

Chapter 1

Numeric Attributes

Last Modified: 2008-09-07 15:51

1.1 Dataset & Samples

A given dataset generally represents a subset or a *sample* of the entire set of possible observations; the universe of all possible observations is also called a *population*. Usually we are interested in the *parameters* or certain characteristics of the entire population (e.g., the mean age). However looking at the entire population is typically not be feasible or may be too costly. The goal of exploratory analysis is to make inferences about the population parameters using information from one or more samples and by computing the appropriate sample *statistics*. In other words, we estimate the population parameters via sample statistics. However, in our discussion below, for simplicity, we will assume that the data represents the entire population, as opposed to being a sample. This way we do not have to distinguish between the population parameters and sample statistics. Note that this is not an unrealistic assumption in data mining, since typically we do not know much about the distributions and parameters of the population to begin with, and the dataset we collect is typically taken to represent the whole population.

1.2 Types of Attributes

Many datasets can be represented in the form of tables. A table is a $n \times d$ matrix, where n is the number of rows and d is the number of columns. The rows denote a collection of instances, which can also be called as examples, records, objects, transactions, feature vectors, etc. The columns denote a collection of attributes, which can also be called dimensions, variables, features, properties, fields, etc. For example, consider a sample of the demographic data shown in Table 1.1, which records information like Age, Sex, Marital Status, Education and Income for individuals in a population. Note that some data may be missing (denoted with “NULL”); for instance, we may not know the age for individual with ID 249.

Attributes can be of different types depending on their domain, i.e., depending on the types of values they take on. A *categorical* attribute is one that has a set-valued domain composed of a set of finite symbols. For example, Sex, Marital Status, and Education are categorical attributes, since $domain(\text{Sex}) = \{M, F\}$, $domain(\text{MaritalStatus}) = \{Single, Married\}$, and $domain(\text{Education}) = \{HighSchool, BS, MS, PhD, -Other\}$. Categorical attributes may be of two types:

ID	Age	Sex	Marital Status	Education	Income
248	54	M	Married	High School	\$100,000
249	NULL	F	Married	High School	\$12,000
250	29	M	Single	B.S.	\$23,000
251	7	M	NULL	Other	\$0
...

Table 1.1: Sample of Demographic Data

- *Nominal*: An attribute is called nominal if its values cannot be ordered. Only test for equality is allowed (for example Sex).
- *Ordinal*: An attribute is called ordinal if its values can be ordered in some way (for example, Education, since we can claim that someone who has a BS is more educated than someone with High School diploma. Similarly, a PhD is obtained after an MS and an MS after a BS. Thus, there is an order to the domain values, even though we may not be able to quantify the difference between successive values).

A *numeric* attribute is one that has a real-valued or interger-valued domain. For example, Age and Income in Table 1.1 are numeric attributes, since $domain(Age) = domain(Income) = \mathbb{R}^+$ (the set of all positive real values). Numeric attributes may be of two types:

- *Interval*: For these kinds of attributes only differences (addition or subtraction) make sense. For example, temperature measured in C° or F° is interval-scaled, since it is not very meaningful to say that $20C^\circ$ is “twice” as cold as $10C^\circ$. On the other hand, we can take their difference.
- *Ratio*: Most numeric attributes are such that we can take their differences as well as ratios. For example, we can say that someone who is 20 years old is twice as old as someone who is 10 years old.

1.3 Single Attribute Analysis

Here we assume that our data matrix has n rows, representing several instances, but only one numeric attribute, which we treat as a random variable \mathbf{x} . The set of values of \mathbf{x} is given by the vector $(x_1, x_2, \dots, x_n)^T$. Here the superscript T denotes matrix transpose, since we assume that all vectors are column vectors by default.

1.3.1 Measures of Central Tendency

Mean The *mean* (μ) of the random variable \mathbf{x} is the average value, given as

$$\mu = \frac{\sum_{i=1}^n x_i}{n} \quad (1.1)$$

The mean is also defined as the *expected value* of the random variable \mathbf{x} , given as

$$\mathbf{E}[\mathbf{x}] = \sum_{i=1}^n x_i p(x_i) = \sum_{i=1}^n x_i \frac{1}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Note that $p(x_i) = \frac{1}{n}$ since all n values are assumed to be equally likely. The mean is also the *1st moment* of \mathbf{x} , a special case of the *r-th moment* of the random variable \mathbf{x} , which is defined as:

$$\mathbf{E}[\mathbf{x}^r] = \sum_{i=1}^n x_i^r p(x_i) = \frac{\sum_{i=1}^n (x_i)^r}{n}$$

Whereas mean is a useful, intuitive statistic, it has one major problem. It is not *robust*, i.e, it is very sensitive to extreme values. For instance, a single very large value can skew the mean.

