Inference algorithms

Statistical inference

- Judge the accuracy of an estimation or prediction algorithm
  - Efron & Hastie 2016
- Reliability
- Uncertainty

ISO definition of accuracy: the closeness of a measurement to the true value
Two components: bias, variance

Different inference problems

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

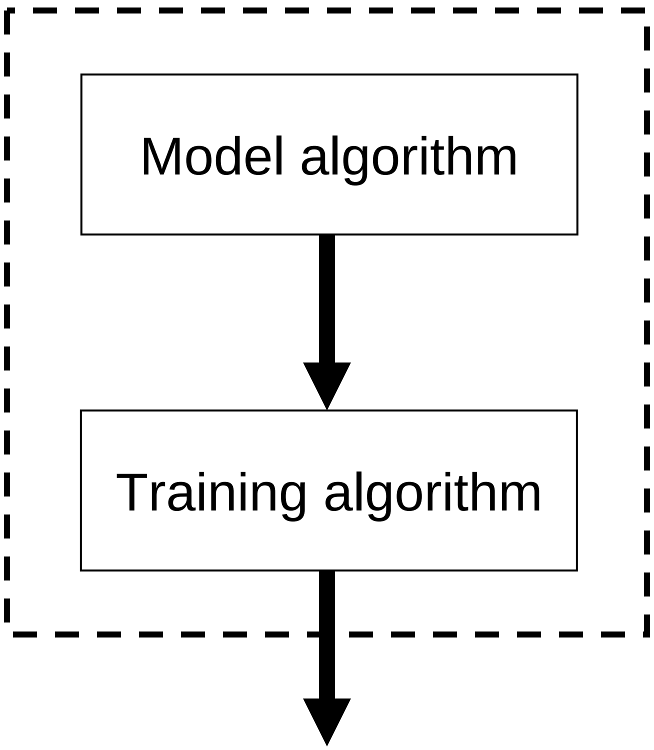
Model selection
Infer the data generating process from among a set of candidate data-generating processes

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)

