lm() inference algorithms

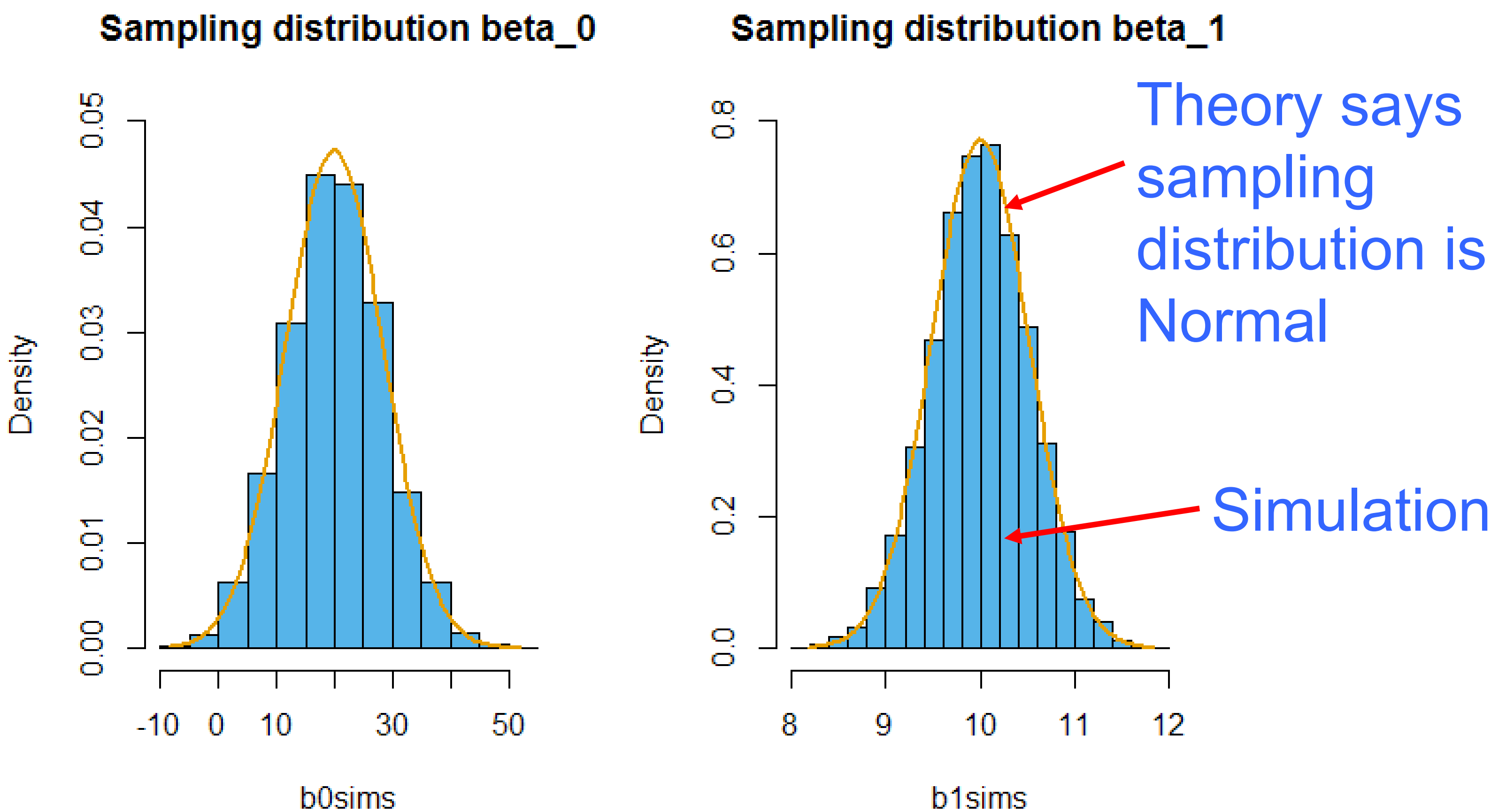
Sampling distribution for parameters β_0, β_1

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



Plug-in principle

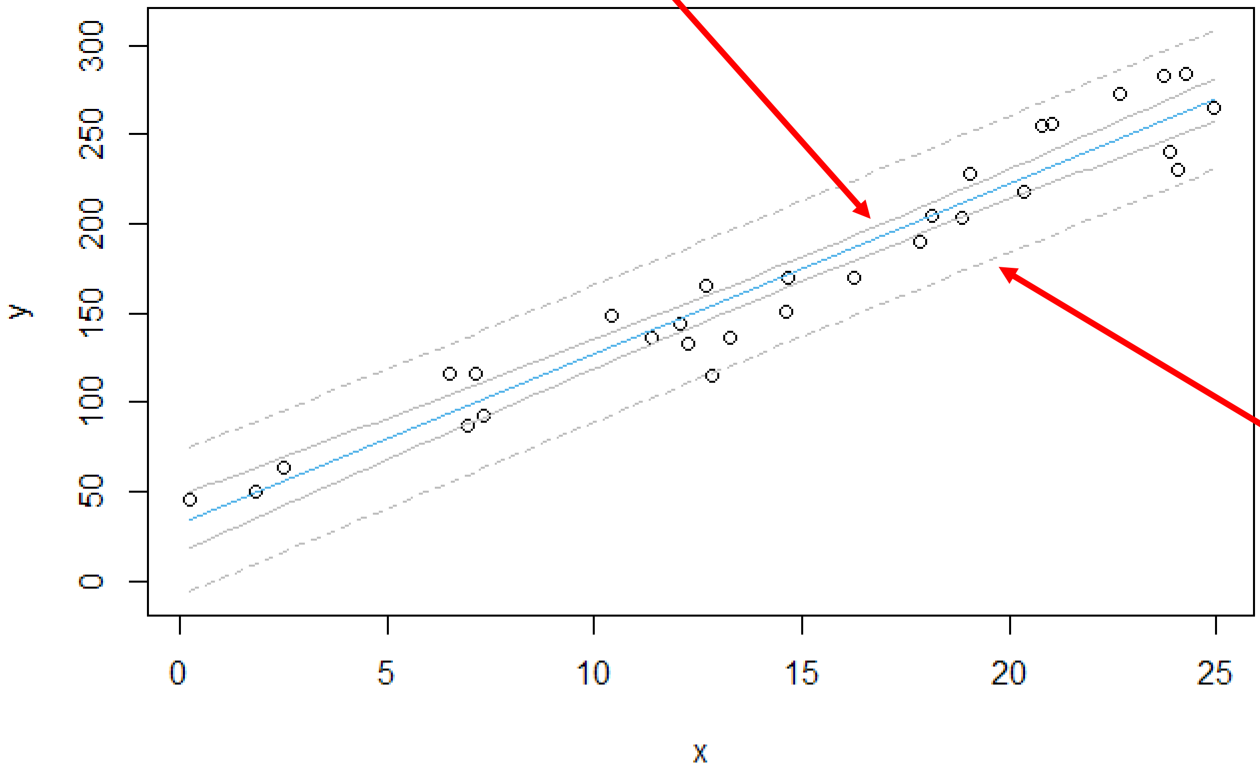
- We don't have access to the true sampling distribution or its parameter values
- Plug in the residual standard error from the sample to estimate the parameters (σ) of the sampling distribution

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

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)
```