Multivariate Summary Statistics

Arnab Maity

NCSU Department of Statistics ~ 5240 SAS Hall ~ 919-515-1937 ~ amaity[at]ncsu.edu

Contents

Review of univariate framework

Consider the `iris` data in `R`.[1] For simplicity, let us only consider the `setosa` species and the `Sepal.length` variable. Suppose we want to estimate the mean sepal.length of the setosa flower. Such a statistical problem has four main components.

(1) **Population:** A group of individuals/objects/items of interest. In our example, the population consists of *all* setosa flowers.

(2) **Parameter:** A summary of the population. In our case, the parameter is the true mean of the sepal length of all setosa flowers.

(3) **Sample:** A subset of the population. Naturally, it is often impossible to observe data on the whole population due to time/resource constraints. Thus, we usually collect data on a subset of the population. In our example, the sample consists of measurements on 50 setosa flowers.

(4) **Statistic:** A summary computed from a sample. We use such summaries to *estimate* the unknown parameter. In our case, we estimate the population mean by the *sample mean*.

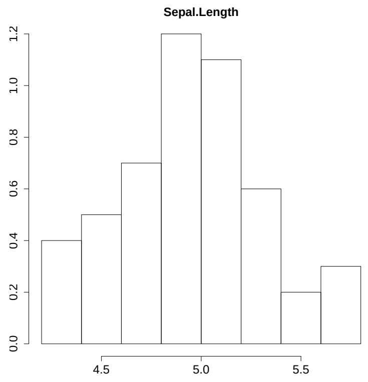

Figure 1: Histogram of Sepal.length in the iris dataset for setosa species.

We denote a hypothetical sample of size n as a collection of random variables $X_1, X_2, \ldots, X_n$, where X_i denotes the `sepal length` of i-th flower. A common assumption is that the population mean (true mean sepal length) is μ and the population variance is σ^2. In general, we assume that $X_1, \ldots, X_n$ form a random sample.

Random sample

The collection of random variables, $X_1, \ldots, X_n$, is called a random sample of size n if $X_1, \ldots, X_n$ are *independent* and each X_i has the *same distribution*.

Thus we have

$$E(X_i) = \mu \text{ and } var(X_i) = \sigma^2 \text{ for all } i.$$

Since we want to know about the population mean μ, a natural way to estimate this is to use the sample mean

$$\bar{X} = (X_1 + \ldots + X_n)/n.$$

Specifically, we call $\bar{X}$ an *estimator* of μ.[2]

Estimator

An estimator is a *formula/rule* that one can apply to *any possible sample*. It does not depend on the true value of the parameter.

[1] *Source:* Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179-188.

[2] Note that we do not need actual numeric data to define a reasonable estimator.

Now we consider the actual sample (observed numeric data) at hand. We have 50 observations (values taken from the `iris` dataset):[3]

$$x_1 = 5.1, x_2 = 4.9, \ldots, x_{50} = 5.$$

Thus the observed value of the sample mean for this sample is

$$\bar{x} = 5.006.$$

The specific value $\bar{x} = 5.006$ is called an *estimate* of μ.

Estimate

An estimate is a *numeric value* that is obtained by applying an estimator to a *specific sample* at hand.

An estimate alone does not give us any indication of how reliable it is. Typically, along with the estimate, one also reports the *standard error* of the estimate.

Standard error

Standard error of an estimator is defined as

$$SE(Estimator) = \sqrt{var(Estimator)}.$$

In our case, the estimator is $\bar{X}$. The standard error of $\bar{X}$ is computed as[4]

$$SE(\bar{X}) = \sqrt{var(\bar{X})} = \sqrt{\sigma^2/n}.$$

Notice that $SE(\bar{X})$ depends on σ^2, the unknown population variance. In practice, we estimate the population variance σ^2 by the observed *sample variance* s^2.[5] Thus the estimated standard error is

$$\widehat{SE}(\bar{X}) = \sqrt{s^2/n}.$$

In `R`, we can compute the estimate and its standard error as follows.

```r
# A snapshot of iris data
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

[3] Notice that we used **lower case letters (e.g., x_i) to denote the observed data but upper case letters (e.g., X_i) to denote random varibles**. We will use this convention throughout this course to differtiate between random variables and observed (numeric) values of the random variables in a particular sample.

[4] Recall that

$$\begin{aligned} var(\bar{X}) &= var[\tfrac{1}{n}(X_1 + \ldots + X_n)] \\ &= \tfrac{1}{n^2}[var(X_1) + \ldots + var(X_n)] \\ &= \tfrac{1}{n^2}[n\sigma^2] = \sigma^2/n \end{aligned}$$

[5] An *estimator* of σ^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The corresponding *estimate* is the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

