

# Today

- Questions from homework & recap
  - pair programming (Q1: logistic growth optim)
- Frequentist inference algorithms
  - sampling distribution algorithm
  - p-value algorithm
  - coverage algorithm
- Confidence vs Prediction intervals

# Different inference problems

## Estimation

Infer a property of a population (e.g. mean) from a sample

## Model selection

Infer the data generating process from among a set of candidate data-generating processes

## Hypothesis test (association)

Infer that  $y$  is associated with  $x$

## Causation

Infer that  $x$  causes  $y$

Infer the size of an effect due to an experimental intervention (estimation)

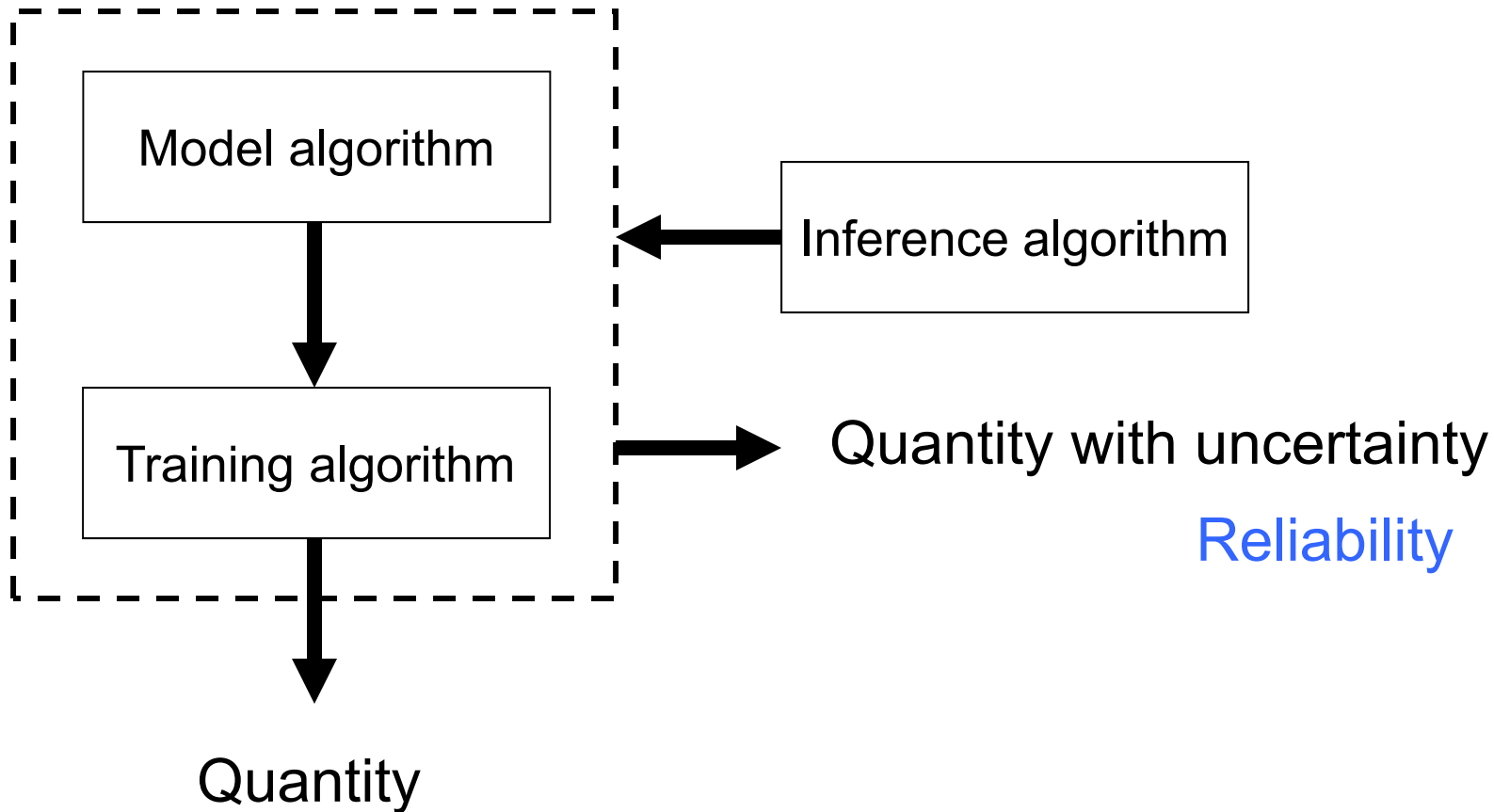
Infer that an experimental intervention had an effect (H-test)

