Chapter 3

Categorical Attributes

In this chapter we present methods to analyze categorical attributes. Since 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 those 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 $dom(X) = \{a_1, a_2, \cdots, a_m\}$. The data $\mathbf{D}$ is thus a $n \times 1$ 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

$$p_1 + p_0 = 1$$

Since 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

$$P(X = x) = f(x) = p^x(1-p)^{1-x} \tag{3.1}$$

Mean and Variance The expected value of X is given as

$$\mu = E[x] = 1 \cdot p + 0 \cdot (1-p) = p \tag{3.2}$$

and the variance of X is given as

$$\begin{aligned} \sigma^2 = var(X) &= E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p - p^2 = p(1-p) \end{aligned} \tag{3.3}$$

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 variance is given as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p} \tag{3.4}$$

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.

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \\ &= \frac{n_1}{n}(1-\hat{p})^2 + \frac{n-n_1}{n}(-\hat{p})^2 \\ &= \hat{p}(1-\hat{p})^2 - (1-\hat{p})\hat{p}^2 \\ &= \hat{p}(1-\hat{p})(1-\hat{p}+\hat{p}) \\ &= \hat{p}(1-\hat{p}) \end{aligned} \tag{3.5}$$

The sample variance could also have been obtained directly from (3.3), by substituting $\hat{p}$ for p.

Example 3.1: Consider the `sepal length` attribute (X_1) in 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 $\{\texttt{Long}, \texttt{Short}\}$. 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

$$f(N = n_1 | \, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \tag{3.6}$$

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

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

Since 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 of `Long` via the binomial distribution

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

In this case, since 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 above tends to a normal distribution with $\mu = 13$ and $\sigma = \sqrt{11.9} = 3.45$. Thus with confidence over 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 2.4.1).

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 a m-dimensional Bernoulli random variable $\mathbf{X} = (A_1, A_2, \ldots, A_m)$, where each A_i is a Bernoulli variable with parameter p_i denoting the probability of observing symbol a_i. However, since 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 $j \neq i$. In other words, if $X = a_i$, then $\mathbf{X} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th standard basis vector $\mathbf{e}_i \in \mathbb{R}^m$

$$\mathbf{e}_i = (\overbrace{0, \cdots, 0}^{i-1}, 1, \overbrace{0, \cdots, 0}^{m-i})^T \tag{3.7}$$

In $\mathbf{e}_i$, only the i-th 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

$$\mathbf{X}(v) = \mathbf{e}_i \text{ if } v = a_i \tag{3.8}$$

The range of $\mathbf{X}$ thus 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

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = \prod_{j=1}^{m} p_j^{e_{ij}} \tag{3.9}$$

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

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

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)	$n_4 = 12$

Table 3.1: Discretized `sepal length` Attribute

Example 3.3: Let us consider the `sepal length` attribute (X_1) in 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

$$\boldsymbol{\mu} = E[\mathbf{X}] = \sum_{i=1}^{m} \mathbf{e}_i f(\mathbf{e}_i) = \sum_{i=1}^{m} \mathbf{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \mathbf{p} \tag{3.10}$$

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

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \mathbf{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\mathbf{p}} \tag{3.11}$$

where n_i is the number of occurrences of 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 $\sum_{i=1}^{m} n_i = n$, 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.

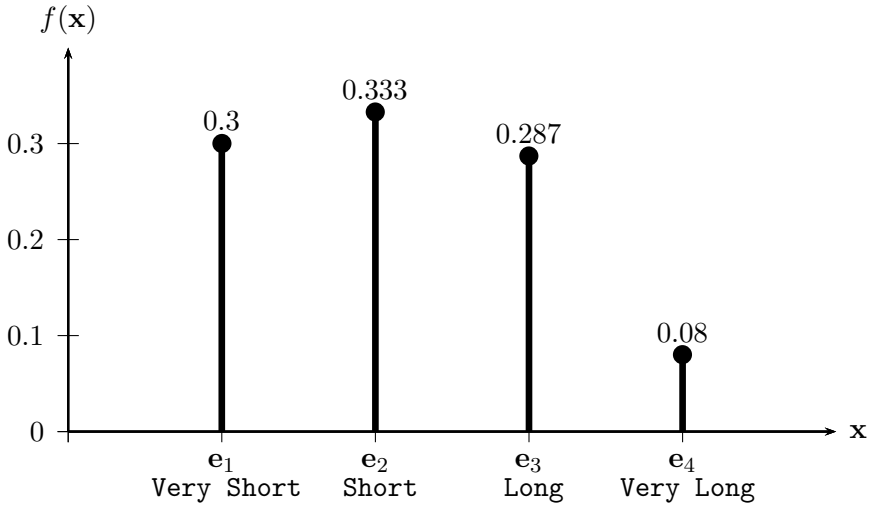

Figure 3.1: Probability Mass Function: `sepal length`

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. Since

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 a 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 bX.

Let us first consider the variance along each Bernoulli variable A_i. By (3.3), 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 (2.32), 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 $E[A_i A_j] = 0$, since 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 above expressions for variance and covariance, the $m \times m$ covariance

matrix for $\mathbf{X}$ is given as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_m & -p_2p_m & \cdots & p_m(1-p_m) \end{pmatrix} \tag{3.12}$$

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

$$-p_ip_1 - p_ip_2 - \cdots + p_i(1-p_i) - \cdots - p_ip_m = p_i - p_i\sum_{j=1}^{m} p_j = p_i - p_i = 0 \tag{3.13}$$

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

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

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

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

$$\boldsymbol{\Sigma} = \mathbf{P} - \mathbf{p} \cdot \mathbf{p}^T \tag{3.14}$$

Sample Covariance Matrix The sample covariance matrix can be obtained from (3.14) in a straightforward manner

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \tag{3.15}$$

where $\widehat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) = \text{diag}(\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_m)$ 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.

	X
x_1	`Short`
x_2	`Short`
x_3	`Long`
x_4	`Short`
x_5	`Long`

(a)

	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

(b)

	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

(c)

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

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 $\{\texttt{Long}, \texttt{Short}\}$. We model X_1 as the multivariate Bernoulli variable $\mathbf{X}_1$ defined as

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

According to (3.11) and (3.15), the sample mean is

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

and the sample covariance matrix is

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

To show that the same results 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 (2.50) on the centered column vectors. We have

$$\begin{aligned} \sigma_1^2 &= \frac{1}{5} Z_1^T Z_1 = 1.2/5 = 0.24 \\ \sigma_2^2 &= \frac{1}{5} Z_2^T Z_2 = 1.2/5 = 0.24 \\ \sigma_{12} &= \frac{1}{5} Z_1^T Z_2 = -1.2/5 = -0.24 \end{aligned}$$

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 results above 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

$$f(\mathbf{N} = (n_1, n_2, \ldots, n_m)|\mathbf{p}) = \binom{n}{n_1 n_2 \ldots n_m} \prod_{i=1}^{m} p_i^{n_i} \tag{3.16}$$

We can see that this is a direct generalization of the binomial distribution in (3.6). The term

$$\binom{n}{n_1 n_2 \ldots n_m} = \frac{n!}{n_1! n_2! \ldots n_m!}$$

denotes the number of ways of choosing n_i occurrences of each symbol a_i from a sample of size n, with $\sum_{i=1}^{m} n_i = n$.