```r
# Get the species only take the setosa flowers and extract
# only sepal.length (the first column)
species <- iris$Species
SL <- iris[species == "setosa", 1]

# sample size, Sample mean, Sample variance
n <- length(SL)
xbar.SL <- mean(SL)
s2.SL <- var(SL)

# Standard error of xbar
SE <- sqrt(s2.SL/n)

# output
out <- c(xbar.SL, SE)
names(out) <- c("Mean.SL", "SE")
out
```

```
##    Mean.SL         SE
## 5.00600000 0.04984957
```

Random vector and sample mean

Let us now consider the multivariate problem: estimate the mean of all the four variables: Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width for the setosa flowers.

The basic framework presented in the last section remains the same; however, now we have four measurements for each flower:

$$\begin{pmatrix} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{pmatrix} = \begin{pmatrix} SL \\ SW \\ PL \\ PW \end{pmatrix}.$$

So each observation is not a scalar random variable; instead each observation is a *random vector*.

Random vector

The vector $\boldsymbol{X}_{p\times 1} = (X_1, \ldots, X_p)^T$ is called a random vector if each element X_i is a random variable.

In our particular example, $p = 4$. So our random sample in this case consists of the following random vectors:[6]

$$\boldsymbol{X}_1 = \begin{pmatrix} \text{sepal length of 1st flower} \\ \text{sepal width of 1st flower} \\ \text{petal length of 1st flower} \\ \text{petal width of 1st flower} \end{pmatrix} = \begin{pmatrix} SL_1 \\ SW_1 \\ PL_1 \\ PW_1 \end{pmatrix}, \ldots, \boldsymbol{X}_n = \begin{pmatrix} SL_n \\ SW_n \\ PL_n \\ PW_n \end{pmatrix}.$$

Our parameter of interest is the *mean vector*

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} E(SL) \\ E(SW) \\ E(PL) \\ E(PW) \end{pmatrix} = \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix},$$

where μ_{SL} = population mean of `sepal length`, μ_{SW} = population mean of `sepal width`, and so on. *Thus the parameter is a* 4×1 *vector.*

Similar to the univariate case, the estimator of the vector $\boldsymbol{\mu}$ is the *sample mean*.

Sample mean

Given a set of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the sample mean is defined as

$$\bar{\boldsymbol{X}} = \frac{1}{n}(\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

[6] Recall that be default we take vectors as column vectors.

The observed sample (numeric data for our particular sample) are

$$x_1 = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}, \ldots, x_{50} = \begin{pmatrix} 5 \\ 3.3 \\ 1.4 \\ 0.2 \end{pmatrix}.$$

Thus the estimate of μ is

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n).$$

In R we can compute this estimate as follows:[7]

```
setosa <- iris[species == "setosa", 1:4]
head(setosa)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

```
xbar <- colMeans(setosa)
xbar
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
```

How to quantify the variability in $\bar{X}$?[8] To understand this, we need quantify the variability of a random vector. This is done by computing the *variance-covariance matrix*.

Variance-covariance matrix

Let us first discuss the concept of *covariance between two scalar random variables*. Suppose X_1 and X_2 are two scalar random variables. One way to measure the *degree of linear relationship* between X_1 and X_2 is to compute the *covariance* between them.

Covariance

We define the covariance between two random variables X_1 and X_2 as

$$cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2,$$

where $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$.

[7] The `colMeans` function computes mean of each column. Can you do this using `apply`? Using `sweep`?

[8] In other words, how to define a concept like "standard error" in this case?

Covariance measures the *strength of linear relationship* between X_1 and X_2. The quantity $cov(X_1, X_2)$ takes positive values if larger values of X_1 pair with larger values of X_2, and takes negative if larger values of X_1 pair with smaller values of X_2. Zero or "small" values of covariance indicate that there is no *linear* relationship (i.e., slope is zero) between X_1 and X_2.

