Live Session 1 - Discrete Response Models

Jeffrey Yau

September 5, 2017

Agenda

1. Introdution (30 minutes, depending on the number of students attending the sessions)
2. A Discussion of weekly workflow and tips for success in this course (20 minutes)
3. An overivew of topics covere in this lecture (5 minutes)
4. Discussion of the analysis of two binary variables (35 minutes)

---

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*?
   - Which 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

2a. Quick Introduction for this course (5 minutes)

- Professor Yau to give an overivew of this course, addressing various perspectives though this course primarily focuses on statistical models for various types of response variales:
  1. the traditional statistical perspective
  2. the modern statistical perspective
  3. machine learning perspective
- This course focuses on statistical model building from the modern perspective used in data science, covering steps from defining a (business, policy, scientific, etc) problem that can be addressed using data at hand, conducting EDA as a pre-model-building step, understading the underlying statistical assumptions of a model under consideration, specifying a model, engineering features, estimating a model using functions in R, testing hypothesis, evaluating a model, conducting model diagnostics, and testing model assumptions. On occasion basis, I will also compare models covered in this course with machine learning techniques that can be used to solve similar problems.

- **The design of this course:** I designed this course to study each model as mathematically rigorous as practically possible. Each model is introduced with mathematical formulation, assumptions examined, estimation (in R) illustrated, statistical inference extensively studied, and practical topics, such as feature engineering (aka variable creation and transformation as termed in the statistical literature), model performance measurement, model interpretation, and model selection from data science perspective, extensively discussed.

based on multiple pillars - async lectures, assigned readings, live sessions, and exercises, of learning the materials. Each of these pillars are tightly integrated, with a lot of repetitions among different pillars. The main reason of this design is that the materials covered in this course are conceptually difficult, and the various forms of learning (i.e. from watching asyn lectures - passive learning - to working through difficult analytic and empirical problems - active learning) can maximize the chance of mastering this materials. Different people absorb materials differently, but regardless of how

2b. Weekly Workflow (5 minutes)

A typical week of the course proceeds as follows:

```
- Before live session: Watch all async content, study the assigned readings, attempt some of the end-of

- In live session: Please come to the live session prepared. **Live sessions are not lectures.** We wil

- After live session: Review the materials covered in the previous week and continue to attempt some of

- During the live sessions, I will also share some of my professional experience ("stories") in statist
```

2c. Professors' Expectation and How to Succeed in this Class (10 minutes)

Please take out the syllabus and review some of the highlights together

Here are some strategies based on our past experience and how I designed the course

- Review materials taught in the new version of w203, especially the part on linear regression modeling.
- **On watching the async lecture:** Different people learn differently; as such, it is difficult to give a "general" rule of how one should watch the async lectures. That said, I think it is very rare that students can watch each lecture in one sitting and master the materials. It is more likely that you will have to watch a section; pause; read the corresponding sections in the text and work through some examples; and then rewatch the section of the lecture to ensure that you understand the materials. This is why I followed so closely the assigned textbooks, which I chose by considering many different texts. This is also why mastering the concepts and methods in this course is very time consuming.
- Study your readings. More importantly, do the exercises in the texts. This is an advanced course; it is very difficult, if not impossible, to learn statistical modeling by simply reading textbooks or watching videos; doing hands-on work is critical to learn the materials taught in this course (and any statistical modeling and machine learning course). Note that it is also impossible to cover every single concept in a ninety-minute async lecture; you will have to build upon the materials taught in the async lecture by studying the readings and doing many exercises.
- Come to the live session prepared. I expect that you should watch the async video lecture, study the assigned readings, work on the examples in the book, and attempt a at least a few end-of-chapter exercises. In the live session, we may occasionally ask questions that you can only answer if you keep up with the readings.

