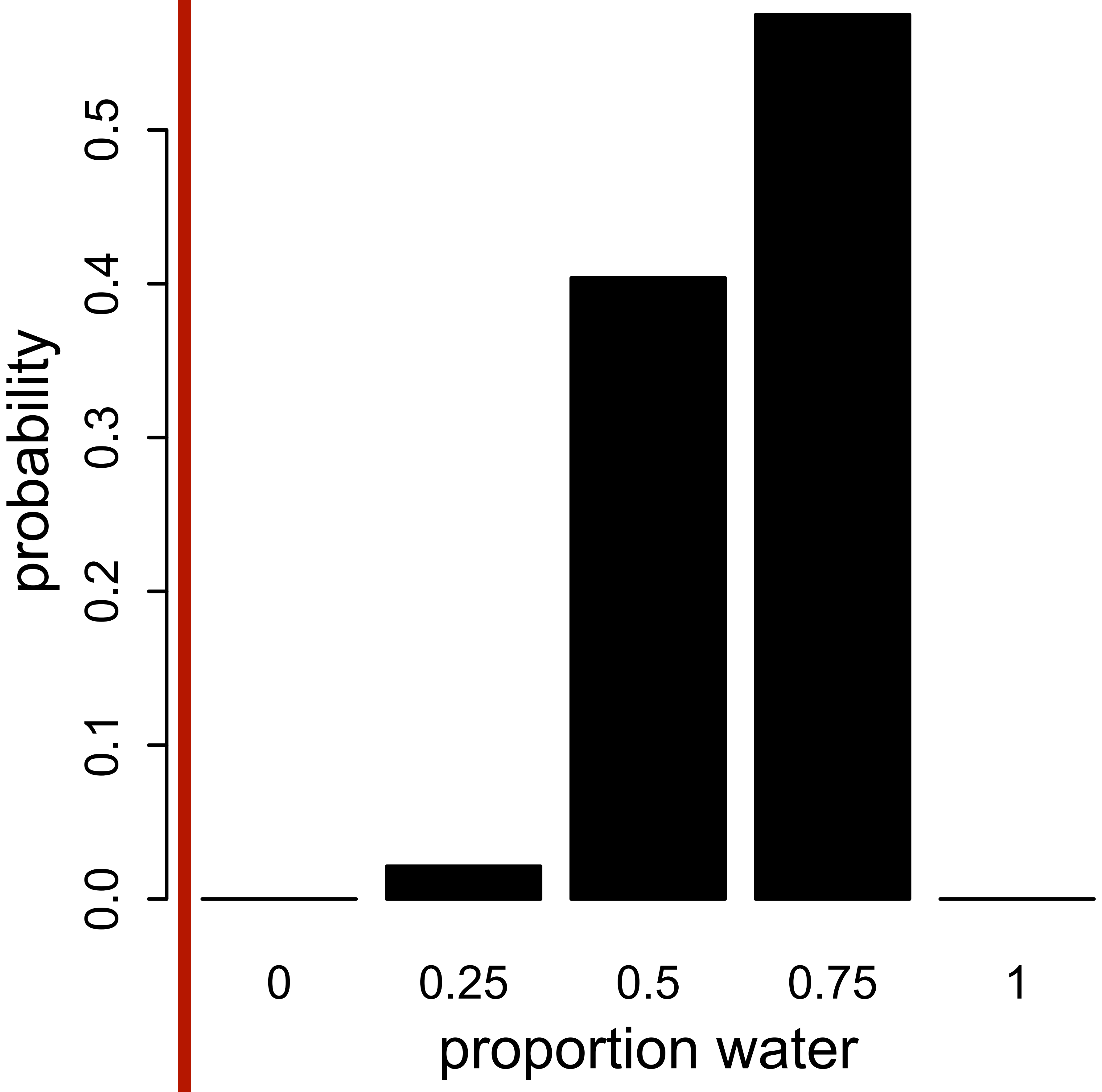
Possible proportion	Ways to produce sample	Probability of proportion
0	0	0
0.25	27	0.02
0.5	512	0.40
0.75	729	0.57
1	0	0



Probability

Posterior distribution

Possible proportion	Ways to produce sample	Probability of proportion
0	0	0
0.25	27	0.02
0.5	512	0.40
0.75	729	0.57
1	0	0

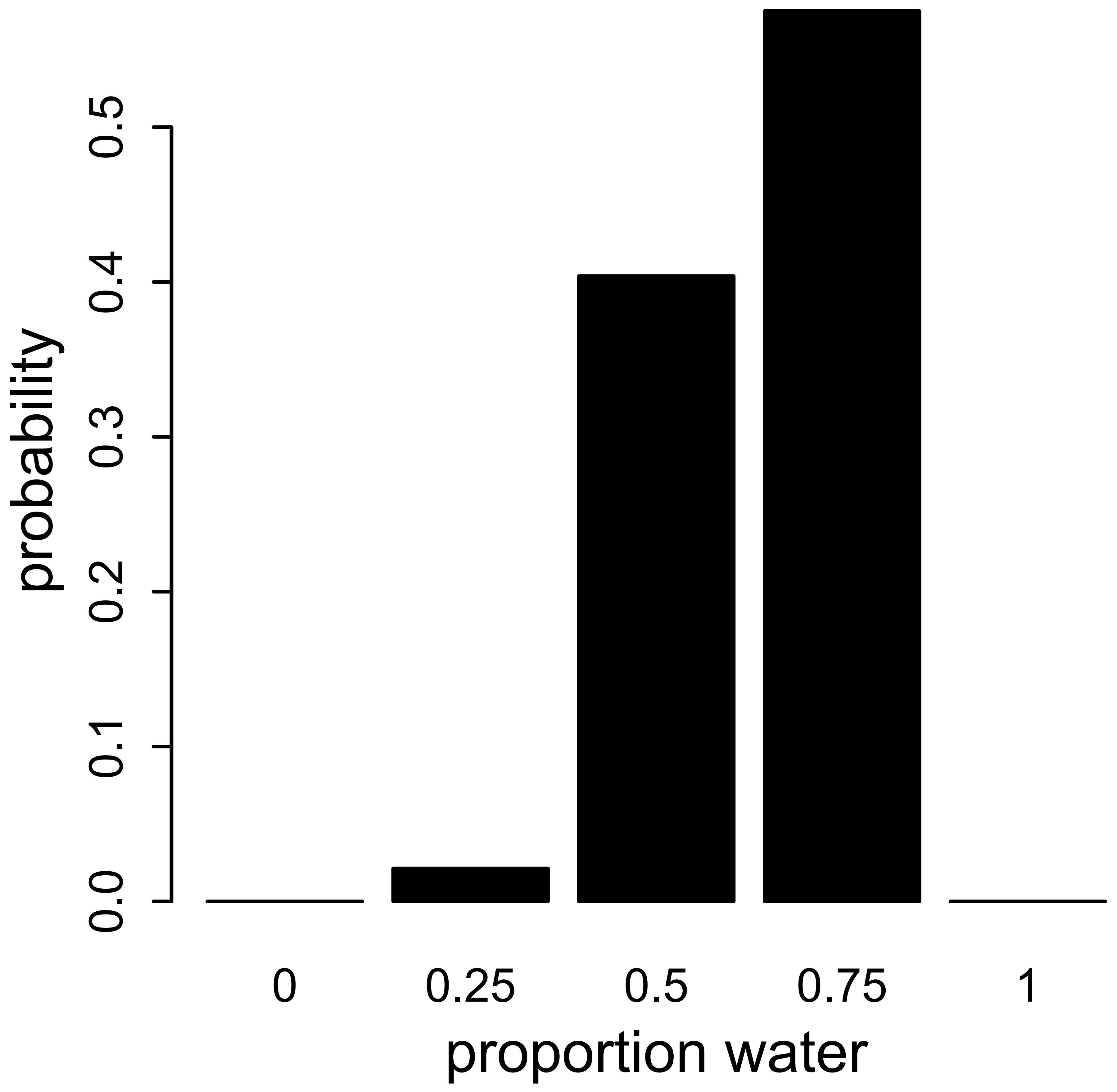


Probability

R code 2.1

```
sample <- c("W","L","W","W","W","L","W","L","W")
W <- sum(sample=="W") # number of W observed
L <- sum(sample=="L") # number of L observed
p <- c(0,0.25,0.5,0.75,1) # proportions W
ways <- sapply( p , function(q) (q*4)^W * ((1-q)*4)^L )
prob <- ways/sum(ways)
cbind( p , ways , prob )
```

```
        p ways       prob
[1,] 0.00    0 0.00000000
[2,] 0.25   27 0.02129338
[3,] 0.50  512 0.40378549
[4,] 0.75  729 0.57492114
[5,] 1.00    0 0.00000000
```



Workflow

(1) Define generative model of the sample

(2) Define a specific estimand

(3) Design a statistical way to produce estimate

(4) Test (3) using (1)

(5) Analyze sample, summarize

Test Before You Est(imate)

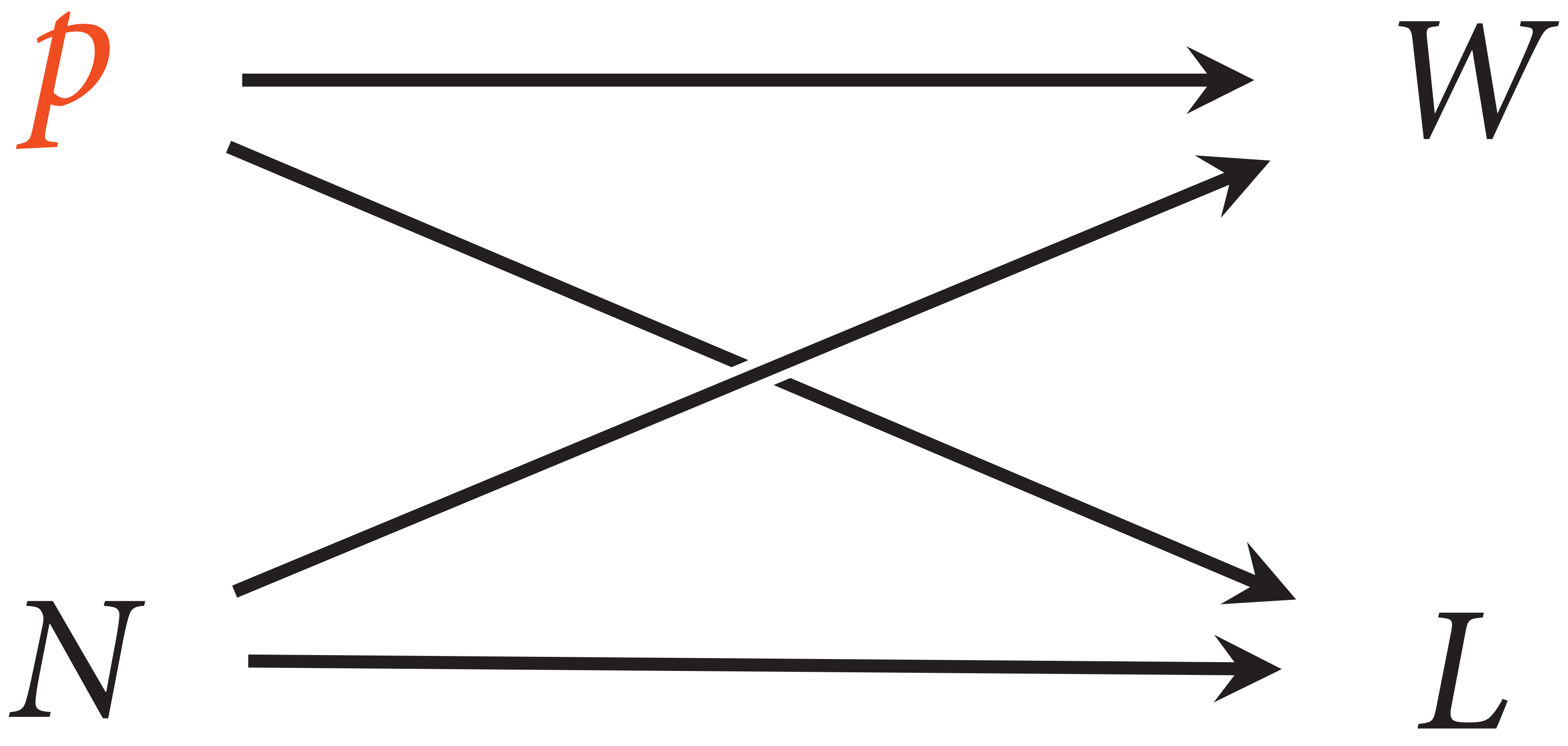
(1) Code a generative simulation

(2) Code an estimator

(3) Test (2) with (1)

Extremely powerful, fun

Generative simulation



$$W, L = f(p, N)$$

```r
# function to toss a globe covered p by water N times
sim_globe <- function( p=0.7 , N=9 ) {
    sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
}
```

R code 2.3

```r
# function to toss a globe covered p by water N times
sim_globe <- function( p=0.7 , N=9 ) {
    sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
}
```

R code 2.3

Possible observations

Number of tosses

Probability of each possible observation

```r
# function to toss a globe covered p by water N times
sim_globe <- function( p=0.7 , N=9 ) {
    sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
}
```
R code 2.3

```r
sim_globe()
```
R code 2.4

```
[1] "L" "W" "W" "W" "L" "L" "L" "W" "L"
```

```
sim_globe()
```

R code 2.4

```
[1] "L" "W" "W" "W" "L" "L" "L" "W" "L"
```

```
replicate(sim_globe(p=0.5,N=9),n=10)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "W"  "L"  "L"  "W"  "W"  "L"  "L"  "W"  "W"  "L"
 [2,] "W"  "L"  "W"  "L"  "W"  "L"  "L"  "W"  "L"  "L"
 [3,] "W"  "L"  "L"  "L"  "L"  "W"  "L"  "W"  "W"  "W"
 [4,] "W"  "W"  "L"  "W"  "L"  "W"  "W"  "W"  "W"  "W"
 [5,] "L"  "W"  "W"  "W"  "W"  "W"  "L"  "W"  "L"  "L"
 [6,] "L"  "W"  "L"  "L"  "W"  "L"  "W"  "W"  "W"  "W"
 [7,] "W"  "W"  "W"  "L"  "W"  "W"  "W"  "L"  "L"  "L"
 [8,] "L"  "W"  "L"  "L"  "L"  "W"  "L"  "W"  "W"  "W"
 [9,] "W"  "L"  "L"  "W"  "L"  "W"  "W"  "W"  "L"  "L"
```

R code 2.4

```
sim_globe()
```

```
[1] "L" "W" "W" "W" "L" "L" "L" "W" "L"
```

Test the simulation on extreme settings

R code 2.5

```
sim_globe( p=1 , N=11 )
```

```
 [1] "W" "W" "W" "W" "W" "W" "W" "W" "W" "W" "W"
```

R code 2.6

```
sum( sim_globe( p=0.5 , N=1e4 ) == "W" ) / 1e4
```

```
[1] 0.505
```

IF YOU TEST
NOTHING
YOU MISS
EVERYTHING

Code the estimator

$$\text{Ways for } p \text{ to produce } W,L = (4p)^W \times (4{-}4p)^L$$

```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

Code the estimator

$$\text{Ways for } p \text{ to produce } W,L = (4p)^W \times (4-4p)^L$$

```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

Code the estimator

$$\text{Ways for } p \text{ to produce } W,L = (4p)^W \times (4-4p)^L$$

```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

Code the estimator

$$\text{Ways for } p \text{ to produce } W,L = (4p)^W \times (4-4p)^L$$

```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

Code the estimator

$$\text{Ways for } p \text{ to produce } W,L = (4p)^W \times (4-4p)^L$$

```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