## Prediction

Predict the value of a new observation or population state (extrapolation or interpolation)

Predict the population state in the future (forecast/extrapolation)

# Algorithms in data science



"Dumb" - doesn't say about reliability

# Algorithms in data science

- Inference algorithm
  - looking back: considering all the ways data could have happened
    - frequentist (sampling distribution)
    - likelihood (probability accounting)
    - Bayesian (likelihood + belief updating)
  - looking forward: predicting new data and testing against them
    - cross validation, AIC, machine learning

# Frequentist probability

- Long-term frequency
  - e.g. tossing a coin
  - $P(\text{heads}) = \lim_{n \rightarrow \infty} \text{heads} / n$

# Sample vs population statistic

- Population statistic
  - e.g. mean weight
  - there is a true value
  - “fixed” not random
- Sample statistic
  - e.g. mean of sample
  - random variable

# Sampling distribution

- Frequentist notion of looking back: considering all the ways data could have happened
- Imaginary repeated sampling from the data generating process

# Sampling distribution algorithm

- Data generating process repeated many times, each time calc sample statistic

repeat very many times

sample  $n$  units from the population

calculate the sample statistic

plot sampling distribution (histogram) of the sample statistic



# Make the algorithm

How much does  
this species  
weigh?



repeat very many times

sample  $n$  units from the population

calculate the sample statistic

plot sampling distribution (histogram) of the sample statistic

# Sampling distribution

132 orange-spotted warblers. 1 indicates infected

pathogen <-

```
c(1,1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,0,1,0,0,0,0,  
  1,1,1,0,1,1,0,1,1,0,0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,1,1,  
  0,1,0,0,0,1,0,0,0,1,0,0,1,0,0,1,1,0,1,1,0,1,1,0,0,0,0,0,0,0,0,1,0,1,1,1,0,  
  1,0,1,0,0,0,0,0,0,1,1,0,0,0,1,1,1,1,0,0,1)
```

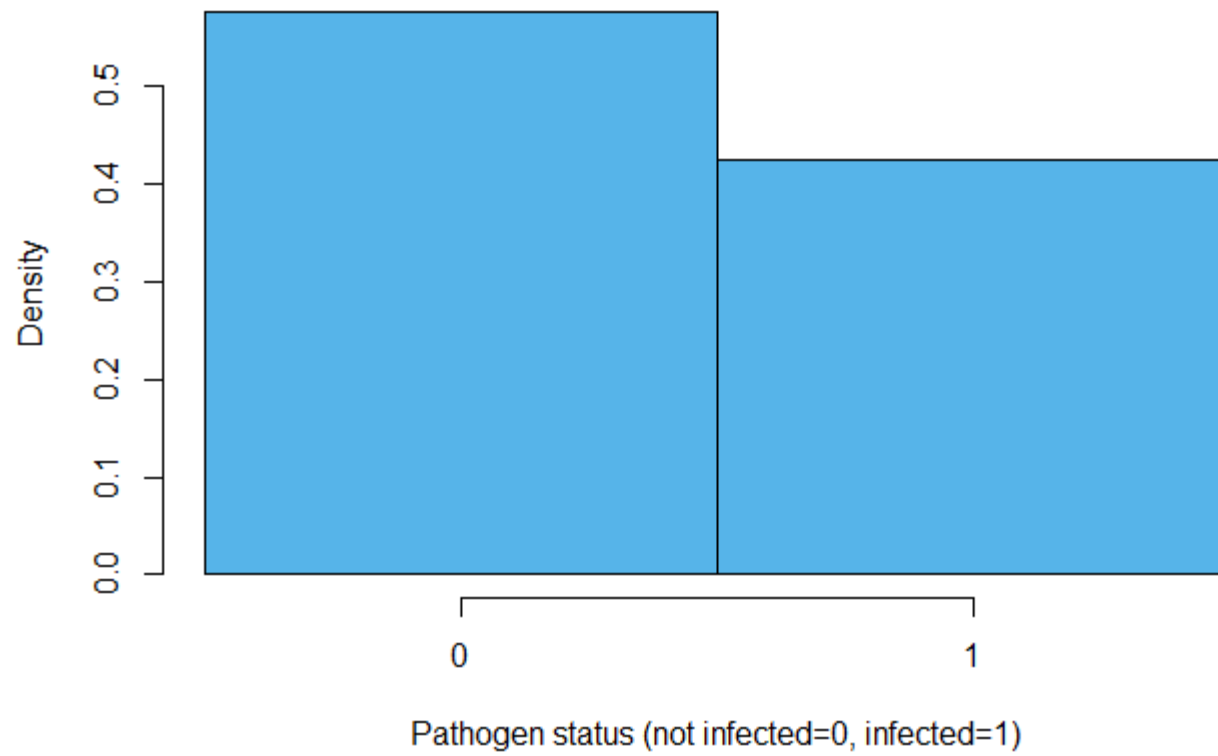
Take a sample:

```
sample(pathogen,10)
```

```
0 1 0 0 0 0 0 1 0 0
```

```
pathogen prevalence = 0.2
```

