The Multinomial Distribution

Rebecca C. Steorts

Computing set up

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(margins)
library(ggtern)
library(viridis)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "70%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Learning goals

- Introduce multinomial data
- Introduce the multinomial distribution
- Visualize the distribution
- Write the distribution as an exponential family in canonical form.
- Exercises to continue working with the exponential family.

Multinomial data

Multinomial data arises naturally when we count outcomes across multiple mutually exclusive categories over a fixed number of independent trials.

Examples of Multinomial data

1. A researcher surveys 500 students, asking: what is your primary method of transportation to campus? Answers in the survey include the following: walking, biking, car, bus, and other. In this example, we count the modes of transportation for each student.
2. In an election, 10,000 people vote for one of 3 candidates: A, B, or C. In this example, we count the number of votes for each candidate.
3. In a study of 1,000 participants, individuals that own cars, rate the importance of air conditioning in four categories: "not important" to "very important."

Multinomial Simulation

```r
set.seed(42)

n_trials <- 10000
probs <- c(0.2, 0.5, 0.3)
categories <- c("A", "B", "C")


# each observation is a 0/1 draw for the category
sim_data <- rmultinom(n = n_trials, size = 1, prob = probs)
# for each observation, we update if it belongs with category 1,2 or 3
category_indices <- apply(sim_data, 2, function(x) which(x == 1))
# we create a label mapping for A, B, and C
category_labels <- categories[category_indices]

# Create a data frame
df <- data.frame(Category = factor(category_labels,
                                   levels = categories))
```

Multinomial Simulation

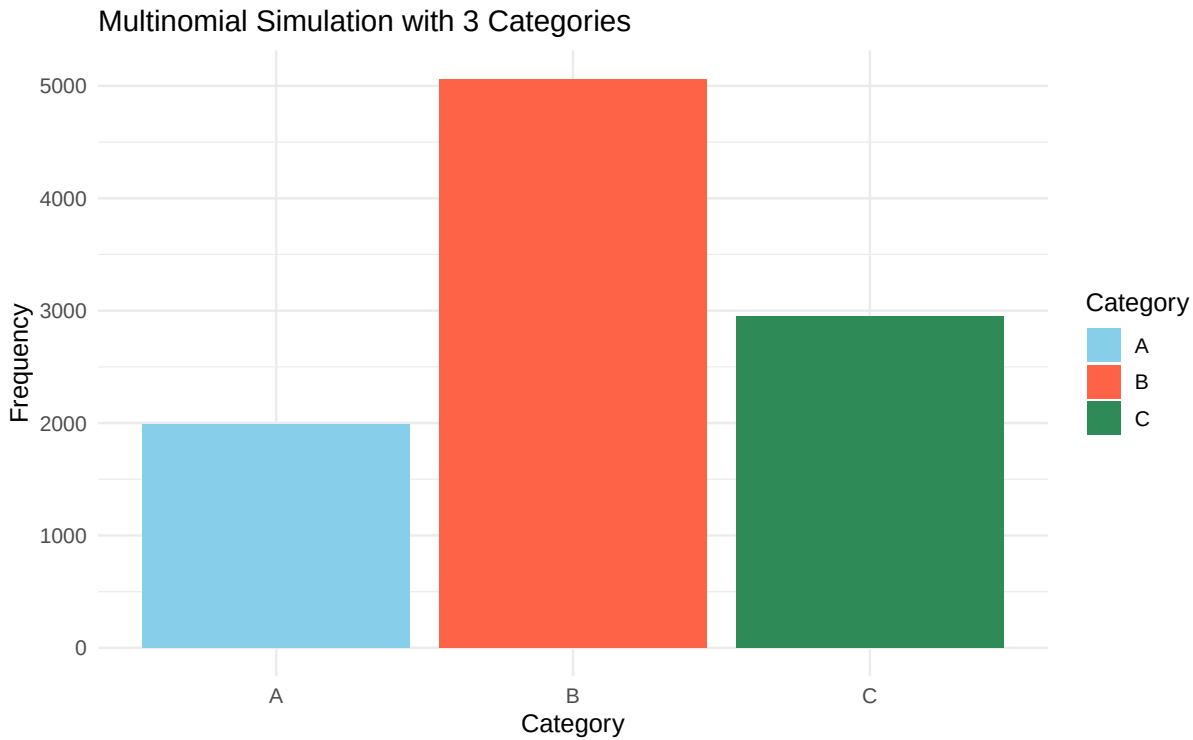
```
table(df$Category)
```

```
##
##    A    B    C
## 1990 5061 2949
```

```
prop.table(table(df$Category))
```

```
##
##       A       B       C
## 0.1990 0.5061 0.2949
```

Multinomial Simulation



Multiomial Distribution

Let $X = (X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n, \mathbf{p})$, where:

- $n \in \mathbb{N}$ is the number of data points,
- $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ with $\sum_{i=1}^{k} p_i = 1$, and $p_i \geq 0$,
- X_i is the number of outcomes in category i, with $\sum_{i=1}^{k} X_i = n$.

Then, the probability mass function is given by:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Or, using product notation:

$$P(\mathbf{x}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i}$$

Background: Exponential Family

A probability distribution belongs to the canonical exponential family if its density or pmf can be written in the form:

$$f(x \mid \theta) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right)$$

where:

- $T(x)$ is the vector of sufficient statistics,
- $\eta(\theta)$ is the vector of natural (canonical) parameters,
- $A(\theta)$ is the log-partition function,
- $h(x)$ is the base measure.

