CHAPTER 3 Categorical Attributes

In this chapter we present methods to analyze categorical attributes. Because categorical attributes have only symbolic values, many of the arithmetic operations cannot be performed directly on the symbolic values. However, we can compute the frequencies of these values and use them to analyze the attributes.

3.1 UNIVARIATE ANALYSIS

We assume that the data consists of values for a single categorical attribute, X. Let the domain of X consist of m symbolic values $dom(X) = \{a_1, a_2, \ldots, a_m\}$. The data $\mathbf{D}$ is thus an $n \times 1$ symbolic data matrix given as

$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where each point $x_i \in dom(X)$.

3.1.1 Bernoulli Variable

Let us first consider the case when the categorical attribute X has domain $\{a_1, a_2\}$, with $m = 2$. We can model X as a Bernoulli random variable, which takes on two distinct values, 1 and 0, according to the mapping

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

The probability mass function (PMF) of X is given as

$$P(X = x) = f(x) = \begin{cases} p_1 & \text{if } x = 1 \\ p_0 & \text{if } x = 0 \end{cases}$$

where p_1 and p_0 are the parameters of the distribution, which must satisfy the condition

$$p_1 + p_0 = 1$$

Because there is only one free parameter, it is customary to denote $p_1 = p$, from which it follows that $p_0 = 1 - p$. The PMF of Bernoulli random variable X can then be written compactly as

$$P(X = x) = f(x) = p^x (1-p)^{1-x}$$

We can see that $P(X = 1) = p^1(1-p)^0 = p$ and $P(X = 0) = p^0(1-p)^1 = 1 - p$, as desired.

Mean and Variance

The expected value of X is given as

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

and the variance of X is given as

$$\begin{aligned} \sigma^2 = var(X) &= E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p - p^2 = p(1-p) \end{aligned} \tag{3.1}$$

Sample Mean and Variance

To estimate the parameters of the Bernoulli variable X, we assume that each symbolic point has been mapped to its binary value. Thus, the set $\{x_1, x_2, \ldots, x_n\}$ is assumed to be a random sample drawn from X (i.e., each x_i is IID with X).

The sample mean is given as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p} \tag{3.2}$$

where n_1 is the number of points with $x_i = 1$ in the random sample (equal to the number of occurrences of symbol a_1).

Let $n_0 = n - n_1$ denote the number of points with $x_i = 0$ in the random sample. The sample variance is given as

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \\ &= \frac{n_1}{n}(1-\hat{p})^2 + \frac{n-n_1}{n}(-\hat{p})^2 \\ &= \hat{p}(1-\hat{p})^2 + (1-\hat{p})\hat{p}^2 \\ &= \hat{p}(1-\hat{p})(1-\hat{p}+\hat{p}) \\ &= \hat{p}(1-\hat{p}) \end{aligned}$$

The sample variance could also have been obtained directly from Eq. (3.1), by substituting $\hat{p}$ for p.

Example 3.1. Consider the `sepal length` attribute (X_1) for the Iris dataset in Table 1.1. Let us define an Iris flower as `Long` if its sepal length is in the range $[7, \infty]$, and `Short` if its sepal length is in the range $[-\infty, 7)$. Then X_1 can be treated as a categorical attribute with domain {`Long`, `Short`}. From the observed sample of size $n = 150$, we find 13 long Irises. The sample mean of X_1 is

$$\hat{\mu} = \hat{p} = 13/150 = 0.087$$

and its variance is

$$\hat{\sigma}^2 = \hat{p}(1 - \hat{p}) = 0.087(1 - 0.087) = 0.087 \cdot 0.913 = 0.079$$

Binomial Distribution: Number of Occurrences

Given the Bernoulli variable X, let $\{x_1, x_2, \ldots, x_n\}$ denote a random sample of size n drawn from X. Let N be the random variable denoting the number of occurrences of the symbol a_1 (value $X = 1$) in the sample. N has a binomial distribution, given as

$$f(N = n_1 \,|\, n, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \tag{3.3}$$

