# ماتریس داده‌ها

Data Matrix

# 1.1-Data Matrix (ماتریس داده‌ها)

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Depending on the application domain, rows may also be referred to as entities, **instances**, examples, records, transactions, objects, points, feature-vectors, tuples, and so on. Likewise, columns may also be called **attributes**, properties, features, dimensions, variables, fields, and so on.

The number of instances n is referred to as the **size of the data**, whereas the number of attributes d is called the **dimensionality** of the data.

Table 1.1. Extract from the Iris dataset

| | **Sepal length** $X_1$ | **Sepal width** $X_2$ | **Petal length** $X_3$ | **Petal width** $X_4$ | **Class** $X_5$ |
|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 5.9 | 3.0 | 4.2 | 1.5 | Iris-versicolor |
| $\mathbf{x}_2$ | 6.9 | 3.1 | 4.9 | 1.5 | Iris-versicolor |
| $\mathbf{x}_3$ | 6.6 | 2.9 | 4.6 | 1.3 | Iris-versicolor |
| $\mathbf{x}_4$ | 4.6 | 3.2 | 1.4 | 0.2 | Iris-setosa |
| $\mathbf{x}_5$ | 6.0 | 2.2 | 4.0 | 1.0 | Iris-versicolor |
| $\mathbf{x}_6$ | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| $\mathbf{x}_7$ | 6.5 | 3.0 | 5.8 | 2.2 | Iris-virginica |
| $\mathbf{x}_8$ | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\mathbf{x}_{149}$ | 7.7 | 3.8 | 6.7 | 2.2 | Iris-virginica |
| $\mathbf{x}_{150}$ | 5.1 | 3.4 | 1.5 | 0.2 | Iris-setosa |

The analysis of a single attribute is referred to as **univariate analysis**, whereas the simultaneous analysis of two attributes is called **bivariate analysis** and the simultaneous analysis of more than two attributes is called **multivariate analysis**.

# 1.2-Attributes (مشخصه‌ها)

## مشخصه‌های عددی (Numeric Attributes)

• Interval-scaled: For these kinds of attributes only differences (addition or subtraction) make sense. For example, attribute .
• Ratio-scaled: Here one can compute both differences as well as ratios between values. For example, for attribute Age.

$$domain(\texttt{Age}) = \mathbb{N} \qquad domain(\texttt{petal length}) = \mathbb{R}^+$$

## مشخصه‌های دسته‌ای (Categorical Attributes)

Nominal: The attribute values in the domain are unordered

$$domain(\texttt{Sex}) = \{\texttt{M}, \texttt{F}\}$$

Ordinal: The attribute values are ordered

$$domain(\texttt{Education}) = \{\texttt{HighSchool}, \texttt{BS}, \texttt{MS}, \texttt{PhD}\}$$


a.fahim@ut.ac.ir

# 1.3-Data: Algebraic and Geometric View (داده‌ها از نظر جبری و هندسی)

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}^T \in \mathbb{R}^d$$

$$\mathbf{e}_j = (0, \ldots, 1_j, \ldots, 0)^T$$

Any other vector in $\mathbb{R}^d$ can be written as a linear combination of the standard basis vectors.

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{id}\mathbf{e}_d = \sum_{j=1}^{d} x_{ij}\mathbf{e}_j$$

$\mathbf{D} \in \mathbb{R}^{n \times d}$ $\quad \mathbf{x}_i^T \in \mathbb{R}^d$ $\quad X_j \in \mathbb{R}^n$

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} — \mathbf{x}_1^T — \\ — \mathbf{x}_2^T — \\ \vdots \\ — \mathbf{x}_n^T — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

