Mohammed J. Zaki$^1$    Wagner Meira Jr.$^2$

$^1$Department of Computer Science  
Rensselaer Polytechnic Institute, Troy, NY, USA  

$^2$Department of Computer Science  
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 2: Numeric Attributes
Univariate analysis focuses on a single attribute at a time. The data matrix $D$ is an $n \times 1$ matrix,

$$
D = \begin{pmatrix}
X \\
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
$$

where $X$ is the numeric attribute of interest, with $x_i \in \mathbb{R}$.

$X$ is assumed to be a random variable, and the observed data a random sample drawn from $X$, i.e., $x_i$’s are independent and identically distributed as $X$.

In the vector view, we treat the sample as an $n$-dimensional vector, and write $X \in \mathbb{R}^n$. 
Empirical Probability Mass Function

The *empirical probability mass function (PMF)* of $X$ is given as

$$
\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x)
$$

where the indicator variable $I$ takes on the value 1 when its argument is true, and 0 otherwise. The empirical PMF puts a probability mass of $\frac{1}{n}$ at each point $x_i$.

The *empirical cumulative distribution function (CDF)* of $X$ is given as

$$
\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)
$$

The *inverse cumulative distribution function* or *quantile function* for $X$ is defined as follows:

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

The inverse CDF gives the least value of $X$, for which $q$ fraction of the values are lower, and $1 - q$ fraction of the values are higher.
The *mean or expected value* of a random variable $X$ is the arithmetic average of the values of $X$. It provides a one-number summary of the *location* or *central tendency* for the distribution of $X$.

If $X$ is discrete, it is defined as

$$
\mu = E[X] = \sum_x xf(x)
$$

where $f(x)$ is the probability mass function of $X$.

If $X$ is continuous it is defined as

$$
\mu = E[X] = \int_{-\infty}^{\infty} xf(x) \, dx
$$

where $f(x)$ is the probability density function of $X$. 

---

Zaki & Meira Jr. (RPI and UFMG)  
Data Mining and Analysis  
Chapter 2: Numeric Attributes
The sample mean is a statistic, that is, a function \( \hat{\mu} : \{x_1, x_2, \ldots, x_n\} \rightarrow \mathbb{R} \), defined as the average value of \( x_i \)'s:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

It serves as an estimator for the unknown mean value \( \mu \) of \( X \).

An estimator \( \hat{\theta} \) is called an unbiased estimator for parameter \( \theta \) if \( E[\hat{\theta}] = \theta \) for every possible value of \( \theta \). The sample mean \( \hat{\mu} \) is an unbiased estimator for the population mean \( \mu \), as

\[
E[\hat{\mu}] = E \left[ \frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu
\]

We say that a statistic is robust if it is not affected by extreme values (such as outliers) in the data. The sample mean is not robust because a single large value can skew the average.
Sample Mean: Iris *sepal length*

\[ \hat{\mu} = 5.843 \]
The *median* of a random variable is defined as the value \( m \) such that

\[
P(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \frac{1}{2}
\]

The median \( m \) is the “middle-most” value; half of the values of \( X \) are less and half of the values of \( X \) are more than \( m \).

In terms of the (inverse) cumulative distribution function, the median is the value \( m \) for which

\[
F(m) = 0.5 \quad \text{or} \quad m = F^{-1}(0.5)
\]

The *sample median* is given as

\[
\hat{F}(m) = 0.5 \quad \text{or} \quad m = \hat{F}^{-1}(0.5)
\]

Median is robust, as it is not affected very much by extreme values.
The *mode* of a random variable $X$ is the value at which the probability mass function or the probability density function attains its maximum value, depending on whether $X$ is discrete or continuous, respectively.

The *sample mode* is a value for which the empirical probability mass function attains its maximum, given as

$$\text{mode}(X) = \arg \max_x \hat{f}(x)$$
Empirical CDF: sepal length
The median is 5.8, since

\[
\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)
\]
The *value range* or simply *range* of a random variable $X$ is the difference between the maximum and minimum values of $X$, given as

$$r = \max\{X\} - \min\{X\}$$

The *sample range* is a statistic, given as

$$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}$$

Range is sensitive to extreme values, and thus is not robust.