Median The *median* of \mathbf{x} is the middle-most value, i.e., half the points have values less than the median, and the other half have values more than the median. If \mathbf{x} is sorted in increasing order, the median is the value at position $\frac{n+1}{2}$ if n is odd, and the average value of the elements in positions $\frac{n}{2}$ and $\lceil \frac{n+1}{2} \rceil$ if n is even. Unlike the mean, median is robust, since it is not affected very much by extreme values.

Mode The *mode* of \mathbf{x} is the most frequently occurring value. The mode may not be a very useful measure of central tendency for raw data values, since by chance an unrepresentative element may be the most frequent element. It is more useful when we consider ranges of values and find out the peaks.

1.3.2 Measures of Dispersion

Range The *range* of the random variable \mathbf{x} is the difference between the maximum and minimum values, given as $range(\mathbf{x}) = \max\{x_i\} - \min\{x_i\}$. By definition it is sensitive to extreme values, and thus not robust.

Variance and Standard Deviation The *variance* (σ^2) of \mathbf{x} is the average squared deviation of the data values x_i from their mean μ ¹, defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} \quad (1.2)$$

The variance is also defined as the expected value of the *second moment about the mean*, i.e.,

$$\mathbf{E}[(\mathbf{x} - \mu)^2] = \sum_{i=1}^n (x_i - \mu)^2 p(x_i) = \sum_{i=1}^n (x_i - \mu)^2 \frac{1}{n} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

In general, the *r-th moment about the mean* is defined as

$$\mathbf{E}[(\mathbf{x} - \mu)^r] = \sum_{i=1}^n (x_i - \mu)^r p(x_i) = \frac{\sum_{i=1}^n (x_i - \mu)^r}{n}$$

The *standard deviation* (σ) is just the square-root of the variance, i.e.,

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \quad (1.3)$$

¹From an estimation view-point one has to use the denominator of $n - 1$ for sample variance, to obtain an unbiased estimator for the population variance. Here since we assume that \mathbf{x} represents the entire population, we can simply divide by n . In any case, for large n , the difference is negligible.

The *standard score* (z) of a value x_i for the attribute \mathbf{x} is the number of standard deviations away the value is from the mean, given as:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (1.4)$$

Whereas standard deviation is an absolute measure of dispersion, the *coefficient of variation*, given as $\frac{\sigma}{\mu}$, gives a relative measure of dispersion. It tells us the magnitude of the deviation relative to the magnitude of the mean. Likewise, the standard score is also a relative measure of dispersion with respect to the standard deviation.

Inter-Quartile Range Like the mean, variance and standard deviation are both non-robust. A more robust measure of dispersion is the *inter-quartile range*. *Quartiles* divide the data into 4 equal parts. If we consider the values of \mathbf{x} in sorted order, the first quartile (Q_1) is the value to the left of which 25% of the points lie, the second quartile (Q_2) is the value to the left of which 50% of the points lie (i.e., the same as the median), the third quartile (Q_3) is the value to the left of which 75% of the points lie, and the fourth quartile (Q_4) is the value to the left of which 100% of the points lie (i.e, the maximum value in \mathbf{x}). The *inter-quartile range* is given as $IQR = Q_3 - Q_1$, which indicates how much we have to go on either side of the median to include exactly half of the data points. *IQR* is robust by definition.

In general, median and quartiles are examples of *fractiles* or *quantiles*. Given the frequency distribution of an attribute X , the q -th quantile is the value such that the fraction q of the points lie to the left. Thus the median is the 0.5-quantile, whereas the first quartile is the 0.25-quantile.

1.4 Two Attribute Analysis

Here we assume that our table has n rows, and two numeric attributes, which can be represented by two random variable \mathbf{x} and \mathbf{y} , with corresponding value vectors $(x_1, x_2, \dots, x_n)^T$ and $(y_1, y_2, \dots, y_n)^T$, respectively. Another way to look at the two variables is to consider the data as a $n \times 2$ matrix. Row i then represents the pair of values (x_i, y_i) for \mathbf{x} and \mathbf{y} corresponding to point i . As before, we assume that the data represents the entire population, since otherwise we would have to separately define the population and sample means.