The mean and covariance matrix of $\mathbf{N}$ are given as n times the mean and covariance matrix of $\mathbf{X}$. That is, the mean of $\mathbf{N}$ is given as

$$\boldsymbol{\mu}_{\mathbf{N}} = E[\mathbf{N}] = nE[\mathbf{X}] = n \cdot \boldsymbol{\mu} = n \cdot \mathbf{p} = \begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix} \tag{3.17}$$

and its covariance matrix is given as

$$\boldsymbol{\Sigma}_{\mathbf{N}} = n \cdot (\mathbf{P} - \mathbf{p}\mathbf{p}^T) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_m & -np_2p_m & \cdots & np_m(1-p_m) \end{pmatrix} \tag{3.18}$$

Likewise the sample mean and covariance matrix for $\mathbf{N}$ are given as

$$\hat{\boldsymbol{\mu}}_{\mathbf{N}} = n\hat{\mathbf{p}} \qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = n(\widehat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T) \tag{3.19}$$

3.2 Bivariate Analysis

Assume that the data comprises two categorical attributes, X_1 and X_2, with

$$dom(X_1) = \{a_{11}, a_{12}, \cdots, a_{1m_1}\}$$
$$dom(X_2) = \{a_{21}, a_{22}, \cdots, a_{2m_2}\}$$

We are given n categorical points of the form $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ with $x_{i1} \in dom(X_1)$ and $x_{i2} \in dom(X_2)$. The dataset is thus a $n \times 2$ symbolic data matrix

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

We can model X_1 and X_2 as multivariate Bernoulli variables $\mathbf{X}_1$ and $\mathbf{X}_2$ with dimensions m_1 and m_2, respectively. The probability mass functions for $\mathbf{X}_1$ and $\mathbf{X}_2$ are given according to (3.9)

$$P(\mathbf{X}_1 = \mathbf{e}_{1i}) = f_1(\mathbf{e}_{1i}) = p_i^1 = \prod_{k=1}^{m_1} (p_i^1)^{e_{ik}^1}$$
$$P(\mathbf{X}_2 = \mathbf{e}_{2j}) = f_2(\mathbf{e}_{2j}) = p_j^2 = \prod_{k=1}^{m_2} (p_j^2)^{e_{jk}^2}$$

where $\mathbf{e}_{1i}$ is the i-th standard basis vector in $\mathbb{R}^{m_1}$ (for attribute X_1) whose k-th component is e_{ik}^1, and $\mathbf{e}_{2j}$ is the j-th standard basis vector in $\mathbb{R}^{m_2}$ (for attribute X_2) whose k-th component is e_{jk}^2. Further the parameter p_i^1 denotes the probability of observing symbol a_{1i}, and p_j^2 denotes the probability of observing symbol a_{2j}. Together they must satisfy the conditions: $\sum_{i=1}^{m_1} p_i^1 = 1$ and $\sum_{j=1}^{m_2} p_j^2 = 1$.

The joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ is modeled as the $d' = m_1 + m_2$ dimensional vector variable $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, specified by the mapping

$$\mathbf{X}\left((v_1, v_2)^T\right) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \mathbf{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix} \tag{3.20}$$

provided that $v_1 = a_{1i}$ and $v_2 = a_{2j}$. The range of $\mathbf{X}$ thus consists of $m_1 \times m_2$ distinct pairs of vector values $\{(\mathbf{e}_{1i}, \mathbf{e}_{2j})\}$, with $1 \le i \le m_1$ and $1 \le j \le m_2$. The joint PMF of $\mathbf{X}$ is given as

$$P\big(\mathbf{X} = (\mathbf{e}_{1i}, \mathbf{e}_{2j})\big) = f(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = p_{ij} = \prod_{r=1}^{m_1} \prod_{s=1}^{m_2} p_{ij}^{e_{ir}^1 \cdot e_{js}^2}$$

where p_{ij} the probability of observing the symbol pair (a_{1i}, a_{2j}). These probability parameters must satisfy the condition $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$. The joint PMF for $\mathbf{X}$ can be expressed as the $m_1 \times m_2$ matrix

$$\mathbf{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m_2} \\ p_{21} & p_{22} & \cdots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} & p_{m_1 2} & \cdots & p_{m_1 m_2} \end{pmatrix} \tag{3.21}$$

bins	domain	counts
$[2.0, 2.8]$	Short (a_1)	47
$(2.8, 3.6]$	Medium (a_2)	88
$(3.6, 4.4]$	Long (a_3)	15

Table 3.3: Discretized `sepal width` Attribute

Example 3.7: Consider the discretized `sepal length` attribute (X_1) in Table 3.1. We also discretize the `sepal width` attribute (X_2) into three values as shown in Table 3.3. We thus have

$$dom(X_1) = \{a_{11} = \texttt{VeryShort}, a_{12} = \texttt{Short}, a_{13} = \texttt{Long}, a_{14} = \texttt{VeryLong}\}$$
$$dom(X_2) = \{a_{21} = \texttt{Short}, a_{22} = \texttt{Medium}, a_{23} = \texttt{Long}\}$$

The symbolic point $\mathbf{x} = (\texttt{Short}, \texttt{Long}) = (a_{12}, a_{23})$, is mapped to the vector

$$\mathbf{X}(\mathbf{x}) = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{23} \end{pmatrix} = (0, 1, 0, 0 \mid 0, 0, 1)^T \in \mathbb{R}^7$$

where we use $\mid$ to demarcate the two sub-vectors $\mathbf{e}_{12} = (0, 1, 0, 0)^T \in \mathbb{R}^4$ and $\mathbf{e}_{23} = (0, 0, 1)^T \in \mathbb{R}^3$, corresponding to symbolic attributes `sepal length` and `sepal width`, respectively.

Mean The bivariate mean can easily be generalized from (3.10), as follows

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}\right] = \begin{pmatrix} E[\mathbf{X}_1] \\ E[\mathbf{X}_2] \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix} \tag{3.22}$$

where $\boldsymbol{\mu}_1 = \mathbf{p}_1 = (p_1^1, \cdots, p_{m_1}^1)^T$ and $\boldsymbol{\mu}_2 = \mathbf{p}_2 = (p_1^2, \cdots, p_{m_2}^2)^T$ are the mean vectors for $\mathbf{X}_1$ and $\mathbf{X}_2$. $\mathbf{p}_1$ and $\mathbf{p}_2$ also represent the probability mass functions for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Sample Mean The sample mean can also be generalized from (3.11), by placing a probability mass of $\frac{1}{n}$ at each point

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{m_1} n_i^1 \mathbf{e}_{1i}\\ \sum_{i=1}^{m_2} n_i^2 \mathbf{e}_{2i}\end{pmatrix} = \frac{1}{n}\begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \\ n_1^2 \\ \vdots \\ n_{m_2}^2\end{pmatrix} = \begin{pmatrix}\hat{p}_1^1 \\ \vdots \\ \hat{p}_{m_1}^1 \\ \hat{p}_1^2 \\ \vdots \\ \hat{p}_{m_2}^2\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2\end{pmatrix} = \begin{pmatrix}\hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2\end{pmatrix} \quad (3.23)$$

where n_j^i is the observed frequency of symbol a_{ij} in the sample of size n, and $\hat{\boldsymbol{\mu}}_i = \hat{\mathbf{p}}_i = (p_1^i, p_2^i, \ldots, p_{m_i}^i)^T$ is the sample mean vector for $\mathbf{X}_i$. $\hat{\mathbf{p}}_i$ is also the empirical PMF for attribute $\mathbf{X}_i$.

