Introduction to Categorical Data Analysis

Abdus Sattar, Ph.D
sattar@case.edu

Categorical Data Analysis

Chapter 1: Outline

- **Part I**: Introduction to categorical variable
- **Part II**: **One** binary variable summary and inference
  1. Estimation of a proportion, π
  2. Confidence Interval (CI) for π
- **Part III**: **Two** binary variables summary and inference
  1. Estimation of π_1 - π_2, π_1/π_2, $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$
  2. CI for difference of Proportions, Relative Risk, Odds Ratio

Random variable

- A **random variable** is a characteristic measured on individuals in the sample and population. Age, sex, race, BMI, blood pressure, inflammatory biomarkers, and CVD risk factors, are examples of random variables.
- The random variable may be **outcome(Y)** of an experiment or **explanatory factor(X)**. Example: effects of statin drug (X) on IMT (Y) in a clinical trial.
- Random variables can be classified as:
  1. **Discrete**: sex, race, cancer stage (I, II, III, IV), obesity (underweight, normal, overweight, obese)
  2. **Continuous**: age, BMI, blood pressure, IL6, Viral load, IMT measures.

Measurement Scale: Nominal, Ordinal, and Interval

Table 1.1: Classification of variables by measurement scale

Measurement Scales			
Scale	**Discrete?**	**Definition**	**Examples**
Nominal	Yes	Set of categories, no ordering implied	Sex, race, hypertension (yes/no)
Ordinal	Yes	Ordering implied, but differences in scores not meaningful	Severity score (mild, moderate, severe), cancer stages (I, II, III, IV)
Interval	No	Differences between values are meaningful	Height, age, blood pressure

Patients with primary biliary cirrhosis of the liver

Table 1.2: Baseline data for selected variables of an RCT (N=312, Mayo Clinic)

PID	Treatment	Sex	Age	Bilirubin	Cholesterol	Histologic
201	1	1	58.8	14.5	261	4
209	1	1	56.5	1.1	302	3
217	1	2	70.1	1.4	176	4
221	1	1	54.7	1.8	244	4
222	2	1	38.1	3.4	279	3
Scale	**?**	**?**	**?**	**?**	**?**	**?**

Treatment: 1=DP, 2=Placebo; Sex: 1=Male, 2=Female;
Bilirubin(mg/dl); Cholesterol(mg/dl); Histologic state of disease: 1-4

Choice of analytic method depend on the scale

Table 1.3: Choice of analytic method depends on the scale

	Type of Analysis		
Scale	**Summary Statistics**	**Comparing two groups**	**Measuring Association**
Nominal	Frequency tables	Chi-square test	Contingency coefficient Kappa
Ordinal	Frequency tables	Chi-square test for trend, Nonparametric test	Spearnman's r, Kendall's tao
Interval	Mean, SD	t-test, Nonparametric test	Spearman's r or Pearson's r

Outcome type determines regression models

Table 1.4: Regression models depends on outcome type

Outcome classification	Outcome type	Regression model
Numerical	Continuous	Linear
	Count	Poisson model
	Time-to-event	Proportional hazards
Categorical	Binary	Logistic
	Ordinal	Proportional odds
	Nominal	Polytomous logistic