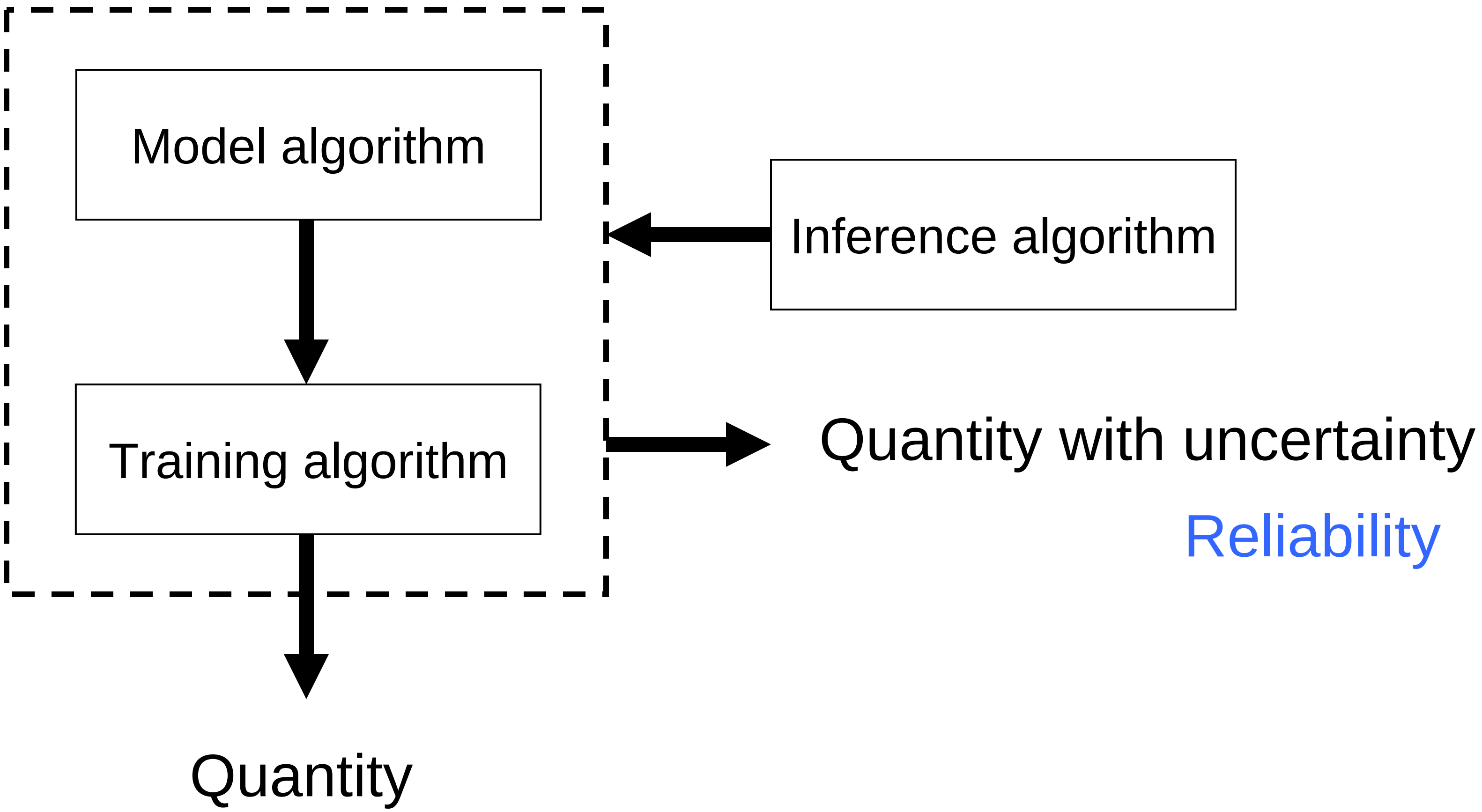
Algorithms in data science



Quantity

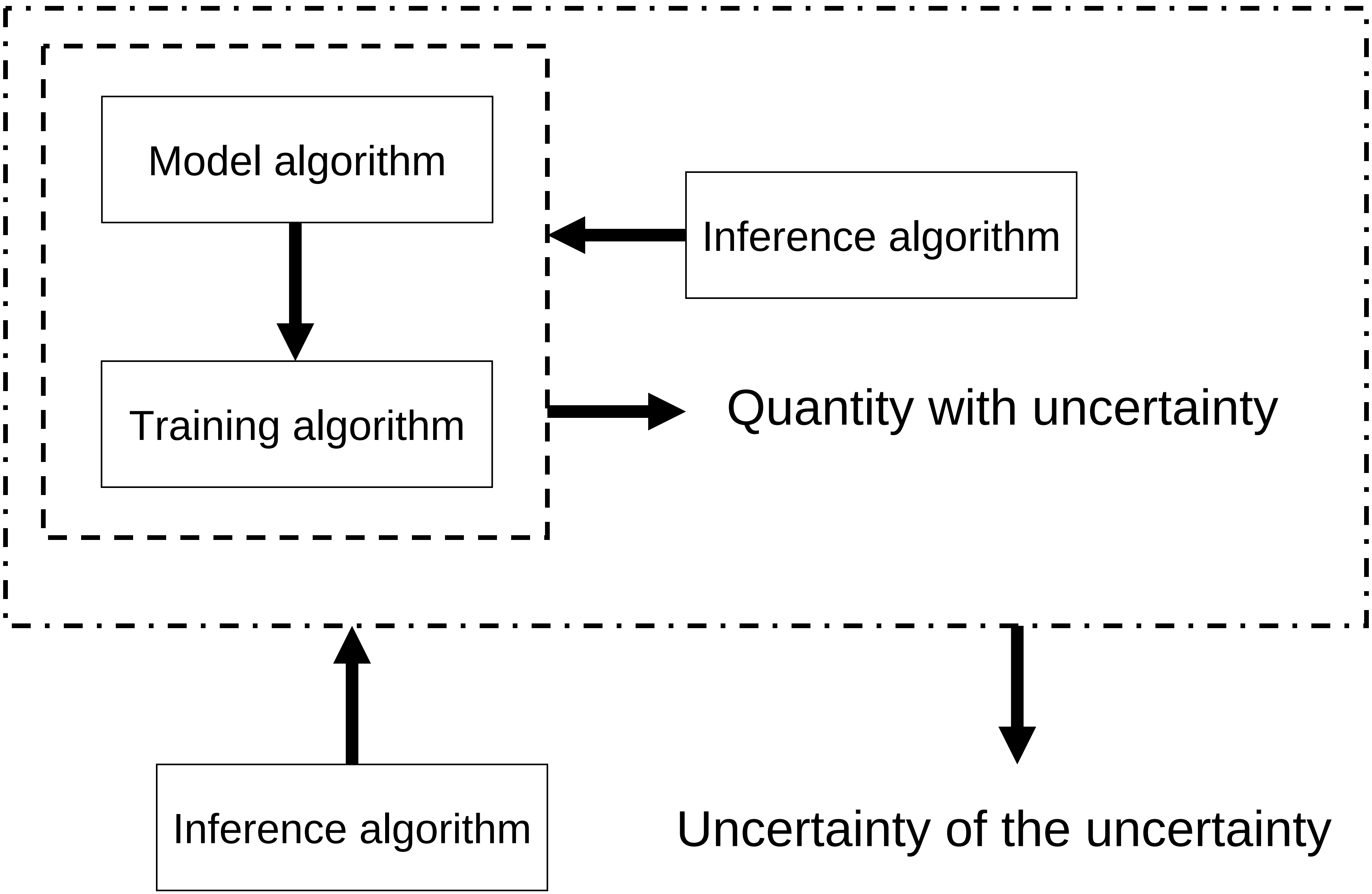
"Dumb" - doesn't say about reliability

Algorithms in data science



"Dumb" - doesn't say about reliability

Algorithms in data science



Inference algorithms

- Looking back: considering all the ways data could have happened
- Looking forward: predicting new data and testing against them

These are two big ideas in data science

Inference algorithms

- Looking back: considering all the ways data could have happened
