CHAPTER 2 Numeric Attributes

In this chapter, we discuss basic statistical methods for exploratory data analysis of numeric attributes. We look at measures of central tendency or location, measures of dispersion, and measures of linear dependence or association between attributes. We emphasize the connection between the probabilistic and the geometric and algebraic views of the data matrix.

2.1 UNIVARIATE ANALYSIS

Univariate analysis focuses on a single attribute at a time; thus the data matrix **D** can be thought of as an $n \times 1$ matrix, or simply a column vector, given as

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

where X is the numeric attribute of interest, with $x_i \in \mathbb{R}$. X is assumed to be a random variable, with each point x_i $(1 \leq i \leq n)$ itself treated as an identity random variable. We assume that the observed data is a random sample drawn from X, that is, each variable x_i is independent and identically distributed as X. In the vector view, we treat the sample as an n-dimensional vector, and write $X \in \mathbb{R}^n$.

In general, the probability density or mass function $f(x)$ and the cumulative distribution function $F(x)$, for attribute X, are both unknown. However, we can estimate these distributions directly from the data sample, which also allow us to compute statistics to estimate several important population parameters.

Empirical Cumulative Distribution Function

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

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{2.1}$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

is a binary *indicator variable* that indicates whether the given condition is satisfied or not. Intuitively, to obtain the empirical CDF we compute, for each value $x \in \mathbb{R}$, how many points in the sample are less than or equal to x. The empirical CDF puts a probability mass of $\frac{1}{n}$ at each point x_i. Note that we use the notation $\hat{F}$ to denote the fact that the empirical CDF is an estimate for the unknown population CDF F.

Inverse Cumulative Distribution Function

Define the *inverse cumulative distribution function* or *quantile function* for a random variable X as follows:

$$F^{-1}(q) = \min\{x \mid \hat{F}(x) \geq q\} \qquad \text{for } q \in [0,1] \tag{2.2}$$

That is, the inverse CDF gives the least value of X, for which q fraction of the values are higher, and $1-q$ fraction of the values are lower. The *empirical inverse cumulative distribution function* $\hat{F}^{-1}$ can be obtained from Eq. (2.1).

Empirical Probability Mass Function

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

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x) \tag{2.3}$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

The empirical PMF also puts a probability mass of $\frac{1}{n}$ at each point x_i.

2.1.1 Measures of Central Tendency

These measures given an indication about the concentration of the probability mass, the “middle” values, and so on.

Mean

The *mean*, also called the *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.

The mean or expected value of a discrete random variable X is defined as

$$\mu = E[X] = \sum_{x} x f(x) \tag{2.4}$$

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

The expected value of a continuous random variable X is defined as

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

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

Sample Mean The *sample mean* is a statistic, that is, a function $\hat{\mu} : \{x_1, x_2, \ldots, x_n\} \to \mathbb{R}$, defined as the average value of x_i's:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.5}$$

It serves as an estimator for the unknown mean value μ of X. It can be derived by plugging in the empirical PMF $\hat{f}(x)$ in Eq. (2.4):

$$\hat{\mu} = \sum_x x \hat{f}(x) = \sum_x x \left(\frac{1}{n} \sum_{i=1}^{n} I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample Mean Is Unbiased An estimator $\hat{\theta}$ is called an *unbiased estimator* for parameter θ if $E[\hat{\theta}] = \theta$ for every possible value of θ. The sample mean $\hat{\mu}$ is an unbiased estimator for the population mean μ, as

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \tag{2.6}$$

where we use the fact that the random variables x_i are IID according to X, which implies that they have the same mean μ as X, that is, $E[x_i] = \mu$ for all x_i. We also used the fact that the expectation function E is a *linear operator*, that is, for any two random variables X and Y, and real numbers a and b, we have $E[aX + bY] = aE[X] + bE[Y]$.

Robustness 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 unfortunately not robust because a single large value (an outlier) can skew the average. A more robust measure is the *trimmed mean* obtained after discarding a small fraction of extreme values on one or both ends. Furthermore, the mean can be somewhat misleading in that it is typically not a value that occurs in the sample, and it may not even be a value that the random variable can actually assume (for a discrete random variable). For example, the number of cars per capita is an integer-valued random variable, but according to the US Bureau of Transportation Studies, the average number of passenger cars in the United States was 0.45 in 2008 (137.1 million cars, with a population size of 304.4 million). Obviously, one cannot own 0.45 cars; it can be interpreted as saying that on average there are 45 cars per 100 people.