Our scientific observation



True prevalence is 0.424

# Sampling distribution algorithm 1

for each possible combination of  $n$  sample units  
    sample  $n$  units from the population  
    calculate the sample statistic  
plot sampling distribution (histogram) of the sample statistic

for pathogen prevalence

There are  $3e14$  possible samples.

Too hard! It would take 100 years to compute!

# Sampling distribution algorithm 2

Invoke the law of large numbers

repeat very many times

- sample  $n$  units from the population

- calculate the sample statistic

- plot sampling distribution (histogram) of the sample statistic

## for pathogen prevalence

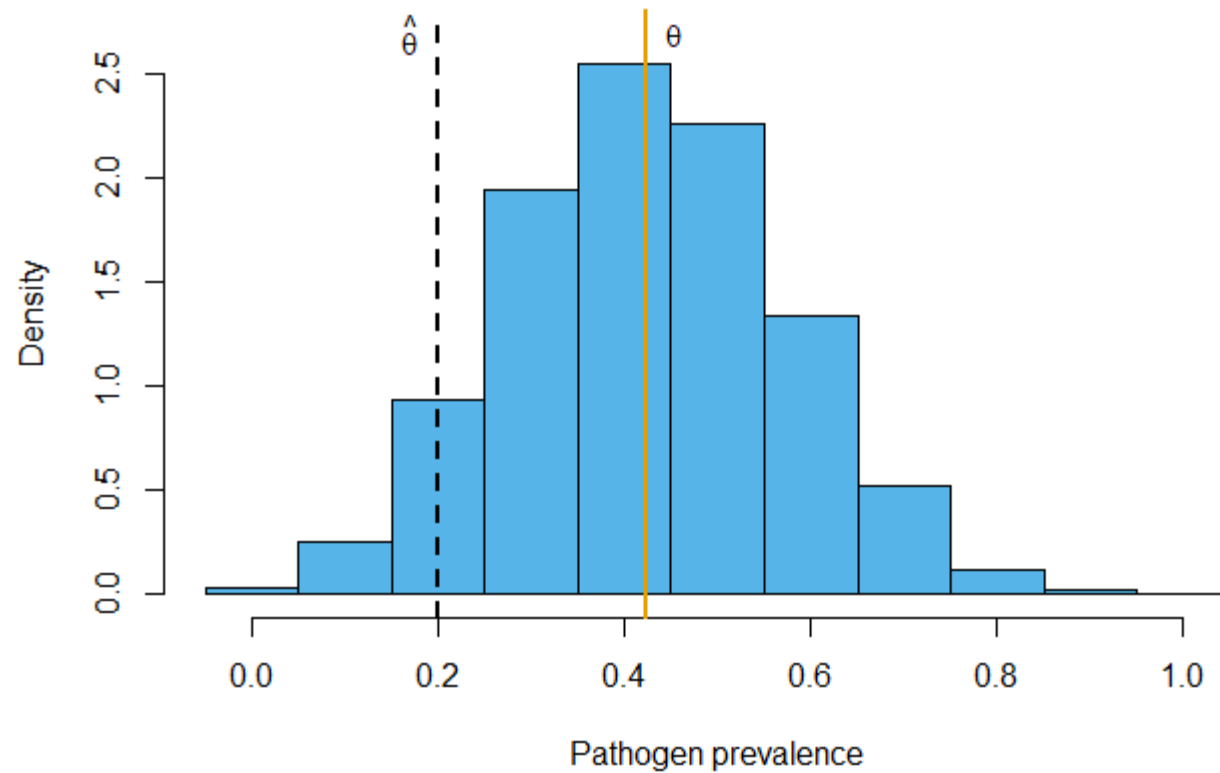
for a large number of repeated samples

- randomly sample 10 birds from the population

- calculate the prevalence in the sample

- plot sampling distribution (histogram) of prevalence

# The sampling distribution for prevalence

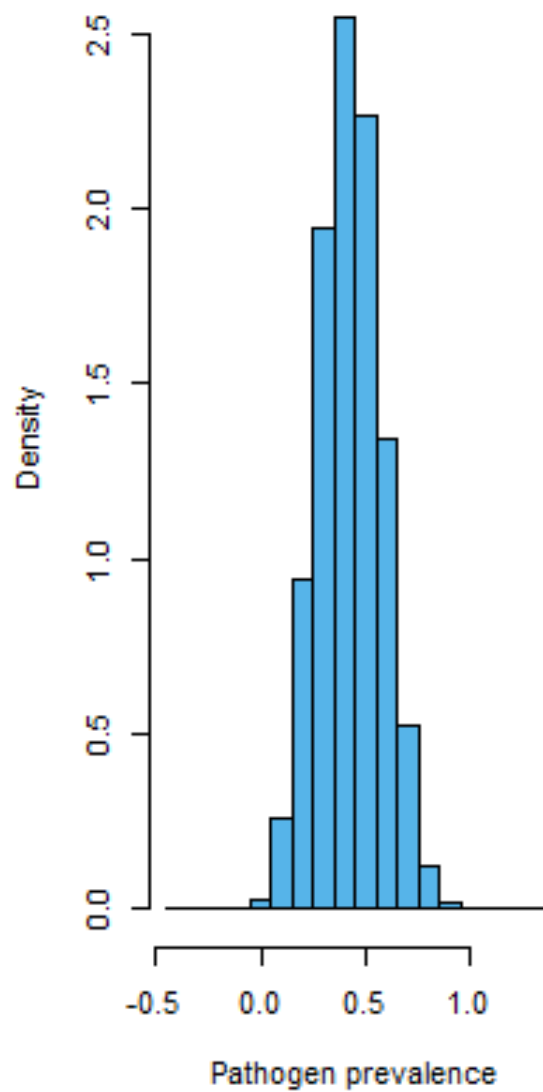


# Confidence interval

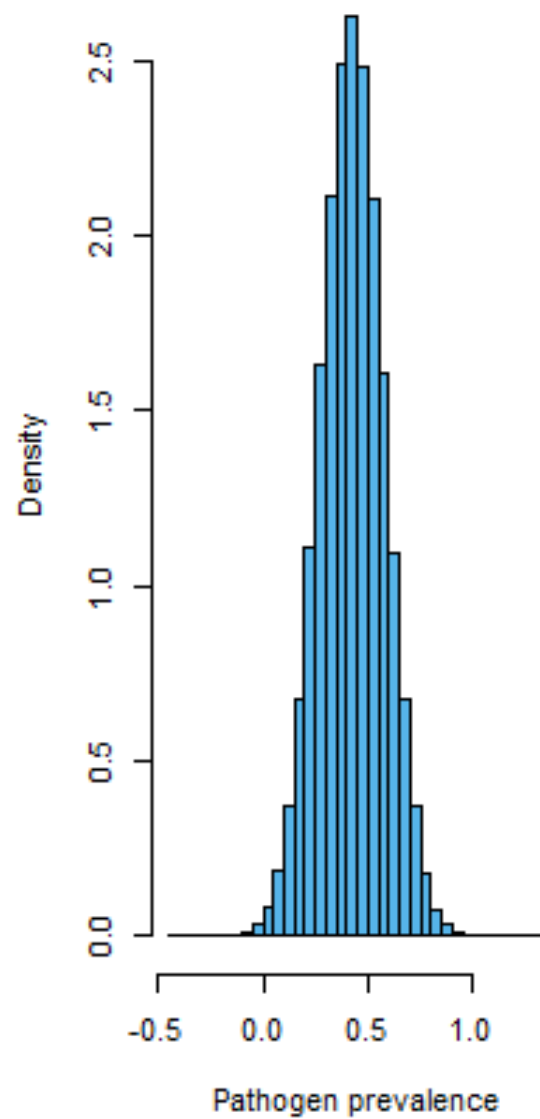
- An interval calculated by some procedure that would **contain** (or **cover**) the true population value 95% of the time, **if sampling and calculating an interval were repeated a very large number of times**

