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

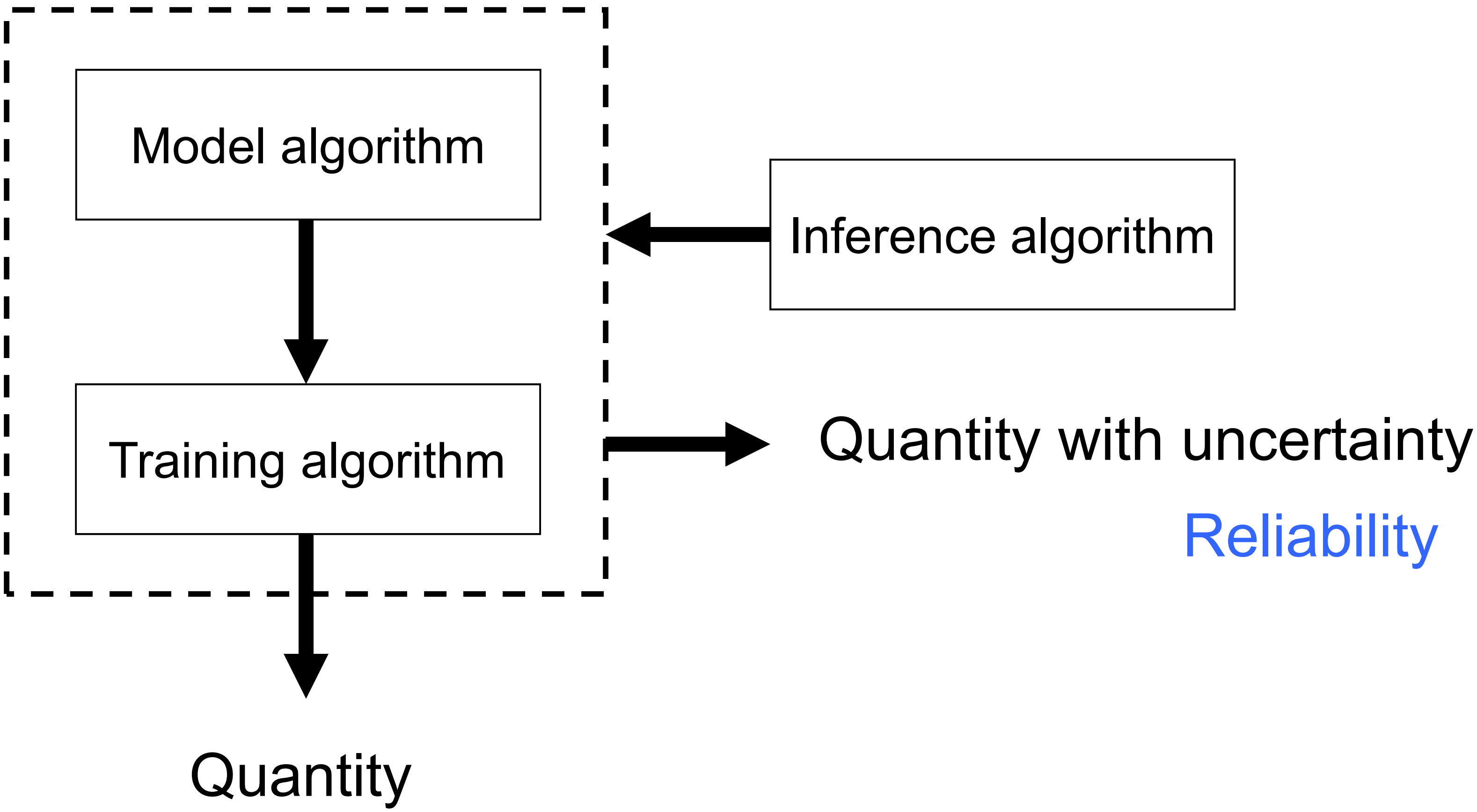
Model selection
Infer the data generating process from among a set of candidate data-generating processes