Covariance Matrix The covariance matrix for $\mathbf{X}$ is the $d' \times d' = (m_1 + m_2) \times (m_1 + m_2)$ matrix given as

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22}\end{pmatrix} \quad (3.24)$$

where $\boldsymbol{\Sigma}_{11}$ is the $m_1 \times m_1$ covariance matrix for $\mathbf{X}_1$, and $\boldsymbol{\Sigma}_{22}$ is the $m_2 \times m_2$ covariance matrix for $\mathbf{X}_2$, which can be computed using (3.14). That is

$$\boldsymbol{\Sigma}_{11} = \mathbf{P}_1 - \mathbf{p}_1\mathbf{p}_1^T$$
$$\boldsymbol{\Sigma}_{22} = \mathbf{P}_2 - \mathbf{p}_2\mathbf{p}_2^T$$

where $\mathbf{P}_1 = \text{diag}(\mathbf{p}_1)$ and $\mathbf{P}_2 = \text{diag}(\mathbf{p}_2)$. Further, $\boldsymbol{\Sigma}_{12}$ is the $m_1 \times m_2$ covariance matrix between variables $\mathbf{X}_1$ and $\mathbf{X}_2$, given as

$$\begin{aligned}\boldsymbol{\Sigma}_{12} &= E[(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T] \\ &= E[\mathbf{X}_1\mathbf{X}_2^T] - E[\mathbf{X}_1]E[\mathbf{X}_2]^T \\ &= \mathbf{P}_{12} - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T \\ &= \mathbf{P}_{12} - \mathbf{p}_1\mathbf{p}_2^T \\ &= \begin{pmatrix} p_{11} - p_1^1 p_1^2 & p_{12} - p_1^1 p_2^2 & \cdots & p_{1m_2} - p_1^1 p_{m_2}^2 \\ p_{21} - p_2^1 p_1^2 & p_{22} - p_2^1 p_2^2 & \cdots & p_{2m_2} - p_2^1 p_{m_2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} - p_{m_1}^1 p_1^2 & p_{m_1 2} - p_{m_1}^1 p_2^2 & \cdots & p_{m_1 m_2} - p_{m_1}^1 p_{m_2}^2\end{pmatrix}\end{aligned} \quad (3.25)$$

where $\mathbf{P}_{12}$ represents the joint PMF for $\mathbf{X}$ given in (3.21).

Incidentally, each row and each column of $\mathbf{\Sigma}_{12}$ sum to zero. For example, consider row i and column j

$$\begin{aligned}\sum_{k=1}^{m_2}(p_{ik} - p_i^1 p_k^2) &= \left(\sum_{k=1}^{m_2} p_{ik}\right) - p_i^1 = p_i^1 - p_i^1 = 0 \\ \sum_{k=1}^{m_1}(p_{kj} - p_k^1 p_j^2) &= \left(\sum_{k=1}^{m_1} p_{kj}\right) - p_j^2 = p_j^2 - p_j^2 = 0\end{aligned} \tag{3.26}$$

which follows from the fact that summing the joint mass function over all values of $\mathbf{X}_2$, yields the marginal distribution of $\mathbf{X}_1$, and summing it over all values of $\mathbf{X}_1$ yields the marginal distribution for $\mathbf{X}_j$. Combined with the fact that $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ also have row and column sums equalling zero via (3.13), the full covariance matrix $\mathbf{\Sigma}$ has rows and columns that sum up to zero.

Sample Covariance Matrix The sample covariance matrix is given as

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{11} & \widehat{\mathbf{\Sigma}}_{12} \\ \widehat{\mathbf{\Sigma}}_{12}^T & \widehat{\mathbf{\Sigma}}_{22} \end{pmatrix} \tag{3.27}$$

where

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{11} &= \widehat{\mathbf{P}}_1 - \hat{\mathbf{p}}_1 \hat{\mathbf{p}}_1^T \\ \widehat{\mathbf{\Sigma}}_{22} &= \widehat{\mathbf{P}}_2 - \hat{\mathbf{p}}_2 \hat{\mathbf{p}}_2^T \\ \widehat{\mathbf{\Sigma}}_{12} &= \widehat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1 \hat{\mathbf{p}}_2^T\end{aligned}$$

Here $\widehat{\mathbf{P}}_1 = \text{diag}(\hat{\mathbf{p}}_1)$ and $\widehat{\mathbf{P}}_2 = \text{diag}(\hat{\mathbf{p}}_2)$, and $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ specify the empirical probability mass functions for $\mathbf{X}_1$, and $\mathbf{X}_2$, respectively. Further, $\widehat{\mathbf{P}}_{12}$ specifies the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$, given as

$$\widehat{\mathbf{P}}_{12}(i,j) = \hat{f}(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = \frac{1}{n}\sum_{k=1}^{n} I_{ij}(\mathbf{x}_k) = \frac{n_{ij}}{n} = \hat{p}_{ij} \tag{3.28}$$

where I_{ij} is the indicator variable

$$I_{ij}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k1} = \mathbf{e}_{1i} \text{ and } \mathbf{x}_{k2} = \mathbf{e}_{2j} \\ 0 & \text{otherwise} \end{cases}$$

Taking the sum of $I_{ij}(\mathbf{x}_k)$ over all the n points in the sample yields the number of occurrences, n_{ij}, of the symbol pair (a_{1i}, a_{2j}) in the sample. One issue with the across-attribute covariance matrix $\widehat{\mathbf{\Sigma}}_{12}$, is the need to estimate a quadratic number of parameters. That is, we need to obtain reliable counts n_{ij} to estimate the parameters p_{ij}, for a total of $O(m_1 \times m_2)$ parameters to estimate, which can be a problem if

the categorical attributes have many symbols. On the other hand, estimating $\widehat{\mathbf{\Sigma}}_{11}$ and $\widehat{\mathbf{\Sigma}}_{22}$ requires that we estimate m_1 and m_2 parameters, corresponding to p_i^1 and p_j^2, respectively. In total, computing $\mathbf{\Sigma}$ requires the estimation of $m_1 m_2 + m_1 + m_2$ parameters.

