Categorical and multinomial distributions

David J. H. Shih

Intended learning outcomes

- Recognize and apply categorical and multinomial distributions in statistical models
- Derive the multinomial distribution
- Derive posterior predictive distributions for statistical models

Categorical distribution

Definition

A random variable X with support $\mathcal{X} = \{1, 2, \dots, K\}$ has a categorical distribution if

$$P(X = k \mid \boldsymbol{\theta}) = \theta_k, \quad k \in \mathcal{X}.$$

where $\boldsymbol{\theta} = [\theta_k]$ and θ_k is the probability of success for outcome k s.t. $\sum_k \theta_k = 1$.

Other representations are

$$\begin{aligned} P(X = x \mid \boldsymbol{\theta}) &= \prod_k^K \theta_k^{I(x=k)} \\ p(X = x \mid \boldsymbol{\theta}) &= \text{Categorical}(x \mid \boldsymbol{\theta}) \\ X &\sim \text{Categorical}(\boldsymbol{\theta}) \\ X &\sim \boldsymbol{\theta}. \end{aligned}$$

Categorical distribution

- Generalization of Bernoulli distribution
- Representation for empirical probability mass function

Example

Let $X = \{1, 2, 3, 4\}$ represent the suit (Club, Diamond, Heart, Spade) of a playing card drawn from a cut deck.

$$X \sim \boldsymbol{\theta}$$

Dirichlet distribution

Define a random vector $\boldsymbol{X} = [X_1, \ldots, X_K]$ with $K \geq 2$ such that

$$X_k \in [0, 1] \quad \text{and} \quad \sum_k^K X_k = 1.$$

$\boldsymbol{X}$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ if

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_k^K x_k^{\alpha_k - 1},$$

where $\alpha_k > 0$ and $B(\boldsymbol{\alpha})$ is the multivariate beta function:

$$B(\boldsymbol{\alpha}) = \frac{\prod_k^K \Gamma(\alpha_k)}{\Gamma\left(\sum_k^K \alpha_k\right)}.$$

Dirichlet-categorical model

Given data $X \in \{1, \ldots, K\}$, define model

$$\begin{aligned} X &\sim \boldsymbol{\theta} \\ \boldsymbol{\theta} &\sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \end{aligned}$$

with parameter $\boldsymbol{\theta}$ and hyperparameter $\boldsymbol{\alpha}$.

The posterior distribution of $\boldsymbol{\theta}$ is

$$p\left(\boldsymbol{\theta} \mid x\right) = \frac{p\left(x \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(x\right)} \quad \text{(Bayes' theorem)},$$

where we need to solve $p\left(x \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$ and $p\left(x\right)$.

We first solve the numerator:

$$
\begin{aligned}
p\left(x \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) &= \theta_x \left(B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha_k - 1} \right) \\
&= \left(\prod_k \theta_k^{I(x=k)} \right) B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha_k - 1} \\
&= B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha_k + I(x=k) - 1} \\
&= B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha'_k - 1},
\end{aligned}
$$

where $\alpha'_k = \alpha_k + I(x = k)$.

Now, we solve the denominator:

$$
\begin{aligned}
p\left(x\right) &= \int_{\Theta} p\left(x \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \quad \text{(total law of prob.)} \\
&= \int_{\Theta} B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha'_k - 1} d\boldsymbol{\theta} \\
&= B(\boldsymbol{\alpha})^{-1} \int_{\Theta} \prod_k \theta_k^{\alpha'_k - 1} d\boldsymbol{\theta} \\
&= B(\boldsymbol{\alpha})^{-1} B(\boldsymbol{\alpha}') \quad \text{(Dirichlet distribution integrates to 1)}
\end{aligned}
$$

Finally, we solve the posterior by substituting in the numerator and the denominator.

$$\begin{aligned} p\left(\boldsymbol{\theta} \mid x\right) &= \frac{B(\boldsymbol{\alpha})^{-1} \prod_k \theta_k^{\alpha'_k - 1}}{B(\boldsymbol{\alpha})^{-1} B(\boldsymbol{\alpha}')} = \frac{\prod_k \theta_k^{\alpha'_k - 1}}{B(\boldsymbol{\alpha}')} \\ &= \text{Dirichlet}\left(\theta \mid \boldsymbol{\alpha}'\right), \end{aligned}$$

where $\alpha'_k = \alpha_k + I(x = k)$.

