Exercise 5.6.14 If A is an $m \times n$ matrix, it can be proved that there exists a unique $n \times m$ matrix $A^\#$ satisfying the following four conditions: $AA^\#A = A$; $A^\#AA^\# = A^\#$; $AA^\#$ and $A^\#A$ are symmetric. The matrix $A^\#$ is called the **generalized inverse** of A, or the **Moore-Penrose** inverse.

a. If A is square and invertible, show that $A^\# = A^{-1}$.

b. If rank $A = m$, show that $A^\# = A^T(AA^T)^{-1}$.

c. If rank $A = n$, show that $A^\# = (A^TA)^{-1}A^T$.

5.7 An Application to Correlation and Variance

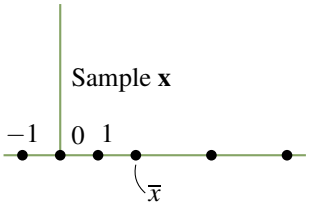
Suppose the heights $h_1, h_2, \ldots, h_n$ of n men are measured. Such a data set is called a **sample** of the heights of all the men in the population under study, and various questions are often asked about such a sample: What is the average height in the sample? How much variation is there in the sample heights, and how can it be measured? What can be inferred from the sample about the heights of all men in the population? How do these heights compare to heights of men in neighbouring countries? Does the prevalence of smoking affect the height of a man?

The analysis of samples, and of inferences that can be drawn from them, is a subject called *mathematical statistics*, and an extensive body of information has been developed to answer many such questions. In this section we will describe a few ways that linear algebra can be used.

It is convenient to represent a sample $\{x_1, x_2, \ldots, x_n\}$ as a **sample vector**[15] $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ in $\mathbb{R}^n$. This being done, the dot product in $\mathbb{R}^n$ provides a convenient tool to study the sample and describe some of the statistical concepts related to it. The most widely known statistic for describing a data set is the **sample mean** $\overline{x}$ defined by[16]

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean $\overline{x}$ is "typical" of the sample values x_i, but may not itself be one of them. The number $x_i - \overline{x}$ is called the **deviation** of x_i from the mean $\overline{x}$. The deviation is positive if $x_i > \overline{x}$ and it is negative if $x_i < \overline{x}$. Moreover, the sum of these deviations is zero:

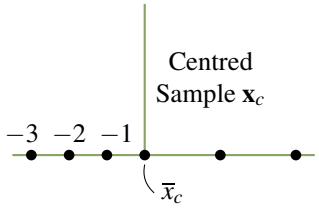


$$\sum_{i=1}^{n}(x_i - \overline{x}) = \left(\sum_{i=1}^{n} x_i\right) - n\overline{x} = n\overline{x} - n\overline{x} = 0 \tag{5.6}$$

This is described by saying that the sample mean $\overline{x}$ is *central* to the sample values x_i.

If the mean $\overline{x}$ is subtracted from each data value x_i, the resulting data $x_i - \overline{x}$ are said to be **centred**. The corresponding data vector is

$$\mathbf{x}_c = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{bmatrix}$$



and (5.6) shows that the mean $\overline{x}_c = 0$. For example, we have plotted the sample $\mathbf{x} = \begin{bmatrix} -1 & 0 & 1 & 4 & 6 \end{bmatrix}$ in the first diagram. The mean is $\overline{x} = 2$,

[15] We write vectors in $\mathbb{R}^n$ as row matrices, for convenience.

[16] The mean is often called the "average" of the sample values x_i, but statisticians use the term "mean".

and the centred sample $\mathbf{x}_c = \begin{bmatrix} -3 & -2 & -1 & 2 & 4 \end{bmatrix}$ is also plotted. Thus, the effect of centring is to shift the data by an amount $\overline{x}$ (to the left if $\overline{x}$ is positive) so that the mean moves to 0.