In fact, N is the sum of the n independent Bernoulli random variables x_i IID with X, that is, $N = \sum_{i=1}^{n} x_i$. By linearity of expectation, the mean or expected number of occurrences of symbol a_1 is given as

$$\mu_N = E[N] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i] = \sum_{i=1}^{n} p = np$$

Because x_i are all independent, the variance of N is given as

$$\sigma_N^2 = var(N) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} p(1 - p) = np(1 - p)$$

Example 3.2. Continuing with Example 3.1, we can use the estimated parameter $\hat{p} = 0.087$ to compute the expected number of occurrences N of `Long` sepal length Irises via the binomial distribution:

$$E[N] = n\hat{p} = 150 \cdot 0.087 = 13$$

In this case, because p is estimated from the sample via $\hat{p}$, it is not surprising that the expected number of occurrences of long Irises coincides with the actual occurrences. However, what is more interesting is that we can compute the variance in the number of occurrences:

$$var(N) = n\hat{p}(1 - \hat{p}) = 150 \cdot 0.079 = 11.9$$

As the sample size increases, the binomial distribution given in Eq. 3.3 tends to a normal distribution with $\mu = 13$ and $\sigma = \sqrt{11.9} = 3.45$ for our example. Thus, with confidence greater than 95% we can claim that the number of occurrences of a_1 will lie in the range $\mu \pm 2\sigma = [9.55, 16.45]$, which follows from the fact that for a normal distribution 95.45% of the probability mass lies within two standard deviations from the mean (see Section 2.5.1).

3.1.2 Multivariate Bernoulli Variable

We now consider the general case when X is a categorical attribute with domain $\{a_1, a_2, \ldots, a_m\}$. We can model X as an m-dimensional Bernoulli random variable $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, where each A_i is a Bernoulli variable with parameter p_i denoting the probability of observing symbol a_i. However, because X can assume only one of the symbolic values at any one time, if $X = a_i$, then $A_i = 1$, and $A_j = 0$ for all $j \neq i$. The range of the random variable $\mathbf{X}$ is thus the set $\{0, 1\}^m$, with the further restriction that if $X = a_i$, then $\mathbf{X} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith standard basis vector $\mathbf{e}_i \in \mathbb{R}^m$ given as

$$\mathbf{e}_i = (\overbrace{0, \ldots, 0}^{i-1}, 1, \overbrace{0, \ldots, 0}^{m-i})^T$$

In $\mathbf{e}_i$, only the ith element is 1 ($e_{ii} = 1$), whereas all other elements are zero ($e_{ij} = 0, \forall j \neq i$).

This is precisely the definition of a *multivariate Bernoulli variable*, which is a generalization of a Bernoulli variable from two outcomes to m outcomes. We thus model the categorical attribute X as a multivariate Bernoulli variable $\mathbf{X}$ defined as

$$\mathbf{X}(v) = \mathbf{e}_i \text{ if } v = a_i$$

The range of $\mathbf{X}$ consists of m distinct vector values $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$, with the PMF of $\mathbf{X}$ given as

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = p_i$$

where p_i is the probability of observing value a_i. These parameters must satisfy the condition

$$\sum_{i=1}^{m} p_i = 1$$

The PMF can be written compactly as follows:

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = \prod_{j=1}^{m} p_j^{e_{ij}} \tag{3.4}$$

Because $e_{ii} = 1$, and $e_{ij} = 0$ for $j \neq i$, we can see that, as expected, we have

$$f(\mathbf{e}_i) = \prod_{j=1}^{m} p_j^{e_{ij}} = p_1^{e_{i0}} \times \cdots p_i^{e_{ii}} \cdots \times p_m^{e_{im}} = p_1^0 \times \cdots p_i^1 \cdots \times p_m^0 = p_i$$