Multiple categorical random variables

Given observed iid data $Z_i \in \{1, \ldots, K\}$ for $i \in \{1, \ldots, N\}$ and known parameter $\boldsymbol{\theta}$, suppose

$$Z_i \sim \boldsymbol{\theta}.$$

In other words,

$$\begin{aligned}
P\left(Z_i = k \mid \boldsymbol{\theta}\right) &= \theta_k \\
P\left(Z_i = z_i \mid \boldsymbol{\theta}\right) &= \prod_k \theta_k^{I(z_i = k)} \\
\prod_i P\left(Z_i = z_i \mid \boldsymbol{\theta}\right) &= \prod_i \prod_k \theta_k^{I(z_i = k)} \\
&= \prod_k \theta_k^{\sum_i^N I(z_i = k)}
\end{aligned}$$

For $k = 1 \ldots K$, define $x_k = \sum_i^N I(z_i = k)$. Then,

$$\prod_i P\left(Z_i = z_i \mid \boldsymbol{\theta}\right) \;=\; \prod_k^K \theta_k^{x_k}$$

Define random vector $\boldsymbol{X} = [X_k]$, where $X_k = \sum_i^N I(Z_i = k)$.

$$P\left(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\right) \;\propto\; \prod_k^K \theta_k^{x_k},$$

where each $\boldsymbol{x}$ can arising from different number of occurrences of each realization $\boldsymbol{z} = [z_i]$.

Example where $K = 4$ and $N = 6$:

$\boldsymbol{z}$	$\boldsymbol{x}$
[1 1 1 1 1 1]	[6 0 0 0]
[2 2 2 2 2 2]	[0 6 0 0]
...	
[1 1 1 2 2 2]	[3 3 0 0]
[1 2 1 1 2 2]	[3 3 0 0]
...	[3 3 0 0]
...	
[1 2 2 1 3 4]	[2 2 1 1]
[2 1 2 1 3 4]	[2 2 1 1]
[3 1 2 1 4 2]	[2 2 1 1]
[4 1 2 1 3 2]	[2 2 1 1]
...	[2 2 1 1]
...	

There are $N!$ ways of drawing random vector $[Z_i]$, which consists of N random variables.

We cannot distinguish among all ways of arranging all the Z_i such that $Z_i = 1$, so we divide by $X_1!$.

Similarly for $X_2, X_3, ...$

Therefore, each $\boldsymbol{x}$ has $\frac{N!}{x_1!\, x_2! \ldots x_K!}$ occurrences, which means that

$$P\left(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}\right) \;=\; \frac{N!}{\prod_k^K x_k!} \prod_k^K \theta_k^{x_k}.$$

Multinomial distribution

Definition

A random vector $\boldsymbol{X} = [X_1, \ldots, X_K]$ with support

$$\mathcal{X} = \left\{ [x_1, \ldots, x_K] : x_k \geq 0, \sum_k x_k = N \right\}$$

has a multinomial distribution if

$$\begin{aligned} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}) &= \frac{N!}{\prod_k x_k!} \prod_k \theta_k^{x_k} \\ &= \frac{\Gamma(N+1)}{\prod_k \Gamma(x_k+1)} \prod_k \theta_k^{x_k}, \end{aligned}$$

where θ_k is probability of success for outcome k s.t. $\sum_k \theta_k = 1$.
Usually, each x_k is a non-negative integer.

Multinomial distribution

Given $\boldsymbol{\theta} = [\theta_k]$, we can also write

$$\begin{aligned} p\left(X = x \mid \boldsymbol{\theta}\right) &= \text{Multinomial}\left(\boldsymbol{x} \mid N, \boldsymbol{\theta}\right) \\ X &\sim \text{Multinomial}\left(N, \boldsymbol{\theta}\right). \end{aligned}$$

Multinomial theorem

Theorem

Given positive integers N and K. Let $\mathcal{A}$ be a set of vectors $\boldsymbol{x} = [x_1, \ldots, x_K]$ s.t. each x_k is a nonnegative integer and $\sum_k = N$. Then, for any real numbers $p_1, \ldots, p_K$,

$$\left(\sum_k p_k \right)^N = \sum_{\boldsymbol{x} \in \mathcal{A}} \frac{N!}{\prod_k x_k!} \prod_k p_k^{x_k}.$$

Applying the multinomial theorem with on $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$, we get

$$\frac{N!}{\prod_k x_k!} \prod_k \theta_k^{x_k} = \left(\sum_k \theta_k \right)^N = 1^N,$$

since $\sum_k \theta_k = 1$. This proves that the multinomial distribution satisfies Kolmogorov's axiom 2.

Posterior predictive distribution

Given data $\boldsymbol{x} = [x_i]$, if a model with parameter $\theta \in \Theta$ has a posterior distribution $p(\theta \mid \boldsymbol{x})$, then the posterior prediction distribution for a new observation $\tilde{x}$ is

$$p(\tilde{x} \mid \boldsymbol{x}) = \int_{\Theta} p(\tilde{x} \mid \theta)\, p(\theta \mid \boldsymbol{x})\, d\theta,$$

which follows from the application of the total law of probability.