```r
compute_posterior( sim_globe() )
```

R code 2.9

```
  poss ways  post                 bars
1 0.00    0 0.000
2 0.25  243 0.291 ######
3 0.50  512 0.612 ############
4 0.75   81 0.097 ##
5 1.00    0 0.000
```

(1) Test the estimator where the answer is known

(2) Explore different sampling designs

(3) Develop intuition for sampling and estimation

PAUSE

More possibilities

4-sided globe

[0 0.25 0.5 0.75 1]

More possibilities

4-sided globe

[0 0.25 0.5 0.75 1]

10-sided globe

[0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1]

More possibilities

4-sided globe

[0 0.25 0.5 0.75 1]

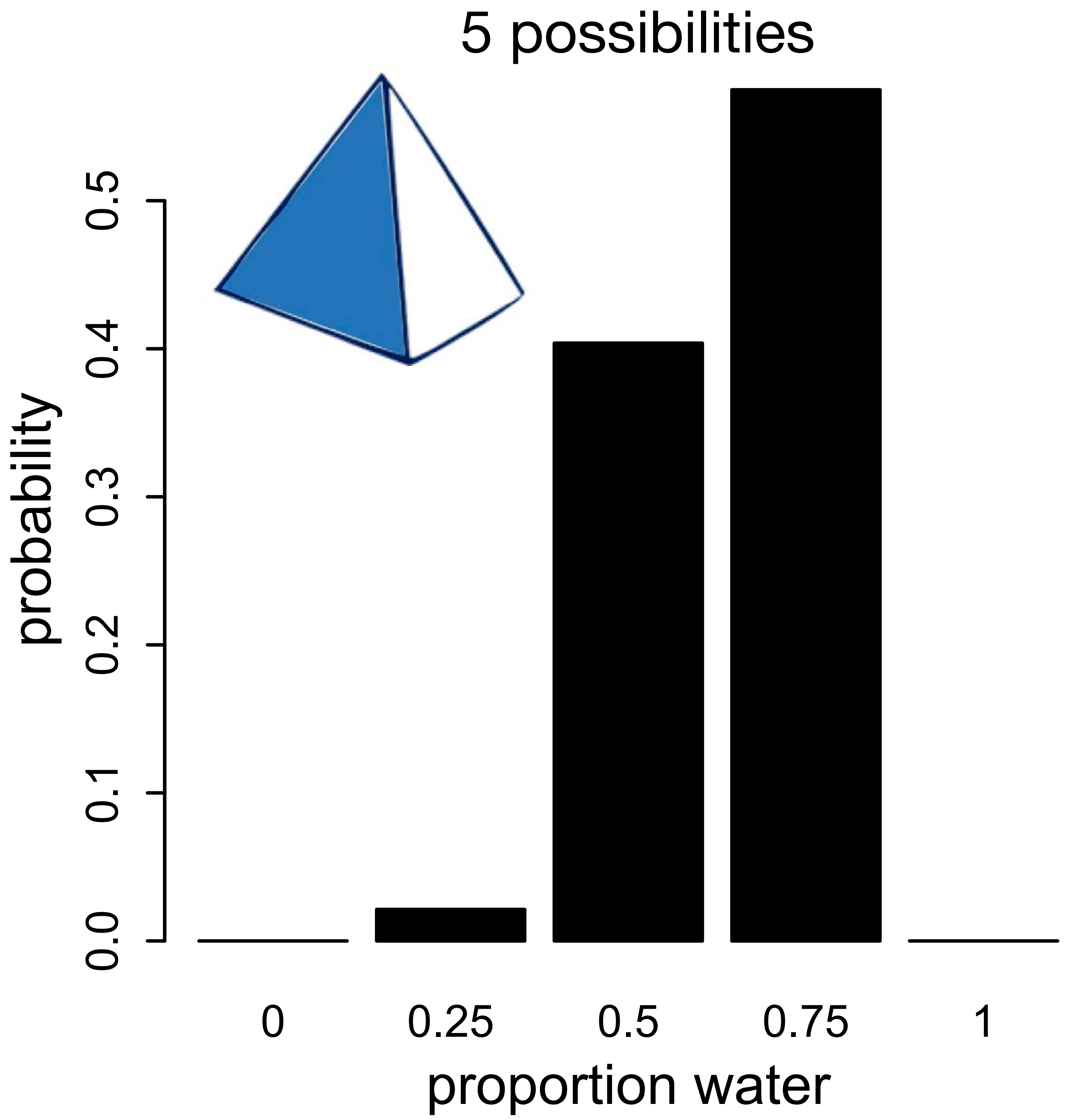
10-sided globe

[0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1]

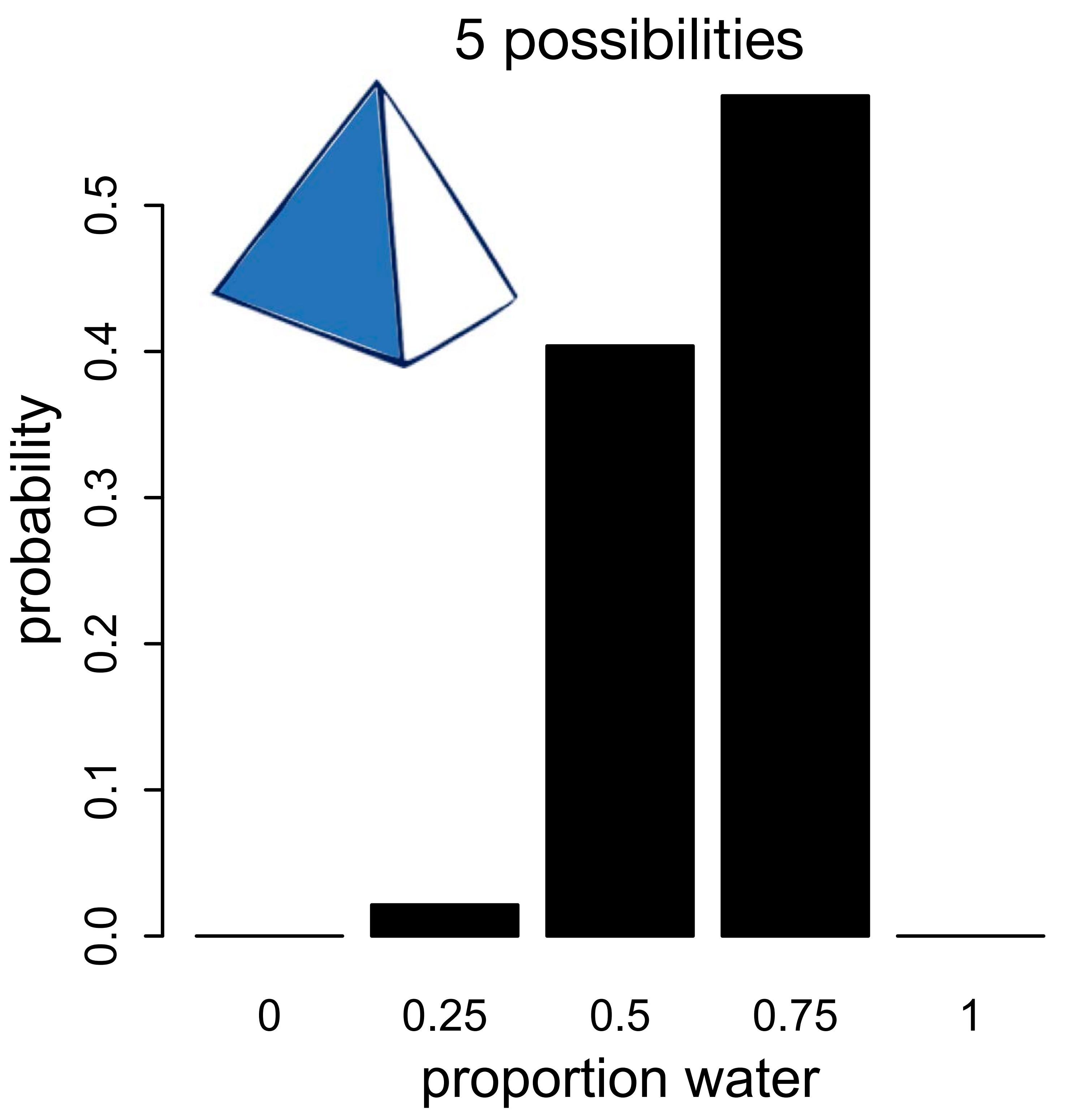
20-sided globe

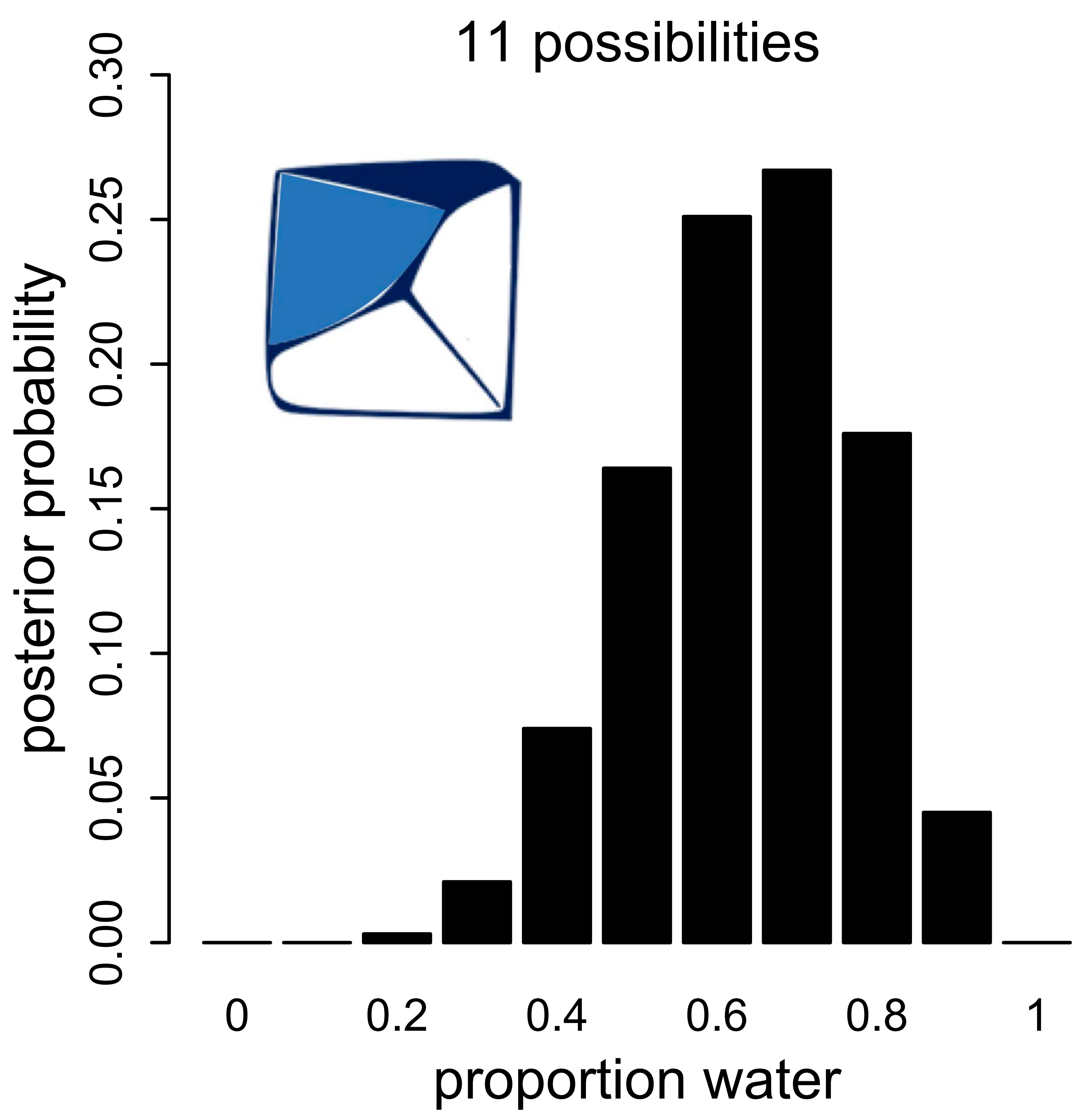
[0 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1]

More possibilities

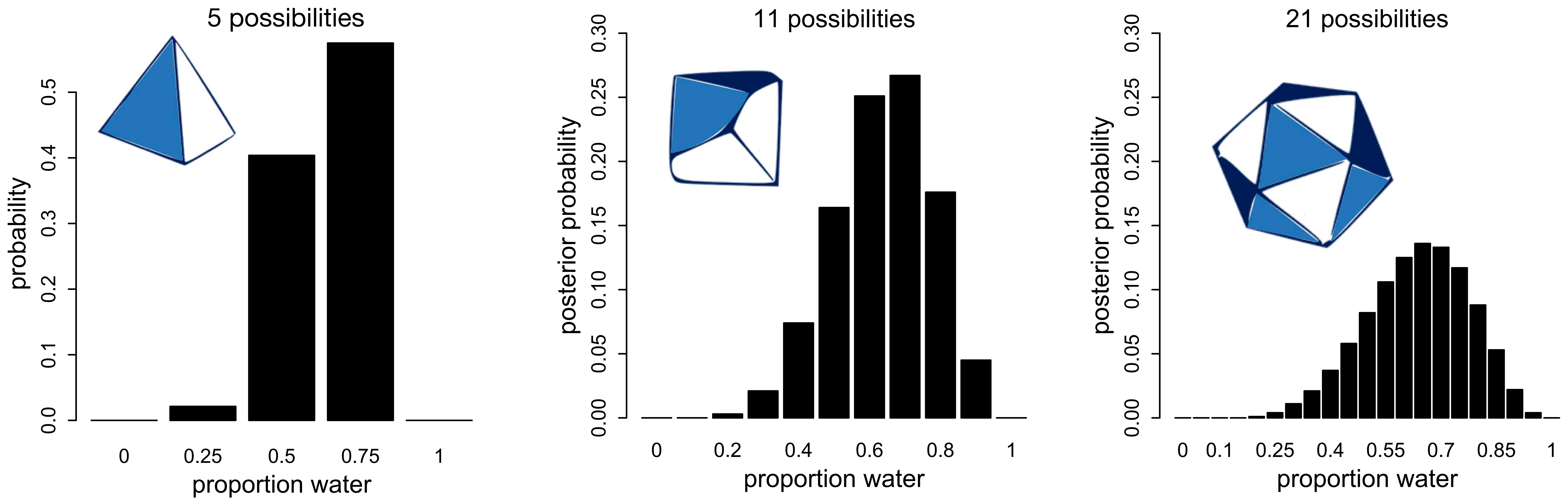


More possibilities





More possibilities



Infinite possibilities