Mean The 2-dimensional mean is given as the vector

$$\boldsymbol{\mu} = (\mu_{\mathbf{x}}, \mu_{\mathbf{y}})^T \quad (1.5)$$

obtained by taking the means along each attribute. The two-attribute median and mode can also be defined by taking the median and mode along each attribute.

Covariance The two attribute variance is called *covariance*²; it measures how much the two attributes deviate from their respective means. It is defined as follows:

$$\sigma_{\mathbf{xy}} = \frac{\sum_{i=1}^n (x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})}{n} \quad (1.6)$$

The covariance measures how \mathbf{x} and \mathbf{y} co-vary, i.e, how likely are values of \mathbf{x} to be above (or below) their mean when values of \mathbf{y} are above (or below) their mean. As such it measures only linear relationship between

²As in the case of variance, the sample covariance has denominator $n - 1$, to obtain an unbiased estimator for the population covariance.

\mathbf{x} and \mathbf{y} . Covariance is also a special case of the r -th and s -th *product moment about the mean*, which is given as

$$\mathbf{E}[(\mathbf{x} - \mu_{\mathbf{x}})^r (\mathbf{y} - \mu_{\mathbf{y}})^s] = \sum_{i=1}^n (x_i - \mu_{\mathbf{x}})^r (y_i - \mu_{\mathbf{y}})^s p(x_i, y_i) = \frac{\sum_{i=1}^n (x_i - \mu_{\mathbf{x}})^r (y_i - \mu_{\mathbf{y}})^s}{n}$$

Note that $p(x_i, y_i) = \frac{1}{n}$ since each of the n points (x_i, y_i) is equally likely (assuming, as we do, that the data represents the entire population).

Correlation The *correlation coefficient* between variables \mathbf{x} and \mathbf{y} is given as:

$$\rho_{\mathbf{xy}} = \frac{\sigma_{\mathbf{xy}}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} = \frac{\sum (x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})}{\sqrt{\sum (x_i - \mu_{\mathbf{x}})^2 \sum (y_i - \mu_{\mathbf{y}})^2}} \quad (1.7)$$

Correlation is *standardized* covariance, i.e, it is obtained by normalizing the covariance by the standard deviation of each variable.

The *Cauchy-Schwartz* inequality states that for any vectors \mathbf{a} and \mathbf{b}

$$|\mathbf{a}^T \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\| \quad (1.8)$$

Let \mathbf{a} and \mathbf{b} be the n -dimensional vectors given as

$$\mathbf{a} = \begin{pmatrix} x_1 - \mu_{\mathbf{x}} \\ x_2 - \mu_{\mathbf{x}} \\ \vdots \\ x_n - \mu_{\mathbf{x}} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} y_1 - \mu_{\mathbf{y}} \\ y_2 - \mu_{\mathbf{y}} \\ \vdots \\ y_n - \mu_{\mathbf{y}} \end{pmatrix} \quad (1.9)$$

The correlation coefficient can then be re-written as follows:

$$\rho_{\mathbf{xy}} = \frac{\mathbf{a}^T \mathbf{b}}{\sqrt{\mathbf{a}^T \mathbf{a}} \sqrt{\mathbf{b}^T \mathbf{b}}} = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (1.10)$$

It follows immediately from the Cauchy-Schwartz inequality that $|\rho_{\mathbf{xy}}| \leq 1$. Also worth mentioning is that the correlation coefficient is identical to the angle between the two vectors, i.e.,

$$\rho_{\mathbf{xy}} = \cos \theta = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \implies \theta = \cos^{-1}(\rho_{\mathbf{xy}}) \quad (1.11)$$

\mathbf{x}	\mathbf{y}
1	0.8
5	2.4
9	5.5

(a)

$\mu_{\mathbf{x}}$	$\mu_{\mathbf{y}}$
$\frac{15}{3} = 5$	$\frac{8.7}{3} = 2.9$

(b)

$\mathbf{a} = \mathbf{x} - \mu_{\mathbf{x}}$	$\mathbf{b} = \mathbf{y} - \mu_{\mathbf{y}}$
-4	-2.1
0	-0.5
4	2.6

(c)