  - Frequentist: sampling distribution
  - Bayesian/likelihood: P(data|model)

Sampling distribution

132 orange-spotted warblers. 1 indicates female

```
f <- 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)
```

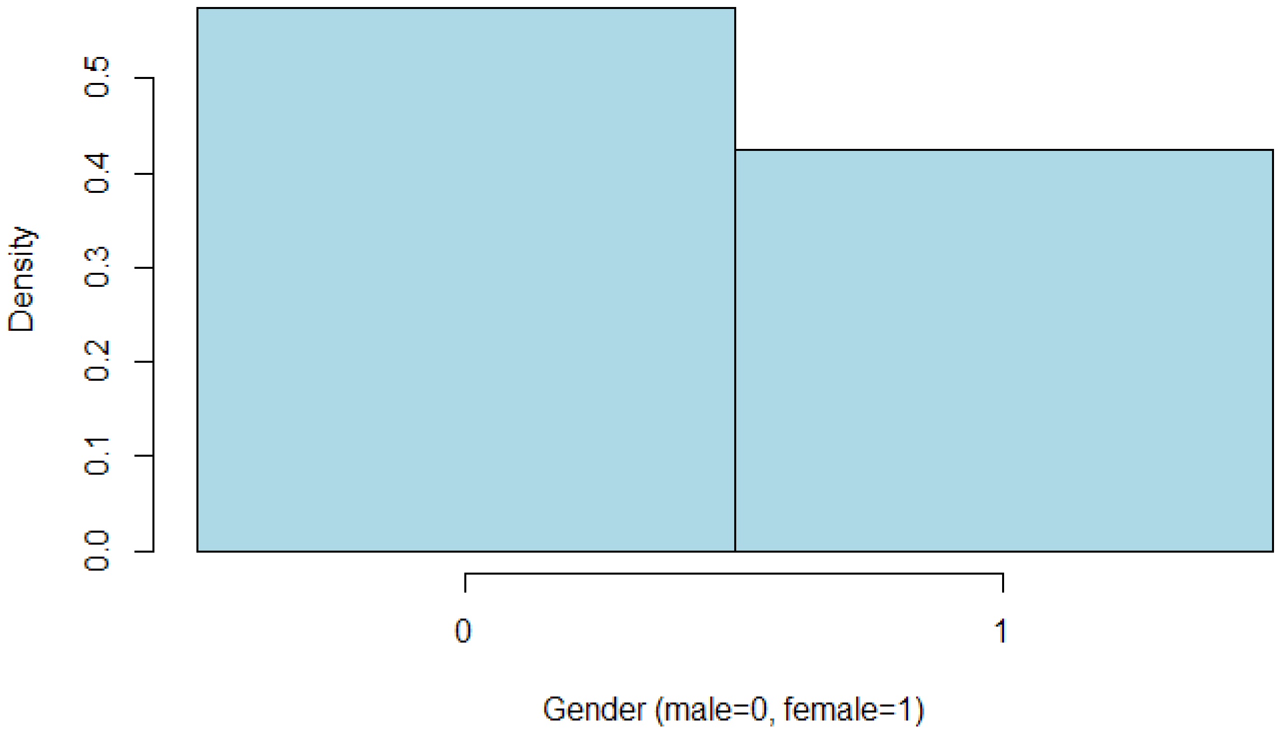
Take a sample:

```
sample(f,10)
```

0 1 0 0 0 0 0 1 0 0

sex ratio = 0.2

Our scientific observation



True sex ratio 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 bird sex ratio

There are 3e14 possible samples.
Too hard! It would take 100 years to compute!

