

STAT 238 - Bayesian Statistics

Lecture Twelve

Spring 2026, UC Berkeley

Aditya Guntuboyina

18 Feb 2026

Today we shall generalize the Beta-Binomial analysis from last week to Dirichlet-Multinomial inference. First we shall look at the Multinomial distribution (which is a generalization of the Binomial distribution) and the Dirichlet distribution (which is a generalization of the Beta distribution).

1 Multinomial Distribution

The multinomial distribution is a generalization of the Binomial distribution. We first recall the binomial distribution. We say that $X \sim \text{Bin}(n, p)$ if

$$\mathbb{P}\{X = x\} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

X represents the count of successes in n repetitions of a simple binary experiment with two outcomes success (probability p) and failure (probability $1-p$).

The multinomial distribution is obtained by considering n repetitions of an experiment with k outcomes (for some $k \geq 1$) with probabilities p_1, \dots, p_k (we need $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$). Let X_i denote the count of the i -th outcome among the n repetitions of the experiment (in other words, X_i is the number of times the i -th outcome happens in the n trials). The joint distribution of (X_1, \dots, X_k) is written as $\text{Multinomial}(n; p_1, \dots, p_k)$. It corresponds to the probabilities:

$$\mathbb{P}\{X_1 = x_1, \dots, X_k = x_k\} = \begin{cases} \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^k p_i^{x_i}, & \text{if } x_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^k x_i = n, \\ 0, & \text{otherwise.} \end{cases}$$

The following consequences of $(X_1, \dots, X_k) \sim \text{Multinomial}(n; p_1, \dots, p_k)$ are straightforward to see:

1. To derive the distribution of X_i i.e., $\mathbb{P}\{X_i = x\}$, we can write

$$\begin{aligned} \mathbb{P}\{X_i = x\} &= \mathbb{P}\{x \text{ trials had outcome } i, (n-x) \text{ trials had outcome not } i\} \\ &= \frac{n!}{x!(n-x)!} p_i^x (1-p_i)^{n-x}. \end{aligned}$$

In other words $X_i \sim \text{Bin}(n, p_i)$. In essence, here we consider the variant of the original experiment where, for each trial, we just record if the original outcome was 'not i ' or

i. This converts the setup to binary trials and X_i represents the number of successes (success = outcome is i) so $X_i \sim \text{Bin}(n, p_i)$.

2. The joint distribution of (X_i, X_j) (for fixed $i \neq j$) can be derived in the following way:

$$\begin{aligned} & \mathbb{P}\{X_i = x_1, X_j = x_2\} \\ &= \mathbb{P}\{x_1 \text{ trials were } i, x_2 \text{ trials were } j, n - x_1 - x_2 \text{ trials were neither } i \text{ nor } j\} \\ &= \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} p_i^{x_1} p_j^{x_2} (1 - p_i - p_j)^{n - x_1 - x_2} \end{aligned}$$

provided $x_1, x_2 \in \{0, 1, \dots, n\}$ with $x_1 + x_2 \leq n$ (otherwise, the probability equals 0).

3. It can be checked (see e.g., https://en.wikipedia.org/wiki/Multinomial_distribution) that

$$\mathbb{E}X_i = np_i, \quad \text{var}(X_i) = np_i(1 - p_i), \quad \text{and} \quad \text{cov}(X_i, X_j) = -np_i p_j \text{ if } i \neq j.$$

2 Dirichlet Distribution

The Dirichlet distribution is a generalization of the Beta distribution. Recall the $\text{Beta}(a, b)$ distribution is a distribution over $p \in [0, 1]$ corresponding to the density:

$$f(p) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1 - p)^{b-1} I\{0 \leq p \leq 1\}.$$

This density function is well-defined only when both a and b are strictly positive. However one can still define the $\text{Beta}(a, b)$ distribution even when one or both of a and b are zero. This can be done by a limiting argument:

1. $\text{Beta}(0, 0) = \lim_{\epsilon \downarrow 0} \text{Beta}(\epsilon, \epsilon) = \text{Bernoulli}(0.5)$. In other words, $\text{Beta}(0, 0)$ becomes a discrete two-point distribution assigning equal probability to 0 and 1.
2. For fixed $b > 0$, we have $\text{Beta}(0, b) = \lim_{\epsilon \downarrow 0} \text{Beta}(\epsilon, b) = \text{Bernoulli}(0) = \delta_{\{0\}}$. In other words, $\text{Beta}(0, b)$ is the point mass as 0.
3. For fixed $a > 0$, we have $\text{Beta}(a, 0) = \lim_{\epsilon \downarrow 0} \text{Beta}(a, \epsilon) = \text{Bernoulli}(1) = \delta_{\{1\}}$. In other words, $\text{Beta}(a, 0)$ is the point mass as 1.

The limiting statements above can be made rigorous by moment calculations (e.g., by showing that the moments of $\text{Beta}(\epsilon, \epsilon)$ converge to the moments of $\text{Bernoulli}(0.5)$ as $\epsilon \downarrow 0$).

The Beta distribution can be viewed as the distribution over the probabilities (p and $1 - p$) of an experiment with two outcomes. The Dirichlet distribution is the distribution over the probabilities p_1, \dots, p_k of an experiment with k outcomes. Specifically given a_1, \dots, a_k , the $\text{Dirichlet}(a_1, \dots, a_k)$ distribution corresponds to the density:

$$\frac{\Gamma(a_1 + \dots + a_k)}{\Gamma(a_1) \dots \Gamma(a_k)} p_1^{a_1-1} \dots p_k^{a_k-1} I\{p_1, \dots, p_k \geq 0, \sum_i p_i = 1\}.$$

Note that this should be viewed as density on $(k - 1)$ -dimensional space (as opposed to k -dimensional space) because of the restriction $p_1 + \dots + p_k = 1$. In other words, for every $A \subseteq \mathbb{R}^k$, we have

$$\begin{aligned} & \mathbb{P}\{(p_1, \dots, p_k) \in A\} \\ &= \frac{\Gamma(a_1 + \dots + a_k)}{\Gamma(a_1) \dots \Gamma(a_k)} \int_S p_1^{a_1-1} \dots p_{k-1}^{a_{k-1}-1} (1 - p_1 - \dots - p_{k-1})^{a_k-1} \\ & I\{(p_1, \dots, p_{k-1}, 1 - p_1 - \dots - p_{k-1}) \in A\} dp_1 \dots dp_{k-1} \end{aligned}$$

where

$$S := \{(p_1, \dots, p_{k-1}) : p_i \geq 0, p_1 + \dots + p_{k-1} \leq 1\}.$$

It can be checked that if $(p_1, \dots, p_k) \sim \text{Dirichlet}(a_1, \dots, a_k)$, then (see e.g., https://en.wikipedia.org/wiki/Dirichlet_distribution)

$$\begin{aligned} \mathbb{E}p_i &= \frac{a_i}{a_1 + \dots + a_k} \\ \text{var}(p_i) &= \left(\frac{a_i}{a_1 + \dots + a_k} \right) \left(\frac{(a_1 + \dots + a_k) - a_i}{a_1 + \dots + a_k} \right) \left(\frac{1}{a_1 + \dots + a_k + 1} \right) \\ \text{cov}(p_i, p_j) &= - \left(\frac{a_i}{a_1 + \dots + a_k} \right) \left(\frac{a_j}{a_1 + \dots + a_k} \right) \left(\frac{1}{a_1 + \dots + a_k + 1} \right). \end{aligned}$$

As for the case of the Beta, the following facts about the Dirichlet can also be proved using moment calculations:

1. $\text{Dirichlet}(0, \dots, 0) = \lim_{\epsilon \downarrow 0} \text{Dirichlet}(\epsilon, \dots, \epsilon)$ is the discrete uniform distribution on e_1, \dots, e_k with e_i is the vector which has 1 in the i -th location and 0 everywhere else i.e.,

$$\text{Dirichlet}(0, \dots, 0) = \frac{1}{k} \sum_{i=1}^k \delta_{\{e_i\}}.$$

2. If $k = 4$ and $a_2, a_4 > 0$, then $\text{Dirichlet}(0, a_2, 0, a_4) = \lim_{\epsilon \downarrow 0} \text{Dirichlet}(\epsilon, a_2, \epsilon, a_4)$ is supported on $\{(p_1, p_2, p_3, p_4) : p_1 = 0, p_2 \geq 0, p_3 = 0, p_4 \geq 0, p_1 + p_2 + p_3 + p_4 = 1\}$, and the distribution of (p_2, p_4) is $\text{Dirichlet}(a_2, a_4)$. Informally, we write

$$\text{Dirichlet}(0, a_2, 0, a_4) = \text{Dirichlet}(a_2, a_4).$$

3. More generally, let $a = (a_1, \dots, a_k)$ with each $a_i \geq 0$. Let $I = \{i : a_i > 0\}$ and let m be the cardinality of I . Then $\text{Dirichlet}(a) := \lim_{\epsilon \downarrow 0} \text{Dirichlet}(a_1 + \epsilon, \dots, a_k + \epsilon)$ satisfies the following.

- a) The support of $\text{Dirichlet}(a)$ equals $\{(p_1, \dots, p_k) : p_i \geq 0, \sum_{i=1}^k p_i = 1, p_i = 0 \text{ for all } i \notin I\}$.
- b) If $p \sim \text{Dirichlet}(a)$, then the subvector $p_i, i \in I$ satisfies:

$$(p_i, i \in I) \sim \text{Dirichlet}(a_i, i \in S).$$

Note that the above is a Dirichlet distribution for an m -dimensional probability vector.

3 Dirichlet Prior and Multinomial Likelihood

We saw the following basic fact in Lecture 8:

$$p \sim \text{Beta}(a, b) \quad \text{and} \quad X | p \sim \text{Bin}(n, p) \implies p | X = x \sim \text{Beta}(a + x, b + n - x).$$

The generalization of this is:

$$\begin{aligned} (p_1, \dots, p_k) \sim \text{Dirichlet}(a_1, \dots, a_k) \quad \text{and} \quad (X_1, \dots, X_k) | (p_1, \dots, p_k) \sim \text{Multinomial}(n; p_1, \dots, p_k) \\ \implies (p_1, \dots, p_k) | X_1 = x_1, \dots, X_k = x_k \sim \text{Dirichlet}(a_1 + x_1, a_2 + x_2, \dots, a_k + x_k). \end{aligned}$$

So the posterior mean estimate of p_i is given by:

$$\mathbb{E}(p_i | X_1 = x_1, \dots, X_k = x_k) = \frac{a_i + x_i}{(a_1 + x_1) + \dots + (a_k + x_k)} = \frac{a_i + x_i}{n + a_1 + \dots + a_k}$$

where we used $x_1 + \dots + x_k = n$.

In the special case where all $a_i = 0$, then the posterior mean of p_i is simply x_i/n ; so we are estimating the probability of the i -th outcome by simply the proportion of times the i -th outcome appeared in the n trials. This is just the frequentist MLE. So the posterior mean estimate coincides with the frequentist MLE when the prior is $\text{Dirichlet}(0, \dots, 0)$.

More precisely, the posterior corresponding to the $\text{Dirichlet}(0, \dots, 0)$ prior equals $\text{Dirichlet}(x_1, \dots, x_k)$. A key property of this posterior is that if $x_i = 0$ for some i , then, by the properties of Dirichlet distributions with some zero hyperparameters discussed in the previous section, the posterior places all its mass on the set $\{p_i = 0\}$. In other words, any outcome that is not observed in the sample is assigned zero probability under the posterior and is therefore completely excluded from future consideration.