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(heads) = lim n -> Inf, 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(1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,0,1,0,0,0,0,
      1,1,1,0,1,1,0,1,1,0,0,1,1,0,0,1,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,1,1,
      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)
```

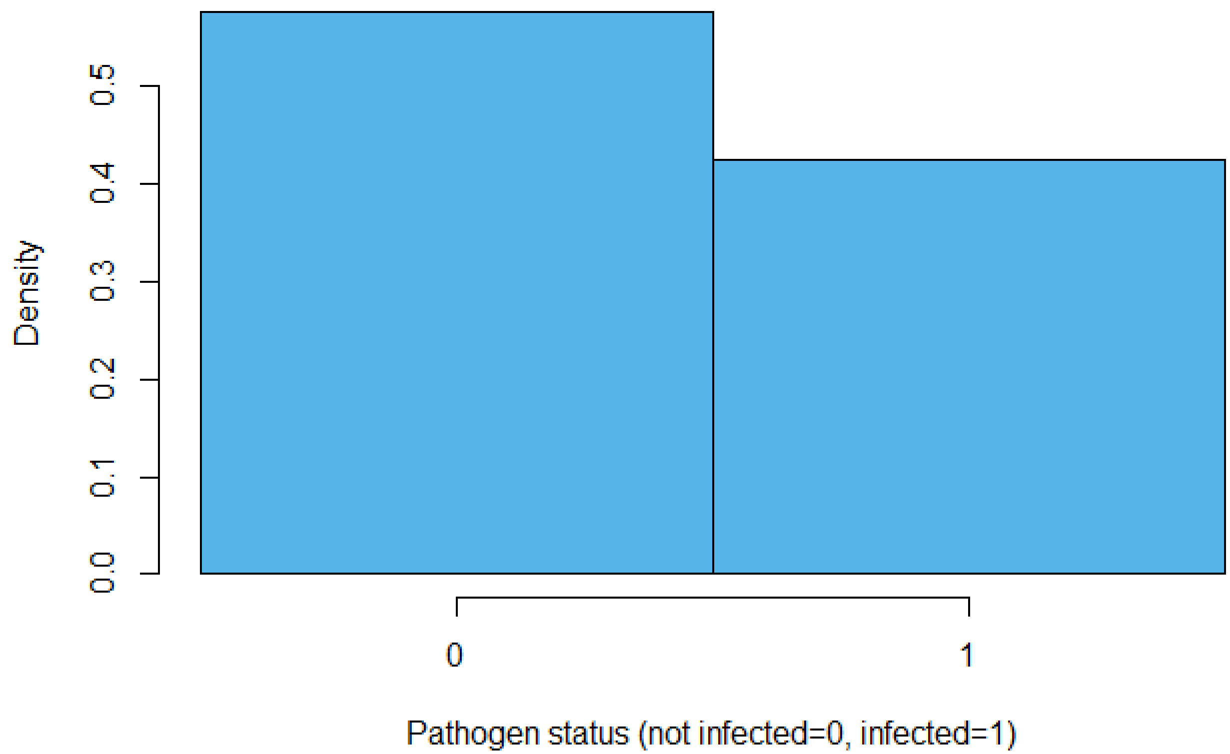
Take a sample:

```
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

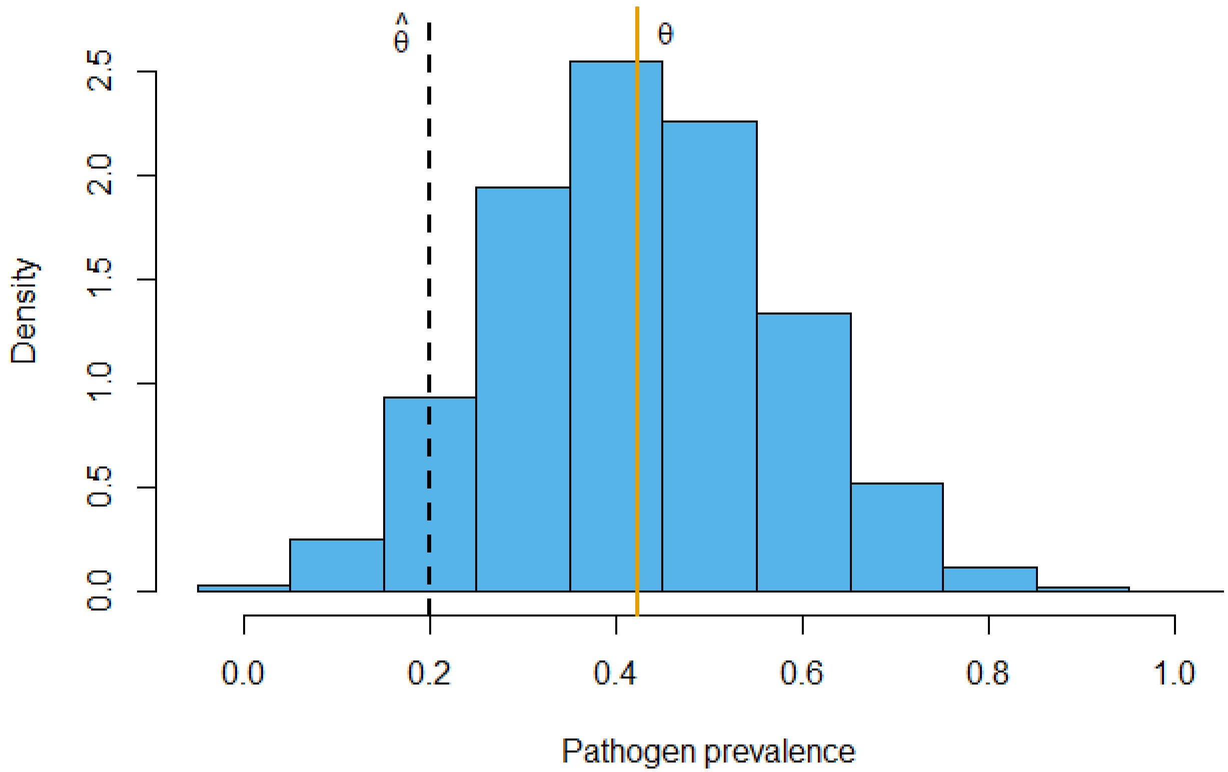
Invoke the law of large numbers

