Live Session - Week 1: Discrete Response Model - Lecture 1

Devesh Tiwari

January 10, 2017

Agenda

1. Introdution (30 minutes, depending on the number of students attending the sessions)
2. An overivew of topics covere in this lecture
3. Discussion of the analysis of two binary variables

1. Introduction (30 minutes, depending on the number of students attending the sessions)

1. Instructor's self introduction
2. Students' self introduction: each student takes turn introducing himself/herself (3 minutes each), addressing the questions below
   - Did you take the new or old version of *w203*?
   - What is your cohort?
   - What company are you working for, and what's your role?
   - Do you use machine learning or statistic modeling in your current work? If so, what techniques do you use?
   - Why do you take this course?
3. Course Overview, Other Reminders, Q&A

2. Topics covered in this lecture

- 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
- Two Binary variables
  - Contingency table
  - MLE
  - C.I.s for the difference of two probabilities

- Relative Risks
- Odds
- Odds ratios (OR)
- log(OR)
- Estimation and inference

Understanding and exploring random variables with two possible outcomes

Motivation:

We want to answer the following-type of questions:

1. Does the vaccine "help"" to prevent a specific disease (assuming an experiment was conducted and done correctly? (Does the vaccine group vs the placebo group have different exposure to the disease?)
2. Does the job training affect productivity?
3. Does the newly introduced tools reduce the number of person-hours needed?
4. Does the exercise group 1 have reduce weight more than the exercise group 2? ... the list goes on.

Review: Random variables

Question: What is a random variable? What are the different components of a random variable and why are we interested in them?

Random variables have a probability distribution function. This function "maps" given values of a random variable to the relative likelihood we observe it. Perhaps the most famous random variable is the normal, or gaussian, distribution which takes the form of a "bell curve." That bell curve is described mathematically as well, in the form of a function where the mean and variance are parameters we are interested in estimating.

Discrete random variables also have a PDF. Consider the following Bernoulli distribution where the probability of observing a 1 (or a success) is π and the probability of observing a zero (or a failure) is $1 - \pi$.

$$P(Y_j = y) = \pi_j^y (1 - \pi_j)^{1-y}$$

We are interested in estimating the parameter, π, which denotes the probability of a trial resulting in a success. One was to do this is to use maximum likelihood estimation.

MLE, confidence intervals, and tests

Discussion questions: Think about the following questions in groups or alone

1. What is the Wald CI or Wald standard error? Does the formula look familiar? Why does the book caution us against using the Wald interval?
2. What does it mean for a confidence interval to be too liberal or too conservative?
3. What is the Agresti-Caffalo confidence interval? How is it calculate and to what extent does it address the shortcomings of the Wald interval?
4. What is the likelihood ratio test and the LR - interval?

Take home exercise: Interpreting outcomes: Relative risks and odds ratio

In 2009, political scientists tested whether Latino voters were more likely to vote in a special election if they received mail that encouraged them to vote. In this study, 3.13% of voters in the control group voted and 3.78% of voters in one of the treatment group voted. How would you report this outcome? Would your interpretation of the intervention's efficacy change if the control group's turnout rate were 40% and if the treatment group's turnout rate were 40.65%?