Median

The *median* of a random variable is defined as the value m such that

$$P(X \le m) \ge \frac{1}{2} \text{ and } P(X \ge m) \ge \frac{1}{2}$$

In other words, 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 therefore the value m for which

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

The *sample median* can be obtained from the empirical CDF [Eq. (2.1)] or the empirical inverse CDF [Eq. (2.2)] by computing

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

A simpler approach to compute the sample median is to first sort all the values x_i ($i \in [1, n]$) in increasing order. If n is odd, the median is the value at position $\frac{n+1}{2}$. If n is even, the values at positions $\frac{n}{2}$ and $\frac{n}{2} + 1$ are both medians.

Unlike the mean, median is robust, as it is not affected very much by extreme values. Also, it is a value that occurs in the sample and a value the random variable can actually assume.

Mode

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 [Eq. (2.3)] attains its maximum, given as

$$\text{mode}(X) = \arg\max_{x} \hat{f}(x)$$

The mode may not be a very useful measure of central tendency for a sample because by chance an unrepresentative element may be the most frequent element. Furthermore, if all values in the sample are distinct, each of them will be the mode.

Example 2.1 (Sample Mean, Median, and Mode). Consider the attribute `sepal length` (X_1) in the Iris dataset, whose values are shown in Table 1.2. The sample mean is given as follows:

$$\hat{\mu} = \frac{1}{150}(5.9 + 6.9 + \cdots + 7.7 + 5.1) = \frac{876.5}{150} = 5.843$$

Figure 2.1 shows all 150 values of `sepal length`, and the sample mean. Figure 2.2a shows the empirical CDF and Figure 2.2b shows the empirical inverse CDF for `sepal length`.

Because $n = 150$ is even, the sample median is the value at positions $\frac{n}{2} = 75$ and $\frac{n}{2} + 1 = 76$ in sorted order. For `sepal length` both these values are 5.8; thus the sample median is 5.8. From the inverse CDF in Figure 2.2b, we can see that

$$\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)$$

The sample mode for `sepal length` is 5, which can be observed from the frequency of 5 in Figure 2.1. The empirical probability mass at $x = 5$ is

