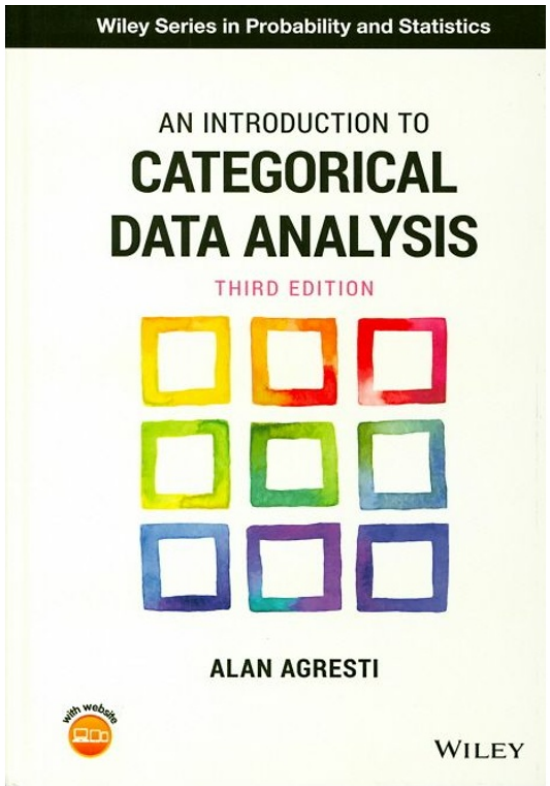
STS3016: Categorical Data Analysis

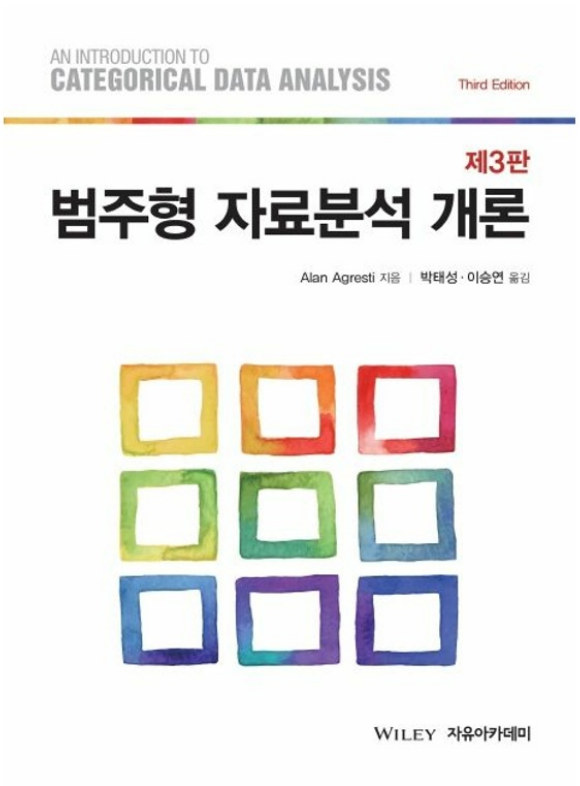
1. Introduction

Fall 2025

Syllabus

- mcho@inha.ac.kr (office hour) / 5N441
- *An Introduction to Categorical Data Analysis*, Agresti Alan (2018), 3rd edition, Wiley.





Syllabus

- Prerequisites : statistical inference, linear regression, matrix algebra & R
- Midterm exam (35%), Final exam (40%), Homework (20%), and Attendance (5%)

- **Class**

  Review : 10 minutes

  Lecture, Exercise & Lab

  Summary & Preview : 5 minutes

  Solve homework assignment questions (after submission)

Categorical Data Analysis

Frequentist : θ fixed
Bayesian : θ random $\sim$ prob. dist.