The globe is a polyhedron with an infinite number of sides

The posterior probability of any "side" p is proportional to:

$$p^W(1-p)^L$$

Infinite possibilities

The globe is a polyhedron with an infinite number of sides

The posterior probability of any "side" p is proportional to:

$$p^W (1-p)^L$$

Only trick is normalizing to probability. After a little calculus:

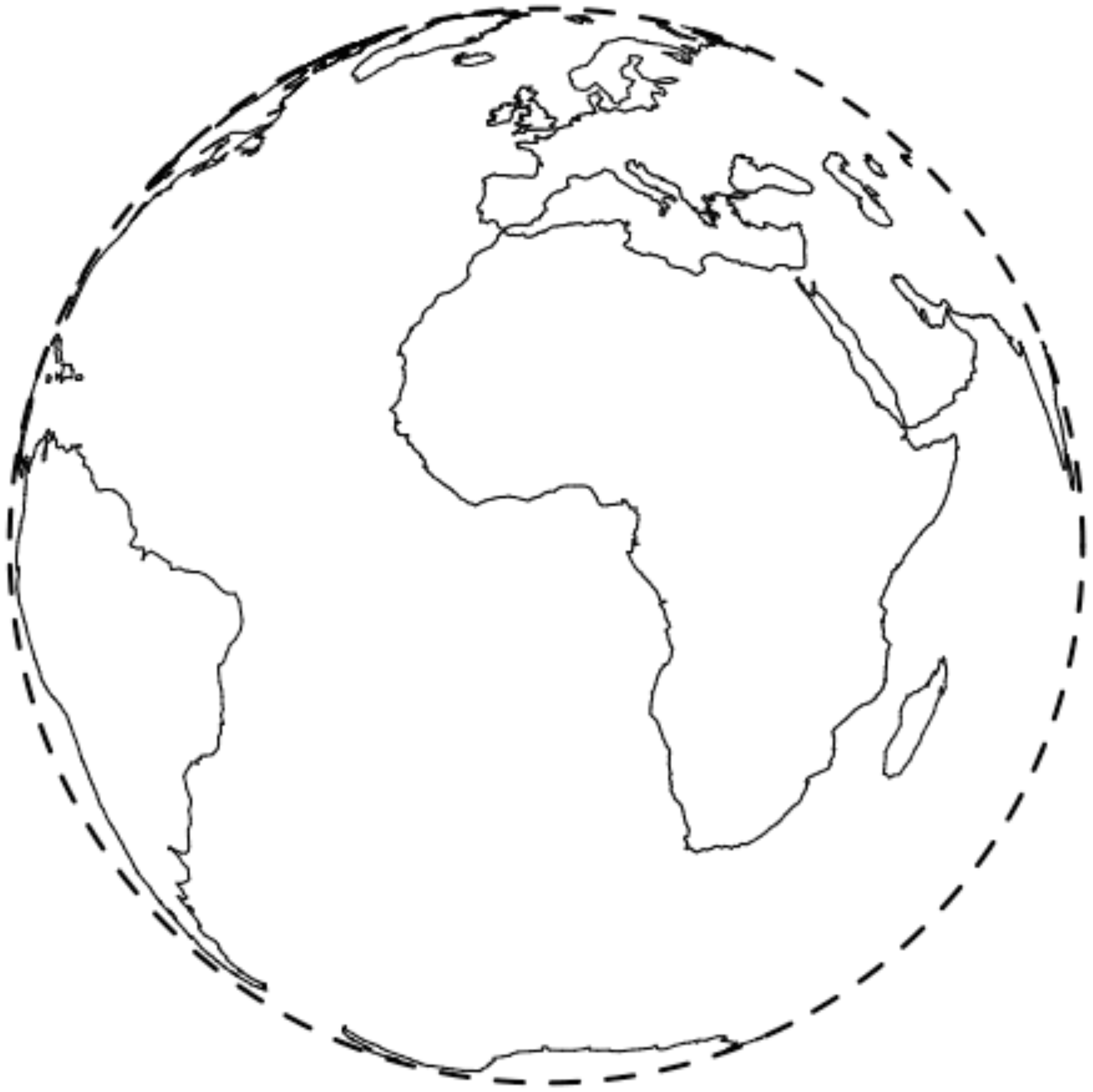
$$\text{Posterior probability of } p = \frac{(W+L+1)!}{W!L!} p^W (1-p)^L$$

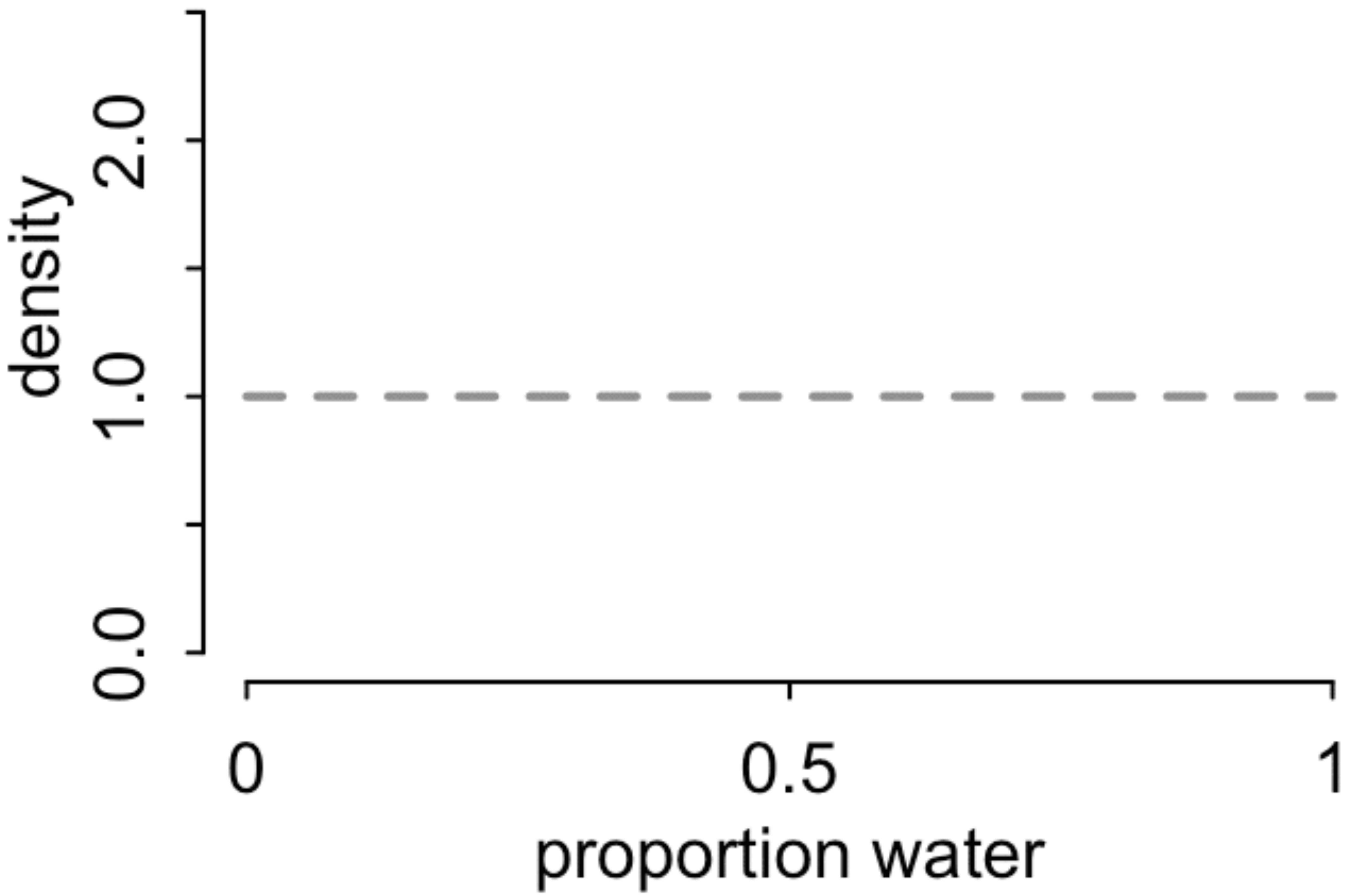
Infinite possibilities

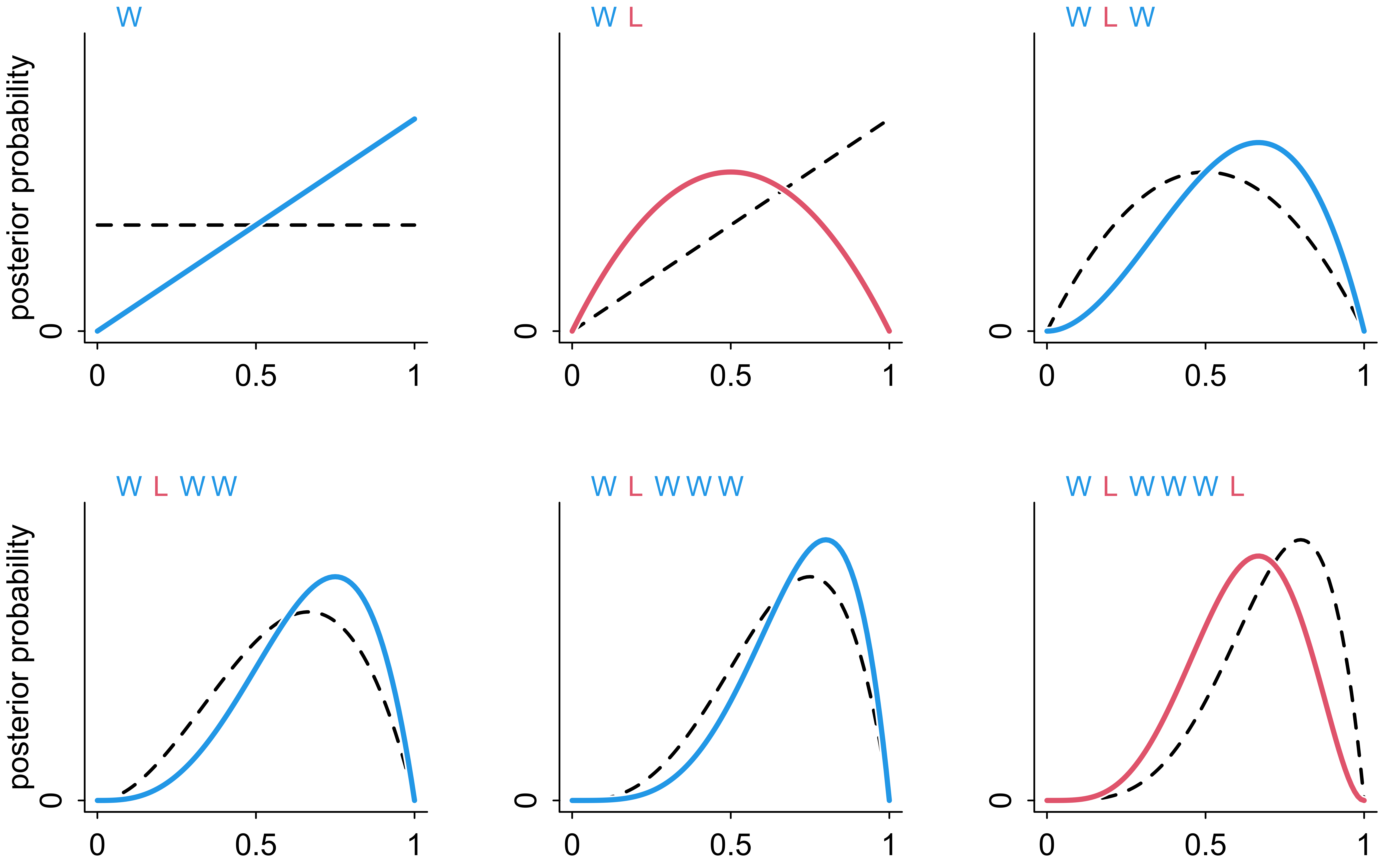
The "Beta" distribution

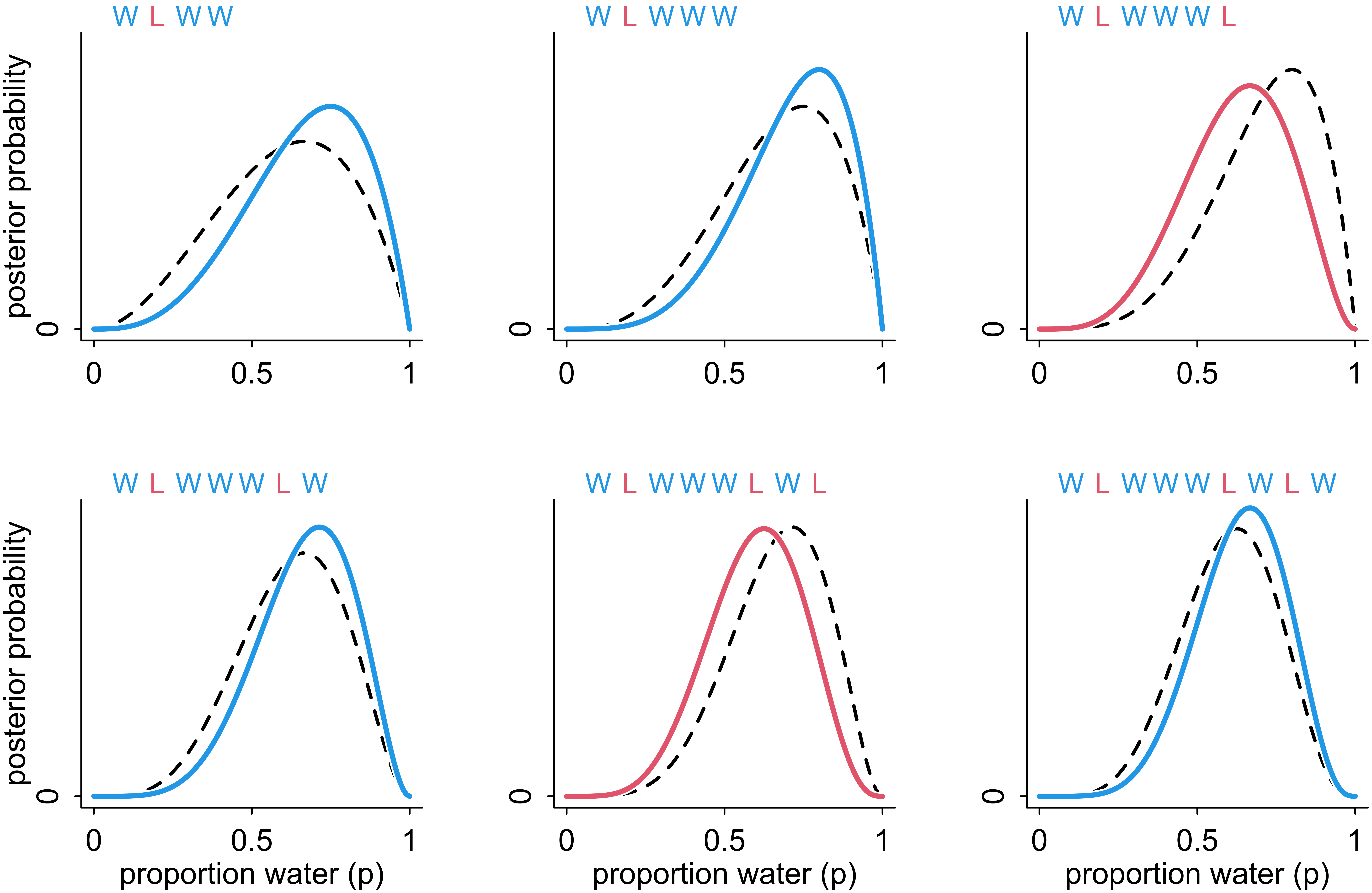
$$\text{Posterior probability of } p = \underbrace{\frac{(W+L+1)!}{W!L!}}_{\textit{Normalizing constant}} \overbrace{p^{W}(1-p)^{L}}^{\textit{relative number of ways to observe sample}}$$

Ten tosses of the globe

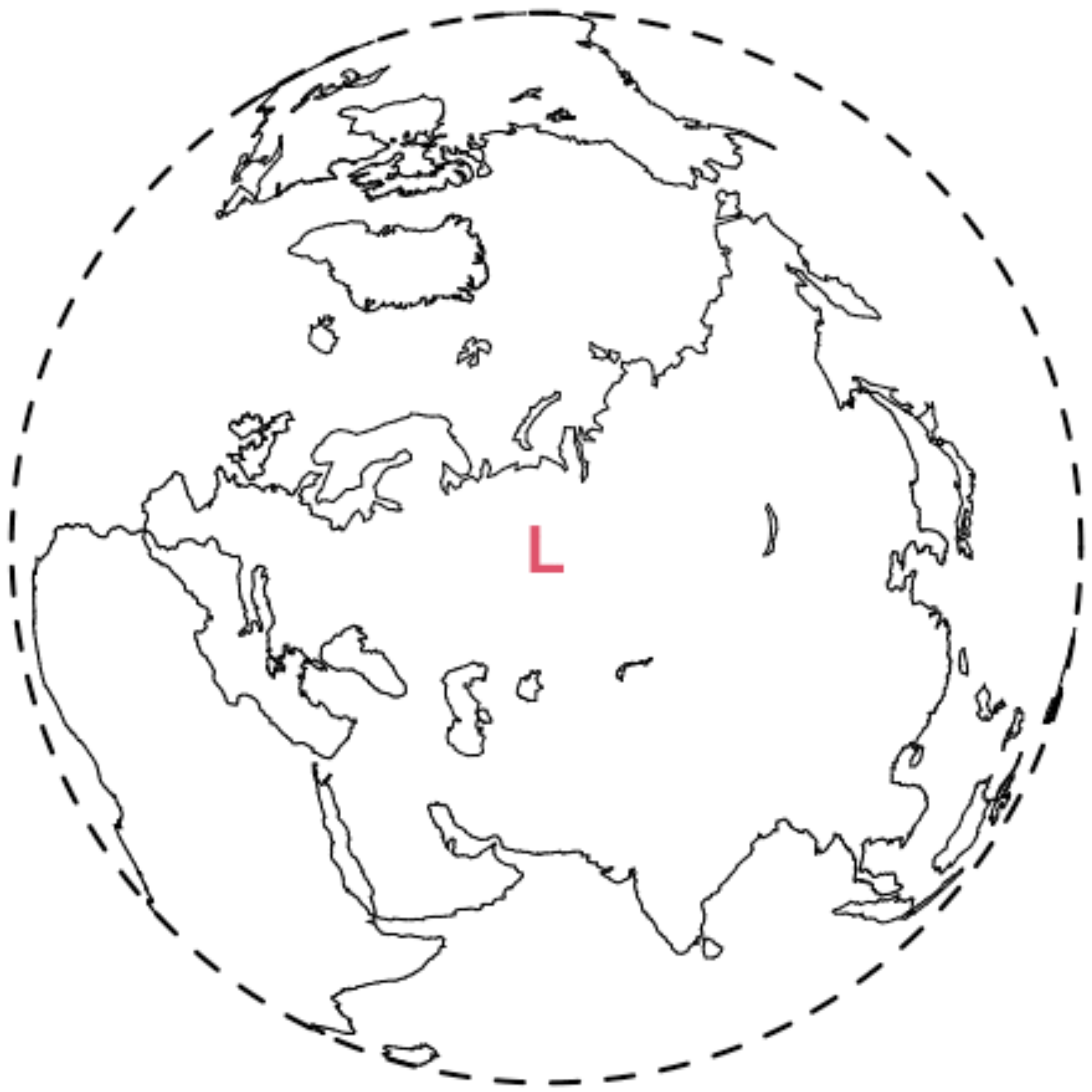


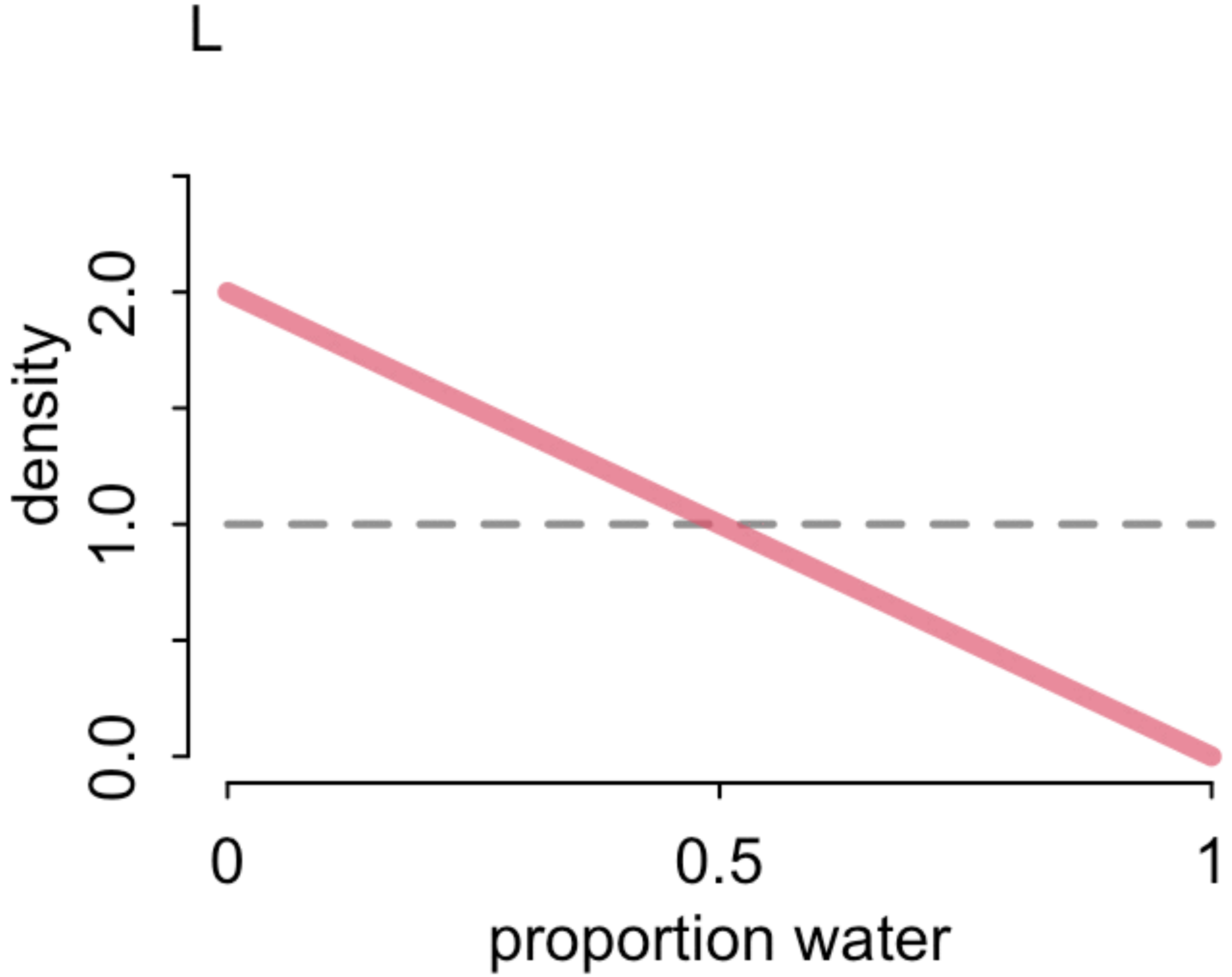



W
W L
W L W
W L W W
W L W W W
W L W W W L
posterior probability