$$\hat{f}(5) = \frac{10}{150} = 0.067$$

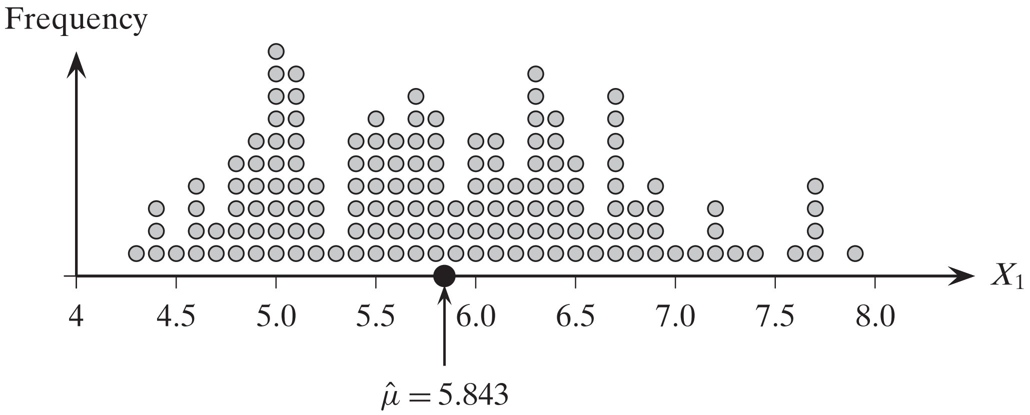
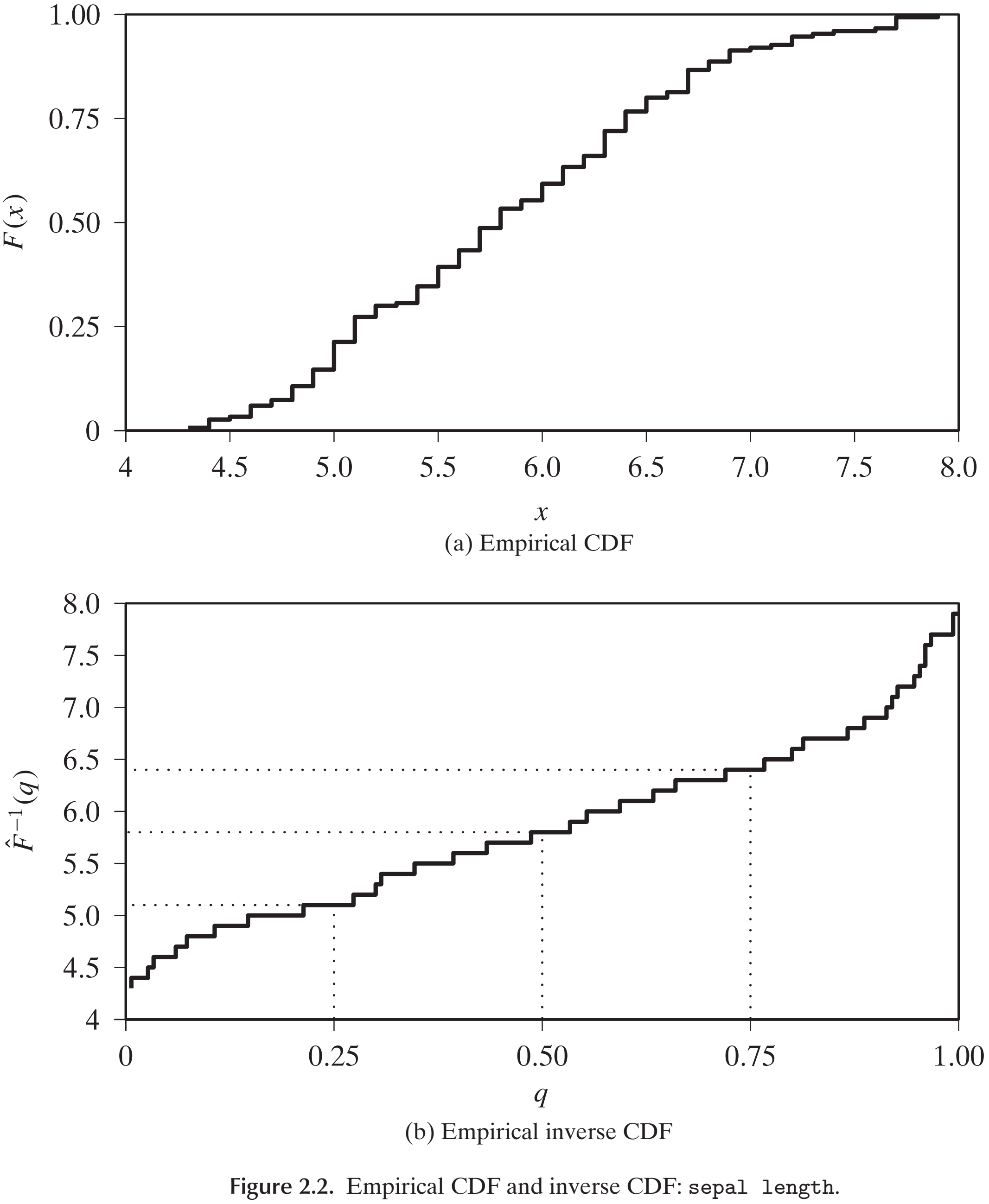


Figure 2.1. Sample mean for `sepal length`. Multiple occurrences of the same value are shown stacked.



(a) Empirical CDF

(b) Empirical inverse CDF

Figure 2.2. Empirical CDF and inverse CDF: `sepal length`.

2.1.2 Measures of Dispersion

The measures of dispersion give an indication about the spread or variation in the values of a random variable.

Range

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 (value) range of X is a population parameter, not to be confused with the range of the function X, which is the set of all the values X can assume. Which range is being used should be clear from the context.

The *sample range* is a statistic, given as

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

By definition, range is sensitive to extreme values, and thus is not robust.

Interquartile Range

Quartiles are special values of the quantile function [Eq. (2.2)] that divide the data into four equal parts. That is, quartiles correspond to the quantile values of 0.25, 0.5, 0.75, and 1.0. The *first quartile* is the value $q_1 = F^{-1}(0.25)$, to the left of which 25% of the points lie; the *second quartile* is the same as the median value $q_2 = F^{-1}(0.5)$, to the left of which 50% of the points lie; the third quartile $q_3 = F^{-1}(0.75)$ is the value to the left of which 75% of the points lie; and the fourth quartile is the maximum value of X, to the left of which 100% of the points lie.

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

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25) \tag{2.7}$$

IQR can also be thought of as a *trimmed range*, where we discard 25% of the low and high values of X. Or put differently, it is the range for the middle 50% of the values of X. IQR is robust by definition.

The *sample IQR* can be obtained by plugging in the empirical inverse CDF in Eq. (2.7):

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

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. More formally, variance is the expected value of the squared deviation from the mean, defined as

$$\sigma^2 = var(X) = E[(X-\mu)^2] = \begin{cases} \displaystyle\sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases} \tag{2.8}$$

The *standard deviation*, σ, is defined as the positive square root of the variance, σ^2.