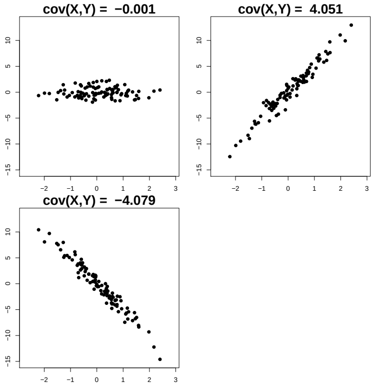

Figure 2: Examples of positive, negative and near zero covariance

Notice that each if the random variable also has its own variance, that is, $var(X_1)$ and $V(X_2)$. Thus to get a complete picture of variability of X_1 and X_2, we need to look at all these quantities:

$$var(X_1), var(X_2), \text{ and } cov(X_1, X_2).$$

Clearly, as the number of variables increases, the number of such quantities increases as well.

In the multivariate world, there is a nice way to summarize the variability of a set of random variables using matrices. Let us consider a random vector

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

The "variability" of $\boldsymbol{X}$ can be summarized by the 2×2 matrix

$$\boldsymbol{\Sigma} = cov(\boldsymbol{X}) = \begin{pmatrix} var(X_1) & cov(X_1, X_2) \\ cov(X_2, X_1) & var(X_2) \end{pmatrix}.$$

This matrix is called the *variance-covariance matrix* of $\boldsymbol{X}$.[9]

Variance-covariance matrix

Suppose we have a $p \times 1$ random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. The variance-covariance matrix of $\boldsymbol{X}$ is defined as the following $p \times p$ matrix:

$$\boldsymbol{\Sigma} = cov(\boldsymbol{X}) = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_p) \\ cov(X_2, X_1) & var(X_2) & \ldots & cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & \ldots & var(X_p) \end{pmatrix}.$$

Typically, the population variance-covariance matrix is unknown. We can estimate $\boldsymbol{\Sigma}$ by the *sample variance-covariance matrix*, denoted by $\boldsymbol{S}$.

[9] Notice that $\boldsymbol{\Sigma}$ is symmetric since $cov(X_1, X_2) = cov(X_2, X_1)$.

Sample covariance matrix

Given a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the sample covariance is

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T.$$

Here S is an *estimator* of $\mathbf{\Sigma}$. Given a numeric sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the corresponding *estimate* is

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

In R, we can compute sample covariance matrix directly by using the formula above or using the `cov()` function. In our specific example, we can compute S as below (rounded to 3 decimal places).

```
setosa <- iris[species == "setosa", 1:4]
colnames(setosa) <- c("SL", "SW", "PL", "PW")
S <- cov(setosa)
# Rounded to 3 digits
round(S, 3)
```

```
##       SL    SW    PL    PW
## SL 0.124 0.099 0.016 0.010
## SW 0.099 0.144 0.012 0.009
## PL 0.016 0.012 0.030 0.006
## PW 0.010 0.009 0.006 0.011
```

Much like the univariate case, we can compute

$$cov(\bar{\mathbf{X}}) = \mathbf{\Sigma}/n,$$

and we can estimate this quantity by replacing $\mathbf{\Sigma}$ by its estimator $\mathbf{S}$,

$$\widehat{cov(\bar{\mathbf{X}})} = \mathbf{S}/n.$$

In our example, we estimate $cov(\bar{\mathbf{X}})$ as follows.

```
# Sample size (number of setosa flowers)
n <- nrow(setosa)
# S/n, rounded to 5 digits
round(S/n, 5)
```

```
##         SL      SW      PL      PW
## SL 0.00248 0.00198 0.00033 0.00021
## SW 0.00198 0.00287 0.00023 0.00019
## PL 0.00033 0.00023 0.00060 0.00012
## PW 0.00021 0.00019 0.00012 0.00022
```

Linear combination of variables

Given a random vector, $\boldsymbol{X} = (X_1, \dots, X_p)^T$, often we are interested in weighted sums of the random variables. Such sums are called *linear combinations*.

Linear combination

A *linear combination* of a collection of random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_p$, is defined as

$$a_1 X_1 + a_2 X_2 + \dots + a_p X_p,$$

where $a_1, \dots, a_p$ are constants.
Define the vector of constants $\boldsymbol{a} = (a_1, \dots, a_p)^T$ and the random vector $\boldsymbol{X} = (X_1, \dots, X_p)^T$. Then the linear combination can be written as

$$a_1 X_1 + a_2 X_2 + \dots + a_p X_p = \boldsymbol{a}^T \boldsymbol{X}.$$

One linear combination

Using the `iris` data, suppose we want to estimate the difference between mean sepal length and mean sepal width for the setosa flowers. In this case, we are interested in the parameter $\mu_{SL} - \mu_{SW}$.[10] Thus we can write

$$\mu_{SL} - \mu_{SW} = E(SL) - E(SW) = E(SL - SW) = E(\boldsymbol{a}^T \boldsymbol{X}) = \boldsymbol{a}^T \boldsymbol{\mu},$$

where $\boldsymbol{a} = (1, -1, 0, 0)^T$. So here we are interested in estimating the mean of a linear combination.

