CQS Summer Institute: Machine Learning and Statistics in R

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN, USA

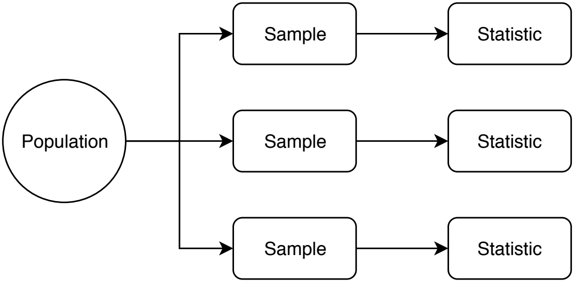
August 17, 2018

Course Overview

- Syllabus and R code:
- https://github.com/biostatmatt/cqs-ml-stat-r
- Monday: Intro and Data Management
- Tuesday: Supervised Learning Part 1
- Wednesday: Supervised Learning Part 2
- Thursday: Unsupervised Learning
- Friday: Statistical Inference

Statistical Inferences

- Statistical inferences are statements about an unobservable property of a population (i.e., a *population parameter*), that are based on a sample from that population.
- Inferences are based on sample statistics
- Under model assumptions, the distribution of statistics across samples can be deduced.
- In order to make inferences about a population parameter using a sample statistic, the *sampling distribution* of the statistic must depend on the parameter. The sampling distribution can then be used to make statistical inferences.


Population
Sample
Sample
Sample
Statistic
Statistic
Statistic

The Wald confidence interval

- Confidence intervals satisfy the following:

$$P(\hat{\theta} + C_L < \theta < \hat{\theta} + C_H) = 1 - \alpha$$

  where θ is the parameter of interest, $\hat{\theta}$ is a sample statistic, C_L and C_H define the lower and upper bounds, and $1 - \alpha$ is the confidence level or *coverage* of the interval.

The Wald confidence interval

- Using model assumptions, approximations, and asymptotic arguments, it is generally possible to find a sample statistic $\hat{\theta}$ that is approximately normally distributed with mean θ and known variance σ^2
- The following expressions are then approximately valid

$$P(\phi_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma^2} < \phi_{1-\alpha/2}) = 1 - \alpha$$
$$P(\hat{\theta} - \sigma^2\phi_{1-\alpha/2} < \theta < \hat{\theta} + \sigma^2\phi_{1-\alpha/2}) = 1 - \alpha$$

- Thus, the last expression gives an approximate $100\% \times (1 - \alpha)$ *Wald* confidence interval for θ

The Wald confidence interval

- Wald-type confidence intervals are ubiquitous, and are the default for many statistical routines.
- Confidence intervals are probability statements about about population parameters; statistical inferences.
- However, if any of the model assumptions, approximations, or asymptotic arguments are not valid, this probability statement may be incorrect, i.e., the nominal and actual coverages may be different.
- Unfortunately, in practice, we can not directly validate either the model assumptions or the coverage accuracy.

Predictions vs. Inferences

Inferences

- Statistical inferences are statements about an unobservable property of a population (i.e., a *population parameter*), that are based on a sample from that population.
- Parameters cannot be observed or measured directly
- Accuracy of inferences cannot be assessed directly
- Dependent on model assumptions, which should be “checked”

Predictions:

- Predictions are statements about observable quantities generated by a member of a population, e.g., whether an event will occur within some period of time, that are based on a sample from that opulation.
- Accuracy of predictions can be assessed directly
- Thus, if predictions are sufficiently accurate, the validity of model assumptions is irrelevant

Simulation as a tool

Using simulation to evaluate effect of invalid assumptions:

Inferences

- Specify a "true" population model
- Simulate a sample from that population
- Make inferences using invalid assumptions
- Evaluate accuracy of inferences
- E.g., compare nominal vs. actual coverage

Predictions

- Specify a "true" population model
- Simulate a sample from that population
- Create prediction model using invalid assumptions
- Evaluate accuracy of predictions
- E.g., compare nominal test error vs. actual test error

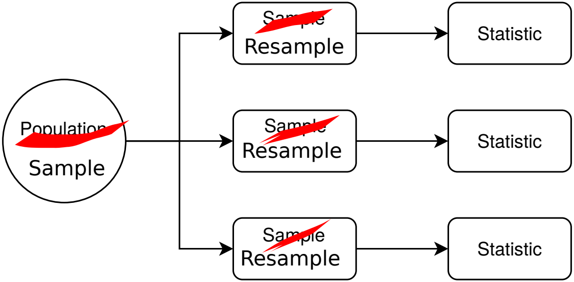
Modeling assumptions for linear regression

- associations can be modeled linearly
- no unmeasured confounders
- errors are additive
- errors are normally distributed
- errors have constant variance (homoscedasticity)

Invalid Modeling Assumptions: inferences-vs-predictions.R

Bootstrap

- Bootstrap is a very powerful and general tool
- Easy to use
- Helps us approximate sampling distributions
- Depends on fewer assumptions
- Make inferences more robust to bad model assumptions



Population
Sample
Sample
Resample
Sample
Resample
Sample
Resample
Statistic
Statistic
Statistic

Bootstrap terminology

- θ - unknown parameter
- $x_1, \ldots, x_N$ - original sample
- $x_{11}^*, \ldots, x_{1N}^*$ - first bootstrap sample
- $x_{B1}^*, \ldots, x_{BN}^*$ - B^{th} bootstrap sample
- $\hat{\theta}$ - original sample statistic
- $\hat{\theta}_1^*$ - first boostrap sample statistic
- $\hat{\theta}_B^*$ - B^{th} boostrap sample statistic

Bootstrap confidence interval

- Still need to satisfy the following:

$$P(\hat{\theta} + C_L < \theta < \hat{\theta} + C_H) = 1 - \alpha$$

- Bootstrap simply substitutes $\hat{\theta}$ for θ, and $\hat{\theta}^*$ for $\hat{\theta}$:

$$P(\hat{\theta}^* + C_L < \hat{\theta} < \hat{\theta}^* + C_H) = 1 - \alpha$$
$$P(C_L < \hat{\theta} - \hat{\theta}^* < C_H) = 1 - \alpha$$

- But, the distribution of $\hat{\theta} - \hat{\theta}^*$ is easy to find using the boostrap. And C_L and C_H are simply the $\alpha/2$ and $1 - \alpha/2$ percentiles of that distribution.
- This method makes a different kind of approximation
- Often more robust to invalid modeling assumptions

Boostrap: inferences-vs-predictions.R

Sample bias

- Both predictions and inferences can be inaccurate if constructed using a sample that is not representative of the population.
- Selection bias - occurs when a sample is selected in a biased fashion, e.g., self-selected study participants (e.g., parturients who elect to use N_2O as an analgesic during delivery)
- Sample bias often undetectable; must be avoided.
- Inference vs. Prediction: either may be more or less robust

Other types of bias

- Recall bias
- Observer bias

Sample bias: inferences-vs-predictions.R

Wrap up

- Course completion certificates
- Class photo
- Course evaluations
- First time! We need feedback:
  - Content (more math? more supervised learning?)
  - Format (presentation, R code, labs)
  - Schedule/location/refreshments
- Thank you!