We can also write the variance as the difference between the expectation of X^2 and the square of the expectation of X:

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned} \tag{2.9}$$

It is worth noting that variance is in fact the *second moment about the mean*, corresponding to $r = 2$, which is a special case of the *rth moment about the mean* for a random variable X, defined as $E[(\mathbf{x}-\mu)^r]$.

Sample Variance The *sample variance* is defined as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{2.10}$$

It is the average squared deviation of the data values x_i from the sample mean $\hat{\mu}$, and can be derived by plugging in the empirical probability function $\hat{f}$ from Eq. (2.3) into Eq. (2.8), as

$$\hat{\sigma}^2 = \sum_x (x-\hat{\mu})^2 \hat{f}(x) = \sum_x (x-\hat{\mu})^2 \left(\frac{1}{n}\sum_{i=1}^{n} I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

The *sample standard deviation* is given as the positive square root of the sample variance:

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

The *standard score*, also called the *z-score*, of a sample value x_i is the number of standard deviations the value is away from the mean:

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

Put differently, the z-score of x_i measures the deviation of x_i from the mean value $\hat{\mu}$, in units of $\hat{\sigma}$.

Geometric Interpretation of Sample Variance We can treat the data sample for attribute X as a vector in n-dimensional space, where n is the sample size. That is, we write $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$. Further, let

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

denote the mean subtracted attribute vector, where $\mathbf{1} \in \mathbb{R}^n$ is the n-dimensional vector all of whose elements have value 1. We can rewrite Eq. (2.10) in terms of the magnitude of Z, that is, the dot product of Z with itself:

$$\hat{\sigma}^2 = \frac{1}{n}\, \|Z\|^2 = \frac{1}{n}\, Z^T Z = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{2.11}$$

The sample variance can thus be interpreted as the squared magnitude of the centered attribute vector, or the dot product of the centered attribute vector with itself, normalized by the sample size.

Example 2.2. Consider the data sample for `sepal length` shown in Figure 2.1. We can see that the sample range is given as

$$\max_i\{x_i\} - \min_i\{x_i\} = 7.9 - 4.3 = 3.6$$

From the inverse CDF for `sepal length` in Figure 2.2b, we can find the sample IQR as follows:

$$\hat{q}_1 = \hat{F}^{-1}(0.25) = 5.1$$
$$\hat{q}_3 = \hat{F}^{-1}(0.75) = 6.4$$
$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = 6.4 - 5.1 = 1.3$$

The sample variance can be computed from the centered data vector via Eq. (2.11):

$$\hat{\sigma}^2 = \frac{1}{n}(X - \mathbf{1} \cdot \hat{\mu})^T (X - \mathbf{1} \cdot \hat{\mu}) = 102.168/150 = 0.681$$

The sample standard deviation is then

$$\hat{\sigma} = \sqrt{0.681} = 0.825$$

Variance of the Sample Mean Because the sample mean $\hat{\mu}$ is itself a statistic, we can compute its mean value and variance. The expected value of the sample mean is simply μ, as we saw in Eq. (2.6). To derive an expression for the variance of the sample mean, we utilize the fact that the random variables x_i are all independent, and thus

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i)$$

Further, because all the x_i's are identically distributed as X, they have the same variance as X, that is,

$$var(x_i) = \sigma^2 \text{ for all } i$$

Combining the above two facts, we get

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2 \tag{2.12}$$

Further, note that

$$E\left[\sum_{i=1}^{n} x_i\right] = n\mu \tag{2.13}$$

Using Eqs. (2.9), (2.12), and (2.13), the variance of the sample mean $\hat{\mu}$ can be computed as

$$\begin{aligned} var(\hat{\mu}) = E[(\hat{\mu}-\mu)^2] = E[\hat{\mu}^2] - \mu^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \frac{1}{n^2}E\left[\sum_{i=1}^{n} x_i\right]^2 \\ = \frac{1}{n^2}\left(E\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] - E\left[\sum_{i=1}^{n} x_i\right]^2\right) = \frac{1}{n^2}\, var\left(\sum_{i=1}^{n} x_i\right) \\ = \frac{\sigma^2}{n} \end{aligned} \tag{2.14}$$

In other words, the sample mean $\hat{\mu}$ varies or deviates from the mean μ in proportion to the population variance σ^2. However, the deviation can be made smaller by considering larger sample size n.