To estimate this parameter, we first calculate the differences[11]

$$D_1 = SL_1 - SW_1 = \boldsymbol{a}^T \boldsymbol{X}_1, \dots, D_n = SL_n - SW_n = \boldsymbol{a}^T \boldsymbol{X}_n.$$

So an estimator of $\mu_{SL} - \mu_{SW}$ is[12]

$$\bar{D} = \frac{1}{n} \sum_i D_i = \frac{1}{n} \sum_i (SL_i - SW_i) = \overline{SL} - \overline{SW} = \boldsymbol{a}^T \bar{\boldsymbol{X}}.$$

The estimator above is expected since we can estimate μ_{SL} by $\overline{SL}$ and μ_{SW} by $\overline{SW}$. So it is natural to estimate $\mu_{SL} - \mu_{SW}$ by $\overline{SL} - \overline{SW}$.

To compute the variance of the estimator, we notice

$$var(\bar{D}) = \frac{var(SL - SW)}{n} = \frac{var(SL) + var(SW) - 2cov(SL, SW)}{n}$$

Note that the terms $var(SL), var(SW)$ and $cov(SL, SW)$ are from the population covariance matrix $\boldsymbol{\Sigma}$. It can be shown that in this case

$$var(\bar{D}) = \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}/n.$$

[10] Recall that our random vector is

$$\boldsymbol{X} = \begin{pmatrix} SL \\ SW \\ PL \\ PW \end{pmatrix}.$$

[11] Recall that our random sample is

$$\boldsymbol{X}_1 = \begin{pmatrix} SL_1 \\ SW_1 \\ PL_1 \\ PW_1 \end{pmatrix}, \dots, \boldsymbol{X}_n = \begin{pmatrix} SL_n \\ SW_n \\ PL_n \\ PW_n \end{pmatrix}.$$

[12] Recall, estimator of $\boldsymbol{\mu}$ is

$$\bar{\boldsymbol{X}} = \begin{pmatrix} \overline{SL} \\ \overline{SW} \\ \overline{PL} \\ \overline{PW} \end{pmatrix}.$$

General result

If $\boldsymbol{Y}$ is a random vector with mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$, and $\boldsymbol{a}$ is a vector, then

$$E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\mu} \text{ and } var(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{a}.$$

This result does *not* depend of the distribution of $\boldsymbol{Y}$.

In our particular case, we want to estimate $\boldsymbol{a}^T\boldsymbol{\mu}$, and our estimator is $\boldsymbol{a}^T\bar{\boldsymbol{X}}$. We know that $cov(\bar{\boldsymbol{X}}) = \boldsymbol{\Sigma}/n$. So we can verify that

$$var(\boldsymbol{a}^T\bar{\boldsymbol{X}}) = \boldsymbol{a}^T cov(\bar{\boldsymbol{X}})\boldsymbol{a} = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}/n.$$

Since we do not know $\boldsymbol{\Sigma}$, we replace $\boldsymbol{\Sigma}$ by $\boldsymbol{S}$.

Using R, we can compute the estimates as below.[13]

```
# Define the coefficient/contrast vector a
a <- c(1, -1, 0, 0)
a
```

```
## [1]  1 -1  0  0
```

```
# Estimate a^T\mu by a^T X-bar
t(a) %*% xbar
```

```
##       [,1]
## [1,] 1.578
```

```
# Estimate the variance of the estimator
t(a) %*% (S/n) %*% a
```

```
##             [,1]
## [1,] 0.001390122
```

Multiple linear combinations

Suppose we want to know how different `sepal width`, `petal length` and `petal width` are from `sepal length` *on average*. Specifically, we want to estimate

$$\begin{pmatrix} \mu_{SW} - \mu_{SL} \\ \mu_{PL} - \mu_{SL} \\ \mu_{PW} - \mu_{SL} \end{pmatrix}.$$

Thus the vector of contrasts can be written as[14]

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix} = \boldsymbol{A\mu}$$

[13] We have already computed `xbar` and `S` in previous sections.

