2

Basics

2.1 Data Types 🅰☺

Environmental scientists are familiar with numerical data, especially with continuous numerical data – for example temperature, pressure, pollution concentration, specific humidity, streamflow and sea level height. Numerical data can also be discrete, being recorded in integers instead of real numbers. Figure 2.1 shows there are other types of data, namely categorical data. *Categorical* variables represent data which can be placed into groups or categories. Categorical data can, in turn, be either nominal or ordinal. *Nominal* data do not have order, for example true/false, colour (red, green, blue), country of birth (USA, China, Russia, Liechtenstein, etc.), animals (cats, dogs, elephants, etc.). *Ordinal* data have categories with some natural order, such as weather type (sunny, cloudy, rainy), education level (elementary school graduate, high school graduate, some college and college graduate) and the Likert scale used in customer surveys (strongly disagree, disagree, neutral, agree, strongly agree). Unlike in the environmental sciences, where data are predominantly continuous, data in commercial or computer science application areas tend to be predominantly categorical and/or discrete. This has an important bearing as data methods developed for commercial or computer science applications tend to be predominantly designed for categorical and/or discrete data. Some were later adapted to work with continuous data.

2.2 Probability 🅰☺

Environmental data contain fluctuations – for example, the atmosphere has fluctuations ranging from large-scale weather systems to small-scale turbulence. Thus, an understanding of random variables and probability theory is essential for analysing environmental data.

We start with a simple example for illustrating the basic concepts of probability. Suppose one has 100 days of weather observations, where there two

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

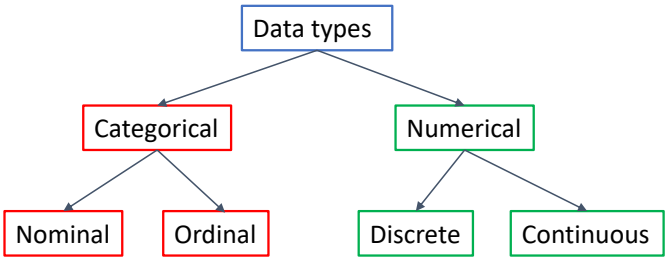

Figure 2.1 Main types of data.

variables in the daily weather, namely temperature and precipitation. For simplicity, the temperature variable has three classes – cold (c), normal (n) and hot (h) – while the precipitation has two classes – dry (d) and wet (w). Table 2.1 shows the distribution of the weather data – for example, out of 100 days, there are 15 days of cold dry weather, 5 days of cold wet weather, and so on. The bottom row gives the sum over the numbers in the column above, for example the total number of cold days is $15 + 5 = 20$. The rightmost column gives the sum over the row, for example the total number of dry days is $15 + 35 + 10 = 60$.

Table 2.1 Distribution of 100 days of weather observations, with the corresponding probability distribution $P(x, y)$ listed in the table to the right. $P(x)$ and $P(y)$ are the marginal distributions.

	cold	norm	hot	sum
dry	15	35	10	60
wet	5	15	20	40
sum	20	50	30	100

$P(x, y)$	$x = \mathrm{c}$	$x = \mathrm{n}$	$x = \mathrm{h}$	$P(y)$
$y = \mathrm{d}$	0.15	0.35	0.10	0.60
$y = \mathrm{w}$	0.05	0.15	0.20	0.40
$P(x)$	0.20	0.50	0.30	1

Next, we want to obtain the probability distribution $P(x, y)$, where x and y are the temperature and precipitation variables, respectively. $P(c, d)$, the probability of cold dry weather, is simply the number of observations with cold dry weather divided by N, the total number of observations, that is, $P(c, d) = 15/100 = 0.15$. The probability table is shown on the right side of Table 2.1. Strictly speaking, probability is defined only in the limit as $N \to \infty$, so our finite N only allows us to get an estimate of the true probability. $P(x, y)$ is called a *joint probability* or *joint distribution* as it depends on both x and y. The bottom row $P(x)$ and the rightmost column $P(y)$ are called *marginal distributions*, as they appear on the margins of probability tables. They are obtained by summing over $P(x, y)$,

$$P(x) = \sum_y P(x, y), \qquad P(y) = \sum_x P(x, y), \tag{2.1}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

as one can check by summing over the rows and columns of $P(x, y)$ in Table 2.1. From $P(y)$, the probability of dry days, $P(d)$, is 0.60, while the probability of wet days is 0.40. Note that

$$\sum_y P(y) = 0.60 + 0.40 = 1, \quad \text{and} \quad \sum_x P(x) = 0.20 + 0.50 + 0.30 = 1, \quad (2.2)$$

as the sum of the probabilities over all the events must equal one.

$P(x|y)$, the *conditional probability* of x given y, is the probability of observing x when the value y is already known. For instance, if $x = c$ and $y = d$, $P(c|d)$ is the probability of getting cold temperature under dry conditions. Since the joint probability of getting x and y, that is, $P(x, y)$, equals the probability of getting y, that is, $P(y)$, multiplied by the conditional probability of getting x given y, that is, $P(x|y)$, we can write

$$P(x, y) = P(x|y)P(y). \quad (2.3)$$

Thus,

$$P(x|y) = \frac{P(x, y)}{P(y)}, \quad \text{if } P(y) > 0, \quad \text{otherwise 0.} \quad (2.4)$$

Using the values from Table 2.1, $P(c|d)$, the conditional probability of having cold conditions given it is dry, is $P(c, d)/P(d) = 0.15/0.60 = 0.25$.

Similarly,

$$P(x, y) = P(y|x)P(x), \quad (2.5)$$

where $P(y|x)$ is the conditional probability of y given x. Thus,

$$P(y|x) = \frac{P(x, y)}{P(x)}, \quad \text{if } P(x) > 0, \quad \text{otherwise 0.} \quad (2.6)$$

Using Table 2.1, $P(d|c)$, the conditional probability of having dry conditions given it is cold, is $P(c, d)/P(c) = 0.15/0.20 = 0.75$.

Combining (2.3) and (2.5) gives

$$P(y|x)P(x) = P(x|y)P(y). \quad (2.7)$$

Thus,

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_y P(x, y)} = \frac{P(x|y)P(y)}{\sum_y P(x|y)P(y)}, \quad (2.8)$$

upon invoking (2.1) and (2.3). This equation is called *Bayes theorem* or Bayes rule, having originated from the work of Thomas Bayes (1702–1761), an English mathematician and Presbyterian minister. For more details on Bayes theorem, see Section 2.14.

If the probability of getting x is not affected at all by the given value of y, we have

$$P(x|y) = P(x). \quad (2.9)$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

Then, (2.3) simplifies to

$$P(x, y) = P(x)P(y), \tag{2.10}$$

and x and y are said to be *independent* events. If one computes $P(x)P(y)$ from Table 2.1, one finds the product not equal to $P(x, y)$, so x and y are not independent events in that dataset.

Keeping the same $P(x)$ and $P(y)$ from Table 2.1, one can check that Table 2.2 indeed satisfies (2.10), so x and y are independent. Thus, in this example, the probability of getting dry or wet weather is unaffected by whether the temperature is cold, normal or hot, and similarly, the probability of getting cold, normal or hot weather is unaffected by whether it is dry or wet.

Table 2.2 Probability distribution $P(x, y)$, with x and y being independent.

$P(x,y)$	$x=\mathrm{c}$	$x=\mathrm{n}$	$x=\mathrm{h}$	$P(y)$
$y=\mathrm{d}$	0.12	0.30	0.18	0.60
$y=\mathrm{w}$	0.08	0.20	0.12	0.40
$P(x)$	0.20	0.50	0.30	1

2.3 Probability Density A☺

Thus far, only probabilities of discrete events have been considered. Next, we extend the concept of probability to continuous variables. Suppose the probability of a real variable x lying within the interval $(x, x + \delta x)$ is denoted by $p(x)\delta x$ for $\delta x \to 0$ (Fig. 2.2(a)), then $p(x)$ is called the *probability density* or

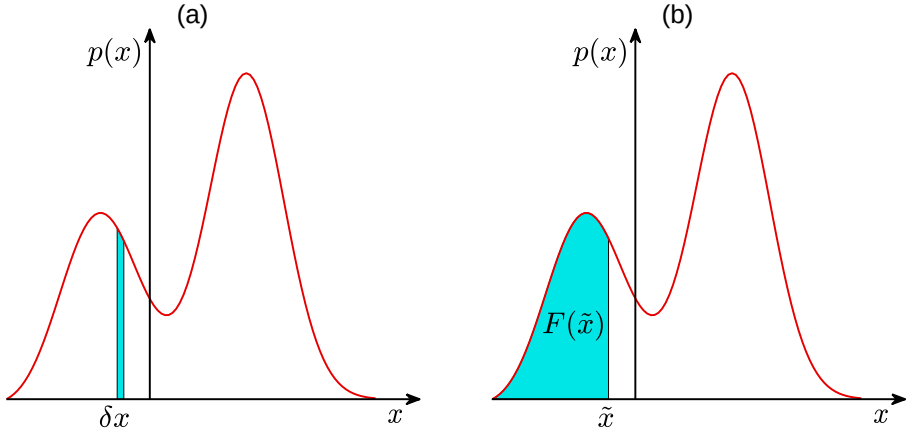

Figure 2.2 (a) The probability of x lying within the interval $(x, x + \delta x)$ is given by the area of the narrow vertical band of height $p(x)$ and width δx. The two peaks in $p(x)$ indicate the two regions of higher probability. (b) The cumulative distribution $F(\tilde{x})$ is given by the shaded area under the curve.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

probability density function (PDF) over x. The probability of x lying within the interval (a, b) is obtained by integrating the PDF:[1]

$$P(x \in (a, b)) = \int_a^b p(x)dx. \tag{2.11}$$

As probabilities cannot be negative, it follows that

$$p(x) \geq 0. \tag{2.12}$$

The requirement that the sum of probabilities over all discrete events equals one is replaced in the continuous case by

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{2.13}$$

Note that $p(x)$ is not prohibited from exceeding 1.

The *cumulative distribution function* (CDF) $F(\tilde{x})$ is defined to be

$$F(\tilde{x}) = P(x \leq \tilde{x}) = \int_{-\infty}^{\tilde{x}} p(x)dx. \tag{2.14}$$

In Fig. 2.2(b), $F(\tilde{x})$ is seen as the area under the curve $p(x)$, stretching over the interval $-\infty < x \leq \tilde{x}$. It follows from taking the derivative of F that

$$p(x) = \frac{dF(x)}{dx}. \tag{2.15}$$

From (2.14) and (2.11), we have

$$F(b) - F(a) = \int_a^b p(x)dx = P(x \in (a, b)). \tag{2.16}$$

The *complementary cumulative distribution function* (CCDF) or simply the *tail distribution* is

$$\tilde{F}(\tilde{x}) \equiv 1 - F(\tilde{x}) = P(\tilde{x} < x) = \int_{\tilde{x}}^{\infty} p(x)dx. \tag{2.17}$$

If there are two continuous variables x and y, the joint probability density distribution is $p(x, y)$. The *marginal probability density* distributions are defined similar to the marginal probability distributions for discrete variables in (2.1), but with integration replacing summation, that is

$$p(x) = \int p(x, y)\, \mathrm{d}y, \qquad p(y) = \int p(x, y)\, \mathrm{d}x, \tag{2.18}$$

where the integrations are over the domains of y and x, respectively.

Similar to (2.6) for discrete variables, the *conditional probability density* distribution can be defined by

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{2.19}$$

for $p(y) > 0$, and analogously for $p(y|x)$.

[1] In this book, we try to follow the convention of using the capital letter P to denote a probability and the small letter p to denote a probability density.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.4 Expectation and Mean A☺

Let x be a random variable that takes on discrete values. For example, x can be the outcome of a die cast, where the possible values are $x_i = i$, with $i = 1, \ldots, 6$. The *expectation* or expected value of x from a population is given by

$$\mathrm{E}[x] = \sum_i x_i P_i, \tag{2.20}$$

where P_i is the probability of x_i occurring. If the die is fair, $P_i = 1/6$ for all i, so $\mathrm{E}[x]$ is 3.5. We also write

$$\mathrm{E}[x] = \mu_x, \tag{2.21}$$

with μ_x denoting the *mean* of x for the population, that is, the *population mean*.

The expectation of a sum of random variables satisfies

$$\mathrm{E}[ax + by + c] = a\,\mathrm{E}[x] + b\,\mathrm{E}[y] + c, \tag{2.22}$$

where x and y are random variables, and a, b and c are constants.

For a random variable x that takes on continuous values over a domain Ω, the expectation is given by an integral,

$$\mathrm{E}[x] = \int_\Omega x\, p(x)\, \mathrm{d}x, \tag{2.23}$$

where $p(x)$ is the PDF. For any function $f(x)$, the expectation is

$$\begin{aligned} \mathrm{E}[f(x)] &= \int_\Omega f(x)p(x)\, \mathrm{d}x \quad \text{(continuous case)}, \\ \mathrm{E}[f(x)] &= \sum_i f(x_i)P_i \quad \text{(discrete case)}. \end{aligned} \tag{2.24}$$

In real world problems, one normally cannot compute the mean by using the formula for the population mean, that is, (2.20) or (2.23), because one does not know P_i or $p(x)$. One can sample only N measurements of x $(x_1, \ldots, x_N)$ from the population. The *sample mean* $\overline{x}$ or $\langle x \rangle$ is calculated by

$$\overline{x} \equiv \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{2.25}$$

that is, simply taking the average of the N measurements, which is in general different from the population mean μ_x. However, one can show that the expectation of the sample mean equals the population mean. Thus, as the sample size increases, the sample mean approaches the population mean. In general, the sample mean should be regarded as a statistical estimator of the true population mean.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.5 Variance and Standard Deviation A☺

The fluctuations around the mean value are commonly characterized by the variance of the population (i.e. the *population variance*),

$$\begin{aligned} \mathrm{var}(x) &\equiv \mathrm{E}[(x-\mu_x)^2] = \mathrm{E}[x^2 - 2x\mu_x + \mu_x^2] \\ &= \mathrm{E}[x^2] - 2\mu_x \mathrm{E}[x] + \mu_x^2 = \mathrm{E}[x^2] - \mu_x^2, \end{aligned} \tag{2.26}$$

where (2.22) and (2.21) have been invoked. The *population standard deviation* σ is the positive square root of the population variance, that is,

$$\sigma^2 = \mathrm{var}(x). \tag{2.27}$$

The *sample standard deviation* s is the positive square root of the *sample variance*, given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{2.28}$$

The denominator of $N-1$ (instead of N) is a bias correction introduced by Friedrich Bessel (1784–1846) to ensure the expectation of the sample variance equals the population variance. As the sample size increases, the sample variance approaches the population variance. For large N, there is negligible difference in the result when using $N-1$ or N in the denominator of (2.28), so either form can be used. The sample variance is a statistical estimator of the population variance.

Note that the standard deviation has the same unit as the variable x. For instance, if x is wind speed measured in $\mathrm{m\,s^{-1}}$, then its mean and standard deviation will also have units of $\mathrm{m\,s^{-1}}$, while the variance will have units of $\mathrm{m^2 s^{-2}}$. The mean is a measure of the location of the data, while the standard deviation is a measure of the scale or spread of the data.

