Categorical Response and Probability Distributions

Demetris Athienitis

UF UNIVERSITY of FLORIDA

Categorical data

Definition

A categorical variable is any variable whose measurement scale consists of a set of categories

- Nominal variables: Unordered categories
- Ordinal variables: Ordered categories

Remark

Methods developed for ordinal variables utilize ordering information and therefore can not be used for nominal variables. Alan Agresti has a supplemental book *Analysis of Categorical Ordinal Data*.

Examples

- Nominal data
  - Favorite color: Red, green, yellow, etc.
  - Type of music: Rap, pop, country, etc.
- Ordinal data
  - Rate your pain on a scale of 0-10
  - Political beliefs: Liberal, moderate, conservative

Response variables vs. explanatory

Frequently we distinguish between response and explanatory variables.

- Explanatory variables are used to explain changes in the response variable.
- Explanatory variables also referred to as *independent* variables or *predictors*. Response variables are referred to as *dependent* variables.
- We focus on methods where the response (or dependent) variable is categorical and the explanatory variables are either categorical or continuous.

Section 2

Probability distributions for categorical data

There are two very important classes of categorical distributions (for this class)

- Bernoulli, Binomial, Multinomial distributions
- Poisson distribution (to be seen later)

Bernoulli distribution

Y be a random variable that only takes two values

$$Y = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1 - \pi \end{cases}$$

- The PMF defined as

$$P(Y = y) \equiv p(y) = \pi^y (1 - \pi)^{1-y}, \qquad y = 0, 1 \quad 0 \leq \pi \leq 1$$

- Denote by $Y \sim \text{Bernoulli}(\pi)$
- $E(Y) = \pi$ and $V(Y) = \pi(1 - \pi)$

Example

Suppose we roll a die and let Y be an indicator of whether a 5 is rolled

$$Y = \begin{cases} 1 & \text{if outcome is 5} \\ 0 & \text{otherwise} \end{cases}$$

Then, $Y \sim \text{Bernoulli}(1/6)$ with mean $1/6$ and variance $5/36$.

Binomial distribution

Let $Y_1, Y_2, \ldots, Y_n$ be Bernoulli random variables such that

1. the random variables are independent of each other
2. each Bernoulli has identical success probability π

Now let $Y = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \pi)$ where the PMF is given by

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n$$

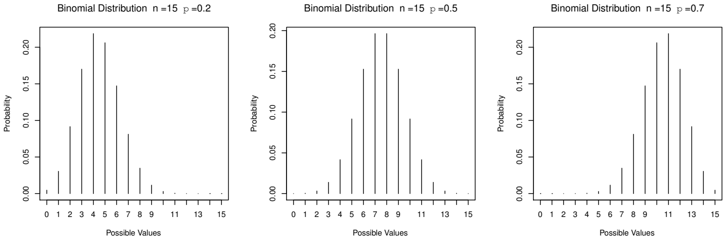
where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$

Binomial distribution

Using the rules for means and variances we can see that

- $E(Y) = n\pi$
- $V(Y) = n\pi(1-\pi)$

Let's look at the PMF for the binomial for different π values



Example

- A die is rolled 4 times and the number of 5s is observed (y)
- $Y \sim \text{Binomial}(4, 1/6)$ and therefore the PMF is

y	$p(y)$
0	0.4823
1	0.3858
2	0.1157
3	0.0154
4	0.0008

- Find the probability that there is at least one 5

$$P(Y \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 0.5177$$

In R, one would simple use

```
1-pbinom(0,4,1/6)
```

Binomial random variables

Interest frequently lies in the proportion of successes

$$\hat{\pi} = \frac{Y}{n} = \frac{\sum Y_i}{n}$$

Using the rules of means and variances we can see

$$E(\hat{\pi}) = \pi$$
$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

Multinomial random variables

An extension of a binomial with categories $c \geq 2$

- Probability π_i of being in category i, such that $\sum_{i=1}^{c} \pi_i = 1$
- The probability that y_1 are category 1, y_2 are category 2, etc. is

$$P(Y_1 = y_1, \ldots, Y_c = y_c) = \left(\frac{n!}{y_1! y_2! \ldots y_c!} \right) \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_c^{y_c}$$

where $n = \sum_{i=1}^{c} y_i$

We learned

- What is a categorical response variable
- Binomial distribution