Confidence = **reliability** of the **procedure**

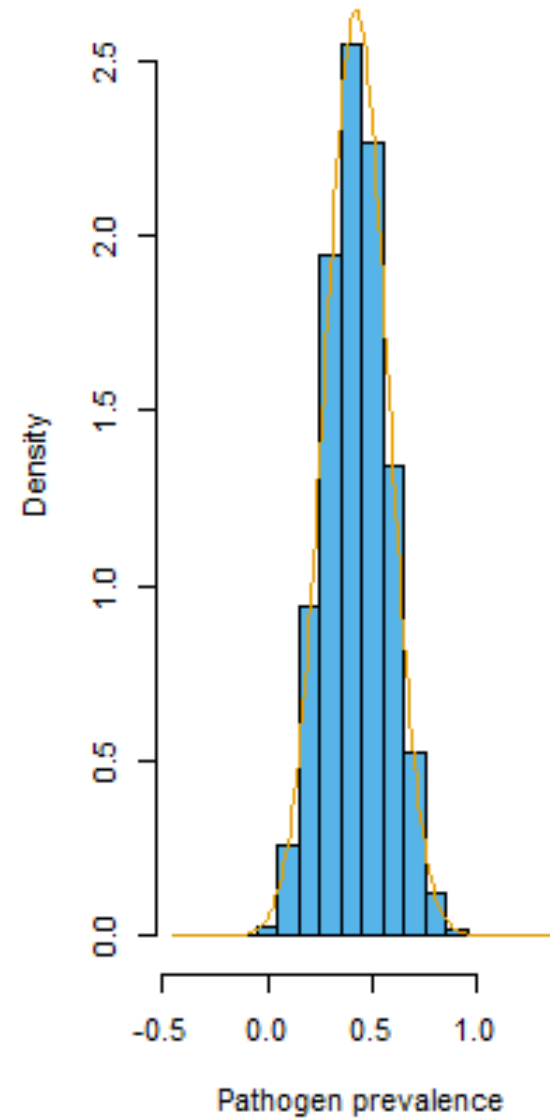
True sampling distribution



Approximating Normal

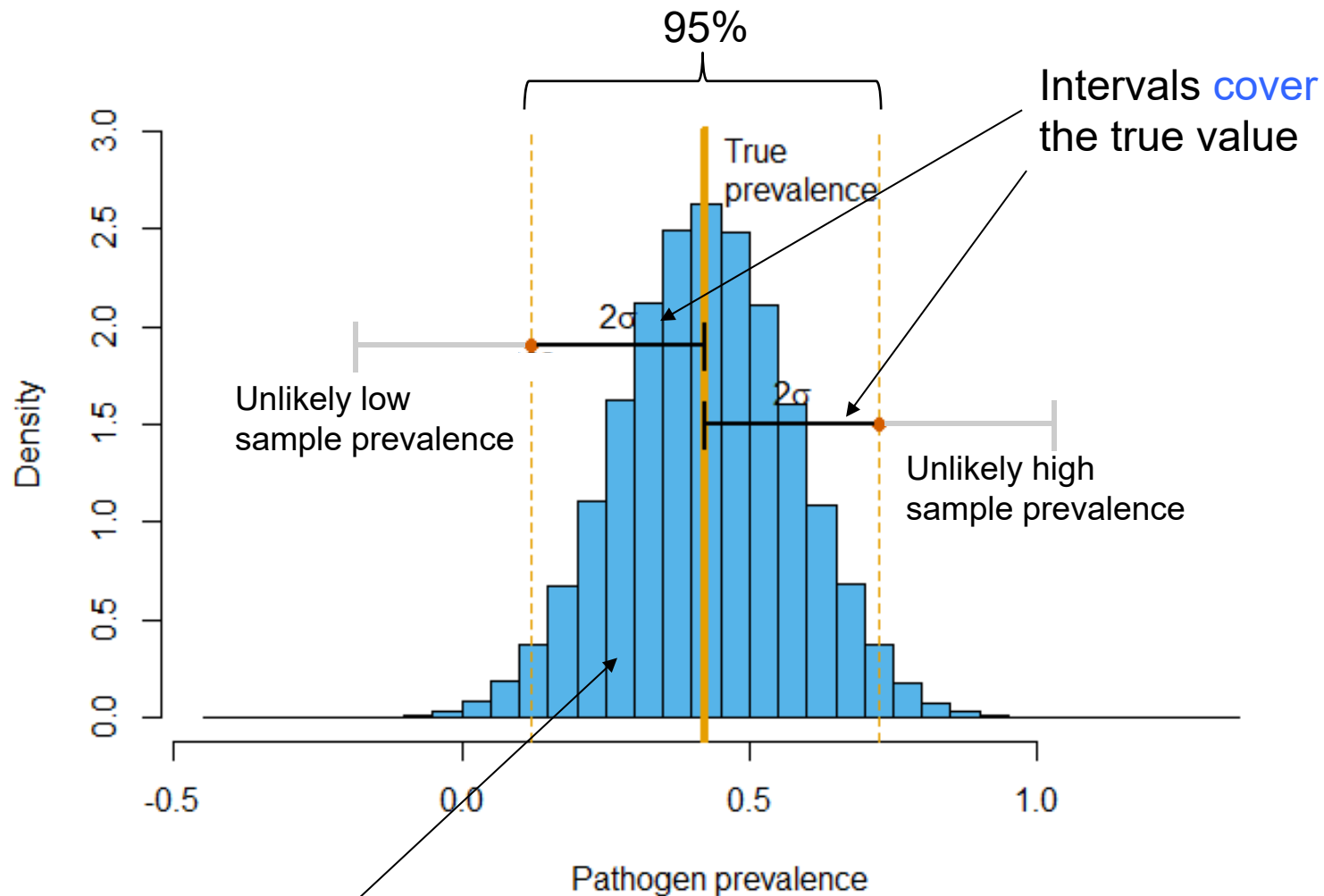


Normal overlaid on true





# Construct an interval to cover true value



Normal distribution approximating the true sampling distribution

# Plug in principle

- We don't know the true sampling distribution or its parameters
- Plug in the sample instead as an estimate
  - in this example we can use the standard error of the sample as an estimate of the standard deviation of the sampling distribution