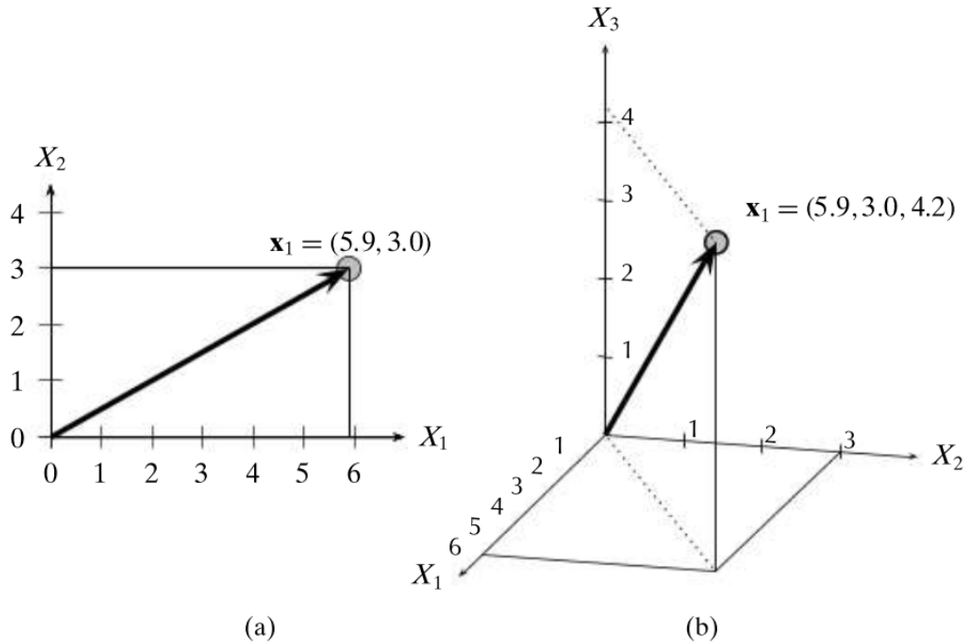



**Figure 1.1.** Row $\mathbf{x}_1$ as a point and vector in (a) $\mathbb{R}^2$ and (b) $\mathbb{R}^3$.

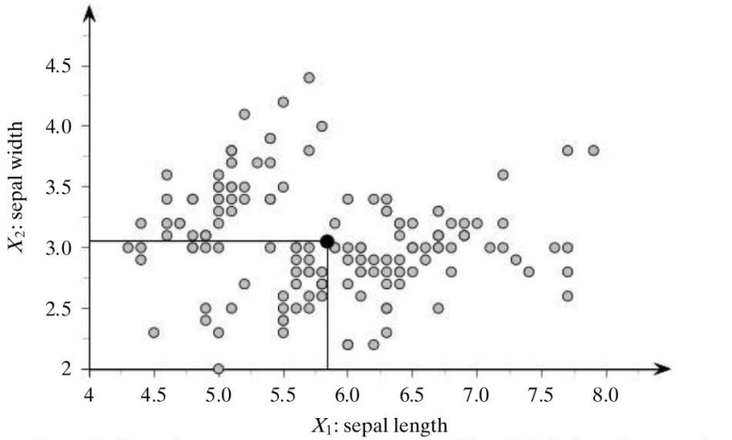



**Figure 1.2.** Scatterplot: `sepal length` versus `sepal width`. The solid circle shows the mean point.

## Distance and Angle

$$\mathbf{a}^T\mathbf{b} = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = a_1b_1 + a_2b_2 + \cdots + a_mb_m = \sum_{i=1}^{m} a_ib_i$$

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_m^2} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

$$\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|} = \left(\frac{1}{\|\mathbf{a}\|}\right)\mathbf{a}$$

$$\|\mathbf{a}\|_p = \left(|a_1|^p + |a_2|^p + \cdots + |a_m|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{m} |a_i|^p\right)^{\frac{1}{p}} \quad \text{for any } p \neq 0.$$

Euclidean distance

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{(\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b})} = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

The cosine of the smallest angle between vectors a and b, also called the cosine similarity

$$\cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right)^T\left(\frac{\mathbf{b}}{\|\mathbf{b}\|}\right) \qquad -1 \leq \cos\theta \leq 1$$

The Cauchy–Schwartz inequality . $|\mathbf{a}^T\mathbf{b}| \leq \|\mathbf{a}\| \cdot \|\mathbf{b}\|$

Orthogonality

$\mathbf{a}^T\mathbf{b} = 0.$

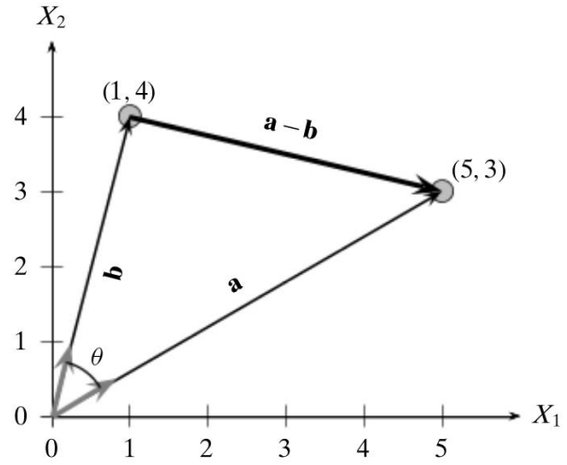



**Figure 1.3.** Distance and angle. Unit vectors are shown in gray.

$$mean(\mathbf{D}) = \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

$$\text{var}(\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 = \frac{1}{n}\left(\sum_{i=1}^{n}\|\mathbf{x}_i\|^2\right) - \|\boldsymbol{\mu}\|^2$$