$p(\tilde{x} \mid \boldsymbol{x})$ may be used as likelihood function to evaluate the probability of held-out observations $\tilde{x}_i$, or as a sampling distribution to sample new observations $\tilde{x}_i$.

Posterior predictive distribution

In Bayesian modelling, we often sample from $p\left(\tilde{x} \mid \boldsymbol{x}\right)$ and evaluate how the empirical distribution of posterior predictive sample $[\tilde{x}_i]$ agrees with the empirical distribution of $\boldsymbol{x}$.

Iterative update under categorical-Dirichlet model

Let $X = \{1, 2, 3, 4\}$ represent the suit (Club, Diamond, Heart, Spade) of a playing card drawn from a cut deck (i.e. $K = 4$).

$$X \sim \boldsymbol{\theta}$$
$$\boldsymbol{\theta} \sim \text{Dirichlet}\,(\boldsymbol{\alpha})$$

Draw a card, which has suit x_1, and update $p\,(\boldsymbol{\theta} \mid x_1)$. Put card back into the deck. Repeat for $x_2, \ldots x_N$. What is the probability distribution for the suit of the $N + 1$ card, X_{N+1}?

$$
\begin{aligned}
p\left(\boldsymbol{\theta}\right) &= \mathsf{Dirichlet}\left(\boldsymbol{\theta} \mid [\alpha_k]\right) \\
p\left(\boldsymbol{\theta} \mid x_1\right) &= p\left(x_1 \mid \theta\right) p\left(\theta\right) p\left(x_1\right)^{-1} \\
&= \mathsf{Dirichlet}\left(\boldsymbol{\theta} \mid [\alpha_k + I(x_1 = k)]\right) \\
p\left(\boldsymbol{\theta} \mid x_1, x_2\right) &= p\left(x_2 \mid \theta, x_1\right) p\left(\theta \mid x_1\right) p\left(x_2 \mid x_1\right)^{-1} \\
&= \mathsf{Dirichlet}\left(\boldsymbol{\theta} \mid [\alpha_k + I(x_1 = k) + I(x_2 = k)]\right) \\
&\ldots \\
p\left(\boldsymbol{\theta} \mid x_{1:N}\right) &= \frac{p\left(x_N \mid \theta, x_{1:(N-1)}\right) p\left(\theta \mid x_{1:(N-1)}\right)}{p\left(x_N \mid x_{1:(N-1)}\right)} \\
&= \mathsf{Dirichlet}\left(\boldsymbol{\theta} \mid [\alpha_k + y_k]\right),
\end{aligned}
$$

where $y_k = \sum_i^N I(x_i = k)$.

$$
\begin{aligned}
p\left(X_{N+1} = l \mid x_{1:N}\right) &= \int_{\Theta} p\left(X_{N+1} = l \mid \boldsymbol{\theta}\right)\, p\left(\boldsymbol{\theta} \mid x_{1:N}\right)\, d\boldsymbol{\theta} \\
&= \int_{\Theta} \theta_l\, p\left(\boldsymbol{\theta} \mid x_{1:N}\right)\, d\boldsymbol{\theta} \\
&= \int_0^1 \int_{\Theta_{-l}} \theta_l\, p\left(\boldsymbol{\theta} \mid x_{1:N}\right)\, d\boldsymbol{\theta}_{-l}\, d\theta_l \\
&= \int_0^1 \theta_l \int_{\Theta_{-l}} p\left(\boldsymbol{\theta} \mid x_{1:N}\right)\, d\boldsymbol{\theta}_{-l}\, d\theta_l \\
&= \int_0^1 \theta_l\, p\left(\theta_l \mid x_{1:N}\right)\, d\theta_l \quad \text{(total law of prob.)} \\
&= \mathsf{E}_{p(\theta_l \mid x_{1:N})}\left(\theta_l\right) \quad \text{(def'n of expectation)}
\end{aligned}
$$

Under Dirichlet $(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$,

$$\mathsf{E}\left(\theta_l\right) = \frac{\alpha_l}{\sum_k \alpha_k}.$$

Therefore,

$$\begin{aligned} p\left(X_{N+1} = l \mid x_{1:N}\right) &= \mathsf{E}_{p(\theta_l \mid x_{1:N})}\left(\theta_l\right) \\ &= \mathsf{E}_{\mathsf{Dirichlet}(\theta_l \mid [\alpha'_k])}\left(\theta_l\right) \\ &= \frac{\alpha_l^{(N)}}{\sum_k \alpha_k^{(N)}} \end{aligned}$$

where $\alpha_k^{(N)} = \alpha_k + \sum_i^N I(x_i = k)$.

Summary

"In God we trust; all others bring data." - W. Edwards Deming

Casella & Berger 2002, section 4.6, pages 177-182.

Intended learning outcomes

- Recognize and apply categorical and multinomial distributions in statistical models
- Derive the multinomial distribution
- Derive posterior predictive distributions for statistical models