Often one would like to compare two very different variables, for example sea surface temperature and fish catch, which have different units and very likely different magnitudes. To avoid 'comparing apples with oranges', one usually standardizes the variables. The *standardized variable*

$$z = (x - \overline{x})/s \tag{2.29}$$

is obtained from the original variable by subtracting the sample mean and dividing by the sample standard deviation.[2] The standardized variable is also called the standard score, z-score, z-value, normal score, normalized variable or standardized anomaly (where *anomaly* means deviation from the mean value). The advantage of standardizing variables is that now the sea surface temperature and fish catch standardized variables will both have no units and have sample means of zero and sample standard deviations of one.

[2] In situations where the population mean μ_x or standard deviation σ are known, they are used instead of the sample mean $\overline{x}$ and standard deviation s in (2.29).

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.6 Covariance A☺

For two random variables x and y, with mean μ_x and μ_y, respectively, their *covariance* is defined by

$$\text{cov}(x, y) = \text{E}[(x - \mu_x)(y - \mu_y)]. \tag{2.30}$$

For brevity, we will use 'covariance' instead of 'population covariance' when there is no ambiguity. Note that cov is symmetric with respect to its two arguments, that is, $\text{cov}(x, y) = \text{cov}(y, x)$. The variance in (2.26) is simply a special case of the covariance, with

$$\text{var}(x) = \text{cov}(x, x). \tag{2.31}$$

Covariance is a measure of the joint variability of x and y. If high values of x occur together with high values of y, then $(x - \mu_x)(y - \mu_y)$ will be positive; similarly, if low values of x occur together with low values of y, then $(x-\mu_x)(y-\mu_y)$ will involve multiplying two negative numbers and so will also be positive – leading to a positive covariance. If high values of x occur together with low values of y, and vice versa – the covariance will be negative. Thus, a positive covariance indicates a tendency of similar behaviour between x and y, whereas a negative covariance indicates opposite behaviour. For instance, if the covariance between temperature and snow amount is negative, then high temperature tends to occur with low snow and low temperature with high snow.

One can show that if x and y are independent, then their covariance is zero. However, the converse is not true in general – for example, if x is uniformly distributed in $[-1,\ 1]$ and $y = x^2$, one can show that $\text{cov}(x, y)$ is zero, even as y depends on x non-linearly. Thus, covariance only measures the linear joint variability between two variables.

The sample covariance computed from the data by

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) \tag{2.32}$$

approaches the population covariance as $N \to \infty$.

The magnitude of the covariance is not too useful since it is not normalized and, therefore, depends on the magnitudes of the variables. For instance, if x and y are measured in units of centimetres instead of metres, the covariance computed using measurements in cm will be 10^4 times that using metres. The normalized version of the covariance, the correlation coefficient (Section 2.11), is widely used as its magnitude reveals the strength of the linear relation.

If instead of just two variables x and y, we have d variables, that is, $\mathbf{x} = x_1, ..., x_d$, then the covariance coefficient generalizes to the *covariance matrix*:

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

$$\begin{aligned}\mathrm{cov}(\mathbf{x}) &= \mathrm{E}\left[(\mathbf{x}-\mathrm{E}[\mathbf{x}])(\mathbf{x}-\mathrm{E}[\mathbf{x}])^{\mathrm{T}}\right] \\ &= \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_d) \\ \mathrm{cov}(x_2,x_1) & \mathrm{var}(x_2) & \cdots & \mathrm{cov}(x_2,x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_d,x_1) & \mathrm{cov}(x_d,x_2) & \cdots & \mathrm{var}(x_d) \end{bmatrix},\end{aligned} \tag{2.33}$$

where the superscript T denotes the transpose of a vector or matrix.

Another way to generalize (2.32) is by letting

$$\begin{aligned}\mathrm{cov}(\mathbf{x},\mathbf{y}) &= \mathrm{E}\left[(\mathbf{x}-\mathrm{E}[\mathbf{x}])(\mathbf{y}-\mathrm{E}[\mathbf{y}])^{\mathrm{T}}\right] \\ &= \begin{bmatrix} \mathrm{cov}(x_1,y_1) & \mathrm{cov}(x_1,y_2) & \cdots & \mathrm{cov}(x_1,y_d) \\ \mathrm{cov}(x_2,y_1) & \mathrm{cov}(x_2,y_2) & \cdots & \mathrm{cov}(x_2,y_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_d,y_1) & \mathrm{cov}(x_d,y_2) & \cdots & \mathrm{cov}(x_d,y_d) \end{bmatrix}.\end{aligned} \tag{2.34}$$

Clearly, $\mathrm{cov}(\mathbf{x})$ in (2.33) is equivalent to the special case $\mathrm{cov}(\mathbf{x},\mathbf{x})$ in (2.34). There is no standard nomenclature for $\mathrm{cov}(\mathbf{x})$ and $\mathrm{cov}(\mathbf{x},\mathbf{y})$, as both are referred to as covariance matrices. Some authors call $\mathrm{cov}(\mathbf{x})$ the variance matrix or the variance-covariance matrix and $\mathrm{cov}(\mathbf{x},\mathbf{y})$ the covariance matrix, while others call $\mathrm{cov}(\mathbf{x})$ the covariance matrix and $\mathrm{cov}(\mathbf{x},\mathbf{y})$ the cross-covariance matrix. In this book, we will use 'covariance matrix' to denote either $\mathrm{cov}(\mathbf{x})$ or $\mathrm{cov}(\mathbf{x},\mathbf{y})$.

2.7 Online Algorithms for Mean, Variance and Covariance C☺

In recent decades, there has been increasing interest in *online learning* problems where data become available in a sequential order and the models are to be updated with the continually arriving new data. The traditional *batch learning* approach, which trains the model with the complete training dataset, is very inefficient in the online learning situation – to update the model with one additional data point, the model has to be retrained with the complete dataset containing N points. In contrast, an online learning algorithm would update the model with only the single new data point – and all previous data points can be erased from the computer memory. Obviously, when one has to update a model frequently with newly arrived data, an online learning algorithm would have a huge advantage over a batch learning algorithm in terms of computer time and memory.

First, consider the sample mean. The batch algorithm is given by (2.25), where one has to use all N data points for the computation. To develop an online algorithm, we first define the sample mean computed with N data points to be

$$\overline{x}_N \equiv \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2.35}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

which can be rewritten as

$$N\,\overline{x}_N = \sum_{i=1}^{N} x_i \;= \sum_{i=1}^{N-1} x_i + x_N \tag{2.36}$$

$$= (N-1)\,\overline{x}_{N-1} + x_N = \; N\,\overline{x}_{N-1} + x_N - \overline{x}_{N-1}. \tag{2.37}$$

Thus, the online algorithm for the sample mean is given by

$$\overline{x}_N = \overline{x}_{N-1} + \frac{x_N - \overline{x}_{N-1}}{N}. \tag{2.38}$$

This means that if one has $\overline{x}_{N-1}$, the sample mean for the first $N-1$ data points, and a new data point x_N, then the updated sample mean for the N data points can be obtained simply from $\overline{x}_{N-1}$ and x_N. The earlier data points $x_1, \ldots, x_{N-1}$ are not needed in this update of the sample mean. The ability to delete old data can be very helpful as datasets can grow to enormous size as time passes.

Let us count the number of operations in the two approaches. In the batch algorithm (2.35), there are $N-1$ additions followed by one division. In the online algorithm (2.38), there are one subtraction, one division and one addition. When N becomes large, the batch algorithm becomes much slower than the online algorithm.

For online updating of the sample variance, the Welford algorithm (Welford, 1962; Knuth, 1998, vol. 2, p. 232) involves updating the mean with (2.38) and updating the sum of squared differences

$$S_N \equiv \sum_{i=1}^{N} (x_i - \overline{x}_N)^2, \tag{2.39}$$

by

$$S_N = S_{N-1} + (x_N - \overline{x}_{N-1})(x_N - \overline{x}_N), \qquad N \geq 2, \tag{2.40}$$

with the sample variance

$$s_N^2 = \frac{S_N}{N-1}\,. \tag{2.41}$$

Similarly, for an online algorithm to compute the sample covariance, define

$$C_N \equiv \sum_{i=1}^{N} (x_i - \overline{x}_N)(y_i - \overline{y}_N). \tag{2.42}$$

With

$$\overline{y}_N = \overline{y}_{N-1} + \frac{y_N - \overline{y}_{N-1}}{N}, \tag{2.43}$$

one can show that

$$C_N = C_{N-1} + (x_N - \overline{x}_{N-1})(y_N - \overline{y}_N) \tag{2.44}$$

$$= C_{N-1} + \frac{N-1}{N}(x_N - \overline{x}_{N-1})(y_N - \overline{y}_{N-1}), \tag{2.45}$$

with the sample covariance being $C_N/(N-1)$.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.8 Median and Median Absolute Deviation A ☺

In the last few decades, there has been increasing usage of *robust statistics* to alleviate weaknesses in traditional statistical estimators (Wilcox, 2004). Traditional statistical methods commonly make assumptions (e.g. the random variables obey a Gaussian distribution) that may not be valid for some datasets, leading to poor statistical estimates. Statistical methods that perform poorly when the underlying assumptions are not satisfied are called *non-robust*. Robust methods are designed to work well with a broad range of datasets.

Another weakness is referred to as being *non-resistant* to outliers in the data – an *outlier* being an extreme data value arising from a measurement or recording error, or from an abnormal event. For instance, someone entering data by hand may misread '.100' as '100', and ends up entering a number a thousand times larger than the actual value. Non-resistant methods yield poor estimates when given even a small number of outliers. Resistant methods are designed to work well even when the datasets contain outliers.

It is desirable to have methods that are *robust* and *resistant*. Some authors, such as Wilks (2011), make a distinction between robustness and resistance. However, since most methods that are robust are also resistant, and vice versa, many authors do not make a distinction between robustness and resistance and simply refer to all such methods as robust methods.

While the mean and standard deviation are the most common estimators of location and scale (or spread) of the data, they are not resistant to outliers. Suppose student A made seven repeated measurements in a laboratory experiment, recording the values (arranged in ascending order): 1.0, 1.2, 1.2, 1.3, 1.5, 1.7 and 1.8. His lab partner, student B, also recorded the same measurements but mistakenly typed in '18.' instead of '1.8' for the final data point. The mean computed by A was 1.386, but was 3.700 by B. The computed standard deviation was 0.291 by A and 6.310 by B. Clearly, the mean and standard deviation are non-resistant to outliers.

A robust/resistant alternative to the mean is the *median*, defined as the middle value of a population or a sample of measurements sorted in ascending order. In the above example of seven measurements, the middle is the fourth measurement, namely 1.3, as there are three measurements above and three below. What happens if there is an even number of data points? Suppose we drop the seventh data point and are left with six measurements. Then the third and fourth are the two middle points, and we take the average of these two values, that is, $(1.2 + 1.3)/2 = 1.25$, as the median. Thus, the median is defined to be the middle value if N, the number of data points, is odd, and to be the average of the two middle values if N is even.

Let us return to the example with the students each recording seven measurements. Student A's mean was 1.386 and his median was 1.3, while student B's mean was 3.700 and his median was 1.3. Thus, with a completely erroneous seventh data point, student B managed to obtain the same median value as student A.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

The *breakdown point* of a statistical estimator is the proportion of incorrect data points, for example data points with arbitrarily high or low values, the estimator can handle before giving a completely incorrect result. For the mean, the breakdown point is 0, since the mean cannot handle even one single incorrect data point. In contrast, the median has a breakdown point of 50%. For instance, if the above example has values recorded as 1.0, 1.2, 1.2, 1.3, 999, 999 and 999, the median will still be 1.3.

A robust and resistant substitute for the standard deviation is the *median absolute deviation* (MAD), defined by

$$\text{MAD} = \text{median}(|x - \text{median}(x)|), \tag{2.46}$$

with a breakdown point of 50%. The deviations, $x - \text{median}(x)$, are around the median instead of the mean, and computing the absolute value avoids squaring the deviations in the standard deviation formula (2.28), which amplifies the large deviations.

For student A, with median = 1.3, his deviations $x - \text{median}(x) = -0.3$, -0.1, -0.1, 0.0, 0.2, 0.4 and 0.5, and the absolute deviations sorted in ascending order are 0.0, 0.1, 0.1, 0.2, 0.3, 0.4 and 0.5, with MAD = 0.2. For student B, the absolute deviations arranged in ascending order are 0.0, 0.1, 0.1, 0.2, 0.3, 0.4 and 16.7, again with MAD = 0.2.

Unlike the mean and standard deviation, there are no simple online learning algorithms for the median and MAD, that is, algorithms where one can erase the old data as new data arrive to update the estimator.

In this book, MAD stands for median absolute deviation around the median. Other estimators with the same acronym MAD can be defined, for example, mean absolute deviation around the mean, mean absolute deviation around the median or even median absolute deviation around the mean (though this final one is not commonly used).

2.9 Quantiles A☺

Often one is interested in finding a value x_α where $P(x \leq x_\alpha) = \alpha$ for a given value of α, with $0 \leq \alpha \leq 1$. For instance, one may want to know the value $x_{0.95}$ where 95% of the distribution lies below the value $x_{0.95}$ – that is, finding the 95th percentile. As the cumulative probability distribution function $F(x_\alpha) \equiv P(x \leq x_\alpha)$ from (2.14) is a monotonically increasing function, it has an inverse function $F^{-1}(\alpha)$. $F^{-1}(\alpha)$ is the value of x_α where $F(x_\alpha) = \alpha$. We call $q_\alpha \equiv x_\alpha$ the α *quantile* of F.

Figure 2.3 illustrates how to obtain a quantile value from a cumulative distribution function $F(x)$. Along the ordinate axis, we locate the value 0.95. To find $F^{-1}(0.95)$, we simply look at the intersection between the cumulative distribution curve and the horizontal line with ordinate = 0.95, leading to the value of the quantile $q_{0.95}$ along the abscissa. $F^{-1}(0.5) = q_{0.5}$ is simply the median, with 50% of the x values distributed above and 50% below the median.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

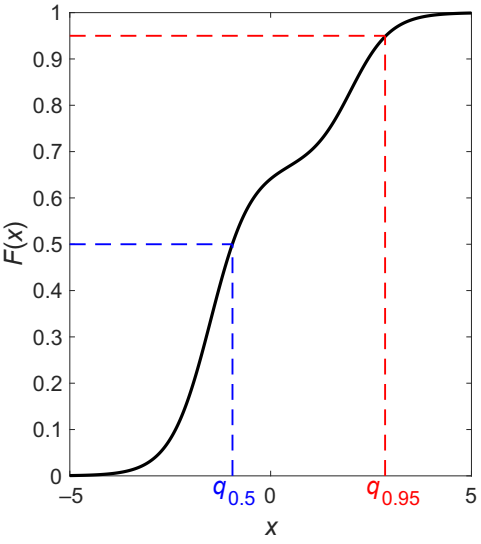

Figure 2.3 A cumulative distribution function $F(x)$. By inverse mapping from the ordinate to the abscissa, one can obtain the quantiles q_α. The 95th percentile $q_{0.95}$ and the median $q_{0.5}$ are shown.

The median splits up the cumulative distribution into two equal halves. Other common ways to split up the cumulative distribution into quantiles include *terciles*, with $q_{0.333}$ and $q_{0.667}$ splitting the distribution into three equal parts, and *quartiles*, with $q_{0.25}$, $q_{0.5}$ and $q_{0.75}$ splitting the distribution into four equal parts. The *interquartile range* (IQR), defined by the separation between the third quartile and the first quartile,

$$\text{IQR} = q_{0.75} - q_{0.25}, \tag{2.47}$$

is often used to characterize the spread or scale of the data, as it measures the spread of the middle 50% of the data. Since it ignores the top and bottom 25% of the data, it is resistant to outliers.