posterior probability
0
0.5
1

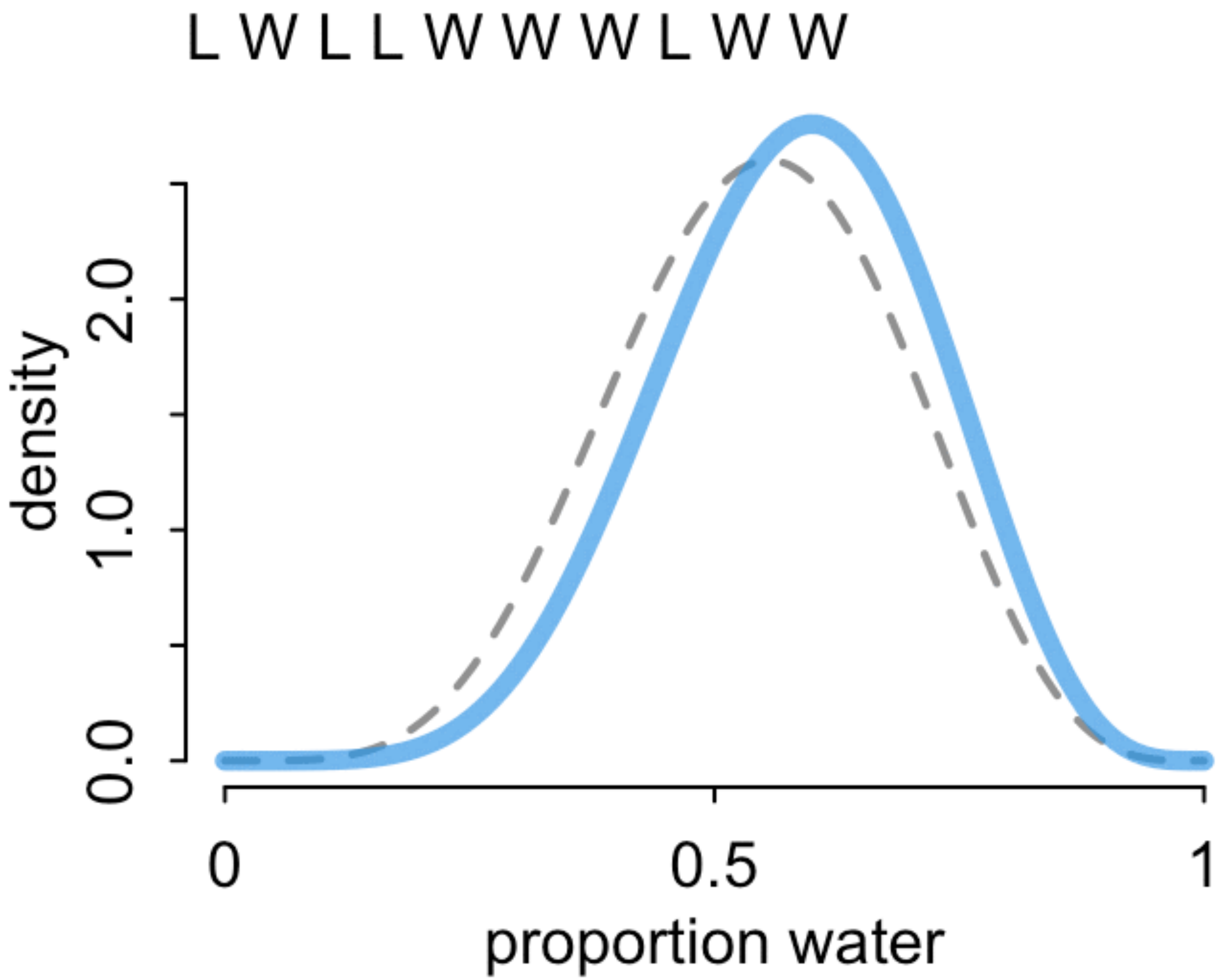

W L W W
W L W W W
W L W W W L
W L W W W L W
W L W W W L W L
W L W W W L W L W
posterior probability
0
0
0.5
1
proportion water (p)

(1) No minimum sample size

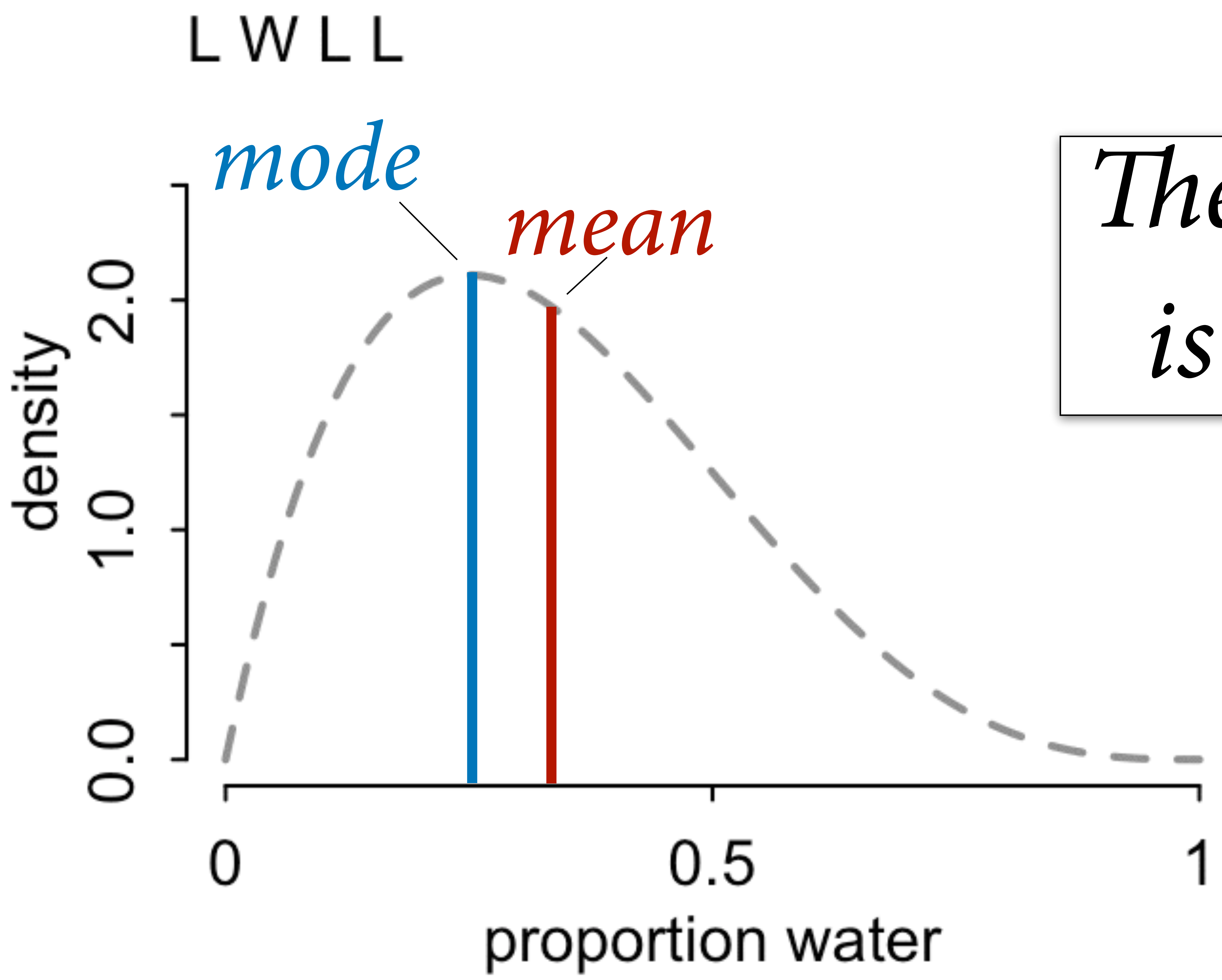




(2) Shape embodies sample size



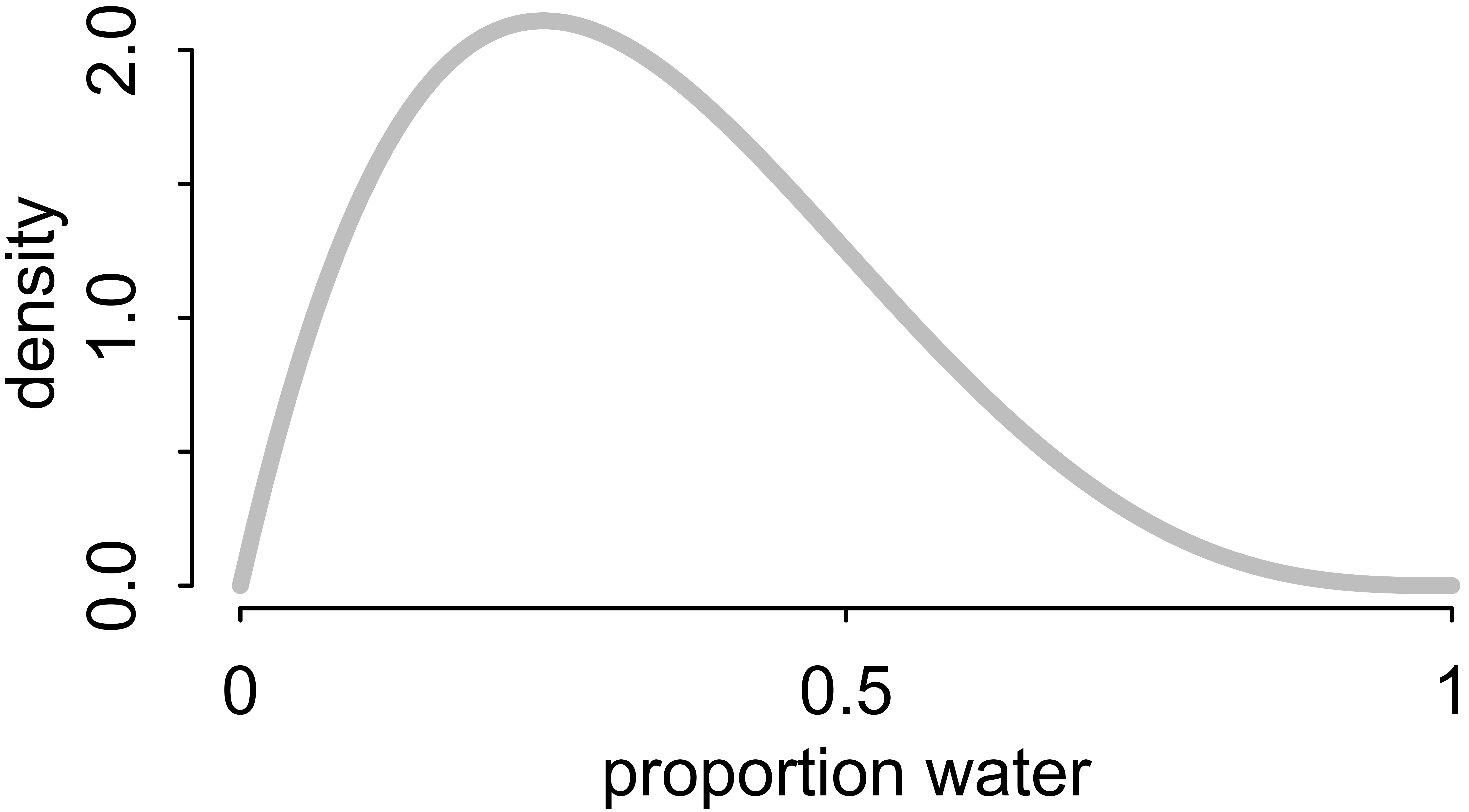
(3) No point estimate



*The **distribution** is the estimate*

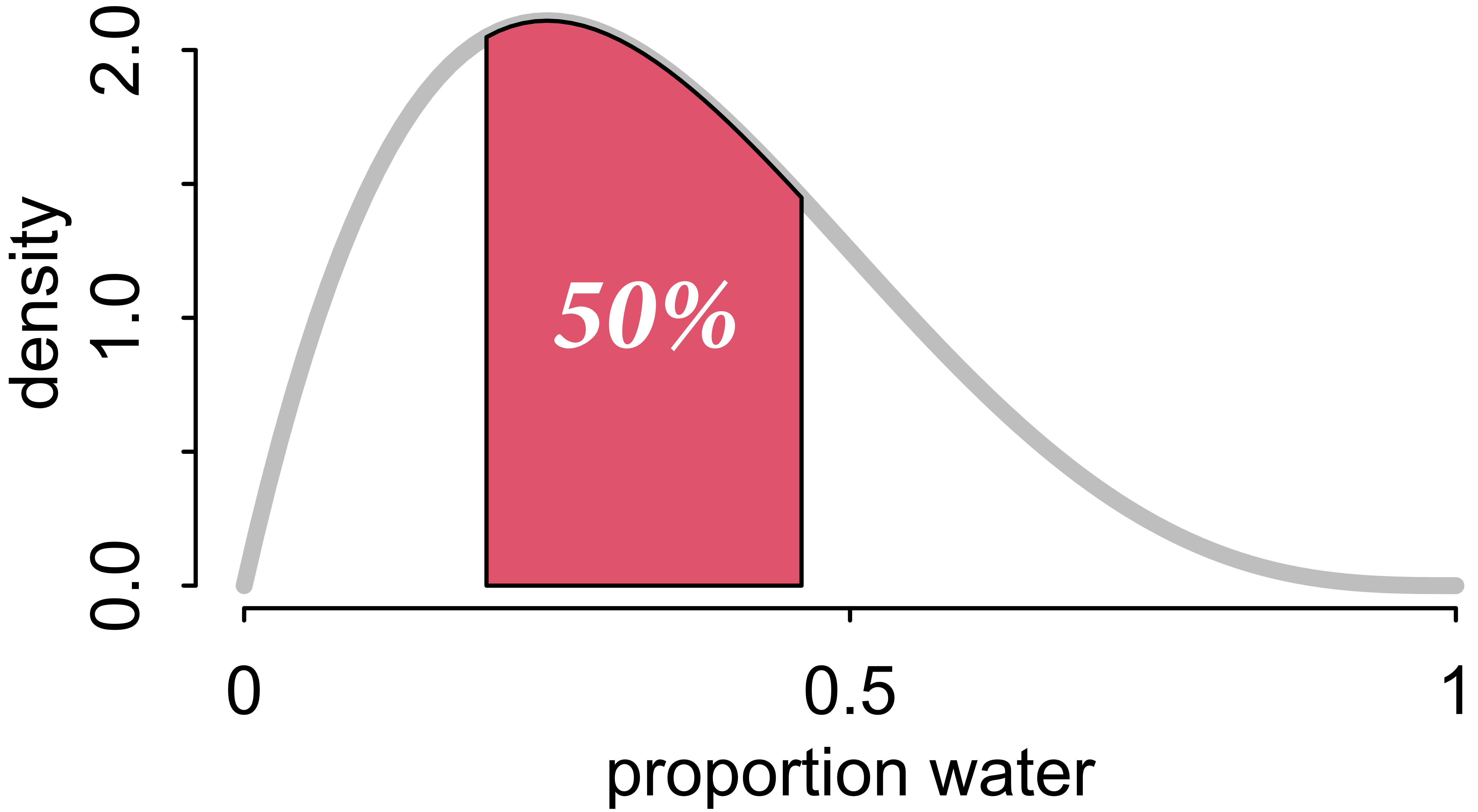
Always use the entire distribution

(4) No one true interval



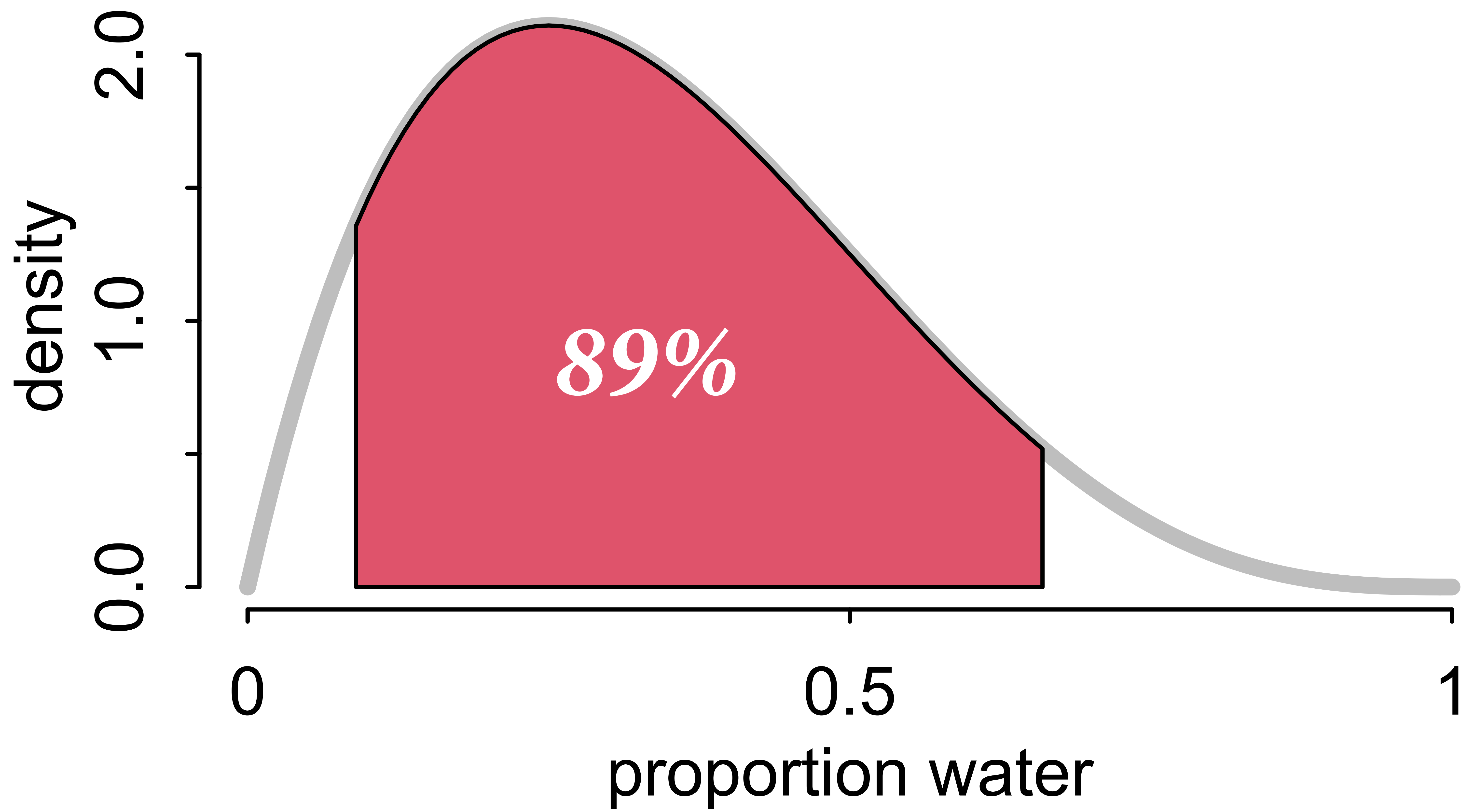
Intervals communicate shape of posterior

(4) No one true interval



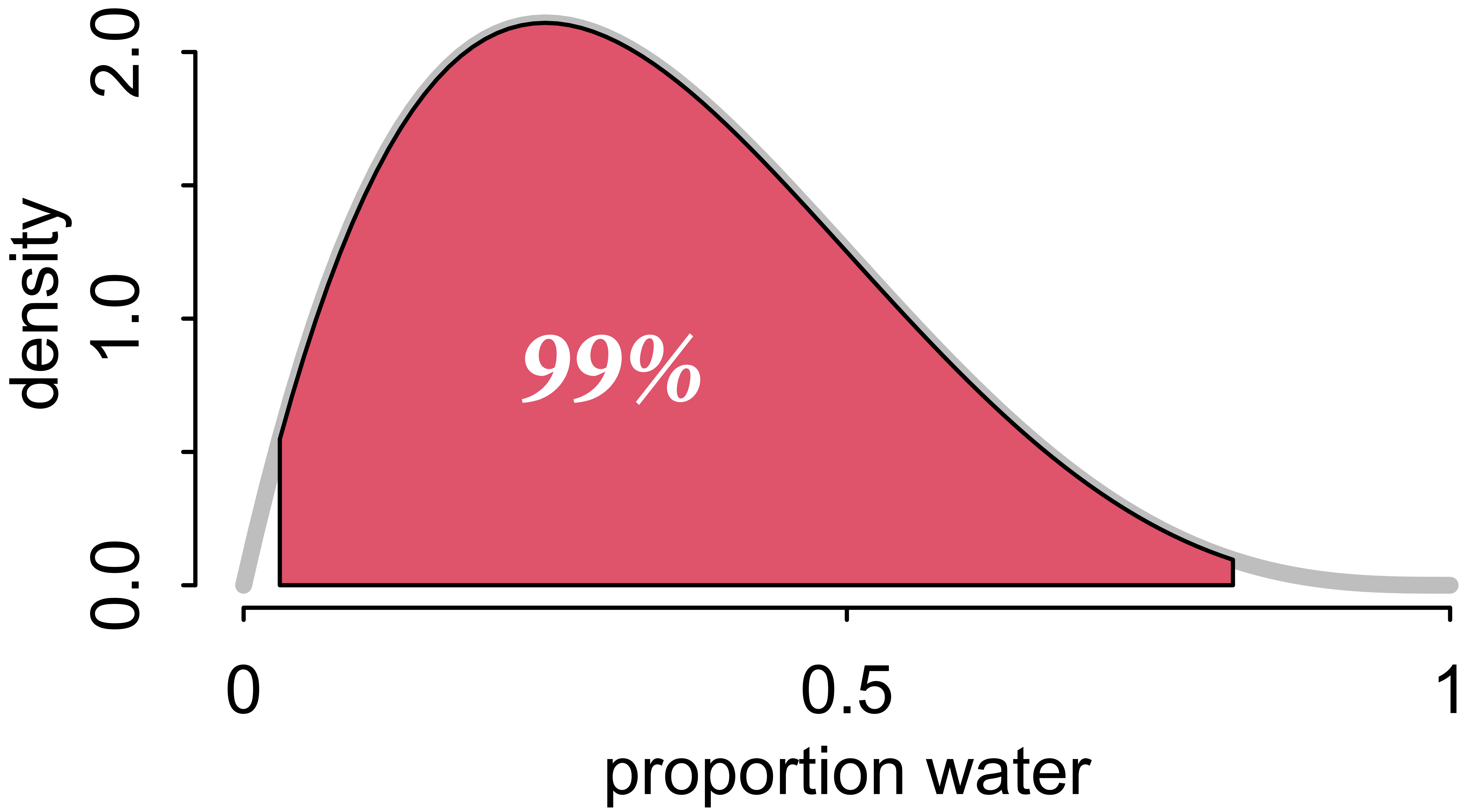
Intervals communicate shape of posterior

(4) No one true interval



Intervals communicate shape of posterior

(4) No one true interval



Intervals communicate shape of posterior

95% is obvious superstition. Nothing magical happens at the boundary.

Letters From My Reviewers

"The author uses these cute 89% intervals, but we need to see the 95% intervals so we can tell whether any of the effects are robust."