Figure 1.1: Correlation Example

As an example, let our sample consist of $n = 3$ points in $d = 2$ dimensions, as shown in Figure 1.1(a). The means along each dimension are shown in Figure 1.1(b), and the mean subtracted vectors are shown in Figure 1.1(c). We can compute the correlation coefficient as follows:

$$\rho_{xy} = \frac{(-4, 0, 4) \cdot (-2.1, -0.5, 2.6)^T}{\sqrt{(-4)^2 + 0^2 + 4^2} \sqrt{(-2.1)^2 + (-0.5)^2 + 2.6^2}} = \frac{18.8}{\sqrt{32} \sqrt{11.42}} = 0.98 \quad (1.12)$$

The angle between the two mean subtracted 3-dimensional vectors $\mathbf{a} = (-4, 0, 4)^T$ and $\mathbf{b} = (-2.1, -0.5, 2.6)^T$ is $\cos^{-1}(0.98) = 10.4^\circ$. Figure 1.2, shows the three points in the $d = 2$ dimensional space comprised of two attributes, and it also shows the two $n = 3$ dimensional mean-subtracted vectors.

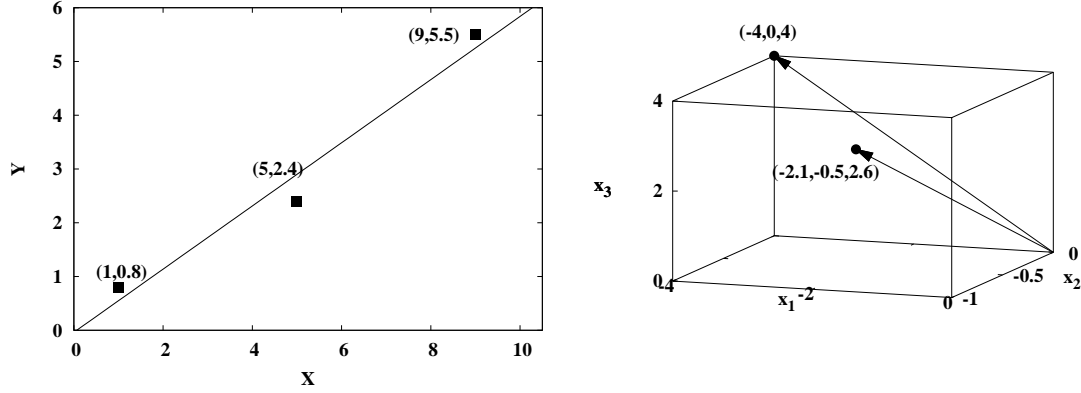


Figure 1.2: Geometric Interpretation of Correlation: The best fitting line, through the three points in the original $d = 2$ dimensional space, is shown to indicate a good linear fit. The right figure shows the two $n = 3$ dimensional vectors.

Covariance Matrix The 2-dimensional covariance information can be summarized via a 2×2 covariance matrix given as:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix} \quad (1.13)$$

Since $\sigma_{xy} = \sigma_{yx}$, $\mathbf{\Sigma}$ is a symmetric matrix. Also, using (1.7), we can re-write $\sigma_{xy} = \rho_{xy} \sigma_x \sigma_y$ to obtain:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\ \rho_{xy} \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} \quad (1.14)$$

Thus $\det(\mathbf{\Sigma}) = \sigma_x^2 \sigma_y^2 - \rho_{xy}^2 \sigma_x^2 \sigma_y^2 = (1 - \rho_{xy}^2) \sigma_x^2 \sigma_y^2$. Since $|\rho_{xy}| \leq 1$, we have $\rho_{xy}^2 \leq 1$. This implies that $\det(\mathbf{\Sigma}) \geq 0$, i.e., the determinant is never negative.

1.5 Multiple Attribute Analysis

We now generalize mean and variance to multiple attributes. We now assume that we have n points over d attributes. We assume that each point (row) is given as a d -dimensional vector $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^d)^T$. Each attribute or random variable (column) can also be thought of as the vector $\mathbf{x}^j = (x_1^j, x_2^j, \dots, x_n^j)^T$.