For a five-part split, *quintiles* use $q_{0.2}$, $q_{0.4}$, $q_{0.6}$ and $q_{0.8}$. For a 10-part split, *deciles* use $q_{0.1}$, $q_{0.2}$, ..., $q_{0.9}$. For a 100-part split, *percentiles* use $q_{0.01}$, $q_{0.02}$, ..., $q_{0.99}$.

Next, we examine how quantiles can be computed from a dataset $\{x_1, \ldots, x_N\}$. We first sort the data points into ascending order, that is, $x_{(1)}, \ldots, x_{(N)}$, with $x_{(1)}$ the smallest and $x_{(N)}$ the largest value in the original dataset.

Computing quantiles with observed data is not entirely straightforward. The reason is that the quantile q_α usually falls between $x_{(i)}$ and $x_{(i+1)}$ for some integer i. For example, with six data points, the median $q_{0.5}$ falls between $x_{(3)}$ and $x_{(4)}$, so we let $q_{0.5}$ be the average of the two values. In general, q_α need not fall midway between $x_{(i)}$ and $x_{(i+1)}$, so various schemes compute q_α differently. Hyndman and Y. N. Fan (1996) listed nine different schemes for computing quantiles. Fortunately, when $N \geq 100$, the differences between the various schemes become negligible.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.10 Skewness and Kurtosis B☺

As the mean is computed from the first moment of the data and the variance from the second moment, one can proceed onto *skewness*, a third moment statistic. The population skewness coefficient is traditionally defined by

$$\gamma_{\mathrm{p}} = \mathrm{E}\left[\left(\frac{x-\mu_x}{\sigma}\right)^3\right] = \frac{\mathrm{E}[(x-\mu_x)^3]}{\sigma^3}\,, \tag{2.48}$$

where μ_x and σ are the population mean and standard deviation, respectively.

The sample skewness is computed from

$$\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^3}{s^3}, \tag{2.49}$$

where $\overline{x}$ and s are the sample mean and standard deviation, respectively.

The skewness is easily seen to be zero for a symmetric probability distribution, for example the Gaussian distribution (Section 3.4). If the right tail of the Gaussian is made longer or fatter, the skewness becomes positive. If the left tail is longer or fatter, the skewness becomes negative. Note that while symmetry implies zero skewness, the converse is not true, as one can make the left tail fatter and the right tail longer to compensate each other, leaving the skewness at zero. Distributions for variables that are non-negative, for example wind speed, precipitation amount, pollution concentration, and so on tend to have positive skewness.

The cubic power makes the traditional skewness coefficient very non-resistant to outliers, thus rather unreliable to use in practice. A resistant skewness coefficient based on quartiles was introduced in Bowley (1901), generally regarded as the first English-language textbook on statistics, with

$$\gamma_{\mathrm{B}} = \frac{q_{0.75}+q_{0.25}-2q_{0.5}}{q_{0.75}-q_{0.25}}, \tag{2.50}$$

where the denominator is simply the IQR. Bowley's skewness is also called Yule's coefficient or the Yule–Kendall index (Yule, 1912).

From the fourth moment of the data, the population *kurtosis* is defined by

$$\beta = \frac{\mathrm{E}[(x-\mu_x)^4]}{\sigma^4}\,. \tag{2.51}$$

For a Gaussian distribution, $\beta = 3$. Distributions with more outliers than the Gaussian has $\beta > 3$, while those with fewer outliers has $\beta < 3$. Many authors use 'kurtosis' to mean 'excess kurtosis' (i.e. the kurtosis of a distribution relative to that of a Gaussian), that is, $\beta' \equiv \beta - 3$, so the Gaussian has $\beta' = 0$. With the fourth power involved, the traditional kurtosis is obviously not resistant to outliers.

For more resistant higher moments, L-moments are often used (Hosking, 1990; von Storch and Zwiers, 1999).

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.11 Correlation A☺

2.11.1 Pearson Correlation A☺

The (Pearson) correlation coefficient, widely used to represent the strength of the linear relationship between two variables x and y, is defined by

$$\hat{\rho}_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, \tag{2.52}$$

where σ_x and σ_y are the population standard deviations for x and y, respectively.

For a sample containing N pairs of (x, y) measurements or observations, the *sample correlation* is computed by

$$\rho \equiv \rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{N}(y_i - \overline{y})^2\right]^{\frac{1}{2}}}, \tag{2.53}$$

which lies between -1 and $+1$. At the value $+1$, x and y will show a perfect straight-line relation with a positive slope, whereas at -1, the perfect straight line will have a negative slope. With increasing noise in the data, the sample correlation moves towards 0.

This formula for ρ involves two passes with the data, as it requires a first pass to compute the means $\overline{x}$ and $\overline{y}$. Substituting in the formulas for the means (2.25), one can rewrite (2.53) as

$$\rho = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]^{\frac{1}{2}} \left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]^{\frac{1}{2}}}, \tag{2.54}$$

where ρ can be computed by a single pass. For some datasets, this formula can lead to the subtraction of similar numbers, resulting in the loss of significant figures. For instance, consider a number with seven significant figures: 0.1234567. If one is to subtract from it the similar number 0.1234511, one gets 0.0000056, with only two significant figures.

With two variables x and y, a *scatterplot* that plots the data points as dots in the x-y plane is often useful for visualizing the distribution of the data points. In Fig. 2.4, scatterplots of synthetic (x,y) data are shown, along with the corresponding correlation coefficient. The x variable is from a Gaussian distribution with zero mean and unit standard deviation, while $y = x + \text{noise}$ in Figs. 2.4(a), (c) and (e), and $y = -x + \text{noise}$ in Figs. 2.4(b), (d) and (f). The added random noise is Gaussian, with increasing noise lowering the magnitude of the correlation from (a) to (f).

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

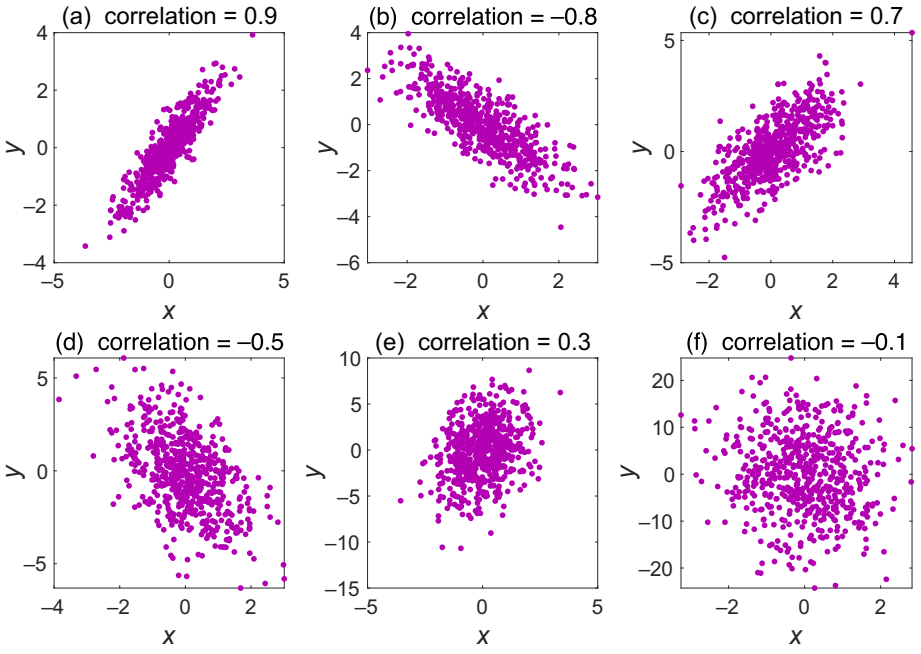

Figure 2.4 Scatterplots showing distribution of (x, y) data and the corresponding Pearson correlation coefficient as the noise level rises from (a) to (f).

It will be instructive to look at scatterplots and correlations with real data. The daily surface air temperature, relative humidity, wind speed and sea level pressure at Vancouver, British Columbia, Canada, from averaged hourly observations by Environment and Climate Change Canada, were downloaded from www.weatherstats.ca. In Fig. 2.5(a), the correlation is -0.33 and, indeed, focusing on where the data density is high, we see lower relative humidity concurring with higher temperature, which is not surprising since Vancouver has rainy winters and dry summers. However, when temperature becomes very low, the relative humidity drops as temperature drops, opposite to our expectation from the negative correlation coefficient. The reason for this behaviour is that in winter there are occasional Arctic air outbreaks, bringing very cold, dry air from the Arctic. The strongest correlation of -0.38 was found between pressure and wind speed in Fig. 2.5(d), as low pressure systems give rise to stormy weather with high wind speeds.

The Pearson correlation assumes a linear relation between x and y; however, the sample correlation is not *robust* to deviations from linearity in the relation, as illustrated in Fig. 2.6(a) where $\rho \approx 0$ though there is a strong (non-linear) relationship between the two variables. Thus, the correlation can be misleading when the underlying relation is non-linear. Furthermore, the sample correlation is not *resistant* to outliers, where in Fig. 2.6(b) $\rho = -0.50$. If the single outlier is

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

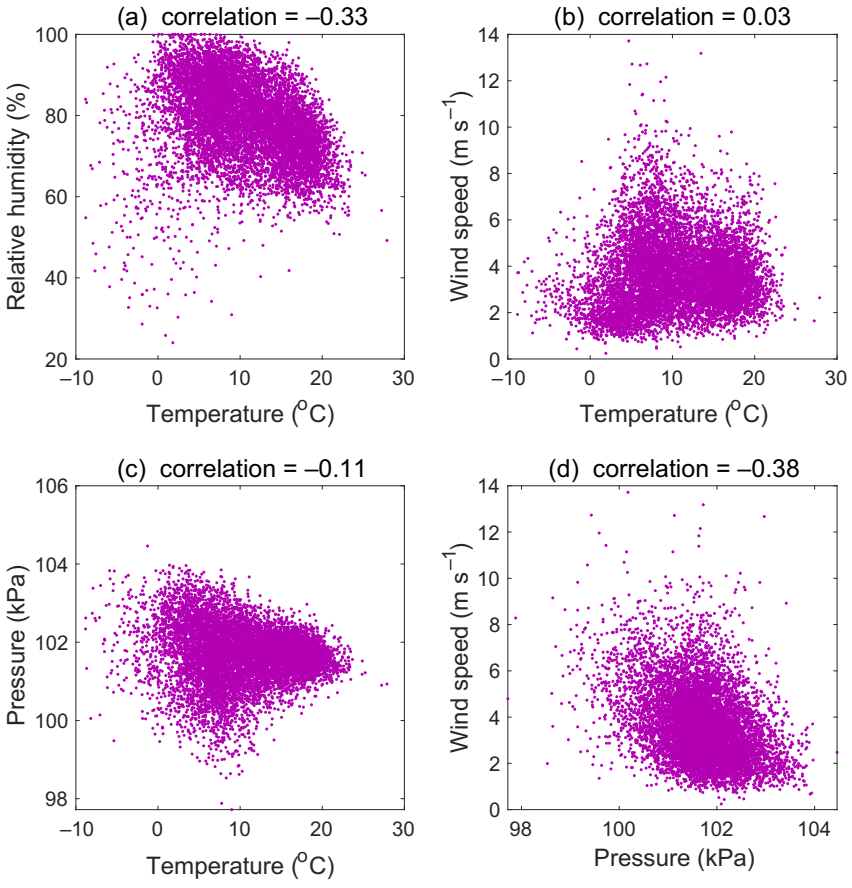

Figure 2.5 Scatterplots and Pearson correlation coefficients of daily weather variables from Vancouver, BC, Canada, with 25 years of data (1993–2017). [Data source: weatherstats.ca based on Environment and Climate Change Canada data.]

removed, ρ changes from -0.50 to $+0.70$, that is, the strong positive correlation was completely masked by one outlier. Later in this chapter, we will study more robust/resistant estimates of the correlation.

If there are d variables, for example d stations reporting the air pressure, then correlations between the variables lead to a correlation matrix

$$\mathbf{C} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \cdots & \rho_{dd} \end{bmatrix}, \tag{2.55}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

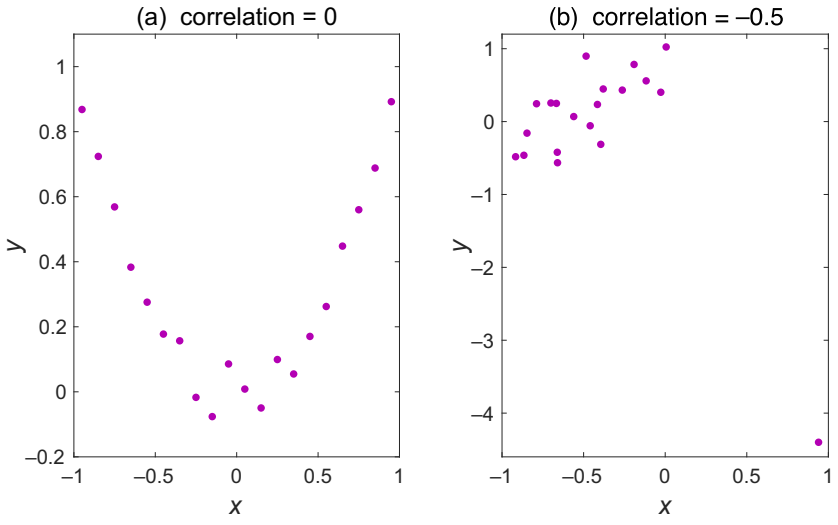

Figure 2.6 (a) An example illustrating that correlation is not robust to deviations from linearity. Here, the strong non-linear relation between x and y is completely missed by the near-zero correlation coefficient. (b) An example showing that correlation is not resistant to outliers. By removing the single outlier on the lower right corner, the correlation coefficient changes from negative to positive.

where ρ_{ij} is the correlation between the ith and the jth variables. The diagonal elements of the matrix satisfy $\rho_{ii} = 1$, and the matrix is symmetric, that is, $\rho_{ij} = \rho_{ji}$. The jth column of $\mathbf{C}$ gives the correlations between the variable j and all other variables. The correlation matrix is simply the normalized version of the covariance matrix $\mathrm{cov}(\mathbf{x})$ in (2.33).

2.11.2 Serial Correlation A☺

Often the observations are measurements at regular time intervals, that is, time series, and there is *serial correlation* in the time series – that is, neighbouring data points in the time series are correlated. Serial correlation is well illustrated by *persistence* in weather patterns, for example, if it rains one day, it increases the probability of rain the following day. Serial correlation in a single time series is called *autocorrelation*. Serial correlation can involve more than one time series, for example rainfall today can increase river runoff tomorrow.

In making statistical estimates, it is common to also estimate the confidence interval (Section 4.4). For instance, for a statistical estimate z, we would like to estimate the interval $[z_{\mathrm{lower}}, z_{\mathrm{upper}}]$ containing z, where there is 95% chance that the true parameter z_{true} lies within this confidence interval. In hypothesis testing (Section 4.1), one would like to know if the observed z is enough to

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

reject the null hypothesis at a certain level. In both cases, the answers depend on the sample size N, that is, larger sample size makes the confidence intervals narrower, or z significant at a more desirable level.