A more robust measure of the dispersion of $X$ is the *interquartile range (IQR)*, defined as

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

The *sample IQR* is given as

$$\hat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$
Variance and Standard Deviation

The *variance* of a random variable $X$ provides a measure of how much the values of $X$ deviate from the mean or expected value of $X$

$$
\sigma^2 = \text{var}(X) = \mathbb{E}[(X - \mu)^2] = \begin{cases} 
\sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\
\int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx & \text{if } X \text{ is continuous}
\end{cases}
$$

The *standard deviation* $\sigma$, is the positive square root of the variance, $\sigma^2$.

The *sample variance* is defined as

$$
\hat{\sigma}^2 = \frac{\text{biased}}{n-1} \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2
$$

and the *sample standard deviation* is

$$
\hat{\sigma} = \sqrt{\frac{\text{unbiased}}{n-1} \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2}
$$
The sample values for $X$ comprise a vector in $n$-dimensional space, where $n$ is the sample size. Let $Z$ denote the centered sample

$$Z = X - \mathbf{1} \cdot \hat{\mu} = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of ones. Sample variance is squared magnitude of the centered attribute vector, normalized by the sample size:

$$\hat{\sigma}^2 = \frac{1}{n} \|Z\|^2 = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$
Variance of the Sample Mean and Bias

Sample mean \( \hat{\mu} \) is itself a statistic. We can compute its mean value and variance

\[
E[\hat{\mu}] = \mu
\]

\[
\text{var}(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}
\]

The sample mean \( \hat{\mu} \) varies or deviates from the mean \( \mu \) in proportion to the population variance \( \sigma^2 \). However, the deviation can be made smaller by considering larger sample size \( n \).

The sample variance is a biased estimator for the true population variance, since

\[
E[\hat{\sigma}^2] = \left( \frac{n - 1}{n} \right) \sigma^2
\]

But it is asymptotically unbiased, since

\[
E[\hat{\sigma}^2] \rightarrow \sigma^2 \quad \text{as} \quad n \rightarrow \infty
\]
\[ x_1, x_2, \ldots, x_n \text{ IID with } x \]

\[ \hat{\mu} = \frac{1}{n} \sum x_i \]

\[ \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \hat{\mu})^2 \]

\[ \frac{1}{n} \|Z\|^2 \]

\[ Z \text{ centred vector} \]

\[ E[X] \]

\[ E[(X-\mu)^2] = E[X^2] - (E[X])^2 \]
Bivariate Analysis

In bivariate analysis, we consider two attributes at the same time. The data $D$ comprises an $n \times 2$ matrix:

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

Geometrically, $D$ comprises $n$ points or vectors in 2-dimensional space

$$x_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$$

$D$ can also be viewed as two points or vectors in an $n$-dimensional space:

$$X_1 = (x_{11}, x_{21}, \ldots, x_{n1})^T$$
$$X_2 = (x_{12}, x_{22}, \ldots, x_{n2})^T$$

In the probabilistic view, $X = (X_1, X_2)^T$ is a bivariate vector random variable, and the points $x_i$ ($1 \leq i \leq n$) are a random sample drawn from $X$, that is, $x_i$’s IID with $X$. 

Zaki & Meira Jr. (RPI and UFMG) Data Mining and Analysis Chapter 2: Numeric Attributes 15 / 34
The bivariate mean is defined as the expected value of the vector random variable $\mathbf{X}$:

$$\mathbf{\mu} = E[\mathbf{X}] = E \left[ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

The sample mean vector is given as

$$\hat{\mathbf{\mu}} = \sum_\mathbf{x} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_\mathbf{x} \mathbf{x} \left( \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$
The covariance between two attributes $X_1$ and $X_2$ provides a measure of the association or linear dependence between them, and is defined as

$$
\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - E[X_1]E[X_2]
$$

If $X_1$ and $X_2$ are independent, then

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

which implies that $\sigma_{12} = 0$.