That an arbitrary interval contains an arbitrary value is not meaningful. Use the whole distribution.

Workflow

(1) Define generative model of the sample

(2) Define a specific estimand

(3) Design a statistical way to produce estimate

(4) Test (3) using (1)

(5) Analyze sample, summarize

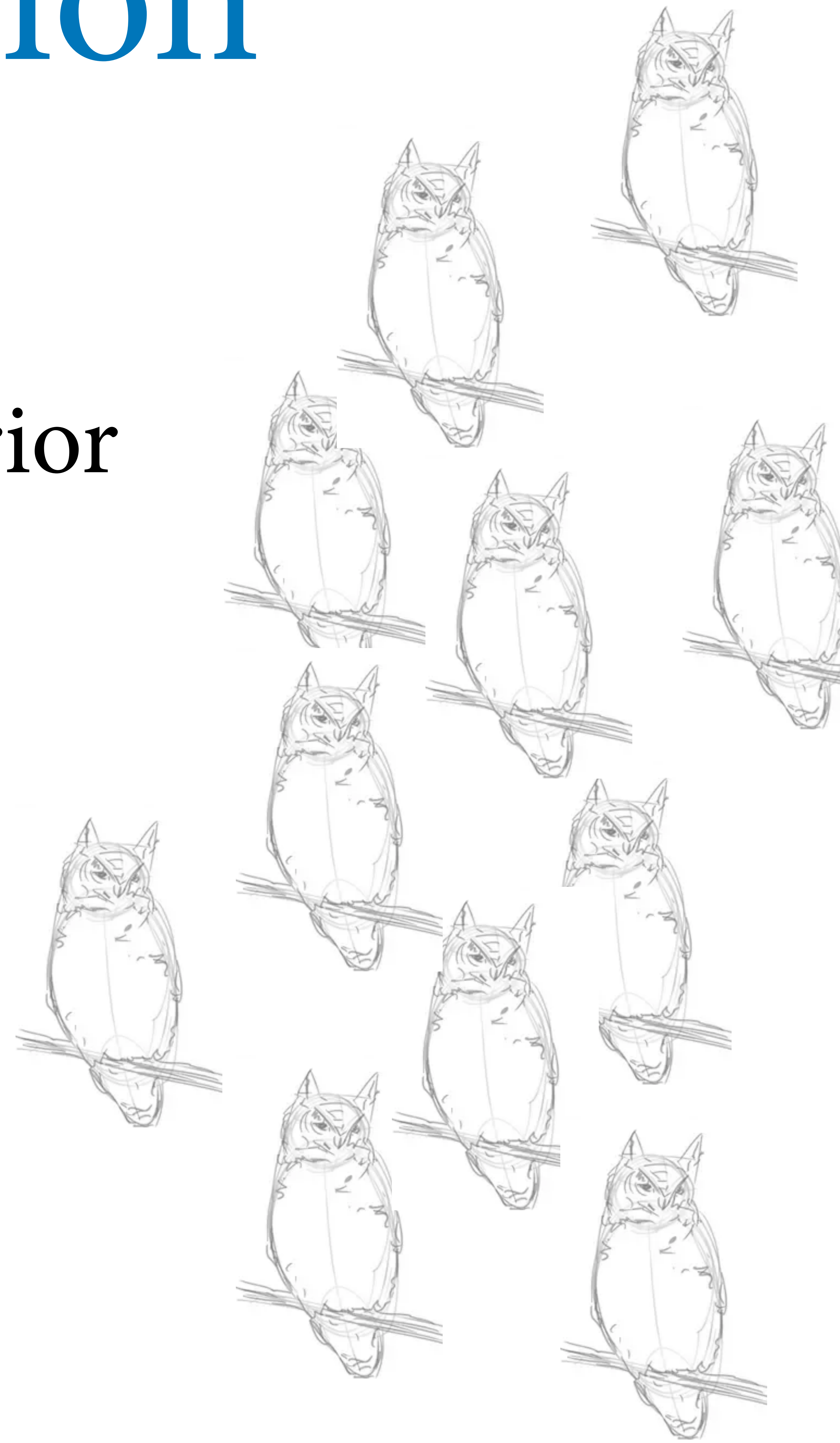
From Posterior to Prediction

Implications of model depend upon **entire** posterior

Must average any inference over entire posterior

This usually requires integral calculus

OR we can just take samples from the posterior

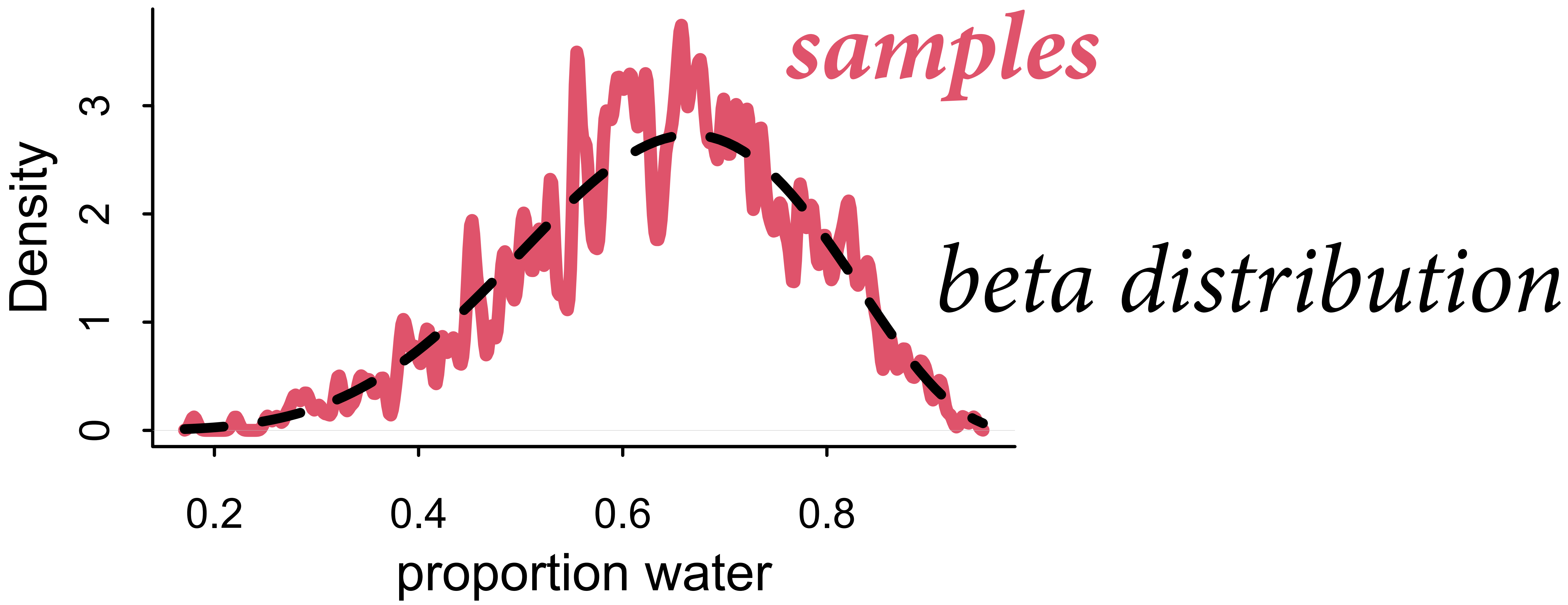


Sampling the posterior

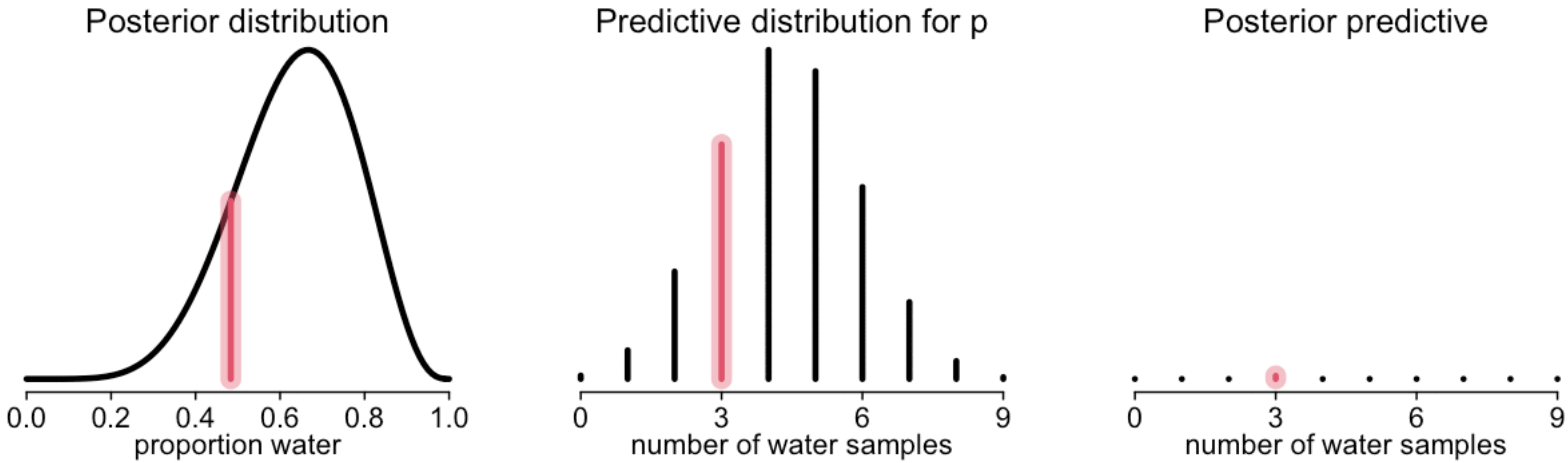
```
post_samples <- rbeta( 1e3 , 6+1 , 3+1 )
```

R code 2.19

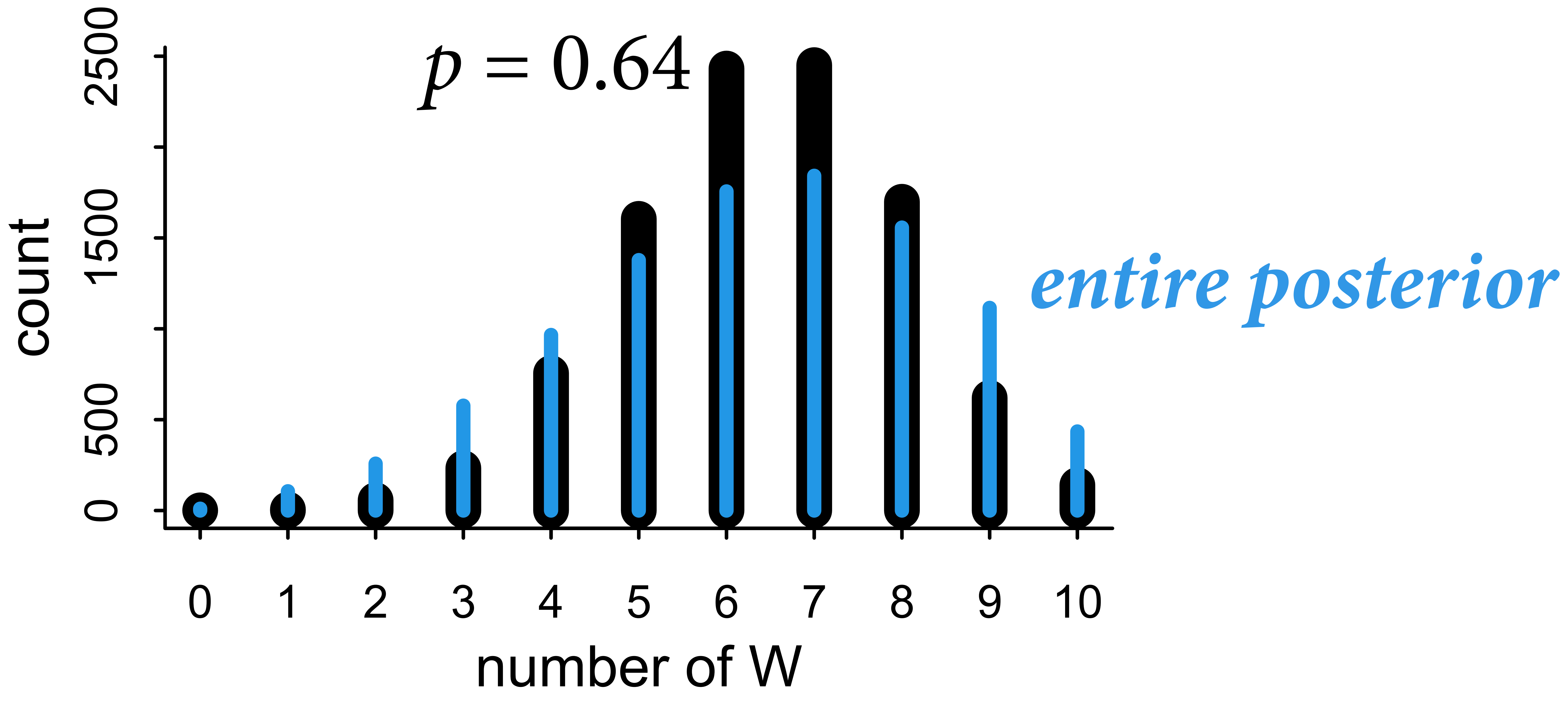
```
dens( post_samples , lwd=4 , col=2 , xlab="proportion water" , adj=0.1 )
curve( dbeta(x,6+1,3+1) , add=TRUE , lty=2 , lwd=3 )
```



Uncertainty ⇨ Causal model ⇨ Implications



