Discrete Response Model
Lecture 4

datascience@berkeley

Introduction to Multinomial Probability Distribution

Introduction

The previous three weeks provided analysis methods for when there were binary responses. The purpose of this week is to generalize some of these previous methods to allow for more than two response categories. Examples include:

- Canadian political party affiliation—Conservative, New Democratic, Liberal, Bloc Quebecois, or Green
- Chemical compounds in drug discovery experiments—positive, blocker, or neither
- Cereal shelf placement in a grocery store—bottom, middle, or top
- Beef grades—Prime, choice, select, standard, utility, and commercial
- Five-level Likert scale—strongly disagree, disagree, neutral, agree, or strongly agree

For these examples, some responses are ordinal (e.g., Likert scale) and some are not (e.g., chemical compounds).

Multinomial Probability Distribution

The multinomial probability distribution is the extension of the binomial distribution to situations where there are more than two categories for a response.

Notation:

- Y denotes the response category with levels of $j = 1, \ldots, J$
- Each category has a probability of $\pi_j = P(Y = j)$.
- n denotes the number of trials
- $n_1, \ldots, n_J$ denote the response count for category j, where $\sum_{j=1}^{J} n_j = n$

The probability mass function for observing particular values of $n_1, \ldots, n_J$ is

$$\frac{n!}{\prod_{j=1}^{J} n_j!} \prod_{j=1}^{J} \pi_j^{n_j}$$

Example: Multinomial Simulation

As a quick way to see what a sample looks like in a multinomial setting, consider the situation (a sample) with **n = 1,000 trials**, $\pi_1 = 0.25$, $\pi_2 = 0.35$, $\pi_3 = 0.2$, $\pi_4 = 0.1$, and $\pi_5 = 0.1$. Below is how we can simulate a sample:

```
> pi.j<-c(0.25, 0.35, 0.2, 0.1, 0.1)
> set.seed(2195) #Set a seed to be able to reproduce the sample
> n.j<-rmultinom(n = 1, size = 1000, prob = pi.j)
> data.frame(n.j, pihat.j = n.j/1000, pi.j)
  n.j pihat.j pi.j
1 242   0.242 0.25
2 333   0.333 0.35
3 188   0.188 0.20
4 122   0.122 0.10
5 115   0.115 0.10
```

Suppose there are m = 5 separate sets of n = 1000 trials.

```
set.seed(9182)
n.j<-rmultinom(n = 5, size = 1000, prob = pi.j)
     [,1] [,2] [,3] [,4] [,5]
[1,]  259  259  237  264  247
[2,]  341  346  374  339  341
[3,]  200  188  198  191  210
[4,]   92  106   89  108  107
[5,]  108  101  102   98   95
```

```
> n.j/1000
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.259 0.259 0.237 0.264 0.247
[2,] 0.341 0.346 0.374 0.339 0.341
[3,] 0.200 0.188 0.198 0.191 0.210
[4,] 0.092 0.106 0.089 0.108 0.107
[5,] 0.108 0.101 0.102 0.098 0.095
```

Notice the variability from one set to another.

Berkeley SCHOOL OF INFORMATION