Another question that arises about samples is how much variability there is in the sample

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

that is, how widely are the data "spread out" around the sample mean $\overline{x}$. A natural measure of variability would be the sum of the deviations of the x_i about the mean, but this sum is zero by (5.6); these deviations cancel out. To avoid this cancellation, statisticians use the *squares* $(x_i - \overline{x})^2$ of the deviations as a measure of variability. More precisely, they compute a statistic called the **sample variance** s_x^2 defined[17] as follows:

$$s_x^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

The sample variance will be large if there are many x_i at a large distance from the mean $\overline{x}$, and it will be small if all the x_i are tightly clustered about the mean. The variance is clearly nonnegative (hence the notation s_x^2), and the square root s_x of the variance is called the **sample standard deviation**.

The sample mean and variance can be conveniently described using the dot product. Let

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$$

denote the row with every entry equal to 1. If $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$, then $\mathbf{x} \cdot \mathbf{1} = x_1 + x_2 + \cdots + x_n$, so the sample mean is given by the formula

$$\overline{x} = \frac{1}{n}(\mathbf{x} \cdot \mathbf{1})$$

Moreover, remembering that $\overline{x}$ is a scalar, we have $\overline{x}\mathbf{1} = \begin{bmatrix} \overline{x} & \overline{x} & \cdots & \overline{x} \end{bmatrix}$, so the centred sample vector $\mathbf{x}_c$ is given by

$$\mathbf{x}_c = \mathbf{x} - \overline{x}\mathbf{1} = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{bmatrix}$$

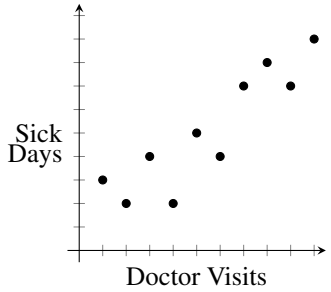
Thus we obtain a formula for the sample variance:

$$s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2 = \frac{1}{n-1}\|\mathbf{x} - \overline{x}\mathbf{1}\|^2$$

Linear algebra is also useful for comparing two different samples. To illustrate how, consider two examples.

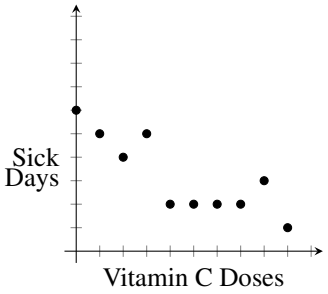
The following table represents the number of sick days at work per year and the yearly number of visits to a physician for 10 individuals.

Individual	1	2	3	4	5	6	7	8	9	10
Doctor visits	2	6	8	1	5	10	3	9	7	4
Sick days	2	4	8	3	5	9	4	7	7	2



The data are plotted in the **scatter diagram** where it is evident that, roughly speaking, the more visits to the doctor the more sick days. This is an example of a *positive correlation* between sick days and doctor visits.

[17] Since there are n sample values, it seems more natural to divide by n here, rather than by $n-1$. The reason for using $n-1$ is that then the sample variance s^2x provides a better estimate of the variance of the entire population from which the sample was drawn.



Now consider the following table representing the daily doses of vitamin C and the number of sick days.

Individual	1	2	3	4	5	6	7	8	9	10
Vitamin C	1	5	7	0	4	9	2	8	6	3
Sick days	5	2	2	6	2	1	4	3	2	5

The scatter diagram is plotted as shown and it appears that the more vitamin C taken, the fewer sick days. In this case there is a *negative correlation* between daily vitamin C and sick days.

In both these situations, we have **paired samples**, that is observations of two variables are made for ten individuals: doctor visits and sick days in the first case; daily vitamin C and sick days in the second case. The scatter diagrams point to a relationship between these variables, and there is a way to use the sample to compute a number, called the correlation coefficient, that measures the degree to which the variables are associated.

