مشخصه‌های دسته‌ای (کیفی)

Categorical Attributes

Bernoulli Variable

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

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

$$\sigma^2=p(1-p)$$

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

$$\begin{aligned}\hat{\sigma}^2&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\\&=\frac{n_1}{n}(1-\hat{p})^2+\frac{n-n_1}{n}(0-\hat{p})^2=\hat{p}(1-\hat{p})^2+(1-\hat{p})\hat{p}^2\\&=\hat{p}(1-\hat{p})(1-\hat{p}+\hat{p})=\hat{p}(1-\hat{p})\end{aligned}$$

Binomial Distribution: Number of Occurrences

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

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

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

Multivariate Bernoulli Variable

one-hot encoding

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

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

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = p_i \qquad \sum_{i=1}^{m} p_i = 1$$

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

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

Mean

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

Covariance Matrix

$$\boxed{\sigma_i^2 = \mathrm{var}(A_i) = p_i(1 - p_i)}$$

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

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

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

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

Sample Mean

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

Sample Covariance Matrix

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

where $\widehat{\mathbf{P}} = \mathrm{diag}(\hat{\mathbf{p}})$, and $\hat{\mathbf{p}} = \hat{\mu} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)^T$

Example 1

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)	$n_4 = 12$

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

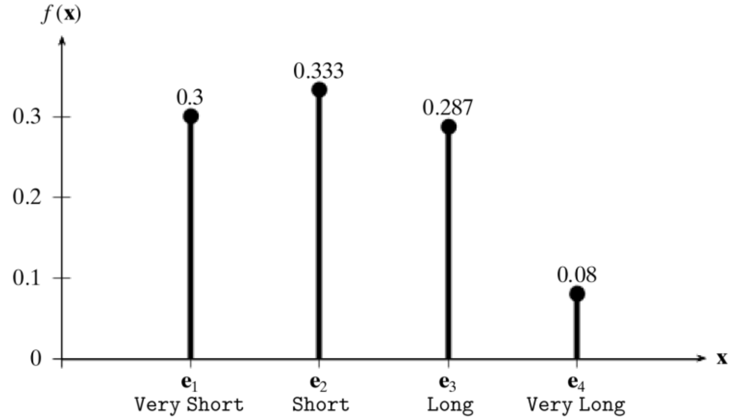


Figure 3.1. Probability mass function: `sepal length`.

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

$$= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.3 & 0.333 & 0.287 & 0.08 \end{pmatrix}$$

$$= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.09 & 0.1 & 0.086 & 0.024 \\ 0.1 & 0.111 & 0.096 & 0.027 \\ 0.086 & 0.096 & 0.082 & 0.023 \\ 0.024 & 0.027 & 0.023 & 0.006 \end{pmatrix}$$

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

Example 2

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

	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

	$\bar{A}_1$	$\bar{A}_2$
$\mathbf{z}_1$	−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

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

$$\sigma_1^2 = \frac{1}{5}\bar{A}_1^T\bar{A}_1 = 1.2/5 = 0.24$$

$$\sigma_2^2 = \frac{1}{5}\bar{A}_2^T\bar{A}_2 = 1.2/5 = 0.24$$

$$\sigma_{12} = \frac{1}{5}\bar{A}_1^T\bar{A}_2 = -1.2/5 = -0.24$$

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

Multinomial Distribution: Number of Occurrences

$$\mathbf{N} = (N_1, N_2, \ldots, N_m)^T$$

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

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

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

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

$$\hat{\boldsymbol{\mu}}_{\mathbf{N}} = n\hat{\mathbf{p}}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = n\left(\widehat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T\right)$$

