STAT 426

1.1 Categorical Response Data

Categorical response data

- We make the distinction between response (or dependent) variables and explanatory (or independent) variables.
- We concentrate on the response variables that are categorical.
- Explanatory variables can be of any type as in ordinary regression models

Categorical response data

Data types: Quantitative Data; Qualitative Data

Categorical variables scales:

- Nominal: Categories do not have a natural order: blood type, gender
- Ordinal: Categories have a natural order: Low / middle / high education level.

Other levels of measurements:

- Interval: There is a numerical distances between any two values.

  Example: Blood pressure level. SAT Score → difference between two values is meaningful.

- Ratio: An interval variable where ratios are valid (presence of absolute zero). Zero is meaningful.

  Example: Distance run by an athlete.

Weight. $4g = 2g \times 2$.

Categorical Data Examples

Categorical data arise from many applications in the real world, specially in social and biomedical sciences. Let's see some examples:

Example

Example 1: Level of income A new business is investigating the power income of their potential customers, to agree upon an affordable price on their products. They use a questionnaire with the question: What is your household income?

- Below \$30,001
- \$30,001 - \$40,000
- \$40,001 - \$50,000
- \$50,001 and above

Interval. / Ordinal.

Quant.

Categorical Data Examples

Example

Example 2: Level of education When filling forms for job applications, admission, training, etc., a respondent is usually requested the level of education. Companies use a questionnaire with the question: What is your highest level of education?

- School SAT
- High School
- BSc
- MSc
- PhD

Ordinal.

Qualit.

Categorical Data Examples

Example

Example 3: Level of satisfaction Businesses normally like to rate a customer service rendered in order to improve their service. They use a questionnaire with the question:
Kindly rate your customer service experience with us:

- Very poor
- Poor
- Neutral
- Good
- Very good

Ordinal.

Categorical Data Examples

Example

Example 4: Level of motivation A company wants to improve employee productivity and uses a questionnaire to study what motivates employees to work better. The specific question is: *What motivates you to work better? (If Others please specify specify)*

- Peer motivation
- Recognition
- Professional growth opportunities
- Friendly work culture
- Others ——

Nominal.

Categorical Data Examples

Example

Example 5: Motives for travelling Travel companies want to improve their marketing strategies and ask their customers the following question: *What are your motives for travelling? (If Others please specify)*

- Business
- Leisure
- Family
- Study
- Health
- Others ——

Nominal.

Levels of measurements

- A variable's level of measurement determines the statistical methods to be used for its analysis.
- Variables hierarchy: Ratio $>$ Interval $>$ Ordinal $>$ Nominal
- Statistical methods applied to variables at a lower level can be used with variables at a higher level but the contrary is not true.

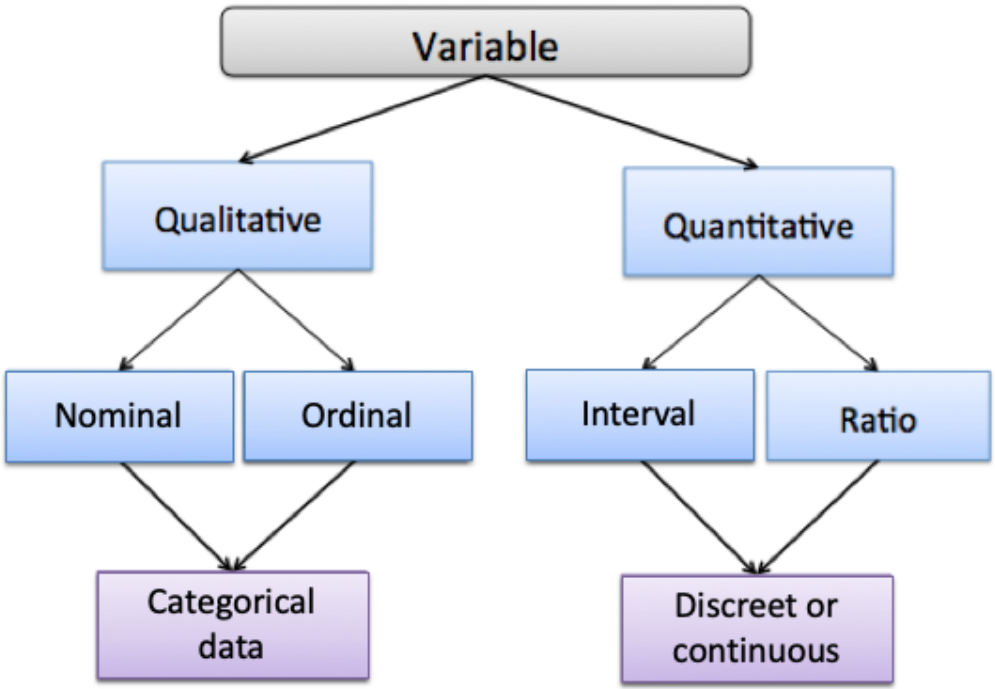
  Example: An ordinal variable can be analyzed as a nominal variable (ignoring the order) but the opposite does not apply.
- In this class we will discuss the analysis of nominal and ordinal variables.
- The methods can also be applied to interval and ratio variables grouped into a small number of ordered categories.

  Example: Years of Education: 0-10 years, 10-12 years, $>$ 12 years and so on.

Levels of Measurements

It is important to make the distinction between:

- Continuous vs. Discrete variables
- Quantitative vs. Qualitative



Models for Categorical Responses

- Models in this class resemble regression models with continuous response variables, but responses can have
  - Binomial
  - Multinomial
  - Poisson
- Featured models are:
  - Logistic regression models (logit): Binary responses and assume a binomial distribution.
  - Generalization of the logistic regression: Multi-category responses (nominal and ordinal).
  - Loglinear models: Count data and assume a Poisson distribution.
  - Generalizations to multivariate categorical responses: Represent associations and interactions among variables.
  - Models for repeated categorical responses: Longitudinal data.

Example (Agresti, Table 2.1)

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10845
Aspirin	5	99	10933

- Is there a relationship between myocardial infarction and aspirin use?
- If so, of what kind? How strong?
- What kinds of quantities can measure the strength of the relationship?

Example (Agresti, Table 2.1)

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10845
Aspirin	5	99	10933

What kind of statistical models might be appropriate?

- What is random?
- What would be the distributions?
- How would you form estimates? Tests? Confidence intervals?

Please review:

- convergence in probability ($\to_p$)
- convergence in distribution ($\overset{d}{\to}$)
- law of large numbers (LLN)
- central limit theorem (CLT)
- consistency and asymptotic normality
- normal approximation to the binomial
- Poisson approximation to the binomial
- normal approximation to the Poisson
- confidence intervals (CIs)
- tests (level, power) and P-values
- marginal and conditional probabilities, densities, and distributions
- Bayes' rule for probabilities