- Form study groups! Do it now (if you haven't done so already)!
- Do not skip the readings or async lectures. In this course, once you fall behind, it is extremely difficult to catch up, as we cover a different statistical model almost every week.
- For the lab, do not use the **"divide-and-conquer"** strategy if you work in a group, although you are encouraged to discuss your work (after you finish them) with your group.
  - Attend office hours to ask questions related to the concepts and techniques covered in async lectures and readings; don't just come to office hours when you have questions on the labs. **Note that for the labs, we will only answer clarifying questions in the office hours.**
  - Labs: After each of the lab due date, we will distribute the solution; as such no late work will be accepted. We will also return the graded lab to you within two weeks or so with general comments for good and not-so-good practices observed in all the labs. Please talk to me or Professor Tiwari during office hours if you have questions about the garded lab **after you carefully review our solutions**.

3. 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
- This week's lecture starts with the simplest case, the Binomial probability model, covering both parameter estimation and statistical inference.
- A lot of time is spent on discussing the confidence interval of this model, a discussion typically not covered in the elementary statistics courses.
- As you will see, we use simulation extensively in this course, as simulation is one of the best ways to "get a feel" of a probability model and gain an understanding of its behavior
- A single binomial probability model is then extended two two binomial probability model, introducing the notion of contingency table. Also introduced is the formulation of a likelihood function and the method of maximum likelihood, a version of which was taught in the parameter estimation module in w203.
- With more than one binomial random variables, the concept of relative risk, odds, and odds ratio become very powerful, as the meaning of the difference between probability of success, $\pi_1 - \pi_2$, is a function of the magnitude of π_1 and π_2.
- All of these concepts are introduced as the preparation for the study of the regression model of categorical response data, with binary response being the simplest case.

In the live session today, I want to focus on (1) some of the derivations that are missing in the book and MLE, (2) confidence intervals, and (3) the concepts of relative risks and odds ratios. I will leave a few take-home exercises as well.

4. A Summary of the Confidence Intervals for the Probability of Success π

Recall that in w203, we've learned that the *typical form* of a confidence interval for a parameter of a probability model is:

estimator ± (distributional value) × (standard deviation of estimator)

Before we begin, what is the interpretation of a (1-α)100% confidence interval (as we learned in w203)?

Is it okay to say that the estimated confidence interval has (1-α)100% probability of containing the ture parameter, calling it θ? If so, please explain? If not, why not?

In fact, the Wald confidence interavl takes this form:

1. Wald Confidence Interval:

$$\hat{\pi} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

What assumption is required for one to construct the Wald confidence interval?

What are the problems with this confidence interval?

Why does a confidence interval method not actually achieve its stated confidence level of binomial random variable?

There has been a lot of research on finding an interval for π, the earliest of these studies began in the early 20^{th} century. However, the mathematical details go outside of the scope of this course. Interested readers can refer to the papers referenced in Chapter 1 of the book. Here, we will only summarize some of the findings from this literature. In practice, these intervals can be computed using functions from various R libraries. Even in that case, it is instructive to construt these intervals "manually", for not all statistical or programming languages come with these readily-available functions.

Alternatives: For $n < 40$, use *Wilson interval.* For $n \geq 40$, use *Agresti-Coull interval.* Note that even for $n < 40$, the *Agresti-Coull interval* is still generally better than the *Wald interval.*

2. Wilson "score" Interval:

$$\hat{\pi} \pm \frac{Z_{1-\frac{\alpha}{2}} n^{1/2}}{n + Z^2_{1-\frac{\alpha}{2}}} \sqrt{\hat{\pi}(1-\hat{\pi}) + \frac{Z^2_{1-\frac{\alpha}{2}}}{4n}}$$

where $\pi = \frac{w + Z^2_{1-\frac{\alpha}{2}}/2}{n + Z^2_{1-\frac{\alpha}{2}}}$

3. Agresti-Coull interval:

$$\pi \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n + Z^2_{1-\frac{\alpha}{2}}}}$$