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

$$dom(X_1) = \{a_{11}, a_{12}, \ldots, a_{1m_1}\}$$

$$dom(X_2) = \{a_{21}, a_{22}, \ldots, a_{2m_2}\}$$

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

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

$$\textstyle\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$$

Mean and Sample Mean

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

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

Covariance and Sample Covariance

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

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

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

$$\begin{aligned} \boldsymbol{\Sigma}_{12} &= E[(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T] \\ &= E[\mathbf{X}_1 \mathbf{X}_2^T] - E[\mathbf{X}_1] E[\mathbf{X}_2]^T \\ &= \mathbf{P}_{12} - \boldsymbol{\mu}_1 \boldsymbol{\mu}_2^T \\ &= \mathbf{P}_{12} - \mathbf{p}_1 \mathbf{p}_2^T \end{aligned}$$

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

$$= \begin{pmatrix} p_{11} - p_1^1 p_1^2 & p_{12} - p_1^1 p_2^2 & \cdots & p_{1m_2} - p_1^1 p_{m_2}^2 \\ p_{21} - p_2^1 p_1^2 & p_{22} - p_2^1 p_2^2 & \cdots & p_{2m_2} - p_2^1 p_{m_2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} - p_{m_1}^1 p_1^2 & p_{m_1 2} - p_{m_1}^1 p_2^2 & \cdots & p_{m_1 m_2} - p_{m_1}^1 p_{m_2}^2 \end{pmatrix}$$

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

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

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

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

Example 3

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)	$n_4 = 12$

Table 3.3. Discretized `sepal width` attribute

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

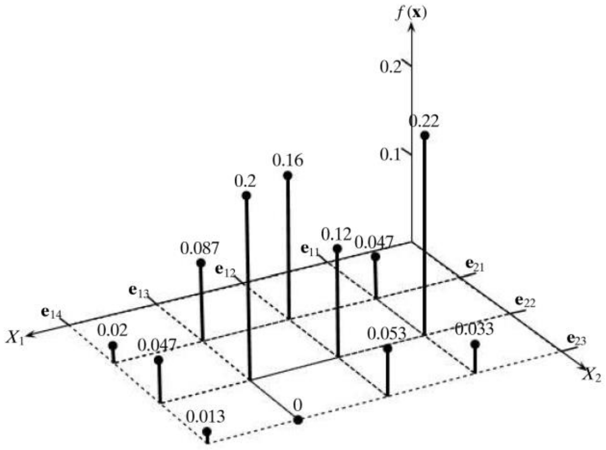


Figure 3.2. Empirical joint probability mass function: `sepal length` and `sepal width`.

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

		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 ($\mathbf{e}_{22}$)	24	18	8
	Long ($\mathbf{e}_{13}$)	13	30	0
	Very Long ($\mathbf{e}_{14}$)	3	7	2

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{12}^T & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix} \qquad \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1)^T$$

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

Attribute Dependence: Contingency Analysis

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2 \qquad e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n}$$

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

		Sepal width (X_2)			
		Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
Sepal length (X_1)	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	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

χ^2 Statistic and Hypothesis Testing

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

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

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

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

a.fahim@ut.ac.ir

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

The **significance level α** corresponds to the least level of surprise we need to reject the null hypothesis.

we reject the null hypothesis if $\boldsymbol{p_v(\chi^2) \le \alpha}$

Note that the value 1− α is also called the **confidence level**. So equivalently, we say that we reject the null hypothesis at the 100(1 −α)% confidence level if $p-value(\chi^2) \le \alpha$.

$$\boxed{\text{p-value}(\chi^2) = P(x \ge \chi^2) = 1 - F_q(\chi^2)}$$

Critical value, v_α

$$P(x \ge v_\alpha) = 1 - F_q(v_\alpha) = \alpha, \text{ or equivalently } F_q(v_\alpha) = 1 - \alpha$$

$$v_\alpha = F_q^{-1}(1-\alpha)$$

$$\chi^2 \ge v_\alpha: \; P(x \ge \chi^2) \le P(x \ge v_\alpha)$$

$$\text{p-value}(\chi^2) \le \text{p-value}(v_\alpha) = \alpha$$

Example 4

Table 3.6. Expected counts

		X_2		
		Short (a_{21})	Medium (a_{22})	Long (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