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

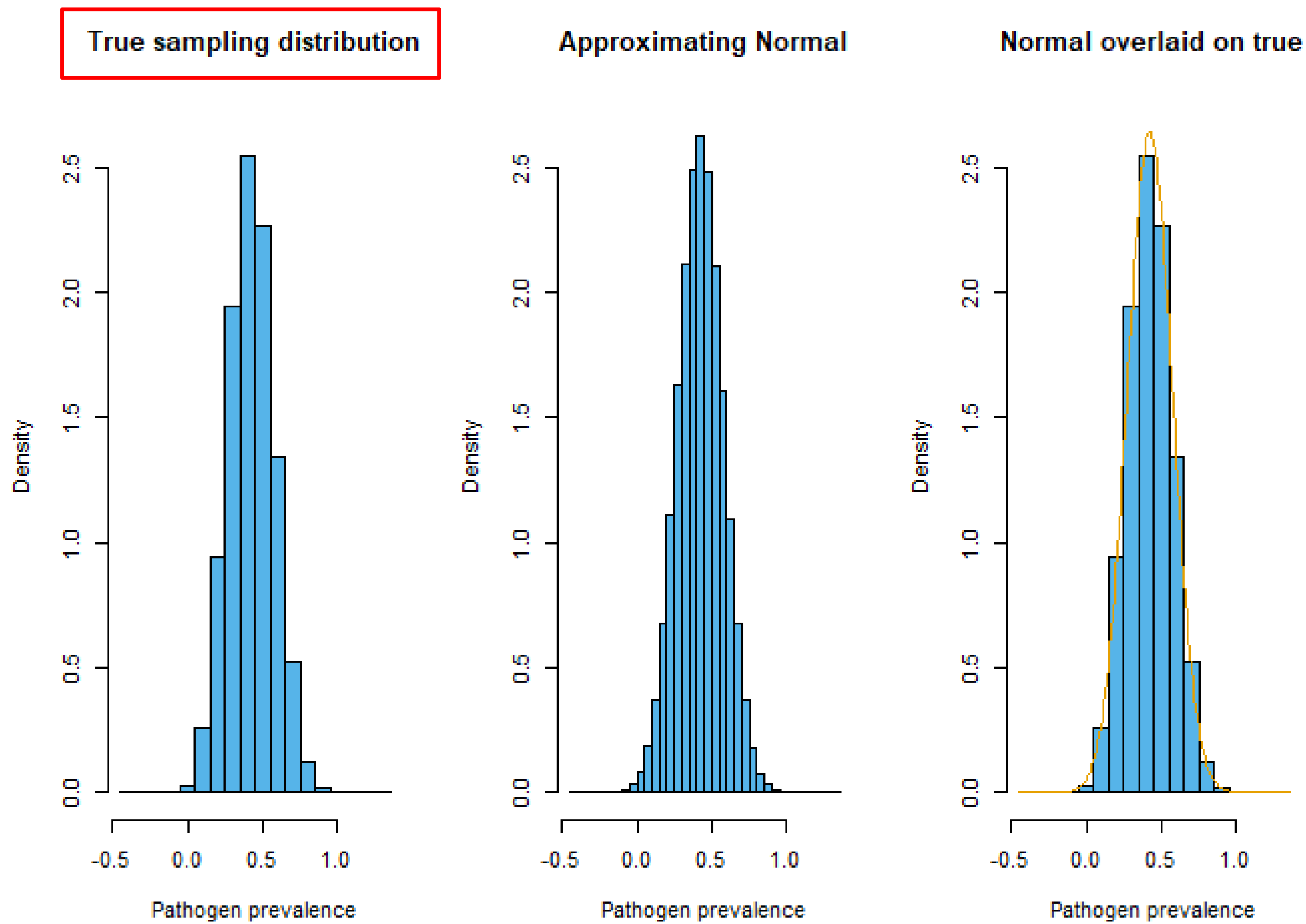
The sampling distribution for prevalence