Sample Variance Is Biased, but Is Asymptotically Unbiased The sample variance in Eq. (2.10) is a *biased estimator* for the true population variance, σ^2, that is, $E[\hat{\sigma}^2] \neq \sigma^2$. To show this we make use of the identity

$$\sum_{i=1}^{n}(x_i-\mu)^2 = n(\hat{\mu}-\mu)^2 + \sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{2.15}$$

Computing the expectation of $\hat{\sigma}^2$ by using Eq. (2.15) in the first step, we get

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right] - E[(\hat{\mu}-\mu)^2] \tag{2.16}$$

Recall that the random variables x_i are IID according to X, which means that they have the same mean μ and variance σ^2 as X. This means that

$$E[(x_i - \mu)^2] = \sigma^2$$

Further, from Eq. (2.14) the sample mean $\hat{\mu}$ has variance $E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$. Plugging these into the Eq. (2.16) we get

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{1}{n} n\sigma^2 - \frac{\sigma^2}{n} \\ &= \left(\frac{n-1}{n}\right)\sigma^2 \end{aligned}$$

The sample variance $\hat{\sigma}^2$ is a biased estimator of σ^2, as its expected value differs from the population variance by a factor of $\frac{n-1}{n}$. However, it is *asymptotically unbiased*, that is, the bias vanishes as $n \to \infty$ because

$$\lim_{n\to\infty} \frac{n-1}{n} = \lim_{n\to\infty} 1 - \frac{1}{n} = 1$$

Put differently, as the sample size increases, we have

$$E[\hat{\sigma}^2] \to \sigma^2 \qquad \text{as } n \to \infty$$

2.2 BIVARIATE ANALYSIS

In bivariate analysis, we consider two attributes at the same time. We are specifically interested in understanding the association or dependence between them, if any. We thus restrict our attention to the two numeric attributes of interest, say X_1 and X_2, with the data $\mathbf{D}$ represented as an $n \times 2$ matrix:

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

Geometrically, we can think of $\mathbf{D}$ in two ways. It can be viewed as n points or vectors in 2-dimensional space over the attributes X_1 and X_2, that is, $\mathbf{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$. Alternatively, it can be viewed as two points or vectors in an n-dimensional space comprising the points, that is, each column is a vector in $\mathbb{R}^n$, as follows:

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

In the probabilistic view, the column vector $\mathbf{X} = (X_1, X_2)^T$ is considered a bivariate vector random variable, and the points $\mathbf{x}_i$ $(1 \le i \le n)$ are treated as a random sample drawn from $\mathbf{X}$, that is, $\mathbf{x}_i$'s are considered independent and identically distributed as $\mathbf{X}$.

Empirical Joint Probability Mass Function

The *empirical joint probability mass function* for $\mathbf{X}$ is given as

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \tag{2.17}$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^{n} I(x_{i1} = x_1, x_{i2} = x_2)$$

where $\mathbf{x} = (x_1, x_2)^T$ and I is a indicator variable that takes on the value 1 only when its argument is true:

$$I(\mathbf{x}_i = \mathbf{x}) = \begin{cases} 1 & \text{if } x_{i1} = x_1 \text{ and } x_{i2} = x_2 \\ 0 & \text{otherwise} \end{cases}$$

As in the univariate case, the probability function puts a probability mass of $\frac{1}{n}$ at each point in the data sample.

2.2.1 Measures of Location and Dispersion

Mean

The bivariate mean is defined as the expected value of the vector random variable $\mathbf{X}$, defined as follows:

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{2.18}$$

In other words, the bivariate mean vector is simply the vector of expected values along each attribute.

The sample mean vector can be obtained from $\hat{f}_{X_1}$ and $\hat{f}_{X_2}$, the empirical probability mass functions of X_1 and X_2, respectively, using Eq. (2.5). It can also be computed from the joint empirical PMF in Eq. (2.17)

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left(\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{2.19}$$

Variance

We can compute the variance along each attribute, namely σ_1^2 for X_1 and σ_2^2 for X_2 using Eq. (2.8). The *total variance* [Eq. (1.4)] is given as

$$var(\mathbf{D}) = \sigma_1^2 + \sigma_2^2$$

The sample variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be estimated using Eq. (2.10), and the *sample total variance* is simply $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$.

2.2.2 Measures of Association

Covariance

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

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] \tag{2.20}$$

By linearity of expectation, we have

$$\begin{aligned}\sigma_{12} &= E[(X_1-\mu_1)(X_2-\mu_2)]\\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2]\\ &= E[X_1X_2] - \mu_2 E[X_1] - \mu_1 E[X_2] + \mu_1\mu_2\\ &= E[X_1X_2] - \mu_1\mu_2\\ &= E[X_1X_2] - E[X_1]E[X_2] \qquad (2.21)\end{aligned}$$

Eq. (2.21) can be seen as a generalization of the univariate variance [Eq. (2.9)] to the bivariate case.

