Lecture 1 Overview

Jeffrey Yau

1/8/2017

Main Topics Covered in Lecture 1:

- An introduction to categorical data, Bernoulli probability model, and Binomial probability model
- Computing the probability of binomial probability model
- Simulating a binomial probability model
- Estimating the Binomial probability model using maximum likelihood estimation (MLE)
- Confidence intervals:
  - Wald confidence interval
  - Alternative confidence intervals
- Hypothesis test for the probability of success
- The case of two binary variable
  - Contingency tables
  - The notions of relative risks, odds, and odds ratios

Required Readings:

BL2015: Christopher R. Bilder and Thomas M. Loughin. Analysis of Categorical Data with R. CRC Press. 2015.

- Ch. 1 (Skip Sections 1.2.6 and 1.2.7)

Optional Readings:

BL2015: Christopher R. Bilder and Thomas M. Loughin. Analysis of Categorical Data with R. CRC Press. 2015.

- Appendix A and B

An Overview of the Lecture

This lecture kicks off the first of the five lectures on the statistical analysis of discrete (response) data. It starts with an introduction of categorical data, the simplest of which is binary case, and the mathematical models for binary variables: *Bernoulli and Binomial* probability models. These models form the foundation to study regression models for discrete data. Importantly, the notions and techniques in the first five lectures provide a complementary perspective to the classification techniques in machine learning.

This lecture begins with the parameter and confidence interval estimation of these probability models and statistical inference. Pay special attention to the assumptions required for the *Binomial* probability model to be validly applied. When studying *Bernoulli and Binomial* probability models, ask "what parameters characterize the models and how they are different from those of the *Normal* probabiliy model that you encountered in your previous probability and statistic courses?" I ask these questions because it is often (mistakenly) thought that the mean and standard deviation are the only parameters of a probability model and are sufficient to characterist a probability distribution.

A few *"toy"* examples are included in the asynchronous lecture to illustrate the assumptions underlying the models. As you are watching the lectures and studying the readings, think of two examples: one that satisfies the assumptions of the *Binomial* probability model and one does not. We may ask you to share your examples in class.

The maximum likelihood estimation (MLE) method is used extensively for parameter estimation. Maximum likelihood estimators come with many desriable statistical properties. One important question you need to ask is "what condidtions are needed in the derviation of the MLE of the parameters of the model, in this case, the *Binomial* probability model?". This is a question that we ask every time we estimate a model. *R*, *Python*, and many other statistical packages make many estimation methods readily available and are extremely easy to implement. The downside of such ease of implementation, however, is that the assumptions of an underlying model being estimated using functions in *R* are often forgotten. In this course, we will keep asking you this type questions, and I hope that you will start forming the habit to this type of questions in each step of the data analysis and modeling, starting from EDA.

This lecture also spends a fair amount of time on confidenece interval estimation and covers a few different types of confidenece intervals. As this perhaps is a relatively more difficult portion of the lecture, you may need to spend more time studying this part of the readings and watching this part of the video a few times. During you study, ask "what are the pros and cons of each of the C.I. estimation method?"

In hypothesis testing, the likelihood ratio test, which you may not come across in your previous study of statistics, is used extensively. This is a very useful technique and will be used throughout this course.

Once the foundation of studying a single binary variable is formed, this lecture proceeds to two binary variables. It is in this section where a number of very important concepts are introduced, including contingency table (and its related estimation and inference concepts and techniques), relative risks, odds, and odds ratios. Contingenc table is useful for characterizing categorical variables in the exploratory data analysis (EDA) step, which we emphasize a lot in this course, and the notion of relative risks, odds, and odds ratios will be used in the interpretation of binary, multinomial, and ordinal logistic regression models.