Table 3.1. Discretized `sepal length` attribute

Bins	Domain	Counts
[4.3, 5.2]	Very Short (a_1)	$n_1 = 45$
(5.2, 6.1]	Short (a_2)	$n_2 = 50$
(6.1, 7.0]	Long (a_3)	$n_3 = 43$
(7.0, 7.9]	Very Long (a_4)	$n_4 = 12$

Example 3.3. Let us consider the `sepal length` attribute (X_1) for the Iris dataset shown in Table 1.2. We divide the sepal length into four equal-width intervals, and give each interval a name as shown in Table 3.1. We consider X_1 as a categorical attribute with domain

$$\{a_1 = \texttt{VeryShort}, a_2 = \texttt{Short}, a_3 = \texttt{Long}, a_4 = \texttt{VeryLong}\}$$

We model the categorical attribute X_1 as a multivariate Bernoulli variable $\mathbf{X}$, defined as

$$\mathbf{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \mathbf{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \mathbf{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \mathbf{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

For example, the symbolic point $x_1 = \texttt{Short} = a_2$ is represented as the vector $(0,1,0,0)^T = \mathbf{e}_2$.

Mean

The mean or expected value of $\mathbf{X}$ can be obtained as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \sum_{i=1}^{m} \mathbf{e}_i f(\mathbf{e}_i) = \sum_{i=1}^{m} \mathbf{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \mathbf{p} \tag{3.5}$$

Sample Mean

Assume that each symbolic point $x_i \in \mathbf{D}$ is mapped to the variable $\mathbf{x}_i = \mathbf{X}(x_i)$. The mapped dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is then assumed to be a random sample IID with $\mathbf{X}$. We can compute the sample mean by placing a probability mass of $\frac{1}{n}$ at each point

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \mathbf{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\mathbf{p}} \tag{3.6}$$

where n_i is the number of occurrences of the vector value $\mathbf{e}_i$ in the sample, which is equivalent to the number of occurrences of the symbol a_i. Furthermore, we have

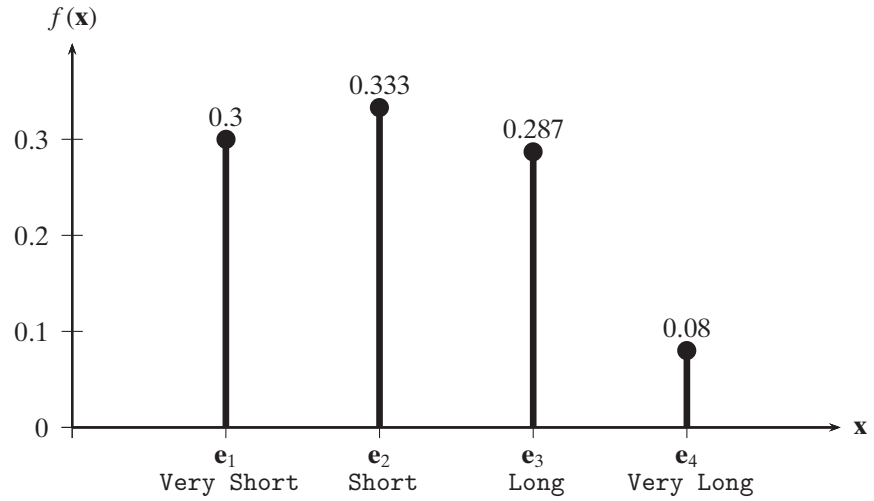

Figure 3.1. Probability mass function: `sepal length`.

$\sum_{i=1}^{m} n_i = n$, which follows from the fact that $\mathbf{X}$ can take on only m distinct values $\mathbf{e}_i$, and the counts for each value must add up to the sample size n.