```
# now simulate posterior predictive distribution
post_samples <- rbeta(1e4,6+1,3+1)
pred_post <- sapply( post_samples , function(p) sum(sim_globe(p,10)=="W") )
tab_post <- table(pred_post)
for ( i in 0:10 ) lines(c(i,i),c(0,tab_post[i+1]),lwd=4,col=4)
```



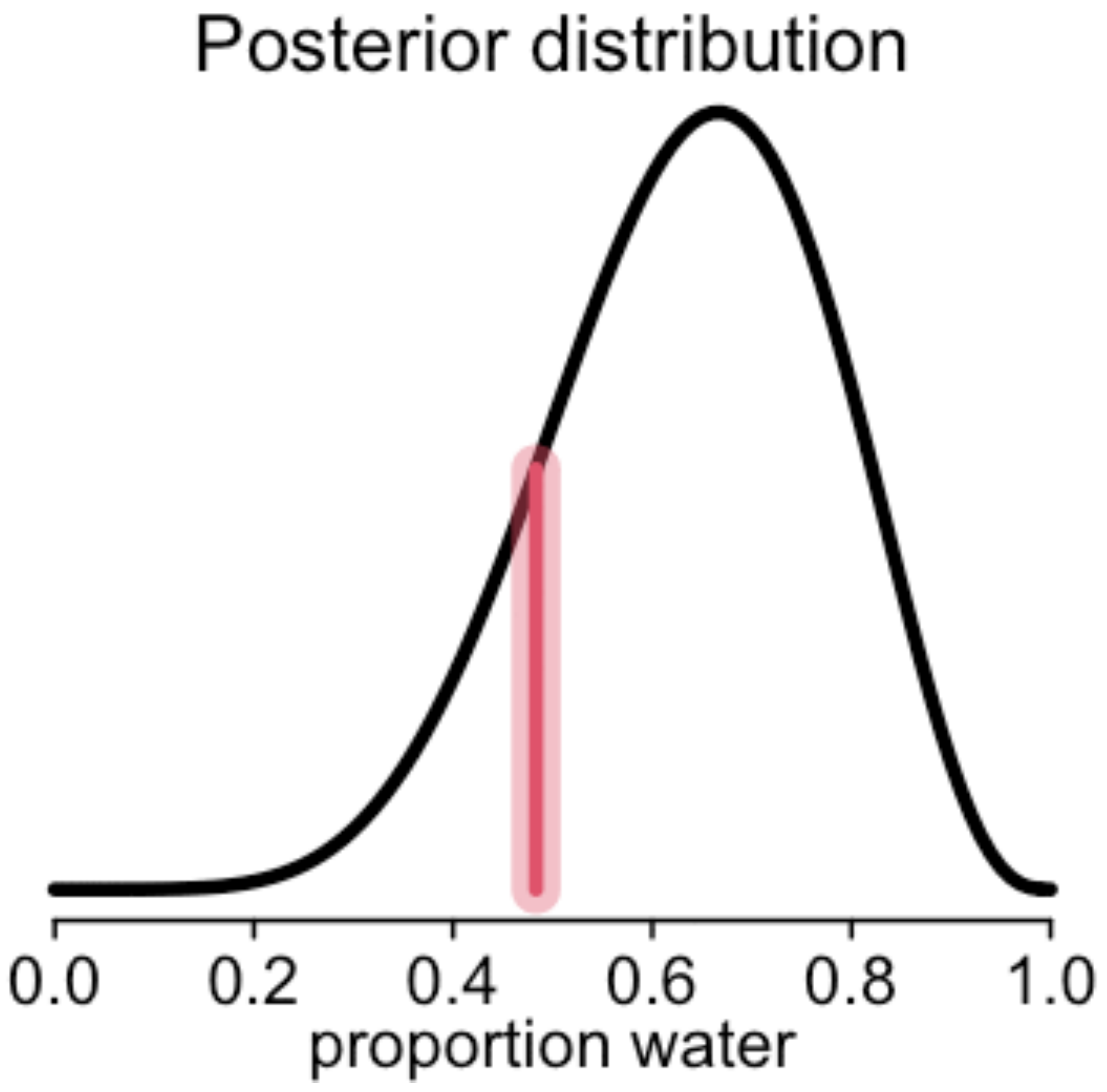
Sampling is Fun & Easy

Sample from posterior, compute desired quantity for each sample, profit

Much easier than doing integrals

Turn a **calculus problem** into a **data summary problem**

MCMC produces only samples anyway



Sampling is Handsome & Handy

Things we'll compute with sampling:

Model-based forecasts

Causal effects

Counterfactuals

Prior predictions

Bayesian data analysis

For each possible explanation of the data,

Count all the ways data can happen.

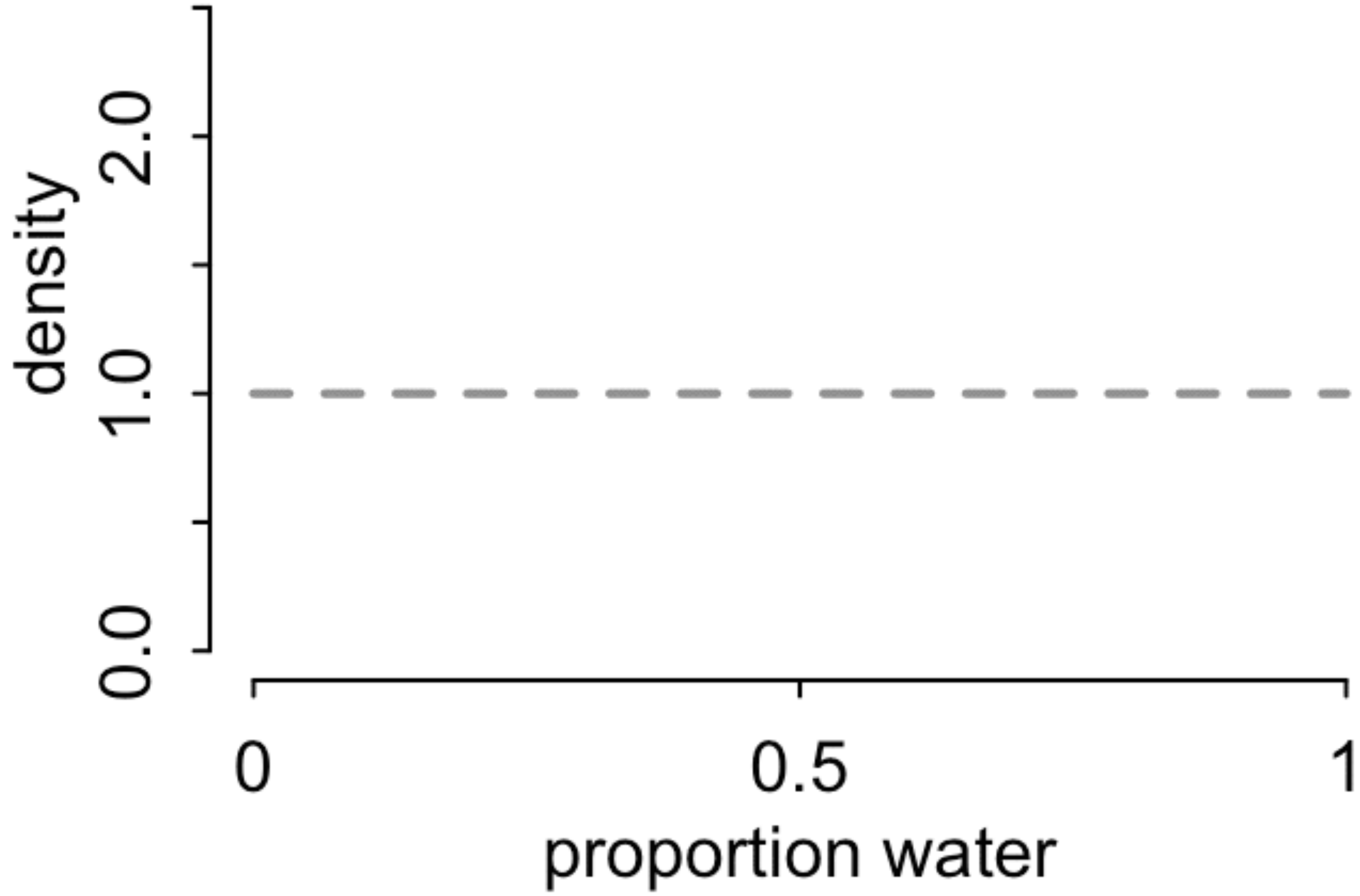
Explanations with more ways to produce the data are more plausible.

Bayesian modesty

No guarantees except ***logical***

Probability theory is a method of logically deducing ***implications of data*** *under assumptions that you must choose*

Any framework selling you more is hiding assumptions



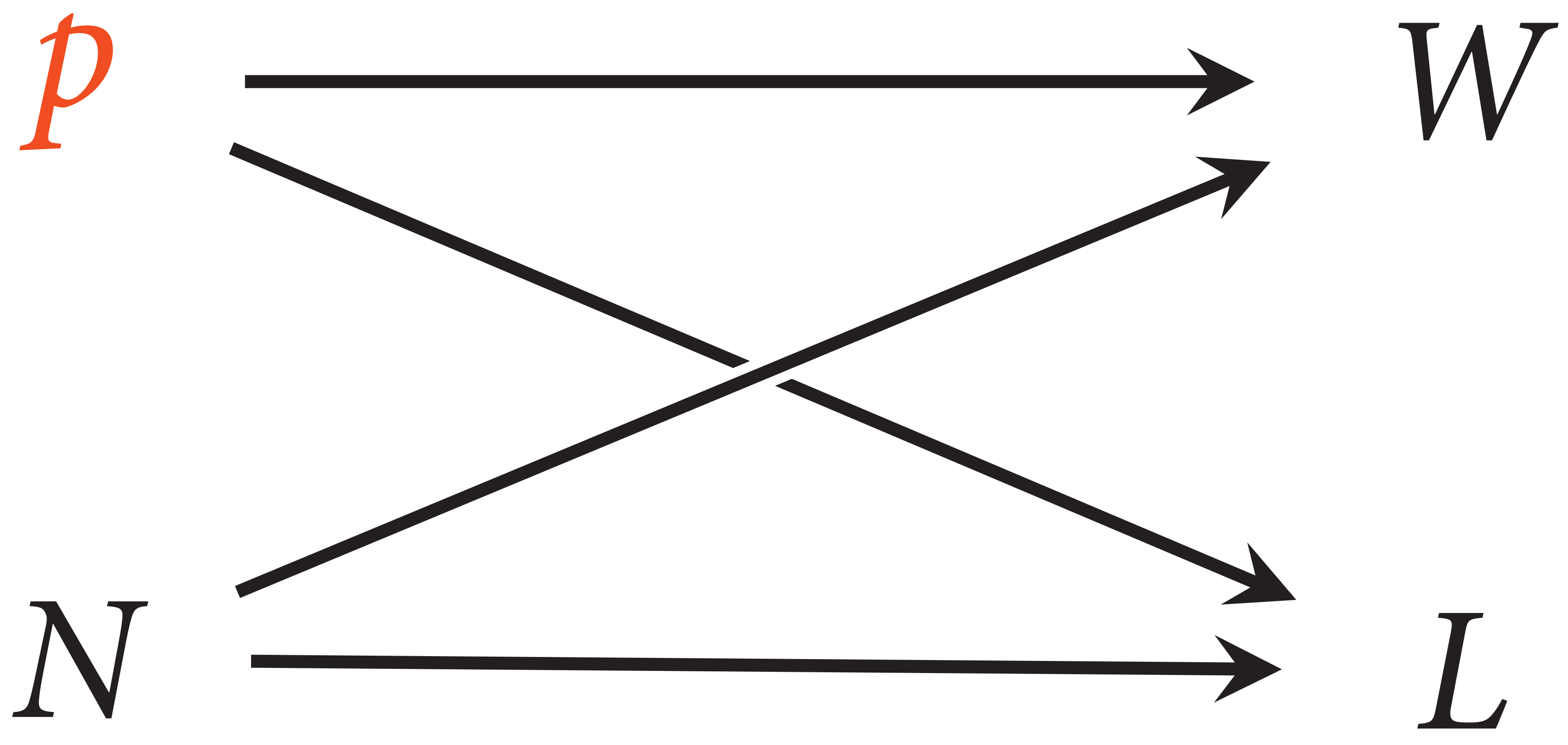
Course Schedule

Week 1	Bayesian inference	Chapters 1, 2, 3
Week 2	Linear models & Causal Inference	Chapter 4
Week 3	Causes, Confounds & Colliders	Chapters 5 & 6
Week 4	Overfitting / Interactions	Chapters 7 & 8
Week 5	MCMC & Generalized Linear Models	Chapters 9, 10, 11
Week 6	Integers & Other Monsters	Chapters 11 & 12
Week 7	Multilevel models I	Chapter 13
Week 8	Multilevel models II	Chapter 14
Week 9	Measurement & Missingness	Chapter 15
Week 10	Generalized Linear Madness	Chapter 16

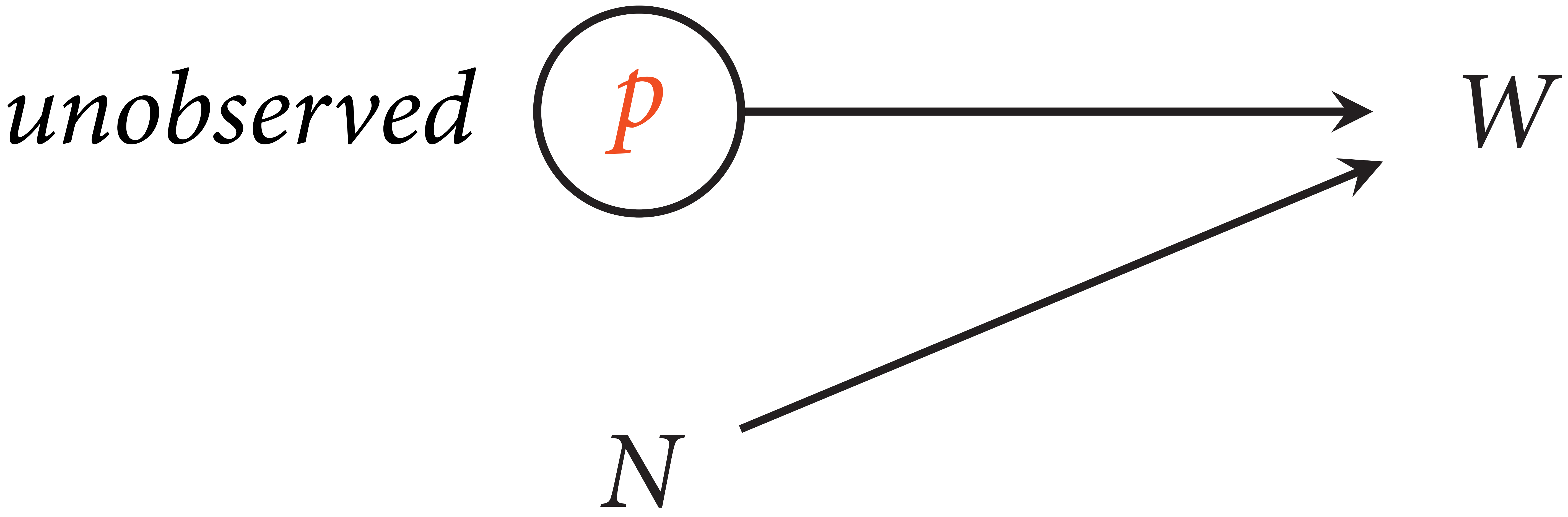
`https://github.com/rmcelreath/stat_rethinking_2023`

BONUS ROUND

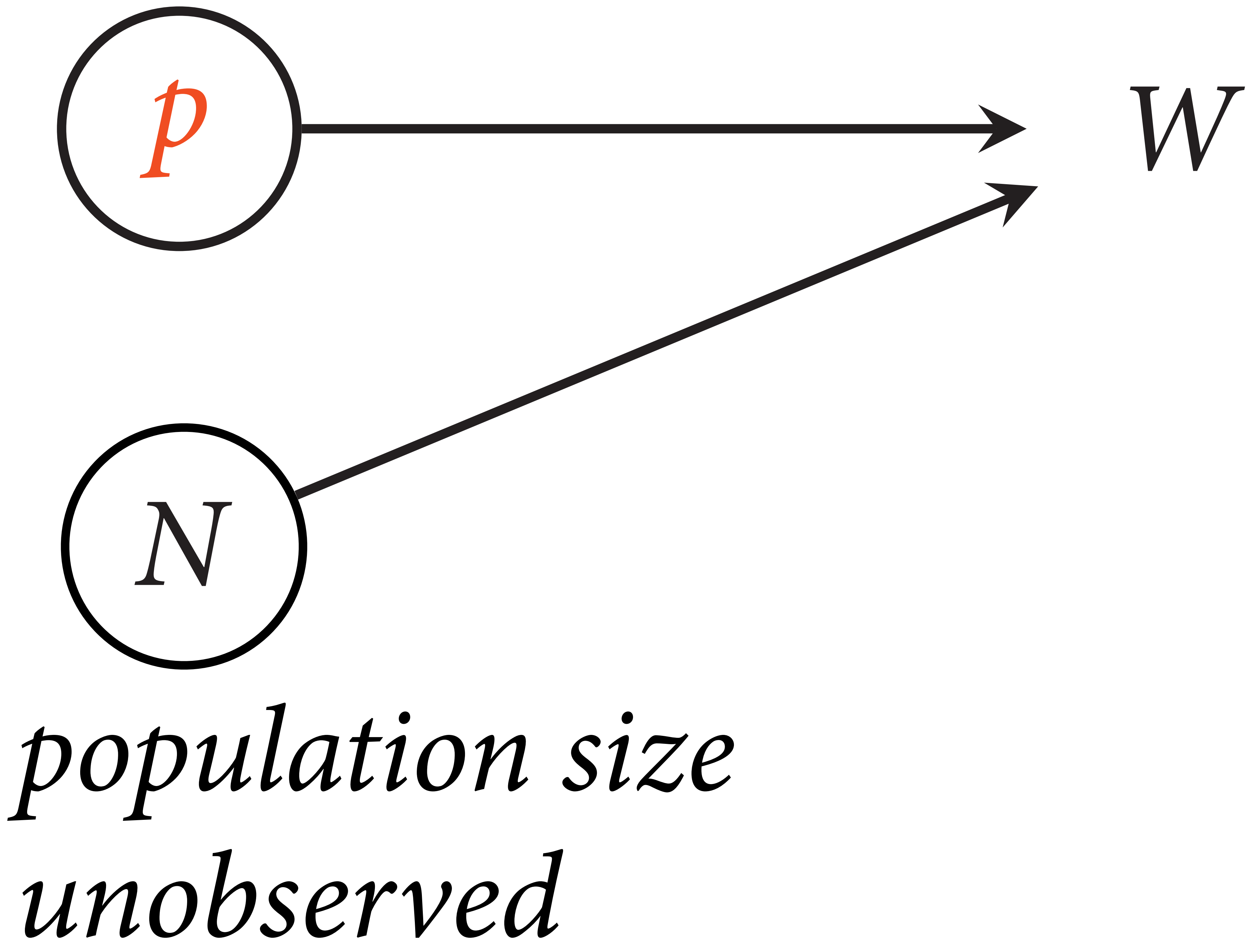
Misclassification



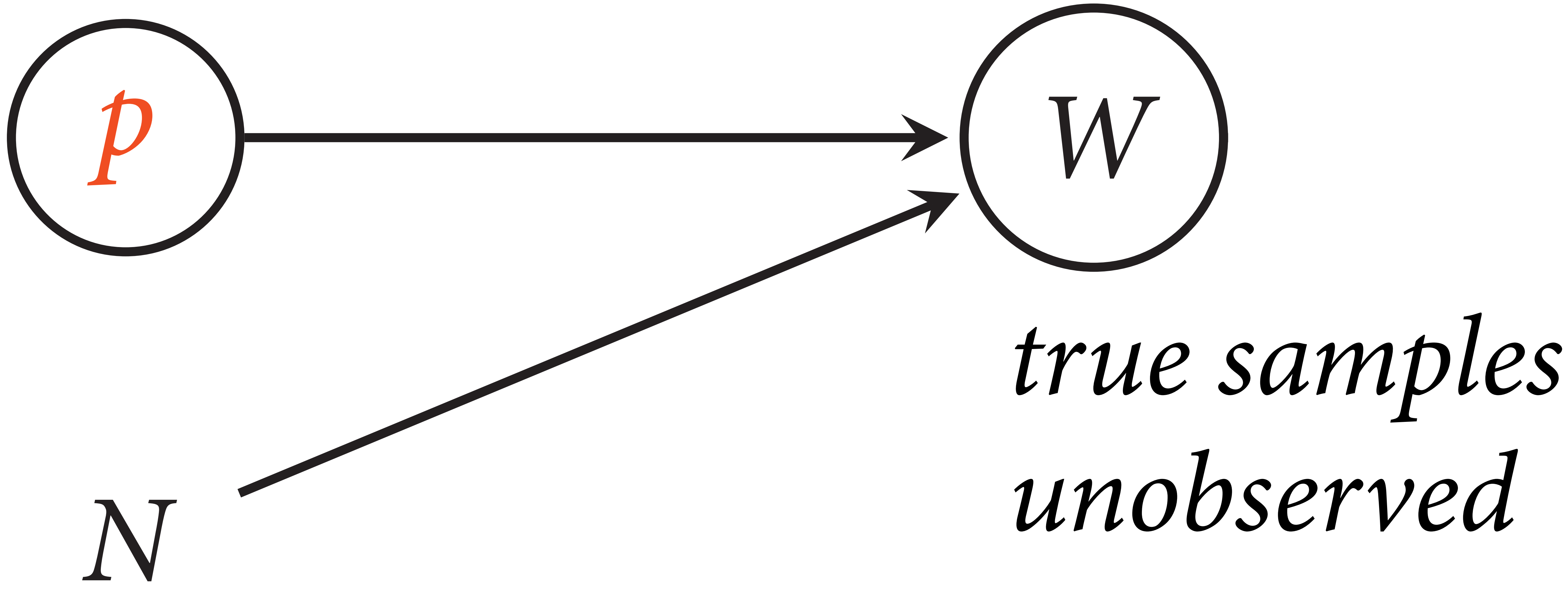
Misclassification



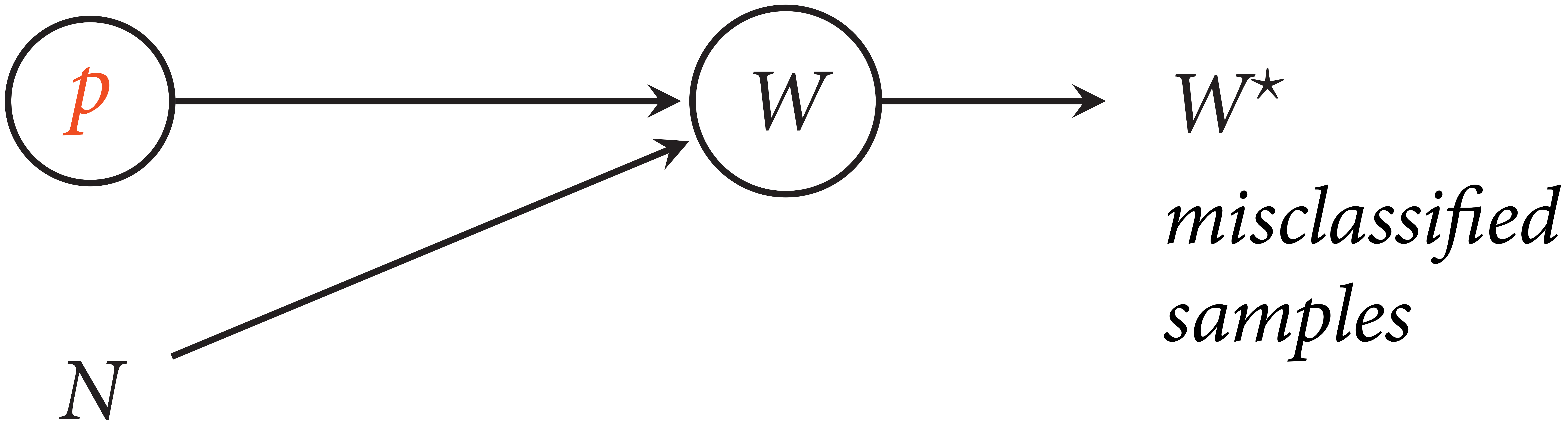
Misclassification



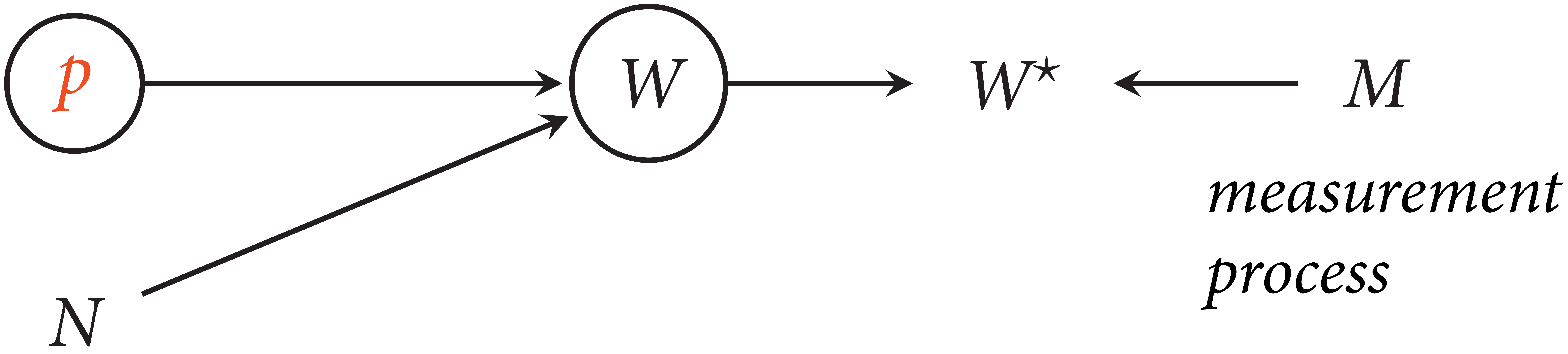
Misclassification



Misclassification



Misclassification



Misclassification simulation

Obey the workflow! Code a generative model:

```
sim_globe2 <- function( p=0.7 , N=9 , x=0.1 ) {
    true_sample <- sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
    obs_sample <- ifelse( runif(N) < x ,
        ifelse( true_sample=="W" , "L" , "W" ) , # error
        true_sample ) # no error
    return(obs_sample)
}
```

R code 2.29

Misclassification simulation

Obey the workflow! Code a generative model:

R code 2.29

```
sim_globe2 <- function( p=0.7 , N=9 , x=0.1 ) {
    true_sample <- sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
    obs_sample <- ifelse( runif(N) < x ,
        ifelse( true_sample=="W" , "L" , "W" ) , # error
        true_sample ) # no error
    return(obs_sample)
}
```

Misclassification simulation

Obey the workflow! Code a generative model:

```
sim_globe2 <- function( p=0.7 , N=9 , x=0.1 ) {
    true_sample <- sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
    obs_sample <- ifelse( runif(N) < x ,
        ifelse( true_sample=="W" , "L" , "W" ) , # error
        true_sample ) # no error
    return(obs_sample)
}
```

R code 2.29

Misclassification simulation

Obey the workflow! Code a generative model:

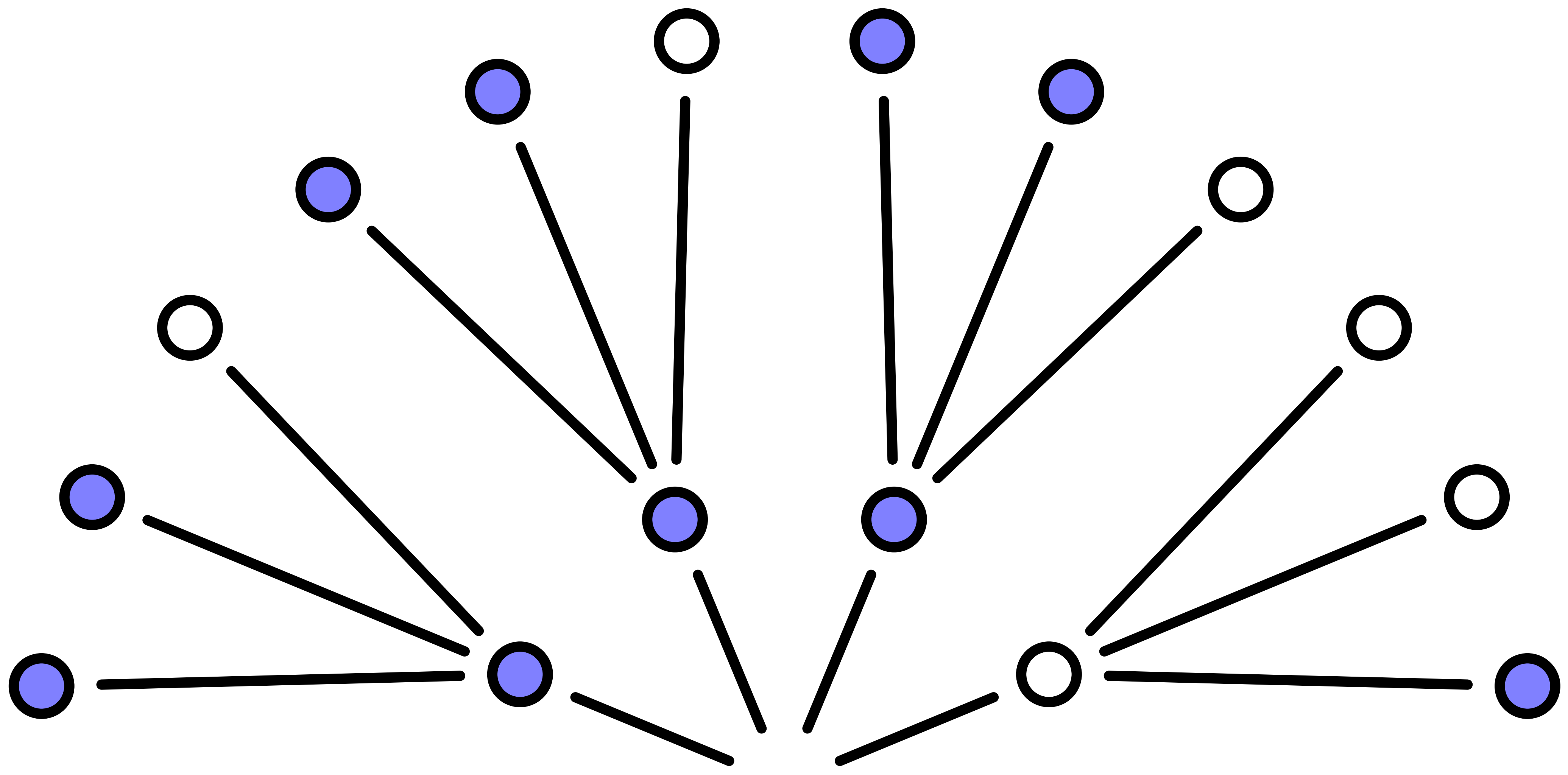
```
sim_globe2 <- function( p=0.7 , N=9 , x=0.1 ) {
    true_sample <- sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
    obs_sample <- ifelse( runif(N) < x ,
        ifelse( true_sample=="W" , "L" , "W" ) , # error
        true_sample ) # no error
    return(obs_sample)
}
```

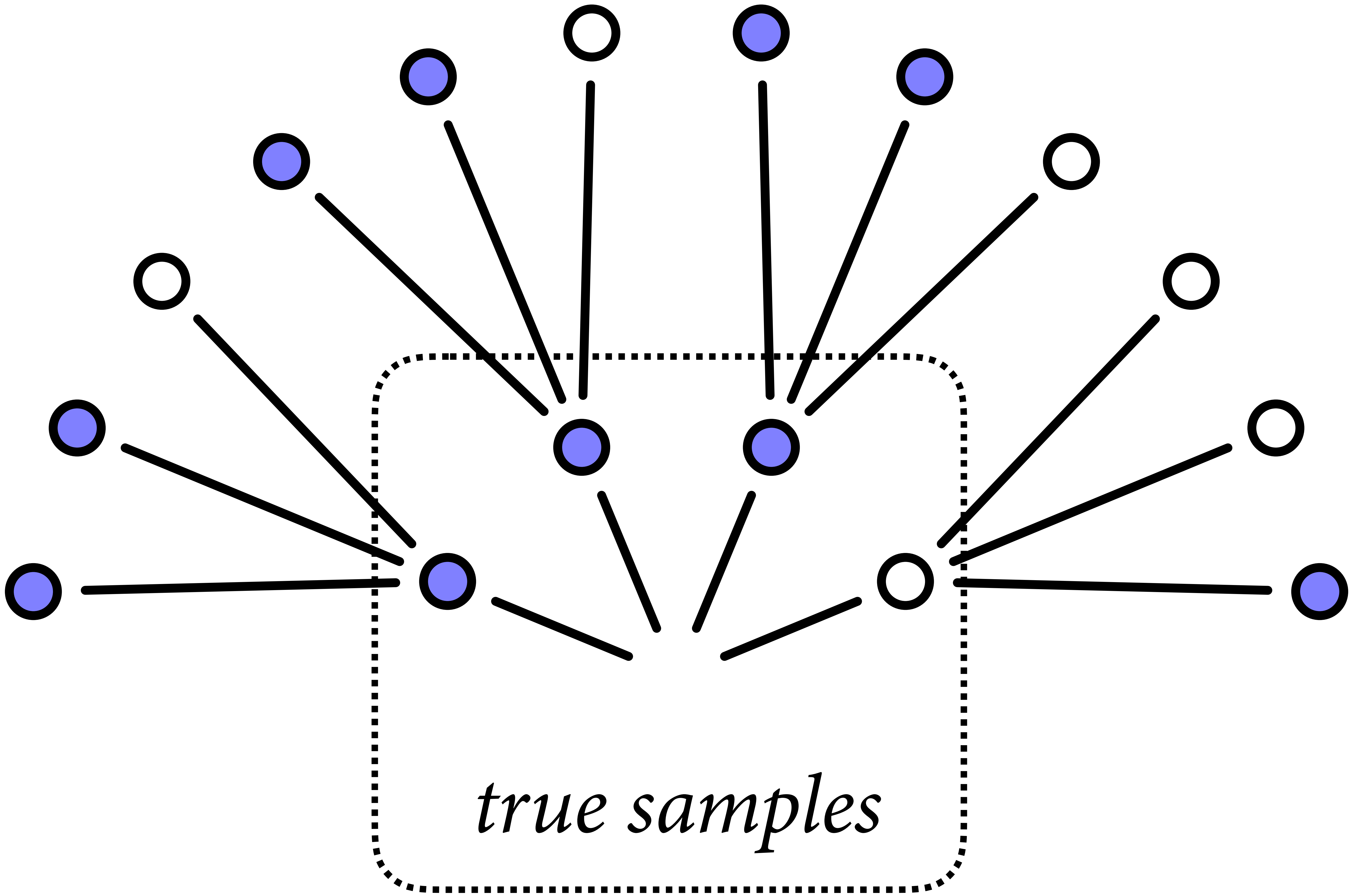
R code 2.29

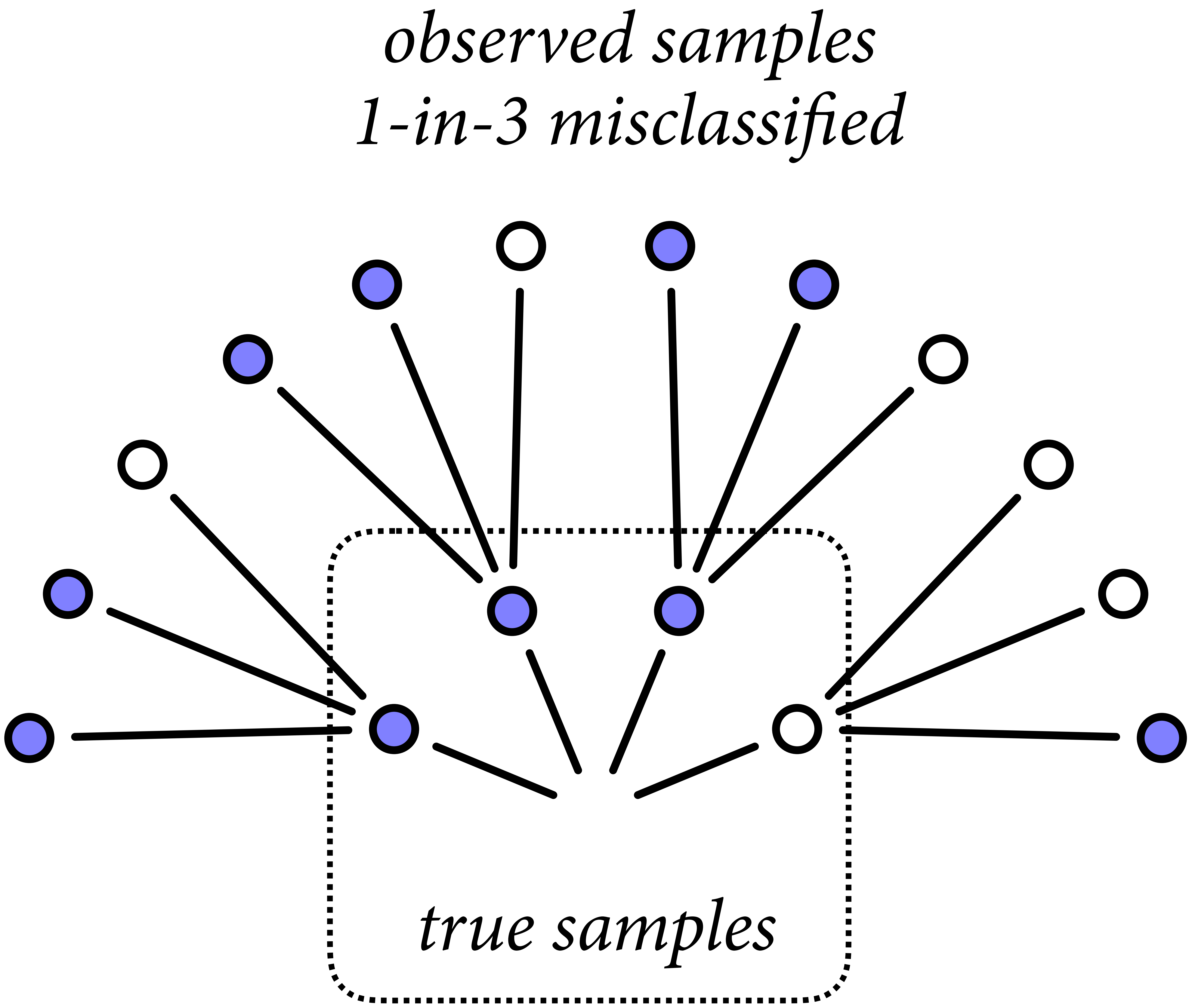
Misclassification estimator

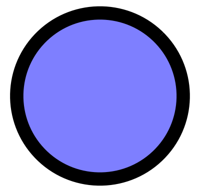