		X_2		
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)
X_1	Very Short ($\mathbf{e}_{11}$)	7	33	5
	Short ($\mathbf{e}_{12}$)	24	18	8
	Long ($\mathbf{e}_{13}$)	13	30	0
	Very Long ($\mathbf{e}_{14}$)	3	7	2

Table 3.4: Observed Counts (n_{ij}): sepal length and sepal width

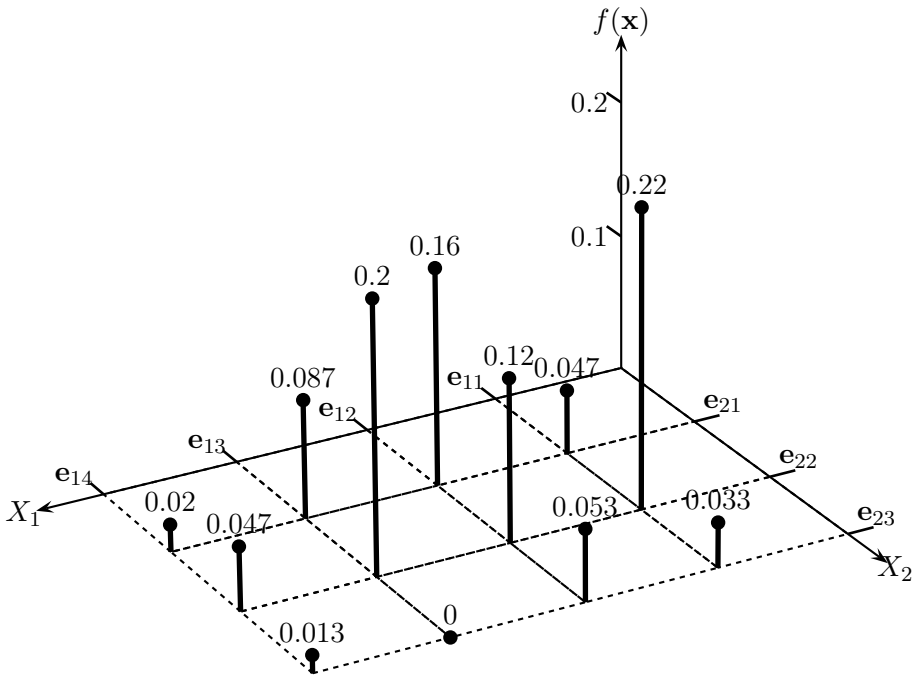

Figure 3.2: Empirical Joint Probability Mass Function: sepal length and sepal width

Example 3.8: We continue with the bivariate categorical attributes X_1 and X_2 in Example 3.7. From Example 3.4, and from the occurrence counts for each of the

values of `sepal width` in Table 3.3, we have

$$\hat{\boldsymbol{\mu}}_1 = \hat{\mathbf{p}}_1 = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \qquad \hat{\boldsymbol{\mu}}_2 = \hat{\mathbf{p}}_2 = \frac{1}{150}\begin{pmatrix} 47 \\ 88 \\ 15 \end{pmatrix} = \begin{pmatrix} 0.313 \\ 0.587 \\ 0.1 \end{pmatrix}$$

Thus the mean for $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1)^T$$

From Example 3.5 we have

$$\widehat{\boldsymbol{\Sigma}}_{11} = \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}$$

In a similar manner we can obtain

$$\widehat{\boldsymbol{\Sigma}}_{22} = \begin{pmatrix} 0.215 & -0.184 & -0.031 \\ -0.184 & 0.242 & -0.059 \\ -0.031 & -0.059 & 0.09 \end{pmatrix}$$

Next, we use the observed counts in Table 3.4 to obtain the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$ using (3.28), as plotted in Figure 3.2. From these probabilities we get

$$E[\mathbf{X}_1\mathbf{X}_2^T] = \widehat{\mathbf{P}}_{12} = \frac{1}{150}\begin{pmatrix} 7 & 33 & 5 \\ 24 & 18 & 8 \\ 13 & 30 & 0 \\ 3 & 7 & 2 \end{pmatrix} = \begin{pmatrix} 0.047 & 0.22 & 0.033 \\ 0.16 & 0.12 & 0.053 \\ 0.087 & 0.2 & 0 \\ 0.02 & 0.047 & 0.013 \end{pmatrix}$$

Furthermore, we have

$$\begin{aligned} E[\mathbf{X}_1]E[\mathbf{X}_2]^T = \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\mu}}_2^T &= \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T \\ &= \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix}\begin{pmatrix} 0.313 & 0.587 & 0.1 \end{pmatrix} \\ &= \begin{pmatrix} 0.094 & 0.176 & 0.03 \\ 0.104 & 0.196 & 0.033 \\ 0.09 & 0.168 & 0.029 \\ 0.025 & 0.047 & 0.008 \end{pmatrix} \end{aligned}$$

We can now compute the across-attribute sample covariance matrix $\widehat{\mathbf{\Sigma}}_{12}$ for $\mathbf{X}_1$ and $\mathbf{X}_2$ using (3.24), as follows

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{12} &= \widehat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T \\ &= \begin{pmatrix} -0.047 & 0.044 & 0.003 \\ 0.056 & -0.076 & 0.02 \\ -0.003 & 0.032 & -0.029 \\ -0.005 & 0 & 0.005 \end{pmatrix}\end{aligned}$$

Once can observe that each row and column in $\widehat{\mathbf{\Sigma}}_{12}$ sums to zero. Putting it all together, from $\widehat{\mathbf{\Sigma}}_{11}$, $\widehat{\mathbf{\Sigma}}_{22}$ and $\widehat{\mathbf{\Sigma}}_{12}$ we obtain the sample covariance matrix as follows

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{11} & \widehat{\mathbf{\Sigma}}_{12} \\ \widehat{\mathbf{\Sigma}}_{12}^T & \widehat{\mathbf{\Sigma}}_{22} \end{pmatrix}$$

$$= \left(\begin{array}{cccc|ccc} 0.21 & -0.1 & -0.086 & -0.024 & -0.047 & 0.044 & 0.003 \\ -0.1 & 0.222 & -0.096 & -0.027 & 0.056 & -0.076 & 0.02 \\ -0.086 & -0.096 & 0.204 & -0.023 & -0.003 & 0.032 & -0.029 \\ -0.024 & -0.027 & -0.023 & 0.074 & -0.005 & 0 & 0.005 \\ \hline -0.047 & 0.056 & -0.003 & -0.005 & 0.215 & -0.184 & -0.031 \\ 0.044 & -0.076 & 0.032 & 0 & -0.184 & 0.242 & -0.059 \\ 0.003 & 0.02 & -0.029 & 0.005 & -0.031 & -0.059 & 0.09 \end{array}\right)$$

In $\widehat{\mathbf{\Sigma}}$, each row and column also sums to zero.

3.2.1 Attribute Dependence: Contingency Analysis

Testing for the independence of the two categorical random variables X_1 and X_2 can be done via *contingency table analysis*. The main idea is to set up a hypothesis testing framework, where the null hypothesis H_0 is to assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, and the alternative hypothesis H_1 is that they are dependent. We then compute the value of the chi-square statistic χ^2 under the null hypothesis. Depending on the p-value, we either accept or reject the null hypothesis; in the latter case the attributes are considered to be dependent.

