

W271 Live Session 1 - Discrete Response Models

Professor Jeffrey Yau

Agenda

| | Estimated Time | Topics |
|---|----------------|---|
| 1 | 5 minutes | Welcome to the course! |
| 2 | 30 minutes | Review several important concepts covered in the first week |
| 3 | 15 minutes | Introduction to the course and course logistics |
| 4 | 30 minutes | Getting to know each other |
| 5 | 10 minutes | Q&A |

Readings

BL2015: Ch. 1 (Skip Sections 1.2.6 and 1.2.7), Appendix B.3 and B.5

Note: For the first five lectures of the course we follow the book very closely to allow for multiple touch points - async video lectures, readings, live session examples, and homework on the same concepts and techniques.

Week 1 Topics

- 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)
- $\log(\text{OR})$
- Estimation and inference

Summary of This Week's Materials

Week 1 of this course begins with the simplest case of discrete variable modeling, starting with the simplest case of Bernoulli and Binomial probability models, covering both parameter estimation and statistical inference. It points out the importance of proposing the appropriate mathematical framework to model the variable of interest based on the “type” of the variable. In this case, the “type” of the (random) variables of interest is “discrete”, rendering continuous probability model inappropriate.

This is the simplest case of modeling categorical variables, as there are only two categories. Therefore, we do not have to worry about whether or not the categories are ordered (or ranked) or count (such as the number of customers arriving at a particular store within a specific hour).

You should have already seen both the Bernoulli and Binomial probability models in W203 and perhaps other statistics courses elsewhere. However, many aspects of this model are typically not obvious and the underlying assumptions are often overlooked (note that all probability and statistical models come with assumptions) in this earlier courses. In this course, we will study how to apply theoretical statistical models in practice.

The Binomial probability model is a reasonable model for distribution of an event happened (or “success”) in a given number of trials, so long as the assumptions listed on page 3 and 4 of our textbook “*Analysis of Categorical Data with R*” are not violated.

The book lists several examples and discusses in each cases whether each of the assumptions is reasonable

Another (perhaps) new aspect of studying Bernoulli and Binomial probability models is to model it in R using the R built-in `dbinom(n,size,prob)` function

We also study how to simulate a probability model and use simulation to evaluate how well procedures perform when these assumptions are violated. (Pages 7 and 8 of the book describe this)

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 to two binomial probability model, introducing the notion of a contingency table. Also introduced is the formulation of a likelihood function and the method of maximum likelihood, a version of which should be covered in w203.

With more than one binomial random variables, the concepts of relative risk, odds, and odds ratios become very powerful, in fact, essential, 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 to prepare our 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) the statistical analysis workflow, (2) confidence intervals, and (3) the concepts of odds ratios.

Statistical Analysis Workflow Revisit

- Postulate a statistical model that conforms with the underlying (business, policy, scientific, etc.) question being asked
- Estimate the parameter of the statistical model
- Check model assumptions
- Conduct statistical inference

In the case of the success parameter π from the Bernoulli distribution, we use MLE:

$$\begin{aligned} L(\pi|y_1 \cdots y_n) &= P(Y_1 = y_1, \dots, Y_n = y_n) \\ &= P(Y_1 = y_1) \times \cdots \times P(Y_n = y_n) \\ &= \prod_{i=1}^n P(Y_i = y_i) \\ &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)} \end{aligned}$$

The MLE is $\hat{\pi} = \frac{w}{n}$, where w denotes the number of successes in n trials.

The variance of the estimate is $\hat{Var}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n}$

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, θ , is:

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

What is the frequentist interpretation of a $(1-\alpha)100\%$ confidence interval (as we learned in w203)?

In fact, if we assume that \hat{p}_i follow the Normal Distribution, that is, $\hat{\pi} \sim N(\pi, \hat{Var}(\hat{\pi}))$, we have the **Wald confidence interval** taking the following 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?

While one of MLE's properties is that it is asymptotically normal, an implicit assumption to construct this confidence interval is that the normal approximation is a good approximation.

What are the problems with this confidence interval?

Regardless of whether or not the above assumption is satisfied, this confidence interval suffers from some problems:

1. The interval can exceed the (0,1) interval, which would not be sensible for probability.

2. When $w = 0$ or $w = 1$, the confidence interval degenerates into a single point equal to $\hat{\pi}$.
3. The coverage is not necessarily equal to the stated level $1 - \alpha$.

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

Example: Computation of Wald Confidence Interval

Let's pause for a minute before writing code.

Suppose we have a Bernoulli probability model with $\pi = 0.6$. In addition, suppose that we “think” that the sample of observations (that is, our data) come from this (theoretical) distribution.