Centered Data Matrix

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1}\cdot\boldsymbol{\mu}^T = \begin{pmatrix}\mathbf{x}_1^T\\ \mathbf{x}_2^T\\ \vdots\\ \mathbf{x}_n^T\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}^T\\ \boldsymbol{\mu}^T\\ \vdots\\ \boldsymbol{\mu}^T\end{pmatrix} = \begin{pmatrix}\mathbf{x}_1^T - \boldsymbol{\mu}^T\\ \mathbf{x}_2^T - \boldsymbol{\mu}^T\\ \vdots\\ \mathbf{x}_n^T - \boldsymbol{\mu}^T\end{pmatrix} = \begin{pmatrix}\bar{\mathbf{x}}_1^T\\ \bar{\mathbf{x}}_2^T\\ \vdots\\ \bar{\mathbf{x}}_n^T\end{pmatrix}$$

## Orthogonal Projection

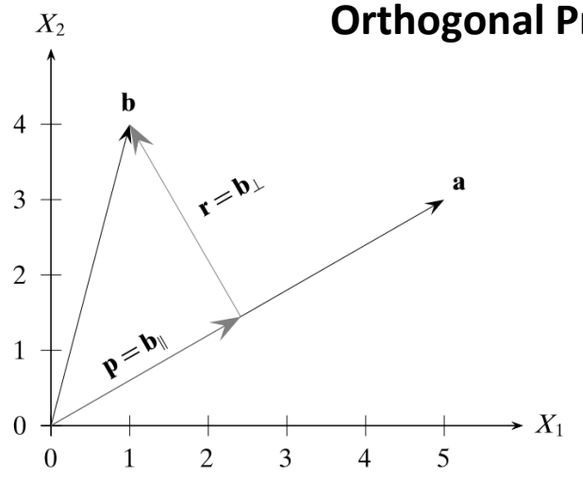



Figure 1.4. Orthogonal projection.

orthogonal decomposition

$$\text{proj}_{\mathbf{a}}(\mathbf{b}) = \left(\frac{\mathbf{b}^T\mathbf{a}}{\mathbf{a}^T\mathbf{a}}\right)$$

$$\mathbf{p} = \text{proj}_{\mathbf{a}}(\mathbf{b})\cdot\mathbf{a}$$

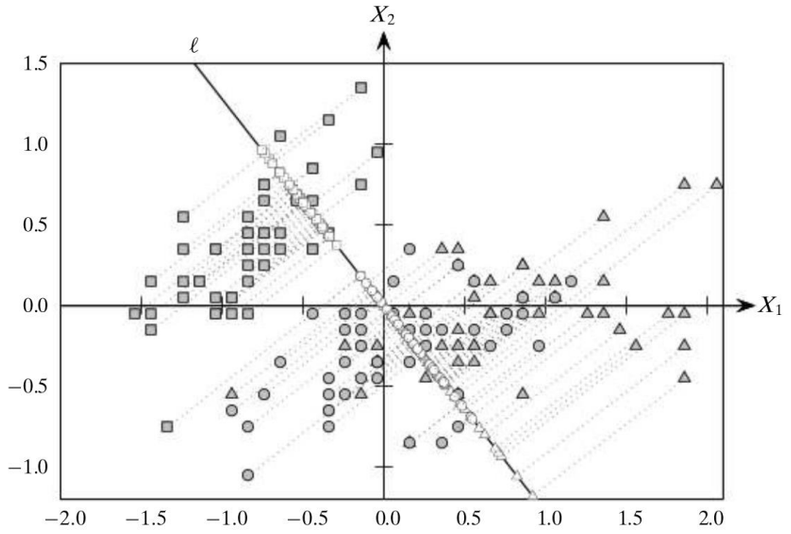



Figure 1.5. Projecting the centered data onto the line $\ell$.

