

Chapter 2

Categorical Attributes

Last Modified: 2008-09-10 17:11

2.1 Single Attribute Analysis

We now treat the case of a random variable \mathbf{x} that takes on categorical values. We assume that the data consists of n points given by the vector $(x_1, x_2, \dots, x_n)^T$. Table 2.1 shows an example with $n = 8$ points.

\mathbf{x}
R
G
R
Y
B
Y
G
R

Table 2.1: Single attribute with $dom(\mathbf{x}) = \{R, G, Y, B\}$

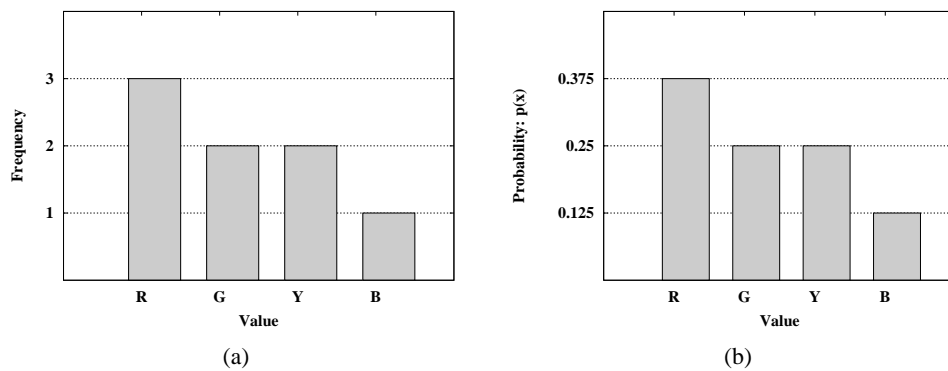


Figure 2.1: Histogram of \mathbf{x} : (a) Frequency Histogram, (b) Empirical Probability Distribution Function

Since categorical attributes typically allow one to test only of equality of values, there are not that many operations one can perform. For example the concept of mean and variance are not defined for categorical data. The mode still makes sense, since it gives the most frequent value. For the example above, R is the mode.

One approach to quickly summarize a categorical attribute is to display the frequency histogram of its values, as shown in Figure 2.1(a), which plots for each value of \mathbf{x} its frequency. These frequencies can be used to create an empirical probability distribution function for the random variable, as shown in Figure 2.1(b), by computing

$$P(\mathbf{x} = x) = \frac{n_x}{n} \quad (2.1)$$

where n_x is the frequency of value x . For example, $P(\mathbf{x} = R) = \frac{3}{8} = 0.375$. One point to keep in mind is that they “shape” of the distribution is meaningless for categorical values, since there is no inherent order to the values displayed. In fact, the values can be shown in any permutation without changing the meaning.

Pseudo-counts In creating the empirical probability distribution, it may happen that a particular value in $dom(\mathbf{x})$ may not appear among the n points. For example, if $dom(\mathbf{x}) = \{R, G, Y, B, O\}$, then for the example above, we will have $n_O = 0$ and consequently $P(\mathbf{x} = O) = 0$. Nevertheless, it is conceivable that O may occur in some larger sample, and it is preferable to assign it some small probability of occurrence. One way to achieve this is to add *pseudo-counts* to each value; the simplest approach is the so-called *Laplace adjustment*, where we add 1 to each value and then adjust the total count accordingly, given as:

$$P(\mathbf{x} = x) = \frac{n_x + 1}{n + |dom(\mathbf{x})|} \quad (2.2)$$

As an example, we now obtain $P(\mathbf{x} = O) = \frac{0+1}{8+5} = \frac{1}{13} = 0.077$.

2.2 Two Attribute Analysis: Contingency Tables

Here we assume that we have two categorical random variables \mathbf{x} and \mathbf{y} , with n 2-dimensional points $(x_i, y_i)^T$. As in the case of numeric attributes where we computed the covariance and correlation, we would like to test whether the two random variables are independent or dependent.

Test for the independence of the two categorical random variables can be done via *contingency table analysis*. The basic idea is to compute the squared deviation between the observed and expected counts for each possible pair of values $(x \in \mathbf{x}, y \in \mathbf{y})$.

Let n_{xy} denote the observed number of occurrences of the pair (x, y) in the sample. These counts are recorded in the so-called contingency table, at the cell indexed by x and y respectively. For example, assume that we have a sample of size $n = 1000$, let the random variable \mathbf{x} represent the attribute Car Size, with $dom(\mathbf{x}) = \{S, M, L\}$ (for Small, Medium and Large), and let \mathbf{y} represent the attribute Car Company, with $dom(\mathbf{y}) = \{A, B, C, D\}$. Assume that the observed frequencies for each pair of values n_{xy} are as shown in Table 2.2. For example, $n_{SA} = 157$.