Sampling distribution algorithm 2

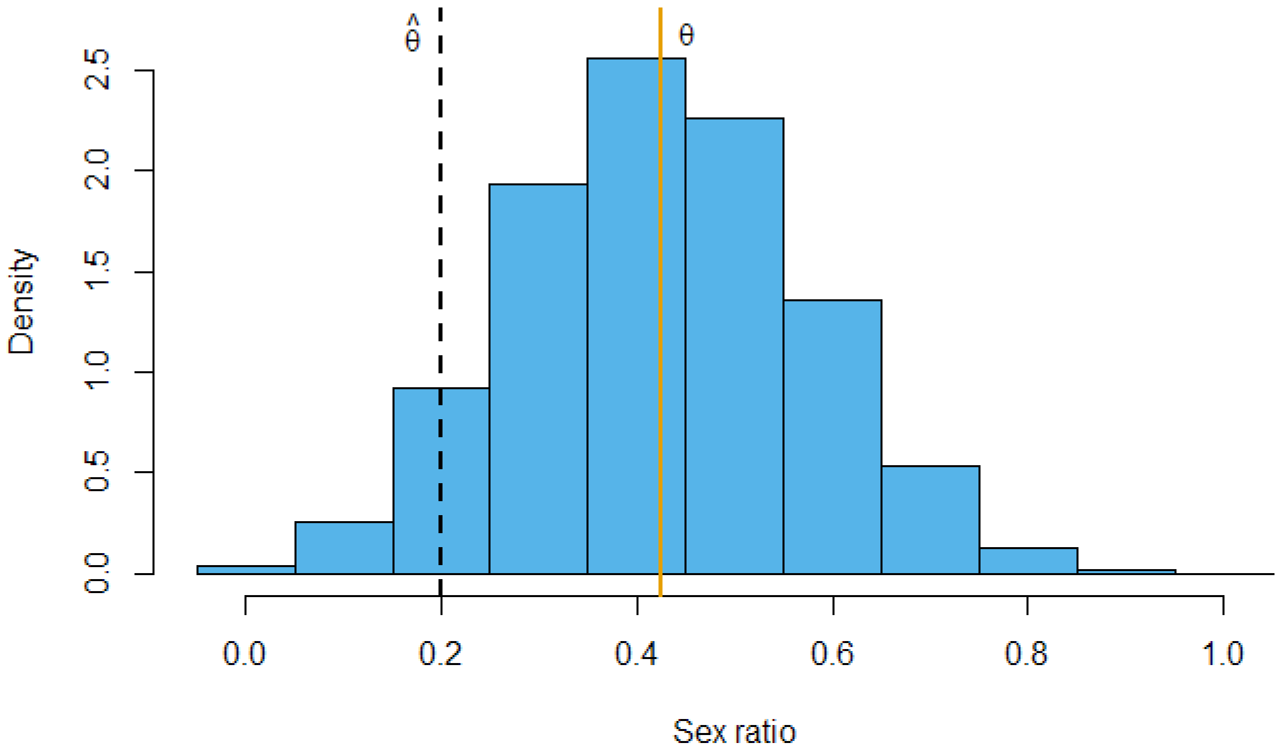
We can invoke the law of large numbers

repeat very many times
    sample n units from the population
    calculate the sample statistic
plot sampling distribution (histogram) of the sample statistic

for bird sex ratio

for a large number of repeated samples
    randomly sample 10 birds from the population
    calculate the sex ratio in the sample
plot sampling distribution (histogram) of sex ratios