## Linear Independence and Dimensionality

linear combination is given as $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k$

$c_i \in \mathbb{R}$

The set of all possible linear combinations of the k vectors is called the span, denoted as $span(\boldsymbol{v_1}, \ldots, \boldsymbol{v_k})$, which is itself a vector space being a subspace of $\mathbb{R}^m$

$$col(\mathbf{D}) = span(X_1, X_2, \ldots, X_d) \qquad row(\mathbf{D}) = span(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

## Linear Independence

We say that the vectors $v_1, \ldots, v_k$ are linearly dependent if at least one vector can be written as a linear combination of the others.

linearly independent if and only if

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \text{ implies } c_1 = c_2 = \cdots = c_k = 0$$

## Dimension and Rank

Let S be a subspace of $\mathbb{R}^m$. A basis for S is a set of vectors in S, say $v_1, \ldots, v_k$, that are linearly independent and they span S, that is, $span(v_1, \ldots, v_k) = S$.
In fact, a basis is a minimal spanning set.
The standard basis for $\mathbb{R}^m$ is an **orthonormal** basis consisting of the vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \qquad \cdots \qquad \mathbf{e}_m = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

The number of vectors in a basis for S is called the dimension of S, denoted as $\dim(S)$. Because S is a subspace of $\mathbb{R}^m$, we must have $\dim(S) \leq m$.

For any matrix, the dimension of its row and column space is the same, and this dimension is also called the **rank** of the matrix.

For the data matrix $D \in \mathbb{R}^{n \times d}$, we have $rank(D) \leq \min(n, d)$

Thus, even though the data points are ostensibly in a d dimensional attribute space (the **extrinsic dimensionality**), if $rank(D) < d$, then the data points reside in a lower dimensional subspace of $\mathbb{R}^d$, and in this case, $rank(D)$ gives an indication about the **intrinsic dimensionality** of the data.

In fact, with dimensionality reduction methods it is often possible to approximate $D \in \mathbb{R}^{n \times d}$ with a derived data matrix $D' \in \mathbb{R}^{n \times k}$, which has much lower dimensionality, that is, $k \ll d$. In this case k may reflect the **true intrinsic dimensionality** of the data.

Each numeric attribute X is a random variable, defined as a function that assigns a real number to each outcome of an experiment (observation or measurement)

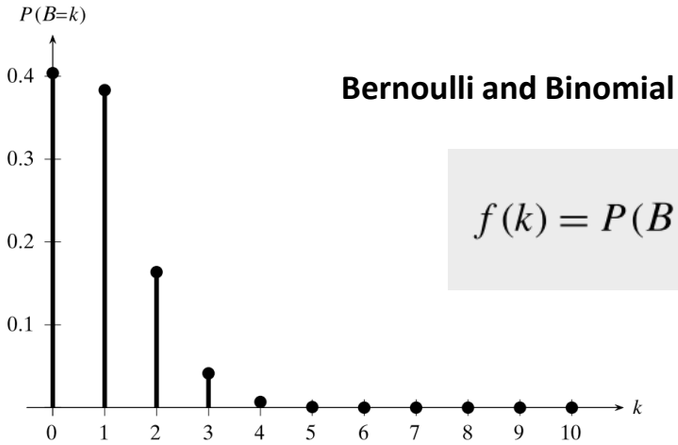
$$X : \mathcal{O} \rightarrow \mathbb{R}$$

where $O$, the domain of X, is the set of all possible outcomes of the experiment, also called the sample space, and R, the range of X, is the set of real numbers.

**Probability Mass Function**

$$\sum_x f(x) = 1 \qquad f(x) = P(X = x) \quad \text{for all } x \in \mathbb{R}$$

$$f(x) \geq 0, \quad \text{for all } x \in \mathbb{R}$$




**Bernoulli and Binomial Distribution**

$$f(k) = P(B = k) = \binom{m}{k} p^k (1-p)^{m-k}$$

Figure 1.6. Binomial distribution: probability mass function ($m = 10$, $p = 0.087$).