Contingency Table A contingency table for $\mathbf{X}_1$ and $\mathbf{X}_2$, is the $m_1 \times m_2$ matrix of observed counts n_{ij} for all pairs of values $(\mathbf{e}_{1i}, \mathbf{e}_{2j})$ in the given sample of size n,

defined as

$$\mathbf{N}_{12} = n \cdot \widehat{\mathbf{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

where $\widehat{\mathbf{P}}_{12}$ is the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$, computed via (3.28). The contingency table is then augmented with row and column marginal counts, as follows

$$\mathbf{N}_1 = n \cdot \hat{\mathbf{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix} \qquad \mathbf{N}_2 = n \cdot \hat{\mathbf{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix}$$

Note that the marginal row and column entries, and the sample size satisfy the following constraints

$$n_i^1 = \sum_{j=1}^{m_2} n_{ij} \qquad n_j^2 = \sum_{i=1}^{m_1} n_{ij} \qquad n = \sum_{j=1}^{m_2} n_j^1 = \sum_{i=1}^{m_1} n_i^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$$

It is worth noting that $\mathbf{N}_1$ and $\mathbf{N}_2$ have multinomial distribution with parameters $\mathbf{p}_1 = (p_1^1, \ldots, p_{m_1}^1)$ and $\mathbf{p}_2 = (p_1^2, \ldots, p_{m_2}^2)$, respectively. Furthermore, $\mathbf{N}_{12}$ also has a multinomial distribution with parameters $\mathbf{P}_{12} = \{p_{ij}\}$ (for $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$).

`sepal length` (X_1)	`sepal width` (X_2)			
	`Short` a_{21}	`Medium` a_{22}	`Long` a_{23}	Row Counts
`Very Short` (a_{11})	7	33	5	$n_1^1 = 45$
`Short` (a_{12})	24	18	8	$n_2^1 = 50$
`Long` (a_{13})	13	30	0	$n_3^1 = 43$
`Very Long` (a_{14})	3	7	2	$n_4^1 = 12$
Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

Table 3.5: Contingency Table: `sepal length` vs. `sepal width`

Example 3.9 (Contingency Table): Table 3.4 shows the observed counts for the discretized `sepal length` (X_1) and `sepal width` (X_2) attributes. Augmenting this with the row and column marginal counts and the sample size yields the final contingency table shown in Table 3.5.

χ^2 Statistic and Hypothesis Testing Under the null hypothesis, $\mathbf{X}_1$ and $\mathbf{X}_2$ are assumed to be independent, which means that their joint probability mass function is given as

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$$

Under this independence assumption, the expected frequency for each pair of values is given as

$$e_{ij} = n \cdot p_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n} \tag{3.29}$$

However, from the sample we already have the observed frequency of each pair of values, n_{ij}. We would like to determine whether there is a significant difference in the observed and expected frequencies for each pair of values, or there is no significant difference. If there is no significant difference between the expected and observed counts, then the independence assumption is valid, and we accept the null hypothesis that the attributes are independent. On the other hand if there is a significant difference, then the null hypothesis should be rejected, and we conclude that the attributes are dependent.

The χ^2 statistic quantifies the difference between observed and expected counts for each pair of values; it is defined as follows

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{3.30}$$

At this point, we need to determine the probability of obtaining the computed χ^2 value. In general, this can be rather difficult if we do not know the sampling distribution of a given statistic. Fortunately, for the χ^2 statistic, it is known that its sampling distribution follows the *chi-squared* density function, with q degrees of freedom

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} e^{-\frac{x}{2}} \tag{3.31}$$

where the Gamma function Γ is defined as

$$\Gamma(k > 0) = \int_0^\infty x^{k-1} e^{-x} dx \tag{3.32}$$

The degrees of freedom, q, represent the number of independent parameters. In the contingency table there are $m_1 \times m_2$ observed counts n_{ij}. However, note that each row i and each column j must sum to n_i^1 and n_j^2, respectively. Further, the sum of the row and column marginals must also add to n, thus we have to to remove $(m_1 + m_2)$ parameters from the independent parameter list. However, doing this

removes one of the parameters, say $n_{m_1 m_2}$, twice, so we have to add back one to the count. The total degrees of freedom is therefore

$$\begin{aligned} q &= |dom(X_1)| \times |dom(X_2)| - (|dom(X_1)| + |dom(X_2)|) + 1 \\ &= m_1 m_2 - m_1 - m_2 + 1 \\ &= (m_1 - 1)(m_2 - 1) \end{aligned}$$

p-value The *p-value* of a statistic θ is defined as the probability of obtaining a value at least as extreme as the observed value, say z, under the null hypothesis, defined as

$$p\text{-}value(z) = P(\theta \geq z) = 1 - F(\theta) \tag{3.33}$$

where $F(\theta)$ is the cumulative probability distribution for the statistic.

The p-value gives a measure of how surprising is the observed value of the statistic. If the observed value lies in a low probability region, then the value is more surprising. In general, the lower the p-value, the more surprising the observed value, and the more the grounds for rejecting the null hypothesis. The null hypothesis is rejected if the p-value is below some *significance level*, α. For example, if $\alpha = 0.01$, then we reject the null hypothesis if $p\text{-}value(z) \leq \alpha$. The significance level α corresponds to the probability of rejecting the null hypothesis when it is true. For a given significance level α, the value of the test statistic, say z, with a p-value of $p\text{-}value(z) = \alpha$, is called a *critical value*. An alternative test for rejection of the null hypothesis is to check if $\chi^2 > z$, since in that case $p\text{-}value(\chi^2) \leq p\text{-}value(z) = \alpha$.

		X_2		
		`Short` (a_{21})	`Medium` (a_{22})	`Short` (a_{23})
X_1	`Very Short` (a_{11})	14.1	26.4	4.5
	`Short` (a_{12})	15.67	29.33	5.0
	`Long` (a_{13})	13.47	25.23	4.3
	`Very Long` (a_{14})	3.76	7.04	1.2

Table 3.6: Expected Counts

Example 3.10: Consider the contingency table for `sepal length` and `sepal width` in Table 3.5. We compute the expected counts using (3.29); these counts are shown in Table 3.6. For example, we have

$$e_{11} = \frac{n_1^1 n_1^2}{n} = \frac{45 \cdot 47}{150} = \frac{2115}{150} = 14.1$$

Next we use (3.30) to compute the value of the χ^2 statistic, which is given as $\chi^2 = 21.8$.

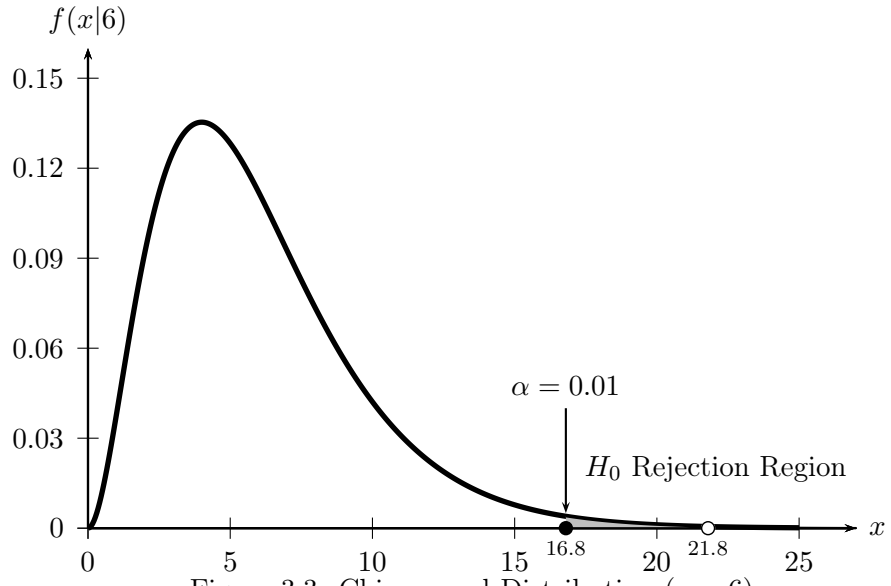

Figure 3.3: Chi-squared Distribution ($q = 6$)