Unfortunately, traditional confidence interval estimates and significance tests assume the N data points are all independent observations. With serial correlation, this assumption is false, as the number of independent observations is smaller and sometimes much smaller than N. For example, suppose the weather is typically three days of rain, alternating with five days of sun, that is, one has a typical rainy event of three days alternating with a sunny event of five days, so over eight days, there are two events. If one has $N = 80$, there are only about 20 events, so the *effective sample size* N_{eff} is only about 20. One needs to use N_{eff} instead of N in the significance tests and confidence interval estimates when there is serial correlation in the data (see Section 4.2.4).

To determine the degree of autocorrelation in a time series, we use the autocorrelation coefficient, where a copy of the time series is shifted in time by a lag of l time intervals and then correlated with the original time series. The lag-l autocorrelation coefficient is given by

$$\rho(l) = \frac{\sum_{i=1}^{N-l} [(x_i - \overline{x})(x_{i+l} - \overline{x})]}{\sum_{i=1}^{N} (x_i - \overline{x})^2}, \tag{2.56}$$

where $\overline{x}$ is the sample mean. There are other estimators of the autocorrelation function, though this version is most commonly used (von Storch and Zwiers, 1999, p. 252). The function $\rho(l)$ has the value 1 at lag 0. From symmetry, one defines $\rho(-l) = \rho(l)$.

The effective sample size can be derived as

$$N_{\text{eff}} = N \left[1 + 2 \sum_{l=1}^{N-1} \left(1 - \frac{l}{N} \right) \rho(l) \right]^{-1}, \tag{2.57}$$

(von Storch and Zwiers, 1999, p. 372; Thiébaux and Zwiers, 1984). Thiébaux and Zwiers (1984) compared several methods for estimating N_{eff}. Their direct estimation approach involves substituting values of $\rho(l)$ into (2.57), Unfortunately, direct estimation involves estimating $\rho(l)$ at large lags (when the true autocorrelation function is effectively zero) and summing over many such terms. Even using the option of truncating the summation at large lags, direct estimation was not among the better methods (Thiébaux and Zwiers, 1984).

A better approach is to assume an auto-regressive (AR) process (Section 11.8). For the simplest AR process of order 1 (abbreviated as AR(1)), when N is large, the effective sample size is approximated by

$$N'_{\text{eff}} \approx N \frac{1 - \rho(1)}{1 + \rho(1)}, \tag{2.58}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

with $\rho(1)$ being the lag-1 autocorrelation coefficient (Zwiers and von Storch, 1995). For $0 \le \rho(1) < 1$, (2.58) gives $0 < N'_{\text{eff}} \le N$. If $\rho(1) = 0$, $N'_{\text{eff}} = N$, as expected for independent data. Sallenger et al. (2012) found that $\rho(1)$ from (2.56) did not give stable estimates for noisy time series; instead, they fitted an AR(1) model and substituted the AR(1) coefficient for $\rho(1)$ in (2.58) to obtain N'_{eff}.

It is possible to have $N'_{\text{eff}} > N$ if $\rho(1) < 0$. To keep the effective sample size within a reasonable range, Zwiers and von Storch (1995) recommended using

$$N_{\text{eff}} = \begin{cases} 2 & \text{if } N'_{\text{eff}} \le 2, \\ N'_{\text{eff}} & \text{if } 2 < N'_{\text{eff}} \le N, \\ N & \text{if } N < N'_{\text{eff}}, \end{cases} \tag{2.59}$$

with N'_{eff} computed from (2.58). How N_{eff} is used in hypothesis testing is further pursued in Section 4.2.4.

For illustration, the autocorrelation function was computed for the daily temperature at Vancouver, BC in Fig. 2.7. That the autocorrelation function in Fig. 2.7(a) has a strong trough at around 180 days and a strong peak at around 360 days merely indicates that the time series has a strong seasonal cycle. For Gaussian *white noise*,[3] 95% of the distribution falls within the interval $[-1.96/\sqrt{N}, 1.96/\sqrt{N}]$, which is marked by the two horizontal lines in Fig. 2.7, (see Section 3.4) (Box, Jenkins, et al., 2015, Section 2.1.6; von Storch and Zwiers, 1999, pp. 252–253), that is, outside of this interval, there is only 5% chance the true autocorrelation is zero.

For the short record of $N = 90$ days during the winter of 2016–2017 (Fig. 2.7(b)), (2.59) gave $N_{\text{eff}} \approx 7.9$, an order of magnitude smaller than N.

2.11.3 Spearman Rank Correlation A☺

For the correlation to be more robust and resistant to outliers, the Spearman rank correlation (Spearman, 1904) is often used instead of the Pearson correlation. If the data $\{x_1, \ldots, x_N\}$ are rearranged in order according to their size (starting with the smallest), and if x is the nth member, then $\text{rank}(x) \equiv r_x = n$. The Spearman correlation ρ_{spearman} is simply the Pearson correlation ρ of r_x and r_y, that is,

$$\rho_{\text{spearman}}(x, y) = \rho(r_x, r_y). \tag{2.60}$$

Spearman correlation assesses how well the relationship between two variables can be described by a monotonic function. If the relation is perfectly monotonic, that is, if $x_i < x_j$, then $y_i < y_j$ for all $i \ne j$, then the Spearman correlation takes on the maximum value of 1. The minimum value of -1 is attained if $x_i < x_j$, then $y_i > y_j$ for all $i \ne j$. Thus, the Pearson correlation measures if the relation between two variables is linear, while Spearman measures if the relation is monotonic (regardless whether it is linear or non-linear).

[3] *White noise* is a random signal having equal intensity at all frequencies. The values at any two times are identically distributed and statistically independent; thus, the autocorrelation $\rho(l) = 0$ for all $l \ne 0$.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

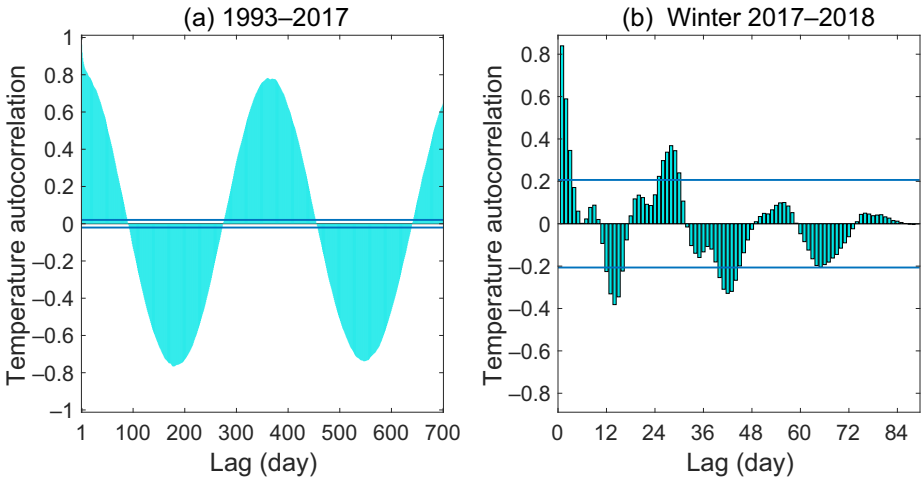

Figure 2.7 Autocorrelation function for the daily temperature at Vancouver, BC during (a) 1993–2017 and (b) winter of 2016–2017 (Dec.-Feb.), with the horizontal lines indicating the 95% confidence interval. [Data source: weatherstats.ca based on Environment and Climate Change Canada data.]

For example, if five measurements of x yielded the values $1, 3, 0, 3$ and 6, then the corresponding r_x values are $2, 3.5, 1, 3.5$ and 5 (where the tied values were all assigned an averaged rank). If measurements of y yielded $2, 3, -1, 7$ and -99 (an outlier), then the corresponding r_y values are $3, 4, 2, 5$ and 1. The Spearman rank correlation is -0.05, while the Pearson correlation is -0.79, which shows the strong influence exerted by an outlier.

In Fig. 2.6(a), the Pearson correlation is 0.00 while the Spearman correlation is -0.02, but in Fig. 2.6(b), it is -0.50 for Pearson versus $+0.44$ for Spearman. If the single outlier at the bottom right corner of Fig. 2.6(b) is removed, it is 0.70 for Pearson and 0.68 for Spearman. Clearly, the Spearman correlation is much more resistant to the outlier than Pearson.

There are alternative robust and resistant correlations, such as the Kendall rank correlation and the biweight midcorrelation.

2.11.4 Kendall Rank Correlation A☺

An alternative approach to rank correlation is via Kendall rank correlation or Kendall's tau (after the Greek letter τ) (Kendall, 1938; Kendall, 1945). Given x_i and y_i $(i = 1, \ldots, N)$, we say that a pair (i, j), $i < j$, is:
- *concordant* when $x_i - x_j$ and $y_i - y_j$ are both non-zero and have the same sign;
- *discordant* when $x_i - x_j$ and $y_i - y_j$ are both non-zero and have opposite signs.

Let C and D be the number of concordant pairs and discordant pairs, respectively. The total number of pairs (i, j) with $i < j$ is $N_0 = N(N-1)/2$.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

Kendall's τ_A was defined originally in Kendall (1938) as

$$\tau_A = \frac{C - D}{N_0}, \tag{2.61}$$

which has a rather simple interpretation, namely the number of concordant pairs minus the number of discordant pairs, divided by the total number of pairs. If all pairs are concordant, then $C = N_0$, $D = 0$ and $\tau_A = 1$, and if all pairs are discordant, then $\tau_A = -1$.

The denominator was adjusted for ties (i.e. $x_i = x_j$ or $y_i = y_j$) in Kendall (1945), and statistical packages usually implement this modified τ (often called τ_B), so τ_B ranges from -1 to $+1$ even with tied data. τ_B is defined by

$$\tau_B = \frac{C - D}{\sqrt{(N_0 - T_x)(N_0 - T_y)}}, \tag{2.62}$$

with

$$T_x = \sum_{j=1}^{g^{(x)}} t_j^{(x)}(t_j^{(x)} - 1)/2, \tag{2.63}$$

$$T_y = \sum_{j=1}^{g^{(y)}} t_j^{(y)}(t_j^{(y)} - 1)/2, \tag{2.64}$$

where $g^{(x)}$ is the number of tied groups in the variable x and $t_j^{(x)}$ is the size of tied group j (e.g. if the value $x = 5.1$ appears twice, $t_j^{(x)} = 2$), and $g^{(y)}$ and $t_j^{(y)}$ are similarly defined for the variable y. When there are no ties, $\tau_A = \tau_B$.

In Fig. 2.6(b), the correlation is -0.50 for Pearson, 0.44 for Spearman and 0.35 for Kendall. If the single outlier at the bottom right corner of Fig. 2.6(b) is removed, it is 0.70 for Pearson, 0.68 for Spearman and 0.50 for Kendall.

The usage of Kendall's τ has been increasing in recent decades, though whether it is better or worse than the Spearman correlation is problem dependent (W. C. Xu et al., 2013).

2.11.5 Biweight Midcorrelation B☺

We have seen one approach in making correlation more robust and resistant, namely using ranks as in the Spearman and Kendall rank correlations. A different approach involves replacing the non-robust/resistant measures in the Pearson correlation, that is, the mean and deviation from the mean, by the corresponding robust/resistant ones, that is, the median and the deviation from the median. This second approach is used in the biweight midcorrelation (Mosteller and Tukey, 1977; Wilcox, 2004, pp. 203–209).

To calculate the biweight midcorrelation function $\text{bicor}(x, y)$, first rescale x and y by

$$p_i = \frac{x_i - M_x}{9\,\text{MAD}_x}, \quad q_i = \frac{y_i - M_y}{9\,\text{MAD}_y}, \quad i = 1, \ldots, N, \tag{2.65}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

where M_x and M_y are the median values of x and y, respectively, and MAD_x and MAD_y (the median absolute deviations) are the median values of $|x_i - M_x|$ and $|y_i - M_y|$, respectively.

Next, define the weights (called 'biweights' by Beaton and Tukey (1974))

$$w_i^{(x)} = \begin{cases} (1 - p_i^2)^2, & \text{if } |p_i| < 1 \\ 0, & \text{if } |p_i| \geq 1, \end{cases} \tag{2.66}$$

$$w_i^{(y)} = \begin{cases} (1 - q_i^2)^2, & \text{if } |q_i| < 1 \\ 0, & \text{if } |q_i| \geq 1. \end{cases} \tag{2.67}$$

The biweight midcorrelation is defined by

$$\text{bicor}(x, y) = \frac{\sum_{i=1}^{N} w_i^{(x)}(x_i - M_x)\, w_i^{(y)}(y_i - M_y)}{\left\{\sum_{i=1}^{N}[w_i^{(x)}(x_i - M_x)]^2\right\}^{\frac{1}{2}} \left\{\sum_{i=1}^{N}[w_i^{(y)}(y_i - M_y)]^2\right\}^{\frac{1}{2}}}. \tag{2.68}$$

Formally, bicor resembles the Pearson correlation (2.53), except for the presence of the weights $w_i^{(x)}$ and $w_i^{(y)}$ and the use of medians M_x and M_y instead of the means. The weights in (2.66) and (2.66) are set to zero for outliers (large $|p_i|$ or $|q_i|$); thus, bicor is resistant to outliers. The biweight midcorrelation, like the Pearson correlation, ranges from -1 (negative association) to $+1$ (positive association).

2.12 Exploratory Data Analysis A☺

In statistics, exploratory data analysis (EDA) was pioneered by John W. Tukey, who wrote the classic book entitled *Exploratory Data Analysis* (Tukey, 1977). Tukey felt that statistics placed too much emphasis on statistical hypothesis testing, so he advocated EDA, which tries to explore and visualize the data, thereby letting the data suggest what hypotheses to test. Besides using more robust/resistant statistics such as the median and the quartiles to summarize a dataset than using the traditional mean and standard deviation, EDA also uses graphical methods extensively to aid in visualizing the structure of the datasets. Graphical methods include scatterplots (Fig. 2.5), histograms, quantile–quantile plots and boxplots.

2.12.1 Histograms A☺

Histograms (from 'historical diagrams'), introduced by Pearson (1895), present the probability distribution of a given variable by plotting the frequencies of observations occurring over the domain of the variable. To construct a histogram, the domain is partitioned into intervals (called 'bins' or 'buckets'), and the frequency, that is, how many observed values fall into each bin, is counted. The frequencies can be simply the raw counts, or normalized, that is, dividing the counts by the total number of observations. The bins are usually of equal

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

width, but can be of unequal width. With normalized frequencies, the area over each bin gives the probability of occurrence within that interval. The bin width cannot be chosen to be too wide, which smooths out important details in the histogram, nor too narrow, which gives a noisy-looking histogram. (Scott, 2015, p. 78) recommends using a bin width $\leq 2.6\,\mathrm{IQR}/(N^{1/3})$, where IQR is the interquartile range and N the sample size. Most histogram packages will have a reasonable default bin width, so the user does not have to specify the bin width.

Figure 2.8 gives an example of using the histogram method on the weather data for Vancouver, BC. The histogram gives the actual distribution of the data, while a Gaussian distribution curve has also been fitted to the data.[4] Comparing the histogram with the Gaussian curve tells us how close the observed distribution is to a Gaussian distribution. Temperature in Fig. 2.8(a) actually shows a bimodal distribution (i.e. having two humps) in contrast to the unimodal Gaussian, while wind speed in Fig. 2.8(c) is also clearly non-Gaussian. The fit for relative humidity (Fig. 2.8(b)) is poor for the right tail as humidity cannot exceed 100%, and the pressure distribution (Fig. 2.8(d)) is more narrowly distributed than a Gaussian. For precipitation (Fig. 2.8(e)), most days have no precipitation – these dry days are omitted in Fig. 2.8(f). Although the Gaussian distribution is called the 'normal distribution', in reality, environmental variables often do not closely resemble Gaussians.