The sample covariance between $X_1$ and $X_2$ is given as

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$
Correlation

The correlation between variables $X_1$ and $X_2$ is the standardized covariance, obtained by normalizing the covariance with the standard deviation of each variable, given as

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

The sample correlation for attributes $X_1$ and $X_2$ is given as

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$$
Geometric Interpretation of Sample Covariance and Correlation

Let $Z_1$ and $Z_2$ denote the centered attribute vectors in $\mathbb{R}^n$:

$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix}$$

$$Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

The sample covariance is given as

$$\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$$

and the sample correlation is

$$\hat{\rho}_{12} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1} \sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\| \|Z_2\|} = \left( \frac{Z_1}{\|Z_1\|} \right)^T \left( \frac{Z_2}{\|Z_2\|} \right) = \cos \theta$$

The correlation coefficient is simply the cosine of the angle between the two centered attribute vectors.
Geometric Interpretation of Covariance and Correlation

\[ \theta \]

\[ Z_1 \]

\[ Z_2 \]

\[ x_1 \]

\[ x_2 \]

\[ x_n \]
The variance–covariance information for the two attributes $X_1$ and $X_2$ can be summarized in the square $2 \times 2$ covariance matrix

$$\Sigma = E[(X - \mu)(X - \mu)^T] = \begin{pmatrix} \sigma^2_1 & \sigma_{12} \\ \sigma_{21} & \sigma^2_2 \end{pmatrix}$$

Because $\sigma_{12} = \sigma_{21}$, $\Sigma$ is symmetric.

The total variance is given as

$$\text{var}(D) = \text{tr}(\Sigma) = \sigma^2_1 + \sigma^2_2$$

We immediately have $\text{tr}(\Sigma) \geq 0$.

The generalized variance is

$$|\Sigma| = \det(\Sigma) = \sigma^2_1 \sigma^2_2 - \sigma^2_{12} = \sigma^2_1 \sigma^2_2 - \rho^2_{12} \sigma^2_1 \sigma^2_2 = (1 - \rho^2_{12}) \sigma^2_1 \sigma^2_2$$

Note that $|\rho_{12}| \leq 1$ implies that $\det(\Sigma) \geq 0$. 
Correlation: sepal length and sepal width

The sample mean is

\[ \hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix} \]

The sample covariance matrix is

\[ \hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix} \]

The sample correlation is

\[ \hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109 \]
In multivariate analysis we consider all the $d$ numeric attributes $X_1, X_2, \ldots, X_d$.

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

In the row view, the data is a set of $n$ points or vectors in the $d$-dimensional attribute space

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \in \mathbb{R}^d$$

In the column view, the data is a set of $d$ points or vectors in the $n$-dimensional space spanned by the data points

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in \mathbb{R}^n$$
Mean and Covariance

In the probabilistic view, the $d$ attributes are modeled as a vector random variable, $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$, and the points $\mathbf{x}_i$ are considered to be a random sample drawn from $\mathbf{X}$, i.e., IID with $\mathbf{X}$.

The \textit{multivariate mean vector} is

$$\mu = E[\mathbf{X}] = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_d)^T$$

The \textit{sample mean} is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

The (sample) covariance matrix is a $d \times d$ (square) symmetric matrix

$$\Sigma = \begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2
\end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix}
\hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\
\hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2
\end{pmatrix}$$
\[ \Sigma \] is a \textit{positive semidefinite} matrix, that is, \( a^T \Sigma a \geq 0 \) for any \( d \)-dimensional vector \( a \).

To see this, observe that

\[
\begin{align*}
a^T \Sigma a &= a^T E[(X - \mu)(X - \mu)^T] a \\
&= E[a^T (X - \mu)(X - \mu)^T a] \\
&= E[Y^2] \\
&\geq 0
\end{align*}
\]

Because \( \Sigma \) is also symmetric, this implies that all the eigenvalues of \( \Sigma \) are real and non-negative, and they can be arranged from the largest to the smallest as follows: \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0 \).