Our goal now is to determine whether these two variables are dependent or independent. We rely on the classical *hypothesis testing* approach to answer this question. Let our *Null Hypothesis*, denoted H_0 be that \mathbf{x} and \mathbf{y} are independent. We need to either accept or reject the null hypothesis, based on some statistical test.

Assuming that the null hypothesis holds, i.e., assuming that \mathbf{x} and \mathbf{y} are independent, we can compute the expected frequency for each pair of values (x, y) , denoted e_{xy} . To do this, we first compute the marginal

Car Size (x)	Car Company (y)				Row Marginals
	A	B	C	D	
S	157	65	181	10	$P_S = 0.413$
M	126	82	142	46	$P_M = 0.396$
L	58	45	60	28	$P_L = 0.191$
Col. Marginals	$P_A = 0.341$	$P_B = 0.192$	$P_C = 0.383$	$P_D = 0.086$	n = 1000

Table 2.2: Contingency Table

probabilities for \mathbf{x} and \mathbf{y} . The (row) marginal probabilities for values of \mathbf{x} are obtained as follows:

$$P(\mathbf{x} = x) = \frac{\sum_y n_{xy}}{n} = \frac{n_x}{n} \quad (2.3)$$

Likewise the (column) marginal probabilities for values of \mathbf{y} are obtained as follows:

$$P(\mathbf{y} = y) = \frac{\sum_x n_{xy}}{n} = \frac{n_y}{n} \quad (2.4)$$

For example, we have $P(\mathbf{x} = S) = \frac{157+65+181+10}{1000} = \frac{413}{1000} = 0.413$, and $P(\mathbf{y} = A) = \frac{157+126+58}{1000} = \frac{341}{1000} = 0.341$.

Assuming \mathbf{x} and \mathbf{y} are independent, we have

$$P(\mathbf{x} = x, \mathbf{y} = y) = P(\mathbf{x} = x) \cdot P(\mathbf{y} = y) \quad (2.5)$$

Therefore the expected number of occurrences for any pair of values (x, y) is given as

$$e_{xy} = P(\mathbf{x} = x, \mathbf{y} = y) \cdot n = P(\mathbf{x} = x) \cdot P(\mathbf{y} = y) \cdot n \quad (2.6)$$

For example $e_{SA} = P(\mathbf{x} = S) \cdot P(\mathbf{y} = A) \cdot 1000 = 0.413 \times 0.341 \times 1000 = 140.8$. Likewise we can compute the expected values in each cell in the contingency table, as shown in Table 2.3.

	Car Size (x)	Car Company (y)			
		A	B	C	D
n_{xy}	S	157	65	181	10
e_{xy}		140.8	79.3	158.2	34.7
n_{xy}	M	126	82	142	46
e_{xy}		135	76	151.7	33.2
n_{xy}	L	58	45	60	28
e_{xy}		65.1	36.7	78.1	16.0

Table 2.3: Expected versus Observed Frequencies

Next we compute the χ^2 statistic, which computes the squared deviation of the observed values from the expected values, given as follows

$$\chi^2 = \sum_{x \in \mathbf{x}} \sum_{y \in \mathbf{y}} \frac{(n_{xy} - e_{xy})^2}{e_{xy}} \quad (2.7)$$

For our sample, we get $\chi^2 = 45.81$.

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, in this case, it is known that the sampling distribution is given by the *chi-squared* distribution, with q degrees of freedom, denoted as $\chi^2(q)$. The degrees of freedom are given as $s - t - 1$, where s is the number of terms in the summation, t is the number of independent parameters replaced by estimates, and we lose one degree of freedom for the total. Thus we have $s = |\text{dom}(\mathbf{x})| \cdot |\text{dom}(\mathbf{y})|$. As for the estimated parameters, not all of the $P(\mathbf{x} = x)$ and $P(\mathbf{y} = y)$ are independent, since the marginal probabilities along each dimension must add to 1. Thus the number of independent estimated parameters are given as $t = |\text{dom}(\mathbf{x})| - 1 + |\text{dom}(\mathbf{y})| - 1$. Finally, the degrees of freedom are computed as

$$q = s - t - 1 = |\text{dom}(\mathbf{x})| \cdot |\text{dom}(\mathbf{y})| - (|\text{dom}(\mathbf{x})| + |\text{dom}(\mathbf{y})| - 2) - 1 = (|\text{dom}(\mathbf{x})| - 1) \cdot (|\text{dom}(\mathbf{y})| - 1)$$

For our sample, we have $q = (3 - 1) \cdot (4 - 1) = 2 \cdot 3 = 6$.

