Analysis of Discrete Data
STAT 504

Lecture 1, 1/13/2005

Instructor:
Aleksandra (Seša) Slavković
sesa@stat.psu.edu

TA:
TBA
TBA@stat.psu.edu

Some common terminology

- Observational units
- Sample vs. Population
- Statistics vs. Parameter
- Random Variables
- Experiment vs. Observational Study

Observational Units

- Observational units are entities whose characteristics we measure.

  Synonyms: “case” or “subject.”

- In the social sciences, observational units are often people or groups of people. In the life sciences, observational units might be animals, bacterial colonies, etc.

Observational Units

- In the physical sciences, observational units might be “samples” of some chemical substance or some manufactured good.

  It is a bad habit to use the word “sample” for each observational unit; in standard statistical terminology the whole group of observational units are referred to as “the sample.”

Population and Sample

- Population: the entire collection of units about which we would like information (e.g. all apartments in State College)

- Sample: the collection of units we actually measure (e.g. 100 apartments in SC)

- Parameter: the true value we hope to obtain (e.g. true proportion of 1bd apartments in SC)

- Statistic: an estimate of the parameter based on observed information in the sample (e.g. observed proportion of 1bd apartments out of 100 sampled apartments)

  Parameters are generally unknown so we estimate them with sample statistics

Random Variables (RVs)

- Random variables are characteristics of the observational units which can have different possible values (this is the practical, not the statistical definition)

  Types
  - Quantitative (numerical, measurement) variables represent an amount or quantity of something (e.g. time spent waiting for the bus)
  - Qualitative variables represent things that can be categorized (e.g. the colors of the cars that pass while you wait for the bus)

  Letters like X or Y represent random variables if its value is not known before the experiment is run.
Discrete vs. Continuous

- **Discrete** random variables can only take on values from a countable set of numbers such as the integers or some subset of integers. (Usually, they can’t be fractions.)

- **Continuous** random variables can take on any real number in some interval. (They can be fractions.)

  **Note:** We consider variables like height to be continuous even though we can only measure them in discrete units (e.g. millimeters); think of a space of possible values a variable can take.

Categorical: Nominal vs Ordinal

- **Nominal** (unordered) random variables have categories where order doesn’t matter (e.g. gender, ethnic background, religious affiliation, …)

- **Ordinal** (ordered) random variables have ordered categories (e.g. social class, health status, …); numerical distance between categories is unknown (“how much better” is a patient with a good condition vs. fair condition)

Measurement scale

- Interval variables have a numerical distance between two values (e.g. income)

  **Note:** Methods applicable for one type of variable can be used for the variables at higher levels too (but not at lower levels)

- Example: grades
  - Nominal: pass/fail
  - Ordinal: A,B,C,D,F
  - Interval: 4,3,2,5,2,1

Explanatory vs. Response Variable

- **Explanatory** (independent) variable attempts to explain (or is purported to cause) differences in a response (outcome, dependent) variable.

  **E.g.** homework scores and exam scores can be explanatory variables for the final grade.

Focus of this class

- Discretely measured responses
  - Nominal variables
  - Ordinal variables
  - Discrete interval variables with few values
  - Continuous variables grouped into a small number of categories

Analysis Grid

<table>
<thead>
<tr>
<th></th>
<th>Continuous Outcome</th>
<th>Categorical Outcome (response, dependent)</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Continuous Explanatory</strong></td>
<td>Regression</td>
<td>Logistic regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loglinear models</td>
<td></td>
</tr>
<tr>
<td><strong>Categorical Explanatory</strong> (Independent)</td>
<td>ANOVA</td>
<td>Chi-Square Test of Independence, loglinear models</td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>ANOVA</td>
<td>Logistic regression/ loglinear models</td>
<td></td>
</tr>
</tbody>
</table>
Types of Studies

- **Randomized Experiment**: we create differences in the explanatory variable and then examine the results.
  - The investigators apply one or more manipulations (i.e. treatments) to the experimental subjects.
  - Subjects are randomly assigned to treatments.

- **Observational Study**: we observe differences in the explanatory variables.

---

Example of a Categorical Variable

- Notation:
  - Population proportion = \( p \)  
  - Population size = \( N \)  
  - Sample proportion = \( \hat{p} = \frac{X}{n} = \# \text{ with trait} / \text{ total #} \)  
  - Sample size = \( n \)

- The Rule for Sample Proportions
  - If numerous samples of size \( n \) are taken, the frequency curve of the sample proportions \( \hat{p} \)'s from the various samples will be approximately normal with the mean \( \mu = p \) and standard deviation \( \sigma = \sqrt{\frac{p(1-p)}{n}} \)

---

Contingency Table

- A statistical tool for summarizing and displaying results for categorical variables.

- A two-way table if for two categorical variables.

- 2x2 Table, for two categorical variables, each with two categories.

- Place the counts of each combination of the two variables in the appropriate cells of the table.

---

Example: Admissions Data

- A university offers only two degree programs: English and Computer Science. Admission is competitive and there is a suspicion of discrimination against women in the admission process. Here is a two-way table of all applicants by sex and admission status:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admit</td>
<td>35</td>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td>Deny</td>
<td>45</td>
<td>40</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>60</td>
<td>140</td>
</tr>
</tbody>
</table>

These data show an association between the sex of the applicants and their success in obtaining admission.

---

Example: Clinical Trial of Effectiveness of an Analgesic Drug

- Source: Koch et al. (1982)

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Active</th>
<th>Poor</th>
<th>Moderate</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>20</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Placebo</td>
<td>11</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Placebo</td>
<td>6</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Placebo</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

Fixed number of patients in two Treatment groups.

Small counts.

Can model causal relationship.

---

Example: Delinquent Children by the Education Level


<table>
<thead>
<tr>
<th>Education Level of Head of Household</th>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>15</td>
<td>20</td>
<td>3</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>Very</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>35</td>
<td>40</td>
<td>20</td>
<td>135</td>
</tr>
</tbody>
</table>

Ordinal and nominal variables.

Fixed total.
Example: Census Data

- Source: American Fact Finder website (U.S. Census Bureau: Block data)

<table>
<thead>
<tr>
<th>Sex Age Race</th>
<th>Male Under 18 years</th>
<th>Male 18 Years &amp; over</th>
<th>Female Under 18 years</th>
<th>Female 18 Years &amp; over</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>6</td>
<td>31</td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td>Black</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Asian</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total minority</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Review of discrete probability

- Based on Dr. Schafer’s notes
- Agresti, Ch. 1

Next Lecture

- Continue Review of Discrete Probability
  - Likelihood/ LogLikelihood/ ML estimation
  - Agresti, ch. 1