To motivate the definition of the correlation coefficient, suppose two paired samples $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$ are given and consider the centred samples

$$\mathbf{x}_c = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{bmatrix} \text{ and } \mathbf{y}_c = \begin{bmatrix} y_1 - \overline{y} & y_2 - \overline{y} & \cdots & y_n - \overline{y} \end{bmatrix}$$

If x_k is large among the x_i's, then the deviation $x_k - \overline{x}$ will be positive; and $x_k - \overline{x}$ will be negative if x_k is small among the x_i's. The situation is similar for $\mathbf{y}$, and the following table displays the sign of the quantity $(x_i - \overline{x})(y_k - \overline{y})$ in all four cases:

Sign of $(x_i - \overline{x})(y_k - \overline{y})$:

	x_i large	x_i small
y_i large	positive	negative
y_i small	negative	positive

Intuitively, if $\mathbf{x}$ and $\mathbf{y}$ are positively correlated, then two things happen:

1. *Large values of the x_i tend to be associated with large values of the y_i, and*
2. *Small values of the x_i tend to be associated with small values of the y_i.*

It follows from the table that, if $\mathbf{x}$ and $\mathbf{y}$ are positively correlated, then the dot product

$$\mathbf{x}_c \cdot \mathbf{y}_c = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

is positive. Similarly $\mathbf{x}_c \cdot \mathbf{y}_c$ is negative if $\mathbf{x}$ and $\mathbf{y}$ are negatively correlated. With this in mind, the **sample correlation coefficient**[18] r is defined by

$$r = r(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\| \, \|\mathbf{y}_c\|}$$

[18]The idea of using a single number to measure the degree of relationship between different variables was pioneered by Francis Galton (1822–1911). He was studying the degree to which characteristics of an offspring relate to those of its parents. The idea was refined by Karl Pearson (1857–1936) and r is often referred to as the Pearson correlation coefficient.

Bearing the situation in $\mathbb{R}^3$ in mind, r is the cosine of the "angle" between the vectors $\mathbf{x}_c$ and $\mathbf{y}_c$, and so we would expect it to lie between -1 and 1. Moreover, we would expect r to be near 1 (or -1) if these vectors were pointing in the same (opposite) direction, that is the "angle" is near zero (or π).

This is confirmed by Theorem 5.7.1 below, and it is also borne out in the examples above. If we compute the correlation between sick days and visits to the physician (in the first scatter diagram above) the result is $r = 0.90$ as expected. On the other hand, the correlation between daily vitamin C doses and sick days (second scatter diagram) is $r = -0.84$.

However, a word of caution is in order here. We *cannot* conclude from the second example that taking more vitamin C will reduce the number of sick days at work. The (negative) correlation may arise because of some third factor that is related to both variables. For example, case it may be that less healthy people are inclined to take more vitamin C. Correlation does *not* imply causation. Similarly, the correlation between sick days and visits to the doctor does not mean that having many sick days *causes* more visits to the doctor. A correlation between two variables may point to the existence of other underlying factors, but it does not necessarily mean that there is a causality relationship between the variables.

Our discussion of the dot product in $\mathbb{R}^n$ provides the basic properties of the correlation coefficient:

Theorem 5.7.1

Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ *and* $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$ *be (nonzero) paired samples, and let* $r = r(\mathbf{x}, \mathbf{y})$ *denote the correlation coefficient. Then:*

1. $-1 \leq r \leq 1$.
2. $r = 1$ *if and only if there exist* a *and* $b > 0$ *such that* $y_i = a + bx_i$ *for each* i.
3. $r = -1$ *if and only if there exist* a *and* $b < 0$ *such that* $y_i = a + bx_i$ *for each* i.

Proof. The Cauchy inequality (Theorem 5.3.2) proves (1), and also shows that $r = \pm 1$ if and only if one of $\mathbf{x}_c$ and $\mathbf{y}_c$ is a scalar multiple of the other. This in turn holds if and only if $\mathbf{y}_c = b\mathbf{x}_c$ for some $b \neq 0$, and it is easy to verify that $r = 1$ when $b > 0$ and $r = -1$ when $b < 0$.