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

```
set.seed(23951)
n_sim = 1
n_trials = 10
p = 0.6
bin <- rbinom(n = n_sim, size = n_trials, prob = p)
bin
```

```
## [1] 6
```

From the simulation, we get a total of 6 successes from 10 draws. Our hypothetical value of the true probability of success is $\pi = 0.6$. Let's use this information to derive the Wald confidence interval.

Recall the Wald Confidence Interval's formula before writing down the code.

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

$w = 6$ and $n = 10$

```
w = 6
n = 10
alpha = 0.05

pi.hat = w/n # Recall that this is an MLE for pi in Bernoulli probability model
var.wald = pi.hat * (1 - pi.hat)/n

wald.CI_lower.bound = pi.hat - qnorm(p = 1 - alpha/2) * sqrt(var.wald)
wald.CI_upper.bound = pi.hat + qnorm(p = 1 - alpha/2) * sqrt(var.wald)

round(data.frame(pi.hat, var.wald, wald.CI_lower.bound, wald.CI_upper.bound),
      4)
```

```
##   pi.hat var.wald wald.CI_lower.bound wald.CI_upper.bound
## 1    0.6   0.024          0.2964          0.9036
```

Is this a “good” confidence interval or not? What do we mean by “good”? Does it achieve its stated confidence level?

- We have never really questioned the validity of confidence interval in elementary statistics courses as well as w203? Why now?
- As it turns out, Wald confidence interval does not perform well when it comes to achieving the true confidence level, say $95\% = 1 - \alpha$ where $\alpha = 0.05$. Let's examine the reason behind it.
- Binomial random variables are discrete random variables, meaning that it can only take a finite number of values; given the number of trials n , there are only $n + 1$ possible intervals, corresponding to $w \in 0, 1, \dots, n$.
- Given the true parameter value π , some of these $n + 1$ intervals contain π , obviously, and some don't.
- More formally, the true confidence level at a given parameter value π , called it $C(\pi)$, following the book's notation, is the sum of the binomial probabilities for all the $n + 1$ intervals that actually contain π :

$$C(\pi) = \sum_{i=1}^n I(w) \binom{n}{w} \pi^w (1 - \pi)^{n-w}$$

where $I(w) = 1$ if the interval contains π and 0 otherwise.