If X_1 and X_2 are independent random variables, then we conclude that their covariance is zero. This is because if X_1 and X_2 are independent, then we have

$$E[X_1X_2] = E[X_1] \cdot E[X_2]$$

which in turn implies that

$$\sigma_{12} = 0$$

However, the converse is not true. That is, if $\sigma_{12} = 0$, one cannot claim that X_1 and X_2 are independent. All we can say is that there is no linear dependence between them, but we cannot rule out that there might be a higher order relationship or dependence between the two attributes.

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

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\hat{\mu}_1)(x_{i2}-\hat{\mu}_2) \qquad (2.22)$$

It can be derived by substituting the empirical joint probability mass function $\hat{f}(x_1, x_2)$ from Eq. (2.17) into Eq. (2.20), as follows:

$$\begin{aligned}\hat{\sigma}_{12} &= E[(X_1-\hat{\mu}_1)(X_2-\hat{\mu}_2)]\\ &= \sum_{\mathbf{x}=(x_1,x_2)^T}(x_1-\hat{\mu}_1)(x_2-\hat{\mu}_2)\hat{f}(x_1,x_2)\\ &= \frac{1}{n}\sum_{\mathbf{x}=(x_1,x_2)^T}\sum_{i=1}^{n}(x_1-\hat{\mu}_1)\cdot(x_2-\hat{\mu}_2)\cdot I(x_{i1}=x_1, x_{i2}=x_2)\\ &= \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\hat{\mu}_1)(x_{i2}-\hat{\mu}_2)\end{aligned}$$

Notice that sample covariance is a generalization of the sample variance [Eq. (2.10)] because

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_1)(x_i-\mu_1) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_1)^2 = \hat{\sigma}_1^2$$

and similarly, $\hat{\sigma}_{22} = \hat{\sigma}_2^2$.

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

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2\sigma_2^2}} \tag{2.23}$$

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

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\sum_{i=1}^{n}(x_{i1}-\hat{\mu}_1)(x_{i2}-\hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\hat{\mu}_1)^2\sum_{i=1}^{n}(x_{i2}-\hat{\mu}_2)^2}} \tag{2.24}$$

Geometric Interpretation of Sample Covariance and Correlation

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

$$Z_1 = X_1 - \mathbf{1}\cdot\hat{\mu}_1 = \begin{pmatrix} x_{11}-\hat{\mu}_1 \\ x_{21}-\hat{\mu}_1 \\ \vdots \\ x_{n1}-\hat{\mu}_1 \end{pmatrix} \qquad Z_2 = X_2 - \mathbf{1}\cdot\hat{\mu}_2 = \begin{pmatrix} x_{12}-\hat{\mu}_2 \\ x_{22}-\hat{\mu}_2 \\ \vdots \\ x_{n2}-\hat{\mu}_2 \end{pmatrix}$$

The sample covariance [Eq. (2.22)] can then be written as

$$\hat{\sigma}_{12} = \frac{Z_1^TZ_2}{n}$$

In other words, the covariance between the two attributes is simply the dot product between the two centered attribute vectors, normalized by the sample size. The above can be seen as a generalization of the univariate sample variance given in Eq. (2.11).

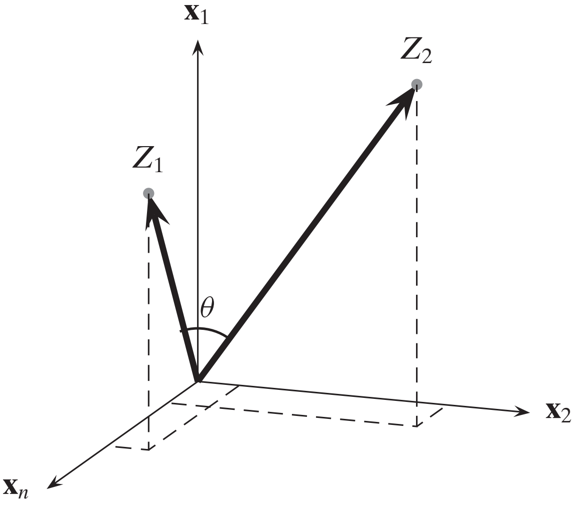


Figure 2.3. Geometric interpretation of covariance and correlation. The two centered attribute vectors are shown in the (conceptual) n-dimensional space $\mathbb{R}^n$ spanned by the n points.