Example 3.4 (Sample Mean). Consider the observed counts n_i for each of the values a_i ($\mathbf{e}_i$) of the discretized `sepal length` attribute, shown in Table 3.1. Because the total sample size is $n = 150$, from these we can obtain the estimates $\hat{p}_i$ as follows:

$$\hat{p}_1 = 45/150 = 0.3$$
$$\hat{p}_2 = 50/150 = 0.333$$
$$\hat{p}_3 = 43/150 = 0.287$$
$$\hat{p}_4 = 12/150 = 0.08$$

The PMF for $\mathbf{X}$ is plotted in Figure 3.1, and the sample mean for $\mathbf{X}$ is given as

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix}$$

Covariance Matrix

Recall that an m-dimensional multivariate Bernoulli variable is simply a vector of m Bernoulli variables. For instance, $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, where A_i is the Bernoulli variable corresponding to symbol a_i. The variance–covariance information between the constituent Bernoulli variables yields a covariance matrix for $\mathbf{X}$.

Let us first consider the variance along each Bernoulli variable A_i. By Eq. (3.1), we immediately have

$$\sigma_i^2 = var(A_i) = p_i(1 - p_i)$$

Next consider the covariance between A_i and A_j. Utilizing the identity in Eq. (2.21), we have

$$\sigma_{ij} = E[A_i A_j] - E[A_i] \cdot E[A_j] = 0 - p_i p_j = -p_i p_j$$

which follows from the fact that $E[A_i A_j] = 0$, as A_i and A_j cannot both be 1 at the same time, and thus their product $A_i A_j = 0$. This same fact leads to the negative relationship between A_i and A_j. What is interesting is that the degree of negative association is proportional to the product of the mean values for A_i and A_j.

From the preceding expressions for variance and covariance, the $m \times m$ covariance matrix for $\mathbf{X}$ is given as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_m & -p_2 p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

Notice how each row in $\mathbf{\Sigma}$ sums to zero. For example, for row i, we have

$$-p_i p_1 - p_i p_2 - \cdots + p_i(1 - p_i) - \cdots - p_i p_m = p_i - p_i \sum_{j=1}^{m} p_j = p_i - p_i = 0 \tag{3.7}$$

Because $\mathbf{\Sigma}$ is symmetric, it follows that each column also sums to zero.

Define $\mathbf{P}$ as the $m \times m$ diagonal matrix:

$$\mathbf{P} = \text{diag}(\mathbf{p}) = \text{diag}(p_1, p_2, \ldots, p_m) = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_m \end{pmatrix}$$

We can compactly write the covariance matrix of $\mathbf{X}$ as

$$\mathbf{\Sigma} = \mathbf{P} - \mathbf{p} \cdot \mathbf{p}^T \tag{3.8}$$

Sample Covariance Matrix

The sample covariance matrix can be obtained from Eq. (3.8) in a straightforward manner:

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \tag{3.9}$$

where $\widehat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}})$, and $\hat{\mathbf{p}} = \hat{\boldsymbol{\mu}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)^T$ denotes the empirical probability mass function for $\mathbf{X}$.

Example 3.5. Returning to the discretized `sepal length` attribute in Example 3.4, we have $\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (0.3, 0.333, 0.287, 0.08)^T$. The sample covariance matrix is given as

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} &= \widehat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \\
&= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.3 & 0.333 & 0.287 & 0.08 \end{pmatrix} \\
&= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.09 & 0.1 & 0.086 & 0.024 \\ 0.1 & 0.111 & 0.096 & 0.027 \\ 0.086 & 0.096 & 0.082 & 0.023 \\ 0.024 & 0.027 & 0.023 & 0.006 \end{pmatrix} \\
&= \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}
\end{aligned}$$

One can verify that each row (and column) in $\widehat{\boldsymbol{\Sigma}}$ sums to zero.