[14] Recall that our original parameter is the mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix}.$$

Since $\bar{\boldsymbol{X}}$ is an estimator of $\boldsymbol{\mu}$, we can simply replace $\boldsymbol{\mu}$ in the quantity above by $\bar{\boldsymbol{X}}$, and say that $\boldsymbol{A}\bar{\boldsymbol{X}}$ is an estimator of $\boldsymbol{A}\boldsymbol{\mu}$.

Similar to before, we can compute the variance-covariance matrix of this estimator as

$$cov(\boldsymbol{A}\bar{\boldsymbol{X}}) = \boldsymbol{A}cov(\bar{\boldsymbol{X}})\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{\Sigma}/n)\boldsymbol{A}^T.$$

Since $\boldsymbol{\Sigma}$ is unknown, we can replace $\boldsymbol{\Sigma}$ by $\boldsymbol{S}$. In our example, we demonstrate these results as follows.

```
# Define the coefficient matrix A
A <- cbind(c(-1, -1, -1), c(1, 0, 0), c(0, 1, 0), c(0, 0, 1))
A
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   -1    1    0    0
## [2,]   -1    0    1    0
## [3,]   -1    0    0    1
```

```
# Estimate A\mu by A X-bar
A %*% xbar
```

```
##        [,1]
## [1,] -1.578
## [2,] -3.544
## [3,] -4.760
```

```
# Estimate the variance-covariance matrix
A %*% (S/n) %*% t(A)
```

```
##              [,1]         [,2]        [,3]
## [1,] 0.0013901224 0.0004075102 0.000480000
## [2,] 0.0004075102 0.0024339592 0.002072653
## [3,] 0.0004800000 0.0020726531 0.002293878
```

Practice

For each of the situations described below, write the parameter and its estimator using vectors/matrices. Clearly define $\boldsymbol{a}$ or $\boldsymbol{A}$ as appropriate.

1. Estimate the average of sepal width and petal width of setosa flowers.

2. Simultaneously estimate the average of sepal width and length, and the mean difference of petal width and length of setosa flowers.

Correlation

A disadvantage of covariance is that it is unbounded, and depends on the unit of measurement. A better measure of linear relationship between two random variables X_1 and X_2 is the correlation coefficient:

$$cor(X_1, X_2) = \frac{cov(X_1, X_2)}{\sqrt{var(X_1)}\sqrt{var(X_2)}}.$$

Correlation coefficient is bounded between -1 and 1. Large positive values indicate a strong positive relationship, and vice versa. Small values indicate absence of no linear relationship. See Figure 3 for an example.

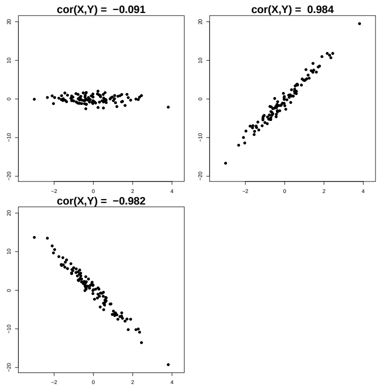

Figure 3: Examples of strong positive, strong negative and near zero correlation

Now suppose we have a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. The correlation matrix of $\boldsymbol{X}$ is

$$cor(\boldsymbol{X}) = \begin{pmatrix} 1 & cor(X_1, X_2) & \ldots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \ldots & cor(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ cor(X_p, X_1) & cor(X_p, X_2) & \ldots & 1 \end{pmatrix}.$$

Note that the diagonal entries are 1 since $cor(X_i, X_i) = 1$.

Typically, $cor(\boldsymbol{X})$ is unknown and can be estimated using the sample by the *sample correlation* matrix $\boldsymbol{R}$.

Given sample data, we can use the `cor()` function to compute $\boldsymbol{R}$.

```
# Sample correlation matrix
R <- cor(setosa)
round(R, 3)
```

```
##       SL    SW    PL    PW
## SL 1.000 0.743 0.267 0.278
## SW 0.743 1.000 0.178 0.233
## PL 0.267 0.178 1.000 0.332
## PW 0.278 0.233 0.332 1.000
```

Note that the correlation between `SL` and `SW` is quite high which indicates a moderate to strong linear relationship between `SL` and `SW`.