The sample correlation [Eq. (2.24)] can be written as

$$\hat{\rho}_{12} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1}\sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\|\ \|Z_2\|} = \left(\frac{Z_1}{\|Z_1\|}\right)^T \left(\frac{Z_2}{\|Z_2\|}\right) = \cos\theta \tag{2.25}$$

Thus, the correlation coefficient is simply the cosine of the angle [Eq. (1.3)] between the two centered attribute vectors, as illustrated in Figure 2.3.

Covariance Matrix

The variance–covariance information for the two attributes X_1 and X_2 can be summarized in the square 2×2 *covariance matrix*, given as

$$\begin{aligned}
\mathbf{\Sigma} &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\
&= E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & X_2 - \mu_2 \end{pmatrix}\right] \\
&= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}
\end{aligned} \tag{2.26}$$

Because $\sigma_{12} = \sigma_{21}$, $\mathbf{\Sigma}$ is a *symmetric* matrix. The covariance matrix records the attribute specific variances on the main diagonal, and the covariance information on the off-diagonal elements.

The *total variance* of the two attributes is given as the sum of the diagonal elements of $\mathbf{\Sigma}$, which is also called the *trace* of $\mathbf{\Sigma}$, given as

$$var(\mathbf{D}) = tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2$$

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

The *generalized variance* of the two attributes also considers the covariance, in addition to the attribute variances, and is given as the *determinant* of the covariance matrix $\mathbf{\Sigma}$, denoted as $|\mathbf{\Sigma}|$ or $\det(\mathbf{\Sigma})$. The generalized covariance is non-negative, because

$$|\mathbf{\Sigma}| = \det(\mathbf{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 = (1 - \rho_{12}^2)\sigma_1^2\sigma_2^2$$

where we used Eq. (2.23), that is, $\sigma_{12} = \rho_{12}\sigma_1\sigma_2$. Note that $|\rho_{12}| \leq 1$ implies that $\rho_{12}^2 \leq 1$, which in turn implies that $\det(\mathbf{\Sigma}) \geq 0$, that is, the determinant is non-negative.

The *sample covariance matrix* is given as

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$

The sample covariance matrix $\widehat{\mathbf{\Sigma}}$ shares the same properties as $\mathbf{\Sigma}$, that is, it is symmetric and $|\widehat{\mathbf{\Sigma}}| \geq 0$, and it can be used to easily obtain the sample total and generalized variance.

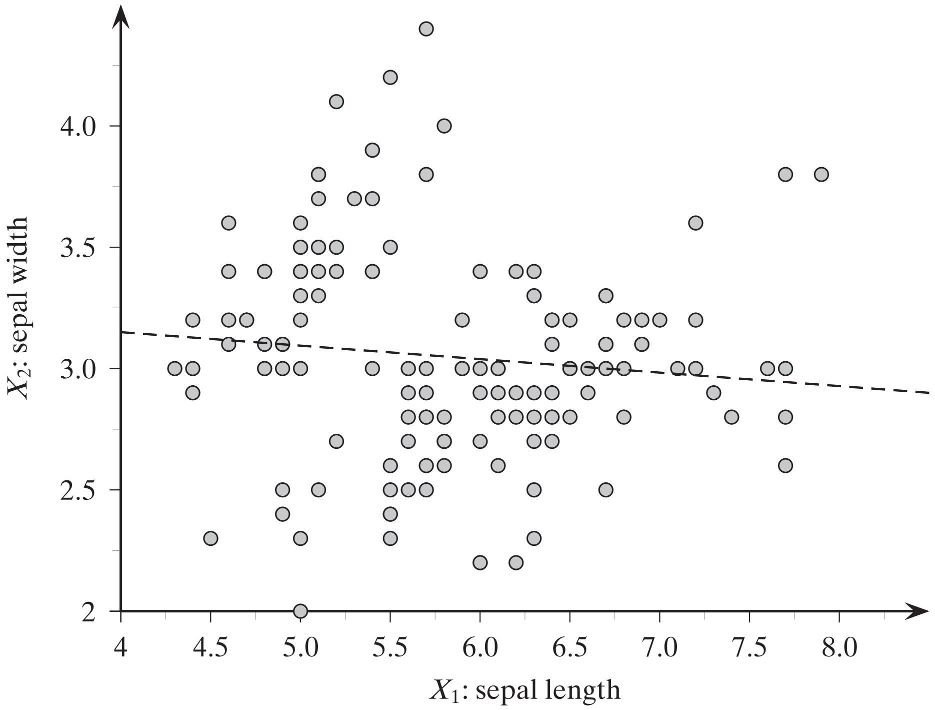


Figure 2.4. Correlation between `sepal length` and `sepal width`.

