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

Professor Jeffrey Yau and Professor 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

An Example of Two Binary Variables

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.

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

```
pi1 <- 0.2
pi2 <- 0.4
n1 <- 100
n2 <- 200

set.seed(2017)
```

Suppose you are given these parameters, we 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] 69
```

Steps to build a contingency table

"'{r, tidy=TRUE, tidy.opts=list(width.cutoff=65)}

```
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   69   131
```

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

```
## Group 1 Group 2
##     100     200
```

```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 0.34  0.66
```

Exercise: Let's spend some time here to discuss the numbers in the contingency table and assumptions used throughout this construction.

Confidence interval for the difference of two probabilities

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

```
##  num [1:2, 1:2] 0.26 0.345 0.74 0.655
##  - 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

Exercise (Breatout Session - 5 minutes): 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.

```r
# 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] -0.19331129  0.02331129
```

```r
# Note: Each interval limit could be calculated one at a time
# as well, store it in a data.frame, and produce a nicer
# output
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 -0.1933113 0.02331129
```

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] -0.18970221  0.02604467
```

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.

Relative 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.

Exercise (Breakout session - 5 minutes): 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.)

Exercise (Breakout session: 8 minutes): 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.

```
## COMMENT TO BE HERE
round(pi.hat1/pi.hat2, 4)
```

```
## [1] 0.7536
```

```
## 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)
```

```
## [1] -1.7925 -0.0446
```

Odds Ratios

If we run out of time, please study page 40 - 43 of the text.

Odds: probability of a success to probability of a failure: $\frac{\pi}{1-\pi}$

Exercise (1 minute): 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 A: Confidence Intervals for the Probability of Success π

Recall the typical form of a confidence interval for aparameter:

estimator $\pm$ (distributional value) $\times$ (standard deviation of estimator)

Different C.I.s for π 1. Wald Confidence:

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

Note that when $\hat{\pi}$ is close to 0 or 1, two problems arise: a. Calculated limits may be less than 0 or greater than 1, which is outside the boundaries for a probability. b. When $\hat{\pi} = 0$ or 1, then $\frac{\hat{\pi}(1-\hat{\pi})}{=} 0$ for $n > 0$. This leads to the lower and upper limits to be exactly the same.

Problems with the Wald Confidence Interval: 1) The Wald Confidence Interval works well only if the sample size is large. 2) The discreteness of the binomial distribution often makes the normal approximation work poorly even with large samples.

The result is a confidence interval that is often too "liberal". This means when *95%* is stated as the confidence level, the true confidence level is often lower.

On the other hand, there are "conservative" intervals. These intervals have a true confidence level larger than the stated level.

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

Appendix B: 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
```