2.12.2 Quantile–Quantile (Q–Q) Plots B☺

A *quantile–quantile plot* (Q–Q plot) is a probability plot, which provides a graphical tool for comparing two probability distributions by plotting their quantiles against each other. For a chosen set of quantiles, a point (x, y) on the plot corresponds to one of the quantiles of the y distribution plotted against the same quantile of the x distribution.

There are two ways to use a Q–Q plot: (i) to compare observed data with a specified theoretical distribution (e.g. a Gaussian distribution) and (ii) to assess whether two sets of observed data obey the same distribution. If the agreement between the two distributions is perfect, then the plot is a straight line.

There are many ways to choose the quantiles for the plot. If the distribution for the observations y_i, $(i = 1, \ldots, N)$, is to be compared with a specified theoretical distribution, one way is to simply use N quantiles. The y data are sorted into ascending order, $y^1, \ldots, y^N$, then the ith ordered value $y^{(i)}$ is plotted against the $(i - \frac{1}{2})/N$ quantile of the theoretical distribution along the x-axis.

The Q–Q plot for Vancouver's daily temperature versus the standard Gaussian distribution (i.e. Gaussian with zero mean and unit standard deviation) shows the temperature to have weaker tails than the Gaussian, especially for high temperatures (Fig. 2.9(a)), which can also be seen in (Fig. 2.8(a)). However the bimodal structure seen in the histogram is much less noticeable in the Q–Q

[4] The Gaussian distribution is specified by two parameters, namely its population mean μ and variance σ^2 (see Section 3.4). From the dataset, compute the sample mean and variance and use these as the parameters of the Gaussian distribution.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

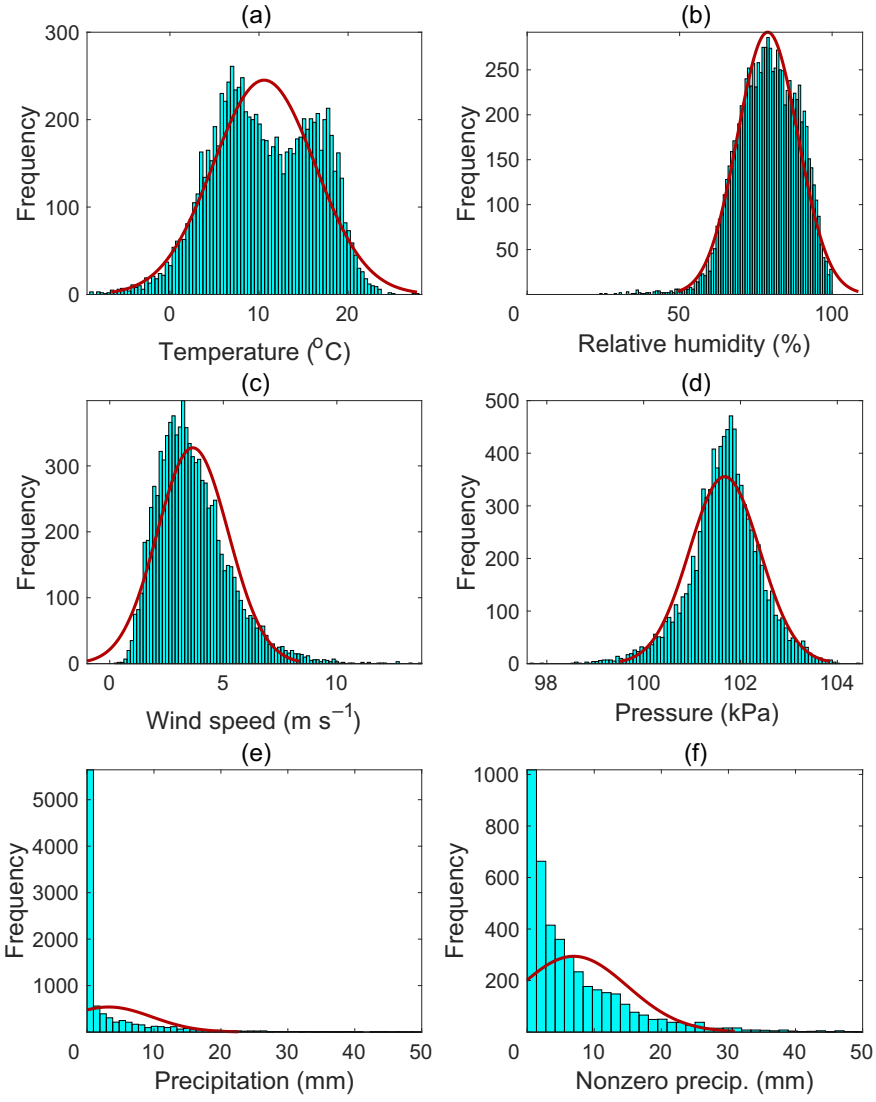

Figure 2.8 Histogram for the distribution of daily (a) temperature, (b) relative humidity, (c) wind speed, (d) sea level pressure, (e) precipitation and (f) non-zero precipitation in Vancouver, BC from 1993 to 2017. A Gaussian distribution curve has also been fitted to the data. Relative humidity is bounded between 0% and 100%, and wind speed is non-negative. Since 53.4% of the days in (e) have no precipitation, the dry days are omitted in (f). [Data source: weatherstats.ca based on Environment and Climate Change Canada data.]

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

plot. Thus, comparing the histogram with the Q–Q plot, the Q–Q plot tends to be better in revealing the departure of the observations from the theoretical distribution near the tails, but the histogram tends to be better in showing the departure away from the tails.

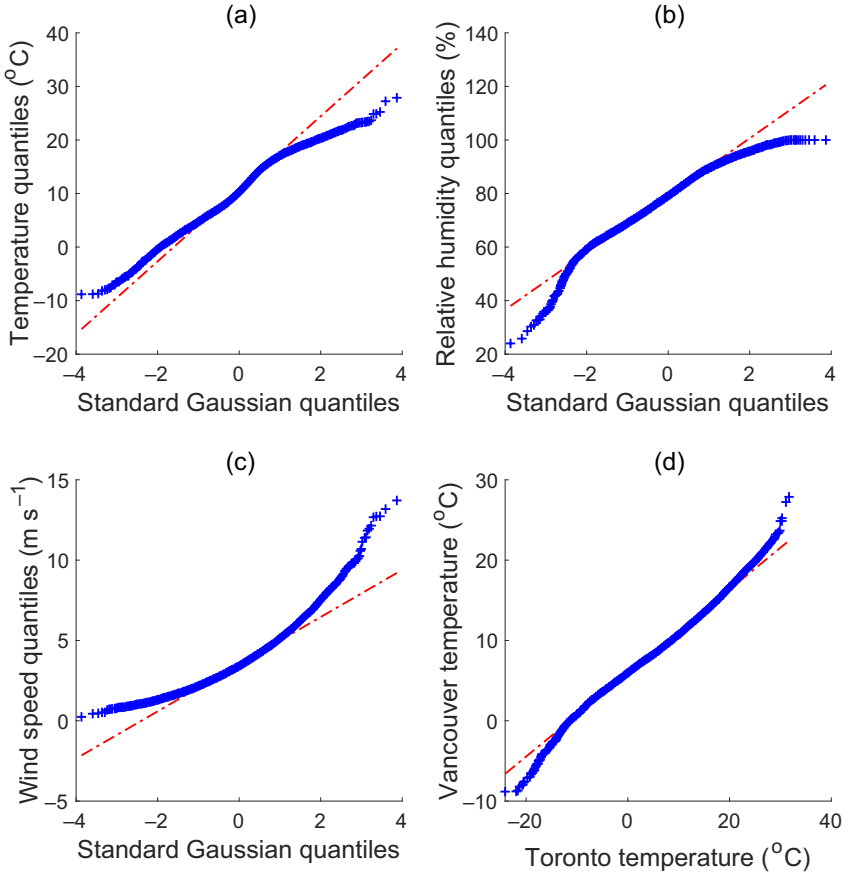

Figure 2.9 Quantile–quantile plots where quantiles of the daily (a) temperature, (b) relative humidity and (c) wind speed in Vancouver, BC from 1993 to 2017 are plotted against the quantiles of the standard Gaussian distribution as indicated by the '+' symbols. If the observed distribution is a perfect Gaussian, the plot will fall on the straight (dot-dashed) line. In (d), the quantiles of the temperature in Toronto, Ontario are plotted against those from Vancouver. [Data source: weatherstats.ca based on Environment and Climate Change Canada data.]

Figure 2.9(b) shows the relative humidity to have a shorter tail than the Gaussian at high values but a longer tail at low values. On the other hand,

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

wind speed (Fig. 2.9(c)) has a longer tail than the Gaussian at high values but a shorter tail at low values.

The Q–Q plot can also be used to assesses whether two sets of observations have the same distribution. One plots the quantile values for the first dataset along the x-axis and the corresponding quantile values for the second dataset along the y-axis. The two datasets can have different numbers of data points, as a Q–Q plot only plots selected quantiles. If the resulting plot is linear, the two datasets obey the same distribution. Figure 2.9(d) shows the Q–Q plot of Toronto's temperature versus Vancouver's temperature. Toronto's temperature, though having a larger range than Vancouver's, has relatively shorter tails.

2.12.3 Boxplots A☺

Tukey (1977) advocated using five numbers to summarize a dataset, that is, the median, the lower and upper quartiles (i.e. $q_{0.25}$ and $q_{0.75}$), and the minimum and maximum values. *Boxplots* (or box-and-whisker plots) arose as a visualization tool for the five-number summary (Tukey, 1977).

The top and bottom of each 'box' are the upper and lower quartiles of the sample data, respectively, with the distance between the top and bottom indicating the interquartile range (IQR). The horizontal line or 'waist line' within each box is the sample median. Skewness is present if the median is not centred in the box.

The whiskers are the (dashed) lines extending above and below each box. The most common convention is to have the whisker above the box extending from $q_{0.75}$ to a furthest observation not more than 1.5 IQR above $q_{0.75}$. Any observation beyond is considered an outlier and is plotted as a '+' or 'o' symbol. Similarly, the lower whisker extends from $q_{0.25}$ to a furthest observation not more than 1.5 IQR below $q_{0.25}$, with any observation beyond plotted as an outlier. For a Gaussian distribution, 99.7% of the distribution lies within the interval $[q_{0.25} - 1.5\ \text{IQR},\ \ q_{0.75} + 1.5\ \text{IQR}]$.

A common variant of the boxplot displays notches on the two sides of the waistline, that is, '>–<' (Fig. 2.10(a)), with the height of the notches indicating the uncertainty of the sample median (McGill et al., 1978). The notches extend

$$\pm\, 1.57\ \text{IQR}/\sqrt{N} \tag{2.69}$$

from the sample median. If the notches from two boxes do not overlap, then the two sample medians are considered different at the 5% significance level.

Figure 2.10 shows boxplots of weather variables at three Canadian cities, Vancouver and Victoria in British Columbia on the west coast and Toronto, Ontario to the east. In (a), there are only 90 data points from Dec. 2016 to Feb. 2017, so the notches are much wider than those in (b), where there are 25 years of daily data. Toronto is seen to have a much larger temperature range in (b) and lower relative humidity in (c) than the two west coast cities, while Victoria is seen to have lower wind speed than the other two cities in (d).

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

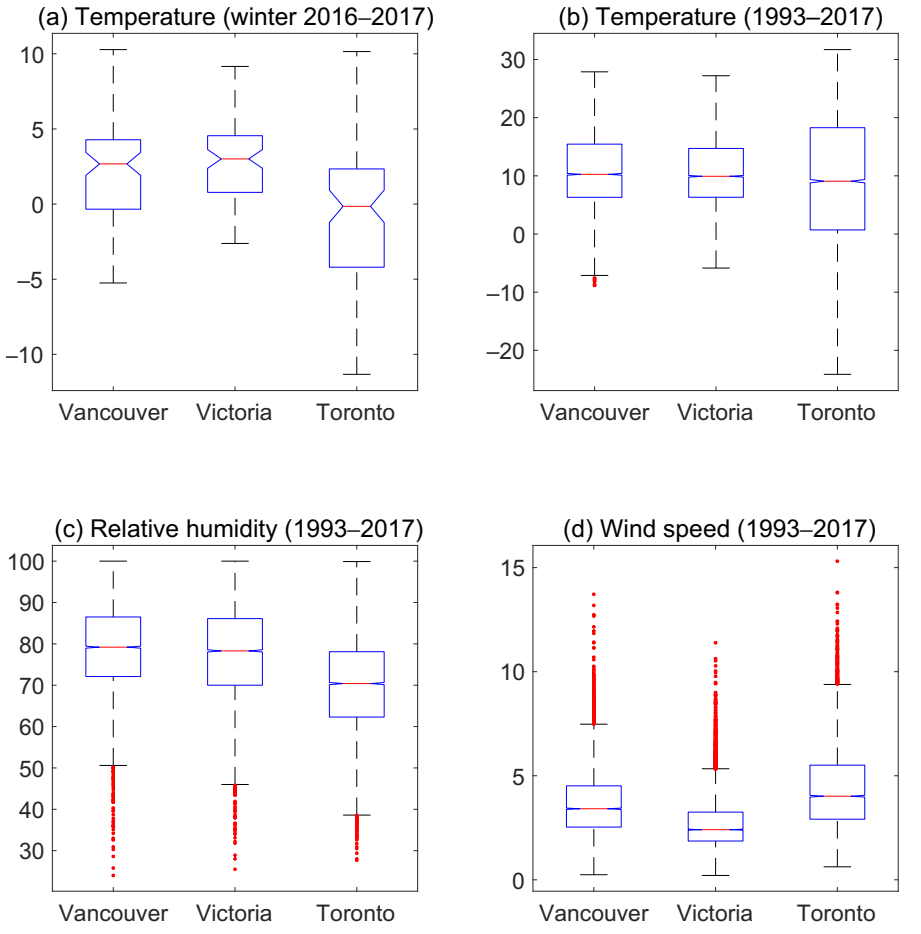

Figure 2.10 Boxplots for the daily weather at three Canadian cities: (a) temperature during the winter of 2016–2017, and (b) temperature, (c) relative humidity and (d) wind speed from 1993 to 2017. [Data source: weatherstats.ca based on Environment and Climate Change Canada data.]

The astute reader may question the notches computed from (2.69) since there is serial correlation in the weather data. Indeed, if we assume the effective sample size $N_{\text{eff}} \approx 8$ for Vancouver temperature (winter 2016–2017) as in Section 2.11.2, then replacing N by N_{eff} in (2.69) would make the width of the notches 3.4 times as wide. Unfortunately, common boxplot packages do not provide an option for replacing N by N_{eff} when computing the notches, leading to notch widths that are too narrow for serially correlated data. In such situations, it is best to turn off the option for displaying notches in the boxplot.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.13 Mahalanobis Distance A☺

For a one-dimensional dataset with mean μ and standard deviation σ, and a particular data point x, the Mahalanobis distance measures the number of standard deviations between the data point x and the mean μ, that is,

$$d_{\mathrm{M}} = \sqrt{(x-\mu)^2}/\sigma. \tag{2.70}$$

For higher-dimensional data, the Mahalanobis distance is defined by

$$d_{\mathrm{M}} = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\,, \tag{2.71}$$

where $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$, $\mathbf{C}$ is the covariance matrix $\mathrm{cov}(\mathbf{x})$ from (2.33) and $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$. Clearly, (2.70) is the special one-dimensional case of the general formula (2.71), as $\mathbf{C}$ reduces to the variance σ^2 in 1-D. If $\mathbf{C}^{-1}$ is the identity matrix, (2.71) reduces to the familiar Euclidean distance.