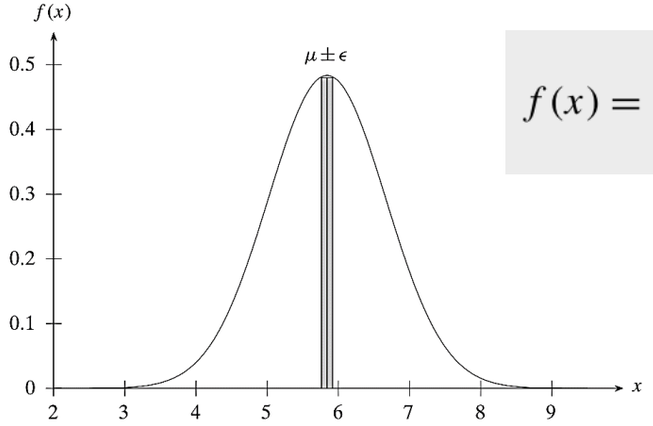
**Probability Density Function**

$$P(X \in [a, b]) = \int_a^b f(x)\, dx$$

$$f(x) \geq 0, \quad \text{for all } x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \qquad f(x) \simeq \frac{P(X \in [x-\epsilon, x+\epsilon])}{2\epsilon}$$

**Gaussian or Normal Distribution**




$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

Figure 1.7. Normal distribution: probability density function ($\mu = 5.84$, $\sigma^2 = 0.681$).

**Cumulative Distribution Function**

$$F(x) = P(X \leq x) \qquad \text{for all } -\infty < x < \infty$$

$$F : \mathbb{R} \rightarrow [0, 1]$$

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u) \qquad F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

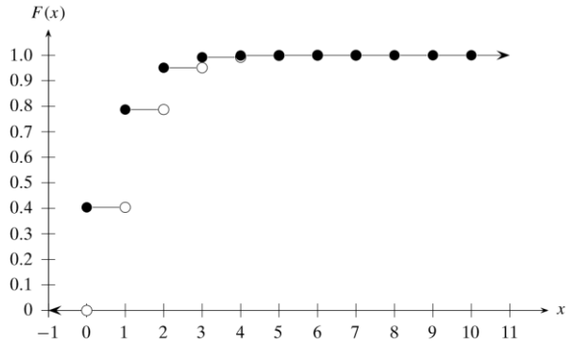



Figure 1.8. Cumulative distribution function for the binomial distribution.

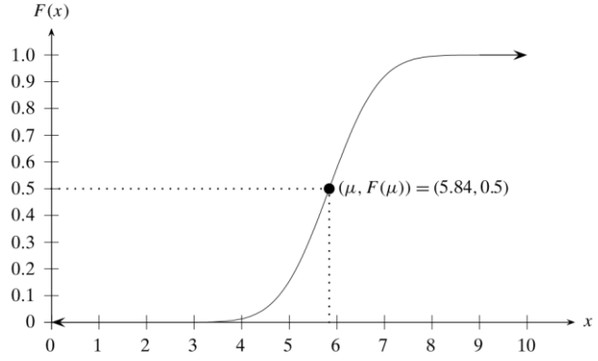



Figure 1.9. Cumulative distribution function for the normal distribution.

**Bivariate Random Variables**

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \mathbf{X} : \mathcal{O} \rightarrow \mathbb{R}^2$$

**Joint Probability Mass Function**

$$f(\mathbf{x}) = f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(\mathbf{X} = \mathbf{x})$$

$$f(\mathbf{x}) = f(x_1, x_2) \geq 0 \qquad \text{for all } -\infty < x_1, x_2 < \infty$$

$$\sum_{\mathbf{x}} f(\mathbf{x}) = \sum_{x_1} \sum_{x_2} f(x_1, x_2) = 1$$

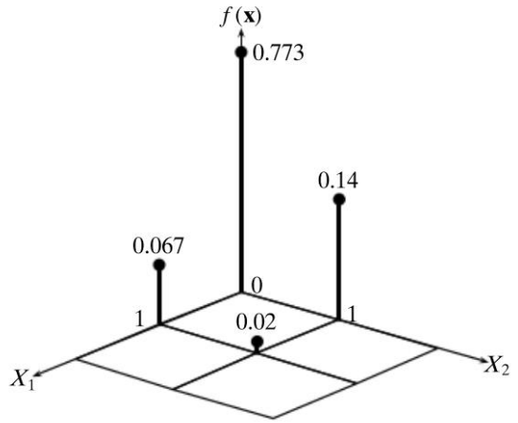



Figure 1.10. Joint probability mass function: $X_1$ (long sepal length), $X_2$ (long sepal width).