Example 2.3 (Sample Mean and Covariance). Consider the `sepal length` and `sepal width` attributes for the Iris dataset, plotted in Figure 2.4. There are $n = 150$ points in the $d = 2$ dimensional attribute space. The sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

The sample covariance matrix is given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The variance for `sepal length` is $\hat{\sigma}_1^2 = 0.681$, and that for `sepal width` is $\hat{\sigma}_2^2 = 0.187$. The covariance between the two attributes is $\hat{\sigma}_{12} = -0.039$, and the correlation between them is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Thus, there is a very weak negative correlation between these two attributes, as evidenced by the best linear fit line in Figure 2.4. Alternatively, we can consider the attributes `sepal length` and `sepal width` as two points in $\mathbb{R}^n$. The correlation is then the cosine of the angle between them; we have

$$\hat{\rho}_{12} = \cos\theta = -0.109, \text{ which implies that } \theta = \cos^{-1}(-0.109) = 96.26^\circ$$

The angle is close to 90°, that is, the two attribute vectors are almost orthogonal, indicating weak correlation. Further, the angle being greater than 90° indicates negative correlation.

The sample total variance is given as

$$tr(\widehat{\mathbf{\Sigma}}) = 0.681 + 0.187 = 0.868$$

and the sample generalized variance is given as

$$|\widehat{\mathbf{\Sigma}}| = \det(\widehat{\mathbf{\Sigma}}) = 0.681 \cdot 0.187 - (-0.039)^2 = 0.126$$

2.3 MULTIVARIATE ANALYSIS

In multivariate analysis, we consider all the d numeric attributes $X_1, X_2, \ldots, X_d$. The full data is an $n \times d$ matrix, given as

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

In the row view, the data can be considered as a set of n points or vectors in the d-dimensional attribute space

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

In the column view, the data can be considered as a set of d points or vectors in the n-dimensional space spanned by the data points

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

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 $\mathbf{X}$, that is, they are independent and identically distributed as $\mathbf{X}$.

Mean

Generalizing Eq. (2.18), the *multivariate mean vector* is obtained by taking the mean of each attribute, given as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

Generalizing Eq. (2.19), the *sample mean* is given as

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

Covariance Matrix

Generalizing Eq. (2.26) to d dimensions, the multivariate covariance information is captured by the $d \times d$ (square) symmetric *covariance matrix* that gives the covariance for each pair of attributes:

$$\boldsymbol{\Sigma} = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

The diagonal element σ_i^2 specifies the attribute variance for X_i, whereas the off-diagonal elements $\sigma_{ij} = \sigma_{ji}$ represent the covariance between attribute pairs X_i and X_j.

Covariance Matrix Is Positive Semidefinite

It is worth noting that $\boldsymbol{\Sigma}$ is a *positive semidefinite* matrix, that is,

$$\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} \geq 0 \text{ for any } d\text{-dimensional vector } \mathbf{a}$$

To see this, observe that

$$\begin{aligned} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} &= \mathbf{a}^T E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T]\mathbf{a} \\ &= E[\mathbf{a}^T(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{a}] \\ &= E[Y^2] \\ &\geq 0 \end{aligned}$$

where Y is the random variable $Y = \mathbf{a}^T(\mathbf{X}-\boldsymbol{\mu}) = \sum_{i=1}^d a_i(X_i - \mu_i)$, and we use the fact that the expectation of a squared random variable is non-negative.

Because $\boldsymbol{\Sigma}$ is also symmetric, this implies that all the eigenvalues of $\boldsymbol{\Sigma}$ are real and non-negative. In other words the d eigenvalues of $\boldsymbol{\Sigma}$ can be arranged from the largest to the smallest as follows: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. A consequence is that the determinant of $\boldsymbol{\Sigma}$ is non-negative:

$$\det(\boldsymbol{\Sigma}) = \prod_{i=1}^d \lambda_i \geq 0 \qquad (2.27)$$

Total and Generalized Variance

The total variance is given as the trace of the covariance matrix:

$$var(\mathbf{D}) = tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2 \qquad (2.28)$$

Being a sum of squares, the total variance must be non-negative.

The generalized variance is defined as the determinant of the covariance matrix, $\det(\boldsymbol{\Sigma})$, also denoted as $|\boldsymbol{\Sigma}|$. It gives a single value for the overall multivariate scatter. From Eq. (2.27) we have $\det(\boldsymbol{\Sigma}) \geq 0$.