Figure 2.11 illustrates why the Mahalanobis distance (and not the Euclidean distance) is the appropriate distance for determining if a data point $\mathbf{x}$ is far from the centre of a dataset. Consider the two points marked by the asterisk and the star. The Euclidean distance (Fig. 2.11(a)) between the asterisk and the centre of the dataset is 7.44, versus 4.48 between the star and the centre. However, in terms of the Mahalanobis distance (Fig. 2.11(d)), the distance from the centre is 2.48 for the asterisk and 4.46 for the star. Thus the Mahalanobis distance correctly indicates the star as being much further from the centre than the asterisk, and should be considered an outlier.

Robust methods have been developed to estimate $\mathbf{C}$ when the data is non-Gaussian. These include the fast-minimum covariance determinant (Fast-MCD) method (Rousseeuw and van Driessen, 1999), the orthogonalized Gnanadesikan–Kettenring (OGK) method (Maronna and Zamar, 2002) and the Olive–Hawkins method (Olive, 2004). Of the three methods, the author's choice is the OGK method, based on some limited tests comparing the accuracy and speed of the three methods.

2.13.1 Mahalanobis Distance and Principal Component Analysis B😐

One can compute the Mahalanobis distance from (2.71), using the sample covariance matrix for $\mathbf{C}$ and the sample mean for $\boldsymbol{\mu}$. However, for the inquisitive reader, it is illuminating to see how the Mahalanobis distance can be derived from principal component analysis (PCA) (Section 9.1). From (9.29), one can write the centred data $\mathbf{x}-\boldsymbol{\mu}$ (Fig. 2.11(b)) as

$$\mathbf{x}-\boldsymbol{\mu} = \sum_i a_i \mathbf{e}_i\,, \tag{2.72}$$

where the summation is over all the PCA modes, that is, over all the principal components (PC) a_i multiplied by their corresponding eigenvectors $\mathbf{e}_i$. The eigenvectors satisfy the eigenequation

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

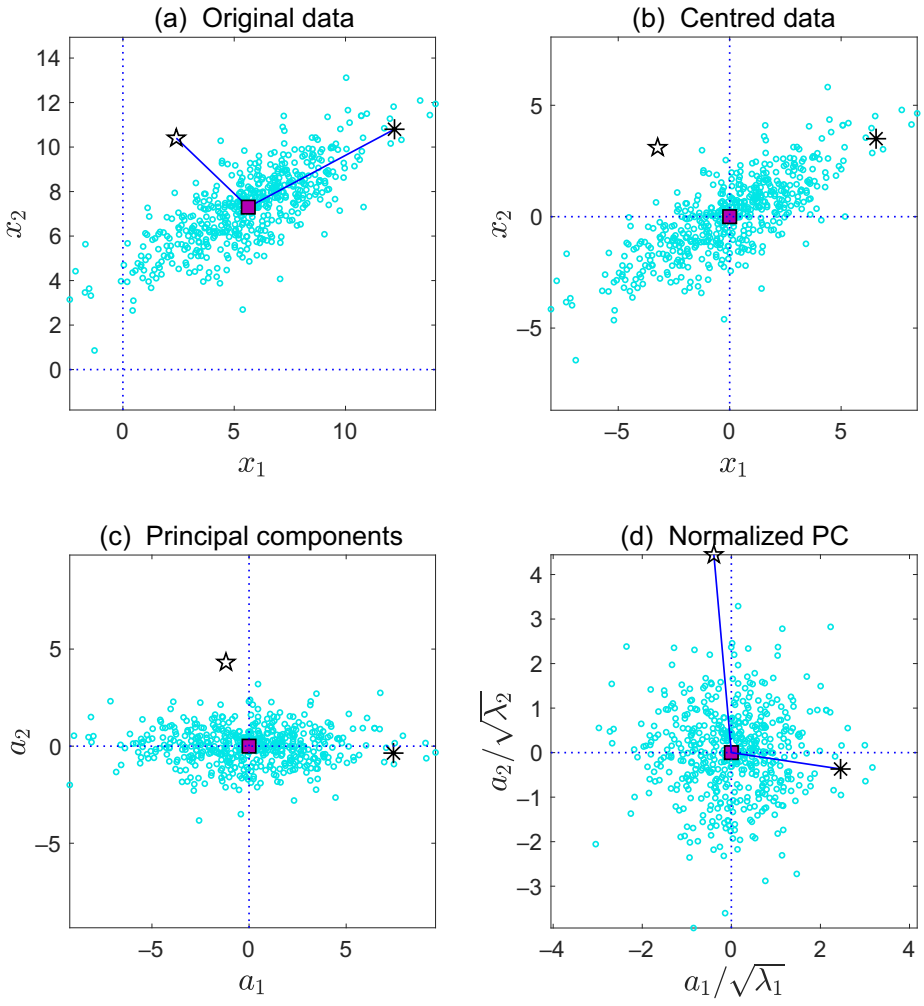

Figure 2.11 Mahalanobis distances versus Euclidean distances, as illustrated by two data points marked by the asterisk and the star. The square marks the centre (i.e. the mean) of the Gaussian dataset containing 500 points. (a) In the original data, the line marking the Euclidean distance from the centre is longer for the asterisk than for the star. (b) Subtracting the mean gives the centred data. (c) Principal components (a_1 and a_2) are obtained by rotating the centred data, so the direction of the maximum variance is along the horizontal axis. (d) Principal components are normalized to have unit variance in each direction. The line connecting the centre and the asterisk/star gives the Mahalanobis distance. Thus in terms of Euclidean distance, the asterisk is further from the centre than the star, but in terms of Mahalanobis distance, the star is further from the centre than the asterisk.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

$$\mathbf{C}\,\mathbf{e}_i = \lambda_i \mathbf{e}_i\,, \tag{2.73}$$

where λ_i are the eigenvalues. The vector $\mathbf{e}_1$ points in the direction of maximum variance of the dataset, while $\mathbf{e}_2$ points in the direction of maximum variance within the space orthogonal to $\mathbf{e}_1$. In general, $\mathbf{e}_i$ points in the direction of maximum variance within the space orthogonal to $\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}$. The PC a_1 is the coordinate in the $\mathbf{e}_1$ direction, while a_2 is the coordinate in the $\mathbf{e}_2$ direction (Fig. 2.11(c)). A common convention is to make all eigenvectors of unit length, then λ_i is the variance of a_i [see (9.37) and (9.39)].

Left multiplying (2.73) by $\mathbf{C}^{-1}$ gives

$$\mathbf{e}_i = \lambda_i \mathbf{C}^{-1}\mathbf{e}_i\,, \tag{2.74}$$

$$\lambda_i^{-1}\mathbf{e}_i = \mathbf{C}^{-1}\mathbf{e}_i\,. \tag{2.75}$$

$$\sum_i a_i \lambda_i^{-1}\mathbf{e}_i = \sum_i a_i \mathbf{C}^{-1}\mathbf{e}_i = \mathbf{C}^{-1}\sum_i a_i \mathbf{e}_i = \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\,, \tag{2.76}$$

upon invoking (2.72). Thus,

$$d_{\mathrm{M}}^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_j \sum_i a_j \mathbf{e}_j^{\mathrm{T}} a_i \lambda_i^{-1}\mathbf{e}_i\,. \tag{2.77}$$

With orthonormal eigenvectors, $\mathbf{e}_j^{\mathrm{T}}\mathbf{e}_i = \delta_{ij}$ (the Kronecker delta function, where $\delta_{ij} = 1$ if $i = j$, and 0 otherwise),

$$d_{\mathrm{M}}^2 = \sum_i a_i^2/\lambda_i\,. \tag{2.78}$$

The Mahalanobis distance squared is simply the sum over the square of the normalized principal components, as $a_i/\sqrt{\lambda_i}$ can be regarded as the normalized PC. Thus, the distance from the origin in the normalized PC space is the Mahalanobis distance (Fig. 2.11(d)).

2.14 Bayes Theorem A☺

Bayes theorem, named after the Reverend Thomas Bayes (1702–1761), plays a central role in modern statistics (Jaynes, 2003; N. D. Le and Zidek, 2006). Historically, it had a major role in the debate around the foundations of statistics, as the traditional '*frequentist*' school and the *Bayesian* school disagreed on how probabilities should be assigned in applications. Frequentists assign probabilities to random events according to their frequencies of occurrence or to subsets of populations as proportions of the whole. In contrast, Bayesians describe probabilities in terms of beliefs and degrees of uncertainty, similarly to how the general public uses probability. For instance, a sports fan prior

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

to the start of a sports tournament believes that team A, with a stellar historical record, has a probability of 70% for winning a game. However, after watching several mediocre games, the fan may modify his estimate of the winning probability to 50%. Similarly, to estimate the probability of a hypothesis, a Bayesian first specifies/guesses some prior probability, then updates it to a posterior probability by incorporating new data (evidence).

We will use a *classification* problem to illustrate the Bayes approach. Suppose a meteorologist wants to classify the approaching weather state as either storm (C_1) or non-storm (C_2). Assume there is some *a priori probability* (or simply *prior probability*) $P(C_1)$ that there is a storm, and some prior probability $P(C_2)$ that there is no storm. For instance, from the past weather records, if 15% of the days were found to be stormy during this season, then the meteorologist may assign $P(C_1) = 0.15$, and $P(C_2) = 0.85$. Now suppose the meteorologist has a barometer measuring a pressure x at 6 a.m. The meteorologist would like to obtain an *a posteriori probability* (or simply *posterior probability*) $P(C_1|x)$, that is, the conditional probability of having a storm on that day given the 6 a.m. pressure x. In essence, he would like to improve on his simple prior probability with the new information x.

The joint probability density $p(C_i, x)$ is the probability density that an event belongs to class C_i and has value x. The joint probability density can be written as

$$p(C_i, x) = P(C_i|x)p(x), \tag{2.79}$$

with $p(x)$ the probability density of x. Alternatively, $p(C_i, x)$ can be written as

$$p(C_i, x) = p(x|C_i)P(C_i), \tag{2.80}$$

with $p(x|C_i)$, the conditional probability density of x, given that the event belongs to class C_i. Equating the right hand sides of these two equations, we obtain

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}, \tag{2.81}$$

which is *Bayes theorem*. The previous form of Bayes theorem encountered in (2.8) was for two discrete variables x and y, whereas here we have a discrete variable C and a continuous variable x. Since $p(x)$ is the probability density of x without regard to which class, it can be decomposed into

$$p(x) = \sum_i p(x|C_i)P(C_i). \tag{2.82}$$

Substituting this for $p(x)$ in (2.81) yields

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{\sum_i p(x|C_i)P(C_i)}, \tag{2.83}$$

where the denominator on the right hand side is seen as a normalization factor for the posterior probabilities to sum to unity. Bayes theorem says that the posterior probability $P(C_i|x)$ is simply $p(x|C_i)$ (the *likelihood* of x given the

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

event is of class C_i) multiplied by the prior probability $P(C_i)$ and divided by a normalization factor. The advantage of Bayes theorem is that the posterior probability is now expressed in terms of quantities that can be estimated. For instance, to estimate $p(x|C_i)$, the meteorologist can divide the 6 a.m. pressure record into two classes and estimate $p(x|C_1)$ from the pressure distribution for stormy days, and $p(x|C_2)$ from the pressure distribution for non-stormy days.

For the general situation, the scalar x is replaced by a vector $\mathbf{x}$, and the classes are $C_1, \ldots, C_k$; then, Bayes theorem becomes

$$P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_i p(\mathbf{x}|C_i)P(C_i)}, \tag{2.84}$$

for $i = 1, \ldots, k$.

If instead of the discrete variable C_i, we have a continuous variable w, then Bayes theorem (2.81) for two continuous variables w and x takes the form

$$p(w|x) = \frac{p(x|w)p(w)}{p(x)}. \tag{2.85}$$

The scalars x and w can be generalized to the vectors $\mathbf{x}$ and $\mathbf{w}$, so Bayes theorem becomes

$$p(\mathbf{w}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{x})}. \tag{2.86}$$

Often a model controlled by some parameters $\mathbf{w}$ is used to model the variables $\mathbf{x}$. The model parameters $\mathbf{w}$ are to be estimated using a dataset D containing the observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Given a prior distribution $p(\mathbf{w})$ and $p(D|\mathbf{w})$ (i.e. the likelihood of observing D given the parameters $\mathbf{w}$), we can obtain a posterior distribution $p(\mathbf{w}|D)$ from Bayes theorem,

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}, \tag{2.87}$$

where $p(D)$ is simply a normalization factor,

$$p(D) = \int p(D|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}. \tag{2.88}$$

The likelihood $p(D|\mathbf{w})$ is treated differently by frequentists, who view the parameters $\mathbf{w}$ as being fixed. The frequentists commonly estimate the value of $\mathbf{w}$ by maximizing the likelihood function (Section 3.5). In contrast, Bayesians view the observed data D as fixed, but $\mathbf{w}$ is given by a distribution $p(\mathbf{w}|D)$.

Figure 2.12 illustrates the relation between $p(\mathbf{w}|D)$, $p(D|\mathbf{w})$ and $p(\mathbf{w})$ where, for simplicity, $\mathbf{w}$ is reduced to a scalar w. Case (a): Little prior information is available for w, that is, a very broad and flat $p(w)$. Case (b): More precise prior information is available from the narrower $p(w)$ distribution.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

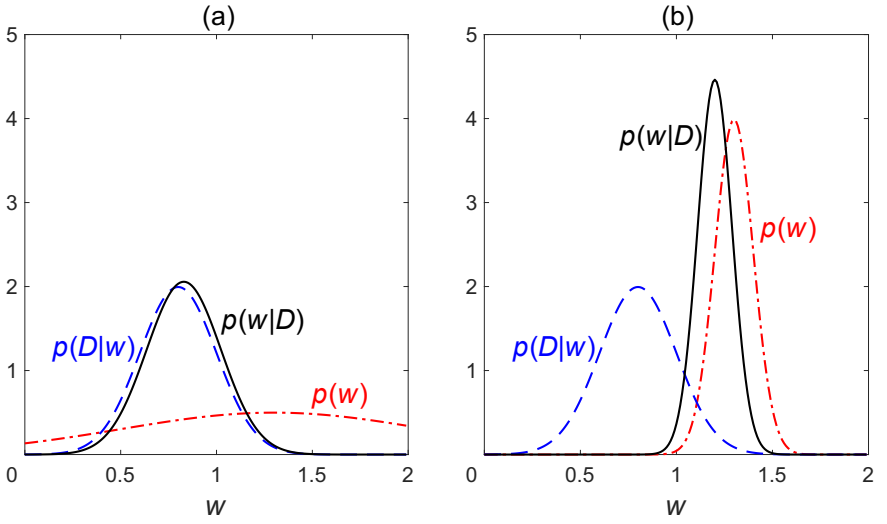

Figure 2.12 Relation between $p(w|D)$, $p(D|w)$ and $p(w)$. (a) A broad and flat distribution of $p(w)$ provides little prior information for estimating w, leading to the posterior distribution $p(w|D)$ being very similar to the likelihood $p(D|w)$. (b) A narrow $p(w)$ distribution leads to a larger difference between $p(w|D)$ and $p(D|w)$. If more data are available, $p(D|w)$ will be narrower and more strongly peaked than that shown in (b), and the $p(w|D)$ distribution will be pulled more towards $p(D|w)$. [Follows Cowan (2007)].