Use the intuition from the generative model to draw out the Garden of Forking Data, build a Bayesian estimator.

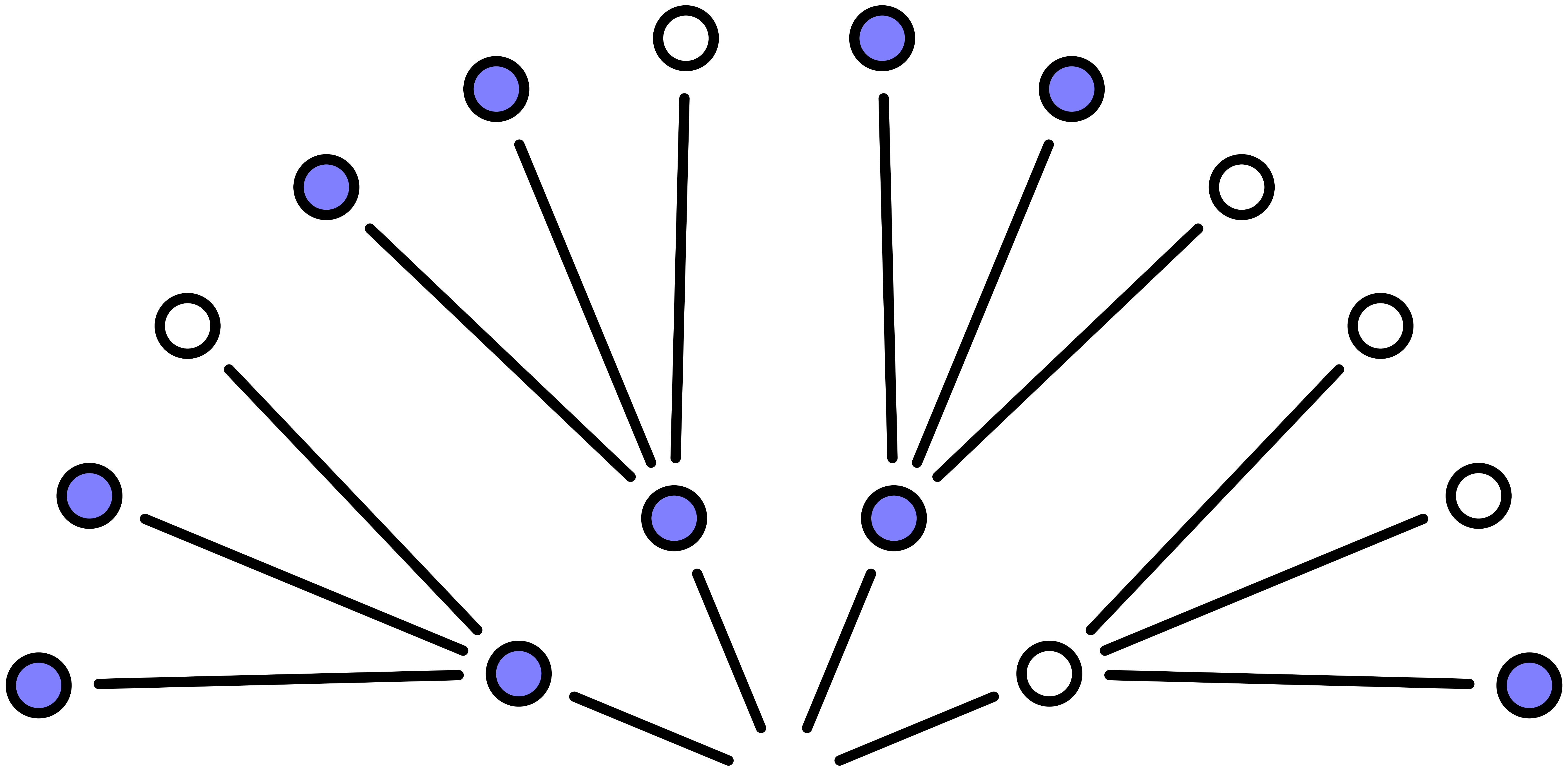
Two stages: (1) true samples, (2) misclassification



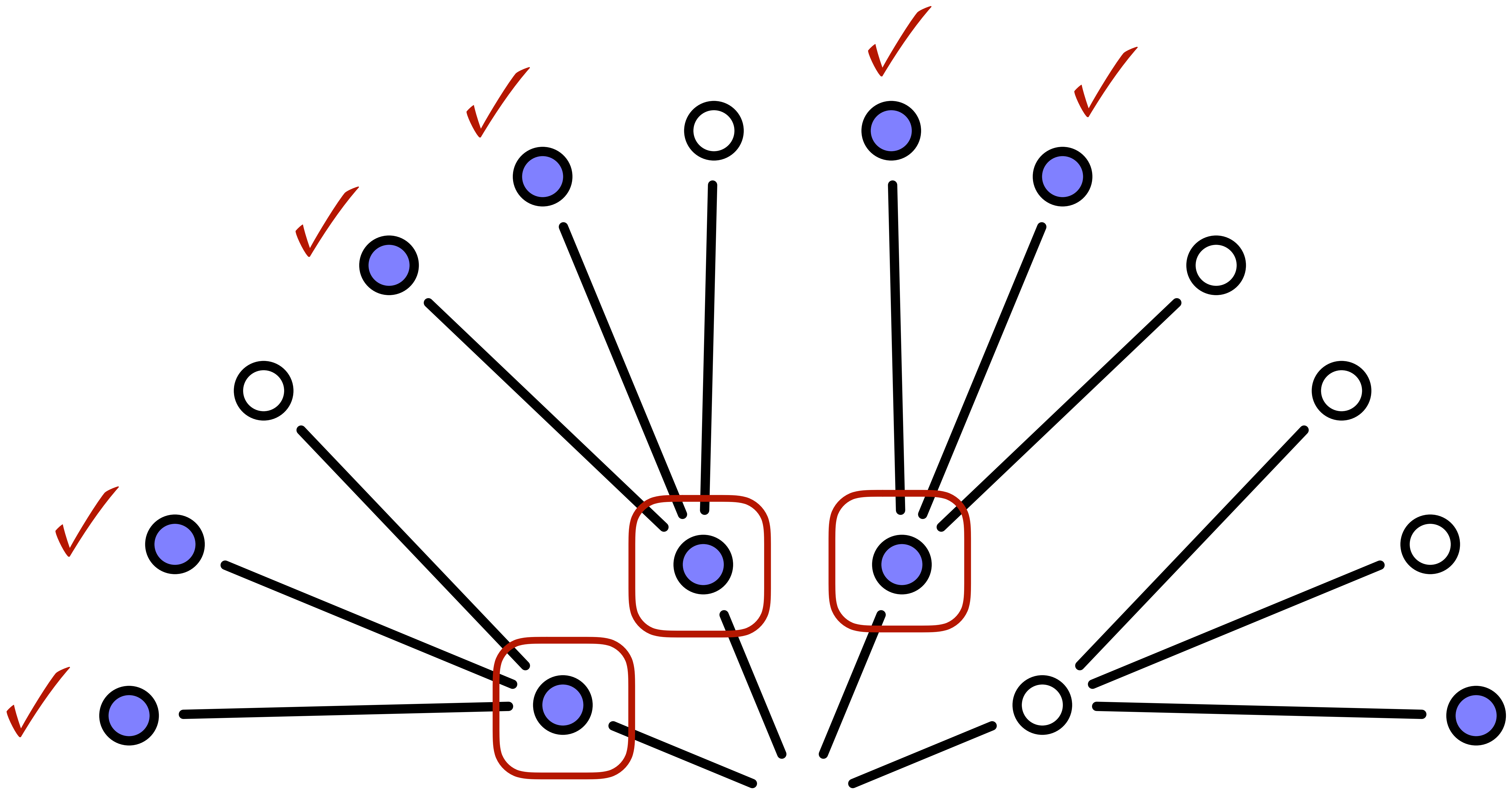

true samples


observed samples
1-in-3 misclassified
true samples

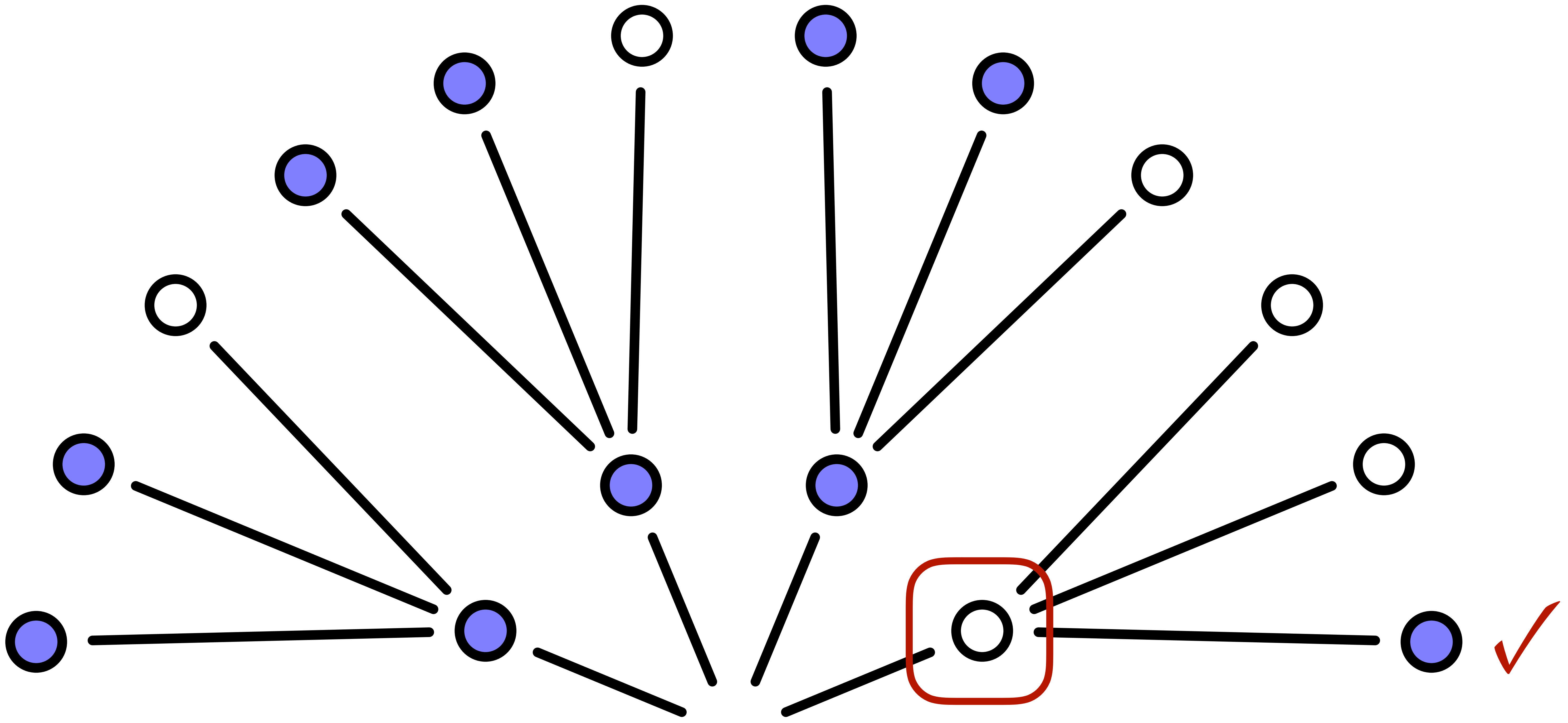
Observe  — How many ways can this happen?



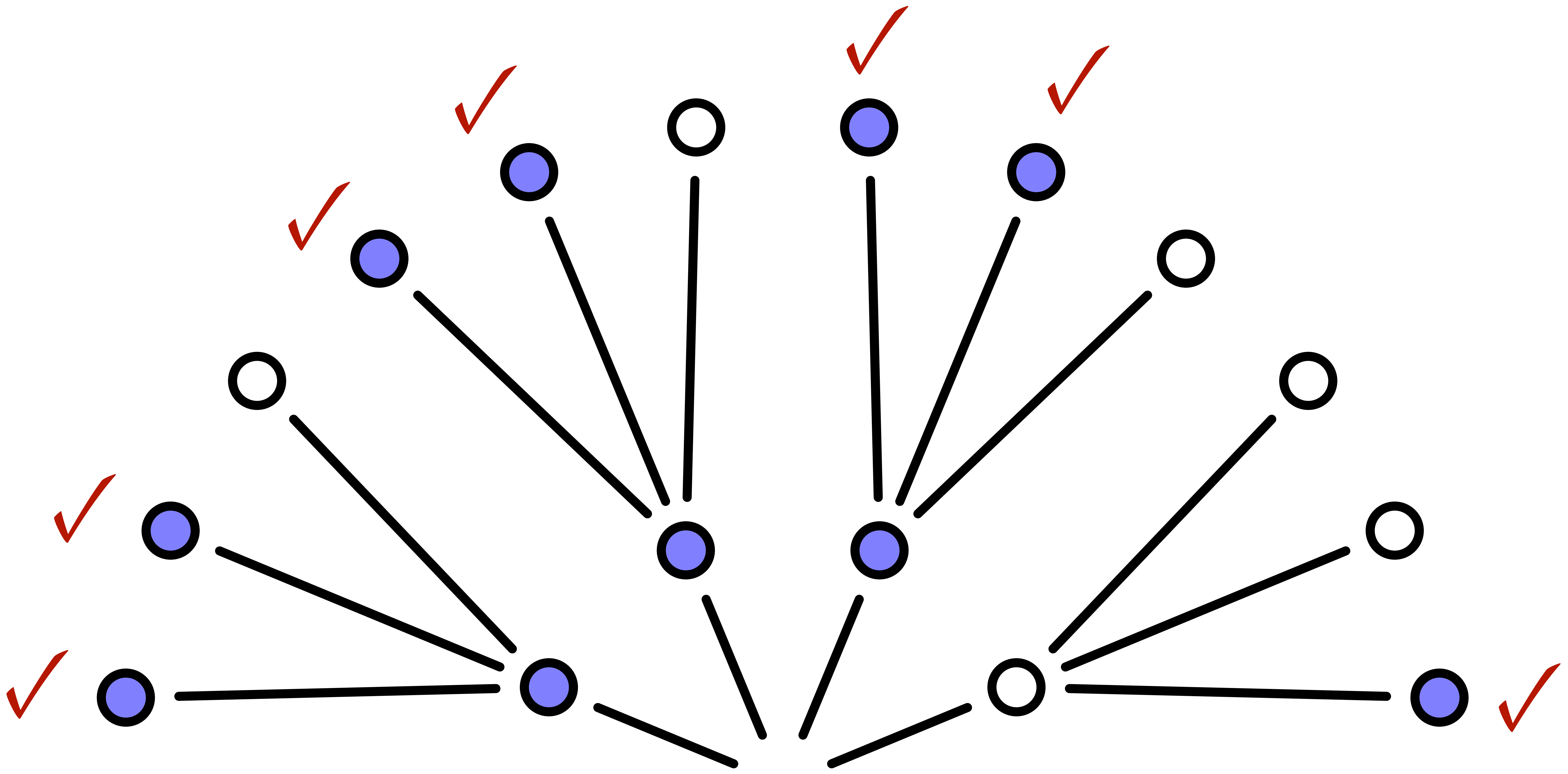
6 ways to observe water, when true sample is water



1 way to observe water, when true sample is land



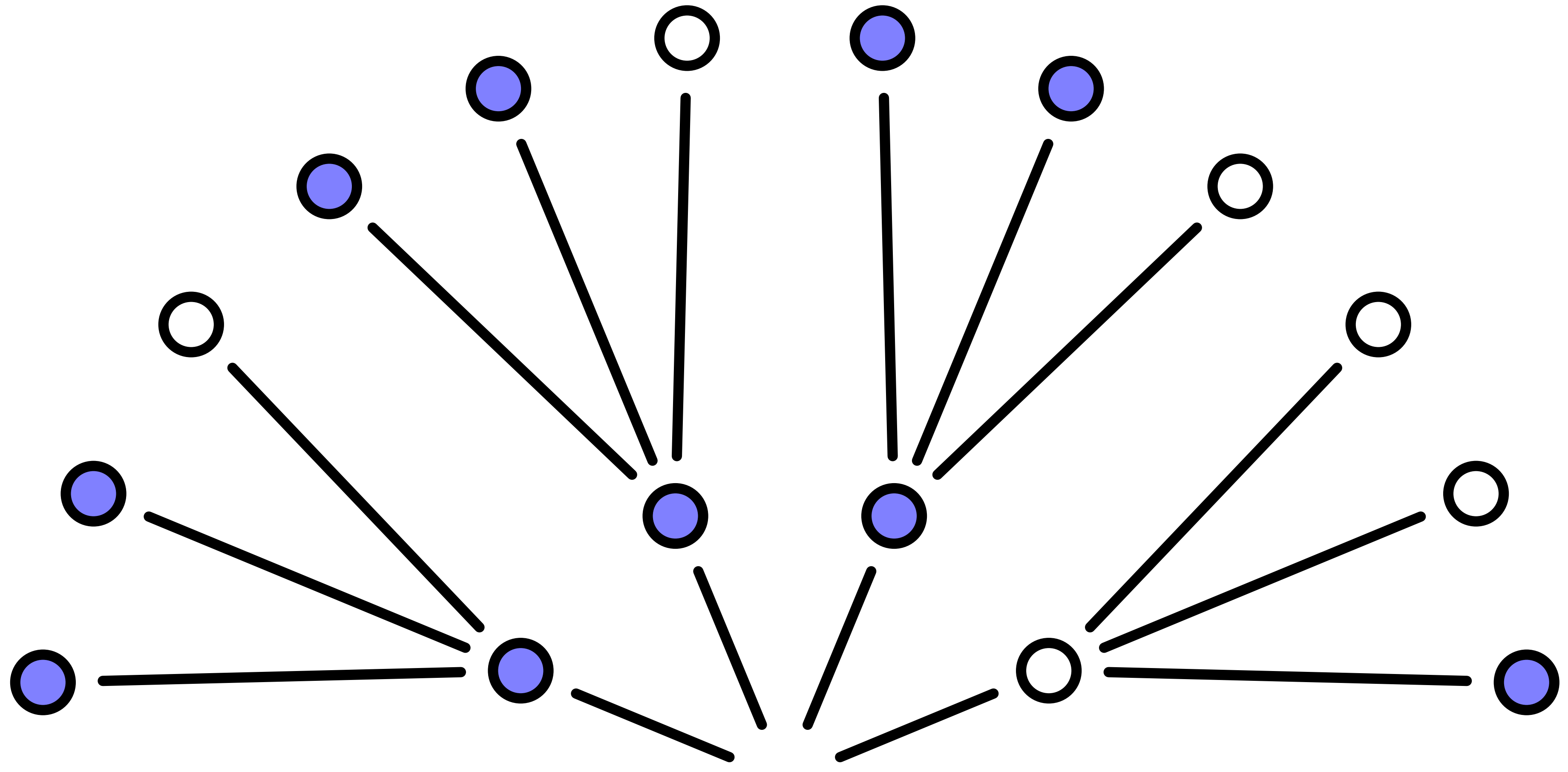
$3{\times}2 + 1{\times}1 = 7$ *ways to observe water*



Misclassification estimator

$$\Pr(\text{water}|p,x) = p(1-x) + (1-p)x$$

$$\Pr(\text{land}|p,x) = (1-p)(1-x) + px$$



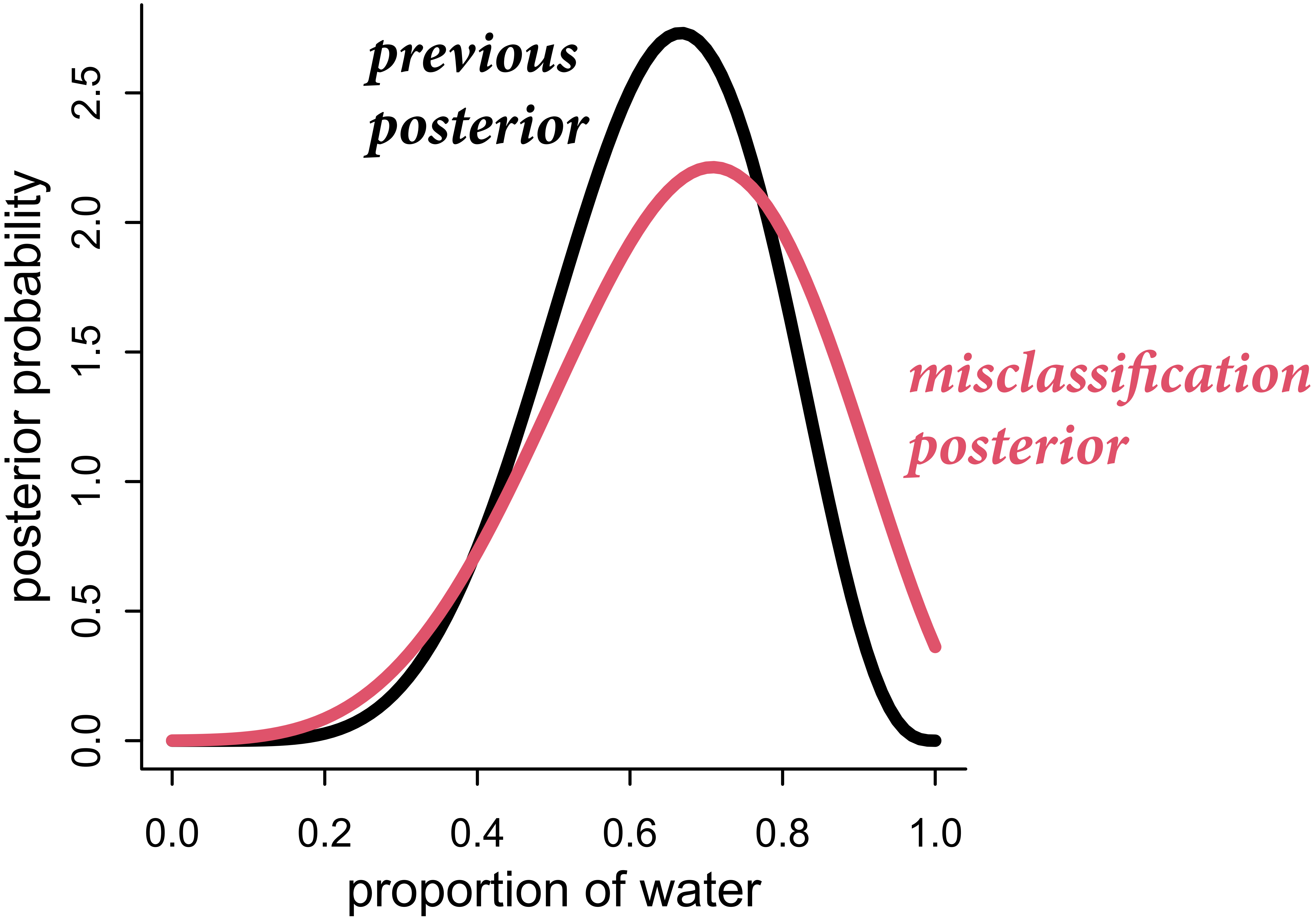
Posterior distribution for p given W,L,x:

$$\Pr(p|W,L,x) = \frac{[p(1-x) + (1-p)x]^W \times [(1-p)(1-x) + px]^L}{Z}$$

$$\Pr(p|W,L,x) = \frac{\overbrace{[p(1-x)+(1-p)x]}^{\textit{probability of each water}}{}^{W} \times \overbrace{[(1-p)(1-x)+px]}^{\textit{probability of each land}}{}^{L}}{Z}$$

$$\Pr(p|W,L,x) = \frac{\overbrace{[p(1-x)+(1-p)x]}^{\textit{probability of each water}}{}^{W} \times \overbrace{[(1-p)(1-x)+px]}^{\textit{probability of each land}}{}^{L}}{\underbrace{Z}_{\substack{\textit{some unpleasant}\\ \textit{normalizing constant}}}}$$

Misclassification posterior



Measurement matters

When there is measurement error, better to model it than to ignore it

Same goes for: missing data, compliance, inclusion, etc

Good news: Samples do not need to be *representative* of population in order to provide good estimates of population

What matters is *why* the sample differs