Today

- Questions from homework
- Frequentist inference algorithms
  - linear model sampling distribution
  - classical confidence intervals
  - bootstrap

Frequentist inference algorithms

- All frequentist inferences are based on the sampling distribution
- The sampling distribution is the frequentist approach to considering all the ways data could have happened (i.e. looking back)

Main concepts from HW videos

- Sampling distribution algorithm (4 lines)
- Sampling distribution imagined but true
- Confidence interval covers true value, x%
  - reliability of procedure
- Plug in principle
- Coverage algorithm, adds line:
  - calculate the interval for the sample statistic

Frequentist inference recipe

1) Make a model (biologically informed)
2) Which quantity of the model corresponds to the science? (map model to question)
    - parameters?
    - f(parameters)? e.g. predictions of y
3) Train the model
4) Get quantity from the trained model (sample statistic)
5) Derive an inference (e.g. 95% CI) from the sampling distribution of the sample statistic
6) Plug in an estimate of the sampling distribution

Linear model parameters

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

$$e_i \sim \mathrm{Normal}(0, \sigma^2)$$

Sampling distribution algorithm

repeat very many times

    sample data from the population

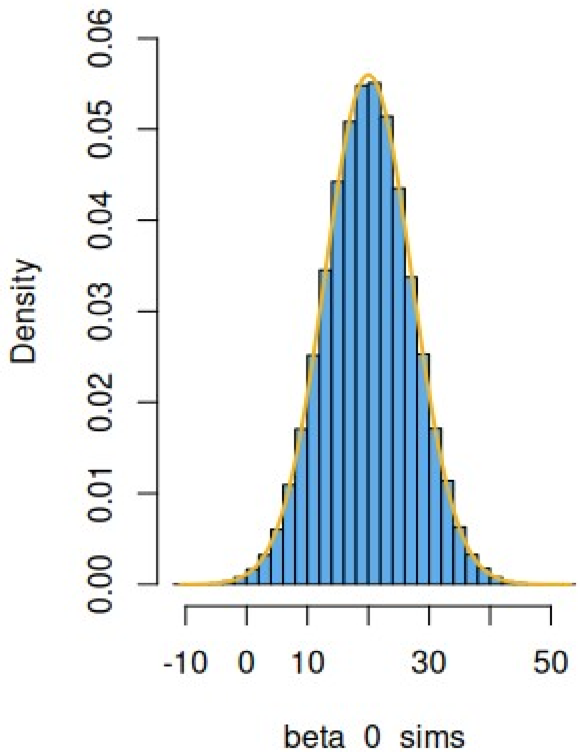
    train the linear model to estimate the parameters

plot sampling distributions (histograms) of the parameter estimates

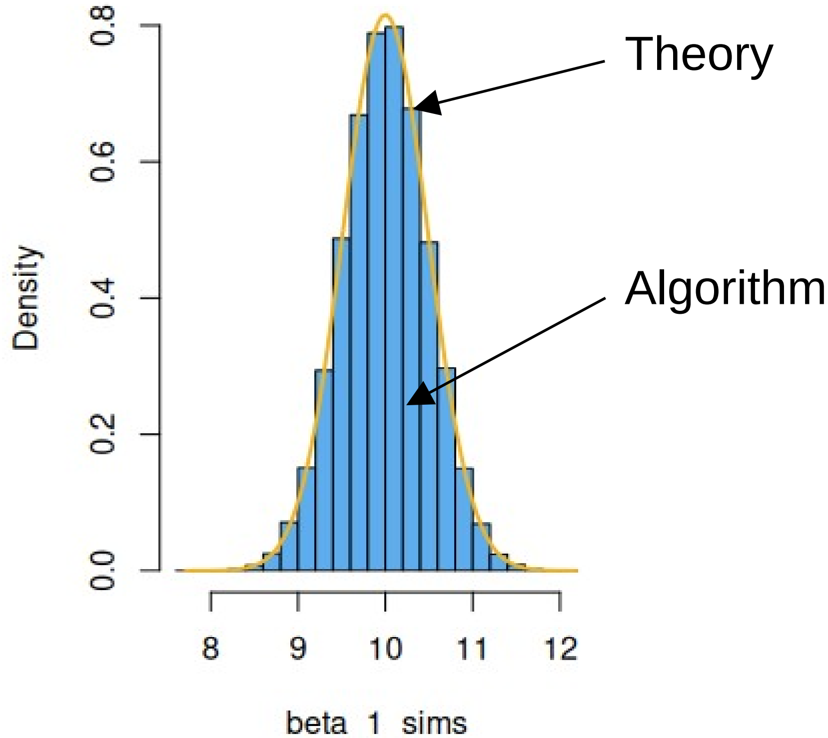
Code: linear_model_sampling_distribution.md