where $\pi = \frac{w+2}{n+4}$

Steps to build a contingency table

```r
# This simulation gives the number of successes, given the number
# of independent binary trials and the (theoretical) probability
# of success

set.seed(2017)

pi1 <- 0.2
pi2 <- 0.4
n1 <- 100
n2 <- 100

w1 <- rbinom(n = 1, size = n1, prob = pi1)
w2 <- rbinom(n = 1, size = n2, prob = pi2)
# w1 w2

c.table <- array(data = c(w1, w2, n1 - w1, n2 - w2), dim = c(2, 2),
    dimnames = list(Group = c("Group 1", "Group 2"), Event = c("True",
        "False")))
c.table
```

```
##          Event
## Group     True False
##   Group 1   26    74
##   Group 2   40    60
```

```r
rowSums(c.table)  #Row sum for Group 1 and Group 2: n1 and n2
```

```
## Group 1 Group 2
##     100     100
```

```r
pi.hat.table <- c.table/rowSums(c.table)
round(pi.hat.table, 2)
```

```
##          Event
## Group     True False
##   Group 1 0.26  0.74
##   Group 2 0.40  0.60
```

```r
alpha<-0.05
str(pi.hat.table) #recall the structure of the array
```

```
##  num [1:2, 1:2] 0.26 0.4 0.74 0.6
##  - attr(*, "dimnames")=List of 2
##   ..$ Group: chr [1:2] "Group 1" "Group 2"
##   ..$ Event: chr [1:2] "True" "False"
```

```r
pi.hat1<-pi.hat.table[1,1] #extract the estimated probability of success for group 1

pi.hat2<-pi.hat.table[2,1] #extract the estimated probability of success for group 2
```

Wald Interval

Take-home Exercise: Recall the the Wald interval suffers from the same problems discussed in the case for one binary random variable, but we demonstrate how it is calculated anyway. List all of the problems with Wald interval with one binomial random variable case.

```
# Wald
var.wald<-pi.hat1*(1-pi.hat1) / sum(c.table[1,]) + pi.hat2*(1-pi.hat2) / sum(c.table[2,])

pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald)
```

```
## [1] -0.2688816 -0.0111184
```

```
# Note: Each interval limit could be calculated one at a time as well, store it in a data.frame, and pr

lower<-pi.hat1 - pi.hat2 - qnorm(p = 1-alpha/2) *
    sqrt(pi.hat1*(1-pi.hat1) / sum(c.table[1,]) +
    pi.hat2*(1-pi.hat2) / sum(c.table[2,]))

upper<-pi.hat1 - pi.hat2 + qnorm(p = 1-alpha/2) *
    sqrt(pi.hat1*(1-pi.hat1) / sum(c.table[1,]) +
    pi.hat2*(1-pi.hat2) / sum(c.table[2,]))

data.frame(lower, upper)
```

```
##        lower      upper
## 1 -0.2688816 -0.0111184
```

Agresti-Caffo Interval - Agresti and Caffo (2000), based on their examination of various types of CIs, recommended that adding one success and one failure for each group results in an interval that does a reasonable job.

```
# Agresti-Caffo
pi.tilde1<-(c.table[1,1]+1)/(sum(c.table[1,])+2) #calculate the adjusted prob of success

pi.tilde2<-(c.table[2,1]+1)/(sum(c.table[2,])+2) #calculate the adjusted prob of success

var.AC<-pi.tilde1*(1-pi.tilde1) / (sum(c.table[1,])+2) + pi.tilde2*(1-pi.tilde2) / (sum(c.table[2,])+2)

pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.AC)
```

```
## [1] -0.265253469 -0.009256335
```

5. The notions of Relative Risks and Odd Ratios

Relativce Risks

- The problem with basing inference on $\pi_1 - \pi_2$ is that it measures a quantity whose meaning changes with the sizes of π_1 and π_2.

Discussion: Come up with an example to illustrate the point made above and explain the meaning of the relative risks in the following code.

Continue with the example above, we can actually estimate the relative risks, which is based on the MLE estimate of the probability of success in each group, using MLE by appealing to the *invariance* property of MLE (see appendix B of Bilder and Loughin's book for more detail.)