**Joint Probability Density Function**

$$P(\mathbf{X} \in W) = \iint_{\mathbf{x} \in W} f(\mathbf{x})\, d\mathbf{x} = \iint_{(x_1, x_2)^T \in W} f(x_1, x_2)\, dx_1\, dx_2$$

$$f(\mathbf{x}) = f(x_1, x_2) \geq 0 \qquad \text{for all } -\infty < x_1, x_2 < \infty$$

$$\int_{\mathbb{R}^2} f(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1\, dx_2 = 1$$

Assuming that X follows a bivariate normal distribution

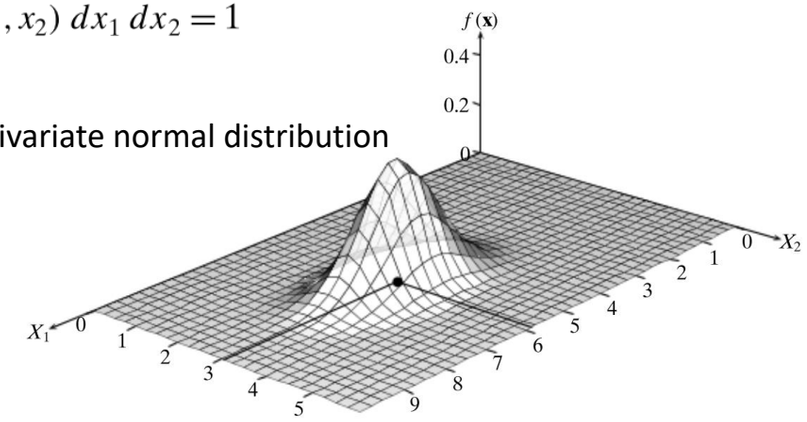



Figure 1.11. Bivariate normal density: $\mu = (5.843, 3.054)^T$ (solid circle).

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{x}-\boldsymbol{\mu})}{2}\right\}$$

**Joint Cumulative Distribution Function**

$$F(\mathbf{x}) = F(x_1, x_2) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2) = P(\mathbf{X} \leq \mathbf{x})$$

**Statistical Independence**

$$P(X_1 \in W_1 \text{ and } X_2 \in W_2) = P(X_1 \in W_1) \cdot P(X_2 \in W_2)$$

$$F(\mathbf{x}) = F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$$

$$f(\mathbf{x}) = f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

**Multivariate Random Variable**

$$\mathbf{X} = (X_1, X_2, \ldots, X_d)^T \qquad \mathbf{X} : \mathcal{O} \rightarrow \mathbb{R}^d$$

$$P(\mathbf{X} \in W) = \int_{\mathbf{x} \in W} \cdots \int f(\mathbf{x})\, d\mathbf{x}$$

$$P\left((X_1, X_2, \ldots, X_d)^T \in W\right) = \int_{(x_1, x_2, \ldots, x_d)^T \in W} \cdots \int f(x_1, x_2, \ldots, x_d)\, dx_1\, dx_2 \ldots dx_d$$

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$$

$$F(x_1, x_2, \ldots, x_d) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d)$$

If $X_1, X_2, \ldots, X_d$ are independent then the following conditions are also satisfied

$$F(\mathbf{x}) = F(x_1, \ldots, x_d) = F_1(x_1) \cdot F_2(x_2) \cdot \ldots \cdot F_d(x_d)$$

$$f(\mathbf{x}) = f(x_1, \ldots, x_d) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_d(x_d)$$

## Random Sample and Statistics

In statistics, the word **population** is used to refer to the set or universe of all entities under study.
However, looking at the entire population may not be feasible or may be too expensive. Instead, we try to make **inferences** about the **population parameters** by drawing a **random sample** from the population, and by computing appropriate **statistics** from the sample that give **estimates** of the corresponding population parameters of interest.

Given a random variable X, a random sample of size n from X is defined as **a set of n independent and identically distributed (IID)** random variables $S_1, S_2, \ldots, S_n$, that is, all of the $S_i$'s are statistically independent of each other, and follow the same probability mass or density function as X.

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

## Multivariate Sample

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_i)$$

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} f_{X_j}(x_{ij})$$

## Statistic

Let $\{S_i\}_{i=1}^{m}$ denote a random sample of size m drawn from a (multivariate) random variable X. A **statistic** $\hat{\theta}$ is some **function over the random sample**, given as:

$$\hat{\theta} : (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m) \rightarrow \mathbb{R}$$

The statistic $\hat{\theta}$ is an estimate of the corresponding population parameter $\theta$.
If we use the value of a statistic to estimate a population parameter, this value is called a **point estimate** of the parameter, and the **statistic** is called **an estimator of the parameter**.