$\alpha = 0.01$

$\chi^2 = 21.8$

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

$\text{p-value}(21.8) = 1 - F_6(21.8) = 1 - 0.9987 = 0.0013$

$v_\alpha = F_6^{-1}(1-\alpha) = F_6^{-1}(0.99) = 16.81$

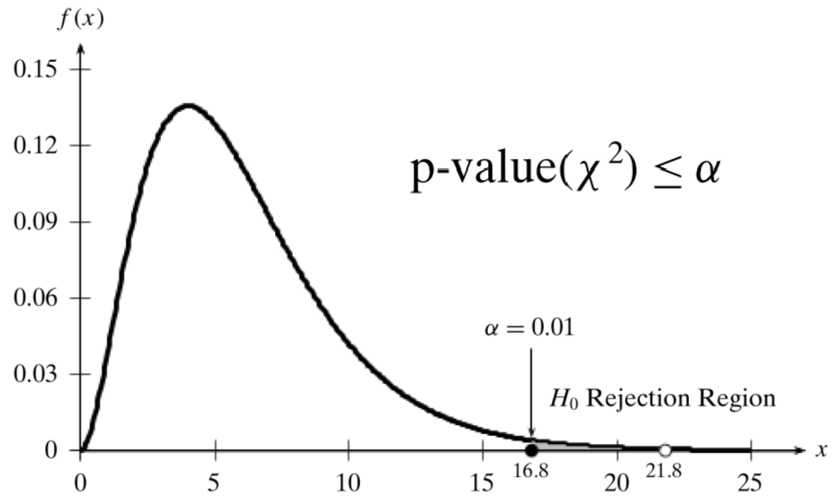


Figure 3.3. Chi-squared distribution ($q = 6$).

As $21.8 > v_\alpha = 16.81$. In effect, we reject the null hypothesis that sepal length and sepal width are independent, and accept the alternative hypothesis that they are dependent.

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

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

Example 5

Example 3.11 (Multivariate Analysis). Let us consider the 3-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.

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

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

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

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

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

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

Multiway Contingency Analysis

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

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

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

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

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

$$q = \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1)$$

$$= \left(\prod_{i=1}^{d} m_i\right) - \left(\sum_{i=1}^{d} m_i\right) + d - 1$$

$\chi^2 = 231.06$

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

$\alpha = 0.01$

$\text{p-value}(231.06) = 7.91 \times 10^{-34}$

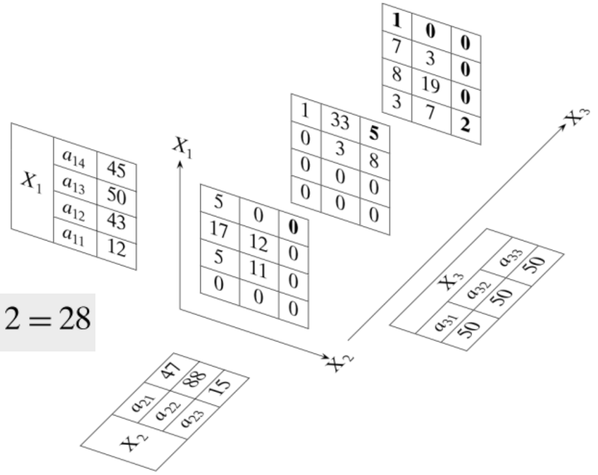


Figure 3.4. 3-Way contingency table (with marginal counts along each dimension).