Let's do some computation to solidify our intuition. Let's continue to use the numbers from the example above: $n = 10$ and $\pi = 0.6$

```
pi = 0.6
alpha = 0.05
n = 10
w = 0:n

wald.CI.true.coverage = function(pi, alpha = 0.05, n) {

  w = 0:n

  pi.hat = w/n
  pmf = dbinom(x = w, size = n, prob = pi)

  var.wald = pi.hat * (1 - pi.hat)/n
  wald.CI_lower.bound = pi.hat - qnorm(p = 1 - alpha/2) * sqrt(var.wald)
  wald.CI_upper.bound = pi.hat + qnorm(p = 1 - alpha/2) * sqrt(var.wald)

  covered.pi = ifelse(test = pi > wald.CI_lower.bound, yes = ifelse(test = pi <
    wald.CI_upper.bound, yes = 1, no = 0), no = 0)

  wald.CI.true.coverage = sum(covered.pi * pmf)

  wald.df = data.frame(w, pi.hat, round(data.frame(pmf, wald.CI_lower.bound,
    wald.CI_upper.bound), 4), covered.pi)

  return(wald.df)
}

wald.df = wald.CI.true.coverage(pi = 0.6, alpha = 0.05, n = 10)
wald.CI.true.coverage.level = sum(wald.df$covered.pi * wald.df$pmf)

# Let's compute the true coverage for a sequence of pi
```

```

pi.seq = seq(0.01, 0.99, by = 0.01)
wald.CI.true.matrix = matrix(data = NA, nrow = length(pi.seq),
  ncol = 2)
counter = 1
for (pi in pi.seq) {
  wald.df2 = wald.CI.true.coverage(pi = pi, alpha = 0.05, n = 10)
  # print(paste('True Coverage is',
  # sum(wald.df2$covered.pi*wald.df2$pmf)))
  wald.CI.true.matrix[counter, ] = c(pi, sum(wald.df2$covered.pi *
    wald.df2$pmf))
  counter = counter + 1
}
str(wald.CI.true.matrix)

## num [1:99, 1:2] 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 ...

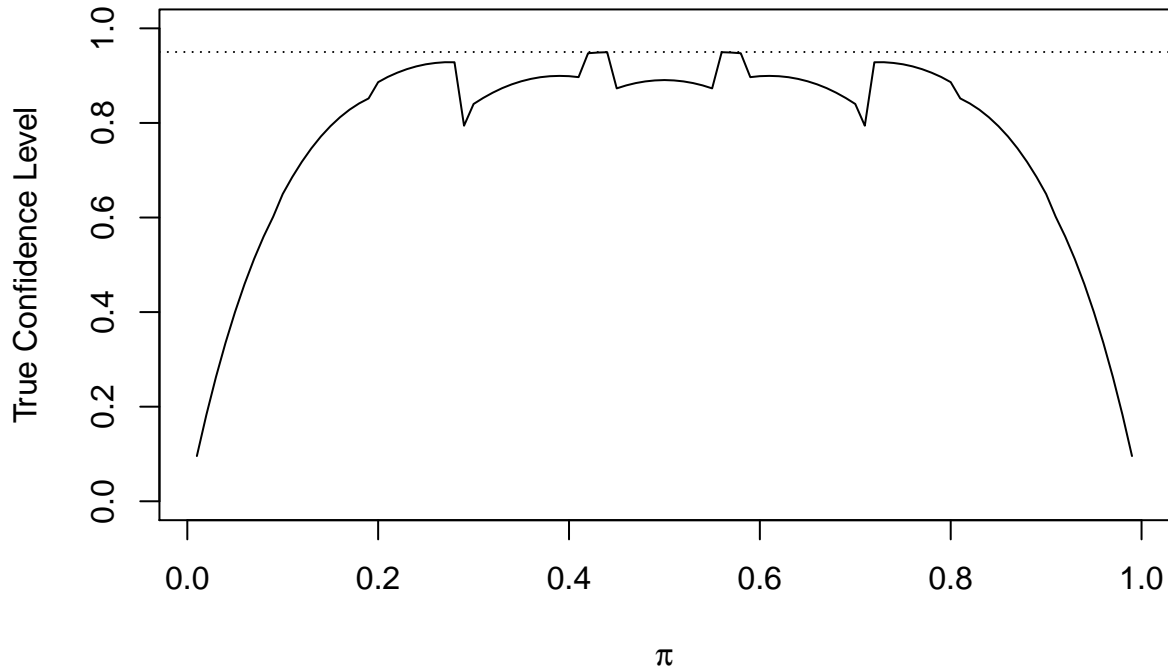
wald.CI.true.matrix[1:5, ]

##      [,1] [,2]
## [1,] 0.01 0.0956
## [2,] 0.02 0.1828
## [3,] 0.03 0.2624
## [4,] 0.04 0.3347
## [5,] 0.05 0.4002

# Plot the true coverage level (for given n and alpha)
plot(x = wald.CI.true.matrix[, 1], y = wald.CI.true.matrix[,
  2], ylim = c(0, 1), main = "Wald C.I. True Confidence Level Coverage",
  xlab = expression(pi), ylab = "True Confidence Level", type = "l")
abline(h = 1 - alpha, lty = "dotted")

```

Wald C.I. True Confidence Level Coverage



There has been a lot of research on finding a confidential interval for π , the earliest of these studies began in the early 20th 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”, since 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:

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

where $\tilde{\pi} = \frac{w + \frac{1}{2} Z_{1-\frac{\alpha}{2}}^2}{n + Z_{1-\frac{\alpha}{2}}^2}$, which can be considered as an “adjusted” estimate of π .

One of the advantages of this interval is that it is bounded between 0 and 1.

3. Agresti-Coull interval:

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

- 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.

Introduction to the course and course logistics

- We will walk through the course syllabus, which includes many details related to this course

Learning Objectives

After successfully completing this course, students will be able to

1. Apply the rigorous process to apply statistical techniques to model categorical response, time series, and panel data.
2. Describe the mathematical formulation and underlying statistical assumptions of some of the most popular statistical models used in the industry to model categorical response, time series, and panel data.
3. Leverage exploratory data analysis (EDA) in the statistical modeling process to generate insights for subsequent statistical modeling as applied to solving data science problems that are often given as (vague) business or policy problems
4. Have ample opportunities to work on a wide range of examples with real-world and simulated data and be able to leverage these examples for other real-world applications
5. Choose from a set of statistical techniques for a given question and to make trade-offs between model complexity, ease of result and model interpretation, and implementation complexity in real-world applications
6. Understand the design principles of solutions and theoretical foundations of the statistical models that make up the solutions are the major focus, as they are essential for data science practitioners.
7. Formulate, choose, apply, implement, evaluate, and test statistical models to capture key patterns exhibited in data.
8. Develop the foundation for learning other, perhaps more advanced, statistical techniques to model categorical response, time series, and panel data.

Getting to know each other

1. Which cohort are you from?
2. What courses have you taken before w271?
3. What company do you work for?
4. What's your role in that company?
5. Why take this course as one of your elective?