Sampling distributions

β_0



β_1



Confidence interval (95%)

$$\beta_0 \pm t_{95}\sigma_0$$

Contains 95% of true sampling distribution

$$\beta_1 \pm t_{95}\sigma_1$$

$$t_{95} \approx 2.0$$

Plug in

$$\hat{\beta}_0 \pm t_{95}\hat{\sigma}_0 \qquad \hat{\sigma}_0 = \mathrm{fn}\left(\hat{\sigma}_e\right) \qquad \hat{\sigma}_e = \mathrm{fn}\left(\mathrm{SSQ}\right)$$

$$\hat{\beta}_1 \pm t_{95}\hat{\sigma}_1 \qquad \hat{\sigma}_1 = \mathrm{fn}\left(\hat{\sigma}_e\right)$$

Hat indicates estimate from the sample

Bootstrap

Sampling distribution algorithm

repeat very many times
    sample data from the population
    train the model to estimate the parameters
plot sampling distribution (histogram) of the parameter estimates

Bootstrap algorithm

repeat very many times
    generate data based on the sample ← plug in
    train the model to estimate the parameters
plot sampling distribution (histogram) of the parameter estimates

Bootstrap algorithms

- Non-parametric bootstrap
  - resample the data
- Empirical bootstrap
  - resample the residuals
- Parametric bootstrap
  - generate data from a distribution
  - use estimated parameters of the distribution

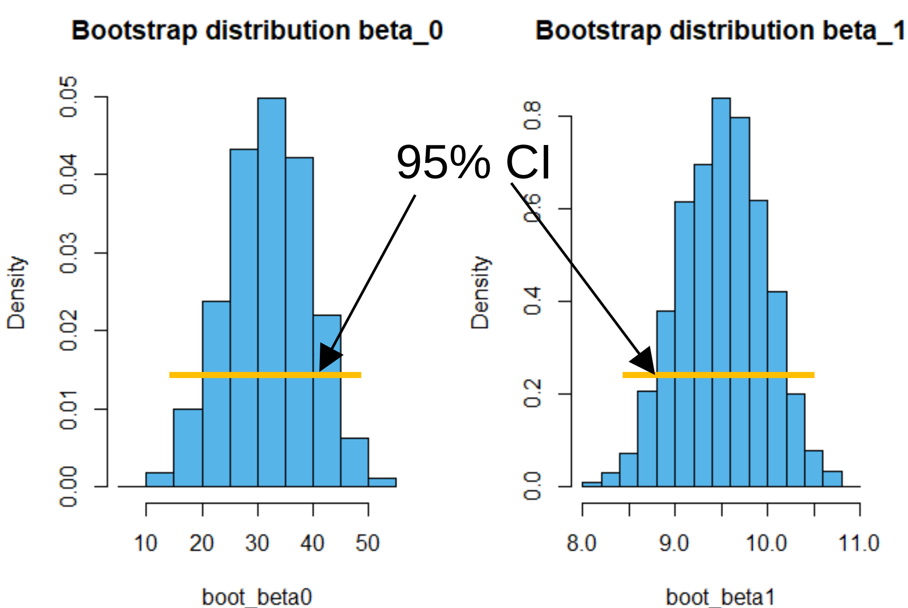
Huge advantage

- Can obtain reliability/uncertainty for any quantity that can be calculated from any fitted model

Code (e.g. empirical bootstrap)

```
for ( i in 1:10000 ) {
    e_boot <- sample(e_fit, replace=TRUE)
    df_boot$y <- coef(fit)[1] + coef(fit)[2]*df_boot$x + e_boot
    fit_boot <- lm(y ~ x, data=df_boot)
    boot_beta0[i] <- coef(fit_boot)[1]
    boot_beta1[i] <- coef(fit_boot)[2]
}
```

plug in



Pseudocode
- Train model, save errors
- For many times
    - Resample errors with replacement
    - Create new y-values at original x values
    - Train the model
    - Keep parameter estimates

Bootstrap: further reading

Brief exposition:

James G, Witten D, Hastie T, Tibshirani R (2021). An Introduction to Statistical Learning: With Applications in R, Second edition. Springer, New York. Chapter 5.2.

Definitive references:

Davison AC, Hinkley DV (1997). Bootstrap Methods and Their Application. Cambridge University Press, Cambridge ; New York, NY, USA.

Efron B, Tibshirani R (1993). An Introduction to the Bootstrap. Chapman & Hall, New York.