2.15 Classification A☺

Once the posterior probabilities $P(C_i|\mathbf{x})$ have been estimated from (2.84), we can proceed to classification: Given an input or predictor vector $\mathbf{x}$, called a *feature vector* in the ML literature, we choose the class C_j having the highest posterior probability, that is,

$$P(C_j|\mathbf{x}) > P(C_i|\mathbf{x}), \quad \text{for all } i \neq j. \tag{2.89}$$

From (2.84), this is equivalent to

$$p(\mathbf{x}|C_j)P(C_j) > p(\mathbf{x}|C_i)P(C_i), \quad \text{for all } i \neq j. \tag{2.90}$$

In the *feature space* (i.e. the space of the predictor variables $\mathbf{x}$), the pattern classifier has divided the space into *decision regions* $R_1, \ldots, R_k$, so that if a feature vector lands within R_i, the classifier will assign the class C_i. The decision region R_i may be composed of several disjoint regions, all of which are assigned the class C_i. The boundaries between decision regions are called *decision boundaries* or *decision surfaces*.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

To justify the decison rule (2.90), consider the probability P_{correct} of a new pattern being classified correctly:

$$P_{\text{correct}} = \sum_{j=1}^{k} P(\mathbf{x} \in R_j, C_j), \tag{2.91}$$

where $P(\mathbf{x} \in R_j, C_j)$ gives the probability that the pattern that belongs to class C_j has its feature vector falling within the decision region R_j, thus classified correctly as belonging to class C_j. Note that P_{correct} can be expressed as

$$\begin{aligned} P_{\text{correct}} &= \sum_{j=1}^{k} P(\mathbf{x} \in R_j | C_j) P(C_j), \\ &= \sum_{j=1}^{k} \int_{R_j} p(\mathbf{x}|C_j) P(C_j) \mathrm{d}\mathbf{x}. \end{aligned} \tag{2.92}$$

To maximize P_{correct}, one needs to maximize the integrand by choosing the decision regions so that $\mathbf{x}$ is assigned to the class C_j satisfying (2.90).

In general, classification need not be based on probability distribution functions, since in many situations, $p(\mathbf{x}|C_i)$ and $P(C_i)$ are not known. The classification procedure is then formulated in terms of *discriminant functions*, which tell us which class we should assign to the given $\mathbf{x}$. For example, in Fig. 2.13(a), $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$, and the two classes are separated by the line $x_2 = x_1$.

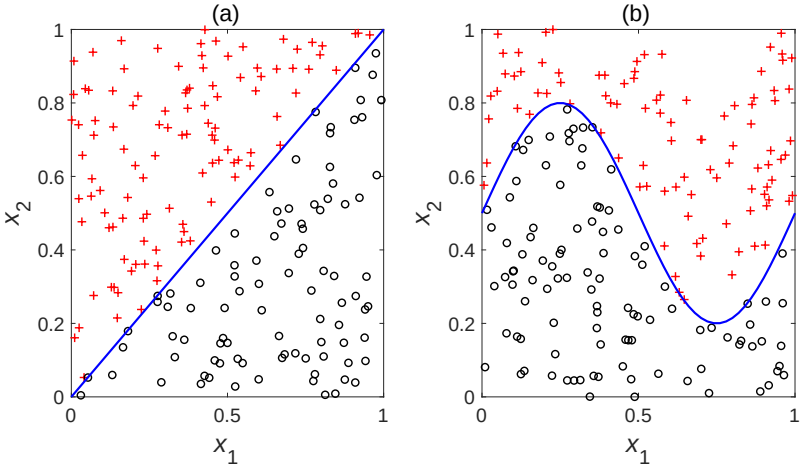

Figure 2.13 (a) A linear decision boundary separating two classes of data denoted by crosses and circles, respectively. (b) A non-linear decision boundary.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

The discriminant function can be simply $y(\mathbf{x}) = -x_1 + x_2$, with C_2 assigned when $y(\mathbf{x}) > 0$, and C_1 otherwise. Thus, the decision boundary is given by $y(\mathbf{x}) = 0$.

When there are more than two classes, the discriminant functions are $y_1(\mathbf{x}), \ldots, y_k(\mathbf{x})$, where a feature vector $\mathbf{x}$ is assigned to class C_j if

$$y_j(\mathbf{x}) > y_i(\mathbf{x}), \quad \text{for all } i \neq j. \tag{2.93}$$

Equation (2.89) can be viewed as a special case of (2.93). An important property of a discriminant function $y_i(\mathbf{x})$ is that it can be replaced by $f(y_i(\mathbf{x}))$ for any monotonic function f, since the classification is unchanged as the relative magnitudes of the discriminant functions are preserved by f. There are many classical *linear discriminant analysis* methods (Duda et al., 2001), where the discriminant function is a linear combination of the predictor variables x_l, that is,

$$y_i(\mathbf{x}) = \sum_l w_{il} x_l + w_{i0} \equiv \mathbf{w}_i^{\mathrm{T}} \mathbf{x} + w_{i0}, \tag{2.94}$$

with parameters $\mathbf{w}_i$ and w_{i0}. Based on (2.93), the decision boundary between class C_j and C_i is obtained from setting $y_j(\mathbf{x}) = y_i(\mathbf{x})$, yielding a hyperplane[5] decision boundary described by

$$(\mathbf{w}_j - \mathbf{w}_i)^{\mathrm{T}} \mathbf{x} + (w_{j0} - w_{i0}) = 0. \tag{2.95}$$

Suppose $\mathbf{x}$ and $\mathbf{x}'$ both lie within the decision region R_j. Consider any point $\tilde{\mathbf{x}}$ lying on a straight line connecting $\mathbf{x}$ and $\mathbf{x}'$, that is,

$$\tilde{\mathbf{x}} = a\mathbf{x} + (1 - a)\mathbf{x}', \tag{2.96}$$

with $0 \leq a \leq 1$. Since $\mathbf{x}$ and $\mathbf{x}'$ both lie within R_j, they satisfy $y_j(\mathbf{x}) > y_i(\mathbf{x})$ and $y_j(\mathbf{x}') > y_i(\mathbf{x}')$ for all $i \neq j$. Since the discriminant function is linear, we also have

$$y_j(\tilde{\mathbf{x}}) = a y_j(\mathbf{x}) + (1 - a) y_j(\mathbf{x}'), \tag{2.97}$$

therefore $y_j(\tilde{\mathbf{x}}) > y_i(\tilde{\mathbf{x}})$ for all $i \neq j$. Thus, any point on the straight line joining $\mathbf{x}$ and $\mathbf{x}'$ must also lie within R_j, meaning that the decision region R_j is simply connected and convex. As we shall see later, with non-linear ML methods such as neural networks and support vector machines, the decision boundaries can be curved surfaces (Fig. 2.13(b)) instead of hyperplanes, and the decision regions need not be simply connected nor convex. Discriminant analysis and other classification methods are explained in detail in Chapter 12.

2.16 Clustering A☺

In machine learning, there are two general approaches, *supervised learning* and *unsupervised learning*. An analogy for the former is students in a Spanish class

[5] A hyperplane is a subspace where the dimension is one less than that of its ambient space.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

where the teacher demonstrates the correct Spanish pronunciation. An analogy for the latter is students working on a team project without supervision. In unsupervised learning, the students are provided with learning rules, but must rely on self-organization to arrive at a solution, without the benefit of being able to learn from a teacher's demonstration.

In classification, the training dataset consists of predictors or features $\mathbf{x}_i$ ($\mathbf{x}$ can be made up of continuous and/or discrete and/or categorical variables) and discrete/categorical response variables C_i, $i = 1, \ldots, N$, N being the number of observations. Here, C_i serves the role of the teacher or *target* for the classification model output $\tilde{C}_i$, that is, $\tilde{C}_i$ is fitted to the given target data, similar to students trying to imitate the Spanish accent of their teacher; thus, the learning is supervised.

For instance, suppose $\mathbf{x}$ contains three variables – air temperature, humidity and pressure, and C can be 'no precipitation', 'rain' or 'snow' a day later. Such a classification model uses three meteorological inputs to predict whether it will be 'no precipitation', 'rain' or 'snow' a day later.

Clustering or cluster analysis is the unsupervised version of classification, that is, we are given the $\mathbf{x}$ data but not the C data. The goal of clustering is to group the $\mathbf{x}$ data into a number of subsets or 'clusters', such that the data within a cluster are more closely related to each other than data from other clusters. After performing clustering on the air temperature, humidity and pressure data, we may indeed find three main clusters. The first cluster of data points may occur where humidity is low and pressure is high, corresponding to days of no precipitation. A second cluster may occur where humidity is high, pressure is low and temperature is high, corresponding to rainy days, while a third cluster may be somewhat similar to the second cluster but occurring at low temperature, corresponding to snowy days. Thus, even without the target C data, we can learn much from the $\mathbf{x}$ data alone.

A simple and popular method for performing clustering is *K-means clustering*. First choose K, the number of clusters. Next, start with initial guesses for the mean positions of the K clusters in the $\mathbf{x}$ space (i.e. the position of a cluster centre, a.k.a. *centroid*, is simply the mean position of all the data points belonging to that cluster). Iterate the following two steps until convergence:

(i) For each data point, find the closest centroid [based on the squared Euclidean distance in (10.2)] and assign the data point to be a member of this cluster.

(ii) For each cluster, reassign the centroid to be the mean position of all data points belonging to that cluster.

This is known as Lloyd's algorithm, and is sometimes referred to as 'naive K-means' as there are faster algorithms. The initial choice for the K centroids often involves randomly picking K data points from the $\mathbf{x}$ data. The initialization can be improved, for example by using the K-means++ method of Arthur and Vassilvitskii (2007), where the centroids are chosen randomly from the data points, but with the probability of choosing a data point being proportional

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

to its squared distance from the closest centroid (among the centres already chosen). See Section 10.2.1 for more details on K-means clustering and Chapter 10 for more choices of clustering methods.

Figure 2.14 illustrates K-means clustering with $K = 3$ clusters for the daily air pressure and temperature data at Vancouver, BC, Canada during 2013–2017. As air pressure and temperature have different units, clustering was performed on the standardized data. The poor initial choice of the centroids did not hinder the convergence of the clustering algorithm. Upon convergence in Fig. 2.14(d), the three centroids make physical sense: In the Pacific Northwest region of North America, summer tends to be sunny while winter has numerous weather systems passing through, so it makes sense to have one cluster representing summer and two clusters representing winter. In winter, during high pressure days, the clear skies increase outgoing long-wave radiation, leading to colder temperatures (as characterized by the the lower-right centroid) than during low pressure days where the clouds reduce the outgoing long-wave radiation (lower-left centroid). As K, the number of clusters, is specified by the user, choosing a different K will in general lead to very different clusters. It turns out $K = 3$ is optimal according to two internal evaluation criteria (see Fig. 10.1).

2.17 Information Theory 🅱☺

Information theory studies the quantification, storage and communication of information. The field started in 1948 with the publication of an article in two parts by Claude E. Shannon (Shannon, 1948a,b), on how best to encode information for transmission. His theory was based on probability theory, and the central concept of information entropy, a measure of the uncertainty in a message, was surprisingly similar to the thermodynamic entropy developed by the physicists Ludwig Boltzmann and J. Willard Gibbs in the 1870s. Information theory has since grown into a large field (Cover and Thomas, 2006) and is connected to machine learning (MacKay, 2003). Information theory has also been applied increasingly to the environmental sciences, especially regarding predictability of dynamical systems (Leung and North, 1990; Kleeman, 2002; DelSole, 2004; DelSole, 2005; Y. M. Tang et al., 2005; DelSole and Tippett, 2007) and the selection of good predictors from a large pool of available predictor variables (Sharma, 2000; May, Dandy, et al., 2008; May, Maier, et al., 2008).

How information is connected to probability can be illustrated by the following example. The statement 'the sun has risen from the east' is much less informative than the statement 'a magnitude 9.0 earthquake has occurred off Japan'. The reason is that the first event has probability 1 (so the statement contains no useful information) while the second event has low probability (so the statement contains potentially life-saving information). Thus, information content is low when the probability of the event occurring is high.

First, start with a discrete random variable X, from which we can draw specific values x. We want to develop a measure of information content, $h(x)$,

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

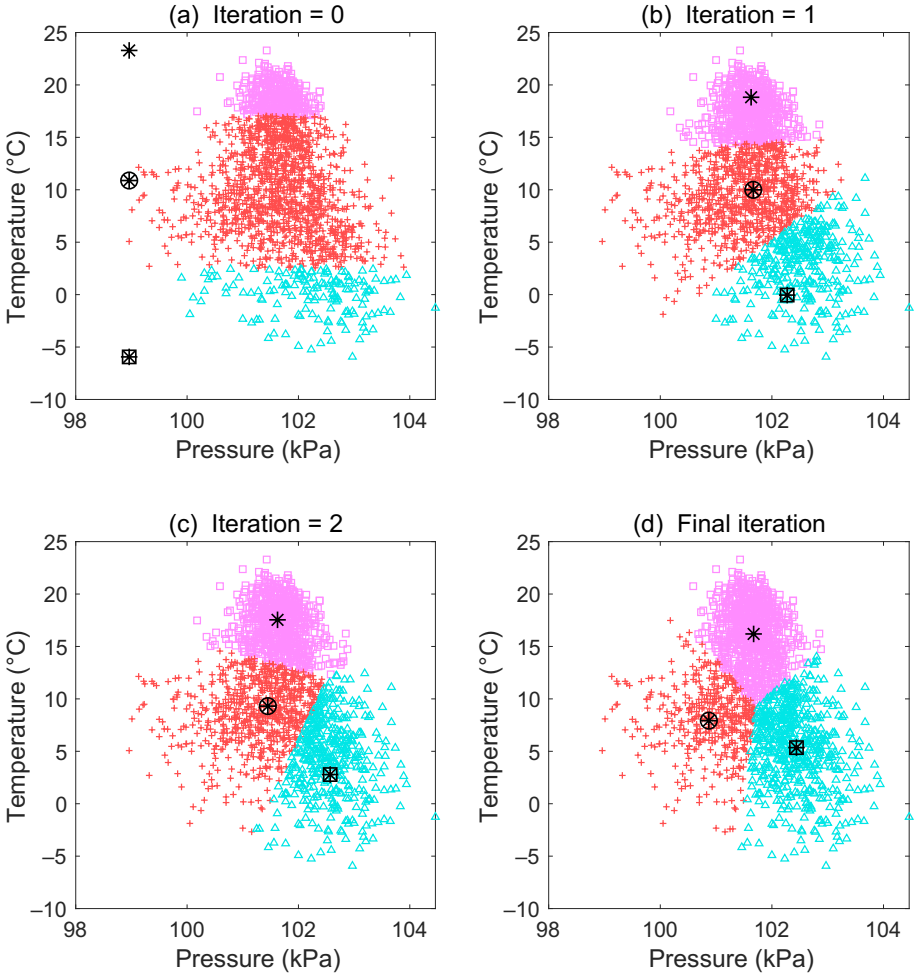

Figure 2.14 (a) The initial guesses for the three centroids are marked by three asterisks. The data points are assigned to clusters based on their nearest centroid. In (b), the centroids have been recalculated based on the mean position of the cluster members in (a), and cluster members in (b) have been reassigned based on their closeness to the centroids in (b). The location of the centroids and their associated cluster members are shown after (c) two iterations and (d) after final convergence of the K-means clustering algorithm.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

where $h(x)$ is a monotonically decreasing function of the probability $P(x)$. If there are two independent events x and y, the information from observing both events should be the sum of the two separate events, that is,

$$h(x, y) = h(x) + h(y). \tag{2.98}$$

Since the probability of observing two independent events obeys $P(x, y) = P(x)\,P(y)$, we can choose $h(x)$ to be

$$h(x) = -\log P(x), \tag{2.99}$$

so that

$$h(x, y) = -\log P(x, y) = -\log[P(x)P(y)] \tag{2.100}$$

$$= -\log P(x) - \log P(y) = h(x) + h(y). \tag{2.101}$$

The minus sign in (2.99) ensures $h(x) \geq 0$ and h is a monotonically decreasing function of P. One can choose any base for the logarithm function, but the two most common choices are base 2 and base e, that is, $\log_2$ or $\log_e$ (ln). With base 2, $h(x)$ has units of *bits* (from 'binary digits'), whereas with base e (i.e. using natural logarithm), $h(x)$ has units of *nats*.