Further, the number of degrees of freedom is given as

$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

The plot of the chi-squared density function with 6 degrees of freedom is shown in Figure 3.3. From the cumulative chi-squared distribution, we obtain

$$p\text{-}value(21.8) = 1 - F(21.8|6) = 1 - 0.9987 = 0.0013$$

At a significance level of $\alpha = 0.01$, we would certainly be justified in rejecting the null hypothesis, since the large value of the χ^2 statistic is indeed surprising. Further, at the 0.01 significance level, the critical value of statistic is

$$z = F^{-1}(1 - 0.01|6) = F^{-1}(0.99|6) = 16.81$$

This critical value is also shown in Figure 3.3, and we can clearly see that the observed value of 21.8 is in the rejection region, since $21.8 > z = 16.81$. In effect, we reject the null hypothesis that `sepal length` and `sepal width` are independent, and accept the alternative hypothesis that they are dependent.

3.3 Multivariate Analysis

Assume that the dataset comprises d categorical attributes X_j ($1 \leq j \leq d$) with $dom(X_j) = \{a_{j1}, a_{j2}, \cdots, a_{jm_j}\}$. We are given n categorical points of the form

$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$ with $x_{ij} \in dom(X_j)$. The dataset is thus a $n \times d$ symbolic matrix

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Each attribute X_i is modeled as a m_i-dimensional multivariate Bernoulli variable $\mathbf{X}_i$, and their joint distribution is modeled as a $d' = \sum_{j=1}^{d} m_j$ dimensional vector random variable

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_d \end{pmatrix}$$

Each categorical data point $\mathbf{v} = (v_1, v_2, \cdots, v_d)^T$ is therefore represented as a d'-dimensional binary vector

$$\mathbf{X}(\mathbf{v}) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \vdots \\ \mathbf{X}_d(v_d) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1k_1} \\ \vdots \\ \mathbf{e}_{dk_d} \end{pmatrix} \tag{3.34}$$

provided $v_i = a_{ik_i}$, the k_i-th symbol of X_i. Here $\mathbf{e}_{ik_i}$ is the k_i-th standard basis vector in $\mathbb{R}^{m_i}$.

Mean Generalizing from the bivariate case, the mean and sample mean for $\mathbf{X}$ are given as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_d \end{pmatrix} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_d \end{pmatrix} \qquad \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \vdots \\ \hat{\boldsymbol{\mu}}_d \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \vdots \\ \hat{\mathbf{p}}_d \end{pmatrix} \tag{3.35}$$

where $\mathbf{p}_i = (p_1^i, \ldots, p_{m_i}^i)^T$ is the PMF for $\mathbf{X}_i$, and $\hat{\mathbf{p}}_i = (\hat{p}_1^i, \ldots, \hat{p}_{m_i}^i)^T$ is the empirical PMF for $\mathbf{X}_i$.

Covariance Matrix The covariance matrix for $\mathbf{X}$, and its estimate from the sample, are given as the $d' \times d'$ ($d' = \sum_{i=1}^{d} m_i$) matrices

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1d} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \boldsymbol{\Sigma}_{1d}^T & \boldsymbol{\Sigma}_{2d}^T & \cdots & \boldsymbol{\Sigma}_{dd} \end{pmatrix} \qquad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\Sigma}}_{12} & \cdots & \widehat{\boldsymbol{\Sigma}}_{1d} \\ \widehat{\boldsymbol{\Sigma}}_{12}^T & \widehat{\boldsymbol{\Sigma}}_{22} & \cdots & \widehat{\boldsymbol{\Sigma}}_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \widehat{\boldsymbol{\Sigma}}_{1d}^T & \widehat{\boldsymbol{\Sigma}}_{2d}^T & \cdots & \widehat{\boldsymbol{\Sigma}}_{dd} \end{pmatrix} \tag{3.36}$$

where $\mathbf{\Sigma}_{ij}$ (and $\widehat{\mathbf{\Sigma}}_{ij}$) is the $m_i \times m_j$ covariance matrix (and its estimate) for attributes $\mathbf{X}_i$ and $\mathbf{X}_j$

$$\mathbf{\Sigma}_{ij} = \mathbf{P}_{12} - \mathbf{p}_1\mathbf{p}_2^T \qquad\qquad \widehat{\mathbf{\Sigma}}_{ij} = \widehat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T$$

Here $\mathbf{P}_{12}$ is the joint PMF and $\widehat{\mathbf{P}}_{12}$ is the empirical joint PMF for $\mathbf{X}_i$ and $\mathbf{X}_j$, which can be computed using (3.28).

Example 3.11 (Multivariate Analysis): Let us consider the three-dimensional subset of the Iris dataset, with the discretized attributes `sepal length` (X_1) and `sepal width` (X_2), and the categorical attribute `class` (X_3). The domains for X_1 and X_2 are given in Table 3.1 and Table 3.3, respectively, and $dom(X_3) = \{\texttt{iris-versicolor}, \texttt{iris-setosa}, \texttt{iris-virginica}\}$. Each value of X_3 occurs 50 times.

The categorical point $\mathbf{x} = (\texttt{Short}, \texttt{Medium}, \texttt{iris-versicolor})$, is modeled as the vector

$$\mathbf{X}(\mathbf{x}) = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{22} \\ \mathbf{e}_{31} \end{pmatrix} = (0,1,0,0 \mid 0,1,0 \mid 1,0,0)^T \in \mathbb{R}^{10}$$

From Example 3.8 and the fact that each value in $dom(X_3)$ occurs 50 times in a sample of $n = 150$, the sample mean for these three attributes is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \\ \hat{\boldsymbol{\mu}}_3 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1 \mid 0.33, 0.33, 0.33)^T$$

Using $\boldsymbol{\mu}_3 = (0.33, 0.33, 0.33)^T$, we can compute the sample covariance matrix for X_3 using (3.15)

$$\widehat{\mathbf{\Sigma}}_{33} = \begin{pmatrix} 0.222 & -0.111 & -0.111 \\ -0.111 & 0.222 & -0.111 \\ -0.111 & -0.111 & 0.222 \end{pmatrix}$$

Using (3.27) we obtain

$$\widehat{\mathbf{\Sigma}}_{13} = \begin{pmatrix} -0.067 & 0.16 & -0.093 \\ 0.082 & -0.038 & -0.044 \\ 0.011 & -0.096 & 0.084 \\ -0.027 & -0.027 & 0.053 \end{pmatrix}$$

$$\widehat{\mathbf{\Sigma}}_{23} = \begin{pmatrix} 0.076 & -0.098 & 0.022 \\ -0.042 & 0.044 & -0.002 \\ -0.033 & 0.053 & -0.02 \end{pmatrix}$$