# Coverage

repeat very many times

- sample  $n$  units from the population

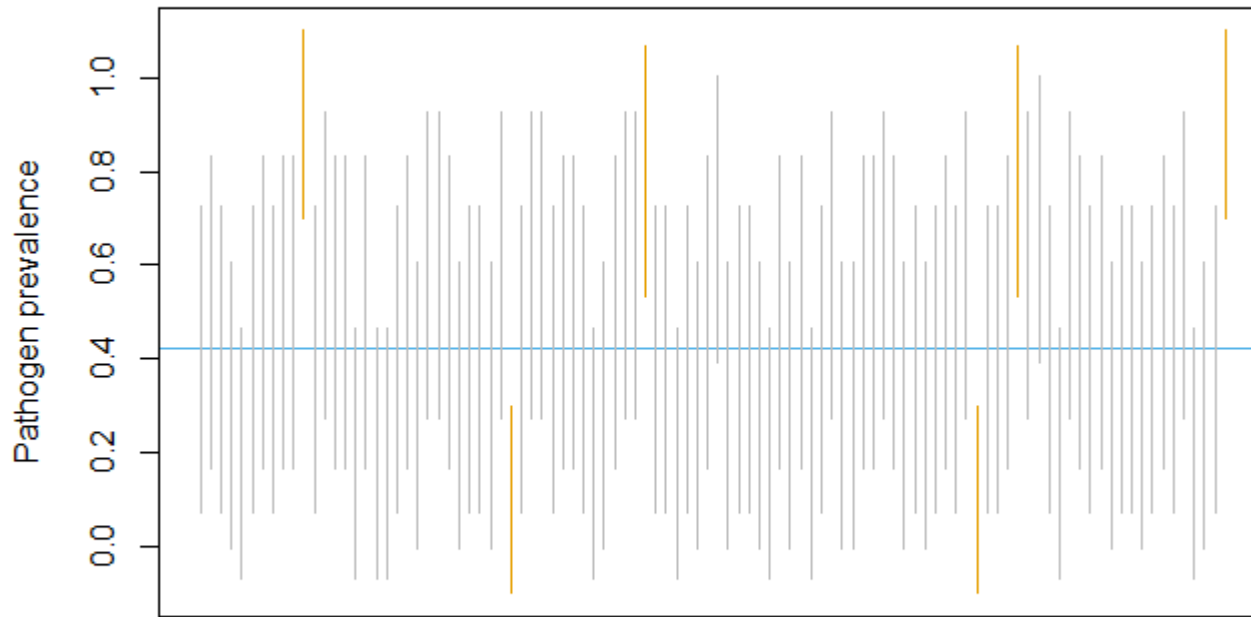
- calculate the sample statistic

- calculate the interval for the sample statistic

- calculate frequency true value is in the interval

**Calibrates** the degree of confidence in the procedure

# First 100 95% confidence intervals



95.6% of the intervals cover the true value  
In first 100, 6 do not cover the true value  
(we expect about 5/100)