2.17.1 Entropy B☺

Since X is a random variable, we are more interested in the average amount of information transmitted, that is, the expectation of $h(x)$ than $h(x)$ itself. The expectation of of (2.99) involves summing $h(x)$ weighted by the probability $P(x)$ over all possible states of x, that is,

$$H(X) = \mathrm{E}[h(x)] = -\sum_x P(x) \log P(x) = -\sum_i P_i \log P_i, \tag{2.102}$$

where $H(X)$ is called the *entropy* and P at the discrete values of x is also written as P_i. For any x with $P(x) = 0$, we will set $P \log P = 0$ since $\lim_{P \to 0} P \log P = 0$.

Readers familiar with the statistical mechanics of Boltzmann and Gibbs will recall that the thermodynamic entropy S is given by

$$S = -k_\mathrm{B} \sum_i P_i \log P_i, \tag{2.103}$$

where k_B is the Boltzmann constant and the summation is over all i states. By switching to natural units, k_B becomes unity, and (2.103) is identical in form to (2.102).

Next, consider the simple example of X being a binary variable (0 or 1). If we write $P(1) = \alpha$ and $P(0) = 1 - \alpha$, then (2.102) gives

$$H(X) = -[P(0) \log_2 P(0) + P(1) \log_2 P(1)] \tag{2.104}$$

$$= -[(1 - \alpha) \log_2 (1 - \alpha) + \alpha \log_2 \alpha]. \tag{2.105}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

The entropy is maximized (Fig. 2.15) when $\alpha = 0.5$, that is, $P(1) = P(0) = 0.5$, and minimized ($H = 0$) when $P(1) = 0$ or 1. For the coin flip analogy, $P(1) = 0$ or 1 means the coin is loaded to always come out 'head' (H) or 'tail' (T). Since the outcome is certain, more flips of the coin do not provide any new information, that is, the information content is 0 at $H = 0$. In contrast, when $\alpha = P(1) = P(0) = 0.5$, there is greatest uncertainty since there is equal probability of getting H or T; thus, information on the outcome of coin tosses is most valuable.

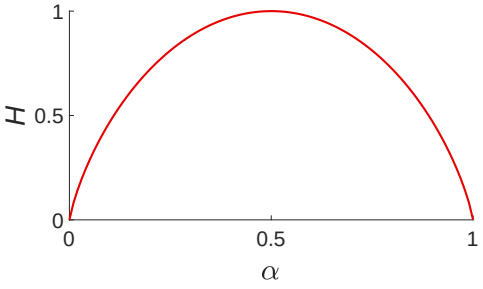

Figure 2.15 Entropy H as a function of α. When $\alpha = 0.5$, the maximum ($H = 1$) is attained.

It is straightforward to define entropy for continuous random variables. With $\mathbf{X}$ denoting a random real vector, the entropy is given by

$$H(\mathbf{X}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{2.106}$$

where the natural logarithm is commonly used and the summation in (2.102) is replaced by an integration over the entire domain of $\mathbf{x}$. However, when computing with sampled data, continuous variables are commonly discretized or quantized, that is, dividing the domain of each variable into bins and counting how many data points fall within each bin to obtain a histogram and, thereby, a sample discrete probability distribution. We will continue our discussion using the discrete variable formulation.

2.17.2 Joint Entropy and Conditional Entropy B☺

Suppose there are two random variables X and Y from which we can draw specific values x and y with joint probability $P(x, y)$. The *joint entropy* between the two random variables is

$$H(X, Y) = \mathrm{E}[-\log P(x, y)] = -\sum_x \sum_y P(x, y) \log P(x, y). \tag{2.107}$$

If the value of x is already known, then the additional information needed to specify y is $-\log P(y|x)$. The average additional information needed to specify y is the *conditional entropy*

$$H(Y|X) = \mathrm{E}[-\log P(y|x)] = -\sum_x \sum_y P(x, y) \log P(y|x). \tag{2.108}$$

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

Using (2.6), it can easily be shown that

$$H(X,Y) = H(X) + H(Y|X), \tag{2.109}$$

that is, the information needed to describe both X and Y is the information needed to describe X plus the additional information needed to describe Y after X is known.

2.17.3 Relative Entropy B 😐

Suppose the unknown true probability distribution is $P(x)$ and we have obtained $Q(x)$, an approximation of the true distribution. A measure of the dissimilarity between the two distributions is the *relative entropy*, also known as the *Kullback–Leibler* (KL) *divergence*, where

$$D_{\mathrm{KL}}(P||Q) = \mathrm{E}\left[\log \frac{P(x)}{Q(x)}\right] = \sum_x P(x) \log \frac{P(x)}{Q(x)}\,. \tag{2.110}$$

This quantity is referred to as a divergence instead of a distance as it is asymmetric, that is, $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$ in general.

$$D_{\mathrm{KL}}(P||Q) = -\sum_x P(x) \log Q(x) + \sum_x P(x) \log P(x) = H_{\mathrm{c}}(P,Q) - H(P), \tag{2.111}$$

where $H_{\mathrm{c}}(P,Q)$ is called the *cross-entropy*, with

$$H_{\mathrm{c}}(P,Q) = -\sum_x P(x) \log Q(x). \tag{2.112}$$

It can be proven that $D_{\mathrm{KL}}(P||Q) \geq 0$, and $D_{\mathrm{KL}}(P||Q) = 0$ if, and only if, $P = Q$ (Cover and Thomas, 2006). In coding theory (Cover and Thomas, 2006), the cross-entropy $H_{\mathrm{c}}(P,Q)$ is the expected number of bits needed to encode data from a source with distribution P while we use model Q to define our codebook. The entropy $H(P)$ is the expected number of bits if we use the true model; thus, the relative entropy $D_{\mathrm{KL}}(P||Q)$ can be interpreted as the expected number of extra bits that must be communicated if a code that is optimal for the incorrect distribution Q is used instead of using a code based on the true distribution P.

Relative entropy has been used in studies of predictability of dynamical systems (Kleeman, 2002; DelSole, 2004), including the predictability of the El Niño-Southern Oscillation (ENSO) (see Section 9.1.5), the dominant mode of interannual climate variability in the equatorial Pacific, with global implications (Y. M. Tang et al., 2005). For instance, Q can be the distribution from climatology (i.e. the expected behaviour from historical observed data) while P can be the distribution from the predictions by a numerical model. The relative entropy $D_{\mathrm{KL}}(P||Q)$ is a measure of the additional information provided by the prediction model over the information from climatology (Kleeman, 2002).

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.17.4 Mutual Information B 😐

Given two random variables X and Y, if we want to detect any linear relation between the two, we can compute the correlation (Section 2.11). But what if the relation is non-linear? Correlation can completely miss a strong non-linear relation. To detect linear and non-linear relations, one can turn to *mutual information* (MI), which determines, via the KL divergence, how dissimilar the joint distribution $P(X, Y)$ is to $P(X)\,P(Y)$, that is, MI is given by

$$I(X, Y) = D_{\mathrm{KL}}(P(X, Y) || \, P(X)\,P(Y)) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \,. \tag{2.113}$$

As D_{KL} is non-negative, $I(X, Y) \geq 0$. When X and Y are independent variables, $P(X, Y) = P(X)\,P(Y)$. The logarithm term in (2.113) becomes $\log 1 = 0$, giving $I(X, Y) = 0$. Thus, the minimum MI occurs when X and Y are independent. As X and Y become more dependent, MI increases.

From (2.5), we can substitute $P(x, y) = P(y|\, x)\,P(x)$ into (2.113), giving

$$I(X, Y) = -\sum_x \sum_y P(x, y) \log P(y) + \sum_x \sum_y P(x, y) \log P(y|\, x) \tag{2.114}$$

$$= -\sum_y P(y) \log P(y) + \sum_x \sum_y P(x, y) \log P(y|\, x), \tag{2.115}$$

that is,

$$I(X, Y) = H(Y) - H(Y|\, X), \tag{2.116}$$

with $H(Y|\, X)$ being the conditional entropy. Thus, MI can be viewed as the reduction in the uncertainty in Y when the value of X is known. Similarly,

$$I(X, Y) = H(X) - H(X|\, Y). \tag{2.117}$$

A problem with applying the information theory approach to continuous variables is that it is difficult to estimate continuous probability density distributions. When computing with sampled data, continuous variables are often discretized or quantized, that is, dividing the domain of each variable into bins and counting how many data points fall within each bin to obtain a histogram as a discrete approximation of the probability density distribution. The final result unfortunately depends on the choice of the bin width – having a wide bin width gives a crude approximation of a continuous distribution, while having a narrow bin width means there are few data points within each bin, thus a noisy histogram. Reshef et al. (2011) has proposed a way to deal with the bin width problem. An alternative approach is to use kernel density estimation (see Section 3.13) to estimate the distribution, but instead of the bin width there is now an adjustable width parameter of the kernel function. A more recent approach using K-nearest neighbour distances to estimate MI (Kraskov et al., 2004) has seen increasing usage.

When there are many available predictors or features, one may want to select the most relevant predictors before building a prediction model. Traditionally, a

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

common approach is to compute the correlation between a predictor variable and the response variable and select the predictors with high correlation. However, predictors non-linearly related to the response variable may be missed using this selection procedure. MI has been proposed as a better measure for predictor selection, as it is not restricted to detecting linear relations (H. C. Peng et al., 2005). In environmental sciences, MI has been proposed for predictor selection in hydrological studies (Sharma, 2000; May, Maier, et al., 2008; May, Dandy, et al., 2008).

Exercises

Some exercises involve working with data files, which are downloadable from our book website (web link given in the Preface).

2.1

In a tropical Atlantic region, the number of occurrences when the daily sea surface temperature condition is cool, normal or warm and when the wind condition is calm or stormy have been recorded in the table below. What is the probability of a day being (a) warm and stormy and (b) cool and stormy? What is the probability of the day being (c) stormy if it is a warm day and (d) stormy if it is a cool day?

	cool	normal	warm
calm	1805	3661	2012
stormy	32	125	228

2.2

A variable y is measured by two instruments placed 50 km apart in the east–west direction. Values are recorded daily for 100 days. The autocorrelation function of y shows the first zero crossing (i.e. the smallest lag at which the autocorrelation is zero) occurring at six days (for both stations). Furthermore, y at one station is correlated with the y at the second station, with the second time series shifted in time by various lags. The maximum correlation occurred with y from the eastern station lagging y from the western station by two days. Assuming a sinusoidal wave is propagating between the two stations, estimate the period, wavelength, and the speed and direction of propagation.

2.3

Prove that the expectation of the sample mean in (2.25) equals the population mean.

2.4

Given two variables x and y with zero population means: (a) Show that the population covariance $\mathrm{cov}(x, y) = \mathrm{E}[xy]$ is zero if x and y are independent. (b) However, the converse is not true in general. Given x uniformly distributed in $[-1,\ 1]$ and $y = x^2$, show that $\mathrm{cov}(x, y)$ is zero even though x and y are not independent.

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press

2.5

Using the data file provided on the book website, compare the Pearson correlation with the Spearman and Kendall rank correlations for the time series x and y (each with 40 observations). Repeat the comparison for the time series x_2 and y_2 (from the same data file as above), where x_2 and y_2 are the same as x and y, except that the fifth data point in y is replaced by an outlier in y_2. Repeat the comparison for the time series x_3 and y_3, where x_3 and y_3 are the same as x and y, except that the fifth data point in x and y is replaced by an outlier in x_3 and y_3. Make scatterplots of the data points in the x–y space, the x_2–y_2 space and the x_3–y_3 space. Also plot the linear regression line in the scatterplots.

2.6

Analyse the monthly sea surface temperature anomalies (i.e. deviations from the mean) for the Niño1+2 region in the eastern equatorial Pacific (0°–10°S, 80°W–90°W) and the Niño 3.4 region in the central equatorial Pacific (5°S–5°N, 170°W–120°W) (shown in Fig. 9.3 and data downloadable from our book website):
(a) For each variable, compute the histogram and compare to the Gaussian distribution fit to the data. (b) For each variable, compute the quantile–quantile plot relative to the standard Gaussian distribution. (c) Compute the quantile–quantile plot between Niño1+2 and Niño3.4 anomalies. (d) Compute the boxplot for the two variables. [Data source: Climatic Research Unit, University of East Anglia]

2.7

In addition to the Niño1+2 and Niño3.4 sea surface temperature anomalies, analyse the Southern Oscillation Index (SOI) (Tahiti pressure minus Darwin pressure, standardized to zero mean and unit standard deviation): (a) For each of the three time series, plot the Pearson autocorrelation function. (b) Compute the Pearson correlation and the Spearman and Kendall rank correlations between Niño1+2 and Niño3.4 anomalies. (c) Compute the Pearson correlation and the Spearman and Kendall rank correlations between Niño1+2 and SOI. (d) Compute the Pearson correlation and the Spearman and Kendall rank correlations between Niño3.4 and SOI. [Data source: Climatic Research Unit, University of East Anglia]

2.8

Suppose a test for the presence of a toxic chemical in a lake gives the following results: if a lake has the toxin, the test returns a positive result 99% of the time; if a lake does not have the toxin, the test still returns a positive result 2% of the time. Suppose only 5% of the lakes contain the toxin. What is the probability that a positive test result for a lake turns out to be a false positive?

2.9

Use K-means clustering to analyse the dataset containing air temperature and humidity at Vancouver, BC, Canada from 2013–2017. Try $K = 2$ and 3,

and try to explain what the clusters represent. [Data source: weatherstats.ca based on Environment and Climate Change Canada data]

2.10

A biological oceanographer collected 100 water samples at various locations, from which the water temperature (T), the nitrate concentration (N), the silicate concentration (S) and the concentration of a marine microorganism (M) were measured. The measurements were discretized into 'below normal', 'normal' and 'above normal', that is, 1, 2 and 3, respectively. The observed number of occurrences are given in the table below. Which of the three environmental variables has the strongest relation with M? Try to determine this using mutual information and using Pearson and Spearman correlation.

$M\backslash T$	1	2	3	$M\backslash N$	1	2	3	$M\backslash S$	1	2	3
1	2	38	3	1	13	9	3	1	3	7	5
2	5	8	5	2	8	33	7	2	13	26	12
3	18	4	17	3	4	8	15	3	9	17	8

https://doi.org/10.1017/9781107588493.003 Published online by Cambridge University Press