Statistics

① summary of data, EDA

② Inference → model

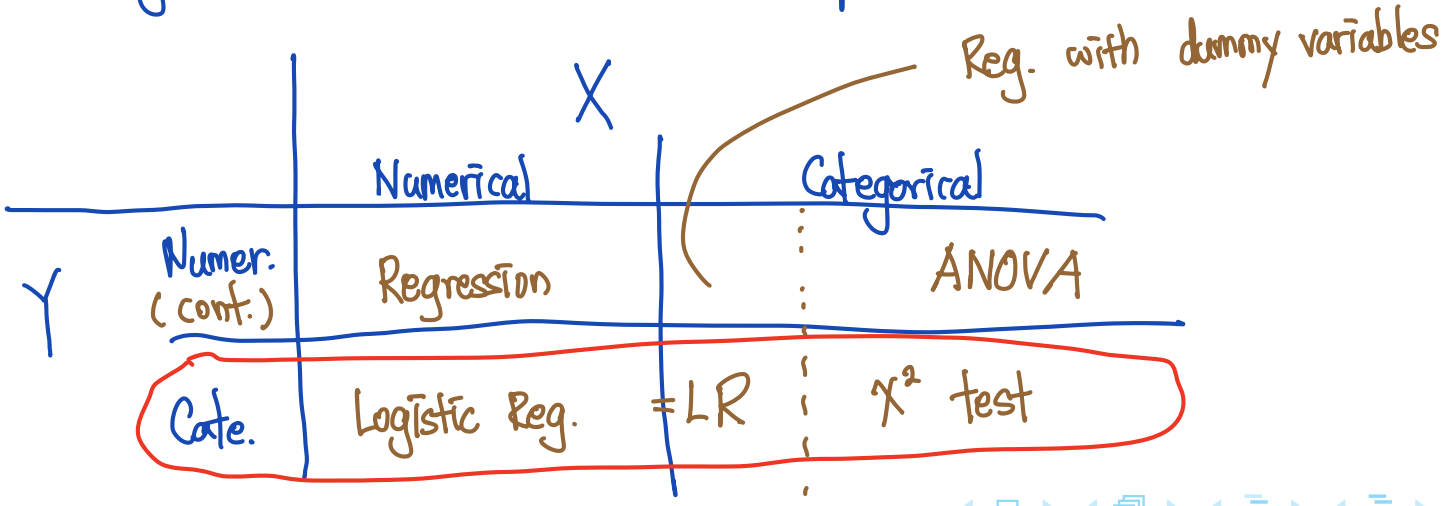
"parameter" θ, e.g., p, β

③ Prediction (Classification)

(parametric methods)

→ probability distribution, likelihood

- Relationships / Associations among variables (correlation, independence)



Tentative Course Schedule

Y

1. Introduction – categorical response data, probability distributions, statistical inference for "p"
   Tests
   Binomial, Multinomial, ...

2. Analyzing Contingency Tables – probability structure, odds ratio, or Chi-squared tests
   to see associations (or independence)
   using proportions

3. Generalized Linear Models (GLMs) – 3 components, data type
   ① $Y \approx$ ③ g ② $(X\beta)$
   $Y \sim$ Normal, Binomial, Poisson, NB, ...
   (Exponential Family)

4. Logistic Regression – statistical inference, predictors, effect, prediction,
   binary Y, g = logit vs. probit
   based on likelihood
   & model selection, ...

Tentative Course Schedule

5. Multicategory Logit Models – nominal / ordinal

6. Loglinear Models for Counts – contingency tables

7. Marginal Modeling – Generalized Estimating Equations (GEE)

8. Random Effects – Generalized Linear Mixed Models (GLMMs)

1. Categorical Response Data
2. Probability Distributions for Categorical Data
3. Statistical Inference for a Proportion
4. Statistical Inference for Discrete Data

Categorical Response Data

A **categorical** variable has a measurement scale consisting of

Quantitative vs. Qualitative

Response vs. Explanatory variables

Binary – Nominal – Ordinal scale distinction

Probability Distributions for Categorical Data

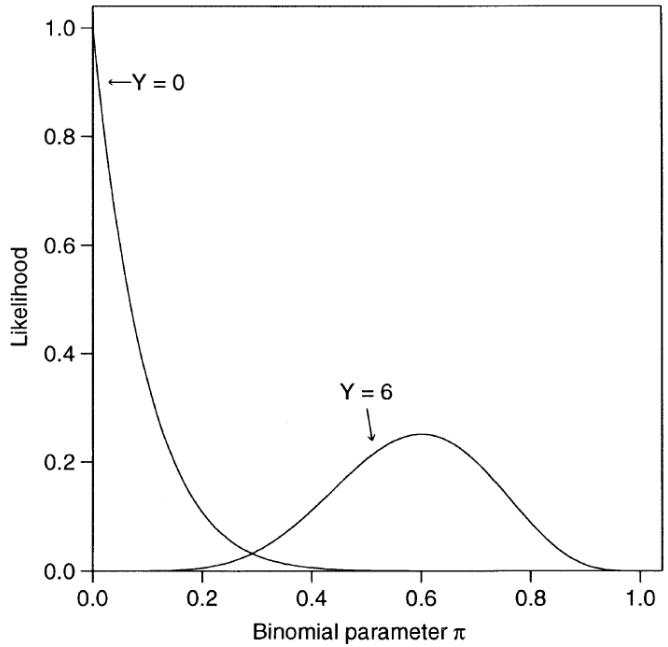
1.

e.g.

2.

Likelihood Function and MLE

Likelihood function:



Maximum likelihood estimator (MLE)

Significance Test About a Binomial Parameter

For binomial, the ML estimator for π

e.g.

Confidence Intervals for a Binomial Parameter

Wald, Likelihood-Ratio, and Score Tests

1. Wald test

2. Score test

3. Likelihood-ratio test

Example

Small-Sample Binomial Inference

Exercise 1.8

When asked to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said yes.

(a) Estimate the population proportion who would say yes. Construct and interpret a 99% confidence interval for this proportion.

(b) Conduct a significance test to determine whether a majority or minority of the population would say yes. Report and interpret the p-value.

Exercise 1.13

Consider $H_0 : \pi = 0.5$ and $H_1 : \pi \neq 0.5$. With $y = 0$ in $n = 25$ trials,

(a) Find l_0 and l_1, the maximized likelihood under H_0, H_1, respectively.

(b) Find the likelihood-ratio test statistic and report the p-value.

(c) Find the likelihood-ratio test statistic and p-value for testing $H_0 : \pi = 0.074$ and $H_1 : \pi \neq 0.074$.

Summary

1. Introduction to CDA
   - Categorical Response Data
   - Probability Distributions
   - Statistical Inference for a Proportion
   - Statistical Inference for Discrete Data

Next Class

2. Analyzing Contingency Tables

- Probability Structure for Contingency Tables
- Comparing Proportions in 2×2 Contingency Tables
- The Odds Ratio
- Chi-Squared Tests of Independence
- Testing Independence for Ordinal Variables
- Association in Three-Way Tables