Take-home Exercise: Below include the codes for the MLE estimate of RR and its associated Wald confidence interval. (1) Run the code. (2) Explain each line of the code and add a comment. (3) Interpret the relative risks. (4) Interpret the confidence interval.

```r
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

## Estimate the Relative Ratios
#round(pi.hat1/pi.hat2, 4)

## COMMENT TO BE HERE
#var.log.rr <- (1-pi.hat1)/(n1*pi.hat1) + (1-pi.hat2)/(n1*pi.hat2)

## COMMENT TO BE HERE
#ci <- exp(log(pi.hat1/pi.hat2)) + qnorm(p = c(alpha/2, 1-alpha/2)*sqrt(var.log.rr))

## COMMENT TO BE HERE
#round(ci,4)

pi3 <- .1
pi4 <- .9

pi4/pi3
```

```
## [1] 9
```

6. Odds Ratios

Odds: probability of a success divided by the probability of a failure: $\frac{\pi}{1-\pi}$

Exercise: Suppose $\pi = 0.1$. (1) What are the corresponding odds? (2) Interpret it in the following two types of statements

a. The odds of success are X. (Fill in X)
b. The probability of failure is X times the probability of success. (Fill in X)

The notion of odds ratios comes in when there are more than one groups and we to calculate (or estimate) the odds separately in each group and then compare them.

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

where the corresponding MLE $\hat{\pi}_j = \frac{w_j}{n_j}$.

Exercise: (1) Do OR have a upper or lower limt? If so, what are they? (2) Estimate the confidence interval for the odd ratio? (3) Interpret the confidence interval.

Appendix:

An Example of Two Binary Variables

Motivation:

Studying binomial probability model allows us 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 new marketing campagin increase sales?

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.

```
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

```
pi1 <- 0.2
pi2 <- 0.4
n1 <- 100
n2 <- 100

set.seed(2017)
```

Suppose you are given these parameters. Simulate two independent Binomial random variables, and $w1$ and $w2$, the number of successes (or event being happened) of group 1 and group 2 are given below.

Recall that the probability mass function of the Binomial random variable is

$$P(W_j = w_j) = \binom{n_j}{w_j} \pi_j^{w_j} (1 - \pi_j)^{n_j - w_j}$$

where $w_j = 0, 1, \ldots, n_j$ where $j = 1, 2$

```
## [1] 26
```

```
## [1] 40
```

Steps to build a contingency table

```
c.table <- array(data = c(w1, w2, n1 - w1, n2 - w2), dim = c(2, 2),
    dimnames = list(Group = c("Group 1", "Group 2"), Event = c("True",
        "False")))
c.table
```

```
##          Event
## Group     True False
##   Group 1   26    74
##   Group 2   40    60
```

```
rowSums(c.table)  #Row sum for Group 1 and Group 2: n1 and n2
```

```
## Group 1 Group 2
##     100     100
```

```
pi.hat.table <- c.table/rowSums(c.table)
round(pi.hat.table, 2)
```

```
##          Event
## Group     True False
##   Group 1 0.26  0.74
##   Group 2 0.40  0.60
```

Take-home Exercise: Let's spend some time here to think about the numbers in the contingency table and assumptions used throughout this construction.

Take-home exercise: 1. Use *prop.test()* function to test the difference in the probabilities. What is the result? Is the result same as what we compute above? What confidence interval does it use? Please read the R documentation when answering this questions. 2. Install the *PropCIs* package and use its *wald2ci()* functions to estimate both the Wald and Agresti-Caffo confidence intervals. linked phrase 3. Read page 31 - 33 (on True confidence levels for the Wald and Agresti-Caffo intervals) and try the code in the text.

Using Probability, Cumulative Probability, Quantile Functions, and Simulation:

Let's start with an example:

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution: Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5=0.2. We can find the probability of having exactly 4 correct answers by random attempts as follows.

```
dbinom(4, size = 12, prob = 0.2)
```

```
## [1] 0.1328756
```

```
paste("The probability of having exactly 4 correct answers is",
    100 * round(dbinom(4, size = 12, prob = 0.2), 2), "%")
```

```
## [1] "The probability of having exactly 4 correct answers is 13 %"
```

To find the probability of having four or less correct answers by random attempts, we apply the function dbinom with x = 0,…,4.

```
round(dbinom(0, size = 12, prob = 0.2) + dbinom(1, size = 12,
    prob = 0.2) + dbinom(2, size = 12, prob = 0.2) + dbinom(3,
    size = 12, prob = 0.2) + dbinom(4, size = 12, prob = 0.2),
    2)
```

```
## [1] 0.93
```

```
# OR, use CPF
round(pbinom(4, size = 12, prob = 0.2), 2)
```

```
## [1] 0.93
```