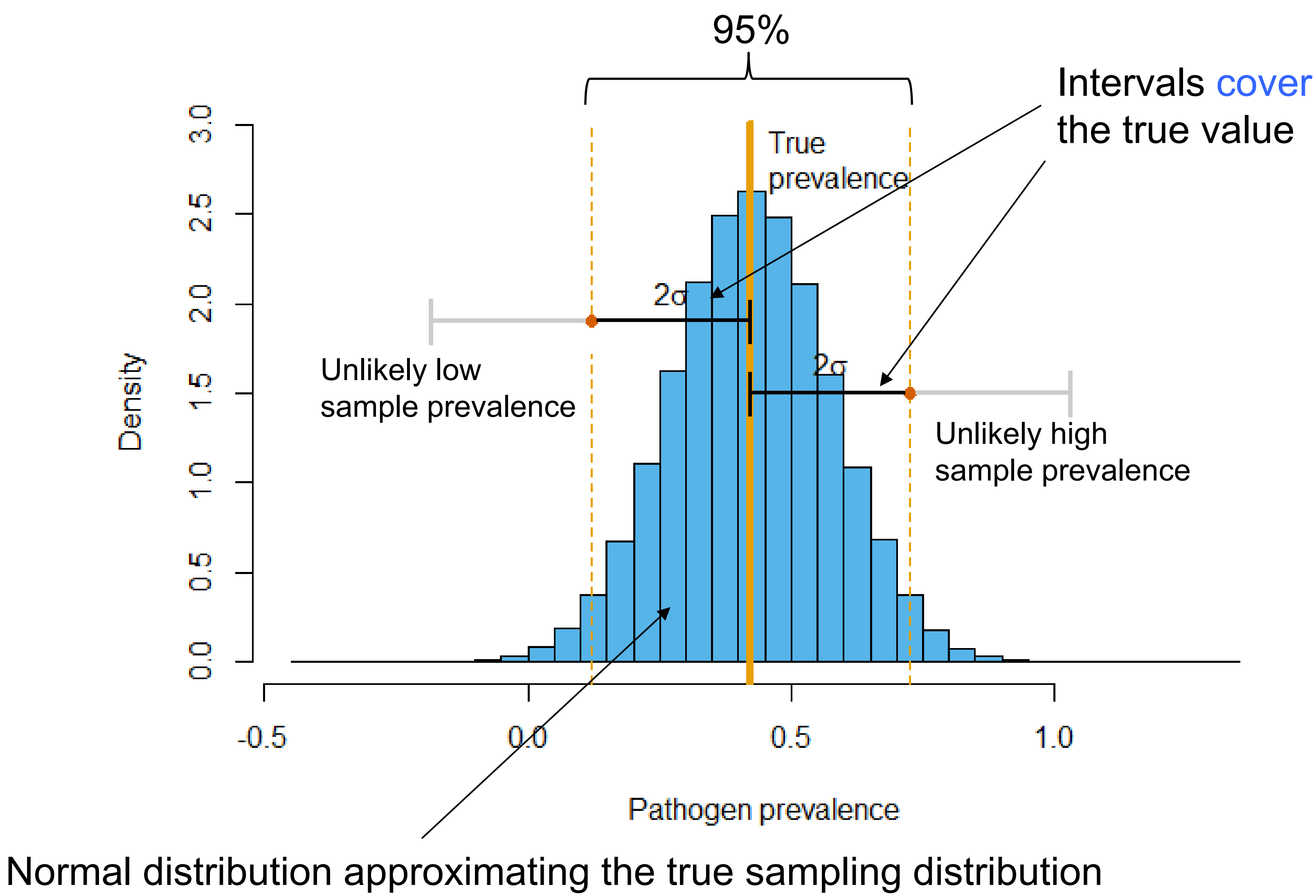
Confidence interval

- 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
Density
0.0
0.5
1.0
1.5
2.0
2.5
-0.5
0.0
0.5
1.0
Pathogen prevalence

Construct an interval to cover true value



Plug in principle

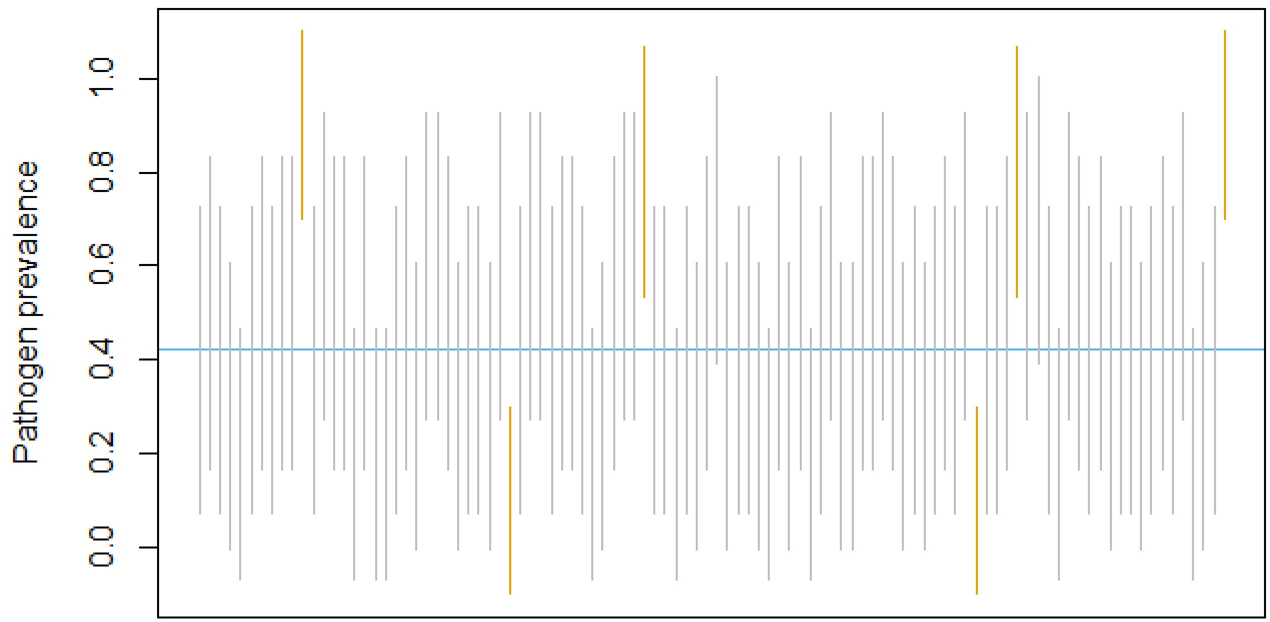
- We don't know the true sampling distribution or its parameters
- Plug in the sample instead as an estimate
  - in this example we can use the standard error of the sample as an estimate of the standard deviation of the sampling distribution