The chi-squared distribution with q degrees of freedom is given as:

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{q/2-1} e^{-x/2} \quad (2.8)$$

where the Gamma Function is defined as

$$\Gamma(\alpha > 0) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad (2.9)$$

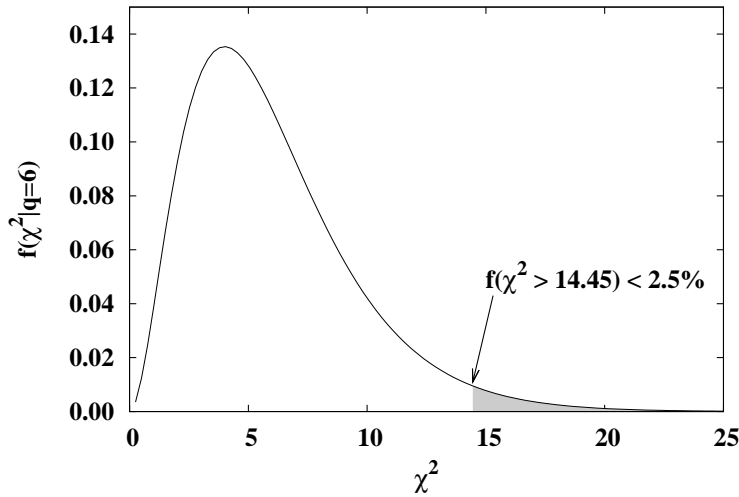


Figure 2.2: Chi-squared Distribution ($q = 6$)

The chi-squared p.d.f. with $q = 6$, namely $f(\chi^2|6)$, is shown in Figure 2.2. Let $\chi_{\alpha,q}^2$ denote the $1 - \alpha$ quantile value for the chi-squared distribution with q degrees of freedom, for which $f(x > \chi_{\alpha,q}^2) < \alpha$. In other words, the probability of obtaining a value greater than $\chi_{\alpha,q}^2$ is at most α . For example, $\chi_{0.025,6}^2 = 14.45$, which is shown in the figure. Furthermore, $\chi_{0.01,6}^2 = 16.81$. It is thus clear that the probability of obtaining such a high value of $\chi^2 = 45.81$ as in our sample data is extremely small. In general we can define the *p-value* of a statistic to be the probability of obtaining a value at least as extreme as the value of the statistic

under the null hypothesis. In other words, the p-value is the area under the curve to the right of the value of the statistic. For example, the p-value of $\chi^2 = 45.81$ is given as $f(x > 45.81) \approx 0$, which indicates that such a large deviation of the observed and expected values is extremely unlikely. Since the p-value is much less than some chosen threshold, say $\alpha = 0.01$, we conclude that our null hypothesis should be rejected, and consequently \mathbf{x} and \mathbf{y} are not independent.

2.3 Multiple Attribute Analysis: Multi-Dimensional Contingency Tables

Here we assume that we have d categorical attributes or random variables $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^d$, and each of the n points is a d -dimensional point $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^d)^T$.

As before, we would like to determine whether these d -dimensions are independent or not. We can test for independence by constructing a d -dimensional (or d -way) contingency table, which records the observed joint frequencies. The size of the d -dimensional contingency table is given as the product of the domains of each dimension, namely $\prod_{i=1}^d |\text{dom}(\mathbf{x}^i)|$, and each cell is indexed by a d -dimensional vector $\mathbf{x} = (x^1, x^2, \dots, x^d)^T$. Thus we denote the observed count in a cell as $n_{\mathbf{x}}$, and the expected count as $e_{\mathbf{x}}$.

Under the independence assumption, the expected counts are obtained by taking the product of the marginal probabilities along each dimension and multiplying it by the sample size, as follows

$$e_{\mathbf{x}} = n \cdot \prod_{i=1}^d P(\mathbf{x}^i = x^i) \quad (2.10)$$

The chi-squared statistic is computed as before:

$$\chi^2 = \sum_{\mathbf{x}} \frac{(e_{\mathbf{x}} - n_{\mathbf{x}})^2}{e_{\mathbf{x}}} \quad (2.11)$$

The degrees of freedom are computed as $s - t - 1$. Here s gives the number of cells, which is $s = \prod_{i=1}^d |\text{dom}(\mathbf{x}^i)|$. The number of estimated free parameters is given as $t = \sum_{i=1}^d |\text{dom}(\mathbf{x}^i)| - 1$. Thus the degrees of freedom are given as

$$q = s - t - 1 = \prod_{i=1}^d |\text{dom}(\mathbf{x}^i)| - \sum_{i=1}^d |\text{dom}(\mathbf{x}^i)| + (d - 1) \quad (2.12)$$

To reject the null hypothesis, we have to check whether the p -value of the statistic is smaller than the desired threshold α (say $\alpha = 0.01$). Note that the d -dimensional analysis indicates whether all d attributes taken together are independent or not. In general we may have to conduct k -way analysis to test if any k attributes of interest are independent or not.