Mean The multi-dimensional mean vector is obtained by taking the mean of each attribute, given as

$$\boldsymbol{\mu} = (\mu_{\mathbf{x}^1}, \mu_{\mathbf{x}^2}, \dots, \mu_{\mathbf{x}^d})^T \quad (1.15)$$

Covariance The multi-dimensional covariance is captured by the $d \times d$ symmetric *covariance matrix* that gives the covariance for each pair of attributes, defined as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{\mathbf{x}^1}^2 & \sigma_{\mathbf{x}^1\mathbf{x}^2} & \cdots & \sigma_{\mathbf{x}^1\mathbf{x}^d} \\ \sigma_{\mathbf{x}^2\mathbf{x}^1} & \sigma_{\mathbf{x}^2}^2 & \cdots & \sigma_{\mathbf{x}^2\mathbf{x}^d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{\mathbf{x}^d\mathbf{x}^1} & \sigma_{\mathbf{x}^d\mathbf{x}^2} & \cdots & \sigma_{\mathbf{x}^d}^2 \end{pmatrix} \quad (1.16)$$

The diagonal elements $\sigma_{\mathbf{x}^j}^2$ denote the single attribute variances, whereas the off-diagonal elements $\sigma_{\mathbf{x}^i\mathbf{x}^j}$ denote the covariance between attributes \mathbf{x}^i and \mathbf{x}^j .

Let $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^d)^T$ denote the d -dimensional vector of random variables corresponding to the d attributes. Then, the covariance matrix can also be written as:

$$\boldsymbol{\Sigma} = \mathbf{E}[(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})] \quad (1.17)$$

$$= \begin{pmatrix} \mathbf{E}[(\mathbf{x}^1 - \mu_{\mathbf{x}^1})^2] & \mathbf{E}[(\mathbf{x}^1 - \mu_{\mathbf{x}^1})(\mathbf{x}^2 - \mu_{\mathbf{x}^2})] & \cdots & \mathbf{E}[(\mathbf{x}^1 - \mu_{\mathbf{x}^1})(\mathbf{x}^d - \mu_{\mathbf{x}^d})] \\ \mathbf{E}[(\mathbf{x}^1 - \mu_{\mathbf{x}^1})(\mathbf{x}^2 - \mu_{\mathbf{x}^2})] & \mathbf{E}[(\mathbf{x}^2 - \mu_{\mathbf{x}^2})^2] & \cdots & \mathbf{E}[(\mathbf{x}^2 - \mu_{\mathbf{x}^2})(\mathbf{x}^d - \mu_{\mathbf{x}^d})] \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{E}[(\mathbf{x}^1 - \mu_{\mathbf{x}^1})(\mathbf{x}^d - \mu_{\mathbf{x}^d})] & \mathbf{E}[(\mathbf{x}^2 - \mu_{\mathbf{x}^2})(\mathbf{x}^d - \mu_{\mathbf{x}^d})] & \cdots & \mathbf{E}[(\mathbf{x}^d - \mu_{\mathbf{x}^d})^2] \end{pmatrix} \quad (1.18)$$

Whereas the covariance matrix gives all pair-wise attribute covariances, the *generalized variance*, gives a single value for the overall multivariate scatter, defined as $\det(\boldsymbol{\Sigma})$, which represents the *determinant* of the covariance matrix. The determinant $\det(\boldsymbol{\Sigma})$ is also denoted as $|\boldsymbol{\Sigma}|$. It is worth noting that $\boldsymbol{\Sigma}$ is a *positive-semidefinite* matrix, i.e., $\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \geq 0$ for any d -dimensional vector \mathbf{a} . This also means that all upper left $k \times k$ sub-matrices of $\boldsymbol{\Sigma}$ have non-negative determinants ($1 \leq k \leq d$). Another consequence of interest later is that all the eigenvalues of $\boldsymbol{\Sigma}$ are also non-negative.

1.6 Normal Distribution

The normal distribution is one of the most important statistical distributions, especially since many physically observed variables follow an approximately normal distribution, and due to the central limit theorem, the sampling distribution of the mean follows a normal distribution. The normal distribution also plays an important role in data mining, especially in clustering and density estimation.