Multinomial distribution

Let $\mathbf{X} = (X_1, \ldots, X_k) \sim \text{Multinomial}(n, \mathbf{p})$, with:

- $\sum_{i=1}^{k} X_i = n$,
- $\sum_{i=1}^{k} p_i = 1$, and $p_i > 0$.

The pmf is:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^{k} p_i^{x_i}$$

Multinomial distribution

Since $\sum x_i = n$ and $\sum p_i = 1$, we can express everything in terms of the first $k-1$ components:

$$x_k = n - \sum_{i=1}^{k-1} x_i,$$

$$p_k = 1 - \sum_{i=1}^{k-1} p_i.$$

Now, we subsitute this into the pmf to find that

$$P(\mathbf{x}) = \frac{\mathbf{n}!}{\mathbf{x_1}! \cdots \mathbf{x_{k-1}}! \left(\mathbf{n} - \sum_{\mathbf{i=1}}^{\mathbf{k-1}} \mathbf{x_i}\right)!} \prod_{\mathbf{i=1}}^{\mathbf{k-1}} \mathbf{p_i^{x_i}} \left(\mathbf{1} - \sum_{\mathbf{i=1}}^{\mathbf{k-1}} \mathbf{p_i}\right)^{\mathbf{n} - \sum_{\mathbf{i=1}}^{\mathbf{k-1}} \mathbf{x_i}}$$

Exponential Family

We rewrite the product terms as exponentials:

$$\prod_{i=1}^{k-1} p_i^{x_i} = \exp\left(\sum_{i=1}^{k-1} x_i \log p_i\right)$$

$$\left(1 - \sum_{i=1}^{k-1} p_i\right)^{n-\sum x_i} = \exp\left(\left(n - \sum x_i\right) \log p_k\right)$$

Combining these:

$$\exp\left(\sum_{i=1}^{k-1} x_i \log p_i + \left(n - \sum x_i\right) \log p_k\right)$$

$$= \exp\left(\sum_{i=1}^{k-1} x_i \log\left(\frac{p_i}{p_k}\right) + n \log p_k\right)$$

Exponential Family

Now write the full expression:

$$P(\mathbf{x}) = \underbrace{\frac{n!}{x_1! \cdots x_{k-1}!(n - \sum x_i)!}}_{\mathbf{h(x)}} \cdot \exp\left(\sum_{\mathbf{i=1}}^{\mathbf{k-1}} \mathbf{x_i} \log\left(\frac{\mathbf{p_i}}{\mathbf{p_k}}\right) + \mathbf{n} \log \mathbf{p_k}\right)$$

This matches the exponential family form:

$$P(\mathbf{x}) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^\top T(\mathbf{x}) - A(\boldsymbol{\eta})\right),$$

where the terms are given on the next slide.

Exponential Family Form

Sufficient statistics:

$$T(\mathbf{x}) = (x_1, \ldots, x_{k-1})$$

Natural (canonical) parameters:

$$\eta_i = \log\left(\frac{p_i}{p_k}\right), \quad i = 1, \ldots, k-1$$

Base measure:

$$h(\mathbf{x}) = \frac{n!}{x_1! \cdots x_{k-1}!(n - \sum x_i)!}$$

Exponential Family Form

Log-partition function:

Using:

$$p_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{k-1} e^{\eta_j}}, \quad p_k = \frac{1}{1 + \sum_{j=1}^{k-1} e^{\eta_j}}$$

Then:

$$\log p_k = -\log \left(1 + \sum_{j=1}^{k-1} e^{\eta_j} \right)$$

So the log-partition function is:

$$A(\boldsymbol{\eta}) = -n \log p_k = n \log \left(1 + \sum_{j=1}^{k-1} e^{\eta_j} \right)$$

Exponential Family

Thus, the final canonical form of the exponential family is given by

$$P(\mathbf{x}) = \frac{n!}{x_1! \cdots x_{k-1}!(n - \sum x_i)!} \exp\left(\sum_{i=1}^{k-1} x_i \eta_i - n \log\left(1 + \sum_{j=1}^{k-1} e^{\eta_j}\right)\right)$$

Connections to GLMs

There are two main approaches to GLMs for Multinomial data.

1. Nominal data: This is used when there is no natural ordering to the data among the response categories.

Specifically, one category is chosen as the baseline (often the first category).

The logits for the remaining categories are defined as follows:

$$\text{logit}(p_j) = \log(\frac{p_j}{p_1}) = {x_j}^T \beta_j \quad j = 2, \ldots K$$

.

The $(K-1)$ logit equations are used to simultaneousl estimate the β_j parameters.

Connections to GLMs

2. Ordinal data: This is used when there is an obvious natural ordering among the response categories.

This is assessed by a crude method that sets cut points $C_1, \ldots, C_{K-1}$ that correspond to the K ordinal categories with associated probabilities $p_1, \ldots, p_k$ (that sum to 1).

Exercise 1

Show the Binomial distribution is a special case of the Multinomial distribution.

Exercise 2

Using the log-partition function, give the mean of the Multinomial distribution.

Exercise 3

Using the log-partition function, give the covariance matrix of the Multinomial distribution.