Combined with $\widehat{\mathbf{\Sigma}}_{11}$, $\widehat{\mathbf{\Sigma}}_{22}$ and $\widehat{\mathbf{\Sigma}}_{12}$ from Example 3.8, the final sample covariance matrix is the 10×10 symmetric matrix given as

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{11} & \widehat{\mathbf{\Sigma}}_{12} & \widehat{\mathbf{\Sigma}}_{13} \\ \widehat{\mathbf{\Sigma}}_{12}^T & \widehat{\mathbf{\Sigma}}_{22} & \widehat{\mathbf{\Sigma}}_{23} \\ \widehat{\mathbf{\Sigma}}_{13}^T & \widehat{\mathbf{\Sigma}}_{23}^T & \widehat{\mathbf{\Sigma}}_{33} \end{pmatrix}$$

3.3.1 Multi-way Contingency Analysis

For multi-way dependence analysis, we have to first determine the empirical joint probability mass function for $\mathbf{X}$

$$\hat{f}(\mathbf{e}_{1i_1}, \mathbf{e}_{2i_2}, \ldots, \mathbf{e}_{di_d}) = \frac{1}{n} \sum_{k=1}^{n} I_{i_1 i_2 \ldots i_d}(\mathbf{x}_k) = \frac{n_{i_1 i_2 \ldots i_d}}{n} = \hat{p}_{i_1 i_2 \ldots i_d}$$

where $I_{i_1 i_2 \ldots i_d}$ is the indicator variable

$$I_{i_1 i_2 \ldots i_d}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } x_{k1} = \mathbf{e}_{1i_1}, x_{k2} = \mathbf{e}_{2i_2}, \ldots, x_{kd} = \mathbf{e}_{di_d} \\ 0 & \text{otherwise} \end{cases}$$

The sum of $I_{i_1 i_2 \ldots i_d}$ over all the n points in the sample yields the number of occurrences, $n_{i_1 i_2 \ldots i_d}$, of the symbolic vector $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$. Dividing the occurrences by the sample size results in the probability of observing those symbols. Using the notation $\mathbf{i} = (i_1, i_2, \ldots, i_d)$ to denote the index tuple, we can write the joint empirical PMF as the d-dimensional matrix $\widehat{\mathbf{P}}$ of size $m_1 \times m_2 \times \cdots \times m_d = \prod_{i=1}^{d} m_i$, given as

$$\widehat{\mathbf{P}}(\mathbf{i}) = \left\{\hat{p}_{\mathbf{i}}\right\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \le i_1 \le m_1, \ldots, 1 \le i_d \le m_d$$

The d-dimensional contingency table is then given as

$$\mathbf{N} = n \times \widehat{\mathbf{P}} = \left\{n_{\mathbf{i}}\right\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \le i_1 \le m_1, \ldots, 1 \le i_d \le m_d$$

The contingency table is augmented with the marginal count vectors $\mathbf{N}_i$ for all d attributes $\mathbf{X}_i$

$$\mathbf{N}_i = n\hat{\mathbf{p}}_i = \begin{pmatrix} n_1^i \\ \vdots \\ n_{m_i}^i \end{pmatrix}$$

where $\hat{\mathbf{p}}_i$ is the empirical PMF for $\mathbf{X}_i$.

χ^2-Test We can test for a d-way dependence between the d categorical attributes by assuming as the null hypothesis H_0 that they are d-way independent. The alternative hypothesis H_1 is that they are not d-way independent, i.e., they are dependent in some way. Note that d-dimensional contingency analysis indicates whether all d attributes taken together are independent or not. In general we may have to conduct k-way contingency analysis to test if any subset of k attributes are independent or not.

Under the null hypothesis, the expected number of occurrences of the symbol tuple $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$ is given as

$$e_{\mathbf{i}} = n \cdot \hat{p}_{\mathbf{i}} = n \cdot \prod_{j=1}^{d} \hat{p}_{i_j}^{j} = \frac{n_{i_1}^1 n_{i_2}^2 \ldots n_{i_d}^d}{n^{d-1}} \tag{3.37}$$

The chi-squared statistic measures the difference between the observed counts $n_{\mathbf{i}}$ and the expected counts $e_{\mathbf{i}}$

$$\chi^2 = \sum_{\mathbf{i}} \frac{(n_{\mathbf{i}} - e_{\mathbf{i}})^2}{e_{\mathbf{i}}} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \cdots \sum_{i_d=1}^{m_d} \frac{(n_{i_1,i_2,\cdots,i_d} - e_{i_1,i_2,\cdots,i_d})^2}{e_{i_1,i_2,\cdots,i_d}} \tag{3.38}$$

The χ^2 statistic follows a chi-squared density function with q degrees of freedom. For the d-way contingency table we can compute q by noting that there are ostensibly $\prod_{i=1}^{d} |dom(X_i)|$ independent parameters (the counts). However, we have to remove $\sum_{i=1}^{d} |dom(X_i)|$ degrees of freedom, since the marginal count vector along each dimension $\mathbf{X}_i$ must equal $\mathbf{N}_i$. However, doing so removes one of the parameters d times, so we need to add back $d-1$ to the free parameters count. The total number of degrees of freedom is given as

$$\begin{aligned} q &= \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1) \\ &= \prod_{i=1}^{d} m_i - \sum_{i=1}^{d} m_i + d - 1 \end{aligned} \tag{3.39}$$

To reject the null hypothesis, we have to check whether the p-value of the observed χ^2 value is smaller than the desired significance level α (say $\alpha = 0.01$) using the chi-squared density with q degrees of freedom (3.31).

Example 3.12: Consider the 3-way contingency table in Figure 3.4. It shows the observed counts for each tuple of symbols (a_{1i}, a_{2j}, a_{3k}) for the three attributes `sepal length` (X_1), `sepal width` (X_2) and `class` (X_3). From the marginal counts

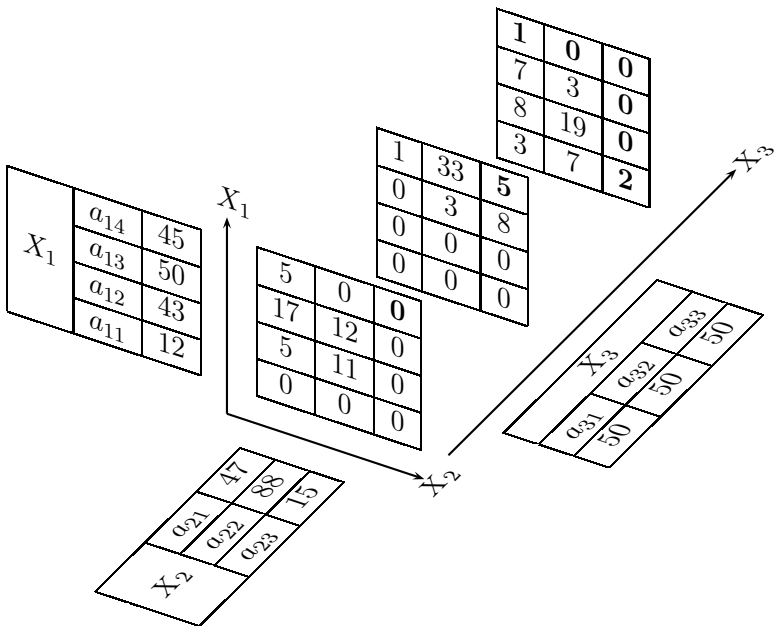

Figure 3.4: 3-way contingency table, with marginal counts along each dimension

		$X_3(a_{31}/a_{32}/a_{33})$		
		X_2		
		a_{21}	a_{22}	a_{23}
X_1	a_{11}	1.25	2.35	0.40
	a_{12}	4.49	8.41	1.43
	a_{13}	5.22	9.78	1.67
	a_{14}	4.70	8.80	1.50

Table 3.7: 3-way Expected Counts

for X_1 and X_2 in Table 3.5, and the fact that all three values of X_3 occur 50 times, we can compute the expected counts (3.37) for each cell. For instance

$$e_{(4,1,1)} = \frac{n_{14} \cdot n_{21} \cdot n_{31}}{150^2} = \frac{45 \cdot 47 \cdot 50}{150 \cdot 150} = 4.7$$

The expected counts are the same for all three values of X_3 and are given in Table 3.7.