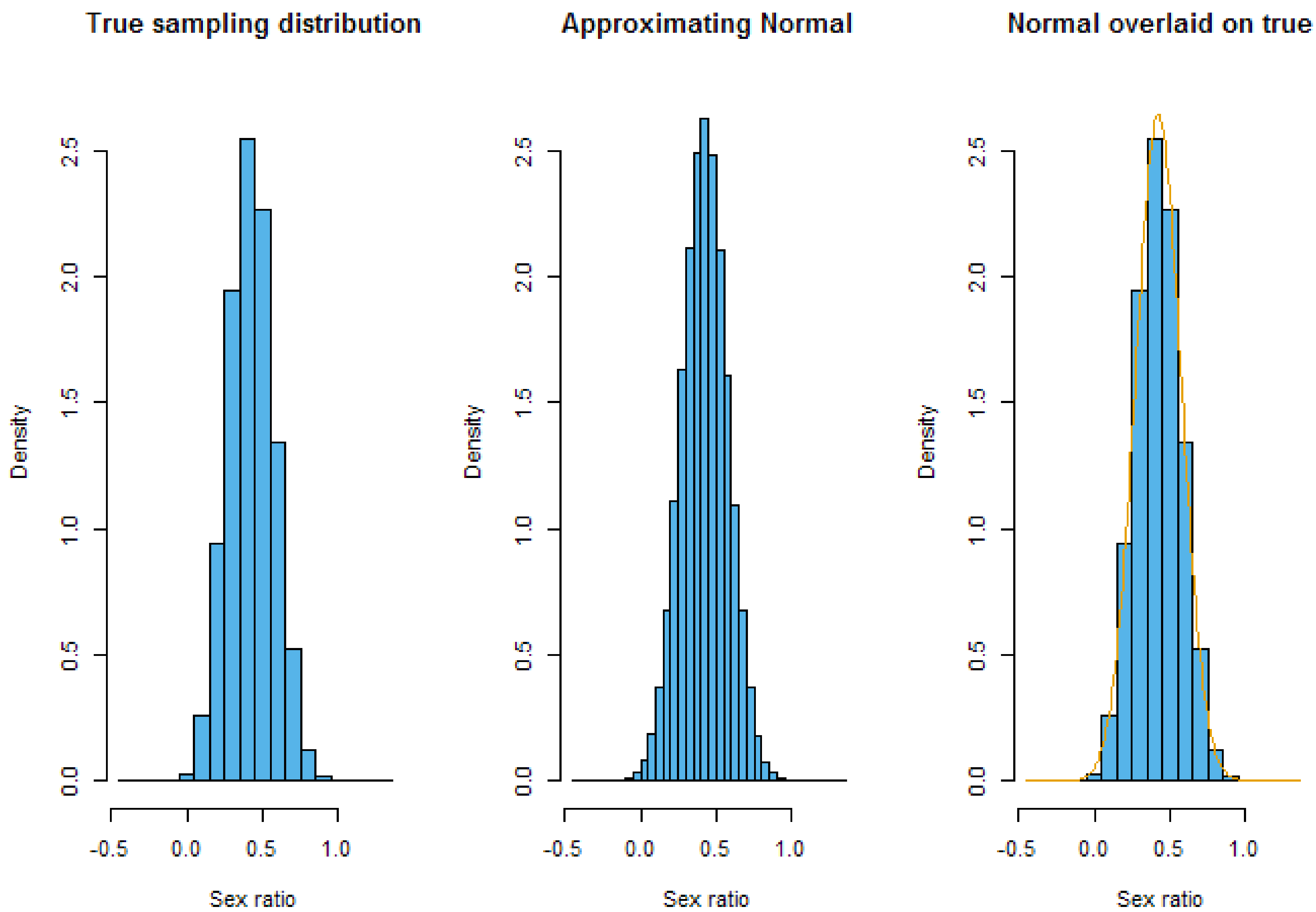
The sampling distribution for sex ratio



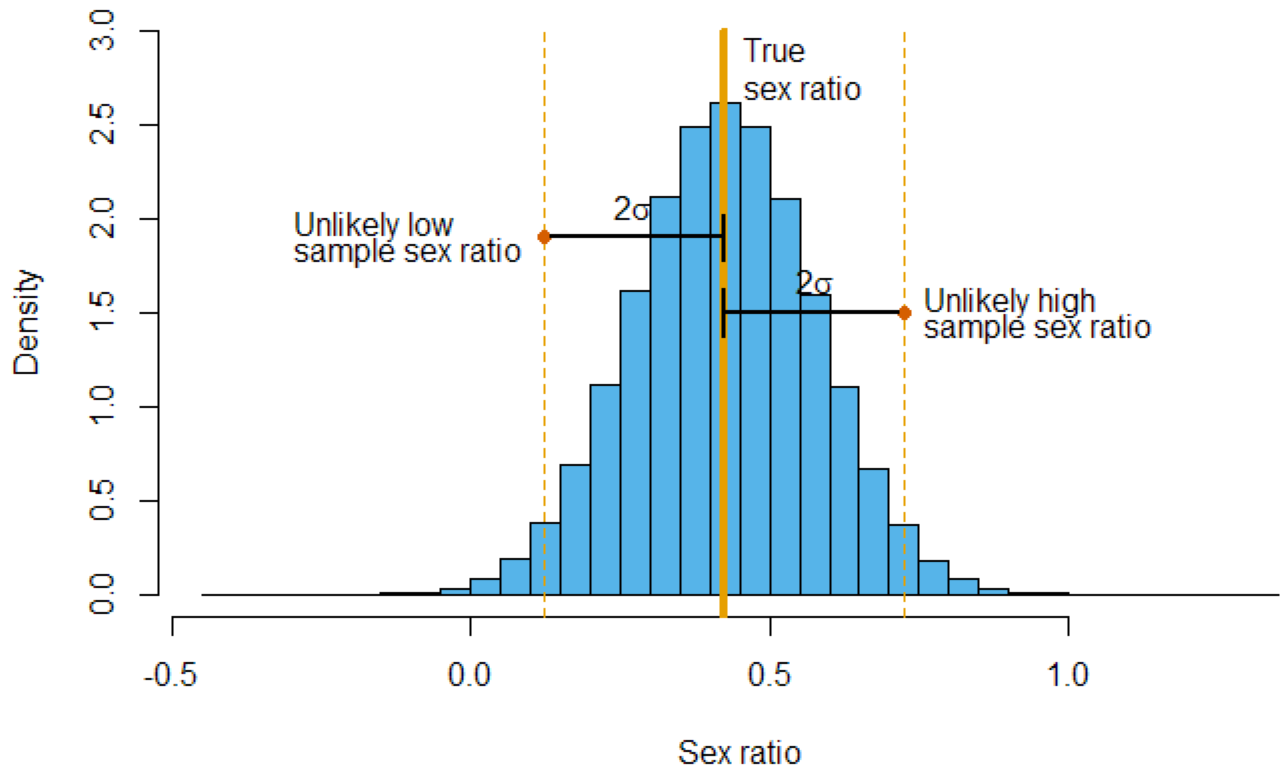
Confidence interval

- 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
Sex ratio
0.0
0.5
1.0
1.5
2.0
2.5
-0.5
0.0
0.5
1.0

Using the normal approx for inference



Plug in principle

- We don't know the true sampling distribution or its parameters
- Plug in the sample instead as an estimate