Finally, $\mathbf{y}_c = b\mathbf{x}_c$ means $y_i - \overline{y} = b(x_i - \overline{x})$ for each i; that is, $y_i = a + bx_i$ where $a = \overline{y} - b\overline{x}$. Conversely, if $y_i = a + bx_i$, then $\overline{y} = a + b\overline{x}$ (verify), so $y_1 - \overline{y} = (a + bx_i) - (a + b\overline{x}) = b(x_1 - \overline{x})$ for each i. In other words, $\mathbf{y}_c = b\mathbf{x}_c$. This completes the proof. □

Properties (2) and (3) in Theorem 5.7.1 show that $r(\mathbf{x}, \mathbf{y}) = 1$ means that there is a linear relation with *positive* slope between the paired data (so large x values are paired with large y values). Similarly, $r(\mathbf{x}, \mathbf{y}) = -1$ means that there is a linear relation with *negative* slope between the paired data (so small x values are paired with small y values). This is borne out in the two scatter diagrams above.

We conclude by using the dot product to derive some useful formulas for computing variances and correlation coefficients. Given samples $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$, the key observation is the following formula:

$$\mathbf{x}_c \cdot \mathbf{y}_c = \mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y} \tag{5.7}$$

Indeed, remembering that $\overline{x}$ and $\overline{y}$ are scalars:

$$\begin{aligned}\mathbf{x}_c \cdot \mathbf{y}_c &= (\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1}) \\ &= \mathbf{x} \cdot \mathbf{y} - \mathbf{x} \cdot (\overline{y}\mathbf{1}) - (\overline{x}\mathbf{1}) \cdot \mathbf{y} + (\overline{x}\mathbf{1})(\overline{y}\mathbf{1}) \\ &= \mathbf{x} \cdot \mathbf{y} - \overline{y}(\mathbf{x} \cdot \mathbf{1}) - \overline{x}(\mathbf{1} \cdot \mathbf{y}) + \overline{xy}(\mathbf{1} \cdot \mathbf{1}) \\ &= \mathbf{x} \cdot \mathbf{y} - \overline{y}(n\overline{x}) - \overline{x}(n\overline{y}) + \overline{x}\,\overline{y}(n) \\ &= \mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y}\end{aligned}$$

Taking $\mathbf{y} = \mathbf{x}$ in (5.7) gives a formula for the variance $s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2$ of $\mathbf{x}$.

Variance Formula

If x is a sample vector, then $s_x^2 = \frac{1}{n-1}\left(\|\mathbf{x}_c\|^2 - n\overline{x}^2\right)$.

We also get a convenient formula for the correlation coefficient, $r = r(\mathbf{x},\ \mathbf{y}) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\|\ \|\mathbf{y}_c\|}$. Moreover, (5.7) and the fact that $s_x^2 = \frac{1}{n-1}\|\mathbf{x}_c\|^2$ give:

Correlation Formula

If $\mathbf{x}$ and $\mathbf{y}$ are sample vectors, then

$$r = r(\mathbf{x},\ \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y} - n\overline{x}\,\overline{y}}{(n-1)s_x s_y}$$

Finally, we give a method that simplifies the computations of variances and correlations.

Data Scaling

Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$ be sample vectors. Given constants a, b, c, and d, consider new samples $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$ where $z_i = a + bx_i$, for each i and $w_i = c + dy_i$ for each i. Then:

a. $\overline{z} = a + b\overline{x}$

b. $s_z^2 = b^2 s_x^2$, so $s_z = |b|s_x$

c. If b and d have the same sign, then $r(\mathbf{x},\ \mathbf{y}) = r(\mathbf{z},\ \mathbf{w})$.