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)

lm() inference algorithms

Sampling distribution for parameters β_0, β_1

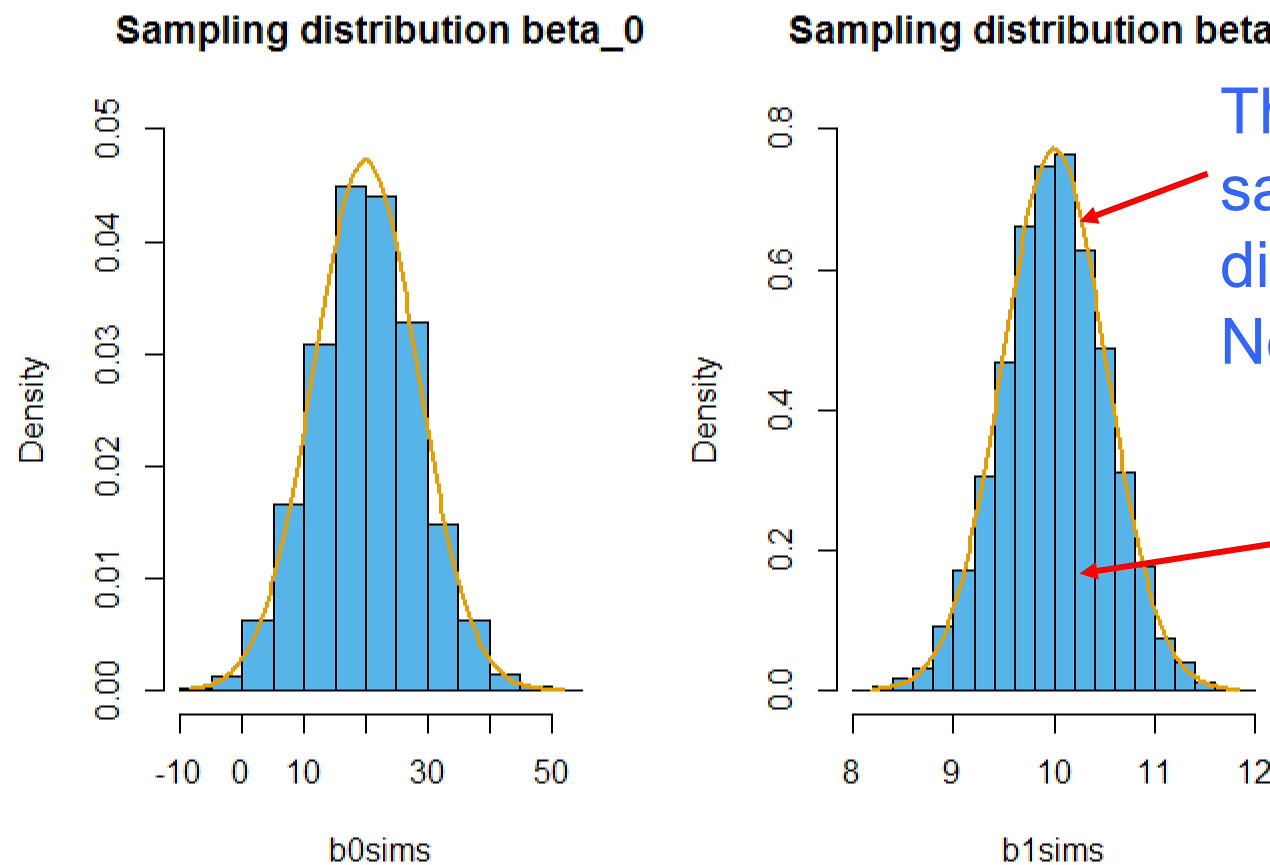
repeat very many times
- sample data from the population
- fit the linear model
- estimate the parameters

plot sampling distribution (histogram) of parameter estimates

Sampling distribution for any other quantities
(e.g. mean of y given x) is similar

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Population: normal distribution of errors