The value of the χ^2 statistic (3.38) is given as

$$\chi^2 = 231.06$$

Using (3.39), the number of degrees of freedom is given as

$$q = 4 \cdot 3 \cdot 3 - (4 + 3 + 3) + 2 = 36 - 10 + 2 = 28$$

In Figure 3.4 the counts in bold are the dependent parameters. All other counts are independent. In fact, any 8 distinct cells could have been chosen as the dependent parameters.

For a significance level of $\alpha = 0.01$, the critical value of the chi-square distribution is $z = 48.28$. The observed value of $\chi^2 = 231.06$ is much greater than z, and is thus extremely unlikely to happen under the null hypothesis. We conclude that the three attributes are not 3-way independent, but rather there is some dependence between them. However, this example also highlights one of the pitfalls of multi-way contingency analysis. We can observe in Figure 3.4 that many of the observed counts are zero. This is due to the fact that the sample size is small, and we cannot reliably estimate all the multi-way counts. Consequently, the dependence test may not be reliable as well. Figure 3.4

3.4 Distance and Angle

With the modeling of categorical attributes as multivariate Bernoulli variables, it is possible to compute the distance or the angle between any two points $\mathbf{x}_i$ and $\mathbf{x}_j$

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{e}_{1i_1} \\ \vdots \\ \mathbf{e}_{d\, i_d} \end{pmatrix} \qquad \mathbf{x}_j = \begin{pmatrix} \mathbf{e}_{1j_1} \\ \vdots \\ \mathbf{e}_{d\, j_d} \end{pmatrix}$$

The different measures of distance and similarity rely on the number of matching and mismatching values (or symbols) across the d attributes $\mathbf{X}_k$. For instance, we can compute the number of matching values q via the dot product

$$q = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^{d} (\mathbf{e}_{ki_k})^T \mathbf{e}_{kj_k} \tag{3.40}$$

On the other hand, the number of mismatches is simply $d - q$. Also useful is the norm of each point

$$\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i = d \tag{3.41}$$

Euclidean Distance The Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given as

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j} = \sqrt{2(d - q)}$$

Thus the maximum Euclidean distance between any two points is $\sqrt{2d}$, which happens when there are no common symbols between them, i.e., when $q = 0$.

Hamming Distance The *Hamming distance* between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the number of mismatched values

$$\delta_H(\mathbf{x}_i, \mathbf{x}_j) = d - q = \frac{1}{2}\delta(\mathbf{x}_i, \mathbf{x}_j)^2$$

Hamming distance is thus equivalent to the squared Euclidean distance divided by 2.

Cosine Similarity The cosine of the angle between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given as

$$\cos\theta = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} = \frac{q}{d}$$

Jaccard Coefficient The *Jaccard Coefficient* is a commonly used similarity measure between two categorical points. It is defined as the ratio of number of matching values to the number of distinct values that appear in both $\mathbf{x}_i$ and $\mathbf{x}_j$, across the d attributes

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{q}{2(d-q)+q} = \frac{q}{2d-q} \tag{3.42}$$

Here we utilize the observation that when the two points do not match for $\mathbf{X}_k$, then there a contribution of two to the distinct symbol count, otherwise if they match then the number of distinct symbols increases by 1. Over the $d - q$ mismatches and q matches, the number of distinct symbols is $2(d-q)+q$.

Example 3.13: Consider the 3-dimensional categorical data from Example 3.11. The symbolic point (`Short`, `Medium`, `iris-versicolor`) is modeled as the vector

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{22} \\ \mathbf{e}_{31} \end{pmatrix} = (0,1,0,0 \mid 0,1,0 \mid 1,0,0)^T \in \mathbb{R}^{10}$$

and the symbolic point (`VeryShort`, `Medium`, `iris-setosa`) is modeled as

$$\mathbf{x}_2 = \begin{pmatrix} \mathbf{e}_{11} \\ \mathbf{e}_{22} \\ \mathbf{e}_{32} \end{pmatrix} = (1,0,0,0 \mid 0,1,0 \mid 0,1,0)^T \in \mathbb{R}^{10}$$

The number of matching symbols is given as

$$\begin{aligned} q = \mathbf{x}_1^T \mathbf{x}_2 &= (\mathbf{e}_{12})^T \mathbf{e}_{11} + (\mathbf{e}_{22})^T \mathbf{e}_{22} + (\mathbf{e}_{31})^T \mathbf{e}_{32} \\ &= \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= 0 + 1 + 0 = 1 \end{aligned}$$

The Euclidean and Hamming distances are given as

$$\begin{aligned} \delta(\mathbf{x}_1, \mathbf{x}_2) &= \sqrt{2(d-q)} = \sqrt{2 \cdot 2} = \sqrt{4} = 2 \\ \delta_H(\mathbf{x}_1, \mathbf{x}_2) &= d - q = 3 - 1 = 2 \end{aligned}$$

The cosine and Jaccard similarity are given as

$$\begin{aligned} \cos\theta &= \frac{d}{q} = \frac{1}{3} = 0.333 \\ J(\mathbf{x}_1, \mathbf{x}_2) &= \frac{q}{2d - q} = \frac{1}{5} = 0.2 \end{aligned}$$

3.5 Annotated References

3.6 Exercises and Projects

1. Show that after transforming a d-dimensional categorical dataset into a d'-dimensional binary space via the transformation in Section 3.3, each vector has length $\sqrt{d}$. Show that the maximum Euclidean distance between any two d'-dimensional in the new space is $\sqrt{2d}$. Finally show that the cosine similarity between any two vectors in the new space lies in the range $\cos\theta \in [0, 1]$, and consequently $\theta \in [0°, 90°]$.

2. Prove identity (**??**)

$$E[(\mathbf{V}_1 - \boldsymbol{\mu}_1)(\mathbf{V}_2 - \boldsymbol{\mu}_2)^T] = E[\mathbf{V}_1 \mathbf{V}_2^T] - E[\mathbf{V}_1] E[\mathbf{V}_2]^T$$

3. Let X and Y be two categorical variables. Let the domain of X be $\{x_1, x_2, x_3\}$ and the domain of Y be $\{y_1, y_2, y_3, y_4\}$ Given the contingency table below compute the χ^2 correlation between them. Are they dependent or independent?