ST3241 Categorical Data Analysis I

An Introduction

Some Objectives Of This Course

- Analyzing binomial and Poisson variables in real data.
- Visualizing and analyzing categorical data.
- How to use SAS and / or R for the above purposes.
- Provide hands-on practice with real life data.

Some Topics To Be Covered

- Introduction to Categorical Data
- Two-way Contingency Tables
- Three-way Contingency Tables
- Generalized Linear Models

Some Topics To Be Covered

- Logistic Regression
- Loglinear Models
- Multicategory Logit Models
- Models for Matched Pairs (if Time Permits)

Software To Be Used

- SAS:
  - Only for use in Statistics Computer Labs
  - You can't download them in your own PC.
- R:
  - This is a freeware and you can install it in your own PC
  - Website: http://www.r-project.org

Let's Begin With An Example

	Belief in Afterlife	
Gender	Yes	No or or Undecided
Females	435	147
Males	375	134

- 1091 people responded to a survey by their gender and their belief in an afterlife.
- Is there any association between gender and their belief in afterlife?

Another Example

- For the 23 space shuttle flights (Ft) that occurred before the Challenger mission disaster in 1986, the following table shows the temperature (Temp) at the time of flight and whether at least one primary O-ring suffered thermal distress (TD).

The Data

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

- Is there any association between Temperature and thermal distress?

What is Categorical Data

- A *categorical* variable is one for which the measurement scale consists of a set of categories
- Example
  - Political philosophy: liberal, moderate, conservative.
  - Choice of breakfast cereal: hot, cold, none.
  - Test for Alzheimer's disease: symptoms present, symptoms absent.
- One and only one category should be applied to each subject.

Where Can We Have Categorical Data

- Social sciences : opinions on issues
- Health sciences : response to treatments/drugs
- Behavioral sciences : e.g. diagnose mental illness
- Public health : AIDS awareness
- Zoology : animals food preferences
- Education : students' response to exams
- Marketing : consumer preferences
- Almost everywhere

Distinction in Categorical Data

- Ordinal variable
  - Categories are ordered
  - e.g. response to a treatment : excellent, good, fair, poor.
  - e.g. company's inventory level : too low, about right, too high.
- Nominal variable
  - Categories can not be ordered.
  - e.g. religious affiliation : Catholic, Jewish, Muslim, Hindu, Others
  - e.g. mode of transport to work : MRT, bus, taxi, car, others.

Notes

- For nominal variables, the order of listing is irrelevant, and the statistical analysis should not depend on that ordering.
- Methods designed for ordinal variables utilize category ordering and thus they can't be used for nominal variables.

Probability Distributions Involved

- Poisson Distribution
- Binomial Distribution

An Example

- AYE is heavily used by commercial as well as private vehicles. A group of researchers catalogue for the next year all accidents resulting in a fatality in a particular part of that road in order to study the rate of fatal automobile accidents.

Probability Model For The Study

- The Poisson distribution is a potential probability model for the number of fatal accidents in a given week.
- Let Y = the no. of fatal accidents in a week, then

$$P[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0, 1, 2, \cdots$$

where μ is the average no. of fatal accidents in a week

Notes

- A key feature of the Poisson distribution is that its variance increases as the mean does.
- The assumption of Poisson model is too simplistic though it produces useful results in a wide variety of categorical data analysis.

A Variation of The Example

- Suppose the researchers decided to count the number of fatal accidents in every N accidents.
- Estimate the probability of fatality of an accident.

Probability Model

- Let Y = no. of fatal accidents out of N accidents
- π= probability of fatality of an accident

$$P[Y = y] = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y}, y = 0, 1, \cdots, N$$

- Exercise: can you find the relation between these two models?

Notes

- Some experiments have more than two possible outcomes. For instance, one might summarize the outcome in each accident using the categories uninjured, injury not requiring hospitalization, injury requiring hospitalization, fatal.
- The probability distribution to be used is then multinomial.
- Standard procedures for categorical data analysis assume a Poisson, binomial or multinomial model.

Inference Problem

- The parameters of the probability models (e.g., μ and π) are unknown.
- Use observed data to estimate the parameters.