1.6.1 Univariate Normal Distribution

A random variable, or in our case, an attribute, \mathbf{x} has a normal distribution with mean μ and variance σ^2 (with $\mu \in \mathbb{R}$ and with $\sigma \in \mathbb{R}^+$) if \mathbf{x} has a continuous distribution whose *probability density function* (p.d.f.) is given as follows:

$$f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (1.19)$$

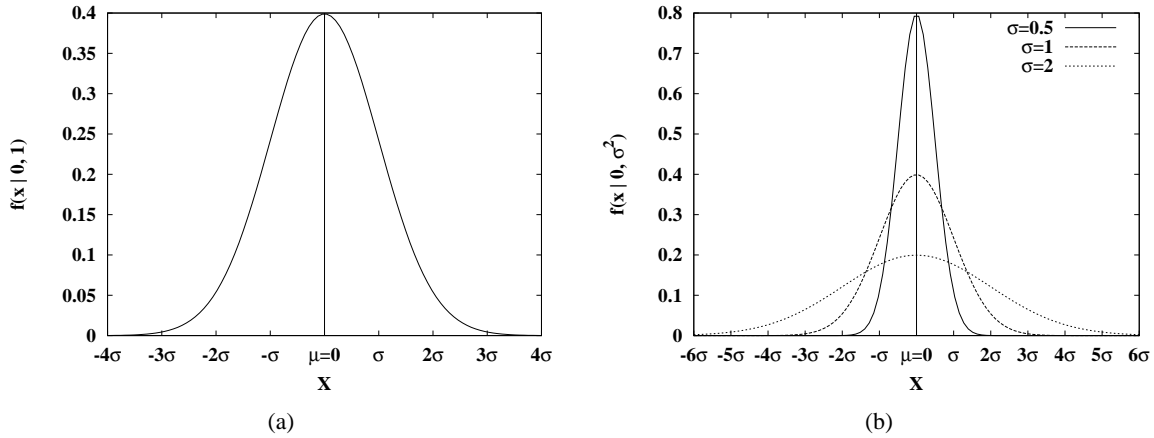


Figure 1.3: (a) Standard Normal Distribution, (b) Normal Distribution with Different Variances ($\mu = 0$).

Figure 1.3(a) plots the p.d.f. of the standard normal distribution, with $\mu = 0$ and $\sigma^2 = 1$. The normal distribution has a characteristic *bell* shape. Figure 1.3(b) shows the effect of variance on the shape of the distribution. Since the normal distribution is symmetric, the mean μ is also the median, as well as the mode, of the distribution. Note that term $(x_i - \mu)^2$ in (1.19) measures the distance of a value x_i from the mean of the distribution μ .

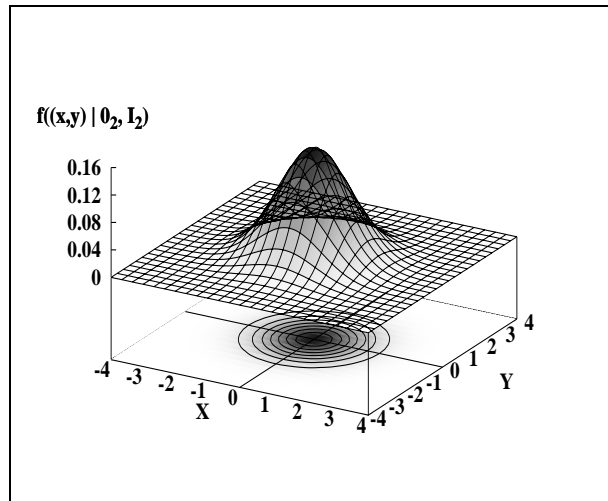


Figure 1.4: Standard Bivariate Normal Distribution

1.6.2 Multivariate Normal Distribution

Given d random variables or attributes $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^d$, and the d -dimensional point $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^d)^T$, we say that these random variables have a multivariate normal distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, whose p.d.f. is given as follows:

$$f(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2}} \quad (1.20)$$

Let $\mathbf{0}_d = (0_1, 0_2, \dots, 0_d)^T$ denote the zero vector with d components (or dimensions), and let \mathbf{I}_d denote the $d \times d$ identity matrix (whose diagonals are all 1's and the off-diagonal elements are all 0's). Figure 1.4 plots the p.d.f. of the standard bivariate (i.e., $d = 2$) normal distribution, with

$$\boldsymbol{\mu} = \mathbf{0}_2 = (0 \ 0)^T$$

and

$$\boldsymbol{\Sigma} = \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This corresponds to the case where the two attributes are independent, and follow the standard normal distribution. In general, $\boldsymbol{\Sigma}$ will be a positive semi-definite matrix. Also, just as in the univariate case, the term

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

in (1.20) measures the distance, called the *Mahalanobis distance*, of \mathbf{x}_i from the mean of the distribution $\boldsymbol{\mu}$.