Coverage

repeat very many times

    sample n units from the population

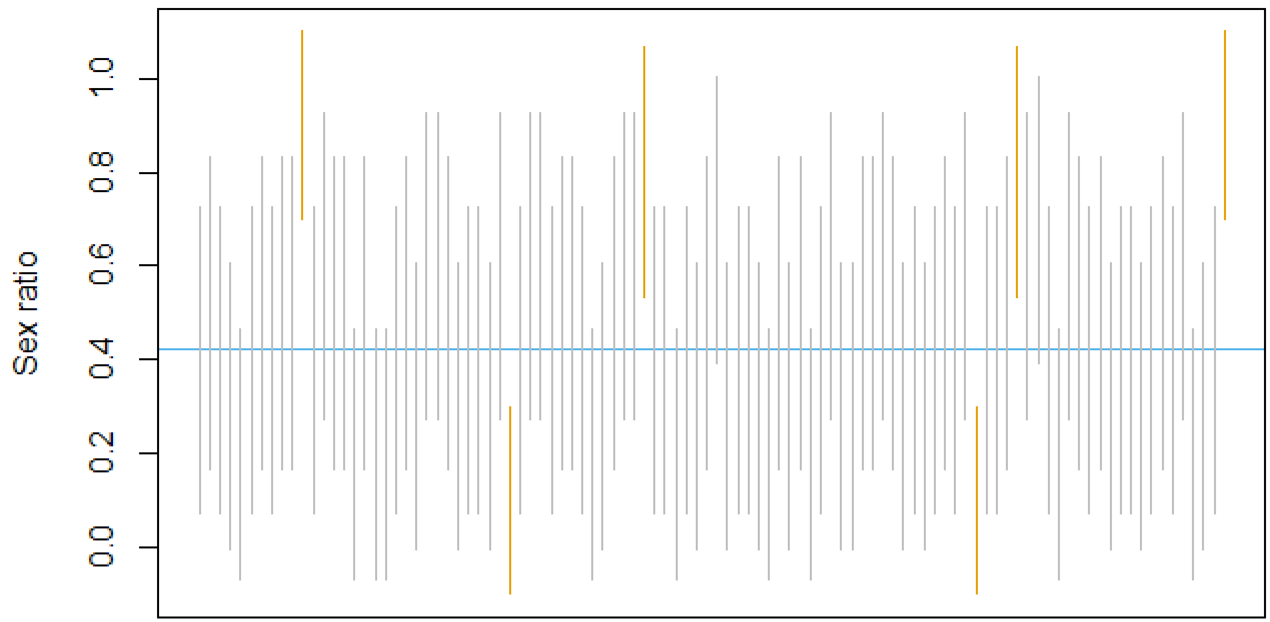
    calculate the sample statistic

    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

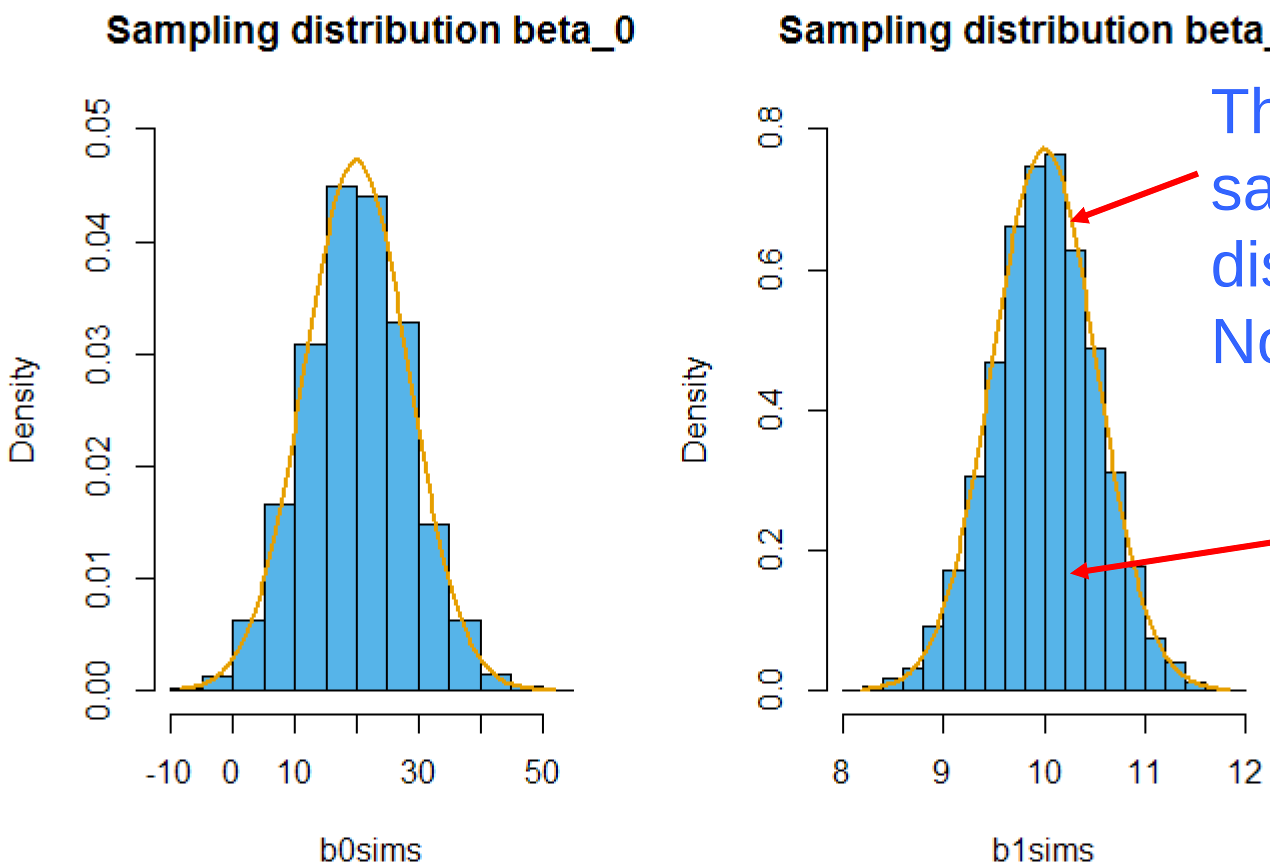
    estimate the parameters

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

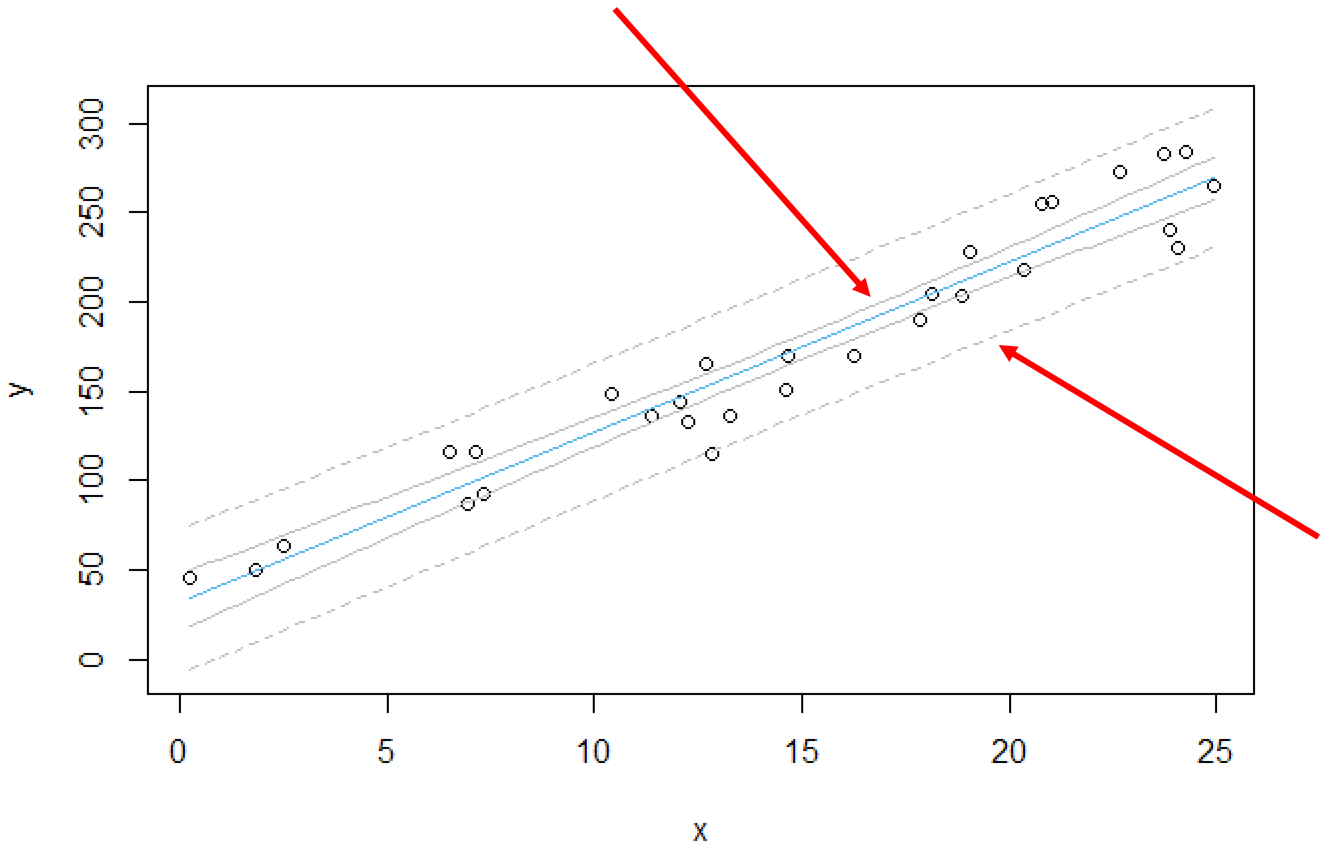
- We don't have access to the true sampling distribution or its parameter values
- Plug in the residual standard error from the sample to estimate the parameters (σ) of the sampling distribution

P-values

- The probability of a sample statistic as large **or larger** than the one observed **given that some hypothesis is true**
- Obtained from the sampling distribution of the parameters (*t* standardized)
- t is β in standard error units

Confidence vs prediction intervals

CI: uncertainty in mean response



PI: uncertainty in individual response

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)
```