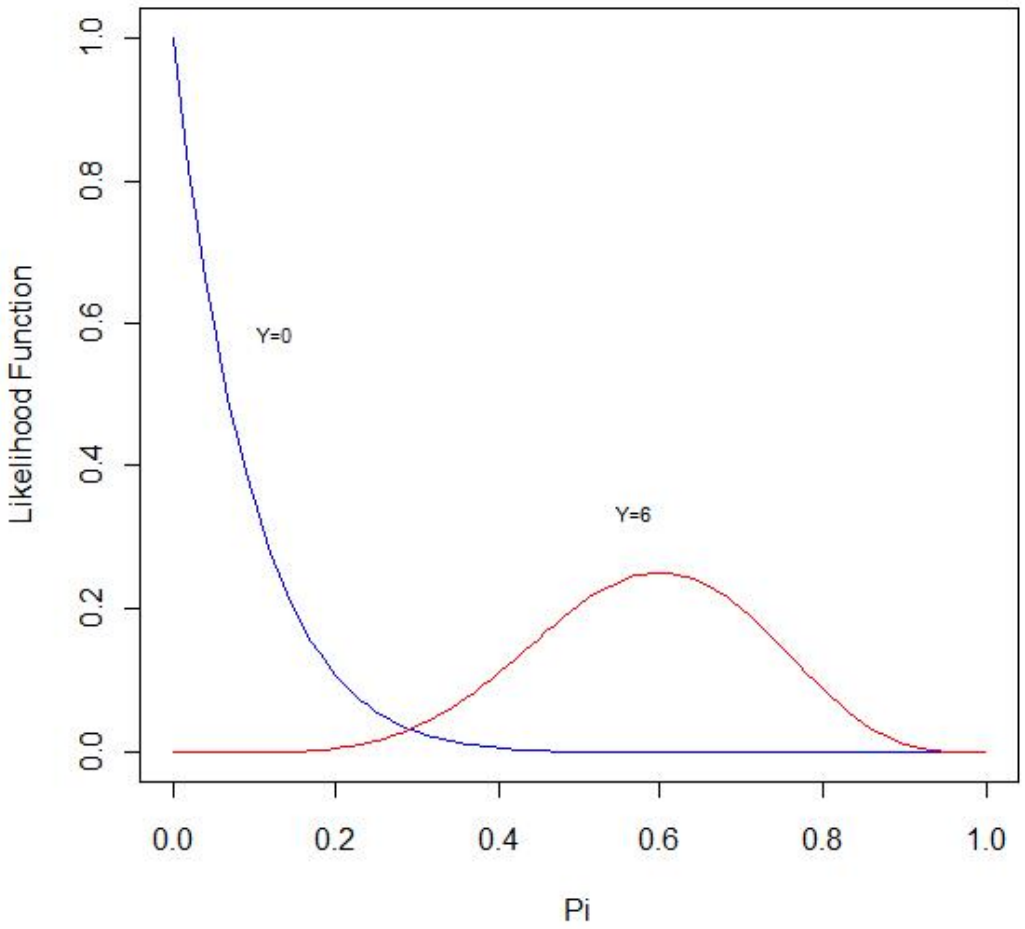
Maximum Likelihood Estimation (MLE)

- Consider the binomial model with $N = 10$ and let the observed count be $Y = 0$.
- Then

$$P[Y = 0] = \frac{10!}{0!10!}\pi^0(1-\pi)^{10} = (1-\pi)^{10}$$

- The probability of the observed data, expressed as a function of the parameter is called a *likelihood function*:

$$L(\pi) = P[Y = y] = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y}$$


Likelihood Function
0.0
0.2
0.4
0.6
0.8
1.0
Y=0
Y=6
0.0
0.2
0.4
0.6
0.8
1.0
Pi

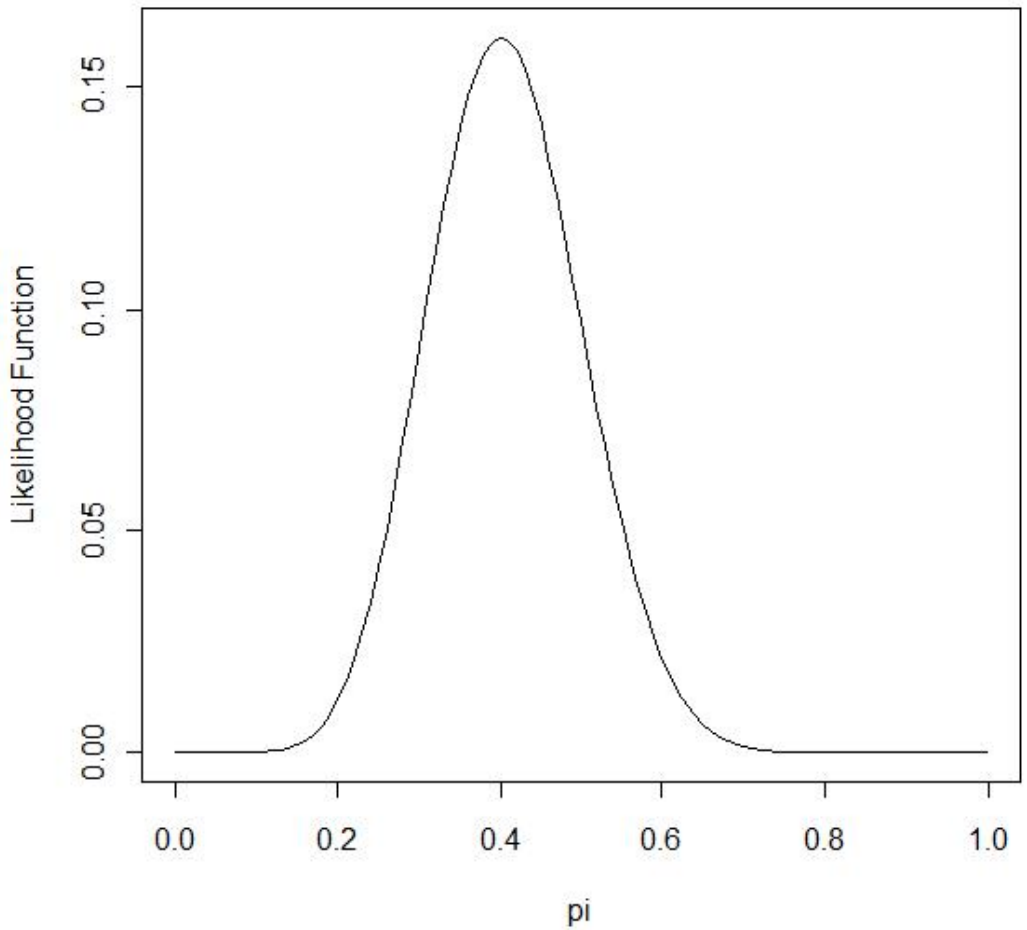
MLE

- The maximum likelihood estimator (MLE) is defined to be the parameter value, for which the likelihood function is maximized.
- For the binomial model, the MLE of π is the sample proportion, $\hat{\pi} = y/N$

Another Example

- A class of 25 students were asked whether he/she is a vegetarian.
- 10 students answered "yes".
- We have the likelihood function

$$L(\pi) = \frac{25!}{10!15!}\pi^{10}(1-\pi)^{15}$$


Likelihood Function
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
pi

Inference About π

- The estimate of the probability of a student being vegetarian,

$$\hat{\pi} = 10/25 = 0.4$$

- We want to test $H_0 : \pi = 0.5$ against the alternative $H_1 : \pi \neq 0.5$.

Testing of Hypothesis

- To test $H_0 : \pi = \pi_0$ against $H_1 : \pi \neq \pi_0$
- Note that, $E(\hat{\pi}) = \pi, var(\hat{\pi}) = \pi(1-\pi)/N$
- For The test statistic

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/N}}$$

- Reject H_0 at the significance level α, if

$$|Z| > z_{\alpha/2}$$

Confidence Interval

- A large sample $(1-\alpha)100\%$ confidence interval for π is given by

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

In Our Example

- $\hat{\pi} = 0.4, \pi_0 = 0.5$
- Therefore, the test statistic

$$Z = \frac{0.4 - 0.5}{\sqrt{0.5 \times 0.5/25}} = -1$$

- For $\alpha = 0.05, z_{\alpha/2} = 1.96$ and we do not reject H_0.
- Similarly, the 95% confidence interval is

$$0.4 \pm 1.96 \times \sqrt{\frac{0.4 \times 0.6}{25}} = 0.4 \pm 0.192 = (0.208, 0.592)$$

Notes

- Though the formula for test statistic and confidence interval is quite simple, they do not work well for very small or very large values of π.
- Since these are based on large sample assumptions, they do not work well for small N either.
- There are exact procedures available for small values of N.