It is worth emphasizing that whereas the modeling of categorical attribute X as a multivariate Bernoulli variable, $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, makes the structure of the mean and covariance matrix explicit, the same results would be obtained if we simply treat the mapped values $\mathbf{X}(x_i)$ as a new $n \times m$ binary data matrix, and apply the standard definitions of the mean and covariance matrix from multivariate numeric attribute analysis (see Section 2.3). In essence, the mapping from symbols a_i to binary vectors $\mathbf{e}_i$ is the key idea in categorical attribute analysis.

Example 3.6. Consider the sample $\mathbf{D}$ of size $n = 5$ for the `sepal length` attribute X_1 in the Iris dataset, shown in Table 3.2a. As in Example 3.1, we assume that X_1 has only two categorical values {`Long`, `Short`}. We model X_1 as the multivariate Bernoulli variable $\mathbf{X}_1$ defined as

$$\mathbf{X}_1(v) = \begin{cases} \mathbf{e}_1 = (1,0)^T & \text{if } v = \texttt{Long}(a_1) \\ \mathbf{e}_2 = (0,1)^T & \text{if } v = \texttt{Short}(a_2) \end{cases}$$

The sample mean [Eq. (3.6)] is

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (2/5, 3/5)^T = (0.4, 0.6)^T$$

and the sample covariance matrix [Eq. (3.9)] is

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T &= \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \\
&= \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{pmatrix} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}
\end{aligned}$$

Table 3.2. (a) Categorical dataset. (b) Mapped binary dataset. (c) Centered dataset.

(a)

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

(b)

	A_1	A_2
$\mathbf{x}_1$	0	1
$\mathbf{x}_2$	0	1
$\mathbf{x}_3$	1	0
$\mathbf{x}_4$	0	1
$\mathbf{x}_5$	1	0

(c)

	Z_1	Z_2
$\mathbf{z}_1$	−0.4	0.4
$\mathbf{z}_2$	−0.4	0.4
$\mathbf{z}_3$	0.6	−0.6
$\mathbf{z}_4$	−0.4	0.4
$\mathbf{z}_5$	0.6	−0.6

To show that the same result would be obtained via standard numeric analysis, we map the categorical attribute X to the two Bernoulli attributes A_1 and A_2 corresponding to symbols Long and Short, respectively. The mapped dataset is shown in Table 3.2b. The sample mean is simply

$$\hat{\boldsymbol{\mu}} = \frac{1}{5}\sum_{i=1}^{5}\mathbf{x}_i = \frac{1}{5}(2,3)^T = (0.4, 0.6)^T$$

Next, we center the dataset by subtracting the mean value from each attribute. After centering, the mapped dataset is as shown in Table 3.2c, with attribute Z_i as the centered attribute A_i. We can compute the covariance matrix using the inner-product form [Eq. (2.30)] on the centered column vectors. We have

$$\sigma_1^2 = \frac{1}{5}Z_1^T Z_1 = 1.2/5 = 0.24$$

$$\sigma_2^2 = \frac{1}{5}Z_2^T Z_2 = 1.2/5 = 0.24$$

$$\sigma_{12} = \frac{1}{5}Z_1^T Z_2 = -1.2/5 = -0.24$$

Thus, the sample covariance matrix is given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

which matches the result obtained by using the multivariate Bernoulli modeling approach.

Multinomial Distribution: Number of Occurrences

Given a multivariate Bernoulli variable $\mathbf{X}$ and a random sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn from $\mathbf{X}$. Let N_i be the random variable corresponding to the number of occurrences of symbol a_i in the sample, and let $\mathbf{N} = (N_1, N_2, \ldots, N_m)^T$ denote the vector random variable corresponding to the joint distribution of the number of occurrences over all the symbols. Then $\mathbf{N}$ has a multinomial distribution, given as

$$f\big(\mathbf{N} = (n_1, n_2, \ldots, n_m) \mid \mathbf{p}\big) = \binom{n}{n_1 n_2 \ldots n_m}\prod_{i=1}^{m} p_i^{n_i}$$