Multivariate Bernoulli variables

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

The number of matching symbols:
$$s = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^{d} (\mathbf{e}_{ki_k})^T \mathbf{e}_{kj_k}$$

The number of mismatches d −s
$$\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i = d$$

Euclidean Distance

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

Hamming Distance

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

Cosine Similarity

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

Jaccard Coefficient

It is defined as the ratio of the number of matching values to the number of distinct values that appear in both x_i and x_j , across the d attributes:

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

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

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

The Euclidean and Hamming distances are given as

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \sqrt{2(d-s)} = \sqrt{2 \cdot 2} = \sqrt{4} = 2$$
$$\delta_H(\mathbf{x}_1, \mathbf{x}_2) = d - s = 3 - 1 = 2$$

The cosine and Jaccard similarity are given as

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

$$[x_{\min}, v_1],\ (v_1, v_2],\ \ldots,\ (v_{k-1}, x_{\max}]$$

Equal-Width Intervals

Partition the range of X into k equal-width intervals.

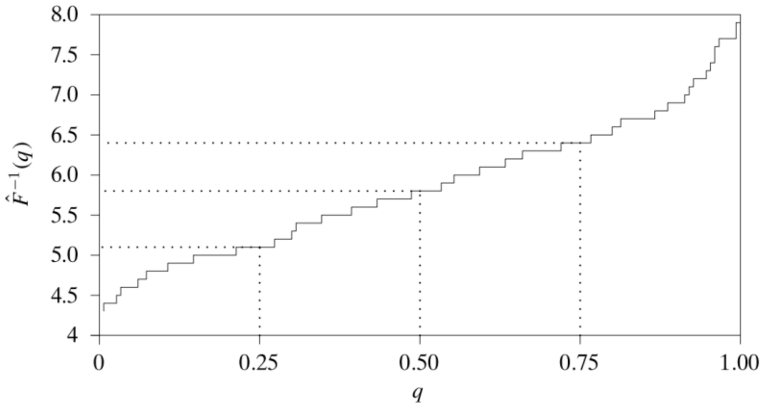
$$w = \frac{x_{\max} - x_{\min}}{k}$$

$$v_i = x_{\min} + i w,\ \text{for } i = 1, \ldots, k-1$$

Equal-Frequency Intervals

$$v_i = \hat{F}^{-1}(i/k) \text{ for } i = 1, \ldots, k-1$$

$$\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}, \text{for } q \in [0, 1].$$



Example 3.14. Consider the `sepal length` attribute in the Iris dataset. Its minimum and maximum values are

$$x_{\min} = 4.3 \qquad x_{\max} = 7.9$$

We discretize it into $k = 4$ bins using equal-width binning. The width of an interval is given as

$$w = \frac{7.9 - 4.3}{4} = \frac{3.6}{4} = 0.9$$

and therefore the interval boundaries are

$$v_1 = 4.3 + 0.9 = 5.2 \qquad v_2 = 4.3 + 2 \cdot 0.9 = 6.1 \qquad v_3 = 4.3 + 3 \cdot 0.9 = 7.0$$

The four resulting bins for `sepal length` are shown in Table 3.1, which also shows the number of points n_i in each bin, which are not balanced among the bins.

For equal-frequency discretization, consider the empirical inverse cumulative distribution function (CDF) for `sepal length` shown in Figure 3.5. With $k = 4$ bins, the bin boundaries are the quartile values (which are shown as dashed lines):

$$v_1 = \hat{F}^{-1}(0.25) = 5.1 \qquad v_2 = \hat{F}^{-1}(0.50) = 5.8 \qquad v_3 = \hat{F}^{-1}(0.75) = 6.4$$

The resulting intervals are shown in Table 3.8. We can see that although the interval widths vary, they contain a more balanced number of points. We do not get identical counts for all the bins because many values are repeated; for instance, there are nine points with value 5.1 and there are seven points with value 5.8.

Table 3.8. Equal-frequency discretization: `sepal length`

Bin	Width	Count
$[4.3, 5.1]$	0.8	$n_1 = 41$
$(5.1, 5.8]$	0.7	$n_2 = 39$
$(5.8, 6.4]$	0.6	$n_3 = 35$
$(6.4, 7.9]$	1.5	$n_4 = 35$