The verification is left as an exercise. For example, if $\mathbf{x} = \begin{bmatrix} 101 & 98 & 103 & 99 & 100 & 97 \end{bmatrix}$, subtracting 100 yields $\mathbf{z} = \begin{bmatrix} 1 & -2 & 3 & -1 & 0 & -3 \end{bmatrix}$. A routine calculation shows that $\overline{z} = -\frac{1}{3}$ and $s_z^2 = \frac{14}{3}$, so $\overline{x} = 100 - \frac{1}{3} = 99.67$, and $s_z^2 = \frac{14}{3} = 4.67$.

Exercises for 5.7

Exercise 5.7.1 The following table gives IQ scores for 10 fathers and their eldest sons. Calculate the means, the variances, and the correlation coefficient r. (The data scaling formula is useful.)

	1	2	3	4	5	6	7	8	9	10
Father's IQ	140	131	120	115	110	106	100	95	91	86
Son's IQ	130	138	110	99	109	120	105	99	100	94

Exercise 5.7.2 The following table gives the number of years of education and the annual income (in thousands) of 10 individuals. Find the means, the variances, and the correlation coefficient. (Again the data scaling formula is useful.)

Individual	1	2	3	4	5	6	7	8	9	10
Years of education	12	16	13	18	19	12	18	19	12	14
Yearly income (1000's)	31	48	35	28	55	40	39	60	32	35

Exercise 5.7.3 If $\mathbf{x}$ is a sample vector, and $\mathbf{x}_c$ is the centred sample, show that $\overline{x}_c = 0$ and the standard deviation of $\mathbf{x}_c$ is s_x.

Exercise 5.7.4 Prove the data scaling formulas found on page 326: (a), (b), and (c).

Supplementary Exercises for Chapter 5

Exercise 5.1 In each case either show that the statement is true or give an example showing that it is false. Throughout, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ denote vectors in $\mathbb{R}^n$.

a. If U is a subspace of $\mathbb{R}^n$ and $\mathbf{x}+\mathbf{y}$ is in U, then $\mathbf{x}$ and $\mathbf{y}$ are both in U.

b. If U is a subspace of $\mathbb{R}^n$ and $r\mathbf{x}$ is in U, then $\mathbf{x}$ is in U.

c. If U is a nonempty set and $s\mathbf{x}+t\mathbf{y}$ is in U for any s and t whenever $\mathbf{x}$ and $\mathbf{y}$ are in U, then U is a subspace.

d. If U is a subspace of $\mathbb{R}^n$ and $\mathbf{x}$ is in U, then $-\mathbf{x}$ is in U.

e. If $\{\mathbf{x}, \mathbf{y}\}$ is independent, then $\{\mathbf{x}, \mathbf{y}, \mathbf{x}+\mathbf{y}\}$ is independent.

f. If $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is independent, then $\{\mathbf{x}, \mathbf{y}\}$ is independent.

g. If $\{\mathbf{x}, \mathbf{y}\}$ is not independent, then $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is not independent.

h. If all of $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are nonzero, then $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is independent.

i. If one of $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ is zero, then $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is not independent.

j. If $a\mathbf{x}+b\mathbf{y}+c\mathbf{z}=\mathbf{0}$ where a, b, and c are in $\mathbb{R}$, then $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is independent.

k. If $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is independent, then $a\mathbf{x}+b\mathbf{y}+c\mathbf{z}=\mathbf{0}$ for some a, b, and c in $\mathbb{R}$.

l. If $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is not independent, then $t_1\mathbf{x}_1+t_2\mathbf{x}_2+\cdots+t_n\mathbf{x}_n=\mathbf{0}$ for t_i in $\mathbb{R}$ not all zero.

m. If $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is independent, then $t_1\mathbf{x}_1+t_2\mathbf{x}_2+\cdots+t_n\mathbf{x}_n=\mathbf{0}$ for some t_i in $\mathbb{R}$.

n. Every set of four non-zero vectors in $\mathbb{R}^4$ is a basis.