The total variance is given as: \( \text{var}(D) = \prod_{i=1}^{d} \sigma_i^2 \)

The generalized variance is \( \text{det}(\Sigma) = \prod_{i=1}^{d} \lambda_i \geq 0 \)
Eigenvalues & Eigenvectors of $\Sigma$

$\Sigma$: 1) Square  
2) Symmetric  
3) Positive Semi-definite

$d \times d$

$\Sigma \vec{a} \rightarrow \vec{b}$

Linear transformation

$\Sigma \vec{u} = \lambda \vec{u}$

Vector $\vec{u}$ is invariant under $\Sigma$ multiplication except for scaling
If $\Sigma$ is full rank, then $\text{rank } = d$.

$$\sum \tilde{u}_i = \lambda_i \tilde{u}_i$$

$d$ equations

$$\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_d \leftarrow \text{eigenvectors}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$$

$k \ll d$

$\tilde{u}_1 \leftarrow \text{dominant eigenvector}$

$1^{st} \text{ PC} \leftarrow \text{principal component}$

only for $\Sigma$
\[ \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \Sigma I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]

\[ \Sigma u = \lambda u \]

\[(\Sigma u - I \lambda u) = 0 \]

\[(\Sigma - \lambda I) \vec{u} = 0 \]

\[\det(\Sigma - \lambda I) = 0 \]

characteristic equation to solve for \( \lambda \)
\[ \Sigma - 2I = \begin{pmatrix} 2-x & 1 \\ 1 & 1-2 \end{pmatrix} \]

\[ \det(\Sigma - 2I) = (2-x)(1-2) - 1 = 0 \]

\[ = 2 - 3x + x^2 - 1 = 0 \]

\[ \frac{x^2 - 3x + 1}{a^2 \cdot b \cdot c} = 0 \]

\[ 3 \pm \sqrt{9 - 4} \]

\[ \lambda_1 = \frac{3 + \sqrt{5}}{2} \]

\[ \lambda_2 = \frac{3 - \sqrt{5}}{2} \]
Sample Covariance Matrix: Inner and Outer Product

Let \( Z \) represent the centered data matrix

\[
Z = D - 1 \cdot \hat{\mu}^T = \begin{pmatrix}
\vdots & \vdots & \vdots & \vdots \\
- & - & \vdots & - \\
- & - & \vdots & - \\
\vdots & \ddots & \ddots & \ddots \\
- & - & - & -
\end{pmatrix} = \begin{pmatrix}
Z_1 \\
Z_2 \\
\vdots \\
Z_d
\end{pmatrix}
\]

where \( z_i = x_i - \hat{\mu} \) is a centered point, and \( Z_i = X_i - 1 \cdot \hat{\mu}_i \) is a centered attribute.

Inner product and outer product form for sample covariance matrix:

\[
\hat{\Sigma} = \frac{1}{n} (Z^T Z) = \frac{1}{n} \begin{pmatrix}
Z_1^T Z_1 & Z_1^T Z_2 & \cdots & Z_1^T Z_d \\
Z_2^T Z_1 & Z_2^T Z_2 & \cdots & Z_2^T Z_d \\
\vdots & \vdots & \ddots & \vdots \\
Z_d^T Z_1 & Z_d^T Z_2 & \cdots & Z_d^T Z_d
\end{pmatrix}
\]

i.e., \( \hat{\Sigma} \) is given as the pairwise *inner or dot products* of the centered attribute vectors, normalized by the sample size, or as a sum of rank-one matrices obtained as an *outer product* of each centered point.
Data Normalization

If the attribute values are in vastly different scales, then it is necessary to normalize them.

**Range Normalization:** Let $X$ be an attribute and let $x_1, x_2, \ldots, x_n$ be a random sample drawn from $X$. In *range normalization* each value is scaled by the sample range $\hat{r}$ of $X$:

$$x'_i = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

After transformation the new attribute takes on values in the range $[0, 1]$.

**Standard Score Normalization:** Also called $z$-normalization; each value is replaced by its $z$-score:

$$x'_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

where $\hat{\mu}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance of $X$. After transformation, the new attribute has mean $\hat{\mu}' = 0$, and standard deviation $\hat{\sigma}' = 1$. 

