Categorical Data Analysis
Lecture Note 1

Instructor: Seokho Lee

Hankuk University of Foreign Studies

1 An Introduction to the Analysis of Categorical Data

1.1. What is meant by "categorical data"?

A *categorical* variable has a measurement scale that consists of a set of categories.

Examples:

- Attitude toward gun control: favor or oppose
- Gender: male or female
- Education: Did not graduate from high school, high school, some college, bachelor's degree, master's degree, Ph.D.
- Cholesterol level: high, moderate, low
- Heart disease: yes, no

1.2. Discrete (Count) Data

Often data consist of the counts of the number of events of certain types. If the numbers are small, the methods for categorical data are often used for statistical analysis. We will also study methods of analysis specifically designed for count data.

Examples:

- Number of deaths from tornadoes in a state during a given year: 0, 1, 2, 3, $\cdots$
- Number of automobile accidents inside HUFS during a week

1.3. Response and Explanatory Variables

The *response* variable is often called the *dependent* variable.

The *explanatory* variables are often called *predictors* or *independent* variables.

Examples: Distinguish between response and explanatory variables.

- Job satisfaction (happy, so-so, unhappy) and Education
- Cholesterol level (high, moderate, low) and Heart Disease (yes, no)
- Number of deaths from tornadoes and number of tornadoes

1.4. Nominal and Ordinal Data

Categorical data with ordered scales are *ordinal* variables. Otherwise they are *nominal*.

Examples: Identify whether each variable is nominal or ordinal

- Education
- Political party affiliation (Democrat, Republican, Other)
- Cholesterol level (high, moderate, low)
- Hospital location (Yongin, Sungnam, Suwon, Seoul)

Methods designed for nominal variables:

- can be used for both nominal and ordinal variables.
- do not use the ordering for ordinal variables. This can cause a loss of power.
- give the same results no matter in which order the categories are listed.

Methods designed for ordinal variables:

- cannot be used with nominal variables.
- make use of the category ordering.
- would give different results if the categories were differently ordered.

1.5. Examples of Categorical Data Analyses

- In a sample of 50 adult Americans, only 14 correctly described the Bill of Rights as the first ten amendments to the U.S. Constitution. Estimate the proportion of Americans that can give a correct description of the Bill of Rights.
- In 1954, 401,974 children were participants in a large clinical trial for polio vaccine. Of these 201,229 were given the vaccine, and 200,745 were given a placebo. 110 of the children who received a placebo got polio and 33 of those given the vaccine got polio. Was the vaccine effective?
- The Storm Prediction Center tracks the number of characteristics of tornadoes. Construct a model relating the number of deaths to the number of tornadoes or to the number of killer tornadoes.
- A study was carried out to compare two treatments for a respiratory disorder. The goal was to compare the proportions of patients responding favorably to test and placebo. A confounding factor is that the study was carried out at two centers which had different patient populations. We wish to examine the association between treatment and response while adjusting for the effects of the centers.

1.5. **Examples of Categorical Data Analyses** (continue)

- Stillbirth is the death of a fetus at any time after the twentieth week of pregnancy. A premature birth is the live birth of a child from the twentieth until the thirty-seventh week of pregnancy. Fit loglinear models to investigate the association among the following variables that we recorded in a study of stillbirth in the Australian state of Queensland:
  - Birth status (B) - stillbirth or live birth
  - Gender (G) - male or female
  - Gestational age (A) - $\leq$ 24, 25–28, 29–32, 33–36, 37–41 weeks
  - Race (R) - Aborigine or white
- A female horseshoe crab with a male crab attached to her nest sometimes had other male crabs, called satellite, near her. Fit a logistic regression model to estimate the probability that a female crab has at least one satellite as a function of color, weight, and carapace width.
- 59 alligators are sampled in Florida. The response is primary food type: Fish, Invertebrate, and Other. The explanatory variable is length of the alligator in meters.

Data can be summarized in contingency table. For the second example of the above,

drug\get polio	yes	no	total
vaccine	110	201,119	201,229
placebo	33	200,712	200,745
total	143	401,831	401,974

In general, two-way contingency tables can be cross-classified as

$X \backslash Y$	1	$\cdots$	j	$\cdots$	J	total
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
total	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

The sample data are the cell counts: n_{ij}
We define row totals, column totals, and the grand total:

$$n_{i+} = \sum_{j=1}^{J} n_{ij} \qquad n_{+j} = \sum_{i=1}^{I} n_{ij} \qquad n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$