Theory says sampling distribution is Normal

Simulation

Plug-in principle

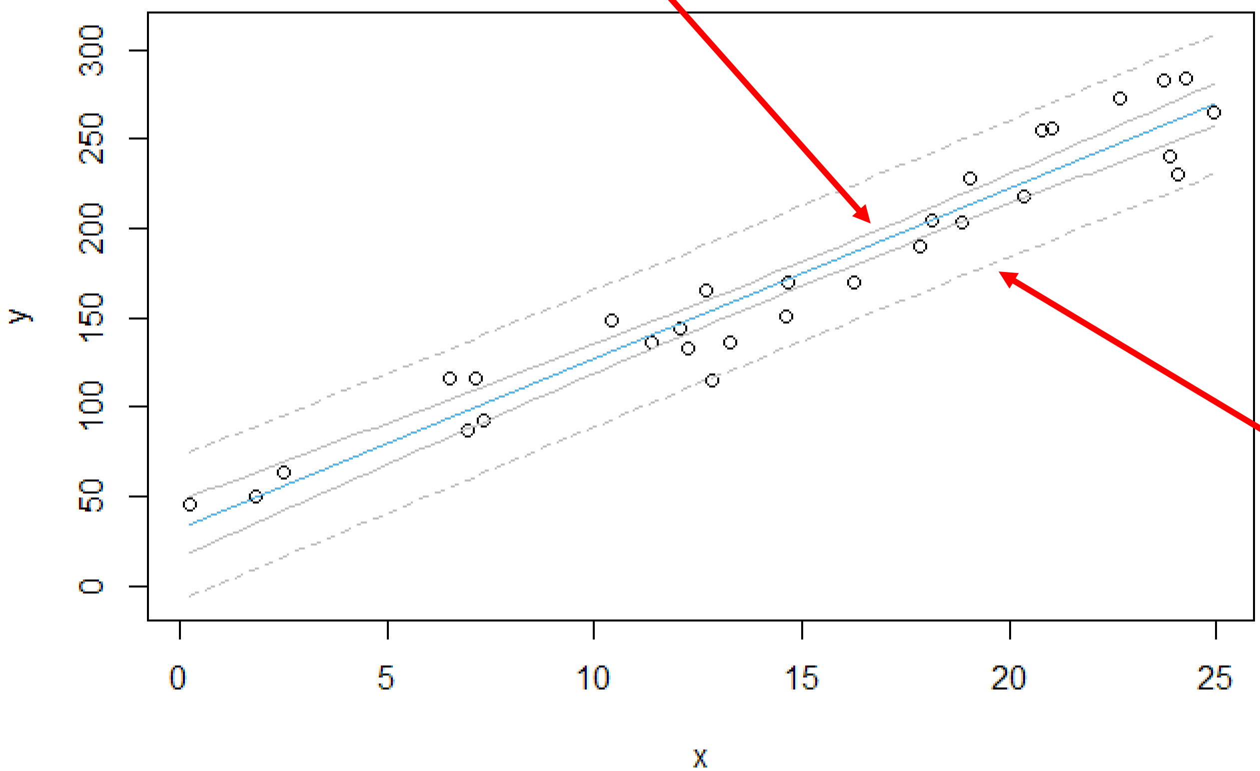
- We don't have access to the true sampling distribution or its parameter values
- Plug in the residual standard error from the sample to estimate the parameters (σ) of the sampling distribution

P-values

- The probability of a sample statistic as large **or larger** than the one observed **given that some hypothesis is true**
- p-value for lm parameters:
- Obtained from the sampling distribution of the parameters (*t* standardized)
- *t* is β in standard error units
- hypothesis is null (beta = β, sd=s.e.)

Confidence vs prediction intervals

CI: uncertainty in mean response (estimation uncertainty)



PI: uncertainty in individual response (estimation uncertainty + data generating process)

Robustness

- Normality of e_i is not that crucial
- More relevant: sampling distributions for β are Normal
  - central limit theorem says whatever the e_is, the sampling distribution will tend Normal
- Most problematic: when e_i is asymmetrical or heteroscedastic

R code - most common inferences

```
plot(x,y)
fit <- lm(y ~ x)
summary(fit)
confint(fit)
newd <- data.frame(x = seq(min(x), max(x), length.out=100))
pred_w_ci <- cbind(newd,predict(fit, newd, interval = "confidence"))
pred_w_pi <- cbind(newd,predict(fit, newd, interval = "prediction"))
lines(pred_w_ci[c(1,nrow(pred_w_ci)),c("x","fit")],col="#56B4E9")
lines(pred_w_ci[,c("x","lwr")],col="grey")
lines(pred_w_ci[,c("x","upr")],col="grey")
lines(pred_w_pi[,c("x","lwr")],col="grey",lty=2)
lines(pred_w_pi[,c("x","upr")],col="grey",lty=2)
plot(fit,1:6)
```

Open lab

- Problems from previous homework
- Prediction intervals
  - lm() with your dataset
  - prediction intervals for y|x
  - also plot confidence intervals for y|x