Zaki & Meira Jr. (RPI and UFMG)  Data Mining and Analysis  Chapter 2: Numeric Attributes  27 / 34
Normalization Example

Since Income is much larger, it dominates Age. The sample range for Age is \( \hat{r} = 40 - 12 = 28 \), whereas for Income it is \( 2700 - 300 = 2400 \).

For range normalization, the point \( \mathbf{x}_2 = (14, 500) \) is scaled to

\[
\mathbf{x}_2' = \left( \frac{14 - 12}{28}, \frac{500 - 300}{2400} \right) = (0.071, 0.035)
\]

For z-normalization, we have

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\mu} )</td>
<td>27.2</td>
<td>2680</td>
</tr>
<tr>
<td>( \hat{\sigma} )</td>
<td>9.77</td>
<td>1726.15</td>
</tr>
</tbody>
</table>

Thus, \( \mathbf{x}_2 = (14, 500) \) is scaled to

\[
\mathbf{x}_2' = \left( \frac{14 - 27.2}{9.77}, \frac{500 - 2680}{1726.15} \right) = (-1.35, -1.26)
\]
The normal distribution plays an important role as the parametric distribution of choice in clustering, density estimation, and classification.

A random variable \( X \) has a normal distribution, with the parameters mean \( \mu \) and variance \( \sigma^2 \), if the probability density function of \( X \) is given as follows:

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

The term \((x - \mu)^2\) measures the distance of a value \( x \) from the mean \( \mu \) of the distribution, and thus the probability density decreases exponentially as a function of the distance from the mean.

The maximum value of the density occurs at the mean value \( x = \mu \), given as \( f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \), which is inversely proportional to the standard deviation \( \sigma \) of the distribution.
Normal Distribution: $\mu = 0$, and Different Variances

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

For $\sigma = 0.5$, $1$, and $2$. The graph shows the distribution with different variances. The peak of the distribution shifts to the left as the variance decreases, and the width of the distribution increases as the variance increases.
Multivariate Normal Distribution

Given the \(d\)-dimensional vector random variable \(X = (X_1, X_2, \ldots, X_d)^T\), it has a multivariate normal distribution, with the parameters mean \(\mu\) and covariance matrix \(\Sigma\), if its joint multivariate probability density function is given as follows:

\[
f(x|\mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp \left\{ -\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right\}
\]

where \(|\Sigma|\) is the determinant of the covariance matrix.

The term

\[
(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)
\]

measures the distance, called the Mahalanobis distance of the point \(x\) from the mean \(\mu\) of the distribution, taking into account all of the variance–covariance information between the attributes.
Standard Bivariate Normal Density

Parameters: \( \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \)
Geometry of the Multivariate Normal

Compared to the standard multivariate normal, the mean $\mu$ translates the center of the distribution, whereas the covariance matrix $\Sigma$ scales and rotates the distribution. The eigen-decomposition of $\Sigma$ is given as

$$\Sigma u_i = \lambda_i u_i$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ are the eigenvalues and $u_i$ the corresponding eigenvectors. This can be expressed compactly as follows:

$$\Sigma = U\Lambda U^T$$

where

$$\Lambda = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_d \\
\end{pmatrix}$$

and

$$U = \begin{pmatrix} u_1 & u_2 & \cdots & u_d \end{pmatrix}$$

The eigenvectors represent the new basis vectors, with the covariance matrix given by $\Lambda$ (all covariances become zero). Since the trace of a square matrix is invariant to similarity transformation, such as a change of basis, we have

$$\text{var}(D) = \text{tr}(\Sigma) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i = \text{tr}(\Lambda)$$
Bivariate Normal for Iris: sepal length and sepal width

\[ \hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix} \]

\[ \hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix} \]

We have

\[ \hat{\Sigma} = U \Lambda U^T \]

\[ U = \begin{pmatrix} -0.997 & -0.078 \\ 0.078 & -0.997 \end{pmatrix} \]

\[ \Lambda = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix} \]

Angle of rotation is:

\[ \cos \theta = e_1^T u_1 = -0.997 \]

or \( \theta = 175.5^\circ \)